

第7章 连续变量的统计描述与参数估计

学习目标

- 了解什么是**统计描述**：用少量数字（即描述指标）和图表概括大量原始数据的统计方法
- 掌握**统计描述的指标**，了解指标的含义和用途
- 会利用SPSS的相关功能对数据进行统计描述
- 了解什么是**参数估计**
- 掌握**参数估计的指标**，了解指标的含义和用途
- 会利用SPSS的相关功能对数据进行参数估计
- 掌握**Bootstrap方法**

主要内容

- 7.1 连续变量的统计描述指标体系
- 7.2 连续变量的参数估计指标体系
- 7.3 案例
- 7.4 Bootstrap方法



南京大學
NANJING UNIVERSITY

7.1 连续变量的统计描述指标体系

- 针对不同测量尺度的变量，有不同的描述指标和统计图形与之对应：
 - 连续变量（Scale）：均值、中位数、标准差、方差、偏度系数、峰度系数、直方图、箱图、P-P图、Q-Q图、线图、散点图等。
 - 分类变量（Nominal或Ordinal）：众数（Mode）、比、率、条图、饼图等。

连续变量的统计描述指标体系

- **集中趋势（Central Tendency）**：描述数据的平均水平，如均数（Mean）、几何均数（G）、中位数（Median）、众数（Mode）等。
- **离散趋势（Dispersion Tendency）**：描述数据与平均水平的偏离，如方差（Variance）、标准差（Std.Deviation）、全距（Range）等。
- **分布特征（Distribution Tendency）**：描述数据的分布情况，如偏度系数和峰度系数。

7.1.1 集中趋势的描述指标

算术均数（Arithmetic Mean）

- 简称均数（Mean），有总体均数（ μ ）和样本均数（ \bar{X} ）之分
- 适用于对称分布，特别是正态分布的资料，不适用于偏态分布的资料（为什么呢？）
- 均数的计算方法

中位数（Median）

- 中位数是一组观察值的**位置平均数**：将一组数据按从小到大排好，处于中间位置上的数。
 - **注意：观察值必须先排好序。**
- **用于描述偏态分布资料的中心趋势**，它不受两端特大、特小值的影响，当分布末端无确切数据时也可计算。

其他集中趋势描述指标

- 截尾均数 (Trimmed Mean)

- 数据排序后按照一定比例去掉两端的数据再求均数。
- 适用于两端有极端值的资料。
- 常用5%截尾均数(左右两端各去掉5%)。

其他集中趋势描述指标

- 几何均数（Geometric Mean, G ）
 - 适用于呈倍数关系的等比资料或对数正态分布的资料，尤其是对数正态分布的计量资料。
 - 应用中应注意观察值不能同时有正有负。
 - 同一资料算得的几何均数小于算术均数。

其他集中趋势描述指标

- 众数 (Mode)

- 样本数据中出现频次最高的数值

- 适用于单峰数据，反映出出现频次最高的数据情况

- 调和均数 (Harmonic Mean)

- 观察值倒数的均数的倒数，较少使用

7.1.2 离散趋势的描述指标

引例

- 两组数据：
 - A公司6个月的销售额（单位：万元）：36, 33, 39, 36, 37, 35
 - B公司6个月的销售额（单位：万元）：23, 29, 38, 35, 40, 51
 - 两个公司的平均销售额相等，均为36万元/月，但两组数据的离散程度不同。

离散趋势指标与集中趋势指标的关系

- 对连续变量的描述，需要将集中趋势和离散趋势结合起来，才能对其分布有全面的认识。
- 离散趋势指标是用来衡量集中趋势指标代表性好坏的，一般来讲，离散程度越低，集中趋势指标代表数据平均水平的程度越高。

全距（Range , R ）

- 又称**极差**，即最大和最小观察值之间的间距，用全距描述资料的离散程度简单明了。
- 不能反映观察值的整个变异度。
- 用于预备性检查，大体上了解数据的分布范围，以确定随后的分析方法。

四分位间距

- **四分位数**：将一组数据按从小到大排好，进行四等分，三个分割点上对应的数据称为四分位数，分别为下四分位数 Q_L 、中位数和上四分位数 Q_u 。
- **四分位间距** $Q = Q_u - Q_L$ 。
- 四分位间距比极差稳定，但仍未考虑每个观察值的变异度。
- 适用于偏态分布的资料，特别是末端无确切数据时。

