第1章　記述統計から推測統計へ

Lambert Adolphe Jacques Quetelet (1796–1874)

1.1　1変量データの記述

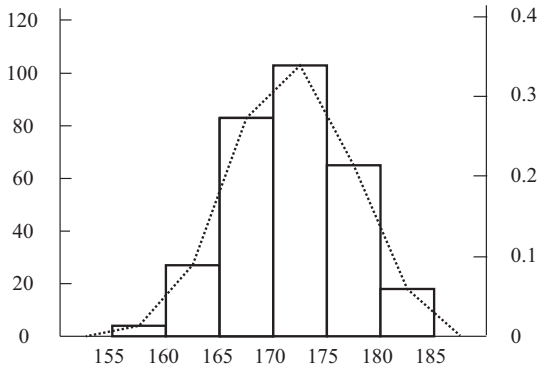
1変量データ (1次元データ) 数列 $x_1, x_2, \ldots, x_n$　(n をデータの大きさという)

度数分布表

x	a_1	$\cdots$	a_i	$\cdots$	a_m	合計
度数 f	f_1	$\cdots$	f_i	$\cdots$	f_m	n

例 題 1.1 (度数分布表・ヒストグラム・度数折れ線 (度数多角形))

階 級	155 −160	160 −165	165 −170	170 −175	175 −180	180 −185	合計
階級値 x	157.5	162.5	167.5	172.5	177.5	182.5	
度 数 f	4	27	83	103	65	18	300
相対度数	0.013	0.090	0.277	0.343	0.217	0.060	1.000



代表値 観測値 $x_1, x_2, \ldots, x_n$ を1つの値で代表させる.

- **mean or average (平均値):** 相乗平均・調和平均など別の定義もいろいろあるので, はっきり区別したいときは**算術平均**と呼ぶ.

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

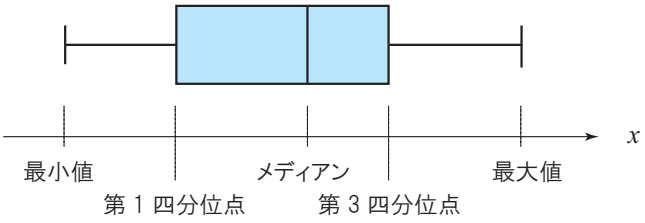
あるいは, 度数分布表から,

$$\bar{x} = \frac{1}{n}\sum_i a_i f_i = \sum_i a_i \frac{f_i}{n}.$$

- **median (中央値):** 観測値 $x_1, x_2, \ldots, x_n$ を大きさの順に並べたとき, 順位がちょうど真ん中にある量.
- **mode (最頻値):** 観測値 $x_1, x_2, \ldots, x_n$ の中に同じ値が重複して現れる場合, 現れる度数が最も多い観測値. 観測値を度数分布表にまとめたとき, (相対) 度数が最も大きくなる階級の階級値もモードという. モードは 2 つ以上あることもある.

分布のばらつき データのばらつき, 広がり具合を数値化する:

- **box plot (箱ひげ図):**



- **variance (分散):**

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{x}^2$$

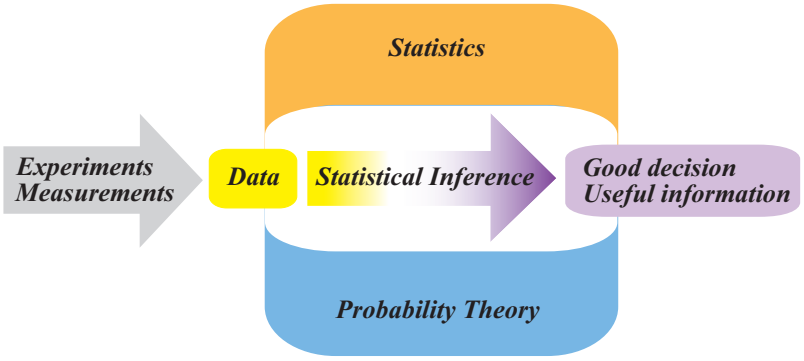
- **standard deviation (標準偏差):** 分散の正の平方根

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

変量 x を明示したいときは, σ_x^2, σ_x のように書く. 度数分布表を用いれば,

$$\sigma^2 = \frac{1}{n}\sum_i (a_i - \bar{x})^2 f_i = \sum_i (a_i - \bar{x})^2 \frac{f_i}{n} = \sum_i a_i^2 \frac{f_i}{n} - \bar{x}^2$$

1.2 Inferential Statistics



1.3 What is a Random Variable?

統計の対象として観測される量は**確率変数 (random variable)** としてモデル化される. 習慣によって, 確率変数には $X, Y, Z, T, \ldots$ のように大文字を用いる.

- **Discrete random variables (離散型確率変数)**
  (1) コインを 3 回投げるとき表の出る回数.
  (2) 授業開始時の出席者数.
- **Continuous random variables (連続型確率変数)**
  (1) 円の内部から 1 点をランダムに選んだとき, その点と中心との距離.
  (2) 新生児の体重.

確率変数とその実現値 確率変数 X は特定の数を表すのではない. その取りうる個別の値を X の実現値という. 確率変数 X が統計の対象として観測する量であるのなら, 実現値とは観測された一つの値のことである.

1.4 Distributions of Discrete Random Variables

例 題 1.2 コインを 3 回投げて, 表の出る回数を X とする. X は $\{0, 1, 2, 3\}$ の範囲を動く確率変数である. このとき,

$$P(X=0)=\frac{1}{8}, \quad P(X=1)=\frac{3}{8}, \quad P(X=2)=\frac{3}{8}, \quad P(X=3)=\frac{1}{8},$$

が成り立つ. 各値を取る確率を一覧表にしてもよい.

x	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

このように X の取りうる値それぞれに対して, それをとる確率を与えることで, X の値の出やすさの確率的な傾向 (確率分布) はすっかり明らかにされる.

離散型確率変数の分布 離散型確率変数 X の取りうる値を網羅して $\{a_1, \ldots, a_i, \ldots\}$ とする. 各値を取る確率を一覧表にしたものを確率分布という.

x	a_1	$\cdots$	a_i	$\cdots$
$P(X=x)$	p_1	$\cdots$	p_i	$\cdots$

あるいは, $P(X = a_i)$ の一般式を書くことができれば (たとえば, 二項分布など), 一覧表を書かなくても確率分布がわかる. $p_i = P(X = a_i)$ とおくと,

$$p_i \geq 0, \qquad \sum_i p_i = 1$$

が成り立つ. ($p_i = 0$ となる a_i は除外してよいが, $p_i = 0$ を許しておく方が便利.)

離散型確率変数の平均値と分散

$$\mathbf{E}[X] = m_X = \sum_i a_i p_i = \sum_i a_i P(X = a_i),$$
$$\mathbf{V}[X] = \sigma_X^2 = \sum_i (a_i - m_X)^2 p_i = \sum_i a_i^2 p_i - m_X^2.$$

分散については, 次のように書くと便利 (連続型にも通用する).

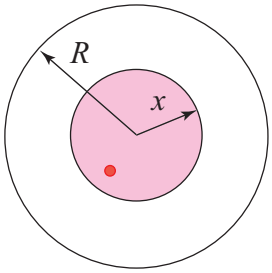
$$\mathbf{V}[X] = \mathbf{E}[(X - m_X)^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

例 1.2 (続)

$$\mathbf{E}[X] = \frac{3}{2}, \qquad \mathbf{V}[X] = \frac{3}{4}.$$

1.5 Distributions of Continuous Random Variables

例 題 1.3 半径 R の円の内部から 1 点をランダムに選んだとき, その点と中心との距離を X とする. X は $[0, R]$ に値をとる連続型確率変数になる. 特定の実数 a に対して $X = a$ となる確率は $P(X = a) = 0$ であるから, 離散型のように確率分布を与えることはできない.



分布関数 $F(x) = P(X \leq x)$ を考える. $x < 0$ のとき $F(x) = 0$, $x > R$ のとき $F(x) = 1$ は明らか. そこで, $0 \leq x \leq R$ とする. $X \leq x$ はランダムに選んだ 1 点と中心 O との距離が x 以下となることを意味するが, それはランダム点が O を中心とする半径 x の円板から選ばれたことを意味する. ランダムに 1 点を選ぶという行為から, 円の面積比を考えるのが合理的である.

$$F(x) = P(X \leq x) = \frac{\pi x^2}{\pi R^2} = \frac{x^2}{R^2}.$$

分布関数を微分して,

$$f(x) = \begin{cases} \dfrac{2}{R^2}\, x, & 0 \leq x \leq R, \\ 0, & \text{その他.} \end{cases}$$

これを確率変数 X の（確率）密度関数という.

連続型確率変数の分布 連続型確率変数 X の分布は, (確率) 密度関数 $f(x) = f_X(x)$ を用いて与える. 分布関数 $F_X(x) = P(X \le x)$ と密度関数 $f_X(x)$ の関係は,

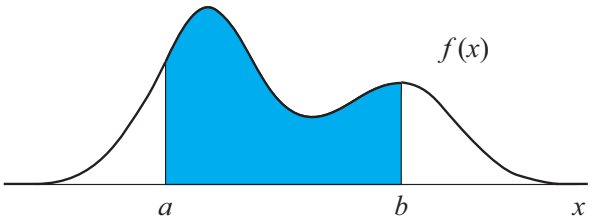
$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)dt \quad \Leftrightarrow \quad \frac{d}{dx} F_X(x) = f_X(x).$$

ここで,

$$f(x) \ge 0, \qquad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

確率を面積で与えることになる:

$$P(a \le X \le b) = \int_a^b f(x)dx, \qquad a < b,$$



連続型確率変数の平均値と分散

$$\mathbf{E}[X] = m_X = \int_{-\infty}^{+\infty} xf(x)dx,$$
$$\mathbf{V}[X] = \sigma_X^2 = \int_{-\infty}^{+\infty} (x - m_X)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - m_X^2.$$

離散型と同様に,

$$\mathbf{V}[X] = \mathbf{E}[(X - m_X)^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

例 1.3 (続)

$$\mathbf{E}[X] = \frac{2}{3} R, \qquad \mathbf{V}[X] = \frac{1}{18} R^2.$$

HW 1 サイコロを 2 個投げるとき, 出る目の和 X の確率分布, 平均値, 分散を求めよ.

HW 2 サイコロを 2 個投げるとき, 出る目の大きいほうを L, 小さいほうを S とする. ただし, 同じ目が出たときは $L = S$ とする. L, S それぞれの確率分布, 平均値, 分散を求めよ.

HW 3 長さ L の棒をランダムに折って長いほうの断片の長さを X とする.

(1) X の密度関数を求めよ.

(2) (1) を用いて, 長いほうの断片の長さが短いほうの 2 倍以上になる確率を求めよ.

(3) X の平均値と分散を求めよ.

第2章　基本的な離散分布

Siméon-Denis Poisson (1781–1840)

2.1　Binomial Distribution (二項分布)

表が出る確率が p であるコインを n 回投げたとき, 表の出る回数 X の分布

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\qquad k=0,1,2,\dots,$$

を二項分布といい, $B(n,p)$ で表す. 特に, $B(1,p)$ を成功確率 p のベルヌーイ分布という.

例 題 2.1 $B(4,1/2)$ と $B(4,1/4)$ を図示せよ.

k	0	1	2	3	4
$P(X=k)$	$\frac{1}{2^4}$	$\frac{4}{2^4}$	$\frac{6}{2^4}$	$\frac{4}{2^4}$	$\frac{1}{2^4}$

k	0	1	2	3	4
$P(X=k)$	$\frac{81}{4^4}$	$\frac{108}{4^4}$	$\frac{54}{4^4}$	$\frac{12}{4^4}$	$\frac{1}{4^4}$

定 理 2.2 二項分布 $B(n,p)$ の平均値と分散は

$$m=np,\qquad \sigma^2=np(1-p)$$

確率母関数 $\{0,1,2,\dots\}$ に値をとる確率変数に対して $p_k=P(X=k)$ $(k=0,1,2,\dots)$ とおく. このとき,

$$f(x)=\sum_{k=0}^{\infty}p_kx^k$$

を X のまたは確率分布 $\{p_0,p_1,\dots\}$ の母関数という.

