第 6 章教学要求：

1. 理解总体、个体、简单随机样本及其分布的概念.

2. 了解直方图和条形图、经验分布函数. 会求经验分布函数.

3. 了解样本均值、样本方差与样本标准差、样本原点矩、平均数、众数、中位数和极差等数字特征，并会根据数据计算这些数字特征.

4. 了解 $\chi^2$ 分布、$t$ 分布、$F$ 分布和分位数的概念. 会查表计算上 $\alpha$ 分位数.

5. 理解统计量的概念，掌握来自正态总体的抽样分布.

# 第 6 章　样本与抽样分布

在概率论中，一切的分析和运算都是基于分布已知这个假设进行的. 但在实际问题中，情况往往并非如此，我们常常对所要研究的随机变量知道不多或知之甚少，这时需要经试验或观测，获得反映随机变量信息的数据，并以概率论为理论基础，对数据进行整理、分析，从而对研究对象的性质和统计规律做出合理、科学的估计和推断. 这就是数理统计基本的和主要的任务.

数理统计研究统计的一般原理与方法.

本章主要介绍数理统计中的基本概念、基本分布和正态抽样分布及性质.

## §6.1 总体与样本

### 6.1.1 总体和个体

在数理统计中，概括性地说，研究对象的全体称为**总体**，总体中的每个元素称为**个体**.

例如，研究……学生，……总体，……个体.

在实际中，我们说研究……学生，一般是带有“目的性”的. 如：我们想研究学生的……，根据这种“目的性”，我们研究对象的全体就具体为“……”、个体则为“……”. 因此，也说——**总体**是研究对象的某数量指标.

记数量指标为 $\boldsymbol{X}$，则 $\boldsymbol{X}$ 是随机变量.

例如，若 $\boldsymbol{X}$ 是表示学生的……数量指标，那么每个个体的指标值即为 $\boldsymbol{X}$ 的取值.

注意到，在进行研究时，个体的指标值事先是不知道的，我们一般是通过“随机抽样”的方式来获得个体的指标值及有关情况——即总体 $\boldsymbol{X}$ 的取值及其分布的. 因此，数量指标 $\boldsymbol{X}$ 是一个随机变量.

总体中所包含的个体的数量称为**总体容量**. 根据总体容量的有限或无限，分为**有限总体**和**无限总体**.

### 6.1.2 样本和简单随机样本

通常人们以随机抽样的方式来了解总体分布.

把从总体中抽取出的一部分个体称为总体的一个**样本**，样本中的个体称为**样品**，样本中所包含的个体的数目称为**样本容量**.

通过样本来了解总体，则样本应该具有代表性.

如何获得具有代表性的样本？

获得具有代表性的样本最常采用的方法是：在相同的条件下，对总体 $X$ 进行 $n$ 次重复且独立的随机观测，把 $n$ 次观测的结果按试验的次序依次记为 $X_1, X_2, \cdots, X_n$ . 采用这种有放回抽取得到的样本 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量，与总体 $X$ 有相同的分布，因此具有代表性，这个样本称为**简单随机样本**. 样本的一次观测值记为 $x_1, x_2, \cdots, x_n$ ，称为样本的一组**样本值**.

如果没有特别说明，今后讨论中提到的样本均指简单随机样本，并简称样本.

怎样才能获得简单随机样本呢？

对有限总体，采用有放回抽取方式得到的样本是简单随机样本，但有放回抽取在实际应用中有时不方便.

采用不放回抽取方式取得的样本不是简单随机样本，当总体容量比样本容量大很多时，可以把它当作简单随机样本.

对无限总体，抽走少量样本不影响总体的构成或影响很小，因此常采用不放回抽取方式.

