経済統計分析 9
分散分析

今日のおはなし.

- 検定 statistical test のいろいろ
  - 2変数の関係を調べる手段のひとつ
  - 適合度検定
  - 独立性検定
  - 分散分析

- 今日のタネ
  - 吉田耕作. 2006. 直感的統計学. 日経BP.
  - 中村隆英ほか. 1984. 統計入門. 東大出版会.

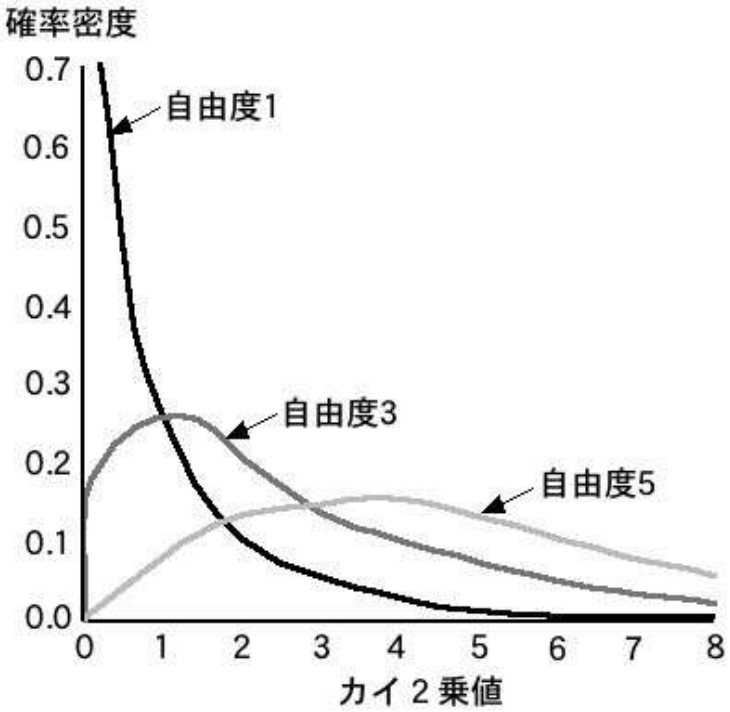
仮説検定の手続き

- 仮説検定のロジック
  - もし帰無仮説が正しければ, 検定統計量が既知の分布に従う
  - 計算された検定統計量の値から, 実現する確率（p値）が求まる

- 手続き
  1. 仮説を立てる.
  2. 有意水準を決める.
  3. 検定統計量（test statistics）を計算する.
  4. p値を求めて, 棄却/受容を判定する.

X^2分布 Chi-squared

- 自由度mのカイ2乗分布
  - m個の独立した標準正規分布に従う確率変数の2乗和の分布
  - 互いに独立な標準正規分布に従う確率変数をZ_1, Z_2, Z_3とおくと.
  - $Z_1^2 + Z_2^2 + Z_3^2$ は自由度3のカイ2乗分布に従う



http://www.ias.tokushima-u.ac.jp/linguistik/2005/joho/joho2005_chi.html

F分布

- 自由度m_1, m_2のF分布
  - 自由度m_1のカイ2乗分布に従う確率変数をm_1で割ったものと，自由度m_2のカイ2乗分布に従う確率変数をm_2で割ったものの比は自由度m_1, m_2のF分布に従う
  - いま，確率変数U_1が自由度m_1のカイ2乗分布に従い，確率変数U_2が自由度m_2のカイ2乗分布に従うとすると，

$$\frac{U_1 / m_1}{U_2 / m_2}$$ は自由度m_1, m_2のF分布に従う

- カイ2乗分布，F分布の出番
  - 2乗して和をとっている　→分散に関係しそう
  - 分散の比を調べたりしそう

適合度検定（例）

- サザエさん症候群（Blue Monday）の検定
  - 吉田耕作『直感的統計学』p.285-286
  - 曜日ごとの不良率を，各曜日に100個ずつ取り出して調べてみた

曜日	月曜	火曜	水曜	木曜	金曜	合計
不良数	10	3	0	0	2	15

  - 不良率が曜日によって異なるかどうかを有意水準5%で検定しよう．
- 検定のイメージ
  - 不良率が曜日によって同じ（帰無仮説）なら，同じ回数だけ起こるはず
  - しかし，サンプル誤差はありうるから，少しはずれるかもしれない
  - 不良率が曜日によらないなら，毎日不良品が3個（=15/5）あるはず
  - それぞれの曜日の「ずれ」の和の大きさで判断しよう
  - 「ずれ」をそのまま足すと，正と負が相殺してしまう　→2乗和をとる．

適合度検定(例)

- 実際の手続き

曜日	月曜	火曜	水曜	木曜	金曜	合計
不良数	10	3	0	0	2	15
理論値	3	3	3	3	3	
「誤差」^{2}	7^2	0^2	3^2	3^2	1^2	
揃え	$7^2/3$	$0^2/3$	$3^2/3$	$3^2/3$	$1^2/3$	22.66

