統計初歩

1. はじめに

統計に関しては 1 年次のドキュメンテーション演習でも取り扱いましたが，この実験でも復習の意味も含めて簡単に取り上げたいと思います．また，プログラミング序論の復習の意味も含めて統計値を求める C 言語プログラムを作成する課題を用意しました．

2. データの種類と統計量

実験で観測されるデータには様々なものがあります．たとえば，情報コミュニケーション工学科 2 年生各学生の身長とか，血液型，この学科が好きか嫌いかの度合いなどのデータです．データの統計量（平均値など）を求める際に重要となるのがデータ水準です．

観測される変数は数値で表現されることになります．このときに観測される変数と数値を対応させる基準がデータ水準であり，数値自体が持つ意味を区別します．データ水準，名義尺度，順序尺度，間隔尺度，比尺度があります．次にこれらについて説明します．

(1) 名義尺度

血液型を表す変数は A,B,O,AB のように単に他のものと区別するためのラベル（名義）を値として取ります．コンピュータで処理を行うために，たとえば A は 1，B は 2 のように数値を割り当てることもありますが，この数値もあくまでも単なるラベルです．このような変数は名義尺度であると言います．

(2) 順序尺度

この学科が好きか嫌いかといった値は，たとえば「この学科が好きですか？」という問いに対して，「1:とても好き，2:好き，3:嫌い，4:とても嫌い」から選ぶという質問から得ることができます．それぞれの回答には便宜上 1～4 という数値が割り当たっていますが，先程のように単なるラベルではないことがわかるでしょうか．数値が小さいほど好きで，大きいほど嫌いということを表しています．このように順序（大小）関係に意味がある変数は順序尺度であると言います．

(3) 間隔尺度，比尺度

これらは数値自体に意味があるものです．数値の差だけに意味あるものが間隔尺度で，数値の比にも意味あるものが比尺度です．温度（℃）は 10℃から 15℃に上がったときに 50%温度が高くなったと言うことはできません．0℃が基準ではないので比を取る意味がないからです．
したがって間隔尺度です．一方，身長は 100cm から 150cm に伸びたとき 50%背が伸びたと言うことができます．したがって比尺度です．なお，間隔尺度は距離尺度とも呼ばれます．

統計とは観測された一つ一つの値を個別に検討するためのものではなく，観測値全体の状況や傾向を知るためのものです．これを知るためのものが様々な統計量です．ここで，先のデータ水準によって，利用すべき統計量が異なる点に注意すべきことが重要です．たとえば，名義尺度の変数の平均値とはなんでしょうか．何の意味もないものであることは，ちょっと考えればわかると思います．表 1 にそれぞれのデータ水準にあった統計量を示します．なお，ある水準の統計値は，それより上位の水準でも利用可能です．たとえば，順序尺度水準で利用できる最大値は比尺度水準でも利用できます．

表 1　各データ水準で利用できる統計量

データ水準	統計量
名義尺度	有効データ数，最頻値，度数分布
順序尺度	最大値，最小値，中央値，第 1 四分位数，第 3 四分位数
間隔尺度	平均値，分散，不偏分散，範囲，四分偏差，歪度，尖度
比尺度	幾何平均，調和平均，変動係数

3. 度数分布

得られた観測値の傾向を掴むためにまず行うことは度数分布表にまとめ，その後，ヒストグラムなどのグラフを描くことです．図 1 に示す値は 2002 年度の NPB パシフィックリーグの選手のうち，規定打席に達した打者の安打数です．これを度数分布表にまとめたものが表 2，ヒストグラムに表したのが図 2 です．

