

心理统计学第八讲

讲师 司马紫衣



高途学院APP下载



高途学院公众号





第十一章 卡方检验

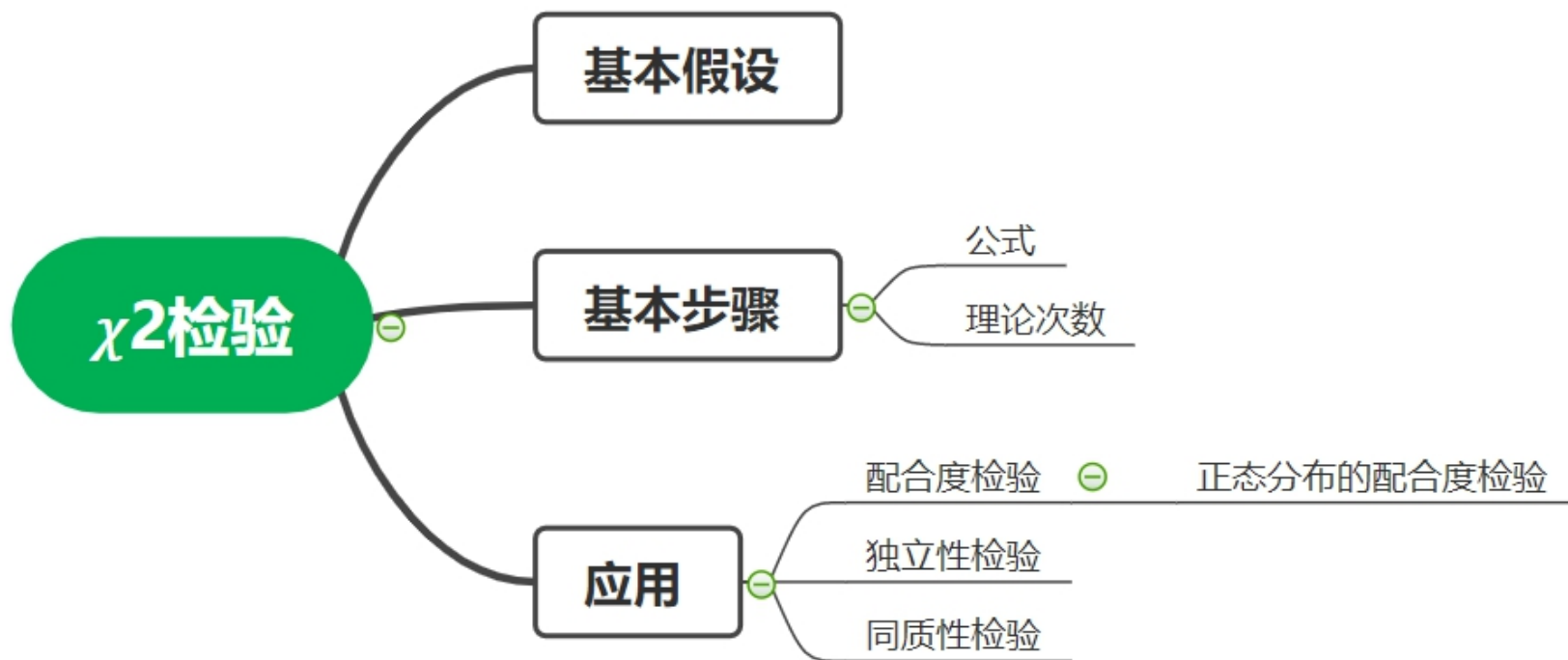


概览

- ◆ 基本原理（选择）
- ◆ 配合度检验（选择、简答、综合）
- ◆ 独立性检验（选择）
- ◆ 同质性检验（选择）



本章结构





一、基本原理

(一) 基本原理

- 处理问题：实际观察频数与理论频数的分布是否相一致

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- 数据类型：计数数据
- 数据形态：不做假定
- 属于非参数检验。
- 分析依据是 χ^2 分布。
- 又称列联表分析、交叉表分析、百分比检验。