### 6.1.3 样本的联合分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X$ 的样本.

如果总体 $X$ 的分布函数为 $F_X(x)$ ，那么样本 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_X(x_i) . \tag{6.1}$$

当 $X$ 为连续型随机变量且概率密度函数为 $f_X(x)$ 时，样本 $X_1, X_2, \cdots, X_n$ 的联合概率密度函数为

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_X(x_i) . \tag{6.2}$$

当 $X$ 为离散型随机变量且概率分布为 $f_X(x) = P\{X = x\}$（$x$ 取遍 $X$ 的所有可能的取值）时，样本 $X_1, X_2, \cdots, X_n$ 的联合概率分布为

$$f(x_1, x_2, \cdots, x_n) = P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} f_X(x_i) . \tag{6.3}$$

**例 6.1** 设总体 $X$ 服从参数为 $p$ 的 0-1 分布，则 $X$ 的概率分布为

$$f_X(x)=P\{X=x\}=p^x(1-p)^{1-x},\ x=0,1.$$

于是样本 $X_1,X_2,\cdots,X_n$ 的联合概率分布为

$$f(x_1,x_2,\cdots,x_n)=\prod_{k=1}^{n}f_X(x_k)=p^{s_n}(1-p)^{n-s_n},$$

其中 $x_k(k=1,2,\cdots,n)$ 取1或0，而 $s_n=\sum_{k=1}^{n}x_k$ .

**例 6.2** 设总体 $X\sim N(\mu,\sigma^2)$，则样本 $X_1,X_2,\cdots,X_n$ 的联合概率密度函数为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f_X(x_i)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

### 6.1.4 直方图与条形图

1. 直方图

在实际统计工作中，首先接触到的是一系列数据，样本数据的整理是统计研究的基础．整理数据最常用的方法之一是给出数据的频数分布表或频率分布表．在研究连续性随机变量总体的分布时，通常用由一组矩形组成的图形直观地表示频数、频率分布表，这个图形就称为样本的**频率直方图**，简称**直方图**．

绘制直方图的一般步骤见书 P165.

**例 6.3**(例 6.3 P166)

2．条形图

如果样本数据很少，以及在研究离散型随机变量总体的分布时，一般用条形图来直观地反映频数、频率分布表．

条形图的制作方法：（1）统计出样本观测值 $x_1,x_2,\cdots,x_n$ 的频数与频率表；（2）在 $x$ 轴上的每个 $x_i$ 处画出以样本数据中出现 $x_i$ 的频数或频率为高度的一条线段．

**例 6.4**(例 6.4 P167)

### 6.1.5 经验分布函数

根据总体 $X$ 的样本值 $x_1,x_2,\cdots,x_n$，按如下方法可以得到一个函数．

将总体 $X$ 的样本观测值 $x_1,x_2,\cdots,x_n$ 由小到大重新排序为

$$\underbrace{x_{(1)},\cdots,x_{(1)}}_{n_1},\underbrace{x_{(2)},\cdots,x_{(2)}}_{n_2},\cdots,\underbrace{x_{(m)},\cdots,x_{(m)}}_{n_m},$$

其中 $x_{(1)}<x_{(2)}<\cdots<x_{(m)}$，$n_1+n_2+\cdots+n_m=n$．定义函数

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}n_i, & x_{(k)}\le x<x_{(k+1)}\ (k=1,2,\cdots,m-1),\\ 1, & x\ge x_{(m)}.\end{cases}\qquad (6.4)$$

$F_n(x)$称为总体 $X$（关于样本观测值 $x_1,x_2,\cdots,x_n$）的**经验分布函数**或**样本分布函数**.

对于给定的样本值 $x_1,x_2,\cdots,x_n$，经验分布函数 $F_n(x)$ 与总体分布函数 $F(x)$ 有类似的性质.

**例 6.5**（例 6.5 P$_{168}$）

**定理 6.1（格利文科（Glivenko）定理）** 对于任意实数 $x$，经验分布函数 $F_n(x)$ 以概率 1 一致收敛于总体分布函数 $F(x)$，即

$$P\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|=0\}=1.\qquad (6.5)$$

格利文科定理表明，当样本容量充分大，经验分布函数可以很好地近似总体分布函数. 这一结论是数理统计依据样本来推断总体特征的理论基础.

作业（P$_{169}$）：1.　2.-3.

## §6.2 样本的数字特征

### 6.2.1 样本的基本数字特征

设 $X_1,X_2,\cdots,X_n$ 是来自总体 $X$ 的一个样本，$x_1,x_2,\cdots,x_n$ 是样本的观测值.

**样本均值**　$\bar{X}=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i$；

观测值　$\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$.

