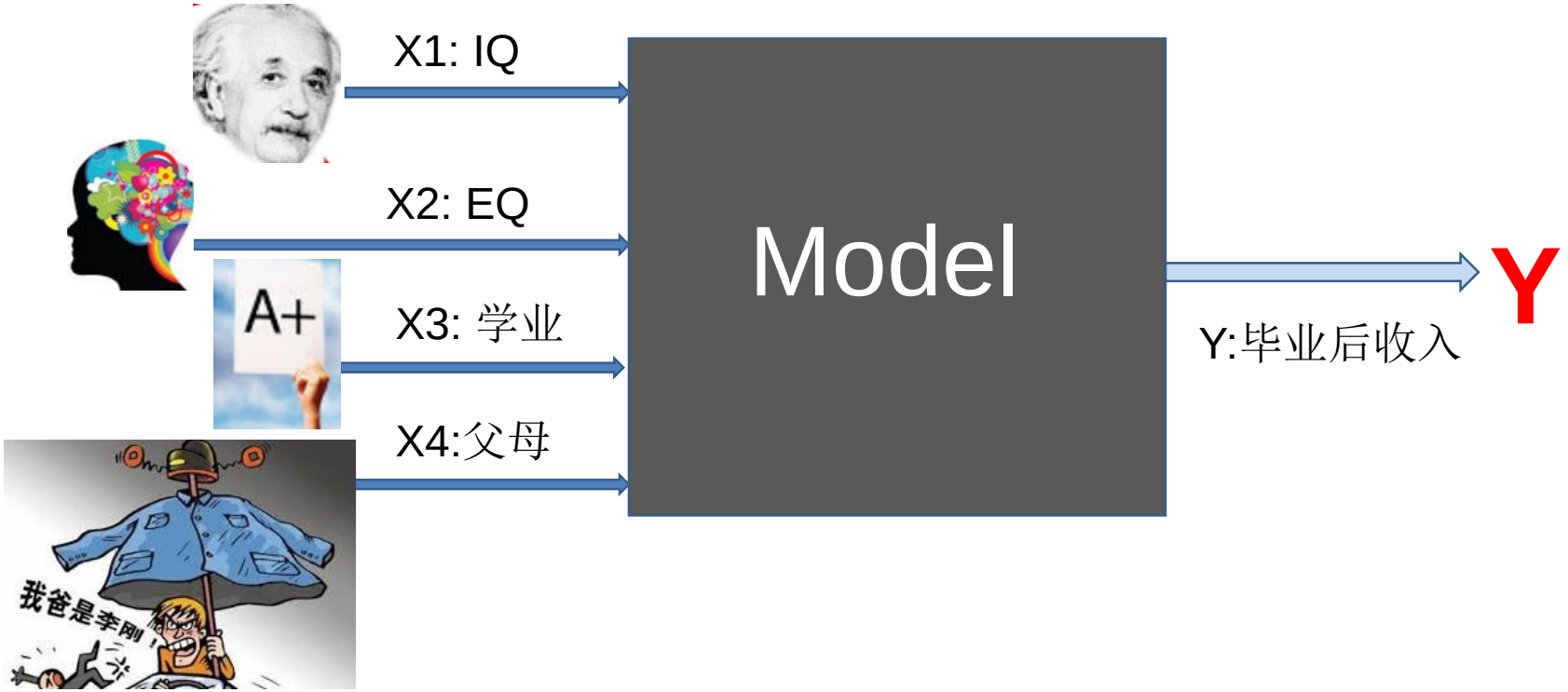
第六章：数据分析初步

Data Analysis I

王树佳 | 深圳大学经济学院

sjwang123@163.com

模型



输入

输出

数据分析初步

本章探讨

1. 如何描述每个变量的基本特征？
2. X与Y的关系，或X对Y的影响
   A. X对Y是否有显著影响？
   B. X与Y是什么关系？
   C. X对Y的影响有多大？

Contents

1. 描述性统计
   - 定性变量
   - 定量变量
2. 两个变量之间的关系
   - 定量对定量
   - 定量对定性

案例：毕业生薪水调查

介绍：为了研究毕业生工作后收入情况及影响因素，英国某商学院在2004年针对1999年毕业的大学生进行了一项调查，以研究他们毕业5年后的情况变化。

抽样：把所有1999年毕业的商学院学生的名字和地址列表（包括全日制学生和在职学生）作为抽样框， 然后按照随机数字表，随机抽取了400名毕业生进行问卷调查。

调查：首先设计好一份问卷，包括毕业情况、毕业后的第一份工作和现在的工作情况。然后写一封信，说明调研的目的以及请求被访者的合作。把信和问卷一并寄给被抽中的毕业生。如果6个星期后没有回复，再次寄出一封信和问卷，并电话询问。

调查数据

总共收回了357份问卷。把问题进行编码，录入电脑，保存在EXCEL文件SURVEY2004.xls中。

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID-CODE	GENDER	MONTHS	DEGREE	TYPE	Salary99	JOB99	AREA99	Salary04	JOB04	AREA04	NO-JOBS
2	1	0	2	2	1	14160	3	7	29756	1	7	2
3	2	0	3	2	0	14004	1	8	21776	6	7	3
4	3	1	1	3	1	13680	1	7	22808	2	8	2
5	4	1	3	3	0	13476	1	5	23480	1	6	4
6	5	1	3	2	1	14604	1	7	30260	1	7	1
7	6	1	3	1	0	11588	7	5	24392	6	6	1
8	7	0	10	1	1	14544	4	8	31928	4	8	3
9	8	1	2	3	0	13332	3	8	22160	3	9	3
10	9	1	3	2	0	14436	4	7	29276	5	8	2
11	10	1	3	2	0	14076	2	7	28412	2	8	2

数据编码

名称	说明（DESCRIPTION）
ID-CODE	编号
GENDER	性别: 0=Female, 1=Male
MONTHS	毕业后几个月开始第一份工作
DEGREE	毕业等级: 1 - First Class Honours（一等荣誉学士学位） 2 - Upper Second Class Honours(二级甲等荣誉学士学位) 3 - Lower Second Class Honours(二级乙等荣誉学士学位) 4 - Third Class Honours（三等荣誉学士学位） 5 - Other.
TYPE	是否工读交替制课程(sandwich course)? 0 = Yes, 1 = No.

数据编码

名称	说明（DESCRIPTION）
SAL99	第一份工作的年薪
JOB99	第一份工作所从事的行业（INDUSTRY） 行业（INDUSTRY）编号见后面
AREA99	第一份工作所在地区(REGION) 地区(REGION)编号见后面
SAL04	2004年年薪
JOB04	2004年所从事的行业（INDUSTRY）
AREA04	2004年工作所在地区(REGION)
NO-JOBS	2004年前转换工作的次数

数据编码

REGION	
1	Northern England.
2	Yorks/Humberside.
3	North West England.
4	East Anglia.
5	East Midlands.
6	West Midlands.
7	Greater London.
8	South East [Excl GL].
9	South West.

INDUSTRY	
1	Manufacturing.
2	Public Utilities.
3	Retailing.
4	Financial Sector.
5	Public Administration.
6	Education.
7	Higher Degree.

数据分析主要内容

一、描述性统计

1. 样本分布情况：如男女比例多少？一级荣誉学位占百分之多少？行业分布如何？等等
2. 变量基本情况：如第一年（第五年）平均年薪多少？平均多久找到第一份工作？平均跳槽次数多少？等等

二、哪个因素对毕业后的收入有影响？关系如何？

分别探讨每个自变量对因变量的影响。

三、探讨多个因素一起对毕业后收入的影响。

将在下一讲讨论

统计软件Minitab

Minitab软件是现代质量管理统计的领先者，全球六西格玛实施的共同语言，以无可比拟的强大功能和简易的可视化操作深受广大质量学者和统计专家的青睐。Minitab 1972年成立于美国的宾夕法尼亚州州立大学（Pennsylvania State University），到目前为止，已经在全球100多个国家，4800多所高校被广泛使用。

网址：http://www.minitab.com.cn/

Contents

描述性统计(Descriptive Statistics)

描述性统计就是对原始数据进行初步归纳和描述，以便发现数据可能存在的一些规律。

描述性统计的内容:

- 频数分布(分类变量）
- 数据的集中趋势（平均值）
- 数据离散程度（方差、标准差等）
- 数据的分布形状

描述性统计的常用方法:

- 数值描述
- 图形表示（可视化）

变量的分类

1. 定量数据(Quantitative，Measured)

说明现象的数量特征。

案例：MONTHS，SAL99，SAL04，NO-JOBS

2. 定性数据(Qualitative，Attribute)