補 題 2.3 確率母関数について次が成り立つ.

(1) $f(0)=p_0,\quad f(1)=1.$

(2) $\mathbf{E}[X]=f'(1).$

(3) $\mathbf{V}[X]=f''(1)+f'(1)-\{f'(1)\}^2.$

2.2 Geometric Distribution (幾何分布)

表が出る確率が p であるコインを投げ続けるとき, 表が初めて出るまでに出た裏の回数 X の分布は

$$P(X=k)=p(1-p)^k, \qquad k=0,1,2,\ldots.$$

この分布をパラメータ p の幾何分布という. (待ち時間の分布として重要)

定 理 2.4 パラメータ p の幾何分布の平均値と分散は

$$m=\frac{1-p}{p}, \qquad \sigma^2=\frac{1-p}{p^2}.$$

2.3 Poisson Distribution (ポアソン分布)

確率変数 X がパラメータ $\lambda>0$ のポアソン分布に従うとは,

$$P(X=k)=\frac{\lambda^k}{k!}\,e^{-\lambda}, \qquad k=0,1,2,\ldots.$$

定 理 2.5 パラメータ λ のポアソン分布の平均値と分散は

$$m=\lambda, \qquad \sigma^2=\lambda.$$

定 理 2.6 (ポアソンの少数の法則) 二項分布 $B(n,p)$ は $np=\lambda$ (これは平均値である) を保ったまま, $n\to\infty$, $p\to 0$ とすると, パラメータ λ のポアソン分布に収束する.

例 題 2.7 (栗ようかんに入っている栗の個数) 1 本当たり 3 個の栗が行き渡るように材料を調整して, 大鍋で栗ようかんを作った. 大鍋を適当にかき混ぜて, 大きな柄杓で 1 本分をすくい取るとき, (1) その 1 本に全く栗が入っていない確率を求めよ. [0.05] (2) 栗が 5 個以上入っている確率を求めよ. [0.18]

HW 4 「ナンバーズ 3 ストレート」では 000〜999 の数を 1 つ指定する. 掛け金は 200 円であり, 当たればしかるべき賞金がもらえる. 週 5 日毎日買ったとして, 当たるまでの平均待ち時間 (週) を求めよ. [199.8 週]

HW 5 メールの着信がまったくランダムに起こるとして, ある 20 分間に全くメールの着信がない確率を求めよ. ただし, メール着信は 1 時間に平均 3 回起こることが経験から知られている.

HW 6 50 名のクラスに 5 月 5 日生まれの学生は何人くらいいるだろうか? 1 年を 365 日, どの日に生まれる確率も同じと仮定すると, 5 月 5 日生まれの学生の人数 X は二項分布 $B(50,1/365)$ に従う. ポアソンの少数の法則を用いて, $P(X=k)$ $(k=0,1,2,3,4)$ を計算せよ. [0.87198, 0.11945, 0.00818, 0.00037, 0.00001; 厳密値は次の通り: 0.87182, 0.11976, 0.00806, 0.00035, 0.00001]

HW 7 X をパラメータ λ のポアソン分布に従う確率変数とする.

(1) $P(X=0) \geq P(X=1)$ となるようなパラメータ λ の範囲を求めよ.

(2) X のモード, つまり $P(X=k)$ が最大になるような k を求めよ.

第3章　基本的な連続分布

Johann Carl Friedrich Gauss (1777–1855)

3.1　Uniform Distribution (一様分布)

区間 $[a,b]$ からどの点も同等な確からしさで 1 点を選ぶときのモデルとして現れる.

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & \text{その他}\end{cases}$$

定 理 3.1 $[a,b]$ 上の一様分布の平均値と分散は,

$$m=\frac{a+b}{2},\qquad \sigma^2=\frac{(b-a)^2}{12}$$

3.2　Exponential Distribution (指数分布)

ランダム到着の待ち時間をモデル化するときに現れる. $\lambda>0$ を定数として

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$$

定 理 3.2 パラメータ λ の指数分布の平均値と分散は,

$$m=\frac{1}{\lambda},\qquad \sigma^2=\frac{1}{\lambda^2}$$

3.3　Normal Distribution (正規分布)

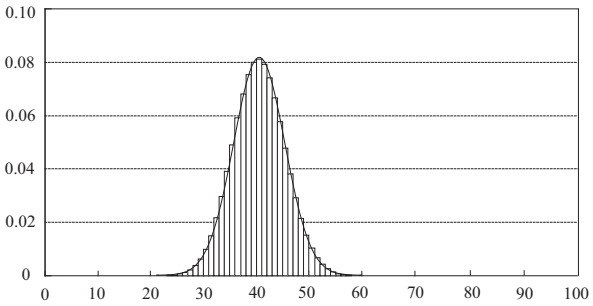
$N(m,\sigma^2)$: 平均 m, 分散 σ^2 の正規分布 (またはガウス分布)

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

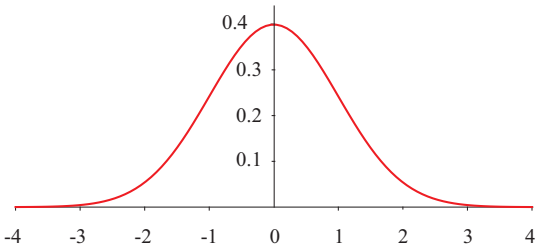
定 理 3.3 (de Moivre–Laplace の定理) 二項分布は, 同じ平均と分散をもつ正規分布で近似できる.

$$B(n,p)\approx N(np,np(1-p)),\qquad 0<p<1,\quad n\to\infty.$$

例 題 3.4 $B(100, 0.4) \approx N(40, 6.197^2)$



3.4 Standard Normal Distribution (標準正規分布) $N(0, 1)$



定 理 3.5 (標準化あるいは規準化) $X \sim N(m, \sigma^2)$ のとき,

$$aX + b \sim N(am + b, a^2\sigma^2), \quad \text{特に}, \quad Z = \frac{X - m}{\sigma} \sim N(0, 1)$$

例 題 3.6 $Z \sim N(0, 1)$ とする.

(1) 次の確率を求めよ. $P(Z \leq 1.15), \quad P(Z \leq -1.23)$ [0.8749, 0.1093]

(2) 次の等式を満たす a を求めよ. $P(Z \geq a) = 0.33, \quad P(Z < a) = 0.75$ [0.44, 0.67]

(3) $X \sim N(2, 5^2)$ のとき, $P(X \leq 0)$ を求めよ.

例 題 3.7 公平なコインを 400 回投げたとき, 表が 225 回以上出る確率を正規分布近似を用いて求めよ (連続補正 (半目補正) に注目).

HW 8 公平なコインを 500 回投げて, 表がちょうど 250 回出る確率を求めよ.

HW 9 (偏差値) 受験者全員の平均点を m, 標準偏差を σ とするとき,

$$(\text{偏差値}) = 50 + 10 \times \frac{x - m}{\sigma}$$

受験者数が多数の時, 得点の分布は正規分布に近いと想定されることが多い. 偏差値は, 20 以下にも 80 以上にもなり得るが, そのような極端な値の出る確率を求めよ.

HW 10 ある大学では過去のデータによると, 入学試験の合格者のうち 4%が入学を避退するという. 1000 人の定員のところ 1050 人を合格にするとき, 定員割れを起こす確率を求めよ. [0.0901]

標準正規分布表 $\quad I(z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_0^z e^{-x^2/2} dx$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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	0.4773	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	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

1–3 章 演習問題（期末試験対策）

演習問題 1 地図帳で目的地を探すとき, 目的地がいつもページの端によっている気がする. 30cm×40cm の地図で, 目的地が周辺 5cm の範囲に見つかる確率を求めよ. [1/2]

演習問題 2 ある 2 人は午後 0 時から午後 0 時 50 分の間に公園に到着し, そこで 10 分間だけ休憩するのが日課である. ただし, 公園に到着する時刻はお互いにランダムであるとする. この 2 人が公園で遭遇する確率を求めよ. どのように確率を定義するか, 明確に述べて答えよ. [9/25]

演習問題 3 3 辺の長さが $3, 4, 5$ の直角三角形の内部に 1 点 P をランダムに選ぶとき, P と斜辺 (長さ 5 の辺) との距離が 1 以下になる確率を求めよ. [95/144]

演習問題 4 (離散型または連続型) 確率変数 X に対して, 分布関数が $F(x) = F_X(x) = P(X \leq x)$ で定義される. ここで, x はすべての実数を走る. コイン 3 個を同時に投げるとき, 表の枚数を X とする. X の分布関数を求め, そのグラフを示せ.

演習問題 5 長さ L の棒をランダムに折ってできる短いほうの断片の長さを Y とする. 確率変数 Y の分布関数, 密度関数, 平均値, 分散を求めよ. [$F_Y(x) = 0\,(x < 0); = 2x/L\,(0 \leq x \leq L/2); = 1\,(x > L/2)$. $f_Y(x) = 2/L\,(0 \leq x \leq L/2); = 0\,(\text{otherwise})$. $\mathbf{E}[Y] = L/4$. $\mathbf{V}[Y] = L^2/48$.]

演習問題 6 半径 R の円の内部から 1 点をランダムに選び, その点と円周までの最短距離を X とする. X の平均値と分散を求めよ. [$\mathbf{E}[X] = R/3$. $\mathbf{V}[X] = R^2/18$.]

演習問題 7 中心を O とする半径 R の円の内部にランダムに 1 点を選び, その点を通る中心を O とする円の面積を X とする. X の分布関数, 密度関数, 平均, 分散を求めよ.

演習問題 8 (マメ知識：ポアソン分布では, 偶数のほうが出やすい) バス停に並んでいる客の人数がパラメータ λ のポアソン分布に従うとする. その人数が偶数である確率と奇数である確率とではどちらが大きいか？[指数関数のテーラー展開を思い出すとよい.]

演習問題 9 $N \geq 4$ を自然数とする. 1 番から N 番まで通し番号のついた N 枚のカードから, 同時に 4 枚のカードを抜き取り, その中の最大の番号を X とする. $4 \leq k \leq N$ に対して, 確率 $P(X = k)$ を求めて, 平均値 $\mathbf{E}[X]$ を計算せよ. [$4(N+1)/5$]

演習問題 10 (1) $X \sim N(20, 4^2)$ に対して, $P(X > 17.8)$ を求めよ. [0.7088]
(2) $X \sim N(50, 10^2)$ のとき, $P(X > a) = 0.985$ を満たす a を求めよ. [28.3]

演習問題 11 サイコロを 60 回投げるとき, 1 の目が 12 回以上出る確率を求めよ. 次に, サイコロを 600 回投げるとき, 1 の目が 120 回以上出る確率を求め, 先の答えと比較せよ. [二項分布の正規分布近似を用いよ.]

演習問題 12 大規模な選抜試験が実施され, 上位 5%が合格となる. 試験の結果, 平均点は 68 点, 標準偏差が 8 点であった. 受験者全体の得点分布は正規分布であると仮定できるとして, 合格するための最低点を求めよ. [81.12 点あるいは 82 点]

第4章　母数の推定 I

— 二項母集団の母比率

Jacob Bernoulli (1654–1705)

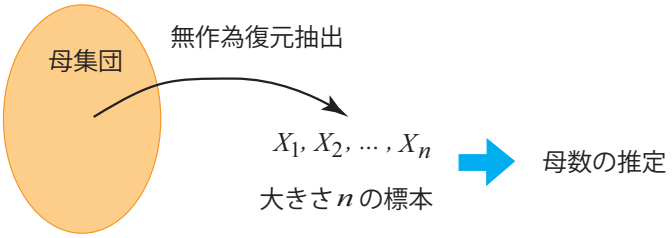
4.1　Sampling (標本抽出)

調査対象の集団 (母集団) に対して, 全数調査が不可能である場合に, その一部分 (標本) を調査して全体の性質を推定することが重要である.

