経済統計概論レポート課題解答例

１．以下の問に答えよ．

(ア) 標本数が５０，標本平均が３，標本分散は４であった．このとき，母集団平均の９５％信頼区間を求めよ．

（解答例）

標本数が十分大きいので，$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ は正規分布で近似できる．また，σ は $\hat{\sigma}=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$ で近似できる．従って，$P\left(-1.96\le\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\le 1.96\right)=0.95$ の性質を使って９５％信頼区間を構成する．（）の中を変形すると $P\left(\overline{X}-1.96\dfrac{\hat{\sigma}}{\sqrt{n}}\le\mu\le\overline{X}+1.96\dfrac{\hat{\sigma}}{\sqrt{n}}\right)=0.95$ となる．σ の推定値としてＳを使用しても良いから，９５％信頼区間は，$3-1.96\dfrac{2}{\sqrt{50}}\le\mu\le 3+1.96\dfrac{2}{\sqrt{50}}$．計算すると，

$2.45\le\mu\le 3.55$ である．

(イ) ６００世帯を調査する視聴率調査で視聴率（「新選組！」初回視聴率）は２６．３％＜関東地区＞であった，ここから母集団に対する真の視聴率の９５％信頼区間を求めよ．

（解答例）

比率の推定を使用する．スライドシリーズ６の３１枚目によると，ｎを標本数，ｍを視聴している世帯数とすると，９５％信頼区間は，

$$\frac{m}{n}-1.96\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}\Big/\sqrt{n}\le\mu\le\frac{m}{n}+1.96\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}\Big/\sqrt{n}$$

であるから，$m/n=0.263$，$n=600$ を代入して，

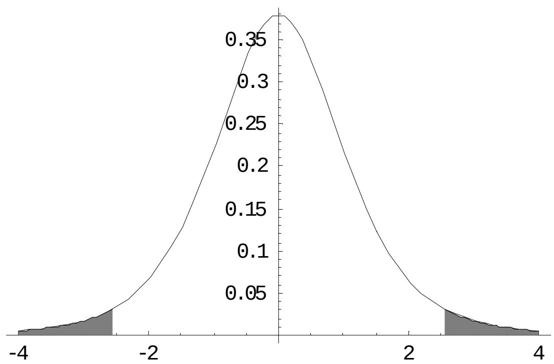
$$0.263-1.96\sqrt{0.263\times 0.737}\Big/\sqrt{600}\le\mu\le 0.196+1.96\sqrt{0.263\times 0.737}\Big/\sqrt{600}$$

となる．計算すると$0.228 \le \mu \le 0.298$であるから，９５％信頼区間は22.8%から29.8%ということになる．

(ウ) 標本のデータ値は，１，３，２，６，９であった，母集団の分布は正規分布であることがわかっている．このとき，母平均の９５％信頼区間を求めよ．

（解答例）

$\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$が自由度$n-1$のｔ分布に従う．ただし，$S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$である．自由度（５－１）のｔ分布の密度関数は下の通りであり，黒く塗ってある部分の確率合計は 0.05 である．この黒塗りの部分との境目の値は，教科書のｔ分布表の自由度（ｄｆ）４の行の.0.025 の列を参照すればよい．参照すると±2..776 であるから，９５％信頼区間は，$P\left(-2.776 \le \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \le 2.776\right)=0.95$を元に構成すればよい．



従って，$P\left(\overline{X}-2.776\dfrac{S}{\sqrt{n}} \le \mu \le \overline{X}+2.776\dfrac{S}{\sqrt{n}}\right)=0.95$である．標本平均は，４．２．

$$S^2=\frac{1}{5-1}\left\{(1-4.2)^2+(3-4.2)^2+(2-4.2)^2+(6-4.2)^2+(9-4.2)^2\right\}$$
$$=10.7$$
，

$S=\sqrt{10.7}=3.27$，$n=5$を代入して，$4.2-2.776\dfrac{3.27}{\sqrt{5}} \le \mu \le 4.2+2.776\dfrac{3.27}{\sqrt{5}}$．計算すると，$0.148 \le \mu \le 8.253$である．(答えは小数点以下１桁がプラスマイナス１で正解)

(エ) 集団 A については標本数が３１，平均が１０，標本分散は９であった．集団 B については標本数が５１，平均が１２，標本分散は１６であった．集団 A，B の母分散は等しいと仮定して，集団 A の母平均と集団 B の母平均が等しいか，等しくないか検定を行え.

