『Rで学ぶ統計解析』演習問題略解

2　記述統計学

問題 2.1　C 君の母親は，医師からの勧めに従って，毎朝きまった時刻に血圧を測定して記録しているが，その記録をみながら「血圧が毎日変動して不安定だ」といって心配している．統計解析を勉強している C 君は，1 日の同じ時刻に繰り返し測定しても，ある程度の変動は避けられないのではないかと考えて，5 回繰り返して測定してみたところ，次の測定値が得られた (血圧の単位は mmHg).
(1) 最高血圧と最低血圧のそれぞれについて，中央値，平均値，標準偏差を求めよ．

最高血圧	116	128	120	116	118
最低血圧	71	70	72	68	69

(2) 後の章で学習する理論によれば，測定値が 平均値± 2 × (標準偏差) 程度の範囲で変動している場合には，特に心配しなくてもよいといわれている．C 君は，母親にどのようにアドバイスしたらよいか考えよ．

解答　**(1)** `saikou<-c(116,128,120,116,118)` として，`median(saikou)`, `mean(saikou)`, `sd(saikou)` により，最高血圧の中央値=118, 平均値=119.6, 標準偏差=4.98 という計算結果が得られる．最低血圧についても同様にして，中央値＝ 70, 平均値＝ 70, 標準偏差＝ 1.58 となる．

(2) 平均値± 2 × (標準偏差) を計算すると，最高血圧についてはおよそ [110, 130], 最低血圧についてはおよそ [67, 73] となる．したがって，最高血圧，最低血圧がそれぞれこの範囲に入っていれば一喜一憂しないようにアドバイスするのがよい．

問題 2.2　次の表は．OECD(経済協力開発機構) 加盟国の 2008 年の GDP(国内総生産；名目) および 1 人あたり GDP を示したものである (通貨の単位は米ドル，出典：(財) 矢野恒太記念会，『日本国勢図会 2010/2011 年版』，2010).

(1) この表では国名は GDP の降順に並んでいるが，1 人あたりの GDP の降順に並べてみよ．日本は何番目になるか？

(2) GDP および 1 人あたりの GDP それぞれについて，ヒストグラムを描き，また平均，標準偏差，四分位値，などを求めよ．

(3) GDP のパレート図を描け．

解答　**(1)** `GDP<-c(143694,48997,…,168)`, `GDP1<-c(47186,38371,…,52568)` として与えられたデータを入力する．1 人あたりの GDP を降順に並べるには，たとえば `rev(sort(GDP1))` とすればよい．日本は 19 番目となる．

参考：`rank(GDP1)` とすれば，1 人あたりの GDP の“昇順”に並べた場合の各国の順番が求められ，日本が 12 番目であることがわかる．

(2) ヒストグラムは `hist(GDP)`, `hist(GDP1)` によって描ける．平均値などは `summary(GDP)`, `summary(GDP1)` によって求められる．

(3) 略

問題 2.3　第 9 章の例 9.1 には，都道府県別の交通事故件数と車両保有台数のデータが示されている．このデータについて，散布図を描き，相関係数を求めよ．

	GDP(億ドル)	1 人あたり (ドル)
アメリカ合衆国	143,694	47,186
日本	48,997	38,371
ドイツ	36,559	44,519
フランス	28,565	44,550
イギリス	26,526	43,237
イタリア	23,031	38,455
スペイン	15,945	34,971
カナダ	14,996	44,950
メキシコ	10,852	10,183
オーストラリア	10,337	48,049
韓国	9,291	19,115
オランダ	8,729	53,094
トルコ	7,300	10,270
ポーランド	5,283	13,861
ベルギー	5,049	47,151
スイス	5,003	64,885
スウェーデン	4,790	51,954
ノルウェー	4,518	94,763
オーストリア	4,129	49,527
ギリシャ	3,503	31,174
デンマーク	3,408	62,054
フィンランド	2,706	50,931
アイルランド	2,663	59,944
ポルトガル	2,436	22,929
チェコ	2,161	20,719
ハンガリー	1,542	15,363
ニュージーランド	1,278	29,693
スロバキア	950	17,566
ルクセンブルグ	576	117,967
アイスランド	168	52,568

解答 交通事故件数を `kensu` に，また車両保有台数を `daisu` に入力して，`plot(daisu,kensu)` によって散布図を描く．`cor(daisu,kensu)` によって相関係数を計算すると，およそ 0.87 となり，かなり強い正の相関がある．

問題 2.4　次のデータは，2010 年 (平成 22 年)1 年間におけるわが国の空港の乗降客数 (国内線と国際線の合計) を示すものである (出典：国土交通省ホームページ『空港管理状況調書』；単位は千人；年間の乗降客数が 1000 人未満の空港は省いてある．また，乗降客数が 1000 人以上で 10 万人未満の空港については，空港名を省略し，該当空港数のみを示す)．

(1) 乗降客数の降順に空港を並べて，ランキングを作成せよ．

(2) 乗降客数の階級区分を工夫して，空港数のヒストグラムを作成せよ．

空港	乗降客数	空港	乗降客数	空港	乗降客数
成田国際	30,780	鹿児島	4,968	出雲	702
中部国際	9,271	那覇	14,526	岡山	1,348
関西国際	14,220	旭川	1,198	佐賀	340
東京国際	64,221	帯広	552	対馬	264
新千歳	16,748	秋田	1,062	福江	130
稚内	169	山形	163	屋久島	155
釧路	709	山口宇部	796	奄美	527
函館	1,582	中標津	169	徳之島	149
仙台	2,826	女満別	711	久米島	238
新潟	942	青森	1,031	宮古	1,093
大阪国際	14,789	花巻	285	石垣	1,693
広島	2,810	大館能代	122	札幌	205
高松	1,423	庄内	356	三沢	263
松山	2,381	福島	277	百里	145
高知	1,273	八丈島	195	小松	2,118
福岡	16,345	富山	949	美保	469
北九州	1,200	能登	155	徳島	808
長崎	2,331	静岡	593	名古屋	445
熊本	2,875	神戸	2,224		
大分	1,536	南紀白浜	124		
宮崎	2,683	鳥取	305	その他 24 空港	各 10 万人未満

解答　**(1)** 乗降客数を `kyakusu` に入力して，`rev(sort(kyakusu))` とすれば，降順に並べられる．

ランキングは，東京国際，成田国際，新千歳，福岡，大阪国際，那覇，……，福江，南紀白浜，大館能代となる．

(2) 単に `hist(kyakusu)` とすれば，階級の幅が 10000 となり，大多数の空港が乗降客数 10000 以下の階級に入ってしまう．そこで，たとえば乗降客数の常用対数をとって，`hist(log10(kyakusu))` としてみると，もっと適切なヒストグラムが得られる．

問題 2.5　次の数値は，2010 年 7 月の毎日，気象庁 (東京都千代田区) において観測されたその日の最高気温である (出典：気象庁ホームページ；単位は °C)．

一方，下記のデータは，同じ期間の毎日午後 2 時台の東京電力管内の電力消費量 (1 時間当たりの

29.5, 31.0, 28.8, 31.6, 30.3, 30.6, 27.8, 30.1, 27.8, 31.1,
28.9, 29.1, 27.2, 31.3, 31.4, 31.9, 32.1, 31.7, 34.5, 34.5,
36.3, 36.1, 35.7, 35.8, 34.4, 33.3, 33.8, 34.2, 27.9, 29.2, 32.2

平均値) を示したものである (出典：東京電力ホームページ；単位は万 kW)．

(1) これらのデータに対して散布図を描き，また相関係数を計算せよ．

(2) 上記の 2 つのホームページには，他の期間，他の種類の多くのデータが掲載されている．それらのデータの中から興味あるものを抽出して，同様の解析を試みよ．

4900, 4916, 4049, 4066, 4984, 5022, 4716, 4916, 4640, 4146,
3725, 4689, 4367, 4679, 4899, 5249, 4492, 4228, 4793, 5726,
5918, 5965, 5999, 5213, 4715, 5550, 5666, 5596, 4659, 4953, 4712

解答　**(1)** 最高気温を kion, 電力消費量を denryoku に入力して，plot(kion,denryoku) によって散布図を描く．相関係数は cor(kion,denryoku) によって計算すると，およそ 0.68 となり，相関はそれほど強くはない．これは，気温は東京都千代田区で測定しているのに対して，東京電力管内ははるかに広い地域に及んでいることなどを反映しているのであろう．

(2) 略

問題 2.6　ある大学では，サンプリング調査を行って，学生生活の実態を調べた．調査表を配布したのは，男子・女子とも学部在学生総数の 25% であり，回答率は男子 40%，女子 50% で，十分多数の回答が得られた．調査項目のうちで，学生の家庭の年収額については，調査結果が次のように整理されて報告されている (金額の単位は十万円；数値は，男子，女子それぞれの回答者数に対する百分率)．

	45 未満	45〜75	75〜95	95〜105	105〜125	125〜155	155 以上
男子	17.6	17.3	14.4	17.1	10.6	10.5	12.5
女子	11.8	16.8	16.3	15.2	11.3	9.0	19.6

(1) 男子，女子それぞれの家庭の年収額分布について，累積相対度数 (%) を求め，図示せよ．2 つの図を比較して，どのようなことがわかるか述べよ．

(2) 男子，女子それぞれのデータについて，平均値，標準偏差を求めたいが，上記の数値だけでは計算できない．不足している数値は何か？　それらの数値を適当に補って，計算せよ．

解答　**(1)** 累積相対度数は cumsum 関数を使って計算できる．男子，女子についてこれらを (階段状のグラフとして) 図示すると，男子のグラフがいたるところで女子のグラフの上側にあり，女子のほうが年収額の多い家庭の出身者が多いことがわかる．

(2) 平均値，標準偏差は式 (2.2), (2.6) を使って計算することになるが，そのためにはすべての階級の階級値が必要になる．両端を除く階級の階級値は，それぞれの階級の中央の値とすればよいが，45 未満および 155 以上の階級については，階級値をどう選べばよいかが自明でない．特に根拠はないが，たとえば，45 未満の階級の階級値は 40, 155 以上の階級の階級値は 180 などとする．

問題 2.7　近年の日本では，少子高齢化が進んでいるといわれている．しかし，それには地域差があるのではないかと思われる．そこで，一例として，秋田県と愛知県の「年齢 5 歳階級別人口」(国勢調査による平成 17 年 10 月 1 日現在の数値；単位は千人) を取り上げて，解析してみよう．

(1)「65 歳以上の人口が県全体の人口の中で占める割合」を「高齢化指数」とする定義がある．この定義に従って，両県の高齢化指数を計算せよ．

(2) 上記の指数には，「少子化」の程度が反映されているとはいえない．少子化の程度も反映されるものとして，累積相対度数が考えられる．そこで，両県の人口分布の累積相対度数を求め，図示せよ．これらを比較して，どのようなことがわかるか述べよ．

解答　**(1)** 高齢化指数は，秋田：0.269, 愛知：0.173 となり，秋田県のほうが高齢化が相当進んでいることがわかる．

−	0〜4	5〜9	10〜14	15〜19	20〜24	25〜29	30〜34	35〜39	40〜44
秋田	41	48	54	55	49	59	65	61	68
愛知	354	366	349	378	443	511	618	541	478
−	45〜49	50〜54	55〜59	60〜64	65〜69	70〜74	75〜79	80 歳以上	
秋田	77	91	95	75	79	81	70	78	
愛知	418	459	568	486	402	329	242	276	

(2) 両県の人口分布データを表すベクトルを `akita`, `aichi` とすると，累積相対度数は `cumsum(akita)/sum(akita)` 等によって求められる．これらを (階段状のグラフとして) 図示すると，愛知のグラフがいたるところで秋田のグラフの上側にあり，問題 2.6 の男子学生と女子学生の家庭の年収額と同様な関係になっている．すなわち，「65 歳以上の人口の割合」だけでなく，何歳以上の人口の割合も秋田のほうが大きく，秋田県は「高齢化」とともに「少子化」も進んでいることがわかる．

問題 2.8　次の表は，世界の 20 ヵ国の 1 人あたり GNI(単位は米ドル) と自動車保有台数 (千人あたりの台数) を示す (出典：総務省統計局ホームページ「世界の統計 2011」，第 3 章および第 8 章．2008 年のデータであるが，一部例外がある．なお，GNI は GDP に“海外からの所得の純受取”を加えたものである)．

国名	1 人あたり GNI	自動車保有台数	国名	1 人あたり GNI	自動車保有台数
日本	39,574	591	ブラジル	8,332	198
インド	1,080	15	イギリス	44,187	526
韓国	19,487	346	イタリア	37,936	673
タイ	3,884	134	ギリシャ	29,983	560
中国	3,316	37	スペイン	34,830	606
トルコ	9,871	138	ドイツ	44,887	554
マレーシア	7,921	334	フランス	45,085	598
アメリカ合衆国	46,236	809	ポーランド	13,614	495
カナダ	44,549	605	ロシア	11,451	245
メキシコ	9,855	264	オーストラリア	45,402	687

(1) このデータについて散布図を作成し，相関係数を求めよ．

(2) 上記のホームページ (あるいは他の情報源) から 2009 年以降のデータを取得して，(1) で作成した散布図の上にプロットし，各国の時間変化を観察せよ．

解答　**(1)** 相関係数は 0.92 で，正の強い相関があることがわかる．

(2) 略

問題 2.9　ある大学のある学科では．入学試験の成績と入学後の成績の相関を調べることになった．入学者は 50 名で，入学試験の成績も入学後の成績も，それぞれ 50 名中の順位で示されている．下記のデータは，入学試験の順位の 1 番から 50 番までの学生の入学後の順位を示したものである．相

関を調べよ.

3, 6, 2, 7, 10, 19, 18, 1, 9, 8,
25, 39, 17, 4, 11, 5, 36, 45, 13, 16,
27, 44, 33, 48, 50, 41, 15, 28, 20, 38,
12, 29, 47, 14, 23, 49, 22, 46, 26, 21,
40, 32, 37, 43, 24, 30, 35, 31, 42, 34

解答　入試の順位と入学後の順位を次のようにして入力する.

```
nyushi<-1:50
nyugakugo<-c(3,6,2,…,42,34)
```

スピアマンの順位相関係数を `cor(nyushi,nyugakugo)` によって計算すると 0.58 となり，相関はそれほど強くないことがわかる.

3　実験的推測統計

問題 3.1　(1) sample 関数を使い，コイン投げを 10 回繰り返した結果を「おもて」「うら」の文字列で表示する R の命令を書け．

(2) コイン投げを 100 回繰り返しておもての出る相対度数を計算する R の命令を書け．

(3) コイン投げを 100 回繰り返しておもての出る相対頻度を計算するという実験を 100 回繰り返して，そのヒストグラムを描く R の命令を書け．

(4) (3) のヒストグラムの縦軸を相対度数とし，平均 0.5, 標準偏差 0.05 の正規分布の密度関数を重ねて描く R の命令を書け．

ヒント　使う関数は，sample(), mean(), sapply(), hist(), curve(), dnorm(), 使うオプションは replace=T, freq=F, add=T.

