2025 年 9 月 25 日

『データサイエンス指向の統計学［改訂版］』

（大内俊二 著，学術図書出版社）

問・章末問題解答

第 1 章

問 1.1 教師あり学習：決定木（けっていぎ）分析・判別分析・ロジスティック回帰・サポートベクターマシンなど．
教師なし学習：クラスター分析・アソシエーション分析・主成分分析など．

章末問題

1.1 略．

1.2 Amazon のレコメンデーションシステム，NewsPicks のニュースキュレーションサービスなど．

1.3 略．

1.4 解答例：個々には個人情報が守られる規制がなされているが，ビッグデータの中にあるいろいろなデータを融合することで，個人情報が明らかになってしまうという危険性は拭えない．データ活用における倫理観や道徳観の欠如から生ずる社会問題．

1.5 略．

1.6 略．

1.7 略．

1.8 ビジネス用語としては「最新のデジタル技術を駆使したデジタル化時代に対応するための企業の変革」，広義には「IT の浸透が人々の生活をあらゆる面でより良い方向に変化させる」の意味で使われる．経済産業省が 2018 年 12 月に発表した DX 推進ガイドラインでは，「企業がビジネス環境の激しい変化に対応し，データとデジタル技術を活用して，顧客や社会のニーズを基に，製品やサービス，ビジネスモデルを変革するとともに，業務そのものや，組織，プロセス，企業文化・風土を変革し，競争上の優位性を確立すること」とより詳細に定義している．

第 2 章

問 2.1 略．

問 2.2 $\dfrac{60}{50} = 1.2$ より，20 %増えた．

問 2.3 2 $(= 10 - 8)$ パーセントポイント（ポイントといわれることが多い）増えた．2 パーセント増えたとはいわないので注意しよう．

章末問題

2.1 (1) D　(2) A　(3) D　(4) A　(5) B　(6) B　(7) B

2.2 5 項選択回答形式は，意見の特徴を 3 項選択回答形式よりもより細かくとらえることができ，「どちらでもない」への回答が，3 項選択回答形式よりも減る傾向が

ある．極端な判断を好まない一般的な日本人ではとくにその傾向が強くなると考えられる．

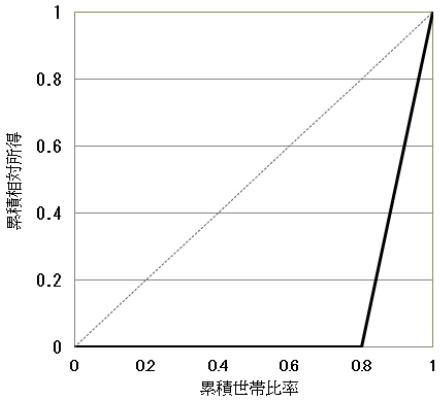
2.3 (1) 正　(2) 誤　(3) 誤　(4) 誤　(5) 正　(6) 誤　(7) 正

2.4 空き家数を総住宅数で割った空き家率で比較すべきである．

第 3 章

問 3.1 年間収入が完全に平等に配分されている場合，表 3.4 の 5 つの階級の年間収入比率はすべて 0.2 となり，累積年間収入比率が累積世帯比率（階級が 1 つ上がるにつれ 0.2 ずつ増えてゆく）に一致するため．

問 3.2 五分位階級の場合，下図のようになる．十分位階級にするなど階級の数が増えると，最後の階級の線分の始点は点 $(1, 0)$（右下のコーナー）に近づき，その線分は横軸に垂直になってゆく．

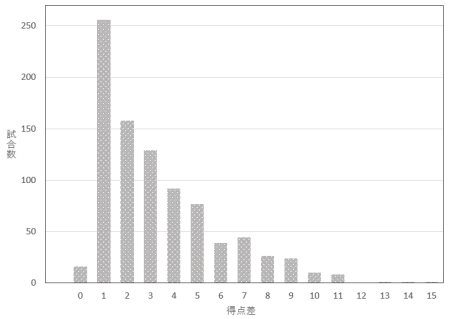


問 3.3 略．

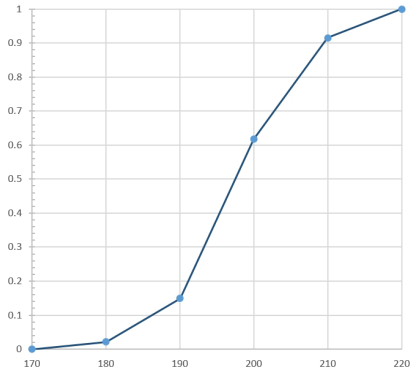
問 3.4 略．

章末問題

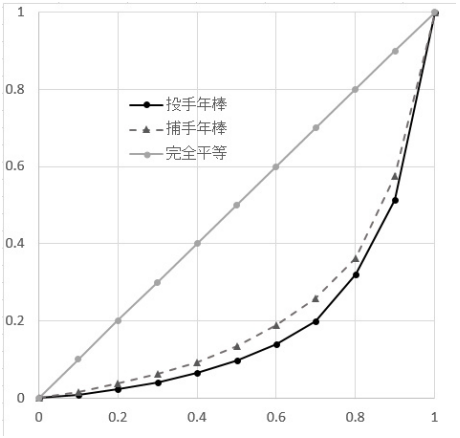
3.1 得点差の棒グラフ



3.2 (2) 累積相対度数の折れ線グラフ　　(3) 38 %

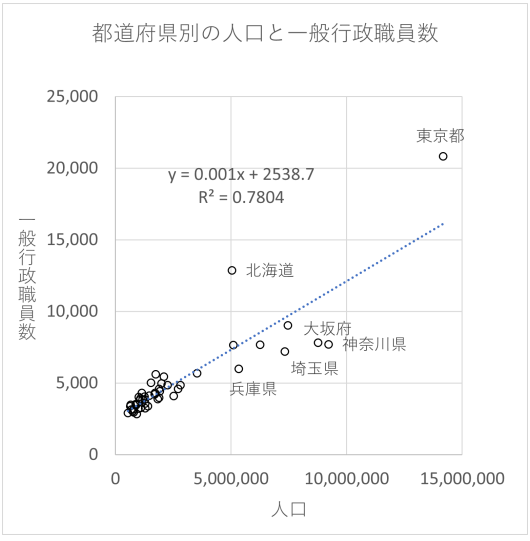


3.3 投手と捕手の年俸のローレンツ曲線



参考までに投手と捕手の年俸のジニ係数を R で計算すると，それぞれ 0.619, 0.554（ジニ係数の値は近似方法によって計算結果が違ってくることに注意）となり，このデータからは投手の年俸の格差のほうが捕手のそれより大きいといえる．

3.4 下図は，令和 6 年の都道府県別の人口と令和 5 年の一般行政職員数の散布図である．散布図作成の Excel による操作方法については，補章 p.217 を見よ．



散布図の点が右上がりの直線的に散らばっているので，都道府県の一般行政職員数は，その人口の一次関数でおおよそ説明できるといえる．東京と北海道は，人口に対する般行政職員数が他の府県に比べ多い．一方，その数が目立って少ないのは，神奈川県・埼玉県・兵庫県・大阪府である．

第 4 章

問 4.1 $w_i = \dfrac{1}{n} \quad (i = 1, 2, \ldots, n)$

問 4.2 (4.7) の左辺の分数部分 $= \dfrac{\frac{410}{400} \times 400 \times 20 + \frac{130}{140} \times 140 \times 30 + \frac{140}{120} \times 120 \times 50}{(4.4)}$

$= \dfrac{410}{400} \dfrac{400 \times 20}{(4.4)} + \dfrac{130}{140} \dfrac{140 \times 30}{(4.4)} + \dfrac{140}{120} \dfrac{120 \times 50}{(4.4)}$

問 4.3 $(1 + r)^5 = 1.15$ より $r = \sqrt[5]{1.15} - 1 \fallingdotseq 1.028 - 1 = 0.028 = 2.8\ \%$

$\sqrt[5]{1.15}$ は Excel ではセルに「$= 1.15\hat{}(1/5)$」と入力すれば求まる．

問 4.4 略．

問 4.5 「2 所得の分布状況」にある図 9 に所得金額階級別世帯数の相対度数分布が示されている．この分布は右に長い裾をひいており，その平均所得金額は 524 万 2 千円，中央値は 405 万円で，平均所得金額以下の割合は 62.2 %となっている．

問 4.6 3 つのデータすべてにおいて，算術平均 = 中央値 = 最頻値 = 5 となる．したがって代表値では 3 つのデータは識別できない．

問 4.7 偏差の合計 $= (x_1 + x_2 + \cdots + x_n) - n\bar{x} = n\bar{x} - n\bar{x} = 0$．したがって偏差の算術平均も常に 0 になる．

問 4.8 平均 $\bar{x} \fallingdotseq 3.21$, 分散 $s^2 = \dfrac{(1 - 3.21)^2 \times 19 + (2 - 3.21)^3 \times 3 + \cdots + (13 - 3.21)^2 \times 1}{48}$

$\fallingdotseq 6.00$, 標準偏差 $s = \sqrt{6.00} \fallingdotseq 2.45$（$\sqrt{6.00}$ は Excel の関数 SQRT(6.00) で計算できる）．

問 4.9 それぞれの変動係数を求めると，$\mathrm{CV_A} = 10\ \%$，$\mathrm{CV_B} = 0.2\ \%$ となるので，銘柄 A の変動のほうが大きかったといえる．

問 4.10 略．

問 4.11 略．

問 4.12 $T_{\text{数学}} = 10z_{\text{数学}} + 50 = 10 \times \dfrac{2}{3} + 50 \fallingdotseq 57,$

$T_{\text{国語}} = 10z_{\text{国語}} + 50 = 10 \times 2 + 50 = 70.$

章末問題

4.1 (1) A 高校 60 点, B 高校 55 点.
(2) A 高校 75 点, B 高校 70 点.
(3) A 高校 63 点, B 高校 67 点.
(4) 自分の予想通りの結果になったか.

