建模培训：统计建模过程

汪四水

苏州大学数学科学学院

2016年7月14日

导航

什么是统计

统计是指对某一现象有关的数据进行搜集、整理、计算和分析等一系列活动。在实际应用中，常有以下**三种涵义**：统计工作、统计资料和统计学。

- 统计工作(statistical work)：对统计资料的搜集、整理、分析和提供数量资料的工作总称。
- 统计资料或数据(statistical data)：是统计工作的成果，用来反映总体现象的数据资料的总称。
- 统计学(statistics)：搜集、整理和分析统计数据资料的理论与方法的科学。

什么是统计

什么是**数理统计学**/统计学

- 研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议。

什么是统计——总体、样本、统计量、统计推断

总体

- 研究对象的全体（集合）
- 一元总体，多元总体，无限维总体（随机过程，时间序列）

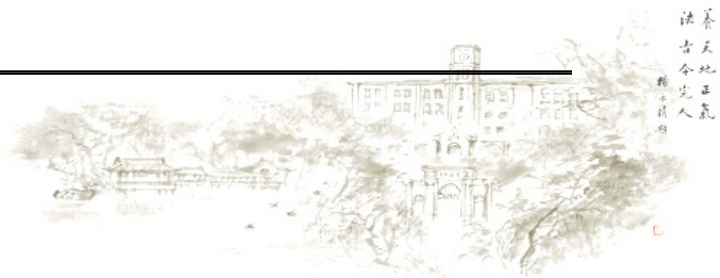
数据的类型

名义型、顺序型（次序型）、区间型、比率型

什么是统计——总体

总体与统计方法简介

	数据类型 名义型　　顺序型　　区间型　　比率型
一元	描述性统计： 平均位置（平均值，中位数，众数） 离散程度（方差，标准差，极差，四分位极差，变异系数） 分布形状（偏度，峰度）
多元	条件，边缘分布，联合分布，独立；回归模型，判别分析，聚类分析，主成分分析，因子分析，典型相关，对应分析，结构方程模型……
随机过程/时间序列	泊松过程，更新过程，马尔可夫过程，维纳过程，时间序列等



什么是统计——样本

数据获取的方法

- 抽样调查：如市场调查等
- 试验设计：如质量工程，农田试验等
- 记录数据：如股票，地震数据等

数据（样本）的符号表示： $x_1, x_2, \ldots, x_n$

什么是统计——统计量

数据中信息提取：统计量

$$T = T(x_1, x_2, ...x_n)$$

常用基本统计量

- 样本均值
- 样本方差
- 偏度，峰度
- 相关系数

什么是统计——统计推断（参数估计，假设检验等）

参数的估计

- 点估计：方法有：矩法，极大似然法，最小二乘法，贝叶斯法等
- 区间估计：方法有：枢轴量法，大样本法，自助法（bootstrap）
- 估计量的评判标准：无偏性，渐近无偏性，相合性，均方差

