2006．8．8

2006 年度　統計工学　期末試験問題

[1]　n 個の確率変数 $X_{11}, X_{12}, \cdots, X_{1n}$ が独立に同一の正規分布 $N(\mu_1, \sigma^2)$ にしたがい，m 個の確率変数 $X_{21}, X_{22}, \cdots, X_{2m}$ が独立に同一の正規分布 $N(\mu_2, \sigma^2)$ にしたがうとする．また，$X_{11}, X_{12}, \cdots, X_{1n}$ と $X_{21}, X_{22}, \cdots, X_{2m}$ は独立とする．

このとき，帰無仮説 $H_0 : \mu_1 = \mu_2$，対立仮説 $H_1 : \mu_1 \neq \mu_2$ に対する検定統計量を示し，それが帰無仮説のもとで自由度 $n + m - 2$ の t 分布にしたがうことを，t 分布の定義（テキスト p.31，2.2.2）に照らして示せ．

[2]　東急田園都市線沿線の賃貸マンション 100 件のデータを用いて，賃貸料を目的変数 y とし，説明変数として専有面積(㎡)を x_1，徒歩(分)を x_2 にとった重回帰分析を行った．その結果，x_1 の偏回帰係数は正値で，x_2 のそれは負値であった．また，x_1 と x_2 の相関係数は正値であった．

ここで，説明変数として専有面積 x_1 のみを用いた単回帰分析を行ったとすると，そのときの x_1 の単回帰係数と，上の重回帰分析での x_1 の偏回帰係数との大小関係を論ぜよ．結論だけでなく理由を示すこと．

[3]　厚生労働省が毎年発表しているその年の平均寿命の算出法を述べよ．

[4]　次の 3 元分割表は，ある年度に高速道路で発生した前面衝突事故について，運転者がシートベルトを着用していた 400 例と，着用していなかった 400 例のそれぞれについて，運転者の性別，事故での生死を分類したものである．

(1)　シートベルト着用割合が男女で異なるかどうかを検定せよ．

(2)　(1)の結果を踏まえて，シートベルト着用の生存への効果を検定せよ．

	男性		女性		
	生存	死亡	生存	死亡	合計
シートベルト着用	90	62	115	133	400
非着用	156	172	27	45	400

【解答例】

[1]　$X_{11}, X_{12}, \cdots, X_{1n}$ の算術平均を $\overline{X}_1$，平方和を $S_1 = \sum_{j=1}^{n}(X_{1j} - \overline{X}_1)^2$ とする．

同様に，$X_{21}, X_{22}, \cdots, X_{2m}$ の算術平均を $\overline{X}_2$，平方和を $S_2 = \sum_{j=1}^{m}(X_{2j} - \overline{X}_2)^2$ とする．

S_1 と S_2 から併合分散 $V = (S_1 + S_2) / (n + m - 2)$ を求めたとき，平均値の差に関する検定統計量は

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)V}} \tag{1}$$

である．この統計量が帰無仮説 $H_0 : \mu_1 = \mu_2$ のもとで，自由度 $n + m - 2$ の t 分布にしたがう理由は次の通りである．

テキスト p.29，定理 2.2 より，S_1 / σ^2 は自由度 $n - 1$ のカイ 2 乗分布にしたがい，同様に S_2 / σ^2 は自由度 $m - 1$ のカイ 2 乗分布にしたがう．仮定より S_1 と S_2 は独立であるから，カイ 2 乗分布の再生性（テキスト p.28）より，$(S_1 + S_2) / \sigma^2$ は自由度 $n + m - 2$ のカイ 2 乗分布にしたがう．

一方，仮定より $\overline{X}_1$ と $\overline{X}_2$ は独立で，帰無仮説 $H_0 : \mu_1 = \mu_2$ のもとでは $\overline{X}_1 - \overline{X}_2$ は正規分布 $N(0, \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2)$ にしたがう．よって

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{2}$$

は標準正規分布にしたがう．

また，テキスト p.30，定理 2.3 より，$\overline{X}_1$ と S_1 は独立で，$\overline{X}_2$ と S_2 は独立である．よって $\overline{X}_1 - \overline{X}_2$ と $S_1 + S_2$ は独立である．

いま，確率変数 X が標準正規分布にしたがい，確率変数 Y が自由度 n のカイ 2 乗分布にしたがい，かつ，X と Y が独立のとき，$T = X \Big/ \sqrt{Y/n}$ の分布を自由度 n の t 分布と定義している．

(1) 式の統計量 T は，(2)式の量 U と $Y = (S_1 + S_2) / \sigma^2$ を用いて

$$T = \frac{U}{\sqrt{Y/(n + m - 2)}}$$

と書ける．これより題意を得る．

[2]　専有面積(㎡) x_1 の平方和を S_{11}，徒歩(分) x_2 の平方和を S_{22}，さらに両者の偏差積和を S_{12} とする．また，賃貸料 y と専有面積(㎡) x_1 との偏差積和を S_{1y}，賃貸料 y と徒歩(分) x_2 との偏差積和を S_{2y} とする．

