統計学 2019（心理とビジネスを学ぶ人のための）

［松本治彦］

その 2　平均値の標準誤差より終わりまで

2019 年 1 月

宇部フロンティア大学出版会

平均値の標準誤差

平均値、元データのバラつき、データの件数と誤差の関係は、標準誤差（SE；Standard Error）で表す。（平均値の標準誤差）＝（元データの標準偏差）÷√（平均値の計算に用いたデータの件数）である。サンプルサイズとは、集団全体から抜き出されたデータの件数。例えば、75 万個の「4 人のデータから求めた平均値」の平均値は 300 万人の高校生全体の金額の平均値と一致する。この 75 万個の平均値について、分散あるいは標準偏差がどうなっているかということだが、この「4 人のデータから求めた平均値」の標準偏差のことを標準誤差と呼ぶ。

なお、標準誤差はデータから算出された平均値のみに対して存在しているわけではない。データから算出した「割合の標準誤差」はもちろん、データから算出された「標準偏差の標準誤差」というややこしいものもある。いずれにせよ、自分の関心のある「真の値」の代替として、限られたデータからそれに近いものを算出する、という行為を行う背景には、「必ずいま例として挙げた 75 万個の平均値のように、膨大な数の「あり得たはずの値」が存在している。その「あり得たはずの値」の分布における標準偏差が標準誤差である。一方、標準偏差はもとのデータそのもののバラつきを示す指標である。そして複数のデータから求められた平均値のバラつき（標準誤差）は、必ず元のデータのバラつき（標準偏差）よりも小さいものになる。また求めるのに用いたデータの件数、すなわちサンプルサイズが増えれば増えるほど標準誤差は小さくなる。データの件数が多くなればなるほど、元のデータのうち真の平均値より大きいものだけ、あるいは逆に小さいものだけがサンプルに含まれる確率よりも、真の平均値より大きい値のものと小さい値のものが混在してくる確率のほうが大きくなる。そうするとデータの件数が増えれば増えるほど、真の平均値付近に「データの平均値」はどんどん集まってくることになる。ゆえに、データの件数が増えれば増えるほど、データから求めた平均値のバラつき（標準誤差）は、もとのデータのバラつき（標準偏差）よりも小さくなってくる。データの件数が大きくなればなるほど標準誤差が標準偏差よりも小さくなるという関係性を数学的に表現すると、

平均値の標準誤差＝（元データの標準偏差）/√（平均値の計算に用いたデータの件数）

という関係になる。

平均値と標準偏差を使えば「サンプルサイズ設計」ができる

これまでのデータから求めた平均値と標準偏差を用いて「次の調査でどれくらいの標準誤差にするためにどれくらいのデータの件数（すなわちサンプルサイズ）が必要か」という見積もりを行うことができる。このようなデータの件数の見積もりのことは専門用語でサンプルサイズ設計と呼ぶ。

先ほどの関係式に「標準偏差が 1 千円」という情報を加え、横軸が 1 地域ごとのサンプルサイズ、縦軸がそこから求めた平均値の標準誤差のグラフができる。サンプル数が 4 だと、標準誤差は 500 円、サンプルサイズが 100 人だと、SE は 100 円、サンプルサ

イズが 2500 人ずつのサンプルが得ることができれば、SE は 20 円となる。
平均値が 4 千円、SE が 100 円の結果で、平均値±2SE の範囲「だいたい 3800 円～4200 円」。平均値が約 4 千円という結果がどこかの地域で得られたとして、それが数十円の単位まで正確、という必要はおそらくない、各地域で 2500 人は大げさすぎ、一方、たった 4 人で標準誤差が 500 円という状況では「平均 4 千円」の結果が得られたとしても、それは「平均予算が 3 千円～5 千円という範囲となり、あまり参考にならない。
このように最終的に得られるであろう誤差と、調査にかかる手間や予算を天秤にかけて、必要なデータの数を見積もるのがサンプルサイズ設計である。こうしたサンプルサイズ設計の考え方が理解できていれば、「とりあえず全数調査」とか「とりあえずビックデータ」といった考え方が適切でない状況が分かる。

平均値

平均とは、複数個のデータを代表する一つの統計量である（データの分布が正規分布とみなせることが前提)。統計用語では、「代表値」の一種である。
例えば、平均が 50 であるとすれば 50 がこのデータ集団の代表である。実際に 50 という値のデータがなくても、この付近にデータが密集し、50 から±方向に離れるにつれて分布がまばらになってゆく。
平均を用いてはならない場合としては、データの尺度が間隔・比率尺度でない場合、データの中に極端値が存在する場合、データの分布が正規分布とみなせない場合である。このときには、名義尺度データあるいは順位尺度データの処理へ進む。

偏差平方和、標準偏差、標準誤差

数値データのばらつきの程度を調べるのに使われる。

偏差平方和（残差平方和ともいう）

偏差を平方した合計である。偏差とは、平均からの隔たりであり、測定値－平均値である。したがって、
偏差＝測定値－平均値、偏差平方＝偏差×偏差、偏差平方和＝Σ偏差平方となる。

標準偏差 SD（standard deviation ）

標準偏差とは、データが平均からどの程度ズレているかを表す統計量である。すなわち、標準偏差の値は、データ 1 個分の標準的なズレ幅を示す。統計用語では「散布度」の一種。もし、データの分布が正規分布であれば「平均±標準偏差」の範囲にデータ全体の約 68%がおさまる。例えば、100 個のデータがあり、平均が 50、標準偏差が 10 であるとすると、50±10（つまり、40～60）の範囲にほぼ 68 個のデータがおさまる。
標準偏差の計算式としては、SD＝√｛(偏差平方和）/（N－1)｝＝√｛(データの平方和）/（N－1）－N×（平均値）^{2}/（N－1)｝＝不偏推定値（unbiased estimate）
または、SD＝√（偏差平方和/N）＝√｛データの平方和/N－（平均値）^{2}｝
データ処理上は、どちらか一方を一貫して用いるなら支障はない。なお、SD^2 は分散（varinace)。

標準誤差　SE

　標準誤差は、標準偏差（不偏推定値の方）を√N で割ったもの。

　標準誤差（SE）＝標準偏差の不偏推定値/√N＝√｛偏差平方和/N（N−1)｝

数値データのばらつきの程度を表す場合、1 群のみのときは、(平均値±標準偏差)、2 群以上で平均値の比較を行うときには、(平均値±標準誤差）である。

平均と標準偏差の意味

　本来、平均とは真の値のことである。すなわち、データにまったく誤差が加わらないときには、すべてのデータが集中する一点の値を表している。したがって、真の値が一点に定まらないときの平均は意味がない。正規分布は真の値が一点に定まることの証である。また、そのときデータの値は、「データ＝真の値±偶然誤差」として定義することができる。この真の値の推定が平均値、偶然誤差の推定は SD であり、ある観察場面における標準的データは、(平均値±SD）という値をとることを意味している。

「頻度分布、正規分布、正規分布のあてはめ、

平均μに対する最良の推定値は、で与えられる x バーであり、

標準偏差σの最良の推定値としては、　　から s を計算すればよい。これら２つの統計量は、観測値に関する最初の２個の累乗和から計算され、次の点に関して正規分布と特別の関連をもっている。すなわち、母集団分布が正規型とすれば、その分布に関して標本が提供する情報は、全部この２つの統計量に要約されている。・・・しかし、分布が正規型と著しく異なっている時には、この２つの統計量は、ほとんどあるいは全く役に立たないことがある。」

<u>平均の分布での標準偏差は標準誤差となる！</u>

　平均値に関する統計的な処理の基礎になるのは次の基本的命題である。ある量が分散 σ^2 の正規分布に従うならば、その量に関する大きさが n の無作為標本の平均は、分散が σ^2/n の正規分布に従う。もとの分布が正確には正規分布でないときでも、平均の分布は標本の大きさが増すにつれて一般に正規型に近づくという事実があるので、この命題の効用は幾分か増大する。したがって、もとの分布が正規型であるという十分な確証はなくても、平均の分布が正規型に近づかないような例外的な分布ではないと考えられる根拠があれば、この方法を広く適用してもさしつかえない。

　このことから、もしも母集団の分散がわかっていれば、与えられた大きさの無作為標本の平均の分散を求めることができて、それによってある定められた値と標本平均との差が有意であるかどうかを検定することができる。その差が標準誤差より何倍も大きければ、それは確かに有意である。標準誤差の２倍をもって有意性の限界にとるのが慣例であるが、これは χ^2 分布に関して既に用いた対応する限界 P＝0.05 とほぼ同等である。

「・統計学は平均のことを「mean」、しかし、エクセルでは「Average」を使っている
分散＝平均平方和

・標準偏差は通常、測定誤差を表すと考えられている。そのため、測定回数を増やしても、測定値の標準偏差は変化しない特徴がある。

・測定機器の検出限界などを決める時には、標準濃度が 0 のブランクの測定値の「平均値±3SD」を基準にすることが多く、3σと呼ばれる（約 99.7%）。

・標準誤差 standard error SE—標準偏差をデータ数ｎの平方根で除した値 SE=σ/√ｎ
標準誤差は通常、母平均値を推定するときの推定誤差と考えられる。そのため、測定値の標準偏差が同じでも、データ数（測定回数）が 4 倍になると、SE は半分になる。つまり、データ数の増加により情報量が増えるので、推定精度が上がったと考えることができる。

Ｖ－4　正規分布

1）正規分布とは

生物現象など自然界で観察される多くの計測値は、何であれ平均値に近いほどその出現率が高く、平均値からその両側に値が遠ざかるにしたがって出現頻度は少なくなる。

このうち、同じものを何度も繰り返し計測し、平均値からのずれ（誤差）の大きさを求め、その出現度数を描いてみると、平均値を中心として左右対称の釣鐘状の分布型になることが多い。

1812 年に数学者ガウスは、この純粋な条件で繰り返し計測したときに、一貫して現れる分布型を発見し、それを正規分布（normal distribution）と名付けた。発見者の名前をとってガウス分布（Gaussian distribution）ともいわれる。

この曲線は誤差曲線とも呼ばれます。その理由は、ある寸法を目標にして何かを作るとき、ちょっとした手のはずみなどで、目標の寸法よりもわずかばかり大きくなってしまったり、逆に小さくなってしまったりする‘誤差’が生じます。この誤差の大きさは、正規分布に従うことが知られているからです。つまり、物を作るときには、人によって、あるいは機械によって、目標より平均して大きめの物を作ったり、あるいは小さめの物を作ったりする‘くせ’があります。この誤差の平均値は 0 でないのが普通ですが、誤差の大きさは、その平均値を中心にして左右対称な正規分布にしたがいます。