解答

プログラム例

```
# (1)
sample(c("おもて","うら"), 10, replace=T)
# (2)
coin < - sample(c("おもて","うら"), 100, replace=T)
head < - which(coin == "おもて")
length(head) / 100
# (3)
jikken < - sapply(rep(100,100), function(n) {
        coin < - sample(c("おもて","うら"), n, replace=T)
        head < - which(coin == "おもて")
        length(head) / 100
    })
hist(jikken)
# (4) は (3) のヒストグラム表示を変えるだけ
hist(jikken, freq=F)
curve(dnorm(x, 0.5, 0.05), add=T)
```

問題 3.2　10 万人都市で，内閣支持率が 20% という想定の下で，1000 人を対象とした架空調査を 10000 回実施し，推定支持率のヒストグラムを描いて，調査結果のばらつきについて気がついたことを説明せよ．

ヒント　sample で標本抽出して mean で平均を計算することを sapply で繰り返す．

解答

プログラム例

```
population < - c(rep("支持",20000), rep("不支持",80000))
    # 支持率 20% の 10 万人母集団
z < - sapply(rep(1000,10000), function(x) {
        ss = sample(population, x)
        mean(ss == "支持")
    })
hist(z)
```

ほぼ ±4% の範囲に収まっている．誤差は最大でも 2 割，といってもよい．

問題 3.3　テレビの視聴率調査を実験する．あるテレビ番組を 100 万世帯のうち 10 万世帯が見ているという状況を作り，そこから大きさ 600 の標本をランダムに選んで，その相対視聴世帯数を計算する R の命令を書け (600 は実際のテレビ視聴率調査における対象世帯数である)．それを使って，架空の視聴率調査を 1000 回繰り返し，1000 通りの調査結果をヒストグラムにまとめよ．その図から，結果がどれくらいばらつくかを調べよ．同じような実験を，5 万世帯が見ているテレビ番組の視聴率調査という設定で実施し，その結果を考察せよ．

ヒント　視聴率 10% の世帯は `c(rep(1,100000),rep(0,900000))` によって表現できる．`sample` で標本抽出して `mean` で平均を計算することを `sapply` で繰り返す．結果はヒストグラム表示するとわかりやすい．

解答

プログラム例

```
## 視聴率調査
population < - c(rep(1,100000),rep(0,900000))
    # 視聴率 10% の 100 万人母集団の作成
mean(sample(population, 600))
## 視聴率調査の精度計算
ww < - sapply(rep(600,1000), function(n) mean(sample(population, n)))
hist(ww, freq=F)
curve(dnorm(x,mean(ww),sd(ww)), add=T, col=2)
## 視聴率調査 (視聴率 5% の場合は，population を次の式で置き換えるだけ
population < - c(rep(1,50000),rep(0,950000))    # 視聴率 10% の 100 万人母集団の作成
```

参考までに，正規分布の密度関数を重ねて描いてある．`col=2` は線の色を赤くするオプションである．視聴率が 10% の場合の誤差は 3% から 4% 程度なのに対して，視聴率が 5% の場合は 2% から 3% と小さくなる．

問題 3.4　平均 10, 標準偏差 2 の正規分布に従う架空データを 100 個生成し，それら 100 個の平均値と標準偏差を計算する R の命令を書け．100 個のデータのヒストグラムを描き，計算された平均と標準偏差をもつ正規分布のグラフをヒストグラムに重ねて描け．同じ実験を架空データを変えて実施せよ．複数回の実験結果をまとめよ．また，生成する個数を 1000 として同じ実験を実施せよ．

ヒント　平均 m, 標準偏差 s の正規分布に従うデータを `n` 個生成する関数は `rnorm(n,m,s)`．正規分布の密度関数を `[a,b]` の範囲で描くのは `curve(dnorm(x,m,s))`, ただし，平均と標準偏差はデータから計算したものを使う．重ねて描くためには「`add=T`」オプションが必要．また，ヒストグラムの縦軸を相対度数とする必要がある (「`freq=F`」オプションを使う)．

解答

プログラム例

```
## 正規母集団からの標本抽出実験
n < - 100                  # n = 1000 の実験は，ここを n < - 1000 とする
w < - rnorm(n, 10, 2)
c(mean(w), sd(w))
hist(w, freq=F)
```

```
curve(dnorm(x,mean(w),sd(w)), add=T, col=2)
```

4　確率論の基礎知識

練習問題 4.1　パラメータ n, p の 2 項分布の平均と分散を計算せよ．

解答　平均は

$$E(X) = \sum_{i=0}^{n} i \times \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!} p^i (1-p)^{n-i}$$
$$= np \sum_{i=1}^{n} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} = np$$

分散は，

$$E(X(X-1)) = \sum_{i=0}^{n} i(i-1) \times \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=2}^{n} \frac{n!}{(i-2)!(n-i)!} p^i (1-p)^{n-i}$$
$$= n(n-1))p^2 \sum_{i=2}^{n} \binom{n-2}{i-2} p^{i-2} (1-p)^{n-i} = n(n-1)p^2$$

を利用して，次のように計算される．

$$V(X) = E(X(X-1)) + E(X) - (E(X))^2 = np(1-p)$$

練習問題 4.2　パラメータ p の幾何分布の平均と分散を計算せよ．

解答　平均は

$$E(X) = \sum_{k=1}^{\infty} k \times p(1-p)^{k-1} = p \sum_{k=1}^{\infty} \sum_{i=1}^{k} (1-p)^{k-1} = p \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} (1-p)^{k-1} = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{p}$$

分散を計算するには，まず

$$E(X(X-1)) = p \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1} = 2p \sum_{k=2}^{\infty} \sum_{i=1}^{k-1} i \times (1-p)^{k-1} = 2p \sum_{i=1}^{\infty} i \sum_{k=i+1}^{\infty} (1-p)^{k-1}$$
$$= 2 \sum_{i=1}^{\infty} i(1-p)^i = 2 \frac{1-p}{p^2} = \frac{2}{p^2} - \frac{2}{p}$$

したがって，

$$V(X) = E(X(X-1)) + E(X) - (E(X))^2 = \frac{1-p}{p^2}$$

練習問題 4.3　パラメータ λ のポアソン分布の平均と分散を計算せよ．

解答　平均は

$$E(X) = \sum_{k=0}^{\infty} k \times \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \mathrm{e}^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \mathrm{e}^{-\lambda} = \lambda$$

分散は

$$E(X(X-1)) = \sum_{k=0}^{\infty} k(k-1) \times \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} = \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} \mathrm{e}^{-\lambda}$$
$$= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \mathrm{e}^{-\lambda} = \lambda^2$$

を利用して，次のように計算される．

$$V(X)E(X(X-1))+E(X)-(E(X))^2=\lambda$$

練習問題 4.4　パラメータ a,b のガンマ分布のモーメント母関数を計算し，その平均と分散を計算せよ．

解答　モーメント母関数は定義通りの計算．

$$\begin{aligned}M(\theta)=E(\mathrm{e}^{\theta X})&=\int_0^\infty \mathrm{e}^{\theta x}\times\frac{b^a x^{a-1}}{\Gamma(a)}\mathrm{e}^{-bx}\mathrm{d}x=\int_0^\infty\frac{b^a x^{a-1}}{\Gamma(a)}\mathrm{e}^{-(b-\theta))x}\mathrm{d}x\\&=\left(\frac{b}{b-\theta}\right)^a\int_0^\infty\frac{(b-\theta)^a x^{a-1}}{\Gamma(a)}\mathrm{e}^{-(b-\theta))x}\mathrm{d}x\\&=\left(\frac{b}{b-\theta}\right)^a\end{aligned}$$

最後の等式は，被積分関数がパラメータ $a,b-\theta$ のガンマ分布の密度関数であることを利用している．4 番目の等式のように，定積分がわかっているような被積分関数に持ち込むような変形をすれば，部分積分のような道具を使わなくてすむ．計算ミスをなくす工夫である．

微分すると，

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}M(\theta)&=a\frac{b^a}{(b-\theta)^{a+1}}\\\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}M(\theta)&=a(a+1)\frac{b^a}{(b-\theta)^{a+2}}\end{aligned}$$

なので，平均値は a/b，2 次モーメントは $a(a+1)/b^2$，したがって分散は a/b^2．

練習問題 4.5　X,Y が独立ならば，任意の 2 つの関数 $g(x),h(x)$ に対して，

$$E(g(X)h(Y))=E(g(X))E(h(Y))$$

が成り立つことを証明せよ．

解答　連続の場合で証明する．離散の場合も同様．X,Y の結合密度関数はそれぞれの（周辺）密度関数 $f_X(x),f_Y(x)$ の積で表される．したがって，

$$\begin{aligned}E(g(X)h(Y))&=\int_{-\infty}^\infty\int_{-\infty}^\infty g(x)h(y)\times f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y\\&=\int_{-\infty}^\infty g(x)f_X(x)\mathrm{d}x\int_{-\infty}^\infty h(y)f_Y(y)\mathrm{d}y=E(g(X))E(h(Y))\end{aligned}$$

練習問題 4.6　X_i がパラメータ a_i,b のガンマ分布に従い $(i=1,2,...,n)$，互いに独立ならば，$X_1+X_2+\cdots+X_n$ はパラメータ $\sum_{i=1}^n a_i,b$ のガンマ分布に従うことを示せ．このことから，2 つ目のパラメータ (尺度パラメータ) b の値が共通のガンマ分布は再生性をもつことがわかる．

解答　独立な確率変数の和のモーメント母関数は，個々の確率変数のモーメント母関数の積に等しいということを使う．パラメータ a_i,b のガンマ分布のモーメント母関数は

$$M_i(\theta)=\left(\frac{b}{b-\theta}\right)^{a_i}$$

なので，独立な確率変数の和である $X_1 + X_2 + \cdots + X_n$ のモーメント母関数 $M(\theta)$ は

$$\begin{aligned} M(\theta) &= \left(\frac{b}{b-\theta}\right)^{a_1}\left(\frac{b}{b-\theta}\right)^{a_2}\cdots\left(\frac{b}{b-\theta}\right)^{a_n} \\ &= \left(\frac{b}{b-\theta}\right)^{a_1+a_2+\cdots+a_n} \end{aligned}$$

これは，パラメータ $a_1 + a_2 + \cdots + a_n, b$ のガンマ分布のモーメント母関数に他ならない．

練習問題 4.7　式 (4.75) が成り立つことを確かめよ．

解答　右辺を変形する．$E(X \mid Y)$ を Y の関数として $g(Y)$ とおくと

$$V(E(X \mid Y)) = E(g(Y)^2) - (E(g(Y)))^2 = E(g(Y)^2) - (E(X))^2$$

$$V(X \mid Y) = E(X^2 \mid Y) - g(Y)^2 \Rightarrow E(V(X \mid Y)) = E(X^2) - E(g(Y)^2)$$

したがって，

$$V(E(X \mid Y)) + E(V(X \mid Y)) = E(X^2) - (E(X))^2 = V(X)$$

問題 4.1　必ずしも互いに排反でない事象 A, B について

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

が成り立つことを確かめよ．これを確率の和の公式 (あるいは加法定理) という．

ヒント　集合として，$A \cup B$ が $A \cap B^c$ と $A^c \cap B$ と $A \cap B$ の和集合に分解できることを使う．
解答　A, B は互いに排反な 2 つの事象の和に分解できる：

$$A = (A \cap B) \cup (A \cap B^c), B = (A \cap B) \cup (A^c \cap B)$$

これを使うと，$A \cup B$ は互いに排反な 3 つの事象の和に分解できる：

$$A \cup B = (A \cap B) \cup (A^c \cap B) \cup (A \cap B^c)$$

これより，

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A^c \cap B) + P(A \cap B^c) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

問題 4.2　命題 4.3 に関連して，次の式が成り立つことを確かめよ．

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

これはベイズの公式と呼ばれる．

また，この定理を使って，次の確率を計算せよ．某社では，ある装置を製造するために多数必要な同一部品を 3 つの会社 $\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_3$ から購入している．購入数量の比率は $50\%, 30\%, 20\%$ であり，1 台の装置に使用する割合もこの比率である．また各社の納入部品 1 個が使用開始後 1 年以内に故障する確率は，それぞれ 0.015, 0.010, 0.020 であるものとする．完成した装置を使用開始後 1 年で部品 1 個が故障したとして，それが $\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_3$ 社製のものである確率を求めよ．

解答　条件付き確率の定義から

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)}$$

が成り立つ．分母分子はそれぞれ

$$P(A) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(A \mid B_j)P(B_j)$$
$$P(A \cap B_i) = P(A \mid B_i)P(B_i)$$

と書き換えられるので，これらを代入することによって，ベイズの定理が証明される．

A を部品が 1 年以内に故障するという事象，B_1, B_2, B_3 をそれぞれ会社 B_1, B_2, B_3 で生産されるという事象とする．このとき，条件式から，

$$P(B_1) = 0.5, P(A \mid B_1) = 0.015$$
$$P(B_2) = 0.3, P(A \mid B_2) = 0.01$$
$$P(B_3) = 0.2, P(A \mid B_3) = 0.02$$

が与えられているので，これをベイズの定理の式に代入すればよい．

$$P(A \mid B_1)P(B_1) : P(A \mid B_1)P(B_1) : P(A \mid B_1)P(B_1) = 7.5 : 3 : 4$$

なので，

$$P(B_1 \mid A) = \frac{15}{29}, P(B_2 \mid A) = \frac{6}{29}, P(B_3 \mid A) = \frac{8}{29}$$

問題 4.3　X は非負の値だけをとる確率変数で，$F(x)$ はその分布関数である．X の期待値 $E(X)$ について次の式が成り立つことを確かめよ．

$$E(X) = \int_0^\infty x \mathrm{d}F(x) = \int_0^\infty (1 - F(x))\mathrm{d}x$$

ヒント　$x = \int_0^x \mathrm{d}u$.

