数据分析高级课程 项目1 讲义

<2018年 9月12日更新 V1.0>

本文档主要描述数据分析高级课程项目1优达公开课讲解的讲义，旨在为学员提供更加详细的指导内容。

修订记录

日期	版本	修改人	修改原因
2018年 9月12日	V1.0 <定稿>	老孟、Kylie	➔ 完成第一版资料

/ 1 统计学回顾

// 1.1 衡量数据4维度

- Center 集中趋势测量:
  - 均值Mean：数据的算数平均值
  - 中位数Median：数据按照大小排列后中间的值
  - 众数Mode：数据中出现次数最多的值
- Spread 离散程度测量：
  - 标准差（STD）：衡量数据偏离“算数”平均值的程度
- Shape 数据的形状：使用直方图观察数据的分布情况，可分为3类
  - Right Skewed（右偏态）
  - Left Skewed（左偏态）
  - Symmetric（对称分布）
- Outliers 异常值：与其他数值相比差异较大的值

// 什么是数理统计学（扩展）

概率论是数理统计学的基础，而数理统计学是概率论的重要应用。

什么是数理统计学：

1. 使用概率论和数学的方法；
2. 要就怎样收集带有随机误差的的数据；
3. 并在设定的模型（统计模型）之下；
4. 对这种数据进行分析（统计分析）；
5. 以对所研究的问题做出推断（统计推断）。

数理统计学应用过程：

1. 当我们研究一个问题时，首先要通过适当的观察或实验取得必要的数据；
2. 然后对所得数据进行分析；
3. 以对所提问问题做出尽可能正确的结论；

> 摘自《概率论与数理统计》陈希儒

/ 2 假设检验

// 2.1 假设检验

设定假设检验：

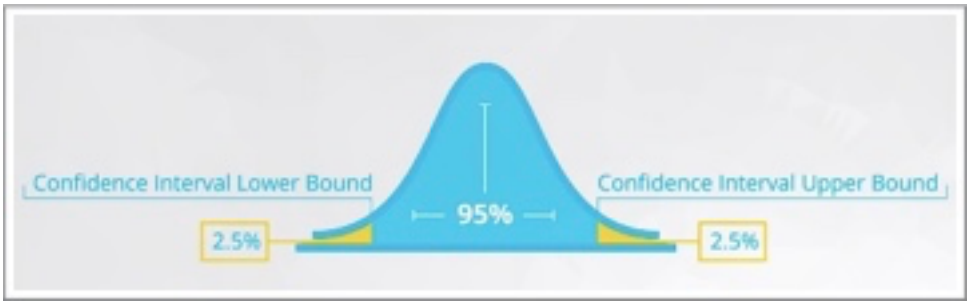
- H0零假设（null hypothesis）：针对两个测量的现象是没有相关性的，或者是检验的样本和总体之间没有相关性（收集数据之前认为真的条件）。
- H1对立假设（alternative hypothesis）：和零假设的内容完全对立的假设（与零假设是竞争且不重合的关系）。
- 显著性水平（significance level）：零假设为真时，发生不接受零假设的概率，即发生第一类错误的概率，用α表示。

Wiki的解释（扩展）：

- 在推论统计学中，零假设（英语：null hypothesis，又译虚无假设、原假设，符号：H_0）是做统计检验时的一类假设。零假设的内容一般是希望能证明为错误的假设，或者是需要着重考虑的假设。在相关性检验中，一般会取“两者之间无关联”作为零假设，而在独立性检验中，一般会取“两者之间非独立”作为零假设。
- 与零假设相对的是备择假设（或对立假设），即希望证明是正确的另一种可能。从数学上来看，零假设和备择假设的地位是相等的，但是在统计学的实际运用中，常常需要强调一类假设为应当或期望实现的假设。如果一个统计检验的结果拒绝零假设（结论不支持零假设），而实际上真实的情况属于零假设，那么称这个检验犯了第一类错误。反之，如果检验结果支持零假设，而实际上真实的情况属于备择假设，那么称这个检验犯了第二类错误。通常的做法是，在保持第一类错误出现的机会在某个特定水平上的时候（即显著性差异值或α值），尽量减少第二类错误出现的概率。
- [/链接/](https://zh.wikipedia.org/wiki/零假设)