4.2 基準となる年の死亡状況が今後変化しないと仮定したときに，各年齢の者が平均的に見て今後何年生きられるかという期待値を表したものを平均余命（へいきんよみょう）といい，特に 0 歳の平均余命を平均寿命という．自分が同年齢の平均的な日本人であると仮定し，あと何年くらい生きることができるかは，平均寿命から自分の年齢を引いた値ではなく，自分の年齢の平均余命で判断すべきである．

4.3 男子：平均 12.01 kg，中央値 11.95 kg.
女子：平均 11.64 kg，中央値 11.50 kg.
男女いずれも，平均のほうが大きいので体重の分布は右に歪んでいるといえる．

4.4 ある地域の最高気温の分布など考えてみよ．自然界では最大値は右に，最小値は左に集中する傾向があるといわれている．

4.5 $(1+r)^3 = \dfrac{101}{100}\times\dfrac{103}{101}\times\dfrac{105}{103}$ より $r = \sqrt[3]{1.05}-1 \fallingdotseq 1.016-1 = 0.016 = 1.6\,\%$

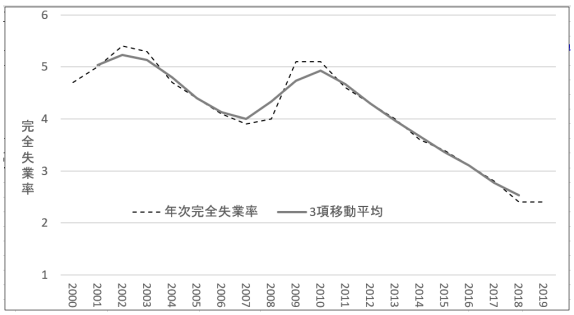
4.6 (1)平均の速さ $= \dfrac{\text{移動距離}}{\text{要した時間}} = \dfrac{400}{\frac{200}{40}+\frac{200}{50}} \fallingdotseq 44.4\,\mathrm{km/h}$

(2) 略.

4.7 (1) ラスパイレス $= \dfrac{500\cdot 1000+400\cdot 2000+1100\cdot 500}{300\cdot 1000+500\cdot 2000+1000\cdot 500}\times 100 \fallingdotseq 102.8\ \%$

(2) パーシェ $= \dfrac{500\cdot 1500+400\cdot 2800+1100\cdot 600}{300\cdot 1500+500\cdot 2800+1000\cdot 600}\times 100 \fallingdotseq 103.3\ \%$

4.8



4.9 平均 $= (175\times 1+185\times 2+195\times 14+205\times 22+215\times 7+225\times 1)/47 \fallingdotseq 202.4$ (千円)，中央値 $= 200+10\times\dfrac{24-17}{22} \fallingdotseq 203.2$，最頻値 $= 205$.

〈度数分布（ヒストグラム）からの中央値の求め方〉 総度数が 47 だから，中央値は前から 24 番目のデータで，200 以上 210 未満の階級に入る．この階級の下限は 200 で，中央値の位置は 200 より大きく 210 未満にあるので，$200+h$ $(0<h<10)$ と表せる．h の値を求めるために，この階級にデータが一様に分布すると仮定する．当該階級の直前の階級までには $1+2+14=17$ 個あるから，当該階級にある中央値までに入る度数は $(24-17)$ 個．また，ヒストグラムのこの階級の度数を柱の横幅で比例配分すると，$10:h=22:(24-17)$ が成り立つ．この比例式より $h=10\times\dfrac{24-17}{22}$ を得る．

4.10 最大値 $=1,163$（東京），最小値 $=951$（秋田），標準偏差 $s \fallingdotseq 52.7$

4.11 平均 $\fallingdotseq 68.8$ 万円，中央値 $=60+10\times\dfrac{24-4}{25}=68.0$ 万円，最頻値 $=65$ 万円，標準偏差 $\fallingdotseq 8.39$ 万円． 分布は正の方向に歪んでいる (右に長い裾をひく分布)．ちなみに同年の法文経系公立大学（昼間部）の平均は 54.3 万円，標準偏差は 3.32 万円．

4.12 (1) $\overline{x}_1=\overline{x}_2$，　$s_1<s_2$　(2) $\overline{x}_3<\overline{x}_4$, $s_3=s_4$

4.13 平均と標準偏差，それぞれの意味を考えて判断する．

4.14 範囲で比較すると，$R_{\mathrm{A}}=4<R_{\mathrm{B}}=8$．標準偏差で比較すると，$s_{\mathrm{A}} \fallingdotseq 1.41 < s_{\mathrm{B}} \fallingdotseq 2.38$.

4.15 分布が正の方向に顕著に歪んでおり（右に長い裾をひく分布），“31”という外れ値と判断される値も存在するので，平均値と標準偏差ではなく，中央値 Me と四分位範囲 IQR を用いる．$\mathrm{Me}=3.5$, $\mathrm{IQR}=7-3=4$.

〈四分位数の求め方〉 まずデータを小さい順に並べる．いまの場合，データの大きさが 14（偶数）なので，データを半分に分け，大きいほうのグループの中央値 7 を第 3 四分位数 Q_3，小さいほうのグループの中央値 3 を第 1 四分位数 Q_1 とすればよい．

4.16 平成 19 年度の CV $\fallingdotseq 20.7$ %，令和 3 年度の CV $\fallingdotseq 16.4$ %

4.17 身長の CV $\fallingdotseq 3.41$ %，体重の CV $\fallingdotseq 14.27$ %

4.18 (1) 算術平均 4, 標準偏差 2.

(2) (1) のデータを x_1，x_2，x_3，x_4，x_5 とし，問題 (a), (b), (c) のデータを u_1，u_2，u_3，u_4，u_5 とする．

(a) $u_i = x_i + 10$ $(i = 1, 2, \ldots, 5)$ と表せるので，算術平均 $\overline{u} = \overline{x} + 10 = 14$，標準偏差 $s_u = s_x = 2$

(b) $u_i = 0.1x_i$ $(i = 1, 2, \ldots, 5)$ より，算術平均 0.4，標準偏差 0.2.

(c) $u_i = 0.1x_i + 10$ $(i = 1, 2, \ldots, 5)$ より，算術平均 10.4，標準偏差 0.2.

4.19 $\bar{x}_{\mathrm{F}} = 1.8\bar{x}_{\mathrm{C}} + 32$, $s_{\mathrm{F}} = 1.8s_{\mathrm{C}}$ （章末問題 *4.18* の (2) の性質を用いる．）

4.20 (1) $n\overline{x} = \sum_{i=1}^{n} x_i$ より，$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = 0$ だから

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right) = \frac{1}{ns_x}\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

(2) $${s_z}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^2 = \frac{1}{{s_x}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$= \frac{1}{{s_x}^2}\cdot{s_x}^2 = 1 \text{ より，} s_z = 1.$$

(3) 標準得点の定義式 $z_i = \dfrac{x_i - \overline{x}}{s_x}$ の分子と分母は同じ単位をもつので，割り算することによって単位がなくなるから．

(4) $\overline{h} = 10\overline{z} + 50 = 50$, $s_h = 10s_z = 10$ （章末問題 *4.18* の (2) の性質を用いる．）

4.21 数学の標準得点 $= \dfrac{65 - 50}{15} = 1$ （偏差値 $= 1 \times 10 + 50 = 60$），英語の標準得点 $= \dfrac{72 - 65}{10} = 0.7$（偏差値 $= 0.7 \times 10 + 50 = 57$）だから数学のほうがよい．

4.22 この試験の平均 $\overline{x} = 1.0$，標準偏差 $s \fallingdotseq 9.95$ だから，$T_{100} \fallingdotseq 149$, $T_0 \fallingdotseq 49$．この問題のように，データの分布が極端に偏っている場合には，偏差値は意味のない値をとる．