- 理論値と実現値の差を理論値で割ったものを2乗して足す
- 「ずれ」の総和とみなすことができる
- もし帰無仮説が正しければ,この「ずれ和」は自由度4のカイ2乗分布に従うことが分かっている
  - カイ2乗分布は2乗和で定義されていたことを思い出そう.
- 自由度4のカイ2乗分布の上側5%点は9.488　→帰無仮説を棄却
- 「曜日によって不良率が異なる」という仮説を棄却

適合度検定

- 目的
  - 度数データが与えられているとき, 理論的度数分布と一致するかどうかを検定する
- 状況
  - 母集団がk個のカテゴリに分類できる
  - n個からなるサンプルのうち, カテゴリiに属する個数を X_i と書く
  - カテゴリiに属する理論的な確率を p_i と書く
  - つまり, カテゴリiの理論的度数は np_i となる
- 検定統計量

$$Q = \sum_{i=1}^{k} \frac{\left(X_i - np_i \right)^2}{np_i} \sim \chi^2(k-1)$$

適合度検定(練習問題)

- 不良品個数が次のようであったら, 曜日効果は認められるか

曜日	月曜	火曜	水曜	木曜	金曜	合計
不良数	8	4	2	2	4	

- 検定統計量は6となり, 帰無仮説を棄却しない.

独立性の検定（例）

- 教授はエライか検定
  - 吉田耕作『直感的統計学』p.302-303
  - 教授の階級と査読付き論文数の同時度数分布（人）を作ってみた

本数	講師	助教授	准教授	正教授	合計
0	8	18	16	6	48
1～2	0	2	2	2	6
3～4	0	0	3	0	3
5以上	0	0	1	2	3
合計	8	20	22	10	60

  - 論文数と教授の階級が関係ないかどうか検定しよう
- 検定のイメージ
  - 論文数が階級によって同じ（帰無仮説）なら，分布が同じになるはず
  - 適合度検定と似たような発想で．

独立性の検定（例）

- 実際の手続き
  - 階級に関わらず，論文数の分布が周辺分布に等しいと仮定すると
  - 理論的な度数分布は

本数	講師	助教授	准教授	正教授	合計
0	6.40	16.00	17.60	8.00	48
1～2	0.80	2.00	2.20	1.00	6
3～4	0.40	1.00	1.10	0.50	3
5以上	0.40	1.00	1.10	0.50	3
合計	8	20	22	10	60

  - 適合度検定と同じく，仮説的な度数分布との差の2乗を理論値で除したものの2乗和をとったものが検定統計量　= 13.204
  - 自由度9のカイ2乗分布に従うから，有意水準1%で帰無仮説を受容

独立性の検定

- 目的
  - 2次元の度数データが与えられているとき，理論的度数分布と一致するかどうかを検定する
- 状況
  - 母集団が$k \times m$個のカテゴリに分類できる（「分割表」と呼ぶ）
  - n個からなるサンプルのうち，カテゴリ(i, j) に属する個数を $X_{i,j}$ と書く
  - カテゴリ (i, j) に属する理論的な確率を $p_i p_j$ と書く
    - 分布が独立であれば，同時確率は周辺確率の積となる
    - 周辺確率は周辺度数から求める
  - つまり，カテゴリ (i, j) の理論的度数は $n\, p_i\, p_j$ となる
- 検定統計量

$$Q = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{\left(X_{i,j} - np_i p_j \right)^2}{np_i p_j} \sim \chi^2 (k-1)(m-1)$$

独立性の検定(練習問題)

- 管理職のレベルと高血圧の関係が以下のようであるとき, 職階と高血圧は独立に分布しているといえるか
  - 自由度2のカイ2乗分布の上側5%点は5.991.
  - 吉田耕作『直感的統計学』p.300

	重役級	部長級	課長級	合計
高血圧	80	140	80	300
正常	40	160	400	600
合計	120	300	480	900

  - 検定統計量は144で, 帰無仮説を棄却.

分散分析(例)

- 貯蓄率は職業によって異なるか?
  - 中村ほか『統計入門』pp.224-226
  - 貯蓄率を職業別に尋ねてみた

職業						
A	21	21	15	13		
B	16	20	20	18	23	23
C	15	18	16	16	15	

  - 貯蓄率が職業によって異なるかどうかを検定してみよう
  - [注意] 今回はカテゴリではなくて連続変数を扱っていますよ.
- 検定のイメージ
  - 貯蓄率が平均的に等しければ(帰無仮説), 職業別の平均からの分散と, 全体の平均からの分散は等しくなるはず
  - 平均からの乖離が正規分布に従うなら, F分布が利用できる
    - F分布は分散の比で定義されたことを思い出そう.

分散分析（例）

- 職業ごとの平均値を出してみると

職業	平均						
A	17.5	21	21	15	13		
B	20.0	16	20	20	18	23	23
C	16.0	15	18	16	16	15	

- 職業ごとに平均値が異なるとすると，偶然変動の2乗和は95.