標本を 1 個取り出せば, 観測値 x が 1 個得られる. 観測値は取り出された標本ごとに違った数値となるが, 母集団をよくかき混ぜて無作為に標本を選ぶのなら, 観測値 x の現れ方に母集団分布が反映する. そこで, 母集団分布に従う確率変数を X として, 観測値 x を X の実現値とみなすことができる.

Random Sampling with Replacement (無作為復元抽出) 母集団から 1 個の標本を無作為に取り出して得られる値は, 母集団分布に従う確率変数である. 取り出した標本を元に戻して, 同じ操作で次々に標本を取り出すことにすれば, 1 回目の標本 X_1, 2 回目の標本 X_2, ..., n 回目の標本 X_n のように確率変数の列が得られる. このような標本の取り出し方を**無作為復元抽出**といい, $X_1, X_2, \ldots, X_n$ を母集団から得られた n 個の **(無作為) 標本**という.

Estimate of Population Parameters (母数の推定) 母集団分布そのものを標本調査によって推定することは困難な問題であり, 実用上知りたいのは母集団分布を特徴づける統計量やパラメータである. そのような量を母数と総称する. 特に, 母集団分布の平均値を母平均, 分散を母分散と呼ぶ. 母平均や母分散などの基本的な母数の推定がこれからのメインテーマである.



注意 非復元抽出では毎回の標本調査のあと母集団が変化するが, 母集団が巨大なら「非復元抽出 ≈ 復元抽出」と考えてよい. つまり, 母集団が巨大なら n 個の無作為標本を得たいときに, まとめて n 個を取り出しても実用上の誤差は無視してよい.

4.2 Point Estimation

一般に, 標本の関数 $f(X_1, X_2, \ldots, X_n)$ で母数を推定する方式を**点推定 (point estimation)** という. 母平均の点推定として, 標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

が母平均の推定量 (点推定) として適当である. その根拠として次の 2 性質がある.

定 理 4.1 (標本平均の不偏性) $\mathbf{E}[\bar{X}] = m$.

定 理 4.2 (標本平均の一致性) 大きさ n の無作為標本 $\bar{X}$ について,

$$P\left(\lim_{n\to\infty} \bar{X} = m\right) = 1$$

これは次の一般的な結果から従う.

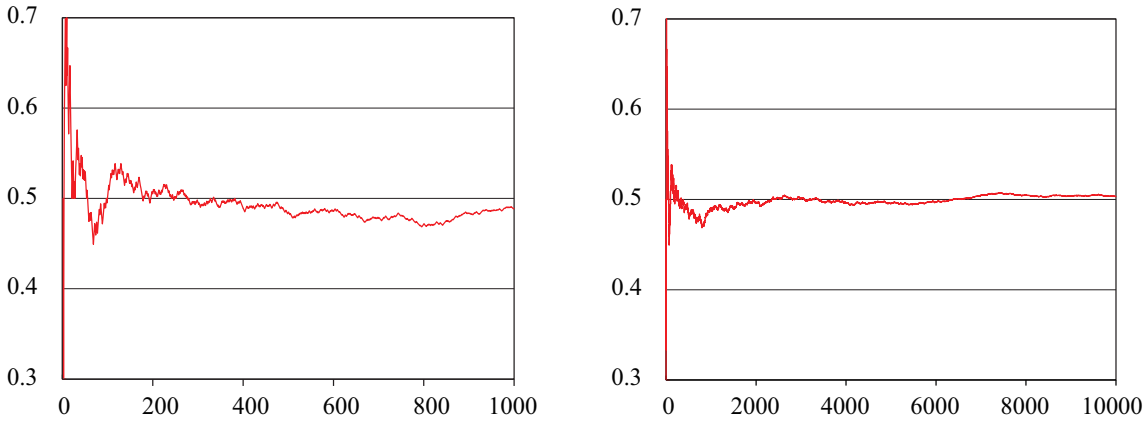
定 理 4.3 (Strong law of large numbers (大数の強法則)) $X_1, X_2, \ldots$ を独立で同分布な確率変数列とし, その平均値を m とする. このとき,

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = m\right) = 1$$

例 題 4.4 (コイン投げのシミュレーション) いつも通り, コイン投げの結果を表なら 1, 裏なら 0 として数値化する. コインを投げ続けて, その結果 $x_1, x_2, \ldots$ に対して

$$t_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

とおく. t_n は初めの n 回のコイン投げで, 表の出た相対頻度である.



4.3 Binomial Population

ある属性 E によって, 2 つの集団に分かれているような母集団を**二項母集団**といい, 属性 E をもつ集団の比率 p を母比率という. 母比率の推定を扱う. まず, 各個体には, 属性 E をもつときは 1, 属性 E をもたないときは 0 の値を与えておく. 母平均 $m = p$ に注意しておく.

大きさ n の標本を $X_1, X_2, \ldots, X_n$ とおく. 各 k に対して,

$$X_k = \begin{cases} 1, & k \text{ 番目の標本が属性 } E \text{ をもつ}, \\ 0, & k \text{ 番目の標本が属性 } E \text{ をもたない}, \end{cases}$$

である. 母平均の点推定には標本平均を用いる. 今の場合, 標本平均と呼ばずに, **標本比率**と呼んで,

$$\hat{p} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

と書く. つまり, 母比率の点推定としては標本比率 $\hat{p}$ を用いる.

例 題 4.5 (Audience Rating Survey (視聴率調査)) テレビ局では視聴率の獲得にしのぎを削っているようである. 果たして, コンマ以下の数字に意味はあるのだろうか?

2016 年 4 月 25 日 (月) 〜 5 月 1 日 (日) ドラマ (関東地区) 視聴率ベスト 10

番組名	放送局	放送日 放送開始時刻 – 分数	視聴率 (%)*
連続テレビ小説・とと姉ちゃん	ＮＨＫ総合	16/04/27(水) 8:00 - 15	24.6
真田丸	ＮＨＫ総合	16/05/01(日) 20:00 - 45	17.0
日曜劇場・９９．９・刑事専門弁護士	ＴＢＳ	16/05/01(日) 21:00 - 54	16.2
世界一難しい恋	日本テレビ	16/04/27(水) 22:00 - 60	13.1
警視庁捜査一課９係	テレビ朝日	16/04/27(水) 21:00 - 54	12.0
土曜ワイド劇場・再捜査刑事・片岡悠介	テレビ朝日	16/04/30(土) 21:00 - 126	11.4
横山秀夫サスペンス・刑事の勲章	ＴＢＳ	16/04/25(月) 21:00 - 114	10.4
トットてれび	ＮＨＫ総合	16/04/30(土) 20:15 - 30	10.1
グッドパートナー無敵の弁護士	テレビ朝日	16/04/28(木) 21:00 - 54	9.9
ラヴソング	フジテレビ	16/04/25(月) 21:00 - 54	9.4
連続テレビ小説・とと姉ちゃん／他	ＮＨＫ総合	16/04/29(金) 12:45 - 15	9.4

* ビデオリサーチ社による番組平均世帯視聴率

日本の放送エリアは全部で 32 ありますが, それぞれの放送エリアごとに視聴率調査が行なわれています. ビデオリサーチでは, 関東地区をはじめ全国 27 地区の調査エリアで, PM システムによる調査とオンラインメータシステムによる調査を実施しています. (日本全国をひとつの調査エリアとした視聴率調査は実施していません) また, 調査対象世帯数は, PM システムによる調査の関東地区・関西地区・名古屋地区で 600 世帯, それ以外のオンラインメータシステムによる調査地区は 200 世帯です. (ビデオリサーチ社のウェッブページから. 2016.5 現在)

参考: 藤平芳紀「視聴率の正しい使い方」(朝日新書)

4.4 Interval Estimation of Binomial Parameter

標本比率 $\hat{p}$ は, 標本の取り方によって変動する (あたりまえ!) ので, 確率変数として扱う. さらに, $\hat{p}$ が母比率 p に丁度一致する確率はゼロに近い. そこで, $\hat{p}$ の変動を評価して, 母比率を信頼度もこめて推定することが重要になる. これを達成するのが**区間推定 (interval estimation)** である.

● **$\hat{p}$ の分布を調べる.**

(1) $\displaystyle\sum_{k=1}^{n} X_k$ は二項分布 $B(n,p)$ に従う.

(2) n が大きいとき, $B(n,p)$ は同じ平均と分散をもつ正規分布 $N(np, np(1-p))$ で近似できる (ドモアブル–ラプラスの定理). 実用上 $pn \geq 5$, $n(1-p) \geq 5$ ならよい.

(3) したがって, n が大きいときは

$$\hat{p} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \Longleftrightarrow \qquad \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

(4) 2 次不等式の近似あるいは大数の法則による議論 (詳細は教科書) によって, 分母の p を $\hat{p}$ で置き換える:

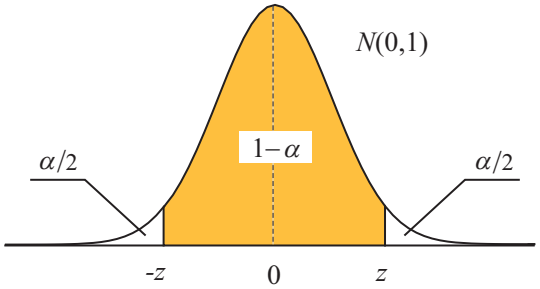
$$\Longleftrightarrow \qquad \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1).$$

両側 α 点 = 片側 $\alpha/2$ 点 与えられた α に対して, $Z \sim N(0,1)$ (標準正規分布) が

$$P\left(-z \leq Z \leq z\right) = 1-\alpha$$

を満たすような z を $N(0,1)$ の**両側 α 点**という.

z	1.00	1.64	1.96	2.00	2.58	3.00	3.29
α	0.317	0.100	0.050	0.045	0.010	0.003	0.001
$1-\alpha$	0.683	0.900	0.950	0.955	0.990	0.997	0.999



● **二項母集団における母比率の区間推定** 母比率 p に対する信頼係数 $1-\alpha$ の信頼区間

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ \right] \qquad \text{または} \qquad \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

区間の端点を信頼限界と呼ぶ. 信頼係数としては

$$90\%\,(\alpha = 0.1, z = 1.64) \qquad 95\%\,(\alpha = 0.05, z = 1.96) \qquad 99\%\,(\alpha = 0.01, z = 2.58)$$

などが習慣的に用いられる.

α	1	大	小	0
信頼係数 $(1-\alpha)$	0%	小	大	100%
信頼区間の幅	0 (点推定)	小 (シャープな推定)	大 (アバウトな推定)	∞

信頼区間の意味 標本調査の結果, 観測値 $x_1 \ldots, x_n$ が得られたとする (二項母集団のときは, $x_k = 0$ または $= 1$). 標本比率 $\hat{p}$ を計算して, 上の公式を用いると信頼区間が得られる. この**信頼区間が母平均を含んでいるか含んでいないかはどちらかであるが, これはわからない.** コイン投げと同じである. 言えることは, 「確率 $1-\alpha$ で信頼区間は母平均を含み, 確率 α で含まない」ということだけである. 「信頼区間の中点が母比率に近い確率が高い」とか「信頼区間の端の方は母比率から外れている確率が高い」などというのは理論を知らないことさらしているだけだが, 世間には意外と多いので注意.

