

1. 確率分布（二項分布とポアソン分布）

今回は、2 項分布とポアソン分布を紹介する。ともに、頻度（人数、回数など）の分布のための理論分布である。

1. 2 項分布

2 種類の結果の可能性がある実験を、同じような状況で独立に複数回繰り返すことを考える。独立に繰り返すということは、すでに起こった結果に次の実験が影響されないことを意味する。このような実験を、結果が 2 種類であることから 2 項実験とかベルヌイ実験と呼ぶ。このような実験として、想像しやすいものの 1 つがコイン投げの実験である。その他にも、サイコロを振って 1 の目が出るかでないかの実験をした場合や、宝くじを買って 1 等が当たるか当たらないかを繰り返した場合も、2 項実験の事例であると考えることができる。この繰り返し実験の結果として 2 つの結果のうち一方が起こった回数の分布のことを「2 項分布」と呼ぶ。

この分布を考える際に必要な事項はなんであろうか。分布を考える上で必要となる実験の特徴は何かということである。この場合、次の 2 つの事項を考えればよい。1 つは 1 回の実験で一方の結果が発生する確率であり、もう一つは実験の回数である。ここで、その指定された事象の発生する確率を p とする。また、実験の回数は n で表すこととする。

ここで、2 種類の結果を分かりやすくするため、一方を「成功」、もう一方を「失敗」と呼ぼう。もちろん、どの結果を「成功」とするかは自由に決めることができる。コイン投げをして、表の出る回数に注目している場合は、「表が出る」ことを「成功」とすればよいし、宝くじを買って 1 等が当たらない回数に注目している場合には、「1 等が当たらない」ことを「成功」として、分布を考える。一般化するために、成功の回数を表す変数を X で表したときに、その変数 X がある具体的な値 k をとる確率は次の式で与えられる；(X のことを 2 項分布に従う確率変数という)

$$P(X = k) = {}_n C_k p^k (1 - p)^{n-k}$$

ここで、 ${}_n C_k$ は n 個の中から k 個を取り出す組み合わせの数を表している。なお、上記のような 2 項分布に従う確率変数の平均と分散は、それぞれ np と $np(1-p)$ になる。

1.1. 2 項分布と社会調査

標本調査により意見を聞く際の賛成の数のモデルとして、2 項分布が使用される。母集団から単純無作為抽出（全ての人が同じ確率で抽出されることと、選ばれる人の組み合わせについても等確率であることが保証される抽出法）で標本抽出が行われ、ある項目への賛否を問うとする。ここで非常に大きな母集団、または、復元抽出であるとし、母集団でその項目を支持する人の割合を p とする。大きさ n の標本を抽出し、支持する人の人数を求め、その分布を考えると、それが 2 項分布になる。

なお、実際のこの母集団比率についての統計的推測をする際には、2 項分布の分布的特性を直接使うことは少ない。それは、一般の標本調査の場合には、標本の大きさ(n)はかなり大きく、2 項分布を正規分布で近似できるからである。すなわち、標本が大きい場合、支持する人の人数の分布は、平均が np 、分散が $np(1-p)$ の正規分布になる。

2. ポアソン分布

そう頻繁には起こらない現象を、一定の期間観測して発生した数の分布のモデルとして、ポアソン分布がある。ポアソン分布は、平均の値だけを決めると分布が決まるという非常に単純な分布モデルである。いま平均を μ とし、その逆数を $\theta(=1/\mu)$ とすると、

$$P(X = k) = e^{-\theta} \frac{\theta^k}{k!}$$

で表される。ポワソン分布に従う確率変数の平均は μ であるが、分散も実は μ になる。

ポアソン分布の事例としては、交通事故の件数、製造過程における不良品の個数、サッカーの得点などがあげられる。サッカーの得点の事例を後で紹介する。

2.1. 2 項分布とポアソン分布の関係

頻繁に起こらないということは、2 項分布において、成功の確率 p が 0 に近いことを意味し、それを長期に観測するということは、2 項分布の実験回数 n を無限大に近づけることを意味する。2 項分布の平均は np であるが、この平均が一定であるとして、 p を 0 に、 n を無限大に近づけたときの分布がポアソン分布になる。

ここでいう頻繁に起こらないということは、2 項分布において、成功の確率 p が 0 に近いことを意味し、それを長期に観測するということは、2 項分布の実験回数 n を無限大に近づけることを意味する。2 項分布の平均は np であるが、この平均が一定であるとして、 p を 0 に、 n を無限大に近づけたときの分布がポアソン分布になる。

分布関数の式でその関係を示そう。いま $\mu = np$ とし、この値が一定のもとで n を無限大に

飛ばしたときの 2 項分布の分布関数の極限を考える。いま、 $p = \mu/n$ なので、

$$\begin{aligned}\Pr(X = k) &= {}_n C_k p^k (1-p)^{n-k} = {}_n C_k (\mu/n)^k (1-\mu/n)^{n-k} \\ &= \frac{n!}{(n-k)!k!} (\mu/n)^k (1-\mu/n)^{n-k} \\ &= \frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n} \frac{\mu^k}{k!} \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-k}\end{aligned}$$

とかけ、ここで $n \rightarrow \infty$ の極限を考える。よく知られた結果である $\lim_{n \rightarrow \infty} (1 - \mu/n)^n = e^{-\mu}$ を利用すれば、最初の $k-1$ 個の積と最後の項が 1 に収束することから、

$$\lim_{n \rightarrow \infty} {}_n C_k p^k (1-p)^{n-k} = e^{-\mu} \frac{\mu^k}{k!}$$

が得られる。これがポアソン分布の確率関数である。

次にこのことを利用して、サッカーの得点分布としてポアソン分布が適していることの説明を行う。いま、一定の時間で得点が入るか入らないかの 2 項実験を想定する。なお、得点の入る可能性はその試合を通じて一定であると仮定する。まず、前半と後半 45 分で考えると、 $n=2$ の 2 項実験である。10 分刻みとすると、 $n=9$ である。刻み時間を短くすると、その一定の時間で得点が入る確率 p は 0 に近づく。一方、2 項実験としての試行回数 n はどんどん大きくなる。また刻み時間が小さくなるとその時間に 2 点入る可能性は排除できるので、試合での得点そのものと得点が入った回数の和は等しくなる。以上のことから、ポアソン分布がサッカーの得点分布の近似として適切であろうと考えることができる。

では、実際のデータへの当てはめを行ってみよう。表 1 のデータは、ある年の J リーグ 306 試合におけるホームチームの得点の分布である。本文にもあったように、サッカーはホームであるかアウェーであるかにより得点の可能性が大きく違うスポーツである。そのため、ホームチームだけの結果を使用している。

表 1 J リーグホームチーム得点分布

得点	0	1	2	3	4	5	6	7	計
試合数	69	105	72	36	15	5	1	3	306

このデータにポアソン分布を当てはめてみよう。なお、1 試合当りの平均得点数は 1.533 となり、これを平均とするポアソン分布の確率を求める。先に示した式で計算をしてもよいが、もし Excel が使えるのであれば、Excel に用意されている関数=poisson()を利用すれ

ばより簡単に計算できる。具体的には、=poisson(得点, 1.533,0)として計算する。関数内の最後の引数の0は、確率を計算するためのスイッチで、1を指定すると累積確率が返される関数である。

表2 ポアソン分布の当てはめ結果

得点	試合数	相対度数	理論確率	期待度数
0	69	0.225	0.216	66.1
1	105	0.343	0.331	101.3
2	72	0.235	0.254	77.6
3	36	0.118	0.130	39.7
4	15	0.049	0.050	15.2
5	5	0.016	0.015	4.7
6	1	0.003	0.004	1.2
7	3	0.010	0.001	0.3
8以上	0	0.000	0.000	0.1
計	306			

表2の理論確率という列が、ポアソン分布を想定した場合のその得点の確率で、それに306をかけて算出された数字が期待度数である。念のため、適合度検定¹も行うと、検定統計量の値は、

$$\begin{aligned} \sum_i \frac{(O_i - E_i)^2}{E_i} &= \frac{(69 - 66.1)^2}{66.1} + \frac{(105 - 101.3)^2}{101.3} + \frac{(72 - 77.6)^2}{77.6} \\ &\quad + \frac{(36 - 39.7)^2}{39.7} + \frac{(15 - 15.2)^2}{15.2} + \frac{(9 - 6.2)^2}{6.2} \\ &= 1.04 \end{aligned}$$

となる(5点以上を一つのセルとしている)。いま、セルの数が6で、推定するパラメータは平均の1個であるので、自由度は4になる。ここで有意確率(p値)を計算すると0.90となる。このことから、データへ適合しているといえるであろう。

なお、2項分布でnが大きくなると正規分布で近似できることを説明したが、ポアソン分布の場合も同様で、平均が大きくなると正規分布で近似できる。サッカーの得点分布は、ポアソン分布に近いことを示したが、バスケットボールのように大量の得点が入る場合の得点の分布は、正規分布に近くなる。データはNBAなどのサイトで公開されているので各

¹ 適合度の検定については、「7. カイ二乗検定」という教材で説明する。

自で得点分布（ヒストグラムの作成などで）を確かめてみることを勧める。

練習問題

問 1（二項分布の性質）

- (1) 一の目から六の目が同じサイコロを 6 回振ったとき、一の目は何回出そうだろうか。
その期待値（平均）を次の 1～6 のうちから最も近いものを一つ選べ。

1. 1 回 2. 2 回 3. 3 回 4. 5 回 5. 5 回 6. 6 回

- (2) 一の目から六の目が同じサイコロを 6 回振ったとき、出る目の分散はどのくらいだろうか。次の 1～6 のうちから最も近いものを一つ選べ。

1. $\frac{1}{6}$ 2. $\frac{2}{6}$ 3. $\frac{3}{6}$ 4. $\frac{4}{6}$ 5. $\frac{5}{6}$ 5. $\frac{6}{6}$