一、基本原理

(二) 基本概念

1. 观察频数 f_o

在实验或调查中得到的计数资料，又称实际数、实计数、实际频数。

2. 理论次数 f_e

根据某种理论、原理、计算出来的次数，又称期望次数。



一、基本原理

(三) 基本假设

1. 分类相互排斥，互不包容。
2. 观测值之间相互独立。
3. 每个单元格中理论次数的大小至少在5个以上。或者在自由度很大时，每一个类别的理论次数不小于1，且不超过20%的理论次数可以小于5。



一、基本原理

(四) 计算公式

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

1. 公式呈现近似 χ^2 分布， f_e 越大，越接近。
2. χ^2 值越大，代表统计量和理论值差越大。
3. 公式随着自由度变化，变化趋势和 χ^2 分布一样。



一、基本原理

(五) 基本步骤

1. 计算理论次数
 - 虚无假设成立时候的数值。

2. 列出假设

$$H_0: f_o = f_e$$

$$H_1: f_o \neq f_e$$

3. 计算自由度;

4. 计算 χ^2 值;

5. 查单侧表 χ^2 检验。



一、基本原理

(六) 小期望次数的连续性矫正

1. 单元格合并法
2. 增加样本法
3. 去除样本法
4. 使用矫正公式



二、配合度检验

- 目的：检验一个因素多项分类的实际频数和理论次数是否接近。
 - ✓ 无差假说：各项分类的实际计数没有差异，概率相等。
 - ✓ 假设分布：各项分类的次数分布符合某种已知分布。
 - ✓ 假设比率或百分比：各项分类的次数分布符合某种比例。



二、配合度检验

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- 理论次数（ f_e ）= 总数 × 各分组的概率
- 自由度（df）= 分类项数 - 使用统计量的个数
- 如果小样本的话（期望次数的小于5个）应该使用耶茨连续性校正公式进行校正。

$$\chi^2 = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$



- 已知552名学生的身高次数分布，问这些学生的身高是否符合正态？

[illegible]



二、配合度检验

- ① 计算 f_e ， $f_{ej} = N * y_j * \frac{i}{s}$ y_j 是第 j 个分组的组中值 X_{cj} 所对应的Z分数的 y 值， $Z_{cj} = \frac{X_{cj} - \bar{X}}{s}$
- ② 设： $H_0: f_{0j} = f_{ej}$ $H_1: f_{0j} \neq f_{ej}$
- ③ $\chi^2 = \sum \frac{(f_{0j} - f_{ej})^2}{f_{ej}} = 3.905$
- ④ 计算自由度：理论次数小于5的两组就近合并，剩9组，计算理论次数用了“平均数、标准差、总数”三个统计量，所以 $df = 9 - 3 = 6$
- ⑤ 查表不显著。



三、独立性检验

- 目的：检验两个或两个以上因素各种分类间是否有关联的问题。
- 数据格式： $R \times C$ 表（列联表） 的格式。

类别X	类别Y			合计
	Y1	Y2	Y3	
X1	x_1y_1	x_1y_2	x_1y_3	f_{x1}
X2	x_2y_1	x_2y_2	x_2y_3	f_{x2}
合计	f_{y1}	f_{y2}	f_{y3}	N



三、独立性检验

1. 通用公式

$$\chi^2 = \sum \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \text{ 或者 } \chi^2 = N \left(\sum \frac{f_{oi}^2}{f_{xi}f_{yi}} - 1 \right)$$

- ✓ $f_e = \frac{f_{xi}f_{yi}}{N}$
- ✓ $df = (R - 1)(C - 1)$
- ✓ 其中 f_{oi} 为该格的次数; f_{xi} 为该格所在行的总次数; f_{yi} 为该格所在列的总次数。



