

第五章



统计量及其分布

引论：数理统计学

□ 统计学：

统计学是一门数据收集、表达、整理与分析，并进行推断的科学。在信息时代，统计学扮演着非常重要的角色。

- 农业（田间试验设计和统计分析）；
- 工业（试验设计、质量控制、可靠性）；
- 经济与金融（经济预测、企业管理、金融风险定量分析）
- 医学与制药（病因、药品研制、疫苗、安全性、效果）
- 互联网（数据挖掘、推荐算法、搜索）。

□ 统计学两大核心问题：

如何收集数据？如何分析数据以获得信息和知识？

引论：数理统计学

➤ 大数据：新资源、新机会

- ✓ 由于科技的进步，各领域产生了海量和复杂的数据，人类进入**大数据 BigData** 时代。
- ✓ 大数据是人类自身产生的一种**新的“自然”资源**，一种基础性生产要素。与支撑传统经济发展的自然资源，如土地、石油、水等不同，这种人造“自然”资源越用越多，越用越有价值。数据资源是人力资源和物质资源外的第三大资源！
- ✓ 大数据是一场革命，庞大的数据资源使得各个领域开始**量化进程**（学术界、商业、政府）。

统计学与数据科学

- 大数据时代，一切皆可量化，一切皆可记录。如何利用全面、及时、经济的网络电子化数据，通过使用新的分析与挖掘技术，产生新的认识，是我们的**重大机遇**。
- 如何分析如此庞大和复杂的数据？如何在海量数据中寻找新的现象和发现新的规律？
- 数据科学必须提高到与自然科学并列的高度。用数据的方法研究科学，用科学的方法研究数据。
- 大数据时代，数据的重要作用更加凸显，许多国家都把大数据提升到国家战略的高度。

大数据时代的统计学

- 大数据不能被直接拿来使用，统计学是数据分析的灵魂
- 大数据告知信息但不解释信息。像股票市场，即使把所有的数据都公布，不懂的人依然不知道数据代表的信息。
- 大数据对数据分析提出了全新挑战：
- ✓ 传统统计方法应用到大数据上，巨大计算量和存储量使其难以承受；
- ✓ 对结构复杂、来源多样的数据，如何建立有效的统计学模型？
- 传统处理分析能力不能完全知道这些海量数据中蕴含的规律。大数据是统计的新战场、新领域。

大数据时代的统计学

➤ 对统计学而言，这或许是最好的时代

信息技术的发展让海量数据触手可及。大数据相关产业高速发展，如生物医药、金融、移动互联网、车载信息技术。数据科学的巨大需求正在悄然兴起，统计学正大有可为。

➤ 对统计学而言，这或许是最坏的时代

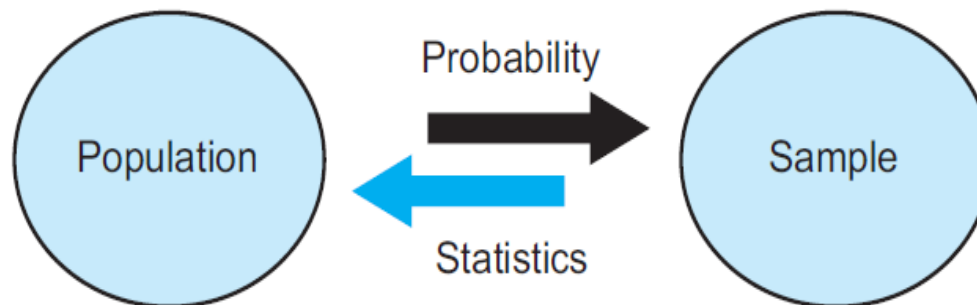
大数据相关科学与产业发展太过迅速，随之而来的是具有复杂结构的海量的**庞大数据集**。现有的统计模型无法拟合非结构化数据；无法算出大规模数据下的最大似然估计。在大数据相关产业与科学面前，**统计学正面临着严峻挑战**。

From Probability to Statistics

Probability and Statistics are **inverse processes** of each other.

Probability: Given the rules that govern uncertainty (distribution), what can we say about outcomes?

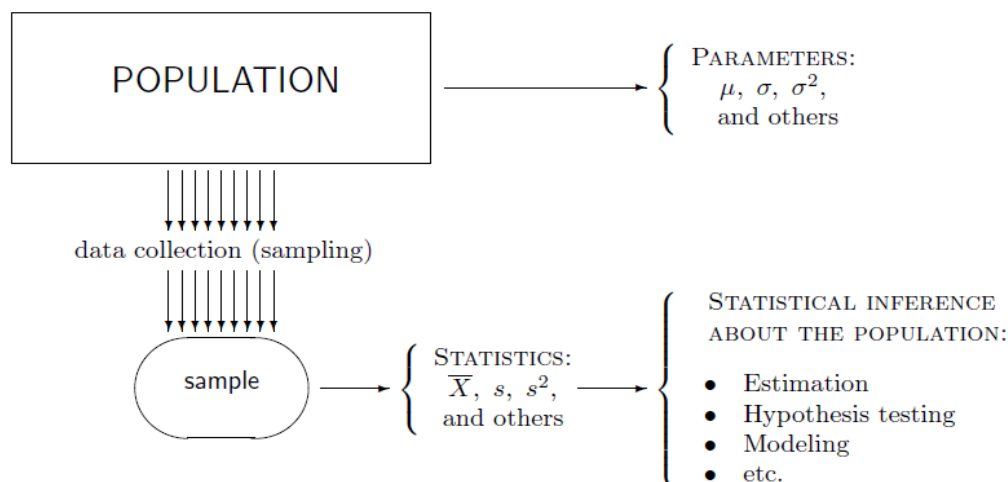
Statistics: Given a set of outcomes, what can we say about the distribution?



数理统计 vs 概率论

- **概率论** 侧重随机现象的内在逻辑体系和一般规律；
给定分布，研究分布的性质、规律与数字特征。
- **统计学** 研究数据的收集、整理与分析，通过数据来研究规律、发现规律。对随机现象(未知分布或参数)作出估计或推断。概率论是统计学的基础。

如：某股票价格，服从对数正态分布 $\ln S \sim N(\mu, \sigma^2)$
如何根据观测值对其中的参数作出估计？



本章主要内容

- § 5.1 总体与样本
- § 5.2 经验分布函数
- § 5.3 统计量及其分布
- § 5.4 三大抽样分布

第一节：总体与样本

总体：研究对象的全体（某项数量指标），
数量化后对应某RV X , 其分布称为总体的分布。

个体：总体中的每个元素（成员）。

例：研究清华大学学生身高，则
总体：清华全体学生（的身高）；
个体：每一个清华学生（的身高）。

例：研究某厂产品质量，则
总体：该厂全部产品（的质量指标）；
个体：每个产品（的质量指标）。

样本与样本的二重性

□ 样本：

从总体中随机抽取的部分个体的集合。

记为： X_1, \dots, X_n . n 称为样本容量。

□ 样本的二重性：

➤ 随机性

抽样前，无法预知样本的具体值，用RV X_1, \dots, X_n 表示。

➤ 确定性

取样后，得到 n 个确定的观测值，用 x_1, \dots, x_n 表示。

简单随机样本

要使得推断可靠，要选择正确的样本，使样本能很好地代表总体（否则，可能存在统计被误用的情况）。

□ 简单随机样本：

对总体进行 n 次重复的、独立的观测所得到的样本。

- 随机性：每个个体有同等机会入选；
- 独立性：每一样本的取值不影响其它样本的取值。
- ✓ 对有限总体，有放回抽样可得简单随机样本；
- ✓ 对无限总体（或大容量有限总体），无放回抽样亦可（近似）得到简单随机样本。

样本联合分布函数

设总体 X 具有分布函数 $F(x)$, X_1, \dots, X_n 为简单随机样本, 则样本的联合分布就是随机向量 (X_1, \dots, X_n) 的联合分布:

$$F(x_1, \dots, x_n) = \prod_{i=1}^n F(x_i).$$

若总体 X 具有密度函数 $f(x)$, 则随机向量 (X_1, \dots, X_n) 的密度函数为

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i).$$

第二节：经验分布函数

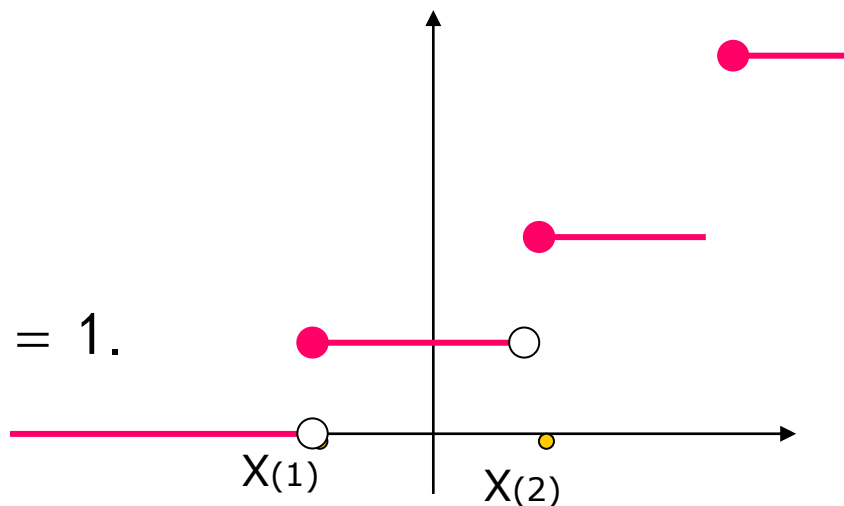
总体的分布函数为 $F(x)$. n 个观测值： x_1, x_2, \dots, x_n

定义 $F_n(x) = \frac{1}{n} \{\# x_i \leq x\}$.

称 $F_n(x)$ 为经验分布函数。

满足：

非减、右连续、 $F_n(-\infty) = 0$, $F_n(+\infty) = 1$.

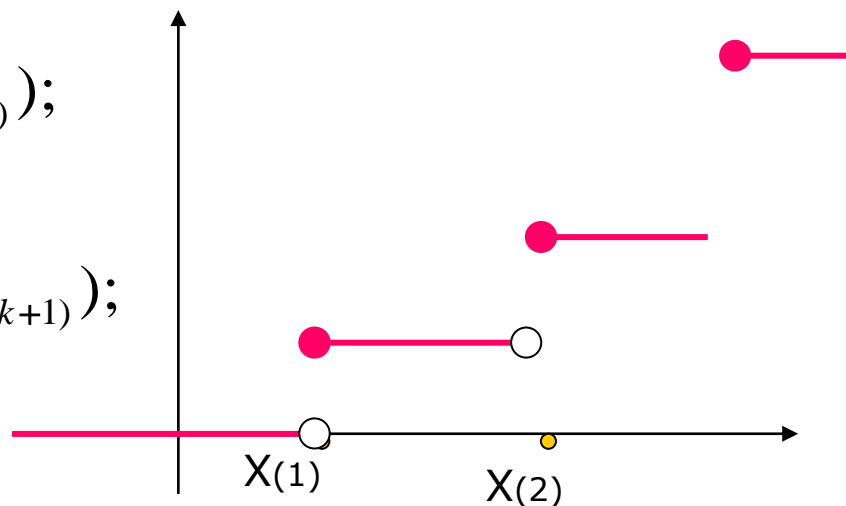


第二节：经验分布函数

总体的分布函数为 $F(x)$. n 个观测值： x_1, x_2, \dots, x_n

从小到大排序： $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ --- 有序样本；

$$\text{定义 } F_n(x) = \begin{cases} 0, & x < x_{(1)}; \\ 1/n, & x \in [x_{(1)}, x_{(2)}); \\ \dots & \\ k/n, & x \in [x_{(k)}, x_{(k+1)}); \\ \dots & \\ 1, & x \geq x_{(n)}. \end{cases}$$



Note: If there are r observations with the same value x , F_n has a jump of height r/n at x .

