SSH 科目「のぞみ」テキスト

統計入門/続・統計入門

大阪府立大手前高等学校

統計入門

1　はじめに

このテキストは1年次「プレ・サイエンス探求」の一環である「統計入門講座」のために作成した。5回の講座を通じて,「記述統計」と呼ばれる部分を主に扱い,「統計的推測」への入り口まで到達することを目標とする。引き続き，2年次SSH科目「のぞみ」において統計的推測について取り上げる予定である。

授業では，できるだけ手を動かして実際に作業をすることを重視する。やがて将来，大規模なデータをコンピュータを利用して扱うようになるとしても，そこで原理的にどのようなことが行われているのかを理解しておくことは結果の解釈をするためにも不可欠と考えるからである。そのような原理の理解のためには，小規模な「おもちゃのデータ」を簡単な電卓程度の補助のもとに自分の手で扱い，触ってみることが近道である。

2　データ（data）

データとは，実験，観察，調査などによって収集された情報を数値その他で具体的に表現したものをいう。データの中には，性別・居住する都道府県・職業など，数値以外の表現がなされるものもある。数値で表されるデータを**量的データ**，後者のように通常は数値で表されないデータを**質的データ**という。このテキストでは当面，量的データを中心に扱う。

例 2.1. 量的データには，たとえば長さ，面積，体積，重さ，時間，温度，金額，などがある。

例 2.2. 質的データには，たとえば性別，居住地，職業，天候（晴れ，雨，曇り），などがある。

数値で表されるデータが収集される場面は非常に幅広く，人間活動のさまざまな分野にかかわりを持つ。他書を参考に，いくつか例を挙げてみる。

例 2.3. さまざまな分野からのデータの例

(1) 淀川の月間流出水量の10年間120ヶ月分のデータ
(2) 全国100地点で採取した雨水の酸性度
(3) シェークスピアの戯曲に使われている単語の長さの分布
(4) 糖尿病で治療を受けている患者30人の，それぞれの空腹時血糖値
(5) 結核菌を注射されたモルモット80匹の生存時間
(6) ボストンマラソンに女性の参加が認められた1972年から現在までの，女性の優勝記録

このように，データはそれぞれ具体的な意味を持つ数値であるが，当面の間はさまざまな背景を持つデータに対して共通に適用できる考え方を学ぶため，個々の意味を忘れて数値の集まりをデータと考えて取り扱う。

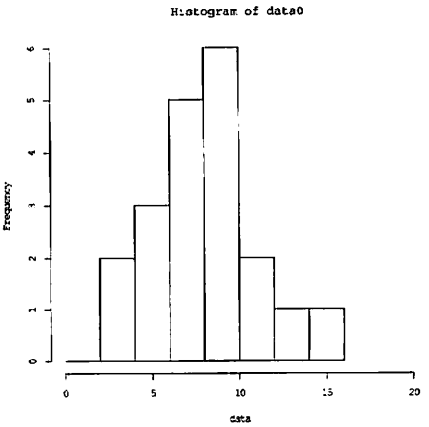
3　分布（distribution）

収集したデータは，通常，ただその数値を列挙しただけでは特徴がつかみにくい。データの特徴をつかむために最初に考えるべきことは，一組のデータ・セットから度数分布を求めヒストグラムを作ることである。

3, 7, 8, 4,
12, 9, 7, 8,
10, 3, 6, 8, $\Longrightarrow$
7, 5, 6, 9,
11, 5, 8, 15

階級	度数
2 以上 4 未満	2
4 以上 6 未満	3
6 以上 8 未満	5
8 以上 10 未満	6
10 以上 12 未満	2
12 以上 14 未満	1
14 以上 16 未満	1
合計	20

$\Longrightarrow$



左端のデータ・セットから度数分布が作られ，右端のヒストグラムが出来上がった。この過程を逆にたどろうとしても，右端のヒストグラムから左端のデータ・セットを復元することはできない。この意味で，最初のデータ・セットが最も詳しく情報量が多いはずだが，生の数値の羅列をみてデータの特徴をつかむのは難しい。**データは適切に加工され表現されて始めて意味を読み取ることができるようになる。**分布の形をよく見ることはその第一歩である。

4　分布の特徴を少数の数値で表す

分布の形を見ることは重要だが，複数の分布を比較したり，さらに加工して吟味するためには，数値で表現したほうが便利である。データの特徴を少数の数値で表現するには「どのあたりに」「どの程度広がって」分布しているかを表せばよい。そのための方法を 2 通り扱う。

平均と分散・標準偏差を用いる方法

平均によって「どのあたりに」を表現し，分散やその平方根である標準偏差で「どの程度広がって」を表現する。理論的に扱いやすく，統計的推測の理論を学ぶ際にも活躍する。

中央値と四分位数および最大・最小値による 5 数要約を用いる方法

中央値で「どのあたりに」を表現し，四分位数で「どの程度広がって」を表現する。

第一回の目標

No.1 で説明なしに現れた言葉について，No.2 以下で定義を理解し，簡単なデータ・セットについて計算する。最後に再度 No.1 を読み，その意味がわかるようになることが目標である。

5　度数分布およびヒストグラムの作成

例 5.1. データ・セット 一組のデータの集まりをデータ・セットという。

58	58	53	40	49	47	56	46	60	46	61	40	37	82	46	53	56
45	55	54	48	51	50	71	43	39	50	53	54	45	39	63	47	41
48	48	61	51	58	45	52	47	51	41	37	70	56	37	44	38	72
63	47	55	46	45	42	44	67	49	51	52	62	45	40	67	46	43
38	37	44	56	61	57	46	51	43	43	59	40	70	49	52	43	44
37	48	54	53	42	42	45	65	39	48	69	49	36	43	55		

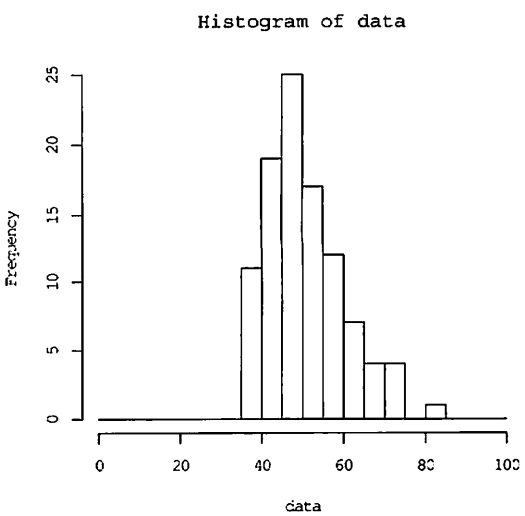
度数分布表（相対度数・累積度数・相対累積度数つき）

データの値のとりうる範囲をいくつかの区間（**階級**, class）に分割し，各階級に含まれるデータの個数（**度数**, frequency）を表にしたものを**度数分布表**という。各階級内のデータの個数の，データ総数に対する割合を**相対度数**という。**累積度数**とは，ある階級以下の度数の合計であり，累積度数のデータ総数に対する割合を**累積相対度数**という。右図は，例 5.1 に示したデータ・セットの度数分布表（相対度数・累積度数・相対累積度数つき）である。

階　級	度数	相対度数 (%)	累積度数	累積相対度数 (%)
35 以上 40 未満	11	11.0	11	11.0
40 以上 45 未満	19	19.0	30	30.0
45 以上 50 未満	25	25.0	55	55.0
50 以上 55 未満	17	17.0	72	72.0
55 以上 60 未満	12	12.0	84	84.0
60 以上 65 未満	7	7.0	91	91.0
65 以上 70 未満	4	4.0	95	95.0
70 以上 75 未満	4	4.0	99	99.0
75 以上 80 未満	0	0.0	99	99.0
80 以上 85 以下	1	1.0	100	100.0
合計	100	100.0		

ヒストグラム

度数分布表にもとづき，データの取りうる値を横軸にとり，右図のようにグラフで表したものをヒストグラム（柱状グラフ, histogram）という。各階級に対して，階級の幅を柱の幅にとり，面積が度数に比例するように柱の高さを定める。通常は階級の幅を一定にし，度数を柱の高さとする。



演習 5.1. 次のデータ・セットの階級幅を 10 としたときの度数分布表をつくり，ヒストグラムを書きなさい。

データ・セット

82	66	63	74	71	58	60	56	73	55	62	47	74	73	50
48	68	58	59	67	70	51	38	68	47	61	48	61	59	74

度数分布表（相対度数・累積度数・相対累積度数つき）

階級	度数	相対度数	累積度数	累積相対度数
0 以上 10 未満				
10 以上 20 未満				
20 以上 30 未満				
30 以上 40 未満				
40 以上 50 未満				
50 以上 60 未満				
60 以上 70 未満				
70 以上 80 未満				
80 以上 90 未満				
90 以上 100 以下				
合計				

ヒストグラム

6　平均・分散・標準偏差

6.1　データ・セットからの平均・分散・標準偏差の計算

n 個のデータからなるデータ・セット $x_1, x_2, \cdots, x_n$ の平均（mean）$\bar{x}$ を次式で定義する。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

また，各データ x_i の値から平均を引いた値 $x_i - \bar{x}$ を，x_i の偏差（deviation）という。偏差の平方の平均を分散（variance）といい，S^2 と表す。また，分散の正の平方根を標準偏差（standard deviation）といい，S と表す。

$$\text{分散}\quad S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\},\ \text{標準偏差}\ S = \sqrt{S^2}$$

演習 6.1. 次のデータ・セットの平均・分散・標準偏差を求めなさい。

データ・セット

4,8,1,2,9,5,4,7,6,3

解答

データ値 x_i	偏差 $x_i - \bar{x}$	偏差の平方 $(x_i - \bar{x})^2$
計		

和の記号 $\sum$

和 $a_1 + a_2 + a_3 + \cdots + a_n$ を，記号 $\sum$ を用いて $\displaystyle\sum_{i=1}^{n} a_i$ と書く。この記号を用いたとき，平均，分散，標準偏差は次のように表せる。（和の記号 $\sum$ については数学 B 教科書「数列」の章を参照）

$$\text{平均}\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \text{分散}\ S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ \text{標準偏差}\ S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

6.2 度数分布からの平均・分散・標準偏差の計算

度数分布からの平均値の計算

度数分布表からは，もとのデータの個々の値を知ることはできない。そこで，たとえば階級「80以上90未満」のデータが7個あるとき，階級を代表する値（**階級値**）を一つ定め（たとえば80と90の平均値85），「階級値85のデータが7個」と考えて，この7個のデータの合計を 85×7 とする。このようにして平均値の近似値を求め，それを度数分布表から求めた平均値 $\bar{x}$ として用いる。

演習 6.2. 次の度数分布表をもとに，平均・分散・標準偏差を求めなさい。なお，階級値としては，各階級の下端と上端の平均値を用いなさい。

度数分布表

階級	度数 f_i	階級値 x_i	$f_i x_i$	偏差 $x_i - \bar{x}$	$f_i(x_i - \bar{x})^2$
0以上10未満	0				
10以上20未満	0				
20以上30未満	0				
30以上40未満	11				
40以上50未満	44				
50以上60未満	29				
60以上70未満	11				
70以上80未満	4				
80以上90未満	1				
90以上	0				
合計	100				

7　分布と平均値・分散・標準偏差

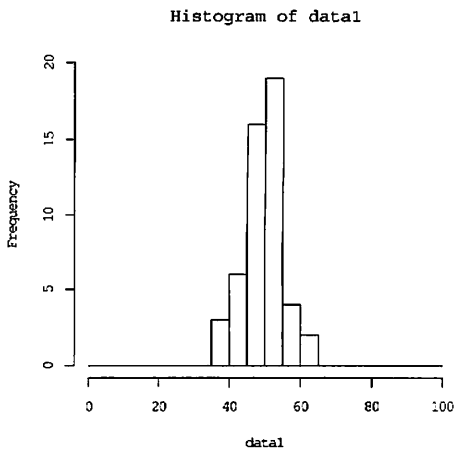
演習 7.1. いくつかのデータセットに対するヒストグラム，平均値，分散，標準偏差を示す。

問 1 ヒストグラムの図中に直線 $x = \bar{x}$ を書き込め。ただし，x はデータの値を表す変数で横軸の値，$\bar{x}$ は平均値を表す。(たとえば平均が 62 ならば，横軸 x 軸として直線 $x = 62$ を書き入れる。)

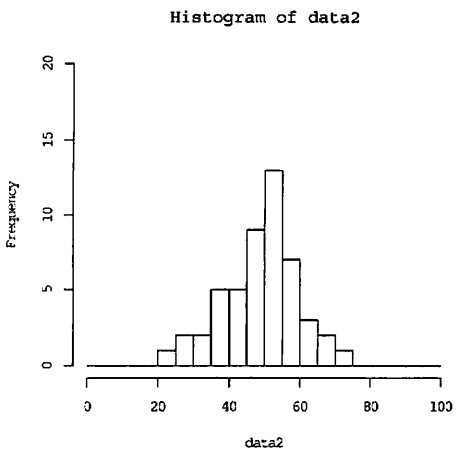
問 2 分布の形と平均値の位置の間にどのような関係があるかを観察せよ。

問 3 (1) から (4) について，標準偏差の違いとヒストグラムの形の違いの間にどのような関係があるかを観察せよ。

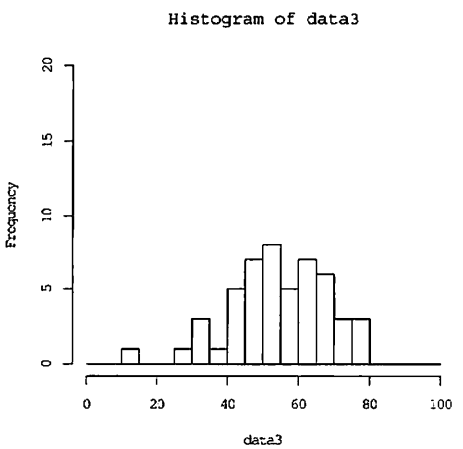
(1) 平均 50.1，分散 27.9，標準偏差 5.3



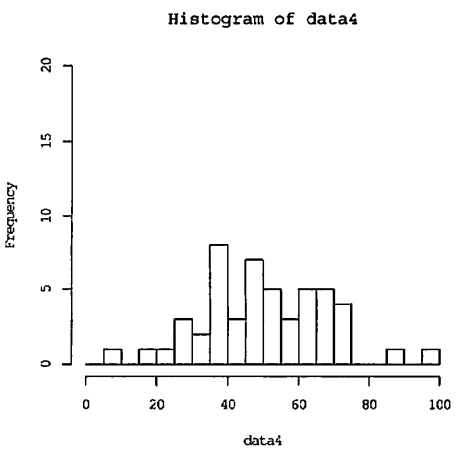
(2) 平均 49.4，分散 105.5，標準偏差 10.3



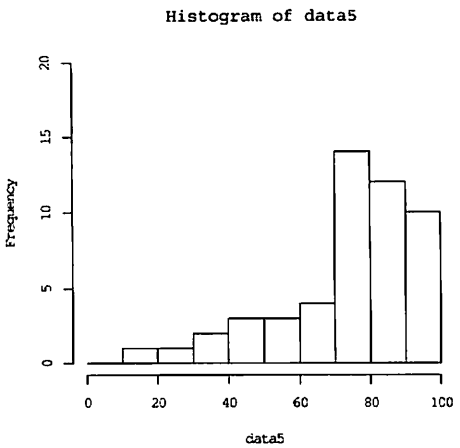
(3) 平均 54.2，分散 194.1，標準偏差 13.9



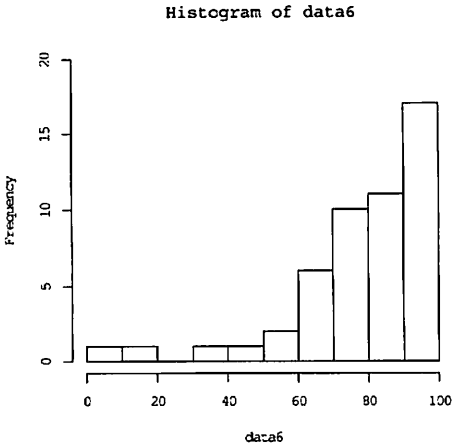
(4) 平均 50.5，分散 319.9，標準偏差 17.9



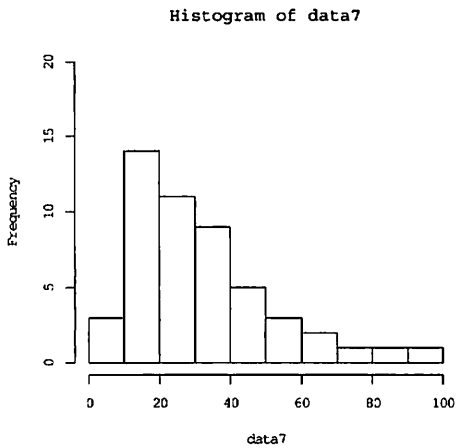
(5) 平均 74.2，分散 361.6，標準偏差 19.0



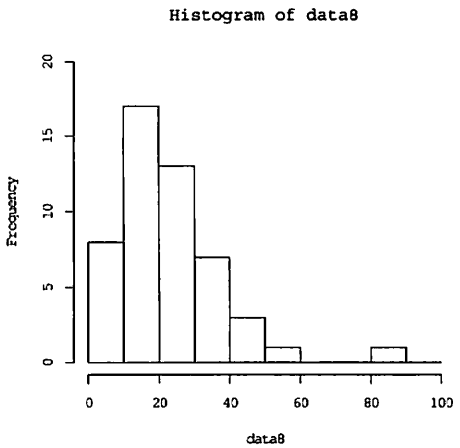
(6) 平均 78.4，分散 416.9，標準偏差 20.4



(7) 平均 31.4，分散 408.9，標準偏差 20.2



(8) 平均 22.9，分散 219.9，標準偏差 14.8



参考 このテキスト作成にあたり，人工的なデータ・セットの生成，ヒストグラムの作成，後に出てくる箱ひげ図の作成その他のために統計計算とグラフィックスのための言語・環境である「R」を用いている。R は無料で入手，利用できる。詳しく知りたい人は以下の Wiki などを出発点として調べればよい。

RjpWiki　http://www.okada.jp.org/RWiki/

8　中央値，四分位数および箱ひげ図

データを代表する値としてよく用いられるものに，平均のほかに，中央値（median）がある。

定義 データを大きさの順に並べたとき，中央にくるデータの値を**中央値**（メジアン,median）という。データの個数が奇数個なら中央にくる1個のデータの値を用い，データの個数が偶数個なら中央にくる2個のデータの値の平均値を用いる。

演習 8.1. 次のデータ・セットの中央値を求めよ。

(1) 3,5,1,10,7,8,9,3,5　　(2) 9,2,5,7,3,8,2,3,1,4,6,8

演習 8.2. 演習 7.1 に示した人工的なデータ・セットに対する中央値はそれぞれ次のようになる。前節のヒストグラムの中に，直線 $x =$ 中央値 を書きいれよ。なお，すでに書き入れてある直線 $x =$ 平均 と区別がつくように直線のそばに「平均」「中央値」と書き添えておくこと。

(1) 50.1　(2) 50.2　(3) 54.1　(4) 48.3　(5) 78.8　(6) 84.3　(7) 26.3　(8) 20.3

以上の作業の後，分布に対して平均の位置と中央値の位置の現れ方の違いについて観察せよ。

四分位数

データの代表値として平均を用いた場合には，データの散らばりを表す数値として，平均からの偏差をもとに分散・標準偏差を用いた。

データの代表値として中央値を用いた場合には，データの散らばりを表す数値として，データを値の小さいほうから4等分し，下から4分の1のところにある値（第1四分位数）および下から4分の3のところにある数（第3四分位数）を中央値とともにもちいることが多い。正確には次のように定義される。