安打数は比尺度の変数です．比尺度の変数（間隔尺度も同様です）の度数分布表を作成するときには，どのくらいの階級数に分けるべきかを考えなければなりません．この目安はスタージェスの公式から得ることができます．

$$C = 1 + \frac{\log n}{\log 2}$$

ここで n は有効な観測値（有効データ数）の数です．この値を元に，階級幅が切りのよい数値になるように階級数を決めてください．ただし，この値を鵜呑みにするのではなく，分布をよく表すように階級数を調整することも重要です．また，階級の最小値が Min，最大値が Max であるとき，その中央値

$$\frac{Min + Max}{2}$$

を，その階級の代表する値として，階級値と呼びます．

度数分布表を元にヒストグラムを書くときの注意すべきことは，棒の間に空間を空けないことと，階級ごとの棒の面積が，階級ごとの度数と比例するように書くことです．

名義尺度の場合は，当然ですが階級に分ける必要はありません．そして，ヒストグラム（正確にはヒストグラムではなく，単なる棒グラフです）に表す際には棒の間を開けます．順序尺度の場合は，とりうる値が少ない場合は名義尺度と同様に，多い場合は間隔尺度・比尺度と同様に扱うとよいでしょう．

165，150，193，171，140，153，135，148，150，
122，148，126，112，130，134，110，145，144，
107，140，137，116，126，111，115，98，117，
86，90

図 1　2002 年度 NPB パリーグ規定打席以上選手の案打数

表 2　度数分布表

階級（本）	度数	相対度数(%)	累積度数	相対累積度数(%)
～100	3	10	3	10
101～120	7	24	10	34
121～140	9	31	19	66
141～160	7	24	26	90
161～180	2	7	28	97
180～	1	3	29	100
計	29			

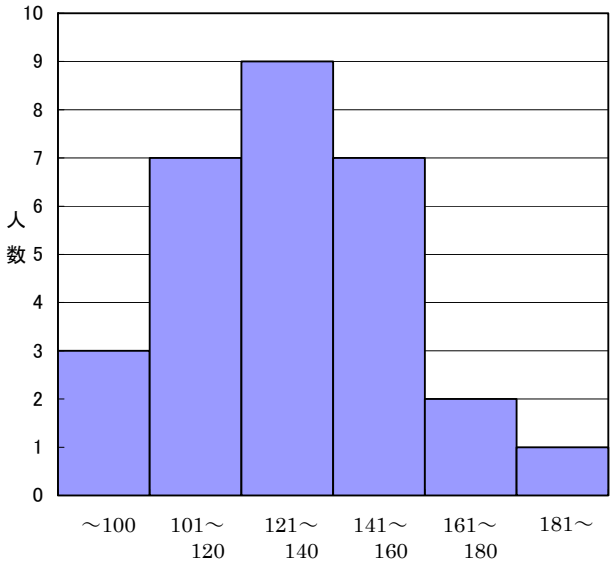

図 2　ヒストグラム

4. 様々な統計量

度数分布表やヒストグラムを描くことで，観測値全体のだいたいの傾向はつかむことができます．さらに，傾向を表す数値がほしい場合には，次に示す様々な統計量を計算します．本節では，一つの変数に対する観測値の傾向を示す統計量について説明します．

4.1 代表値

代表値とは観測値の分布を代表とする値で，最頻値，中央値，平均値などがあります．

(1) 最頻値（モード）

最頻値は度数の最も大きな値です．たとえば，{1,2,2,3,3,3,4,10,12}というデータの再頻値は 3 です．また，{1,2,2,3,3,3,4,10,12}の最頻値は 2 と 3 になります．また，度数分布が与えられた場合の最頻値は，度数の最も多い階級の階級値とします．表 1 の度数分布表では階級[121,140]の度数が最も多いため，その階級値 130.5 が最頻値となります．ただし，分布が左右対称でないことも考えられるため，その階級の下限値を l，その階級の次の階級の度数を f_{+1}，前の階級の度数を f_{-1}，階級の幅を h としたときに

$$Mo = l + \frac{f_{+1}}{f_{-1} + f_{+1}} h$$

で求まる値を最頻値と定義することもあります.