第 5 章

問 5.1 大企業と若者の定義が曖昧である．例えば，大企業は「資本金の額又は出資の総額が 3 億円を越え，かつ常時使用する従業員の数が 300 人を越える会社及び個人」，若者は「15 歳から 34 歳までの日本在住の日本人」など，明確に規定する．正社員・パート・アルバイト・派遣社員・契約社員など雇用形態も限定する必要があろう．

問 5.2 層別抽出法では，全地域から標本が抽出されるが，2 段抽出法では，標本がいくつかの地域に片寄ってしまう．

章末問題

5.1 次の理由によって生ずる誤差．回答者によるもの：質問に対する虚偽の回答・質問の意味の取り違え・回答者の不在や回答拒否による調査不能など，調査・分析者によるもの：調査員の不正・調査データを集計する際の集計ミス・コンピュータを用いてデータ解析を行う際の調査データの入力ミスなど．

5.2 例えば，入試制度の改善には賛成だが入学定員の増員には反対であるという人は，回答に困ってしまう．この質問のように，1 つの質問で 2 つ（以上）のことを聞いている質問をダブルバーレル質問 (*double-barrelled question*) という．アンケート調査では，1 つの質問で 2 つ以上のことを聞いてはいけない．

5.3 (d)

5.4 仮定から大きさ 1000 の標本の半分が賛成者で，残り半分が反対者であるとする．賛成者は 80 %が調査に協力するので，賛成者の有効回答数は $500 \times 0.8 = 400$，同様に反対者の有効回答数は $500 \times 0.4 = 200$ となり，賛成者と反対者の比は $2:1$ となり，母集団における同比から大きくずれてしまう．調査において回答率は，見逃せない重要な問題である．

5.5 農林水産省の Web サイトにある「食育に関する意識調査報告書」の「6 標本抽出方法」には，層化二段無作為抽出法の具体的な手順が詳細に紹介されている．

`https://www.maff.go.jp/j/syokuiku/ishiki/r03/pdf/houkoku_5.pdf`

（閲覧日:2021 年 10 月 15 日）

第 6 章

問 6.1 $\Omega = \{(0,0),\ (1,0),\ (0,1),\ (1,1)\}$

問 6.2 (1) $\Omega = \{1,2,3,4,5,6,7,8,9,10\}$
(2) $B = \{1,2,5,10\}$, $A \cap B = \{1,5\}$, $A \cup B = \{1,2,3,5,7,9,10\}$, $A^c = \{2,4,6,8,10\}$

問 6.3 $\dfrac{3}{7}$

問 6.4 (1) $P(A \cup B) = 0.7$
(2) $P(A \cap B) = 0$
(3) $P(A^c) = 0.6$
(4) $P(A^c \cap B) = P(B) = 0.3$
(5) $P(A^c \cap B^c) = 1 - P(A \cup B) = 0.3$

問 6.5 $P(A \cap B) = \dfrac{3}{52}$, $P(A)P(B) = \dfrac{13}{52} \times \dfrac{12}{52} = \dfrac{3}{52}$ より $P(A \cap B) = P(A)P(B)$ が成り立つ．

問 6.6 $P(A \cup B) = 0.4 + 0.2 - 0.4 \times 0.2 = 0.52$

問 6.7 (I) $\left(\dfrac{3}{5}\right)^3 = \dfrac{27}{125}$ (II) $\dfrac{3}{5} \times \dfrac{2}{4} \times \dfrac{1}{3} = \dfrac{1}{10}$

章末問題

6.1 (1) $\dfrac{\frac{\sqrt{3}}{2}r}{r} = \dfrac{\sqrt{3}}{2}$ (2) $\dfrac{\frac{3}{4}\pi r^2}{\pi r^2} = \dfrac{3}{4}$

6.2 (1) $P(A \cap B) = P(A)P(B \,|\, A) = 0.6 \times 0.8 = 0.48$
(2) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.5 - 0.48 = 0.62$
(3) $P(A \,|\, B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.48}{0.5} = 0.96$

6.3 (1) $P(A \cap B) = 0.0198$, $P(A^c \cap B) = 0.098$
(2) $P(A \,|\, B) = \dfrac{0.0198}{0.0198 + 0.098} \fallingdotseq 0.168$

6.4 $P(A \,|\, B) = 0.3$, $P(A \,|\, B^c) = 0.8$, $P(B) = 0.03$, $P(B^c) = 0.97$ だから
求める確率 $P(B \,|\, A) = \dfrac{0.03 \times 0.3}{0.03 \times 0.3 + 0.97 \times 0.8} \fallingdotseq 0.011$.

第 7 章

問 7.1 第 1 式が成り立つのは明らか．$p(1) + p(2) + \cdots + p(6) = 6 \times \dfrac{1}{6} = 1$

問 7.2 実際に何回かサイコロ振った結果，各々の目が何回出たか数えた度数または相対度数を各目ごとに示したものが度数分布である．一方，確率分布は 1 から 6 のすべての目が出ることは同様に確からしいと仮定して想定した確率モデルである．

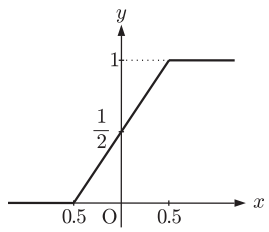
問 7.3 $P(X=x)=\dfrac{1}{2}$ $(x=0,\ 1)$．離散型確率分布は

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

のように表に示すことも多い．

問 7.4 確率 $P(0<X<0.25)=\displaystyle\int_0^{0.25} f(x)\,dx$ は，図の網掛け部分（長方形）の面積だから $1\times 0.25=0.25$．

問 7.5



問 7.6 $E[X]=1\times\dfrac{1}{6}+2\times\dfrac{1}{6}+3\times\dfrac{1}{6}+4\times\dfrac{1}{6}+5\times\dfrac{1}{6}+6\times\dfrac{1}{6}=3.5$

問 7.7 $E[X]=2\times\dfrac{1}{3}+\dfrac{1}{2}\times\dfrac{2}{3}=1$ 万円

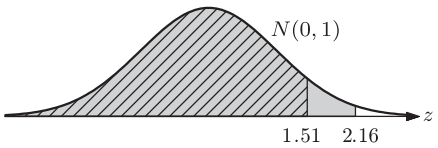
$V[X]=1\times\dfrac{1}{3}+\left(\dfrac{1}{2}\right)^2\times\dfrac{2}{3}=0.5$ 万円^2

問 7.8 $E[g(X)]=(x_1-\mu)^2p(x_1)+(x_2-\mu)^2p(x_2)+\cdots+(x_k-\mu)^2p(x_k)$ となり，分散 (7.11) に一致する．

問 7.9 $P(X=3)={}_5\mathrm{C}_3\left(\dfrac{1}{3}\right)^3\left(\dfrac{2}{3}\right)^2=10\times\left(\dfrac{4}{243}\right)\fallingdotseq 0.165$

問 7.10 $P(X\geq 18)=\displaystyle\sum_{x=18}^{20}{}_{20}\mathrm{C}_x\,0.5^x\,0.5^{20-x}=1-\mathsf{BINOM.DIST}(17,20,0.5,\mathsf{TRUE})\fallingdotseq 0.0002$

問 7.11 (1) $P(Z<1.24)=\mathsf{NORM.DIST}(1.24,0,1,\mathsf{TRUE})\fallingdotseq 0.8925$
標準正規分布用の関数 $\mathsf{NORM.S.DIST}(1.24,\mathsf{TRUE})$ を使ってもよい．
(2) $\underbrace{P(Z<2.16)}_{\text{網掛け部分の面積}}-\underbrace{P(Z<1.51)}_{\text{斜線部分の面積}}\fallingdotseq 0.0501$



問 7.12 $\mathsf{NORM.S.DIST}(2,\mathsf{TRUE})-\mathsf{NORM.S.DIST}(-2,\mathsf{TRUE})\fallingdotseq 0.9545$

問 7.13 $\overline{X}$ は $N\left(1180,\ \dfrac{20^2}{25}\right)$ に従う．$P(\overline{X}>1170)=1-P(\overline{X}\leq 1170)=1-\mathsf{NORM.DIST}(1170,1180,4,\mathsf{TRUE})\fallingdotseq 0.9938$

問 7.14 標本サイズ $n=64$ は十分に大きいので，中心極限定理より $\overline{X}_{64}$ の標本分布は正規分布 $N\left(80,\ \dfrac{10^2}{64}\right)$ で近似できる．この正規分布のもとで，確率 $P(\overline{X}_{64}\leq 77)$ は Excel の関数 $\mathsf{NORM.DIST}(77,80,10/8,\mathsf{TRUE})$ で求まり，その近似値は 0.0082 となる．

問 7.15 $P(X < -3 \text{ または } 3 < X) = 2 \times \mathsf{NORM.DIST}(-3, 0, 1, \mathsf{TRUE}) \fallingdotseq 0.0027$, $P(T < -3 \text{ または } 3 < T) = 2 \times \mathsf{T.DIST}(-3, 3, \mathsf{TRUE}) \fallingdotseq 0.0577$，後者は前者の約 21 倍となっている．

章末問題

7.1 142 円

7.2

x	2	3	4	5	6	7	8	9	10	11	12	計
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

7.3 (1) 1 回目に赤球が出れば $X = 1$，白球は 4 個しかないから X の最大値は 5 である．X のとりうる値は 1, 2, 3, 4, 5.

