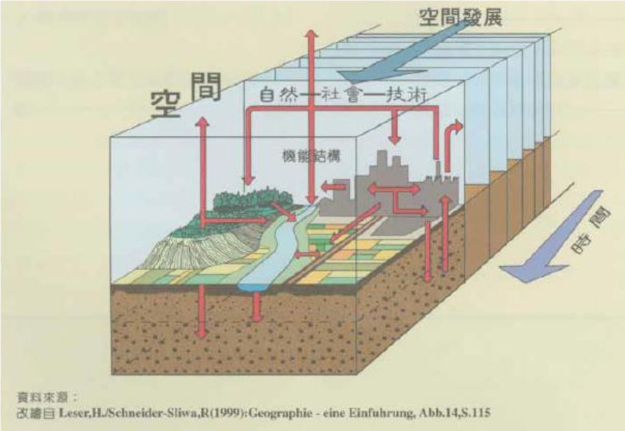
地理学研究中的概率函数和统计假设检验

——河北师范大学资环学院　　胡引翠

地理系统和地理要素的随机性质



复杂系统

不能从数量上确定地理系统状态及地理要素的确定性变化规律。

自然界中的有两类现象：

1. 确定性现象

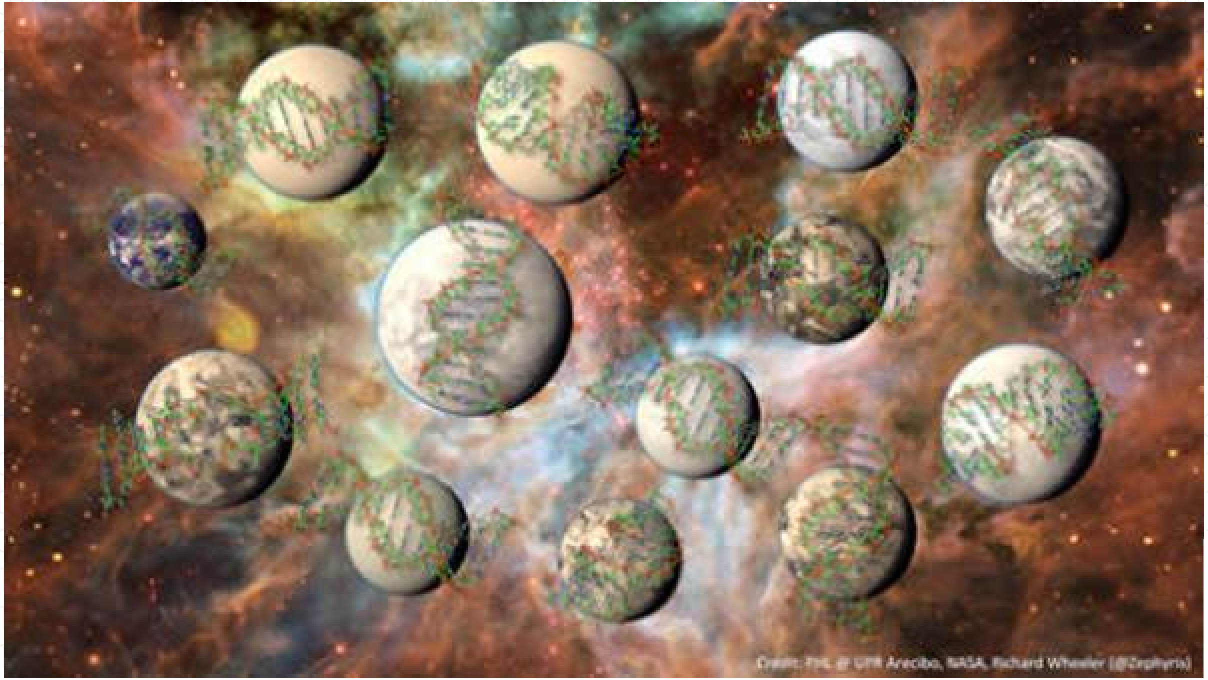
- 每天早晨太阳从东方升起;
- 水在标准大气压下加温到100°C沸腾;

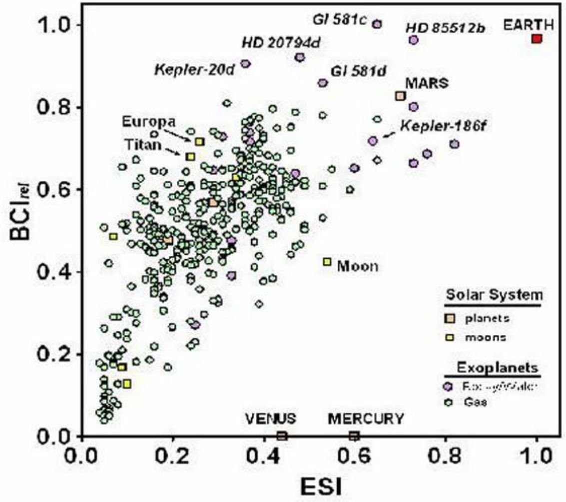
2. 随机现象

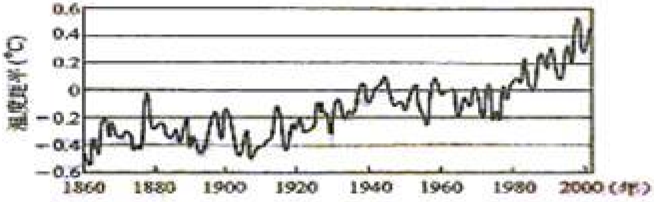
- 掷一枚硬币，正面朝上？反面朝上？
- 一天内进入某超市的顾客数;
- 某种型号电视机的寿命;

- **随机现象**：在一定的条件下，并不总出现相同结果的现象称为随机现象.
- **特点**：1. 结果不止一个;
  2. 事先不知道哪一个会出现.
- **随机现象的统计规律性**：随机现象的各种结果会表现出一定的规律性，这种规律性称之为统计规律性.

美国天文学家认为银河系可能拥有大约1亿颗可支持生命存在的行星。生命复杂性指数BCI和地球相似性指数ESI。







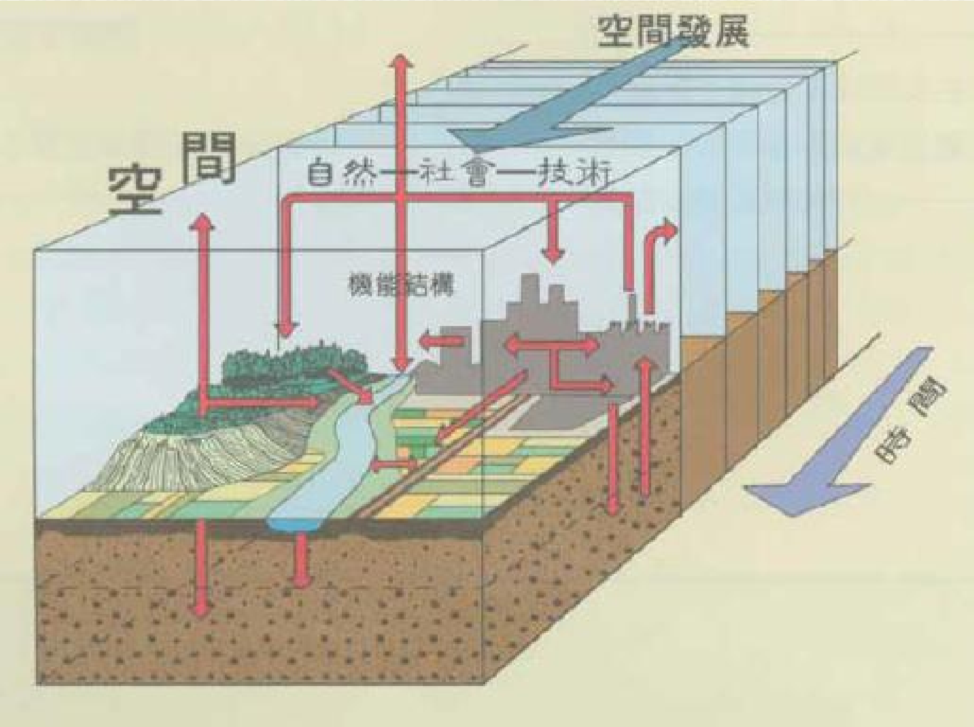
1860－2000年全球气温变化趋势

本章内容

01
地理学中的概率函数

02
地理数据的空间分布

03
地理学研究中的统计假设检验



1、地理学中的概率函数：

随机变量：在多次随机实验中，试验的**每一种可能**对应着一个数值，按习惯，我们将这些数值**与一个变量的取值联系**起来，这样**与随机试验结果联系**在一起的变量我们称之为随机变量。

