統計学入門

劉　慶豊[1]
小樽商科大学

平成21 年10 月5 日

[1]E-mail: qliu@res.otaru-uc.ac.jp

1　統計学とは何：統計学とデータ

定義 1 (全く不厳密だが私なりの定義) 統計学はデータを解析する学問で、データ *(*ある種の情報）を根拠に数学的な方法を用いて、不確実性を持った自然・社会現象の背後にある規則を発見し解明することを目的とする。応用数学の一分野として数学ではあるが、応用面から言うと極簡単な数学しか使わない。暗記できる数学、ルールによる数学であり、外国語と同じくある種の言語である、「文法」さえ覚えればマスタできる。

1.1　統計学の応用例

幾つか架空の応用例をあげて、統計学の役割を理解してもらう。

1.1.1　応用例一：天気情報にある降水確率

何故「明日の降水確率は 60%」って分かるのか？

- 過去の毎日の天気状況を把握しておく：過去の観測、解析結果と天気状況に関する情報の収集と蓄積
- 明日の天気の状況を推測する：衛星観測、観測結果の解析：雲の状態、風の強さ、周辺の天気状況など
- 確率の計算：過去において似たような状況では何回中何回１時間当たり１ mm 以上雨が降ったのかを計算する

例えば、明日とほぼ同じ天気状況であった日は過去において 100 日あったとしよう、その 100 日の中 60 日降水があったら、「明日の降水確率が 60%」

1.1.2　応用例二：ダイエットグッズの宣伝の真実

おもてに出た情報：

- A さん、使用前７８キロ、一ヶ月間使用後６０キロ
- B さん、使用前６５キロ、一ヶ月間使用後５５キロ

一つの疑い：１００人を使用させてその中から体重が減った A さんと B さんを選んだ。

- A さん出産後１５キロ減

- B さん大怪我して入院していた

本当に開示してほしいのは統計データ

- 使用人数、使用期間
- 平均体重（使用前後)、平均効果（平均的に体重が幾ら減ったのか)：平均による比較の任意さ。
- 分散（体重減の散らばり具合、人によって体重の変化がどう違うのか）
- より客観的な結論付け。統計理論に基づいた検定結果：本当に効果があったのか。

1.1.3　応用例三：商品市場に関するアンケート調査の解析

市場細分のための ×××boy、××× ステーションと W×× などのゲーム機市場のアンケート調査

アンケート項目

ID	性別	年齢	身長	体重	購入時期	現在の使用頻度	職業	学歴	年収	購入機種型番
1										
2										
3										

既知の商品情報

型番	単価	バッテリの性能	折り畳み式か	ゲームの種類	ゲームの特徴	···
DI						
DII						
DIII						

例えば、若い OL 向けの新商品を開発したい場合、商品の特性と購買者の特性の間の因果関係を調べ、どのように既存の商品を改良すれば若い OL が好むのかを分析する。新商品の開発を企画する時の参考となる。

1.1.4　応用例四：品質管理、管理図の原理

ボルトの生産流れ作業の制御：サンプルの平均に異常が発生した時生産をストップし、生産ラインを調整する。

異常と判断する臨界値を決めるのは統計学の役割。

1.1.5　応用例五：美容室の割引制度の効果

某々ヘアーサロン

1.1.6　応用例六：立地分析

○○チェーン店新しく出店したい
既存の店に関するデータを収集する、

1.1.7　やや分かりにくい応用例：最適ポートフォリオの決定

一番単純でよく使われる方法：
各資産（株）の収益率の分散と共分散を推定して、それを情報として利用し、ターゲット収益率のもとで、収益率の分散を最小にする組み合わせを割り出す。そのような組み合わせは最適ポートフォリオとなる。
このような最適ポートフォリオ導出するため利用するテクニックは統計学的なツールである。

1.1.8　挙げきれない例

化粧品や薬の効果に関する検証、国民経済の予測、老齢化社会に関する予測、ロケットの軌道の予測、株価と為替レートとの関係、来店客数の予測、売上の予測などなど。

1.2　統計学的に纏めてみよう

1.2.1　データの種類

1. 質的なデータ：性質を表すデータ、数値だけでは表せない。

   例：赤か緑か、男か女か、大学生か高校生か、雨が降ったか降らなかったか、太ったか痩せたか、薬を飲んだか飲まなかったか · · ·

2. 量的なデータ：数量を表すデータ、数値で表せる。

   例：降水量は何ミリ、体重は何キロ、売り上げは何円 · · ·

1.2.2　母集団と標本

1. 母集団：分析したい対象の全体を母集団と呼ばれる。

   例：人類存在以来の天気、薬を飲んだすべての人、ゲーム機を買ったすべての人 · · ·

2. 標本：母集団の一部。

   50 年前から今日までの天気、三種類のゲーム機を買った人 · · ·

1.2.3　データを収集する方法

1. 全数調査：母集団全体に関して調査する。
2. 標本調査：選んだ標本に関して調査する。

2　前回のおさらいと補足

- 講義概要
- 統計学の定義
- ダイエットグッズ等の例
- Excel の基本操作表の作り方の練習

3　表の作り方の練習

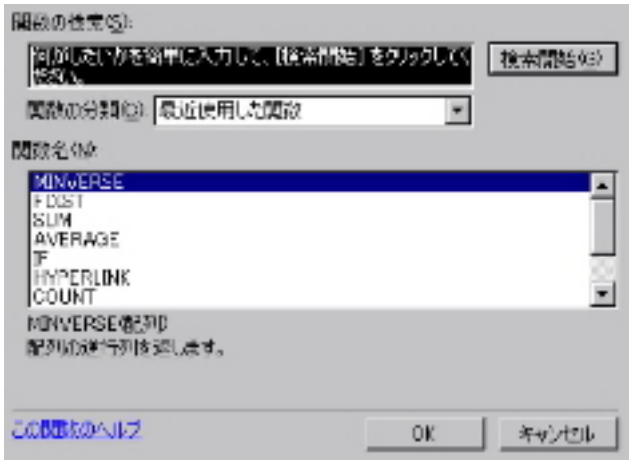
Excel のいろいろな関数に慣れる。複雑な表を作成し、デザインを工夫する。

3.1　いろいろな関数

sum,mean,max,min,round,if... などの関数の使い方を覚える。

入力したいセルをアクティブに（選択）して、ツール場にある Σ ボタンの横にある ▼ をクリックしてください。上にあるメニューが表示される。メニューの中にいろいろな関数がある。

その一番下にある [その他の機能] を押したら、次のウィンドが表示される。



枠の中に全ての関数が入っている。どれかをクリックしたらその関数に関する説明が下に表示される。また使いたい関数は画面の中にある指示に従えば検索できる。

以下で幾つかよく使われる関数について説明する

SUM 総和を計算する。=sum(始まりのセルの番地:終わりのセルの番地)。例:=sum(a1:a5)。

AVERAGE 平均を計算する。入力のルールは sum と殆ど同じ。。

MAX 最大値を求める。同上。

MIN 最小値を求める。同上。[1]

ROUND 指定した小数点以下の桁数まで四捨五入する。=round(数字またはセルの番地、桁数)。例：=round(123.246,2)、=round(b1,3)。

[1]少なくともここまで頑張って理解してください。

	A	B	C	D
1	2	1.264	入力法	出力
2	3		=sum(a1 :a9)	35
3	6		=average(a1 :a9)	3.888889
4	3		=max(a1 :a9)	7
5	4		=min(a1 :a9)	2
6	5		=round(123.246,2)	123.25
7	7		=round(b1 ,2)	1.26
8	2		=if(a1 >2,1 ,0)	0
9	3		=if(a1 =2,"準優勝","")	準優勝

IF 条件を判断する。=if(条件式、真であれば表示する値、偽であれば表示する値)。
例：=if(a1>2,1,0)、=if(a1=1,"優勝","")。

3.2　表のデザイン（参考）

- 文字の色やサイズを変えてみる。
- 罫線：セルや表を選択 > メニューバー > 書式 > セル > 罫線 > 外枠（内枠)。罫線の太さや色なども変えられる。
- 列や行の書式：一つの列または複数の列を選択 > 列 > 幅など。
- などなど

練習 2 *http://www.geocities.jp/qfliuwind/*にあるデータ *DATA01* を使って、最小値、最大値、平均体重、標準体重、標準体重との差の項目を設けて[2]さらに各生徒に標準、標準ではないという二つのランクも付けて表を作って下さい。ただし、標準体重 = (身長$^2/10000) * 22$、二つのランクの決め方は標準体重との差を x とし、標準 $(-5 <= x <= 5$ 同じく $abs(x) <= 5)$、標準ではないは標準のその他で $(x < -5$ または $x > 5)$ とする[3]。時間があれば、標準、痩せ、肥満三つのランクを自分で決めて、表を作ってみてください。出来るところまでで良い、結果を *Word* に貼り付けてプリントアウトし提出してください。

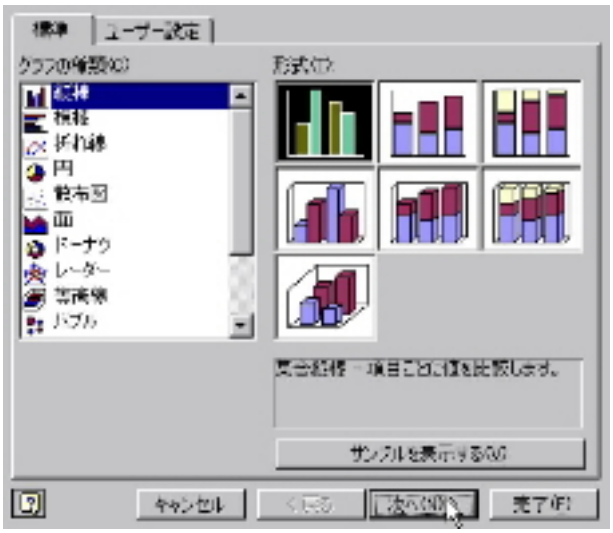
[2]少なくともここまで頑張って完成させてください。
[3]ここで付けた体重のランクはあくまでも教育のためで、科学的な根拠がない。

3.3　グラフを作ってみよう

ヒストグラムの作成の準備として、以下のデータ DATA02 を使ってグラフの作り方を説明する。グラフウィザードをクリック > グラフの種類を選択 > データの範囲を選択 (データ範囲（D)=Sheet1!A1:E7、または =a1:e7) > タイトルなどを入力 > 作成場所を選ぶ > 完了。完成したグラフの修正ができる。

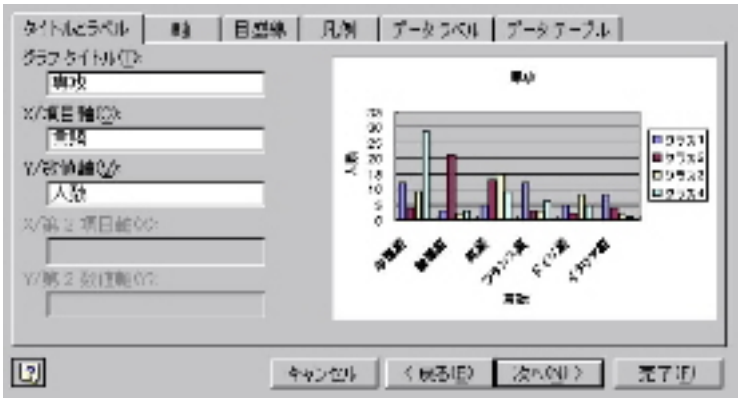
	A	B	C	D	E
1	専攻	クラス1	クラス2	クラス3	クラス4
2	中国語	12	4	9	29
3	韓国語	3	21	2	3
4	英語	5	13	15	9
5	フランス語	12	3	3	6
6	ドイツ語	5	2	8	5
7	イタリア語	8	4	2	1

グラフウィザードをクリック



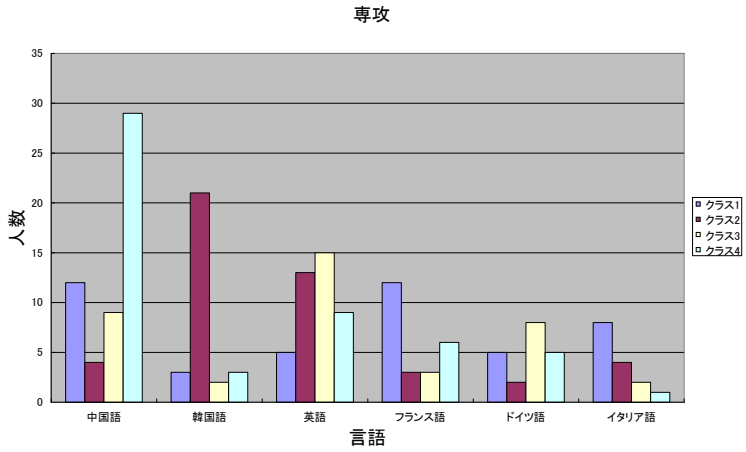
グラフの種類を選択

(データ範囲（D)=Sheet1!A1:E7、または
=a1:e7)



タイトルなどを入力

作成場所を選ぶ



完了

練習 3 同じデータを使って、グラフのタイプタイプとデータの範囲を変えながら一種類以上の図を作成してください。プリントアウトし提出してください。

練習 4 *DATA01* の身長のデータを使って、散布図を作成し提出してください。

4　データを視覚的に分かりやすくするための度数分布表とヒストグラム

大学生男子 50 人の身長のデータ（DATA01）をそのままながめていても目がちかちかするだけでわかることは少ない。イメージが付きやすくするために度数分布表とい

うものを作ってみましょう。

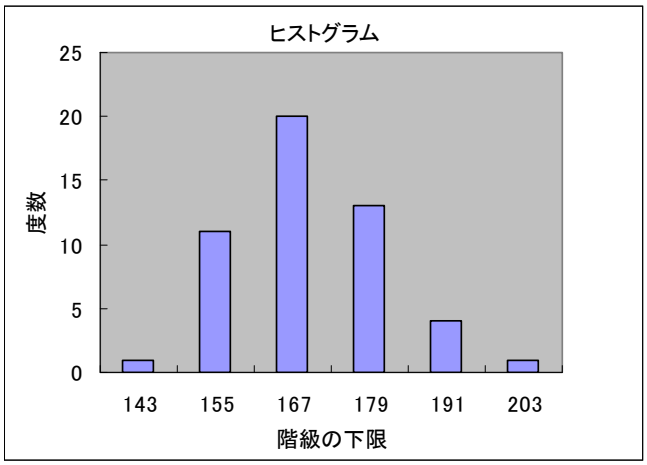
分布表の例

階級	度数	累積度数	相対度数	累積相対度数
143−152	9	9	18%	18%
152−161	10	19	20%	38%
161−170	14	33	28%	66%
170−179	12	45	24%	90%
179−188	2	47	4%	94%
188−198	3	50	6%	100%

定義 5 度数：各階級に入っているデータの数．相対度数：度数/全体のデータ数。

定義 6 累積度数：下の階級からの度数の合計。相対累積度数：累積度数/全体のデータ数。

さらにそれを棒グラフにして、視覚的にもっと分かりやすくなる。このような各階級の度数を棒グラフにしたものをヒストグラムという。



練習 7 *DATA01* の中にある体重のデータの度数分布表とヒストグラムを作成してください。*10* 階級にしてください。分布表に度数、累積度数、相対度数、相対累積度数の列を入れてください

5　代表値の意味合い

今回はいろいろな代表値紹介する。

定義 8 (代表値) 代表値はデータを根拠に計算したデータの特性を代表できる数値である。

調査などで収集してきたデータの特性を代表値で纏める。全数調査の場合では、代表値はそのまま母集団（分析対象のすべて）の特性を表す。標本（母集団の一部分）調査の場合、標本の代表値を計算して、母集団の特性を推測するために利用する。

定義 9 (記述統計) データの特性を代表値、度数分布表、ヒストグラムなどで表す、記述する統計学である。

定義 10 (推測統計) 様々な数学のツールを利用して標本のデータを分析し、母集団の性質を推測する統計学である。

授業計画の 07 までは記述統計である。記述統計の概念を正確に理解し覚えるのは推測統計を勉強するための基本である。記述統計のつまらなさを我慢できれば、推測統計の面白さが分かることが出来る。

5.1　様々な平均

1. 算術平均：
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
算数平均の性質：$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$

2. 切落し平均：データの大きいものや小さいものを切り落としてから平均を計算する異常に大きいまたは小さい値が全体に与える必要以上の影響を取り除く役割がある。**異常値に対して頑健である**。
例：国際試合の不正を防ぐため、それぞれ違う国から来た 5 人の裁判がいる場合、点数の最大値と最小値を切り落として合計（$\bar{x} \times 3$）を得点とする。

3. 移動平均：時系列データ $\{x_1, x_2, \cdots, x_t\}$ によく使われる。t 時点の近くの値で t 時点の平均を計算する。

3 項平均：

$$\bar{x}_t = \frac{x_{t-1} + x_t + x_t}{3}$$

4 項平均

$$\bar{x}_t = \frac{x_{t-1} + x_t + x_{t+1} + x_{t+2}}{4}$$

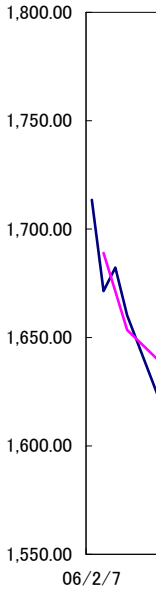
データ全体に関して移動平均を取ると、新しいデータの系列が出来る：３項平均の場合 $\{\bar{x}_2, \bar{x}_3, \cdots, \bar{x}_{n-1}\}$、４項平均の場合 $\{\bar{x}_3, \bar{x}_4, \cdots, \bar{x}_{n-2}\}$。株や為替レートチャートなどによく使われる。移動平均を項数が多ければ滑らかになり（ギザギザが消えていく）、**より長期的な動きを反映する**。

