软件工程统计方法

统计量及其抽样分布

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

内 容 提 纲

- 统计学导论
- 描述统计
- 概率计算基础
- 随机变量及其分布
- 统计量及其抽样分布
- 参数估计
- 参数假设检验
- 非参数假设检验
- 方差分析
- 回归分析

本 节 内 容

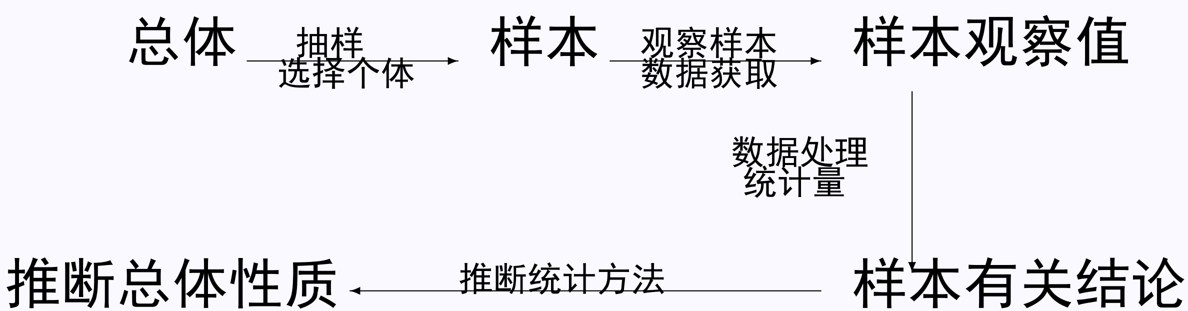
- 总体分布与样本分布
- 统计量与抽样分布
  - 切比雪夫不等式
  - 大数定理
  - 中心极限定理
- 重要抽样分布
  - χ^2分布
  - t分布
  - F分布

1 推断统计

- 统计推断是数理统计理论的主要部分。现行的统计推断理论, 是建立在概率论的基础上的。
- 只有在总体未知或总体分布已知, 但含有未知参数时, 统计推断才有意义。
- 例如, 假定在一大群人中, 身高服从正态分布$\mathbb{N}(\mu,\sigma^2)$, 其中μ,σ是未知参数, 即推断的对象。
  - 批量生产的一种电子元件, 在一定条件下, 我们假定元件寿命的概率分布为指数分布, 其概率密度为$f(x)=\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0$, 参数$\theta>0$未知, 为统计推断对象。

推断统计

- 统计推断：总体的概率分布$F(x;\theta)$含有未知参数θ, 从总体中随机抽样, 得样本值$x_1,\cdots,x_n$去获得对未知参数θ的了解, 包括估计问题和检验问题。
- R. A. Fisher把推断统计学的内容概括为：抽样分布、参数估计和假设检验。
- 推断统计学的基本步骤包括观察现象, 收集资料, 创建方法, 分析推断等。



2 抽样方法

在数据获取阶段，研究者通常需要考虑如有抽取一个好的样本：能够有效反映总体性质和合理的获取成本。好的样本也是相对而言的。一个样本对于某一研究目的是好的，但对于另外一个研究目的可能就是很糟糕的。在获取方式固定的前提下，高质量的数据和低成本的代价往往是矛盾的。在有限的资源下，尽可能获取高质量的数据，以提高统计结论的准确性。

抽样方法有很多，通常可以分为概率抽样和非概率抽样。非概率的抽样方法主要依据研究目的、研究对象和调查资源的受限情况制定的方法，不依据某种随机原则。概率抽样方法，也称随机抽样。随机抽样是随机并不“随便”，后面具有严格的科学含义。

抽样方法

- 简单随机抽样(Simple Random Sampling)
  简单随机抽样是指从一个单元数为N(可能很大)的总体中逐个等概率无放回抽取个体，直至达到需要的n个个体为止。简单随机抽样是最简单的一种抽样方法，也是其它概率抽样方法的一个基础。
- 系统抽样(Systematic Sampling)
  当总体中的个体数较多时，可将总体分成均衡的几个部分，然后按预先定出的规则，从每一部分抽取一个个体，得到需要的样本，这种抽样叫做系统抽样。系统抽样方法采用随机的方式将总体中的个体编号。为简便起见，有时可直接采用个体所带有的号码，如考生的准考证号，街道上各户的门牌号，等等。

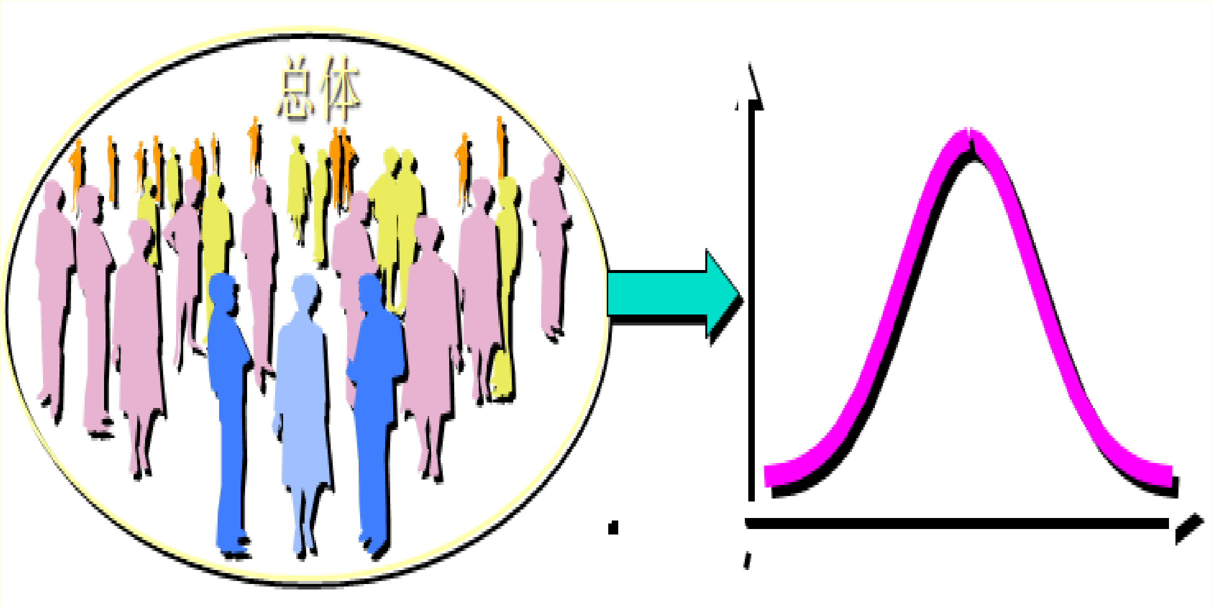
抽样方法

- 整群抽样(Cluster Sampling)
  整群抽样先对总体聚类分组, 这样的组称为群, 再随机抽取群, 被抽中的群的所有个体组成样本。整群抽样时只需要把群作为抽样框, 而不需要把数量庞大的个体作为抽样框, 因此能大大降低抽样的成本, 提高抽样效率。
- 分层抽样(Stratified Sampling)
  分层抽样先按对观察指标影响较大的某种特征, 将总体分为若干个类别(层), 再从每一层内随机抽取一定数量的个体, 合起来组成样本。分层是从分布不均匀的研究人群中抽取有代表性样本的方法。先按照某些人口学特征或某些标志将研究人群(如行政区、不同高校等)分为若干组, 然后从每层抽取一定数量的随机样本。