什么是统计——统计推断（参数估计，假设检验等）

统计假设检验

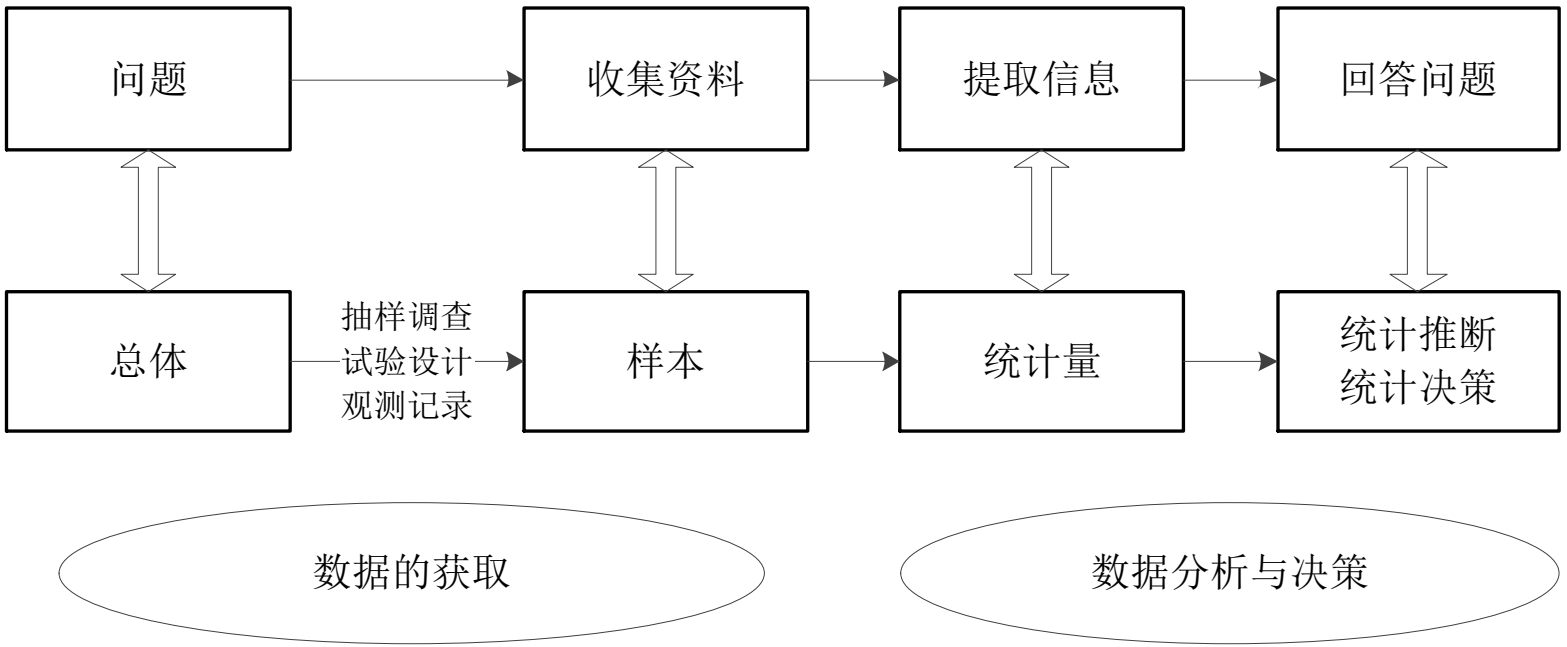
- 步骤
  1) 提出原假设H_0，备择假设H_1
  2) 给定显著性水平α（常取0.05，0.01）
  3) 找到检验统计量T
  4) 判断（小概率事件原理）

什么是统计——统计计算的工具

统计计算的工具

- 统计软件：SAS,SPSS,SPLUS,R...
  办公软件：EXCEL
- 符号软件：mathematica maple
- 数值软件：matlab,scilab
- 编程语言：C,C++,Basic

统计建模流程


问题
收集资料
提取信息
回答问题
总体
抽样调查
试验设计
观测记录
样本
统计量
统计推断
统计决策
数据的获取
数据分析与决策

统计建模流程——step1：理论建模

step1：理论建模（问题的形成及建模）

这部分事实上与数据无关，主要是用统计、概率、数学语言去描述问题，然后形成统计模型去表达该问题（主要考虑到随机性）

统计建模流程——step2：收集数据

step2：收集数据（抽样调查，试验设计，观测数据）

注意事项

- 数据是观测到的，还是实验得到的？
- 如何收集有代表性的数据？
- 有没有没回答的问题？（抽样调查中常出现）
- 有没有缺失值？
- 分类数据还是连续数据？
- 数据是如何编码的？
- 数据测量的单位（量纲）？
- 有没有异常数据？

统计建模流程——step3:基于数据统计建模

step3：统计建模：确定总体的概率分布，常常包括：

- 系统误差部分（非随机部分）：如非参数成分，参数成分
- 随机误差部分：随机成分。
- 重点考虑系统误差部分：确定性部分与协变量的关系

方法（二者常常结合使用）：

- 图形法（直观但不精确）：统计图形；
- 数值法（精确但不直观）：统计量法，如回归分析，贝叶斯分析，变量的选择...，等等；

统计建模流程——step4：统计推断

step4：推断或统计决策

用基于数据所得到的统计模型，去验证理论分析所得到的模型，看是否一致，以及是否有差别？并解释为什么？

统计建模流程——例子（问题、总体）

问题

一家物流公司，在全国各地有很多站点，如何得知这些站点间的距离？(站点很多，一个一个去测很费时，有没有其他办法知道各站点间的距离？)

