



下载APP



01 | 统计基础（上）：系统掌握指标的统计属性

2020-12-02 张博伟

A/B测试从0到1

[进入课程 >](#)**讲述：张博伟**

时长 15:06 大小 13.84M



你好，我是博伟。

在学习、解决技术问题的时候，我们都知道有这么一句话“知其然知其所以然”。那么，A/B 测试的“所以然”是什么呢？在我看来，就是 A/B 测试背后的计算原理，知道 A/B 测试为什么要这么设计，最佳实践中为什么要选择这样的指标、那样的检验方法。

那说到 A/B 测试背后的计算原理，我们首先得知道，A/B 测试的理论基础是假设检验（Hypothesis Testing）。可以说，假设检验，贯穿了 A/B 测试从实验设计到分析测试结果整个流程。



如果要一句话解释“假设检验”的话，就是选取一种合适的检验方法，去验证在 A/B 测试中我们提出的假设是否正确。现在，你只要知道“假设检验”中，最重要也最核心的

是“检验”就可以了，因为选取哪种检验方法，取决于指标的统计属性。

也就是说，理解指标的统计属性，是我们掌握假设检验和 A/B 测试的前提，也是“知其所以然”的第一步。

而至于深入理解并用好“假设检验”的任务，我们就留着下一讲去完成吧。

指标的统计属性，指的是什么？

在实际业务中，我们常用的指标其实就是两类：

均值类的指标，比如用户的平均使用时长、平均购买金额、平均购买频率，等等。

概率类的指标，比如用户点击的概率（点击率）、转化的概率（转化率）、购买的概率（购买率），等等。

很明显，这些指标都是用来表征用户行为的。而用户的行为是非常随机的，这也就意味着这些指标是由一系列随机事件组成的变量，也就是统计学中的随机变量（Random Variable）。

“随机”就代表着可以取不同的数值。比如，一款社交 App 每天的使用时间，对轻度用户来说可能不到 1 小时，而对重度用户来说可能是 4、5 小时以上。那么问题来了，在统计学中，怎么表征呢？

没错，我们可以用**概率分布（Probability Distribution）**，来表征随机变量取不同值的概率和范围。所以，A/B 测试指标的统计属性，其实就是要看这些指标到底服从什么概率分布。

在这里，我可以先告诉你结论：**在数量足够大时，均值类指标服从正态分布；概率类指标本质上服从二项分布，但当数量足够大时，也服从正态分布。**

看到这两个结论你可能会很多问题：

什么是正态分布？什么是二项分布？

“数量足够大”具体是需要多大的数量？

概率类指标，为什么可以既服从二项分布又服从正态分布？

不要着急，我这就来——为你解答。

正态分布 (Normal Distribution)