定義 データ・セットの中央値を M とする。

(1) 第1四分位数 Q_1 を，M より小さいデータからなるデータ・セットの中央値と定義する。
(2) 第3四分位数 Q_3 を，M より大きいデータからなるデータ・セットの中央値と定義する。

なお,「M より小さい」「M より大きい」といったとき，M 自身は含まない。また，第2四分位数は中央値である。

演習 8.3. 次のデータ・セットに対し，最小値 $Min.$，第 1 四分位数 Q_1，中央値 $Med.$，第 3 四分位数 Q_3，最大値 $Max.$ を求めよ。

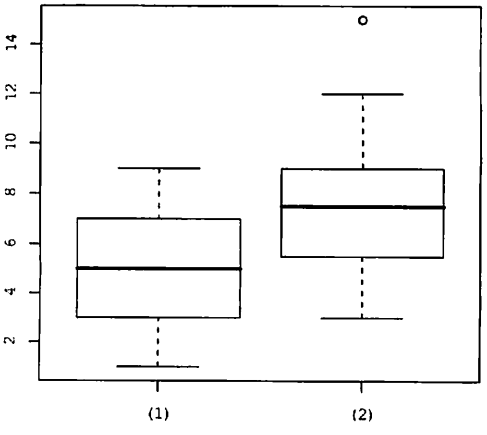
(1) 7,4,2,9,7,8,3,5,7,1,3

(2) 3,7,8,4,12,9,7,8,10,3,6,8,7,5,6,9,11,5,8,15

5 数要約と箱ひげ図

データがどのあたりに，どれくらいの広がりをもって分布しているかを上の 5 つの数値によって要約して表し，それを視覚的に捉えやすく表現したものとして，**箱ひげ図** (boxplot) がある。

例えば，上の演習問題の二種類のデータ・セットの箱ひげ図を並べると右図のようになる。

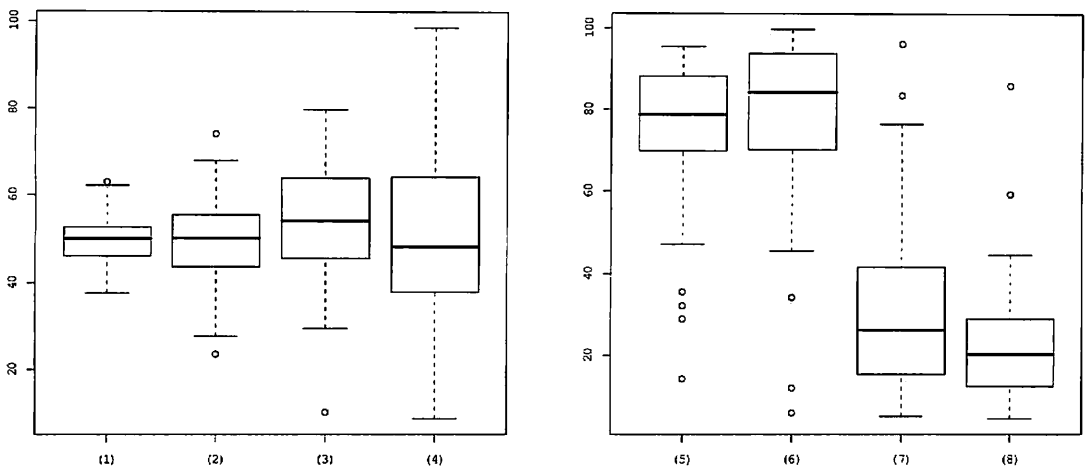


ここではデータの値を縦軸にとり，箱の下端は第 1 四分位数，上端は第 3 四分位数を表している。箱の途中に引かれた太線の位置が中央値にあたる。(1) ではひげの下端が最小値，ひげの上端が最大値である。ただし，ひげの長さは箱の長さ $Q_3 - Q_1$ の 1.5 倍以下とし，それを超えて極端に大きいデータや小さいデータは外れ値 (outlier) として個々のデータを小さな○印で表している。(2) には大きいほうに外れ値が一つあり，このときひげの上端は外れ値を除いた中での最大値を表している。

演習 8.4. 演習 8.3 の (2) のデータについて，最大値と第 3 四分位数との差が $Q_3 - Q_1$ の 1.5 倍よりも大きいことを確かめよ。

9　分布の形をみる　ヒストグラムと箱ひげ図の比較

No.4 の演習問題にあげた 8 つの分布について，それぞれを箱ひげ図で表したものを以下に示す。No.4 とこのプリントを並べてヒストグラム，平均値，標準偏差の値と，箱ひげ図，中央値，四分位数などを比較観察してみよう。



例えば，次のようなことが観察できる。

(1) 標準偏差が大きいほどデータが広がって分布していることが読み取れる。

(2) ヒストグラムが左右対称に近い分布の箱ひげ図では，ひげの長さが上下ともほぼ同じ長さである。ヒストグラムが一方の裾が長い分布の箱ひげ図では，対応する側のひげが他方より長い。

(3) (5) (6) (7) (8) のデータセットでは端の方に分布の峰（ピーク）がありデータが多く集まっている。このことは，箱ひげ図よりもヒストグラムからの方がより容易に読み取ることができる。

ヒストグラムにおいて山の頂上のようにデータの分布の峰（ピーク）になっているところでのデータの値を最頻値（モード,mode）という。度数分布表では，度数最大の階級の階級値が最頻値である。ただし，峰が二つできるような分布に対しては最頻値は適切な代表値とはいえない。

演習 9.1. No.2 の例 5.1 における度数分布表についてモードを求めよ。

演習 9.2. No.4 の演習 7.1 のデータ (6)(7) について，平均（mean），メジアン（median），モード（mode）を小さいほうから順に並べよ。（たとえば，$mean < median < mode$ のように答えよ。）

10 補足問題

問題 10.1. 次の式を，$\sum$ 記号を使わない形に書き直せ。ただし，計算はしなくてもよい。

(1) $\displaystyle\sum_{k=1}^{5} a_k$　(2) $\displaystyle\sum_{k=1}^{5} a_{2k-1}$　(3) $\displaystyle\sum_{k=1}^{7}(3k+1)$　(4) $\displaystyle\sum_{k=3}^{7} k^2$　(5) $\displaystyle\sum_{k=1}^{10} 3$

問題 10.2. （データの定数倍と平均）n 個のデータ $x_1, x_2, \cdots, x_n$ からなると定数 a に対し，$y_i = ax_i$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の平均値 $\bar{y}$ と $x_1, x_2, \cdots, x_n$ の平均値 $\bar{x}$ の間には $\bar{y} = a\bar{x}$ という関係が成り立つことを示せ。

問題 10.3. （データの平行移動と平均）データ・セット $x_1, x_2, \cdots, x_n$ と定数 b に対し，$y_i = x_i + b$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の平均値 $\bar{y}$ と $x_1, x_2, \cdots, x_n$ の平均値 $\bar{x}$ の間には $\bar{y} = \bar{x} + b$ という関係が成り立つことを示せ。

問題 10.4. （データの変換と平均）データ・セット $x_1, x_2, \cdots, x_n$ と定数 a，b に対し，$y_i = ax_i + b$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の平均値 $\bar{y}$ を $x_1, x_2, \cdots, x_n$ の平均値 $\bar{x}$ および a，b を用いて表せ。

問題 10.5. （データの定数倍と標準偏差）データ・セット $x_1, x_2, \cdots, x_n$ と定数 a に対し，$y_i = ax_i$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の標準偏差 S_y と $x_1, x_2, \cdots, x_n$ の標準偏差 S_x の間には $S_y = |a|S_x$ という関係が成り立つことを示せ。

問題 10.6. （データの平行移動と標準偏差）データ・セット $x_1, x_2, \cdots, x_n$ と定数 b に対し，$y_i = x_i + b$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の標準偏差 S_y と $x_1, x_2, \cdots, x_n$ の標準偏差 S_x の間には $S_y = S_x$ という関係が成り立つことを示せ。

問題 10.7. （データの変換と標準偏差）n 個のデータ $x_1, x_2, \cdots, x_n$ からなるデータ・セットと定数 a，b に対し，$y_i = ax_i + b$ によって新しいデータ・セット $y_1, y_2, \cdots, y_n$ を定義する。このとき，$y_1, y_2, \cdots, y_n$ の標準偏差 S_y を $x_1, x_2, \cdots, x_n$ の標準偏差 S_x および a，b を用いて表せ。

問題 10.8. （標準得点および偏差値）n 個のデータ $x_1, x_2, \cdots, x_n$ からなるデータ・セットの平均を $\bar{x}$，標準偏差を S_x とする。変換 $y_i = x_i + b$ によって作ったデータ・セット $y_1, y_2, \cdots, y_n$ の平均を 0 にし，さらに変換 $z_i = y_i/c$ によって作ったデータ・セット $z_1, z_2, \cdots, z_n$ の標準偏差を 1 にしたい。このとき，定数 b，c をどのように選べばよいか。また，このとき z_i を x_i の式で表せ。このようにして作った $z_1, z_2, \cdots, z_n$ を $x_1, x_2, \cdots, x_n$ の標準化 (standardization)，あるいは**標準得点** (standard score) または **Z 得点**という。さらに，$T_i = 10z_i + 50$ という変換によって平均 50 点，標準偏差 10 点となるようにしたとき，T_i を**偏差値**という。

問題 10.9. （分散の計算式）n 個のデータ $x_1, x_2, \cdots, x_n$ からなるデータ・セットの分散を S^2 とするとき，次の等式が成り立つことを示せ。(左の等号は定義，右の等号成立が問題)

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

問題 10.10. $x_1, x_2, \cdots, x_n$ に対し，変数 t の 2 次関数 $f(t) = \dfrac{1}{n}\left\{(x_1 - t)^2 + (x_2 - t)^2 + \cdots + (x_n - t)^2\right\}$ は $t = \bar{x}$ （平均）において最小値 S^2 （分散）をとることを示せ。

和の記号 $\sum$

次の等式の左辺の記号の意味を，右辺で表される和として定義する。

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

例 次に，$\sum$ 記号を用いて表した式を，$\sum$ 記号を用いずに書き換えたらどうなるかを示す。なお，記号の説明のために，あえて足し算を計算せずにそのままにしている。

(1) $\displaystyle\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$

$\sum$ の下に書いてある i の値が始まり，上に書いてある i の値で終わり。

(2) $\displaystyle\sum_{k=1}^{5} (2k+1) = 3 + 5 + 7 + 9 + 11$

$\sum$ の後に書いてある式 $2k+1$ に，$k=1$ を代入した値，$k=2$ を代入した値，と順に足して行き，$k=5$ を代入した値を足したところで終わる。

(3) $\displaystyle\sum_{l=3}^{7} (3l-1) = 8 + 11 + 14 + 17 + 20$

$\sum$ の下に書いてある l の値は 3 だから，$l=3$ を代入した値から始まる。

(4) $\displaystyle\sum_{i=1}^{5} 7 = 7 + 7 + 7 + 7 + 7$

$i=1$ のときの値 7，$i=2$ のときの値 7，$i=3$ のときの値 7，$i=4$ のときの値 7，$i=5$ のときの値 7 を足す。結局，7 を 5 個足した。

$\sum$ の性質

(1) $\displaystyle\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

(2) $\displaystyle\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$　　(c は定数)

(3) $\displaystyle\sum_{i=1}^{n} c = nc$　　(c は定数)

問題 これらの等式が成り立つことを確かめてみること。

※ $\sum$ 記号は，数学 B で学習する「数列」の章で頻繁に利用する。

平均値・分散と2次関数の最小値

n 個のデータ $x_1, x_2, \cdots, x_n$ の平均値を t についてのある2次関数を最小にする t の値として，また，分散をその2次関数の最小値として特徴付けることができる。

問題10.10. $x_1, x_2, \cdots, x_n$ に対し，変数 t の2次関数 $f(t) = \frac{1}{n}\{(x_1-t)^2+(x_2-t)^2+\cdots+(x_n-t)^2\}$ は $t = \bar{x}$（平均）において最小値 S^2（分散）をとることを示せ。

統計入門補足3　　年　組　番 名前

偏差値を求める

試験の点数の偏差値は，次の3段階の手順により求められる。(問題10.8の解答)

(1) 平均点と標準偏差を求める。(←第1回プリントの内容)

(2) 各点数から平均点を引き，標準偏差で割る。これを**標準得点**という。

標準得点の平均は0点，標準偏差は1になる。(←問題10.2〜10.7によりわかる)

(3) 標準得点を10倍して50点を足す。この値を**偏差値**という。

偏差値の平均は50，標準偏差は10となる。(←問題10.2〜10.7によりわかる)

問題1 5人が10点満点のテストを受けた結果，次の点数を得た。このとき，平均，標準偏差，それぞれの標準得点，偏差値を求めよ。

番号 i	(1)	(2)	(3)	(4)	(5)
点数 x_i	3	7	2	3	5

平均値

標準偏差　下の表を利用し，標準偏差を求めよ。

標準得点と偏差値の計算　それぞれの標準得点と偏差値を求め，表を完成せよ。

番号	点数	偏差	$(\text{偏差})^2$	標準得点	偏差値
(1)	3				
(2)	7				
(3)	2				
(4)	3				
(5)	5				
計					

問題 1　ある試験では，平均点が 62 点，標準偏差が 12.5 点であった。この試験で，A 君の得た点数は 75 点である。このとき，A 君の偏差値はいくらか。

問題 2 いろいろ試してみよ。

＿＿＿＿＿＿＿＿ では，平均点が ＿＿＿点，標準偏差が ＿＿＿点 であった。

この試験で ＿＿＿点 を得たとき，偏差値は ＿＿＿＿ である。

＿＿＿＿＿＿＿＿ では，平均点が ＿＿＿点，標準偏差が ＿＿＿点 であった。

この試験で ＿＿＿点 を得たとき，偏差値は ＿＿＿＿ である。

＿＿＿＿＿＿＿＿ では，平均点が ＿＿＿点，標準偏差が ＿＿＿点 であった。

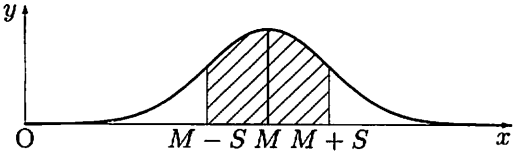
この試験で ＿＿＿点 を得たとき，偏差値は ＿＿＿＿ である。

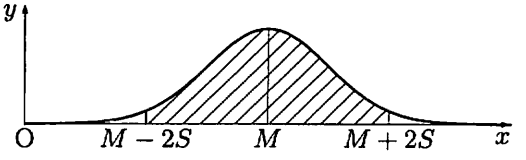
偏差値からわかること

試験の点数をもとに偏差値を算出したとき，偏差値の平均は 50，標準偏差は 10 になっている。したがって，偏差値が 60 であるということは，その試験において，平均点よりも標準偏差 1 つぶんだけ高い点数だったということを表している。また，偏差値 70 であるということは，その試験において，平均点よりも標準偏差の 2 倍分だけ高い点数だったということを表している。

得点分布はさまざまな形をとりうるが，もしも得点分布が**正規分布**と呼ばれるきれいな形の分布になっているならば，**（平均点 ± 標準偏差）の幅の中に全データの 68.27% が含まれる。**また，**（平均点 ± 標準偏差の 2 倍）の幅の中に全データの 95.45% が含まれる。**

下図は正規分布を表す曲線である。M を平均，S を標準偏差とすると，下図左の斜線部は全体の 68.27% を占め，下図右の斜線部は全体の 95.45% を占める。





偏差値の場合には，平均 50，標準偏差 10 になるように標準化されているので，**もしその試験の得点分布が正規分布に従うならば，**「偏差値 70 以上」であることは，全受験生の中で上位から約 2.5 パーセント以内にいることを表している。

11　2次元のデータ

前節までは，例えば，40人のクラスのそれぞれの身長を測って得た40個の数値の集まり，というように，1人の対象に対して1つの数値が対応しているような場合について扱った。今回は，40人のクラスのそれぞれの身長と体重を測って身長と体重の間の関係を調べるというように，2種類の数値が組みになっている場合を扱う。

上の例では，1人の生徒に対して2つの値（身長，体重）が対応している。このように，2つの数値が組になっているデータを**2次元データ**という。もっと多くの属性について（例えば，座高，握力，垂直とび，etc.）を調べれば，3次元データ，4次元データ，… などを考えることもある。そのような場合でも，そのうちの2つの項目を取り出してその関係を見ることが基本になるので，ここでは2次元データについて2種類のデータの間の関係を調べることを考える。

12　相関図（散布図）

次のデータは教科書から取った。（数研出版　改訂版数学B，p.122）

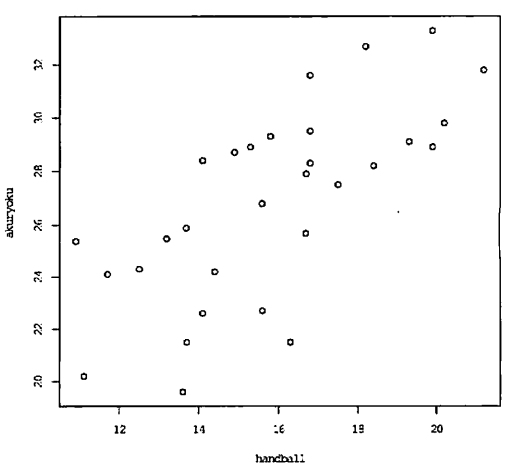
例 12.1. 生徒30人のハンドボール投げの記録と，握力の記録

番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
距離	19.3	15.6	18.2	13.2	16.7	14.9	18.4	14.4	17.5	11.7	16.8	14.1	16.3	19.9	16.8
握力	29.1	26.8	32.7	25.5	27.9	28.7	28.2	24.2	27.5	24.1	28.3	28.4	21.5	33.3	31.6

番号	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
距離	13.6	10.9	16.8	13.7	21.2	15.8	14.1	15.6	13.7	11.1	15.3	16.7	12.5	19.9	20.2
握力	19.6	25.4	29.5	25.9	31.8	29.3	22.6	22.7	21.5	20.2	28.9	25.7	24.3	28.9	29.8

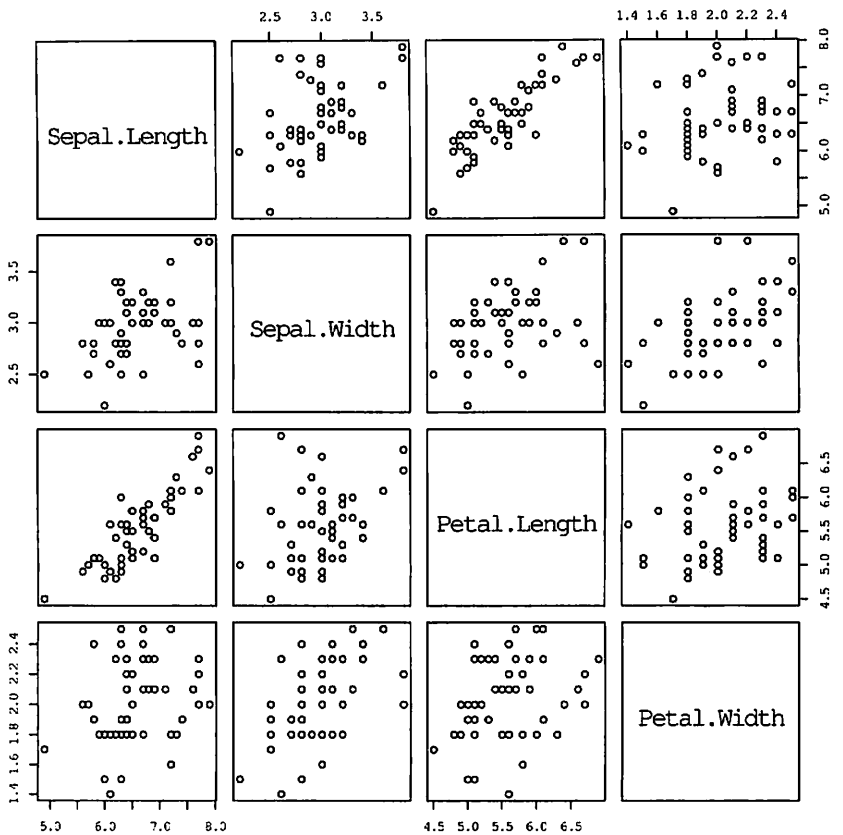
2行目がハンドボール投げの記録で単位は m，3行目が握力の記録で単位は kg である。