随机变量

随机变量在一定条件下，因随机因素影响而在试验结果中取不同数值的量，随机变量具有偶然性与规律性。

设X为一随机变量，x是任意实数，称函数

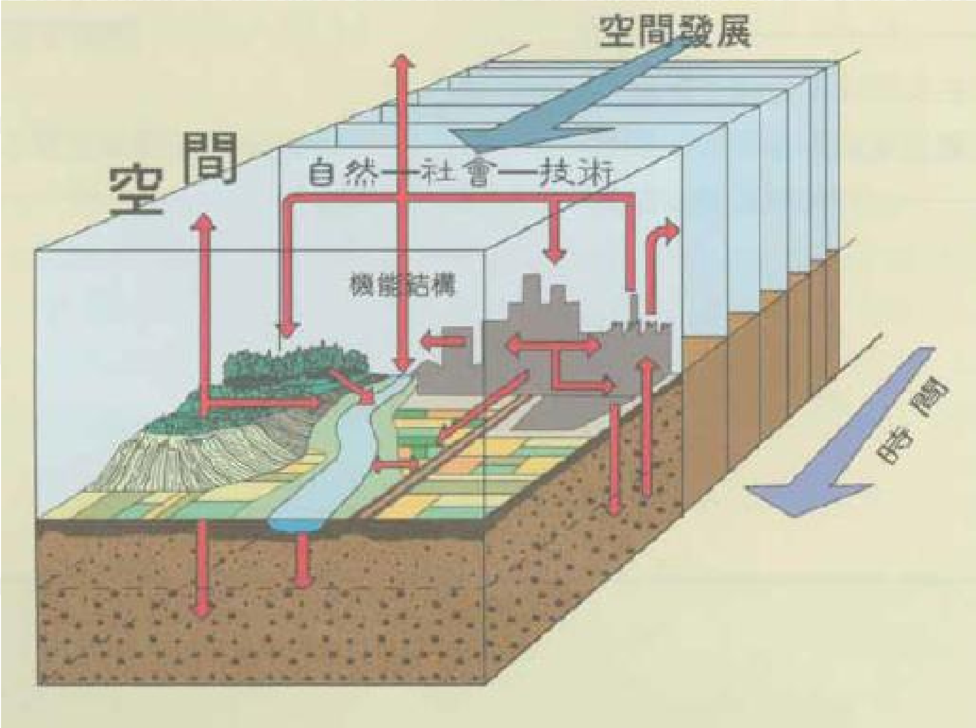
$$F(x)=P(X\le x)\quad(-\infty<x<+\infty)$$

为X的分布函数。

概率的公理化定义

- **非负性公理：** $P(A) \geq 0$;
- **正则性公理：** $P(\Omega) = 1$;
- **可列可加性公理：** 若 $A_1, A_2, \cdots\cdots, A_n \cdots\cdots$

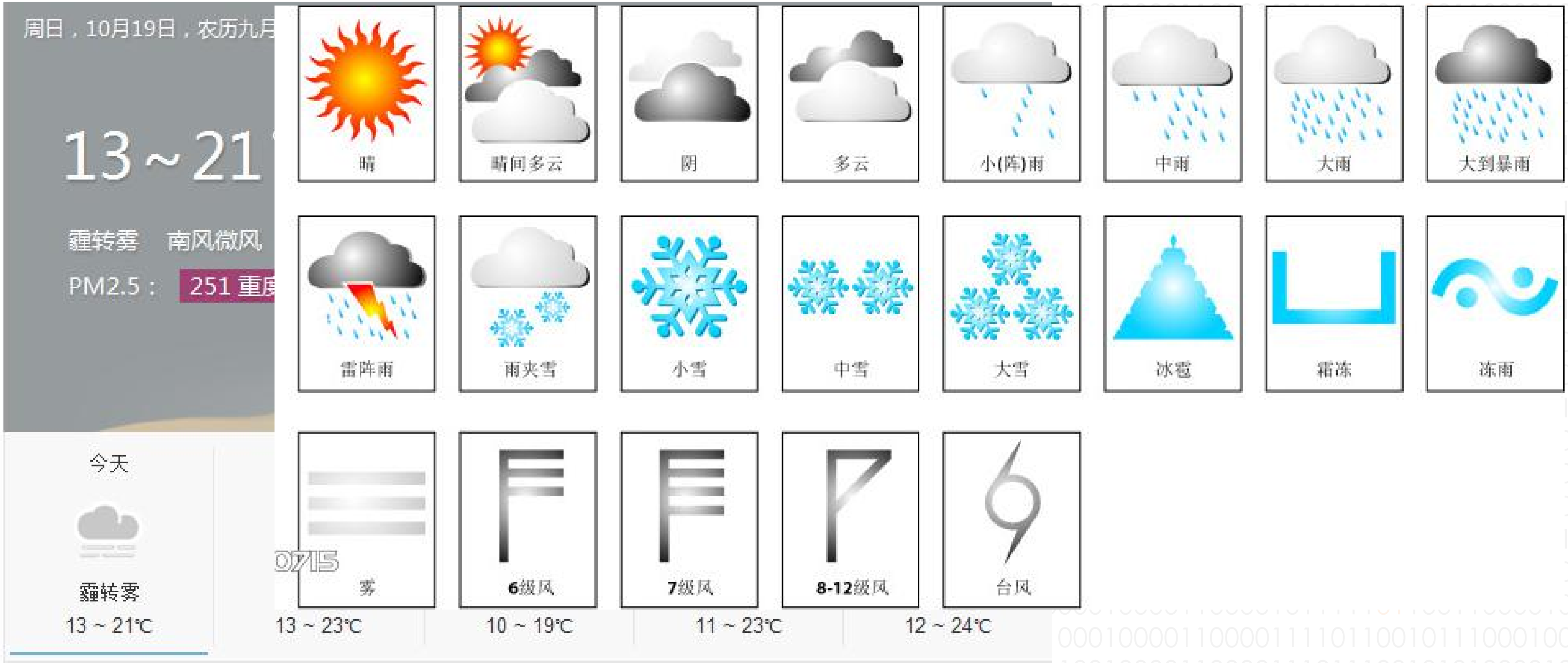
互不相容，则 $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$



1、地理学中的概率函数：

—地理数据离散型分布

随机变量所能取的值可以按一定次序列举，且具有确定的概率。


周日，10月19日，农历九月
13 ~ 21
霾转雾　南风微风
PM2.5：　251 重度
今天
霾转雾
13 ~ 21℃
13 ~ 23℃
10 ~ 19℃
11 ~ 23℃
12 ~ 24℃
晴
晴间多云
阴
多云
小(阵)雨
中雨
大雨
大到暴雨
雷阵雨
雨夹雪
小雪
中雪
大雪
冰雹
霜冻
冻雨
雾
6级风
7级风
8-12级风
台风

离散型随机变量的分布函数

随机变量取值

设可能取的值是xi，相应的概率分别是Pi=P(X=xi)(i=1,2,…,n)

随机事件概率

不同随机事件的概率P写成各事件相应的随机变量X的函数：P=f(x)

设离散型随机变量X的概率分布为pk=P(X=xk)(k=0,1,2, …)，则X的分布函数为

$$F(x)=P(X\le x)=\sum_{x_k\le x} p_k$$

例 2　设随机变量 X 的分布列为

X	-1	2	3
p_k	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

求 X 的分布函数

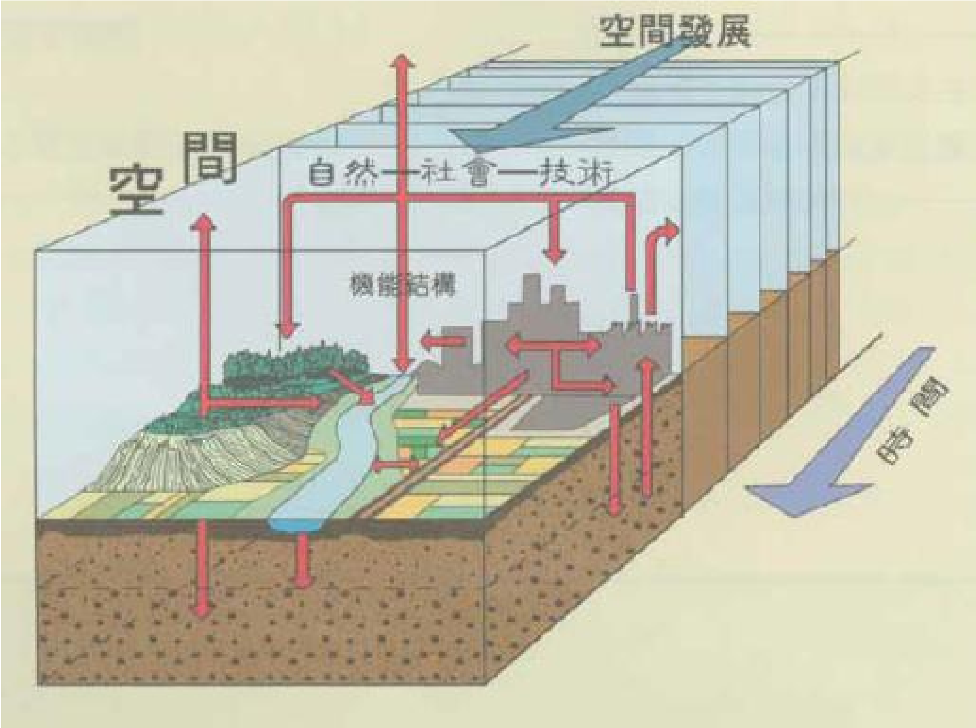
解：

当 $x<-1$ 时，$\{X\le x\}$ 是不可能事件 所以 $F(x)=0$

当 $-1\le x<2$ 时，$F(x)=P(X\le x)=P(X=-1)=\frac{1}{6}$

因此 **X** 的分布函数为

$$F(x)=\begin{cases}0, & x<-1\\ \frac{1}{6}, & -1\le x<2\\ \frac{2}{3}, & 2\le x<3\\ 1, & 3\le x\end{cases}$$



