問2

確率 p で起こる事象が初めて起こるまでの回数、つまり幾何分布を扱う問題。

平均が 1/p であることを応用すれば解ける。

■ 記述5

初めて3種類のカードが揃うまでの期待値は、次の3つの期待値の和を取れば良い。

① 初めて1種類のカードが揃う回数

② 1種類のカードがある状態で、初めて2種類のカードが揃う回数

③ 2種類のカードがある状態で、初めて3種類のカードが揃う回数

それぞれが幾何分布に従うことを用いると、答えは、

$$\underbrace{1}_{①\ p=1} + \underbrace{\frac{1}{2/3}}_{②\ p=\frac{2}{3}} + \underbrace{\frac{1}{1/3}}_{③\ p=\frac{1}{3}} = 5.5$$

■ 記述6

x と y それぞれの値を、記述5と同じ枠組みによって求め、差を取れば良い。

$$x = 5.5 + \frac{1}{1/4} = 9.5$$

$$y = \frac{1}{1} + \frac{1}{3/4} + \frac{1}{1/2} + \frac{1}{1/4} = \frac{25}{3}$$

よって、

$$x - y = 9.5 - \frac{25}{3} = \frac{7}{6}$$

問4

シンプルな適合度検定の問題。ぜひささっと解きたい。

［1］

分割表より、

$$190 \cdot \frac{135}{300} = 85.5$$

［2］

安定の解き方と簡略法、どちらもできると良い。

期待度数表

85.5	49.5
104.5	60.5

より.

$$\frac{(93-85.5)^2}{85.5} + \frac{(42-49.5)^2}{49.5} + \frac{(97-104.5)^2}{104.5} + \frac{(68-60.5)^2}{60.5}$$

$$= \frac{(7.5)^2}{85.5} + \frac{(7.5)^2}{49.5} + \frac{(7.5)^2}{104.5} + \frac{(7.5)^2}{60.5}$$

分子は等しくなる

$$= 3.262$$

or

$$\frac{300(93 \cdot 68 - 42 \cdot 97)^2}{190 \cdot 110 \cdot 135 \cdot 165} = 3.262$$

［3］

自由度1のカイ二乗分布表を見ると、3.262 は 0.10 と 0.05 の間に位置するので、有意水準 10% で棄却されるが、有意水準 5% では棄却されない。②が正解

問6

主成分分析とAICによるモデル選択についての設問。最後の正誤問題以外は易しい。正誤問題について押さえるべき点は2点ある：(i) 因子負荷量は主成分と元の変量の相関係数と一致する。 (ii) AIC基準は、モデル同定の一致性を満たさない。

［1］

寄与率は固有値を正規化したものである。累積寄与率が初めて 0.8 を超えるような因子の個数をカウントすれば良い。③ 第5主成分

［2］

表から固有ベクトルの1次元目と2次元目の値を取ってきて、プロットしてみれば解ける。①

［3］

AICが1番小さいのは②モデル4

［4］

① 主成分分析は標準化していない変量に対しても適用できる。

② 主成分分析は、分散共分散行列の固有値問題を解く方法と、相関係数行列の固有値問題を解く方法の両方で実施できる。特に相関係数行列で解くと、**因子負荷量（主成分負荷量）が主成分と元の変量の相関係数に一致する。**

③ ×

④ モデル同定の一致性とは、サンプルを増やすことで正しいモデルが選択される確率が1に収束することである。AIC は成り立たないらしい。（モデルパラメータ数でペナルティを課しているだけなので、なんとなく成り立たなそうではある）

⑤ AIC は全データに対して1度学習を実行するだけで算出できるが、交差検証法は学習と検証を繰り返し実施するので、交差検証法の方が計算量が大きくなりがち。

問8

L1正則化による平滑化の問題。重回帰の正則化は散々やったが、平滑化の正則化は初であり、数式の形から**どの式が 0 を取りやすいのかを判断する**ことが求められる。

［1］

式を見ると、一見普通の回帰式の誤差のように見えるが全然違う。サンプル i ごとに β_i が設定されるので、正則化項がない場合 $\beta_i = y_i$ となる。