1次元データの場合にヒストグラムを書き，分布の様子を見ることが大切であったように，2次元データの場合にも，視覚的に捉えやすく加工し，目で見ることが大切である。特に，2つの変数間の関係を調べたい場合には，横軸と縦軸にそれぞれの項目をとり，1つの2次元データを1つの点として書き込んで，それら全体の散らばりかたをみることが有効である。右図は，横軸にハンドボール投げの距離をとり，縦軸に握力をとって図示したものである。このような図を**相関図**または**散布図**という。この例の場合，右上がりに点が分布している様子が見て取れる。これは，「ハンドボールの飛距離が大きいほど握力も強い」というおおまかな関係がある，ということを表している。



例 12.2. 次は，以前紹介した「R」についてくる「iris[1]」というデータ・セットをもとに作った。このデータはあやめ（アイリス，iris）について，その萼片（がくへん，sepal）の長さと幅，花弁（かべん，petal）の長さと幅，そして品種の5項目を150の個体について調べた5次元データからなる。

このデータからverginica一種だけを抜き出し，萼片の長さ（Sepal.Length），萼片の幅（Sepal.Width），花弁の長さ（Petal.Length），花弁の幅（Petal.Width）の4つの変数について，その中の2つずつの相関図（散布図）をすべて書き出したものが次図である。



たとえば，左端の列の上から3段目（1列3行）の相関図では，萼片の長さ（Sepal.Length）が横軸に，花弁の長さ（Petal.Length）が縦軸にとってある。この両者には，かなり強い直線的な関係がありそうだ。一方，1列目の3行目にある萼片の長さ（Sepal.Length）と花弁の幅（Petal.Width）の相関図では，あまり強い直線的な関係はなさそうである。

一方の変量が増加すると他方も増加する傾向があるとき**正の相関関係**があるといい，一方が増加すると他方が減少する傾向があとき**負の相関関係**があるという。どちらの傾向も認められないとき**相関関係がない**という。

[1]統計的推測の理論を確立したフィッシャーが次の論文の中で用いたデータ。統計の入門書ではしばしば例として取り上げられている。Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179-188. （taxonomy　分類学）

13　相関係数

目で見て傾向をつかんだら，次に考えることは，相関の正負や程度を数値で表すことである。天下りだが，次のように定義する。

定義 13.1. n 個の 2 次元データからなるデータ・セットが与えられたとする。

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

ここから第一成分だけを取り出したデータ・セットを x で表し，その平均を $\bar{x}$ とする。また第 2 成分だけを取り出したデータ・セットを y で表し，その平均を $\bar{y}$ とする。このとき，x と y の共分散 S_{xy} を次式で定義する。

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

上では，x_1 と y_1 とが対応しているというように，対応関係が分かるように座標のような書き方をした。右のように，表の形でデータを示すことも多い。このとき，1 行目の平均が $\bar{x}$，2 行目の平均が $\bar{y}$ である。

変量 1	x_1	x_2	$\cdots$	x_n
変量 2	y_1	y_2	$\cdots$	y_n

定義式が読み取れたかどうか，次の問題で確認する。

問題 13.1. 次の 2 次元データ (x, y) のセットについて，相関図を書き，x と y の共分散を求めよ。

(1)

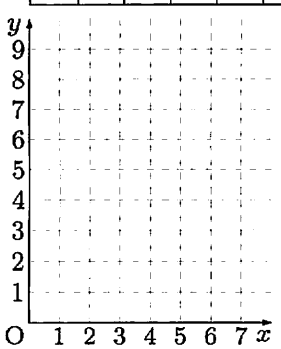
x	1	2	4	4	5
y	2	4	8	7	9

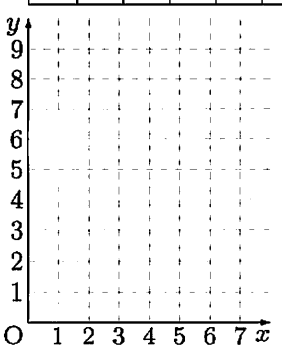
(2)

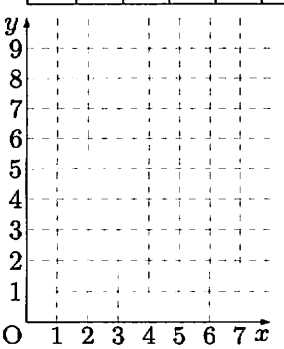
x	1	2	2	4	5
y	8	4	5	2	1

(3)

x	1	2	4	5	5
y	2	8	4	1	8







定義 13.2. n 個の 2 次元データからなるデータ・セットが与えられたとする。

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

ここから第一成分だけを取り出したデータ・セットを x で表し，その平均を $\bar{x}$ とする。また第 2 成分だけを取り出したデータ・セットを y で表し，その平均を $\bar{y}$ とする。このとき，x と y の相関係数 r_{xy} を次式で定義する。

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

ここで，S_{xy} は x と y の共分散，S_x は x の標準偏差，S_y は y の標準偏差である。より具体的に書くと，

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

問題 13.2. 問題 13.1 の各データ・セットについて，相関係数を求めよ。

(1)　　(2)　　(3)

相関係数（続き）

相関係数には次の性質がある。

(1) 相関係数 r_{xy} のとりうる値の範囲は，$-1 \leqq r_{xy} \leqq 1$ である。
(2) r_{xy} の値が 1 に近いほど，正の相関が強くなる。このとき，相関図の点は右上がりに分布する。
(3) r_{xy} の値が -1 に近いほど，負の相関が強くなる。このとき，相関図の点は右下がりに分布する。
(4) r_{xy} の値が 0 に近いほど，相関は弱くなる。

(1) の証明は問題にまわして，(2) (3) (4) について観察しよう。

例 13.3. 例 12.2 における散布図行列のうち，いくつかについて相関係数を計算すると次のようになる。プリント No.1 の例 12.2 を横に置き，見比べてほしい。

(1) Sepal.Length と Sepal.Width の相関係数 0.4572278
(2) Sepal.Length と Patal.Length の相関係数 0.864227
(3) Sepal.Length と Patal.Width の相関係数 0.2811077
(4) Sepal.Width と Petal.Length の相関係数 0.4010446
(5) Sepal.Width と Petal.Width の相関係数 0.537728
(6) Petal.Length と Petal.Width の相関係数 0.3221082

これによると最も相関が強いのが Sepal.Length と Patal.Length であり，最も相関が弱いのが Sepal.Length と Patal.Width である。相関図から読み取れる感覚的な把握と一致しているだろうか。

問題 13.3. 相関係数や，その定義に使われている共分散をプリント No.2 のように定めたとき，上のような性質を持つのはなぜか。その理由を考えよ。

14 補足問題

問題 14.1. 2次元のデータ・セット (x_i, y_i), $i = 1, 2, \cdots, n$, に対して相関係数 r_{xy} は $-1 \leqq r_{xy} \leqq 1$ を満たすことを $n = 3$ の場合に証明せよ。

15　これまでの振り返り

これまで4回の統計入門講座で，およそ次のようなことを取り上げました。

(1) 1次元のデータセットについて，データの特徴を記述する

(2) 2次元のデータセットについて，2つの変量の間の直線的な関係の強さを記述する

1次元のデータセットの特徴をつかむには，**度数分布**から**ヒストグラム**を作成して，分布の形を見ることが重要でした。特に，データがどのあたりにどの程度広がっているか，という特徴に注目して，それを数値で表すために，**平均値**と**標準偏差**，あるいは，**中央値**と**四分位点**，という2つの方法を取り上げました。中央値と四分位点を図示するものとして，**箱ひげ図**も扱いました。

また，各データから平均値を引いて，標準偏差で割るという操作を加えることにより，データセットの平均を0，標準偏差を1に変換する**標準化**という方法について学びました。こうして標準化によって得た**標準得点**を10倍して50を足すことにより，平均値50，標準偏差10となるようにし，普段テストの点数として見慣れている100点満点の数値に近い印象を与えるように作り変えたものが**偏差値**でした。

2次元のデータセット（各個体について2つの変量の値が組になっているようなデータセット）について，2つの変量の関係を把握するには，まず**相関図**（散布図）をみることが重要でした。一方が大きいほど他方も大きい（一方が小さいほど他方も小さい）というような関係があるときに，**正の相関関係がある**（負の相関関係がある）といい，このような直線的な相関関係の強さを表すものとして**相関係数**を定義しました。相関係数は-1以上1以下の値をとり，1に近いほど正の相関関係が強いことを示し，-1に近いほど負の相関関係が強いことを示しています。

16　これからの見通し

ここまでは，調査や実験などで手に入れたデータの特徴を記述する段階でした。これからは，手に入れた一部のデータ（標本）から，全体のデータ（母集団）について推定したり，立てた仮説が正しいと判断できるかどうかを検定することを目標とします。

例えば，選挙の際にテレビ局が行う出口調査などは，一部の有権者の投票結果（標本）から母集団についての投票結果を知ろうという試みです。

この目的のためには，確率の考え方が欠かせません。そこで，ここからしばらくは確率変数や確率分布など，統計の中で必要となる確率の考えについて学びます。

17　確率変数と確率分布

さいころを3回投げて，1の目が出た回数をXで表すとする。このとき，Xは0,1,2,3のいずれかの値をとる変数である。Xのとりうる値のおのおのに対して，その値をとる確率が定まる。これ

は，下で定義する確率変数の例となっている。

問題 17.1. さいころを 3 回投げて，1 の目が出た回数を X とし，X が値 n をとる確率を $P(X=n)$ と書く。このとき，$P(X=0)$，$P(X=1)$，$P(X=2)$，$P(X=3)$ をそれぞれもとめよ。

定義 17.1. 変数 X がとりうる値のおのおのに対してその値をとる確率が定まるとき，この変数を**確率変数**という。X のとりうる値 x_k に対して，$X=x_k$ となる確率 $p_k=P(X=x_k)$ を対応させるとき，この対応関係を**確率分布**（または単に**分布**）といい，確率変数 X はこの分布に**従う**という。

確率分布は，次のように表の形で表すこともある。

X	x_1	x_2	$\cdots$	x_n
P	p_1	p_2	$\cdots$	p_n

問題 17.2. 先の問題 17.1 の確率変数 X が従う確率分布を表の形で表せ。

問題 17.3. 袋の中に赤玉が 1 個，白玉が 2 個入っている。この中から無作為に玉を 1 個とりだし，色を調べてもとに戻す。これを 5 回繰り返し，赤玉が出た回数を X とする。この X は確率変数である。確率変数 X が従う確率分布を求め，表の形に書け。

18　確率変数の期待値（平均値）

期待値については，授業ですでに学んだ。

定義 18.1. 確率変数 X が，次の表で与えられる確率分布に従うとする。

X	x_1	x_2	$\cdots$	x_n
P	p_1	p_2	$\cdots$	p_n

このとき，X の期待値（平均値）$E(X)$ を次式で定義する。

$$E(X) = \sum_{k=1}^{n} x_k p_k = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$$

問題 18.1. 問題 17.3 で求めた確率分布に従う確率変数 X について，期待値 $E(X)$ を求めよ。

問題 18.2. 大小 2 個のさいころを投げたとき，大きいさいころの目を X，小さいさいころの目を Y とすると，それぞれは確率変数である。また，2 個のさいころの目の和 $Z = X + Y$ も確率変数である。

(1) 確率変数 X，Y の期待値を求めよ。

(2) 確率変数 $Z = X + Y$ の期待値を求めよ。

結果として，期待値について次の等式が成り立っている。

$$E(X + Y) = E(X) + E(Y)$$

これは，確率変数 X，Y に対して一般に成り立つ等式である。

19 確率変数の分散，標準偏差

確率変数 X の期待値を $E(X) = \mu$ とおく。確率変数 $(X-\mu)^2$ の期待値 $E((X-\mu)^2)$ を，X の分散といい，$V(X)$ と表す。

$$V(X) = E((X-\mu)^2) = \sum_{k=1}^{n} p_k(x_k - \mu)^2$$

また，分散の正の平方根を標準偏差といい，$\sigma(X)$ と表す。$\sigma(X) = \sqrt{V(X)}$

確率変数の分散や標準偏差の定義は，データセットの分散（偏差の平方の平均）や標準偏差の定義と対応している。標準偏差が大きいほど，確率変数 X の値の散らばりの度合いが大きく，標準偏差が小さいほど，確率変数 X の値が期待値の近くに集まり散らばりの度合いが小さい。

問題 19.1. 1 個のさいころを投げたときに出る目の数を X とするとき，確率変数 X の分散と標準偏差を求めよ。

問題 19.2. 袋の中に赤玉 3 個と白玉 2 個が入っている。この中から 3 個の玉を同時に取り出すとき，赤玉の個数を X とする。確率変数 X の期待値，分散，標準偏差を求めよ。

定理 19.1. 確率変数 X に対して，次の等式が成り立つ。$V(X) = E(X^2) - \{E(X)\}^2$

証明

問題 19.3. 問題 19.2 における分散を，上の定理を用いて計算し，問題 19.2 で求めた結果と同じになることを確かめよ。

20　二項分布

問題 3.1 では，さいころを 3 回投げて，1 の目が出た回数 X を考えた。その結果，確率変数 X の確率分布表は次のようになった。

X	0	1	2	3	計
P	${}_3\mathrm{C}_0\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3$	${}_3\mathrm{C}_1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2$	${}_3\mathrm{C}_2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1$	${}_3\mathrm{C}_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0$	1

一般に，1 回の試行で事象 A の起こる確率が p であるとき，この試行を n 回行う反復試行において事象 A が起こる回数を X とするとき，確率変数 X の従う確率分布を**二項分布**といい，$B(n,p)$ で表す。たとえば上の例は，二項分布 $B\left(3,\frac{1}{6}\right)$ である。

問題 20.1. 確率分布 $B(n,p)$ に従う確率変数 X について，その確率分布表を作れ。$q=1-p$ とせよ。

確率分布 $B(n,p)$ に従う確率変数 X の期待値，分散，標準偏差を，単に確率分布 $B(n,p)$ の平均，分散，標準偏差という。

二項分布の平均，分散，標準偏差について，次の事実が証明できる。

定理 20.1. 確率変数 X が二項分布 $B(n,p)$ に従うとき，$q=1-p$ とすると，

$$E(X)=np,\quad V(X)=npq,\quad \sigma(X)=\sqrt{npq}$$

証明は省略し，ここでは使うことに重点を置く。

問題 20.2. 二項分布 $B(1,p)$ に従う確率変数について，上の定理が成り立つことを示せ。

問題 20.3. 二項分布 $B\left(3, \frac{1}{6}\right)$ に従う確率変数 X の平均 $E(X)$，分散 $V(X)$，標準偏差 $\sigma(X)$ を求めよ。

問題 20.4. 次の二項分布の平均，分散と標準偏差を求めよ。

(1) $B\left(12, \frac{1}{4}\right)$　　(2) $B\left(9, \frac{1}{2}\right)$　　(3) $B\left(8, \frac{2}{3}\right)$

問題 20.5. 1 個のさいころを 10 回投げるとき，1 の目が出る回数を X とする。X の期待値，標準偏差を求めよ。

問題 20.6. 1 枚の硬貨を 10 回投げるとき，表が出る回数を X とする。X の期待値，標準偏差を求めよ。

問題 20.7. 不良品率 3% の製品の山から 50 個取り出したときの不良品の個数 X は二項分布に従うという。このとき，X の期待値，標準偏差を求めよ。

問題 20.8. 白玉 3 個と黒玉 2 個が入っている袋から玉を 1 個取り出し，もとに戻す操作を 100 回行う。白玉の出る回数の期待値と標準偏差を求めよ。

統計入門用語集 No.1

年　組　番 名前

データ

実験，観察，調査などによって収集された情報を数値その他で具体的に表現したもの。この入門講座では数値データを中心に扱う。一組のデータの集まりをデータセットという。

平均

データセット $x_1, x_2, \cdots, x_n$ の平均 $\bar{x}$ を次式で定義する。

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

偏差

データセット $x_1, x_2, \cdots, x_n$ の平均を $\bar{x}$ とするとき，各データから平均を引いた値を偏差という。データ x_1 の偏差は $x_1 - \bar{x}$，データ x_2 の偏差は $x_2 - \bar{x}$，$\cdots$ である。一般に，データ x_i の偏差は $x_i - \bar{x}$ $(i = 1, 2, \cdots, n)$ である。

分散

データセット $x_1, x_2, \cdots, x_n$ に対し，偏差の 2 乗の平均を分散といい，S^2 で表す。すなわち，

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

標準偏差

データセット $x_1, x_2, \cdots, x_n$ に対し，分散の正の平方根を標準偏差といい S で表す。すなわち，

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

標準得点

データセット $x_1, x_2, \cdots, x_n$ に対し，各データ x_i の標準得点 z_i を次式で定義する。

$$z_i = \frac{x_i - \bar{x}}{S}$$

このとき，標準得点 $z_1, z_2, \cdots, z_n$ の平均は 0，標準偏差は 1 となる。

偏差値

データセット $x_1, x_2, \cdots, x_n$ に対し，各データ x_i の偏差値 T_i を次式で定義する。

$$T_i = 50 + 10z_i \quad (z_i \text{は標準得点})$$

このとき，偏差値 $T_1, T_2, \cdots, T_n$ の平均は 50，標準偏差は 10 となる。

2 次元のデータセット

1 クラス 40 人の生徒それぞれの身長と体重を組にしたデータ (身長，体重) のように，2 つの数値が組になったデータの集まりを 2 次元のデータセットという。同様に，3 つの数値が組になったデータの集まりを 3 次元のデータセットという。以下同様に，n 次元のデータセットを考えることもできる。

共分散

2 次元のデータセット $(x_1, y_1), \cdots, (x_n, y_n)$ に対し，共分散 S_{xy} を次式で定義する。

$$S_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

ここで，$\bar{x}$ はデータセット $x_1, x_2, \cdots, x_n$ の平均値，$\bar{y}$ はデータセット $y_1, y_2, \cdots, y_n$ の平均値である。

相関図（散布図）

2 次元のデータセット $(x_1, y_1), \cdots, (x_n, y_n)$ に対し，座標平面上に (x_k, y_k) $(k = 1, \cdots, n)$ を座標に持つ点を書き込んで（プロットして）できる図を相関図（散布図）という。

正の相関関係・負の相関関係

2 次元のデータセット $(x_1, y_1), \cdots, (x_n, y_n)$ に対し，x_k，y_k のうち一方のとる値が大きいほど他方のとる値も大きい，という傾向があるとき，x_k と y_k の間には正の相関関係があるという。また，x_k，y_k のうち一方のとる値が大きいほど他方のとる値が小さい，という傾向があるとき，x_k と y_k の間には負の相関関係があるという。正の相関関係があるとき，相関図にプロットされたデータを表す点はおおむね右上がりに分布し，負の相関関係があるとき，相関図にプロットされたデータを表す点はおおむね右下がりに分布する。

相関係数

2 次元のデータセット $(x_1, y_1), \cdots, (x_n, y_n)$ に対し，共分散 S_{xy} を，$x_1, x_2, \cdots, x_n$ の標準偏差 S_x と $y_1, y_2, \cdots, y_n$ の標準偏差 S_y の積 $S_x S_y$ で割った値を，相関係数といい r_{xy} で表す。すなわち，

