

A/B实验

基本原理详解

引子



马云、李嘉诚、王健林、马化腾还有我，我们五人的资产加起来足以撼动整个亚洲甚至世界的经济体系，**平均财富超过2000亿**



统计学基础概念-平均值

平均值*(The average value)*有算术平均值，几何平均值，平方平均值〔均方根平均值〕，调和平均值，加权平均值等，其中以算术平均值最为常见。

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

平均值举例

| | 1号考生 | 2号考生 | 3号考生 | 4号考生 | 5号考生 | 6号考生 |
|----|------|------|------|------|------|------|
| A组 | 95分 | 85分 | 75分 | 65分 | 55分 | 45分 |
| B组 | 73分 | 72分 | 71分 | 69分 | 68分 | 67分 |

$Average(A组) = 70分$

$Average(B组) = 70分$

| | 第一次测验 | 第二次测验 | 第三次测验 | 第四次测验 | 第五次测验 |
|-----|-------|-------|-------|-------|-------|
| A同学 | 50分 | 100分 | 100分 | 60分 | 50分 |
| B同学 | 73分 | 70分 | 75分 | 72分 | 70分 |

$Average(A同学) = 72分$

$Average(B同学) = 72分$

统计学基础概念-方差

统计中的方差 (variance) 是每个样本值与全体样本值的平均数之差的平方值的平均数

方差计算公式

$$s^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + (x_3 - M)^2 + \cdots + (x_n - M)^2}{n}$$

公式描述： 公式中M为数据的平均数，n为数据的个数， s^2 为方差。

方差举例

| | 1号考生 | 2号考生 | 3号考生 | 4号考生 | 5号考生 | 6号考生 |
|----|------|------|------|------|------|------|
| A组 | 95分 | 85分 | 75分 | 65分 | 55分 | 45分 |
| B组 | 73分 | 72分 | 71分 | 69分 | 68分 | 67分 |

$Average(A组) = 70分$

$Average(B组) = 70分$

$Variance(A组) = 291$

$Variance(B组) = 4$

总结平均值与方差

平均值是描述**集中趋势**，描述的是整体能力。比如GDP高算大国，人均高算强国

方差是描述**离散程度**，或者说是波动大小或者说是变异性。方差大波动大不稳定，方差小波动小稳定

(2) 同质与变异

一个总体中有许多个体，他们之所以成为研究对象，必定存在共性（比如性别、年龄、职业等属性），这些共性即称为**同质性**；

然而，同一总体内的个体也会存在差异，这是绝对存在的，这些差异就是我们前面强调的**变异（variation）**

没有同质性就构不成一个总体供人们研究；总体内没有变异性就无需统计学。

很有道理!!!



统计学基础概念-标准差

标准差 (Standard Deviation)，是方差的算术平方根，用 σ 表示。标准差也被称为标准偏差，或者实验标准差。

$$\text{样本方差: } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\text{母体方差: } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$\text{样本标准差: } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$$\text{母体标准差: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

标准差和均值的**量纲 [单位]**是一致的，在描述一个波动范围时标准差比方差更方便。比如一个班男生的平均身高是170cm,标准差是10cm,那么方差就是100cm²。可以进行的比较简便的描述是本班男生身高分布是170±10cm，方差就无法做到这点。再举个例子，从正态分布中抽出的一个样本落在 $[\mu-3\sigma, \mu+3\sigma]$ 这个范围内的概率是99.7%，也可以称为“正负3个标准差”。如果没有标准差这个概念，我们使用方差来描述这个范围就略微绕了一点。万一这个分布是有实际背景的，这个范围描述还要加上一个单位，这时候为了方便，人们就自然而然地将这个量单独提取出来了。

统计学基础概念-标准差

| | 1号考生 | 2号考生 | 3号考生 | 4号考生 | 5号考生 | 6号考生 |
|----|------|------|------|------|------|------|
| A组 | 95分 | 85分 | 75分 | 65分 | 55分 | 45分 |
| B组 | 73分 | 72分 | 71分 | 69分 | 68分 | 67分 |

$Average(A组) = 70分$

$Variance(A组) = 291$

$Standard\ Deviation(A组) = 17.078分$

$Average(B组) = 70分$

$Variance(B组) = 4$

$Standard\ Deviation(B组) = 2.160分$

统计学基础概念-大数定律

在**随机**事件的**大量重复**出现中，往往呈现几乎**必然**的规律，这个规律就是大数定律。通俗地说，这个定理就是，在试验不变的条件下，重复试验多次，随机事件的频率近似于它的概率。偶然中包含着某种必然



概率是**频率**随样本趋于无穷的**极限**

期望是**平均数**随样本趋于无穷的**极限**

统计学基础概念-抽样



盖洛普民意调查

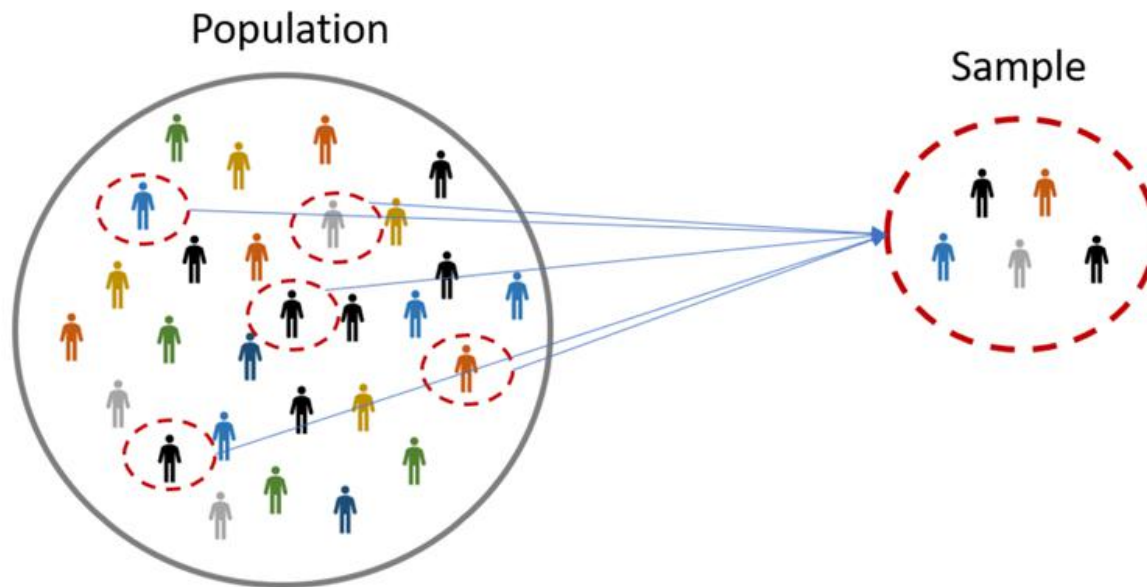


人口小普查



1. 全量成本太高、周期太长
2. 没有办法全量

统计学基础概念-抽样



抽样 [Sampling] 是一种推论统计方法，它是指从目标总体 [Population, 或称为总体] 中抽取一部分个体作为样本 [Sample]，通过观察样本的某一或某些属性，依据所获得的数据对总体的数量特征得出具有一定可靠性的估计判断，从而达到对总体的认识。