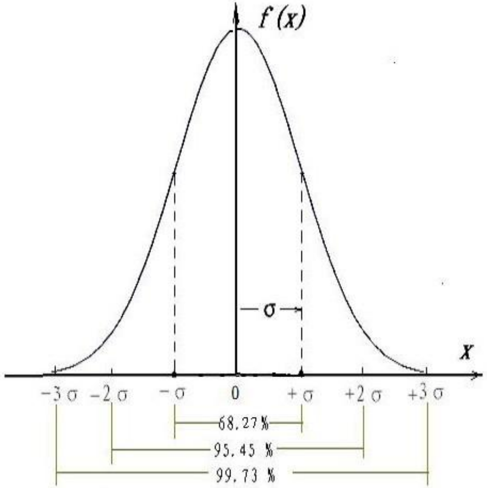
1、地理学中的概率函数：

—地理数据离散型分布

随机变量所能取的值可以按一定次序列举，且具有确定的概率。

—地理数据连续型分布

随机变量所能取的值可以连续地充满一个区间或任何实数。



f(x)
σ
X
-3σ -2σ -σ 0 +σ +2σ +3σ
68.27%
95.45%
99.73%

1.3.2 连续型随机变量的分布函数 $F(x)=P(X\le x)=\int_{-\infty}^{x}p(t)dt$

例 3　　设随机变量 X~U（1，5），求 X 的分布函数

解：随机变量 X 的概率密度为

$$p(x)=\begin{cases}\frac{1}{4}, & 1\le x\le 5\\ 0, & 其它\end{cases}$$

当 x<1 时，p(x)=0，所以 F(x)=0

当 $1\le x\le 5$ 时，$p(x)=\frac{1}{4}$, $F(x)=\int_{-\infty}^{x}p(t)dt=\int_{1}^{x}\frac{1}{4}dt=\frac{x-1}{4}$

当 5<x 时，p(x)=0，所以 F(x)=0+1+0=1

因此 **X** 的分布函数为:

$$F(x)=\begin{cases}0, & x<1\\ \frac{x-1}{4}, & 1\le x\le 5\\ 1, & 5<x\end{cases}$$

2、地理数据的空间分布

（1）地理数据离散型分布

两个重要的分布

二项分布
泊松分布

排列与组合公式

- 从 n 个元素中任取 r 个，求取法数.
- 排列讲次序，组合不讲次序.
- **全排列**：$P_n = n!$
- $0! = 1$.
- **重复排列**：n^r
- **选排列**：

$$P_n^r = \frac{n!}{(n-r)!} = n(n-1)\cdots\cdots(n-r+1)$$

组 合

组合： $C_n^r = \binom{n}{r} = \dfrac{n!}{r!(n-r)!} = \dfrac{P_n^r}{r!}$

加法原理

完成某件事情有 n 类途径，在第一类途径中有m_1种方法，在第二类途径中有m_2种方法，依次类推，在第 n 类途径中有m_n种方法，则完成这件事共有 $m_1+m_2+\ldots+m_n$种不同的方法.

乘法原理

完成某件事情需先后分成 n 个步骤，做第一步有m_1种方法，第二步有 m_2 种方法，依次类推，第 n 步有m_n种方法，则完成这件事共有 $m_1\times m_2\times\ldots\times m_n$种不同的方法.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{k} \text{或} C_n^k$$

$$\binom{1}{0} = \binom{0}{0} = 1$$

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{(3)(2)(1)}{(2)(1)(1)} = 3$$

$$\binom{10}{5} = \frac{10!}{5!(10-5)!} = \frac{(10)(9)(8)(7)(6)5!}{5!(5)(4)(3)(2)(1)} = 252$$

$$(a+b)^n=\binom{n}{0}a^0b^n+\binom{n}{1}a^1b^{n-1}+\binom{n}{2}a^2b^{n-2}+\dots$$

$$+\binom{n}{n-1}a^{n-1}b^1+\binom{n}{n}a^nb^0$$

$$=\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}$$

有一些随机现象，在单次试验观测中所出现的结果只可能有两种，就是说，它的基本事件组中只包含两个基本事件，记为 A 和 B。设它们各自的概率分别为 p 和 q，根据概率归一化条件有

$$p+q=1 \tag{1.2.1}$$

现在，要对这样的一个随机现象的 N 次独立试验结果来做整体的察看。求在这 N 次独立试验序列中有 n_1 次出现事件 A(自然也就是有 N-n1 次出现 B)的概率。

由互不相容事件的"或"的概率加法定理: $P_N(n_1)=\dfrac{N!}{n_1!(N-n_1)!}p^{n_1}q^{N-n_1}$

(1.2.2)

式中因子 $\dfrac{N!}{n_1!(N-n_1)!}$ 是以各种不同次序在 N 次实验中有 n_1 次 A 出现的组合数，通常记为 。

利用二项式定理及(1-2-1)式，不难证明(1-2-2)式给出的概率分布函数满足归一化条件：

$$\sum_{n_1=0}^{N} P_N(n_1) = \sum_{n_1=0}^{N} C_N^{n_1} p^{n_1} q^{N-n_1} = (p+q)^N = 1$$

也正由于概率分布函数 $P_N(n_1)$恰是二项式展开中 P 的第 n_1 次幂的通项，所以这一分布叫做二项式分布。

设某毒理试验采用白鼠共3只，它们有相同的死亡概率 π，相应不死亡概率为 $1-\pi$。

记试验后白鼠死亡的例数为X，分别求X＝0、1、2和3的概率

表 7-1　3 只白鼠各种试验结果及其发生概率

死亡数 X	存活数 $3-X$	试验结果 甲	乙	丙	试验结果的概率	X 取值概率 $P(X)=\binom{3}{k}\pi^k(1-\pi)^{3-k}$
0	3	生	生	生	$(1-\pi)(1-\pi)(1-\pi)$	$P(X=0)=\binom{3}{0}\pi^0(1-\pi)^3$
1	2	死	生	生	$\pi(1-\pi)(1-\pi)$	$P(X=1)=\binom{3}{1}\pi^1(1-\pi)^2$
		生	死	生	$(1-\pi)\pi(1-\pi)$	
		生	生	死	$(1-\pi)(1-\pi)\pi$	
2	1	死	死	生	$\pi\pi(1-\pi)$	$P(X=2)=\binom{3}{2}\pi^2(1-\pi)^1$
		死	生	死	$\pi(1-\pi)\pi$	
		生	死	死	$(1-\pi)\pi\pi$	
3	0	死	死	死	$\pi\pi\pi$	$P(X=3)=\binom{3}{3}\pi^3(1-\pi)^0$

$$P(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

右侧$\binom{n}{k}\pi^k(1-\pi)^{n-k}$为二项式$[\pi+(1-\pi)]^n$展开式的各项

- Poisson(泊松)分布

- 取名于法国数学家 SD Poisson(1781-1840)

泊松分布的概念

- 当二项分布中n很大，p很小时,二项分布就变成为Poisson分布，所以Poisson分布实际上是二项分布的极限分布。
- 由二项分布的概率函数可得到泊松分布的概率函数为：

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k}\approx\frac{(np)^k}{k!}\mathrm{e}^{-np},$$

$$(k=0,1,2,\cdots,n).$$

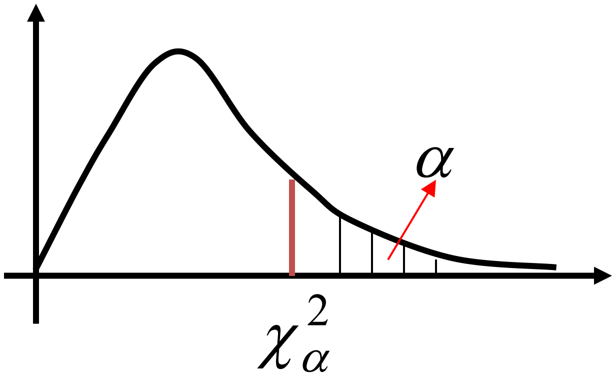
Poisson分布主要用于描述在单位时间(空间)中稀有事件的发生数

例如:

1. 放射性物质在单位时间内的放射次数;
2. 在单位容积充分摇匀的水中的细菌数;
3. 野外单位空间中的某种昆虫数等。

2、地理数据的空间分布

（2）地理数据连续型分布



5个重要的分布

正态分布、标准正态分布
对数正态分布、加马分布
卡方分布

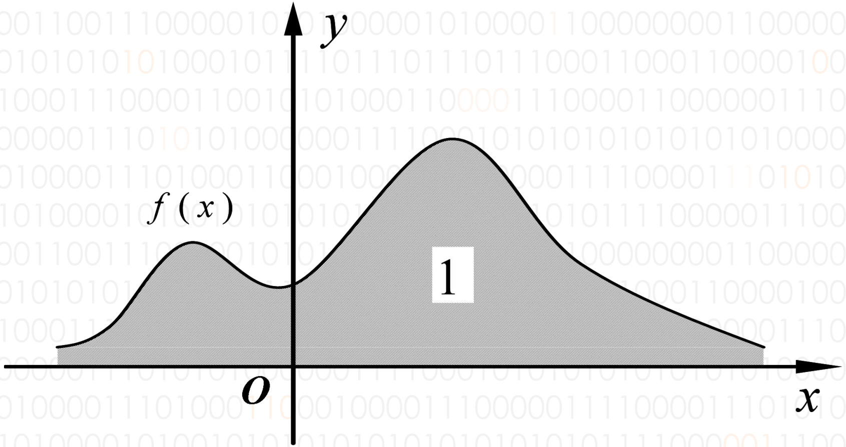
连续型随机变量的定义及其概率密度的性质

定义：设$F(x)$是随机变量X的分布函数，若存在非负可积函数$f(x)$,使得对任意实数x，有

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t$$

称X为连续型随机变量，称$f(x)$为X的概率密度函数，或密度函数，也称概率密度。

性质：1. $f(x) \geq 0$ 2. $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$

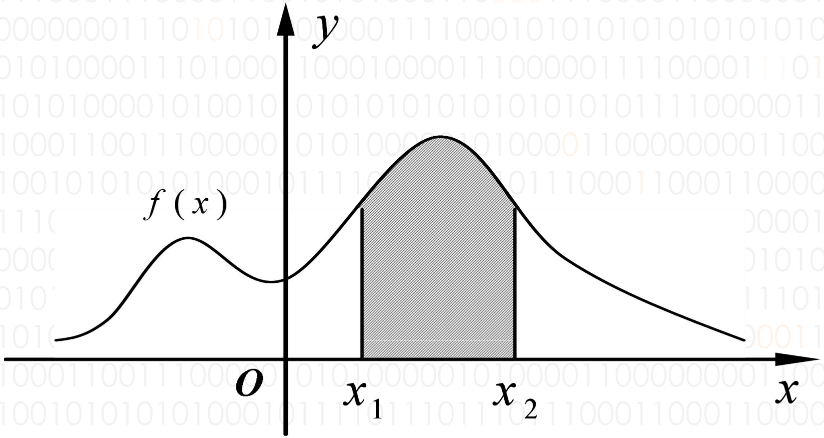


从图形上来看，性质1表示X的概率密度$f(x)$位于x轴上方，

性质2表示$f(x)$与x轴所围区域面积等于1.

3.对于任意实数 $x_1, x_2, (x_1 < x_2)$，有

$$P(x_1 < X \le x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$$



从图形上来看，性质3表示X落在区域 $(x_1, x_2]$ 的概率等于相应的曲边梯形的面积。

4.若$f(x)$在点x处连续，则 $F'(x) = f(x)$

对于连续型随机变量X来说，通过$F(x)$求导得$f(x)$，通过$f(x)$积分得$F(x)$。

5.连续型随机变量取任一指定实数值的概率为零.

即

$$P\{X=x_0\}=0$$

由性质5，易得：

$$P(x_1<X\le x_2)=P(x_1\le X\le x_2)=P(x_1<X<x_2)$$

$$=P(x_1\le X<x_2)=\int_{x_1}^{x_2}f(x)\mathrm{d}x$$

注：对离散型随机变量，上式不成立。

正态分布

定义：若连续型随机变量X的概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<+\infty,$$

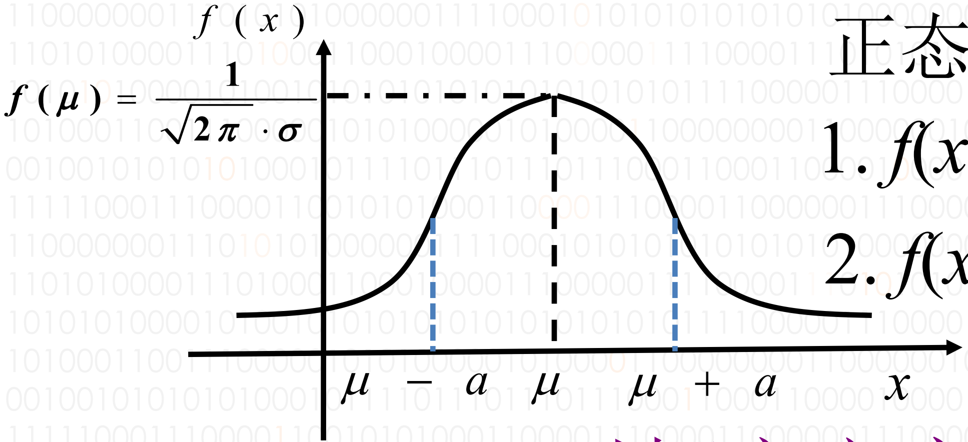
其中μ，$\sigma^2(\sigma>0)$为常数,则称X服从参数为μ和 σ^2的正态分布

记为 $X\sim N(\mu,\sigma^2)$

- 正态分布最早由**Gauss**在研究测量误差时所得到，所以正态分布又称为**Gauss分布**。

•正态分布是概率论中最具有应用价值的分布之一，大量的随机变量都服从正态分布．如人的身高、体重，气体分子向任一方向运动的速度，测量误差等许多随机变量，都服从正态分布．

•大量相互独立且有相同分布的随机变量的累积也近似服从正态分布

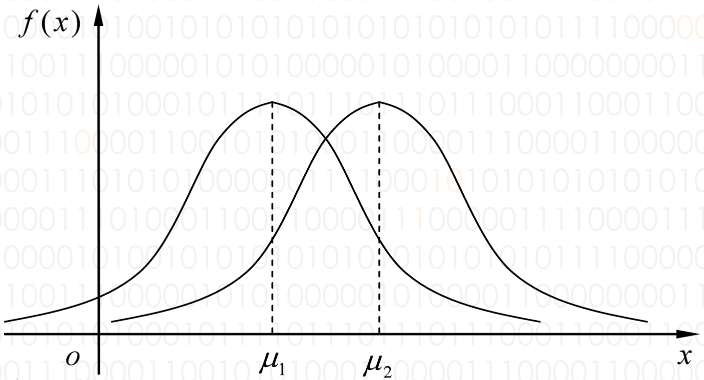


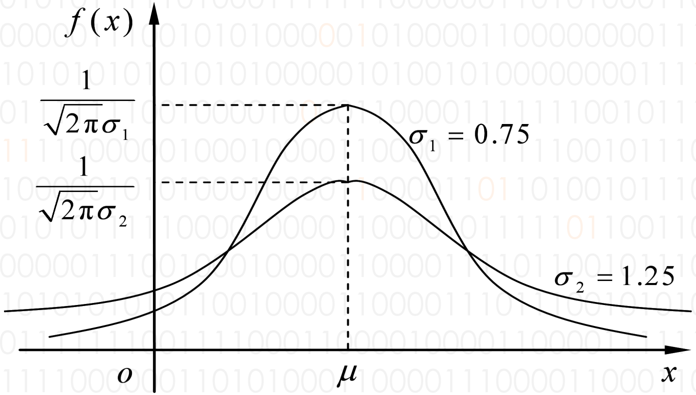
正态分布的图形具有如下特点:

1. $f(x)$为关于$x=\mu$的对称钟形曲线

2. $f(x)$为在$x=\mu$取得最大值

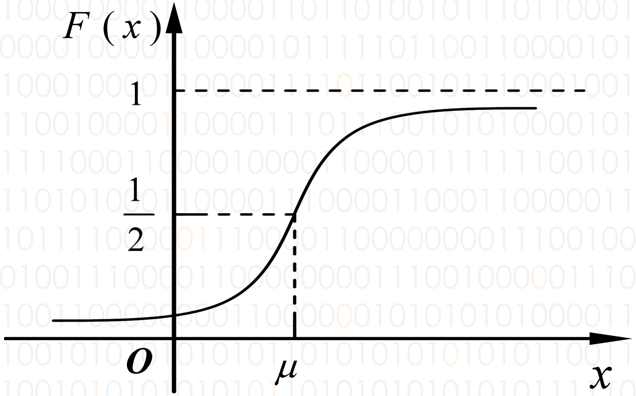
μ，σ对概率密度曲线的影响





正态分布的分布函数:

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$



特别地，当 $\mu=0,\sigma^2=1$ 时，称X服从标准正态分布。

记为 $X\sim N(0,1)$

其概率密度为: $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},-\infty<x<+\infty,$

相应的分布函数记为:

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

一般正态分布的标准化

定理：

如果 $X \sim N(\mu, \sigma^2)$，则 $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

概率计算：

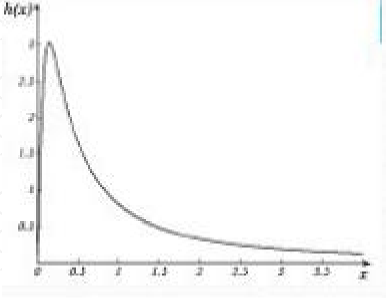
若 $X \sim N(\mu, \sigma^2)$

$$\boxed{P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

查标准正态分布表

对数正态分布

- 如果随机变量X的函数Y=lnX服从正态分布，则称X服从参数为μ 和 σ^2的对数正态分布。



$$f(x)=\begin{cases}\dfrac{1}{\sigma x\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}\left(\dfrac{\ln x-\mu}{\sigma}\right)^2\right] & x>0\\ 0 & x\le 0\end{cases}$$

