当今大数据时代，概率论的重要性日趋显现，对于我们而言，即使不是去应付考研，也要去了解其基本内涵，防止“掉队”。

其实不仅是考研，高考理数中，就全国 1 卷而言，概率与统计在高考的考察难度与比重也是逐渐增大，当然这是对接大学高等数学、概率论与数理统计、随机过程的趋势。我觉得这是好事，圆锥曲线那种区分度基本靠庞大且繁琐的计算量，简直是高考数学的毒瘤，对思维训练几乎无益处，但一直作为高考理数的次压轴或者压轴，折磨一代又一代高中生。然而，上大学之后根本就没遇到过那个玩意。

目前的人工智能、机器学习、深度学习用到概率统计的地方可太多了，比如**极大似然估计**、贝叶斯统计、**概率分布**、假设检验、最大熵模型、各种回归算法等等。提高概率统计的重心与地位，是目前大数据信息化时代的趋势，也是对未来人才的能力要求。

咳……扯得有点远了，继续说回考研，说了那么多，其实就是想告诉你，考研中的概率论一定会越来越具有深度，只想套公式就能做出来的日子一去不复返了！那考研概率论中哪一部分是难点呢？我觉得其一在概率分布与概率密度的含义理解，其二在数理统计的基本概念上，这篇先讲第二个难点（也没有讲完，内容有点多），以后有机会再详细探讨第一点。

先明确几个概念：

1.**总体**：研究对象的全体称为总体；

2.**样本**：n 个相互独立且与总体 X 具有相同概率分布的随机变量 X_1、$X_2,\cdots,X_n$ 的整体 $(X_1$、$X_2,\cdots,X_n)$ 称为来自总体 X，容量为 n 的一个**简单随机样本**，简称为**样本**，一次抽样结果的 n 个具体数值 $(x_1$、$x_2,\cdots,x_n)$，称为样本 X_1、$X_2,\cdots,X_n$ 的一个**观测值**（或样本值）；

3.**统计量**：设 X_1、$X_2,\cdots,X_n$ 为来自总体 X 的样本，$\mathrm{g}(x_1$、$x_2,\cdots,x_n)$ 为 n 元函数，如果 g 中不含任何未知参数，则称 $\mathrm{g}(X_1$、$X_2,\cdots,X_n)$ 为样本 X_1、$X_2,\cdots,X_n$ 的一个统计量，若 $(x_1$、$x_2,\cdots,x_n)$ 为样本值，则称 $\mathrm{g}(x_1$、$x_2,\cdots,x_n)$ 为 $\mathrm{g}(X_1$、$X_2,\cdots,X_n)$ 的**观测值**。

希望各位能好好体会这三个概念，搞清楚样本与观测值的区别，明确统计量作为随机样本的函数，因此统计量也是随机变量。

明确了这三个概念，后面的内容就好展开了，对于期望、方差这些概念可能大家都很熟悉，这是总体 X 的最重要的数字特征，但是现实中我们很难遍历总体，例如分析中国人饮食习惯，总不能一个个问遍全中国人吧，那怎么办呢？样本应运而生，我们可以进行抽样调查，只要保证足够随机，这样可以拿样本近似代替总体，大大降低了调研成本。

那现在问题来了，你咋能保证用样本数字特征代替总体的数字特征不会有偏差，这就涉

及到**无偏性**的概念，通过构造某一个统计量作为样本数字特征，其实就是样本均值，样本均值既然是统计量，因此它也是随机变量，随机变量自然就有期望，因此要求样本均值的期望等于总体期望，这样代替才有意义；同样的道理，样本方差也是构造出来的一个统计量，用以代替（或者说估计）总体方差，以保证无偏性，下面给出样本均值和样本方差的定义：

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

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

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

其中$\sum_{i=1}^{n}X_i=n\bar{X}$

很多人对于样本方差除以 n-1 有疑问，看了前面的内容应该有所了解了，其实就是为了满足无偏性，不信的话可以试试求样本方差的期望是否等于总体方差（后面会给出推导过程）。还有一种解释是，这可以认为是一种“修正”或者“调整”，修正的依据是自由度，在均值给定的条件下，只要知道 n-1 个样本，必然可以推算出第 n 个样本值来，因此自由度为 n-1，这里可以把自由度理解为限制条件（本质上就是矩阵里的秩）。

了解了样本均值和样本方差这两个统计量，再推导一下它俩的性质：

设总体X的期望$EX=\mu$，方差$DX=\sigma^2$，$X_1,\dots,X_n$是取自总体X，容量为 n 的一个样本，$\bar{X}$，S^2分别为样本的均值和方差，则

$EX_i=\mu, DX_i=\sigma^2(i=1,2,3\cdots,n)$，

$$E\bar{X}=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{n\mu}{n}=\mu$$

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

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

$$=\frac{1}{n-1}\sum_{i=1}^{n}\left[DX_i+\left(EX_i\right)^2\right]-\frac{n}{n-1}\left[D\bar{X}+\left(E\bar{X}\right)^2\right]$$

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

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

好了，讲了已经够多了，就先到这吧，下次再谈数理统计基本概念。