计算机应用数学

Applied Mathematics for Computer Science

主讲教师：宋骐

https://songqi1990.github.io/

中国科学技术大学
University of Science and Technology of China

1

1.2 离差及其相关定理

2

§1.2.1 马尔可夫定理
§1.2.2 切比雪夫定理
§1.2.3 参数估计与假设检验
§1.2.4 切诺夫界

3

离差
Deviation from the Mean

- 考虑随机变量可能永远不会取期望附近的值
- 判断抽样和测量的精度->偏离预期值的概率
- 极端（extreme）离差的概率广泛应用于工程领域
  - 一堵墙能抵御多长时间的海啸
  - 组装设备一个月能容忍多少组件故障

  …

§1.2.1 马尔可夫不等式

马尔可夫不等式（Markov’ s Inequality)

如果X是一个非负随机变量，那么对任意$x > 0$

$$P[X \geq a] \leq \frac{E(X)}{a}$$

证明：假设Y_a是基于X和a的一个随机变量：

$$Y_a = \begin{cases} 0, if\ X < a \\ a, if\ X \geq a \end{cases}$$

对于变量Y_a，其期望可以表示为$E[Y_a] = \Pr(X \geq a) * a$，且$Y_a \leq X$

$\Rightarrow E[Y_a] \leq E(X)$

$\Rightarrow \Pr(X \geq a) * a \leq E(X)$

$\Rightarrow \Pr(X \geq a) \leq \frac{E(X)}{a}$

马尔可夫不等式：

如果R是一个非负随机变量，那么对任意$x > 0$，

$$P[X \geq a] \leq \frac{E(X)}{a}$$

推论：

如果X是一个非负随机变量，那么对任意$c \geq 1$

$$P[X \geq c \cdot E[X]] \leq \frac{1}{c}$$

令$a = c \cdot E[X]$即可得到此推论

例 n 个人把自己的帽子放进了一个房间，他们的帽子全部混在了一起，然后每个人再随机地取回一顶帽子，请问恰好k个人拿到自己帽子的概率不超过?

解法1： 设X为正确拿到自己帽子的人数，令$X = x_1 + x_2 + \cdots + x_n$, $x_i = 1$表示第i个人得到自己的帽子，否则$x_i = 0$:

$$E(X) = \mathrm{E}(x_1) + \mathrm{E}(x_2) + \cdots E(x_n),$$

注x_i, x_j不是相互独立，$n-1$拿到自己的帽子，最后一个人肯定也拿到自己帽子

解：
$$E(x_i) = 1 * \frac{1}{n} + 0 * \left(1 - \frac{1}{n}\right) = \frac{1}{n}$$
$$E(X) = \mathrm{E}(x_1) + \mathrm{E}(x_2) + \cdots E(x_n) = 1$$

基于马尔可夫不等式：

$$P[X \geq k] \leq \frac{E(X)}{k} = \frac{1}{k}$$

上界太大

解法2： 设X表示k人中每个人都拿到了自己的帽子，Y表示事件其它的$n-k$个人中没有人拿到自己的帽子。 x_i表示第i个人得到自己的帽子, 则

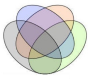
$$P(XY) = \mathrm{P(X)P(Y|X)}$$

$$P(X) = \mathrm{P}(x_1, \cdots x_k)$$
$$= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\cdots P(x_k|x_1 \cdots x_{k-1})$$
$$= \frac{1}{n} \cdot \frac{1}{n-1} \cdots \frac{1}{n-k+1} = \frac{(n-k)!}{n!}$$

第1.2节 离差及其相关定理 第10页

$P(XY) = \mathrm{P(X)P(Y|X)}$

$P(Y|X)$ 为$n-k$个人随机从$n-k$中找帽子，没有人选中的概率



考虑至少一个人选中的概率：

$P(\cup x_i)$

$$= \sum_i P(x_i) - \sum_{i_1<i_2} P(x_{i_1}x_{i_2}) + \cdots$$

$$+ (-1)^{j+1} \sum_{i_1<i_2<i_j} P\left(x_{i_1}x_{i_2}\cdots x_{i_j}\right) + \cdots + (-1)^{n-k+1} P(x_1x_2\cdots x_{n-k})$$