（解答例）

＜理屈＞

帰無仮説は「両集団の母平均が等しい」である．標本数が十分大きいので，

$\dfrac{\overline{X}_{A集団}-\mu_{A集団}}{\sigma/\sqrt{n_{A集団}}}$，$\dfrac{\overline{X}_{B集団}-\mu_{B集団}}{\sigma/\sqrt{n_{B集団}}}$ はともに標準正規分布で近似でき無相関である．ここで，添え字の「A集団」はA集団内で計算した標本平均，母平均，標本数を示し，添え字「B集団」はB集団内で計算した標本平均，母平均，標本数を示す．

$Z_1=\dfrac{\overline{X}_{A集団}-\mu_{A集団}}{\sigma/\sqrt{n_{A集団}}}$，$Z_2=\dfrac{\overline{X}_{B集団}-\mu_{B集団}}{\sigma/\sqrt{n_{B集団}}}$ とおく．$\overline{X}_{A集団}-\mu_{A集団}=\left(\sigma/\sqrt{n_{A集団}}\right)Z_1$，$\overline{X}_{B集団}-\mu_{B集団}=\left(\sigma/\sqrt{n_{B集団}}\right)Z_2$ と書き換え，差をとると，$\left(\overline{X}_{A集団}-\overline{X}_{B集団}\right)-\left(\mu_{A集団}-\mu_{B集団}\right)=\left(\sigma/\sqrt{n_{A集団}}\right)Z_1-\left(\sigma/\sqrt{n_{B集団}}\right)Z_2$. さらに帰無仮説が正しいとすると，$\mu_{A集団}=\mu_{B集団}$ であるから，代入して，

$\overline{X}_{A集団}-\overline{X}_{B集団}=\left(\sigma/\sqrt{n_{A集団}}\right)Z_1-\left(\sigma/\sqrt{n_{B集団}}\right)Z_2$.

$E\left[\overline{X}_{A集団}-\overline{X}_{B集団}\right]=\left(\sigma/\sqrt{n_{A集団}}\right)E[Z_1]-\left(\sigma/\sqrt{n_{B集団}}\right)E[Z_2]=0$． Z_1,Z_2 は無相関だから，

$V\left[\overline{X}_{A集団}-\overline{X}_{B集団}\right]=\dfrac{\sigma^2}{n_{A集団}}V[Z_1]+\dfrac{\sigma^2}{n_{B集団}}V[Z_1]=\left(\dfrac{1}{n_{A集団}}+\dfrac{1}{n_{B集団}}\right)\sigma^2$． さらに，$\overline{X}_{A集団}-\overline{X}_{B集団}=\left(\sigma/\sqrt{n_{A集団}}\right)Z_1-\left(\sigma/\sqrt{n_{B集団}}\right)Z_2$ は無相関な正規分布の線形和で近似

できるから，それ自身正規分布で近似できる．以上を総合すると，$\overline{X}_{A集団}-\overline{X}_{B集団}$ は

$N\left[0,\left(\dfrac{1}{n_{A集団}}+\dfrac{1}{n_{B集団}}\right)\sigma^2\right]$ で近似できる．よって，$\dfrac{\overline{X}_{集団A}-\overline{X}_{集団B}}{\sigma\sqrt{\dfrac{1}{n_{A集団}}+\dfrac{1}{n_{B集団}}}}$ は標準正規分布

で近似できる．σはその推定値

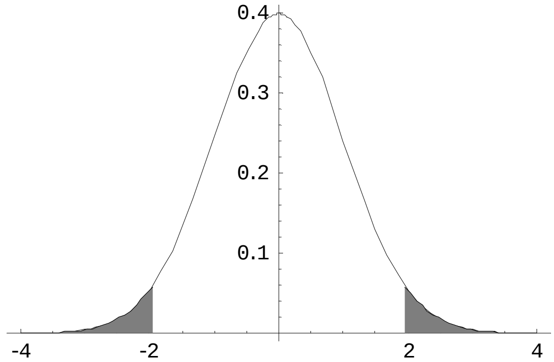
$\hat{\sigma}^2=\dfrac{1}{n_{A集団}+n_{B集団}}\left\{\sum_{集団A}\left(X_i-\overline{X}_{A集団}\right)^2+\sum_{集団B}\left(X_i-\overline{X}_{B集団}\right)^2\right\}$ で推定して，統計量は

$\dfrac{\overline{X}_{集団A}-\overline{X}_{集団B}}{\hat{\sigma}\sqrt{\dfrac{1}{n_{A集団}}+\dfrac{1}{n_{B集団}}}}$ で，境界値は標準正規分布表に基づいて決めることになる．

＜実際の解法＞

帰無仮説は「両集団の母平均が等しい」，対立仮説は「両集団の母平均が異なる」であるので, 両側検定となる. 統計量は $\dfrac{\overline{X}_{集団A}-\overline{X}_{集団B}}{\hat{\sigma}\sqrt{\dfrac{1}{n_{A集団}}+\dfrac{1}{n_{B集団}}}}$ で, 有意水準５%で検定すると,

境界値は帰無仮説の場合の分布の両側合計５%のはずれを考慮して決める.



上図の網掛け部が合計確率５%になるように決める. よって, ±1.96 が境界値である.

