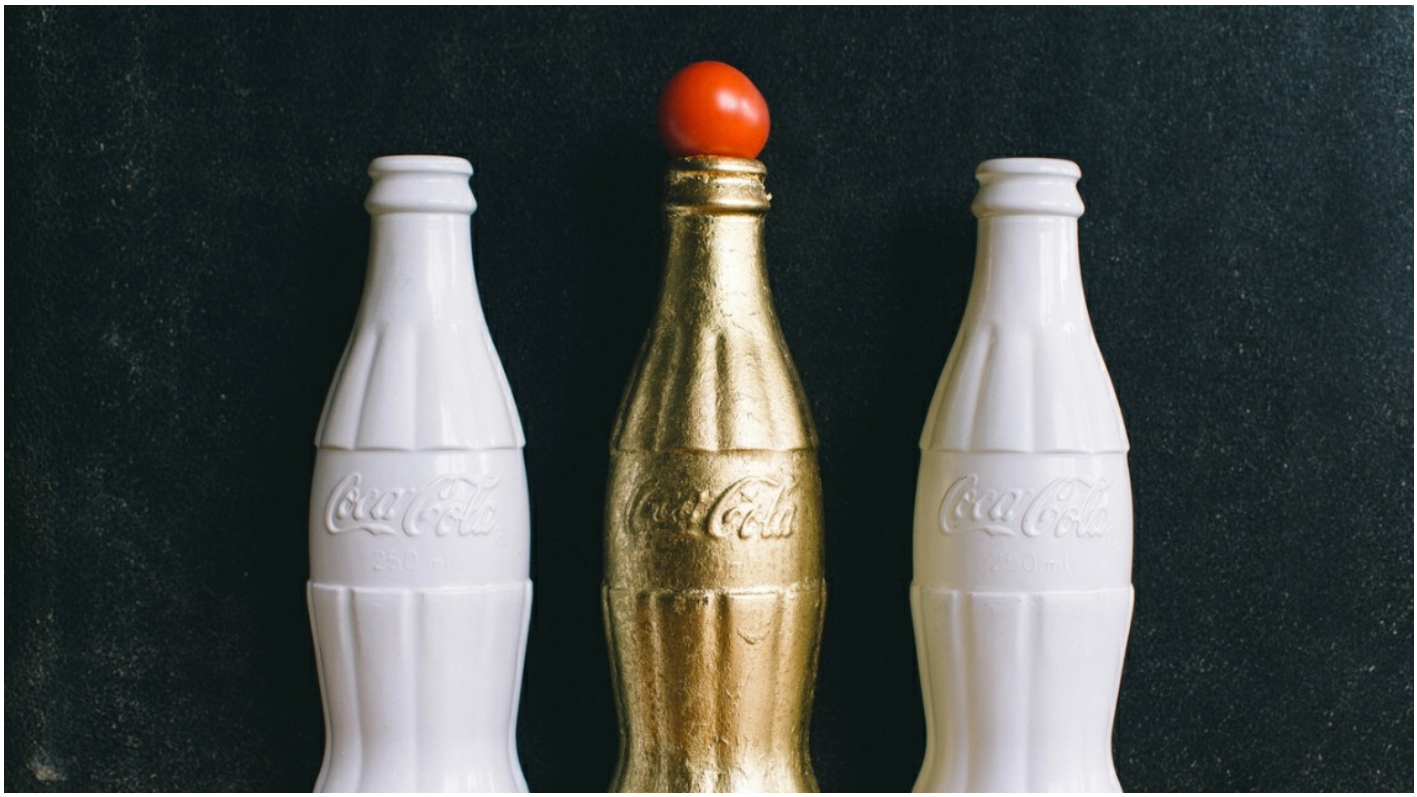
31讲统计意义（下）：如何通过显著性检验，判断你的AB测试结果是不是巧合



你好，我是黄申，今天我们接着来聊显著性检验。

上一节，我介绍了差异显著性检验的概念，它是指从统计的角度来说，差异的产生有多大的概率、是不是足够可信。这点和数值差异的大小是有区别的。既然我们不能通过差异的大小来推断差异是否可信，那么有没有什么方法，可以帮助我们检验不同数据分布之间，是否存在显著差异呢？具体的方法有不少，比如方差分析（F检验）、t检验、卡方检验等等。我这里以方差分析为例，来讲这个方法是如何帮助我们解决AB测试中的问题。

