Chpt.6 Sampling and Distribution

第六章 样本及抽样分布

目录—Content

- 随机样本
- 直方图和箱线图
- 统计量和抽样分布

Introduction

概率论：

- 假设(已知)随机变量服从某一分布，研究它的性质、特点与规律，如数字特征、分布函数的特性；
- 人类客观对实践的总结形成了概率论；

但是我们可以问：

- 概率所描述的知识是如何获取的？

  比如，假如X服从正态分布，如何获取其参数？

- 实际中，如何判断一个随机变量是否服从某一分布。

  比如，如何判断X是否为正态分布？

Introduction

数理统计：

- 随机变量其分布未知或者不完全知道(如：是正态分布，但不知参数)，期望通过重复的、独立的观察得到许多数据，以概率论为理论基础，通过对这些数据的分析，估计分布的参数，乃至推断出随机变量的分布。
- 统计要进行抽样、需要推断，这些工作形成了一定的理论：统计推断理论.

统计推断的理论基础：

- 概率论

  描述了一些随机现象，以及研究随机现象的手段；

- 大数定理

  1 频率稳定性：事件A发生的频率以概率收敛到概率p

$$\frac{n_A}{n} \xrightarrow{p} p \ (n \to \infty)$$

  2 算术均值稳定性：

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \xrightarrow{p} \mu \quad (n \to \infty)$$

- 中心极限定理

  大量相互独立的随机因素的综合影响，尽管这诸多的因素之分布是未知的，但是他们的和服从正态分布。

6.1 随机样本

[定义]：

- 对某一数量（或几个）指标进行随机实验、观察，将试验的全部可能的观察值称为总体。
- 每个可能的观察值称为个体
- 总体中所包含的个体的总数称为总体的容量。容量有限的称为有限总体，容量为无限的称为无限总体。

总体是对对象某些指标的所有观察的值:

观察全校本科生的身高，得到12000个身高观测值；

扔硬币10000次，观察反正面的情况，得到10000个数值。

由此见到这些值有些会相同的，数目也可以是无限种的。

随机变量与总体的区别：

[1] 随机变量与基本随机事件相对应，随机变量显然只是一组互异的值，进一步对应每个值(或一个区间)出现的可能性大小

总体是从另一个角度，所有试验结果一一罗列出，所以可能出现大量相等的值。

Example: 随机抛一枚硬币，X表示随机变量，正面为0，反面为1

对此试验进行观察10000次，总体为

X	0	1
p	0.5	0.5

序号	1	2	3	4	…	10000
取值	0	1	0	0	…	1

随机变量与总体的区别：

[2] 总体中的每个值是对随机变量X的观察值，这样一个总体

对应一个随机变量；

总体的研究 ⟷ 随机变量的X的研究

随机变量的分布、数字特征就称为总体的分布、数字特征

[3] 总体是从统计的角度看

随机变量是从概率角度看的

问题：在实际中总体的分布是未知的，如何解决？

途径一：逐个观察总体中的每个个体

不现实、不可行（具有破坏性、无限则不可能）

途径二：选取有代表性的个体

不知道总体，难以选择有代表性的个体

抽样：对总体进行一次观察并记录其结果（取值是多种可能），称为一次抽样；对X独立进行n次观察，并将结果按顺序记为 $X_1, \cdots, X_n$

样本：随机抽取部分个体，以用于推断总体的特性。

这是从理论上将抽样，实际中一经完成，得到一组实数值，

$x_1, \ x_2, \ \cdots, x_n$ 称为样本值。

从总体中抽取样本必须满足：

(1) 随机性 为使样本具有充分的代表性，抽样必须是随机的，应使总体中的每一个个体都有同等的机会被抽取到.

(2) 独立性 各次抽样必须是相互独立的，即每次抽样的结果既不影响其它各次抽样的结果，也不受其它各次抽样结果的影响.

称这种随机的、独立的抽样为简单随机抽样

由此得到的样本称为简单随机样本.

从总体中抽取样本必须满足：

若从总体中进行放回抽样，属于简单随机抽样，得到的样本就是简单随机样本；

若从有限总体中进行不放回抽样，则不是简单随机抽样。当总体容量N很大而样本容量n较小($n/N\le 10\%$) 时，可近似看作放回抽样，从而可近似看作简单随机抽样，得到的样本也可近似地作为简单随机样本.

6.1 样本概念

从总体中抽取容量为n的样本，就是对代表总体的随机变量X随机地、独立地进行n次观测，**每次观测的结果仍可以看作一个随机变量。**

n次观测的结果就是n个随机变量：$X_1 \ldots\ldots X_n$，它们相互独立，并与总体X服从相同的分布.