注记:

经验分布函数是下述RV的分布函数:

Y	$x_{(1)}$	$x_{(2)}$	\cdots	$x_{(n)}$
P	$1/n$	$1/n$	\cdots	$1/n$

由Bernoulli大数律, $F_n(x) \xrightarrow{P} F(x)$. (频率 \xrightarrow{P} 概率)

(样本有随机性, 经验分布函数亦随机)

定理: 设总体的分布函数为 $F(x)$, X_1, \dots, X_n 是来自该总体的样本, $F_n(x)$ 是其经验分布函数, 则当 $n \rightarrow +\infty$ 时, 有

$$P\left\{ \sup_{-\infty < x < +\infty} |F_n(x) - F(x)| \rightarrow 0 \right\} = 1.$$

了解*

表明: 当 n 很大时, 经验分布函数是总体分布函数的一个
一致良好近似。

样本数据的整理与显示

Before you do anything with a data set, look at it!

- Probability model, i.e., a family of dist to be used;
- Statistical methods suitable for the given data;
- Presence or absence of outliers;
- Presence or absence of heterogeneity;
- Existence of time trends and other patterns;
- Relation between two or several variables;
- ...

样本数据的整理与显示

样本数据的整理是统计研究的基础。

1. 频数—频率分布表

整理数据的最常用方法之一是给出其**频数分布表**或**频率分布表**。

2. 样本数据的图形显示

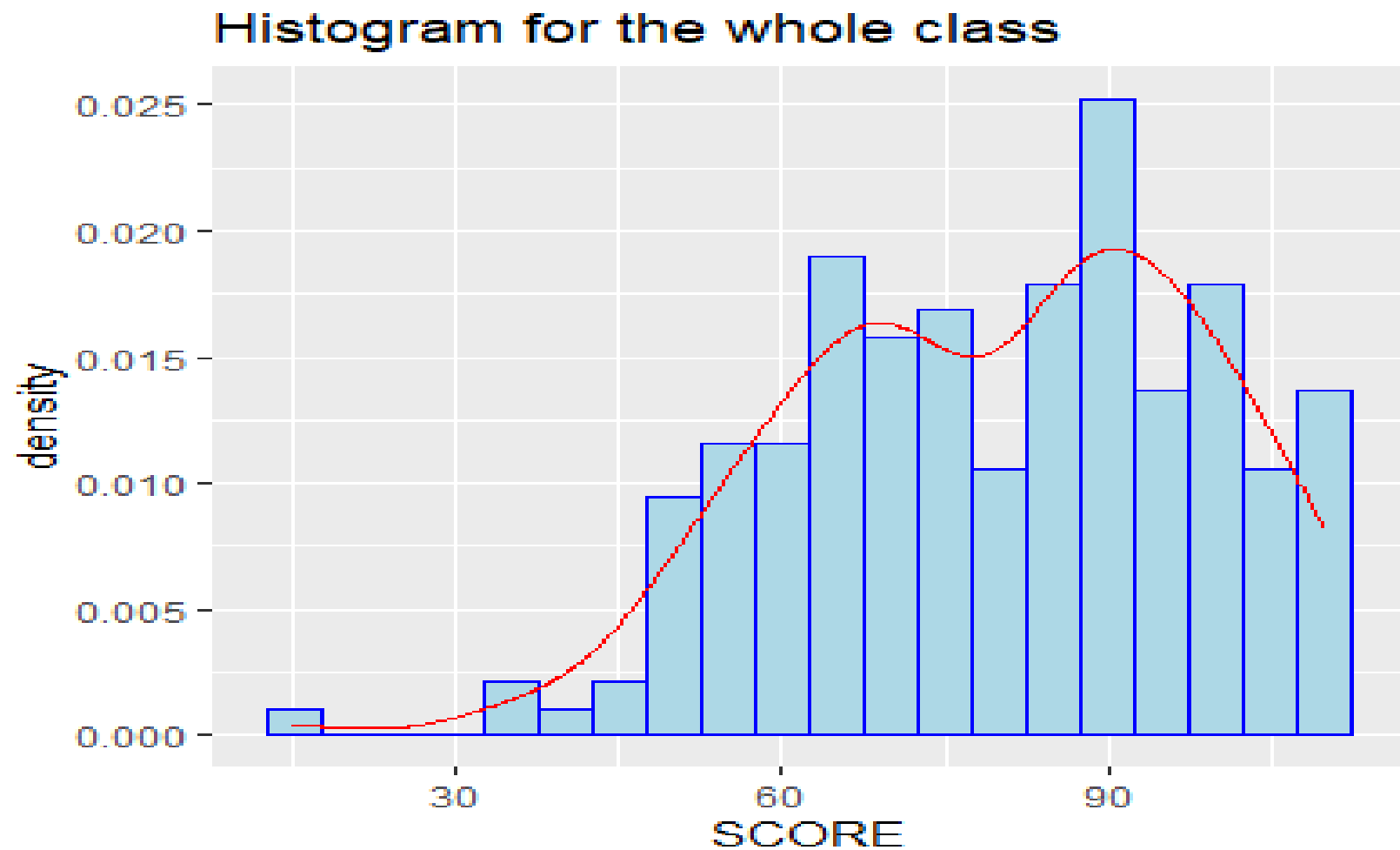
□ **直方图 (Histogram)**

直方图是频数分布的图形表示，

- 横坐标表示所关心变量的取值区间，
- 纵坐标有三种表示方法：**频数**，**频率**，**频率/组距**。

□ **茎叶图 (Stem-and-leaf plots)**

Histogram for Midterm Test



第三节：统计量及其分布

一、统计量

目标：通过样本，构造适当的函数，对总体进行估计与推断.

定义：设 X_1, \dots, X_n 是来自总体 X 的一组样本， $T(X_1, \dots, X_n)$ 是样本 X_1, \dots, X_n 的函数，**且不含未知参数**，则称 T 为统计量，统计量的分布称为**抽样分布**。



统计量起到从样本到总体的推断中的**桥梁**作用。

注记：统计量具有二重性：

- **随机性：** $T(X_1, \dots, X_n)$ 为 RV (X_1, \dots, X_n 为RV)；
- **确定性：** $T(x_1, \dots, x_n)$ 为统计量的观测值
(x_1, \dots, x_n 为样本观测值)

例:

设总体 $X \sim N(\mu, \sigma^2)$, 其中 μ 已知, σ^2 未知。则

$$\frac{1}{3} \sum_{i=1}^3 X_i, \quad X_1 X_2 + 2\mu \quad \text{为统计量};$$

$(X_1^2 + X_2^2 + X_3^2) / \sigma^2$ 不是统计量(含有未知参数)。

注记:

尽管统计量不依赖于未知参数, 但是它的分布一般是依赖于未知参数的。统计量的概率分布称为**抽样分布**。

二、样本均值 (Sample Mean)

定义：设 X_1, \dots, X_n 是来自该总体 X 的样本， x_1, \dots, x_n 为观测值，
样本均值定义为

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{为样本均值的观测值。}$$

性质1: $\sum_{i=1}^n (X_i - \bar{X}) = 0$ (样本所有“偏差”之和为0)

证: $\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - n\bar{X} = 0.$

数据与样本均值之差

性质2: $\min_c \sum_{i=1}^n (x_i - c)^2$ 的最优解为: $c = \bar{x}$.

观测值与样本均值的偏差平方和最小。

比较: $\min_c E(X - c)^2$ 的最优解为: $c = EX$.

$$\begin{aligned}\text{证: } \sum_{i=1}^n (x_i - c)^2 &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - c)]^2 \\&= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - c)^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(\bar{x} - c) \\&= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - c)^2 + 2(\bar{x} - c) \sum_{i=1}^n (x_i - \bar{x}) \\&= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - c)^2 \geq \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

等号当且仅当 $c = \bar{x}$ 时成立。

样本均值的抽样分布

定理:

设总体 X 的均值、方差存在, 即 $E(X) = \mu, D(X) = \sigma^2$, X_1, \dots, X_n 是来自该总体 X 的样本, x_1, \dots, x_n 为观测值, \bar{X} 为样本均值。则

$$(1) E(\bar{X}) = \mu,$$

$$(2) D(\bar{X}) = \sigma^2 / n.$$

The square root law:

The standard deviation of the sample mean is proportional to $1/\sqrt{n}$.

$$\text{证: (1) } E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

$$(2) D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \left(\sum_{i=1}^n D(X_i)\right) = \sigma^2 / n.$$

样本均值的抽样分布

定理：设 X_1, \dots, X_n 是来自总体 X 的样本, \bar{X} 为样本均值。

(1) 若总体为 $N(\mu, \sigma^2)$, 则 \bar{X} 的精确分布为 $N(\mu, \sigma^2 / n)$;

(2) 对任意分布, 若 $E(X) = \mu, D(X) = \sigma^2$,

则 (n 较大时) \bar{X} 的近似分布为 $N(\mu, \sigma^2 / n)$;

证: (1) 由正态分布的可加性由卷积公式推得

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2) \Rightarrow \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2 / n).$$