三、独立性检验

1. 通用公式

$$\chi^2 = \sum \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \text{ 或者 } \chi^2 = N \left(\sum \frac{f_{oi}^2}{f_{xi}f_{yi}} - 1 \right)$$

- ✓ $f_e = \frac{f_{xi}f_{yi}}{N}$
- ✓ $df = (R - 1)(C - 1)$
- ✓ 其中 f_{oi} 为该格的次数; f_{xi} 为该格所在行的总次数; f_{yi} 为该格所在列的总次数。

类别X	类别Y			合计
	Y1	Y2	Y3	
X1	x1y1	x1y2	x1y3	f_{x1}
X2	x2y1	x2y2	x2y3	f_{x2}
合计	f_{y1}	f_{y2}	f_{y3}	N



三、独立性检验

2. 四格表 (2×2列联表) 独立性检验

(1) 独立样本

- $$\chi^2 = \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)}$$

- $$df = (R - 1)(C - 1)$$

- 若列联表中理论次数小于5, 一般需要进行耶茨校正:

$$\chi^2 = \frac{N\left(|AD - BC| - \frac{N}{2}\right)^2}{(A + B)(A + C)(B + D)(C + D)}$$

- 注: 此独立指的是变量的独立

类别X	类别Y	
	Y1	Y2
X1	A	B
X2	C	D



三、独立性检验

(2) 相关样本

- $$\chi^2 = \frac{(A-D)^2}{A+D}$$

- $$df = (R - 1)(C - 1)$$

若列联表中某格的理论次数小于5，同样需要校正：

- $$\chi^2 = \frac{(|A-D| - 1)^2}{A+D}$$

类别X	类别Y	
	Y1	Y2
X1	A	B
X2	C	D



四、同质性检验

目的：检验不同人群母总体在某一个变量的反应是否有显著差异，即双样本在单一变量的分布情形是同质还是异质。从而了解多个样本的数据是否可以合并。



四、同质性检验

示例：从四所幼儿园分别随机抽出6岁儿童若干,各自组成一个实验组,进行识记测验。测验材料是红、绿、蓝三种颜色书写的字母,以单位时间内的识记数量为指标,结果如下。问四组数据是否可以合并分析。

分 组	红色字母	绿色字母	蓝色字母
1	24	17	19
2	15	12	9
3	20	20	14
4	10	25	28



四、同质性检验

分组	红色字母	绿色字母	蓝色字母	小计	自由度	χ^2
1	24(20)	17(20)	19(20)	60	2	1.3
2	15(12)	12(12)	9(12)	36	2	1.5
3	20(18)	20(18)	14(18)	54	2	1.33
4	10(21)	25(21)	28(21)	63	2	8.86
					$\Sigma df = 8$	$\Sigma \chi^2 = 12.99$
合并	69	74	70	213	2	0.20

变异原因	χ^2	自由度	P
合并 χ^2	0.20	2	$P > 0.05$
异质性 χ^2	12.79	6	$P < 0.05$
总计	12.99	8	



真题演练

下列关于 χ^2 检验和方差分析的表述，正确的有（ ）。（多选）
(312, 2013)

- A. 两者都是无方向性的检验方法
- B. χ^2 检验和方差分析都必须做齐性检验
- C. χ^2 检验属于非参数检验，方差分析属于参数检验
- D. χ^2 检验与方差分析都要求测量数据呈正态分布



真题演练

下列关于 χ^2 检验和方差分析的表述，正确的有（ ）。 （多选）
(312, 2013)

- A. 两者都是无方向性的检验方法
- B. χ^2 检验和方差分析都必须做齐性检验
- C. χ^2 检验属于非参数检验，方差分析属于参数检验
- D. χ^2 检验与方差分析都要求测量数据呈正态分布

答案：AC



真题演练

某研究将工人、农民、公务员、商人按生活满意度分为满意、不满意、介于两者之间三类，现欲研究不同职业类型人员之间的生活满意度的差异。分析该研究数据最恰当的统计方法是（ ）。(312, 2012)

