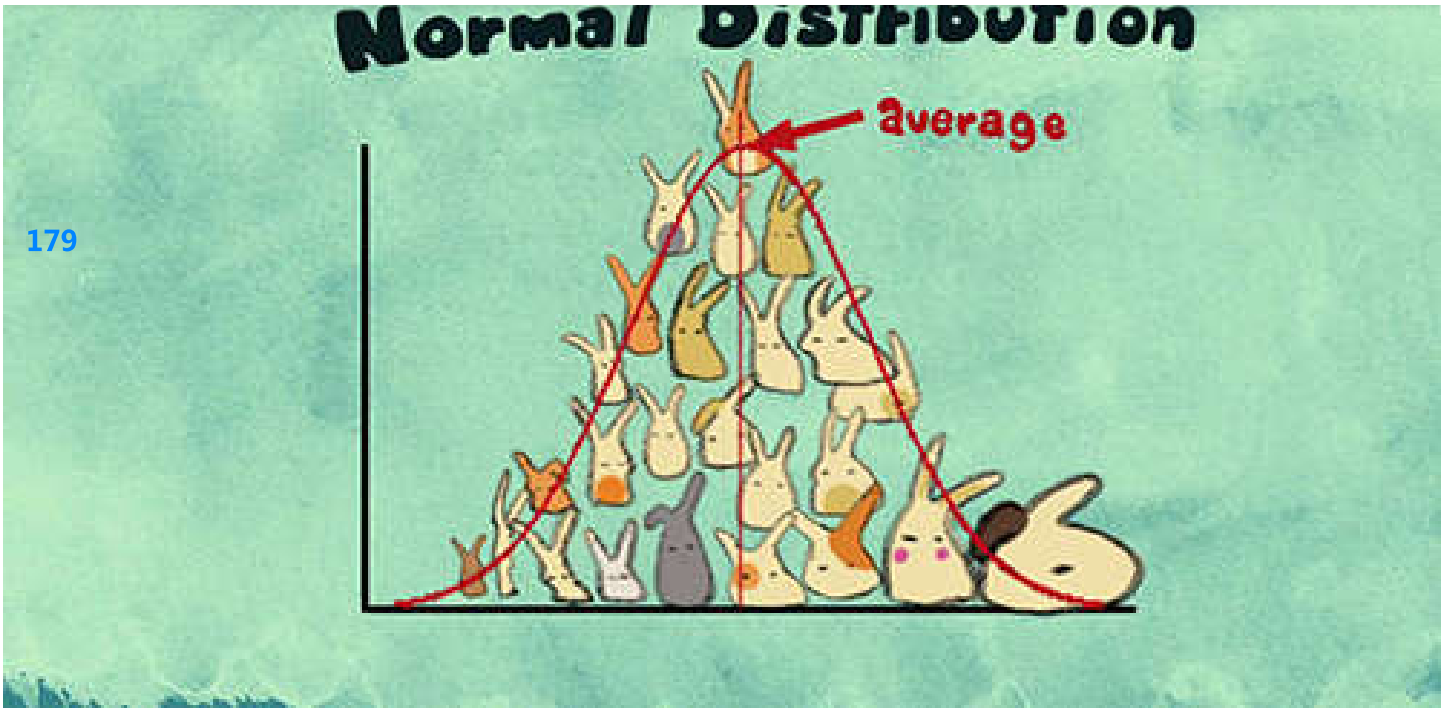


179

中心极限定理通俗介绍

朱雀
为了以后不用介绍自己而努力

关注他

179 人赞了该文章

中心极限定理是统计学中比较重要的一个定理。 本文将通过实际模拟数据的形式，形象地展示中心极限定理是什么，是如何发挥作用的。

什么是中心极限定理（Central Limit Theorem）

> 中心极限定理指的是给定一个任意分布的总体。我每次从这些总体中随机抽取 n 个抽样，一共抽 m 次。 然后把这 m 组抽样分别求出平均值。 这些平均值的分布接近正态分布。

我们先举个**栗子**

现在我们要统计全国的人的体重，看看我国平均体重是多少。当然，我们把全国所有人的体重都调查一遍是不现实的。所以我们打算一共调查1000组，每组50个人。 然后，我们求出第一组的体重平均值、第二组的体重平均值，一直到最后一组的体重平均值。中心极限定理说：这些平均值是呈现正态分布的。并且，随着组数的增加，效果会越好。 最后，当我们再把1000组算出来的平均值加起来取个平均值，这个平均值会接近全国平均体重。

