



下载APP



## 02 | 统计基础（下）：深入理解A/B测试中的假设检验

2020-12-10 张博伟

A/B测试从0到1

[进入课程 >](#)**讲述：张博伟**

时长 19:59 大小 18.31M



你好，我是博伟。

在上节课学习 A/B 测试指标的统计属性时，我用一句话给你简单解释了下假设检验：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。

这句话其实很抽象，所以今天这一讲，我们就具体展开下，看看假设检验是什么，以及如何利用假设检验来做出推断。

### 假设检验 (Hypothesis Testing) 是什么？



假设检验，顾名思义，就是要检验我们提出的假设是不是正确的，在事实上能否成立。

在统计中，我们很难获取总体数据（Population）。不过，我们可以取得样本数据（Sample），然后根据样本数据的情况产生对总体数据的假设。所以，我们所说的假设检验，其实就是检测通过样本数据产生的假设在总体数据（即事实）上是否成立。

在 A/B 测试的语境中，假设一般是指**关于实验组和对照组指标的大小的推断**。

为了更加形象地帮你理解假设检验，这节课我就从一个推荐系统的案例出发，从中抽象出假设检验的基本原理和相关概念，让你在实践中学习理论，同时把理论应用到实践中去。

新闻 App 中的推荐系统是重要的组成部分，可以根据用户过往的浏览记录来推荐用户喜欢的内容。最近，工程团队改进了推荐系统的算法，就想通过 A/B 测试来验证改进的效果。

实验组中使用新算法，对照组中使用旧算法，然后通过点击率来表征算法的效果：推荐效果越好，点击率越高。那么，我们提出的假设就是：实验组（新算法）的点击率比对照组（旧算法）的点击率高。

你可能会有些疑惑，我们提出的“假设”，和假设检验中的“假设”是相同的吗？

其实不完全相同。

## 假设检验中的“假设”是什么？

为什么这么说呢？因为在假设检验中的“假设”是一对：零假设（Null Hypothesis）和备择假设（Alternative Hypothesis），它们是完全相反的。在 A/B 测试的语境下，零假设指的是实验组和对照组的指标是相同的，备择假设指的是实验组和对照组的指标是不同的。

为了更好地理解零假设和备择假设，我们可以回到推荐系统的案例中，把最开始提出的假设转化成假设检验中的零假设和备择假设：

零假设是，实验组和对照组的点击率是相同的。

备择假设是，实验组和对照组的点击率是不同的。

你可能会问，我们最开始提出的假设不是“实验组的点击率比对照组的点击率高”吗？为什么备择假设中仅仅说两组的点击率不同，却没说谁大谁小呢？

要回答这个问题，我们就得先了解单尾检验（One-tailed Test）和双尾检验（Two-tailed Test）这两个概念。

单尾检验又叫单边检验（One-sided Test），它不仅在假设中说明了两个比较对象不同，并且还明确了谁大谁小，比如实验组的指标比对照组的指标大。

双尾检验又叫双边检验（Two-sided Test），指的是仅仅在假设中说明了两个比较对象不同，但是并没有明确谁大谁小。

回到推荐系统案例中的最初假设，我们已经明确了实验组的点击率比对照组的高，那就应该选用单尾检验。但是，我们的备择假设却变成了两组的点击率不同，这是双尾检验的假设，为什么呢？

这就是理论和实践的不同之处，也是为什么我们觉得 A/B 测试的理论好掌握，但实践总出问题的原因。这里，我先告诉你结论，再给你说明为什么。结论是：**在 A/B 测试的实践中，更推荐使用双尾检验。**

更推荐你使用双尾检验的原因，主要有两个。

第一个原因是，双尾检验可以让数据自身在决策中发挥更大的作用。

我们在实践中使用 A/B 测试，就是希望能够通过数据来驱动决策。我们要尽量减少在使用数据前产生的任何主观想法来干扰数据发挥作用。所以，双尾检验这种不需要我们明确谁大谁小的检验，更能发挥数据的作用。