3 总体分布

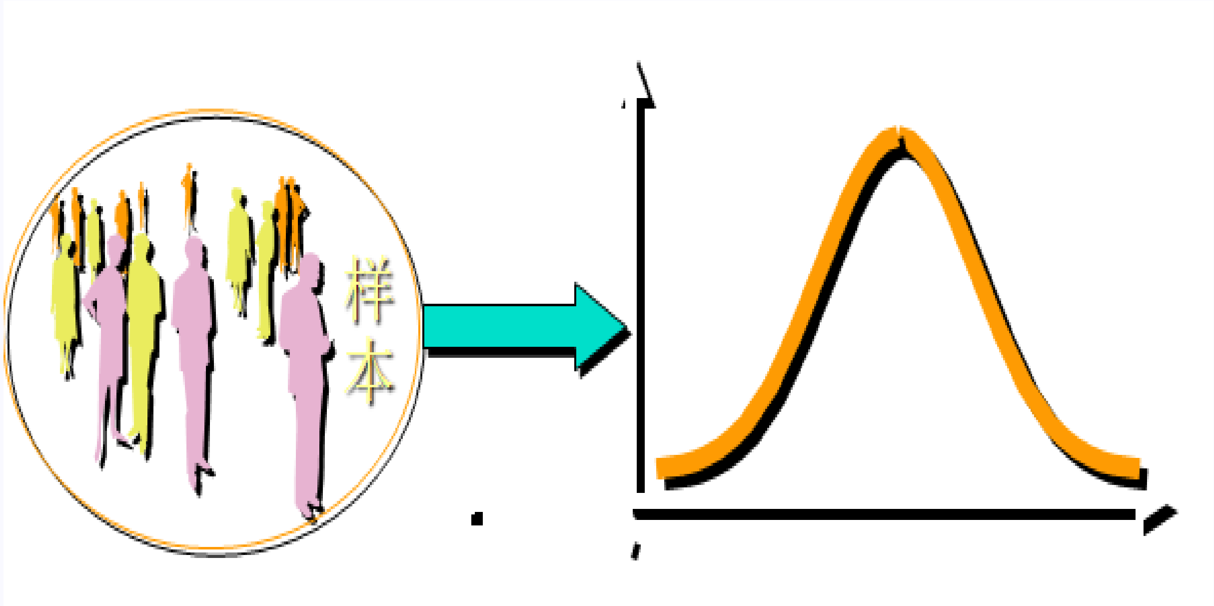
推断统计学中，我们需要对比理解三个容易混淆的概念：总体分布、样本分布和抽样分布。

- 总体分布：总体中所有个体观察值所构成的分布。
  - 注意：不一定是正态分布。
- 总体分布通常是未知的：
  - 分布形式和参数都未知。
  - 分布形式已知，参数未知。



4 样本分布

- 一个样本中个体观察值的分布。
- 样本分布通常也称经验分布。
- 当样本容量n逐渐增大时, 样本分布逐渐接近总体的分布。当n为总体中个体数量时与总体分布完全一致。



频率分布与经验分布

总体的分布函数也叫理论分布函数。统计学的一个重要问题是利用样本来估计或者推断总体的分布函数。那么我们首先需要表达样本分布。

设$X_1,\cdots,X_n$是样本, x_i是样本值, 将x_i排列合并相同元素得x_j^*, m_j为x_j^*的出现次数, 构造样本分布。

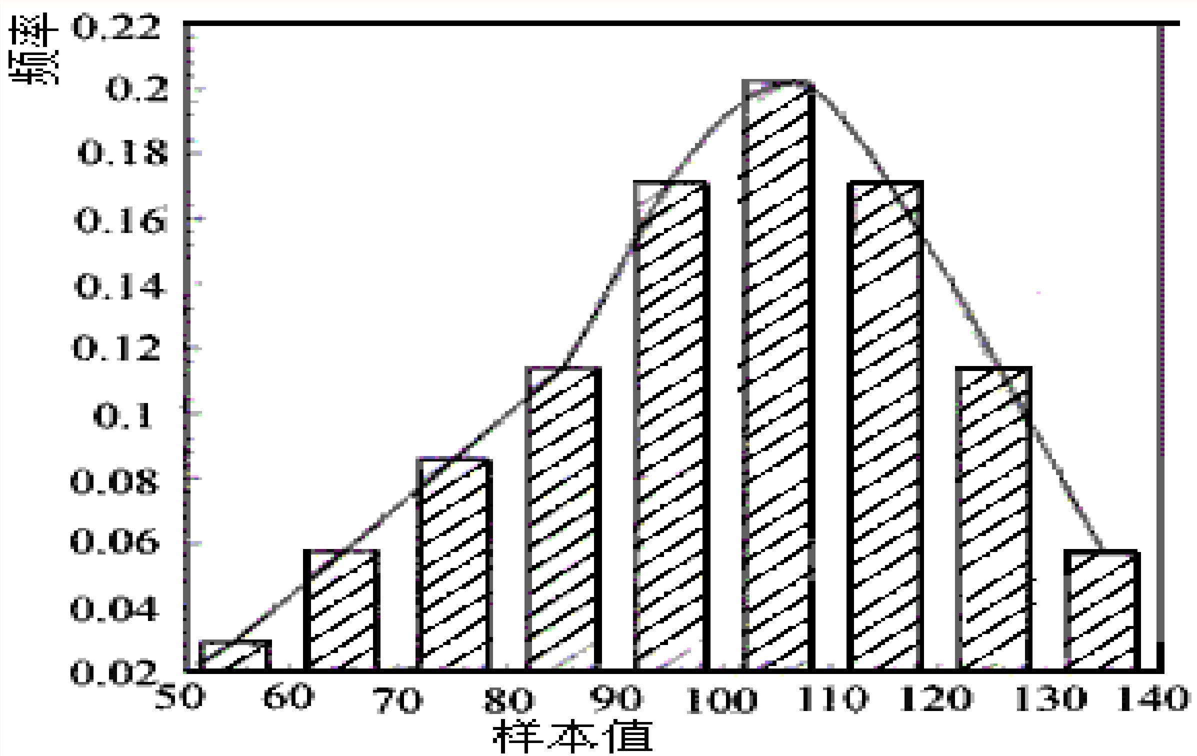
- 频数和频率分布:

	x_1^*	$\cdots$	x_n^*
频数m_j	m_1	$\cdots$	m_k
频率$\frac{m_j}{n}$	$\frac{m_1}{n}$	$\cdots$	$\frac{m_k}{n}$

- 经验分布函数:

