# Chapter 6
# 样本及抽样分布

世界银行（WB）
有关各国发展的部分统计信息及有关报告

联合国统计司（UNSD）
国民经济核算、国际贸易...

美联储
银行资产和负债、股票债券

世界贸易组织（WTO）
国际贸易统计、贸易报告等

国际货币基金组织（IMF）
有关IMF贷款、汇率及其他经济

路透社
全球主要股票、债券市场交易情况

# 2020年第七次全国人口普查主要数据公报

**一、总人口**

全国总人口为1443497378人。其中：

普查登记的大陆31个省、自治区、直辖市和现役军人的人口共1411778724人。

香港特别行政区人口为7474200人。

澳门特别行政区人口为683218人。

台湾地区人口为23561236人。

**二、人口增长**

全国人口与2010年第六次全国人口普查的1339724852人相比，增加72053872人，增长5.38%，年平均增长率为0.53%。

**三、家庭户人口**

全国共有家庭户494157423户，集体户28531842户，家庭户人口为1292809300人，集体户人口为118969424人。平均每个家庭户的人口为2.62人，比2010年第六次全国人口普查的3.10人减少0.48人。

**四、性别构成**

全国人口中，男性人口为723339956人，占51.24%；女性人口为688438768人，占48.76%。总人口性别比（以女性为100，男性对女性的比例）为105.07，与2010年第六次全国人口普查基本持平。

# 一些统计数据

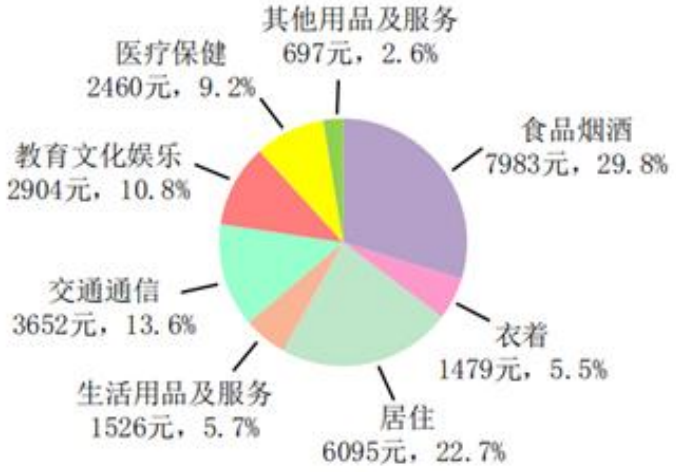
- 国家统计局统计数据：https://www.stats.gov.cn/sj/zxfb/202402/t20240228_1947915.html
- 2023年广东国民经济和社会发展统计公报 https://h5.drcnet.com.cn/docview.aspx?version=edu&docid=7427481&chnid=3650

图1　2019-2023年国内生产总值及其增长速度


图2 2019-2023年三次产业增加值占国内生产总值比重

图18　2019-2023年全国居民人均可支配收入及其增长速度


图19　2023年全国居民人均消费支出及其构成



# 一些统计数据


2023年研究生教育报名474万人，招生130.2万人，在学研究生388.3万人，毕业生101.5万人。普通、职业本专科招生1042.2万人，在校生3775.0万人，毕业生1047.0万人。

# 广东省国民经济和社会发展统计

表1 2023年年末常住人口及构成

| 指标 | 年末常住人口（万人） | 比重（%） |
| --- | --- | --- |
| 常住人口 | 12706 | 100.00 |
| 其中：城镇 | 9583 | 75.42 |
| 乡村 | 3123 | 24.58 |
| 其中：男性 | 6689 | 52.64 |
| 女性 | 6017 | 47.36 |
| 其中：0-15岁 | 2437 | 19.18 |
| 16-59岁 | 8460 | 66.58 |
| 60岁及以上 | 1809 | 14.24 |

年末常住人口12706万人。全年出生人口103万人，出生率8.12%；死亡人口68万人，死亡率5.36%；自然增长人口35万人，自然增长率2.76%。

# 广东省国民经济和社会发展统计

表3 2023年分区域主要指标

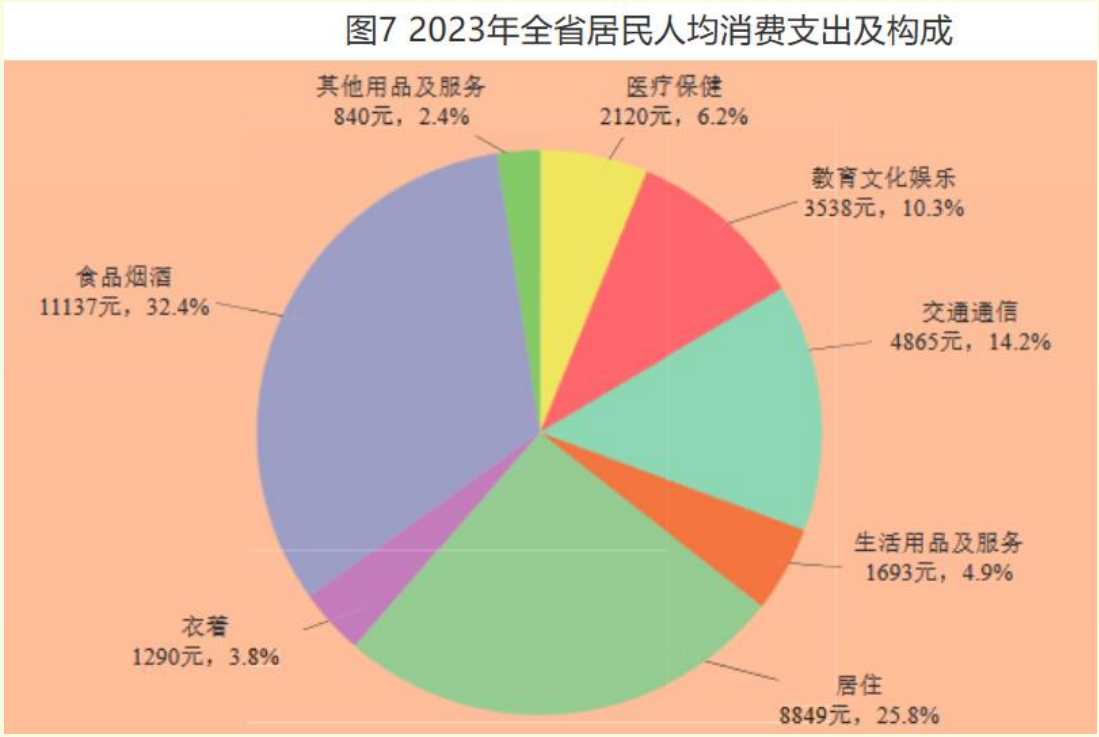
| 区 域 | 地区生产总值（亿元） | 比上年增长（%） | 规模以上工业增加值增长（%） | 固定资产投资增长（%） | 社会消费品零售总额增长（%） | 地方一般公共预算收入增长（%） |
|---|---|---|---|---|---|---|
| 珠三角 | 110214.70 | 4.8 | 4.2 | 2.4 | 5.7 | 3.9 |
| 粤　东 | 8390.78 | 5.0 | 6.0 | -3.2 | 3.6 | 16.4 |
| 粤　西 | 9362.60 | 3.4 | 1.8 | 8.6 | 5.6 | 4.4 |
| 粤　北 | 7705.08 | 4.6 | 5.6 | 4.4 | 3.6 | 4.1 |

