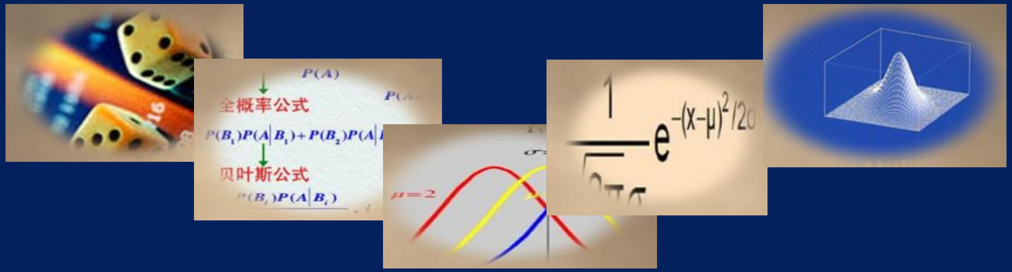
第六章 抽样分布

6.1 随机样本



杨建芳

概率论与数理统计的关系

概率论

分布已知，研究性质、特点和规律性

基础 →

← 重要应用

数理统计

分布未知或不完全知道，通过统计对其分布做出推断

数理统计学是一门关于数据资料的收集、整理、分析和推断的学科。

- 参数估计
- 假设检验
- 方差分析
- 回归分析等

内容提要

- 样本和抽样分布
  - 基本概念
    - 总体、个体、容量、样本
    - 统计量：$\overline{X}$、S^2、S、A_k、B_k
  - 三大分布（构造、性质、分位点）
    - χ^2分布
    - t分布
    - F分布
  - 正态总体的样本均值和方差的分布
    - 单正态总体
      - $\overline{X}$ 的分布
        - 方差已知
        - 方差未知
      - S^2的分布
    - 双正态总体
      - $\overline{X}-\overline{Y}$ 的分布
        - 方差已知
        - 方差未知
      - S_1^2/S_2^2 的分布

随机样本

- 总体：试验的全部可能观察值称为总体(或母体)
- 个体：每一个可能观察值称为个体
- 容量：总体中所包含的个体的个数称为总体的容量（有限、无限）

例如，研究杭州电子科技大学二年级学生的年龄分布时，这些同学年龄的全体就组成了总体，其中每个同学的年龄就是个体。

- 在统计学里，我们所关心的往往是研究对象的某一项(或某几项)数量指标，例如上述例子中的大学二年级学生的年龄。
- 总体中的每一个个体是随机试验的一个观察值，因此它是某一随机变量X的值。这就意味着一个总体对应一个随机变量X，对总体的研究就是对一个随机变量X的研究。

总体 $\xleftrightarrow{\text{一一对应}}$ 随机变量X

如何理解总体和随机变量一一对应？

假设杭州电子科技大学二年级学生共4000人，所有可能出现的年龄不妨设为18，19，20，21，由这些学生年龄全体构成总体为：

$\{x_1, x_2, \dots x_{4000}\}$={19,20,20,19,18,21,20,…,19}

不妨设其中18岁占10%，19岁占40%，20岁占45%，21占5%，

若用X来表示二年级同学年龄的分布，则其分布律为：

X	**18**	**19**	**20**	**21**
p_i	0.1	0.4	0.45	0.05

类似地，若考虑研究对象的多项指标时，如在研究某地区中学生的营养状况时，若关心的数量指标是身高和体重，我们用X和Y分别表示身高和体重，那么此总体就可用二维随机变量(X,Y)或其联合分布函数$F(x,y)$来表示.

二、简单随机样本

在实际中，总体的分布一般是未知的，或只知道它具有某种形式而其中包含着未知参数。在数理统计中，人们都是通过从总体中抽取部分个体，根据获得的数据对总体分布作出推断。被抽出的部分个体叫做总体的一个样本。

由于我们是利用抽样来对总体的分布进行推断,所以抽样必须是随机的. 例如前面提到二年同学年龄分布的例子，抽样时就希望每个同学等可能地被抽到,只有这样才能通过抽样比较客观地了解总体。

最常用的一种抽样方法叫作“**简单随机抽样**”，它要求抽取的样本满足下面两点:

1. 代表性： $X_1,X_2,\ldots,X_n$中每一个与所考察的总体有相同的分布.

