直前確認事項

統計検定準 1 級

1 分布

- 幾何分布 (2019#2)

  ベルヌーイ試行で、初めて成功するまでの回数 (最後の 1 回を含む)

  $$P(X=n)=(1-p)^{n-1}p$$
  $$E(X)=1/p$$
  $$V(X)=1/p^2-1/p$$

- 超幾何分布 (2017#9, 2018#2)

  つぼの中に N 個の玉 (うち M 個が赤) がある。n 個取り出したときに k 個が赤である事象

  $$P(X=k)=\frac{\binom{M}{k}\times\binom{N-M}{n-k}}{\binom{N}{n}}$$
  $$E(X)=np \quad (p:=M/N)$$
  $$V(X)=\frac{N-n}{N-1}\,npq$$

- 指数分布

  $$P(X=x)=\frac{1}{\lambda}e^{-\frac{x}{\lambda}}$$
  $$E(X)=\lambda$$
  $$V(X)=\lambda^2$$

- ポアソン分布

  $$P(X=n)=e^{-\lambda}\frac{\lambda^n}{n!}$$
  $$E(X)=\lambda$$
  $$V(X)=\lambda$$

  再生性 $X_j\sim P_o(\lambda_j)\Rightarrow X_1+X_2\sim P_o(\lambda_1+\lambda_2)$

- 2 変量正規分布

$$f_2(x_1,\ x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right]$$

$$E(X) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

条件付き期待値・分散

$$E[X_2|X_1 = x_1] = \mu_2 + \rho\sigma_2\frac{x_1-\mu_1}{\sigma_1}$$

$$V[X_2|X_1 = x_1] = \sigma_2^2(1-\rho^2)$$

- Γ 分布

$Ga(\alpha,\ \frac{1}{\beta})$

$$f(x) := \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \qquad (x > 0)$$

平均 $\frac{\alpha}{\beta}$, 分散 $\frac{\alpha}{\beta^2}$, モード $\frac{\alpha-1}{\beta}$

Γ 関数について

- $\Gamma(z) := \int_0^\infty t^{z-1}e^{-t}dt$
- $\Gamma(1) = 1$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- $\Gamma(z+1) = z\Gamma(z)$
- $\Gamma(n+1) = n!$

- β 分布

  $X \sim Be(a,\ b)$ のとき
  - $f(x) := \dfrac{1}{B(a,\ b)}\, x^{a-1}(1-x)^{b-1}$

    ちなみに $B(a,\ b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
  - $E(X) = \dfrac{a}{a+b}$
  - モード $\dfrac{a-1}{a+b-2}$

- 生存関数 (2016#8)

$$S(x) := 1 - F(x) \quad (F(x) : cdf)$$
$$h(x) := \frac{f(x)}{1-F(x)} (= (-\log S(x))')$$

  $h(x)$ を**ハザード関数**という。

2 確率変数

(2018#2)

- 確率変数 X と Y が独立 $\Longleftrightarrow P(XY) = P(X)P(Y)$ (WB p2)
- 確率変数 X と Y が独立ならば (WB p16)

$$E[XY] = E[X]E[Y]$$

- $V[X \pm Y] = V[X] + V[Y] \pm 2\mathrm{Cov}[X,\ Y] = V[X] + V[Y] \pm 2\rho\sqrt{V[X]V[Y]}$

  ただし、$\rho := \dfrac{\mathrm{Cov}[X,\ Y]}{\sqrt{V[X]}\sqrt{V[Y]}}$
- $\mathrm{Cov}[X_i,\ X_j] = E[X_iX_j] - E[X_i]E[X_j]$ (WB p15)