$S^2_{A集団}=\dfrac{1}{n-1}\sum_{集団A}\left(X_i-\overline{X}_{A集団}\right)^2$，$S^2_{B集団}=\dfrac{1}{n-1}\sum_{集団B}\left(X_i-\overline{X}_{B集団}\right)^2$ であるから,

$$\hat{\sigma}^2 = \frac{1}{n_{A集団} + n_{B集団}} \left\{ \sum_{集団A} \left(X_i - \overline{X}_{A集団} \right)^2 + \sum_{集団B} \left(X_i - \overline{X}_{B集団} \right)^2 \right\}$$

$$= \frac{1}{n_{A集団} + n_{B集団}} \left\{ \left(n_{A集団} - 1 \right) S^2_{A集団} + \left(n_{B集団} - 1 \right) S^2_{B集団} \right\}$$

$$= \frac{1}{31+51} \{ 31 \times 9 + 50 \times 16 \} = 13.16$$

よって，統計量の値は，$\dfrac{\overline{X}_{集団A} - \overline{X}_{集団B}}{\hat{\sigma}\sqrt{\dfrac{1}{n_{A集団}} + \dfrac{1}{n_{B集団}}}} = \dfrac{10-12}{3.63\sqrt{\dfrac{1}{31}+\dfrac{1}{51}}} = -2.42$．したがって，

統計量値＜－１．９６なので，統計量値は棄却域に入り，有意水準５％で帰無仮説は棄却され，対立仮説が採択される．したがって，二つの集団の母平均は異なるといえる．

(オ) （エ）と同じ状況で，集団 A，B の母分散は等しいと仮定して，集団 A の母平均より集団 B の母平均が大きいかどうか検定を行え．

（解答例）

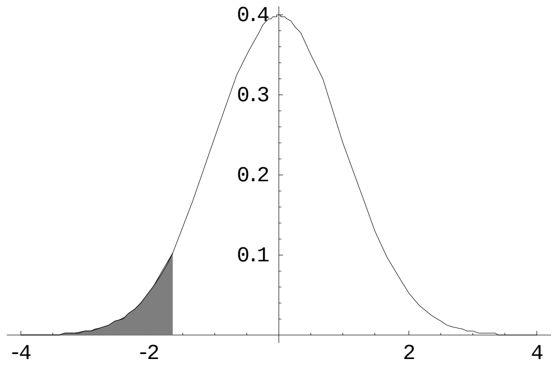
$\left(\overline{X}_{A集団} - \overline{X}_{B集団} \right) - \left(\mu_{A集団} - \mu_{B集団} \right) = \left(\sigma / \sqrt{n_{A集団}} \right) Z_1 - \left(\sigma / \sqrt{n_{B集団}} \right) Z_2$ より，

$\left(\overline{X}_{A集団} - \overline{X}_{B集団} \right) = \left(\mu_{A集団} - \mu_{B集団} \right) + \left(\sigma / \sqrt{n_{A集団}} \right) Z_1 - \left(\sigma / \sqrt{n_{B集団}} \right) Z_2$．従って，

$E\left(\overline{X}_{A集団} - \overline{X}_{B集団} \right) = \left(\mu_{A集団} - \mu_{B集団} \right)_2$，$V\left(\overline{X}_{A集団} - \overline{X}_{B集団} \right) = \sigma^2 \left(1/n_{A集団} + 1/n_{B集団} \right)$

である．対立仮説は，$\mu_{A集団} < \mu_{B集団}$ であるから，

$E\left(\overline{X}_{A集団} - \overline{X}_{B集団} \right) = \left(\mu_{A集団} - \mu_{B集団} \right) < 0$ となり，$\overline{X}_{A集団} - \overline{X}_{B集団} < 0$ の場合に棄却するのが妥当となる（片側検定）．したがって，統計量値が負の場合に棄却することになる．数で言うと，網掛けの部分が棄却域であるから，境界値は－1.64 である．統計量値がこの値より小さい場合に帰無仮説を棄却することになる．ところが，統計量値は（エ）によると－２．４２なので棄却域に入り，帰無仮説は有意水準５％で棄却される．



(カ) （エ）と同じ状況で，集団 A，B の母分散は等しくないと仮定して，集団 A の母平均より集団 B の母平均が等しいかどうか検定を行え.

（解答例）

＜理屈＞

$\dfrac{\overline{X}_{A集団}-\mu_{A集団}}{\sigma_{A集団}\big/\sqrt{n_{A集団}}}$，$\dfrac{\overline{X}_{\mathrm{B}集団}-\mu_{\mathrm{B}集団}}{\sigma_{\mathrm{B}集団}\big/\sqrt{n_{\mathrm{B}集団}}}$ はともに無相関の標準正規分布で近似できる．ただし，母分散が異なるので，それぞれの集団の母分散を $\sigma^2_{A集団},\sigma^2_{B集団}$ としてある.

$Z_1=\dfrac{\overline{X}_{A集団}-\mu_{A集団}}{\sigma_{A集団}\big/\sqrt{n_{A集団}}}$，$Z_2=\dfrac{\overline{X}_{B集団}-\mu_{B集団}}{\sigma_{\mathrm{B}集団}\big/\sqrt{n_{B集団}}}$ とおく．

$\left(\overline{X}_{A集団}-\overline{X}_{\mathrm{B}集団}\right)-\left(\mu_{A集団}-\mu_{\mathrm{B}集団}\right)=\dfrac{\sigma_{A集団}}{\sqrt{n_{A集団}}}Z_1-\dfrac{\sigma_{B集団}}{\sqrt{n_{B集団}}}Z_2$ ．帰無仮説 $\mu_{A集団}=\mu_{\mathrm{B}集団}$ の下では，$\overline{X}_{A集団}-\overline{X}_{\mathrm{B}集団}=\dfrac{\sigma_{A集団}}{\sqrt{n_{A集団}}}Z_1-\dfrac{\sigma_{B集団}}{\sqrt{n_{B集団}}}Z_2$ である.