分区域看，全年珠三角地区生产总值增长4.8%，粤东、粤西、粤北分别增长5.0%、3.4%、4.6%。粤港澳大湾区建设扎实推进。

# 广东省国民经济和社会发展统计


全年社会消费品零售总额47494.86亿元，比上年增长5.8%。

# 广东省国民经济和社会发展统计




全年广东居民人均消费支出34331元,全年城镇居民人均消费支出39333元,全年农村居民人均消费支出22209元

# 2023年12月房价环比涨跌幅

表1：2023年12月70个大中城市新建商品住宅销售价格指数

| 城市 | 环比 | 同比 | 1-12月平均 | 城市 | 环比 | 同比 | 1-12月平均 |
|---|---|---|---|---|---|---|---|
| | 上月=100 | 上年同月=100 | 上年同期=100 | | 上月=100 | 上年同月=100 | 上年同期=100 |
| 北　京 | 100.0 | 101.7 | 103.5 | 唐　山 | 99.3 | 98.3 | 98.4 |
| 天　津 | 99.6 | 102.3 | 99.9 | 秦皇岛 | 99.8 | 97.1 | 95.9 |
| 石家庄 | 100.3 | 101.0 | 99.2 | 包　头 | 99.6 | 97.1 | 97.1 |
| 太　原 | 100.4 | 101.2 | 98.5 | 丹　东 | 99.7 | 97.8 | 97.3 |
| 呼和浩特 | 99.7 | 100.0 | 98.8 | 锦　州 | 99.2 | 99.1 | 98.1 |
| 沈　阳 | 99.8 | 98.7 | 97.0 | 吉　林 | 99.3 | 99.4 | 97.6 |
| 大　连 | 99.4 | 95.9 | 95.6 | 牡丹江 | 99.7 | 96.5 | 96.6 |
| 长　春 | 100.1 | 97.9 | 95.9 | 无　锡 | 99.6 | 97.6 | 98.7 |
| 哈尔滨 | 99.3 | 98.4 | 96.8 | 徐　州 | 99.0 | 97.9 | 99.8 |
| 上　海 | 100.2 | 104.5 | 104.4 | 扬　州 | 99.4 | 97.3 | 99.5 |
| 南　京 | 98.8 | 96.9 | 99.8 | 温　州 | 99.3 | 97.0 | 95.4 |
| 杭　州 | 99.8 | 102.2 | 104.2 | 金　华 | 99.4 | 96.0 | 96.8 |
| 宁　波 | 99.7 | 101.2 | 102.8 | 蚌　埠 | 99.5 | 99.6 | 99.2 |
| 合　肥 | 99.8 | 100.7 | 102.6 | 安　庆 | 99.2 | 98.4 | 98.7 |
| 福　州 | 99.4 | 98.2 | 98.6 | 泉　州 | 99.8 | 98.2 | 97.2 |
| 厦　门 | 98.9 | 96.7 | 97.4 | 九　江 | 99.6 | 98.2 | 99.4 |
| 南　昌 | 99.3 | 98.9 | 101.2 | 赣　州 | 99.9 | 96.8 | 98.0 |
| 济　南 | 99.6 | 101.8 | 102.8 | 烟　台 | 99.4 | 99.0 | 99.7 |
| 青　岛 | 99.2 | 99.8 | 101.6 | 济　宁 | 99.2 | 97.7 | 97.1 |
| 郑　州 | 99.5 | 98.4 | 98.9 | 洛　阳 | 99.7 | 98.2 | 97.5 |
| 武　汉 | 98.9 | 100.3 | 99.0 | 平顶山 | 99.6 | 97.6 | 98.1 |
| 长　沙 | 99.9 | 102.4 | 103.4 | 宜　昌 | 99.9 | 97.5 | 96.4 |
| 广　州 | 99.0 | 97.0 | 98.8 | 襄　阳 | 99.4 | 98.6 | 98.3 |
| 深　圳 | 99.1 | 96.4 | 97.7 | 岳　阳 | 99.5 | 97.8 | 95.8 |
| 南　宁 | 100.3 | 98.7 | 97.9 | 常　德 | 99.4 | 97.8 | 97.0 |
| 海　口 | 99.4 | 102.1 | 102.0 | 韶　关 | 99.4 | 97.9 | 98.7 |
| 重　庆 | 99.4 | 102.2 | 101.1 | 湛　江 | 99.0 | 100.2 | 98.5 |
| 成　都 | 100.1 | 104.9 | 107.0 | 惠　州 | 99.6 | 95.3 | 96.8 |
| 贵　阳 | 99.9 | 99.2 | 98.4 | 桂　林 | 99.6 | 97.9 | 97.2 |
| 昆　明 | 99.5 | 99.4 | 98.7 | 北　海 | 99.4 | 101.4 | 97.9 |
| 西　安 | 100.5 | 104.5 | 102.3 | 三　亚 | 99.7 | 103.1 | 101.8 |
| 兰　州 | 99.4 | 99.3 | 98.1 | 泸　州 | 99.4 | 97.8 | 98.0 |
| 西　宁 | 99.0 | 97.8 | 99.0 | 南　充 | 99.8 | 99.0 | 100.0 |
| 银　川 | 99.1 | 100.5 | 101.6 | 遵　义 | 99.6 | 100.0 | 100.5 |
| 乌鲁木齐 | 99.5 | 100.2 | 100.7 | 大　理 | 99.9 | 97.6 | 97.1 |

# 数理统计的基本概念

**数理统计的分类**

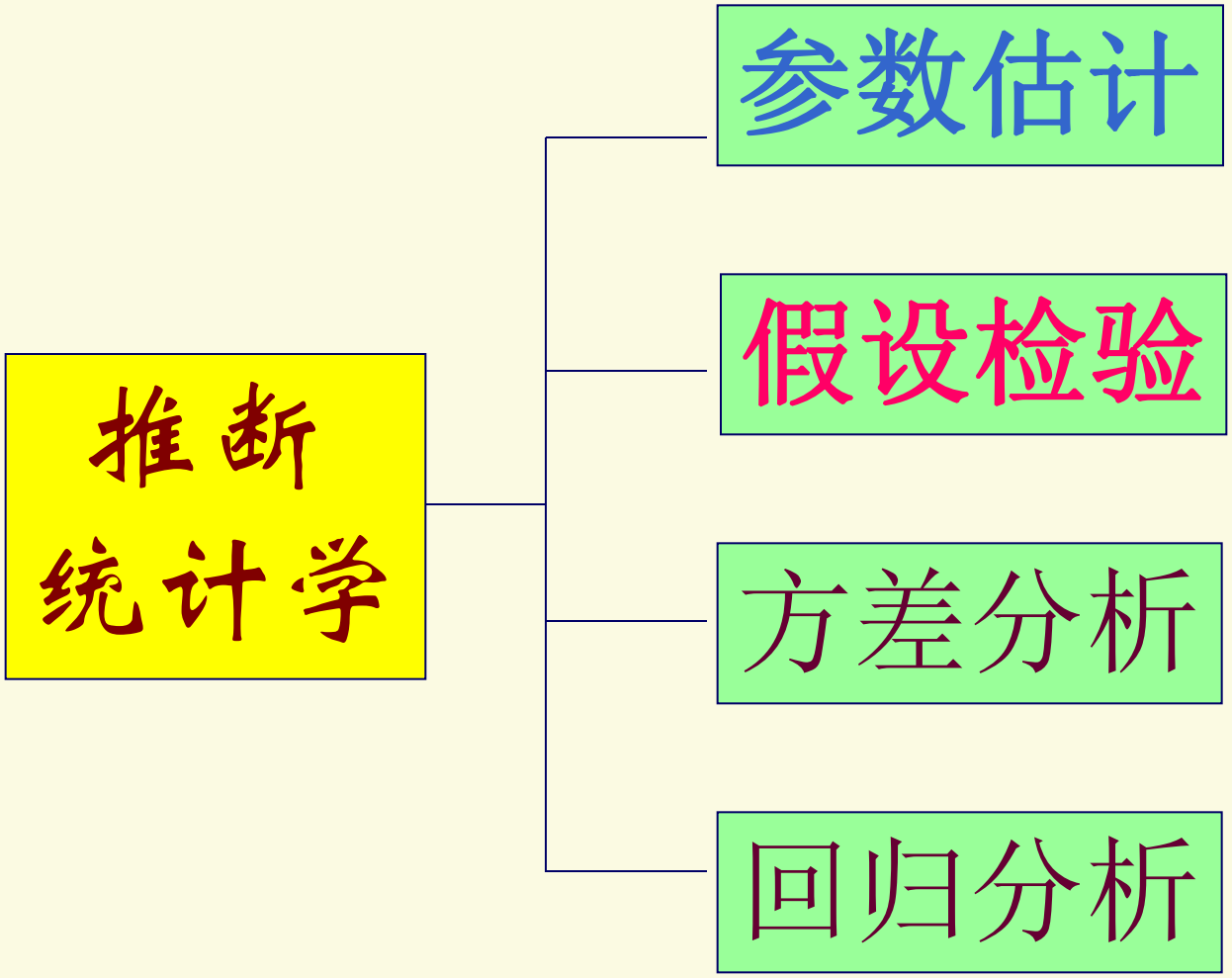
**描述统计学**——

对随机现象进行观测、试验，以取得有代表性的观测值

**推断统计学**——

对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性

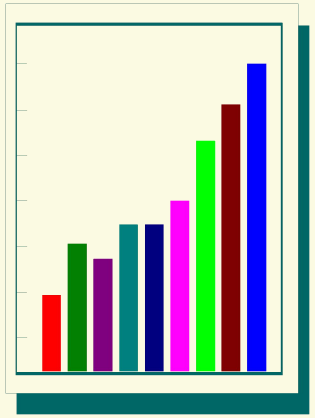
# 推断统计学



# 数理统计

数理统计学是一门应用性很强的学科. 它是研究怎样以有效的方式收集、 整理和分析带有随机性的数据，以便对所考察的问题作出推断和预测.

由于大量随机现象必然呈现它规律性，只要对随机现象进行足够多次观察，被研究的规律性一定能清楚地呈现出来.