(2)

x	1	2	3	4	5	計
$P(X=x)$	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{6}{35}$	$\frac{3}{35}$	$\frac{1}{35}$	1

(3) $E[X] = 1 \times \dfrac{3}{7} + 2 \times \dfrac{2}{7} + 3 \times \dfrac{6}{35} + 4 \times \dfrac{3}{35} + 5 \times \dfrac{1}{35} = 2$（回）

$V[X] = 1^2 \times \dfrac{3}{7} + 1^2 \times \dfrac{6}{35} + 2^2 \times \dfrac{3}{35} + 3^2 \times \dfrac{1}{35} = 1.2$ より

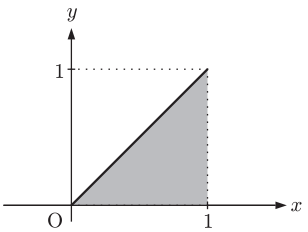
$\sigma = \sqrt{V[X]} = \sqrt{1.2} \fallingdotseq 1.1$（回）

7.4 (1) 確率変数 X は自然数の値をとる．X の実現値が k になるのは，サイコロを繰り返し投げて $k-1$ 回連続して 1 以外の目が出たあとに 1 の目が出る場合であるから，その確率は $P(X=k) = \left(1 - \dfrac{1}{6}\right)^{k-1} \cdot \dfrac{1}{6} = \dfrac{1}{6}\left(\dfrac{5}{6}\right)^{k-1}$ となる．X の確率分布は $P(X=k) = \dfrac{1}{6}\left(\dfrac{5}{6}\right)^{k-1}$ $(k = 1, 2, \ldots)$ である．

(2) サイコロを k 回投げて一度も 1 の目が出ないのは，X の実現値が k より大きいときであり，それは k 回連続して 1 以外の目が出る場合なので求める確率は $P(X > k) = \left(\dfrac{5}{6}\right)^k$ となる．

(3) $E[X] = \displaystyle\sum_{k=1}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} k \cdot \frac{1}{6}\left(\frac{5}{6}\right)^{k-1} = \frac{1}{6} \cdot \frac{1}{(1-5/6)^2} = 6$

7.5 平均 $E[X] = \displaystyle\int_0^1 x\,dx$ は，右図の網掛けの三角形の面積だから $\dfrac{1}{2}$．対称な分布だから，平均は中央値 $= \dfrac{1}{2}$ に一致する．積分の計算で求める場合，$E[X] = \displaystyle\int_0^1 x\,dx = \left[\frac{1}{2}x^2\right]_0^1 = \frac{1}{2}(1-0) = \frac{1}{2}$



分散 $V[X] = \displaystyle\int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \left[\frac{1}{3}\left(x - \frac{1}{2}\right)^3\right]_0^1 = \frac{1}{3}\left(\frac{1}{8} + \frac{1}{8}\right) = \frac{1}{12}$

7.6 (1) $\displaystyle\int_{-1}^{1} kx^2\,dx = 1$ より $k = \dfrac{3}{2}$，$P(0 \le X \le \dfrac{1}{2}) = \displaystyle\int_0^{\frac{1}{2}} \frac{3}{2}x^2\,dx = \frac{3}{2}\left[\frac{1}{3}x^3\right]_0^{\frac{1}{2}} = \frac{1}{2}\left(\frac{1}{8} - 0\right) = \frac{1}{16}$

(2) $E[X] = \displaystyle\int_{-1}^{1} x \cdot \frac{3}{2}x^2\,dx = \frac{3}{2}\int_{-1}^{1} x^3\,dx = 0$

$V[X] = \displaystyle\int_{-1}^{1} (x-0)^2 \cdot \frac{3}{2}x^2\,dx = \frac{3}{2}\left[\frac{1}{5}x^5\right]_{-1}^{1} = \frac{3}{5}$

7.7 (1) $P\Big(X > -\dfrac{1}{2}\Big) = 1 - P\Big(X \le -\dfrac{1}{2}\Big) = 1 - F\Big(-\dfrac{1}{2}\Big) = 1 - \Big(\dfrac{1}{2}\Big)^2 = \dfrac{3}{4}$

(2) $f(x) = \dfrac{d}{dx}F(x) = 2(x+1),\ -1 < x < 0$

7.8 (1) $E[X] = 0,\ V[X] = E\big[X^2\big] = \dfrac{5}{2}$

(2) $V\left[\dfrac{X+Y}{2} + 3\right] = \dfrac{1}{2^2}V[X+Y] = \dfrac{1}{2^2}\left(V[X] + V[Y]\right) = \dfrac{1}{2}V[X] = \dfrac{5}{4}$

7.9 (1) $E[X] = \dfrac{3}{2}$万円, $V[X] = \dfrac{25}{8}$ (2) $E[Y] = 1$ 万円, $V[Y] = \dfrac{1}{2}$

(3) $\mathrm{Cov}[X, Y] = -\dfrac{5}{16}$ (4) $E[Z] = \dfrac{5}{4}$万円, $V[Z] = \dfrac{3}{4}$

7.10 $P(X \ge 30) = 1 - \mathsf{BINOM.DIST}(29, 50, 1/4, \mathsf{TRUE}) \fallingdotseq 1.64 \times 10^{-7}$

7.11 $1 - P(X = 0) = 1 - \mathsf{BINOM.DIST}(0, 10, 0.05, \mathsf{FALSE}) \fallingdotseq 0.4$

7.12 正答数 X は $B\left(10,\ \dfrac{1}{3}\right)$ に従う．得点の合計が正となるのは $3X - (10 - X) > 0$ が成り立つときだから，求める確率 $P(X > 2.5) = 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\} = 1 - \mathsf{BINOM.DIST}(2, 10, 1/3, \mathsf{TRUE}) \fallingdotseq 0.701$.

7.13 X は二項分布 $B(100,\ 0.2)$ に従う．$E[X] = 20,\ V[X] = 16$.
$Y = 25 + 0.5X$ と書け，$E[Y] = 25 + 0.5E[X]$，$V[Y] = 0.5^2 V[X]$ が成り立つ．$E[Y] = 35,\ V[Y] = 4$

7.14 (1) $P(1 \le X \le 3) = \dfrac{1}{3}$，$x = 1$ に関して対称な分布だから $E[X] = 1$

$E[X] = \displaystyle\int_{-2}^{4} x \cdot \frac{1}{6}\,dx$ を計算してもよい．

(2) 二項分布 $B\left(3,\ \dfrac{1}{3}\right)$ (3) $P(N = 2) = {}_3\mathrm{C}_2\left(\dfrac{1}{3}\right)^2 \dfrac{2}{3} = \dfrac{2}{9}$

7.15 0.171

7.16 $P(70 < T) = 1 - P(T < 70) \fallingdotseq 0.0228$

7.17 合格点を c とする．$P(X \le c) = 0.3$ を満たす c は Excel の関数 $\mathsf{NORM.INV}(0.3, 45, 9, \mathsf{TRUE})$ で求まり $c \fallingdotseq 40.28$，よって合格点 41 点．

7.18 (1)

k	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	計
$P(\overline{X} = k)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	1

(2) $E\big[\overline{X}\big] = 2,\ V\big[\overline{X}\big] = \dfrac{1}{3}$

7.19 日本に住む 20 歳の男子の胸囲を X とする．X は $N(86.9, 4.80^2)$ に従うから $\overline{X}$ は $N\left(86.9, \dfrac{4.80^2}{16}\right)$ に従う．$P(\overline{X} \le 85) = \mathsf{NORM.DIST}(85, 86.9, 4.80/4, \mathsf{TRUE}) \fallingdotseq 0.0567$

7.20 (1) $P(\overline{X} > 1170) = 1 - \mathsf{NORM.DIST}(1170, 1180, 20/5, \mathsf{TRUE}) \fallingdotseq 0.994$

(2) 条件を式で表すと $P(\overline{X} > 1175) \ge 0.9$ となる．この式は $Z = \dfrac{\overline{X} - 1180}{20/\sqrt{n}}$ とおくと，$P\left(Z > -\dfrac{\sqrt{n}}{4}\right) \ge 0.9$ と変形できる．$\mathsf{NORM.INV}(0.1, 0, 1, \mathsf{TRUE})$