解答

$$E(X) = \int_0^\infty \left(\int_0^x \mathrm{d}t \right) \mathrm{d}F(x) = \left(\int_0^\infty \int_t^\infty \mathrm{d}F(x) \right) \mathrm{d}t = \int_0^\infty (1 - F(t))\mathrm{d}t$$

あるいは，幾何学的に $1 - F(t)$ を積分するということは，$y = 1$ と y 軸と $y = F(x)$ に囲まれた部分の面積を計算していることになるが，それを

$$\int_0^\infty (1 - F(t))\mathrm{d}t = \int_0^\infty \left(\int_{F((t)}^1 \mathrm{d}u \right) \mathrm{d}t = \int_0^1 F^{-1}(u)\mathrm{d}u$$

と書き換えてみれば，これは分布関数の逆関数 $F^{-1}(x)$ を 0 から 1 までの積分したものと考えてよい．そうすると，$0 = u_0 < u_1 < \cdots < u_n = 1$ として，

$$\int_0^1 F^{-1}(u)\mathrm{d}u \approx \sum_i F^{-1}(u_i)(u_{i+1} - u_i)$$

と表される．$u_i = F(x_i)$ とおくと，

$$\sum_i F^{-1}(u_i)(u_{i+1} - u_i) = \sum_i x_i(F(x_{i+1}) - F(x_i))$$
$$= \sum_i x_i \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i}(x_{i+1} - x_i)$$

ここで，n を十分に大きくとると，$\frac{F(x_{i+1})-F(x_i)}{x_{i+1}-x_i} \approx f(x_i)$ となり，和を積分で近似できる．

$$\sum_i x_i \frac{F(x_{i+1})-F(x_i)}{x_{i+1}-x_i}(x_{i+1}-x_i) \approx \sum_i x f(x_i) \times (x_{i+1}-x_i) \to \int_0^\infty x f(x)\mathrm{d}x$$

問題 4.4　X が指数分布をする確率変数ならば，任意の非負定数 x, y について

$$P(X > x+y \mid X > x) = P(X > y)$$

が成り立つことを示せ．

解答　パラメータ λ の指数分布の分布関数は $1-\mathrm{e}^{-\lambda x}$ で与えられるので，

$$P(X > x+y \mid X > x) = \frac{P(X > x+y)}{P(X > x)} = \frac{\mathrm{e}^{-\lambda(x+y)}}{\mathrm{e}^{-\lambda x}} = \mathrm{e}^{-\lambda y} = P(X > y)$$

問題 4.5　あるコールセンターに次々にかかってくる電話の統計をとったところ，1 分間に平均 5 本で，着信の間隔は指数分布をしていることがわかった．このとき，1 時間の間にかかってくる電話の本数は平均が $300(=5\times 60)$ のポアソン分布に従うことを示せ．

ヒント　$X_1, X_2, \ldots$ を互いに独立に平均が 1/5[分] の指数分布に従う確率変数とし，$S_n = X_1 + X_2 + \cdots + X_n$ とおく．1 時間の間にかかってくる電話の本数が n 本であるという事象は，次の不等式が成り立つことに相当する：$S_n \le 60 < S_{n+1}$．$S_n = s(\le 60)$ という条件の下では，$X_{n+1} > 60 - s$ が上記の不等式が成り立つことと同等である．s を $0 \le s \le 60$ の範囲で動かして，(連続分布に関する) 全確率の公式を使うとよい．

解答　ヒントに従って，$X_1, X_2, \ldots$ を互いに独立に平均が 1/5[分] の指数分布に従う確率変数として，$S_n = X_1 + X_2 + \cdots + X_n$ とおくと，S_n はパラメータ $n, b(=5)$ のガンマ分布に従う．1 時間の間にかかってくる電話の本数を N としたとき，$N = n$ という事象は $S_n(=s) \le 60$ で，かつ $S_{n+1} > 60$，あるいは $X_{n+1} > 60 - s$ という事象と同じである．そこで，$0 < S_n \le 60$ で条件を付けた全確率の公式を使えばよい．S_n の密度関数を $f_{S_n}(s)$ とすると，

$$\begin{aligned} P(N=n) &= \int_0^{60} P(X_{n+1} > 60-s) f_{S_n}(s)\mathrm{d}s \\ &= \int_0^{60} \mathrm{e}^{-b(60-s)} \frac{b^n s^{n-1}}{(n-1)!} \mathrm{e}^{-bs}\mathrm{d}s = \frac{(60b)^n}{n!}\mathrm{e}^{-60b} \end{aligned}$$

これは，N がパラメータ $60b = 300$ のポアソン分布に従うことを意味する．

問題 4.6　例 4.12 で述べたとおり，パラメータが n, p の 2 項分布は，p が小さいとき，n が大きくなるとパラメータが np のポアソン分布で近似できることが知られている．いくつかの p と n について，R を使って確率関数を計算して，近似の程度を調べよ．確率関数を計算する R の関数は，2 項分布：`dbinom(k,n,p)`，ポアソン分布：`dpois(k,np)` である．

解答　確率関数は棒グラフが適切であるが，違いが微妙なので，折れ線で比較する．$(n,p) = (10,0.1), (20,0.05), (100,0.01), (100,0.05)$ など，比較してみると，n が大きくなるにつれて近似の程度がよくなっていることが視認できる．

プログラム例

```
n = 100; p = 0.05; nn = min(n,20)
plot(dbinom(0:nn,n,p), type="b")
lines(dpois(0:nn, n*p),col=2)
```

問題 4.7　区間 [0,1] 上の一様分布に従う乱数を R で生成するには `runif` という関数を使う．それを m 個生成して平均値を計算するということを n 回繰り返して，ヒストグラムを描く，という実験を行う R のプログラムを書け．そのプログラムを使い，$m = 2, 4, 12$ としたときに，ヒストグラムの形状がどうなるか，調べよ．ただし，$n = 10000$ とせよ．また，平均が 0.5, 分散が $1/12m$ の正規分布の密度関数を重ねて描き，比較せよ．

ヒント　`mean(runif(m))` を `sapply` で繰り返す．ヒストグラムは `hist` 関数を使って描くが，正規分布の密度関数を重ねるので，「`freq=F`」オプションが必要．正規分布の密度関数は `curve(dnorm(x,1/2,sqrt(1/12/m)),add=T)` で描くことができる．

解答　ヒント通り．$m = 2$ の時，ヒストグラムは二等辺三角形のようになる．$m = 4$ で正規分布の密度関数らしくなり，$m = 12$ でほとんど正規分布の密度関数と重なる．

プログラム例

```
## 中心極限定理
n = 10000
m = 12
z = sapply(rep(m,n), function(m) mean(runif(m)))
hist(z, freq=F)
curve(dnorm(x, 0.5, sqrt(1/12/m)), add=T)
```

5　推測統計の確率モデル概説，標本分布

問題 5.1　$X_1, X_2, ..., X_n$ はパラメータ p の 2 項母集団からの大きさ n の独立標本としたとき，標本平均の従う分布を求めよ．

解答　$X_1 + X_2 + \cdots + X_n$ はパラメータ n, p の 2 項分布に従うので，標本平均を $\bar{X}$ とすると，

$$P\left(\bar{X} = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^k, \quad k = 0, 1, ..., n$$

問題 5.2　$X_1, X_2, ..., X_n$ をパラメータ a, b のガンマ分布 (4.6.8 項参照) を母集団分布とする母集団からの大きさ n の独立標本とするとき，その標本平均 $\bar{X}$ はどのような分布に従うか，モーメント母関数を使って計算せよ．

解答　ガンマ分布のモーメント母関数は

$$M(\theta) = \left(\frac{b}{b-\theta}\right)^a$$

したがって，標本平均のモーメント母関数は

$$M_{\bar{X}}(\theta) = \left(\frac{b}{b-\theta/n}\right)^{an} = \left(\frac{bn}{bn-\theta}\right)^{an}$$

これは，パラメータ an, bn のガンマ分布のモーメント母関数になっているので，その平均，分散はガンマ分布の平均，分散の公式を適用して，

$$E\left(\bar{X}\right) = \frac{an}{bn} = \frac{a}{b}$$
$$V\left(\bar{X}\right) = \frac{an}{(bn)^2} = \frac{1}{n}\frac{a}{b^2}$$

もちろん，モーメント母関数を微分することによってモーメントを計算し，平均，分散を導いても同じ結果が得られる．

問題 5.3　$X_1, X_2, ..., X_n$ は互いに独立に平均 μ, 分散 σ^2 の正規分布に従う確率変数で $\bar{X}$ はそれらの標本平均とするとき，任意の i について，$\bar{X}$ と $X_i - \bar{X}$ は互いに独立であることを示せ．

ヒント　2 つの正規分布に従う確率変数が独立であることと無相関であることは同じ．
解答　$\bar{X}$ も $X_i - \bar{X}$ も正規分布に従う確率変数の線形和なので，正規分布に従う．その共分散を計算すると，

$$C(\bar{X}, X_i - \bar{X}) = C(\bar{X}, X_i) - V(\bar{X}) = \frac{1}{n}\sum_{j=1}^{n} C(X_j, X_i) - \frac{\sigma^2}{n} = \frac{1}{n}V(X_i) - \frac{\sigma^2}{n} = 0$$

正規分布に従う 2 つの確率変数の共分散がゼロならば，それらは互いに独立．

問題 5.4　命題 5.2 を利用して，正規分布の不偏分散 $\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ の分散が $2\sigma^4/(n-1)$ となることを示せ．

解答　命題 5.2 より，

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

は自由度 $n-1$ のカイ 2 乗分布に従うことがわかっている．したがって W の分散は $2(n-1)$．不偏分散を V とすると，$V=\sigma^2 W/(n-1)$ という関係にあるので，V の分散は $2\sigma^4/(n-1)$.

問題 5.5　自由度 n の t 分布に従う確率変数 X に対して，X^2 は自由度 $1,n$ の F 分布に従うことを示せ．

ヒント　t 分布は標準正規分布に従う確率変数とカイ 2 乗分布に従う確率変数の比で表されることを使う．

解答　累積分布関数を計算すると，

$$\begin{aligned}P(X^2\le x)&=P\left(|X|\le\sqrt{x}\right)=2\int_0^{\sqrt{x}}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}\left(\frac{t^2}{n}+1\right)^{-(n+1)/2}\mathrm{d}t\\&=\int_0^x\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}\left(\frac{u}{n}+1\right)^{-(n+1)/2}\frac{\mathrm{d}u}{\sqrt{u}},\quad(t^2=u\text{ と置いた})\end{aligned}$$

これは自由度 $1,n$ の F 分布の累積分布関数に他ならない．

意味的には，次のように考えられる．標準正規分布に従う確率変数を Z，自由度 n のカイ 2 乗分布に従う確率変数を W とおくと，$Z/\sqrt{W/n}$ は自由度 n の t 分布に従う．Z^2 は自由度 1 のカイ 2 乗分布に従う．したがって，F 分布の定義より，$\left(Z/\sqrt{W/n}\right)^2=Z^2/(W/n)$ は自由度 $1,n$ の F 分布に従う．

問題 5.6　標準正規分布に従う確率変数の 2 乗がガンマ分布に従うことを示せ．R で標準正規分布に従う乱数を使ってこのことを確かめよ．

ヒント　変数変換を使う．R で「`rnorm(n)^2`」とすれば正規分布に従う乱数の 2 乗を生成することができる．

解答　X を標準正規分布に従う確率変数とし，$g(x)=x^2$ とした場合，$Y=X^2$ の分布関数は

$$\begin{aligned}P(Y\le x)&=P(X^2\le x)=P\left(-\sqrt{x}\le X\le\sqrt{x}\right)=2P\left(0\le X\le\sqrt{x}\right)\\&=\sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}}\mathrm{e}^{-u^2/2}\mathrm{d}u\end{aligned}$$

変数変換

$$u^2=v\Leftrightarrow u=\sqrt{v},\quad\mathrm{d}u=\frac{\mathrm{d}v}{2\sqrt{v}}$$

を適用して

$$\begin{aligned}F_Y(x)&=\sqrt{\frac{2}{\pi}}\int_0^x\frac{1}{2\sqrt{v}}\mathrm{e}^{-v/2}\mathrm{d}v\\f_Y(x)&=\frac{\mathrm{d}}{\mathrm{d}x}F_Y(x)=\sqrt{\frac{2}{\pi}}\frac{1}{2\sqrt{x}}\mathrm{e}^{-x/2}\end{aligned}$$

ここで，$\sqrt{\pi}=\Gamma(1/2)$ ということを使うと

$$f_Y(x)=\frac{1}{\Gamma(a)}b^a x^{a-1}\mathrm{e}^{-bx},\quad(a=b=1/2)$$

となり，これはパラメータ a,b のガンマ分布の密度関数を表している．

問題 5.7　パラメータ a,b のベータ分布に従う確率変数を X とするとき，$X/(2a)/((1-X)/(2b))$ は，自由度 $2a,2b$ の F 分布に従うことを示せ．

ヒント　変数変換を使う．

解答　パラメータ a, b のベータ分布の累積分布関数を $B_{a,b}(x)$ と記す．

$$Y = \frac{X/(2a)}{(1-X)/(2b)}$$

とおくと，

$$\begin{aligned} P\left(Y = \frac{X/(2a)}{(1-X)/(2b)} \leq y\right) &= P\left(X \leq \frac{y/(2b)}{1/(2a) + y/(2b)}\right) \\ &= B_{a,b}\left(\frac{ay}{b+ay}\right) \end{aligned}$$

となるので，Y の密度関数は

$$\begin{aligned} &\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\frac{ay}{b+ay}\right)^{a-1} \left(\frac{b}{b+ay}\right)^{b-1} \frac{ab}{(b+ay)^2} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\frac{b}{a}\right)^{b} y^{a-1} \left(y + \frac{b}{a}\right)^{-(a+b)} \end{aligned}$$

これは自由度 $2a, 2b$ の F 分布の密度関数に他ならない．

問題 5.8　自由度 m, n の F 分布の上側 $100\alpha\%$ 点を $F_\alpha(m,n)$ としたとき，つねに $F_\alpha(m,n)F_{1-\alpha}(n,m) = 1$ が成り立つことを示せ．

ヒント　F 分布は 2 つのカイ 2 乗分布に従う確率変数と関係がある．

解答　自由度 m のカイ 2 乗分布に従う確率変数を X, 自由度 n のカイ 2 乗分布に従う確率変数を Y とすると，$(X/m)/(Y/n)$ は自由度 m, n の F 分布に従い，$(Y/n)/(X/m)$ は自由度 n, m の F 分布に従う．したがって，

$$\alpha = P\left(\frac{X/m}{Y/n} \geq F_\alpha(m,n)\right) = P\left(\frac{Y/n}{X/m} \leq \frac{1}{F_\alpha(m,n)}\right) \tag{5.1}$$

$$= 1 - P\left(\frac{Y/n}{X/m} \geq \frac{1}{F_\alpha(m,n)}\right) \tag{5.2}$$

最右辺の等号は自由度 n, m の F 分布の上側 $100(1-\alpha)\%$ 点 $F_{1-\alpha}(n,m)$ が $1/F_\alpha(m,n)$ に等しいことを意味する．

6　統計的推定問題

練習問題 6.1　パラメータ $p = 0.4$ の 2 項母集団からの大きさ $n = 100$ の標本から $\hat{p} = 0.35$ という推定値を得た．この結果から真の値の 95% 信頼区間を正規近似で求めよ．また，2 項分布の累積分布関数を使って，95% 信頼区間を求めよ．2 項分布の累積分布関数は R の関数 `pbinom(k,n,p)` によって計算することができる．

訂正　「パラメータ $p = 0.4$ の」は削除.