例： 某零件宽度 $X \sim N(0.9000,\ 0.0030^2)$,现规定限度是 0.9000 ± 0.0050 求零件的废品率。

解： 正品率

$$P\{|X-0.9000| \le 0.0050\}$$
$$= 2\Phi\left(\frac{0.0050}{0.0030}\right) - 1 = 90.44\%$$

故废品率 $= 100\% - 90.44\% = 9.56\%$

Γ－分布（加马分布）

$$p_{\Gamma}(x)=\begin{cases}\dfrac{\lambda^{\alpha}x^{\alpha-1}d^{-\lambda x}}{\Gamma(\alpha)} & x>0\\ 0 & x\le 0\end{cases}$$

其中 $\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx \qquad (\alpha>0,\lambda>0)$

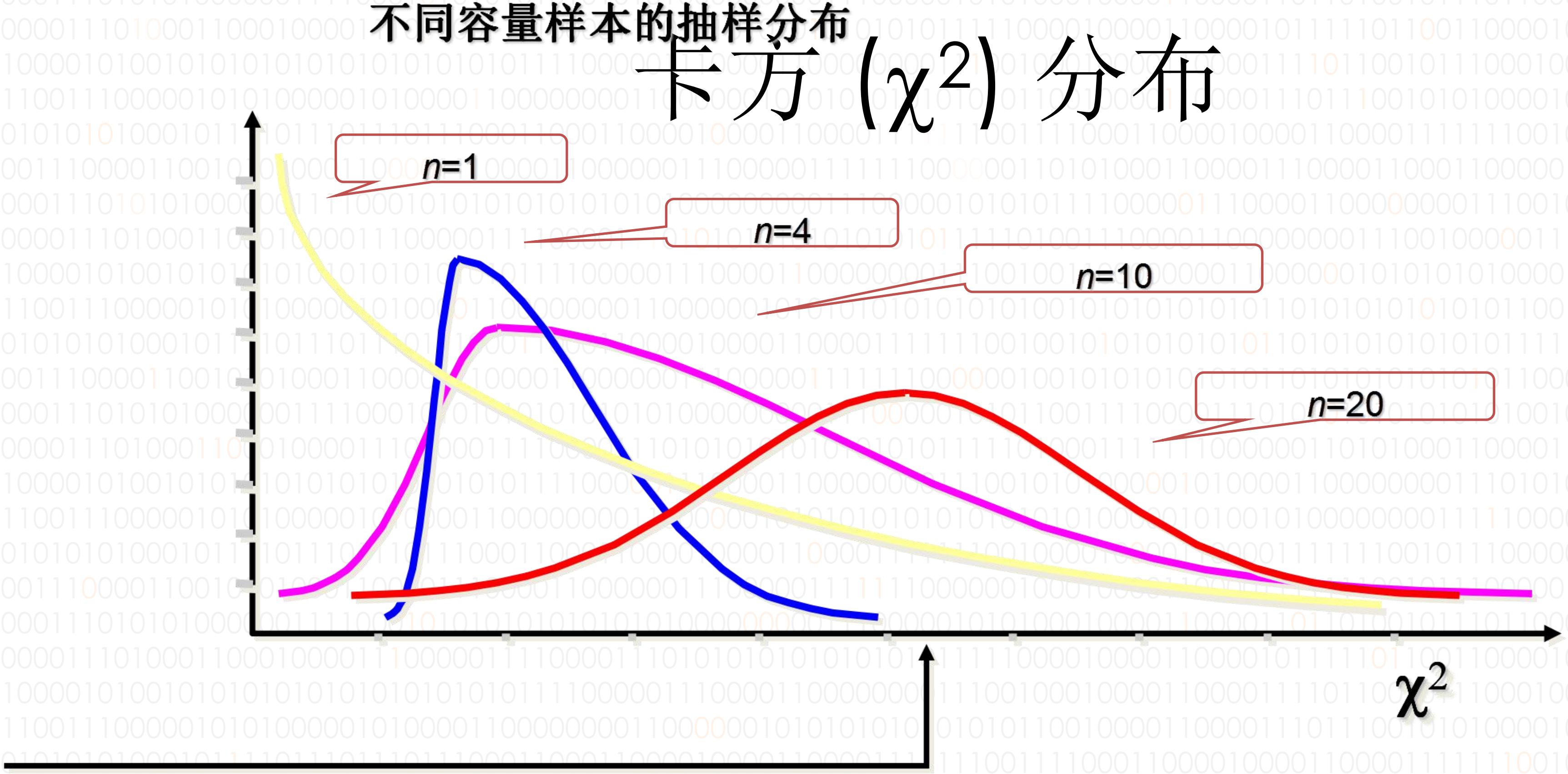
则称ξ服从 Γ－分布，记为 $\xi\sim\Gamma(\lambda,\alpha)$或$G(\lambda,\alpha)$

x^2分布

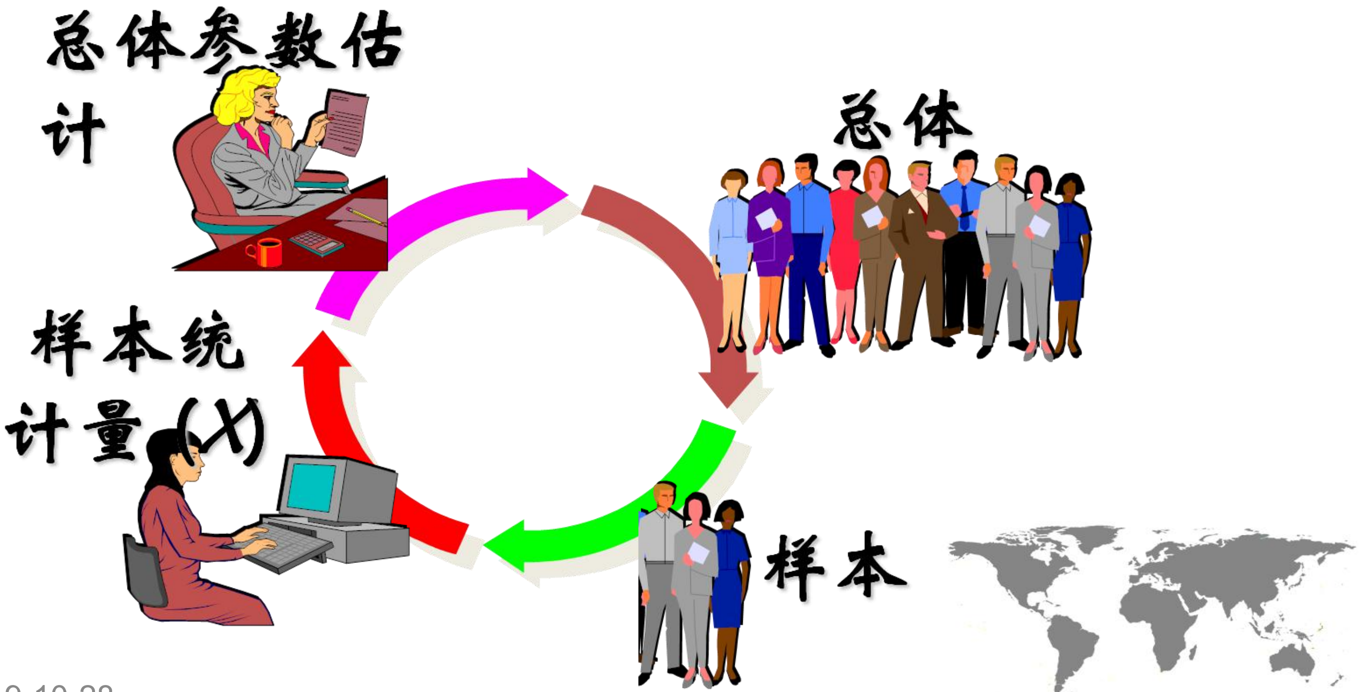
$$p_{x^2}(x)=\begin{cases}\dfrac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} & x>0\\ 0 & x\le 0\end{cases}$$

★ $\Gamma-$分布的一个特例， 其中$\lambda=\dfrac{1}{2},\alpha=\dfrac{n}{2}$($n$为自然数)

卡方 (χ^2) 分布



3、地理学研究中的统计假设检验



3、地理学研究中的统计假设检验

（1）空间类型的抽样设计

（2）三种重要的抽样分布

（3）样本统计假设检验

3、地理学研究中的统计假设检验

（1）空间类型的抽样设计

- 抽样技术
  - 非概率抽样
    - 便利抽样
    - 判断抽样
    - 配额抽样
    - 滚雪球抽样
  - 概率抽样
    - 简单随机抽样
    - 系统抽样
    - 分层抽样
    - 整群抽样