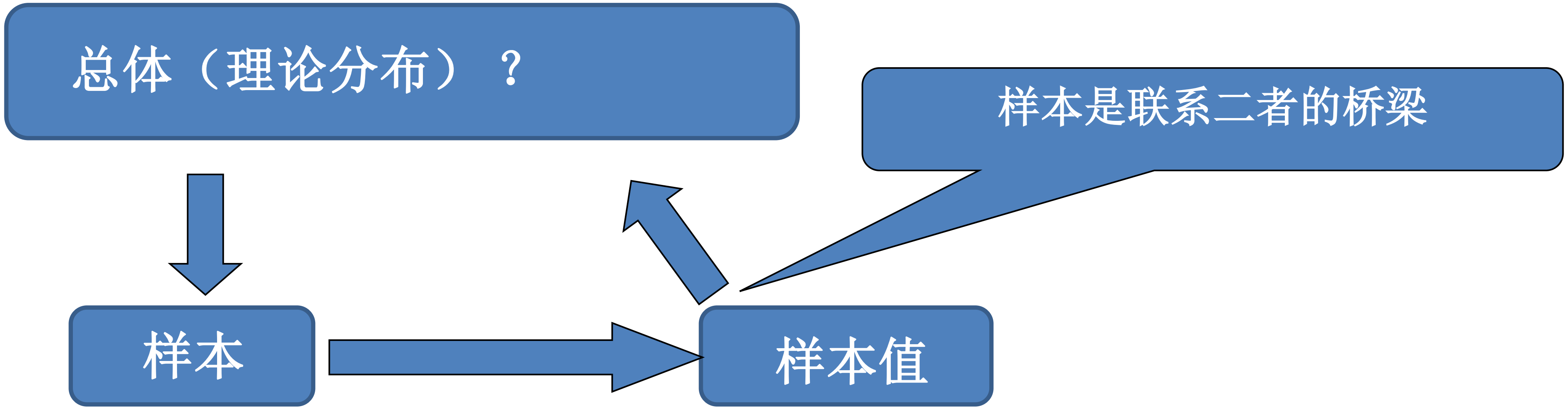
2. 独立性： $X_1,X_2,\ldots,X_n$是相互独立的随机变量.

由简单随机抽样得到的样本称为**简单随机样本**，简称样本.

它们的观察值称 $x_1,x_2, \ldots, x_n$为样本值。

今后如不加声明，均指简单随机样本。

总体、样本、样本值的关系



统计是从手中已有的资料——样本值

去推断未知总体的情况——总体分布$F(x)$。

总体分布决定了样本取值的规律， 因而可以由样本值去推断总体.

1、为什么可以把总体看成是一个随机变量？

如何理解个体与总体具有一样的分布？

2、简单随机样本有什么特点？有什么意义？

第六章 抽样分布

6.3　抽样分布（1）

统计量

杨建芳

统计量

样本是总体的代表和反映，但我们在抽取样本后，并不直接利用样本进行推断，而需要对样本进行一番“加工”和“提炼”，把样本中所包含的关于我们所关心的事物的信息集中起来，这便是针对不同的问题构造样本的某个函数，称之为统计量。

定义：设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量。

称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值。

严格地说，一个统计量就是样本 $(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)$ 的一个函数，且要求它不包含总体的任何未知参数。因此，统计量也是一个随机变量。

常用统计量

总体均值
$E(X)$

1.样本均值 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$

总体方差
$D(X)$

2.样本方差 $S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\dfrac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)$

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

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

常用统计量

总体k阶原点矩：

$E(X^k)$

3.样本k阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$， $k = 1, 2, \cdots$

总体k阶中心矩：

$E([X - E(X)]^k)$

4.样本k阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$， $k = 1, 2, \cdots$

容易看出，一阶样本原点矩即为 $\bar{X}$.

常用统计量的观察值

设X_1，$X_2,\cdots$，X_n是来自总体X的一个样本，x_1，x_2，$\cdots$，x_n是相应于样本X_1，X_2，$\cdots$，X_n的观察值.

1）**样本均值**：$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ $\longrightarrow$ 观察值：$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

2）**样本方差**：$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$,样本标准差$S=\sqrt{S^2}$ $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

3）**样本k阶（原点）矩**：$A_k=\frac{1}{n}\sum_{i=1}^{n}{X_i}^k,k=1,2,\cdots$ $a_k=\frac{1}{n}\sum_{i=1}^{n}{x_i}^k,k=1,2,\cdots$

4）样本k阶（中心）矩：$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k,k=2,3\cdots$ $b_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k,k=2,3\cdots$