$P\left(x_{i_1}x_{i_2}\cdots x_{i_j}\right) = \frac{(\mathrm{n-k-j})!}{(n-k)!}$ j个人选到自己的,$n-k-j$个人随机选

$$\sum_{i_1<i_2<i_j} P\left(x_{i_1}x_{i_2}\cdots x_{i_j}\right) = \frac{C_{n-k}^{j}(\mathrm{n-k-j})!}{(n-k)!} = \frac{1}{j!}$$

$$\mathrm{P(Y|X)} = 1 - P(\cup x_i) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots \frac{(-1)^{n-k}}{(n-k)!} = \sum_{i=0} \frac{(-1)^i}{i!}$$

27 September 2025 10

10

第1.2节 离差及其相关定理 第11页

$$P(XY) = \mathrm{P(X)P(Y|X)}$$

由于$P(X) = \frac{(n-k)!}{n!}$， $\mathrm{P(Y|X)} = \sum_{i=0}^{n-k} \frac{(-1)^i}{i!}$

因此指定的 k人拿到自己帽子而其它人没有捡到自己帽子的概率

$$P(XY) = \frac{(n-k)!}{n!} \sum_{i=0}^{n-k} \frac{(-1)^i}{i!}$$

只有k个人找到帽子概率为

$$C_n^k P(XY) = \frac{\mathrm{n!}}{\mathrm{k!(n-k)!}} \frac{(n-k)!}{n!} \sum_{i=0}^{n-k} \frac{(-1)^i}{i!} \le \frac{e^{-1}}{k!} < \frac{1}{k!}$$

基于Taylor 级数$\mathrm{e}^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$

27 September 2025 11

11

第1.2节 离差及其相关定理 第12页

例 假设某学校的大学生平均IQ为150，请问这学校的学生IQ大于200的概率?

解： 设X表示学校学生IQ，从假设条件$\mathrm{E}(X) = 150$ 根据马尔可夫定理：

$$P(X \ge 200) \le \frac{\mathrm{E(X)}}{200} = \frac{3}{4}$$

让我们观察另外一个事实：某校没有大学生智商低于100，令$\mathrm{Y} = X - 100$，Y为非负，$E(\mathrm{Y}) = 50$，我们有：

$$P(X \ge 200) = P\ (Y \ge 100) \le \frac{\mathrm{E}(Y)}{100} = \frac{1}{2}$$

当对X-a（非负数）使用马尔可夫定理将会获得更小的上界

27 September 2025 12

12

§1.2.2 切比雪夫不等式

切比雪夫不等式（Chebyshev’s Inequality)

设随机变量R和正实数x，

$$P[|X-E[X]| \geq c] \leq \frac{Var[X]}{c^2}$$

证明：基于马尔可夫不等式

$$Pr(X \geq a) \leq \frac{E[X]}{a}$$

令$E[X]=\mu$，$Var[X]=\sigma^2$，取$X=(X-\mu)^2$，$a=c^2$，代入上式：

$$Pr\left((X-\mu)^2 \geq c^2\right) \leq \frac{E\left[(X-\mu)^2\right]}{c^2} = \frac{\sigma^2}{c^2}$$

$$\rightarrow Pr(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

例 假设各人群中，IQ的标准差大约为15，国民平均IQ为100，请问IQ为300或更高的可能性?

解： 设X表示国民IQ，根据马尔可夫定理，可得粗略边界：

$$P(X \geq 300) \leq \frac{E(X)}{300} = \frac{1}{3}$$

基于切比雪夫定理获得更严格的边界：

$$P(X \geq 300) = P\ (X-100 \geq 200)$$

$$\leq \frac{Var(X)}{200^2} = 15^2/200^2 \approx 1/178$$

178人中至多有一位IQ大于等于300

例 假设n个学生，一年共有d天，设M为生日匹配的学生对的数量，请估计匹配数目的可能性?

