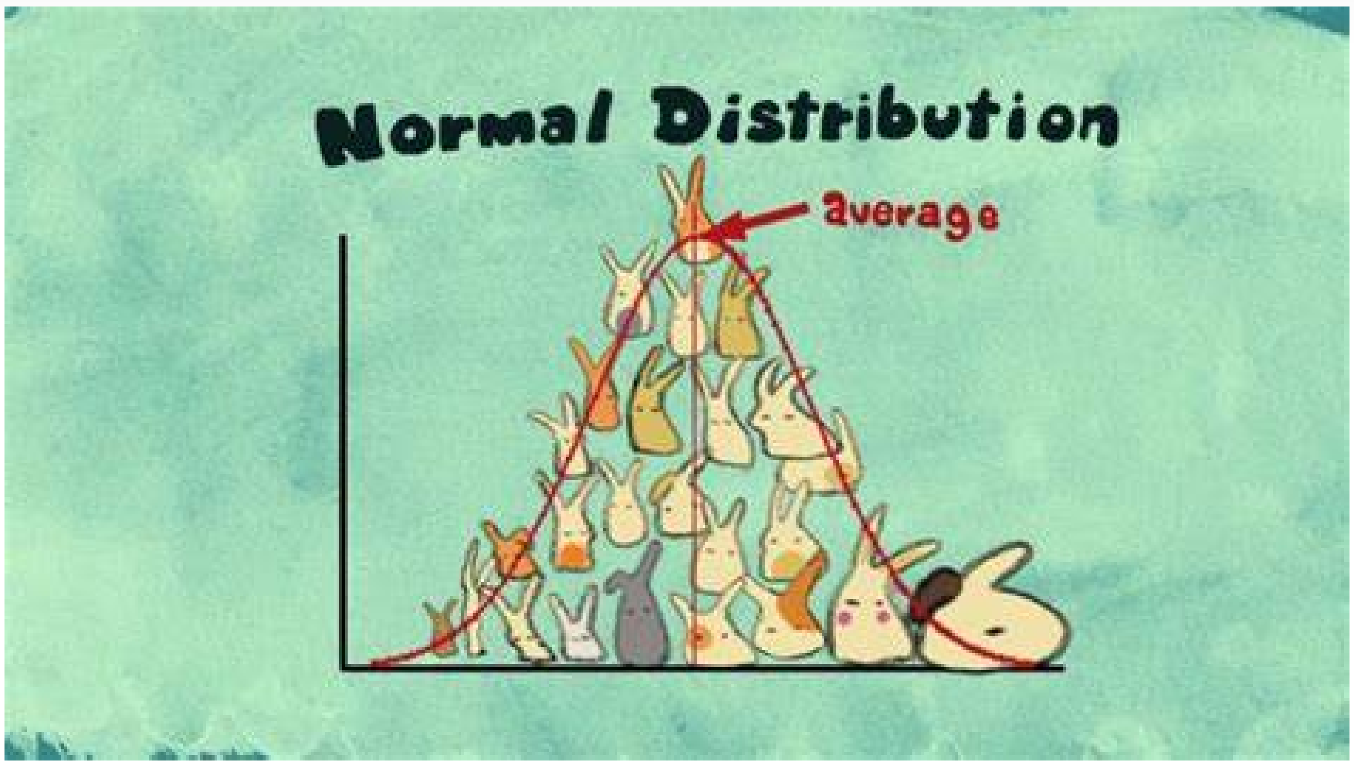

深入浅出|中心极限定理（Central Limit Theorem）及证明

荒野求生

在介绍统计学中最重要的定理之一-中心极限定理-之前，我们先来想一个问题：统计学的目的是什么？根据<*Mathematical statistics with application 7th Edition*>书中所写的：

> 统计学的目的是基于从总体中的样本所获得的信息，对总体进行推断，并且提供推断的准确性。

这其中有几个关键词：**总体**，**样本**，**推断**。总体的含义就是所研究对象的所有可能的数据，比如，全世界每个人的身高，工厂上个月生产出来的每个灯泡的寿命等等。样本的概念是从总体中衍生出来的，比如，全世界任意20个人的身高，工厂上个月任意100个灯泡的寿命，样本就是总体的一个子集。通常情况下总体的数据是难以获得的，而样本是容易得到的，所以统计学的目的就是从样本数据来推断总体。