3 母関数

- 確率母関数

$$G(s) := E[s^X] = \sum s^X p(x)$$

このとき

$$G'(1) = E[X],\ G''(1) = E[X(X-1)]$$

- モーメント母関数

$$m(\theta) := E[e^{\theta X}] = G(e^\theta)$$

このとき

$$m'(0) = E[X],\ m''(0) = E[X^2]$$

モーメント母関数は連続分布に用いられることが多い

独立な確率変数の和のモーメント母関数はモーメント母関数の積 (WB p10)
$X_1,\ X_2$: 独立な確率変数
$X_1,\ X_2$ のモーメント母関数をそれぞれ $m_1(\theta),\ m_2(\theta)$ とすると
$X_1 + X_2$ のモーメント母関数 $m(\theta)$ は

$$m(\theta) = m_1(\theta)m_2(\theta)$$

4 実験計画法

- ネイマン配分法 (2016#3)
  層の大きさ × 層内の標準偏差 の比で配分
  → 分散を最小にする
- フィッシャーの 3 原則
  - 繰り返し
  - ランダム化
  - 局所管理

5 推定

- 母比率の区間推定 (2 級 p119)
  標本比率 $\hat{p} = \frac{x}{n}$ を用いると

$$E[\hat{p}] = p,\ V[\hat{p}] = \frac{p(1-p)}{n}$$

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,\ 1)$$

  信頼率 95% の信頼区間 ($\sqrt{}$ 内の p を $\hat{p}$ で置き換える)

$$\hat{p} - 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

- 母分散の区間推定 (2 級 p116)

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-1)}$$

- 多項分布の差の信頼区間 (WB p.73)
  $P(A_i) = p_i$ $(i = 1,\ 2,\ \cdots,\ k)$ となる多項分布について

$$(\hat{p}_i - \hat{p}_j) \pm 1.96 \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n} + \frac{\hat{p}_j(1-\hat{p}_j)}{n} + \frac{2\hat{p}_i\hat{p}_j}{n}}$$

6 検定

- 母平均の差の検定 (2 級 p152)

$$t := \frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\,\hat{\sigma}} \sim t(m+n-2)$$

ここで、プールした分散

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{m+n-2}$$

- 適合度検定

$$\chi^2 = \sum \frac{(x_i - m_i)^2}{m_i} \sim \chi^2(n-1)$$

- 逸脱度

$$G^2 = 2\sum^{i}\sum^{j} x_{ij} \log \frac{x_{ij}}{m_{ij}}$$

x_{ij} : 観測値, m_{ij} : 期待度数
G^2 の自由度は χ^2 の自由度に等しい

- 無相関の検定

$$T := \frac{|r|\sqrt{n-2}}{1-r^2} \sim t(n-2)$$

r : 標本相関係数

- 回帰係数の推定値の検定 (重回帰)(WB p130)
$H_0 : \beta_k = 0$ として検定

$$t = \frac{\hat{\beta}_k}{\hat{\beta}_k \text{の } SD} \sim \mathrm{t}(n-d-1)$$

(n : データサイズ、d : 説明変数数)

- 母分散の検定 (2 級 p147)
  母分散 σ^2 が特定の値 σ_0^2 に等しいかを検定する

  帰無仮説 H_0: $\sigma^2 = \sigma_0^2$　対立仮説 H_1: $\sigma^2 \neq \sigma_0^2$

  H_0 の下で、標本 $x_i \sim N(\mu,\ \sigma_0^2)$ なので、次の検定統計量を用いればよい

$$\chi^2 = \frac{\sum (x_i - \overline{x})^2}{\sigma_0^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

- 母分散の比の検定 (2 級 p155)
  2 つの母集団からの標本が $x_i \sim N(\mu_x,\ \sigma_x^2)\ \ (i = 1,\ \cdots,\ m)$,
  $y_j \sim N(\mu_y,\ \sigma_y^2)\ \ (j = 1,\ \cdots,\ n)$ とする

  帰無仮説 H_0: $\sigma_x^2/\sigma_y^2 = 1$　対立仮説 H_1: $\sigma_x^2/\sigma_y^2 \neq 1$

  H_0 の下で、次の検定統計量を用いればよい

$$F := \frac{\hat{\sigma}_x^2}{\sigma_x^2} \cdot \frac{\sigma_y^2}{\hat{\sigma}_y^2} = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \sim F(m-1,\ n-1)$$