其中要注意的几点：

每组的平均值也会组成一个正态分布。（神奇！）

2. **样本每组要足够大，但也不需要太大**
   取样本的时候，一般认为，每组大于等于30个，即可让中心极限定理发挥作用。

179
话不多说，我们现在来一步步**看到**中心极限定理是如何起作用的。

用实际数据来展示中心极限定理

注：我们使用python语言以及iPython Notebook来生成和展现数据。不懂的童鞋可以略过代码

第一步， 生成数据

假设我们现在观测一个人掷骰子。这个骰子是公平的，也就是说掷出1~6的概率都是相同的：1/6。他掷了一万次。我们用python来模拟投掷的结果：

```
import numpy as np
random_data = np.random.randint(1, 7, 10000)
print random_data.mean() # 打印平均值
print random_data.std()  # 打印标准差
```

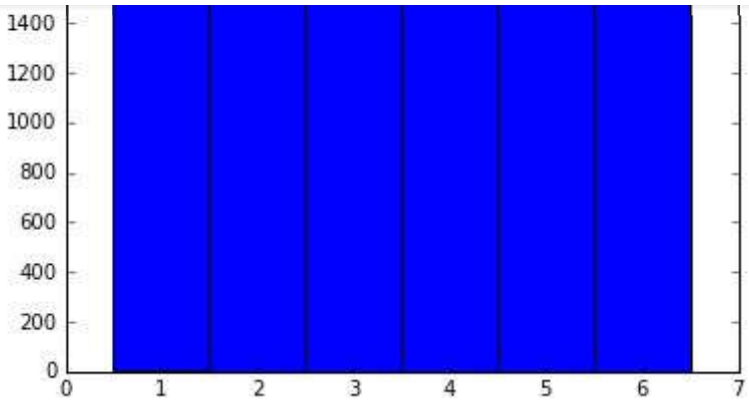
生成出来的平均值：**3.4927***（每次重新生成都会略有不同）*
生成出来的标准差：**1.7079**

平均值接近3.5很好理解。 因为每次掷出来的结果是1、2、3、4、5、6。 每个结果的概率是1/6。所以加权平均值就是3.5。

第二步，画出来看看

我们把生成的数据用直方图画出来直观地感受一下：

179



可以看到1~6分布都比较平均，不错。

第三步，抽一组抽样来试试

我们接下来随便先拿一组抽样，手动算一下。例如我们先从生成的数据中随机抽取10个数字：

```
sample1 = []
for i in range(0, 10):
    sample1.append(random_data[int(np.random.random() * len(random_data))])

print sample1 # 打印出来
```

这10个数字的结果是： **[3, 4, 3, 6, 1, 6, 6, 3, 4, 4]**
平均值：**4.0**
标准差：**1.54**

可以看到，我们只抽10个的时候，样本的平均值（4.0）会距离总体的平均值（3.5）有所偏差。有时候我们运气不好，抽出来的数字可能偏差很大，比如抽出来10个数字都是6。那平均值就是6了。 为什么会出现都是6的情况呢？因为我比较6...哦不是，因为这就是随机的魅力呀！

不过不要担心，接下去就是见证奇迹的时刻。

第四步，见证奇迹的时刻

我们让中心极限定理发挥作用。现在我们抽取1000组，每组50个。
我们把每组的平均值都算出来。

```
samples = []
samples_mean = []
samples_std = []
```

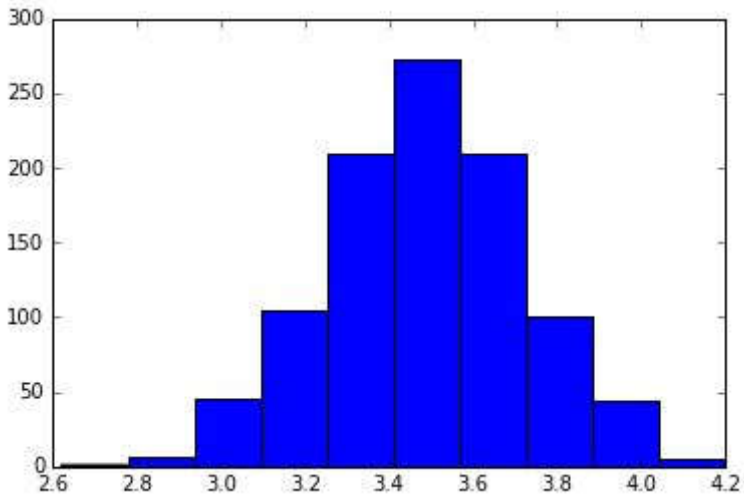
```
    for j in range(0, 50):
        sample.append(random_data[int(np.random.random() * len(random_data))])
    sample_np = np.array(sample)
    samples_mean.append(sample_np.mean())
    samples_std.append(sample_np.std())
    samples.append(sample_np)

samples_mean_np = np.array(samples_mean)
samples_std_np = np.array(samples_std)

print samples_mean_np
```