正規分布は、その分布に従うあるグループの平均値と標準偏差が分かっていれば、その分布に関する全てが分かります。例えば、

正規分布曲線下の面積は、－∞～＋∞　で　1

$\mu-\sigma$ と $\mu+\sigma$ の間の正規分布曲線の面積は全体の約 68%

$\mu-2\sigma$ と $\mu+2\sigma$ の間の正規分布曲線の面積は全体の約 95%

$\mu-3\sigma$ と $\mu+3\sigma$ の間の正規分布曲線の面積は全体の約 99.7%

$\mu-4\sigma$ と $\mu+4\sigma$ の間の正規分布曲線の面積は全体の約 99.99%

という具合となります。「平均値が μ、標準偏差が σ である正規分布」を N（μ、σ^2）と略して記号で表す習慣があります。N は normal distribution の頭文字です。

例）東京オリンピックで優勝した日本女子バレーボールチームの平均身長は 171cm、この当時の女子の平均身長（μ）を 156cm、標準偏差（σ）を 5cm と仮定すると、図 11 から、171cm 以上は③図の $\mu+3\sigma$ 以上にあたる。すなわち約 0.15%である。わが国の昭和 20 年代に約 200 万人の出生数があるのですが、女子がその約半分とすると 100 万人です。したがって、171cm 以上は 1,500 人しかいないことになります。その中から運動神経もある程度発達していて訓練に耐えることができ、なおかつバレーボールが好きな人をさがすのは、大変であったことがわかります。

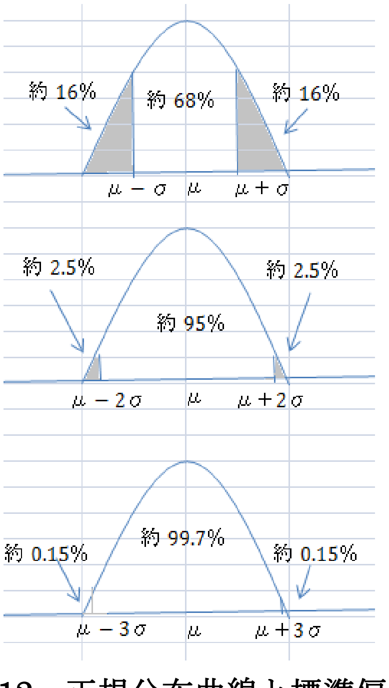


図 12　正規分布曲線と標準偏差

2）正規分布の標準化

先程の例で平均値よりも大きく、175 ㎝よりも小さい人は何%でしょう。これに答えるには、日本の当時の女子の身長について、細かい数表を作っておく必要があります。また、平均値や標準偏差は、各測定値によって単位が異なっています。例えば、身長は cm、体重は kg、ヘモグロビン濃度は mg/dl などです。これらについて、片っ端から数表を作ることが必要になります。しかし、実際にそのような作業をすることは現実的ではありません。そこで、正規分布に従うものならどんなもの

にでも適用できる数表があれば便利ですよね。それでは、そのような便利な数表を作るにはどうしたらよいのでしょうか。

それには「正規分布は平均値と標準偏差によって決まる」という性質を利用するのです。つまり、平均値を固定して、標準偏差をものさしにしてばらつきの大きさを表してやれば、数表は1つで済むことになります。

次の図を見てください。

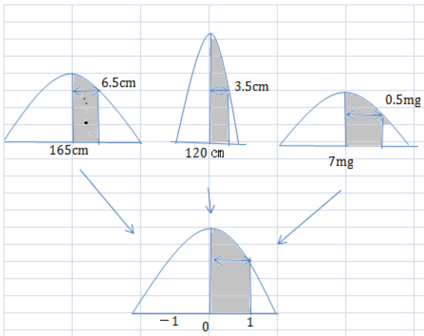


図 13　正規分布の標準化

（a）は青年男子の身長で平均値が 165 cm、標準偏差は 6 cm、（b）は小学 3 年生の身長で平均値は 124 cm、標準偏差は 4 cm、（c）は 1 錠の胃腸薬に含まれるパントテン酸カルシウムの量で平均値は 5 mg、標準偏差は 0.4 mgです。この 3 つの分布は、どれも正規分布なのですが、平均値も違うし標準偏差も異なります。単位も異なります。しかし、ともに正規分布である共通点を利用して、1 つの数表が使えるように工夫できそうです。正規分布であるという共通の点は、平均値の両側に標準偏差だけの幅をとると、その幅の中の面積（図で 2 重斜線の部分）は 0.6826 です。というように、平均値から標準偏差を単位としてある幅をとると、その範囲の面積がどんな正規分布の場合にも等しい値になるのです。そこで、(d) のような正規分布を考えます。この正規分布は平均が 0、標準偏差が 1 です。つまり N（0、1^2）です。この正規分布と他の 3 つの正規分布と比べてみると、例えば、(a) では 165 cmのところを 0 とみなし、横軸の目盛を 6 cmを単位（1 とする）にして書き直すと、(d) と全く同じになります。つまり、165 cmのところが 0 になり、171 cmのところが 1 になり、177 cmのところが 2 に、159 cmのところが−1 になるわけです。

したがって、(d) の正規分布 N（0、1^2）についての詳しい数表があれば、(a) の図形のどの部分の面積もわかることになります。(b) の場合も (c) の場合も全く同じことです。

このように平均値が 0、標準偏差が 1 になるように統計量を考えて、各測定値が平均 0、標準偏差 1 になるような正規分布を作成すると、各測定値の分布上の位置が、比較できて便利です。

問題 9

図 12 で左上の分布で 180 ㎝は下の分布に置き換えると、どのような値になるのですか、下記の空欄を使って答えてください。

このような平均 0、標準偏差が 1 の正規分布を規準正規分布あるいは、標準正規分布といいます。

z を使って、$z = \frac{x - \mu}{\sigma}$

このようにすると、各測定値の単位に関係なく分布上の位置を示したり、検定に利用できます。

z 分布表を使った例）

今、μ ＝156cm、σ ＝5cm の身長分布で 169cm 以上の人の割合を求めたいとします。

この場合まず、169cm に対応する z を求めます。

z ＝（169－156）/5＝2.6

付表 1 の正規分布表より、z ＝2.6 より右側の面積は 0.0047、したがって、割合は 0.47%となります。また、同じ μ 、σ の集団で 150cm 以下の割合は、

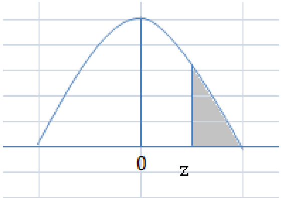
z ＝（150－156）/5＝－1.2

付表 1 より 0.1151 となります。したがって、割合は 11.51%となります。

＊ 注意；正規分布は左右対称なので、z が負でも面積は±1.2 と同じことなります。

（この部分に、付表 1 の見方を記した部分を記載してください。）

付表１　正規分布表



ｚから外側（色の暗い部分）の面積です。ｚが負の場合も、対称性から同面積です。

z	0.00	0.01	0.02	・・・・	0.08	0.09
0.0	0.5000	0.4960	0.4920		0.4681	0.4641
0.1	0.4602	0.4563	0.4522		0.4286	0.4247
0.2	0.4207	0.4168	0.4129		0.3897	0.3859
0.3	0.3821	0.3783	0.3745		0.3520	0.3483
0.4	0.3446	0.3409	0.3372		0.3156	0.3121
0.5	0.3085	0.3050	0.3015		0.2810	0.2776
0.6	0.2743	0.2709	0.2676		0.2483	0.2451
0.7	0.2420	0.2389	0.2358		0.2177	0.2148
0.8	0.2119	0.2090	0.2061		0.1894	0.1867
0.9	0.1841	0.1884	0.1788		0.1635	0.1611
1.0	0.1587	0.1562	0.1539		0.1401	0.1379
1.1	0.1357	0.1335	0.1314		0.1190	0.1170
1.2	0.1151	0.1131	0.1112		0.1002	0.0985
1.3	0.0968	0.0951	0.0934		0.0838	0.0823
1.4	0.0808	0.0793	0.0778		0.0694	0.0681
1.5	0.0668	0.0655	0.0643		0.0571	0.0559
1.6	0.0548	0.0537	0.0526		0.0465	0.0455
1.7	0.0446	0.0436	0.0427		0.0375	0.0367
1.8	0.0359	0.0351	0.0344		0.0301	0.0294
1.9	0.0287	0.0281	0.0274		0.0239	0.0233
2.0	0.0228	0.0222	0.0217		0.0188	0.0183
2.1	0.0179	0.0174	0.0170		0.0149	0.0143
2.2	0.0139	0.0136	0.0132		0.0113	0.0110
2.3	0.0107	0.0104	0.0102		0.0087	0.0084
2.4	0.0082	0.0080	0.0078		0.0066	0.0064
2.5	0.0062	0.0060	0.0059		0.0049	0.0048
2.6	0.0047	0.0042	0.0044		0.0037	0.0036
2.7	0.0035	0.0034	0.0033		0.0027	0.0026
2.8	0.0026	0.0025	0.0024		0.0020	0.0019
2.9	0.0019	0.0018	0.0018		0.0014	0.0014
3.0	0.0013	0.0013	0.0013		0.0010	0.0016
3.1	0.0010	0.0009	0.0009		0.0007	0.0007
3.2	0.0007	0.0007	0.0006		0.0005	0.0005
3.3	0.0005	0.0005	0.0005		0.0004	0.0003
3.4	0.0003	0.0003	0.0003		0.0003	0.0002
3.5	0.0002	0.0002	0.0002		0.0002	0.0002

3）偏差値から正規化を考えてみる！

偏差値とは、ある集団におけるある個体のある種の測定値 x_i を次のような式で標準化した値 y_i です。$\frac{y_i-50}{10}=\frac{x_i-\bar{x}}{SD}$ すなわち、 $y_i=\frac{10}{SD}(x_i-\bar{x})+50$

ここで、SD は標準偏差。偏差値 y_i はもとの測定値 x_i を平均 50、標準偏差 10 になるように標準化したものです。これから考えると基準正規分布は、

$$\frac{z-0}{1}=\frac{x_i-\bar{x}}{SD}\to \quad \frac{x-\mu}{\sigma}$$

とおけるのです。つまり、みなさんはすでに、正規分の基準化よりも難しい計算を使っているのです。

4）標本平均の分布（中心極限定理）

身長が正規分布に従うことは以前に述べたとおりですが、その身長の集団から n 人を選び出し（抽出）、その標本平均を求めます。この作業を繰り返し行うと、標本平均の集団が出来上がりますが、その集団はどんな集団になるのでしょうか。この答えは重要な定理となっているので、以下に紹介します。

