目次

Ⅰ 確率

上級編では、初級編、中級編で学んだ内容を踏まえ、統計学の基礎について説明します。ここでは統計的推測に欠かせない確率の考え方について、基本的な事項を説明します。

1. ランダムネスと確率

統計学においては、判断はそのデータが得られる確率に基づいて行われます。これが統計的推測の基礎です。「確率」と関係の深い言葉に「ランダムネス」があります。ここでのランダムネスとは、何が次に起こるか確定的に予想できないことをいいます。たとえば、これからコインを投げるとき、次に表が出るか裏が出るかをいうことはできません。しかし、全体としてはランダムネスには法則性があります。次に何が出るか言い当てられないにしても、表、裏が半々ずつでることは、予想ができます。確率論はランダムネスの法則を扱う数学理論です。

2. 確率の公理

確率論の公理は次の三つです。

(a)$P(\Omega) = 1$

(b)$0 \leq P(A) \leq 1$

(c) 事象Aと事象Bが互いに排反ならば、

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

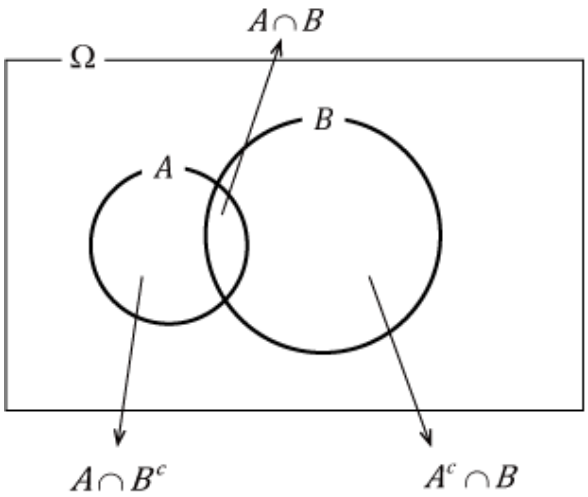
① 加法定理

AとBが排反、すなわち$A \cap B = \emptyset$であったとき、確率の公理(c)より、

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

となります。これを加法定理と呼びます。

ここで、AとBが排反ではない場合を考えます。$A \cup B$は、$A \cap B^c$, $A^c \cap B$, $A \cap B$の三つの事象の和事象であり、これらの事象は互いに排反であるため、



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

となります。また、$A = (A \cap B^c) \cup (A \cap B)$であり、$A \cap B^c$と$A \cap B$は排反事象であるから、

$$P(A) = P(A \cap B^c) + P(A \cap B)$$

です。同様に

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

となるので、

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

となります。これは三つ以上の事象でも成り立ちます。

② 乗法定理

AとBに対して、条件付き確率を$P(B|A)$とするとき、条件付き確率の定義式

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

を変形すると

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

となります。この式は、乗法定理と呼ばれます。

3. ベイズの定理

いくつかのつぼに黒と白の玉が混ざって入っているものとし、このつぼのどれかからいくつかの玉が抜き出されたとします。ただし、どのつぼから抜き出されたかは知られてい

ません。今、抜き出した玉という結果からどのつぼから取り出したかという原因を推定したいと考えます。

ここで、Aを得られた結果、H_1，H_2，…，H_nを原因とします。知りたいのはAが起こったとき原因がH_iである確率$P(H_i|A)$ですが、知ることができるのは原因に対する結果の確率$P(A|H_i)$である場合がほとんどであり、これを直接計算するのは困難です。

ベイズの定理は結果に対する原因の確率$P(H_i|A)$を計算する公式を与えます。

H_iは互いに排反で、これ以上の原因はないとします。乗法定理から

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

が成り立ちます。右辺の分子を乗法定理を用いて書き換えることにより、次のようになります。

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}$$

右辺の分母は、H_iが排反であるという条件から、次のようになります。

$$\begin{aligned} P(A) &= P((A \cap H_1) \cup (A \cap H_2) \cup \cdots \cup (A \cap H_n)) \\ &= P(A \cap H_1) + P(A \cap H_2) + \cdots P(A \cap H_n) \\ &= P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + \cdots + P(H_n)P(A|H_n) \\ &= \sum_{j=1}^{n} P(H_j)P(A|H_j) \end{aligned}$$

これにより、次の式が成り立ちます。

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{\sum_{j=1}^{n} P(H_j)P(A|H_j)}$$

この式をベイズの定理といいます。

$P(H_i)$を事前確率、$P(H_i|A)$を事後確率といいます。

Ⅱ 確率変数

第Ⅰ章では確率の基本事項を説明しましたが、それだけでは、あまりに一般的、抽象的で、具体的な現象を記述したり、分析したりするには不十分です。ここでは確率的に動く量(変数)、またその値の出方の様子について説明します。

1. 確率変数と確率分布

「乱数サイ」の目は$\{0,1,2,\dots,9\}$をとります。この場合離散的な変数Xが$\{0,1,2,\dots,9\}$のいずれかの値をとることになります。正しいさいころでは、乱数サイの目はランダムネスをもっており、予測はできません。このように変数Xがある特定の値をとることが偶然に支配されており、かつ、ある特定の値xをとる確率が与えられているとき、その変数を確率変数といいます。

離散型確率変数のとる値が$x_1, x_2, \cdots, x_n$である場合、確率分布$P(X = x_i) = f(x_i)(i = 1,2,\cdots,n)$に対して、次の式が成り立ちます。ここで、$P(X = x)$は確率変数$X$が値$x$をとる確率を意味しています。

$$0 \leq f(x_i) \leq 1$$

$$\sum_{i=1}^{n} f(x_i) = 1$$

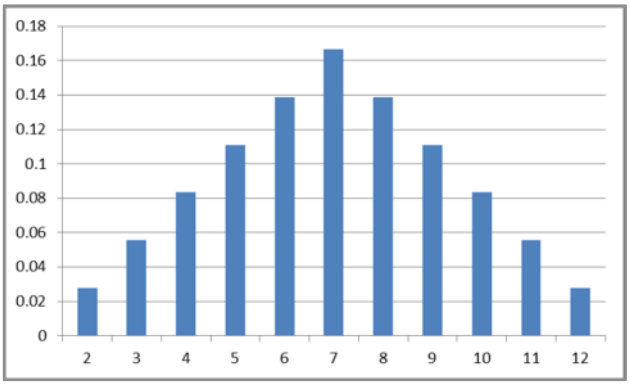
乱数サイの場合は、

$$f(x_i) = 0.1, \quad x_i = i \quad (i = 0,1,\cdots,9)$$

です。

2個のさいころの目の和の出る確率を考えると、下の図表のようになります。

目の和	2	3	4	5	6	7	8	9	10	11	12
場合の数	1	2	3	4	5	6	5	4	3	2	1
確率	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



このfを離散型の確率分布といいます。確率分布は確率の「重み」の分布の様子を表しています。

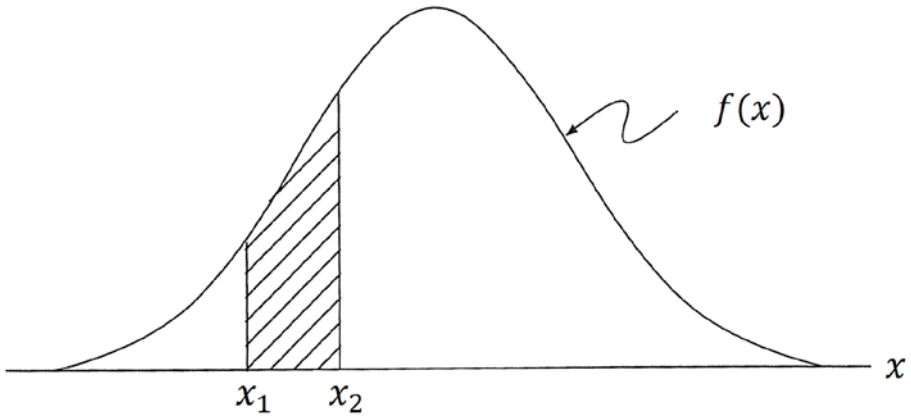
次に連続型の確率変数を定義してみましょう。例えば、日本の高校1年生の男子生徒の身長について考えることにします。A君が165.5cm、B君が168.2cmであったとしましょう。身長は遺伝や生育環境により決まるものですが、高校1年生という集団を考える場合、ある確率でこのような身長の生徒が出現すると考えることが、確率論を用いた推測を行う場合の基本となります。

身長のような連続量の場合は、連続確率変数Xを考える必要があります。連続確率変数の場合は、ある特定の値xになる確率は考えることができず、次のようにある範囲を考える必要があります。すなわち、変数Xがある特定の値をとることが偶然に支配されており、かつ、ある特定の範囲$(x, x+\Delta x)$の値をとる確率が与えられているとき、その変数を連続確率変数と呼びます。

$$\lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = f(x)$$

$$F(x) = P(X \leq x)$$

関数$f(x)$をXの確率密度関数、$F(x)$を確率分布関数(累積密度関数)といいます。



確率密度関数と確率の関係

密度関数は上の図で表され、この場合、Xがx_1からx_2の範囲の値をとる確率は斜線の面積の部分になります。

2. 確率変数の期待値と分散

① 期待値

確率変数はいろいろな値をとりますが、それらの値を代表する平均が考えられます。これが期待値です。

期待値の意味は、Xを例えば、宝くじの確率的な結果と考えると理解しやすいでしょう。人は誰でも主観的には最高額を期待しますが、それは飽くまで根拠のない「期待」であって、客観的な予想はそれよりも低いでしょう。確率論でいう期待値とは、この得られるであろう客観的な予想額のことです。

一般に、確率変数Xに対して、それがとる値の重みつき平均を、確率変数の期待値といい、$E(X)$と書きます。すなわち、

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) \quad (\text{離散型})$$

$$E(X) = \int_{x_{\min}}^{x_{\max}} x f(x)\, dx \quad (\text{連続型})$$

を期待値(平均値)とよび、μで表します。

② 分散

次に分散について考えてみましょう。

Xを1個のさいころを振ったときに表れる目、Yを2個のさいころを振ったときの2個の目X_1, X_2の相加平均$Y = \frac{(X_1+X_2)}{2}$とすると、それぞれの期待値は、

$$E(X) = \frac{21}{6} = \frac{7}{2}$$

$$E(Y) = E\left\{\frac{(X_1 + X_2)}{2}\right\}$$

$$= \frac{\{E(X_1) + E(X_2)\}}{2}$$

$$= \frac{\{\left(\frac{7}{2}\right) + \left(\frac{7}{2}\right)\}}{2}$$

$$= \frac{7}{2}$$

となり、両者は等しい値となります。しかし、Xの確率分布は1,2,…,6の上に一様にばらついているのに対し、Yの確率分布は期待値 $\frac{7}{2}$ に集中しています。

期待値は確率変数の重要な指標ですが、期待値が同じでもばらつきがことなれば、確率分布は異なります。集中やばらつきを表すために、期待値$E(X)$からのずれの量$X-E(X)$を考えます。期待値をμと表すと、$X-\mu$はそれ自体確率変数であるので、目安としてその平均(期待値)を考えると

$$E(X-\mu)=E(X)-\mu=\mu-\mu=0$$

となり、打ち消しあってしまうため、二乗をとり、

$$V(X)=E\{(X-\mu)^2\}$$

と分散を定義します。上の定義から、必ず$V(X)\geq 0$であり、$V(X)$の値が大きいほど、Xのばらつきは大きいことを表しています。

分散の計算は、定義から

$$V(X)=\sum_{i=1}^{n}(x_i-\mu)^2 f(x_i) \quad (\text{離散型})$$

$$V(X)=\int_{x_{min}}^{x_{max}}(x-\mu)^2 f(x)\,dx \quad (\text{連続型})$$

となり、σ^2と表します。なお、実際の計算では

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

$$=E(X^2)-\mu^2$$

$$=E(X^2)-\{E(X)\}^2$$

を用いるほうが計算は簡単です。

1個のさいころを振る例では、

$$E(X^2)=1^2\times(\tfrac{1}{6})+2^2\times(\tfrac{1}{6})+\cdots+6^2\times(\tfrac{1}{6})=\frac{91}{6}$$

であるから、$V(X)=\frac{91}{6}-(\frac{7}{2})^2=\frac{35}{12}$

2個のさいころを振ったときの2個の目X_1, X_2の相加平均$Y=\frac{(X_1+X_2)}{2}$の場合は

$$E(Y^2)=1^2\cdot\left(\frac{1}{36}\right)+\left(\frac{3}{2}\right)^2\cdot\left(\frac{2}{36}\right)+\cdots+\left(\frac{11}{2}\right)^2\cdot\left(\frac{2}{36}\right)+6^2\cdot\left(\frac{1}{36}\right)$$

$$=\frac{1957}{4\cdot 36}=\frac{329}{24}$$

$$V(Y) = \frac{329}{24} - \left(\frac{7}{2}\right)^2 = \frac{35}{24}$$

となります。$V(X) > V(Y)$となっていることが分かります。

Ⅲ 確率分布

第Ⅱ章では確率変数とその確率分布の一般的ルールについて説明しましたが、現象にはそれぞれ、それを表すのにふさわしい確率分布があります。

ここではそれぞれの現象に当てはまる代表的な確率分布について簡単に説明します。

1. 超幾何分布

2種類A、BからなるN個のものがあり、個数は$A = M$個、$B = N - M$個であるとします。この集団から勝手にn個取り出したときに、Aがx個、Bが$n - x$個であるとすると、xの最小値は$0(n < N - M$のとき$)$あるいは$n - (N - M)(n \geq N - M$のとき$)$であり、最大値は$n(n < M$のとき$)$あるいは$M(n \geq M$のとき$)$となります。その確率は、組合せを計算してみると、

$$f(x) = {}_{M}C_{x} \cdot \frac{{}_{N-M}C_{n-x}}{{}_{N}C_{n}}$$

$$x = \mathrm{Min}(0, n - (N - M)), \ldots, \mathrm{Max}(n, M)$$

で与えられます。この確率分布を超幾何分布といいます。

ある湖の中にいる魚1000匹のうち、200匹に赤い色の標識をつけて放流されているとします。今、湖から魚を5匹捕ったとき、そのうち標識がついた魚が1匹である確率はどれくらいでしょうか。

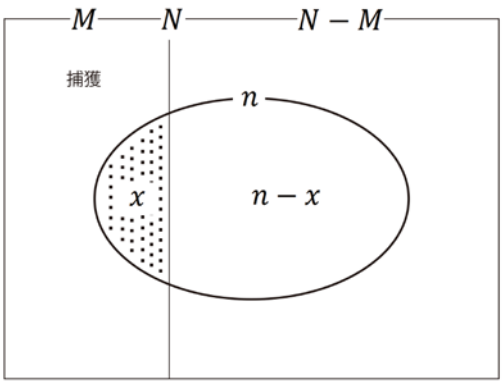
1匹である確率は、$N = 1000,\ M = 200,\ n = 5$として計算すると

$$f(1) = {}_{200}C_{1} \cdot \frac{{}_{800}C_{4}}{{}_{1000}C_{5}} = 0.41063$$

となります。0匹である確率は、

$$f(0) = \frac{{}_{800}C_{5}}{{}_{1000}C_{5}} = 0.32686$$

から、可能性は1匹の方が高いことが分かります。$x = 0,1$で0.7を超えるので、$f(1)$が最大の確率となります。また、$x = 1$は比例式$N{:}\,M = n{:}\,x$が成り立つ場合であることに注意します。実際の場合は、Nは未知ですが、$M, n,\ x$は制御ないし、観察できるため、これらの比例式より魚の個体数Nが推定できます。これは捕獲再捕獲法といわれ、資源調査に使われます。



ここで、n個のものを取り出すときには、1個ずつ元に戻さずに取り出すと考えています。これを非復元の場合といいます。このときはもとの集団の構成はその都度変わっていきますが、それを組合せの数で計算したのが超幾何分布です。

復元の場合には、超幾何分布ではなく二項分布になります。

2. 二項分布とベルヌーイ試行

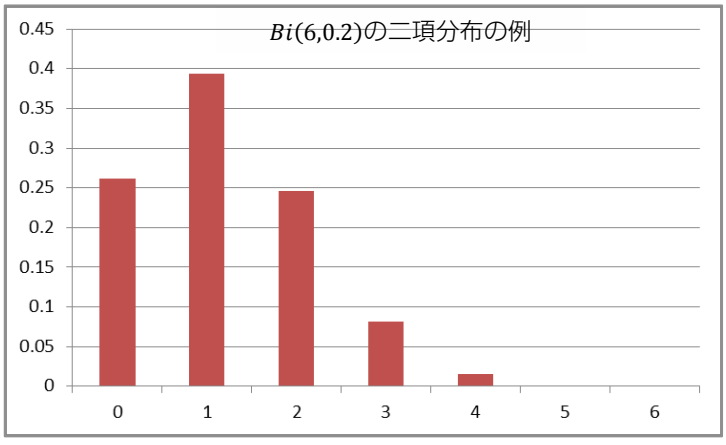
2種類の可能な結果(例えば、成功S、失敗F)を生じる実験又は観測があり、それらの確率がそれぞれp、$1-p$であるとします。これを同じ条件でかつ独立にn回繰り返すことを考えましょう。これをベルヌーイ試行といいます。

今、Sがx回、Fが$n-x$回生じるとすれば、$x=0,1,\dots,n$であり、その確率は、

$$f(x)={}_nC_xP^x(1-P)^{n-x},\quad x=0,1,\dots,n$$

で与えられます。

上記の式は、pのx個の積、$1-p$の$n-x$個の積にSがn回中どのx回で起こるかの場合の数${}_nC_x$を乗じています。この確率分布を二項分布といい、$Bi(1,p)$で表します。また、$Bi(n,p)$をベルヌーイ分布ということがあります。



Xが二項分布$Bi(n,p)$に従っているならば、その期待値と分散は、

$$E(X)=np,\quad V(X)=np(1-p)$$

となります。ベルヌーイ試行で1回当たりの成功率がpでありその試行回数がnであるなら、平均的に$n\times p$回の成功が生じることは直感に合うでしょう。また、分散は$p=\frac{1}{2}$のときに最大となりますが、このとき現象の予測がしにくいことは、例えば雨が降る確率予報が50%のときに傘を持っていくかどうか迷うことからも、理解できるでしょう。

二項分布は平均が最も確率が高い分布となります。

3. ポアソン分布

二項分布において、nが大きい一方で、pが小さい現象を考えてみましょう。

例えば、不動産業において契約成立に到達する確率は極めて小さいとし、仮に$p=0.002$とします。1000件のあっせん申し込みに対し、成立件数が3件となる確率はどうなるでしょう。$n=1000,\ p=0.002$として、

$$f(3)={}_{1000}C_3(0.002)^3(0.998)^{997}$$

となります。$E(X)=np=2$であるので、$x=0,1,2,3$くらいまでの確率は小さくないはずです。この計算には次の定理が使えます。

$np\to\lambda$となるように$n\to\infty$、$p\to 0$となる極限では、各xについて

$${}_nC_xP^x(1-P)^{n-x}\to\frac{e^{-\lambda}\lambda^x}{x!}$$

が成り立ちます。これをポアソンの小数の法則といいます。

これにより計算すると、$\lambda=np=2$であるので、

$$f(3)=e^{-2}\cdot\frac{2^3}{3!}=0.180447$$

となります。

今、$\lambda>0$として

$$f(x)=\frac{e^{-\lambda}\lambda^x}{x!},\quad x=0,1,2,\cdots$$

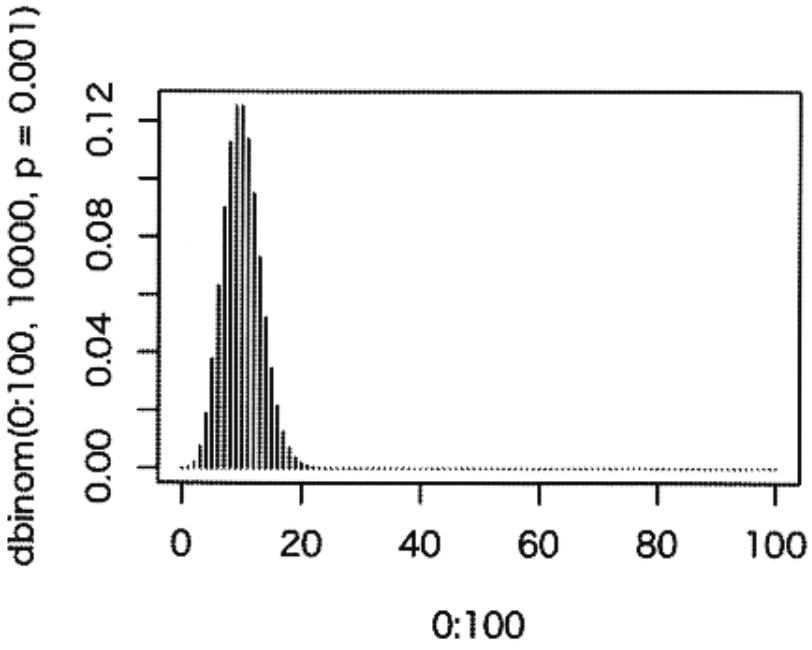
となり、この分布をポアソン分布といい、$Po(\lambda)$と表します。

Xがポアソン分布$Po(\lambda)$に従うならば、その期待値と分散は

$$E(X)=\lambda,\quad V(X)=\lambda$$

となります。このようにポアソン分布においては、期待値と分散が等しくλとなるのが特

徴です。



ポアソン分布の例

ポアソン分布は、交通事故にあう確率、大量生産の不良品数、破産件数など、発生する確率は小さくても回数を大きくすると、ある程度現実に観察されるような事象に用いられます。

