統計学ノートNo. 9
仮説検定Ⅱ

担当教員：蛭川雅之

研究室：紫英館3階345研究室

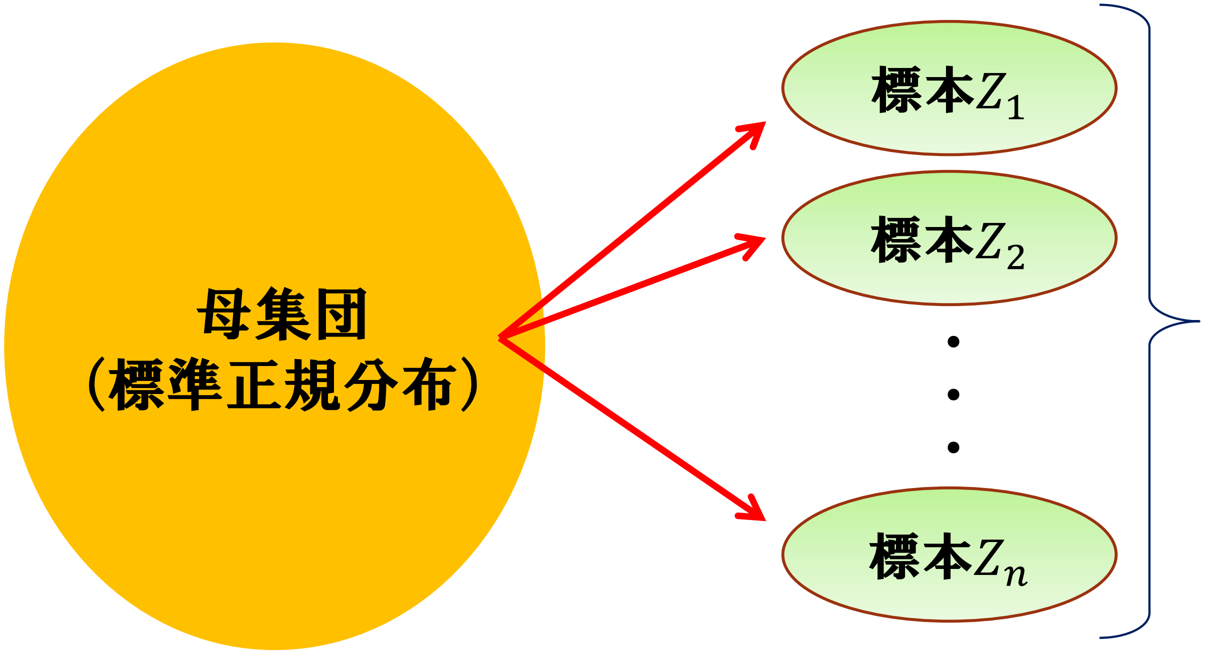
メール： hirukawa@econ.ryukoku.ac.jp

オフィスアワー：毎週金曜日3講時

１　カイ2乗分布

正規分布の2乗和の分布

- **「標準正規分布から独立に取り出したn個のデータの2乗和」という統計量を考える。**



$$\sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

正規分布の2乗和の分布（つづき）

- $\sum_{i=1}^{n} Z_i^2$ の分布は正規分布か？
  - これは負の値をとらないため、直感的にその分布が正規分布でないことはわかる。
  - もちろん t 分布でもない。

正規分布の2乗和の分布（つづき）

- では、どのような分布に従うのか？
  - $\sum_{i=1}^{n} Z_i^2$ は**自由度nのカイ2乗（χ^2）分布**に従う。
  - このことを