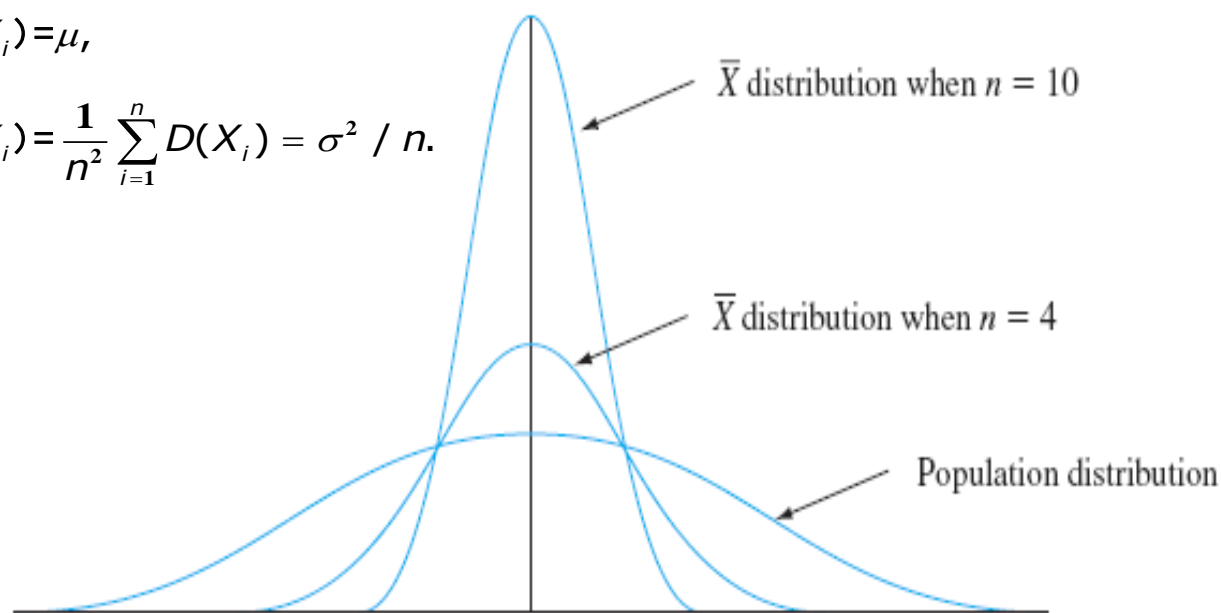
(2) 由中心极限定理 (n 充分大时)

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \stackrel{\text{近似}}{\sim} N(0,1) \Rightarrow \bar{X} \stackrel{\text{近似}}{\sim} N(\mu, \sigma^2 / n).$$

样本均值的抽样分布

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \mu,$$

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \sigma^2 / n.$$



A normal population distribution and \bar{X} sampling distributions

Sample mean becomes more and more concentrated around the expected value μ as n increases.

Special Case: Sample Proportion

If $X \sim B(n, p)$, then the **sample proportion**

$$\bar{p} = X/n$$

has **approximately normal** distribution (for large n)

$$\bar{p} \sim N(p, p(1-p)/n).$$

(np and $n(1-p)$ must each be at least 5)

相当于总体是 0-1 两点分布: $X = X_1 + \dots + X_n$

三、样本方差 (Sample Variance)

定义：设 X_1, \dots, X_n 是来自该总体 X 的样本， x_1, \dots, x_n 为观测值，

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



$S^* = \sqrt{S^{*2}}$ 为样本标准差， $s^* = \sqrt{s^{*2}}$ 为其观测值。

(无偏)样本方差： $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.



样本方差简便计算

偏差平方和

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n X_i^2 + \sum_{i=1}^n \bar{X}^2 - 2 \sum_{i=1}^n X_i \bar{X} \\&= \sum_{i=1}^n X_i^2 + \sum_{i=1}^n \bar{X}^2 - 2n\bar{X}^2 \\&= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\&= \sum_{i=1}^n X_i^2 - \frac{(\sum X_i)^2}{n}.\end{aligned}$$

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

定理:

设总体 X 的均值、方差存在, 即 $E(X) = \mu, D(X) = \sigma^2$, X_1, \dots, X_n 是来自该总体 X 的样本, x_1, \dots, x_n 为观测值, \bar{X}, S^2 分别为样本均值和样本方差。

则 $E(S^2) = \sigma^2$.

证明:

注意到: $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2,$

及 $E(X_i^2) = [E(X_i)]^2 + D(X_i) = \mu^2 + \sigma^2,$

$$E(\bar{X}^2) = [E(\bar{X})]^2 + D(\bar{X}) = \mu^2 + \sigma^2 / n,$$

有 $E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right]$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right]$$
$$= \frac{1}{n-1} \left[n(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2 / n) \right]$$
$$= \frac{1}{n-1} \left[(n-1)\sigma^2 \right] = \sigma^2.$$

四、样本矩 (Sample Moment)

定义：设 X_1, \dots, X_n 是来自该总体 X 的样本， x_1, \dots, x_n 为观测值，

样本 k 阶原点矩：
$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k,$$

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

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

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

样本矩与总体矩的关系

定理：设总体 X 的 k 阶矩存在： $E(X^k) = \mu_k$ ，则

$$(1) A_k \xrightarrow{P} \mu_k.$$

$$(2) g(A_1, \dots, A_k) \xrightarrow{P} g(\mu_1, \dots, \mu_k) \text{ (} g \text{ 为连续函数)}.$$

证：样本 X_1, \dots, X_n 独立、同分布(与总体 X 同分布)

则 X_1^k, \dots, X_n^k 独立、同分布(与 X^k 同分布)且

$$E(X_1^k) = \dots = E(X_n^k) = \mu_k,$$

$$\text{由辛钦大数律, } A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mu_k.$$

四、样本矩

- 当总体关于分布中心对称时，用 \bar{x} 和 s 刻画样本特征有代表性；
- 当其不对称时，只用 \bar{x} 和 s 就显得很不够。需要一些刻画分布形状的统计量，如**样本偏度**和**样本峰度**，它们都是样本中心矩的函数。

四、样本矩

$$B_3 / B_2^{3/2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{B_2^{1/2}} \right)^3.$$

定义：设 X_1, \dots, X_n 是样本，则称统计量

$\beta_s = B_3 / B_2^{3/2}$ 为样本偏度；

$\beta_t = B_4 / B_2^2 - 3$ 为样本峰度。

消除量纲
的影响

- **样本偏度** (是否为0) 反映了总体分布密度曲线的**对称性**信息
(明显大于0时，样本的右尾长，及样本中有几个较大的数)。
- **样本峰度** 反映了总体分布密度曲线在其峰值附近的**陡峭**程度 (明显大于0时，比正态陡峭，称为尖顶型)。

标准正态的偏度和峰度分别是多少？

对比：分布的偏度系数和峰度系数

□ 如何刻画分布偏离对称性的程度？

定义：设随机变量 X 的前三阶矩存在，则称

以正态分布为基准

$$\beta_s(X) = E\left(\frac{X - \mu}{\sigma}\right)^3$$

为随机变量的偏度系数(是无量纲的量)

$\beta_s(X) > 0$: 正偏（右偏）； $\beta_s(X) < 0$: 负偏（左偏）。

□ 如何刻画分布的尖峭程度和尾部粗细程度？

定义：设随机变量 X 的前四阶矩存在，则称

$$\beta_k(X) = E\left(\frac{X - \mu}{\sigma}\right)^4 - 3 \quad \text{为随机变量的峰度系数 (是无量纲的量)}$$

$\beta_k(X) > 0$: 比 $N(0,1)$ 更尖峭，尾部更粗。

$\beta_k(X) < 0$: 比 $N(0,1)$ 更平坦，尾部更细。

五、次序统计量 (Order Statistics)

将样本 X_1, X_2, \dots, X_n 从小到大排序(对每个 ω)得到:

$$X_{(1)}(\omega) < X_{(2)}(\omega) < \dots < X_{(n)}(\omega),$$

称 $X_{(i)}(\omega)$ 为第 i 个次序统计量。

特别, $X_{(1)}(\omega)$ 为最小次序统计量;

$X_{(n)}(\omega)$ 为最大次序统计量。

第 i 个次序统计量:

将样本观测值由小到大排列后得到的第 i 个观测值

定理：设总体有密度 $p(x)$,分布函数 $F(x)$,则第 k 个
次序统计量 $X_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} \cdot [1-F(x)]^{n-k} \cdot p(x).$$

证：对任意 x ,考虑 $X_{(k)} \in (x, x + \Delta x]$. 

$P\{X_{(k)} \in (x, x + \Delta x]\} = P\{X_1, \dots, X_n \text{ 中恰有 } (k-1) \text{ 个落在 } (-\infty, x],$
恰有一个落在 $(x, x + \Delta x]$, 恰有 $(n-k)$ 个落在 $(x + \Delta x, +\infty)\}$

$$= \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} \cdot [F(x + \Delta x) - F(x)] \cdot [1 - F(x + \Delta x)]^{n-k}.$$

两边除 Δx , 并令 $\Delta x \rightarrow 0$, 即得。

多项分布

特例： $k=1, p_1(x) = n[1-F(x)]^{n-1} \cdot p(x) \Rightarrow F_1(x) = 1 - [1-F(x)]^n.$

$k=n, p_n(x) = n[F(x)]^{n-1} \cdot p(x) \Rightarrow F_n(x) = [F(x)]^n.$

与以前结果一致

例：

设有样本容量为9 的来自 $U(0, 1)$ 的简单随机样本。
样本中位数的分布如何？

利用 $k=5$ 时的公式，有

$$f_{(5)}(x) = \frac{9!}{4!4!} x^4 (1-x)^4 = 630x^4(1-x)^4.$$

五、次序统计量

□ 对任意多个次序统计量可给出其联合分布。

***不要求**

□ 次序统计量的函数在实际中经常用到。如

➤ 样本极差 $R_n = X_{(n)} - X_{(1)}$,

➤ 样本中程 $[X_{(n)} - X_{(1)}]/2$ 。

六、样本中位数与样本分位数

设 $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ 为次序统计量；

样本中位数定义为 $M_{0.5} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n \text{ 为奇数;} \\ \frac{1}{2} \left(X_{(n/2)} + X_{(n/2+1)} \right), & n \text{ 为偶数。} \end{cases}$

(如： $n = 7, M_{0.5} = X_{(4)}$ ； $n = 8, M_{0.5} = (X_{(4)} + X_{(5)}) / 2$ ；)

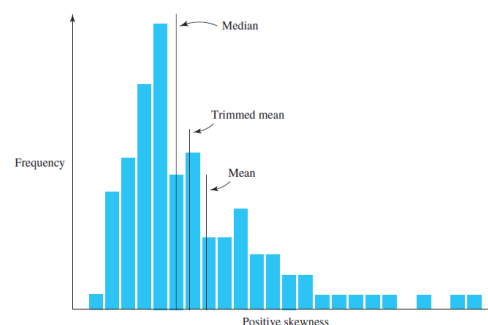
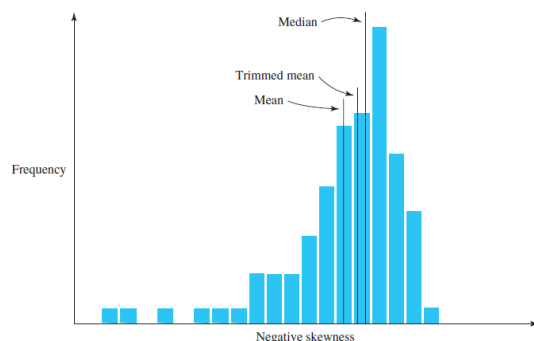
样本 p 分位数定义为 $p = \begin{cases} X_{([np]+1)}, & np \text{ 不是整数;} \\ \frac{1}{2} \left(X_{(np)} + X_{(np+1)} \right), & np \text{ 是整数。} \end{cases}$

