統計学セミナー　第５回資料

北海道対がん協会　細胞診センター検査科　和田　恒之

　統計学的仮説検定の説明を前回まで行ってきたが、今回は相関と回帰について取り上げる。
相関と回帰は同じようなものとしてとらえられがちだが、相関は変数間の関連の強さを表したもの、回帰はある変数のばらつきがどの程度他方の変数のばらつきによって説明できるかを示すという違いが有り、相関には関連の向きは問われないが、回帰では必ず向きが存在する。この二つが相関と回帰で必ず覚えておかなくてはならない点である。
一般的に散布図を示しそれに直線を引いて、相関係数、y=bx+a の式を横に記して相関図として提示されることが多いと思うが、この y=bx+a は回帰式である。よって相関図の中に回帰式を提示することはその意味合いから考えると変では有るが。計算の結果、回帰と同じ式が導かれてしまうことがこのような状況になっていると思われる。

相関 (Correlation)

　統計学的仮説検定よりも多くの人になじみがある言葉だと思うが、改めて相関の定義を確認しておきたい。変数同士の関連の強さを表すことを相関といい、「かなりの程度の規則性をもって、増減を共にする関係」を相関関係と呼ぶ。一般的に相関を論ずるときに変数の数は二つのことが多いと思うが、これを単相関と呼び、三つ以上の変数をもって関連を論ずるときは重相関と呼び方が区別されている。
　本稿では単相関の説明を主とする。

単相関（以下、相関と記述する）
　変数同士の関連の強さを見るのに判りやすいのが増減どちらかの傾きを持つ直線的関係だと思う。実は相関の定義には必ずしも直線である必要はなく二次、三次曲線のような変曲点を持つ曲線でも、またその他の階段状の直線であっても相関関係は成り立つ（規則性が認められれば良い）、それを数学的に記述すると対象とするモデルによっては非常に複雑な式になるものや、相関の評価が難いといった欠点が有るので一般的には増減に関して直線関係が有るときに相関を論ずることが多い。

相関係数

相関では相関係数が相関の強さを見る指標となる。そして、その相関係数には

1　ピアソンの積率相関係数（Pearson's Correlation Coefficient）　r
2　スピアマンの順位相関係数（Spearman's rank Correlation Coefficient）　r_s もしくは ρ（ロー）
3　ケンドールの順位相関係数（Kendall's rank Correlation Coefficient）　τ（タウ）

以上の三つがあるが、ピアソンの積率相関係数とスピアマンの順位相関係数がよく使われている。

ピアソンの積率相関係数（Pearson's Correlation Coefficient)
　相関係数といえば r だが、厳密にはこのピアソンの積率相関係数が r と表現される。この相関係数 r は二変数の直線的な関係の強さを示す指標となっている。さらにピアソンの積率相関係数を計算するには二変数が二次元正規分布に従っていなくてはならないという前提条件がある。
　この説明で勘のよい方はパラメトリック検定をイメージするかもしれないが、このピアソンの積率相関係数はまさしくパラメトリックな相関係数なのである。

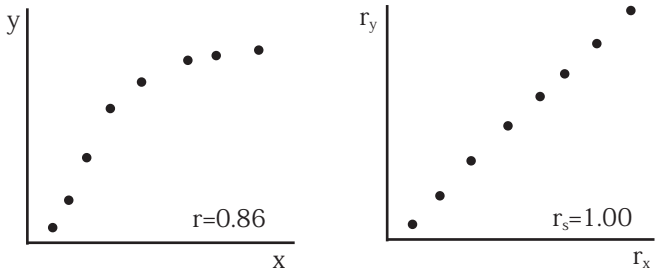
対して、直線に乗らない相関関係を扱うには２つの方法がある。ひとつは直線的関係になるように対数変換などを行う事。もうひとつは分布の形を問わないノンパラメトリックな相関係数を使用する事である

