概率论与数理统计讲座

杜鸿飞

数学科学学院

2018 年 12 月 13 日

统计简史

早期记录

- 公元前2000年左右，夏朝就进行了人口调查统计；
- 公元前550年左右，古罗马的监察官为了课税和决定能参战的男子人数，每5年做一次人口和财产的登记；
- 公元前300年左右，《印度经典》要求村里的会计保存村里人口、土地使用和农作物收成等数据。

近代统计学

- 统计学*STATISTICS*的词根*STATUS*在拉丁语中是国家的意思，由18世纪德国学者G.Achenwall创造，意为：由国家来收集、处理和使用数据；
- 18世纪中那些搞政治权术的人认为统计学是作为国家权术的一种科学，其作用是成为政府的耳目；
- 然而原始数据是含有杂质且令人困惑的，要使其易懂并能用于决策，需要对原始数据进行归纳整理！

统计简史

现代统计学

- 近代统计学以比利时数学家 凯特勒(A.Quetlet, 1796 - 1874)为代表， 将概率论与统计学结合起来，把统计学用于社会科学，1844年他利用男子身高服从正态分布这一特性，找出了法国躲避征兵的人的身高大小范围，他把应征人的身高分布与一般人的身高分布比较，找出了2000个为躲避征兵而假称低于最低身高的人。
- 1834年创立了英国皇家统计学会，当时认为统计学是“与人类有关的事实，可以由数量表示，并经过大量的累积重复可以导出一般规律”。
- 1908年， Gosset (笔名Student)发表了关于t分布的论文，创立了 小样本代替大样本的方法，开创了统计学的新纪元。
- 当代，统计学+机器学习，对数据的利用达到了一个新的巅峰。

统计结果有什么用处?

一些特别的统计数据

原因	天数	原因	天数
未结婚(男性)	3500	30%超重	1300
未结婚(女性)	1600	20%超重	900
吸香烟(男性)	2250	咖啡	6
吸香烟(女性)	800	家有烟雾警报	-10
危险工作,事故	300	带有气垫的轿车	-50
医疗 X-射线	6	移动冠状动脉监护器	-125

表: 不同原因引起的寿命损失[1]

[1]《统计与真理—怎样运用偶然性》 [美] C.R.劳/著

样本与统计量

以下是随机变量(向量)

总体(X)、样本 $(X_1, X_2, \cdots, X_n)$

统计量，如样本均值$\overline{X}$、样本方差S^2、样本原点矩(A_k)、样本中心矩(M_k)。

以下是数值

总体矩(如$E(X), D(X)$)、样本值$(x_1, x_2, \cdots, x_n)$

统计值，如样本均值$\overline{x}$、样本方差s^2、样本原点矩(a_k)、样本中心矩(m_k)。

构造统计量的目的

由样本构造的统计量，用于估计总体参数。

四个统计分布构造定理

假设随机变量相互独立

1. 标准正态分布：

$$X \sim N(\mu, \sigma^2) \Longrightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

2. χ^2分布：

$$X_i \sim N(0,1), i = 1, 2, \cdots, n \Longrightarrow \chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

3. t分布：

$$X \sim N(0,1), Y \sim \chi^2(n) \Longrightarrow T = \frac{X}{\sqrt{Y/n}} \sim t(n)$$

4. F分布：

$$X \sim \chi^2(n_1), Y \sim \chi^2(n_2) \Longrightarrow F = \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$$

抽样分布定理

假设总体为正态分布$X \sim N(\mu,\sigma^2)$，则$\overline{X}$与S^2相互独立

1. 样本均值： $\overline{X} \sim N(\mu,\sigma^2/n) \Longrightarrow U = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. 样本方差： $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. 样本均值与样本方差： $T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

判断统计量所服从分布的方法：

先不考虑系数，将每一部分化为标准形式分布，然后按构造定理进行组合

(1) 样本的线性组合，一般可化为标准正态分布；

(2) 样本平方的组合，一般可化为χ^2分布；

(3) 两项相除，分母为样本平方的组合，分子若为线性组合，则化为t分布；

(4) 两项相除，分母为样本平方的组合，分子若为平方的组合，则化为F分布。

例1：设总体$X \sim N(0,1)$, $Y \sim N(0,4)$相互独立，X_1, X_2, X_3, X_4 和 $Y_1, Y_2, \cdots, Y_9$分别为来自总体X和Y的样本，确定统计量 $z = \frac{3(X_1+X_2+X_3+X_4)}{\sqrt{\sum_{i=1}^{9} Y_i^2}}$的分布。

解：由正态分布的可加性可知

$$X_1 + X_2 + X_3 + X_4 \sim N(0,4) \Longrightarrow U = \frac{X_1 + X_2 + X_3 + X_4}{2} \sim N(0,1)$$

$$Y_i \sim N(0,4), i = 1,2,\cdots,9 \Longrightarrow V = \sum_{i=1}^{9} \left(\frac{Y_i}{2}\right)^2 \sim \chi^2(9)$$