解： 令$B_1, B_2, \cdots, B_n$是n个人的生日，令$S_{i,j}=1(i \neq j)$表示$B_i=B_j$，给定i，j则 $E(S_{i,j}) = P(B_i=B_j) = \frac{1}{d}$

生日匹配的对数:$M = \sum_{1 \leq i<j \leq n} S_{i,j}$

$$E(M) = \sum_{1 \leq i<j \leq n} E(S_{i,j}) = C_n^2 \cdot 1/d$$

生日匹配事件两两独立：

$$Var(M) = \sum_{1 \leq i<j \leq n} Var(S_{i,j}) = C_n^2 \cdot 1/d(1-1/d)$$

$$P\left(M-E(M) \geq x^2\right) \leq C_n^2 \cdot 1/d(1-1/d)/x^2$$

n=95,d=365,x=7，95人中有超过75%概率，6-19对生日匹配学生

例 假设产品合格率为p，为了估计p，随机选择n个产品,假设我们希望估计合格率与p相差在0.04内的概率至少为95%，需要抽查n至少为多少?

解： 定义x_i为第i个产品为合格产品的变量,$x_i = 1$表示产品合格

令 $X_n = x_1 + x_2 + \cdots + x_n$，我们使用$X_n/n$来估计$p$（statistical estimate）

为满足$P\left(\left|\frac{X}{n} - p\right| \le 0.04\right) \ge 0.95$

x_i两两独立, X服从二项式分布: $Var(X) = np(1-p) \le n\frac{1}{4}$

$$Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} \le 1/4n$$

为满足$P\left(\left|\frac{X}{n} - p\right| \ge 0.04\right) \le \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \le 1 - 0.95$

$n \ge 3125$

§1.2.3 参数估计与假设检验

中心极限定理

设从均值为μ,方差为σ^2(有限）的任意总体中抽取样本量为n的样本，当n充分大，样本均值的抽样分布服从均值为μ、方差为σ^2/n正态分布。

$$\bar{X} \approx \mu + \frac{\sigma}{\sqrt{n}} \cdot N(0,1).$$

- 独立随机变量的标准化的和随样本量变大会趋向正态分布
- 不要求随机变量本身是正态分布的，在一定条件下，我们可以使用对正态分布成立的方法去应对非正态分布。

样本方差

- **总体方差**：给定一组数$x_1, x_2, \cdots, x_N$, $\mu = E(x_i)$, $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
- **样本方差**：给定一组数$x_1, x_2, \cdots, x_n$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $s^2 = \frac{\sum_{i=1}^{n}(x_i-x)^2}{n-1}$

例 设$x_1, x_2, \ldots, x_3$是均值为x、方差为σ^2的随机变量X的n个观测值的随机样本，证明：样本方差s^2是总体方差σ^2的一个无偏估计，其中：
a) 被抽样总体为正太分布
b) 被抽样总体的分布未知

证明 a)：当样本来自正态分布时：

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

其中χ^2是自由度为$v = (n-1)$的卡方随机变量

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

$$\Rightarrow E(s^2) = E(\frac{\sigma^2}{(n-1)}\chi^2) = \frac{\sigma^2}{(n-1)}E(\chi^2)$$

$$\Rightarrow E(s^2) = \sigma^2 \qquad \boxed{E(\chi^2) = v = (n-1)}$$

z检测

根据中心极限定理

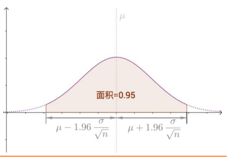
$$\bar{X} \approx \mu + \frac{\sigma}{\sqrt{n}} \cdot N(0,1).$$