4. 一様分布

6個の目から成るさいころを振った場合に出る目Xを考えると、

$$P(X = x) = \frac{1}{6} \quad (x = 1,2, \dots ,5,6)$$

と想定できます。このように値となる確率が等しい分布を離散一様分布といいます。一般的には、$1,2, \dots , n-1, n$となるn個の値については、

$$P(X = x) = \frac{1}{n} \quad (x = 1,2, \dots , n)$$

となり、平均と分散は

$$E(X) = \frac{n+1}{2}$$

$$V(X) = \frac{n^2 - 1}{12}$$

となります。

5. 正規分布

正規分布は代表的な連続型の確率分布で、自然界や人間社会の中の数多くの現象に対して当てはまり、統計学の理論上も応用上も非常に重要な分布です。

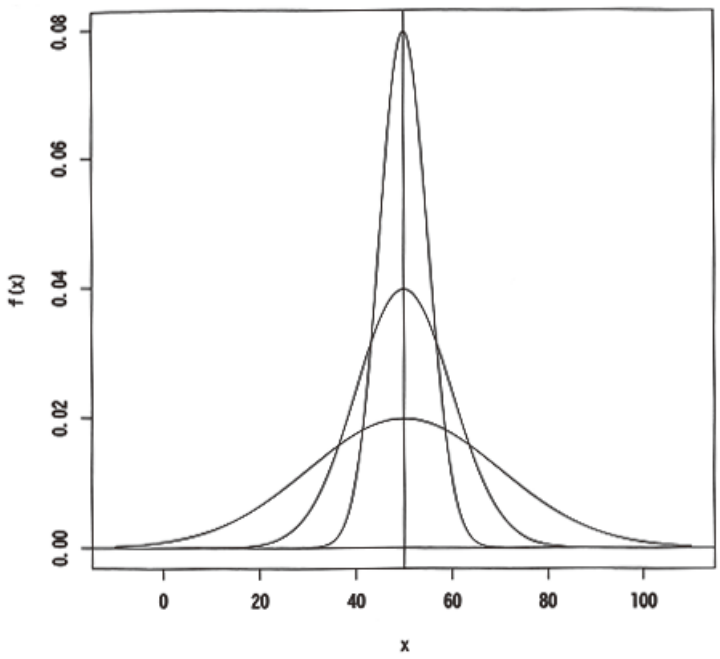
正規分布の確率密度関数は、

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$$

で与えられます。したがって、正規分布はパラメータμとσにより確率が求められます。また、この分布の平均と分散を求めると、

$$E(X) = \mu, \quad V(X) = \sigma^2$$

となります。このことから、平均μ、分散σ^2の正規分布といい、$N(\mu, \sigma^2)$と表しています。



分散10,100,400の正規分布（平均50）の確立密度関数

正規分布の主な特徴は次の二つです。

(a) Xが正規分布$N(\mu, \sigma^2)$に従っているとき、その線形変換$Y = aX + b$は$N(a\mu +$

$b, a^2\sigma^2)$に従う。

(b) 標準化係数$Z = \frac{X-\mu}{\sigma}$は正規分布$N(0, 1^2)$に従う。これを標準正規分布という。

この標準正規分布から

$$P(-1 \leqq Z \leqq 1) = P(\mu - 1\sigma \leqq X \leqq \mu + 1\sigma) = 0.6827$$

$$P(-2 \leqq Z \leqq 2) = P(\mu - 2\sigma \leqq X \leqq \mu + 2\sigma) = 0.9545$$

$$P(-3 \leqq Z \leqq 3) = P(\mu - 3\sigma \leqq X \leqq \mu + 3\sigma) = 0.9973$$

となり簡単に確率が計算できます。

また、正規分布は一般のランダム系列からその和や平均としても生じます。さいころを多数回振った場合の目の和、一様乱数の和などは、その回数nが大きいときにはほぼ正規分布に従って分布します。nが大きくなるときにひとりでに正規分布が出現することを中心極限定理といいます。

6. 指数分布

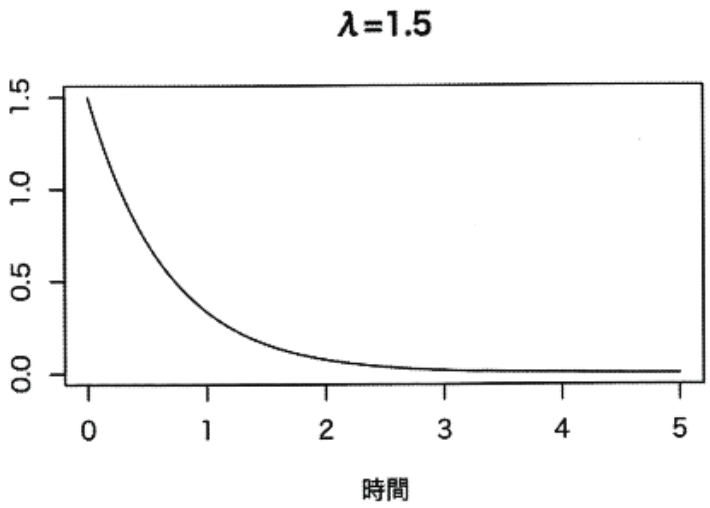
指数分布は、確率密度関数

$$f(x) = \lambda e^{-\lambda x} \ (x \geqq 0), \quad 0(x < 0)$$

で定義される連続型の分布です。平均と分散は

$$E(X) = \frac{1}{\lambda}, \qquad V(X) = \frac{1}{\lambda^2}$$

となります。



指数分布の例

指数分布の特徴は、$x = 0$で確率が最大であること、つまり、小さい値ほど出やすいことです。この分布は連続的な待ち時間分布の性質を持ちます。故障率が一定のシステムの偶発的な故障までの待ち時間、すなわち寿命、耐用年数などがこの例となります。また、災害までの年数も同様であると言われています。指数分布によって生起までの年数が分布する希少事象は、近い将来起こっても不自然ではありません。確率が小さいことと遠い将来にしか起こらないことは同じではありません。

7. χ^2分布

正規確率変数XからもとめたZについては

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

となっています。実際の分析では、偏差を扱うことが多いので、平均からの偏差の二乗$Z^2 = \frac{(X-\mu)^2}{\sigma^2}$について考えてみましょう。平均は$E(Z^2) = 1$ですが、$Z^2$も確率変数であるので、その分布は0以上のさまざまな値をとります。この分布が自由度1のχ^2分布です。$E(Z^2) = 1$、$V(Z^2) = 2\times1$

特に標本分散S^2について、これを確率変数として表すと

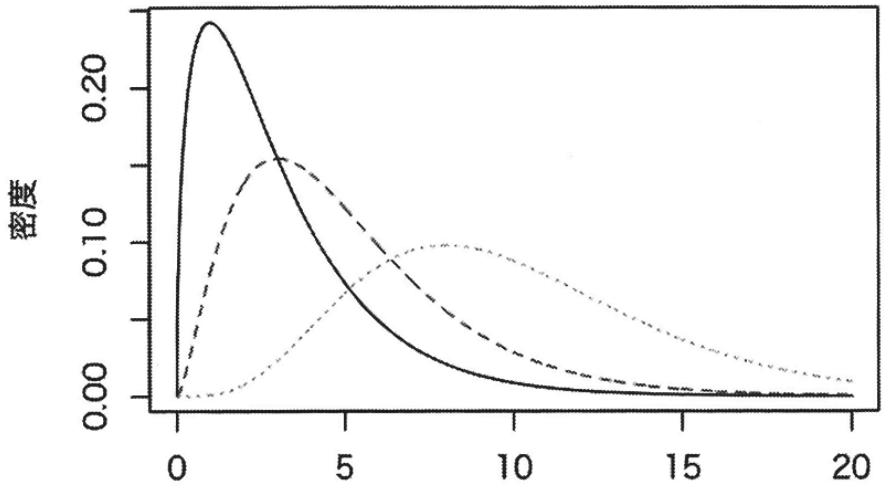
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

とした場合に

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

となるので、分散の大きさについて検定することができます。



自由度 $df = 3, 5, 10$の場合のχ^2分布

8. t分布

成人男性の身長の平均を推定することを考えてみましょう。

正規確率変数Xから求めたZについては

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

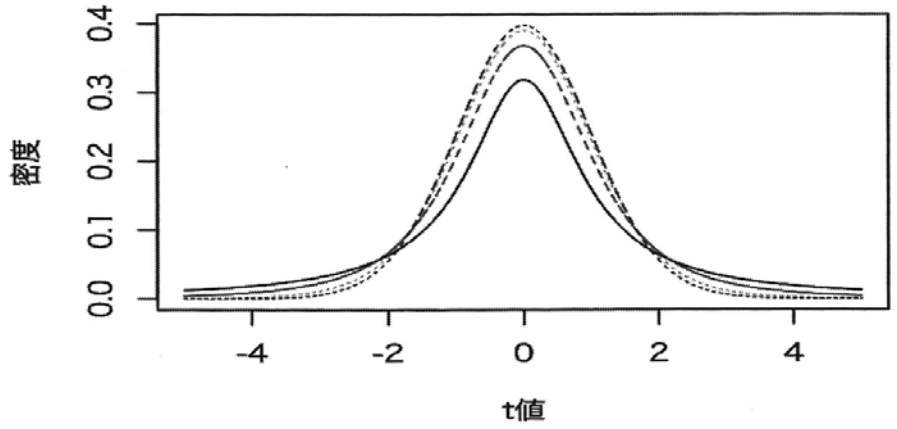
となっています。また、平均については中心極限定理から

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

となっています。そこで、

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

となる確率変数を考えてみましょう。これはZにおいて母分散が不明であるので、代わりに標本分散を用いた場合に対応しています。この確率変数の分布は対称ですが正規分布に比べて裾が長いt分布となります。



自由度 $df = 1, 3, 10, 50$の場合の t 分布

nが大きくなるとt分布は標準正規分布に近づきます。t分布については、自由度が1である場合には平均と分散はともに存在しませんが、自由度$n-1$が3以上であれば、

$$E(t) = 0$$

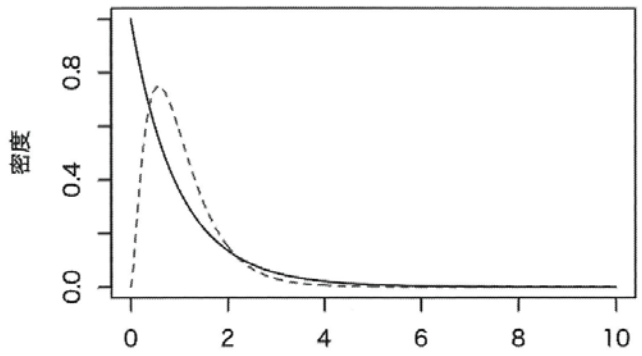
$$V(t) = \frac{n}{n-2}$$

となります。

9. F分布

成人男性と成人女性について、身長に差があるかどうか検定することを考えてみましょう。男性の身長Y_mと女性の身長Y_fについて、正規分布に従うとしてモデルを仮定してみましょう。この場合には、モデルにおいて説明された平均変動と平均残差の比について統計的に比較できます。

今、二つの独立な確率変数UとVについて、それぞれが自由度k_1のχ^2分布と自由度k_2のχ^2分布に従っているとします。この場合に比$F = \frac{U/k_1}{V/k_2}$ を考えると、比は自由度$(k_1,\ k_2)$のF分布に従います。



自由度 (*df1* ,*df2*) = (2, 50), (5, 50)の場合のF分布

この分布は二つの自由度k_1とk_2により分布の形が異なります。Fの平均$E(F)$は自由度$k_2 > 2$の場合には次のようになります。

$$E(F) = \frac{k_2}{(k_2 - 2)}$$

また、Fの分散$V(F)$は、自由度$k_2 > 4$の場合には次のようになります。

$$V(F) = \frac{2{k_2}^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

Ⅳ 標本分布

母集団から無作為抽出等により標本を抽出したとき、標本から計算される統計量が従う確率分布を標本分布といいます。この章では、標本分布から計算される平均と母集団平均の関係などについて説明します。

この標本分布を考える前に、まず標本抽出の方法について説明しましょう。

1. 標本抽出

① 単純無作為抽出法と系統抽出法

単純無作為抽出法は、母集団から同じ確率で無作為に標本を抽出する方法です。単純無作為抽出法と類似した抽出法に系統抽出法があります。系統抽出法は、まず母集団の要素全てに番号をつけ、次に第1番目の標本を無作為に抽出します。第2番目以降は番号について同じ間隔で抽出する、という方法です。

② 層化抽出法

層化抽出法はあらかじめ母集団をいくつかの層に分割し、各層から必要な大きさの標本を無作為に抽出する方法です。このような抽出法を行う理由の一つは、層内が等質的な標本であれば誤差分散が小さくなることです。例えば、家計を調べるときに性別や年代や職業などの特性のいずれかで分けられた層に分割し、その中から無作為に対象者を抽出する場合をいいます。

③ クラスター抽出法と多段抽出法

大規模な標本調査を行う場合には、調査対象を直接抽出することが難しい場合があります。このような場合に、抽出単位を何段階かに分けて、まず、第一次抽出単位をある確率で抽出し、次に抽出した第一次抽出単位の中から、更にある確率で第二次抽出単位を抽出します。たとえば、全国学校調査では、県を抽出し、その県から学校を抽出し、その学校から学級(クラス)を抽出し、そこから児童(生徒)を抽出します。このような手順で指定した段数までを行うのが多段抽出法です。段数が多くなるほど、平均などの推定精度は悪くなります。そのために、層化抽出法と多段抽出法を組み合わせた層化多段抽出法なども使われます。

クラスター抽出法では、母集団を網羅的に分割し小集団(クラスター)を構成します。次にいくつかのクラスターを抽出し、その成員全員を対象者とします。あらかじめクラスターごとの名簿があれば、時間と費用が節約できます。ただし、精度は低下します。この方法はエリアマーケティングなどに用いられます。

2. 標本分布

母集団から無作為抽出等により標本を抽出したとき、標本から計算される統計量が従う確率分布を標本分布といいます。母集団分布がどのようなものであっても、抽出方法に偏りがなく、標本サイズnが十分大きい場合は、標本平均は近似的に正規分布に従うと考えられます。

① 確率変数の和の分布

母集団からの確率的な抽出を考えたとき、標本の各観測値$x_i(i=1,\dots,n)$は母集団分布に従う確率変数$X_i(i=1,\dots,n)$の実現値ですが、正しい確率的な抽出を行っている限り、x_iは独立同分布の実現値であるとみなすことができます。したがって、同時確率密度関数$f(x_1,x_2,\dots,x_n)$はそれぞれの確率密度関数$g(x_i)$の積と考えてよいでしょう。

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}g(x_i)$$

二つの確率変数の和の場合と同様にn個の独立同分布の確率変数の和の平均と分散は、$E(X_i)=\mu,\ V(X_i)=\sigma^2$とすると、次のようになります。

$$E(X_1+X_2+\cdots+X_n)=n\mu$$

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

② 平均の期待値とその分散

標本平均に対応する確率変数は、

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

です。これは、和の平均から

$$E(\bar{X})=\frac{n\mu}{n}=\mu$$

となり、期待値が母平均μと一致します。傾向として、$\bar{X}$はμを過大にも過小にも推定せず、平均的に正しく推定します。更にnが大きくなるとμに集中する傾向が見られます。なぜなら

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

だからです。したがって、nが大きくなると$\bar{X}$の分散は0に近づき、$\bar{X}$はμ(母平均)に収束していきます。

V 推定

推定とは、標本をもとにその標本が抽出されたもとの母集団分布の母数を求めることです。

たとえば、ある集団の所得分布が対数正規分布に従うといっても、その平均、分散が知られていなければ、現実の経済分析や政策評価に用いることはできません。

また、人間の身体計測値が正規分布$N(\mu,\ \sigma^2)$に従うといっても、その平均μ、分散σ^2が分からなければ、洋服のメーカーや公衆衛生学者にとっては何の意味もありません。

更には単位時間当たりの電話の呼び出し数、料金ゲートへの車の到着台数、交通事故や火災件数などがポアソン分布$Po(\lambda)$に従うといっても、λが分からなければ、電話会社、高速道路株式会社、生命・損害保険会社にとっては、何の役にも立ちません。

このように母集団の母数は実際の問題では未知であり、これらを標本から定める必要があります。これを母数の推定といいます。

1. 点推定と区間推定

母集団平均μを標本平均$\bar{x}$で推定する場合のように、母集団の未知の母数θを推定する場合、それをある一つの値$\hat{\theta}$で指定する方法を点推定と呼びます。

これに対して、適当な区間を与え、その範囲に母数が含まれるという表現を用いる方法があります。これを区間推定と呼びます。まず、点推定について説明し、その後区間推定について説明します。

2. 点推定

① 母集団分布と母数

標本から得られる観測値$x_i(i=1,\ldots,n)$は、実験を繰り返すたびに変動する確率変数です。例えば、光速を測定する実験を行ったとすると、観測値は測定誤差によって変動することから、真の光速μを平均とする正規分布$N(\mu,\ \sigma^2)$に従うと考えられます。ここで分散σ^2の大小は測定の精度によって定まりますが、観測された標本平均$\bar{x}$はどの程度母集団平均μに近いといえるのでしょうか。

一般に標本$x_i(i=1,\ldots,n)$に従う確率分布を$f(x|\theta)$と表し、その分布を特定する母数(パラメータ)θを推定するために利用される推定量を$\hat{\theta}=T(x_1,\cdots,x_n)$と表します。

光速の例では、母集団分布は正規分布、その母数は$\theta=(\mu,\ \sigma^2)$と2次元です。母集団分布の平均と分散は重要な母数であり、母集団平均、母集団分散は略して母平均、母分散と呼ばれることがあります。母平均の推定量$\hat{\mu}$としては標本平均$\bar{x}$が代表的ですが、この

他にも標本の中央値や刈込み平均$\bar{x}_\alpha$など、$\hat{\mu}$には多くの候補があります。

母集団の分布をどのように想定するかは、現実の観測や実験の内容を確認し、経験や理論的考察を通じて導かれるものです。たとえば、光速の例や、身長、テストの点数などについては、正規分布とみなしてよいでしょう。また、介護が必要な世帯員がいる割合を知るために無作為に抽出した世帯に対して面接調査を行って、要介護の家族がいるかを聞いたような場合には、二項分布を想定してよいでしょう。