正态分布是 A/B 测试的指标中最主要的分布，是计算样本量大小和分析测试结果的前提。

在统计上，如果一个随机变量 x 的概率密度函数 (Probability Density Function) 是：

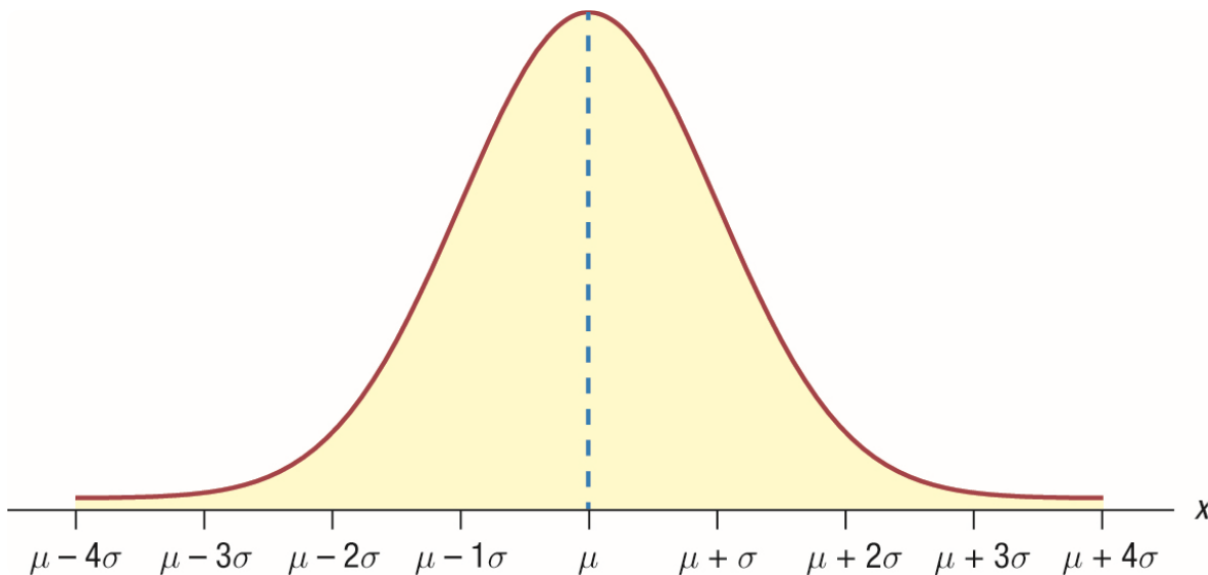
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{n}}$$

那么， x 就服从正态分布。

其中， μ 为 x 的平均值 (Mean)， σ 为 x 的**标准差** (Standard Deviation)， n 为随机变量 x 的个数， x_i 为第 i 个 x 的值。

随机变量 x 服从正态分布时的直方图 (Histogram) 如下：




直方图是表征随机变量分布的图表，其中横轴为 x 可能的取值，纵轴为每个值出现的概率。通过直方图你可以看到，**距离平均值 μ 越近的值出现的概率越高。**

除了平均值 μ ，你还能在直方图和概率密度函数中看到另一个非常重要的参数：**标准差 σ** 。 σ 通过计算每个随机变量的值和平均值 μ 的差值，来表征随机变量的离散程度（偏离平均值的程度）。