**样本方差**　$S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$；

观测值　$s^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2$.

**样本标准差**　$S=\sqrt{S^2}=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}$；

观测值　$s=\sqrt{s^2}=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$.

**样本 $k$ 阶原点矩**　$A_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^k,\quad k=1,2,\cdots$；

观测值　$a_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^k,\quad k=1,2,\cdots$.

**样本 $k$ 阶中心矩**　$B_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^k,\quad k=2,3,\cdots$；

观测值 $b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,\quad k=1,2,\cdots.$

显然，$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)=0$，样本一阶原点矩是样本均值.

容易证明：$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$.

显然，样本的二阶中心矩与样本方差之间有如下关系：

$$B_2=\frac{n-1}{n}S^2.$$

6.2.2 样本的简易数字特征

这里介绍平均数、众数、中位数和极差. 设 $x_1,x_2,\cdots,x_n$ 是采集到的一组样本数据.

1. 平均数

平均数是样本数据的平均值.

2. 众数

众数是指样本数据中出现频数最高的数值，用 $M_0$ 表示.

众数可以不唯一. 比如出现频数最高的位置特征有两个，那么众数就有两个. 如数据 2, 1, 2, 3, 1, 3, 1, 4, 3, 5, 2, 6 的众数有 1, 2, 3 三个. 如果数据的分布没有明显的集中趋势或不存在频数最高峰值，众数也可能不存在. 例如，测得 10 名运动员的体重（单位：公斤）分别为 76, 77, 82, 84, 83, 83.5, 81, 82.5, 90, 95，其值各不相同，故认为不存在众数. 由于这些数都集中在 83 左右，故可以说：83 公斤左右的人比较集中.

**例 6.6**（例 6.6 P171）

3. 中位数

中位数是指样本数据按大小排序后处于“中间”位置上的数值，用 $M_e$ 表示.

考虑到观测数有偶数或奇数两种情况，中位数用如下方法确定.

设 $x_1,x_2,\cdots,x_n$ 为来自总体 $X$ 的样本观测值，将样本观测值由小到大进行排序，记为 $x_{(1)},x_{(2)},\cdots,x_{(n)}$，则中位数定义为

$$M_e=\begin{cases}x_{(\frac{n+1}{2})}, & \text{当}n\text{为奇数},\\ \frac{1}{2}(x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}), & \text{当}n\text{为偶数}.\end{cases}\qquad(6.6)$$

**例 6.7**（例 6.7 P172）

4. 极差

极差是指样本数据的最大值与最小值之差.

例如，

作业($P_{172}$)：1.(2)、(4)-(7)　3.　4.　6.　5.

## §6.3 三个常用的抽样分布

$\chi^2$ 分布、$t$ 分布和 $F$ 分布都是由正态分布导出的分布，是试验统计中的常用分布，在数理统计中占有十分重要的地位.

1. $\chi^2$ 分布（卡方分布）

**定义 6.1**　若 $X_1, X_2, \cdots, X_n$ 相互独立且都服从 $N(0,1)$，则随机变量

$$X = X_1^2 + X_2^2 + \cdots + X_n^2$$

所服从的分布称为自由度为 $n$ 的 $\chi^2$ **分布**，记作 $X \sim \chi^2(n)$.

$\chi^2(n)$ 分布的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{6.7}$$

其中 $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \ (\alpha > 0)$ 是 $\Gamma$ 函数.

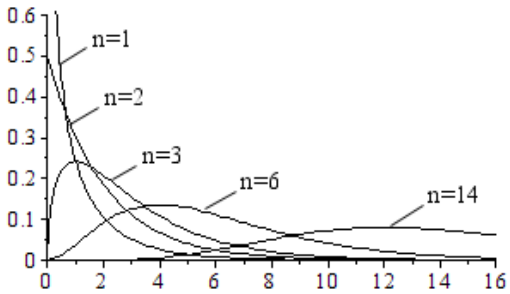


图 6.3　$\chi^2(n)$ 分布的概率密度函数图像

从图 6.3 可以看出，随着自由度 $n$ 的增大，$\chi^2(n)$ 分布的密度函数图形逐渐变得对称.

$\chi^2$ 分布有如下两条重要性质.