例：TOPIX の日次データの移動平均。[4]

日付	終値	3項移動平均	5項移動平均	7項移動平均
18/04/2006	1741.75			
17/04/2006	1719.05	1734.96		
14/04/2006	1744.07	1735.63	1738.31	
13/04/2006	1743.77	1743.58	1743.99	1748.44
12/04/2006	1742.89	1752.28	1755.65	1754.43
11/04/2006	1770.18	1763.47	1763.58	1762.52
10/04/2006	1777.34	1777.08	1769.96	1762.80
07/04/2006	1783.72	1778.91	1770.59	1763.64
06/04/2006	1775.67	1768.48	1766.49	1765.32
05/04/2006	1746.05	1757.12	1761.95	1759.32
04/04/2006	1749.65	1750.11	1750.83	1752.08
03/04/2006	1754.64	1744.15	1741.04	1741.77
31/03/2006	1728.16	1736.49	1734.13	
30/03/2006	1726.68	1722.13		
29/03/2006	1711.54			

TOPIX の日次データの移動平均

[4]以下の表と次のグラフは Yahoo Japan Finance のホームページからダウンロードした TOPIX のデータをもとに作成したものである。



移動平均の動き

日付	終値	3項移動平均	5
18/04/2006	1741.75		
17/04/2006	1719.05	1734.96	
14/04/2006	1744.07	1735.63	
13/04/2006	1743.77	1743.58	
12/04/2006	1742.89	1752.28	
11/04/2006	1770.18	1763.47	
10/04/2006	1777.34	1777.08	
07/04/2006	1783.72	1778.91	
06/04/2006	1775.67	1768.48	
05/04/2006	1746.05	1757.12	
04/04/2006	1749.65	1750.11	
03/04/2006	1754.64	1744.15	
31/03/2006	1728.16	1736.49	
30/03/2006	1726.68	1722.13	
29/03/2006	1711.54		

TOPIX の日次データの

4. 幾何平均：

$$\bar{x} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_{n-1} \times x_n}$$

幾何平均はよく平均増加率の計算に使われる。例：なぜだろうか。逆に考えれば分かる。3 年間の間で各年それぞれ $r_1 = 23\%, r_2 = 27\%, r_3 = 28\%$ で増加したとする、10 年間を通して増加したのは

$$R = (1 + r_1)(1 + r_2)(1 + r_3) - 1 \approx 1$$

となる。そして平均増加率は

$$r = \sqrt[3]{(1 + r_1)(1 + r_2)(1 + r_3)} - 1 \approx 26\%$$

と定義する。ここで、中国人の年収が 3 年間で 2 倍（200%）になった。この 10 年間の間で年間平均増加率 r はいくらだろうか。それは

$$r = \sqrt[3]{1+R} - 1 = \sqrt[3]{2} - 1 \approx 26\%$$

になる。平均増加率はどんな意味があるだろう。理解するために、逆に平均的に毎年 25% 増加すれば、3 年で何倍になるのを考えよう。それは $1 + r$ を 3 回を掛けることになる：

$$(1+r)^3 = (1+26\%)\,(1+26\%)\,(1+26\%) \approx 2.$$

なるほど 2 に戻った、これは「平均的に」の意味だ。

5.2 レポート課題（提出期限４月２６日）

1. http://www.geocities.jp/qfliuwind/の DATA01 の中にある体重のデータの度数分布表とヒストグラムを作成してください。10 階級にしてください。分布表に度数、累積度数、相対度数、相対累積度数の列を入れてください。

6　講義に関するアンケート調査

1. 難しい過ぎるか？簡単すぎるか？ちょうど良いのか？難しいならどこが難しいか？

2. 改善すべき点

3. こんなことが知りたい

4. Excel の内容に関する要望

5. 統計学の内容に関する要望

7　代表値の意味合い（続き）

前回で紹介した平均の続きに、他の代表値について説明する。

7.1　中央値（メディアン median）

定義 11 データを小さいものから順に並べて、その真ん中にある値は中央値である。

例 12

$$X = \{4, 7, 2, 30, 9, 7, 1\}$$

$$X^* = \{1, 2, 4, 7, 7, 9, 30\}$$

7 が真ん中に位置するので中央値である。

中央値は算術平均より異常値の影響を受けにくい。

例 13 某社の株価の先週一週間月曜日から金曜日までの五日間の株価はそれぞれ 100, 90, 10, 110, 110 円となっているとする。水曜日の 10 円の高値はその会社の収益実績に関する噂によるもので、異常値である。平均を計算すれば、84 になるが、中央値は 100 となっている。明らかに、平均値は異常値の影響を大きく受けていて、株価の平均でこの会社の実力を評価すると過小評価につながる。一方では中央値の方はより忠実にこの会社の実力を反映している。

7.2 最頻値（モード mode）

定義 14 度数が最大な値が最頻値である。同じことで、出現頻度が一番高い値である。

例 15

$$X = \{2, 3, 2, 4, 6, 4, 6, 6, 7\}$$

とする、6 の度数（出現頻度、出現した回数）が一番高く（高く、多く）、3 になっているので、6 が最頻値である。

7.3 レンジ（range）

定義 16 データの最大値と最小値との差。同じくデータの取る値の範囲で理解してもいい。

$$R = \max(X) - \min(X)$$

例 17

$$X = \{2, 3, 2, 4, 6, 4, 6, 6, 7\}$$

X の最小値は 2 で、最大値は 7 なので、レンジ $= 7 - 2 = 5$。

レンジはデータ散らばりの尺度の一つである。データの範囲をあらわしているが、レンジの範囲内でデータはどういう具合で散在しているのかに関する情報を含んでいない。

7.4 平均偏差（mean deviation）

定義 18 $X = \{x_1, x_2, \cdots, x_n\}$ とし、平均偏差は

$$d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

ただし $\bar{x}$ は算術平均である。

数式での定義しか出来ないが、解釈すれば、平均偏差は個々のデータが平均からの乖離を全データに渡って評価する一つの統計量である。

例 19 $X = \{x_1, x_2, \cdots, x_{10}\} = \{2, 3, 2, 4, 6, 4, 6, 6, 3, 4\}$ とする、

$$\bar{x} = \frac{2+3+2+4+6+4+6+6+7}{10} = 4$$

$$\begin{aligned} d &= \frac{|2-4| + |3-4| + |2-4| + \cdots + |7-4|}{10} \\ &= \frac{2+1+2+0+2+0+2+2+3}{10} = \frac{7}{5} \end{aligned}$$

意味としては、各データは平均的に $\bar{x}$ から 7/5 離れている。

7.5 Excel で計算する時のコマンド

- レンジ：= max(データ範囲）-min(データ範囲）

  例 20 = *max(a1:a10)-min(a1:a10)*

- 中央値：= MEDIAN(データ範囲)

  例 21 =*MEDIAN(a1:a10)*。

- 最頻値：= MODE(データ範囲)

  例 22 =*MODE(a1:a10)*

8　代表値の意味合い（続き）

前回の続きで散らばりの尺度（レンジ、平均偏差も散らばりの尺度）となっている代表値を幾つか紹介する。

8.1　分散（Variance)

定義 23 分散はデータが平均からの乖離の具合を図る尺度である。

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$
$$= \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

ただし、$\bar{x}$ は算術平均である。

計算する時は $\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$ で行うほうが簡単。すなわち**分散は二乗の平均引く平均の二乗**となる。

もう一つの定義は

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

である。n が大きい場、以上の二つは殆ど同じ値になる。

8.2　標準偏差（Standard Deviation）

定義 24 分散の平方根である。

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

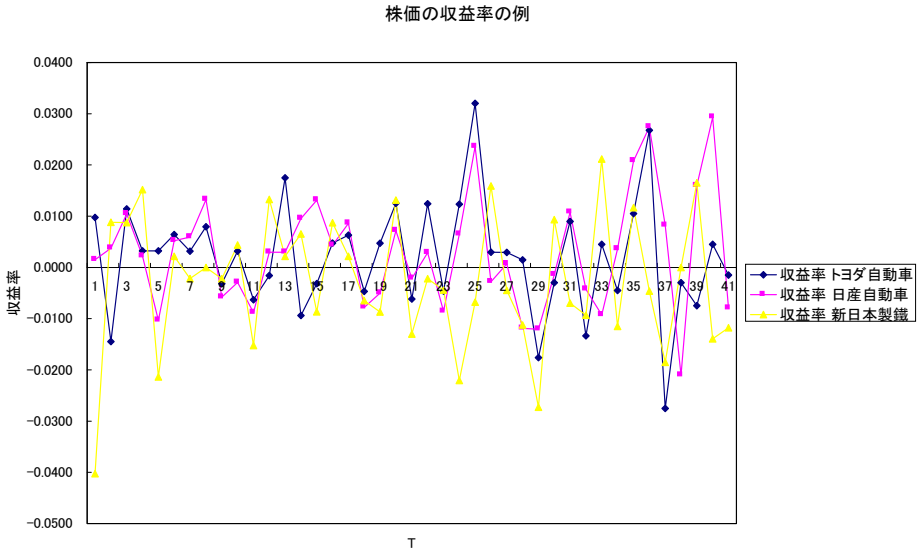
或いは

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

分散と標準偏差が違うものですが、基本的に同じ役割を果たしている。

8.3　分散や標準偏差の応用例

収益率の期待値[5]（標本平均で推定）とあわせて、分散（標本の分散で推定）や標準偏差（標本の標準偏差で推定）を利用して、株式の投資価値を評価する方法がある。Yahoo!ファイナンスからダウンロードした 2006 年 3 月 1 日から 4 月 28 日までの三つの会社の株式データで計算した結果：トヨダ自動車、日産自動車、新日本製鐵の株式のの収益率と収益率の分散はそれぞれ、(0.002, 0.0001), (0.003, 0.0001) と (-0.003, 0.0002) である。期待値が高く、分散（リスク）が小さいものが良いと言う観点から、この中一番パフォマンスの良いのがトヨダ自動車である。



8.4　変動係数

変動係数は標準偏差を平均で割ったもので、もとのデータの単位の影響を取り除ける。

$$C.V. = \frac{S}{\bar{x}}$$

[5]ここで厳密な説明なしで紹介するだけに留まる、後半の講義で期待値や分散などの推定と予測に関して詳しく説明する。

8.5　標準化データ、偏差値

定義 25 標準化データはもとのデータを標準偏差で割ったもので、標準化の操作によりデータの単位の影響が取り除かれる。

$$z_i = \frac{x_i}{S_x}$$

ただし、S_x は x の標準偏差をあらわしている。

偏差値は日本でよく成績の評価に使われている。

$$Ti = 50 + 10zi$$

8.6　Excel で計算する時のコマンド

分散：S^2： =VARP(データ範囲)，s^2： =VAR(データ範囲)
標準偏差：S： =STDEVP(データ範囲)，s: =STDEV(データ範囲)

練習 26 各自で *Yahoo* ファイナンスから *3* つの企業の株価データをダウンロードして、各企業の株価の収益率の平均と分散を計算してください。平均と分散を基準にどの企業の株式のパフォマンスが良いのか説明してください。さらに、*3* つの企業でペアーを三種類組んで、各ペアーの中の二つの株式を同じ割合で購入するとする。どの組み合わせがよりいいパフォマンスを持っているのか、平均と分散を計算し、その結果を用いて評価してください。

8.7　分散の説明の補足

分散は統計学の中では極めで重要な概念であるため、理解を深めるために例を挙げて説明します。

例 27 三つのデータのセットがあるとする。

$$\begin{array}{cccc} X & 1 & 5 & 3 \\ Y & 13 & 1 & 20 \\ Z & 15 & 50 & 120 \end{array}$$

分散の定義式

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n} \\ &= \frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2 \end{aligned}$$

に従って、計算する。

$$\begin{aligned} \bar{x} &= \frac{1+5+3}{3} = 3 \\ S_x^2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{(1-3)^2+(5-3)^2+(3-3)^2}{3} \\ &= \frac{(-2)^2+2^2+0^2}{3} = \frac{4+4}{3} = 2.67 \end{aligned}$$

または同じく

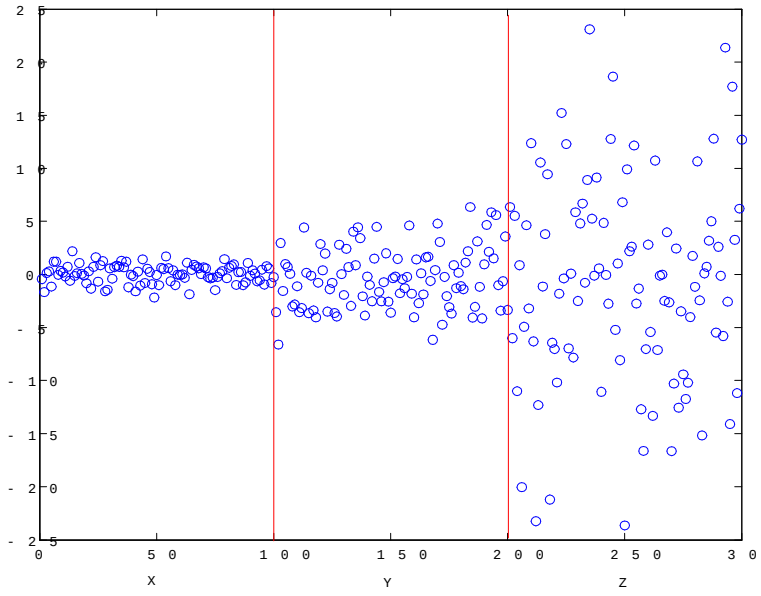
$$S_x^2 = \frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2 = \frac{1^2+5^2+3^2}{3}-3^2 = 2.67$$

同じように

$$S_y^2 = \frac{13^2+1^2+20^2}{3}-\left(\frac{13+1+20}{3}\right)^2 = 61.56$$

$$S_z^2 = \frac{15^2+50^2+120^2}{3}-\left(\frac{15+50+120}{3}\right)^2 = 1905.6$$

明らかに $S_z^2 > S_y^2 > S_x^2$. 意味としては Z のデータはもっとも散らばっている。この事実はデータを眺めるだけでも分かる。X のデータは $1,5,3$ で、Z のデータ $15,50,120$ のように大きく離れていない。あまり厳密ではないが、データの分散はデータセットの中にある個々のデータは互いに大きく離れているかどうかを評価する指標である。



例 28 分散の概念を直感的に理解するために下のグラフを見よう。$X = \{x_1, x_2, \cdots, x_{100}\}, Y = \{y_1, y_2, \cdots, y_{100}\}, Z\{y_1, y_2, \cdots, y_{100}\}$ の平均は 0 でそれぞれの分散 $1,\ 3,\ 10$ である。グラフで見ると三者の違いは明らかである。分散が大きいほどデータの散らばりの程度が大きくなる。

8.8　散布図を作ってみよう（次回のための予習）

以前習ったグラフの作り方でグラフの種類を散布図と選択すれば作成できる。DATA01 の中の身長と体重のデータを使って作成してください。散布図から何を見て取れるのか考えましょう。

9　二変数の関係を明確にする散布図、分割表と相関係数

今まで一変数に関する整理方法と統計量を説明してきた。度数分布表やヒストグラムなどの整理法方と平均や分散などの統計量で一変数の特性を明らかにすることが出来る。今回から、二変数の場合について考えよう。二変数の場合、当然一変数に使われた手法や統計量は二つの変数におのおの適用できるが、二つの変数の間の関係を明確にするために、一変数の場合と違った手法や統計量が必要となる。

9.1　散布図

二つの変数をそれぞれ縦軸と横軸にして使ったグラフは散布図となる。散布図から二つの変数間の線形関係が見えてくる。図１はDATA01の中の身長と体重のデータで作った散布図である。当然なことで、身長が高ければ体重が重いはずである。散布図で見るとこのような身長と体重の関係は右上がりのラインを一本引けるように見える。

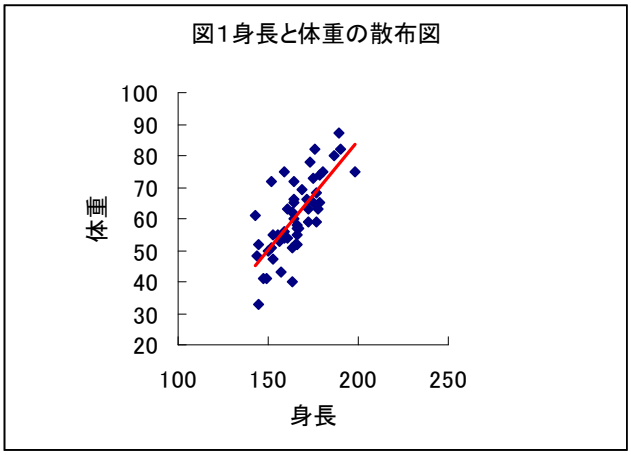


図２は日本の1980年から2000年までの一人当たりGDP(国内総生産）と乳児死亡率のデータの散布図である。一人当たりGDPの成長と伴に、乳児死亡率も下がっていることが分かる。左下がりの直線を引けるように見える。