平均値 μ 、分散 σ^2 の任意の分布型（どんな分布でもよい）をした母集団から、大きさ n の標本 χ_1、χ_2、χ_3、…..、χ_n を選んだとき、標本平均

$$\bar{x}=\frac{x_1+x_2+x_3+\cdots\cdots\cdots\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

の分布は、n が大きくなると正規分布 N（μ 、σ^2/n）に近づく。

この定理は、後述の区間推定や検定で使うので重要です。

なぜ､元のデータが平均値付近になくても､それらを足し合わせた値の集まりは中心(平均値）付近に集まり、そこから左右対称な分布になるのか？

これは、ド・モアブルは見つけた「コインを何枚か投げてそのうち何枚が表になるか」という確率は、投げる枚数が多くなると正規分布に近づくという事実を考えればよい。

問題 10

この正規分布の標準偏差（標準誤差）について下記の空欄を使って答えてください。

ここで、もう一度、平均値、偏差、偏差平方和、分散、標準偏差および標準誤差の関係を整理しておいてください。

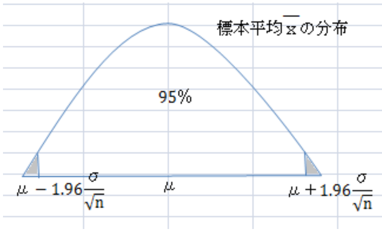


図 14　標本平均の分布

なお、標本平均の分布は n が 25 より大きいときはよく正規分布に近似し、x がどんなにゆがんでいても、n が 50 より大きいと正規分布によく近似することが知られています。

Ｖ－5　推定と検定

１）推定とは

例えば、インフルエンザにかかった子どもが脳症をおこして死亡する確率（何%とか）は、どれくらいか、知りたいことがあります。その確率はすべての子どもをインフルエンザに罹患させなければわからない。これは、実際には不可能です。

そこで、この場合には全体の一部のみから全体を知ろうとする。これが推定です。

２）母集団と標本

知りたいことは確率だけとは限らない。平均値を知りたいこともある。これらの知りたい値を

母数 parameter といいます。母数を知ろうとする対象が**母集団 population**、そこから抽出する一部を**標本 sample** と呼びます。

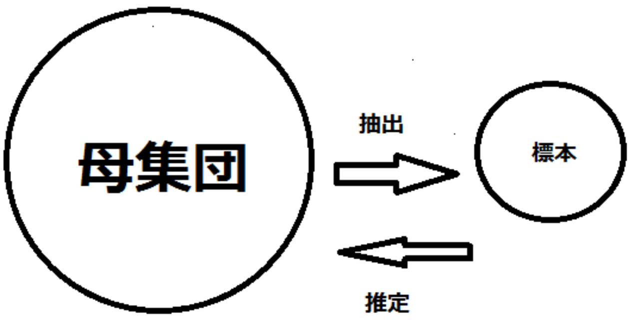


図 15　母集団と標本

３）平均値μの区間推定

　中心極限定理より、大きさｎの標本を無作為抽出して得られる$\bar{x}$の分布は、平均μ、標準偏差$\frac{\sigma}{\sqrt{n}}$の正規分布に近似することがわかっている。このことを利用すると、$\bar{x}$が$\mu - 1.96\frac{\sigma}{\sqrt{n}}$と$\mu + 1.96\frac{\sigma}{\sqrt{n}}$との間に入る確率は95％となる。これを式に表すと

$P\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$になります。この式を変形して$\bar{x}$と$\mu$を入れ替えると、$P\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$となります。ここで、$\sigma$をSDで置き換えると、$\bar{x} - 1.96\frac{SD}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{SD}{\sqrt{n}}$、この式が$\mu$の95％信頼区間となります。

＊正規分布の項では、$\mu \pm 2\sigma$で約95％としていましたが、ここではより95％に近づけるために、　2 →1. 96を用います。

この式の意味；

　標本平均を求めるため標本を抽出する作業を何回も繰り返して、そのたびごとに信頼区間をつくると、100回のうち95回はその区間に母平均（μ）が含まれているということです。けっして、1回だけ求めた信頼区間に母平均（μ）が入る確率が95％という意味ではありません。

問題11

　上記で、標準偏差がSD/√n になっているのはなぜですか、下記の空欄を使って答えてください。

４）ｔ分布（スチューデント分布ともいう）による区間推定

　先ほど、μの区間推定で標準偏差σのかわりに標本推定値ｓを使用しました。これは、標本が大きいからです。標本の大きさｎが25より大きいときはｓを使用して差し支えありません。

しかし、標本の大きさが5とか10のときには、σをｓで置き換えることはできないので、ｔ分布を用います。ｔ分布は平均が0の正規分布に似た左右対称の分布で、自由度（df＝ｎ－1）によって曲線が異なります。

$t=\frac{\bar{x}-\mu}{\frac{SD}{\sqrt{n}}}$　自由度は独立な変数の個数でここでは調査対象をｎとしたときｎ－1となります。df＝5のときに、ｔ分布で左右の面積が5%となるｔ値は2.571です。ｎが大きくなるにつれてｔ分布の曲線は次第に正規分布に近づき、ｔ＝2571は基準正規分布で左右の面積が5%ととなるｚの値1.96に近づく。

df＝5のとき、μの95%信頼区間は　$\overline{x}-2.571\frac{SD}{\sqrt{n}}<\mu<\overline{x}+2.571\frac{SD}{\sqrt{n}}$　で一般には

$\overline{x}-t_0\frac{s}{\sqrt{n}}<\mu<\overline{x}+t_0\frac{s}{\sqrt{n}}$　となります。ここで t_0 は自由度dfと目的とする信頼区間によって決まるもので、ｔ分布表から求めることができます。例えば、

df＝5で信頼区間が95%のときには t_0＝2.571
df＝10で信頼区間が95%のときには t_0＝2.228
df＝5で信頼区間が99%のときには t_0＝4.032
df＝10で信頼区間が99%のときには t_0＝3.168

__問題12__

今、10人の男子学生の身長が、166、167、170、175、173、169、177、171、175、178（cm）とすると、標本平均は172.1cm、標本標準偏差は4.15 cmになります。このときに、下記の空欄を利用してｔ分布を使ってμに対する95%の信頼区間を求めください。

＊　ここまでの説明で、正規分布でμ＋2σがμ＋1.96σとなること、さらにμ＋1.96SD/√ｎ→μ＋ｔSD/√ｎの展開、区間推定、母平均の検定の作り変え」

自由度の説明：

今、正規分布から取り出された2つのデータがあるとします。データが2つあれば、ｔ分布を利用して母平均μの区間推定ができます。区間推定をするには、まず2つのデータから標本平均を求め、それを使って標本標準偏差ｓを計算するのが第一段階です。しかし、よく考えてみると、ｓを計算するに際して、本当の平均値μがわからないので、データから架空の平均値を作り出して、その平均値で後の計算をやっています。本来な

ら、標本平均が母平均と同じになるように標本を選ばなければなりません。ということは、2 つの標本を選ぶ場合には、2 つめの標本は 1 つめとの平均がちょうど母平均と同じになるように選ばなければならないので、勝手に選ぶことができません。自動的に決まってしまうということです。すなわち、2 つの標本を選ぶ場合には、自由が許されるのは最初の 1 つだけで、2 つめは自由が許されないのが本来の姿ということになります。そういう意味で、n が 2 のときには、自由度は 1 になります。n が 3 以上のときも同じことです。したがって自由度は df＝n－1 になるのです。

自由度という考え方は「データの数から、そのデータで作り出して使用した平均値の数を差し引いたもの」と考えておけば間違いありません。

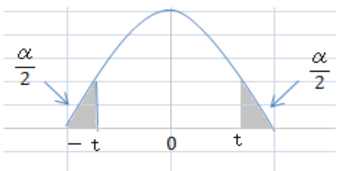
問題 13

18 歳男子 100 人の座高の平均値は 88.8 ㎝、標準偏差は 3.50 ㎝であった。下記の空欄を利用して μ に対する 95%信頼区間を求めてください。

問題 14

上記で μ に対する 99%信頼区間を求めてください（下記の空欄）。

付表２　ｔ分布表



ｔ、－ｔの外側（色の暗い部分）の面積を加えてαとなるが、そのときの自由度dfとｔの値です。

df＼α	0.10	0.05	0.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
15	1.753	2.131	2.947
20	1.725	2.086	2.845
25	1.708	2.060	2.787
30	1.697	2.042	2.750
40	1.684	2.021	2.704
50	1.676	2.009	2.678
60	1.671	2.000	2.660
70	1.667	1.994	2.648
80	1.664	1.990	2.639
90	1.662	1.987	2.632
100	1.660	1.984	2.626
120	1.658	1.980	2.617
∞	1.645	1.960	2.576

p 値と信頼区間の意味う

例えば「カラスは基本的に黒い」という仮説を主張したいときに、あえてまず考えた。「自説を完全に覆すような仮説」、つまり「カラスが黒いかどうかは半々」というようなものを帰無仮説と呼ぶ。主張したいことを「無に帰す」仮説という意味。そして、帰無仮説が成立していると仮定した状態で、実際のデータまたはそれ以上に帰無仮説に反するようなデータが得られる確率を p 値と呼ぶ。Probability の意味。この場合、カラスが黒いか白いかは半々のときに、100 回連続で黒いカラスが見つかったというような観察結果が得られる確率が、1 兆分の 1 の 1 兆分の 1 よりも小さいといいうのが今回の場合の p 値である。これが小さければ「その帰無仮説はあり得ない」と考える方が自然である。どれくらい p 値が小さければ「あり得ない」と考えるかという目安として、分野にもよるが、概ね 5%未満、つまり帰無仮説のもとでは 20 回に 1 回程度しか起こらないようなデータが得られたとすれば「あり得ない」と考えるのが慣例。

なぜ、5%を境目とするか、特に数学的な根拠はないが、統計学者フィッシャーがかつて「p 値を 5%で判断するのが便利だ」と書いたことがきっかけになっているらしい。

「黒いカラスが 9 割」「黒いカラスが 8 割」という帰無仮説を唱えてもよいのだが、「完全に台無しにする」仮説以外の帰無仮説についても、どこまでならありえない仮説で、どこからは否定しきれない仮説なのか、という区間を示す。これが信頼区間の真の意味である。信頼区間が平均値±2SE で表すことができる、というのは計算上たまたまそうであるというだけで、本来の意味としては「あり得ない帰無仮説」と「否定しきれない帰無仮説」の境目がどこからどこまで、という範囲を示すというのが、イエジ・ネイマンたちの定義による信頼区間の意味。