回答的结果是定性（文字型）的，说明的是事物的品质、属性特征。

案例：GENDER，DEGREE，TYPE，JOB99，AREA99，JOB04，AREA04

Contents

定性变量的描述性统计

首先要进行赋值，将其数量化。

例如性别：女=0，男=1。

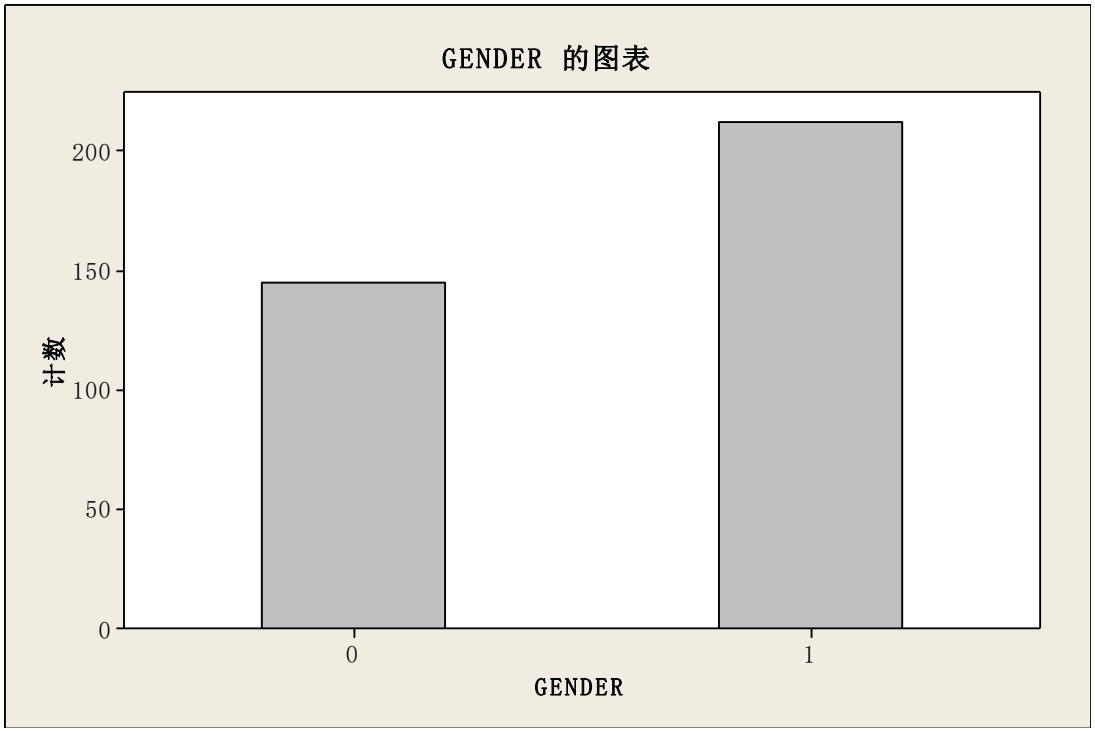
数值描述：统计-表格-单变量计数（Minitab操作）

离散变量计数：性别

GENDER	计数	百分比
0	145	40.62
1	212	59.38
N=	357	

定性变量的描述性统计

图形描述：图形-条形图-唯一值计数-简单



定性变量的描述性统计

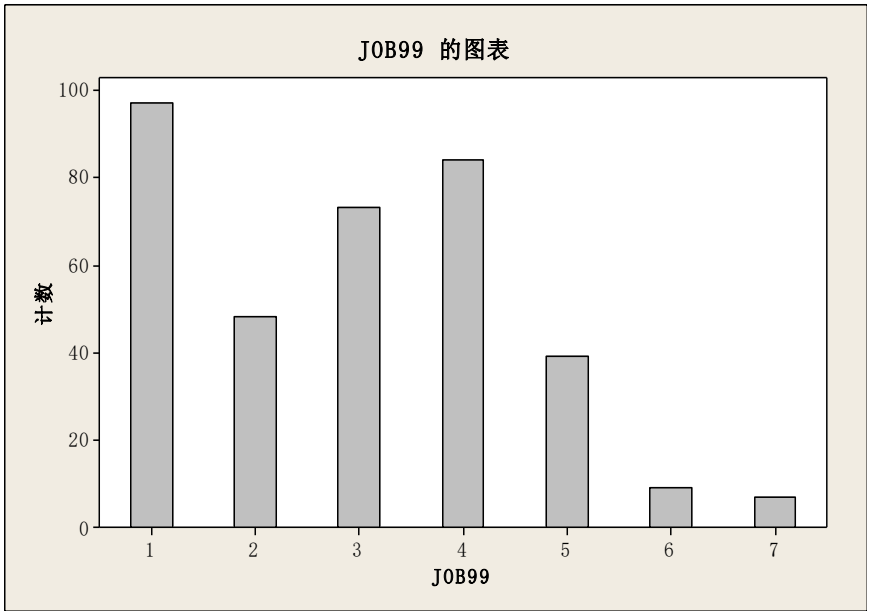
例：1999年和2004年从事行业

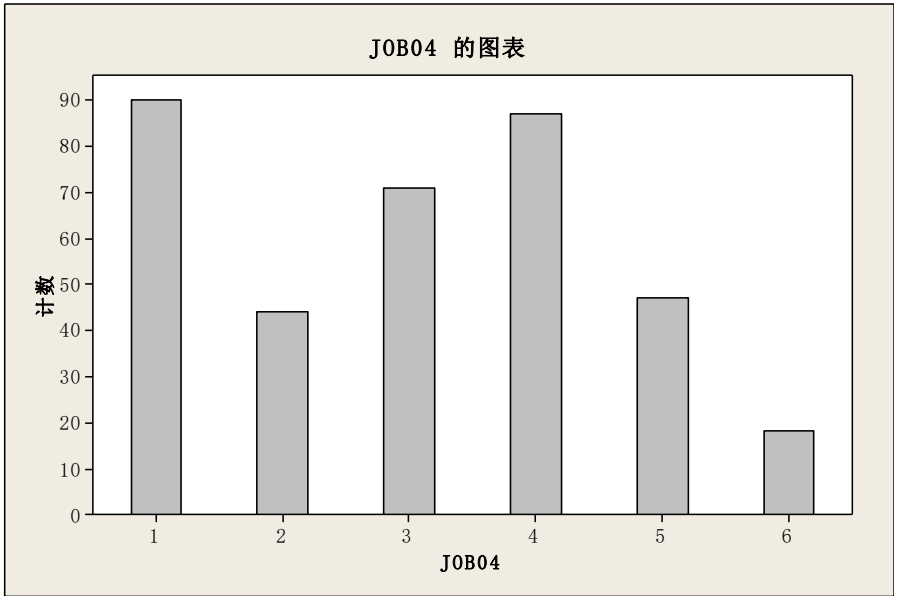
JOB99	计数	百分比
1	97	27.17
2	48	13.45
3	73	20.45
4	84	23.53
5	39	10.92
6	9	2.52
7	7	1.96
N=	357	

JOB04	计数	百分比
1	90	25.21
2	44	12.32
3	71	19.89
4	87	24.37
5	47	13.17
6	18	5.04
N=	357	

定性变量的描述性统计

例：1999年和2004年从事行业





解释？合理？

Contents

1. 描述性统计
   - 定性变量
   - 定量变量
2. 两个变量之间的关系
   - 定量对定量
   - 定量对定性

定量变量的描述性统计

图形描述：

直方图(Histogram)、箱线图(Box plot)