9.2　分割表

分割表の例：身長と体重。身長が高くて体重が重い人数が身長が高くて体重が低い人より多い…。分割表からある程度二変数間の関係が見えるが、視覚的に利用しにくい。

	155cm以下	155〜175cm	175cm以上	周辺度数分布
50kg以上	1	13	10	24
50kg以下	6	11	4	21
周辺度数分布	7	24	14	45

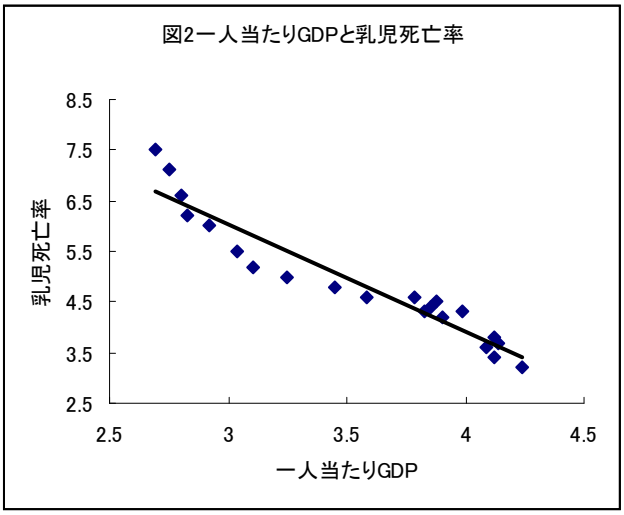


図 1: データの出所：GDP のデータは内閣府 SNA、乳児死亡率は厚生労働省大臣官房統計情報部人口動態・保健統計課、人口のデータは総務省統計局から。

9.3　共分散と相関係数

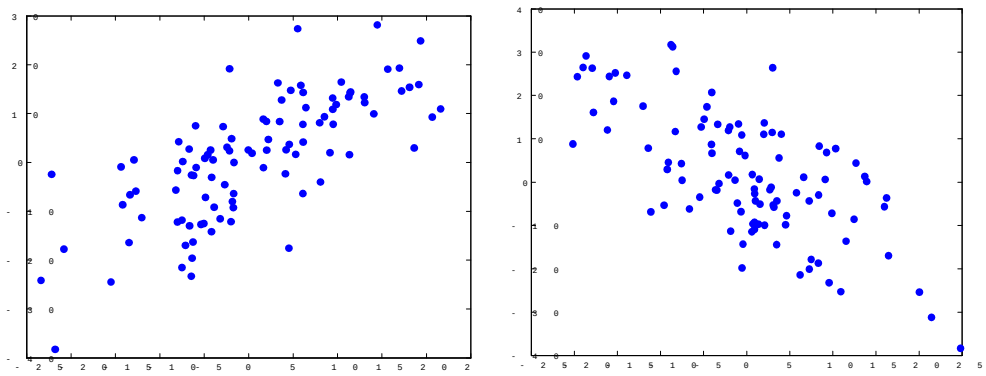
共分散と相関係数は二変数間の相関関係を示す統計量である。二変数 X と Y がとする、この場合よく S_{xy} で X と Y の共分散を ρ_{xy} で X と Y の相関係数を表す。

9.3.1　共分散

定義 29

$$\begin{aligned} S_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y} \end{aligned}$$

例 30 図 3 の中に共分散がプラスとマイナスの例を示している。



正の共分散を持つ　　負の共分散を持つ

図 3 異なった共分散を持ったデータの散布図

9.3.2　相関係数

定義 31 相関係数は標準化された共分散である。

$$\begin{aligned}\rho_{xy} &= \frac{S_{xy}}{S_x S_y} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}\end{aligned}$$

一つ重要な性質：証明を省略しますが、任意の二つの変数の相関係数は

$$-1 \leq \rho \leq 1$$

である。図 4 は異なった相関係数を持ったデータの散布図である。

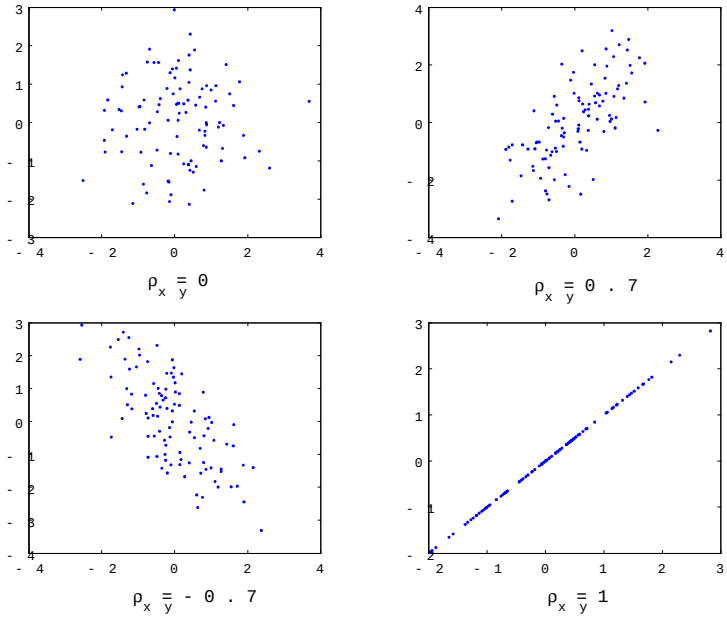


図４異なった相関係数を持ったデータの散布図

例 32 肥料と農産物の出来高の関係を相関係数で見ることが出来る。(以下のデータは架空のデータ)。

例 33

田んぼの *ID*	a	b	c	d	e	f	g	h	i	j
肥料の量 *(g)*	10	15	20	25	30	35	40	45	50	55
出来高 *(kg)*	6	15	19	18	7	19	24	22	34	35

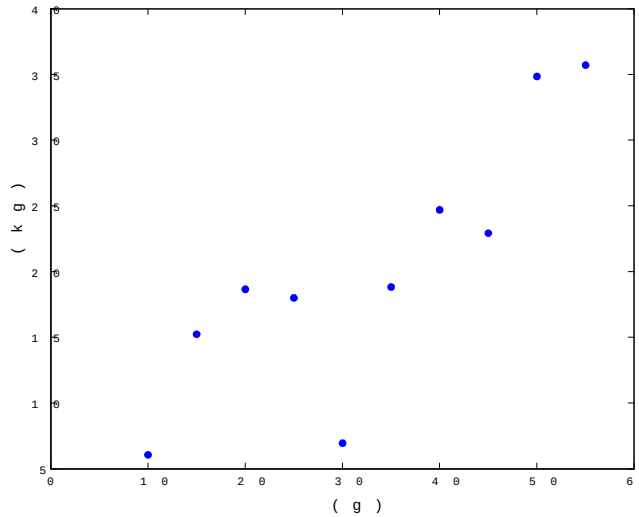


図 5 肥料と農産物の出来高との関係

相関係数を計算したら、

$$\rho_{xy} = 0.84$$

従って、この種の肥料は農産物の増産に大きく寄与していると結論付けられる。

練習 34 $x = \{4, -3, 5, 1, 5\}, y = \{1, -3, 3, 0, 1\}$ とする。x と y の平均、分散及び共分散を手計算で計算してください。レポートではないが、提出してください

9.3.3　Excel で計算する時のコマンド

1. 共分散：=COVAR(データ 1 の範囲，データ 2 の範囲)
2. 相関係数：=CORREL(データ 1 の範囲，データ 2 の範囲)

9.3.4　散布図の作り方の練習

DATA01 の身長と体重のデータで散布図を作成してください。

10　散布図に直線を当てはめる

今日は回帰分析の考え方の基本となっていることを簡単に説明する。回帰分析に関しては推測統計の内容となっているため、後の講義で説明する。

10.1　数学の準備

二つの変数 x, y の関係を数式で表すことが出来る。例えば、$y = a + bx$。このような数式は x の値を横軸の座標とし、y を縦軸の座標とすれば、グラフで表すことが出来る。$a = 3, b = 2$ として、例を挙げる。$y = 3 + 2x$ となる。そしてグラフを書くために、まず幾つかの点を決める。

$x = 0$ の時、$y = 3 + 2 \times 0 = 3$,

$x = 2$ の時、$y = 3 + 2 \times 2 = 7$...,

同じようにして、幾つかの点の座標を決める。ここでは以下のように 4 つの点を決めた。

x	0	2	5	7
y	3	7	13	17

グラフにこの 4 つの点を書いて繋げば一本の直線になる。

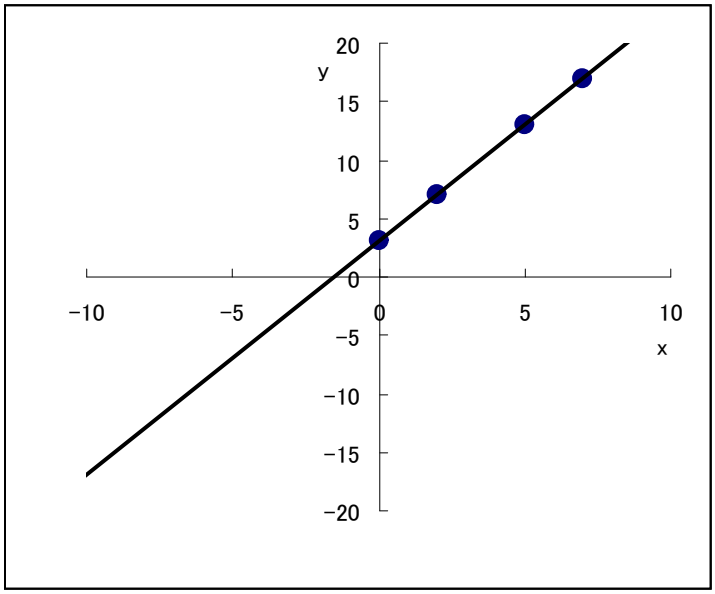


図 1　一次式と直線

数学の中でよく直線で二つの変数間の線形関係を表す。

10.2　データの散布図に直線を当てはめよう

以下のように身長と体重のデータがあるとしよう。

学籍番号	身長(cm)	体重(kg)
161111	172	59
161112	166	58
161113	164	65
161114	175	73
161115	149	41
161116	144	48
161117	161	63
161118	159	56
161119	172	63
161120	167	57
161121	166	52
161122	187	80

散布図を作成する。

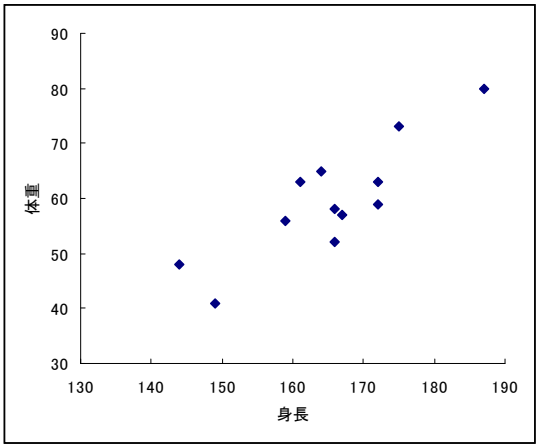
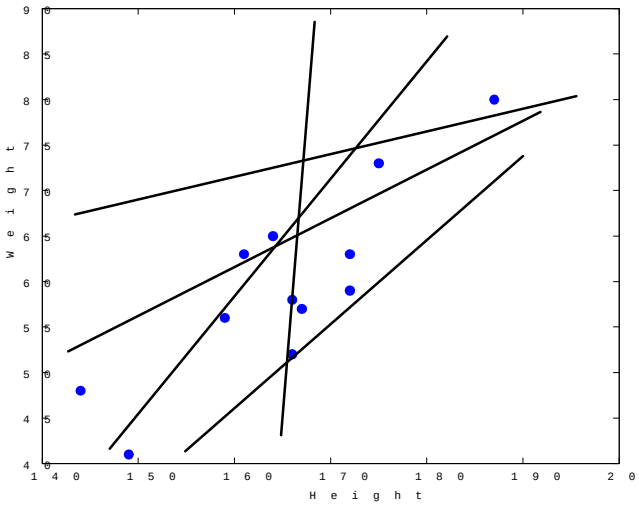


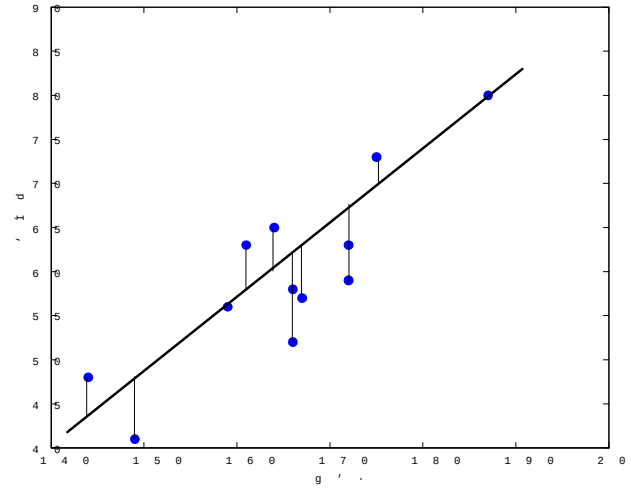
図２身長と体重の散布図

一本の直線を引きたいが、どうやって引けば身長と体重の関係をよく代表できるの

かを考えよう。何本も直線を引けるが、どれがいいか基準をないと分からない。



そこで自然な発想で、データの点から直線への縦の距離の二乗（縦の距離そのものを使うこともあるが、その場合では後で出で来る計算が難しくなる）の総和が一番小さくすることの出来る直線が一番良いと決める。



身長と体重を x と y として、各生徒の身長と体重をそれぞれ x_i と y_i であらわす, ただし $i = 1, 2, 3, ..., 12$. 引いた直線を $\hat{y} = a + bx$ とする。それで、各データから直線への縦の距離は

$$\begin{aligned} d_i &= y_i - \hat{y} \\ &= y_i - (a + bx_i) \end{aligned}$$

となる。その二乗の総和を S とあらわして、

$$S = \sum_{i=1}^{12} (y_i - (a + bx_i))^2$$

となる。

一般の場合においては

$$S = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 . \tag{1}$$

定義 35 (最小二乗法) 上述したように、データの点から直線までの縦の距離の二乗和 S を最小にするように直線を決める方法は最小二乗法である。このように求めた直線の係数は

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{2}$$

$$a = \bar{y} - b\bar{x} \tag{3}$$

11　係数 a と b の求め方（参考）

S を最小にするように a と b を決める。1 式を展開したら

$$S = \sum_{i=1}^{n} y_i^2 + na^2 + \left(\sum_{i=1}^{n} x_i^2\right) b^2 + \left(\sum_{i=1}^{n} x_i\right) 2ab - \left(\sum_{i=1}^{n} y_i\right) 2a - \left(\sum_{i=1}^{n} x_i y_i\right) 2b \tag{4}$$

もちろん x_i と y_i はデータである。

まず 3 式を a の二次式としてみる。a に関して整理すると

$$S = \left(na^2 - \left(\sum_{i=1}^{n} y_i\right) 2a + \left(\sum_{i=1}^{n} x_i\right) 2ab - \left(n\bar{y}^2 - 2nb\bar{x}\bar{y} + nb^2\bar{x}^2\right) \right) + \left(n\bar{y}^2 - 2nb\bar{x}\bar{y} + nb^2\bar{x}^2\right) + \sum_{i=1}^{n} y_i^2 + \left(\sum_{i=1}^{n} x_i^2\right) b^2 - \left(\sum_{i=1}^{n} x_i y_i\right) 2b$$

さらに整理すると

$$S = n(a - (\bar{y} - b\bar{x}))^2 + \cdots$$

ゆえに S を最小にするのは

$$a = \bar{y} - b\bar{x} \tag{5}$$

となる。

次に、3 式を b の二次式としてみる。b に関して整理すると

$$S = \sum_{i=1}^{n} x_i^2 \left(b - \left(\frac{\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \right) \right)^2 + \cdots$$

となる。ゆえにゆえに S を最小にするのは

$$b = \frac{(\sum_{i=1}^{n} x_i y_i) - a \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{6}$$

となる。さらに 5 式の結果を a の代わりに 6 式に代入すれば

$$b = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} \tag{7}$$

$$a = \bar{y} - b\bar{x} \tag{8}$$

が得られる。

または、微分の計算が出来るなら、1 式の偏微分で以下の連立方程式を構成して、それを解けば同じ結果を得られる。

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

11.1　上述した例の係数の計算

まず必要となる要素の計算を行う

i	x_i	y_i	x_i^2	x_iy_i
1	172	59	29584	10148
2	166	58	27556	9628
3	164	65	26896	10660
4	175	73	30625	12775
5	149	41	22201	6109
6	144	48	20736	6912
7	161	63	25921	10143
8	159	56	25281	8904
9	172	63	29584	10836
10	167	57	27889	9519
11	166	52	27556	8632
12	187	80	34969	14960
	$\sum_{i=1}^{12} x_i$	$\sum_{i=1}^{12} y_i$	$\sum_{i=1}^{12} x_i^2$	$\sum_{i=1}^{12} (x_iy_i)$
	1982	715	328798	119226
	$\bar{x}$	$\bar{y}$		
	165.17	59.58		