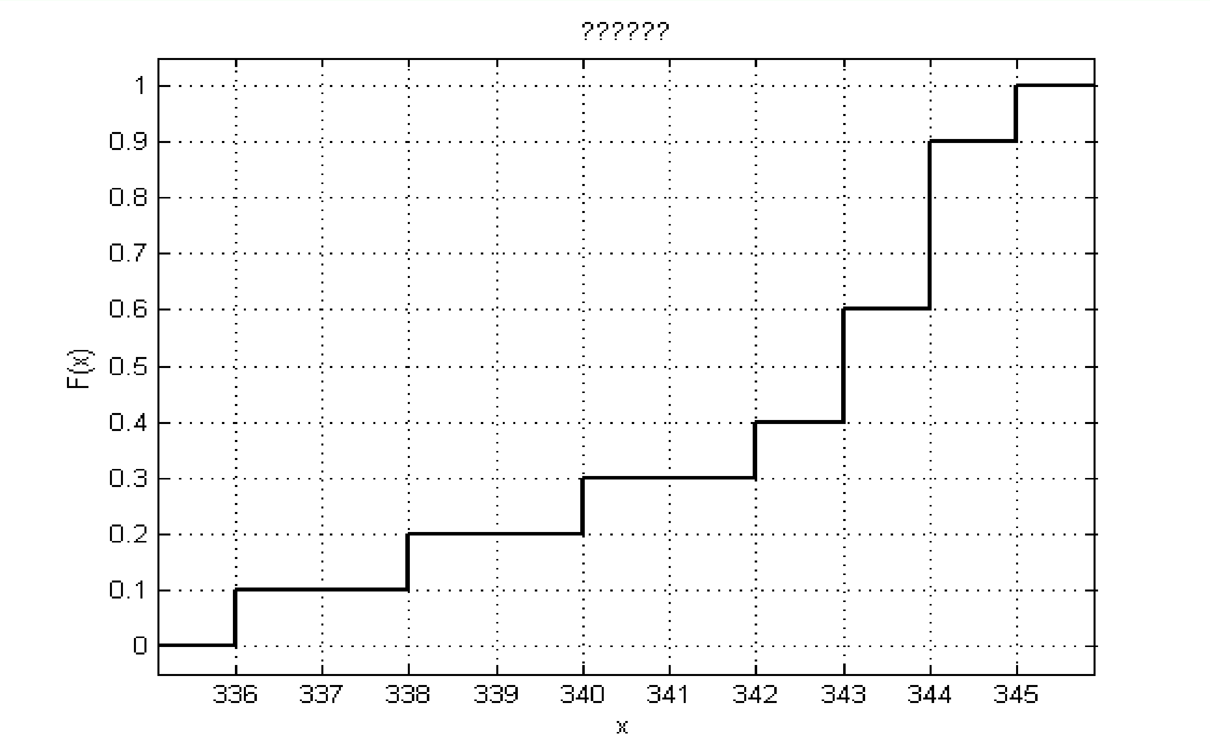
$$F(x)=\begin{cases} 0 & x<x_1^* \\ \sum_{j=1}^{l}\frac{m_l}{n} & x_l^*\le x<x_{l+1}^* \\ 1 & x>x_k^* \end{cases}$$

频率分布图



思考：当样本容量n逐渐增大时, 频数/频率分布图将如何变化?

经验分布图



思考：当样本容量n逐渐增大时，经验分布图将如何变化?

5 统计量

统计量是样本的函数, 它依赖且只依赖于样本$x_1, x_2, \cdots, x_n$。统计量只依赖于样本, 而不能依赖于任何其他未知的量, 特别是它不能依赖于总体分布中所包含的未知参数。样本矩一类常用的样本统计量。

Definition 1 (样本矩, Sample Moment) 设样本$x_1, \cdots, x_n$,对自然数k, 样本的k阶矩定义如下:

- 样本k阶原点矩:

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{1}$$

- 样本k阶中心矩:

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \tag{2}$$

统 计 量

许多最常用的统计量，都可由样本矩构造。

- 样本均值为样本的一阶原点矩A_1, 它代表样本的平均程度, 记为$\overline{X}$。

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{3}$$

- 样本方差为修正后的二阶中心矩, 即$\frac{n}{n-1}B_2$, 它代表样本的分散程度, 记为S^2。

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{4}$$

- 样本偏度为三阶中心矩和二阶中心矩比值

$$S^3=\frac{B_3}{B_2^{\frac{3}{2}}} \tag{5}$$

- 样本峰度为四阶中心矩和二阶中心矩比值

$$S^4=\frac{B_4}{B_2^2} \tag{6}$$

统计量相关定理

Lemma 1 设$X_1,\cdots,X_n$为某总体的一样本, $g(x)$为x的函数且$E(g(X_i))$和$Var(g(X_i))$存在, 则

$$E(\sum_{i=1}^{n} g(X_i)) = n(E(g(X_1))) \tag{7}$$

$$Var(\sum_{i=1}^{n} g(X_i)) = n(Var(g(X_1))) \tag{8}$$

Theorem 1 设某总体的均值为μ, 方差为σ^2。$X_1,\cdots,X_n$为的总体一样本, 则

(1) $E(\overline{X}) = \mu$

(2) $Var(\overline{X}) = \frac{\sigma^2}{n}$

(3) $E(S^2) = \sigma^2$

(1)(2)是关于均值的抽样分布的理论基础。(3)是S^2作为方差无偏估计的理论基础。

抽 样 分 布

- 样本统计量是随机变量。
  - 样本均值$\overline{X}$, 样本方差s^2等。
- 样本统计量的概率分布是一种理论分布, 称为抽样分布。
  - 在重复选取容量为n的样本时, 由该统计量的所有可能取值形成的频数分布。
  - 抽样分布来自容量相同的所有可能样本。
- 抽样分布提供了样本统计量长远而稳定的信息, 是进行推断的理论基础, 也是抽样推断科学性的重要依据。

6 均值抽样分布

在重复选取容量为n的样本时, 由样本均值的所有可能取值形成的频数分布, 是一种理论分布。

Example 1 设一总体为四本书, 四本书平均每页的错别字为: $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$。

则总体的均值如下:

$$\mu = \frac{\sum_{i=1}^{4} x_i}{4} = 2.5 \tag{9}$$

总体的方差如下:

$$\sigma^2 = \frac{\sum_{i=1}^{4}(x_i - \mu)^2}{4} = 1.25 \tag{10}$$

均值抽样分布

现从总体中抽取$n=2$的简单随机样本, 在有放回抽样条件下, 共有16个样本。

所有样本的结果为

1, 1	1, 2	1, 3	1, 4
2, 1	2, 2	2, 3	2, 4
3, 1	3, 2	3, 3	3, 4
4, 1	4, 2	4, 3	4, 4

则样本均值$\overline{X}$的所有结果为

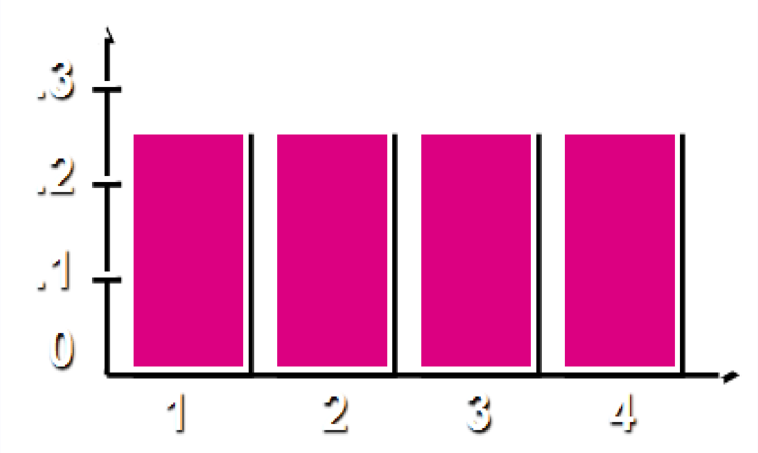
1	1.5	2	2.5
1.5	2	2.5	3
2	2.5	3	3.5
2.5	3	3.5	4

