经济统计 – 10

刁莉男
diaoln@jlu.edu.cn
吉林大学商学院

June 6, 2012

复习

- 1、ANOVA检验；
- 2、ANOVA表；
- 3、配对处理均值的统计推断（构造总体均值差的置信区间）。

提纲

12、方差分析-双因素方差分析

无交互作用的双因素方差分析。

有交互作用的双因素方差分析。

13、相关系数与回归

正态相关分析

等级相关分析

1、双因素方差分析

双因素方差分析

- 自变量(Independent Variable);
- 因变量(Dependent Variable);
- 当方差分析中涉及两个分类型自变量时，为双因素方差分析。
- 两自变量对因变量的影响独立，为无交互作用的双因素方差分析；
- 两自变量搭配对因变量产生新影响效应，为有交互作用的双因素方差分析。

双因素方差分析：例子

有4种品牌的电视在5个地区销售，销售量数据如下，现请分析电视品牌（自变量1）和销售地区（自变量2）对销售量（因变量）是否有影响（$\alpha = 0.05$）？

	地区1	地区2	地区3	地区4	地区5
品牌1	365	350	343	340	323
品牌2	345	368	363	330	333
品牌3	358	323	353	343	308
品牌4	288	280	298	260	298

2、双因素方差分析：无交互作用；

无交互作用双因素方差分析

- 行因素(k)、列因素(r), $k * r$个观测值。
- 总平均值$\bar{X}_G = \frac{1}{kr}\Sigma_1^k\Sigma_1^r X_{ij}$.
- 行平均值：行因素第i个水平下观测值的平均值，$\bar{X}_{i\cdot} = \frac{1}{r}\Sigma_{j=1}^r X_{ij}$.
- 列平均值：列因素第j个水平下观测值的平均值，$\bar{X}_{\cdot j} = \frac{1}{k}\Sigma_{i=1}^k X_{ij}$.

例子

	地区					
	1	2	3	4	5	行均值
品牌1	365	350	343	340	323	344.2
品牌2	345	368	363	330	333	347.8
品牌3	358	323	353	343	308	337.0
品牌4	288	280	298	260	298	284.8
列均值	339	330.25	339.25	318.25	315.5	328.45

步骤1：提出假设

- 对行因素提出假设：

  $H_0: \mu_1 = \mu_2 = ... = \mu_i = ... = \mu_k$；

  $H_1: \mu_1, \mu_2, ..., \mu_k$不全相等；

- 对列因素提出假设：

  $H_0: \mu_1 = \mu_2 = ... = \mu_j = ... = \mu_r$；

  $H_1: \mu_1, \mu_2, ..., \mu_r$不全相等；

步骤2：构造统计量

分解总误差平方和。

- 总误差平方和：

  $SSTotal = \Sigma_{i=1}^{k}\Sigma_{j=1}^{r}(X_{ij} - \bar{X}_G)^2,$
- 行因素误差平方和：

  $SSR = \Sigma_{i=1}^{k}\Sigma_{j=1}^{r}(\bar{X}_{i\cdot} - \bar{X}_G)^2,$
- 列因素误差平方和：

  $SSC = \Sigma_{i=1}^{k}\Sigma_{j=1}^{r}(\bar{X}_{\cdot j} - \bar{X}_G)^2,$
- 随机误差项平方和：

  $SSE = \Sigma_{i=1}^{k}\Sigma_{j=1}^{r}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_G)^2.$

步骤2：构造统计量

分解总误差平方和。

$SSTotal = SSR + SSC + SSE$

自由度：

$df_{Total} = df_R + df_C + df_E$

- $df_{Total} = kr - 1$,
- $df_R = k - 1$,
- $df_C = r - 1$,
- $df_E = (k-1)(r-1)$.

步骤2：构造统计量

计算各平方和均方。

- 行因素均方MSR：$MSR = \frac{SSR}{k-1}$,
- 列因素均方MSC：$MSC = \frac{SSC}{r-1}$,
- 随机误差均方MSE：$MSE = \frac{SSE}{(k-1)(r-1)}$.