- A. 相关分析 B. χ^2 检验 C. 因素分析 D. t检验



真题演练

某研究将工人、农民、公务员、商人按生活满意度分为满意、不满意、介于两者之间三类，现欲研究不同职业类型人员之间的生活满意度的差异。分析该研究数据最恰当的统计方法是（ ）。(312, 2012)

- A. 相关分析 B. χ^2 检验 C. 因素分析 D. t检验

答案：B



真题演练

近一个世纪以来，某城市的居民患抑郁症、焦虑症、强迫症的比例非常接近。近期，临床心理学家为了考察该城市居民的心理健康状况，进行了一项调查研究。结果发现，抑郁症患者85人，焦虑症患者124人，强迫症患者91人。

请问该城市居民三种神经症患者比例是否发生明显变化

$$(\chi^2_{0.05(2)}=5.99, F_{0.05(3,2)}=9.55, Z_{0.05}=1.96) \quad (312, 2016)$$



真题演练

解：

H0: 该城市居民三种神经症患者比例没有变化

H1: 该城市居民三种神经症患者比例发生明显变化

根据已知题意，分类项数为3， $f_e = \frac{85+124+91}{3} = 100$ ，

采用 χ^2 计算公式

$$\chi^2 = \sum \frac{(f_{oi} - f_{ei})^2}{f_{ei}} = \frac{(85-100)^2}{100} + \frac{(124-100)^2}{100} + \frac{(91-100)^2}{100} = 8.82$$