// 2.2 置信区间

- 置信区间: 对总体参数的一个区间估计，在一定的置信水平的可信度下该区间内参数值可以被纳入其中范围，范围的上下限由假设检验得出下（如下图），范围与假设检验描述有关。



- 置信区间的意义：如果我们重复取样，每次取样后都用这个方法构造置信区间，我们有 95% 的信心说真实值落在此区间内。

- 表示方式：（点估计 - 边际误差，点估计+边际误差）
    - 边际误差：置信区间宽度的一半，通过对样本估计值的加减，达到置信区间的最终结果。
    - 点估计：从样本统计量估计得到的总体参数，因为样本统计量为数轴上某一点值，估计的结果也以一个点的数值表示。

置信水平和置信系数（扩展）：

置信水平和置信系数的概念是Neyman区间估计理论的基本概念：被估计的参数μ虽然未知，但是是一个常数，没有随机性，而区间[μ1(X), μ2(X)]则是随机的。如果把这个区间估计反复使用多次，则有时它包含μ，有时不包含μ。当次数充分大时，包含μ的频率接近置信系数。因此，一个置信系数为0.95的区间估计[μ1(X), μ2(X)]可以理解为：当把[μ1(X), μ2(X)]使用100次时，平均约有95%次其结果是正确的，即包含了被估计的μ。

P值说明：

- P值的定义：当零假设为真，真实值落在置信区间的概率。
    - P值是衡量样本数据和零假设关系的值。
    - P的取值区间为[0,1]
    - P值很小（通常是小于等于0.05）说明样本数据有足够证据拒绝零假设，P值大则反之。
    - [/简单的介绍/](https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/)

关于P值和显著性水平（α）的说明：

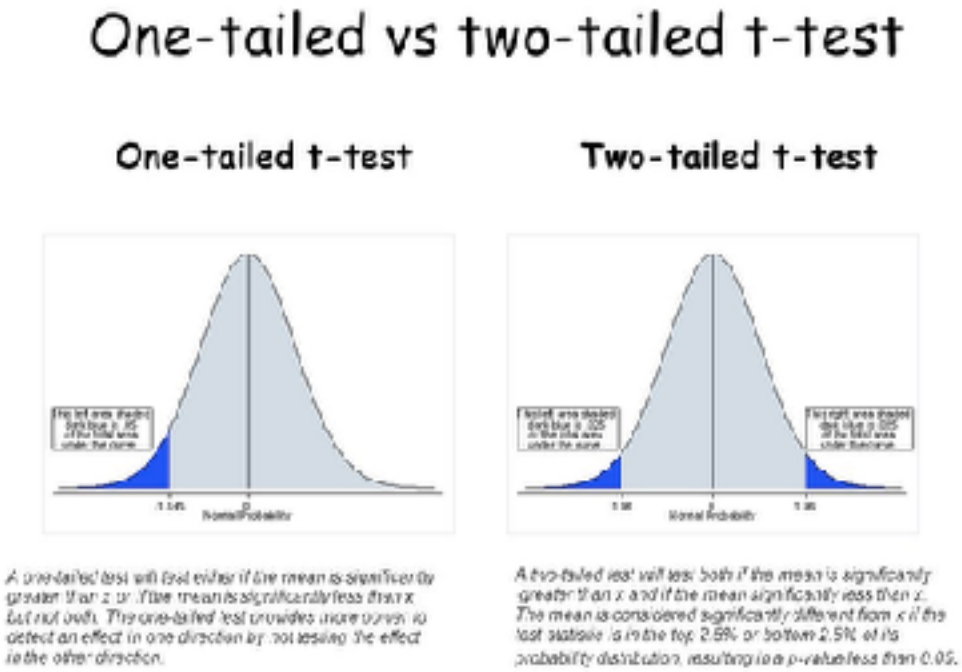
- α叫做显著性水平（significance level），是犯第一类错误的概率。
- 1-α为置信水平。
- 在进行假设检验中使用P值和α（比如0.05）做比较，但含义有区别：
    - α由J.Neyman-E.Pearson提出。需要事先给定原假设H0与备择假设H1。
    - P值由R.A.Fisher提出，为计量落在零假设置信区间的概率。P值为如果零假设为真，观察到统计量落在零假设置信区间的概率，注意这里和备则假设没有关系。
- [/对比参考链接/](https://www.zhihu.com/question/21429785)

