文科概率统计

3.6 抽样分布

3.6.2 总体与样本

数理统计是一个应用极广的数学分支，它以概率论为理论基础，以试验和观测结果为依据，对随机现象的统计规律性作出种种合理的估计和推断.

由于在大量的重复试验下，任何随机现象都会呈现出其确定的统计规律性，而实际中允许的试验又总是有限的，所以从全部研究对象中抽取由一部分个体组成的局部，并通过这个局部的特性去推断总体的特性，便成了数理统计的重要任务之一.

定义 在数理统计中被研究对象的全体称为**总体**.

定义 组成总体的每个单位称为**个体**.

例如，在人口普查中，全部人口就是总体，而其中每一个人就是一个个体.

研究某城市中学生的身高分布情况，此时全体中学生的身高是一个总体，而每个中学生的身高则是一个个体.

在实际应用中，人们所关心的不是总体中每个个体的具体性能，而是它的某一数量指标，对于一个确定的由数量指标构成的总体来说，由于每个个体的取值是不同的，从总体中任取一个个体，其取值是不能预先确定的，所以总体的任何一个数量指标都是一个随机变量. 因此，通常用随机变量 X 表示总体，即总体是指某个随机变量 X 可能取值的全体.

定义 从总体 X 中抽取一个个体，就是对代表总体的随机变量 X 进行一次试验(或观测)，记为 X_i，其具体的取值记为 x_i. 从总体中抽取若干个个体 $X_i(i=1,2,\cdots,n)$ 的过程，称为**抽样**；抽出的这些个体 X_i 所成的集体，称为**样本**(或**子样**)，记为 $(X_1,X_2,\cdots,X_n)$；样本中所含个体的个数 n 称为**样本容量**；每个抽中的个体 X_i 称为**样本点**. 其具体的取值 x_i 称为**样本观测值**.

3.6.3 简单随机样本

由于要从样本来推断总体的分布并进行各种分析，因此要求样本能够很好地反映总体的特征.

定义 如果从总体 X 中进行独立的重复试验，得到容量为 n 的样本 $(X_1, X_2, \cdots, X_n)$ 满足下面的两个条件：

(1) $X_i (i=1,2,\cdots,n)$ 与 X 有相同的分布函数 $F(X)$；

(2) $X_1, X_2, \cdots, X_n$ 相互独立.

那么样本 $(X_1, X_2, \cdots, X_n)$ 称为**简单随机样本**. 这种抽样称为**简单随机抽样**. 抽样后，简单随机样本 $(X_1, X_2, \cdots, X_n)$ 的样本值（或称为样本观测值）就成为 n 个具体的数值 $(x_1, x_2, \cdots, x_n)$（如图 3.34 所示）.

今后，凡是不加特别说明的，所提到的抽样与样本，都是指简单随机抽样与简单随机样本.

3.6.5 统计量的概念

样本是总体的代表和反映，也是统计推断的依据. 为了对总体的分布或数字特征进行各种统计推断，还需要对样本作加工处理，把样本中应关心的事物和信息集中起来，针对不同的问题构造出样本的不同函数，这种样本的函数称为统计量.

定义 由样本$(X_1, X_2, \cdots, X_n)$所确定的函数$f(X_1, X_2, \cdots, X_n)$称为**统计量**.

若$(x_1, x_2, \cdots, x_n)$是一个样本观测值，则称$f(x_1, x_2, \cdots, x_n)$是统计量$f(X_1, X_2, \cdots, X_n)$的一个**观测值**.

显然，统计量不仅是一个随机变量，而且还不含有未知参数.

例 3.6.2 设(X_1, X_2, X_3)是由服从正态分布$N(\mu, \sigma^2)$的总体 X 中抽取的一个容量为 3 的样本，其中μ、σ是未知参数，因此$\frac{X_1 + X_2 + X_3}{3} - \mu$，$\frac{X_1 + X_2 + X_3}{\sigma}$都不是统计量，而$X_1 + X_2 + 5$，$X_1^2 + X_2^2$都是统计量.