実際に計算してみると、「97.0%のカラスが黒い」という仮説のもとでは偶然 100 羽連続して黒いカラスに出会う確率は 4.8%だが、「97.1%のカラスが黒い」という仮説のもとではこの確率が 5.3%となる。つまり p 値が 5%を下回るかどうか、というところで仮説を判断するのだとすれば、「97.1%のカラスが黒い」という仮説から「100%のカラスが黒い」というところまでの仮説はすべて否定しきれない。これが信頼区間の考え方。

こうした統計的仮設検定の考え方や p 値、信頼区間を使って、「いまデータが得られる範囲で 97.1%のカラスが黒い～100%のカラスが黒い」と考えて問題はなさそうだという結論を得ることができた。ちなみに、東南アジアには灰色のカラスがいるし、突然変異で真っ白の羽をもつアルビノのカラスというのもいるので、「すべてのカラスが黒い」ことはないのだが、少なくとも現実的な意思決定として「いま我々は、次に出会うカラスを黒と考えておいたほうがよさそうだ」ということは示せる。

ここで、「スポーツをすれば出世する」というような仮説を検証するためには、もう少し本格的な統計的仮設検定の方法を知らなければいけない。

検定

検定とは

「差がある」という仮説を検定する場合、差の程度が不明なため、そのままでは検定できない。そこで考え出されたのが、その逆の「差がない」という仮説を検定して、それに何らかの矛盾が見つかれば、もとの「差がある」という仮説を採用する。逆に、明らかな矛盾がないときにはその判定を保留する。このような論法で検定を行います。

ここで、「差がない」という仮説は本来「無」に帰すべきものとして「帰無仮説」(null hypothesis) と呼び、H_0と略す。また、もとの「差がある」という仮説は、「対立仮説」(alternative hypothesis) と呼び、H_1と略す。

仮説の設定、

例えば、2 つのグループ (A,B) 間で「差がある」という仮説について検定する場合、まず「2 つのグループ間には差がない ; A=B」とう仮説 (帰無仮説) を設定し、元々証明したい「差がある ; A≠B」という仮説 (対立仮説) は、一旦伏せておく。

例題) 具体的な例で学ぶ統計学の項目の最初に紹介した男子学生の集団 111 人の身長データを使って、検定を行ってみます。この集団の標本平均は 169.3cm、標本標準偏差は 5.4cm です。今、この世代の身長の全国平均が 165cm であるとすると、この集団の身長の平均が全国平均と異なるかどうかを検定する。

（この続きは、自分で以下に記入してください。）

有意水準（危険率）

実際の検定には、帰無仮説を誤って捨てる確率α（第１種の過誤の確率で危険率とも言う）を示し、そのときの検定統計量から確率を出して、

その値がαよりも小さいときに

↓

「有意水準αで統計的に有意である」といい、帰無仮説を棄却する。

検定統計量より得られる**実際の確率をＰ値**（P-value）と呼び、

Ｐ値がα以下となるような統計量の範囲を

棄却域という。

一般に有意水準αには5%（0.05)、1%（0.01）が用いられる。

統計的判断における２種類のエラー（過誤）

第１種の過誤（α）→　有意差があると判断した場合におこる。
本当は、帰無仮説が正しいのに、実際の確率Ｐがα以下のため、帰無仮説が違っているとして棄却する誤りを第１種過誤または、αエラーと呼ぶ。

第２種の過誤（β）→　有意差がない（判定保留）と判断した場合におこる。
本当は帰無仮説が誤っているのに、実際の確率Ｐがαより大きいため、帰無仮説を棄却しない、すなわち、対立仮説を採用しない誤りを第２種過誤または、βエラーと呼ぶ。
その誤りを起こす確率は、データ数に依存する。

片側検定と両側検定

正規分布やｔ分布を利用する場合、片側、両側検定の区別が問題なります。

両側検定の方が検定が厳密である。

例）正規分布（付表 1）を用いた片側検定で有意水準 0.05 に対するｚは 1.65 である。これは両側検定の場合の有意水準 0.1 の位置（ｚの値）に相当する。（両側検定で有意水準 0.05 に対するｚは 1.96）

したがって片側検定では、検定統計量の偏りより少なくても「統計的に有意」となるので、判定が甘くなる。

片側検定は、あらかじめ変化（差）の向きが理論的に片側にだけ起こると想定される場合に行う。例えば、降圧剤の効果を調べる実験で、投与後の血圧の上昇を想定しなくてよい場合に用いる。
両側検定は、処理効果がどちら向きの変化をもたらすかを予想できないときに用いる。通常は、変化の向きを予め予測できないとして、両側検定を用いる。

帰無仮説のもとでの検定統計量の分布の上限、あるいは下限の部分のみの確率を考慮して行われる検定方式を片側検定と呼ぶ。検定の有意水準をαとした場合、上側と下側の有意水準はそれぞれα/2 とするのが普通です。このような検定法は、両側検定（both sided test, two tailed test）と呼ばれている。検定の検出力は片側検定（one sided test, one tailed test）の方が高くなる。しかし、片側検定での対立仮説について、事前に確定的な情報が使用できればよいが、単なる思い込みによるものであれば、検定結果は過大に評価されるものとなる。事前情報が無い場合は、両側検定を行わなければならない。
検定の手順
ある 2 つのグループからそれぞれ抽出した標本について調べたところ、ある統計量に差があり、その差が偶然なのかそれとも意味があるのかを検定する場合の手順を以下に示す。
１）標本の分布などを仮定する（ここでは当然、正規分布になることを考えている）
２）「2 つのグループに差がない」と否定したい仮説「帰無仮説」をたてる。次に「2 つのグループには差がある」の帰無仮説を否定したときに成り立つ仮設「対立仮説」をたてる。
３） 偶然と判断する基準「有意水準」を決める（通常は 0.05（5%））。
４）帰無仮説のもとで、2 つのグループの標本の統計量の差を評価するための検定統計量（たとえば、ｔ値）の理論的な分布と有意水準に基づく棄却域を計算する。
５）標本から求めた検定統計量（ｔ値など）が棄却域に入っていれば、帰無仮説を棄却して対立仮説を採用する。帰無仮説が棄却できない場合は判断を保留する。
６）対立仮説を採用した場合には、具体的に差の内容を確認する。

【ここで、区間推定と検定を整理してみます】

母集団と標本；
集団の中である知りたい値を母数（parameter）という。母数を知ろうとする対象を母集団（population）と呼び、そこから抽出する一部を標本（sample）と呼ぶ。統計学では、その標本から母集団の母数を推定する。

正規分布；

　身長や知能指数のなどの分布は正規分布に従う。この分布は平均値を中心とした左右対称の釣鐘状の分布です。この分布によると、「平均値±1.96×標準偏差」の範囲にすべてのデータの 95%が存在するなどの性質が分かっています。この性質を利用して推定や検定を行います。

ｔ分布；

　標本数が少ない時には、正規分布の代わりにｔ分布を使って推定とか検定を行います。

区間推定；

「真の母平均は値 A と値 B の間（区間）に“おそらく”含まれているだろう」というような考え方で母平均の推定値をあらわすことができる。
このような考え方で、標本から母集団の母数を推定することを区間推定という。母集団の正規分布が仮定できる場合、”おそらく“の程度を信頼度（信頼係数）といい、一般に利用されるのは 0.95（95%）である。また、設定した信頼度（95%）で母数を含む区間のことを（95%）信頼区間という。

ｔ分布を使った区間推定；

データ数がｎ、標本平均が$\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i$、標本標準偏差が$SD=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$

のとき、次のように母平均μの取り得る範囲を推定することができます。μの 95%の信頼区間は

$$\bar{x} - t_0 \frac{SD}{\sqrt{n}} < \mu < \bar{x} + t_0 \frac{SD}{\sqrt{n}}$$

ここで、t_0はｔ分布表より求めることができます。自由度とはｎ－1 です。

検定（平均値の差の検定）；

2 つの学校における体重の差やヘモグロビン濃度（貧血検査）の差、あるいは 75g 経口ブドウ糖負荷後 2 時間値の 2 つの地域における差など、2 つの集団における平均値の差の検定を必要とする場合がしばしばある。
「差がある」という仮説を検定する場合、差の程度が不明なため、そのままでは検定できない。そこで考え出されたのが、その逆の「差がない」という仮説を検定して、それに何らかの矛盾が見つかれば、もとの「差がある」という仮説を採用する。逆に、明らかな矛盾がないときにはその判定を保留する。このような論法で検定を行います。

ここで、「差がない」という仮説は本来「無」に帰すべきものとして「帰無仮説」(null hypothesis）と呼び、H_0と略す。また、もとの「差がある」という仮説は、「対立仮説」（alternative hypothesis）と呼び、H_1と略す。

仮説の設定としては、

例えば、2 つのグループ（A,B）間で「差がある」という仮説について検定する場合、まず「2 つのグループ間には差がない；A=B」とう仮説（帰無仮説）を設定し、元々証明したい「差がある；A≠B」という仮説（対立仮説）は、一旦伏せておく。

実際の検定には、帰無仮説を誤って捨てる確率α（第 1 種の過誤の確率で危険率とも言う）を示し、そのときの検定統計量から確率Ｐ値を出して、

その値がαよりも小さいときに

↓

「有意水準αで統計的に有意である」といい、帰無仮説を棄却する。

検定統計量より得られる**実際の確率をＰ値**（P-value）と呼び、

Ｐ値がα以下となるような統計量の範囲を

棄却域という。

一般に有意水準αには 5%（0.05）、1%（0.01）が用いられる。
ｔ分表からｔ$_0$を求めるには、TINV（確率、自由度）で求める。
計算から求めるｔは、

$$\frac{\overline{x}_1-\overline{x}_2}{\sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

です。

パラメトリック検定とノンパラメトリック検定

パラメトリックとは、「平均・標準偏差・データ数」などのパラメータ（母数）を使った」という意味。例えば、A 校の学生 6 人と B 校の学生 4 人が 400m 競争をして勝負する場合、まず単純にそれぞれの平均タイムで勝負するという方法が考えられる。これは平均タイムを学校全体の様子を代表する値として利用するので、「パラメトリックな勝負」であるといえる。ただし、学生の中にたまたまオリンピック選手が 1 人まじっていたら、1 人だけ極端なタイム（外れ値）を出して平均に大きく影響してしまう。このような外れ値がある場合、平均などは学校全体の代表値としては不正確になる。