➢ 無限母集団と有限母集団

実験や観測が同じ条件で繰り返せる場合は、観測値$(x_1, \cdots, x_n)$は互いに独立と考えてかまいません。また、各x_iの分布は同一で、母集団分布そのものです。このような母集団の想定を無限母集団と呼び、$(x_1, \cdots, x_n)$を無作為標本と呼びます。

母集団が有限な場合、たとえば、上の要介護者のいる世帯の割合を調べるような場合で、大都市ではなく全部で$N = 5000$世帯の町を対象として行われるような場合には、独立に同一の分布に従うという想定は成り立ちません。通常の調査方法は非復元単純無作為抽出であり、N枚のカードから$N = 1200$枚を抜き出すように世帯が選ばれます。同一の世帯が繰り返して抽出することがないため、要介護の世帯数は二項分布ではなく超幾何分布に従います。なお、このような有限母集団の場合でも、毎回N枚のカードを戻しながら抽出する世帯を選び、重複を排除しないという復元単純無作為抽出法を採用した場合には、無限母集団と同じく、二項分布が適切なモデルとなります。

② 大数の法則と中心極限定理

標本平均$\bar{x}$は重要な推定量であるだけでなく、さまざまな形で利用されます。そこで、標本平均に関する二つの大定理について、ここで再度確認しておきましょう。

➢ 大数の法則

正しいコインを10回投げることを考えてみましょう。このように1回の実験で2種類の事象(この場合、表か裏)のいずれかが生じ、しかもそのような事象が起こる確率が常に一定(この場合$\frac{1}{2}$)であるような試行をベルヌーイ試行と呼びます。「成功」を表とし、表が出た回数の割合について考えてみましょう。i回目のコイン投げで、表が出た場合1、裏が出た場合0をとる確率変数x_iを考えます。10回のコイン投げで、表の出た回数(頻度)は、和

$$r = x_1 + x_2 + \cdots + x_{10}$$

となります。表が出た回数の割合 $\hat{p} = \frac{r}{10}$ は観測された成功率であって、$\hat{p} = 0, 0.1, 0.2 \cdots$となります。一般に$n$をコイン投げの回数とするとき、$\frac{r}{n}$は相対度数です。$r$は確率変数で、$n = 10$、$P = 0.5$の二項分布$Bi(10, 0.5)$、すなわち

$$f_{10}(x) =_{10} C_x \left(\frac{1}{2}\right)^{10}, \quad x = 0,1,2,\cdots,10$$

に従い、その期待値、分散は

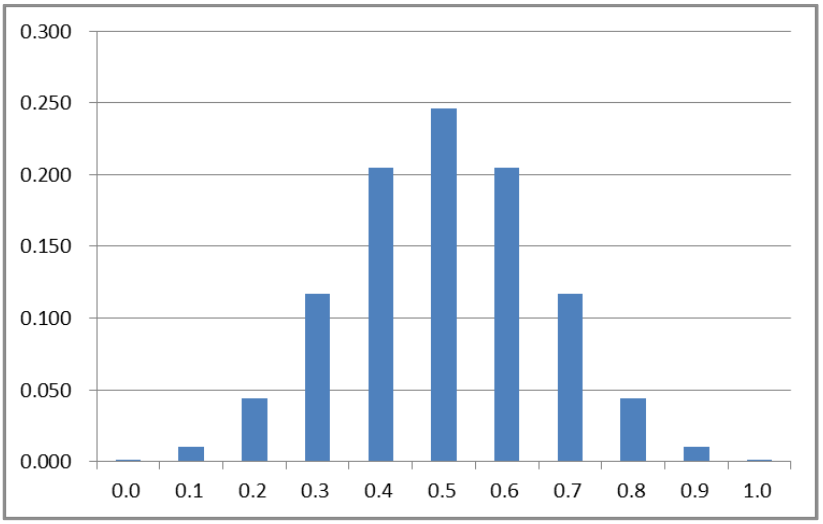
$$E(r) = np = 5, \qquad V(r) = np(1-p) = 2.5$$

であるので、割合r/nの期待値、分散は

$$E\left(\frac{r}{n}\right) = p = 0.5, \qquad V\left(\frac{r}{n}\right) = p(1-p) = 0.025$$

です。ここで、$P = 0.5$は真の成功率となっています。成功の割合が$\frac{x}{10}$となる確率は$f_{10}(x)$で、実際に計算すると次のようになります。

観測された成功率 x/10	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
確率$f_{10}(x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001



期待値である真の成功率0.5及びその周辺の発生確率が高いですが、表の割合が0.2以下、及び0.8以上であるような確率も11.0%近くあります。

ここでコイン投げの回数を増やして$\frac{r}{n}$の期待値$E\left(\frac{r}{n}\right) = p$、及びその周辺が発生する確率がどのようになっていくか調べてみましょう。期待値$P = 0.5$の周辺としては、0.5 ± 0.1の範囲をとり、0.4から0.6までとします。

$$P\left(0.4 \le \frac{r}{10} \le 0.6\right) = \sum_{x=4}^{6} f_{10}(x) = 0.65625$$

$$P\left(0.4 \le \frac{r}{10} \le 0.6\right) = \sum_{x=20}^{30} f_{50}(x) = 0.88108$$

$$P\left(0.4 \leq \frac{r}{10} \leq 0.6\right) = \sum_{x=40}^{60} f_{100}(x) = 0.96780$$

となります。nを増やしていくと確率は上がり、$n = 100$では、表の観測された成功率$\hat{p} = \frac{r}{n}$が0.4から0.6までの確率は96%を超え、ほとんどの値が真の成功率$P = 0.5$の周囲に集中します。

これを式の形で表すと、任意の$\epsilon > 0$に対して

$$\mathrm{P_r}\{|\bar{x} - \mu| > \epsilon\} \to 0 \quad \left(n \to \infty\right)$$

となり、nが大きくなると標本平均$\bar{x}$は母平均μに近づくことを表しています。これを大数の法則といいます。

つまり、大標本では観測された標本平均を真の母集団平均とみなしてよいということになります。

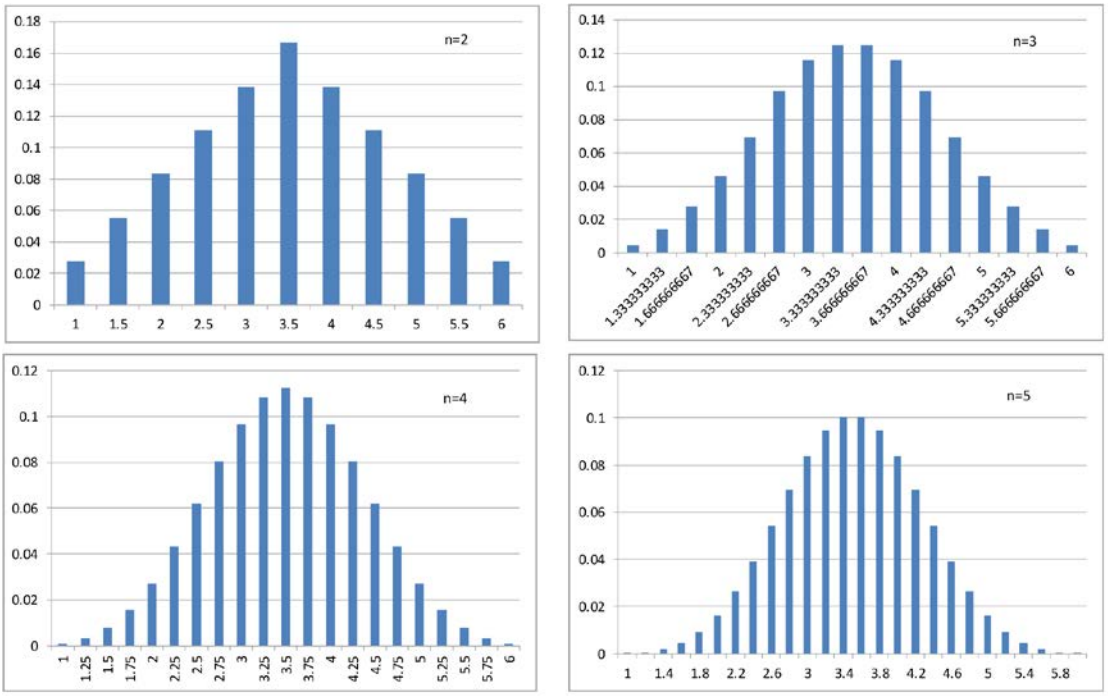
- 中心極限定理

もうひとつの大定理、中心極限定理は、大まかに言えば、母集団分布が何であっても、nが大きいときには、平均$\bar{x}$の確率分布は正規分布に従うと考えてよいというものです。

$\bar{x}$の期待値は$E(\bar{x}) = \mu$とnによらず一定ですが、分散は$V(\bar{x}) = \frac{\sigma^2}{n}$と次第に小さくなります。そこで$\bar{x}$の代わりに$\sqrt{n}(\bar{x} - \mu)$を考えると、その期待値は$0$、分散は$\sigma^2$と一定となります。ここで$n \to \infty$とすると、$\sqrt{n}(\bar{x} - \mu)$の確率分布は正規分布$N(0,\ \sigma^2)$に近づくことが証明できます。これを次のように表すことがあります。

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{D} N(0, \sigma^2) \ \left(n \to \infty\right)$$

さいころの例で考えてみましょう。さいころの目の出方の確率分布は離散型の一様分布であり、$\mu = \frac{7}{2}$、$\sigma^2 = \frac{35}{12}$です。この確率分布は正規分布とはかなり違います。この母集団からランダムに$n = 2$の標本X_1,X_2を取り出したときの標本平均$\frac{X_1+X_2}{2}$とは、つまり、さいころを2回振ったときに出る目の平均値です。このnを増やしていくと、下の図のようになり、正規分布に近づいていくことが分かります。



③ 点推定の基準

推定量は一つとは限らず複数考えることができます。たとえば、身長の分布は$N(\mu,\ \sigma^2)$であることが知られています。今、平均値の母数θを知りたいとしたときに、推定量$\hat{\theta}$の候補としては、標本平均、中央値、刈り込み平均などいくつか挙げられます。

刈込み平均$\bar{x}_\alpha$は、観測値のうち大きな値と小さな値を除いて、残りの観測値の平均値として定義されるものであり、αは両側から刈込む比率を表します。正確には$[n\alpha]$個の観測値を両側から取り除いて、残りの$n-2[n\alpha]$個の観測値から平均を計算します。

これらの候補が実用上意味を持つためには、標本分布が真の母数θの周辺に集中していることを示すいくつかの基準を満たす必要があります。

➢ 不偏性

$\hat{\theta}$はθの推定量であるので、その分布はθの周りに分布していなければなりません。この一つの基準が不偏性です。不偏性とは、推定量$\hat{\theta}$の期待値が$E(\hat{\theta}) = \theta$と、母数に等しくなること、すなわち、推定に平均的に過大・過小の偏りがないことを表しています。これを満たす推定量を不偏推定量といいます。

✧ 平均の不偏推定量

母平均μの推定を考えましょう。$x_1, \cdots, x_n$を無作為標本として、母集団の平均と分散をμ、σ^2とします。このとき、標本平均$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$について次の性質が成り立ちます。

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$V(\bar{x}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(x_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

これは、どのような母集団であっても標本平均$\bar{x}$が母平均μの不偏推定量であることを示しています。

✧ 分散の不偏推定量

母集団分散σ^2の推定量としては標本分散

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

を考えるのが自然です。しかし、その期待値は母集団分散σ^2とは異なり、不偏推定量とはなりません。このことを確かめるには偏差平方和を$T_{xx} = \sum(x_i - \bar{x})^2$とおくと、

$$\begin{aligned} T_{xx} &= \sum (x_i - \bar{x})^2 \\ &= \sum [(x_i - \mu) - (\bar{x} - \mu)]^2 \\ &= \sum (x_i - \mu)^2 - 2(\bar{x} - \mu)\sum (x_i - \mu) + n(\bar{x} - \mu)^2 \\ &= \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \end{aligned}$$

となるので、ここで期待値を計算すると次の結果が得られます。

$$\begin{aligned} E(T_{xx}) &= \sum E[(x_i - \mu)^2] - nE[(\bar{x} - \mu)^2] = \sum V(x_i) - nV(\bar{x}) = n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2 \end{aligned}$$

結局$E[\hat{\sigma}^2] = \left[\frac{(n-1)}{n}\right]\sigma^2 \neq \sigma^2$であり、不偏推定量とはなりません。同時に偏差平方和を$n-1$で割った次式のs^2がσ^2の不偏推定量となることが分かります。

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

➢ 一致性

一致性とは、標本の大きさnが大きくなるに従い、$\hat{\theta}$が真の母数θに近づく性質のことをいいます。

$$\hat{\theta} \xrightarrow{P} \theta$$

この条件を満たす場合、$\hat{\theta}$を一致推定量と呼びます。大数の法則から平均$\bar{x}$はnが大きくなると母平均μに近づきます。母集団分布が対称な場合には、中央値mも刈込み平均$\bar{x}_\alpha$も母平均の一致推定量であることが知られています。

$\hat{\sigma}^2 \xrightarrow{P} \sigma^2$となることも簡単に確認できます。

まず$\hat{\sigma}^2 = \sum \frac{(x_i-\mu)^2}{n} - (\bar{x}-\mu)^2$の第2項は$(\bar{x}-\mu) \xrightarrow{P} 0$なので$(\bar{x}-\mu)^2 \xrightarrow{P} 0$となります。また、$u_i = (x_i-\mu)^2$とおくと、$u_i$は互いに独立で同じ分布に従う確率変数なので、大数の法則が適用できます。これから第1項は$\bar{u} \xrightarrow{P} E(u) = \sigma^2$となります。

$\hat{\sigma}$とsはいずれも母集団標準偏差σの一致推定量です。

➢ 漸近正規性

分析を行うためには、推定量の標本分布が知られていることが望ましいですが、正確な標本分布を知ることは簡単ではありません。しかし、このようなときでも中心極限定理から漸近分布($n \to \infty$における分布)は正規分布であることが多いことが知られています。中心極限定理より、$\bar{x}$の漸近分布は母集団分布に関係なく$N(\mu,\ \sigma^2)$であり、$\bar{x}$は漸近正規推定量です。

➢ 有効性

二つの推定量が、両方とも不偏推定量であり、一致推定量であった場合には、分散が小さな推定量の方が優れています。ただし推定量$\hat{\theta}$の標準偏差は未知なので、その推定量が用いられます。これを標準誤差と呼び、$\mathrm{se}(\hat{\theta})$あるいは単にse、s. e.などと表します。推定値とその標準誤差を組にして表示することもよくあります。

標本平均の場合は、$V(\bar{x}) = \frac{\sigma^2}{n}$なので、$\sigma^2$をその推定値$s^2$で置き換えて$\mathrm{se}(\bar{x}) = \frac{s}{\sqrt{n}}$となります。母集団が正規分布でなくても、$n$が大きければ、近似的に$N(\mu,\ \mathrm{se}^2)$とみなすことができ、やはり$\mathrm{se}(\bar{x}) = \frac{s}{\sqrt{n}}$となります。

3. 区間推定

これまで説明してきた点推定は、θをある一つの値として推定しますが、区間推定はθに

対して確率の考え方を用いた推定を行います。区間推定とは、真の母数の値θがある区間$[L,\ U]$に入る確率を$1-\alpha$(αはθが区間に入らない確率)以上になるように保証する方法であり、

$$P(L \leq \theta \leq U) \geq 1-\alpha$$

となる確率変数L, Uを求めるものです。L, Uはそれぞれ、下側信頼限界、上側信頼限界と言われます。$1-\alpha$は信頼係数と呼び、区間$[L, U]$を$100(1-\alpha)$%信頼区間と呼びます。$1-\alpha$は通常、99%、95%に設定されることが多いです。

① 母集団平均の区間推定

➢ 正規分布の推定

母平均μの信頼区間を構成する方法をいくつかの類型に分類して紹介します。ここでは、標本$x_1, \cdots, x_n$は独立に正規分布$N(\mu, \sigma^2)$に従うものとします。

✧ σ^2が既知のとき

この場合、標本平均$\bar{x}$は正規分布$N(\mu,\ \frac{\sigma^2}{n})$に従う($\bar{x} \sim N(\mu,\ \frac{\sigma^2}{n})$)ため、95%信頼区間は下の式のようになります。

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

✧ σ^2が未知だがnが大きいとき

この場合、標本から計算された標準偏差sを上の式のσに代用します。これは、点推定のところで説明した一致性を根拠にしています。

分析結果では不偏分散s^2の平方根が与えられることが多いですが、$n > 100$であれば、nで割った$\hat{\sigma}^2$との差は1%以下なので、計算上は大きな差はありません。

➢ t分布の利用

σ^2が未知でnが十分に大きくないときは、t分布を利用します。正規分布と対比させると

$$\sigma^2\text{が既知のとき}\quad z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{(標準正規分布)}$$

$$\sigma^2\text{が未知のとき}\quad t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1} \quad \text{(自由度}n-1\text{の}t\text{分布)}$$

となります。要するにsをσに代入し、正規分布に代えてt分布を用います。区間推定のためには自由度$\nu = n-1$のt分布で上側確率が$\frac{\alpha}{2}$となる値t_0を求めます。このとき$\mathrm{P_r}\{|t| \leq t_0\} = 1-\alpha$となるので、$-t_0 \leq \frac{(\bar{x}-\mu)}{(s/\sqrt{n})} \leq t_0$を解けば信頼区間が求められます。$100(1-\alpha)$%信頼区間の公式は次のとおりです。

$$\bar{x} - t_0 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_0 \frac{s}{\sqrt{n}}$$

信頼係数としては95%がよく使われます。t分布表の抜粋を下に掲げておきます。この表の2.5%点の行を見ていくと、自由度が60のときt_0は約2.0となり、それより大きいと次第に1.96に近づいていくことが分かります。自由度が240より大きければ、正規分布とほとんど同じ結論を得ることになります。nが大きいときにsをσに代用してよいというのはこの性質によるものです。

自由度	10	20	30	60	120	240	正規分布
0.5%点	3.17	2.85	2.75	2.66	2.62	2.60	2.5758
2.5%点	2.23	2.09	2.04	2.00	1.98	1.97	1.9600
5.0%点	1.81	1.72	1.70	1.67	1.66	1.65	1.6449

➢ 二つの母平均の差の区間推定

教材A、Bそれぞれを使った授業を受ける生徒の成績を二つの母集団として、実際に試験を受けた生徒たちの成績を二つの標本と考えると、母平均の差を推定することができます。

今、A、B二つの母集団からの無作為標本を$x_1, \cdots, x_m, y_1, \cdots, y_n$とし、それぞれの母平均$\mu_1$、$\mu_2$の推定量として$\bar{x} = \sum \frac{x_i}{m}$と$\bar{y} = \sum \frac{y_i}{n}$を利用します。母集団の分布に関しては、正規分布が利用できる場合、又は中心極限定理によって$\bar{x} \sim N\left(\mu_1, \frac{{\sigma_1}^2}{m}\right)$と$\bar{y} \sim N\left(\mu_2, \frac{{\sigma_2}^2}{n}\right)$とみなせる場合を考えます。このとき、$\bar{x}$と$\bar{y}$の差$d = \bar{x} - \bar{y}$を推定量とします。

一般に、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ます。ここでは確率変数$\bar{x}$と$\bar{y}$が正規分布に従うので、dも正規分布に従います。その平均は$\delta = \mu_1 - \mu_2$、分散は$V(d) = {\sigma_1}^2/m + {\sigma_2}^2/n = \sigma^2(1/m + 1/n)$となります。ここで$\bar{x}$と$\bar{y}$は独立であるので、共分散は0で$V(d)$には現れません。

以上の結果から、$d = \bar{x} - \bar{y}$は$\delta = \mu_1 - \mu_2$の不偏推定量となっていることが分かります。なお、正規分布に従う確率変数の一次式が正規分布に従うことは重要な性質です。