若将样本$X_1 \ldots\ldots X_n$看作一个 n 维随机变量$(X_1 \ldots\ldots X_n)$, 则

(1) 当总体X是离散随机变量, 且概率分布为p(x)时， $(X_1 \ldots\ldots X_n)$的概率分布

$$p(x_1,\cdots,x_n)=p(x_1)p(x_2)\cdots p(x_n)$$

(2) 当总体X是连续随机变量, 且概率密度为f(x)时， $(X_1 \ldots\ldots X_n)$的概率密度

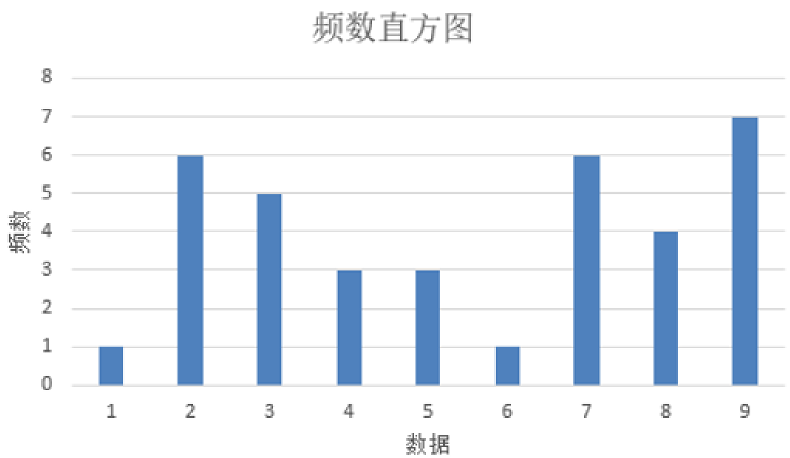
$$f(x_1,\cdots,x_n)=f(x_1)f(x_2)\cdots f(x_n)$$