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

相関係数のとりうる値の範囲は $-1 \leqq r_{xy} \leqq 1$ であり，相関係数の値が 1 に近いほど正の相関が強く，-1 に近いほど負の相関が強い。

確率変数（離散確率変数）

変数 X がとりうる値のおのおのに対してその値をとる確率が定まるとき，この変数を確率変数という。

【例】サイコロを投げて出る目の値 X は確率変数である。X のとりうる値は $1,2,3,4,5,6$ であり，そのおのおのの値をとる確率はすべて $\frac{1}{6}$ である。

【例】くじ引きを一回引いてあたる確率が p であるとする。同一の条件でこのくじを繰り返し n 回引いたとき，当たりくじを引く回数 X は確率変数である。X のとりうる値は $0,1,2,\cdots,n$ であり，値 k をとる確率 $P(X=k)$ は，${}_n\mathrm{C}_k p^k(1-p)^{n-k}$ である。

確率分布（離散確率変数の）

確率変数 X のとりうる値 x_k に対して，$X=x_k$ となる確率 $p_k=P(X=x_k)$ を対応させるとき，この対応関係を確率分布（または単に分布）といい，確率変数 X はこの分布に従うという。

確率分布は，次のように表の形（確率分布表）で表すこともある。

X	x_1	x_2	$\cdots$	x_n
P	p_1	p_2	$\cdots$	p_n

確率変数の期待値

上の確率分布表に従う確率変数 X の期待値（平均値）$E(X)$ を次式で定義する。

$$E(X)=\sum_{k=1}^{n} x_k p_k = x_1p_1+x_2p_2+\cdots+x_np_n$$

2 個の確率変数 X，Y と定数 a，b に対して，期待値は次の性質を持つ。

$$E(X+Y)=E(X)+E(Y), \qquad E(aX+b)=aE(X)+b$$

確率変数の分散

確率変数 X に対し，確率変数 $(X-\mu)^2$（ここで $\mu=E(X)$）の期待値 $E((X-\mu)^2)$ を X の分散といい，$V(X)$ と表す。

$$V(X)=E((X-\mu)^2)=\sum_{k=1}^{n} p_k(x_k-\mu)^2$$

確率変数の標準偏差

確率変数 X に対し，分散の正の平方根を標準偏差といい，$\sigma(X)$ と表す。

$$\sigma(X)=\sqrt{V(X)}$$

続・統計入門

1　はじめに

1年次の後期に土曜学習活動日を利用して「統計入門講座」を6回実施しました。以下の文章は，第4回目が終わった時点で一度振り返ったときのものですが，再掲しておきます。

最初の4回で，およそ次のようなことを取り上げました。

(1) 1次元のデータセットについて，データの特徴を記述する

(2) 2次元のデータセットについて，2つの変量の間の直線的な関係の強さを記述する

1次元のデータセットの特徴をつかむには，**度数分布**から**ヒストグラム**を作成して，分布の形を見ることが重要でした。特に，データがどのあたりにどの程度広がっているか，という特徴に注目して，それを数値で表すために，**平均値**と**標準偏差**，あるいは，**中央値**と**四分位点**，という2つの方法を取り上げました。中央値と四分位点を図示するものとして，**箱ひげ図**も扱いました。

また，各データから平均値を引いて，標準偏差で割るという操作を加えることにより，データセットの平均を0，標準偏差を1に変換する**標準化**という方法について学びました。こうして標準化によって得た**標準得点**を10倍して50を足すことにより，平均値50，標準偏差10となるようにし，普段テストの点数として見慣れている100点満点の数値に近い印象を与えるように作り変えたものが**偏差値**でした。

2次元のデータセット（各個体について2つの変量の値が組になっているようなデータセット）について，2つの変量の関係を把握するには，まず**相関図**（**散布図**）をみることが重要でした。一方が大きいほど他方も大きい（一方が小さいほど他方も小さい）というような関係があるときに，**正の相関関係がある**（**負の相関関係がある**）といい，このような直線的な相関関係の強さを表すものとして**相関係数**を定義しました。相関係数は -1 以上1以下の値をとり，1に近いほど正の相関関係が強いことを示し，-1 に近いほど負の相関関係が強いことを示しています。

その後，今後の目標を以下のように定めました。

ここまでは，調査や実験などで手に入れたデータの特徴を記述する段階でした。これからは，手に入れた一部のデータ（標本）から，全体のデータ（母集団）について推定したり，立てた仮説が正しいと判断できるかどうかを検定することを目標とします。

この目的のためには，確率の考え方が欠かせません。そこで，ここからしばらくは確率変数や確率分布など，統計の中で必要となる確率の考えについて学びます。

このように当面の目標を定めて，数学Aの授業で扱った内容の繰り返しも一部含みながら，**確率変数の期待値**（**平均**），**分散**，**標準偏差**について取り上げ，最後に**二項分布**について簡単に触れるところまで進みました。

2年次の授業「のぞみ」では，1年次の土曜に行った統計入門講座に続く内容を扱います。1年次の終わりに駆け足で取り上げた「二項分布」から始めることにします。

2　二項分布

数学 A「場合の数と確率」の章中の「反復試行の確率」の節で，次のような問題を扱った。

問題 2.1. 1 個のさいころを 5 回続けて投げるとき，6 の目がちょうど 2 回出る確率を求めよ。

6 の目が出る回数を X で表すと，X のとりうる値は全部で $0,1,2,3,4,5,6$ の 7 通りある。上の問題は，$X=2$ となる確率 $P(X=2)$ を問うている。変数 X のとりうる値 r のおのおのに対して，X が値 r をとる確率 $P(X=r)$ が対応する。このような変数 X を，**確率変数**と呼んだ。確率変数のとりうる値と，その値をとる確率との対応を**確率分布**といい，その対応を表にしたものを確率分布表と呼んだ。

期待値の計算をする際には，確率分布が必要となった。

問題 2.2. 1 枚の硬貨を 3 回続けて投げるとき，表が出る回数の期待値を求めよ。

次に，一般に二項分布の定義を述べる。上の問題にみえるように，すでに二項分布の例には出会っていることが分かるだろう。

定義 2.1. 確率変数 X のとりうる値が $0,1,2,\cdots,n$ の $n+1$ 通りであり，X が値 r をとる確率 $P(X=r)$ が次式で与えられるとき，この確率変数 X は**二項分布** $B(n,p)$ **に従う**という。

$$P(X=r)={}_n\mathrm{C}_r p^r q^{n-r},\ 0\leqq p\leqq 1,\ p+q=1$$

問題 2.3. 次の確率変数 X は，いずれも二項分布に従う。どのような二項分布に従うのかを，記号 $B(n,p)$ を用いて表せ。ただし，さいころは 1 から 6 までの目のそれぞれが出る確率が 1/6 である理想的なさいころとし，硬貨は投げたときに表と裏のでる確率がそれぞれ 1/2 であるような理想的な硬貨であるとする。

(1) 1 枚の硬貨を 300 回続けて投げるとき，表がでる回数 X

(2) 1 個のさいころを 100 回続けて投げるとき，5 以上の値が出る回数 X

(3) ある工場で生産する製品は，1000 個中 23 個が不良品であることが経験的に分かっている。これ前提として，この工場で生産された製品を無作為に 50 個とりだしたとき，この中に含まれる不良品の個数 X

3　二項分布の利用

確率変数 X が二項分布 $B(n,p)$ に従うとする。二項分布を利用するにあたって必要な手順を，次の 2 段階に分けて考える。

(1) 確率変数 X の確率分布を求める。(確率分布表を作る)

(2) 確率分布を使う。

確率分布表を作るためには，面倒な四則計算をかなりの分量実行しなければならない。数学 A の「場合の数と確率」の章で期待値を扱った部分や，数学 C の教科書の「確率と確率分布」の章では，比較的簡単な例について確率分布を求めることが問題となっている。

ここでは，後半の確率分布の使い方の簡単な例を取り上げてみよう。(30 数年前の高校数学参考書からの引用)

例 3.1. ある工場に同一の機械が 20 台ある。どの機械も 6 日に 1 台の割合で使用できなくなる。この機械が 5 台以上使えない日は 1ヶ月の間に何日ぐらいであると考えられるか。ただし，二項分布 $B(20,\frac{1}{6})$ の確率分布表の一部は次のようになっている[1]。

X	0	1	2	3	4	5	6	$\cdots$
P	0.02	0.10	0.20	0.24	0.20	0.13	0.06	$\cdots$

問題 3.1. ある工場で生産する製品は，1000 個中 23 個が不良品であることが経験的に分かっている。これ前提として，この工場で生産された製品を無作為に 50 個とりだしたとき，この中に含まれる不良品の個数が 2 個以上になる確率を求めよ。ただし，二項分布 $B(50,\frac{23}{1000})$ の確率分布表の一部は次のようになっている。

X	0	1	2	3	4	5	6	$\cdots$
P	0.31	0.37	0.21	0.08	0.02	0.00	0.00	$\cdots$

[1]表の中で，たとえば $X=3$ となる確率は ${}_{20}C_3(\frac{1}{6})^3(\frac{5}{6})^{17}$ を計算した値（の近似値）なのだな，ということはわかるようになっておいてほしい。

硬貨を投げて表が出る確率が $\frac{1}{2}$ であるとしても，10 回硬貨を投げたときちょうど 5 回表が出るとは限らない。しかし，だから表が出る回数について何が起こるかまったくわからないわけではなく，10 回中 4 回，5 回，6 回出ることのほうが，10 回中 0 回，1 回，9 回，10 回というような極端に偏って出ることよりも起こりやすいということはいえるだろう。

硬貨を n 回投げて，表が r 回でたとき，$\frac{r}{n}$ を表の出た回数の**相対度数**という。次の問題は，表の出た回数の相対度数と，数学的確率 $\frac{1}{2}$ とのずれに関するものである。

問題 3.2. 1 枚の硬貨を 10 回投げたとき，表が r 回出たとする。このとき，相対度数 $\frac{r}{10}$ と数学的確率 $\frac{1}{2}$ との差が 0.1 以下，すなわち

$$\left|\frac{r}{10}-\frac{1}{2}\right| \leqq 0.1$$

である確率を求めよ。ただし，二項分布 $B(10, \frac{1}{2})$ の確率分布表の一部は次のようになっている。

X	$\cdots$	3	4	5	6	$\cdots$
P	$\cdots$	0.117	0.205	0.246	0.205	$\cdots$

投げる回数を増やせば，相対度数は数学的確率に近づく。上の問題で投げる回数を 100 回にすれば，相対度数と数学的確率との差が 0.1 以下になる確率はもっと大きくなるだろう。しかし，たとえ二項分布 $B(100, \frac{1}{2})$ の確率分布表を与えておいたとしても，これを手で計算するのは面倒である。

4　二項分布の期待値（平均）と標準偏差

確率変数 X の期待値（平均），分散，標準偏差の定義を思い出しておこう。X は次の確率分布に従うものとする。

X	x_1	x_2	$\cdots$	x_n
P	p_1	p_2	$\cdots$	p_n

確率変数の期待値

上の確率分布表に従う確率変数 X の期待値（平均値）$E(X)$ を次式で定義する。

$$E(X) = \sum_{k=1}^{n} x_k p_k = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$$

確率変数の分散

確率変数 X に対し，確率変数 $(X-\mu)^2$（ここで $\mu = E(X)$）の期待値 $E((X-\mu)^2)$ を X の分散といい，$V(X)$ と表す。

$$V(X) = E((X-\mu)^2) = \sum_{k=1}^{n} p_k (x_k - \mu)^2$$

確率変数の標準偏差

確率変数 X に対し，分散の正の平方根を標準偏差といい，$\sigma(X)$ と表す。

$$\sigma(X) = \sqrt{V(X)}$$

さて，数学 A の教科書では，定義に基づいて期待値の計算をした。問題 2.2 は，二項分布 $B(3, \frac{1}{2})$ に従う確率変数の期待値を求める問題に他ならない。二項分布 $B(n,p)$ について確率分布表が文字 n, p を含んだ式で表せるのだから，数学でいつもするように，文字 n, p を含んだ式のまま定義に基づいて期待値（平均）や分散，標準偏差を求める式をいったん作ってしまえば，以後はそれを公式として使うことができる。結論からいうと，次のようになる。

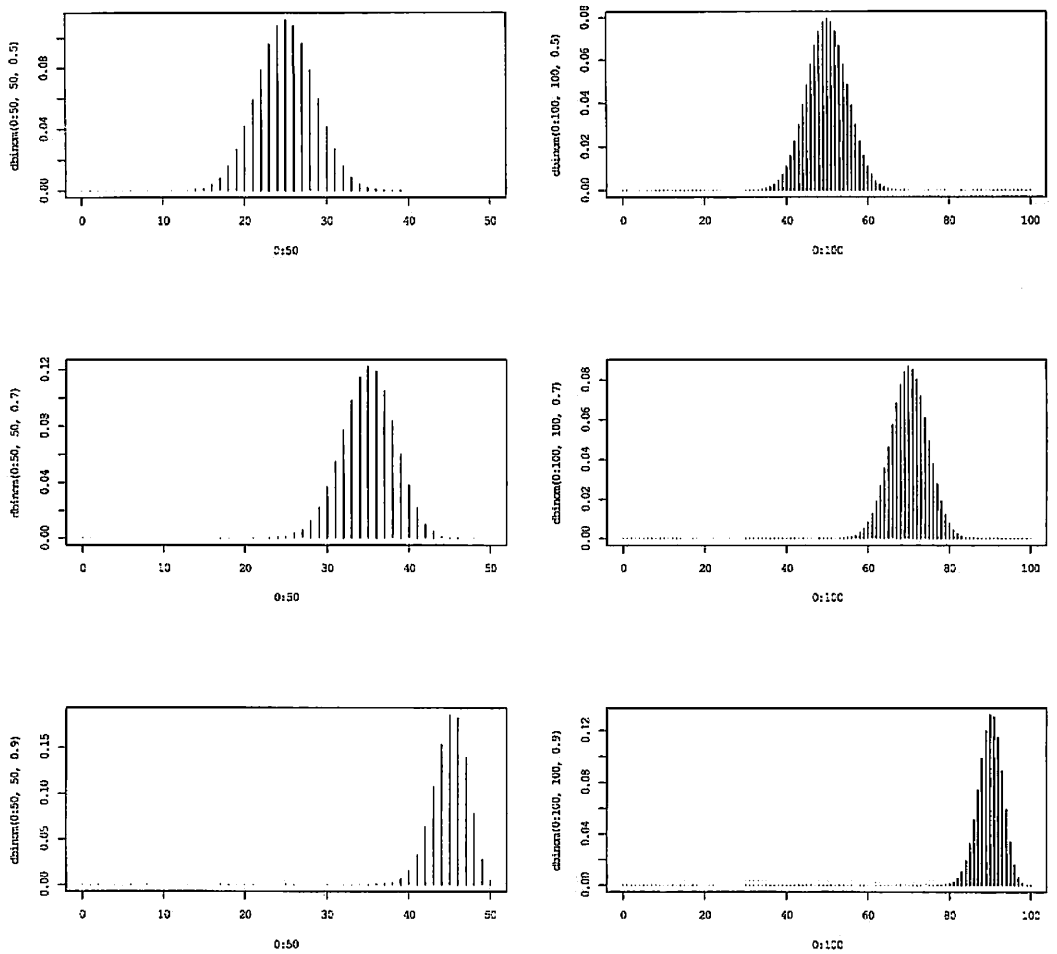
定理 4.1. 確率変数 X が二項分布 $B(n,p)$ に従うとき，$q = 1-p$ とすると，

$$E(X) = np, \quad V(X) = npq, \quad \sigma(X) = \sqrt{npq}$$

問題 4.1. 100 個のさいころを同時に投げたとき，1 の目が出る個数の期待値を求めよ。

証明は，数学 C の教科書 p.122 から p.132 までを参照。このテキストでは統計の考え方を理解するための説明に重点を置き，もう少し使える数学的な道具立てが揃ってから，余裕があれば証明にもどろうと思う。

二項分布を棒グラフの形にしたものを眺めておこう。



左の列が，上から順に $B(50, 0.5)$，$B(50, 0.7)$，$B(50, 0.9)$ に従う確率分布であり，右の列が上から順に $B(100, 0.5)$，$B(100, 0.7)$，$B(100, 0.9)$ に従う確率分布である。

問題 4.2. 上にグラフを示したそれぞれの二項分布に対して，公式を用いて平均，分散，標準偏差を計算し，その値とグラフの様子とを観察して整合性をもつことを確かめよ。

5　離散型確率変数と連続型確率変数

さいころを投げて出た目の数や，硬貨を5回投げて表が出た回数など，とりうる値が0,1,2,3,…というようにとびとびの値をとる確率変数を，特に**離散型確率変数**という。それに対して，金属棒の長さの測定値などのように，連続的な値をとると考えることのできる確率変数がある。長さの測定値は，測定を繰り返したとき，毎回完全に同一の値をとることはなく，測定誤差を含んで測定のたびにわずかに異なるのが普通である。測定値は，真の値に近い値をとる確率が高く，真の値から大きく離れた値をとる確率は小さい，というように，確率変数と考えることができる。このような連続的な値をとる確率変数を，**連続型確率変数**という。

6　連続型確率変数と確率

離散型確率変数 X の場合には，さいころを投げて2の目が出る確率は $\frac{1}{6}$，というように，X が一つの値をとる確率 $P(X=2)$ というものを考えた。X のとりうる値のすべてに対して，その値をとる確率を表にしたものを確率分布といった。

連続型確率分布の場合には，とりうる値が連続的に無数にあるため，確率変数が特定の1つの値をとる確率，という考え方をすると辻褄が合わなくなってしまう。そのため，連続型確率変数 X については，X の値がある区間に入る確率，という考え方をする。たとえば，大手前高校の全生徒の中から無作為に一人選んで身長を測ったとき，その値が165cm以上170cm未満である確率を求める，というような問いの立て方をする。(無作為に一人選んで身長を測ったとき，その値がちょうど170cmである確率は，という問いの立て方は，連続型確率変数として扱ったときにはしない。) この例で，身長の測定値を X としたとき，X の値が $165 \leqq X < 170$ の範囲にある確率を次のように表す。

$$P(165 \leqq X < 170)$$

7　連続型確率変数の分布—確率密度関数

連続型確率変数の場合には，とりうる値が連続的であるので，離散的確率分布のときのような確率分布表を作ることができない。連続的確率変数の分布を表す方法を考えよう。

まず，統計入門の最初のほうで取り上げた，データの整理の仕方を思い出そう。そこでは，データの取りうる値を幾つかの階級に分けて，ある階級に属するデータの個数を調べて度数分布表を作った。この考え方は，連続的確率変数において，変数の値がある区間に属する確率を調べるという考え方に近い。

例 7.1. 次のようなデータがあったとしよう。

58	58	53	40	49	47	56	46	60	46	61	40	37	82	46	53	56
45	55	54	48	51	50	71	43	39	50	53	54	45	39	63	47	41
48	48	61	51	58	45	52	47	51	41	37	70	56	37	44	38	72
63	47	55	46	45	42	44	67	49	51	52	62	45	40	67	46	43
38	37	44	56	61	57	46	51	43	43	59	40	70	49	52	43	44
37	48	54	53	42	42	45	65	39	48	69	49	36	43	55		