スピアマンの順位相関係数（Spearman's rank Correlation Coefficient）

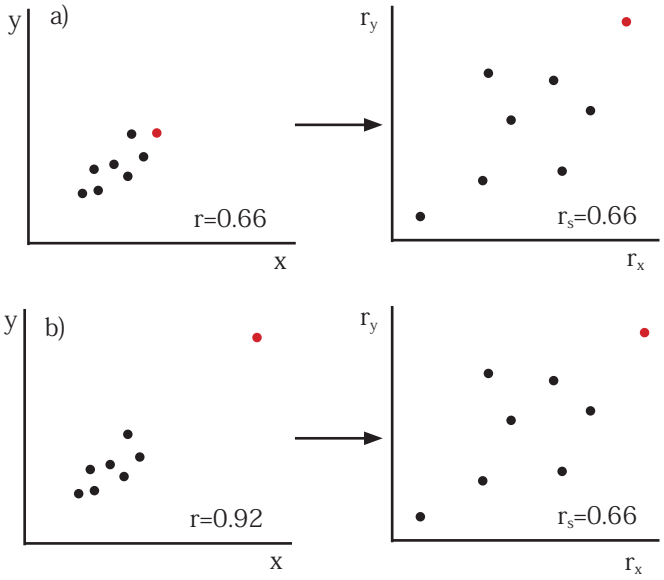
ピアソンの積率相関係数には正規分布に従うことを求めたり、外れ値があると精度が落ちるといった欠点があるが、これは統計学的仮説検定で触れたパラメトリック、ノンパラメトリックのような考え方をすると、それぞれの相関係数が理解しやすい。ピアソンがパラメトリックであればノンパラメトリックに相当するのがスピアマンの順位相関係数（ケンドールも同様）である。これらの順位相関係数は直線関係に乗らない相関関係や分布が歪んでいる、外れ値が存在するといった場合に使用すると有効である。

スピアマンの順位相関係数はその特性として下図のような単調増加 (減少）曲線であれば直線関係でなくても適用可能という点である。しかし、単調な曲線でない時は、あまり精度は良くない。また分布の歪みという点に対しても下図 a）と b) で赤丸はそれぞれ位置が異なるのでピアソンの積率相関係数はそれぞれ異なるが、順位相関としてみると赤丸の位置はどちらも同じ場所に位置するので r_s は両者とも同じである。変数の分布にあまり左右されないという点はノンパラメトリック検定の長所を表していることが分かると思う。

Spearman の順位相関係数の特性
直線関係でなくても、単調増加 (減少）曲線（左図）であれば適用可
順位を取ることで右図のようにすることが可能

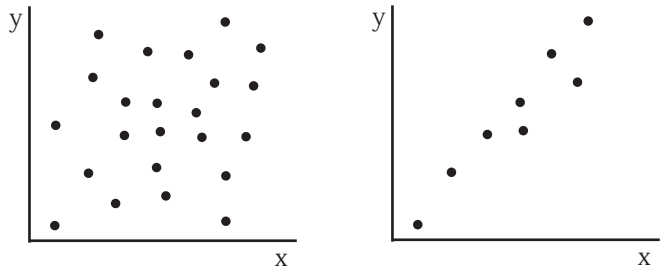


標本分布の偏りに左右されにくい
赤丸の点の位置が a)、b) では大きく異なり相関係数 r も違うが順序相関で見ると赤丸の位置はどちらも同じ所にあるので両者の r_s は変わらない



相関の検定

相関係数の計算そのものは、さして難しい物ではないが求めた相関係数が、標本母集団に適用できるのか否かという問題が出てくる。つまり、標本母集団では相関が無いが抽出した標本ではたまたま相関があるようなデータであった時それは相関していると言えるだろうか？ということである（下図）。



大本の母集団は左図のような一様分布をしているのにサンプリングした結果、たまたま右図のような分布になった。このサンプリング標本から得られた相関係数は母集団を反映するだろうか？（極端例であるが）

この問題には母相関係数の検定（無相関検定と呼ぶこともある）を行うことで知ることができる。統計学的仮説検定において二群間に差はないという帰無仮説を立て、それを棄却できるかどうかを見るが。同じ考えで相関は 0 である（標本は無相関）という帰無仮説を検定する。それが棄却されれば相関は 0 ではないという対立仮説が採用され得られた相関係数は支持される。

無相関検定は母相関係数が 0 という帰無仮説の下、次式の検定統計量が、自由度 n-2 の t 分布に従うことを利用して検定を行なう。

$$t_0=r(n-2)^{1/2}/(1-r^2)^{1/2}$$

R であれば X,Y の二変数に対して相関係数を求める時、cor.test(X,Y,method="pearson") と打ち込む。method= 以下の部分を "pearson" でピアソンの積率相関係数が、"spearman" でスピアマンのスピアマンの順位相関係数、"kendall" でケンドールの順位相関係数を求めることができ、同時に母相関係数が 0 の帰無仮説を両側検定した有意確率と 95% 信頼区間が計算される。** 但し信頼区間が計算されるのはピアソンの積率相関係数のみ