方差分析

方差分析（Analysis of Variance, ANOVA），也叫**F检验**。这种方法可以检验两组或者多组样本的均值是否具备显著性差异。它有四个前提假设，分别是：

- 随机性：样本是随机采样的；
- 独立性：来自不同组的样本是相互独立的；
- 正态分布性：组内样本都来自一个正态分布；
- 方差齐性：不同组的方差相等或相近。

根据第三个前提，我们假设数据是正态分布，那么分布就有两个参数，一个是平均数，一个是方差。如果我们仅仅知道两个分组的平均值，但并不知道它们的方差相差多大，那么我们所得出的两个分布是否有显著差异的结论就不可靠了。

为了突出重点，我们先假设咱们的数据都符合上述四个前提，然后我来详细讲解一下方差分析的主要思想。最后，我会通过Python语言来验证各个假设和最终的F检验结果。

这里，我使用之前提到的A/B测试案例，通过方差分析来检验多种算法所产生的用户转化率有没有显著性差异。我们把“转化率”称为“因变量”，把“算法”称为“因素”。这里我们只有算法一个因素，所以所进行的方差分析是单因素方差分析。在方差分析

中，因素的取值是离散型的，我们称不同的算法取值为“水平”。如果我们比较算法a和b，那么a和b就是算法这个因素的两个水平。

我们假设只有两种算法a和b参与了A/B测试。为了检验这些算法导致的转化率，是不是存在显著的差异，我们进行一个为期10天的测试，每天都为每种算法获取一个转化率。具体的数据我列在下面这张表格中。

算法	1	2	3	4	5	6	7	8	9	10
a	0.29	0.36	0.32	0.29	0.34	0.24	0.27	0.29	0.31	0.27
b	0.29	0.33	0.31	0.30	0.31	0.26	0.25	0.30	0.28	0.29

我使用Y_{ij}来表示这种表格中的数据，i表示第i次采样（或第i天），j表示第j种水平（或第j种算法）。以上面这张表格为例，$Y_{51}=0.34$。

如果我们把每种算法导致的转化率看作一个数据分布，那么方差分析要解决的问题就是：这两个转化率分布的均值，是不是相等。如果我把两种数据分布的均值记做μ1和μ2，那么原假设H0就是μ1=μ2。而对立假设H1就是μ1 <> μ2。

之前我们提到，差异是不是显著性，关键要看这个差异是采样的偶然性引起的，还是分布本身引起的。方差分析的核心思想也是围绕这个展开的，因此它计算了三个数值：SST、SSM和SSE。SST表示所有采样数据的**因变量方差**（Total Sum of Squares），我把它的计算公式列在这里。

$$SS_T = \sum\sum\left(Y_{ij} - \overline{\overline{Y}}\right)^2$$

在这个公式中， Y_{ij}如前所说，表示了第i天第j种算法所导致的转化率。而$\overline{\overline Y}$表示了10天里，2种算法全部20个数据的平均值。SSM表示数据分布所引起的方差，我们称它为**模型平方和**（Sum Of Squares for Model），它的计算公式如下：

$$SS_M = \sum n_j\left(\overline{Y_j} - \overline{\overline{Y}}\right)^2$$

在这个公式中，n_{j}为水平j下的观测数量，在我们的案例中为10。$\overline Y_{j}$为第j个水平的平均值，在案例中为算法a或算法b在这10天的平均值。$\overline Y_{j}-\overline{\overline Y}$表示的是某个算法的采样均值和所有采样均值之间的差异，n_{j}是相应的权重。我们这里的两个算法都被测试了10天，所以权重相同。根据我们的案例，SSM是0.00018。SSE表示采样引起的方差，我们称它为**误差平方和**（Sum of Squaress for Error）。它的计算公式如下：

$$SS_E = \sum\sum\left(Y_{ij} - \overline{Y_j}\right)^2$$

根据我们的案例，SSE是0.01652。我们刚刚介绍的三个统计量，SST、SSM和SSE这三者的关系其实是这样的：

$$SS_T = SS_M + SS_E$$

你可以把这三者的公式代入，自己证明一下等式是否成立。由此可以看出，SST是由SSM和SSE构成的。如果在SST中，SSM的占比更大，那么说明因素对因变量的差异具有显著的影响；如果SSE的占比更大，那么说明采样误差对因变量的差异具有更显著的影响。我们使用这两部分的比例来衡量显著性，并把这个比例称为F值。具体公式如下：