表 1 最小二乗法の準備計算

そして、結果を公式に代入する：

$$b = \frac{(\sum_{i=1}^{n} x_iy_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} = \frac{119226 - 12 \times 165.17 \times 59.58}{328798 - 12 \times 165.17^2} = 0.80$$

$$a = \bar{y} - b\bar{x} = 60 - 0.80 \times 165.17 = -72.14$$

求める直線は

$$\hat{y} = -72.14 + 0.80x$$

この式は身長と体重の関係を近似的に示している。およそ、身長が 1cm 増えれば体重が 0.8kg 増加する。

練習 36 $x = \{2, 5, 6, 9\}, y = \{4, 6, 8, 9\}$ とする。最小二乗法で近似直線の係数を求めよう。（ヒント：まずエクセルで表一を真似して表を作って、準備の計算を行ってください。）

12　散布図に直線を当てはめる

今日は回帰分析の考え方の基本となっていることを簡単に説明する。回帰分析に関しては推測統計の内容となっているため、後の講義で説明する。

12.1　数学の準備

二つの変数 x, y の関係を数式で表すことが出来る。例えば、$y = a + bx$。このような数式は x の値を横軸の座標とし、y を縦軸の座標とすれば、グラフで表すことが出来る。$a = 3, b = 2$ として、例を挙げる。$y = 3 + 2x$ となる。そしてグラフを書くために、まず幾つかの点を決める。

$x = 0$ の時、$y = 3 + 2 \times 0 = 3$,

$x = 2$ の時、$y = 3 + 2 \times 2 = 7$...,

同じようにして、幾つかの点の座標を決める。ここでは以下のように 4 つの点を決めた。

x	0	2	5	7
y	3	7	13	17

グラフにこの 4 つの点を書いて繋げば一本の直線になる。

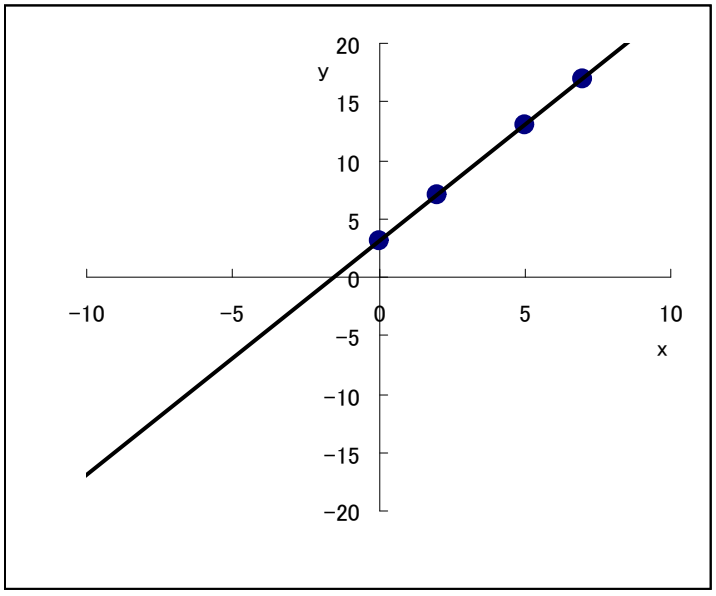


図 1　一次式と直線

数学の中でよく直線で二つの変数間の線形関係を表す。

12.2　データの散布図に直線を当てはめよう

以下のように身長と体重のデータがあるとしよう。

学籍番号	身長(cm)	体重(kg)
161111	172	59
161112	166	58
161113	164	65
161114	175	73
161115	149	41
161116	144	48
161117	161	63
161118	159	56
161119	172	63
161120	167	57
161121	166	52
161122	187	80

散布図を作成する。

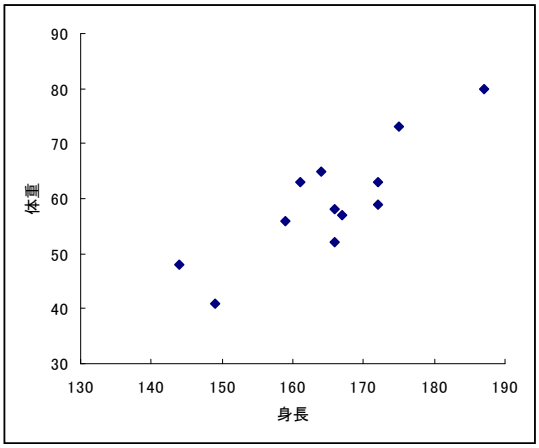
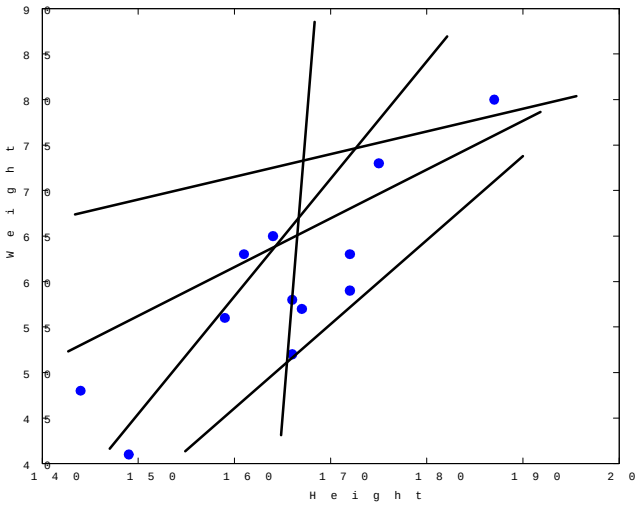


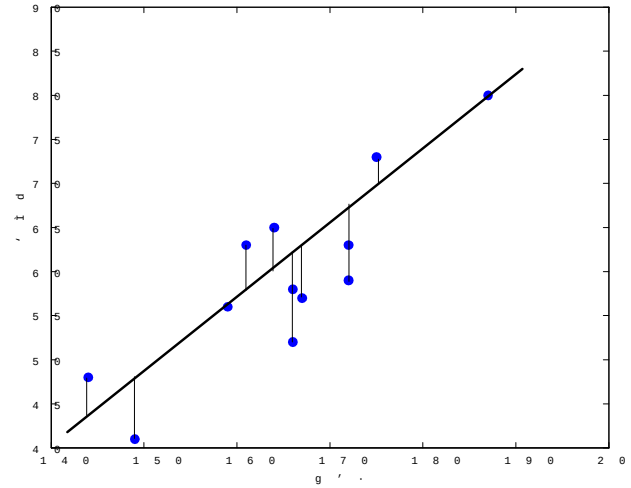
図２身長と体重の散布図

一本の直線を引きたいが、どうやって引けば身長と体重の関係をよく代表できるの

かを考えよう。何本も直線を引けるが、どれがいいか基準をないと分からない。



そこで自然な発想で、データの点から直線への縦の距離の二乗（縦の距離そのものを使うこともあるが、その場合では後で出で来る計算が難しくなる）の総和が一番小さくすることの出来る直線が一番良いと決める。



身長と体重を x と y として、各生徒の身長と体重をそれぞれ x_i と y_i であらわす, ただし $i = 1, 2, 3, ..., 12$. 引いた直線を $\hat{y} = a + bx$ とする。それで、各データから直線への縦の距離は

$$\begin{aligned} d_i &= y_i - \hat{y} \\ &= y_i - (a + bx_i) \end{aligned}$$

となる。その二乗の総和を S とあらわして、

$$S = \sum_{i=1}^{12} \left(y_i - (a + bx_i)\right)^2$$

となる。

一般の場合においては

$$S = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 . \tag{9}$$

定義 37 (最小二乗法) 上述したように、データの点から直線までの縦の距離の二乗和 S を最小にするように直線を決める方法は最小二乗法である。このように求めた直線の係数は

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{10}$$

$$a = \bar{y} - b\bar{x} \tag{11}$$

13　係数 a と b の求め方（参考）

S を最小にするように a と b を決める。1 式を展開したら

$$\begin{aligned} S = \sum_{i=1}^{n} y_i^2 + na^2 + \left(\sum_{i=1}^{n} x_i^2\right) b^2 + \left(\sum_{i=1}^{n} x_i\right) 2ab \\ - \left(\sum_{i=1}^{n} y_i\right) 2a - \left(\sum_{i=1}^{n} x_i y_i\right) 2b \end{aligned} \tag{12}$$

もちろん x_i と y_i はデータである。

まず 3 式を a の二次式としてみる。a に関して整理すると

$$\begin{aligned} S = \left(na^2 - \left(\sum_{i=1}^{n} y_i\right) 2a + \left(\sum_{i=1}^{n} x_i\right) 2ab - \left(n\bar{y}^2 - 2nb\bar{x}\bar{y} + nb^2\bar{x}^2\right) \right) \\ + \left(n\bar{y}^2 - 2nb\bar{x}\bar{y} + nb^2\bar{x}^2\right) + \sum_{i=1}^{n} y_i^2 + \left(\sum_{i=1}^{n} x_i^2\right) b^2 - \left(\sum_{i=1}^{n} x_i y_i\right) 2b \end{aligned}$$

さらに整理すると

$$S = n(a - (\bar{y} - b\bar{x}))^2 + \cdots$$

ゆえに S を最小にするのは

$$a = \bar{y} - b\bar{x} \tag{13}$$

となる。

次に、3 式を b の二次式としてみる。b に関して整理すると

$$S = \sum_{i=1}^{n} x_i^2 \left(b - \left(\frac{\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \right) \right)^2 + \cdots$$

となる。ゆえにゆえに S を最小にするのは

$$b = \frac{(\sum_{i=1}^{n} x_i y_i) - a \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{14}$$

となる。さらに 5 式の結果を a の代わりに 6 式に代入すれば

$$b = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} \tag{15}$$

$$a = \bar{y} - b\bar{x} \tag{16}$$

が得られる。

または、微分の計算が出来るなら、1 式の偏微分で以下の連立方程式を構成して、それを解けば同じ結果を得られる。

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

13.1 上述した例の係数の計算

まず必要となる要素の計算を行う

i	x_i	y_i	x_i^2	x_iy_i
1	172	59	29584	10148
2	166	58	27556	9628
3	164	65	26896	10660
4	175	73	30625	12775
5	149	41	22201	6109
6	144	48	20736	6912
7	161	63	25921	10143
8	159	56	25281	8904
9	172	63	29584	10836
10	167	57	27889	9519
11	166	52	27556	8632
12	187	80	34969	14960
	$\sum_{i=1}^{12} x_i$	$\sum_{i=1}^{12} y_i$	$\sum_{i=1}^{12} x_i^2$	$\sum_{i=1}^{12} (x_iy_i)$
	1982	715	328798	119226
	$\bar{x}$	$\bar{y}$		
	165.17	59.58		

表 1 最小二乗法の準備計算

そして、結果を公式に代入する：

$$b = \frac{(\sum_{i=1}^{n} x_iy_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} = \frac{119226 - 12 \times 165.17 \times 59.58}{328798 - 12 \times 165.17^2} = 0.80$$

$$a = \bar{y} - b\bar{x} = 60 - 0.80 \times 165.17 = -72.14$$

求める直線は

$$\hat{y} = -72.14 + 0.80x$$

この式は身長と体重の関係を近似的に示している。およそ、身長が 1cm 増えれば体重が 0.8kg 増加する。

練習 38 $x = \{2, 5, 6, 9\}, y = \{4, 6, 8, 9\}$ とする。最小二乗法で近似直線の係数を求めよう。（ヒント：まずエクセルで表一を真似して表を作って、準備の計算を行ってください。）

14　復習

最小二乗法の意味合い

最小二乗法の目的：データの点から近似直線への縦の距離の二乗和を最小にする近似直線の係数を求める。

基準が無いと、近似直線が無限に画ける、その中にデータの点からの縦の距離が最小のものがある。その直線の係数が最小二乗法で求まる。

15　実習

練習 39 $x = \{4, 6, 9\}, y = \{6, 6, 9\}$ とする。x と y の平均、分散、共分散、相関係数を手計算で計算しよう、最小二乗法で近似直線の係数を求めよう、近似直線の式を書いてください。（ヒント：まず表一を真似して表を作って、準備計算を行ってください。）近似直線のグラフを手で書いてください。

15.1　レポート課題 3

DATA03 をダウンロードし、まずデータの基本統計量と共分散および相関係数を Excel で計算して下さい。ビールの消費量と平均気温の間はどのような関係があるのかを計算の結果を持って説明しよう。

そして、仮に 2007 年の 5 月に生産計画を立てることになっているとする。2007 年 7 月の平均気温の予測値は 37 度として、2007 年にこのブランドのビールをどのぐらい生産すればいいでしょう？Excel で分析ツールで回帰分析を行って回答しよう。

提出締め切り：12 月 6 日

16　補足：Excel を用いた回帰分析

ツール → 分析ツール → 回帰分析 →OK→ 入力 Y 範囲に非説明変数（分析したいまたは予測したい変数）の範囲、入力 X 範囲に説明変数（非説明変数の変化を説明できる変数）の範囲を指定 →OK。

出力の結果の見方

概要

回帰統計	
重相関 R	0.641785
重決定 R2	0.411887
補正 R2	0.390883
標準誤差	20.70401
観測数	30

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	8405.914	8405.914	19.60993	0.000132
残差	28	12002.37	428.656		
合計	29	20408.28			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	94.39179	171.995	0.548805	0.587489	−257.924	446.7081	−257.924	446.7081
X 値 1	20.47776	4.624284	4.42831	0.000132	11.00534	29.95019	11.00534	29.95019

黄色の部分の値は求めたい係数です。たとえば y をビールの消費量として、x を 7 月の平均気温とする。両者の関係は

$$y = a + b \times x + u$$

（ただし、u は撹乱項と呼ばれる。気温以外の不確実な要素の影響を表す。）と仮定して回帰分析を行った場合、a の推定量は切片の値で 94.4 ぐらい、b は表の中の「X 値　1」に対応している値で 20.5 ぐらい。推定された式は

$$y = 94.4 + 20.5x$$

となる。

17　数学の準備

集合と記号集合論で使われる記号 $\cup, \cap$。
集合は物の集まり。

例 40 クラスを表す集合で、A をクラス *1* のメンバーとする。$A =$ { 酒井さん、中村さん、田中さん、...。}。

例 41 整数の集合、$A = \{2, 1, 5\}, B = \{3, 6, 5\}$。

$A \cup B$ は A と B の和集合を表す。上の例の場合 $A \cup B = \{2, 1, 5, 3, 6\}$。
$A \cap B$ は A と B の積集合で両方がともに持つ要素より構成される。同じ例で $A \cap B = \{5\}$。

18　確率に関する説明

日常生活でよく気軽に「確率」という言葉を使っている。たとえば、

- 今日雨が降っているから、彼女遅刻する確率が高いね；
- 両端の車両は席がたくさん空いて、座れる確率が高い；
- 男の子と女の子生まれてくる確率が同じ

などなど。今日はこのような言葉で表したあいまいな確率と違って、数学の表現を利用してより厳密的に確率の意味を説明する。

18.1　確率に関する例

定義を与える前に、幾つかの例で印象を付けてもらう。

例 42 コインを投げて 100 回、1000 回を投げれば、大体表と裏を同じ回数５０回、５００回ぐらい出る。経験的に、コインを投げる場合、表と裏をでる確率は同じで *1/2* である。

例 43 正確に作ったサイコロを投げる。*1* が出る確率は *1/6* である。

例 44 太陽は西から沈む確率は *100* %（１）である。

18.2　確率と確率に関連する幾つかの概念

試行 実験、行動、操作、自然現象など。

事象 試行によって生じた結果、ある種の集合として考えられる。

主観的確率 経験したことのない事象の起きる可能性を主観的に判断して、確信度としての確率。たとえば、柔道の選手が「俺明日勝つ確率 90 %あるぜ！」と言う。明日の試合は以前の試合と同じではない、選手の話は自分の確信度を表しているしかない。

統計的確率 仮にある試行を同じ条件で互いに無関係に n 回繰り返すことができるとする。n 回の中である事象が m 回起きたとする。n が無限大に近づくにつれ、m/n はある定数 p に収束する。この事象が起こる確率は p となる。p は統計的確率である。

18.3　確率の公理

どんなふうに確率を定義しても以下の三つの公理を満たさなければならない。

1. 任意の事象 A の確率 $P(A)$ は $0 \leq P(A) \leq 1$ である。
2. 確実に起こる事象 Ω の確率 $P(\Omega) = 1$。
3. 事象 A と B が必ず同時に起こらない場合、A または B 何れか起こる確率は $P(A \cup B) = P(A) + P(B)$。この場合、事象 A と B が排反するという。

18.4　確率に関する計算例

例題 45 コインを 2 回投げる、2 回表が出る確率はいくら。

解答 46 1 回投げて表と裏が出る確率が同じ 0.5 とする。2 回投げて、出るかの性のあるすべてのパターンを考える。

$$\left\{\begin{array}{l} \text{表} \left\{\begin{array}{l} \text{表} \\ \text{裏} \end{array}\right. \\ \text{裏} \left\{\begin{array}{l} \text{表} \\ \text{裏} \end{array}\right. \end{array}\right.$$

