Chpt.6 Sampling and Distribution

第六章 样本及抽样分布

Introduction

概率论:

- 假设已知随机变量服从某一分布，研究它的性质、特点与规律，如求数字特征、随机变量函数的分布，介绍常用的各种分布等;

但是:

- 概率论中所描述的知识是如何获取的?
  比如，假如X服从正态分布，如何获取其参数?

- 实际中，如何判断一个随机变量是否服从某一分布?
  比如，如何判断X是否为正态分布?

Introduction

数理统计：

- 随机变量其分布未知或者不完全知道(如：是正态分布，但不知参数)，希望通过重复的、独立的观察得到许多数据，以概率论为理论基础，通过对这些数据的分析，对所研究的随机变量的分布做出种种推断。

- 举例：研究南开大学的学生身高

- 统计需要进行抽样、推断，这些工作形成了统计推断理论.

统计推断的理论基础：

- **大数定理**

1 频率稳定性：事件A发生的频率以概率收敛到概率p

$$\frac{n_A}{n} \xrightarrow{p} p \ (n \to \infty)$$

2 算术均值稳定性：$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \xrightarrow{p} \mu \quad (n \to \infty)$

- **中心极限定理**

大量相互独立的随机因素的综合影响，尽管这诸多的因素之分布是未知的，但是它们的和近似服从正态分布。

总体、个体

[定义]：在数理统计中，我们一般研究数量指标

- 对某一个数量指标进行随机实验或观察，试验的全部观察值称为<u>总体</u>。
- 每个观察值称为<u>个体</u>
- 总体中所包含的个体的数目称为总体的<u>容量</u>。容量有限的称为<u>有限总体</u>，容量无限的称为<u>无限总体</u>。

总体是对象某些指标的所有观察值:

- 扔硬币1000次，观察正面朝上的情况，得到1000个数值。
- 200位南开大学学生的身高。
- 天津市每天的最高气温。

随机变量 vs 总体：

[1] 随机变量是一组互异的值，以及对应每个值(或一个区间)出现的可能性大小。随机变量是从概率角度看的.

总体将所有试验结果一一罗列，可能出现大量相等的值。总体是从统计的角度看的。

Example: 随机抛一枚硬币1000次，随机变量X表示正面朝上的次数。

X	0	…	i	…	1000
P	$C_{1000}^{0}\left(\frac{1}{2}\right)^{1000}$	…	$C_{1000}^{i}\left(\frac{1}{2}\right)^{1000}$	…	$C_{1000}^{i}\left(\frac{1}{2}\right)^{1000}$

抛硬币，观察1000次，0表示正面朝上，1表示反面朝上，总体为

实验序号	1	2	3	4	…	1000
观察值	0	1	0	0	…	1

随机变量 vs 总体：

上例中的总体对应哪个随机变量？

0-1分布的随机变量

[2] 总体中的每个值是对随机变量X的观察值，也就是说一个总体对应一个随机变量；

总体的研究 ⟷ 随机变量的研究

随机变量的分布、数字特征就称为总体的分布、数字特征

注：今后将不区分总体与对应的随机变量，统称为总体X

在实际中总体的分布是未知的，你觉得应如何研究？

A 逐个观察总体中的每个个体

B 选取有代表性的个体

提交

从总体中抽取样本必须满足：

抽样是从总体中抽取部分个体，用于推断总体的特性。被抽出的部分个体称为总体的一个**样本**。

(1) 随机性　为使样本具有充分的代表性，抽样必须是随机的，应使总体中的每一个个体都有同等的机会被抽取到.

(2) 独立性　各次抽样必须是相互独立的，即每次抽样的结果既不影响其它各次抽样的结果，也不受其它各次抽样结果的影响.

称这种随机的、独立的抽样为简单随机抽样

由此得到的样本称为简单随机样本.