设$(X_1, X_2, \cdots, X_n)$是总体X中的一个样本，下面是数理统计中常用的几个统计量及其观测值：

（1）样本均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，它的观测值为 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

（2）样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$，它的观测值为

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

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

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

3.6.6 抽样分布

定义 在数理统计中把统计量的分布称为**抽样分布**.

一般来说,要确定某个统计量的分布是比较困难的,有时甚至是不可能的.但是对于来自正态总体的几个常用统计量的分布,已得到了一系列重要的结果,下面不加证明地介绍在统计推断中常用的几个统计量的分布.

1. 设总体 $X \sim N(\mu,\sigma^2)$,且 $(X_1,X_2,\cdots,X_n)$ 是 X 中样本容量为 n 的样本,样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$,则

(1) $$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right).$$

(2) 统计量 $U=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 服从标准正态分布(此统计量称为 U **统计量**),即

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

2. 设总体 $X \sim N(\mu, \sigma^2)$，且 $(X_1, X_2, \cdots, X_n)$ 是 X 中样本容量为 n 的样本，样本均值 $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$，样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$，

令

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

则称统计量 T 服从自由度为 $n-1$ 的 t 分布（此统计量亦称为 t 统计量），即

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

其密度曲线如图 3.36 所示(本课程略去了定理的证明).

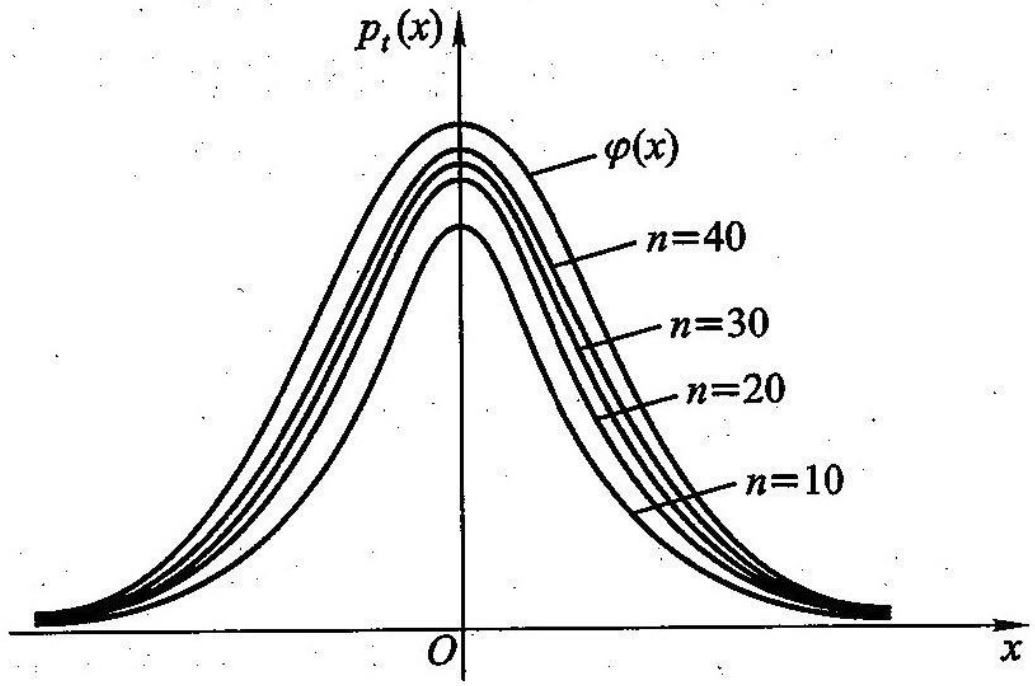


图 3.36

由 t 分布的密度函数可知,t 分布的密度曲线是关于 y 轴对称的,当 $n>30$ 时,t 分布近似于标准正态分布,即它们的密度曲线几乎是相同的.