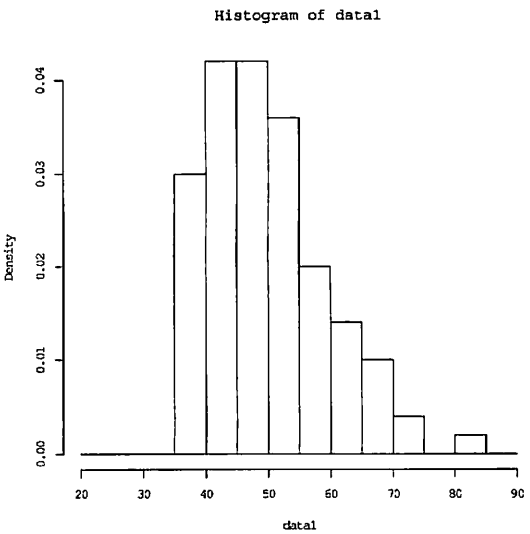
このデータの度数分布，および相対度数分布表は下図左のようになる。ここで，相対度数は，各階級に属するデータの個数を，データの総数で割ったものである。また，密度の欄には，相対度数を階級の幅で割った値が書いてある。このとき，

(密度) × (階級の幅) = (相対度数)

となっている。

この密度をヒストグラムにしたものが，下図右である。

階　級	度数	相対度数	密度
35 以上 40 未満	11	0.11	0.022
40 以上 45 未満	19	0.19	0.038
45 以上 50 未満	25	0.25	0.050
50 以上 55 未満	17	0.17	0.034
55 以上 60 未満	12	0.12	0.024
60 以上 65 未満	7	0.07	0.014
65 以上 70 未満	4	0.04	0.008
70 以上 75 未満	4	0.04	0.008
75 以上 80 未満	0	0.00	0.000
80 以上 85 以下	1	0.01	0.002
合計	100	1.00	



上図右の，密度のヒストグラムは，次のような特徴をもっている。

(1) 一つの階級に対応する「棒」の面積は，その階級の相対度数に一致

(2) すべての「棒」の面積の合計は 1

(3) このデータから無作為に 1 つのデータを取り出しその値を X とするとき，$55 \leqq X < 70$ となる確率 $P(55 \leqq X < 70)$ はこの範囲 $55 \leqq X < 70$ の部分の「棒」の面積の和である。

このように，面積を利用して確率分布を表すことができる。この考え方を，連続型確率変数の分布を表すために使う。

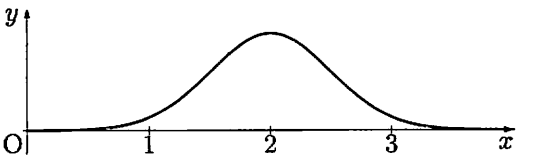
8　正規分布

確率密度関数 $f(x)$ が次式で与えられるような確率分布を**正規分布**という。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

数学の授業で現時点までに扱った内容では，まだこの式の表す内容を十分に理解することはできない。たとえば，式の中に現れる e という文字は**自然対数の底**を表すが，これは数学 III の授業で扱う。e はある定数であり，その近似値は $e=2.71828\cdots$ である。

ここでは，上の関数 $f(x)$ のグラフに基づきながら話を進めよう。次のグラフは，$m=2$，$\sigma=0.5$ の場合の $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ のグラフである。



$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ は次の性質を持つ。

(1)　$f(x)\geqq 0$ である。

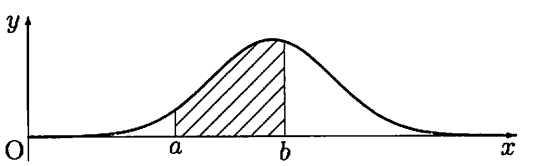
(2)　$f(x)$ のグラフと x 軸で挟まれた部分の面積は 1 になる。

グラフと x 軸とで挟まれた部分の面積は 1 になるという事実を計算によって示そうとすると，やはり授業で扱った範囲の数学的な知識・技術ではまだ足りない。ここでは，これら事実を認めて，統計の考え方を理解することに重点を置く。

「確率密度関数 $f(x)$ が次式で与えられるような分布を**正規分布**という」という 1 行の意味を，もう少し丁寧に書くと次のようになる。

実数値を取る変数 X について，X の値が $a\leqq X\leqq b$ となる確率を $P(a\leqq X\leqq b)$ と書く。

定義 8.1. 連続的な値をとる変数 X について，確率 $P(a\leqq X\leqq b)$ が $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ のグラフを用いた次の斜線部の面積で表されるとき，この X は**正規分布 $N(m,\sigma^2)$ に従う**という。

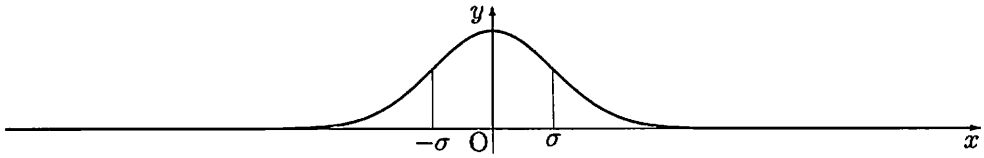


このとき，$f(x)$ を正規分布 $N(m,\sigma^2)$ の**確率密度関数**といい，m を**平均**，σ を**標準偏差**という。

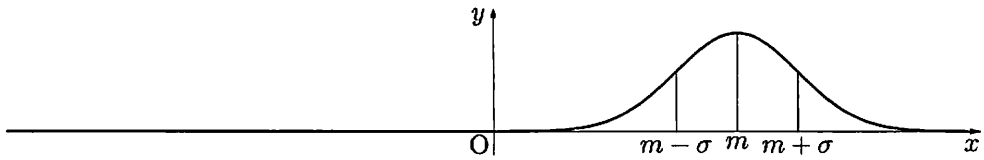
$m=0$ の場合，正規分布 $N(0,\sigma^2)$ の確率密度関数は次のようになる。

$$f_0(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}\quad (m=0\text{ の場合の正規分布の密度関数})$$

式の右辺に，変数 x は x^2 という形でのみ現れているから，$f_0(x)=f_0(-x)$ が成り立つ。これは，関数 $f_0(x)$ が偶関数であることを表しており，そのグラフは y 軸に関して対称である。実際，そのグラフを書くと次のようになる。

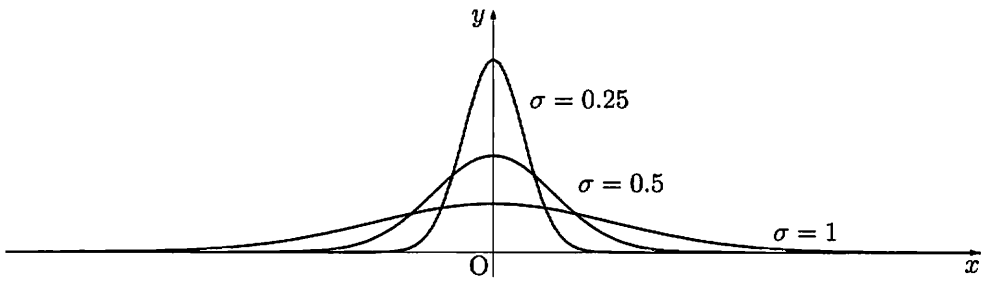


ここで，$f_0(x)$ の式の中で，x のところを $x-m$ で置き換えたものが一般の $f(x)$ である。このとき，グラフは x 軸方向へ m だけ平行移動したものになる。



$f_0(x)$ のグラフを平行移動した $f(x)$ のグラフは直線 $x=m$ に関して対称である。

データの整理でみたように，分布が対称のとき中央値と平均値とは一致した。そのことと，正規分布の確率密度関数のグラフが直線 $x=m$ について対称であり，対称の中心 m を平均と呼んでいることとが整合性を持つ。$m=0$ とし，σ の値を変えて $f(x)$ のグラフを書くと次のようになる。



σ の値が小さいほど，グラフの広がりは狭くなり，平均の近くに集中する傾向を持つ。これは，標準偏差の意味と整合性を持つ。

連続型確率変数の期待値（平均）$E(X)$ と分散 $V(X)$ は積分を用いて定義される。積分を知らないので，今は式の意味はわからなくてもよい。正規分布の場合の結果をその後にまとめておく。

定義 8.2. 連続型確率変数 X が，確率密度関数 $f(x)$ の表す確率分布に従うとき，X の期待値（平均）$E(X)$，分散 $V(X)$ を次式で定義する。ここで，$\alpha \leqq X \leqq \beta$ は X のとりうる値の範囲とする。

$$E(X)=\int_\alpha^\beta xf(x)\,dx\quad V(X)=\int_\alpha^\beta (x-m)^2f(x)\,dx$$

定理 8.3. 確率変数 X が正規分布 $N(m,\sigma^2)$ に従うとき，期待値 $E(X)=m$，分散 $V(X)=\sigma^2$，標準偏差 $\sigma(X)=\sigma$ が成り立つ。

9　正規分布の標準化

確率変数 X が正規分布 $N(m,\sigma^2)$ に従うとする。X の値から，平均 m を引いた新しい確率変数 $Y=X-m$ をつくると，この確率変数 Y の平均は 0 になる。さらに，確率変数 Y の値を標準偏差で割った新しい確率変数 $Z=\dfrac{Y}{\sigma}$ を作ると，この確率変数の平均は 0，標準偏差は 1 になる。

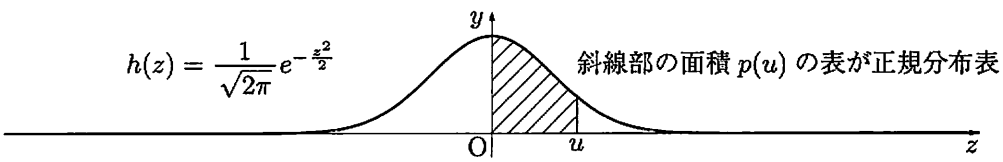
一般に正規分布 $N(m,\sigma^2)$ に従う確率変数 X が与えられたとき，

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

とおくことにより，平均 0，標準偏差 1 の正規分布（これを**標準正規分布**という）に変換することができる。標準正規分布 $N(0,1)$ の確率密度関数 $h(z)$ は次のようになる。(以上の証明には積分の知識が必要である。)

$$h(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

標準正規分布については，そのグラフにおける次図斜線部の面積をいろいろな u の値に対して表にしたもの（正規分布表）が用意されている。



したがって，正規分布に従う確率変数がある区間内の値をとる確率を求めるには，まず「平均を引いて標準偏差で割る」という操作で標準正規分布に従う確率変数に作り変え，次に，正規分布表を使う，という手順を踏めばよい。

10　正規分布表の使い方

例 10.1. 標準正規分布に従う確率変数 Z が，$0.8 \leqq Z \leqq 1.2$ の範囲内の値をとる確率を，正規分布表を利用して求めよ。

解答. 正規分布表により，$P(0 \leqq Z \leqq 1.2)=p(1.2)=0.3849$，$P(0 \leqq X \leqq 0.8)=p(0.8)=0.2881$ だから，

$$P(0.8 \leqq Z \leqq 1.2)=p(1.2)-p(0.8)=0.3849-0.2881=0.0968$$

である。　□

一般の正規分布に従う確率分布について，標準正規分布に変換した後に正規分布表を利用する例を挙げる。

例 10.2. ある高等学校の 2 年生男子生徒 180 人の身長が，平均 171.2cm，標準偏差 5cm の正規分布に大体従うものとする。このとき，次の問いに答えよ。

(1) 身長が 165cm から 175cm までの生徒の人数はおよそ何人か。

(2) 身長の高いほうから 50 人にはいるのは，約何 cm 以上の生徒か。

解答. (1) 身長を X とし，$Z = \dfrac{X - 171.2}{5}$ とおくと，Z は標準正規分布 $N(0,1)$ に従う。また，

$$165 \leqq X \leqq 175 \iff -1.24 \leqq Z \leqq 0.76$$

であるから，

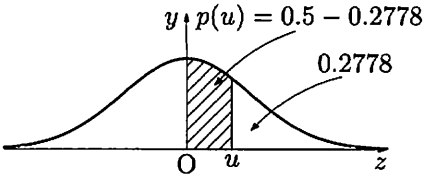
$$\begin{aligned} P(165 \leqq X \leqq 175) &= P(-1.24 \leqq Z \leqq 0.76) \\ &= p(1.24) + p(0.76) \quad (\text{正規分布の対称性を使った}) \\ &= 0.3925 + 0.2764 = 0.6689 \end{aligned}$$

$180 \times 0.6689 = 120.402$. よって，約 120 人

(2) $50/180 = 0.27777\cdots$ であるから，高いほうから 50 人にはいるとは，高いほうから約 27.78%以内にはいることである。$p(u) = 0.5 - 0.2778 = 0.2222$ となる u を正規分布表から探すと，$u = 0.59$ のときがもっとも近い。ここで，

$$Z \geqq 0.59 \iff \frac{X - 171.2}{5} \geqq 0.59 \iff X \geqq 174.15$$

であるから，約 174cm 以上の生徒が高いほうから 50 人にはいる。



問題 10.1. ある県における 17 歳の女子高校生の身長が，平均 157.4cm，標準偏差 4.9cm の正規分布に大体従うものとする。

(1) 身長 170cm 以上の生徒は，1000 人中何人くらいか。

(2) 身長 154cm の生徒は，50 人の中で身長の高い方からおよそ何番目くらいか。

(3) 200 人中，身長の低いほうから 50 人の中に入るのは，約何 cm 以下の生徒か。

問題 10.2. 出生男児の体重が，平均 3.08kg，標準偏差 0.34kg の正規分布に大体従うという。

(1) 無作為に選んだ 1000 人の出生男児のうち，体重が 2.9kg 以上 3.1kg 以下である男児はおよそ何人くらいか。

(2) 体重が重いほうから 5%以内に入るのは，およそ何 kg 以上の体重の男児か。

問題 10.3. ある錠剤中に含まれる有効成分量が，平均 200 単位，標準偏差 50 単位の正規分布に従っているとき，有効成分量が 180 単位以上 220 単位以下である錠剤は全体の約何%を占めていると考えられるか。

問題の解答

問題 10.4. ある県における 17 歳の女子高校生の身長が，平均 157.4cm，標準偏差 4.9cm の正規分布に大体従うものとする。

(1) 身長 170cm 以上の生徒は，1000 人中何人くらいか。

(2) 身長 154cm の生徒は，50 人の中で身長の高い方からおよそ何番目くらいか。

(3) 200 人中，身長の低いほうから 50 人の中に入るのは，約何 cm 以下の生徒か。

解答. 身長 X (cm) は，正規分布 $N(157.4, 4.9^2)$ に従う。$Z = (X - 157.4)/4.9$ と変換すると，確率変数 Z は標準正規分布 $N(0,1)$ に従う。

(1) $X \geqq 170 \iff Z \geqq (170 - 157.4)/4.9 = 2.571\cdots$ であるから，

$$\begin{aligned} P(X \geqq 170) &= P(Z \geqq 2.57) \\ &= 0.5 - P(0 \leqq Z \leqq 2.57) \\ &= 0.5 - p(2.57) \\ &= 0.5 - 0.4949 \text{ (正規分布表より)} \\ &= 0.0051 \end{aligned}$$

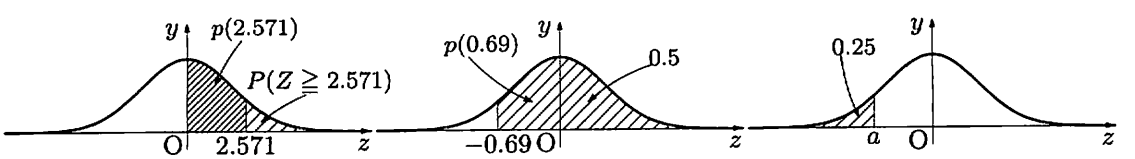
$0.0051 \times 1000 = 5.1$ より，身長 170cm 以上の生徒は約 5 人である。

(2) $X \geqq 154 \iff Z \geqq (154 - 157.4)/4.9 = -0.6938\cdots$ であるから，

$$\begin{aligned} P(X \geqq 154) &= P(Z \geqq -0.69) \\ &= 0.5 + p(0.69) \\ &= 0.5 + 0.2549 = 0.7549 \end{aligned}$$

$50 \times 0.7549 = 37.745$ より，身長 154cm 以上の生徒が約 38 人いると考えられる。よって，高いほうからおよそ 38 番である。

(3) 200 人中低いほうから 50 人は，低いほうから 25%にあたる。標準正規分布に従う確率変数 Z について，$P(Z \leqq a) = 0.25$ となる a は負の値であり，$P(Z \leqq a) = 0.25 \iff p(-a) = 0.5 - 0.25 = 0.25$ だから，正規分布表から $p(u) = 0.25$ に最も近い u を探すと $u = 0.67$ のときであることがわかる。従って，$a = -0.67$。このとき，$Z \leqq -0.67 \iff X \leqq 157.4 - 0.67 \times 4.9 = 154.117$ である。したがって，低いほうから 50 人に入るのは約 154cm 以下の生徒である。 □



問題 10.5. 出生男児の体重が，平均 3.08kg，標準偏差 0.34kg の正規分布に大体従うという。

(1) 無作為に選んだ 1000 人の出生男児のうち，体重が 2.9kg 以上 3.1kg 以下である男児はおよそ何人くらいか。

(2) 体重が重いほうから 5%以内に入るのは，およそ何 kg 以上の体重の男児か。

解答. 体重 X（kg）は正規分布 $N(3.08, 0.34^2)$ に従う。$Z = (X - 3.08)/0.34$ と変換すると，確率変数 Z は標準正規分布 $N(0,1)$ に従う。

(1) $2.9 \leqq X \leqq 3.1 \iff (2.9 - 3.08)/0.34 \leqq Z \leqq (3.1 - 3.08)/0.34$ だから，

$$\begin{aligned} P(2.9 \leqq X \leqq 3.1) &= P(-0.529\cdots \leqq Z \leqq 0.058\cdots) \\ &= p(0.53) + p(0.06) \\ &= 0.2019 + 0.0239 \\ &= 0.2258 \end{aligned}$$

$1000 \times 0.2258 = 225.8$ より，約 226 人である。

(2) 正規分布表で $p(u) = 0.5 - 0.05 = 0.45$ に最もちかい u を探すと，$u = 1.64$（または $u = 1.65$）である。$Z \geqq 1.64 \iff X \geqq 3.6376$ だから，約 3.64kg 以上の男児である。 □

問題 10.6. ある錠剤中に含まれる有効成分量が，平均 200 単位，標準偏差 50 単位の正規分布に従っているとき，有効成分量が 180 単位以上 220 単位以下である錠剤は全体の約何%を占めていると考えられるか。

解答. 有効成分量を X とすると，確率変数 X は正規分布 $N(200, 50^2)$ に従う。$Z = (x - 200)/50$ と変換すると，確率変数 Z は標準正規分布 $N(0,1)$ に従う。ここで，

$$180 \leqq X \leqq 220 \iff -0.4 \leqq Z \leqq 0.4$$

であるから，$P(180 \leqq X \leqq 220) = P(-0.4 \leqq Z \leqq 0.4) = 2p(0.4) = 0.3108$。したがって，全体の約 31%を占めている。

11　正規分布における区間と確率との対応

11.1　正規分布に従うデータの95%を含む範囲

前項の問題を解く際には，正規分布 $N(m,\sigma^2)$ に従う確率変数 X に対して，変換 $Z=(X-m)/\sigma$ を行うことにより，標準正規分布 $N(1,0)$ に従う確率変数 Z に作り変えて，正規分布表を利用した。

この変換は，言葉でいえば,「平均を引いて，標準偏差で割る」という変換だから，その結果得られた確率変数 Z の値は,

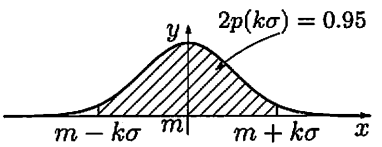
X の値が，平均から標準偏差の何倍ずれているか

を表している。平均や標準偏差の異なるさまざまな正規分布にしたがう確率変数 X 対して，標準化すれば正規分布表からある範囲に値をとる確率がわかるということは，どのような正規分布に従う確率変数 X 対しても「平均から標準偏差の何倍以内のずれ」という表し方で範囲を指定すれば，その範囲内に値をとる確率は正規分布の平均や標準偏差によらず一定であるということを表している。