✧ 分散が既知のとき

このとき

$$z = \frac{d - \delta}{\sqrt{{\sigma_1}^2/m + {\sigma_2}^2/n}} \sim N(0,1)$$

となるので、標準正規分布の上側2.5%点$z_0 = 1.96$を用いて、次の式のように信頼区間が

求められます。

$$d-1.96\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}} \leq \delta \leq d+1.96\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}$$

✧ 分散が未知だが等しいとき

この場合、$\sigma_1^2=\sigma_2^2=\sigma^2$を共通の推定値$s^2$で置換えた確率変数が自由度$m+n-2$の$t$分布に従うことを利用して、信頼区間を構成します。

$$t=\frac{d-\delta}{\sqrt{\frac{1}{m}+\frac{1}{n}}s} \sim t_{m+n-2}$$

推定値s^2についてはx,yのそれぞれから不偏分散が計算できるため、それらを合成します。$s_1^2=\sum\frac{(x_i-\bar{x})^2}{(m-1)}$、$s_2^2=\sum\frac{(y_i-\bar{y})^2}{(n-1)}$とすると、合成された分散は

$$s^2=\frac{(m-1)s_1^2+(n-1)s_2^2}{(m-1)+(n-1)}$$

と求められます。この推定量は不偏です。自由度が二つ減るのは平方和の計算でμ_1とμ_2に代えて$\bar{x}$と$\bar{y}$を用いているためです。

➢ 対応がある標本の場合

例えば、本人と父親の身長の組のようなデータの場合、子の身長と親の身長は互いに独立な観測値ではありません。このように各標本が観測値の対$(x_i,y_i)(i=1,\ \ldots,n)$として与えられる場合を対比較あるいは対応のある場合と呼びます。

この問題では差$d_i=x_i-y_i$を観測値と考えれば、平均の差に関する区間推定は1標本の場合に帰着されるので、dの平均$\bar{d}$と分散s^2を求めればよいことになります。

② 母集団分散の区間推定

光速の測定に関して、分散の信頼区間を求めてみましょう。正規分布からの標本$x_i \sim N(\mu,\ \frac{\sigma^2}{n})(i=1,\cdots,n)$については、次の統計量は自由度$n-1$の$\chi^2$分布に従います。

$$\chi^2=\frac{\sum(x_i-\bar{x})^2}{\sigma^2}=\frac{(n-1)s^2}{\sigma^2} \sim {\chi^2}_{n-1}$$

自由度$\nu=\ n-1$のχ^2分布の下側と上側の$\frac{100\alpha}{2}$点をそれぞれχ_L^2、χ_U^2と書くと、

$\mathrm{P_r}\left\{\chi_L^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_U^2\right\}=1-\alpha$なので、次の信頼区間が得られます。

$$\frac{(n-1)s^2}{\chi_U^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_L^2}$$

③　母集団比率の区間推定

母集団における要介護者がいる世帯の比率pを推定したい場合、母集団は有限でもその大きさNが十分大きいときには、標本のうち該当する世帯の数xは二項分布$B(n,p)$に従うと想定します。xの期待値と分散はそれぞれ

$$E(x)=np,\ \ V(x)=npq\quad(\text{ただし}q=1-p)$$

です。

正確には、母集団の大きさNが小さい場合には非復元抽出を適用するとxの分布は超幾何分布であり、その期待値と分散は次のとおりです。

$$E(x)=np,\ \ V(x)=\frac{N-n}{N-1}npq\fallingdotseq(1-f)npq$$

ここで$f=n/N$は抽出率であり、これが非復元単純無作為抽出の場合に用いられる式です。有限母集団の場合に必要な修正項である$\frac{N-n}{(N-1)}\fallingdotseq 1-f$は有限母集団修正と呼ばれています。この式からも$N$がある程度大きければ、二項分布の想定は妥当であることが分かります。

ここでpの推定量として標本平均$\hat{p}=\frac{x}{n}$を用いると、その平均と分散は次のとおりで、特に$\hat{p}$はpの不偏推定量です。

$$E(\hat{p})=p,\ \ V(\hat{p})=\frac{pq}{n}$$

更にnが大きい場合、二項分布に関する中心極限定理によって次のzは近似的に標準正規分布に従います。

$$z=\frac{\hat{p}-p}{\sqrt{pq/n}}\sim N(0,1)$$

従って、標準正規分布の上側 $\frac{100\alpha}{2}$%点をz_0 とすると、次の表現が得られます。

$$\mathrm{P_r}\left\{\hat{p}-z_0\sqrt{\frac{p(1-p)}{n}}\le p\le\hat{p}+z_0\sqrt{\frac{p(1-p)}{n}}\right\}=1-\alpha$$

左辺の{ }内の不等式をpについて解けば信頼区間が得られますが、ここでは、zの分母が$\hat{p}$の標準偏差であることに着目し、もともとnが大きいので$\hat{p}\xrightarrow{p}p$という大数の法則を利用して、これを推定値すなわち標準誤差で置換えます。その結果、次の形の信頼区間が得られます。

$$\hat{p}-z_0\sqrt{\frac{\hat{p}\hat{q}}{n}}\le p\le\hat{p}+z_0\sqrt{\frac{\hat{p}\hat{q}}{n}}\quad\text{ただし}\hat{q}=1-\hat{p}$$

ここでは、(a)有限母集団を無限母集団とみなす、(b)中心極限定理を利用して正規分布で近似する、(c)推定量の標準偏差を標準誤差で置換える、という三つの近似が使われています。

④ 回帰係数の区間推定

最も基本的な単回帰モデルでは、説明変数xによって応答変数yが次の関係式で定められます。

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \quad (i = 1, \cdots, n)$$

ここで誤差項ϵ_iは互いに独立と仮定されます。標準的な仮定では説明変数xはあらかじめ選ばれて固定された値とされますが、その仮定を緩めてxも確率的に変動するとしてもかまいません。ただし、その場合にはxはϵと独立であり、推定したい母数$(\alpha,\ \beta,\ \sigma^2)$とは無関係な確率分布に従うものとします。例えば、無作為に抽出したn世帯に関して、今月の収入xと消費支出yとの関係を表す消費関数$y = \alpha + \beta x$を回帰分析を用いて推定するときは(x, y)が両方とも確率的に変動しますが、この場合でも標準的な回帰分析モデルの結果が利用できます。

- 回帰係数の推定方法

最小二乗法を用いるとxの係数は$\hat{\beta} = \frac{T_{xy}}{T_{xx}}$と定められます。ただし、

$T_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$は$x$と$y$の偏差の積和、$T_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})^2$は$x$の残差平方和です。分散$s_x^2 = \frac{T_{xx}}{(n-1)}$と共分散$s_{xy} = \frac{T_{xy}}{(n-1)}$を用いて、$\hat{\beta} = \frac{s_{xy}}{s_x^2}$と表現することもできます。なお、$y$の分散と偏差平方和もそれぞれ$s_y^2$、$T_{yy}$という記号で表すこととします。

誤差項ϵ_iを想定しているモデルでは、最小二乗法によって得られた$\hat{\beta}$は確率変数であり、βの不偏推定量です。このことを確かめるために$T_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})y_i$及び$\sum_{i=1}^{0}(x_i - \bar{x}) = 0$という性質を利用し、$y_i$に$\alpha + \beta x_i + \epsilon_i$を代入して整理すると

$$T_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i) = \beta \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i$$

が得られます。最後の式の第2項に確率変数ϵ_iが現れていることから、確かに$\hat{\beta}$は確率変数であることが分かります。更に$w_i = (x_i - \bar{x})/T_{xx}$と表すと、$\hat{\beta} = \beta + \sum_{i=1}^{n} w_i \epsilon_i$という表現が得られます。この形から、正規分布に従うϵ_iの1次式である$\hat{\beta}$はやはり正規分布に従うことが導かれます。更にその期待値と分散は次の式で評価されます。

$$E(\hat{\beta}) = \beta + \sum_{i=1}^{N} w_i\, E(\epsilon_i) = \beta$$

$$V(\hat{\beta}) = \sum_{i=1}^{N} w^2 V(\epsilon_i) = \left(\sum_{i=1}^{N} w_i^2\right)\sigma^2 = \frac{\sigma^2}{T_{xx}}$$

この式は$\hat{\beta}$がβの不偏推定量であることを表しています。以上をまとめると次の表現が得られます。

$$\hat{\beta} = \frac{T_{xy}}{T_{xx}} \sim N(\beta, \sigma^2 (T_{xx})^{-1})$$

誤差項ϵ_iの分散σ^2を推定するには、まず各観測値$(i = 1, \ldots, n)$に対応して、y_iの予測値$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$と残差$e_i = y_i - \hat{y}_i$を求めます。残差$e_i$は誤差$\epsilon_i$に近いと考えれば、これから$V(\epsilon) = \sigma^2$が推定できることは理解できるでしょう。次の$s^2$が$\sigma^2$の不偏推定量となります。

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

次の式で対比されるように、zの式に現れる未知のσ^2をs^2で置換えた統計量tは自由度$\nu = n - 2$のt分布に従うことが導かれます。この式のtはt-値あるいはt-比と呼ばれます。

$$z = \frac{\hat{\beta} - \beta}{\sigma / \sqrt{T_{xx}}} \sim N(0,1) \qquad \text{及び} \quad t = \frac{\hat{\beta} - \beta}{s / \sqrt{T_{xx}}} \sim t_{n-2}$$

$\hat{\beta}$の信頼区間は、自由度$n - 2$のt分布から上側$100\alpha/2\%$点t_0を求めれば、$\mathrm{P_r}\{|t| \le t_0\} = 1 - \alpha$となるので、次の信頼区間が得られます。

$$\hat{\beta} - t_0 \frac{s}{\sqrt{T_{xx}}} \le \beta \le \hat{\beta} + t_0 \frac{s}{\sqrt{T_{xx}}}$$

ここで、$\mathrm{se}(\hat{\beta}) = s/\sqrt{T_{xx}}$なので、この記号を用いて次のように簡単に表せます。

$$\hat{\beta} - t_0 \mathrm{se}(\hat{\beta}) \le \beta \le \hat{\beta} + t_0 \mathrm{se}(\hat{\beta})$$

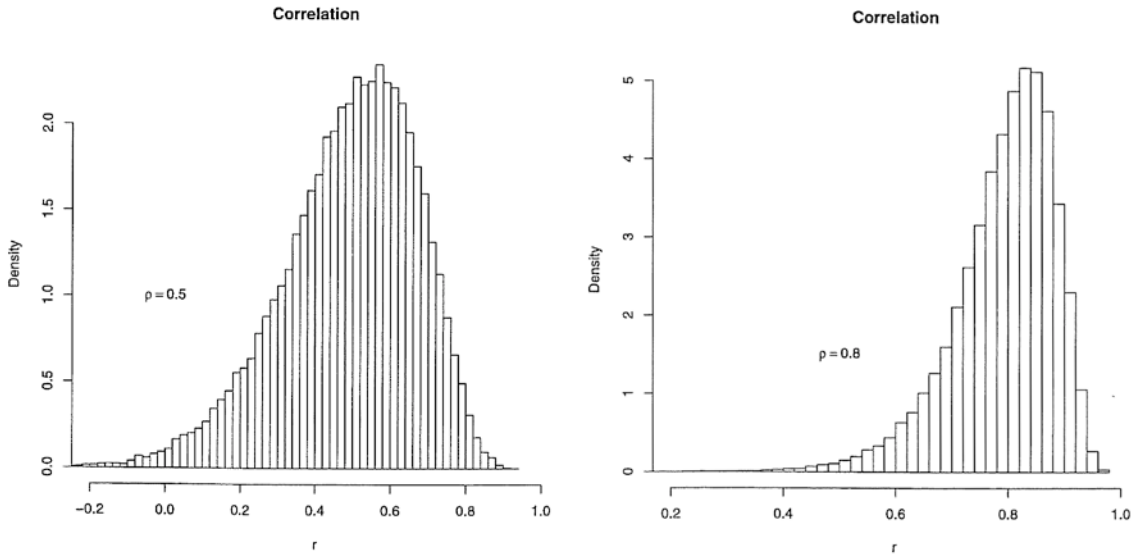
なお、分散s^2の推定に現れる残差平方和は、次の式を利用して計算することができます。

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = T_{yy} - \hat{\beta} T_{xy} = (n-1)\{s_y^2 - \hat{\beta} s_{xy}\}$$

⑤ 母集団相関係数の区間推定

標本の相関係数$r = s_{xy}/(s_x s_y)$の確率分布は、母集団が正規分布の場合であっても複雑であり、近似的な議論が必要です。$(x_i, y_i)(i = 1, \cdots, n)$が相関係数を$\rho$とする正規分

布に従うとき、大きさ$n = 20$の標本を実験的に発生して、そのたびにrを求めてヒストグラムを描くと下の図のようになります。

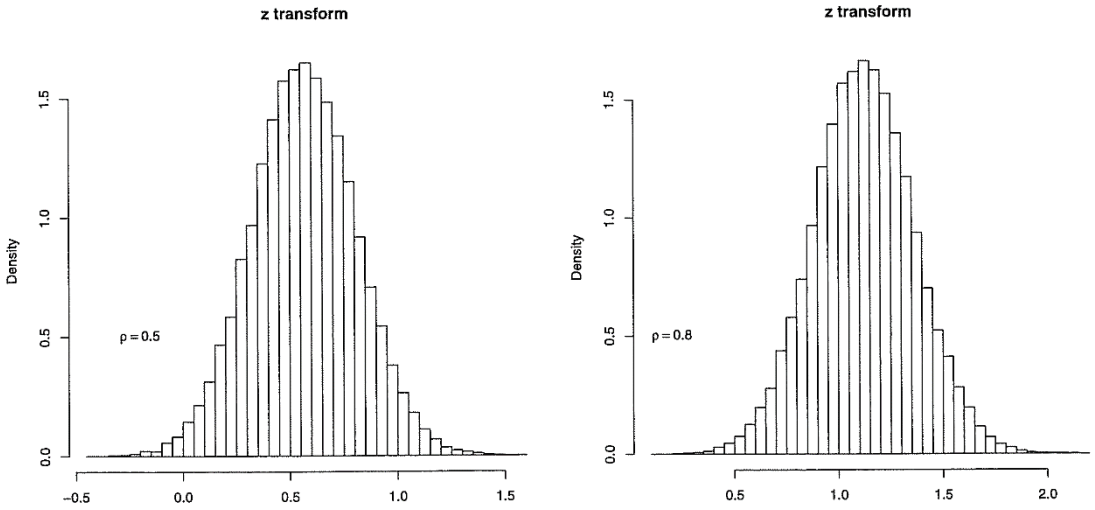


相関係数の分布(左：ρ=0.5, 右：ρ=0.8)

この図でも明らかなように、rの分布はρがゼロから離れると非対称性が強くなります。中心極限定理が働いて正規分布に近くなるとはいえ、nが極めて大きい場合を除いて非対称な分布となります。このときは次の図のように

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

と変換したzの分布はρの値によらず対象となります。更にzは近似的に平均、分散の正規分布に従うことが知られています。なお、必ずしも$(x,\ y)$が正規分布でなくてもこの近似を使うことができます。



zの分布 (左 : $\rho = 0.5$, 右 : $\rho = 0.8$)

練習問題 （解答は P.70 です）

問 1 2011年に首都圏の勤労者世帯$n = 2500$を対象に実施された調査で、当該月の消費支出の平均が$\bar{x} = 32.8$万円、標準偏差が$s = 29.5$万円となった。

(1) 首都圏の勤労者世帯を母集団とする平均消費支出額μについて95%信頼区間を求めよ。

(2) 調査世帯数が$n = 25$で、$\bar{x} = 32.8$、$s = 29.5$となった場合にμの95%信頼区間を求め、その信頼性について論評せよ。

Ⅵ 仮説検定

仮説検定とは、統計的仮説の有意性の検定のことをいいます。仮説のもとで期待した値と観測した結果との違いを、これらの差が単に偶然によって起こったものかどうかという観点から、確率の基準で評価します。

1. 仮説検定の考え方

① 有意水準

仮説検定の目的は、母集団について仮定された命題を標本に基づいて検証することです。

たとえば、ある水族館で飼育されているタコには、サッカーの試合で勝利チームを予言する能力があると言われているとします。20試合中14試合の結果を的中させたとすると、このタコには予言能力があると言えるでしょうか。ここで、「このタコには予言能力はなく、無作為にチームを選んでいる」と仮定してみましょう。

この仮説の下でn回中x回的中する確率は$p = \frac{1}{2}$の二項分布$B(n, p)$を用いて

$p_r(x|H_0) = {}_nC_x \left(\frac{1}{2}\right)^n$と求められますが、20試合中14試合の結果を的中する確率は、${}_{20}C_{14} \left(\frac{1}{2}\right)^{20} = 0.03696$となります。ここで14回以上的中する確率は$P(x \geqq 14) = 1 - 0.9423 = 0.0577$となり、$x = 14$という標本は仮説からすればかなりはずれた値です。

このように、仮説のもとで発生する確率が小さい事象が観測された場合には、仮説の妥当性に疑いが生じます。このとき、仮説は棄却されるといいます。つまり、仮説検定とは、仮説が有意であるか否かに応じて、それを棄却するかあるいは棄却しないかを決定することです。

ここでは、0.0577を「まれである」と判断しましたが、一般にあらかじめどの程度の希少確率を考えるかにより、有意か否かが変わります。この基準の確率を有意水準といい、αで表します。例えば、$\alpha = 0.1$とすると、0.0577はまれだと判断されますが、$\alpha = 0.01$とした場合にはまれではないことになり、仮説は棄却されないことになります。

有意水準には1%(0.01)又は5%(0.05)がよく使われます。

② 帰無仮説と対立仮説

先ほどの「このタコには予言能力はなく、無作為にチームを選んでいる」という仮説$\left(p = \frac{1}{2}\right)$は有意水準$\alpha = 0.1$で棄却されました。棄却されたことで判断が終わるという考え方もありますが、pについて何らかの積極的な判断をしたいのであれば、あらかじめ、もう一つの仮説を$p \neq \frac{1}{2}$と立てておき、もう一つの仮説が採択されたとしましょう。

このとき、もとの仮説$p = \frac{1}{2}$を帰無仮説(H_0)、これに対立する仮説を対立仮説(H_1)といいます。

帰無仮説を棄却するかしないかの決定に関しては、次の二つの誤りがあります。帰無仮説が正しいときに誤って棄却する誤りを第1種過誤、帰無仮説が誤っているのに受容する誤りを第2種過誤といいます。

品質管理のための抜き取り検査などでは、第1種過誤は当然合格するはずの良製品に不合格の判定を下してしまう誤り(生産者のリスク)、第2種過誤は当然不合格であるはずの不良製品に合格の判定を下してしまう誤り(消費者のリスク)のように考えられます。

判断	真実: H_0が正しい	真実: H_0が誤り(H_1が正しい)
棄却	第1種過誤	正しい判断
受容	正しい判断	第2種過誤

③　棄却域と両側・片側検定

仮説検定の考え方と具体的な計算をよりよく理解するために、t検定の例を説明します。

正規分布に従う母集団$N(\mu, \sigma^2)$から大きさ$n = 25$の標本を抽出して、$\bar{x} = 13.7,\ s = 2.3$を得ました。これをもとに、帰無仮説$H_0: \mu = 15$ を対立仮説$H_1: \mu \neq 15$に対して、有意水準$\alpha = 0.05$で検定します。

検定に用いる統計量(検定統計量)として、t統計量を計算すると、

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{13.7 - 15}{2.3/\sqrt{25}} = -2.83$$

となります。有意水準$\alpha = 0.05$に対応するt分布のパーセント点は、$t_{0.025}(24) = 2.06$なので、$|-2.83| > 2.06$より、有意水準5%で帰無仮説を棄却します。

考えてみると、$\bar{x} = 13.7,\ \mu = 15$から、その差は1.3となっています。$\bar{x}$の分散の推定値は$\frac{s^2}{n}$、標準偏差は$\frac{s}{\sqrt{n}} = \frac{2.3}{\sqrt{25}} = 0.46$なので、この差1.3は標準偏差の3倍近く$(t = 2.83)$でかなり大きいといえます。なぜなら、$\pm 2.06$以上の差は確率0.05というまれなもので、2.83はそれ以上に異常な差であるからです。したがって、$\mu = 15$と考えることに無理があります。これが、ここでのt統計量の具体的な意味です。