解答　分散 $p(1-p)/n$ の推定値として $0.35 * 0.65/100 = 0.002\,275 = 0.0477^2$ を得る．標準正規分布の両側 5% 点は，`qnorm(0.975)` で計算することができて，1.96 である．したがって，2 項母集団パラメータの正規近似による 95% 信頼区間は

$$0.35 \pm 1.96 \times 0.0477 \Leftrightarrow [0.257, 0.443]$$

正規近似を使わない場合は，「`pbinom(kb,100,0.35)-pbinom(ka,100,0.35)> =0.95`」を満たす `ka,kb` の中で，`kb-ka` が最小となるものを探せばよい．通常は下側 2.5% 点 (なければ，それ以下の確率が 2.5% 以下になる最大の値) を `ka`，上側 2.5% 点 (なければ，それ以上の確率が 2.5% 以下になる最小の値) を `kb` とする．`pbinom(44,100,0.35)=0.9754`，`pbinom(25,100,0.35)=0.0021` となるので，2 項母集団パラメータの 95% 信頼区間は

$$[0.25, 0.44]$$

ただし，`pbinom(44,100,0.35)-pbinom(24,100,0.35)=0.963` となり，95% 信頼区間といっても信頼性は 96% と，やや高い．

問題 6.1　パラメータ μ, σ^2 の正規母集団で，

$$T_1(\boldsymbol{X}) = \sum_{i=1}^{n} X_i, \quad T_2(\boldsymbol{X}) = \sum_{i=1}^{n} X_i^2$$

とおくと，$T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$ は μ, σ^2 の十分統計量になることを示せ.

解答　パラメータ μ, σ^2 の正規母集団からの大きさ n の標本 $x_1, x_2, ..., x_n$ に基づく対数尤度関数は

$$\begin{aligned} l(\mu, \sigma^2; \boldsymbol{x}) &= -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \\ &= -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left(T_2(\boldsymbol{x}) - 2\mu T_1(\boldsymbol{x}) + n\mu^2 \right) \end{aligned}$$

となるので，尤度関数の中で n 個の標本は $T_1(\boldsymbol{x}), T_2(\boldsymbol{x})$ という形でしか含まれていない．これは，μ, σ^2 の推定において，$T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$ が十分統計量であることを意味する (命題 6.2).

問題 6.2　パラメータ μ, σ の正規母集団から大きさ 10 の標本を抽出したところ，平均は 510.1, 不偏分散は 24.4 であった．95% 信頼区間の半分の長さを 2 以内に収めたいとしたとき，あといくつの標本をとればよいか．もし，σ^2 が 20 とわかっている場合は，その結果はどう変わるか.

解答　95% 信頼区間の半分の長さは，$1.96\sigma/\sqrt{n}$ で与えられる．σ^2 を不偏分散 $24.4 = 4.94^2$ で代用すると

$$1.96 \frac{4.94}{\sqrt{n}} < 2$$

を満たす最小の n はいくつかという問題になる．$n > 23.4$ なので，答えは，あと 14 個必要．$\sigma^2 = 20 = 4.47^2$ とすると，10 個でよい．

問題 6.3　標準正規分布に従う確率変数 X と定数 $p(0 < p < 1)$ とに対して，$P(X \in [r, s]) = p$ を満たす区間 $[r, s]$ の中で $s - r$ が最小になるのは $r + s = 0$, すなわち，原点を挟んで対称な区間であることを示せ．

ヒント　ラグランジュ乗数を使う．あるいは，$s = r + g(r)$ とおいて，条件式を r で微分して考える．

解答

ラグランジュ乗数 λ を使って，目的関数を

$$s - r - \lambda(\Phi(s) - \Phi(r) - p)$$

と表し，r, s で微分して 0 とおくと

$$-1 + \lambda\phi(r) = 0, 1 - \lambda\phi(s) = 0 \Rightarrow \phi(r) = \phi(s)$$

が導かれる．標準正規分布の密度関数の対称性から，$r + s = 0$ となる r, s が解になることがわかる．

問題 6.4　$\theta > 0$ とし，$X_1, X_2, ..., X_n$ を区間 $U(0, \theta)$ 上の一様分布を母集団分布とする母集団からの独立標本とする．

$$T(\boldsymbol{X}) = c \max\{X_1, X_2, ..., X_n\}$$

が θ の不偏推定量となる c の値を求めよ．そのときの $T(\boldsymbol{X})$ の分散を求めよ．

解答

$$\begin{aligned} P(T(\boldsymbol{X}) < x) &= P\left(X_1 < \frac{x}{c}, X_2 < \frac{x}{c}, ..., X_n < \frac{x}{c}\right) \\ &= \left(\frac{x}{c\theta}\right)^n \end{aligned}$$

したがって，

$$E(T(\boldsymbol{X})) = \int_0^{c\theta} x \times \frac{nx^{n-1}}{(c\theta)^n} \mathrm{d}x = \frac{n}{n+1} c\theta = \theta \Leftrightarrow c = \frac{n+1}{n}$$

分散は

$$E(T(\boldsymbol{X})^2) = \int_0^{c\theta} x^2 \times \frac{nx^{n-1}}{(c\theta)^n} \mathrm{d}x = \frac{n}{n+2} (c\theta)^2 = \frac{(n+1)^2}{n(n+2)} \theta^2$$

より，

$$V(T(\boldsymbol{X})) = \frac{(n+1)^2}{n(n+2)} \theta^2 - \theta^2 = \frac{1}{n(n+2)} \theta^2$$

問題 6.5　次の確率分布に対して，データ 1 つあたりのフィッシャーの情報量を求めよ．(1) パラメータ p のベルヌイ分布，(2) 平均 a の指数分布．

解答　(1)

$$\begin{aligned} \log f(x; p) &= x \log p + (1 - x) \log(1 - p) \\ \frac{\partial}{\partial p} \log f(x; p) &= \frac{x}{p} - \frac{1 - x}{1 - p} \\ \frac{\partial^2}{\partial p^2} \log f(x; p) &= -\frac{x}{p^2} - \frac{1 - x}{(1 - p)^2} \end{aligned}$$

したがって，

$$I_1(p) = E\left(-\frac{\partial^2}{\partial p^2}\log f(X;p)\right) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

(2)

$$\log f(x;a) = -\log a - \frac{x}{a}$$
$$\frac{\partial}{\partial a}\log f(x;a) = -\frac{1}{a} + \frac{x}{a^2}$$
$$\frac{\partial^2}{\partial a^2}\log f(x;a) = \frac{1}{a^2} - \frac{2x}{a^3}$$

したがって，

$$I_1(p) = E\left(-\frac{\partial^2}{\partial a^2}\log f(X;a)\right) = \frac{1}{a^2}$$

問題 6.6　母集団分布がパラメータ θ を含む次の式で与えられるとき，(1) 標本の大きさ n が十分に大きいときの θ の最尤推定量 $\hat{\theta}_n$ の分散を求めよ．(2) n が十分に大きいとき，標本の中央値 X_{med} の分散は近似的に $\pi^2/(4n)$ で与えられる．このことを利用して，最尤推定量と中央値の推定精度について議論せよ．

$$f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$$

ヒント　最尤推定量は漸近有効性をもつことを使う．

解答　**(1)** フィッシャーの情報量を計算すればよい．対数尤度とその微分は

$$l(\theta;x) = -\log(1+(x-\theta)^2) - \log\pi$$
$$\frac{\partial}{\partial\theta}l(\theta;x) = -\frac{2(x-\theta)}{1+(x-\theta)^2}$$
$$\frac{\partial^2}{\partial\theta^2}l(\theta;x) = -\frac{2}{1+(x-\theta)^2} + \frac{4(x-\theta)^2}{(1+(x-\theta)^2)^2}$$
$$= \frac{2}{1+(x-\theta)^2} - \frac{4}{(1+(x-\theta)^2)^2}$$

したがって，

$$E\left(-\frac{\partial^2}{\partial\theta^2}l(\theta;X)\right) = -\int_{-\infty}^{\infty}\frac{2}{\pi(1+(x-\theta)^2)^2}\mathrm{d}x + \int_{-\infty}^{\infty}\frac{4}{\pi(1+(x-\theta)^2)^3}\mathrm{d}x$$
$$J_k = \int_0^{\infty}\frac{1}{\pi(1+x^2)^k}\mathrm{d}x$$

とおくと

$$J_k = 2k\int_0^{\infty}\frac{x^2}{\pi(1+x^2)^{k+1}}\mathrm{d}x = 2kJ_k - 2kJ_{k+1} \Rightarrow J_{k+1} = \frac{2k-1}{2k}J_k$$
$$J_1 = 0.5$$

したがって，

$$E\left(-\frac{\partial^2}{\partial\theta^2}l(\theta;X)\right) = 2(4J_3 - 2J_2) = \frac{1}{2}$$

これより，標本の大きさが十分に大きいとき，$\hat{\theta}_n$ の分散は $2/n$

(2) $V(X_{\mathrm{med}}) \approx \pi^2/(4n)$ なので，

$$\frac{V(\hat{\theta}_n)}{V(X_{\mathrm{med}})} \approx \frac{8}{\pi^2} \approx 0.81$$

問題 6.7　(**分散比の区間推定**) パラメータ μ_i, σ_i^2 の正規母集団 $(i = 1, 2)$ から，それぞれ大きさ m, n の標本 $X_1, X_2, ..., X_m; Y_1, Y_2, ..., Y_n$ をとり，それらの不偏分散をそれぞれ S_X^2, S_Y^2 とする．このとき，2 つの母集団の分散の比 σ_1^2/σ_2^2 の $100(1-\alpha)\%$ 信頼区間は次で与えられることを示せ．ただし，$F_\alpha(m, n)$ は自由度 m, n の F 分布の上側 $100\alpha\%$ 点を表すものとする．

$$\left[\frac{S_X^2}{S_Y^2}\frac{1}{F_{\alpha/2}(m-1,n-1)}, \frac{S_X^2}{S_Y^2}F_{\alpha/2}(n-1,m-1)\right]$$

ヒント　$(S_X/\sigma_1)^2$ は自由度 $m-1$ のカイ 2 乗分布に従う．また，$F_\alpha(m,n)F_{1-\alpha}(n,m) = 1$ という関係がある．

解答　$(m-1)S_X^2/\sigma_1^2$ は自由度 $m-1$ のカイ 2 乗分布に従い，$(n-1)S_Y^2/\sigma_2^2$ は自由度 $n-1$ のカイ 2 乗分布に従う．したがって，$\left(S_X^2/\sigma_1^2\right) / \left(S_Y^2/\sigma_2^2\right)$ は自由度 $m-1, n-1$ の F 分布に従う．このことから，

$$\begin{aligned}\alpha &= P\left(F_{1-\alpha/2}(m-1,n-1) < \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} < F_{\alpha/2}(m-1,n-1)\right)\\ &= P\left(\frac{S_X^2}{S_Y^2}\frac{1}{F_{\alpha/2}(m-1,n-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\alpha/2}(m-1,n-1)}\right)\end{aligned}$$

が成り立つ．ここで，F 分布のパーセント点に対して成り立つ $F_\alpha(m,n)F_{1-\alpha}(n,m) = 1$ という関係を適用すると，表記の $100(1-\alpha)\%$ 信頼区間が得られる．

問題 6.8　$X_1, X_2, ..., X_n$ を正規母集団 $\mathrm{N}(\mu, \sigma^2)$ からの独立標本として，

$$T_1(\boldsymbol{X}) = \sum_{i=1}^n X_i, \quad T_2(\boldsymbol{X}) = \sum_{i=1}^n X_i^2$$

とおくと，$T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$ は未知パラメータ μ, σ^2 の十分統計量になることを示せ．

訂正　問題 6.1 と同問のため，削除．

問題 6.9　n 個のデータを $x_1, x_2, ..., x_n$，その中央値を X_{med} としたとき，n が奇数ならば，

$$\frac{1}{n}\sum_{i=1}^n |x_i - \theta|$$

を最小とする θ は $\theta = X_{\mathrm{med}}$ であることを示せ．n が偶数のとき，$\theta = X_{\mathrm{med}}$ は上の式を最小にすることを示せ．ただし，n 個のデータがすべて異なる場合を考えよ．

解答　$x_1, x_2, ..., x_n$ の順序統計量を $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ とする．$x_{(k)} < \theta < \theta' < x_{(k+1)}$ とすると，

$$\sum_{i=1}^n |x_i - \theta'| - \sum_{i=1}^n |x_i - \theta| = (2k-n)(\theta' - \theta)$$

なので，$k < n/2$ ならば θ は大きいほど平均絶対偏差は小さく，$k > n/2$ ならば θ は小さいほど平均絶対偏差は小さくなる．また，$\theta = x_{(k)}, \theta' = x_{(k+1)}$ とすると，

$$\begin{aligned}\sum_{i=1}^n |x_i - \theta'| - \sum_{i=1}^n |x_i - \theta| &= -2x_{(k+1)} + (2k+2-n)\theta' - (2k-n)\theta\\ &= (2k-n)(x_{(k+1)} - x_{(k)})\end{aligned}$$

となるので，$k < n/2$ ならば θ は大きいほど平均絶対偏差は小さく，$k > n/2$ ならば θ は小さいほど平均絶対偏差は小さくなる．以上から，n が奇数の場合は，中央値が平均絶対偏差を最小にする．

偶数の場合，$k = n/2$ のところで，平均絶対偏差の差が 0 になるので，$x_{(n/2)} < \theta < x_{(n/2+1)}$ のどの値をとっても平均絶対偏差を最小にする．したがって，特に $\theta = X_{\mathrm{med}}$ は平均絶対偏差を最小にする．

7　統計的検定問題

問題 7.1　母集団分布の密度関数が $\exp(\theta a(x) + b(\theta) + c(x))$ で与えられる母集団で，帰無仮説を $H_0 : \theta = \theta_0$, 対立仮説を $H_1 : \theta = \theta_1 (> \theta_0)$ とする θ の仮説検定を考えたとき，尤度比検定の棄却域が

$$\sum a(x_i) > d$$

の形で与えられることを示せ.