$E\left[\overline{X}_{A集団}-\overline{X}_{\mathrm{B}集団}\right]=\dfrac{\sigma_{A集団}}{\sqrt{n_{A集団}}}E[Z_1]-\dfrac{\sigma_{B集団}}{\sqrt{n_{B集団}}}E[Z_2]=0$，$Z_1,Z_2$ は無相関だから，

$V\left[\overline{X}_{A集団}-\overline{X}_{\mathrm{B}集団}\right]=\dfrac{\sigma^2_{A集団}}{n_{A集団}}V[Z_1]+\dfrac{\sigma^2_{B集団}}{n_{B集団}}V[Z_2]=\dfrac{\sigma^2_{A集団}}{n_{A集団}}+\dfrac{\sigma^2_{B集団}}{n_{B集団}}$ ，また，

$\overline{X}_{A集団} - \overline{X}_{\mathrm{B}集団} = \dfrac{\sigma_{A集団}}{\sqrt{n_{A集団}}} Z_1 - \dfrac{\sigma_{B集団}}{\sqrt{n_{B集団}}} Z_2$ は無相関な正規分布の線形和で近似できるから，それ自身正規分布で近似できる．以上を総合すると，$\overline{X}_{A集団} - \overline{X}_{B集団}$ の帰無仮説の下での分布は $N\left[0, \dfrac{\sigma^2_{A集団}}{n_{A集団}} + \dfrac{\sigma^2_{B集団}}{n_{B集団}}\right]$ で近似できる．したがって，検定統計量は，

$\dfrac{\overline{X}_{A集団} - \overline{X}_{B集団}}{\sqrt{\dfrac{\hat{\sigma}^2_{A集団}}{n_{A集団}} + \dfrac{\hat{\sigma}^2_{B集団}}{n_{B集団}}}}$ でその分布は標準正規分布ということになる．ただし，

$\hat{\sigma}^2_{A集団} = \dfrac{1}{n} \sum_{A集団} \left(X_i - \overline{X}_A\right)^2$，$\hat{\sigma}^2_{B集団} = \dfrac{1}{n} \sum_{B集団} \left(X_i - \overline{X}_B\right)^2$ である．

＜実際の解法＞

帰無仮説は「両集団の母平均が等しい」，対立仮説は「両集団の母平均が異なる」であるので，両側検定となる．統計量は $\dfrac{\overline{X}_{A集団} - \overline{X}_{B集団}}{\sqrt{\dfrac{\hat{\sigma}^2_{A集団}}{n_{A集団}} + \dfrac{\hat{\sigma}^2_{B集団}}{n_{B集団}}}}$ で，有意水準５％で検定すると，境界値は帰無仮説の場合の分布の両側合計５％のはずれを考慮して決める．この時の境界値は ± 1.96 である．統計量の値を計算すると，$\dfrac{\overline{X}_{A集団} - \overline{X}_{B集団}}{\sqrt{\dfrac{\hat{\sigma}^2_{A集団}}{n_{A集団}} + \dfrac{\hat{\sigma}^2_{B集団}}{n_{B集団}}}} = \dfrac{10-12}{\sqrt{\dfrac{9\frac{30}{31}}{31} + \dfrac{16\frac{50}{51}}{51}}} = -2.61$ で，境界値の外側なので，帰無仮説を棄却する．

(キ) （エ）と同じ状況で母集団分散が集団 A と集団 B で等しいかどうか検定を行え．

（解答例）

母集団が正規分布に従うと仮定して F 検定を行う．帰無仮説：$\sigma^2_A = \sigma^2_B$，対立仮説：

$\sigma_A^2 \neq \sigma_B^2$である．統計量は,

$$\frac{S_{A集団}^2}{S_{B集団}^2} = \frac{\dfrac{1}{n_A - 1}\sum_{A集団}^{n}\left(X_i - \overline{X}_A\right)^2}{\dfrac{1}{n_B - 1}\sum_{B集団}^{n}\left(X_i - \overline{X}_B\right)^2} \cong \frac{\chi^2(n_A - 1)/(n_A - 1)}{\chi^2(n_B - 1)/(n_B - 1)} = F(n_A - 1, n_B - 1)$$

である．境界値は$F(30,50)$の 2.5%点と 97.5%点である．教科書の表は片側検定用の表なので，両側検定用の表を作成した．表には$F(n1, n2)$の 97.5%点を示した.