- (3) 一の目から六の目が同じサイコロを 6 回振ったとき、一の目が 4 回出る確率はどれくらいか。次の 1～7 のうちから最も近いものを一つ選べ。

1. $\left(\frac{1}{6}\right)^6$ 2. $6 \times \left(\frac{1}{6}\right) \times \left(\frac{5}{6}\right)^5$ 3. $15 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^4$
4. $20 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^3$ 5. $15 \times \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right)^2$ 6. $6 \times \left(\frac{1}{6}\right)^5 \times \left(\frac{5}{6}\right)$
7. $\left(\frac{5}{6}\right)^6$

問 2（ポアソン分布の性質）

ある交差点では、1 年間に平均 4 件の交通事故が発生している。この交差点での交通事故件数がポアソン分布に従うとすると、1 年間の交通事故件数の標準偏差はいくつか。次の 1～5 のうちから 1 つ選べ。

1. 1 件 2. 2 件 3. 3 件 4. 4 件 5. 5 件 6. 6 件

練習問題の解答

問 1 (二項分布の性質)

(1) 解答 : 1

一の目が出ることを「成功」、それ以外の目が出ることを「失敗」と考えると、同じサイコロを 6 回振ることは実験の回数 $n=6$ ・成功する確率 $p=\frac{1}{6}$ の二項分布を考えることと同じになる。二項分布の平均は $n \times p$ で求めることができるので、「一の目が出る」という事象は平均して、 $6 \times \frac{1}{6} = 1$ 回ほど生じることが分かる。

(2) 解答 : 5

二項分布の分散は $n \times p \times (1-p)$ で求めることができるので、出る目の分散は、 $6 \times \frac{1}{6} \times \left(1 - \frac{1}{6}\right) = 6 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{6}$ となる。

(3) 解答 : 5

「一の目が 4 回出る」ということは、成功回数が $X=4$ となることだから、その確率は二項分布の公式より、 $P(X=4) = {}_6C_4 p^4 (1-p)^{6-4} = 15 \times \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right)^2$ となる。

問 2 (ポアソン分布の性質)

解答 : 2

ポワソン分布は平均と分散が同じ値になる。ここで平均は 4 件なので、1 年間の交通事故件数の分散も 4 件となる。標準偏差は分散の正の平方根を計算したものであるため、1 年間の交通事故件数の標準偏差は $\sqrt{4} = 2$ 件となる。

2. 正規分布

1. 正規分布とは？

数値データの理論分布モデルとしてよく使用されるのが正規分布である。正規分布は対称な単峰な分布である。正規分布は平均と分散(または、標準偏差)が決まると完全に形が決まり、平均 μ 、分散 σ^2 の正規分布の曲線を表す関数(確率密度関数)は

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

である。ここで、 π は円周率(=3.141...)、 e は自然対数の底(=2.718...)で、ともに数値である。このような正規分布を一般に $N(\mu, \sigma^2)$ 表す。また、平均0、分散1の正規分布を標準正規分布と呼ぶ。

次の図1は、平均や分散が異なる3種類の正規分布を示している。それぞれの正規分布は、平均の値が中央となる対称な分布である。この図で、平均と分散の違いによる分布の違いを確認してほしい。平均により、分布の中心が定まり、分散が中心からの広がり的大小を定める。

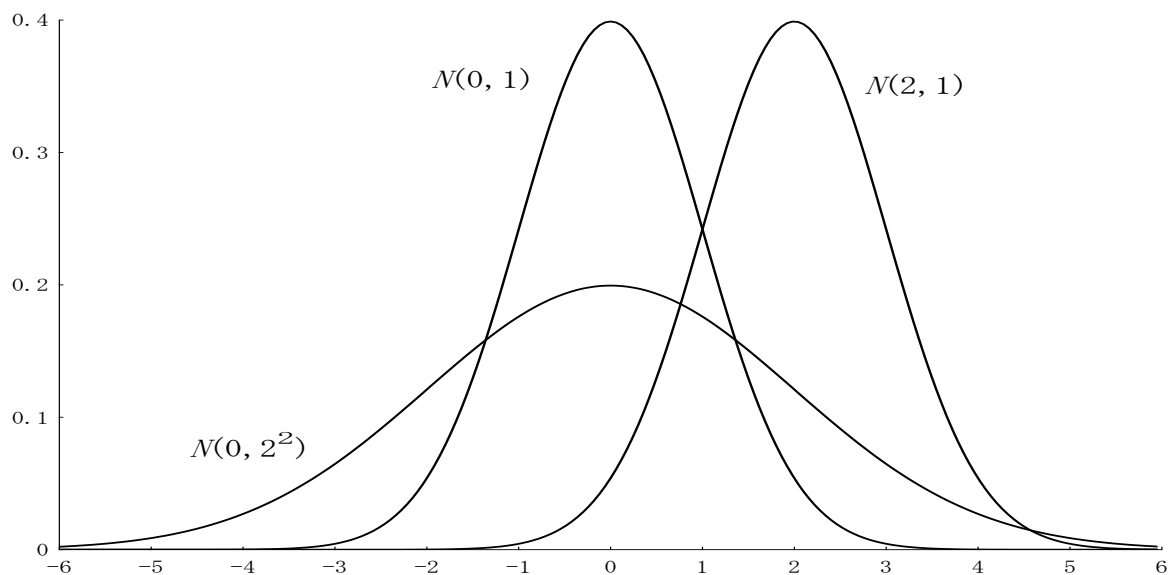


図1 正規分布の形状

2. 正規分布の性質：シグマの法則

正規分布のよく知られた性質としてシグマの法則がある。平均(μ)や分散(σ^2)がどのような値であろうと、次のような区間の割合(確率)に関する式が成り立つ。

$$\Pr[\mu - \sigma < X < \mu + \sigma] = 0.683$$

$$\Pr[\mu - 2\sigma < X < \mu + 2\sigma] = 0.954$$

$$\Pr[\mu - 3\sigma < X < \mu + 3\sigma] = 0.997$$

ただし、 $X \sim N(\mu, \sigma^2)$ である。すなわち、平均の周りプラスマイナス1標準偏差に入る割合が68.3%、平均の周り2標準偏差に入る割合が95.4%、3標準偏差の場合99.7%になる。

任意の区間の確率を得るためには別節で説明する正規分布の数表を用いるか、表計算ソフトに組み込まれている関数(Excelの場合、**NORMDIST**や**NORMSDIST**)を利用すればよい。ただし、数表で用意されている確率は標準正規分布に関するものです。そのため基準化を行い、その値を基に数表を引かなければならない。Excelで計算する場合で**NORMDIST**を使う場合には基準化する必要はない。詳細はExcelのヘルプを参照のこと。

3. 正規分布に従う確率変数の和の分布

正規分布に従う確率変数の和も、正規分布に従うことが知られている。同じ分布に従う独立な確率変数の和の分布が元の分布になることを再生性と呼ぶが、正規分布は再生性を持つ分布である。また、正規分布は、独立でない確率変数の和も正規分布になるという特徴を持っている。以下に 2 つの変数の和の分布を示す。

X と Y が独立な場合の結果

$$X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2) \text{ で、 } Z = X + Y \text{ のとき、} \\ Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

X と Y に相関(ρ)がある場合

$$X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2) \text{ で、 } Z = X + Y \text{ のとき、} \\ Z \sim N(\mu_x + \mu_y, \sigma_x^2 + 2\rho\sigma_x\sigma_y + \sigma_y^2)$$

なお、差の分布についても触れておきたい。独立な変数の差の分布は、和の場合と同じように正規分布になる。差の平均は平均の差になるが、分散は 2 つの変数の分散の和になるので注意が必要である。

X と Y が独立な場合の差の分布

$$X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2) \text{ で、 } Z = X - Y \text{ のとき、} \\ Z \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

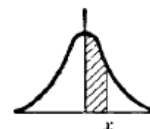
X と Y に相関(ρ)がある場合の差の分布

$$X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2) \text{ で、 } Z = X - Y \text{ のとき、} \\ Z \sim N(\mu_x - \mu_y, \sigma_x^2 - 2\rho\sigma_x\sigma_y + \sigma_y^2)$$

4. 正規分布の数値表の使い方

正規分布に関する確率を求める場合、統計ソフトウェアや Excel を使えば簡単に計算できるが、統計数値表を使うこともあり、ここではその使用方法を説明する。

正規分布表 $x \rightarrow \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3718	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4083	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319

上記は、正規分布表の一部である。この表から正規分布に従う確率変数 X について $\Pr(0 < X < x)$ となる確率を知ることができる。この表の表側で、 x の小数点 1 ケタまでの数値を選び、表頭で小数点 2 ケタ目の数値を選ぶ。その交差しているところの数値が $\Pr(0 < X < x)$ になる。

たとえば、 $\Pr(0 < X < 0.57)$ を求めてみよう。表側で 0.5 の行を選び、次に表頭の 0.07 の列を選び、その交差する位置にある 0.2157 が求める確率となる。

では、数値表を使って具体的な確率を求めてみよう。いま、日本人の成人男子の身長は平均は 171.5cm で、標準偏差は 5.7cm である。いま身長が正規分布するとして、175.0cm を超える割合を求めてみよう。まず、175.0cm の値を基準化した値を求める。

$$\frac{175.0 - 171.5}{5.7} = 0.614...$$

これから、求める確率は標準正規分布で 0.61 を超す確率を求めることになる。いま数値表から 0.61 に対応する数値を確認すると 0.2291 である。この値は、0 から 0.61 の間の確率になる。求めるべき確率は 0.61 を超える確率である。よって、0 を超す確率の 0.5 から 0.2291 を引いた 0.2709 が求める確率となり、日本人の成人男子で身長が 175.0cm を超す割合は約 27%であるといえる。

練習問題

問 1 (正規分布の性質)

