10　標本分散

10.1　標本分散と不偏分散

定義 10.1.1 母集団 X からとった大きさ n の標本 $X_1, \ldots, X_n$ と標本平均 $\bar{X}$ に対して次で定まる確率変数 $\bar{V}$ を、大きさ n の標本分散と言います：

$$\bar{V} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2 = \frac{1}{n}\sum_{j=1}^{n}X_j^2 - \bar{X}^2.$$

また、

$$\tilde{V} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

は不偏標本分散（または単に不偏分散）、その実現値は不偏分散と呼ばれます。母集団の分散を v とすると

$$E[\bar{V}] = \frac{n-1}{n}v, \quad E[\tilde{V}] = v.$$

10.2　母分散が不明な場合の母平均の区間推定

問題 10.2.1 母分散は不明であるとします。この中から大きさ 50 のサンプルを取って調べたところ平均値が 169.6、分散が 4.07^2 でした。このときに母平均 m の信頼度 95 パーセントの信頼区間を求めて下さい。

母分布：	unknown
母平均：	m : unknown
母分散：	unknown
サンプルサイズ：	50（large）
サンプル平均：	169.6
サンプル分散：	4.07^2

大きさ 50 の大きなサンプルを取っているので母分散はサンプルの分散で代用出来ます。更に中心極限定理によれば標本平均 $\bar{X}$ は正規分布 $N\left(m, \frac{4.07^2}{50}\right)$ で近似されます。そこで

$$P[|\bar{X} - m| \le d] = 0.95$$

となる様な $d > 0$ を求めると（計算略）$d \approx 1.13$ が分かります。

従って今回のサンプル平均 169.6 に関して信頼度 95 パーセントで $|169.6 - m| \le 1.13$ が成り立ちますので、これを m に関する条件に読み替えれば

$$169.6 - 1.13 \le m \le 169.6 + 1.13$$

であり、求める信頼区間は $[168.47, 170.73]$ になります。 □

10.3　なぜ $n-1$ で割ると自然な結果となるのか

標本分散：　$\bar{V} = \dfrac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2}{3}$

不偏標本分散：　$\tilde{V} = \dfrac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2}{2}$

$$\begin{aligned}
(X_1 - \bar{X})^2 &= X_1 - \frac{X_1 + X_2 + X_3}{3} = \frac{2}{3}X_1 - \frac{1}{3}X_2 - \frac{1}{3}X_3 = Y_1 \\
(X_2 - \bar{X})^2 &= X_2 - \frac{X_1 + X_2 + X_3}{3} = -\frac{1}{3}X_1 + \frac{2}{3}X_2 - \frac{1}{3}X_3 = Y_2 \\
(X_1 - \bar{X})^2 &= X_1 - \frac{X_1 + X_2 + X_3}{3} = -\frac{1}{3}X_1 - \frac{1}{3}X_2 + \frac{2}{3}X_3 = Y_3
\end{aligned}$$

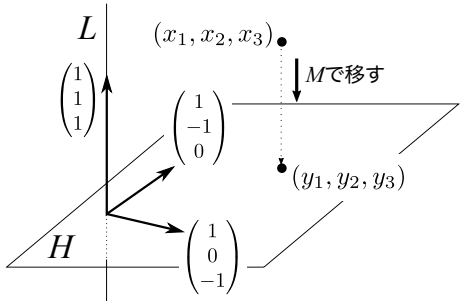
$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = M \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

固有値	固有ヴェクター
0	$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$
1	$\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$

$$M \underbrace{\left\{ p\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + q\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + r\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right\}}_{\text{3 次元空間全体}} = pM\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + qM\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + rM\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

$$= p \cdot 0 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + q \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + r \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

$$= \underbrace{q \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + r \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}}_{\text{平面 } H \text{ 全体}}$$



実はこの一次変換 M は平面 H への正射影変換なのです。

つまり、入力は 3 次元でしたが、出力は（1 次元分潰れて）2 次元になってしまっているのです。たとえ X_1, X_2, X_3 がそれぞれ自由な値を取ったとしても、Y_1, Y_2, Y_3 の 3 つがそれぞれ自由な値を取ることはできないのです。気づけば簡単なことですが、$Y_1 + Y_2 + Y_3$ は必ず 0 になってしまいますよね？　ということは、Y_1, Y_2 が決まれば自動的に Y_3 も決まってしまうのです。

だから一見 3 つあるように見える Y_1, Y_2, Y_3 も、自由度の観点から見れば 2 つであると考えられるわけです。『だから』3 で割るのではなく、2 で割った方が『自然』なのです。

正確な確率論的観点から言うと、X_1, X_2, X_3 は独立ですが、Y_1, Y_2, Y_3 は独立ではありません。『ランダム』と云う言葉を聞くと、つい『独立』であるような印象をもってしまいますが、別の概念ですから注意する必要があります。

10.4　問題演習

基本演習 10.1 [教科書　問題 16.8　改題] 過去の記録から、ある高専 3 年男子学生の 100 メートル走の記録の度数分布は正規分布に従うことがわかっています。

学生の中から 40 人を無作為に選び、100 メートル走の記録をとったところ平均値が 15.3 秒、分散が $(1.88 \text{ 秒})^2$ でした。全学生の平均値を信頼度 95 %で推定して下さい。ただし母分散はサンプルの分散で代用して下さい。また、サンプルの不偏分散で代用した場合にどうなるかも計算してみて下さい。

基本演習 10.2 [問題集　5.16] ある動物用の新しい飼料を試作し、任意抽出された 100 匹にこの飼料を毎日与えて 1 週間後に体重の変化を調べました。増加量の平均は 2.57kg、標準偏差は 0.35kg でした。この増加量について以下の問いに答えて下さい。

(1) 母平均を信頼度 95 %で区間推定して下さい。

(2) 標本平均と母平均の差を 95 %の確率で 0.05kg 以下にするには標本数をいくらにすれば良いでしょうか。

基本演習 10.3 全国一斉にある教科のテストが行われました。受験生から 100 名を抽出し、その得点の平均と標準偏差を求めたところそれぞれ 58.3、12.4 でした。全受験生の得点の平均の 95 %信頼区間を求めて下さい。

基本演習 10.4 ある学校の生徒 50 人を無作為に選び、1 週間に数学を何時間勉強するか聞いたところ、50 人の平均は 18.2 時間、不偏分散は 30.25 時間でした。この学校の生徒の１週間あたりの平均数学学習時間の 95 %信頼区間を求めて下さい。

基本演習 10.5 ある年度の学校保健統計調査によると、全国 2 万人の 17 歳女子の身長の平均は 157.9cm、不偏分散は 5.35^2 でした。17 歳女子の身長の平均の 95 %信頼区間を求めて下さい。

基本演習 10.6 自動車衝突事故の対物保険についてのある研究によると、ある特殊の破損を受けた 120 台の車体を無作為に選んだところ、それらの修理費の平均は 14.4 万円で、標準偏差は 1.7 万円でした。この種の修理の平均費用を信頼度 99 パーセントで区間推定してください。ただし $\sqrt{120} \approx 10.95$ としてください。

基本演習 10.7 ある母集団から大きさ 50 の無作為サンプルを抽出して、そのサンプルの平均値が 54.3、分散が 24.5 でした。母平均の信頼度 99 %の信頼区間を求めて下さい。