统计学基础概念-点估计和区间估计

点估计 (point estimation) 是指用样本数据来估计总体参数，估计结果使用一个点的数值表示“最佳估计值”，因此称为点估计。由样本数据估计总体分布所含未知参数的真实值，所得到的值，称为估计值

区间估计 (interval estimate) 是在点估计的基础上，给出总体参数估计的一个区间范围
[置信区间] 比如 $[\bar{x} - \delta, \bar{x} + \delta]$ ，该区间通常由样本统计量加减估计误差得到，通常区间估计会附带指出该区间的**置信度**，比如点击率有95%概率落在25%到30%之间

想估算全校男生的平均身高，抽样了100名学生，获取平均身高为170cm，那么**点估计**就是认为全校男生的平均身高为170cm。而**区间估计**会认为95%的男生的身高会落到【170-3, 170+3】区间



举个栗子

统计学基础概念-中心极限定理

中心极限定理(Central Limit Theorem)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

举个例子：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：

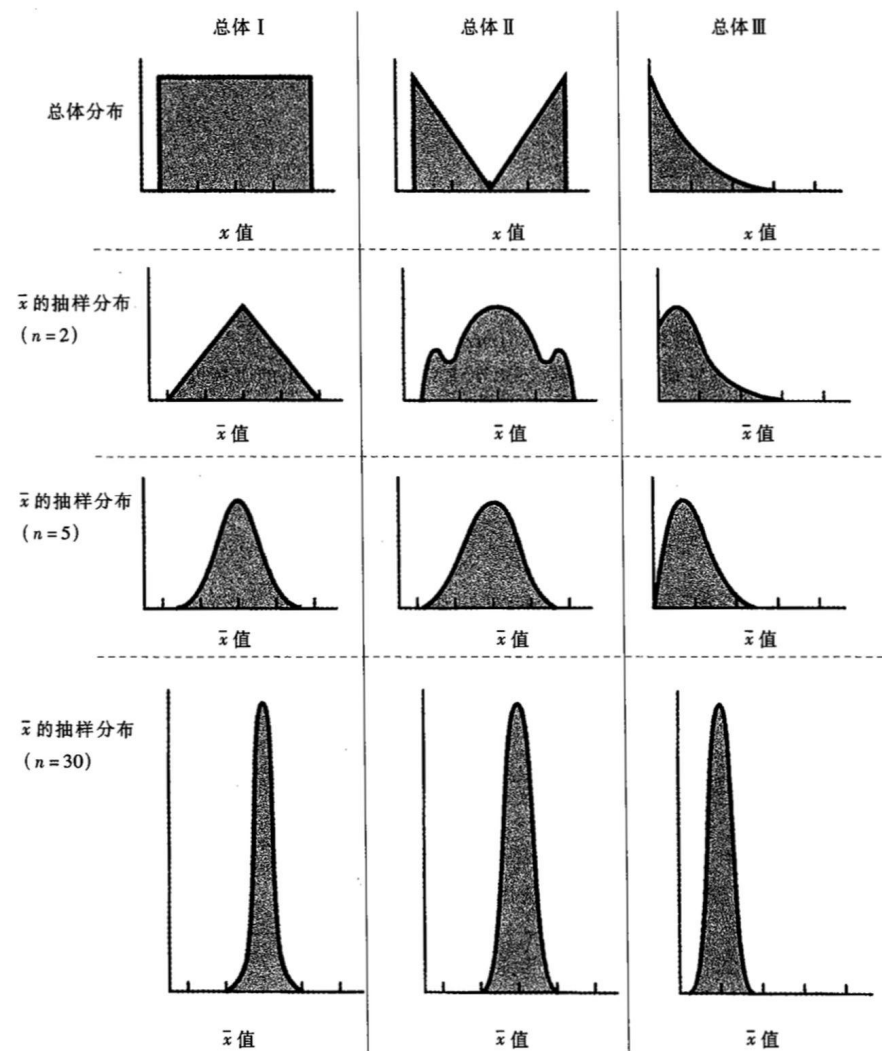
总体本身的分布不要求正态分布

上面的例子中，人的体重是正态分布的。但如果我们的例子是掷一个骰子〔平均分布〕，最后每组的平均值也会组成一个正态分布。〔神奇！〕

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。

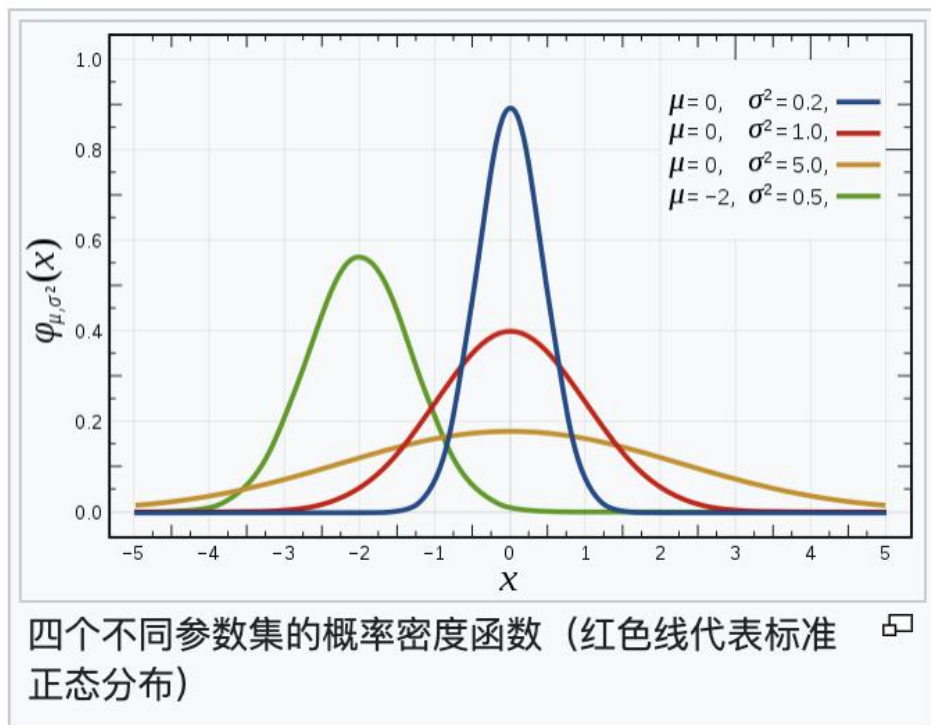
统计学基础概念-中心极限定理



这个统计学基础理论意味着我们能**根据有限样本推断总体**。结合正态分布的其他知识，我们可以轻松计算出给定平均值的值的概率。同样的，我们也可以根据观察到的**样本均值**估计**总体均值**的概率