6.2 直方图和箱线图

直方图

A={9,7,6,9,9,3,3,2,4,8,8,2,7,5,7,2,4,5,8,3,7,9,8,1,7,3,4,9,2,2,7,

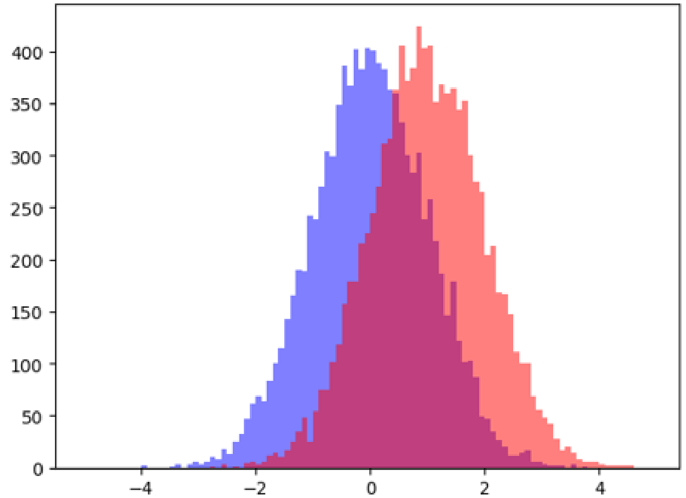
9,9,2,3,5}



算出频率，可以画出频率分布直方图。

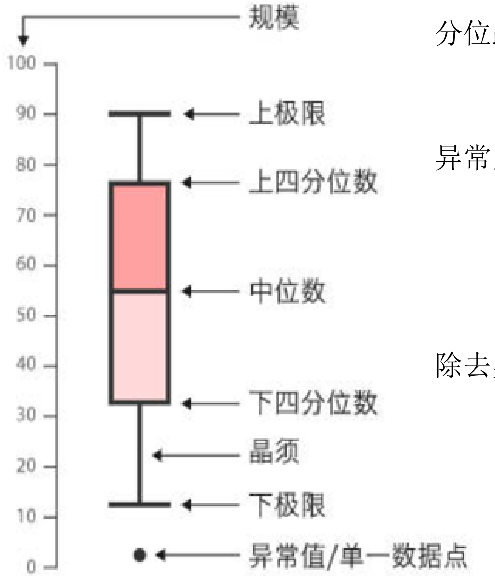
6.2 直方图和箱线图

随机生成10000个数据点，蓝色的均值为0，方差为1；红色的均值为1，方差为1；间隔都是是0.1。



6.2 直方图和箱线图

箱形图主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。



一三（上下）四分位数之间的距离成为四分位距离

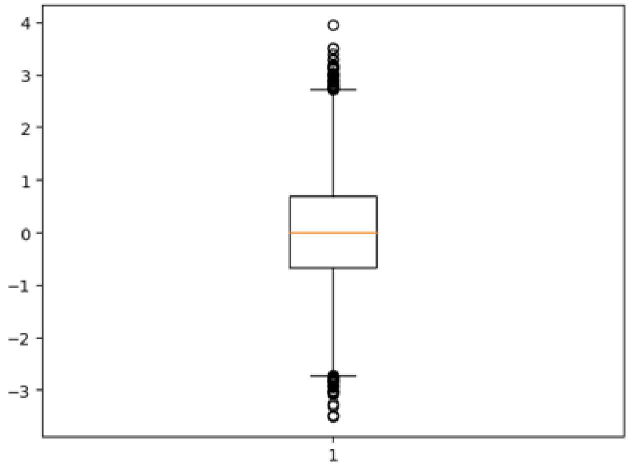
异常点：小于下四分位数-1.5四分位距离

大于上四分位数+1.5四分位距离

除去异常点后作出的箱线图成为修正箱线图

6.2 直方图和箱线图

随机生成10000个数据点，均值为0，方差为1。



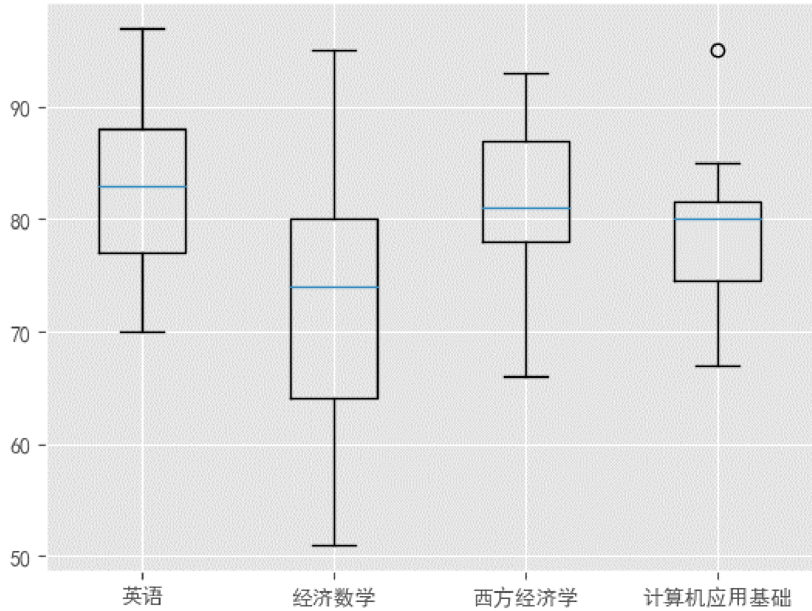
6.2 直方图和箱线图

英　　语　　= {76,90,97,71,70,93,86,83,78,85,81}

经　济　数　学 = {65,95,51,74,78,63,91,82,75,71,55}

西 方 经 济 学 = {93,81,76,88,66,79,83,92,78,86,78}

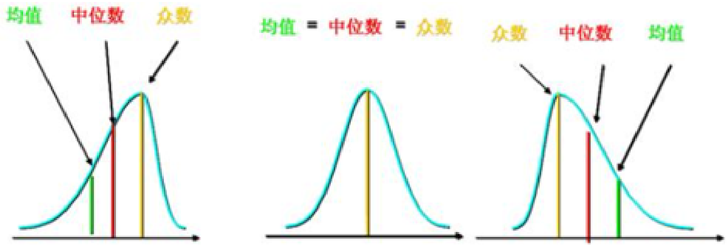
计算机应用基础 = {85,78,81,95,70,67,82,72,80,81,77}



6.2 直方图和箱线图

Mean: 期望(或均值)，试验中每次可能结果的概率乘以其结果的总和。

Mode: 众数，数据中出现最多的数。



Var: 方差，它刻画了随机变量取值在其中心位置附近的分散程度，也就是随机变量取值与平均值的偏离程度。

背景： 为了对总体X的某些概率特征(分布、均值、方差)作出推断，需要考虑各种适用的样本的，由函数满足的性质进一步得到一定的推断。

如大数定理（辛钦）：$X_1 \ldots\ldots X_n$独立同分布，且$E(X_i)=\mu$，则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$

样本是总体X的代表和反映，容量为n的样本$X_1 \ldots\ldots X_n$，可以看作是一个n维随机变量$(X_1 \ldots\ldots X_n)$，则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$

如果有一组样本值 $x_1, x_2, \cdots, x_n$，那么我们可以估计得到 $\mu \approx \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$

来自总体X的n个样本$X_1\ldots\ldots X_n$ 构成n维随机变量 $(X_1\ldots X_n)$，其函数 $g(X_1,\cdots,X_n)$ ，若其中不含任何未知量，则称其为<u>统计量</u>。

统计量都是随机变量，由样本$X_1\ldots\ldots X_n$的观测值$x_1\ldots\ldots x_n$，算得的函数值 $g(x_1,\cdots,x_n)$ 是统计量 $g(X_1,\cdots,X_n)$ 的观测值.

