第三章统计数据的预处理

① 异常数据
② 缺失数据

数据预处理

- 把混在原始数据中的“异常数据”排除、把真正有用的“信息”提取出来，有助于推断统计得出正确分析结论。

  1：异常数据取舍

  2：未检出值和/或缺失值估算

- 采用异常数据进行推断统计得到的结论误导带给科研与统计控制判断出错的隐患不可小视。

一、异常数据

- 单个异常值：是指单个样本观测数据组内隐含的个别异常数据。同义词有：可疑值、异常值、极端值、端值、离群值、逸出值、奇异值、超限值、粗值...

- 异常均数：三个以上（k≥3）样本多均数要作统计分析比较时，无疑也要检查其中是否隐含可疑均数。

- 研究者对7例糖尿病患者给某种药物后，测量其血中胰岛素(/ml,X1)和血糖(mg%,X2)

患者编号	1	2	3	4	5	6	7
胰岛素（X1）	24	17	18	12	15	**121**	10
血 糖（X2）	142	170	194	213	214	**238**	249

- 作者采用直线相关分析

$$\gamma = 0.3140, P > 0.05$$

- 结论：血液中胰岛素与血糖两者含量之间无直线相关

Correlations

		胰岛素x1	血 糖x2
胰岛素x1	Pearson Correlation	1	.314
	Sig. (2-tailed)		.493
	N	7	7
血糖x2	Pearson Correlation	.314	1
	Sig. (2-tailed)	.493	
	N	7	7

- 剔出第6对数据前后的Pearson相关系数，前者是0.314，后者是-0.936，显示有相关性！

Correlations

		胰岛素xa	血 糖xb
胰岛素xa	Pearson Correlation	1	-.936**
	Sig. (2-tailed)		.006
	N	6	6
血糖xb	Pearson Correlation	-.936**	1
	Sig. (2-tailed)	.006	
	N	6	6

**. Correlation is significant at the 0.01 level

异常数据的判别法

* 物理判别法：根据人们对客观事物已有的认识，判别由于外界干扰、人为误差等原因造成实测数据偏离正常结果，在实验过程中随时判断，随时剔除

* 统计判别法：给定一个置信概率，并确定一个置信限，凡超过此限的误差，就认为它不属于随机误差范围，将其视为异常数据剔除

* 能用物理判别法判定异常数据有时不易做到，此时只能用统计判别法

统计判别法

- 拉依达准则
- 肖维勒准则
- 格拉布斯准则
- 狄克逊准则
- t检验（罗马诺夫斯基准则）
- 极差法

统计判断对异常数据的区分

- 异常数据有两种情况：
- 1. 异常值不属于该总体，抽样抽错了，从另外一个总体抽出一个(一些)数据，其值与总体平均值相差较大；
- 2. 异常值虽属于该总体，但可能是该总体固有随机变异性的极端表现，比如说超过3σ的数据，出现的概率很小。

- 犯错误1：将本来属于该总体的、出现的概率小的、第二种情况的异常值判断出来舍去，就会犯错误。----去真
- 犯错误2：不属于该总体但数值又和该总体平均值接近的数据被抽样抽出来，统计检验方法判断不出它是异常值，就会犯另外一种错误。----存伪

统计判别法之一：拉依达准则

- 如果实验数据的总体x是服从正态分布的，则 $$p(|x-u|>3\sigma)\le 0.003$$
- 根据上式对于大于μ+3σ或小于μ-3σ的实验数据作为异常数据，予以剔除。
- 剔除后，对余下的各测量值重新计算偏差和标准偏差，并继续审查，直到各个偏差均小于3σ为止。
- 无需查表，使用简便

例： 对某一长度L测量10次，其数据如下：

次数	1	2	3	4	5	6	7	8	9	10
L(cm)	10.35	10.38	10.3	10.32	10.35	10.33	10.37	10.31	10.34	20.33

试用拉依达准则剔除坏值。

解：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{10}(L_i - \overline{L})^2}{10-1}} = 3.16\,cm$$

$$3\sigma = 3.16 \times 3 = 9.48cm$$

$$\Delta L_{10} = L_i - \overline{L}$$

$$= \mathbf{20.33 - 11.34}$$

$$= 8.99 < 3\sigma = 9.48$$

20.33不能用拉依达准则剔除

例： 对某一长度L测量10次，其数据如下：

次数	1	2	3	4	5	6	7	8	9	10	11
L(cm)	10.35	10.38	10.3	10.32	10.35	10.33	10.37	10.31	10.34	20.33	10.37