$\fallingdotseq -1.28$ より $P(Z > -1.28) = 0.9$ が成り立つから，$\dfrac{\sqrt{n}}{4} \geq 1.28$（標準正規分布のグラフを描いて考える）より $n \geq (4 \times 1.28)^2 = 26.2144$，よって 27 個以上．

7.21 まず X が $N(\mu, 150^2)$ に従うことから，$\overline{X}$ は $N\left(\mu, \dfrac{150^2}{n}\right)$ に従う．「通過の平均時間が母平均から 30 秒以上はずれる」という事象を式で表すと $|\overline{X} - \mu| \geq 30$ となるから，問題文より $P(|\overline{X} - \mu| \geq 30) = 0.1$ を満たす n の値を求めればよい．この式は $Z = \dfrac{\overline{X} - \mu}{150/\sqrt{n}}$ とおくと，$P\left(|Z| > \dfrac{\sqrt{n}}{5}\right) = 0.1$ と変形できる．$P(Z < z_{0.05}) = 0.05$ を満たす $z_{0.05}$ の値は，NORM.INV(0.95, 0, 1) により求まり，その近似値は -1.645 となる．したがって，$\dfrac{\sqrt{n}}{5} > 1.645$ を満たす最小の n の値を求めればよい．$n > 67.65\cdots$ より，68 人．

7.22 二項分布：$P(4900 \leq T \leq 5100) = \mathrm{BINOM.DIST}(5100, 10000, 0.5, \mathrm{TRUE})$
$-\mathrm{BINOM.DIST}(4900, 10000, 0.5, \mathrm{TRUE}) \fallingdotseq 0.95449$
正規分布：$P(4900 \leq T \leq 5100) = \mathrm{NORM.DIST}(5100, 5000, 50, \mathrm{TRUE})$
$-\mathrm{NORM.DIST}(4900, 5000, 50, \mathrm{TRUE}) \fallingdotseq 0.95450$

7.23 $E[X] = \sum_{x=0}^{n} x \cdot {}_n\mathrm{C}_x\, p^x\, (1-p)^{n-x} = \sum_{x=1}^{n} x \cdot {}_n\mathrm{C}_x\, p^x\, (1-p)^{n-x}$．$t = x - 1$ とおくと，t は $0, 1, 2, \cdots, n-1$ の値をとり，${}_n\mathrm{C}_x = \dfrac{n}{x} \dfrac{(n-1)!}{(n-1-t)!\, t!} = {}_{n-1}\mathrm{C}_t$ が成り立つから，$\sum_{x=1}^{n} x \cdot {}_n\mathrm{C}_x\, p^x\, (1-p)^{n-x} = np \sum_{t=0}^{n-1} {}_{n-1}\mathrm{C}_t\, p^t\, (1-p)^{(n-1)-t}$．$\sum_{t=0}^{n-1} {}_{n-1}\mathrm{C}_t\, p^t\, (1-p)^{(n-1)-t} = \{p + (1-p)\}^{n-1} = 1$ だから $E[X] = np$．

第 8 章

問 8.1 $\mathrm{SE}(\overline{X}) = \dfrac{6.30}{\sqrt{753}} \fallingdotseq 0230.$（$\sqrt{753}$ は Excel の関数 SQRT(753) で計算できる．）

問 8.2 $z_{0.025} = \mathrm{NORM.INV}(0.975, 0, 1) \fallingdotseq 1.96,$
$z_{0.05} = \mathrm{NORM.INV}(0.95, 0, 1) \fallingdotseq 1.645,$
$z_{0.005} = \mathrm{NORM.INV}(0.995, 0, 1, \mathrm{TRUE}) \fallingdotseq 2.576$
標準正規分布用の関数を用いると，例えば $z_{0.025}$ は NORM.S.INV(0.975) で求まる．

問 8.3 (8.5) より信頼限界は $8.59 \pm 1.96 \dfrac{0.69}{\sqrt{100}}$ となるので，μ の信頼係数 95 %の信頼区間は $[8.455,\ 8.725]$.

問 8.4 $t_{0.025}(4) = \mathrm{T.INV}(0.975, 4) \fallingdotseq 2.776, \quad t_{0.05}(4) = \mathrm{T.INV}(0.95, 4) \fallingdotseq 2.132.$

問 8.5 信頼限界 $0.12 \pm 1.645 \dfrac{\sqrt{0.12 \times 0.88}}{\sqrt{200}}$ より，求める信頼区間は $[0.082,\ 0.158]$.

章末問題

8.1 ヒント
(1) $\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (X_i - \mu + \mu - \overline{X})^2 = \sum_{i=1}^{n} [(X_i - \mu)^2 + 2(X_i - \mu)(\mu - \overline{X})$
$+ (\mu - \overline{X})^2] = \sum_{i=1}^{n} (X_i - \mu)^2 + 2n(\overline{X} - \mu)(\mu - \overline{X}) + n(\mu - \overline{X})^2$

(2) $E\left[U^2\right] = \dfrac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2\right]$

$= \dfrac{1}{n-1}\left(\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right] - nE\left[(\overline{X}-\mu)^2\right]\right) = \dfrac{1}{n-1}\left(n\sigma^2 - n\dfrac{\sigma^2}{n}\right)$

8.2 （1）$\overline{x} = 1.0$

（2）$s^2 = 0.8$

（3）連立方程式 $np = 1.0$，$np(1-p) = 0.8$ を解くことより，n の推定値は 5，p の推定値は 0.2 となる．

8.3 p.99 の (7.32) と (7.33) より，$E[X] = np$, $V[X] = np(1-p)$ だから $E[\hat{p}] = E\left[\dfrac{X}{n}\right] = \dfrac{1}{n}E[X] = \dfrac{1}{n}\cdot np = p$，$V[\hat{p}] = V\left[\dfrac{X}{n}\right] = \dfrac{1}{n^2}V[X] = \dfrac{1}{n^2}\cdot np(1-p) = \dfrac{p(1-p)}{n}$．この証明で利用した期待値と分散の性質については，p.95 の (7.17) と (7.18) において，$a = \dfrac{1}{n}$，$b = 0$ として考えよ．

8.4 (1) $N\left(\mu,\ \dfrac{9}{n}\right)$

(2) 信頼限界 $\overline{X} \pm 1.96\dfrac{3}{\sqrt{n}}$ より，区間の幅＝上側限界−下側限界 $= 2\times 1.96\dfrac{3}{\sqrt{n}} = \dfrac{11.76}{\sqrt{n}}$

(3) $\dfrac{11.76}{\sqrt{n}} \leq 1$ より $(11.76)^2 \leq n \iff 138.3 \leq n$　∴ 139 以上にすればよい．

8.5 $\overline{x} = 350.68$（Excel の関数 AVERAGE を用いる）だから信頼限界 $350.86 \pm 1.96\dfrac{0.10}{\sqrt{10}}$ より，求める信頼区間 [350.80, 350.92].

8.6 信頼限界 $115{,}000 \pm 1.645\dfrac{35{,}000}{\sqrt{100}}$ より，求める信頼区間 [109,243, 120,758].

8.7 信頼限界 $158.30 \pm 2.576\dfrac{5.12}{\sqrt{770}}$ より，求める信頼区間 [157.83, 158.78].

8.8 信頼限界 $0.48 \pm 1.96\dfrac{\sqrt{23.17}}{\sqrt{240}}$ より，求める信頼区間 [−0.13, 1.09].

8.9 $\overline{x} = 13.39$, $u \fallingdotseq 1.003$（Excel の関数 STDEV.S を用いる）また自由度 9 の t 分布の上側 2.5 %点 $t_{0.025}(9) \fallingdotseq 2.262$ だから，信頼限界 $13.39 \pm 2.262\dfrac{1.003}{\sqrt{10}}$ より，信頼係数 95 %の信頼区間 [12.67, 14.11]．同様の考え方で自由度 9 の t 分布の上側 5 %点 $t_{0.05}(9) \fallingdotseq 1.833$ より，信頼係数 90 %の信頼区間 [12.81, 13.97]．2 つの信頼区間に大差はない．

8.10 自由度 12 の t 分布の上側 2.5 %点 $t_{0.025}(12) \fallingdotseq 2.179$ だから，信頼限界 $96.3 \pm 2.179\dfrac{1.8}{\sqrt{13}}$ より，信頼係数 95 %の信頼区間 [95.21, 97.39].

