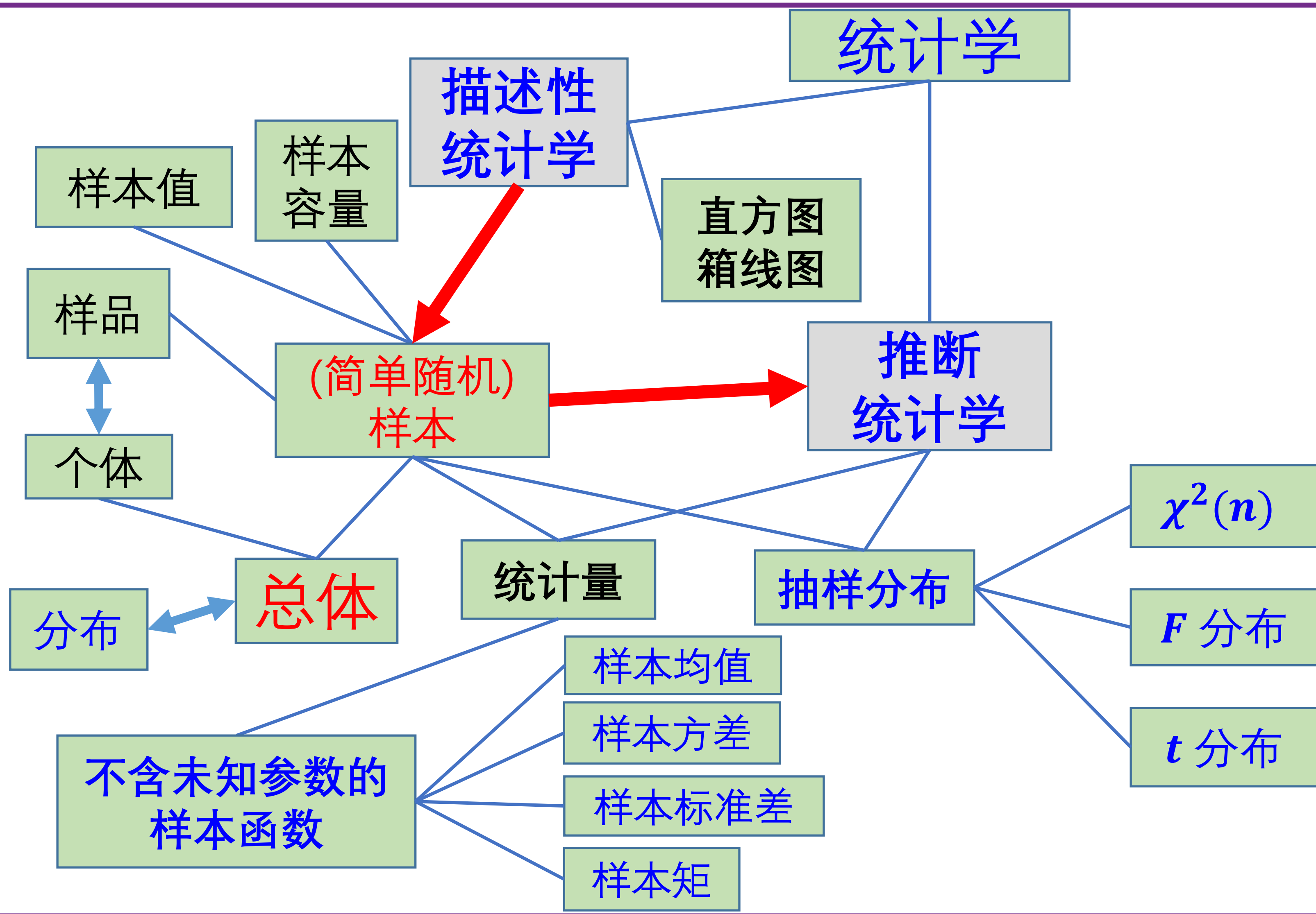
PEKING UNIVERSITY

粒子物理与核物理实验中的数据分析

补充：样本和抽样分布

杨振伟

样本及抽样分布的概念框图

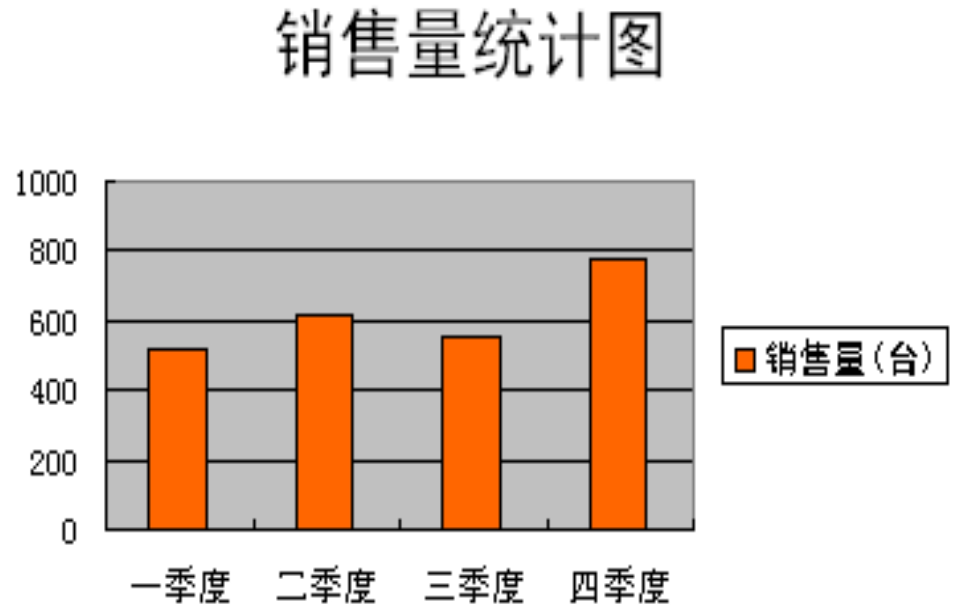


本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布

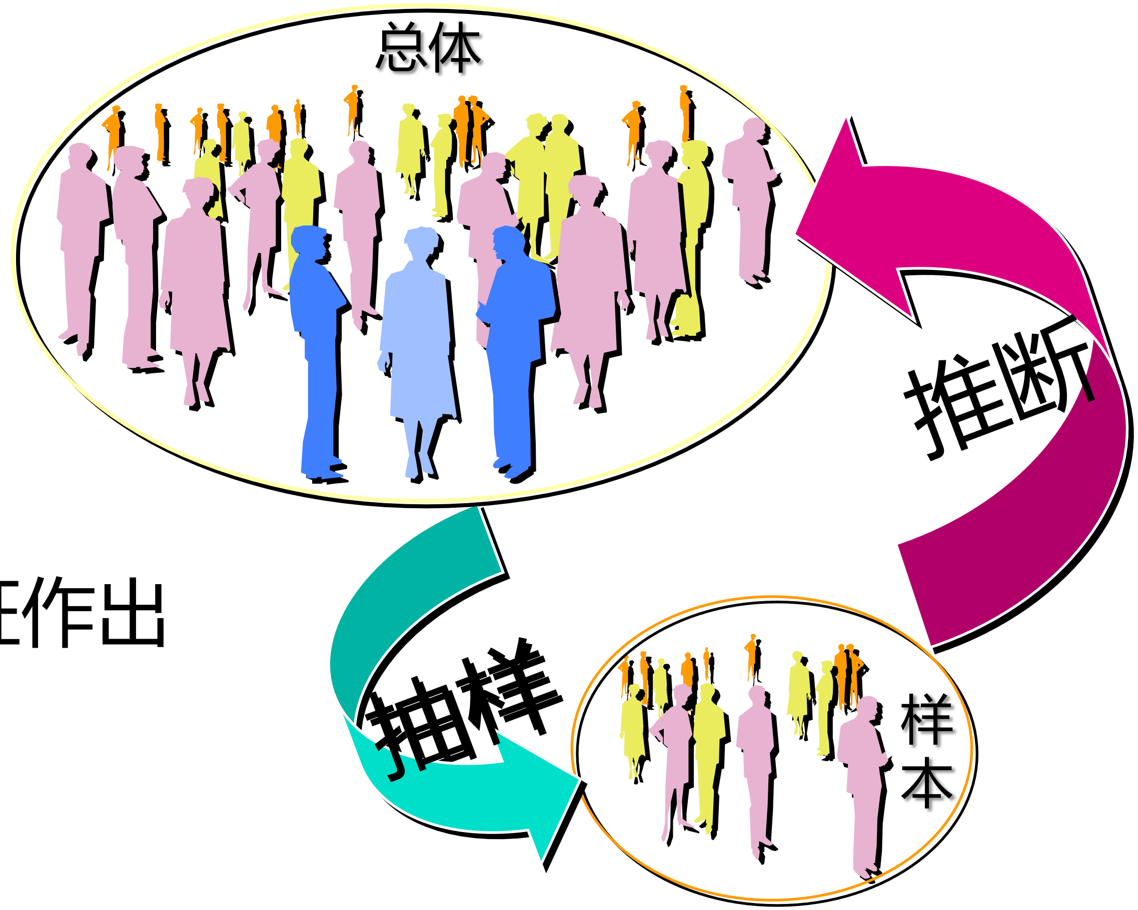
描述性统计学（descriptive statistics）

1. 内容
   - 搜集数据
   - 整理数据
   - 展示数据
   - 描述性分析
2. 目的
   - 描述数据特征
   - 找出数据的基本规律



推断统计(inferential statistics)

1. 内容
   - 参数估计
   - 假设检验
2. 目的
   - 对总体特征作出推断



描述性统计学和推断统计学

反映客观现象的数据

概率论

（包括分布理论、大数定律和中心极限定理等）

描述统计

（统计数据的搜集、整理、显示和分析等）

推断统计

（利用样本信息和概率论对总体数量特征进行估计和检验等）

总体内在的数量规律性

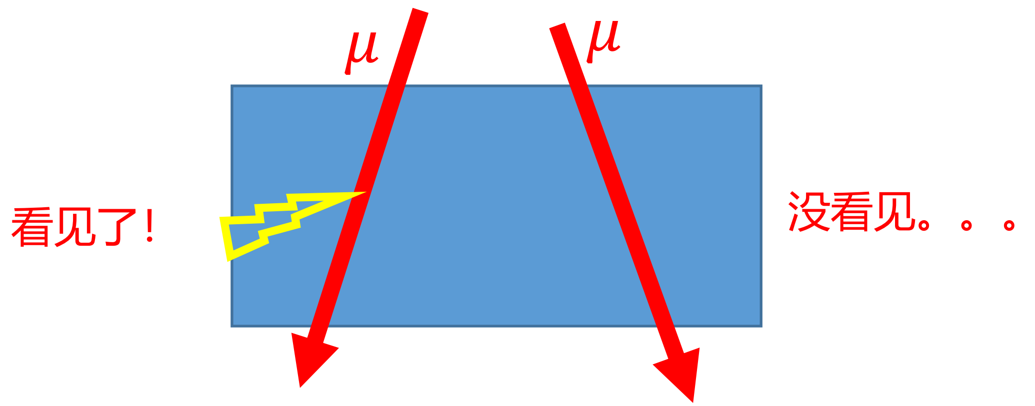
描述统计是整个统计学的基础，推断统计是现代统计学的核心和标志。

本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布

为什么需要随机样本?