简单随机抽样

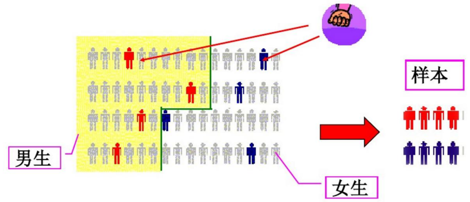
- 在简单随机抽样中，总体中每一个个体都有一个已知且相等的抽中概率
  - 首先确定一个抽样框架，其中的每一个个体被分配了一个唯一的号码
  - 然后产生出随机的数字来确定那些个体被包括进样本中
    - 盲选 **Blind Draw**
    - 随机数表 **the table of random number**
- 优点是易于理解，样本结果可以推断总体，大多数统计推论方法都假定数据是由简单随机抽样法法获得的
- 局限性：抽样框难以构建；数据收集时间和成本高；比其他概率抽样精确度低，标准差较大。

系统抽样

- 在系统抽样中，通过选择一个随机的起点，然后从抽样框架中连续地每隔i个个体选出一个个体，从而选出样本。
- 这种方法成本较低，因为只需要做一次随机抽样
- 可以在不了解抽样框的组成的情况下进行

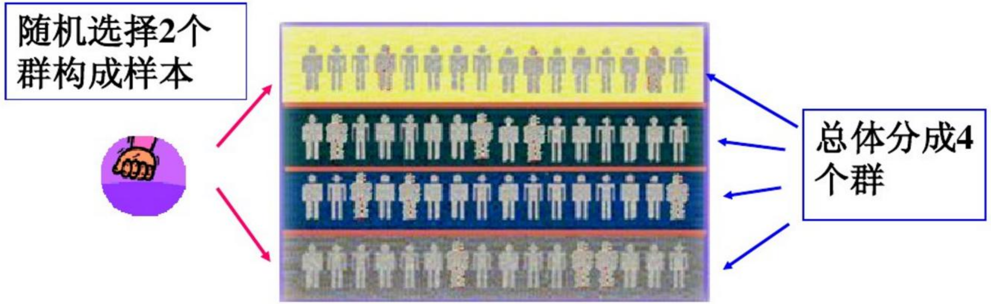
分层抽样

- 分层抽样是一个两阶段过程，总体被分割为子总体，或称为“层”后，再用随机方法，从每一层中选出个体。
  - 各层间应相互独立，并且全体上没有遗漏；分层抽样可以确保子总体在样本中都得以体现。



整群抽样

- 首先将目标总体分为相互排斥且没有遗漏的子总体，或称群，然后根据一种概率抽样技术，选出各群的一个随机样本
  - 可分为单阶段整群抽样与二阶段整群抽样；
  - 与分层抽样的关键差别在于，在整群抽样中，只有一个子总体的样本被选出，而在分层抽样中，为了进一步的抽样，所有的子总体都被选出来了；
  - 整群抽样的目的是通过降低成本来增加抽样效率，分层抽样的目的是增加精确度。
  - 在每个群的中的各体，应尽可能的异质性，但各群本身应尽可能的同质。



3、地理学研究中的统计假设检验

（2）抽样分布

1. χ^2 分布

设有一标准正态变量 z，即 z~n(0, 1)的正态分布，($z_1, z_2, \cdots z_n$)为该分布上的样本值。则其平方和($z_1^2+z_2^2+\ldots+z_n^2$)之统计量，称 χ^2

χ^2 分布具有下列重要性质：

（1）当 n 大于 30 时，可使用正态分布进行变换

（2）设 χ_1^2 与 χ_2^2 为独立随机变量，并且是自由度为 n_1 与 n_2 的 χ^2 分布，则 $\chi^2=\chi^2+\chi^2$

亦为自由度是(n_1+n_2)的 χ^2 分布；

（3）统计量可表示为：$\chi^2=\sum_{j=1}^{k}\frac{(f_j-F_j)^2}{F_j}$

3、地理学研究中的统计假设检验

（2）抽样分布

2.t 分布

设随机变量 ξ 与 η 相互独立，且 ξ 服从 N(0,1)分布，而 $\eta=\sqrt{\frac{x^2}{n}}$（$x^2$ 是服从自由度为 n 的 x^2 分布随机变量。则随机变量

$$t=\frac{\xi}{\eta}=\frac{\xi}{\sqrt{\frac{x^2}{n}}}=\frac{\xi}{\frac{x}{\sqrt{n}}}$$

其密度函数为

$$p_t(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

称此式中的t为密度函数的分布为具有自由度为的分布（或“学生”分布）

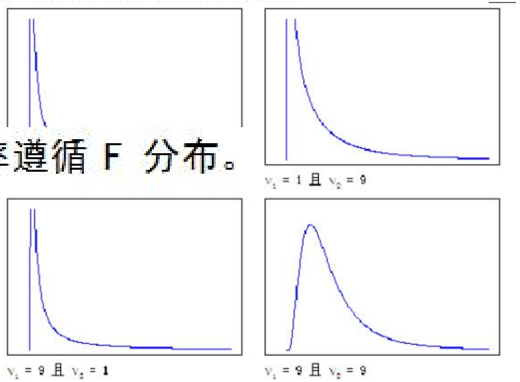
3、地理学研究中的统计假设检验

（2）抽样分布

F 分布

在假设检验中用于确定两个总体方差是否相等。F 分布是两个独立的带有卡方分布的随机变量的抽样分布，每个变量被其自由度所除。F 分布也称为 Snedecor 的 F 分布和 Fisher-Snedecor 分布。

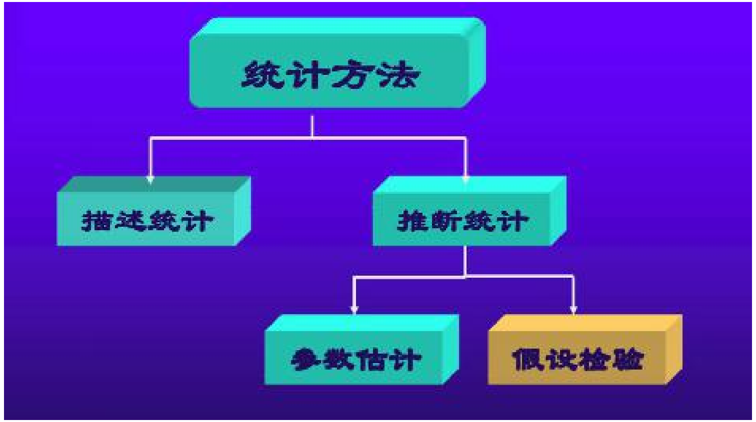
F 分布常用于方差分析中。组间变异性与组内变异性的比率遵循 F 分布。



3、地理学研究中的统计假设检验

（3）假设检验

3、地理学研究中的统计假设检验



假设检验在统计方法中的地位

3、地理学研究中的统计假设检验

假设检验的基本问题

- 假设的陈述
- 两类错误与显著性水平
- 统计量与拒绝域

3、地理学研究中的统计假设检验

（1）假设的陈述

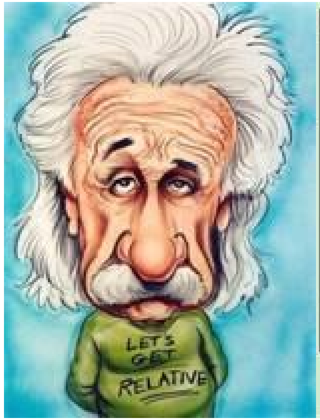


图5　战国城市分布图

对总体参数所做的陈述

总体参数包括：均值、方差等

（分析之前必须陈述）

3、地理学研究中的统计假设检验

（3.1）假设的陈述

什么是假设检验?

先对总体参数或分布形式提出假设，然后利用样本判断是否成立的过程

3、地理学研究中的统计假设检验

（3.1）假设的陈述

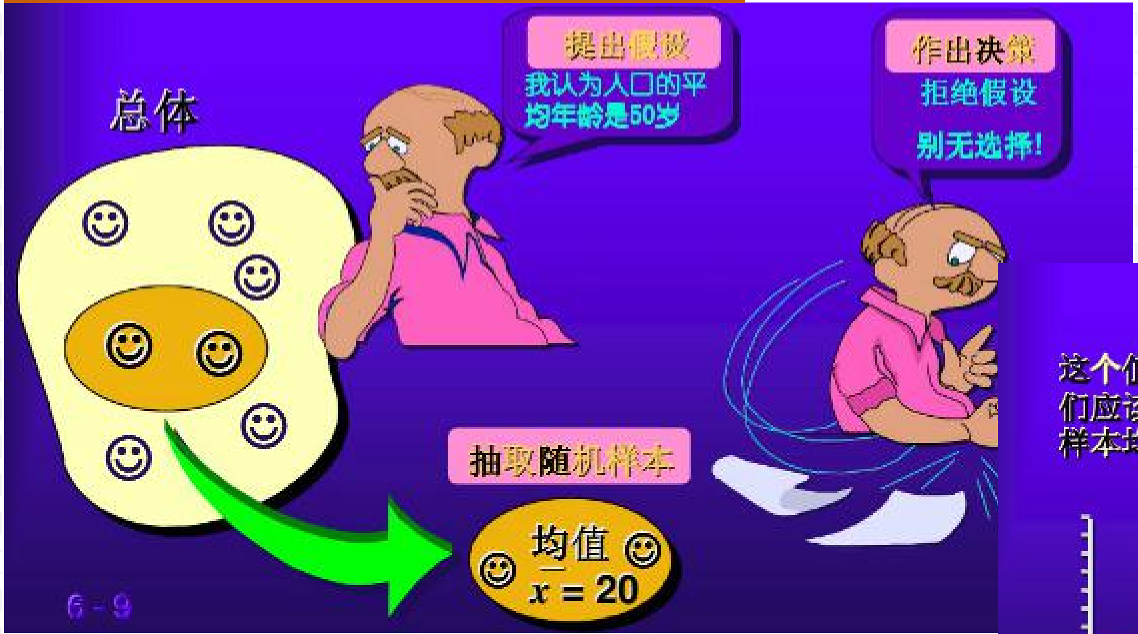
什么是假设检验?