对给定的 $\alpha(0<\alpha<1)$ 和 n，称满足等式

$$P(T \geqslant t_\alpha(n)) = \int_{t_\alpha(n)}^{+\infty} p_t(x)\,\mathrm{d}x = \alpha$$

的 $t_\alpha(n)$ 为 t **分布的临界值**（图 3.37）.

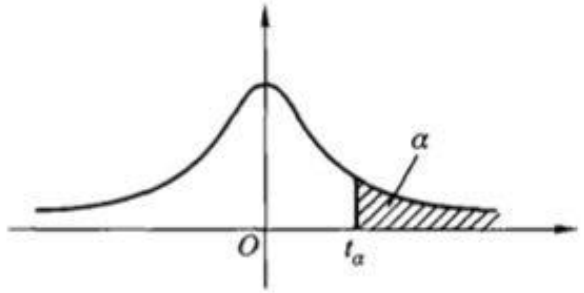
对于不同的 α 和 n，临界值 $t_\alpha(n)$ 的值可由 t 分布临界值表（附表 3）查得.

例 3.6.3 设 $n=20,\alpha=0.01$，查自由度为 20 的 t 分布临界值表（附表 3），则有

$$t_\alpha(n) = t_{0.01}(20) = 2.528\ 0.$$

附表3　t 分布临界值表

$$P\{t(k) > t_\alpha\} = \alpha$$



k \ t_α \ α	0.25	0.10	0.05	0.025	0.01	0.005
1	1.000 0	3.077 7	6.313 8	12.706 2	31.820 7	63.657 4
2	0.816 5	1.885 6	2.920 0	4.320 7	6.964 6	9.924 8
3	0.764 9	1.637 7	2.353 4	3.182 4	4.540 7	5.840 9
4	0.740 7	1.533 2	2.131 8	2.776 4	3.746 9	4.604 1
5	0.726 7	1.475 9	2.015 0	2.570 6	3.364 9	4.032 2
6	0.717 6	1.439 8	1.943 2	2.446 9	3.142 7	3.707 4
7	0.711 1	1.414 9	1.894 6	2.364 6	2.998 0	3.499 5
8	0.706 4	1.396 8	1.859 5	2.306 0	2.896 5	3.355 4
9	0.702 7	1.383 0	1.833 1	2.262 2	2.821 4	3.249 8
10	0.699 8	1.372 2	1.812 5	2.228 1	2.763 8	3.169 3
11	0.697 4	1.363 4	1.795 9	2.201 0	2.718 1	3.105 8
12	0.695 5	1.356 2	1.782 3	2.178 8	2.681 0	3.054 5
13	0.693 8	1.350 2	1.770 9	2.160 4	2.650 3	3.012 3
14	0.692 4	1.345 0	1.761 3	2.144 8	2.624 5	2.976 8
15	0.691 2	1.340 6	1.753 1	2.131 5	2.602 5	2.946 7
16	0.690 1	1.336 8	1.745 9	2.119 9	2.583 5	2.902 8
17	0.689 2	1.333 4	1.739 6	2.109 8	2.566 9	2.898 2
18	0.688 4	1.330 4	1.734 1	2.100 9	2.552 4	2.878 4
19	0.687 6	1.327 7	1.729 1	2.093 0	2.539 5	2.860 9
20	0.687 0	1.325 3	1.724 7	2.086 0	2.528 0	2.845 3

3.（中心极限定理）设总体 X 具有有限的数学期望 $E(X)=\mu$，有限的方差 $D(X)=\sigma^2>0$，当样本容量 n 充分大时（$n\geqslant 50$），$\overline{X}$ 近似地服从正态分布 $N\left(\mu,\frac{\sigma^2}{n}\right)$，即

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\text{近似服从标准正态分布 } N(0,1).$$