正規分布に関する次の記述 **a** から **c** がある。

- a.** 平均と標準偏差がいくつであっても、平均のところで密度関数が最大値を取る。
- b.** 標準偏差が大きいほど、密度関数の最大値は小さくなる。
- c.** 平均 0 の正規分布 $N(0, \sigma_1^2)$ と $N(0, \sigma_2^2)$ があり、 $\sigma_1 < \sigma_2$ とする。このとき、0 以下の任意の $x(x \geq 0)$ に対し、 $N(0, \sigma_1^2)$ における x 以上の確率 P と、 $N(0, \sigma_2^2)$ における x 以上の確率 Q について、常に $P \leq Q$ が成り立つ。

この記述について、次の選択肢 1～5 のうち、最も適切なものを一つ選べ。

- 1. **a** のみ正しい 2. **b** のみ正しい 3. **a** と **b** のみ正しい
- 4. **a** と **c** のみ正しい 5. **a** と **b** と **c** すべてが正しい

問 2 (正規分布に従う確率変数の和の分布)

ある展望台へのエレベータの最大許容重量は 600kg である。成人男性の体重がおおよそ平均 70kg、標準偏差 8kg の正規分布に従うとして、成人男性 8 人グループの合計体重が 600kg を超えて全員が同じエレベータに乗れない確率 (% 表示) はおおよそいくらか。次の 1～5 から最も近いものを一つ選べ。ただし、各人の体重は独立とする。

- 1. 4% 2. 8% 3. 16% 4. 92% 5. 98%

問 3 (二項分布の正規分布への近似)

一目の目から六目の目が出る確率が同じサイコロを 720 回振ったとき、一目の目が 135 回以上出る確率はどれくらいか。次の 1～5 のうちから最も近いものを一つ選べ。ただし、確率はパーセント表示を用いる。

- 1. 約 20% 2. 約 16% 3. 約 7% 4. 約 5% 5. 約 1%

練習問題の解答

問 1 (正規分布の性質)

解答：5

正規分布は平均と分散のみで密度関数の形状が決まり、平均で密度関数は最も高くなるので、**a** は正しい。正規分布の密度関数の最大値は標準偏差が大きくなるほど小さくなるので、**b** は正しい。**c** が正しいかどうかは図 1 (p.1) を見て考えると分かり易い。

問 2 (正規分布に従う確率変数の和の分布)

解答：1

成人男性の体重は平均 70kg、分散 $8^2=64\text{kg}$ の正規分布に従うとしたので、成人 8 人の合計体重は、平均 $70\times 8=560\text{kg}$ 、分散： $64\times 8=512\text{kg}$ の正規分布に従う¹。合計体重が 600kg を超える確率を計算するためには、(i) 600kg の標準化得点 (z) を求め、(ii) 正規分布表から z に対応する確率を求める必要がある。600kg の標準化得点 (z) を求めると、以下ようになる。

$$z = \frac{600 - \text{合計体重の平均値}}{\text{合計体重の標準偏差}} = \frac{600 - 560}{\sqrt{512}} = 1.767$$

正規分布表から、1.767 に対応する確率を求めると 0.0384 となる。よって、1 が正しい。

問 3 (二項分布の正規分布への近似)

解答：3

サイコロの目の出現回数の確率は二項分布に従うが、試行数（サイコロを振った数）が 720 と多いので、正規分布で近似することができる。このとき、出目の出現回数の確率は、

$$\text{平均：試行数} \times \text{一の目が出る確率} = 720 \times \frac{1}{6} = 120$$

$$\text{分散：試行数} \times \text{一の目が出る確率} \times \text{一の目が出ない確率} = 720 \times \left(\frac{1}{6}\right) \times \left(1 - \frac{1}{6}\right) = 100$$

の正規分布に従う。

¹ 各人の体重は独立なので、合計体重の平均と分散は一人ずつの体重を順次足していくことで求めることができる。これは 8 人分の掛け算をすることと等しい。

一の目が135回以上出る確率は、(i) 135回の標準化得点(z)を求め、(ii) 正規分布表から z に対応する確率を求める必要がある。135回の標準化得点(z)を求めると、以下のようになる。

$$z = \frac{135 - 120}{\sqrt{100}} = 1.5$$

正規分布表から、1.5に対応する確率を求めると0.0668となる。よって、3が正しい。

3. 統計的推定

本章の目的

- ・ 統計的推定の考え方を理解する
- ・ 比率の区間推定の考え方を理解する
- ・ 平均値の区間推定の考え方を理解する

Key Words : 点推定、区間推定、95%信頼区間

1. 統計的推定

● 統計的推定とは？

- 統計的推定とは標本調査の結果をもとに、母集団の特性を推定することである。

例：視聴率調査

視聴率調査は関東在住の 600 世帯を無作為に抽出し、調査対象者のテレビ視聴行動を調べている。600 世帯のうち何世帯が A という番組を見ているかを調べることで、A という番組の視聴率を明らかにするのである。この調査は無作為抽出にもとづく標本調査なので、関東圏の視聴率を知るためには調査結果をもとに統計的な推定を行う。

● 2 種類の統計的推定

➤ 点推定

- ☆ 標本調査の結果をもとに 1 つの数値を計算し、その値を母集団の特性の推定値にすること

例：視聴率調査

標本調査の結果、600 世帯のうち 30 世帯が A という番組を見ていた。この場合、関東圏における番組 A の視聴率は 5% ($=30/600$) として推定される

➤ 区間推定

- ☆ 標本調査の結果をもとに、あるていど幅をもたせた数値を母集団の特性の推定値にすること
- ☆ ここでは、標本誤差を明示的にとりいれたかたちで母集団の特性の推定値をもとめる。具体的には、区間推定は以下の数式で求める

標本から計算した値±信頼度×標準誤差

＊信頼度 : 推定の精度を決める値

＊標準誤差 : 標本誤差をもとに計算する

例：視聴率調査

標本調査の結果、600 世帯のうち 120 世帯が B という番組を見ていた。この場合の関東圏における番組 B の視聴率の区間推定値は 17%～23%となる。これは、関東圏における番組 B の視聴率は 17%～23%のどこかにあることを意味している。

2. 比率の区間推定

● 比率の区間推定

- 標本調査の結果から母集団における比率（母比率）を推定する場合に用いられる。
- 比率の区間推定値は以下の形で与えられる

$$\text{標本比率} \pm z_{\text{信頼係数}} \times \sqrt{\frac{\text{標本比率} \times (1 - \text{標本比率})}{\text{標本の大きさ}}}$$

● 95%信頼区間

- 信頼係数は区間推定の精度を決めるものである。信頼係数を決めることで、求めた区間推定値のなかに母集団の特性が含まれる確率が決まる。信頼係数は確率で与えられ、 α と表現される。このとき、求めた区間推定値のなかに母集団の特性が含まれる確率は $1 - \alpha$ になる。
- 多くの場合、信頼係数 α は 0.05 となる。このとき、区間推定値のなかに母集団の特性が含まれる確率が 95% となる。このような形で決められた区間推定値を 95% 信頼区間という。
- 比率の区間推定の場合、 $z_{\text{信頼係数}}$ の値は 1.96 となる。
 - ◇ 標本規模が十分に大きい場合、標本比率の標本分布は標準正規分布¹になる。標準正規分布の場合、-1.96 から 1.96 の区間にデータの 95% の数値が含まれる。したがって、この区間を考えれば、母比率は 95% の確率で含まれることになる。

例題：番組 B の 95%信頼区間を求める

関東圏の視聴率を調べるために、600 世帯を無作為に抽出した。600 世帯を対象に、X 月 Y 日に B という番組を視聴している人数を調査したところ、120 名であった。この調査結果をもとに、関東圏における番組 B の視聴率の 95% 信頼区間を求める

- 標本における番組 B の視聴率は 20% (=120/600)
- 標本比率をもとに、母比率の 95% 信頼区間を求めると、以下のようになる。

$$0.2 \pm 1.96 \times \sqrt{\frac{0.2 \times (1 - 0.2)}{600}} = 0.2 \pm 1.96 \times 0.016 = 0.2 \pm 0.032$$

☆%表示に戻すと、番組 B の視聴率の 95% 信頼区間は 20% ± 3.2% となる。

☆したがって、関東圏における番組 B の視聴率は 95% の確率で、16.8% ~ 23.2% の間にあるといえる。

¹ 標準正規分布とは、平均 0、分散（標準偏差）1 となる正規分布である。

2. 平均値の区間推定値

- 標本調査の結果から母集団の平均値（母平均）を推定する場合に用いられる。
- ここでは、母集団の分散（母分散）がわからない場合における、母平均の 95%信頼区間の求めかた²を紹介する。

$$\text{標本平均} \pm t(n-1, 0.05) \times \frac{\text{標本の標準偏差}}{\sqrt{\text{標本の大きさ}}}$$

- 信頼係数は、自由度＝標本の大きさ－1 と信頼度（95%）をもとに、t 分布表³から求めることができる。

例題：R 大学生の身長の平均値を推定する

R 大学の学生から無作為に 16 名を抽出し、その身長を調べたところ、以下のような結果が得られた。

平均値	164.5cm
標準偏差	8.5

この標本調査の結果から、R 大学生における身長の母平均の 95%信頼区間を求める。

- 母平均の 95%信頼区間を求める式に、数値を代入すると以下ようになる。

$$164.5 \pm 2.131 \times \frac{8.5}{\sqrt{16}} = 164.5 \pm 2.131 \times 2.125 = 164.5 \pm 4.528$$

