第六章　数理统计的基本概念

§6.1总体、样本与统计量

一、基本概念：概率论中问题的讨论，常常从概率存在、$r \cdot vX$的分布、数字特征已知信息出发，讨论它们的种种性质。但在实际问题中,人们事先并不知道事件概率, r·vX的概率分布和数字特征,对它们进行估计与推断构成数理统计的基本问题。

数理统计 $\left\langle \begin{array}{l} \text{试验设计:}\quad\quad\text{怎样抽样的科学} \\ \text{统计推断:估计问题与假设检验} \end{array} \right.$

（未知参数或未知概率分布）

例1.从5000个产品中随机地抽检一个产品，结果可能合格，也可能不合格，X表示合格品个数

X	0	1
P	1－P	P

(X=1)———合格，（X=0）———不合格，但是，P=？事先未知，即0-1分布未知。

问（1）如何求出或近似地求出p值？

（2）若人们根据以往生产经验提出假设："H_0：P=0.65"，那么，是同意这个假设，还是否定这个假设呢？应该用什么方法检验？

（U检，χ^2-检， t-检， F-检）

统计手法：从研究对象的全体元素中**随机地抽取一小部分**进行**观察（or试验）**，然后**用观察得到的资料（or数据）为出发点**，以**概率论的理论为基础对上述问题进行估计或推断**，称之为**统计推断**。

1 总体（母体）：在数理统计中，**把研究对象的全体元素构成的集合**。而**把组成总体的每个元素称为个体**。

有限总体、无限总体。某城市在一定条件下培养的大学生的数集合为无限总体.

便于叙述，一旦**所考察的数量指标明确**后，我们将**总体与其数量指标相应的概率分布等同起来**, 即**总体是一个概率**

分布或服从某个概率分布的 $r\cdot vX$

2 样本：从**总体*X*中随机抽检*n*个体（随机抽样）**，则得 n组观察值$x_1,\ldots x_n$，**称此E为随机抽样，简称抽样**。n为样本容量，若**离开特定的某次抽样即将抽样结果一般化（随机化）**，则抽得结果为n个$r\cdot v$，称这n个$r\cdot vX_1, X_2,\ldots, X_n$,为来自总体X的一个容量为n样本or（$X_1, X_2,\ldots, X_n$）为来自总体X的样本。

n维$r \cdot v\ (X_1, \ldots, X_n)$的$dfF(x_1, \ldots, x_n)$

为**样本的分布**，对应**样本值$(x_1, \ldots, x_n)$**为**样本点，样本点的全体称之为样本空间。**

目的：**数理统计任务即研究如何根据样本来推断总体。**为使抽得样本很好反映总体特性，通常**假定总体X的n次观察是**

在**相同条件下独立重复进行的**。

常用的样本是简单随机样本。

定义1：若$X_1,\ldots,X_n$为来自总体X一个样本，且满足：

（1）X_1，X_2，…，X_n**是相互独立的*n*个*r·v*;**

（2）X_1，X_2，…，X_n**与总体*X*同分布**。

则称$X_1\ldots$，X_n为**简单随机样本**。

注：今后若无特别说明，一般而言，样本X_1，…，X_n是指简单随机样本。

怎样构造函数$T=T(X_1, \ldots, X_n)$，以便把样本中所包含有关信息集中起来，然后再利用之进行统计推断且T为 $r\cdot v$。

3 统计量：（样本$(X_1, \ldots, X_n)$的Borel函数）

定义2 设$X_1, \ldots, X_n$为总体X的容量为n的样本，$T(X_1, \ldots, X_n)$是定义在样本空间上

且不依赖于未知参数的$c \cdot f$, **则称**

$\boldsymbol{T=T(X_1, \ldots, X_n)}$ **为一个统计量。**

例2 **构造统计量与非统计量，**总体

$$X \sim N\left(\mu, \sigma^2\right)$$

μ未知,σ^2已知

$$f_1\left(X_1,\cdots X_n\right) = \sum_{i=1}^{n}\left(X_i - \mu\right)^2 / \sigma^2$$

