概率论与数理统计
第六章 数理统计的基本概念

金玲飞

复旦大学计算机学院
Email: lfjin@fudan.edu.cn

2019.12.10

统计学：是一门关于确定性和带随机性数据资料的收集，整理，分析和推断的科学。

按是否使用概率分为

- 描述统计学：运用图表，表格等方法
- 推断统计学：运用概率论和数学的方法。即数理统计学。

应用广泛：社会，经济，医学，生物，气象等等。

在终极的分析中，一切知识都是历史。
在抽象的意义下，一切科学都是数学。
在理性的世界里，所有的判断都是统计学。

---C.R.劳

《统计与真理-怎样运用偶然性》

数理统计

数理统计：使用概率论和数学的方法，研究

1. 如何有效的收集带有随机性的数据；
2. 如何分析数据；
3. 如何在给定的模型下，进行统计推断。

数理统计

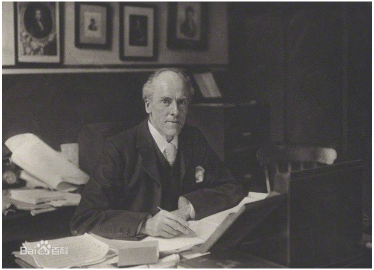
数理统计：使用概率论和数学的方法，研究

1. 如何有效的收集带有随机性的数据；
2. 如何分析数据；
3. 如何在给定的模型下，进行统计推断。

由于抽取的部分具有一定的随机性，因此得到的推断有一定的不确定性。我们必须对信息进行加工，使得推断出错的概率尽可能小。

Karl Pearson， 1857～1936

- 生于伦敦。是英国数学家，生物统计学家，数理统计学的创立者，对生物统计学、气象学、社会达尔文主义理论和优生学做出了重大贡献。



Ronald Aylmer Fisher 1890- 1962

- 生于伦敦, 卒于澳大利亚
- 英国统计学家、生物进化学家、数学家、遗传学家和优生学家。
- 现代统计科学的奠基人之一。主要贡献包括方差分析，极大似然统计推断和许多抽样分布的导出。
- 他是达尔文以来最伟大的生物进化学家。

目录

- 总体与样本，
- 统计量
- χ^2分布，t分布，F分布。
- 正态总体的抽样分布。

总体和样本

总体: 研究对象的全体。
所研究的对象的某个或某几个数量指标的全体，是一个具有确定分布的（一维或多维）随机变量，记为X。

X的分布函数和数字特征称为总体的分布函数和数字特征。“总体X”，“总体F(x)”。

总体和样本

总体: 研究对象的全体。
所研究的对象的某个或某几个数量指标的全体，是一个具有确定分布的（一维或多维）随机变量，记为X。

X的分布函数和数字特征称为总体的分布函数和数字特征。“总体X”，“总体F(x)”。

个体：组成总体的每一个元素，即该随机变量的一个可能的取值。

总体和样本

总体: 研究对象的全体。
所研究的对象的某个或某几个数量指标的全体，是一个具有确定分布的（一维或多维）随机变量，记为X。

X的分布函数和数字特征称为总体的分布函数和数字特征。“总体X”，“总体F(x)”。

个体：组成总体的每一个元素，即该随机变量的一个可能的取值。

样本：从总体中抽出的部分个体。

在数理统计中，总体X的分布函数$F(x)$总是未知，统计推断的主要任务之一是确定总体的分布，为此必须从总体中抽取一部分个体进行试验，推断总体$F(x)$ 的具体形式。

在数理统计中，总体X的分布函数$F(x)$总是未知，统计推断的主要任务之一是确定总体的分布，为此必须从总体中抽取一部分个体进行试验，推断总体$F(x)$ 的具体形式。

简单随机抽样的两个特点

- 随机性：每个个体被抽中的机会是均等的
- 独立性：抽取一个个体后不影响总体

在数理统计中，总体X的分布函数$F(x)$总是未知，统计推断的主要任务之一是确定总体的分布，为此必须从总体中抽取一部分个体进行试验，推断总体$F(x)$ 的具体形式。

简单随机抽样的两个特点

- 随机性：每个个体被抽中的机会是均等的
- 独立性：抽取一个个体后不影响总体

样本$X_1, \ldots, X_n$独立且与总体X有相同的分布

- Ex: 有放回的抽样，不放回抽样（总体相对抽样数很大）

定义 (6.1.1 简单随机样本)