// 单位检验和双尾检验

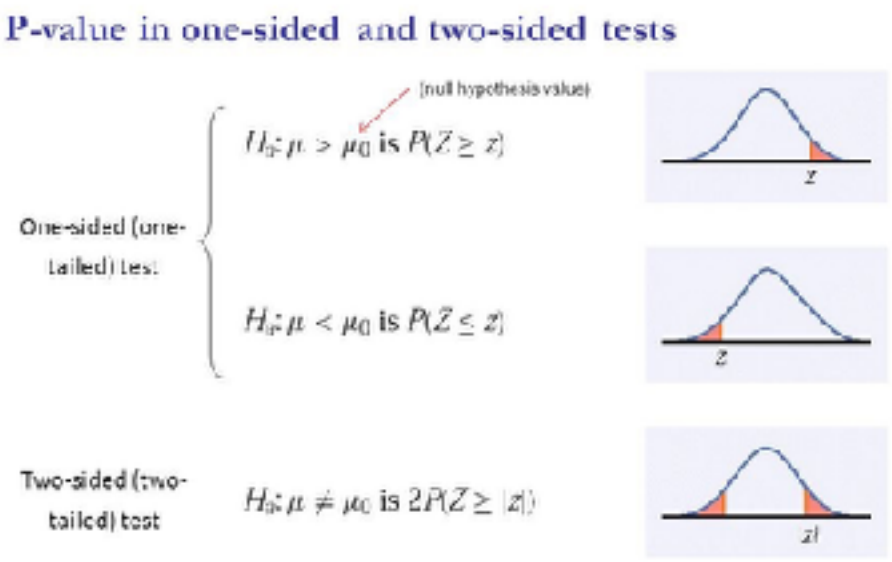
在统计显著性的测试中，单尾检验与双尾检验是根据数据集推断总体参数的两种方法。

- 双尾检验适用于估计值可能大于也可能小于（≠）参考值的情况。
- 单位检验适用于估计值只在一个方向超过（要么大于，要么小于）参考值的情况。
- [/wiki参考资料/](https://en.wikipedia.org/wiki/One-_and_two-tailed_tests)

单双尾检验：



单双尾检验和假设检验的关系：



// 2.4 如何判断假设检验

临界值检验：

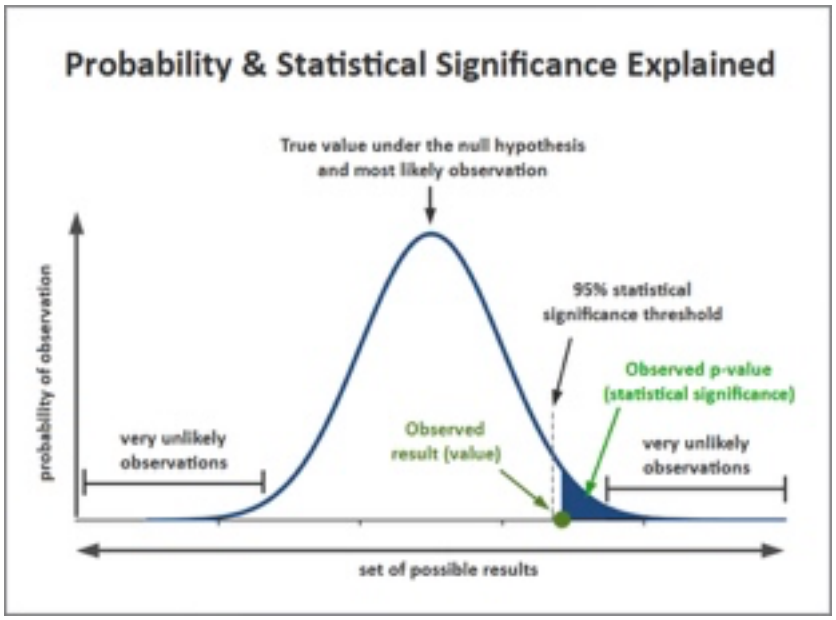
- 利用显著性水平确定临界值（查表）以及拒绝法则。
- 利用检验统计量的值及拒绝法则确定是否拒绝零假设。