問題 11.1. 正規分布 $N(m,\sigma^2)$ に従う確率変数 X について，$m-\sigma \leqq X \leqq m+\sigma$ となる確率 $P(m-\sigma \leqq X \leqq m+\sigma)$ を正規分布表を利用して求めよ。

逆に，正規分布に従う確率変数について,「確率 0.95 で X の値が含まれるような範囲はどのような範囲か」という問いの立て方をすることがある。これを正規分布表から読み取ってみよう。

問題 11.2. 正規分布 $N(m,\sigma^2)$ に従う確率変数 X について，$P(m-k\sigma \leqq X \leqq m+k\sigma)=0.95$ となるような k の値を，正規分布表を利用してもとめよ。



正解は $k=1.96$ である。**どのような正規分布においても，データの95%は平均値からのずれが標準偏差の1.96倍以内に収まっていると考えてよい。**この性質は後に重要になる。

問題 11.3. 偏差値とは，平均 50，標準偏差 10 となるようにデータを変換したものであった。ある試験の受験者全体の得点分布が正規分布に従うと仮定したとき，次の問いに答えよ。

(1) 偏差値 $50\pm\alpha$ の間に全受験生のほぼ95%が含まれるようにしたい。α の値を求めよ。

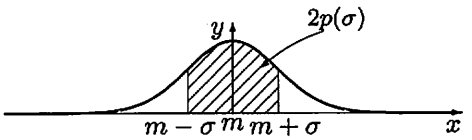
(2) この試験で偏差値 70 以上の受験生は，全受験生の何%程度いると考えられるか。

11.2 代表的な区間と確率との対応

平均 m，標準偏差 σ の正規分布 $N(m, \sigma^2)$ に従う確率変数 X について，その値が平均を中心とするある区間 $m - k\sigma \leqq X \leqq m + k\sigma$ に入る確率を $k = 1, 2, 3$ の場合にまとめておこう。標準化「平均を引いて，標準偏差で割る」という操作をした後，正規分布表から読み取るといういつもの手順で求めることができる。

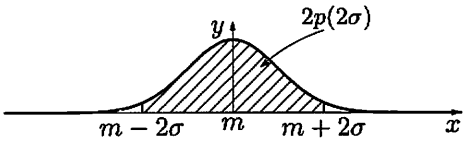
平均からのずれが標準偏差の 1 倍以内である確率

$P(m - \sigma \leqq X \leqq m + \sigma) =$ ☐



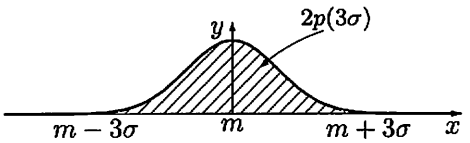
平均からのずれが標準偏差の 2 倍以内である確率

$P(m - 2\sigma \leqq X \leqq m + 2\sigma) =$ ☐



平均からのずれが標準偏差の 3 倍以内である確率

$P(m - 3\sigma \leqq X \leqq m + 3\sigma) =$ ☐



逆に，X の値が含まれる確率が指定された値になるような区間を調べておこう。

確率 50%（確率 0.5）で X の値が含まれるような区間

$P(m - k\sigma \leqq X \leqq m + k\sigma) = 0.5$ となるのは，$k =$ ☐ のときである。

確率 90%（確率 0.9）で X の値が含まれるような区間

$P(m - k\sigma \leqq X \leqq m + k\sigma) = 0.9$ となるのは，$k =$ ☐ のときである。

確率 95%（確率 0.95）で X の値が含まれるような区間

$P(m - k\sigma \leqq X \leqq m + k\sigma) = 0.95$ となるのは，$k =$ 1.96 のときである。

確率 98%（確率 0.98）で X の値が含まれるような区間

$P(m - k\sigma \leqq X \leqq m + k\sigma) = 0.98$ となるのは，$k =$ ☐ のときである。

確率 99%（確率 0.9）で X の値が含まれるような区間

$P(m - k\sigma \leqq X \leqq m + k\sigma) = 0.99$ となるのは，$k =$ ☐ のときである。

12　二項分布の正規分布による近似

これまでに，離散型確率分布として二項分布 $B(n,p)$ が，また，連続型確率分布として正規分布 $N(m,\sigma^2)$ が出てきた。この節では，n が十分大きいとき，二項分布を正規分布で近似することができることを例により観察する。二項分布と正規分布はそれぞれ次のようなものだった（復習）。

二項分布 $B(n,p)$

確率変数 X が二項分布 $B(n,p)$ に従うとき，

(1) X の取りうる値は，$0,1,2,\cdots,n$ の $n+1$ 個

(2) $P(X=k)={}_n\mathrm{C}_k p^k q^{n-k}$ $(q=1-p)$

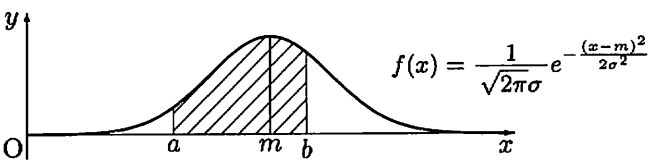
(3) 平均（期待値）$E(X)=np$，分散 $V(X)=npq$，標準偏差 $\sigma(X)=\sqrt{npq}$ $(q=1-p)$

正規分布 $N(m,\sigma^2)$

確率変数 X が二項分布 $N(m,\sigma^2)$ に従うとき，

(1) X の取りうる値は実数

(2) X の値が $a \leqq X \leqq b$ を満たす確率 $P(a \leqq X \leqq b)$ は次の斜線部の面積に一致する。



(3) 平均（期待値）$E(X)=m$，分散 $V(X)=\sigma^2$，標準偏差 $\sigma(X)=\sigma$

主張は次のようなものである。

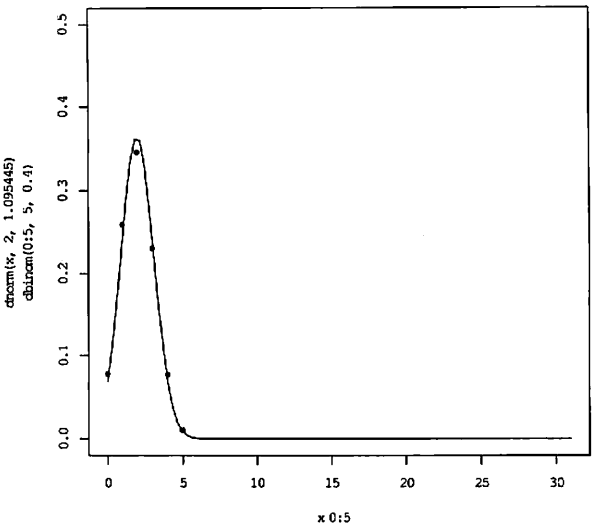
定理 12.1. 二項分布 $B(n,p)$ に従う確率変数 X は，n が十分大きいとき，近似的に正規分布 $N(np,npq)$ に従う。ただし，$q=1-p$ である。

$B(n,p)$ と $N(np,npq)$ は同じ平均，同じ分散（したがって同じ標準偏差）を持つ。この定理は，n が大きければ二項分布 $B(n,p)$ は同じ平均，同じ分散の正規分布で近似できることを主張している。

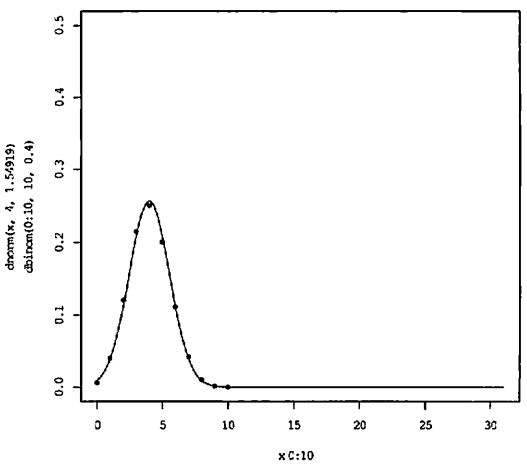
この主張を，二項分布の確率をグラフ用紙にプロットしたものと，正規分布の確率密度関数のグラフを描いたものとを重ねて比べることにより例で観察しよう。

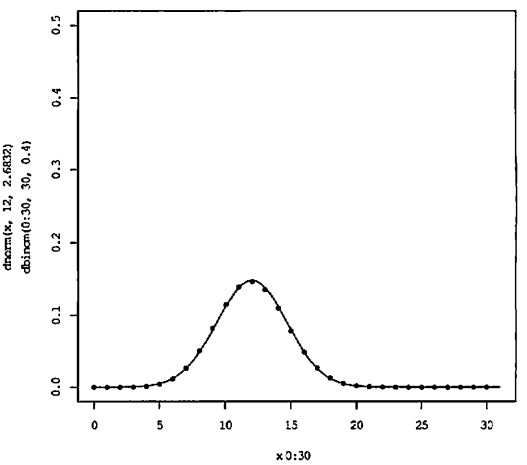
次図は，二項分布 $B(n,p)$ に従う確率変数 X の値が $0,1,2,\cdots,n$ となる確率を黒点でプロットしたものと，正規分布 $N(np,npq)$（ただし，$q=1-p$）の確率密度関数のグラフを重ねて書いたものである。

上段の図は $n=5$，$p=0.4$ の場合である。確率の最大値付近での値が，二項分布と正規分布でずれている様子がわかる。



$n = 10$, $n = 30$ の場合は下図になる（左が $n = 10$, 右が $n = 30$）。$p = 0.4$ は共通である。n が大きくなるにつれてこのずれは目立たなくなり，$n = 30$ の場合には，黒点のプロットはほとんど曲線上にのっているように見える。





さらに標準化して $Z = \dfrac{X - np}{\sqrt{npq}}$ と変換すると，Z は近似的に標準正規分布 $N(0,1)$ に従う。

硬貨を 100 回投げて表が出た回数を X とすると，確率変数 X は二項分布 $B(100, 0.5)$ に従う。この確率変数の値が 49 以上 51 以下になる確率は，$({}_{100}\mathrm{C}_{49} + {}_{100}\mathrm{C}_{50} + {}_{100}\mathrm{C}_{51})\left(\dfrac{1}{2}\right)^{100}$ であるが，計算は大変面倒である。このようなときに，正規分布による近似が使える。

例 12.2. 硬貨を 100 回投げて，表が出た回数を X とする。相対度数 $\frac{X}{100}$ と数学的確率 $\frac{1}{2}$ との差が 0.01 以下になる確率 $P\left(\left|\frac{X}{100}-\frac{1}{2}\right| \leqq 0.01\right)$ を，小数第 3 位を四捨五入して求めよ。

(解)　X は二項分布 $B(100, 0.5)$ に従う。X の平均は $100\times0.5=50$，分散は $100\times0.5\times0.5=25$，標準偏差は $\sqrt{25}=5$ である。X は近似的に正規分布 $N(50,25)$ に従うと考えられるので，$Z=(X-50)/5$ とおくと，Z は近似的に標準正規分布 $N(0,1)$ に従う。このとき，

$$\begin{aligned}\left|\frac{X}{100}-\frac{1}{2}\right| \leqq 0.01 &\iff 49 \leqq X \leqq 50\\ &\iff -0.2 \leqq Z \leqq 0.2\end{aligned}$$

となるので，

$$\begin{aligned}P\left(\left|\frac{X}{100}-\frac{1}{2}\right| \leqq 0.01\right) &= P(-0.2 \leqq Z \leqq 0.2)\\ &= 2p(0.2)\\ &= 0.1586 \quad (\text{正規分布表より})\end{aligned}$$

したがって，求める確率はおよそ 0.16 である。

問題 12.1. 硬貨を 100 回投げて，表が出た回数を X とする。相対度数 $\frac{X}{100}$ と数学的確率 $\frac{1}{2}$ との差が 0.05 以下になる確率 $P\left(\left|\frac{X}{100}-\frac{1}{2}\right| \leqq 0.05\right)$ を，小数第 3 位を四捨五入して求めよ。

例 12.3. 硬貨を n 回投げて，表が出た回数を X とする。相対度数 $\frac{X}{n}$ と数学的確率 $\frac{1}{2}$ との差が 0.01 以下になる確率 $P\left(\left|\frac{X}{n}-\frac{1}{2}\right| \leqq 0.01\right)$ が 0.99 以上になるのは，n がおよそいくら以上のときか。100 未満を切り上げて答えよ。

問題 12.2. 硬貨を n 回投げて，表が出た回数を X とする。相対度数 $\frac{X}{n}$ と数学的確率 $\frac{1}{2}$ との差が 0.01 以下になる確率 $P\left(\left|\frac{X}{n}-\frac{1}{2}\right| \leqq 0.01\right)$ が 0.95 以上になるのは，n がおよそいくら以上のときか。100 未満を切り上げて答えよ。

13　母集団と標本

ある集団について何かを知りたいとき，集団に属する全個体について調査する方法と，集団の中から一部を選び出して調査した結果から集団全体について推測する方法とがある。

集団全体を調査するとき，それを**全数調査**あるいは**悉皆調査**(しっかい)などという。それに対して，集団の中から一部を選び出して調査するとき，**標本調査**という。標本調査を行う場合，知りたいと思う集団全体を**母集団** (population) といい，そこから選び出した母集団の一部を**標本** (sample) という。また，標本を選び出すことを**標本抽出** (sampling) という。標本に含まれる要素の個数を**標本の大きさ**という。標本調査によって集団全体を推測するときには，標本にかたよりがないように，無作為に標本を抽出することが大切になる。

14　母集団分布と母数

たとえば，17歳の大阪府民全体を母集団にとり，その身長について知りたいとする。母集団から無作為に一人を選び出してその身長を調べ，得られた値を X とすると，X は確率変数であり，母集団によって定まるある確率分布に従う。これを**母集団分布** (population distribution) という。

統計的推測の理論では，データはある確率分布に従う確率変数の実現値であり，現象の背後には確率分布がある，という考え方をする。データをできるだけうまく説明するような確率分布を見つけることが，推測統計の目標となる。母集団が現に目の前に存在していなくても，現象の背後に確率分布を想定できれば統計的推測の考え方が適用できる。例えば，目の出方に偏りがあるかも知れないサイコロが1個あるとして，このサイコロの目の出方の偏りを知りたいとしよう。何回かサイコロを振って出た目を調べることが，標本を抽出して調べることに相当する。では母集団に相当するものは何か?サイコロを振って出た目を調べた記録があらかじめ作られていなくてもよい。そのサイコロを投げて出る目を X としたとき，この X が従う確率分布を母集団分布と考える。知りたいものはこの母集団分布である。

母集団分布に対して，期待値や分散など，分布の特徴を表す数値が定まる。これらを**母数** (parameter) という。たとえば母集団が正規分布に従うことがわかっているとき，平均 m と標準偏差 σ が決まれば，正規分布 $N(m, \sigma^2)$ が決まる。母集団について知ることとは，母集団分布について知ることであり，そのためには，母数を知ることが必要である。母集団分布の平均を**母平均**，母集団分布の分散を**母分散**，母集団の標準偏差を**母標準偏差**という。

15　標本の抽出

標本の抽出にはいくつかの方法がある。**単純ランダム・サンプリング（単純無作為抽出）**とは，母集団の対象に番号を振っておき，乱数表などを用いて母集団の各要素が等確率かつ独立に選ばれるように抽出するものである。

また，**層化抽出法**（stratified sampling）とは，調査対象を幾つかのグループ（層という）に分け，各層ごとに独立に標本を抽出する方法である。

たとえば，大阪府民全体に対して調査を行うとき，あらかじめ10代，20代，30代，… と幾つかの層に分けておいて，各層の人口に比例するように選び出す人数を割り当てた上で，10代の府民の中から割り当てられた人数を無作為抽出し，20代の府民の中から割り当てられた府民を無作為抽出し，… というように，世代ごとに割り当てられた人数の調査対象者を抽出することが層化抽出法にあたる。

以下では，標本の抽出は単純ランダム・サンプリングによって行われているものとする。母集団から大きさ n の標本を一組取り出したところ，調べたい値が $\{x_1, x_2, \cdots, x_n\}$ であったとしよう。再度，標本の抽出を行って大きさ n の標本を一組取り出したならば，一般にはさきほどの値とは異なった n 個の値 $\{x'_1, x'_2, \cdots, x'_n\}$ を得る。このように，大きさ n の標本の取りうる値は確率的に決まる。一般に，大きさ n の標本は確率変数の組 $\{X_1, X_2, \cdots, X_n\}$ となる。特に，**母集団から大きさ1の標本 X を取り出したとき，母集団分布の定義により，値 X は母集団分布に従う確率変数である。**（標本 X というとき，標本に付随する調べようとしている値を X で表している。たとえば，大阪府民の身長について調べるために A さんが標本として選ばれたとき，A さんの身長の値を標本 X と表す。）

16 標本平均の標本分布

母平均を知るために，標本の平均を計算することを考えよう。一つの高校の男子生徒の平均身長を調べるために，単純ランダム・サンプリングにより抽出した10人の男子生徒の平均身長を利用するような場合がこれにあたる。10人を抽出して平均身長を計算するという手順を何回も繰り返すと，10人からなる標本の平均値が幾つも得られる。これら，標本の平均値たちは，それ自体で何らかの分布を持つ。

母集団から，大きさ n の標本 $\{X_1, X_2, \cdots, X_n\}$ を取り出したとする。このとき，個々の X_1，X_2，…，X_n は母集団分布に従う確率変数である。このとき，これら n 個の値の平均を**標本平均**といい，$\bar{X}$ で表す。

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

X_1，X_2，…，X_n のそれぞれが確率変数だから，それらの平均である $\bar{X}$ もある確率分布に従う確率変数になる。標本平均 $\bar{X}$ が従う確率分布を，**標本平均の標本分布**という。

標本平均の標本分布を手がかりにして，母集団分布の平均（期待値）について知りたい。そのために，標本平均の標本分布と母集団分布との関係について調べておこう。現実の場面では母集団についてよくわからないから標本を抽出して調べるのであり，あらかじめ母集団分布についてわかっていないのが普通である。しかし，母集団分布や母平均と標本平均の標本分布との関係を，母集団分布がわかっている場合について調べ，それらの関係についての知見を蓄積することにより，その知見を母集団分布がわからない場合に応用することができる。このような観点から，X_1，X_2，…，X_n が**独立に同一の確率分布に従うとき，それらの平均 $\bar{X}$ がどのような確率分布に従うのかを調べてみよう。**

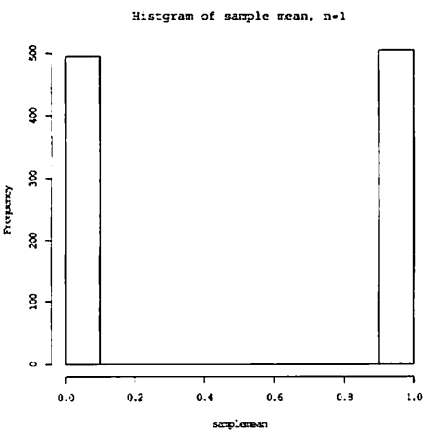
17　母集団分布と標本平均の標本分布の関係

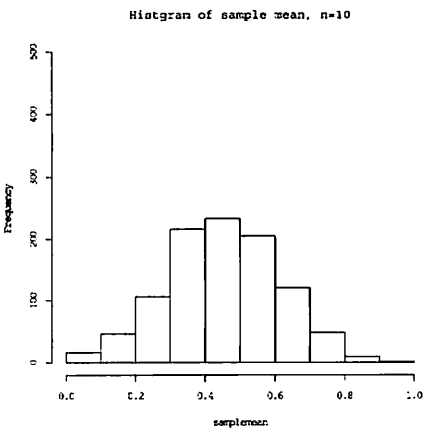
母集団分布と標本平均の標本分布との関係を例によって観察する。観察のポイントは，標本の大きさ n と標本平均の標本分布との関係である。