8.11 信頼限界 $0.15 \pm 1.96 \times \sqrt{\dfrac{0.15(1-0.15)}{500}}$ より，求める信頼区間 [0.119, 0.181]，目標値 0.20 はこの信頼区間の外側にある．

8.12 (1) 母集団は無限母集団とみなせ，標本サイズ 100 も大きい．したがって (8.10) を用いることができる．信頼限界 $0.4 \pm 1.96 \times \sqrt{\dfrac{0.4\times 0.6}{100}}$ より，[0.304, 0.496]

(2) 求める標本サイズを n とすると，条件は $1.96 \times \sqrt{\dfrac{0.4\times 0.6}{n}} = 0.03$ と表すことができる．この式の両辺を 2 乗し，n について解くことから $n = 1025$ 以上．

8.13 (1) $E[X_i] = \mu + E[\varepsilon_i] = \mu$, $V[X_i] = V[\varepsilon_i] = \sigma^2$.
(2) 29 個の測定値の平均と不偏分散は，それぞれ $\overline{x} \fallingdotseq 5.482$，$u^2 \fallingdotseq 0.0427$．自由度 28 の t 分布の上側 2.5 %点 $t_{0.025}(28) \fallingdotseq 2.048$ だから，信頼限界 $5.482 \pm 2.048\dfrac{0.2067}{\sqrt{29}}$ より信頼係数 95 %の信頼区間 [5.40, 5.56]．なお，現在知られている地球の密度は $5.517\,\mathrm{g/cm^3}$ である．

第 9 章

問 9.1 P 値 $= 0.0013$ は 1 %より小さいので，H_0 は有意水準 1 %で棄却される．

問 9.2 P 値$= 2 \times P(T \geq 3.076) \fallingdotseq 0.0082$.
標本サイズ n が大きくなると P 値は小さくなることに注意せよ．

問 9.3 標本サイズ $n = 450$ は十分に大きいので，H_0 のもとで $Z = \dfrac{\sqrt{n}(\overline{X} - 157.9)}{5.33}$ は標準正規分布 $N(0,\ 1)$ に従うとみなせる．

$$\text{検定統計量 } Z \text{ の実現値} = \frac{\sqrt{450}(156.8 - 157.9)}{5.33} \fallingdotseq -4.378.$$
$$\text{P 値}= 2 \times P(Z \leq -4.378) \fallingdotseq 1.20 \times 10^{-5}.$$

この値は有意水準 1 %より小さいので H_0 は有意水準 1 %で棄却される．

問 9.4 母分散 σ^2 の推定値 $u^2 = \dfrac{9(2.841 + 2.835)}{18} \fallingdotseq 2.838$ より，

$$T \text{ の実現値} = \frac{24.39 - 21.68}{\sqrt{2.838(\frac{1}{10} + \frac{1}{10})}} \fallingdotseq 3.597.$$
$$\text{P 値}= 2 \times P(T \geq 3.597) = 2\big(1 - \mathsf{T.DIST}(3.597, 18, \mathsf{TRUE})\big) \fallingdotseq 0.002$$

となり有意水準 1 %より小さいので H_0 は棄却される．

問 9.5 T の実現値 $= \dfrac{\sqrt{8} \times (-0.8)}{\sqrt{0.437}} \fallingdotseq -3.423$.
P 値$= P(T \leq -3.423) = \mathsf{T.DIST}(-3.423, 7, \mathsf{TRUE}) \fallingdotseq 5.55 \times 10^{-3}$
となり有意水準 5 %より小さいので $\mathrm{H}_0 : \delta = 0$ は棄却される．このデータからは効果があったといえる．

問 9.6 P 値$= P(\chi^2 \geq 4.18) = 1 - \mathsf{CHISQ.DIST}(4.18, 5, \mathsf{TRUE}) \fallingdotseq 0.524$．この値は有意水準 5 %より大きいから帰無仮説は棄却されない．したがって，実験データからはこのサイコロが正しいサイコロではないとはいえない．

サイコロの目	1	2	3	4	5	6	計
観測度数	26	38	38	28	37	33	200
期待度数	33.3	33.3	33.3	33.3	33.3	33.3	200

問 9.7 P 値$= P(\chi^2 \geq 8.249) = 1 - \mathsf{CHISQ.DIST}(8.249, 1, \mathsf{TRUE}) \fallingdotseq 0.004$ は有意水準 1 %より小さいので関連がないという仮説は棄却される．このデータからは関連性がないとはいえない．

章末問題

9.1 二項母集団を想定し，その母集団における A を選んだ人の割合 p をとする．
$\mathrm{H}_0 : p = 0.5$（差がない），$\mathrm{H}_1 : p \neq 0.5$（差がある）
検定統計量 $X =$ A を選んだ人の数．H_0 のもとで X は二項分布 $B(12, 0.5)$ に従う．
P 値$= 2 \times P(X \geq 11) = 2(1 - \mathsf{BINOM.DIST}(10, 12, 0.5, \mathsf{TRUE})) \fallingdotseq 0.0063$ は 0.01 より小さいので，有意水準 1 %で H_0 は棄却され，このデータからは差があると判断できる．

9.2 (1) $\mathrm{H}_0 : \mu = 7.3,\ \mathrm{H}_1 : \mu < 7.3$

(2) $\overline{x} \fallingdotseq 6.97,\ u \fallingdotseq 0.957$ より T の実現値 $= \dfrac{\sqrt{15}(6.97-7.3)}{0.957} \fallingdotseq -1.34$. P 値$= P(T \leq -1.34) = \mathsf{T.DIST}(-1.34, 14, \mathsf{TRUE}) \fallingdotseq 0.101$ となり，H_0 は有意水準 5 %で棄却されない．

9.3 H_0 のもとで，検定統計量 $T = \dfrac{\sqrt{n}(\overline{X}-0)}{U}$ は自由度 107 の t 分布に従う．T の実現値 $= \dfrac{\sqrt{108}(0.714-0)}{4.704} \fallingdotseq 1.577$. P 値$= 2 \times P(T \geq 1.577) \fallingdotseq 0.118$ となり，有意水準 5 %より大きいので H_0 は有意水準 5 %で棄却されない．

次のように考えてもよい．$n = 108$ は大きいので，H_0 のもとで $Z = \dfrac{\sqrt{n}(\overline{X}-0)}{4.704}$ は標準正規分布 $N(0,\ 1)$ に従う．検定統計量 Z の実現値 $= \dfrac{\sqrt{108}(0.714-0)}{4.704} \fallingdotseq 1.577$. P 値$= 2 \times P(Z \geq 1.577) \fallingdotseq 0.115$ となり，有意水準 5 %より大きいので H_0 は有意水準 5 %で棄却されない．

9.4 H_0 のもとで, $T = \dfrac{\overline{X}}{U/\sqrt{5}}$ は自由度 4 の t 分布に従う．$\overline{x} = 1, u \fallingdotseq 2.55$ より, T の実現値 $\fallingdotseq 0.877$. P 値$= P(T \geq 0.877) = 1 - \mathsf{T.DIST}(0.877, 4, \mathsf{TRUE}) \fallingdotseq 0.215$ となり，有意水準 0.05 より大きいので帰無仮説は棄却されない．ゆえに，このデータからはこのプログラムは役立ったとはいえない．

9.5 $\mathrm{H}_0 : p = 0.2,\ \mathrm{H}_1 : p < 0.2$（目標を達成していない）の検定．$\overline{x} = 0.15$ だから，H_0 のもとで，検定統計量 Z の実現値 $= \dfrac{0.15-0.2}{\sqrt{\frac{0.2(1-0.2)}{500}}} \fallingdotseq -2.8$. P 値$= P(Z \leq -2.8) \fallingdotseq 0.0026$ となり，帰無仮説は有意水準 0.05 で棄却される．このデータからは目標視聴率を達成できているとはいえない．

9.6 H_0 のもとで検定統計量 $T = \dfrac{\overline{X}-\overline{Y}}{\sqrt{U^2(\frac{1}{100}+\frac{1}{120})}}$ は自由度 218 の t 分布に従う．$u^2 = \dfrac{99 \cdot 0.56^2 + 119 \cdot 0.53^2}{218} \fallingdotseq 0.2958$ より，T の実現値 $\fallingdotseq 2.173$. P 値$= P(T \geq 2.173) = 1 - \mathsf{T.DIST}(2.173, 218, \mathsf{TRUE}) \fallingdotseq 0.015$ となり有意水準 5 %より小さいので H_0 は棄却される．20 年前より速いといってよい．

〈大標本での方法〉 $m = 100,\ n = 120$ は大きいので，H_0 のもとで検定統計量 $Z = \dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{{u_1}^2}{100}+\frac{{u_2}^2}{120}}}$ は $N(0,\ 1)$ に従う．Z の実現値 $= \dfrac{7.42-7.26}{\sqrt{\frac{0.56^2}{100}+\frac{0.53^2}{120}}} \fallingdotseq 2.162$. P 値$= P(Z \geq 2.162) = 1 - \mathsf{NORM.DIST}(2.162, 0, 1, \mathsf{TRUE}) \fallingdotseq 0.015$ となり有意水準 5 %より小さいので H_0 は棄却される．