研究规律得用随机变量

实际应用可以直接用观测值了

常用统计量及其观测值：

1. 样本均值 $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$ 观测值为 $\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$

(2) 样本方差

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

观测值为

$$s^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_i-\overline{x})^2$$

(3) 样本标准差

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_i-\overline{X})^2}$$

观测值为

$$s=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_i-\overline{x})^2}$$

常用统计量及其观测值：

(4) 样本k阶原点矩

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

观测值为

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

(5) 样本k阶中心矩

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

观测值为

$$b_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k$$

统计与统计量讨论

统计研究相当一个N重独立实验；而统计量就是定义在这个N重独立实验上的，具有代表性的随机变量（此处的代表，指的是被研究的统计量可以有效代表 单次实验的某些特点）

换而言之，统计量是一个由多个（联合分布）随机变量运算得到的复合随机变量，但是该随机变量（的数字特征）却又能够有效反映单个随机变量的部分特征。

强化概念：统计量是随机变量，拥有数字特征

我们来比较一下

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

假 $\quad E(X_i)=\mu \qquad D(X_i)=\sigma^2$

则 $\quad \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad E(\overline{X})=\mu,\ \ D(\overline{X})=\frac{1}{n}\sigma^2$

$$S^2\xrightarrow{P}\sigma^2 \qquad Y_n^2\xrightarrow{P}\sigma^2$$

$$s^2\approx\sigma^2 \qquad y_n^2\approx\sigma^2$$

假如我们要用样本估计σ²时，用 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$

而不用 $Y_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 。盖因前者均值为σ²

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

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

$$=\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-2\overline{X}\frac{1}{n-1}\sum_{i=1}^{n}X_i+\frac{n}{n-1}\overline{X}^2$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-2\frac{n}{n-1}\overline{X}^2+\frac{n}{n-1}\overline{X}^2$$

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

$$E(X_i^2) = D(X_i) + (EX_i)^2$$

$$E(X_i^2) = \sigma^2 + \mu^2$$

$$E(\overline{X}^2) = D(\overline{X}) + (E\overline{X})^2$$

$$E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

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

$$= \frac{n}{n-1}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{1}{n}\sigma^2 + \mu^2\right)$$

$$= \sigma^2$$

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

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

$$= \frac{n-1}{n} S^2$$

$$E(Y_n^2) = \frac{n-1}{n} ES^2$$

$$= \frac{n-1}{n} \sigma^2$$

6.3 统计量和抽样分布

样本是随机变量

统计量是样本的函数，从而统计量也是随机变量

统计量的分布称为抽样分布

为什么要研究抽样分布：

- 一般而言，总体分布已知，抽样分布也是知道的，但是确切得到是困难的；
- 从另外一个角度，我们希望由统计量的分布（特别是在观测值得到后），估计、推断出总体的一些特征。我们希望研究统计分布，以便作出统计推断。

几类抽样分布

- 样本均值分布 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,\ 1)$

- χ^2分布 $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$ $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

- T分布 $t=\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

- F分布 $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$

注意：以上是对X~N(μ，σ²)而言的；取得 n 个样本

一.样本均值分布

假设总体分布的均值与方差都是已知的，那么我们可以对来自总体的多个样本的均值做出估计。

假设 $X_1,\cdots,X_n$ 是来自总体X的独立样本, 样本均值为

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

。一般情况下我们知道

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,\ 1)$$

假设 $X_1,\cdots,X_n$ 是来自正态总体 $X\sim N(\mu,\sigma^2)$ 的独立样本,

则样本均值 $\bar{X}$ 服从正态分布 $N(\mu,\frac{\sigma^2}{n})$，标准量服从标准正态分布

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

二. 卡方分布

前面我们更多地讨论了样本均值的分布，关于样本方差有

何种分布？

一般的总体不会得到很直接的结果;

对于正态总体 $X \sim N(\mu, \sigma^2)$ 则统计量 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$