探测器效率 —— 假设搭建一个宇宙线缪子（$\mu^{\pm}$）探测器。要求探测效率$\varepsilon > 95\%$.



探测器效率:

$$\varepsilon = \frac{n_{\text{obs}}}{n_{\text{tot}}}$$

$$n_{\text{tot}} \to \infty$$

为了研究效率，需要采集一定的宇宙线事例（样本）。

总体和样本

总体 —— 研究对象全体元素组成的集合

所研究的对象的某个(或某些)数量指标的全体，它是一个随机变量(或多维随机变量). 记为X .

X 的分布和数字特征称为总体的分布和数字特征.

总体的三层含义:
1）研究对象的全体；2）数据；3）分布

总体

有限总体

无限总体

个体、样本、样本空间

个体 —— 组成总体的每一个元素,

即总体的每个数量指标, 可看作随机变量 X 的某个取值.

用X_i表示.

样本 —— 从总体中抽取的部分个体.

样本取自总体
样本不唯一

用$(X_1, X_2, \cdots, X_n)$表示， n 为样本容量.

称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为n的样本观测值

样本空间 —— 样本所有可能取值的集合.

简单随机样本

若总体 X 的样本$(X_1, X_2, \cdots, X_n)$满足:

(1) $X_1, X_2, \cdots, X_n$ 与X 有相同的分布

(2) $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本.