一方、ノンパラメトリックとは、「(平均などの) パラメータを使わない」という意味である。上と同じように 400m 競争を例にとると、タイムを無視してゴール順に 1 位 10 点、2 位 9 点、、、、、、10 位 1 点というように得点配分し、A 校と B 校の獲得した得点で勝負を決めるような場合が考えられる。この場合、比較に平均タイムなどを用いていないため、「ノンパラメトリックな勝負」といえる。このような方法の場合、オリンピック選手が大差で勝っても、ふつうの人がぎりぎり勝っても 1 位の 10 点にはかわりはないため、外れ値の影響を排除することができる。ただし、得られた平均得点は学校全体の代表値としては外れ値がない場合の平均タイムに比べて情報量が少なく、両校の差がどの程度あるのかはよくわからない。

検定方法の選択では、パラメトリックとノンパラメトリックのどちらを選ぶかは、標本から予想できる母数が信頼できるかどうかに相当する。なお、実用上は「標本から母集団が正規分布に従うと予想できるか否か」と考えても差し支えない。

・標本が正規分布をしている場合、平均・標準偏差・データ数の 3 つの値（パラメーター）で、分布の様子を決定できるため、この 3 つのパラメーターを用いた検定を行うことになる（パラメトリック検定）

・量的データであっても、外れ値などが存在し、正規分布が仮定できない（つまり、上述の 3 つのパラメーターから分布の様子が予想できない）場合には、これらのパラメーターを用いない検定を行うことになる（ノンパラメトリック検定)。ノンパラメトリック検定には、いろいろな種類があるが、例えばウィルコクソンの順位和検定のように、数値の順位を用いたものなどがよく利用される。

統計的仮設検定の考え方

統計的仮設検定（Statistical hypothesis testing）でも「すべて」を証明することはできないが、確率を導入することで「すべての」かわりに「ほとんどの」を考えられるようにする。まずは実際にデータを得ることができるように「言及する範囲」を決める。「いま自分たちが集められるデータの範囲」の中での仮設の妥当性に焦点を当てる。次に 100 か 0 かの全称的な議論ではなく、「ほとんどの」という仮説でもなく、自分の主張を完全に覆してしまうような「カラスが黒いかどうかは半々」という仮説について吟味する。実際のデータを基に、その「自説を完全に覆してしまう仮説」がどれだけあり得ない確率でしか成立しないことなのかを示すことができれば、自説が「さすがに完全に覆されるような主張とは言い難い」ことがわかる。さらに、完全に覆してしまう「黒いかどうか半々」というだけでなく、「9 割のカラスが黒かったとして」という少し自説に近い仮説についても吟味してみる。この「9 割のカラスが黒い」という仮説ですらあり得ない確率でしか成立しないのであれば、やはり 9 割以上のカラスが黒い、と考えた方が自然。

そして最後に、損か得かの問題に落とし込む。この統計的仮設検定の考え方は誰も損

も得もしないような悠久の真理を議論するうえで用いる意味はそれほど大きくない。だが、医学における命、教育学における生徒の学力、ビジネスにおけるお金、というように、損得のかかった状況でより最善な方を選ぼうとすると、統計的仮設検定は威力を発揮する。損得のかかる状況になれば、ぼんやり者とて確率的に妥当な方向へ意思決定の舵をきる。

母平均の検定①

全国 10 歳女子の平均と標準偏差は、それぞれ $\mu=140$ ㎝、 $\sigma=5$ ㎝の正規分布となる。今、ある 10 歳のクラス 25 人の身長の分布は、正規分布になるものとし、その平均値$\bar{x}$は 137 ㎝である。このクラスの身長は全国水準と違うと言えるのか。

（1）仮説設定；「全国水準と違う」という仮説は、違いの程度を特定できない。そこで、その逆の「全国水準と同じ」という仮説（帰無仮説 H_0； $\mu=140$ ㎝）を採用し、もとの仮説（対立仮説 H1； $\mu\neq140$ ㎝）を一旦、伏せておく。

（2）検定統計量を求める；この場合、25 人の身長の平均$\bar{x}=137$ ㎝を統計検定料とする。

（3）確率 P を求める；H_0が正しい場合、標本平均$\bar{x}$は平均値μ、標準偏差$\frac{\sigma}{\sqrt{n}}$の正規分布に従うことを利用する。 $z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{137-140}{\frac{5}{\sqrt{25n}}}=-3$、付表 1 から z=3.0 面積は 0.0013 となる。両側の面積は 0.0026 となる。つまり、確率 P は 0.26％となる。

（4）判定；H_0が起こる確率は 0.26％である。このようなまれな現象が実際に起こった考えるよりは、H_0が正しくないと考える方が妥当である。したがって、H_0を棄却して対立仮説 H1 を採用する。すなわち、そのクラスの身長は、全国平均と比べて差があり、それは有意水準 0.05 で統計学的に有意であると判定する。

（ここで、付表 1 の見方を以下に記入してください。）

母平均と検定②

ある看護学校の学生 15 人の身長を計測し、平均が 162.0 ㎝、標準偏差 SD が 5.0 ㎝であった。20 歳の全国平均は 157.5 ㎝である。この学生の集団は、有意水準 5%で、一般よりも身長が高いと言えるのか。

（1）仮説の設定；帰無仮説 H_0；μ＝157.5 ㎝)、対立仮説 H1；$\mu \neq$157.5 ㎝

（2）検定統計量を求める；この場合、15 人の身長の平均$\bar{x}$＝162.0 ㎝を統計検定料とする。

（3） t の値を求める；この場合、 t 分布表から有意性を検定を行う。

$$t = \frac{\bar{x} - \mu}{\frac{SD}{\sqrt{n}}} = \frac{162.0 - 157.5}{\frac{5.0}{\sqrt{15}}} \fallingdotseq 3.49$$

自由度 df=n－1=15－1=14 なので、付表 2 より面積が 0.05（5%）の t 値は 2.145 である。

（4）判定;計算で求めた t ＝3.49 の方が t 分布表より求めた t 0＝2.145 よりも大きい。したがって、帰無仮説は棄却され、この集団の平均身長は全国平均よりも高いと結論付けることができる。ただし、有意水準 0.05 である。

（ここで、付表 2 の見方を以下に記入してください。）

2つの集団における平均値の差の検定

対応のある場合とない場合とは？

t 検定には対応のあるデータに対する検定と、対応のないデータに対する検定の 2 種類があります。ここで、”対応のある“とか”対応のない“とはどういうことなのでしょうか？

その答えは、同一個体（人でも動物でも機械でもよい）におけるある処置の前と後の状態（前後のある測定値）を比較するのが対応のある場合となります。一方、異なる個体間である測定値を比較するのが対応のない場合となります。

例えば、手術の前後で体温が変化するかどうかを検定するため、5 人の患者の手術直前と直後の体温を測定したとします。この場合、1 人の患者に対して術前の体温と術後の体温の 2 つのデータが対になって得られます。このように 2 つのデータ集団が同一個体から得られる場合を対応があるといいます。

<u>対応のある場合</u>

例題）表のように、手術前後の体温の比較をしてみます。

表 12 対応のあるt検定の方法

患者	術前の体温	術後の体温	差	差の偏差	差の偏差平方
イ	36.7	35.5	−1.2	0.4	0.16
ロ	36.5	35.6	−0.9	0.7	0.49
ハ	36.5	35.6	−0.9	0.7	0.49
ニ	36.4	35.6	−0.8	0.8	0.64
ホ	36.6	35.1	−1.5	0.1	0.01

この場合、表のように同一個体の差を求め、その平均値、偏差、偏差平方、偏差平方和、標準誤差を求める。そして、t ＝差の平均値/差の標準誤差を求めて、t 分布表（付付表 2 の t 値と比較する。危険率 5%で有意差があるかどうか検定してください。下記に、差の平均、差の偏差平方和について、式を立てて、求めてみてください。その後、差の標準誤差について、式立てをしてください。なお、差の標準誤差の答えは 0.281、t は 2.84 です。ところで、自由度はこの場合、いくつですか？下記の空欄を使って答えてください。

対応のある場合の一般式

先程も説明しましたが、ある集団について、冬と夏の血圧を測定して両者を比較する場合や、同一人の左右の握力を測定して比較する場合、同集団で夏と冬の食事の内容（タンパク質、脂肪、摂取エネルギーなど）を比較する場合なども、2 変数は独立ではなく、対応のあるデータとなります。この場合に用いる t を求めるための一般式を下記に示します。

ここでは、冬の血圧値；x_1,x_2・・・・,x_n、夏の血圧値；y_1,y_2,・・・,y_n、各個人の冬と夏の血圧値の差；$z_1=x_1-y_1,z_2=x_2-y_2$,・・・,$z_n=x_n-y_n$、その平均を$\bar{z}$、標準偏差を SD_z とすると、 t は次のように求めることになります。

$t=\frac{\bar{z}}{\frac{SD}{\sqrt{n}}}$この値と t 分布表より自由度 df から求めた t_0 とを比較する。

対応のない場合

対応のない場合は 2 つのデータ集団の個体は同一ではなく、対応のある場合のように、2 つの測定値の差というものはありません。したがって、2 つの集団それぞれの平均値を比較することになります。

平均値の差の検定

2 つの学校における体重の差のように、2 つの集団における平均値の差の検定を必要とする場合がしばしば生じる。この場合、通常は 2 つの母集団の平均、標準偏差とも不明なことが多く、基本的には t 検定を行う。ここでは、問題としている変数ｘ、ｙ、その母標準偏差をそれぞれσ_x、σ_yとすると、次の条件があります。

（1） ｘ、ｙが独立であること

（2） ｘ、ｙがともに正規分布に従うこと

（3） $\sigma_x=\sigma_y=\sigma$であること

以上の条件のもとで t 検定を行います。例えば、母集団と標本が次の表のようだとすれば、次式が $df=n_1+n_2-2$ の t 分布に従います。

表 13　対応ない場合の母集団と標本の統計量

	母集団		標本		
	平均	標準偏差	平均	標準偏差	データ数
A グループ	μ_1	σ	$\bar{x}$	SD_x	n_1
B グループ	μ_2	σ	$\bar{y}$	SD_y	n_2

$$t = \frac{\overline{x} - \overline{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)SD_x^2 + (n_2-1)SD_y^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

今、男女 10 人の身長を測定すると、次のようになった。

男子；178、168、170、174、164、171、169、171、170、180

女子；155、162、159、147、162、151、160、151、162、161

男子； $\overline{x}$ ＝171.5、SD_x＝4.72

女子； $\overline{y}$ ＝157.0、SD_y＝5.58

ここで、帰無仮説 H_0；$\mu_1=\mu_2$、対立仮説 H_1；$\mu_1\neq\mu_2$ として検定すると t の値は次のようになる。

$$t = \frac{171.5 - 157.0}{\sqrt{\frac{(10-1)4.72^2 + (10-1)5.58^2}{10+10-2}\left(\frac{1}{10}+\frac{1}{10}\right)}} = 6.28$$