服从自由度为n的 χ^2 分布. 事实上

$$Y_i = (X_i - \mu)/\sigma \sim N(0,1) \qquad (i = 1,2,\cdots,n)$$

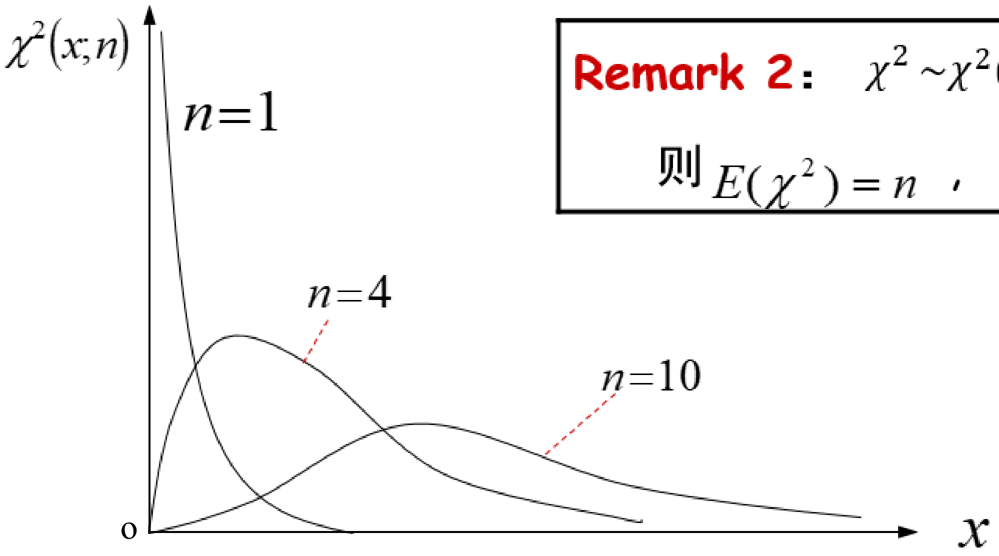
且相互独立，由以下定理知

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = Y_1^2 + \cdots + Y_n^2 \sim \chi^2(n)$$

[定理6.3] 设随机变量 $X_1,\cdots,X_n$ 是来自标准正态总体 $X \sim N(0,1)$ 的独立样本。则随机变量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$

服从自由度为n的 χ^2 分布，记为 $\chi^2(n)$，其概率密度：

$$f_{\chi^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$



Remark 2： $\chi^2 \sim \chi^2(n)$

则 $E(\chi^2)=n$ ， $D(\chi^2)=2n$

Remark 1： χ^2分布具有可加性，也就是说，

$\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$ 且它们相互独立，

则 $\chi_1^2+\chi_2^2 \sim \chi^2(n_1+n_2)$

证明$D(\chi^2(1))=2$

- 设X~N(0,1), 计算$D(X^2)$,即证明 $D(\chi^2(1))=2$

用平方关系来算, $D(X^2)=E(X^4)-(E(X^2))^2$

先算 $E(X^4)$, 令$f(x)$ 是N (0,1)的密度函数

$$E(X^4)=\int_{-\infty}^{+\infty}x^4f(x)dx=\int_{-\infty}^{+\infty}x^3\cdot xf(x)dx$$

$$f'(x)=f(x)\left(-\frac{x^2}{2}\right)'=-xf(x)$$

$$E(X^4)=\int_{-\infty}^{+\infty}-x^3\cdot f'(x)dx=-x^3\cdot f(x)|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}-3x^2\cdot f(x)dx$$

$$=0+\int_{-\infty}^{+\infty}-3x\cdot xf(x)dx=\int_{-\infty}^{+\infty}3x\cdot f'(x)dx$$

$$=3x\cdot f(x)|_{-\infty}^{+\infty}+3\int_{-\infty}^{+\infty}f(x)dx=3$$

$E(X^2)=D(X)+E^2(X)=1+0=1$ 所以 $(E(X^2))^2=1$

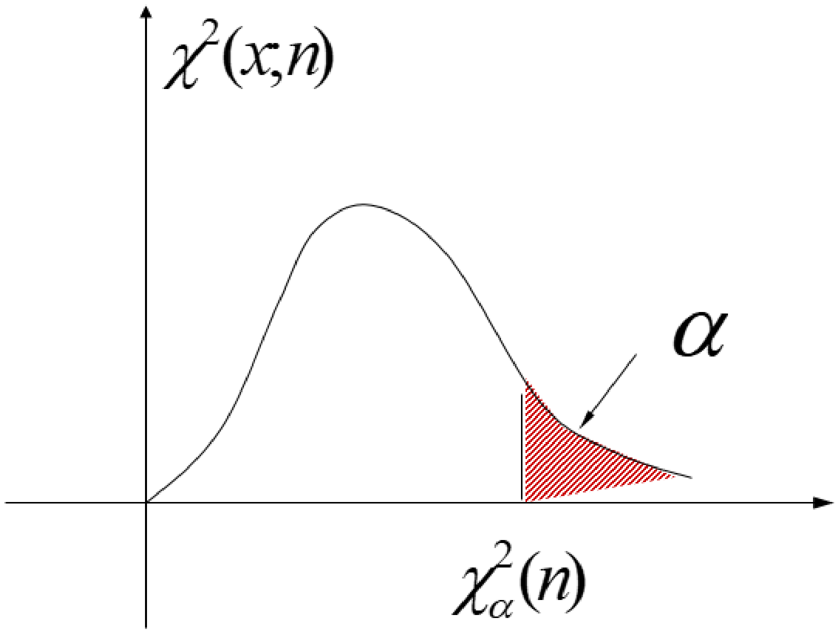
则$D(X^2)=E(X^4)-(E(X^2))^2=3-1=2$

6.3 统计量和抽样分布

对不同的自由度n及不同的数 $\alpha(0<\alpha<1)$ 满足的

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(x)dx = \alpha$$

值 $\chi_\alpha^2(n)$ 称为χ^2分布的**α分位数**。



上面结果的作用：

假设正态分布X~N(μ，σ²)，N个样本 $X_1,\cdots,X_n$，观测值为x_1，x_2，$\cdots,x_n$，知道 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 服从χ^2分布 (χ^2分布 要求X~N(0,1))