如果总体均值、方差已知，可以用z统计量进行z检验

$$z = \frac{x-\mu}{\sigma/\sqrt{n}} \sim N\ (0,1)$$

$$z = \frac{x-\mu}{\sigma/\sqrt{n}} = 1.96,\quad SE_X = \sigma/\sqrt{n}$$

$$\Rightarrow x \in [\mu - 1.96SE_X, 1.96SE_X + \mu]$$



t检测

如果总体均值、方差已知，可以用z统计量进行z检验

$$z=\frac{x-\mu}{\sigma/\sqrt{n}}\sim N\ (0,1)$$

如果总体均值、方差未知，可以用t统计量进行t检验

$$t=\frac{x-\mu}{s/\sqrt{n}}=\frac{x-\mu}{\sigma/\sqrt{n}}/\frac{s}{\sigma}=\frac{z}{\sqrt{\frac{\frac{s^2(n-1)}{\sigma^2}}{n-1}}}=z/\sqrt{Y/\ (n-1)}$$

$$Y=\frac{s^2(n-1)}{\sigma^2}=\frac{\sum_{i=1}^{n}(x_i-x)^2}{n-1}(n-1)/\sigma^2=\frac{\sum_{i=1}^{n}(x_i-x)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(x_i-\mu)^2-n(\mu-x)^2}{\sigma^2}=\sum_{i=1}^{n}((x_i-\mu)/\sigma)^2-(\frac{x-\mu}{\sigma/\sqrt{n}})^2$$

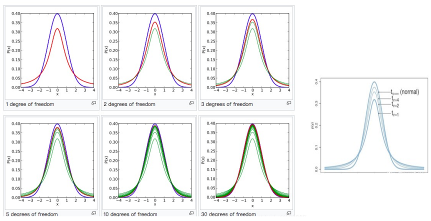
$$E((x_i-\mu)/\sigma)=0, Var((x_i-\mu)/\sigma)=1 \qquad Y\sim\chi^2(n-1)$$

t分布

$$t=\frac{z}{\sqrt{\frac{Y}{n-1}}}$$

$z\sim N(0,1)$
$Y\sim\chi^2(n-1)$

蓝色曲线为正态分布



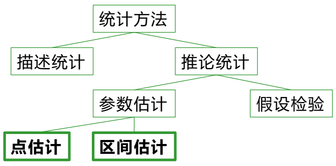
n采样，$v=n-1$为自由度，在自由度为 1, 2, 3, 5, 10, 30比较于标准正态分布，当样本为30，已经接近正态分布，当自由度趋于无穷，t分布为正态分布。

- 统计方法
  - 描述统计
  - 推论统计
    - **参数估计**
    - **假设检验**

参数估计和假设检验是统计推断的两个组成部分，都是利用样本对总体进行某种推断，但推断的角度不同。

- **参数估计讨论的是用样本统计量估计总体参数的方法**
- **假设检验讨论的是用样本信息去检验对总体参数的某种假设是否成立的程序和方法。**

25



26

参数估计

点估计：用样本统计量来估计总体参数，样本统计量为数轴上某一点值，估计的结果也以一个点的数值表示。

- **对总体平均数 μ 的估计，用样本均数;**
- **对总体参数 σ^2 的估计，常用样本方差;**
- **对总体相关系数ρ的估计，常用样本相关系数r。**

优点：提供总体参数的估计值。

缺点：点估计总是以误差的存在为前提，又不能提供正确估计的概率。

27

参数估计

区间估计：根据估计量以一定可靠程度推断总体参数所在的区间范围。

用数轴上的一段距离表示未知参数**可能落入的范围**，它虽不具体指出总体参数等于什么，但能指出总体的**未知参数落入某一区间的概率**有多大。

优点：可以解释总体参数落入某置信区间可能的概率。

缺点：如何平衡**成功估计的概率大小**及**估计范围大小**。

28

区间估计

- **置信区间**(置信间距)：指在一定可靠程度上，总体参数所在的区域距离或区域长度。
- **置信界限**(临界值):置信区间的上下两端 点值。
- **显著性水平**(意义阶段/信任系数):指估计总体参数落在某一区间时，可能犯错误的 概率，用符号α表示。
- **置信度**(置信水平):1- α 。

29

一般规定

- 正确估计的概率(置信水平) 为 .95 或 .99
- 显著性水平为 .05 或 .01
- .05 或 .01 属于小概率事件，原理：小概率事件在一次抽样中是不可能出现的
- α=0.01表示反复抽样1000次，则得到的1000个区间中不包含真值的仅为10个左右