- 母比率に関する検定 (2 級 p148)
  - 母集団から n 人を無作為に抽出して支持率を調査した場合、支持者数 x の分布は $Bin(n,\ p)$ である
  - n が大きい場合、二項分布は正規分布で近似できる
  - $H_0 : p = p_0$ として、以下の z を検定統計量とすればよい

$$z := \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \sqrt{n}\,\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sim N(0,\ 1)$$

- 母比率の差の検定 (2 級 p156)
  - 各母集団からサイズ n_i のサンプルを取ったとき、該当者数が x_i とする
  - $x \sim Bin(n,\ p) \longrightarrow n$ が大きいとき正規分布で近似
  - $H_0 : p_1 - p_2 = 0$ として、以下の z を検定統計量とすればよい

$$z := \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,\ 1)$$

7 時系列

- 共分散定常
  - $E[Y_t] = \mu$
  - $\mathrm{Cov}[Y_t,\ Y_{t-h}] = \gamma_{|h|}$
- ホワイトノイズ
  - 共分散定常
  - $h \neq 0 \Rightarrow \gamma_{|h|} = 0$
- AR(1) 過程
  - $Y_t = c + \phi_1 Y_{t-1} + U_t\ \ (t = 1,\ \cdots,\ T)$ U_t : ホワイトノイズ
  - $|\phi_1| < 1$ ならば AR(1) は共分散定常
  - $\mu = \dfrac{c}{1 - \phi_1}$
  - $\gamma_0 = \dfrac{\sigma^2}{1 - \phi_1^2},\ \gamma_h = \phi_1^h \dfrac{\sigma^2}{1 - \phi_1^2}$
  - 次数選択には偏自己相関係数を利用
- MA(1) 過程
  - $Y_t = \mu + U_t + \theta_1 U_{t-1}$
  - MA(1) は共分散定常
- DW 比

$$DW = 2(1 - \hat{\gamma_1}) \quad (\hat{\gamma_1} : 1 \text{ 次の自己相関係数の推定量})$$

DW 検定 (WB p143)

誤差項はホワイトノイズでなければならない → 自己相関の有無を検定

- DW 比が 2 に近くなれば H_0 を受容 (自己相関なし)
- DW 比が 0 に近くなれば H_0 を棄却 (U_t に 1 次の正の自己相関あり)
- DW 比が 4 に近くなれば H_0 を棄却 (U_t に 1 次の負の自己相関あり)

8 モデル選択

- AIC と BIC
  - $AIC = -2\log L + 2k$ $\quad$ (L : 最大尤度, k : パラメータ数)
  - $BIC = -2\log L + k\log n$
  - AIC は n が大きくなっても真のモデルを選ばないことあり
  - BIC は n が $n \to \infty$ のとき、確率 1 で真のモデルを選ぶ

9 多変量解析

- 決定係数

$$R^2 := \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{S_R}{S_y}$$

$$\underbrace{\sum (y_i - \bar{y})^2}_{S_y,\ \text{全変動}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{S_R,\ \text{回帰で説明可能}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{S_e,\ \text{回帰で説明不可}}$$

- 自由度調整済み決定係数

$$1 - R^{*2} = (1 - R^2) \times \frac{n-1}{n-d-1} \qquad (n : \text{サンプル数}, d : \text{説明変数数})$$

一般に $R^{*2} < R^2$

10 特性値

- 変動係数 (2016#1) $\dfrac{\sqrt{V(X)}}{E(X)}$

11 主成分分析

- 主成分負荷量 (2018#7, WB p197) 元の変数と主成分との相関係数

$$\frac{\sqrt{\lambda_j}u_{k,j}}{\sqrt{s_{k,k}}} \qquad \begin{array}{ll} s_{ij}: & \text{分散共分散行列} \\ u_{k,j}: & \text{第 } j \text{ 固有ベクトルの第 } k \text{ 成分} \end{array}$$