- 在μ已知，可估计σ；或反之。
- 当均值μ未知时，考虑用 $(n-1)S^2=\sum_{k=1}^{n}(X_i-\bar{X})^2$ 来代替 $\sum_{k=1}^{n}(X_i-\mu)^2$，一方面得到分布估计，另一方面可以估计σ。

[定理6.4] S^2是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的n个样本的方差，那么

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

三. T分布（学生分布）William Gosset

但是如果我们不知道σ时，尽管 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 可以估计，但还是无法由此估计μ

可以想到用样本标准差S来代替σ，即得到 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$

但是一般我们根本不知道此服从何种分布，除非是正态分布。

三. T分布（回顾中心极限定理 VS. T分布）

当总体均值与方差已知 $E(X)=\mu, D(X)=\sigma^2$ ，$\{X_i\}$ 是来自总体的独立样本，我们知道 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 近似为标准正态分布 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$

由此可以估计 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 或 $\bar{X}$ 。

如果其中已知σ 我们就可以估计μ，或者反过来, 知道μ估计σ 。

如果X是正态分布，那就可以确切得到 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,\ 1)$

三. T分布

[定理6.1] 若总体服从正态分布 $X \sim N(\mu,\sigma^2)$, $\{X_i\},S$

分别是来自总体的样本与样本标准差,

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

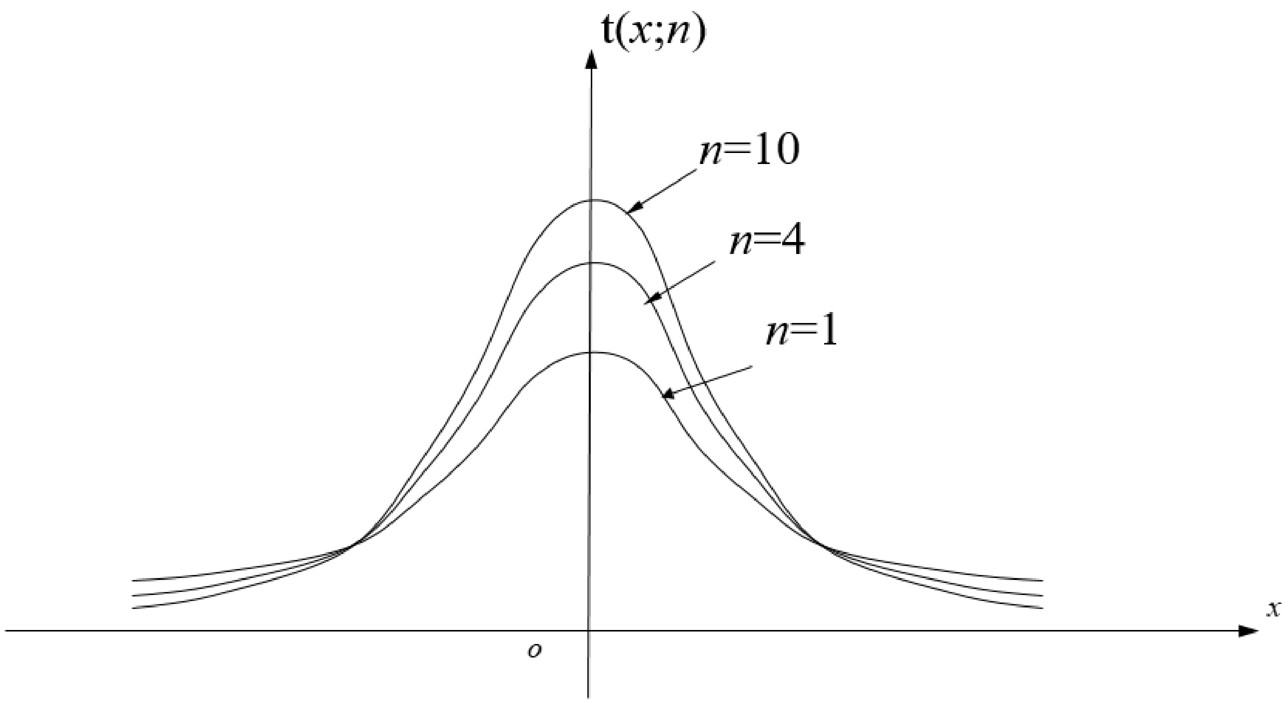
那么随机变量 $t = \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ 服从自由度为n-1的t分布，记