T检验临界值表

自由度	显著性水平（a）			自由度	显著性水平（a）	
（df）	0.10	0.05	0.01	（df）	0.10	0.05
n-m-1				n-m-1		
1	6.314	12.706	63.657	301	1.650	1.968
2	2.920	4.303	9.925	302	1.650	1.968

P值检验：

- 利用检验统计量的值计算P值。
- 如果P值小于等于显著性水平，则拒绝零假设。



P值的计算：

[/P值的计算(Chi Squre之后查表）/](**https://www.wikihow.com/Calculate-P-Value**)

/ 3 统计学检验方法

// 3.1 T检验与Z检验

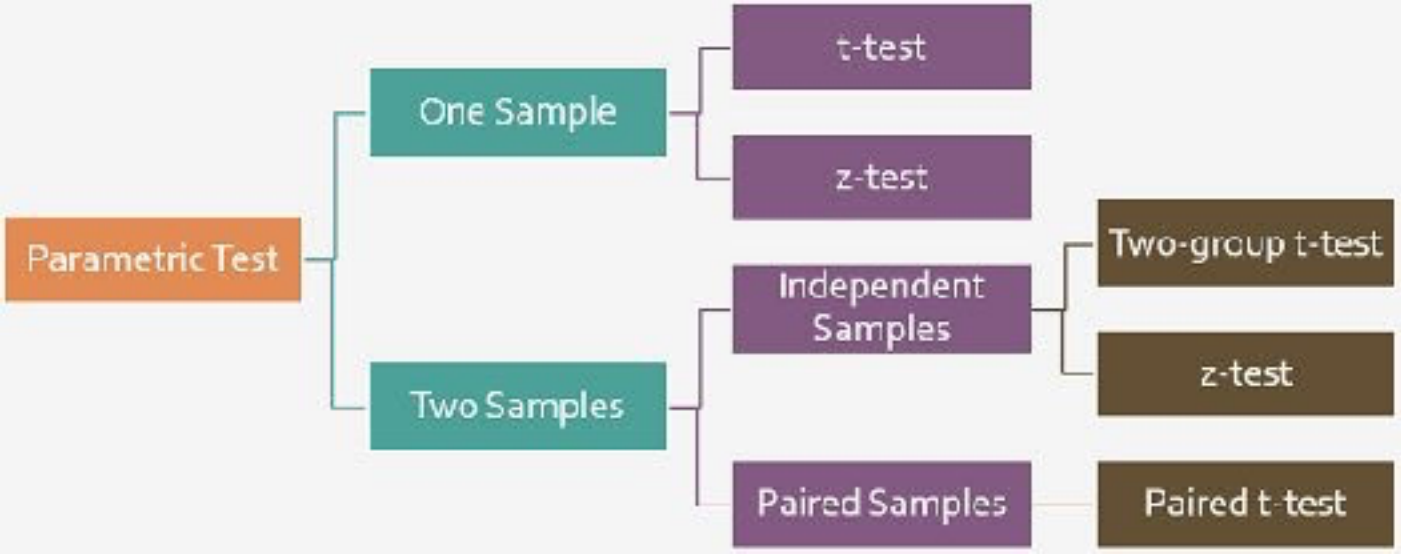
Z检验与单样本T检验的区别：

- 选择：
  - Z测试是在知道总体方差的情况下，对两组数据做对比的测试方法。（例子：For example, the manager of a candy manufacturer wants to know whether the mean weight of a batch of candy boxes is equal to the target value of 10 ounces. From historical data, they know that the filling machine has a standard deviation of 0.5 ounces, so they use this value as the population standard deviation in a 1-sample Z-test.）
  - T测试是在不知道总体方差的情况下，对两组数据做对比的测试方法。
- 假设前提：
  - 单样本Z检验与单样本T检验的假设条件相同。
  - 配对样本T检验和独立样本T检验的条件各有不同。
  - [/检验假设说明/](http://www.psychology.emory.edu/clinical/bliwise/Tutorials/TOM/meanstests/assump.htm)
- 使用Z检验的前提：
  - 知道总体参数
  - 知道总体标准差（当样本大于30时，可使用样本统计值代替总体标准差）
  - 总体分布为正态分布（根据中心极限定理，样本均值的抽样分布符合正态分布。）
- 使用T检验的前提：
  - 单样本T检验知道总体参数（双样本不要求）
  - 不知道总体标准差（当样本小于30时，不能用样本统计量代替总体标准差）
  - 样本分布为T分布（当样本数量很大时，接近正态分布。T检验不要求知道总体分布。）
- 简化判断：样本数大于30使用Z检验（总体参数和总体标准差可由样本近似得出），小于30使用T检验。