もしくは

$$\sqrt{\lambda_j}u_{k,j}$$

12 一般化線形モデル

2 値応答のモデルとしてロジスティック回帰とプロビットモデルがある

- ロジスティック回帰モデル (2019#10)
  生起確率を π として

$$\log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  $\mathrm{logit}(q) := \log\dfrac{q}{1-q}$ を**ロジット関数**という

  $\dfrac{q_i}{1-q_i}$ は**オッズ**

- プロビットモデル

$$\pi = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

  $\Phi(x) : N(0,1)$ の *cdf*

  限界効果 (x_j の効果の大きさ)(WB p.149)

$$\begin{aligned}\frac{\partial\pi}{\partial x_j} &= \frac{\partial}{\partial x_j}\Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \\ &= \phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\cdot\beta_j\end{aligned}$$

- ポアソン回帰モデル (WB p.150)

$$\log\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

13 因子分析

- 因子負荷量 (WB p224)

$$V[x_{ij}] = a_j^2 + d_j^2 = 1$$

a_j : 因子負荷量, d_j : 独自因子

14 分散分析

- 一要因の平均の区間推定

$$\mu_A \pm t_{0.025}(n-1) \times \sqrt{\frac{V_E}{n}}$$

- ブロック因子 応答に影響を与えるが、その効果に興味がないもの
  ブロック因子を取り入れた分析を**乱塊法**という

15 ベイズ統計

- 点推定
  ベイズ推定　期待値
  MAP 推定　モード

- 共役事前分布
  - ベータ二項モデル
  - ガンマ・ポアソンモデル

- MAP 推定量
  事後分布が $Be(a,\ b)$ のとき MAP 推定量は $Be(a,\ b)$ のモード、つまり

$$\text{MAP} = \frac{a-1}{a+b-2}$$

16 多次元尺度法 (MDS)

- 計量 MDS(WB p232)
  k 次元空間の n 個の点 $x_1,\ x_2,\ \cdots,\ x_n$ に対して、互いの距離の 2 乗の行列 D(距離行列) があったとする。このとき

$$B = -\frac{1}{2}\left(I_n - \frac{1}{n}J_n\right) D \left(I_n - \frac{1}{n}J_n\right) = X^t X$$

  B を $U\Lambda^t U$ で対角化する。このとき

$$X = U\Lambda^{\frac{1}{2}} \quad (\text{サイズ } n \times r)$$

  によって、x_j $(j = 1,\ \cdots,\ n)$ が再現される (エッカート・ヤング分解)。

17 分割表

- オッズ (WB p261)
  標本オッズ比

$$OR = \frac{x_1(n_2 - x_2)}{x_2(n_1 - x_1)}$$

  がオッズ比 ψ の自然な推定量。標本サイズが十分に大きければ、標本オッズ比の対数 (標本対数オッズ比) の推定誤差は

$$\sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}}$$

  母オッズ比の対数は、この誤差を標準偏差として正規分布で区間推定 $(\log\widehat{OR}_1,\ \log\widehat{OR}_2)$ される。よって母オッズ比の区間推定は

$$(\widehat{OR}_1,\ \widehat{OR}_2) = (\exp(\log\widehat{OR}_1),\ \exp(\log\widehat{OR}_2))$$

  で得られる
- 逸脱度 (WB p263)

$$G^2 = 2\sum_i \sum_j x_{ij} \log\frac{x_{ij}}{m_{ij}}$$

18 漸近理論

- 不偏推定量 推定量の期待値が真の値に等しい ($E[\hat{\theta}] = \theta$)
- 一致推定量 推定量が真の値に確率収束する ($\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1$)
- 有効推定量 クラメール・ラオの不等式の等号を満たすような不偏推定量
- 漸近有効性 一致推定量の分散が漸近的にクラメール・ラオの不等式の下限に達する ($\lim_{n\to\infty} nV[\hat{\theta}] = J_1(\theta)^{-1}$)
- 漸近正規性 $Z := \sqrt{n}(\hat{\theta} - \theta)$ は正規分布 $N(0,\ J_1(\theta)^{-1})$ に分布収束する
- 正規分布の標本分散は、分散の最尤推定量 (WB p60)
- 十分統計量 統計量 $T = T(X)$ が以下を満たす

$$P(X = x|T(X) = t,\ \theta) = P(X = x|T(X) = t)$$

- **フィッシャー情報量** f を pdf もしくは pmf として

$$J_n(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f(X_1,\ \cdots,\ X_n;\ \theta)\right)^2\right]$$