今、帰無仮説はそのままにして、対立仮説を$H_1': \mu < 15$とすると、$\bar{x}$が$\mu = 15$より相当に小さくなった場合にだけ、帰無仮説を棄却すればよいので、

$$t = \frac{\bar{x} - 15}{s/\sqrt{n}}$$

において、tが十分負になったときにだけ、帰無仮説を棄却すれば良いことになります。$t_{0.05}(24) = 1.71$なので、$-2.83 < -1.71$より、この場合も有意水準5%で帰無仮説を棄却します。

帰無仮説を棄却すべき統計量の値の集合を棄却域といい、棄却しない領域を採択域といいます。採択域は、$t = 0$の周辺の領域(15に近い$\bar{x}$の値に対応)になりますが、棄却域は対立仮説がH_1のような両側対立仮説のときはtの値が著しく0から左右に外れた領域$|t| > t_{\frac{\alpha}{2}}(n-1)$となります。これを両側検定といいます。確率$\alpha$を左右に按分しています。$H_1'$のような片側対立仮説に対しては、$\bar{x}$が十分小さいところに棄却域が定められ、片側検定と言います。

一般に、両側検定は母数θの値がある目標値θ_0と等しいかどうかだけを調べる場合に用いられます。たとえば、工場で新しく機械を購入したとしましょう。機械が正しく働いていれば、材料や運転条件によるばらつきがあるにしても、製品は目標値の近くのものであるはずです。製品が目標値から大きく異なることは、機械が正しく働いていないことを意味しています。このような場合は両側検定となります。

片側検定は母数の大きさが論理的・経験的に予測される場合に使われます。たとえば、英語の特別授業の効果を調べる場合を考えましょう。英語の特別授業の前後での英語の試験の点数の平均を比べる場合、特別授業に効果があれば試験後の点数が良くなっているはずです。このような場合、われわれが知りたいのは授業前後の得点が異なっていることだけではなくて、授業後の得点が向上したかどうかです。このような場合には、対立仮説を不等号で与える片側検定を用います。

伝統的な手順では、仮説を棄却する基準としてあらかじめ固定した優位水準αが用いられますが、関連する判断基準にP-値があります。P-値は確率値、又は観測された有意水準とも呼ばれます。P-値は帰無仮説が正しいときに「検定統計量が観測された値と同じかそれ以上に極端な値をとる確率」と定義されます。片側検定の場合にこの定義は明確です。

$\bar{x}$を検定統計量とする片側検定で$\bar{x}$が大きなときに棄却する場合、観測値を$\bar{x}_{obs}$とするとP-値は$\mathrm{P_r}(\bar{x} \geq \bar{x}_{obs}|H_0)$と評価されます。もし$P$-値が0.03となったとすると、有意水準5%なら帰無仮説は棄却されますが、有意水準1%であれば棄却されません。

両側検定の場合には定義が必ずしも明確ではなく、片側検定のP-値の2倍とすることがあります。また「観測結果と比べて出現する確率が小さい事象の確率」とすることもあります。この場合、標本$x = (x_1, \ldots, x_n)$が得られる確率密度を$f(x)$と表すとき、事象$E = \{x|f(x) \leq f(x_{obs})\}$の$H_0$のもとでの確率を$P$-値と呼びます。ただし、正規分布のように対称な場合にはどの定義でも同じ値となりますが、非対称な場合や離散分布の場合には定義次第で異なる例もあり、必ずしも確立した概念とは言えません。

2. 仮説検定

最も広く使われている検定の例として、まず、母集団の分布が正規分布である場合の仮説検定について説明します。また、正規母集団以外のものに関しても中心極限定理から正規母集団についての検定を漸近的(nが大きい場合の近似)に使うことができます。

① 母集団の平均に関する仮説の検定

正規分布の平均に関する問題は、分散が既知の場合は正規分布による検定を行い、分散

が未知の場合にはt分布による検定を行います。

➢ 正規分布の平均に関する検定(z検定)

✧ 分散既知、両側対立仮説の場合

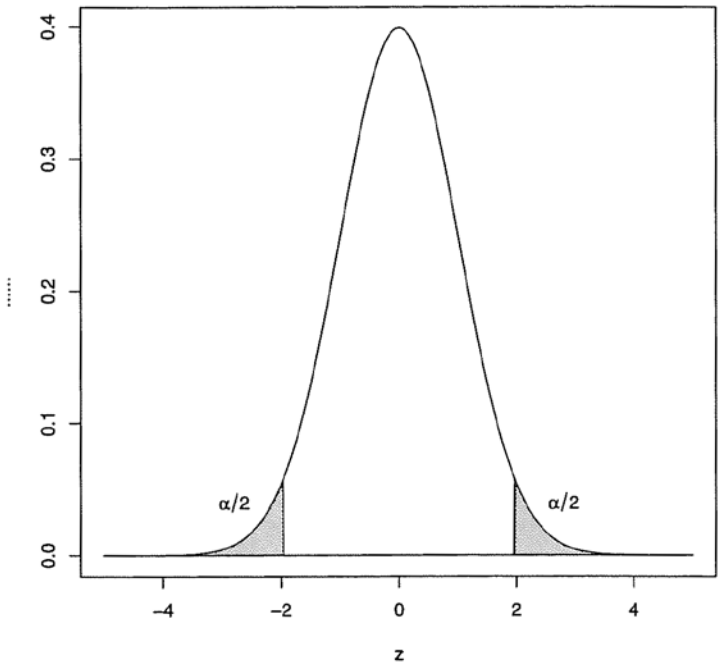
以下、観測値$x_1, \cdots, x_n$が独立に同じ正規分布に従うことを$x_1, \cdots, x_n \sim N(\mu, \sigma^2)$と表します。

帰無仮説は$H_0: \mu = \mu_0$、対立仮説は$H_1: \mu = \mu_1 (\mu_1 \neq \mu_0)$、検定統計量として標本平均$\bar{x}$を利用します。仮説$H_0$のもとで$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$となるので、これを$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$と標準化します。有意水準を5%$(\alpha = 0.05)$とすると、標準正規分布では$\mathrm{P_r}\{|z| \geq 1.96\} = 0.05$なので、観測された$\bar{x}$から求めた$z$の絶対値が1.96より大きいとき、仮説$H_0$を棄却します。有意水準が1%$(\alpha = 0.01)$であれば、$\mathrm{P_r}\{|z| \geq 2.58\} = 0.01$を利用します。以上をまとめると、$\bar{x}$に関する次の棄却域が得られます。

$$\text{有意水準5\%のとき} \quad |\bar{x} - \mu_0| \geq 1.96\frac{\sigma}{\sqrt{n}}$$

$$\text{有意水準1\%のとき} \quad |\bar{x} - \mu_0| \geq 2.58\frac{\sigma}{\sqrt{n}}$$

形式的に判断するには、初めから$\bar{x}$を標準化した$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$を検定統計量として、その値を1.96又は2.58と比較するのが簡単です。zを用いて棄却域を表示すると下の図のようになります。この図の密度関数は仮説H_0が正しいときのzの確率分布、すなわち標準正規分布であり、影をつけた部分の面積が有意水準αに対応します。なお、両側検定仮説のP-値は、観測されたz_{obs}より外側の確率$P\{|z| \geq |z_{obs}|\}$と定義されます。

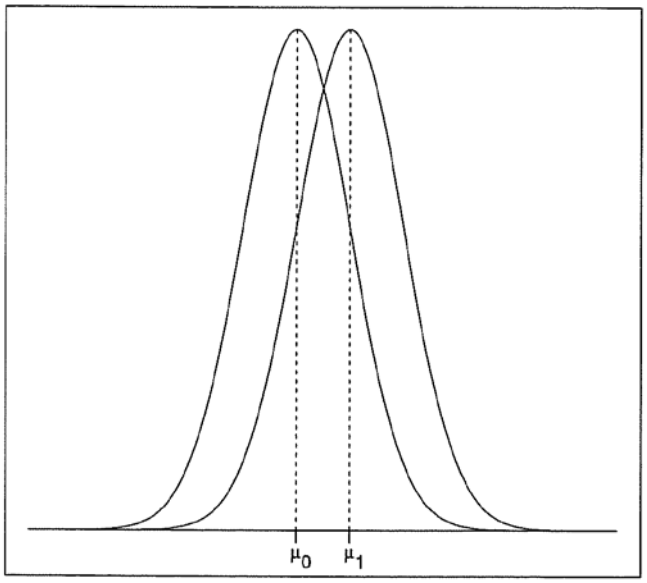


両側検定の棄却域

✧ 分散既知、片側対立仮説の場合

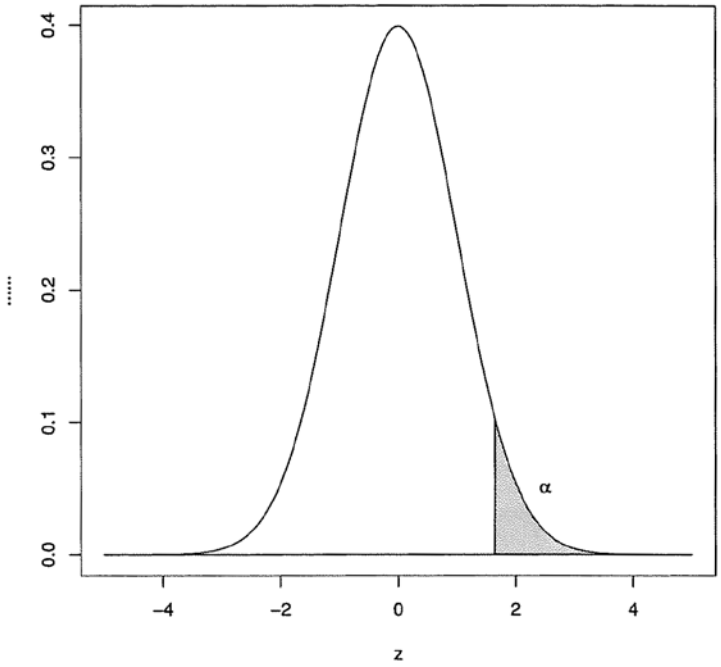
片側対立仮説の場合も、両側検定と同様の手順が利用できます。帰無仮説は$H_0: \mu = \mu_0$であり、これは対立仮説にはよりません。片側対立仮説はH_1としては$\mu > \mu_0$の場合と$\mu < \mu_0$の場合がありますが、そのどちらかは具体的な問題によって決定されます。どちらの場合も考え方は同じなので、ここでは$\mu > \mu_0$の場合について解説します。より正確には片側対立仮説は$H_1: \mu = \mu_1,\ \mu > \mu_0$と表現されます。

下の図に示すように、検定統計量の分布はH_0の場合は$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$であり、$H_1$の場合は$\bar{x} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$となるので、$H_0$の場合より$H_1$の場合に$\bar{x}$が大きな値をとる可能性が高いです。実際、最適な棄却域は$\bar{x}$が大きな値をとる確率αとなるように定められます。



2つの仮説

$\bar{x}$ではなくzを用いて棄却域を表現すると下の図のようになります。



片側検定の棄却域

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_0 \quad \text{有意水準 5\%のときは} z_0 = 1.645$$

なお、$\mathrm{P_r}\{z \geq 2.33\} = 0.01$なので、有意水準を1%としたときの棄却域は$\bar{x} \geq \mu_0 + 2.33\left(\frac{\sigma}{\sqrt{n}}\right)$と定められます。

✧ 正規分布の場合の検出力

有意性検定の理論は、帰無仮説について下す決定が誤っていることを前提としています。帰無仮説H_0が正しければとうてい出そうもない検定量の値が出れば、帰無仮説を誤りとして棄却します。したがって、逆に言えば、帰無仮説が正しくても、出そうもない棄却域の値がたまたま出てしまい、帰無仮説を棄却することがありえます。これは第1種過誤ですが、この確率αはそもそもの仮説検定の考え方そのものであり、初めからある程度小さく抑えられています。

一方で帰無仮説が誤りであるにもかかわらず、たまたま統計量の値が棄却域に入らなかったために、帰無仮説を棄却しない誤りが生じる可能性もあります。これが第2種過誤です。この確率βは対立仮説の各値が正しいという条件のもとに棄却域に入らない確率を求めればよいことになります。αとβは検定方法、つまり、棄却域の取り方によりますが、棄却域を狭くすればαは小さくなりますが、βは大きくなります。

有意性検定では、αを先に固定しています。その条件でβをなるべく小さくする、つまり$1-\beta$をなるべく大きくすることが求められます。この$1-\beta$の確率を検出力といいます。

$$\mathrm{P_r}(\mathrm{reject}|H_1) = \mathrm{P_r}\left(\bar{x} \geq \mu_0 + z_0\frac{\sigma}{\sqrt{n}}\middle|\mu_1\right) = \mathrm{P_r}\left(z \geq \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + z_0\right)$$

➢ 正規分布の平均に関する仮説の検定(t検定)

以上では分散σ^2が既知の場合を説明しましたが、通常、分散は未知であることが多いです。この場合には与えられたデータから標本平均$\bar{x}$とともに、標本不偏分散$s^2 = \sum \frac{(x_i-\bar{x})^2}{(n-1)}$を計算します。分散としては$n$で割る$\sigma^2 = \sum \frac{(x_i-\bar{x})^2}{n}$が与えられる場合もありますが、$n$がある程度大きければ、どちらの分散を用いても結果に大きな差はありません。例えば$n = 100$ならnで割るのと$n-1$で割るのでは1%程度の差しかありませんが、実際の観測値の誤差は通常はもっと大きいからです。ここではnが大きいときと、そうでないときで分けて考えます。

✧ nが大きいとき

この場合は、大数の法則から$s \fallingdotseq \sigma$と考えてかまいません。同時に中心極限定理から、母集団の分布が正規分布とは多少異なっていても、標本平均は近似的に正規分布に従うことも保証されています。したがって、分散が既知の場合の正規分布による検定が利用できることになり、両側検定なら有意水準5%の棄却域は$|\bar{x}-\mu_0| \geq 1.96\frac{\sigma}{\sqrt{n}}$で与えられます。

経済、社会、心理などの分野では標本の大きさnが数百から数千以上となっているので、ほとんどの場合このような正規近似が利用できます。

✧ nが小さいとき

大数の法則は利用できず$s \fallingdotseq \sigma$が成り立たないので、$\bar{x}$を標準化した$z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$が標準正規分布に従うことが保証されません。この場合には、もし「観測値が正規分布に従う」ならば、t分布が利用できます。定義から、確率変数$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$は自由度$n-1$の$t$分布に従います。このようにして計算される$t$の値を$t$-値と呼び、これに基づく検定を$t$検定と呼びます。両側検定の棄却域は$|t| \geq t_0$という形で与えられますが、$t_0$の値は自由度$\nu = n-1$に依存して決まる点で正規分布を用いた検定とは異なります。

t検定の形は次の対比で覚えることができます。

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \qquad t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

推定のところでも説明しましたが、t分布の5%点、2.5%点の値はnが大きくなると正規分布の場合とあまり差がないことが分かります。自由度が240より大きければ、正規分布とほとんど同じになります。

➢ 二つの正規分布の平均の差に関する仮説の検定

ここでは、二つの正規分布の平均に差があるかどうかを検討する問題を扱います。例えば、ラットにたんぱく質を含んだ餌を与えて体重がどの程度増加するかを調べる実験を行い、ビーフ低水準とビーフ高水準の2種類の餌の効果が異なるかについて考えてみましょう。

✧ 分散が等しい場合を想定する理由

分散が等しい場合には仮説の意味は明確ですが、分散が異なる場合には、平均を比較することにあまり意味がない状況もあることに注意が必要です。たとえば、ラットの発育の個体差があまりに大きい場合には、2種の餌の効果を平均で比べることに意味があるでしょうか。

そのため、ここでは、比較的分散が近い場合を想定します。二つの母集団からの無作為標本を$x_1,\cdots,x_m,y_1,\cdots,y_n$として、標本平均$\bar{x}=\sum\frac{x_i}{m}$と$\bar{y}=\sum\frac{y_i}{n}$を利用します。まず、分散が既知の場合について説明します。

✧ 分散が既知の場合

正規分布に従う確率変数$\bar{x}\sim N\left(\mu_1,\frac{\sigma^2}{m}\right)$と$\bar{y}\sim N\left(\mu_2,\frac{\sigma^2}{n}\right)$の差$d=\bar{x}-\bar{y}$も正規分布に従うことを利用します。定義から$d$の平均は $\delta=\mu_1-\mu_2$ 、分散は$V(d)=\frac{\sigma^2}{m}+\frac{\sigma^2}{n}=\sigma^2(\frac{1}{m}+\frac{1}{n})$となります。ここで$\bar{x}$と$\bar{y}$は独立なので、共分散はゼロで$V(d)$には現れません。

仮説H_0:「体重を増加させる効果はない」のもとでは$\delta=0$となるので、有意水準5%の棄却域は次の式で与えられます。

$$\textbf{両側対立仮説の場合}\quad |d|\geq 1.96\sqrt{\frac{1}{m}+\frac{1}{n}}\sigma \quad H_1:\delta\neq 0$$

なお、分散が既知の場合には、それぞれの分散が等しくなくてもdを利用して検定を行うことができます。すなわちxの分散をσ_1^2、yの分散をσ_2^2とすると、標本平均の分布は$\bar{x}\sim N(\mu_1,\sigma^2/m)$及び$\bar{y}\sim N(\mu_2,\sigma^2/n)$となり、これらは独立なので、$d$の分散は$\sigma_1^2/m+\sigma_2^2/n$となります。したがって、両側対立仮説の棄却域(有意水準5%)は次のように定められます。

$$|d|\geq 1.96\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}$$

✧ 分散が未知で等しい場合

正規分布に従う確率変数$\bar{x}\sim N\left(\mu_1,\frac{\sigma^2}{m}\right)$と$\bar{y}\sim N\left(\mu_2,\frac{\sigma^2}{n}\right)$の差$d=\bar{x}-\bar{y}$を標準化した$z$は標準正規分布に従い、分散の推定値$s^2$を$\sigma^2$に代用した$t$は自由度$m+n-2$の$t$分布に従います。

$$z=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sigma}\sim N(0,1)\quad 及び\quad t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{1}{m}+\frac{1}{n}}s}\sim t_{m+n-2}$$

ここで分散s^2の推定にはx,yのそれぞれから求められる不偏分散を合成します。$s_1^2=\sum\frac{(x_i-\bar{x})^2}{(m-1)}$, $s_2^2=\sum\frac{(y_i-\bar{y})^2}{(n-1)}$として、

$$s^2=\frac{(m-1)s_1^2+(n-1)s_2^2}{(m-1)+(n-1)}=\frac{\sum(x_i-\bar{x})^2+\sum(y_i-y)^2}{m+n-2}$$

が、分散の推定量となります。

✧ 分散が未知で等しくない場合

現実の観測値では分散の推定値が一致することはありません。もし推定値の差が大きく、母集団の分散σ_1^2とσ_2^2が等しいと想定することに無理がある場合には、次のようにtを定義します。

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

このtは帰無仮説のもとで「近似的に」次の自由度fをもつt分布に従うことが知られています。

$$f = \frac{(g_1+g_2)^2}{\frac{g_1^2}{(n_1-1)}+\frac{g_2^2}{(n_2-1)}} \qquad \text{ただし } g_i = \frac{s_i^2}{n_i}$$

このtを用いる方法はWelch(ウェルチ)の検定と呼ばれます。
なお、分散が大きく異なる二つの標本で平均を比較することそのものに意味があるかどうかを確認することも重要です。

➢ 対応がある標本の仮説の検定

たとえば、あるダイエット処方の効果を確かめるために、無作為に選んだ$n = 10$人の被験者に処方を適用したときの、処方前と1か月後の体重の関係から、この処方に効果があるかどうかを考えてみましょう。