百分位数（Percentile）

- 百分位数是一个位置指标，用 P_x 表示。
- P_{50} 实际上就是中位数， P_{25} 实际上就是下四分位数 Q_L ， P_{75} 实际上就是上四分位数 Q_u 。
- $P_{97.5}-P_{2.5}$ 也可以用来反映一组数据的离散程度。

方差 (Variance)

标准差 (Standard Deviation)

- 离均差平方和 (sum of squares of deviations from mean, **SS**) 可用来描述数据的变异度。
- **SS的均数 (即方差)** 不受观察值个数的影响, 用来描述数据的离散程度更好。

- 总体方差:
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- 样本方差:
$$s^2 = \frac{\sum (x_i - \bar{X})^2}{n-1}$$

方差 (Variance)

标准差 (Standard Deviation)

- 因方差的单位是原单位的平方，所以使用不方便。
方差的算术平方根，即标准差，是一个更好的指标。
- 标准差也有总体标准差和样本标准差之分

- 总体标准差:
$$\sigma = \sqrt{\frac{\sum (x_i - \bar{X})^2}{N}}$$

- 样本标准差:
$$s = \sqrt{\frac{\sum (x_i - \bar{X})^2}{n-1}}$$

方差 (Variance)

标准差 (Standard Deviation)

- 适用范围:

- 方差和标准差适合于对称分布，特别是正态分布及近似正态分布资料。

变异系数 (Coefficient of Variation, CV)

- 当比较多组资料的变异度时，如果出现下列情况，可利用变异系数进行变异度的比较：
 - 均数相差悬殊
 - 数据资料的单位不同
- 变异系数实际上是标准差占均数的百分比：

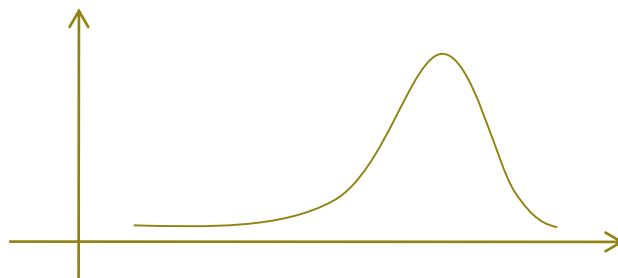
$$CV = S / \bar{X}$$

7. 1. 3 分布特征的描述指标

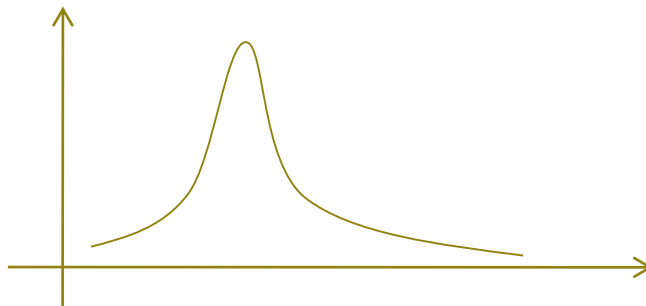
偏度

- **偏度 (Skewness)**：用来描述变量取值分布形态的统计量，指分布不对称的方向和程度。

左偏（负偏）



右偏（正偏）



偏度系数

- 偏度系数: $g_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 / s^3$

g_1 的取值范围: $-3 < g_1 < 3$:

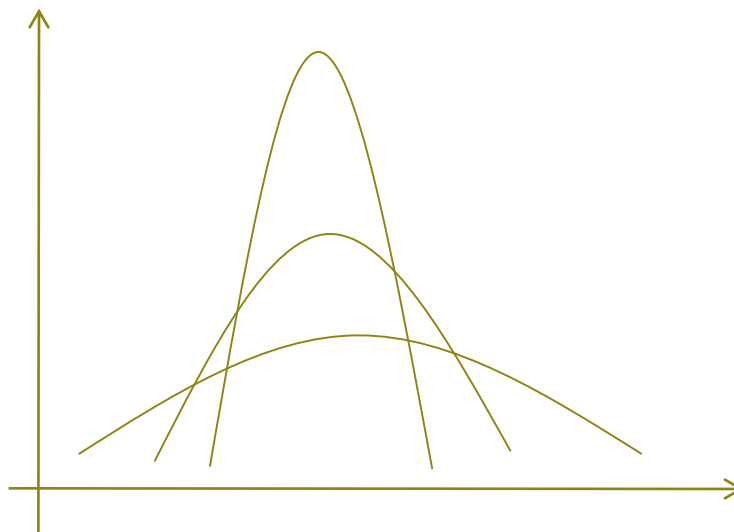
$g_1 < 0$: 左偏 (负偏) ; $g_1 > 0$: 右偏 (正偏)

g_1 越接近0, 分布的偏斜度越小

g_1 越接近 ± 3 , 分布的偏斜度越大

峰度

- **峰度(Kurtosis):** 用来描述变量取值分布形态陡缓程度的统计量，是指分布图形的尖峭程度或峰凸程度。
- 峰度有三种：
 - 尖顶峰
 - 正态峰
 - 平顶峰



峰度系数

- 峰度系数: $g_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 / s^4 - 3$

$g_2 > 0$: 尖顶峰

$g_2 = 0$: 正态峰

$g_2 < 0$: 平顶峰



南京大學
NANJING UNIVERSITY

7. 1. 4 SPSS中的相应功能

Analyze（分析）→Descriptive Statistics （描述统计）

- **Frequencies**过程（频率）：产生原始数据的频数表，计算多种描述指标，绘制相关统计图。
- **Descriptives**过程（描述）：适合连续变量，计算多种描述指标，不能绘图。
- **Explore** 过程（探索）：对连续数据资料分布状况不清楚时进行探索分析，计算多种描述指标，绘制多种统计图等。

Frequencies过程（频率）

- 产生频数表，对连续变量和分类变量资料都适用。
- 计算常用的统计指标、按要求给出某百分位点的数值。
- 常用的条图，饼图等统计图。



Frequencies过程（频率）

频率: 统计



百分位值

☐ 四分位数(Q)

☐ 分割点(U): 10 相等组

☐ 百分位数(P):

添加(A)

更改(C)

删除(R)

集中趋势

☐ 平均值(M)

☐ 中位数(D)

☐ 众数(O)

☐ 合计(S)

☐ 值为组的中点(L)

离散

☐ 标准偏差(T) ☐ 最小值(I)

☐ 方差(V) ☐ 最大值(X)

☐ 范围(N) ☐ 平均值的标准误差(E)

分布

☐ 偏度(W)

☐ 峰度(K)

继续

取消

帮助

频率: 图表



图表类型

☒ 无(O)

☐ 条形图(B)

☐ 饼图(P)

☐ 直方图(H):

☐ 在直方图上显示正态曲线(S)

图表值

☒ 频率(F) ☐ 百分比(C)