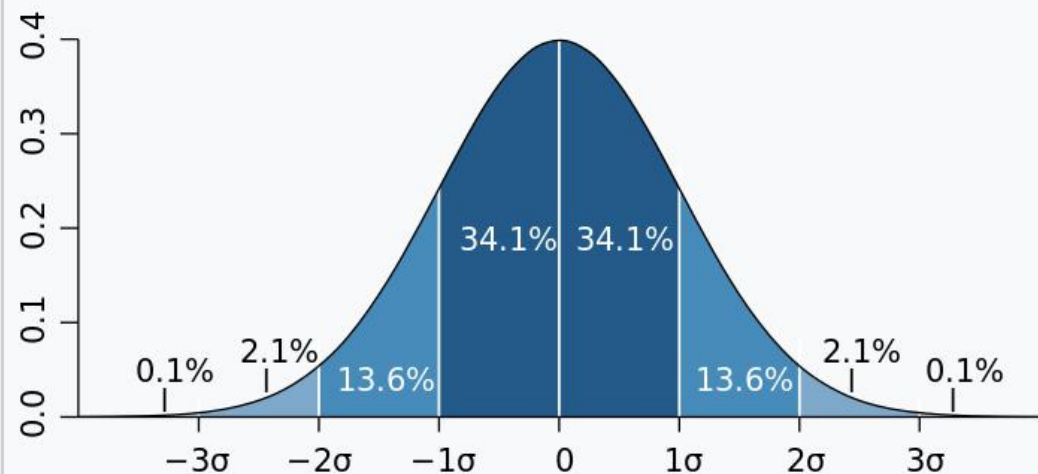
统计学基础概念-正态分布

正态分布 (Normal distribution)，也称“常态分布”，又名高斯分布 (Gaussian distribution)，是一个在数学、物理及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。正态曲线呈钟型，两头低，中间高，左右对称因其曲线呈钟形，因此人们又经常称之为钟形曲线。



若随机变量 X 服从一个数学期望为 μ 、方差为 σ^2 的正态分布，记为 $N(\mu, \sigma^2)$ 。其概率密度函数为正态分布的期望值 μ 决定了其位置，其标准差 σ 决定了分布的幅度。当 $\mu = 0, \sigma = 1$ 时的正态分布是标准正态分布

统计学基础概念-正态分布

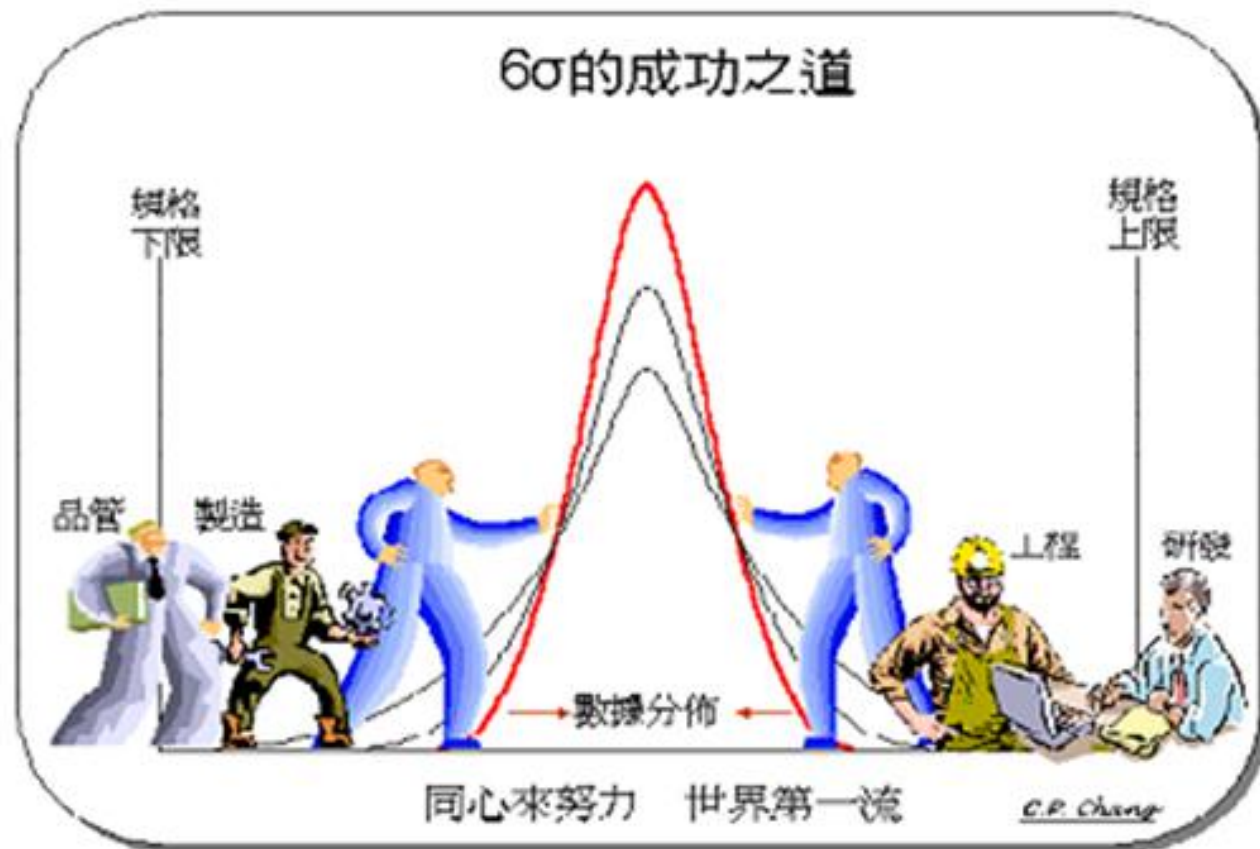


深蓝色区域是距平均值小于一个标准差之内的数值范围。在正态分布中，此范围所占比率为全部数值之68%，根据正态分布，两个标准差之内的比率合起来为95%；三个标准差之内的比率合起来为99%。