回帰のL1正則化のように、絶対値で囲まれた部分の多くが0となるような効果が働くことに注意すると、$\beta_t = \beta_{t+1}$ がいくつかの t において成立することが想像できる。これより、④が正解

［2］

平滑化のプロットを見ると、角ばった線になっており、しばらく直線ののち角にあたり、またしばらく直線ののち角に当たるというをくり返している。この「角ばった」というのが $\beta_{i+2} - \beta_{i+1}$ と $\beta_{i+1} - \beta_i$ の差が 0 となるような i が多いということを意味するのに気付ければ解ける。（だいぶ難しい）

① βが0に近くなるはずだが、そうはなっていないのでバツ

② これもβが0に近くなっていないのでバツ

③ これもβが0に近くなっていないのでバツ

④ $(\beta_{i+2} - \beta_{i+1}) - (\beta_{i+1} - \beta_i) = \beta_{i+2} - 2\beta_{i+1} + \beta_i$ が0に近づくので正解

⑤ ？？？

問10

ロジスティック回帰のモデルの式がわかっているかをメインで問う問題。他も計算問題なので、ぜひ解けておきたい。

［1］

Yi は 0 か 1 を取りうるので、ベルヌーイ分布である。よって①が正解

［2］

ロジスティック回帰は、オッズ比の対数を線型回帰でフィッティングするモデルなので、①が正解

［3］

下記方程式を解けば良い。

$$\log\left(\frac{0.5}{1-0.5}\right) = 15.0429 - 0.2322x$$

答えは④

［4］

下記方程式を解けば良い。

$$\log\left(\frac{\pi}{1-\pi}\right) = 15.0429 - 0.2322 \cdot 31$$

$$= 7.8447$$

$$\Leftrightarrow \frac{\pi}{1-\pi} = e^{7} \cdot e^{0.8447}$$

分けてから付表使う。

$$= (2.7183)^{7} \cdot 2.3396$$

$$\therefore \pi \fallingdotseq 0.9996$$

⑤

問12

時系列分析で出てくる2つのモデル、ARモデル・MAモデルの性質を問う問題。［1］は二次方程式の解の公式と、数値の大小関係を見極めれば解ける。［2］はコレログラムから一定以上のラグで自己相関係数が0になることを議論する。

［1］

設問より、AR(2) モデルが定常となる必要十分条件は、次の方程式の解の絶対値が1より大きいことである。

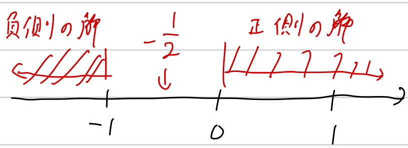
$$1 - a_1 z - a_2 z^2 = 0$$

$$\Leftrightarrow \quad a z^2 + a z - 1 = 0 \qquad (\because a_1 = a_2 = a)$$

判別式 $D = a^2 + 4a > 0$ より、これは2つの解を持つ。

$$z = \frac{-a \pm \sqrt{a^2+4a}}{2a} = \frac{-1 \pm \sqrt{1+4/a}}{2}$$

$\sqrt{1+4/a}$ は、$(1, \infty)$ の範囲をとるので、$-\frac{1}{2}$ を中心とした次の解となる。



全ての解の絶対値が1より大きいことが求めたい条件なので、

$$\frac{-1 + \sqrt{1+4/a}}{2} > 1 \qquad 1 + 4/a > 9$$

$$\therefore \quad a < \frac{1}{2} \qquad \frac{4}{8} > a$$

［2］は解答参照

論述問1

前半は分散分析で、後半はt検定と検定の多重性の問題。

［1］

もし制約がない場合に、母数が一意に定まらないため推定できない、ということを記述すれば良い。母数 μ, α_i と母平均 μ_i の関係は、うまく等号が成り立つような式をモデルの式から導出する。（ $\mu_i = \mu + \alpha_i$ じゃだめなの？と思った）