職業	平均						
A	17.5	3.5	3.5	-2.5	-4.5		
B	20.0	-4.0	0.0	0.0	-2.0	3.0	3.0
C	16.0	-1.0	2.0	0.0	0.0	-1.0	

- 全体の平均は18なので，全体的な変動の2乗和は，140
- 職業ごとの変動の2乗和は $4(-0.5)^2 + 6(2.0)^2 + 5(-2.0)^2 = 45$
- 全変動（140）＝職業変動（45）＋偶然変動（95）
- $F = (45/2)/(95/12) = 2.84$

1元配置分散分析 ANOVA: Analysis of Variance

- 目的
  - サンプルがいくつかのカテゴリに分類されるとき, カテゴリごとの平均値が全て等しいかどうかを検定する
- 状況
  - カテゴリ i には観測値が n_i 個だけあり, カテゴリは m 個ある. 総数はn
  - カテゴリ i の j 番目の観測値の値は x_{ij} と書く
  - 標本平均を上付き線で表す
- 変動の分解:誤差の2乗和
  - 全変動:全体の平均との偏差2乗和

$$\text{全変動} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{X}\right)^2$$

  - 級間変動

$$\text{級間変動} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\bar{X}_i - \bar{X}\right)^2 = \sum_{i=1}^{m} n_i\left(\bar{X}_i - \bar{X}\right)^2$$

1元配置分散分析

- 変動の分解
  - 級内変動

$$\text{級内変動} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{X}_i\right)^2$$

  - このとき, 全変動 = 級内変動 + 級間変動
- 帰無仮説
  - 全ての平均が等しい　→級間の分散 = 級内の分散
- 検定統計量
  - 各観測値が独立に正規分布に従うと仮定するとき,

$$F\text{比} = \frac{\text{級間変動}/(m-1)}{\text{級内変動}/(n-m)} \sim F(m-1, n-m)$$

分散分析表

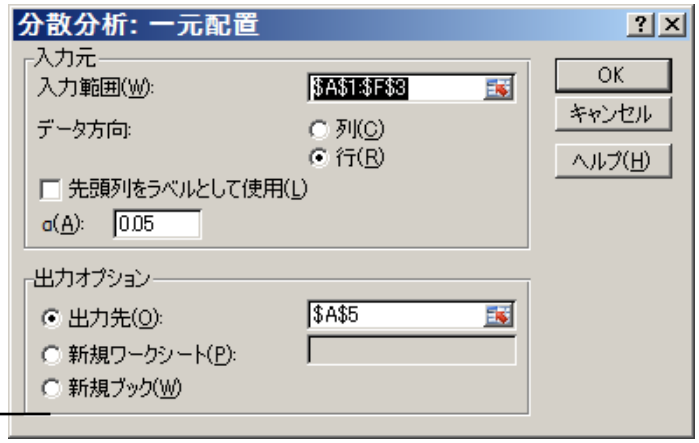
- 分散分析表

	平方和	自由度	分散	F比
級間	$S_A = \sum_{i=1}^{m} n_i \left(\bar{X}_i - \bar{X} \right)^2$	m - 1	$V_A = \frac{S_A}{m-1}$	V_A / V_E
級内	$S_E = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{X}_i \right)^2$	n - m	$V_E = \frac{S_E}{n-m}$	
全体	$S = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{X} \right)^2$			

- MS-Excelで分散分析を行うと, このような出力が得られる.
- 自分で変動を計算して, F検定してもよいんですよ(fdist関数, finv関数).
- やってみよう(練習問題).

MS-Excelで分散分析

- MS-Excel 2007でやってみた
  - データ→データ分析→分散分析：一元配置
  - 出力（桁だけそろえた）



分散分析: 一元配置

概要

グループ	標本数	合計	平均	分散
行 1	4	70	17.5	17
行 2	6	120	20	7.6
行 3	5	80	16	1.5

分散分析表

変動要因	変動	自由度	分散	観測された分散比	P-値	F 境界値
グループ間	45	2	22.500	2.842	0.098	3.885
グループ内	95	12	7.917			
合計	140	14				

2元配置分散分析

- 1元配置分散分析ではカテゴリが1種類
- 2元配置分散分析ではカテゴリが2種類
  - 2つのカテゴリで定義されるcellごとに級内変動を計算

    検証するモデルを

    $$X_{ij} = \mu + \mu_{Ai} + \mu_{Bi} + e_{ij}$$

    とすると，偶然誤差は

    $$x_{ij} - \hat{X}_{ij} = x_{ij} - \bar{X}_{i,\cdot} - \bar{X}_{\cdot,j} + \bar{X}$$

  - このばあいでも，総変動は，それぞれのカテゴリについての級間変動と，上で定義した偶然誤差（級内変動）の和に分解される

    - でも，計量経済学では，分散分析はあんまり用いられない気がする
    - ダミー変数で回帰すればいいような……？