总体：为简化问题，譬如苏州市内物流公司站点的距离

理论建模：考虑站点间的距离与直线距离间的关系，
设x表示站点间的直线距离，
y表示站点间的实际距离。
如何获得二者间的关系？

统计建模流程——例子（问题、总体）

x, y间有什么潜在要满足的关系吗？

i) $x = 0 \Rightarrow y = 0$

ii) 若两站点间本身就是直线关系，则有$x = y$，否则$y \geqslant x$

iii) 一般来说，y应随x的增加而增加，但由于路况的不同，即使有相同的x，也有可能y值是不同的。

iv) 期望x, y成比例增加，即x扩大一倍，y也应该扩大一倍

统计建模流程——例子（问题、总体）

理论建模

考虑如下模型：

1. $y = x$（满足i,iv，但不满足ii,iii）
2. $y = x + \varepsilon$，ε为随机项（不满足ii）
3. $y = \alpha + x + \varepsilon$（ii满足了，但i不满足）
4. $y = \beta x + \varepsilon$，$\beta \geqslant 1$为常数，可满足所有要求。

注意：以上建模过程并不需要任何数据——即理论建模

统计建模流程——例子（收集数据、样本）

数据的收集

1. **已有的数据**：考虑数据是观测的，还是通过实验设计获得的？
2. **若是要设计获取数据**：
   - 若有很多站点，如何选择一个小范围的站点（抽样问题）。
   - 若连接两站点的路线有多条，重复是需要的。
   - 测量y值的人如何分配？（随机化，区组化）

例如：

x	9.5	5	23	15.2	11.4	11.8	12.1	22	28.2	12.1
y	10.7	6.5	29.4	17.2	18.4	19.7	16.6	29	40.5	14.2

x	9.8	19	14.6	8.3	21.6	26.5	4.8	21.7	18	28
y	11.7	25.6	16.3	9.5	28.8	31.2	6.5	25.7	26.5	33.1

统计建模流程——例子（提取信息、统计分析）

提取信息：基于数据的统计分析（统计模型+数据分析）

用数据得出的模型（经验模型）

1. 根据数据，可建立什么统计模型？（图形法，数值法）
2. 统计模型与概念模型一致吗？

统计建模流程——例子（提取信息、统计分析）

SAS程序

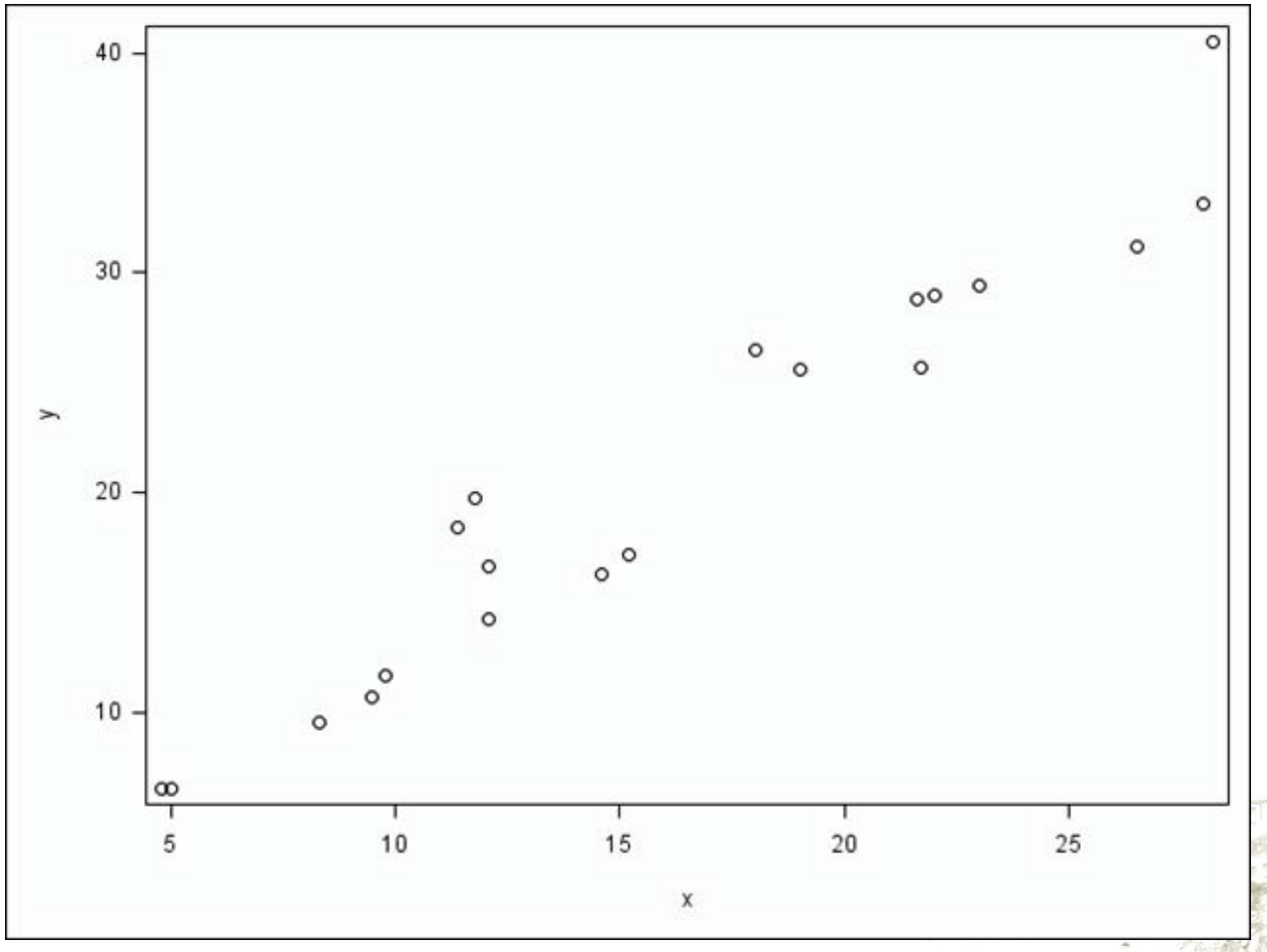
```
data dt;
input x y;
datalines;
9.5 10.7
5 6.5
23 29.4
15.2 17.2
11.4 18.4
11.8 19.7
12.1 16.6
22 29
28.2 40.5
12.1 14.2
9.8 11.7
19 25.6
14.6 16.3
8.3 9.5
21.6 28.8
26.5 31.2
4.8 6.5
21.7 25.7
18 26.5
28 33.1
run;
```