となる。大学生では男は女より体格が大きいので、片側検定で行ってみる。t 分布の自由度 df は 18、有意水準 α＝0.05 のとき、片側検定なの付表 2 の α が 0.1 のところを見て、自重度 df＝18 でみる。そうすると、t_0 は 1.734 を読み取ることができる。計算で求めた t 値は 6.28 で t_0 よりも大きい。したがって帰無仮説は棄却され、男の身長は女よりも大きいと結論付けることができる。

演習問題

1．次の資料は、男女学生の体重です。体重が正規分布に従うと仮定して、t 分布により、男女の体重（μ 、μ）が等しいという仮説を検定してください。

　ポイント：片側検定で行う。有意水準は 0.05 とする。

男：65，58，53，63，70，68，48，62，70，56

女：51，53，47，42，49，51，55，51，55，52

男の平均＝61.3、女の平均＝50.6、男の標準偏差＝7.44、女の標準偏差＝3.89

t＝4.03

２．次の資料は妊娠前と妊娠後の体重です。妊娠前後の体重に差がないという仮説を検定してください。

妊娠前：52.0，51.0，42.0，51.0，42.0，49.0，51.0，44.0，46.1，47.0

妊娠後：61.5，57.3，50.7，60.2，52.5，53.6，57.1，56.0，55.6，54.7

差の平均は-8.41、差の標準偏差は 2.24、ｔ＝-11.87

<u>少ないデータのための t 検定とフィッシャーの正確検定</u>

t 検定の「t」は「test」に由来しているらしい。ｔ検定の発明者であるウィリアム・ゴセットはオックスフォード大学で化学と数学を専攻し、それに回帰分析や相関係数の発明者であるカール・ピアソンのもとで統計学を学んだ。彼の仕事はギネス社で統計学や化学の知識を生かし醸造工程と原材料を改良する事であった。つまり、彼は民間で働く統計家として最も古い時代の人物です。彼の発明した t 検定のことを「student の t 検定」と呼ばれています、これは彼が会社に秘密で研究成果を公表するために student というペンネームを用いたことに由来します。

z 検定と t 検定の基本的な考え方は共通しており、どちらも「平均値の差」が「平均値の差の標準誤差」の何倍になるのか、という値が確率的にどれほどあり得ないかを示す p 値を求める。

理論上、分散の「真の値」とはデータの「真の平均値からのズレの二乗の平均値」です。ただし実際には「真の平均値」はわからないので、「データの平均値」との差の二乗を用いて*計算された推定値を使う*。しかし、データが少なければ少ないほど「データの平均値」は「真の平均値」から離れた値になってしまいがちである。このため、データが少なければ少ないほど、サンプルの分散は「真の分散」より小さめの値になってしまうし、サンプルの分散を用いて計算されたサンプルの標準誤差も当然小さめの値になってしまう。

例えば、成人男性の身長（真の平均値が 170 ㎝）、得られた 3 人のサンプルがそれぞれ 172、174、176 ㎝。平均は 174 ㎝。「真の平均値」からのズレとして分散を考えた場合と、「データの平均値」からのズレとして分散を考えた場合とどうなるのか。「真の

平均値」からのズレで計算すると約 18.7.「データの平均値」からのズレで計算すると約 2.7.「データの平均値」がデータから計算される以上、データの値とその平均値は独立したものではない。データがたまたま本来の分布の中で大きなのもに偏っていれば当然その平均値も大きなものになるし、逆もまた成立する。したがって、「データの平均値」と「真の平均値」が一致していない限り、「データとそこから求めた平均値」は「データと真の平均値」よりも全体としては必ず近い値になる。それが限られたデータだけで分散を求めるとどうしても値が小さくなりがちであるという理由である。

そこでゴゼットとフィッシャーは「データから求められた分散と、そのデータの数の間にどのような関係があるか」を数学的に整理した。測地学の研究で有名なヘルメルトによって発見され、カール・ピアソンによって名付けられた χ^2 分布を用いれば、計算に用いたデータの数ごとに異なる、データから求められた分散が真の分散からどの程度異なるものになるのか、という分布が計算できると明らかにした。χ^2 分布というのは、平均値が 0、分散が 1（すなわち標準偏差も 1）の正規分布に従う X という変数を考えた時、この変数の二乗をいくつか足し合わせたものが従う分布である。この χ^2 分布は、足し合わせる X 二乗の数（自由度）によって分布の形状が異なる。自由度が無限大の時には正規分布に完全に一致し、数百～数千という大きな自由度では「ほぼ正規分布」と呼んで問題ない。このカイ二乗分布の自由度ごとに、「平均値の差」が「平均値の差の標準誤差」の何倍以内に収まる確率が何%となるのか、を計算するための分布が t 分布である。

<u>データの数が限られた場合は「フィッシャーの正確検定」</u>で

クロス集計でみると、どのセルにもできれば 10、最低でも 5 以上の数字が入る場合は z 検定を行って問題ない、というのが慣例的な目安である。

なお、データ数が極端に少ない場合、フィッシャーの直接確率検定を使う手がある。「直接確率」とは、正規分布への近似ではなく、正確な確率計算を用いて p 値を算出するという意味である。

表 14　データが少数の場合

	主任以上	役職なし	合計
体育会出身	4 人（66.7%）	2 人（33.3%）	6 人
その他	1 人（25%）	3 人（75%）	4 人
合計	5 人（50%）	5 人（50%）	10 人

計算してみると、出世者 5 人中 4 人が体育会出身者となる p 値は 26.2%、ただし体育会の出世率のほうが高い場合のみを考える片側検定の p 値である。

「出世者 5 人中 4 人が体育会出身」という実際に得られた組み合わせが偶然得られる確率は 23.8%、これ以下の確率でしか起こらない状況として、「出世者 5 人中 2 人だけが体育会」と「出世者 5 人中 1 人だけ体育会」を足し合わせ、合計 52.4%（5 人の確率

2.4、4人23.8、2人23.8、1人2.4%）というのが両側検定のp値である。つまり、たった10人のデータから、実際に得られたデータあるいはそれ以上に起こりにくいような体育会とその他出身者の出世率の差が偶然に得られる確率は 52.4%、このように 2回に1回以上の確率で偶然生じるような程度の違いであれば、当然「たまたまそうなっただけかもしれない」と疑ってかかったほうがいい。

t検定について最低限知っておけばよいこと

t検定とは数十件程度のデータでも正確にz検定を行えるようにしたものであり、数百～数千件といったデータに対してはt検定とz検定の結果はよく一致する。

t 検定は、z 検定と同様に「平均値の差」が「平均値の差の標準誤差」の何倍かを考えてそれがどれほどあり得ないかp値を求めるものである。

フィッシャーの正確検定は「組み合わせの数」を使って数十件程度のデータでも正確に割合に意味があるのかp値を求めるものである。

もう一度、検定とは

t検定は、1群の平均値について、母集団の平均との比較をするとき、あるいは2群の平均値の差について検定するときに使う。なお､2群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は 3 群以上の平均値の差の検定に使う。F 検定は等分散性の検定に使う。これは2群の母平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル）、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに4以下の時は、フィッシャーの直接確率計算法を利用する。

もう一度、パラメトリックな検定（正規分布を仮定している）

t検定は、1群の平均値について、母集団の平均との比較をするとき、あるいは2群の平均値の差について検定するときに使う。なお､2群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は 3 群以上の平均値の差の検定に使う。F 検定は等分散性の検定に使う。これは2群の母平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル）、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに4以下の時は、フィッシャーの直接確率計算法を利用する。

もう一度、ノンパラメトリックな検定

平均値の差の検定やｔ検定は、正規分布を仮定して成り立つもの（パラメトリック法)。しかし、統計的な検定を行う場合、このよう母部集団の分布を仮定して検定することができる場合がすべてではない。例えば、質問紙調査の回答としてしばしばみられる、「１．とてもそう思う」「２．ややそう思う」「３．あまりそう思わない」「４．まったくそう思わない」といった順位尺度データの場合には、2 つの質問項目の回答に有意な差があるかどうかについて知る方法として、母集団の分布形を仮定しないノンパラメトルック法を用いる。独立 2 標本の代表値の差の検定は、マンホイットニー法（U テスト)、対応のある 2 群の代表値の差の検定はウィルコクソン（の符号付順位和)検定を使う。2 群以上の代表値の差の検定にはクラスカルウォリス検定を使う。抽出した標本の元となる母集団と特定の理論分布が一致するかどうかを検定するにはコルモゴロフスミルノフ法（K−S 法）を使う。このとき、累積相対度数を比較する。また、正規分布からのずれの大きさによって外れ値の有無を判定するには、グラッブス・スミルノフ棄却検定を使う。

相関

相関係数は 2 つの変量ｘ、ｙの関連の強さを表す指標で−1 から１の間の値を取る（負の相関、正の相関、相関なし)。また、両者に直線的相関が想定できるとき、2 変数の関係は、直線の式で表すことができる。 ｙ＝bx＋a　　（この式の場合、ｙが従属変数、ｘが独立変数になる）

この直線を回帰直線、直線の式を回帰式という。また、この式の係数の決定方法が最小 2 乗法である。

多変量解析

複数個の変量をもつ多変量データを変量間の相互関係を考慮に入れて分析する一連の統計的手法が多変量解析である。医学、看護学分野では、1 人の人間がもつデータは、体格、血圧をはじめ、血圧データ、心理データなど、単に 1 時点をとってみても、非常に多数あり、そのほとんどは相互に関連しあっているといえるもの。したがって、人間に関する医学、看護学データの分析にとって、多変量解析は必要不可欠なもの。

主成分分析は、多くの変量が与えられたとき、それらの変量のもつ情報をできるだけ多く表現できるような合成得点を求める分析方法。例えば、人の「身長」「体重」という 2 つの変量を考えると、身長も体重も大きい人は「体格」がよいとされる。このことから、これら 2 変量をもとに「体格（の良さ、悪さ)」という主成分を求め、「体格」だけで、「身長」「体重」の 2 変量がもつ情報をある程度表現しようとしたもの。

因子分析は、数多くの変量の中に潜む共通の因子を探り出す分析方法。例えば、上記

の例で人の「身長」「体重」「胸囲」などの変異から、「体格増大の素因」「肥満傾向」などの基礎として隠れた因子を探り出すもの。典型的なのものには、性格特性の分析がある。