$$f_2 = X_i + 1$$

$$f_3 = \sum_{i=1}^{n} X_i / \sigma^2$$

$$f_4 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

几个重要统计量：

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

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

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

k阶原点矩：$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$　　k为自然数

k阶中心矩： $B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$,k为自然数

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

为样本二阶中心距。

顺序统计量：设$x_1,\ldots x_n$为样本的一组观察值，将之按大小次序排列得到

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

$X_{(i)}$总以 $x_{(i)}$作为它的取值

称$X_{(i)}$为第i个顺序统计量，

$X_{(1)}$与$X_{(n)}$分别为最小顺序统计量与最大顺序统计量。

样本中位数

$$M=\begin{cases} X\left(\dfrac{n+1}{2}\right) & n\text{为奇数} \\ \dfrac{1}{2}\left\{X\left(\dfrac{n}{2}\right)+X\left(\dfrac{n}{2}+1\right)\right\} & n\text{为偶数} \end{cases}$$

样本极差：　称$R = X_{(n)} - X_{(1)}$为样本极差。

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

经验分布函数。

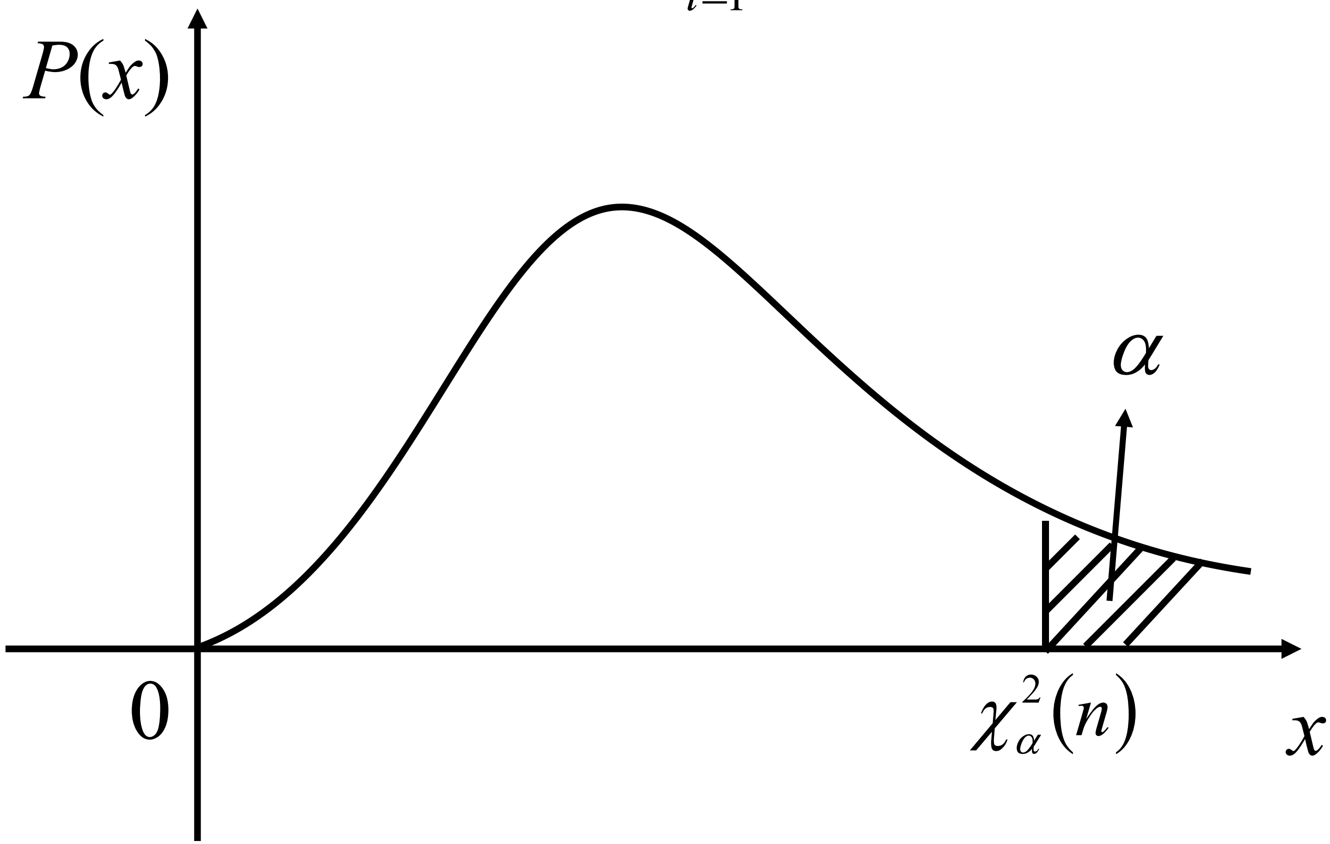
§6.2 三大分布 χ^2, t, F 分布与抽样分布th

一、三大分布

1、 $\chi^2 -$分布

设$X_1,\ldots,X_n$独立同分布， $X_i \sim N(0,1)\left(i = \overline{1,n}\right)$

则它们的平方和 $Z=\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$



χ^2变量性质：若 $\chi^2 \sim \chi^2(n)$,则有

（1）$E\chi^2 = n, D\chi^2 = 2n$

（2）χ^2 −分布之可加性：

若 X，Y 独立且 X~$\chi^2(n_1)$，Y~$\chi^2(n_2)$，

则 $Z = X + Y \sim \chi^2(n_1 + n_2)$（卷积公式）

（3）n 很大时，$Z \sim N(n,2n)$,

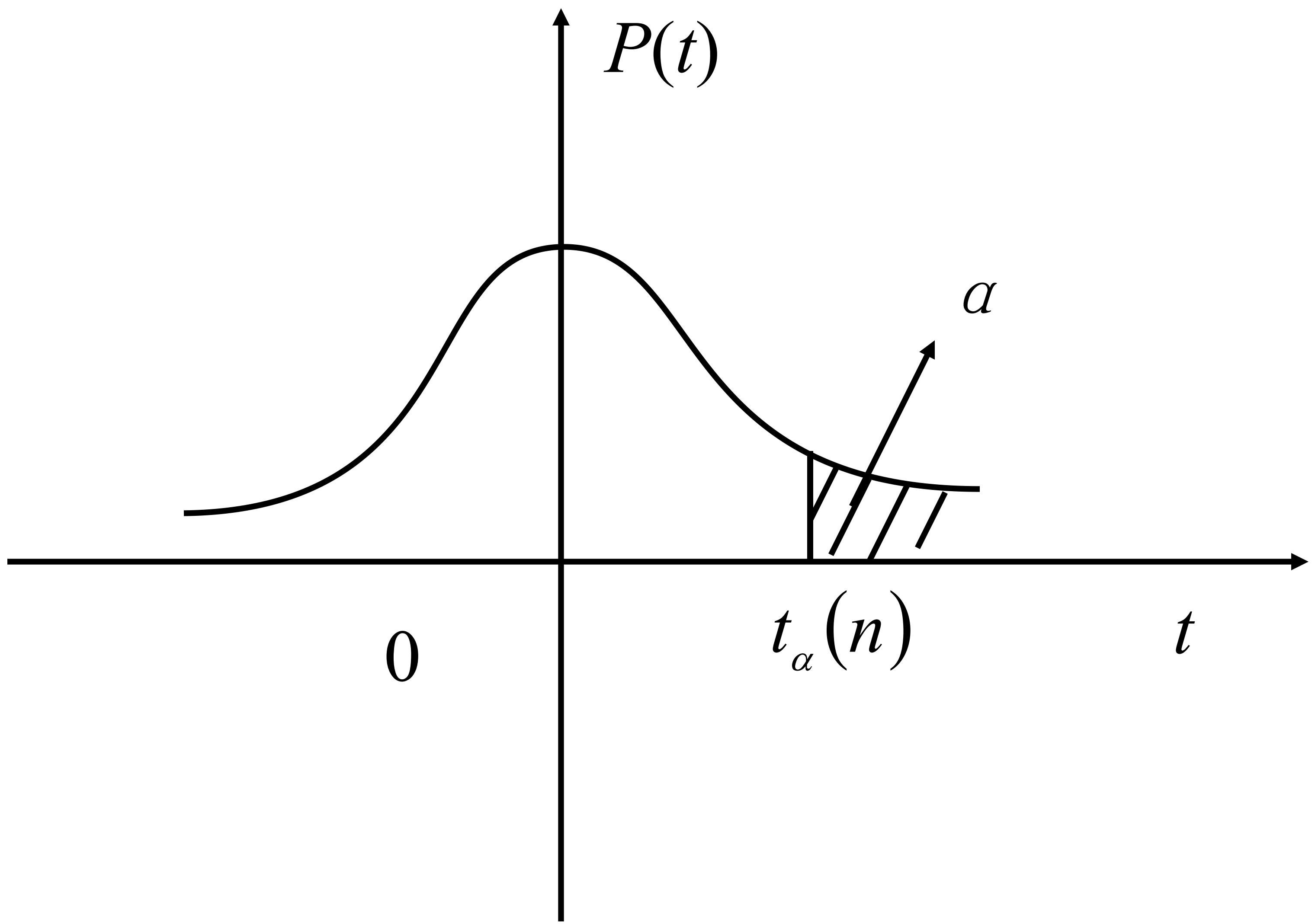
$$\frac{\chi^2-n}{\sqrt{2n}}\sim N(0,1)$$（中心极限 th）。

