概率论与数理统计

第六章　样本及抽样分布

练习：

1. 甲乙两电影院在竞争1000名观众，假设每位观众在选择时随机的，且彼此相互独立，问甲至少应设多少个座位，才能使观众因无座位而离去的概率小于1%.

解：设X表示来甲电影院的人数，至少设N个座位.

则$X \sim B(1000, 0.5)$　$E(X)=500$, $D(X)=250$,

由中心极限定理 $\dfrac{X-500}{\sqrt{250}} \overset{近似}{\sim} N(0,1)$

故 $$\boldsymbol{P\{X > N\}} = P\left\{\frac{X-500}{\sqrt{250}} > \frac{N-500}{\sqrt{250}}\right\}$$

$$=1-\Phi\left(\frac{N-500}{\sqrt{250}}\right) < 1\%$$

即$\Phi\left(\dfrac{N-500}{\sqrt{250}}\right)>99\%$

因　$\boldsymbol{\Phi}(2.327)=0.99$

所以　$\dfrac{\boldsymbol{N}-500}{\sqrt{250}}>2.327,$

解得　$\boldsymbol{N}>536.79$

即　$\boldsymbol{N}=537$

2. 一系统是由n个相互独立起作用的部件组成，每个部件正常工作的概率为0.9，且必须至少由80％的部件正常工作，系统才能正常工作，问n至少为多大时，才能使系统正常工作的概率不低于 0.95 ？

解：设X表示正常工作的部件个 数

则$X \sim B(n,0.9)$ $\quad E(X)=0.9n,\ \ D(X)=0.09n,$

由中心极限定理 $\dfrac{X-0.9n}{\sqrt{0.09n}} \overset{近似}{\sim} N(0,1)$

故 $\boldsymbol{P\{n \geq X \geq 0.8n\}} = P\left\{\dfrac{n-0.9n}{\sqrt{0.09n}} \geq \dfrac{X-0.9n}{\sqrt{0.09n}} \geq \dfrac{0.8n-0.9n}{\sqrt{0.09n}}\right\}$

$$= 2\Phi\left(\frac{0.1n}{0.3\sqrt{n}}\right) - 1 \geq 0.95$$

即$\Phi\left(\frac{0.1n}{0.3\sqrt{n}}\right) \geq 0.975$

因 $\Phi(1.960) = 0.975$

所以 $\frac{0.1n}{0.3\sqrt{n}} \geq 1.960$

解得 $n \geq 34.571$

即 $n = 35$

第六章　样本及抽样分布

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

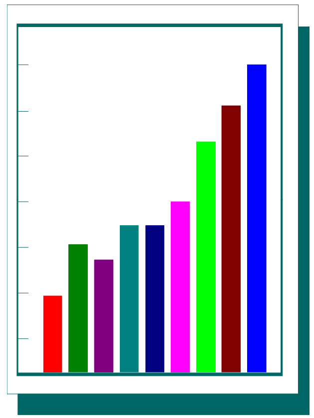
- 参数估计（第七章）
- 假设检验（第八章）
- 方差分析（第九章）
- 回归分析（第九章）

§1　随机样本

- 总体和样本
- 小结

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.

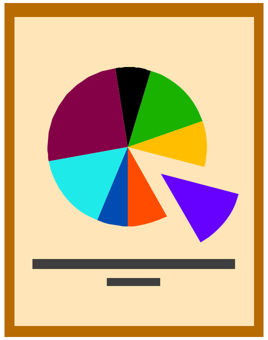


客观上，只允许我们对随机现象进行次数不多的观察试验 ，我们只能获得局部观察资料.

数理统计的任务就是研究有效地收集、整理、分析所获得的有限的资料，对所研究的问题，尽可能地作出精确而可靠的结论.

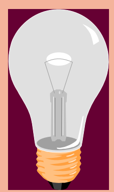
在数理统计中，不是对所研究的对象全体(称为总体)进行观察，而是抽取其中的部分(称为样本)进行观察获得数据（抽样），并通过这些数据对总体进行推断.