重回帰分析は、ある変数を、いくつかの変数によって予測する式を作成する。

判別分析は、いくつかの変量をもつ個体が、いくつかの母集団のうちのどれかに属しているかを判別する分析方法。変量間の相関関係をもとに、母集団ごとの差が明確に現れるような値が得られるように各変量に重みをつけ、そうして得られた重みづき得点により母集団を判別する。

クラスター分析は個体のもつ変量の類似度をもとに、個体をいくつかの集団（クラスター）にまとめ上げていこうという手法の総称。類似度は、個体間の距離の算出。

Ⅵ　分散分析（Analysis of Variance）

英語の原語を略して通称 ANOVA（アノヴァ）と呼んでいます。分散分析も t 検定と同様に平均の有意差の検定をします。t 検定は 2 グループの平均しか比べることができないのですが、分散分析は 3 グループ以上の平均も比べることができます。

分散分析の考え方：

ここでは簡単のため、2 条件 A・B の平均の有意差検定について考えてみます。分散分析では、条件 A と条件 B のデータを同一の母集団からの標本とみなし（帰無仮説）、この母集団の真の値の推定値として条件 A と条件 B の「平均の平均」を求める。すなわち、（X_A＋X_B）/2。これを大平均と呼ぶ。分散分析はこの大平均からのズレに注目する。そうすると、

データの大平均からのズレ＝平均の差の影響によって生じたズレ＋偶然の影響によって生じたズレ

分散分析は、このようなズレの分解を全データについて行う。そして、データのズレ（すなわちデータの値）を支配したのは、平均の差の影響力なのか、それとも偶然の影響力なのかを比較する。もし、前者の影響力が大きければ、平均の差は文字どおり偶然以上であることになる（有意である）。

Ⅶ　相関・予測

相関・予測の分析は条件間の関連を見ます。したがって、各条件のデータどうしは独立であってはならず、対応のあるデータでなければならない。対応のあるデータを入手するには、1 人の被験者または対象者から 2 個以上のデータをとればよい。例えば、1 被験者を身長と体重である。このとき、データの種類や単位は違っていてもかまわない。

相関・予測の分析は、このような対応するデータどうしの間に、次のような 3 点の規則的関係を見出そうとしている。

① 正の相関

いわゆる正比例の関係があるかどうかを見る。すなわち、一方のデータの数値が大きい場合に、それに対応する他方のデータの数値も大きい（小さい場合には小さい）なら、正の相関がある。

② 負の相関

いわゆる逆比例の関係があるかどうかをみる。一方のデータの数値が大きい場合に、それと反対に他方のデータの数値が小さい（小さい場合に大きい）なら、負の相関がある。

③ 予測式

データどうしの相関関係を、一次方程式として表現する。この方程式に一方のデータの任意の数値を代入すれば、他方のデータの数値を予測的に求めることができる。

変数

変数（variable）とは「変化する数値」という意味である。相関・予測の分析では、データを変数とみなす。前頁の例では、サーブの練習本数のデータが一つの変数であり、成功率のデータがもう一つの変数である。したがって前例の場合は「2 変数の相関と予測を分析した」という言い方をする。このように、変数という言い方をするのは、相関・予測の分析がデータどうしの対応関係を数学の関数関係になぞらえて分析しようとするからである。

予測と回帰

予測は回帰ともいわれる。回帰（regression）とは「もどす」ことであり、予測関係を逆方向から見たときの言い方である。

このように、予測も回帰も、言っていることは同じであり、見る方向が異なるだけである。ここでは、「予測」という言い方を多用するが、専門的には「回帰」という言い方が好まれる。

<u>相関係数の計算</u>

変数間の相関・予測が直線的であることをチェックしたら、次に相関係数を計算する。相関係数とは、2 変数の相関の強さと方向を表す統計量である（3 変数以上の場合には重相関係数がある）。この統計量は K.Pearson の考案によるので、特に「ピアソンの相関係数」と呼ぶこともある。

<u>記号</u>

相関係数の記号は“ r ”である。なお、r はデータの相関係数である。このデータの背後に存在する無限データ集団（母集団）の相関係数を表すときには、r に相当するギリシャ文字“ ρ ”を用いる。

ｒの性質

ｒの値は0をはさんで、－1～＋1の範囲で変化する。ｒの値が－の場合は、「負の相関」であり、散布図上のデータは右下がりの直線に収束する傾向を示す。これに対して、ｒの値が＋の場合は「正の相関」であり、散布図上のデータは右上がりの直線に収束する傾向を示す。完全相関（ｒ＝－1、ｒ＝＋1）の場合は、全データが一本の直線上に乗るが、それ以外は、直線への収束には程度の違いがある。なお、ｒ＝0は「無相関」であり、データはどのような直線へ収束するきざしもみせない。

相関係数の計算

変数Aと変数Bのデータを、下表のように入手したとする。

表15　データ・リスト

対象者	変数A	変数B
a	1	6
b	2	4
c	3	5
d	4	3
e	5	1

表16　データ表示(N=5)

	変数A	変数B
$\bar{x}$	3.0	3.8
SD	1.4	1.7

① 偏差積和を計算する。

偏差とは、データと平均との差のことです。偏差は各対象者ごとに2個できるので、それを掛け合わせて（積)、全被験者について合計した値が「偏差積和」となります。

偏差積和＝（1－3.0）×（6－3.8）＋（2－3.0）×（4－3.8）＋（3－3.0）×（5－3.8）＋（4－3.0）×（3－3.8）＋（5－3.0）×（1－3.8）

＝（－4.4）＋（－0.2）＋（0）＋（－0.8）＋（－5.6）

＝－11.0

ここでは、たまたまマイナスの積が多く出たので、和もマイナスになりました。この「偏差の積」のイメージとしては各データと平均とのズレを面積として表現したものと考えることができます。

① データ1組分の偏差積和に変換する。

上で計算した偏差積和は、変数Aと変数Bの対応するデータ5組分の値です。これを1組分の値にします。この値が、共分散といいます。

$$\text{共分散}=\text{データ1組分の偏差積和}=\frac{\text{偏差積和}}{N}=\frac{-11.0}{5}=-2.2$$

② 下式によってｒを計算する。

$$r=\frac{\text{データ1組分の偏差積和（共分散）}}{SD_A\times SD_B}=\frac{SD_{AB}}{SD_A\times SD_B}=\frac{-2.2}{1.4\times1.7}=-0.92$$

ここで、相関係数 r ＝－0.92 は二つの情報をもっている。一つは値の大きさ(0.92)が意味する、もう一つは値の符号（マイナス）が意味することです。

r の値の大きさが意味すること

上の r の計算式における分子の「データ 1 組分の偏差積和＝共分散」とは、変数 A と変数 B のデータ 1 組分の偏差で出来る平均的面積です。これに対して、分母の「$SD_A \times SD_B$」は各変数の標準偏差を組みとなしたときに出来る面積です。したがって、データの各組の値のとり方が「$SD_A : SD_B$」と同比率であれば、分子は分母に収束し、「r ＝±1」となります。このとき、散布図上のデータの点は完全に直線上に並ぶ。このように変数 A と変数 B の標準的な座標（SD_A、SD_B）を仮定して、実際のデータの組みがそれにどの程度、近似した値のとり方をするか、という程度を面積比とした値が r です。 r ＝－0.92 とは完全な直線に対する 9 割程度の近似を意味しています。

r の値の符号が意味すること

次に、 r ＝－0.92 はマイナスに出ているが、これは偏差積和がマイナスになっていたからです。マイナスの積を出すデータの組が多かった (①においてデータ 5 組のうち 4 組がマイナスの積を与えた)。このように、マイナスの積が多く出るということは、平均の座標からみて、その左上と右下にデータが集まっていることを意味しています。

さて、検定の結果は、有意でした（F（1,3）＝16.53、 p ＜.05)。すなわち、 r ＝－0.92 は「真の相関係数 0」＋「偶然のユレ」としてはめったに出現することのない値です。したがって、実質的に相関係数は 0 でないと認めてよいであろう。しかし、この有意性は、その相関が強い相関であることまで保証しない。有意であっても、取るに足らない弱い相関もある。そこで、有意性検定において有意であった場合は、次の相関の強さを判定する必要がある。

判定の基準

相関の強さの判定は経験的であり、一般に、以下の基準を用いている。

注意：相関係数の値と相関の強さの関係

先の判定基準からわかるように、相関係数の値と相関の強さとは直線的関係にない。例えば、「弱い相関」の段階は 0.2～0.4 の 2 ポイントの幅であるが、「中程度の相関」の段階は 0.4～0.7 の 3 ポイントの幅にとってある。したがって、相関係数の値が二倍であるからといって、相関の強さも二倍であると言えない。この点、相関の強さを正確に理解したいなら、説明率または決定係数（coefficient of determination）といわれる指数を計算しておくほうが、誤ったイメージをもたないですむ。

説明率の計算

説明率とは、一方の変数が他方の変数をどのていど説明するかを表す統計量です。この説明率と相関の強さとは同義である。専門用語では「決定係数」と呼ぶこともある。説明率の計算は、下式による。

$$R^2 = 説明率（\%）= r^2 \times 100$$

例えば、$r = 0.35$ の説明率は $0.35^2 \times 100 = 12\%$ である。$r = 0.35$ は $r = 0.70$ と比べると相関係数としては二分の一であるが、説明率を計算してみると、実は相関の強さとしてはもっと弱く、四分の一以下であることがわかる（$r = 0.70$ の説明率は 49%）。ちなみに、先の判定基準を説明率に置き換えてみると、強い相関は 50%以上、中程度の相関は 15~50%、弱い相関は 5~15%、ほとんど無相関は 5%以下となる。

e. 予測式の算出

分析の目的

予測の分析の目的は、一方の変数の値から他方の変数の値を予測する方程式を求めることである。この方程式を予測式または回帰式という。予測と回帰のちがいは、たんに見方の違いである。例えば、「変数 A を変数 B から予測する」は「変数 A を変数 B に回帰させる」ことでもある。以下「予測」に統一するが、「回帰」に読み替えてもよい。