步骤2：构造统计量

- 行因素对因变量影响显著性检验统计量：
  $$F_R = \frac{MSR}{MSE} \sim F(k-1,(k-1)(r-1)),$$
- 列因素对因变量影响显著性检验统计量：
  $$F_C = \frac{MSC}{MSE} \sim F(r-1,(k-1)(r-1)),$$

步骤3：决策规则

拒绝域：

- 行因素：
  如果$F_R > F_\alpha$，则拒绝原假设，表明行因素对因变量影响显著；
- 列因素：
  如果$F_C > F_\alpha$，则拒绝原假设，表明列因素对因变量影响显著。

ANOVA表

	SS	df	MS	F 值	P 值	Fcrit
行	SSR	k-1	MSR	FR		
列	SSC	r-1	MSC	FC		
误差	SSE	(k-1)(r-1)	MSE			
总和	SSTotal	kr-1				

例子

- 行因素（品牌）：

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$，品牌对销售无影响；

  $H_1 : \mu_1, \mu_2, \mu_3, \mu_4$不完全相等，品牌对销售有影响。

- 列因素（地区）：

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$；

  $H_1 : \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$不完全相等。

例：ANOVA表

	SS	df	MS	F 值	P 值	Fcrit
行	13004.55	3	4334.8	18.11	0.0000	3.49
列	2011.7	4	502.9	2.10	0.143	3.26
误差	2872.7	12	239.4			
总和	17888.95	19				

例：结果分析

- $F_R = 18.11 > F_\alpha = 3.49$，或$p = 0.000 < \alpha$，拒绝对行因素的原假设，$\mu_1, \mu_2, \mu_3, \mu_4$之间差异显著；
- $F_C = 2.10 < F_\alpha = 3.26$，或$p = 0.143 > \alpha$，无法拒绝对列因素的原假设，$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$之间无显著差异。

3、双因素方差分析：有交互作用；

例

交管部门对不同路段不同时间对行车时间的影响，经多次抽样得到数据如下，现分析路段、时段以及路段和时段交互作用对行车时间的影响（$\alpha = 0.05$）。

例

	路段1	路段2
时段1	26	19
	24	20
	27	23
	25	22
	25	21
时段2	20	18
	17	17
	22	13
	21	16
	17	12

有交互作用双因素方差分析

- 总观测值，$n(=20)$，每个观测值X_{ijl},
- 行因素$k(=2)$；行因素第i个水平样本均值：$\bar{X}_{i\cdot\cdot}$,
- 列因素$r(=2)$；列因素第j个水平样本均值：$\bar{X}_{\cdot j\cdot}$,
- 行因素中，每个水平行数为$m(=5)$；行因素第i个水平，列因素第j个水平组合的样本均值$\bar{X}_{ij}$.

有交互作用双因素方差分析：方差

- SSTotal=SSR+SSC+SSRC+SSE
- $SSTotal = \Sigma_{i=1}^{k}\Sigma_{j=1}^{r}\Sigma_{l=1}^{m}(X_{ijl} - \bar{X}_G)^2$,
- $SSR = rm\Sigma_{i=1}^{k}(\bar{X}_{i\cdot} - \bar{X}_G)^2$,
- $SSC = km\Sigma_{j=1}^{r}(\bar{X}_{\cdot j} - \bar{X}_G)^2$,
- $SSRC = m\Sigma_{i=1}^{k}\Sigma_{j=1}^{r}(\bar{X}_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_G)^2$.