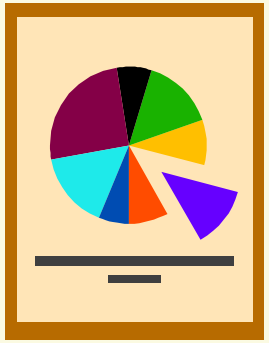


客观上， 只允许我们对随机现象进行次数不多的观察试验 ， 我们只能获得局部观察资料.

# 数理统计（Cont.）

在数理统计中，不是对所研究的对象全体（ 称为总体)进行观察，而是抽取其中的部分(称为样本)进行观察获得数据（抽样），并通过这些数据对总体进行推断.

数理统计方法具有“部分推断整体”的特征 .



# 基本概念

一个统计问题总有它明确的研究对象.

- 总体 研究对象的全体
- 个体 总体中每个对象称为个体.


研究某批灯泡的质量

该批灯泡寿命的全体就是总体


考察国产 轿车的油耗

所有国产轿车每公里耗油量的全体就是总体

我们关心的是总体中的个体的某项指标(如人的身高、灯泡的寿命,汽车的耗油量…）.

由于每个个体的出现是随机的，所以相应的数量指标的出现也带有随机性。从而可以把这种数量指标看作一个随机变量X，因此随机变量X的分布就是该数量指标在总体中的分布。

总体就可以用一个随机变量及其分布来描述.

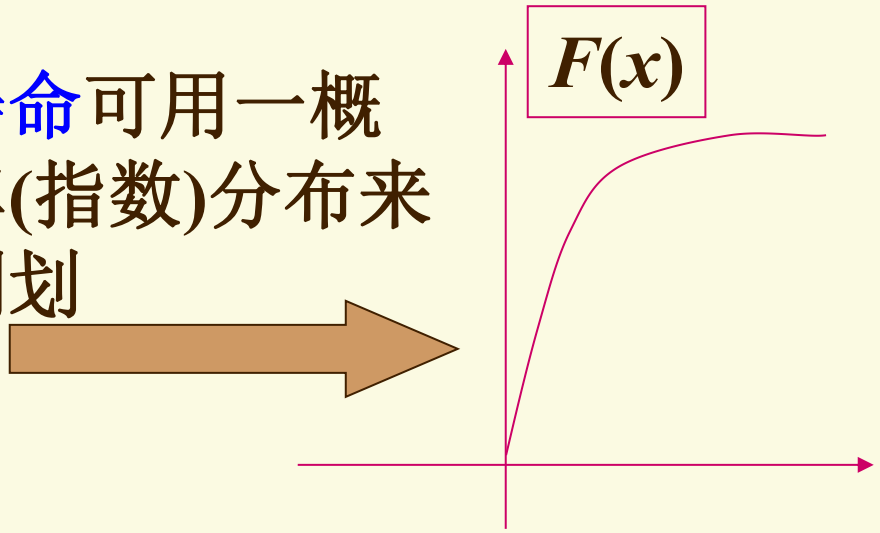
因此在理论上可以把总体与概率分布等同起来.

# 基本概念(Cont.)

- 总体可以用一个随机变量 $X$ 或其分布来描述

如:研究某批灯泡的寿命时，我们关心的就是寿命，那么，寿命这个总体就可以用随机变量$X$表示，或用其分布函数$F(x)$表示.


寿命可用一概率(指数)分布来刻划



# 基本概念(Cont.)

类似地，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.

统计中，总体这个概念的要旨是：总体就是一个概率分布.

# 基本概念(Cont.)

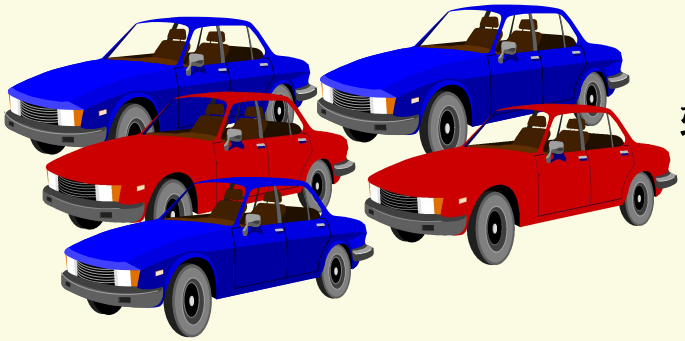
- 个体

总体中每个对象称为个体.

- 样本

为推断总体分布及各种特征,按一定规则从总体中抽取若干个体进行观察试验以获得有关总体的信息。所抽取的部分个体称为样本. 样本中所包含的个体数目称为样本容量.

# 基本概念(Cont.)



从国产轿车中抽5辆进行耗油量试验

样本容量为5

抽到哪 5 辆是随机的!

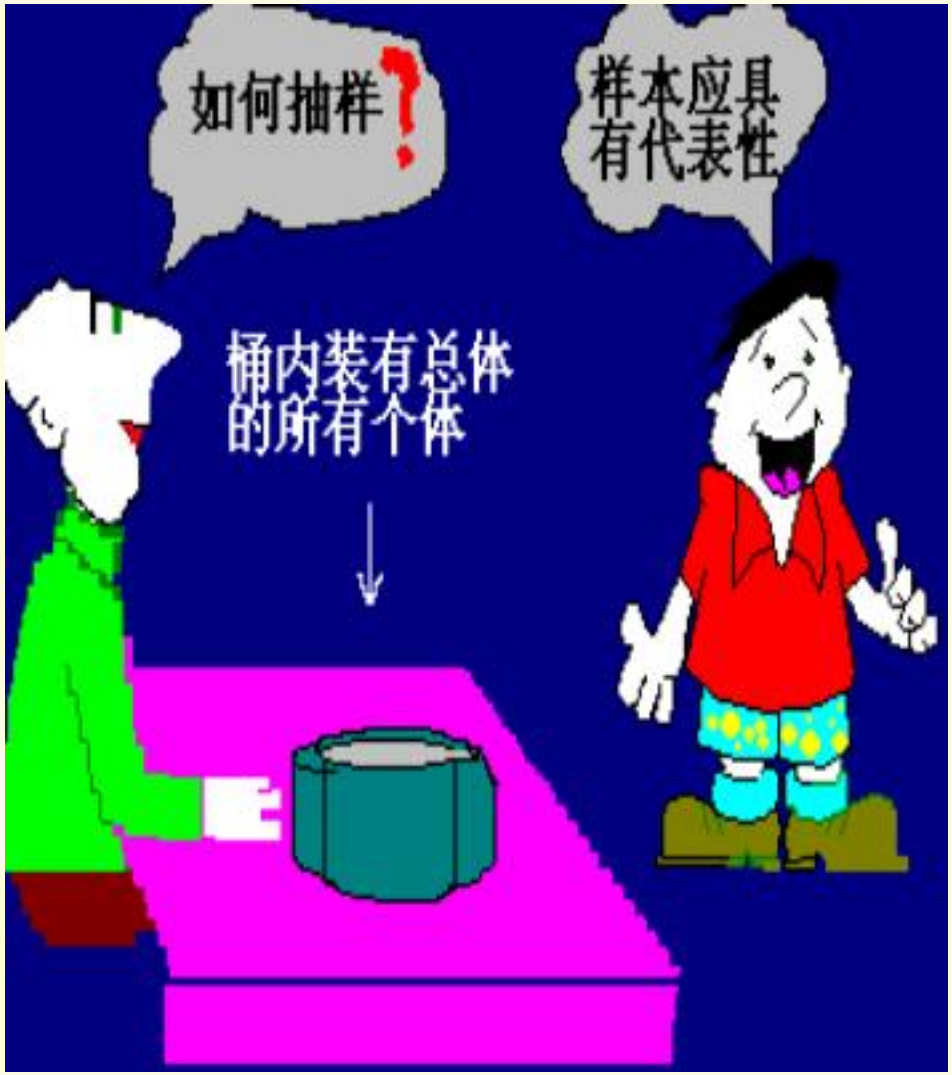
样本是随机变量

容量为 $n$ 的样本可以看作$n$ 维随机变量$(X_1, X_2, \cdots, X_n)$.

一旦取定一组样本，得到的是 $n$ 个具体的数 $x_1, x_2, \cdots, x_n$ , 称为样本$(X1, X2, \ldots, Xn)$的一组观测值，简称样本值 .

# 简单随机样本

抽取样本的目的是为了利用样本对总体进行统计推断,这就要求样本能很好的反映总体的特性且便于处理. 因此：



# 简单随机样本(Cont.)

抽取的样本$X_1, X_2, \cdots, X_n$ 满足下面两点:

1.独立性： $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量;

2.代表性： $X_i$ ($i$ =1,2,$\cdots$,$n$) 与所考察的总体 $X$ 具有相同的分布.

简单随机样本是应用中最常见的情形,今后，说到“$X_1, \cdots, X_n$ 是来自某总体的样本”时，若不特别说明，就指简单随机样本.

# 总体与样本

总体：随机变量$X$，或分布函数$F(x)$

样本：$n$ 维随机变量$(X1, X2, \ldots, Xn)$.