且U, V独立，故由t分布构造定理可知

$$Z = \frac{U}{\sqrt{V/9}} = \frac{3(X_1 + X_2 + X_3 + X_4)}{\sqrt{\sum_{i=1}^{9} Y_i^2}} \sim t(9)$$

例2：设 $X_1, X_2, \cdots, X_(n+m)$ 是来自总体$X \sim N(0, \sigma^2)$ 的容量为$n+m$ 的简单随机样本，问 $Y = k\frac{\sum\limits_{i=1}^{n} X_i^2}{\sum\limits_{i=n+1}^{n+m} X_i^2}$ 中k取何值可服从F分布，为什么？

解：由于$X_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n+m$

$$U = \sum_{i=1}^{n} \left(\frac{X_i}{\sigma}\right)^2 \sim \chi(n)^2, \qquad V = \sum_{i=n+1}^{n+m} \left(\frac{X_i}{\sigma}\right)^2 \sim \chi(m)^2$$

由F分布构造定理有

$$Y = \frac{U/n}{V/m} = \frac{\sum\limits_{i=1}^{n} \left(\frac{X_i}{\sigma}\right)^2 /n}{\sum\limits_{i=n+1}^{n+m} \left(\frac{X_i}{\sigma}\right)^2 /m} = \frac{m}{n} \frac{\sum\limits_{i=1}^{n} X_i^2}{\sum\limits_{i=n+1}^{n+m} X_i^2} \sim F(n, m)$$

所以，当$k = m/n$时，Y服从F分布。

矩估计和极大似然估计

注意事项

估计量是随机变量，大写；估计值是数值，小写

估计量或估计值，加“上三角”表示估计而非真值，如：$\hat{\theta} = \overline{X}$

假设检验一般步骤

(1) 写出似然函数$L(\theta)$，连续型为联合概率密度，离散型为联合分布律；
(2) 似然函数取对数，$lnL(\theta)$；
(3) 似然函数对θ求偏导，令导数为0得方程(组)；
(4) 求解方程(组)得极大似然估计量。

例3：设测量误差均服从零均值的正态分布，进行4次独立测量，各次误差为$0.1, -0.08, 0.04, -0.07$。求测量误差方差的极大似然估计值。

解：设测量误差$X \sim N(0, \sigma^2), x_1, x_2, \cdots, x_n$ 为样本观测值，则似然函数为

$$L(\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}} = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n} x_i^2}$$

$$\ln L(\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial \ln L(\sigma)}{\partial \sigma} = 0 \Longrightarrow -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n} x_i^2 = 0 \Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

代入当前4次测量数据，得测量误差方差的极大似然估计值为

$$\hat{\sigma}^2 = \frac{1}{4}\left(0.1^2 + (-0.08)^2 + 0.04^2 + (-0.07)^2\right) = 0.005725$$

例4：设总体X的概率密度为 $f(x)=\begin{cases} e^{(-x-\theta)}, & x\geq\theta \\ 0, & x<\theta. \end{cases}$ 其中θ为未知参数，$X_1,X_2,\cdots,X_n$是来自X的简单随机样本，(1)写出参数θ的似然函数；(2) 当样本观测值为$1.5,2,1.8,2.2,3$ 时，求θ的极大似然估计值$\hat{\theta}$。

解：(1)θ的似然函数为

$$L(\theta;x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\begin{cases} e^{n\theta-\sum\limits_{i=1}^{n}x_i}, & x_i\geq\theta(i=1,2,\cdots,n) \\ 0, & \text{else.} \end{cases}$$

(2) 当样本观测值为$1.5,2,1.8,2.2,3$ 时，似然函数为

$$L(\theta;1.5,2,\cdots,3)=\begin{cases} e^{5\theta-10.5}, & \theta<1.5 \\ 0, & \text{else.} \end{cases}$$

当θ取1.5时，似然函数达到极大值 $\hat{\theta}=1.5$。

估计量的优良性准则

无偏性、有效性、相合性

1. 无偏：估计量的期望等于真实值，$E(\hat{\theta}) = \theta$
2. 最有效：该无偏估计量$\hat{\theta}_0$的方差最小，$D(\hat{\theta}_0) < D(\hat{\theta})$
3. 相合：$\lim_{n\to\infty} P\{|\hat{\theta}_n - \theta| < \varepsilon\} = 1$

