

正态分布

2023年12月27日 10:20

1. 中心极限定理

若一个结果是由大量的、不相干的、随机因素叠加导致的，那么这个结果，一定表现为正态分布，而正态分布的中心，也就是其对称轴必然等于总体均值。



例如，某一省份几十万考生的英语原始成绩，必然呈现正态分布。而对称轴的位置，必然是所有考生总体的平均分。

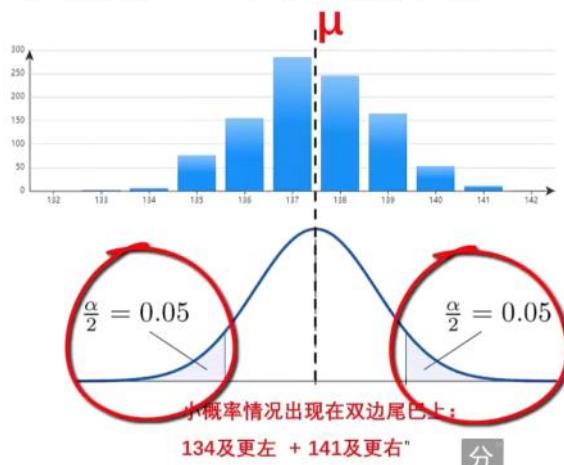
数学表述：

所研究的随机变量如果是有大量独立的随机变量相加而成，那么它的分布将近似于正态分布。

通俗表述：

如果一个结果是由大量的不相干的原因累加导致的，每个原因的作用范围又非常有限，那么这个结果一定是表现为正态分布。例如：我国某年高考中全部考生的英语成绩卷面原始分，必定服从正态分布。

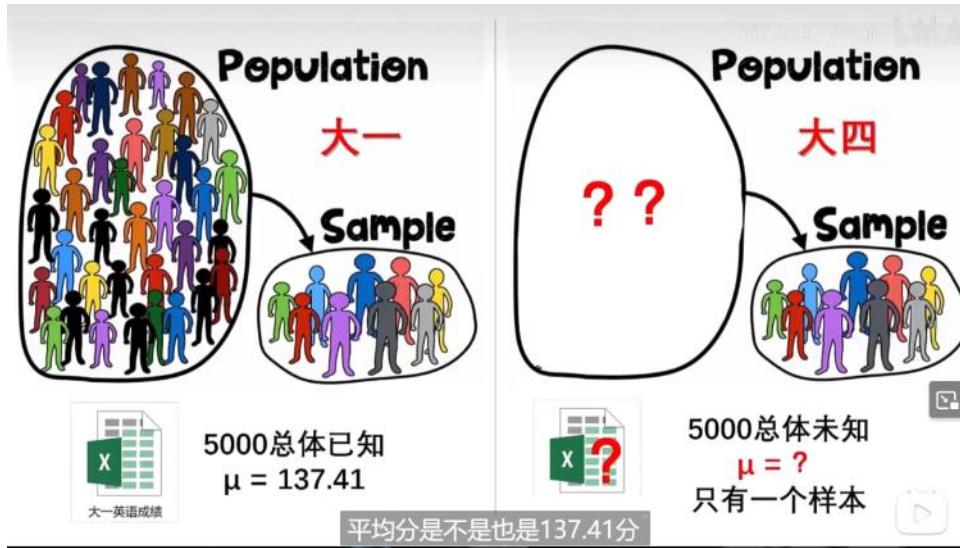
中心极限定理 – 大一英语成绩平均分



p值表 ($\alpha=0.1$)

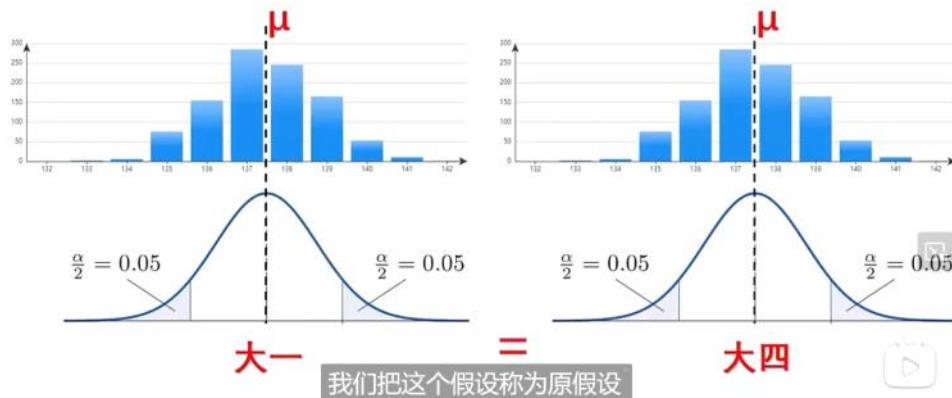
抽样均值	抽到此值概率 (p)	是否极端状况
132	0	是
133	0.2%	是
134	0.6%	是
135	7.6%	否
136	15.5%	否
137	28.5%	否
138	24.6%	否
139	16.5%	否
140	5.3%	否
141	1.1%	是
142	0.1%	是
143	0	是

已知大一的平均分是137.41，大四的平均分是多少？



假设检验的概念

H_0 : 大四的平均分和大一的平均分是一样的。 (Null Hypothesis 原假设)

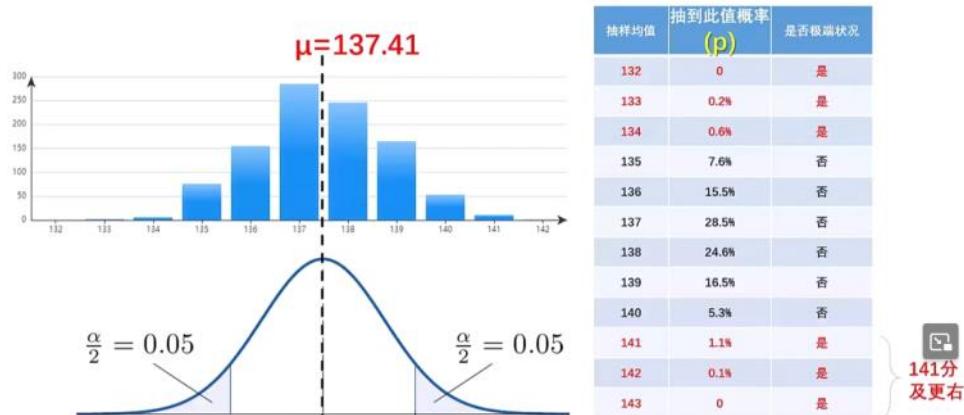


假设记为 H_0

大四随机抽样几次，样本容量为20，得到平均分为141.3，落在大一模型的尾巴上（柱状图1.2%），非常极端的小概率事件，所以推翻原假设。

大四抽样一次 抽样均值 $\bar{X} = 141.3$

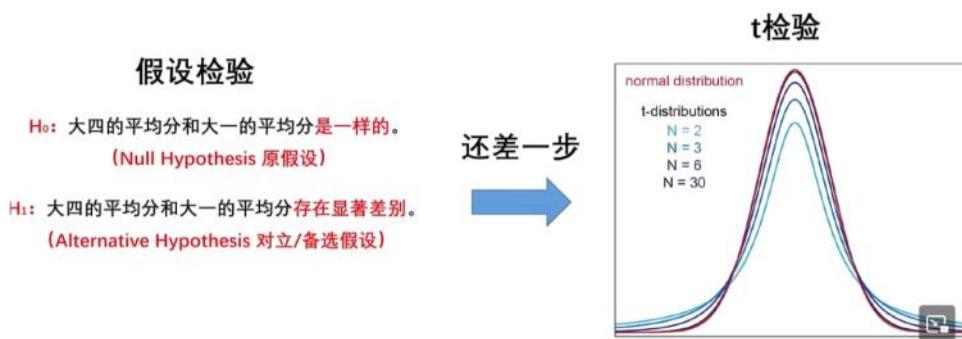
p值表 ($\alpha=0.1$)



~~H₀~~: 大四的平均分和大一的平均分是一样的。
(Null Hypothesis 原假设)

H₁: 大四的平均分和大一的平均分存在显著差别。
(Alternative Hypothesis 对立/备选假设)

正式表述：在90%与10%这种“大概率”与“小概率”的人为划分水平下“大四的英语高考平均分和大一平均分显著不同”。
(significance level 显著水平 $\alpha=0.1$)



正态分布是具有两个参数 μ 和 σ^2 的连续型随机变量的分布，第一参数 μ 是遵从正态分布的随机变量的均值，第二个参数 σ^2 是此随机变量的方差，所以正态分布记作 $N(\mu, \sigma^2)$ 。

t检验

2023年12月27日 10:25

全校平均分: 138.45

王老师班平均分: 137.8

138, 138, 136, 138, 133, 135, 144, 148, 139, 142, 136, 137, 138, 134, 121, 134, 147, 133, 142, 143

t检验 问题的专业提法:

“王老师班的平均分和全校平均分有没有显著差别?”

t检验专业问法: 王老师班的平均分和全校平均分有没有显著差别?



A校区



B校区



1000总体已知
 $\mu = 40$

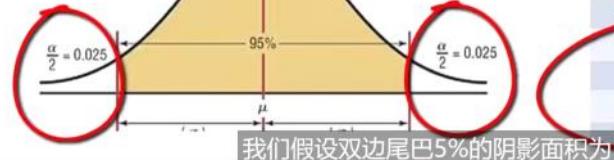
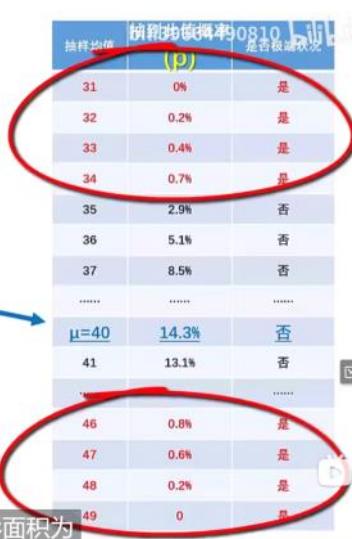
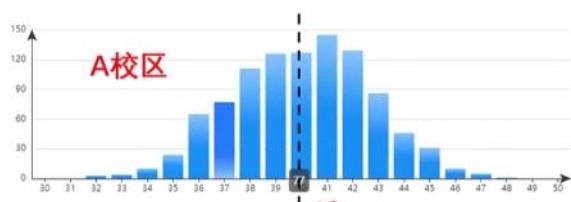


总体未知
 $\mu = ?$

不同校区间的原材料

A校区1000个人平均每月吃40个包子, B校区每月多少个?

A校区-B校区 包子平均数 假设检验



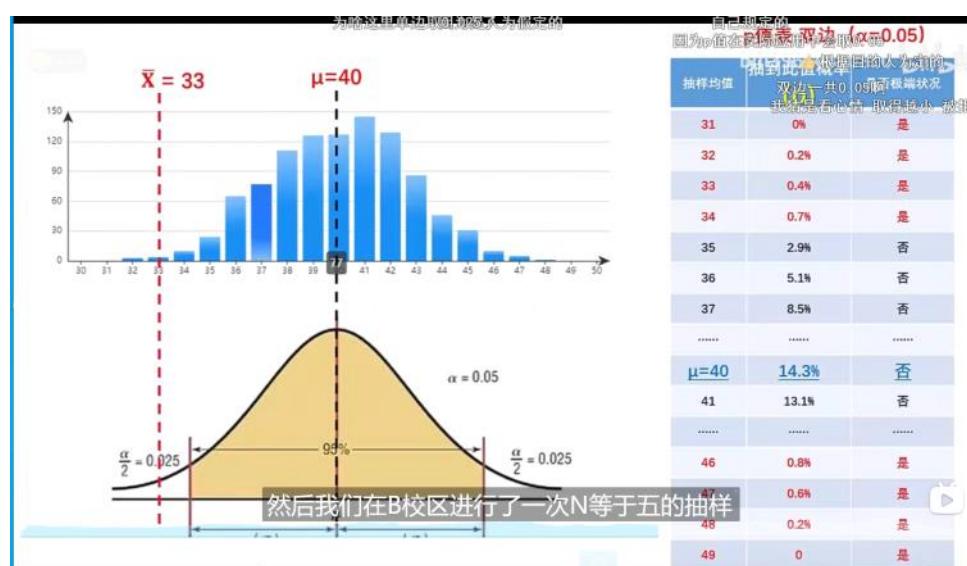
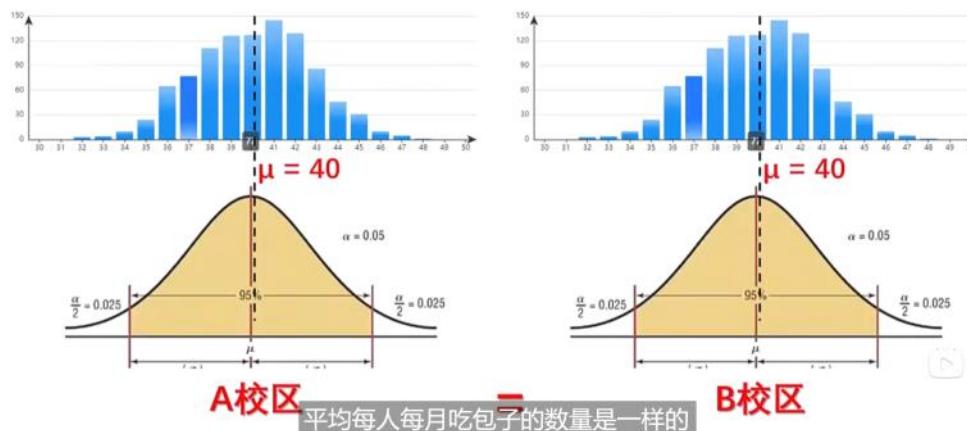
假设双边尾巴5% (单边2.5%) 为极端情况

通常,许多的科学领域中产生 **P值的结果 ≤ 0.05 被认为是统计学意义的边界线**,但是这显著性水平还包含了相当高的犯错可能性。结果 $0.05 \geq P > 0.01$ 被认为是具有统计学意义,而 $0.01 \geq P \geq 0.001$ 被

认为具有高度统计学意义。

A校区 B校区 包子数量假设检验

H_0 : B校区和A校区平均每人每月吃包子数量一样的。



在B校区进行了N=5的抽样，计算出平均值为33，落在左边尾巴为2.5%的阴影当中，是一个极端情况，拒绝原有假设。

从 均值抽样 到 t检验

Student's t-test

- William Sealy Gosset figured out how to test if a batch of beer was *significantly different* than the standard.



While working for the
Guinness brewing company,
he was forbidden to publish
academic research, so
published his method under
the pseudonym 'student'.

William Sealy Gosset
(1876–1937)

得到了一个纯粹意义上的统计T值

$$\frac{\sum_{i=1}^n x_i}{n} - \mu$$

即：样本均值-总体均值

标准差s

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

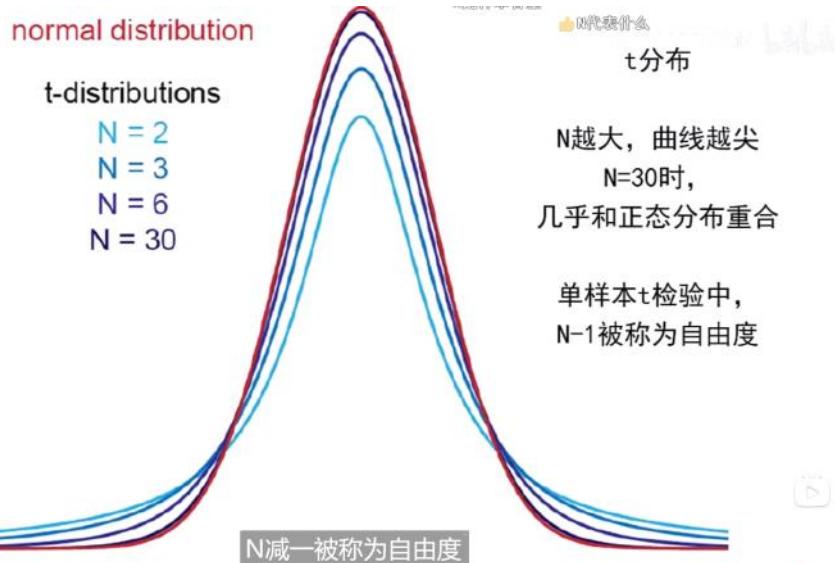
<https://zhuanlan.zhihu.com/p/88236209>

1) 样本标准差S:估计总体标准差

$$\text{标准误差 } se = \frac{S(\text{样本标准差})}{\sqrt{n}(\text{样本大小})}$$

2) $t = \frac{\text{样本均值} - \text{总体均值}}{\text{标准误差}}$

t检验公式



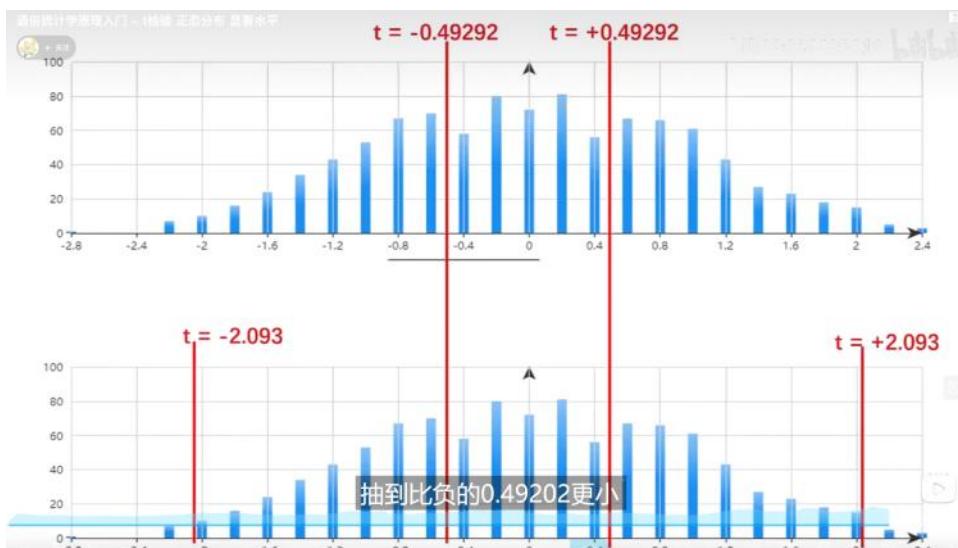
t分布图N是样本容量t值只和样本容量n相关，
根据t值，查找t表格，(n-1自由度，自由度查表) 得到p值

做出假设 H_0 王老师班的英语成绩和全校英语成绩没有显著性差异

```
> score=c(138,138,136,138,133,135,144,148,139,142,136,137  
> t.test(score,mu=138.45)
```

```
One Sample t-test

data: score
t = -0.49202, df = 19, p-value = 0.6283
alternative hypothesis: true mean is not equal to 138.45
95 percent confidence interval:
135.0349 140.5651
sample estimates:
mean of x
137.8
```



H_0 : 王老师班的平均分和全校平均分没有显著性差异。

(Null Hypothesis 原假设)

不能拒绝 H_0

无显著差异!

说服有力，
不扣绩效!

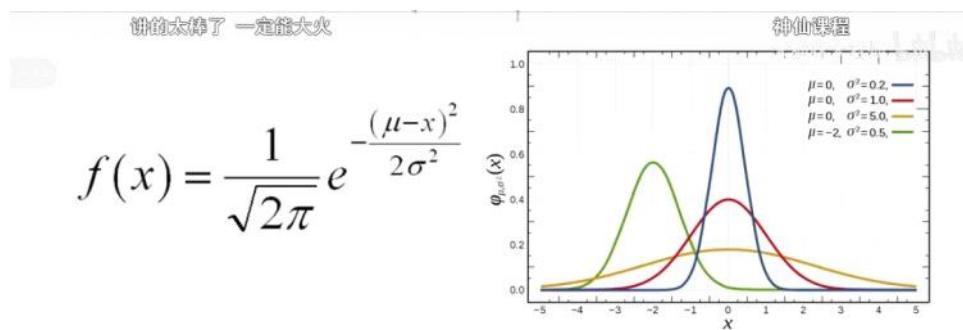
从全校随机抽取20个人的平均分为
137.8，属于正常的大概率事件。

和全校平均分有显著差异



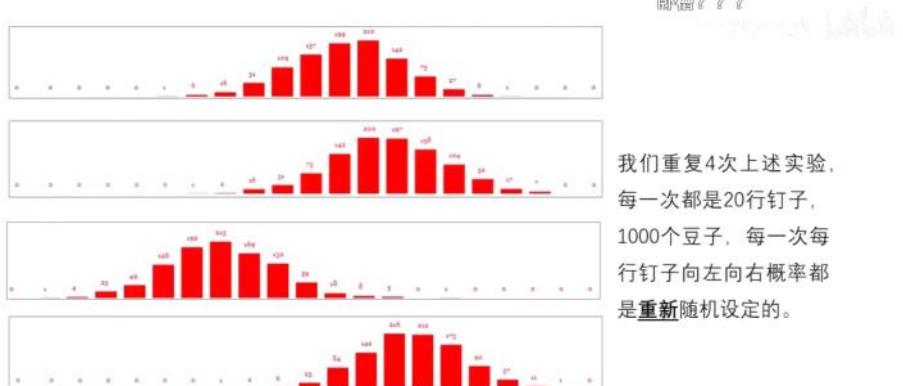
伯努利实验

2023年12月27日 16:29

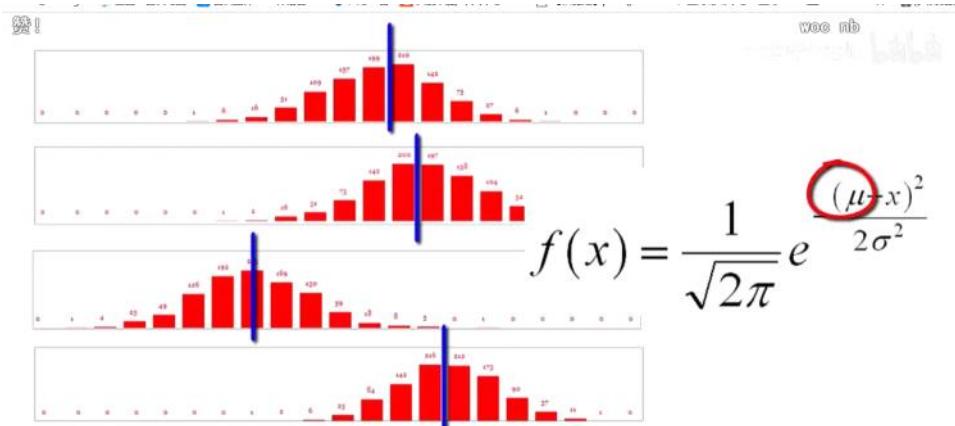


这个公式是正态分布曲线的**概率密度函数**。大家不用害怕，本节课不要求大家掌握这个公式。本课程系列只对公式中几个参数的含义进行感性理解。

我们看到，公式里有一个 π ，圆周率。你也许感到奇怪，正态曲线里没有圆形或圆弧啊，怎么会出来个 π 呢。这个问题，我也回答不了。我只能说，这就是数学的神秘之处，也是发现这些公式的数学家的伟大之处。我们向那些骨灰级的数学家们致敬！

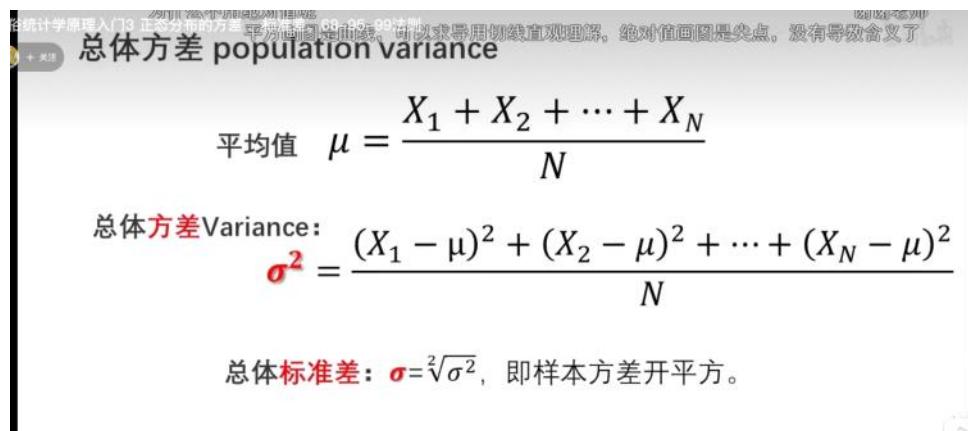


4次实验结果显示，豆子的轮廓都是光滑的正态分布曲线，只是曲线对称轴的位置有所不同。

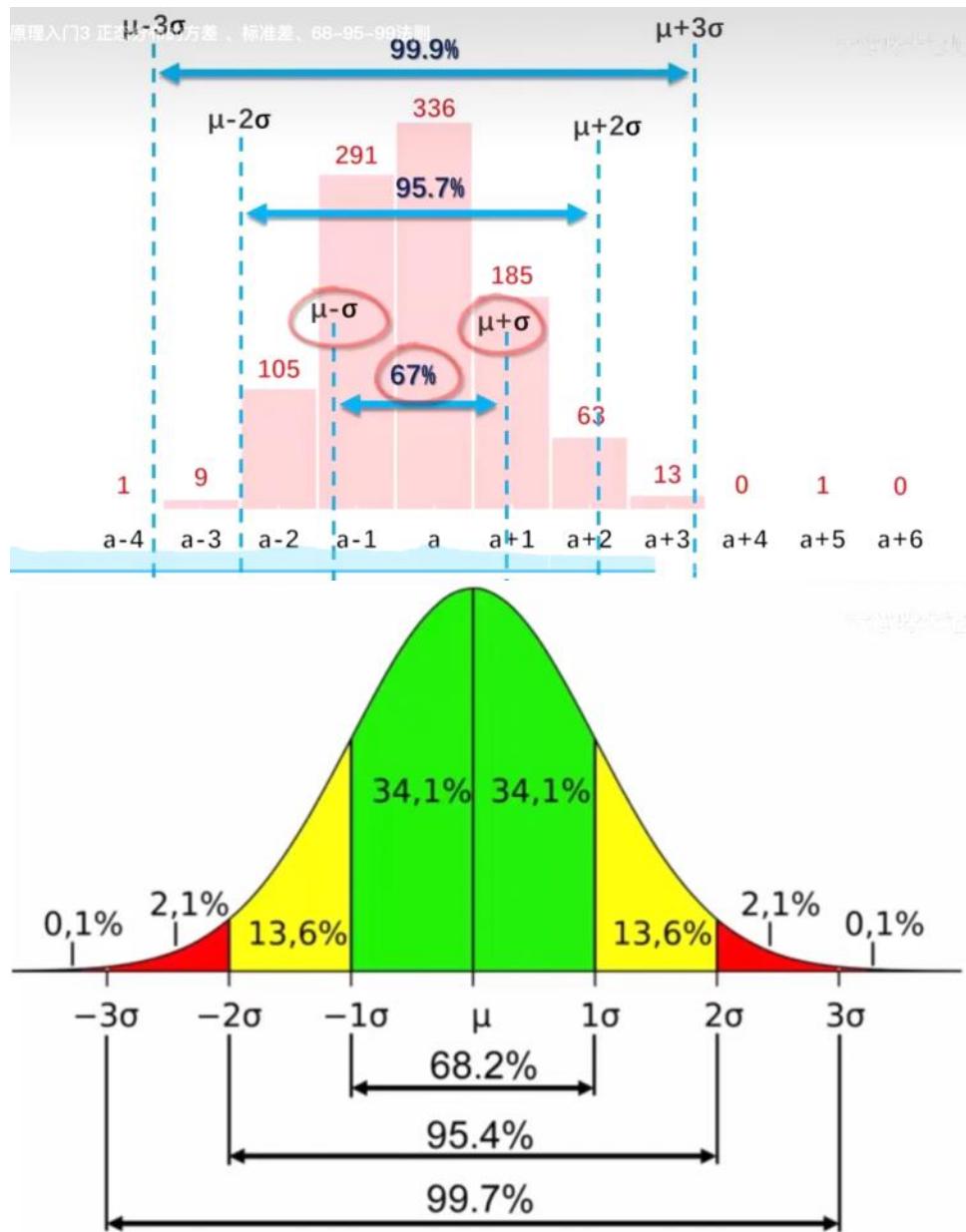


方差

2023年12月27日 16:53

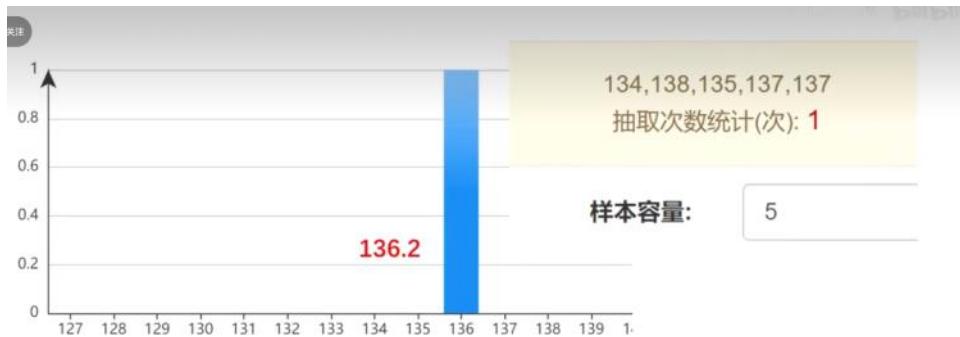


68-95-99原则

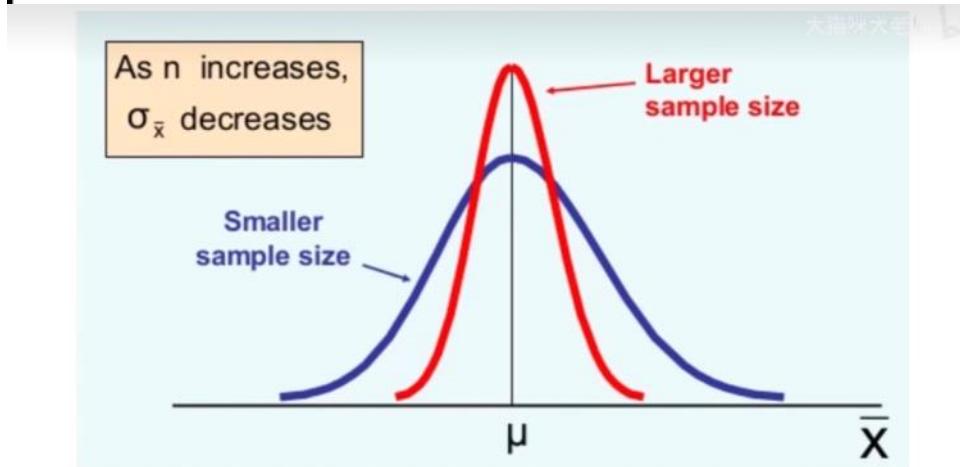
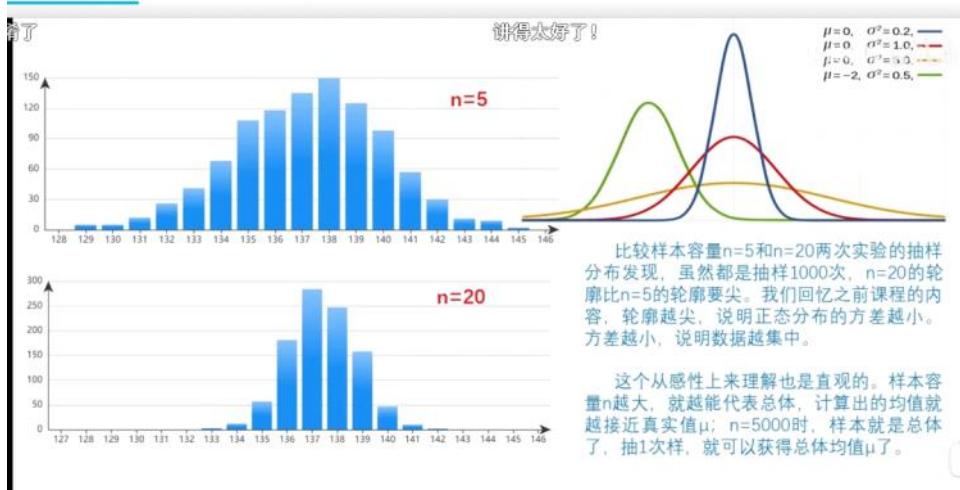


中心极限定理

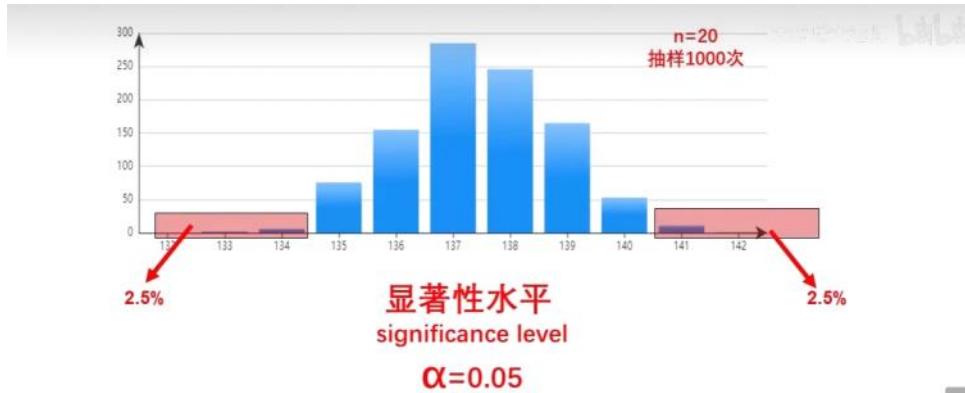
2023年12月27日 17:14



现在，我们把这5000个成绩加载到实验程序中，设定抽样的样本容量为5，我们手动抽样一次，抽出的5个分数在这里显示，这5个分数的平均分计算出来，是136.2分，那么我们在横坐标轴上，找到136附近的区间段，在这里画一个高度为1的柱状图，相当于我们在伽尔顿板实验中，在136的槽子里放了1颗豆子。



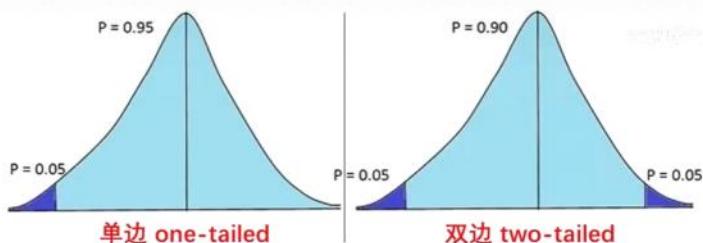
那么，假设你要通过抽样来估计大一新生英语平均分的话，你会选择n=5还是n=20呢？当然是选择样本容量n=20，这样抽样分布就会更集中，方差更小。



上面的表述有点啰嗦。我们简化一下说法。这个双边尾巴5%，也就是每一边2.5%的阈值，是人为规定的。这个阈值水平，叫做“显著性水平”，英语叫significance level，记为 α 。 α 也可以是2%，1%甚至0.1%等等，这个要看具体案例的具体分析。



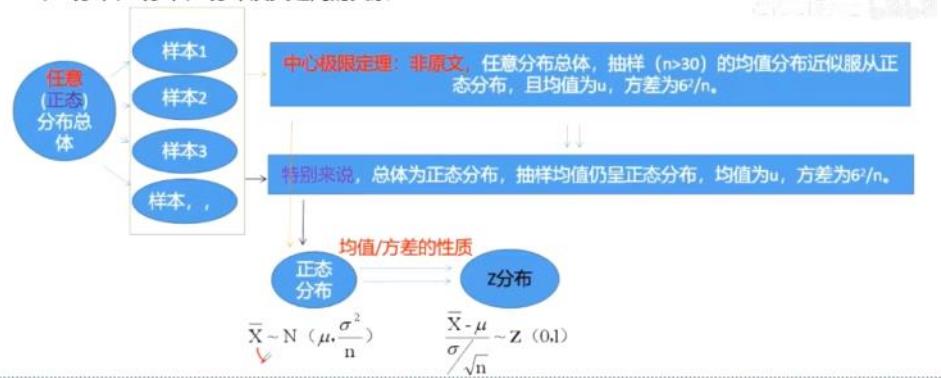
例如，假如这个大学的校长比较严格，觉得 $\alpha=5\%$ 太宽松了，觉得大一大二的平均分差个3分还是比较常见的，觉得才差个3分就算极端情形实在是说不过去。后来校长就说，把 $\alpha=0.05$ 紧一紧，以后， $\alpha=0.01$ 才算极端。



$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \quad \text{单样本 one sample} \quad \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \text{双样本 two sample}$$

当然，本节课只是展示了检验假设的其中一种形式，即，把双边尾巴作为“极端”的情形。接下来的t检验课程中，我们会学习单样本单边检验、双样本单边检验等更实用的形式。有了本节课的基础，到时候大家就肯定不会被这些绕口的概念绕晕了。我们下节课再见。

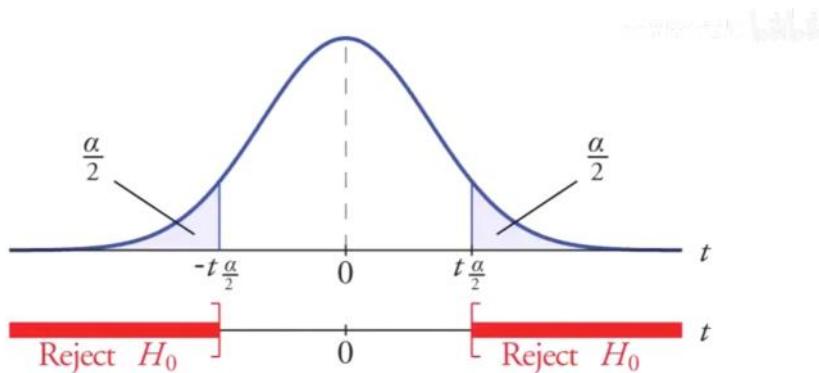
二、N分布、Z分布、t分布及其之间的关系



我们看到我们最原始的这个样本的均值是 X_8

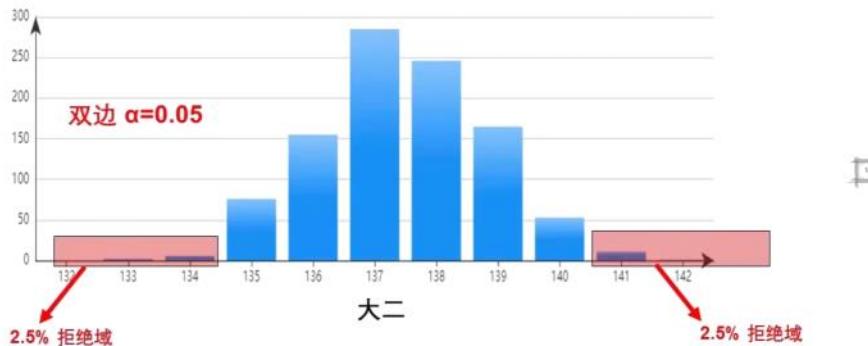
关键一步：从均值抽样分布到t分布

2023年12月31日 19:22

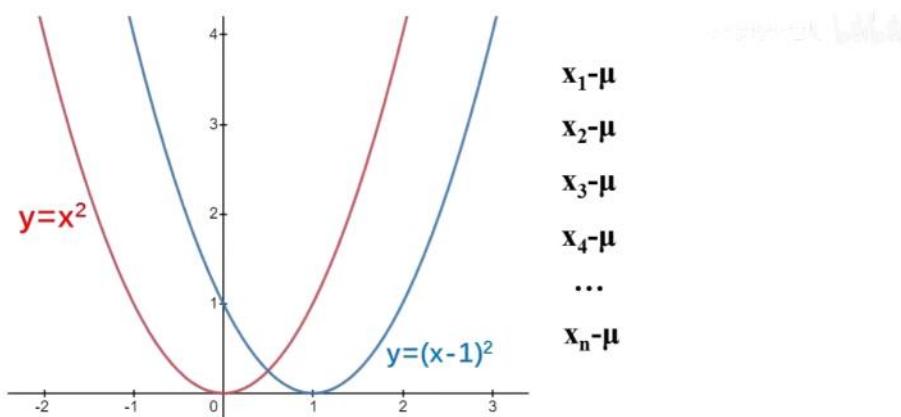


本节课，我们将正式引入t分布。t分布是感性理解t检验的重要一步。t检验，就是t值假设检验。在学习本课前，希望大家再复习一下之前的几节课，以便无缝衔接，理解本节课的内容。之前所有的课程，都是为了本节课做铺垫。本节课中，我们将直接运用前面学过的一些术语或概念，不再做通俗解释了。

H_0 : 大一和大二的英语成绩平均分没有显著差别。



我们先假设 H_0 : 大一和大二的英语成绩平均分没有显著差别。然后设定显著性水平 α 为双边0.05，并划出拒绝域。此时，从大一新生中抽样一次，获得一个样本均值，141分。141分落入了拒绝域。抽样一次，极端事件便发生了。所以，我们拒绝 H_0 ，接受 H_1 。即：在双边 $\alpha=0.05$ 水平下，大一和大二的英语均值存在显著差别。



第一步，处理对称轴。在中学的解析几何中，我们都学过曲线的对称轴，都知道对称轴是y轴，也就是 $x=0$ 的时候，最方便处理数据。怎样才能把均值抽样分布的对称轴移到y轴呢？很简单，就是把excel表总体中的每个数据，都减去总体均值 μ 。这样，每次抽样的均值就是之前的样本均值减去 μ 。

A	B
1	126.00 -11.41
2	136.00 -1.41
3	139.00 1.59
4	141.00 3.59
5	124.00 -13.41
6	136.00 -1.41
7	147.00 9.59
8	143.00 5.59
9	137.00 -0.41
10	141.00 3.59
11	150.00 12.59
12	133.00 -4.41
13	131.00 -6.41
14	134.00 -3.41
15	132.00 -5.41



A	B
1 126.00	-11.41
2 136.00	-1.41
3 139.00	1.59
4 141.00	3.59
5 124.00	-13.41
6 136.00	-1.41
7 147.00	9.59
8 143.00	5.59
9 137.00	-0.41
10 141.00	3.59
11 150.00	12.59
12 133.00	-4.41
13 131.00	-6.41
14 134.00	-3.41
15 132.00	-5.41
16 133.00	-4.41
4991 138.00	0.00
4992 150.00	12.59
4993 141.00	3.59
4994 134.00	-3.41
4995 130.00	-7.41
4996 135.00	-2.41
4997 125.00	-12.41
4998 129.00	-8.41
4999 145.00	7.59
5000 133.00	-4.41

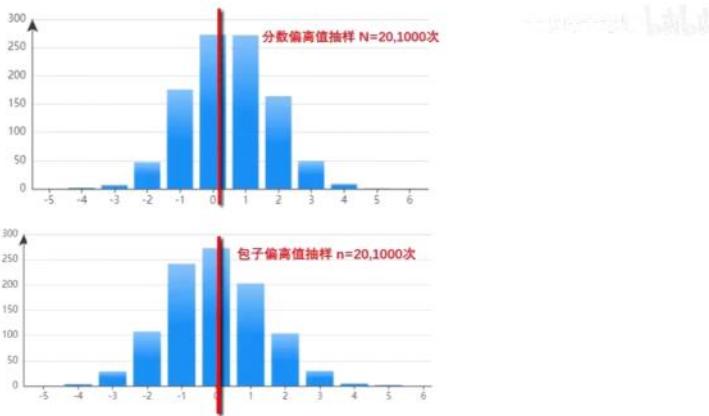


分数偏离值表

现在，我们打开英语成绩excel表，把所有的分数值都减去总体均值 $\mu=137.41$ 分，并四舍五入为整数，因为我们这个虚拟仿真程序姑且只能抽样整数，然后另存为一个新的excel表，这个新表，代表着每个同学的成绩偏离总体均值 μ 的程度，我们姑且把它叫做“偏离值表”。



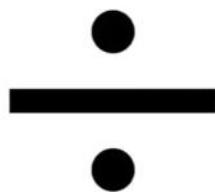
然后，对这个偏离值表进行均值抽样实验。最终仍然得到一个类似正态分布的均值抽样分布。它的对称轴，变成了y轴。这个在直观上也是很容易理解的，即总体中每一个数值偏离总体均值的程度，肯定是在0的左右波动的。



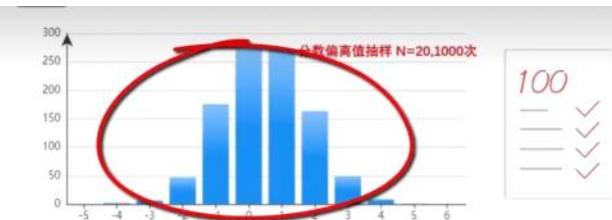
这时，我们再比较这两个不同案例的抽样分布，发现他们更相似了，对称轴都是y轴了。假如不仔细考虑的话，你可能会认为，这两个分布现在已经一模一样了。我们不要忘了，它们的单位还是不一样的。一个是“分”，一个是“只”。单位不一样，就是本质上不一样，还需要继续利用数学来抽象。

分区

100
— ✓
— ✓
— ✓
— ✓



现在进行第二步。怎么样才能把单位去掉呢？你猜对了。小学数学就学过的，用除法，就可以把单位约掉。



分



只



那么，要把上一步得到的这个偏离值，除以一个什么东西，才能把单位约掉呢？



从统计学的定义

样本标准差不是总体标准差，分母为 $n-1$ 会更接近总体标准差

除以标准差的一个主要功能是“标准化”

各位实践证明，用 $n-1$ 比用 n 更能保证

$$\text{标准差 } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

一次抽样样本: -9, -8, 0, -9, -6, 0, 7, 5, 1, 10, 6, 4, -8, -8, 2, -6, 8, 8, 5, 4

样本均值: 0.3



中学数学里，我们恰好学过一个概念，叫标准差。我们每抽样一次，都可以算出一个样本的标准差，记作小 s 。它表示本次抽样样本的内部，各个数值偏离样本均值的离散程度。对于样本标准差，本课程不做严格数学意义上的表述和讲解，谨供大家感性理解。但最重要的是，它是一个带单位的量。



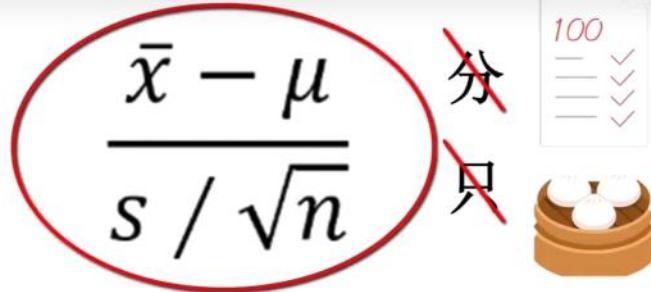
t值是什么？

t值是样本均值偏离总体均值的程度/样本内部的离散程度

均值分布的标准差

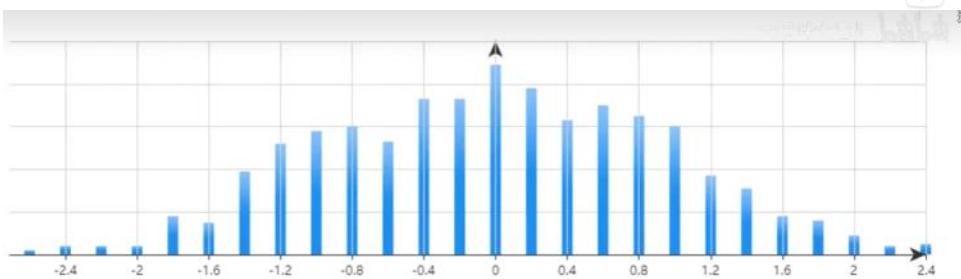
$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

分
只





现在, 我们用这个样本均值偏离总体均值的程度, 除以样本内部的离散程度, 就约掉了“分”和“只”这些不同案例中的单位。我们再在分母上除以一个根号n (别问为啥, 否则又成了数学课了), 就得到一个新公式。这个新公式计算出来的这个值, 不考虑各种总体的千变万化的 μ 值, 也不考虑各种案例的五花八门的单位, 它是抽象出来的一个纯数学的值, 我们把它叫做 **t值**。

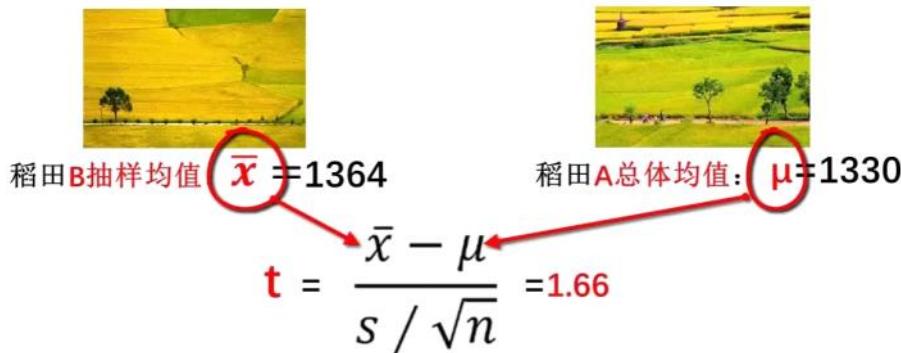


现在, 我们有了一个统一的、万能的、包罗万象的t分布, 就再也不必为每一个具体的案例制造一个均值抽样分布了。我们根据t分布, 也造出一个概率表, 就可以对任意总体和样本进行假设检验了。

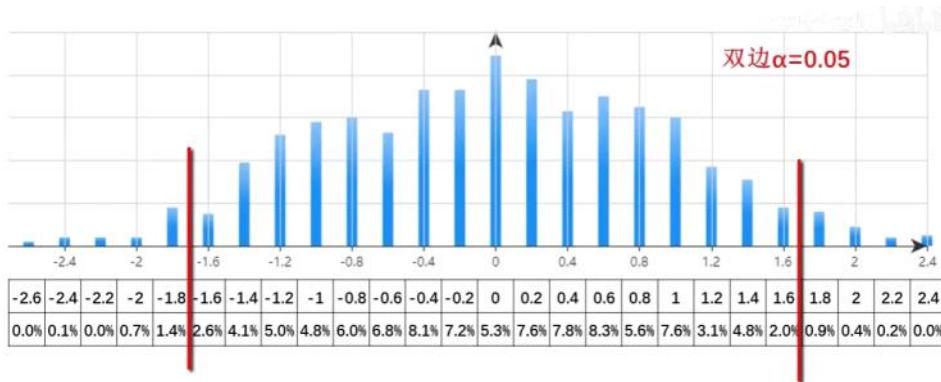


那么, 我们想知道, 在双边 $\alpha=0.05$ 的显著水平下,
稻田B的亩产量和稻田A的亩产量 **有没有显著差别?**

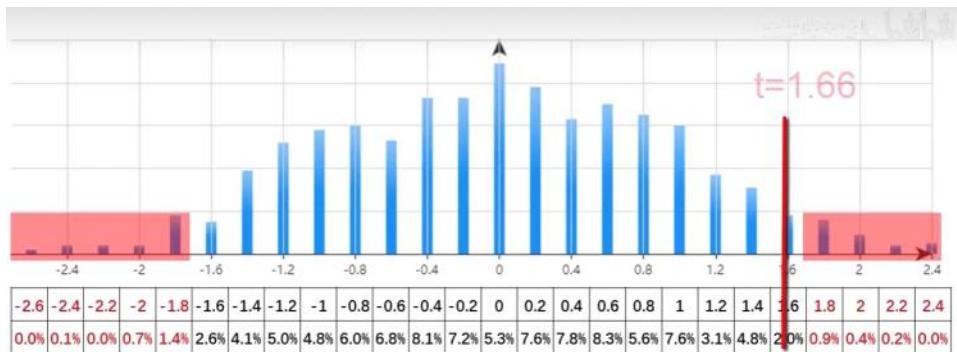
稻田B抽样: 1300, 1400, 1350, 1380, 1420, 1530, 1480, 1390, 1240, 1290, 1440, 1260, 1220, 1470, 1320, 1380, 1240, 1360, 1500, 1310



根据这些信息, 已经可以计算出稻田B抽样的t值了。根据公式, $t=1.66$ 。
下面, 要看这个t值是否落入了t分布的拒绝域。



首先, 在t分布的概率表中, 划出双边 $\alpha=0.05$ 的临界值。我们从左边数出累计2.5%的概率, 划出临界值为-1.8。再从右边数出累计2.5%的概率, 划出临界值为1.8。这样就获得了拒绝域。

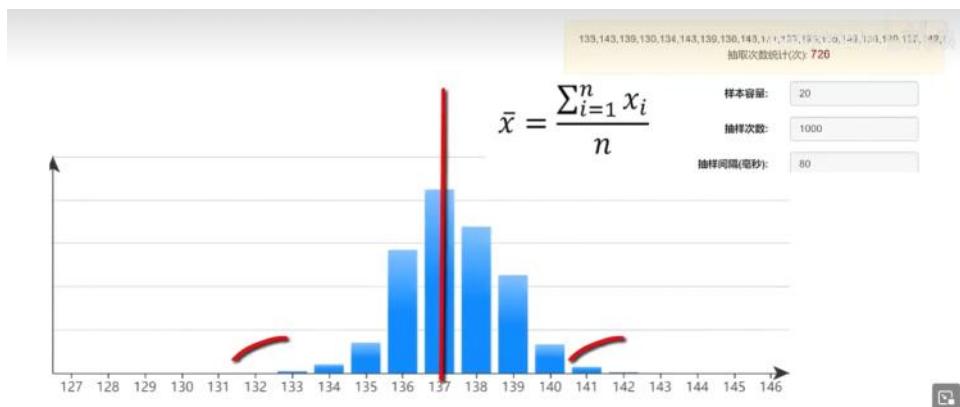


H_0 : 稻田B的亩产量和稻田A的亩产量没有显著差别。

现在t值有了, 拒绝域也有了。我们发现, 本次抽样t值没有落入拒绝域。因此, 在双边 $\alpha=0.05$ 的显著水平下, 不能拒绝 H_0 , 仍然相信: 稻田B的亩产量和稻田A的亩产量确实没有显著差别。

单边假设t检验

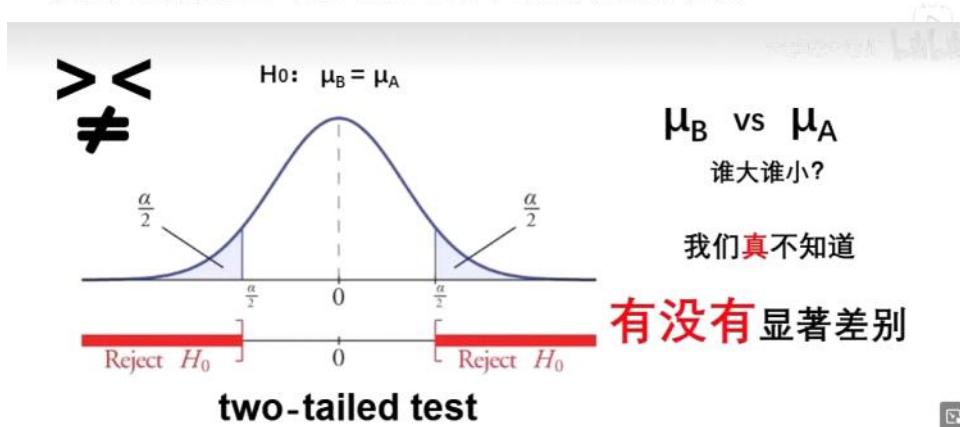
2023年12月31日 19:42



大家好，在之前的课程中，我们学习了均值抽样分布，并通过抽样分布学习了假设检验的概念。例如，一个学生总体A的英语平均分是 $\mu_0=137$ 分，对这个总体进行均值抽样的话，均值抽样分布的对称轴必然为137分，分布的双边尾巴上的，是比137分低很多分或者高很多分的“极端”情况。



142分落入了右边拒绝域，因此我们拒绝了 H_0 。假如样本均值是129分，落入了左边的拒绝域，我们也会去拒绝 H_0 。不管落入左边还是右边的拒绝域，我们都拒绝 H_0 ，得到结论是“ H_1 ：总体B和总体A的均值有显著性差别”。热爱思考的同学可能早就想问这么一个问题：那总体B和总体A，到底谁的均值大谁的均值小呢？

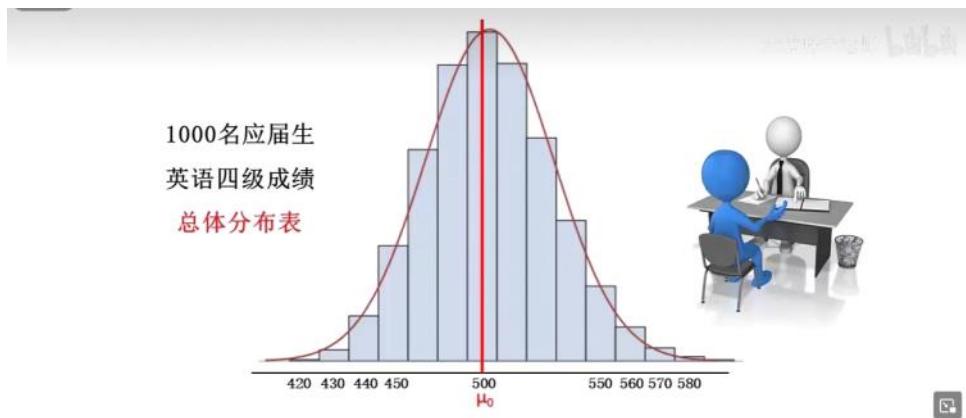


答案是：我们不知道。上面的假设检验，就叫做“双边检验”，也叫“双尾检验”，英语叫two-tailed test。因为我们开始的假设是 $\mu_B = \mu_A$ ，所以总体B的抽样均值，比 μ_A 大也好，比 μ_A 小也好，只要不和 μ_A 相等，就足够检验了我们的这个假设了。所以基于这个“ $\mu_B = \mu_A$ ”的 H_0 假设，我们只能得出到底“有没有显著差别”的结论，得不出“谁大谁小”的结论。