例 題 4.6 (視聴率調査) 標本数 600 から視聴率の推定値 14.1% が得られた. 信頼係数 95% の信頼区間は,

$$0.141 \pm 1.96 \times \sqrt{\frac{0.141(1-0.141)}{600}} \approx 0.141 \pm 0.0278$$

例 題 4.7 視聴率調査において, 信頼係数 95% の信頼区間の長さが 0.01 以下になるためには, どれほどの標本数が必要か? [38416]

HW 11 世論調査により 952 人から回答を得て, 内閣支持率 51% がわかった (NHK 放送文化研究所 2017 年 3 月 10–12 日). 90% 信頼区間を求めよ. $[0.51 \pm 0.027]$

HW 12 世論調査において, 信頼係数 90% の信頼区間の長さが 0.02 以下になるためには, どれほどの標本数が必要か? [6724]

HW 13 視聴率調査において信頼区間を考慮した上で, 順位について考察せよ.

William Sealy Gosset (1876–1937)

第5章　母数の推定 II

— 母平均と母分散の推定

5.1　標本平均の分布

定 理 5.1 (平均値の乗法性と分散の加法性) 独立な確率変数 X, Y に対して,

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y], \qquad \mathbf{V}[X+Y] = \mathbf{V}[X] + \mathbf{V}[Y]$$

定 理 5.2 (標本平均に関する基本定理) 正規母集団 $N(m, \sigma^2)$ から取り出した大きさ n の標本 $X_1, X_2, \ldots, X_n$ の標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

に対して,

$$\bar{X} \sim N\left(m, \frac{\sigma^2}{n}\right) \qquad \Longleftrightarrow \qquad \frac{\bar{X} - m}{\sigma/\sqrt{n}} \sim N(0,1)$$

平均値 m, 分散 σ^2 の一般の母集団でも, n が十分大きいとき, 近似的に成り立つ.

(注意) 大数の法則 $P\Big(\lim_{n\to\infty} \bar{X} = m\Big) = 1$ は上の主張からもわかる.

定 理 5.3 (中心極限定理) $X_1, X_2, \ldots$ を独立で同分布な確率変数列とし, その平均値を $m = 0$, 分散を $\sigma^2 = 1$ とする. このとき,

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k \le x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt.$$

この事実から, n が十分に大きいとき, $\dfrac{1}{\sqrt{n}}\displaystyle\sum_{k=1}^{n} X_k$ は近似的に $N(0,1)$ に従う.

5.2　母平均の区間推定 (母分散が既知)

$X_1, X_2, \ldots, X_n$: 母平均 m (未知), 母分散 σ^2 (既知) をもつ母集団から取り出された標本

● **母平均の区間推定** 母平均 m に対する信頼係数 $1-\alpha$ の信頼区間は,

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}} \qquad z \text{ は } N(0,1) \text{ の両側 } \alpha \text{ 点 (= 上側 } \alpha/2 \text{ 点)} \tag{5.1}$$

● **二項母集団の母比率** 母比率 p に対する信頼係数 $1-\alpha$ の信頼区間は,

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.2}$$

であった. これは, (5.1) の特別な場合とみなすことができる. 二項母集団では母分散は未知ではあるが, 母比率 p を用いて $p(1-p)$ で与えられることはわかっている. (5.2) は, (5.1) において, 母分散 σ^2 を標本比率 $\hat{p}$ を用いて $\sigma^2 = \hat{p}(1-\hat{p})$ と推定した式で置き換えたものである.

例 題 5.4 ある工場のロットから, ランダムに 200 個の標本を選んで不純物量を測定したとき, 平均 2.2 g の不純物が含まれていた. この工場の工程から, 不純物量の標準偏差は 1.5 g であることが経験的に知られている. このロット全体では, 不純物を平均何 g 含んでいるといえるだろうか? 信頼区間を求めよ. [95%信頼区間は 2.2 ± 0.208]

HW 14 ある生産ラインで 1 万個の製品を作った. ランダムに選んだ 40 個の製品の平均重量は 156g であった. この生産ラインの機械的特性から, 生産される製品の重量の標準偏差は 8g である. 生産した 1 万個の製品の平均重量の信頼区間を求めよ. [95% 信頼区間は 156 ± 2.48]

HW 15 HW14 において, 95%信頼区間の幅を 1g 以下にするためには何個の標本をとる必要があるか? [984]

5.3 母平均の区間推定 (母分散未知の場合)

$X_1, X_2, \ldots, X_n$: 母平均 m (未知), 母分散 σ^2 (未知) をもつ母集団から取り出された標本

● **不偏分散と標本分散**

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

前者を**不偏分散**, 後者を**標本分散**という. (文献によっては, 前者も標本分散と呼んでいるので, いささか混乱するので注意せよ) 標本分散は母分散の不変推定量ではない: $\mathbf{E}[S^2] \neq \sigma^2$.

定 理 5.5 不偏分散 U^2 は不偏性を満たす: $\mathbf{E}(U^2) = \sigma^2$.

ただし, 標本数 n が大きくなれば, S^2 と U^2 の差はわずかである.

定 理 5.6 正規母集団 $N(m, \sigma^2)$ から取り出した n 個の標本を $X_1, \ldots, X_n$ に対して,

$$T = \frac{\bar{X} - m}{U/\sqrt{n}} \sim t_{n-1} \quad \text{自由度 } (n-1) \text{ の } t\text{-分布}$$

正規母集団でなくとも, 標本数が大きいときは近似として成り立つ.

自由度 n の t-分布

$$\frac{1}{\sqrt{n}\,B\left(\frac{n}{2},\frac{1}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\,\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \tag{5.3}$$

(1) Γ はガンマ関数.

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt,\qquad x>0.$$

(2) B はベータ関数.

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},\qquad x>0,\ y>0.$$

(3) $N(0,1)$ に比べて, すそ野が厚い.

(4) 自由度 $n=\infty$ の t-分布は標準正規分布 $N(0,1)$ に一致する.

(5) 実用上, $n\geq 30$ で標準正規分布 $N(0,1)$ で代用.

● **母平均の区間推定** 母平均 m に対する信頼係数 $1-\alpha$ の信頼区間は,

$$\bar{X}\pm t\frac{U}{\sqrt{n}}\qquad t \text{ は } t_{n-1} \text{ の両側 } \alpha \text{ 点}$$

例 題 5.7 ある薬品を精製する実験を同一条件下で 8 回行ったところ, 生成物の重量は次のようになった. この方法で得られる生成物の平均重量の 90%信頼区間を求めよ.

32.5 31.8 33.0 32.4 32.2 31.3 32.9 32.1

[$\bar{x}=32.275$, $u^2=0.3135=0.56^2$, $t_7=1.895$ などから 32.275 ± 0.375]

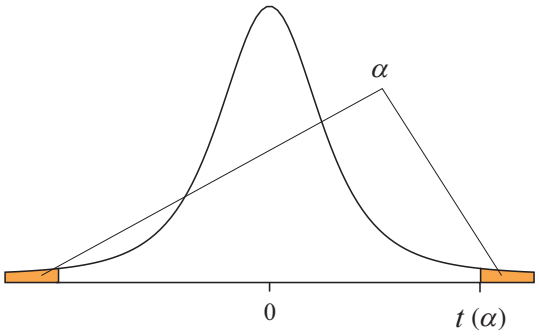
HW 16 ある製品を抜き取り調査してその寿命を測定した結果, 以下の数値を得た. 母集団の平均寿命の 95% 信頼区間を求めよ. [33 ± 4.17]

23 42 33 29 34 41 30 36 34 28

HW 17 (5.3) において $n\to\infty$ を計算して, 自由度 $n=\infty$ の t-分布は標準正規分布 $N(0,1)$ に一致することを示せ. [$\Gamma(1/2)=\sqrt{\pi}$ を用いよ.]

t 分布表 (両側 α 点：$P(|T| \geq t_n(\alpha)) = \alpha$)

$n\backslash\alpha$	0.100	0.050	0.020	0.010
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
∞	1.645	1.960	2.326	2.576



第6章　Testing Hypotheses

Sir Ronald Aylmer Fisher (1890–1962)

6.1　仮説検定の基本

1. 母数に関する帰無仮説 (null hypothesis) H_0 と対立仮説 (alternative hypothesis) H_1 を決める.
2. 関連する確率変数 T (検定統計量) を選び, 仮説 H_0 の下で, この確率変数の分布を調べる.
3. 有意水準 (significance level) $0 < \alpha < 1$ と棄却域 (critical region) を決める.
   - 有意水準とは, H_0 が真なのに誤りであると判定してしまう誤り確率のこと. 慣習では, 10%, 5%, 1% などが用いられる.
   - 棄却域とは, T の実現値として稀と判断される領域で, T がその領域に値をとる確率がちょうど α になる ($P(T \in W) = \alpha$) ように決める. 両側検定か片側検定か (これは H_1 で決まる. 明示すること) によって, 棄却域の取り方が異なる.
4. 標本から T の実現値 t を計算し, W に落ちる ($t \in W$) かどうかを判定する.
   - $t \in W$ のとき. 検定統計量 T の実現値が棄却域に落ちたので, H_0 から想定される揺らぎを超えた稀な値が実現したということ. 実現値は「有意水準 α で有意」であり, 「H_0 を棄却 (reject) し, H_1 を採択 (accept)」する.
   - $t \notin W$ のとき. 実現値 T は棄却域に落ちないので, 「有意水準 α で有意ではない」したがって, 「H_0 を棄却できない (あるいは, 採択する)」となる.

例 題 6.1 コインを 400 回投げたところ, 表が 223 回出た. コインは公正といえるだろうか?

1. このコインで表が出る確率を p とする. 帰無仮説と対立仮説は
$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}$$
2. 400 回投げて表の出る回数を X とする. H_0 のもとで, $X \sim B(400, 1/2) \approx N(200, 10^2)$. 規準化して,
$$Z = \frac{X - 200}{10} \sim N(0, 1)$$
これを検定統計量とする.

3. 有意水準を $\alpha = 0.05$ とする. 棄却域は, 正規分布曲線の両側から合わせて 5% 分を切り取る (両側検定). 両側 5% 点 (= 上側 2.5% 点) は 1.96 なので,

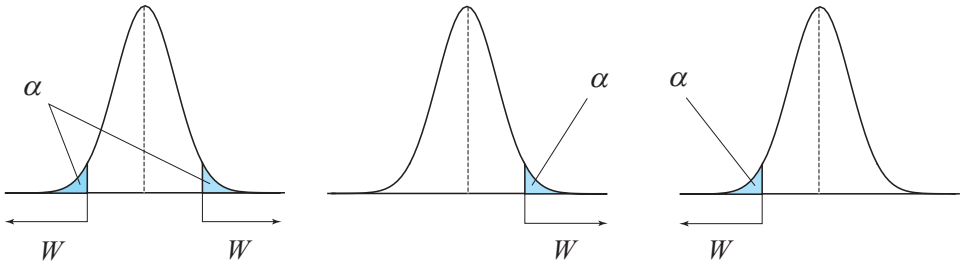
$$W : |z| \geq 1.96$$

4. 実験結果 $x = 223$ から Z の実現値

$$z = \frac{223 - 200}{10} = 2.3$$

が得られる. これは棄却域に落ちるから, H_0 を棄却する. 結論は, 「有意水準 5% の両側検定で H_0 を棄却する」となる. したがって, このコインは公平ではないとの判断に至る.

5. 有意水準 1% では, 両側 1% 点が 2.58 であることより, 実現値 $z = 2.3$ は棄却域に落ちない. 結論は「有意水準 1% の両側検定で H_0 を棄却できない」となる. このことを「高度に有意ではない」ともいう.



$N(0,1)$ の両側 α 点

α	0.317	0.100	0.050	0.045	0.010	0.003	0.001
z	1.00	1.64	1.96	2.00	2.58	3.00	3.29
$1-\alpha$	0.683	0.900	0.950	0.955	0.990	0.997	0.999

6.2　母平均の検定 (母分散既知の場合)

母平均 m, 母分散 σ^2 の母集団から取り出した大きさ n の標本の標本平均について,

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N\left(m, \frac{\sigma^2}{n}\right) \quad \Longleftrightarrow \quad \frac{\bar{X} - m}{\sigma/\sqrt{n}} \sim N(0,1)$$

ただし, $\sim$ は厳密ではなく, n が大きい時に近似的に成り立つ (近似の根拠は中心極限定理による. 正規母集団 $N(m, \sigma^2)$ なら近似は不要).