分母自由度 n2 \ 分子自由度 n1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	50	100
1	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	976.72	982.55	986.91	990.35	993.08	1001.40	1008.10	1013.16
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.44	39.44	39.45	39.46	39.48	39.49
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.28	14.23	14.20	14.17	14.08	14.01	13.96
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.68	8.63	8.59	8.56	8.46	8.38	8.32
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.46	6.40	6.36	6.33	6.23	6.14	6.08
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.30	5.24	5.20	5.17	5.07	4.98	4.92
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.60	4.54	4.50	4.47	4.36	4.28	4.21
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.13	4.08	4.03	4.00	3.89	3.81	3.74
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.80	3.74	3.70	3.67	3.56	3.47	3.40
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.55	3.50	3.45	3.42	3.31	3.22	3.15
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.21	3.15	3.11	3.07	2.96	2.87	2.80
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.98	2.92	2.88	2.84	2.73	2.64	2.56
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.82	2.76	2.72	2.68	2.57	2.47	2.40
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.70	2.64	2.60	2.56	2.44	2.35	2.27
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.60	2.55	2.50	2.46	2.35	2.25	2.17
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.34	2.28	2.23	2.20	2.07	1.97	1.88
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22	2.14	2.08	2.03	1.99	1.87	1.75	1.66
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.08	2.00	1.94	1.89	1.85	1.71	1.59	1.48

ここから，$F(30,50)$の 97.5%点は 1.87, 2.5%点は$F(50,30)$の 97.5%点の逆数なので，$1/1.97 = 0.51$である．一方統計量の値は，$\dfrac{S_A^2}{S_B^2} = \dfrac{9}{16} = 0.56$で，棄却域にはいらないので有意水準５%で帰無仮説は棄却できない．従って，集団Ａの分散とＢの分散は異ならない．したがって，等分散を仮定した平均差の検定を使う方がよい.

(ク) 集団 A は標本数が１０，標本平均が６，標本分散は４，集団 B は標本数が５，標本平均が３，標本分散は１である．集団 A,B ともに母集団の分布は正規分布で分散は等しいと仮定して，集団 A と集団 B の母平均が等しいか検定せよ.

（解答例）

帰無仮説 $\mu_{A集団} = \mu_{B集団}$，対立仮説 $\mu_{A集団} \neq \mu_{B集団}$ として検定を行う．統計量 $\dfrac{\overline{X}_{A集団} - \overline{X}_{B集団}}{S\sqrt{\dfrac{1}{n_{A集団}} + \dfrac{1}{n_{B集団}}}}$ は帰無仮説の下で自由度$n_{A集団} + n_{B集団} - 2$の t 分布に従う. この場

合は，自由度は１０＋５－２＝１３である．したがって，有意水準５%の場合の境界値は±2.161 である．

$$S^2 = \frac{1}{n_A + n_B - 2}\left\{\sum_{\text{A集団}}\left(X_i - \overline{X}_A\right)^2 + \sum_{B\text{集団}}\left(X_i - \overline{X}_B\right)^2\right\}$$
$$= \frac{1}{n_A + n_B - 2}\left\{(n_A - 1)S^2_{A\text{集団}} + (n_{\mathrm{B}} - 1)S^2_{\mathrm{B}\text{集団}}\right\}$$
$$= \frac{1}{13}\{9\times 4 + 4\times 1\} = \frac{40}{13} = 3.08$$

$S = \sqrt{1.9} = 1.75$ より，統計量の値は，$\dfrac{6-3}{1.75\sqrt{\dfrac{1}{10}+\dfrac{1}{5}}} = 3.13$ となり，棄却域にはいるので，帰無仮説は棄却される．

(ケ) ６００世帯に対する視聴率調査の結果，今年の NHK 大河ドラマ「新撰組！」の初回視聴率は２６．３%，昨年の「武蔵　MUSASHI」の初回視聴率は２１．７ %であった（いずれも関東地区）．一般には今年の大河ドラマは昨年より好調と言われているが，母集団で考えた視聴率で「新撰組！」の第１回目が「武蔵 MUSASHI」の第１回目を上回っているとまでいえるかどうかを検定せよ．

（解答例）

サンプル調査数をｎ，「新選組！」の視聴世帯数 m_1，「武蔵　MUSASHI」の視聴世帯数を m_2，「新選組！」の真の視聴率を p_1，「武蔵　MUSASHI」の真の視聴率を p_2 とする．帰無仮説は「二つの視聴率が等しい，$p_1 = p_2$」，対立仮説は「$p_1 > p_2$」ととる． $\dfrac{\sqrt{n}(m_1/n - p_1)}{\sqrt{p_1(1-p_1)}}$，$\dfrac{\sqrt{n}(m_2/n - p_2)}{\sqrt{p_2(1-p_2)}}$ はともに標準正規分布する．

$Z_1 = \dfrac{\sqrt{n}(m_1/n - p_1)}{\sqrt{p_1(1-p_1)}}$，$Z_2 = \dfrac{\sqrt{n}(m_2/n - p_2)}{\sqrt{p_2(1-p_2)}}$ とすると，

$$\frac{m_1}{n}-\frac{m_2}{n}=(p_1-p_2)+\left(\frac{\sqrt{p_1(1-p_1)}}{\sqrt{n}}Z_1-\frac{\sqrt{p_2(1-p_2)}}{\sqrt{n}}Z_2\right)$$

であるが,

$$E\left[\frac{m_1}{n}-\frac{m_2}{n}\right]=(p_1-p_2)+\left(\frac{\sqrt{p_1(1-p_1)}}{\sqrt{n}}E[Z_1]-\frac{\sqrt{p_2(1-p_2)}}{\sqrt{n}}E[Z_2]\right)$$

$$=p_1-p_2$$

$$V\left[\frac{m_1}{n}-\frac{m_2}{n}\right]=\left(\frac{p_1(1-p_1)}{n}V[Z_1]+\frac{p_2(1-p_2)}{n}V[Z_2]\right)$$

$$=\frac{p_1(1-p_1)+p_2(1-p_2)}{n}$$

であるから,

$$\frac{m_1}{n}-\frac{m_2}{n}\cong N\left(p_1-p_2,\frac{p_1(1-p_1)+p_2(1-p_2)}{n}\right)$$

である. 帰無仮説 $p_1=p_2$ の下では, $\frac{m_1}{n}-\frac{m_2}{n}\cong N\left(0,\frac{p_1(1-p_1)+p_2(1-p_2)}{n}\right)$ となる.