例） 二種類の機械で HbA1c を測定した際の相関計数を求める。同じデータを用いてピアソンとスピアマン両方の相関係数を計算した。なお、R で計算を行なうと無相関検定の t 値、p 値も計算されるので共にみてみたい。

装置 A	5.5	5.3	5.4	4.7	5.3	5.2	5.6	4.8	5.2	4.9
装置 B	5.4	5.3	5.3	4.6	5.2	5.0	5.4	4.7	5.1	4.8

```
Pearson's product-moment correlation
 data:  Dataset$var1 and Dataset$var2
 t = 14.6701, df = 8, p-value = 4.578e-07
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  0.9228058 0.9958611
 sample estimates:
       cor
 0.9819163
```

```
Spearman's rank correlation rho
data:  Dataset$var1 and Dataset$var2
S = 2.5307, p-value = 2.377e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.9846626
```

p-value が有意確率でこの数値からどちらの方法でも帰無仮説は棄却される。つまり、この二標本には何がしかの相関があると考えられるという結果であり。cor、rho の下、最終行の数値がそれぞれの方法から計算された相関係数で、どちらも 0.98 である事がわかる。

ちなみに相関係数 0 は無相関、1 および -1 は正か負の完全相関だがその間の数値に対して、かつては強い相関、弱い相関と表現されていたが最近はこの表現はあまりしない方向に進んでいるようである、相関係数は標本データから計算されただけなので、相関の因果関係や擬似相関といったことについては何も示唆してはくれない単純に相関係数だけを出して相関が有る無いというのはあまり感心できるものではない。だからこそ無相関検定、回帰の項で説明する寄与率などを使い結果の解釈を行うべきである。仮説検定でも触れたが結果をどう解釈するかはその研究者のサイエンスに基づいて行うのは相関でも同じである。

相関を調べるにあたって通常は二変数がどのような分布をするか、まずは散布図を描くと思う。もしくは統計学的仮説検定の流れの中で相関を調べる場面も有ると思うが、どちらにしても散布図からデータの分布がどうなっているかをつかんでから使用する相関係数を選択する事が必要であろう。乱暴な言い方をすれば、二変数がほぼ直線に乗るような分布をしていればピアソンもスピアマンどちらを使っても得られる相関係数に大きな差は出てはこないのが一般的である（前項の例題）、しかし相関（回帰も）は統計学の一分野であるので、得られた標本から母集団を推定するという基本命題から逃れる事はできない。従って二変数がそれぞれどのような分布型をした標本なのか､それによって使用される統計学的手法が決定されるという手順が相関にも常に存在するのである。

補足

カテゴリ変数の関連性の指標（カテゴリ変数版の相関係数）

二つのカテゴリ変数によりクロス集計表を作成する目的には独立性の検定（χ 2 検定等）の他に、二つのカテゴリ変数同士の関連度合いを見たいときがある。つまりカテゴリ変数における相関係数を出したいという目的である。
カテゴリ変数であるので相関という言葉は使えず正しくは連関という言葉になるが、この指標としてはユールの連関係数（Q 係数）、ファイ係数､、クラメールの C 等がある。

Q 係数、ファイ係数は共に 2 x 2 分割表の行・列の連関度を見る指標で求め方は以下のとおり

a	b	R_1
c	d	R_2
C_1	C_2	N

Q 係数 (Yule's Coefficient of Association)
分割表の各セルの成分を次式に代入する
Q=（ad-bc)/(ad+bc)
周辺度数に左右されずに -1 ≦ Q ≦ 1 となる。

ファイ φ 係数　(Phi Coefficient)
2 x 2 分割表の χ 2 値から求めるか、各セルの成分から次式によって計算する。
$\phi = ad-bc/(R_1R_2C_1C_2)1/2$
φ 係数のとりうる値は周辺度数 (R_1,R_2,C_1,C_2) で変化して、$R_1=C_1(R_2=C_2)$ または $R_1=C_2(R_2=C_1)$ のときに最大となって
-1 ≦ φ ≦ 1 の範囲をとる。
この φ 係数はピアソンの積率相関係数（間隔尺度)、スピアマンの順位相関係数（順序尺度）と数学的に同値である。

クラメールの C 係数 (Cramer's Contingency Coefficient)
l x m 分割表において、 χ 2 値がデータ総数によって価が変わるので、異なる分割表同士の比較が困難となる。そのため χ 2 値をデータ数に応じて標準化したのがクラメールの C 係数である。これは次式で求められる。
C= χ 2/N(q-1)　(0 ≦ C ≦ 1)
分母の N(q-1) は N はデータ総数、q は l,m のどちらか小さいほうの値となる。