假如你从大学毕业了，创业开了个工厂，自己当老板，现在需要招聘员工。现在各个招聘单位都会对员工的英语水平有一定要求。但是，你的要求比较特殊，你既不要英语太差的，也不要英语太好的。你觉得，英语太差的，不够用，看个英语说明书也看不懂；但你同时也觉得，英语太好的，也没用，你工厂也不需要直接和国外打交道，英语太好的，你还要多给工资，不划算。



下面接着编故事。你工厂开了几年，直接国际化了，你非常需要英语非常好的员工。这一年，你又来这个高校里招聘。你说，英语四级成绩太差的，我一律不招。我要英语四级成绩“好一点”的。那么，还是这个成绩总体的正态分布，你需要划一条线，划出成绩最差的5%，作为你的拒绝域。你该怎么划线呢？



我们仍然用上节课稻田亩产量的例子。有稻田A和稻田B，已知稻田A的亩产量为 $\mu_A=1330$ 斤，现在对稻田B进行一次样本容量为20的抽样，计算得样本均值 \bar{x} 为1364斤。请问，稻田B的亩产量比稻田A大还是小呢？显然，肉眼可见，1364大于1330。但是，怎么通过单边t检验来得出“ $\mu_B > \mu_A$ ”的结论呢？

欲擒故纵
H₁ H₀
欲扬先抑

H₁ : $\mu_B > \mu_A$

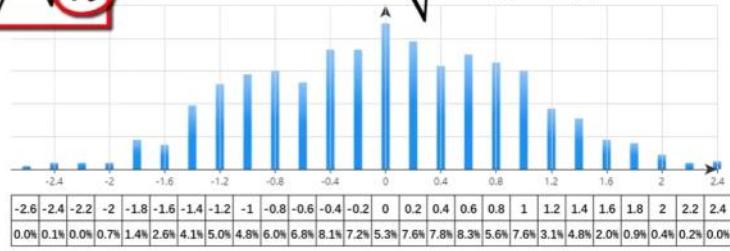
H₀ : $\mu_B \leq \mu_A$

假设检验有个常见套路，就是虽然你想得到的结论是H₁，但你得先假设H₀成立，然后通过检验拒绝H₀，再证实H₁。例如，稻田B的抽样均值为1364斤，我们想得到的结论H₁是“ $\mu_B > \mu_A$ (1330斤)”。但我们要先做原假设“H₀: $\mu_B \leq \mu_A$ (1330斤)”，然后尝试拒绝这个H₀来证实H₁。

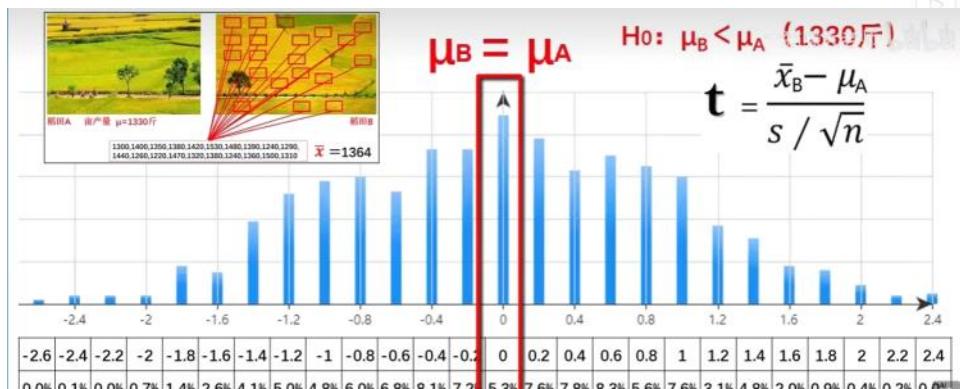
单边t检验公式

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

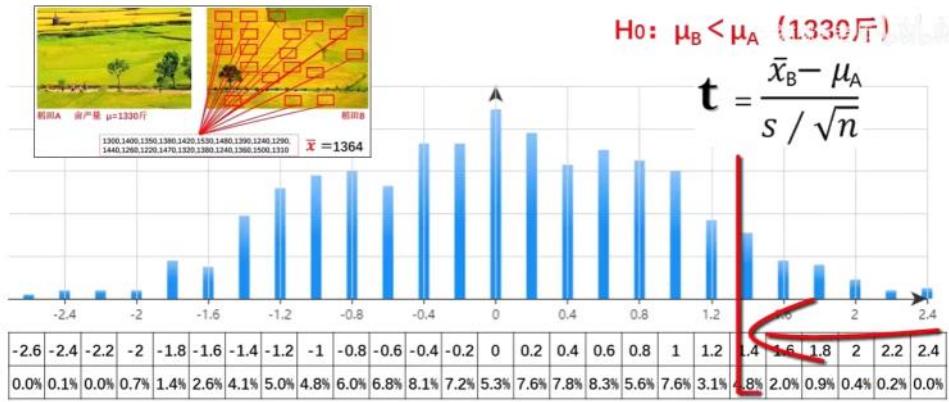
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$



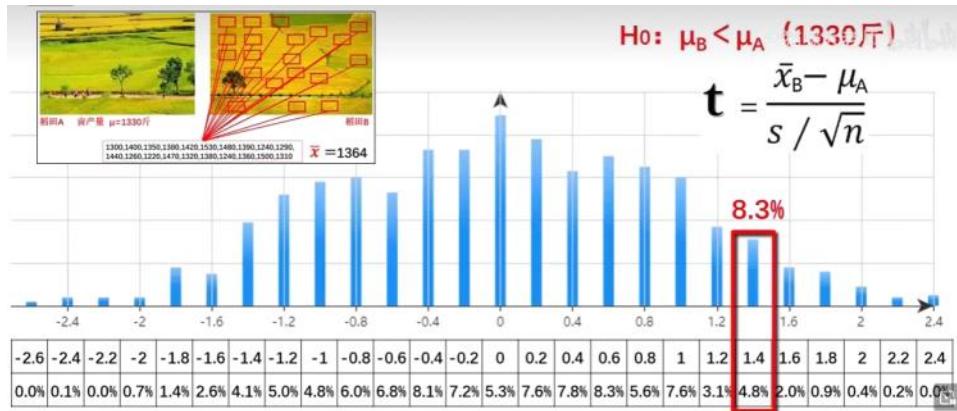
这是上节课得出的万能的t分布和t值表。在单边划线之前，我们先回忆一下t值的公式。这是一个分式。分母肯定是个正数，因为标准差和n都是正数。分子是正是负就不一定了。这是因为，样本均值 \bar{x} 与总体均值 μ 相减，若 \bar{x} 大于 μ ，则t值就是个正数，若 \bar{x} 小于 μ ，则t值就是个负数。



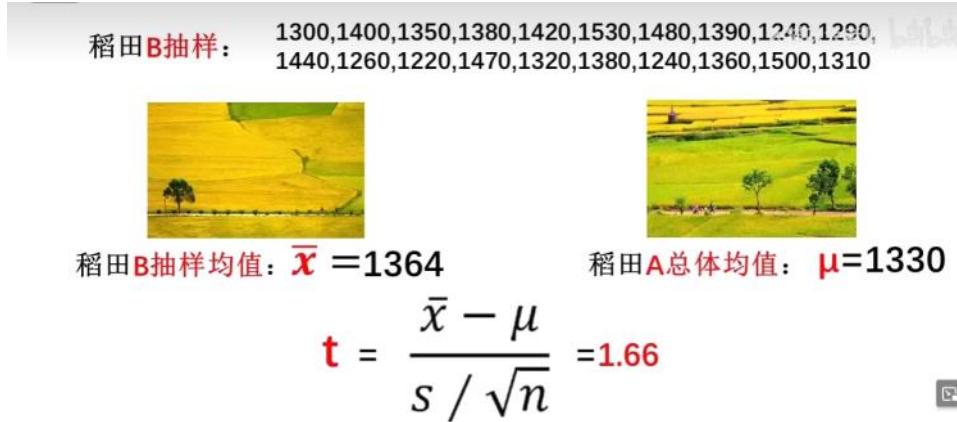
在本例中，t=0这一点，代表 $\mu_B = \mu_A$ (1330斤)。那么现在，请你划出最不能代表“H₀: $\mu_B \leq \mu_A$ (1330斤)”的5%的t值区域。怎么划呢？ μ_B 要是小于 μ_A 的话，根据t值公式，t应该是个负数。所以，t值越小于0，越往左边尾巴尖，就越能代表“H₀: $\mu_B \leq \mu_A$ (1330斤)”。反之，t值越大于0，越往右边尾巴尖，就越不能代表“H₀: $\mu_B \leq \mu_A$ (1330斤)”。所以，我们从右边尾巴尖这里，往左数出5%的t值，作为H₀的拒绝域。



在本例中, $t=0$ 这一点, 代表 $\mu_B=\mu_A$ (1330斤)。那么现在, 请你划出最不能代表“ $H_0: \mu_B < \mu_A$ (1330斤)”的5%的t值区域。怎么划呢? μ_B 要是小于 μ_A 的话, 根据t值公式, t 应该是个负数。所以, t 值越小于0, 越往左边尾巴尖, 就越能代表“ $H_0: \mu_B < \mu_A$ (1330斤)””。反之, t 值越大于0, 越往右边尾巴尖, 就越不能代表“ $H_0: \mu_B < \mu_A$ (1330斤)””。所以, 我们从右边尾巴尖这里, 往左数出5%的t值, 作为 H_0 的拒绝域。



我们开始数。2.2这里的概率0.2%, 累计0.2%, 2这里概率是0.4%, 累计0.6%, 1.8这里是0.9%, 累计1.5%, 1.6这里是2.0%, 累计2.5%, 1.4这里是3.5%, 超过5%了。所以, 单边右边的 $\alpha=0.05$ 临界值为1.6。超过1.6再往对称轴方向数一点点, 就超过5%了。所以, 单边右边 $\alpha=0.05$ 的拒绝域就是包括 $t=1.6$ 在内的再往右的所有区域。



那么, 我们稻田B的样本计算出来的t值是多少呢? t值还是原来的t值1.66, 不管是双边还是单边的, 只要样本一样, t值都是一样的。t值等于1.66, 落入了拒绝域。所以我们在单边 $\alpha=0.05$ 的显著水平下, 拒绝此样本可以代表“ $H_0: \mu_B < \mu_A$ (1330斤)”, 接受此样本可以代表“ $H_1: \mu_B > \mu_A$ (1330斤)”, 即“稻田B的亩产量显著大于稻田A”。

上节课



稻田A

大猫咪大老师

$$H_0: \mu_B = \mu_A$$

稻田B

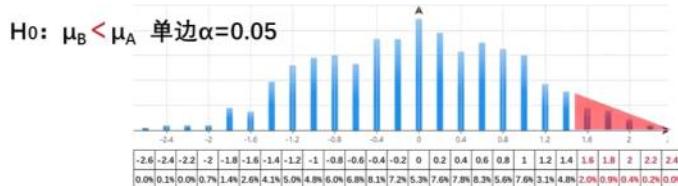
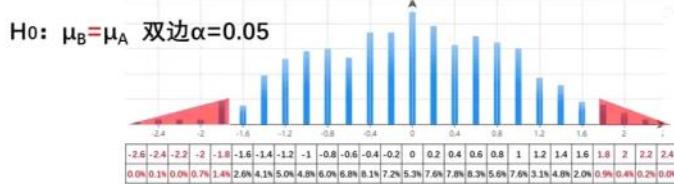
H_0 : 稻田B的亩产量和稻田A的亩产量没有显著差别。
(双边 $\alpha=0.05$)

本节课

$$H_1: \mu_B > \mu_A \text{ (1330斤)}$$

H_1 : 稻田B的亩产量显著大于稻田A

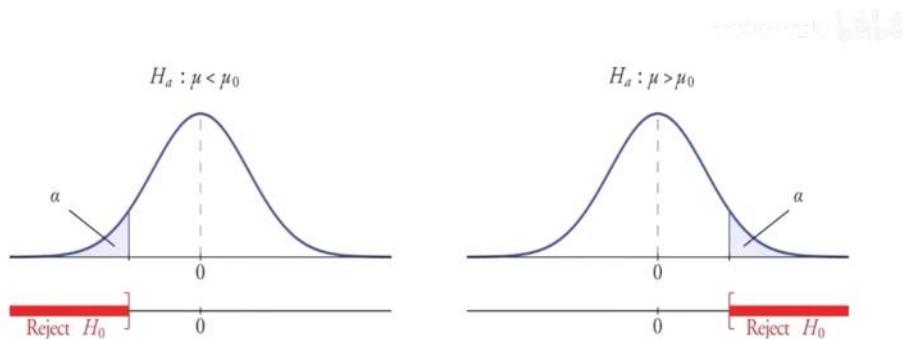
热爱思考的同学又要问了。为什么是同样的样本和同样的 t 值。在上节课双边检验中得出“稻田B和稻田A的亩产量没有显著差别”，在本节课中，又得出了“稻田B的亩产量显著大于稻田A”的结论。这显然不是矛盾的吗？



原因是，我们的原假设变了，所以拒绝域也变了。双边变成了单边，显著水平还是0.05，拒绝域从两边尾巴尖的面积，挪到了一边的尾巴尖的面积。这单边的尾巴，肯定比原来的双边的尾巴要粗，也就是说，单边的拒绝域更大的，更容易拒绝 H_0 。这也就是为什么在实际统计工作中，很多人都不用0.05这个显著水平，而使用0.01，或者0.001等更小的显著水平。

自由度

2023年12月31日 20:35



大家好。我们已经学过了单双边t检验的基本概念和原理。这节课，我将尝试通俗讲解一下t检验里的一个重要概念，“自由度”，Degrees of freedom，简写为df。



有一年过年，爷爷给我20块钱压岁钱。我准备带爷爷去小饭馆搓一顿。我有两个目标：第一，要点4个菜；第二，4个菜的价钱加起来正好是20元，我要把压岁钱全部花光，一分钱都不剩。



所以说，虽然表面上是点4个菜，但为了满足总价等于20块钱这个限制条件，我其实只能自由的选择3个菜，而第4个菜的价钱，是被前3个菜决定了的，第4个菜不能自由选择。换句话说，我点4个菜的自由度其实为4-1=3。

$\{x_1, x_2, x_3, x_4\}$

样本容量n=4

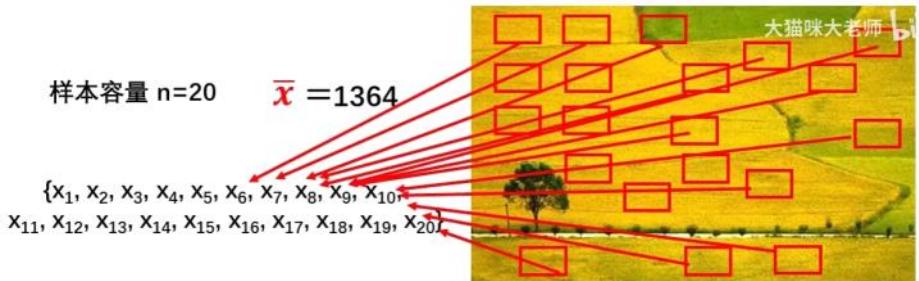
$$x_1 + x_2 + x_3 + x_4 = 20$$

其中 3 个可自由变动

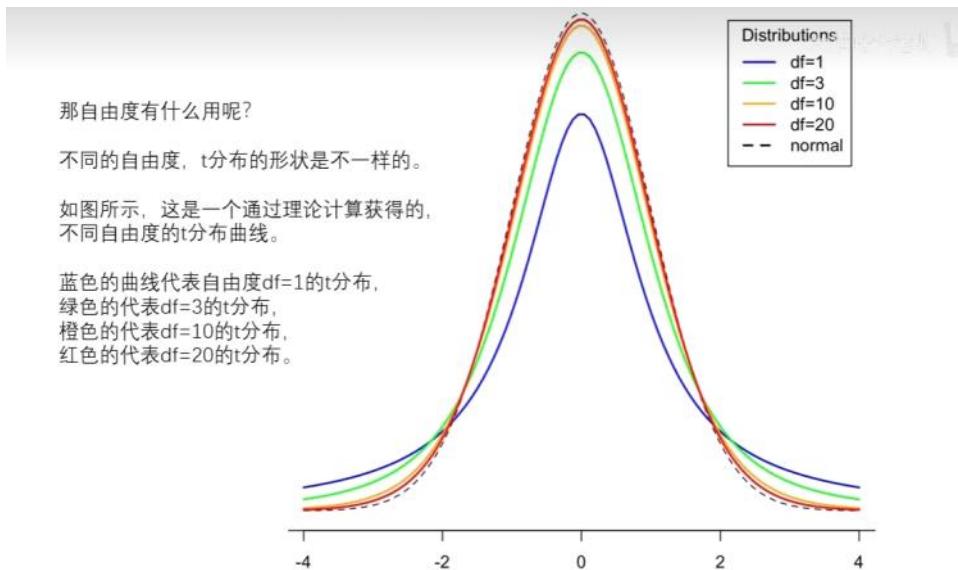
$$\bar{x} = (x_1 + x_2 + x_3 + x_4) / 4 = 5$$

第 4 个因限制而确定
可计算得出

假如我们把点菜看成抽样，那么这个样本就可以记作： $\{x_1, x_2, x_3, x_4\}$ ，样本容量n=4。限制条件是，总价必须等于20块钱，可以记作： $x_1 + x_2 + x_3 + x_4 = 20$ ；变换一下形式，就可以写成样本均值 $\bar{x} = (x_1 + x_2 + x_3 + x_4) / 4 = 5$ 。样本中 x_1, x_2, x_3, x_4 这4个变量，能自由变动的，只有3个。一旦其中3个变量确定了，第4个变量便也就确定了。所以说，第4个变量是不可以自由变动的。于是，在样本容量n=4，样本均值 $\bar{x}=5$ 的条件限制下，这个样本的自由度df为n-1=3。



回到我们之前讲过稻田亩产量的抽样，其样本容量为n=20，样本均值为 $\bar{x}=1364$ 。这个样本由20个变量组成，这一组20个变量中，能自由变动的变量只有19个；一旦其中19个变量确定了，为了满足 $\bar{x}=1364$ 这个限制条件，第20个变量就只能计算出来了，是确定了的，是不能自由变动的。所以，在样本容量n=20，样本均值 $\bar{x}=1364$ 的条件下，这个样本的自由度df为n-1=19。



那自由度有什么用呢？

不同的自由度，t分布的形状是不一样的。

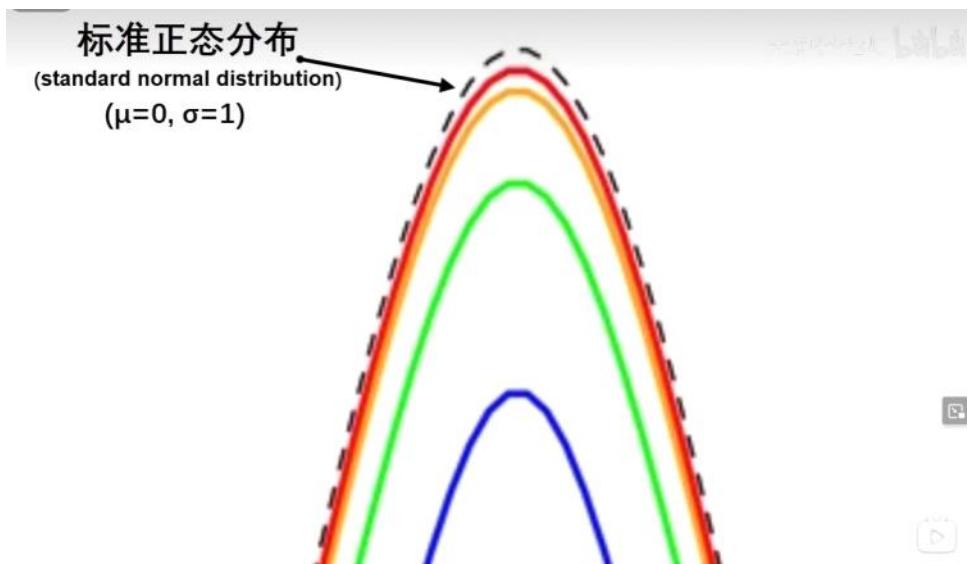
如图所示，这是一个通过理论计算获得的，不同自由度的t分布曲线。

蓝色的曲线代表自由度df=1的t分布，

绿色的代表df=3的t分布，

橙色的代表df=10的t分布，

红色的代表df=20的t分布。

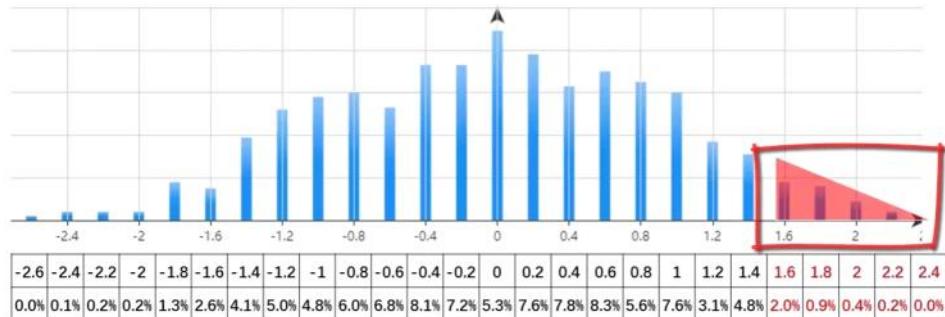


t|临界值表

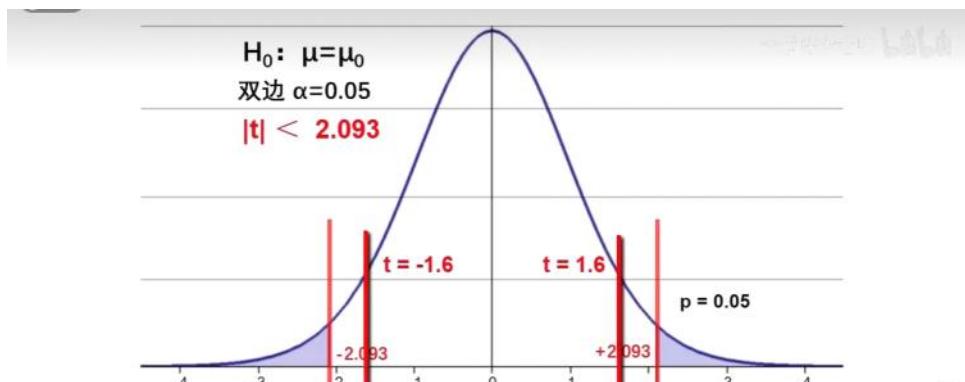
2023年12月31日 20:44

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.994	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.303	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.265	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.688	0.863	1.069	1.333	1.740	2.110	2.567	2.986	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.088	2.526	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.506	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300

$H_0: \mu < \mu_0$ n=20, df=19, 单边右边 $\alpha=0.05$



临界值的概念，其实我们之前的课程中已经接触过了。例如，这是一个样本容量n=20、自由度df=19、单样本抽样1000次形成的一个t值分布图。假如我们人为划定单边右边5%为极端情况，也就是说，我们的原假设 $H_0: \mu < \mu_0$ ，显著水平为单边右边 $\alpha=0.05$ 。我们要把这1000次抽样中，最右边的50次抽样，也就是这5%的阴影面积，作为极端情形。这部分的阴影面积，就叫做拒绝域。拒绝域的这里有一条边界线，这条边界线所代表的t值，就是临界值。

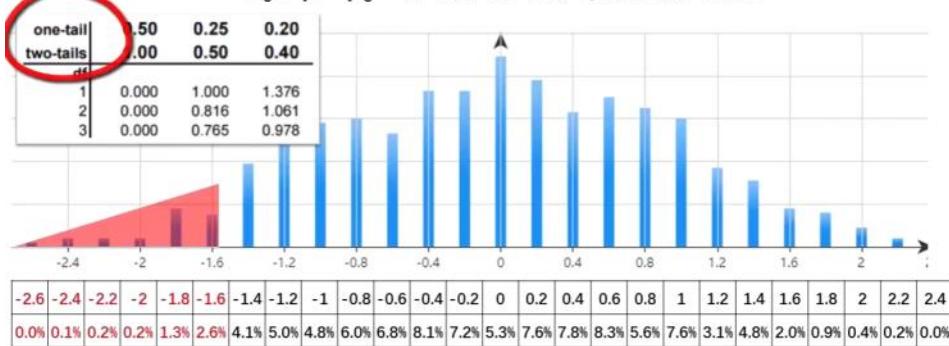


假如抽样一次，由样本计算得到 $t=2.3$ 或者 $t=-2.3$ ，t的绝对值都大于临界值2.093，于是t比左右正负两边的临界值都还极端，于是拒绝 H_0 。假如 $t=1.6$ 或者 $t=-1.6$ ，t的绝对值都小于2.093，于是t不比正负任何一个临界值极端，于是接受 H_0 。所以，双边检验的时候，不管抽样计算出的t值是正还是负，都取其绝对值，再和临界值进行比较。若比临界值大，则极端，则拒绝 H_0 ；若比临界值小，则不极端，则接受 H_0 。

$H_0: \mu > \mu_0$ n=20, df=19, 单边左边 $\alpha=0.05$



$$H_0: \mu > \mu_0 \quad n=20, \text{ df}=19, \text{ 单边左边} \alpha=0.05$$

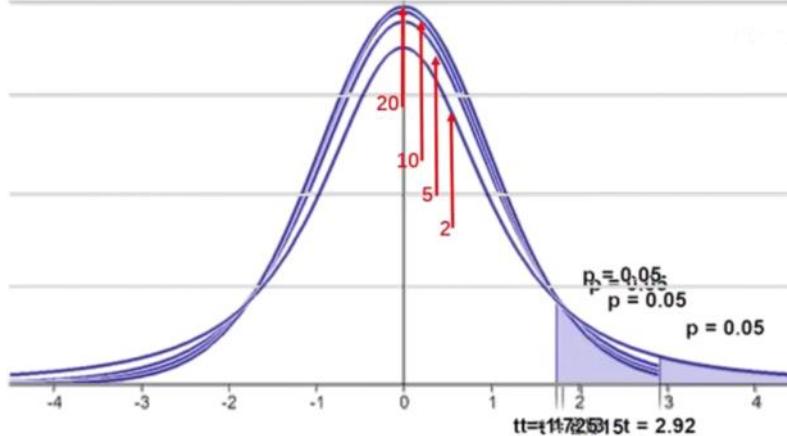


但是，在t临界值表中，好像并不区分“单边左边”或“单边右边”，只能区分“单边”或“双边”。原因我们刚才讲过了，因为t分布反正是左右正负对称的，只看绝对值就可以了。不过，我们心中要时刻记着，此时对应的原假设为 $H_0: \mu > \mu_0$ ，拒绝域在左尾，左尾的t值都是负数。表中查到的临界值，都要在心中加上一个负号。

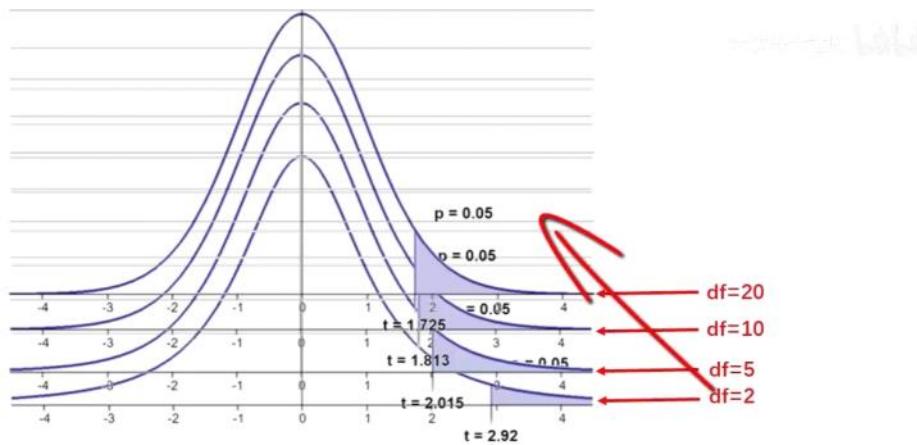
性质2. 同一显著水平下，自由度越大，临界值就越小。

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01
df									
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66
2	0.000	0.816	1.061	1.386	1.886	2.920	4.343	6.965	9.925
3	0.000	0.765	0.978	1.250	1.638	2.353	3.142	4.541	5.841
4	0.000	0.741	0.941	1.190	1.533	2.132	2.716	3.747	4.604
5	0.000	0.727	0.920	1.156	1.476	2.015	2.511	3.365	4.032
6	0.000	0.718	0.906	1.134	1.440	1.943	2.417	3.143	3.707
7	0.000	0.711	0.896	1.119	1.415	1.895	2.315	2.996	3.499
8	0.000	0.706	0.889	1.108	1.397	1.860	2.316	2.896	3.355
9	0.000	0.703	0.883	1.100	1.383	1.833	2.212	2.821	3.250
10	0.000	0.700	0.879	1.093	1.372	1.812	2.188	2.764	3.169
11	0.000	0.697	0.876	1.088	1.363	1.796	2.111	2.718	3.106
12	0.000	0.695	0.873	1.083	1.356	1.782	2.109	2.681	3.055
13	0.000	0.694	0.870	1.079	1.350	1.771	2.100	2.650	3.012
14	0.000	0.692	0.868	1.076	1.345	1.761	2.095	2.624	2.977
15	0.000	0.691	0.866	1.074	1.341	1.753	2.081	2.602	2.947
16	0.000	0.690	0.865	1.071	1.337	1.746	2.071	2.583	2.921
17	0.000	0.689	0.863	1.069	1.333	1.740	2.061	2.567	2.898
18	0.000	0.688	0.862	1.067	1.330	1.734	2.050	2.552	2.878
19	0.000	0.688	0.861	1.066	1.328	1.729	2.040	2.539	2.861
20	0.000	0.687	0.860	1.064	1.325	1.725	2.036	2.528	2.845
21	0.000	0.686	0.859	1.063	1.323	1.721	2.030	2.518	2.831

第二：同一显著水平下，自由度越大，临界值就越小。例如，我们选取单边0.05这一列，可以看出，从上至下，随着自由度的增加，t临界值在逐步减小。我们挑出其中 $df=2, 5, 10, 20$ 的t分布。



并做出其单边右边0.05的拒绝域，然后把图片重叠在一起进行比较。可以看到，自由度越大，曲线中间就越尖。曲线中间越尖的话，曲线下的面积就朝中间集中，但曲线下方的面积永远等于1，中间面积多了，两边尾巴面积就变少了，尾巴更细了。

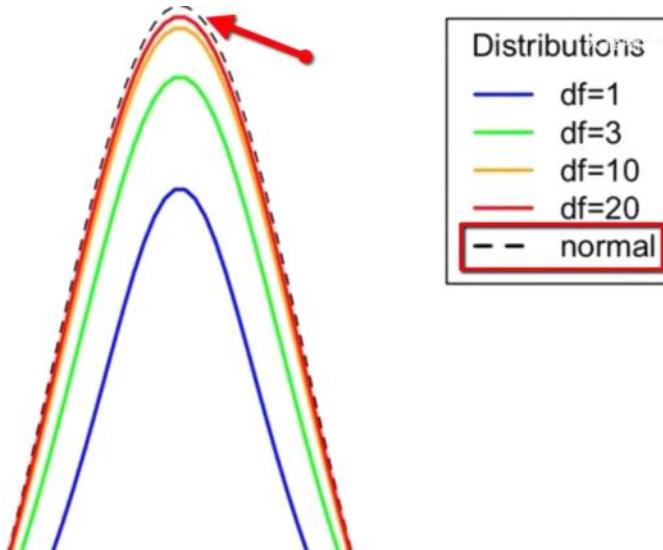


现在把重叠的图片分开，可以把尾巴尖看得更清楚一点。自由度从2增加到20，尾巴越来越细。因此，要想从尾巴尖上划出面积为0.05的拒绝域的话，分界线必然朝对称轴移动，也就是临界值变小了。所以说，同一显著水平下，自由度越大，临界值就越小。

性质3. 自由度增加，相邻t分布越来越相似，并趋近于标准正态分布。

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

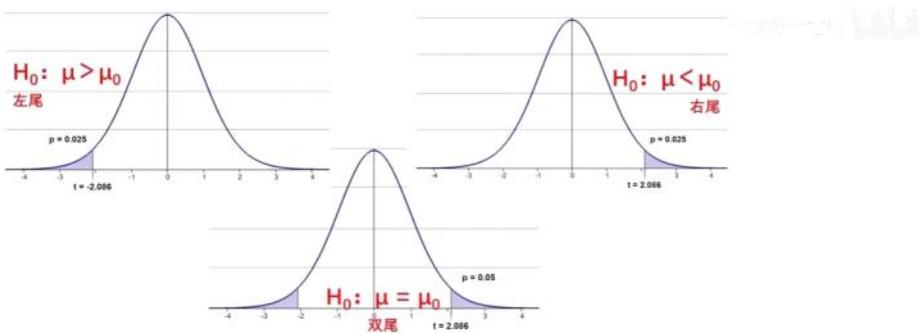
第三：随着自由度的增加，相邻自由度的t分布越来越相似，并最终趋近于标准正态分布。表现在数值上，就是当自由度很大时， $df=n$ 和 $df=n+1$ 的临界值很接近。例如，单边 $\alpha=0.05$ 时， $df=4$ 和 $df=5$ 的临界值相差还比较大，有0.1之多，但 $df=29$ 和 $df=30$ 的临界值就只差0.002了。



性质3. 自由度增加, 相邻t分布越来越相似, 并趋近于标准正态分布。

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

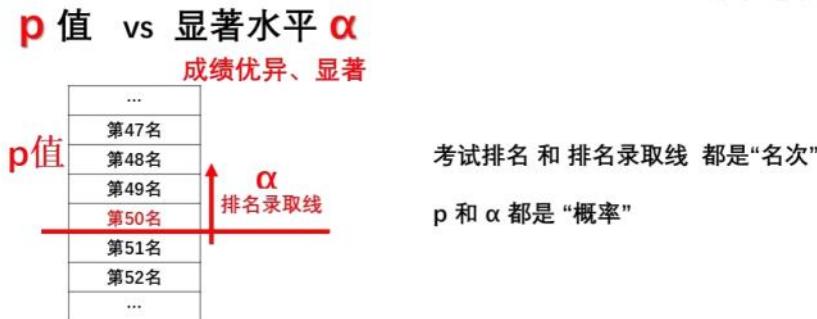
df=60和df=80的自由度相差20, 单边 $\alpha=0.05$ 的临界值仅仅相差0.007。所以, t临界值表的最后, 没有把不同的自由度以“+1”递增的方式都列出来, 因为自由度+1的t分布实在是差别不大, 没必要都列出来。而是只列出了60, 80, 100, 1000这几个自由度, 最后列出了一个z, z分布就是标准正态分布。



注意, 这时强烈建议大家按照单双边, 自己画出一个t分布, 给临界值标上正负号, 然后按照临界值画出拒绝域。最后, 根据样本和公式算出t值。看t值是否落入拒绝域, 进而判断是否拒绝 H_0 。

p值

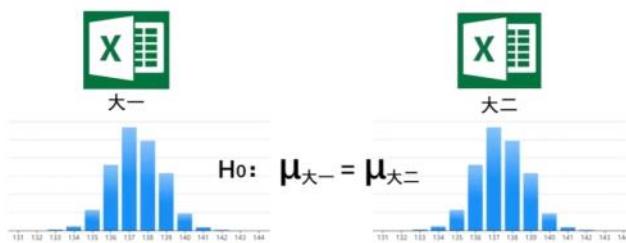
2023年12月31日 21:19



显著水平 α ，只不过也是一个特殊的p值。p和 α 的关系，相当于考试名次和排名录取线的关系。例如，一次考试1000人参加，每个人都会有个成绩名次。这个名次，就相当于p值。而考试只选拔录取前50名。那么这个“前50名”，就是排名录取线，相当于 α 。假如你的排名，比录取排名还要高的话，就说明你的成绩就越“优异”，越“显著”。所以，考试排名和排名录取线都是名次，p和 α 都是概率。

p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

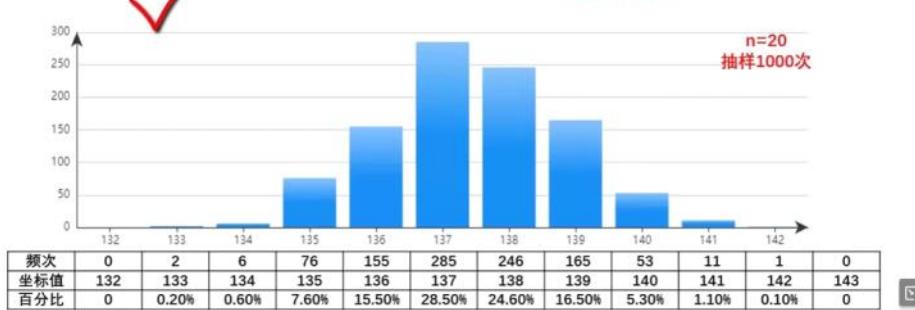
H_0 : 大一新生和大二老生的高考英语成绩平均分没有显著差别。



下面，我们来“咬文嚼”一下，把这个较为学术的p值的定义，和我们之前课程中学过的内容来对应一下。我们回到之前关于大一和大二高考英语成绩平均分的例子。在这个例子中，我们假设已经知道了大二的成绩总体excel表，并可以计算得到大二的平均分 $\mu_{\text{大二}}$ 。我们的原假设 H_0 为， $\mu_{\text{大一}} = \mu_{\text{大二}}$ 。原假设为真的时候，我们粗略的认为，大一和大二是同一个总体，所以大一和大二的均值抽样分布是完全一样的分布。

p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

H_0 : 大一新生和大二老生的高考英语成绩平均分没有显著差别。



如果原假设为真，那么大一和大二的均值抽样分布完全一样。这时需要特别注意，原假设是双边检验。此时，我们抽样1次，算出均分为141分，这个抽样只对应了双边中右边的一条线，还应当找出左边的那条线。根据对称轴是137分，抽样所得的141分是从对称轴往右边尾巴尖第4个柱状图的位置。那么，左边这条线，也应当是从对称轴往左数，第4个柱状图的位置，是133分。

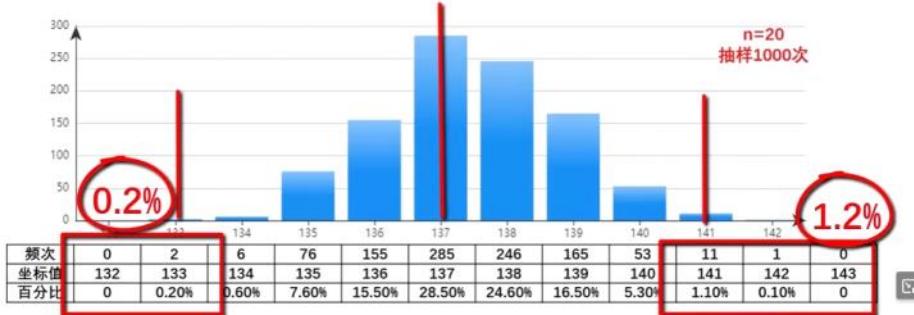
p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

H_0 : 大一新生和大二老生的高考英语成绩平均分没有显著差别。



p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

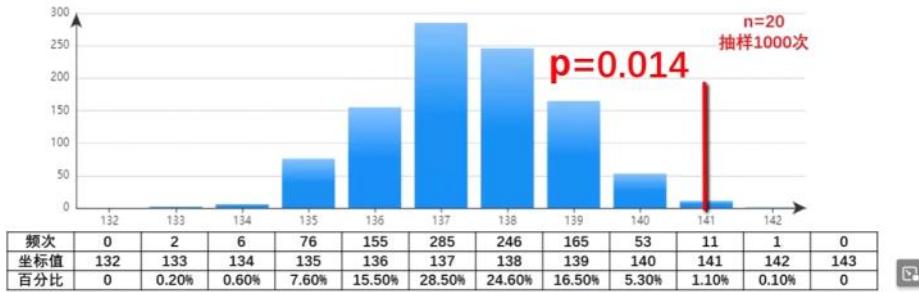
H_0 ：大一新生和大二老生的高考英语成绩平均分没有显著差别。



所以，虽然抽样得到的样本是141分，因为是双边检验，你心中还要有个对称轴另一边的“假想样本”133分。所以，比样本（含）更极端的结果，包括两个部分，右边是141分，142分，143分，等等。但到143分，抽样次数就为0了，所以比143分再高的均分的概率都是0。那么右边的累计概率是 $1.10\% + 0.10\% + 0 = 1.2\%$ 。同样的，左边的累计概率是 $0.20\% + 0 = 0.2\%$ 。那么本次抽样一次，得均分141分的双边p值为，两边加起来， $1.2\% + 0.2\% = 0.014$ 。

p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

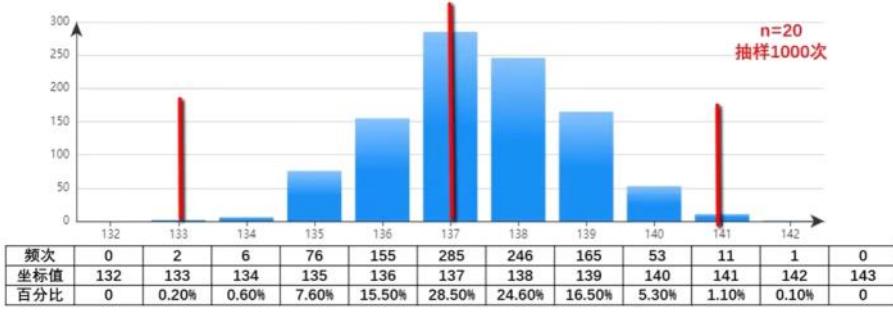
H_0 ：大一新生和大二老生的高考英语成绩平均分没有显著差别。



得出了抽样1次的p值，下一步就是和显著水平 α 比较一下。如果一开始设置的 $\alpha=0.05$ ，则 $p < \alpha$ ，拒绝 H_0 。若一开始设置的 $\alpha=0.01$ ，则 $p > \alpha$ ，接受 H_0 。这里可以看出，用p值来进行假设检验的好处就是，不用再去比较什么临界值，也不用再划出什么拒绝域了。拿p和 α 进行比较，两个纯小数，谁大谁小，一目了然。

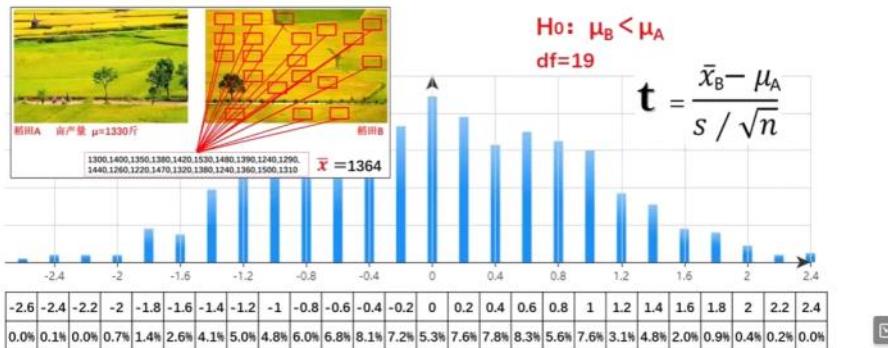
p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。

H_0 ：大一新生和大二老生的高考英语成绩平均分没有显著差别。



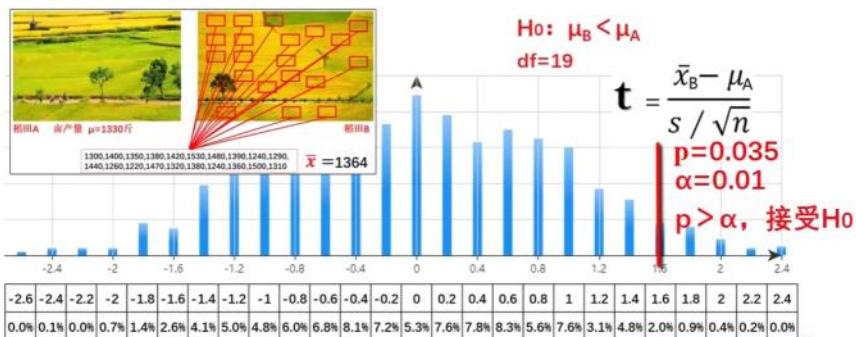
当然，上面需要说明两点。第一，这是一个粗略的抽样1000次形成的均值分布，理论上，141分和133分是以137分为对称轴的，两端的累计频次应该是相同的，但本例中右边是1.2%，左边是0.2%，两边不相等。

p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。



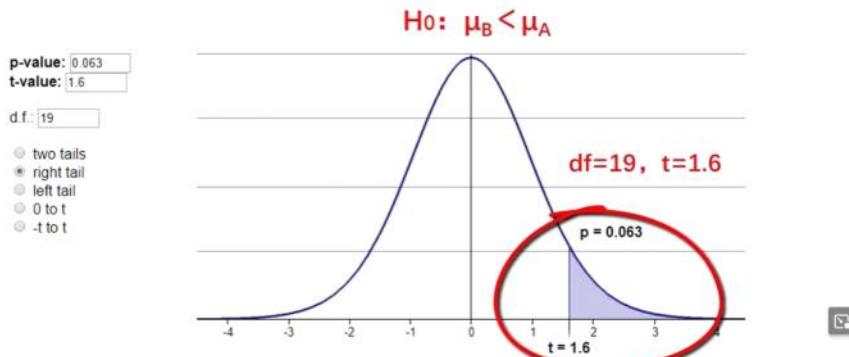
下面，我们再把p值的概念，放到真实的t检验中来对应一下。例如，这是之前我们讲过的，比较稻田A和稻田B亩产量的单边t检验。原假设 $H_0: \mu_B < \mu_A$ ，则拒绝域应该在单边右尾，也就是说右尾是极端方向。

p值就是，当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率。



假设根据样本算出来的t值是1.6。那么，在这个粗略的t值分布表中，从t=1.6到更极端方向的累计概率是：0.02+0.009+0.004+0.002+0=0.035。于是，此次单边右尾检验的p值为0.035。那么，假如事先规定的 $\alpha=0.05$ ，则 $p < \alpha$ ，于是拒绝 H_0 。假如 $\alpha=0.01$ ，则 $p > \alpha$ ，于是接受 H_0 。当然，我们这个演示用的t值表实在不够精确，可能会导致错误的结论。

<http://www.statdistributions.com/t/>



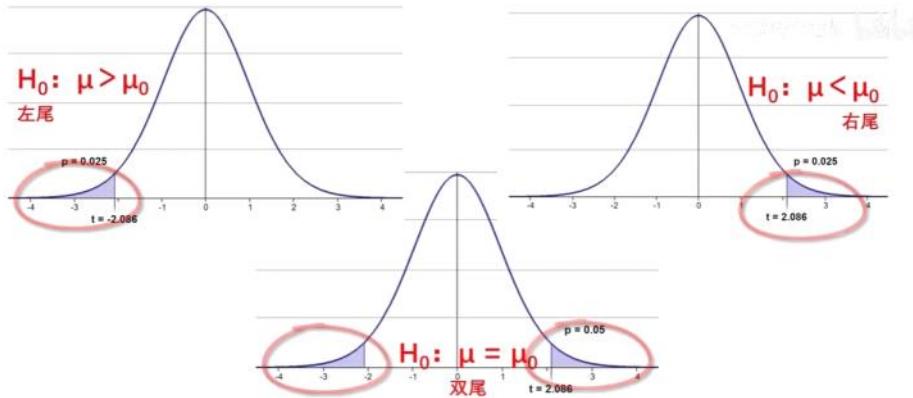
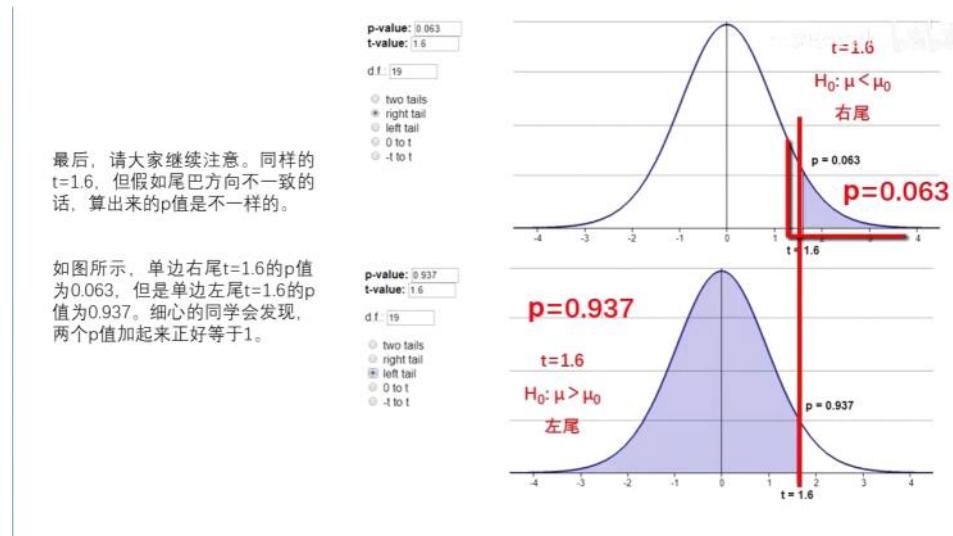
我们用上节课告诉大家的这个网站，来算一下精确的p值。首先，由原假设 H_0 中的“ $<$ ”号，确定是单边右边检验，于是这里选“right tail”。自由度这里输入“19”。然后，t-value这里输入“1.6”。这时，程序算出了 $p=0.063$ ，并在t分布曲线中做出了阴影部分的面积。现在，我们知道了t=1.6对应的精确的p值为0.063。无论 $\alpha=0.05$ ，还是 $\alpha=0.01$ ， p 都大于 α 。于是我们接受 H_0 。这说明，我们从那个粗略的t值表划出来的 $p=0.035$ ，还真是不靠谱。

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

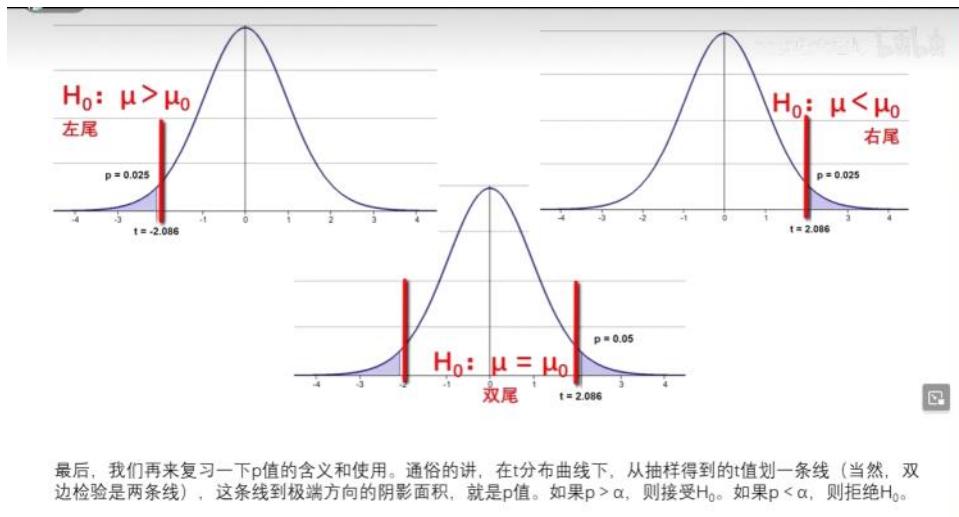

$$p\text{-value} = \frac{\Gamma\left(\frac{(n+1)}{2}\right)}{\sqrt{n\cdot\pi}\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^t \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}} dx$$

手工算p?

第二，可能有同学会问，因为t值公式比较简单，我们用高级点的计算器就可以把t值算出来。那么p值可以直接算出来吗？答案是可以。但不建议用手动计算。这个是p值的计算公式，手动算应该是不大容易算出来吧。



所以，在使用统计软件计算p值时，都要明确设定，你的假设检验，到底是“双尾”还是“单尾”，“单尾”的话，具体是“左尾”、还是“右尾”。不同的尾巴方向，代表着你的原假设中的“大于号”，“小于号”，还是“等于号”。这关乎p值的计算结果和结论的正确与否。



最后，我们再来复习一下p值的含义和使用。通俗的讲，在t分布曲线下，从抽样得到的t值划一条线（当然，双边检验是两条线），这条线到极端方向的阴影面积，就是p值。如果 $p > \alpha$ ，则接受 H_0 。如果 $p < \alpha$ ，则拒绝 H_0 。



我们之前讲过的假设检验的基本套路是，若想得到一个结论 (H_1)，首先写出这个结论的对立结论，并将其作为原假设 H_0 。然后通过抽样数据，来拒绝 H_0 ，从而证得 H_1 。如此看来，p值越小，越容易拒绝 H_0 。所以，一般来说，p值越小越好，最好是越接近于0才更好。

第1类错误第2类错误

2024年1月2日 10:07

通俗统计学原理11 - 第1类错误 Type I error (False Positive "假阳性") vs 第2类错误 Type II error

第1类错误 第2类错误

Type I error Type II error

大家好。在解释抽样样本p值和显著水平 α 的时候，我们经常会遇到两个概念：第1类错误和第2类错误。英语分别叫做Type I error，和Type II error。鉴于我们在之前课程中已经掌握了足够的术语或概念，下面，我直接给出这两类错误的定义。

第1类错误：原假设 H_0 实际为真时，拒绝了 H_0 。

Type I error: Rejecting H_0 when it's actually TRUE.

第2类错误：原假设 H_0 实际为假时，接受了 H_0 。

Type II error: Accepting H_0 when it's actually FALSE.