T检验：

- 单样本T检验：在已知总体平均数的前提下，对样本进行检验。
- 配对样本T检验：2个存在相依关系样本的检验。
- 独立样本T检验：2个不存在相依关系样本的检验。

Z检验与T检验的关系：



- [/原文链接/](https://keydifferences.com/difference-between-t-test-and-z-test.html)
- [/What is a Ztest/](http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/what-is-a-z-test/)
- [/What is a Ttest/](http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/types-of-t-tests/)

Z值计算公式：

Z值的计算公式

Z值是某一特征值与均值之间标准偏差的数量，其是一个相对量。

Z值的计算公式为：

$$z=\frac{(x-\mu)}{\sigma}$$

其中：x-某一特征值；μ-总体均值；σ-总体的标准差

在实际中都是通过抽样来估计总体，则Z值的计算公式变化为：

$$z=\frac{(x-\bar{x})}{s}$$

T值计算公式：

[/wikipedia/](https://en.wikipedia.org/wiki/Student%27s_t-test)

One-sample *t*-test [edit]

In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test are $n - 1$. Although the parent population does not need to be normally distributed, the distribution of the population of sample means $\bar{x}$ is assumed to be normal. By the central limit theorem, if the sampling of the parent population is independent and the second moment of the parent population exists then the sample means will be approximately normal in the large sample limit.[15] (The degree of approximation will depend on how close the parent population is to a normal distribution and the sample size, n.)

Dependent *t*-test for paired samples [edit]

This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test.

$$t = \frac{\bar{X}_D - \mu_0}{\frac{s_D}{\sqrt{n}}},$$

For this equation, the differences between all pairs must be calculated. The pairs are either one person's pre-test and post-test scores or between pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The average (X_D) and standard deviation (s_D) of those differences are used in the equation. The constant μ_0 is non-zero if we want to test whether the average of the difference is significantly different from μ_0. The degree of freedom used is $n - 1$, where n represents the number of pairs.

Independent two-sample t-test [edit]

Equal sample sizes, equal variance [edit]

Given two groups (1, 2), this test is only applicable when:

- the two sample sizes (that is, the number n of participants of each group) are equal;
- it can be assumed that the two distributions have the same variance;

Violations of these assumptions are discussed below.

The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$$

where

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

Here s_p is the pooled standard deviation for $n = n_1 = n_2$ and $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of t is the standard error of the difference between two means.

For significance testing, the degrees of freedom for this test is $2n - 2$ where n is the number of participants in each group.