一般来说, 对有限总体, 放回抽样所得到的样本为简单随机样本, 但使用不方便, 常用不放回抽样代替. 而代替的条件是

$$N/n \geq 10$$

N 为总体中个体总数， n 为样本容量。

本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布

数据的展示

实验采集的原始数据往往是一个个离散的样本点，很难从中直接得到关键信息。只有经过整理后，用常用的表示方法展示后，数据中的信息才会变得直观。

常用的表示方法有列表法和图示法。

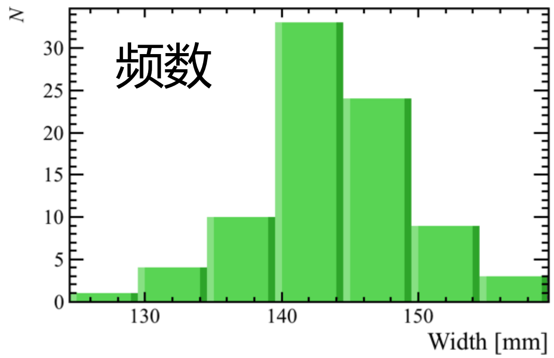
列表法：频数分布表、频率分布表 ……

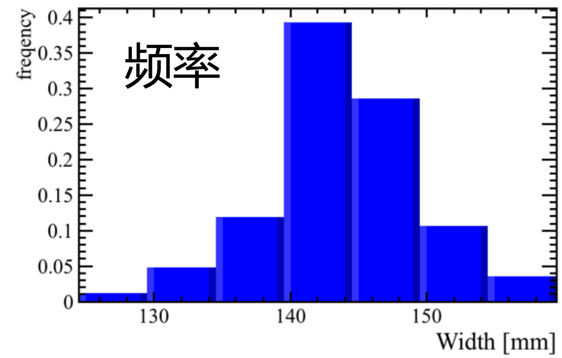
图示法：频数直方图、频率直方图、箱线图 ……

频数直方图与频率直方图

直方图中每个小矩形的高，有时直接用频数或单位组距的频数，有时用频率或单位组距的频率，分别称为频数直方图和频率直方图。作图时要在直方图纵坐标注明。

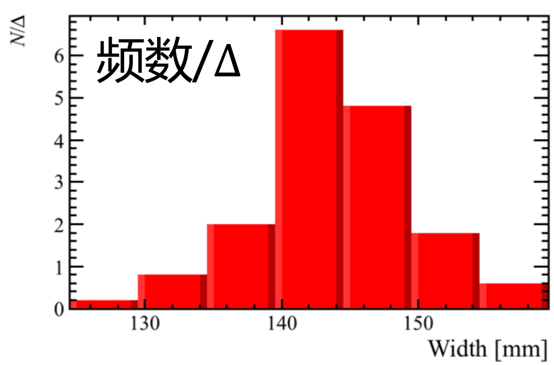
频数累加等于样本总量n

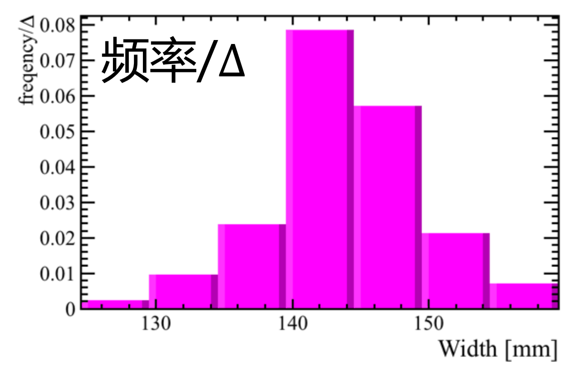




频率累加等于1

面积累加等于样本总量n





面积累加等于1

箱线图

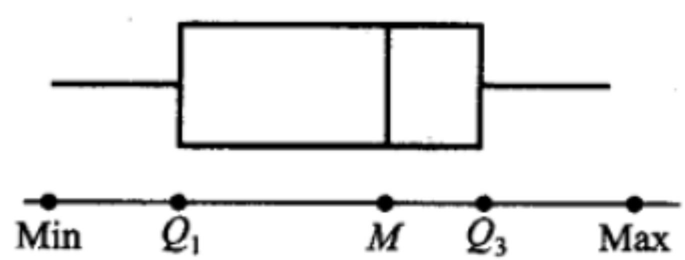
数据集的箱线图是箱子和直线组成的图形，它的定义用到了数据的最小最大值和几个样本分位数。

样本分位数：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，它具有以下性质：

(1) 至少有np个观察值小于或等于x_p；

(2) 至少有$n(1-p)$个观察值大于或等于x_p。

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



箱线图

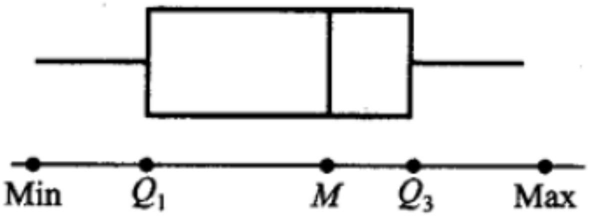
常用样本分位数：

分位数	p	符号	名称
$x_{0.25}$	$p=0.25$	Q_1	第一四分位数
$x_{0.5}$	$p=0.5$	Q_2或M	中分位数
$x_{0.75}$	$p=0.75$	Q_3	第三四分位数

箱线图是基于数据最小值Min、最大值Max、三个常用分位数Q_1, M, Q_3，用箱子和线画出的图形。做法如下：

(1) 画一水平数轴，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点上方箱子内部画一条垂直线段。

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



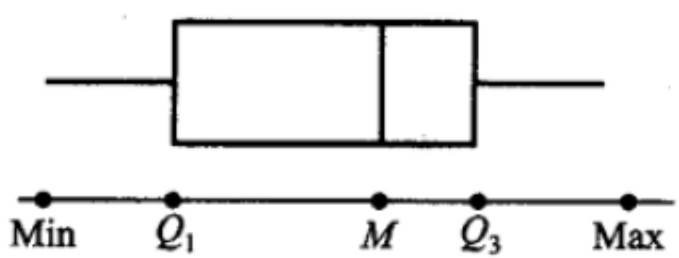
箱线图

自箱线图可形象看出数据集的某些重要性质：

(1) 中心位置：中位数所在的位置即数据集的中心。

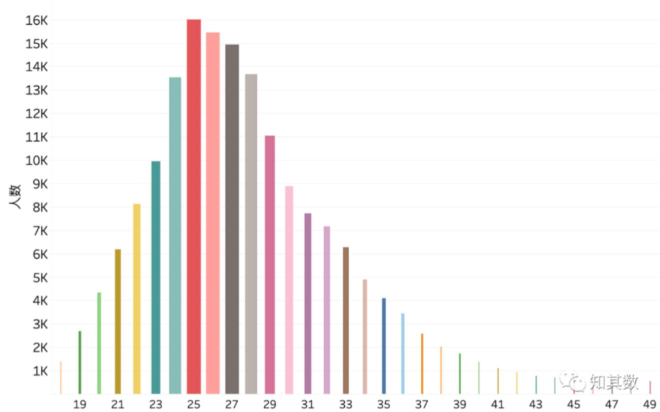
(2) 散布程度：全部数据都落在[Min, Max]之内，在区间[Min, Q_1]，$[Q_1, M]$，$[M, Q_3]$，[Q_3,Max]的数据个数各占$1/4$。区间较短时，表示落在该区间的点比较集中；反之则较为分散。