正态分布曲线简单理解为取值范围的概率，例如上图红色曲线， $[-1, 1]$ 取值概率为积分(面积)大小

| 数字比率 标准差值 | 概率 | 包含之外比例 | |
|--------------------|----------------------|---------------------|---|
| | 百分比 | 百分比 | 比例 |
| 0.318 639 σ | 25% | 75% | 3 / 4 |
| 0.318 639 σ | 25% | 75% | 3 / 4 |
| 0.674 490 σ | 50% | 50% | 1 / 2 |
| 0.994 458 σ | 68% | 32% | 1 / 3.125 |
| 1 σ | 68.268 9492% | 31.731 0508% | 1 / 3.151 4872 |
| 1.281 552 σ | 80% | 20% | 1 / 5 |
| 1.644 854 σ | 90% | 10% | 1 / 10 |
| 1.959 964 σ | 95% | 5% | 1 / 20 |
| 2 σ | 95.449 9736% | 4.550 0264% | 1 / 21.977 895 |
| 2.575 829 σ | 99% | 1% | 1 / 100 |
| 3 σ | 99.730 0204% | 0.269 9796% | 1 / 370.398 |
| 3.290 527 σ | 99.9% | 0.1% | 1 / 1000 |
| 3.890 592 σ | 99.99% | 0.01% | 1 / 10 000 |
| 4 σ | 99.993 666% | 0.006 334% | 1 / 15 787 |
| 4.417 173 σ | 99.999% | 0.001% | 1 / 100 000 |
| 4.5 σ | 99.999 320 465 3751% | 0.000 679 534 6249% | 1 / 147 159.5358 3.4 / 1 000 000 (每一边) |
| 4.891 638 σ | 99.9999% | 0.0001% | 1 / 1 000 000 |
| 5 σ | 99.999 942 6697% | 0.000 057 3303% | 1 / 1 744 278 |
| 5.326 724 σ | 99.999 99% | 0.000 01% | 1 / 10 000 000 |
| 5.730 729 σ | 99.999 999% | 0.000 001% | 1 / 100 000 000 |
| 6 σ | 99.999 999 8027% | 0.000 000 1973% | 1 / 506 797 346 |
| 6.109 410 σ | 99.999 9999% | 0.000 0001% | 1 / 1 000 000 000 |
| 6.466 951 σ | 99.999 999 99% | 0.000 000 01% | 1 / 10 000 000 000 |
| 6.806 502 σ | 99.999 999 999% | 0.000 000 001% | 1 / 100 000 000 000 |
| 7 σ | 99.999 999 999 7440% | 0.000 000 000 256% | 1 / 390 682 215 445 |

统计学基础概念-正态分布

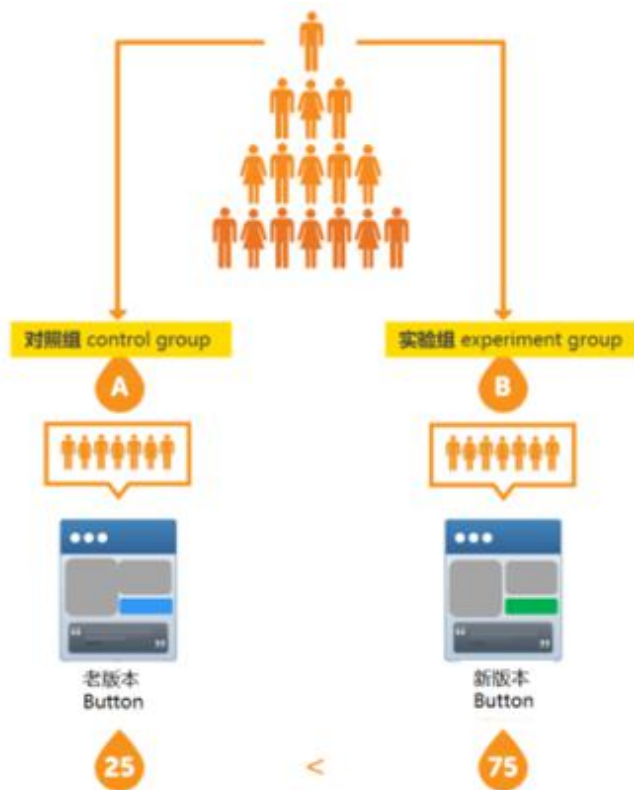


精益六西格玛

统计学基础概念小结

| 概念 | 意义 |
|--------|--|
| 平均值 | 衡量 集中趋势 ，反映整体实力 |
| 方差 | 衡量 离散程度 、变异情况、稳定情况 |
| 标准差 | 衡量离散程度、变异情况、稳定情况，和 平均值同一量纲 |
| 大数定理 | 样本足够大的条件下， 频率为概率 ， 平均值为期望 |
| 抽样 | 样本 推断 总体 |
| 点估计 | 样本估计总体，只有一个 估计值 |
| 区间估计 | 样本估计总体，有一个 区间范围和置信度 |
| 置信区间 | 区间估计中的 区间范围 |
| 置信度 | 有多大概率落在置信区间，这个概率就是 置信度 |
| 中心极限定理 | 抽样分布为 正态分布 |
| 正态分布 | 常见的 概率分布 ，期望值 μ 决定了其位置，其标准差 σ 决定了分布的幅度 |

什么是AB实验



- AB测试的概念来源于生物医学的双盲测试
- 互联网产品线上实验：是一种将多个产品、运营策略相互比较，以确定哪个策略更好的方法通过实验方法，**科学量化策略效果、提升决策效率。**
- 最常见的是AB实验，所以很多时候我们用AB实验指代「实验」
- 实验+数据是用来矫正人的判断力，让产品尽量科学高效的迭代

A/B实验

=

对照

+

随机

+

大样本

AB实验孰优孰劣



AB实验实质为**独立双样本**检验。按照中心极限定理，每个样本的均值分布是符合正态分布的，所以问题即是比较两个正态分布，正好正态分布的和与差也符合**正态分布**