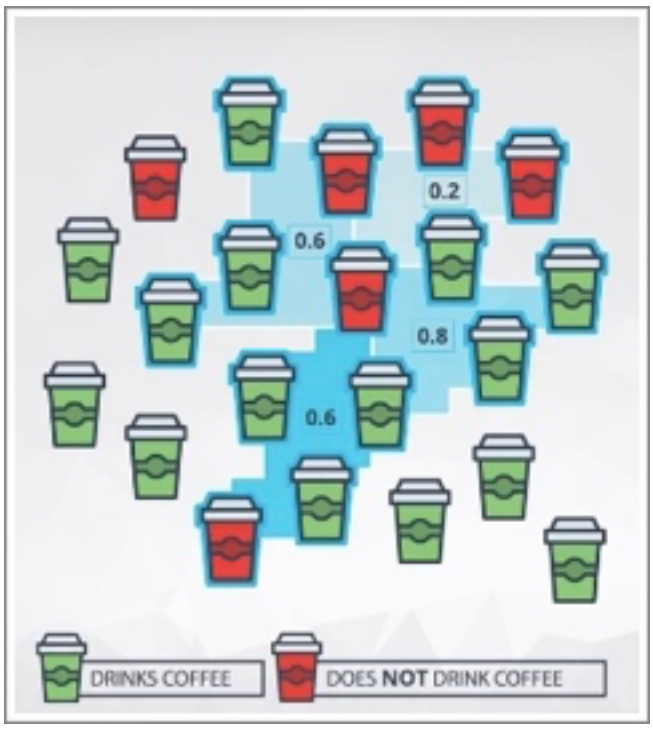
// 3.2 中心极限定律

中心极限定理：在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量之和近似服从正态分布的条件。

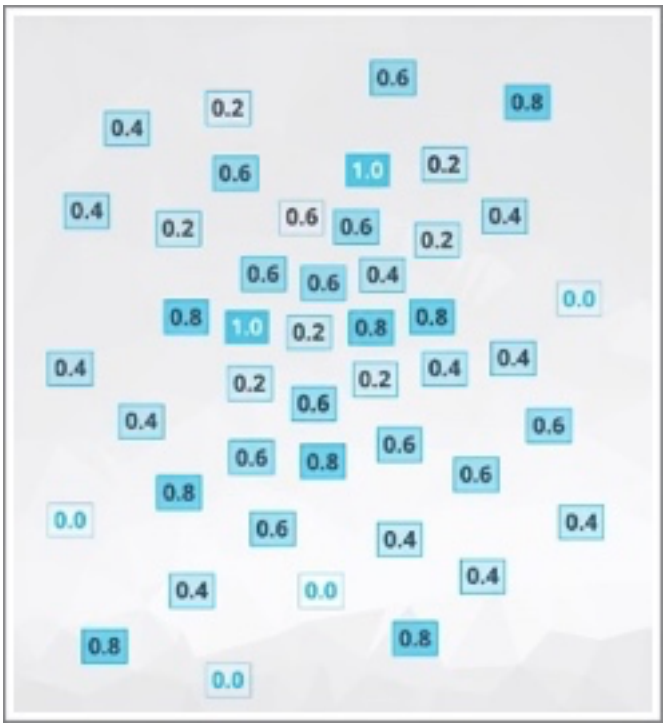
> 摘自WIKIPEDIA [/WIKI链接/](HTTPS://ZH.WIKIPEDIA.ORG/WIKI/中心极限定理)

中心极限定理举例：

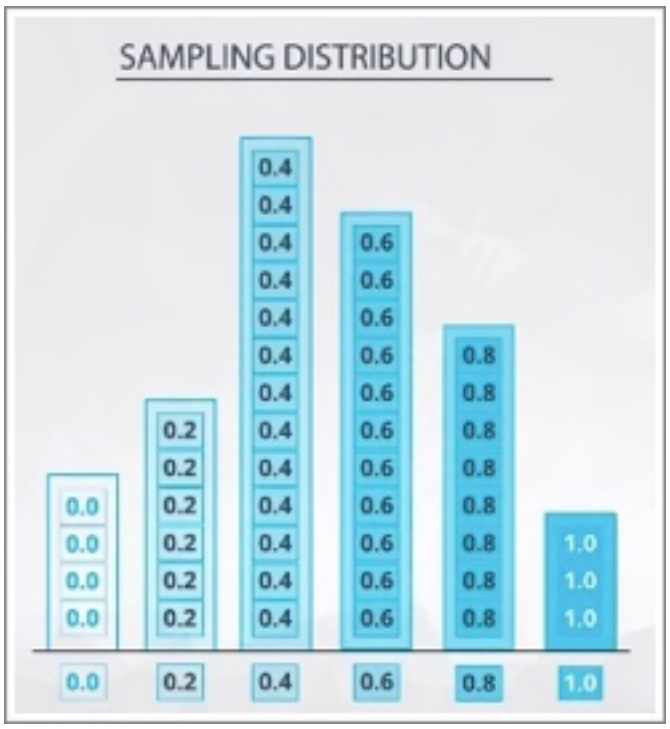
- 班上20个同学（总体），我们找5位（样本）询问是否喝咖啡，图中4个深浅不一的蓝色区域，是4次抽样的随机结果：



- 如果随机取很多次5个样本的话，通过每一次5个样本的信息，我们都可以计算出这5个同学喝咖啡的比例。将每次保抽样的比率简化为比例数字展示如图：



- 根据中心极限定律，当抽样分布样本n足够大，统计值的抽样分布会趋于正态分布。虽然本例中的n=5不是够大，我们仍然可以假设它的抽样分布近似正态分布。根据正态分布我们求出样本均值的置信区间：