あるいは、適当な条件の下で

$$J_n(\theta) = -E\left[\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1,\ \cdots,\ X_n;\ \theta)\right)\right]$$

- **フィッシャー・ネイマンの分解定理**
T が十分統計量 $\iff \exists h,\ g \quad f(x;\theta) = h(x)g(T(x),\ \theta)$
- 最尤推定量は一致性と漸近有効性あり (WB p65)
最良な推定量の 1 つ

- モーメント法

$$\mu_1 := \int xf(x;\theta)dx$$
$$\mu_k := \int (x - \mu_1)^k f(x;\theta)dx$$

に対して

$$\hat{\mu}_1 := \hat{X} = \frac{1}{n}\sum X_i$$
$$\hat{\mu}_k := \frac{1}{n}\sum (X_i - \overline{X})^k$$

で置き換えて、連立方程式を解いてパラメータ θ_j を求める方法

19　ノンパラメトリック法

- ウィルコクソンの順位和検定

  以下、群 A の数値が群 B より低いのでは、という状況
  - 群 A と群 B をあわせて、昇順で順位づけする
  - 各群の順位の和を W_A, W_B とする
  - W_A の取りうる場合の数 N_A を求める
  - W_A の観測値を w_A とするとき、$n(W_A \leq w_A)$ を求める
  - 以下の値を P 値とする

$$P(W_A \leq w_A) := \frac{n(W_A \leq w_A)}{N_A}$$

- ウィルコクソンの符号付き順位検定

  データサイズを N とする
  - データを絶対値の昇順で順位づけする
  - 観測値が負値なら順位に負号をつける
  - 順位の正値の和 T_+ を検定統計量とする

    例) 符号付き順位の集合 $\tilde{D}: 1,\ -2,\ 3,\ -4,\ 5,\ 6$

$$\Longrightarrow T_+ = 1 + 3 + 5 + 6 = 15$$

  - 符号付き順位の場合の数は 2^N 通り
  - 以下の値を P 値とする (t_+ は T_+ の観測値)

$$P(T_+ \leq t_+) := \frac{n(T_+ \leq t_+)}{2^N}$$

- クラスカル・ウォリス検定
  - 一元配置分散分析のノンパラメトリック版
  - k 群のデータに差があるかを検定する
  - 各データに全群での順位を昇順でつける
  - 第 j 群の順位和を R_j、サンプルサイズを n_j とするとき、以下の H を検定統計量とする

$$H := \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

    N が大きいとき、近似的に $\chi^2(k-1)$ に従う
    (12 や 3 は k に無関係)

  - 順位の中央値 $\tilde{N} = \frac{N+1}{2}$、各群の順位の平均 $\overline{R}_j = \frac{R_j}{n_j}$ とするとき、H は以下のようにも表される

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\overline{R}_j - \tilde{N})^2$$

20 クラスター分析

  - ウォード法

$$d(C_1,\, C_2) := \sum_{z \in C_1 \cup C_2} d(z,\, \overline{z})^2 - \sum_{x \in C_1} d(x,\, \overline{x})^2 - \sum_{y \in C_2} d(y,\, \overline{y})^2$$

    ただし、$\overline{z} := \dfrac{|C_1|}{|C_1| + |C_2|} \overline{x} + \dfrac{|C_2|}{|C_1| + |C_2|} \overline{y}$

    ある 2 つのクラスターを合併すると仮定し、合併後のクラスター内のサンプルの重心からの距離の二乗和から、合併前の 2 つのクラスター内のサンプルの重心からの距離の二乗和を引いた値が最小となるクラスター同士を合併する