df=3-1=2，因此比较值 $\chi^2 > \chi^2_{(2)0.05}$ ，拒绝H0。

所以该城市居民三种神经症患者比例发生明显变化的差异在0.05水平上具有统计学意义。



第十二章 非参数检验

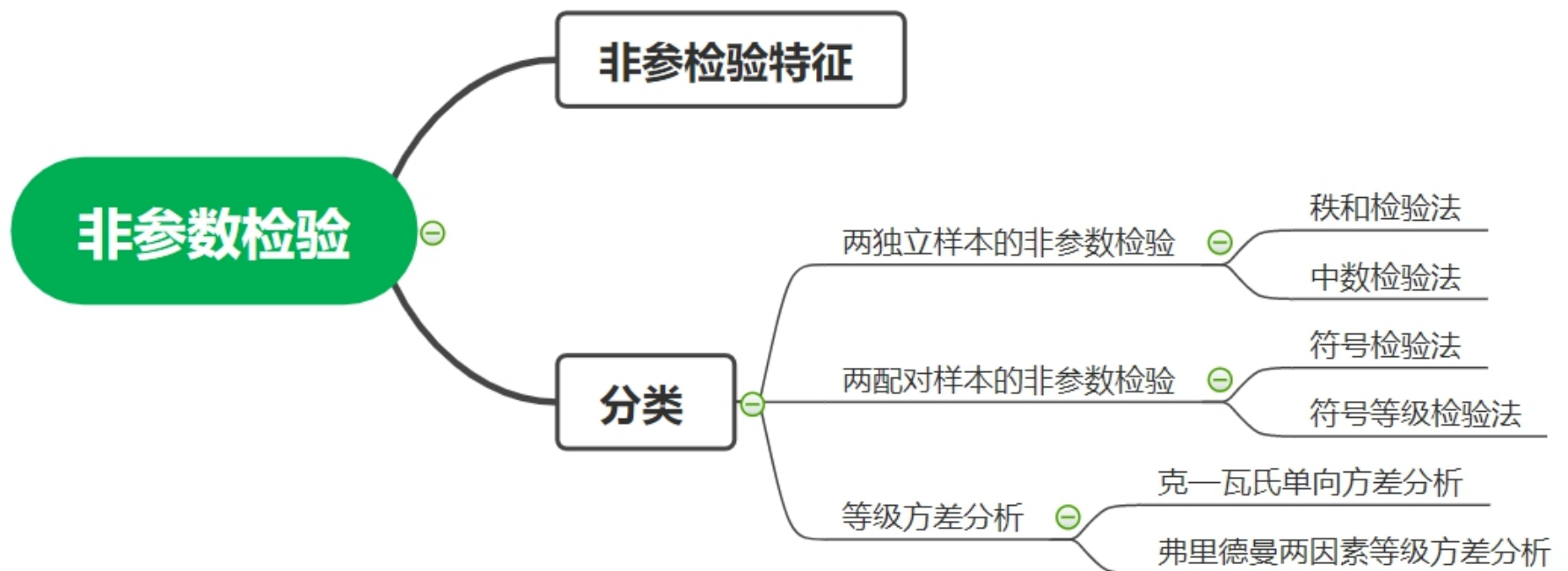


概览

- ◆ 非参数检验特征（选择、简答）
- ◆ 具体的非参数检验（选择）



本章结构





一、非参数检验特征

1. 不需要严格的前提假设。这是它最大的优点。
2. 特别适合顺序变量。
3. 很适合小样本，而且计算简明迅速。
4. 最大的缺点是未能充分利用资料的全部信息。在把等距、等比数据转化为顺序变量的时候，一部分信息就丢失了，这使得非参数检验的精度有所降低。
5. 不能处理交互作用。



二、具体的非参数检验

(一) 两独立样本的非参数检验

1. 秩和检验法

- 对应于参数检验中的“独立样本的t检验”