这一共1000个平均值大概是这样的：**[3.44, 3.42, 3.22, 3.2, 2.94 ... 4.08, 3.74]**（我肯定不会把1000个数字都写完，又没有稿费可以骗）

然后，我们把这1000个数字用直方图画出来：



TADA! 完美地形成了正态分布。

结果打印如下：

平均值：**3.48494**
标准差：**0.23506**

实际应用

在实际生活当中，我们不能知道我们想要研究的对象的平均值，标准差之类的统计
定理在理论上保证了我们可以用只抽样一部分的方法，达到推测研究对象统计参数

本文最先发布于朱曦炽个人博客：中心极限定理通俗介绍

编辑于 2017-03-10

179

「求求你，可怜可怜我吧」

赞赏

8 人已赞赏

中心极限定理

文章被以下专栏收录

Quant的碎碎念

欢迎来到Quant的碎碎念。这里是我在研究Quant中所产生的一些研究结果、感悟等...

关注专栏

推荐阅读

抽样分布篇之二：中心极限定理

在认识抽样分布之前，要先了解一下中心极限定理。 虽然在数理统计的教科书中，在讲完随机变量的分布和数字特征后才开始讲中心极限定理，但实际中心极限定理的提出和应用却早于正态分布，而...

张自达 发表于张老师漫谈...

写下你的评论...


哈里 10 个月前

179

您好！很感谢你的讲解。直观形象。但是我还是不理解一个问题：当次数m比较大的时候，对于抽取样本的n个值求平均，在对m求平均，可以看做是总体均值的估计。但为什么，当m=1的时候也可以这样近似估计呢。

赞


哈里 10 个月前

不好意思 我好像搞明白了，是每次抽样分布的均值不是抽样的均值

赞

张振宇 10 个月前

但是为什么会这样呢？

1


朱雀 (作者) 回复 张振宇 10 个月前

问得好！我也不知道呢。。。可以网上搜搜看

赞 查看对话

张振宇 回复 朱雀 (作者) 10 个月前

我就是在找的过程中看到了你的文章，真的很难找啊，看来得去发帖问了☹

赞 查看对话

mgc 10 个月前

请问一下，那个栗子中。1000组算出来的平均值加起来取个平均值与直接求1000*50人的平均值应该是一样的吧？

1


朱雀 (作者) 回复 mgc 10 个月前

对呀

赞 查看对话


方鸿渐师弟 9 个月前

1

汨罗江畔的顽石 回复 张振宇 7 个月前

我觉得是因为切比雪夫不等式啊，掷筛子有他的期望3.5，每组实验的期望满足切比雪夫不等式，所以每组实验期望在3.5附近概率大，越远离3.5，概率越小。

2 查看对话

Brooklyn 7 个月前

这是我看过最好的中心极限定理解释，可惜不能充值知乎币，给赞好啦

赞

旅人 7 个月前

感谢你的讲解。有一个问题想问一下，比如二项分布以正态分布为极限分布的那个定理，近似成正态分布后在事件发生次数小于0的范围内概率不为0，但是实际上事件发生次数应该大于等于0，这个情况应该怎么理解呢

赞

朱雀 (作者) 回复 旅人 7 个月前

你理解为这是近似的就可以了。

赞 查看对话

吹雪菠萝 回复 张振宇 7 个月前

受答主启发，最后为何会成正态，我认为直观上可以这样理解:
掷筛子，每组五十次，一共1000组，
而掷筛子本身服从均匀分布，于是可以算出总体期望为3.5。据大数定律所以肯定有很多很多很多的组的样本均值接近（收敛于）3.5。而其它类似样本均值为5或2的就比较少。
于是从图上看，越接近3.5，组的数目就越大，于是看上去就成为了正态分布。

3 查看对话

少思凡 回复 吹雪菠萝 7 个月前

赞同！通过大数定理可以更直观地理解中心极限定理！

赞 查看对话

上个注册 5 个月前

谢谢楼主的讲解，你提到一般每组大于30个就可以让中心极限定理发挥作用，也就是m > 30，那么请问n呢？也就是说需要多少组呢？

赞

m>30之后，可以近似认为sample的均值符合正态分布。和n无关。n越多只是你越来越能”看见“正态分布的样子而已。

1 查看对话

179 st小菜鸟 4 个月前

大侠，请教下，直方图怎么画出来的？不是在Python吧？

赞

朱雀 (作者) 回复 st小菜鸟 4 个月前

当然是在python咯

赞 查看对话

黎再子 4 个月前

写得很好，我是看一篇叫“正态分布的前世今生”有点看不懂来知乎一下，下面有人评论说为什么会这样，大概看完那篇东西会有更全面的理解哈。

赞

Juper 回复 朱雀 (作者) 3 个月前

楼主摆个m=10的统计结果出来就更清楚了

赞 查看对话