总体k阶矩与样本k阶矩

1）**总体**X的k阶（原点）矩：$\mu_k = E(X^k)$,（期望存在）

比较：**样本**k阶（原点）矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k$

2）**总体**X的k阶（中心）矩：$E\{[X - E(X)]^k\}, \mathrm{k} = 2,3, \cdots$

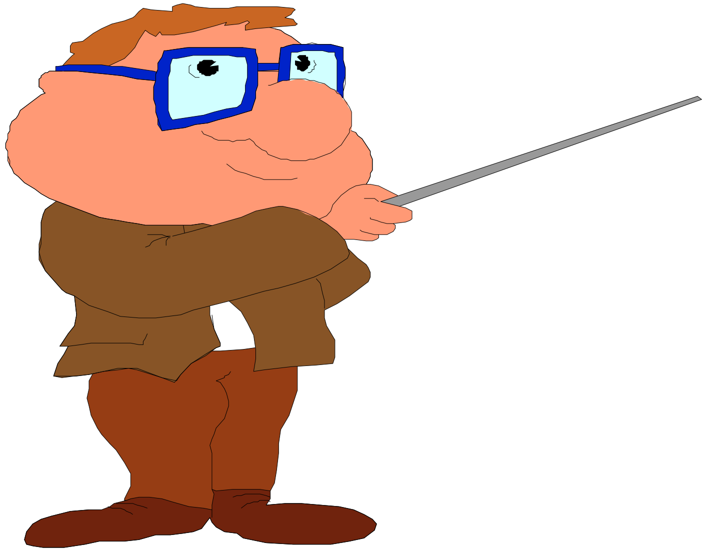
比较：**样本**k阶（中心）矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, k = 2,3 \cdots$

总体k阶矩与样本k阶矩的关系

1）当$n \to \infty$时，$A_k \xrightarrow{P} \mu_k$（样本矩**依概率**收敛总体矩）

2）当$n \to \infty$时，$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k)$（样本矩的连续函数依概率收敛总体矩的函数）

思考



统计量满足什么条件？

为什么统计量中不能含有未知参数？

练习题

设总体X前期服从0-1分布，X_1，X_2，$\cdots$，X_5是来自总体X的一个简单随机样本，$\bar{X}$是样本均值，p是介于0与1之间的常数但未知，则下列各选项中不是统计量的为（　）.

（A）$\min\{X_1, X_2, \cdots, X_5\}$　　（B）$X_1-(1-p)\bar{X}$

（C）$max\{X_1, X_2, \cdots, X_5\}$　　（D）$5\bar{X}$

参考答案：（B）

注：大写改成小写则称为相应的观察值；

总体矩是常数，样本矩是随机变量；

统计量是样本函数，实质上是随机变量，对应的分布，称为抽样分布；

重点掌握样本均值、样本方差。

第六章 抽样分布

6.3 抽样分布（2）

χ^2（卡方）分布

杨建芳

三个常用统计量的分布

- χ^2分布
- t 分布
- F 分布

1. χ^2（卡方）分布的构造定义

★设X_1，$X_2, \cdots$， X_n服从标准正态分布$N(0,1)$，且相互独立.

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$\boldsymbol{\chi^2}$分布.（也记$\chi^2 = {X_1}^2 + {X_2}^2 + \cdots + {X_n}^2 \sim \chi^2(n)$）

★自由度n指的是随机变量χ^2等式右端的独立变量的个数.

★或X_1，X_2，……，X_n是来自总体为标准正态分布$N(0,1)$的样本（样本有独立性）.

2. χ^2（卡方）分布的概率密度（了解）

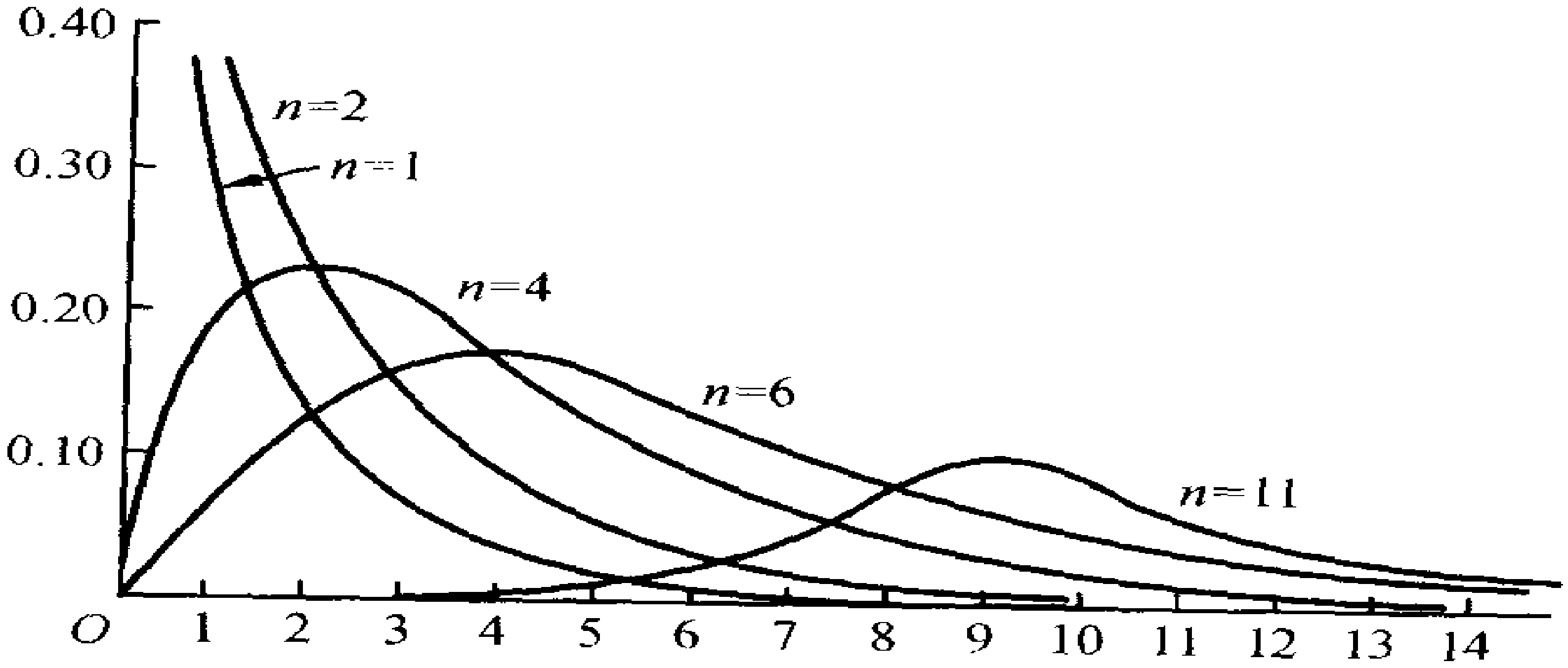
★ $\chi^2(n)$分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-y/2}, & y>0\\ 0, & y\le 0\end{cases}$$