(2) 中央値（メディアン）

中央値はデータを小さい方から大きい方へ整列させたときに，ちょうど真ん中に位置する値です．有効データ数を n，整列後の観測値を $x_i (i = 1,2,\ldots,n)$ とすると，

$$n \text{ が奇数の場合} \quad Me = X_m, m = \frac{n+1}{2}$$

$$n \text{ が偶数の場合} \quad Me = \frac{X_m + X_{m+1}}{2}, m = \frac{n}{2}$$

として求めることができます．たとえば{1,2,2,2,3,4,4,10,12}の中央値は 3 です．度数分布が与えられた場合の中央値は，累積度数を取り総度数（有効データ数）の半分以上で，最も近い累積度数にあたる階級値とします．表 1 の度数分布では総度数が 100 ですので，その半分である 50 以上で最も近い累積度数は 66 となり，その階級の階級値 130.5 が中央値となります．また，最頻値のときと同じ理由で，総度数の半分以上で最も近い累積度数にあたる階級に中央値が存在するとし，その階級の下限値を l，度数を f'，幅を h，l より下の度数を F'，総度数を F としたときに

$$Me = l + \frac{\frac{F}{2} - F'}{f'} h$$

で求まる値を中央値と定義することもあります.

(3) 算術平均値

算術平均値はもっとも一般的な平均値です．有効データ数が n，各観測値を $x_i (i = 1,2,\ldots,n)$ とすると

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

から求まります．たとえば{1,2,2,3,3,3,4,10,12}の算術平均値は 4.44 です．また，度数分布表から算術平均値を求める場合は，各階級の度数を f_i，各階級の階級値を x_i，階級数を m とすると，

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} f_i x_i$$

から求まります.

(4) 調和平均

速度の平均を求める場合，算術平均を取るとおかしなことが起こります．たとえば，行きの速度が $15_{km/h}$，帰りの速度が $20_{km/h}$ であるとき，往復全体の平均速度を算術平均として求めると $17.5_{km/h}$ になります．片道距離が 10_{km} だとすると，本来は行きが 40 分で帰りが 30 分で合計 1 時間 10 分なのですが，算術平均時速からだと 48 分となってしまいます．平均時速は距離を時間で割ったものですから，本来の平均時速は

$$(10_{km} + 10_{km}) \div \left(\frac{10_{km}}{15_{km/h}} + \frac{10_{km}}{20_{km/h}}\right) = 2 \div \left(\frac{1_{km}}{15_{km/h}} + \frac{1_{km}}{20_{km/h}}\right)$$

が正しい式です．つまり逆数の平均の逆数で,

$$Hm = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

から求まります．観測値の逆数が間隔尺度であるときに用いるようにします.

(5) 幾何平均

幾何平均は次の式を満たすGmで定義されます.

$$\log Gm = \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

幾何平均は対数正規分布に従うデータの代表値として適しています.

4.2 散布度

散布度は観測値の分布の広がりを表す値です.

(1) 範囲

範囲とは，有効データ数をn，整列後の観測値を$x_i(i=1,2,\ldots,n)$としたとき，最大値$Max=\max(x_1,x_2,\ldots x_n)$と最小値$Max=\min(x_1,x_2,\ldots x_n)$の差

$$R = Max - Min$$

で定義される値です．たとえば{1,2,2,3,3,3,4,10,12}の範囲は 11 です.

(2) 第 1 四分位数，第 3 四分位数

第 1 四分位数Q_1は小さい順に並べたときに小さいほうから 25%のところに位置する値，第 3 四分位数Q_3は 75%のところに位置する値で，有効データ数をn，整列後の観測値を$x_i(i=1,2,\ldots,n)$としたとき,

$$Q_1 = x_f,\ f = \left[\frac{n}{4}\right]$$

$$Q_3 = x_{n-f+1}$$

で定義されます．$[x]$はxを超えない最大の整数を与えます.

(3) 四分偏差

四分偏差は四分領域とも呼び，第 1 四分位数と第 3 四分位数の差の半分

$$Q = \frac{Q_3 - Q_1}{2}$$

です.

(4) 分散，不偏分散，標準偏差

観測値の散らばりを表すときに単に平均値からのずれ（変移）の総和を取ると正の値と負の値が相殺して 0 になってしまいます．そこで，変移の二乗和をとったもので広がりを表すことが多く，これが分散です．有効データ数がn，各観測値が$x_i(i=1,2,\ldots,n)$であるとき,

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

から求まります．また，不偏分散は観測値が属する母集団の散らばりの推定値で,

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

から求まります.