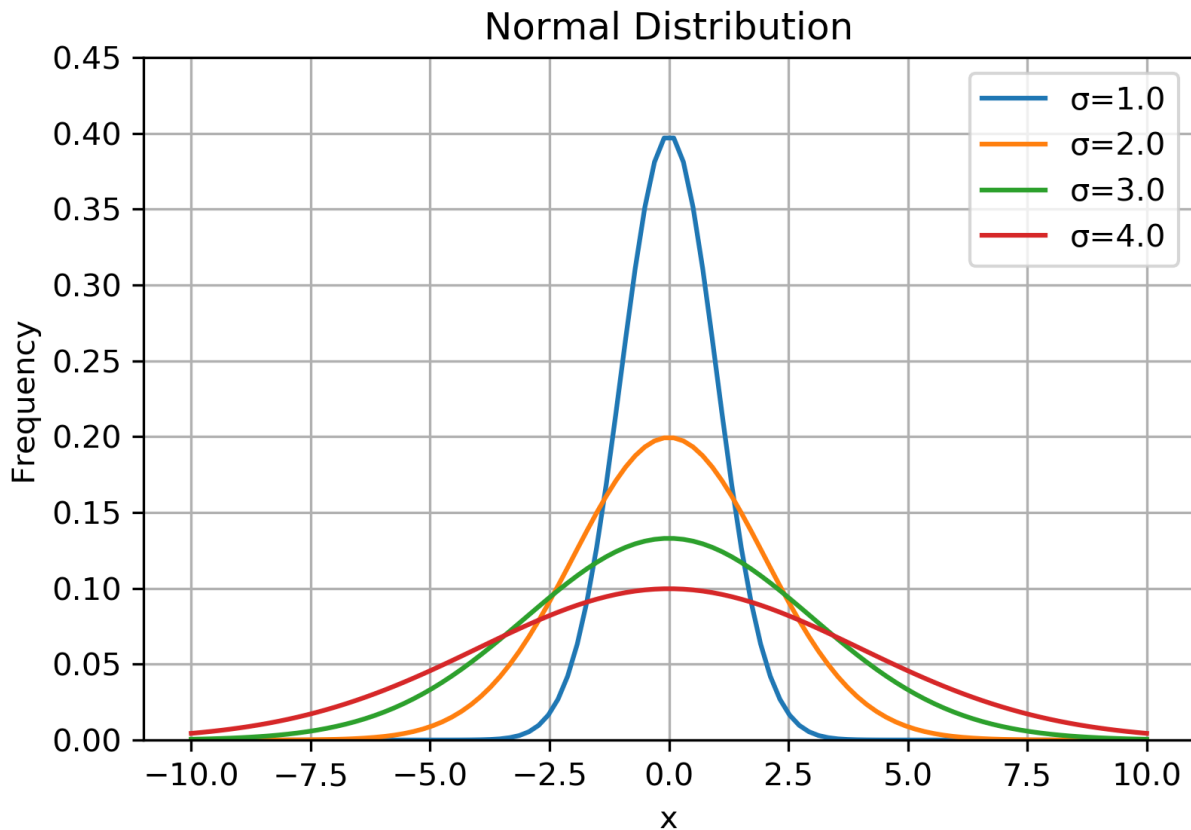
接下来，我们就来看看标准差 σ 是怎么影响随机变量的分布的。

为了方便理解，我们用 Python 做一个简单的模拟，选取服从正态分布的随机变量 x ，其平均值 $\mu=0$ ；分别把 x 的标准差 σ 设置为 1.0、2.0、3.0、4.0，然后分别做出直方图。对应的 Python 代码和直方图如下：

 复制代码

```
1 from scipy.stats import norm
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 ## 构建图表
6 fig, ax = plt.subplots()
7 x = np.linspace(-10,10,100)
8 sigma = [1.0, 2.0, 3.0, 4.0]
9 for s in sigma:
10     ax.plot(x, norm.pdf(x,scale=s), label='σ=%.1f' % s)
11
12 ## 加图例
13 ax.set_xlabel('x')
14 ax.set_ylabel('Frequency')
15 ax.set_title('Normal Distribution')
```

```
16 ax.legend(loc='best', frameon=True)
17 ax.set_ylim(0,0.45)
18 ax.grid(True)
```



通过这个直方图去看标准差 σ 对随机变量分布的影响，是不是就更直观了？ σ 越大， x 偏离平均值 μ 的程度越大， x 的取值范围越广，波动性越大，直方图越向两边分散。

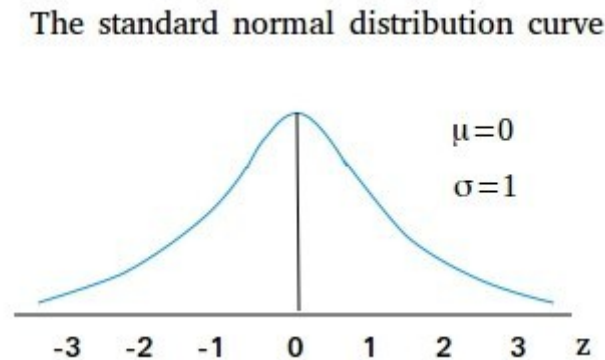
咱们再举个生活中的例子来理解标准差。在一次期末考试中，有 A 和 B 两个班的平均分都是 85 分。其中，A 班的成绩范围在 70~100 分，通过计算得到成绩的标准差是 5 分；B 班的成绩范围在 50~100 分，计算得到的成绩标准差是 10 分。你看，A 班的成绩分布范围比较小，集中在 85 分左右，所以标准差也就更小。

说到标准差，你应该还会想到另一个用来表征随机变量离散程度的概念，就是**方差** (Variance)。其实，方差就是标准差的平方。所以，标准差 σ 和方差在表征离散程度上其实是可以互换的。