★其中$\Gamma(x)$（伽马）函数的定义为

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

3. χ^2（卡方）分布的图形



4. χ^2（卡方）分布的性质

★1） χ^2分布的可加性：

设$\chi_1{}^2 \sim \chi^2(n_1)$，$\chi_2{}^2 \sim \chi^2(n_2)$，且$\chi_1{}^2$，$\chi_2{}^2$相互独立，则有

$\chi_1{}^2 + \chi_2{}^2 \sim \chi^2(n_1 + n_2)$.

★**证**：设$X_i \sim N(0,1)$，独立，记$\chi_1{}^2 = X_1{}^2 + X_2{}^2 + \cdots + X_{n_1}{}^2$

$\chi_2{}^2 = X_1{}^2 + X_2{}^2 + \cdots + X_{n_2}{}^2$,

两式相加，$\chi_1{}^2 + \chi_2{}^2 = X_1{}^2 + X_2{}^2 + \cdots + X_{n_1}{}^2 + X_1{}^2 + X_2{}^2 + \cdots + X_{n_2}{}^2$

独立变量的个数相加，故$\chi_1{}^2 + \chi_2{}^2 \sim \chi^2(n_1 + n_2)$.

4. χ^2（卡方）分布的性质

★ 2） χ^2分布的期望与方差：

设$\chi^2 \sim \chi^2(n)$，则有$E(\chi^2)=n$，$D(\chi^2)=2n.$

★证：设$X_i \sim N(0,1)$，$i=1,2,\cdots,n$,相互独立，记$\chi^2={X_1}^2+{X_2}^2+\cdots+{X_n}^2$,则

$$E\left({X_i}^2\right)=\mathrm{D}(X_i)+[E(X_i)]^2=1+0=1$$

故$E(\chi^2)=E({X_1}^2+{X_2}^2+\cdots+{X_n}^2)=nE({X_i}^2)=n$

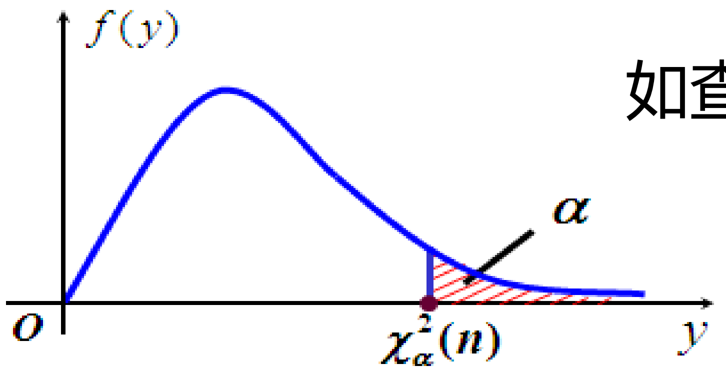
又$D(\chi^2)=D({X_1}^2+{X_2}^2+\cdots+{X_n}^2)=nD({X_i}^2)$

而$D\left({X_i}^2\right)=E\left({X_i}^4\right)-\left[E\left({X_i}^2\right)\right]^2=\int_{-\infty}^{+\infty}x^4\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx-1=2$，故$D(\chi^2)=2n.$