**性质 1**　若 $X \sim \chi^2(n), Y \sim \chi^2(m)$，且 $X$ 与 $Y$ 相互独立，则 $X + Y \sim \chi^2(n+m)$.

**性质 2**　若 $X \sim \chi^2(n)$，则 $E(X) = n, D(X) = 2n$.

**定义 6.2** 设 $X \sim \chi^2(n)$，对于任意给定的 $\alpha(0<\alpha<1)$，使

$$P\{X > \chi_\alpha^2(n)\} = \alpha \tag{6.8}$$

成立的 $\chi_\alpha^2(n)$ 称为自由度为 $n$ 的 $\chi^2$ 分布的**上 $\alpha$ 分位数**.

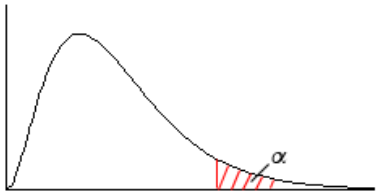


图 6.4 $\chi^2$ 分布的上 $\alpha$ 分位数示意图

$\chi^2$ 分布的上 $\alpha$ 分位数可通过查 $\chi^2$ 分布表(见附录 4)得到.

当 $n>45$ 时，上 $\alpha$ 分位数的近似值

$$\chi_\alpha^2(n) \approx n + z_\alpha\sqrt{2n}\,, \tag{6.9}$$

其中 $z_\alpha$ 是标准正态分布的上 $\alpha$ 分位数.

2. ***t*** 分布

**定义 6.3** 若 $X$ 与 $Y$ 相互独立，且 $X \sim N(0,1)$, $Y \sim \chi^2(n)$，则随机变量

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

所服从的分布称为自由度为 $n$ 的 ***t* 分布**，记作 $T \sim t(n)$.

$t(n)$ 分布的概率密度函数为

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < +\infty\,. \tag{6.10}$$

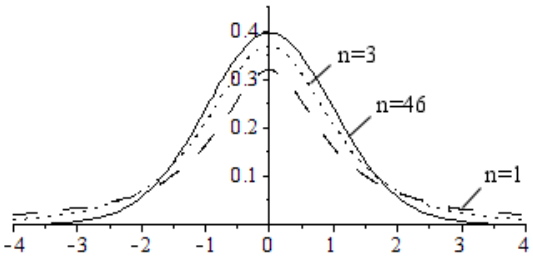


图 6.5 $t(n)$ 分布的概率密度函数图像

从图 6.5 可以看出，$t(n)$ 分布的密度函数的图形关于 $y$ 轴对称，且当 $n$ 很大时，几乎与标准正态分布的密度函数曲线重合.

$t$ 分布的性质：若 $T \sim t(n)$，则 $E(T)=0(n>1), D(T)=\dfrac{n}{n-2}(n>2)$.

**定义 6.4**　设 $T \sim t(n)$，对于任意给定的 $\alpha(0<\alpha<1)$，使

$$P\{T > t_\alpha(n)\} = \alpha \tag{6.11}$$

成立的 $t_\alpha(n)$ 称为自由度为 $n$ 的 $t(n)$ 分布的**上 $\alpha$ 分位数**.

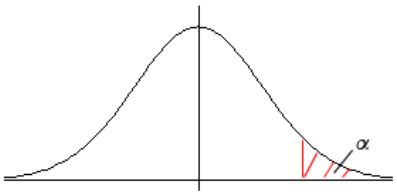


图 6.6　$t$ 分布的上 $\alpha$ 分位数示意图

$t(n)$ 分布的上 $\alpha$ 分位数可通过查 $t$ 分布表(见附录 5)得到，且由对称性知

$$t_{1-\alpha}(n) = -t_\alpha(n). \tag{6.12}$$

当 $n>45$，$t$ 分布接近标准正态分布，$t_\alpha(n) \approx z_\alpha$.

### 3. $F$ 分布

**定义 6.5**　若 $X$ 与 $Y$ 相互独立，且 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，则随机变量

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

所服从的分布称为第一自由度为 $n_1$、第二自由度为 $n_2$ 的 $F$ **分布**，记作 $F \sim F(n_1, n_2)$.