上の表で見れば、表表、表裏、裏表、裏裏 4 パターンの中の 1 パターンだから、明らかに確率は 1/4。

例題 47 上の例の続きで $A =$ { 表が 2 回 }, $B =$ { 裏が一回以上 } とする。$P\{A \cup B\}$ はいくらになる*?*

解答 48 $\{A \cup B\}$ = { 表が 2 回または裏が一回以上 }、調べてみれば全部である。$P\{A \cup B\} = 1$。

例題 49 $C =$ { 表が 2 回 } かつ $D =$ { 裏が 2 回 } の確率求めてください。

解答 50 求める確率は $P(C \cap D)$ で、$C =$ { 表表 }、$D =$ { 裏裏 }、C と D には共通要素がないため、$C \cap D = \phi$、ただし ϕ は空を表す、空集合と呼ぶ。$P\{\phi\} = 0$。言葉で言うと、表が 2 回と裏が 2 回が同時に起きる可能性はあり得ない。

練習 51 例題の例の続きで、$P\{B\}$ を計算して、そして $P\{A \cup B\} = P\{A\} + P\{B\}$ を確かめよう。

練習 52 コインを 3 回投げる、2 回表が出る確率はいくら*?*{ 表 *1* 回以上または裏 *2* 回 } の確率はいくら*?*{ 表 *1* 回以上かつ裏 *2* 回 } の確率はいくら*?*

18.5　実験

コインを投げる。10 回と 50 回を投げて、毎回の結果は表か裏かを記録して、表になった回数と裏の回数を数えてください。

18.6　Excel で実験（試行を行う）

Excel でコインを投げる：

1. ツール → アドイン → 分析ツール
2. ツール → 分析ツール → 乱数発生 →（変数の数 1、乱数の数 10、分布ベルヌーイ、$P = 0.5$）→OK。
3. 1 の数と 0 の数を数える。

同じ実験を乱数の数を 50 にしてもう一回。1000 にしてもう一回。

19　確率変数

確率変数が取る値はおのおのに一定な確率が対応している。

前回で説明した、コイン投げの実験では表を 1 とし裏を 0 としたら、コイン投げの結果を確率変数 X のとる値とそれに対応している確率は $X = 0$ の確率は $0.5, X = 1$ の確率も 0.5 である。サイコロの場合は $x = 1$ の確率は $\frac{1}{6}, x = 2$ の確率も $\frac{1}{6}, ...$。

19.1　離散確率変数

上述した二つの確率変数の例は確率変数 X は 0 と 1 または $0, 1, 2, 3, 4, 5, 6$ と飛び飛びとした値しかとらない、たとえば 0 と 1 の間の値を取ることがない、この場合は離散確率変数と呼ぶ。

確率変数が取る値とそれに対応する確率との関係は以下の表で表せる。

X が取る値	x_1	x_2	x_3	$\cdots$	x_4
対応する確率 $P(x)$	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\cdots$	$P(x_4)$

ただしここでは x_i が昇順で並べてられているとする。ここでは $P(x)$ は**確率度数関数**と呼ばれる。$P(x_1), P(x_2) ...$ は x の具体的な値に対応する確率を表す。

分布関数：分布関数を $F(x)$ で表すと集合 $\{X \leq x\}$ の確率で $F(x) = P(\{X \leq x\})$ となる。離散確率変数の場合は

$$\begin{aligned} F(x) &= P(\{X \leq x\}) \\ &= \sum_{x_i < x} P(x_i) \end{aligned}$$

上式の和記号の意味は x より小さい x_i に関して和を取る意味である。たとえば上の表の例では、$F(x_3) = P(X \leq x_3) = P(x_1) + P(x_2) + P(x_3)$。ただしここでは x_i が昇順で並べてられているとする。

確率変数の**期待値**：確率変数 X の期待値を $E(X)$ で表す。離散確率変数の場合

$$E(X) = \sum_{i=1}^{n} x_i P(x_i).$$

確率変数の**分散**：確率変数 X の分散を $V(X)$ で表して、離散確率変数の場合

$$V(X) = E\left[(X - E(X))^2\right] = \sum_{i=1}^{n} \left[(x_i - E(X))^2 P(x_i)\right].$$

定義 53 (確率分布) ある確率変数 X の取る値に対応する確率がある関数 $P(x)$ に対応するとき、この確率変数 X の**確率分布**は $P(x)$ に従うという。

19.1.1　離散確率分布の例

1. ベルヌーイ分布：

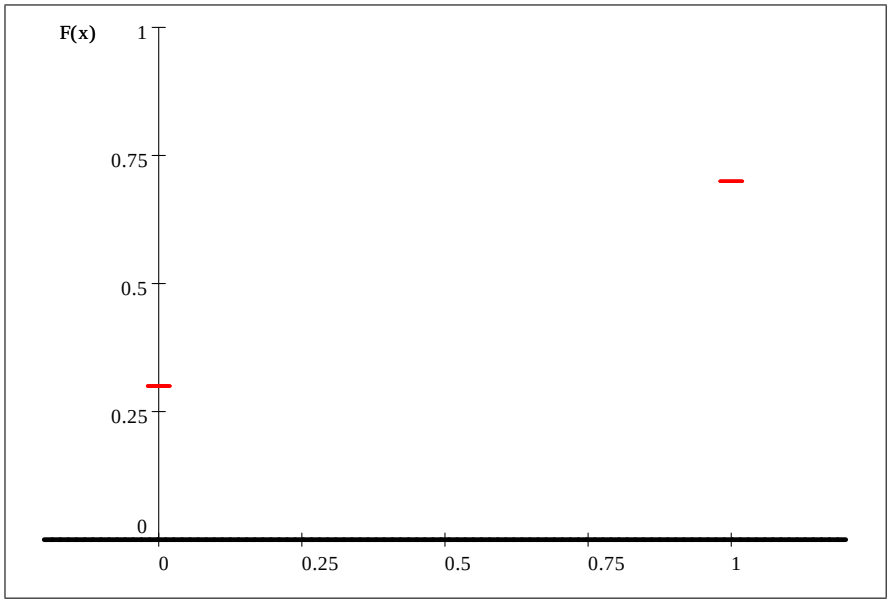
   確率度数関数

   $$\begin{cases} P(x) = p & x = 1 \\ P(x) = 1 - p & x = 0 \end{cases}$$

   期待値 $E(X) = 1 \times p + 0 \times (1 - p) = p$. 分散は

   $$\begin{aligned} V(X) &= (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p \\ &= p^2 - p^3 + p - 2p^2 + p^3 = p - p^2 \end{aligned}$$

確率度数関数のグラフ

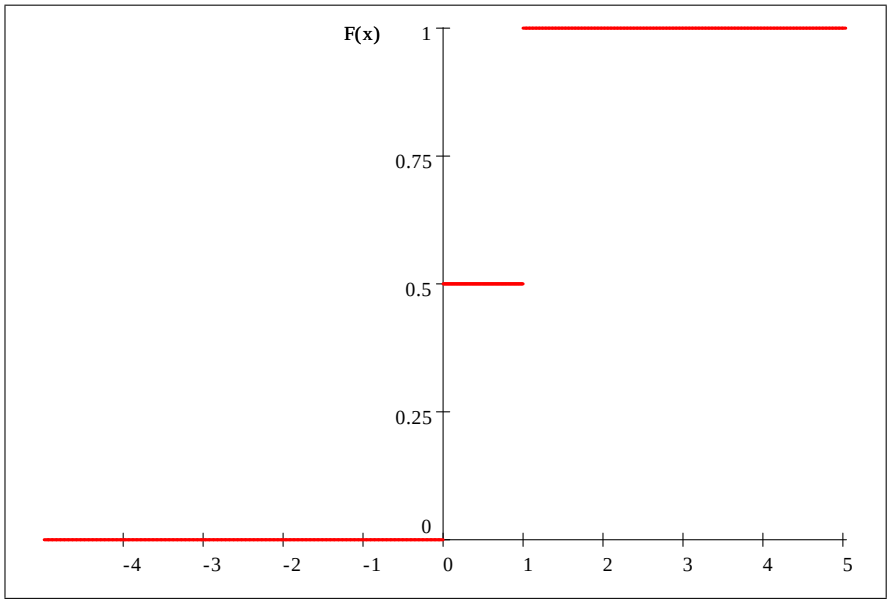


ベルヌーイの確率度数関数　$p = 0.7$

確率分布関数

$$F(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

確率分布関数のグラフ



ベルヌーイ分布の分布関数 $p = 0.5$

コイン投げの結果はベルヌーイ分布のひとつの例になる。コイン一回投げた場合の出方の分布は $p = 0.5$ のベルヌーイ分布である。

2. 二項分布：

確率度数関数

$$P(x) = C_n^x p^x (1-p)^{n-x} \qquad x = 0, 1, 2, ..., n.$$

Y_i を 0 になる確率が p のベルヌーイ分布に従うとする。$X = \sum_{i=1}^{n} Y_i$

$$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = np$$

$$V(X) = V\left(\sum_{i=1}^{n} Y_i\right) = n(p - p^2)$$

ただしここでは C_n^x は組み合わせの計算を表している。

$$C_n^x = \frac{P_n^x}{P_x^x} = \frac{n!/(n-x)!}{x!}$$
$$= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots 3 \times 2}$$

確率分布関数

$$F(x) = \sum_{i=1}^{x} P(x_i)$$

コイン投げはひとつ二項分布の例になる。コインを n 回を投げて x 回表が出る確率は $p = 0.5$ の二項分布になる。

3. ポアソン分布

確率度数関数

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

ポアソン分布の度数関数はに $p = \lambda/n$ の二項分布の度数関数の極限として導出できる。

証明. $p = \lambda/n$ として、証明する。

$$\begin{aligned} C_n^x p^x (1-p)^{n-x} &= \frac{n!/(n-x)!}{x!}\left(\frac{m}{n}\right)^x \left(1-\frac{m}{n}\right)^{n-x} \\ &= \frac{\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-(x-1))}{n}}{x!} m^x \left(1-\frac{m}{n}\right)^n \left(1-\frac{m}{n}\right)^{-x} \end{aligned}$$

ここで $\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-(x-1))}{n} \to 1, \left(1-\frac{m}{n}\right)^{-x} \to 1$ と $\left(1-\frac{m}{n}\right)^n \to e^{-m}$ を利用すれば、ポアソン分布の度数関数 $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ が導かれる。 □

$$E(X) = \lambda.$$
$$V(X) = \lambda.$$

確率分布関数

$$F(x) = \sum_{i=1}^{x} P(x_i)$$

19.1.2 ポアソン分布の応用例

ポアソン分布はパラメーター n を無限大にして確率 p を極めて小さくした二項分布の極限として考えられるため、起きる確率が極めて小さい出来事を分析するとき使わ

れる。たとえば、車の事故、地震が起きる確率など。たとえば、10 年間に平均的に 5 度強の地震は 10 回起きるとする。次の十年間に 5 度強の地震が 2 回以下起きる確率は

$$P(X \leq 2) = F(2) = \sum_{i=0}^{2} \frac{e^{-\lambda_i} \lambda^{x_i}}{x_i!}$$
$$= \frac{e^{-\lambda} \lambda^0}{0!} + \frac{e^{-\lambda} \lambda^1}{1!} + \frac{e^{-\lambda} \lambda^2}{2!}$$

ここで $\lambda = 10$ を代入して

$$= \frac{e^{-10} 10^0}{0!} + \frac{e^{-10} 10^1}{1!} + \frac{e^{-10} 10^2}{2!}$$
$$= 0.0028.$$

二回以下起きる確率は 0.0028 である。

演習問題：$p = 0.3, n = 10$ の二項分布の期待値、分散および $P(6)$ を計算してください。

20　確率変数と確率分布

二項分布と正規分布を例に確率分布に関して説明する。

20.1　ヒストグラムと確率度数（密度）関数（Probability Density Funtion PDF）

20.1.1　二項分布のヒストグラムと確率度数関数のグラフ

20 回コインを投げて、表が出た回数を記録し x とする、この実験を繰り返し 1000 回を行った、1000 個の x の値が得られる。x の 1000 個のデータを利用してヒストグラム（相対度数で描いたもの）を作成して、その上に二項分布の確率度数関数を重ね合わせたグラフ。

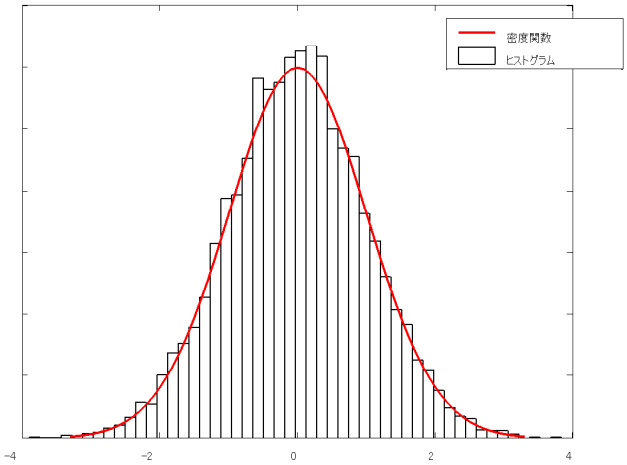
二項分布の確率度数関数

$$P(c)^{6} = C_n^c p^c (1-p)^n - c \qquad x = 0, 1, 2, ..., n. \tag{17}$$

[6]理解しやすくするため確率度数関数、密度関数、累積分布関数を $P(c), f(c), F(c)$ などで表しているが、統計学の慣習では $P(x), f(x), F(x)$ で表記する。

20.1.2　正規分布のヒストグラムと確率度数関数のグラフ

シミュレーションで作った標準化した身長のデータのヒストグラム（相対度数で描いたもの）を作成して、その上に標準正規分布の確率密度関数を重ね合わせたグラフ。正規分布の確率密度関数のグラフは鐘のまたは富士山の形をしている。



正規分布

正規分布の確率密度関数

$$f(c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(c-\mu)^2}{2\sigma^2}} \tag{18}$$

20.1.3　ヒストグラムと度数関数（密度関数）の意味合いの比較

ヒストグラム

ヒストグラムから確率変数の実現値（実験の結果）の度数が占める全体の度数の割合を示している。

度数関数（密度関数） $f(x)$

度数関数や密度関数のグラフは確率の大きさ を示している。

確率変数 X のある値 c での度数関数と密度関数の値が大きければ大きいほど、c に対応している確率大きい。

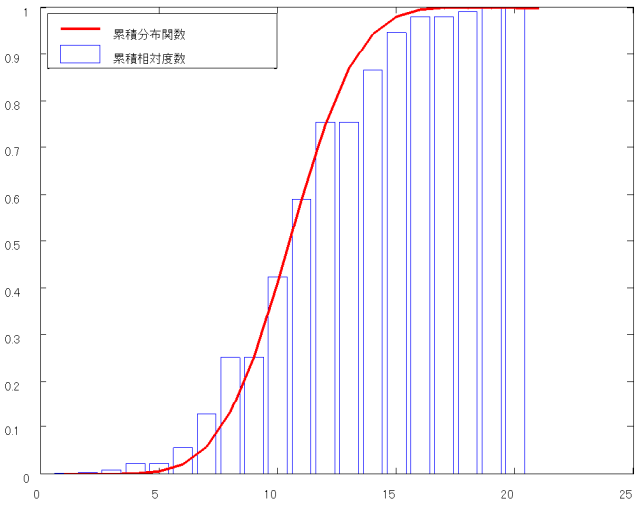
図 1 で言えば、値 10 に対応している確率がもっとも大きい。

図 2 の場合だと、0 に対応している確率密度がもっとも高い。0 近辺の値が起きやす

い。（連続確率関数の場合、特定の値に確率を対応させることができなく、区間に対して確率が振り分けられている。特定の値に対応させているのは確率密度である。）

21　累積相対度数と累積分布関数 (Cumulative Distribution Function CDF)

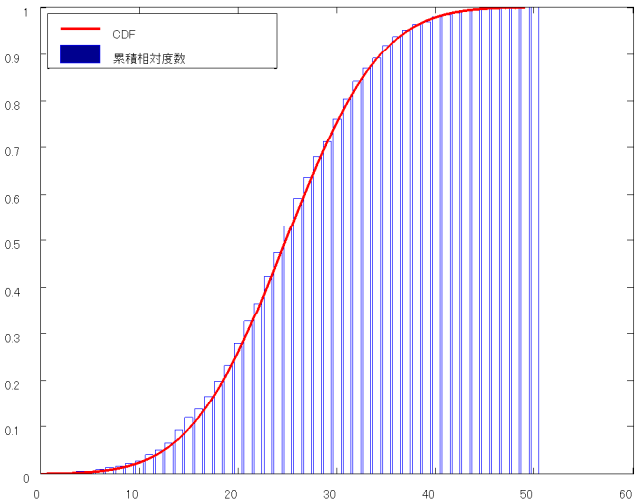
21.0.4　二項分布の累積相対度数と累積分布関数のグラフ



二項分布の累積分布関数

$$F(c) = P(X \leq c) = \sum_{i=1}^{c} P(x_i) = \sum_{i=1}^{c} C_n^{x_i} p^{x_i} (1-p)^{n-x_i} \qquad x = 0, 1, 2, ..., n. \quad (19)$$