接下来我们通过一个实际例子来介绍中心极限定理：

> 一个工厂所生产的灯泡的平均寿命是1000小时，方差是25个小时。我买了36个灯泡装在家里，那么这个9个灯泡的平均寿命超过1005小时的概率是多少

求解问题的第一步是将实际问题抽象出数学模型。在实际应用中，数学模型的建立远远要比解题方法更重要。首先定义随机变量：

$$Y_i = \text{第}i\text{个灯泡的使用寿命}，\quad i = 1,2,3\ldots36$$

由题中信息可得： $E\left(Y_i\right) = 1000, V\left(Y_i\right) = 25$

$$\bar{Y} = 36\text{个灯泡寿命的平均值}，\quad \bar{Y} = \frac{1}{36}\sum_{i=1}^{36} Y_i$$

那么问题的求解转化为求解 $P(\bar{Y} > 1005)$ 。

我们并不知道随机变量 $\bar{Y}$ 的概率分布，因此我们需要将其转化为概率分布已知的随机变量，这里就需要用到中心极限定理：

> 设定有 n 个独立且完全相同的随机变量 $Y_1, Y_2, Y_3, \ldots, Y_n$ ，他们的期望 $E(Y_i) = \mu$ ，方差 $V(Y_i) = \sigma^2$ 。定义随机变量：

> $$U_n = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
> 那么，当 n 趋向于无穷大时，随机变量 U_n 趋向于标准正态分布。

回到我们的问题，求解 $P(\bar{Y} > 1005)$ ，我们并不清楚 $\bar{Y}$ 这个随机变量的概率分布，但是根据中心极限定理，我们知道 $P(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}})$ 的概率分布近似为标准正态分布（当 n 足够大时），其中 μ 为总体的均值，在此题中为1000， σ 为标准差，在此题中为5， n 为样本数量，在此题中为样本灯泡的数量36。那么有：

$$P(\bar{Y} > 1100) = P(\frac{\bar{Y}-1000}{5/\sqrt{36}} > \frac{1005-1000}{5/\sqrt{36}}) = P(U_{36} > 6)$$

其中 U_{36} 近似服从标准正态分布 N（0，1），从可以求得 $P(U_{36} > 3)$ 的概率。

中心极限定理实际上是揭示了任意一个总体中样本均值的分布规律。

通常在教科书中，在描述完中心极限定理后，会出现3个字：证明略。接下来本文使用随机变量特征函数的方式来对其进行证明。

首先，引入随机变量特征函数的概念。对于随机变量 Y ，定义其特征函数为：

$\varphi(t) = E(e^{itY})$ ，其中 t 为任意实数。

那么对于 U_n ，其特征函数为：

$$U_n = \frac{n\bar{Y}-\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n}\eta_i}{\sigma\sqrt{n}}, \quad \eta_i = Y_i - \mu$$

$$\varphi(t) = E(e^{itU_n}) = E(e^{it\frac{\eta_1}{\sigma\sqrt{n}}} \cdot e^{it\frac{\eta_2}{\sigma\sqrt{n}}} \cdot \ldots \cdot e^{it\frac{\eta_n}{\sigma\sqrt{n}}}) = \left[\phi(\frac{t}{\sigma\sqrt{n}})\right]^n$$

$\phi(t)$ 为 η_i 的特征函数。

$\phi(\frac{t}{\sigma\sqrt{n}})$ 在0点处的泰勒展开形式为：

$$\begin{aligned}\phi(\frac{t}{\sigma\sqrt{n}}) &= \phi(0) + \phi'(0)\frac{t}{\sigma\sqrt{n}} + \frac{\phi''(0)}{2!}(\frac{t}{\sigma\sqrt{n}})^2 + o((\frac{t}{\sigma\sqrt{n}})^2) \\ &= 1 + 0 - \frac{t^2}{2n} + o((\frac{t}{\sigma\sqrt{n}})^2)\end{aligned}$$

所以， $\varphi(t)$ 为：

$$\varphi(t) = \left(1 - \frac{t^2}{2n} + o((\frac{t}{\sigma\sqrt{n}})^2)\right)^{(-\frac{2n}{t^2})\times(-\frac{t^2}{2})} = e^{-\frac{t^2}{2}}, n \to +\infty$$

而标准正态分布 $N(0,1)$ 特征函数也为 $e^{-\frac{t^2}{2}}$ ，根据特征函数的唯一性定理，所以 $U_n, n \to +\infty$ 服从标准正态分布，证毕。

编辑于 2020-09-04 10:30