5. χ^2（卡方）分布的上α分位点

★设$\chi^2 \sim \chi^2(n)$，对于给定的正数$\boldsymbol{\alpha}$（$0 < \boldsymbol{\alpha} < \mathbf{1}$），满足条件$P\{\chi^2 > {\chi_\alpha}^2(n)\} = \alpha$的点${\chi_\alpha}^2(n)$称为卡方分布的**上$\alpha$分位点.**

★式$P\{\chi^2 > {\chi_\alpha}^2(n)\} = \int_{{\chi_\alpha}^2(n)}^{+\infty} f(y)dy$



如查表 ${\chi_{0.025}}^2(8) = 17.534$

${\chi_{0.975}}^2(8) = 2.180$

了解

当$\boldsymbol{n}$充分大时$(n > 40)$，近似地有

$$\chi_\alpha^2(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$$

其中 z_α 是 $N(0,1)$ 的上 α 分位点。

例如， $z_{0.05} = 1.645$，

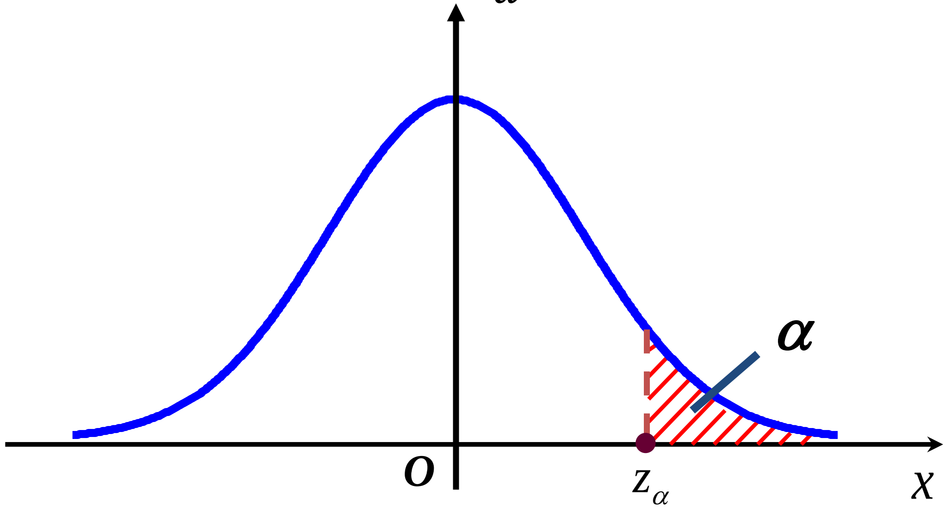
$$\chi_{0.05}^2(50) \approx \frac{1}{2}(1.645 + \sqrt{99}) = 67.221.$$

精确， $\chi_{0.05}^2(50) = 67.505$.

6. 标准正态分布的上α分位点

★设$X\sim N(0,1)$，对于给定的正数$\boldsymbol{\alpha}$（$0<\boldsymbol{\alpha}<\mathbf{1}$），满足条件$P\{X>z_\alpha\}=\alpha$的点$z_\alpha$称为标准正态分布的**上$\alpha$分位点.**（没有自由度）

★式$P\{X>z_\alpha\}=\int_{z_\alpha}^{+\infty}\varphi(x)dx$



1）$z_{1-\alpha}=-z_\alpha$ （对称性）

2）查表 $z_{0.025}=1.96$

$z_{0.05}=1.645$

例 1

★设X_1，X_2，…，X_n服从正态分布$N(\mu,\sigma^2)$，且相互独立.则随机变量$Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\sim\chi^2(n)$.

平方和？标准正态分布？

即$Y=\dfrac{1}{\sigma^2}[(X_1-\mu)^2+(X_2-\mu)^2+\cdots+(X_n-\mu)^2]\sim\chi^2(n)$

例 1 证明

证明：因X_1，X_2，……，X_n服从正态分布$N(\mu,\sigma^2)$，

则 $\frac{X_i-\mu}{\sigma} \sim N(0,1)$,或$(\frac{X_i-\mu}{\sigma})^2 \sim \chi^2(1), i=1,2,\cdots,n$

且相互独立,由卡方分布的定义或可加性，

得$Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n).$

什么样的统计量服从χ^2分布，即χ^2分布的构造？

χ^2分布的性质？

如何理解χ^2分布上α分位点？

练习题

※设X_1，X_2，X_3，X_4，X_5为来自总体$N(0,1)$的样本，若

$Y = a[(X_1 + X_2)]^2 + b[(X_3 - X_4 + X_5)]^2$服从$\chi^2$分布，则常数$a, b$的值（　　）.