- 无偏性、有效性，涉及期望、方差的计算，可与第四章结合分析
- 相合性，涉及依概率收敛，可用第五章的大数定律进行分析，或切比雪夫不等式进行分析。

估计量的优良性准则

例5：设$\hat{\mu}_1,\hat{\mu}_2$ 是总体均值 μ 的两个估计量，已知 $E(\hat{\mu}_1)=\mu, E(\hat{\mu}_2)=\mu$， $D(\hat{\mu}_1)=1, D(\hat{\mu}_2)=4, \rho_{\hat{\mu}_1\hat{\mu}_2}=1/2$，若统计量$\hat{\mu}_3=c_1\hat{\mu}_1+c_2\hat{\mu}_2(c_1>0, c_2>0)$，问：要使形如 $\hat{\mu}_3$ 的统计量为 μ 的无偏估计量且最有效，常数 c_1, c_2 应满足什么条件？（写出式子即可，不必解出具体结果）

解：要使$\hat{\mu}_3$ 为 μ 的无偏估计量，应有

$$E(\hat{\mu}_3)=E(c_1\hat{\mu}_1+c_2\hat{\mu}_2)=c_1E(\hat{\mu}_1)+c_2(\hat{\mu}_2)=c_1\mu+c_2\mu=\mu$$

从而有 $c_1+c_2=1$

$$cov(c_1\hat{\mu}_1, c_2\hat{\mu}_2)=c_1c_2cov(\hat{\mu}_1,\hat{\mu}_2)$$

$$=c_1c_2\rho_{\hat{\mu}_1\hat{\mu}_2}D(\hat{\mu}_1)D(\hat{\mu}_2)=c_1c_2\times\frac{1}{2}\times 1\times 4=2c_1c_2$$

$$D(\hat{\mu}_3)=D(c_1\hat{\mu}_1+c_2\hat{\mu}_2)=c_1^2D(\hat{\mu}_1)+c_2^2D(\hat{\mu}_2)+2cov(c_1\hat{\mu}_1,c_2\hat{\mu}_2)$$

$$=c_1^2+4c_2^2+2c_1c_2$$

故应有 $c_1+c_2=1$ 且使得 $c_1^2+4c_2^2+2c_1c_2$ 最小。

统计学

抽样与统计量

矩估计和极大似然估计

估计量的优良性

区间估计与假设检验

线性回归

区间估计与假设检验 → 单正态总体 $X \sim N(\mu, \sigma^2)$

基本思想： 对待估(检验)参数，选择优良估计量，据此化为常见统计分布之一。

μ, σ 的优良估计量：

- μ　　无论σ^2已知还是未知，$\overline{X} \Longrightarrow \mu$
- σ^2　　μ 已知时，$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \Longrightarrow \sigma^2$
- σ^2　　μ 未知时，$S^2 \Longrightarrow \sigma^2$

可以证明：每种情况下的统计量都是该参数的无偏、有效、相合估计量！

重点掌握：抽样分布定理6.2.4！

区间估计与假设检验 $\rightarrow$ 单正态总体$X \sim N(\mu, \sigma^2)$

假设检验基本步骤:

1. 给出原假设H_0和对立假设H_1 【这一步很 关键，尤其是对立假设的符号】
2. 给出适当的检验统计量【可尝试自己构造统计量，方法见上页】
3. 根据对立假设给出拒绝域【注意单侧检验用α，双侧检验用$\alpha/2$】
4. 根据统计值给出检验结果。

例6：已知某电子元件的长度服从正态分布，且方差为0.01。从一批次的产品中任取10个，测得$s^2=0.012$，则在显著性水平$\alpha=0.05$下，这批次电子元件的精度（即标准差）是否正常？$(\chi^2_{0.975}(10)=3.247, \chi^2_{0.025}(10)=20.483, \chi^2_{0.975}(9)=2.700, \chi^2_{0.025}(9)=19.023)$

思考：1. 检验的是什么？2. 如何给原假设与对立假设？3. 统计量怎么选择？

分析1：检验的是什么？

- 题目中“测得$s^2=0.012$”，未提及样本均值，可能是对方差进行检验；
- 题目中所给的分位数均为χ^2分布，可能是采用χ^2检验法；
- 题目中明确提出判断“精度（即标准差）是否正常”，可以确认是对方差进行检验。

例6：已知某电子元件的长度服从正态分布，且方差为0.01。从一批次的产品中任取10个，测得$s^2 = 0.012$，则在显著性水平$\alpha = 0.05$下，这批次电子元件的精度（即标准差）是否正常？($\chi^2_{0.975}(10) = 3.247, \chi^2_{0.025}(10) = 20.483, \chi^2_{0.975}(9) = 2.700, \chi^2_{0.025}(9) = 19.023$)

思考：1. 检验的是什么？2. 如何给原假设与对立假设？3. 统计量怎么选择？

分析2：原假设和对立假设

- (1) 提出的原假设对目标有利，而对立假设对目标不利；(2) 若检验参数有偏向，则用单侧检验，且注意此时拒绝域中的分位数对应α而非$\alpha/2$。例如对产品平均寿命μ进行检验，原假设应为$H_0 : \mu \geq \mu_0$，对立假设为$H_1 : \mu < \mu_0$。
- 本题中元件长度偏长或偏短均不利，因此应为双侧假设检验，即原假设为等式，对立假设为$\neq$。