2. 中数检验法

- 对应着参数检验中的“两独立样本平均数之差的t检验”
- 虚无假设：两个独立样本是从具有相同中数的总体中抽取的。



二、具体的非参数检验

(三) 等级方差分析

1. 克-瓦氏单向方差分析
 - 对应“**完全随机设计**”的方差分析
2. 弗里德曼两因素等级方差分析
 - 对应“**随机区组设计**”的方差分析



真题演练

下列关于非参数检验的表述，正确的有（ ）。 （多选） （312, 2017）

- A. 适用于小样本
- B. 变量总体须服从二项分布
- C. 适用于顺序数据资料
- D. 未能充分利用资料的全部信息

某项跟踪研究在一年前后分别测量了同一组小样本的焦虑水平，假设结果为非正态分布，现欲比较前后两次焦虑分数有无差异，应选用（ ）。

（中国科学院大学，2016）

- A. t 检验
- B. F 检验
- C. 中数检验法
- D. 符号等级检验法



真题演练

下列关于非参数检验的表述，正确的有（ ）。 （多选） （312, 2017）

- A. 适用于小样本
- B. 变量总体须服从二项分布
- C. 适用于顺序数据资料
- D. 未能充分利用资料的全部信息

答案：ACD

某项跟踪研究在一年前后分别测量了同一组小样本的焦虑水平，假设结果为非正态分布，现欲比较前后两次焦虑分数有无差异，应选用（ ）。

（中国科学院大学，2016）

- A. t 检验
- B. F 检验
- C. 中数检验法
- D. 符号等级检验法



真题演练

下列关于非参数检验的表述，正确的有（ ）。 （多选） （312, 2017）

- A. 适用于小样本
- B. 变量总体须服从二项分布
- C. 适用于顺序数据资料
- D. 未能充分利用资料的全部信息

答案：ACD

某项跟踪研究在一年前后分别测量了同一组小样本的焦虑水平，假设结果为非正态分布，现欲比较前后两次焦虑分数有无差异，应选用（ ）。

（中国科学院大学，2016）

- A.t 检验
- B.F 检验
- C.中数检验法
- D. 符号等级检验法

答案：D



第十三章 线性回归

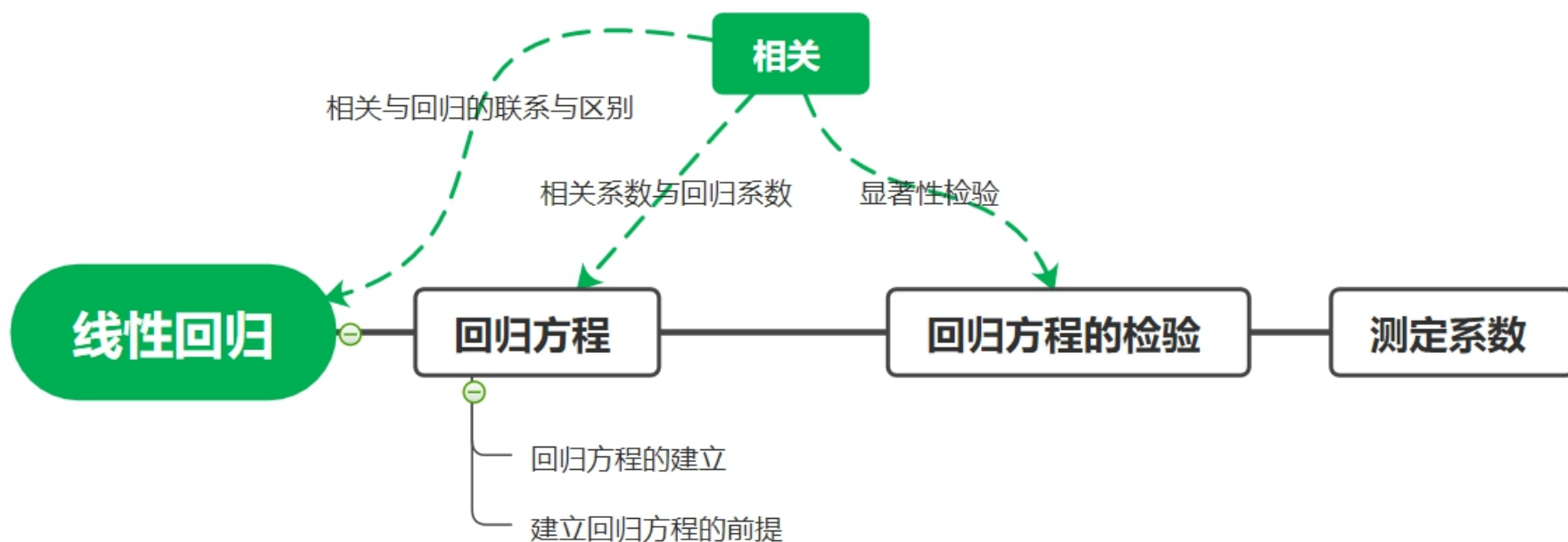


概览

- ◆ 相关与回归（选择、简答）
- ◆ 回归模型的建立、检验（选择、简答）
- ◆ 测定系数（选择、简答）



本章结构





一、一元线性回归概述

(一) 线性回归

1. 回归分析是用一定的数学模型表示变量相关关系的方法。
2. 一元线性回归是当只有一个自变量并且大体成一次函数的线性关系的回归分析。



一、一元线性回归概述

(二) 一元线性回归模型

1. $\hat{Y} = a + bX$

- a : 截距。
- b : 回归系数, 斜率。又叫做 \hat{Y} 对 X 的回归系数, 用 $b_{Y.X}$ 或 $X \rightarrow Y$ 表示。

2. $\hat{X} = a' + b'Y$

- a' : 截距。
- b' : 回归系数, 斜率。又叫做 \hat{X} 对 Y 的回归系数, 用 $b_{X.Y}$ 或 $Y \rightarrow X$ 表示。