例 題 6.2 (両側検定) ある機械部品の長さは規格によって 25 mm と定められている. 部品の長さの微小な狂いはやむをえないが, 規格より長すぎても短すぎても困る. ある製造ラインでは, 管理状況から, 部品の長さは標準偏差 0.8 mm の正規分布にしたがっているとしてよい. 16 個のサンプルで実際に長さを調べたところ長さの平均値は 25.45 mm であった. この製造ラインは適正に部品を作っているといえるだろうか? [有意水準 5% の両側検定で $H_0 : m = 25$ を棄却 (実現値 $2.25 \geq 1.96$). 有意水準 1% では棄却されない.]

例 題 6.3 (片側検定) 従来部品の寿命は 120 時間であるが, 新製法では部品の寿命が長くなることが期待される. 実際, 25 個のサンプルで寿命を調べたところ, 平均寿命は 120.8 時間であった. 部品の製造工程の管理状況から, 新製法での部品の寿命は標準偏差 2.2 時間の正規分布にしたがっているとしてよい. 新製法は期待通りであろうか. 仮説検定で判断せよ. [新しい部品の平均寿命を m とおく. 有意水準 5% の片側検定で $H_0 : m = 25$ を棄却 (実現値 $1.82 \geq 1.64$).]

HW 18 (両側検定) コインが公平かどうかを確かめるために, 100 回振ったところ表が 63 回出た. このコインは公平であるといえるか. [有意水準 5% の両側検定で $H_0 : p = 1/2$ を棄却 (実現値 $2.6 \geq 1.96$). 有意水準 1% でも棄却される. よって高度に有意.]

HW 19 (両側検定) ある調味料の製造ラインでは, 各製品の砂糖の含有量は $m = 60$ (g) になるように調整している. しかしながら, 原料の不均一や製造ラインの狂いなどから, m の値は $50 \sim 70$ の間を変動するが, これまでの経験から標準偏差は常に一定で $\sigma = 3$ となっている (母分散既知). ある時点で, 製品を 25 個抜き取って, 調査したところ, 砂糖の含有量の平均値は 61.43 であった. その時点で製造ラインは $m = 60$ を保持していると考えてよいか? [有意水準 5% の両側検定で $m = 60$ を棄却 (実現値 $2.38 \geq 1.96$). 有意水準 1% では棄却されない.]

HW 20 (片側検定) ある食品の製造ラインでは, 製品 100g 中に含まれる砂糖が 2g 以下になるように調整している. ただし, 2g を多少越しても出荷して問題はない. あるロットから選んだ 200 個の標本は, 平均 2.2g の砂糖を含んでいた. 一方, この工場の工程から, 砂糖の含有量の標準偏差は 1.5g であることが経験的に知られている. 製造ラインに狂いが生じているかどうかを判定せよ. [有意水準 5%の片側検定で「狂いが生じている」]

6.3 2 種類の過誤 (Two Types of Error)

帰無仮説 H_0 をめぐって, 次の 4 つの場合がある.

採否 \ 真偽	H_0 は真	H_0 は偽
H_0 を採択	正しい判断	第 2 種の誤り
H_0 を棄却	第 1 種の誤り	正しい判断

α: 第 1 種の誤り (Type I error) 確率 = 有意水準

β: 第 2 種の誤り (Type II error) 確率

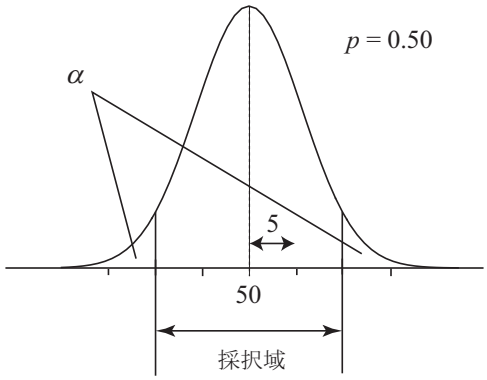
第 1 種の誤り = 生産者危険 = あわて者の間違い
第 2 種の誤り = 消費者危険 = ぼんやり者の間違い

例 題 6.4 コインを 100 回投げたとき, 表が 58 回出た. コインは公平といえるだろうか?

仮説検定を行う.

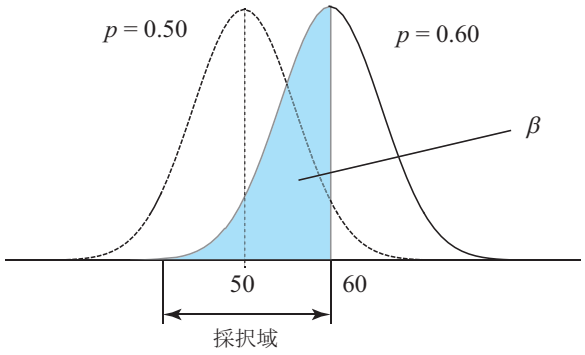
$$H_0 : p = 0.5 \qquad H_1 : p \neq 0.5$$

として, 有意水準 $\alpha = 0.05$ の両側検定を行う. $B(100, 0.5) \approx N(50, 5^2)$ を用いて, $B(100, 0.5)$ の分布と採択域を示したものが次の図である.



実現値 58 は採択域に落ちるので, H_0 は採択され, このコインは公平であると結論される. この結論を誤る確率が第 2 種誤り確率である.

コインが公平ではない場合, 可能な p は無限にあり, 第 2 種誤り確率を簡単に評価することはできない. 仮に, $p = 0.6$ としてみよう. $B(100, 0.6) \approx N(60, 24) \approx N(60, 5^2)$ なので, $B(100, 0.6)$ の分布はおおむね $B(100, 0.5)$ を右に 10 だけ平行移動したものである. 重ねて書いたものが次の図である. 採択域に実現値が現れる確率は, 網掛け部分の面積であり, これが第 2 種誤り確率 β である. おおよそ $\beta = 0.5$ でたいへん大きい.



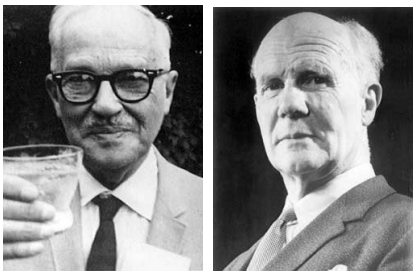
注意

(1) α 小さい $\Longleftrightarrow$ 採択域が大きい $\Longleftrightarrow$ β 大きい

(2) α, β とも小さくするためには, 標本数 n を大きくする.

(3) 「H_0 を採択する」とは言うが, はっきり否定するだけの状況ではないという消極的な採択である. そこで「H_0 を棄却できない」と言う表現も多用される.

第7章　母平均の検定



Jerzy Neyman (1894–1981)
Egon Sharpe Pearson (1895–1980)

7.1　母平均の検定 (母分散既知の場合)

●標本平均に関する基本定理 母平均 m, 母分散 σ^2 の母集団から取り出した大きさ n の標本の標本平均について, n が大きいときは,

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N\left(m, \frac{\sigma^2}{n}\right) \quad \Longleftrightarrow \quad \frac{\bar{X} - m}{\sigma/\sqrt{n}} \sim N(0,1)$$

が近似的に成り立つ (中心極限定理). 正規母集団のときは近似は不要.

7.2　母平均の検定 (母分散未知の場合: T-検定)

●基礎となる理論的結果 正規母集団 $N(m, \sigma^2)$ から取り出した n 個の標本を $X_1, \ldots, X_n$ とするとき, **不偏分散**が

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

で定義される. 標本平均 $\bar{X}$ に対して,

$$T = \frac{\bar{X} - m}{U/\sqrt{n}} \sim t_{n-1} \quad \text{自由度 } n-1 \text{ の } t\text{-分布}$$

例 題 7.1 正味 500(g) と書いてある製品を 9 個選んで調べたところ標本平均 494, 不偏分散 8^2 を得た. この製品は, 明記されたとおりの内容になっているか? [有意水準 $\alpha = 0.05$ の両側検定によって, $t = -2.25 > -2.306$ より H_0 を採択. ちなみに, $N(0,1)$ を誤用すると, $-2.25 < -1.96$ から H_0 を棄却することになる.]

例 題 7.2 (片側検定) ある製造ラインで大量の製品を作っており, その重量は正規分布に従っている. 規定値は 50kg であるが, 製品の平均重量が 50kg を切っているときはラインを直ちに止めて調整する必要がある. ある日に製造された大量の製品から 12 個をサンプリングして重量 (kg) を測定した結果, 平均値 $\bar{x} = 48.6$, 不偏分散 $u^2 = 1.6^2$ を得た. ラインを止める必要があるかを判断せよ.　　[有意水準 5% の片側検定で $H_0 : m = 50$ を棄却 (実現値 $-3.03 \leq -1.796$)]

HW 21 ある英語の資格試験の全国平均は 66 点であった. A 塾から 10 名が受験した. 結果は

78 72 65 86 58 64 76 88 74 59

であり, その平均点 72 点が 66 点を大きく上回ると A 塾は主張している. 検定によって A 塾の主張を確認せよ. [有意水準 5% の片側検定で「上回っているとは言えない」]

7.3 P 値 (P-value)

伝統的な仮説検定では, 有意水準 α を示して H_0 の棄却・採択を述べる. が, ユーザーにとって, 実現値が帰無仮説 H_0 からどのくらい外れているかを数量的に詳しく知りたいこともある. 実現値 t に対して, H_0 の下で,

$$P = \text{実現値 } t \text{ を含めて, それ以上に起こりにくい実現値が得られる確率}$$

を実現値 t の P 値という. この値をどう判断するかは, 個別事情によるもので, 数理統計学の枠外の話となる.

例 題 7.3 A 君は公平なコインを作成したつもりだ. 確認のため 80 回振ったところ表が 32 回出た. このコインは公平であるといえるか. P 値を示せ. [0.0734]

HW 22 ある機械部品の寿命は規格によって 250 時間と定められている. ある製造ラインでは, 管理状況から, 部品の長さは標準偏差 2.25 時間の正規分布にしたがっているとしてよい. 25 個のサンプルで実際に長さを調べたところ長さの平均値は 248.8 時間であった. この製造ラインの部品は規格を満たしているといえるだろうか? P 値を示せ. [0.0076]

7.4 確率変数の和

定 理 7.4 (平均値の線形性) 確率変数 X, Y と定数 a, b に対して,

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

定 理 7.5 (分散の加法性) 独立な確率変数 X, Y と定数 a, b に対して,

$$\mathbf{V}[aX + bY] = a^2\mathbf{V}[X] + b^2\mathbf{E}[Y].$$

定 理 7.6 (独立な正規確率変数の和) 2 つの確率変数 $X \sim N(m_1, \sigma_1^2)$ $Y \sim N(m_2, \sigma_2^2)$ が独立であれば, 定数 a, b に対して,

$$aX + bY \sim N\Big(am_1 + bm_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

7.5 母平均の差の検定

定 理 7.7 2 つの正規母集団 $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ から独立に取り出した大きさ n_1, n_2 の標本平均を $\bar{X}_1$, $\bar{X}_2$ とするとき,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

例 題 7.8 (母分散が既知の場合) ある物質の融点を測定した. 技術者 A は 5 回測定して平均 1264.6 度を得た. 技術者 B は 8 回測定して平均 1263.9 度を得た. 過去の経験によれば A の測定値の標準偏差は 0.7 度, B の測定値の標準偏差は 0.6 度である. さらに 2 人とも測定結果は正規分布に従うとしてよい. 2 人の測定結果に有意の差はあるか検定せよ. [$H_0 : m_1 = m_2$, $H_1 : m_1 \neq m_2$. $z = 1.85$ を得る. 有意水準 5% の両側検定で H_0 は棄却されない.]