$x_1, x_2, \cdots, x_n$ ，为样本值 .

- 若总体 $X$ 的分布函数为$F(x)$, 则其样本的联合分布函数为

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

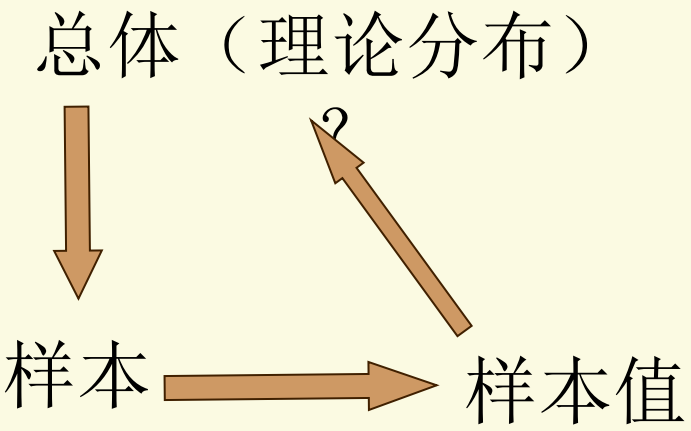
- 若总体$X$的概率密度为$p(x)$,则其样本的联合概率密度为

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

# 总体与样本(Cont.)

事实上我们抽样后得到的资料都是具体的、确定的值.如我们从某班大学生中抽取10人测量身高,得到10个数，它们是样本取到的值而不是样本. 我们只能观察到随机变量取的值而见不到随机变量.

# 总体与样本(Cont.)




统计是从手中已有的资料--样本值，去推断总体的情况---总体分布$F(x)$的性质.

样本是联系二者的桥梁

总体分布决定了样本取值的概率规律，也就是样本取到样本值的规律，因而可以由样本值去推断总体.

# 直方图：连续性随机变量

**例1** 下面给出了84个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm），

数据的“频率直方图”.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 141 | 148 | 132 | 138 | 154 | 142 | 150 | 146 | 155 | 158 | 150 | 140 |
| 147 | 148 | 144 | 150 | 149 | 145 | 149 | 158 | 143 | 141 | 144 | 144 |
| 126 | 140 | 144 | 142 | 141 | 140 | 145 | 135 | 147 | 146 | 141 | 136 |
| 140 | 146 | 142 | 137 | 148 | 154 | 137 | 139 | 143 | 140 | 131 | 143 |
| 141 | 149 | 148 | 135 | 148 | 152 | 143 | 144 | 141 | 143 | 147 | 146 |
| 150 | 132 | 142 | 142 | 143 | 153 | 149 | 146 | 149 | 138 | 142 | 149 |
| 142 | 137 | 134 | 144 | 146 | 147 | 140 | 142 | 140 | 137 | 152 | 145 |

步骤：

1. 找出最小值126，最大值158，现取区间[124.5,159.5]；

2. 将区间[124.5，159.5]等分为7个小区间，小区间的长度记成 $\Delta$, $\Delta = (159.5 - 124.5)/7 = 5$, $\Delta$称为组距；

3. 小区间的端点称为组限, 数出落在每个小区间的数据的频数 $f_i$, 算出频率 $f_i / n$.

列表如下：

| 组 限 | 频 数 | 频 率 | 累计频率 |
|---|---|---|---|
| 124.5~129.5 | 1 | 0.0119 | 0.0119 |
| 129.5~134.5 | 4 | 0.0476 | 0.0595 |
| 134.5~139.5 | 10 | 0.1191 | 0.1786 |
| 139.5~144.5 | 33 | 0.3929 | 0.5715 |
| 144.5~149.5 | 24 | 0.2857 | 0.8572 |
| 149.5~154.5 | 9 | 0.1071 | 0.9643 |
| 154.5~159.5 | 3 | 0.0357 | 1.0000 |

现在自左向右依次在各个小区间上作以 $\frac{f_i}{n}\Big/\Delta$ 为高的小矩形，这样的图形叫频率直方图.

# 频率直方图

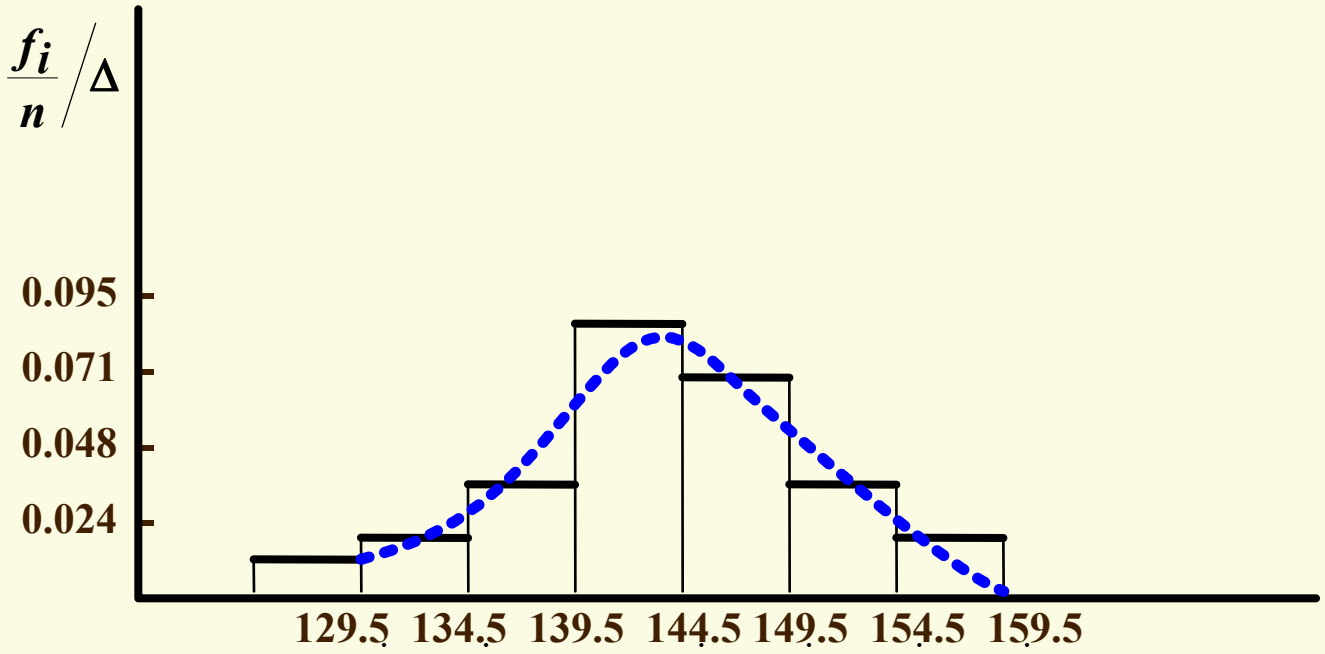

$\frac{f_i}{n}\Big/\Delta$
0.095
0.071
0.048
0.024
129.5
134.5
139.5
144.5
149.5
154.5
159.5

图 6-1

# 箱线图

设有容量为 $n$ 的样本观察值 $x_1, x_2, \cdots, x_n$，样本 $p$ 分位数 $(0 < p < 1)$ 记为 $x_p$，它具有以下的性质：

⑴ 至少有 $np$ 个观察值小于或等于 $x_p$；

⑵ 至少有 $n(1-p)$ 个观察值大于或等于 $x_p$.

样本 $p$ 分位数可按以下法则求 得. 将 $x_1, x_2, \cdots, x_n$ 按从小到大的顺序排列 成 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

$1^{o}$ 若 $np$ 不是整数，则只有一个数据满足定 义中的两点要求，这一数据位于大于 $np$ 的最小整数处，即为位于 $[np]+1$ 处的数.

# 箱线图（Cont.）

$2^{o}$ 若$np$是整数， 就取位于 $[np]$和$[np]+1$处的中位数.

综上，

$$x_p = \begin{cases} x_{([np]+1)}, & \text{当}np\text{不是整数}, \\ \dfrac{1}{2}[x_{(np)} + x_{(np+1)}], & \text{当}np\text{是整数}. \end{cases}$$

# 箱线图（Cont.）

特别，当 $p = 0.5$时，0.5分位数 $x_{0.5}$也记为$Q_2$或$M$称为样本中位数， 即有

$$x_{0.5} = \begin{cases} x_{([\frac{n}{2}]+1)}, & \text{当}np\text{不是整数}, \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{当}np\text{是整数}. \end{cases}$$

0.25分位数 $x_{0.25}$ 称为第一四分位数，又记为$Q_1$;

0.75分位数 $x_{0.75}$ 称为第三四分位数，又记为$Q_3$.

**例2** 设有一组容量为18的样本如下（已经排过序）

122 126 133 140 145 145 149 150 157

162 166 175 177 177 183 188 199 212

求样本分位数：$x_{0.2}$，$x_{0.25}$，$x_{0.5}$.