21.0.5　標準正規分布の累積相対度数と累積分布関数のグラフ



正規分布の累積分布関数

$$F(c) = P(X \leq c) = \int_{-\infty}^{c} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{20}$$

21.0.6　累積相対度数関数と累積分布関数の意味合いの比較

累積相対度数

k 番目の階級の累積相対度数は k 番目の階級より低い階級（k を含む）の 度数の総和 である。

累積分布関数$F(x)$：

ある値 c に対応する累積分布関数の値は c より小さい値（c を含む）に対応する 確率の総和 である。

21.0.7　確率度数関数 $P(c)$（確率密度関数 $f(c)$）と累積分布関数 $F(c)$ の関係

離散確率変数：ある二つの値 a と b（$a < b$）が存在して、確率変数 X がこの二つの値 a と b を取れるが、a と b の間の値を取れない場合、X が離散確率変数であるという。
例：コイン投げの結果、サイコロの出る目。

連続確率変数：確率変数 X が取れる値の中から異なる任意の二つの値 c と d を取り出して、c と d の間の任意の値を取れるなら、X が連続確率変数である。例：人の身長、リンゴの重さ。

離散確率変数の場合

累積分布関数は確率密度関数の和の形になっている。数式で表すと

$$F(c) = P(X \leq c) = \sum_{i=1}^{c} P(x_i) \tag{21}$$

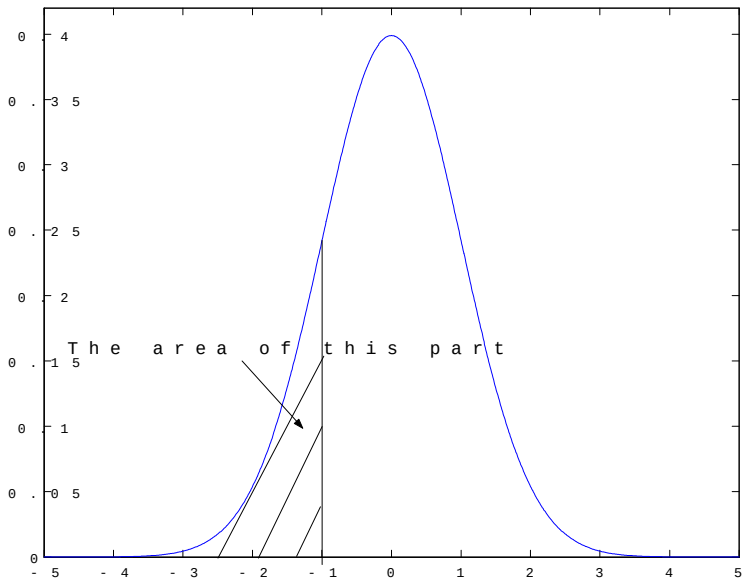
。

連続確率変数の場合

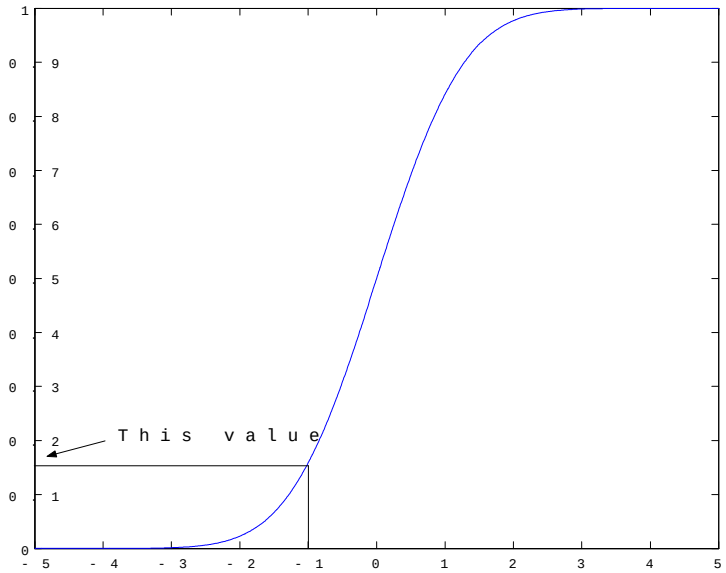
累積分布関数は確率密度関数の積分の形になっている。数式で表すと

$$F(c) = P(X \leq c) = \int_{-\infty}^{c} f(x)\, dx \tag{22}$$

積分が図形の**面積**の計算に対応していることから、連続確率変数の場合に関して、標準正規分布を例にグラフに描けば以下のようになる：



$F(-1)$ の値を密度関数のグラフで表す



$F(-1)$ の値を累積分布関数のグラフで表す

21.0.8 期待値と分散

前回は既に離散確率変数の期待値と分散に関して説明した。離散確率変数の場合

$$E(X) = \sum_{i=1}^{n} x_i P(x_i).$$

確率変数の**分散**：確率変数 X の分散を $V(X)$ で表して、離散確率変数の場合

$$V(X) = E\left[(X - E(X))^2\right] = \sum_{i=1}^{n} \left[(x_i - E(X))^2 P(x_i)\right].$$

連続関数の場合は

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

$$V(X) = E\left[(X - E(X))^2\right] = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx.$$

統計学用語のまとめ

標本の場合	度数	度数分布	累積相対度数	平均	分散（データから）
母集団の場合	確率	確率密度関数	累積分布関数	期待値	分散（密度関数から）

22　二項分布と正規分布のまとめ

1. 二項分布：

   確率度数関数

$$P(x) = C_n^x p^x (1-p)^{n-x} \qquad x = 0, 1, 2, ..., n. \tag{23}$$

   累積分布関数

$$F(x) = \sum_{i=1}^{x} P(x_i) \tag{24}$$

   期待値と分散

$$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = np \tag{25}$$

$$V(X) = V\left(\sum_{i=1}^{n} Y_i\right) = n(p - p^2). \tag{26}$$

2. 正規分布：

   確率密度関数

$$f(c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(c-\mu)^2}{2\sigma^2}} \tag{27}$$

   累積分布関数

$$F(c) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{28}$$

   期待値と分散

正規分布の期待値はμ 分散がσ^2：

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)\,dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= \mu \\ V(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= \sigma^2 \end{aligned}$$

22.1 練習問題

練習 54 講義資料 *10* の *Excel* の実験の小節を参考にして、*Excel* で標本数 10 の $p = 0.5$ のベルヌーイ分布の乱数を 100 列発生してください。各列の 1 の数を $x_i, i = 1, 2, ..., 100$ として、x_i のヒストグラムを作ってください。データ区間を $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ とする。

練習 55 *Excel* で正規分布の乱数を 1000 個発生して、データ区間を -4.2 から 4.2 まで間隔を 0.4 にしてヒストグラムを作ってください。

練習 56 二項分布の確率度数関数などの公式を利用して、$p = 0.3, n = 6$ の二項分布の期待値、分散および $P(3)$ を計算してください。

23 確率に関する計算の演習

確率に関する概念を理解するために練習問題を解く。公式などの参考資料は後ろのページにある。

練習 57 C_5^2 を筆算で計算してください。

練習 58 コインを *10* 回を投げて、*5* 回、*7* 回表が出る確率を計算してください。ステップを書いてから、計算は *Excel* を利用しても良い。

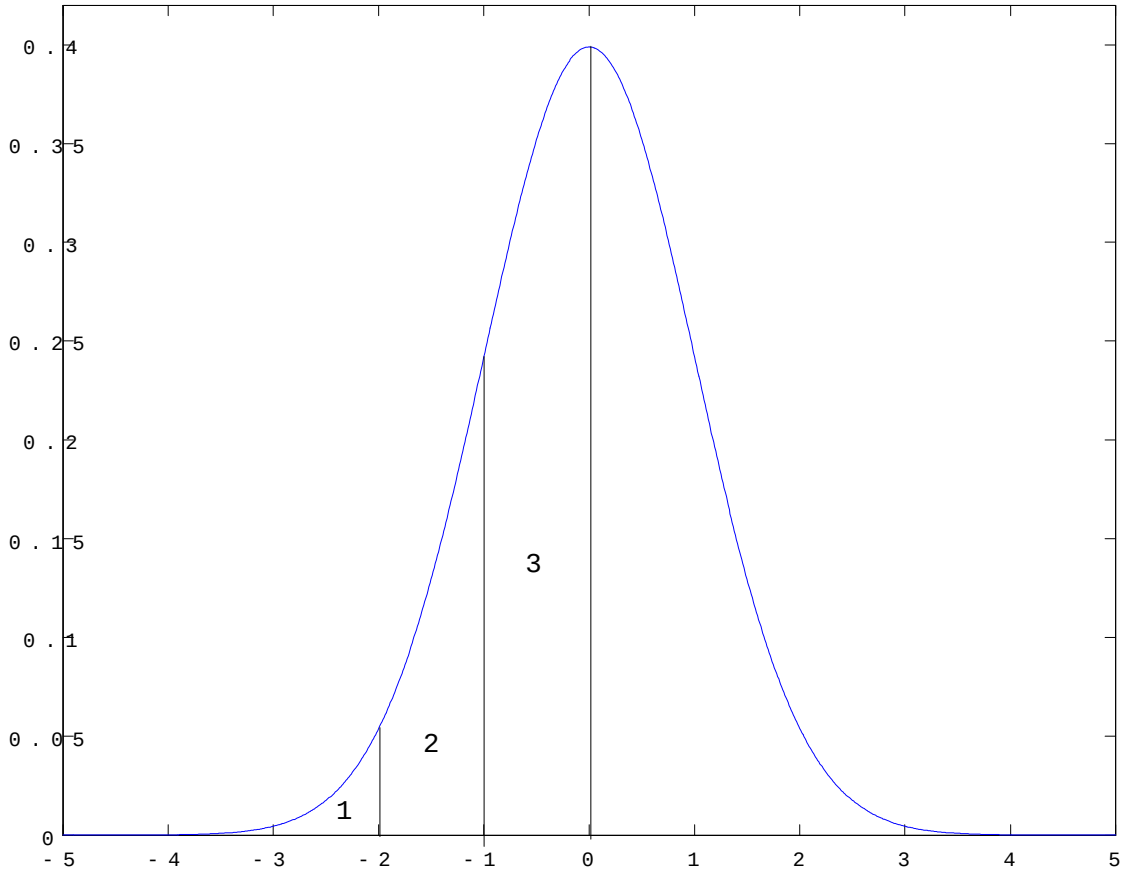


図 2: 図 1

練習 59 *Excel* を利用して、$p = 0.5, n = 10$ の二項分布の $x = 0, 1, 2, ..., 10$ の確率を計算して、その結果を用いて二項分布の確率度数関数と累積分布関数のグラフ（棒グラフ）を作成してください。

練習 60 図 *1* のような密度関数を持つ確率変数 X があるとする。0 までの図形を三つの部分に分けた。-2 以下の部分 (1)、-2 より大きくて -1 以下の部分 (2) と -1 より大きくて 0 以下の部分 (3) の面積はそれぞれ 0.0228、0.1359 と 0.3413 である。$P(X \leq -1)$ と $P(X \leq 0)$ を求めてください。$f(0), f(1)$ の値をグラフから読み取ろう。

練習 61 上の問題と同じ確率変数 X の確率累積分布関数は図 *2* のようになる。$F(-1)$ と $F(0)$ は幾つになるか。それは前の問題の $P(X \leq -1)$ と $P(X \leq 0)$ とはどのような関係を持つのかについて答えてください。

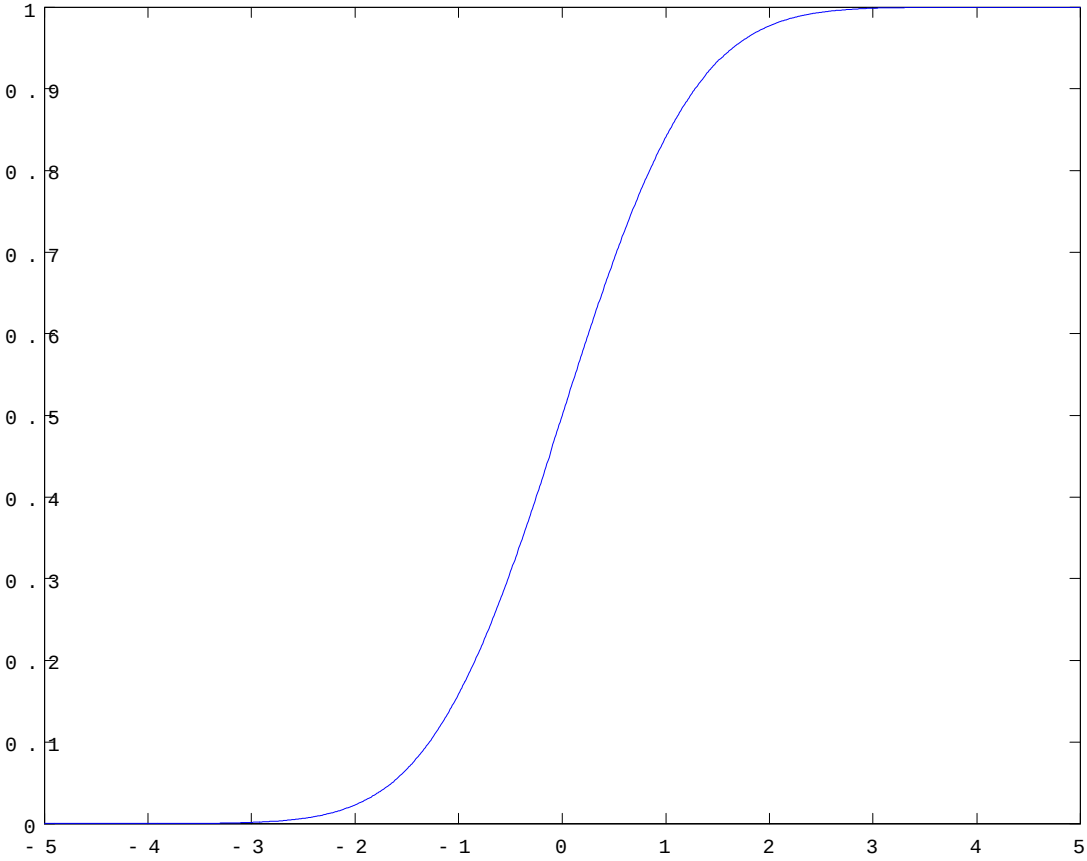


図 3: 図 2

練習 62 台風が平均的に年に *10* 回上陸するとし、次の *1* 年で台風が *5* 回しか上陸しない確率はいくらになるのかを計算してください。ヒント：ポアソン分布に従うとして計算する。

23.1 Excel のコマンドの説明

組合せ C_n^x または ${}_nC_x$ の計算：$= combin\,(n, x)$。
例：C_6^3 は $= combin\,(6, 3)$ となる。

24 二項分布と正規分布のまとめ

1. 二項分布：

   確率度数関数

$$P\,(x) = C_n^x p^x\,(1-p)^{n-x} \qquad x = 0, 1, 2, ..., n. \tag{29}$$

   累積分布関数

$$F\,(x) = \sum_{i=1}^{x} P\,(x_i) \tag{30}$$

   期待値と分散

$$E\,(X) = E\left(\sum_{i=1}^{n} Y_i\right) = np \tag{31}$$

$$V\,(X) = V\left(\sum_{i=1}^{n} Y_i\right) = n(p - p^2). \tag{32}$$

2. 正規分布：

   確率密度関数

$$f\,(c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(c-\mu)^2}{2\sigma^2}} \tag{33}$$

   累積分布関数

$$F\,(c) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \tag{34}$$

   期待値と分散

   正規分布の期待値はμ **分散が**σ^2。

3. ポアソン分布

確率度数関数

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{35}$$

ポアソン分布の度数関数はに $p = \lambda/n$ の二項分布の度数関数の極限として導出できる。

$$E(X) = \lambda. \tag{36}$$

$$V(X) = \lambda. \tag{37}$$

25　二項分布とポアソン分布の意味合いの再確認

二項分布の確率度数関数は成功する確率が p の実験を n 回繰り返したとき x 回成功する確率の計算となっている。

$$P(x) = C_n^x p^x (1-p)^{n-x} \qquad x = 0, 1, 2, ..., n. \tag{38}$$

ポアソン分布の確率度数関数の計算

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{39}$$

はあまり起こらない事象に関する分析でよく利用される。その値は、ある期間内で（1年や 5 日間、10 時間など）ある事象が平均的に λ 回起こる、その事象が同じ長さの期間内で x 回起こる確率の近似計算として使われる。

26　確率密度関数に関する計算

図 1 のような密度関数を持つ確率変数 X があるとする。0 までの図形を三つの部分に分けた。-2 以下の部分 (1)、-2 より大きくて -1 以下の部分 (2) と -1 より大きくて 0 以下の部分 (3) の面積はそれぞれ 0.0228、0.1359 と 0.3413 である。$P(X \leq -1)$ と