9.7 帰無仮説 $\mathrm{H}_0 : P(\mathrm{A}) = \dfrac{4}{10},\ P(\mathrm{B}) = \dfrac{2}{10},\ P(\mathrm{O}) = \dfrac{3}{10},\ P(\mathrm{AB}) = \dfrac{1}{10}$ のもとで，検定統計量 χ^2 は近似的に自由度 3 の χ^2 分布に従う．χ^2 の実現値 $= 32.15$. P 値$= P(\chi^2 \geq 32.15) = 1 - \mathsf{CHISQ.DIST}(32.15, 3, \mathsf{TRUE}) \fallingdotseq 4.87 \times 10^{-7}$．この値は有意水準 1 %より小さいから H_0 は棄却される．このデータからは，イギリス人の血液型の分布は日本人のそれと異なるといえる．

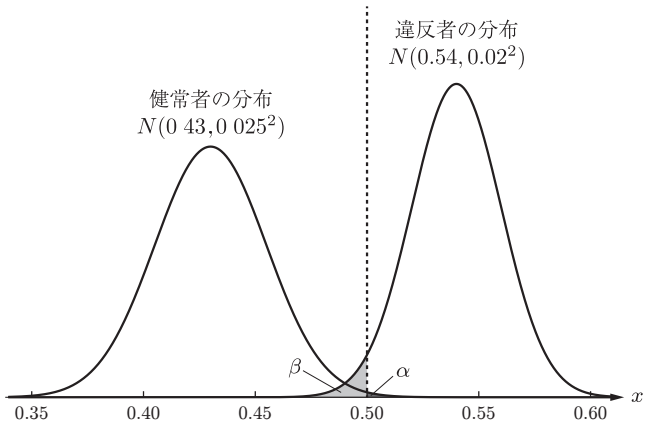
9.8 P 値$= P(\chi^2 \geq 14.33) = 1 - \mathsf{CHISQ.DIST}(14.33, 2, \mathsf{TRUE}) \fallingdotseq 0.0008$. この値は有意水準 5 %より小さい．適合しているという仮説は有意水準 5 %で棄却される．

9.9 P 値$= P(\chi^2 \geq 22.50) = 1 - \mathsf{CHISQ.DIST}(22.50, 1, \mathsf{TRUE}) \fallingdotseq 2.10 \times 10^{-6}$．この値は有意水準 1 %より小さい．差はないという仮説は有意水準 1 %で棄却される．

9.10 P 値= $P(\chi^2 \geq 32.98) = 1 - \mathsf{CHISQ.DIST}(32.98, 2, \mathsf{TRUE}) \fallingdotseq 6.89 \times 10^{-8}$．この値は有意水準 5 %より小さい．関連がないという仮説は有意水準 5 %で棄却される．p.151 の 補 参照．

9.11 P 値= $P(\chi^2 \geq 6.107) = 1 - \mathsf{CHISQ.DIST}(6.107, 4, \mathsf{TRUE}) \fallingdotseq 0.191$．この値は有意水準 5 %より大きい．関連がないという仮説は有意水準 5 %で棄却されない．

9.12 H 値を X とする．α は $N(0.43, 0.025^2)$ のもとで確率 $P(0.5 < X)$ を，β は $N(0.54, 0.02^2)$ のもとで確率 $P(X < 0.5)$ を計算すればよい．Z を標準正規分布に従う確率変数とするとき，$\alpha = P(2.8 < Z) \fallingdotseq 0.00256$，$\beta = P(Z < -2) \fallingdotseq 0.0228$．



第 10 章

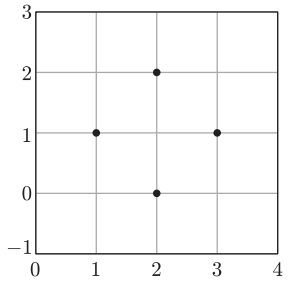
問 10.1 Excel では，(10.2) で与えられる共分散 s_{xy} は関数 COVARIANCE.P，相関係数 r は関数 CORREL を用いて求める． $s_{xy} \fallingdotseq 7,831,804$， $r \fallingdotseq 0.985$, $s_y \fallingdotseq 10177$

問 10.2 回帰直線 $y = 266.83 + 12.826x$

問 10.3 $R^2 = (\text{相関係数})^2 \fallingdotseq 0.978^2 \fallingdotseq 0.956$，求めた回帰直線は電力消費量をよく説明しているといえる．

章末問題

10.1 $\overline{x} = 2$, $\overline{y} = 1$ だから， 共分散 $= \dfrac{1}{4}\big\{(1-2)(1-1) + (2-2)(0-1) + (2-2)(2-1) + (3-2)(1-1)\big\} = 0$ より 相関係数 $= 0$.



10.2 p.158 の (10.3) より $r = \dfrac{1}{n}\sum_{i=1}^{n} u_i v_i$ だから $\dfrac{1}{n}\sum_{i=1}^{n}\Big(u_i \pm v_i\Big)^2 = \dfrac{1}{n}\sum_{i=1}^{n} {u_i}^2 \pm$

$\dfrac{2}{n}\sum_{i=1}^{n} u_i v_i + \dfrac{1}{n}\sum_{i=1}^{n} {v_i}^2 = 1 \pm 2r + 1 = 2(1 \pm r)$ と書ける．$\dfrac{1}{n}\sum_{i=1}^{n}\Big(u_i \pm v_i\Big)^2$ は常に 0 以上だから，$2(1 \pm r) \geq 0$．この不等式を解いて $-1 \leq r \leq 1$．

10.3 児童の場合，年齢が上がるにつれて身長は伸びるし 50 m 走は速くなるため見かけ上の相関が生じる．学年で層別し，それぞれの層（学年）のデータで相関関係を調べる必要がある．

10.4 (1) 点 (x_i, y_i) はすべて右下がりの直線上にあるので相関係数は -1．
(2) 点 (y_i, w_i) はすべて右上がりの直線上にあるので相関係数は 1．
(3) 平均点 w によって受験生を良いほうから並べると edcba となり，評点 y によって順位づけした場合と同じになり，甲の評価が全く反映されない結果になる．
(4) 甲の評点のばらつきにくらべ，乙のそれがだいぶ大きいことが 1 つの原因である．
注意：いくつかの異なるテストの得点や評点を比較したり合計したりする場合には，原則として，相互の平均や標準偏差が等しくなるように調整しておくことが望ましい．
(5) 64, 57, 50, 43, 36
(6) 36, 43, 50, 57, 64
(7) 全員の評点が 50 となり，差がつかなくなる．

10.5 $\dfrac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \dfrac{1}{n}\sum_{i=1}^{n}(\widehat{\alpha} + \widehat{\beta}\, x_i) = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\alpha} + \widehat{\beta}\dfrac{1}{n}\sum_{i=1}^{n} x_i = \widehat{\alpha} + \widehat{\beta}\,\overline{x} \overset{(10.8)\text{ より}}{=} \overline{y}$

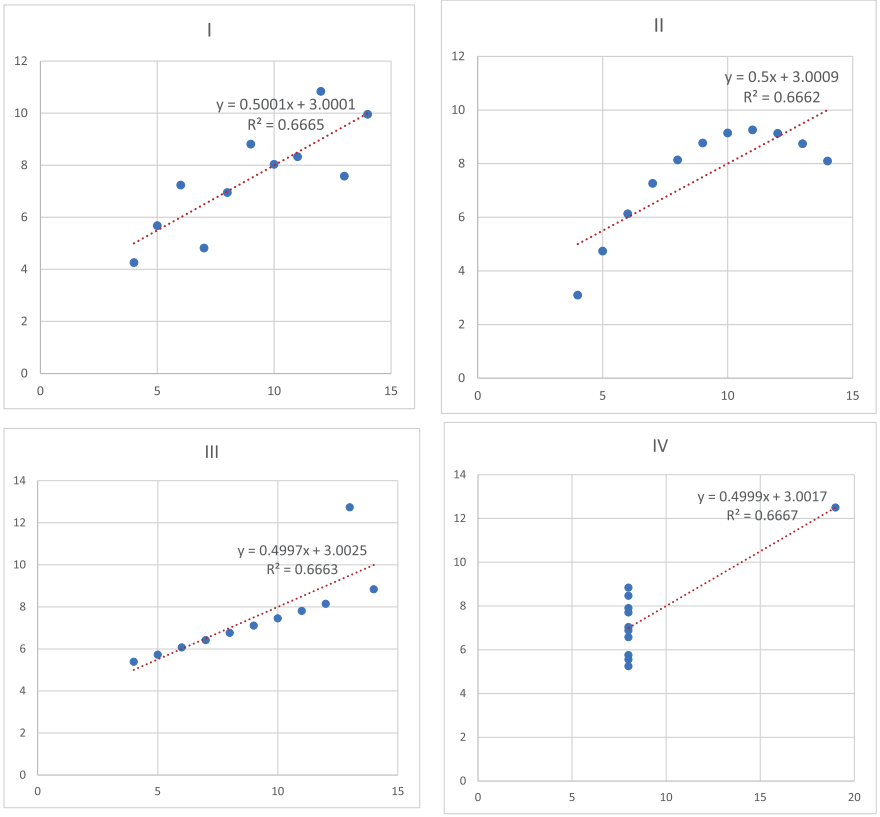
10.6 総変動平方和 $= S_T$, 回帰平方和 $= S_R$, 残差平方和 $= S_e$ とおくと
$S_T = S_R + S_e \geq S_R \geq 0$ が成り立つから，$S_T > 0$ のとき各辺を S_T で割ることから $1 \geq \dfrac{S_R}{S_T} \geq 0$，$\dfrac{S_R}{S_T} = R^2$ だから $0 \leq R^2 \leq 1$．$R^2 = 1$ となるのは $S_R = S_T$，すなわち実測値 y_i の変動が予測値 $\widehat{y_i}$ の変動で完全にとらえられる場合であり，散布図のすべての点が一直線上にあるときである．