统计学基础概念-小概率事件



初从文,三年不中;后习武,校场发一矢,中鼓吏,逐之出;遂学医,有所成。
自撰一良方,服之,卒

- 《杨一笑传》

译文：最开始读书,三次未考中,后转而学习武术,习武场练箭,射中了击鼓手,被开除,于是他又去学医,学到一些知识,就自己编撰了一药方,按药方自己服下试验,结果死了

小概率事件的意义重大, 因为, 有这样一个推理, 小概率事件通过上面的定义, 它是很难发生的, 但是, 如果在一次抽样试验中, 它发生了, 说明这件事违反常理, 进一步, 说明假设不成立。这就是**小概率反证法**。需要注意, 小概率事件在一次试验中发生的机会非常小, 但是, 如果做了许多次试验, 它必然发生。举例: 如果, 置信区间为95%, 做了100次试验, 则小概率事件发生的大概次数为5次

统计学基础概念-假设检验

假设检验是依据反证法思想，首先对总体参数提出某种假设〔原假设〕，然后利用样本信息去判断这个假设是否成立的过程。在A/B Test中一般有两种假设：

- 原假设 H_0 ：我们反对的假设〔样本与总体或样本与样本间的差异是由抽样误差引起的，不存在本质差异〕
- 备择假设 H_1 ：我们坚持的假设〔样本与总体或样本与样本间存在本质差异〕

假设检验的目的就是拒绝原假设，它的核心是**证伪**

在A/B Test中，我们的目标不是要估算全部用户的转化率，而是选出实验组和对照组中的更优方案。因此A/B Test的估计量不再是 P ，而是 $P_2 - P_1$ 〔实验组和对照组的转化率之差〕。原假设是 $P_2 - P_1 = 0$ 〔即两者没差别〕，因为只有当你怀疑实验组和对照组不一样，你才有做实验的动机，所以我们支持的备择假设是 $P_2 - P_1 \neq 0$ 〔两者有差别〕

统计学基础概念-两类错误

| | | 客观真实情况 | |
|--------|-------|---------|----------|
| | | H0正确 | H0不正确 |
| 假设检验结果 | H0被拒绝 | I类错误：弃真 | 正确判断 |
| | H0被接受 | 正确判断 | II类错误：纳伪 |

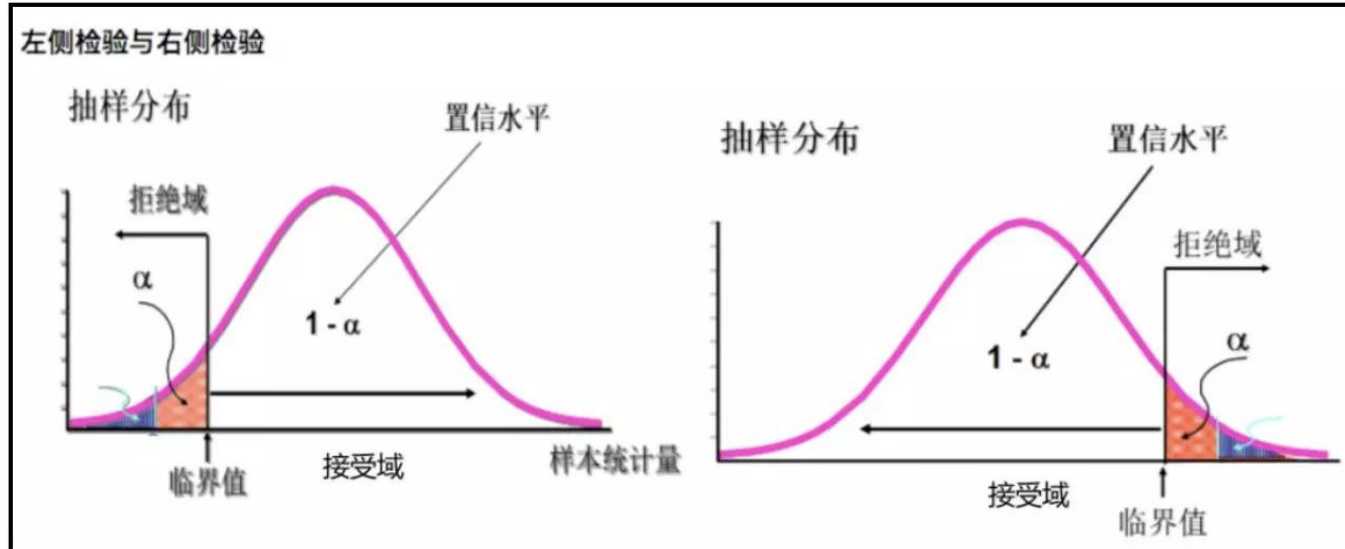
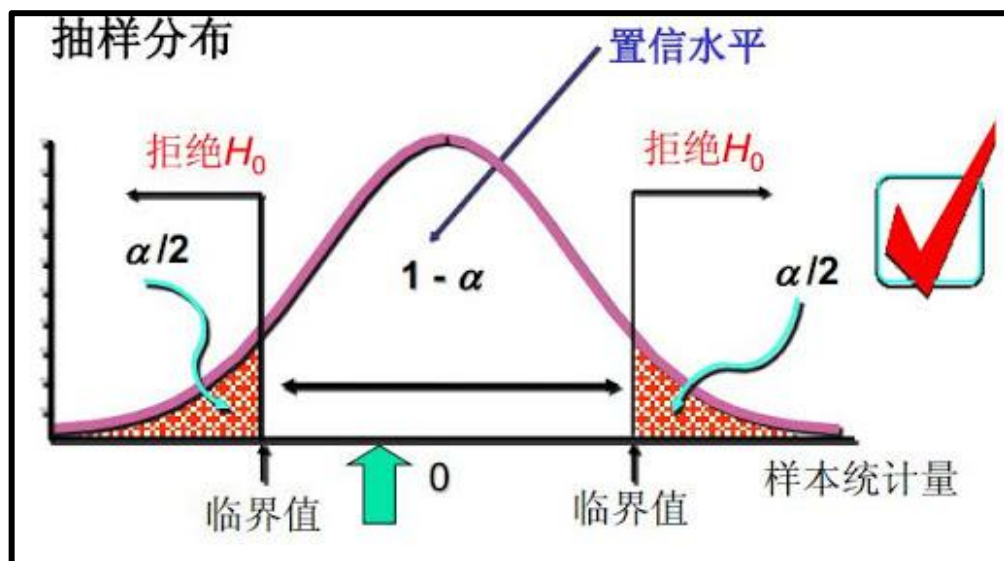
一类错误 α [做无用功] 成本高，尽量避免一类错误，阈值为0.05或者0.01，称为显著性水平，置信度为 $1 - \alpha = 95\%$