$P(X \leq 0)$ の値を答えなさい。$f(0), f(1)$ の値をグラフから読み取ろう。

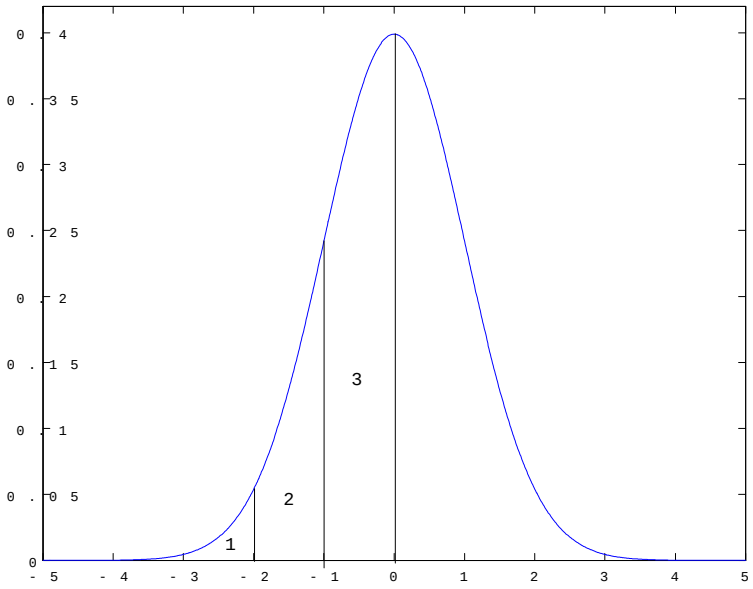


図 1

27　標準正規分布表

標準正規分布に従う確率変数 X があるとする。$P(X \leq c)$ を計算するとき、

$$P(X \leq c) = F(c) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{40}$$

で計算できるが、しかし c が変わるたびに計算が必要になるため、手間がかかる。そこで、標準正規分布表が用意されている。標準正規分布表から $P(X \leq c)$ の値が読み取れる。読み方は例題と標準正規分布表の配布資料を参考しよう。

標準分布表に載っている値は期待値 $\mu = 0$ 分散 $\sigma^2 = 1$ の標準正規分布に対応している。期待値が 0 ではない分散が 1 ではない正規分布の場合、その確率変数の値を標準化してから、表から確率を読み取る。

27.1　確率変数の標準化

期待値 $E(X)=\mu$、分散 $V(X)=\sigma^2$ の確率変数 X があるとする。X から期待値を引いて、標準偏差で割って、できた新しい確率変数

$$Z=\frac{X-\mu}{\sigma}$$

は標準化された確率変数で、その期待値 $E(Z)=0$、分散 $V(Z)=1$ となる。

例題 63 確率変数 X が標準正規分布（期待値 $\mu=0$ 分散 $\sigma^2=1$）に従う。X が 1.64 より小さい確率 $F(1.64)$（同じことで $P(X\leq 1.64)$）はいくら、$P(0<X\leq 1.64)$ はいくらなのか、標準正規分布表利用して答えよう。

まず、標準正規分布表の最初の列から 1.6 を見つけて、そして最初の行から 0.04 を見つける。1.6 が所在の行と 0.04 が所在の列の交点の値は X が 1.64 より小さい確率 $P(X\leq 1.64)=0.95$ となる。

次は $P(0<X\leq 1.64)=P(X\leq 1.64)-P(X\leq 0)$ なので、先と同じ要領で $P(X\leq 0)$ の値を捜し出し、$P(X\leq 1.64)$ の結果も利用すれば、計算できる。

結果は

$P(0<X\leq 1.64)=P(X\leq 1.64)-P(X\leq 0)=0.95-0.5=0.45$。

例題 64 確率変数 X が期待値 $\mu=1$ 分散 $\sigma^2=4$ の正規分布に従う。X が 4.28 より小さい確率はいくらなのか、標準正規分布表利用して答えよう。

期待値 $\mu=1$ 分散 $\sigma^2=4$ の正規分布確率変数 X を標準化すれば標準正規分布（期待値 $\mu=0$ 分散 $\sigma^2=1$）の確率変数になる。標準化された確率変数を Z と記する。

$$Z=\frac{X-\mu}{\sqrt{\sigma^2}}=\frac{X-1}{2}$$

Z に関して標準正規分布表を適用すればいい。$x=4.38$ のとき（小文字の x で X の実現値を表す）

$$z=\frac{x-1}{2}=\frac{4.28-1}{2}=1.64$$

ゆえに、 $P(X\leq 4.28)=P(Z\leq 1.64)=0.95$。

28　期待値と分散の推定

推定には点推定と区間推定があるが、本講義では点推定だけ説明する。

28.1　推定方法

確率変数 X の実現値 $\{X_1, X_2, X_3, ..., X_n\}$ を X の分布に従う母集団からの**無作為標本**と考えて、この母集団の期待値 $E(X)$（μ と記する）と分散 $Var(X) = \sigma^2$ をその無作為標本を用いて推定することについて考える。

定義 65 (無作為（at random）標本) 母集団のすべての個体が均等な機会でかつ互いに無関係（独立）に抽出されるように抽出された標本。公平なくじ引きで考えれば分かりやすいと思う。

期待値μ の推定方法：無作為標本の算術平均（標本平均）

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

を用いて推定する。

分散σ^2 の推定方法：

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

を用いて推定する。

例：確率変数 X:関西外大の学生の身長を確率変数と見なす。

母集団：関西外大の学生全員の身長。

無作為標本：関西外大の学生全員のくじを作り、くじ引きで 100 人を選び、身長を測って無作為標本 $\{X_1, X_2, X_3, ..., X_{100}\}$ とする。

$E(X)$ または μ：関西外大の学生の平均身長、母集団の平均。μ の推定量

$$\bar{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$$

で、σ^2 の推定量

$$S^2 = \frac{1}{100-1}\sum_{i=1}^{100}(X_i - \bar{X})^2$$

となる。

定義 66 (不偏推定量) ある母数 Γ の一つの推定量が γ とする。$E(\gamma) = \Gamma$ であれば、γ が Γ の不偏推定量であるという。

$\bar{X}$ と S^2 はそれぞれ μ と σ^2 の不偏推定量となっている。$\bar{X}$ に関して証明する。

$$E\left(\bar{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right)$$

どの X_i も X の標本であるため、その期待値 $E\left(X_i\right) = \mu$ である、この事実を利用して上式は

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

29　推定の根拠になる統計学の定理

上述した推定方法の正当性の根拠になる統計学の定理を紹介する。

定理 67 (大数の法則) 独立同一な分布に従う確率変数の平均は，サンプル数が大きくなるに従いその期待値に近づく。すなわち、各 X_i が平均（期待値）μ と分散 σ^2 を持つ独立同一な分布に従うとき，

$$\lim_{n\to\infty} \bar{X} = \mu \ .$$

言い換えれば、無作為標本の平均が母集団の平均（期待値）に収束する。

定義 68 (一致推定量) ある母数 Γ の一つ推定量が γ とする。$\lim_{n\to\infty} \gamma = \Gamma$ であれば、γ が Γ の一致推定量であるという。

大数の法則より、上述した $\bar{X}$ は期待値 μ の一推定量となる。

30　エクセルで大数の法則を確認する

1000 個の期待値が 2 分散が 4 の正規分布の乱数を発生して、1 番目の乱数までの平均、2 番目まで、3 番目まで、...、1000 番目までの平均を全部計算してください。そして、得られた 1000 個の平均の値で折れ線のグラフを作ってください。最後に平均がだんだん期待値 2 に近づくことを確認しよう。

31　平均値の検定

まず検定の理論の中で使われる概念を説明する。この節から、標本数がかなり大きいまたは母集団が正規分布に従うとする。

31.1　仮説

帰無仮説 H_0：棄却（否定）したい仮説。
対立仮説 H_1：採択し（認め）たい仮説。
例：速読の訓練の効果を調べたいとする。訓練を受けていると受けていない 100 人ずつの二グループの人に 1000 文字の小説を読ませて、
H_0：訓練を受けたグループの平均所要時間と受けていないグループの平均所要時間が同じ；
H_1:訓練を受けたグループの平均所要時間と受けていないグループの平均所要時間が異なる。

31.2　検定の根拠となる定理

定理 69 (中心極限定理) 独立同一な分布に従う確率変数の平均は，サンプル数が大きくなるに従いその期待値に近づく。すなわち、各 X_i が平均（期待値）μ と分散 σ^2 を持つ独立同一な分布に従うとき，n が大きくなるにつれ，$\sqrt{n}$ 倍した標準化した $\bar{X}$ が標準正規分布に収束する（近づく）。

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1).$$

31.3　母集団の標準偏差 σ が既知の場合の検定方法

普通は母集団の標準偏差 σ が未知で，前回で説明した方法，標本の標準偏差で推定しますが，ここでは説明しやすくするために、取り敢えず σ が既知として話を進める。

31.4　片側検定

平均の検定には**片側検定**と**両側検定**がある、片側検定は平均がある値より大きいかどうか、または小さいかどうかに関して別々に検定する。両側検定の場合、両方を同時に検定する。今から片側検定について説明するが、基本的な考え方が同じである。

ここで，例を挙げながら説明する。

例 70 外大の学生の身長 X を例に説明する。外大の学生の身長の期待値（全員の平均）に関して検定することを考える。100 人をくじ引きで選んで（無作為標本)、身長を測り、その平均は $\bar{X} = 175$ になったとする。全員の身長の標準偏差 $\sigma = 10$ とする。期待値 μ が 170 より大きいかどうかを検定する。

この場合

帰無仮説 H_0：期待値 $\mu = 170$；

対立仮説 H_1：期待値 $\mu > 170$。

基本的な考え方は、仮に X が平均（期待値）が $\mu = 170$、標準偏差が $\sigma = 10$ の正規分布 $N(170, 100)$ に従うとする、その場合、$\bar{X}$ が 175 を超える確率がいくらになるのかを計算する。もし、この確率がとても小さくて、前もって決めた許容値 α（**有意水準**と呼ばれる、通常 $\mathbf{1}\%$ か $\mathbf{5}\%$ とする）よりも小さいなら、$\mu = 170$ の仮説が偽であると認識する（**棄却する**)。逆の場合、真と認識する。$\sqrt{n}$ 倍した標準化した $\bar{X}$ を Z_0 とする、Z_0 が標準正規分布に収束することを考え、標準正規分布の性質を利用する。

一般的な場合**検定の手続き**は以下のようになる：

1. 仮説を立てる。

   例では、帰無仮説 H_0：期待値 $\mu = 170$；

   対立仮説 H_1：期待値 $\mu > 170$。

2. 有意水準 α を決める。

   例では $\alpha = 5\%$ とする。

3. 有意水準 α に対応している標準正規分布の点を標準正規分布表から読み取る、すなわち Z が標準正規分布に従うとして、$P(Z > Z^*) = \alpha$ に対応している Z^* がいくらなのかを読み取る。

   例では標準正規分布の $P(Z > Z^*) = 5\%$ に対応している Z^* の値を読み取る、これは $P(Z \leq Z^*) = 1 - 5\%$ に対応している Z^* の値と同じ。

   例の場合約 $Z^* = 1.65$。

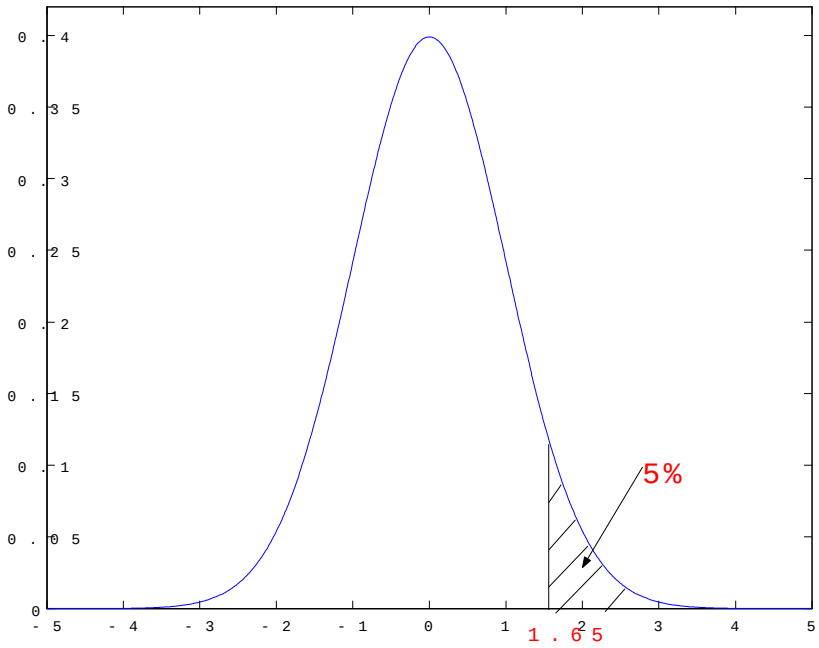
4. $Z_0 = \sqrt{n}\left(\bar{X} - \mu\right)/\sigma$ を計算する。

   例では $Z_0 = \sqrt{100}\,(175 - 170)/10 = 5$。

5. ステップ 3 で求めた Z_0 とステップ 2 で求めた Z^* と比較する、$Z_0 > Z^*$ であれば、帰無仮説が否定（棄却）される。$Z_0 \leq Z^*$ であれば帰無仮説が採択される（認められる）。

   例では $Z_0 = 5, Z^* = 1.65$ で $Z_0 > Z^*$ なので、帰無仮説が否定（棄却）される、すなわち、$\mu = 170$ ではない。対立仮説が採択され、$\mu > 170$ であろうと考えられる。

グラフで示すなら、Z_0 が標準正規分布の裾の端にはいて Z^* よりも右にあったら、$\mu = 170$ の仮説が偽であると認識する。



31.5　演習問題

ある美容室が割引サービスを行った、この割引サービスによって、一日平均の来客数が増えたかどうかを調べたい。この美容室の普段の平均来客数が 10 人、来客数の分散 $\sigma^2 = 9$ だとする。割引サービスを実施後、25 日間来客数を集計して平均を計算して $\bar{X} = 12$ だとする。検定を行って一日平均の来客数が増えたかどうかを判断してください。ヒント：H_0：$\mu = 10$；H_1：$\mu > 10$。

31.6　課題（締め切り　7月5日）

仮にDATA01の中の身長のデータを日本人の無作為標本のデータとする。日本人の平均身長μは160 cmより高いかどうか検定してください。ただし、$\sigma = 6$とする。

32　平均値の検定

32.1　両側検定：σが既知の場合

ナットのサイズの平均の検定を例に説明する。

例として、工場から送ってきた大量なナットを検査することを考える。納品の中から100個のナットを無作為に抽出して、サイズを図り平均を計算した$\bar{X} = 3.2\,\mathrm{cm}$、正常なときの分散$\sigma^2 = 4$で、平均は$\mu = 3$、時には需要者側から$\mu$と$\sigma^2$の値が良品であるかどうかの判断基準として提供される。$\bar{X}$の値を利用して、$\mu = 3$であるかどうかの検定を考える。

一般的な場合**検定の手続き**を上の例を用いながら説明する。

1. 仮説を立てる。

   例では、ナットを入荷する先の側に立って考えれば、ある意味で不良品を検出したいので、棄却したい仮説、帰無仮説はH_0：期待値$\mu = 3\,\mathrm{cm}$；ナットのサイズが標準より大きくなっても小さくなっても良くないので、対立仮説H_1：期待値$\mu > 3\,\mathrm{cm}$または$\mu < 3\,\mathrm{cm}$、すなわち$\mu \neq 3cm$。

2. 有意水準αを決める。「より大きい」、と「より小さい」両方に半々に分ける、$1/2 * \alpha$になる。

   例では$\alpha = 5\%$とする。半分にして、$1/2 * \alpha = 2.5\%$になる。

3. 半分の有意水準$1/2 * \alpha$に対応している標準正規分布の点を標準正規分布表から読み取る、すなわちZが標準正規分布に従うとして、$P\left(Z > Z^*\right) = 1/2 * \alpha$に対応している$Z^*$がいくらなのかを読み取る。

   例では標準正規分布の$P(Z > Z^*) = 2.5\%$に対応しているZの値を読み取る、これは$P(Z \leq Z^*) = 1 - 2.5\%$に対応している$z$の値と同じ。

   例の場合約$Z^* = 1.96$。

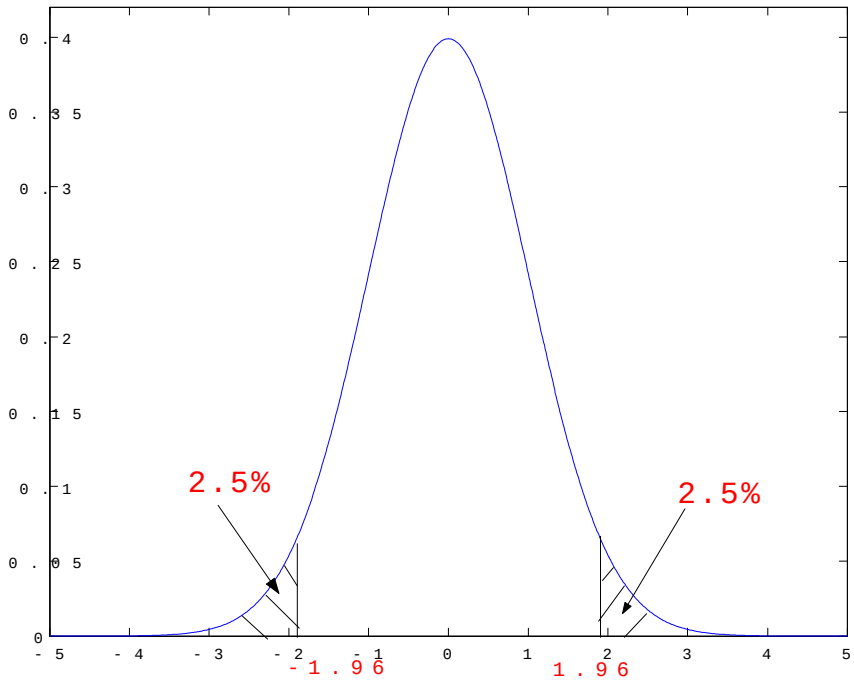
4. $Z_0 = \sqrt{n}\left(\bar{X} - \mu\right)/\sigma$を計算する。