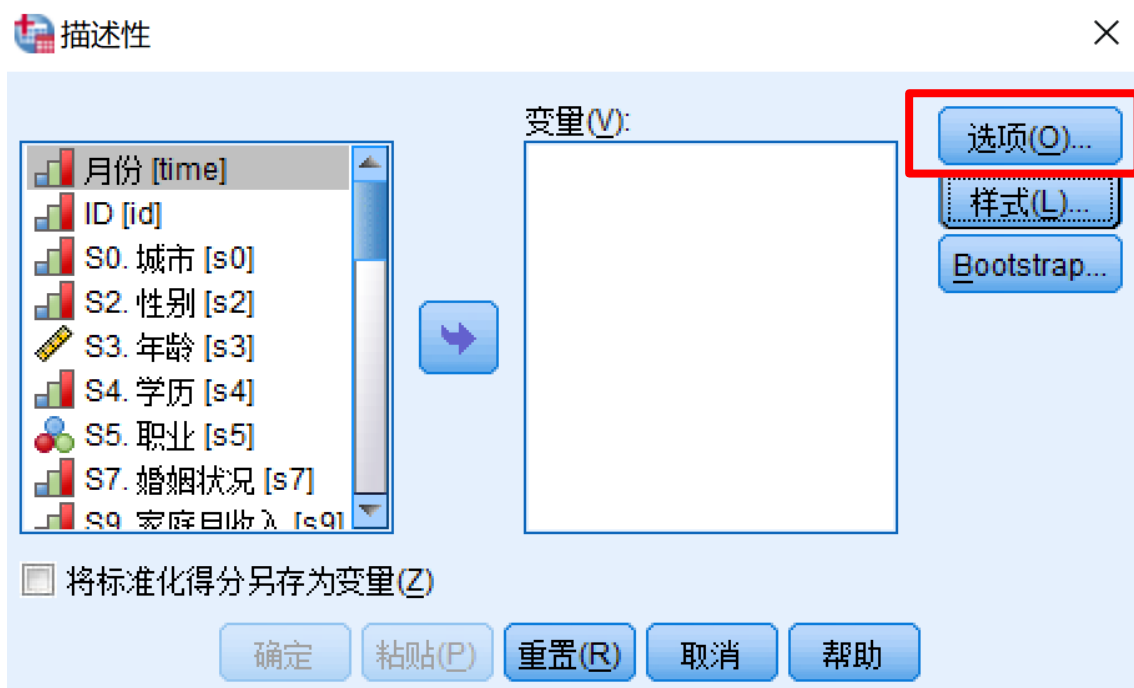
继续

取消


帮助

Descriptive 过程（描述）

- 进行一般性的统计描述，适用于正态分布资料
- 与Frequencies的**区别**：
 - 不能绘制统计图
 - 计算的统计量较少
 - **用于连续变量**



Descriptive 过程（描述）

 描述: 选项 ×

☒ 平均值(M) ☐ 合计(S)

离散

☒ 标准偏差(I) ☒ 最小值(N)
☐ 方差(V) ☒ 最大值(X)
☐ 范围(R) ☐ 平均值的标准误差(E)

分布

☐ 峰度(K) ☐ 偏度(W)

显示顺序

☒ 变量列表(B)
☐ 字母顺序(A)
☐ 按平均值的升序排序(C)
☐ 按平均值的降序排序(D)