样本均值的频数

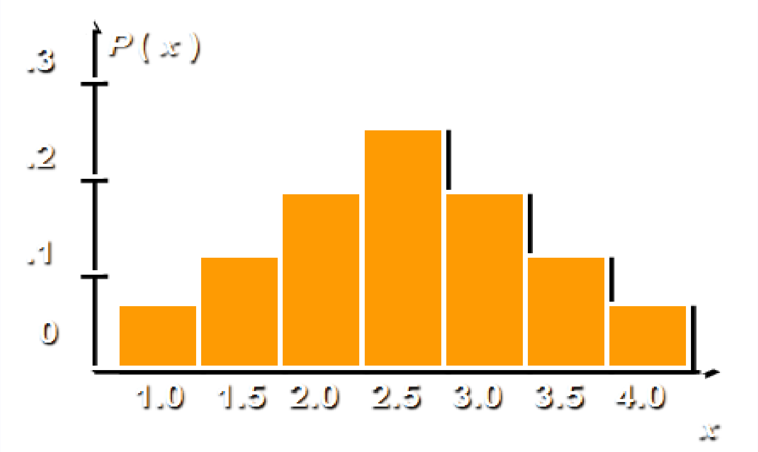
$\overline{x}$	1	1.5	2	2.5	3	3.5	4
f_i	1	2	3	4	3	2	1

均值抽样分布

总体分布



抽样分布



例 子(有放回抽样)

设总体为$\{1,2,3,4\}$, 则总体的均值和方差分别为$\mu = 2.5$和$\sigma^2 = 1.25$。从总体中抽取$n = 2$的简单随机样本, 在有放回抽样条件下, 共有16个样本。

所有样本的结果和相应样本均值

1, 1	1, 2	1, 3	1, 4
2, 1	2, 2	2, 3	2, 4
3, 1	3, 2	3, 3	3, 4
4, 1	4, 2	4, 3	4, 4

1	1.5	2	2.5
1.5	2	2.5	3
2	2.5	3	3.5
2.5	3	3.5	4

样本均值$\overline{X}$的均值为

$$\frac{1}{16}\sum_{i=1}^{16}\overline{x}_i = 2.5$$

样本均值$\overline{X}$的方差为

$$\frac{1}{16}\sum_{i=1}^{16}(\overline{x}_i - 2.5)^2 = \frac{5}{8} = 0.625$$

例 子(无放回抽样)

设总体为$\{1,2,3,4\}$, 则总体的均值和方差分别为$\mu=2.5$和$\sigma^2=1.25$。从总体中抽取$n=2$的简单随机样本, 在无放回抽样条件下, 共有12个样本。

所有样本的结果和相应样本均值

-	1, 2	1, 3	1, 4
2, 1	-	2, 3	2, 4
3, 1	3, 2	-	3, 4
4, 1	4, 2	4, 3	-

-	1.5	2	2.5
1.5	-	2.5	3
2	2.5	-	3.5
2.5	3	3.5	-

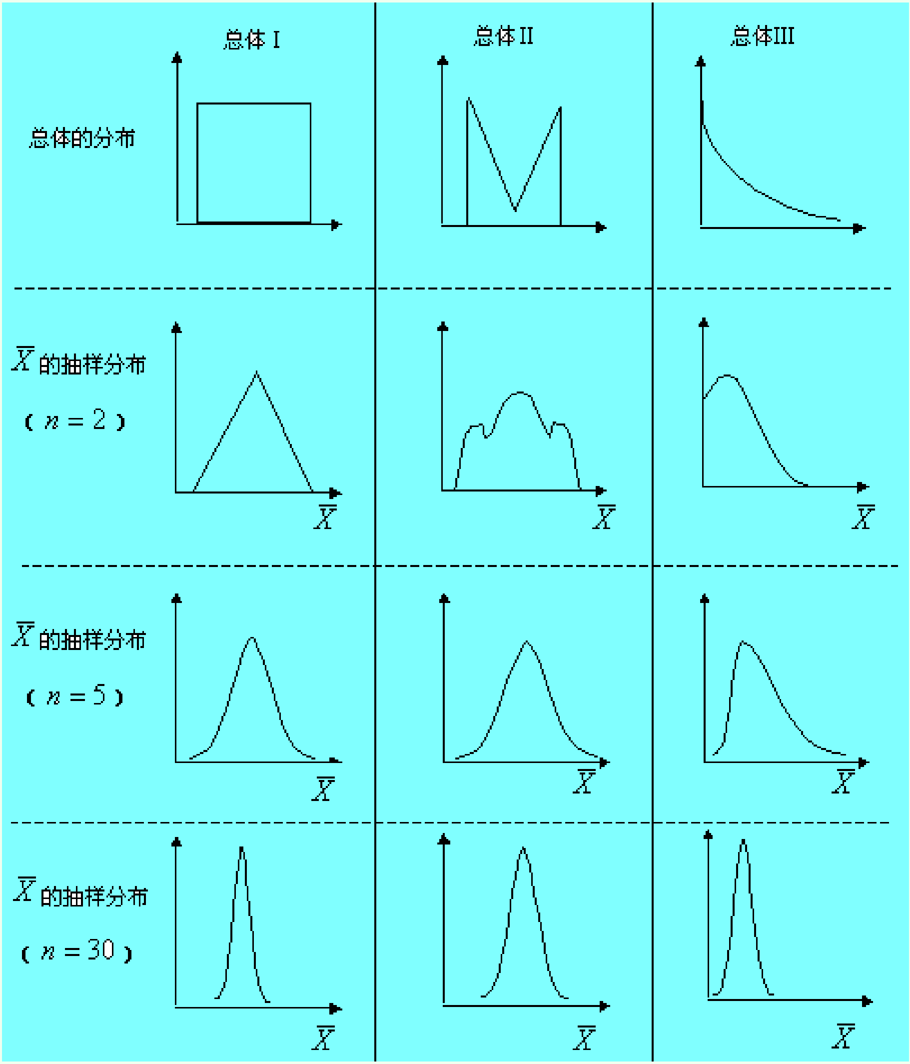
样本均值$\overline{X}$的均值为$\frac{1}{12}\sum_{i=1}^{12}\overline{x}_i=2.5$

样本均值$\overline{X}$的方差为

$$\frac{1}{12}\sum_{i=1}^{12}(\overline{x}_i-2.5)^2=\frac{5}{12}=\frac{5}{8}\cdot\frac{4-2}{4-1}$$

注意：无放回抽样时样本均值的方差等于有放回抽样时的方差乘以有限总体校正系数$FPC=\frac{N'-n'}{N'-1}$, 其中N'为总体中个体数目, n'为每次抽样的个体数目。当n'远小于N'(经验上为$\frac{n}{N}<0.05$)时, 有限总体不放回抽样等同于放回抽样, 此时可以忽略有限总体校正系数。