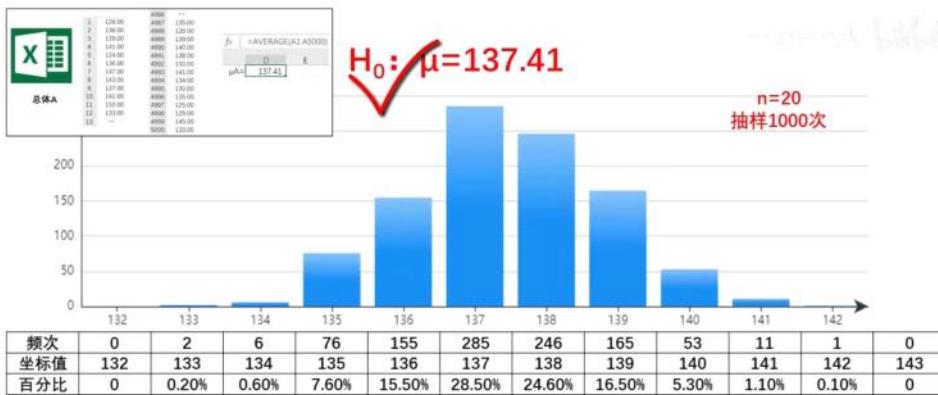
第1类错误是指，原假设 H_0 实际为真时，但拒绝了 H_0 。第2类错误是指，原假设 H_0 实际为假时，但接受了 H_0 。

为了说明这两个概念，我们仍然回到高考英语成绩均值抽样的实验。例如，我们拥有了总体A的英语成绩excel表，并可以看到总体中5000个考生的每个人的分数，并可以计算出总体的真实均分 $\mu_A=137.41$ 分。

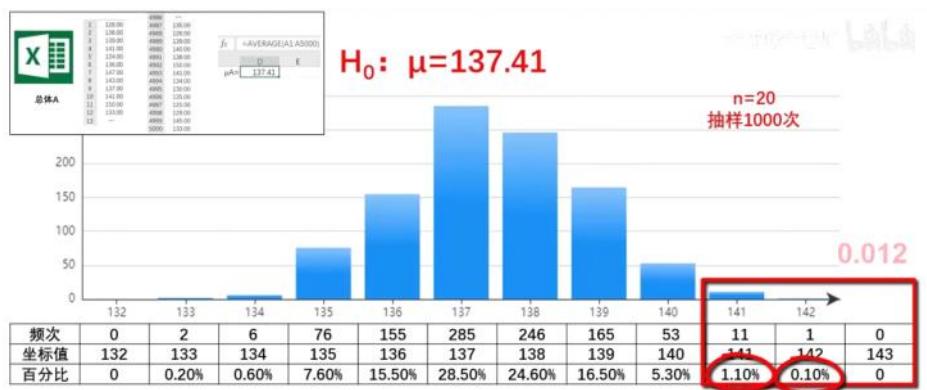
行号	成绩
1	126.00
2	136.00
3	139.00
4	141.00
5	124.00
6	136.00
7	147.00
8	143.00
9	137.00
10	141.00
11	150.00
12	133.00
13	...
5000	133.00

$\mu_A = 137.41$

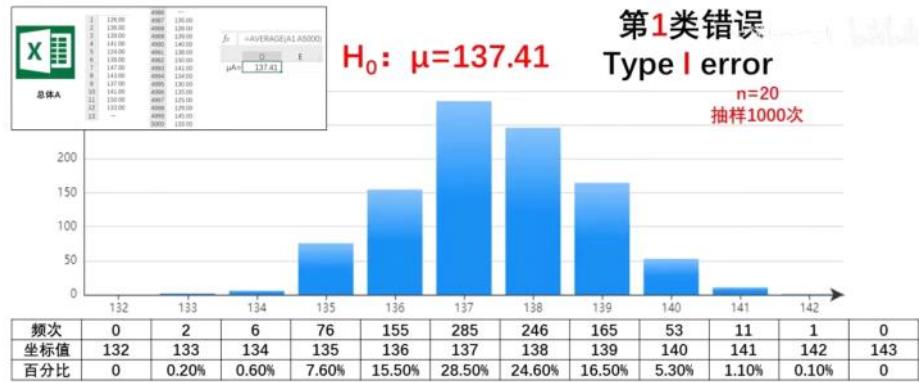




接下来，我们睁着眼，对这个 $\mu=137.41$ 分的总体进行样本容量n=20的1000次均值抽样，获得了这个抽样分布。这说明，即使 H_0 原假设千真万确是真的，总体的均分 μ 就是137.41分，我们竟然还能抽到均值为133分、134分、141分、142分这样极端的样本。这些样本均值偏离真实均值太多了。

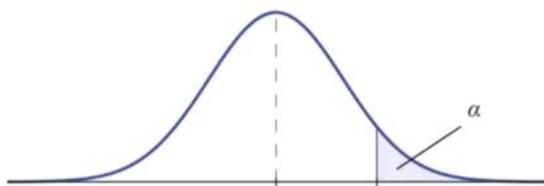


假设我们划定比141分（含）还极端的抽样为拒绝域，则拒绝域的面积为 $0.011+0.001=0.012$ 。也就是说，我们设定了右边 $\alpha=0.012$ 为显著水平。那么，假如抽到了均值为141分的样本，则落入拒绝域，于是就拒绝了 H_0 。这就是所谓，我们眼睁睁的从 H_0 为真的总体中抽样，最后反而拒绝了 H_0 ，认为 H_0 为假。



这就叫，犯了第1类错误。这时，你心里就慌了。原来假设检验这种操作，是有风险的，明明 H_0 是真的，都有可能获得极端抽样，或者说，有可能抽样得到一个很小的p值，然后就错误的拒绝了 H_0 。于是你自然而然的问到，我们有多大的概率犯第1类错误呢？

第1类错误：原假设 H_0 实际为真时，拒绝了 H_0 。



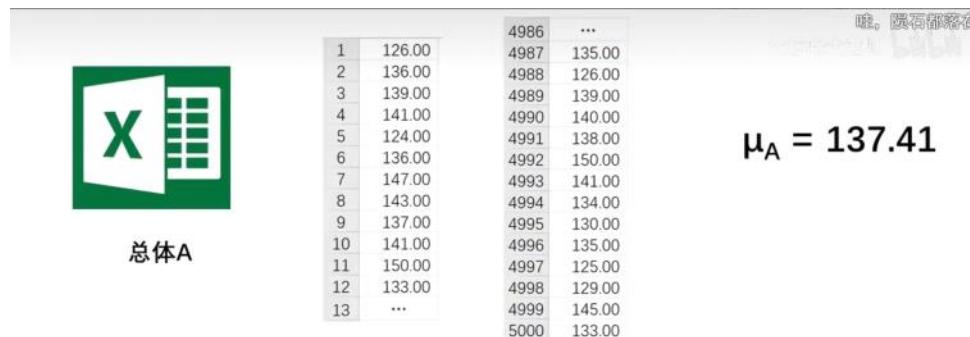
“犯第1类错误的概率 = 显著水平 α ”

一般存在这样的一种说法：显著水平 α ，就是犯第1类错误的概率。本人觉得这种说法存在讨论空间。我们已经学过，显著水平是人为划定的、 H_0 为真的情况下，抽样抽到的极端情况的概率。本人对显著水平本身定义还是持赞成态度的，可是直接把显著水平和犯第1类错误的概率等同起来，觉得不大合适。

I类错误——弃真错误，发生的概率为 α ，否定了真实的原假设。避免方法：可通过 α 水平控制，降低 α 水平

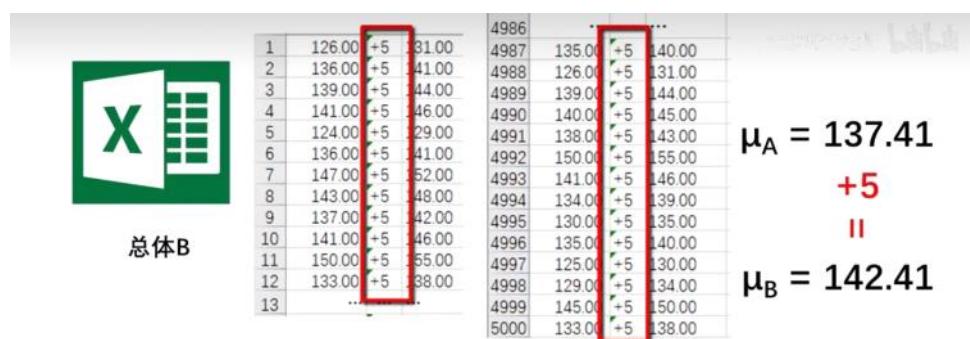
https://blog.csdn.net/garbageSystem/article/details/122603832?ops_request_misc=&request_id=&biz_id=102&utm_term=AB%E6%B5%8B%E8%AF%95&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduweb~default-3-122603832.nonecase&spm=1018.2226.3001.4187

第2类错误



$$\mu_A = 137.41$$

还是刚才的总体A的excel成绩表，我们把它改造一下。用流行的说法就是，我们假想一个平行时空，在这个平行时空里，也有这么5000个学生，姓名都一样。



$$\mu_A = 137.41$$

+5

II

$$\mu_B = 142.41$$

但这个平行时空里学生的英语水平稍微高一点，每个考生的英语成绩都多考了5分。我们把平行时空里的这个总体记作总体B，则总体B的真实均值肯定是 $\mu_B = 137.41 + 5 = 142.41$ 分。



总体B

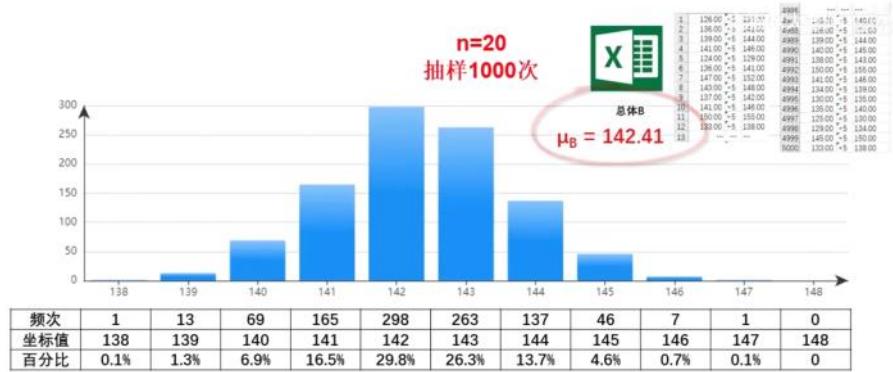
1	126.00	± 5	131.00	4986
2	136.00	± 5	141.00	4987	135.00	± 5	140.00
3	139.00	± 5	144.00	4988	126.00	± 5	131.00
4	141.00	± 5	146.00	4989	139.00	± 5	144.00
5	124.00	± 5	129.00	4990	140.00	± 5	145.00
6	136.00	± 5	141.00	4991	138.00	± 5	143.00
7	147.00	± 5	152.00	4992	150.00	± 5	155.00
8	143.00	± 5	148.00	4993	141.00	± 5	146.00
9	137.00	± 5	142.00	4994	134.00	± 5	139.00
10	141.00	± 5	146.00	4995	130.00	± 5	135.00
11	150.00	± 5	155.00	4996	135.00	± 5	140.00
12	133.00	± 5	138.00	4997	125.00	± 5	130.00
13	4998	129.00	± 5	134.00
				4999	145.00	± 5	150.00
				5000	133.00	± 5	138.00

$\mu_A = 137.41$
 $+5$
 $\mu_B = 142.41$
 11

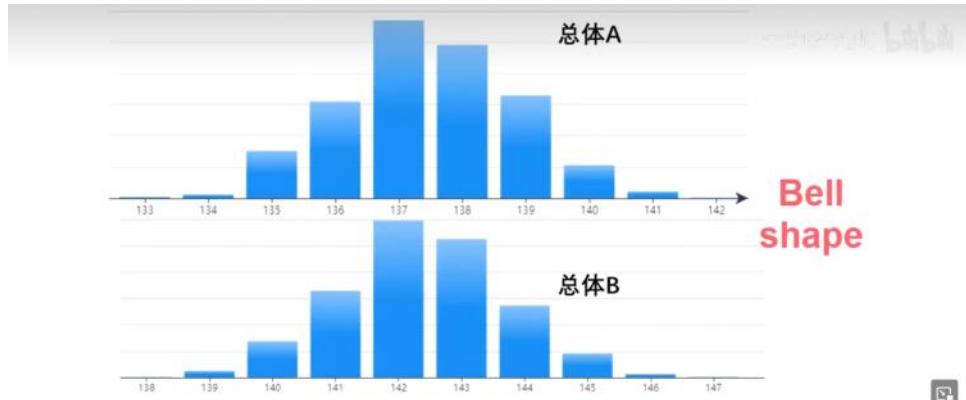
假设：总体B的均值仍然是137.41分

$$H_0: \mu = 137.41$$

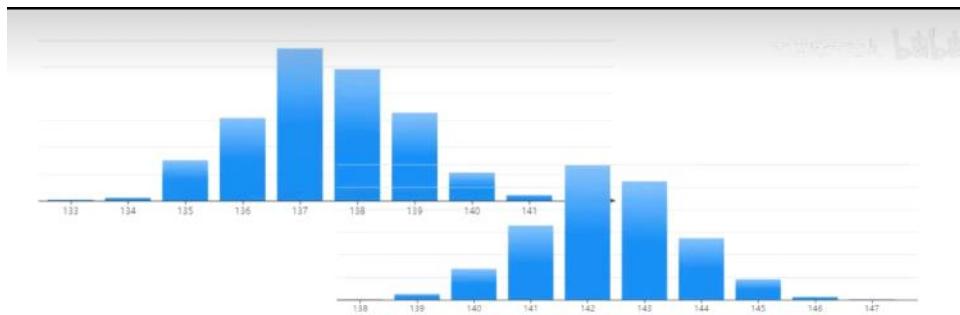
现在我们知道了总体B的真实均值为 $\mu=142.41$ 分，并且也有总体B的excel表。但是，这时我们做个假设，我们假设总体B的均值仍然是 137.41 分，写出来是， $H_0: \mu=137.41$ 分。你可能会说，这不是睁眼说瞎话吗？是的，我们就是要睁眼说瞎话。所以，这个原假设 H_0 ，肯定为假。因为我们眼睁睁的制造出来的总体B，总体B的真实均值就是 142.41 分。



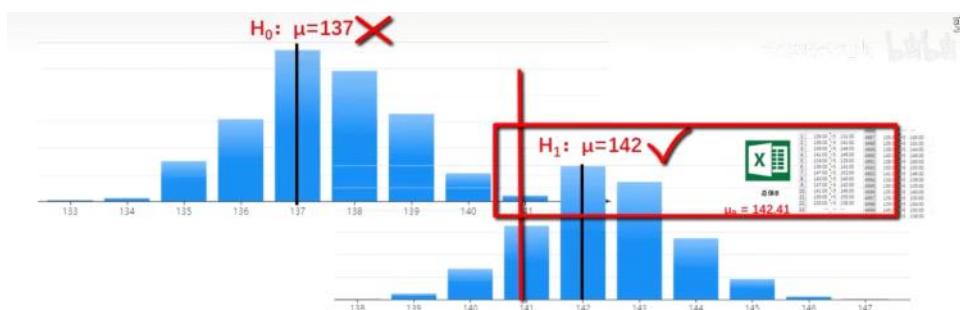
这时，我们对总体B进行样本容量n=20的均值抽样，获得总体B的均值抽样分布。



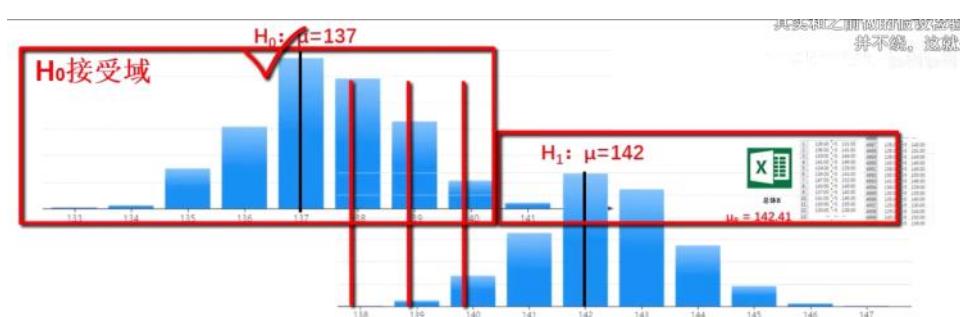
我们把总体A和总体B的均值抽样分布放在一起，轮廓看起来是相同的。这是当然的。首先，均值抽样分布肯定是正态分布，所以，两个轮廓都是铃铛形状，bell shape。



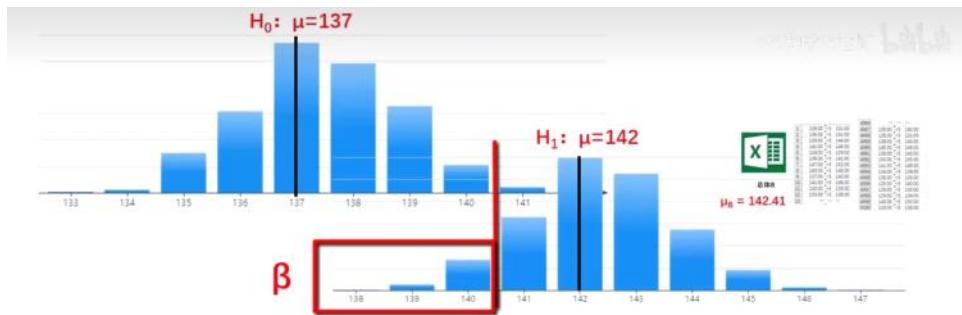
两个分布的坐标对齐了，例如，140分这里对齐了，141分这里对齐了，所有的坐标分数都对齐了。注意，大家跟上我的思路，我们开始弯弯绕了。均分137分的分布，是我们此次的原假设，我们在这里标上 H_0 , $\mu=137$ 。



下面我说的话，可能更拗口了，大家跟上我。我说，这个抽样分布，是对总体B的抽样，也就是对 H_1 总体的抽样。 H_1 的均分我们是知道的，是 $\mu=142$ 分，所以 H_1 这个备用假设为真。原假设 H_0 这个分布，是我们按照加5分之前的总体假想的，所以， H_0 为假。但是，本次假设检验，是从 H_1 的总体中抽的样，然后对着 H_0 为真的分布进行检验。于是，和之前讲第1类错误时一样，我们把 H_0 分布中大于141分（含）的抽样划为 H_0 的拒绝域。

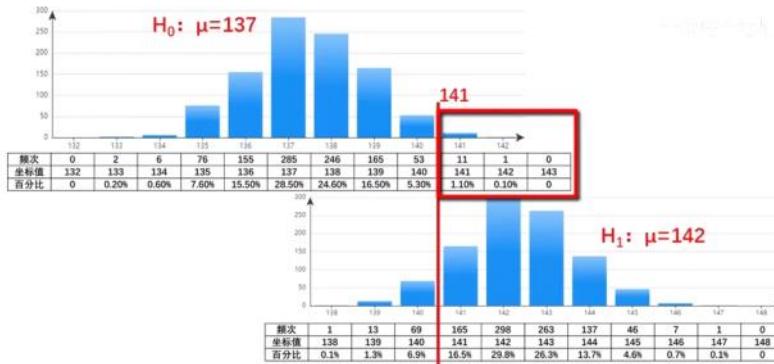


这时，你若从 H_1 总体中抽样1次，得均分141分，或者更大的142分，143分等等，落入了 H_0 的拒绝域。你若因此拒绝了 H_0 ，而接受了 H_1 的话，恭喜你，你做出了一个实际上正确的选择。但是，你若从 H_1 总体中抽样得到的均分是140分，139分，138分等等，比141分小的分数，这些都落入了 H_0 的接受域。按照假设检验的套路，你该接受 H_0 。但 H_0 是假的啊，我们明明知道的啊。此时，你就犯错了吧。这时，你犯的错误，就叫做第2类错误。

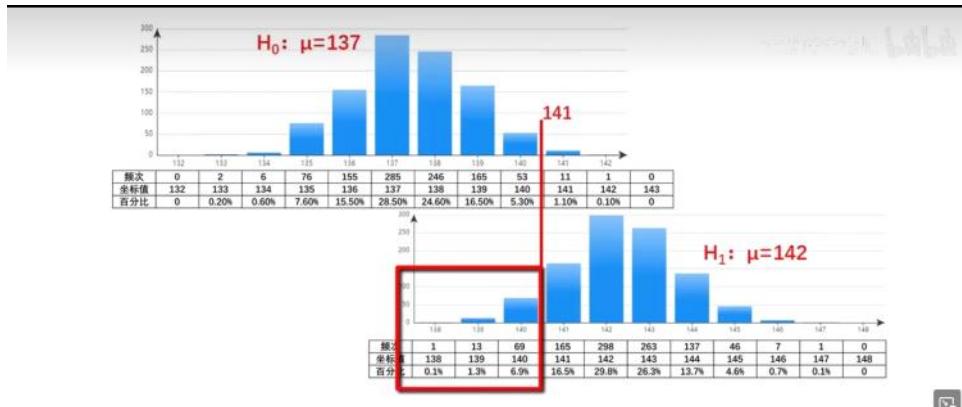


第2类错误: 原假设 H_0 实际为假时, 接受了 H_0 。

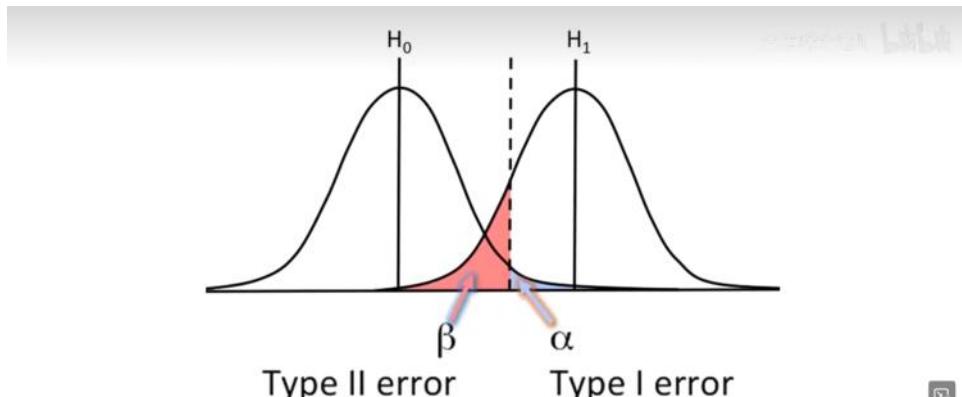
第2类错误就是, H_0 实际为假时, 你却接受了 H_0 。此时, 你又自然而然的问, 有多大概率犯第2类错误呢? 这个概率, 得从 H_1 的分布中算。从图上可以看出, H_0 的拒绝域的这个边界线, 在 H_1 的分布中, 往左划出来的这块面积百分比, 就是犯第2类错误的概率。这个概率, 用字母 β 表示。



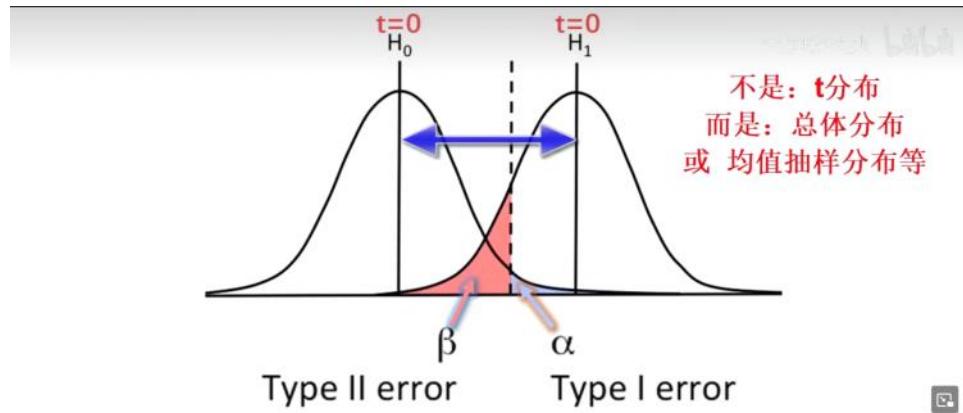
现在, 我们把 H_0 和 H_1 的抽样分布概率表放在一起。我们人为划一条显著水平线或者叫临界值线, 141分。在 H_0 的分布中, 比141分(含)还高的分数, 占 H_0 分布的总百分比为, $\alpha=0.011+0.001=0.012$ 。根据刚才讲过的说法, $\alpha=0.012$ 就是犯第1类错误的概率, 也就是 H_0 为真, 但我们拒绝了 H_0 的概率。



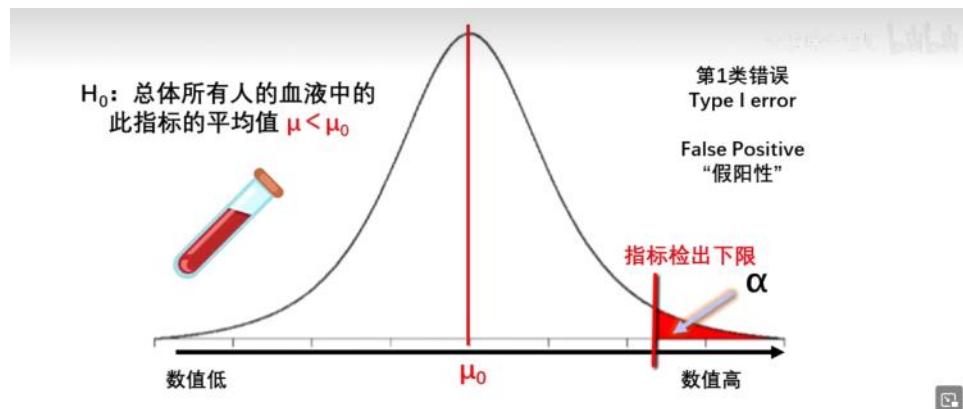
同时, 这条线把 H_1 分布也划分成了两部分。 H_1 分布中比141分(不含)还低的分数的总百分比为:
 $\beta=0.069+0.013+0.001=0.083$ 。 $\beta=0.083$ 就是犯第2类错误的概率, 也就是 H_0 为假, 但我们接受了 H_0 的概率。



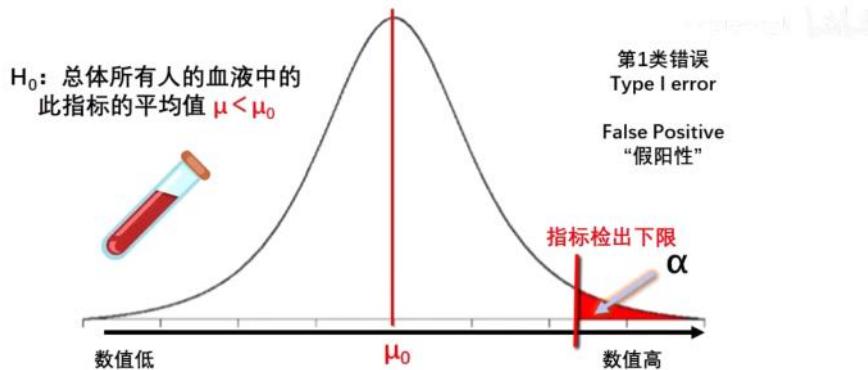
现在, 我们再回过头来, 看本节课封面的这张图片, 应该就不难理解了。 α 就是当 H_0 实际上为真时, 但拒绝了 H_0 , 从而犯第1类错误的概率。 β 就是, 当 H_0 实际上为假, H_1 为真时, 但接受了 H_0 , 从而犯了第2类错误的概率。



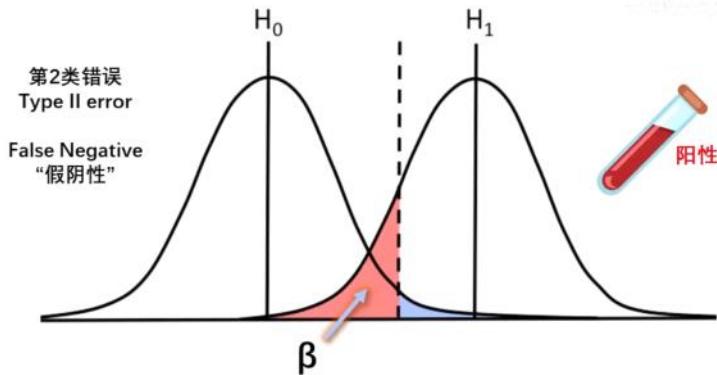
注意，这两条铃铛形状的曲线，也不是t分布曲线，而是 H_0 和 H_1 的总体分布，或者均值抽样分布等等，大家千万不要搞错。因为假如是t分布的话，所有t分布的对称轴肯定都是 $t=0$ ，两条曲线对称轴应该重合才对。而此图中 H_0 和 H_1 的对称轴，分别代表各自不同的总体均值 μ 。



这个原假设就是 H_0 : 总体所有人的血液中的此指标的平均值 $\mu < \mu_0$ 。如图所示，这是一个指标均值 $\mu = \mu_0$ 的正态分布。数轴上，从左往右，表示指标的数值由低到高。我们从指标最高的尾巴尖上，划出一个临界值。这个临界值，叫做“指标检出下限”。‘检出下限’右边的阴影部分就是拒绝域，面积为 α 。



假如抽样1次，指标小于检出下限，则认为符合 H_0 ，也就是所谓的“阴性”或“negative”。假如抽样1次，指标大于检出下限，则落入拒绝域，拒绝 H_0 ，则此抽样就为“阳性”或“positive”。此时，若一个样本实际上是阴性，但由于检测条件、检测仪器等随机因素，导致样本指标落入拒绝域，被当做阳性，那么，便是“假阳性”，“False Positive”，也就是犯了第1类错误。

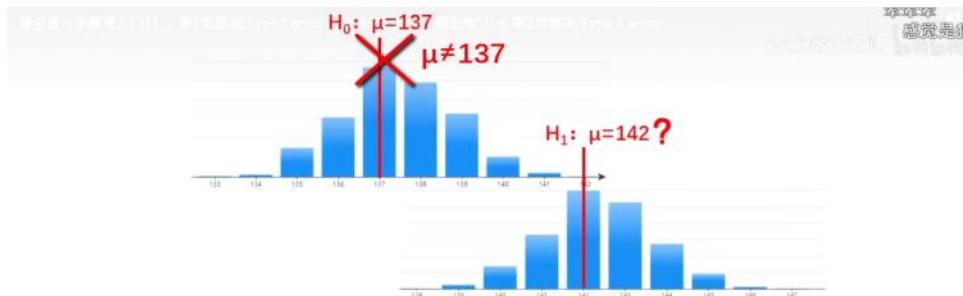


当然，与之对应的，就是本来是阳性，但是没有检出来，被当做了阴性。这就是犯了第2类错误，也就是图中红色面积为 β 的阴影部分。所以，第2类错误也叫“False Negative”，直译过来就是“假阴性”。

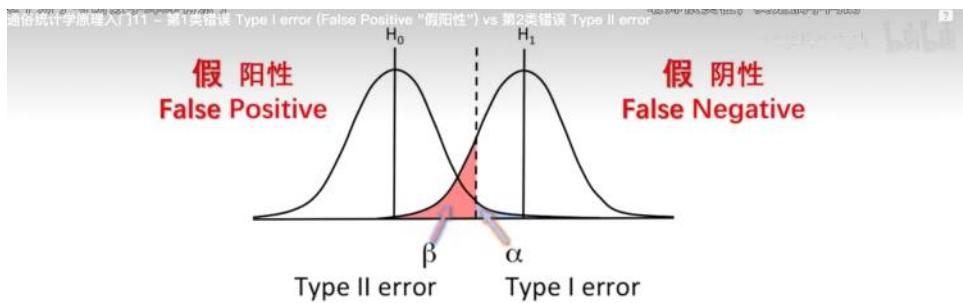
Type I and Type II Error		
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \alpha$
Not rejected	Correct decision True negative Probability = $1 - \beta$	Type II error False negative Probability = β

图表来源: <https://www.scribbr.com/statistics/type-i-and-type-ii-errors/> Scribbr

最后，我们来总结一下。原假设 H_0 为真时，若拒绝 H_0 ，就是犯了第1类错误，错误的拒绝 H_0 的概率为显著水平 α 。原假设 H_0 为真时，若接受 H_0 ，则做出了正确的决定，正确的接受 H_0 的概率为 $1 - \alpha$ 。原假设 H_0 为假时，若拒绝 H_0 ，则做出了正确的决定，正确的拒绝 H_0 的概率为 $1 - \beta$ 。原假设 H_0 为假时，若接受 H_0 ，则犯了第2类错误，错误的接受 H_0 的概率为 β 。



因为我们心中有很多疑问。例如：为什么当 H_0 为假时，也就是当 μ 其实不等于137时， μ 就非要等于142呢？ μ 为什么不能等于143？144？甚至， μ 可以等于左边的135？134？再如： H_1 是不是应当包括 H_0 的所有对立面，而不应该取一个确定的值？这些疑问，我都回答不到了。



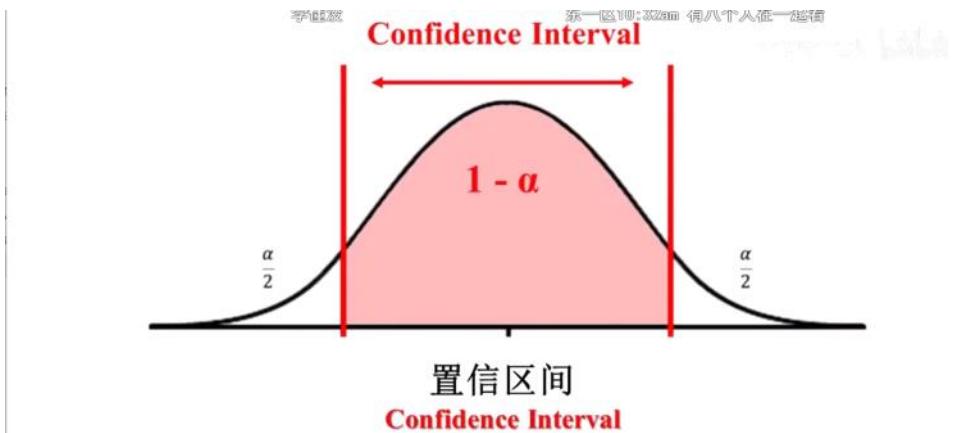
第1类错误: 原假设 H_0 实际为真时, 拒绝了 H_0 。

第2类错误: 原假设 H_0 实际为假时, 接受了 H_0 。

本节课之所以讲解“第1类错误”和“第2类错误”两个概念, 是因为大家在网上经常会碰到它们。本人水平有限, 只能鹦鹉学舌, 讲不出个所以然来, 可能还会讲错。只是希望大家不要把这两个概念, 和我们之前讲过的p值和显著水平 α 过多的掺和在一起。好, 这节课就到这里, 我们下节课见。

置信区间confidence interval

2024年1月2日 10:50



本节课，我们来学一下置信区间的概念。置信区间，英语叫做Confidence Interval。请大家把汉语和英语都好好记一下，因为统计学中，还有其他区间，大家不要搞混。



我们仍然回到之前的高考英语成绩的例子。这个excel表里，储存着5000名学生的英语成绩。但现在，我们没有权限打开这个表。Access restricted。所以无法直接计算出5000个成绩的真实均值 μ 。但我们有权限从表中进行抽样。



假如现在，你从表中随机抽出了20个学生的成绩。这20个成绩的平均分算出来是 \bar{x} 等于137.65分。这能说明什么呢？我们想要知道的是，这5000名学生总体的真实均分 μ 。但限于条件，我们只能抽样1次，得到样本均分为137.65分。这个样本均分就叫做对总体均分 μ 的一次估计 (estimate)。因为单次抽样均分在坐标轴上表现为一个点。因此，这个样本均值也叫做对总体均分 μ 的点估计。



样本容量 $n=20$

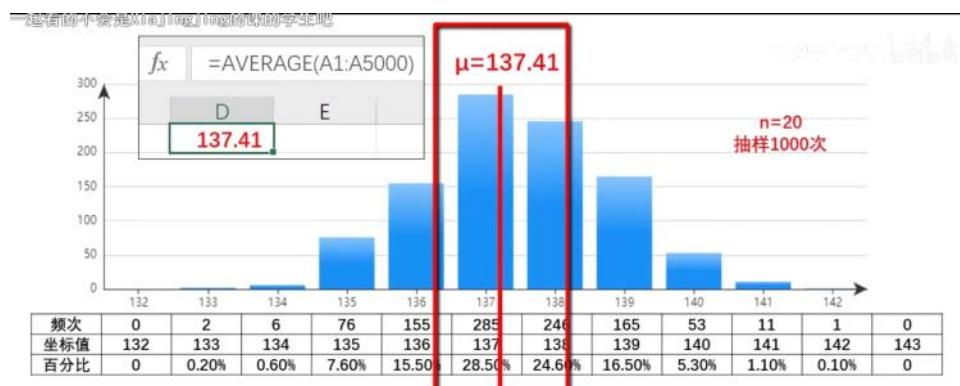
样本均值 $\bar{x}=137.65$ 分

对总体均分 μ 的 点估计
(point estimate)

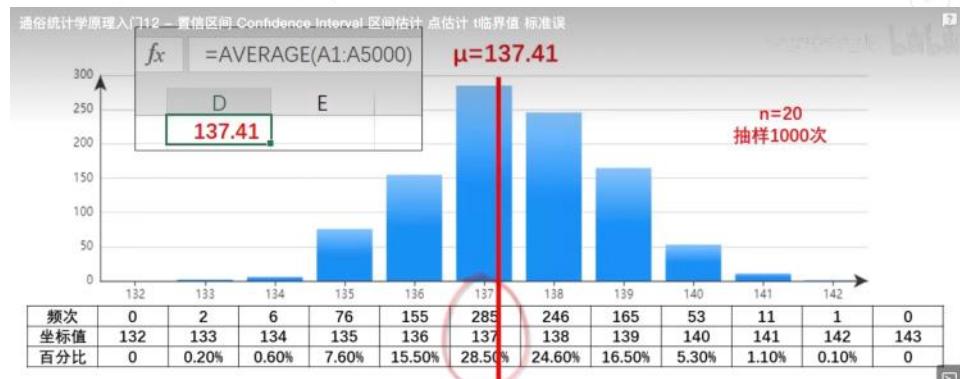
$\mu=137.65$ 左右

[137.65 - ?, 137.65 + ?]

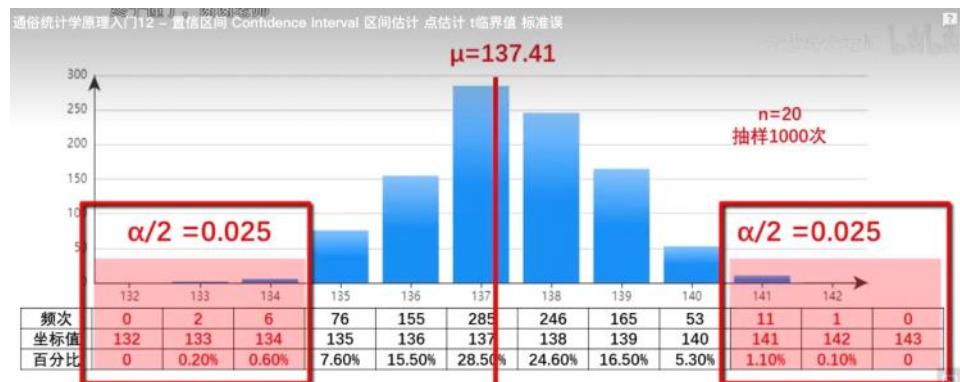
而这个“左右”，实际上就是引入了一个区间的概念。区间，就表示 μ 在 137.65 分上下浮动。那上下浮动的范围是多少呢？1分？3分？还是10分呢？这个浮动范围，我们总不能瞎蒙的。那怎么办呢？



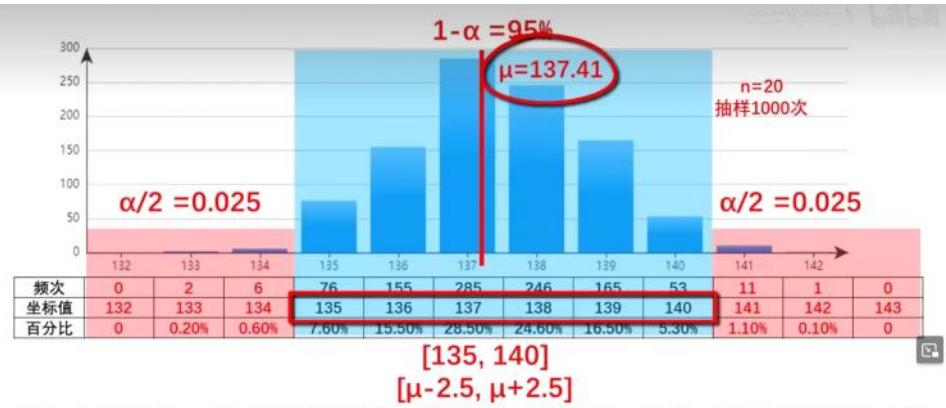
这时，我们回到均值抽样分布实验中。假设，我们现在有权限打开excel表了，并且算出总体的真实均值 $\mu=137.41$ 分。然后，对这个excel表进行了1000次，样本容量 $n=20$ 的抽样，获得如图所示的均值抽样分布。我们发现，均值抽样分布的对称轴，正是总体的真实均值 μ 。



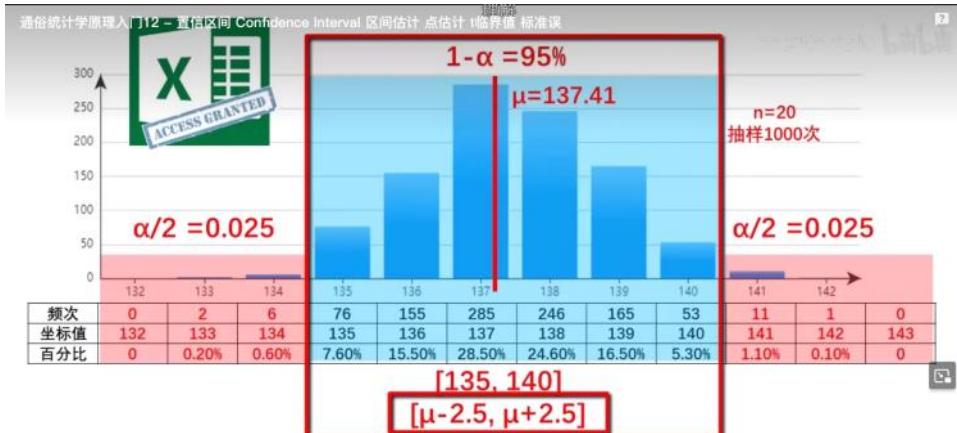
在1000次抽样中，除了能抽到样本均值正好等于总体均值 μ 以外，还可能抽到从133分到141分等等，偏离 μ 的样本均值。换句话说，明明知道了真实的 μ ，也不是每一次抽样均值都是 μ 。所有的抽样均值，其实是分布在 μ 左右的一个区间里。



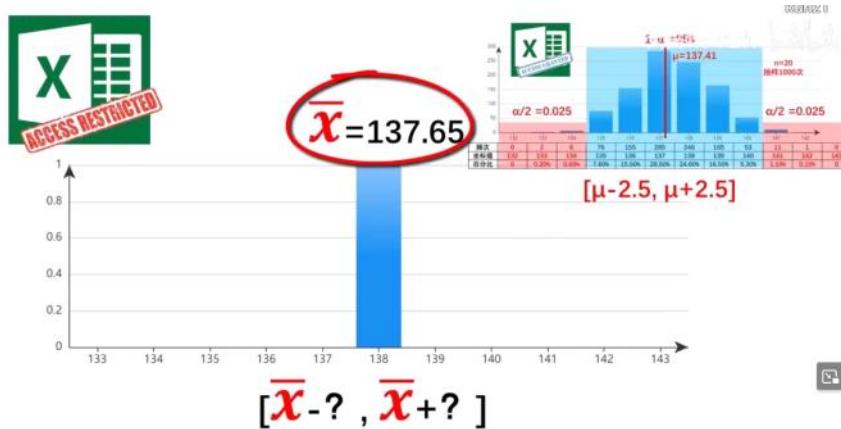
那么这个区间，在 μ 左右浮动的范围多少呢？这时，我们要再用到拒绝域和显著水平了。例如，我们把左右两边尾巴共 $\alpha=0.05$ 的抽样划出去，作为拒绝域。



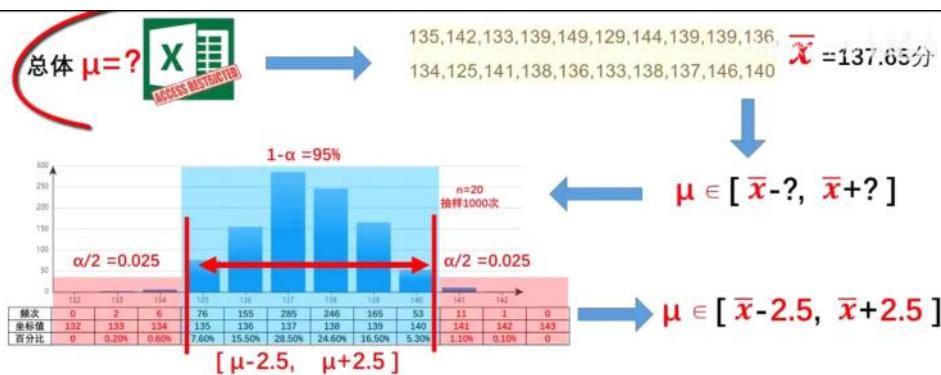
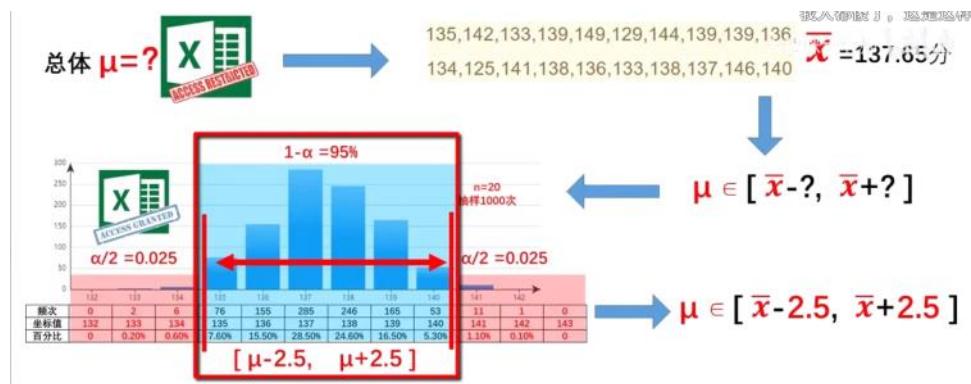
那么，中间剩下的 $1-\alpha=95\%$ ，就是我们的接受域。这个接受域，实际上就是一个区间。在我们这个演示用的，不精确的表格中，这个区间的粗略范围是[135,140]分。对称轴代表的总体均值 $\mu=137.41$ 分，我们粗略的把这个区间~~固定~~作在137.41分上下浮动了2.5分。于是，这个接受域的区间，可以表达为 $[\mu-2.5, \mu+2.5]$ 分。



注意，下面我说的这句话，又比较拗口了。在明明知道总体均值 μ 的情况下，对总体进行1000次抽样，并不是每次抽样均值都等于 μ 的，而是有95%的抽样，落在了 $[\mu-2.5, \mu+2.5]$ 这么一个区间内。这个区间的浮动范围，是在 μ 左右2.5分。



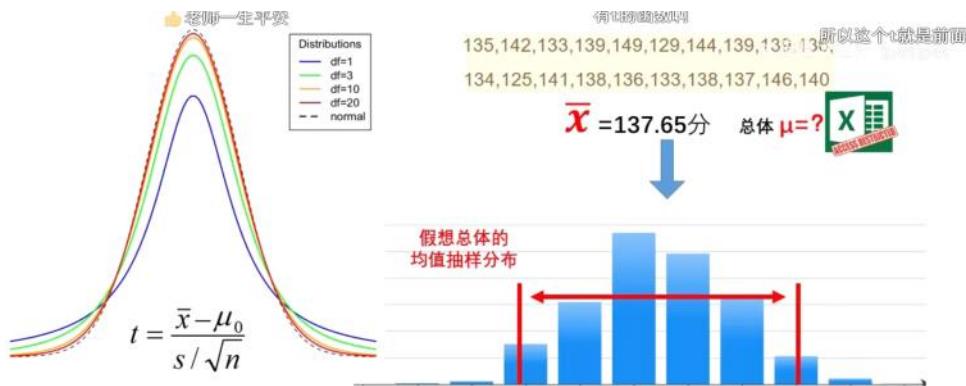
那么，在不知道总体均值 μ ，而且只能抽样1次的情况下，得到一个样本均值137.65分。我们假如想估计一个区间，说真实总体均值 μ ，可能在这个样本均值左右浮动的话。那么，浮动的范围是多少分呢？就干脆也用刚才这个2.5分吧，作为这个区间的浮动范围。



由此看来，构造置信区间的关键，不是抽样均值，因为抽样均值一算就出来了，没啥不好懂的。关键是这个左右浮动的范围。大家到这里，肯定有疑问。这个2.5，是怎么来的呢？在这个例子中，我们是从抽样分布中数出来的2.5分。之所以能数出来，是因为这个总体其实就是我们自己事先造出来的，我们有权限看到所有数据，也能通过程序，模拟1000次抽样。

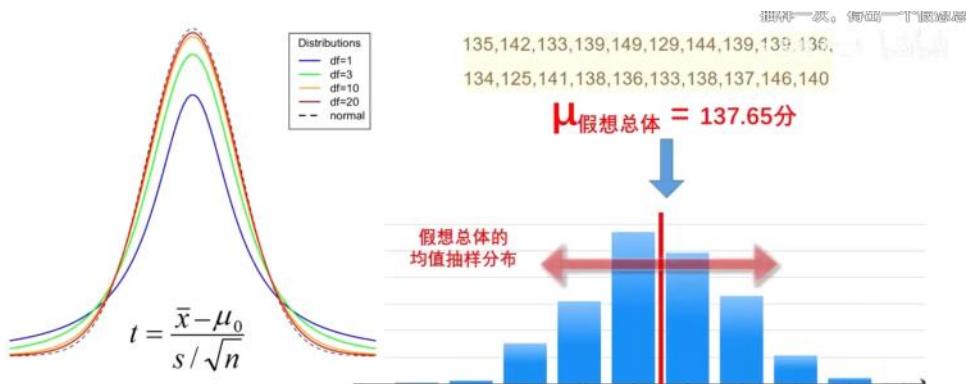


认真听课的同学，肯定觉得这句话似曾相识。对，这就是第6节，“从均值抽样分布到t分布”时，我们讲过的。所以，我们不可能，也不用对每一个案例都抽样1000次。



任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

反推出均值抽样分布后，我们就可以得到95%的接受区域，也就知道了置信区间上下浮动的范围。所以，现在请记住，我们的任务是，通过单次抽样，反推均值抽样分布的95%接受区域。

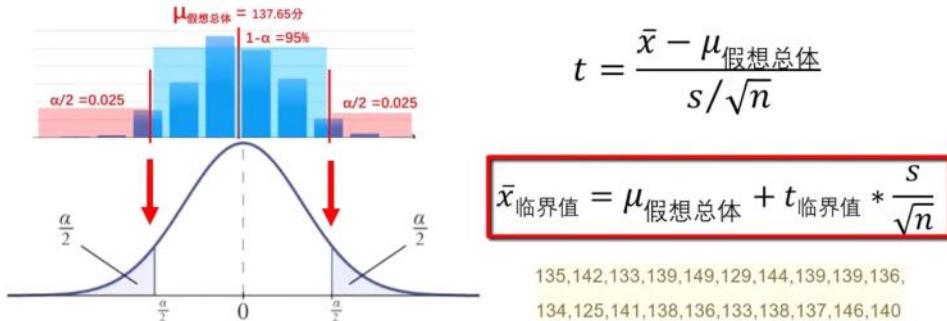


任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

为了避免混淆，我们把这个假想总体的均值记作 $\mu_{\text{假想总体}} = 137.65$ 分。下面这句话，也比较拗口，大家注意听。我们通过单次抽样和t分布，反推出来的这个均值抽样分布，就是从这个均值为137.65分的假想总体中，抽样100次形成的抽样分布。那么，这个假想抽样分布的95%的接受域在哪里呢？

那个假想总体并不等于抽样的平均值啊

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间



抽样分布中的 \bar{x} 临界值，对应着t分布中的t临界值。这个对应关系，是通过t值公式来实现的。注意，这里的 \bar{x} 是假想总体中的抽样均值。这时，我们通过小学数学知识，把t值的公式变换一下，便得到 \bar{x} 临界值的计算公式。其中，t临界值可以通过查表得到，样本标准差s，可以通过样本算出来。

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间											
因为这里的样本为20个 所以自由度：20（样本）-1=19											
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.770	3.747	4.604	7.173	8.610
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.329	1.725	2.090	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792

$$\bar{x}_{\text{临界值}} = \mu_{\text{假想总体}} + t_{\text{临界值}} * \frac{s}{\sqrt{n}}$$

双边 $t = \pm 2.093$

我们先来查表。在t临界值表中，找到自由度df=19，双尾 $\alpha=0.05$ ，也就是中间95%的t临界值为2.093。注意，t值表中只给出正的t临界值，我们心中要知道，其实有 ± 2.093 两个t临界值。

s公式

-1, 样本n为20

20-1df
是谁在讲洋文

贝塞尔修正

先想t分布弄明白。你就能看懂了
因为样本总量是20。-1得自由度

df=19是由N-1得来

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

135,142,133,139,149,129,144,139,139,136,

134,125,141,138,136,133,138,137,146,140

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad s = 5.547$$

$$\bar{x}_{\text{临界值}} = \mu_{\text{假想总体}} + t_{\text{临界值}} * \frac{s}{\sqrt{n}}$$

然后，样本标准差s，通过公式算出。得s等于5.547。

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

$$\bar{x}_{\text{临界值}} = \mu_{\text{假想总体}} + t_{\text{临界值}} * \frac{s}{\sqrt{n}} \quad s = 5.547 \quad n = 20$$

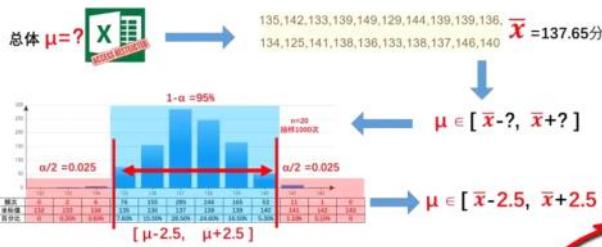
$$\mu_{\text{假想总体}} = 137.65 \text{分} \quad t_{\text{临界值}} = \pm 2.093$$

$$\bar{x}_{\text{临界值}} = 137.65 \pm 2.093 * \frac{5.547}{\sqrt{20}}$$

$$\bar{x}_{\text{临界值}} = 137.65 \pm 2.596$$

于是，我们把t=±2.093, s = 5.547, n=20，代入到公式中，得到x临界值等于137.65加减2.596分。

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

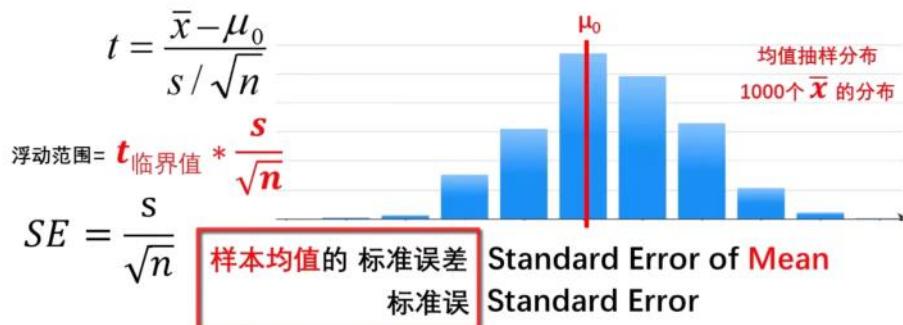


$$\bar{x}_{\text{临界值}} = 137.65 \pm 2.596$$

$$\text{浮动范围} = t_{\text{临界值}} * \frac{s}{\sqrt{n}}$$

这个2.596，就是点估计上下浮动的精确范围。这就回答了之前的问题：“总体μ的置信区间中，上下浮动范围到底怎么来的”，就是用t临界值乘以样本标准差再除以样本容量的平方根算出来的。

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间

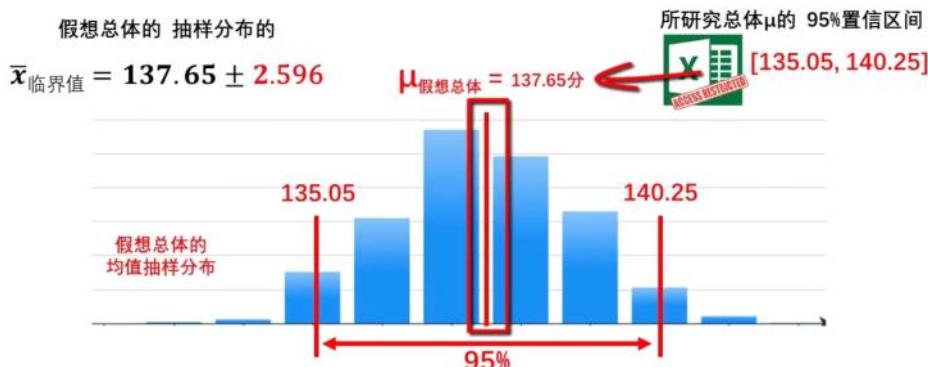


之前我们讲t值公式的时候，这个样本标准差除以样本容量的平方根，当时没有详细展开。现在大家能坚持听到这节课的话，说明大家已经比较专业了。今天，我们给出这个 s 除以根号n的术语。这个东西，叫做“样本均值的标准误差”，简称“标准误”。英语叫做standard error of mean，简称standard error，缩写为SE。

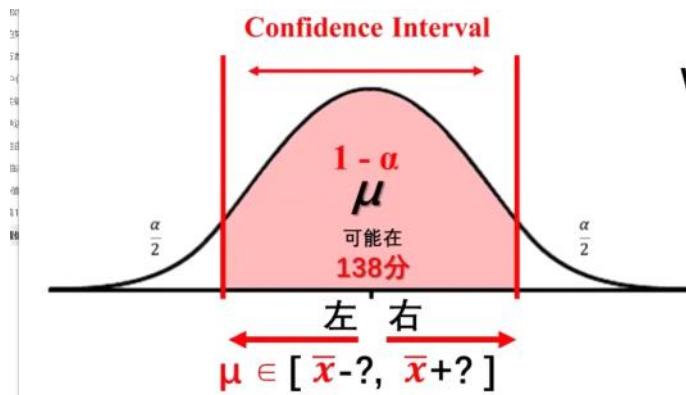


汉语和英语的缩写，都有一定的问题，把“均值”这个关键词给省略了。所以，大家一定要明白，标准误不是所研究总体的标准差，而是所有抽样均值的标准差。例如，这是1000个样本均值的抽样分布，这1000个 \bar{x} 作为一个总体的话，这个总体的标准差，就是标准误。所以，t值其实就可以解释为，一个抽样的均值，偏离了抽样分布的对称轴多少个标准误。

任务目标：通过单次抽样，反推均值抽样分布的95%接受域区间



现在，我们回到任务目标。我们已经完全反推出了这个假想的抽样分布。假想抽样分布的对称轴，就是从所研究总体中，抽样1次得到的样本均值137.65分。假想抽样分布的95%的接受域区间，由137.65加减2.596分算出来，也就是从135.05分到140.25分。这个区间，就是用来估计所研究总体均值 μ 的95%的置信区间。

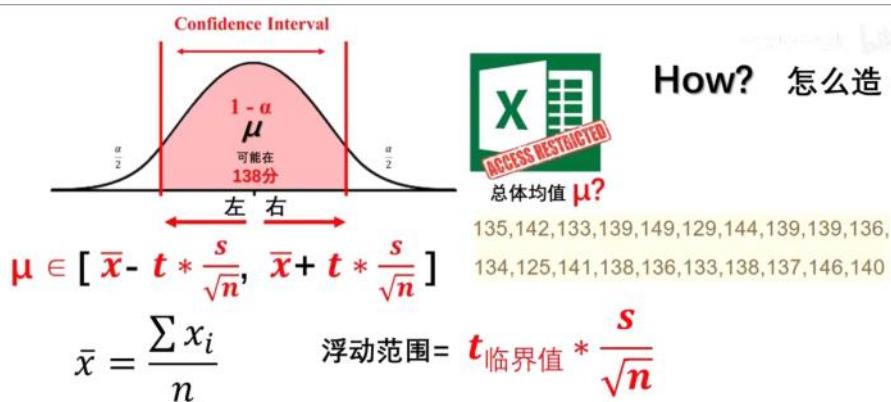


What? 是什么



总体均值 μ ?