ここから, $\left(\frac{m_1}{n}-\frac{m_2}{n}\right)\Big/\sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{n}}\cong N(0,1)$ となる. p_1, p_2 の推定値は,

それぞれ $m_1/n, m_2/n$ とすればよいから, 検定統計量は,

$$\left(\frac{m_1}{n}-\frac{m_2}{n}\right)\Big/\sqrt{\frac{m_1/n(1-m_1/n)+m_2/n(1-m_2/n)}{n}}$$

$$=(0.263-0.217)\Big/\sqrt{\frac{0.263\times0.737+0.217\times0.783}{600}}=1.87$$

である, 有意水準５%とすれば, これが１．６４を上回るかである. この場合, 棄却域に入っているので, 帰無仮説は棄却でき,「新撰組！」は「ＭＵＳＡＳＨＩ」の初回視聴率を上回っている.

(コ) 今度は昔の幕末を扱った大河ドラマと比較する. ６００世帯に対する視聴率調査の結果, 今年の NHK 大河ドラマ「新撰組！」の初回視聴率は２６．３%, １９７

４年の「勝海舟」の初回視聴率は４３０世帯に対する視聴率調査の結果，３０．５％であった．母集団で考えた視聴率で「勝海舟」の第１回目が「新撰組！」の第１回目を上回っているとまでいえるかどうか検定せよ．

（解答例）

帰無仮説は「二つの視聴率が等しい，$p_1 = p_2$」，対立仮説は「$p_1 < p_2$」である．（ケ）とほぼ同様であるが，調査世帯数の変化を考慮する必要がある．現在の調査世帯数をn_1，74 年当時の調査世帯数をn_2とすると，

$$\frac{m_1}{n_1}-\frac{m_2}{n_2}=\left(p_1-p_2\right)+\left(\frac{\sqrt{p_1\left(1-p_1\right)}}{\sqrt{n_1}}Z_1-\frac{\sqrt{p_2\left(1-p_2\right)}}{\sqrt{n_2}}Z_2\right)$$

となり，帰無仮説が正しければ，$\left(\frac{m_1}{n_1}-\frac{m_2}{n_2}\right)\Big/\sqrt{\frac{p_1\left(1-p_1\right)}{n_1}+\frac{p_2\left(1-p_2\right)}{n_2}} \cong N(0,1)$ となる．p_1, p_2 の推定値は，それぞれ $m_1/n, m_2/n$ とすればよいので，統計量は $\left(\frac{m_1}{n_1}-\frac{m_2}{n_2}\right)\Big/\sqrt{\frac{m_1/n_1\left(1-m_1/n_1\right)}{n_1}+\frac{m_2/n_2\left(1-m_2/n_2\right)}{n_2}}$ である．境界値は片側検定となり，対立仮説の下では検定統計量の分子負にになりがちであるから，棄却域は負の範囲となる．具体的には，有意水準を５％とすれば，標準正規分布の 95%点が境界値となる．したがって，－１．６４である．統計量の値は,

$$\left(\frac{m_1}{n_1}-\frac{m_2}{n_2}\right)\Big/\sqrt{\frac{m_1/n_1\left(1-m_1/n_1\right)}{n_1}+\frac{m_2/n_2\left(1-m_2/n_2\right)}{n_2}}$$
$$=\left(0.263-0.305\right)\Big/\sqrt{\frac{0.263\times 0.737}{600}+\frac{0.305\times 0.695}{430}}=-1.47$$

となり,「新選組！」の初回視聴率を「勝海舟」の初回視聴率が上回ったとまではいえない．

２．相関係数の意味について以下の手順に従って考察せよ．

（ア）　２次元ベクトル$\vec{x}=\left(x_1,x_2\right)$と$\vec{y}=\left(y_1,y_2\right)$との内積は$\vec{x}\cdot\vec{y}=x_1y_1+x_2y_2$である．また，このふたつのベクトルのなす角をθとすると，$\cos\theta=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}=\dfrac{x_1y_1+x_2y_2}{\sqrt{x_1^2+x_2^2}\sqrt{y_1^2+y_2^2}}$である．以下のようなデータについて属性１と属性２の値の標本相関係数を式で表し答えよ．

	属性1の値	属性2の値
標本1	x1	y1
標本2	x2	y2

（答え）

$\bar{x},\bar{y}$をｘ，ｙのそれぞれの標本平均とすると，

$$r_{xy}=\frac{(x1-\bar{x})(y1-\bar{y})+(x2-\bar{x})(y2-\bar{y})}{\sqrt{(x1-\bar{x})^2+(x2-\bar{x})^2}\sqrt{(y1-\bar{y})^2+(y2-\bar{y})^2}}$$