**解** (1)因为 $np = 18 \times 0.2 = 3.6,$

$x_{0.2}$位于第$[3.6]+1=4$处，即有 $x_{0.2} = x_{(4)} = 140.$

(2)因为 $np = 18 \times 0.25 = 4.5,$

$x_{0.25}$位于第$[4.5]+1=5$处，即有 $x_{0.25} = 145.$

(3)因为 $np = 18 \times 0.5 = 9$, $x_{0.5}$是这组数中间两个数的平均值，即有 $x_{0.5} = \frac{1}{2}(157+162) = 159.5.$

数据集的箱线图是由箱子和直线组成的图形，它是基于以下五个数的图形概括：最小值 Min，第一四分位数 $Q_1$，中位数$M$，第三四分位数 $Q_3$和最大值 Max. 它的作法如下：

(1) 画一水平数轴，在轴上标上 Min，$Q_1$，$M$，$Q_3$，Max．在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位 于 $Q_1$，$Q_3$ 的上方.

在$M$点的上方画一条垂直线 段.线段位于箱子内部.

(2)自箱子左侧引一条水平 线 Min;在同一水平高度自箱子右侧引一条水平线直至最大值.

如图所示.

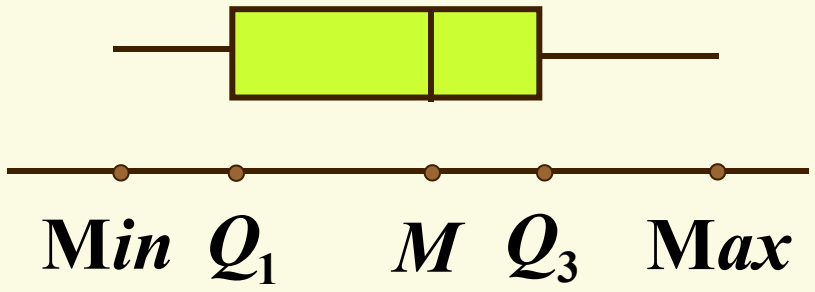



图 6-2

**例3** 以下是8个病人的血压（收缩压，mmHg）数据（已经过排序），试作出箱线图.

102　110　117　118　122　123　132　150

**解**　因为$np = 8 \times 0.25 = 2$,　故

$$Q_1 = \frac{1}{2}(110 + 117) = 113.5.$$

因为$np = 8 \times 0.5 = 4$,　故

$$x_{0.5} = Q_2 = \frac{1}{2}(118 + 122) = 120.$$

因为$np = 8 \times 0.75 = 6$,　故

$$x_{0.75} = Q_3 = \frac{1}{2}(123 + 132) = 127.5.$$

$\mathbf{Min} = 102, \quad \mathbf{Max} = 150,$

作出箱线图如图所示.

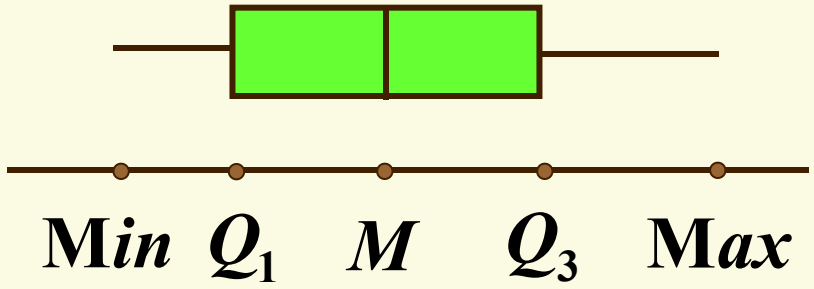



图 6-3

**例4** 下面分别给出了25个男子和25个女子的肺活量（以升计. 数据应经过排序）

女子组 **2.7 2.8 2.9 3.1 3.1 3.1 3.2 3.4 3.4**

**3.4 3.4 3.4 3.5 3.5 3.5 3.6 3.7 3.7**

**3.7 3.8 3.8 4.0 4.1 4.2 4.2**

男子组 **4.1 4.1 4.3 4.3 4.5 4.6 4.7 4.8 4.8**

**5.1 5.3 5.3 5.3 5.4 5.4 5.5 5.6 5.7**

**5.8 5.8 6.0 6.1 6.3 6.7 6.7**

试分别画出这两组数据的箱线图.

**解** 女子组 $\text{Min} = 2.7,\ \text{Max} = 4.2,\ M = 3.5,$

因 $np = 25 \times 0.25 = 6.25,\ Q_1 = 3.2.$

因 $np = 25 \times 0.75 = 18.75,\ Q_3 = 3.7.$

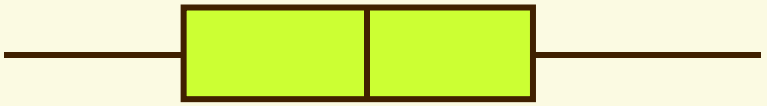
男子组 $\text{Min} = 4.1,\ \text{Max} = 6.7,\ M = 5.3,$

因 $np = 25 \times 0.25 = 6.25,\ Q_1 = 4.7.$

因 $np = 25 \times 0.75 = 18.75,\ Q_3 = 5.8.$

作出箱线图如图所示.

男子



女子

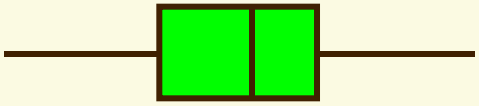


图 6-4

## 疑似异常值

在数据集中，某一个观察值不寻常地大于或小于该数据集中的其他数据,称为疑似异常值.

第一四分位数$Q_1$与第三四分数$Q_3$之间的距离：

$$Q_3 - Q_1 = IQR$$

称为四分位数间距.

若数据小于 $Q_1 - 1.5IQR$ 或大于$Q_3 + 1.5IQR$，则认为它是疑似异常值 .

## 修正箱线图

(1′) 同(1)；

(2′) 计算$IQR = Q_3 - Q_1$，若一个数据小于$Q_1 - 1.5IQR$或大于$Q_3 + 1.5IQR$，则认为它是一个疑似异常值．画出疑似异常值， 并以* 表示;

(3′) 自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值，又自箱子右侧引一水平线直至数据集中除去疑似异常值后的最大值.

**例5** 下面给出了某医院21个病人的住院时间（以天计），试画出修正箱线图（数据已经过排序）.

1 2 3 3 4 4 5 6 6 7 7 9 9

10 12 12 13 15 18 23 55

**解** Min = 1, Max = 55, $M = 7$,

因 $21 \times 0.25 = 5.25$, 得 $Q_1 = 4$,

又 $21 \times 0.75 = 15.75$, 得 $Q_3 = 12$,

$IQR = Q_3 - Q_1 = 8$, $Q_3 + 1.5IQR = 12 + 1.5 \times 8 = 24$,

$Q_1 - 1.5IQR = 4 - 12 = -8.$

观察值55 > 24, 故55 是疑似异常值,且仅此一个疑疑似异常值.

作出修正箱线图如图所示.

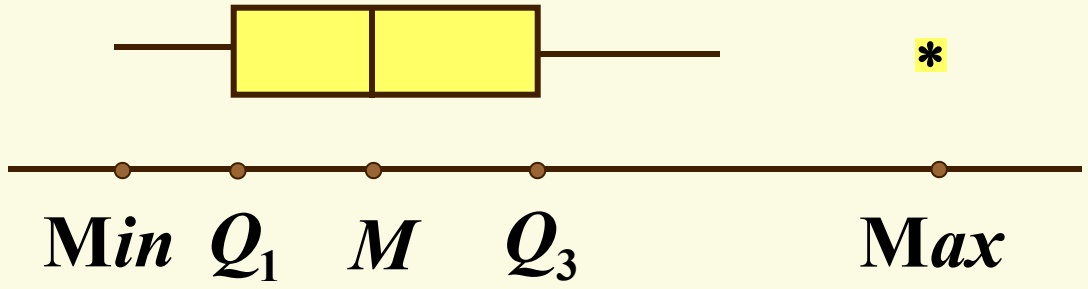



图 6-5

# 三、小结

## 1. 频率直方图作图步骤

(1) 找出最小值和最大值；

(2) 将选定区间分为$k$个小区间；

(3) 算出频率$f_i/n$. 在各个小区间上作以 $\dfrac{f_i}{n}\Big/\Delta$ 为高的小矩形.

## 2. 箱线图作图步骤

(1) 画一水平数轴， 在轴上标上 Min，$Q_1$， $M$，$Q_3$，Max．在数轴上方画一个上、 下侧平行于数轴的矩形箱子，箱子的左右两侧分别位 于 $Q_1$， $Q_3$ 的上方.

在$M$点的上方画一条垂直线 段.线段位于箱子内部.

(2)自箱子左侧引一条水平 线 Min；在同一水平高度自箱子右侧引一条水平线直至最大值.