ヒント　分散 1 の正規分布の密度関数を書き直すと，

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

解答　帰無仮説の下での対数尤度は

$$\log L(\theta_0; \boldsymbol{x}) = \sum (\theta_0 a(x_i) + b(\theta_0) + c(x_i))$$

対立仮説の下での対数尤度は

$$\log L(\theta_1; \boldsymbol{x}) = \sum (\theta_1 a(x_i) + b(\theta_1) + c(x_i))$$

これより，

$$\frac{L(\theta_1; \boldsymbol{x})}{L(\theta_0; \boldsymbol{x})} > c \Leftrightarrow \log L(\theta_1; \boldsymbol{x}) - \log L(\theta_0; \boldsymbol{x}) > \log c$$

$$\log L(\theta_1; \boldsymbol{x}) - \log L(\theta_0; \boldsymbol{x}) = (\theta_1 - \theta_0) \sum a(x_i) + b(\theta_1) - b(\theta_0)$$

したがって，

$$d = \exp\left(\frac{\log c - (b(\theta_1) - b(\theta_0))}{\theta_1 - \theta_0}\right)$$

とすればよい.

問題 7.2　見かけ上はなんの細工もないさいころを 100 回振ったら 1 の目が 30 回出た．このさいころの 1 の目の出る確率は 6 分の 1 より大きい，という仮説を検定せよ.

解答　さいころの 1 の目が出る確率を p とすると，帰無仮説は $H_0 : p = \frac{1}{6}$, 対立仮説は $H_1 : \dot{p} > \frac{1}{6}$ である．帰無仮説の下で 1 の目が 100 回中 30 回以上出る確率は

$$\sum_{i=30}^{100} \binom{100}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{100-i} \approx 7 \times 10^{-4}$$

なので，有意水準 1% でも有意となり，帰無仮説は棄却され，対立仮説が採択される.

問題 7.3　(**平均値の差の検定**) 手作りパン屋では，パン生地を 40 グラムずつに分けてロールパンを作っている．ベテラン職人 A が切り分けた生地の重さと，新米職人 B が切り分けた生地の重さを量ったものが次の記録である．ベテラン職人の切り分けた生地の重さのほうがばらつきが小さい，という仮説を検定せよ.

職人	計測値
A	39.0, 39.6, 39.9, 40.4, 39.8, 39.7, 40.0, 40.4, 40.0, 39.4
B	39.5, 40.7, 40.6, 39.3, 38.9, 40.4, 41.6, 41.6, 42.3, 39.1

解答　職人 A の分散を σ_A^2，職人 B の分散を σ_B^2 とする．帰無仮説は $H_0 : \sigma_A^2 = \sigma_B^2$, 対立仮説は $H_1 : \sigma_A^2 < \sigma_B^2$ とする．データから，不偏分散は職人 A が 0.184，職人 B が 1.398 と計算される．8.1.2 項の統計量 $T_5(X, Y)$ を使うと，統計量の値は $1.398/0.184 = 7.6$ となる．自由度 $9, 9$ の F 分布を当てはめると，その上側 1% 点は 5.35 なので，有意水準 1% でも帰無仮説は棄却される．したがって，対立仮説が採択される．なお，検定統計量の p 値は 0.0029 である．

問題 7.4　ある電気製品の寿命は指数分布をしていると考えられている．新型製品の寿命は旧型製品の寿命の 1.5 倍であると宣伝されている．はたしてそうなのか調べてみたい．そこで，新型製品 $m = 20$ 台，旧型製品 $n = 10$ 台について寿命試験をしてみたところ，標本平均がそれぞれ $\bar{x} = 1500$ 時間，$\bar{y} = 1000$ 時間となった．新型，旧型製品の寿命分布の母平均をそれぞれ μ_X, μ_Y として，これらの母平均の比に関する推定および検定を考えてみよう．

(1) 上記の標本平均を確率変数と考えて $\bar{X}, \bar{Y}$ とすると，それらはどんな分布をするか？ (ヒント：平均が 2 の指数分布は自由度が 2 のカイ 2 乗分布に一致する．)

(2) 次式で定義される F はどのような分布をするか？　$F = (\bar{X}/\mu_X)/(\bar{Y}/\mu_Y)$

(3) μ_X/μ_Y の $100(1-\alpha)\%$ 信頼区間を求めよ．

(4) 帰無仮説を $\mu_X = \mu_Y$ 対立仮説を $\mu_X = 1.5\mu_Y$ として検定を行え．

(5) 他の形の帰無仮説および対立仮説を考えて検定を行え．

解答　**(1)** 平均が μ_X の指数分布に従う確率変数を X とすると，$2X/\mu_X$ は平均が 2 の指数分布になり，自由度が 2 のカイ 2 乗分布に従う．$m\bar{X}$ は m 個の独立な，パラメータ μ_X^{-1} の指数分布に従う確率変数の和になるので，$2m\bar{X}/\mu_X$ は自由度が $2m$ のカイ 2 乗分布に従い，そのモーメント母関数は

$$\left(\frac{1/2}{1/2-\theta}\right)^{2m}$$

である．$\bar{X}$ のモーメント母関数はこの式の θ を $\theta\mu_X/(2m)$ で置き換えればよいので，

$$\left(\frac{1/2}{1/2-\theta\mu_X/(2m)}\right)^{m} = \left(\frac{m/\mu_X}{m/\mu_X-\theta}\right)^{m}$$

これは，パラメータ $m, m/\mu_X$ のガンマ分布のモーメント母関数なので，$\bar{X}$ はパラメータ $20, 20/\mu_X$ のガンマ分布に従う．同様に，$\bar{Y}$ は，パラメータ $10, 10/\mu_Y$ のガンマ分布に従う．

(2) (1) と同様に考えて $2n\bar{Y}/\mu_Y$ は自由度 $2n$ のカイ 2 乗分布に従う．$2m\bar{X}/\mu_X$ が自由度 $2m$ のカイ 2 乗分布に従うことと併せて，

$$F = \frac{2m\bar{X}/\mu_X/(2m)}{2n\bar{Y}/\mu_Y/(2n)} = \frac{\bar{X}/\mu_X}{\bar{Y}/\mu_Y}$$

は，自由度 $2m, 2n(= 40, 20)$ の F 分布に従う (命題 5.6)．

(3) (2) より，

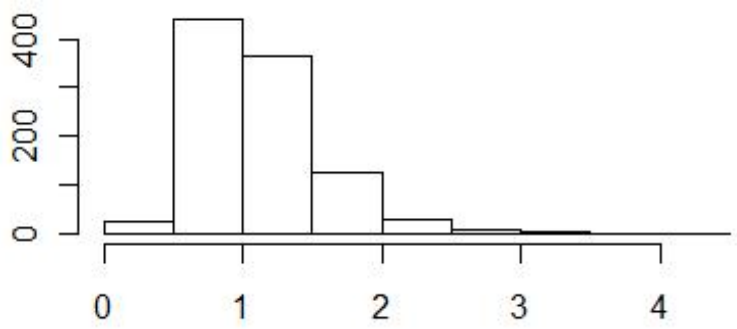
$$P\left(\frac{\bar{X}/\mu_X}{\bar{Y}/\mu_Y} \le x\right) = P\left(\frac{1}{x}\frac{\bar{X}}{\bar{Y}} \le \frac{\mu_X}{\mu_Y}\right) = F_{2n,2m}(x)$$

となる．したがって，μ_X/μ_Y の $100(1-\alpha)\%$ 信頼区間は

$$\left[\frac{1}{F_{\alpha/2}(2m,2n)}\frac{\bar{X}}{\bar{Y}}, \frac{1}{F_{1-\alpha/2}(2m,2n)}\frac{\bar{X}}{\bar{Y}}\right]$$

$\bar{X}/\bar{Y} = 1.5$ の場合，いくつかの α について信頼区間を計算してみると，次のようになる．

$$\begin{array}{ll} \alpha = 0.1 \text{ の場合} & [0.752, 2.758] \\ \alpha = 0.05 \text{ の場合} & [0.656, 3.102] \\ \alpha = 0.01 \text{ の場合} & [0.496, 3.898] \end{array}$$



(4) 帰無仮説の下で $F = \bar{X}/\bar{Y}$ は自由度 $40, 20$ の F 分布に従う．対立仮説は $\mu_X = 1.5\mu_Y$ なので，棄却域は $F > F_\alpha(2m, 2n)$ である．今の場合，$F = 1.5$ だが，$F_{0.1}(40, 20) = 1.71$ なので，10% でも有意にならず，帰無仮説は棄却されない．

参考のために，平均 1000 の指数乱数 20 個の標本平均と，10 個の標本平均の比がどれくらいばらつくかを R で実験した結果を示す．指数乱数がかなりばらつくことと，比をとっていることにより，この程度の標本サイズでは，あまり精度のよい推定ができないことがわかる．

プログラム例

```
na = 20; ma = 1/1000; nb = 10; mb = 1/1000
nn = 1000
z = sapply(1:nn, function(x) mean(rexp(na,ma))/mean(rexp(nb,mb)) )
hist(z)
c(length(which(z> 1.5))/nn, 1-pf(1.5,40,20))
```

(5) 寿命が長くなるという場合の自然な対立仮説は $H_1 : \mu_X > \mu_Y$ であろう．

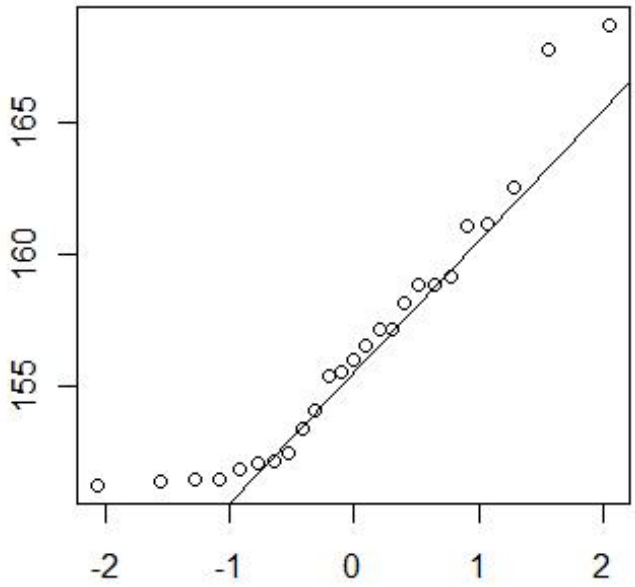


図 1　(問題 8.1) 計測データの Q–Q プロット

8　さまざまな推定・検定

問題 8.1　次のデータはある業者から納入される品物の計測データである．このデータは正規分布に従っているといえるか，P–P プロットと Q–Q プロットの図を描いて調べよ．

151.3,157.2,156.0,152.2,151.4,154.1,158.9,158.9,155.4,156.6,
151.9,153.4,167.8,168.7,157.2,158.2,155.6,152.5,161.2,152.1,
151.5,159.2,161.1,162.6,151.5

解答　Q–Q プロットは図 1, P–P プロットは図 2 のようになる．P–P プロットでは直線の周りに分布しているといえなくはないが，Q–Q プロットを見れば，小さいデータが多く出現していて，正規分布とは別の分布を想定したほうが適切ではないかと思われる．

R のプログラムは以下の通り．

プログラム例

```
data = c(151.3,157.2,156.0,152.2,151.4,154.1,158.9,158.9,155.4,156.6,
    151.9,153.4,167.8,168.7,157.2,158.2,155.6,152.5,161.2,152.1,
    151.5,159.2,161.1,162.6,151.5)
hist(data)
qqnorm(data, xlim=c(-2.5,2.5))
qqline(data)
n = length(data)
qqplot((1:n-0.5)/n, pnorm(data,mean(data),sd(data)), xlim=c(0,1), ylim=c(0,1))
abline(0,1)
```

問題 8.2　交通事故による 1 日の死者数はポアソン分布に従うといわれているので検証してみたい．

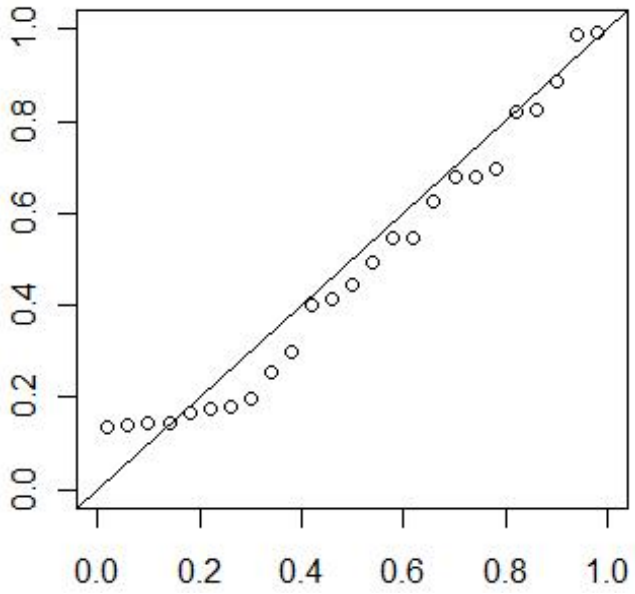


図 2　(問題 8.1) 計測データの P–P プロット

2010 年の神奈川県内交通事故の死者数は次のようになっている．平均値を計算し，その平均値をポアソン分布のパラメータの推定値として，期待度数を計算し，カイ 2 乗適合度検定を実施せよ．

死者数	0 人	1 人	2 人	3 人	4 人
日数	226	104	28	6	1

解答　データ数は 365 平均値は 0.5 なので，パラメータ $\lambda = 0.5$ のポアソン分布の確率関数に 365 を掛けて期待度数を得る．3 人と 4 人をまとめて 3 人以上とし，クラス数 $m = 4$ の度数分布表を作ると，検定に用いる自由度はポアソン分布のパラメータを推定したことにより $(m-1)-1 = 2$ になる．統計量の値 1.09 を自由度 2 のカイ 2 乗分布の上側 5% 点 $\chi^2_{0.05}(2) = 6.0$ と比べると，十分に小さく，採択域に入っているので，1 日の死者数がポアソン分布に従うといってもよいことが確かめられた．

死者数	0 人	1 人	2 人	3 人以上
日数	226	104	28	7
期待度数	221.7	110.5	27.6	5.2

平均は約 0.5 パラメータ 0.5 のポアソン分布で 4 以上の値をとる確率は 0.0017 データ数 365 とした場合の期待度数は 1 未満なので，3 人と 4 人のクラスを 3 人以上というクラスに合併する．後の手順は例 8.5 と同じ．`chisq.test` を使わないで，定義通りに統計量を計算し，その p 値を `pchisq` 関数を使って計算することもできる．その結果，統計量の値は 1.09 p 値は 0.78 となり，ポアソン分布に従っているという仮説は棄却されない．平均値プログラムは以下の通り．

プログラム例

```
data = c(226,104,28,6,1)
n = sum(data)              # データの個数
m = sum(data * (0:4)) / n     # データの平均値
obs = c(data[1:3],7)          # 右端の 2 つのクラスをまとめる
exp = c(dpois(0:2,m), 1-ppois(2,m))
chisq.test(obs, p=exp)
## 同じことを関数で実現
```