4. 设总体 $X \sim N(\mu_1, \sigma_1^2)$，且 $(X_1, X_2, \cdots, X_{n_1})$ 是 X 中样本容量为 n_1 的样本，总体 $Y \sim N(\mu_2, \sigma_2^2)$，且 $(Y_1, Y_2, \cdots, Y_{n_2})$ 是 Y 中样本容量为 n_2 的样本，样本均值

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i,$$

则统计量 $\bar{X} - \bar{Y}$ 服从正态分布 $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$，即

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

5. 设 $(X_1, X_2, \cdots, X_{n_1})$ 来自总体 $X, EX = \mu_1, DX = \sigma_1^2$，$(Y_1, Y_2, \cdots, Y_{n_2})$ 来自总体 $Y, EY = \mu_2, DY = \sigma_2^2$. 样本均值 $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$. 当 n_1、n_2 充分大时 $(n_1 \geqslant 50, n_2 \geqslant 50)$，则统计量

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

近似服从 $N(0,1)$.

例 3.6.4 学生某科成绩 X 服从 $N(80,100)$，现抽取 $n=25$ 的一个样本，试问：

（1）样本平均成绩大于 84 分的概率是多少？

（2）平均成绩在 78～82 分之间的概率是多少？

解 （1）由第 1 条结论

$$X \sim N(80,100), \quad n=25,$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N(80,2^2),$$

即

$$\frac{\bar{X}-80}{10/\sqrt{25}} \sim N(0,1).$$

故

$$P(\bar{X}>84) = P\left(\frac{\bar{X}-80}{10/\sqrt{25}}>2\right) = 1-\Phi(2),$$

其中 $\Phi(x)$ 是标准正态分布的分布函数，查标准正态分布函数数值表（附表 2），就可以得到所求概率 $=1-0.977\ 2=0.022\ 8$（图 3.38）.