HW 23 A 組 36 名, B 組 40 名に同じ試験をしたところ, A 組の平均点は $\bar{x}_A = 64.5$, B 組の平均点は $\bar{x}_B = 61.2$ であった. A 組は B 組よりも成績がよいといえるか. ただし, 成績は両組とも母分散 11^2 の正規分布に従うものとする.

定 理 7.9 分散が等しい 2 つの正規母集団 $N(m_1, \sigma^2)$, $N(m_2, \sigma^2)$ から独立に取り出した大きさ n_1, n_2 の標本平均を $\bar{X}_1$, $\bar{X}_2$, 不偏分散を U_1^2, U_2^2 とする.

$$U^2 = \frac{(n_1 - 1)U_1^2 + (n_2 - 1)U_2^2}{n_1 + n_2 - 2}$$

とおくとき,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) U^2}}$$

は自由度 $n_1 + n_2 - 2$ の t 分布に従う.

例 題 7.10 (母分散は未知であるが等分散である場合) 2 つの環境 A, B のもとである作物の試験栽培を行った. 環境 A からは 6 個のサンプル, 環境 B からは 8 個のサンプルをとって収穫高を調べた結果は次の通りである.

$$A : 6.2 \quad 6.0 \quad 5.9 \quad 6.2 \quad 6.1 \quad 5.8$$
$$B : 6.0 \quad 5.8 \quad 5.7 \quad 6.2 \quad 6.4 \quad 5.9 \quad 5.8 \quad 6.3$$

両者の収穫高は同じ分散をもつ正規分布に従うと仮定してよい. 環境 A,B に有意の差はあるか検定せよ. [$\bar{x}_A = 6.0333$, $u_A^2 = 0.1633^2$, $\bar{x}_B = 6.0125$, $u_B^2 = 0.2207^2$, $u^2 = 0.1987^2$, $t = 0.1937$. 一方, t_{12}-分布の上側 2.5%点は 2.179. 有意水準 5% の両側検定で有意差を認めない.]

第8章　ベイズ推定

Thomas Bayes (1702–1761)

8.1 Conditional Probability

定 義 8.1 A, B を 2 つの事象とする. $P(A) > 0$ のとき,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

を A の下での B の条件付確率という. 事象 A が起こったことを知った上で, 事象 B の起こる確率と解釈される.

例 題 8.2 (Drawing lots) 箱の中に 10 本のくじが入っていて, そのうち 2 本が当たりとなっている. 2 人が順番に 1 本ずつくじを引くとき, 先に引くのが有利か, 後のほうが有利か? [実は, 何番目に引いても当たる確率は同じである.]

例 題 8.3 サイコロを 2 個振って出る目のうち大きい方を X, 小さい方を Y とする (同じ目が出た場合は $X = Y$ とする). $P(X \geq 5|Y = 2)$ と $P(X + Y \geq 8|X \geq 4)$ を求めよ.　[4/9, 5/9]

HW 24 2 つの事象 E, F に対して, $P(E) = \dfrac{1}{3}$, $P(F) = \dfrac{1}{2}$, $P(E \cap F) = \dfrac{1}{4}$ がわかっている. 次の確率を求めよ.　[2/3, 1/12, 1/4, 1/2, 1/6, 3/7]

$$P(E^c), \quad P(E \cap F^c), \quad P((E \cup F^c)^c), \quad P(E|F), \quad P(E|F^c), \quad P(E \cap F|E \cup F)$$

8.2 Independence of Events

定 義 8.4 2 つの事象 A, B が独立であるとは,

$$P(A \cap B) = P(A)P(B)$$

を満たすときにいう. 事象の有限または無限列 $A_1, A_2, \ldots$ が独立であるとは, そこから取り出した任意有限個の事象 $A_{i_1}, A_{i_2}, \ldots, A_{i_n}$ $(i_1 < i_2 < \cdots < i_n)$ に対して

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_n})$$

が成り立つときにいう.

定 理 8.5 $P(A) > 0$ とするとき, 2 つの事象 A, B が独立であるための必要十分条件は $P(B) = P(B|A)$ である.

例 題 8.6 壺の中に $112, 121, 211, 222$ という番号のついた 4 個の玉が入っている. この壺から 1 個の玉を取り出して番号を読むとき, 1 位の数字が 1 である事象を A_1, 10 位の数字が 1 である事象を A_2, 100 位の数字が 1 である事象を A_3 とする. A_1, A_2, A_3 のいずれの 2 つも独立であるが, 3 つの事象は独立ではない.

HW 25 A, B, C が独立で, $P(A) = a$, $P(B) = b$, $P(C) = c$ とする. 次の確率を a, b, c を用いて表せ. $[a(1-b),\ a+b-ab,\ a+b+c-ab-bc-ca+abc,\ a]$

$$P(A \cap B^c), \qquad P(A \cup B), \qquad P(A \cup B \cup C), \qquad P(A|B \cup C)$$

8.3 Bayes' Formula

定 理 8.7 (Bayes' formula) $\Omega = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$ のとき, 任意の事象 B に対して,

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$$

「結果から原因を知る公式」として解釈される.

例 題 8.8 (1) ある国では, 病気 A の感染者は 500 人に 2 人の割合であるという. 検査 B は, 感染者の 95%に陽性反応を示すが, 非感染者の 2% にも陽性反応が出てしまう. ある人がこの検査を受けて陽性反応が出た. この人が感染者である確率を求めよ. [0.160]

(2) 次に, 非感染者の $100p$ % に陽性反応が出るとして, この検査を受けて陽性反応が出た人が感染者である確率を求めよ. この確率が p とともにどのように変化するか? $[1.9/(1.9+498p)]$

HW 26 ある地域では, 病気 A の感染者は 1000 人に 2 人の割合であるという. 検査 B は, 感染者の 90%に陽性反応を示すが, 非感染者の 5% にも陽性反応が出るという.

(1) この検査を受けて陽性反応が出た人が感染者である確率を求めよ. [0.0348...]

(2) この検査を受けて陰性反応が出た人が非感染者である確率を求めよ. [0.9997...]

HW 27 (条件付き確率は直感にあわないかも) 1 から 10 の番号が付いている 10 枚のチケットがある. このうち 1 番と 2 番が当たりくじとなっている. 一郎は 4 枚のチケットを買った.

(1) 一郎が「1 番をもっている」と告げたとき, 残りの 6 枚にあたりが残っている確率を求めよ. [2/3]

(2) 一郎が「少なくとも 1 枚の当たりをもっている」と告げたとき, 残りの 6 枚にあたりが残っている確率を求めよ. [4/5]

4–8 章 演習問題（期末試験対策）

演習問題 13 X_1, X_2 を区間 $[0,1]$ から取り出した標本とする. つまり, それらは独立で $[0,1]$ 上の一様分布に従う. 標本平均 $\bar{X} = (X_1 + X_2)/2$ が不偏推定量であることは既知. a を $0 < a < 1$ を満たす定数とするとき, 重み付き平均を $A = aX_1 + (1-a)X_2$ で定義する.

(1) $\mathbf{E}[A] = 1/2$ を示せ. つまり, A も母平均の不偏推定量である.

(2) $\mathbf{V}[A] \geq \mathbf{V}[\bar{X}]$ を示せ. つまり, $\bar{X}$ のほうが推定量として A より優れている.

演習問題 14 X_1, X_2 を区間 $[0,1]$ から取り出した標本とする. つまり, それらは独立で $[0,1]$ 上の一様分布に従う. それらの相乗平均を $Y = \sqrt{X_1 X_2}$ とする. $\mathbf{E}[Y] = 4/9$ を示せ. つまり, Y は母平均の不偏推定量ではない.

演習問題 15 公正なコインを 500 回投げたとき, 表は何回くらい出ると予想されるか? 知るところを述べよ.

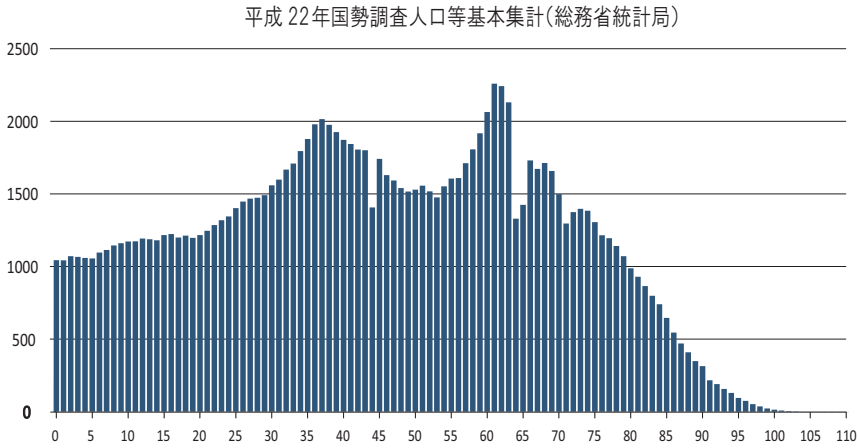
演習問題 16 平均 m が未知, 標準偏差 $\sigma = 3$ の母集団から, 取り出した 10 個の標本は次のようであった.

12 14 16 13 12 19 15 11 17 16

母平均の 90% 信頼区間, 95% 信頼区間を求めよ. [14.5 ± 1.56, 14.5 ± 1.86]

演習問題 17 人口 4000 人の町で子供の遊び場をめぐって賛否が割れている. 無作為に選んだ 100 人の意見は, 賛成 38 人, 反対 62 人であった. 町民の過半数が反対と判定してよいだろうか？[有意水準 5%の両側検定すれば「反対」と判定される]

演習問題 18 日本人の平均年齢は 44.5 歳, 標準偏差は 23.5 歳である (平成 22 年 10 月). あるサークルのメンバー 25 名の平均年齢は 32 歳である. このサークルは日本人の無作為標本といえるだろうか? 考察せよ.



演習問題 19 女子学生 1000 名の学校からランダムに選ばれた 200 人の平均身長は 157.7 cm であった. 全国の同じ年齢の女子の平均値は 158.6 cm, 標準偏差は 4.63 cm である. このクラスの平均身長は全国平均と異なると考えてよいか？[有意水準 1%の両側検定で「異なる」と判定される]

演習問題 20 ある工場で作られる製品の不良率は 8% であるという. ある日の結果は, 良品 177 個, 不良品 23 個であった. 生産工程などに異常がないと言ってよいかどうかを仮説検定で判断せよ. [有意水準 5% の両側検定で「異常なし」有意水準 5% の片側検定で「異常あり」]

演習問題 21 ある日に製造された大量の製品から 10 個をサンプリングして重量 (kg) を測定した結果,

53.2 61.5 48.1 51.3 55.7 47.2 54.5 57.9 53.8 49.2

となった. 規定値は 50kg であるが, この日に生産した製品の平均重量は規定に沿っているか？[$\bar{x} = 53.24$, $u^2 = 20.10$, $t = 2.285$. 有意水準 5% の両側検定で「規定に沿っていない」と判定される]

演習問題 22 ある国では, 病気 A の感染者は 1000 人に 4 人の割合であるという. 検査 B は, 感染者の 90%に陽性反応を示すが, 非感染者の 5% にも陽性反応が出てしまう.

(1) ある人がこの検査を受けて陽性反応が出た. この人が感染者である確率を求めよ. [0.0674]

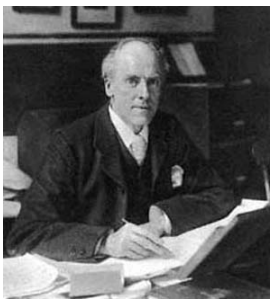
(2) ある人がこの検査を受けて陰性反応が出た. この人が非感染者である確率を求めよ. [0.9938]

演習問題 23 ある国では, $100x$ % が病気 A に感染しているという ($0 \leq x \leq 1$). 検査 B は, 感染者の 90%に陽性反応を示すが, 非感染者の 5% にも陽性反応が出てしまう. ある人がこの検査を受けて陽性反応が出た. この人が感染者である確率を x を用いて表し, x とともにどのように変化するか観察せよ.