回帰 (Regression)

回帰という言葉の意味は一巡して元に戻るということだが、統計学では変数間の関数関係一般を指す言葉として用いられていて、線形回帰、ロジスティック回帰、Cox 回帰などと拡張して使われている。ここでは基本的な線形回帰 (simple linear regression) の説明として、検量線をモデルにしてみたい。

臨床検査技師には改めて説明するまでもないが、既知濃度の標準物質を測ったときの吸光度、その濃度によってほぼ完全に（通常は 98% 以上）説明されるときに、その関係を使いサンプルを測定したときの吸光度からサンプルの濃度を逆算するために使う直線を検量線と呼ぶ。

濃度のように量がわかっている（はっきりしている）データを「独立変数」x、吸光度のように誤差を含む可能性がある測定値を「従属変数」 y として y=bx+a という式の係数 a,b を最初二乗法で推定し、サンプルの測定値 y から x=(y-a)/b によって濃度 x を求める。
測定値から濃度を推定するときは、回帰式をそのまま使わず逆算する形になることに注意をしてほしい。

最初に相関で向きは問わないが回帰では向きがあると書いたのに、この検量線の説明では回帰式の逆算を使っているのは矛盾するのではないのかという指摘が有るかもしれないが、ほぼ完全にという条件（98% 以上）がついていて、逆算を行っても実用上問題ないレベルまで回帰式の係数を設定しているから可能なのであり。すべての回帰で逆算が可能であるというわけではない。例えば、背の高い親の子はみな背が高いかという検討では、独立変数である親に対して従属変数としての子の身長はバラツキが大きいため、背が高いという傾向が見られるが決して 98% 以上の確率で回帰式を設定できないことは自明であろう。

では逆算が可能となる回帰直線の適合度の目安である 98% とは何を見ればよいのか？これは相関係数の二乗が 0.98 以上あることがその目安となる。この r^2 は寄与率、決定係数（Coefficient of Determination) と呼ばれるものである。r^2 が 0.98 ということはその平方根をとると基となる相関係数 r は 0.989 以上、実質 0.99 以上必要ということがわかる。ここまで相関係数があれば逆算しても実用に耐えられることが理解できるであろう。一方、逆算を必要としない回帰直線であれば寄与率が 0.98 以上である必要はなく、相関係数が 0.92 や 0.95 位でも十分に説得力を持つ回帰式として扱うことが可能である。
余談であるが相関係数の優劣のようなもの（あちらよりもこちらの方が相関が良い）を問う場面では相関係数そのものの大小よりも、この寄与率を参考にするほうが良い。例えば相関係数 0.92 と 0.95 の寄与率はそれぞれ 0.8464 と 0.9025 となり。後者が確かに相関係数は大きいのだが、寄与率から見ればどちらもまだまだ（どっちもどっち、場合によっては比べる意味が疑わしいという解釈）と判断できるであろう。

回帰においては仮説検定では使わなかった言葉が多く出てくるので整理をしておく。
直線回帰 y=bx+a の式において
y は従属変数、目的変数、応答変数、基準変数、外的基準、等々
x は独立変数、説明変数、回帰変数、等々
b は回帰係数 regression coefficient
a は切片 intercept

この回帰式とは別に
95% 予測区間（期待値の標準誤差が 95% の信頼性をもってこの区間に含まれるであろう範囲）
95% 信頼区間（データの 95% がこの範囲に入るであろうという範囲）
という言葉も使われる。

この y,x の呼び方は統計学の流儀によって前述のように複数の呼び方が混在している、また、回帰式そのものの提示以外で論文中に y、x と描いてしまうと、読んでいる人間は執筆者がどういった考えでこの解析を考えているのかが判りづらいため、通常は目的変数は何、説明変数は何等と記載される。その言葉が何を意味するかを知らなければ解説書、論文等を呼んでいても混乱してしまうので前述の呼び方は最低限覚えておいたほうが良い。

回帰式は予測式であり x を与えて y を予測する形になっていて、予測される y は常にゆらぎ（誤差）を持っている可能性がある。そのため予測される y の妥当性を表すために予測区間、信頼区間が使われる。

具体的な直線回帰を求めるには、前述した最小二乗法（Least Squares Method) を用いる。その原理は n 個の点に対して回帰直線を引いた際に、「各点 (x_i,y_i) から回帰直線までの垂直距離の二乗和 S（回帰からの偏差平方和）が最小となる時」の直線式を求める。