* 自由度は **16－1＝15** なので、t 分布表から信頼度を求めると、**2.131** となる

☆R 大学生の身長平均値は 95%の確率で、約 160cm～169cm の間にある、といえる。

² 母分散の値がわかっていることは滅多にないので、ここでは母分散が未知の場合を紹介する。ちなみに、母分散の値がわかっている場合には、母平均の 95%信頼区間は以下の式で求めることができる。

$$\text{標本平均} \pm 1.96 \times \frac{\text{母集団の標準偏差}}{\sqrt{\text{標本の大きさ}}}$$

³ t 分布表のみかたについては付録を参照のこと。

付録：t分布表の見方

- 自由度をもとに、信頼度の値を求める
- 95%信頼区間を求める場合には、「.025」($=0.05 \div 2$) という列と計算した自由度に対応する行の交点に信頼度の値がある。

例：95%信頼区間・自由度が15の場合

→ 「.025」($=0.05 \div 2$) という行と「15」という列の交点の数値 **2.131** をみる

α 2 α	.250 (.500)	.200 (.400)	.150 (.300)	.100 (.200)	.050 (.100)	.025 (.050)	.010 (.020)	.005 (.010)	.0005 (.0010)
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.794	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.779	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.764	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.750	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.736	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
31	.682	.853	1.054	1.309	1.696	2.040	2.453	2.744	3.633
32	.682	.853	1.054	1.309	1.694	2.037	2.449	2.738	3.622
33	.682	.853	1.053	1.308	1.692	2.035	2.445	2.733	3.611
34	.682	.852	1.052	1.307	1.691	2.032	2.441	2.728	3.601
35	.682	.852	1.052	1.306	1.690	2.030	2.438	2.724	3.591
36	.681	.852	1.052	1.306	1.688	2.028	2.434	2.719	3.582
37	.681	.851	1.051	1.305	1.687	2.026	2.431	2.715	3.574
38	.681	.851	1.051	1.304	1.686	2.024	2.429	2.712	3.566
39	.681	.851	1.050	1.304	1.685	2.023	2.426	2.708	3.558
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
41	.681	.850	1.050	1.303	1.683	2.020	2.421	2.701	3.544
42	.680	.850	1.049	1.302	1.682	2.018	2.418	2.698	3.538
43	.680	.850	1.049	1.302	1.681	2.017	2.416	2.695	3.532
44	.680	.850	1.049	1.301	1.680	2.015	2.414	2.692	3.526
45	.680	.850	1.049	1.301	1.679	2.014	2.412	2.690	3.520
46	.680	.850	1.048	1.300	1.679	2.013	2.410	2.687	3.515
47	.680	.849	1.048	1.300	1.678	2.012	2.408	2.685	3.510
48	.680	.849	1.048	1.299	1.677	2.011	2.407	2.682	3.505
49	.680	.849	1.048	1.299	1.677	2.010	2.405	2.680	3.500
50	.679	.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
240	.676	.843	1.039	1.285	1.651	1.970	2.342	2.596	3.332
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

練習問題

問 1 (信頼区間の性質)

ある県における政党 A の支持率について調査を行い、支持率の信頼区間を求めたところ、信頼区間の幅がやや広がった。次回の調査のときには信頼区間の幅を約半分にしたい。次の 1~5 のうちから適切な方法を 1 つ選べ。

1. 標本の大きさ n を固定して、信頼係数を 2 倍にする。
2. 標本の大きさ n を固定して、信頼係数を $1/2$ 倍にする。
3. 信頼係数を固定して、信頼の大きさ n を 2 倍にする。
4. 信頼係数を固定して、信頼の大きさ n を 4 倍にする。
5. 信頼係数を固定して、信頼の大きさ n を $1/2$ 倍にする。

問 2 (信頼区間の解釈)

ある母集団の母平均 μ についての 95%信頼区間は $(48, 65)^4$ であった。次の記述 a から c がある。

- a. 標本の 95% が含まれる区間が、48 から 65 である。
- b. 標本平均が 48 から 65 の間にある確率は 0.95 である。
- c. 同じ大きさを持つランダムな標本をとり、それぞれについて 95%信頼区間を求める手続きを多数行くと、 μ は、この手順でできる区間のうちの約 95% の区間に含まれる。

この記述について、次の 1~5 のうちから最も適切なものを一つ選べ。

1. a のみ正しい
2. b のみ正しい
3. c のみ正しい
4. a と b のみ正しい
5. b と c のみ正しい

⁴ $(48, 65)$ は $48 < \mu < 65$ を意味する。

練習問題の解答

問 1 (信頼区間の性質)

解答：4

比率の信頼区間を計算する場合、区間の幅を決めるのは信頼係数そのものではなく、 z 信頼区間なので、1 と 2 は誤り。信頼区間を計算する式を見ると、標準誤差は

$\sqrt{\frac{1}{\text{標本の大きさ}}}$ で小さくなることがわかる。したがって、標本の大きさを 4 倍にする

と信頼区間の幅は $\sqrt{\frac{1}{4}} = \frac{1}{2}$ となる。よって、4 が正しい。

問 2 (信頼区間の解釈)

解答：3

母平均の 95% 信頼区間は、その区間内に母集団の平均値が含まれる確率が 95% となるものなので、a と b は誤り。c は 95% 信頼区間の定義を正確に述べたものなので正しい。よって、3 が正しい

4. 平均値の差の検定

本章の目的

- ・ 2つの集団の間の差を検討する方法として、t検定を理解する。
- ・ 3つ以上の集団の間の差を検討する方法として、(一元配置)分散分析を理解する

Key Words : t検定、(一元配置)分散分析、t分布、F分布

1. t検定

● t検定の目的

- t検定は、2つの集団の間で平均値に違いがあるかどうかを検討する際に用いられる方法である。
- 特に、母集団の分散が未知であるが、集団間で分散の大きさが大きく異なる場合に用いられる。

● t検定の手順

- 1. 統計的仮説の設定

帰無仮説 : 2つの集団の母集団は等しい ($\mu_A = \mu_B$)

対立仮説

両側仮説の場合 : 2つの集団の母集団は等しい ($\mu_A \neq \mu_B$)

片側仮説の場合 : 集団Aの母平均は集団Bよりも大きい(小さい) ($\mu_A > (<) \mu_B$)

- 2. 標本統計量の計算

データから以下にあるt統計量を計算する

$$t = \frac{\text{集団Aの平均値} - \text{集団Bの平均値}}{\text{合併した標準偏差} \times \sqrt{\frac{1}{\text{集団Aのケース数}} + \frac{1}{\text{集団Bのケース数}}}}$$

$$\text{合併標準偏差} = \sqrt{\frac{(\text{集団Aのケース数} - 1) \times \text{集団Aの不偏分散} + (\text{集団Bのケース数} - 1) \times \text{集団Bの不偏分散}}{\text{集団Aのケース数} + \text{集団Bのケース数} - 2}}$$

- 3. 有意水準の設定

☆ 有意水準は一般的には5% (0.05) に設定される

☆ 自由度を計算し、t分布表¹をもとに臨界値を求める

- ・ 自由度は次の式にもとづき計算する

$$\text{自由度} = \text{集団Aのケース数} + \text{集団Bのケース数} - 2$$

- ・ 両側仮説と片側仮説では、臨界値の求め方が異なる

¹ t分布表の見方については付録を参照。

- 4. 帰無仮説の正否の判断
 - t 値と臨界値を比較して、帰無仮説の正否の判断する
 - ☆臨界値 $\geq$ t 値 $\Rightarrow$ 帰無仮説を棄却しない
 - ☆臨界値 $<$ t 値 $\Rightarrow$ 帰無仮説を棄却し、対立仮説を採択する

例：身長の男女差を検討する

男性と女性を無作為に抽出し、その身長を測定したところ、以下のような結果になった。
母集団でも身長に男女差があるかどうかを t 検定で検討する。

	男性	女性
平均値	172.25	156.36
不偏標準偏差	27.64	10.65
ケース数	8	11

- 1. 統計的仮説の設定
 - 帰無仮説：男性と女性の身長の母平均は等しい
 - 対立仮説：男性の身長の母平均は女性よりも大きい（片側仮説²⁾）
- 2. 標本統計量の計算

$$t = \frac{172.25 - 156.36}{4.20 \sqrt{\frac{1}{9} + \frac{1}{11}}} = 8.14$$

$$\text{なお、合併標準偏差} = \sqrt{\frac{(8-1) \times 27.64 + (11-1) \times 10.65}{8+11-2}} = 4.20$$

- 3. 有意水準の設定
 - ☆有意水準を 5% に設定
 - ☆自由度は $8+11-2=17$ なので、t 分布表から臨界値を求めると、**1.740³⁾**になる
- 4. 帰無仮説の正否の判断
 - t 値 (**8.14**) は臨界値 (**1.740**) よりも大きいので、帰無仮説を棄却し、対立仮説を採択する。つまり、男性の身長の母平均は女性よりも大きい、といえる。
 - なお、この場合には、「男性の身長は女性のものよりも**有意に**大きい」と表現することもある。

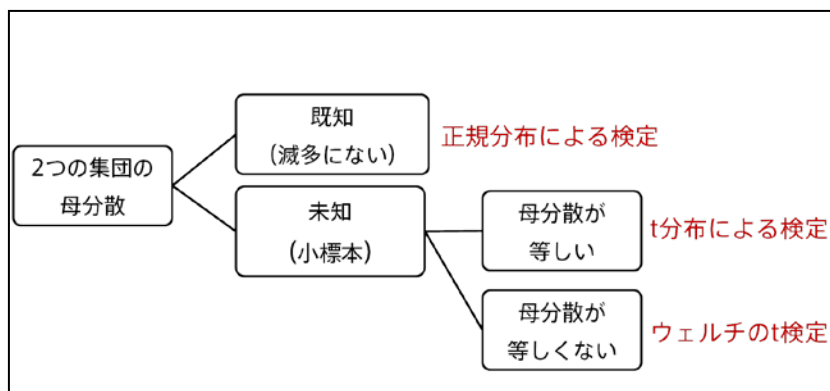
²⁾ 世の中では女性よりも男性のほうが平均身長は高いと信じられているので、ここでは片側仮説をもちいる。なお、男性と女性のうち、どちらの平均身長が高いかわからない場合には両側仮説「男性と女性の身長の母平均は等しくない」を対立仮説にする。