数值描述：

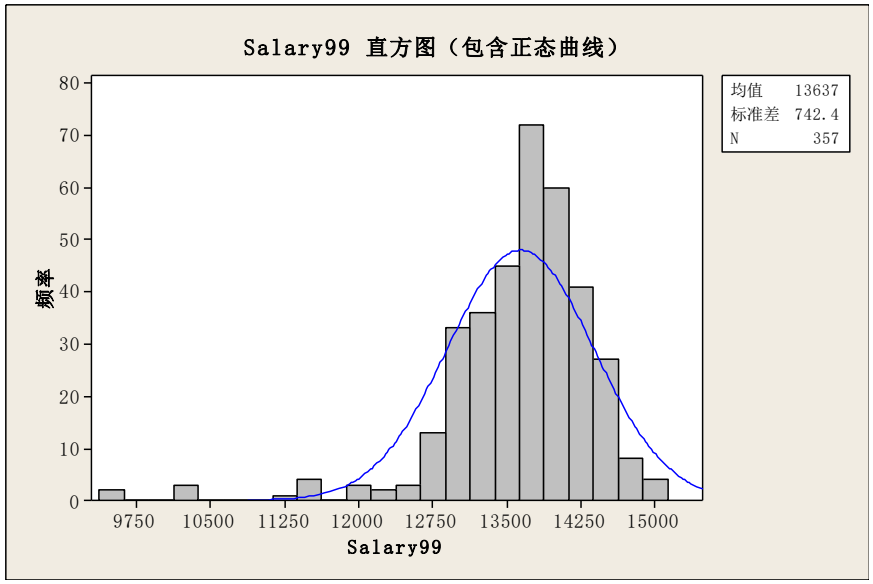
平均数、中位数、众数、最大值、最小值、四分位数，方差、标准差，偏度、峰度等

这些数据可以反映出样本数据的集中趋势、离差程度、形状等分布特征，是描述性统计的重要内容。

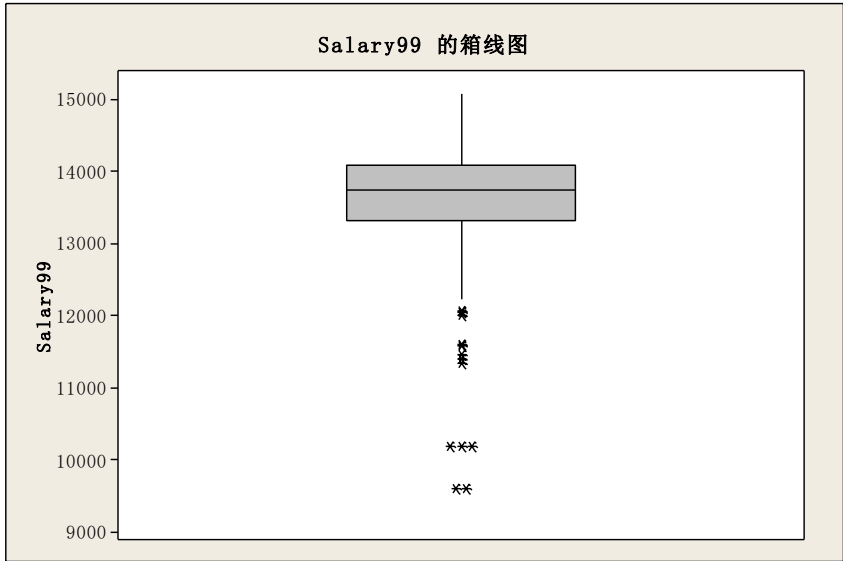
定量变量的描述性统计

Salary99：

直方图



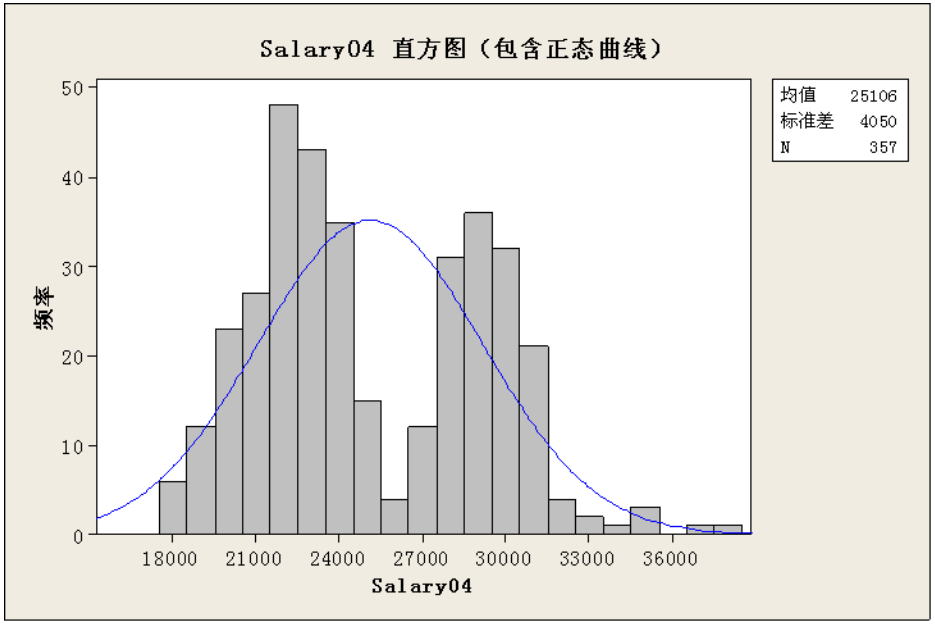
箱线图



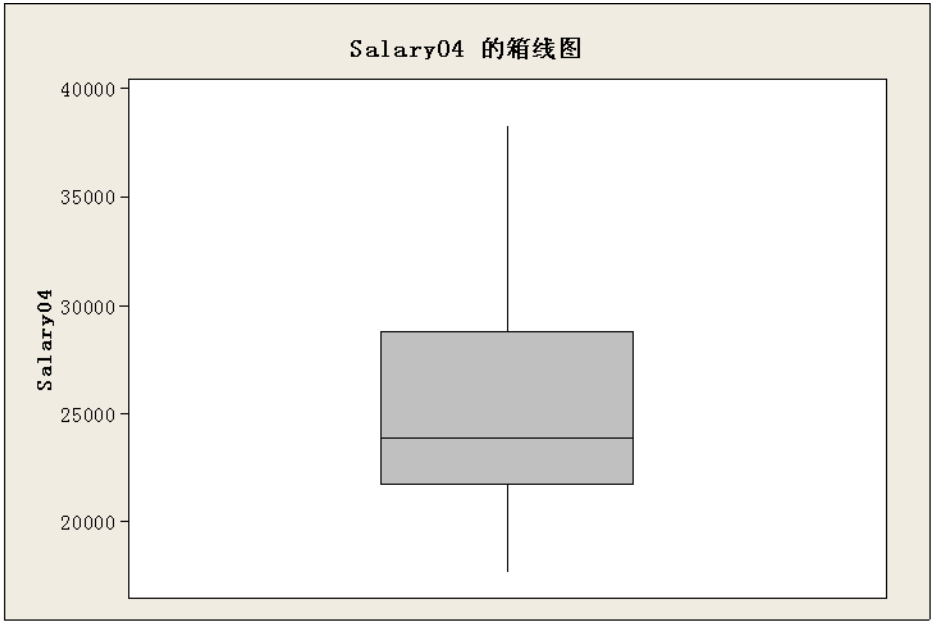
定量变量的描述性统计

Salary2004：

直方图



箱线图



定量变量的描述性统计

描述性统计: Salary99

变量	平均值	标准差	变异系数	最小值	下四分位数	中位数	上四分位数	最大值
Salary99	13637	742	5.44	9588	13320	13740	14088	15072

变量	模式	众数的 N	偏度	峰度
Salary99	14004	7	-2.07	7.91

描述性统计: Salary04

变量	平均值	标准差	变异系数	最小值	下四分位数	中位数	上四分位数	最大值
Salary04	25106	4050	16.13	17720	21776	23888	28814	38300

变量	模式	众数的 N	偏度	峰度
Salary04	21776	6	0.34	-0.78

定量变量的描述性统计

1．数据分布的中心（集中趋势）(Measures of central tendency)

衡量样本数据分布的集中趋势的统计指标

- 平均数(mean)
- 中位数(median)
- 众数(mode)

定量变量的描述性统计

2．数据分布的变异程度 (Measures of dispersion)

衡量样本数据分布的广度（数据的取值范围）和变异程度的统计指标.

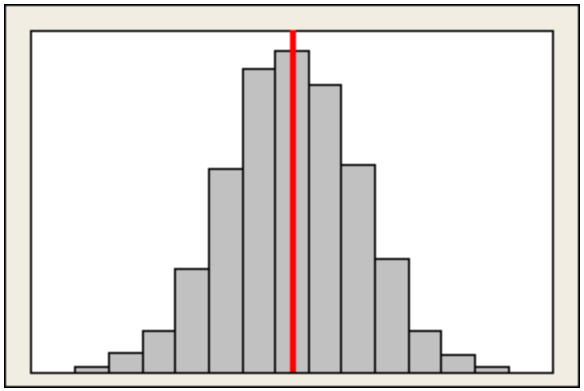
- 极差(Range)：R =最大值 - 最小值 ；
- 分位差，如四分位差(Interquartile range)；
- 平均差（离差绝对值的算术平均）；
- 方差σ^2，标准差σ (最常用）。

定量变量的描述性统计

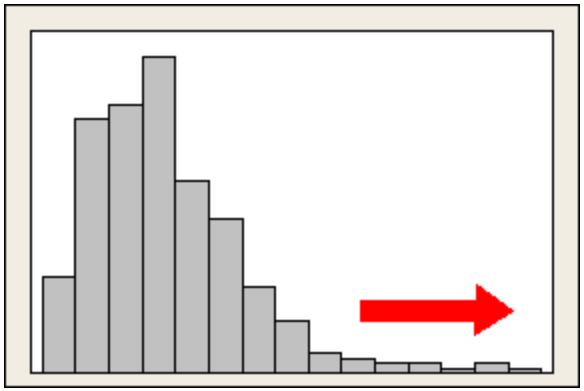
3．分布的形状（Distribution shape）

对称性：偏度(Skewness)

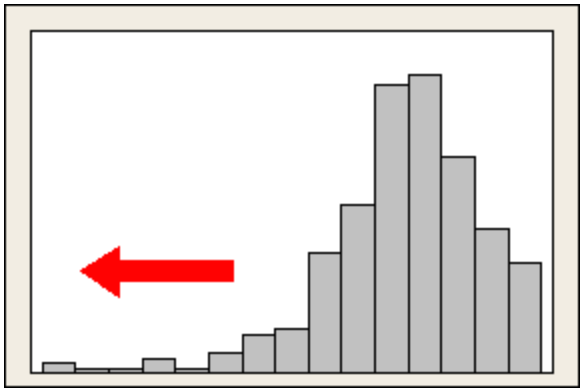
接近0：对称



正值：偏右



负值：偏左



定量变量的描述性统计

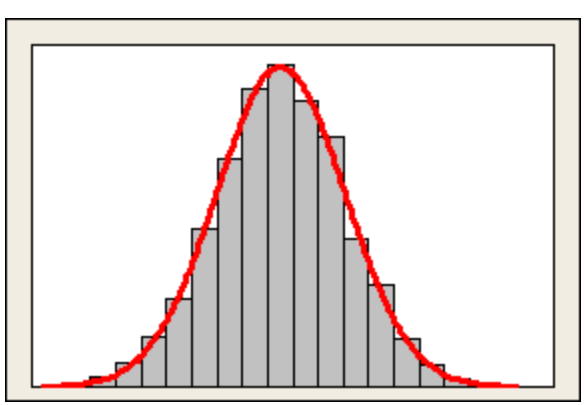
3．分布的形状（Distribution shape）

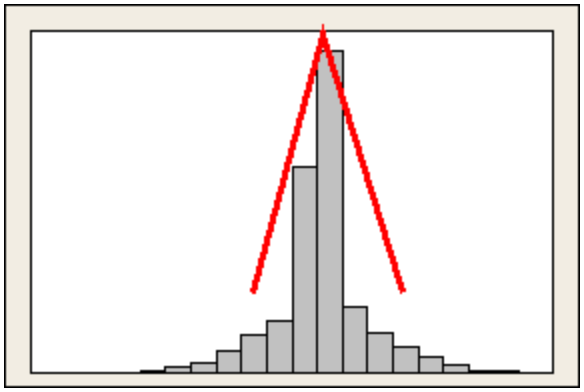
扁平性：峰度(Kurtosis)

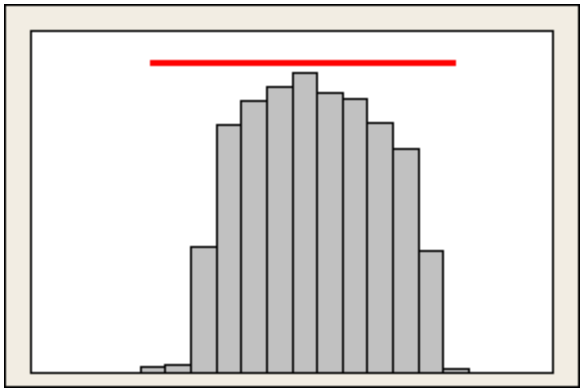
接近于 0
正态分布

正值：＞ 0
比正态分布尖峰

负值：＜ 0
比正态分布扁平







定量变量的描述性统计

3．分布的形状（Distribution shape）

对称性：偏度(Skewness)