30

总体平均数的估计

- 1. 利用抽样的方法抽取样本，计算平均值X和标准差S
- 2. 计算样本平均数的标准误差SE_X
  - 当总体方差已知时：$SE_X = \frac{\sigma}{\sqrt{n}}$，当总体方差未知时：$SE_X = \frac{S_n}{\sqrt{n-1}}$
- 3. 确定显著性水平和置信水平
- 4. 根据样本平均数的抽样分布查表，确定理论值
- 5. 确定置信区间
  - 当理论值为$Z_{\alpha/2}$时，置信区间为$[X - Z_{\alpha/2} \cdot SE_X, X + Z_{\alpha/2} \cdot SE_X]$
  - 当理论值为$T_{\alpha/2}$时，置信区间为$[X - T_{\alpha/2} \cdot SE_X, X + T_{\alpha/2} \cdot SE_X]$
- 6. 解释总体平均数的置信区间

31

总体平均数的估计

总体分布为正态分布且总体方差已知

例1：已知某农场某批次冬瓜重量的总体方差为5.89公斤， 从该农场随机抽取15 个冬瓜，其平均重量为22.4公斤， 试求该农场冬瓜平均重量的95%和99%的置信区间。

32

总体分

解：

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

33

总体平均数的估计

解： $SE_X = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5.89}}{\sqrt{15}} = \frac{2.43}{3.87} = 0.63$

95%的置信区间的显著性水平$\alpha = 0.05$，$Z_{\alpha/2} = 1.96$

所以95%的置信区间为

$22.4-1.96\times0.63 \le u \le 22.4+1.96\times0.63$,即 $[21.2, 23.6]$

99%的置信区间的显著性水平$\alpha = 0.01$，$Z_{\alpha/2} = 2.58$

所以99%的置信区间为

$22.4-2.58\times0.63 \le u \le 22.4 + 2.58\times0.63$,即 $[20.8, 24.0]$

总体平均数的估计

总体分布为非正态分布且总体方差已知

例2：已知某批次产品木材产品长度的方差为436.8cm，现从批次产品中抽取58件，测得该组产品长度的平均数为198.4cm，试求批次产品平均长度的95%和99% 的置信区间。

当样本容量 n>30时,此时样本抽样分布渐近正态分布。这时可依正态分布进行估计，否则不能对总体平均数进行估计。

总体平均数的估计

解： 由于样本容量大于30，该样本的抽样分布为渐进正态分布：

$SE_X = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{436.8}}{\sqrt{58}} = \frac{20.9}{7.6} = 2.75$

所以95%的置信区间为

$198.4-1.96\times2.75 \le u \le 198.4+1.96\times2.75$,即 $[193.01, 203.79]$

所以99%的置信区间为

$198.4-2.58\times2.75 \le u \le 198.4+2.58\times2.75$,即 $[191.3, 205.5]$

总体平均数的估计

总体分布为正态分布且总体方差未知

例3：从某市抽取20 名7 岁女童，经测量，这20名女童的平均身高为116cm，标准差为5cm，试求该市7岁女童总平均身高的95%和99%的置信区间。

无论样本容量n的大小，从该总体抽取的样本所形成的分布均服从自由度为n-1的t分布，对总体平均数的估计可依t分布进行估计

总体分布

解： 总 均数分布

符合t分

SI

	α				
v	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.030	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

总体平均数的估计

解： 总体方差未知，总体分布为正态分布，故样本均数分布符合t分布，可以依t分布对总平均身高进行估计

$SE_X = \frac{s}{\sqrt{n-1}} = \frac{5}{\sqrt{19}} = 1.15$ df=n-1=19

$t_{0.05/2(19)} = 2.093$ $t_{0.01/2(19)} = 2.861$

所以95%的置信区间为

$116-2.093\times1.15 \le u \le 116+2.093\times1.15$,即 [113.59, 118.41]

所以99%的置信区间为

$116-2.861\times1.15 \le u \le 116+2.861\times1.15$,即 [112.71, 119.29]