以上の表記のもと，賃貸料を目的変数 y とし，専有面積(㎡) x_1 と徒歩(分) x_2 を説明変数にした重回帰分析で，得られる x_1 の偏回帰係数 β_1 の最小２乗推定値は

$$\hat{\beta}_1 = \frac{S_{1y} - S_{12} S_{2y} / S_{22}}{S_{11} - S_{12}^2 / S_{22}}$$

である．一方，説明変数として専有面積 x_1 のみを用いた単回帰分析を行ったとすると，そのときの x_1 の単回帰係数 β の最小２乗推定値は

$$\hat{\beta} = \frac{S_{1y}}{S_{11}}$$

である．簡単な計算より，両者の関係は

$$\hat{\beta} = \hat{\beta}_1 + \frac{S_{12}}{S_{11}} \hat{\beta}_2 \qquad (3)$$

と表現できることがわかる．ここに $\hat{\beta}_2$ は，専有面積(㎡) x_1 と徒歩(分) x_2 を説明変数にした重回帰分析での x_2 の偏回帰係数の推定値である．

いま，問題文条件より $\hat{\beta}_1$ は正値，$\hat{\beta}_2$ は負値で，さらに S_{12} は正値であるから，(3)式右辺第２項は負値である．よって $\hat{\beta}_1 > \hat{\beta}$ が成り立つ．この右辺第２項の絶対値が大きいと，左辺の $\hat{\beta}$ は負値になることもある．

この不等式の技術的解釈は次の通りである．いま，専有面積(㎡) x_1 と徒歩(分) x_2 の間には正の相関があるから，専有面積の大きい物件は徒歩も大きい傾向にある．問題文にあるように $\hat{\beta}_2$ は負値，すなわち，面積を固定したときの徒歩の偏回帰係数は負値である．そのため，徒歩を説明変数にしない単回帰分析では，専有面積が大きくなることで，徒歩も大きくなり，その徒歩の効果が専有面積の効果に含まれてしまう．その結果として，徒歩を固定したときの面積の効果よりも割り引かれるのである．

[3]　2006 年に発表される 2005 年の平均寿命については次のようである．

厚生労働省は国民統計として，2005 年 1 月 1 日時点で生存している i歳の人数 n_i（i ＝0,1,2,…）と，その n_i 人のうちで 2005 年 12 月 31 日までに死亡した人数 d_i のデータを有している．これより，2005 年における i歳での 1 年あたりの死亡率は $h_i = d_i / n_i$ と計算される．これをもとに i歳での累積ハザード関数値は

$$H_i = \sum_{j=0}^{i} h_i$$

と算出される．累積ハザード関数値と分布関数値（その年齢までに死亡している割合を表す）の関係は $F_i = 1 - \exp(-H_i)$ である．分布関数が推定されれば，寿命は非負の値をとるので，テキスト p.9 にある平均 μ と分布関数との関係式

$$\mu = \int_0^{\infty} (1 - F(t))\, dt$$

をもとに平均寿命が算出される．

つまり，厚生労働省の言うところの「簡易生命表は，2005 年当時の死亡状況が今後も変わらないと仮定し，各年齢層についてあと何年生きられるかを計算したもので，零歳の子供が平均どれだけ生きられるかを示した数字が『平均寿命』となる」は上記の算出方法を意味しているのである．

[4] (1) シートベルト着用と男女の二元表を作成すると

	男性	女性	行和
シートベルト着用	152	248	400
非着用	328	72	400
列和	480	320	800

となる．2×2 分割表でのカイ２乗適合度検定統計量は，テキスト p.157，練習問題 6.3 の公式が使えるので

$$\chi^2 = \frac{(152 \times 72 - 248 \times 328)^2 \times 800}{400 \times 400 \times 480 \times 320} = 161.3$$

自由度 1 のカイ２乗分布上側 5%点 3.84 よりもはるかに大きく有意である．

シートベルト着用割合が男女で有意に異なるといえる．

(2) (1)の結果より，男女それぞれで，シートベルト着用の生存への効果がない(すなわち，着用の有無にかかわらず生存割合が等しい)としたときの期待度数を算出する．それには上の二元表に加えて，生死と男女の二元表を作成する．

	男性	女性	行和
生存	246	142	388
死亡	234	178	412
列和	480	320	800

たとえば，男性でシートベルト着用して生存した人の期待度数は，男性でシートベルを

着用した 152 名と，男性で生存した 246 名、および男性の総数 480 名より

$$\frac{152\times 246}{480}=77.9$$

と計算される．他も同様で，それを３元表にまとめると

	男性		女性	
	生存	死亡	生存	死亡
シートベルト着用	77.9	74.1	110.05	137.95
非着用	168.1	159.9	31.95	40.05

となる．

各セルの観察度数を n_{ijk}，期待度数を m_{ijk} と記せば，３元表でのカイ２乗適合度統計量は

$$\chi^2=\sum_i\sum_j\sum_k\frac{(n_{ijk}-m_{ijk})^2}{m_{ijk}}$$

で与えられる．上で求めた数値を代入すると，7.42 となる．

このときの自由度は 2 であり，その上側 5%点は 5.99 である．

よって，男女ともにシートベルト着用の効果が有意に認められる．