(3) 关于对称性：若中位数位于箱子的中间位置，则数据分布较为对称。若Min离M的距离较Max离M的距离大，则表示数据分布向左倾斜，反之则向右倾斜，且能看出分布尾部的长短。

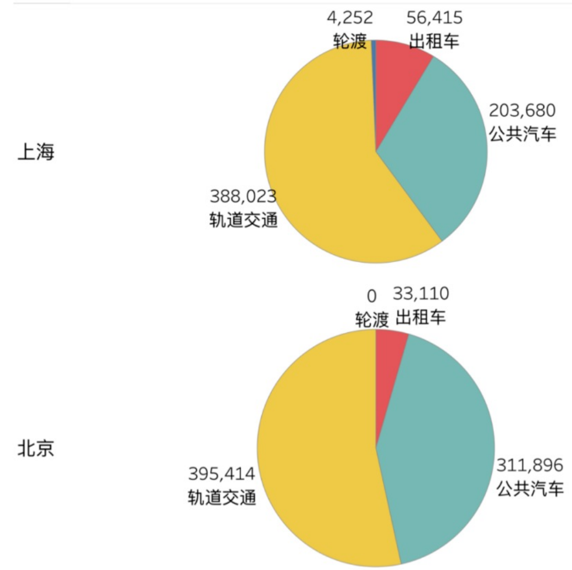


数据展示示例

新出生人口数按母亲年龄分类统计（抽样）

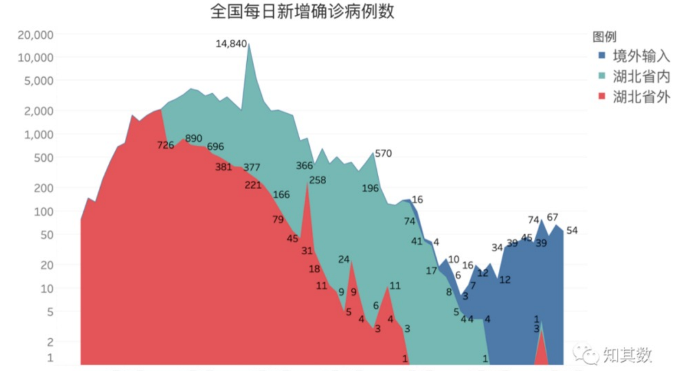


城市客运量分析（总量前四名）

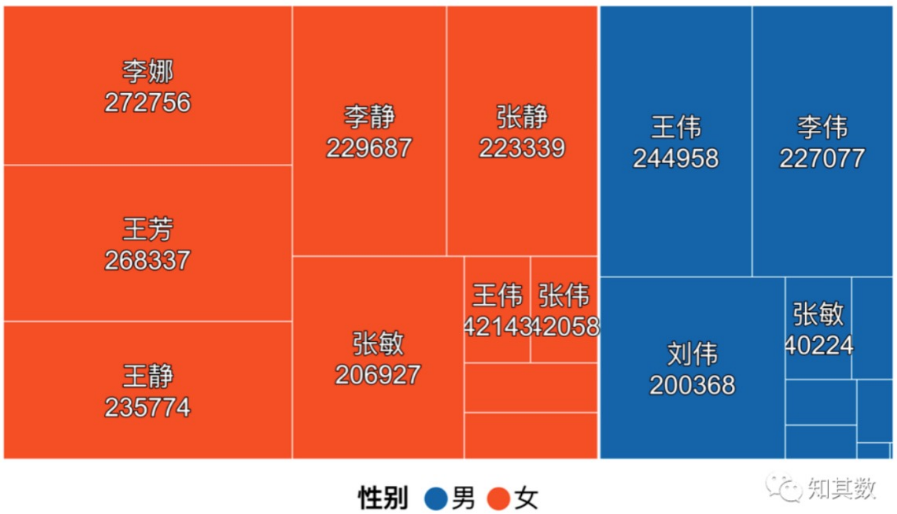


主要城市市内客运总量

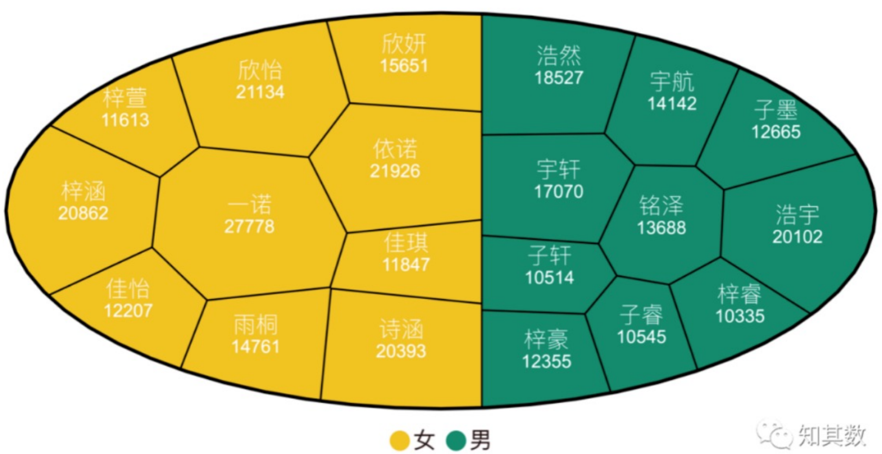
全国每日新增确诊病例数



数据展示示例



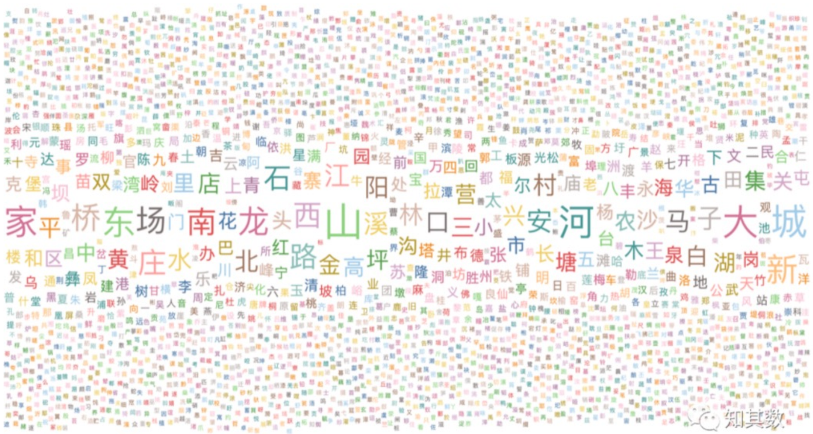
"李娜"和"王伟"是赢家

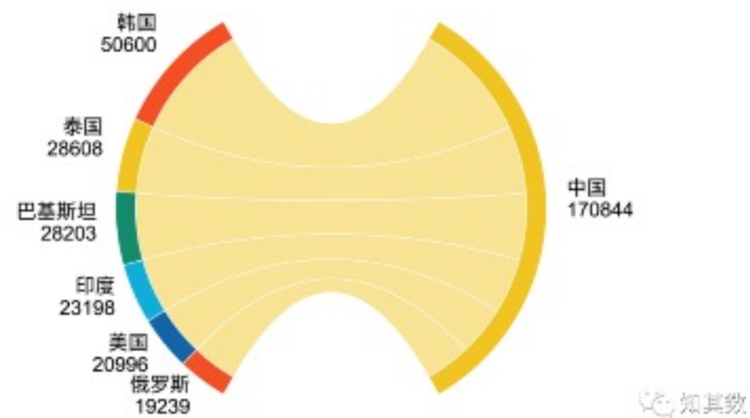


新父母宠爱
"依诺"和"浩宇"

数据展示示例

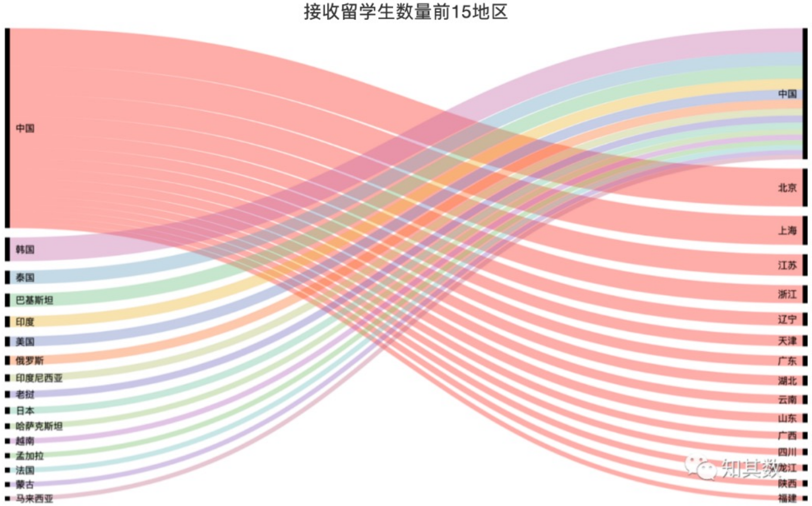
中国乡级行政区名用字统计图

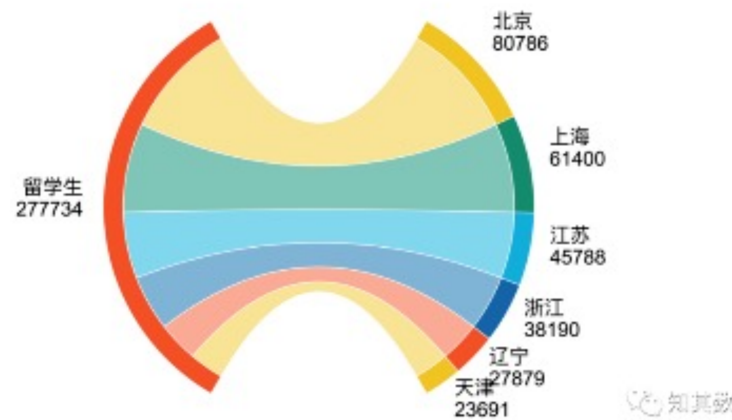




留学生来华数量前15国家

接收留学生数量前15地区





本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布

统计量的概念

利用样本的函数进行统计推断

定义 样本$(X_1, X_2, \cdots, X_n)$的不含有未知参数的连续函数 $g(X_1, X_2, \cdots, X_n)$称为统计量。

样本是随机变量，统计量也是随机变量

函数	参数 μ, σ^2	是否统计量
$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$	已知	是
	未知	否

常用的统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

(1) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 样本均值

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

(3) $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ 样本标准差

(4) $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 样本的k阶原点矩 $A_1 = \bar{X}$

(5) $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ 样本的k阶中心矩 $B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$

定理

定理1：若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为零，即 $\sum_{i=1}^{n}(X_i-\bar{X})=0$。

证明：

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

定理2：数据观察值与样本均值的偏差平方和最小，即在形如$\sum_{i=1}^{n}(X_i-c)^2$的函数中，$\sum_{i=1}^{n}(X_i-\bar{X})^2$最小。

证明：

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