设$X_1, \ldots, X_n$是n个相互独立的随机变量，若其中每个都与总体X具有相同的分布，则称$X_1, \ldots, X_n$是来自总体X的容量为n的简单随机样本，简称样本 *Sample*。

定义 (6.1.1 简单随机样本)

设$X_1,\ldots,X_n$是n个相互独立的随机变量，若其中每个都与总体X具有相同的分布，则称$X_1,\ldots,X_n$是来自总体X的容量为n的简单随机样本，简称样本 *Sample*。

- 抽样前，样本是随机变量 $X_1,\ldots,X_n$。
- 抽取进行试验后，是数 $x_1,\ldots,x_n$，即样本观察值。

定义 (6.1.1 简单随机样本)

设$X_1, \ldots, X_n$是n个相互独立的随机变量，若其中每个都与总体X具有相同的分布，则称$X_1, \ldots, X_n$是来自总体X的容量为n的简单随机样本，简称样本 *Sample*。

- 抽样前，样本是随机变量 $X_1, \ldots, X_n$。
- 抽取进行试验后，是数 $x_1, \ldots, x_n$，即样本观察值。

数理统计的基本任务是利用样本对总体的未知分布（或者分布的某些特征）进行统计推断。

统计量

样本来自总体，包含总体的信息，需要对样本进行加工，提取有用的信息做推断。一种有效的方法是构造样本函数。

定义 (6.2.1 统计量 Statistic)

设$X_1, \ldots, X_n$是来自总体X的样本，$g(x_1, \ldots, x_n)$是n元连续函数且不含任何未知参数，则称$g(X_1, \ldots, X_n)$是统计量。若$x_1, \ldots, x_n$是样本观察值，则称$g(x_1, \ldots, x_n)$是该统计量的观察值。

统计量

样本来自总体，包含总体的信息，需要对样本进行加工，提取有用的信息做推断。一种有效的方法是构造样本函数。

定义 (6.2.1 统计量 Statistic)

设$X_1, \ldots, X_n$是来自总体X的样本，$g(x_1, \ldots, x_n)$是n元连续函数且不含任何未知参数，则称$g(X_1, \ldots, X_n)$是统计量。若$x_1, \ldots, x_n$是样本观察值，则称$g(x_1, \ldots, x_n)$是该统计量的观察值。

- 统计量不含任何未知参数。如$\frac{1}{\sigma}\sum_{i=1}^{n}(X_i - \mu)^2$。

常用的统计量

设$X_1, X_2, \ldots, X_n$是来自总体X的样本，$x_1, x_2, \ldots, x_n$是样本观察值，常用统计量

- 样本平均值（样本均值）：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

常用的统计量

设$X_1, X_2, \ldots, X_n$是来自总体X的样本，$x_1, x_2, \ldots, x_n$是样本观察值，常用统计量

- 样本平均值（样本均值）：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- 样本标准差：$S = \sqrt{S^2}$

常用的统计量

设$X_1, X_2, \ldots, X_n$是来自总体X的样本，$x_1, x_2, \ldots, x_n$是样本观察值，常用统计量

- 样本平均值（样本均值）：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- 样本标准差：$S = \sqrt{S^2}$
- 样本k阶矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
- 样本k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

样本方差S^2与样本二阶中心矩S_n^2

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

样本方差S^2与样本二阶中心矩S_n^2

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

定理 (6.2.1)

设总体X的数学期望和方差存在，并设$E(X) = \mu$，$D(X) = \sigma^2$。若$X_1, \ldots, X_n$是来自总体X的样本，则有

$$E(\overline{X}) = \mu, D(\overline{X}) = \frac{\sigma^2}{n}, E(S^2) = \sigma^2$$

- 样本k阶矩依概率收敛到相应的总体k阶矩。
- 样本方差S^2依概率收敛到总体的方差σ^2。

定义 (6.2.3 顺序统计量 Order statistic)