$$F = \frac{SS_M / (s-1)}{SS_E / (n-s)}$$

在这个公式中，s是水平的个数，n为所有样本的总数量，s-1为分布的自由度，n-s为误差的自由度。你可能对自由度这个概念比较陌生，这里我稍微解释一下。

自由度（degree of freedom），英文缩写是df，它是指采样中能够自由变化的数据个数。对于一组包含n个数据的采样来说，如果方差是一个固定值，那么只有n-1个数据可以自由变化，最后一个数的取值是给定的方差和其他n-1个数据决定的，而不由它自己随意变化，所以自由度就是n-1。这也是为什么在计算一组数的方差时，我们在下面这个公式中使用的除数是n-1，而不是n。

回到方差分析，对于SSM来说，如果SSM是固定的，那么对于s个水平来说，只能有s-1个组数据自由变化，而最后一组数据必须固定，所以对应于SSM的自由度为s-1。对于SSE来说，如果SSE是固定的，那么对于n个采样、s个水平数据来说，只有n-s个数据是可以自由变化的。因为每个水平中，都要有一个数据需要保证该组的平均值$\overline Y_{j}$而无法自由变化。

在我们的案例中，s为不同算法的个数，也就是水平的个数s为2，采样数据的个数n为20，所以分布的自由度为2-1=1，误差的自由度为20-2=18。

在我们的案例中，F=(0.00018/(2-1))/(0.01652/(20-2))=0.196125908。有了F值，我们需要根据F检验值的临界表来查找对应的P值。我列出了这张表的常见内容，你可以看看。

F检验临界值表（α=0.05(a)）

自由度	自变量数目（m）									显著性水平：α=0.05
(df)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n -m -1										
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190

通过这张表以及n和m的值，我们可以找到，在显著性水平α为0.05的时候，F值的临界值。如果大于这个临界值，那么F检验的P值就会小于显著性水平α，证明差异具有显著性。

在咱们的案例中，n=20，m=s-1=1，所以对应的F值为4.414。而我们计算得到的F值为0.196，远远小于4.414，因此说明差异没有显著性。虽然算法a所导致的平均转化率要比算法b的相对高出约2%（要注意，2%的相对提升在转化率中已经算很高了），但是由于差异没有显著性，所以这个提升的偶然性很大，并不意味着算法a比算法b更好。

如果需要，你可以在网上相关的统计资料里查找到完整的F检验临界值表。

使用Python代码进行验证

除了手动的计算，我们还可以用一些Python的代码来验证手动计算是不是准确。

首先，我们要确保自己安装了Python的扩展包statsmodels。如果没有安装，你可以在命令行中输入下面这行：

```
pip install -U statsmodels
```

我们可以把下列数据输入一个oneway.csv文件。

```
algo,ratio
a,0.29
a,0.36
a,0.32
a,0.29
a,0.34
a,0.24
a,0.27
a,0.29
a,0.31
a,0.27
b,0.29
b,0.33
b,0.31
b,0.30
b,0.31
b,0.26
b,0.25
b,0.30
b,0.28
b,0.29
```

安装完了statsmodels，并建立了数据文件oneway.csv，我们就可以运行下面这段Python代码来进行F检验了。

```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import scipy.stats as ss


# 读取数据，d1对应于算法a，d2对应于算法b
df = pd.read_csv("/Users/shenhuang/Data/oneway.csv") #设置为你自己的文件路径
d1 = df[df['algo'] == 'a']['ratio']
d2 = df[df['algo'] == 'b']['ratio']


# 检测两个水平的正态性
print(ss.normaltest(d1))
print(ss.normaltest(d2))


# 检测两个水平的方差齐性
args = [d1, d2]
print(ss.levene(*args))


# F检验的第一种方法
print(ss.f_oneway(*args))


# F检验的第二种方法
model = ols('ratio ~ algo', df).fit()
anovat = anova_lm(model)
print(ano
```