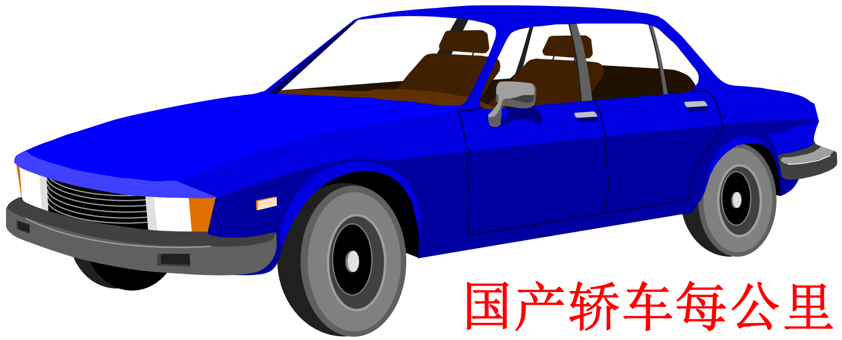
数理统计方法具有“部分推断整体”的特征.



在数理统计研究中，人们往往研究有关对象的某一项(或几项)数量指标，为此，对这一指标进行随机试验，观察试验结果全部观察值，从而考察该数量指标的分布情况.这时，每个具有的数量指标的全体就是总体.每个数量指标就是个体.

某批灯泡的寿命





国产轿车每公里的耗油量

该批灯泡寿命的全体就是总体

国产轿车每公里耗油量的全体就是总体

一、总体和样本

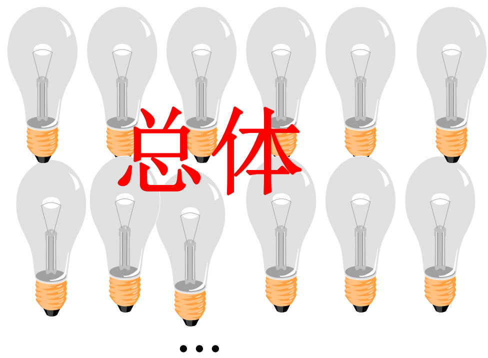
1.总体

一个统计问题总有它明确的研究对象.

研究对象的全体称为**总体**,

总体中每个成员称为**个体**,

总体中所包含的个体的个数称为总体的**容量**.



研究某批灯泡的质量

$$\text{总体}\begin{cases}\text{有限总体}\\ \text{无限总体}\end{cases}$$

实例1 某工厂10月份生产的灯泡寿命所组成的总体中,个体的总数就是10月份生产的灯泡数，这是个有限总体;而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体,它包括以往生产和今后生产的灯泡寿命.

实例2 在考察某大学一年级男生的身高这一试验中,若一年级男生共2 000人，每个男生的身高是一个可能观察值,所形成的总体中共含2 000个可能观察值，是一个有限总体.

实例3 考察某一湖泊中某种鱼的含汞量,所得总体也是有限总体.

有些有限总体，它的容量很大，我们可以认为它是一个无限总体.

实例4 考察全国正在使用的某种型号灯泡的寿命所形成的总体，由于可能观察值的个数很多,就可以认为是无限总体.

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.

例如:研究某批灯泡的寿命时，关心的数量指标就是寿命，那么，此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 X 可用一概率（指数）分布来刻划

寿命总体是指数分布总体

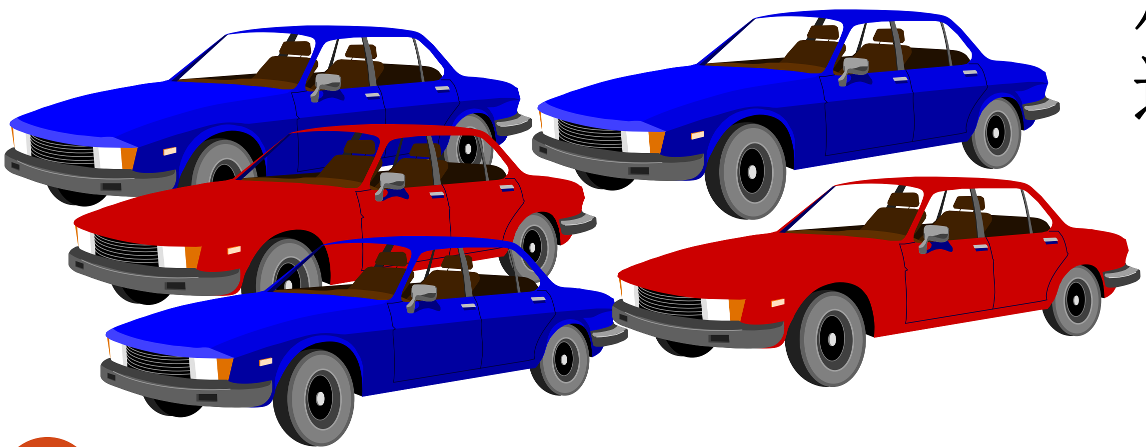
鉴于此，常用随机变量的记号或用其分布函数表示总体. 如说总体X或总体$F(x)$.