二类错误 β [错过机会] 也需要控制，尽量为10%到20%，统计功效为80%~90%。功效 [power]：正确拒绝原假设的概率，记作 $1 - \beta$ 。

| 指标 | 标记 | 意义 |
|----------------|-------------------------|----------------------------------|
| 一类错误、显著度、显著性水平 | α [常取0.05或者0.01] | 认为如果观测值发生的概率低至 α , 就拒绝零假设 |
| 置信度、置信水平 | $1 - \alpha$ [常为95%] | 置信区间展现的是真实值有一定概率落在测量结果的周围的程度 |
| 二类错误 | β [常取20%] | 纳伪的概率 |
| 统计功效 | $1 - \beta$ [常取80%] | 如果备择假设是真的, 我们有多大概率能接受备择假设 |

统计学基础概念-单侧检验和双侧检验

| 假设 | 研究的问题 | | |
|-------|------------------|------------------|------------------|
| | 双侧检验 | 左侧检验 | 右侧检验 |
| H_0 | $\mu = \mu_0$ | $\mu \geq \mu_0$ | $\mu \leq \mu_0$ |
| H_1 | $\mu \neq \mu_0$ | $\mu < \mu_0$ | $\mu > \mu_0$ |



假设检验举例



用万能的扔硬币来举例:

我们的**原假设 H_0 :硬币是均匀的**, **备择假设 H_1 :硬币不是均匀的**。

当扔硬币1次, 正面朝上了, 如果硬币是均匀的, 那么发生这件事的概率是0.5;

当扔硬币2次, 两次正面都朝上, 如果是均匀的硬币, 那么发生这件事的概率是 0.5×0.5 , 为0.25; 接着你扔了3次, 4次, 每次都正面朝上。当扔硬币5次的时候, 仍然是正面朝上, 如果硬币是均匀的, 那么发生这件事的概率只有 $0.5^5 = 0.03$

这是一个**非常小的概率**事件, 因为如果硬币是均匀的, 是不太可能发生这样极端的事情的。但是这样极端的事情却发生了, 这使你怀疑原假设的正确性, 因为一枚不均匀的硬币极有可能投出这样的结果, 因此你**拒绝了原假设, 接受了备择假设**, 认为这是一枚不均匀的硬币。

假设检验步骤

1. 提出原假设与备择假设 (A/B两个假设)
2. 从所研究总体中抽取两个随机样本 (对照组和实验组)
3. 构造**检验统计量** (计算P值)
4. 根据显著性水平确定拒绝域临界值 (一般是0.05或者0.01)
5. 计算检验统计量与临界值进行比较 (判断是否显著)

P值 (P value) : 在假设原假设 H_0 正确时 (即A/B两个假设无显著性差异), 出现现状[实验数据]或更差的情况的概率

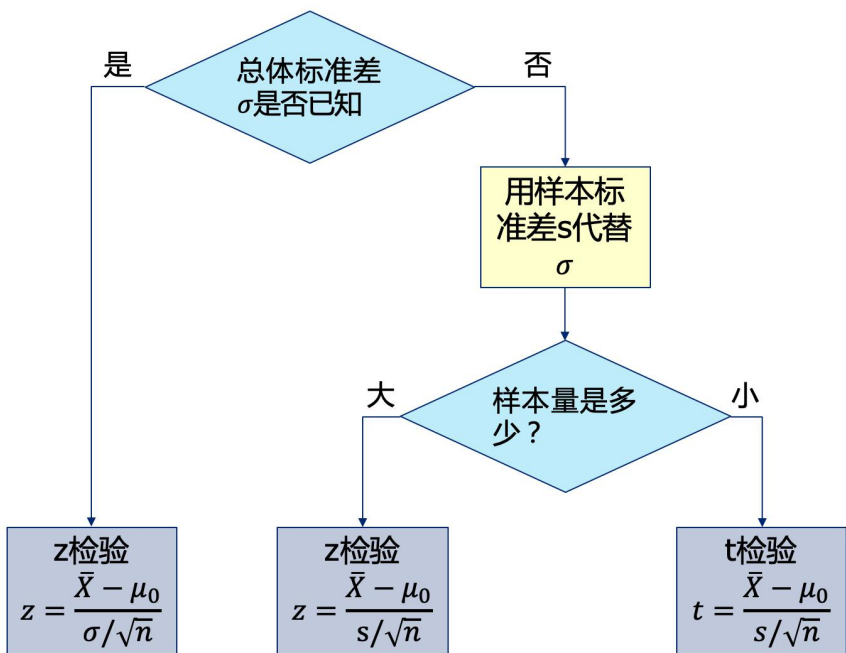
如果根据AB两组实验的**观察值来判断p值和统计功效**, 如果落在了拒绝域出现了**小概率事件**, 那么就可以推翻之前的假设

Z检验和T检验

A/B两组样本抽样**均值对比**的假设检验方法主要有**Z检验**和**T检验**，它们的区别在于Z检验面向总体数据和大样本数据，而T检验适用于小规模抽样样本。下面分别介绍Z检验和T检验。

1.Z-Test 用于大样本 $[n>30]$ ，或总体方差已知；

2.T-Test 在小样本 $[n<30]$ ，且总体方差未知时，适用性优于Z-Test，而在大样本时，T-Test 与 Z-Test 结论趋同



T检验-t统计量计算

检验统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

μ : 两个总体均值

\bar{x} : 两个样本均值

s : 样本标准差

σ : 总体标准差 : 当总体标准差已知时, 用 σ 参与计算更精准。

n : 两个样本量

T检验-查表法

自由度=样本个数-样本数据受约束条件的个数，即 $df = n - k$ (df 自由度， n 样本个数， k 约束条件个数，一般为1)。

通俗点说，一个班上有50个人，我们知道他们语文成绩平均分为80，现在只需要知道49个人的成绩就能推断出剩下那个人的成绩。你可以随便报出49个人的成绩，但是最后一个人的你不能瞎说，因为平均分已经固定下来了，自由度少一个了

| df | 单尾概率 | | | | | |
|-----|-------|-------|-------|--------|--------|--------|
| | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| | 双尾概率 | | | | | |
| | 0.50 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
| 1 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 0.765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 0.741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 0.727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 0.718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 0.711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 0.706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 0.703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 0.700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 0.697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 0.695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 0.694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 0.692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 0.691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 0.690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 0.689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 0.688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 0.688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 0.687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 0.686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 0.686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 0.685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 0.685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 0.684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 0.684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 0.684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 0.683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 0.683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 0.683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 0.681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 0.679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 0.677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| ∞ | 0.674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