扁平性：峰度(Kurtosis)

$$\text{Skew} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^3$$

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^4 - 3$$

描述性统计: 小结

1. 定性变量：图示？数值？
2. 定量变量：图示？数值？

Contents

1. 描述性统计
   - 定性变量
   - 定量变量
2. 两个变量之间的关系
   - 定量对定量
   - 定量对定性

两个变量之间的关系

因变量与自变量

因变量根据研究目的确定，自变量由影响因变量的因素确定。

（1）定量对定量（MvM）：线性回归

（2）定量对定性（MvA）：假设检验及方差分析

（3）定性对定量（AvM）：Logistics 回归

（4）定性对定性（AvA）：列联表

探索变量之间关系一般方法

Step1: 初步分析

利用数据从图形和数值两方面对变量之间的关系进行简单的描述性分析。

Step2: 假设检验

如果初步分析无法得出结论，则需要进行假设检验。

Step3: 结论及解释

如果结论为两个变量之间存在相互关系，则需要对其关系进行详细描述，如存在什么样的关系式，密切程度如何等等。

Contents

1. 描述性统计
   - 定性变量
   - 定量变量
2. 两个变量之间的关系
   - 定量对定量
   - 定量对定性

探讨两个定量变量之间的关系

Step1: 初步分析

初步分析

A. 图形：绘制散点图

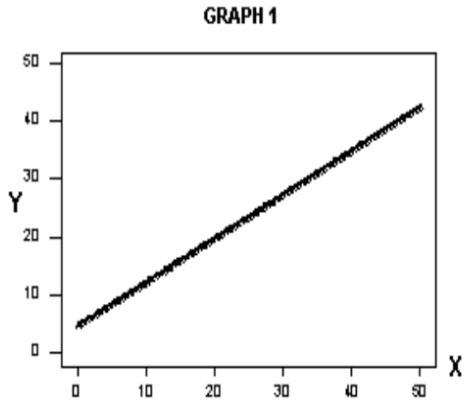
散点图直观判断：函数类型（线性？曲线？）以及关系的密切程度。

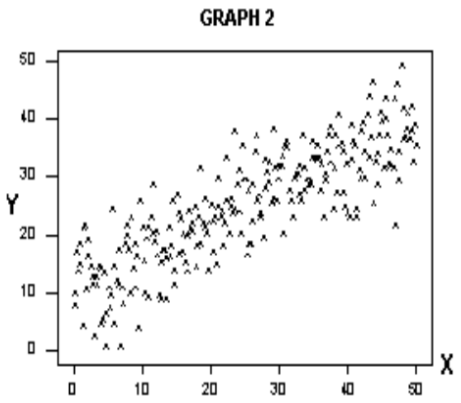
B. 数值：计算可决系数R^2。

可决系数R^2可以从数值上判断线性关系的密切程度。

$R^2=0$：无线性关系; $R^2=1$：线性关系

散点图

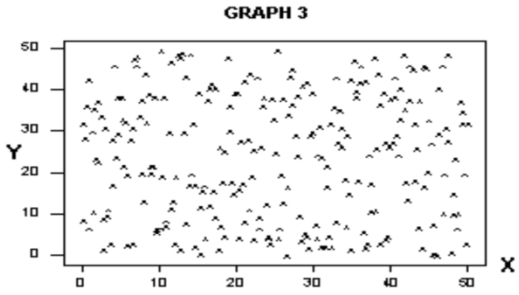


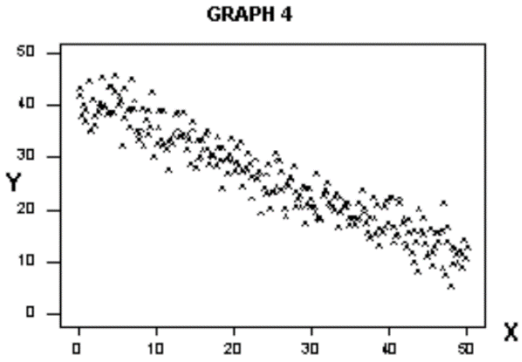


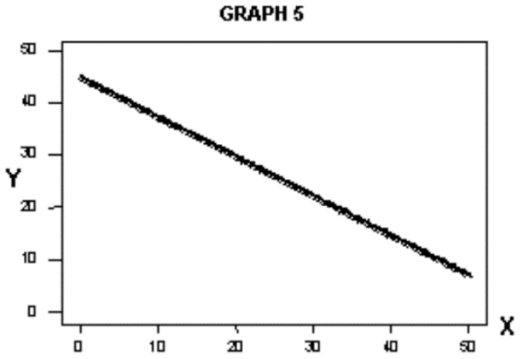
Y与X关系如何？

模型：$y = \alpha + \beta x + \varepsilon$

数据：(x_i, y_i)







方差分析表

离差来源	离差平方和	自由度	均方差	F
回归（R）	SSR	1	MSR=SSR/1	F=MSR/MSE
残差（E）	SSE	*n*-2	MSE=SSE/(*n*-2)	
合计（T）	SST	*n*-1		-