抽样分布



总体 I
总体 II
总体 III
总体的分布
$\overline{X}$ 的抽样分布
（ $n=2$ ）
$\overline{X}$ 的抽样分布
（ $n=5$ ）
$\overline{X}$ 的抽样分布
（ $n=30$ ）
$\overline{X}$

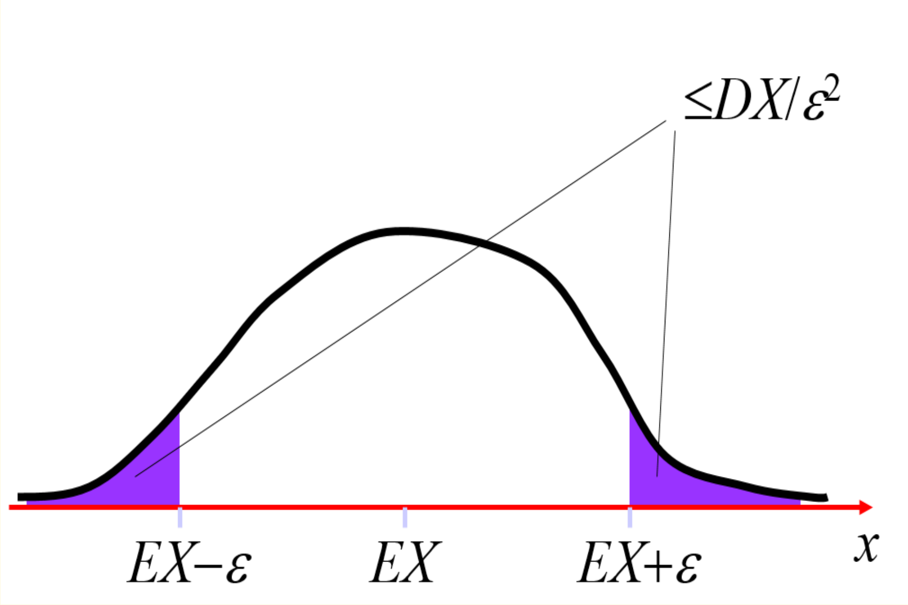
7 切比雪夫不等式

Theorem 2 (切比雪夫不等式) 设随机变量具有数学期望$E(X) = \mu$, 方差$Var(X) = \sigma^2$, 则对于任意$\varepsilon > 0$, 有

$$P\{|X-\mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} \tag{11}$$

即等价于

$$P\{|X-\mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{\varepsilon^2} \tag{12}$$



切比雪夫不等式

证明: 设X的概率密度为$f(x)$, 则

$$
\begin{aligned}
& P\{|X-\mu| \geq \varepsilon\} \\
& = \int_{|x-\mu|\geq\varepsilon} f(x)dx \\
& \leq \int_{|x-\mu|\geq\varepsilon} \frac{(x-\mu)^2}{\varepsilon^2} f(x)dx \\
& \leq \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \\
& = \frac{Var(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}
\end{aligned}
$$

注明：比雪夫不等式表明$|X-\mu| \geq \varepsilon$的概率由σ控制(ε给定)。

切比雪夫不等式

Example 2 设X是掷一颗骰子所出现的点数，若给定ε=1,2, 试计算$P(|X-EX|\geq\varepsilon)$, 并对比切比雪夫不等式。

$$P(X=k)=\frac{1}{6},k=1,2,3,4,5,6$$

$$E(X)=\frac{7}{2},E(X^2)=\frac{91}{6},D(X)=\frac{35}{12}$$

$$P(|X-EX|\geq 1)=\frac{2}{3}\leq\frac{D(X)}{\varepsilon}=\frac{35}{12}$$

$$P(|X-EX|\geq 2)=\frac{1}{3}\leq\frac{D(X)}{\varepsilon}=\frac{35}{48}$$

切比雪夫不等式

Example 3 已知事件A的随机变量$X \sim \mathbb{B}(n, 0.75)$, 估计事件$A$发生频率在$0.74-0.76$之间的概率大于*0.90*的最小实验次数$n$.

$$E(X) = np = 0.75n, Var(X) = np(1-p) = 0.1875n$$

$$f_n(A) = \frac{X}{n}$$

$$P\{0.74 < \frac{X}{n} < 0.76\} = P\{|X - 0.75n| < 0.01n\}$$

$$\geq 1 - \frac{0.1875n}{(0.01n)^2} = 1 - \frac{1875}{n} \geq 0.90$$

解得$n \geq 18750$。

切比雪夫不等式

Example 4 设某大楼有*10000*盏电灯，夜晚每一盏灯开灯的概率是*0.7*。假定开关时间彼此独立，估计夜晚同时开着的灯数在*6800*与*7200*之间的概率.

8 大数定律

大数定律又称大数法则、大数律，是个数学与统计学的概念，意指数量越多，则其平均就越趋近期望值。人们发现，在重复试验中，随着试验次数的增加，事件发生的频率趋于一个稳定值；人们同时也发现，在对物理量的测量实践中，测定值的算术平均也具有稳定性。辛钦定理和伯努利大数定理都概括了这一现象，都称为大数定律。

Theorem 3 (伯努利大数定理) 设x_n是n重伯努利试验中事件A出现的次数，p是事件在每次试验中A出现的概率, 则对任意的$\varepsilon > 0$,有

$$\lim_{n\to+\infty} P(|\frac{x_n}{n} - p| < \varepsilon) = 1 \tag{13}$$

证明: $x_n \sim \mathbb{B}(n,p)$, 则

$$E(\frac{x_n}{n}) = \frac{1}{n}E(n_A) = \frac{np}{n} = p, Var(\frac{x_n}{n}) = \frac{1}{n^2}Var(x_n) = \frac{pq}{n}$$

$$\forall\varepsilon > 0 : P\{|\frac{x_n}{n} - p| < \varepsilon\} \geq 1 - \frac{pq}{n\varepsilon^2}$$

伯努利大数定理建立了在大量重复独立试验中事件出现频率的稳定性, 正因为这种稳定性, 概率的概念才有客观意义。

大 数 定 律

历史上最早的大数定律是伯努利在1713年建立的。概率论的研究到现在约有300多年的历史，最终以事件的频率稳定值来定义其概率。

作为概率这门学科的基础，其“定义“的合理性这一悬而未决的带根本性的问题,由伯努利于1713年发表的这个“大数定律”给予了解决，被称为概率论的第一篇论文,为概率论的公理化体系奠定了理论基础。

之所以被成为“定律”,是这一规律表述了一种全人类多年的集体经验.因此，对后来的类似定理统称为大数“定律”。

推断统计
抽样方法
总体分布
样本分布
统计量
均值抽样分布
切比雪夫不等式
大数定律
中心极限定理
χ^2 分布函数
t 分布函数
F 分布函数

第 30 页 共 100 页
返 回
全 屏
关 闭

大 数 定 律

其它相关的大数定理。

Theorem 4 (独立同分布的大数定律) 设$X_1, X_2, \cdots, X_n,$ 是相互独立有相同分布的随机变量序列,各有数学期望$EX_i = \mu, (i = 1, 2, \cdots)$ 方差有$DX_i = \sigma^2 (i = 1, 2, \cdots)$,则对任意的$\varepsilon > 0$,有

$$\lim_{n \to +\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon\} = 1 \tag{14}$$