17.1　母集団分布が二項分布 $B(1,0.5)$ の場合

この場合，母平均は $1\times 0.5=0.5$ である。母集団から大きさ n の標本 $\{X_1,X_2,\cdots,X_n\}$ を抽出する。X_i たちは独立で，同じ二項分布 $B(1,0.5)$ に従う。その標本平均 $\bar{X}$ がどのような確率分布に従うかを調るため，何回も抽出を繰り返し，標本平均を計算して，標本平均のヒストグラムを描こう。

標本抽出は 1000 回行う。実際にはコンピュータに二項分布 $B(1,0.5)$ に従う確率変数の値を繰り返し発生させて以下のヒストグラムを作った。



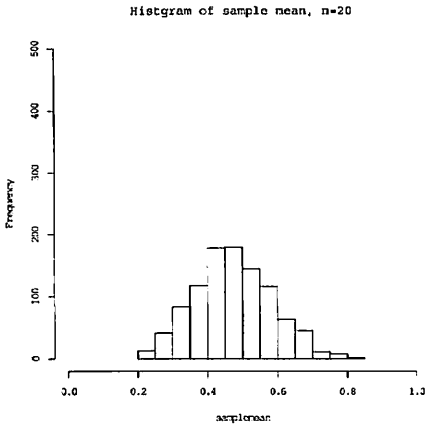


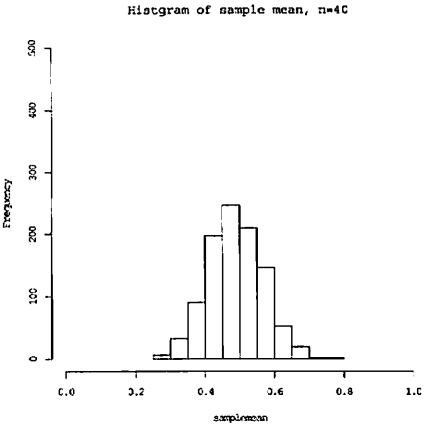
上図左は $n=1$，すなわち大きさ 1 の標本を抽出した場合である。このとき，標本平均は，標本の値そのものであり，標本平均の標本分布は母集団分布そのものとなる。標本平均 $\bar{X}$ の取りうる値は 0 と 1 しかなく，それぞれが確率 0.5 で起こる。(なお，0 以上 1 以下の区間を 10 等分し，0 以上 0.1 未満，0.1 以上 0.2 未満，と階級をとって，最後の階級のみ 0.9 以上 1 以下として 1 を含めているため，値 0 をとった度数が 0.0 以上 0.1 未満のところに現れ，値 1 をとった度数が 0.9 以上 1 未満のところに現れている。)

上図右は，大きさ 10 の標本を抽出した場合である。このとき，標本平均は，抽出した 10 個の値の平均であるから，0.0，0.1，0.2，$\cdots$，1.0 までの 11 通りの値をとる。

大きさ 1 の標本を抽出した場合でも，大きさ 10 の標本を抽出した場合でも，標本平均 $\bar{X}$ の平均(期待値) $E(\bar{X})$ は 0.5 となり母平均に等しい。しかし，標本平均の分布の形は異なる。大きさ 10 の標本を抽出したときの標本平均 $\bar{X}$ の標本分布では，その値が母平均の近くに集まった山形になっていることがわかる。

標本の大きさ $n = 20$ の場合（下図左）と，$n = 40$ の場合（下図右）の標本平均のヒストグラムを挙げる。標本が大きいほど，分布は母平均の近くに集まり，標準偏差が小さくなる傾向が見て取れる。

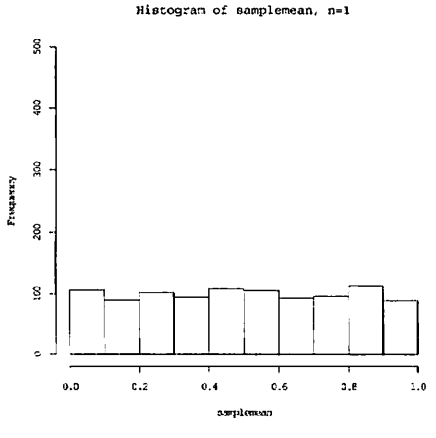


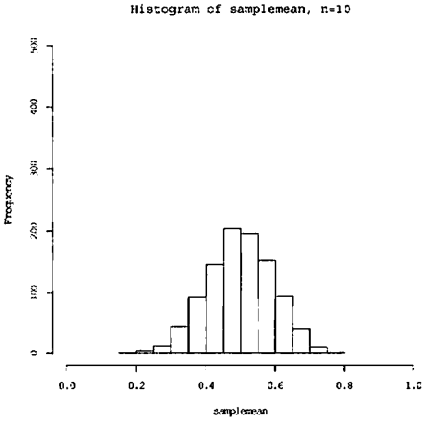


17.2 母集団分布が一様分布の場合

確率密度関数がある区間で一定の値をとるような分布を一様分布という。取りうる値が有限個（N 個）で，各値をとる確率が $1/N$ であるとき，**離散型の一様分布**といい，取りうる値が連続的な区間（たとえば $a \leqq x \leqq b$）で，確率密度関数が $f(x) = \dfrac{1}{b-a}$ であるとき，**連続型の一様分布**という。理想的なさいころを投げて出る目 X は，値 1,2,3,4,5,6 を取る離散型の一様分布に従うと考えられる。

区間 $0 \leqq x \leqq 1$ の値をとる連続型の一様分布が母集団分布であるときに，そこから抽出した大きさ n の標本について，標本平均の標本分布がどのようになるかをコンピュータによる実験によって確かめてみる。下図左は $n = 1$ の場合，右は $n = 10$ の場合である。





実は，**母集団分布が正規分布でなくても，抽出する標本の大きさ n を大きくすれば，標本平均の標本分布は正規分布に近づいていくことが知られている（中心極限定理）**。$n = 10$ から 20 程度で正規分布に十分近くなることが多い。正規分布の重要性はこの事実による。

12 推定

12.1 点推定と区間推定

標本から，母集団の平均や分散など，母数を推定する方法について考える。推定には，**点推定**，**区間推定**という 2 つのタイプの推定方法がある。

点推定とは母数を 1 つの値として推定する方法である。たとえば，1 つの高校の 2 年生の身長の平均を無作為抽出した 50 人の身長から推定するときに，「母集団の平均は 169cm」のような推定を行うのが点推定である。区間推定とは，母数を含む区間の形で推定する方法である。同じ例を用いると，「母平均 m は区間 $165 \leqq m \leqq 174$ に含まれる」のような推定を行うのが区間推定である。

以下，大きさ n の標本を確率変数の組 $\{X_1, X_2, X_3, \cdots, X_n\}$ で表す。それぞれの X_i は母集団と同じ確率分布に従う。標本値から母数を推定する方法と，その特徴について調べる。

12.2 母平均の点推定

標本 $\{X_1, X_2, X_3, \cdots, X_n\}$ から母平均 m を点推定するもっとも自然な方法は，標本平均

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

を用いる方法である。一般に，推定に用いる値を**推定量**という。母平均の推定量として標本平均を使う場合について考える。

標本の平均を計算しても，それが母集団の平均と完全に一致しているというようなことは期待できない。完全に一致はしなくても，推定値として用いることの妥当性を与えるような，いくつかの良い性質を持っている。これについて見ていこう。

標本平均 $\bar{X}$ はそれ自体確率変数である。すなわち，標本を取り直せば，各 X_i の値は母集団分布と同じ確率分布に従って変わり，標本平均の値も変わる。何回も標本平均を取り直せば，標本平均の値がある確率分布に従ってそのつど定まる。これらの標本平均は，あるときは母平均より大きく，あるときは母平均より小さい。そのようなばらつきはあるにせよ，何度も標本を取り直して得た標本平均を更に平均すれば，なおいっそう母平均に近づくだろう（そうであってほしい）。すなわち，標本平均 $\bar{X}$ という確率変数の期待値 $E(\bar{X})$ は，母平均 m に一致してほしい。これを保障するのが次の定理である。

定理 12.1. 期待値が m であるような同一の確率分布に従う確率変数の組 $\{X_1, X_2, \cdots, X_n\}$ に対して，

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

とおくとき，確率変数 $\bar{X}$ の期待値 $E(\bar{X})$ について次式がなりたつ。

$$E(\bar{X}) = m$$

証明. 一般に，確率変数 X，Y と期待値について，次の性質が成り立っていた。

$$E(X+Y)=E(X)+E(Y),\quad E(kX)=kE(X)\ (k\text{ は定数})$$

各 X_i が母集団分布に従い $E(X_i)=m$ だから，上の性質を用いて，

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}(X_1+X_2+\cdots X_n)\right) \\ &= \frac{1}{n}E(X_1+X_2+\cdots X_n) \\ &= \frac{1}{n}\{E(X_1)+E(X_2)+\cdots E(X_n)\} \\ &= \frac{1}{n}(m+m+\cdots+m)=m \end{aligned}$$

となる。すなわち，$E(\bar{X})=m$ が成り立つ。 □

この性質を，推定量 $\bar{X}$ は偏りがないという意味で**不偏推定量**であるという。

推定量が備えてほしい性質の 2 番目として，標本の大きさを大きくするほど，推定値が正確である確率が高くなってほしい，というものがある。

第 12 節例題と問題において，硬貨を n 回投げて表の出た回数を X とするとき，相対度数 $\frac{X}{n}$ と確率 $\frac{1}{2}$ との差が 0.01 以下になる確率が 0.99 以上になるためには何回以上投げなければならないか，また，この確率が 0.95 以上になるためには何回以上投げなければならないかを調べた。1 枚の硬貨を投げたとき，表が出る回数は確率分布 $B(1,0.5)$ に従う。これを母集団分布とする。硬貨を n 回投げて表，裏を調べることは，大きさ n の標本を抽出して値を調べたことに相当する。このとき，相対度数 $\frac{X}{n}$ が標本平均に相当する。標本平均と母平均の差が 0.01 以下になる確率が 0.95 以上になるためには標本の大きさ n がいくら以上でなければならないか，また，この確率が 0.99 以上になるためには標本の大きさ n がいくら以上でなければならないかを求めたのが，問題であり例題であった。これらの問題や例題は，標本平均と母平均との差があらかじめ定めた小さな定数以下になる確率は，標本を大きくすればするほど 1 に近づくという事実の一つの例となっている。

より正確に書くと，どんなに小さな正の定数 ϵ に対しても，標本の大きさ n を限りなく大きくしたとき，標本平均と母平均の差が ϵ 以下となる確率は 1 に近づく。

$$\text{任意の正の定数}\epsilon\text{に対して}, n\to\infty\ \text{のとき}, P(|\bar{X}-m|\leqq\epsilon)\to 1$$

この性質を，標本平均は標本数を大きくするほど母平均に一致していくという意味で**一致推定量**であるという。推定量の望ましさを表す基準は他にもあるが，ここでは以上 2 点に止めておく。

12.3 区間推定

区間推定とは，標本から母数（母平均や母分散など）がその中に入ると考えられる区間 $[L,U]$ を求める推定方法である。

母平均がその中に入ると考えられる区間，という点をもう少し詳しく述べる。以下，母数として母平均を例にとり説明する。

0.95（95%）という値をあらかじめ定めておく。標本を抽出するごとに区間推定によって求めた区間 $[L,U]$ は変化する（すなわち L，U は確率変数である）けれども，繰り返し標本を抽出しなおして区間 $[L,U]$ を求めなおしたとき，95%の割合でそれらの区間の中に真の平均値が入っている，というような性質を持つように区間を求める方法を，区間推定という。このとき，この例における値 0.95 を**信頼係数**といい，区間推定によって求めた区間 $[L,U]$ の L を**下側信頼限界**，U を**上側信頼限界**，$[L,U]$ を **95%信頼区間**という。

信頼係数は，0.95 のほかに，0.99 などもよく用いられる。一般に，区間の中に母平均が入らない確率のほうを文字（たとえば α）で表して，信頼係数を $1-\alpha$ のように書くことが多い。

記号で表せば，確率変数 L，U を用いた区間 $[L,U]$ が $100(1-\alpha)\%$ **信頼区間**であるとは，次式が成り立つことである。

$$P(L \leqq m \leqq U) \geqq 1-\alpha$$

この条件を満たすできるだけ狭い区間を求めたいので，推定の際には

$$P(L \leqq m \leqq U) \fallingdotseq 1-\alpha$$

を満たす区間を求めると考えて解く。

12.4 正規母集団の母平均の区間推定（母分散が既知の場合）

母集団が正規分布 $N(m,\sigma^2)$ に従うという仮定の下で，母平均を区間推定することを考える。

母集団について，その現象の性質から正規分布に従うことが経験的にわかっているが，母平均や母分散などどの正規分布かを定めるパラメータがわからないので推定したい，という状況である。

この場合，母平均も母分散もともに未知であると考えるのが自然だが，未知なものが 2 つある状況をいきなり考えるよりも，未知なものが 1 つの場合をまず考え，それを手がかりに未知なものが 2 つある状況へ進むほうがわかりやすいので，**母平均 m は未知だが，母分散 σ^2 は既知**という場合をまず考える。

次の定理が成り立つ。

定理 12.2. 母集団分布が正規分布 $N(m,\sigma^2)$ に従うとする。ここから大きさ n の標本 $\{X_1,X_2,X_3,\cdots,X_n\}$ をとり，標本平均を $\bar{X}$ で表す。このとき，確率変数 $\bar{X}$ は正規分布 $N(m,\sigma^2/n)$ に従う。

標本平均 $\bar{X}$ の平均（期待値）が m に等しいことは，標本平均が一致推定量であることを述べた際に触れた。また，標本を構成するそれぞれの確率変数 $X_1, X_2, X_3, \cdots, X_n$ は互いに独立であるから，分散 $V(X)$ の性質から，次が成り立つ。

$$V(X_1 + X_2 + X_3 + \cdots + X_n) = V(X_1) + V(X_2) + V(X_3) + \cdots + V(X_n) = n\sigma^2$$

さらに，分散の性質 $V(cX) = c^2V(x)$（c は定数）から，

$$V(\bar{X}) = V\left(\frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

最後に，$\bar{X}$ が正規分布に従うことは，そもそも正規分布の定義をグラフに基づいた直観的な理解にとどめている現段階では，正規分布の確率密度関数を用いた証明をすることができない。正規分布に従う互いに独立な確率変数の和が正規分布に従うこと，および正規分布に従う確率変数の定数倍が正規分布に従うことは，ここでは証明なしに認めることにする。以上をまとめると，$\bar{X}$ は平均 m，分散 σ^2/n の正規分布に従うことになる。

定理 12.3. 母集団分布が正規分布 $N(m, \sigma^2)$ に従うとする。ここから大きさ n の標本 $\{X_1, X_2, X_3, \cdots, X_n\}$ をとり，標本平均を $\bar{X}$ で表す。このとき，95%信頼区間 $[L, U]$ は次のように求められる。

$$[L, U] = \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

【注意】この定理の信頼区間は，$\bar{X}$ と，n と，σ という 3 つの文字を含んでいる。このうち，n は標本の大きさ，$\bar{X}$ は標本平均だから，標本からわかる値である。それに対して，σ は母標準偏差だから，標本からはわからない。このように，標本からはわからない母標準偏差（母分散の正の平方根）を用いているということが，「母分散が既知の場合」という但し書きの意味である。

証明. 母集団分布が $N(m, \sigma^2)$ であるとき，そこから抽出した大きさ n の標本の標本平均 $\bar{X}$ は正規分布 $N(m, \sigma^2/n)$ に従う。正規分布の性質を調べた際に見たように，$\bar{X}$ の値が平均 m から標準偏差 $\sigma/\sqrt{n}$ の 1.96 倍以内に入る確率が約 95% であった。すなわち，

$$P\left(m - 1.96\frac{\sigma}{\sqrt{n}} \leqq \bar{X} \leqq m + 1.96\frac{\sigma}{\sqrt{n}}\right) \fallingdotseq 0.95$$

ここで，不等式を m を中央の項に据えて書き直すと，

$$m - 1.96\frac{\sigma}{\sqrt{n}} \leqq \bar{X} \leqq m + 1.96\frac{\sigma}{\sqrt{n}} \iff \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

したがって，

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \fallingdotseq 0.95$$

すなわち，$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ が 95%信頼区間である。 □

12.5 母分散の点推定（不偏分散）

区間推定の際，母分散が未知の場合には，代わりに標本から求めた母分散の推定量を用いる。

記述統計では，データセットの分散とは，平均からの偏差平方和をデータの個数で割ったものであった。標本 $\{X_1, X_2, X_3, \cdots, X_n\}$ から母分散を点推定する際にも，可能なら母平均 m からの偏差平方和を標本の個数で割った値を使いたい。

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

しかし，母平均もわからないのだから，その代わりに標本平均 $\bar{X}$ を用いざるを得ない。n で割る部分は後回しにして，母平均との差の平方和と，標本平均との差の平方和とではどの程度の違いがあるのかを調べてみよう。

命題 12.4. $\bar{X}$ や X_i は上と同様とし，m を母平均とする。このとき，次式が成り立つ。

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

証明. $X_i - \bar{X} = (X_i - m) - (\bar{X} - m)$ より，

□

命題 12.5. 上記の状況のもとに，$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = (n-1)\sigma^2$

証明. 各 X_i は母集団分布に従うので，その分散は母分散に一致し，$\sigma^2 = E((X_i - m)^2)$ $(i = 1, 2, 3, \cdots, n)$。標本平均 $\bar{X}$ は平均 m，分散 σ^2/n である分布に従うので，$E((\bar{X} - m)^2) = \sigma^2/n$。

$$\begin{aligned} E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) &= E\left(\sum_{i=1}^{n}(X_i - m)^2 - n(\bar{X} - m)^2\right) \\ &= \sum_{i=1}^{n}E\left((X_i - m)^2\right) - nE\left((\bar{X} - m)^2\right) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2 \end{aligned}$$

□

標本平均からの偏差平方和の期待値が，母分散の $n-1$ 倍になった。n 倍ではないことに注意してほしい。ここで，全体を $n-1$ で割ると，

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\sigma^2$$

を得る。この等式は，確率変数 $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2$ の期待値が母分散に等しいことを言っており，言い換えれば，確率変数 $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2$ の値を母分散の推定量として用いれば，それが不偏推定量になっていることを表している。

定義 12.6. 大きさ n の標本 $\{X_1, X_2, X_3, \cdots, X_n\}$ に対し，母分散 σ^2 の不偏推定量 $\hat{\sigma}^2$ を次式で定義する。

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

この推定量を，標本の**不偏分散**という。

これで母集団について母平均も母分散もわからないときに区間推定を行う材料が揃った。しかし，まだ解決しなければならない問題が一つ残っている。

12.6 t 分布

正規母集団からの標本に基づく区間推定を行う際，標本平均 $\bar{X}$ が正規分布 $N(m, \sigma^2/n)$ に従い，さらにそれを標準化（平均を引いて標準偏差で割る）して

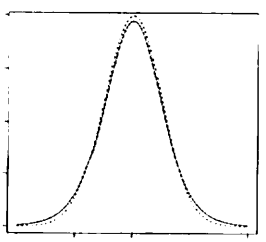
$$Z=\frac{\bar{X}-m}{\sigma/\sqrt{n}}$$

に変換したとき，この Z が標準正規分布 $N(0,1)$ に従うという事実が重要であった。このことから，母平均の 95%信頼区間が「標本平均 ± 1.96 母標準偏差」であるという結果が導かれた。