一、一元线性回归概述

(三) 线性回归的基本假设

1. 线性关系假设：

✓ X 与 Y 在总体上具有线性关系，最基本假设。

2. 正态性假设： Y 服从正态分布。

3. 独立性假设

✓ 一个 X 对应的 Y 值与另一个 X 对应的 Y 值间独立。

✓ 不同 X 产生的误差相互独立，误差与 X 间也相互独立。

4. 误差等分散性假设

✓ X 产生的误差，呈随机化的常态分布，变异量应相等。

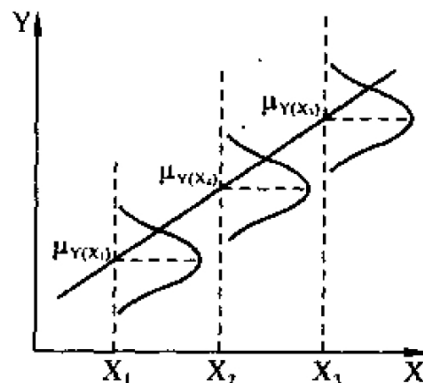


图 12-3 回归样本示意图



二、相关与回归的关系

1. 区别

- ① 相关是分析两变量的关系密切程度；回归是分析两变量的预测关系。
- ② 相关是双向关系；回归是单向关系。
- ③ 相关用相关系数表示密切程度；回归用回归系数表示密切程度。
- ④ 相关用散点图表示；回归是用线形图表示。

2. 联系

- ① 广义上，相关包含回归。
- ② 做回归方程前需要做散点图确定两个变量存在相关关系。
- ③ 相关系数与回归系数可以互相换算。

$$b_{Y \cdot X} = r \times \frac{S_Y}{S_X} \quad b_{X \cdot Y} = r \times \frac{S_X}{S_Y} \quad r = \sqrt{b_{Y \cdot X} \times b_{X \cdot Y}}$$



三、回归模型

(一) 建立回归模型

1. 平均数法

- 将数据按照奇偶分成两组，然后分别代入回归方程，形成二元一次方程组后分别解出a和b。

被试	A	B	C	D	E	F	G	H	I	J
Y	1	1	3	3	4	5	6	7	8	9
X	0	2	1	5	4	2	6	2	5	7

$$(1+3+4+6+8) = 5a + (0+1+4+6+5) b$$

$$(1+3+5+7+9) = 5a + (2+5+2+2+7) b$$

- $\hat{Y} = -0.4 + 1.5x$



三、回归模型

2. 最小二乘法

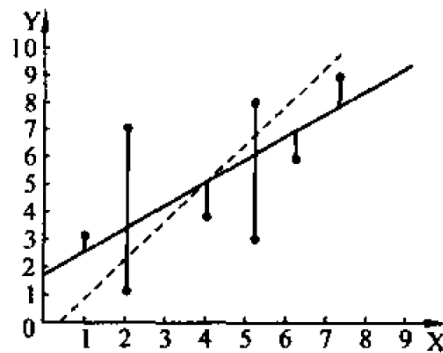
- 误差平方和最小原理，即散点图中每一点沿Y轴方向到直线的距离（ $Y - \hat{Y}$ ）的平方和最小。

- $$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

- $$a = \bar{Y} - b\bar{X}$$

- 或者

- $$b_{Y.X} = r \times \frac{S_Y}{S_X}$$





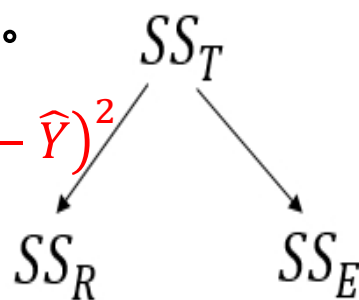
三、回归模型

(二) 检验回归模型

1. 方差分析

- 利用方差分析的方法对回归模型进行有效性检验。

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$



变异来源	平方和	自由度	均方	F	p
回归	SS_R	$df_R = 1$	$MS_R = SS_R / df_R$	MS_R / MS_E	$< \text{or } > \alpha$ (如0.05)
残差	SS_E	$df_E = N-2$	$MS_E = SS_E / df_E$		
总变异	SS_t	$df_t = N-1$			