この場合は、二つの測定値xとyの平均に差があるかどうかを分析するために、異なる母集団から得られた標本x、yを扱うのではなく、処方前後に同じ人の体重を測定しているというように、各標本が対$(x_i, y_i)(i = 1, \dots, n)$として与えられています。これを対比較あるいは対応のある場合と呼びますが、対を作ることによって処理効果を個体間の差から分離して判断できる点で優れた方法とされています。同一の固体の観測値(x, y)には正の相関があるとすれば、その差$d = x - y$の分散は$\sigma_d^2 = V(d) = V(x) + V(y) - 2\mathrm{cov}(x, y)$となり、独立な2標本の場合の分散$V(d) = V(x) + V(y)$より小さくなることからも、対比較の有効性が分かります。この問題では$d_i = x_i - y_i$を観測値と考えると、平均の差に関する検定の問題は、1標本の場合に帰着されます。すなわちx、yの期待値をそれぞれμ_1、μ_2とするとき、仮説$H_0: \mu_1 = \mu_2$のもとで、dは平均0、分散σ_d^2の正規分布に従うので、母集団が一つの場合のt検定を適用すればよいことになります。

$\bar{d} = \sum_{i=1}^{n} \frac{d_i}{n}$, $s_d^2 = \sum_{i=1}^{n} \frac{(d_i-\bar{d})^2}{(n-1)}$を用いて$t = \frac{\sqrt{n}\bar{d}}{s_d}$とすると、これは自由度$n-1$の$t$分布に従います。

② 母集団の比率に関する仮説の検定

例えば全国の有権者から2400人を無作為に選んで面接調査を実施したところ、1250人がある政策に賛成と回答したとき、この結果から、この政策は有権者全体の過半数が支持しているといえるかどうか考えてみましょう。このように比率に関する仮説を検定する場面も多く発生します。母集団から無作為にn人抽出して調査した場合、支持者数xの分布は二項分布$B(n, p)$と考えることができます。nが大きい場合、これまで説明したとおり、二項分布は正規分布で近似できるので、比率に関する検定にも正規分布に関する議論が応

用できます。厳密には超幾何分布ですが、母集団の人数Nが大きいときには二項分布と考えてよいことも既に説明しました。

➢ 仮説$H_0: p = p_0$の検定

この仮説の下では次のzは標準正規分布に従います。

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \sqrt{n}\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

有意水準を5%とする検定の棄却域は両側検定仮説なら$|z| \geq 1.96$、片側検定仮説なら$|z| \geq 1.645$(又は$z \leq -1.645$)となります。

➢ 二つの母集団の比率に関する仮説の検定

では、無作為に抽出した若年男性480人の労働時間に関する調査結果を過去1月以内の労働時間の長さで下の表のように分類したところ、この結果から、年代によって労働時間の差があると考えられるでしょうか。

	長時間	短時間	計(人)
25～29歳	116	76	192
30～34歳	244	44	288
計(人)	360	120	480

この表から、若年男性が長時間働く比率は25～29歳で$\frac{116}{192} = 0.6041$、30～34で$\frac{244}{288} = 0.8472$となっています。この比率にはかなりの差がありますが、各年齢層に属する192人及び288人を調査しただけで母集団の比率が異なると言えるでしょうか。

この例では、各年齢層で長時間働く人の数は、母集団比率pを母数とする二項分布$B(n,p)$に従います。二つの母集団から得られた観測値$x \sim B(n_i, p_i)(i = 1,2)$に基づいて、仮説$H_0: p_1 = p_2$を検定する手順は次の通りです。$n_1$,$n_2$がいずれも大きいとき、それぞれの比率の推定量$\hat{p}_i = \frac{x_i}{n_i}$は近似的に正規分布$N\left(p_i, \frac{p_i(1-p_i)}{n_i}\right)$に従うので、それらの差も正規分布に従います。

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \quad \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

通常は次のzが近似的に標準正規分布に従うことを利用して仮説の検定を行います。

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

この式の分母は、未知のp_1, p_2を$\hat{p}_1, \hat{p}_2$で置き換えたものです。この手順は大きなn_1,n_2に対しては大数の法則によって正当化されます。

仮説$H_0: p_1 = p_2$の下では、上の式の分子は$\hat{p}_1 - \hat{p}_2$です。この形の検定は、より一般的な仮説$H_0: p_1 = p_2 + \delta$に対しても分子を$(\hat{p}_1 - \hat{p}_2) - \delta$に置き換えるだけで利用できます。

③ 分割表に関する適合度と独立性の検定(χ^2検定)

上の表は若年層の年代別に労働時間の長さで分類して集計したものです。このようにいくつかの基準で分類された表を分割表と呼びます。ここでは分類基準が一つの場合と二つの場合に分けて、分割表の適合度と独立性の検定について解説します。

➢ $1\times k$分割表

ある遺伝子の組合せでは3種類の組合せAA、Aa、aaの発生頻度は1:2:1とされています。これについて実験を行い、観測度数O_iと期待度数E_iが次の表で与えられたとします。

	AA	Aa	aa	計
O_i	35	67	30	132
E_i	33	66	33	132

ここでは、$n = 132$が度数合計、AA、Aa、aaの順番に番号を振って、期待度数は$E_1 = E_3 = \left(\frac{1}{4}\right)n,\ E_2 = \left(\frac{1}{2}\right)n$と求めています。

一般に、ある変数Aを因子としてk個の水準(カテゴリ)に分類された$1\times k$分割表は次の形で与えられます。

	A_1	A_2	…	A_k	計
観測度数	O_1	O_2	…	O_k	n
期待度数	E_1	E_2	…	E_k	n

ここで期待値E_iは仮説$H_0: \mathrm{P_r}(A_i) = p_i (p_1 + \cdots + p_k = 1)$のもとで$E_i = np_i$と計算されます。期待度数の計算に利用した仮説$H_0$のもとで、次の$\chi^2$は自由度$(k-1)$の$\chi^2$分布に従うことが知られています。ただし、これは$n$が大きいときの近似的な結論です。

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

すなわち、帰無仮説を$H_0: \mathrm{P_r}(A_i) = p_i (p_1 + \cdots + p_k = 1)$とするとき、$\chi^2$分布表からパーセント点をよみ、$\chi^2 > \chi^2_\alpha(k-1)$であれば、「観測度数は理論確率分布$p_i$に適合している」という仮説$H_0$は有意水準$\alpha$で棄却されます。

自由度が水準の数kより一つ小さくなるのは、度数の合計が一定という制約があるためです。

遺伝の例では$\chi^2 = \frac{(35-33)^2}{33} + \frac{(67-66)^2}{66} + \frac{(30-33)^2}{33} = \frac{27}{66} \fallingdotseq 0.41$となります。自由度2で$\chi^2_{0.05} = 5.991$であるので、仮説は棄却されません。また、仮説が正しくない場合はχ^2は大きな値をとる傾向があるので、この場合のP-値を計算すると、$\mathrm{P_r}(\chi^2 \geqq 0.41) \fallingdotseq 0.815$となり、仮説は棄却されません。

このように仮定された理論上の確率分布について、標本から求められた母数が適合するか否かを検定するのが適合度のχ^2検定です。

➢ 2元分割表

一般に二つの因子を用いてr行c列に分類した$r \times c$分割表は次の形で与えられます。

	B_1	B_2	…	B_c	行和
A_1	f_{11}	f_{12}	…	f_{1c}	$f_{1.}$
A_2	f_{21}	f_{21}	…	f_{2c}	$f_{2.}$
…			…		…
A_r	f_{r1}	f_{r1}	…	f_{rc}	$f_{r.}$
列和	$f_{.1}$	$f_{.2}$	…	$f_{.c}$	$f_{..}$

ここでf_{ij}(観測度数O_{ij})は(i,j)セルの観測度数、$f_{i.} = \sum_j f_{ij}$はi行の合計(行和)、$f_{.j} = \sum_i f_{ij}$はj列の合計(列和)、$f_{..} = n$は標本の大きさです。

この場合、二つの基準A、Bが独立であるという仮説は、$H_0: \mathrm{P_r}(A_i \cap B_j) = \mathrm{P_r}(A_i)\mathrm{P_r}(B_j)$で表されます。$A_i, B_j$の確率の推定値は、それぞれ$\frac{f_{i.}}{n}$、$\frac{f_{.j}}{n}$なので、$(A_i,\ B_j)$の期待度数は$E_{ij} = n(\frac{f_{i.}}{n})(\frac{f_{.j}}{n}) = \frac{f_{i.}f_{.j}}{n}$です。

仮説H_0のもとで、次のχ^2は近似的に自由度$(r-1)(c-1)$のχ^2分布に従うことが知られています。行の和と列の和が固定されていることから、自由度はそれぞれから1を引いて掛け合わせた$(r-1)(c-1)$となります。

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

労働時間の長さの分割表について、期待度数を計算すると次の表のようになります。

	観測度数O_{ij}			期待度数E_{ij}		
	長時間	短時間	計(人)	長時間	短時間	計(人)
25～29歳	116	76	192	144	48	192
30～34歳	244	44	288	216	72	288
計(人)	360	120	480	360	120	480

これから

$$\chi^2 = \frac{(116-144)^2}{144} + \frac{(76-48)^2}{48} + \frac{(244-216)^2}{216} + \frac{(44-72)^2}{72}$$
$$= \frac{(-28)^2}{144} + \frac{(28)^2}{48} + \frac{(-28)^2}{216} + \frac{(28)^2}{72} \fallingdotseq 36.30$$

が得られます。自由度1のχ^2分布では、上側5%点は3.84、1%点は6.63なので、いずれも有意な結果であり、帰無仮説H_0:「労働時間の比率と年代は独立である(つまり、年代によらず労働時間比率は一定である。)」は棄却され、20代と30代では、労働時間の比率に明確な違いがあると結論できます。なお、P-値は$\mathrm{P_r}(x^2 \geqq 36.30) = 1.69\times10^{-9}$とほとんどゼロになります。

これを独立性のχ^2検定といいます。

④ 回帰係数に関する検定

単回帰の論理的なモデルは次のように与えられます。

$$y_i = \alpha + \beta x_i + \epsilon_i,\ \ \epsilon_i \sim N(0,\sigma^2)\ \ (i = 1,\cdots,n)$$

ここで誤差項ϵ_iは互いに独立と仮定されます。

回帰モデルでは、係数βについて$H_0: \beta = \beta_0$という帰無仮説の検定がよく利用されます。経済分析での消費関数を例にとると、βは経済学では限界消費性向と呼ばれる重要な指標です。

回帰分析では、変数xは変数yの変化に「影響を与えない」という仮説$H_0: \beta = 0$の検定が広く用いられます。一般の仮説$H_0: \beta = \beta_0$を検定するにはt-値$t = \frac{\hat{\beta}-\beta_0}{s/\sqrt{T_{xx}}}$を求めて、

t分布のパーセント点と比較すれば分かります。

たとえば、ダイエットにおける減量の大きさは体重に依存するでしょうか。このことを確かめるために体重の変化を従属変数、処方前の体重を説明変数として回帰分析を行ったところ、次の結果が得られたとします。

概要

回帰統計	
重相関 R	0.354498
重決定 R2	0.125669
補正 R2	0.016377
標準誤差	2.984795
観測数	10

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	10.24400735	10.24401	1.149849412	0.31486
残差	8	71.27199265	8.908999		
合計	9	81.516			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	−11.9125	8.902518376	−1.3381	0.217648542	−32.4417	8.616788	−32.4417	8.616788
X 値 1	0.142124	0.13253991	1.07231	0.314859535	−0.16351	0.447761	−0.16351	0.447761

この結果を見ると、xの回帰係数は$\hat{\beta} = 0.1421$とプラスになっていて、体重xが重いほど体重の増加yは大きいので、ダイエットの効果が小さいことを示しています。たとえば体重60kgの人の体重増加は$\hat{y} = \hat{\alpha} + \hat{\beta}\times 60 = -11.9 + 0.142\times 60 = -3.4\text{kg}$であるのに対して、体重$80\text{kg}$の人は$-11.9 + 0.142\times 80 = -0.5\text{kg}$とダイエットの効果が小さくなり、予想に反した結果となっています。

ここでt-値(仮説$H_0: \beta = 0$に対する$t = \frac{\hat{\beta}}{\text{se}(\hat{\beta})}$)を見ると、$1.07$と小さく、$P$-値も$0.315$となっています。すなわち、有意水準$5\%$で有意でないだけではなく、有意水準を$30\%$としても棄却されません。つまり、この観測結果からは、「減量の大きさは体重と無関係である」との仮説を否定する根拠は得られず、体重の説明変数としての説明力は弱いことが分かります。

このように回帰係数のt-値は、通常「係数がゼロ」すなわち「説明変数が応答変数に影響を与えない」という帰無仮説に対して計算されます。

練習問題 （解答は P.70 です）

問 1 ある製品の特性xは、技術的に管理される平均μと、一定の標準偏差$\sigma = 10$をもつ正規分布にしたがうことが知られている。ある日に製造された製品から無作為に$n = 16$個を抽出して検査したところ、その平均は$\bar{x} = 207$であった。

(1) この日に製造された製品が標準的な規格$\mu = \mu_0 = 200$を満たしているかどうか、有意水準5%で検定せよ。

(2) この日の製品全体の平均μについて信頼係数95%の信頼区間を求めよ。

(3) σは未知で、標本から計算された不偏分散が$s^2 = 100$とする。このとき、仮説$H_0: \mu = 200$を有意水準5%で検定せよ。

Ⅶ 回帰分析

統計分析手法の中で最も広く利用されているのは回帰分析です。推定や検定の手法もこれに用いられています。現在では汎用ソフトウエアによって簡単に計算することができますが、その結果を読むためにも、回帰分析の基礎知識は必要不可欠です。

1. 単回帰モデル

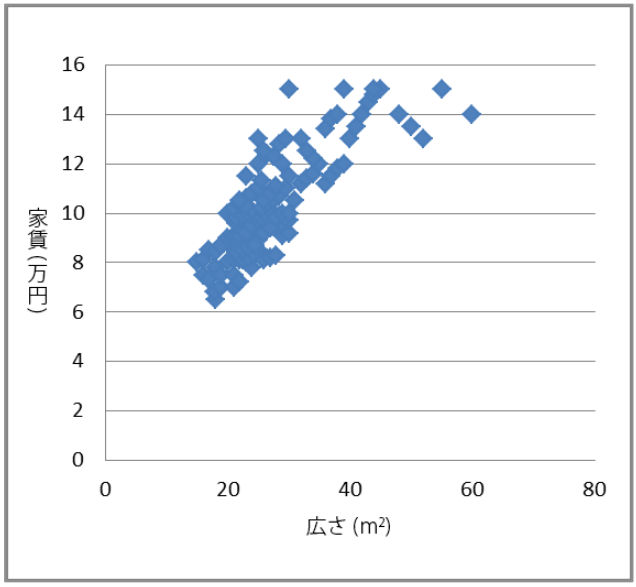
回帰分析とは、2変数のx,yのデータがあるとき、回帰方程式と呼ばれる説明の関係を定量的に表す式を求めることを目的としています。

相関関係は、「家賃が高いから広い」のか「広いから家賃が高い」のかについての情報は提供してくれません。ただ、二つの変数の間に関係があることを示しているだけです。ここで、「部屋が広いと家賃が高くなるのではないか」という一方向の関係を仮定したとします。この仮定が正しそうな場合には、相関関係を想定して分析した場合より、多くの知見が得られることになります。そこで、この仮説

「部屋の大きさxが分かると、家賃yがおおよそ説明(予測)できる」

というモデル化を考え、その妥当性について検討してみましょう。このように原因から結果への関係を因果関係といいます。相関関係では二つの変数間には方向性はありませんが、因果関係では方向性があります。

部屋の大きさと家賃についての散布図を描いたところ、下の図のようになりました。



因果関係を考える場合には、x軸に原因となる変数をとり、y軸には結果となる変数をとった散布図を作成します。

散布図から分かるように、同じ部屋の大きさでも家賃に差があり、必ずしも部屋の大きさ

のみでは家賃を説明できません。そこで、家賃を部屋の大きさで説明できる部分と説明できない部分に分けて考えることにします。つまり、原因から決まる予測値に誤差が加わってデータとなるという加法的な分解モデルを考えます。この原因で説明できる部分をxの関数$f(x)$として表し、説明できない部分を誤差εとして表すと、

$$y_i = f(x) + \varepsilon_i \qquad (i = 1,2,\cdots,n)$$

となります。このモデルで説明する変数xを説明変数(独立変数、予測変数、共変量)と言い、説明される変数yを被説明変数(従属変数、目的変数、応答変数、基準変数)と言います。最も単純な近似として、

$$\hat{y} = f(x) = \alpha + \beta \times x$$

を採用すると、

$$y_i = \hat{y} + \varepsilon_i = \alpha + \beta x_i + \varepsilon_i \qquad (i = 1,2,\cdots,n)$$

となります。ここで、$\hat{y}_i$を予測値と呼び、ε_iを誤差と呼びます。この式が単回帰モデルで、傾きβを回帰係数といいます。

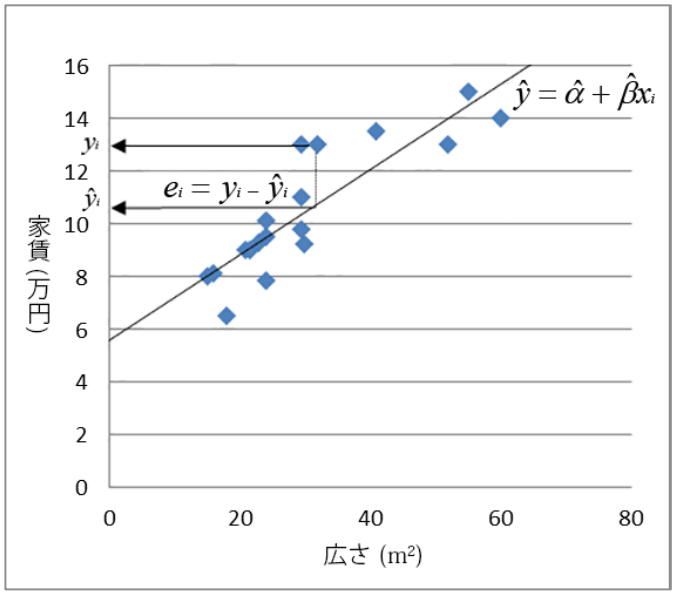
実際には、切片と傾きの組$(\alpha,\ \beta)$はデータに依存して決まりますが、回帰モデルにおける$(\alpha,\ \beta)$と区別するために、あるデータセットに対して求められる値を$(\hat{\alpha},\ \hat{\beta})$と表記し、誤差$\varepsilon_i$も$e_i$と表記して残差と呼びます。

① 最小二乗法

この$(\hat{\alpha},\ \hat{\beta})$を求める方法として、最小二乗法による当てはめを考えましょう。最小二乗法とは、x_iから予想される$\hat{y}$の値$(\hat{\alpha} + \hat{\beta}x_i)$と現実の値$y_i$の差(残差$e_i$)の二乗和$S(\hat{\alpha},\hat{\beta})$

$$S(\hat{\alpha},\ \hat{\beta}) = \sum_{i=1}^{n} \{y_i - (\hat{\alpha} + \hat{\beta}x_i)\}^2$$

を最小にする$(\hat{\alpha},\ \hat{\beta})$を求める方法です。



$S(\hat{\alpha}, \hat{\beta})$は、$(\hat{\alpha},\ \hat{\beta})$の2次関数であり、最小値が存在します。そこで、最小を求めるために、それぞれ$\hat{\alpha}$と$\hat{\beta}$で偏微分して0とおくと、

$$\begin{cases} n\hat{\alpha} + \left(\sum_{i=1}^{n} x_i\right)\hat{\beta} = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\hat{\alpha} + \left(\sum_{i=1}^{n} {x_i}^2\right)\hat{\beta} = \sum_{i=1}^{n} x_i y_i \end{cases}$$