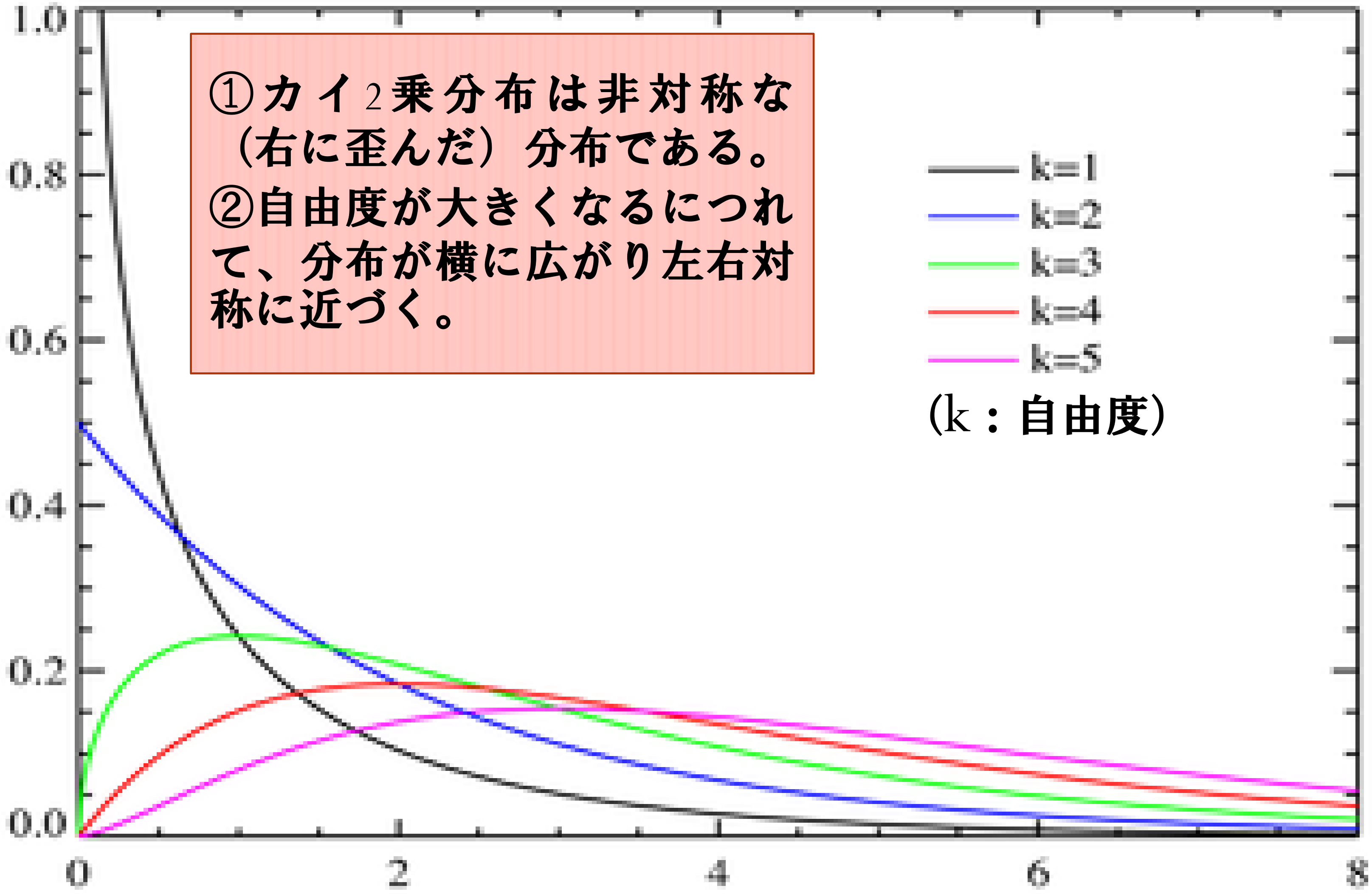
$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

と表記する。

正規分布の2乗和の分布（つづき）

- ところで**自由度**とは？
  - 大雑把に言うと（独立に値を選ぶことのできる）データ数である。
  - t分布と同様、自由度を決めるとカイ2乗分布の形状が決まる。

正規分布の2乗和の分布（つづき）



カイ2乗分布表

- **カイ2乗分布表**は所与の自由度に対するカイ2乗分布の$100q\%$分位点を与える。
  - Excelの関数chiinvを利用してもよい。
- 右裾の確率が5%または1%となる分位点がピアソンのカイ2乗検定に利用される。