10.7 (1) 算術平均 $\overline{x} = 15$, $\overline{y} = 22.75$
(2) 標準偏差 $s_x \fallingdotseq 3.42$, $s_y \fallingdotseq 1.787$
(3) 相関係数 $r \fallingdotseq 0.942$
(4) $R^2 \fallingdotseq 0.888$
(5) y の x への回帰直線 $y = 15.36 + 0.493x$
(6) 24.73 cm

10.8 (1) 算術平均 $\overline{x} = 147.3$, $\overline{y} = 22.1$
(2) 標準偏差 $s_x \fallingdotseq 8.36$, $s_y \fallingdotseq 1.45$
(3) 身長と前腕の長さの共分散 $s_{xy} \fallingdotseq 10.27$
(4) 身長と前腕の長さの相関係数 $r \fallingdotseq 0.85$
(5) y の x への回帰直線 $y = 0.43 + 0.147x$
(6) $R^2 \fallingdotseq 0.723$
(7) $y \fallingdotseq 22.5$

10.9 (1) 4 つのすべてのデータセットに対して，$\overline{x} = 9$, $\overline{y} \fallingdotseq 7.50$, ${u_x}^2 = 11$, ${u_y}^2 \fallingdotseq 4.12$.
(2) 4 つのすべてのデータセットに対して，$r \fallingdotseq 0.816$, 回帰直線 $y = 3.00 + 0.500x$.

(3)



補　章

問 A.1 $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

問 A.2 (1) $\sum_{i=1}^{n}(cx_i + d) = c\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} d = c\sum_{i=1}^{n} x_i + nd$. この式の両端の式を n で割ることから示せる.

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

問 A.3 $\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) + \cdots + (x_n - \overline{x})$
$= (x_1 + x_2 + x_3 + \cdots + x_n) - n\overline{x} = n\overline{x} - n\overline{x} = 0$

問 A.4 (1) $1 + 2^2 + 3^2 + 4^2 + 5^2 = 55$

(2) $1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 + 19 = 100$

問 A.5 $P(X \geq 5) = 1 - P(X \leq 4) = 1 - \mathsf{POISSON.DIST}(4, 2.5, \mathsf{TRUE}) \fallingdotseq 0.109.$

問 A.6 例えば, $P(X=1)P(Y=1) = \dfrac{12}{20}\cdot\dfrac{12}{20} = \dfrac{9}{25} \neq \dfrac{6}{20} = P(X=1,\ Y=1)$ となるので独立ではない.

章末問題

A.1 $P(X > b) = \displaystyle\int_b^{\infty} \lambda e^{-\lambda x}\, dx = e^{-\lambda b}$, $P(X > a+b) = \displaystyle\int_{a+b}^{\infty} \lambda e^{-\lambda x}\, dx$
$= e^{-\lambda(a+b)}$ だから, $P(X > a+b \mid X > b) = \dfrac{P(X > a+b)}{P(X > b)} = \dfrac{e^{-\lambda(a+b)}}{e^{-\lambda b}}$

$= e^{-\lambda a}$. 一方, $P(X > a) = \displaystyle\int_a^\infty \lambda e^{-\lambda x}\, dx = e^{-\lambda a}$.

よって, $P(X > a+b \mid X > b) = P(X > a)$ が成り立つ.

A.2 (1) 平均 $\overline{x} = 1.0$, 分散 $s^2 \fallingdotseq 0.97$

(2) 車の台数 k の理論度数は, Excel の関数 POISSON.DIST(k , 1.0, FALSE)∗197 で求まる.

車の台数 (10 秒間ごとの)	0	1	2	3	4	計
観測度数	72	73	36	12	4	197
理論度数	72.47	72.47	36.24	12.08	3.02	196.28

A.3 (1) $E[X] = \dfrac{2}{3}$, $E[Y] = 1$

(2) $(x - \dfrac{2}{3})(y-1)P(X = x,\ Y = y)$ の値が, $(x,\ y) = (1,\ 0),\ (1,\ 2)$ 以外では 0 になるので, $\mathrm{Cov}[X,Y] = E\left[\left(X - \dfrac{2}{3}\right)(Y-1)\right] = \left(1 - \dfrac{2}{3}\right) \times (-1) \times \dfrac{1}{3} + \left(1 - \dfrac{2}{3}\right) \times (2-1) \times \dfrac{1}{3} = 0.$

(3) $P(X = 0,\ Y = 0) = 0$, 一方 $P(X=0)P(Y=0) = \dfrac{1}{3} \times \dfrac{1}{3} = \dfrac{1}{9}$.

(4) (2) の結果より, $\mathrm{Cov}[X,Y] = 0$ だから X と Y は無相関であるが, (3) の結果より, X と Y は独立ではない.

A.4 $T = 2X - 3Y$ とおくと, $E[T] = 2E[X] - 3E[Y] = -5$, $V[T] = 2^2 V[X] + 3^2 V[Y] = 13$ だから, T は正規分布の再生性 (p.194) により正規分布 $N(-5,\ 13)$ に従う. 求める確率は $P(T \geq 0) = 1 - \mathsf{NORM.DIST}(0, -5, \mathsf{SQRT}(13), \mathsf{TRUE}) \fallingdotseq 0.083$.

A.5 (1) $N(100,\ 0.1)$ に従う. (2) $N(100,\ 0.01)$

(3) 求める測定回数を n とすると, $\overline{X}$ は $N\left(100,\ \dfrac{0.1}{n}\right)$ に従うから

$Z = \dfrac{\sqrt{n}(\overline{X} - 100)}{\sqrt{0.1}}$ は $N(0,\ 1)$ に従う.

$P(|\overline{X} - 100| < 0.1) = P\left(\dfrac{\sqrt{0.1}}{\sqrt{n}}|Z| < 0.1\right) = P(|Z| < \sqrt{0.1n}) \geq 0.95$

したがって $\sqrt{0.1n} \geq 1.96$ より $\quad n \geq \dfrac{1.96^2}{0.1} \fallingdotseq 38.4. \quad \therefore$ 39 回

A.6 対数の性質：$\log uv = \log u + \log v$, $\log \dfrac{u}{v} = \log u - \log v$, $\log u^p = p \log u$ （ただし, $u > 0$, $v > 0$, p は実数）を使って変形する.

A.7 (1) 略.

(2) $x_1, x_2, \ldots, x_n$ の最大値を $x_{(n)}$ とするとき

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n} & (x_{(n)} \leq \theta) \\ 0 & (\text{その他}) \end{cases}$$

(3) θ が値 $x_{(n)}$ をとるとき $L(\theta)$ は最大になるから, $\widehat{\theta} = X_{(n)}$

A.8 $\displaystyle\int_0^1 p \cdot \pi(p \mid x)\, dp = \int_0^1 p \cdot \frac{1}{B(16,17)} p^{15}(1-p)^{16}\, dp$

$= \dfrac{1}{B(16,17)} \displaystyle\int_0^1 p^{16}(1-p)^{16}\, dp = \dfrac{B(17,17)}{B(16,17)} = \dfrac{16}{33}$

ただし, $B(\cdot\,,\,\cdot)$ はベータ関数で, a, b が 1 以上の整数のとき

$B(a,b) = \dfrac{(a-1)!\,(b-1)!}{(a+b-1)!}$ と計算される.

A.9 $(x_i - x_j)^2 = \{(x_i - \overline{x}) - (x_j - \overline{x})\}^2 = (x_i - \overline{x})^2 - 2(x_i - \overline{x})(x_j - \overline{x}) + (x_j - \overline{x})^2$ と変形し，$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = ns^2$，$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x})(x_j - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})\sum_{j=1}^{n}(x_j - \overline{x}) = 0$ となることを使い，目的の式を変形すると s^2 になる．この結果から分散 s^2 は，x_i，x_j $(i,\ j = 1, 2, \ldots, n)$ のすべての組み合わせの差 $(x_i - x_j)^2$ の平均であることがわかる．