① $\sum_{i=1}^{4} \alpha_i = 0$ のとき、 $\mu_i = \mu + \alpha_i$ に $\sum_{i=1}^{4}$ をとって、

$$\sum_{i=1}^{4} \mu_i = \sum_{i=1}^{4} \mu + \sum_{i=1}^{4} \alpha_i$$

$$\Longleftrightarrow \sum_{i=1}^{4} \mu_i = 4\mu + 0$$

$$\therefore \mu = \frac{1}{4} \sum_{i=1}^{4} \mu_i$$

② $\sum_{i=1}^{4} n_i \alpha_i = 0$ のとき $n_i \mu_i = n_i \mu + n_i \alpha_i$ に $\sum_{i=1}^{4}$ をとって、

$$\sum_{i=1}^{4} n_i \mu_i = \mu \underbrace{\sum_{i=1}^{4} n_i}_{= n \text{とおく。}} + \underbrace{\sum_{i=1}^{4} n_i \alpha_i}_{0}$$

$$\therefore \mu = \frac{1}{n} \sum_{i=1}^{4} n_i \mu_i$$

[2]

一般的な一元配置分散分析を実施する。二乗和の計算が厄介なので、計算ミスだけ気をつける。

$$S_A = 4 \cdot 0.5^2 + 5 \cdot 2.4^2 + 5 \cdot 0.4^2 + 4 \cdot 2^2$$
$$= 1 + 12 \cdot 2.4 + 0.8 + 16$$
$$= 46.6$$

$$S_E = 3.5^2 + 0.5^2 + 1.5^2 + 2.5^2$$
$$+ 2.6^2 + 0.4^2 + 2.4^2 + 0.4^2 + 0.6^2$$
$$+ 1.6^2 + 0.6^2 + 0.4^2 + 0.6^2 + 2.4^2$$
$$+ 1 + 0 + 3^2 + 2^2$$
$$= 57.4$$

	平方和	自由度	分散	F
A	46.6	3	15.53	3.788
誤差	57.4	14 $= \nu$	4.1	
計	104.0	17		

$F_{0.05}(3, 14) < F(3, 10) < 3.788$ より,

5%有意水準で有意である.

母分散 σ^2 の不偏推定量は、表の誤差分散の値から、4.1

[3]

検定統計量 T が t 分布に従うように、係数 c を求める。平均 $\overline{Y_1}$ $\overline{Y_2}$ $\overline{Y_3}$ $\overline{Y_4}$ が漸近的に正規分布に従うことを利用する。

$$\overline{Y_1} \sim N\left(\mu_1,\ \frac{1}{n_1}\sigma^2\right)$$
$$\overline{Y_2} \sim N\left(\mu_2,\ \frac{1}{n_2}\sigma^2\right)$$
$$\overline{Y_3} \sim N\left(\mu_3,\ \frac{1}{n_3}\sigma^2\right)$$
$$\overline{Y_4} \sim N\left(\mu_4,\ \frac{1}{n_4}\sigma^2\right)$$

これより、

$$\frac{\overline{Y_1}+\overline{Y_2}}{2} - \frac{\overline{Y_3}+\overline{Y_4}}{2} = \frac{1}{2}\overline{Y_1} + \frac{1}{2}\overline{Y_2} - \frac{1}{2}\overline{Y_3} - \frac{1}{2}\overline{Y_4}$$
$$\sim N\left(\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2},\ \frac{\sigma^2}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}\right)\right)$$

帰無仮説の元で

$$\sim N\left(0,\ \frac{\sigma^2}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}\right)\right)$$

したがって、 $c = \frac{\hat{\sigma}}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}}$ として、

$$T = \frac{1}{c}\left(\frac{\overline{Y_1}+\overline{Y_2}}{2} - \frac{\overline{Y_3}-\overline{Y_4}}{2}\right) \sim t(\overset{14}{\nu})$$

なぜかはナゾ

表から値を代入すると、

$$c = \frac{\sqrt{4.1}}{2}\sqrt{\frac{1}{4}+\frac{1}{5}+\frac{1}{5}+\frac{1}{4}} \fallingdotseq 0.960$$ より、

$$T = \frac{1}{0.960}\left(\frac{30.5+27.6}{2} - \frac{30.4+32.0}{2}\right) = -2.24$$

これに対して両側検定を実施する。

→ 解答のように、5% と 10% 両方やるのが良さそう。

[4] は解答参照