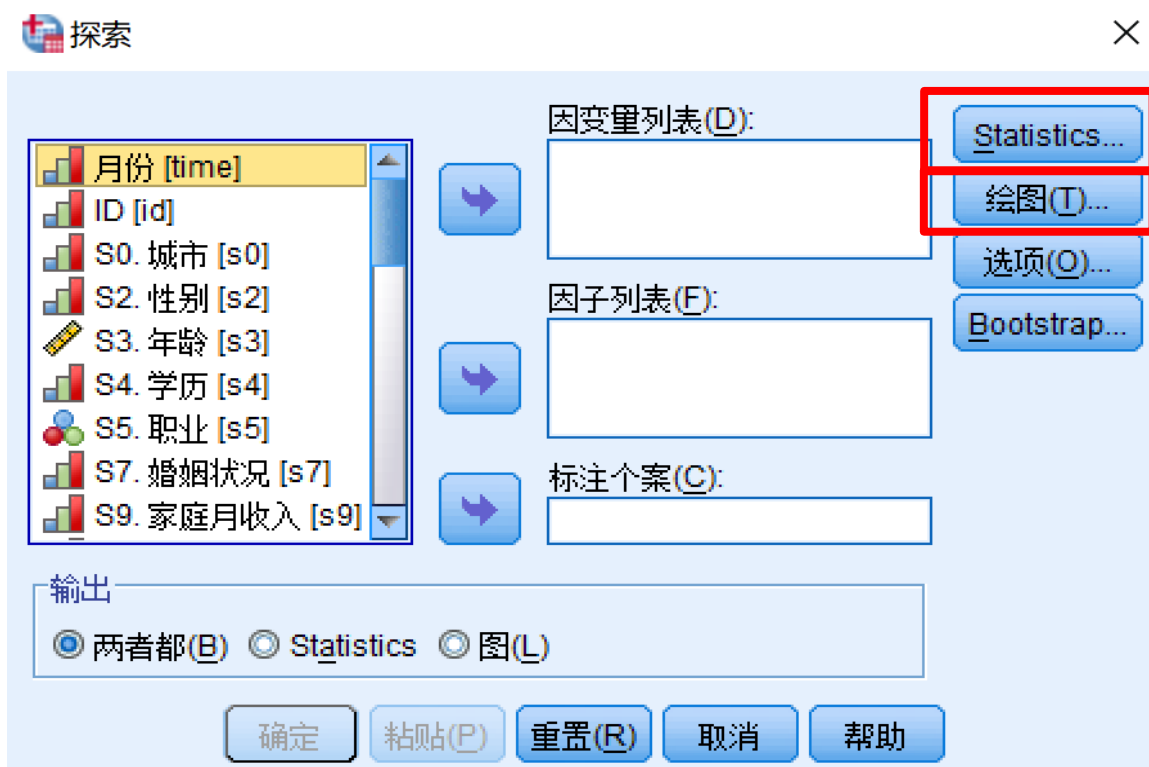
继续

取消

帮助

Explore 过程（探索）

- 若对资料的性质、分布特点等完全不清楚，利用该过程对数据进行初步了解
 - 主对话框：可加入分组变量
 - 计算多个统计量
 - 给出多种统计图
 - 进行简单的参数估计



Explore 过程（探索）

探索: 统计



☒ 描述性(D)

平均值的置信区间(C): 95 %

☐ M-估计量(M)

☐ 界外值(O)

☐ 百分位数(P)

继续

取消

帮助

探索: 图



箱图

☒ 按因子级别分组(F)

☐ 不分组(D)

☐ 无(N)

描述性

☒ 茎叶图(S)

☐ 直方图(H)

☐ 带检验的正态图(O)

伸展与级别 Levene 检验

☒ 无(E)

☐ 幂估计(P)

☒ 已转换(T) 幂(W): 自然对数

☐ 未转换(U)

继续

取消

帮助



南京大學
NANJING UNIVERSITY

7.2 连续变量的参数估计指标体系

总体均值的点估计

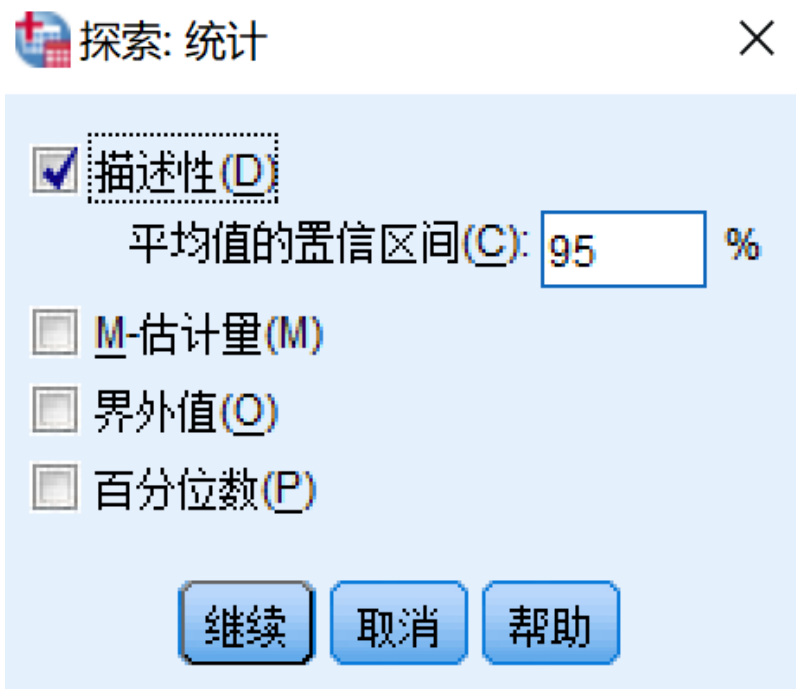
- 总体均值的点估计：选定一个适当的样本统计量值作为总体均值的估计值，如将样本均值作为总体均值的点估计值。

总体均值的区间估计

- **总体均值的区间估计：** 给出一个具有较大可信度 95%（或99%）的包含总体均值的区间, 该区间称为总体均值的**置信度**为95%（或99%）的**置信区间**。

SPSS中的相应功能

- **Explore（探索）** 过程中
 - 可以直接给出总体均值**95%**的置信区间(区间估计)
 - 提供总体均值的**M-估计量**(点估计)