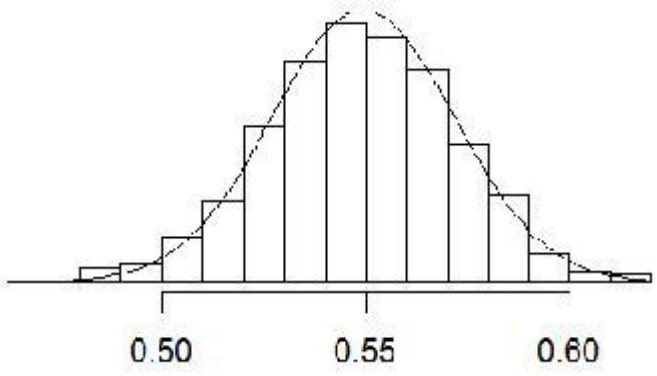


図 3　(問題 8.2) フィッシャーの Z 変換 ($\rho = 0.5$)

```
chi2 = sum((obs - n*exp)^2 / (n*exp))
1 - pchisq(chi2, 3)
```

問題 8.3　$\mu = \nu = 0, \sigma^2 = \tau^2 = 1$ で，相関係数が $\rho = 0.5$ の 2 変量正規分布に従う乱数を $n = 20$ 個生成してそれらの標本相関係数を計算し，(8.33) 式で定義されたフィッシャーの Z を計算するという実験を $N = 1000$ 回繰り返してヒストグラムを描き，正規分布のように分布していることを確かめよ．さらに，それに平均 $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$, 分散 $1/(n-3)$ の正規分布の密度関数を重ねて描き，フィッシャーの Z 変換のもっともらしさを確かめよ．$\rho = 0.9$ として同じ実験を行え．

ヒント　rr を N 個の標本相関係数のベクトルとすると，「hist((log(1+rr)-log(1-rr))/2, freq=F)」によって，Z 変換したデータのヒストグラムが描ける．ここで，「freq=F」は縦軸を相対度数とするオプション指定で，こうすることにより，正規分布の密度関数を描いたときに形状を比較ができる．

解答　相関のある正規乱数の生成法は実験 8.4 に書いてある．$n = 20$ 個の相関のある 2 変量正規乱数を生成してその相関係数を計算する，という計算を繰り返して N 個の相関係数を集めるのは sapply 関数を使う．あとはヒントに従う．結果は $\rho = 0.5$ の場合が図 3, $\rho = 0.9$ の倍が図 4 に示す通りである。これらの図より，確かに正規分布の密度関数で近似的に表現できている．なお，hist(rr) としたときに，どのような (歪んだ) グラフになるか試してみるとよい．

プログラムは以下の通り．

プログラム例

```
n = 20
r = 0.5
N = 10000
rr = sapply(rep(n,N), function(n) {
    z = rnorm(n)
    w = r*z+sqrt(1-r^2)*rnorm(n)
    cor(z,w)})
hist((log(1+rr)-log(1-rr))/2, freq=F)
curve(dnorm(x,(log(1+r)-log(1-r))/2, 1/sqrt(n-3)), add=T)
```

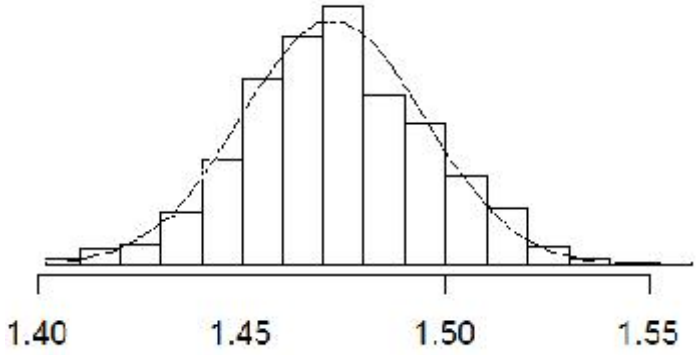


図 4　(問題 8.2) フィッシャーの Z 変換 ($\rho = 0.9$)

9　回帰分析

使用するデータは，`http://www.asakura.jp/download.html` からダウンロードできる．

問題 9.1　次の表は，1891〜2010 年の各年の世界の年平均気温の基準値からの偏差 (単位：°C) を示したものである．ただし，基準値は，1981〜2010 年の 30 年平均値である (出典：気象庁ホームページ)．このデータについて回帰分析を行え．

年	平均気温 (基準年との差)									
1891〜	−0.63	− 0.71	− 0.75	− 0.70	− 0.68	− 0.47	− 0.49	− 0.66	− 0.56	− 0.49
1901〜	−0.58	− 0.70	− 0.77	− 0.83	− 0.70	− 0.60	− 0.78	− 0.82	− 0.82	− 0.78
1911〜	−0.81	− 0.73	− 0.70	− 0.53	− 0.43	− 0.64	− 0.71	− 0.55	− 0.58	− 0.51
1921〜	−0.43	− 0.56	− 0.54	− 0.56	− 0.46	− 0.36	− 0.47	− 0.48	− 0.60	− 0.38
1931〜	−0.34	− 0.38	− 0.53	− 0.38	− 0.46	− 0.48	− 0.37	− 0.34	− 0.37	− 0.32
1941〜	−0.26	− 0.26	− 0.23	− 0.12	− 0.26	− 0.40	− 0.43	− 0.41	− 0.42	− 0.49
1951〜	−0.36	− 0.30	− 0.23	− 0.46	− 0.47	− 0.56	− 0.28	− 0.23	− 0.29	− 0.33
1961〜	−0.24	− 0.22	− 0.19	− 0.49	− 0.43	− 0.36	− 0.36	− 0.38	− 0.27	− 0.30
1971〜	−0.41	− 0.29	− 0.17	− 0.44	− 0.39	− 0.48	− 0.19	− 0.28	− 0.16	− 0.13
1981〜	−0.09	− 0.21	− 0.06	− 0.24	− 0.25	− 0.17	− 0.01	− 0.03	− 0.10	+ 0.04
1991〜	−0.03	− 0.17	− 0.14	− 0.07	+ 0.01	− 0.09	+ 0.09	+ 0.22	0.00	0.00
2001〜	+0.12	+ 0.16	+ 0.16	+ 0.12	+ 0.17	+ 0.16	+ 0.13	+ 0.05	+ 0.16	+ 0.19

解答　Excel で入力データを整え，1 列目に「年」，2 列目に「気温差」というラベルを付ける．その項目名を含めて 101 行分のデータをコピーして R に渡す．R のプログラムは以下の通り．`lm` 関数の `data=`オプションは，一時的に `attach(temp)` を実行したのと同じ効果をもたらす．

プログラム例

```
temp < - read.table("clipboard", header=T)
plot(temp, type="l")
model < - lm(気温差 ~ 年, data=temp)
abline(model)
summary(model)
```

実行結果は図 5 のようになる．切片は −13.62，回帰係数は 0.0068，したがって，100 年間で約 0.7° の気温上昇ということになる．また，回帰分析の詳細な結果を見るまでもなく，切片，傾きと

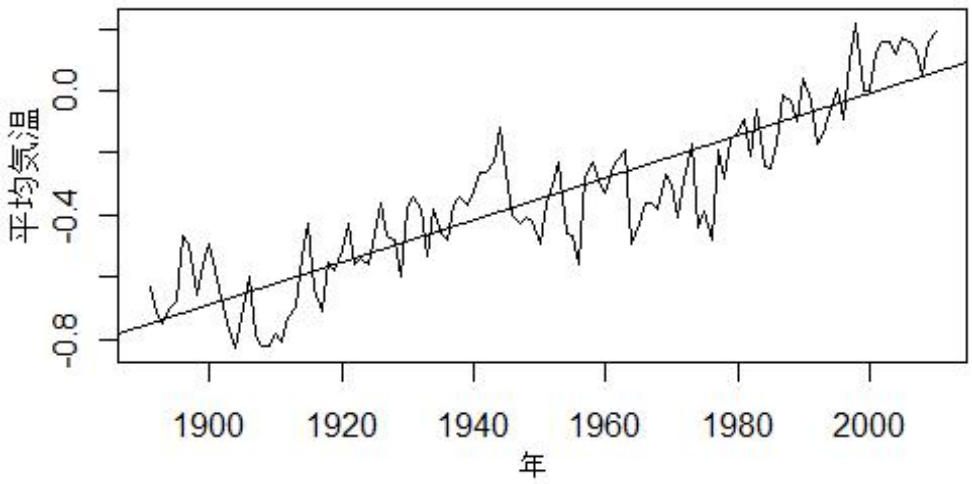


図 5　(問題 9.1) 世界の平均基本変化

も高度に有意，t-値の絶対値が 20 を超える．

問題 9.2　次の表は，圧力と水の沸点との関係を示したものである (圧力の単位は mmHg, 沸点温度の単位は °C. 1mmHg = 133.322Pa)．沸点と圧力の関係を表す回帰式を求めよ．

圧力	沸点	圧力	沸点
680	96.92	730	98.88
685	97.12	735	99.07
690	97.32	740	99.26
695	97.52	745	99.44
700	97.72	750	99.64
705	97.91	755	99.82
710	98.11	760	100.00
715	98.30	765	100.18
720	98.50	770	100.36
725	98.69	775	100.55

解答

前問と手順は同じ．切片は 70.99 回帰係数は 0.038 となった．なお，圧力を P, 沸点を T とすると，

$$T = \frac{1730}{8.07 - \log_{10} P} - 233.4$$

という近似式がある (アントワン近似式)．$680 \leq P \leq 775$ の範囲でほとんど直線になる (図 6)．この近似式を重ねて描く命令も追加した R のプログラムは以下の通り．

プログラム例

```
boil < - read.table("clipboard", header=T)
plot(boil)
model < - lm(沸点 ~ 圧力, data=boil)
abline(model)
curve(1730/(8.07-log10(x))-233.4, add=T, col=2)
```

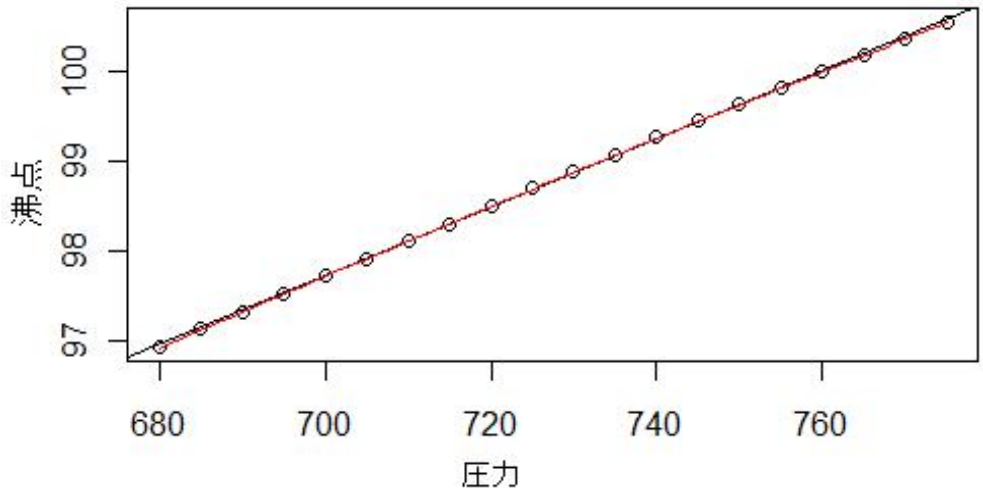


図 6　(問題 9.2) 圧力と沸点の関係

問題 9.3　金属の電気抵抗は，温度が上昇すると増加する．常温付近で，温度の変化する範囲がそれほど広くない (たとえば 0〜100 °C) 場合には，電気抵抗の増加は温度の増加にほぼ比例する．次の表は，銅線の電気抵抗を測定した結果を示したものである．このデータについて回帰分析を行い，電気抵抗と温度の関係式を求めよ．

温度 (°C)	抵抗 (Ω)
12.0	14.7
20.5	15.2
31.0	15.9
39.5	16.3
50.5	17.0
61.0	17.7
69.0	18.1
81.5	18.9
92.0	19.5

解答　前問と手順は同じ．切片は 13.98, 回帰係数は 0.06 となった．R のプログラムは以下の通り．

プログラム例

```
regist < - read.table("clipboard", header=T)
plot(regist)
model < - lm(抵抗 ~ 温度, data=regist)
abline(model)
```

問題 9.4　次の表は，日銀券の発行高を示したものである (出典：日銀ホームページ；単位：億円)．各年 (暦年) について，毎月の月末の発行高の平均値が示されている (ただし，2011 年については，1〜10 月の 10 ヵ月の平均である)．このデータについて回帰分析を行え．

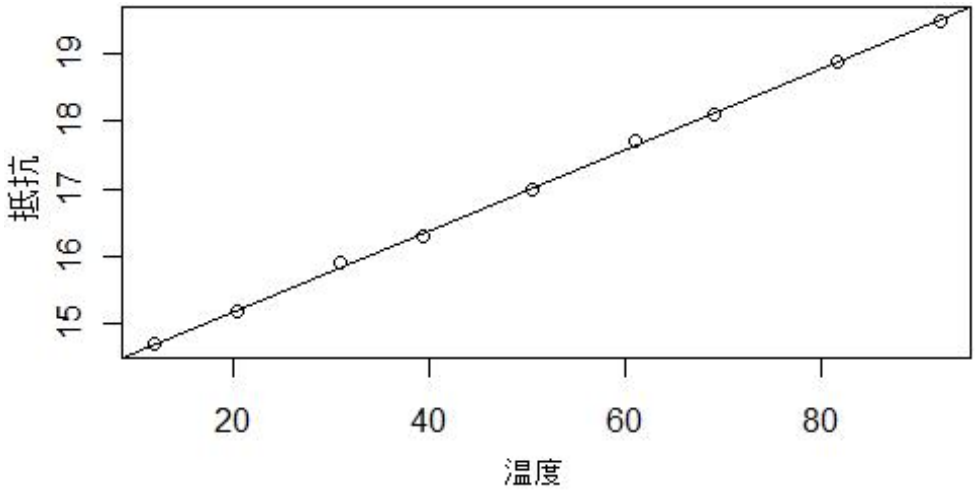


図 7　(問題 9.3) 温度と金属の電気抵抗の関係

年	発行高	年	発行高
2000	561,632	2006	750,898
2001	601,260	2007	761,537
2002	678,085	2008	767,004
2003	711,254	2009	770,081
2004	722,771	2010	777,525
2005	743,803	2011	794,310

解答　データをプロットしてみると，明らかに 2003 年前後から傾向が変化している．機械的に 1 つの回帰式を当てはめるのは適切でない．2004 年以降のデータだけを使って，前問と同じ手順で実行すると，切片は -1814.87, 回帰係数は 0.94 となった．R のプログラムは以下の通り．