（A） $a = 1/2, b = 1/3$　　（B） $a = 2, b = 3$

（C） $a = 1/3, b = 1/2$　　（D） $a = 3, b = 2$

※参考答案：（A）.

第六章 抽样分布

6.3 抽样分布（3）

t分布

杨建芳

1. t分布的构造定义

★设$X \sim N(0,1)$，$Y \sim \chi^2(n)$，且X,Y相互独立.则随机变量$t=\dfrac{X}{\sqrt{Y/n}}$

服从自由度为n的t分布.记为$t \sim t\ (n)$，也称学生氏（Student）分布.

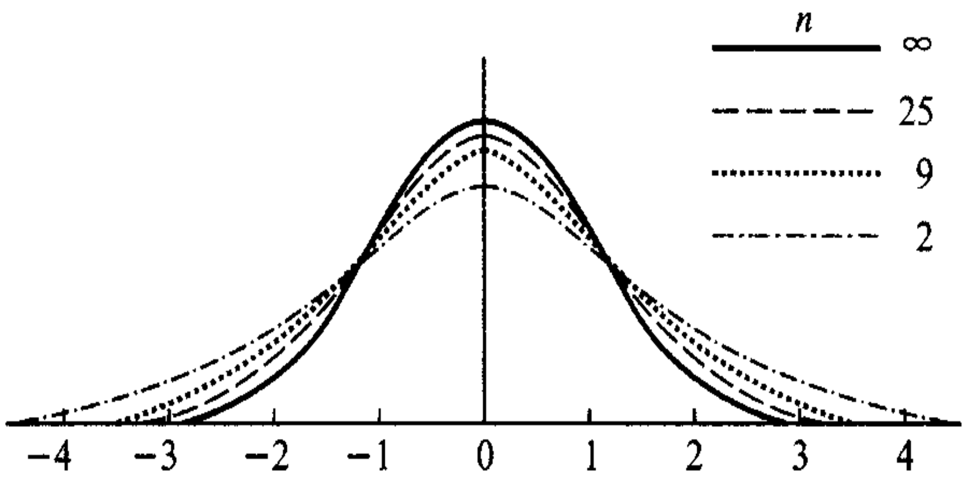
★自由度$\boldsymbol{n}$与分母中随机变量Y中的χ^2分布的自由度一样.

威廉·戈塞，1908年，都柏林，健力士酿酒厂.因为不能以他本人的名义发表，所以论文使用了学生（Student）这一笔名.

2. t分布的概率密度、图形、性质

★ t分布的概率密度为（了解）

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



性质：

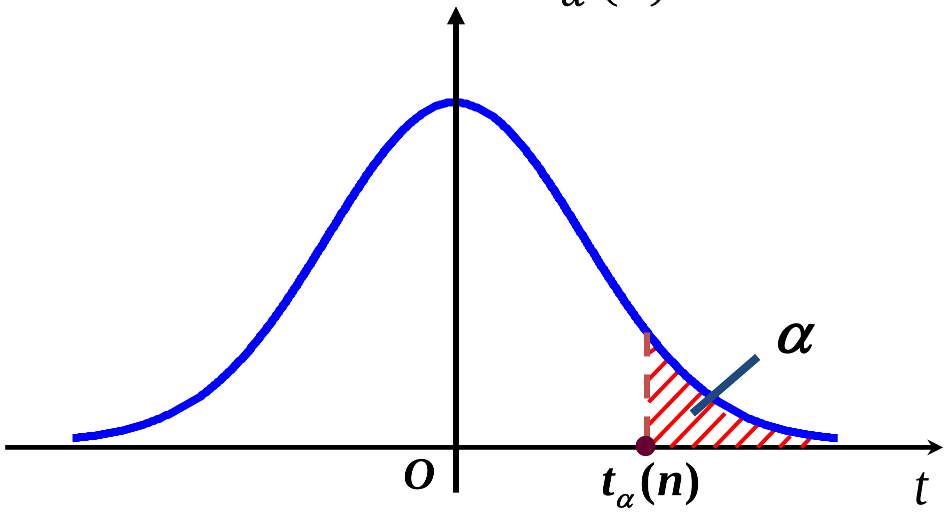
★偶函数. 当较大时, t分布接近于标准正态分布

★ $\lim\limits_{n\to\infty} h(t)=\varphi(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$

3. t分布的上α分位点

★设$t \sim t(n)$，对于给定的正数$\boldsymbol{\alpha}$（$0 < \boldsymbol{\alpha} < \mathbf{1}$），满足条件$P\{t > t_\alpha(n)\} = \alpha$的点$t_\alpha(n)$称为t分布的**上α分位点.**

★式$P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} h(t)dt$



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

2）查表 $t_{0.025}(15) = 2.1315$

$t_{0.05}(15) = 1.7531$

当n充分大时，$t_\alpha(n) \approx z_\alpha$.