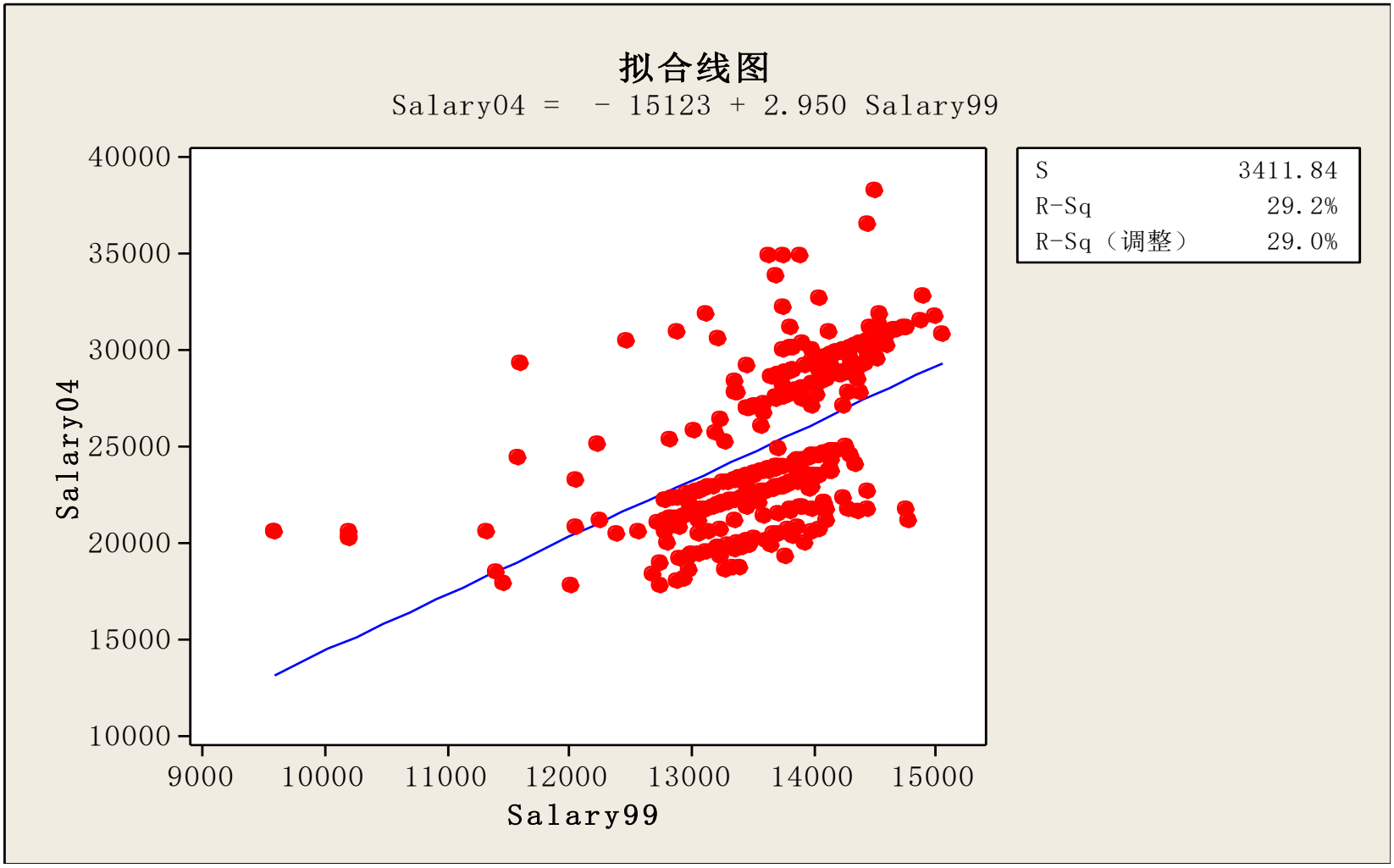
可决系数：$R^2 = SSR/SST$

R^2=因变量被自变量解释的程度

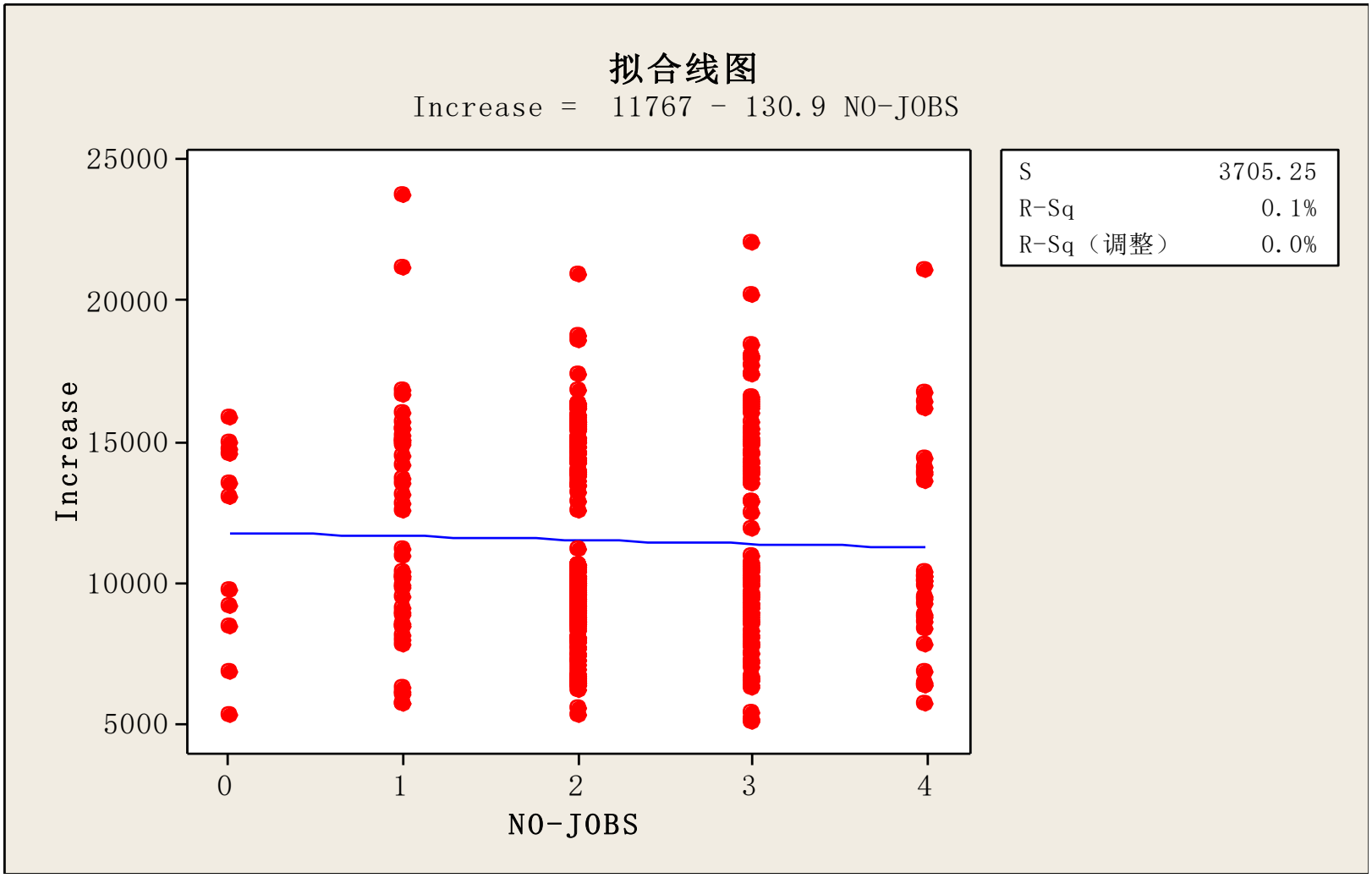
如何得出初步分析结论？

1. 如果数据点非常集中于直线，且R^2接近1，则可认为存在线性关系；
2. 如果数据点沿水平方向均匀分布，且R^2接近0，则可认为两变量没有关系；
3. 否则，则初步分析不能有明确结论，要进行假设检验。

问题1：2004年薪水与起薪有没有关系？



问题2：收入增长与跳槽次数有没有关系？



探讨两个定量变量之间的关系

Step2: 假设检验

相关性检验（F-检验）

1. 如果初步分析无法得出明确结论，则进行假设检验（Hypothesis testing）以明确结论。
2. 对于两个变量都是定量变量(即M v M型)，采用检验为F-检验，即检验两个变量之间是否存在显著关系（回归方程的显著性检验）。

假设检验问题：

$\boldsymbol{H_0}$: 两个变量之间没有线性关系（即总体$\boldsymbol{R^2 = 0}$）

$\boldsymbol{H_1}$: 存在显著的线性关系（即总体$\boldsymbol{R^2 > 0}$）

F-检验的步骤

1．提出假设

$H_0: R^2 = 0$，两个变量之间没有线性关系

$H_1: R^2 > 0$，两个变量之间存在显著线性关系

2．计算F统计量及p-值

3．判别：给定显著性水平α (比如0.05），

- 若p-值<α，则拒绝零假设H_0，自变量x与因变量y之间存在显著的线性关系;
- 若p-值 $\geq \alpha$，则接受零假设H_0，自变量x与因变量y之间不存在线性关系。

4. 结论

问题3：2004年薪水与起薪有没有关系？

回归分析: Salary04 与 Salary99

回归方程为

Salary04 = - 15123 + 2.95 Salary99

自变量	系数	系数标准误	T	P
常量	-15123	3326	-4.55	0.000
Salary99	2.9499	0.2436	12.11	0.000

S = 3411.84　R-Sq = 29.2%　R-Sq（调整） = 29.0%

方差分析

来源	自由度	SS	MS	F	P
回归	1	1707642265	1707642265	146.70	0.000
残差误差	355	4132425655	11640636		
合计	356	5840067919			

问题3：2004年薪水与起薪有没有关系？

假设检验（Hypothesis testing）

1. 提出假设

   $H_0: R^2 = 0$，两个变量之间没有线性关系；

   $H_1: R^2 > 0$，两个变量之间存在线性关系。

2. 计算F统计量及p-值：

   Minitab：统计-回归 或：统计-回归-拟合线图

   得到F-统计量：F =146.70，p-值=0.000

3. 判别：因p-值<0.05，则拒绝零假设H_0

4. 结论：2004年薪水与起薪之间存在显著的线性关系。

问题3：结论

2004年薪水与1999年起薪之间存在着显著的线性关系，回归方程式如下：

Salary04 = - 15123 + 2.95 Salary99

回归系数为2.95，表示1999年每增加一英镑，五年后（2004年）薪水平均将增加2.95英镑。

可决系数为 R^2=29.2% ，表明起薪对毕业生工作五年后的薪水有29.2%的影响。

注：1.是否显著；2.影响大小；3.解释程度。

定量对定量: 小结

1. 初步分析
   散点图；可决系数R^2
2. 假设检验
   F-检验，四个步骤
3. 结论描述
   三方面：1）是否显著；2）影响大小；3）解释程度

Contents

1. 描述性统计
   - 定性变量
   - 定量变量
2. 两个变量之间的关系
   - 定量对定量
   - 定量对定性

第一种情形: M v A(2)

自变量为二值变量(Binary): 只有两个可能结果

- 毕业后收入与性别有关系吗？
- 毕业生薪水与课程类型有关系吗？
- 长期抽烟会得癌症吗？

一个定性变量对一个定量变量的影响

Step1: 初步分析

M v A(2)

初步分析

A. 图形：箱线图

把两个水平的因变量值做箱线图进行比较。

判断：

1）如果两个箱线图完全分离，可认为两个水平有显著差异，即两个变量显著相关。

2）如果两个箱线图完全重合，则可认为两个水平无显著差异，即两个变量没有相关。

3）如果两个箱线图部分重合（既不完全分离，也没有完全重合），则初步分析不能下结论，要进行假设检验。

初步分析

B. **数值**：平均值、标准差等

分别求出两个水平的相关描述性指标，如单位个数、平均值、标准差等，通过这些数据比较两个水平之间的差异程度。

判断：

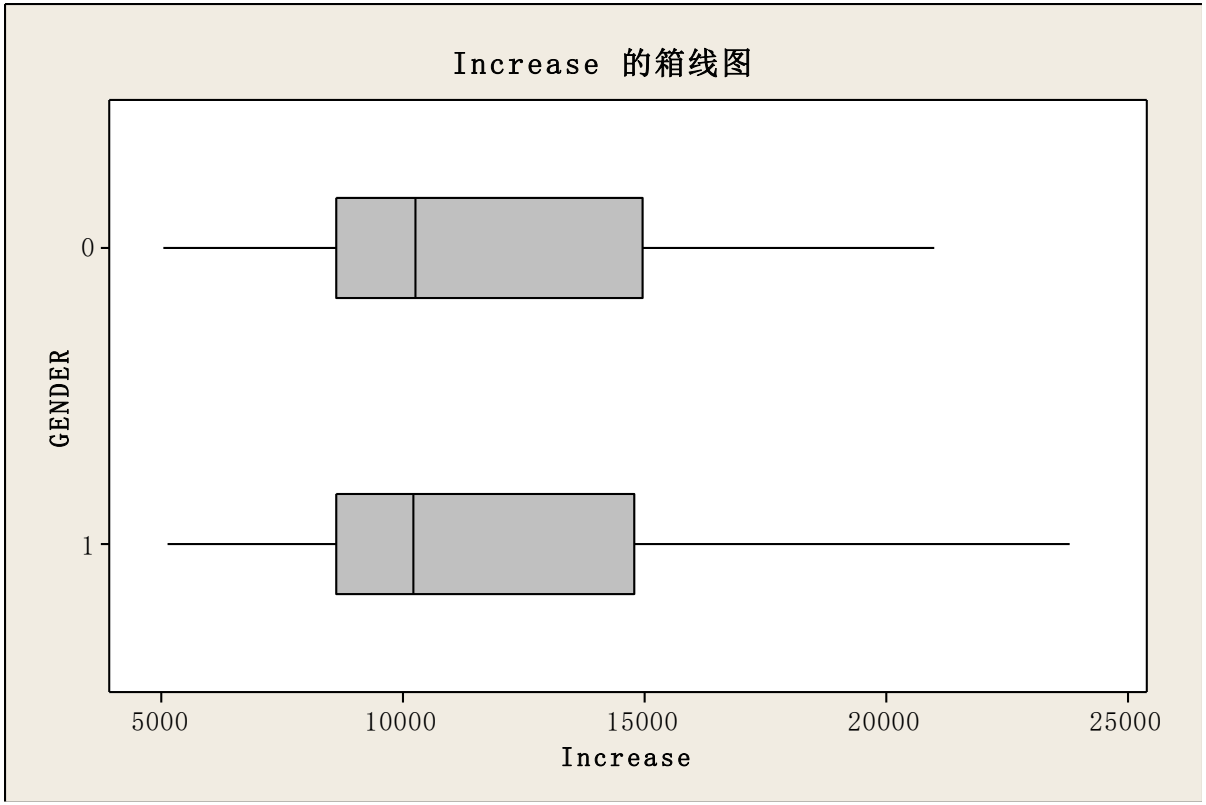
1）如果两个均值相差很大（如差值在3个标准误以上），则认为两个水平有显著差异，即两个变量显著相关。