プログラム例

```
bank < - read.table("clipboard", header=T)
plot(bank)
n < - length(bank[,1])
x < - bank[4:n,1]
y < - bank[4:n,2]/10000
plot(x, y)
model < - lm(y ~x)
abline(model)
```

問題 9.5　二酸化炭素 (CO_2) は，地球温暖化に大きな影響力をもっている．次の表は，気象庁の大気環境観測所 (岩手県大船渡市三陸町綾里) において観測された大気中二酸化炭素の年平均濃度 (単位：ppm) を示したものである (出典：気象庁ホームページ；2010 年の数値は暫定値)．このデータについて回帰分析を行い，濃度増加の長期的な傾向を調べよ．

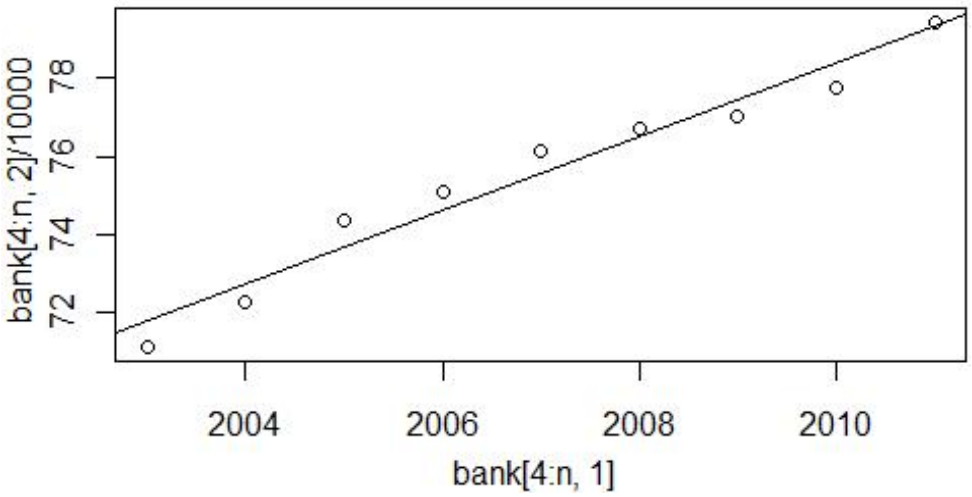


図 8　(問題 9.4) 日銀券の発行高

年	年平均濃度	年	年平均濃度	年	年平均濃度		
1987	351.2	1993	359.3	1999	371.2	2005	382.5
1988	354.0	1994	361.7	2000	372.6	2006	385.3
1989	355.9	1995	363.6	2001	373.4	2007	386.6
1990	356.7	1996	365.1	2002	375.9	2008	388.5
1991	358.1	1997	366.4	2003	378.6	2009	389.7
1992	358.4	1998	369.4	2004	380.3	2010	393.3

解答　前問と手順は同じ．切片は −3190, 回帰係数は 1.78 となった．R のプログラムは以下の通り．

プログラム例

```
carbon < - read.table("clipboard", header=T)
plot(carbon)
model < - lm(carbon$ 年平均濃度 ~carbon$ 年)
abline(model)
```

問題 9.6　次の表は，1980〜1990 年の各月における日銀券の平均発行高 (季節調整済み) を示したものである (出典：日本銀行ホームページ；単位は億円).

(1) このデータについて回帰分析を行え．当てはめた回帰直線と元データとの差 (残差) をプロットして，その傾向を観察せよ.

(2) 次に，まずデータの対数をとってから，回帰分析を行ってみよ．この結果についても残差のプロットを行い，(1) の結果と比較せよ.

(3) (2) の分析結果を基に，この期間の日銀券の発行高の増加率は年平均で何%くらいになっているか考えよ.

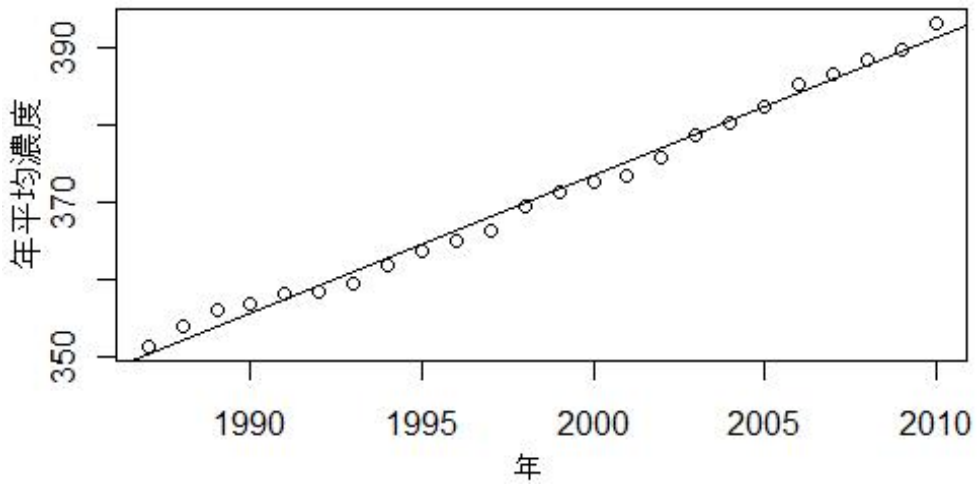


図 9　(問題 9.5) 大気中二酸化炭素の年平均濃度 (単位：ppm)

	1980	1981	1982	1983	1984	1985
Jan	149290	154297	162152	173582	180403	193772
Feb	150610	154498	164143	176415	181975	193688
Mar	152116	155210	165118	177102	182462	194620
Apr	153825	154960	166731	177448	183694	195133
May	152982	156890	168675	177585	183680	195561
Jun	153472	158161	169171	178484	184490	196065
Jul	152391	159092	170050	178338	186360	197285
Aug	152403	160027	171393	179903	186144	197183
Sep	152319	160453	172282	180323	188224	198686
Oct	152792	161303	173955	181817	188558	199782
Nov	154065	162508	174127	180164	191834	200745
Dec	153465	164109	174529	181596	192504	204070

	1986	1987	1988	1989	1990	
Jan	204689	222582	243794	269910	305238	
Feb	204988	225245	249722	278858	306755	
Mar	205539	227190	253436	280406	307577	
Apr	206620	228367	256841	282062	323003	
May	208133	230983	259236	285045	311943	
Jun	210622	232412	258905	287105	311124	
Jul	212372	233994	260680	289165	312791	
Aug	215357	237076	261474	291624	313791	
Sep	216353	238919	263053	293711	315943	
Oct	217832	241289	265498	295054	316387	
Nov	220363	243352	266381	297990	316000	
Dec	219520	243233	268020	299816	317828	

解答　**(1)** 年月データの代わりに通算月 (1 から 132 まで) を説明変数として回帰分析を行う．プログラムは以下の通り．

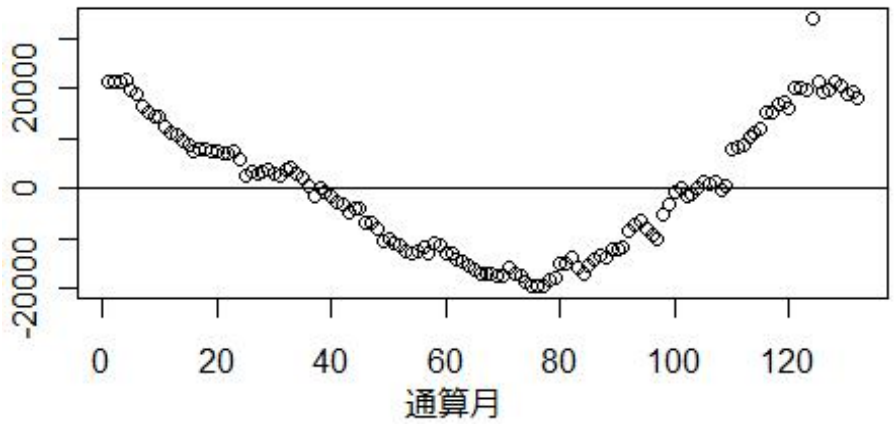


図 10　図 9.6(a) 日銀券の月別発行高の線形回帰誤差

プログラム例

```
banka < - read.table("clipboard", header=T)
attach(banka)
model < - lm(発行高 ~ 通算月)
plot(banka[,3])              ## データプロットと回帰直線
abline(model)
## 予測値の計算と残差プロット
ahat < - model$coefficients[1]
bhat < - model$coefficients[2]
res < - ahat + bhat*banka[,1]
plot(banka[,3]-res)
abline(h=0)
```

残差プロットは次のようになり，直線で回帰させることはあまり適切でないことがわかる．

(2) 対数は log 関数を使う．残差の計算は実験 9.1 の「ahat + bhat * 保有車両数」を計算する R の関数「predict(model)」があるので，それを使ってもよい．プログラムは以下の通り．

プログラム例

```
logbank < - log(発行高)
logkaiki < - lm(logbank ~ 通算月)
summary(logkaiki)
plot(logbank-predict(logkaiki))
abline(h=0)
```

結果は次の通りで，前の図と同じ傾向を示すが，縦軸の数値を見れば，大きく改善されていることがわかる．元データをプロットしたときに単調な逓増，あるいは逓減傾向にある場合はデータの対数をとったものが直線傾向を示すことが多いので，試してみる価値はある．

(3) 対数変換されたデータの近似直線の傾きは「logkaiki$coefficients[2]」で知ることが出来，この場合は 0.006 であった．

発行高の年平均増加率は，i 月の発行高を y_i とすると，

$$\frac{y_{i+12} - y_i}{y_i}$$

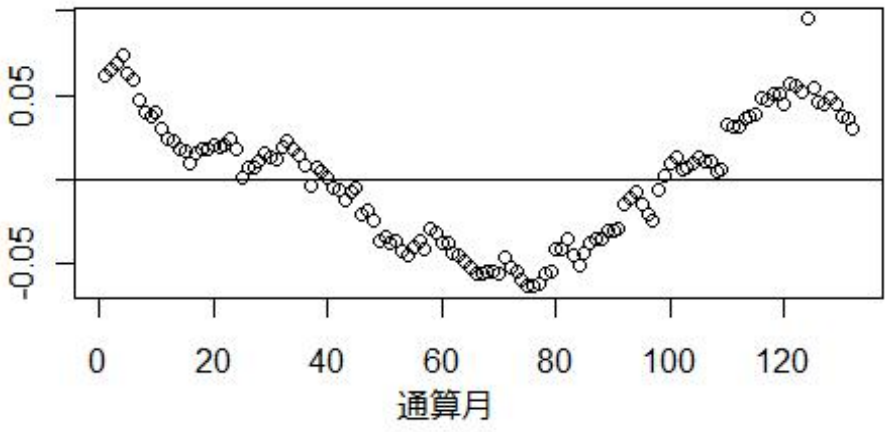


図 11　図 9.6(b) 日銀券の発行高 (対数変換) の線形回帰誤差

その対数をとったもの $z_i = \log y_i$ が傾き 0.006 で増加しているので，

$$\frac{y_{i+12} - y_i}{y_i} = \frac{\mathrm{e}^{z_{i+12}}}{\mathrm{e}^{z_i}} - 1 = \mathrm{e}^{z_{i+12} - z_i} - 1 = \mathrm{e}^{0.072} - 1 = 0.0747$$

したがって，年平均増加率は約 7.47% になる．

10　分散分析

問題 10.1　次のデータは都府県 4 ヵ所の 7 日分の放射線量測定データである．一元配置モデルを用いた分散分析の手順を実行し，平均値の違いについて分析せよ．

地点	1 日目	2 日目	3 日目	4 日目	5 日目	6 日目	7 日目
H	54	30	51	30	30	33	30
I	88	83	83	82	82	84	79
T	58	57	56	57	57	56	55
O	44	44	45	45	50	49	48

解答　この問題の因子は地点で，全体のばらつきに比べて各水準ごとのばらつきが小さいので，平均値に差があることは明白である．それを分散分析によって数値的に確かめる，という問題．例 10.1 にならって全体の散布図を描き，例 10.2 にならって分散分析を実施する．プログラムは以下の通り．読み込むデータ行列は，上の表の行と列を入れ替えたものとしている．その結果，p 値がほとんどゼロ (6×10^{-13}) なので，各水準ごとの平均値は明らかに異なるということが数値的にも確かめられる．

プログラム例

```
data < - read.table("clipboard", header=T)
data1 < - c(data[,2], data[,3], data[,4], data[,5])
level < - c(rep(1,7), rep(2,7), rep(3,7), rep(4,7))
level < - factor(level)
summary(aov(data1 ~ level))
## 実行結果
            Df Sum Sq Mean Sq F value    Pr(> F)
level        3 8309.4 2769.81  84.759 6.547e-13 ***
Residuals 24  784.3   32.68
## 以下の 5 行は確認のため
plot(level,data1,xaxp=c(1,4,3),xlab="地点")
points(1:4,c(mean(data[,2]),mean(data[,3]),mean(data[,4]),mean(data[,5])),pch=16)
lines(1:4,c(mean(data[,2]),mean(data[,3]),mean(data[,4]),mean(data[,5])))
lines(1:4,c(mean(za),mean(zb),mean(zc),mean(zd)))
abline(h=mean(data1), lty=2)
```

問題 10.2　インターネットを利用して購入した 9 銘柄の放射線測定器を使って，同一試料から放射される放射線量率を 10 回ずつ繰り返し測定して，次の表のような結果を得たものとする (この表の数値は，(独) 国民生活センターが行った試験結果を参考にして人工的に作ったものであり，実際の

測定値を示すものではない).

銘柄	10 回の測定値 (マイクロシーベルト/時)									
1	4.736	4.750	4.886	4.918	4.931	4.875	4.861	4.682	4.937	4.926
2	4.215	4.628	4.847	4.457	4.878	4.408	4.873	4.817	4.807	4.443
3	2.495	2.627	2.379	2.432	2.761	2.557	2.087	2.339	2.439	2.603
4	4.317	2.141	1.568	4.903	3.808	1.702	3.453	2.722	1.910	2.489
5	2.575	5.810	3.167	4.500	1.563	1.113	2.222	4.254	1.672	6.060
6	5.241	4.890	5.775	4.918	5.866	5.842	5.517	4.941	4.505	6.068
7	6.663	4.383	5.105	4.691	4.038	3.025	5.302	6.033	3.965	4.431
8	5.002	3.968	4.805	4.285	4.682	4.197	4.151	3.806	4.218	3.553
9	2.596	3.033	2.748	2.652	2.514	2.770	2.804	2.886	2.820	3.092

(1) 各銘柄の測定値の平均，標準偏差，および変動係数を計算し，箱ひげ図 (第 2 章参照) を描け.

(2) このデータに対して分散分析を行うことの妥当性を検討せよ.

(3) このデータに対して，(上記のような妥当性を検討することなく) 分散分析を行って，どのような“結論”が得られるか試みよ.