T检验-正态分布估计

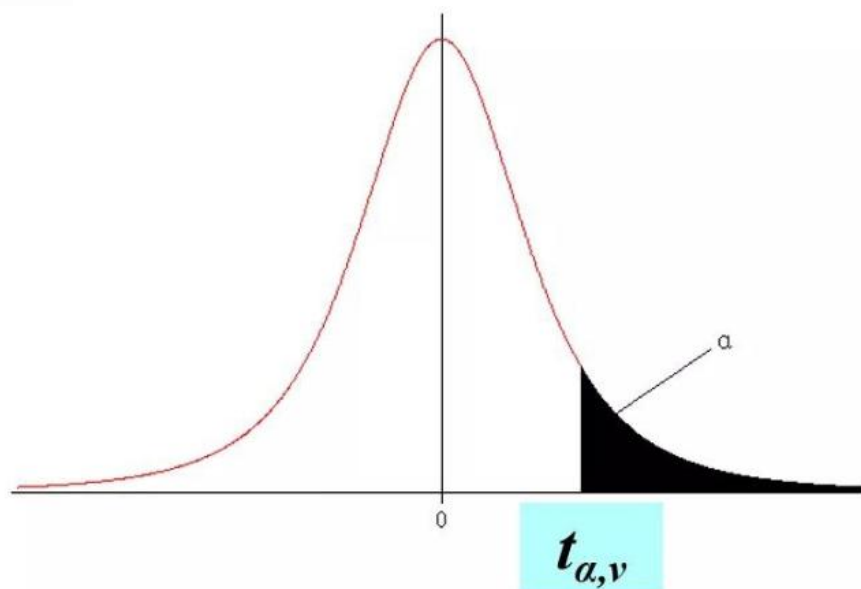
正态分布估计（常用方法）

- 样本量足够大时，t分布逼近一个标准正态分布
- 给定对应的t统计量T，P值的计算公式为

$$2\Phi(-|T|)$$

其中，

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$



实战

| A键盘 | B键盘 |
|-----|-----|
| 6 | 6 |
| 6 | 5 |
| 2 | 8 |
| 7 | 5 |
| 8 | 7 |
| 9 | 5 |
| 3 | 7 |
| 4 | 5 |
| 5 | 9 |
| 6 | 6 |
| 3 | 7 |
| 6 | 6 |
| 6 | 8 |
| 2 | 8 |
| 7 | 5 |
| 4 | 5 |
| 5 | 9 |
| 3 | 8 |
| 4 | 5 |
| 5 | 8 |
| 6 | 6 |
| 8 | 7 |
| 3 | 1 |
| 4 | 3 |
| 5 | 5 |

问题：这是两款键盘布局不一样的手机[A版本，B版本]，你作为公司的产品经理，想在正式发布产品之前知道，哪个键盘布局对用户体验更好呢？我们随机抽取实验者，将实验者分成2组，每组25人，A组使用键盘布局A，B组使用键盘布局B。让他们在30秒内打出标准的20个单词文字消息，然后记录打错字的数量。

零假设：A版本和B版本没有差别

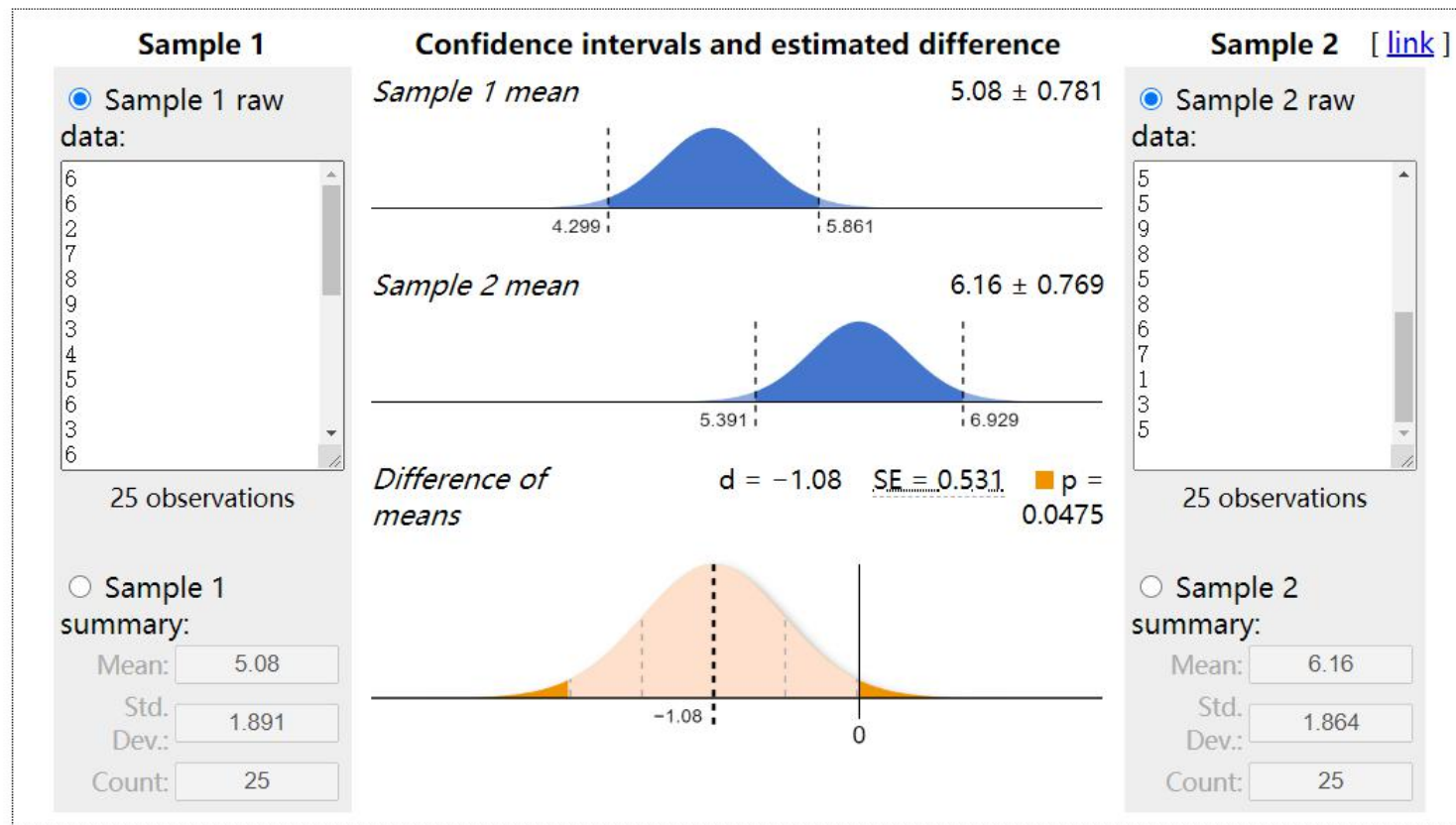
备选假设：A版本和B版本有差别

A键盘 均值：5.08 标准差：1.891 样本数量：25

B键盘 均值：6.16 标准差：1.864 样本数量：25

统计检验量值=2.0337032827228，查表介于0.02和0.05间，可以认为出现了显著性，A键盘和B键盘不一样，结合平均值，可以认为A键盘打字效果更好

Question: Does the average value differ across two groups?