有了方差和标准差，我们就可以描述业务指标的离散程度了，但要计算出业务指标的波动范围（我会在第 4 讲展开具体的计算方法），我们还差一步。这一步就是 z 分数。

要解释 z 分数，就要引出一种特殊的正态分布，也就是标准正态分布（Standard Normal Distribution），其实就是平均值 $\mu=0$ 、标准差 $\sigma=1$ 的正态分布。

标准正态分布的直方图如下所示：



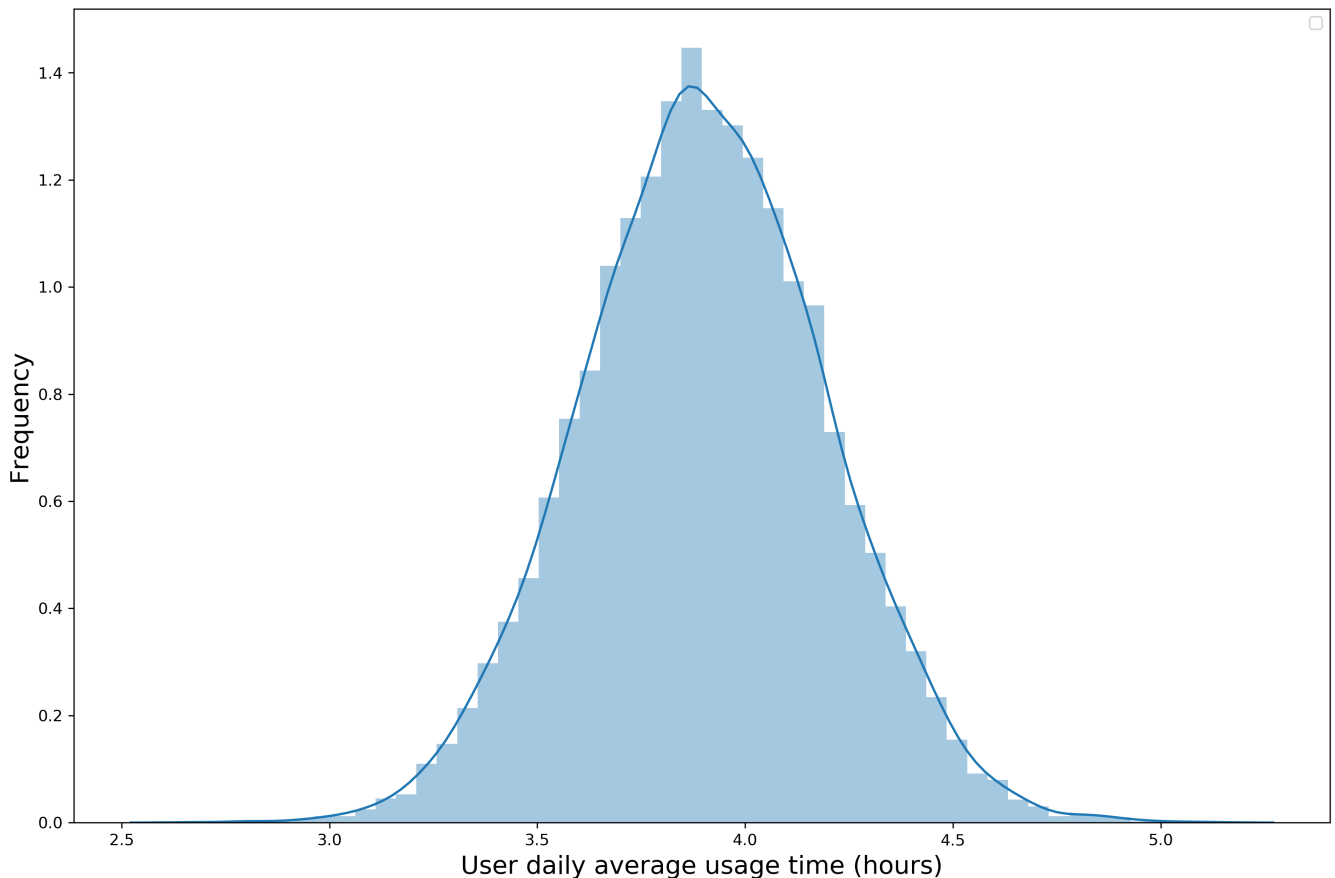
这里的横轴就是 z 分数（Z Score），也叫做标准分数（Standard Score）：