定期試験

1. 日時：7 月 19 日 (水)1・3 講時. いつもの時間帯で受験すること.
2. 教科書・参考書・ノート・計算機等の持ち込み不可. 鉛筆と消しゴムだけで解答する.
3. 期末試験は 1 回だけ実施し, 欠席者・成績不良者に対する再試験はしない.
4. やむを得ない事情 (病気・忌引等) で定期試験を欠席し, 追試験を希望する者は正規の手続きに従って取り扱う.
5. 配布プリントの「宿題」と「演習問題」レベルが自力で解けるように, 本などをよく読んで準備してください. なお, 過去問等はウェッブページに掲載している.

第9章　χ^2-検定



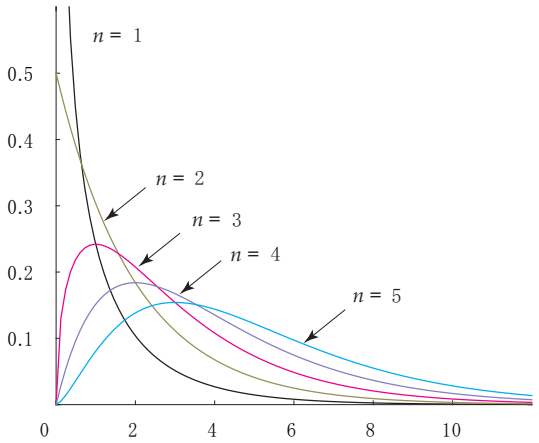
Karl Pearson (1857–1936)

9.1　χ^2-分布

密度関数が

$$f_n(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

で与えられる確率分布を**自由度 n のカイ 2 乗分布** (χ^2-分布) という. (χ^2 は一つの文字として扱う.) 自由度を明記して, χ_n^2 と書くこともある. ここで, $\Gamma(t)$ はガンマ関数.



χ^2-分布に従う確率変数 (1) $X_1, X_2, \ldots, X_n$ が独立同分布な確率変数で, 標準正規分布 $N(0,1)$ に従うとき,

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2$$

は自由度 n の χ^2-分布に従う.

(2) $X_1, X_2, \ldots, X_n$ が独立同分布な確率変数で, 正規分布 $N(m, \sigma^2)$ に従うとき,

$$\chi_{n-1}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{(標本平均)}$$

は自由度 $n-1$ のカイ 2 乗分布に従う. 上式の χ_{n-1}^2 は標本分散を計算する途中に現れる.

定 理 9.1 自由度 n の χ^2-分布 χ_n^2 の平均値と分散は,

$$m = n, \qquad \sigma^2 = 2n.$$

9.2 分布の適合度検定 (Goodness of Fit Test)

母集団の属性が $A_1, A_2, \ldots, A_k$ の k 種類に分けられている. n 個の標本から, それぞれに属するものが $X_1, X_2, \ldots, X_k$ 個得られたとする.

属性	A_1	A_2	$\cdots$	A_k	合計
理論分布	p_1	p_2	$\cdots$	p_k	1
観測度数	X_1	X_2	$\cdots$	X_k	n

観測度数から, 各属性の現れる理論分布 $p_1, p_2, \ldots, p_k$ が妥当かどうかを検定する.

定 理 9.2 (Pearson の χ^2-検定) $m_i = np_i$ とおくとき,

$$\chi_{k-1}^2 = \sum_{i=1}^{k} \frac{(X_i - m_i)^2}{m_i}$$

は, $m_1, \ldots, m_k$ が大きいとき ($m_i = np_i \geq 5$), 自由度 $k-1$ のカイ 2 乗分布に近似的に従う.

例 題 9.3 次の表は, サイコロを 120 回投げて出た目を記録したものである. このサイコロは公平と言えるだろうか?

目	1	2	3	4	5	6	合計
回数	24	18	16	22	23	17	120

[$\chi^2 = 2.9$. χ_5^2-分布の上側 5%点は 11.07. 有意水準 5% でサイコロは公平であると判断する.]

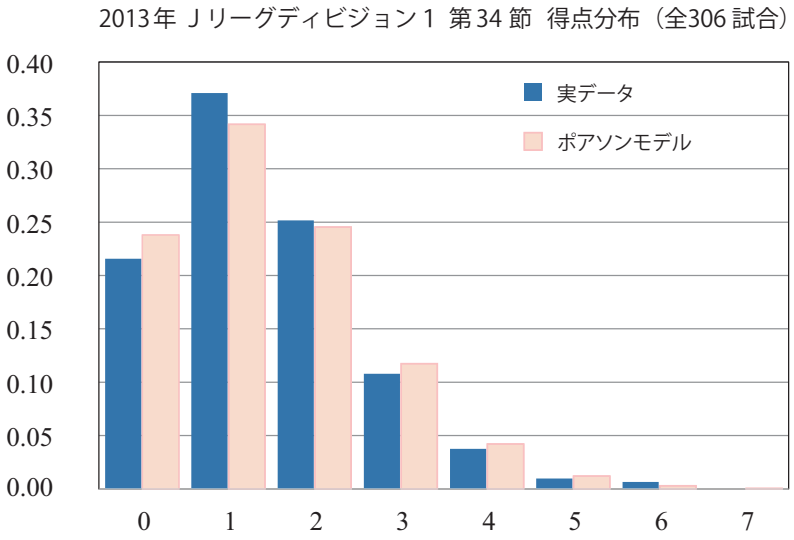
例 題 9.4 次の表は, サッカーの試合において, 1 試合 1 チーム当たりのゴール数を調べた結果である (2013 年 J リーグ・ディビジョン 1 ・第 34 節 18 チーム総当たり全 306 試合).

ゴール数	0	1	2	3	4	5	6	7 以上	合計
試合数	132	227	154	66	23	6	4	0	612
ポアソン分布	0.2379	0.3416	0.2453	0.1174	0.042	0.0121	0.0029	0.0006	1
同上理論予想	145.6	209.1	150.1	71.8	25.8	7.4	1.8	0.4	612

1 試合 1 チーム当たりのゴール数について, 平均値は 1.436, 分散は 1.367 となっている. パラメータ $\lambda = 1.436$ のポアソン分布による理論値を併記した.

(i) $m_i = np_i \geq 5$ となるようにゴール数を 0, 1, ..., 5 以上の 6 クラスに分ける.

(ii) ポアソン分布特有の事情によって, 自由度 $6 - 1 - 1 = 4$ のカイ 2 乗分布を用いる.



HW 28 次の表は, あるクラブの部員の血液型を調べた結果である. 日本人の血液型の分布は $4:3:2:1$ であると言われている. このクラブの部員の構成は, これに従っていると言えるだろうか? $[\chi^2 = 3.01.\ \chi_3^2(0.05) = 7.815.$ 従っていると言える.$]$

血液型	A	O	B	AB	合計
人数	47	23	21	9	100

HW 29 ある映画で観客の人数を調べたら, 男 45 人, 女 55 人であった. このことからこの映画は女性に人気が高いと言えるだろうか? (1) 二項母集団の母比率の検定 (2) 適合度検定, の 2 つの方法で確かめよ.

9.3 独立性の検定

定 理 9.5 2 種類の属性 $A = \{A_1, \ldots, A_r\}$, $B = \{B_1, \ldots, B_s\}$ が独立であるとき,

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\dfrac{X_{ij}}{n} - \dfrac{X_{i\cdot}}{n} \dfrac{X_{\cdot j}}{n} \right)^2}{\dfrac{X_{i\cdot}}{n} \dfrac{X_{\cdot j}}{n}}$$

は, n が大きいとき $(X_{ij} \geq 5)$, 自由度 $(r-1)(s-1)$ のカイ 2 乗分布に近似的に従う.

	B_1	B_2	$\cdots$	B_s	合計
A_1	X_{11}	X_{12}	$\cdots$	X_{1s}	$X_{1\cdot}$
A_2	X_{21}	X_{22}	$\cdots$	X_{2s}	$X_{2\cdot}$
$\vdots$			$\cdots$		$\vdots$
A_r	X_{r1}	X_{r2}	$\cdots$	X_{rs}	$X_{r\cdot}$
合計	$X_{\cdot 1}$	$X_{\cdot 2}$	$\cdots$	$X_{\cdot s}$	n

例 題 9.6 予防接種と発病の関係について次の結果を得た. 予防接種の効果は認められるか? [$\chi^2 = 10.34$. 有意水準 1% とすると, $\chi_1^2(0.01) = 6.6349$. 効果を認める.]

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

演習問題

演習問題 24 人口 150 万人のある都市で, 子供を 5 人持つ 3868 家庭を無作為抽出して, 子供 5 人の性別を調べた. この結果から, 子供を 5 人持つ家庭では男女の性比が 1:1 であると言えるだろうか? 二項分布と比較せよ. [$\chi^2 = 17.58$. $\chi_5^2(0.01) = 15.0863$. 有意水準 1 %で性比 1:1 を棄却. ちなみに性比 51:49 を検定すると, $\chi^2 = 7.97$ となり, 有意水準 5 %で仮説を採択する.]

男:女	0:5	1:4	2:3	3:2	4:1	5:0	合計
家庭数	92	603	1137	1254	657	125	3868

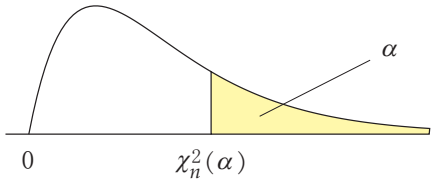
演習問題 25 次の表は, 野球の試合で 1 試合当たり両チーム合わせてのホームラン数を調べた結果である (2016 年プロ野球公式戦 楽天戦 全 143 試合). 平均値 1.448, 分散 1.786. パラメータ $\lambda = 1.448$ のポアソン分布による理論値を併記した. ホームラン数はポアソン分布に従っているかどうかを判定せよ.

ホームラン数	0	1	2	3	4	5	6	7 以上	合計
試合数	40	42	36	14	6	3	2	0	143
ポアソン分布	0.2350	0.3403	0.2464	0.1189	0.0431	0.0125	0.0030	0.0006	0.9999
同上理論予想	33.61	48.67	35.24	17.01	6.16	1.78	0.43	0.09	142.98

演習問題 26 あるグループで 1 日のテレビ視聴時間を調べて次の結果を得た. 年齢と時間数に関係があるだろうか?

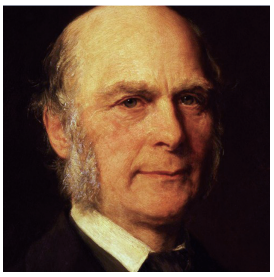
	24 歳以下	25〜35 歳	36 歳以上	合計
2 時間以内	37	155	78	270
2〜3 時間	24	59	25	108
3 時間以上	29	56	77	162
合計	90	270	180	540

カイ・スクエア分布: $P(\chi_n^2 \geq \chi_n^2(\alpha)) = \alpha$



$n\backslash\alpha$	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

値は小数第 4 位以下 ($n = 1$ では表示桁未満) を四捨五入してある.



Sir Francis Galton (1822–1911)

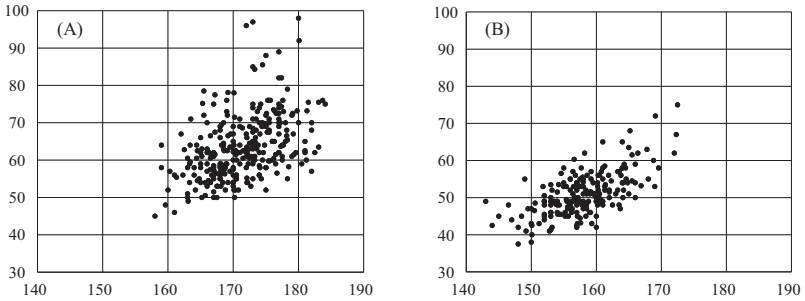
第10章　多変量の統計

10.1　2変量データの記述

2 変量データ (2 次元データ): $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

● **scatter diagram (散布図)** データを xy-座標平面に図示したもの

例 題 10.1 身長 (x) と体重 (y) の散布図. クラス (A) とクラス (B) に対する結果.