试用拉依达准则剔除坏值。

解：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{11}(L_i - \overline{L})^2}{11-1}} = 3.01cm$$

$$3\sigma = 3.01 \times 3 = 9.03cm$$

$$\Delta L_{10} = L_i - \overline{L}$$

$$= 20.33 - 11.25$$

$$= 9.08 > 3\sigma = 9.03$$

20.33用拉依达准则剔除

- 对于服从正态分布的测量结果，其偏差出现在±3σ附近的概率已经很小，如果测量次数不多，偏差超过±3σ几乎不可能，因而，用拉依达判据剔除疏失误差时，往往有些疏失误差剔除不掉。
- 另外，仅仅根据少量的测量值来计算σ，这本身就存在不小的误差。
- 因此拉依达准则不能检验样本量较小的情况。（显著性水平为0.1时，n必须大于10）

统计判别法之二：肖维勒准则

- 肖维勒准则又称为等概率原则，以正态分布为前提, 假设多次重复测量所得n个测量值中, 某个测量值的残余误差$|v_i| = |x_n - \bar{x}| > Z_c \sigma$,则剔除此数据。
- 实用中Zc<3, 所以在一定程度上弥补了3σ准则的不足，另外考虑了测量次数的因素，在一定程度上比拉依达准则更合理。
- Zc是一个与测量次数相关的系数，可以查表获取。
- 肖维勒准则可用于n<10时粗大误差的判定。

Zc系数表

n	Zc	n	Zc	n	Zc
3	1.38	11	2.00	25	2.33
4	1.54	12	2.03	30	2.39
5	1.65	13	2.07	40	2.49
6	1.73	14	2.10	50	2.58
7	1.80	15	2.13	100	2.80
8	1.86	16	2.15		
9	1.92	18	2.20		
10	1.96	20	2.24		

统计判别法之三：格拉布斯准则

- 格拉布斯准则是在未知总体标准差情况下，对正态样本或接近正态样本异常值的一种判别方法。
- 某个测量值的残余误差$|v_i| = |x_n - \bar{x}| > T\sigma$，则判断此值中含有粗大误差, 应予剔除。
- T值与重复测量次数n和置信概率α均有关，因此格拉布斯准则是比较好的判定准则。
- 格拉布斯准则理论较严密，概率意义明确，可用于严格要求的场合,当n=20-100时，判别效果较好。
- T值通过查表获得。

$T_0(n, \alpha)$值表

σ \ n	3	4	5	6	7	8	9	10
0.05	1.15	1.46	1.67	1.82	1.94	2.03	2.11	2.18
0.025	1.15	1.48	1.71	1.89	2.02	2.13	2.21	2.29
0.01	1.15	1.49	1.75	1.94	2.10	2.22	2.32	2.41

σ \ n	11	12	13	14	15	16	17	18
0.05	2.23	2.29	2.33	2.37	2.41	2.44	2.47	2.50
0.025	2.36	2.41	2.46	2.51	2.55	2.59	2.62	2.65
0.01	2.48	2.55	2.61	2.66	2.71	.275	2.79	2.82

σ \ n	19	20	21	22	23	24	25	30
0.05	2.53	2.56	2.58	2.60	2.62	2.64	2.66	2.75
0.025	2.68	2.71	2.73	2.76	2.78	2.80	2.82	2.91
0.01	2.85	2.88	2.91	2.94	2.96	2.99	3.01	3.10

σ \ n	35	40	45	50	60	70	80	100
0.05	2.82	2.87	2.92	2.96	3.03	3.09	3.14	3.21
0.025	2.98	3.04	3.09	3.13	3.20	3.26	3.31	3.38
0.01	3.18	3.24	3.29	3.34				

* 采用格拉布斯方法判定异常数据的过程如下：
* 1. 选定危险率 α
* α 是一个较小的百分数，例如1%，2.5%，5%，它是采用格拉布斯方法判定异常数据出现误判的几率。
* 2. 计算T值
* 如果$x_{(n)}$是可疑数据，则令
*

$$T = \frac{x_{(n)} - \bar{x}}{\sigma}$$

- 3. 根据n及α，查表得到$T_0(n, \alpha)$值
- 4. 如果$T \geqslant T_0(n, \alpha)$,则所怀疑的数据是异常数据，应予剔除。如果$T_0(n, \alpha)$，则所怀疑的数据不是异常数据，不能剔除。
- 5. 余下数据重复操作至无异常数据
- 格拉布斯准则可以检验较少的数据

狄克逊准则

- 亦称Q检验法，狄克逊准则是通过极差比判定和剔除异常数据。
- 该准则认为异常数据应该是最大数据和最小数据，因此该其基本方法是将数据按大小排队，检验最大数据和最小数据是否异常数据。

- 将实验数据x_i按值的大小排成顺序统计量
- $x(1) \leqslant x(2) \leqslant x(3), \ldots\ldots \leqslant x(n)$
- 计算f_0值
- $$f_0 = \frac{x_n - x_{n-1}}{x_n - x_1} \quad 或 \quad \frac{x_2 - x_1}{x_n - x_1}$$
- 根据狄克逊系数表将f_0与$f(n, \alpha)$进行比较
- 如果$f_0 > f(n, \alpha)$，说明$x(n)$*离群远*，则判定该数据为异常数据，予以剔除。