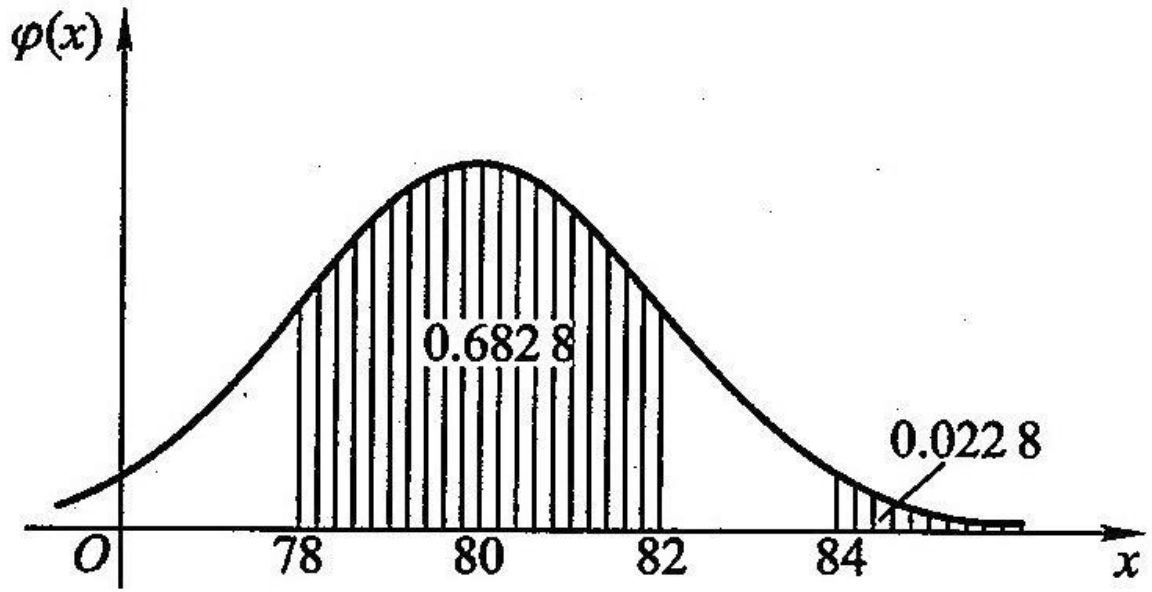


图 3.38

$$(2)\ P(78<\bar{X}<82)=P\left(-1<\frac{\bar{X}-80}{2}<1\right)$$

$$=\Phi(1)-\Phi(-1)=2\Phi(1)-1$$

$$=0.682\ 6.$$

例 3.6.5 设 $X \sim N(150, 20^2)$，$Y \sim N(125, 25^2)$，从 X、Y 中各抽取一个样本容量为 5 的样本，其样本平均值为 $\bar{X}$ 和 $\bar{Y}$，试求 $\bar{X} - \bar{Y} \leqslant 0$ 的概率.

解 已知 $\mu_1 = 150, \mu_2 = 125, \sigma_1 = 20, \sigma_2 = 25, n_1 = n_2 = 5$.

由第 4 条结论可知：

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right), \text{即}$$

$$\bar{X} - \bar{Y} \sim N\left(150 - 125, \frac{20^2}{5} + \frac{25^2}{5}\right) = N(25, 205).$$

所以
$$P(\bar{X} - \bar{Y} \leqslant 0) = P\left(\frac{(\bar{X} - \bar{Y}) - 25}{\sqrt{\frac{20^2}{5} + \frac{25^2}{5}}} \leqslant \frac{-25}{\sqrt{\frac{20^2}{5} + \frac{25^2}{5}}}\right)$$

$$= \Phi\left(-\frac{25}{\sqrt{205}}\right) = \Phi(-1.746) = 1 - \Phi(1.746)$$

$$= 1 - 0.959\ 9 = 0.040\ 1.$$

注 查标准正态分布函数数值表(附表 2)得 $\Phi(1.746) = 0.959\ 9$

例 3.6.6 根据历史数据,已知顾客在商店 A 所花费的平均时间为 55 min,顾客在商店 B 所花费的平均时间为 49 min,假定每个总体(商店 A,商店 B)的标准差均为 15 min. 现一市场分析员在每个商店各观测了 75 名顾客. 用 X 表示顾客在 A 商店所花费的时间,Y 表示顾客在 B 商店所花费的时间,求 $P(\overline{X}-\overline{Y}>0)$.

解 由题已知:$EX=\mu_1=55$,$EY=\mu_2=49$,$\sigma_1^2=\sigma_2^2=15^2$,$n_1=n_2=75$.

由第 5 条结论可知

$$U=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$ 近似服从 $N(0,1)$.

所以
$$P(\bar{X}-\bar{Y}\leqslant 0)=P\left(\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leqslant\frac{-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$$
$$=\Phi\left(-\frac{55-49}{\sqrt{\frac{15^2}{75}+\frac{15^2}{75}}}\right)=\Phi(-2.449)$$
$$=1-\Phi(2.449)=1-0.992\,9=0.007\,1,$$
即
$$P(\bar{X}-\bar{Y}>0)=1-P(\bar{X}-\bar{Y}\leqslant 0)=0.992\,9.$$

标准正态分布表： $\Phi(0)=0.5$、$\Phi(1)=0.8413$、 $\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、 $\Phi(1.96)=0.975$、 $\Phi(2)=0.9772$

设总体 $X\sim N(12,4)$, 在总体 X 中随机取一个样本容量为 4 的样本 (X_1,X_2,X_3,X_4)，$\overline{X}$ 为样本均值，试求 $P\{|\overline{X}-12|<1\}$.

标准正态分布表： $\Phi(0)=0.5$、$\Phi(1)=0.8413$、 $\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、 $\Phi(1.96)=0.975$、 $\Phi(2)=0.9772$

设总体 $X\sim N(12,4)$, 在总体 X 中随机取一个样本容量为 4 的样本

(X_1,X_2,X_3,X_4)， $\bar{X}$ 为样本均值，试求 $P\{|\bar{X}-12|<1\}$.

