2016 年度 森林統計学 第 14 回資料 1　教科書の問題の推奨問題の一覧と補足

第 2 章 標本データの記述（問題 p.31〜p.34）

・ヒストグラムについて：8.，9.，10.

・平均と標準偏差の計算：11.，12.，13.，14.

・平均と標準偏差の理論的背景：17.，20.，21.

・中央値・範囲・四分位数：25.，26.，27.

・総合問題：28.，29.（教科書 p.29，例 2 も参照）

注）分散や標準偏差を計算するとき、各データと平均値の差を求める必要があるが、より計算が簡便になる方法として教科書 p.31 の欄外に「分類されていないデータから s を計算するとき」に便利な式が紹介されている。この方法は分散分析の計算などでもよく用いられる。資料末尾に「平方和の計算の簡便式」として紹介する。

第 4 章 確率分布（問題 p.91〜p.92）

・確率分布と平均・標準偏差：5.，9.，10.（5.と 10.は組になっている）

・期待値：13.，18.

第 5 章 主要な確率分布（問題 p.117〜p.120）

・2 項分布について：6.，9.

・独立試行について：10.，11.

・2 項分布の平均値と標準偏差：14.，15.

・正規分布表の利用：17.，19.，21.，22.

・一般問題：36.（注意〜答に誤植あり；Web 資料の誤植リストを参照）

第 6 章 標本抽出（問題 p.133〜p.135）

・標本抽出（ランダムサンプリング）の方法：1.，2.，3.，5.，8.

・正規母集団からの標本平均の分布：11.，12.

・非正規母集団からの標本平均の分布：14.，15.，17.，18.

・一般問題：（これに相当するものを課題 4 として行なっている。あえて推奨するとすれば問題 20.）

第 7 章 推定（問題 p.153〜p.157）

・2 節（母平均 μ の推定）：1.，3.，4.

・3 節（大標本法〜σ の不偏推定値 s の利用）：9.，12.，13.

・5 節（小標本法〜t 分布）：28.，29.

・一般問題：30.

第 8 章 仮説の検定（問題 p.185〜p.190）

・1 節（2 種類の過誤）: 1.，3.

・2 節（平均値の検定）: 6.，7.，9.[，11.]，13. （注: 11. は管理図に関するもの）

・[3 節（割合の検定）: 14.，19.]

・4 節（正規分布による平均値の差の検定）: 22.，24.

・6 節（小標本法〜t 検定）: 33.，34.，35.

・一般問題：37.

※ [ ]は授業では詳細を説明しなかった項目に関するもの。それぞれ、11.は管理図, 14.と19.は割合 p の検定に関する問題。

第 12 章 母数によらない検定（問題 p.255〜p.257）

・3 節（2 つの中央値の差の検定）: 6.，9.

<u>平方和の計算の簡便式</u>

教科書 p.31 の欄外には、分類されていないデータから標準偏差 s を計算する式（教科書 p.21 の式（9））

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \quad \cdots\cdots\cdots\cdots \text{（9）式}$$

は、以下のようにも計算できるとしている。

$$s=\sqrt{\frac{\sum X_i{}^2-\frac{1}{n}\left(\sum X_i\right)^2}{n-1}} \quad \cdots\cdots\cdots\cdots \text{p.31 欄外の式}$$

これは要するに各データと平均値の差の二乗和（$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$；これを「偏差平方和」あるいは「平方和」という）が $\sum_{i=1}^{n}X_i{}^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2$ でも計算できる、ということで、言い換えると、平均値 $\overline{X}$ の定義（教科書 p.18,（3）式参照）から $\sum_{i=1}^{n}X_i=n\cdot\overline{X}$ とも書けるので

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}X_i{}^2-n\cdot\overline{X}^2$$

ということである。つまり、平均値がわかっていれば、個々のデータの二乗和（$\sum_{i=1}^{n}X_i{}^2$）さえ求めておけば平方和を求めることができる、ということである。

以下に証明を示す。

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}\left(X_i^{\,2}-2X_i\overline{X}+\overline{X}^2\right)$$

$$=\sum_{i=1}^{n}X_i^{\,2}-2\overline{X}\sum_{i=1}^{n}X_i+\overline{X}^2\sum_{i=1}^{n}1$$ （注：$\overline{X}$ は定数なので Σ の外に出せる）

$$=\sum_{i=1}^{n}X_i^{\,2}-2\overline{X}\left(n\cdot\overline{X}\right)+n\cdot\overline{X}^2$$ （注：$\sum_{i=1}^{n}X_i=n\cdot\overline{X}$ ）

$$=\sum_{i=1}^{n}X_i^{\,2}-n\cdot\overline{X}^2$$

具体例：　3 つのデータ 3，4，5 から s を求める（$n=3$，$X_1=3$，$X_2=4$，$X_3=5$）。

① 平均値 $\overline{X}=\dfrac{\sum_{i=1}^{n}X_i}{n}=\dfrac{X_1+X_2+X_3}{3}=\dfrac{3+4+5}{3}=4$