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

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-c)^2\geq\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

定理

定理3：设 $X_1, X_2, \dots, X_n$ 是来自某个总体的样本， $\bar{X}$ 是样本均值。（1）若总体分布为$N(\mu, \sigma^2)$，则$\bar{X} \sim N(\mu, \sigma^2/n)$。（2）若总体分布未知或不是正态分布，但$E[X] = \mu$，$\mathrm{Var}[X] = \sigma^2$存在，则$n$较大时，$\bar{X}$的渐进分布为$N(\mu, \sigma^2/n)$。

定理4：设总体X具有二阶矩，即$E[X] = \mu$，$\mathrm{Var}[X] = \sigma^2 < +\infty$，$X_1, X_2, \dots, X_n$是从这个总体得到的样本，$\bar{X}$和$S^2$分别是样本均值和样本方差，则

$$E[\bar{X}] = \mu, \qquad \mathrm{Var}[\bar{X}] = \sigma^2/n$$

$$E[S^2] = \sigma^2$$

样本方差S^2与样本二阶中心矩S_n^2

1） 关系式 $$S^2=\frac{n}{n-1}S_n^2$$

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

$$=\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$$

➔ 常用计算公式

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

样本方差S^2与样本二阶中心矩S_n^2

2） $E(S_n^2) = \dfrac{n-1}{n}\sigma^2, \qquad E(S^2) = \sigma^2$

推导 设 $E(X) = \mu$, $D(X) = \sigma^2$, 则

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

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

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

$$= \frac{1}{n}n(\sigma^2 + \mu^2) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

$$E(S^2) = E\left[\frac{n}{n-1}S_n^2\right] = \frac{n}{n-1}E(S_n^2) = \sigma^2$$

本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布

抽样分布

统计量是仅依赖于样本的随机变量，所以它必有一个概率分布

抽样分布 —— 统计量 $T_n = g(X_1, X_2, \cdots, X_n)$的分布称为抽样分布。

(1) χ^2分布

(2) t 分布

(3) F 分布

(4) 正态总体样本均值和样本方差的分布

抽样分布：χ^2分布

定义 设$X_1, X_2, \cdots, X_n$相互独立，且都服从标准正态分布$N(0,1)$，则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