我们假设用于A/B测试的两个算法是相互独立且随机的，所以这里只检测了正态分布性和方差齐性。

其中，ss.normaltest分别测试了两个水平的正态分布性，两次结果如下：

```
NormaltestResult(statistic=0.16280747339563784, pvalue=0.9218214431590781)
NormaltestResult(statistic=0.4189199849120419, pvalue=0.8110220857858036)
```

ss.normaltest的原假设是数据符合正态分布，两次检验P值都是远远大于0.05的，所以原假设成立，这两者都符合正态分布。

而ss.levene分析了两者的方差齐性，同样P值都是远远大于0.05，因此符合方差齐的前提。

```
LeveneResult(statistic=0.7944827586206901, pvalue=0.38450823419725666)
```

ss.f_oneway和anova_lm都可以进行F检验。ss.f_oneway给出的结果比较简洁。

```
F_onewayResult(statistic=0.19612590799031476, pvalue=0.663142430745588)
```

而anova_lm提供了更多的信息，但是两种F检验函数都证明了我们之前的手动推算结果是正确的。

```
            df   sum_sq   mean_sq         F    PR(>F)
algo       1.0  0.00018  0.000180  0.196126  0.663142
Residual  18.0  0.01652  0.000918       NaN       NaN
```

总结

方差分析可以帮助我们检测差异的显著性，它分析的内容是受一个或多个因素影响的因变量在不同水平分组的差异。不过单因素的方差分析要求因变量属于正态分布总体，并具有方差齐性。如果因变量的分布明显的是非正态，或者方差的差异很显著，那么我们就不能直接使用这种方法。对于方差不齐的情况，我们可以选择适当的函数，例如对数、倒数等等，对原始数据进行转换，直到方差齐性变得显著，或者剔除明显属于“均值±标准差”之外的数据。

当然，对于非正态分布的数据，我们也可以使用非参数的分析。非参数检验是在总体的方差知道很少的情况下，利用样本数据对总体分布形态等进行推断的方法。名字中的“非参数”的由来，就是因为这种检验方法在推断过程中不涉及有关总体分布的参数，而只是进行分布位置、分布形状之间的比较，因此不受总体分布的限定，适用范围比较广。常见的非参数检验包括二项分布检验、K-S检验、卡方检验等等。

思考题

请尝试使用Python语言实现你自己的方差分析函数，然后通过测试数据来比较你实现的函数和Python扩展包里的函数（例如statsmodels.stats.anova.anova_lm或scipy.stats.f_oneway）。

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

精选留言

强哥

我们这面的ab test计算显著性用的是t检验，不知道跟f检验的区别是什么？对于非参数检验的方法可以用bootstraping吧！分析师对这方面比较有研究

2019-03-01 11:34

作者回复

t检验可以用pairwise的t检验，一般用于两组对比，而f检验可以进行多组（多个水平）的检验

2019-03-02 01:17

mickey

请问，显著性水平α为什么要取0.05?

2019-02-26 15:09

作者回复

这是个好问题，0.05是业界的常见标准，约定俗成。也不一定要取0.05，0.1或者是0.02，0.01，只要不是太大都可以。至于多“大”算“大”，看你愿意接受偶然性的程度，没有明显的界定，当然0.3，0.5这种就肯定太大了。

2019-02-27 01:38

mickey

算法a所导致的平均转化率要比算法b的相对高出约2% 是怎么计算出来的?

2019-02-26 13:39

作者回复

(a均值-b均值) / (b均值)，是相对百分比

2019-02-27 01:36

yaya

笔记：两组样本的差异可能是由

1.采样造成的差异

2.数据分布不同造成的差异

如果要判断更多的是由哪种差异造成的，可以计算他们的比值。

采样的差异计算 各个数据到每个分布中心的距离和比如对第j水平来说，就是数据到j水平的距离，所有采样的差异就是所有水平的差异之和

分布造成的差异计算，就是各水平均值到所有均值的差异和

这两个差异我能理解，但是他们对应的量纲应该是不同的，就是他们不是同一基准下的差异但是为什么引入自由度就可以了呢？采样差异的自由度计算为什么要保证各水平均值不变，我没能理解

2019-02-25 19:45

作者回复

量纲在物理学里用得比较多，我的理解是它代表了不同的含义，比如这里代表转化率的单位。假设转化率的定义是#click/#pageview，那么这里的两个水平的量纲都是这种次数的比例，量纲应该是相同的。

对于自由度的计算，如果平均值发生了变化，那么方差就要重新计算了。

2019-02-26 01:34