統計学ノートNo.11
全体のまとめ・復習

担当教員：蛭川雅之

研究室：紫英館3階345研究室

メール： hirukawa@econ.ryukoku.ac.jp

オフィスアワー：毎週金曜日3講時

1　記述統計（1変量）

度数分布表・ヒストグラム

1. **作成方法**
2. **用語の意味・関連**
   - **階級**
   - **階級値**
   - **度数**
   - **相対度数**
   - **累積度数**
   - **累積比率**

統計量

1. **代表例**
   - **位置に関する尺度：平均値、中央値**
   - **ちらばりに関する尺度：分散、標準偏差**
2. **計算方法**
   - **生データから**
   - **度数分布表から**

2　記述統計（２変量）

散布図

1. **作成方法**
2. **正・負の相関との関連**
   - **およそ右上がり？**
   - **それとも右下がり？**

共分散・相関係数

1. **計算方法**
2. **正の相関・負の相関・無相関**
3. **相関係数**

- **ピアソン相関係数とスピアマン相関係数の違いは？**
- **取りうる値は？**

３　加工されたデータに対する平均値・標準偏差等の計算

平均値・分散等はどうなるのか？

1. **各データに定数bを加える。**
2. **各データを$a(>0)$倍する。**
3. **各データを$a(>0)$倍しさらに定数bを加える。**
4. **応用：**
   - **標準化 ＝ 平均値を引いて標準偏差で割る**
   - **２変量への拡張**

4　離散確率変数

離散確率変数とは…

1. **定義**
   - **実現値は以下のいずれか。**
     a. **有限個：**{0,1}, {1,2,3,4,5,6}, …
     b. **可算無限個：**{0,1,2,3, … }, …
2. **用語・計算方法の確認**
   - **<u>１変量</u>：確率分布、（累積）分布関数、期待値、分散、標準偏差**
   - **<u>２変量</u>：同時確率分布、周辺確率分布、共分散、相関係数、独立性**

代表的な離散確率分布

1. **離散分布の例：**
   - **ベルヌーイ分布**
   - **二項分布**
   - **ポアソン分布**
2. **期待値・分散は？**

5　連続確率変数

連続確率変数とは...

1. **ある区間（例：$[0,1], (-\infty, \infty)$）内の任意の実数をとる。**
2. **確率は区間に対して定義される。**
   - **連続確率変数が特定の一点をとる確率はゼロ！**

正規分布

1. **形状**
2. **表記法：$N(\mu, \sigma^2)$**
3. **標準正規分布**
4. **正規分布表の使い方：**
   - **直接確率を計算する。**
   - **分位点を求める。**

6　母集団と標本

標本分布

1. **部分（標本）から全体（母集団）へ！**
2. **なぜ標本平均をとるのか？**
   - **大数の法則**
3. **正規母集団からの標本平均の分布**
   - **$N(\mu, \sigma^2)$から取り出したn個のデータの標本平均$\bar{X}$に対して、$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$が成り立つ。**

7　推定

推定の種別

- **推定は以下の2種類に大別される。**

1. **点推定：母数を唯一の値で推定する。**
2. **区間推定：母数をある区間で推定する。**

点推定：望ましい性質

- **理想的には、以下の性質を同時に満たす推定量が望ましい。**
  1. **不偏性**
  2. **一致性**
  3. **有効性**
- **実際には、状況に応じていずれかの基準を優先する。**

正規母集団の母平均μに関する95%信頼区間

1. **母分散σ^2が既知の場合：**

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

正規母集団の母平均μに関する95%信頼区間（つづき）

2. **母分散σ^2が未知の場合：**

$$\bar{X} - t_{0.025}\frac{\hat{\sigma}}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{0.025}\frac{\hat{\sigma}}{\sqrt{n-1}}$$

- **ただし、**
  - $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
  - $t_{0.025}\left(= t_{0.025}(n-1)\right)$**は**$t(n-1)$**の上側確率が**2.5%**となる分位点。**

母集団比率pに関する95%信頼区間

- **十分大きいnに対し、**

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

  - **ただし、$\hat{p}$は標本比率 。**

8　仮説検定

仮説検定の手順

1. **帰無仮説H_0および対立仮説H_1を立てる。**
   - **帰無仮説＝初めに仮定したこと**
   - **対立仮説＝帰無仮説の一部または全部の否定**
2. **帰無仮説H_0が正しいと仮定して検定統計量の確率分布を求める。**