如： $n = 10, p = 0.35, M_{0.35} = X_{(4)}$ ；

$n = 20, p = 0.45, M_{0.45} = (X_{(9)} + X_{(10)}) / 2$ 。

样本均值 vs 样本中位数

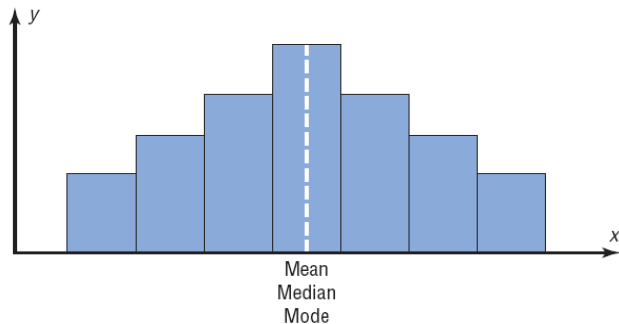
- 样本均值在概括数据方面具有一定的优势。
- 但当数据中含有**极端值 (离群值)**时，使用**中位数**比使用均值更好。**中位数**的这种抗干扰性称为**稳健性**。
- A **symmetric** sample has a sample mean and a sample median roughly equal.
- A sample with **negative skewness** has a sample mean smaller than the sample median.
- A sample with **positive skewness** has a sample mean larger than the sample median



Mean, Median, Mode

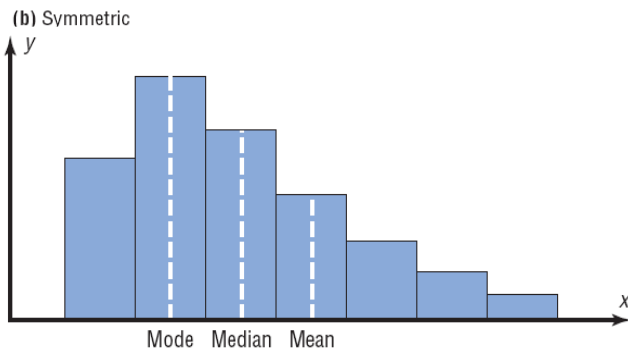
- ▣ **The Mean** $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$
 - The mean for a data set is unique.
 - The mean is affected by extremely high or low values, called outliers.
- ▣ **The Median** $\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$
 - The center or middle value of a data set.
 - The median is affected less than the mean by extremely high or extremely low values.
- ▣ **The Mode**
 - The mode is the most typical case (occur most often)
 - The mode is not always unique (can have more than one mode, or may not exist).

Shape and Mean, Median, Mode



Evenly distributed on both sides of mean. When dist is unimodal, mean, median, and mode are the same (at the center of the dist.)

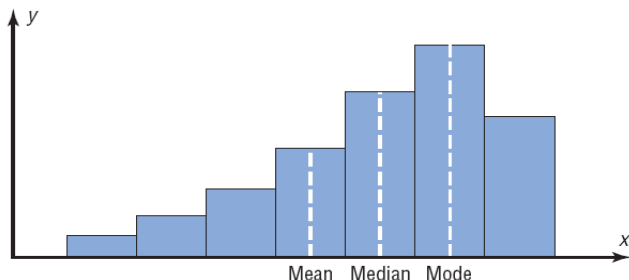
Examples: IQ scores, heights of adult males.



Majority of the data values fall to the left of the mean, with the tail to the right. The mean is to the right of the median, and the mode is to the left of the median.

Examples: Income

(a) Positively skewed or right-skewed



Majority of the data values fall to the right of the mean, with the tail to the left. The mean is to the left of the median, and the mode is to the right of the median.

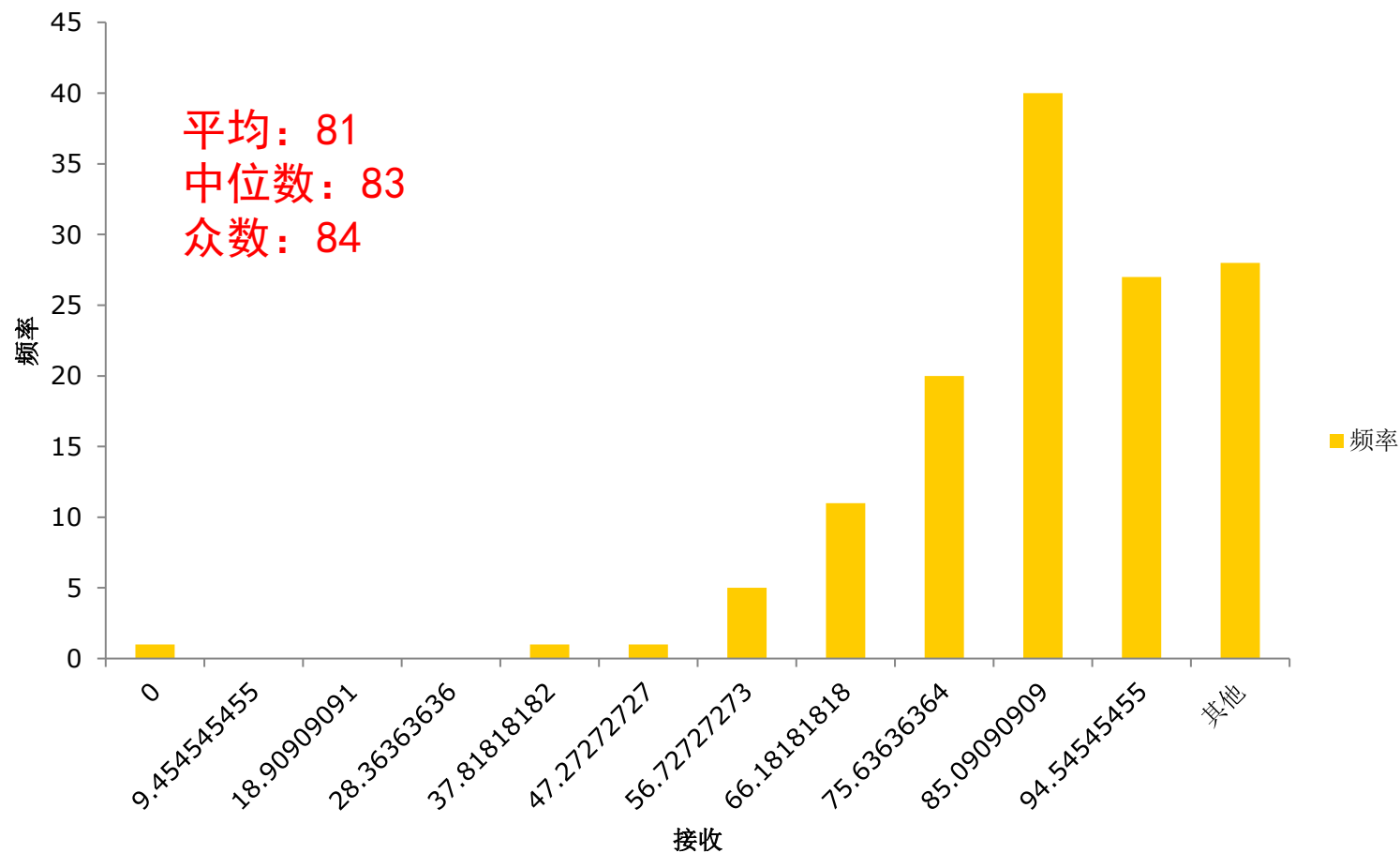
Examples: Student scores of THU

(c) Negatively skewed or left-skewed

Real Data (Student scores of THU)



直方图

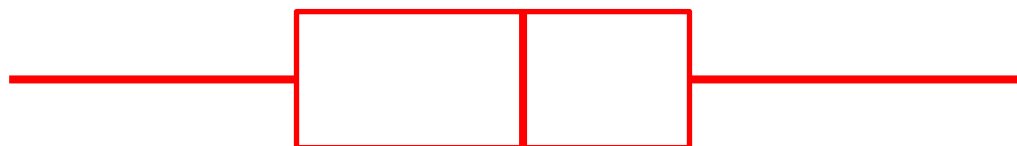


五数概括与箱线图 (Five-number summary and boxplot)