$$z \text{ score} = \frac{x - \mu}{\sigma}$$

事实上，任何一个正态分布都可以通过标准化（Standardization）变成标准正态分布。而标准化的过程，就是按照上面这个公式把随机变量 x 变为 z 分数。不同 z 分数的值，代表 x 的不同取值偏离平均值 μ 多少个标准差 σ 。比如，当 z 分数等于 1 时，说明该值偏离平均值 1 个标准差 σ 。

我们再用一个社交 App 业务指标的例子，来强化下对正态分布的理解。

现在有一个社交 App，我们想要了解用户日均使用时间 t 的概率分布。根据现有的数据，1 万个用户在一个个月内每天使用 App 的时间，我们做出了一个直方图：



可以看出，这 1 万个用户的日均使用时间 t ，大约在 3-5 小时这个范围，而且是近似正态分布的钟形曲线，说明 t 的分布也可以近似为正态分布。

中心极限定理 (Central Limit Theorem)

这其实是均值类变量的特性：当样本量足够大时，均值类变量会趋近于正态分布。这背后的理论基础，就是中心极限定理。

🔗 **中心极限定理**的数学证明和推理过程十分复杂，但不用害怕，我们只要能理解它的大致原理就可以了：**不管随机变量的概率分布是什么，只要取样时的样本量足够大，那么这些样本的平均值的分布就会趋近于正态分布。**

那么，这个足够大的样本量到底是多大呢？

统计上约定俗成的是，样本量大于 30 就属于足够大了。在现在的大数据时代，我们的样本量一般都能轻松超过 30 这个阈值，所以均值类指标可以近似为正态分布。

到这里，“数量足够大”具体是需要多大的数量，以及什么是正态分布，这两个问题我们就都明白了。接下来，我们再学习下什么是二项分布，之后我们就可以理解为什么概率类

指标可以既服从二项分布又服从正态分布了。

二项分布 (Binomial Distribution)

业务中的概率类指标，具体到用户行为时，结果只有两种：要么发生，要么不发生。比如点击率，就是用来表征用户在线上点击特定内容的概率，一个用户要么点击，要么不点击，不会有第三种结果发生。

这种只有两种结果的事件叫做二元事件 (Binary Event)。二元事件在生活中很常见，比如掷硬币时只会出现正面或者反面这两种结果，所以统计学中有专门有一个描述二元事件概率分布的术语，也就是**二项分布** (Binomial Distribution)。

