データ解析実習

2012.10.22　月曜、４・５限

担当教員：高木英至

1

■本日の授業

- 基礎概念導入
- 前半(４限)：基本的概念
  - 統計的推測：推定と検定
  - 仮説検定
- 後半(５限)
  - 平均値の差の検定
  - 一元配置の分散分析

2

■１．統計的推測

- （１）標本抽出（sampling）
  - 母集団（population）：情報を得るべき対象の全体
  - 標本（sample）：研究のために選択された母集団の一部
- 代表性のある標本（representative sample）＝母集団を正確に反映する標本
  - 無作為性（randomness）＝代表性のある標本を得る唯一の方法（無作為抽出、random sampling）
  - →　確率論の応用
  - →　母集団に対する統計的推測が可能になる。

3

（２）統計的推測(statistical inference)

- 1.分布型の推測　－　適合度検定
- 2.母数の推測
  - a.推定
    - 点推定(point estimation)
    - 区間推定(interval estimation)
  - b.検定

4

（3）　信頼区間と標本誤差

- **［例］比率（e.g.,内閣支持率）の信頼区間**

P_0 : **標本の比率**

N **：標本サイズ（サンプル数）**

$Z_{\alpha/2}$: **正規偏差**

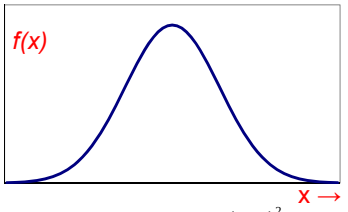
ε **：比率の標本誤差**

$$\varepsilon = Z_{\alpha/2}\sqrt{\frac{P_0(1-P_0)}{N}}$$

5

推定の考え方－まず、

- 正規分布(normal distribution)　テキストp.103



- 連続分布
- ←→離散分布
- 対称分布
- 単峰形

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ: 期待値　　σ: 標準偏差

6

中心極限定理(テキスト　p.110)

- **期待値μ、分散σ²の母集団からＮ個の観測値を標本として抽出したとき、Ｎが大きくなるほど、標本平均値（ $\overline{X}$ ）の分布は次の正規分布に近づく**

$$N\left(\mu, \frac{\sigma^2}{N}\right)$$

7

- **確率95% で**　z **は** [-1.96, 1.96]

$$p \sim N(\mu_p, \sigma^2/N) = N(\mu_p, {\sigma_p}^2)$$

- **とすると、*z*に対応する*p*は**　$\left[\mu_p - z\sqrt{\frac{\sigma^2}{N}}, \mu_p + z\sqrt{\frac{\sigma^2}{N}}\right]$
- σ_p**には**　$\sqrt{\frac{p_o(1-p_o)}{N}}$　**をあてる。**(p.132)
- μ_p**には**p_0**を推定値としてあてる。**
- p_0=0.5, N=1600 **とすれば、**
- **確率95% で***p* **は** [47.55%, 52.45%]
  - ε=±2.45%

8

より一般的に

- 有意水準αのとき、（例：α=0.05）
- 平均値x'の信頼度(1-α)×100%（例：95%）の信頼区間は、正規分布を用いて、

$$\left[\bar{x} - Z_{\alpha/2}\sigma_{\bar{x}}, \bar{x} + Z_{\alpha/2}\sigma_{\bar{x}}\right]$$

- ただし上式は常には使えない
- $\sigma_{\bar{x}}$ が分らないことが多い
- → t 分布を使う

■2．仮説検定

- 1.仮説を立てる。
  - 帰無仮説（H_0,null hypothesis）：検定されるべき仮説、検定仮説
  - 対立仮説（H_1,alternative hypothesis）：帰無仮説の対立事象を主張する仮説
- 2.測定
  - 測定結果がその仮説の下では希にしか起こらないものであるとき
    - → 仮説を否定する（帰無仮説を棄却し対立仮説を採用する）
  - 測定結果がその仮説の下でもある程度の（小さくない）確率で起こり得るとき
    - → 仮説（帰無仮説）は否定できない、と判断する。
  - 帰無仮説を否定できるか否かは確率問題

		棄却	棄却しない
母集団で帰無仮説は	真	第一種の過誤（α過誤）	正しい判断
	偽	正しい判断	第二種の過誤（β過誤）

有意水準（significance level）：第一種の過誤が生じる確率（危険率）
＝ 帰無仮説を棄却する（対立仮説を採用する）ことが誤りである確率
慣例的に、0.05 （5%）以下の有意水準を設定する。（5%, 2.5%, 1%, 0.5%, ･･･）

学年と性別は関連する。
男は1年が多く、女は2年以上が多い傾向がある。

性別 と 学年 のｸﾛｽ表

			学年		
			1年	2年以上	合計
性別	男	度数	122	16	138
		期待度数	115.2	22.8	138.0
	女	度数	100	28	128
		期待度数	106.8	21.2	128.0
合計		度数	222	44	266
		期待度数	222.0	44.0	266.0