第二个原因是，双尾检验可以帮助我们全面考虑变化带来的正、负面结果。

在实践中，我们期望改变可以使指标朝着好的方向变化，但是万一指标实际的变化与期望的正好相反呢？这就可以体现双尾检验的优势了。双尾检验可以同时照顾到正面和负面的结果，更接近多变的现实情况。但是单尾检验只会适用于其中一种，而且通常是我们期望的正面效果。

所以正因为我们选择双尾测试，在备择假设中我们才只说了两组不同，并没有说谁大谁小。

## 假设检验中的“检验”都有哪些，该怎么选取？

现在，我们知道了假设检验中的“假设”包括零假设和备择假设两种，那么“检验”都包括什么呢？

其实，检验有很多种，单尾检验和双尾检验，是从“假设”的角度来分类的。除此之外，常见的“检验”还可以根据比较样本的个数进行分类，包括单样本检验（One-Sample Test）、双样本检验（Two-Sample Test）和配对检验（Paired Test）。那么问题来了，在测试中到底该选择哪种检验方法呢？

答案是：**在 A/B 测试中，使用双样本检验。**

其中的原因其实很简单，我给你解释下它们各自的适用范围，你就知道了。

当两组样本数据进行比较时，就用双样本检验。比如 A/B 测试中实验组和对照组的比较。

当一组样本数据和一个具体数值进行比较时，就用单样本检验。比如，我想比较极客时间用户的日均使用时间有没有达到 15 分钟，这个时候，我就可以把一组样本数据（抽样所得的极客时间用户的每日使用时间）和一个具体数值（15）来进行比较。

当比较同一组样本数据发生变化前和发生变化后时，就用配对检验。比如，我现在随机抽取 1000 个极客时间的用户，给他们“全场专栏一律 1 折”这个优惠，然后在这 1000 个人中，我们会比较他们在收到优惠前一个月的日均使用时间，和收到优惠后一个月的日均使用时间。

看到这里，你可能会问，我还听说过 **T 检验**（T Test）和 **Z 检验**（Z Test），那这两个检验在 A/B 测试中该怎么选择呢？

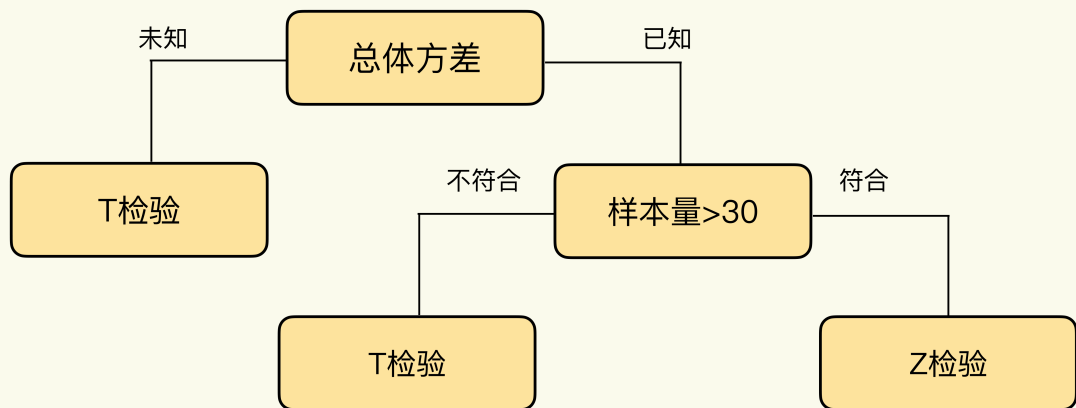
选择 T 检验还是 Z 检验，主要看样本量的大小和是否知道总体方差（Population Variance）：

当我们不知道总体方差时，使用 T 检验。

当我们已知总体方差，且样本量大于 30 时，使用 Z 检验。

我还给你画了张图，你一看就明白了。

## A/B测试 如何选取实验对象单位



那么这些理论具体到 A/B 测试实践中，一个经验就是：**均值类指标一般用 T 检验，概率类指标一般用 Z 检验（比例检验）。**

为什么会有这样的经验呢？

因为上节课我讲了，样本量大的情况下均值类指标是正态分布，正态分布的总体方差的计算需要知道总体中各个数据的值，这在现实中几乎做不到，因为我们能获取的只是样本数据。所以总体方差不可知，选用 T 检验。