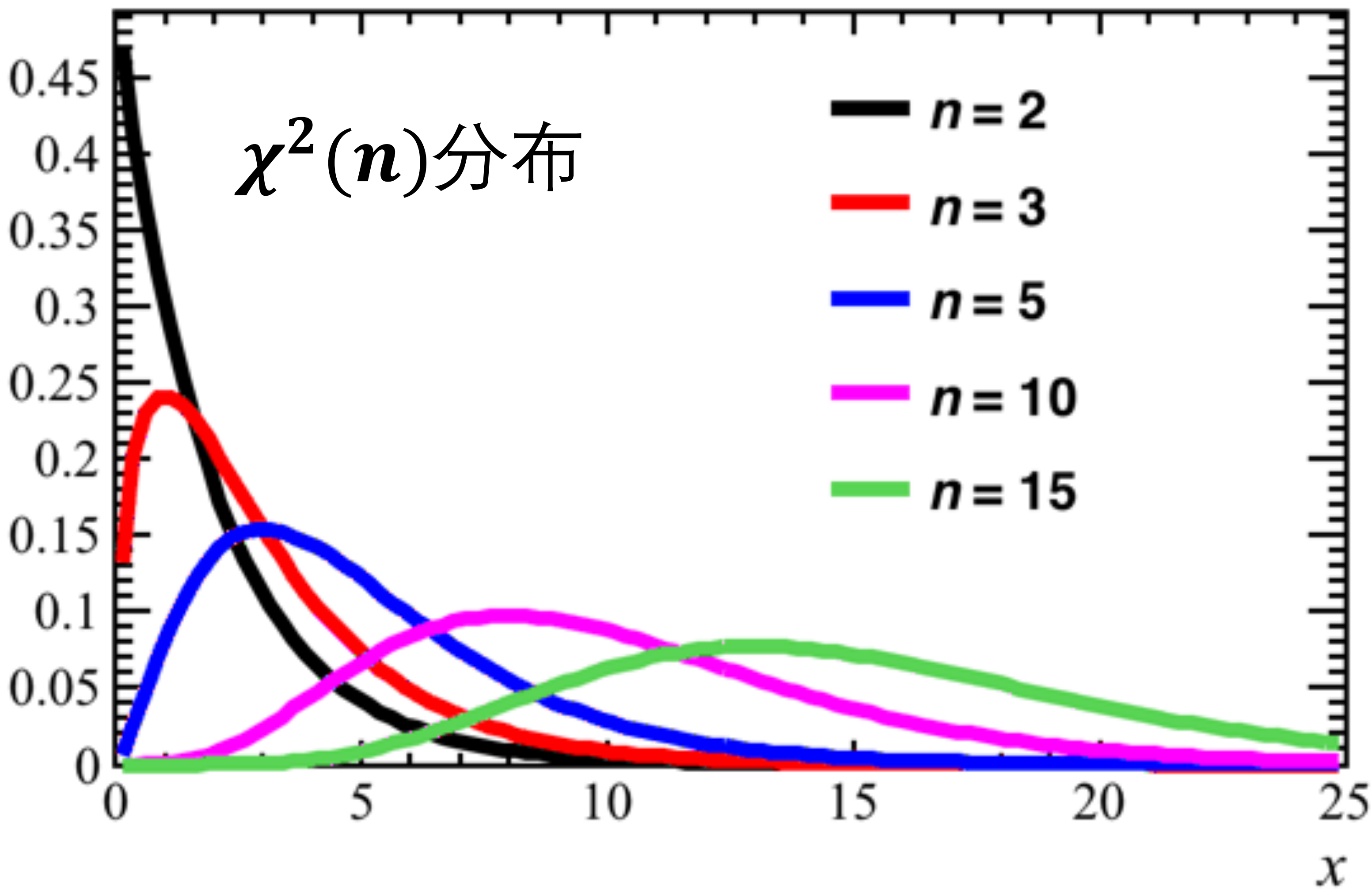
（n为自由度，即求和中独立变量的个数）

$\chi^2(n)$是$Ga(n/2,1/2)$，概率密度为

$$f(y) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}$$

$$Ga(\alpha,\lambda):\ f(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x \geq 0$$

$\chi^2(n)$分布



$\chi^2(n)$分布
n = 2
n = 3
n = 5
n = 10
n = 15
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
x

$\chi^2(n)$分布的证明

证明：

1. 第二章证明过：若$X \sim N(0,1)$，则 $X^2 \sim \chi^2(1)$ ，即 $Ga(1/2,1/2)$。

2. 伽玛分布的可加性：

若$X_1 \sim Ga(\alpha_1, \lambda), X_2 \sim Ga(\alpha_2, \lambda)$，且$X_1, X_2$相互独立，则

$$Z = X_1 + X_2 \sim Ga(\alpha_1 + \alpha_2, \lambda).$$

证明如下： 首先$Z \in (0, \infty)$。当$z \leq 0$时$f_Z(z) = 0$。

利用傅里叶卷积公式，

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z - x_2) f_Y(x_2) \mathrm{d}x_2$$

给定 z 时，$x_2 \in (0, z)$，积分变为

$$f_Z(z) = \int_0^z f_X(z - x_2) f_Y(x_2) \mathrm{d}x_2$$

$\chi^2(n)$分布的证明

$$f_Z(z)=\int_0^z f_X(z-x_2)f_Y(x_2)\mathrm{d}x_2$$

$$=\int_0^z \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)}(z-x_2)^{\alpha_1-1}e^{-\lambda(z-x_2)}\frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)}x_2^{\alpha_2-1}e^{-\lambda x_2}\mathrm{d}x_2$$

$$=\frac{\lambda^{\alpha_1}\lambda^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^z (z-x_2)^{\alpha_1-1}x_2^{\alpha_2-1}e^{-\lambda(z-x_2)-\lambda x_2}\mathrm{d}x_2$$

$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^z (z-x_2)^{\alpha_1-1}x_2^{\alpha_2-1}\mathrm{d}x_2$$

变量替换$t=x_2/z$，
$x_2=zt$，$\mathrm{d}x_2=z\mathrm{d}t$

$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}z^{\alpha_1+\alpha_2-1}\int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}\mathrm{d}t$$