³⁾ ここで求めた臨界値は片側仮説に対応したものである。

- 2つの平均値の差の検定について

- 厳密にいうと、次の2つの基準をもとに、もちいるべき検定が変化する

- ✧ 母集団の分散が既知か未知か？
- ✧ 集団間の分散が等しいか等しくないか？



2. (一元配置) 分散分析

- (一元配置) 分散分析の目的

- 分散分析は、2つ以上の集団の間で平均値に違いがあるかどうかを検討する際に用いられる方法である。

- ✧ ちなみに、3つ以上の集団間の平均値の差をt検定で検討する場合、比較する集団数が増えるにしたがって、分析結果に誤り（第1種の過誤⁴）を含む確率が上昇する

- 分散分析は集団間でそもそも差があるのかどうかを検討する方法であり、集団間のどこに差があるのかどうかを検討するには多重比較⁵（事後検定）の方法を用いる必要がある。なお、多重比較の方法は本章の水準を超えるので割愛する。

- 分散分析の手順

- 1. 統計的仮説の設定

帰無仮説：すべての集団でその母平均は等しい

対立仮説：少なくとも1つの母平均が他の集団の母平均と異なる

⁴ 第1種の過誤とは、帰無仮説が正しいときに対立仮説を採択する誤りのことである。有意水準は分析において第1種の過誤を犯す確率と等しい。ちなみに、対立仮説が正しいときに帰無仮説を採択する誤りのことを第2種の過誤という。

⁵ 比較する集団が3つの場合でも、集団間の差のあり方は多様でありうる（例えば、「AとBで差がある」・「BとCで差がある」etc..）。分散分析で明らかにできるのは集団間にそもそも差があるのかどうかであり、差のあり方を明らかにするには多重比較を行う必要がある。なお、多重比較の方法についても、さまざまな方法（テューキーのHSD法、ボンフェローニの方法、シェッフェの方法etc..）が存在する

➤ 2. 分散分析表の作成

データから以下の分散分析表⁶を作成する。

分散分析表を作成することで、F 値を求める。

変動因	平方和 (SS)	自由度 (df) ⁷	平均平方 (MS)	F値
群間 (Between)	$SS_{\text{Between}} = \sum_{j=1}^J n_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	$df_{\text{Between}} = \text{集団数} - 1$	$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{df_{\text{Between}}}$	$\frac{MS_{\text{Between}}}{MS_{\text{Within}}}$
群内 (Within)	$SS_{\text{Within}} = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2$	$df_{\text{Within}} = \text{データ数} - \text{集団数}$	$MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}}$	
全体 (Total)	$SS_{\text{Total}} = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2$	$df_{\text{Total}} = \text{データ数} - 1$		

➤ 3. 有意水準の設定

✧ 有意水準は一般的には 5% (0.05) に設定される

✧ 自由度を計算し、F 分布表⁷をもとに臨界値を求める

・ 臨界値を求めるために、次の 2 つの自由度を計算する

$$\text{因子の自由度} = \text{集団数} - 1$$

$$\text{残差の自由度} = \text{ケース総数} - \text{集団数}$$

➤ 4. 帰無仮説の正否の判断

F 値と臨界値を比較して、帰無仮説の正否の判断する

☆ 臨界値 $\geq$ F 値 $\Rightarrow$ 帰無仮説を棄却しない

☆ 臨界値 $<$ t 値 $\Rightarrow$ 帰無仮説を棄却し、対立仮説を採択する

例：教授法についての実験

教授法が受講生の統計学の得点にどのような影響を与えるのかに興味を持ったある研究者は次のような 3 つの条件のもと実験を行った。

1. 統計学の知識をただひたすら詰め込む
2. 統計学の問題を解かせながら、統計学を理解させる
3. 統計学について何も教えない

その後、統計学の試験を行い、各条件の受講生の得点のデータをところ、次のような結果が得られた。教授法によって、受講生の成績が異なるかどうかを分散分析で検討する。

⁶ 分散分析表の「群間」は「因子」、「群内」は「残差（誤差）」と表現されることもある。

⁷ F 分布表の見方については付録を参照。

	詰め込み条件	問題解決条件	何もしない条件
	70	67	34
	53	83	46
	69	62	35
	73	67	35
	72	76	35
	77	89	41
	57	86	48
	55	68	48
	56	75	54
	64	88	50
平均点	64.6	76.1	42.6
人数	10	10	10
全平均	61.1		

➤ 1. 統計的仮説の設定

帰無仮説：3つの条件の母平均には差がない

対立仮説：3つの条件の母平均のいずれかが異なる

➤ 2. 分散分析表の作成

データをもとに分散分析表を作成すると、次のようになる。

	平方和	自由度	平均平方	F値
群間	5795.0	2	2897.50	37.69
群内	2075.7	27	76.88	
全体	7870.7	29		

➤ 3. 有意水準の設定

✧ 有意水準を5% (0.05) に設定

✧ 2つの自由度を計算すると、以下のようになる

$$\text{因子の自由度} = 3 - 1 = 2$$

$$\text{残差の自由度} = 30 - 3 = 27$$

- 2つの自由度をもとに臨界値を求めると、**3.354**になる

➤ 4. 帰無仮説の正否の判断

F値 (37.69) は臨界値 (3.354) よりも大きいので、帰無仮説を棄却し、対立仮説を採択する。つまり、実験条件によって成績が異なるといえる。

ただし、実験条件間のどこに差があるのは多重比較の方法を用いなければわからない。(テューキーの HSD 法を用いると、「詰め込み条件－問題解決条件」・「詰め込み条件－何もしない条件」・「問題解決条件－何もしない条件」の間に平均値の差があることが分かる。つまり、問題解決条件 > 詰め込み条件 > 何もしない条件の順で点数が上がるのがわかる。)

付録：t 分布表と F 分布表の見方

1. t 分布表の見方

- 自由度と設定した有意水準をもとに、対応する臨界値を見つける
- 片側検定と両側検定でやり方が異なるので注意が必要

例 1：有意水準を 5% に設定・自由度は 17・対立仮説は片側仮説

→ 「.050」という行と「17」という列の交点の数値 **1.740** をみる

例 2：有意水準を 5% に設定・自由度は 17・対立仮説は両側仮説

→ 「.025」(=0.05÷2) という行と「17」という列の交点の数値 **2.110** をみる

α 2 α	.250 (.500)	.200 (.400)	.150 (.300)	.100 (.200)	.050 (.100)	.025 (.050)	.010 (.020)	.005 (.010)	.0005 (.0010)
1	1.000	1.376	1.963	3.078	6.114	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.220	4.303	6.965	9.925	31.599
3	.765	.978	1.250	1.638	2.153	3.782	4.541	5.841	12.924
4	.741	.941	1.190	1.533	2.132	3.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.115	3.771	3.365	4.032	6.869
6	.718	.906	1.134	1.440	2.103	3.767	3.143	3.707	5.959
7	.711	.896	1.119	1.415	2.095	3.765	2.998	3.499	5.408
8	.706	.889	1.108	1.397	2.090	3.764	2.896	3.355	5.041
9	.703	.883	1.100	1.383	2.087	3.762	2.821	3.250	4.781
10	.700	.879	1.093	1.372	2.085	3.762	2.764	3.169	4.587
11	.697	.876	1.088	1.363	2.084	3.761	2.718	3.106	4.437
12	.695	.873	1.083	1.356	2.082	3.760	2.681	3.055	4.318
13	.694	.870	1.079	1.350	2.081	3.760	2.650	3.012	4.221
14	.692	.868	1.076	1.345	2.080	3.760	2.624	2.977	4.140
15	.691	.866	1.074	1.341	2.080	3.760	2.601	2.947	4.073
16	.690	.865	1.071	1.338	2.080	3.760	2.581	2.921	4.015
17	.689	.864	1.069	1.336	1.740	2.110	2.562	2.898	3.965
18	.688	.862	1.067	1.333	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.332	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
31	.682	.853	1.054	1.309	1.696	2.040	2.453	2.744	3.633
32	.682	.853	1.054	1.309	1.694	2.037	2.449	2.738	3.622
33	.682	.853	1.053	1.308	1.692	2.035	2.445	2.733	3.611
34	.682	.852	1.052	1.307	1.691	2.032	2.441	2.728	3.601
35	.682	.852	1.052	1.306	1.690	2.030	2.438	2.724	3.591
36	.681	.852	1.052	1.306	1.688	2.028	2.434	2.719	3.582
37	.681	.851	1.051	1.305	1.687	2.026	2.431	2.715	3.574
38	.681	.851	1.051	1.304	1.686	2.024	2.429	2.712	3.566
39	.681	.851	1.050	1.304	1.685	2.023	2.426	2.708	3.558
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
41	.681	.850	1.050	1.303	1.683	2.020	2.421	2.701	3.544
42	.680	.850	1.049	1.302	1.682	2.018	2.418	2.698	3.538
43	.680	.850	1.049	1.302	1.681	2.017	2.416	2.695	3.532
44	.680	.850	1.049	1.301	1.680	2.015	2.414	2.692	3.526
45	.680	.850	1.049	1.301	1.679	2.014	2.412	2.690	3.520
46	.680	.850	1.048	1.300	1.679	2.013	2.410	2.687	3.515
47	.680	.849	1.048	1.300	1.678	2.012	2.408	2.685	3.510
48	.680	.849	1.048	1.299	1.677	2.011	2.407	2.682	3.505
49	.680	.849	1.048	1.299	1.677	2.010	2.405	2.680	3.500
50	.679	.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
240	.676	.843	1.039	1.285	1.651	1.970	2.342	2.596	3.332
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