狄克逊系数$f(n,a)$与f_0的计算公式

n	$f(n,a)$		f_0的计算公式	
	$a=0.01$	$a=0.05$	$x_{(1)}$可以时	$x_{(n)}$可以时
3	0.988	0.941	$\frac{x_{(2)}-x_{(1)}}{x_{(n)}-x_{(1)}}$	$\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(2)}}$
4	0.889	0.765		
5	0.780	0.642		
6	0.698	0.560		
7	0.637	0.507		
8	0.683	0.554	$\frac{x_{(2)}-x_{(1)}}{x_{(n-1)}-x_{(1)}}$	$\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(2)}}$
9	0.635	0.512		
10	0.597	0.477		
11	0.679	0.576	$\frac{x_{(3)}-x_{(1)}}{x_{(n-1)}-x_{(1)}}$	$\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-x_{(3)}}$
12	0.642	0.546		
13	0.615	0.521		
14	0.641	0.546	$\frac{x_{(3)}-x_{(1)}}{x_{(n-1)}-x_{(1)}}$	$\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-x_{(3)}}$
15	0.616	0.525		
16	0.595	0.507		
17	0.577	0.490		
18	0.561	0.475		
19	0.547	0.462		
20	0.535	0.450		
21	0.524	0.440		
22	0.514	0.430		
23	0.505	0.421		
24	0.497	0.413		
25	0.489	0.406		

t检验准则（罗马诺夫斯基准则）

t检验准则与狄克逊准则相似，也是检验最大实验数据和最小实验数据。首先将实验数据按大小排列

$x(1) \leqslant x(2) \leqslant x(3), \ldots\ldots \leqslant x(n)$

对最小数据和最大数据分别进行检验，如果

$$\left|x_{(1)} - \bar{x}^*\right| > K(n,\alpha)\sigma^* \qquad 或 \left|x_{(n)} - \bar{x}^*\right| > K(n,\alpha)\sigma^*$$

则x(1)或x(n)是异常数据，应予剔除

式中 $\bar{x}^*$ 及 σ^* 分别为不包括$x_{(1)}$或$x_{(n)}$的n-1个数据的均值和标准差。

t检验中的$K(n, \alpha)$

n \ α	0.01	0.05	n \ α	0.01	0.05	n \ α	0.01	0.05
4	11.40	4.97	13	3.23	2.29	22	2.91	2.14
5	6.53	3.04	14	3.17	2.26	23	2.90	2.13
6	5.04	3.04	15	3.12	2.24	24	2.88	2.12
7	4.36	2.78	16	3.08	2.22	25	2.86	2.11
8	3.96	2.62	17	3.04	2.20	26	2.85	2.10
9	3.71	2.51	18	3.01	2.18	27	2.84	2.10
10	3.54	2.43	19	3.00	2.17	28	2.83	2.09
11	3.41	2.37	20	2.95	2.16	29	2.82	2.09
12	3.31	2.33	21	2.93	2.15	30	2.81	2.08

应注意的问题：

- ①　所有的检验法都是人为主观拟定的，至今无统一的规定。以数据按正态分布为前提的，当偏离正态分布和测量次数少时检验不一定可靠。
- ② 若有多个可疑数据同时超过检验所定置信区间，应逐个剔除，重新计算，再行判别。若有两个相同数据超出范围时，应逐个剔除。
- ③在一组测量数据中，可疑数据应很少。反之，说明系统工作不正常。
- ④为了减少犯错误的概率，可以将3种以上统计检验法结合使用，根据多数方法的判断结果，确定可疑值是否为异常值

- 拉依达准则不能检验样本量较小的情况，格拉布斯准则则可以检验较少的数据。在国际上，常推荐格拉布斯准则和狄克逊准则。
- 但对于异常数据一定要慎重，不能任意的抛弃和修改。往往通过对异常数据的观察，可以发现引起系统误差的原因，进而改进过程和试验。