となります。これを二元連立一次方程式として解くと、

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} = r_{xy}\frac{\sqrt{s_{yy}}}{\sqrt{s_{xx}}} = r_{xy}\frac{s_y}{s_x}$$

となります。得られた$\hat{\alpha}$，$\hat{\beta}$による1次式を回帰方程式、あるいは回帰直線、$\hat{\beta}$はその傾きで偏回帰係数と呼ばれます。最小二乗法の性質として、以下のことを言うことができます。

- 予測値$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$の平均は観測値の平均と等しくなる

$$\bar{\hat{y}} = \bar{y}$$

- 残差$e_i = y_i - \hat{y}_i$の平均は0となる

$$\bar{e} = 0$$

- 予測値の偏差と残差の偏差の積和は0となる

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(e_i - \bar{e}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i = 0$$

- 観測値と予測値の相関係数は、(予測値の標準偏差)/(観測値の標準偏差)と等しくなる

$$s_{y\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i + e_i - \bar{y})(\hat{y}_i - \bar{y})}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y})}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1} = s_{\hat{y}\hat{y}}$$

となるので、データyと予測値$\hat{y}$の相関係数は、

$$r_{y\hat{y}} = \frac{s_{y\hat{y}}}{\sqrt{s_{yy}}\sqrt{s_{\hat{y}\hat{y}}}} = \frac{\sqrt{s_{\hat{y}\hat{y}}}}{\sqrt{s_{yy}}} = \frac{s_{\hat{y}}}{s_y}$$

となります。このyと$\hat{y}$の相関係数を重相関係数と呼び、通常の相関係数と区別して、Rを用います。

② 決定係数

被説明変数の変動がどの程度説明変数で説明できたかを基準にして、被説明変数yについての平均からの変動(偏差の二乗和)を求めると、

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i + e_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}{e_i}^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i$$

となりますが、$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i$より、

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}{e_i}^2$$

となります。これは、データの変動がモデルで説明された変動(予測変動)とモデルで説明できなかった変動(残差変動)とに分解されることを表しています。変動の大きさを$\sum_{i=1}^{n}(y_i - \bar{y})^2$で割って基準化すると

$$1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{n}{e_i}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$1 = \frac{s_{yy}}{s_{yy}} = \frac{s_{\hat{y}\hat{y}}}{s_{yy}} + \frac{s_{ee}}{s_{yy}} = R^2 + (1 - R^2)$$

となります。したがって、R^2が1に近いほど、yは$\hat{y}$に近づき、R^2が1なら正確に$y = \hat{\alpha} + \hat{\beta}x$が成立し、$y$は$x$から完全に決定されます。つまり、$R^2$は回帰式がどの程度、データを説明しているかを表しているので、これを決定係数と呼びます。また、$\hat{y} = \hat{\alpha} + \hat{\beta}x$であるので、

$$R^2 = {r^2}_{y\hat{y}} = {r^2}_{yx}$$

となり、最小二乗法を用いた単回帰分析では、決定定数は相関係数の二乗と等しくなります。

③ 自由度調整済み決定係数

モデルがどの程度、従属変数(被説明変数)を説明しているかを決定係数R^2で見ることができることが分かりましたが、決定係数の大きさR^2はデータの大きさnにも依存します。つまり、データの大きさnが小さければ、決定係数R^2は大きくなり、たとえば、$n = 2$で

あれば、どのようなデータに対しても$R^2 = 1.0$となります。これは、決定係数がデータの大きさとは無関係な変動に基づいて定義されているためです。そこで、「データ1個当たり」の考え方に基づく決定係数を考えます。

n個の予測値$\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{n-1}, \hat{y}_n$は

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

として求められていました。これは、$\hat{\beta}$を求めれば、$\hat{\alpha}$も$\hat{y}_i (i = 1,2, \cdots, n)$も全て決まることを意味しています。したがって、自由に決めることができるのは$\hat{\beta}$のみです。このように見かけ上のnではなく実態としての個数を自由度といいます。この自由度を用いて求められるのが自由度調整済み決定係数R^{*2}です。

今、説明変数の個数をp個(単回帰モデルでは$p = 1$)用いて分析して得られた残差について、残差分散を計算します。データの変動の自由度が$n - 1$で、予測値の変動の自由度がpです。三つの変動の自由度については

$$n - 1 = \text{予測モデルの自由度} + \text{残差の自由度}$$

より、残差の自由度 $= (n - 1) - p$となります。最小二乗解では、予測値と残差の積和は0となることが分かっているので、残差の分散は、

$$s_{ee} = s^2 = \frac{\sum {e_i}^2}{(n - 1) - p}$$

となります。データの分散を$s_{yy} = s_y^2$とすると、自由度調整済み決定係数 R^{*2}は、

$$R^{*2} = 1 - \frac{s_{ee}}{s_{yy}}$$

として定義され、$R^2 > R^{*2}$となります。

先に散布図でみた家賃データについて広さを説明変数とする回帰モデルを計算してみると、

$$\hat{y}_i = 4.4547 + 0.2099x_i$$

$$R^2 = 0.6595$$

$$R^{*2} = 0.6571$$

となりました。この式は、切片が44,547円、広さが1㎡広くなると2,099円高くなるという

関係を説明しており、この回帰式でデータの変動のうち、約65%を説明できることを意味しています。

2. 重回帰モデル

次に複数の説明変数を用いる方法について説明します。

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (i = 1,2,\cdots,n)$$

このようにyがxの線形関数である場合を線形回帰、線形モデルと言います。このモデルにおいて、説明変数として、部屋の大きさ(㎡)だけではなく、かかる時間(徒歩)(分)を用いれば、家賃をより適切に説明できる可能性があります。

$$\text{家賃}\left(\text{円}\right) = \beta_0 + \beta_1 \times \text{部屋の大きさ}\left(\text{㎡}\right) + \beta_2 \times \text{徒歩}\left(\text{分}\right) + \text{誤差}$$

このように応答変数を説明するために、複数の説明変数を用いるモデルを重回帰モデルといいます。

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

ここで、$(\beta_0, \beta_1, \ldots, \beta_p)$のそれぞれは、それ以外の説明変数の値を固定し、対応する説明変数の値を1単位増加させたときに、yがどの程度増加するかを示しているとも解釈できるので、これを偏回帰係数といいます。

しかし、実際には説明変数間にも相関関係があり、ある説明変数の値が1単位変化すると他の説明変数の値も変化するので、偏回帰係数の大小については、回帰式に含めた説明変数の組と合わせて検討しなければなりません。

この偏回帰係数の最小二乗推定値は、$b = \left(b_1, \ldots, b_p\right)^t$として、

$$S_x b = s_{yx}$$

となる連立方程式の解として求められます。ここでS_xは説明変数間の分散共分散行列で大きさ$p \times p$であり、s_{yx}は、応答変数と説明変数との共分散を要素とするp次元ベクトルです。定数項b_0は

$$b_0 = \bar{y} - b_1 \bar{x}_1 - \cdots - b_p \bar{x}_p$$

と求めます。

分散共分散行列S_xの逆行列S_x^{-1}を求めることができれば、偏回帰係数の最小二乗解$\hat{b}$は

$$\hat{b} = S_x^{-1} s_{yx}$$

として求まります。通常はこの逆行列は対角行列はとりません。

① モデルの候補を挙げる

家賃を応答変数として、部屋の大きさと徒歩時間を説明変数とした重回帰モデルを用いて分析してみましょう。単回帰モデルでは説明変数が1個でしたが、ここでは、候補となるモデルとしては四つあります。

- モデル0:家賃は二つの説明変数を用いて説明できない
- モデル1:家賃は部屋の大きさを用いて説明できる
- モデル2:家賃は徒歩時間を用いて説明できる
- モデル3:家賃は部屋の大きさと徒歩時間を用いて説明できる

この四つのモデルは説明変数の個数をpとすると2^p個のモデルがあることから設定されます。われわれはこの2^p個のモデルから一つのみを選択する必要があります。誤差について何も仮定しない場合には、モデルを選択するための基準として、変動に基づく決定係数R^2や残差分散に基づく自由度調整済み決定係数R^{*2}を用いて一つのモデルを選択することになります。

② 誤差ϵが正規分布に従う場合

自由度調整済み決定係数は、説明変数の組が応答変数をどの程度説明しているかを示しています。応答変数に含まれる誤差について何ら仮定を置かないことで、さまざまな場面に応用できますが、個別の事情への適用にしかなりません。そこで、誤差に何らかの分布を仮定することで、回帰モデルの有効性を検証しましょう。そのモデルとは、以下を仮定するものです。

- 誤差ϵは平均0、分散σ^2の正規分布$N(0,\sigma^2)$に従う
- 異なる誤差ϵ_iと$\epsilon_k(i \neq k)$は独立である

この仮定のもとで回帰モデルを設定します。

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \quad \epsilon \sim N(0,\sigma^2)$$

このとき、上の仮定のもとで

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$V(Y) = \sigma^2$$

となります。また、ϵについての分布からn個のϵ_iについて

$$\epsilon_i \sim N(0,\sigma^2) \qquad (i = 1,2,\cdots,n)$$

から、$\hat{Y}_i$をモデルによる予測値、$e_i = Y_i - \hat{Y}_i$を残差、$\bar{Y}$を平均とすると

$$\sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim x_{n-1-p}^2$$

$$\sum_{i=1}^{n}\frac{\left(\hat{Y}_i-\bar{Y}\right)^2}{\sigma^2}\sim x_p^2$$

であることが示され、予測値の分散$S_{\hat{y}}^2$と残差の分散S_e^2の比は、自由度$(p, n-1-p)$のF分布

$$F=\frac{S_{\hat{y}}^2}{S_e^2}\sim F_{(p,n-1-p)}$$

に従うため、この値を用いてモデルを評価することができます。ただし、F-値を用いて、二つのモデルから一つのモデルを選択する場合には、モデルが階層的になっている必要があります。つまり、一方のモデルの説明変数の組が他方のモデルの説明変数を全て含んでいることが必要です。このように誤差に適切な分布を仮定することで、モデルについての仮説検定

- 帰無仮説 H_0:この説明変数の組では応答変数は説明できない
- 対立仮説 H_1:少なくとも一つの説明変数で応答変数は説明できる

を組み立てることができます。したがって、複数の階層的なモデルから、決定係数R^2、自由度調整済み決定係数R^{*2}とF-値の3個の基準をもとにモデルの選択が可能になります。

重回帰モデルでの分散分析表

変動要因	平方和	自由度	分散	F-値
モデル	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	p	$s_{\hat{y}}^2=\sum_{i=1}^{n}\frac{(\hat{y}_i-\bar{y})^2}{p}$	$F=\frac{s_{\hat{y}}^2}{s_e^2}$
残差	$\sum_{i=1}^{n}{e_i}^2$	$(n-1)-p$	$s_e^2=\sum_{i=1}^{n}\frac{e_i^2}{[(n-1)-p]}$	
合計	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$(n-1)$	$s_y^2=\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n-1}$	

さらに、説明変数の組$(x_1, x_2, \dots, x_p)$において、説明変数の有意性を検討する検定は、

$$H_0\text{:ある説明変数}x_j\text{では説明できない}\leftrightarrow H_0\text{:}\beta_j=0$$

とする帰無仮説になるので、$\hat{\beta}_j$に基づくt-値

$$t_j=\frac{\hat{\beta}_j-0}{\mathrm{se}(\hat{\beta}_j)}\sim t_{n-p-1}$$

を用いて検定を行うことができます。

ここで、家賃(単位: 10円)を応答変数とし、部屋の大きさと徒歩時間を説明変数とした重回帰モデルを用いて分析してみましょう。

この場合、候補となるモデルは四つあります。

- モデル0:家賃は二つの説明変数を用いて説明できない
- モデル1:家賃は部屋の大きさを用いて説明できる
- モデル2:家賃は徒歩時間を用いて説明できる
- モデル3:家賃は部屋の大きさと徒歩時間を用いて説明できる

概要

回帰統計	
重決定 R2	0.7893
補正 R2	0.787
標準誤差	1178
観測数	188

分散分析表

	自由度	観測された分散比	有意 F
回帰	2	346.4	<2.2e-16
残差	185		
合計	187		

	係数	標準誤差	t	P-値
切片	2825.46	406.64	6.948	6.12E-11
X 値 1	285.49	10.99	25.986	<2e-16
X 値 2	-60.96	29.13	-2.092	0.0378

予測式と偏回帰係数の標準誤差を示すと以下のようになります。

家賃(10円単位)≒2825.46+285.49×大きさ＋(-60.96)×徒歩
(標準誤差) (406.64) (10.99) (29.13)

偏回帰係数を検討する場合にt-値を用いる場合がありますが、その場合には説明変数間の相関関係を考慮する必要があります。さらに、説明変数の個数が3個以上になると、説明変数間の相関関係を捉えにくくなります。このような場合には、説明変数の相関係数行列R_xの行列式

$$0 \leq |R_x| \leq 1$$

を求め、その絶対値が0に近ければ、変数間に線形従属関係が疑われるので、係数の評価などでは注意が必要であり、逆に絶対値が1に近ければ、係数についてはそのまま比較できるでしょう。

この例で行列式の値を求めると、$\left|R_{(大きさ、\ 徒歩)}\right| = 0.9936$となり、1に近く、偏回帰係数の大きさはそのまま解釈できそうです。大きさのt-値は25.986、徒歩のt-値は-2.092で、いずれも「この説明変数では説明できない」という帰無仮説を5%の優位水準で棄却することができます。

また、F-値は346.4と十分大きいので、「この説明変数の組では応答変数は説明できない」

という帰無仮説は棄却されます。

➢ 多重共線性

標準化された変数y, x_1, x_2について説明変数x_1とx_2との間の相関係数が$r_{x_1,x_2} = 0.90$であり、偏回帰係数が$\frac{1}{2}$であったとします。しかし、説明変数間の相関が高い

ので、回帰モデルを

$$y \fallingdotseq \frac{1}{2}x_1 + \frac{1}{2}x_2 \fallingdotseq 1x_1 + 0x_2 \fallingdotseq \frac{3}{2}x_1 - \frac{1}{2}x_2$$

としても決定係数などには大きな差が出ないでしょう。このように説明変数間の相関関係を無視して偏回帰係数の大きさを比較してはいけません。偏回帰係数の大きさは、設定したモデルのもとでのみ意味を持ちます。この問題は、しばしば、説明変数間の多重共線性問題として取り扱われます。

練習問題 （解答は P.71 です）

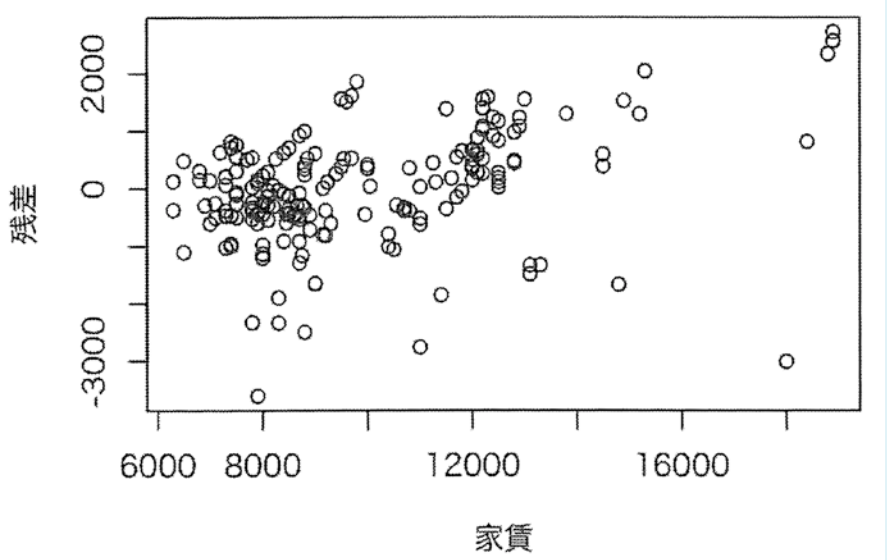
問１ 賃貸マンションのデータを用いて、説明変数として徒歩と大きさと築年数を用いて重回帰分析を行った結果、以下のような結果を得た。家賃(単位10円)と残差の散布図は以下のようになった。この結果から、この分析の妥当性について検討せよ。

相関係数行列

	徒歩	大きさ	築年数
徒歩	1.00000000	−0.08027299	0.04352430
大きさ	−0.08027299	1.00000000	−0.04826252
築年数	0.04352430	−0.04826252	1.00000000

重回帰分析の結果

残差				
最小値	第1四分位数	中央値	第3四分位数	最大値
-3608.6	-461.6	25.3	541.2	2724.5
偏回帰係数:				
説明変数	推定値	標準誤差	t値	$\mathrm{P_r}(>\|t\|)$
切片	3816.099	355.465	10.736	$< 2e^{-16}$
徒歩	-52.150	24.266	-2.149	0.0329
大きさ	281.737	9.153	30.781	$< 2e^{-16}$
築年数	-116.953	12.831	-9.115	$< 2e^{-16}$
残差の標準誤差	980.8	自由度184		
決定係数	0.8548	自由度調整済決定係数	0.8524	
F比:	361.1	自由度1=3	$p - value$	
		自由度2=184	$< 2.2e^{-16}$	



家賃(単位10円)と残差の散布図

Ⅷ 時系列データの分析

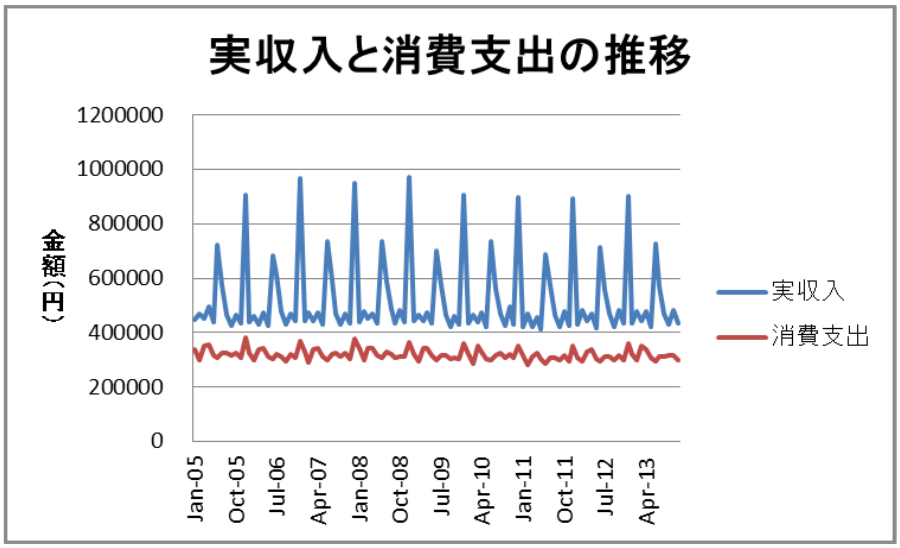
最後に、利用されることの多い時系列データの見方について紹介します。

1. 時系列データ

時系列データは、複数時点から収集したデータをもとに未来の傾向を探るために活用されます。このデータの特徴は、データが時点順に並べられており、この順序は変更できないことです。前章で扱った家賃データでは、データ1件1件を記録するとき、その順番は任意ですが、時系列データでは、その記録される順番は時点順です。

下の表は、家計調査の勤労者世帯の実収入と消費支出の時系列データです。時系列データは横軸に時点、縦軸に対象となる変数をとった折れ線グラフなどでデータを見て、次の時点の傾向などを考察します。また、ボーナス支給などがあるため、季節変動も考慮する必要があります。

年月	実収入	消費支出
2005年1月	448635	338924
2005年2月	469673	300222
2005年3月	451360	353317
…	…	…
2013年9月	431931	315443
2013年10月	482684	316555
2013年11月	436293	300994



① 指数化

ある時点tの実収入y_tと次の期のy_{t+1}の変化を考える場合、$y_{t+1}-y_t$という差を見る方法と、$\frac{(y_{t+1}-y_t)}{y_t}$という変化率で見る方法がありますが、時系列データでは、多くの場合、