那么概率类指标是二项分布，二项分布的总体方差的计算不需要知道总体中各个数据的值，可以通过样本数据求得总体方差。而且现实中 A/B 测试的样本量一般都远大于 30，所以选用 Z 检验。这里的**比例检验** (Proportion Test) 是，专指用于检验概率类指标的 Z 检验。

讲了这么多检验，我现在来总结一下：**对于 A/B 测试来说，要选用双尾、双样本的比例检验（概率类指标）或 T 检验（均值类指标）。**

再次回到我们的案例中来，由于点击率为概率类指标，所以这里选用双尾、双样本的比例检验。

### 如何利用假设检验做出推断？

选取了正确的假设和检验方法，接下来就要检验我们的假设是不是正确了，这在 A/B 测试中就是分析测试结果这一步啦。

### A/B 测试可能出现的结果

假设检验会推断出两种结果：

- 1. 接受零假设，拒绝备择假设，也就是说实验组和对照组的指标是相同的。
- 2. 接受备择假设，拒绝零假设，也就是说实验组和对照组的指标是不同的。

但是请注意，这两个结果只是假设检验根据样本数据，通过一系列统计计算推断出的结果，并不代表事实情况（总体数据情况）。如果考虑到事实情况的话，结合假设检验的推断结果会有四种可能：

A/B测试的可能结果		
	两组指标事实上不同	两组指标事实上相同
假设检验推断出两组指标不同	推断正确	第一类错误（Type I Error）或假阳性（False Positive）
假设检验推断出两组指标相同	第二类错误（Type II Error）或假阴性（False Negative）	推断正确

可以看出，只有当假设检验推断的情况和事实完全相符时，推断才正确，否则就会出现两类错误。

**第一类错误 (Type I Error):** 统计上的定义是拒绝了事实上是正确的零假设。在 A/B 测试中，零假设是两组的指标是相同的，当假设检验推断出两组指标不同，但事实上两组指标相同时，就是第一类错误。我们把两组指标不同称作阳性 (Positive)。所以，第一类错误又叫假阳性 (False Positive)。

发生第一类错误的概率用 $\alpha$ 表示，也被称为**显著水平** (Significance Level)。“显著”是指错误发生的概率大，统计上把发生率小于 5% 的事件称为小概率事件，代表这类事件不容易发生。因此显著水平一般也为 5%。

**第二类错误 (Type II Error):** 统计上的定义是接受了事实上是错误的零假设。在 A/B 测试中，当假设检验推断出两组指标相同，但事实上两组指标是不同时，就是第二类错误。我们把两组指标相同称作阴性 (Negative)，所以第二类错误又叫假阴性 (False Negative)。发生第二类错误的概率用 $\beta$ 表示，统计上一般定义为 20%。

这两种错误的概念读起来可能比较拗口，也不太容易理解，那么我就举一个新冠病毒核酸检测的例子来给你具体解释一下。

我们在这里的零假设是：被测试者是健康的，没有携带新冠病毒。

把携带新冠病毒作为阳性，没有携带作为阴性。如果一个健康的人去检测，结果检测结果说此人携带新冠病毒，这就犯了第一类错误，拒绝了事实上正确的零假设，是假阳性。如果一个新冠肺炎患者去检测，结果检测结果说此人没有携带新冠病毒，这就犯了第二类错误，接受了事实上错误的零假设，是假阴性。

现在我们了解了假设检验推断的可能结果，那么，如何通过假设检验得到测试结果呢？

实践中常用的有两种方法：P 值 (P Value) 法和置信区间 (Confidence Interval) 法。

## P 值法

在统计上，P 值就是当零假设成立时，我们所观测到的样本数据出现的概率。在 A/B 测试的语境下，P 值就是当对照组和实验组指标事实上是相同时，在 A/B 测试中用样本数据所观测到的“实验组和对照组指标不同”出现的概率。