# 统计量

由样本推断总体特征, 需要对样本值进行“加工”, “**提炼**”. 这就需要构造一些样本的函数, 它把样本中所含的信息集中起来.

**统计量的定义**

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.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值, 则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

# 统计量(Cont.)

**实例** 设$X_1, X_2, X_3$是来自总体$N(\mu, \sigma^2)$的一个样本, 其中$\mu$为已知, $\sigma^2$为未知, 判断下列各式哪些是统计量, 哪些不是?

$T_1 = X_1,$ $T_2 = X_1 + X_2 e^{X_3},$

$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$

$T_4 = \max(X_1, X_2, X_3),$ $T_5 = X_1 + X_2 - 2\mu,$

是

$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2).$

不是

# 常用统计量

它反映了总体均值的信息

1. 样本平均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

2. 样本方差

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

它反映了总体方差的信息

3 样本标准差

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

# 常用统计量

4. 样本k阶原点矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad k=1,2,\ldots$$

它反映了总体$k$阶矩的信息

5 样本k阶中心矩

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

它反映了总体$k$阶中心矩的信息

# 统计量的分布

统计量既然是依赖于样本的，而后者又是随机变量，故统计量也是随机变量，因而就有一定的分布，这个分布叫做统计量的“抽样分布”．

当总体的分布函数已知时，抽样分布是确定的，然而要求出统计量的精确分布，一般来说较难，下面介绍几个常用的统计量的分布。

# 常用统计量的分布

## （一）$\chi^2$分布

定义：设随机变量$X_1, X_2, \cdots X_n$相互独立，$X_i \sim N(0,1)\ (i=1,2,\cdots,n)$

则称 $\chi_n^2 = \sum_{i=1}^{n} X_i^2$

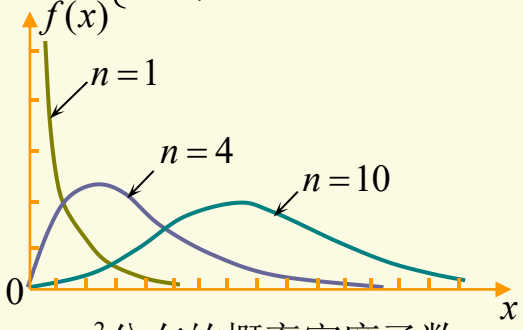
服从自由度为$n$的$\chi^2$分布，记为$\chi_n^2 \sim \chi^2(n)$

自由度指该式右端包含的独立变量的个数

定理：$\chi^2(n)$分布的概率密度为：

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

其中$\Gamma(\alpha)=\int_0^{+\infty} x^{\alpha-1}e^{-x}\mathrm{d}x$




$\chi^2$分布的概率密度函数

# 回顾：连续型随机变量函数的分布

**例** 设 $X$ 具有概率密度 $f_X(x)$， 求 $Y=X^2$ 的概率密度.

**解** 设 $Y$ 和 $X$ 的分布函数分别为 $F_Y(y)$ 和 $F_X(x)$，

注意到 $Y=X^2\geq 0$，故当 $y\leq 0$ 时，$F_Y(y)=0$;

当 $y>0$ 时，

$$F_Y(y)=P(Y\leq y)=P(X^2\leq y) \quad ①$$

求导可得

$$=P(-\sqrt{y}\leq X\leq\sqrt{y})=F_X(\sqrt{y})-F_X(-\sqrt{y}), \quad ②$$

$$f_Y(y)=\frac{dF_Y(y)}{dy}=\begin{cases}\dfrac{1}{2\sqrt{y}}\left[f_X(\sqrt{y})+f_X(-\sqrt{y})\right], & y>0;\\ 0, & y\leq 0.\end{cases}$$

# 连续型随机变量函数的分布(Cont.)

若 $f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad -\infty<x<\infty$

即 $X\sim N(0,1)$

则 $Y=X^2$ 的概率密度为

$$f_Y(y)=\begin{cases}\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}, & y>0;\\ 0, & y\le 0.\end{cases}$$

$\Gamma$分布的定义，$X\sim\Gamma(\partial,\theta)$

$$f(x)=\begin{cases}\frac{1}{\theta^{\partial}\Gamma(\partial)}x^{\partial-1}e^{-x/\theta} & x>0\\ 0, & \text{其他}\end{cases}$$

此时$Y\sim\chi^2(1)$，即为$\Gamma(\frac{1}{2},2)$分布

# 常用统计量的分布

- 自由度(degree of freedom, df)在数学中是指能够自由取值的随机变量的个数.在统计学中，自由度指的是计算某一统计量时，取值不受限制的变量个数。
- 如有3个变量x、y、z，但x+y+z=18，因此其自由度等于2。
- 通常df=n-k。其中n为样本含量，k为被限制的条件数或变量个数，或计算某一统计量时用到其它独立统计量的个数。自由度通常用于抽样分布中。

# 常用统计量的分布（Cont.）

- Γ Functions伽马函数

定义

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx \qquad \text{Where a >0}$$

函数特征

- Γ(n) = (n-1)!  for natural number n.
- Γ(x+1) = x Γ(x)
- Γ( ½ ) = $\pi^{1/2}$

# 卡方分布是由英国统计学家Karl Pearson (1857-1936)于1900年提出来的。

Karl Pearson（1857～1936），生卒于伦敦，公认为**统计学之父**。

K. Pearson 1879年毕业于剑桥大学数学系，曾参与激进的政治活动。出版几本文学作品，并且作了三年的律师实习。1884年进入**伦敦大学学院 (University College, London)**，教授数学与力学，从此待在该校一直到1933年。

K. Pearson 最重要的学术成就，是为现代统计学打下基础。自从达尔文演化论问世后，关于演化的本质争论不断，在这方面他深受 Galton（达尔文表哥，「优生学」一词的发明者）与 Weldon 影响。 Weldon 1893年提出「所谓变异，遗传与天择事实上只是『算术』」的想法。这促使 K. Pearson 在1893-1912年间写出18篇〈在演化论上的数学贡献〉的文章，而这门「算术」，也就是今日的统计。许多熟悉的统计名词如标准差，成分分析，卡方检定都是他提出的。

K. Pearson、Galton 与 Weldon 为了推广统计在生物上的应用，于1901年创立统计的元老期刊《Biometrika》， 由 K. Pearson 主编至死，但是 K. Pearson 的主观强，经常对他本人认为有「争议」的文章， 删改或退稿，并因此与英国本世纪最有才华的统计学家 Fisher 结下梁子。

1906年 Weldon 死后，K. Pearson 不再注意生物问题，而专心致志于将统计发展成一门精确的科学。

## $\chi^2$分布的一些重要性质：

$D(X_i^2)=E(X_i^4)-\left[E(X_i^2)\right]^2=2$

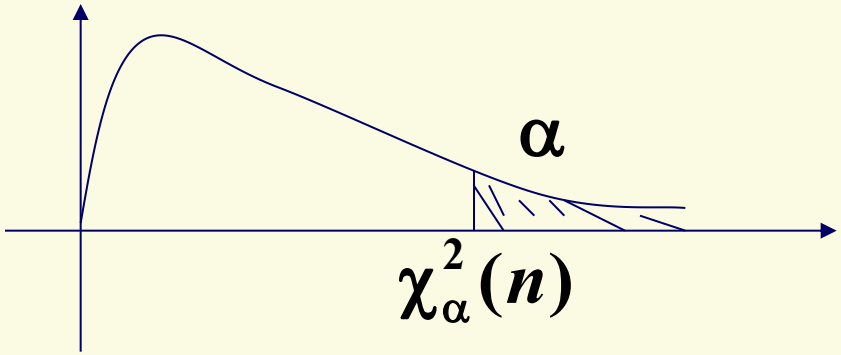
1. 设$\chi^2 \sim \chi^2(n)$,则有$E(\chi^2)=n, D(\chi^2)=2n$

2. 设$Y_1 \sim \chi^2(n_1), Y_2 \sim \chi^2(n_2)$,且$Y_1, Y_2$相互独立，则有$Y_1+Y_2 \sim \chi^2(n_1+n_2)$

性质2称为$\chi^2$分布的可加性，可推广到有限个的情形：

设$Y_i \sim \chi^2(n_i)$,且$Y_1, Y_2, \cdots, Y_m$相互独立，则$\sum_{i=1}^{m} Y_i \sim \chi^2\left(\sum_{i=1}^{m} n_i\right)$