南京大學
NANJING UNIVERSITY

7.3 案例

案例：信心指数的统计描述1

- 教材自带数据：CCSS_Sample.dav
- 要求：对数据文件中的信心总指数index1、现状指数index1a和预期指数index1b进行统计描述，并计算出95%个体参考值范围。
- 步骤：
 - 进入“频率”菜单进行统计描述
 - 取消左下方“显示频率表格”复选框
 - 选择三个变量到“变量”框
 - 点击“统计量”按钮设置统计量

案例：信心指数的统计描述1



案例：信心指数的统计描述1

频率: 统计 ×

百分位值

☒ 四分位数(Q)

☐ 分割点(U): 10 相等组

☒ 百分位数(P):

添加(A) 2.5

更改(C) 97.5

删除(R)

集中趋势

☒ 平均值(M)

☒ 中位数(D)

☐ 众数(O)

☐ 合计(S)

☐ 值为组的中点(L)

离散

☒ 标准偏差(T) ☒ 最小值(I)

☒ 方差(V) ☒ 最大值(X)

☐ 范围(N) ☐ 平均值的标准误差(E)

分布

☐ 偏度(W)

☐ 峰度(K)

继续 取消 帮助

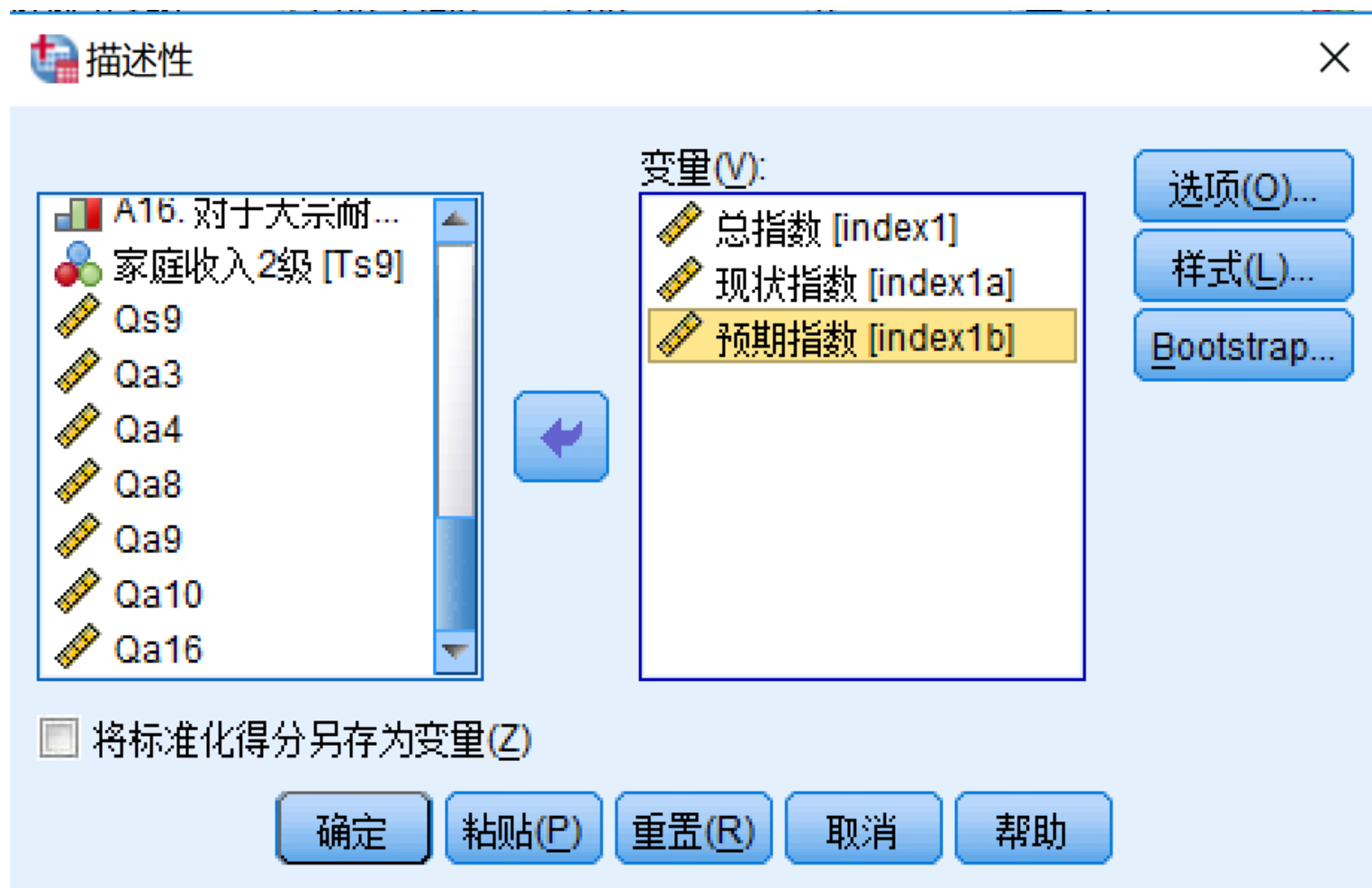
案例：信心指数的统计描述1

統計資料				
		总指数	现状指数	预期指数
N	有效	1147	1147	1147
	遺漏	0	0	0
平均數		95.8935	99.2227	94.0598
中位數		93.7280	88.0359	96.8570
標準偏差		20.99710	28.43333	23.11645
變異數		440.878	808.454	534.370
最小值		.00	.00	.00
最大值		156.21	176.07	145.29
百分位數	2.5	46.8640	44.0180	48.4285
	25	85.9174	88.0359	84.7499
	50	93.7280	88.0359	96.8570
	75	109.3494	110.0449	108.9641
	97.5	132.7814	154.0629	133.1784

案例：信心指数的统计描述2

- 教材自带数据：CCSS_Sample.dav
- 要求：对数据文件中的信心总指数index1、现状指数index1a和预期指数index1b进行统计描述。
- 步骤：
 - 进入“描述”菜单进行统计描述
 - 选择三个变量到“变量”框
 - 点击“选项”按钮进行设置

案例：信心指数的统计描述2



案例：信心指数的统计描述2

描述: 选项

☒ 平均值(M) ☐ 合计(S)

离散

☒ 标准偏差(I) ☒ 最小值(N)

☐ 方差(V) ☒ 最大值(X)

☐ 范围(R) ☐ 平均值的标准误差(E)

分布

☐ 峰度(K) ☐ 偏度(W)

显示顺序

☒ 变量列表(B)

☐ 字母顺序(A)

☐ 按平均值的升序排序(C)