（イ）　（ア）で求めた属性１と属性２の値の標本相関係数の式から，標本相関係数はどのようなベクトルどうしの角の cos とみなせるか答えよ．

（答え）

属性１の値からその標本平均を引いた値のベクトルと，属性２の値からその標本平均を引いた値のベクトルとのなす角のｃｏｓである．

（ウ）　（イ）の結果から，標本相関係数が－１から１の間にあることと，属性１と属性２の間の関係が深いとなぜ標本相関係数の絶対値が大きくなるかを説明せよ．また，標本相関係数が１になる場合，－１になる場合はそれぞれどんな場合か答えよ．

（答え）

$-1\leq\cos\theta\leq1$より，相関係数は－１から１の間になる．属性１と属性２の間の関係が深い

と，属性１の値からその標本平均を引いた値のベクトルと，属性２の値からその標本平均を引いた値のベクトルが似てくる．ということは，それらのなす角が小さくなるか，どちらかのベクトルを逆にとったベクトルとのなす角が小さくなる．前者の場合，ｃｏｓの値は大きくなり１に近い，後者の場合，ｃｏｓの値は小さくなり－１にちかい．いずれの場合も，絶対値が大きくなる．標本相関係数が１になる場合は，二つのベクトルが平行で同じ方向であるから，すなわち，$(y_i-\bar{y})=\beta(x_i-\bar{x})\quad(\beta>0)$となる場合である．つまり，$y_i=\beta(x_i-\bar{x})+\bar{y}\quad(\beta>0)$となり，ｙがｘの一次式で完全にかけ，一次式の傾きがプラスの場合である．標本相関係数が－１になる場合は，二つのベクトルが平行で逆方向であるから，，すなわち，$(y_i-\bar{y})=\beta(x_i-\bar{x})\quad(\beta<0)$となる場合である．つまり，$(y_i-\bar{y})=\beta(x_i-\bar{x})\quad(\beta<0)$となり，ｙがｘの一次式で完全にかけ，その一次式の傾きが負の場合である．

（エ）　（ア）をｎ次元ベクトルに拡張すると，ｎ次元ベクトル$\vec{x}=(x_1,x_2,\cdots,x_n)$と$\vec{y}=(y_1,y_2,\cdots,y_n)$との内積は，$\vec{x}\cdot\vec{y}=x_1y_1+x_2y_2+\cdots+x_ny_n$である．また，この二つのベクトルのなす角をθとすると$\cos\theta=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}=\dfrac{x_1y_1+x_2y_2+\cdots+x_ny_n}{\sqrt{x_1^2+x_2^2+\cdots+x_n^2}\sqrt{y_1^2+y_2^2+\cdots+y_n^2}}$となる．ここから，標本相関係数はどのようなベクトルどうしの角の cos とみなせるか答えよ．

属性１の値からその標本平均を引いた値のベクトルと，属性２の値からその標本平均を引いた値のベクトルとのなす角のｃｏｓである．ただ，ｎ次元に拡張されているので，データの標本数も２より大きくてもよくなる．

（オ）　二つのｎ次元ベクトル$\vec{x}=(x_1,x_2,\cdots,x_n)$と$\vec{y}=(y_1,y_2,\cdots,y_n)$において，$\vec{y}=(y_1,y_2,\cdots,y_n)$の$\vec{x}=(x_1,x_2,\cdots,x_n)$に対する射影は

$P(\vec{y}|\vec{x})=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}=\dfrac{|\vec{y}|\cos\theta}{|\vec{x}|}\vec{x}$ である．このとき，$\vec{e}=\vec{y}-P(\vec{y}|\vec{x})$ とすると，$\vec{e}\cdot\vec{x}=0$ であることを示せ. さらに，$|\vec{y}|^2=|P(\vec{y}|\vec{x})|^2+|\vec{e}|^2$ であることを示せ. そして，$\cos^2\theta$ を $|\vec{y}|^2,|\vec{e}|^2$，及び，$|\vec{y}|^2,|P(\vec{y}|\vec{x})|^2$ で表せ.

（解答）

$$\vec{e}\cdot\vec{x}=\{\vec{y}-P(\vec{y}|\vec{x})\}\cdot\vec{x}=\left\{\vec{y}-\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}\right\}\cdot\vec{x}=\vec{x}\cdot\vec{y}-\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}\cdot\vec{x}.$$

$\vec{x}\cdot\vec{x}=|\vec{x}|^2$ より，$\vec{e}\cdot\vec{x}=\vec{x}\cdot\vec{y}-\vec{x}\cdot\vec{y}=0$

$\vec{y}=P(\vec{y}|\vec{x})+\vec{e}$ であるから，

$$|\vec{y}|^2=\vec{y}\cdot\vec{y}=\{P(\vec{y}|\vec{x})+\vec{e}\}\cdot\{P(\vec{y}|\vec{x})+\vec{e}\}=P(\vec{y}|\vec{x})\cdot P(\vec{y}|\vec{x})+P(\vec{y}|\vec{x})\cdot\vec{e}+\vec{e}\cdot P(\vec{y}|\vec{x})+\vec{e}\cdot\vec{e}$$

ところが，$P(\vec{y}|\vec{x})\cdot\vec{e}=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}\cdot\vec{e}=0$，$\vec{e}\cdot P(\vec{y}|\vec{x})=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{e}\cdot\vec{x}=0$，$P(\vec{y}|\vec{x})\cdot P(\vec{y}|\vec{x})=|P(\vec{y}|\vec{x})|^2$，

$\vec{e}\cdot\vec{e}=|\vec{e}|^2$ であるから，$|\vec{y}|^2=|P(\vec{y}|\vec{x})|^2+|\vec{e}|^2$.