ここで，母標準偏差を，標本から求めた不偏分散の正の平方根で置き換えたとき，

$$t=\frac{\bar{X}-m}{\hat{\sigma}/\sqrt{n}}$$

はもはや正規分布に従うとは期待できない。分母にある $\hat{\sigma}$ 自体が確率変数であり，定数ではないため，ここが定数である母分散が既知の場合とは状況が異なるのである。上式 t は，**自由度 $n-1$ の t 分布**と呼ばれる確率分布に従うことが知られている。t 分布の確率密度関数のグラフは，正規分布の確率密度関数のグラフと似ているが，わずかに異なる。右図では，実線が t 分布，点線が正規分布を表している。



t 分布の確率密度関数を具体的に書くことはしない。そこには高校では扱わないガンマ関数というものが現れる。

進むにつれて，使える数学的道具立てが高校の段階ではまだまだ足りず，いろいろな事柄を認めながら進まなければならなくなってきた。今はおよその話の流れを掴み取ってほしい。

12.7　正規母集団の母平均の区間推定（母分散が未知の場合）

母集団が正規分布に従う場合について，母平均，母分散とも未知の場合に，母平均の区間推定をしたい。信頼度95%の信頼区間を求める方法にたどり着くまでの流れをまとめると以下のようになる。

(1) 母分散 σ^2 が既知の場合には，標本平均 $\bar{X}$ を標準化した $Z=\dfrac{\bar{X}-m}{\sigma^2/\sqrt{n}}$ は標準正規分布 $N(0,1)$ に従った。

(2) 母分散 σ^2 が既知の場合には，母平均の95%信頼区間は $\left[\bar{X}-1.96\dfrac{\sigma}{\sqrt{n}},\bar{X}+1.96\dfrac{\sigma}{\sqrt{n}}\right]$ だった。これは，標準正規分布 $N(0,1)$ においては区間 $[-1.96,1.96]$ の中に全体の95%の値が入ることに基づいていた。

(3) 母分散 σ^2 が未知の場合には，母分散の不偏推定量 $\hat{\sigma}^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ を用いる。$n-1$ で割っていることに注意。

(4) このとき，母分散が既知の場合の標準化に用いた式で，母標準偏差 σ をその推定量 $\hat{\sigma}$ で取り替えたもの $\dfrac{\bar{X}-m}{\hat{\sigma}^2/\sqrt{n}}$ は自由度 $n-1$ の t 分布に従う。母分散が既知の場合には標準化後の分布は標本の大きさに依らなかったが，母分散が未知の場合には，変換後の分布もなお標本の大きさに依存することに注意。

(5) 自由度 $n-1$ の t 分布においても，区間 $[-T,T]$ の中に全体の95%の値が入るような定数 T が定まる。その値 T を，$t_{0.025}(n-1)$ と書き表す。添え字の0.025は，この値 $t_{0.025}(n-1)$ より大きな値をとる確率が2.5%であることを表しており，また，カッコ内の $n-1$ は，t 分布の自由度を表している。t 分布は平均を中心とする対称な分布であり，値 $t_{0.025}(n-1)$ より大きな値をとる確率が2.5%のとき，値 $-t_{0.025}(n-1)$ より小さな値をとる確率も2.5%であり，これら両端を切り落とすと，区間 $[-t_{0.025}(n-1),t_{0.025}(n-1)]$ に値が入る確率が0.95となる。

(6) この数値を用いれば，あとは母分散が既知の場合と同様に，95%信頼区間は $\left[\bar{X}-t_{0.025}(n-1)\dfrac{\hat{\sigma}}{\sqrt{n}},\bar{X}+t_{0.025}(n-1)\dfrac{\hat{\sigma}}{\sqrt{n}}\right]$ となることが導ける。

12.8　母集団比率の区間推定

母平均の点推定，母分散の点推定，母平均の区間推定について扱ってきた。平均，分散以外の母数の推定として，母比率の推定を取り上げる。

母比率とは，母集団の中にある条件を満たすものと満たさないものとの 2 種類があり場合に，条件を満たすものの全体に対する比率を指す。たとえば，ある選挙で投票した有権者全体の中で，候補者 A に投票した者は何パーセントになるかを一部の有権者に対するアンケート調査から推定する場合などは，母比率の推定に当たる。また，ある工場で生産する製品のなかで，不良品が何パーセントあるかを抽出した標本の中の不良品の割合から推定する場合も母比率の推定にあたる。

この状況のモデルとして，袋の中に赤玉と白玉が入っている状態を考える。赤玉を「条件を満たすもの」と考える。母比率 $p = \dfrac{\text{赤玉の個数}}{\text{玉の総数}}$ である。ここから無作為に玉を一つ取り出して色を調べて記録し，玉を戻す。これを n 回繰り返すことにより，大きさ n の標本を得る。

標本における赤玉の比率を，母比率の点推定量として用いることができる。

$$\hat{p} = \frac{\text{標本のうち赤が出た回数}}{n}$$

赤が出た回数は二項分布 $B(n,p)$ に従うが，n が十分大きいときには二項分布 $N(np, np(1-p))$ で近似することができた。この回数を n で割った値，すなわち母比率の推定量 $\hat{p}$ は近似的に二項分布 $N\left(p, \dfrac{p(1-p)}{n}\right)$ に従う。よって，$E(\hat{p}) = p$ であり，これは推定量 $\hat{p}$ が不偏推定量であることを示している。

標本比率が正規分布に従うのであれば，それを用いた区間推定を行うのはもうおなじみの手順である。母比率 p の 95%信頼区間は，標本比率から標準偏差 $\sqrt{\dfrac{p(1-p)}{n}}$ の ± 1.96 倍の範囲内に母比率が入る確率が 0.95 であることからわかる。

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leqq p \leqq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \fallingdotseq 0.95$$

このままでは困る。母比率を推定したいのに，その信頼区間の限界に母比率自体を使ったのではどうどうめぐりで計算のしようがない。そこで，信頼限界（区間の両端）に現れる p を，その点推定量 $\hat{p}$ で取り替えることにより，次の信頼区間を得る。（すべて近似の話である）

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqq p \leqq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \fallingdotseq 0.95$$

定理 12.7. 標本の大きさ n が十分大きいとき，母比率 p に対する 95%信頼区間は，標本比率 $\hat{p}$ を用いて次のように求められる。

$$\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

以上で，推定の話をひとまず終わる。

13　仮説検定

13.1　仮説検定の考え方

1枚の硬貨があるとする。表の出る確率と，裏の出る確率は同じだろうと思って10回投げてみると，7回表が出た。A君はこの結果を見て,「この硬貨は表のほうが出やすいに違いない」と主張する。それに対してB君は,「表と裏の出る確率が等しい場合でも，10回投げたときいつも表と裏がそれぞれ5回出るわけではない。表裏の出方にはばらつきがあり，表が多く出ることもあれば，裏が多く出ることもある。今回，たまたま表が多めに出ただけで，これだけでこの硬貨が表の出やすい硬貨であるとはいえない」と主張して，意見が一致しない。A君の主張が妥当かどうかを判定する一つの方法が,「仮説検定」という方法である。

硬貨の例について考える。「硬貨の表のでる確率が0.5であるという仮定のもとに，10回中7回表がでる確率を計算する。もしこの確率がきわめて小さいならば，表のでる確率が0.5という仮定は間違っていると考えられる。もしこの確率がきわめて小さいとはいえないならば，表のでる確率が0.5という仮定は必ずしも間違っているとはいえない。」これが仮説検定の考え方である。

A君としては,「表のほうが出やすい」と主張したい。そのために，あえて「表でる確率が0.5である」と仮定するところからはじめる。この仮定を，帰無仮説という。以下，H_0 と表す。

帰無仮説 H_0 : 表のでる確率が0.5に等しい

これに対して，A君が本当に主張したいことを表した，帰無仮説とは対立する仮説を対立仮説という。以下，H_1 と表す。今の場合は，

対立仮説 H_1 : 表の出る確率は0.5よりも大きい

この帰無仮説のもとに,「10回中7回以上表が出る」という事象が起こる確率を求める。

$$ {}_{10}\mathrm{C}_7(0.5)^{10} + {}_{10}\mathrm{C}_8(0.5)^{10} + {}_{10}\mathrm{C}_9(0.5)^{10} + {}_{10}\mathrm{C}_{10}(0.5)^{10} = 0.171875 $$

すなわち，約17パーセントである。何パーセント以下のときに,「きわめて小さい」と考えるのか，という基準は，あらかじめ定めておかなければならない。よく利用されるのは，5%，1%である。これらの値を，有意水準という。今，有意水準を5%にとったとしよう。このとき，上の17%という値は「きわめて小さい」とはいえない。帰無仮説 H_0 のもとで「10回中7回表」という事象は起こっても不思議はない。そこで,「この硬貨は表のほうが出やすい」とは必ずしもいえない，と判断する。

それに対して，もしも「10回中9回表が出た」とする。この場合はどうか。上と同じ帰無仮説 H_0 のもとで,「10回中9回以上表が出る」という事象が起こる確率を求める。

$$ {}_{10}\mathrm{C}_9(0.5)^{10} + {}_{10}\mathrm{C}_{10}(0.5)^{10} = 0.0107421875 $$

約1パーセントとちょっとである。有意水準を5%にとったとき，この数字は「きわめて小さい」といえる。めったに起こらないことが起こったと考えるよりも，むしろ，最初に仮定した帰無仮説が

間違っていたと判断したほうが無理がないと考え，帰無仮説は成り立たない（このサイコロは表裏が出る確率が同じとはいえない）と判断する。これを，**有意水準 5%で帰無仮説が棄却されたという。**帰無仮説が棄却されたことが，A 君にとっては，思ったとおりの結論が出たことになっている。帰無仮説が棄却されたとき，**対立仮説が採択されたともいう。**

この考え方は，背理法による証明の考え方と似ている。

問題 13.1. A 君はこれまでの数学のテストでは 3 題中 2 題解けるという状態が続いていた。ところが，今回のテストではこれまでと同程度の問題に対して 8 題中 7 題を解くことができた。このことから，A 君の実力が上がったと判断してよいか。また，続けて次回のテストでも，今回と同等以上の割合で問題が解けたとしたらどうか。有意水準 5%で検定せよ。

解答

①帰無仮説と対立仮説の設定

帰無仮説を，H_0：「A 君が問題に正解する確率は 2/3 である」とし，

対立仮説を，H_1：「A 君が問題に正解する確率は 2/3 より大きい」とする。

②帰無仮説のもとで，8 題中 7 題 以上 正解する確率を計算し，有意水準と比べる

③問題前半部分に対する結論

④帰無仮説のもとで，2 回続けて，8 題中 7 題以上正解する確率を計算し，有意水準と比べる

⑤問題後半部分に対する結論

13.2　母平均の検定

例 13.1. ある工場で生産している長さ 7mm の規格のボルトについて，工場の製作機器が正常に動いている場合，過去のデータから製品の長さは平均 7mm，標準偏差 0.35mm の正規分布に従うことがわかっているとする。ある日，その日に製作したボルトから無作為に 100 本を取り出して長さを調べたところ，平均値が 7.06mm であった。この日，工場の製作機器の作り出すボルトの長さの平均値が規格からはずれているといえるか。有意水準 5%で検定せよ。

① 帰無仮説と対立仮説の設定

帰無仮説 H_0 と，対立仮説 H_1 と以下のようにとる。

帰無仮説 H_0 :「この日製作したボルトの母平均は 7mm である」

対立仮説 H_1 :「この日製作したボルトの母平均は 7mm ではない」

前項の問題（実力が上がったか）では，正解率が帰無仮説で設定した値よりも有意に上がったかどうかに関心があり，現に実施したテストでの正解率は平素より高かったので，対立仮説として「正解率が 2/3 より高い」という仮説を設定した。2/3 より低い可能性は考慮に入れていない。それに対して，この例題では，製品が規格どおりに作れているかいないかに関心があり，規格よりも長いほうに離れていても，短いほうに離れていても不正確であることには変わりはない。「規格どおりに作れていないのではないか」という疑問を検定にかけようとしているので，対立仮説では，「7mm より長い」ではなく，「7mm ではない」としている。一般に，「ある母数 θ が特定の値 a に等しい」という帰無仮説に対して，「母数 θ が値 a より大きい（または，小さい）」という対立仮説を置いて行う検定を，**片側検定**といい，「母数 θ が値 a とは異なる」という対立仮説を置いて行う検定を**両側検定**という。前項の問題は片側検定であり，ここでの例は両側検定である。

② 帰無仮説のもとで，ボルトの長さが平均から 0.06mm 以上離れる確率と有意水準を比べる

「ボルトの長さが 7.06mm 以上になる確率を求める」とせずに，「ボルトの長さが平均から 0.06mm 以上離れる確率を求める」とした。ここに，片側検定と両側検定との違いが現れている。

さて，帰無仮説を仮定すると，母集団は正規分布 $N(7, 0.35)$ に従う。この母集団から取り出した大きさ 100 の標本の平均値 $\bar{X}$ がどのような分布に従うかは，12 節に現れた次の定理によりわかる。

> 母集団分布が正規分布 $N(m, \sigma^2)$ に従うとする。ここから大きさ n の標本 $\{X_1, X_2, X_3, \cdots, X_n\}$ をとり，標本平均を $\bar{X}$ で表す。このとき，確率変数 $\bar{X}$ は正規分布 $N(m, \sigma^2/n)$ に従う。

今の場合，大きさ 100 の標本の平均値 $\bar{X}$ は，平均 7mm，標準偏差 $0.35/10 = 0.035$ mm の正規分布 $N\left(7, \dfrac{0.35^2}{100}\right)$ に従う。このような分布に従う確率変数 $\bar{X}$ の値が，平均 7mm から 0.06mm 以上ずれる確率 $P(|\bar{X} - 7| \geqq 0.06)$ が，5%よりも大きいか小さいかを知りたい。

$$P(|\bar{X} - 7| \geqq 0.06) \geqq 0.05 \text{ か？}$$

ここで，確率変数 $\bar{X}$ を標準化した確率変数を Z とする。この手順はすでに何度も利用した。「平均を引き，標準偏差で割る」という操作をすることによって，平均 0，標準偏差 1 の正規分布に従う確率変数 Z に変換する。

$$Z = \frac{\bar{X} - 7}{0.035}$$

このとき，

$$|\bar{X} - 7| \geqq 0.06 \iff |Z| \geqq 1.71$$

であるから，「$P(|\bar{X} - 7| \geqq 0.06) \geqq 0.05$ か？」を知るには，

$$P(|Z| \geqq 1.71) \geqq 0.05 \text{ か？}$$

を知ればよい。ところで，標準正規分布に従う確率変数 Z については，

$$P(|Z| \leqq 1.96) = 0.95, \text{ すなわち, } P(|Z| \geqq 1.96) = 0.05$$

だった（正規分布においては，平均からのずれが標準偏差の 1.96 倍の範囲内にデータ約 95%が含まれる，という以前繰り返し取り上げた事実）ので，

$$P(|Z| \geqq 1.71) \geqq P(|Z| \geqq 1.96) = 0.05$$

であることがわかる。すなわち，ボルトの長さが平均から 0.06mm 以上離れる確率は，有意水準 5%よりも大きい。

③**結論** 帰無仮説のもとで，無作為抽出したボルトの平均値が 7mm から 0.06mm 以上ずれる確率は有意水準 5%よりも大きいので，帰無仮説は棄却できない。すなわち，この有意水準のもとで，製作されたボルトの長さの平均値が規格からずれているとはかならずしもいえない。

【補足 1】正規分布表を用いて求めると，$P(|Z| \geqq 1.71) = 0.087$ である。もし，2 日続けて上と同じ結果が出たとしたら，帰無仮説のもとでそのようなことが起こる確率は $0.087^2 = 0.0076$ であり，1%にも満たない。すなわち，有意水準を 5%にとった場合でも，1%にとった場合でも，仮説は棄却され，工場の製作機械が不調であると判断される。ある 1 日の結果で検定を行い製作機械は不調とはいえないという結果を得たからといって，翌日それと同じ現象がくりかえし起きても「昨日調べたようにこの程度のずれなら製作機械は不調とはいえない」などと考えてはいけない。続けて複数回起きるということは，ただ 1 回起きるということとはまったく異なる事態を表している。

【補足 2】上の問題では，母標準偏差が 0.35 とわかっているという前提で話を進めた。母標準偏差がわかっていない場合には，確率変数の標準化の際に，母標準偏差 σ の代わりにその推定値 $\hat{\sigma}$ を使うこととなる。その場合，変換後の確率変数 Z は，正規分布ではなく t 分布という確率分布に従うことは，以前一度触れた。ここではこれ以上立ち入らず，検定のごく基本的な考え方をつかむに止めておく。

参考文献

[1] 東京大学教養学部統計学教室編『統計学入門』(東京大学出版会，1991)

[2] 東北大学統計グループ『これだけは知っておこう! 統計学』(有斐閣ブックス，2002)

[3] 竹村彰通『統計　第2版』(共立講座21世紀の数学14，2007)

[4] C.R. ラオ『統計学とは何か　偶然を生かす』(丸善，1993)

[5] 鄭躍軍，金明哲，村上征勝『文化情報学ライブラリ　データサイエンス入門』(勉誠出版，2007)

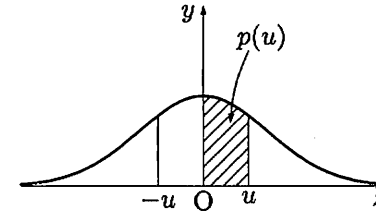
[6] 小島寛之『完全独習　統計学入門』(ダイヤモンド社，2006)

[7] デイヴィッド・ムーア，ショージ・マッケイブ『実データで学ぶ，使うための統計入門』(日本評論社，2008)

[8] 金明哲『Rによるデータサイエンス』(森北出版，2007)

[9] 山田剛史，杉澤武俊，村井潤一郎『Rによるやさしい統計学』(オーム社，2008)

正規分布表



u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0..0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0..4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49897	0.49900

あとがき

大手前高校では，2学年前期にSSH科目「理想（のぞみ）」を設置しています。1学年後期のSSH科目「信念（まこと）」がレポート作成能力やプレゼンテーション能力など，科学的コミュニケーション能力を育むことを目的としているのに対し，「理想（のぞみ）」では統計やデータ分析など，将来科学的方法論を身に着けていくための土台作りを行うこと目的としています。

このテキストは，SSH指定初年度の入学生である63期生に初めて「理想（のぞみ）」の授業を行うにあたり作成した教材をまとめたものです。現在の教育課程では，生徒は，中学校においても統計をほとんど学習することなく高校へ入学してきています。そこで，初年度の取り組みでは1学年後期に土曜学習活動日などを利用して本テキスト前半「統計入門」の授業を行い，その準備の下に，2学年前期に後半「続・統計入門」の授業を行うという方法をとりました。

テキスト作成にあたっては，グラフや図を用いできるだけ視覚的に捉えられるように心がけました。また，ある統計的手法を説明する際にもその意味が理解できるよう，説明の仕方に工夫を加えました。最後はなんとか推定・検定の考え方までたどり着きました。

教材作成にあたり，数回にわたって大阪府立大学の林利治先生にご助言を頂きました。また，林先生には1学年後期の初めに，統計学習への動機付けを行う意味で生徒への特別講義を行っていただきました。これらのご指導・ご助言に感謝いたします。

新しい学習指導要領では，中学校にも「資料の活用」など統計的内容が置かれ，高等学校でも必履修科目の中に「データの分析」という単元が置かれています。改善の余地の多い本テキストですが，今後の統計教育実践の参考になる部分があれば幸いです。

大阪府立大手前高校数学科
SSH科目「理想（のぞみ）」

のぞみ テキスト　統計入門／続・統計入門

平成22年（2010年）11月22日　初版第1刷

発行　大阪府立大手前高等学校
大阪市中央区大手前2-1-11
電話　06(6941)0051
FAX　06(6941)3163