2）如果两个均值相差很小（如差值不到1个标准误），则可认为两个水平无显著差异，即两个变量没有相关。

3）如果相差不大不小，则数值初步分析不能下结论，要进行假设检验。

问题4：薪水增加与性别有关吗？

A. 图形：箱线图。Minitab：图形-箱线图，选择“含组”，在“尺度”中选择“转置值和类别尺度”



问题4：薪水增加与性别有关吗？

B. 数值：平均值、标准差等

Minitab：统计-基本统计量-显示描述性统计，变量选“Increase”，按变量选“Gender”

变量	GENDER	平均值	标准误	标准差	最小值	中位数	最大值
Increase	0	11444	297	3572	5036	10244	20972
	1	11486	261	3796	5108	10208	23780

初步分析结论？

一个定性变量对一个定量变量的影响

Step2: 假设检验

M v A(2)

MvA(2)假设检验（t-检验）

1. 如果初步分析无法得出明确结论，则进行假设检验（Hypothesis testing）以明确结论。
2. 比较两个水平上的因变量值是否存在显著差异（即MvA(2)型），使用双样本t-检验。

假设检验问题：

$H_0: \mu_1 = \mu_2$, 两个变量之间没有关系

$H_1: \mu_1 \neq \mu_2$, 两个变量之间有关系

双样本t-检验的步骤

1．提出假设

$H_0: \mu_1 = \mu_2$，两个变量之间没有关系

$H_1: \mu_1 \neq \mu_2$，两个变量之间有关系

2．计算t-统计量及p-值

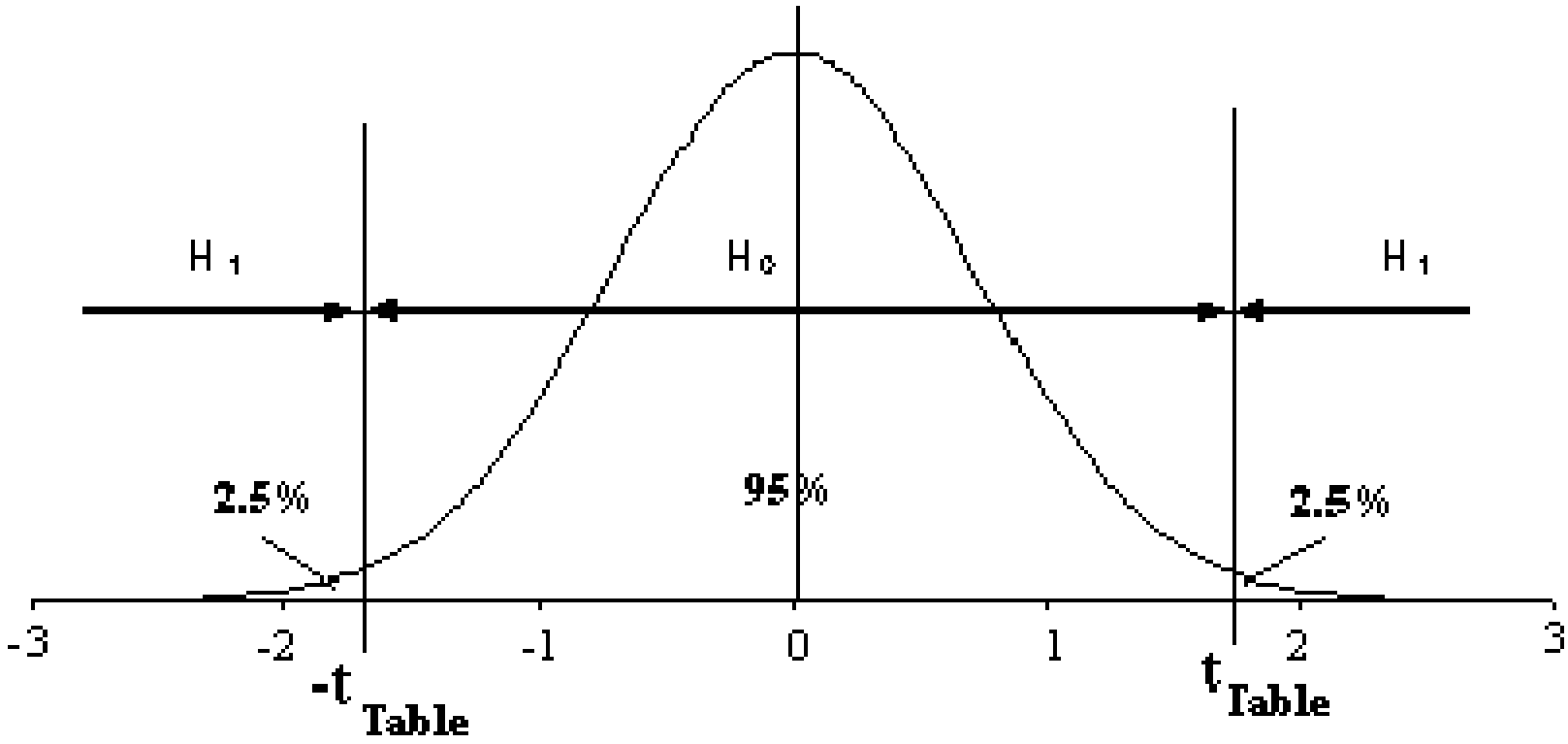
3．判别：给定显著性水平α (比如0.05），

A. 若p-值<α，则拒绝零假设H_0，自变量x与因变量y之间存在显著关系，即自变量两个水平之间存在显著差异;

B. 若p-值 $\geq \alpha$，则接受零假设H_0，自变量x与因变量y之间不存在显著关系，即自变量两个水平之间不存在显著差异。

4. 结论

t-检验图

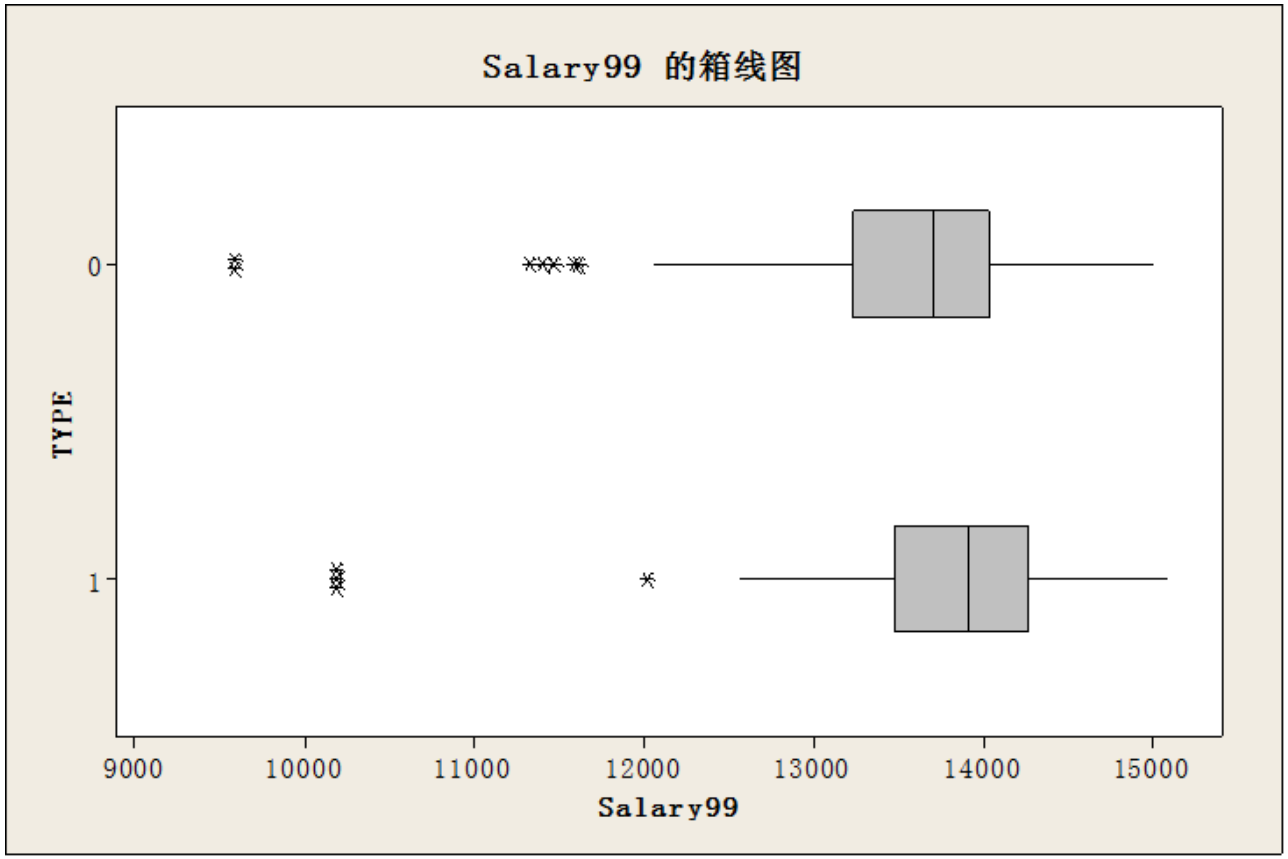


案例：起薪与课程类别有关系吗？

1. 初步分析
2. 假设检验
3. 结果描述

问题5：起薪与课程类别有关系吗？

初步分析：图形（箱线图）



问题5：起薪与课程类别有关系吗？

初步分析：数值（数字特征）

变量	TYPE	平均值	标准误	标准差	最小值	中位数	最大值
Salary99	0	13584	43.1	698	9588	13704	15000
	1	13785	86.7	840	10188	13908	15072

初步分析结论？

问题5：起薪与课程类别有关系吗？

1．提出假设：$H_0: \mu_1 = \mu_2$,两个变量之间没有关系；

$H_1: \mu_1 \neq \mu_2$,两个变量之间有显著关系。

2．用Minitab软件的“双样本t-检验”计算t值:

统计-基本统计量-双样本t：

样本：选“Salary99”

下标：选“TYPE”

问题5：起薪与课程类别有关系吗？

双样本 T 检验和置信区间: Salary99, TYPE

Salary99 双样本 T

TYPE	N	平均值	平均值 标准差	平均值 标准误
0	263	13584	698	43
1	94	13785	840	87

差值 = mu (0) - mu (1)

差值估计: -200.7

差值的 95% 置信区间: (-392.0, -9.4)

差值 = 0 (与 ≠) 的 T 检验: **T 值 = -2.07 P 值 = 0.040 自由度 = 141**

问题5：起薪与课程类别有关系吗？

假设检验（Hypothesis testing）

1. 提出假设

$H_0: \mu_1 = \mu_2$, 两个变量之间没有关系

$H_1: \mu_1 \neq \mu_2$, 两个变量之间有关系

2. 计算t-统计量及p-值：

T值=-2.07，P值=0.040，自由度=1413。

3. 判别：因p-值=0.040<0.05，拒绝零假设H_0。

4. 结论：课程类别对毕业生起薪有显著影响。

问题5：结果描述

经假设检验，我们有理由认定课程类型对商学院毕业生的起薪有显著影响。

非三明治课程毕业生平均起薪显著高于三明治课程毕业生。

三明治课程毕业生平均起薪为13584英镑，非三明治课程毕业生平均起薪为13785英镑，三明治课程毕业生高出201英镑。

小结：第一种情形: M v A(2)

1. 初步分析
   A. 图形：箱线图
   B. 数值：自变量两个水平的因变量值数字特征
2. 假设检验：双样本t-检验
3. 结论描述

第二种情形: M v A(3+)

自变量有三个或以上水平

- 年收入与毕业等级(DEGREE)：M v A(5)
- 年收入与从事行业(JOB99,JOB04)：M v A(7)
- 年收入与工作所在地(AREA99,AREA904)：M v A(9)

一个定性变量对一个定量变量的影响

Step1: 初步分析

M v A(3+)

问题6：年收入增长与毕业等级有关系吗？

初步分析

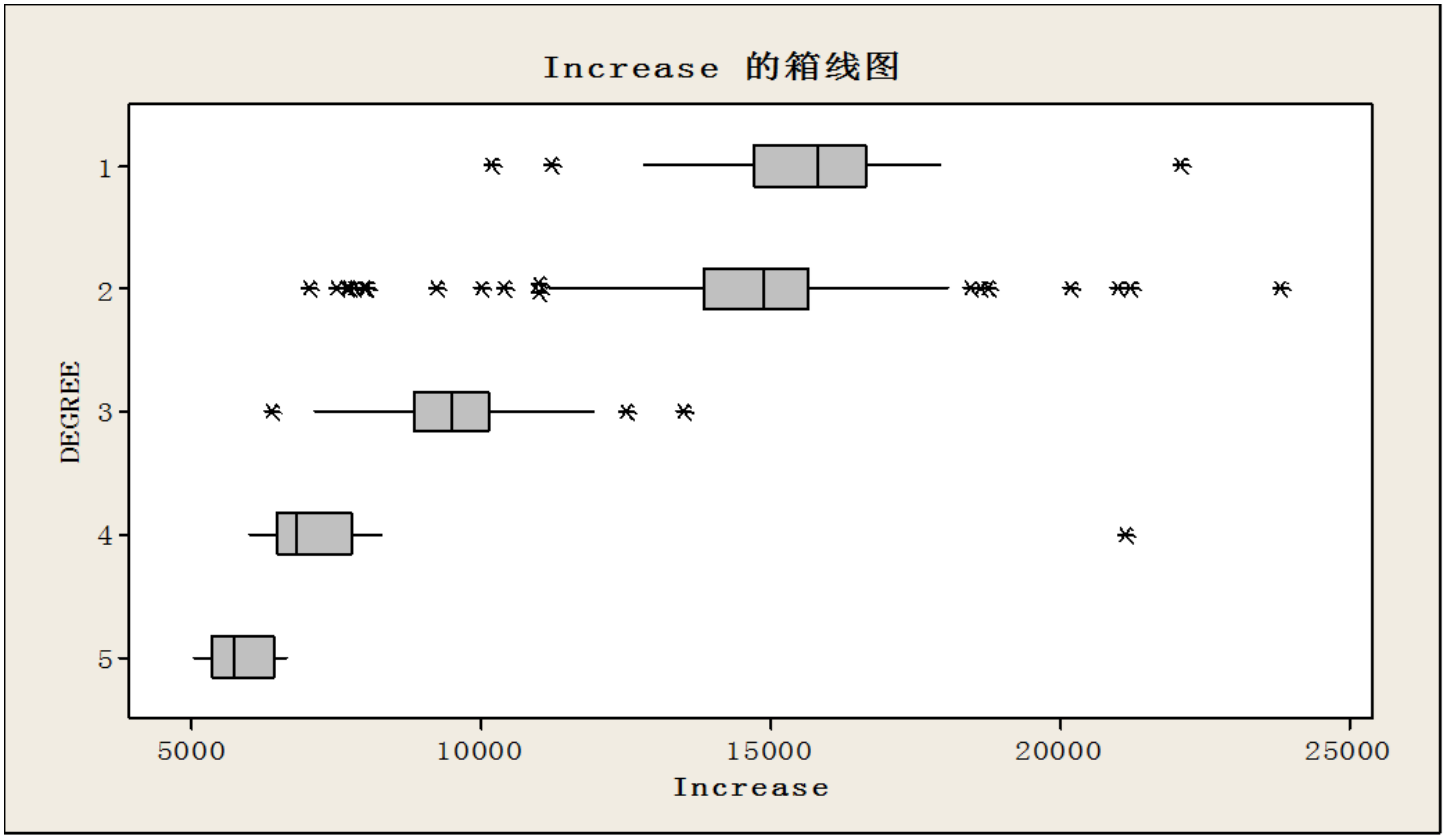
1. 数值:统计-基本统计量-显示描述性统计

变量	DEGREE	平均值	标准误	标准差	变异系数	最小值	中位数	最大值
Increase	1	15643	463	2267	14.49	10172	15788	22076
	2	14559	215	2547	17.49	7016	14840	23780
	3	9479	79	925	9.76	6348	9470	13460
	4	7369	386	2377	32.25	5996	6812	21092
	5	5867	132	575	9.80	5036	5708	6632

问题6：年收入增长与毕业等级有关系吗？

初步分析

2. 图形



问题6：初步分析结论

从图形和数据描述均可看出，毕业生工作五年以后的收入增幅与毕业成绩等级存在明显关系。

毕业成绩越好，毕业后的年收入增长就越快，而且差距明显，成绩最好的毕业生的平均年收入增幅是成绩最差（肄业）的毕业生的2.67倍，两者相差19776英镑。

成绩等级	平均年收入
1	15643
2	14559
3	9479
4	7369
5	5867

问题7：年收入增长与从事行业有关系吗？

初步分析

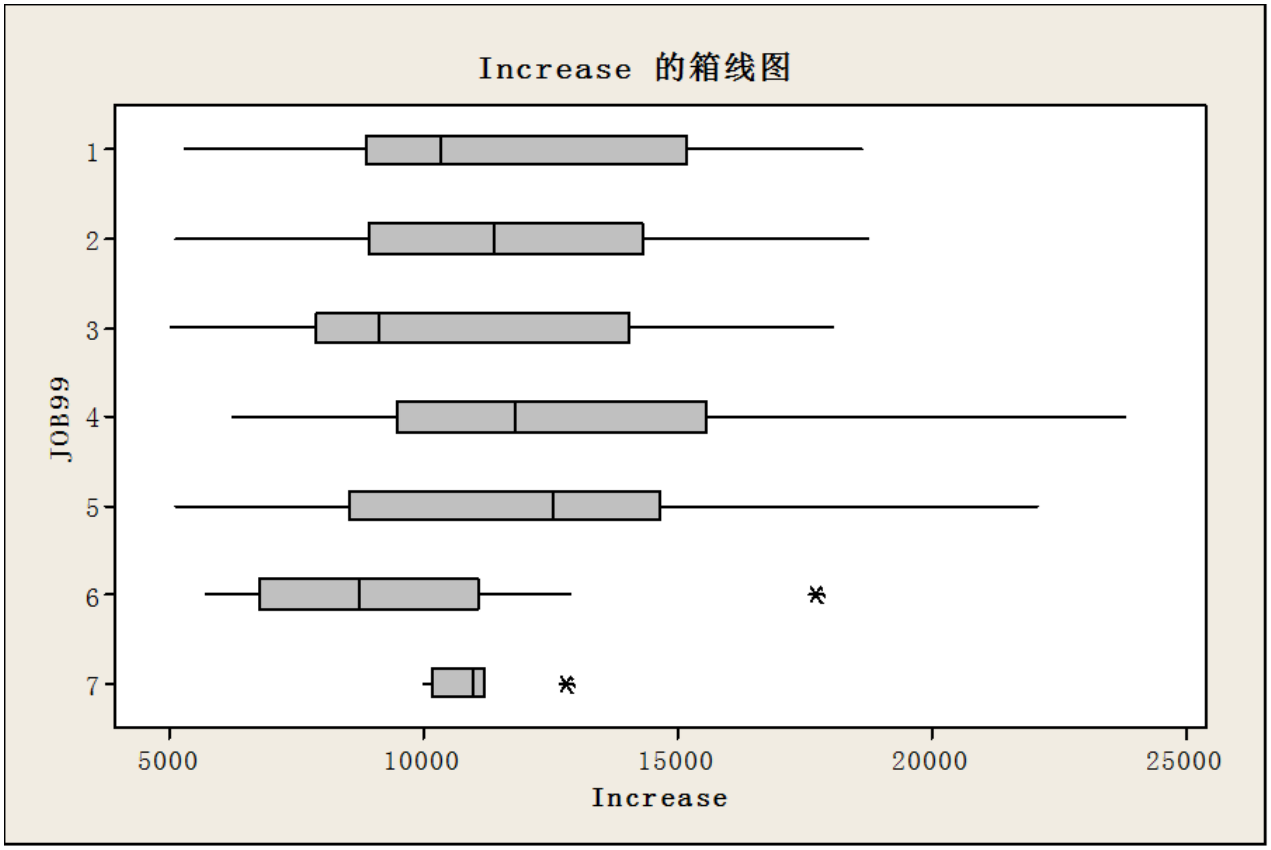
1. 数值:统计-基本统计量-显示描述性统计

变量	JOB99	平均值	标准误	标准差	变异系数	最小值	中位数	最大值
Increase	1	11499	353	3477	30.24	5276	10340	18596
	2	11554	454	3145	27.22	5132	11354	18740
	3	10323	403	3447	33.39	5036	9128	18044
	4	12501	442	4051	32.41	6248	11798	23780
	5	11786	681	4254	36.09	5108	12536	22076
	6	9407	1247	3741	39.77	5708	8732	17720
	7	10922	358	947	8.68	9980	10964	1280