有交互作用双因素方差分析：自由度

- $df_{Total} = df_R + df_C + df_{RC} + df_E$
- $df_{Total} = rkm - 1,$
- $df_R = k - 1,$
- $df_C = r - 1,$
- $df_{RC} = (r-1)(k-1),$
- $df_e = rk(m-1).$

ANOVA表

	SS	df	MS	F 值	P 值	Fcrit
行	174.05	1	174.05	44.1	0.000	4.49
列	92.45	1	92.45	23.45	0.0001	4.49
交互	0.05	1	0.05	0.012	0.91	4.49
误差	63.2	16	3.95			
总和	329.75	19				

13章 相关系数与回归

1、变量间的关系

变量间的三种关系

- 函数关系；
- 独立关系；
- <u>相依关系</u>。

相依关系的分析方法

- 相关分析：测量两个变量之间的关系。
  - 正态相关分析；
  - 等级相关分析；
- 回归分析。

2、相关系数：正态相关分析；

正态相关分析

- 前提条件：有关变量的联合分布是正态分布。
- 包括简单相关分析（两个变量）、偏相关分析和复相关分析（多个变量）。
- 用相关系数来度量相关关系；
- 用样本相关系数(r)来估计和检验总体相关系数(ρ)。

Pearson相关系数：

- $r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)s_X s_Y}$

例：研究销售人员销售量和打电话次数是否存在相关关系，数据如下：

Name	Calls	Sales	Name	Calls	Sales
Tom	20	30	Calos	10	40
Jeff	40	60	Rich	20	40
Brian	20	40	Mike	20	50
Greg	30	60	Mark	20	30
Susan	10	30	Soni	30	70

检验相关系数显著性

- $H_0 : \rho = 0, H_1 : \rho \neq 0$
- 统计量： $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- df=n-2;

检验相关系数显著性

- $H_0 : \rho = 0, H_1 : \rho \neq 0$
- 统计量：
  $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.759\sqrt{10-2}}{\sqrt{1-0.759^2}} = 3.297$$
- df=n-2=8;
- $t_{0.05}(8) = 2.306;$
- 拒绝域： $t > t_{0.05}(8)$ 或者 $t < -t_{0.05}(8)$
- p-value=0.011,
- 拒绝原假设。

3、相关系数：等级相关分析；

Spearman等级相关系数

- 使用前提：不要求知道变量的分布，更不要求服从正态分布；
- 等级（Rank，秩）：将观测值$X_1, X_2, ..., X_n$按递增顺序排序，X_i在排列中的顺序号称作X_i的秩。

Spearman等级相关系数

- Spearman等级相关系数：
  n个对象两个标志(X, Y)，r_i表示X_i的等级，而s_i表示Y_i的等级，$d_i = r_i - s_i$，
  $r_s = \frac{\Sigma(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\Sigma(r_i - \bar{r})^2 \Sigma(s_i - \bar{s})^2}} = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$。
- Spearman相关系数显著性检验
  与Pearson相关系数检验相同。
- 上例中，
  $corr_{Spearman} = 0.7398$, p-value=0.0144.

例：

Nr.	Calls	Sales	R_C	R_S	$R_C - \bar{R}_C$	$R_S - \bar{R}_S$
1	20	30	5	2	-0.5	-3.5
2	40	60	10	8.5	4.5	3
3	20	40	5	5	-0.5	-0.5
4	30	60	8.5	8.5	3	3
5	10	30	1.5	2	-4	-3.5
6	10	40	1.5	5	-4	-0.5
7	20	40	5	5	-0.5	-0.5
8	20	50	5	7	-0.5	1.5
9	20	30	5	2	-0.5	-3.5
10	30	70	8.5	10	3	4.5

检验相关系数显著性

- $H_0 : \rho = 0, H_1 : \rho \neq 0$
- 统计量：
  $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7398\sqrt{10-2}}{\sqrt{1-0.7398^2}} = 3.11$$
- df=n-2=8;
- $t_{0.05}(8) = 2.306$;
- 拒绝域： $t > t_{0.05}(8)$ 或者 $t < -t_{0.05}(8)$
- p-value=0.0144
- 拒绝原假设。

本次重点

- 1、无交互作用双因素方差分析；
- 2、有交互作用双因素方差分析；
- 3、正态相关分析；
- 4、等级相关分析。