標準偏差はこれら分散，または，不偏分散の平方根で，平均値と同じ次元を持ちます．そして，データが正規分布に従うとき，平均値±標準偏差の範囲内には全データの 68.27%が，平均値±標準偏差×2 の範囲内には全データの 95.45%が，平均値±標準偏差×3 の範囲内には全データの 99.73%が含まれるという性質があります.

4.3 分布形状

代表値や散布度が同じであっても，その形状が異なることがあります．この形状を現すものとして分布の歪度と尖度があります．

(1) 歪度

歪度は分布の左右対称性の歪，つまり左右にどれくらい歪んだ分布になっているかを表し，有効データ数が n，各観測値が $x_i\,(i=1,2,\ldots,n)$ であるとき，

$$Sk=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^3}{nV^{1.5}}$$

から得られます．Sk が 0 のときは左右対称，正の時は左に偏った分布，負のときは右に偏った分布ということになります．

(2) 尖度

尖度は分布の尖り具合を表し，

$$Kw=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{nV^{2}}-3$$

から得られます．Kw が 0 のとき正規分布と同程度，正のときが正規分布より尖っている（全体的に中心に集中している），負のときが正規分布よりなだらかであるということになります．

5. 相関係数

前節で述べたのは一つ変数に対する観測値の傾向を表す値でした．これに対して，二つの変数に対する観測値の関係を知りたいときに用いるのが相関係数です．相関係数は二つの変数間の関係が深いかどうかを表す値で，その絶対値が 0 に近いほど関係が薄く，1 に近いほど関係が深いということになります．相関係数は比較する変数のデータ尺度によって用いる値が異なります．

(1) 間隔尺度間，比尺度間の相関係数

二変数が間隔尺度，比尺度の場合，ピアソンの積率相関係数を用います．それぞれ有効データ数が n の二変数に対する観測値群 X と Y があるとき，ピアソンの積率相関係数は

$$r=\frac{X\text{とY}\text{の共分散}}{\text{Xの標準偏差}\times\text{Yの標準偏差}}$$

から求まります．共分散とは，

$$\mathrm{C}=\frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(\mathrm{y}_i-\bar{\mathrm{y}}\right)$$

から求まる値です．図 3 は図 1 と同じ人の二塁打数です．安打数と二塁打数の相関係数をとってみると 0.76 となります．これは，安打数と二塁打数の関係がかなりあることを示しおり，安打数が多い人は二塁打も多いと言えるということです．ここで，安打数と盗塁数の相関係数を計算してみると 0.4 となり，これは関係がややある程度になります．本塁打数と盗塁数の相関係数は-0.1 となり，ほとんど関係がないという結果が得られます．どのくらいの値であるとき，どのくらい関係があると言ってよいかを表 3 に示します．

27，23，46，31，25，35，32，40，27，18，25，
33，19，31，26，12，31，31，19，40，33，18，
23，14，28，21，27，12，15

図 3　2002 年 NPB パシフィックリーグ規定打席以上選手の二塁打数

表 3　相関係数の解釈目安

相関係数の絶対値	解釈の目安
0.0～0.2	ほとんど相関関係はない
0.2～0.4	やや相関関係がある
0.4～0.7	かなり相関関係がある
0.7～1.0	強い相関関係がある

二つのデータ群の相関関係を見るためにグラフ化する場合は相関図で表します．これは散布図とも呼びます．安打数と二塁打数の関係を表す相関図を図 4 に示します．この図からも安打数が増えれば二塁打数が増えていることがわかると思います．