解： $X\sim N(12,4), \bar{X}\sim N(12,\frac{4}{4})=N(12,1)$,所以 $\frac{\bar{X}-12}{1}\sim N(0,1)$

$$P\{|\bar{X}-12|<1\}=P\{-1<\frac{\bar{X}-12}{1}<1\}=\Phi(1)-\Phi(-1)=2\Phi(1)-1$$

=2*0.8413-1=0.6826。

标准正态分布表： $\Phi(0)=0.5$、$\Phi(1)=0.8413$、 $\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、 $\Phi(1.96)=0.975$、 $\Phi(2)=0.9772$

若总体 $X\sim N(2,\ \sigma^2)$,(X_1,X_2,X_3,X_4)是来自总体 X 的简单随机样本，

$\overline{X}$ 为样本均值，试求 $P\{\overline{X}-2>0.98\sigma\}$.

标准正态分布表： $\Phi(0)=0.5$、$\Phi(1)=0.8413$、 $\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、 $\Phi(1.96)=0.975$、 $\Phi(2)=0.9772$

若总体 $X\sim N(2,\ \sigma^2),(X_1,X_2,X_3,X_4)$ 是来自总体 X 的简单随机样本，

$\overline{X}$ 为样本均值，试求 $P\{\overline{X}-2>0.98\sigma\}$.

解：由 $X\sim N(2,\ \sigma^2)$，知 $\overline{X}\sim N(2,\ \frac{\sigma^2}{4})$，则

$$P(\overline{X}-2>0.98\sigma)=P(\frac{\overline{X}-2}{\sigma/2}>\frac{0.98\sigma}{\sigma/2})=P(\frac{\overline{X}-2}{\sigma/2}>1.96)=1-\Phi(1.96)=0.025.$$

标准正态分布表：　$\Phi(0)=0.5$、$\Phi(1)=0.8413$、　$\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、　$\Phi(1.96)=0.975$、　$\Phi(2)=0.9772$

设总体 $X\sim N(52,\ 6^2)$，从总体 X 中抽取一个样本容量为 36 的样本，$\overline{X}$ 为样本均值，求 $E(\overline{X})$，$D(\overline{X})$，并计算样本均值落在 50.5 到 53.5 之间的概率.

标准正态分布表：　$\Phi(0)=0.5$、$\Phi(1)=0.8413$、　$\Phi(1.5)=0.9332$、

$\Phi(1.65)=0.95$、　$\Phi(1.96)=0.975$、　$\Phi(2)=0.9772$

设总体 $X\sim N(52,\ 6^2)$，从总体 X 中抽取一个样本容量为 36 的样本，

$\overline{X}$ 为样本均值，求 $E(\overline{X})$，$D(\overline{X})$，并计算样本均值落在 50.5 到 53.5 之间的概率.

解：　因为总体 $X\sim N(52,\ 6^2)$，

故　$\overline{X}\sim N(52,\ \frac{6^2}{36})$，即 $\overline{X}\sim N(52,\ 1)$，

因此　$E(\overline{X})=52$，　$D(\overline{X})=1$；

同时　$P\{50.5<\overline{X}<53.5\}=\Phi(\frac{53.5-52}{1})-\Phi(\frac{50.5-52}{1})$

$=\Phi(1.5)-\Phi(-1.5)=0.8664$.

假设总体 X 服从正态分布 $N(\mu,\sigma^2)$，$X_1,X_2,\ldots,X_9$ 是取自总体 X 的简单随机样本，$\overline{X}$ 为样本均值，若 $P(X-\mu<a)=P(\overline{X}-\mu<b)$，则 $\frac{a}{b}=$

假设总体 X 服从正态分布 $N(\mu,\sigma^2)$，$X_1,X_2,\ldots,X_9$ 是取自总体 X 的简单随机样本，$\overline{X}$ 为样本均值，若 $P(X-\mu<a)=P(\overline{X}-\mu<b)$，则 $\frac{a}{b}=$ 3