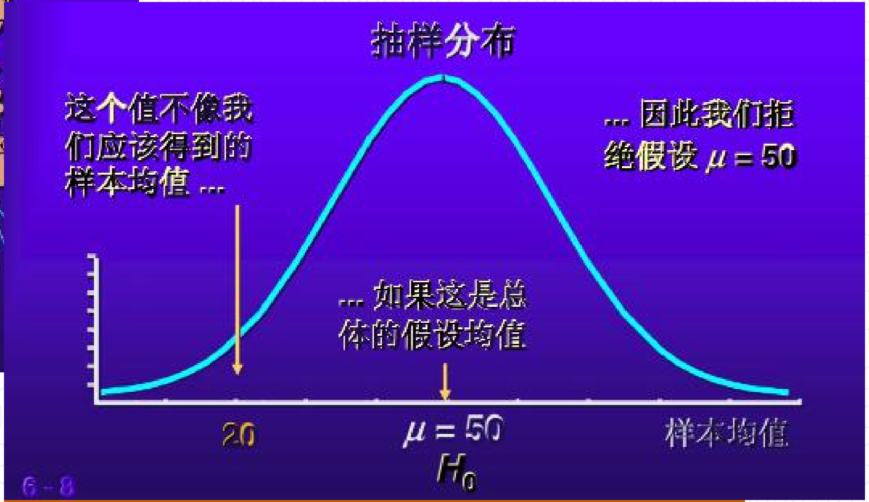
逻辑上运用反证法，统计上利用小概率原理

"一个小概率事件在一次实验中几乎是不可能发生的"

小概率指$p<5\%$。

假设检验的过程





假设检验的基本思想

假设检验的基本概念：

原假设和备择假设

- **原假设**：用H0表示，即虚无假设、零假设、无差异假设（研究者想收集证据予以反对的假设，总有符号=，≤，≥）；
  备择假设：用H1表示，是原假设被拒绝后替换的假设（研究者想收集证据予以支持的假设，总有符号≠，<，>）。

- 若证明为H0为真，则H1为假；H0为假，则H1为真。

- 对于任何一个假设检验问题**所有可能的结果**都应包含在**两个假设**之内，非此即彼。

假设检验的基本概念：

检验统计量

- 用于假设检验问题的统计量称为检验统计量。

- 与参数估计相同，需要考虑：

  总体是否正态分布；

  大样本还是小样本；

  总体方差已知还是未知。

假设检验的基本概念：

显著性水平

- 用样本推断H_0是否正确，必有犯错误的可能。
  原假设H_0正确，而被我们拒绝，犯这种错误的概率用α表示。把α称为假设检验中的显著性水平(Significant level), 即决策中的风险。
- 显著性水平就是指当原假设正确时人们却把它拒绝了的概率或风险。
- 通常取α = 0.05或α=0.01或α=0.001, 那么, 接受原假设时正确的可能性(概率)为:95%, 99%, 99.9%。