Theorem 5 (切比雪夫大数定律) 设$X_1, X_2, \cdots, X_n,$ 是相互独立的随机变量序列,各有数学期望$EX_i (i = 1, 2, \cdots)$ 和有限的方差,并且方差有$DX_i (i = 1, 2, \cdots)$ 共同的上界，即$DX_i \le c$ 则对任意的$\varepsilon > 0$,有

$$\lim_{n \to +\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i| < \varepsilon\} = 1 \tag{15}$$

大 数 定 律

Theorem 6 (辛钦大数定理) 设$X_1, X_2, \cdots, X_n,$ 是独立同分布的随机变量序列,只要数学期望$EX_i = \mu (i = 1, 2, \cdots)$ 存在, 则对任意的$\varepsilon > 0$,有

$$\lim_{n \to +\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon\} = 1 \tag{16}$$

伯努利大数定理是辛钦大数定理的特殊情况。

大数定律提供了通过试验来确定事件概率的方法, 既然频率$\frac{n_A}{n}$与概率p有较大偏差的可能性很小, 我们便可以通过做试验确定某事件发生的频率并把它作为相应的概率估计。这种方法即是在后面要介绍的参数估计法, 参数估计的重要理论基础之一就是大数定律。

9 中心极限定理

中心极限定理，是概率论中讨论随机变量和的分布以正态分布为极限的一组定理。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量之和近似服从正态分布的条件。

中心极限定理有着有趣的历史。这个定理的第一版被法国数学家棣莫弗发现，他在1733年发表的卓越论文中使用正态分布去估计大量抛掷硬币出现正面次数的分布。这个超越时代的成果险些被历史遗忘，所幸著名法国数学家拉普拉斯在1812年发表的巨著中拯救了这个默默无名的理论.拉普拉斯扩展了棣莫弗的理论，指出二项分布可用正态分布逼近。但同棣莫弗一样，拉普拉斯的发现在当时并未引起很大反响。直到十九世纪末中心极限定理的重要性才被世人所知。

Theorem 7 (德莫佛—拉普拉斯中心极限定理) 设$Y_n \sim \mathbb{B}(n,p), 0 < p < 1, n = 1, 2, \cdots$, 则对任意实数$x$, 有

$$\lim_{n\to+\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_x^{-\infty} e^{-\frac{t^2}{2}} dt \tag{17}$$

例子

Example 5 设某工厂有*400*台同类机器，各台机器发生故障的概率都是*0.02*，各台机器工作是相互独立的. 试求机器出故障的台数不小于*2*的概率.

解:设故障台数为X,则$X \sim \mathbb{B}(400, 0.02)$.
(1) 二项分布计算

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.9972$$

(2) 泊松分布计算

$$\lambda = np = 400 * 0.02 = 8$$

查表得

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.000335 - 0.002684 \approx 0.9969$$

(3)正态分布计算

$$\sqrt{npq} = \sqrt{400 * 0.02 * 0.98} = 2.8$$

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \Phi(\frac{1 - np}{\sqrt{npq}}) = \Phi(\frac{7}{2.8}) = 0.9938$$

例 子

Example 6 某保险公司的老年人寿保险有*1*万人参加，每人每年交*200*元. 若老人在该年内死亡，公司付给受益人*1*万元。设老年人死亡率为*0.017*，试求

(1) 保险公司在一年内这项保险亏本的概率。

(2) 保险公司在一年内这项保险盈利10万元以上的概率。

中心极限定理

中心极限定理虽然表述形式简洁，但是严格证明它却非常困难。中心极限定理就像一张大蜘蛛网，棣莫弗和拉普拉斯编织了它的雏形，可是这张网上漏洞太多，一个多世纪来，数学家们就像蜘蛛一样前赴后继，努力想把所有的漏洞都补上。在十九世纪，泊松(Poission)、狄利克莱(Dirichlet)、柯西(Cauchy)、贝塞尔(Bessel)这些大蜘蛛都曾经试图对把这张网上的漏洞补上。

把漏洞补上的严格方案的雏形是从切比雪夫1887年的工作开始的，不过切比雪夫的证明存在一些漏洞。马尔科夫和李雅普诺夫都是切比雪夫的学生，马尔科夫沿着老师的基于矩法的思路在蜘蛛网上辛勤编织，但洞还是补得不够严实；李雅普诺夫不像马尔可夫那样深受老师的影响，他沿着拉普拉斯当年提出的基于特征函数的思路，于1901年给出了一个补洞的方法，切比雪夫对这个方法大加赞赏，李雅普诺夫的证明被认为是第一个在一般条件下的严格证明。

中心极限定理

1901年，俄国数学家里李雅普诺夫用更普通的随机变量定义中心极限定理并在数学上进行了精确的证明。

Theorem 8 (独立同分布的中心极限定理) 设$X_1, X_2, \cdots, X_n$, 是相互独立有相同分布的随机变量序列,且有期望和方差:$E(X_i) = \mu, Var(X_i) = \sigma^2 (i = 1, 2, \cdots)$ 则对任意实数x,

$$\lim_{n\to\infty} P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le x) = \frac{1}{\sqrt{2\pi}} \int_x^{-\infty} e^{-\frac{t^2}{2}} dt = \Phi(x) \tag{18}$$

中心极限定理说明, 对于任意分布, 当样本容量足够大时, 样本均值的抽样分布近似正态分布$N(\mu, \sigma^2/n)$。样本容量足够大的数值通常还与原始总体分布与正态分布的相近程度。原始总体分布与正态分布越相近, 对样本容量的要求越低。在总体分布信息未知的情况下, 经验上通常$n \ge 30$认为样本容量足够大, 更严格的要求则为$n \ge 50$。特别地, 当原始总体服从正态分布时, 独立同分布的正态样本均值服从正态分布(此时并不要求$n \to \infty$)。

Theorem 9 (独立同分布的正态样本) 设$X_1, \cdots, X_n$是总体$N(\mu, \sigma^2)$的样本, $\overline{X}$和S^2分别是样本均值和样本方差, 则有

$$\overline{X} \sim N(\mu, \sigma^2/n) \tag{19}$$

例 子

Example 7 设某种电器元件的寿命服从均值为*100*小时的指数分布, 现随机取得*16*只, 设它们的寿命是相互独立的, 求这*16*只元件的寿命的总和大于*1920*小时的概率.