このようにして求めた回帰直線だが、データの配置によっては何通りもの回帰直線の残差平方和が大差ないという状況がありえる。選択した独立変数と従属変数が実は全く無関係であった時、データの重心を通るどのような傾きの直線を引いても残差平方和はほとんど同じになってしまう。また、身長と体重のように、どちらも誤差を含む可能性のある測定値の場合、どちらを独立変数、どちらを従属変数とみなすかが問題となる。一般的にはどちらかによって他方が決まるという方向性（因果の向きという）が仮定できれば、それを独立変数と見なしても良いとされているが、回帰分析を行なうと独立変数に測定誤差が有る可能性が消去（排除）されてしまうことを知っておいた方が良い。よって、測定誤差が大きい可能性のある変数を独立変数とした回帰分析はできれば避けたほうが良い。そして従属変数、独立変数の設定は因果の向きに基づいて正しく設定しなくてはならない。

回帰係数の検定

そして、回帰直線の係数の推定値の安定性を評価することが必要となってくる。そのために使われるのが t 値と呼ばれる統計量である。Ｅｘｃｅｌの分析ツール内の回帰分析でも概要の表中に係数と一緒に「ｔ」という表示がある。切片の段の数値が a、X 値 1 が b の係数値の有意性を検定するための数値になる。

検定にあたっての帰無仮説は以下のようになる。
回帰係数の有意性の検定　test for significance of the regression ; test for regression slope
H_0：回帰係数（ｂ）が０である（H_0：$\beta = 0$ の検定　β は傾き b の母数）

ｙ切片の検定 test for regression intercept
H_0：切片（ａ）が０である（H_0：$\alpha = \alpha_0$ の検定　α はｙ切片 a の母数）
これら二つの統計量ｔは、自由度ｎ－２のｔ分布に従う。

例題　次のデータから回帰直線式を求め、回帰係数の検定を行なってみる。

X	1	3	4	5	6	7	9	10	11	12
Y	4	4	5	6	5	7	6	8	9	7

Ｅｘｃｅｌでは分析ツール内の回帰分析を選び、変数を指定するが最初が従属変数（目的変数）, 次が独立変数（説明変数）の順になっている（ｙ→Ｘの順）ので指定の際は注意が必要である。例題のデータでは、独立変数が X、従属変数が Y として扱われる。オプションについては必要であればチェックをすればよい、変数の指定だけでも回帰分析については必要な物が計算される。

概要

回帰統計	
重相関 R	0.864922
重決定 R2	0.748089
補正 R2	0.716601
標準誤差	0.885478
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	18.62742	18.62742475	23.75729139	0.001233
残差	8	6.272575	0.784071906		
合計	9	24.9			

	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	3.416388	0.617695	5.530868508	0.000553264	1.991981	4.840795	1.991981	4.840795
X 値 1	0.394649	0.080968	4.874145196	0.001233474	0.207937	0.581361	0.207937	0.581361

同じデータを使いRでも計算してみる。Rで回帰分析は線形回帰モデルへのデータの適用となりlm(Y^X)とデータを指定すれば回帰直線の推定値が得られる。ここでもYは従属変数、Xが独立変数である。検定結果も知りたい時はsummary(lm( ))とすることで一度に表示させることができる。

入力例（データはDatasetに格納）

```
RegModel.2 <- lm(Y~X, data=Dataset)
summary(RegModel.2)                    **summary(lm(Y~X, data=Dataset)) と指定しても結果は同じ
```

結果は以下の通り

```
Call:
lm(formula = Y ~ X, data = Dataset)
Residuals:
      Min        1Q    Median        3Q       Max
-1.15217  -0.73829   0.09699   0.63043   1.24247
Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 3.41639   0.61769     5.531     0.000553 ***
X           0.39465   0.08097     4.874     0.001233 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8855 on 8 degrees of freedom
Multiple R-squared: 0.7481,        Adjusted R-squared: 0.7166
F-statistic: 23.76 on 1 and 8 DF,  p-value: 0.001233
```

まず、回帰係数a,bはExcelでは切片がaで3.41、bはX値1で0.39である。RではEstimate(推定）の列で、aがintercept(切片)、bがXで表示されていてExcelと同じであることが分かる。よって回帰式はy＝3.41+0.39 xとなる。次に回帰係数の検定であるが「t」、「t value] が統計量であり、その有意確率をt分布表から知るのだがExcel、R共にp値が計算されているので、この数値を見れば判断ができる。a,bどちらの回帰係数もその確率は＜0.01であるので帰無仮説は棄却されるので、この回帰は有意であると判断できる。

第5回　終了