Verdict: Sample 2 mean is greater

Hypothesis: ☒ d = 0 ☐ d ≤ 0 ☐ d ≥ 0

Confidence: 95%

统计学基础概念小结

| 概念 | 意义 |
|---------|--|
| 小概率事件 | 衡量 集中趋势 ，反映整体实力 |
| 反证法 | 衡量 离散程度 、变异情况、稳定情况 |
| 假设检验 | 衡量离散程度、变异情况、稳定情况，和 平均值同一量纲 |
| 一类错误 | 样本足够大的条件下， 频率为概率 ， 平均值为期望 |
| 二类错误 | 样本 推断 总体 |
| 统计功效 | 样本估计总体，只有一个 估计值 |
| 效应值 | 可以通俗理解为两个差异的 量化衡量标准 ，比如实验组的付费率为 35%，对照组为 30%，那这个 5% 就可以理解为效应值 |
| P值 | 样本估计总体，有一个 区间范围和置信度 |
| Z检验 | 两组样本抽样均值对比的假设检验方法，适用于 总体方差已知或者大样本 [$n \geq 30$] |
| T检验 | 两组样本抽样均值对比的假设检验方法，适用于 总体方法未知或者小样本 [$n < 30$] |
| 卡方检验 | 关联性分析 ，比如实验质量监控，看两组实验用户数是否有差异 |
| Delta方法 | 检验 比率指标 [分子、分母同时为样本均值] |
| 多重比较修正 | 独立实验次数增多， H_0 概率提升。多重比较修正的核心在于 控制H_0出现的次数 |

0.05可信吗？


为什么是0.05？

我们前面提到，在显著性检验中，当 p 值小到一定程度时，我们就认为原假设不成立。可是为什么这条线就划在了0.05这里？这个问题有一个很无趣的答案：这是费希尔老爷子随口一说的。为了避免像错怪格格巫一样的错误，我们希望尽可能保守一些，因此显著性的界限也应该比较小。但是另一方面，这个界限也不能太小，不然社会投入到科研的资源无法满足能得到显著性结果的样本量。

费希尔的随口一说之中似乎也包含了某种神奇的直觉。有学者提出，对于过去近百年中生物医学和社会科学（运用统计学方法最普遍的学科）研究中常见的效应大小和样本量而言，0.05这个界限恰好在任何实验都做不出显著性结果和假阳性发现满天飞之间找到了一点微妙的平衡。当然，科学研究在不断地发展，当代的许多新领域（如基因组学）中的海量数据和测试已经对0.05这条金标准作出了挑战，统计学家也发展出了新的对策。这里我们先按下不表，在后续文章中将会——道来。

另外，0.05的存在也是「前计算机时代」的一个历史遗留产品。九十年代以前，计算机和统计软件还没有被广泛使用，人们进行统计学分析时，往往需要借助统计学表格，把根据样本算出的统计量与表格中的临界值进行比较。由于篇幅所限，表格自然不能列出所有的 p 值，因此当时的人们都倾向于报告 $p < 0.05$ 的结果。随着统计软件的流行，如今获得精确的 p 值已不是难事，人们也不再采用这样模糊的表述了。但是0.05这个门槛儿却成为了一种文化，被科学界保留了下来。

一些工具

 **AB Testguide**

How many visitors do you need?

Conversion rate: Control via your test page, in %

Expected improvement over control relative, in %

Unique visitors on your test page per week

Max number of weeks for AB-test used to calculate minimum expected relative improvement vs Control

Hypothesis:

☒ One-sided
☐ Two-sided

Power:

☐ 75%
☒ 80%
☐ 90%
☐ 95%

Required confidence level (1 - alpha):

Minimum sample size:

28903

unique visitors per test variation

Power: Confidence level: CR A: CR B:

AB test duration

Minimum test duration 57.81 weeks *

Round up to a **AB-test period of 58 weeks** (discrete number of business cycles)

Minimum improvement needed

Minimum Conversion rate B of 3.26 % needed to run AB-test in max 4 weeks, with 1000 unique visitors per week on your page *

This amounts to a **minimum relative improvement of 63% needed** in order to run the test at the required levels of power and confidence.

* assuming an AB-test with two variants (Control and B)
Both calculations use R-function power.prop.test as described in
37signals.com/blog/ab-testing-tech-note-determining-sample-size

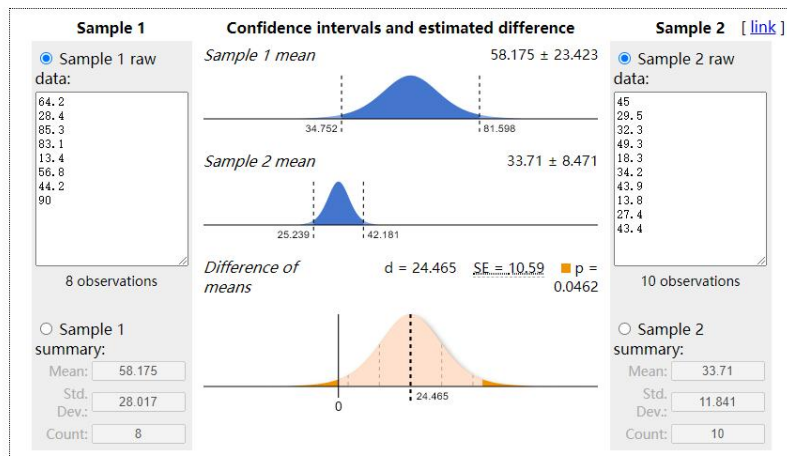
<< Back to the power and confidence visualisation tool

Recommended reading

<http://www.wlv-vector.com/blog/2014/05/a-clear-picture-of-power-and-significance-in-ab-tests/>

<https://abtestguide.com/abtestsize/>

Question: Does the average value differ across two groups?



Verdict: Sample 1 mean is greater

Hypothesis: ☒ d = 0 ☐ d ≤ 0 ☐ d ≥ 0

Confidence:

<https://www.evanmiller.org/ab-testing/t-test.html>



没有银弹

THANK YOU