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.

2. 样本

总体分布一般是未知，或只知道是包含未知参数的分布，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，这一抽取过程称为“抽样”，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.



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

对总体X在相同的条件下，进行n次重复、独立观察，其结果依次记为X_1，X_2,…，X_n.

这样得到的随机变量$X_1, X_2, \cdots X_n$是来自总体X的一个简单随机样本，与总体随机变量具有相同的分布. n称为这个样本的容量.

一旦取定一组样本X_1，… ,X_n ,得到n个具体的数$(x_1,x_2,\ldots,x_n)$，称为样本的一次观察值，简称样本值 .

最常用的一种抽样叫作“简单随机抽样”，其特点：

1. 代表性：$X_1,X_2,\ldots,X_n$中每一个与所考察的总体有相同的分布.

2. 独立性：$X_1,X_2,\ldots,X_n$是相互独立的随机变量.

定义:

设X是具有分布函数F的随机变量，若$X_1, X_2, \cdots, X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数F(或总体F、或总体X）得到的容量n为的简单随机样本，简称样本，它们的观察值$x_1, x_2, \cdots, x_n$称为样本值，又称为X的n个独立的观察值.

由简单随机抽样得到的样本称为简单随机样本，它可以用与总体独立同分布的n个相互独立的随机变量$X_1, X_2, \ldots, X_n$表示.

若总体的分布函数为$F(x)$、概率密度函数为$f(x)$,则其简单随机样本的联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = F(x_1)\, F(x_2) \ldots F(x_n)$$

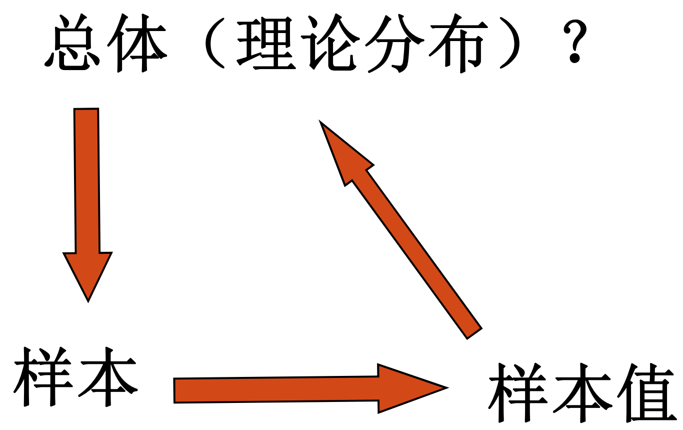
其简单随机样本的联合概率密度函数为

$$f^*(x_1, x_2, \cdots, x_n) = f(x_1)\, f(x_2) \ldots f(x_n)$$

简单随机样本是应用中最常见的情形，今后，当说到“$X_1,X_2,\ldots,X_n$是取自某总体的样本”时，若不特别说明，就指简单随机样本.

3. 总体、样本、样本值的关系

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.



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.

二、小结

研究对象的全体称为总体

总体中每个成员称为个体

设X是具有分布函数F的随机变量，若$X_1, X_2, \cdots, X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数F(或总体F、或总体X）得到的容量n为的简单随机样本.

简称样本.

§3　抽样分布

- 统计量与经验分布函数
- 统计学三大抽样分布
- 几个重要的抽样分布定理
- 小结

一、统计量与经验分布函数

1. 统计量

由样本值去推断总体情况，需要对样本值进行“加工”，这就要构造一些样本的函数，它把样本中所含的（某一方面）的信息集中起来.

这种不含任何未知参数的样本的函数称为统计量. 它是完全由样本决定的量.

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.

请注意 ：

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 .

例1 $X \sim N(\mu,\sigma^2)$，μ,σ^2 是未知参数,

$(X_1, X_2, \cdots, X_n)$ 是一样本, 则

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

是统计量, 其中 $X_i \sim N(\mu,\sigma^2)$

但 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 不是统计量.