为 $t \sim t(n-1)$ 。分子是标准正态分布，分母是卡方分布

t(n)其概率密度为：

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

Remark 3： t 分布的分布曲线关于 $t=0$ 对称；



Remark1: 其中伽玛函数 $$\Gamma(\alpha)=\int_0^{\infty}u^{\alpha-1}e^{-u}du \qquad (\alpha>0)$$

有如下性质：

$$(1)\quad \Gamma(\alpha+1)=\alpha\Gamma(\alpha)$$

$$(2)\quad \Gamma(n)=(n-1)!$$

$$(3)\quad \Gamma(\frac{1}{2})=\sqrt{\pi}$$

Remark2: 由定理可以知道总体为正态分布 $X\sim N(\mu,\sigma^2)$，

随机变量 $t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 服从自由度为n-1的 t 分布。

$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 中如μ是非统计量(未知)，由对 t 的分析可以估计出μ，而且是在σ也未知的情况下

Remark 3：t 分布的形式如上所言。之所以叫做 t 分布是因为在1908年，Dublin城的W.S.Gosset用笔名:Student发表了一篇文章提出了该分布。

Remark 5：由 t 分布的出处，我们可以知道它是对 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 的一个近似，而后者在n无限大时趋近于标准正态分布。故此，可以推测当自由度n无限增大时, t分布将趋近于标准正态分布 $N(0,1)$ 或者

$$f_t(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \rightarrow \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$$

T分布的分位点

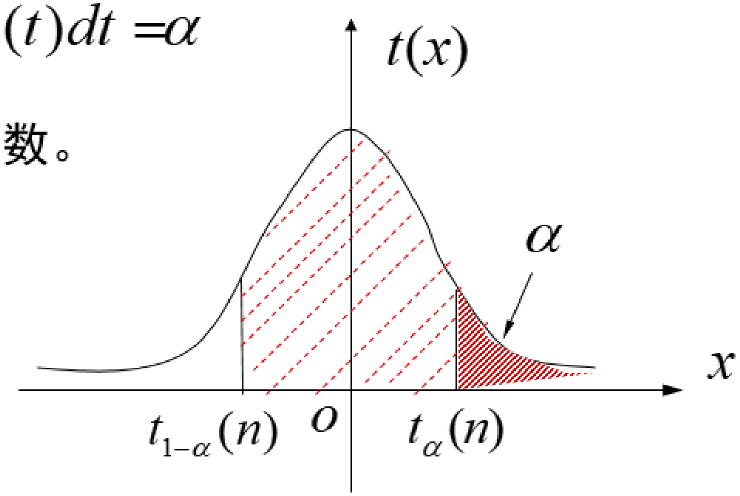
[分位点] 随机变量X, 对一个正数α (0< α <1) , 满足

$P\{X > x_\alpha\} = \alpha$ 的值 X_α称为X分布的α分位数.

对不同的自由度n及不同的数 α (0< α <1) , 满足

$$P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{\infty} f_t(t)dt = \alpha$$

的$t_\alpha(n)$ 值称为t分布的α分位数。



由 $f_t(t)$的对称性知：

$$t_{1-\alpha}(n) = -t_\alpha(n)$$

[定理6.2] 若随机变量 $X \sim N(0,1)$，$Y \sim \chi^2(n)$

则 $t = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$，概率密度

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

注意到：此与前面的定义有相似之处。

前面定理说 $X \sim N(\mu, \sigma^2)$，则 $t = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

Problem：两者有什么关系吗？

和P45说的 卡方分布一样，这也是一个大拓展！

四．F分布

有时，需要比较来自两个总体的样本的方差，

$$F = {S_1^2}\Big/{S_2^2}$$

[定理6.5] 如果 S_1^2 和 S_2^2 是两个 n_1, n_2 独立样本的方差，来自两个具有相同方差（但是未知）的独立正态总体，那么 ${S_1^2}\Big/{S_2^2}$ 服从参数为 n_1-1, n_2-1 的F分布，记为

$$F \sim F(n_1-1, n_2-1)$$

四．二个正态总体的统计量的分布

[定理6.6] 若随机变量 $U \sim \chi^2(n_1), V \sim \chi^2(n_2)$, U与V独立, 则

随机变量 $F = \dfrac{U/n_1}{V/n_2}$ 服从自由度为 (n_1, n_2) 的 F 分布，记

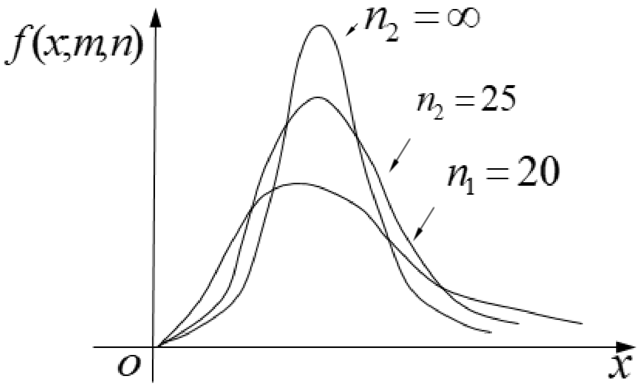
为 $F \sim F(n_1, n_2)$，其概率密度为

$$
f_F(z) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} \dfrac{z^{\frac{n_1}{2}-1}}{(n_1 z + n_2)^{\frac{n_1+n_2}{2}}} & z > 0 \\ 0 & z \le 0 \end{cases}
$$

