2009 年度 森林統計学 第 13 回資料 2　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）

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

・[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))

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

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、ということだが、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、つまりは

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

ということである。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

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. $\sum_{i=1}^{n}X_i^{\,2}-n\cdot\overline{X}^2$ を使う方法

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)。

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 つの平均値の差の検定」で用いる検定値の計算式

「2 つの平均値の差の検定」では、[1]正規分布による検定、[2]t 検定、[3]Welch の t 検定 の 3 種類の方法があり、条件によりいずれか最も適切なものを選択して検定を行なう。検定の手順はいずれも同じである。

①基礎統計値を算出する。

データ数(n_1, n_2)、標本平均($\bar{x}_1$, $\bar{x}_2$)、標本標準偏差(s_1, s_2)を算出する。

あわせて標本平均の標準偏差($s_{\bar{x}_1}$, $s_{\bar{x}_2}$)も算出する。

注）教科書 6 章の定理 1, 定理 2 により $s_{\bar{x}_1} = \dfrac{s_1}{\sqrt{n_1}}$, $s_{\bar{x}_2} = \dfrac{s_2}{\sqrt{n_2}}$ である。

②検定値を求める。

検定値の計算式の基本式は[1], [2], [3]どれも同じで、異なるのは平均値の差の分布の標準偏差の算出方法だけである。

$$\text{検定値}(z,t) = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$ ([1]の検定値は z; [2], [3]は t)

注）ただし $\mu_1 = \mu_2$ という帰無仮説の場合 $\mu_1 - \mu_2 = 0$ なので、上式の分子は $\bar{x}_1 - \bar{x}_2$ になる。

[1], [3]: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{{s_{\bar{x}_1}}^2 + {s_{\bar{x}_2}}^2}$

[2]: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{{s_P}^2}{n_1} + \dfrac{{s_P}^2}{n_2}}$, ただし ${s_p}^2 = \dfrac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$

注）$s_{\bar{x}_1}$, $s_{\bar{x}_2}$ は 2 乗されて分散になっていることに注意。${s_p}^2$ は合成された分散。

③棄却限界値と検定値を比較して判定を行なう。

対応する分布表([1]は正規分布表−教科書 p.295, 表 IV; [2], [3]は t 分布表−教科書 p.296, 表 V)を用いて設定した α(と t 分布表の場合は自由度 ν)に相当する棄却限界値を読み取り、検定値と比較するというやり方はどれも共通。検定値の絶対値が棄却限界値より大きければ、有意差があると判定する(帰無仮説を棄却する)。

自由度は[2]では $\nu = n_1 + n_2 - 2$、[3]では次式で求める(小数点以下は切捨てでよい)。

$$\nu = \frac{\left(\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}}$$

(注: $s_{\bar{x}_1}, s_{\bar{x}_2}$ を用いて $\nu = \dfrac{\left({s_{\bar{x}_1}}^2 + {s_{\bar{x}_2}}^2\right)^2}{\dfrac{\left({s_{\bar{x}_1}}^2\right)^2}{n_1 - 1} + \dfrac{\left({s_{\bar{x}_2}}^2\right)^2}{n_2 - 1}}$ とも表せる)