例 1

★设X_1，X_2，X_3是来自总体$N(0,4)$的样本.试确定随机变量

$$Y=\frac{\sqrt{2}X_3}{\sqrt{{X_1}^2+{X_2}^2}}$$的分布.

分母出现标准正态分布的平方和？

分子：标准正态分布？

例 1 解

解：1）分子化成标准正态分布形式: $U = \frac{X_3}{2} \sim N(0,1)$

2）分母找χ^2形式：$\frac{X_1}{2} \sim N(0,1)$，$\frac{X_2}{2} \sim N(0,1)$

由独立性，卡方的定义，$V = (\frac{X_1}{2})^2 + (\frac{X_2}{2})^2 \sim \chi^2(2)$

3）显然U,V相互独立.则由t分布的定义:$\frac{U}{\sqrt{V/2}} = \frac{X_3/2}{\sqrt{({X_1}^2+{X_2}^2)/8}} \sim t(2)$

即：随机变量 $Y = \frac{\sqrt{2}X_3}{\sqrt{{X_1}^2+{X_2}^2}}$服从自由度为2的t分布.

什么样的统计量服从t分布，即t分布的构造？

t分布的性质？

如何理解t分布上α分位点？

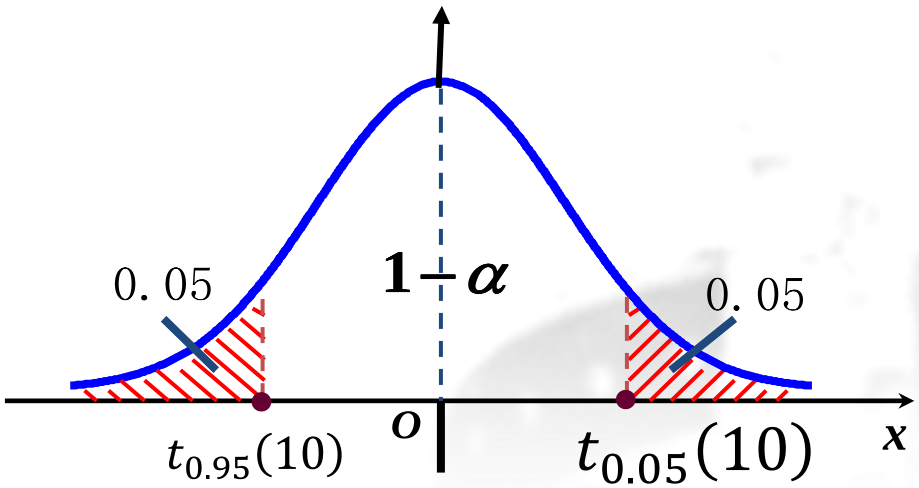
上$1-\alpha$分位点上α分位点关系？

练习题

※设$X \sim t(10)$,若$P\{X > 1.8125\} = 0.05$,则$t_{0.95}(10) = ($ $)$.

(A)1.8125 (B)0.95 (C)-1.8125 (D) -0.95

※参考答案：(C)



第六章 抽样分布

6.3 抽样分布（4）

F分布

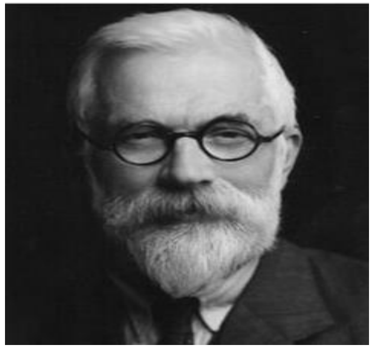
杨建芳

1. F分布的构造定义

★设$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且U,V相互独立.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$F\sim F\ (n_1,n_2)$.

★第一自由度n_1与分子中χ^2分布的自由度一致，第二自由度n_2与分母中χ^2分布的自由度一致.