总体平均数的估计

总体分布为非正态分布且总体方差未知

例4：某校进行一次考试，从中抽取40名考生经计算，这40 名考生的平均成绩为82分，标准差为7分，试求全体考生平均成绩的95%和 99%的置信区间。

只有当样本容量 n>30时，此时样本抽样分布服从自由度为n-1的t分布，这时可依t 分布对总体平均数进行估计，否则不能对总体平均数进行估计。

总体平均数的估计

解： 由于n>30，可以依t分布对全体考生评价成绩进行估计

$SE_X = \frac{S}{\sqrt{n-1}} = \frac{7}{\sqrt{39}} = 1.12$ df=n-1=39

$t_{0.05/2(40)} = 2.021$ $t_{0.01/2(40)} = 2.704$

所以95%的置信区间为

$82-2.021\times1.12 \le u \le 82+2.021\times1.12$,即 [79.74,84.26]

所以99%的置信区间为

$82-2.704\times1.12 \le u \le 82+2.704\times1.12$,即 [78.97,85.03]

总体标准差的估计

- 1. 与总体平均数的估计过程类似
- 2. 由于总体标准差未知，可以使用其无偏估计量S_{n-1}作为替代

$$SE_X = \frac{S_{n-1}}{\sqrt{2n}}$$

- 3. 总体为正态分布时，若n>30，则可以通过正态分布来估计，否则，总体标准差无法估计

总体标准差的估计

例5：某区一次英语统考中，随机抽取40名考生，计算其英语成绩的标准差为15.6，试求该区英语统考成绩总标准差的95%和99%的置信区间。

总体标准差的估计

解：由于n>30，可以依正态分布进行估计

$$SE_S = \frac{S_{n-1}}{\sqrt{2n}} = \frac{S\times\sqrt{n}/\sqrt{n-1}}{\sqrt{80}} = \frac{15.6\times\sqrt{40}/\sqrt{39}}{\sqrt{80}} = 1.77$$

因此，95%的置信区间为

$15.6-1.96\times1.77 \le \sigma \le 15.6+1.96\times1.77$,即 $[12.15, 19.05]$

所以99%的置信区间为

$15.6-2.58\times1.77 \le \sigma \le 15.6+2.58\times1.77$,即 $[11.03, 20.17]$

总体方差的估计

- 1. χ^2分布：从正态分布的总体中，随机抽取容量为n的样本，其样本方差与总体方差的分布

$$\chi^2 = \frac{\sum(X-X)^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{(n-1)S_{n-1}^2}{\sigma^2} \ldots$$

- 2. 利用理论χ^2值与样本方差来确定总体方差的置信区间：

$$\frac{nS^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{nS_n^2}{\chi^2_{1-\alpha/2}} \quad 或 \quad \frac{(n-1)S_{n-1}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S_{n-1}^2}{\chi^2_{1-\alpha/2}}$$

总体方差的估计

例6：**在某市进行的一次智力测验中，随机抽取20名12岁学生，经计算其治理测验的方差为72.25，试求该市12岁学生智力测验分数总体方差的95%和99%的置信区间。**

46

n \ α	0.995	0.99	0.975	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928

47

总体方差的估计

解：**我们认为智力测验分数服从正态分布，由该总体中抽出的样本的估计总体方差时服从χ^2分布。**

df=n-1=19，由χ^2分布值表，$\chi^2_{.05/2}=32.9$，$\chi^2_{1-.05/2}=8.91$，

$\chi^2_{.01/2}=38.6, \chi^2_{1-.01/2}=6.84$

所以95%的置信区间和99%的置信区间分别为

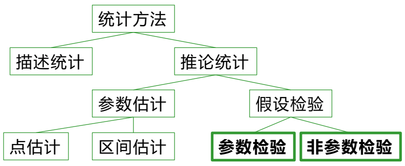
$$\frac{20\times72.25}{32.9}\le\sigma^2\le\frac{20\times72.25}{8.91} \qquad \frac{20\times72.25}{38.6}\le\sigma^2\le\frac{20\times72.25}{6.84}$$