例6：已知某电子元件的长度服从正态分布，且方差为0.01。从一批次的产品中任取10个，测得$s^2 = 0.012$，则在显著性水平$\alpha = 0.05$下，这批次电子元件的精度（即标准差）是否正常？$(\chi^2_{0.975}(10) = 3.247, \chi^2_{0.025}(10) = 20.483, \chi^2_{0.975}(9) = 2.700, \chi^2_{0.025}(9) = 19.023)$

思考：1. 检验的是什么？2. 如何给原假设与对立假设？3. 统计量怎么选择？

分析3：统计量的选择

- 本题对方差进行检验，但期望μ未知，故S^2是方差σ^2的优良估计量；
- 根据抽样分布定理，有：

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

例6：已知某电子元件的长度服从正态分布，且方差为0.01。从一批次的产品中任取10个，测得$s^2=0.012$，则在显著性水平$\alpha=0.05$下，这批次电子元件的精度（即标准差）是否正常？$(\chi^2_{0.975}(10)=3.247, \chi^2_{0.025}(10)=20.483, \chi^2_{0.975}(9)=2.700, \chi^2_{0.025}(9)=19.023)$

解：设电子元件的长度为$X\sim N(\mu,\sigma^2)$，由题意知需检验
$H_0:\sigma^2=0.01, \qquad H_1:\sigma^2\neq 0.01$
由于总体期望μ未知，在原假设成立的条件下，检验统计量为

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

原假设H_0的拒绝域为：$\chi^2>\chi^2_{\frac{\alpha}{2}}(n-1)$ 或 $\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1)$
由于 $\chi^2=9\times\frac{0.012}{0.01}=10.8, \chi^2_{0.025}(9)=19.023, \chi^2_{0.975}(9)=2.700$

$$\chi^2_{1-\frac{\alpha}{2}}(n-1)<\chi^2<\chi^2_{\frac{\alpha}{2}}(n-1)$$

所以不能拒绝原假设，即认为在显著性水平$\alpha=0.05$下，这批次电子元件的精度（即标准差）正常。

统计学
抽样与统计量
矩估计和极大似然估计
估计量的优良性
区间估计与假设检验
线性回归

线性回归

需掌握内容

- 能绘制散点图，并判断函数形式；
- 能估算参数a, b, σ^2，注意公式不要混淆；
- 能采用R检验法判断线性关系是否显著；
- 能进行简单预测（将数据代入回归方程即可预测）。

公式的记忆：对应总体矩或样本矩，如

$$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \rightleftharpoons M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$D(X) = E\{[X - E(X)]^2\} = E(X^2) - E(X)^2$$

$$l_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\rightleftharpoons cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\} = E(XY) - E(X)E(Y)$$

例7：下表列出了在不同挂物质量$X(g)$下弹簧长度$Y(cm)$的测量值

x_i	5	10	15	20	25	30
y_i	7.25	8.12	8.95	9.90	10.9	11.8

(1) 作散点图，能否从直观上认为X与Y有明显的线性关系？

(2) 若计算得$\bar{x} = \frac{1}{6}\sum\limits_{i=1}^{6} x_i = 17.5, \bar{y} = \frac{1}{6}\sum\limits_{i=1}^{6} y_i = 9.49, \sum\limits_{i=1}^{6} x_i^2 = 2275, \sum\limits_{i=1}^{6} y_i^2 = 554.66, \sum\limits_{i=1}^{6} x_i y_i = 1076.2$，试检验$X$与$Y$的线性相关关系是否显著？

$(\alpha = 0.01, R_{0.01}(4) = 0.917, R_{0.01}(5) = 0.847\ )$

解：(1) 作散点图，直观上可看出有明显线性关系。

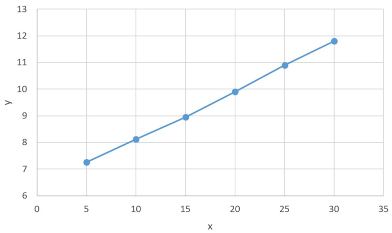

图: 散点图

(2) 由于$l_{xy}=\sum_{i=1}^{6}x_iy_i-6\bar{x}\bar{y}=79.75$,

$$l_{xx}=\sum_{i=1}^{6}x_i^2-6\bar{x}^2=437.5,\quad l_{yy}=\sum_{i=1}^{6}y_i^2-6\bar{y}^2=14.3$$

$$R=\frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}=\frac{79.75}{\sqrt{437.5}\sqrt{14.3}}>R_{0.01}(4)=0.917$$

故认为X与Y的线性关系显著。