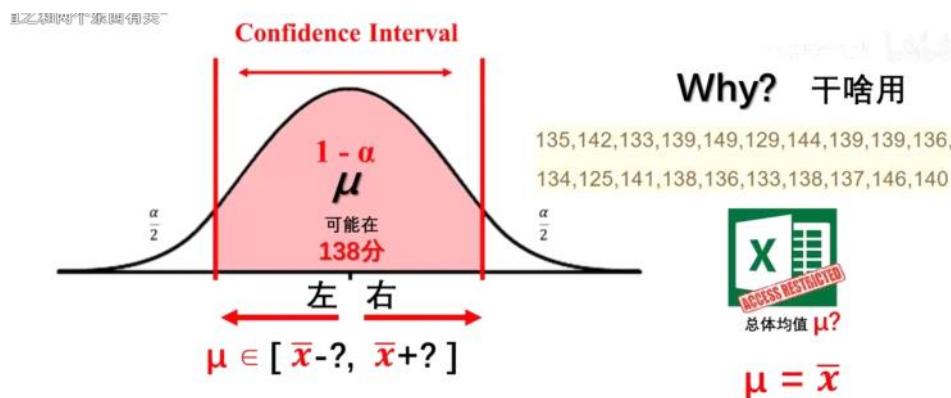
现在，我们来总结回答一下本节课开始提出的三个问题。第一个问题：置信区间是什么？置信区间，是对一个未知总体的均值 μ 的区间估计。区间估计由两部分组成：一个是点估计，也就是从未知总体中抽样1次所得的样本均值，另一个是上下浮动的范围。



How? 怎么造



总体均值 μ ?
135, 142, 133, 139, 149, 129, 144, 139, 139, 136,
134, 125, 141, 138, 136, 133, 138, 137, 146, 140



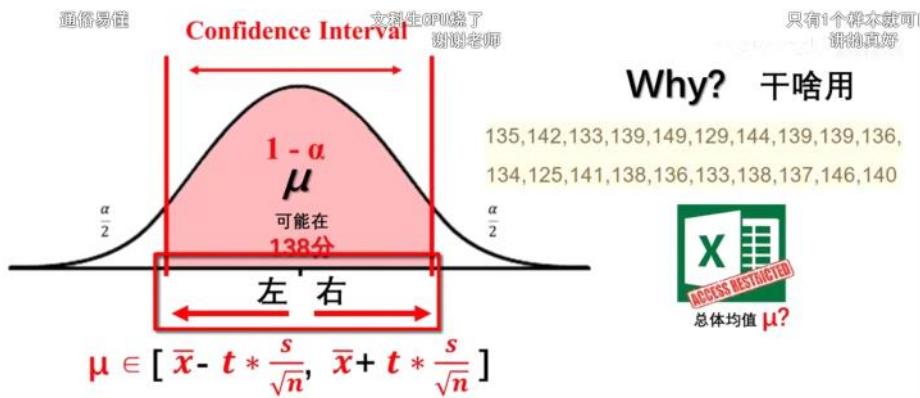
Why? 干啥用

135, 142, 133, 139, 149, 129, 144, 139, 139, 136,
134, 125, 141, 138, 136, 133, 138, 137, 146, 140



总体均值 μ ?

第三个问题：为什么用置信区间？有3个理由。第一，由于各种限制，很多时候，我们只能对一个未知总体抽样1次。我们总不能说总体均值 μ ，就直接等于样本均值 \bar{x} 。我们要加一个上下浮动的范围，形成一个区间估计。



第二，在总体未知的情况下，这个样本就是我们获得的关于总体的唯一的数据，要好好利用，不要只算出个样本t值来就扔一边了。我们还可以算出样本标准差s，并且进一步来估算出标准误。置信区间的宽度，其实就是2t个标准差。

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

Why? 干啥用

没有实际意义

$t=2.09$ 是几只包子? $H_0: \mu=\mu_0$ 只包子

$t=2.86$ 是多少分? $H_0: \mu=\mu_0$ 分

第三，其实也是很直观的一个理由。t值是从所有案例中抽象出来的，对称轴等于0（因为减去了 μ_0 ），没有单位（因为做了除法），没有任何实际意义的一个值。你不知道 $t=2.09$ 代表着几只包子？也不知道 $t=2.86$ 代表着多少分？不仅如此，要算出一个t值的话，你还得有个原假设，才能有个 μ_0 。

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

Why? 干啥用

135, 142, 133, 139, 149, 129, 144, 139, 139, 136,
 134, 125, 141, 138, 136, 133, 138, 137, 146, 140

$t=2.09$ 是几只包子? $H_0: \mu=\mu_0$ 只包子

$t=2.86$ 是多少分? $H_0: \mu=\mu_0$ 分



$$\mu \in [\bar{x} - t * \frac{s}{\sqrt{n}}, \bar{x} + t * \frac{s}{\sqrt{n}}]$$

但很多情况下，我们就是想抽一次样，并没有想去做什么假设检验，也并没有想和哪个假想的 μ_0 去比较。我们只是想利用这个抽样，做出一个区间估计。



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

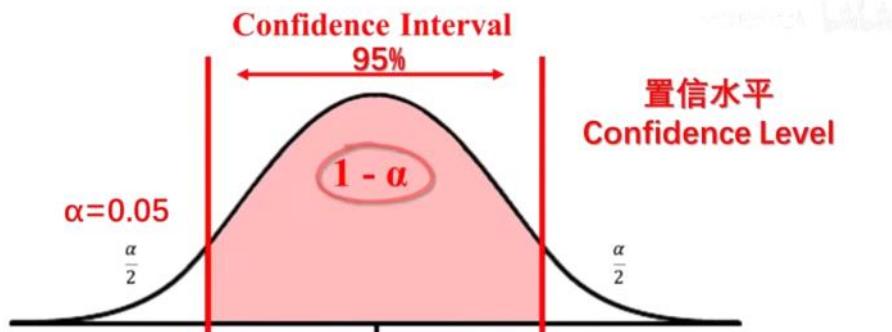
135,142,133,139,149,129,144,139,139,136,
134,125,141,138,136,133,138,137,146,140

95% 置信区间
 $\mu \in [138 - 2.5, 138 + 2.5]$ 分

例如，在一次全校考试后，我们通过20个同学的抽样均分，算出一个置信区间，我们就有95%的信心（confidence）说：全校均分 μ 应该在138分上下2.5分左右。这里，又有实际的数值，138，又有浮动范围，上下2.5，而且还带有单位，“分”。所以说，在这种情况下，置信区间比一个光秃秃的t值更能表达实际意义。

置信水平confidence level 区间估计

2024年1月2日 14:33



本节课，我们来详细解释一下置信水平这个概念。置信水平的英语叫做 Confidence Level。置信水平是一个百分比。通过上节课置信区间的学习，我们其实已经知道了，在确定了显著水平 α 的情况下，置信水平就是 $1-\alpha$ 。例如，假如 $\alpha=0.05$ 的话，置信水平就是 $1-\alpha=95\%$ 。

Confidence level is **NOT** the probability that a **specific** confidence interval contains the population parameter.

- <https://blog.minitab.com>

置信水平**不是某个置信区间包含真实总体均值 μ 的概率。**

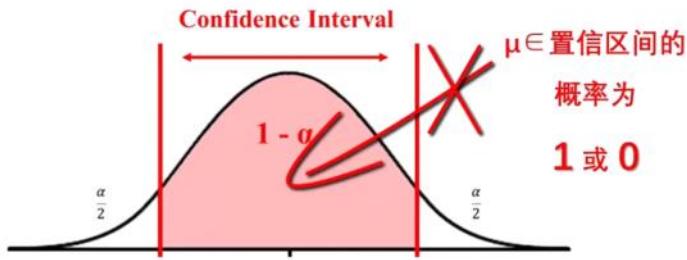
“这个置信区间**可能**包含了 (contain) 总体均值 μ ”

“这个置信区间**有95%的概率**包含了总体均值 μ ” **(错误)**

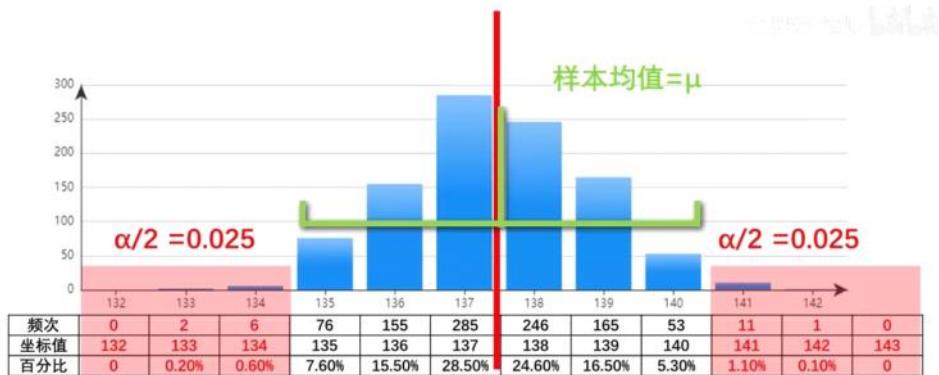
那么，95%这个置信水平应该如何理解呢？很多参考资料上特别说明：置信水平不是某个置信区间包含真实总体均值 μ 的概率。我们构造了一个具体的 (specific) 区间，并用这个具体的区间来估计总体均值 μ 的范围。我们上节课中说，“这个置信区间**可能**包含了 (contain) 总体均值 μ ”。但我们并没有说，“这个置信区间**有95%的概率**包含了总体均值 μ ”。

总体永远是未知的

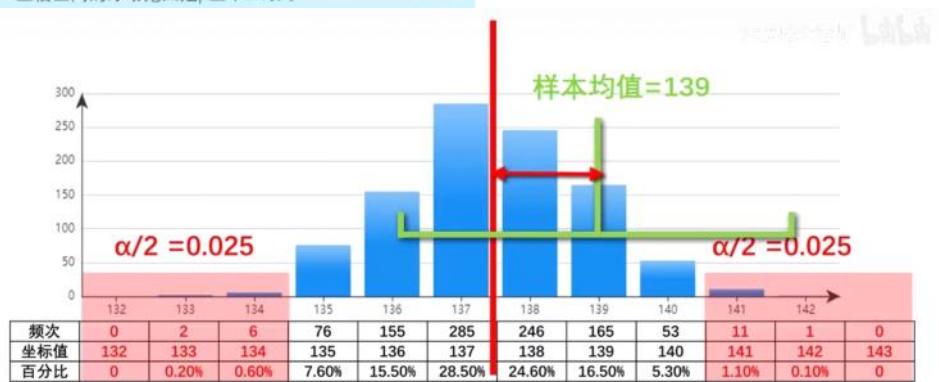
总体均值 μ 是一个未知但存在的常数。



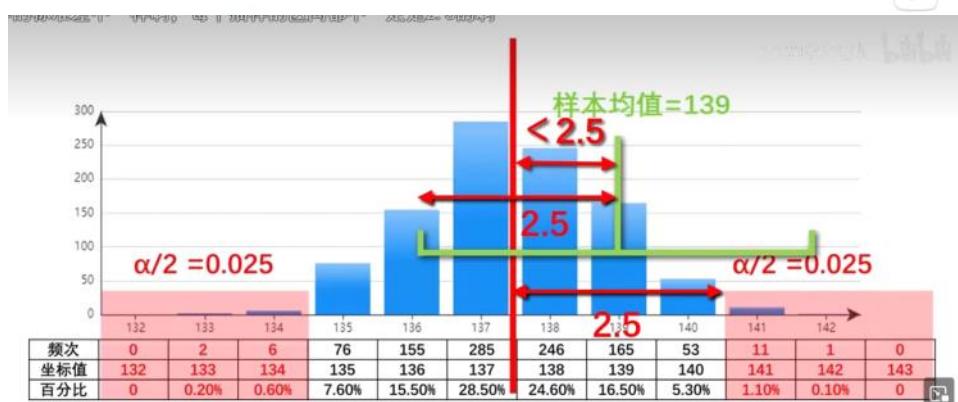
一种解释是：这是因为，真实世界的统计案例中，所研究总体的全貌永远是未知的，所以真实的总体均值 μ ，也是一个未知的但确实存在的常数。对于仅通过1次抽样而造出来的一个具体的置信区间，因为我们不知道总体均值 μ 等于多少，所以， μ 要么就在这个区间内，要么就不在这个区间内。概率是1或者0，不存在其他的什么概率。



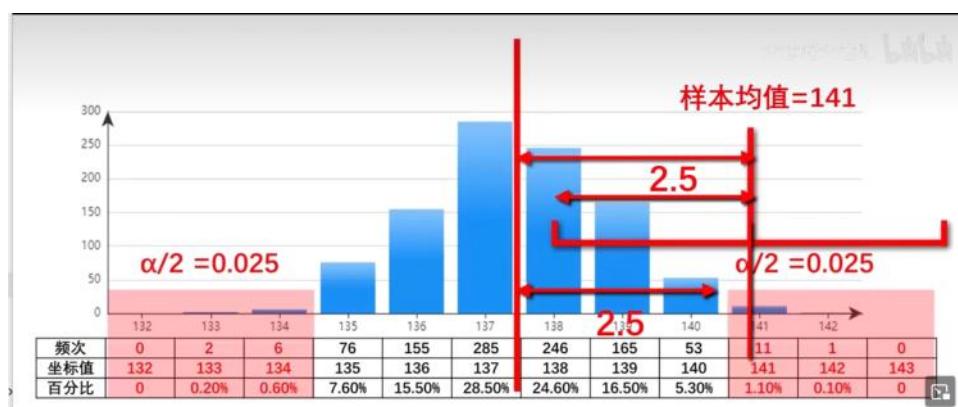
例如，这是一个从5000名学生的英语成绩总体中，抽样1000次形成均值抽样分布。左右两边红色的阴影部分，是双边 $\alpha=0.05$ 的拒绝域。中间红色对称轴，代表着总体真实均值 $\mu=137.41$ 分。假如抽样1次，样本均值正好等于总体均值 μ 。我们以这个正好等于 μ 的样本均值，构造一个95%的置信区间。为了方便演示和讲解，我们仍然粗略的认为置信区间的浮动范围是 μ 上下2.5分。



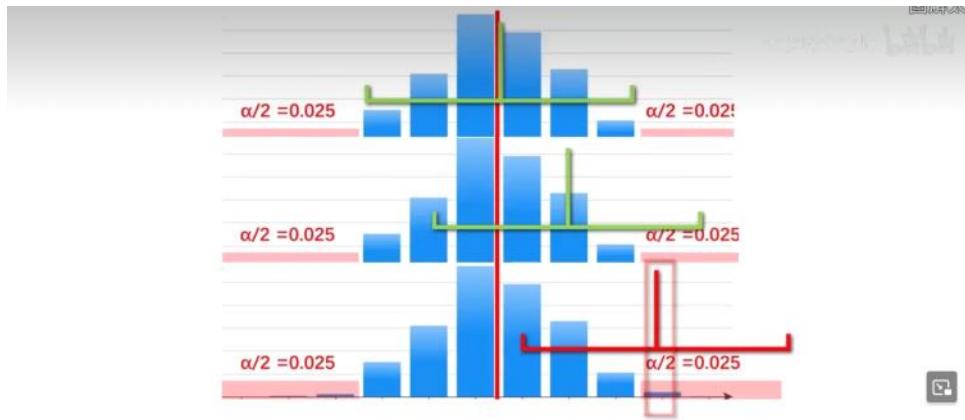
这时，假如再抽样一次，获得样本均值139分，则比总体均值 μ 往右边偏了一些。我们以这个均值139分，再构造一个置信区间，置信区间上下浮动范围，仍然取值粗略的2.5分。



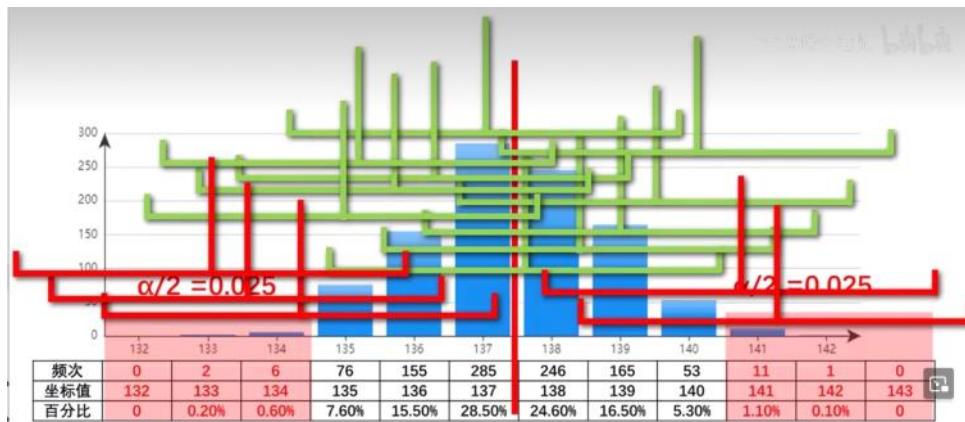
下面的话，比较拗口，大家注意听。因139分，还在接受域内，而这一半接受域的宽度为2.5分，所以，这个偏移量，必然小于2.5分。与此同时，置信区间的宽度的一半，也是2.5分。既然偏移量小于2.5分，则说明置信区间左边的端点，还没有移过总体均值 μ 。所以，这个置信区间，也必然包含了总体均值 μ 。



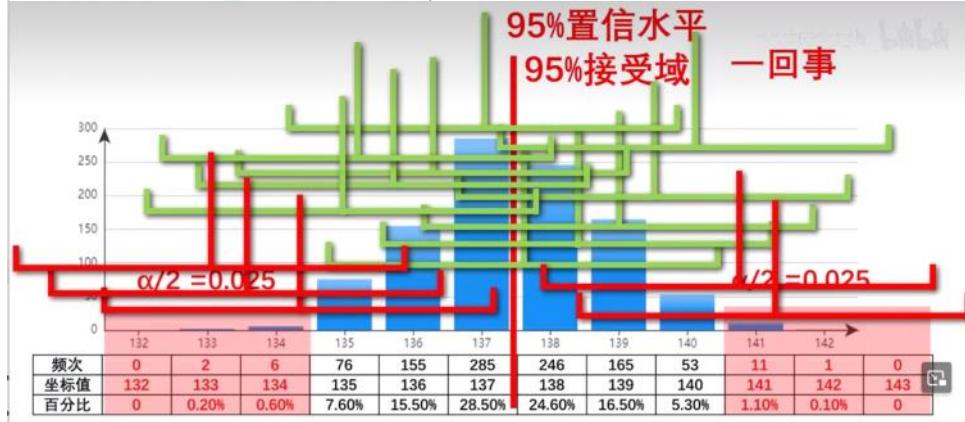
下面，再继续抽样一次，得样本均值为141分。141分偏离 μ 的偏移量，已经超过2.5分了。下面说的话，也比较拗口，大家注意听。因为141分偏离了接受域宽度的一半2.5分，于是样本均值就落入了拒绝域。与此同时，因为偏移量大于置信区间宽度的一半2.5分，所以置信区间的左端点，已经移过了总体均值 μ 。所以，这个置信区间，必然没有包含总体均值 μ 了。所以，我们用红色的“山字形”来表示这个置信区间。



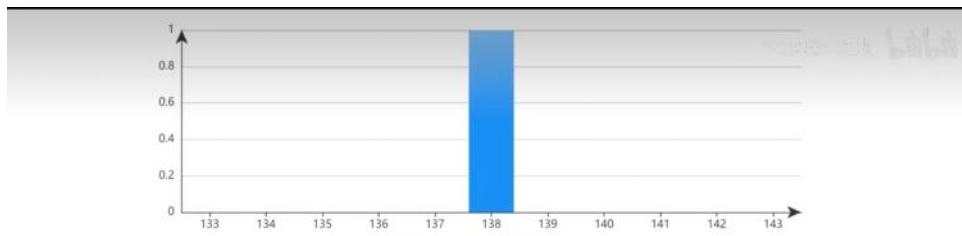
讲到这个份上，不知道大家看出点什么道道来没有？因为置信区间的宽度，和接受域的宽度，是一回事，所以，只要抽样均值落在接受域内，则构造出来的置信区间，必然包含总体均值 μ 。而只要抽样均值落在接受域外，也就是落在了拒绝域内，则构造出来的置信区间，必然不包含总体均值 μ 。



如此看来，置信区间的置信水平95%，和抽样分布的95%的接受域，这两个95%，是一回事。所以说，95%置信水平，不是指某一个具体的置信区间包含总体均值 μ 的概率，而是假如有能力进行重复大量的抽样，并构造大量的置信区间的话，其中95%的置信区间，必然包含总体均值 μ 。



如此看来，置信区间的置信水平95%，和抽样分布的95%的接受域，这两个95%，是一回事。所以说，95%置信水平，不是指某一个具体的置信区间包含总体均值 μ 的概率，而是假如有能力进行重复大量的抽样，并构造大量的置信区间的话，其中95%的置信区间，必然包含总体均值 μ 。



$$[\bar{x} - 2.5, \bar{x} + 2.5]$$

$$\text{浮动范围} = t_{\text{临界值}} * \frac{s}{\sqrt{n}}$$



需要再次指出的是，为了讲解方便和直观理解，我们把所有置信区间的浮动范围，都统一取值为粗略的2.5分。但实际上，根据置信区间浮动范围的公式可以得知，每次抽样的样本标准差 s 都不尽相同，所以，每个置信区间~~的宽度~~也不尽相同。特此说明。好，本节课就到这里，我们下节课见。

单样本t检验

2024年1月2日 14:58

某高中，王老师班，20名同学英语成绩： $(\bar{x} = 135.8)$ 全校均分 $\mu_0 = 137$

136, 136, 134, 136, 131, 133, 142, 146, 137, 140, 134, 135, 136, 132, 119, 132, 145, 131, 140, 141

我们仍然以英语成绩为例。例如，有一个高中，王老师负责一个高三班级的英语课程，班级内有20位同学。在此次全年级英语考试后，全校的英语平均分为137分，王老师班的平均分为135.8分。

某高中，王老师班，20名同学英语成绩： $(\bar{x} = 135.8)$ 全校均分 $\mu_0 = 137$

136, 136, 134, 136, 131, 133, 142, 146, 137, 140, 134, 135, 136, 132, 119, 132, 145, 131, 140, 141

我们把全校的英语成绩看做一个总体，总体均分 $\mu_0 = 137$ 分。然后，我们把王老师班看做全校的一个抽样，样太容
量 $n=20$ ，样本均分 $\bar{x}=135.8$ 分。 μ_0 是已知确定的，是137分。但当我们把王老师班拿出来说事，就相当于暂时假设王老师班不是来自这个学校，而是来自一个假想的总体，这个假想总体的均值，记为 μ 。

某高中，王老师班，20名同学英语成绩： $(\bar{x} = 135.8)$ 全校均分 $\mu_0 = 137$

136, 136, 134, 136, 131, 133, 142, 146, 137, 140, 134, 135, 136, 132, 119, 132, 145, 131, 140, 141

1. 王老师的班级均分和全校总体均分有无显著区别？ $H_0: \mu = \mu_0$
(H_0 : 王老师班这个样本，来自均分“ $\mu = 137$ 分”的一个总体。)
2. 王老师的班级均分是否显著低于全校总体均分？ $H_0: \mu < \mu_0$
(H_0 : 王老师班这个样本，来自均分“ $\mu < 137$ 分”的一个总体。)
3. 王老师的班级均分是否显著高于全校总体均分？ $H_0: \mu > \mu_0$
(H_0 : 王老师班这个样本，来自均分“ $\mu > 137$ 分”的一个总体。)

下面我们问3个问题，每个问题其实对应着一个原假设 H_0 。第1个问题：王老师的班级均分和全校总体均分有无显著区别。里面的“有无”，就是“等于还是不等于”的问题，显然是一个双边假设检验。我们把原假设写出来是， $H_0: \mu = \mu_0$ 。原假设的含义是：王老师班这个样本，来自均分 $\mu = 137$ 分的一个总体，也就是说，来自本学校。

某高中，王老师班，20名同学英语成绩： $(\bar{x} = 135.8)$ 全校均分

136, 136, 134, 136, 131, 133, 142, 146, 137, 140, $\mu_0=137$
134, 135, 136, 132, 119, 132, 145, 131, 140, 141

1. 王老师的班级均分和全校总体均分有无显著区别？ $H_0: \mu=\mu_0$

(H_0 : 王老师班这个样本，来自均分 " $\mu=137$ 分" 的一个总体。)

2. 王老师的班级均分是否显著低于全校总体均分？ $H_0: \mu < \mu_0$

(H_0 : 王老师班这个样本，来自均分 " $\mu < 137$ 分" 的一个总体。)

3. 王老师的班级均分是否显著高于全校总体均分？ $H_0: \mu > \mu_0$

(H_0 : 王老师班这个样本，来自均分 " $\mu > 137$ 分" 的一个总体。)

第2个问题和第3个问题，里面直接出现了小于号和大于号，显然都是单边假设检验。假如把字面意思写成原假设的话，那么，第2个问题的原假设就是 $H_0: \mu < \mu_0$ ，意思是：王老师班这个样本，来自均分 $\mu < 137$ 分的一个总体。第3个问题的原假设就是 $H_0: \mu > 137$ ，意思是：王老师班这个样本，来自均分 $\mu > 137$ 分的一个总体。假如大

某高中，王老师班，20名同学英语成绩：

136, 136, 134, 136, 131, 133, 142, 146, 137, 140,
134, 135, 136, 132, 119, 132, 145, 131, 140, 141

班级均分 全校均分

$(\bar{x} = 135.8)$ vs $\mu_0=137$

单样本 t 检验

One Sample t -test

上述3个问题中，都只有一个样本，就是王老师这个班级。拿一个样本均值和一个总体均值进行比较的t检验，叫做“单样本t检验”，英语叫做“One sample t-test”。注意，是“单样本”，不是“独立样本”。“独立样本”是另外一个东西，名字本身带有误导性，后面课程中我们马上就讲到。

1. 王老师的班级均分和全校总体均分有无显著区别？ $H_0: \mu=\mu_0$

(H_0 : 王老师班这个样本，来自均分 " $\mu=137$ 分" 的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

我们先计算第1个问题：王老师的班级均分和全校总体均分有无显著区别？原假设是 $H_0: \mu=137$ 。我们通过如下代码，将数据输入到软件中。首先，给王老师班的这个样本命名，记作wang_class。然后，用c()命令，将样本组合成一个向量或数组。c是combine，“组合”的英文的缩写。向量的概念，大家高中数学应该学过，实在忘了，就当它是个数组就行。这里这个加号，是当你的一行代码没有结束，但按了回车键时，程序自动加上的。它是一个换行提示符，不是你敲进去的。

1. 王老师的班级均分和全校总体均分有无显著区别？ $H_0: \mu=\mu_0$

(H_0 : 王老师班这个样本，来自均分 " $\mu=137$ 分" 的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

Student's t-Test

所以，这两行代码，就把王老师班这一个样本，一组数据，赋值给了wang_class这个变量。然后，我们用t.test()函数，来进行t检验。t.test()函数，在R软件里特指Student's t-Test。我们在后面课程中，会专门用一节课，来讲解Student's t-Test发明的历史。大家现在只需要知道，统计学中，**不仅仅只有Student's t-Test一种t检验**，还有**其他类型的t检验**。本系列课程讲到今天，用的都是Student's t-Test。

1. 王老师的班级均分和全校总体均分有无显著区别? $H_0: \mu = \mu_0$

(H_0 : 王老师班这个样本, 来自均分 " $\mu=137$ 分" 的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

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

`t.test()`函数, 需要输入3个参数, 第1个参数是样本, 也就是`wang_class`, 通过这个样本, 程序会算出样本均值 \bar{x} 、样本标准差 s 和样本容量 n ; 第2个参数是已知总体的均值, 对应着公式中的 μ_0 , 也就是已知全校总体的均分137分; 第3个参数, `alternative`, 即备用假设, 也就是 H_1 的意思, 这个参数告诉程序, 这个t检验, 是一个双边检验, 还是一个单边左尾检验, 还是一个单边右尾检验。这里, 我们给`alternative`赋值一个字符串"two.sided" 代表这是一个双边t检验。

12

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

One Sample t-test

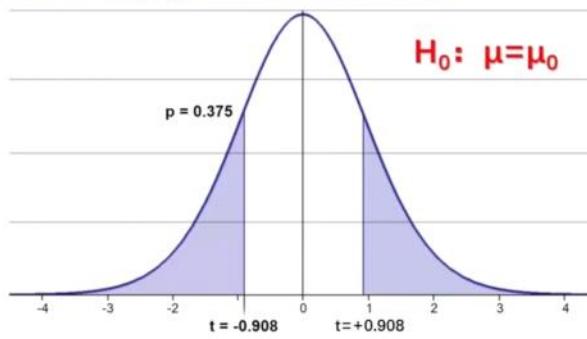
```
data: wang_class  
t = -0.90834, df = 19, p-value = 0.3751  
alternative hypothesis: true mean is not equal to 137  
95 percent confidence interval:  
133.0349 138.5651  
sample estimates:  
mean of x  
135.8
```

在`t.test()`函数输入完毕后, 回车, 就可以看到这样的计算结果。首先, 程序告诉我们, 这是一个单样本t检验, One Sample t-test。没错, 因为输入到`t.test()`函数里的, 只有一个样本。样本数据, `data`, 就是`wang_class`这个变量。`t`值算出来等于-0.90834。单样本t检验的自由度`df`是样本容量 $n=20$, 减去1, 等于19。`p`值算出来是0.3751。这些概念, 之前课程中都讲过了, 现在, 再来回忆一下。

`t = -0.90834, df = 19, p-value = 0.3751`

p-value: 0.375
t-value: -0.90834

d.f.: 19
 two tails
 right tail
 left tail
 0 to t
 -t to t



下面, 我们把计算出来的结果, 在t分布中画出来。在这个自由度为19的t分布中, `t`值等于-0.90834。因为第1个问题的原假设是 $H_0: \mu = \mu_0$ 137分, 所以, 这是一个双边t检验。于是, 我们还要找到t分布对称轴另一边的 $t=+0.90834$ 。在这两个对称的`t`值处, 分别划两条线, 比这两条线还往两边极端尾巴方向上去的, 曲线下的阴面积是0.375, 也就是`p`值等于0.375。

16

```

> wang_class=c(136,136,134,136,131,133,142,146,137,140,
+ 134,135,136,132,119,132,145,131,140,141)
> t.test(wang_class, mu=137, alternative="two.sided")

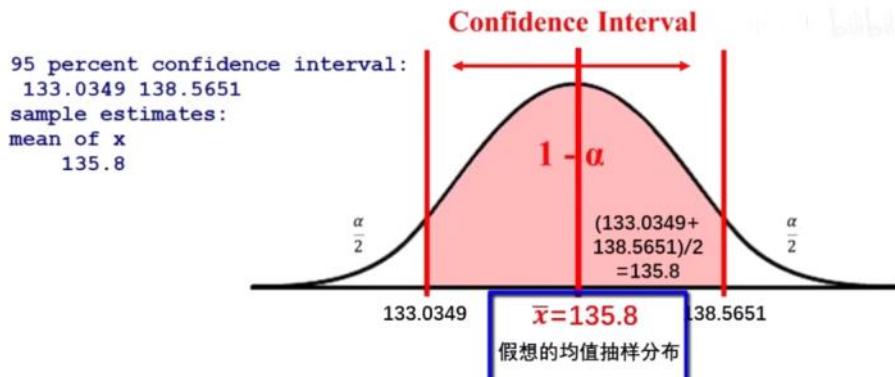
One Sample t-test

data: wang_class
t = -0.90834, df = 19, p-value = 0.3751
alternative hypothesis: true mean is not equal to 137
95 percent confidence interval:
133.0349 138.5651
sample estimates:
mean of x
135.8

```

然后，再看计算结果中的下面一行，alternative hypothesis就是备用假设的意思。因为我们在t.test()函数中注明了，这是一个双边检验，所以程序自动认定，我们的备用假设 H_1 ，就是 $\mu \neq 137$ 。这就是程序提醒我们，自己设定的到底是双边还是单边，单边的话，是右边还是左边。所以，这里“not equal”，就是双边，对应“two.sided”。再下面一行，就是95%的置信区间了。意思是，通过王老师班这个样本算出来的，对王老师班所代表的一个总体的均分 μ 的一个区间估计。这个区间是从133.0349分到138.5651分。

17



我们把这个置信区间也画出来。95%置信区间，置信水平为95%，则显著水平 α 为0.05。置信区间的对称轴，是 $(133.0349+138.5651)/2=135.8$ 分，也就是王老师的班级均分。这个置信区间的示意图中的分布，是我们假想的、王老师班级可能代表的、一个总体均值 $\mu=135.8$ 分的均值抽样分布图。这里的sample estimates，字面意思是样本估计，就是指对这个假想总体 μ 的点估计。mean of x，就是王老师班的样本均值，135.8分。

1.王老师的班级均分和全校总体均分有无显著区别

- 王老师的班级均分和全校总体均分有无显著区别？
 $H_0: \mu = \mu_0$ (王老师班这个样本，来自均分“ $\mu=137$ 分”的一个总体。)

```

> wang_class=c(136,136,134,136,131,133,142,146,137,140,
+ 134,135,136,132,119,132,145,131,140,141)
> t.test(wang_class, mu=137, alternative="two.sided")

```

```

One Sample t-test

data: wang_class
t = -0.90834, df = 19, p-value = 0.3751
alternative hypothesis: true mean is not equal to 137
95 percent confidence interval:
133.0349 138.5651
sample estimates:
mean of x
135.8

```

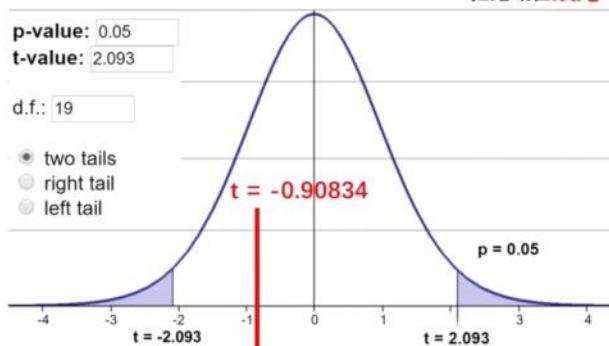
以上，就是通过R软件，对第1个问题， $H_0: \mu=137$ 分的一个计算结果。但是，程序并没有告诉我们，应该拒绝 H_0 ，还是接受 H_0 。这个终极问题，是需要我们的**人脑**，结合实际情况做出判断的。我们根据计算结果，再复习一下之前学过的知识，进而做出判断。首先，应该确定一个显著水平。例如，我们选择常规的 $\alpha=0.05$ 。然后，我们用3种方法进行判断。

$t = -0.90834, df = 19, p\text{-value} = 0.3751$

$H_0: \mu = \mu_0$

拒绝域在双尾

df=19
双边 $\alpha=0.05$
 $t_{\text{临界值}}=2.093$



查完了表，我们脑中，要自行补足这样一个t分布图。自由度为19的t分布中，双边 $\alpha=0.05$ ，有两个t临界值， ± 2.093 。然后，划出双边尾巴的两块拒绝域。通过样本算出来的 $t=-0.90834$ ，显然没有落入拒绝域。于是，我们的结论是，在双边 $\alpha=0.05$ 的显著水平下，无法拒绝 H_0 ，从而接受 H_0 ，认为王老师的班级均分和全校总体均分没有显著差别。这是按第1种方法，t临界值查表，来判断是否拒绝 H_0 。

21

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

```
One Sample t-test           $\alpha=0.05$   
data: wang_class           p >  $\alpha$  接受 $H_0$   
t = -0.90834, df = 19, p-value = 0.3751  
alternative hypothesis: true mean is not equal to 137  
95 percent confidence interval:  
 133.0349 138.5651  
sample estimates:  
mean of x  
 135.8
```

第2种判断方法：直接看p值。这就比较简单了。 $p=0.3751$ ， p 大于显著水平 $\alpha=0.05$ ，所以， p 不显著，于是，接受 H_0 。对p值用法还有疑问的同学，请复习第10节《p值的含义》。

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

```
One Sample t-test           $\mu_0 \in [133.0349, 138.5651]$   
data: wang_class           接受 $H_0$   
t = -0.90834, df = 19, p-value = 0.3751  
alternative hypothesis: true mean is not equal to 137  
 $\alpha=0.05$  95 percent confidence interval:  
 133.0349 138.5651  
sample estimates:  
mean of x  
 135.8
```

第3种判断方法：看置信区间是否包含 μ_0 ，也就是137分。置信水平是95%，也就是显著水平是0.05。置信区间 $[133.0349, 138.5651]$ 包含了总体均值 $\mu_0=137$ 分，因此，接受 H_0 。对此还有疑问的同学，请复习第13节《置信水

1. 王老师的班级均分和全校总体均分有无显著区别? $H_0: \mu = \mu_0$

(H_0 : 王老师班这个样本, 来自均分“ $\mu = 137$ 分”的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="two.sided")
```

```
One Sample t-test  
  
data: wang_class  
t = -0.90834, df = 19, p-value = 0.3751  
alternative hypothesis: true mean is not equal to 137  
95 percent confidence interval:  
133.0349 138.5651  
sample estimates:  
mean of x  
135.8
```

我们总结一下, R只负责计算, 算出了t值, p值, 置信区间, 样本均值等结果。但R不负责判断, R也不负责给你设定显著水平。具体是拒绝还是接受 H_0 , 您得自己拿主意。我们通过3种判断方法, 得出了相同的结论, 即, 在 $\alpha=0.05$ 的显著水平下, 接受 H_0 , 认为: 王老师的班级均分和全校总体均分没有显著区别。第1个问题, 回答完毕。

2. 王老师的班级均分是否显著低于全校总体均分?

2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu \leq \mu_0$

(H_0 : 王老师班这个样本, 来自均分“ $\mu \leq 137$ 分”的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="greater")
```

$H_1: \mu > 137$

拒绝域在右尾的单边假设检验

现在, 我们通过R来计算第2个问题。第2个问题的原假设是, $H_0: \mu < 137$, 那么, 备用假设就是 $H_1: \mu > 137$ 。数据还是刚才的数据, wang_class这个变量不用改变。我们只把t.test()函数中的alternative参数变一下, 刚才alternative是双边检验“two.sided”, 现在, 我们把它改成“greater”。意思是告诉程序, 我们的备用假设是 $H_1: \mu > 137$, 等同于告诉程序, 我们的原假设是 $H_0: \mu < 137$ 。这时要注意: 原假设 H_0 中是小于号, 那么拒绝域在分

布的右尾, 所以, 这是一个右尾的单边假设检验。假如忘了为什么的同学, 请复习第7节《单边假设检验》。

2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu \leq \mu_0$

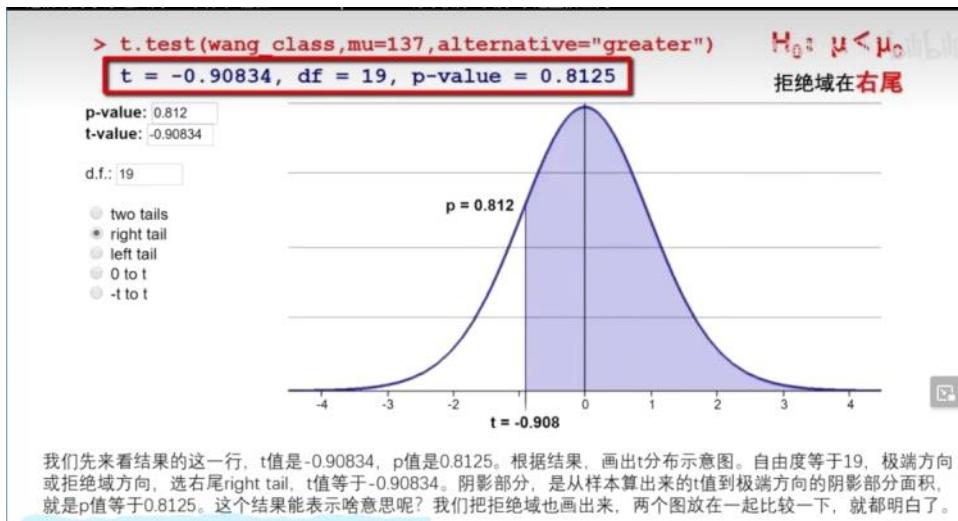
(H_0 : 王老师班这个样本, 来自均分“ $\mu \leq 137$ 分”的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="greater")
```

```
One Sample t-test
```

```
data: wang_class  
t = -0.90834, df = 19, p-value = 0.8125  
alternative hypothesis: true mean is greater than 137  
95 percent confidence interval:  
133.5157 Inf  
sample estimates:  
mean of x  
135.8
```

代码改好后, 回车, 得到如下计算结果。

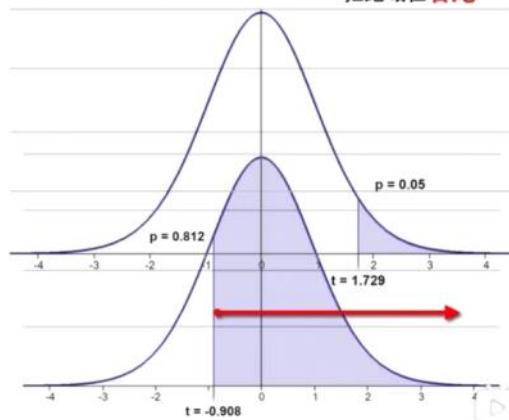


我们先来看结果的这一行, t值是-0.90834, p值是0.8125。根据结果, 画出t分布示意图。自由度等于19, 极端方向或拒绝域方向, 选右尾right tail, t值等于-0.90834。阴影部分, 是从样本算出来的t值到极端方向的阴影部分面积, 就是p值等于0.8125。这个结果能表示啥意思呢? 我们把拒绝域也画出来, 两个图放在一起比较一下, 就都明白了。

$> t.test(wang_class, mu=137, alternative="greater")$ $H_0: \mu < \mu_0$
 t = -0.90834, df = 19, p-value = 0.8125
 拒绝域在右尾

上图是拒绝域的示意图, 拒绝域的边界线, 就是单边右尾 $\alpha=0.05$ 的临界值, $t=1.729$ 。下图是样本计算出来的t值: $t=-0.908$ 。极端方向是单边右边, 则抽样得到的t值-0.908, 不比临界值+1.729极端, 所以, 不拒绝 H_0 。或者说, 抽样得到的t值-0.908, 落入了接受域, 所以, 也不拒绝 H_0 。再或者, $p=0.8125$, 大于 α , 所以, 也不拒绝 H_0 。

无论如何判断, 结论都是相同的, 即, 在 $\alpha=0.05$ 的显著水平下, “接受” H_0 , 认为: 王老师的班级均分135.8分“显著低于”全校总体均分137分。



2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu < \mu_0$

(H_0 : 王老师班这个样本, 来自均分“ $\mu < 137$ 分”的一个总体。)

```

> wang_class=c(136,136,134,136,131,133,142,146,137,140,
+ 134,135,136,132,119,132,145,131,140,141)
> t.test(wang_class, mu=137, alternative="greater")
  
```

One Sample t-test $H_1: \mu > 137$

```

data: wang_class
t = -0.90834, df = 19, p-value = 0.8125
alternative hypothesis: true mean is greater than 137
95 percent confidence interval:
 133.5157      Inf
sample estimates:
mean of x
 135.8
  
```

再看结果的这一行。这是提醒你, 你通过t.test()函数设定的备用假设, 是true mean is greater than 137, 意思是 H_1 是 $\mu > 137$, 等价于 $H_0: \mu < 137$ 。

2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu \leq \mu_0$

(H_0 : 王老师班这个样本, 来自均分 " $\mu < 137$ 分" 的一个总体。)

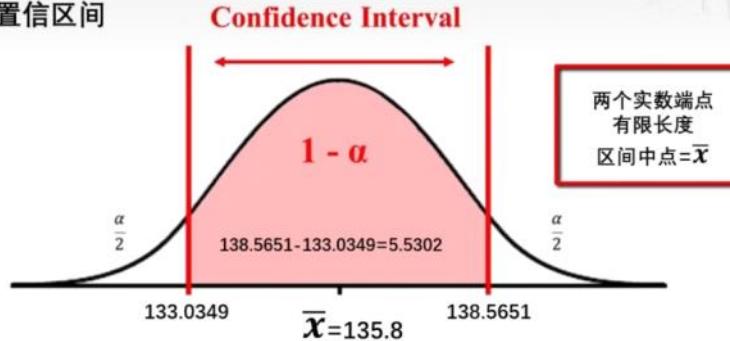
```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="greater")
```

One Sample t-test

```
data: wang_class  
t = -0.90834, df = 19, p-value = 0.8125  
alternative hypothesis: true mean is greater than 137  
95 percent confidence interval:  
 133.5157      Inf  
sample estimates:  
mean of x  
135.8
```

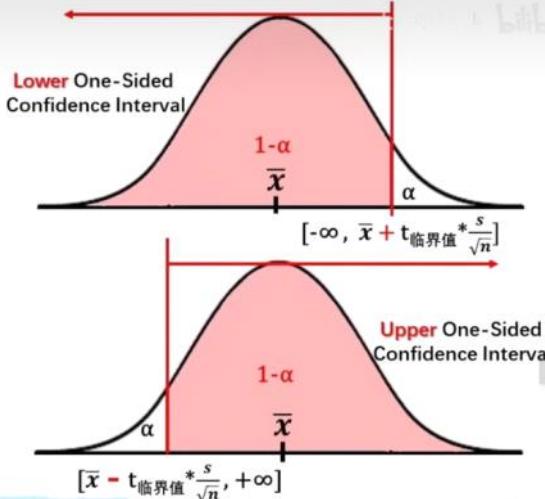
再下面, 是95%的置信区间, $[133.5157, +\infty]$ 。这个讲起来就非常之别扭, 非常的绕, 不留神还会讲错, 可能会把我自己绕进去。所以, 本节课只做简单说明, 不要求大家掌握。

双边的置信区间



之前在第12节《置信区间》中讲到的, 都是双边的置信区间, 有两个确定的实数端点, 有限的长度, 区间中点就是 \bar{x} , 比较好理解。例如, 这是刚才第1个问题中双边检验算出的置信区间, $[133.0349, 138.5651]$, 区间的宽度是 $138.5651 - 133.0349 = 5.5302$, 区间的中点, 是样本均值135.8分。

单边的置信区间



32

单边的置信区间

2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu < \mu_0$

(H_0 : 王老师班这个样本, 来自均分 " $\mu < 137$ 分" 的一个总体。)

alternative hypothesis: true mean is greater than 137
95 percent confidence interval: $H_1: \mu > 137$
133.5157 Inf

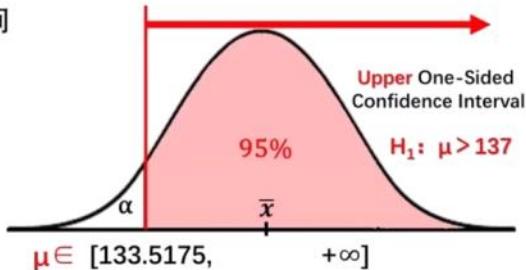
R Documentation:

conf.int

a confidence interval for the mean appropriate to
the specified alternative hypothesis.

那单边置信区间如何理解呢? 我们回到第2个问题中。原假设是 $H_0: \mu < \mu_0$ 。备用假设是 $H_1: \mu > \mu_0$ 。注意, 根据R软件文档, 计算结果中的置信区间, 是" a confidence interval for the mean appropriate to the specified alternative hypothesis", 翻译过来, 就是"这是一个使备用假设 H_1 成立的置信区间"。 H_1 成立的话, 就是说, 王老师班这个样本, 来自一个总体均值 $\mu > 137$ 分的总体。

单边的置信区间



alternative hypothesis: true mean is greater than 137

95 percent confidence interval:

133.5177 Inf

而这个单边置信区间, 就是对这个 μ 的估计。既然 $H_1: \mu > 137$ 分成立, 那么, 只要 μ 大于137分, H_1 都成立。 μ 可以是1000分, 10000分, 甚至是无穷大, H_1 都成立。所以, 结果中置信区间这里, 是 $[133.5175, +\infty]$ 。画出来的话, 是这样的, 是一个upper one-sided confidence interval。这就是对本例中, 单边t检验的, 单边置信区间的解。再说一遍, 不要求掌握。

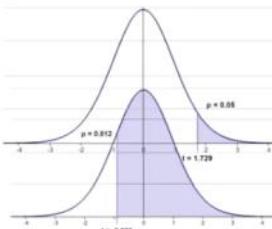
2. 王老师的班级均分是否显著低于全校总体均分? $H_0: \mu < \mu_0$

(H_0 : 王老师班这个样本, 来自均分 " $\mu < 137$ 分" 的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class, mu=137, alternative="greater")
```

One Sample t-test

```
data: wang_class  
t = -0.90834, df = 19, p-value = 0.8125  
alternative hypothesis: true mean is greater than 137  
95 percent confidence interval:  
133.5157 Inf  
sample estimates:  
mean of x  
135.8
```



现在, 我们再回过头来看一下第2个问题本身。这是一个拒绝域在单边右尾的单样本t检验。样本均值是135.8分, 是明显小于 $\mu_0=137$ 分的。而原假设又是, 样本所代表的总体均值 $\mu < 137$ 分。这真所谓是, 明明看起来是对的, 还要去假设它是对的。所以, 原假设肯定大概率是成立的。p值很大, 等于0.8125, 也说明了 H_0 是大概率成立的。p=0.8125, 说明抽样抽到135.8分, 一点都不极端, 一点也不显著。所以, 接受 H_0 。至此, 第2个问题回答完毕。

3. 王老师的班级均分是否显著高于全校总体均分?