图形法：

```
proc sgplot data=dt;
scatter x=x y=y;
run;
```

统计建模流程——例子（提取信息、统计分析）

图形：



统计建模流程——例子（提取信息、统计分析）

数值法：

```
1 proc reg data=dt;
2 model y=x;
3 run;
```

回归系数显著，截距项不显著（可去掉）

表：方差分析

源	自由度	平方和	均方	F 值	$Pr > F$
模型	1	1648.26305	1646.26305	277.73	$< .0001$
误差	18	106.82645	5.93480		
校正合计	19	1755.08950			

表：参数估计

变量	自由度	参数估计	标准误差	t值	$Pr > \|t\|$
Intercept	1	0.37908	1.34401	0.28	0.7811
x	1	1.26943	0.07617	16.67	$< .0001$

统计建模流程——例子（提取信息、统计分析）

去掉截距项：

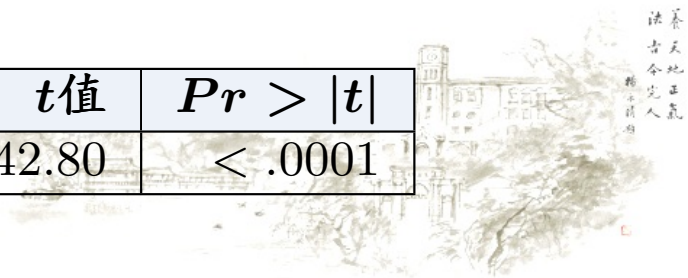
```
1 /*去掉截距项;*/
2 proc reg data=dt;
3 model y=x/noint;
4 run;
```

表：方差分析

源	自由度	平方和	均方	F 值	$Pr > F$
模型	1	10346	10346	1832.10	$< .0001$
误差	19	107.29859	5.64729		
未校正合计	20	10454			

表：参数估计

变量	自由度	参数估计	标准误差	t值	$Pr > \|t\|$
x	1	1.28907	0.03012	42.80	$< .0001$



统计建模流程——例子（提取信息、统计分析）

$$\hat{\beta} = 1.28907$$

x, y间的关系为：

$$y = 1.28907x$$

统计建模流程——例子（回答问题、统计推断）

$$y = 1.28907x$$

1. 表明该模型符合理论模型
2. 该模型符合实际情况
3. 将该模型应用于具体计算：测出直线距离，计算实际距离