（4）$\chi^2(n)$ 上侧 α 分位数，若对某些 $\alpha(0<\alpha<1)$ 和不同 n，查附表，∃一个满足等式 $P\left[\chi^2\geq\chi_\alpha^2(n)\right]=\alpha$ 的临界值的数值 $\chi_\alpha^2(n)$，则称临界值 $\chi_\alpha^2(n)$ 的数值为 $\chi_\alpha^2(n)$ 的上侧 α 分位数。

2．t — 分布　:设 X，Y 独立且 $X \sim N(0,1)$, $Y \sim \chi^2(n)$:, 则称 T=X/$\sqrt{Y/n}$ 服从自由度为 n 的 t 分布，记 T~t(n)

（1）T 变量 pdfp(t)曲线十分近似标准正态变量 pdf 曲线,(当$n \geq 30$ 时)

(2)若给出了α和自由度 n，有临界值$t_\alpha(n)$满足等式 P$\left[T \geq t_\alpha(n)\right] = \alpha$ 临界值$t_\alpha(n)$的数值，则称临界值$t_\alpha(n)$的数值为$t(n)$的上侧α分位数。


$P(t)$
α
0
$t_\alpha(n)$
t

3.F-分布：设$X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$且r·vX与Y独立，则称

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

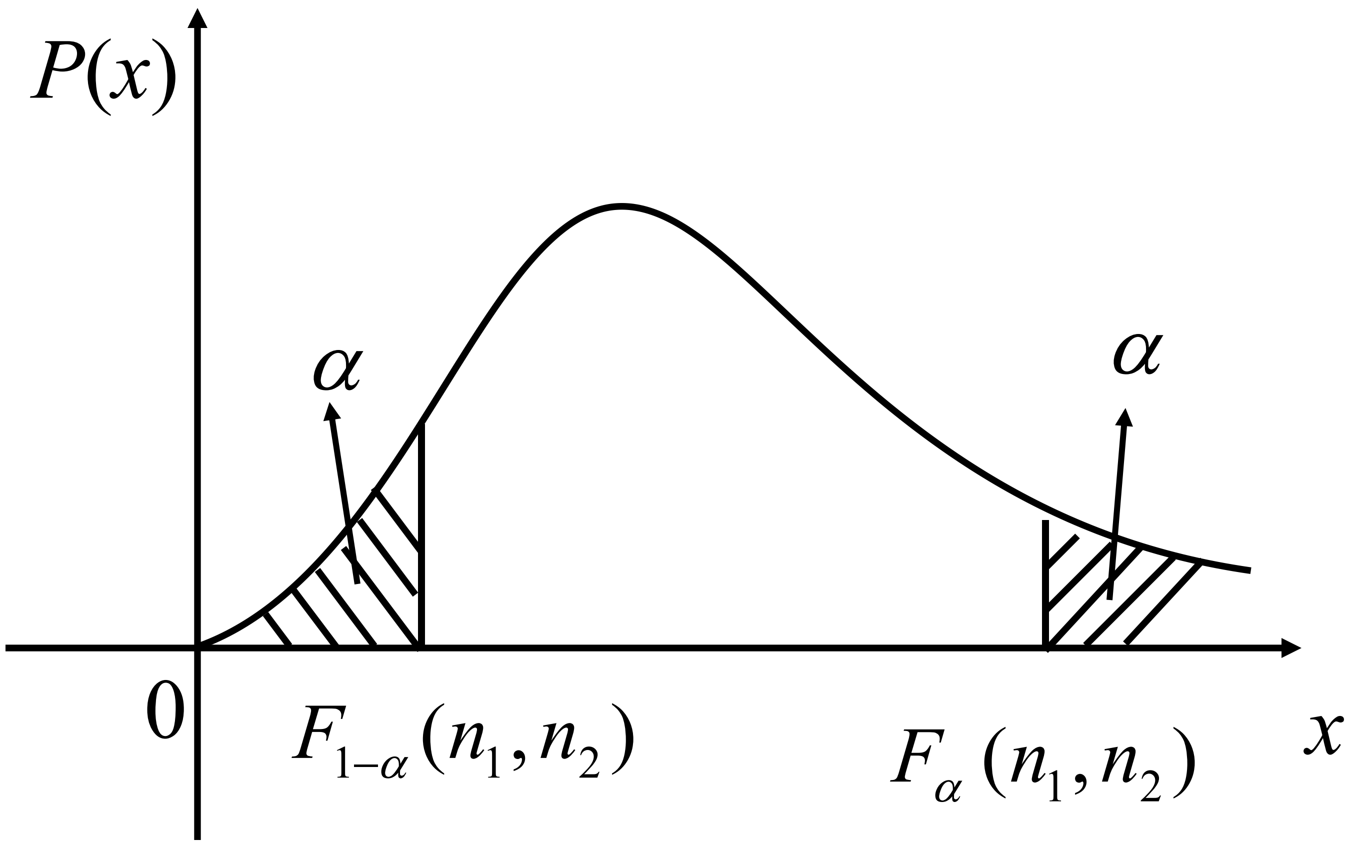
服从第一自由度为n_1，第二自由度为n_2的F-分布，记$F \sim F(n_1,n_2)$。