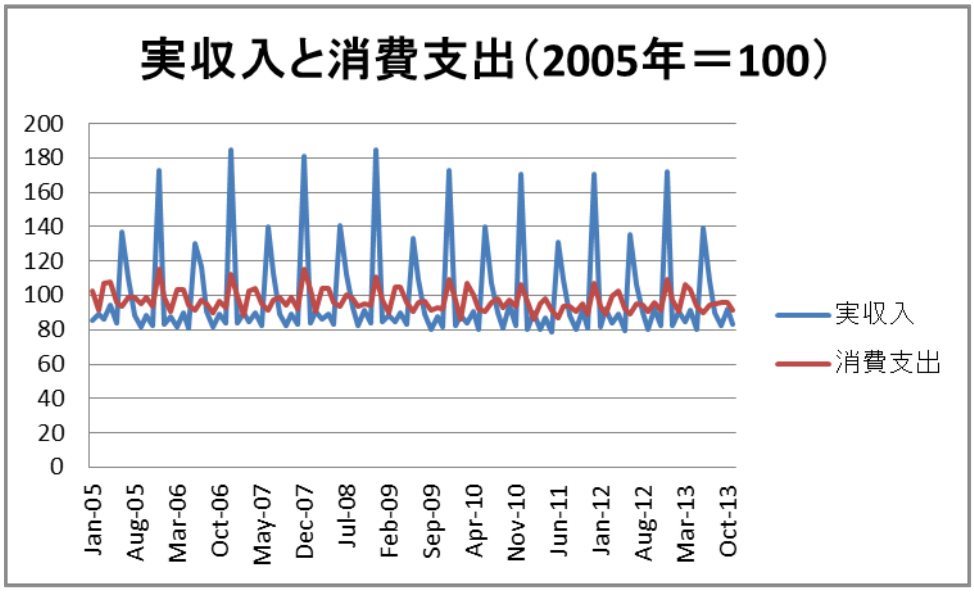
変化率をみます。

複数の時系列データを比較する場合は、ある時点を基準時として指数化が行われることがあります。観測値を$y = (y_1, y_2, \cdots, y_{T-1}, y_T)$とし、基準時$t_0$の値を$y_0$とすると、

$$I_t = \frac{y_t}{y_0} \qquad (t = 1,2, \cdots, T)$$

によって、各時点の指数I_tが求められます。基準時は、多くの場合、西暦の末尾が0又は5の年が用いられます。

下の図は、実収入と消費支出について、2005年平均を基準時として、指数化したものです。上の図では実収入と消費支出の金額のレベル差がありましたが、指数化することによって、レベルが揃えられ、比較しやすくなっています。

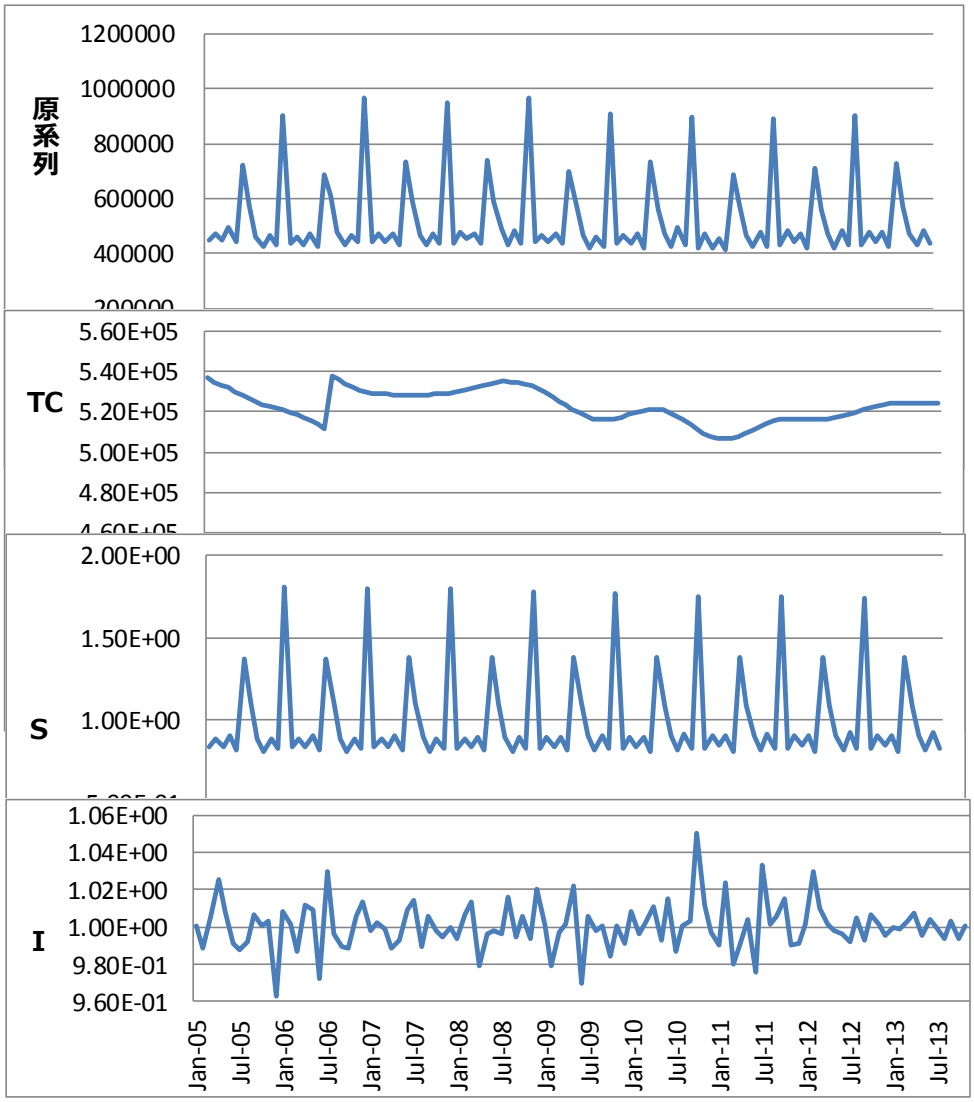


② 移動平均

時系列データは、

- 傾向変動(T)…長期にわたる連続的な変化、基本的な変動方向(傾向)を表す。
- 循環変動(C)…数年程度の周期の一定しない上下運動、景気循環などを表す。
- 季節変動(S)…1年周期の規則的な変動、季節性表す。
- 不規則変動(I)…上記以外の変動で特に規則的でない変動を表す。

の四つからなると考えられています。



時系列データの変化を見るときは、短期的な上下の細かい変動ではなく、長期的な傾向変動(トレンド)を見るようにします。

そのため、データを

$$y_t = TC_t + S_t + I_t \qquad (t = 1,2,\cdots,T)$$

のように、三つに分解し、傾向変動を抜き出してみようと考えます。その方法の一つが移動平均です。

この場合には、時点tでのma_tとしてk時点前からk時点後までの$2k+1$個の値$(y_{t-k}, y_{t-k+1}, \cdots, y_{t-1}, y_t, y_{t+1}, \cdots, y_{t+k-1}, y_{t+k})$を用いて、

$$ma_t = \sum_{s=t-k}^{t+k} \frac{y_s}{2k+1}$$

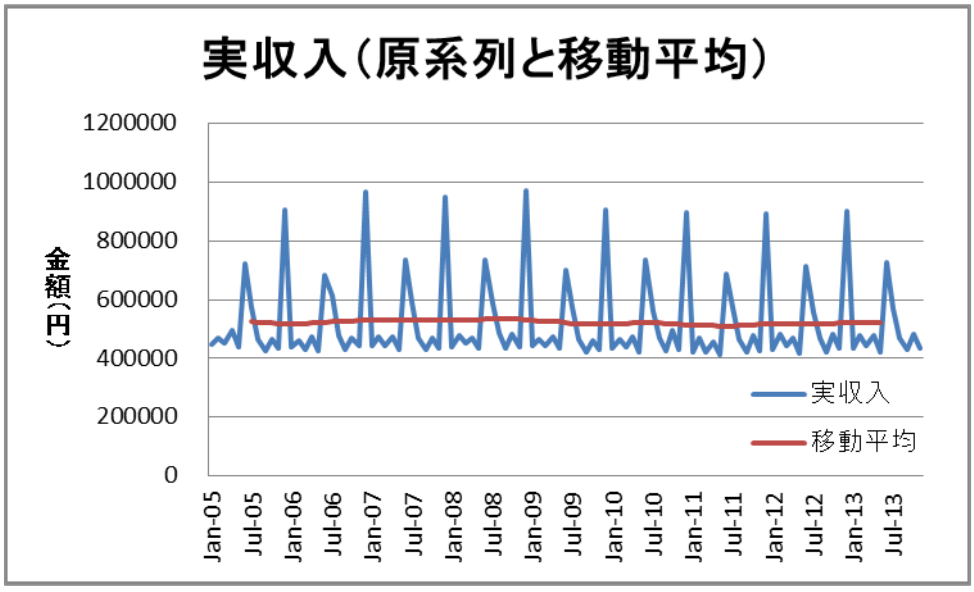
とし、$(2k+1)$移動平均といいます。

しかし、多くの経済時系列データは月次データや四半期データなど偶数の周期をもちます。月次データの場合は、最初の項をy_{t-6}の代わりに$\frac{y_{t-6}}{2}$、最後の項としてy_{t+6}の代わりに$\frac{y_{t+6}}{2}$を用いて

$$ma_t = \frac{\left(\frac{y_{t-6}}{2} + y_{t-5} + \cdots + y_{t+5} + \frac{y_{t+6}}{2}\right)}{12}$$

として、求められます。

下の図は、上記の方法で、実収入のデータの移動平均を求めたものです。原系列に比べると、季節的な変動や短期的な不規則が打ち消され、傾向変動が見やすくなっていることが分かります。



なお、経済時系列データの多くでは、季節性を取り除いて、基調を見るために、季節調整値として、TCI系列を公表しています。

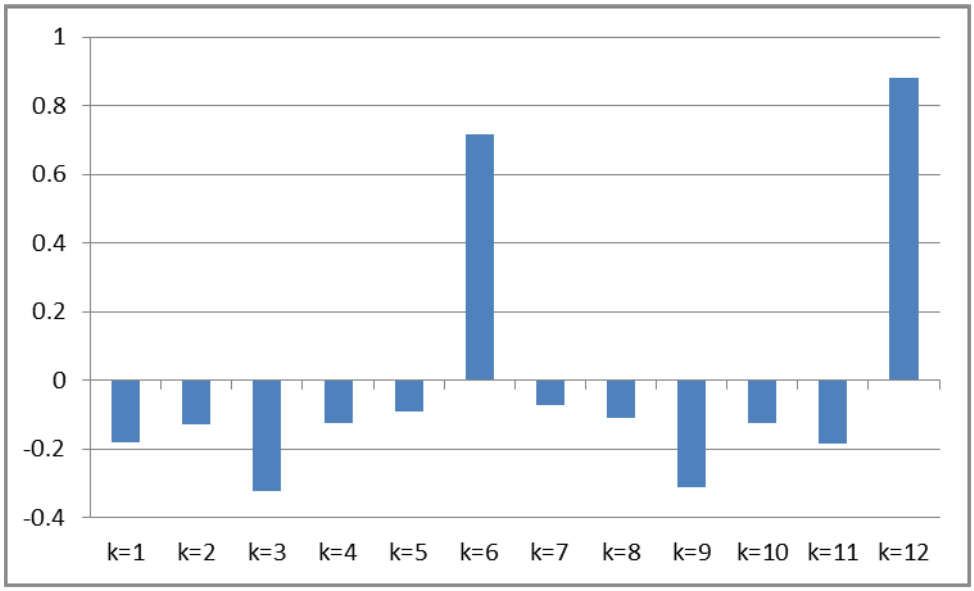
③ 自己相関係数とコレログラム

二つの変数の関係を表す指標としての相関係数については、前章で説明しましたが、この考え方を時系列データに応用すると、一つの時系列データに対して多数の相関を考えることができます。

ある時点tにおけるデータy_tに対して1時点前のデータをy_{t-1}、2時点前のデータをy_{t-2}、k時点前のデータをy_{t-k}と表したとき、時系列y_tとk時点前にずらした時系列y_{t-k}を別の時系列と考え、k時点前との関係を表す自己相関係数を定義することができます。

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

このr_kを時点差kを横軸にとってプロットしたものがコレログラムです。下の図は、実収入のコレログラムです。



このデータでは、6か月と12か月に周期性があることが読み取れます。

2. 指数の作成と利用

異なった時点間又は地域間における数量(生産、出荷、輸出など)や価格を比較する目的で作成した指数が広く利用されています。一般に、多数の同種のデータを比較するために、ある値を基準にして他の値を基準値に対する比で表したものを指数と呼びます。数量指数や価格指数はその代表的なものです。ここでは、物価指数を例として、基本的な考え方を説明します。

基準時点の価格をp_{0i}、比較時点の価格をp_{ti}、基準時点の購入数量q_{0i}、比較時点の購入数量q_{ti}とすると、物価指数の代表的な算式は下のとおりです。

- ラスパイレス式

$$I^L = \frac{\sum_{i=1}^{m} p_{ti}q_{0i}}{\sum_{i=1}^{m} p_{0i}q_{0i}} = \frac{\sum_{i=1}^{m} \frac{p_{ti}}{p_{0i}} p_{0i}q_{0i}}{\sum_{i=1}^{m} p_{0i}q_{0i}} = \sum_{i=1}^{m} \frac{p_{ti}}{p_{0i}} w_{0i}$$

- パーシェ式

$$I^P = \frac{\sum_{i=1}^{m} p_{ti}q_{ti}}{\sum_{i=1}^{m} p_{0i}q_{ti}} = \frac{\sum_{i=1}^{m} p_{ti}q_{ti}}{\sum_{i=1}^{m} \frac{p_{0i}}{p_{ti}} p_{ti}q_{ti}} = \frac{1}{\sum_{i=1}^{m} \frac{p_{0i}}{p_{ti}} w_{ti}}$$

➢フィッシャー式

$$I^F = \sqrt{I^L \times I^P} = \sqrt{\frac{\sum_{i=1}^{m} p_{ti}q_{0i}}{\sum_{i=1}^{m} p_{0i}q_{0i}} \times \frac{\sum_{i=1}^{m} p_{ti}q_{ti}}{\sum_{i=1}^{m} p_{0i}q_{ti}}}$$

物価指数を作成するためには、個々の品目の価格とウエイトが必要です。個々の品目の価格については毎月調査し、ウエイトについては一定の期間固定し、同じウエイトを用います。ラスパイレス式では、基準時のウエイトを使用し、パーシェ式では比較時のウエイトを用います。

消費者物価指数や企業物価指数といった代表的な物価指数は、西暦の末尾が0又は5の年を基準年とするラスパイレス式で計算されています。

解答と解説

■練習問題　推定をする（問題は p.37）

問 1　(1) 勤労者世帯の消費支出のように、極端な正のゆがみを持った(右の裾が長い)分布でも、$n = 2500$と十分に大きい場合には、標本平均$\bar{x}$は正規分布に従うと考えてよい。また、母集団の標準偏差は標本の標準偏差にほぼ等しい。95%信頼区間は$|\bar{x} - \mu| \leq \frac{1.96s}{\sqrt{n}} = \frac{1.96\times 29.5}{\sqrt{2500}}$ の解として$31.6 < \mu < 34.0$となる。

(2) 消費支出のような強いゆがみを持った母集団分布からの標本平均の分布は、調査世帯数が$n = 25$程度では、正規分布には十分近いとは言えない。したがって、正規分布を利用した信頼区間の公式$\bar{x} \pm \frac{1.96s}{\sqrt{25}}$や1.96を自由度$25 - 1 = 24$の$t$分布の上側の2.5%点である2.06で置換えた$\bar{x} \pm \frac{2.06s}{\sqrt{25}}$を構成しても、参考程度の意味しか持たない。

■練習問題　仮説を検定する（問題は p.52）

問 1　(1)$H_0: \mu = 200$を両側対立仮説$\mu \neq 200$に対して検定すればよい。ここでは分散が既知だから、次のzを求めるのが簡単である。

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{207 - 200}{10/\sqrt{16}} = 2.80$$

両側検定の棄却域$|z| \geq 1.96$と比較して、仮説は棄却される。

なお、$\mathrm{P_r}(|z| \geq 2.80) = 0.0051$として求められる$P$-値も十分小さいため、生産工程に何らかの障害が発生していることが疑われる。

(2) 標準偏差$\sigma = 10$をもつ正規分布にしたがうことを前提とするから、$z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0, 1)$を利用すればよい。信頼係数95%の信頼区間は$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 207 \pm 1.96\frac{10}{\sqrt{16}} = [202.1,\ 211.9]$である。

$\mu = 200$はこの区間に含まれないから、前問で見たとおり、帰無仮説$H_0: \mu = 200$は「有意水準5%、両側対立仮説」を用いると棄却される。

(3)σは未知だからt 検定を用いる。

$t = \frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t_{n-1}$を利用する。自由度は$16 - 1 = 15$、両側対立仮説を想定して、上側2.5%点は$t_0 = 2.13$である。これと$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = 2.80$を比較して仮説は棄却される。

なお信頼係数95%の信頼区間は$\bar{x} \pm 2.13\frac{s}{\sqrt{n}} = 207 \pm 2.13\frac{10}{\sqrt{16}} = [201.7,\ 212.3]$であり$\mu = 200$は含まれていないが、このことは仮説検定の結果と整合的である。

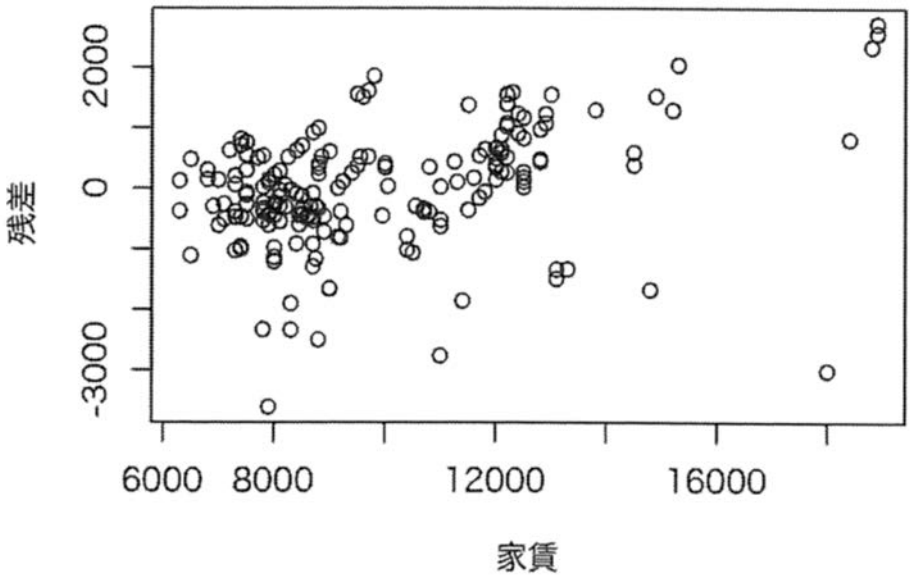
■練習問題　線形モデルの活用（問題は p.62）

問１　・相関係数行列

この相関係数行列から説明変数間には相関が余りないことが分かる。それは、相関係数行列の行列式の絶対値が1.0に近いことからも確認できる。

・重回帰分析の結果

決定係数の値$R^2 = 0.8584$により応答変数の家賃は説明されていると解釈できる。誤差の分布が$N(0, \sigma^2)$と仮定した場合のF値も大きく、帰無仮説は棄却されよう。次に係数について検討すると、相関係数行列や行列式の値から、偏回帰係数については、おおむね個別に帰無仮説を検定することができよう。説明変数としては大きさ、築年数、徒歩共に5%で有意であるが、説明変数として築年数と大きさの組のみを用いた場合でも同程度の結果が得られるであろう。



家賃(単位10円)と残差の散布図

応答変数と残差の散布図から，右上がりの傾向が見られる．これより，高い家賃について過小に予測されている可能性がある。