简单随机抽样

- 对有限总体进行放回抽样，属于简单随机抽样。
- 对有限总体进行不放回抽样，理论上不是简单随机抽样。
- 当总体容量N很大而样本容量n较小($n/N \le 10\%$) 时，在实际中可将不放回抽样近似看作简单随机抽样。
- 对于无限总体, 一般采取不放回抽样.

6.1 样本概念

从总体中抽取容量为n的样本，就是对总体所对应的随机变量X独立地进行n次观测，每次观测的结果仍可以看作一个随机变量。

n次观测结果对应n个随机变量：$X_1 \ldots\ldots X_n$，它们相互独立，并与总体X服从相同的分布。实际中抽样一经完成，就会得到一组实数值x_1，x_2，$\cdots, x_n$ 称为样本值。

若将样本$X_1 \ldots\ldots X_n$看作一个 n 维随机变量$(X_1 \ldots\ldots X_n)$, 则

(1) 当总体X是离散随机变量, 且概率分布为p(x)时， $(X_1 \ldots\ldots X_n)$的概率分布

$$p(x_1, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

(2) 当总体X是连续随机变量, 且概率密度为f(x)时， $(X_1 \ldots\ldots X_n)$的概率密度

$$f(x_1, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

6.2 统计量

来自总体X的n个样本$X_1\ldots\ldots X_n$ 构成n维随机变量 $(X_1\ldots X_n)$，$g(X_1,\cdots,X_n)$ 是$(X_1\ldots X_n)$的函数，若g中不含任何未知参数，则称 $g(X_1,\cdots,X_n)$ 为统计量。

统计量 $g(X_1,\cdots,X_n)$是随机变量。

设样本$X_1\ldots\ldots X_n$的一组观测值为 $x_1,\cdots,x_n$，算得的函数值

$g(x_1,\cdots,x_n)$ 称为统计量 $g(X_1,\cdots,X_n)$的观测值。

研究规律要用随机变量

实际应用可以直接用观测值

常用统计量及其观测值：

(1) 样本均值 $\overline{X}=\dfrac{1}{n}\sum_{k=1}^{n}X_k$ 观测值为 $\overline{x}=\dfrac{1}{n}\sum_{k=1}^{n}x_k$

(2) 样本方差 $S^2=\dfrac{1}{n-1}\sum_{k=1}^{n}(X_i-\overline{X})^2$

为什么这里分母是n-1?

观测值为 $s^2=\dfrac{1}{n-1}\sum_{k=1}^{n}(x_i-\overline{x})^2$

(3) 样本标准差 $S=\sqrt{S^2}=\sqrt{\dfrac{1}{n-1}\sum_{k=1}^{n}(X_i-\overline{X})^2}$

观测值为 $s=\sqrt{\dfrac{1}{n-1}\sum_{k=1}^{n}(x_i-\overline{x})^2}$

常用统计量及其观测值：

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

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

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

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

对比样本方差和样本的二阶中心矩

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n} (X_i - \overline{X})^2 \qquad B_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

经验分布函数

经验分布函数 我们还可以作出与总体分布函数 $F(x)$ 相应的统计量——经验分布函数.它的作法如下:设 $X_1,X_2,\cdots,X_n$ 是总体 F 的一个样本,用 $S(x)$,$-\infty<x<\infty$ 表示 $X_1,X_2,\cdots,X_n$ 中不大于 x 的随机变量的个数.定义经验分布函数 $F_n(x)$ 为

$$F_n(x)=\frac{1}{n}S(x),\quad -\infty<x<\infty.$$

例:设总体 F 具有一个样本值 1,2,3,则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & 若\ x<1,\\ \frac{1}{3}, & 若\ 1\leqslant x<2,\\ \frac{2}{3}, & 若\ 2\leqslant x<3,\\ 1, & 若\ x\geqslant 3.\end{cases}$$

例题：设总体X的一组样本值为：

4.5，2，1，1.5，3.5，4.5，6.5，5，3.5，4

求样本均值，样本方差，经验分布函数。

解：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(4.5+2+1+1.5+3.5+4.5+6.5+5+3.5+4)$