☐ 按平均值的降序排序(D)

继续 取消 帮助

案例：信心指数的统计描述2

描述性統計資料					
	N	最小值	最大值	平均數	標準偏差
总指数	1147	.00	156.21	95.8935	20.99710
现状指数	1147	.00	176.07	99.2227	28.43333
预期指数	1147	.00	145.29	94.0598	23.11645
有效的 N (listwise)	1147				


案例：信心指数的统计描述3

- 教材自带数据：CCSS_Sample.dav
- 要求：分月份对总指数index1进行统计描述，以详细了解其分布情况。
- 步骤：
 - 进入“探索”菜单进行统计描述
 - 设置因变量和因子
 - 点击“统计量”按钮设置统计量

案例：信心指数的统计描述3



案例：信心指数的统计描述3

 探索: 统计 ×

☒ 描述性(D)

平均值的置信区间(C): %

☒ M-估计量(M)

☒ 界外值(O)

☒ 百分位数(P)

继续

取消

帮助

案例：信心指数的统计描述3

描述性統計資料

月份		統計資料	標準錯誤
总指数	200704	平均數	98.3363
		95% 平均數的信賴區間 下限	96.1866
		上限	100.4861
		5% 修整的平均值	98.9930
		中位數	101.5387
		變異數	357.994
		標準偏差	18.92074
		最小值	31.24
		最大值	140.59
		範圍	109.35
		內四分位距	23.43
		偏斜度	-.535
		峰度	.768
			.141
			.281
	200712	平均數	94.1391
		95% 平均數的信賴區間 下限	91.5752
		上限	96.7030
		5% 修整的平均值	95.2468
		中位數	93.7280
		變異數	516.071

案例：信心指数的统计描述3

M-Estimators

月份		Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
总指数	200704	99.6194	100.3020	99.5448	100.3332
	200712	95.7921	96.5184	95.7521	96.5143
	200812	91.0241	91.2941	91.0482	91.2996
	200912	100.3076	100.0637	100.6882	100.0618

- a. The weighting constant is 1.339.
- b. The weighting constant is 4.685.
- c. The weighting constants are 1.700, 3.400, and 8.500
- d. The weighting constant is $1.340 \cdot \pi$.