次序统计量的应用之一是五数概括与箱线图。

在得到有序样本后，容易计算如下五个值：

1. 最小观测值
2. 最大观测值
3. 中位数
4. 第一四分位数
5. 第三四分位数

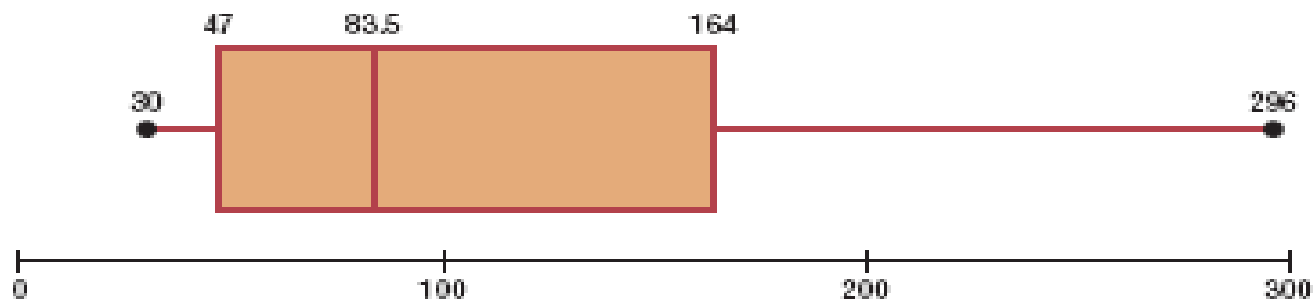


最小 第一四分位数 中位数 第三四分位数 最大

五数概括就是指用这五个数来大致描述一批数据的轮廓。

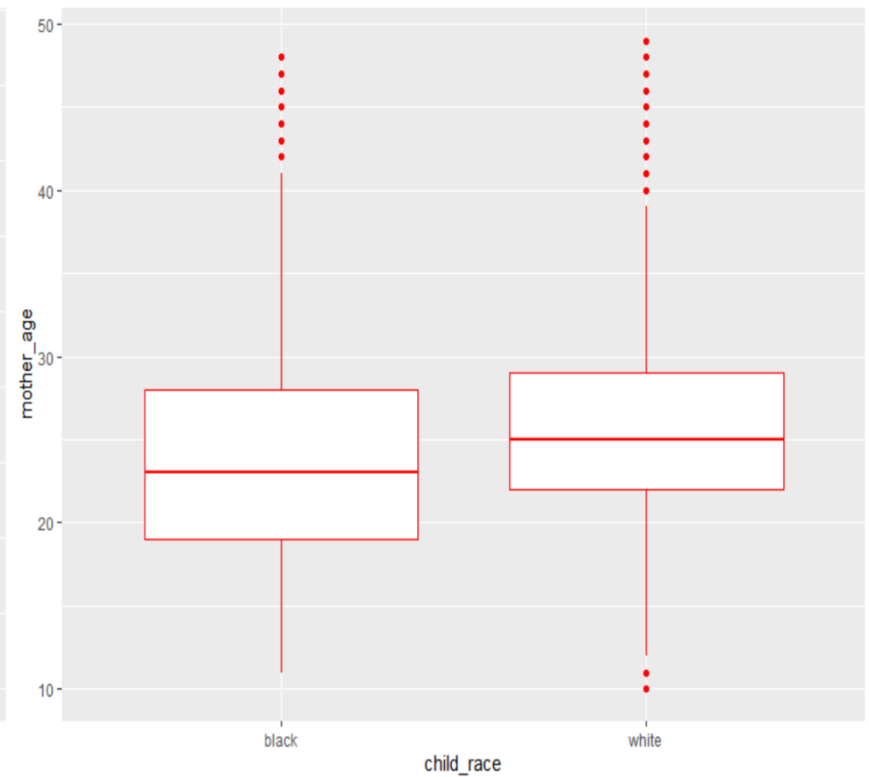
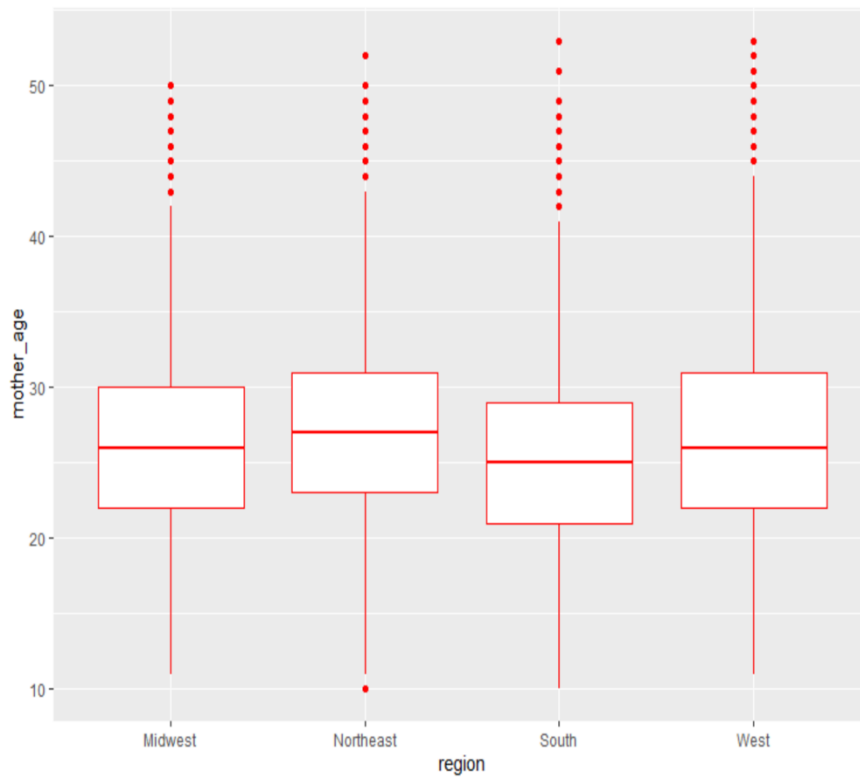
Procedure for constructing boxplot

1. Find the **five-number summary**: maximum and minimum, $Q1$ and $Q3$, and the median.
2. Draw a **horizontal axis with a scale** such that it includes maximum and minimum data values.
3. Draw a **box** whose vertical sides go through $Q1$ and $Q3$, and draw a vertical line through median.
4. Draw **line from minimum to left** side of box and **line from maximum to the right** side of box.



The distribution is somewhat positively skewed.

US birth data (mother age)



第四节：三大抽样分布

(Three Sampling Distributions)

很多统计推断是基于**正态分布**假设的，以**标准正态变量**为基石而构造的三个著名统计量在实际中有广泛的应用。

这三个统计量有明确背景，被称为统计学中的**三大抽样分布**

注意：

三大分布均从正态变量衍生出来。

如果数据不服从正态分布，使用三大分布有时是不合适的。

一、 χ^2 分布

1. 定义：设 X_1, \dots, X_n 是来自总体 $N(0,1)$ 的样本
(即它们相互独立，同分布，均服从 $N(0,1)$)，则称

$$\chi^2 := X_1^2 + \dots + X_n^2$$

的分布为自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$.

χ^2 的密度：

$$X_i \sim N(0,1) \Rightarrow X_i^2 \sim \mathbf{Ga}(1/2, 1/2)$$

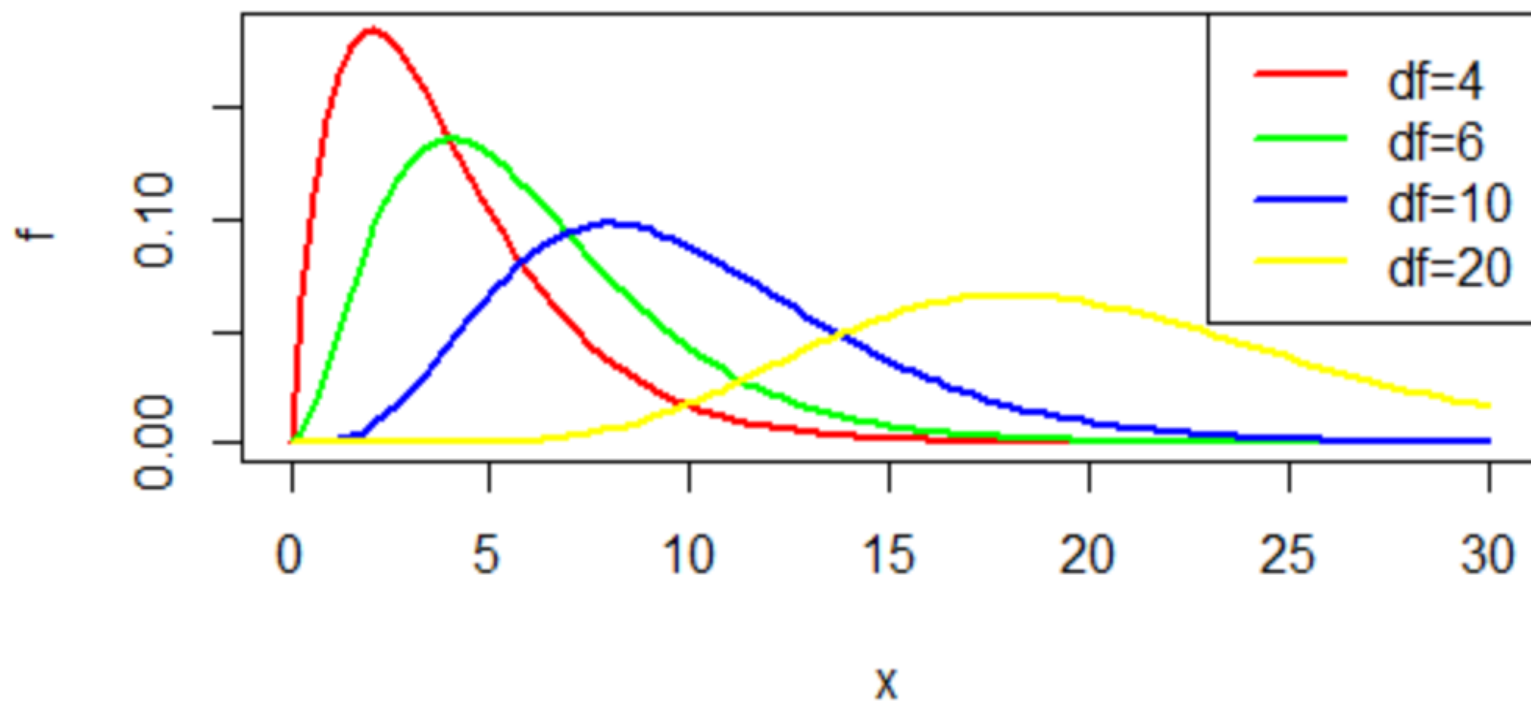
因 X_1^2, \dots, X_n^2 独立,由卷积公式(或 Γ 分布的可加性)

$$\chi^2 = X_1^2 + \dots + X_n^2 \sim \mathbf{Ga}(n/2, 1/2).$$

(Γ 分布: page 115)

χ^2 分布是取非负值的偏态分布

Chi-square Density



2. χ^2 分布的均值与方差

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

证：因 $X_i \sim N(0,1)$

$$\Rightarrow D(X_i) = E(X_i^2) - [E(X_i)]^2 \Rightarrow E(X_i^2) = 1.$$

$$\begin{aligned} E(\chi^2) &= E(X_1^2 + \cdots + X_n^2) \\ &= E(X_1^2) + \cdots + E(X_n^2) = n. \end{aligned}$$

$$\text{又 } D(\chi^2) = D(X_1^2) + \cdots + D(X_n^2),$$

$$D(X_i^2) = \underline{E(X_i^4)} - [E(X_i^2)]^2 = 3 - 1 = 2,$$

$$\Rightarrow D(\chi^2) = 2n.$$

回顾：分位数

设连续型**RV** X 的密度函数为 $f(x)$,分布函数为 $F(x)$,
对任意 $p \in (0,1)$,称满足条件

$$F(x_p) = \int_{-\infty}^{x_p} f(x)dx = p$$

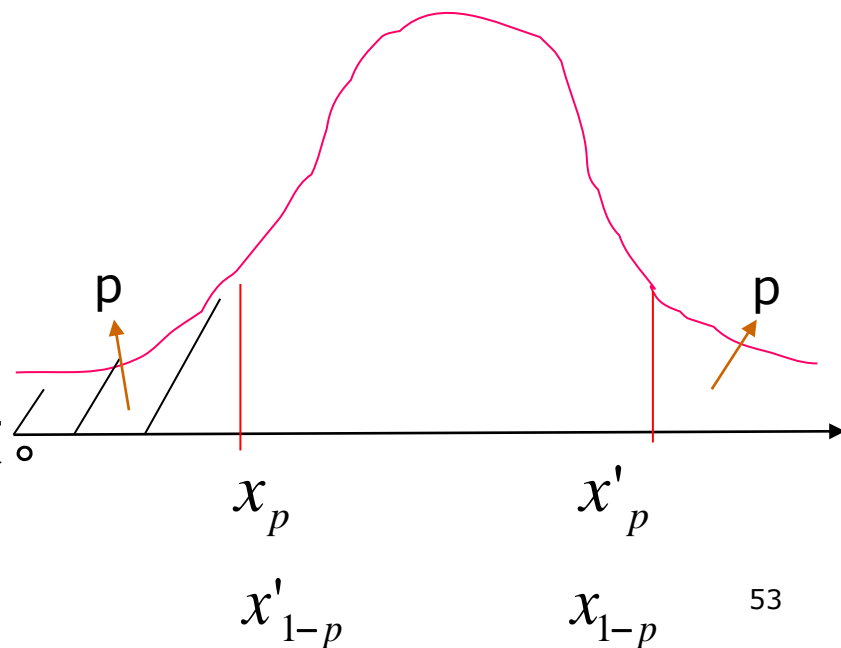
的数 x_p 为此分布的(下侧) p 分位数。

同理，称满足条件

$$1 - F(x'_p) = \int_{x'_p}^{+\infty} f(x)dx = p$$

数 x'_p 为此分布的上侧 p 分位数。

关系： $x'_{1-p} = x_p$, $x_{1-p} = x'_p$.