SPSS实现

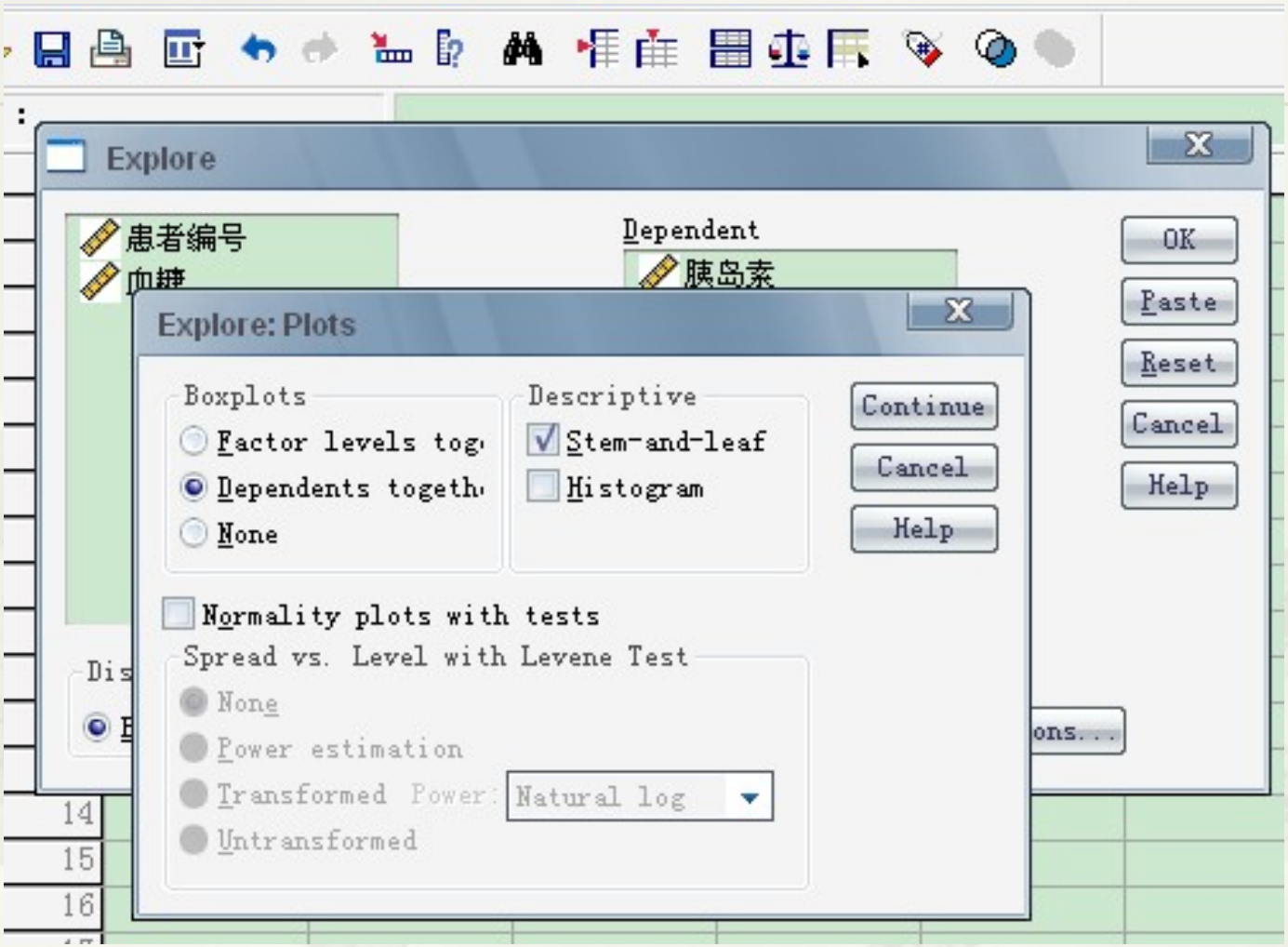
- 研究者对7例糖尿病患者给某种药物后，测量其血中胰岛素(/ml,X1)和血糖(mg%,X2)

患者编号	1	2	3	4	5	6	7
胰岛素（X1）	24	17	18	12	15	**121**	10
血 糖（X2）	142	170	194	213	214	**238**	249

- 作者采用直线相关分析 $r = 0.314, P > 0.05$
- 结论：血液中胰岛素与血糖两者含量之间无直线相关

SPSS实现

- 本例为小样本，单击Analyze，后单击Descriptive statistics选择 [Explore]主对话框中，再单击[Plots…]选项→进入[Explore：Plots]对话框：在Boxplots项下点选⊙Dependents Together，在Descriptive项下勾选☑Stem-and-leaf，其余各项可以不要勾选和点选；单击[Continue]返回[Explore]对话框，单击OK, SPSS 运行、输出结果


Explore
患者编号
血糖
Dependent
胰岛素
OK
Paste
Reset
Cancel
Help
Explore: Plots
Boxplots
Factor levels tog
Dependents togeth
None
Descriptive
Stem-and-leaf
Histogram
Continue
Cancel
Help
Normality plots with tests
Spread vs. Level with Levene Test
None
Power estimation
Transformed Power:
Natural log
Untransformed
14
15
16