如果我们在 A/B 测试中观测到“实验组和对照组指标不同”的概率（P 值）很小，比如小于 5%，是个小概率事件，虽然这在零假设成立时不太可能发生，但是确实被我们观测到了，所以肯定是我们的零假设出了问题。那么，这个时候就应该拒绝零假设，接受备择假设，即两组指标是不同的。

与此相反的是，当我们在 A/B 测试中观测到“实验组和对照组指标不同”的概率（P 值）很大，比如 70%，那么在零假设成立时，我们观测到这个事件还是很有可能的。所以这个时候我们接受零假设，拒绝备择假设，即两组指标是相同的。

在统计中，我们会用 P 值和显著水平 $\alpha$ 进行比较，又因为 $\alpha$ 一般取 5%，所以就用 P 值和 5% 进行比较，就可以得出假设检验的结果了：

当 P 值小于 5% 时，我们拒绝零假设，接受备择假设，得出两组指标是不同的结论，又叫做结果显著。

当 P 值大于 5% 时，我们接受零假设，拒绝备择假设，得出两组指标是相同的结论，又叫做结果不显著。

至于 P 值具体的计算，我推荐你用工具来完成，比如 Python 或者 R：

比例检验，可以用 Python 的 `proportions_ztest` 函数、R 的 `prop.test` 函数。

T 检验，可以用 Python 的 `ttest_ind` 函数、R 的 `t.test` 函数。

## 置信区间法

置信区间是一个范围，一般前面会跟着一个百分数，最常见的是 95% 的置信区间。这是什么意思呢？在统计上，对于一个随机变量来说，有 95% 的概率包含总体平均值（Population mean）的范围，就叫做 95% 的**置信区间**。

置信区间的统计定义其实不是特别好懂，其实你可以直接把它理解为**随机变量的波动范围**，95% 的置信区间就是包含了整个波动范围的 95% 的区间。

**A/B 测试本质上就是要判断对照组和实验组的指标是否相等，那怎么判断呢？**答案就是计算实验组和对照组指标的差值 $\delta$ 。因为指标是随机变量，所以它们的差值 $\delta$ 也会是随机变量，具有一定的波动性。



这就意味着，我们就要计算出 $\delta$ 的置信区间，然后看看这个置信区间是否包括 0。如果包括 0 的话，则说明 $\delta$ 有可能为 0，意味着两组指标有可能相同；如果不包括 0，则说明两组指标不同。

至于置信区间的具体的计算，我也推荐你使用 Python 或者 R 等工具完成：

比例检验，可以使用 Python 的 `proportion_confint` 函数、R 的 `prop.test` 函数。


T 检验，可以使用 Python 的 `tconfint_diff` 函数、R 的 `t.test` 函数。

现在回到推荐系统的案例中，我会分别用 P 值法和置信区间法来根据 A/B 测试的结果进行判断。

实验组（新推荐算法）：样本量为 43578，其中有 2440 个点击，点击率为 5.6%。

对照组（旧推荐算法）：样本量为 43524，其中有 2089 个点击，点击率为 4.8%。

这时候，我用 R 中的比例检验函数 `prop.test` 来计算 P 值和置信区间。

 复制代码

```
1 prop.test(x = c(2440, 2089), n = c(43578, 43524), alternative = "two.sided", c
```

得到了如下结果：

```
data: c(2440, 2089) out of c(43578, 43524)
X-squared = 28.076, df = 1, p-value = 1.167e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.005023823 0.010966272
sample estimates:
   prop 1    prop 2 
0.05599156 0.04799651
```

可以得出 P 值  $= 1.167e^{-7}$ ，远远小于 5% 且接近于 0，所以我们拒绝零假设，接受备择假设，并且推断出实验组和对照组指标显著不同。

同时，我们也可以得出两组指标差值 $\delta$ 的 95% 置信区间为[0.005,0.011]，不包含 0，也可以推断出两组指标显著不同。

## 小结

今天这节课，我们针对 A/B 测试的理论基础——假设检验，学习了假设、检验，以及相关的统计概念。你只要记住以下两个知识点就可以了。

第一，对于 A/B 测试来说，要选用双尾、双样本的比例检验（概率类指标）或 T 检验（均值类指标）。这决定了你在计算分析 A/B 测试结果时如何选取检验的参数，所以很重要。