仮説検定の手順（つづき）

3. **有意水準（例：5%, 1%）を設定し、この有意水準に基づいて2の確率分布に対する臨界値および棄却域を定める。**

   - **対立仮説の立て方によって両側検定か片側検定かが決まる。**

正規母集団の母平均μに関する検定

- **正規母集団$N(\mu, \sigma^2)$に対し、帰無仮説を**

$$H_0: \mu = \mu_0$$

とする。

- **正規母集団から大きさnの標本$X_1, X_2, \ldots, X_n$を取り出す。**

正規母集団の母平均μに関する検定（つづき）

- **検定統計量は次の通りである。**

1. **母分散σ^2が既知の場合：**

$$z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0,1)$$

2. **母分散σ^2が未知の場合：**

$$t = \frac{\sqrt{n-1}(\bar{X} - \mu_0)}{\hat{\sigma}} \sim t(n-1)$$

母集団比率に関する検定

- **帰無仮説を**

$$H_0: p = p_0$$

とする。

- **大きさnの<u>ベルヌーイ分布</u>に従う標本$X_1, X_2, \ldots, X_n$を取り出す。**

母集団比率に関する検定（つづき）

- **nが十分大きいとき、$H_0: p = p_0$の下で検定統計量**

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

は近似的に標準正規分布$N(0,1)$に従う。

- **ただし、$\hat{p}$は標本比率 。**

母集団比率の差に関する検定

- **帰無仮説を**

$$H_0: p_1 = p_2 \Rightarrow H_0: p_1 - p_2 = 0$$

とする。

- **各グループの標本比率$\hat{p}_1, \hat{p}_2$および$H_0: p_1 = p_2 = p$の下での2グループ全体の標本比率**

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

を計算する。

母集団比率の差に関する検定（つづき）

- **nが十分大きいとき、$H_0: p_1 - p_2 = 0$の下で検定統計量**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

は近似的に標準正規分布$N(0,1)$に従う。

適合度検定

- **確率変数Xが属する母集団はk個の互いに背反なカテゴリー（もしくは事象）$A_1, A_2, \ldots, A_k$に分類できるものとする。**
- **XがカテゴリーA_iに属する確率をp_iとすると、帰無仮説H_0および対立仮説H_1は以下のようになる。**

$$H_0: Pr(X \in A_i) = p_i, i = 1, \ldots, k$$

$$H_1: H_0\text{は正しくない}$$

適合度検定（つづき）

- **大きさnの標本において、H_0の下でのカテゴリーA_iに属する期待度数（＝理論値）は**

$$E_i = np_i$$

である。

- **実際に観測されたカテゴリーA_iに属する度数をO_iとする。**

適合度検定（つづき）

- **十分に大きい**

$$n = \sum_{i=1}^{k} O_i$$

に対し、H_0の下で検定統計量

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

は近似的に自由度$(k-1)$のカイ2乗分布に従う。

独立性検定

- **確率変数(X,Y)が属する母集団は$k \times \ell$個のカテゴリーに分類できるものとする。**
- **帰無仮説H_0および対立仮説H_1は以下のようになる。**

$$H_0: p_{ij} = p_{i.}p_{.j}, i = 1, \dots, k, j = 1, \dots, \ell$$

$$H_1: H_0\text{は正しくない}$$

  - **ただし、**

$$p_{ij} = Pr\left(X \in A_i, Y \in B_j\right)$$

$$p_{i.} = Pr(X \in A_i), p_{.j} = Pr\left(Y \in B_j\right)$$

独立性検定（つづき）

- **実際に観測されたカテゴリー(A_i, B_j)に属する度数をO_{ij}とする。**
- **周辺確率を次の手順で推定する。**

$$\hat{p}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{\ell} O_{ij}\,, \hat{p}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{k} O_{ij}$$

  - **これらを用いて、理論値を**

$$\hat{E}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$$

に置き換える。

独立性検定（つづき）

- **十分に大きい**

$$n = \sum_{j=1}^{\ell} \sum_{i=1}^{k} O_{ij}$$

に対し、H_0の下で検定統計量

$$\hat{Q} = \sum_{j=1}^{\ell} \sum_{i=1}^{k} \frac{\left(O_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

は近似的に自由度$(k-1)(\ell-1)$のカイ2乗分布に従う。

クラメール連関係数

- **k行×ℓ列分割表に対するクラメール連関係数Vは**

$$V = \sqrt{\frac{\hat{Q}}{n\{min(k, \ell) - 1\}}}$$

で与えられる。

クラメール連関係数（つづき）

- **2×2分割表の場合、**

$$V = \sqrt{\frac{\hat{Q}}{n}}$$

となる。

- **$0 \leq V \leq 1$であり、 $V > 0.25$ が2つの要因の関連を示す大まかな目安とされる。**

9　单回帰分析

単回帰モデル

- **線形回帰モデル**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \dots, n$$

において、Xは説明変数（独立変数）、Yは被説明変数（従属変数）と呼ばれる。

  - **特に、説明変数が1つだけの線形回帰モデルを単回帰モデルという。**

最小2乗法

- **最小2乗法とは、残差2乗和**

$$SSR(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{Y_i - (b_0 + b_1 X_i)\}^2$$