2. F分布表のみかた

- 因子の自由度と誤差の自由度をもとに、対応する臨界値を見つける

➤ 因子の自由度 (ν_1) は列に、誤差の自由度 (ν_2) は行に対応する

例：有意水準を 5% に設定・因子の自由度は 2・誤差の自由度は 27

⇒ 「27」という行と「2」という列の交点の数値 **3.354** をみる

$\alpha=0.05$

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	9
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124
41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118
42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112
43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106
44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096
46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091
47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086
48	4.043	3.191	2.798	2.566	2.409	2.295	2.207	2.138	2.082
49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959
240	3.880	3.033	2.642	2.409	2.252	2.136	2.048	1.977	1.919
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880

練習問題

問 1 (分散分析の実施と解釈)

ある工場では触媒を使って化合物を生成している。異なる 3 種類の触媒 A,B,C で化合物の硬度に差が生じるかどうかを実験により調べ、基本統計量を求めたところ、次の表が得られた。ただし、小数点第 3 位で四捨五入した。

	繰り返し数 ⁸	平均	標準偏差
水準 1 : 触媒 A	6	5.32	0.70
水準 2 : 触媒 B	5	5.54	0.71
水準 3 : 触媒 C	4	2.93	0.54
	15	4.75	1.30

誤差が平均 0、分散 σ^2 の正規分布に従うとして、データをもとに一元配置分散分析を行った結果は、以下のようになった。

	自由度	平方和	平均平方	F 値	p 値
触媒	(ア)	18.37	(イ)	20.53	0.0001
残差	12	5.37	(ウ)		

[1] (ア) の値はいくつか。次の 1～5 のうちから一つ選べ。

1. 1 2. 2 3. 3 4. 4 5. 5

[2] (ウ) の値はいくらか。次の 1～5 のうちから一つ選べ。

1. 0.45 2. 1.29 3. 1.69 4. 5.37 5. 6.63

[3] この結果から触媒による硬度には差があるか。次の 1～5 のうちから最も適切なものを 1 つ選べ。

- 触媒 A の平均と触媒 B の平均では差があまりないので、硬度には差がない
- 触媒 C の標準偏差が、他の 2 つの触媒の標準偏差に比べ小さいので、硬度の差がある。
- F 値から、硬度には差がないと考えられる。
- F 値から、硬度には差があると考えられる。
- 触媒ごとに繰り返し数が異なるので、分散分析は適用できない。

⁸ 繰り返し数は実験を行った回数を意味し、各条件のケース数をあらわすものである。

練習問題の解答

問 1 (分散分析の実施と解釈)

〔1〕 解答：2

実験条件は3つで、触媒（因子）の自由度は $3-1=2$ となるので、2が正しい。

〔2〕 解答：1

残差の平均平方は **残差平方和÷残差の自由度** で求めることができるので、 $5.37 \div 12 = 0.45$ となる。したがって、1が正しい。

〔3〕 解答：4

分散分析では F 値をもとに、帰無仮説の正否を判断する。有意水準を 5% とすると臨界値は 3.885 となり、 F 値よりも小さいので、帰無仮説を棄却し対立仮説を採択できる。よって、触媒によって硬度に差があるといえる。

なお、表にある p 値をもちいても同様な結論になる。 p 値はデータ上で帰無仮説が成立する確率を示すものである。有意水準よりも p 値が小さい場合、帰無仮説を棄却し対立仮説を採択する。いま、 p 値は 0.0001 で有意水準 0.05（5%）よりも小さいので、帰無仮説を棄却し対立仮説を採択できる。

5. カイ二乗 (χ^2) 検定

本章の目的

- ・ モデルの適合度を検討する方法として、カイ二乗検定を理解する。
- ・ クロス集計表の独立性を検討する方法として、カイ二乗検定を理解する。

Key Words：カイ二乗検定，期待度数，適合度の検定，クロス集計表の独立性の検定

1. 適合度の検定

● カイ二乗検定の考え方

- カイ二乗検定は、ある確率モデルで予想される結果と標本調査による観測結果が同じものかどうかを検討するために用いられる。

例：サイコロがイカサマかどうかを検討する

イカサマがないサイコロであれば、各目が出る確率はすべて $1/6$ に等しいはずである。したがって、各目が出る確率が $1/6$ よりも大きく外れるのであれば、サイコロはイカサマであるはずである。カイ二乗検定を使うことで、実際のサイコロの出目の確率がイカサマのないもの（各目が出る確率はすべて $1/6$ ）と同じかどうかを検討できる。

- カイ二乗検定では、モデルの予想（期待度数）と観測結果（観測度数）のズレを測る指標として、カイ二乗 (χ^2) 統計量が用いられる。

● 適合度の検定の手順

- 1. 統計的仮説の設定

帰無仮説：観測度数と期待度数は一致する

対立仮説：観測度数と期待度数は一致しない

- 2. 標本統計量の計算

データから以下にあるカイ二乗統計量を計算する

$$\text{カイ二乗統計量} = \sum (\text{観測度数} - \text{期待度数})^2 \div \text{期待度数}$$

* 期待度数：想定する確率モデルによって異なる

- 3. 有意水準の設定

◇ 有意水準は一般的には 5% (0.05) に設定される

◇ 自由度を計算し、カイ二乗分布表¹をもとに臨界値を求める

・ 自由度は次の式にもとづき計算する

$$\text{自由度} = \text{分析対象となるセルの数} - 1$$

¹ カイ二乗分布表の見方については付録を参照。

➤ 4. 帰無仮説の正否の判断

カイ二乗統計量と臨界値を比較して、帰無仮説の正否の判断する

☆臨界値 $\geq$ カイ二乗統計量 $\Rightarrow$ 帰無仮説を棄却しない

☆臨界値 $<$ カイ二乗統計量 $\Rightarrow$ 帰無仮説を棄却し、対立仮説を採択する

例：サイコロがイカサマかどうかを検討する

あるサイコロを 600 回ふったところ、以下の表のような結果が得られた。このサイコロがイカサマのないものなのかどうかをカイ二乗検定で検討する。

出目	1	2	3	4	5	6	合計
回数（観測度数）	85	113	106	92	105	99	600
期待度数	100	100	100	100	100	100	600

*サイコロにイカサマがない場合、各目の出る確率は $1/6$ となる。

サイコロを 600 回ふったので、各目が出る回数は $600 \times (1/6) = 100$ となることが期待される。この回数が期待度数である。

➤ 1. 統計的仮説の設定

帰無仮説：観測度数と期待度数は一致する

対立仮説：観測度数と期待度数は一致しない

➤ 2. 標本統計量の計算

以下のやり方でカイ二乗統計量を計算する

出目	1	2	3	4	5	6	合計
回数（観測度数）	85	113	106	92	105	99	600
期待度数	100	100	100	100	100	100	600
観測度数－期待度数	-15	13	6	-8	5	-1	
(観測度数－期待度数) ²	225	169	36	64	25	1	
(観測度数－期待度数) ² ÷期待度数	2.25	1.69	0.36	0.64	0.25	0.01	5.2

*すべての目に対し、「(観測度数－期待度数)²÷期待度数」をたしたものがカイ二乗統計量となる

➤ 3. 有意水準の設定

✧ 有意水準を 5% (0.05) に設定

✧ 自由度は $6 - 1 = 5$ ²なので、カイ二乗分布表から臨界値を求めると、**11.07**

² サイコロの目の数は 6 なので、分析対象となるセルの数は 6 となる。

➤ 4. 帰無仮説の正否の判断

カイ二乗統計量 (5.2) は臨界値 (11.07) よりも小さいので、帰無仮説は棄却できない。つまり、観測度数と期待度数は異なる、とはいえない。よって、サイコロにイカサマがあるとはいえない。

2. クロス集計表の独立性の検定

● クロス集計表の独立性の検定とは？

➤ クロス集計表³を作成することにより、質的変数の間の関係を検討することができる。クロス集計表の独立性の検定は、分析対象となる変数の間に関係があるかどうかを検討する方法である。

☆ 2つの変数の間に関係がある場合には、変数 X の値の変化によって変数 Y の値が変化する。

☆ 2つの変数の間に関係がない場合には、変数 X の値が変化しても、変数 Y の値が変化しない。このような場合、X と Y は統計的に独立という。

● クロス集計表の独立性の検定の手順

➤ 1. 統計的仮説の設定

帰無仮説：2つの変数の間の関係は独立である⁴。

対立仮説：2つの変数の間の関係は独立でない。

➤ 2. 標本統計量の計算

データから以下にあるカイ二乗統計量を計算する

$$\text{カイ二乗統計量} = \sum (\text{観測度数} - \text{期待度数})^2 \div \text{期待度数}$$

$$\text{期待度数} = (\text{行周辺度数} \times \text{列周辺度数}) \div \text{総数}$$

➤ 3. 有意水準の設定

☆ 有意水準は一般的には 5% (0.05) に設定される

☆ 自由度を計算し、カイ二乗分布表をもとに臨界値を求める

・自由度は次の式にもとづき計算する

$$\text{自由度} = (\text{行の数} - 1) \times (\text{列の数} - 1)$$

➤ 4. 帰無仮説の正否の判断

カイ二乗統計量と臨界値を比較して、帰無仮説の正否の判断する

☆ 臨界値 $\geq$ カイ二乗統計量 $\Rightarrow$ 帰無仮説を棄却しない

☆ 臨界値 $<$ カイ二乗統計量 $\Rightarrow$ 帰無仮説を棄却し、対立仮説を採択する

³ クロス集計表の詳細については、統計検定 3 級用学習コンテンツ「6. 質的変数の関係を明らかにする」を参照のこと。

⁴ 「2つの変数の間に関係はない」ということを意味する。

例：ラーメンとギョウザの好みの関係

A氏は、日本国民を無作為に抽出し、ラーメンとギョウザの好みの関係を調査した。その結果を二重クロス集計表としてまとめたところ、以下の表のようになった。クロス集計表の独立性の検定を使って、ラーメンとギョウザの好みの関係を検討する。