$= 3.6$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{1}{9}\Big((4.5-3.6)^2+(2-3.6)^2+(1-3.6)^2+(1.5-3.6)^2$$
$$+(3.5-3.6)^2+(4.5-3.6)^2+(6.5-3.6)^2+(5-3.6)^2$$
$$+(3.5-3.6)^2+(4-3.6)^2\Big)$$

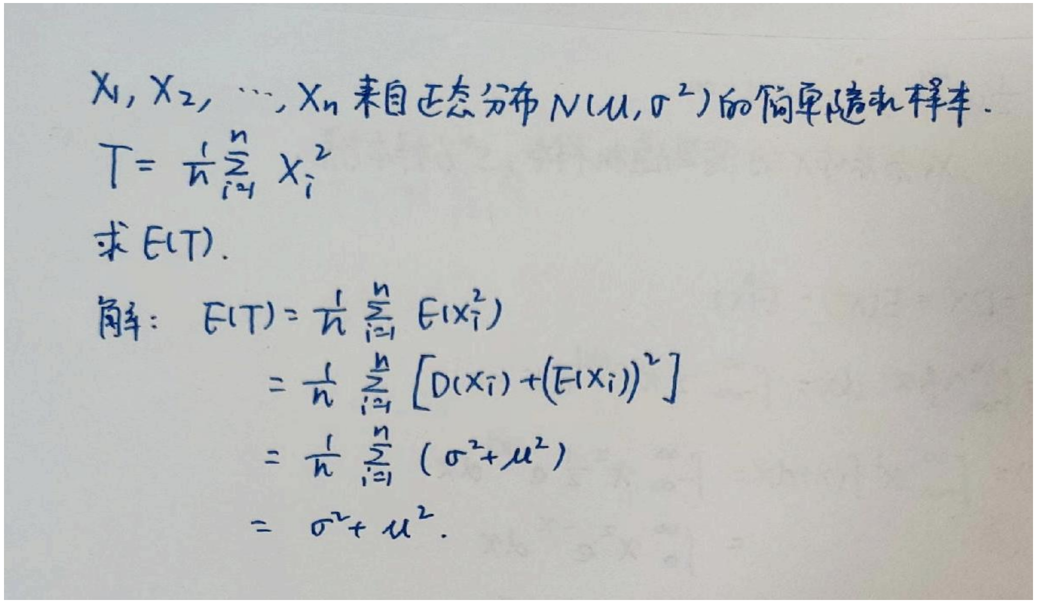
≈ 2.88

将样本值按从小到大排序：

$1 < 1.5 < 2 < 3.5 = 3.5 < 4 < 4.5 = 4.5 < 5 < 6.5$

$$F_{10}(x) = \begin{cases} 0 & x<1 \\ \frac{1}{10} & 1\le x<1.5 \\ \frac{2}{10} & 1.5\le x<2 \\ \frac{3}{10} & 2\le x<3.5 \\ \frac{5}{10} & 3.5\le x<4 \\ \frac{6}{10} & 4\le x<4.5 \\ \frac{8}{10} & 4.5\le x<5 \\ \frac{9}{10} & 5\le x<6.5 \\ 1 & x\ge 6.5 \end{cases}$$

例题：设$X_1, X_2..., X_n$是来自正态分布总体N(μ,σ^2)的简单随机样本，记统计量$T = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} X_i^2$，求E(T)。



$X_1, X_2, \cdots, X_n$ 来自正态分布 $N(\mu, \sigma^2)$ 的简单随机样本.

$T = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

求 $E(T)$.

解：$E(T) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2)$

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

$= \frac{1}{n}\sum_{i=1}^{n} (\sigma^2 + \mu^2)$

$= \sigma^2 + \mu^2.$

思考题

设X是一个随机变量，其均值为μ，方差为σ^2，$X_1, X_2 ..., X_n$是X的简单随机样本，S^2为样本方差。

证明：$E(S^2)=\sigma^2$.