$\chi^2(n)$分布的证明

$$f_Z(z)=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}z^{\alpha_1+\alpha_2-1}\underline{\int_0^1(1-t)^{\alpha_1-1}t^{\alpha_2-1}\mathrm{d}t}$$

$\boxed{\text{贝塔函数}B(\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}}$

$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}z^{\alpha_1+\alpha_2-1}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

$$=\frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}z^{\alpha_1+\alpha_2-1}e^{-\lambda z}$$

$\Rightarrow Z\sim Ga(\alpha_1+\alpha_2,\lambda)$

n个服从$\chi^2(1)$分布的独立随机变量之和服从$\chi^2(n)$。

$\sum_{i=1}^n X_i^2\sim\chi^2(n)$ 其中X_i相互独立，$X_i\sim N(0,1)$。

$\chi^2(n)$分布的性质

1. $E[\chi^2(n)] = n$, $D[\chi^2(n)] = 2n$

2. 若$X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, X_1, X_2相互独立，则 $X_1 + X_2 \sim \chi^2(n_1 + n_2)$

3. 满足$P\left(\chi^2 > \chi^2_\alpha(n)\right) = \int_{\chi^2_\alpha(n)}^{\infty} f(y)\mathrm{d}y = \alpha$ 的点$\chi^2_\alpha(n)$称为 $\chi^2(n)$分布的上 α 分位数。
其中 $f(y)$为$\chi^2(n)$分布的概率密度函数。

可通过查表或者在统计软件（例如ROOT）中得到分位数：
ROOT::Math::chisquared_quantile_c(α, ndf)

抽样分布：F分布

定义

设随机变量 $X \sim \chi^2(n), Y \sim \chi^2(m)$，且 X, Y相互独立。令

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

则称 F 服从第一自由度为n，第二自由度为 m 的 F 分布.

F分布的密度函数

首先考察$Z=X/Y$的密度函数。记$f_X(x)$和$f_Y(y)$分别为$\chi^2(n)$和$\chi^2(m)$的密度函数。独立随机变量的商的密度函数公式，

$$f_Z(z)=\int_{-\infty}^{\infty} y f_X(zy) f_Y(y)\mathrm{d}y$$

$$f_X(x)=\frac{(1/2)^{n/2}}{\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}},\qquad f_Y(y)=\frac{(1/2)^{m/2}}{\Gamma(m/2)}y^{\frac{m}{2}-1}e^{-\frac{y}{2}}$$

$$f_Z(z)=\frac{(1/2)^{\frac{m+n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty} y(zy)^{\frac{n}{2}-1}e^{-\frac{zy}{2}}y^{\frac{m}{2}-1}e^{-\frac{y}{2}}\mathrm{d}y$$

$$=\frac{(1/2)^{\frac{m+n}{2}}z^{\frac{n}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty} y^{\frac{m+n}{2}-1}e^{-\frac{y}{2}(z+1)}\mathrm{d}y$$

做变量替换：$u = y(z+1)/2$。
则$y = 2u/(1+z)$，$\mathrm{d}y = 2du/(1+z)$

$$f_Z(z) = \frac{(1/2)^{\frac{m+n}{2}} z^{\frac{n}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} y^{\frac{m+n}{2}-1} e^{-\frac{y}{2}(z+1)} \mathrm{d}y$$

$$= \frac{(1/2)^{\frac{m+n}{2}} z^{\frac{n}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} \left(\frac{2u}{1+z}\right)^{\frac{m+n}{2}-1} e^{-u} \frac{2}{1+z} \,\mathrm{d}u$$

$$= \frac{z^{\frac{n}{2}-1}(1+z)^{-\frac{m+n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} u^{\frac{m+n}{2}-1} e^{-u} \,\mathrm{d}u$$

$$= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1}(1+z)^{-\frac{m+n}{2}} \qquad \text{其中 } z \in (0,\infty)$$

下面导出 $F=\frac{m}{n}Z$ 的密度函数：$f_F(t)$。对$t>0$，有

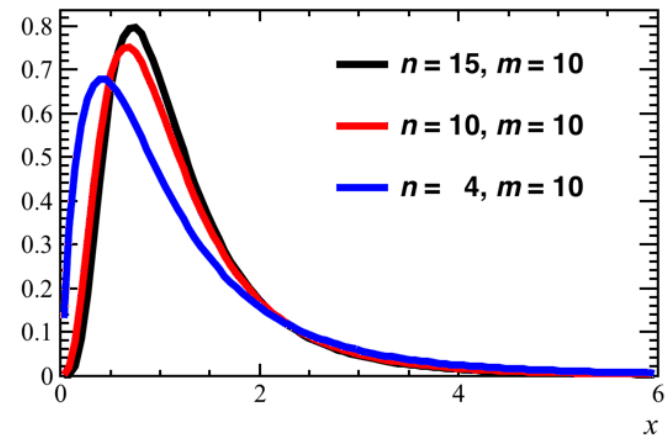
$$f_F(t)=f_Z\left(\frac{n}{m}t\right)\cdot\frac{n}{m}$$

$$=\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{nt}{m}\right)^{\frac{n}{2}-1}\left(1+\frac{nt}{m}\right)^{-\frac{m+n}{2}}\cdot\frac{n}{m}$$

$$=\frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}t^{\frac{n}{2}-1}\left(1+\frac{nt}{m}\right)^{-\frac{m+n}{2}}$$

F 分布： $F(n,m)=\frac{X/n}{Y/m}$

$$f_F(x)=\frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\left(1+\frac{n}{m}x\right)^{-\frac{m+n}{2}}$$



F 分布的性质

1. 若 $F \sim F(n,m)$，则$1/F \sim F(m,n)$

2. 满足$P\big(F > F_\alpha(n,m)\big) = \alpha$的点$F_\alpha(n,m)$称为 $F(n,m)$分布的上 α 分位数。

> 可通过查表或者在统计软件（例如ROOT）中得到分位数：
> ROOT::Math::fdistribution_quantile_c(α, n, m)