设$X_1,\dots,X_n$是来自总体X的样本，由样本建立n个函数：

$$X_{(k)} = X_{(k)}(X_1,\dots,X_n), k = 1,\dots,n$$

其中$X_{(k)}$是这样的统计量：当样本观察值为$x_1,\dots,x_n$时，将$x_1,\dots,x_n$从小到大排成 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，则第$i$个值$x_{(i)}$就是$X_{(i)}$ 的观察值。称$X_{(1)},\dots,X_{(n)}$是样本$X_1,\dots,X_n$的顺序统计量，称$X_{(k)}$是样本$X_1,\dots,X_n$的第k位顺序统计量。

定义 (6.2.3 顺序统计量 Order statistic)

设$X_1,\ldots,X_n$是来自总体X的样本，由样本建立n个函数：

$$X_{(k)} = X_{(k)}(X_1,\ldots,X_n), k = 1,\ldots,n$$

其中$X_{(k)}$是这样的统计量：当样本观察值为$x_1,\ldots,x_n$时，将$x_1,\ldots,x_n$从小到大排成 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，则第$i$个值$x_{(i)}$就是$X_{(i)}$ 的观察值。称$X_{(1)},\ldots,X_{(n)}$是样本$X_1,\ldots,X_n$的顺序统计量，称$X_{(k)}$是样本$X_1,\ldots,X_n$的第k位顺序统计量。

- 最小顺序统计量：$X_{(1)} = \min\{X_1,\ldots,X_n\}$
- 最大顺序统计量：$X_{(n)} = \max\{X_1,\ldots,X_n\}$
- 极差：$R = X_{(n)} - X_{(1)}$

- 在简单随机样本中，$X_1, \cdots, X_n$是独立同分布的，而次序统计量即不独立，也不同分布。

例子

考虑X取值仅为*0*，*1*，*2*的离散均匀分布。现从中抽取容量为*3*的样本，则一切可能取值为$3^3 = 27$种。则可得到$X_{(1)}, X_{(2)}, X_{(3)}$的分布列。

正态总体的抽样分布

统计量的分布，称为抽样分布(sampling distribution)。
若X服从正态分布，就称为正态总体。

定义 (6.2.4 χ^2分布)

设$X_1, \ldots, X_n$相互独立且均服从标准正态分布$N(0,1)$，称随机变量

$$\chi^2 = X_1^2 + \ldots + X_n^2$$

所服从的分布是自由度为n的χ^2分布，记为$\chi^2 \sim \chi^2(n)$。

χ^2分布的概率密度函数

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt, \Gamma(1) = 1, \Gamma(1/2) = \sqrt{\pi}$.

- χ^2分布是一个非负值的偏态分布。
- $n = 2$时，是指数分布。

性质一：设$\chi^2 \sim \chi^2(n)$，则$E(\chi^2) = n$，$D(\chi^2) = 2n$。

性质一：设$\chi^2 \sim \chi^2(n)$，则$E(\chi^2) = n$，$D(\chi^2) = 2n$。

性质二：可加性 若$\chi_1^2 \sim \chi^2(n), \chi_2^2 \sim \chi^2(m)$，且$\chi_1^2, \chi_2^2$独立，则$\chi_1^2 + \chi_2^2 \sim \chi^2(n+m)$。

例子 (6.2.1)

设总体$X \sim N(0,1)$，$X_1, \ldots, X_6$是来自总体X的样本，又设

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2$$

试确定C，使得CY服从χ^2分布。

例子 (6.2.1)

设总体$X \sim N(0,1)$，$X_1, \ldots, X_6$是来自总体X的样本，又设

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2$$

试确定C，使得CY服从χ^2分布。

$$X_1 + X_2 + X_3 \sim N(0,3), X_4 + X_5 + X_6 \sim N(0,3).$$

t分布(Student distribution)

t分布的发现具有划时代意义，打破了正态分布一统天下的局面，开创了小样本统计推断的新纪元。

定义 (6.2.5)

设随机变量$X \sim N(0,1)$， $Y \sim \chi^2(n)$，并且X和Y独立，则称随机变量

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

所服从的分布是自由度为n的t分布，记为$T \sim t(n)$。

t分布(Student distribution)

t分布的发现具有划时代意义，打破了正态分布一统天下的局面，开创了小样本统计推断的新纪元。

定义 (6.2.5)