$F \sim F(n_1, n_2)$ 其中分子的自由度

n_1 为第一自由度；

分母的自由度 n_2 为第二自由度。

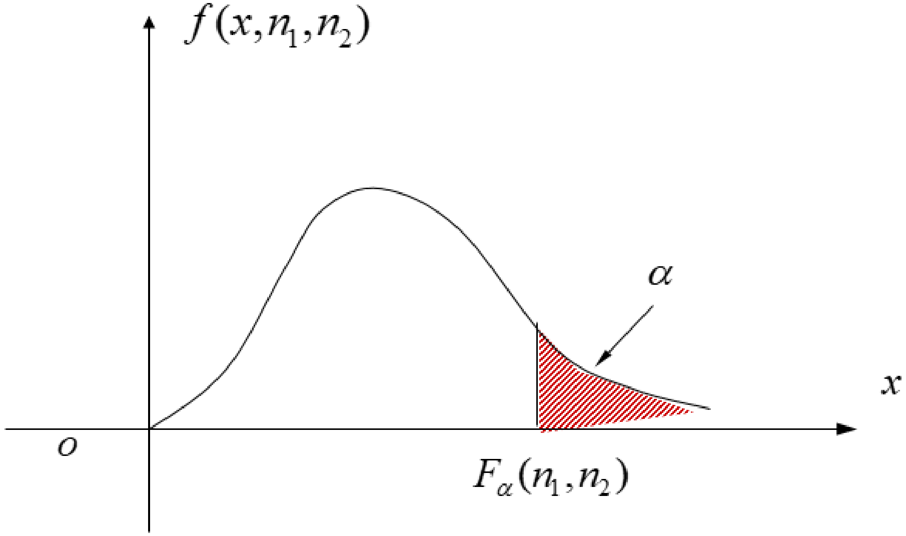


Remark 1：如果 $X \sim F(m, n)$，则 $\frac{1}{X} \sim F(n, m)$。

6.3 统计量和抽样分布

满足 $$\int_{F_\alpha(n_1,n_2)}^{\infty} f(x;n_1,n_2)dx=\alpha\,,\quad 0<\alpha<1$$

称 $F_\alpha(n_1,n_2)$ 为F分布的α分位点。



Remark 2: $$F_\alpha(n_1,n_2)=\frac{1}{F_{1-\alpha}(n_2,n_1)}.$$

6.3 统计量和抽样分布

[定理6.7] 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本，且两个样本独立。

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

分别是两个样本的样本均值，

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(Y_j - \bar{Y}\right)^2$$

是各样本方差

则有：

[1]
$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

[2] $$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

[3] 当$\sigma_1=\sigma_2=\sigma$时

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

其中 $$S_w = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

证明：

[1] 由 $X\sim N(\mu_1,\sigma_1^2), Y\sim N(\mu_2,\sigma_2^2)$, 与 $X_1,X_2,\cdots,X_{n_1}$ $Y_1,Y_2,\cdots,Y_{n_2}$ 分别是来自两个总体的简单样本

$X_1,X_2,\cdots,X_{n_1}$ 的线性组合 $\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$ 服从正态分布 $N(\mu_1,\frac{\sigma_1^2}{n_1})$

同理 $\bar{Y}\sim N(\mu_2,\frac{\sigma_2^2}{n_2})$

又由于 $\bar{X},\bar{Y}$ 独立，其组合是正态分布，所以

$$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

其标准化随机变量 $$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

[2] $$\chi_1^2=\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$$

$$\chi_2^2=\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$$

由于 S_1^2 与 S_2^2 独立，

$$\frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)}\sim F(n_1-1,n_2-1)$$

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}=\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

[3] 随机变量

$$U=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim N(0,1)$$

由上知，统计量

$$\chi_1^2=\frac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1)$$

$$\chi_2^2=\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)$$

又由于S_1^2 与 S_2^2 独立，利用χ^2 分布的可加性知随机变量

$$V=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

随机变量U,V独立, 因此

$$\frac{U}{\sqrt{\dfrac{V}{n_1+n_2-2}}}=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$S_w=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

[定理6.8] 假设 $X_1,\cdots,X_n$ 是来自正态总体 $X\sim N(\mu,\sigma^2)$ 的独立样本，则样本均值 $\bar{X}$ 与样本方差 S^2 独立。

考研的试题中出现过本定理的应用！