案例：信心指数的统计描述3

百分位數

月份			百分位數						
			5	10	25	50	75	90	95
加權平均（定義 1）	总指数	200704	62.4854	78.1067	85.9174	101.5387	109.3494	117.1600	124.9707
		200712	54.6747	62.4854	85.9174	93.7280	109.3494	117.1600	124.9707
		200812	54.6747	62.4854	78.1067	93.7280	101.5387	117.1600	117.1600
		200912	78.1067	78.1067	85.9174	101.5387	109.3494	132.7814	140.5920
Tukey 的樞紐	总指数	200704			85.9174	101.5387	109.3494		
		200712			85.9174	93.7280	109.3494		
		200812			78.1067	93.7280	101.5387		
		200912			85.9174	101.5387	109.3494		

案例：信心指数的统计描述3

極端值

月份				個案編號	數值
总指数	200704	最高	1	105	140.59
			2	158	140.59
			3	184	140.59
			4	194	140.59
			5	288	140.59
		最低	1	258	31.24
			2	230	31.24
			3	248	46.86
			4	140	46.86
			5	72	46.86
	200712	最高	1	407	148.40
			2	408	148.40
			3	311	132.78
			4	372	132.78
			5	382	132.78 ^a
		最低	1	397	.00
			2	577	15.62



南京大學
NANJING UNIVERSITY


7.4 Bootstrap方法

功能

- 通过更多的样本来：
 - 判断原参数估计值是否准确（针对参数点估计）
 - 计算出更准确的参数的置信区间（针对参数区间估计）

案例

- 对CCSS_Sample.sav中总指数的均值、标准差进行Bootstrap方法的参数点估计和区间估计。
- 实现过程：
 - “分析”->“描述统计”->“描述”
 - 将“总指数”选入“变量”框
 - 点击“统计量”按钮后选中均值和标准差
 - 点击“Bootstrap”按钮后，如右设置



The image shows the SPSS Bootstrap dialog box. The 'Execute bootstrap' checkbox is checked. The 'Sample size (N)' is set to 1000. The 'Set Mersenne Twister seed' checkbox is unchecked, and the 'Seed (D)' is set to 2000000. Under the 'Confidence interval' section, the 'Level (%) (L)' is set to 95, and the 'Percentile (C)' radio button is selected. Under the 'Sampling' section, the 'Simple (M)' radio button is selected. The 'Variable (V)' list contains '月份 [time]', 'ID [id]', 'S0. 城市 [s0]', 'S2. 性别 [s2]', 'S3. 年龄 [s3]', and 'S4. 学历 [s4]'. The 'Stratification variable (R)' box is empty. At the bottom, there are buttons for 'Continue', 'Cancel', and 'Help'.

Bootstrap

☒ 执行 bootstrap

样本数(N): 1000

☐ 设置 Mersenne Twister 种子

种子(D): 2000000

置信区间

级别(%) (L): 95

☒ 百分位(C)

☐ 偏差修正加速(B)

抽样

☒ 简单(M)

☐ 分层(I)

变量(V):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]

分层变量(R):

继续 取消 帮助

Bootstrap方法的参数估计结果

Descriptive Statistics

		Statistic	Bootstrap ^a			
			Bias	Std. Error	95% Confidence Interval	
					Lower	Upper
总指数	N	1147	0	0	1147	1147
	Minimum	.00				
	Maximum	156.21				
	Mean	95.8935	-.0132	.6197	94.6473	97.0983
	Std. Deviation	20.99710	-.01762	.55390	19.87022	21.96896
Valid N (listwise)	N	1147	0	0	1147	1147

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

THE END