②-1. $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ を使う方法(教科書(9)式	②-2. $\sum_{i=1}^{n}X_i^{\,2}-n\cdot\overline{X}^2$ を使う方法

②-1.

i	X_i	$X_i-\overline{X}$	$\left(X_i-\overline{X}\right)^2$
1	3	3−4 = −1	$(-1)^2=1$
2	4	4−4 = 0	$(0)^2=0$
3	5	5−4 = 1	$(1)^2=1$

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=2$$

②-2.

i	X_i	$X_i^{\,2}$
1	3	9
2	4	16
3	5	25

$$\sum_{i=1}^{n}X_i^{\,2}=50$$

$$\sum_{i=1}^{n}X_i^{\,2}-n\cdot\overline{X}^2=50-3\cdot4^2=2$$

偏差平方和はいずれも 2 となる。

注）②-2 の方が計算回数が 1 列分少ない。N が大きくなるほど、②-1 の方法よりも②-2 の方法の方が計算の手間が少なくなる（計算間違いが起こる可能性も小さくなる）。また、平均値 $\overline{X}$ の計算が間違っていたとき、②-1 では最初から計算をやり直さないといけないが、②-2 だと最後の式の再計算だけで済む。

③得られた偏差平方和から s を求める

$$s=\sqrt{\frac{\text{偏差平方和}}{n-1}}=\sqrt{\frac{2}{3-1}}=1$$

注）教科書 p.21 の式（9）と教科書 p.31 欄外の式は一見見た目が異なるが、それぞれ

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}},\quad s=\sqrt{\frac{\sum_{i=1}^{n}X_i^{\,2}-n\cdot\overline{X}^2}{n-1}}$$

と書くことができる。すなわち、偏差平方和を自由度(n-1)で割ったものが分散であり、そのルートが標準偏差である。

「2 種類の過誤」の補足

教科書 p.159，表-1（α と β を追記）

	H_0 を採択	H_1 を採択
H_0 が真	正しい （確率：1－ α）	第 1 種の過誤 （確率：α = 有意水準）
H_1 が真	第 2 種の過誤 （確率：β）	正しい （確率：1－ β = 検出力）

①α

「有意水準」（「危険率」という場合もある）。α と β を同時に最小とする α は 0.05 程度であることがわかっている。検定者は有意水準を適宜設定することで、第 1 種の過誤を起こす確率（すなわち α）をコントロールできる。

②β

β は第 2 種の過誤を起こす確率だが、特に 1－β を「検出力」という。「検出力」は、母集団に有意な差がある場合に正しく有意であるという判定をすることができる可能性を意味している。正しい検定方法を用いることが検出力を高めることにつながる。

検定の 3 つの立場

①信頼区間によるもの

「推定」の考え方から信頼区間を求め、信頼区間の中に確率変数（「平均値の差の検定」資料(2)式の x^*すなわち $\bar{x}_1 - \bar{x}_2$）が入るか否かで判定する（入らなければ「有意」）。

②検定値（「平均値の差の検定」資料(1)式の z^*あるいは t^*）によるもの

標準の方法。確率変数（$x^* = \bar{x}_1 - \bar{x}_2$）を標準化の公式で検定値（$z^*$あるいは t^*）に変換し、該当の α（t 検定の場合は α と自由度 ν）に対応する閾値と比較して判定する（閾値より検定値 z^*あるいは t^*が大きければ「有意」）。

注）閾値は教科書では数表から読んでいるが、エクセルの関数を使うと、正規分布の場合 NORMINV 関数、t 分布の場合は TINV 関数で求めることができる。

③P 値（「p 値」と書く場合もある）によるもの

確率変数が検定値（z^*［正規分布による検定の場合］あるいは t^*［t 分布による検定の場合］）より大きな値をとる確率を求めることができ、これを P 値という（2 項分布の独立事象の確率(P あるいは p)とは異なるので注意）。P 値が α より小さければ「有意」となる。

α を事前に定めなくても有意性を判断できるので、研究論文では P 値を示すのが標準的な検定のやり方になっている。

注）エクセルの関数では、場合正規分布の P 値は NORMDIST 関数で、t 分布の P 値は TDIST 関数で求めることができる。