3. χ^2 分布的分位数

设RV $\chi^2 \sim \chi^2(n)$, 给定 α ($0 < \alpha < 1$),

若数 $\chi_{1-\alpha}^2(n)$ 满足:

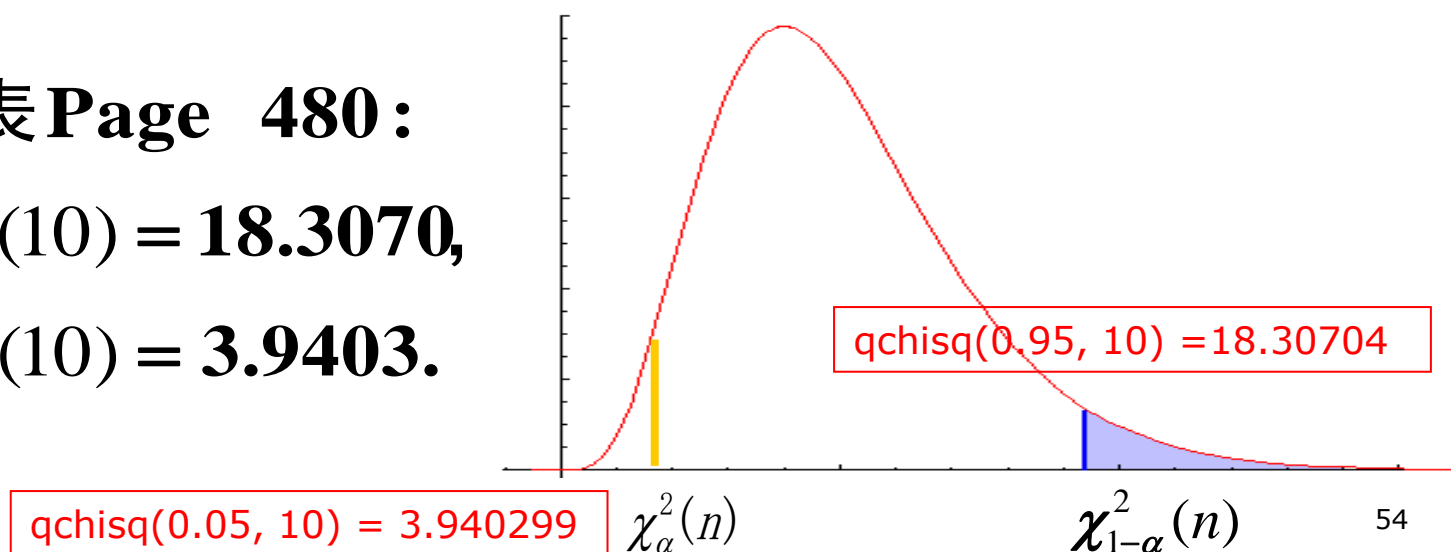
$$P(\chi^2 \leq \chi_{1-\alpha}^2(n)) = 1 - \alpha,$$

则数 $\chi_{1-\alpha}^2(n)$ 称为 χ^2 分布的(下侧) $1 - \alpha$ 分位数。

例: 查表 **Page 480**:

$$\chi_{0.95}^2(10) = 18.3070,$$

$$\chi_{0.05}^2(10) = 3.9403.$$



二、t 分布

问题：若 $X_i \sim N(\mu, \sigma^2)$, 相互独立, 则 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$.

实际问题中, σ 常常未知, 可用样本标准差近似,

问: $\frac{\bar{X} - \mu}{s / \sqrt{n}}$ 的分布是什么? (特别是当样本量 n 较小时)

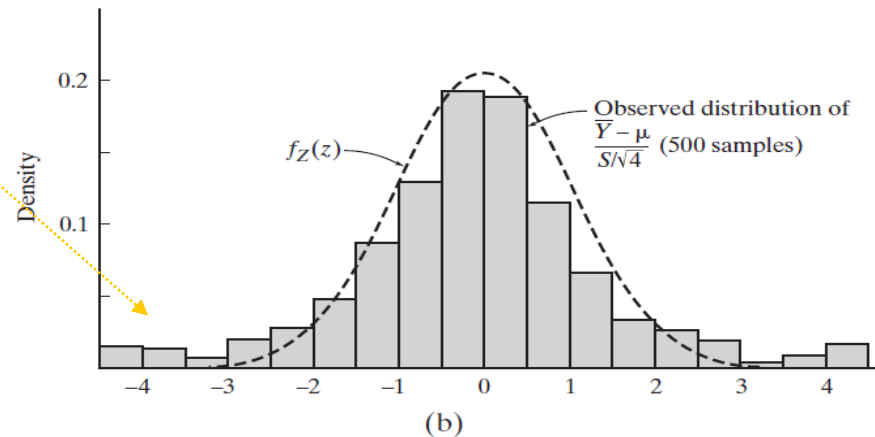
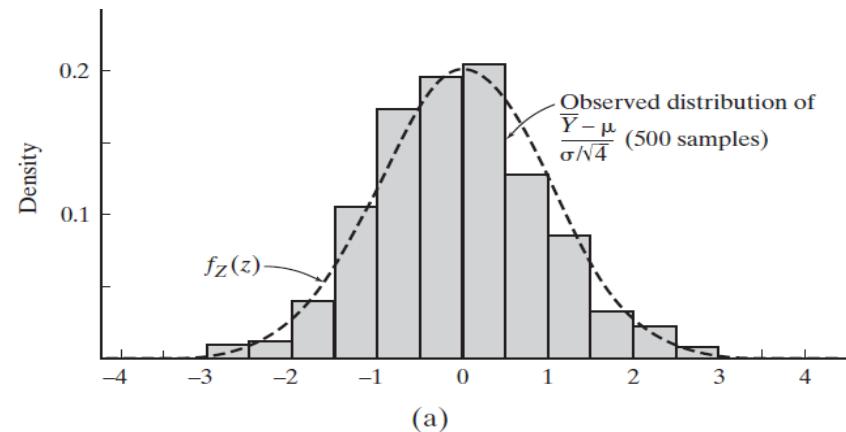
- t 分布是小样本统计分布的第一例, 由Gosset首先发现。之前, 对未知参数的估计都依赖于中心极限定理, 需要大样本才能使用。
- Gosset的方法被其啤酒厂老板认为是商业机密, Gosset只得以笔名Student发表。t 分布亦称为Student分布。

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \text{ and } \frac{\bar{X} - \mu}{s / \sqrt{n}} : \text{ any difference?}$$

Are there probabilistic differences between the two?

Both seemed to have the same general bell-shaped configuration, but the latter has **thicker tails**.

T distribution is one of the major statistical breakthroughs of 20th century.



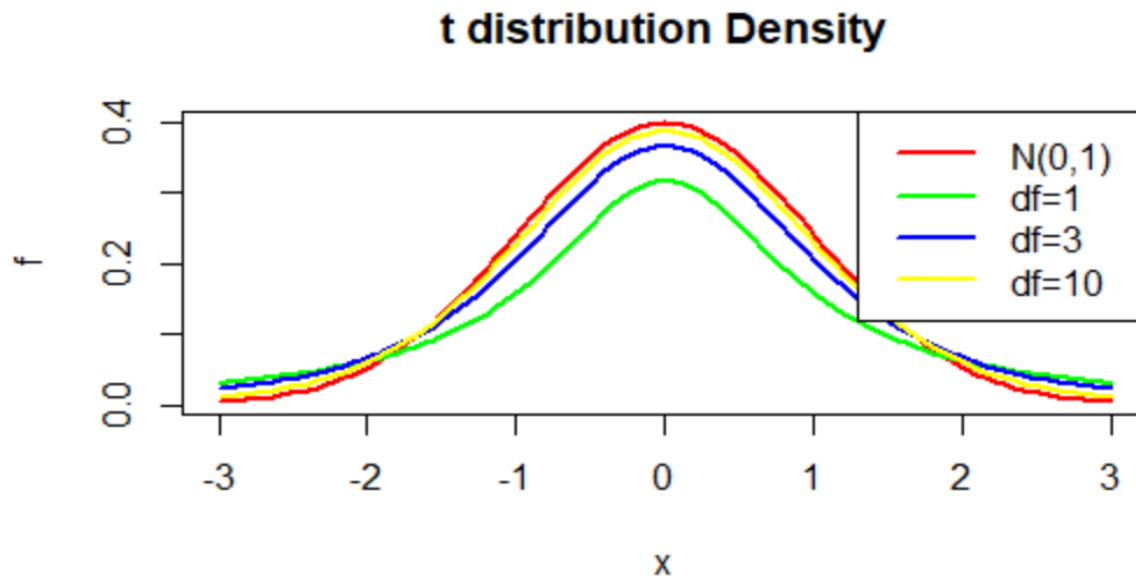
二、t 分布

1. 定义：

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, X 与 Y 独立, 则称

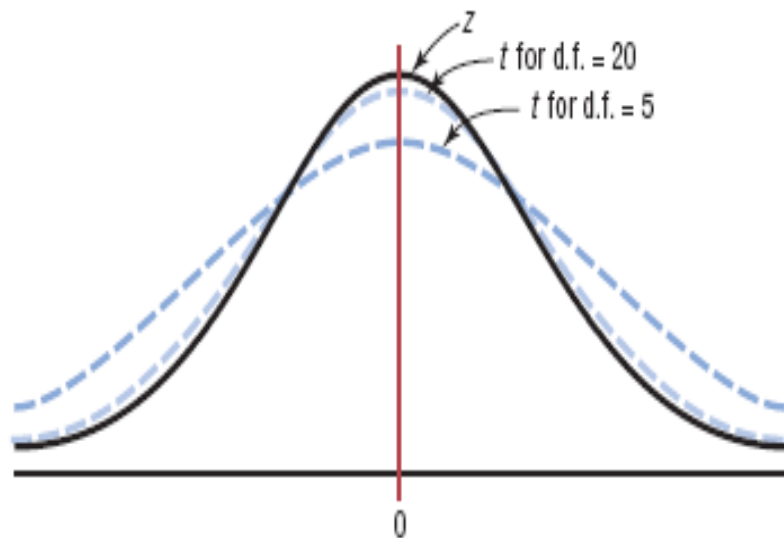
$$t := \frac{X}{\sqrt{Y/n}}$$

的分布为自由度为 n 的 t 分布, 记为 $t \sim t(n)$.



2. t 分布的性质

- (1) 类似于 $N(0, 1)$, 密度关于 $t=0$ 对称;
- (2) 当 n 趋于无穷时, t 分布的密度趋于 $N(0, 1)$ 密度;
- (3) t 分布的“尾”更大, 峰更低;
- (4) 自由度 $n=1$ 为1时, t 分布是标准的Cauchy分布, 期望不存在;
 $n>1$ 时, 期望存在且为 0;
 $n>2$ 时, 方差存在且为 $n/(n-2)$;
 $n\geq 30$ 时, t 分布可用 $N(0, 1)$ 近似。



Why t-distribution is close to normal when n is large?