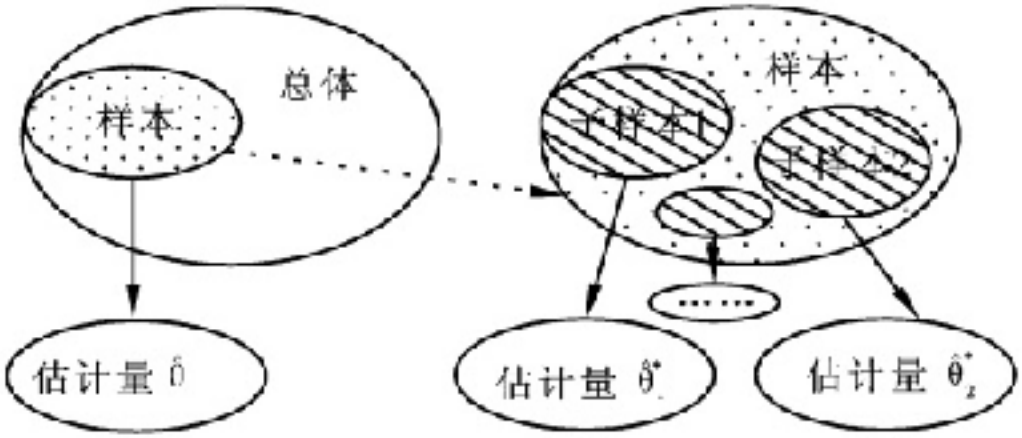
// 3.3 自助法

如果我们收到的样本比较小，那么会不会因为偶然性的原因使得根据样本的统计值推算总体参数变得不准确呢。这种情况下可以使用自助法（Bootstrap）得到较好的结果。自助

自助法步骤：

(1) 采用重复抽样技术从原始样本中抽取一定数量(可自己给定，一般与原始样本相同)的样本，此过程允许重复抽样。

(2) 根据抽出的样本计算待估计的统计量。



(3) 重复上述N次(一般大于1000)，得到N个统计量。
(4) 根据中心极限定理，上述N个统计量符合正态分布。通过计算上述N个统计量的样本方差， 可以估计统计量的置信区间。

自助法使用条件：
Bootstrap是现代统计学较为流行的一种统计方法，在小样本时效果很好（如果原始样本足够大，可以使用抽样分布）。

/ 4 代码示例

// 4.1 项目简介

项目背景：

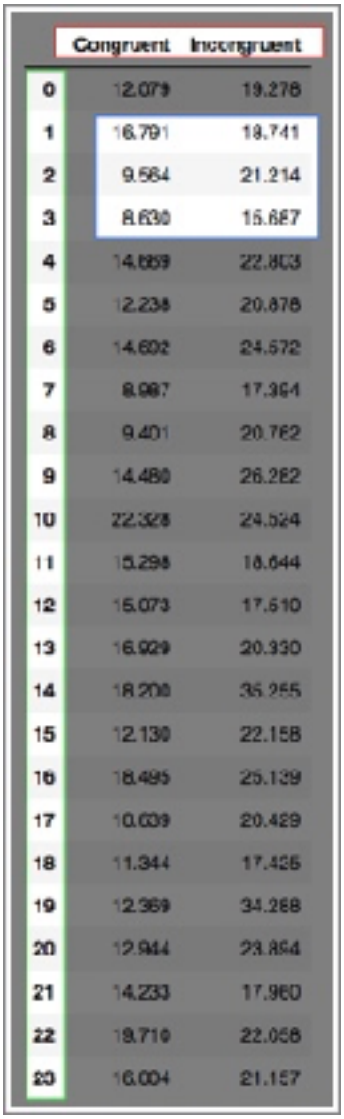
- 斯特普鲁效应：通过控制文字含义与文字颜色是否一致的条件，记录被试人的反应时间，从而分析左右脑认知的影响。
- Congruent ： 文字描述与颜色一致
- Incongruent：文字描述与颜色不一致
- [/wiki说明链接/](https://en.wikipedia.org/wiki/Stroop_effect)