3. 王老师的班级均分是否显著高于全校总体均分？

(H_0 : 王老师班这个样本，来自均分“ $\mu > 137$ 分”的一个总体。)

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=137,alternative="less")
```

One Sample t-test

$H_1: \mu < \mu_0$

```
data: wang_class  
t = -0.90834, df = 19, p-value = 0.1875  
alternative hypothesis: true mean is less than 137  
95 percent confidence interval:  
-Inf 138.0843  
sample estimates:  
mean of x  
135.8
```



下面，我们看第3个问题。原假设 $\mu > \mu_0$ 。所有代码几乎是一样的，只不过把alternative参数改成“less”，意思是备用假设是， $H_1: \mu < \mu_0$ ，等价于 $H_0: \mu > \mu_0$ 。回车，算出结果。

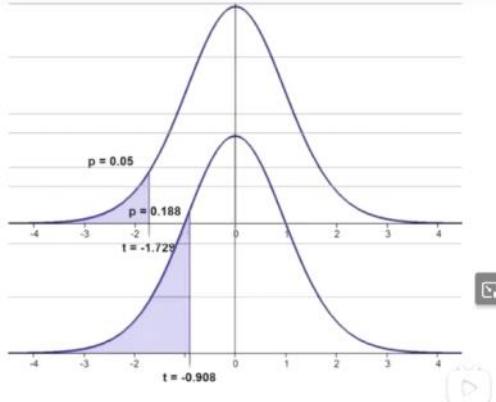
```
> t.test(wang_class,mu=137,alternative="less")  
t = -0.90834, df = 19, p-value = 0.1875
```

$H_0: \mu > \mu_0$

拒绝域在左尾

我们在t分布中，做出拒绝域和p值的示意图。 H_0 为 $\mu > \mu_0$ ，所以，拒绝域在单边左尾。上面这个图，是自由度等于19，单边左边 α 为0.05的拒绝域。

下面这个图，是根据样本算出来的 $t=-0.90834$ ，所对应的p值，即，从 $t=-0.90834$ ，划一条线，到极端方向，也就是左尾的，曲线下的阴影部分面积，为0.1875。



```
> t.test(wang_class,mu=137,alternative="less")  
t = -0.90834, df = 19, p-value = 0.1875
```

$H_0: \mu > \mu_0$

拒绝域在左尾

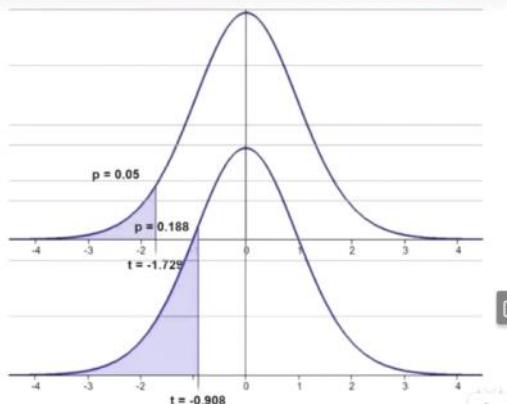
假如根据临界值来判断，样本算出来的 $t=-0.90834$ ，不如临界值-1.729极端，所以，**不拒绝 H_0** 。

假如根据拒绝域来判断，则 $t=-0.90834$ ，没有落入拒绝域，落入了接受域，所以，**不拒绝 H_0** 。

假如根据p值来判断， $p=0.1875$ ，显著水平 $\alpha=0.05$ ，则 $p > \alpha$ ，不极端，**不显著**，所以，**不拒绝 H_0** 。

不管哪种方式，得出的结论都是“**接受**” $H_0: \mu > \mu_0$ ，即，王老师班这个样本代表的总体均分，“**显著高于**”137分。

第3个问题，回答完毕。



原假设	$H_0: \mu = 137$	$H_0: \mu < 137$	$H_0: \mu > 137$
王老师班来自什么样的一个总体	$\mu = 137$ 分的总体	$\mu < 137$ 分的总体	$\mu > 137$ 分的总体
t值	-0.90834	-0.90834	-0.90834
p值	0.3751	0.8125	0.1875
拒绝域位置	双边	单边右边	单边左边
拒绝域 p值 示意图			

现在，我们把3个问题的原假设和结果，放在一起比较。第一，可以发现，t值都是一样的，这是必然的，因为t值，都是通过同样的样本，即王老师班20个同学的分数算出来的。

原假设	$H_0: \mu = 137$	$H_0: \mu < 137$	$H_0: \mu > 137$
王老师班来自什么样的一个总体	$\mu = 137$ 分的总体	$\mu < 137$ 分的总体	$\mu > 137$ 分的总体
t值	-0.90834	-0.90834	-0.90834
p值	0.3751	0.8125	0.1875
拒绝域位置	双边	单边右边	单边左边
拒绝域 p值 示意图			

为什么会出现这种情况呢？一句话，**程序是死的，人脑是活的**。按照正常人的脑回路，你是**不会同时进行三种假设的**。例如，你要觉得王老师班和全校均分137分没啥差别，你就会进行一个双边检验。双边检验的结果是p=0.3751，就是确实没啥差别。既然确定没啥差别了，你为啥还要去检验 $\mu < 137$ 和 $\mu > 137$ 这两个假设呢？没必要的。

2. 王老师的班级均分是否显著低于全校总体均分？ $H_0: \mu < \mu_0$ p=0.8125

(H_0 : 王老师班这个样本，来自均分“ $\mu < 137$ 分”的一个总体。)

接受 H_0 ，则认为：“王老师班的这个样本代表的总体均分 μ ，显著低于137分”。

3. 王老师的班级均分是否显著高于全校总体均分？ $H_0: \mu > \mu_0$ p=0.1875

(H_0 : 王老师班这个样本，来自均分“ $\mu > 137$ 分”的一个总体。)

接受 H_0 ，则认为：“王老师班的这个样本代表的总体均分 μ ，显著高于137分”。

如此看来，之前第2和第3个问题，我们得出的结论，不管是“显著低于”还是“显著高于”的说法，都是错误的，或者至少说，是“非常不严谨”的。所以，写原假设的时候，不出现“显著”这个字眼，因为这时候p值还没算出来，你怎么知道显著不显著？特别是，p值算出来明明不显著的时候，在结论中，更不能使用“显著”这个字眼。

某高中，王老师班，20名同学英语成绩：

($\bar{x} = 135.8$)

136, 136, 134, 136, 131, 133, 142, 146, 137, 140, 134, 135, 136, 132, 119, 132, 145, 131, 140, 141

全校均分

$\mu_0 = 137 - 140$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \begin{aligned} 135.8 - 137 &= -1.2 \\ 135.8 - 140 &= -4.2 \end{aligned}$$



H_0 : 王老师班这个样本，来自均分 $\mu > 140$ 分的一个总体。

H_1 : 王老师班这个样本，来自均分 $\mu < 140$ 分的一个总体。

当然，这次，我们也不要傻傻的提出3个原假设了。我们通过人脑和肉眼观察到，王老师班的 $\bar{x} = 135.8$ 分，显然比 $\mu_0 = 140$ 分小。所以，我们想证明 H_1 : 王老师班这个样本，来自均分 $\mu < 140$ 分的一个总体。但我们要写出 H_0 : 王老师班这个样本，来自均分 $\mu > 140$ 分的一个总体。于是，这是一个拒绝域在左尾的单边t检验。特别注意，假设都用“显著”这个词。

53

```
> wang_class=c(136,136,134,136,131,133,142,146,137,140,  
+ 134,135,136,132,119,132,145,131,140,141)  
> t.test(wang_class,mu=140,alternative="less")  
      H0: μ > 140  
      H1: μ < 140  
One Sample t-test
```

```
data: wang_class  
t = -3.1792, df = 19, p-value = 0.00247  
alternative hypothesis: true mean is less than 140  
95 percent confidence interval:  
-Inf 138.0843  
sample estimates:  
mean of x  
135.8
```



我们把数据输入到R软件中。wang_class的数据不变， μ 变成140分，备用假设 H_1 ，alternative这里赋值“less”。回车看结果。果不其然，t值变大，p值变小。

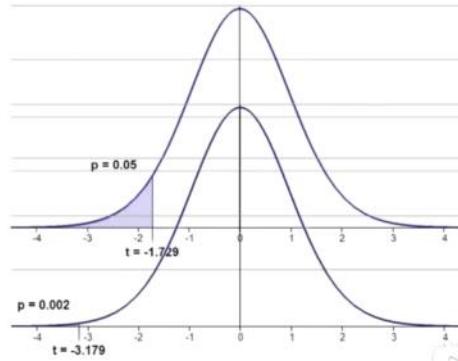
```
> t.test(wang_class,mu=140,alternative="less")  
      H0: μ > μ0  
      t = -3.1792, df = 19, p-value = 0.00247  
      alternative hypothesis: true mean is less than 140  
      拒绝域在左尾
```

我们仍然把结果画到t分布曲线中，进行判断。

第一，抽样算出的 $t = -3.1792$ ，比临界值 $t = -1.729$ 还要极端，于是，拒绝 H_0 ；第二，抽样算出的t值，落入了拒绝域，于是，拒绝 H_0 ；第三， $p = 0.00247$ ， $p < \alpha$ ，结果很显著，于是，拒绝 H_0 。

从这三个判断方法，都“显著的”拒绝了 H_0 ，于是接受 H_1 ，即，这里写的，true mean is less than 140，意思是，王老师班代表的总体的均值 $\mu < 140$ 分。

这时，就可以说，王老师班的均分，显著低于全校均分140分了。



55

做假设的3步

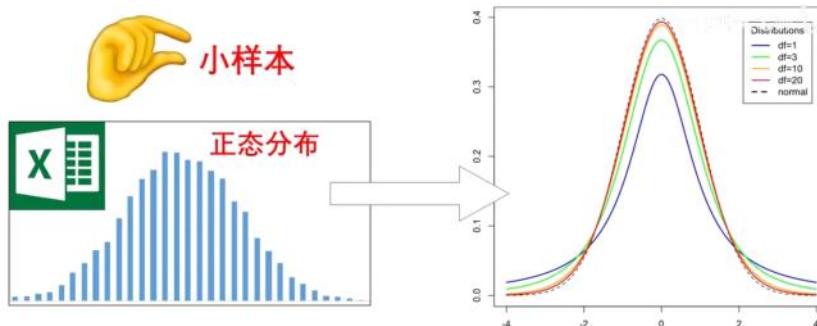
2024年1月2日 17:25

我们想证明a群体小于总体B

- 1.做出相反的假设, $a > B$
- 2.设想满足 $a > B$ 时的拒绝域, 即 $a < B$ 的极端情况
- 3.想办法证明a落入拒绝域

小样本t检验

2024年1月2日 18:38



既然小样本t检验这么有用，那怎么才能保证小样本能够进行t检验呢？那就是，确保小样本来自一个正态分布的总体，换句话说，就是要保证小样本的正态性。所以，小样本研究的第一步，就是检验样本的正态性。

讲到这里

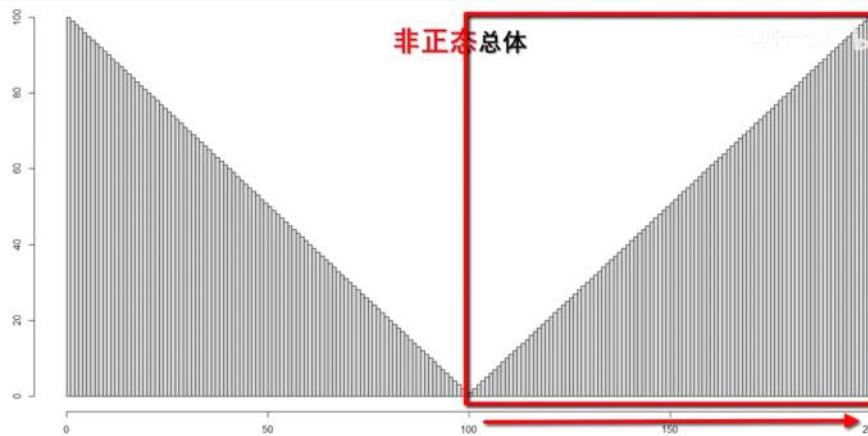
提问人：你真聪明

1. 多小的样本才算小？多大的样本才算大？

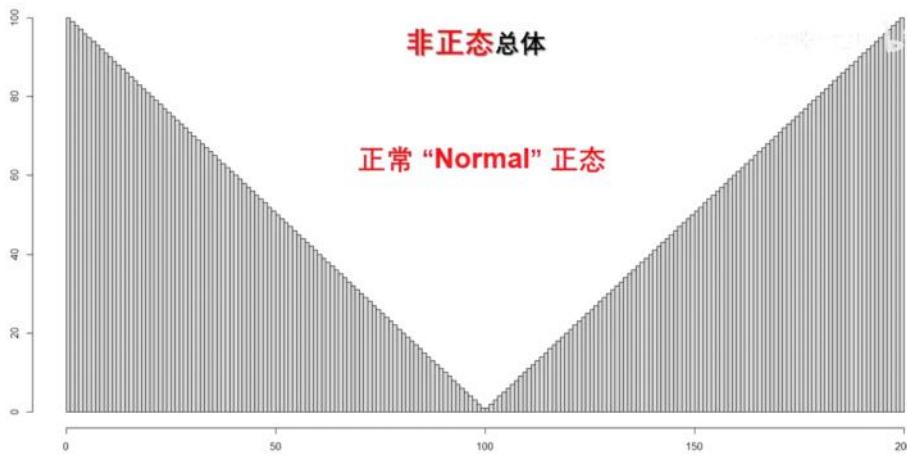


2. 大样本需不需要检验正态性？

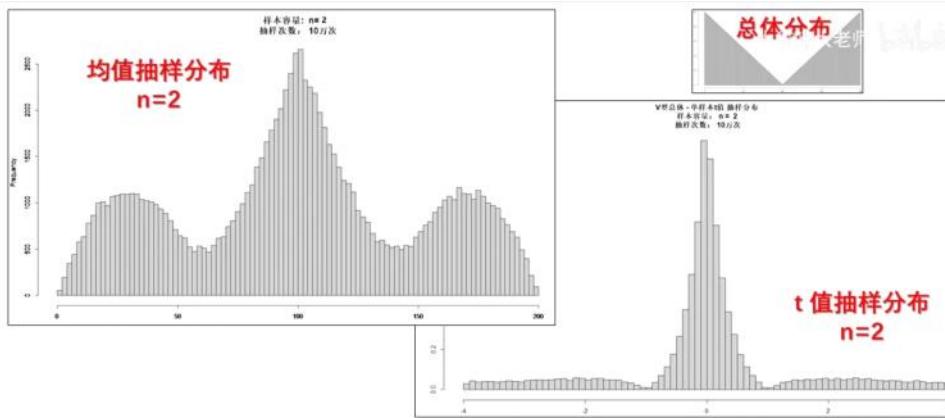
讲到这里，大家自然又会问到两个问题：第一个问题：多小的样本才算小？多大的样本才算大？第二个问题：既然小样本需要检验正态性，那么，大样本需不需要检验正态性呢？下面，我们通过抽样仿真程序，来回答这两个问题。



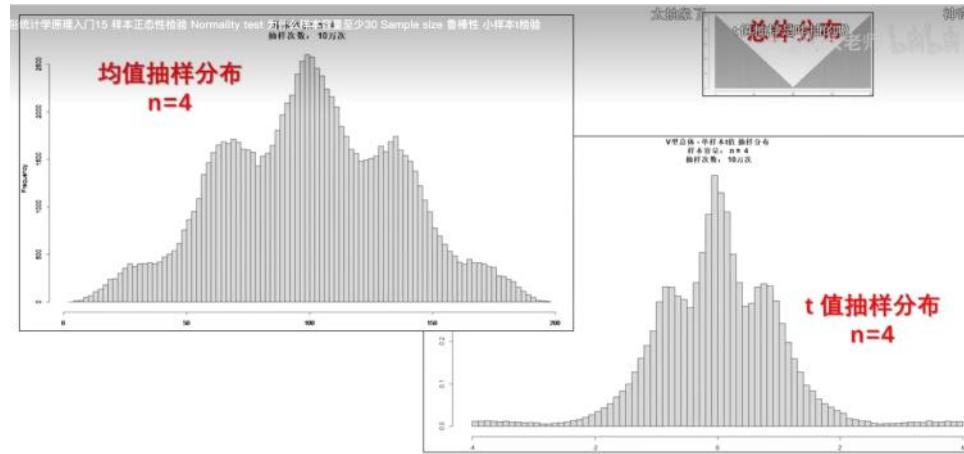
首先，我们来人为构造一个不服从正态分布的总体。例如，我们在R软件中，构造一个V字形的总体分布。这个总体的分布中，分两个部分，左半边从左往右是：100个1, 99个2, 98个3, 97个4, 以此类推，一直到3个97, 2个99, 1个100；右半边从左往右是：1个101, 2个102, 3个103, 以此类推，一直到98个198, 99个199, 100个200。



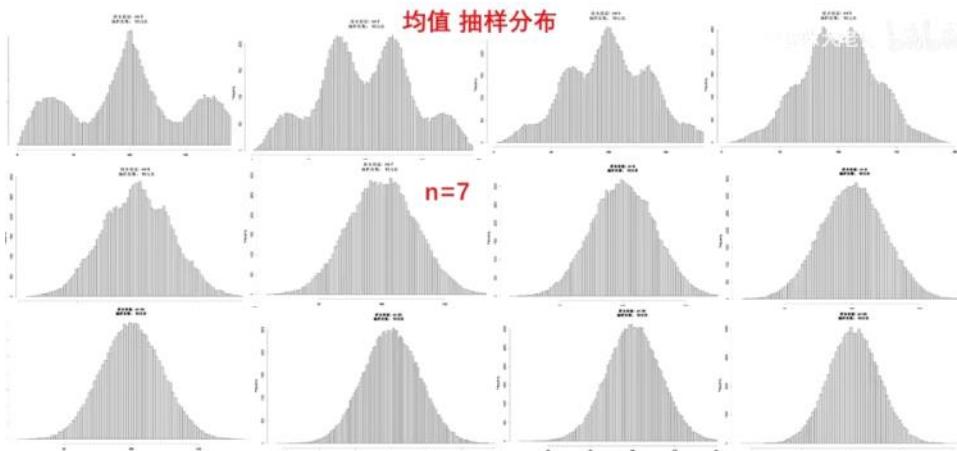
是，这个总体分布呈现V字形，这是非常奇异的一种总体分布，所以，不是一种正常的分布。“正常”的英语就是normal，“正态”的英语，也是normal。所以，我们可以不严格的理解为，不正常的分布，就不是正态分布。现在，我们通过R程序脚本，对这个V字形总体，进行从n=2到n=50的，各个样本容量的均值抽样分布和t值抽样分布。



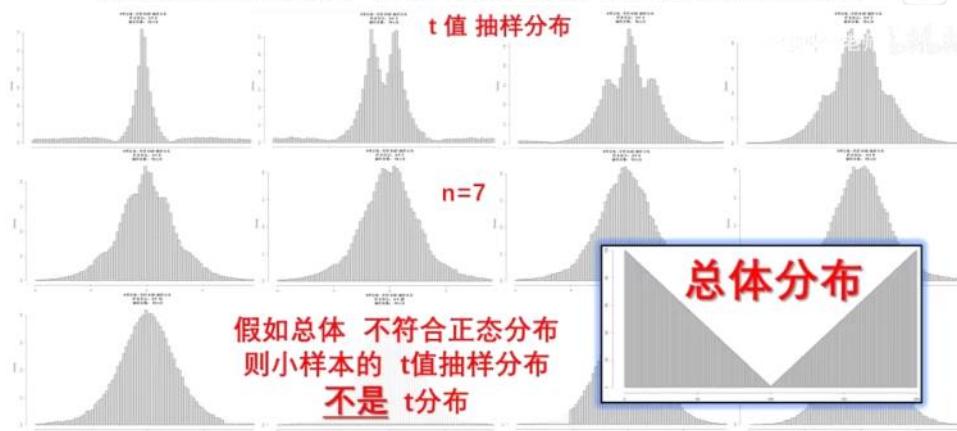
这是对V型总体进行的，样本容量n=2，抽样10万次的均值抽样分布和t值抽样分布。这个t值抽样分布，显然和真正的、铃铛形状的t分布完全不像。



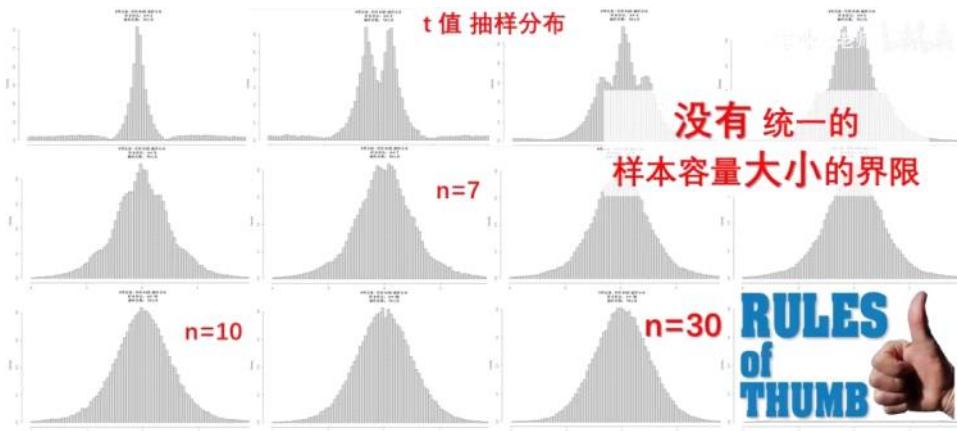
n=4，也不像。



我们把得到的均值抽样分布放到一起比较，发现样本容量从n=7开始，就慢慢接近铃铛形状了。

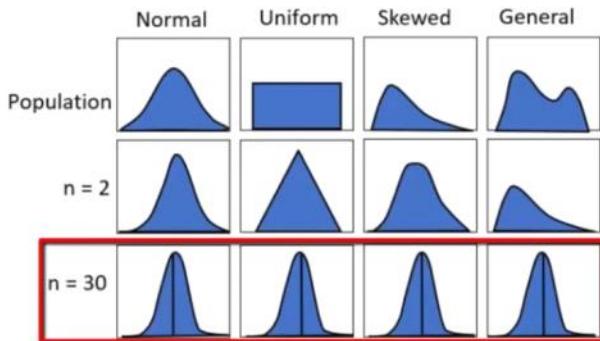


到这里，我们就通过实验演示了，“假如总体不符合正态分布，则小样本的t值抽样分布，根本就不是t分布”。



所以，没有万能的、严格的、统一的样本容量是大是小的界限。但是，按照英语里讲的“rule of thumb”，“拇指法则”或“经验法则”的话，一般取n=30作为界限，样本容量大于30左右，就算作大样本了。通过实验证实，就算V型的这么奇葩的总体分布，n > 30时，t值抽样分布就是非常漂亮的铃铛形状了，就和实际的t分布一致了。

中心极限定理：对于**任意分布**的总体，只要样本容量**足够大**，
均值抽样分布就接近于**正态分布**。



根据中心极限定理，这是必然的。中心极限定理说，对于任意分布的总体，只要样本容量足够大，均值抽样分布就接近于正态分布。多大才算足够大呢，对于任意形状的总体分布，一般认为n=30就足够大了。例如，这个图中，有四种类型的总体分布，样本容量n=2时，均值抽样分布各不相同，都不像铃铛；但当n=30时，就能保证是铃铛形状了。

没有统一的
样本容量**大小的界限**

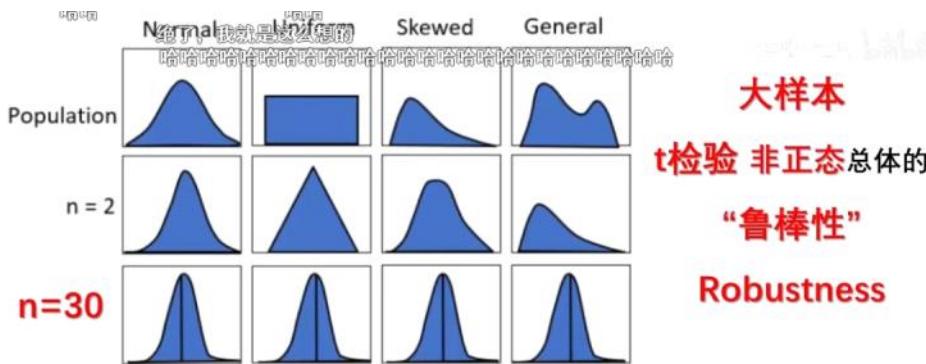
中心极限定理：对于**任意分布**的总体，只要样本容量**足够大**，
均值抽样分布就接近于**正态分布**。

所以，我们就回答了“多小的样本才算小？多大的样本才算大？”这个问题，同时也顺便回答了，为什么样本容量一般至少大于30。这都要拜这个金光闪闪、万能的中心极限定理所赐。

..... 27, 28, 29, **30**, 31, 32



这时，肯定会有同学问，那29算不算大样本呢，28呢，27呢？那样本容量是要正好等于30呢，还是要起码等于31才可以？对于这些问题，我不回答，也不纠结，谁爱纠结谁自己去纠结吧。哈哈。



大样本

t检验 非正态总体的

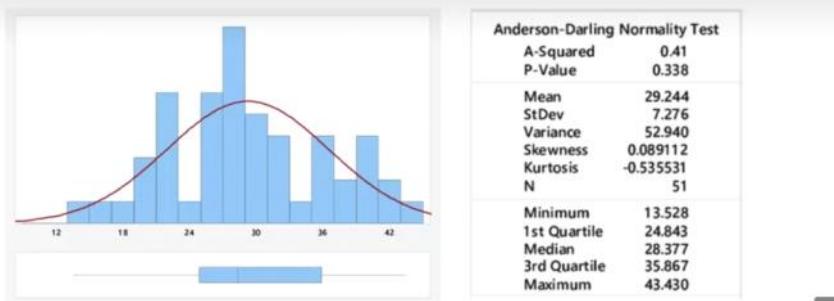
“鲁棒性”

Robustness

实验表明，样本容量大于30时，t分布已不再要求样本的正态性，可以直接进行t检验。这个就叫做大样本容量时，t检验对于非正态性总体的“鲁棒性”，英语叫robustness。个人觉得翻译得很好。Robust，本身是“强壮、稳固”的意思，当然，它还有个意思叫“乐百氏”，但听起来都不够学术。而“鲁棒”就很好，只要n > 30，t检验就很鲁莽，同时又很棒，不管什么奇形怪状的总体分布，t检验都通吃，都能得到稳定的铃铛。

到底要不要 检验 大样本 的正态性？

本课程建议：要



统计软件中的 正态性检验

那么，本课程的回答是：要。原因有二。第一、小样本，大样本，反正都是输到程序里去检验，又不用你自己算。多检验一下，又不花你钱，干嘛这么舍不得呢？万一查出样本有异常，说明总体也可能异常，这本身不也是研究的一部分吗？

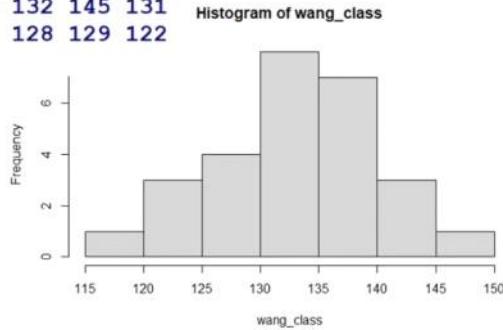
> wang_class

```
[1] 136 136 134 136 131 133 142 146 137
[10] 140 134 135 136 132 119 132 145 131
[19] 140 141 123 122 126 127 128 129 122
```

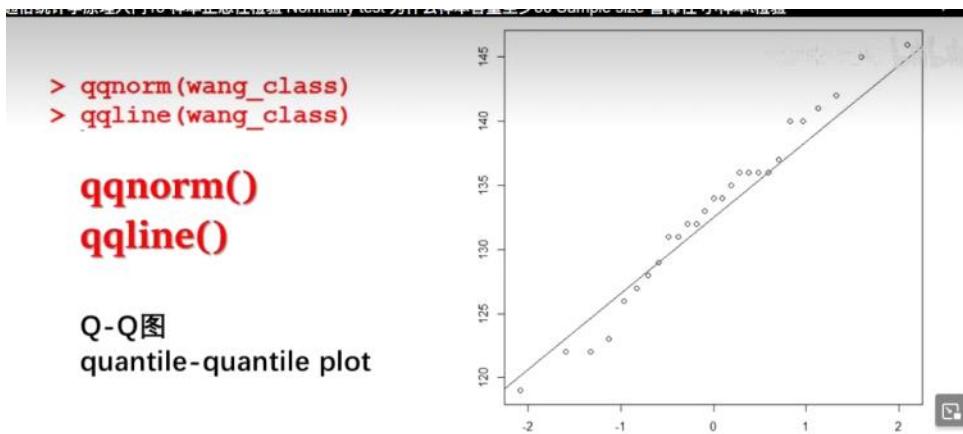
> hist(wang_class)

hist()

直方图 histogram



首先，样本拿来，你起码得用肉眼先瞄一下吧。怎么瞄呢，最基本的，就是R软件里的hist() (histogram) 直方图命令。画出直方图后，观察一下样本的分布是不是大体正常。所谓大体正常，就是中不溜的数值最多，在中间，偏大偏小的，极端的，在两边，也就算是正态了。因为世界上太多的現象都是符合正态分布的，所以，一般来说，样本比较容易满足正态性。例如，这是上节课王老师班的分数，其直方图是这样的，看起来大体上是正态的。



还有高级一点的做图方法，叫Q-Q图 (quantile-quantile plot)，即“分位点-分位点的图”。我们用 `qqnorm()` 和 `qqline()`，做出王老师班的Q-Q图，假如样本中所有的数据点，都落在直线附近的话，就可以认为样本大致符合正态性。Q-Q图的原理和使用，入门课程中也不详细展开了。

常用正态性检验举例

Kolmogorov-Smirnov Test

Shapiro-Wilk Test

Anderson-Darling Test

H_0 : 样本所来自的总体**符合**正态分布。 $p > \alpha$

H_1 : 样本所来自的总体**不符合**正态分布。 $p < \alpha$

需要大家知道的是，通常，样本正态性检验的原假设都是， H_0 : 样本所来自的总体符合正态分布。所以，只要p值大于设定的显著水平，就可以认为样本符合正态性。若p值小于显著水平，则认为 H_1 : 样本所来自的总体**不符合**正态分布。

```

> wang_class
[1] 136 136 134 136 131 133 142 146 137 140 134 135 136 132 119
[16] 132 145 131 140 141 123 122 126 127 128 129 122

> shapiro.test(wang_class)

Shapiro-Wilk normality test

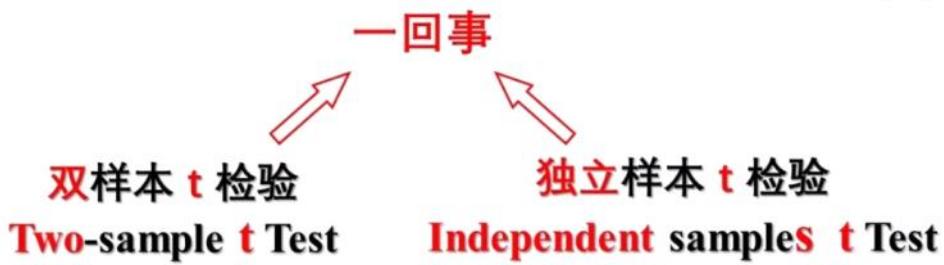
data: wang_class
W = 0.97813, p-value = 0.8178

```

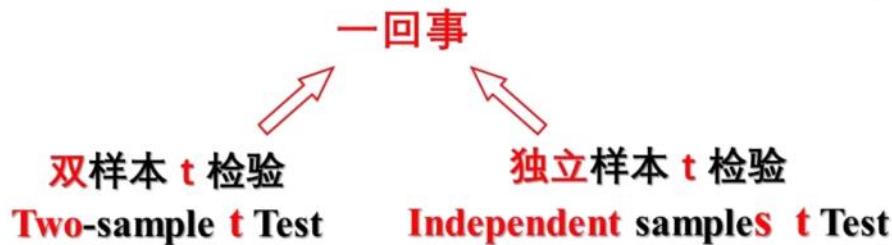
限于入门课的篇幅，我们只举一个Shapiro-Wilk检验的例子。例如，在R软件里，用`shapiro.test(wang_class)`命令，来检验王老师班级分数的正态性。得 $p=0.8178$ 。无论 $\alpha=0.05$ 也好，还是 $\alpha=0.01$ 也好， p 都远远大于 α 。所以，认为样本符合正态性。

双样本t检验

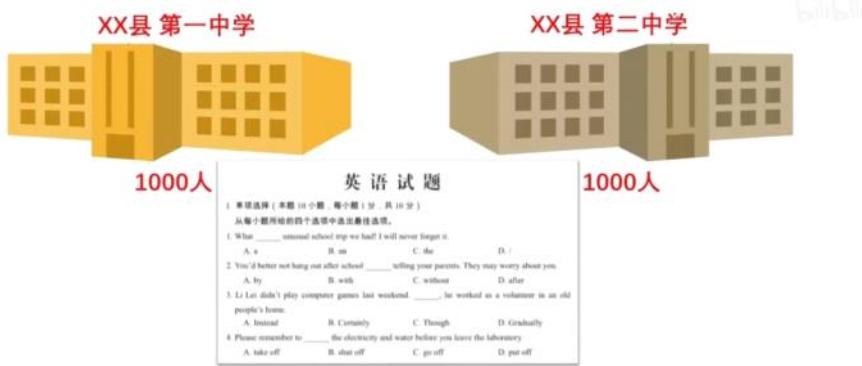
2024年1月2日 18:56



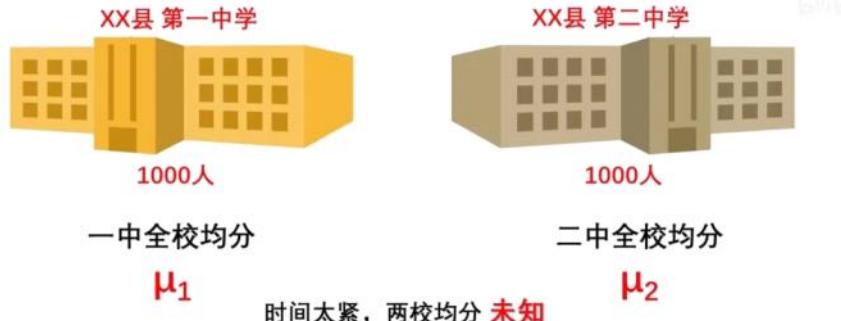
大家好，今天我们来通俗讲解一下双样本t检验。首先，重要的事情先告诉大家。双样本t检验，英语叫two-sample t test，还有一种叫法，叫做独立样本t检验，英语叫independent samples t test。我已经在大家的留言中发现，可能有的同学把这两种叫法当成两回事了。所以，本节课“开宗明义”，告诉大家，双样本t检验，和独立样本t检验，是一回事。



之所以出现这种误解，只能说是术语从英语翻译成汉语时，没有翻译好。“双”和“独立”这两个词，字面上明然是相反的意思，所以很难当成一回事。双样本，大家显然知道是俩样本；“独立”样本，大家就很可能以为是单个样本了。然而，假如从英语来看，two-sample t test 和independent samples t test，就不会产生误解了。



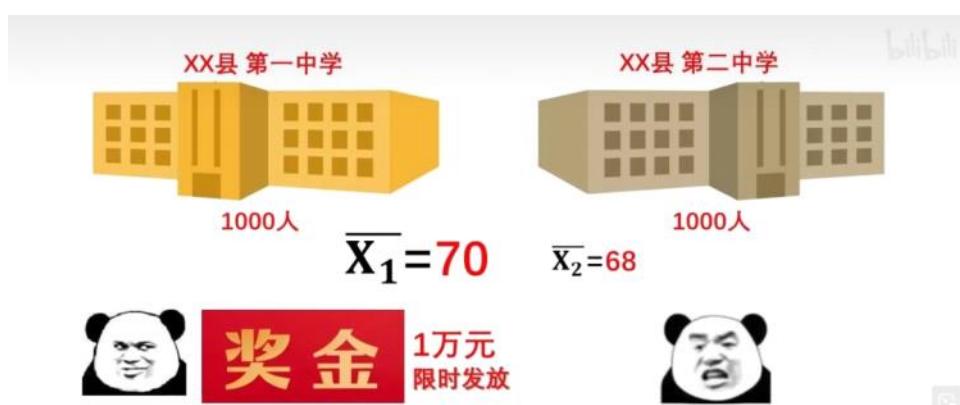
好，我们开始编故事。某县城有两个中学，一中和二中。一中有1000个学生，二中也有1000个学生。现在，县教育局想比较一下两个中学的英语水平谁高谁低，于是组织了一次英语统考，一中和二中的所有学生，考同一套试卷。



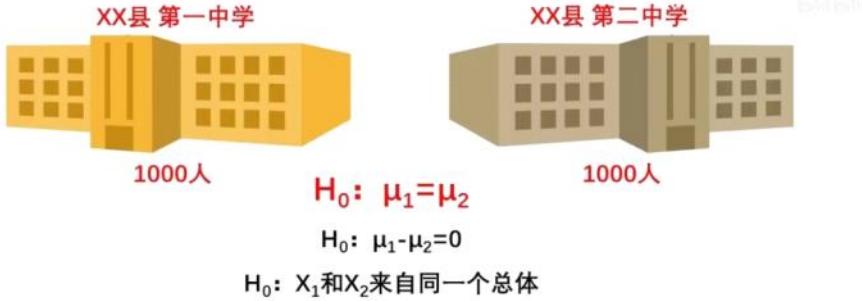
我们把一中的总体平均分记作 μ_1 ，二中的总体平均分记作 μ_2 。一般来说，要比较两个学校的总体水平的话，就是应该比较 μ_1 和 μ_2 谁大谁小。不过，由于各种原因，我们一时半会弄不到两个学校的全部学生的成绩，所以， μ_1 和 μ_2 是未知的。



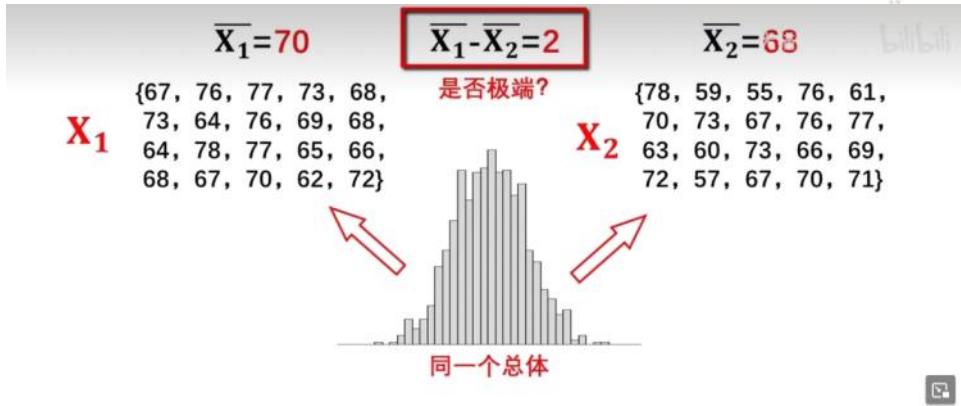
那怎么办呢，我们只能去抽样。比方说，两个学校放学的时候，我们站到学校大门口，随机逮住20个学生，问他们英语统考考了多少分。这样，我们分别从一中和二中，各得到一个样本容量为20的样本，我们把一中的样本记作 X_1 ，二中的样本记作 X_2 。



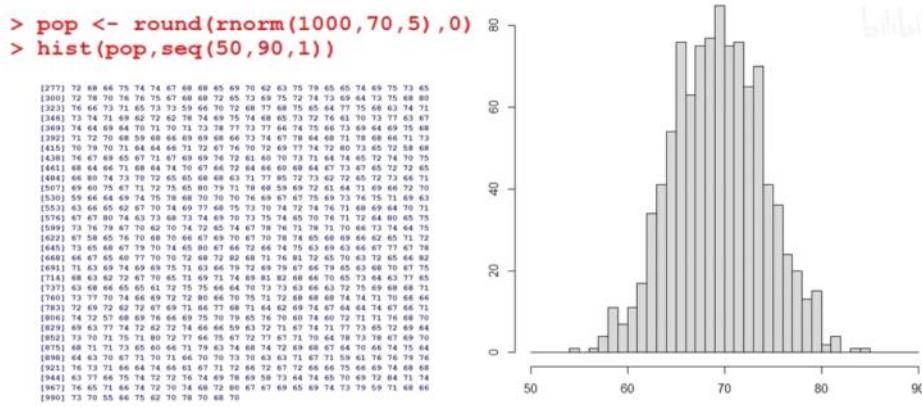
假如我再把这个故事编得更荒唐一点。假如教育局现在有一笔1万元的经费，必须在今天用掉。局长就准备把这1万块钱，奖励给一中和二中之间，英语水平较高的学校的校长。但现在手头只有这两个样本，一中的样本均分显然比二中的样本均分多2分。那假如局长把这1万元奖励给一中校长，那二中校长是否服气呢？



二中校长的说法，有没有道理呢？我们可以通过假设检验和程序仿真抽样来演示一下。我们首先提出原假设， H_0 ：一中和二中的总体均分相等，即 $\mu_1 = \mu_2$ ，或者说 $\mu_1 - \mu_2 = 0$ 。或者更不严格地认为，两个样本来自同一个总体。为什么说“不严谨”，是因为这里姑且认为两个总体的方差是一样的，这样好讲一点。方差不一样的情形，我们留在后面课程中去讲。



所以，假设这两个样本，来自同一个总体的话，那么，这两个样本的均值，相差2分的概率是多大呢？双样本的均值差值为2分，算不算异常呢？算不算极端呢？



我们在R中，用rnorm()命令，制造一个正态分布总体，这个总体有1000个成绩，总体均分为70分，总体的标准差为5分。然后，用hist()命令，画出总体的直方图，如图所示。其对称轴是总体均值70分，总体的高矮胖瘦，由标准差5分和“68-95-99.7法则”决定。

```

样本容量
n=20
sampling_size <- 20
for (i in 1:10000){
  X1 <- smpl_1 <- sample(pop,sampling_size)
  X2 <- smpl_2 <- sample(pop,sampling_size)
  X1_mean <- mean(smpl_1)
  X2_mean <- mean(smpl_2)
  mean_dif <- smpl_1_mean-smpl_2_mean
  statics <- append(statics,mean_dif)
}

```

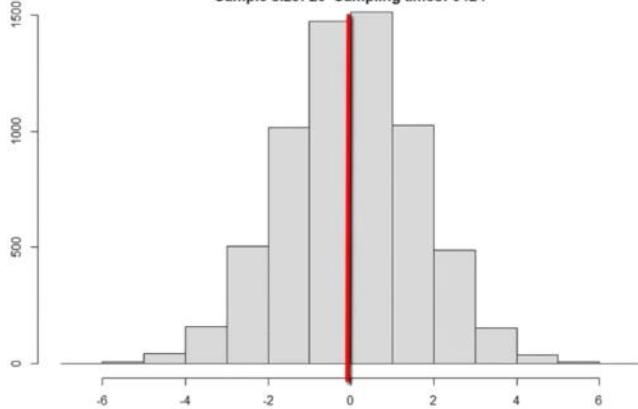
双样本的 均值差值 分布

下面，我们通过程序来模拟抽样。每次抽两个样本，第1个样本记作 X_1 ，包含20个成绩，其平均分记作 \bar{X}_1 。第2个样本记作 X_2 ，也包含20个成绩，其平均分记作 \bar{X}_2 。每次抽出两个样本后，算出样本均值的差值 $\bar{X}_1 - \bar{X}_2$ ，记作 $mean_dif$ 。我们抽1万次，每次计算一个差值，累计在直方图中，最后就可以做出双样本的均值差值的分布图。上面的R代码仅供演示，不作要求。

```

sample_1 mean: 70.35
sample_2 mean: 68.75
mean difference: 1.6
Sample size: 20 Sampling times: 6424

```



好，程序开始。

可以看出，随着抽样次数的增加，差值的分布，最后仍然呈现为熟悉的铃铛形状。而且，对称轴是0分。

这也是可以直观解释的，本来就是从同一个总体中抽出的两个样本，它们均值不会差太多。那均值的差值，就不会和0差太多。

```

sample_1 mean: 70.7
sample_2 mean: 67.6
mean difference: 3.1
Sample size: 20 Sampling times: 10000

```