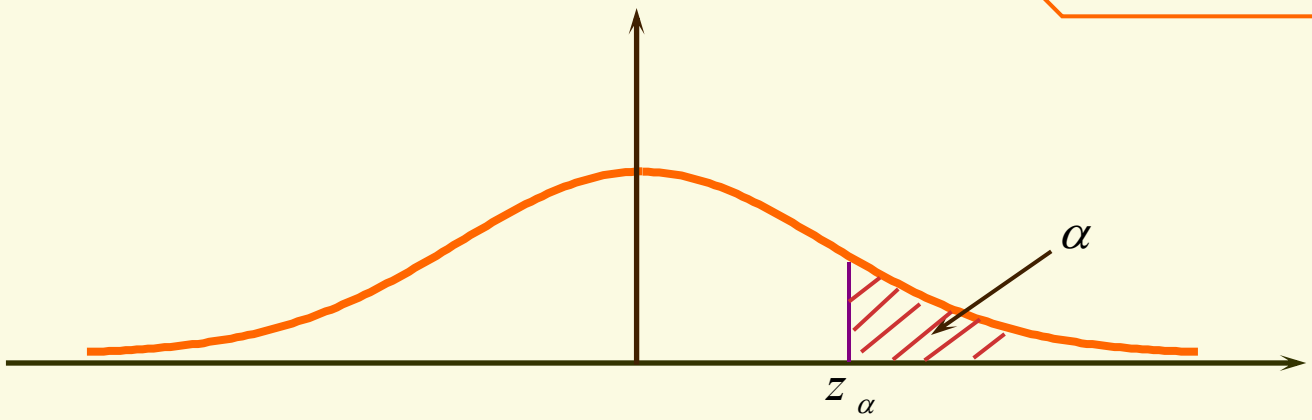
对给定的概率$\alpha, 0<\alpha<1$,称满足条件$\int_{\chi_\alpha^2(n)}^{\infty} f(y;n)dy=\alpha$的点$\chi_\alpha^2(n)$为$\chi^2(n)$分布的上$\alpha$分位数,上$\alpha$分位数$\chi_\alpha^2(n)$的值可查$\chi^2$分布表



标准正态分布的上$\alpha$分位数：设$X \sim N(0,1)$,满足

$P\{X > z_\alpha\} = \alpha \ (0 < \alpha < 1)$的$z_\alpha$。

$$z_{1-\alpha} = -z_\alpha$$



## (二) $t-$分布

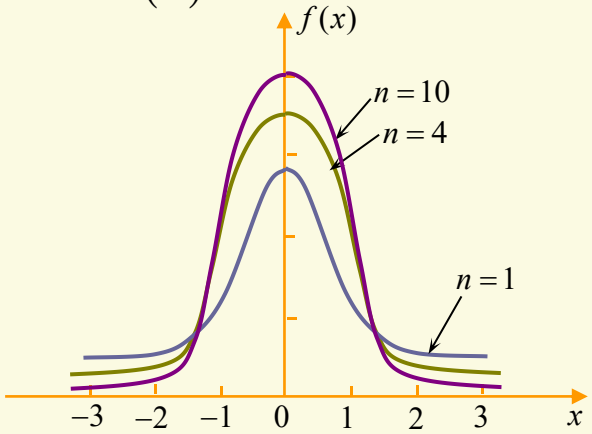
定义：设$X\sim N(0,1),Y\sim\chi^2(n)$,并且$X,Y$相互独立，

则称随机变量$T=\dfrac{X}{\sqrt{Y/n}}$服从自由度为$n$的$t$分布，记为$T\sim t(n)$

定理：$t(n)$分布的概率密度为：$f(t;n)=\dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\dfrac{t^2}{n}\right)^{-\frac{n+1}{2}},-\infty<t<+\infty$

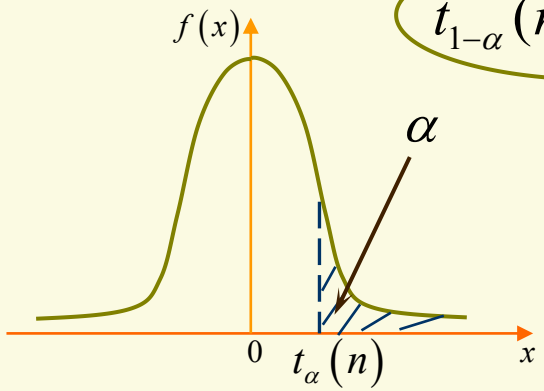
对给定的$\alpha$，$0<\alpha<1$,称满足条件$\int_{t_\alpha(n)}^{\infty}f(t;n)dt=\alpha$的点$t_\alpha(n)$

为$t(n)$分布的上$\alpha$分位数。$t$分布的上$\alpha$分位数可查$t$分布表




$t$分布的密度函数

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




$t$分布的分位数

## (二) $t$–分布

### 性质

1. 具有自由度为$n$的$t$分布$t \sim t(n)$,其数学期望与方差为：$E(t)=0, D(t)=n/(n-2) \quad (n>2)$

2. $t$分布的密度函数关于$t=0$对称.当$n$充分大时,其图形近似于标准正态分布概率密度的图形,

再由$\Gamma$函数的性质有 $\lim\limits_{n\to\infty} h(t)=\dfrac{1}{\sqrt{2\pi}}e^{-t^2/2}$.

即当$n$足够大时，$t \overset{\text{近似}}{\sim} N(0,1)$.

## (二) $t$–分布

学生t-分布可简称为t分布。其推导由威廉·戈塞（William Sealy Gosset，1876.6.13－1937.10.16）于1908年首先发表，当时他还在都柏林的健力士酿酒厂工作。因为不能以他本人的名义发表，所以论文使用了Student这一笔名。之后t检验以及相关理论经由罗纳德·费雪（Sir Ronald Aylmer Fisher，FRS，1890.2.17－1962.7.29）的工作发扬光大，而正是他将此分布称为学生分布。

（http://baike.baidu.com/view/1332600.htm）

'Student' in 1908

## (三) $F$分布

定义：设$X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，且$X, Y$独立，

则称随机变量$F = \dfrac{X/n_1}{Y/n_2}$服从自由度$(n_1, n_2)$的**$F$分布**，记为$F \sim F(n_1, n_2)$

其中$n_1$称为第一自由度，$n_2$称为第二自由度。

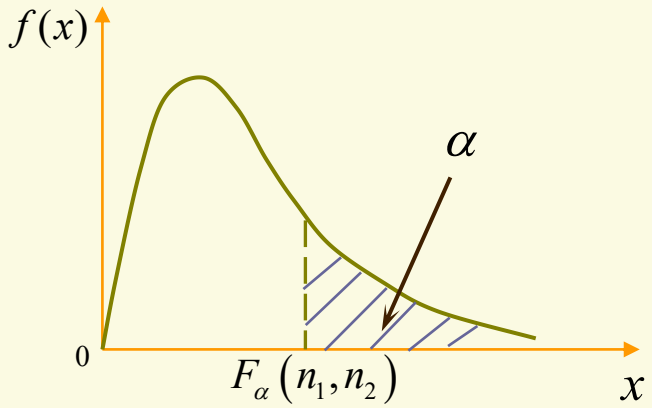
性质：$F \sim F(n_1, n_2)$,则$1/F \sim F(n_2, n_1)$

定理：$F(n_1, n_2)$分布的概率密度为：

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

其中$B(a,b) = \int_0^1 x^{\alpha-1}(1-x)^{b-1} dx = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

对于给定的$\alpha$, $0<\alpha<1$,称满足条件$\int_{F_\alpha(n_1,n_2)}^{\infty} f(x;n_1,n_2)dx=\alpha$的点$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$分布的上$\alpha$分位数。$F_\alpha(n_1,n_2)$的值可查$F$分布表



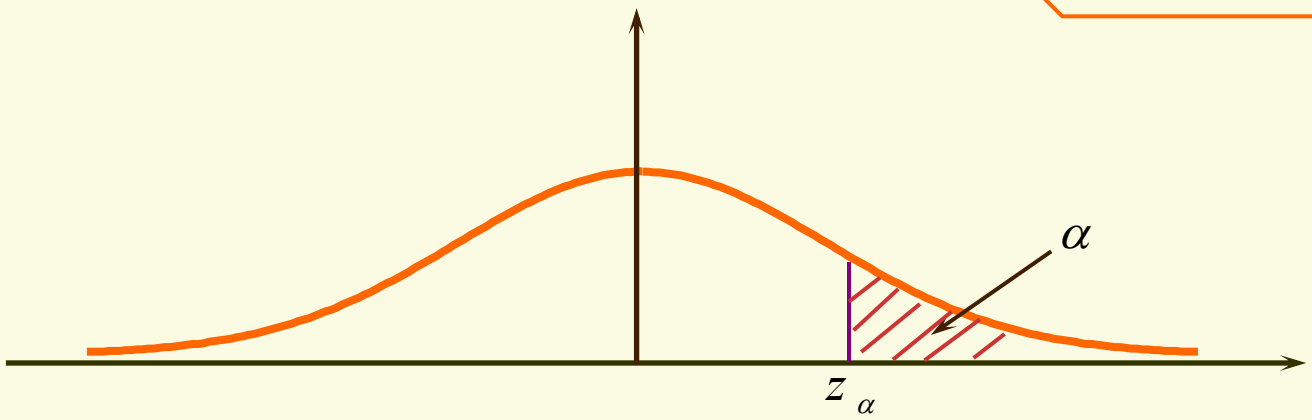

$F$分布的分位数

$$F_{1-\alpha}(n_1,n_2)=[F_\alpha(n_2,n_1)]^{-1}$$

标准正态分布的上$\alpha$分位数：设$X \sim N(0,1)$,满足

$P\{X > z_\alpha\} = \alpha \ (0 < \alpha < 1)$的$z_\alpha$。

$$z_{1-\alpha} = -z_\alpha$$



# 正态总体样本均值和方差的分布

设总体$X$的均值为$\mu$，方差为$\sigma^2$，$X_1, X_2, \cdots, X_n$是来自总体的一个样本，则样本均值$\bar{X}$和样本方差$S^2$有