を最小にする(b_0, b_1)を回帰係数(β_0, β_1)の推定値とする推定法である。

単回帰モデルの最小2乗推定量

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\left(= \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}\right)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

予測値・残差の性質

1. **予測値** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n$**の平均**

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$$

は被説明変数の平均$\bar{Y}$**に等しい。**

2. **回帰直線**$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$**は点**$(\bar{X}, \bar{Y})$**を通る。**

予測値・残差の性質（つづき）

3. **残差$e_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n$の和はゼロである。**

$$\sum_{i=1}^{n} e_i = 0$$

4. **説明変数と残差との積の和はゼロである。**

$$\sum_{i=1}^{n} X_i e_i = 0$$

分散分解

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}_{ESS} + \underbrace{\sum_{i=1}^{n} e_i^2}_{RSS}$$

$$TSS = ESS + RSS$$

重相関係数・決定係数・自由度修正済み決定係数

1. **重相関係数**

$$R = r_{Y\hat{Y}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}}$$

2. **決定係数**

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

重相関係数・決定係数・自由度修正済み決定係数（つづき）

3. **自由度修正済み決定係数**

$$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

k: 定数項を除く説明変数の数

- **これらの取りうる値の範囲、関係は？**

古典的諸仮定

1. **誤差項ϵの平均はゼロである。**
2. **説明変数Xは誤差項ϵと無相関である。**
3. **各観測値に付随する誤差項ϵは相互に無相関である。**
4. **誤差項ϵの分散は均一である。**
5. **誤差項ϵは正規分布に従う。**

最小2乗推定量の統計的特性

- **仮定1-4の下で、**

1. **不偏性**：$E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$

2. **一致性**：$\hat{\beta}_0 \xrightarrow{p} \beta_0,\ \hat{\beta}_1 \xrightarrow{p} \beta_1$

の両方が成り立つ。

ガウス＝マルコフ定理

- **仮定1-4の下で、最小2乗推定量$(\hat{\beta}_0, \hat{\beta}_1)$は、$(\beta_0, \beta_1)$の全ての線形不偏推定量の中で最小の分散を持つ。**

  - **線形推定量とは、ある定数$c_i,\ i = 1, \ldots, n$を用いて$\sum_{i=1}^{n} c_i Y_i$の形に表現できる推定量を指す。**

  - **仮定5（誤差項の正規性）が加わると、最小2乗推定量は最小分散不偏推定量となることが知られている。**

回帰式の標準誤差

- **誤差分散$\sigma^2 = Var(\epsilon)$を$s^2 = RSS/(n-2)$で推定する。**
  - **この推定量の正の平方根**

$$s = \sqrt{s^2} = \sqrt{\frac{RSS}{n-2}}$$

は、しばしば回帰式の標準誤差（Standard Error of Regression；略称SER）と呼ばれる。

標準誤差

- **σ^2をs^2に置き換えることにより、$Var(\hat{\beta}_0)$および$Var(\hat{\beta}_1)$の推定量が得られる。**

$$\widehat{Var}(\hat{\beta}_0) = s^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\}$$

$$\widehat{Var}(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

標準誤差（つづき）

- **これらの正の平方根**

$$SE(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)}$$

$$SE(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}$$

は$\hat{\beta}_0$および$\hat{\beta}_1$の標準誤差（Standard Error；略称SE）と呼ばれる。

回帰係数に関するt検定

- **帰無仮説**

$$H_0: \beta_1 = b$$

に対する検定統計量として*t*統計量

$$t_1 = \frac{\hat{\beta}_1 - b}{SE(\hat{\beta}_1)}$$

を考える。

回帰係数に関するt検定（つづき）

- **標準的な統計パッケージでは、$b = 0$に対応する*t*値**

$$t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

が自動的に計算される。

回帰係数に関するt検定（つづき）

- **帰無仮説H_0が正しい場合、$t_1 \sim t(n-2)$（＝自由度$(n-2)$のt分布）に従う。**
  - **標本数nが十分大きい場合は、漸近的に$t_1 \sim N(0,1)$として差し支えない。**
- **帰無仮説H_0と対立仮説H_1の組がどのように設定されているかにより、棄却域は異なる。**
  - **両側検定か片側検定か？**