例题：设$X_1,X_2...,X_n$是来自二项分布总体B(n,p)的简单随机样本，$\overline{X}$和S^{2}为样本均值和样本方差，记统计量T=$\overline{X}$-S^2。求E(T)

$X_1, X_2, \cdots, X_m$ 为来自二项分布总体 $B(n,p)$ 的简单随机样本.

$\overline{X}$, S^2 为样本均值和样本方差.

$T=\overline{X}-S^2$.

求 $E(T)$.

解：$E(T)=E(\overline{X})-E(S^2)$

$=E(X)-D(X)$

$=np-np(1-p)$

$=np^2$

例题：设总体X的概率密度为$f(x)=\frac{1}{2}e^{|x|}, -\infty<x<\infty, X_1, X_2, \ldots, X_n$为总体X的简单随机样本，$S^2$为样本方差。求$E(S^2)$.

$X.\quad f(x)=\frac{1}{2}e^{-|x|},\quad -\infty<x<\infty$

$X_1, X_2, \cdots, X_n$为总体X的简单随机样本，S^2为样本方差.

求ES^2.

解：$ES^2=DX=E(X^2)-E^2(X)$

$E(X)=\int_{-\infty}^{\infty}xf(x)\,dx=\int_{-\infty}^{\infty}\frac{1}{2}xe^{-|x|}\,dx=0$

$$
\begin{aligned}
E(X^2)&=\int_{-\infty}^{\infty}x^2f(x)\,dx=\int_{-\infty}^{\infty}x^2\frac{1}{2}e^{-|x|}\,dx\\
&=\int_0^{\infty}x^2e^{-x}\,dx\\
&=\int_0^{\infty}-x^2\,d(e^{-x})\\
&=-x^2e^{-x}\Big|_0^{\infty}+\int_0^{\infty}2xe^{-x}\,dx\\
&=\int_0^{\infty}-2x\,d(e^{-x})\\
&=-2xe^{-x}\Big|_0^{\infty}+\int_0^{\infty}2e^{-x}\,dx\\
&=-2e^{-x}\Big|_0^{\infty}\\
&=2.
\end{aligned}
$$

$ES^2=2-0=2.$

例题：求总体N(20,3)的容量分别为10，15的两个独立样本均值差的绝对值大于0.3的概率。

求总体 N(20,3) 的容量分别为 10,15 的两独立样本均值差的绝对值大于0.3的概率

解：设 $\bar{X}$ 是容量为10的样本均值. $\bar{X} \sim N(20, \frac{3}{10})$

$\bar{Y}$ 是容量为15的样本均值. $\bar{Y} \sim N(20, \frac{3}{15})$

$\bar{X}-\bar{Y}$ 也是正态变量, $\bar{X}-\bar{Y} \sim N(0, \frac{1}{2})$

$$\left(\begin{aligned} &E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = 0 \\ &D(\bar{X}-\bar{Y}) = D(\bar{X}) + D(-\bar{Y}) = D(\bar{X}) + D(\bar{Y}) = \frac{1}{2} \end{aligned} \right)$$

$$\begin{aligned} P(|\bar{X}-\bar{Y}| > 0.3) &= P\left(\left| \frac{\bar{X}-\bar{Y}}{\sqrt{2}} \right| > 0.3 \times \sqrt{2} \right) \\ &= 1 - P\left(\left| \frac{\bar{X}-\bar{Y}}{\sqrt{2}} \right| \le 0.3 \times \sqrt{2} \right) \\ &= 1 - [2\Phi(0.3 \times \sqrt{2}) - 1] \\ &= 2 - 2\Phi(0.3 \times \sqrt{2}) \\ &\approx 2 - 2\Phi(0.4243) \\ &= 2 - 2 \times 0.6628 \\ &= 0.6744 \end{aligned}$$