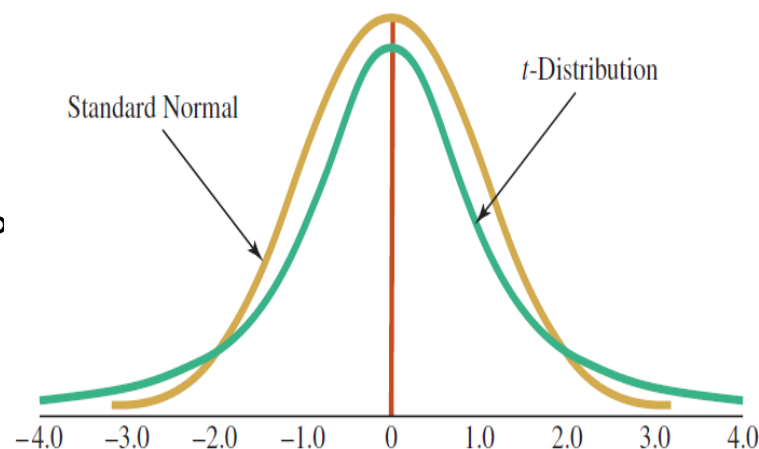
设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, X 与 Y 独立

$$t := \frac{X}{\sqrt{Y/n}} \sim t(n)$$

$$\frac{Y}{n} = \frac{1}{n} (X_1^2 + \dots + X_n^2) \xrightarrow[\text{LLN}]{P} E(X_1^2) = 1.$$

这里, $X_i \sim N(0,1)$,

所以, t 的分布与 X 的分布近似。



3. t 分布的分位数

设RV $t \sim t(n)$, 给定 α ($0 < \alpha < 1$), 若数 $t_{1-\alpha}(n)$ 满足:

$$P(t \leq \underline{t_{1-\alpha}(n)}) = 1 - \alpha,$$

则数 $t_{1-\alpha}(n)$ 称为 t 分布的(下侧) $1-\alpha$ 分位数。

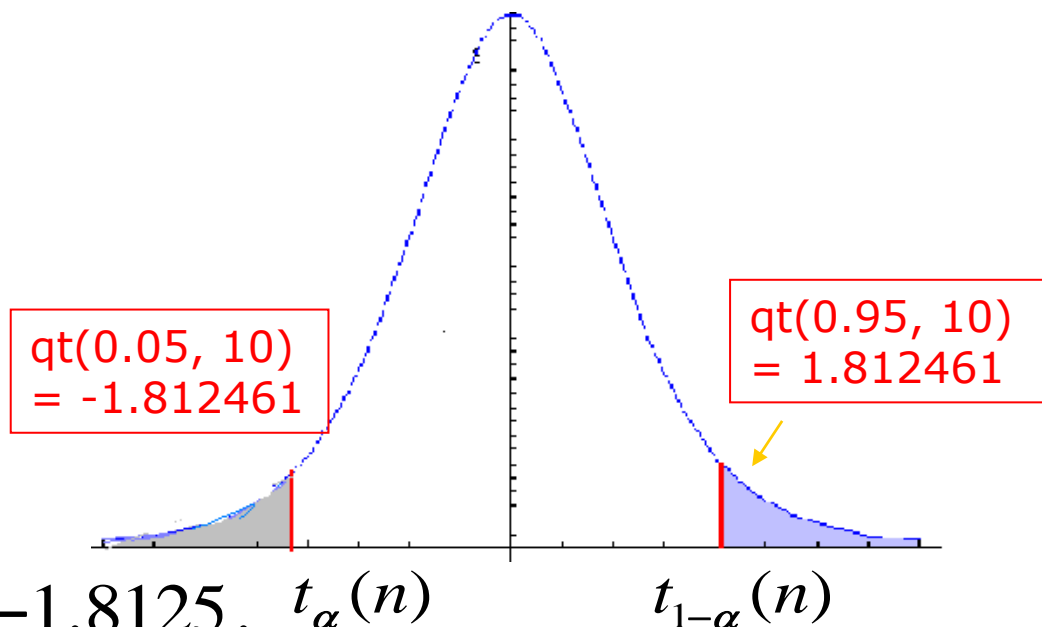
例: 查表 Page 483:

$$t_{0.95}(10) = 1.8125,$$

另, 由对称性

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

$$t_{0.05}(10) = -t_{0.95}(10) = -1.8125.$$

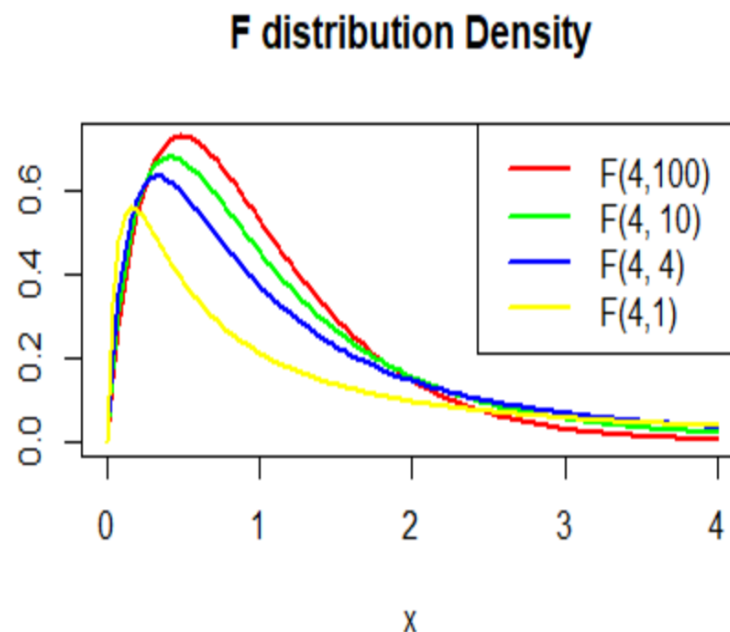


三、F分布

定义：设 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$,
且 X 与 Y 独立，则称

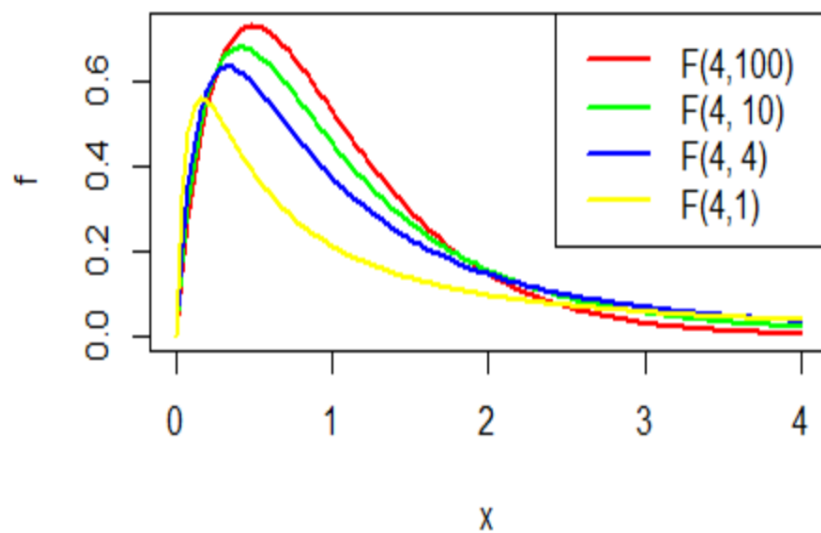
$$F := \frac{X / m}{Y / n}$$

的分布为自由度为 m 和 n 的 F 分布，
记为 $F \sim F(m, n)$,
 m 为第一自由度， n 为第二自由度
 F 分布是取非负值的偏态分布。

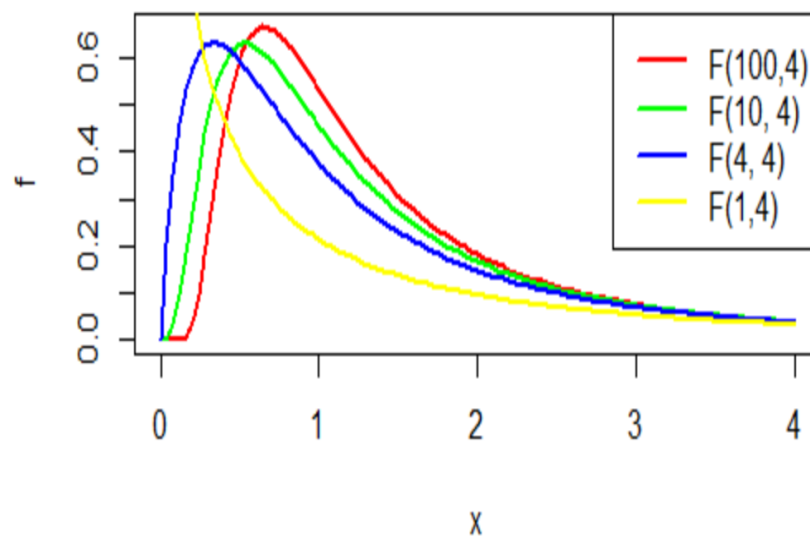


三、F分布

F distribution Density



F distribution Density



三、F分布

密度推导(思想):

(1) 引入 $Z := X / Y$, 利用 **RV** 商的分布公式,
求 Z 的密度;

(2) $F = \frac{n}{m} Z$, 即可知 F 的密度。

例： 设 X_1, X_2, \dots, X_{15} 是总体 $N(0, \sigma^2)$ 的一个样本，

求 $\frac{X_1^2 + X_2^2 + \dots + X_{10}^2}{2(X_{11}^2 + X_{12}^2 + \dots + X_{15}^2)}$ 的分布。

解： 由于 $X_i / \sigma \sim N(0,1)$, 且相互独立，

$$\Rightarrow \frac{1}{\sigma^2} (X_1^2 + X_2^2 + \dots + X_{10}^2) \sim \chi^2(10) ,$$

$$\frac{1}{\sigma^2} (X_{11}^2 + X_{12}^2 + \dots + X_{15}^2) \sim \chi^2(5) ,$$

且两者相互独立， 故

$$Y := \frac{\frac{1}{\sigma^2} (X_1^2 + X_2^2 + \dots + X_{10}^2) / 10}{\frac{1}{\sigma^2} (X_{11}^2 + X_{12}^2 + \dots + X_{15}^2) / 5} \sim F(10, 5).$$