(1) $F(n_1,n_2)$上侧 α 分位数：对某些n_1，n_2，α $(0<\alpha<1)$，查附表得临界值$F_\alpha(n_1,n_2)$的数值，则称$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$的上侧 α 分位数。

(2) α 分位数性质：

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

$$\text{Pr}oof::\because X\sim x^2(n_1),\quad Y\sim x^2(n_2)$$


$P(x)$
α
α
0
$F_{1-\alpha}(n_1, n_2)$
$F_{\alpha}(n_1, n_2)$
x

且X与Y独立。

$$F=\frac{X/n_1}{Y/n_2}=\frac{n_2}{n_1}\cdot\frac{X}{Y}\sim F(n_1,n_2),$$

$$\frac{Y/n_2}{X/n_1}\sim F(n_2,n_1)$$

于是对 $\forall\alpha(0<\alpha<1)$

有上侧 α 分位数 $F_\alpha(n_2,n_1)$使

$$P\left\{\frac{Y/n_2}{X/n_1} \geq F_\alpha(n_2, n_1)\right\} = \alpha$$

$$而 P\left\{\frac{Y/n_2}{X/n_1} > F_\alpha(n_2, n_1)\right\}$$

$$= p\left\{\frac{X/n_1}{Y/n_2} < \frac{1}{F_\alpha(n_2, n_1)}\right\} = \alpha$$

$$= 1 - P\left\{\frac{X/n_1}{Y/n_2} \geq \frac{1}{F_\alpha(n_2, n_1)}\right\}$$

故
$$P\left\{\frac{X/n_1}{Y/n_2} \geq \frac{1}{F_\alpha(n_2,n_1)}\right\} = 1-\alpha$$

又
$$\frac{X/n_1}{Y/n_2} \sim F(n_1,n_2)$$

故$\dfrac{1}{F_\alpha(n_2,n_1)}$表示$F(n_1,n_2)$上侧 1- α分位数

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

可从分位数$F_{\alpha}(n_2,n_1)$求出分位数$F_{1-\alpha}(n_1,n_2)$

例 $\alpha=0.05, n_1=8, n_2=12$

$$F_{0.05}(8,12)=2.85,$$

$$F_{0.95}(12,8)=\frac{1}{2.85}=0.35$$

二、抽样分布：

前提：总体 X 为正态总体，$(X_1,\cdots X_n)$为来自正态总体 X 的简单随机样本。

Th1（样本均值分布）设 $X_1,\cdots,X_n$ 为总体 $N(\mu,\sigma^2)$ 的一个样本，则样本均值

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

推论：设 $\overline{X}$ 为正态总体 $N(\mu,\sigma^2)$ 样本均值，

则 $U=\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$

Prof： $\because X_1,\cdots,X_n$ 为$N\left(\mu,\sigma^2\right)$ 一个简单随机样本，

$$\therefore \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\text{亦为正态变量}$$

$$\text{而}E\overline{X}=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\times n\mu=\mu$$

$$D\overline{X}=\frac{1}{n^2}\sum_{n^2}^{1}DX_i=\frac{1}{n^2}\times n\sigma^2=\frac{\sigma^2}{n}$$

$$\therefore \overline{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$$

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

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

Th2．（样本方差分布）设 $X_1,\cdots X_n$ 为总体 $N(\mu,\sigma^2)$的一个样本, 则样本方差 S^2 与样本均值$\overline{X}$相互独立，且