即 $[43.92, 162.18]$ $\qquad$ $[37.44, 211.26]$

48

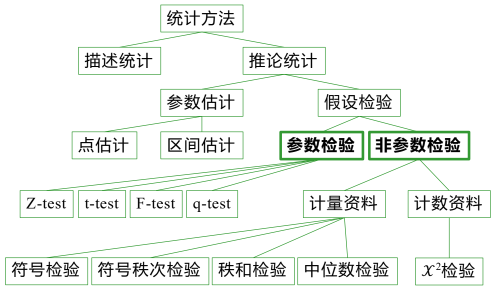
参数估计和假设检验是统计推断的两个组成部分，都是利用样本对总体进行某种推断，但推断的角度不同。

- **参数估计**讨论的是用样本统计量估计总体参数的方法
- **假设检验**讨论的是用样本信息去检验对总体参数的某种假设是否成立的程序和方法。



假设检验

- **参数假设检验：**总体的分布形式已知，需要对总体的未知参数进行假设检验。
- **非参数假设检验：**对总体分布形式所知甚少，需对未知分布函数的形式及其他特征进行假设检验。



§1.2.4 切尔诺夫界
Chernoff Bound

设$x_1, x_2, \cdots, x_n$是相互独立的随机变量，对任意i满足$0 \le x_i \le 1$,令$X = x_1 + x_2 + \cdots + x_n$，对所有$c \ge 1$,

$$P[X \ge cE(X)] \le e^{-\beta(c)E(X)}$$

$$\beta(c) = c\ln c - c + 1$$

二项式尾切尔诺夫界

独立抛1000次硬币，求正面次数超过期望20%及以上的概率边界

解

$$X = x_1 + x_2 + \cdots + x_{1000}$$

根据目标，$c = 1.2, \beta(c) = c\,lnc - c + 1 = 0.0187$

假设硬币均匀，$E(X) = 500$，$P[X \ge 1.2E(X)] \le e^{-0.0187*500} < 0.000083$

假设扔100万次，$E(X) = 500000$，$P[X \ge 1.2E(X)] \le e^{-9392}$

假设超过期望30%，$E(X) = 500$，$P[X \ge 1.2E(X)] \le e^{-0.041*500}$

百万分之一

彩票游戏切尔诺夫界

Pick-4：花1元在0000-9999间选四位数，如果随机摇号选中了号码，获得5000元，中奖率1/10000。如果1000万人参加这个游戏，彩票发行商人不敷出的概率是多少?

解 中奖总人数：$X = x_1 + x_2 + \cdots + x_n$
中奖期望人数：1000人

如果假设玩家挑选数字及中奖数字随机、独立、均匀分布，超过2000人中奖入不敷出，$c = 2, \beta(c) = c\ lnc - c + 1 = 0.386$

$$P[X \geq 2E(X)] \leq e^{-386}$$

c=1.1，$P[X \geq 1.1E(X)] \leq e^{-0.00484*1000} < 0.01$

随机负载均衡切尔诺夫界

总共有24000个任务，平均每个任务耗时$\frac{1}{4}$秒，确定服务器台数m，使得给定时间间隔内不大可能任一服务器被分配超过600秒的负荷而导致超载。

解 我们先确定第一台服务器超载概率。设T为第一台服务器分到负载秒数，计算$P(T \geq 600)$上界

$$T = t_1 + t_2 + \cdots + t_n, n = 24000$$

我们先确定第一台服务器超载概率。设T为第一台服务器分到负载秒数，计算$P(T \geq 600)$上界

假设任务到哪台服务器与任务耗时无关，每个任务耗时不超过1秒，平均每个任务耗时$\frac{1}{4}$秒。任务随机分配到m台服务器，第一台的期望负载为

$$E(X) = 24000 * \frac{1}{4} / m$$

随机负载均衡切尔诺夫界

总共有24000个任务，平均每个任务耗时$\frac{1}{4}$秒，确定服务器台数m，使得给定时间间隔内不大可能任一服务器被分配超过600秒的负荷而导致超载。

解 我们先确定第一台服务器超载概率。设T为第一台服务器分到负载秒数，计算$P(T \geq 600)$上界

$$E(X) = 24000 * \frac{\frac{1}{4}}{m} = 6000/m$$

超负载c=m/10:$P[X \geq 600] \leq e^{-(c\ lnc - c + 1) \cdot 6000/m}$

P[有一台服务器超载]≤ $\sum_{i=1}^{m} P$ [第i台服务器超载]