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

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

$$E(S^2) = D(X) = \sigma^2$$

定理 1 (样本均值的分布)

设 $X_1, X_2, \ldots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}$是样本均值，则有 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

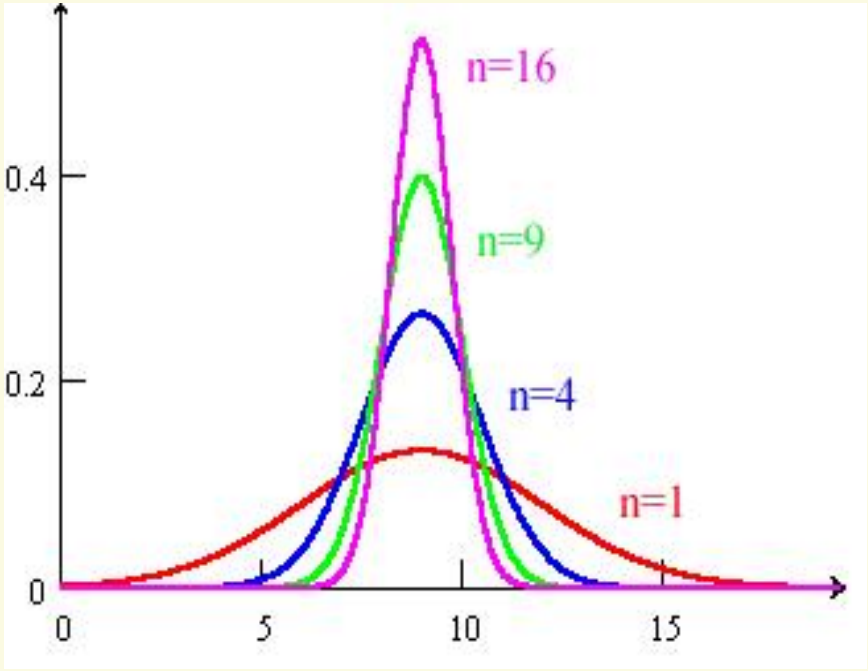
即 $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

（若正态总体的$\mu$，$\sigma^2$已知，可由该定理求样本均值$\bar{X}$）

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

请注意：

$n$取不同值时样本均值 $\bar{X}$ 的分布




基于此，可以证明

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

$$E(S^2) = D(X) = \sigma^2$$

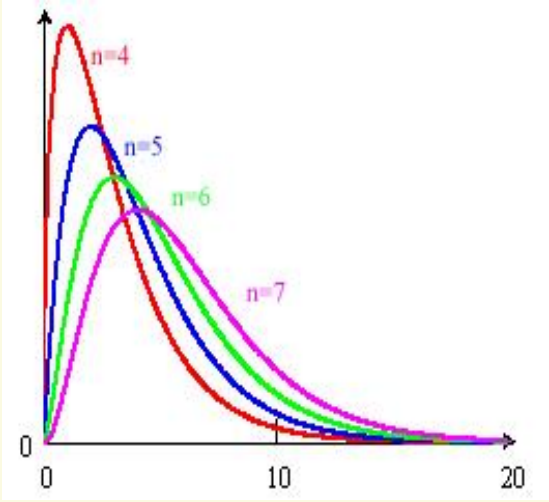
定理 2 (样本方差的分布)

设$X_1,X_2,\ldots,X_n$是来自正态总体 $N(\mu,\sigma^2)$的样本,

$\overline{X}$和$S^2$分别为样本均值和样本方差, 则有

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

**(2) $\overline{X}$与$S^2$独立.**




$n$取不同值时 $\dfrac{(n-1)S^2}{\sigma^2}$ 的分布

(若正态总体的$\mu$, $\sigma^2$已知，可由该定理求样本方差$S^2$)

## 定理 3 (样本均值方差的分布)

设$X_1,X_2,\ldots,X_n$是取自正态总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}$和$S^2$分别为样本均值和样本方差，

则有 $$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

**证 由定理1、2,t分布的定义可得**

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\,, \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$ **且相互独立**

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

这里样本方差$S^2$的值是$s^2 = \dfrac{1}{n-1}\left(\sum_i x_i^2 - n\bar{x}^2\right)$

# 定理 4 (两个正态总体样本均值方差的分布)

定理： 设样本$(X_1,\cdots,X_{n_1})$和$(Y_1,\cdots,Y_{n_2})$分别来自总体$N(\mu_1,{\sigma_1}^2)$和$N(\mu_2,{\sigma_2}^2)$

并且它们相互独立，其样本方差分别为$S_1^2,S_2^2$,

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

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

3° 当${\sigma_1}^2={\sigma_2}^2=\sigma^2$时，$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

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

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

证明:(1) $\dfrac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$, $\dfrac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$

且两者独立，由$F$分布的定义，有：

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

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

且$\bar{X}$与$\bar{Y}$相互独立，所以$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2})$,

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

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

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

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

当$\sigma_1^2=\sigma_2^2=\sigma^2$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$S_W^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, $S_W=\sqrt{S_W^2}$

又由给定条件知：

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

且它们相互独立,故有$\chi^2$分布的可加性知：

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

$U$与$V$相互独立,

于是按$t$分布定义知：

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

例　设总体$X\sim N(0,1)$，$X_1$，$X_2$，…，$X_n$为简单随机样本，试问下列统计量各服从什么分布？

$$(1)\ \frac{X_1-X_2}{\sqrt{X_3^2+X_4^2}};\quad (2)\ \frac{\sqrt{n-1}X_1}{\sqrt{\sum_{i=2}^{n}X_i^2}};\quad (3)\ \frac{(\frac{n}{3}-1)\sum_{i=1}^{3}X_i^2}{\sum_{i=4}^{n}X_i^2}.$$

解 (1)　因为$X_i\sim N(0,1)$，$i=1, 2, \ldots, n$.

所以　$X_1-X_2\sim N(0,2)$，　$\dfrac{X_1-X_2}{\sqrt{2}}\sim N(0,1)$，

$$X_3^2+X_4^2\sim \chi^2(2),$$

故

$$\frac{X_1-X_2}{\sqrt{X_3^2+X_4^2}}=\frac{(X_1-X_2)/\sqrt{2}}{\sqrt{\dfrac{X_3^2+X_4^2}{2}}}\sim t(2).$$

(2) 因为$X_1 \sim N(0,1)$，$\sum_{i=2}^{n} X_i^2 \sim \chi^2(n-1)$

故

$$\frac{\sqrt{n-1}X_1}{\sqrt{\sum_{i=2}^{n} X_i^2}} = \frac{X_1}{\sqrt{\sum_{i=2}^{n} X_i^2 \Big/ (n-1)}} \sim t(n\text{-}1).$$

$$\frac{\sqrt{n-1}X_1}{\sqrt{\sum_{i=2}^{n} X_i^2}}$$

(3) 因为 $\sum_{i=1}^{3} X_i^2 \sim \chi^2(3)$, $\sum_{i=4}^{n} X_i^2 \sim \chi^2(n-3)$,

所以

$$\frac{(\frac{n}{3}-1)\sum_{i=1}^{3} X_i^2}{\sum_{i=4}^{n} X_i^2} = \frac{\sum_{i=1}^{3} X_i^2 \Big/ 3}{\sum_{i=4}^{n} X_i^2 \Big/ (n-3)} \sim F(3, n\text{-}3).$$

$$\frac{(\frac{n}{3}-1)\sum_{i=1}^{3} X_i^2}{\sum_{i=4}^{n} X_i^2}$$

例　若$T\sim t(n)$，　问$T^2$服从什么分布？

例　若$T\sim t(n)$，　问$T^2$服从什么分布？

解：　因为$T\sim t(n)$，可以认为

$$T=\frac{U}{\sqrt{V/n}},$$

其中$U\sim N(0,1)$，　$V\sim\chi^2(n)$，

$$T^2=\frac{U^2}{V/n},\qquad U^2\sim\chi^2(1),$$

$$T^2=\frac{U^2/1}{V/n}\sim F(1,n).$$

**作业：5，9，10（due：6.13）**