```

> sum(statics>=2|statics<=-2)
[1] 2058
p=0.2058

```

$\alpha = \text{双边} 0.05$

$H_0: \mu_1 = \mu_2$

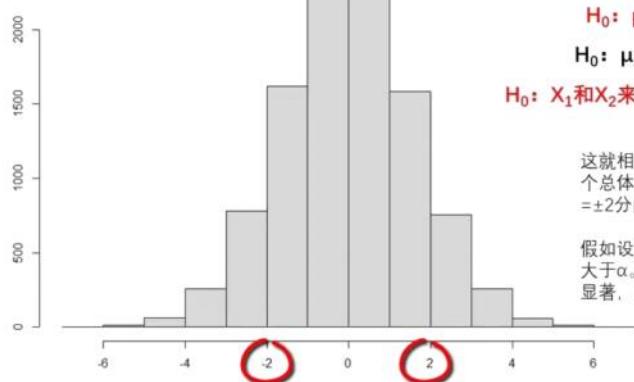
$p > \alpha$

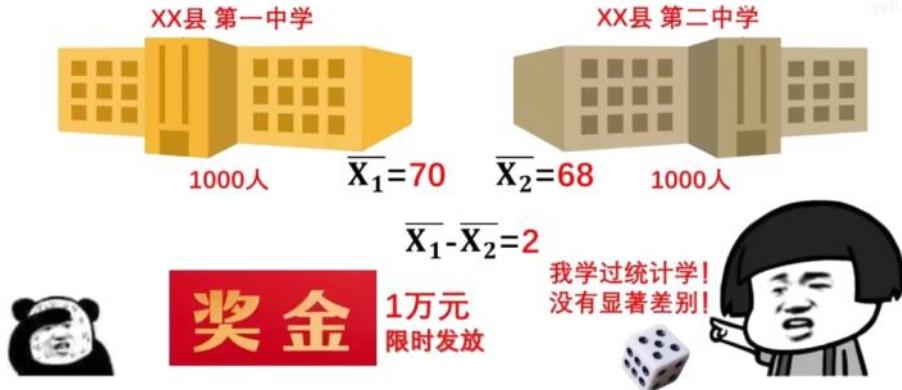
$H_0: \mu_1 - \mu_2 = 0$

$H_0: X_1 \text{ 和 } X_2 \text{ 来自同一个总体}$

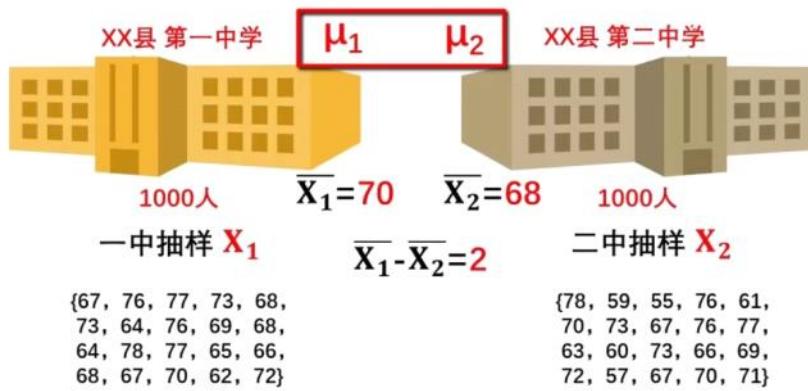
这就相当于说，在 H_0 ：两个样本来自同一个总体的原假设下，抽到样本均值差值 $= \pm 2$ 分的 p 值，等于 $2058/10000 = 0.2058$ 。

假如设定显著水平 α 为双边 0.05 ，那么， p 大于 α 。说明 ± 2 分这个差值，不极端，不显著，因此，**不能拒绝 H_0** 。

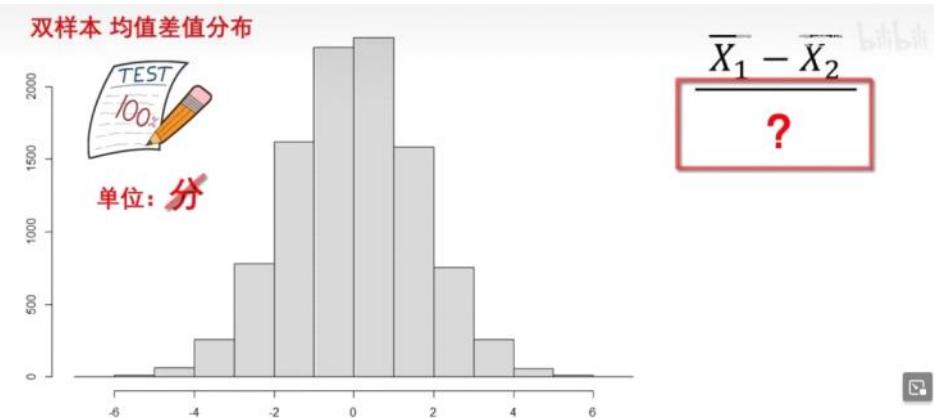




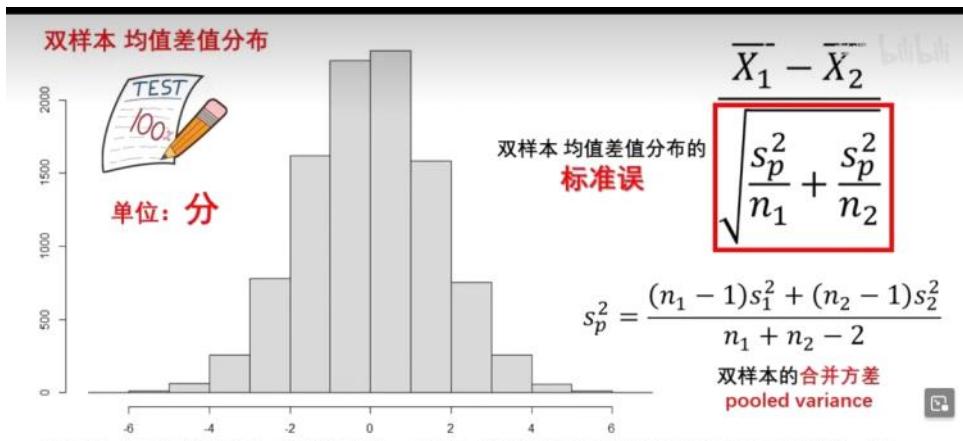
也就是说，二中校长说的没错，在显著水平 $\alpha=0.05$ 的情况下，一中和二中，各抽出一个样本，样本均值的差值等于2分的话，不能说明两个学校的总体均分有显著差别。以上，就是双样本的均值差值的抽样分布，和对应的假设检验。



现在，我们再从头捋一遍，当初为什么要进行双样本的均值差值检验。是因为，要比较两个中学的总体均分 μ_1 和 μ_2 有无显著差别，但暂时没条件计算出真实的 μ_1 和 μ_2 ，而只能从两个总体中分别抽取一个样本 X_1 和 X_2 ，然后计算两个样本的均值的差值。但双样本的均值存在差别，就能说明 μ_1 和 μ_2 存在显著差别吗？



类似的，在双样本的均值差值分布中，我们发现对称轴已经在0了，所以，只需要再做一个除法，把单位约掉就可以了。那么，除以一个什么东西呢？



在双样本的均值差值分布中，我们除以这么一个东西。这个东西叫做双样本的均值差值分布的标准误。其中， n_1 代表 X_1 的样本容量， n_2 代表 X_2 的样本容量。 s_p^2 ，叫做 X_1 和 X_2 这两个样本的“**合并方差**”，英语叫做 pooled variance。

合并方差，用于描述**双样本作为一个整体的、内部的数值离散程度**。关于双样本的这种标准误，入门课程中需要大家掌握。大家只需要知道，它是一个带有单位的量就可以了。

29

双样本t检验公式

X_1 {67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72} $n_1=20$

X_2 {78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71} $n_2=20$

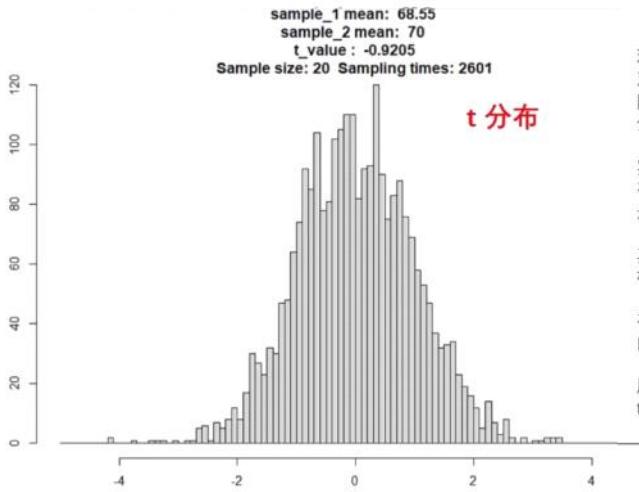
双样本 均值差值分布的
标准误

$$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$\text{双样本的容量都是 } n=20 \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2} / \sqrt{n}}$$



当双样本各自的样本容量相同时，也就是 $n_1=n_2$ 时，公式可以进一步简化。例如，本例中，双样本的样本容量都为 $n=20$ ，那么，公式简化如下。这就是双样本的样本容量都为 n 时的， t 值公式。这个 t 值，就是一个没有单位的，脱离了任何具体案例的，纯数学的一个值。



现在，我们仍然通过程序，来模拟双样本的 t 值抽样分布。可以看到，随着抽样次数的增加，双样本的 t 值分布，逐步接近了铃铛形状。

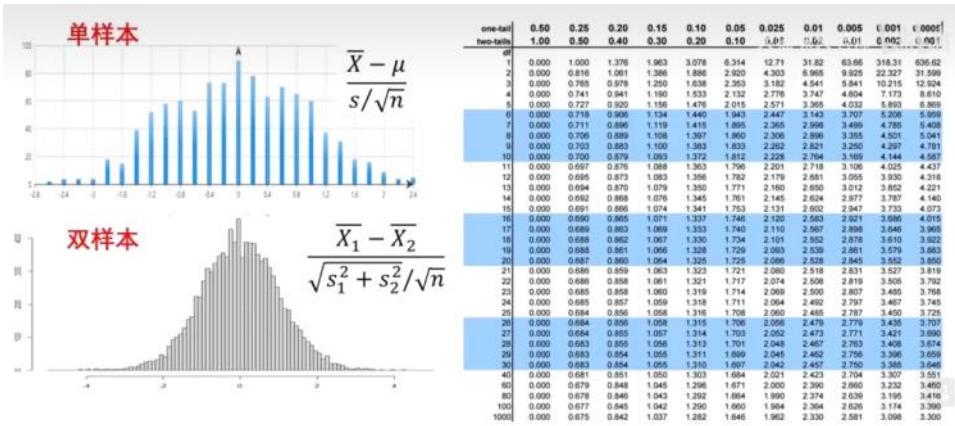
实际上，这就是一个 t 分布。这个 t 分布的**自由度**，是 $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ 。

其中， n_1 和 n_2 ，分别是两个样本的样本容量。

在本例中， $n_1=n_2=20$ ，所以，本例的自由度为 $(20-1)+(20-1)=38$ 。

所以，我们抽样得到了一个 $df=38$ 的 t 分布。

32



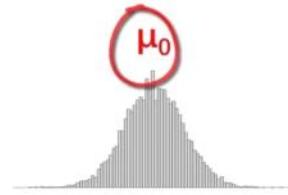
请大家注意，单样本和双样本抽样形成的t分布，是同样的t分布。t临界值表里，并不区分单样本或双样本的。t临界值表里，只区分双尾和单尾，大家不要把概念搞混。

33

单样本t检验

单样本 t检验

检验：单样本 是否来自 已知的总体。



我们简单的对比总结一下。单样本t检验，是比较单个样本均值和一个已知的总体均值，用以检验这个单样本，是否来自这个已知总体均值的总体。

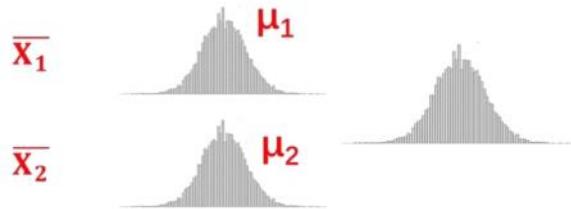
35

例如，我们在第5节《假设检验》中讲到的，已知大二英语成绩的真实均分 $\mu_{大二}$ 。我们想知道大一的总体均分 $\mu_{大一}$ 和 $\mu_{大二}$ 有无差别。但我们手头只有大一的一个样本 $X_{大一}$ ，并不知道 $\mu_{大一}$ 是多少。那么，我们就用单样本t检验，先假设 $H_0: \mu_{大一} = \mu_{大二}$ ，或者说，假设 $X_{大一}$ 这个样本，来自和大二一样的总体。然后，拿 $X_{大一}$ 和 $\mu_{大二}$ ，算出一个 t 值和p值，看看是否拒绝 H_0 。

36

双样本 t 检验

检验：两个样本是否来自同一个总体。



双样本t检验，是比较两个样本的均值，用以检验这两个样本所分别代表的总体均值是否相等，或者说，检验这两个样本，是否来自同一个总体。



37



{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72}

μ_1 未知
 X_1 已知

目标：比较 μ_1 和 μ_2 谁大谁小

$H_0: \mu_1 = \mu_2$



{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71}

μ_2 未知
 X_2 已知

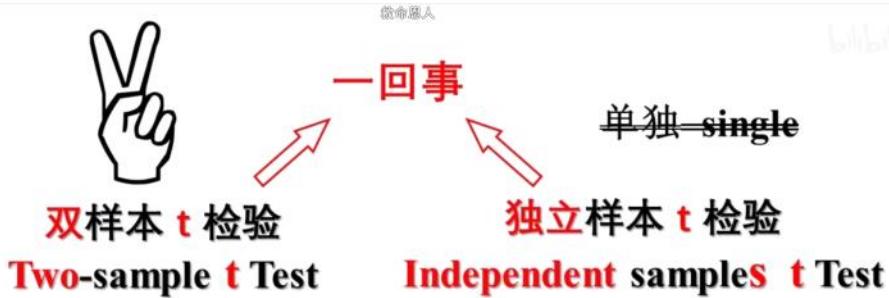
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2 / \sqrt{n}}}$$

例如，在本节课的例子中，两个中学的总体均值 μ_1 和 μ_2 都是未知的，已知的只是从两个总体中抽出来的两个样本 X_1 和 X_2 。我们想知道 μ_1 和 μ_2 有无显著差别，或者说，我们想知道 X_1 和 X_2 ，是否来自同一个总体。我们就用双样本t检验。先假设 $H_0: \mu_1 = \mu_2$ ，或者说，假设 X_1 和 X_2 来自同一个总体。然后，拿 X_1 和 X_2 ，算一个t值和p值，看看是否拒绝 H_0 。

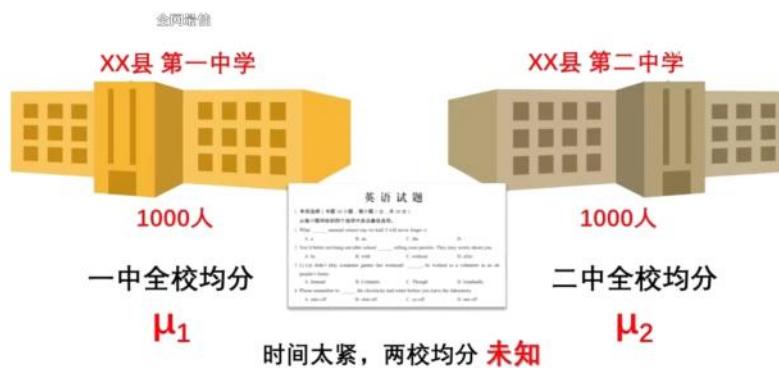
38

独立样本t检验

2024年1月3日 16:52

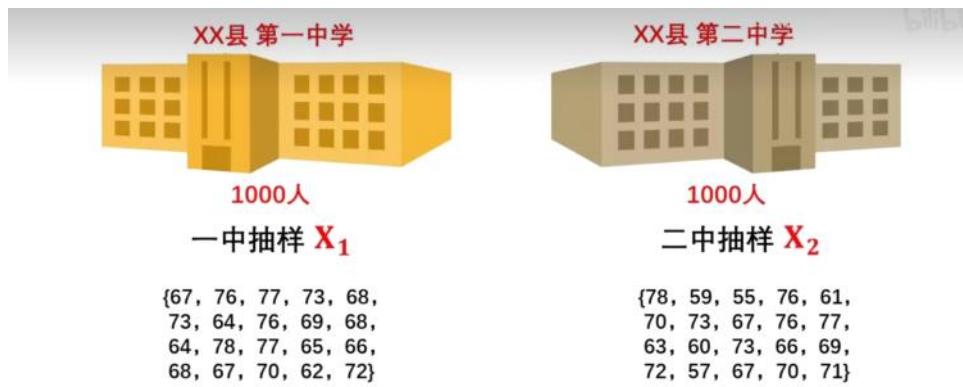


大家好，上一节课中，我们学习了“双样本t检验”和“独立样本t检验”是一回事。独立，independent，不是“单独”，single。所以，独立样本t检验，是两个样本之间的t检验，不能多，也不能少，就是两个。这节课，我们来通俗讲一下样本之间“相互独立”的含义。



我们仍然回到上节课的故事中。一个县城，有两个中学，一中和二中，每个中学都有1000个学生。县教育局想比较一下两个中学的总体英语水平，于是组织了一次统考，一中和二中考同一份试卷。考完之后，由于时间仓促，两个学校各自的总体均分 μ_1 和 μ_2 ，还没有统计出来。

怎么才算互相独立的样本



我们在一中和二中门口，分别随机抽样20个学生，问他们英语考了多少分，于是获得了两个样本， X_1 和 X_2 。注意，随机抽样，是个大学问，别看我们每次“随机”说得这么轻松，但真正做起来，可是需要非常严谨的方法的。本节课只讲独立性，暂时不对“随机抽样”进行展开讲解。



目标：比较 μ_1 和 μ_2 谁大谁小

$$H_0: \mu_1 = \mu_2$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2 / n}}$$

$$t = 1.0526$$

$$p = 0.2992$$

$$\alpha = 0.05$$

我们获得了两个样本：一中样本 X_1 ，样本均分 $\bar{X}_1 = 70$ 分，二中样本 X_2 ，样本均分 $\bar{X}_2 = 68$ 分。样本均分存在差别，但是， μ_1 和 μ_2 是否存在显著差别呢。我们先写出原假设 H_0 ，认为 μ_1 和 μ_2 没有差别，然后通过双样本的 t 值公式，算出一个 p 值， $p > \alpha$ ，于是接受 H_0 ，所以， μ_1 和 μ_2 没有显著差别。

一中抽样 X_1

{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72}

姓名	成绩
张三	67
李四	76
王五	77
马六	73
.....

二中抽样 X_2

{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71}

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
.....

受试 subjects
观测值 observations

现在，我们盯着这两个样本，来看一下，“样本之间相互独立”到底是什么意思。但首先，我们先了解两个术语，“受试”（subjects）和“观测值”（observations）。在本例中，我们抽样抽到的，提供英语成绩的同学，就是受试，是人。同学提供的成绩分数，就是观测值，是数据。有了这两个概念，我们就可以比较容易的理解“相互独立的样本”了（Independent samples）。

Independent Samples

There is **no relationship** between the **subjects** in each sample.

两个样本各自的受试之间，没有任何关系。

1. Subjects in the first sample cannot also be in the second sample.

（两个样本中，第一个样本中的受试，不能同时出现在第二个样本中）

2. No subjects in either sample can influence subjects in the other sample.

（两个样本中，任何一个样本中的受试，不能影响另外一个样本中的受试）

3. No sample can influence the other sample.

（两个样本中，一个样本不能影响另一个样本）

参考资料：<https://libguides.library.kent.edu/SPSS/IndependentTTest>

各种教材中存在多种说法，有简单的，有复杂的，我挑了一个相对好理解的，供大家参考。Independent samples 是指，There is no relationship between the subjects in each sample。翻译一下，两个样本各自的受试之间，没有任何关系。听起来似乎很笼统，具体是指什么呢？具体有三个方面：这三个具体的方面，仍然比较抽象，我们放到一中和二中的故事中来举例子说明。

1. 两个样本中, 第一个样本中的受试, 不能同时出现在第二个样本中

一中样本 X_1

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

全县唯一的
钱二麻

二中样本 X_2

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
钱二麻	66
.....

首先, 来看第一个方面: 第一个样本中的受试, 不能出现在第二个样本中。假如, 我们先在一中门口抽样完毕后, 再去二中门口抽样。在二中抽样时, 怎么发现一个同学, 感觉有点面熟呢, 好像刚才在一中抽过了啊。因为脸上有两颗麻子, 不容易记错。我们一问, 果不其然, 这个同学就是一中的, 名字叫钱二麻, 刚才在一中门口, 已经被抽到过了, 现在人家来二中, 找他女朋友一起去吃饭, 结果就又被抽到了。这就叫, 第一个样本中的受试, 出现在了第二个样本中。

1. 两个样本中, 第一个样本中的受试, 不能同时出现在第二个样本中

一中样本 X_1

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

受试存在重叠
样本之间互不独立

二中样本 X_2

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
钱二麻	66
.....

这时, X_1 和 X_2 这两个样本, 就存在关系了。 X_1 和 X_2 , 就不是互相独立的样本了。实际上, 这就是犯了低级的抽样错误。我们并不是总能碰到钱二麻同学这样不凡的容貌, 好让我们及时发现, 受试重叠了 (overlapping subject)。所以, 抽样的时候, 我们不能只把成绩记下来就拉倒了, 还得把姓名、或学号、或其他能区分受试个体的信息给记下来, 免得两个样本里, 出现重叠的受试, 导致样本之间互不独立。

10

一中抽样 X_1

{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72}

巧

二中抽样 X_2

{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71}

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2}}$$

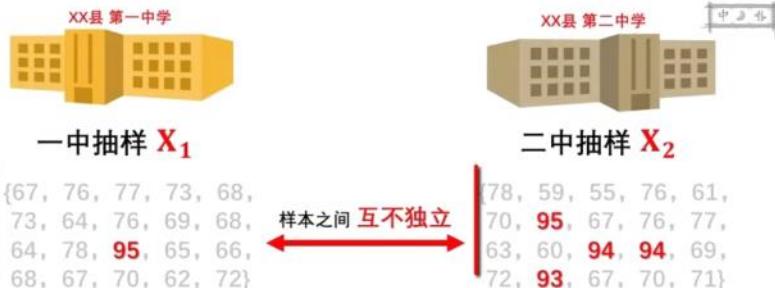
样本容量 $n=20$

重叠无效的观测值

$$1/20 = 5\%$$

你可能会说, 这个例子太扯了吧。怎么会这么巧呢? 事情就是这么巧。概率不就是个“巧”字吗。在本例中, 样本容量本来就很小, $n=20$, 只需要碰上1个巧的, 重叠了, 那可就是1/20, 也就是5%的观测值就无效了。5%的观测值都无效了, 你还能相信算出来的t值和p值吗?

2. 两个样本中，任何一个样本中的受试，不能影响另外一个样本中的受试



那么，在一中和二中的故事中，假如一中也有个天才同学，半小时就做完了试卷，并且通过某种方式，把答案传给了二中的不止一个好朋友。而我们在一中和二中抽样时，恰巧把天才同学和他的好友，都抽到了我们的样本中。那么，虽然两个样本中的受试没有重叠，但 X_2 中多个受试的观测值，受到了 X_1 中这个天才受试的影响。那么， X_1 和 X_2 ，就也不是相互独立的样本。

14

3. 两个样本中，一个样本不能影响另一个样本



下面，看第三个方面：一个样本不能影响另一个样本。注意，这里说的是样本作为整体，而不是样本中的受试个体。假如，有一位王老师，由于某些原因，**同时**在一中和二中担任英语老师。王老师在两个中学上课时，用的是**同一套教材**、放的是**同一套PPT**、布置的是**同样的**课堂练习和课后作业，讲的是**同样的**笑话，等等等等，全都是一样的。

Paired t test

Paired t Test

样本 故意 相互不独立

接下来，再举一个常见的、故意让两个样本相互不独立的例子。这就是大名鼎鼎的“Paired t Test”。

Paired t Test

(学期开始) pretest										
学号	01	02	03	04	17	18	19	20	
前测成绩	65	75	69	85	78	80	69	77	
同一批同学 一个学期后										
(学期结束) posttest										
学号	01	02	03	04	17	18	19	20	
后测成绩	67	76	66	88	81	79	68	80	

例如，一个班级有20名同学，在学期开始时，进行了一次英语摸底考试。然后，学期结束时，又进行了一次考试，试卷和之前摸底考试是同一套试卷。目的就是为了看一看，这20名同学，在经历了一个学期的英语学习后，英语水平，有没有显著提高。学期开始时的考试，叫做前测（pretest），学期结束时的考试，叫做后测（posttest）。

Paired t Test

学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77
后测成绩	67	76	66	88	81	79	68	80
差值	+2	+1	-3	+3	+3	-1	-1	+3

一般来说，在“Paired t Test”中，每个受试，前后两次的成绩，和前后两次成绩的差值，都写在同一列。所以看到这样的数据格式，你就知道了，这是Paired t Test。而Paired t Test的目的，就是比较所有受试的前测和后测的成绩，有没有显著差别。

Paired t Test

“配对 t 检验” “成对 t 检验”

学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77
后测成绩	67	76	66	88	81	79	68	80
差值	+2	+1	-3	+3	+3	-1	-1	+3

“Paired t Test”，常见的翻译叫做“配对t检验”或“成对t检验”。关键词是“Pair”，就是“一对”的意思。每个受试的前测成绩和后测成绩，来自同一个受试，所以，是“一对”，是一个“Pair”。因此，我更愿意把“Paired t Test”叫做“成对t检验”，而不是“配对t检验”。人家本来就是一对，就跟一双鞋、一双袜子一样，是“成对”的。不是人为强行给“配对”的。前测与后测的差值，只能成对的，才能做减法。不成对的，当然不能相减。

Paired t Test

Paired t Test中的，这两行数据，或者说，这两个样本，就不是相互独立的。因为这两个样本太有联系了，它们来自完全相同的20个受试。Paired t Test这种实验设计，本身就是故意从相同的受试中，产生完全关联的、成对的、可以相减的数据，用以检验前测和后测有无显著差别。所以，Paired t Test中的这两个样本，就不是Independent samples，而是Dependent samples，翻译过来，叫做“相互依赖的样本”，或简称“相依样本”。

样本 1
↑
独立样本
相依样本
↓
样本 2

学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77
后测成绩	67	76	66	88	81	79	68	80
差值	+2	+1	-3	+3	+3	-1	-1	+3

Paired t Test, 我们会用一节课来专门讲解。本节课提到它, 只是为了说明, 什么样的样本不是相互独立的。当然, 我们剧透一下, Paired t Test 和 Two-Sample t Test的最大区别是, **自由度不一样**。

Two-Sample t Test

一中样本 X_1
 $n=20$
{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72}

二中样本 X_2
 $n=20$
{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71}

40个受试

$df = (20-1)+(20-1) = 38$

Paired t Test

$n=20$

学号	01	02	19	20
前测成绩	65	75	69	77

$n=20$

学号	01	02	19	20
后测成绩	67	76	68	80

20个受试

$df = 20-1 = 19$

例如, 在一中和二中的Two-Sample t Test中, 假如每个样本的样本容量都是20的话, 两个样本便一共有40个成绩, 来自40个受试, 那么, 这个Two-Sample t Test的自由度是 $(20-1) + (20-1) = 38$ 。但是, 在Paired t Test中, 每个样本也是20个成绩, 一共40个成绩, 但只来自20个受试。所以, 这个Paired t Test的自由度是 $(20-1) = 19$ 。



再次强调, 样本之间的独立与否, 是很难界定的。你不可能通过一种公式去计算两个样本是否相互独立, 而只能通过科学的抽样方法和实验设计, 来最大限度的保证样本之间的独立性。所以, 请把你的**受试群体**、**抽样方法**都描述清楚, 人家才好判断你的样本是不是独立的, 才好去决定, 要不要相信你的结论。



一中样本 X_1

二中样本 X_2

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
钱二麻	66
.....

最后，我们通过一个思考题来说明，为什么不能通过计算，来判断两个样本是否相互独立。在刚才钱二麻同学的故事中，他考了66分，被抽样抽到了两次。因此，钱二麻的66分，同时出现在 X_1 和 X_2 两个样本中。因此，这两个样本，不是相互独立的。



一中样本 X_1

二中样本 X_2

抽样补救

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
.....

然后再在二中门口，重新抽取另外一个同学的成绩，来替换钱二麻的66分。



替换前

一中样本 X_1

二中样本 X_2

区别? 意义?

替换后

一中样本 X_1

二中样本 X_2

样本 互不独立

样本 相互独立

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
.....

姓名	成绩
张三	67
李四	76
王五	77
马六	73
钱二麻	66
.....

姓名	成绩
陈磊	78
刘强	59
胡亮	55
杨光	76
王浩	66
.....

这就很尴尬了。替换前，和替换后，这两个样本在数值上，没有发生任何变化。输入到统计软件中，算出来的t值和p值，也完全没有任何差别。但是，替换前，两个样本就不是相互独立的，替换后，两个样本就是相互独立的了。所以说，样本是否相互独立，不能靠计算得出。但这时，你肯定会问，既然结果都一样，那样本独立不独立，有区别吗？重新抽样补救，有意义吗？

34

替换前

一中样本 X_1		二中样本 X_2	
姓名	成绩	姓名	成绩
张三	67	陈磊	78
李四	76	刘强	59
王五	77	胡亮	55
马六	73	杨光	76
钱二麻	66	钱二麻	66
.....

哲学
&
信仰

替换后

一中样本 X_1		二中样本 X_2	
姓名	成绩	姓名	成绩
张三	67	陈磊	78
李四	76	刘强	59
王五	77	胡亮	55
马六	73	杨光	76
钱二麻	66	王浩	66
.....

样本 互不独立

样本 相互独立

这是一个很玄乎的问题。个人认为，这既是一个哲学问题，也是一个信仰问题。说是哲学问题，是因为，数学、科学，本来就是从哲学分化出来的；数值上没差别，但抽样方法不同，说明认识论不同，是哲学本质上的不同。说是信仰问题，是因为，在科研工作中，无论结果如何，难道不都该遵守科学的方法吗？这是最起码最起码的科研良知，而“良知就是天理”，良知就该是信仰。



方差齐性检验的基本原理 F检验初步

2024年1月7日 17:15

XX县 第一中学

μ_1 未知 X_1 已知
(67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72)

XX县 第二中学

μ_2 未知 X_2 已知
(78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71)

目标：比较 μ_1 和 μ_2 谁大谁小

$H_0: \mu_1 = \mu_2$

$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2 / \sqrt{n}}}$

Student's
双样本 t检验
样本要求

1. 两个样本 都符合正态性
2. 两个样本 相互独立性
3. 两个样本 方差齐性

大家好。我们在之前的课程中已经学习了Student's双样本t检验的基本原理（注意：双样本检验有很多种，本节课如不特殊说明，指的都是Student's双样本t检验）。双样本t检验对于两个样本有一定的要求，本入门课程只讲解最常见的三种要求。第一，两个样本都要符合正态性。第二，两个样本之间相互独立。这两个要求，在之前的课程中，我们已经讲过了。

双样本t检验对双样本的要求：

1. 两个样本都符合正态性
2. 两个样本相互独立性
3. 两个样本方差齐性

方差 齐性

Homogeneity of Variances

Student's
双样本 t检验
样本要求

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} \quad \text{总体 方差}$$

1. 两个样本 都符合正态性

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \quad \text{样本 方差}$$

2. 两个样本 相互独立性

3. 两个样本 方差齐性

今天，我们来看一下第三个要求：两个样本要满足方差齐性。方差齐性，英语叫做Homogeneity of variances。Variance就是方差的意思。总体方差，是表示总体中的各个观测值偏离总体均值的程度，用 σ^2 来表示。样本方差，是表示样本中的各个观测值偏离样本均值的程度，用 s^2 来表示。

方差 齐性

Homogeneity of Variances

Student's
双样本 t检验
样本要求

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} \quad \text{总体 方差}$$

1. 两个样本 都符合正态性

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \quad \text{总体 标准差}$$

2. 两个样本 相互独立性

3. 两个样本 方差齐性

需要注意的是，总体方差 σ^2 的单位是观测值单位的平方。例如，在一中和二中的故事中，一中的总体方差的单位是 分^2 ，分的平方，没有任何现实意义。所以，把总体方差开平方，得到总体标准差 σ ， σ 的单位就和观测值的单位一样了，都是分，这样方便理解。

方差 齐性

Homogeneity of Variances

Student's

bilibili

双样本 t 检验

样本要求

$$\text{分}^2 \quad s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \quad \text{样本 方差}$$

↓

$$\text{分} \quad s = \sqrt{\frac{\sum(X - \mu)^2}{n - 1}} \quad \text{样本 标准差}$$

1. 两个样本 都符合正态性
2. 两个样本 相互独立性
3. 两个样本 方差齐性

同理，样本方差 s^2 的平方开平方，得到样本标准差 s ， s 的单位就和观测值的单位一样了，都是分。

方差 齐性

Homogeneity of Variances

Student's

bilibili

双样本 t 检验

样本要求

$$\begin{array}{ll} \sigma^2 & \sigma \\ \text{总体 方差} & \text{总体 标准差} \\ s^2 & s \\ \text{样本 方差} & \text{样本 标准差} \end{array}$$

1. 两个样本 都符合正态性
2. 两个样本 相互独立性
3. 两个样本 方差齐性

注意，这些符号表示都是统计学中的惯例，大家务必要记住。方差齐性或是标准差齐性，都是同样的原理，下文为了表述简单，如不特殊指明，方差还是标准差，就不再严格区分了。

方差 齐性

Homogeneity of Variances

Student's

bilibili

双样本 t 检验

样本要求

Equal Variances

方差 相等

1. 两个样本 都符合正态性
2. 两个样本 相互独立性
3. 两个样本 方差齐性

Homogeneity，这个单词看起来比较唬人，应该至少是个六级词汇，是“同质、齐次、齐性”的意思。我们换一种说法，方差齐性就是“Equal Variances”。Equal 这个词大家应该都认识，顶多是四级词汇，是“相等”的意思。所以，“方差齐性”，就是“方差相等”。谁的方差相等？注意，是样本所代表的总体的方差相等。当然，完全绝对相等是不可能的，差不多就行，大差不差，就“整齐”了，所以叫做“齐性”。

方差齐性

Homogeneity of Variances

Equal Variances

方差相等

为什么?

Student's

双样本 t 检验

样本要求

1. 两个样本都符合正态性
2. 两个样本相互独立性
3. 两个样本方差齐性

本节课的重点是，向大家通俗讲解，双样本t检验中，样本为什么要满足方差齐性？
我们还是通过讲故事来感性理解。



一中全校均分
 $\mu_1 = 70$



二中全校均分
 $\mu_2 = 70$



还是回到某县城一中和二中的故事中。在一次全县统考中，一中和二中考同一套试卷，一中总共有1000名同学，二中也总共有1000名同学，一中的总体均分 μ_1 是70分，二中的总体均分 μ_2 也是70分。两个学校的均分相等。



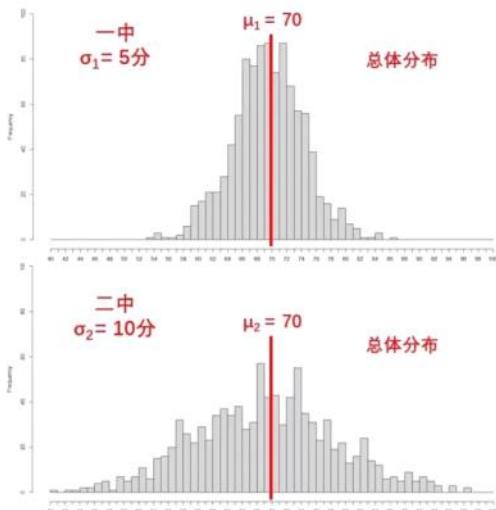
一中全校均分
 $\mu_1 = 70$



二中全校均分
 $\mu_2 = 70$



这时，假如有位母亲，要给孩子选一个中学去上学。她看到两个学校的总体均分一样，心里想，那应该就是两个学校的水平都一样吧，那么，孩子上哪个学校都一样了。

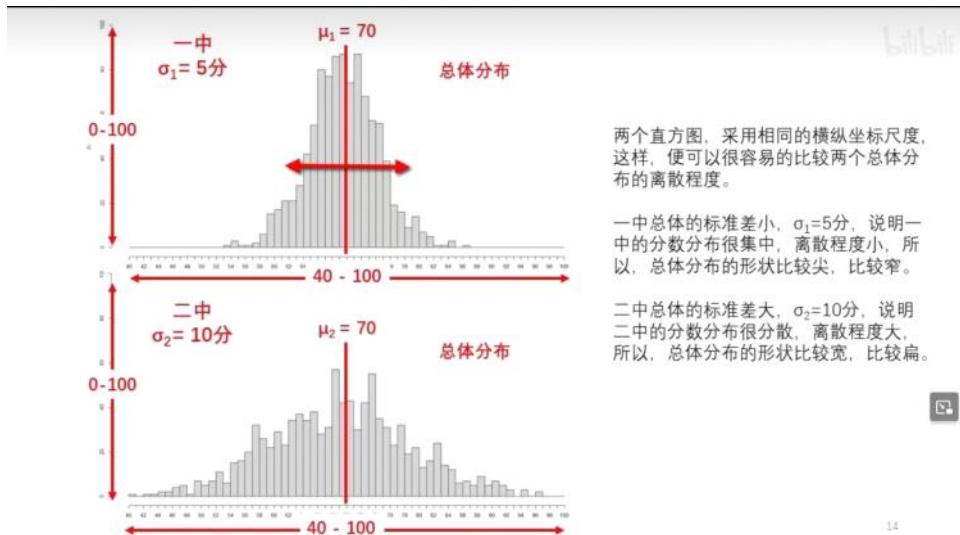


假如这时教育局公布了一中和二中的每个同学的成绩，这位母亲就把数据输入到程序里，算出两个总体的标准差，并做出两个总体的直方图。

一中的总体标准差 σ_1 大概是5分左右，二中的总体标准差 σ_2 大概是10分左右。

一中和二中的总体均分 μ_1 和 μ_2 ，都是在总体分布的对称轴70分。

12

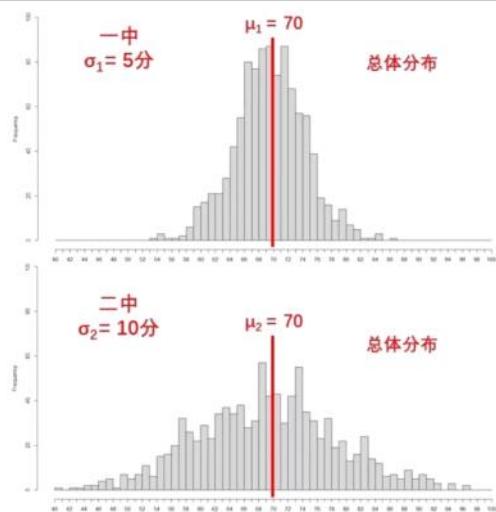


两个直方图，采用相同的横纵坐标尺度，这样，便可以很容易的比较两个总体分布的离散程度。

一中总体的标准差小， $\sigma_1=5$ 分，说明一中的分数分布很集中，离散程度小，所以，总体分布的形状比较尖，比较窄。

二中总体的标准差大， $\sigma_2=10$ 分，说明二中的分数分布很分散，离散程度大，所以，总体分布的形状比较宽，比较扁。

14



此时，面对两个中学的总体分布离散程度，这位正在选择中学的母亲，最终会选择哪个学校呢？

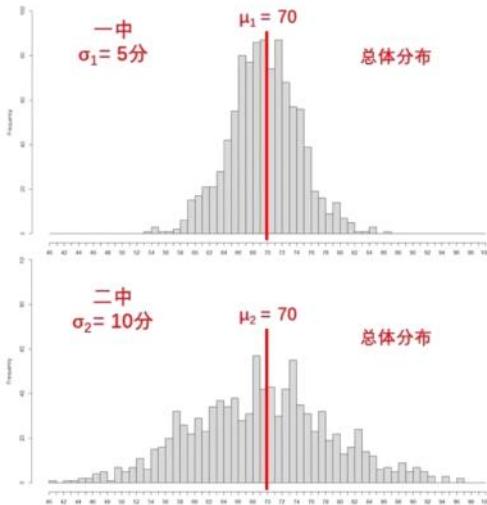
假如这位母亲是一名应试教育参与者的话，假如她关注的是总体均分的话，那么，她应该选择总体标准差较小的一中。

因为当她关注到一个中学的总体均分是70分时，她心中潜在的一种想法是，她的孩子来这个中学后，也能考70分左右。

那么，这两个中学里，学生在哪个中学考到均分70分的概率比较大呢？显然，答案是一中。

一中的总体标准差小，所有学生的成绩都很集中，都靠近均分70分，所以，每一个学生考到70分左右的概率会更大。所以，这位母亲应该选择一中。

15

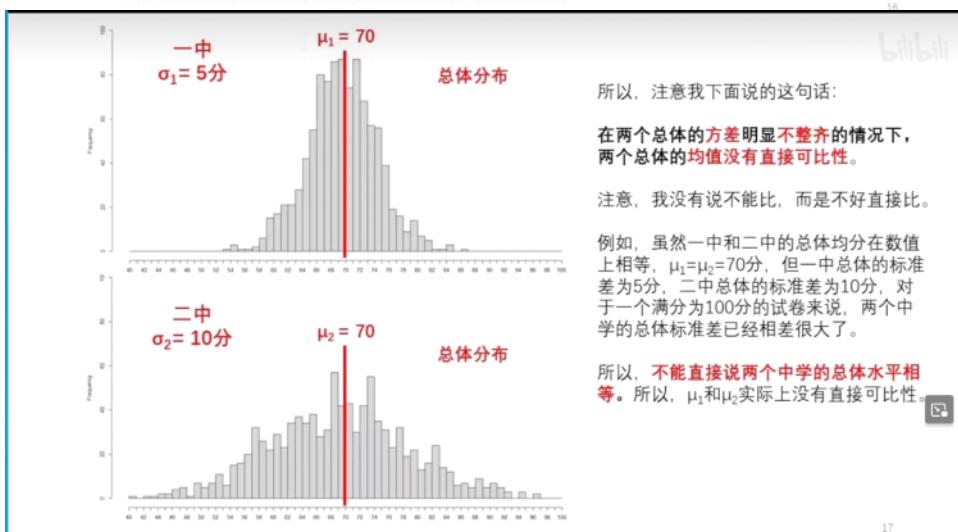


当然，大家可能会说，二中的总体分布比较扁，但考到高分的概率也大啊。分布图中，二中考90分以上的，比一中多多了。

对，这种说法没错。但请时刻提醒自己，在我们杜撰的一中和二中的故事中，这是一个双样本t检验，比较的是两个学校的总体均分，而不是最高分。

假如是比较最高分的话，一中和二中会选出自己的尖子生去PK，那就不用统考了，那叫选拔竞赛，选拔跟抽样是完全不同的两回事。

在我们的故事中，两个学校的水平是以总体均分衡量的。二中这些90分以上的高分，偏离了二中的总体均分很远，属于极端现象。我更愿意认为，这些高分，是学生个体能力的体现，而不是学校总体水平的表现。



所以，注意我下面说的这句话：

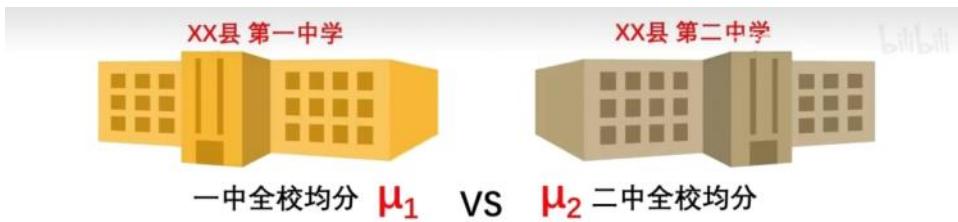
在两个总体的方差明显不整齐的情况下，两个总体的均值没有直接可比性。

注意，我没有说不能比，而是不好直接比。

例如，虽然一中和二中的总体均分在数值上相等， $\mu_1 = \mu_2 = 70$ 分，但一中总体的标准差为5分，二中总体的标准差为10分，对于一个满分为100分的试卷来说，两个中学的总体标准差已经相差很大了。

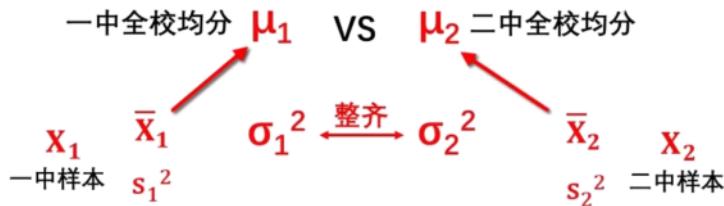
所以，不能直接说两个中学的总体水平相等。所以， μ_1 和 μ_2 实际上没有直接可比性。

17



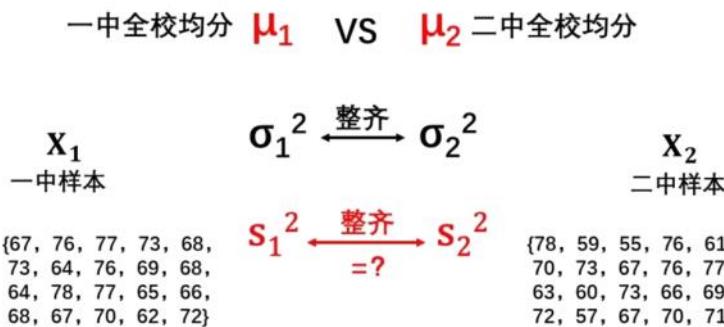
$$\begin{array}{c}
 \mathbf{X}_1 \quad \mathbf{X}_2 \\
 \{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72\} \quad \{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71\}
 \end{array}$$

那么，问题来了。我们不知道一中和二中两个总体的分布，也不知道两个总体的方差，我们手头只有两个样本，那怎么知道两个总体的方差是否整齐呢？



{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72} {78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71}

那我们就分析这两个样本。样本不就是用来代表总体的吗。之前我们学过用样本均值 \bar{X} 来估计总体均值 μ 。其实，我们也可以用样本方差 s^2 来估计总体方差 σ^2 。



所以，要求两个总体的方差整齐，现在变成了要求两个样本的方差整齐。这时，问题又来了，两个样本的方差，怎么看是不是“整齐”的呢？假如两个样本的方差完全相等，那自然最好。但完全相等的情形太少见了。注意我说的下面这句话，就算两个样本是从同一个总体中抽出来的，由于随机性，也不能保证这两个样本的方差就是相等的。

方差齐性检验

Test for Homogeneity of Variances

- F-test $F = \frac{s_1^2}{s_2^2}$ $\sigma_1^2 \longleftrightarrow \sigma_2^2$ 目标：总体方差 是否整齐？
- Levene's test $s_1^2 \longleftrightarrow s_2^2$
- Bartlett's test 手段：分析 样本方差

这时候，我们就需要引入方差齐性检验的概念了。即，通过样本，来检验总体的方差是否整齐。常见的方差齐性检验有：F-test, Levene's test, Bartlett's test等等。每种检验方法的公式都不相同，其中Levene 和 Bartlett检验的公式都非常复杂。不过，F检验的公式较为简单，就是通过两个样本方差的比值，来检验总体的方差是否整齐。

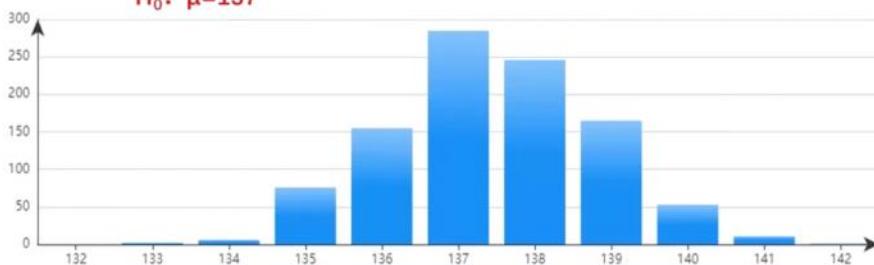


$H_0: \mu=137$

$$\bar{X} = \frac{\sum_i^n X_i}{n}$$

均值抽样分布

“均值检验”



我们明明知道一个excel表中，5000个学生的总体均分是137分，但是每次抽样算出来的样本均值，不可能都是137分，而是所有的均值分布，呈现出一个均值分布，即，原假设 $\mu=137$ 为真的均值分布。然后，再单独抽样1次，看这次抽样的均值在均值分布中是否极端，来判断要不要拒绝原假设。

方差齐性检验

$$\sigma_1^2 = \sigma_2^2$$

$H_0: \text{总体方差 齐齐}$

➤ F-test $F = \frac{s_1^2}{s_2^2}$

$X_1: \{67, 76, 77, 73, 68, 73, 64, 76, 69, 68, 64, 78, 77, 65, 66, 68, 67, 70, 62, 72\}$

➤ Levene's test

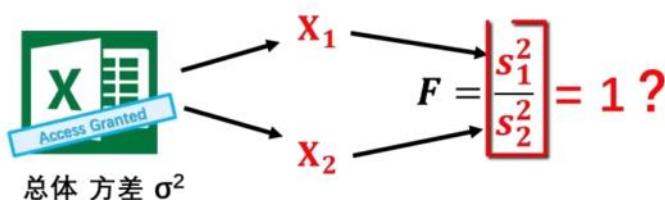
$X_2: \{78, 59, 55, 76, 61, 70, 73, 67, 76, 77, 63, 60, 73, 66, 69, 72, 57, 67, 70, 71\}$

➤ Bartlett's test

类似的，各种方差齐性检验的基本原理，也都差不多：都是在原假设为真的情况下，先反复抽样，每次抽样按照公式计算出一个统计量，例如，计算一个F值，通过反复抽样就可以得到F值的分布；然后，再把本次的样本带入公式，得出本次的F值，看看在F分布中是否极端，进而拒绝或接受原假设。请注意，F检验有很多种形式，本节课用的，是其最简单的一种、用来检验方差齐性的形式。

25

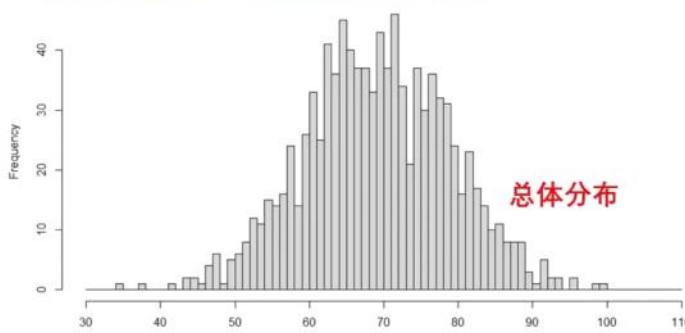
方差齐性检验 F-test



F检验的具体方法是，假如有两个样本，都是从同一个总体中抽出来的，那么，毫无疑问，这两个样本所代表的总体的方差是完全一样的。完美情况下，两个样本的方差，整齐到完全相等，那么F值，也就是 s_1^2 除以 s_2^2 ，应该等于1才对。但这种完美的情况，在现实中是不可能存在的。因为抽样的随机性，两个样本的方差必然存在区别。所以，F值不可能正正好等于1。但我们猜测，F值应该分布在1的左右。

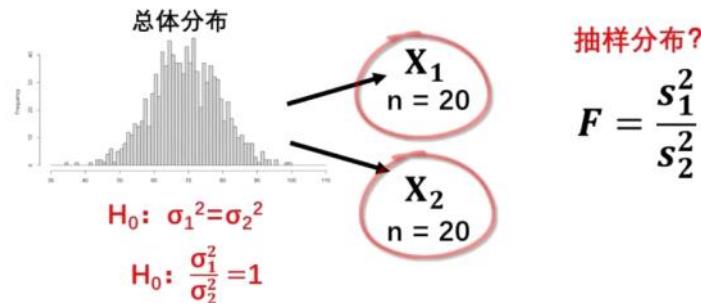
26

population <- rnorm(1000, 70, 10)

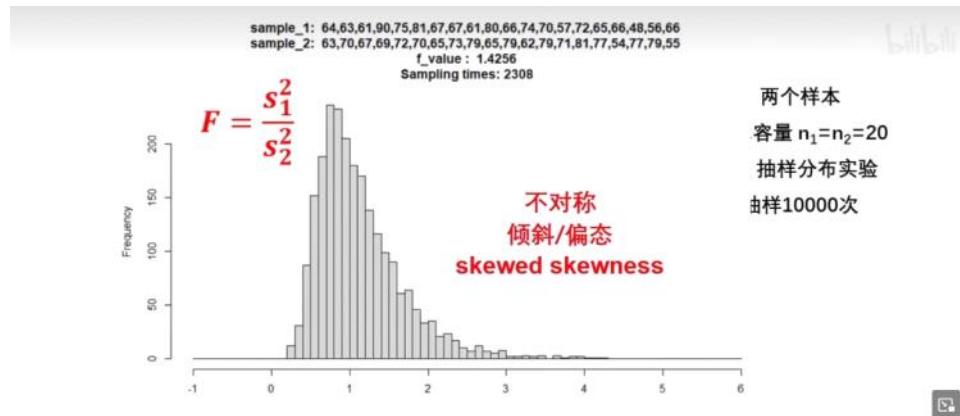


那么，我们就来做一个F值的抽样分布实验。我们先制造一个总体，这个总体共有1000个成绩，总体均分为70分，总体标准差为10分。直方图做出来，总体分布是这样的。

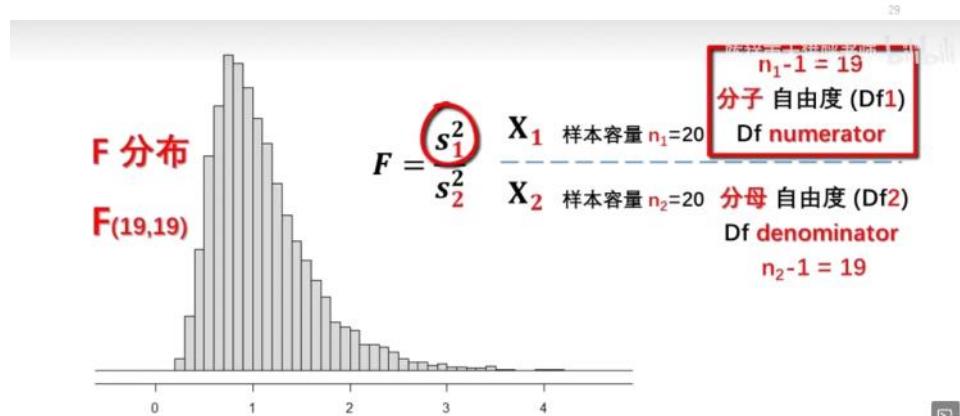
方差齐性检验 F-test



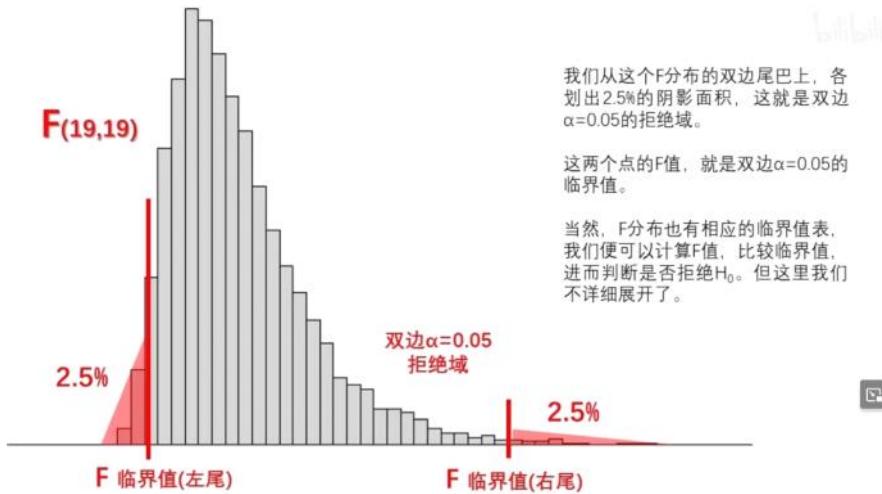
然后，在这同一个总体中，每次抽两个样本，样本容量都是20，然后计算F值。相当于在原假设 $\sigma_1^2 = \sigma_2^2$ 为真的情况下，或者说，在原假设 σ_1^2 比 σ_2^2 等于1为真的情况下，来看一下两个样本的方差比值，F值，是如何分布的。



实验开始。可以看出，F值抽样分布，逐渐呈现出一种不对称的铃铛形状。这种不对称，叫做“倾斜”或者“偏态”。英语中，其形容词叫做skewed，名词叫做skewness。既然不对称，也就没有对称轴。虽然峰值不是1，但我们仍然可以说，大部分的抽样F值，集中在F=1的附近。



这个结果是可以直观理解的。因为两个样本来自同一个总体，样本方差不会差得太多，因此，样本方差的比值，F值，都在1的左右浮动。这就是一个F分布，它有两个自由度，分子自由度和分母自由度。分子自由度是分子的样本容量减去1，记作Df1=20-1=19。分母自由度是分母的样本容量减去1，记作Df2=20-1=19。所以，这个F分布，记作 $F(19,19)$ ，意思是分子和分母自由度都是19的F分布。



```

X1 > yizhong
[1] 67 76 77 73 68 73 64 76 69 68 64 78 77 65 66 68
[17] 67 70 62 72
X2 > erzhong
[1] 78 59 55 76 61 70 73 67 76 77 63 60 73 66 69 72
[17] 57 67 70 71
> var.test(yizhong,erzhong)

F test to compare two variances

data: yizhong and erzhong
F = 0.51101, num df = 19, denom df = 19,
p-value = 0.1524
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2022652 1.2910498
sample estimates:
ratio of variances
0.5110132

```

因为我们不用临界值表，我们直接用R软件来计算p值。现在，回到一中和二中的故事中，在进行双样本t检验之前，先把一中样本 X_1 ，二中样本 X_2 ，输入到R软件中，通过F检验，来检验两个样本的方差齐性。

```

X1 > yizhong
[1] 67 76 77 73 68 73 64 76 69 68 64 78 77 65 66 68
[17] 67 70 62 72
X2 > erzhong
[1] 78 59 55 76 61 70 73 67 76 77 63 60 73 66 69 72
[17] 57 67 70 71
> var.test(yizhong,erzhong)

F test to compare two variances

data: yizhong and erzhong
F = 0.51101, num df = 19, denom df = 19,
p-value = 0.1524
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2022652 1.2910498
sample estimates:
ratio of variances
0.5110132

```

$H_0: \sigma_1^2 = \sigma_2^2$

$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$

F检验的函数是var.test()。yizhong和erzhong这两个变量就是 X_1 和 X_2 这两个样本。函数默认是双尾F检验。结果显示，F值等于0.51101，分子和分母自由度都是19。p=0.1524，大于 $\alpha=0.05$ 。于是，不拒绝 H_0 。 H_0 是什么呢？两个方差相等，或者两个方差的比值等于1。对立假设 H_1 是：两个方差的比值不等于1。

```

X1 > yizhong
[1] 67 76 77 73 68 73 64 76 69 68 64 78 77 65 66 68
[17] 67 70 62 72
X2 > erzhong
[1] 78 59 55 76 61 70 73 67 76 77 63 60 73 66 69 72
[17] 57 67 70 71
> var.test(yizhong,erzhong)

    F test to compare two variances

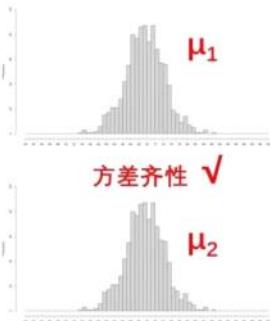
data: yizhong and erzhong
F = 0.51101, num df = 19, denom df = 19,
p-value = 0.1524
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2022652 1.2910498
sample estimates:
ratio of variances
0.5110132

```

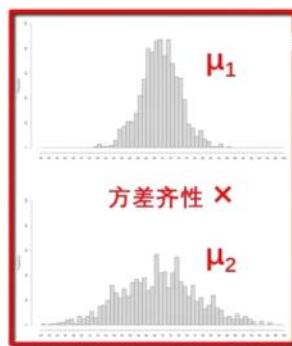
$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

按照计算结果，接受 H_0 ，认为这两个总体具有方差齐性，于是，可以进行双样本t检验。以上，就是在双样本t检验之前，用F检验来进行方差齐性检验的过程。



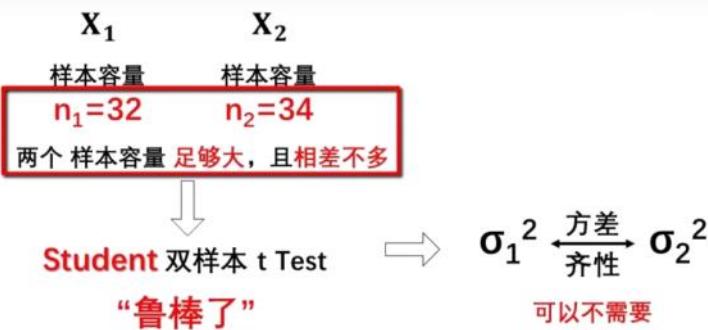
Student 双样本 t Test



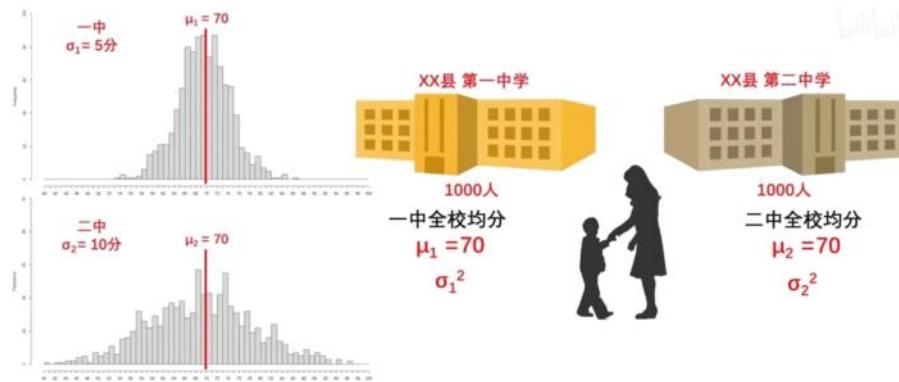
Welch 双样本 t Test

讲到这里，大家肯定有很多疑问。大家可能会问，假如两个总体的方差就是不整齐，但我偏要比较两个总体的均分，可以不可以呢？当然可以，你这么坚持，总有办法成全你。这种情况下，就不叫Student's t Test了，叫做Welch's t Test。但本节课，我们也不详细展开讲解。

35



当然，大家可能会听说，假如两个样本的样本容量足够大，且相差不多的话，例如，一个样本容量 $n_1=32$ ，另一个样本容量 $n_2=34$ 的话，Student 双样本t检验就具有“鲁棒性”，可以不进行方差齐性检验。这种说法也没错，但本节课我们也不展开讲解。



本节课，只是想让大家从感性上理解，双样本t检验为什么会有这种“方差齐性”的要求。请大家牢牢记住这个母亲选学校的故事。比较两个学校的水平，不能只看总体均分的，还要看一下总体的方差。总体方差如果不整齐，直接比较总体均分谁高谁低的话，不具有太大的说服力。

配对t检验

2024年1月7日 19:01

Paired t Test

“配对t检验”

“成对t检验”

名称不要搞混

Group t Test

“成组t检验”

“双样本t检验”

“独立样本t检验”

大家好，之前的课程中，我们已经引入了Paired t Test。中文里，常见的翻译是“配对t检验”或“成对t检验”。因为术语的翻译较为混乱，为避免大家产生误解，本课程特别提醒，你可能还听说过“成组t检验”，英文叫做“Group t Test”，“成组t检验”是之前讲过的“双样本t检验”或“独立样本t检验”，两个样本，就是两组数据，所以叫“成组”。所以，本课程中，“Paired t Test”就是“配对t检验”或“成对t检验”，不是“成组t检验”。大家不要搞混。

2

前测
(学期开始) pre-test

学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77

同一套试卷

后测
(学期结束) post-test

学号	01	02	03	04	17	18	19	20
后测成绩	67	76	66	88	81	79	68	80

同一批同学
一个学期后

王老师是某班级的英语老师，这个班有20名同学。本学期开始时，进行了一次英语摸底考试。然后，学期结束时，又进行了一次考试，试卷和之前的摸底考试是同一套试卷。目的就是为了看一看，这20名同学，在经历了一个学期的英语学习之后，英语水平，有没有显著提高。学期开始时的考试，叫做前测（pre-test），学期结束时的考试，叫做后测（post-test）。

5

学号	01	02	03	04	17	18	19	20	均分
前测成绩 pre	65	75	69	85	78	80	69	77	75.6
后测成绩 post	67	76	66	88	81	79	68	80	77.15
差值 d	+2	+1	-3	+3	+3	-1	-1	+3	\bar{d}

我们把每个同学，或者叫做“受试”的前后测成绩，和成绩的差值，都写在同一列。每个受试的前后测成绩，是“成对”的，是一个“pair”，这就是所谓的“配对”。“成对”的成绩，可以相减，得到一个差值。不“成对”的成绩，当然不能相减。

学号	01	02	03	04	17	18	19	20	均分
前测成绩 <i>pre</i>	65	75	69	85	78	80	69	77	75.6
后测成绩 <i>post</i>	67	76	66	88	81	79	68	80	77.15
差值 <i>d</i>	+2	+1	-3	+3	+3	-1	-1	+3	1.55 \bar{d}

“后测的总体平均水平 显著大于 前测”？

王老师计算了一下前后测的差值这一行，发现，所有受试的差值的均分，是正的1.55分。也就是说，后测平均分与前测平均分的差值，显然是大于0的。那么，王老师可可以说，后测的总体平均水平显著大于前测呢？

学号	01	02	03	04	17	18	19	20	均分
前测成绩 <i>pre</i>	65	75	69	85	78	80	69	77	75.6
后测成绩 <i>post</i>	67	76	66	88	81	79	68	80	77.15
差值 <i>d</i>	+2	+1	-3	+3	+3	-1	-1	+3	1.55 \bar{d}

d: difference

配对t检验 t 值公式

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

这时，我们就要用配对t检验了。配对t检验的t值公式是这个。大家在网上可能会看到不一样的公式，其实这些公式本质上都是一样的东西。本课程选了较为简单的一种公式。公式中，字母*d*就是difference的意思，*d*是一个数组。其中的每一个数值代表每一个受试前后测的成绩差值。 \bar{d} 就是差值数组*d*的平均值，也就是这一行差值的平均数。 s_d 就是差值数组*d*的标准差。然后，分母上，再除以一个根号*n*。

8

```
> pre <- c(65, 75, 69, 85, 77, 80, 82, 70, 73, 79, 81, 69, 75, 78, 74, 76, 78, 80, 69, 77)
> post <- c(67, 76, 66, 88, 78, 77, 83, 77, 74, 81, 80, 73, 78, 78, 79, 80, 81, 79, 68, 80)
> d <- post-pre
> d
[1]  2  1 -3  3  1 -3  1  7  1  2 -1  4  3  0  5  4  3 -1 -1  3
> mean(d)
[1] 1.55
```

mean() 命令：平均值

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

下面，我们用R软件，按照这个公式，先来手工算一下这个t值是多少。首先，我们把前测和后测的成绩输入到程序中，然后，算出差值数组*d*，再算出*d*的平均值，也就是 \bar{d} 。在R软件中，用**mean()**命令来算平均值，**mean**就是均值的意思。得到 \bar{d} 等于1.55分。

9

```
> pre <- c(65,75,69,85,77,80,82,70,73,79,81,69,75,78,74,76,78,80,69,77)
> post <- c(67,76,66,88,78,77,83,77,74,81,80,73,78,78,79,80,81,79,68,80)
> d <- post-pre
> d
[1]  2  1 -3  3  1 -3  1  7  1  2 -1  4  3  0  5  4  3 -1 -1  3
> mean(d)
[1] 1.55
> sd(d)
[1] 2.584875
```

sd() 命令： standard deviation, 标准差

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

公式中的 s_d : 差值数组 d 的 标准差

然后再算分母，先算分母中差值数组d的标准差 s_d ，用sd()命令来算标准差。sd就是standard deviation标准差的意思。这里的字母标记稍微有点混乱，大家要注意。公式里的 s_d ， s 就是标准差的意思，下标d，就是指差值数组d。所以，公式里的s下标d，理解为差值数组d的标准差。R软件里的命令sd()，是standard deviation的缩写。我们用命令sd()，算出差值数组d的标准差为2.58左右。

```

> pre <- c(65,75,69,85,77,80,82,70,73,79,81,69,75,78,74,76,78,80,69,77)
> post <- c(67,76,66,88,78,77,83,77,74,81,80,73,78,78,79,80,81,79,68,80)
> d <- post-pre
> d
[1]  2  1 -3  3  1 -3  1  7  1  2 -1  4  3  0  5  4  3 -1 -1  3
> mean(d)
[1] 1.55
> sd(d)
[1] 2.584875
> t <- mean(d) / (sd(d) / sqrt(20))
> t
[1] 2.681681

```

sqrt() 命令: **square root**, 平方根

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

根号n, 用sqrt()命令算出, sqrt就是square root, 平方根的意思。本例中, 样本容量n=20。最后, 我们在R软件中写出配对t检验的t值为, 差值的平均值 \bar{d} 除以括号里的差值的标准差 s_d 除以根号n, 得t=2.681681。

t临界值表									
cur.prob	t _{0.05}	t _{0.01}	t _{0.05}						
one-tail	0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005	0.001
two-tail	0.50	0.40	0.30	0.10	0.05	0.025	0.01	0.002	0.001
1	0.050	1.376	1.645	2.378	2.024	12.71	31.82	63.66	346.31
2	0.050	1.980	2.358	2.776	2.571	13.73	32.87	65.30	356.70
3	0.050	0.765	0.979	1.250	1.061	3.241	5.451	5.941	15.215
10	0.050	0.490	0.865	1.071	1.327	1.964	3.120	2.851	2.971
20	0.050	0.408	0.714	0.893	1.061	1.727	2.628	2.446	2.536
50	0.050	0.368	0.602	0.767	0.933	1.574	2.101	2.552	2.878
100	0.050	0.348	0.562	0.697	0.861	1.495	2.043	2.447	2.791
200	0.050	0.337	0.532	0.657	0.811	1.455	1.983	2.364	2.704
500	0.050	0.327	0.512	0.637	0.781	1.425	1.943	2.324	2.654
1000	0.050	0.321	0.503	0.627	0.765	1.405	1.918	2.297	2.621
21	0.050	0.660	0.953	1.021	1.793	2.080	2.618	2.831	3.527
22	0.050	0.660	0.953	1.021	1.793	2.080	2.618	2.831	3.527

$$df = n - 1$$

学号	01	02	03	04	17	18	19	20	均分
前测成绩 <i>pre</i>	65	75	69	85	78	80	69	77	75.6
后测成绩 <i>post</i>	67	76	66	88	81	79	68	80	77.15
差值 <i>d</i>	+2	+1	-3	+3	+3	-1	-1	+3	$\frac{1.55}{d}$

配对t检验 自由度 =

这时，我们查表，查t临界值表，来看一下p值是多少。通过手工查表，顺便复习、加深理解。之前课程中讲过，配对t检验的自由度，是单个样本中观测值的数量n减去1，或者说，是受试的个数n减去1，再或者说，是配对的对数n减去1，而不是双样本中的 $(n-1)+(n-1)$ 。再换句话说，虽然我们有前测和后测两组数据，每组数据中都有20个观测值，但这些数据是成对的，只来自20个受试。所以，配对t检验的自由度是 $n-1=20-1=19$ 。

13

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819

t=2.681681

H_0 : “后测均分 小于 前测均分” (单边检验)

我们在t临界值表中, 先找到df=19这一行。然后, 王老师本次t检验的目的, 是想得出“后测的总体平均水平显著大于前测”这个结论, 于是原假设 H_0 应该是其对立面, 也就是“后测均分小于前测均分”。于是, 这是一个单边检验。所以, 我们在t临界值表中, 再找到单边one-tail这一行。然后, 我们算出来的t值是2.681681, 可是, 在df=19这一行, 找不到这个精确的值, 没关系, 我们能找到两个临界值2.539和2.861。而我们算出来的t值2.681681, 正是介于这两个临界值之间。

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819

t=2.681681的 p值, 介于0.005和0.01之间

H_0 : “后测均分 小于 前测均分” (单边检验)

这时我们画一个框, 框出来的单边这一行的 α 值或者p值, 介于0.01和0.005之间。所以, 我们算出来的2.681681这个t值, 所对应的精确的p值, 也应当介于0.005和0.01之间。

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819

t=2.681681的 p值, 介于0.005和0.01之间

$\alpha=0.05$

$p < \alpha$ H_1 : “后测均分 大于 前测均分”

假如我们设定 $\alpha=0.05$ 的话, 显然, $p < \alpha$ 。于是, 我们就可以拒绝 H_0 : “后测均分小于前测均分”, 进而接受 H_1 , 得出结论: “后测均分 显著大于 前测均分”。也就是说, 王老师班的20名同学, 学了一个学期的英语, 英语成绩是有显著提高的。以上, 是我们用手工计算和手工查临界值表来进行配对t检验的过程。

```
> t.test(post,pre,paired=TRUE,alternative="greater")
```

Paired t-test

```
data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5505688      Inf
sample estimates:
mean of the differences
                  1.55
```

bilibili



当然，手工算是为了加深理解。R软件不用我们手工算的。下面，我们直接用一句R软件的命令行来进行配对t检验，这句命令行是：

```
> t.test(post,pre,paired=TRUE,alternative="greater")
```

Paired t-test

```
data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5505688      Inf
sample estimates:
mean of the differences
                  1.55
```

bilibili



首先，paired=TRUE这个参数，说明这是一个配对t检验。然后，大家一定要注意后测post和前测pre这两个参数的顺序，这个顺序一定要和alternative备用假设配合使用。命令行中参数的顺序，组合出我们想要得到的结论，也就是备用假设，所以应当这样读出来理解：“本次配对t检验的备用假设为，后测post比前测pre的分数高，greater”。

```
> t.test(post,pre,paired=TRUE,alternative="greater")
```

Paired t-test

```
data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5505688      Inf
sample estimates:
mean of the differences
                  1.55
```

bilibili



我们回车，看结果输出。首先，程序识别出，这是一个配对t检验。然后，算出来的t值2.6817和我们刚才手工算出来的t值2.681681是一致的。自由度df=20-1=19，p值是0.00738，和我们刚才查表估计的介于0.005和0.01之间，也是一致的。

```

> t.test(post,pre,paired=TRUE,alternative="greater")
Paired t-test

data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.5505688      Inf
sample estimates:
mean of the differences
1.55

```

注意备用假设 H_1 这里，翻译过来是，后测post和前测pre均分之间的真正差值大于0，和我们刚才说的“后测均分大于前测均分”，也是一个意思。这里是单边的95%置信区间，是从0.5505688到正无穷infinity，是指前后测差值的区间估计。单边置信区间，在前面第14节中也已经讲解过了，忘了的同学，可以复习一下。

```

> t.test(post,pre,paired=TRUE,alternative="greater")
Paired t-test

data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.5505688      Inf
sample estimates:
mean of the differences
1.55

```

这里的mean of the differences，注意，differences是个复数，因为差值数组d中一共有20个差值，差值数组的均值是1.55分，和我们之前算出来的也是一样的。综上，用R软件的配对t检验命令行，算出p值=0.00738，在 $\alpha=0.05$ 的显著水平下，我们可以拒绝 H_0 ，接受 H_1 ，认为“后测均分显著大于前测均分”。

学号	01	02	03	04	17	18	19	20	均分
前测成绩 pre	65	75	69	85	78	80	69	77	75.6
后测成绩 post	67	76	66	88	81	79	68	80	77.15
差值 d	+2	+1	-3	+3	+3	-1	-1	+3	\bar{d}

配对t检验的实质就是
两组数据的差值“是否为0”的单样本t检验

下面，我们给出配对t检验的实质。请注意我下面说的这句话，“配对t检验的实质就是，两组数据的差值是否为0的，单样本t检验”。

学号	01	02	03	04	17	18	19	20	均分
前测成绩 <i>pre</i>	65	75	69	85	78	80	69	77	75.6
后测成绩 <i>post</i>	67	76	66	88	81	79	68	80	77.15
差值 <i>d</i>	+2	+1	-3	+3	+3	-1	-1	+3	\bar{d}

两个样本
一个样本

配对t检验的 实质 就是

两组数据的 差值 “是否为0”的 单样本t检验

为什么这么说呢？在我们编的这个故事中，前后测的两组成绩，相当于两个样本，两个样本配对相减，得到一个差值数组d。假如我们把这个差值数组d，看作本次前后测实验，所获得的唯一一个样本。