设随机变量$X \sim N(0,1)$， $Y \sim \chi^2(n)$，并且X和Y独立，则称随机变量

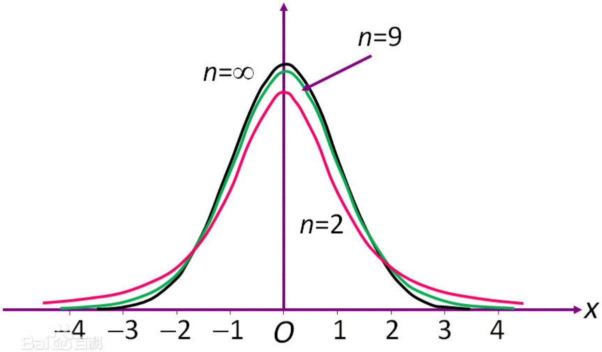
$$T = \frac{X}{\sqrt{Y/n}}$$

所服从的分布是自由度为n的t分布，记为$T \sim t(n)$。

t分布的概率密度函数

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

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



F分布

定义 (6.2.6)

设随机变量$X \sim \chi^2(n_1)$，$Y \sim \chi^2(n_2)$，且X与Y独立，则称随机变量

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

所服从的分布是自由度为n_1, n_2的F分布，记为$F \sim F(n_1, n_2)$。

F分布的概率密度函数

$$\psi(x) = \begin{cases} \dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}x^{n_1/2-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)(1+n_1x/n_2)^{(n_1+n_2)/2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

- F分布是一个非负值的偏态分布。

F分布与两个自由度的次序有关。

- 若$F \sim F(n_1, n_2)$，则$\frac{1}{F} \sim F(n_2, n_1)$；
- 若$T \sim t(n)$，则$T^2 \sim F(1, n)$；
- F分布的数学期望为$\frac{n_2}{n_2-2}$。

例子 (6.2.2)

设正态总体$X \sim N(0,4)$，而$X_1, \ldots, X_{15}$是来自总体X的样本，试求随机变量

$$Y = \frac{X_1^2 + \ldots + X_{10}^2}{2(X_{11}^2 + \ldots + X_{15}^2)}$$

所服从的分布。

例子 (6.2.2)

设正态总体$X \sim N(0,4)$，而$X_1, \ldots, X_{15}$是来自总体X的样本，试求随机变量

$$Y = \frac{X_1^2 + \ldots + X_{10}^2}{2(X_{11}^2 + \ldots + X_{15}^2)}$$

所服从的分布。

$Y \sim F(10,5)$.

一般的，求统计量的分布是个难题。只能确定一些特殊分布的统计量的分布。

定义 (6.3.1 正态总体基本定理)

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

1. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\overline{X}$与S^2相互独立
4. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

例子 (6.3.1)

设$X_1, \ldots, X_n$是来自正态总体$X \sim N(\mu, \sigma^2)$的样本，$\overline{X}$是样本均值，记$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$，求统计量$T = \frac{\overline{X}-\mu}{S_n/\sqrt{n-1}}$的分布。

定理 (6.3.2)

设$X_1,\ldots,X_m$是来自正态总体$X\sim N(\mu_1,\sigma^2)$的样本，$Y_1,\ldots,Y_n$是来自正态总体$N(\mu_2,\sigma^2)$的样本，且两样本相互独立。则有

$$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

其中

$$S_w^2=\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$$

例子

设$X_1,\ldots,X_n$是来自正态总体$X\sim N(\mu,\sigma^2)$的样本，$\overline{X},S^2$为其样本均值和样本方差。又设$X_{n+1}\sim N(\mu,\sigma^2)$，且与$X_1,\ldots,X_n$独立，求统计量

$$\frac{X_{n+1}-\overline{X}}{S}\sqrt{\frac{n}{n+1}}$$

的分布。

例子

设$X_1, \ldots, X_n$是来自正态总体$X \sim N(\mu, \sigma^2)$的样本，$\overline{X}, S^2$为其样本均值和样本方差。又设$X_{n+1} \sim N(\mu, \sigma^2)$，且与$X_1, \ldots, X_n$独立，求统计量

$$\frac{X_{n+1} - \overline{X}}{S} \sqrt{\frac{n}{n+1}}$$

的分布。

$$\frac{X_{n+1} - \overline{X}}{S} \sqrt{\frac{n}{n+1}} \sim t(n-1)$$