解: 记16只元件的寿命分别为$X_1, \cdots, X_{16}$, 则16只电器元件的寿命总和为$X = \sum_{i=1}^{16} X_i$. 由题意知

$$E(X_i) = 100, Var(X_i) = 100^2$$

根据中心极限定理

$$Y = \frac{X - 16 * 100}{4 * 100} = \frac{X - 1600}{400} \sim \mathbb{N}(0, 1)$$

$$P(X > 1920) = 1 - P(X \leq 1920) \approx 1 - \Phi(\frac{1920 - 1600}{400})$$
$$= 1 - \Phi(0.8) = 0.2119$$

重要抽样分布

推断统计中三大抽样分布：

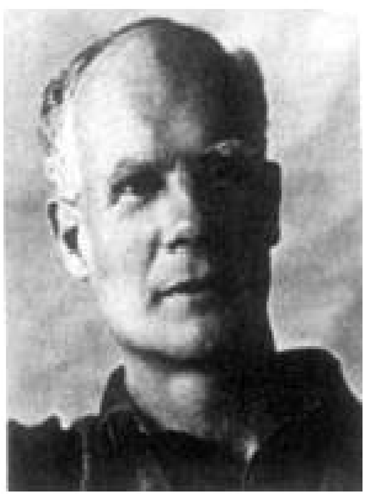
- χ^2-分布
- t-分布
- F-分布

Karl Pearson　W. S. Gosset　R. A. Fisher

重要抽样分布

推断统计中三大抽样分布：

- χ^2-分布
- t-分布
- F-分布



Karl Pearson

W. S. Gosset

R. A. Fisher

10 χ^2 分布函数

第一位剑客就是卡尔·皮尔逊(Karl Pearson)，手中的宝剑就是χ^2分布。

χ^2分布这把宝剑最早的锻造者其实是物理学家麦克斯韦，他在推导空气分子的运动速度的分布的时候，发现分子速度在三个坐标轴上的分量是正态分布，而分子运动速度的平方符合自由度为3的χ^2分布。

麦克斯韦虽然造出了这把宝剑，但是真正把它挥舞得得心应手、游刃有余的是皮尔逊。在分布曲线和数据的拟合优度检验中，χ^2分布可是一个利器，而皮尔逊的这个工作被认为是假设检验的开山之作。皮尔逊继承了高尔顿的衣钵，统计功力深厚，在19世纪末20世纪初很长的一段时间里，一直被数理统计武林人士尊为德高望重的第一大剑客。

χ^2 分布函数

Definition 2 设$X_1, \cdots, X_n$相互独立, 且$X_i \sim N(0,1)$, 则称随机变量

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

所服从的分布为自由度为n的χ^2分布, 记为$\chi^2(n)$。

$\chi^2 \sim \chi^2(n)$, 则其概率密度为

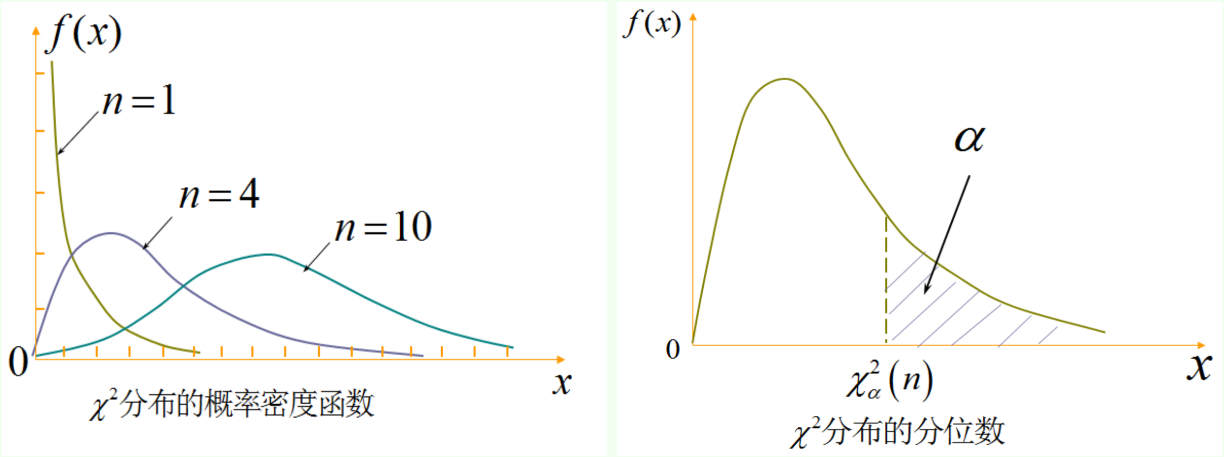
$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & : \ x > 0 \\ 0 & : \ x \le 0 \end{cases}$$

其中$\Gamma(x)$为伽马函数, 定义为

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt, x > 0$$

$$\Gamma(s+1) = s\Gamma(s), \Gamma(n+1) = n!$$

χ^2 分布性质



Theorem 10 $\chi^2 \sim \chi^2(n)$, 则

$$E(\chi^2) = n, D(\chi^2) = 2n$$

Theorem 11 设$X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, 且X_1, X_2相互独立, 则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

对于给定的$0 < \alpha < 1$, 其上分位点$\chi^2_\alpha(n)$可以通过χ^2分布表查询可得。

正态样本均值和方差的分布

Theorem 12 设$X_1, \cdots, X_n$是总体$\mathbb{N}(\mu, \sigma^2)$的样本, 样本均值和样本方差分别为$\overline{X}$和$s^2$, 则有

(1) 均值抽样分布

$$\overline{X} \sim \mathbb{N}(\mu, \frac{\sigma^2}{n})$$

(2) 方差抽样分布

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

例 子

Example 8 设总体$X \sim N(\mu, \sigma^2)$, 其中μ, σ已知. $X_1, \cdots, X_n$是取自总体X的样本, 求

(1) 统计量$X' = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2$的分布。

(2) 设$n = 5$, $a(X_1 - X_2)^2 + b(2X_3 - X_4 - X_5)^2 \sim \chi^2(k)$, 则$a, b, k$分别为多少?

解: (1)令$Y_i = \frac{X_i - \mu}{\sigma}$, 显然$Y_i$独立且$Y_i \sim N(0,1)$, 所以

$$X' = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n Y_i^2 \sim \chi^2(n)$$

(2)$X_1 - X_2 \sim N(0, 2\sigma^2)$, 所以$\frac{(X_1 - X_2)^2}{2\sigma^2} \sim \chi^2(1)$
$2X_3 - X_4 - X_5 \sim N(0, 6\sigma^2)$, 所以$\frac{(2X_3 - X_4 - X_5)^2}{6\sigma^2} \sim \chi^2(1)$
$X_1 - X_2$与$2X_3 - X_4 - X_5$独立, 则有

$$\frac{(X_1 - X_2)^2}{2\sigma^2} + \frac{(2X_3 - X_4 - X_5)^2}{6\sigma^2} \sim \chi^2(2)$$

所以$a = \frac{1}{2\sigma^2}, b = \frac{1}{6\sigma^2}, k = 2$。

11 t 分布函数

第二位剑客是戈塞特(W.S.Gosset)，笔名是大家都熟悉的学生氏(Student)，而他手中的宝剑是t分布。

戈塞特是化学、数学双学位，依靠自己的化学知识进酿酒厂工作，工作期间考虑酿酒配方实验中的统计学问题，追随卡尔·皮尔逊学习了一年的统计学，最终依靠自己的数学知识打造出了t分布这把利剑而青史留名。

1908年，戈塞特提出了正态样本中样本均值和标准差的比值的分布，并给出了应用上极其重要的第一个分布表。戈塞特在t分布的工作是开创了小样本统计学的先河。

t 分布函数

Definition 3 设$X \sim N(0;1)$, $Y \sim \chi^2(n)$, 且X, Y相互独立, 则称随机变量

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

所服从的分布为自由度为n的t分布, 记为$t(n)$.