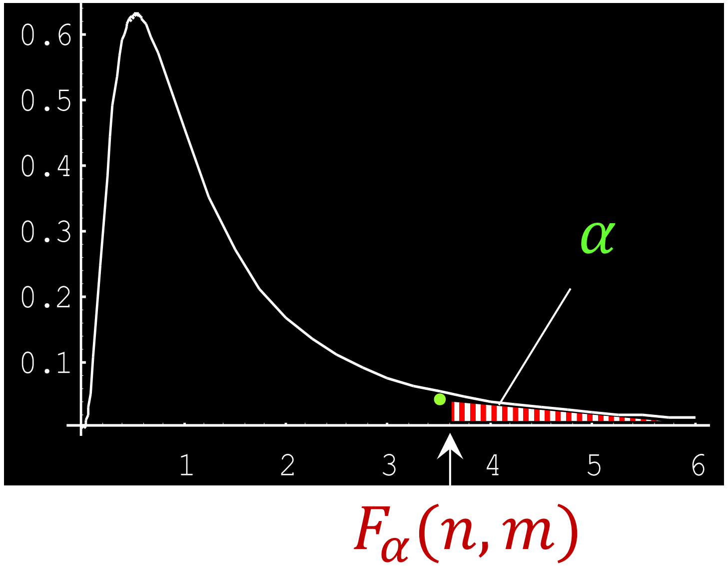
3. $F_{1-\alpha}(n,m) = \dfrac{1}{F_\alpha(m,n)}$

例如 $F_{0.05}(4,5) = 5.19$

求 $F_{0.95}(5,4) =?$

$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}$$

$$F_{0.95}(5,4) = \frac{1}{F_{0.05}(4,5)} = \frac{1}{5.19} = 0.193$$



抽样分布：t分布

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

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

为服从自由度为 n 的 t 分布，即Student 分布，学生氏分布.

t分布的密度函数

$$T=\frac{X}{\sqrt{Y/n}}$$ 与 $F(n,m)=\frac{X^2/n}{Y^2/m}$ 对比

由于 $T^2=\frac{X^2/1}{Y^2/n}$，所以$T^2\sim F(1,n)$。

$$f_{T^2}(t^2)=\frac{\Gamma\left(\frac{n+1}{2}\right)\left(\frac{1}{n}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}(t^2)^{-\frac{1}{2}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$
$$=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}t^{-1}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

下面考虑如何从T^2的密度函数导出T的密度函数。

由于$X \sim N(0,1)$，则 $-X \sim N(0,1)$，所以

$T = X/\sqrt{Y/n}$ 与 $-T$ 服从相同分布。

那么，对任意实数t，有

$$P(0 < T < t) = P(0 < -T < t) = P(-t < T < 0)$$

$$\Rightarrow \qquad P(0 < T < t) = \frac{1}{2} P(T^2 < t^2)$$

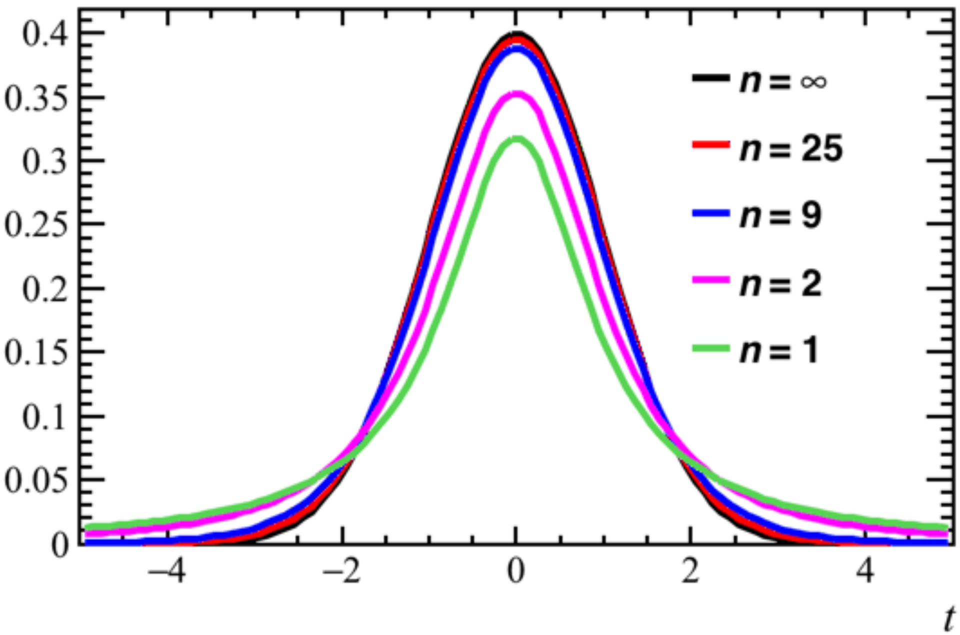
上式左边是T的分布函数减去某常数，右边是T^2的分布函数的1/2。
左右两边求导可得相应的密度函数:

$$f_T(t) = t f_{T^2}(t^2) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad t \in (-\infty, \infty)$$

t分布

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

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$



$n \to \infty$时

$$\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \to e^{-t^2}$$

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\left(\frac{n}{2}\right)^{1/2}} \to 1$$

t 分布的图形($n = \infty$时变成标准正态分布)

t分布的性质

1° $f_n(t)$ 是偶函数,

$$n \to \infty, \qquad f_n(t) \to \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2° 满足$P(T > t_\alpha) = \alpha$的点t_α称为 t 分布的上α分位数；满足$P(|T| > t_{\alpha/2}) = \alpha$的点$t_{\alpha/2}$称为 t 分布的双侧α分位数.

可通过查表或者在统计软件（例如ROOT）中得到分位数:
ROOT::Math::tdistribution_quantile_c(α, n)

正态总体样本均值和样本方差的分布

样本均值的分布

设总体 $X \sim N(\mu, \sigma^2)$，样本为$(X_1, \cdots, X_n)$，

则样本均值（μ和σ^2已知）

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

或

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

即正态总体的样本均值服从高斯分布，与总体相比，均值相同，方差减小为$1/n$.

正态总体样本均值和样本方差的分布

样本方差的分布——σ^2已知

设总体 $X \sim N(\mu,\sigma^2)$，样本为$(X_1,\cdots,X_n)$，

则$\frac{n-1}{\sigma^2}$与样本方差S^2的乘积（σ^2已知）

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

【证明过程复杂，需要先证 $\bar{X}$ 与 S^2 相互独立。】

样本方差的分布— σ^2未知时

设总体 $X \sim N(\mu,\sigma^2)$，样本为$(X_1,\cdots,X_n)$，

则有（μ已知）

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

正态总体样本均值和样本方差的分布

总结： 设总体 $X \sim N(\mu, \sigma^2)$，样本为$(X_1, \cdots, X_n)$，

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

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

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

又设总体 $X' \sim N(\mu', \sigma'^2)$，样本为$(X_1', \cdots, X_{n'}')$，

且样本$(X_1', \cdots, X_{n'}')$与$(X_1, \cdots, X_n)$相互独立。则，

$$\boxed{\frac{S^2/S'^2}{\sigma^2/\sigma'^2} \sim F(n-1, n'-1)}$$

本章要点

- 数据收集、数据描述
- 随机样本
- 直方图
- 样本分布的数字特征
- 几个常用统计量的分布