解答　(1)

平均	4.850	4.637	2.472	2.901	3.294	5.356	4.764	4.267	2.792
標準偏差	0.093	0.240	0.186	1.160	1.777	0.530	1.058	0.452	0.181
変動係数	0.019	0.052	0.075	0.400	0.540	0.099	0.222	0.106	0.065

箱ひげ図は図 12 の通り.

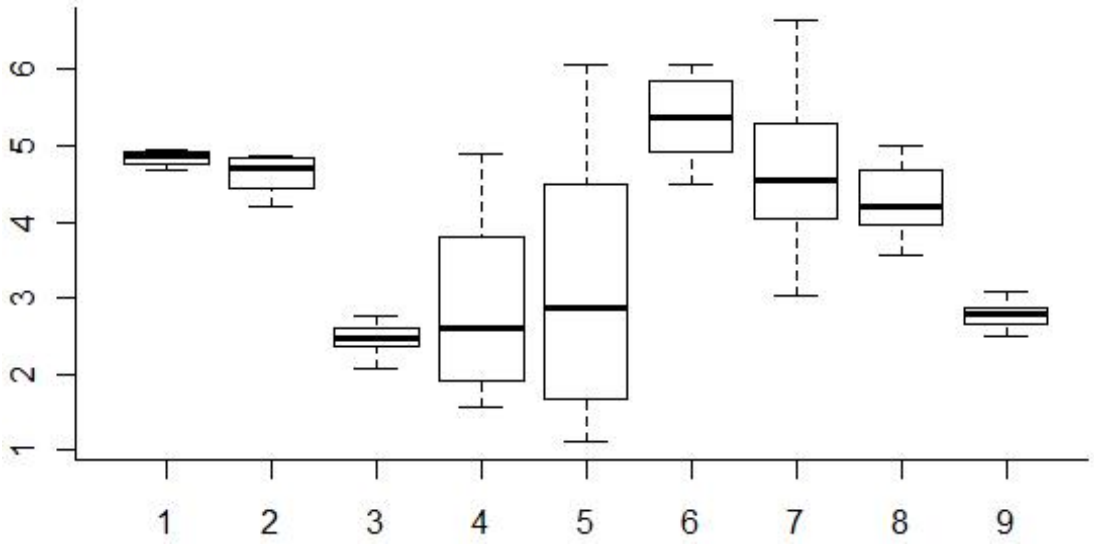


図 12　(問題 10.1) 9 種類の放射線測定器の測定値のばらつき

次のプログラムでは，上の表の行と列を転置したものを入力している.

プログラム例

```
data < - read.table("clipboard", header=T)
brand < - factor(floor((10:99)/10))
```

```
count <- c(); for(i in 1:9) count = c(count, data[,i])
## 基本統計量（平均，標準偏差，変動係数）
apply(data, 2, mean)
apply(data, 2, sd)
apply(data, 2, sd) / apply(data, 2, mean)
## 箱ひげ図
plot(level,count)
```

(2) 分散分析を適用するには各グループの分散が等しいという前提が必要であるが，このデータをみると，明らかに銘柄によってばらつきが異なる．したがって，分散分析を機械的に当てはめても，正しい結論は得られない．

(3) 形式的に分散分析を行うと，(summary(aov(count ~brand))) 次のような結果が得られる．予想通り，銘柄ごとの平均値は変わらない，という仮説は強く棄却される．

プログラム例

```
> summary(aov(count ~brand))
            Df Sum Sq Mean Sq F value    Pr(
> F)
brand      8 90.749 11.3437  16.357 3.869e-14 ***
Residuals 81 56.174  0.6935
```

問題 10.3　次の表は，福島県内 3 ヵ所，茨城県内 1 ヵ所，および千葉県内 1 ヵ所の合計 5 ヵ所の定点観測地点において，ある日の午前 3 時から 12 時まで 1 時間ごとに測定された放射線量を示している．このデータを分析して，観測地点によって，また観測時刻によって，放射線量に有意な差があるかどうかを検討せよ．

地点	1 時間ごとの測定値 (マイクロシーベルト/時)
1	0.123 0.120 0.123 0.118 0.121 0.120 0.125 0.126 0.127 0.125
2	0.247 0.248 0.248 0.248 0.243 0.257 0.252 0.253 0.247 0.250
3	0.257 0.261 0.262 0.262 0.263 0.259 0.262 0.259 0.254 0.251
4	0.108 0.110 0.109 0.110 0.111 0.112 0.108 0.109 0.114 0.107
5	0.081 0.078 0.076 0.075 0.074 0.076 0.079 0.078 0.077 0.080

解答　地点別に時系列グラフをプロットしてみると，地点 2, 3 は他地点に比べて 2 倍くらいの値をとっており，また各地点ごとのばらつきは小さく，平均値に差があることは間違いがないが，念のため分散分析をしてみると，p 値は 10^{-16} とほとんどゼロとなる．

観測時刻の影響を調べるには，地点と観測時刻を要因とする二元配置モデルを適用して分散分析を行う．その結果，時間要因の p 値は 0.87 なので，観測時刻が放射線量の観測値に影響を与えているという仮説は棄却される．

プログラム例

```
>time
< - factor(rep(1:10,5))
> summary(aov(count ~point + time))
          Df Sum Sq Mean Sq F value Pr(
> F)
point      4 0.284352 0.071088 6533.8125
< 2e-16 ***
time       9 0.000048 0.000005    0.4904 0.8713
```

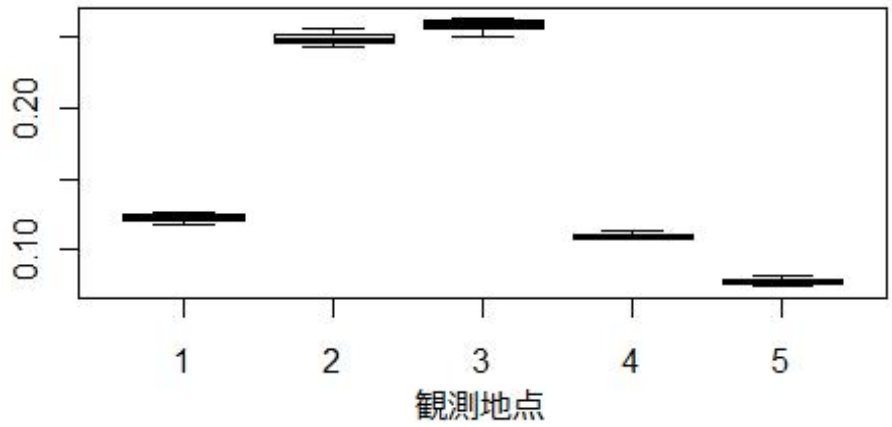


図 13　(問題 10.2) 各地点の測定値の時間変化

```
Residuals 36 0.000392 0.000011
```

問題 10.4　ある市内に住んでいる A, B, C, D の 4 人が，市が管理している市民農園を借りてジャガイモを栽培した．ジャガイモは 3 種類 (男爵，メークイン，インカのめざめ) あって，4 人ともこれらをすべて同じ面積の場所に作付した．市民農園は，1 枚の畑を小さく区切って大勢の市民に貸しているものなので，どの場所でも地力は同じであると考えられる．次の表は，各人の 3 種類のジャガイモの収穫量 (単位：kg) を示している．

	A	B	C	D
男爵	2.9	3.3	3.0	2.8
メークイン	2.8	3.2	2.9	2.7
インカのめざめ	2.7	2.9	3.1	3.0

このデータについて分散分析を行って，以下の問いに答えよ．

(1) 人によって収穫量に有意な差があるかどうか？

(2) ジャガイモの種類によって収穫量に差があるかどうか？

解答　**(1)** 前問と同じように，人を要因としたときの箱ひげ図を描くと図 14 のようになる．作業者 A と B は差がありそうに見えるが，4 人の作業者ごとの平均値には差がない，という帰無仮説は p 値 0.078 となり，10% 有意であっても，5% では有意にならない．

(2) 品種の違いについては，「インカのめざめ」が他の 2 品種に比べて少なめに見えるが，統計的にみて有意な差はないだろうと思われる．実際，分散分析による計算結果から，p 値は 0.77 となり，予想が裏付けられる．

プログラム例

```
## (1) 作業者ごとのばらつき
data < - read.table("clipboard", header=T)
person < - factor(c(rep("A",3),rep("B",3),rep("C",3),rep("D",3)))
count < - c(data[,1],data[,2],data[,3],data[,4])
summary(aov(count ~person))
plot(person, count)
## (2) 品種ごとのばらつき
```

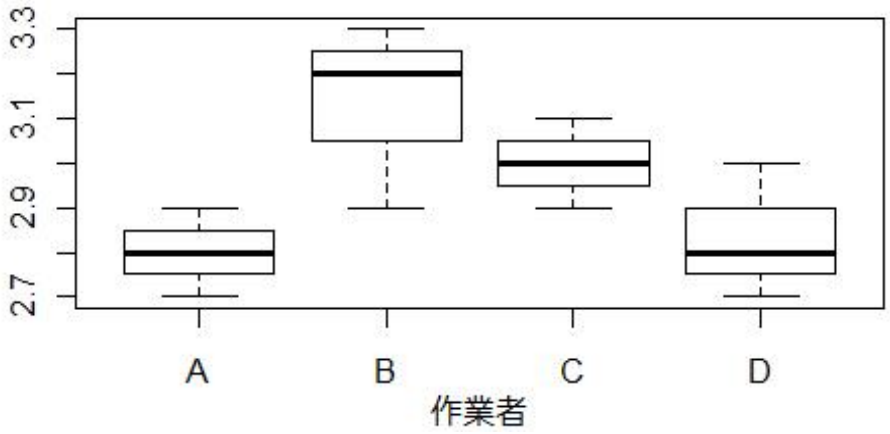


図 14　(問題 10.4) 作業者ごとの収穫量のばらつき

```
brand
< - factor(c(rep("男爵",4),rep("メイクィン",4),rep("インカ",4)))
count
< - c(t(data[1,]),t(data[2,]),t(data[3,]))
summary(aov(count ~brand))
```

問題 10.5　ある化学工場で，製品の収率を改善することを目的として実験を行った．因子としては，反応の温度 (A1, A2, A3, A4 の 4 水準) と触媒の種類 (B1, B2, B3 の 3 種類) を取り上げ，各温度と触媒の組み合わせについて 3 回ずつ測定を行って，次の表の結果を得た (単位は %)．このデータについて分散分析を行い，温度あるいは触媒の種類によって収率に有意な差があるかどうかを調べよ．また，交互作用についても検討せよ．最適な水準の組み合わせはどれか考えよ．

	A1	A2	A3	A4
B1	55.6, 56.6, 51.1	64.0, 60.8, 67.1	58.5, 60.0, 57.7	51.5, 55.7, 48.4
B2	66.3, 65.6, 65.7	72.3, 71.1, 67.1	54.8, 55.1, 55.0	60.7, 59.0, 59.9
B3	52.0, 50.9, 52.6	55.8, 58.3, 56.8	49.8, 52.6, 53.0	43.1, 47.2, 41.5

解答　反応温度の 4 水準ごとにある 9 通りのデータを繰り返しデータと見なして一元配置モデルを適用すると，箱ひげ図は図 15 のようになり，分散分析を実施すると，p 値は 0.002 となり，1% でも有意となる．

触媒の種類の違いによる収率の違いを調べるために，同様に各水準ごとに 12 通りの繰り返しデータがあると考え，一元配置モデルを適用する．箱ひげ図は図 16 のようになり，分散分析を実施すると，p 値は 10^{-4} 未満となり，1% でも有意となる．

交互作用を検出するには，aov 関数のモデル式として「収率~温度 * 触媒」と入力すればよい．

プログラム例

```
data
< - read.table("clipboard", header=T)
温度< - rep(data[,1],3)
触媒< - c(rep("B1",12),rep("B2",12),rep("B3",12))
収率< - c(data[,3],data[,4],data[,5])
summary(aov(収率 ~温度))
```

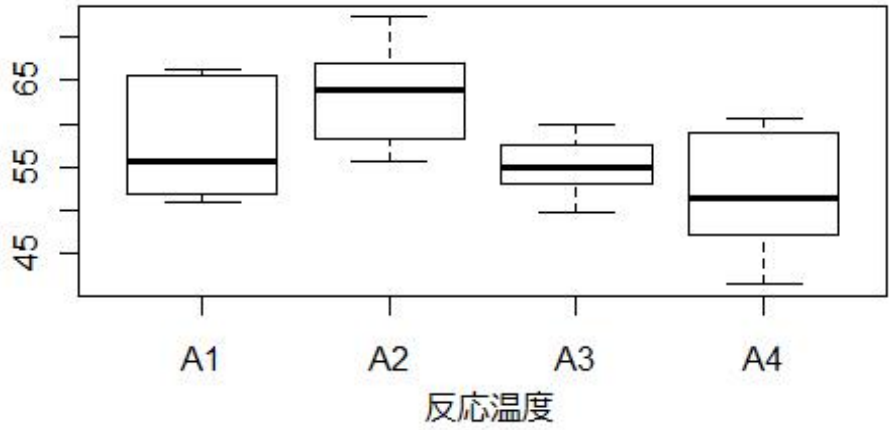


図 15　(問題 10.5) 反応温度ごとの収率のばらつき

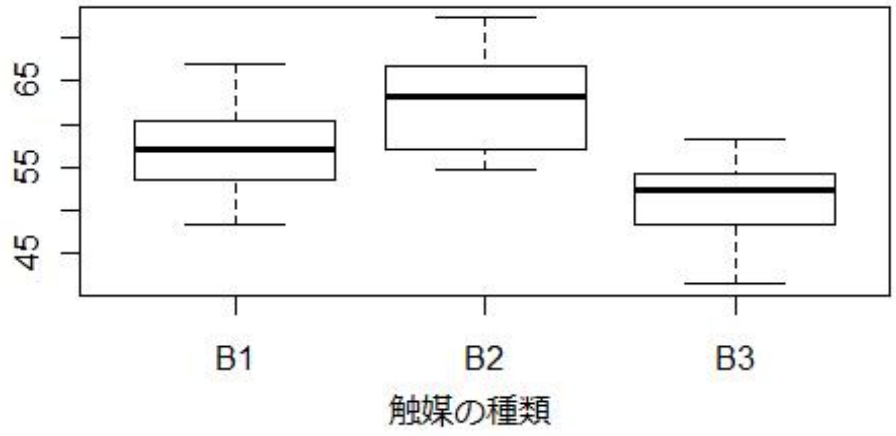


図 16　(問題 10.5) 触媒の種類の違いによる収率のばらつき

```
summary(aov(収率 ~触媒))
summary(aov(収率 ~温度 + 触媒))
summary(aov(収率 ~温度 * 触媒))
```