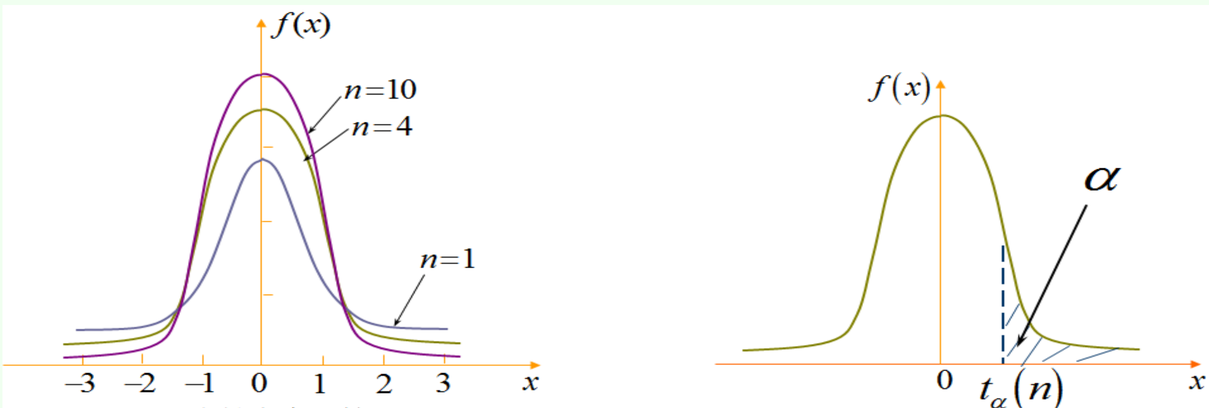
设随机变量$T \sim t(n)$, 则T的概率密度为

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}, -\infty < t < +\infty$$

$$\lim_{n\to\infty} f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$$

即n足够大(经验上$n \geq 30$)时, t分布近似于$N(0,1)$分布。

t 分 布 性 质



t分布的密度函数　　　　t分布的分位数

对于给定的$0 < \alpha < 1$, 其上分位点$t_\alpha(n)$可以通过t分布表查询可得。由t分布的对称性可知$t_{1-\alpha}(n) = -t_\alpha(n)$。

t 分布性质

Theorem 13 设$X_1, X_2, \cdots, X_n$是总体$N(\mu, \sigma^2)$的样本, $\overline{X}, S^2$分别是样本均值和样本方差, 则

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

由$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$可知

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1)$$

12 F 分布函数

第三位剑客是费希尔(R.A.Fisher)，手持F分布这把宝剑，在一片荒芜中开拓出方差分析的肥沃土地。

F分布就是为了纪念费希尔而用他的名字首字母命名的。费希尔剑法飘逸，在三位剑客中当属费希尔的天赋最高，各种兵器的使用都得心应手。费希尔统计造诣极高，受高斯的启发，系统地创立了极大似然估计剑法，这套剑法现在被尊为统计学参数估计中的第一剑法。

F 分布函数

Definition 4 设$X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且X, Y相互独立, 则称随机变量

$$F = \frac{X/n_1}{Y/n_2} = \frac{n_2}{n_1}\frac{X}{Y}$$

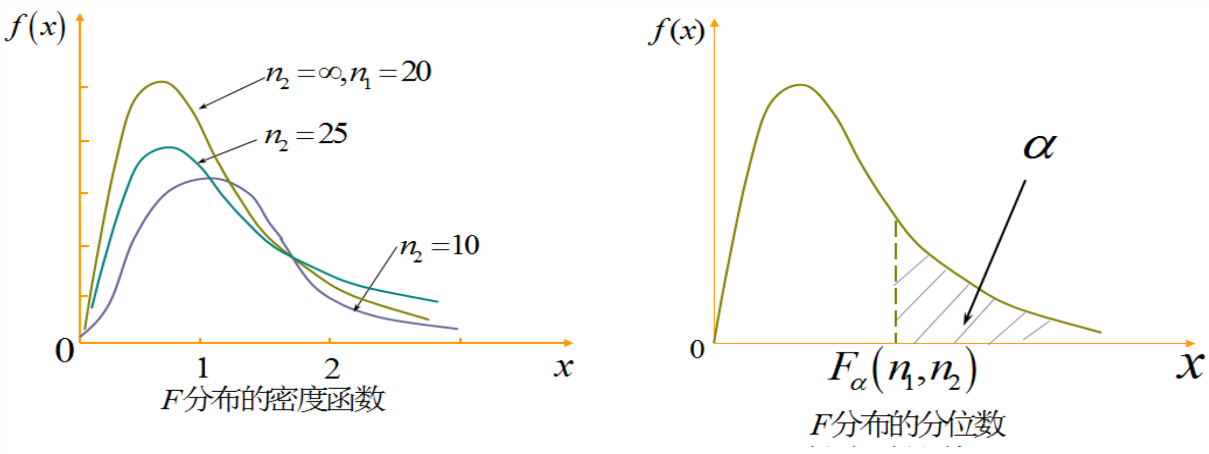
所服从的分布为自由度为(n_1, n_2)的F分布, 记为$F(n_1, n_2)$, 其中n_1为第一自由度, n_2为第二自由度.

设随机变量$F \sim F(n_1, n_2)$, 则F的概率密度为

$$f(x) = \begin{cases} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}(\frac{n_1}{n_2})(\frac{n_1}{n_2}x)^{\frac{n_1}{2}-1}(1+\frac{n_1}{n_2}x)^{-\frac{n_1+n_2}{2}} & : \ x > 0 \\ 0 & : \ x \leq 0 \end{cases}$$

若$F \sim F(n_1, n_2)$, 则$F^{-1} \sim F(n_2, n_1)$.

F 分布性质



对于给定的$0<\alpha<1$, 其上分位点$F_\alpha(n_1,n_2)$可以通过F分布表查询可得.

$$F_{1-\alpha}(n_1,n_2)=\frac{1}{F_\alpha(n_2,n_1)}$$

F 分布性质

Theorem 14 设$X_1, X_2, \cdots, X_{n_1}$与$Y_1, Y_2, \cdots, Y_{n_2}$是分别来自具有相同方差的两正态总体$N(\mu_1, \sigma^2)$， $N(\mu_2, \sigma^2)$的两个独立样本, s_1^2, s_2^2分别为两个样本的方差, 则

$$\frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1) \tag{20}$$

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}} \sim N(0, 1)) \tag{21}$$

F 分布性质

Theorem 15 设$X_1, X_2, \cdots, X_{n_1}$与$Y_1, Y_2, \cdots, Y_{n_2}$是分别来自两正态总体$N(\mu_1, \sigma_1^2)$，$N(\mu_2, \sigma_2^2)$的两个独立样本, $\overline{X}, \overline{Y}$分别为两个样本的均值, s_1^2, s_2^2分别为两个样本的方差, 则

- $$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

- 当$\sigma_1^2 = \sigma_2^2 = \sigma^2$时

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{s_w \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

其中

$$s_w = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

内 容 回 顾

- 总体分布与样本分布
- 统计量与抽样分布
  - 切比雪夫不等式
  - 大数定理
  - 中心极限定理
- 重要抽样分布
  - χ^2分布
  - t分布
  - F分布