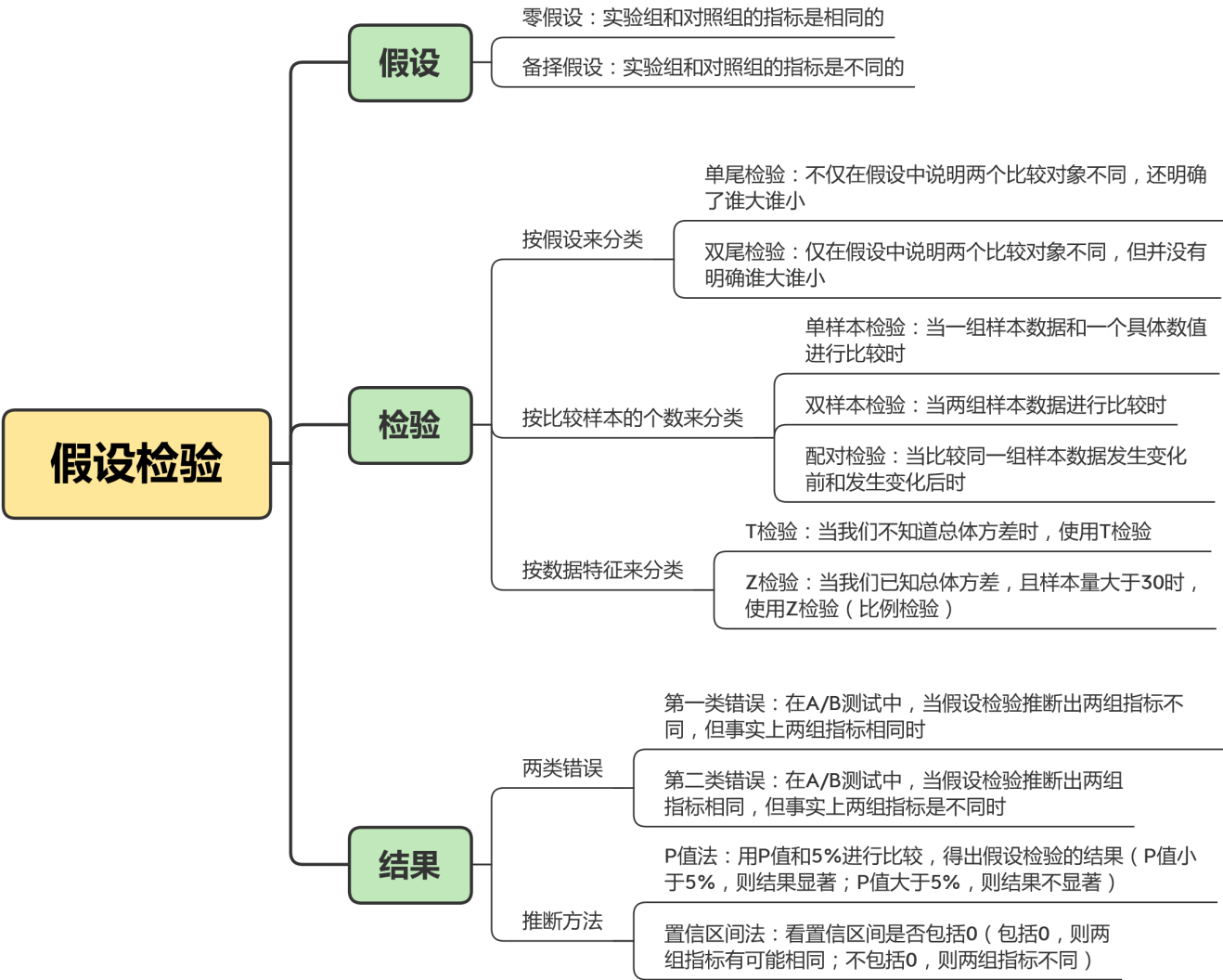
第二，在 A/B 测试实践中，计算样本量大小、指标波动性和分析测试结果的时候，会用到这些统计概念。