カイ2乗分布：応用編

- $N(\mu, \sigma^2)$**から独立に取り出した大きさ**n**の標本**$X_1, X_2, \ldots, X_n$**を標準化したものの**2**乗和**W**は自由度**n**のカイ**2**乗分布に従う。**

$$W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

カイ2乗分布：応用編（つづき）

- **Wの母平均μを標本平均$\bar{X}$に置き換えて得られる2乗和Vは自由度$(n-1)$のカイ2乗分布に従う。**

$$V=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

２　カイ２乗検定

ピアソンのカイ2乗検定

- **ピアソンのカイ2乗検定とは「観察された事象の相対的頻度がある頻度分布に従う」という帰無仮説を検定するものである。**

1. **適合度検定：**
   - **観測された度数分布は理論分布と同じか？**
2. **独立性検定：**
   - **観測された2つの要因は互いに独立か？**

ピアソンのカイ2乗検定（つづき）

- **適合度検定、独立性検定いずれの場合も、検定統計量は以下の形をとる。**

$$Q = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

  - O_i: **カテゴリーiの頻度の観測値**
  - E_i: **帰無仮説の下でのカテゴリーiの頻度の期待値（理論値）**

ピアソンのカイ2乗検定（つづき）

- nが十分に大きいとき、帰無仮説の下で検定統計量Qは近似的に（ある自由度の）カイ2乗分布に従う。
  - 具体的な自由度の求め方は各検定法の項を参照せよ。
- Qが極めて大きい値をとる場合、帰無仮説を棄却する。

3　適合度検定

適合度検定とは…

- **与えられた度数データが特定の理論的度数分布と一致しているかどうかを検定する手法である。**

- **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）$A_1, A_2, \dots, A_k$に分類できるものとする。**
  - **例：サイコロの目**$\Rightarrow k = 6$

適合度検定とは…（つづき）

- 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。

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

$H_1: H_0$は正しくない

検定統計量の導出

- **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。**

検定統計量の導出（つづき）

- **十分大きい標本数**

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

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。

適合度検定：例題

日本人の血液型の分布はA型、O型、B型、AB型の順にそれぞれ40%、30%、20%、10%であると言われている。無作為に抽出された200人の血液型について次のようなデータが得られたとする。このデータは日本人の血液型の分布に従っているといえるかどうか有意水準5%で検定せよ。

血液型	A型	O型	B型	AB型	計
度　数	98	54	31	17	200

適合度検定：例題（つづき）

- **度数分布表を以下のように拡張すると計算の見通しが立てやすい。**
  - **表の右下が検定統計量Qの値である。**

血液型	A型	O型	B型	AB型	計
度　数(O_i)	98	54	31	17	200
H_0下の確率	0.4	0.3	0.2	0.1	1.0
理論値(E_i)	80	60	40	20	200
$(O_i - E_i)^2/E_i$	4.0500	0.6000	2.0250	0.4500	**7.1250**

適合度検定：例題（つづき）

- $\chi^2(3)$の上側5%点は$\chi^2_{0.05}(3) = 7.81$である。
- $Q = 7.125 < 7.81 = \chi^2_{0.05}(3)$であるから、有意水準5%で$H_0$は棄却されない。
  - このデータが日本人の血液型の分布に適合していないとは言い難い。

4　独立性検定

独立性検定とは…

- **適合度検定の2次元拡張版であり、2つの要因の関連性の有無を検定する。**
  - **アンケート調査の分析によく使われる。**
- **各次元がカテゴリーに分類可能であることを前提とする。**

1. **「ある俳優の好き嫌い」に「男女」差があるか？**
2. **「企業の業績の良し悪し」に「企業の分野」間で差があるか？…**

独立性検定とは…（つづき）

- 一組の確率変数(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。
  - 検定に用いられる観測値は通常分割表（＝2次元の度数分布表）の形式で与えられる。