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

Th3．设 $X_1,\cdots X_n$ 为总体$N(\mu,\sigma^2)$的一个样本，$\overline{X}$ 与 S^2 分别为样本均值和样本方差，

则 $$\frac{\overline{X}-\mu}{S}\sqrt{n} \sim t(n-1)$$ 其中$S=\sqrt{S^2}$

$$\because \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (Th\,1)$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \qquad (Th2)$$

且$\overline{X}$与 S^2 独立

$$\therefore \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \text{与} \frac{(n-1)s^2}{\sigma^2} \text{亦独立。}$$

$$t = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} / \sqrt{\frac{(n-1)s^2}{\sigma^2} / n-1}$$

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

t-分布定义

Th4设X_1，…X_{n1}和Y_1， … Y_{n2}分别是来自总体 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$ ，它们相互独立，则：

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

其中

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

s_1^2, s_2^2分别为两个样本的样本方差。

由Th1知：

$$\overline{X}\sim N(\mu_1,\frac{\sigma^2}{n_1}),\qquad \overline{Y}\sim N(\mu_2,\frac{\sigma^2}{n_2})$$

$\because \overline{X}$与$\overline{Y}$ 独立，故

$$\overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2)$$

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

由Th2知：

$$\frac{(n_1-1)s_1^2}{\sigma^2}\sim x^2(n_1-1),\qquad \frac{(n_2-1)s_2^2}{\sigma^2}\sim x^2(n_2-1)$$

又∵二者独立，故由 χ^2—分布可加性：

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}\sim x^2(n_1+n_2-2)$$

而t—分布定义

$$t=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

与

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}$$ 相互独立

由t—分布定义

$$t=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\Bigg/\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}\Big/n_1+n_2-2}$$

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

Th5设$X_1, \ldots, X_{n1}$和$Y_1, Y_2, \ldots$，Y_{n2}分别是总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的两个样本，它们相互独立，则：

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

其中 s_1^2和s_2^2 分别是两个样本的样本方差。

由Th2：

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

$\because \dfrac{(n_1 - 1)s_1^2}{\sigma_1^2}$ 与 $\dfrac{(n_2 - 1)s_2^2}{\sigma_2^2}$ 相互独立。

∴由F定义：

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

例3设$X_1,\ldots,X_n,X_{n+1}$来自 $N(\mu,\sigma^2)$ 样本。

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

求统计量

$$T=\frac{X_{n+1}-\overline{X}}{s^{*}}\cdot\sqrt{\frac{n-1}{n+1}}$$ 分布

解：由题设

$$X_{n+1}-\overline{X}\sim N(0,(1+\frac{1}{n})\sigma^{2})$$

$$\frac{X_{n+1}-\overline{X}}{\sqrt{\frac{n+1}{n}}\sigma}\sim N(0,1),$$ 而 $\overline{X}$ 与 s^{2} 独立

于是 $\dfrac{(n-1)s^2}{\sigma^2}$ 与 $\dfrac{X_{n+1}-\overline{X}}{\sqrt{\dfrac{n+1}{n}}\sigma}$ 独立

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

$$(n-1)s^2=ns^{*2}$$

$$T=\frac{X_{n+1}-\overline{X}}{\sqrt{\frac{n+1}{n}}\sigma}\Bigg/\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/n-1}$$

$$=\frac{X_{n+1}-\overline{X}}{\sqrt{\frac{n+1}{n}}\sigma}\Bigg/\sqrt{\frac{nS^{*2}}{\sigma^2}\Big/n-1}$$

$$=\frac{X_{n+1}-\overline{X}}{S^*}\sqrt{\frac{n-1}{n+1}}\sim t(n-1)$$

补充习题

E9~12,18~20

第六章数理统计初步总结

1、三个概念；

2、三大分布；

3、五个抽样分布定理。