Frequency Stem & Leaf

2.00 1 . 0

3.00 1 . 78

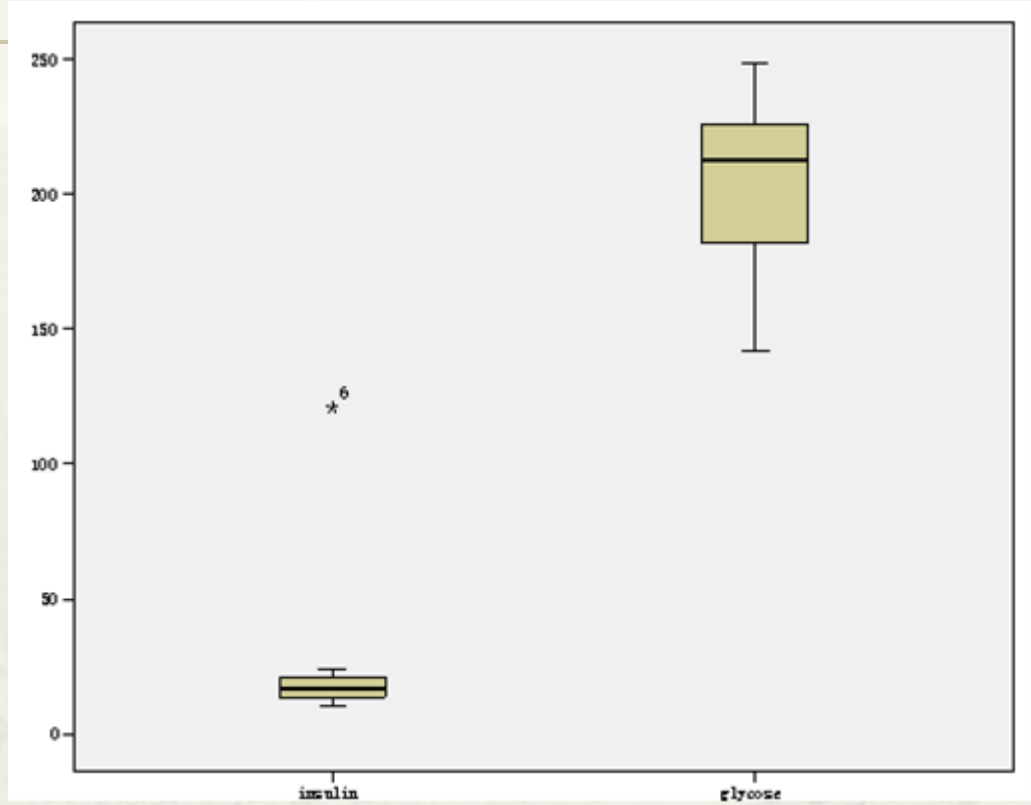
1.00 2 . 4

1.00 **Extremes (>=121)**

Stem width: 10

Each leaf: 1 case(s)

胰岛素检出离群值 **121**



* 叶茎图和箱须图提示有极端值（≥121）

二、缺失数据的处理

缺失数据

在实践工作中，常会因为某些原因导致数据缺失，只能观测到一部分数据，统计学中一般称为缺失数据

原因：

- 信息暂时无法获取
- 信息是被遗漏的
- 某个或某些属性是不可用的
- 某些信息（被认为）是不重要的
- 获取这些信息的代价太大
- 系统实时性能要求较高，即要求得到这些信息前迅速做出判断或决策

数据缺失的机制

- 将数据集中不含缺失值的变量（属性）称为完全变量，数据集中含有缺失值的变量称为不完全变量，Little 和 Rubin定义了以下三种不同的数据缺失机制：
- 1）完全随机缺失：数据的缺失与不完全变量以及完全变量都是无关的。
- 2）随机缺失：数据的缺失仅仅依赖于完全变量。
- 3）非随机、不可忽略缺失：不完全变量中数据的缺失依赖于不完全变量本身，这种缺失是不可忽略的。

缺失数据预处理思想

1. 保留缺失数据不予处理：不对缺失数据做任何处理
2. 直接丢弃含缺失数据的记录，也就是将存在遗漏信息属性值的对象(元组、记录)删除，从而得到一个完备的信息表.
3. 特殊值填充：将缺值作为一种特殊的属性值来处理，它不同于其他的任何属性值．如所有的缺值都用“unknown”填充，这样将可能导致严重的数据偏离，不推荐！

4.可能值插补缺失值

A. 用平均值来代替所有缺失数据

B. K -最近距离邻居法：先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的K个样本，将这K个值加权平均来估计该样本的缺失数据。

C.用回归、贝叶斯形式化方法或判定树归纳确定，这些方法直接处理的是模型参数的估计而不是空缺值预测本身.

与前面的方法相比，它使用现存数据的多数信息来推测空缺值.