2. F分布的性质

(1) 若 $F \sim F(m, n)$, 则 $\frac{1}{F} \sim F(n, m)$

(2) 若 $t \sim t(n)$, 则 $t^2 \sim F(1, n)$

证:(2) 因 $t = \frac{X}{\sqrt{Y/n}}$, $X \sim N(0,1)$, $Y \sim \chi^2(n)$, X 与 Y 独立,

$$\Rightarrow t^2 = \frac{X^2}{Y/n}, X^2 \sim \chi^2(1),$$

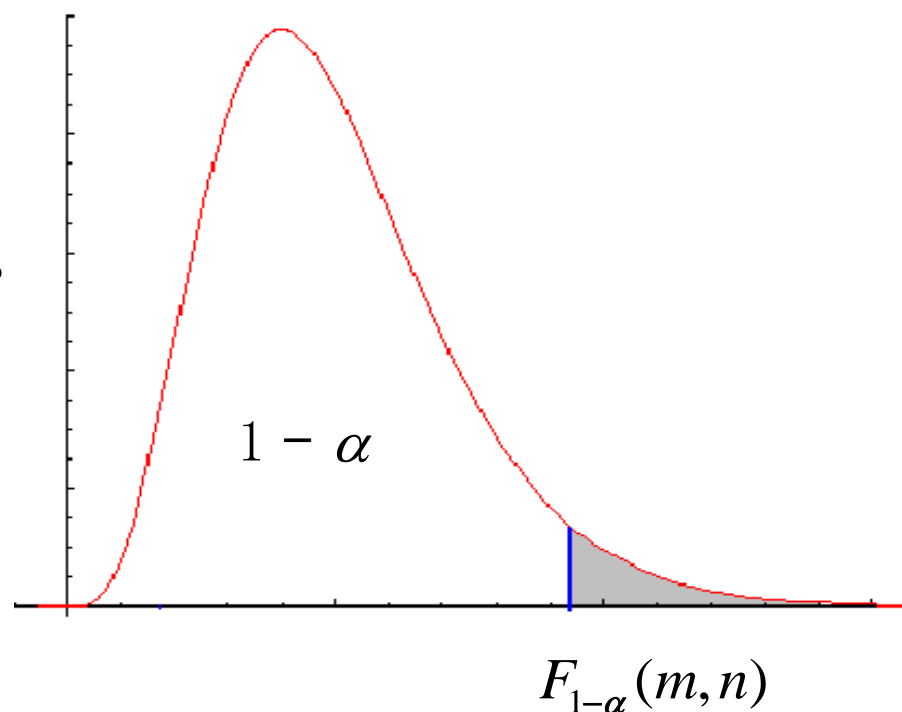
由 F 分布定义即知。

3. F分布的分位数

设RV $F \sim F(m, n)$, 给定 α ($0 < \alpha < 1$),
若数 $F_{1-\alpha}(m, n)$ 满足:

$$P(F \leq \underline{F_{1-\alpha}(m, n)}) = 1 - \alpha,$$

则数 $F_{1-\alpha}(m, n)$ 称为 F 分布的
(下侧) $1 - \alpha$ 分位数。



3. F分布的分位数

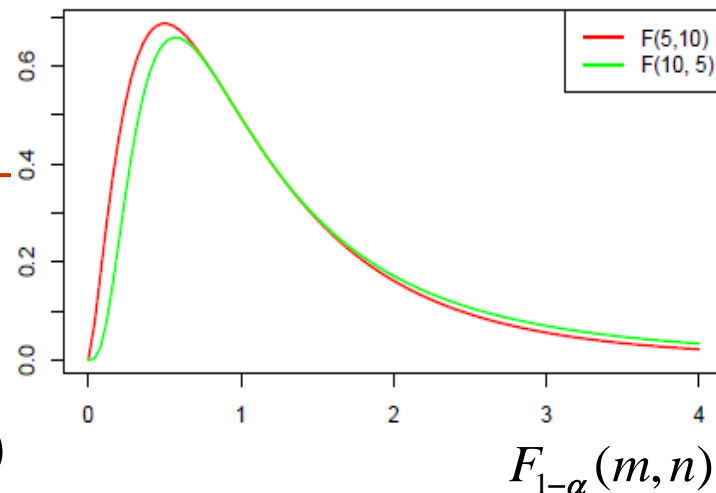
性质: $F_{\alpha}(n, m) = \frac{1}{F_{1-\alpha}(m, n)}$. (*)

证: 因 $F \sim F(m, n) \Rightarrow 1/F \sim F(n, m)$

$$\alpha = P(1/F \leq F_{\alpha}(n, m)) = P(F \geq 1/F_{\alpha}(n, m))$$

对立事件: $P(F \leq 1/F_{\alpha}(n, m)) = 1 - \alpha$

$$\Rightarrow F_{1-\alpha}(m, n) = 1/F_{\alpha}(n, m), \text{即}(*).$$



比较分位数定义:

$$P(F \leq F_{1-\alpha}(m, n)) = 1 - \alpha$$

例: 查表 **Page 486**: $F_{0.95}(10, 5) = 4.74$,

$$F_{0.05}(10, 5) = 1/F_{0.95}(5, 10) = 1/3.33 = 0.3.$$

$$\text{qf}(0.95, 10, 5) \\ = 4.735063$$

$$\text{qf}(0.05, 10, 5) \\ = 0.3006764$$

四、样本均值与样本方差的分布（正态总体）

回顾：

定理：设 X_1, \dots, X_n 是来自总体 X 的样本， \bar{X} 为样本均值。

(1) 若总体为 $N(\mu, \sigma^2)$ ，则 \bar{X} 的精确分布为 $N(\mu, \sigma^2 / n)$ ；

(2) 对任意分布，若 $E(X) = \mu, D(X) = \sigma^2$ ，

则(n 较大时) \bar{X} 的近似分布为 $N(\mu, \sigma^2 / n)$ ；

定理：设总体 X 的均值、方差存在，即 $E(X) = \mu, D(X) = \sigma^2$ ，

X_1, \dots, X_n 是来自该总体的样本， \bar{X}, S^2 分别为样本均值和样本方差。则

$$(1) E(\bar{X}) = \mu, \quad (2) D(\bar{X}) = \sigma^2 / n, \quad (3) E(S^2) = \sigma^2.$$

设 X_1, \dots, X_n 是来自该正态总体 $N(\mu, \sigma^2)$ 的样本,

样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. 以下说法正确的是

☒ A 随机变量 $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$ 服从 $\chi^2(n)$.

☐ B 随机变量 $\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$ 服从 $\chi^2(n)$.

☒ C 随机变量 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ 服从 $N(0, 1)$

定理：设 X_1, \dots, X_n 是来自该正态总体 $N(\mu, \sigma^2)$ 的样本，
其样本均值和样本方差分别为：

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

则


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

$$(2) \bar{X} \text{ 与 } S^2 \text{ 相互独立};$$

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

注记：

$$(1) \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1),$$

$$\text{对比：} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n);$$


$$(2) \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1),$$

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

证明:(1) Step 1

令 $Z_i = \frac{X_i - \mu}{\sigma}$, 则 Z_1, \dots, Z_n 独立同分布, 均为 $N(0,1)$.

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{\bar{X} - \mu}{\sigma},$$

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^n \left[\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma} \right]^2 \\ &= \sum_{i=1}^n [Z_i - \bar{Z}]^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 \quad (*) \end{aligned}$$

Step 2: 作正交变换 $(Y_1, \dots, Y_n)^T = A (Z_1, \dots, Z_n)^T$,

其中A为n阶正交阵($AA^T = I_n$), 且第一行元素均为 $\frac{1}{\sqrt{n}}$.

由多维正态知识, $(Y_1, \dots, Y_n)^T \sim N(0, AI_n A^T)$, 即 $N(0, I_n)$
 $\Rightarrow Y_1, \dots, Y_n$ 独立, 均服从 $N(0, 1)$.

若 $X \sim N(\mu, C)$, 定义 $Y_{k \times 1} = A_{k \times n} X_{n \times 1}$,
则 $Y \sim N(A\mu, ACA^T)$

注意到: $Y_1 = \sum a_{1j} Z_j = \sum \frac{1}{\sqrt{n}} Z_j = \sqrt{n} \bar{Z}$.

$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$ (正交变换下, 长度不变)

由 (*), $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2$
 $= \underline{Y_2^2 + \dots + Y_n^2} \sim \chi^2(n-1).$

证 (续)

$$(2) \quad \text{由 } \bar{Z} = \frac{\bar{X} - \mu}{\sigma}$$

$$\Rightarrow \bar{X} = \sigma \bar{Z} + \mu = \frac{\sigma}{\sqrt{n}} Y_1 + \mu \quad (\text{只依赖于 } Y_1)$$

$$\text{又 } \frac{(n-1)S^2}{\sigma^2} = Y_2^2 + \cdots + Y_n^2$$

$$S^2 = \frac{\sigma^2}{(n-1)} [Y_2^2 + \cdots + Y_n^2] \quad (\text{只依赖于 } Y_2, \dots, Y_n)$$

\bar{X} 与 S^2 相互独立。

证 (续)

$$(3) \text{ 因 } U := \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1), V := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

又两者独立, 由 t 分布的定义有

$$\frac{U}{\sqrt{V/(n-1)}} \sim t(n-1),$$

$$\text{而 } \frac{U}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \bigg/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{\bar{X} - \mu}{S / \sqrt{n}}.$$

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

五、两正态总体的均值差与方差比

问题：

设 X_1, \dots, X_m 是来自总体 $N(\mu_1, \sigma_1^2)$ 的样本，

样本均值： $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$ ，样本方差： $S_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$ 。

设 Y_1, \dots, Y_n 是来自总体 $N(\mu_2, \sigma_2^2)$ 的样本，

样本均值： $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ ，样本方差： $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ 。

两组样本相互独立。

问题：均值差 $\mu_1 - \mu_2$ 和方差比 σ_1^2 / σ_2^2 如何？

？

两总体方差比（正态总体）

结论1:

因 $(m-1) S_x^2 / \sigma_1^2 \sim \chi^2(m-1),$

$$(n-1) S_y^2 / \sigma_2^2 \sim \chi^2(n-1),$$

两总体独立,

P. 288

$$\Rightarrow F := \frac{S_x^2 / \sigma_1^2}{S_y^2 / \sigma_2^2} \sim F(m-1, n-1).$$

两总体均值差（正态总体）

定理：设总体X为 $N(\mu_1, \sigma_1^2)$ ，总体Y为 $N(\mu_2, \sigma_2^2)$ 。独立

(1) 则
$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$
 服从 $N(0, 1)$.

方差已知

标准化

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

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ 服从 } N(0, 1).$$

两总体均值差（正态总体）

定理：（3）设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ， **方差相等，但未知**

则 $G = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_w \sqrt{1/m + 1/n}}$ 服从 $t(m + n - 2)$ 。

其中， $S_w^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$ 。

Information from both groups is combined to estimate common variance using pooled variance (weighted sample variance)

证明： $Z := \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{1/m + 1/n}}$ 服从 $N(0, 1)$ ，

$\frac{(m-1)S_x^2}{\sigma^2}$ 服从 $\chi^2(m-1)$ ， $\frac{(n-1)S_y^2}{\sigma^2}$ 服从 $\chi^2(n-1)$ ，

从而 $W := \frac{(m-1)S_x^2}{\sigma^2} + \frac{(n-1)S_y^2}{\sigma^2}$ 服从 $\chi^2(m+n-2)$ 。（可加性）

由 t 分布的定义， $\frac{Z}{\sqrt{W/(m+n-2)}}$ 服从 $t(m+n-2)$ 。

The End of Chapter 5