		ギョウザの好み		計
		すき	きらい	
ラーメンの好み	すき	27	93	120
	きらい	33	47	80
	計	60	140	200

➤ 1. 統計的仮説の設定

帰無仮説：ラーメンの好みとギョウザの好みは関係がない。

対立仮説：ラーメンの好みとギョウザの好みは関係がある。

➤ 2. 標本統計量の計算

以下のやり方でカイ二乗統計量を計算する

A. 各セルの期待度数を計算する⁵

		ギョウザの好み		計
		すき	きらい	
ラーメンの好み	すき	36 = (120×60) ÷ 200	84 = (140×120) ÷ 200	120
	きらい	24 = (60×80) ÷ 200	56 = (140×80) ÷ 200	80
	計	60	140	200

* 各セルの期待度数は、(行周辺度数×列周辺度数) ÷ 総数で求める

B. 各セルで(観測度数－期待度数)² ÷ 期待度数を計算する

		ギョウザの好み		計
		すき	きらい	
ラーメンの好み	すき	2.25 = (27-36) ² ÷ 36	0.96 = (93-84) ² ÷ 84	120
	きらい	3.38 = (33-24) ² ÷ 24	1.45 = (47-56) ² ÷ 56	80
	計	60	140	200

C. すべてのセルの値を合計し、カイ二乗統計量を計算する

$$\text{カイ二乗統計量} = 2.25 + 0.96 + 3.38 + 1.45 = 8.04$$

⁵ 各セルの期待度数を計算することは、二変数の間に関係がないクロス集計表を考えると等しい。

- 3. 有意水準の設定
 - ✧ 有意水準を 5% (0.05) に設定
 - ✧ 自由度は $(2-1) \times (2-1) = 1^6$ なので、カイ二乗分布表から臨界値を求めると、**3.841**
- 4. 帰無仮説の正否の判断

カイ二乗統計量 (**8.04**) は臨界値 (**3.84**) よりも大きいので、帰無仮説を棄却し、対立仮説を採択する。つまり、ラーメンの好みとギョウザの好みは関係がある ⁷、といえる。

⁶ この例におけるクロス集計表の行数は 2、列数は 2 である。

⁷ クロス集計表の独立性の検定で帰無仮説が棄却された場合、変数間の関係はクロス集計表をもとに解釈する。

付録 カイ二乗分布表のみかた

- 自由度 (ν) と設定した有意水準 (α) をもとに、対応する臨界値を見つける

例 1 : 有意水準を 5% に設定・自由度は 5

→ 「5」という行と「0.05」という列の交点の数値 11.07 を見る

例 2 : 有意水準を 5% に設定・自由度は 1

→ 「1」という行と「0.05」という列の交点の数値 3.84 を見る

ν	α							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.00	0.00	0.00	0.02	2.71	3.84	例 2	6.63
2	0.02	0.05	0.10	0.21	4.61	5.99		9.21
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.55	0.68	1.15	1.61	9.24	11.07	12.83	15.09
6	0.87	1.24	1.64	2.20	例 1	12.59	14.45	16.81
7	1.24	1.69	2.17	2.83	14.07	16.01	18.48	
8	1.65	2.18	2.73	3.49	15.51	17.53	20.09	
9	2.09	2.70	3.33	4.17	16.92	19.02	21.67	
10	2.56	3.25	3.94	4.87	18.31	20.48	23.21	
11	3.05	3.82	4.57	5.58	19.68	21.92	24.72	
12	3.57	4.40	5.23	6.30	21.03	23.34	26.22	
13	4.11	5.01	5.89	7.04	22.36	24.74	27.69	
14	4.66	5.63	6.57	7.79	23.68	26.12	29.14	
15	5.23	6.26	7.26	8.55	25.00	27.49	30.58	
16	5.81	6.91	7.96	9.31	26.30	28.85	32.00	
17	6.41	7.56	8.67	10.09	27.59	30.19	33.41	
18	7.01	8.23	9.39	10.86	28.87	31.53	34.81	
19	7.63	8.91	10.12	11.65	30.14	32.85	36.19	
20	8.26	9.59	10.85	12.44	31.41	34.17	37.57	
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
35	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34
40	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
80	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
90	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
100	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81
120	86.92	91.57	95.70	100.62	140.23	146.57	152.21	158.95
140	104.03	109.14	113.66	119.03	161.83	168.61	174.65	181.84
160	121.35	126.87	131.76	137.55	183.31	190.52	196.92	204.53
180	138.82	144.74	149.97	156.15	204.70	212.30	219.04	227.06
200	156.43	162.73	168.28	174.84	226.02	233.99	241.06	249.45
240	191.99	198.98	205.14	212.39	268.47	277.14	284.80	293.89

練習問題

問 1 (二重クロス集計表の独立性の検定)

ある法案に賛成か反対かについて住民に調査を行い、男女別に集計して、次の 2 行 3 列のクロス集計表を得た。

	賛成	反対	どちらでもない	行計 n_i
男性	102	48	50	200
女性	72	54	24	150
列計 m_j	174	102	74	$N = 350$

〔1〕 帰無仮説「 H_0 : 法案への賛否と性別は独立である」のもとでクロス集計表の各セルの期待度数を計算したい。次の 1~5 のうちから正しい式を一つ選べ。ただし、第 i 行($i=1,2$)、第 j 列($j=1,2,3$)での期待度数を e_{ij} 、観測度数を n_{ij} 、第 i 行の観測度数の和を $n_i = \sum_{j=1}^3 n_{ij}$ 、第 j 列の観測度数の和を $m_j = \sum_{i=1}^2 n_{ij}$ 、総和を $N = \sum_{i=1}^2 \sum_{j=1}^3 n_{ij}$ と表記する。

1. $e_{ij} = n_{ij} + n_i + m_j - N$
2. $e_{ij} = n_i m_j / N$
3. $e_{ij} = (n_{ij} + n_i + m_j) / 6$
4. $e_{ij} = (2 \times 3) \sqrt{n_i} \sqrt{m_j} / \sqrt{N}$
5. $e_{ij} = (n_i / 3 + m_j / 2) / 2$

〔2〕 法案への賛否と性別は独立であるかを見るために、上のデータについてカイ二乗統計量を求めたところ、 $\chi^2 = 7.674$ であった。有意水準を 5% とした検定結果として、次の 1~5 のうちから適切なものを一つ選べ。

1. χ^2 分布表の自由度 6 の行の値から、法案の賛否と性別は独立ではないと判断できる。
2. χ^2 分布表の自由度 6 の行の値から、法案の賛否と性別は独立であると判断できる。
3. χ^2 分布表の自由度 3 の行の値から、法案の賛否と性別は独立ではないと判断できる。
4. χ^2 分布表の自由度 3 の行の値から、法案の賛否と性別は独立であると判断できる。
5. χ^2 分布表の自由度 2 の行の値から、法案の賛否と性別は独立ではないと判断できる。

練習問題の解答

問 1 (二重クロス集計表の独立性の検定)

[1] 解答 : 2

クロス集計表の独立性の検定では、期待度数は (行周辺度数×列周辺度数) ÷ 総数で計算できる。この式をより正確に表現すると $e_{ij} = n_i m_j / N$ となるので、2 が正しい。

[2] 解答 : 5

分析対象となっているクロス集計表は 2 行 3 列なので、自由度は $(2-1) \times (3-1) = 2$ となる。有意水準は 5% なので、自由度をもとに臨界値を求めると 5.99 になる。したがって、カイ二乗統計量は臨界値よりも大きいので、帰無仮説は棄却される。よって、5 が正しい。

6. 回帰分析

- 本章の目的
- ・ 量的変数の間の関係を探る方法として、(単) 回帰分析を理解する。
- ・ 単回帰分析にかかわる統計的検定を理解する

Key Words：単回帰分析，切片と回帰係数，決定係数

1. 相関係数と回帰分析

- 相関係数は、2 つの量的変数の間の関係の方向性や関係の強さを測定するものである。つまり、相関係数は関係のあり方を測定するモノサシとしての役割が強い。

例：勉強時間と試験の得点の関係

相関係数を計算することで、勉強時間と得点の関係の方向性や関係の強さが分かる

- 回帰分析は、一方の変数の値から他方の変数の値を予測する方法である。つまり、回帰分析は予測をおこなう道具としての役割が強い。

例：勉強時間と試験の得点の関係

回帰分析をおこなうことで、勉強時間を 1 時間増やした場合に得点が何点あがるのかということや、勉強時間が 3 時間だった場合に得点が何点になるのかということをも明らかにできる。

2. 単回帰分析の基礎

- 単回帰分析の考え方
 - 回帰分析は独立変数¹ (説明変数) が 1 単位変化したときに従属変数² (目的変数・被説明変数) がどのくらい変化するかを明らかにする分析手法である。
 - ☆ 二つの変数の間に直線的な関係があると考えてデータを分析する。回帰分析の結果、以下の 3 つの関係のうちどれが成立しているかが分かる
 - 正比例の関係
 - 逆比例の関係
 - 無関係の関係
 - 回帰分析では、次の直線の式 (回帰式) をデータにあてはめ、切片および回帰係数を推定³することが目的となる。

$$Y \text{ (従属変数)} = a \text{ (切片)} + b \text{ (回帰係数)} \times X \text{ (独立変数)}$$

¹ 独立変数は原因と考えられる変数を指す。

² 従属変数は結果となる変数を指す。

³ 切片および回帰係数の推定には、最小二乗法という方法を用いる。最小二乗法は、従属変数の値と回帰式の差が最小となるようなかたちで、切片と回帰係数の推定を行う

- 単回帰分析の結果を理解する

- 切片：a

- ✧ 独立変数の値が 0 のときの従属変数の値
- ✧ 単回帰分析の場合、切片は次の式で求めることができる。

$$\text{切片} = \text{従属変数の平均値} - \text{回帰係数} \times \text{独立変数の平均値}$$

- 回帰係数：b

- ✧ 独立変数の値が 1 単位増加した場合の従属変数の変化量
- ✧ 単回帰分析の場合、回帰係数は次の式で求めることができる。

$$\text{回帰係数} = \text{相関係数} \times \frac{\text{従属変数の標準偏差}}{\text{独立変数の標準偏差}}$$

- 決定係数：R²

- ✧ 従属変数のもつ情報のうち、回帰式で説明できるものの割合
 - データに対する回帰式の適合度合いの指標として扱われる
- ✧ 単回帰分析⁴の場合、決定係数は(相関係数)²で求めることができる