这里我们还是结合着社交 App 的例子，来学习下二元分布。

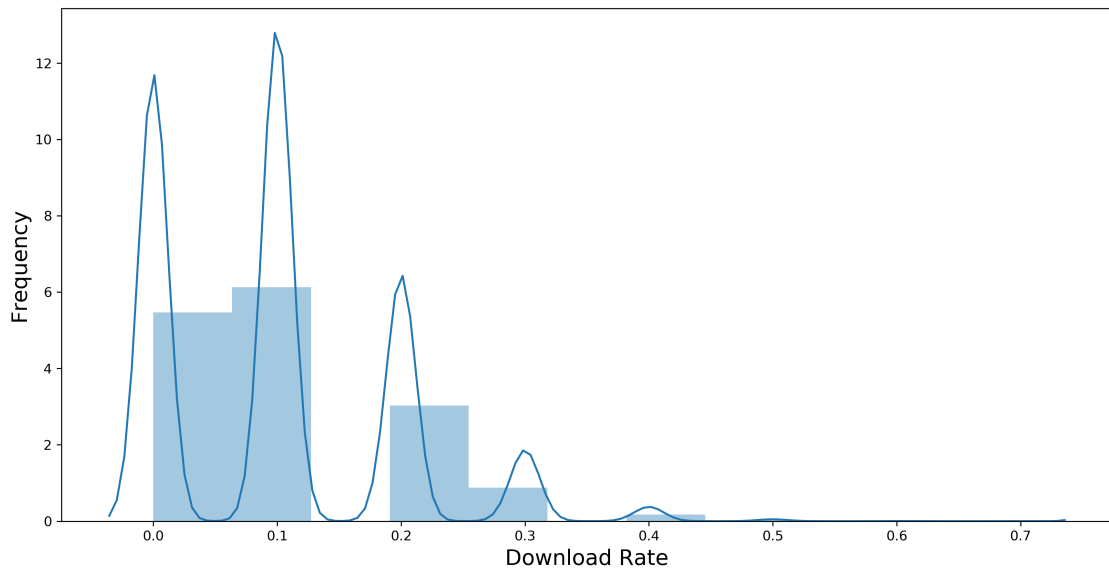
这款社交 App 在网上投放了广告，来吸引人们点击广告从而下载 App。现在我们想通过数据看看 App 下载率的分布情况：

下载率 = 通过广告下载 App 的用户数量 / 看到广告的用户数量。

因为单个二元事件的结果，只能是发生或者不发生，发生的概率要么是 100% 要么是 0%，所以我们要分析下载率就必须把数据进行一定程度的聚合。这里，我们就以分钟为单位来举例，先计算每分钟的下载率，再看它们的概率分布。

我们有一个月的用户及下载数据，一个月一共有 43200 分钟 (60*24*30)，因为我们关注的是每分钟的下载率，所以一共有 43200 个数据点。通过数据分析发现，每分钟平均有 10 个人会看到广告，下载率集中分布在 0-30% 之间。

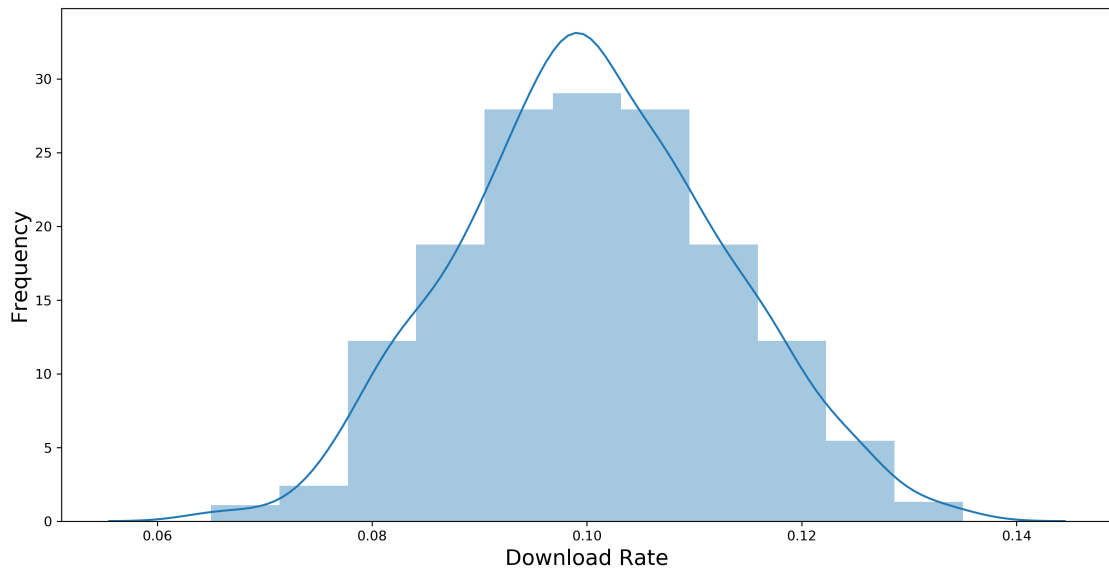
下图是每分钟下载率的概率分布：



你可能会说，概率在某种程度上也是平均值，可以把这里的下载率理解为“看到广告的用户平均下载量”，那我们已经有 43200 个数据点了，样本量远远大于 30，但为什么下载率的分布没有像中心极限定理说的那样趋近于正态分布呢？