个案剔除法(Listwise Deletion)

- 最常见、最简单的处理缺失数据的方法，也是很多统计软件（如SPSS）默认的缺失值处理方法。
- 如果缺失值所占比例比较小，这一方法十分有效。至于具体多大的缺失比例算是“小”比例，专家们意见也存在较大的差距。有学者认为应在5%以下，也有学者认为20%以下即可。
- 这种方法却有很大的局限性。它是以减少样本量来换取信息的完备，会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息。当缺失数据所占比例较大，特别是当缺数据非随机分布时，这种方法可能导致数据发生偏离，从而得出错误的结论。

单一插补

- 单一插补是以估算为基础的方法，是在缺失数据被替代后，对新合成的数据进行相应的统计分析。

  1：均值插补

  2：热卡填充发法

  3：回归插补

  4：回归随机插补

均值插补(Mean Imputation)

- 缺失值是数值型的：平均值来填充该缺失的变量值
- 缺失值是非数值型的，众数来补齐该缺失的变量值。
- 均值替换法也是一种简便、快速的缺失数据处理方法。使用均值替换法插补缺失数据，对该变量的均值估计不会产生影响。但这种方法是建立在完全随机缺失（MCAR）的假设之上的，而且会造成变量的方差和标准差变小。

热卡填充法（Hotdecking）

- 在数据库中找到一个与最相似的对象，然后用这个相似对象的值来进行填充。
- 不同的问题可能会选用不同的标准来对相似进行判定。
- 变量Y与变量X相似，把所有个案按Y的取值大小进行排序。那么变量X的缺失值就可以用排在缺失值前的那个个案的数据来代替了。
- 与均值替换法相比，利用热卡填充法插补数据后，其变量的标准差与插补前比较接近。但在回归方程中，使用热卡填充法容易使得回归方程的误差增大，参数估计变得不稳定，而且这种方法使用不便，比较耗时。

回归插补(Regression Imputation)

- 回归插补首先需要选择若干个预测缺失值的自变量，然后建立回归方程估计缺失值，即用缺失数据的条件期望值对缺失值进行替换。
- 该方法也有诸多弊端，第一，容易忽视随机误差，低估标准差和其他未知性质的测量值，而且这一问题会随着缺失信息的增多而变得更加严重。第二，研究者必须假设存在缺失值所在的变量与其他变量存在线性关系，很多时候这种关系是不存在的。

随机回归插补

- 该方法就是在回归插补值的基础上再加上残差项。
- 残差项的分布可以包括正态分布，也可以是其他的非正态分布。

单一插补法优缺点

- 单一插补法改变了传统方法将缺失值忽略不考虑的习惯，使得各种统计分析均可以在插补后的完整数据集上展开。
- 但单一插补法的缺点也是显而易见的：
  无论采用何种方法，都存在扭曲样本分布的问题(如均值插补会降低变量之间的相关关系，回归插补则会人为地加大变量之间的相关关系)，尽管由于随机回归插补引入随机误差项，能够缓解这一问题，但是随机误差项的确定是比较困难的。

（五）多重插补方法(Multiple Imputation)

* 多重插补建立在贝叶斯理论基础之上，基于EM算法(最大期望算法)来实现对缺失数据的处理。
* 分为三个步骤：

①为每个空值产生一套可能的插补值，这些值反映了无响应模型的不确定性；每个值都可以被用来插补数据集中的缺失值，产生若干个完整数据集合。

②每个插补数据集合都用针对完整数据集的统计方法进行统计分析。

③对来自各个插补数据集的结果，根据评分函数进行选择，产生最终的插补值。

多重插补法的出现，弥补了单一插补法的缺陷。

- 第一，多重插补过程产生多个中间插补值，可以利用插补值之间的变异反映无回答的不确定性，包括无回答原因已知情况下抽样的变异性和无回答原因不确定造成的变异性。
- 第二，多重插补通过模拟缺失数据的分布，较好地保持变量之间的关系。
- 第三，多重插补能给出衡量估计结果不确定性的大量信息，单一插补给出的估计结果则较为简单。

多重插补和贝叶斯估计的思想是一致的，但是多重插补弥补了贝叶斯估计的几个不足

(1)贝叶斯估计以极大似然的方法估计，极大似然的方法要求模型的形式必须准确，如果参数形式不正确，将得到错误结论，即先验分布将影响后验分布的准确性。而多重插补所依据的是大样本渐近完整的数据的理论，在数据挖掘中的数据量都很大，先验分布将极小的影响结果，所以先验分布对结果的影响不大。