```
> d
[1] 2 1 -3 3 1 -3 1 7 1 2 -1 4 3 0 5 4 3 -1 -1 3
> t.test(d, mu=0, alternative="greater")
One Sample t-test 单样本 t检验
data: d
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
0.5505688 Inf
sample estimates:
mean of x
1.55
-----
> t.test(post

```
pre, paired=TRUE, alternative="greater")
Paired t-test 配对 t检验
data: post and pre
t = 2.6817, df = 19, p-value = 0.00738
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.5505688 Inf
sample estimates:
mean of the differences
1.55
```


```

然后，我们把这个差值样本d，和 $\mu=0$ ，做一个单样本t检验，我们来看一下结果。

可以看到，一个差值样本d和 $\mu=0$ 的单样本t检验，和两个样本的配对t检验，结果是一模一样的。

其中的唯一差别就是，配对t检验时，你很省事，不用自己算出差值d，直接把两个样本丢给程序就可以了。

学号	01	02	03	04	17	18	19	20	均分
前测成绩 <i>pre</i>	65	75	69	85	78	80	69	77	75.6
后测成绩 <i>post</i>	67	76	66	88	81	79	68	80	77.15
差值 <i>d</i>	+2	+1	-3	+3	+3	-1	-1	+3	\bar{d}

$$\text{配对 t检验 } t = \frac{\bar{d}}{s_d / \sqrt{n}} \quad t = \frac{\bar{x} - 0}{s_x / \sqrt{n}} \text{ 单样本 t检验}$$

配对t检验的 实质 就是

两组数据的 差值 “是否为0”的 单样本t检验

我们这时回过头来，比较一下配对t检验和单样本t检验的公式。发现它俩是完全一样的，只不过差值样本在配对t检验中用d来表示，在单样本t检验中用x来表示。所以说，“配对t检验的实质就是，两组数据的差值是否为0的，单样本t检验”。

配对 t检验 VS 双样本 t检验

两个角度比较：实验设计、数学计算



上面，我们给出了配对t检验的数学本质。下面，我们通俗的讲解下，为什么非要用配对t检验，用双样本t检验不行吗？我们试着从两个角度：实验设计角度和数学计算角度来说明这个问题。

27

实验设计角度：

配对 t检验设计 – 比较前后测成绩

(只有1个班, n=20)

影响因素 (factor) “一学期英语课”	不存在	存在
测量值 (measure) “英语成绩”	前测成绩	后测成绩

第一：实验设计角度。这需要回到我们的故事本身，或者说，回到我们想要回答的问题本身。王老师想知道，“一学期的英语课”，能否让这个班的20名同学的“英语成绩”产生显著提高。我们把“一学期英语课”叫做影响因素, factor；把“英语成绩”叫做测量值, measure，我们有两个测量值，前测成绩和后测成绩。

实验设计角度：

配对 t检验设计 – 比较前后测成绩

(只有1个班, n=20)

影响因素 (factor) “一学期英语课”	不存在	存在
测量值 (measure) “英语成绩”	前测成绩	后测成绩

我们想研究，“一学期英语课”这个因素，“存在”与“不存在”的两种情况下，同一批受试的“英语成绩”这个测量值有没有差异。前测成绩就是“一学期英语课”这个因素“不存在”时的测量值，后测成绩就是影响因素“存在”时的测量值。然后，比较前后测成绩。这就是配对t检验的设计。听起来比较合理，可行性也比较高，容易操作。



实验设计角度：

双样本 t检验设计 - 仅比较后测

(A班 n=20, B班 n=20, 两个班 前测 无差别)

影响因素 (factor)	不存在 (一学期不上课)	存在 (一学期正常上课)
测量值 (measure)	A班 后测成绩	B班 后测成绩

现在, 假如我们就是要用双样本t检验, 该怎么设计实验呢? 假如王老师有两个班, A班和B班, 每个班都有20名同学, 一共有40名同学。学期开始时, 两个班都进行了前测摸底考试, 发现, 两个班的均分完全一样, 方差也是齐性的。也就是说, 两个班的初始英语水平, 是完全无差别的。然后, 为了实验研究“一学期英语课”这个因素的存在与否, 是否能让本来无差别的两个班的“英语成绩”产生差别, 王老师决定A班这个学期不上英语课了, B班这个学期正常上英语课。

实验设计角度：

双样本 t检验设计 - 仅比较后测

(A班 n=20, B班 n=20, 两个班 前测 无差别)

影响因素 (factor)	不存在 (一学期不上课)	存在 (一学期正常上课)
测量值 (measure)	Control group A班 控制组 对照组 后测成绩	Treatment group B班 实验组 后测成绩

此时, A班就叫做Control group, 翻译过来叫“控制组”或“对照组”, 意思就是, 控制A班不存在“一学期英语课”这个因素; B班叫做Treatment group, 翻译过来叫“实验组”, treatment是“对待”或“治疗”的意思, 就是说, B班存在影响因素, 接受了一学期的“英语对待”或“英语治疗”。学期结束后, A班和B班, 再进行一次后测考试, 考的还是之前摸底考试的同一套试卷。这时, 我们再用t检验比较A班和B班的后测成绩, 便是双样本t检验了。注意, 因为两个班前测无差别, 所以只比较两个班的后测成绩。

实验设计角度：

双样本 t检验设计 - 仅比较后测

(A班 n=20, B班 n=20, 两个班 前测 无差别)

影响因素 (factor)	不存在 (一学期不上课)	存在 (一学期正常上课)
测量值 (measure)	Control group A班 控制组 对照组 后测成绩	Treatment group B班 实验组 后测成绩

此时, A班就叫做Control group, 翻译过来叫“控制组”或“对照组”, 意思就是, 控制A班不存在“一学期英语课”这个因素; B班叫做Treatment group, 翻译过来叫“实验组”, treatment是“对待”或“治疗”的意思, 就是说, B班存在影响因素, 接受了一学期的“英语对待”或“英语治疗”。学期结束后, A班和B班, 再进行一次后测考试, 考的还是之前摸底考试的同一套试卷。这时, 我们再用t检验比较A班和B班的后测成绩, 便是双样本t检验了。注意, 因为两个班前测无差别, 所以只比较两个班的后测成绩。

实验设计角度：

双样本 t检验设计 - 仅比较后测

(A班 n=20, B班 n=20, 两个班 前测 无差别)

影响因素 (factor) “一学期英语课”	不存在 (一学期不上课)		存在 (一学期正常上课)	
测量值 (measure) “英语成绩”	Control group 控制组 对照组	A班 后测成绩	Treatment group 实验组	B班 后测成绩

于是, A班和B班的后测成绩作为双样本, 每个样本的样本容量都是n=20, 且两个样本中的受试不一样, 都是互相独立的, 两个样本一共40个不同的受试。此时, 这个双样本t检验的自由度为 $(20-1) + (20-1) = 38$ 。

配对 t检验设计 - 比较前后测成绩

(只有1个班, n=20)

影响因素 (factor) “一学期英语课”	存在	不存在
测量值 (measure) “英语成绩”	前测成绩	后测成绩

我们来比较这两个实验设计。其实很显然, 这个双样本t检验存在严重缺陷, 就是, 现实世界中, 王老师不可能让A班的学生一个学期不上课。

因此, 这个双样本设计, 是不可行的。

这就是说, 从实验设计角度来看, 由于现实世界中存在很多具体实际的限制, 例如, 不能让学生无缘无故的不上课, 所以, 有的情况下只好采用配对t检验, 而不是双样本t检验。

数学计算角度：

配对 t检验

学号	01	02	03	04	17	18	19	20
前测成绩 pre	65	75	69	85	78	80	69	77
后测成绩 post	67	76	66	88	81	79	68	80

双样本 t检验

赋值同样的数据

学号	01	02	03	04	17	18	19	20
A班 后测成绩	65	75	69	85	78	80	69	77
学号	01	02	03	04	17	18	19	20
B班 后测成绩	67	76	66	88	81	79	68	80

学期结束后, A班和B班的后测成绩如下。为了对比分析, 我们把A班后测成绩设定为配对实验中的前测成绩, B班后测成绩设定为配对实验中的后测成绩。这相当于说, 双样本t检验中, 两个样本的数据, 和配对t检验中两个样本的数据, 一模一样。这样来比较一下, 同样的数据, 在双样本t检验和配对t检验中, 结果有何不同。

```

> t.test(b,a,var.equal=TRUE,alternative="greater")
Two Sample t-test 双样本t检验

data: b and a
t = 0.92659, df = 38, p-value = 0.18 不可以 拒绝原假设
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-1.27028 Inf
sample estimates:
mean of x mean of y
77.15 75.60

```



```

> t.test(post,pre,paired=TRUE,alternative="greater")
Paired t-test 配对t检验

data: post and pre
t = 2.6817, df = 19, p-value = 0.00738 可以 拒绝原假设
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.5505688 Inf
sample estimates:
mean of the differences
1.55

```

我们发现，同样的两个样本，假如使用双样本t检验的话，t值算出来是0.92659，比配对t检验中的t=2.6817小很多，双样本t检验中的p值算出来是0.18，比配对t检验中的p=0.00738倒是大了不少。

从配对t检验到双样本t检验，t值变小了，也就是变得不极端了，那么p值自然也就变大了。

p=0.18，假如仍然设定 $\alpha=0.05$ 的话，那么就拒绝不了原假设了。

配对t检验									
学号	01	02	03	04	17	18	19	20
前测成绩 pre	65	75	69	85	78	80	69	77
后测成绩 post	67	76	66	88	81	79	68	80

t = 2.6817

df = 19

p = 0.00738

有 显著性差异

双样本t检验									
学号	01	02	03	04	17	18	19	20
A班 后测成绩	65	75	69	85	78	80	69	77
B班 后测成绩	67	76	66	88	81	79	68	80

t = 0.92659

df = 38

p = 0.18

没有 显著性差异

于是，同样的两组数据，本来在配对t检验中，还能计算出显著性差异，但是假如换成双样本t检验，就检测不出显著差异了。

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \xrightarrow{\text{分子相同}} t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2}/\sqrt{n}} \quad \begin{matrix} \text{t值} \\ \text{变小} \end{matrix}$$

分子相同
分母变大

配对t检验

双样本t检验

这是为什么呢？我们通过t值公式就可以看出来。从配对t检验到双样本t检验，公式中分子的数值，其实是一样的，都是两组数据差值的均值。分母上，根号n也是相同的，n都等于20。但是，双样本t检验这里，两个样本各自方差之和的平方根，肯定比两个样本差值的标准差，要大得多。小学数学里学过，分子不变，分母变大，值就变小了。t值变小了，就没那么极端了，所以，统计学差异消失了。

配对 t 检验 相依样本 (dependent)									
学号	01	02	03	04	17	18	19	20
前测成绩 pre	65	75	69	85	78	80	69	77
后测成绩 post	67	76	66	88	81	79	68	80
$t = 2.6817$ $df = 19$ $p = 0.00738$ 差异显著									

双样本 t 检验 独立样本 (independent)									
学号	01	02	03	04	17	18	19	20
A班 后测成绩	65	75	69	85	78	80	69	77
学号	01	02	03	04	17	18	19	20
B班 后测成绩	67	76	66	88	81	79	68	80
$t = 0.92659$ $df = 38$ $p = 0.18$ 差异不显著									

这就从数学计算角度说明了，有时候，同样的两组数据，通过配对t检验设计，比通过双样本t检验设计，更能获取显著的统计学差异。这也就是之前第17节“样本独立性”课程中提到的，有时候，通过“相互之间完全不独立的两个样本”或者叫“相依样本”，比通过“相互完全独立的两个样本”，更容易达成想要得到的结论。以上，就是为什么有些情况下，用配对t检验而不用双样本t检验的原因。

39

配对t检验的缺点：考试效应

英语试题									
1. 单项选择 (单题 10 小题，每小题 1 分，共 10 分) 从每个题所给的四个选项中选出最佳选项。									
1. What is the annual school trip we had? I will never forget it. A. a B. an C. the D. /									
2. You'd better not hang out after school _____ telling your parents. They may worry about you. A. by B. with C. without D. after									
3. Li Lei didn't play computer games last weekend. _____, he worked as a volunteer in an old people's home. A. Instead B. Certainly C. Though D. Gradually									
4. Please remember to _____ the electricity and water before you leave the laboratory. A. take off B. shut off C. go off D. put off									

(学期开始) pre-test									
学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77

(学期结束) post-test									
学号	01	02	03	04	17	18	19	20
后测成绩	67	76	66	88	81	79	68	80

前测、后测 同一套试卷

最后，我们再举两个例子来说明一下配对t检验的缺点。第一个例子，考试效应。仍然是这个故事，很多同学可能一开始就想要提问了：就算这一学期没上英语课，同一份试卷，考两次，按常识，第二次后测的成绩，也肯定比第一次前测的成绩要高啊。所谓“一回生、两回熟”嘛。是的，这就是前测后测实验中不可避免的一个问题。

配对t检验的缺点：考试效应

英语试题									
1. 单项选择 (单题 10 小题，每小题 1 分，共 10 分) 从每个题所给的四个选项中选出最佳选项。									
1. What is the annual school trip we had? I will never forget it. A. a B. an C. the D. /									
2. You'd better not hang out after school _____ telling your parents. They may worry about you. A. by B. with C. without D. after									
3. Li Lei didn't play computer games last weekend. _____, he worked as a volunteer in an old people's home. A. Instead B. Certainly C. Though D. Gradually									
4. Please remember to _____ the electricity and water before you leave the laboratory. A. take off B. shut off C. go off D. put off									

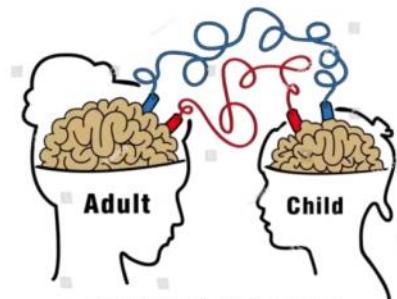
(学期开始) pre-test									
学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77

(学期结束) post-test									
学号	01	02	03	04	17	18	19	20
后测成绩	67	76	66	88	81	79	68	80

当然，我们可以不考同一份试卷，我们可以在后测中，采用所谓的“难度相同”的另一份试卷，但如何证明两份试卷的“难度相同”，本身就是一个问题。不仅如此，假如两套试卷虽然题目不同，但知识点相同，同学们考完前测，赶紧翻书复习，也会使后测成绩提高。这个就叫做“**考试效应对实验设计内部效度的威胁**”，“**A testing threat to the internal validity of research design**”。最后这句话中的术语，不做要求，仅供已经遇到此类术语的同学们参考。



配对t检验的缺点：成熟效应



图片来源: <https://www.shutterstock.com>

第二个例子，成熟效应。不知道随着年龄的增长，大家在回想自己成长历程的时候，有没有这种感觉：自己小时候，有些知识怎么学也学不会，老师怎么讲也听不懂；后来长大了，二三十岁了，回头看看小时候的知识，怎么就那么简单呢？

42

配对t检验的缺点：成熟效应

(学期开始) pre-test

学号	01	02	03	04	17	18	19	20
前测成绩	65	75	69	85	78	80	69	77

(学期结束) post-test

学号	01	02	03	04	17	18	19	20
后测成绩	67	76	66	88	81	79	68	80



于是，在我们这个英语前测后测的故事中，也存在这个问题。有些题目，就算不去上课，不去专门学习，学生成长了一个学期，头脑和智商更成熟了，在后测考试中自然就能做出来了，于是后测成绩提高了。这个就叫做“成熟效应对实验设计内部效度的威胁”，“A maturation threat to the internal validity of research design”。当然，这些术语也不做要求。

t检验的本质信噪比

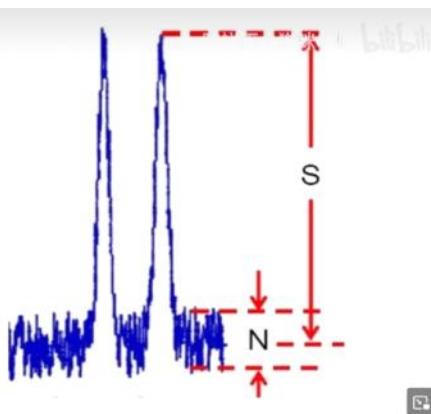
2024年1月7日 19:35

信噪比：信号与噪声的比值

Signal
—
Noise

Signal to Noise Ratio

SNR or S/N



大家好，本节课，我们来讲一下t检验的本质。t检验的真实本质，是一种“信噪比”。“信噪比”，是“信号与噪音的比值”的简称。英语叫做Signal to Noise Ratio，简称SNR或S/N。



举一个例子，一大群人聚在一起聊天，声音吵得不得了。假如你和你的朋友也在这群人里聊天，虽然面对面，但如果你朋友说话声音太小的话，就会被其他人的噪音给淹没了，你压根儿就听不清楚。

这时候，你朋友说的话，就是“**信号**”，是对你有意义的信息。而整个人群聊天的声音，就是**噪音**。你朋友说话的声音越大，信噪比就越大，你就越有可能听得到信号。你朋友的声音越小，甚至比噪音还小，信噪比就越小，信号就有可能淹没到噪音里了，你就听不到信号了，也聊不成天了。

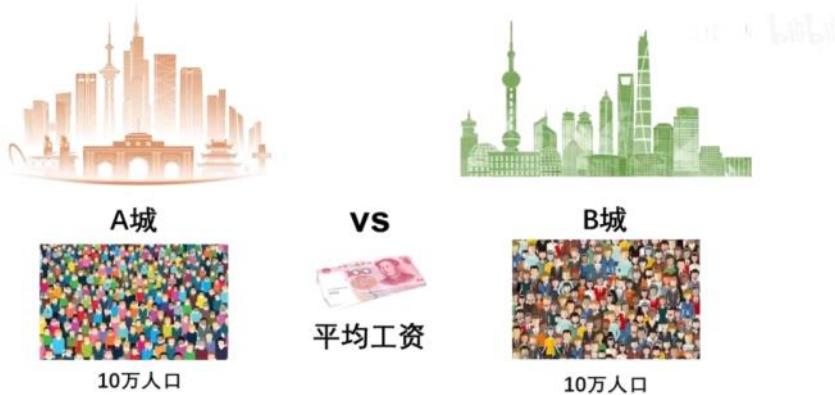


再举个例子，我小时候，家里看的是黑白电视，屋子外面竖个天线，接受电视信号。那时候，电视频道少，电视上有个旋钮，需要手动调频道。扭到有信号的频道，就能收到节目。没信号的频道，屏幕上就显示白花花的雪花。

这个雪花，就是纯粹的噪音，叫做**白噪音**。这时候，没有信号，信号为0，**信噪比就等于0**。



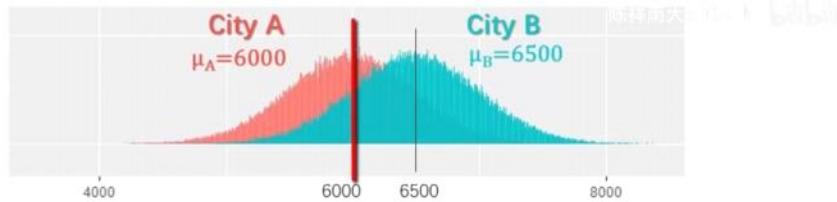
有信号的频道，信噪比就很大，当然雪花可能也有一点，但能够检出图像和声音信号了。以上，是“信噪比”的概念在现实物理世界中的两个例子。



现在，我们把“信噪比”引入到统计学中来。我们仍然编故事。有A和B两个城市，每个城市都有10万人口，我们想比较两个城市人口的月平均工资。

```
city_A <- rnorm(100000, mean=6000, sd=500)
city_B <- rnorm(100000, mean=6500, sd=500)
```

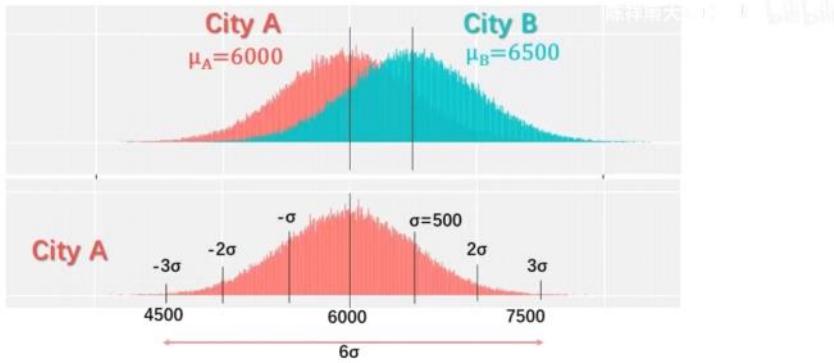
我们用R软件命令，生成两个正态分布的总体。假如，A城10万人口的平均工资为6000元，标准差 σ 为500元。B城平均工资为6500元，标准差 σ 也为500元。标准差保持一样，是为了“方差齐性”，忘了是什么的同学，请复习之前的课程。



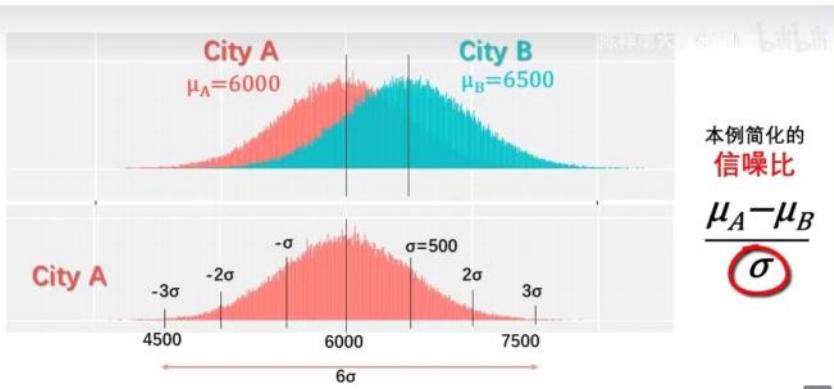
$$\mu_A - \mu_B = 500 \text{ (信号)}$$

“ μ_A 和 μ_B 存在显著差别”？

我们用R软件的直方图命令，把两个城市的总体分布画出来，放在一起比较。可以看出，A城的均值 μ_A ，也就是A城分布对称轴的位置，是6000元，B城 μ_B 是6500元， μ_A 和 μ_B 相差500元。我们的目的是比较两个城市平均工资的差值，那么，这个差值500元，就是我们检出的信号。此时，信号等于500元，我们是否可以说，两个总体的均值存在显著差别呢？



我只能说，不一定。例如，我们只看A城的总体分布，其标准差 $\sigma_A=500$ 元。根据“68-95-99”法则，整个总体99%的数据，横跨6个 σ ，跨度达3000元之多。那么，仅在A城自己内部，随机抽取两个工资数据的话，相差500元以上，都是大概率事件。那么， μ_A 和外部的 μ_B ，也只相差个500元，怎么好意思说“差别显著”呢？



假如我们把 $\mu_A - \mu_B = 500$ 元当作信号，然后，为讲解方便，把信噪比公式中的噪音简化为1个标准差，那么，信噪比等于1，信号和噪声一样大，所以，不能说差别显著。请注意，真实的t检验中的信噪比公式，我们都学过的，是比这个复杂得多的，待会我们马上讲到。

$$\mu_A = 6000 \quad \mu_B = 6500$$

$$\mu_A - \mu_B = 500 \text{ (信号)}$$

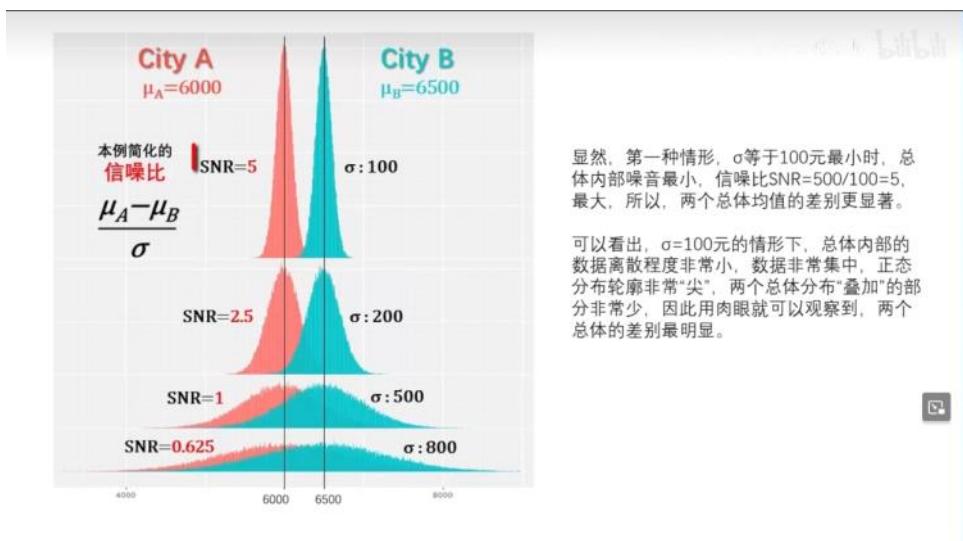
$$\sigma_A = \sigma_B = \begin{cases} 100 \\ 200 \\ 500 \\ 800 \end{cases} \quad \begin{array}{l} \text{方差齐性} \\ \text{内部噪音 改变} \\ 4 \text{ 种情形} \end{array}$$

接下来，我们把这个例子扩展变化一下。仍然保持 $\mu_A=6000$ 元， $\mu_B=6500$ 元不变，也就是信号500元不变，但改变标准差，也就是改变内部噪音。注意，为了保持A城和B城具有“方差齐性”，我们仍然把A城和B城的标准差设置成一样的。我们把总体标准差分别设成100元，200元，500元，800元，再来比较两个总体的分布。



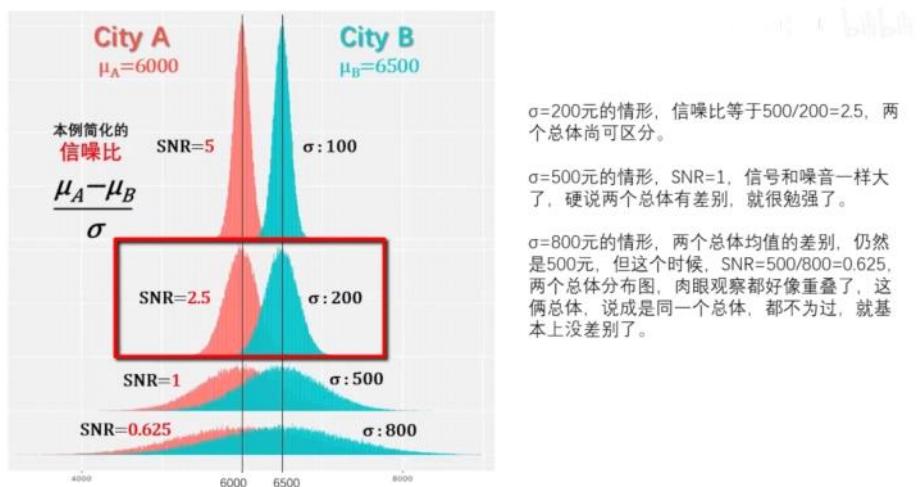
如图所示，四个直方图的坐标尺度完全一致，可以直接相比，可谓一目了然。

四种情形下， μ_A 和 μ_B 的差值，也就是信号，都是500元。但是，四种情形的内部噪音，可谓有天壤之别。



显然，第一种情形， σ 等于100元最小时，总体内部噪音最小，信噪比SNR=500/100=5，最大，所以，两个总体均值的差别更显著。

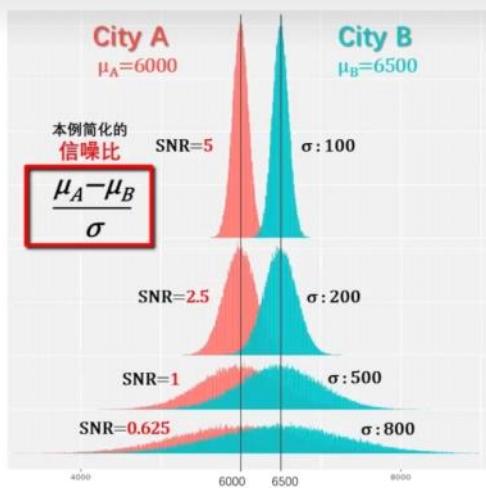
可以看出， $\sigma=100$ 元的情形下，总体内部的数据离散程度非常小，数据非常集中，正态分布轮廓非常“尖”，两个总体分布“叠加”的部分非常少，因此用肉眼就可以观察到，两个总体的差别最明显。



$\sigma=200$ 元的情形，信噪比等于500/200=2.5，两个总体尚可区分。

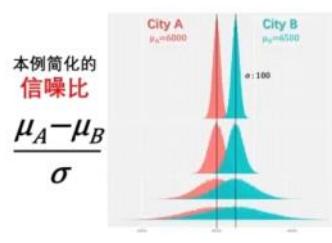
$\sigma=500$ 元的情形，SNR=1，信号和噪音一样大了，硬说两个总体有差别，就很勉强了。

$\sigma=800$ 元的情形，两个总体均值的差别，仍然是500元，但这个时候，SNR=500/800=0.625，两个总体分布图，肉眼观察都好像重叠了，这两个总体，说成是同一个总体，都不为过，就基本上没差别了。



以上，就是信噪比在统计学中，比较两个总体均值是否有显著差别时的应用。信号，就是两个总体均值的差值；噪音，就是总体本身的数据离散程度，也就是标准差 σ 。

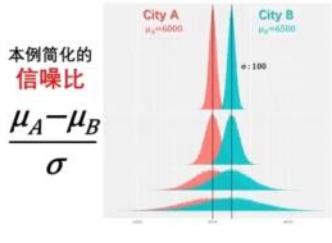
当然，刚才说了，上面例子中，信噪比公式是故意简化过的，分母中只除以一个 σ 。



双样本t检验 信噪比

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad \begin{matrix} \text{信号} \\ \text{噪音} \end{matrix}$$

而真正的t检验中，信噪比是这样的。我们以双样本t检验的t值公式为例。此时，请大家心中明白，t检验时，我们无法获得总体数据，所以只能通过抽样，用样本来估计总体。于是，公式中的分子上，用样本均值去估计总体均值，再用两个样本均值的差异，去估计两个总体均值的差异。于是，分子上就是信号。



双样本t检验 信噪比

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad \begin{matrix} \text{信号} \\ \text{噪音} \end{matrix}$$

分母上，用两个样本的标准差 s 去估计两个总体的标准差 σ ，再用平方和开根号的形式，来综合出一个包含了两个总体内部噪音的数值。这个分母，在第16节“双样本t检验”中，我们介绍过了，叫做“双样本的均值差值分布的标准误”，这一长串术语，不好理解，现在，你可以把它当作两个总体的综合内部噪音，就可以感性理解了。

卡方拟合优度检验 卡方检验

2024年1月8日 10:34

卡方 检验

Chi Square Test

$$\chi^2$$



大家好，今天我们来学习一下卡方检验。卡方检验，Chi Square Test。“卡”，就是希腊字母χ。方，就是“平方”，square。



卡方 拟合优度 检验

Chi Square Test for
Goodness of Fit

卡方 独立性 检验

Chi Square Test of
Independence

入门课中，我们只介绍两种最基本的卡方检验：卡方拟合优度检验 (chi-square goodness of fit test) 和卡方独立性检验 (chi-square test of independence)。这两名字看起来也是相当的唬人，不过没关系，我们仍然来编故事、举例子。故事讲完以后，你就知道这些术语是什么意思了。本节课，我们先介绍较为简单的“卡方拟合优度”。

姥爷的小卖部

什锦糖 姥爷的勾兑比例

高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%

我小时候在农村里，我姥爷开了一个小卖部，里面啥都卖。快过年的时候，卖什锦糖。什锦糖，就是把各种糖掺和在一起卖。姥爷按照一个比例，勾兑了一种什锦糖。高粱饴，单价比较便宜，占比40%，大白兔20%，大虾酥20%，酒心糖15%，巧克力最贵，只占比5%。姥爷把糖掺和在一起，搅匀了，开始卖。各种糖的比例，姥爷公开贴出来，童叟无欺。

姥爷的小卖部

什锦糖 姥爷的勾兑比例

高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%
100多块糖:				13块
				4块

卖了一阵，有村民来反应，说，你这什锦糖的比例不对劲，说称了100多块糖，巧克力才4个，没有5个，不是5%，酒心糖也只有13个，没有15%。我姥爷怀疑，是不是我整天蹲在小卖部里，把巧克力都给挑出来偷吃了。

姥爷的小卖部

什锦糖 姥爷的勾兑比例

高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%
100多块糖:				13块
				4块

我说我冤枉啊，我没有偷。姥爷说，不要狡辩。我说，你掺和糖的时候可能没有搅匀。姥爷说，搅了好几遍，肯定都搅匀了。我又说，每次抓糖的时候，谁能保证这么正正好按照你勾兑的比例呢，多一个两三块，少一个两三块，都很正常。

```
> tang_Catagory <- c('高粱饴', '大白兔', '大虾酥', '酒心糖', '巧克力')
```

高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%

类别变量

性别: {男, 女} 有限

Categorical Variable

星期几: {Mon, Tue, Wed, Thu, Fri, Sat, Sun}

我们用R软件来模拟这个什锦糖的故事。首先指定糖的种类，一共有这5种。注意，我们趁机在这里引入一个概念，叫“类别变量”，Categorical variable。在这个故事中，类别变量就是“糖的种类”，这个变量只有5个取值范围，不可能有第6种糖。再例如：性别也是一个类别变量，一般只能取“男”和“女”两个值。星期几也是一个类别变量，只能取“周一”到“周日”7个值，不能取出一个“星期3.5”来。类别变量是可以穷尽的，取值是有限的。

```
> tang_Catagory <- c('高粱饴', '大白兔', '大虾酥', '酒心糖', '巧克力')
```



类别变量 Categorical Variable

性别: {男, 女} 有限

连续变量 Continuous Variable

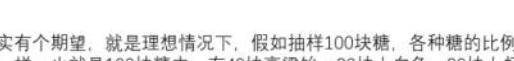
$$\text{分数: } 0, \underbrace{1, 10}_{0.4, 0.57} \text{ 无限}$$

与类别变量相对的，是“连续变量”，Continuous variables。之前t检验中的分数，就是连续变量。例如，0分1分、10分、等等。而任意两个数值中间，还可以取其他值。例如，0和1之间，还可以取0.4分；0.4和1之间，还可以取0.57分，等等。一般来说，连续变量是不能穷尽的，取值是无限的。

按姥爷比例
1000块糖的总体

高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%

为了给大家一个直观的印象，1000块糖的总体，造出来是这个样子的

假如抽样100块糖 期望频次					
	高粱饴	大白兔	大虾酥	酒心糖	巧克力
期望频次 Expectation	40	20	20	15	5
H_0 姥爷总体比例					
高粱饴	40%	20%	20%	15%	5%

在抽样之前，我们心中其实有个期望，就是理想情况下，假如抽样100块糖，各种糖的比例，正好和总体中真实的各种糖的比例一模一样。也就是100块糖中，有40块高粱饴，20块大白兔，20块大虾酥，15块酒心糖5块巧克力。我们把这5种类别的糖“应该”或者“期望”出现的次数，叫做我们对这个样本所期望的频次，英语叫做Expectation。

```
> smpl <- sample(x = tang Population, size=100)
> table(factor(smpl,levels=tang Catagory))
```

高粱饴 大白兔 大虾酥 酒心糖 巧克力
38 24 18 18 2

	高粱饴	大白兔	大虾酥	酒心糖	巧克力
期望频次 Expectation	40	20	20	15	5
观测频次 Observation	38	24	18	18	2

然后，我们从这1000块糖中，随机抽取100块糖，相当于姥爷把1000块糖搅匀了，再一把一把的抓，抓出

然后，我们从这1000块糖中，随机抽取100块糖，相当于姥爷把1000块糖搅匀了，再一把一把的抓，抓出100块糖。结果，这次抽样中，各种糖的块数如下。高粱饴38块，大白兔24块，大虾酥18块，酒心糖18块，巧克力2块。我们这个实际抽样中观测到的各种糖的次数，叫做观测频次，Observation。请注意，这里都是频次，次数，都是整数，不是比例，不是小数，也不是百分比。

类别变量的 总体概率分布

H₀ 原假设 总体

高粱饴	大白兔	大虾酥	酒心糖	巧克力	$\Sigma = 1$
40%	20%	20%	15%	5%	

1个样本
样本容量：100

高粱饴	大白兔	大虾酥	酒心糖	巧克力	
期望频次 Expectation	40	20	20	15	5
观测频次 Observation	38	24	18	18	2

现在，我们引入原假设H₀：总体中的各种糖的比例符合姥爷的勾兑比例。我们把这个比例，叫做“类别变量的总体概率分布”，注意，这里都是小数百分比，而不是频次。这几个百分比，加起来，必须等于1。那么，假如样本容量为100，这就是期望频次，和H₀中的总体分布比例没有差异。但实际观测频次，和期望频次显然是有差异的。

类别变量的 总体概率分布

H₀ 原假设 总体

高粱饴	大白兔	大虾酥	酒心糖	巧克力	$\Sigma = 1$
40%	20%	20%	15%	5%	

1个样本
样本容量：100

高粱饴	大白兔	大虾酥	酒心糖	巧克力	
期望频次 Expectation	40	20	20	15	5
观测频次 Observation	38	24	18	18	2

于是，我们自然会提出这样几个问题：在H₀为真的情况下，抽一次样，观测频次和期望频次之间的差异算不算极端呢？或者说，如何衡量这种差异的大小？再或者说，出现这种水平的差异，是否考虑拒绝H₀？假如，能把这个差异，给算成一个数值，就好办了。

	高粱饴	大白兔	大虾酥	酒心糖	巧克力
期望频次 Expectation	40 E	20	20	15	5
观察频次 Observation	38 O	24	18	18	2

卡方公式 $\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$

类别变量的种类
n=5

注意，公式中，n不是样本容量，而是类别变量的种类，本例中，有5种糖，所以，n=5。不同的教材上，使用的字母可能不同，请大家自行区别。每种类别的观测频次和期望频次之差的平方，除以本类别的期望频次，然后所有类别加起来，就得到了一个总的、观测频次和期望频次之间的差别。

	高粱饴	大白兔	大虾酥	酒心糖	陈年陈皮饼	巧克力
期望频次 Expectation	40	20	20	15	5	
观察频次 Observation	38	24	18	18	2	

卡方公式 $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

类别变量的种类

$n=5$

和方差类似，这里差值也是用的平方，也是因为差值有正有负，加起来可能会抵消，所以用平方的形式。当然，你问我为什么不用绝对值，我回答不了。分子上，除的是期望频次，你也可能问我，为什么不除以期望频次的平方，我也回答不了。大家可以去读一下Karl Pearson的论文原作，本节课就不讨论了。

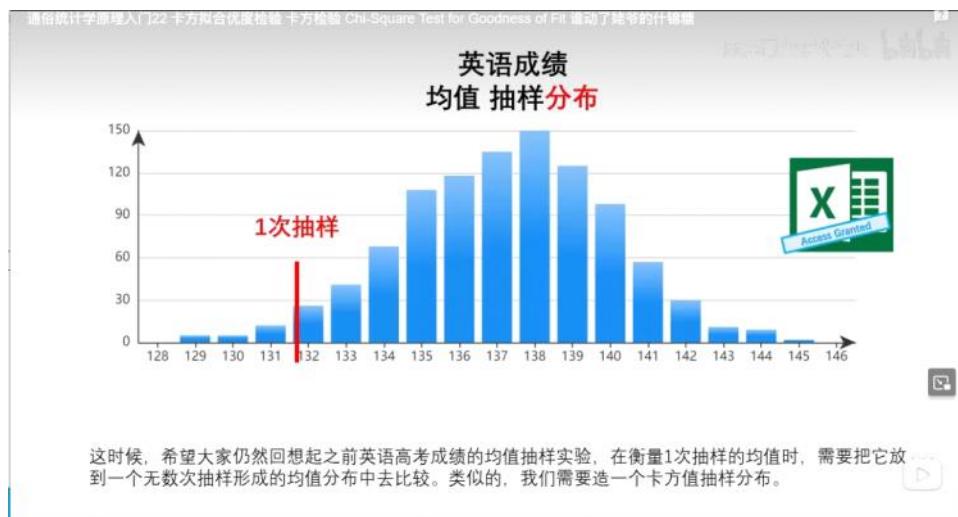
	高粱饴	大白兔	大虾酥	酒心糖	陈年陈皮饼	巧克力
期望频次 Expectation	40	20	20	15	5	
观察频次 Observation	38	24	18	18	2	

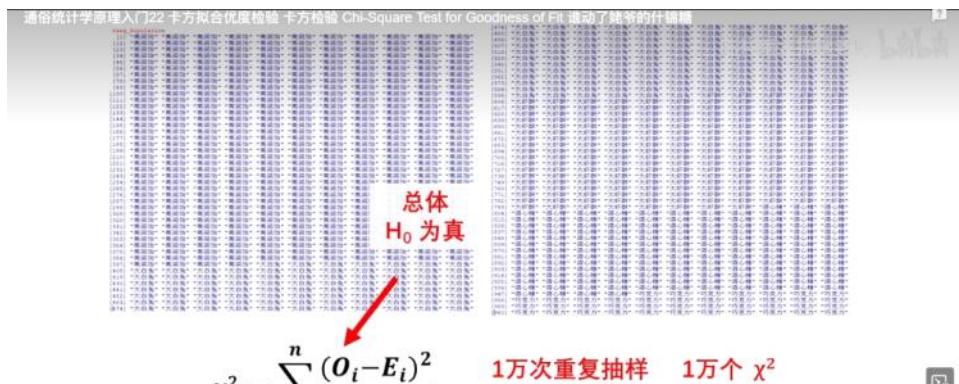
卡方公式 $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

$$\chi^2 = \frac{(38-40)^2}{40} + \frac{(24-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(18-15)^2}{15} + \frac{(2-5)^2}{5}$$

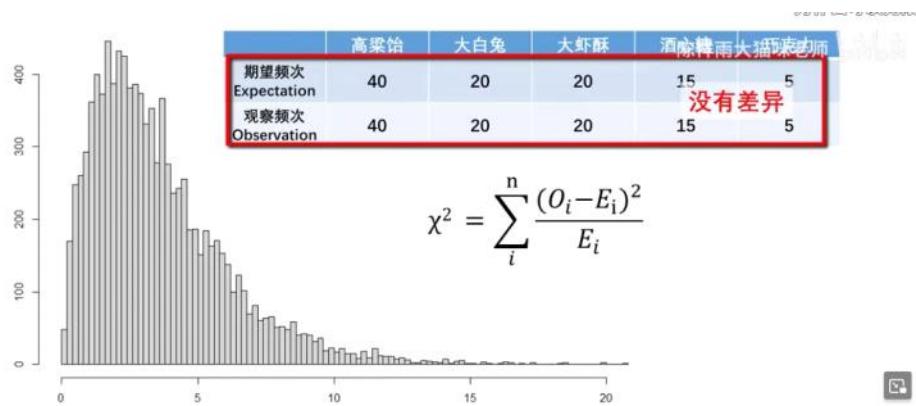
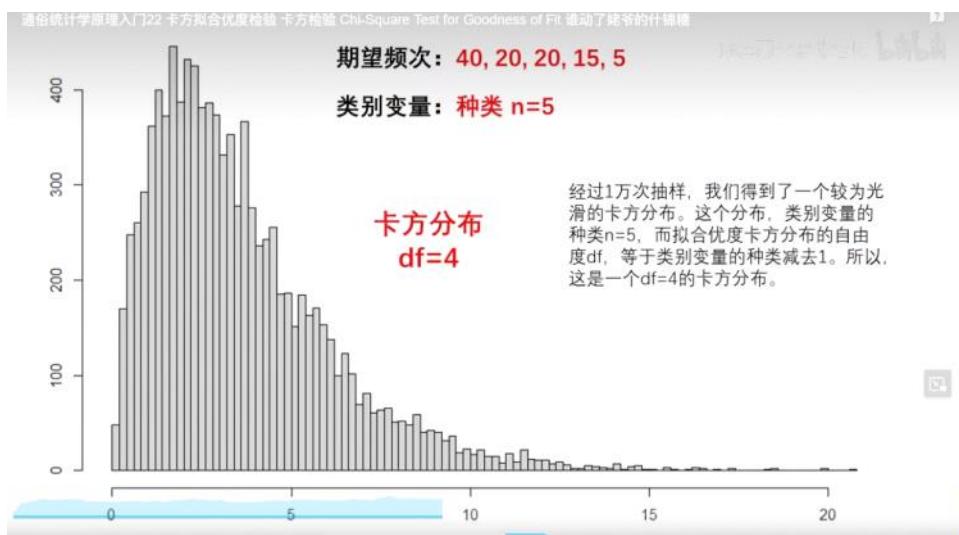
$$= 0.1 + 0.8 + 0.2 + 0.6 + 1.8 = 3.5$$

我们把数据代入公式，公式展开是这个样子。计算得卡方等于3.5。1次抽样，样本中的所有数据浓缩成一个3.5。这个3.5代表着什么呢？

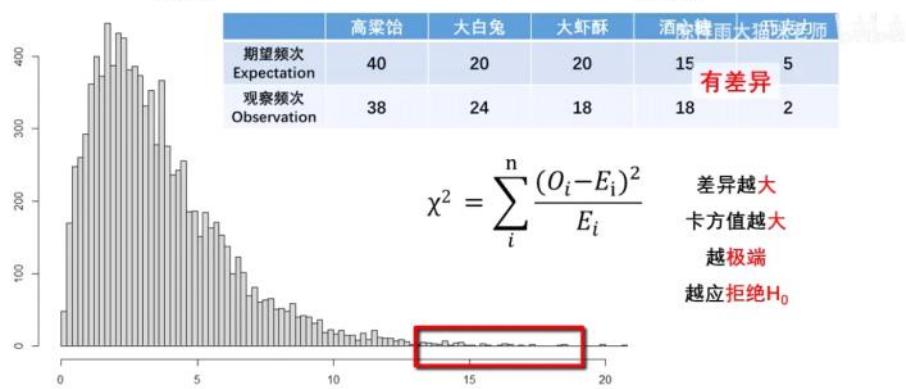




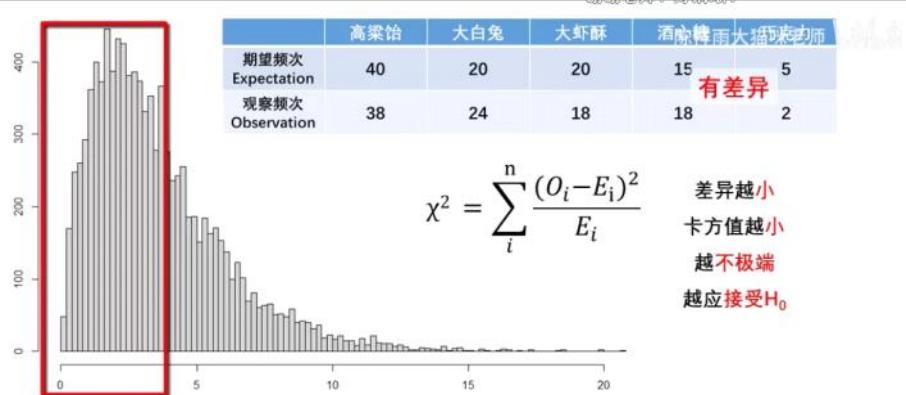
我们从这个 H_0 为真的 1000 块糖的总体中，每次抽 100 块糖，按照这个公式，计算出 1 个卡方值。假如进行 1 万次重复抽样，就可以得到 1 万个卡方值，把 1 万个卡方值，做成直方图，便可以得到卡方分布。



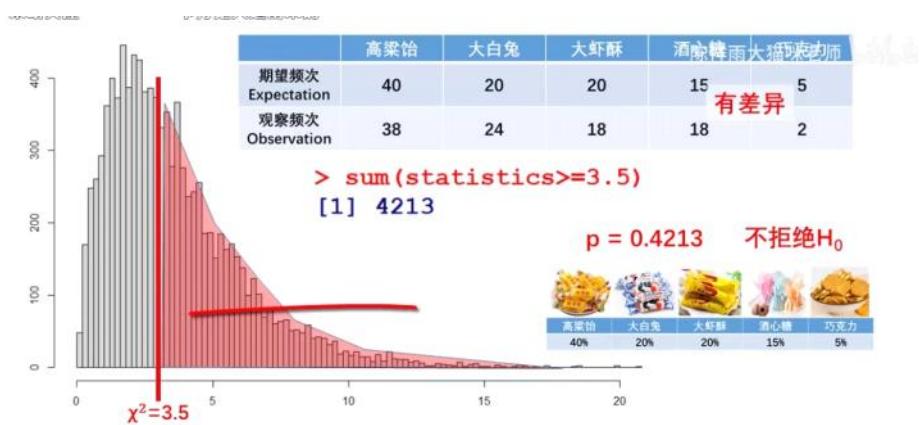
我们来分析一下这个卡方分布。假设每次抽样的观测频次，和期望频次完全一样的话，那么 O_i 等于 E_i ，卡方值都应该是 0。但实际上，由于抽样的随机性，卡方值等于 0 的概率非常小。和均值分布类似，卡方值也是以一种概率分布的形式存在的。由于公式限定，卡方值不可能是负的，所以，卡方值最小就是 0。不过，卡方值理论上可以无限大，所以，卡方分布，是朝右边的尾巴尖倾斜的，英语叫 positively skewed。



卡方值越大，说明观测频次和期望频次的差异越大，所以样本越极端，越应该拒绝 H_0 。



如果卡方值越小，则观测频次越接近期望频次，所以样本越不极端，越应该接受 H_0 。于是，卡方分布中，拒绝域只在右边，不在左边。所以，入门课中，我们姑且把卡方检验看成一个天然的单边检验。



有了拒绝域的方向，我们来算p值。刚才抽样一次，得卡方=3.5，于是在卡方等于3.5这里画一条线，比3.5（含）还极端的样本数量占总共10000次抽样的百分比，就是p值。我们用sum函数，来统计比3.5还大的次数，得4213次。则卡方=3.5的p值为0.4213。假如 α 设定为0.05的话， $p > \alpha$ 。于是，这个样本不极端，不显著，于是无法拒绝 H_0 ，认为总体中类别变量的分布，仍然符合姥爷勾兑的比例。于是，我就是清白的了，我没有偷吃巧克力。

卡方临界值表

> sum(statistics)>=3.5)

[1] 4213

	0.995	0.99	0.975	0.95	0.9	0.5	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000397	0.000157	0.000982	0.00393	0.0158	0.455	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.020	0.051	0.103	0.211	1.386	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.072	0.115	0.216	0.352	0.584	2.366	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.297	0.484	0.711	1.064	3.357	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
	0.412	0.554	0.831	1.145	1.610	4.354	7.269	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	0.872	1.237	1.635	2.204	5.348	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.239	1.690	2.167	2.833	6.346	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	1.646	2.180	2.733	3.490	7.344	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.088	2.700	3.325	4.168	8.343	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	2.558	3.247	3.940	4.865	9.342	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588

刚才的p值，是用R程序数出来的。我们也可以通过查表来估计p值。和t临界值表类似，卡方分布，也有临界值表。在计算机还没有普及的日子里，我们不容易算出一个精确的p值，所以查表是最方便的。本例中， $df=4$ ，于是我们先找到 $df=4$ 这一行。然后，算出来的卡方值是3.5，发现3.5介于表上的3.357和5.989之间。于是，3.5对应的p值应该介于0.5和0.2之间，这和我们估计的0.4213是一致的。

通过统计学原理入门 22.2.1 方差分析与卡方检验

类别变量的 总体分布

H_0 原假设 总体

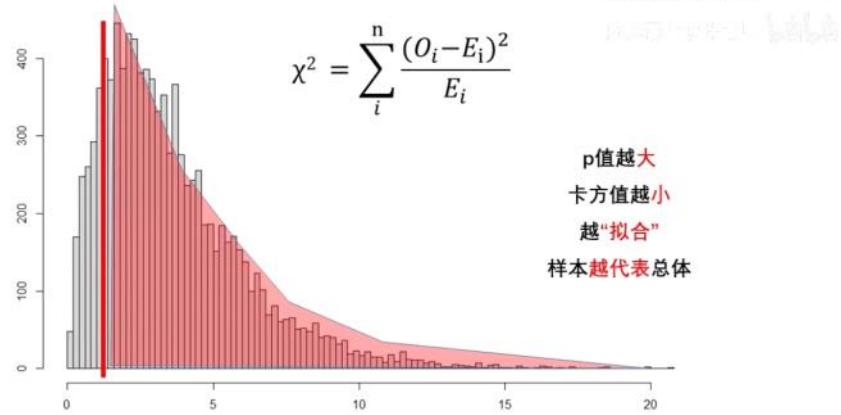
	高粱饴	大白兔	大虾酥	酒心糖	巧克力
40%	20%	20%	15%	5%	

1个样本 样本容量: 100

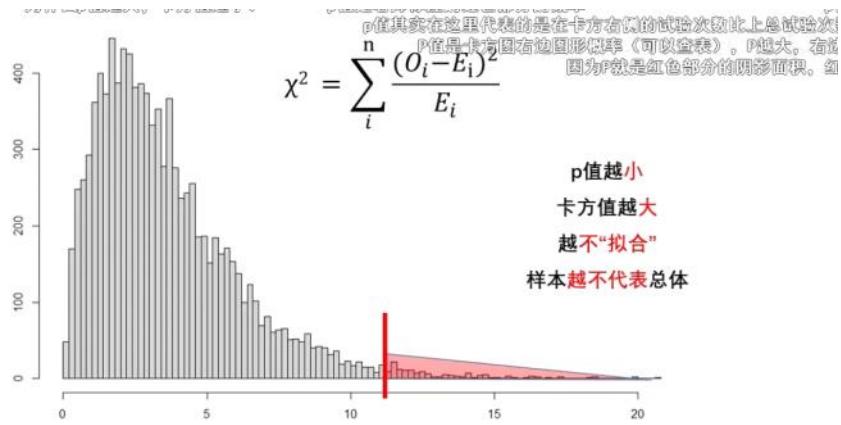
	高粱饴	大白兔	大虾酥	酒心糖	巧克力
期望频次 Expectation	40	20	20	15	5
观测频次 Observation	38	24	18	18	2

“拟合优度” 观测频次，是否“符合”了期望频次？
多么好

现在，我们回过头来，看一下什么叫做“拟合优度”。 H_0 中总体的类别分布是这样。因此，样本容量为100的话，期望频次是这样的。然后，观测频次又是这样的。那么，观测频次，是否“符合”了期望频次呢？这个“符合”，就是“拟合”。“优度”，就是指观测频次，有多么好的“符合”了期望频次，或者说，样本有多么好的代表了总体。这就是“拟合优度”的含义。



而“拟合优度”，是可以用p值来表示的。p值越大，说明样本的卡方值越小，观测频次和期望频次之间的差异也越小，样本越能拟合期望，样本越能代表总体。



反之，p值越小，说明样本卡方值越大，观测频次和期望频次之间的差异也越大，样本越不符合期望，样本越不能代表总体。



当然，各种类别之间的分布比例，也不能“极端不均衡”。例如，各种糖的分布比例是96%，1%，1%，1%，1%。这就比较极端了。在这样的总体中，就不容易抽出一个卡方分布。因为除了高粱饴之外，抽到其他糖的概率都很小。这属于**研究问题本身存在问题**，假如我姥爷按这个比例来勾兑什锦糖，我觉得**这就叫什锦糖了**，这叫一大袋高粱饴里面，不小心掉进去几块其他种类的糖。就这，还要进行卡方抽样实验，没有太大的意义。



所以，卡方拟合优度检验，一般有一个条件，即，每种类别的频次，一般不小于5。100块糖中，假如只能抽出一两块大白兔，一两块大虾酥等等，这些糖，就不要单独设一个类别了，干脆合并起来，叫做“其他种类”就完事了。

类别变量的 总体概率分布					
H_0 原假设	总体	百分比 小数			
		高粱饴	大白兔	大虾酥	酒心糖
		40%	20%	20%	15%
		40	20	20	15
		38	24	18	18
		5	2		

1个样本 样本容量: 100	高粱饴	大白兔	大虾酥	酒心糖	巧克力	频次 次数
	期望频次 Expectation	40	20	20	15	
	观测频次 Observation	38	24	18	18	2

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

第二个需要注意的地方。在“类别变量”的总体比例分布中，各种类别用的是百分比，percentage。百分比，就是概率，probability。所以，姥爷的勾兑比例，就是“类别变量”的概率分布，这个分布，是对总体而言的。但是，在抽样时，无论是期望频次，还是观测频次，用的都是“次”，“次数”，英语叫times，或者frequency。频次，不是百分比，而是数出来的次数。而卡方公式中，用的都是“频次”，而不是用的百分比。这些，大家一定要注意，不要用错。

	高粱饴	大白兔	大虾酥	酒心糖	巧克力	卡方=3.5
期望频次 Expectation	40	20	20	15	5	$p = 0.4$
观测频次 Observation	38	24	18	18	2	不拒绝 H_0

	高粱饴	大白兔	大虾酥	酒心糖	巧克力	卡方=35
期望频次 Expectation	400	200	200	150	50	$p < 0.001$
观测频次 Observation	380	240	180	180	20	拒绝 H_0

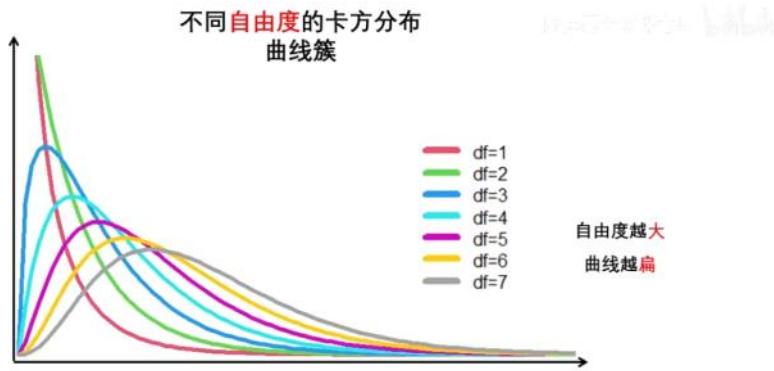
我们举个例子，刚才是样本容量为100，抽样频次是这样的，卡方值是3.5，p值0.4左右，不拒绝 H_0 。假如我们把这个样本中各类糖的频次，扩大10倍，这样，就得到了一个样本容量为1000的样本，样本中各类糖的比例，和刚才是一样的。但是，卡方值算出来，就变成了35了。卡方值35，查表一看，就是非常极端的抽样了，要拒绝 H_0 了。这在直观上也是好理解的，抽样1000块糖，就是把总体全给抽了，总体里一共才20块巧克力，姥爷说好的50块呢？所以，肯定要拒绝 H_0 。

	高粱饴	大白兔	大虾酥	酒心糖	巧克力	卡方=3.5
期望频次 Expectation	40	20	20	15	5	$p = 0.4 > 0.05$
观测频次 Observation	38	24	18	18	2	不拒绝 H_0

	高粱饴	大白兔	大虾酥	酒心糖	巧克力	卡方=35
期望频次 Expectation	400	200	200	150	50	$p < 0.001$
观测频次 Observation	380	240	180	180	20	拒绝 H_0

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

所以，卡方公式中，各个数据，都是频次，而不是比例。比例相同的两个样本，并不能算出相同的卡方值。其根源在于，卡方公式中，分子上是平方，分母上没有平方，这都是纯数学运算的东西，大家可以自己思考下，本节课就不详细展开了。



刚才的故事中，姥爷勾兑的是5种糖，得到的是 $df=4$ 的卡方分布。假如姥爷勾兑的是6种糖的，就可以得到 $df=5$ 的卡方分布。不同自由度的卡方分布，是一组形状类似、但又互相区别的曲线簇。在这个图中，可以看到，自由度越大，卡方曲线越扁，且倾斜度越低，峰值越朝右边移动。

期望的总体类别概率分布
> `expected_p <- c(0.4,0.2,0.2,0.15,0.05) Σ =1`

样本的观测频次
> `chisq.test(x=c(38,24,18,18,2),p=expected_p)`

以上讲解了卡方拟合优度的概念，和卡方分布曲线的产生。下面，我们在R软件中进行卡方拟合优度检验。其程序命令非常简单，即`chisq.test()`命令。拟合优度检验，只需要两个参数，一个是样本的观测频次，一个是期望的总体类别概率分布。需要注意的是，这几个概率加起来，要等于1，否则程序会报错。