这是因为在二项分布中，中心极限定理说的样本量，指的是计算概率的样本量。在社交 App 的例子中，概率的样本量是 10，因为平均每分钟有 10 人看到广告，还没有达到中心极限定理中说的 30 这个阈值。所以，我们现在要提高这个样本量，才能使下载率的分布趋近正态分布。

提高样本量的方法也很简单，可以计算每小时的下载率。因为每小时平均有 600 人看到广告，这样我们的样本量就从 10 提高到了 600。下图是每小时下载率的概率分布：



现在再看这张直方图，每小时下载率的分布是不是就趋近于正态分布了！图中下载率的平均值大约为 10%。

在二项分布中，有一个从实践中总结出的经验公式： $\min(np, n(1-p)) \geq 5$ 。其中， n 为样本大小， p 为概率的平均值。

这个公式的意思是说， np 或者 $n(1-p)$ 中相对较小的一方大于等于 5，只有二项分布符合这个公式时，才可以近似于正态分布。这是中心极限定理在二项分布中的变体。

在我们的例子中，计算每分钟下载率的概率分布时， $np=10*10\%=1$ ，小于 5，所以不能近似成正态分布；计算每小时下载率的概率分布时， $np=600*10\%=50$ ，大于等于 5，所以可以近似成正态分布。

我们可以利用这个公式来快速判断概率类指标是不是可以近似成正态分布。不过你也可以想象在实践中的 A/B 测试，由于样本量比较大，一般都会符合以上公式的。

小结

今天这节课，我们主要学习了 A/B 测试和假设检验的前提，也就是指标的统计属性。我给你总结成了一个定理、两个分布和三个概念：

1. 一个定理：中心极限定理。

2. 两个分布：正态分布和二项分布。
3. 三个概念：方差，标准差和 z 分数。

生活中随机变量的分布有很多种，今天我重点给你介绍了正态分布和二项分布，它们分别对应的是最普遍的两类业务指标：均值类和概率类。

而且你要知道，有了中心极限定理，我们就可以把业务中的大部分指标都近似成正态分布了。这一点非常重要，因为 A/B 测试中的很多重要步骤，比如计算样本量大小和分析测试结果，都是以指标为正态分布为前提的。

同时，你还可以用通过方差和标准差来了解业务指标的离散程度，再结合 z 分数就可以计算出业务指标的波动范围了。只有理解了指标的波动范围，才能够帮助我们得到更加准确的测试结果。

在下节课中，我们继续学习 A/B 测试的统计基础，也就是假设检验及其相关的统计概念。

思考题

我在刚开始接触概率类指标的二项分布时对于其如何能近似成正态分布很迷惑，大家可以在这里聊一聊在学习 A/B 测试的统计过程中有什么难理解的地方，以及如何解决的？

欢迎在留言区写下你的思考和想法，我们可以一起交流讨论。如果你觉得有所收获，欢迎你把课程分享给你的同事或朋友，一起共同进步！

提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 开篇词 | 用好A/B测试，你得这么学

精选留言 (9)

[写留言](#)**那一刻** 置顶

2020-12-03

请问老师两个问题

1.每分钟下载率是怎么计算的呢？我的理解是每分钟内下载人数除以观看人数。但是老师提到每分钟平均有 10 个人会看到广告，所以是以每分钟平均下载人数除以每分钟观看人数么？

...