一方，相関係数が非常に少なかった本塁打数と盗塁数の相関図を書いてみると図 5 のようになり，関係のないことがよくわかります．

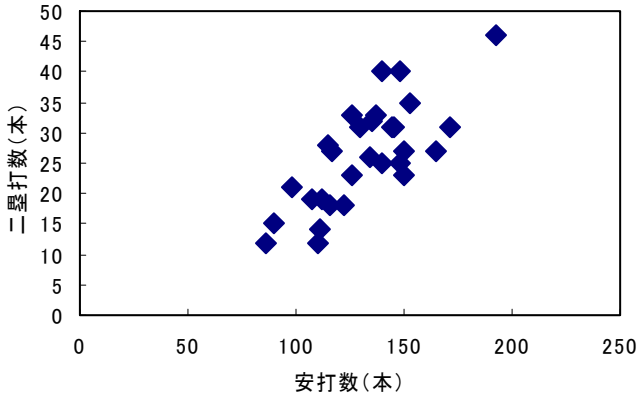

図 4　安打数と二塁打数の相関図

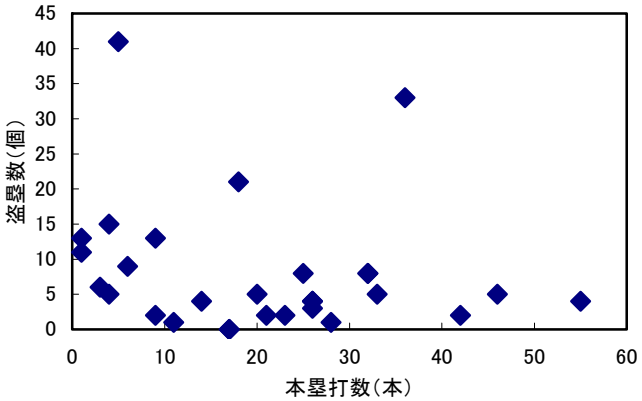

図 5　本塁打数と盗塁数の相関図

(2) 名義尺度間（カテゴリー間）の相関係数

血液型と生まれ月を調査した結果から，血液型と生まれ月に関係があるかを考える場合は，両方とも名義尺度であるため，次のように相関係数を求めます．

変数 X と Y それぞれが k 個と m 個のカテゴリ（名義）を持つときに，その変数同士の関係を見るときにはまず表 4 のようなクロス集計表を作ります．ここで，i 行 j 列の値 o_{ij} は変数 X のカテゴリー i，変数 Y のカテゴリー j に属する度数データです．また，$n_{i\bullet}$ は i 行の合計，$n_{\bullet j}$ は j 列の合計です．このとき，変数 X のカテゴリー i，変数 Y のカテゴリー j の期待値は

$$E_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

で表されます．そして，

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(o_{ij} - E_{ij}\right)^2}{E_{ij}}$$

から求まる値を用いると，次のような属性相関係数が定義されます．

$$\phi = \sqrt{\frac{\chi_0^2}{n}}$$

$$V = \frac{\phi}{\sqrt{t-1}}, t = \min(k, m)$$

ϕ はファイ係数と呼ばれ，2 行 2 列の場合は先のピアソンの積率相関係数に一致します．ファイ係数は t の異なる集計表の間で相関の強さの比較はできません．V はクラメール係数と呼ばれ，k と m の大きさに関係なく 0 から 1 の範囲をとるため，集計表間で相関の強さを比較することができます．

なお，期待値（誤解を恐れずに説明すると，確率的に求められた取られやすい値の平均値）が 1 未満の升目があったり，5 未満の升目が 20%以上あったりする場合は，この相関係数の信頼性は低くなります．

表 4　$k \times m$ カテゴリデータのクロス集計表

	1	2		i		K	
1	o_{11}	o_{21}		o_{i1}		o_{k1}	$n_{\bullet 1}$
2	o_{12}	o_{22}		o_{i2}		o_{k2}	$n_{\bullet 2}$
J	o_{1j}	o_{2j}		o_{ij}		o_{kj}	$n_{\bullet j}$
M	o_{1m}	o_{2m}		o_{im}		o_{km}	$n_{\bullet m}$
	$n_{1\bullet}$	$n_{2\bullet}$		$n_{i\bullet}$		$n_{k\bullet}$	n