● **covariance (共分散)** n 個の 2 変量データ (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) に対して, 変数ごとの平均値と分散

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,; \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

を用いて**共分散**が定義される:

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

(注意) $\sigma_{xy} = \sigma_{yx}$. $\sigma_{xx} = \sigma_x^2$ (したがって, 分散を σ_{xx} と書く流儀もある).

● **correlation coefficients (相関係数)**

$$r = r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}}\sqrt{\sigma_{yy}}}$$

(注意) $r_{xy} = r_{yx}$.

正の相関・負の相関
強い相関・弱い相関・無相関

定 義 10.2 (観測値の規準化 (標準化))

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma_x}, \qquad \tilde{y}_i = \frac{y_i - \bar{y}}{\sigma_y}$$

定 理 10.3 2 変数 x, y に対して, 規準化された変数を $\tilde{x}, \tilde{y}$ とするとき,

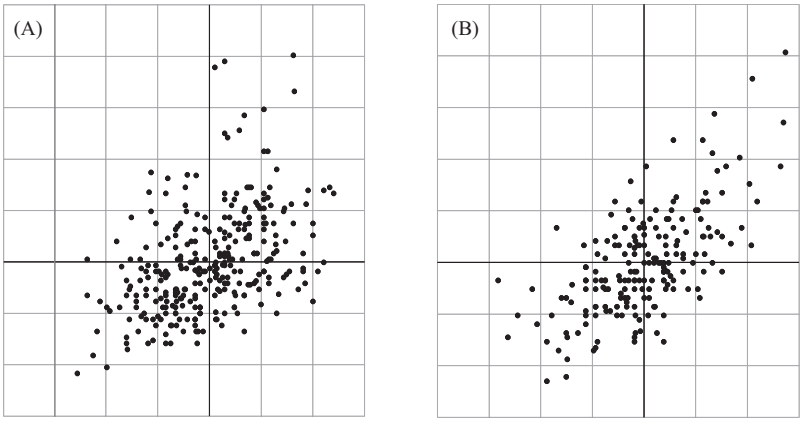
$$r_{xy} = \sigma_{\tilde{x}\tilde{y}} = r_{\tilde{x}\tilde{y}} \tag{10.1}$$

が成り立つ. 特に, 変数 x, y の相関係数は, それらを規準化した変数 $\tilde{x}, \tilde{y}$ の共分散に一致する.

定 理 10.4 相関係数は $-1 \leq r_{xy} \leq 1$ を満たす.

証 明 $\sum\{t(x_i - \bar{x}) + (y_i - \bar{y})\}^2 \geq 0$ がすべての t で成り立つことを用いる. ∎

例 題 10.5 規準化された変数に対する散布図.



	共分散	相関係数
クラス A	20.15	0.45
クラス B	20.23	0.65

HW 30 2 変量データ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が $\sigma_x > 0$, $\sigma_y > 0$ を満たすものとする. このとき, 散布図が右上がりの直線に乗ることと相関係数が $r = 1$ を満たすことは同値であることを示せ. また, 散布図が右下がりの直線に乗ることと相関係数が $r = -1$ を満たすことは同値であることを示せ. [ヒント：定理 10.4 の証明を見直す.]

10.2 Random Vectors

定 義 10.6 2 つの確率変数 X, Y に対して, **covariance (共分散)** が

$$\sigma_{XY} = \mathrm{Cov}\,(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

で定義される. さらに, **correlation coefficient (相関係数)** が次で定義される:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}}$$

定 理 10.7 相関係数は $-1 \leq r_{XY} \leq 1$ を満たす.

例 題 10.8 サイコロを 2 個投げて出た目のうち大きい方 (同じ目のときはその目) を X, 小さい方 (同じ目のときはその目) を Y とする. X, Y のとり得る値について確率を求めて表にしたものが結合分布.

$X\backslash Y$	1	2	3	4	5	6	合計
1	1/36	0	0	0	0	0	1/36
2	2/36	1/36	0	0	0	0	3/36
3	2/36	2/36	1/36	0	0	0	5/36
4	2/36	2/36	2/36	1/36	0	0	7/36
5	2/36	2/36	2/36	2/36	1/36	0	9/36
6	2/36	2/36	2/36	2/36	2/36	1/36	11/36
合計	11/36	9/36	7/36	5/36	3/36	1/36	1

$$\mathbf{E}[X] = \frac{161}{36}, \quad \mathbf{E}[Y] = \frac{91}{36}, \quad \mathbf{V}[X] = \mathbf{V}[Y] = \frac{2555}{36^2}, \quad \mathrm{Cov}\,(X,Y) = \frac{1225}{36^2}, \quad r = \frac{35}{73}$$

HW 31 サイコロを 4 回投げるとき, 1 の目の出る回数を X, 6 の目の出る回数を Y とする. X, Y の相関係数を求めよ. $[r_{XY} = -1/5]$

10.3 Regression Models

2 変量データ (x_i, y_i) を関数 $y = f(x)$ を用いて合理的に表したい (x を説明変数, y を目的変数という). 特に, 1 次関数

$$y = ax + b$$

によるものを **linear regression model (線形回帰モデル)** または y の x への**回帰直線**という.

● **Method of least squares (最小二乗法)** 1 次関数 $y = ax + b$ を想定して, 実際の観測では $x = x_i$ に対する観測値 y_i は偏差をともなって現れると考え, 各観測値 (x_i, y_i) に対して偏差 ϵ_i を

$$y_i = ax_i + b + \epsilon_i$$

によって定義する. 偏差の平方和

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

を最小にするように a, b を定めるのが最小二乗法である. Q は a, b に関して 2 次関数なので, 最小化するのは易しい. 偏微分を計算して,

$$\frac{\partial Q}{\partial a} = 2an(\sigma_x^2 + \bar{x}^2) - 2n(\sigma_{xy} + \bar{x}\bar{y}) + 2bn\bar{x},$$
$$\frac{\partial Q}{\partial b} = 2bn - 2n\bar{y} + 2an\bar{x}$$

が得られる. 連立方程式 $\dfrac{\partial Q}{\partial a} = \dfrac{\partial Q}{\partial b} = 0$ を解くと, 解は 1 つだけであって,

$$a_0 = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad b_0 = \bar{y} - a_0\bar{x} \tag{10.2}$$

求めるべき線形回帰モデルは $y = a_0x + b_0$ で与えられる.

定 理 10.9 2 変量データ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ に対して, x を説明変数, y を目的変数とする線形回帰モデルは

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) = \frac{\sigma_y}{\sigma_x}\, r(x - \bar{x}) \quad \Leftrightarrow \quad \frac{y - \bar{y}}{\sigma_y} = r\,\frac{x - \bar{x}}{\sigma_x} \tag{10.3}$$

で与えられる. また, y を説明変数, x を目的変数とする線形回帰モデルは

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}) = \frac{\sigma_x}{\sigma_y}\, r(y - \bar{y}) \quad \Leftrightarrow \quad \frac{x - \bar{x}}{\sigma_x} = r\,\frac{y - \bar{y}}{\sigma_y} \tag{10.4}$$

で与えられる. ただし, r は相関係数である.

(注意) 定理に述べた 2 つの回帰モデルは, いずれも平均ベクトル $(\bar{x}, \bar{y})$ の定める点を通るが, それらは一般には一致しない (説明変数と目的変数は対称的な役割にない).

例 題 10.10 クラス A,B に所属する学生の身長 (x) と体重 (y) のデータをもとに線形回帰モデルを作ろう. クラス A について,

$$\bar{x} = 171.45, \quad \bar{y} = 63.59,$$
$$\sigma_x^2 = 27.7557, \quad \sigma_y^2 = 73.3508, \quad \sigma_{xy} = 20.1530$$

となっている. したがって, x を説明変数とする線形回帰モデルは,

$$y = 0.73x - 61.57 \tag{10.5}$$

となる. また, y を説明変数とする線形回帰モデルは

$$x = 0.27y + 154.28 \tag{10.6}$$

となる. 回帰直線 (10.6) の傾き $1/0.27 \approx 3.70$ は, 確かに回帰直線 (10.5) の傾きに一致せずそれより大きい. 同様にして, クラス B について計算すると,

$$\bar{x} = 157.98, \quad \bar{y} = 51.05,$$
$$\sigma_X^2 = 28.1218, \quad \sigma_Y^2 = 34.6541, \quad \sigma_{XY} = 20.2323$$

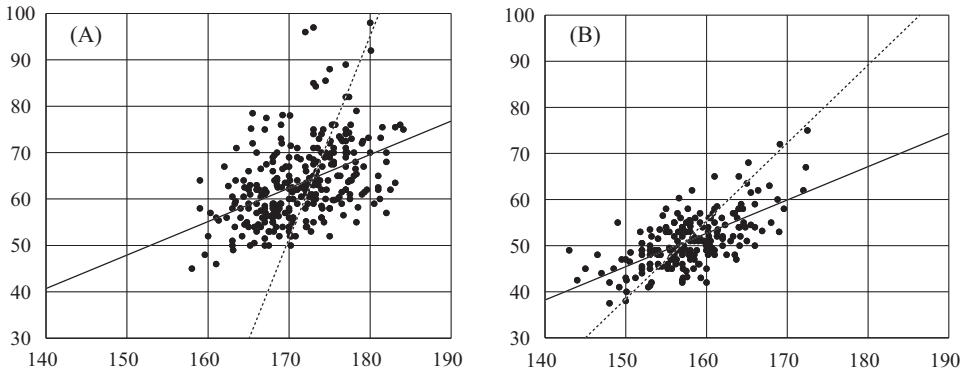
から, x を説明変数とする線形回帰モデルは,

$$y = 0.72x - 62.70$$

となり, y を説明変数とする線形回帰モデルは

$$x = 0.58y + 128.18$$

となる.



HW 32 4 つのデータ $(0,1)$, $(1,3)$, $(3,6)$, $(4,6)$ に対して x を説明変数とする線形回帰モデルを求めよ. $[y-4=1.3(x-2)]$

演習問題

演習問題 27 変数 x, y の共分散と標準偏差に関して, $|\sigma_{xy}| \leq \sigma_x\sigma_y$ を示せ.

演習問題 28 親の形質が子にどのくらい遺伝するかに大変興味を持った Galton は, 親子の身長を調査して分析を行った (1886). 今日「回帰分析」と呼ばれる統計解析のさきがけとして有名な研究である. 下の表は Galton が分析を行ったデータである. ざっと見て正の相関があることは明らかであるが, はたして相関係数はどのくらいだろうか? 数値 (単位はインチ) のはっきりしている網掛け部分のデータを用いて計算してみよ.

	Mid-height parents (x)											
Adult Children (y)		below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
	above							5	3	2	4		14
	73.2						3	4	3	2	2	3	17
	72.2			1		4	4	11	4	9	7	1	41
	71.2			2		11	18	20	7	4	2		64
	70.2			5	4	19	21	25	14	10	1		99
	69.2	1	2	7	13	38	48	33	18	5	2		167
	68.2	1		7	14	28	34	20	12	3	1		120
	67.2	2	5	11	17	38	31	27	3	4			138
	66.2	2	5	11	17	36	25	17	1	3			117
	65.2	1	1	7	2	15	16	4	1	1			48
	64.2	4	4	5	5	14	11	16					59
	63.2	2	4	9	3	5	7	1	1				32
	62.2		1		3	3							7
	below	1	1	1			1		1				5
	sum	14	23	66	78	211	219	183	68	43	19	4	928