若 μ,σ 已知,则为统计量

几个常见统计量

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值．

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

它反映了总体均值的信息

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

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

它反映了总体方差的信息

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

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

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

例如

$$A_1 = \overline{X}$$

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

注 样本方差S^2与样本二阶中心矩S_n^2的不同

1）关系式 $S^2 = \dfrac{n}{n-1} S_n^2$

推导 $\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^{\,2} - 2X_i\overline{X} + \overline{X}^2)$

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

$$= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = n(A_2 - \overline{X}^2)$$

注 $B_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

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

故 $B_2 = A_2 - \overline{X}^2$ $S^2 = \dfrac{n}{n-1}(A_2 - \overline{X}^2) = \dfrac{n}{n-1} S_n^2$

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

推导 $E(X)=\mu\,,D(X)=\sigma^2$ 则

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

$$E(S_n^2)=E(A_2)-E\left(\overline{X}^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\left[D\left(\overline{X}\right)+\left[E\left(\overline{X}\right)\right]^2\right]$$

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

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

统计量的观察值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad \alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \qquad k = 1, 2, \cdots$$

$$b_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k \qquad k = 1, 2, \cdots$$

由以上定义得下述结论:

若总体 X 的k阶矩 $E(X^k) \underline{\underline{记成}}\ \mu_k$存在,

则当$n \to \infty$时, $A_k \xrightarrow{P} \mu_k$, $k=1,2,\cdots$.

证明 因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布,

所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布,

故有 $E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = \mu_k$.

再根据第五章辛钦定理知

辛钦定理

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k, \quad k=1,2,\cdots;$$

由第五章关于依概率收敛的序列的性质知

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k),$$

其中g是连续函数.

以上结论是下一章所要介绍的矩估计法的理论根据.

例2 从一批机器零件毛坯中随机地抽取10件，测得其重量为(单位: 公斤):

210, 243, 185, 240, 215,

228, 196, 235, 200, 199

求这组样本值的均值、方差、二阶原点矩与二阶中心矩.

解　令 $(x_1, x_2, \cdots, x_{10})$

$$= (210,\ 243,\ 185,\ 240,\ 215,\ 228,\ 196,\ 235,\ 200,\ 199)$$

则 $$\overline{x}=\frac{1}{10}(230+243+185+240+215$$
$$+228+196+235+200+199)$$
$$=217.19$$

$$s^2=\frac{1}{9}\sum_{i=1}^{10}(x_i-\overline{x})^2=433.43$$

$$A_2=\frac{1}{10}\sum_{i=1}^{10}x_i^{\ 2}=47522.5$$

$$B_2=\frac{9}{10}s^2=\frac{1}{10}\sum_{i=1}^{10}(x_i-\overline{x})^2=390.0$$

2. 经验分布函数

总体分布函数 $F(x)$ 相应的统计量称为经验分布函数.

经验分布函数的做法如下:

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

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

对于一个样本值，$F_n(x)$的观察值容易求得.（$F_n(x)$的观察值仍以$F_n(x)$表示.）

实例 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

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

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

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2\\ 1, & x\ge 2.\end{cases}$$

一般地，

设 $x_1, x_2, \cdots, x_n$ 是总体 F 的一个容量为 n 的样本值，

先将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列，

并重新编号， $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, \quad k = 1,2,\cdots,n-1. \\ 1, & x \ge x_{(n)}. \end{cases}$$

格里汶科定理

对于任一实数 x，当 $n \to \infty$ 时，$F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} \left|F_n(x) - F(x)\right| = 0\right\} = 1.$$

对于任一实数 x 当 n 充分大时，经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来使用.

作业：课后习题 2、3、4

练习：

1、 在总体$N(52, 6.3^2)$中,随机抽取一个容量为36的样本,求样本均值$\overline{X}$落在50.8到53.8之间的概率.

解 $\overline{X} \sim N(52,\ 6.3^2/36)$

故 $P(50.8 < \overline{X} < 53.8)$

$$= \Phi\left(\frac{53.8-52}{6.3/6}\right) - \Phi\left(\frac{50.8-52}{6.3/6}\right)$$

$$= \Phi(1.7143) - \Phi(-1.1429)$$

$$= 0.8239$$

2、设 $X \sim b(1,p)$, $X_1, X_2, \cdots X_n$, 是来自总体X 的样本，

那么下列选项中不正确的是 ____

A) 当 充分大时， 近似有 $\overline{X} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$

B） $P\left\{\overline{X} = k\right\} = C_n^k p^k (1-p)^{n-k}, k = 0,1,2,\cdots n$

C） $P\left\{\overline{X} = \dfrac{k}{n}\right\} = C_n^k p^k (1-p)^{n-k}, k = 0,1,2,\cdots n$

D） $P\left\{X_i = k\right\} = C_1^k p^k (1-p)^{1-k}, 1 \le i \le n, \quad k = 0,1,$

答案：B