(3) 順序尺度の場合（順序の相関係数）

順位に相関があるかを示す値にはスピアマンの順位相関係数があります．それぞれ有効データ数が n の順序尺度の二変数 X，Y が得られたとき X，Y それぞれに小さい方から順位をつけます．この順位の差を d_i とすると $\sum_{i=1}^{n} d_i^2$ は二変数の順序の一致性を表す指標になります．つま

り順位が同じであれば$\sum_{i=1}^{n} d_i^2 = 0$になります．また，完全に逆順の場合は$\sum_{i=1}^{n} d_i^2 = \frac{(n^3 - n)}{3}$となります．この事実を利用して，取りうる値が-1 から 1 になるように変換した式

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n^3 - n)}$$

から求まる値がスピアマンの順位相関係数です．この値が 0 に近いとき二変数は無相関であり，1 または-1 に近いとき二変数には相関が強いということになります．

たとえば，「大学が好きですか」，「野球が好きですか」という二つの質問を 5 人に聞いた結果が表 5 のような結果が得られたとします．ここで回答は「1-とても好き，2-好き，3-嫌い，4-とても嫌い」という選択肢から選ぶようにしました．つまり，変数は共に 4 つのカテゴリを持つ順序水準ということになります．このデータのスピアマンの順位相関係数を求めると 1-(6*22.25)/(125-5)=-0.1125 となり，大学が好きかを指標とした 5 人の順番と，野球が好きかどうかを指標にした 5 人の順番には，あまり相関がないことになります．見方を変えて，この 5 人に関しては，大学が好きであることと，野球が好きであることの間に相関はあまりないと言ってよいでしょう．

表 5　スピアマンの順位相関係数を求めるための計算表

被験者番号	1	2	3	4	5
X（回答番号）	1	2	2	4	3
Y（回答番号）	1	4	3	1	3
Xi の順位	1	2.5	2.5	5	4
Yi の順位	1.5	5	3.5	1.5	3.5
d_i	-0.5	-2.5	1	3.5	0.5
d_i^2	2.5	6.25	1	12.25	0.25

6. 検定

前節の相関係数はどのくらいの関係があるかを示す統計量でしたが，研究成果を論文に記す場合，検定という手法をよく用います．これは，たとえば先の二変数が独立であることを示したり，二群のデータの平均値に差があることを示したりするときに用います．しかし，検定の説明にはかなりの時間が必要なので，ここでは取り上げません．興味のある人は統計学の本などを読んでみるとよいでしょう．

7. おわりに

本資料では，ドキュメンテーション演習で学んだことの復習も含めて，統計について簡単に説明しました．統計はこれから皆さんが研究を行っていく上で幾度と無く必要となるものですので，一度しっかりと勉強しておくことをお勧めします．

課題

次に示す課題 1 のグラフと考察を記したレポート，および，課題 2 のプログラムソースを提出してください．

今年度から学科方針としてペーパレス化を目指しているため，レポートも電子提出とします．レポートは PDF（推奨），または Microsoft Word 形式のファイルとしてください．1 ページ目は学科所定の表紙としてください．レポートのファイルサイズが 800KB を越えないように気をつけてください．このサイズを越えるとレポートが届かない可能性があります．プログラムソースは拡張子が c または cpp のプレーンテキストファイルとしてください．ファイルは分割せず 1 つのファイルにまとめてください．この 2 つのファイルを添付し，サブジェクトを半角英数字の学籍番号とした電子メールを cs101@hands.ei.tuat.ac.jp 宛に送付してください．

締め切りは 4 月 18 日午後 5 時です（年度の最初であるため特別に提出期限をちょっとだけ延長してあります）．

課題 1

図 3 に示すデータについて，階級幅を変えた複数のヒストグラムを描いてください．そして，階級幅の変えることでグラフから読み取れることにどのような影響が出てくるのかを考察してください．

課題 2

間隔尺度の二変数の相関係数を求めるプログラムを作成してください．プログラムは標準の C 言語，または C++言語を用いて，コンソール上で動作する（Microsoft Windows であれば DOS プロンプト上で動作する）プログラムとしてください．