(2)贝叶斯估计仅要求知道未知参数的先验分布，没有利用与参数的关系。而多重插补对参数的联合分布作出了估计，利用了参数间的相互关系。

SPSS实现

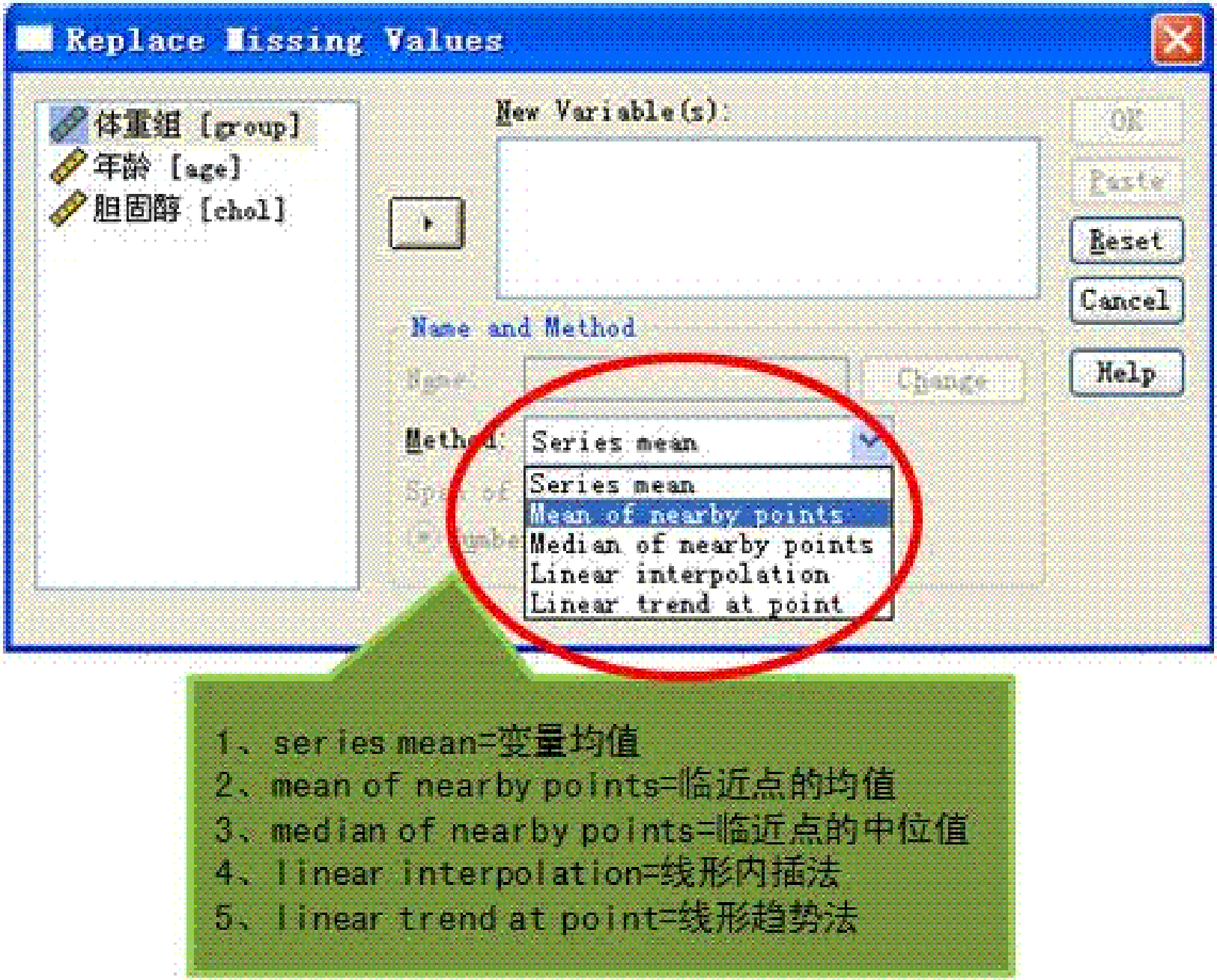
- 1、listwise deletion法

  在SPSS 的统计分析程序中, 打开options按钮, 便会出现缺失值的处理栏(missing values) , 可分别选择下列选项: exclude cases analysis by analysis (剔除正在分析的变量中带缺失值的观察单位) ; exclude case list wise (剔除所有分析变量中带缺失值的
  观察单位)


Linear Regression
体重组 [group]
年龄 [age]
Dependent:
胆固醇 [chol]
Block 1 of 1
Previous
Next
Independent(s):
体重组 [group]
年龄 [age]
Method:
Enter
Selection Variable:
Case Labels:
WLS Weight:
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Save...
Options...
Linear Regression: Options
Stepping Method Criteria
Use probability of F
Entry: .05
Removal: .10
Use F value
Entry: 3.84
Removal: 2.71
Include constant in equation
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Continue
Cancel
Help
此处就是回归分析中
“option”子对话框