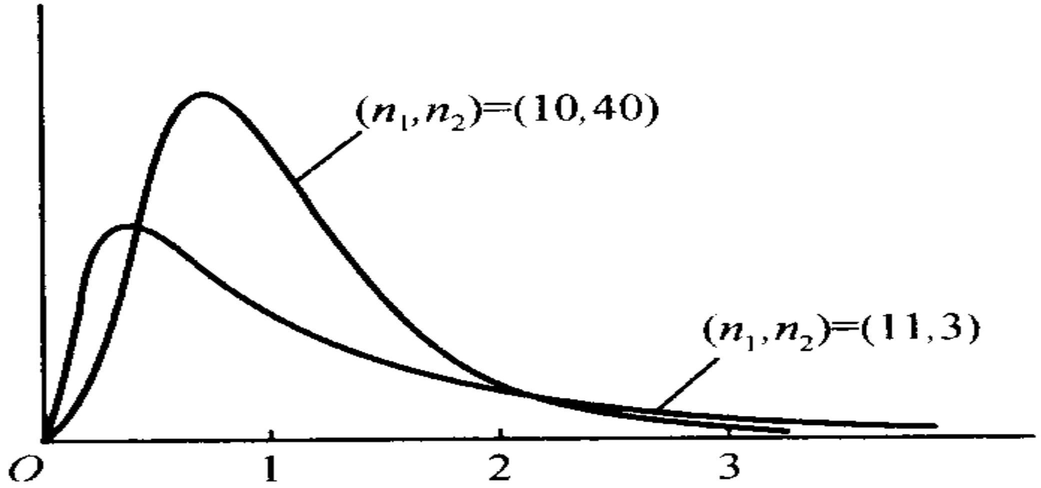


★F分布是1924年英国统计学家费希尔(R.A.Fisher)提出，并以其姓氏的第一个字母命名的.

2. F分布的概率密度、图形、性质

★ F分布的概率密度为（了解）

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



性质：

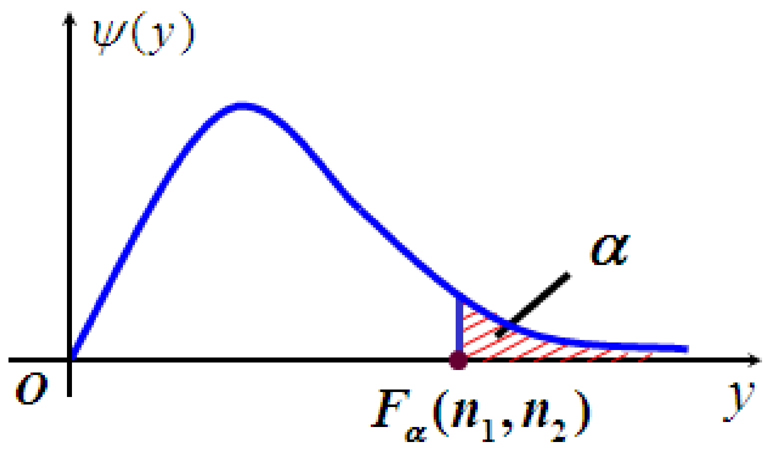
★由F分布的定义，易知：

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

3. F分布的上α分位点

★设$F \sim F\ (n_1, n_2)$，对于给定的正数$\boldsymbol{\alpha}$（$0 < \boldsymbol{\alpha} < \mathbf{1}$），满足条件$P\{F > F_\alpha(n_1, n_2)\} = \alpha$的点$F_\alpha(n_1, n_2)$称为F分布的**上$\alpha$分位点**.

★式$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)dy$



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

2）查表$F_{0.05}(9,15) = 2.59$

$$F_{0.95}(9,15) = \frac{1}{F_{0.05}(15,9)} = \frac{1}{3.01}$$

例 1

★设X_1，X_2，X_3是来自总体$N(0,4)$ 的样本.

试确定随机变量$Y=\frac{2{X_3}^2}{{X_1}^2+{X_2}^2}$ 的分布.

分析：考虑分子分母的分布形式？χ^2分布形式？标准正态分布的平方和？

例 1 解

解：1）分子: $\frac{X_3}{2} \sim N(0,1), U = (\frac{X_3}{2})^2 \sim \chi^2(1)$

2）分母：$\frac{X_1}{2} \sim N(0,1)$，$\frac{X_2}{2} \sim N(0,1), V = (\frac{X_1}{2})^2 + (\frac{X_2}{2})^2 \sim \chi^2(2)$

3）显然U, V相互独立.则由F分布的定义:$\frac{U/1}{V/2} = \frac{2{X_3}^2}{{X_1}^2 + {X_2}^2} \sim F(1,2)$

故：$Y = \frac{2{X_3}^2}{{X_1}^2 + {X_2}^2}$服从自由度为(1,2)的F分布，即$Y \sim F(1,2)$.

什么样的统计量服从F分布，即F分布的构造？

F分布的性质？

如何理解F分布上α分位点？

上$1-\alpha$分位点与上α分位点关系？

练习题

※设$X \sim t(n)$,则X^2的分布是（　　）.

（A） $X^2 \sim F(1,n)$　　（B） $X^2 \sim F(n,1)$

（C） $X^2 \sim \chi^2(1)$　　（D） $X^2 \sim \chi^2(n)$

※参考答案：(A) $X^2 \sim F(1,n)$.

注意符号的含义

随机变量	分布	分位点（数）
X	$N(0,1)$	z_α
χ^2	$\chi^2(n)$	$\chi_\alpha{}^2(n)$
t	$t(n)$	$t_\alpha(n)$
F	$F(n_1,n_2)$	$F_\alpha(n_1,n_2)$

谢谢大家！