$P(\vec{y}|\vec{x})=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}=\dfrac{|\vec{y}|\cos\theta}{|\vec{x}|}\vec{x}$ より，$|P(\vec{y}|\vec{x})|^2=\dfrac{|\vec{y}|^2\cos\theta^2}{|\vec{x}|^2}|\vec{x}|^2$．したがって，

$$\cos\theta^2=\frac{|P(\vec{y}|\vec{x})|^2}{|\vec{y}|^2}=\frac{|\vec{y}|^2-|\vec{e}|^2}{|\vec{y}|^2}=1-\frac{|\vec{e}|^2}{|\vec{y}|^2}.$$

（カ）　（エ）と（オ）を総合すると，

$$(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})=\frac{(\qquad)}{(\qquad)}(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})+\vec{e}$$

ただし，$\vec{e}$ と $(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$ の標本相関係数が０としたとき．$r_{x,y}^2$ を $\sum_{i=1}^{n}(y_i-\bar{y})^2$ と $\sum_{i=1}^{n}e_i^2$ で表すと（　　　　　　　　）である．()内を埋めよ.

$\vec{x}=(x_1-\bar{x},\cdots,x_n-\bar{x})$，$\vec{y}=(y_1-\bar{y},\cdots,y_n-\bar{y})$とおいて，$\vec{y}=P(\vec{y}|\vec{x})+\vec{e}$を考えると，$P(\vec{y}|\vec{x})=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_1-\bar{x},\cdots,x_n-\bar{x})$であるから，

$$(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})+\vec{e}\,.$$

このとき，$\vec{e}$の各要素e_iは，$e_i=y_i-\bar{y}-\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i-\bar{x})$となる．$\vec{e}$の各要素をデータ値とみたときの標本平均は，

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}e_i&=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\\&=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{n}\sum_{i=1}^{n}\bar{y}-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left\{\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\bar{x}\right\}\\&=\bar{y}-\frac{1}{n}(n\bar{y})-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\bar{x}-\frac{1}{n}n\bar{x}\right)\\&=0\end{aligned}$$

よって，，$\vec{e}$と$(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$の標本相関係数は，$\vec{e}\cdot\vec{x}=\sum_{i=1}^{n}e_i(x_i-\bar{x})=0$より，

$$\frac{\sum_{i=1}^{n}(e_i-\bar{e})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(e_i-\bar{e})^2}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\frac{\sum_{i=1}^{n}e_i(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}e_i^2}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}=0$$

となって，題意を満たす．

$r_{x,y}^2$は，（エ）の答えより，$\vec{x}=(x_1-\bar{x},\cdots,x_n-\bar{x})$，$\vec{y}=(y_1-\bar{y},\cdots,y_n-\bar{y})$とおき，二つのベクトルのなす角θとしたときの，$\cos^2\theta$であるから，（オ）の答えより，$r_{x,y}^2=\cos\theta^2=\dfrac{|\vec{y}|^2-|\vec{e}|^2}{|\vec{y}|^2}=1-\dfrac{|\vec{e}|^2}{|\vec{y}|^2}$ となる．$|\vec{y}|^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$，

$|\vec{e}|^2=\sum_{i=1}^{n}e_i^2$ であるから，$r_{x,y}^2=\cos\theta^2=1-\dfrac{|\vec{e}|^2}{|\vec{y}|^2}=1-\dfrac{\sum_{i-1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

（キ）　$\sum_{i=1}^{n}\{(y_i-\bar{y})-\beta(x_i-\bar{x})\}^2$ を最小にするβを求めると，$(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})$というベクトルの先端と$(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$というベクトル上の点$\beta(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$を結んだ距離を最小にするβを求める問題に還元できるので，βは（　　　　　）である．

高校の問題であるとおり，ある点とある直線上の点との距離を最小にする点は，その点からその直線への垂線をおろしたいわゆる垂線の足である．この問題では，点の座標が，$(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})$，直線上の点の座標が，$\beta(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$となっていると考える．垂線の足では，垂線を表すベクトル$[y_1-\bar{y}-\beta(x_1-\bar{x}),\cdots,y_n-\bar{y}-\beta(x_n-\bar{x})]$と直線上のベクトル（直線の方向ベクトル）$(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$が直交する．つまり，垂線の足は$(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})$の$(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$に対する射影と一致している．よって，

$\beta(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})=\dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}|^2}\vec{x}=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_1-\bar{x},\cdots,x_n-\bar{x})$ となり，

$\sum_{i=1}^{n}\{(y_i-\overline{y})-\beta(x_i-\overline{x})\}^2$ を最小にする β は，$\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$ となる．