展开 ∨

作者回复: 你好，对于第一个问题你的理解是正确的，这里的10是分母；对于第二个问题你的理解也是对的，每分钟和每小时的下载率的平均值都是10%，这里因为时间长度不同，下载量不同，但是下载率是相同的。



👍 1

**KayeArt**

2020-12-03

希望老师推荐一些AB测试的论文

展开 ∨

作者回复: 这两篇KDD的文章不错：

<https://exp-platform.com/Documents/2014%20experimentersRulesOfThumb.pdf>

<https://exp-platform.com/Documents/2009-ExpPitfalls.pdf>



👍 1

**何涛(Louis)**

2020-12-03

想问问张老师有没有推荐的参考读物或者网站？多谢！

展开 ∨

作者回复: 这里有一些比较好的英文书籍推荐：

<https://www.alex-birkett.com/ab-testing-books/>



👍 1





fei 20-12-03

求问老师,a/b 测试一般是产品主导还是测试主导呢?因为测试很难及时的获取到一些数据指标,并且也很难制定指标标准,但是产品又不太了解技术实现.

展开 ∨



那一刻

2020-12-03

不知老师可不可以把文中提到的app一个月的用户及下载数据,脱敏后提供一下? 这样方便我们通过数据以及公式看文中的图

展开 ∨

作者回复: 你好, 不好意思这个是实际的用户数据不太方便提供, 不过你可以用Python模拟出参数不同(比如文中提到的每分钟和每小时的情况)的二项分布数据, 然后分别画出直方图, 你就知道二项分布是如何近似成正态分布的啦。



四月. 🐼

2020-12-03

不明白"每分钟下载率的概率分布图"中纵轴的frequency的含义

作者回复: 你好, Frequency (频率) 在这里不是很好理解, 你可以把它近似理解为是x轴上各个下载率的值的出现的次数占总次数的比例, Y的值越大, 说明相对应的下载率(X的值)出现的次数越多, 但是如果你仔细计算会发现Y轴的值有的会超过100%, 因为Y轴的计算还涉及到图中每个小长方体的组距, 具体可以参考这里: <https://baike.baidu.com/item/%E9%A2%91%E7%8E%87%E5%88%86%E5%B8%83%E7%9B%B4%E6%96%B9%E5%9B%BE>

所以我说是近似理解。不过其实你只要掌握近似的意思即可, 这里的数学细节不是本节课的重点。



勇哥

2020-12-03

急啊 目前负责增长, 缺A/B测试工具, 又要给技术讲业务需求



1

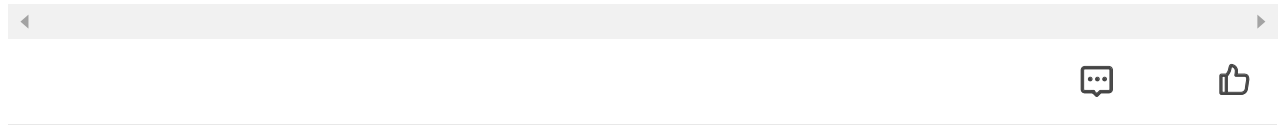


Geek_6e25ac

2020-12-03

作为没有数学基础的学渣，完全听不懂。各种公式听得一脸蒙。有适合产品经理学习的深入浅出的课吗？

作者回复: 可以看一下目录哦，就前两节课是理论，因为学习A/B测试肯定要学一些理论的，后面就全都是实战啦，老师讲得很深入浅出的。



愤怒的熊要吃了你

2020-12-02

在看完老师关于二项分布的描述后确实是懵了，看了好久的不是讲二项分布吗，咋讲着讲着变成正态分布了...

而且老师举的那个例子中“每分钟下载率的概率分布图”我也没看懂，主要是没看懂横纵轴之间的关系。...

展开 ∨

作者回复: 本节课二项分布里的重点是在样本量大的情况下可以近似成正态分布，因为它也是遵从中心极限定理的。感兴趣的话可以查询下相关的统计概念，学习精神可嘉！