ｶｲ2乗検定

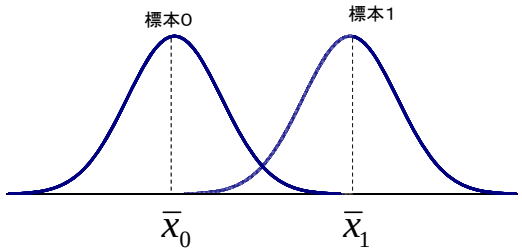
	値	自由度	漸近有意確率（両側）	正確有意確率（両側）	正確有意確率（片側）
Pearson のｶｲ2乗	5.084[b]	1	.024		
連続修正[a]	4.367	1	.037		
尤度比	5.122	1	.024		
Fisher の直接法				.031	.018
線型と線型による連関	5.065	1	.024		
有効なｹｰｽの数	266				

a. 2x2 表に対してのみ計算
b. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 21.17 です。

5%水準で有意

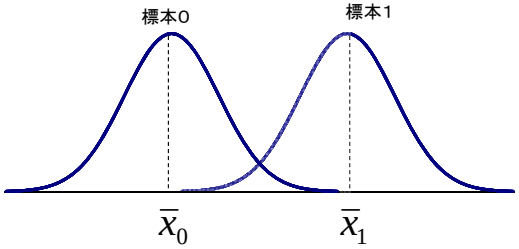
4限は
おしまい

集団間の平均値の差の検定



- 帰無仮説(H_0)：母集団では平均値に差がない
- 考え方１：標本の平均値の分布（正規分布、ｔ分布）を利用する考え方　→　ｔ検定

集団間の平均値の差の検定



- 考え方２：集団による値のバラツキが有意であるかどうか、を検定する
  - 帰無仮説(H_0)：母集団では平均値に差がない
  - 分散分析（Ｆ検定）

平方和の分割

全体の平方和は

$$SS_{total} = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 \qquad (9\cdot1) \qquad p.273$$

群間の平方和は

$$SS_A = \sum_{j=1}^{a} n_j(\bar{y}_j - \bar{y})^2 \qquad (9\cdot2)$$

- 前提
  - 集団数 ＝ a
  - 各集団の標本数：n_j

群内の平方和(残差の平方和)は

$$SS_e = \sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}2)^2 + \ldots$$

$$= \sum_{j=1}^{a} n_j {s_j}^2 \qquad (9\cdot3)$$

$$SS_{total} = SS_A + SS_e$$

平方和の分割

- 全体の平方和に占める群間の平方和の割合が大きいほど、全体の値のバラツキは集団差によって説明できる
- 偶然からでも、標本ではある程度の群間の平方和は生じる
- 群間の平方和が有意に（偶然から生じ得る以上に）大きければ、集団間の平均値の差は有意だと考える

$$\eta = \sqrt{\frac{SS_A}{SS_{total}}} \quad : \quad \text{相関比}$$

17

分散分析表

平方和／自由度

グループ間平均平方／グループ内平均平方

要因	平方和	自由度	平均平方	F値	有意確率
グループ間	SS_A	$a-1$	$SS_A/(a-1)$	$[SS_A/(a-1)]/[SS_e/(n-a)]$	p
グループ内	SS_e	$n-a$	$SS_e/(n-a)$		
全体	SS_{total}	$n-1$	$SS_{total}/(n-1)$		

- 自由度(degrees of freedom)
  - 平均値が固定されたとき、自由に動ける値の数
  - 自由度が大きければ自動的に平方和は大きくなる
- F値の検定： F検定、F分布表
  - 帰無仮説が正しければF値は小さい

18

分散分析表

分散分析

平方和／自由度

グループ間平均平方／グループ内平均平方

開放性

	平方和	自由度	平均平方	F 値	有意確率
ｸﾞﾙｰﾌﾟ間	1108.626	4	277.156	2.601	.037
ｸﾞﾙｰﾌﾟ内	20887.593	196	106.569		
合計	21996.219	200			

- 自由度(degrees of freedom)
  - 平均値が固定されたとき、自由に動ける値の数
  - 自由度が大きければ自動的に平方和は大きくなる
- F値の検定： F検定、F分布表
  - 帰無仮説が正しければF値は小さい

19

多重比較検定

- 多数の平均値があるときの、相互の差の検定
  - t検定は使えない（p.279-）
- → 多重比較検定を使う
  - S-N-K検定、テューキー検定、ダンカン検定、…

20

開放性

Student-Newman-Keuls

学部	度数	α = .05 のサブグループ	
		1	2
教養学部	50	46.5000	
工学部	52	49.6731	49.6731
教育学部	30	50.2000	50.2000
理学部	33	50.7273	50.7273
経済学部	36		53.6389
有意確率		.281	.337

等質なサブグループのグループ平均値が表示されています。

a 調和平均サンプルサイズ = 38.272 を使用

b グループ サイズが等しくありません。グループ サイズの調和平均が使用されます。タイプ I 誤差水
ん。

21

今日は おしまい

22