可以使用的检验方式：

- 配对T检验
    - 实现方式1: 直接根据样本计算，比较24个样本Con和Incon的数值。
    - 实现方式2: 实现方式2: 使用自助法（Bootstrap），通过放回抽样的方式模拟10000次抽样，再计算。

数据说明：

- 2列数据（Congruent和Incongruent）
- 24个样本
- 数据为浮点型



	Congruent	Incongruent
0	12.079	19.278
1	16.791	18.741
2	9.564	21.214
3	8.630	15.687
4	14.669	22.803
5	12.238	20.878
6	14.692	24.572
7	8.987	17.394
8	9.401	20.762
9	14.480	26.282
10	22.328	24.524
11	15.298	18.644
12	15.073	17.510
13	16.929	20.330
14	18.200	35.255
15	12.130	22.158
16	18.495	25.139
17	10.639	20.429
18	11.344	17.425
19	12.369	34.288
20	12.944	23.894
21	14.233	17.960
22	19.710	22.058
23	16.004	21.157

// 4.2 随机抽样

- 使用.sample方法实现随机抽样
- n可以使用shape[0]的方式和样本数量保持一致
- replace = True 是定义冲抽样的参数，如果是False则不回放回抽样的个体。

```python
# 设定每次抽出的内容为bootstrap_sample
# dataframe可以使用.sample进行抽样
# 注意1：第一个参数是抽样的次数
# 注意2：对于自助法，一般将抽样的次数设定为样本的总数(用shape[0]获得)
# 注意3 ：抽取时默认是自助法 replace = Ture
# .sample方法不受seed设置影响影响
bootstrap_sample = coffee_red.sample(coffee_red.shape[0],replace = True)
# 检查下，每次都不一样
bootstrap_sample.head(1)
```

// 4.3 计算T值与P值

- 在python中使用scipy.stats进行计算
- 计算结果是双尾的P值，单尾P值需要把结果除2得到[/sciyp怎样得到单位T值/]
- (https://stackoverflow.com/questions/15984221/how-to-perform-two-sample-one-tailed-t-test-with-numpy-scipy)
- [/scipy.stats官方链接/](https://docs.scipy.org/doc/scipy/reference/stats.html)
- 配对样本T检验示例：

```python
# 引入scipy科学计算模块
from scipy import stats
# 注意1：使用.ttest_ind和.ttest_rel计算相依和独立T检验
# 注意2：小括弧中输入比较的两个数据
# 注意3：会同时输出T值和P值
t, ptwo = stats.ttest_rel(sample['Con'], sample['In'])
```

scipy.stats.ttest_rel

scipy.stats.ttest_rel(*a, b, axis=0, nan_policy='propagate'*) [source]

Calculate the T-test on TWO RELATED samples of scores, a and b.

This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values.

Parameters:

a, b : *array_like*
The arrays must have the same shape.

axis : *int or None, optional*
Axis along which to compute test. If None, compute over the whole arrays, *a*, and *b*.

nan_policy : *{'propagate', 'raise', 'omit'}, optional*
Defines how to handle when input contains nan. 'propagate' returns nan, 'raise' throws an error, 'omit' performs the calculations ignoring nan values. Default is 'propagate'.

Returns:

statistic : *float or array*
t-statistic

pvalue : *float or array*
two-tailed p-value

// 4.4 自助法

```
# import scipy的stats
from scipy import stats

# 制作假数据（通过随机抽样方式，后续掩饰和实际项目不同）
stroopfack = stroop.sample(stroop.shape[0],replace = True)

# 设定记录bootstrap方法得出t，p值的空列表
boot_t= []
boot_p = []

# 使用for in range 循环1000遍
for i in range(1000):
    # 生成sample，自展为True(replace = Ture)
    # 生成的数量和sample数量相同 (stroopfack.shape[0])
    bootstrap_sample = stroopfack.sample(stroopfack.shape[0],
                                         replace = True)
    # 使用stats中的ttest_rel得出配对T检验的T值和Z值
    t, p = stats.ttest_rel(bootstrap_sample['Congruent'],
                           bootstrap_sample['Incongruent'])
    # 生成mean_s 赋值为随机抽取的平均值
    # mean_s = bootstrap_sample.mean()
    # 将这个值附加到列表中
    boot_t.append(t)
    boot_p.append(p)

# 之后对两个列表做平均就可得出使用Bootstrap方法得出的T值和P值
```