问题7：年收入增长与从事行业有关系吗？

初步分析

2. 图形



一个定性变量对一个定量变量的影响

Step2: 假设检验

$M \neq A(3+)$

MvA(3+)假设检验（单因素方差分析）

1. 如果初步分析无法得出明确结论，则进行假设检验（Hypothesis testing）以明确结论。
2. 比较三个或以上水平的因变量值是否存在显著差异（即MvA(3+)型），使用单因素方差分析。

假设检验问题：

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$: 两个变量之间没有关系

H_1: 至少有一个均值不等: 两个变量之间有显著关系

单因素方差分析F-检验的步骤

1．提出假设

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$: 两个变量之间没有关系

H_1: 至少有一个均值不等: 两个变量之间有显著关系

2．计算F统计量及p-值

3．判别：给定显著性水平α (比如0.05），

A. 若p-值<α，则拒绝零假设H_0，自变量x与因变量y之间存在显著关系，即自变量各个水平之间存在显著差异;

B. 若p-值 $\geq \alpha$，则接受零假设H_0，自变量x与因变量y之间不存在显著关系，即自变量各个水平之间不存在显著差异。

4. 结论

问题7：年收入增长与从事行业有关系吗？

假设检验(单因素方差分析F-检验）

1. 提出假设

   H_0: $\mu_1 = \mu_2 = \cdots = \mu_7$: 两个变量之间没有关系

   H_1: 至少有一个均值不等:两个变量之间有关系

2. 计算样本数据的F-值或p-值

   由下表得到： F-值=2.89，或P-值=0.009

3. **判别**：因为$F = 2.89 > F_{0.05} = 2.10$,

   p-值=0.009<0.05，故接受H_1。

4. 结论：可以认为年收入增长与第一份工作所从事的行业存在显著关系。

统计->方差分析->单因子，“响应”为“Increase”，“因子”为“JOB99”

单因子方差分析: Increase 与 JOB99

来源	自由度	SS	MS	F	P
JOB99	6	229979563	38329927	2.89	0.009
误差	350	4648568977	13281626		
合计	356	4878548539			

S = 3644 R-Sq = 4.71% R-Sq（调整） = 3.08%

平均值（基于合并标准差）的单组 95% 置信区间

水平	N	平均值	标准差	-----+---------+---------+---------+----
1	97	11499	3477	(--*---)
2	48	11554	3145	(----*----)
3	73	10323	3447	(----*---)
4	84	12501	4051	(---*--)
5	39	11786	4254	(-----*-----)
6	9	9407	3741	(----------*----------)
7	7	10922	947	(------------*------------)

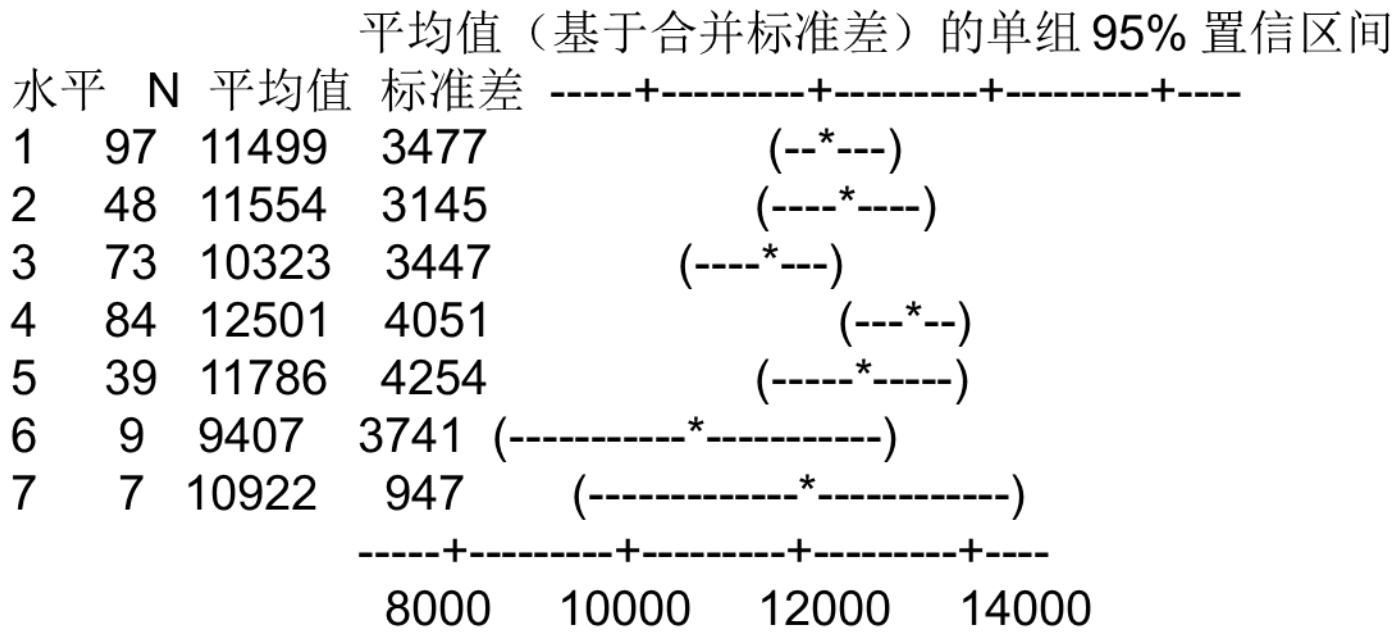
-----+---------+---------+---------+----

8000 10000 12000 14000

合并标准差 = 3644

问题7：结论

假设检验结果表明，年收入增长与第一份工作所从事的行业存在显著关系。其关系为：金融业(4)明显收入增长高于其它行业，零售(3)业明显偏低，而教育(6)和读研(7)则较为分散，有高有低。



小结：
第二种情形: M v A(3+)

1. 初步分析
   A. 图形：箱线图
   B. 数值：自变量各个水平的因变量值数字特征
2. 假设检验：单因素方差分析F-检验
3. 结论描述

Recap

1. 描述性统计
   - 定性变量：图形？数值？
   - 定量变量：图形？数值？
2. 两个变量之间的关系
   - 定量对定量
     - 初步分析
     - 假设检验
   - 定量对定性
     - M v A(2)：初步分析？假设检验？
     - M v A(3+)： 初步分析？假设检验？

讨论课问题

一、关于探讨两个定量变量的关系(MvM)

1. 如何进行初步分析？(看什么？如何下结论?），如何用MINITAB进行相关操作？

2.关于假设检验

（1）在什么情形下应进行假设检验？

（2）如何检验自变量与因变量之间是否存在显著关系?

二、关于探讨定量因变量与定性自变量的关系(MvA(2))

1. 如何进行初步分析？(看什么？如何下结论?），如何用MINITAB进行相关操作？

2.关于假设检验

（1）在什么情形下应进行假设检验？

（2）如何检验自变量与因变量之间是否存在显著关系?

三、关于探讨定量因变量与定性自变量的关系(MvA(3+))

1. 如何进行初步分析？(看什么？如何下结论?），如何用MINITAB进行相关操作？

2.关于假设检验

（1）在什么情形下应进行假设检验？

（2）如何检验自变量与因变量之间是否存在显著关系?

实践课问题

一、关于探讨两个定量变量的关系(MvM)

- 讨论毕业工作五年后年薪增幅与跳槽次数的关系。

二、关于探讨定量因变量与定性自变量的关系

- 讨论年薪增幅与性别的关系；
- 讨论年薪增幅与第一份工作所在地区(REGION)之间的关系 。

数据来源：商学院毕业生薪水调查数据

实践：薪酬的影响因素

美国统计局1995年3月调查了1260个职员的基本情况，以确定影响时新（美元）的关键因素有哪些。**数据**：Ch6 beauty.xls

变量：

（1）工资(Wage)：时薪（单位：美元/小时）

（2）相貌 (Looks)：调查员对被访者相貌吸引力的评价，分五档：

1=很丑,2=比较难看, 3=一般水平，4=比较好看，5=很漂亮或英俊

（3）教育程度(Educ)：受教育程度年限

（4）经验(Exper)：潜在的工作经验(年限)，年龄减去受教育年限再减6(6岁上学)

（5）女性(Female)：1=女性，0=男性

（6）婚姻(Married)：1=已婚，0=未婚

（7）种族(Black)：1=黑人，0=否则

实践：薪酬的影响因素

回答如下问题(如无特殊说明，显著性水平都是0.05)：

1. 教育的回报：分析教育程度（Educ）与工资（Wage）的关系。
2. 工作经验：分析工作经验（exper）与工资（Wage）的关系。
3. 婚姻溢价：婚姻状况(Married)对工资(Wage)是否存在显著影响？
4. 是否存在性别歧视？
5. 分析工资(Wage)与相貌 (Looks)之间的关系。