=mP [第1台服务器超载]≤ $me^{-(c\ lnc - c + 1) \cdot 6000/m}$

$$\begin{aligned} m &= 11: \quad 0.784\ldots \\ m &= 12: \quad 0.000999\ldots \\ m &= 13: \quad 0.0000000760\ldots. \end{aligned}$$

切尔诺夫界证明

设$x_1, x_2, \cdots, x_n$是相互独立的随机变量，对任意i满足$0 \le x_i \le 1$,令 $X = x_1 + x_2 + \cdots + x_n$，对所有$c \ge 1$, $P[X \ge cE(X)] \le e^{-\beta(c)E(X)}$

证明：

$$P[X \ge cE(X)] = P\left[c^X \ge c^{cE(X)}\right]$$

$$\le E\left(c^X\right)/c^{cE(X)} \quad \text{(Markov Bound)}$$

$$\le e^{(c-1)E(X)}/c^{cE(X)} \quad \text{(等待证明)}$$

$$= e^{(c-1)E(X)}/e^{c\ln c E(X)} \quad (c^c = e^{c\ln c})$$

切尔诺夫界证明

证明: $E\left(c^X\right) \le e^{(c-1)E(X)}$

$$E\left(c^X\right) = E(c^{x_1} \cdots c^{x_n}) = E(c^{x_1}) \cdots E(c^{x_n})$$

由于: $E(c^{x_i}) = \sum c^r P(x_i = r) \le \sum (1 + (c-1)r)\, P(x_i = r)$

$$= \sum P(x_i = r) + (c-1) \sum r P(x_i = r)$$

$$= 1 + (c-1)\, E(x)$$

$$\le e^{(c-1)\, E(x)} \quad \text{由于} 1 + z \le e^z$$

$$E\left(c^X\right) = E(c^{x_1} \cdots c^{x_n}) = E(c^{x_1}) \cdots E(c^{x_n}) \le e^{(c-1)E(x_1)} \cdots e^{(c-1)E(x_n)}$$

$$\le e^{(c-1)E(X)}$$

边界比较

设$A_1, A_2, \cdots, A_n$是相互独立的事件，想知道多少事件可能会发生。

设x_i为事件A_i的指示器随机变量，p_i=P$(x_i = 1) = P(A_1)$,令$X = x_1 + x_2 + \cdots + x_n$，表示事件发生个数。

$$E(X) = \sum_{i=1}^{n} p_i \qquad Var(X) = \sum_{i=1}^{n} p_i(1-p_i) \qquad \Rightarrow \sigma_X = \sqrt{\sum_{i=1}^{n} p_i(1-p_i)}$$

马尔可夫定理： 对任意$c > 1$, $P[X \ge cE(X)] \le 1/c$

切比雪夫定理： 对任意$c > 1$, $P[|X - E(X)| \ge c\sigma_X] \le 1/c^2$

切诺夫界定理： 对任意$c > 1$, $P[X \ge cE(X)] \le e^{-(c\ln c - c + 1)E(X)}$

墨菲定律（Murphy’s law）

设$A_1, A_2, \cdots, A_n$是相互独立的事件，设x_i为事件A_i的指示器随机变量。事件发生总数$X = x_1 + x_2 + \cdots + x_n$

则 $P(X = 0) \le e^{-E(X)}$

证明：$P[X = 0] = P[\overline{A_1} \cap \overline{A_2} \cdots \cap \overline{A_n}] = \Pi_{i=1}^{n}(1 - P(A_i))$

$\le \Pi_{i=1}^{n} e^{-P(A_i)}$ $\quad 1 + x \le e^x$

$= e^{-\sum_{i=1}^{n} P(A_i)}$

$= e^{-\sum_{i=1}^{n} E(A_i)}$

$= e^{-E(X)}$

预计10个事件发生，至少有一个事件发生概率不小于1-$e^{-10} > 1 - 1/22000$