计算样本量大小时，会用到：第一类 / 第二类错误及其概率 $\alpha$ 和 $\beta$ 。

计算指标波动性时，会用到：方差和置信区间。

分析 A/B 测试结果时，会用到：各类检验、置信区间、P 值。

本节课中的关于假设检验的概念和知识点比较琐碎，为了方便你日后理解记忆，我也给你准备了下面的导图：



到这里我们的统计篇就告一段落了，现在你应该已经掌握了 A/B 测试所需的基本统计知识啦。其实，前两节的内容比较偏理论，会不太好理解。不过，理论知识的学习，如果只是填鸭式地讲，效果可能并不好。那该怎么掌握这些理论知识呢？在我这些年做 A/B 测试的实践中发现，要想真正把理论知识理解透，化为己用，还是需要自己多思考，多实践。等你有了一些实战后，自然就能自己体悟到理论学习的好处了。而且这时候再回过头来看理论，就会非常容易看懂。

所以，在今天的内容中，如果有哪些地方你还不能理解，那也没关系，不要给自己设置心理障碍，可以先放一放。之后的课程中，我都会运用今天讲到的理论，去解决在 A/B 测试中遇到的问题。你可以在学习的过程中不断回顾这些理论，或者发挥主观能动性，多查阅一些资料。等你学完整个课程，再回头看这两节理论知识，一定会发现理论原来如此简单。

那么接下来，我们就进入“基础篇”模块，去详细学习 A/B 测试的主要流程吧！

## 思考题

这节课涉及的统计概念都是虽然经常听到，但是难理解的，你们在学习统计中有没有对这些概念的理解有独特的心得？可以拿出来分享给大家。

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

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 01 | 统计基础（上）：系统掌握指标的统计属性

下一篇 导读 | 科学、规范的A/B测试流程，是什么样的？

## 精选留言 (10)

写留言



xiaomin 置顶

2020-12-11

请问老师，AB测试是否也能转换成单样本检验？比如AB两组样本，用A样本的均值标准差，和B样本做单样本检验？通常用excel的Z.TEST时会这么干，会有什么问题吗

作者回复: 你好，双样本检验是两个有波动性的随机变量在比较，单样本检验时一个随机变量和一个常数比较，你把其中一个变量简化成一个常数肯定会丢失掉原数据的一些特征嘛，结果肯定没有双样本检测准确的，所以A/B测试是不推荐单样本检测的。



梅不烦



2020-12-10

T检验的样本量不是小于30吗，我觉得都要用z检验吧

展开 ∨

作者回复: 你好，Z检验的适用范围从概念上来讲其实是没有样本量30这个概念的，主要是从是否已知总体方差来判断的（已知用Z检验，未知用T检验），只是在统计中我们习惯说样本量大于30就是很大的样本，就可以用样本方差来近似总体方差，这样我们就知道总体方差，就可以用Z检验了，但其实30只是经验值，大于30的总体方差也是样本方差近似的，所以如果准确的说的话样本量大于30，在总体方差未知的情况下，也要用T检验。



💬 3

👍 3



西西

2020-12-16

如果不只两个实验可以用t或z检验吗？一个对照组两个实验组，用实验组分别和对照组做假设检验吗？

作者回复: 对的！你说的是A/B/n测试，这里面有不止一个实验组，这是后就要用实验组分别和对照组做假设检验。



💬

👍 1



皓昊

2020-12-13

老师，采用python 进行置信区间法检验，得到ci\_low,ci\_upp两个参数，这两个参数都是区间值，如下。这两个参数的区间该怎么理解呢。

```
ci_low,ci_upp=proportion_confint(counts,nobs,alpha=0.05,method='normal')
print('ci_low:{0},ci_upp:{1}'.format(ci_low,ci_upp))...
```