例：勉強時間とテストの成績の関係

10 人の学生に対し、1 日あたりの勉強時間（X）と統計学の試験の点数（Y）を調査したところ、次のような結果が得られた。勉強時間から試験の得点を予測するために、回帰分析を行う。

	1 日の勉強時間	統計学の点数
①	1	45
②	1	40
③	2	70
④	2	70
⑤	2	80
⑥	3	65
⑦	3	85
⑧	4	80
⑨	5	75
⑩	7	90

⁴ 重回帰分析の場合は、決定係数は重相関係数の二乗を計算することで求めることができる。

これらの変数の平均値・標準偏差・相関係数⁵を求めたところ、以下のようになった。

	1 日の勉強時間 (X)	統計学の点数 (Y)
平均値	3	70
標準偏差	1.8	15.5
相関係数	0.72	

このデータから、回帰式を推定すると以下のような結果になる。

- 回帰係数 : $b = 0.72 \times (15.5 \div 1.8) = 6.2$
 - 勉強時間が 1 時間ふえると試験の点数は 6.2 点あがる
- 切片 : $a = 70 - 6.2 \times 3 = 51.4$
 - 勉強時間が 0 時間だと、試験の得点は 51.4 点になる
- 決定係数 : $R^2 = (0.72)^2 = 0.52$
 - 試験の得点の持つ情報のうち、回帰式は約 52% を説明できる

この結果をもとに、勉強時間が 6 時間の人の得点を予測すると、以下のようになる。

$$51.4 + 6.2 \times 6 = 88.6 \text{ (点)}$$

* 回帰式が「得点 = $51.4 + 6.2 \times \text{勉強時間}$ 」なので、勉強時間に 6 時間を代入すると、予測値を求めることができる。

3. 単回帰分析にかかわる統計的検定

- 回帰式の検定
 - 標本調査から得られたデータに対し、単回帰分析を行う場合には、作成した回帰式に意味があるかどうかを検討するために、回帰式の検定を行う必要がある。
 - 回帰式の検定の手順
 - ◇ 1. 統計的仮説の設定

帰無仮説：作成した回帰式はデータにフィットしていない

対立仮説：作成した回帰式はデータにフィットしている
 - ◇ 2. 検定統計量の計算

データから以下の分散分析表を作成する

分散分析表を作成することで、F 値を求める。

⁵ 相関係数の詳細については、統計検定 3 級用学習コンテンツ「5. 量的変数の関係を明らかにする」を参照のこと。

	平方和	自由度	平均平方	F 値
回帰	$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})$	1	$MS_R = SS_R$	$F = \frac{MS_R}{MS_E}$
残差	$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)$	ケース数-2	$MS_E = SS_E / \text{ケース数}-2$	
全体	$SS_T = \sum_{i=1}^n (y_i - \bar{y})$	ケース数-1		

* $\hat{y}_i$ は回帰式にもとづく、 y の予測値を意味する。

$\bar{y}$ は y の平均値を意味する。

◇ 3. 有意水準の設定

有意水準は一般的には 5% (0.05) に設定される

自由度を計算し、F 分布表⁶をもとに臨界値を求める

- ・ 臨界値を求めるために、次の 2 つの自由度を計算する

因子の自由度=1

残差の自由度=ケース数-2

◇ 4. 帰無仮説の正否の判断

F 値と臨界値を比較して、帰無仮説の正否の判断する

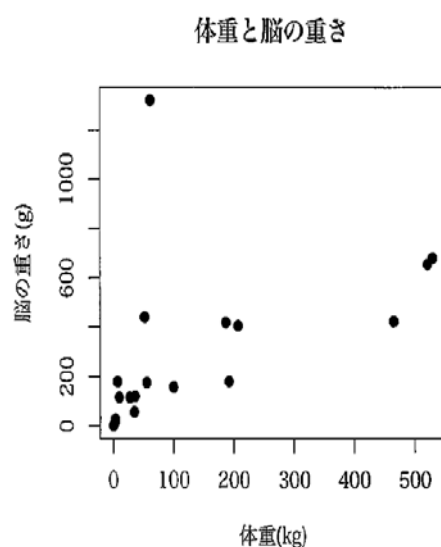
- 臨界値 $\geq$ F 値 $\Rightarrow$ 帰無仮説を棄却しない
- 臨界値 $<$ F 値 $\Rightarrow$ 帰無仮説を棄却し、対立仮説を採択する

⁶F 分布表のみかたについては、統計検定 2 級用学習コンテンツ「4. 平均値の差の検定」を参照のこと。

練習問題

問 1 (単回帰分析)

ある動物学者は、動物の体重(kg)と脳の重さ(g)について、「体重が重くなると脳も重くなる」という仮説を立てた。そこで、象などの大型動物を除いた 23 種類の動物それぞれの体重の平均と脳の重さの平均について、散布図を作成し相関係数を求めたところ、 $r = 0.54$ であった。



同僚の統計学教授に相談したところ、「自然対数 $\log$ を用いて、 $\log$ (体重) を説明変数 x とし、 $\log$ (脳の重さ) を応答変数 (従属変数) y として回帰分析を行なって検討しては」と助言された。そこで、この助言のもとに線形単回帰分析を適用した。

基本統計量	平均	標準偏差
$\log$ (体重)	2.56	2.94
$\log$ (脳の重さ)	4.07	2.32

ある統計解析ソフトによる最小二乗法による結果は次のとおりである。

残差

最小値	第 1 四分位数	中央値	第 3 四分位数	最大値
-0.88	-0.51	-0.26	0.15	1.96

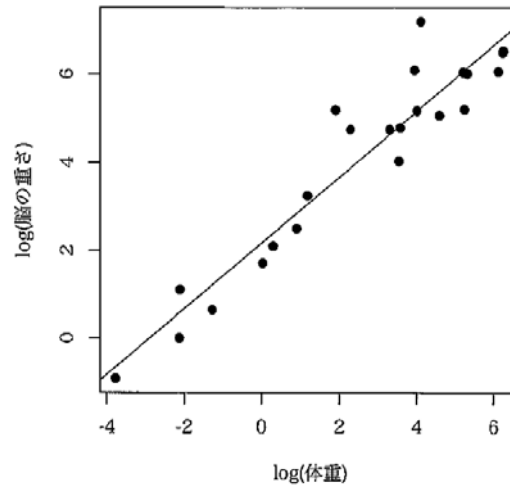
回帰係数

	推定値	標準誤差	t 値	両側 p 値
切片	2.16	0.21	10.23	$< 1.29 \times 10^{-9}$
x	0.75	0.05	13.65	$< 6.52 \times 10^{-12}$

決定係数 R^2 0.90 自由度調整済み決定係数 0.89

F 値 186.4 自由度 (1, 21) p 値 6.52×10^{-12}

脳の重さについての回帰分析



[1] $\log$ (体重) と $\log$ (脳の重さ) についての相関係数はいくらか。次の 1~5 のうちから適切なものを一つ選べ。

1. 0.54 2. 0.75 3. 0.89 4. 0.90 5. 0.95

[2] この結果からデータへの直線の当てはまりの程度を知りたい。どの値を用いるのが、最も適切か。次の 1~5 のうちから一つ選べ。

1. 切片の t 値 10.23 2. 説明変数 x の回帰係数 0.75
3. 説明変数 x の t 値 13.65 4. 決定係数 R^2 0.90
5. F 値 186.4

[3] この結果を用いると、 $\log$ (体重) の値が大きいほど、 $\log$ (脳の重さ) が大きくなる、傾向が読み取れる。回帰式 $y = a + bx$ の係数を読み取り、体重 80kg の動物の脳の重さを予測したい。脳の重さの予測値として、次の 1~5 のうちから最も適切なものを一つ選べ。ただし、 $\log(80\text{kg}) = 4.38$ であり、 e は自然対数の底とする。

1. $\log(5.45) = 1.70$ 2. 5.45 3. 13.52 4. 62.16
5. $e^{5.45} = 231.60$

練習問題の解答

問 1 (単回帰分析)

[1] 解答 : 5

決定係数は相関係数の値の二乗を計算することで求められる。逆に、決定係数の値が分かっているときに相関係数の値を求めるには、決定係数の正の平方根をもとめればよい。

よって、求めるべき相関係数の値は $\sqrt{0.90} = 0.95$ となり、5 が正しい。

[2] 解答 : 4

決定係数はデータに対する回帰式の適合度合いの指標であるので、決定係数の値を見ればよい。よって、4 が正しい。

[3] 解答 : 5

データから推定された回帰式は以下のとおりである。

$$\log(\text{脳の重さ}) = 2.16 + 0.75 \times \log(\text{体重})$$

予測したい動物の体重は 80kg なので、この値を回帰式に代入する。

$$\begin{aligned}\log(\text{脳の重さ}) &= 2.16 + 0.75 \times \log(80\text{kg}) \\ &= 2.16 + 0.75 \times 4.38 \quad (\because \log(80\text{kg}) = 4.38) \\ &\doteq 5.45\end{aligned}$$

応答変数の測定単位は log なので、体重の単位に戻すために、以下の計算を行う。

$$\text{脳の重さ (g)} = e^{5.45} = 231.60$$

よって、5 が正しい。