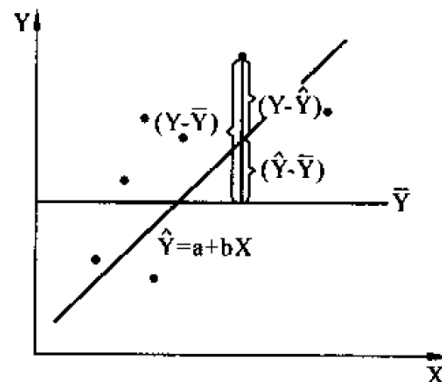


图 12-5 线性回归变异分析示意图



三、回归模型

2. 假设检验

- 对回归系数进行显著性检验。
- 回归系数服从t分布，使用t检验
- $H_0: \beta=0$ β 代表总体回归系数
- $H_1: \beta \neq 0$
- $t = \frac{b-\beta}{SE_b}$

- $SE_b = \sqrt{\frac{S_{YX}^2}{\sum (X - \bar{X})^2}} \quad S_{YX} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N-2}} = \sqrt{MS_E}$

- S_{YX} 代表误差的标准差

➤ 对回归系数的假设检验和对相关系数的假设检验会产生相同结果。



三、回归模型

(三) 测定系数

- r^2 作为回归效果的检验，被称为测定系数，又称决定系数，是指回归平方和在总平方和中所占的比例，或者Y的变异中有 r^2 是有变量X引起的。



真题演练

线性回归分析的前提假设有（ ）。(多选) (312, 2011)

- A. 变量总体服从正态分布
- B. 个体间随机误差相互独立
- C. 自变量的个数多于因变量的个数
- D. 因变量与自变量存在线性关系

建立一元线性回归模型的主要方法有（ ）。(多选) (312, 2018)

- A. 因素分析法
- B. 平均数方法
- C. 最小二乘法
- D. 逐步回归法



真题演练

线性回归分析的前提假设有（ ）。(多选) (312, 2011)

- A. 变量总体服从正态分布
- B. 个体间随机误差相互独立
- C. 自变量的个数多于因变量的个数
- D. 因变量与自变量存在线性关系

答案：ABD

建立一元线性回归模型的主要方法有（ ）。(多选) (312, 2018)

- A. 因素分析法
- B. 平均数方法
- C. 最小二乘法
- D. 逐步回归法



真题演练

线性回归分析的前提假设有（ ）。(多选) (312, 2011)

- A. 变量总体服从正态分布
- B. 个体间随机误差相互独立
- C. 自变量的个数多于因变量的个数
- D. 因变量与自变量存在线性关系

答案：ABD

建立一元线性回归模型的主要方法有（ ）。(多选) (312, 2018)

- A. 因素分析法
- B. 平均数方法
- C. 最小二乘法
- D. 逐步回归法

答案：BC



真题演练

在回归分析中，考察回归效果使用的指标是（ ）。(312, 2008)

- A. 内部一致性系数
- B. 决定系数
- C. 概化系数
- D. 列联系数



真题演练

在回归分析中，考察回归效果使用的指标是（ ）。(312, 2008)

- A. 内部一致性系数
- B. 决定系数
- C. 概化系数
- D. 列联系数

答案：B



真题演练

假定有人研究了学生的大学英语平均成绩对考研英语成绩的预测效果，随机抽取199名考研学生的大学英语平均成绩和考研成绩，计算得二者相关系数为0.81，回归方程有效性检验结果如下表所示。请根据上述资料回答如下问题：（江西师大，2016）

变异来源	平方和	自由度	均方	F 值	Sig.
回归	22500.00	②	⑤	⑦	0.000
残差	①	③	⑥		
总体	32350.00	④			

- (1) 请写出①~⑦序号替代的数值，补充完整该表格。
- (2) 所建方程是否有效？为什么？
- (3) 大学英语成绩可以决定考研成绩变异的比例是多少？

THANKS

- 期待下次相遇 -



高途学院APP下载



高途学院公众号