   例では$Z_0 = \sqrt{100}\left(3.2 - 3\right)/2 = 1$。

5. ステップ 3 で求めた Z_0 とステップ 2 で求めた Z^* と比較する、$Z_0 > Z^*$ または $Z_0 < -Z^*$ であれば、帰無仮説が否定（棄却）される。$-Z^* \leq Z_0 \leq Z^*$ であれば帰無仮説が採択される（認められる）。

   例では $Z_0 = 1, Z^* = 1.96$ で $Z_0 < Z^*$ なので、帰無仮説 $\mu = 3$ が否定（棄却）されず、すなわち、今回の納品は問題がないと判断できる。

グラフで示すなら、Z_0 が標準正規分布の裾の端にはいて Z^* よりも右または $-Z^*$ よりも左にあったら、$\mu = 3$ の仮説が棄却される。



33　試験時間：最終の講義、7月22日。

34　無断欠席した場合試験の成績が0点。

35　公欠の場合:事前に連絡ください。レポート課題を出す。

36　平均値の検定

36.1　練習問題

ある弓道の選手が監督からひとつ新しい方法を教えてもらった。このコツを使う前に、この選手の平均点数は大体9.1であった。このコツを使って、9回打ちました、得点はそれぞれ$8.4, 9.3, 8, 7, 8.9, 9.8, 9.3, 9.2, 9, 8.9$であった。この方法は効果があるのかまたは逆効果があるのかに関して仮説を立てて両側検定を有意水準$\alpha = 1\%$で、行ってください。（ただし）ここでは分散$\sigma^2 = 4$とする。

37　平均値の検定σが未知の場合

これからの検定の話は標本数nが十分に大きいか、または母集団が正規分布に従うとする。

37.1　検定の根拠

ここまではσが既知の場合の平均値の検定を説明してきた。復習になるが、σが既知の場合

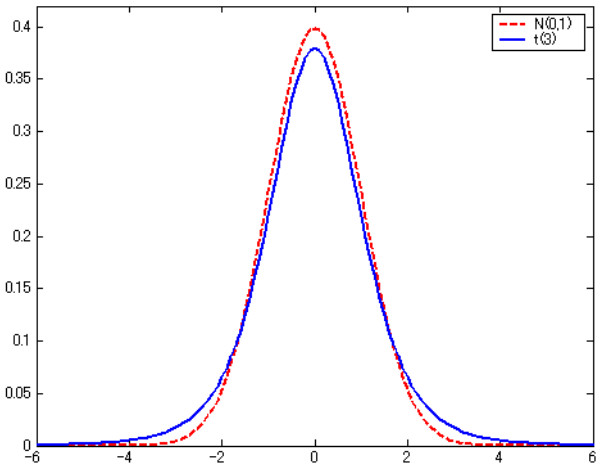
$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

σ極一部の例外を除けば普通の場合は未知である。そのときσの代わりにσの推定量s（標本誤差、標本分散の平方根）を使う。そうすると

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s} \xrightarrow{d} N(0, 1)$$

は成り立たない、代わりに

$$\frac{\sqrt{n-1}(\bar{X} - \mu)}{s} \xrightarrow{d} t_{(n-1)}.$$



となる。ただし、$t_{(n-1)}$ は自由度 $n-1$ の t 分布を表す。

$\frac{\sqrt{n-1}(\bar{X}-\mu)}{s}$ を t で表して、言葉で言うと、t が自由度 $n-1$ の t 分布に従う。

37.2　t 分布

t 分布の密度関数のグラフは正規分布のグラフの形に似ている。

正規分布は平均と分散によって密度関数が決まるが、t 分布は正規分布と違って自由度というパラメーターにより密度関数が決まる。自由度が大きくになるにつれ、t 分布は標準正規分布に近づき、密度関数のグラフが同じようになっていく。

下のグラフは標準正規分布の密度関数のグラフ（赤い点線）と自由度が 3 の t 分布の密度関数のグラフ（ブルーの実線）

37.3　片側検定：

σ が既知の場合、検定の手続きの中で標準正規分布表を利用していたが、σ が未知の場合は t 分布表を利用する。t 分布表の読み方は表を参考してください。説明したように検定の根拠となっている理論が違うが、検定の手順が σ が既知のときとほとんど同じである。具体な手続きを例を挙げながら以下のように説明する。

例 71 外大の学生の身長 X を例に説明する。外大の学生の身長の期待値（全員の平均同じく母集団の平均）に関して検定することを考える。25 人をくじ引きで選んで（無

作為標本)、身長を測り、データとして記録した。期待値 μ が 170 より大きいかどうかを検定する。

1. 仮説を立てる。

   例では、帰無仮説 H_0：期待値 $\mu = 170$；

   対立仮説 H_1：期待値 $\mu > 170$。

2. 有意水準 α を決める。

   例では $\alpha = 5\%$ とする。

3. 自由度を計算する。自由度 $\upsilon = n - 1$。

   例では $\upsilon = 25 - 1 = 24$。

4. 有意水準 α に対応している 自由度が $n-1$ の t 分布 の点を t 分布表 から読み取る、すなわち t が標準正規分布に従うとして、$P\left(t > t^*\right) = \alpha$ に対応している t^* がいくらなのかを読み取る。

   例では自由度 24 の t 分布で、$P(t > t^*) = 5\%$ に対応している t^* の値を読み取る、約 $t^* = 1.71$。

5. データを利用して、$\bar{X}$ と s を計算する。

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

   例では仮に計算した結果 $\bar{X} = 175, s^2 = 100$ とする。

6. $t_0 = \sqrt{n-1}\left(\bar{X} - \mu\right)/s$ を計算する。

   例では $t_0 = \sqrt{24}\left(175 - 170\right)/10 \approx 2.5$。

7. ステップ 6 で求めた t_0 とステップ 4 で求めた t^* と比較する、$t_0 > t^*$ であれば、帰無仮説が否定（棄却）される。$t_0 \leq t^*$ であれば帰無仮説が採択される（認められる)。

   例では $t_0 = 2.5, t^* = 1.71$ で $t_0 > t^*$ なので、帰無仮説が否定（棄却）される、すなわち、$\mu = 170$ ではない。対立仮説が採択され、$\mu > 170$ で外大の学生の平均身長は 170 cm 以上であろう。

37.4　両側検定：

例を省略して手続きだけ示す。

1. 仮説を立てる。

   帰無仮説 H_0：期待値 $\mu = \mu_0$

   対立仮説 H_1：$\mu \neq \mu_0$。

2. 有意水準 α を決める。「より大きい」、と「より小さい」両方に半々に分ける、$1/2 * \alpha$ になる。

3. 自由度を計算する。自由度 $\upsilon = n - 1$。

   例では $\upsilon = 25 - 1 = 24$。

4. 半分の有意水準 $1/2 * \alpha$ に対応している 自由度が $n-1$ の t 分布 の点を t 分布表 から読み取る、すなわち t が標準正規分布に従うとして、$P(t > t^*) = \alpha$ に対応している t^* がいくらなのかを読み取る。

5. データを利用して、$\bar{X}$ と s を計算する。

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

6. $t_0 = \sqrt{n-1}\left(\bar{X} - \mu\right)/s$ を計算する。

7. ステップ６で求めた t_0 とステップ４で求めた t^* と比較する、$t_0 > t^*$ または $t_0 < -t^*$ であれば、帰無仮説が否定（棄却）される。$t_0 \leq t^*$ かつ $t_0 \geq -t^*$ であれば帰無仮説が採択される（認められる）。

練習 72 *DATA01* の中にある身長のデータを日本人の身長の無作為標本として、前半 9 個のデータを使って、日本人の平均身長 $\mu > 160$ であるかどうかの片側検定を行ってください。（手計算）

38　平均値の検定

38.1　両側検定：σ が既知の場合

ナットのサイズの平均の検定を例に説明する。

例として、工場から送ってきた大量なナットを検査することを考える。納品の中から 100 個のナットを無作為に抽出して、サイズを図り平均を計算した $\bar{X} = 3.8\,\mathrm{cm}$、正常なときの分散 $\sigma^2 = 4$ で、平均は $\mu = 3$、時には需要者側から μ と σ^2 の値が良品であるかどうかの判断基準として提供される。$\bar{X}$ の値を利用して、$\mu = 3$ であるかどうかの検定を考える。

一般的な場合**検定の手続き**を上の例を用いながら説明する。

1. 仮説を立てる。

   例では、ナットを入荷する先の側に立って考えれば、ある意味で不良品を検出したいので、棄却したい仮説、帰無仮説は H_0：期待値 $\mu = 3\,\mathrm{cm}$；ナットのサイズが標準より大きくなっても小さくなっても良くないので、対立仮説 H_1：期待値 $\mu > 3\,\mathrm{cm}$ または $\mu < 3\,\mathrm{cm}$、すなわち $\mu \neq 3cm$。

2. 有意水準 α を決める。「より大きい」、と「より小さい」両方に半々に分ける、$1/2 * \alpha$ になる。

   例では $\alpha = 5\%$ とする。半分にして、$1/2 * \alpha = 2.5\%$ になる。

3. 半分の有意水準 $1/2 * \alpha$ に対応している標準正規分布の点を標準正規分布表から読み取る、すなわち Z が標準正規分布に従うとして、$P\left(Z > Z^*\right) = 1/2 * \alpha$ に対応している Z^* がいくらなのかを読み取る。

   例では標準正規分布の $P(Z > Z^*) = 2.5\%$ に対応している Z の値を読み取る、これは $P(Z \leq Z^*) = 1 - 2.5\%$ に対応している z の値と同じ。

   例の場合約 $Z^* = 1.96$。

4. $Z_0 = \sqrt{n}\left(\bar{X} - \mu\right)/\sigma$ を計算する。

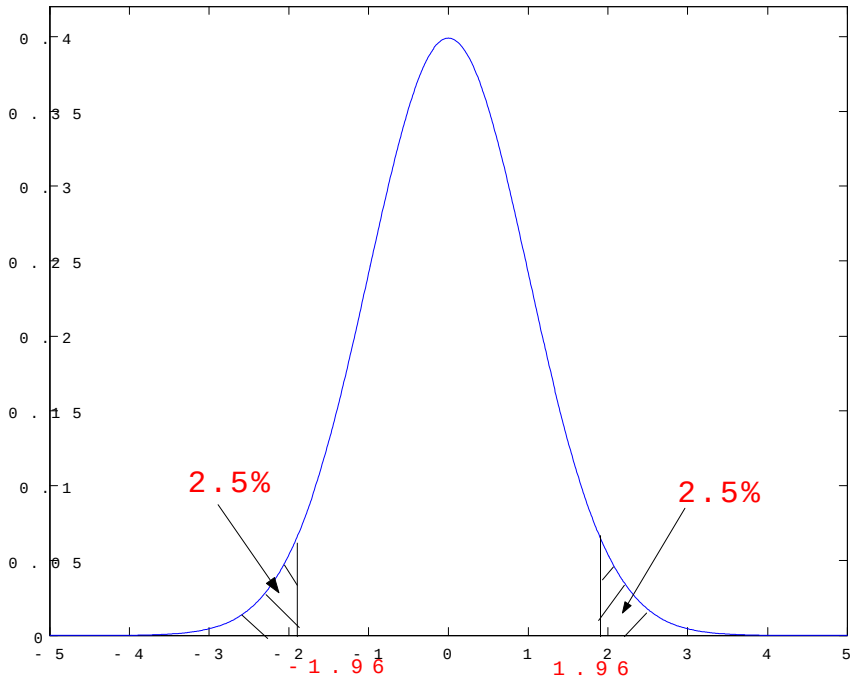
   例では $Z_0 = \sqrt{100}\left(3.2 - 3\right)/2 = 1$。

5. ステップ 3 で求めた Z_0 とステップ 2 で求めた z と比較する、$Z_0 > Z^*$ または $Z_0 < -Z^*$ であれば、帰無仮説が否定（棄却）される。$Z_0 \leq Z^*$ または $Z_0 \geq -Z^*$ であれば帰無仮説が採択される（認められる）。

   例では $Z_0 = 1, Z^* = 1.96$ で $Z_0 < Z^*$ なので、帰無仮説 $\mu = 3$ が否定（棄却）されず、すなわち、今回の納品は問題がないと判断できる。

グラフで示すなら、Z_0 が標準正規分布の裾の端にはいて Z^* よりも右または $-Z^*$ よ

りも左にあったら、$\mu = 3$ の仮説が棄却される。



39　σ が未知の場合の両側検定の演習問題

ある種の部品を大量に入荷した会社は今回入荷した商品を検査したい。商品の中から 4 個抜き取って、サイズを測った。それぞれ 10 cm, 12 cm, 15 cm, 9 cm だった。商品のサイズの分布が正規分布に従うとして、平均が 13 cm ぐらいであれば合格とする。両側検定を行って良品として入荷して良いのかを判断してください。

40　復習

問題 73 以下の二つのデータの系列に関して、分散、標準偏差、共分散、相関係数を計算してください。

$$X = \{3,\ 6,\ 9\}$$
$$Y = \{2,\ 3,\ 8\}$$

問題 74 コインを三回投げて、最初の一回が裏で次の二回が表である確率を計算してください。

問題 75 サイコロを二回転がって、出た目の和が 10 になる確率を計算してください。

問題 76 （架空の例）環境ホルモンが有害であるかどうかを調べるために 100 匹のラットに環境ホルモンが混入している薬剤を飲ませた。普通のラットの平均寿命は 23ヶ月だが、この 100 匹の平均寿命は 21ヶ月となった。ラットの寿命の分布が正規分布に従うとして、分散を 4 とする。有意水準 1% で、環境ホルモンよってラットの寿命が縮んだかどうか検定しなさい。

問題 77 （架空の例）地球温暖化を調べるため、ある地域の 7 月の平均気温を 10 年間に渡り計測した。この 10 年間の 7 月の平均気温の平均は 31 度だった。7 月の平均気温は分散 1 の正規分布に従うとする。昔の例年の 7 月の平均気温が平均的に 29 度であるとして、自分で有意水準を決め、気温が上昇したかに関して、検定しなさい。

41　単回帰分析における検定の紹介

単回帰モデル（41）が一定の条件（本講義の範囲を超えるため詳細を省略する、詳細を知りたい場合統計学入門 [1] を参考にしてください）を満たしているときその係数の最小二乗推定量 $\hat{b}$ は正規分布に従う。その性質を利用して検定できる。

$$Y_i = a + b \times X_i \tag{41}$$

その時、

$$t \equiv \hat{b} \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$$

とすれば、t が自由度 $n-2$ の t 分布に従う。以下の問題の中で Excel で行った回帰分析の結果にある t 検定を紹介する。

41.1　問題

DATA03 をダウンロードし、まずデータの基本統計量と共分散および相関係数を Excel で計算して下さい。ビールの消費量と平均気温の間はどのような関係があるのかを計算の結果を持って説明しよう。

そして、仮に 2007 年の 5 月に生産計画を立てることになっているとする。2007 年 7 月の平均気温の予測値は 37 度として、2007 年にこのブランドのビールをどのぐらい生産すればいいでしょう？Excel で分析ツールで回帰分析を行って回答しよう。

出力の結果の見方

概要

回帰統計	
重相関 R	0.641785
重決定 R2	0.411887
補正 R2	0.390883
標準誤差	20.70401
観測数	30

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	8405.914	8405.914	19.60993	0.000132
残差	28	12002.37	428.656		
合計	29	20408.28			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	94.39179	171.995	0.548805	0.587489	−257.924	446.7081	−257.924	446.7081
X 値 1	20.47776	4.624284	4.42831	0.000132	11.00534	29.95019	11.00534	29.95019

最初の黄色の部分の値は求めたい係数です。たとえば y をビールの消費量として、x を 7 月の平均気温とする。両者の関係は

$$y = a + b \times x + u$$

（ただし、u は撹乱項と呼ばれる。気温以外の不確実な要素の影響を表す。）と仮定して回帰分析を行った場合、a の推定量は切片の値で 94.4 ぐらい、b は表の中の「X 値　1」に対応している値で 20.5 ぐらい。推定された式は

$$y = 94.4 + 20.5x$$

となる。x に 37 を代入して予測値が 852.9 となることが分かる。

出力の結果の表の中の黄色い部分にある t の列にある値は t 値で、p の列に対応している値は t 値が到達した p 値である。簡単に説明すれば p 値が小さければ係数の信頼性が低くなると考えられる。具体的な検定方法はパワーポイントの資料を参照ください。

参考文献

[1] 森棟公夫「統計学入門」第二版、新世社、(2002)。