展开 ∨

作者回复: 你好，python的这个函数是计算出两个比例的两个置信区间，所以输出有4个数，建议用R的prop.test, 得出的结果是两个比例差值的置信区间。



💬

👍 1



皓昊

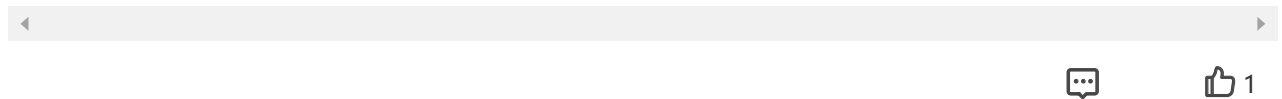
2020-12-11

老师有两个问题:

- 1.在上面的案例中我们只取了单侧P值(5%),单双侧P值如何选择?
- 2.二项分布在满足 $\min(np, n(1-p)) \geq 5$ 时,也会变成正态分布,此时我们是否也应该用T检验.

展开 ∨

作者回复: 你好, 对于第一个问题, 根据Python或者R算出的P值不论是单位还是双尾, 只需要和5%比较就可以啦, 因为你在计算P值前会告诉函数是单尾还是双尾, 所以计算方式也不一样, 并不是算出P值后再做处理的。对于第二个问题, 其实T检验和Z检验都是用于正态分布的变量, 所以选择哪种检验其实主要取决于总体方差是否已知。



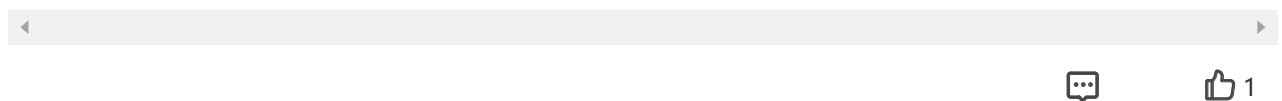
**Shehla**

2020-12-10

非常好, 逻辑清晰, 地铁上听着也很方便, 感谢

展开 ∨

编辑回复: 博伟老师的声音是不是也很好听^\_^

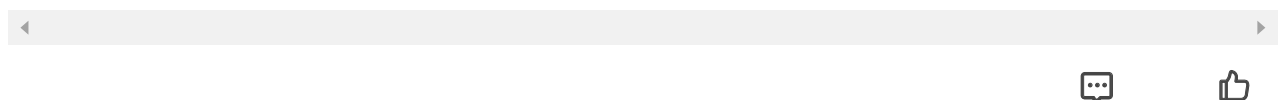


**Geek\_0e2f02**

2020-12-13

t检验就是小样本情况下的z检验, 能用Z检验的地方都可以用t检验, 可以这么理解吗, 老师

作者回复: 你好, 实践中可以这么来用, 但是这种理解不太对哈, Z检验和T检验的根本区别是总体方差是否已知, 在样本量很大时两种检验得到的结果其实是相似的。



**梅不烦**

2020-12-11

老师, 案例中的我用Python实现的, 怎么p值和您的有稍微的差别没关系吧

```
import numpy as np
```

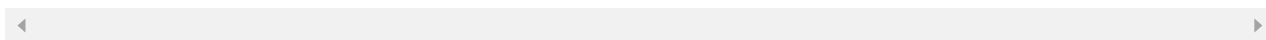
```
from statsmodels.stats.proportion import proportions_ztest # Z检验
```

```
count = np.array([2440, 2089]) # 点击数: 实验组, 对照组
```

```
nobs = np.array([43578, 43524]) # 样本量: 实验组, 对照组...
```

展开 ∨

作者回复: 你好，这个没有关系的哈，不同程序的实现方法会有一些不同，而且P值主要是和5%去比较。



1



**Kendal**

2020-12-10

后面会讲如何控制  $\alpha$ 和 $\beta$ 么？比如不是常见的5%，20%。是通过样本数量来控制么？

作者回复: 嗯嗯在第9节课中会讲到如果控制power ( $1-\beta$ )的！



2



**吴优秀同学**

2020-12-10

- 1.在样本量足够大的情况下t分布近似于在分布，所以如果你不知道该用t检验还是z检验而你样本量够大时，直接用t就得了；
- 2.在工作中我一般用独立性检验也就是卡方检验来验证两个样本比率是否发生变化。python库是`scipy.stats.chi2_contingency`，大家也可以尝试一下。

展开 ∨

作者回复: 嗯嗯优秀！



2