SPSS实现

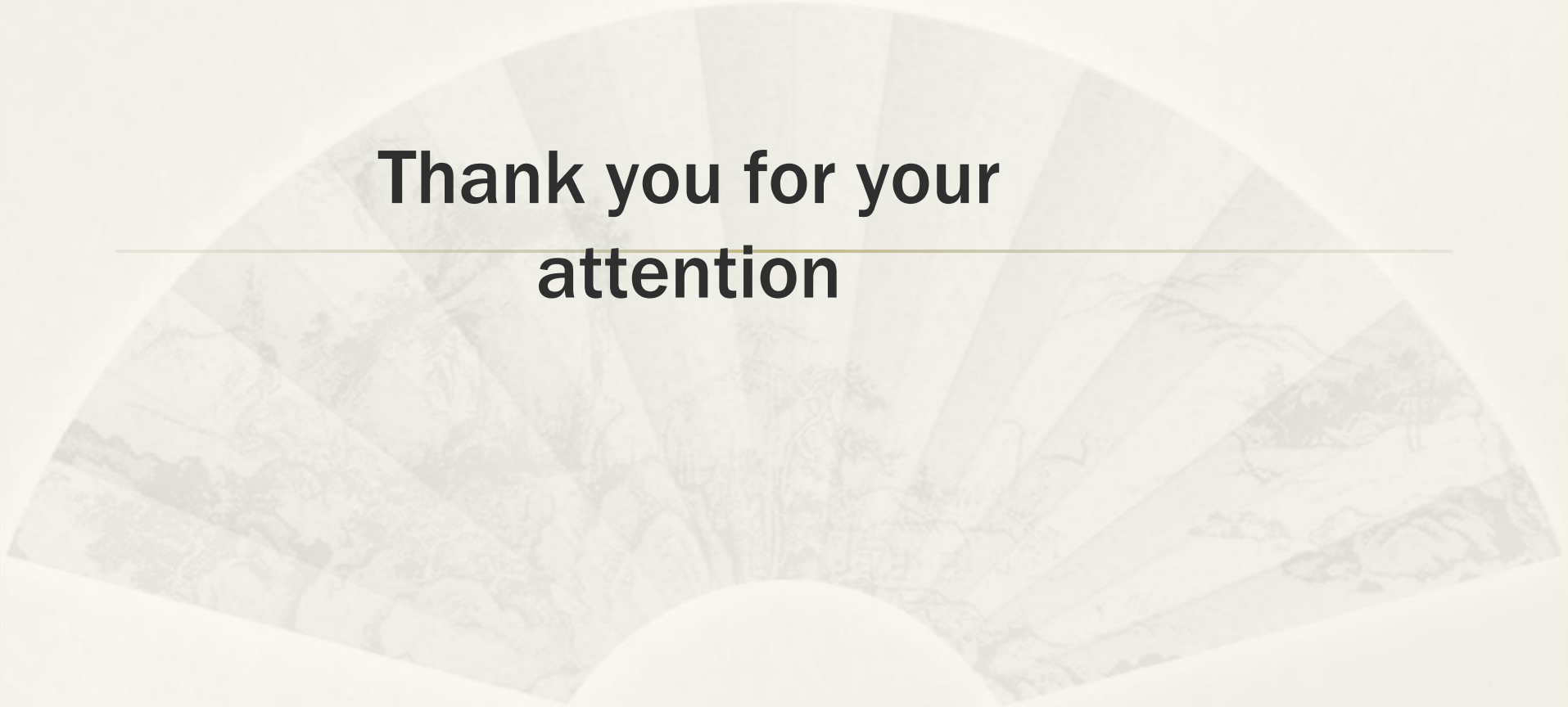
- 2、如果遇到的缺失值形式是完全随机变量,在样本容量不大的情况下,可采用填补的方法(imputation)。
- 点击“transform，此菜单下的“replace missing values”列出了5 种替代的方法
- 通常可填上平均值,或者回归的预测值,这两种方法都有缺点，对最终数据结果影响较大


Replace Missing Values
体重组 [group]
年龄 [age]
胆固醇 [chol]
New Variable(s):
OK
Paste
Reset
Cancel
Help
Name and Method
Name:
Change
Method:
Series mean
Series mean
Mean of nearby points
Median of nearby points
Linear interpolation
Linear trend at point
1、series mean=变量均值
2、mean of nearby points=临近点的均值
3、median of nearby points=临近点的中位值
4、linear interpolation=线形内插法
5、linear trend at point=线形趋势法

- SPSS有个Missing data analysis栏目,增加了EM (expectation and maximization)填补。
- 它的方法是把有同样缺失的样本放在同一组,计算它的协方差矩阵（covariance matrix）,然后再根据每组的样本数来校正它对整个样本的权重(weight),从这里再重新填补每个缺失值，这重方法算是现在比较精确的缺失值填补的方法。


选择进入缺失值分析的变量
Missing Value Analysis
体重组 [group]
年龄 [age]
Quantitative
胆固醇 [chol]
Patterns...
Descriptives...
Estimation
Listwise
Pairwise
EM
Regression
Variables...
EM...
Regression...
选择分类变量
Categorical
missing
Maximum
25
Case Labels:
Use All Variables
OK
Paste
Reset
Cancel
Help
Listwise 框：所选择的任意一个应变量或分组变量中带有缺失值的记录将都不尽如分析
Pairwise 框：在具体计算时用刀的变量具有缺失值的记录将不进入当前分析
EM框：使用EM（期望最大化）迭代方法估计缺失值，推荐
Regression：使用多元（多重）线形回归算法来估计缺失值

总之，缺失值处理方法的选用取决于缺失值的形式、缺失样本总样本的比例等具体情况而定，最终的衡量标准要保证最终数据的客观性与准确性。



Thank you for your attention