予測式の説明

さて、変数 A を変数 B から予測する数式は、一般に、次のような形になる。

$$A` = k \cdot B + y$$

この予測式は、散布図上では、データが収束すべき直線を表している（この直線を予測直線という）。式中の A`は、変数 A の予測された値であり、必ずしも実際の値と一致しないので、A とは区別して表している。k はいわゆる「傾き」である。k が＋なら予測直線は右上がり、k がーならば予測直線は右下がりになる。専門用語では、「回帰係数」と呼び、変数 A が変数 B に依存する(回帰する)程度とみている。なぜなら、k の値が大きければ大きいほど、変数 A の予測値（A`）は変数 B が 1 ポイント上下するたびに大きく変化するからである。また、このため k を「重み」と表現することもある。重みが大きいことは、変数 A に及ぼす変数 B の影響が重大であることを意味している。予測式の最後の y は定数であり、いわゆる「y 切片」である。すなわち、予測直線がタテ軸を切る点を表している。

3 変数以上の相関・予測の分析

これまで、2 変数の相関を分析してきた。今度は、もっと変数の多い場合、すなわち 3 変数以上の相関を分析してみる。

3 変数以上の相関を分析するには、以下の 2 つのコースがある。

コースⅠ：「相関マトリスクの作成」→　「因子分析」

このコースは、変数間に特定の**予測関係が存在しない場合**に選択する。基本的には、2 変数ずつを取り出して、一組ずつ相関係数を計算する（相関マトリスクの作成）。また、その結果に基づいて、多くの変数を少数の因子にまとめる「因子分析」へ発展することもできる。

コースⅡ：「重相関係数の計算」→　「回帰分析」

このコースは、変数間に特定の**予測関係が存在する場合**に選択する。すなわち、全変

数のうち、ある1個が目的変数、あとの残りがすべて予測変数と最初から決まっていなければならない。このコースを選択すると、まず「重相関係数」を計算し、次に予測式を算出する。ただし、この予測式の算出は「回帰分析」として行われるのが一般的である。すなわち、目的変数の予測に貢献する予測変数と貢献しない予測変数を、複数の候補の中から選択しようとする。

このコースⅠは、まず「相関マトリスク」を作成し、次に「因子分析」へ移行することができる。ただし、因子分析への移行の際には、変数の個数が10個以上（できるなら20個以上）は、ほしい。まず、相関マトリスクから説明する。

相関マトリスクの作成

相関マトリスクとは「相関行列」という意味であり、多くの相関係数をヨコ（行）とタテ（列）に並べた表である。相関マトリスクを作成するには、いわば全変数の「総当たり表」を作って、この表のなかで「ぶつかる」2変数どうしの相関係数を一つずつ計算してゆけばよい。例えば、下の表13は4変数の相関マトリスクの例である。

表17　4変数の相関マトリスク（N＝42）

	B	C	D
変数Ａ	.432**	.610**	.279†
変数Ｂ		.214	.186
変数Ｃ			.378*
†p<.10 *p<.05 **p<.01（両側）			

表中の各相関係数に対しては、個別に有意性検定（両側検定）を行う。そして、その結果を記号を付けて表すこと。

因子分析（Factor Analysis）は、多くの変数の相関関係を、少数の因子に要約するための方法です。ただし、そのような多変数の動きを支配する共通の因子がみつかるかどうかはケース・バイ・ケースであり、探索的です。

重相関と重相関係数の定義

重相関（multiple correlation）とは、変数が3個以上の場合で、かつ、この変数間にあらかじめ一定の予測関係が存在する場合の相関関係のことである。すなわち、全ての変数のうち、特定の1個が目的変数、あとの残りが予測変数と最初から決まっていなければならない。また、重相関係数とは、そのような3変数以上の間の相関関係の強さを表す統計量である。重相関係数の記号には一般に“R”が用いられる。これは2変数の場合の相関係数の記号”r“の大文字である。

ａ．重相関係数の計算

Rの計算

3変数Ａ・Ｂ・Ｃがあり、予測関数が明確に存在する（大前提)。ここでは、Ａが目的

変数、B・C が予測変数であるとする。

1）データ表示を行う。

表 18　各変数の平均と標準偏差（満点各 20 点、(N＝50)

	条件 A	条件 B	条件 C
$\overline{x}$	14.9	15.7	13.8
SD	7.6	5.2	6.9

2）2 変数ずつ取り出し、相関係数 r を計算する。

ここでは、すでに計算ずみであり、以下のとおりとする。なお、条件 A の全データを変数 A ということにする。以下同様。

変数 A と変数 B の相関係数；r_{AB}＝0.512
変数 B と変数 C の相関係数；r_{BC}＝0.356
変数 A と変数 C の相関係数；r_{AC}＝0.434

3） r の値を下式に代入し、重相関係数 R を計算する。

$$R=\sqrt{\frac{{r_{AB}}^2+{r_{AC}}^2-2\times r_{AB}\times r_{BC}\times r_{AC}}{1-{r_{BC}}^2}}=\sqrt{\frac{0.512^2+0.356^2-2\times 0.512\times 0.356\times 0.434}{1-0.434^2}}$$

$$=0.533$$

重相関係数 R の性質

① R は 0 から 1 までの正の値しかとらない。
重相関は 3 変数以上の対応関係であるため、その方向性をプラス・マイナスだけに限定することはできない。したがって、R は、重相関の強さだけを表す。そのように、R は強さの情報しか持っていないので、論文には R そのままでなく、強さの意味を表す R^2 を掲載するのが通例である。

② 予測関係が変わると R の値も変わる。
例えば、「目的変数 A、予測変数 B・C」としたときの R の値と「目的変数 B、予測変数 A・C」としたときの R の値とは一致しない。この点、変数間の予測関係が事前に確定していないと、なにを求めているのかわからなくなる。

③ 使用上の制約は r と同様である。
各変数間の相関の直線性、データ分布の正規性、目的変数の（各予測変数に対する）等散布性が満たされないと、R は使用できない。

b．偏相関係数の計算

偏相関の定義

偏相関（partial correlation）とは「部分的相関」という意味である。すなわち、重相関における複数の予測変数のうち、特定の 1 個だけと目的変数との相関関係を表す用語である。例えば、変数 A・B・C があり、A が目的変数、B・C が予測変数である時、この 3 変数は以下のような相関関係によって結ばれている（矢印は予測の方向）。

このとき、例えば、目的変数 A に対する予測変数 B だけの純粋な相関（上図の太線

部分）をみるには、このAとBの相関に加わる変数Cの影響力（AとCの相関、BとCの相関）を取り除かなければならない。偏相関とは、そうした他の変数の影響を取り除いたときの、目的変数と予測変数1個との部分的相関を意味している。なお、上図からわかるように、3変数の場合、偏相関係数は2個求めることができる（AとB、および、AとC)。

偏相関係数の記号

偏相関係数の記号は、通常の相関係数の記号rに特別な添え字を付ける。例えば、“$r_{AB.C}$”のように書く。この添え字「AB.C」の意味は、「変数Aと変数Bの偏相関・変数Cの影響を取り除いたときの」ということです。

<u>計算例</u>

重相関係数を計算したときの例を用いて、さらに偏相関係数を計算してみる。前例では、目的変数A、予測変数B・Cであった。

① 各2変数の相関係数を計算する。

$r_{AB}=0.512$、 $r_{BC}=0.434$、 $r_{AC}=0.356$

② 各相関係数を下式に代入し、偏相関係数を計算する。

まず、変数Aと変数Bの偏相関係数を計算する。

$$r_{AB.C}=\frac{r_{AB}-r_{AC}\times r_{BC}}{\sqrt{1-r_{AC}^2}\times\sqrt{1-r_{BC}^2}}$$

$$=\frac{0.512-0.356\times0.434}{\sqrt{1-0.356^2}\times\sqrt{1-0.434^2}}=0.425$$

次に、同様にして、変数Aと変数Cの偏相関係数を計算する。

$$r_{AC.B}=\frac{r_{AC}-r_{AB}\times r_{BC}}{\sqrt{1-r_{AB}^2}\times\sqrt{1-r_{BC}^2}}$$

$$=\frac{0.356-0.512\times0.434}{\sqrt{1-0.512^2}\times\sqrt{1-0.434^2}}=0.173$$

<u>**c. 予測式の算出と回帰分析**</u>

<u>目的</u>

重相関係数と偏相関係数を計算したら、最後は予測式を算出する。3変数以上の場合に、予測式を求める目的は二つある。１）目的変数を予測すること、および２）予測変数を選別することである。このうち、最初の１）はふつうの目的であり、目的変数の値の変化に関心がある。しかし、一般に研究例は少ない。これに対して、次の２）の目的は、予測変数どうしの有効性の比較に関心がある。すなわち、目的変数を予測するのに、どの予測変数が有効か、どの予測変数が有効でないかをより分ける。この目的をとった場合は実験計画法に似ており、けっきょく「効く要因」の検討になる。3変数の場合を例にとると、そのときの予測式の形は、例えば以下のようになる。

A‘＝1.38B＋0.77C＋11.6

上式の傾き（1.39、0.77）が予測変数 B と C の効き具合を表すことになる。ここでは、変数 B のほうが有力である。すなわち、1 ポイント変化したときに目的変数 A の値を上下する度合いが大きい（1.39＞0.77）。

なお、予測変数 B・C が共に 0 のときも、目的変数 A は 11.6 程度の値をとる。これは、なんの説明もできないので、偶然誤差によるユレであるとみなすしかない。かくして、上式は、実験計画法における分散の分解式に類似した意味を持つ（下式）。

データ A‘＝変数 B の効果＋変数 C の効果＋偶然誤差

このように予測式をデータの模型にみたてて、予測変数どうしの重み（傾き）の大きさを比較する方法を回帰分析と呼ぶ。なぜなら、分析の焦点が目的変数でなく予測変数にあるので、視点は「目的変数←予測変数」（予測）という方向でなく、「目的変数→予測変数」（回帰）という方法になるからである。したがって、回帰分析を目的とした場合の予測式は「回帰式」と呼ぶ方が適切であり、専門家も、その呼び方を好む。

参考文献
1）研究者のための統計的方法、R・A・フィッシャー著、遠藤健児・鍋谷清治共訳、森北出版株式会社、1979.
2）統計学が最強の学問である、西内啓著、ダイヤモンド社、2013.
3）統計学が最強の学区門である[実践編]、西内啓著、ダイヤモンド社、2014.
4）統計学　系統看護学講座（基礎分野）、金森雅夫著、医学書院、2013.
5）宇部フロンティア大学における長期履修学生制度実践と課題、松本治彦著、私学経営（366）、17－25、2005.
6）学生満足度調査に関する一考察、松本治彦、佐藤美幸、網木政江、白石義孝著、宇部フロンティア大学人間社会学部紀要、Vol.1 No.1、宇部フロンティア大学出版会、2010.
7）複数連結汽水湖の水温、塩分の変動、松本治彦、合屋晏秀、李寅鉄、羽田野袈裟善、斎藤隆著、水工学論文集 39、237－242、1995.