独立性検定とは…（つづき）

- (X, Y)が特定のカテゴリー(A_i, B_j)に属する確率を

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

とする。

- 対応するXおよびYの周辺確率をそれぞれ

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

とする。

独立性検定とは…（つづき）

- **帰無仮説H_0および対立仮説H_1は以下のようになる。**

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}, i = 1, \ldots, k, j = 1, \ldots, \ell$$

$$H_1: H_0\text{は正しくない}$$

  - **母集団比率の差の検定と同様、確率の数値は具体的に与えられていない点に注意せよ。**

検定統計量の導出

- **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。**
- **仮に周辺確率がわかっているとしよう。**
  - **大きさnの標本に対し、H_0の下でのカテゴリー(A_i, B_j)に属する期待度数（＝理論値）は**

$$E_{ij} = np_{i.}p_{.j}$$

となるはずである。

検定統計量の導出（つづき）

- このとき、十分大きい標本数

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

に対し、H_0の下で検定統計量

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

は近似的に自由度$(k\ell - 1)$ のカイ2乗分布 に従う。

検定統計量の導出（つづき）

- **実際には、周辺確率は観測できない。**
  - **それらを次の手順で推定する。**

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

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

検定統計量の導出（つづき）

- **周辺確率の推定値を用いて、理論値を**

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

に置き換える。

- 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。

独立性検定：例題

無作為に選ばれた東京・大阪在住の成人男性各100人に対し、自家用車の色として赤を選ぶかどうかを調査した結果を一覧表にまとめたものである。「都市」「自家用車の色」2つの属性に関係が認められるといってよいかどうか有意水準5%で検定せよ。

都市\赤を	選　ぶ	避ける	計
東　京	40	60	100
大　阪	55	45	100
計	95	105	200

独立性検定：例題（つづき）

- **以下の手順で検定を進めるとよい。**

1. **分割表から周辺確率を推定し、次にH_0の下での同時確率を計算する。**

都市\赤を	選　ぶ	避ける	計
東　京	0.2375	0.2625	0.5000
大　阪	0.2375	0.2625	0.5000
計	0.4750	0.5250	1.0000

$100 \div 200 = 0.5000$

$95 \div 200 = 0.4750$

$0.5000 \times 0.4750 = 0.2375$

独立性検定：例題（つづき）

2. 1.**で求めた同時確率を用いて各カテゴリーの理論値を計算する。**

都市\赤を	選　ぶ	避ける	計
東　京	47.5	52.5	100
大　阪	47.5	52.5	100
計	95	105	200

$200 \times 0.2375 = 47.5$

独立性検定：例題（つづき）

3. **分割表および2.で求めた理論値を用いて、各カテゴリーに対応する$(O_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$を計算する。**

- **表の右下が検定統計量$\hat{Q}$の値である。**

都市\赤を	選　ぶ	避ける	計
東　京	1.1842	1.0714	2.2556
大　阪	1.1842	1.0714	2.2556
計	2.3684	2.1429	**4.5113**

独立性検定：例題（つづき）

- **$\chi^2(1)$の上側5%点は$\chi^2_{0.05}(1) = 3.84$である。**
- **$\chi^2_{0.05}(1) = 3.84 < 4.5113 = \hat{Q}$であるから、有意水準5%で$H_0$は棄却される。**
  - **「都市」「自家用車の色」2つの属性に関係が認められる。**

2つの要因の関連性の強さ

- 2つの要因の関連性の有無を検定するのと併せて、それらの関連性の強さを知りたい場合どうするか？

- **クラメール連関係数（Cramér's coefficient of association）**は、k行×ℓ列の分割表における行要素と列要素の関連の強さを示す代表的な指標である。

2つの要因の関連性の強さ（つづき）

- **検定統計量$\hat{Q}$について、**

$$0 \leq \hat{Q} \leq n\{min(k, \ell) - 1\}$$

が成り立つことが知られている。

- **この関係から、クラメール連関係数**

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

が得られる。

2つの要因の関連性の強さ（つづき）

- 2×2分割表の場合、

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

　となる。

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

クラメール連関係数：計算例

- 独立性検定の例題に関し、「都市」「自家用車の色」2つの属性についてクラメール連関係数Vを計算する。

- この例は2×2分割表で与えられているため、

$$V = \sqrt{\frac{\hat{Q}}{n}} = \sqrt{\frac{4.5113}{200}} \approx 0.1502$$

となる。