$F(n_1, n_2)$ 分布的概率密度函数为

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0, \\ 0, & x \le 0. \end{cases} \tag{6.13}$$

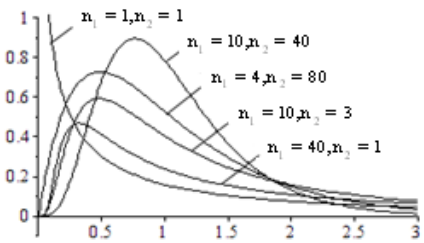


图 6.7　$F(n_1, n_2)$ 分布的概率密度函数图像

$F$ 分布的性质：若 $F \sim F(n_1, n_2)$，则 $\dfrac{1}{F} \sim F(n_2, n_1)$.

**定义 6.6**　设 $F \sim F(n_1, n_2)$，对于任意给定的 $\alpha(0 < \alpha < 1)$，使

$$P\{F > F_\alpha(n_1, n_2)\} = \alpha \tag{6.14}$$

成立的 $F_\alpha(n_1, n_2)$ 称为第一自由度为 $n_1$、第二自由度为 $n_2$ 的 $F$ 分布的**上 $\alpha$ 分位数**.

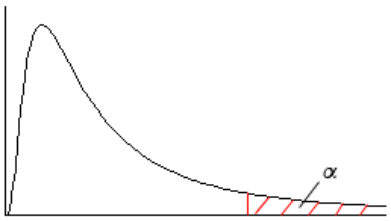


图 6.8　$F$ 分布的上 $\alpha$ 分位数示意图

$F$ 分布的上 $\alpha$ 分位数的值可通过查 $F$ 分布表(见附录 6)得到，且

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}. \tag{6.15}$$

例如，由附表 6 可以查得 $F_{0.05}(10,11) = 2.85$，然后利用(6.15)式得

$$F_{0.95}(11,10) = \frac{1}{F_{0.05}(10,11)} = \frac{1}{2.85} = 0.3509.$$

作业($P_{177}$)：1.-6.

## §6.4 常用统计量及其分布

### 6.4.1 统计量的概念

样本是总体的代表，包含有总体的许多信息，但这些信息比较分散，一般不能直接用于对总体进行统计推断，必须进行整理、加工，将有用的信息集中起来.

以样本作为自变量的样本函数就是样本信息集中的一种表现，人们就是利用它来对总体进行统计推断的.

**定义 6.7**　设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X$ 的样本，$x_1, x_2, \cdots, x_n$ 是样本的观测值，而 $g(t_1, t_2, \cdots, t_n)$ 为一个已知的 $n$ 元函数. 若样本函数 $g(X_1, X_2, \cdots, X_n)$ 中不含有未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为样本的**统计量**，称 $g(x_1, x_2, \cdots, x_n)$ 为**统计量的观测值**.

在 §6.2 中介绍的样本的基本数字特征都是统计量. 但 $\sum_{i=1}^{n} c_i X_i$ (其中 $c_i (i = 1, 2, \cdots, n)$ 是未知参数)不是统计量.

### 6.4.2 来自正态总体的常用统计量及其分布

正态分布是应用最为广泛的一种分布，而由正态总体的样本均值与样本方差所构成的统计量的分布则是统计推断中最常采用的抽样分布.

**定理 6.2** 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本， $\bar{X}$ 为样本均值，则有

(1) $\bar{X} \sim N(\mu, \sigma^2/n)$；

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

(3) $\sum\limits_{i=1}^{n}\left(\dfrac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$. $F(n_1, n_2)$.

**定理 6.3** 若 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，其样本均值和样本方差分别为 $\bar{X}=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i$ 和 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$，则

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

(2) $\bar{X}$ 与 $S^2$ 相互独立；

(3) $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

**定理 6.4** 若 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 是来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 相互独立的两个样本，$\bar{X}, \bar{Y}$ 和 $S_1^2, S_2^2$ 分别是两个样本的样本均值和样本方差，则有

(1) $\bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2, \dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$；

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

(3) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，

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

其中 $S_\omega^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，$S_\omega=\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$.

**例 6.8** (P179)

**例 6.9** (P179)

作业 (P180)：2. 1. 3. -4.