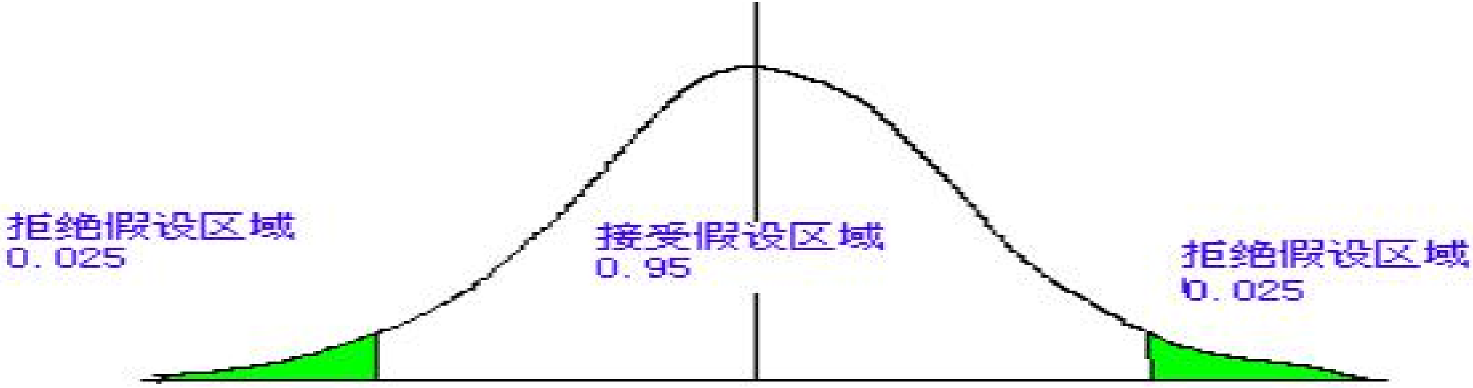
假设检验的基本概念：

接受域与拒绝域

- 接受域：原假设为真时允许范围内的变动，应该接受原假设。

- 拒绝域：当原假设为真时只有很小的概率出现，因而当统计量的结果落入这一区域便应拒绝原假设，这一区域便称作拒绝域。

例：α = 0.05时的接受域和拒绝域



假设检验的基本概念：

双侧检验与单侧检验

假设检验根据实际的需要可以分为：

双侧检验（双尾）：指只强调差异而不强调方向性的检验。

$$H_0: \mu_1 = \mu_0$$

$$H_1: \mu_1 \neq \mu_0$$

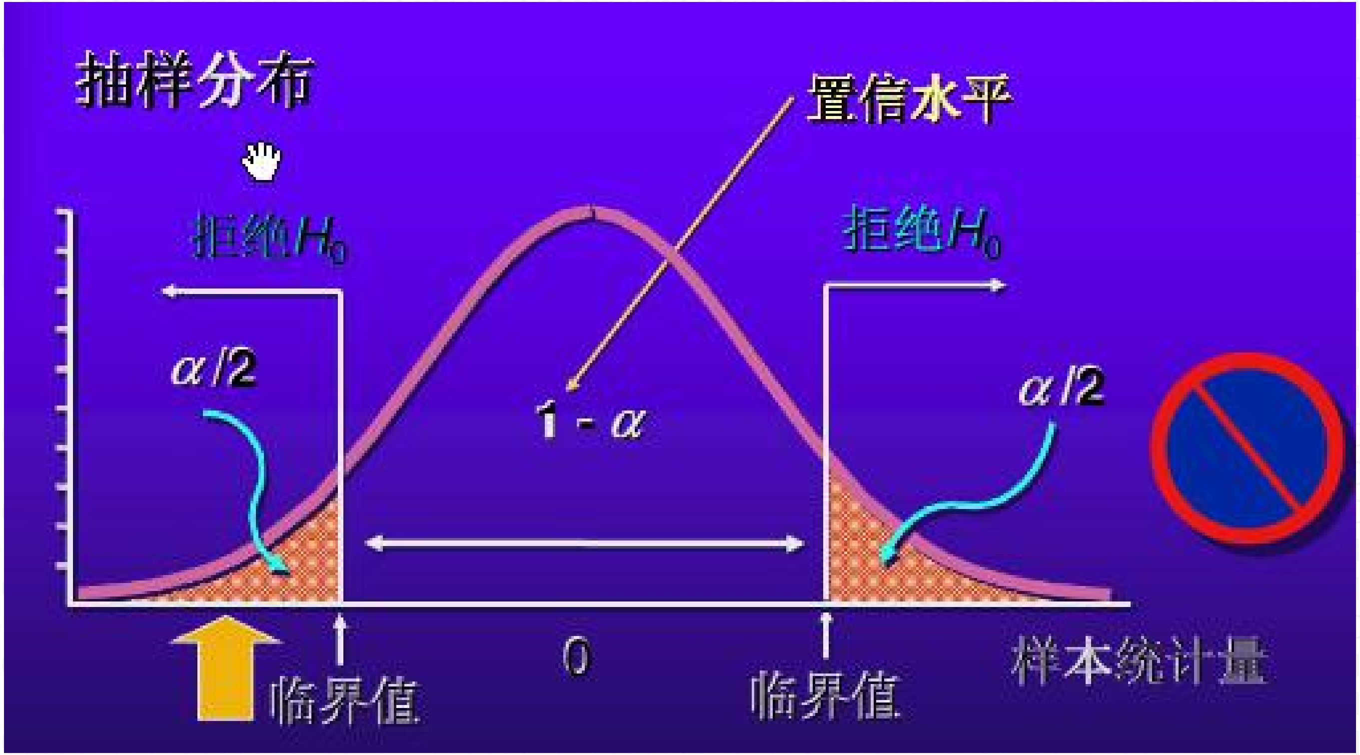
只关注 μ_1，μ_0是否有差异，不关心 μ_1比μ_0大还是小

单侧检验（单尾）：强调某一方向性的检验。

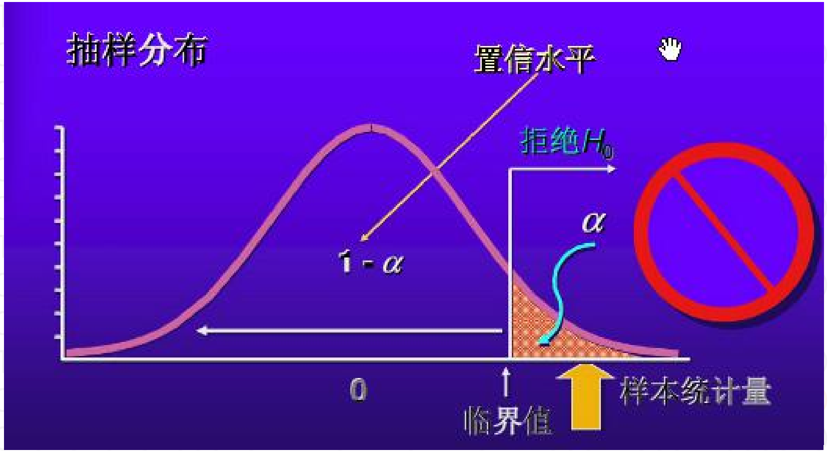
左侧检验 $\begin{cases} H_0: \mu_1 \geq \mu_0 \\ H_1: \mu_1 < \mu_0 \end{cases}$

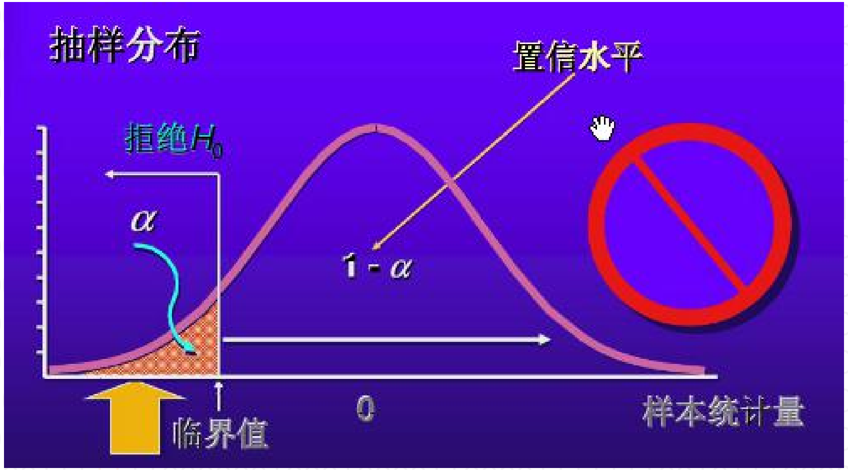
右侧检验 $\begin{cases} H_0: \mu_1 \leq \mu \\ H_1: \mu_1 > \mu \end{cases}$

双侧检验示意图



假设检验中的单侧检验示意图





拒绝域 拒绝域

(a)右侧检验 (b)左侧检验

假设检验的基本概念：

假设检验中的两类错误

- 假设检验是依据样本提供的信息进行推断的,即由部分来推断总体,因而假设检验不可能绝对准确,是可能犯错误的。

两类错误:

- α错误(I型错误): H_0为真时却被拒绝,弃真错误;
- β错误(II型错误): H_0为假时却被接受,取伪错误。

假设检验中各种可能结果的概率:

	接受H_0,拒绝H_1	拒绝H_0，接受H_1
H_0为真	$1-\alpha$(正确决策)	α(弃真错误)
H_0为伪	β(取伪错误)	$1-\beta$(正确决策)

假设检验的步骤

- 建立原假设和备择假设;
- 确定适当的检验统计量,使得在原假设成立时，其分布已知;
- 指定检验中的显著性水平;
- 利用显著性水平根据检验统计量的值建立拒绝原假设的规则;
- 搜集样本数据,计算检验统计量的值;
- 作出统计决策:(两种方法)

  1) 将检验统计量的值与拒绝规则所指定的临界值相比较,确定是否拒绝原假设;

  2)由步骤5的检验统计量计算p值,利用p值确定是否拒绝原假设。

两个区域方差的比较：

两个独立样本正态总体方差显著检验

通过比较两个样本方差，从而判断两总体方差是否相等的问题，即 $\sigma_1^2 = \sigma_2^2$ 。自然地，应用它们的估计量 s_1 和 s_2 的比值来进行判断。如果比值远大于1或远小于1，说明 s_1 和 s_2 之值相差甚大。

为了要具体明确“远大于1或小于1”的数值及其意义，就要研究统计量

$$F = \frac{s_1^2}{s_2^2}$$

的分布。可以证明，在原假设成立的条件下，

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

即服从第一自由度为 n_1-1,第二自由度为 n_2-1 的F 分布。（p73）

两个区域均值（平均数）的比较：

两个独立样本，正态，大样本

假设	H_0： $m_1=m_2$	
统计量	已知$\sigma^2=\sigma_1^2=\sigma_2^2$	u检验法， $u=\dfrac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim N(0,1)$
统计量	总体均方差未知	t分布检验， $t=\dfrac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t$分布(自由度为$n_1+n_2-2$)

$$\sigma=\sqrt{\frac{(n_1-1)s_1^{*2}+(n_2-1)s_2^{*2}}{n_1+n_2-2}}=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}$$

两个以上区域均值（平均数）的比较：

两个以上独立样本，正态，大样本

表 4—7　方差分析表

方差来源	平方和 S	自由度 f	平均离差平方和 $\overline{S}$	F 值	显著性
组间	$S_A = k\sum_{i=1}^{n}(\overline{x}_i - \overline{x})^2$	n-1	$\overline{S}_A = \frac{S_A}{n-1}$	$\frac{\overline{S}_A}{\overline{S}_e}$	
组内	$S_e = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \overline{x}_i)^2$	$n(k-1)$	$\overline{S}_e = \frac{S_e}{n(k-1)}$		
总和	$S = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \overline{x})^2$	$nk-1$			

适合性检验：

事先对总体分布一无所知

设母体遵从的分布函数为F（x），取自母体的样本为x1,x2,···xn，现在要用此组样本来检验假设：

H0 :F(x)=F0(x)　　这里是某个给定的分布函数。

当我们利用子样对总体分布进行检验时，自然想到利用区间中频率和概率的差异来构造统计量。

适合性检验：

事先对总体分布一无所知

具体作法是：

（1） 把样本值 $x_1,x_2,\ldots x_n$ 出现的范围，划分为 k 段：$[b_0,b_1),[b_1,b_2),\ldots[b_{k-1},b_k)$，

其中 $-\infty=b_0<b_1<\ldots<b_{k-1}<b_k=+\infty$，且每段内至少含有 5 个以上的样本（k 个区间可为等分，亦可不等分）。

（2） 求出每个区间 $[b_{i-1},b_i)$ 内的频数，并求出各段中的频率 $f=\dfrac{n_i}{n}$

（3） 算出 $p_i=p(b_{i-1}\le\xi_i<b_i)=F_0(b_i)-F_0(b_{i-1})$，它表示当 H_0 为真时，ξ_i 出现在 $[b_{i-1},b_i)$ 中的概率。

适合性检验：

事先对总体分布一无所知

（4）　利用 ξ_i 落入区间 $[b_{i-1}, b_i)(i=0,1,\dots,k)$ 中频率与概率之差 $\frac{n_i}{n}-p_i$ 来代表第 i 个区间上频率直方图与概率密度曲线的偏差，并构造统计量

$$\chi^2=\sum_{i=1}^{k}\left(\frac{n_i}{n}-p_i\right)^2\frac{n}{p_i}=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$$

其中：k 为地理数据分组的组数，

$n_1,n_2,\dots,n_k$ 为每一组实际观测次数，i=1,2,…,k,n 为观测总次数 $n=\sum_{i=1}^{k}n_i$，

p_i 为各组理论频率，，可用汉字写成：$\chi^2=\sum_{组}\frac{(观测次数-理论次数)^2}{理论次数}$

适合性检验：

事先对总体分布一无所知

就近似的服从自由度为 k-1 的 χ^2 分布。如果 $F_0(x)$中有 r 个参数，它是通过子样估计出来的

这时 χ^2 分布的自由度为 k-r-1。

（5）对于给定的信度α，可由 χ^2 分布按自由度 k-1 查出置信限 $\chi_\alpha^2(k-1)$再由样本

按（4—24）式算出 χ^2 值

当时 $\chi^2 \geqslant \chi_\alpha^2$，则拒绝原假设 H_0：

当时 $\chi^2 < \chi_\alpha^2$，则接受原假设》H_0

?