```
期望的总体类别概率分布  
> expected_p <- c(0.4,0.2,0.2,0.15,0.05) Σ =1
```

```
样本的观测频次  
> chisq.test(x=c(38,24,18,18,2),p=expected_p)
```

```
Chi-squared test for given probabilities  
  
data: c(38, 24, 18, 18, 2)  
X-squared = 3.5, df = 4, p-value = 0.4779  
  
> sum(statistics>=3.5)  
[1] 4213
```

结果显示，这是一个针对给定概率分布的卡方检验，也就是拟合优度检验。统计量卡方值算出来是3.5，和我们刚才手工算的一样。自由度，等于分类变量的种类减去1，等于4。p值等于0.4779，和我们刚才统计出来的0.4213差不多。

例1：学校里有4个食堂，全校师生是否同等的喜欢4个食堂。

H_0 ：全校师生 同等的喜欢 4个食堂。

	第1食堂	第2食堂	第3食堂	第4食堂
H_0 总体分布	0.25	0.25	0.25	0.25
观测频次 Observation	31	37	30	40

调查问卷：“你最喜欢的食堂是哪一个”

	第1食堂	第2食堂	第3食堂	第4食堂
观测频次 Observation	31	37	30	40

下面，我们举两个卡方拟合优度检验的例子。在社会统计中，我们经常做调查问卷。例如，学校里有4个食堂，我想调研下，全校师生是否同等的喜欢4个食堂。调查问卷的问题是，“你最喜欢的食堂是哪一个”，选项是4个食堂。这时，我们的 H_0 是：全校师生同等喜欢4个食堂。也就是说，4个食堂被选为“最喜欢的食堂”的概率都是0.25。

例1：学校里有4个食堂，全校师生是否同等的喜欢4个食堂。

H_0 ：全校师生 同等的喜欢 4个食堂。

	第1食堂	第2食堂	第3食堂	第4食堂
H_0 总体分布	0.25	0.25	0.25	0.25
观测频次 Observation	31	37	30	40

调查问卷：“你最喜欢的食堂是哪一个”

	第1食堂	第2食堂	第3食堂	第4食堂
观测频次 Observation	31	37	30	40

我们在校园里随机采访路人，抽样得到这么一组数据。假如我们没学过卡方拟合优度检验，我们可能会说，第4食堂喜欢的人最多啊，第4食堂应该是最受喜欢的食堂。但我们学了卡方检验，就在R软件中检验一下吧，看看这个较多的40是否只是一种随机现象。

例1：学校里有4个食堂，全校师生是否同等的喜欢4个食堂。

H_0 ：全校师生 同等的喜欢 4个食堂。

```
观测频次  
> chisq.test(x = c(31,37,30,40), p = c(0.25,0.25,0.25,0.25))  
Chi-squared test for given probabilities  
  
data: c(31, 37, 30, 40)  
X-squared = 2, df = 3, p-value = 0.5724
```

我们在程序中输入观测频次，和 H_0 的总体概率分布。注意，期望频次，我们不需要管的。程序会根据观测频次算出样本容量，然后再按这个概率分布，算出期望频次。结果显示， p 等于0.5724。这说明在 H_0 为真的情况下，也就是全校师生对4个食堂同等喜爱的情况下，抽出这么一组结果，一点都不极端。第4食堂的这个40，并不能说明全校师生显著的更喜欢第4食堂。所以，结论就是， H_0 ：全校师生对4个食堂同等喜欢。

例2：判断此色子是不是均匀的 (fair die)

H_0 ：色子抛出各个点数的概率相同 (均匀的)。

	1	2	3	4	5	6
H_0 点数分布	1/6	1/6	1/6	1/6	1/6	1/6



抛色子60次

	1	2	3	4	5	6
观测频次	9	12	11	8	12	8

再举一个例子。概率学中，抛硬币，抛色子，都是最古典的实验模型。之前我们说过，硬币要均匀的，那么，色子，也得是均匀的。不均匀的，那叫出老千。如何判断一个色子是不是均匀的呢。对于均匀的色子，我们心中都有个期望，就是抛出6种点数的概率都是一样的。那么， H_0 就是：每个点数的概率都是六分之一。我们抛同一个色子60次，得到这样一组观测频次，看起来不是那么理想。理想的话，每个点数都该是10次。那么，我们是否就说色子不均匀呢？

例2：判断此色子是不是均匀的 (fair die)

观测频次
> chisq.test(x=c(9,12,11,8,12,8), p=c(1/6,1/6,1/6,1/6,1/6,1/6))
Chi-squared test for given probabilities

data: c(9, 12, 11, 8, 12, 8)
X-squared = 1.8, df = 5, p-value = 0.8761



我们就用卡方拟合优度检验，把数据代入程序中。这是期望的概率分布，这是观测频次。结果显示，这是一个卡方拟合优度检验，卡方值等于1.8，自由度等于6-1=5，p值等于0.8761。p值非常不显著，所以，不拒绝 H_0 。结论，色子是均匀的。

卡方独立性检验

2024年1月8日 19:33



本节课，我们来讲解另外一种卡方检验，卡方独立性检验。两者的差别就是，拟合优度检验中，只有一个类别变量，而独立性检验中，有两个类别变量。大家暂时听不懂没关系，我们还是先来编故事。

猜测：不同年龄的村民，对于不同种类的糖，有着不同的偏好

年龄群体 VS 糖果种类



我们仍然借着姥爷卖什锦糖的故事往下编。姥爷在卖糖的过程中发现，不同年龄的村民，对于不同种类的糖，有着不同的偏好。例如，姥爷觉得，村里的老头老太太们，大多喜欢吃高粱饴；村里的小孩子们，大多喜欢吃巧克力。

姥爷的调查问卷

1. 请选择您的 **年龄群体**

A. 小孩 B. 中年人 C. 老年人

2. 请选择您最喜欢的 **糖果种类**

A. 高粱饴 B. 酒心糖 C. 大虾酥 D. 巧克力



姥爷想验证一下自己的想法，于是就在村里做了调查问卷。注意，姥爷这次不是卖什锦糖了，而只是做了一个调查问卷，调查一下不同年龄群体对于不同糖果种类的偏好。为了方便讲解，这个故事里，只有3种年龄群体，和4种糖果种类。

姥爷的调查问卷

1. 请选择您的 **年龄群体**
A. 小孩 B. 中年人 C. 老年人

2. 请选择您最喜欢的 **糖果种类**
A. 高粱饴 B. 酒心糖 C. 大虾酥 D. 巧克力

序号	年龄群体	陈祥 最喜欢种类
1	小孩	巧克力
2	中年人	巧克力
3	小孩	大虾酥
4	老年人	高粱饴
5	老年人	大虾酥
6	中年人	酒心糖
.....
198	小孩	巧克力
199	老年人	高粱饴
200	中年人	酒心糖

姥爷在村里调查了200个人，或者说，做了一个200人的抽样，一共收集到了200份调查问卷。结果是这样的。当然，数据这样展示的话，是很难看出什么名堂来的。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	56	6	5	7	74
中年人	4	12	13	5	34
小孩	4	12	12	64	92
总计	64	30	30	76	200

Contingency Table (列联表)

我们稍加排序整理，得到一个这样的表格。这个表叫做“列联表”，英语叫做Contingency table。本次抽样，共调查200人，其中老年人74人，中年人34人，小孩92人，总计200人。其中，最喜欢高粱饴的64人，最喜欢酒心糖的30人，最喜欢大虾酥的30人，最喜欢巧克力的76人，总计也是200人。注意，表里的数据，都是频次，次数，不是百分比。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	56	6	5	7	74
中年人	4	12	13	5	34
小孩	4	12	12	64	92
总计	64	30	30	76	200

姥爷观察：老年人大多喜欢高粱饴，小孩大多喜欢巧克力

横着看，竖着看，都比较符合姥爷一开始观察到的：即老年人大多喜欢高粱饴，小孩大多喜欢巧克力。例如，横着看，74个老年人里面，喜欢高粱饴的最多；竖着看，64个喜欢高粱饴的人里面，老年人最多。再例如，横着看，92个小孩里面，喜欢巧克力的最多；竖着看，76个喜欢巧克力的人里面，小孩最多。

		类别变量 1 糖果种类				总计
		高粱饴	酒心糖	大虾酥	巧克力	
类 别 变 量 2	老年人	56	6	5	7	74
	中年人	4	12	13	5	34
	小孩	4	12	12	64	92
	总计	64	30	30	76	200

Contingency Table (列联表)

这个表里，有两个类别变量。一个是糖果种类，一个是年龄群体。糖果种类，只有4种取值，也就是这4种糖果。年龄群体，只有3种取值，分别是3类人群。

		类别变量 1 糖果种类				总计
类 别 变 量 2	老年人	高粱饴	酒心糖	大虾酥	巧克力	
		56	6	5	7	74
		4	12	13	5	34
		4	12	12	64	92
	总计	64	30	30	76	200

猜测：不同年龄的村民，对于不同种类的糖，有着不同的偏好

现在，我们回到姥爷的猜测：姥爷认为，各年龄群体对各糖果种类的偏好不一样。如何用数据来表达“偏好”这一概念呢？本列联表中的数据，都是频次，次数。而次数，并不是一个很好的表达“偏好”的数据形式。例如，中年人和小孩，最喜欢酒心糖的，都是12人；但这并不能代表中年人和小孩，对酒心糖的偏好是一样的。因为本次抽样的200人中，中年人只有34人，而小孩则多达92人，12人的频次虽然相同，但比例并不相同。而“比例”这个数据形式，比“频次”，更能表达“偏好”这一概念。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	56	6	5	7	74
中年人	4	12	13	5	34
小孩	4	12	12	64	92

频次

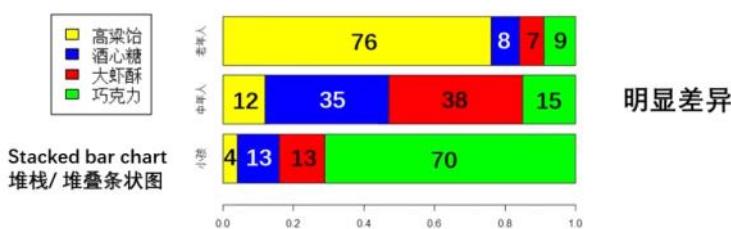
用“百分比”衡量“偏好”

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	76%	8%	7%	9%	100%
中年人	12%	35%	38%	15%	100%
小孩	4%	13%	13%	70%	100%

百分比

所以，我们把这个频次表，换成百分比表。有两点需要注意：第一，百分比，是横向的百分比，或者说，是按各个年龄群体进行计算的百分比。例如，这个76%是由56除以74算出来的，指总共74个老年人中，有76%最喜欢高粱饴。第二，为了方便讲解，这里的百分比，都进行了近似取整。下面，为了更方便的引入卡方独立性检验，我们再把这个样本的百分比表，做成柱状图。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	76%	8%	7%	9%	100%
中年人	12%	35%	38%	15%	100%
小孩	4%	13%	13%	70%	100%



如图所示，这就是姥爷调查问卷的数据，按各类年龄群体的偏好百分比，制成的条状图。这种图叫堆栈或堆叠条状图，英文叫stacked bar chart。每个长条，代表一类年龄群体，长条的长度是100%。



实际上
全村 总体
不同 年龄群体
对不同 种类的糖
偏好 一模一样

但是，我们是学过假设检验的。我们知道，这只是一个抽样而已，而样本是随机产生的，很有可能代表不了总体的真实情况。于是，我们应该猜测，有没有可能，“实际上，全村总体来看，不同年龄的村民，对于不同种类的糖，有着一模一样的偏好，只不过这一次抽样，恰好表现为偏好不一样”呢？



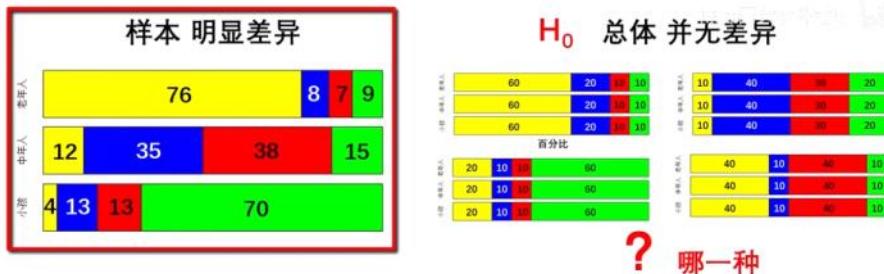
H_0 ：各“年龄群体”中，各“糖果种类”的 百分比/偏好，没有差异

很多人脑海中，会浮现出这么一个图。既然，各类年龄群体对4种糖的偏好百分比没有差异，那么，每种糖的偏好就是100%除以4，等于25%吧。这就是最常见的一个，误区，也是本节课理解的一个难点。注意，没有差异，并不等于平均分配4种糖的偏好。平均25%，只是一种特殊情况而已。那不特殊的情况是什么样的呢？这时候，语言表达有点无力，我们直接看图。



H_0 ：各“年龄群体”中，各“糖果种类”的 百分比/偏好，没有差异

例如，这4种情形，都是满足 H_0 的，即“各年龄群体中，各糖果种类的百分比/偏好，没有差异”。这4种情形中的比例，都不是25%平分的。这还只是4种情形，实际上可以存在无数种情形。



H_0 ：各“年龄群体”中，各“糖果种类”的 百分比/偏好，没有差异

那么，问题又来了。既然 H_0 有无数种情形，那本次抽样，应该属于哪一种情形呢？换句话说，本次抽样的 H_0 ，各种糖的偏好百分比，应该是多少呢？

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	56	6	5	7	74
中年人	4	12	13	5	34
小孩	4	12	12	64	92
总计	64	30	30	76	200

if 各 年龄群体 之间 没有差异
then 不考虑 年龄群体

这个问题，其实也不难。我们再回来看这个表，这是本次抽样的频次表。注意我下面说的这句话，假如我们想得到一个，各年龄群体之间，没有差异的一个百分比，那我们就干脆。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老人	64	6	5	74	200
中年人	4	12	12	5	34
小孩	4	12	12	54	92
总计	64	30	30	76	200

频次表

if 各 年龄群体 之间 没有差异
then 不考虑 年龄群体



把年龄群体这个类别变量给抹去，不考虑年龄群体，只看各类糖果的总计就好了。于是，我们得到这样一个表格。样本中，所有200个人，不再分老人，中年人，或小孩了，所有人都是一样的人。这200个不分类别的人，对各类糖果的偏好频次，是这样的。

	高粱饴	酒心糖	大虾酥	巧克力	总计
老人	64	6	5	74	200
中年人	4	12	12	5	34
小孩	4	12	12	54	92
总计	64	30	30	76	200
百分比	32%	15%	15%	38%	100%

if 各 年龄群体 之间 没有差异
then 不考虑 年龄群体



我们再把频次，换成百分比。例如，高粱饴被选了64次，占200的32%，酒心糖被选了30次，占200的15%，以此类推。这个百分比，就是本次抽样，不考虑年龄群体这个类别变量，或者说，各类年龄群体没有差异时，各类糖的偏好百分比。

样本 明显差异

H_0 总体 并无差异

老人	老人				总计	百分比
	高粱饴	酒心糖	大虾酥	巧克力		
老人	76	8	7	9	200	100%
中年人	12	35	38	15		
小孩	4	13	13	70		
					64	32%
					30	15%
					30	15%
					76	38%
					200	100%

H_0 ：各“年龄群体”中，各“糖果种类”的 百分比/偏好，没有差异

那么，我们就用这个偏好百分比，来作为 H_0 中各种糖的偏好比例。于是，这就是我们的 H_0 ：全村总体中，各年龄群体，对各糖果种类的偏好，都是这个比例，没有差异。那么，在这样一个没差异的总体中，抽到这样一个有明显差异的样本的概率是多少呢？



样本 明显差异					H ₀ 总体 并无差异				
老人人					老人人				
中年人					中年人				
小孩					小孩				
总计					总计				
百分比					百分比				

量化差别

此样本的p值

希望这时大家已经明白了，这就是一个假设检验求p值的问题。如何求出p值呢？首先，我们要量化样本和H₀总体之间的差别。如何量化呢？

样本 明显差异					H ₀ 总体 并无差异				
高粱饴 酒心糖 大虾酥 巧克力					高粱饴 酒心糖 大虾酥 巧克力				
老年人	56	6	5	7	老年人	32	15	15	38
中年人	4	12	13	5	中年人	32	15	15	38
小孩	4	12	12	64	小孩	32	15	15	38
总计	64	30	30	76	总计	64	30	30	200

观测 频次

期望 百分比

这时，请回想一下拟合优度检验中，已经引入的两个概念，期望频次，和观测频次。左图是本次样本的观测频次。右图是由观测频次推算出的，H₀中各种糖的期望百分比。例如，32%是由64除以200得来的。那么，如何用期望百分比，算出期望频次呢？

样本 观测 频次					H ₀ 期望 百分比				
高粱饴 酒心糖 大虾酥 巧克力					高粱饴 酒心糖 大虾酥 巧克力				
老年人	56	6	5	7	老年人	32	15	15	38
中年人	4	12	13	5	中年人	32	15	15	38
小孩	4	12	12	64	小孩	32	15	15	38
总计	64	30	30	76	总计	64	30	30	200

老年人

中年人

小孩

高粱饴 酒心糖 大虾酥 巧克力

其实不难，就是简单的乘除法。例如，本次抽样中，老年人4种糖一共选了74次，高粱饴的期望百分比是32%，于是期望频次就是 $74 \times 32\% = 23.68$ ，老年人酒心糖的期望百分比是15%，于是期望频次就是 $74 \times 15\% = 11.1$ 。以此类推，各类糖的期望频次，就都可以算出来了。假如你听着有点迷糊，请按下暂停键，把每个单元格中期望频次的计算公式，都仔细看下。

	高粱饴	酒心糖	大虾酥	巧克力
老年人	74*32% =23.68	74*15% =11.1	74*15% =11.1	74*38% =28.12
中年人	34*32% =10.88	34*15% =5.1	34*15% =5.1	34*38% =12.92
小孩	92*32% =29.44	92*15% =13.8	92*15% =13.8	92*38% =34.96

现在，我们把期望频次的计算公式，推广到一般情形。例如，这里的 $74 \times 32\%$ 中， 32% ，是 $64/200$ 得来的。于是， $74 \times 32\%$ 可以写成 $74 \times 64/200$ 。这其实也就是说，每个期望频次，都是列联表中，“行总计”乘以“列总计”再除以“样本容量”得到的。

	高粱饴	酒心糖	大虾酥	巧克力
老年人	74*32% =23.68	74*15% =11.1	74*15% =11.1	74*38% =28.12
中年人	34*32% =10.88	34*15% =5.1	34*15% =5.1	34*38% =12.92
小孩	92*32% =29.44	92*15% =13.8	92*15% =13.8	92*38% =34.96
	高粱饴	酒心糖	大虾酥	巧克力
老年人				74
中年人			34* (76/200)	34
小孩				92
总计	64	30	30	76
				200

我们可以根据这个公式，再来算一下中年人的巧克力的期望频次。即：第2行，第4列的期望频次，是第2行的行总计34，乘以第4列的列总计76，再除以样本容量200。期望频次算出来等于12.92。

		高粱饴	酒心糖	大虾酥	巧克力
老年人	56	6	5	7	
中年人	4	12	13	5	
小孩	4	12	12	64	

		高粱饴	酒心糖	大虾酥	巧克力
老年人	23.68	11.1	11.1	28.12	
中年人	10.88	5.1	5.1	12.92	
小孩	29.44	13.8	13.8	34.96	

现在，观测频次有了，期望频次也有了。这两个频次之间的差别，就是样本和 H_0 之间的差别。

高粱饴 酒心糖 大虾酥 巧克力					高粱饴 酒心糖 大虾酥 巧克力				
老年人	56	6	5	7	老年人	23.68	11.1	11.1	28.12
中年人	4	12	13	5	中年人	10.88	5.1	5.1	12.92
小孩	4	12	12	64	小孩	29.44	13.8	13.8	34.96

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

如何量化这种差别呢？在拟合优度检验中，我们也已经学过了，就用这个卡方值，来量化观测频次和期望频次之间的差别。

	高粱饴	酒心糖	大虾酥	巧克力		高粱饴	酒心糖	大虾酥	巧克力
老年人	56	6	5	7	老年人	23.68	11.1	11.1	28.12
中年人	4	12	13	5	中年人	10.88	5.1	5.1	12.92
小孩	4	12	12	64	小孩	29.44	13.8	13.8	34.96

观测 频次

$$\begin{aligned}
 \bar{x}^2 &= \frac{(56-23.68)^2}{23.68} + \frac{(6-11.1)^2}{11.1} + \frac{(5-11.1)^2}{11.1} + \frac{(7-28.12)^2}{28.12} \\
 &+ \frac{(4-10.88)^2}{10.88} + \frac{(12-5.1)^2}{5.1} + \frac{(13-5.1)^2}{5.1} + \frac{(5-12.92)^2}{12.92} \\
 &+ \frac{(4-29.44)^2}{29.44} + \frac{(12-13.8)^2}{13.8} + \frac{(12-13.8)^2}{13.8} + \frac{(64-34.96)^2}{34.96} \\
 &= 44.11 + 2.34 + 3.35 + 15.86 \\
 &+ 4.35 + 9.33 + 12.23 + 4.85 \\
 &+ 21.98 + 0.23 + 0.23 + 24.12 \\
 &= \mathbf{143.02}
 \end{aligned}$$

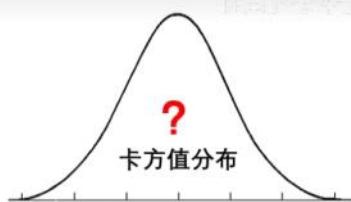
$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

我们把数据代入，展开公式。例如，老年人高粱饴的观测频次是56，期望频次是23.68，是这一项；再如，中年人大虾酥的观测频次是13，期望频次是5.1，是这一项；以此类推。最后，我们算出一个总的卡方值，143.02。那么，这个卡方值，代表什么呢？

	高粱饴	酒心糖	大虾酥	巧克力
老年人	56	6	5	7
中年人	4	12	13	5
小孩	4	12	12	64

观测 频次

$$\chi^2 = 143.02$$



H_0 为真的总体

我们仍然需要把这一次抽样的卡方值，放到无数次抽样的卡方值分布中去看，看看本次的卡方值是否极端，是否属于小概率事件。要想构造一个卡方值分布，我们首先需要构造出一个 H_0 为真的总体，然后从这个总体中反反复抽样，才能获得卡方值分布。

	高粱饴	酒心糖	大虾酥	巧克力	36%	期望 频次
老年人	23.68	11.1	11.1	28.12	36%	
中年人	10.88	5.1	5.1	12.92	17%	
小孩	29.44	13.8	13.8	34.96	48%	
	32%	15%	15%	38%	200	

	高粱饴	酒心糖	大虾酥	巧克力	36%	H_0 为真的 总体
老年人	2368	1110	1110	2812	36%	
中年人	1088	510	510	1292	17%	
小孩	2944	1380	1380	3496	48%	
	32%	15%	15%	38%	20000	

如何构造 H_0 为真的总体呢？其实也不难，我们把这个期望频次表中，每个单元格的数据，都乘以100，就得到一个总频次为20000的数据表。显然，这20000个数据中，各类糖的比例和各类年龄群体的比例，和 H_0 是一模一样的，只不过数目扩大了100倍而已。我们就把这20000个频次，作为完美符合 H_0 的总体。

	高粱饴	酒心糖	大虾酥	巧克力	H_0 为真的 总体
老年人	2368	1110	1110	2812	
中年人	1088	510	510	1292	
小孩	2944	1380	1380	3496	

	高粱饴	酒心糖	大虾酥	巧克力	制造 卡方值分布
老年人	?	?	?	?	
中年人	?	?	?	?	
小孩	?	?	?	?	

N = 200
反复大量抽样

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

我们要从这个 H_0 为真的总体中，进行样本容量为200的，反复大量抽样。每次抽样，都算出一个卡方值，以此制造出一个卡方值分布。

	高粱饴	酒心糖	大虾酥	巧克力	H_0 为真的 总体
老年人	2368	1110	1110	2812	
中年人	1088	510	510	1292	
小孩	2944	1380	1380	3496	

```
population <- c(
  rep("老年人-高粱饴", 2368), rep("老年人-酒心糖", 1110), rep("老年人-大虾酥", 1110), rep("老年人-巧克力", 2812),
  rep("中年人-高粱饴", 1088), rep("中年人-酒心糖", 510), rep("中年人-大虾酥", 510), rep("中年人-巧克力", 1292),
  rep("小孩-高粱饴", 2944), rep("小孩-酒心糖", 1380), rep("小孩-大虾酥", 1380), rep("小孩-巧克力", 3496))
```

首先，我们用最笨最原始的方法，把这个 H_0 为真的总体，输入到R软件中。代码的含义，稍微给大家解释下，例如，这里表示，老年人选高粱饴2368次，这里表示，小孩选酒心糖1380次，以此类推。

```

> smpl <- sample(population, 200)
> smpl
[1] "老年人-巧克力" "小孩-高粱饴" "中年人-高粱饴" "老年人-巧克力" "老年人-酒心糖" "老年人-巧克力" "老年人-酒心糖"
[11] "老年人-酒心糖" "老年人-大虾酥" "老年人-高粱饴" "老年人-巧克力" "小孩-巧克力" "中年人-高粱饴" "老年人-巧克力" "小孩-巧克力"
[21] "小孩-酒心糖" "中年人-高粱饴" "老年人-巧克力" "老年人-酒心糖" "中年人-巧克力" "中年人-大虾酥" "小孩-巧克力" "小孩-高粱饴"
[31] "中年人-巧克力" "小孩-巧克力" "小孩-酒心糖" "老年人-巧克力" "小孩-高粱饴" "小孩-巧克力" "老年人-巧克力" "小孩-高粱饴"
[41] "小孩-高粱饴" "小孩-大虾酥" "小孩-巧克力" "中年人-大虾酥" "老年人-巧克力" "老年人-高粱饴" "小孩-酒心糖" "小孩-巧克力"
[51] "小孩-高粱饴" "小孩-巧克力" "小孩-巧克力" "老年人-高粱饴" "小孩-酒心糖" "小孩-酒心糖" "小孩-巧克力" "小孩-巧克力"
[61] "老年人-巧克力" "小孩-巧克力" "老年人-巧克力" "老年人-高粱饴" "小孩-高粱饴" "小孩-巧克力" "中年人-大虾酥" "小孩-巧克力"
[71] "小孩-高粱饴" "老年人-大虾酥" "小孩-高粱饴" "老年人-酒心糖" "小孩-巧克力" "老年人-巧克力" "小孩-高粱饴" "小孩-巧克力"
[81] "小孩-巧克力" "老年人-巧克力" "中年人-高粱饴" "老年人-高粱饴" "小孩-高粱饴" "小孩-巧克力" "中年人-酒心糖" "小孩-酒心糖"
[91] "老年人-高粱饴" "老年人-巧克力" "老年人-大虾酥" "老年人-高粱饴" "小孩-高粱饴" "老年人-酒心糖" "老年人-巧克力" "老年人-高粱饴"
[101] "老年人-高粱饴" "小孩-巧克力" "小孩-巧克力" "老年人-高粱饴" "小孩-高粱饴" "老年人-酒心糖" "老年人-巧克力" "老年人-高粱饴"
[111] "小孩-高粱饴" "小孩-巧克力" "小孩-巧克力" "中年人-高粱饴" "小孩-高粱饴" "老年人-高粱饴" "老年人-巧克力" "老年人-高粱饴"
[121] "老年人-高粱饴" "小孩-高粱饴" "中年人-高粱饴" "老年人-高粱饴" "小孩-高粱饴" "老年人-高粱饴" "老年人-巧克力" "老年人-高粱饴"
[131] "小孩-巧克力" "老年人-巧克力" "中年人-大虾酥" "老年人-高粱饴" "小孩-高粱饴" "老年人-高粱饴" "老年人-巧克力" "老年人-高粱饴"
[141] "小孩-大虾酥" "老年人-巧克力" "小孩-高粱饴" "小孩-高粱饴" "小孩-高粱饴" "老年人-高粱饴" "老年人-巧克力" "老年人-高粱饴"
[151] "老年人-大虾酥" "小孩-巧克力" "老年人-巧克力" "老年人-高粱饴" "小孩-高粱饴" "中年人-巧克力" "中年人-巧克力" "小孩-酒心糖"
[161] "小孩-高粱饴" "小孩-高粱饴" "老年人-大虾酥" "老年人-高粱饴" "老年人-高粱饴" "老年人-巧克力" "中年人-高粱饴" "小孩-巧克力"
[171] "小孩-巧克力" "小孩-高粱饴" "小孩-酒心糖" "小孩-高粱饴" "小孩-高粱饴" "小孩-高粱饴" "中年人-大虾酥" "中年人-高粱饴"
[181] "小孩-巧克力" "小孩-大虾酥" "老年人-巧克力" "老年人-酒心糖" "小孩-巧克力" "中年人-大虾酥" "老年人-高粱饴" "小孩-高粱饴"
[191] "小孩-高粱饴" "老年人-酒心糖" "老年人-巧克力" "老年人-酒心糖" "老年人-巧克力" "老年人-酒心糖" "小孩-大虾酥" "小孩-巧克力"

```

高粱饴 酒心糖 大虾酥 巧克力				本次样本 观察频次
老年人	19	13	8	29
中年人	9	0	10	8
小孩	41	13	10	40

下面，我们从这个总体中，先抽样1次。得到这样一个样本，包含200个数据。稍作统计，这就是这次样本的观测频次。

观测频次 Observed Frequency				期望频次 Expected Frequency			
老年人	19	13	8	29	23.68	11.1	11.1
中年人	9	0	10	8	10.88	5.1	5.1
小孩	41	13	10	40	29.44	13.8	13.8

然后把观测频次和期望频次，都输入到R软件中，再根据公式，算出这次抽样的卡方值。等于20.50左右。
以上，是1次抽样。

```

statistics <- c()
#抽样1000次
for(i in 1:10000){
  #抽样1次
  smpl <- sample(population, 200)

  o_freq <- c(
    sum(smpl=="老年人-高粱饴"), sum(smpl=="老年人-酒心糖"), sum(smpl=="老年人-大虾酥"), sum(smpl=="老年人-巧克力"),
    sum(smpl=="中年人-高粱饴"), sum(smpl=="中年人-酒心糖"), sum(smpl=="中年人-大虾酥"), sum(smpl=="中年人-巧克力"),
    sum(smpl=="小孩-高粱饴"), sum(smpl=="小孩-酒心糖"), sum(smpl=="小孩-大虾酥"), sum(smpl=="小孩-巧克力"))

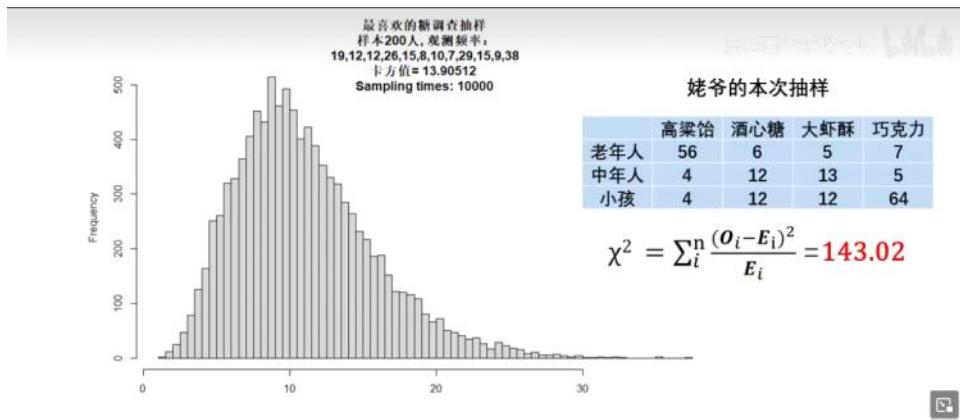
  #计算卡方
  chsq <- sum((o_freq - e_freq)^2/e_freq)

  statistics <- append(statistics , chsq) #统计量加入序列

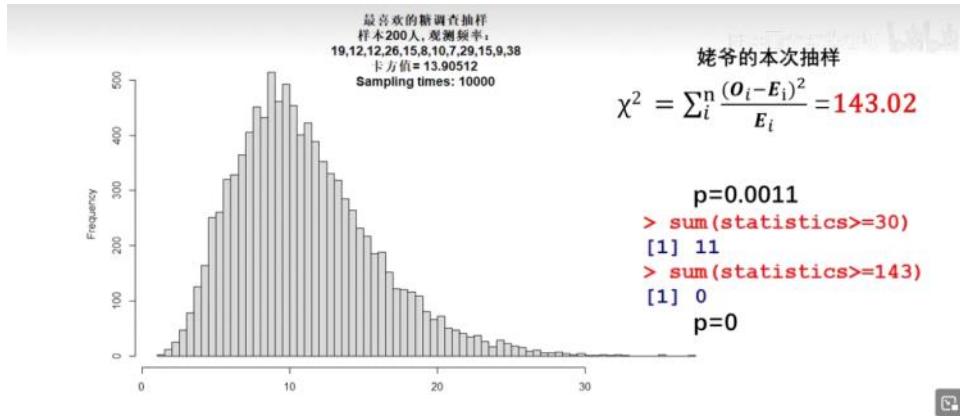
  hist(statistics , 100,main=paste("\n\n最喜欢的糖调查抽样\n",
  "样本200人， 观测频率: \n"))
}

```

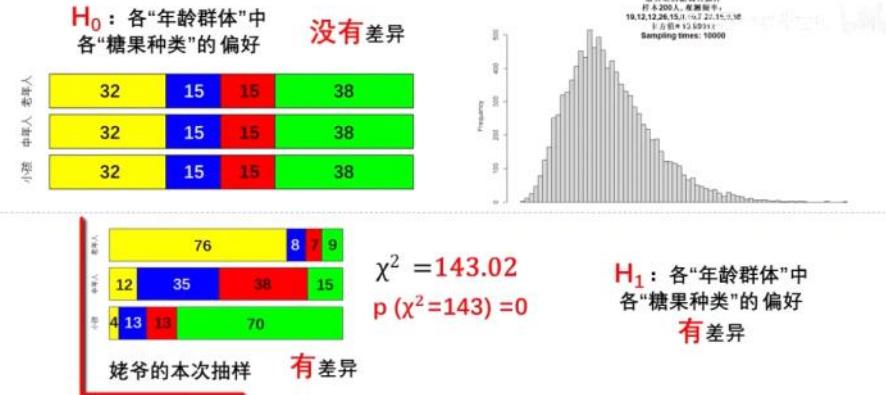
下面，我们进行10000次抽样，每次都得到一个观测频次，然后和期望频次比较，计算出一个卡方值。
然后把10000次抽样的卡方值，画到一个直方图里。



而姥爷的本次抽样, 卡方值为143.02, 在这个 H_0 为真的卡方值分布中, 显然是比尾巴尖30, 还要极端好几倍的。



我们用R程序中的sum()函数数一下, 抽到比143.02 (含) 还极端的卡方值的次数。发现, 在这10000个卡方值中, 大于等于30的, 只有11个, 也就是卡方=30的p值是11/10000, $p=0.0011$ 。10000个卡方值中, 大于等于143的, 一个都没有, 也就是卡方=143的p值=0。



这就说明, 在 H_0 为真的总体中, 也就是, 各年龄群体, 对各糖果种类的偏好, 完全没有差异的总体中, 抽样得到卡方值=143.02, 是非常极端的小概率事件。于是我们就拒绝 H_0 , 进而接受 H_1 , 即: 各年龄群体, 对各糖果种类的偏好, 是有差异的。或者说, 姥爷抽到的这个样本, 不来自这个偏好没有差异的总体。

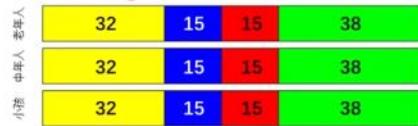
	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人	56	6	5	7	74
中年人	4	12	13	5	34
小孩	4	12	12	64	92
总计	64	30	30	76	200

抽样频次/观测频次



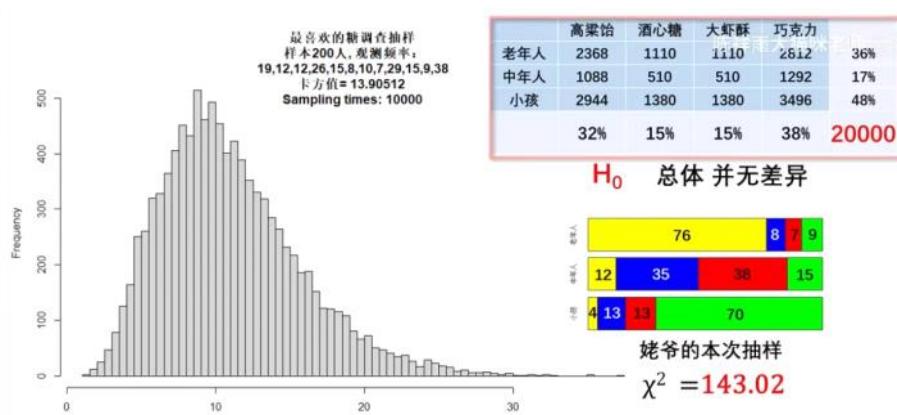
抽样 百分比

H_0 总体 并无差异



总体 百分比

现在，我们再从头捋一遍。姥爷在村里做了个调查问卷，抽样200人，问大家最喜欢吃哪种糖。问卷结果的频次是这样的，画成百分比图是这样的。根据这个图，姥爷猜测，不同年龄群体，对不糖果种类，有不同的偏好。但是，姥爷转念一想，有没有可能，总体中的各年龄群体，对各糖果种类，其实并没有偏好差异，而只是这一次抽样，恰巧表现得有差异呢？



那么，我们就构造出一个 H_0 为真的，20000人的总体，并对这个总体进行反复大量的抽样，得出卡方值分布。再把姥爷这次抽样的卡方值，放到这个卡方值分布中去比较，结果发现姥爷的这次抽样，是极端的小概率事件，于是拒绝 H_0 。结论是，姥爷这个样本所来自的总体中，不同年龄群体，对不同糖果种类，存在显著的偏好差异。以上，就是整个假设检验的思路。现在，我们给这种假设检验命名。

卡方 拟合优度 检验

χ^2 Chi Square Test for
Goodness of Fit

1个类别变量 “糖果种类”



之前我们学过的，姥爷兑兑锦糖比例的故事中，只有一个类别变量，“糖果种类”。只有一个类别变量的卡方检验，叫做卡方拟合优度检验。



本节课的故事中，有两个类别变量：“糖果种类”和“年龄群体”。本故事的研究问题是，不同年龄群体，对不同糖果种类的偏好或者百分比，是否存在差异。下面，我们对这个故事进行抽象。

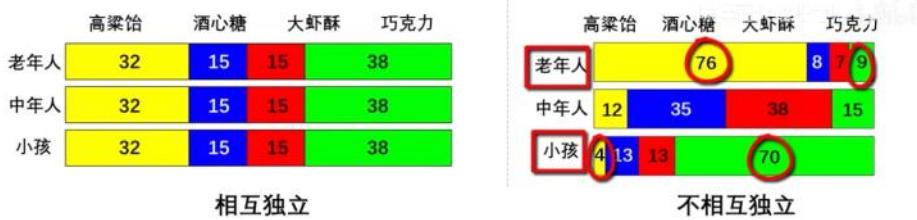


相互独立

不相互独立

当一个类别变量取不同值的时候
另外一个类别变量中，
某一类别的偏好百分比，是否存在差异

请注意我下面说的话，本故事的研究问题是，当一个类别变量取不同值的时候，另外一个类别变量中，某一类别的偏好百分比，是否存在差异。再进一步抽象，就是研究这两个类别变量，是否独立。



相互独立

不相互独立

当一个类别变量取不同值的时候
另外一个类别变量中，
某一类别的偏好百分比，是否存在差异

在本故事中，假如，无论“年龄群体”的取值是“老年人”，“中年人”，还是“小孩”，糖果种类的偏好百分比都不受影响，都是一样的，那么，两个类别变量就是相互独立的。但是，假如“年龄群体”取值“老年人”和“小孩”时，高粱饴或巧克力的偏好存在显著的差别，那就说明，“糖果种类”的偏好是受“年龄群体”取值影响的，于是，这两个类别变量，就不是相互独立的。

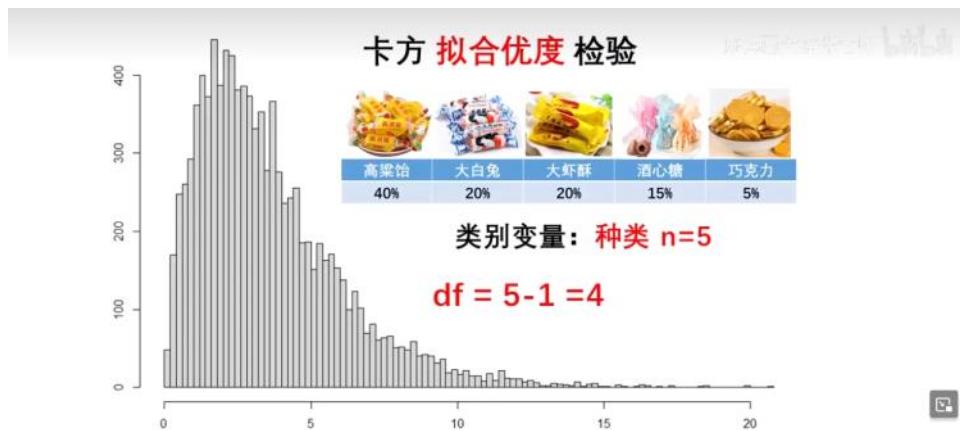
卡方 独立性 检验

Chi Square Test of
Independence

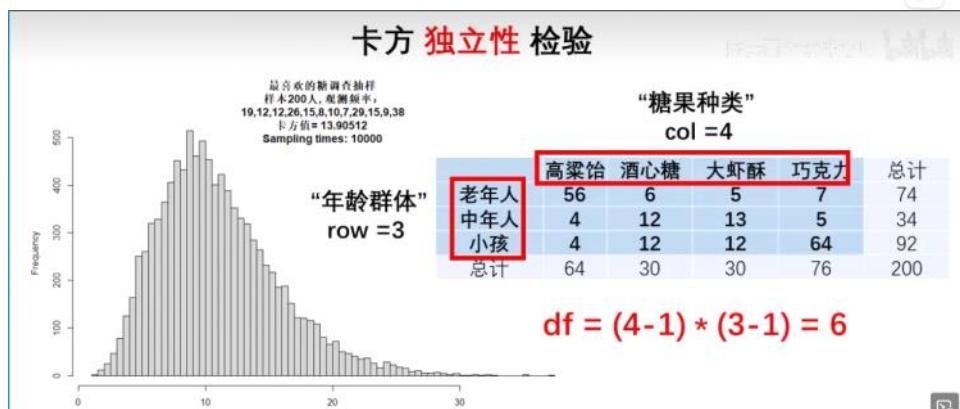
类别变量 1 糖果种类					
类别变量 年龄群 量体 2	高粱饴	酒心糖	大虾酥	巧克力	
	老年人	56	6	5	7
	中年人	4	12	13	5
	小孩	4	12	12	64

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

所以，这种检验两个类别变量是否相互独立的卡方检验，就叫做“卡方独立性检验”，Chi Square Test of Independence。下面，我们来看一下，卡方独立性检验的自由度。

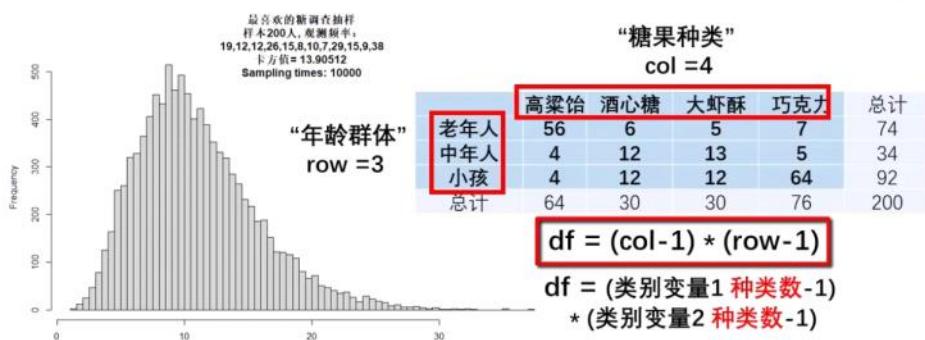


我们先回忆一下拟合优度检验的自由度。因为拟合优度检验只有一个类别变量，其自由度的计算是比较简单的，就是类别变量的种类减去1。



卡方独立性检验中，有两个类别变量。一个是“糖果种类”，有4种取值，对应列联表中的4列；另一个是“年龄群体”，有3种取值，对应列联表中3行。那么，这个卡方分布的自由度被定义为 $df=(4-1)*(3-1)=6$ 。所以，这是一个 $df=6$ 的卡方分布。

卡方 独立性 检验



大家可能听到过这样的说法, 卡方独立性检验的自由度是, 列联表的列数column number-1 乘以行数 row number-1。这里所谓的列数和行数, 其实就是两个类别变量的种类数。我更倾向于用“种类数-1”这种说法。因为列数和行数, 容易产生歧义。假如列联表多了总计一列和总计一行, 解释起来就比较啰嗦了。以上, 就是卡方独立性检验的基本原理。

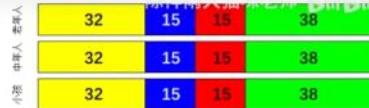
```
> candies_smpl = c(  
+ 56, 6, 5, 7,  
+ 4, 12, 13, 5,  
+ 4, 12, 12, 64)  
  
> rownames = c("老年人", "中年人", "小孩")  
> colnames = c("高粱饴", "酒心糖", "大虾酥", "巧克力")  
  
> candies_mat = matrix(candies_smpl, nrow=3, byrow=TRUE,  
+ dimname = list(rownames, colnames))  
  
> candies_mat  
高粱饴 酒心糖 大虾酥 巧克力  
老年人 56 6 5 7  
中年人 4 12 13 5  
小孩 4 12 12 64
```

下面, 我们演示一下如何用R软件来进行卡方独立性检验。我们只需要把抽样的观测频次输入到程序中就可以了, 不需要计算期望频次等其他任何数据。首先, 把样本这个一维的数组, 赋值给这个变量。然后, 用矩阵matrix命令, 按照我们规定的行数、行与列的名称, 将其转换为一个二维的列联表。

```
> candies_mat  
高粱饴 酒心糖 大虾酥 巧克力  
老年人 56 6 5 7  
中年人 4 12 13 5  
小孩 4 12 12 64  
  
> chisq.test(candies_mat)  
  
Pearson's Chi-squared test  
  
data: candies_mat  
X-squared = 143.02, df = 6, p-value < 2.2e-16
```

然后, 用卡方检验命令chisq.test, 就可以直接出结果了。卡方值算出来是143.02, 和我们手工算的是一样的。自由度等于 $(4-1) * (3-1) = 6$ 。p值是2.2乘以10的负16次方, 也就是小数点后, 有十几个0, 这和我们刚才估算的p=0, 也是一致的。

```
> candies_mat
高粱饴 酒心糖 大虾酥 巧克力
老年人 56 6 5 7
中年人 4 12 13 5
小孩 4 12 12 64
```



```
> chisq.test(candies_mat)
```

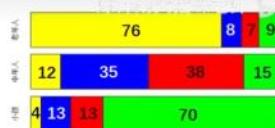
```
Pearson's Chi-squared test
data: candies_mat
X-squared = 143.02, df = 6, p-value < 2.2e-16
```

H_0 : 各“年龄群体”对各“糖果种类”的偏好没有差异

H_0 : “年龄群体”与“糖果种类”两个类别变量相互独立

于是，我们拒绝 H_0 。希望大家现在还记得 H_0 是什么， H_0 是：各年龄群体对各糖果种类的偏好没有差异。或者说，“年龄群体”和“糖果种类”这两个类别变量是相互独立的。

```
> candies_mat
高粱饴 酒心糖 大虾酥 巧克力
老年人 56 6 5 7
中年人 4 12 13 5
小孩 4 12 12 64
```



```
> chisq.test(candies_mat)
```

```
Pearson's Chi-squared test
data: candies_mat
X-squared = 143.02, df = 6, p-value < 2.2e-16
```

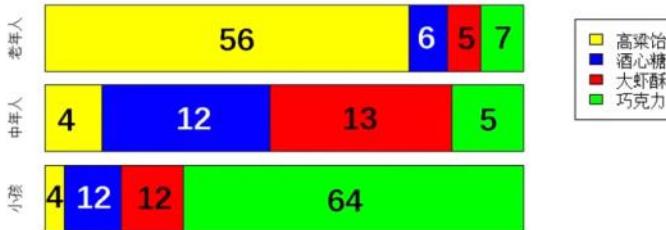
H_1 : 各“年龄群体”对各“糖果种类”的偏好有显著差异

H_1 : “年龄群体”与“糖果种类”两个类别变量不相互独立

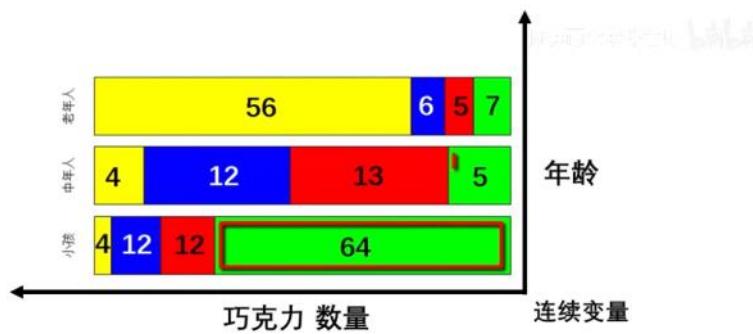
姥爷的猜测

拒绝了 H_0 ，就是接受了对立的 H_1 。 H_1 就是，两个类别变量不是相互独立的，或者说，各“年龄群体”对各“糖果种类”的偏好存在显著差异。于是，姥爷最初的猜测，得到了统计数据的支撑。

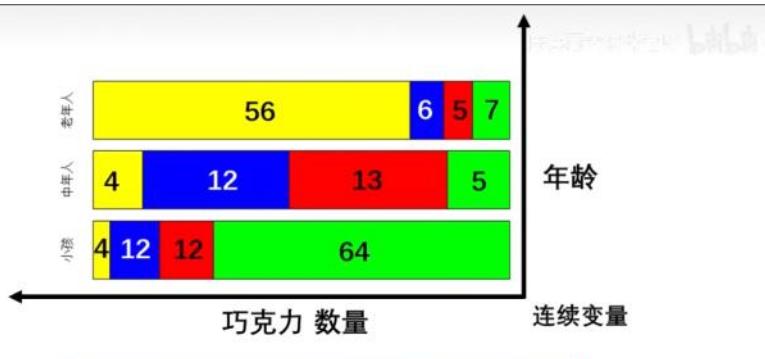
卡方 独立性 检验



故事讲完了，大家可能还是不满意的。就算姥爷证明了，各年龄群体，对各糖果种类存在不同的偏好，又有什么用呢？这里，我非常不严谨，不科学的进行举例说明。例如，姥爷看着这个样本的stacked bar chart，可以告诉养老院门口的小卖部，多备货一些高粱饴；而幼儿园门口的小卖部呢，多备货一些巧克力。



再例如，图中显示，随着年龄的增长，人们对巧克力的偏好逐渐降低。于是我们可能会产生一种猜想：一个人的年龄，和一个人每天吃巧克力的数量，这两个连续变量，是否存在一个线性相关呢？



$$\text{巧克力 数量} = a * \text{年龄} + b ?$$

换句话说，是否存在一个公式模型，可以通过一个人的“年龄”，估算出这个人每天吃巧克力的数量呢。这就是线性拟合，是后面的课程了。

卡方 独立性 检验

H_0 : 偏好无差别

	高粱饴	酒心糖	大虾酥	巧克力	总计	
老年人	56	6	5	7	74	32 15 15 38
中年人	4	12	13	5	34	32 15 15 38
小孩	4	12	12	64	92	32 15 15 38
总计	64	30	30	76	200	

难点1: H_0 中的比例

以上，就是卡方独立性检验的基本原理。其中有几个难点，我们再复习一下：第一， H_0 中的比例，不是简单的平均分配，而是抹去其中一个类别变量后，得到的比例。

卡方 独立性 检验

	高粱饴	酒心糖	大虾酥	巧克力	总计
老年人				74	
中年人			$34*76/200$ =5.1	34	
小孩				92	
总计	64	30	30	76	200

$$E_{ij} = \frac{O_i * O_j}{N}$$

难点2：期望频次的计算

第二，期望频次的计算方法，是行总计乘以列总计，再除以样本容量。

卡方 独立性 检验

类别变量 1 糖果种类

类 别	年 龄	高粱饴 酒心糖 大虾酥 巧克力				
		老年人	56	6	5	7
变 群	中年人	4	12	13	5	
量 体	小孩	4	12	12	64	
2						

难点3：“独立性”的含义

第三，“独立性”检验，是指两个类别变量之间相互独立。注意，是类别变量，不是连续变量。

假设检验

2024年1月9日 10:01

<https://blog.csdn.net/ws19920726/article/details/105831471>

假设检验，也称为显著性检验，通过样本的统计量来判断与总体参数之间是否存在差异（差异是否显著）。即我们对总体参数进行一定的假设，然后通过收集到的数据，来验证我们之前作出的假设（总体参数）是否合理。在假设检验中，我们会建立两个完全对立的假设，分别为原假设 H_0 与备择假设 H_1 。然后根据样本信息进行分析判断，是选择接受原假设还是拒绝原假设。

假设检验基于“反证法”。首先，我们假设原假设为真，如果在此基础上，得出了违反逻辑与常理的结论，则表明原假设是错误的，我们就接受备择假设。

小概率事件

在假设检验中，违反逻辑与常规的结论，就是小概率事件。一般来说，小概率事件在一次试验中是不会发生的。如果发生，则我们便有理由拒绝原假设。

假设检验遵循“疑罪从无”的原则，接受原假设，并不代表原假设一定是正确的，只是没有充分的证据，证明原假设是错误的。

P-Value与显著性水平

为了便于量化，我们可以计算一个概率值（P-Value），该概率值可以认为是支持原假设的概率，也就是样本统计量与总体参数无差异的概率。然后，我们设定一个显著性水平 α （通常取值为0.05）。**当P-Value的值大于 α 时，支持原假设。**

假设检验与置信区间有一定的关联性，只不过假设检验是通过反证的角度来判断是否接受原假设。

假设检验的步骤

步骤如下：

设置原假设与备择假设。

设置显著性水平 α （通常选择 $\alpha=0.05$ ）。

根据问题选择假设检验的方式。

计算统计量，并通过统计量获取P值。

根据P值与 α 值，决定接受原假设还是备择假设。

Z检验 AB测试

2024年1月9日 10:07

Z检验用来判断样本均值是否与总体均值具有显著性差异。Z检验是通过正态分布的理论来推断差异发生的概率，从而比较两个均值的差异是否显著。Z检验适用于：

总体呈正态分布。

总体方差已知。

样本容量较大 (≥ 30)。

$$Z = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- \bar{x} : 样本均值。
- μ_0 : 待检验的总体均值（假设的总体均值）。
- $S_{\bar{x}}$: 样本均值分布的标准差（标准误差）。
- σ : 总体的标准差。
- n : 样本容量。

<https://blog.csdn.net/u19326726>

t检验与Z检验类似，用来判断样本均值是否与总体均值均有显著性差异。不过t检验是基于t分布的，适用于：

总体呈正态分布。

总体方差未知。

样本数量较少 (< 30)

不过，随着样本容量的增大（样本数量 ≥ 30 ），t分布逐渐接近于正态分布。此时，t检验也近似于Z检验。

t统计量计算公式：

$$t = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$$

- \bar{x} : 样本均值。
- μ_0 : 待检验的总体均值（假设的总体均值）。
- $S_{\bar{x}}$: 样本均值的标准差（标准误差）。
- S : 样本的标准差。
- n : 样本容量。

例：鸢尾花的平均花瓣长度为3.5cm，这种说法正确吗？
可以根据假设检验的步骤，进行解决。

设置原假设与备择假设：

原假设： $\mu = \mu_0 = 3.5\text{cm}$ （说法正确）

备择假设： $\mu \neq \mu_0 \neq 3.5\text{cm}$ （说法不正确）

设置显著性水平：

$\alpha = 0.05$

根据问题选择假设检验的方式：

鸢尾花数据呈正态分布，但总体标准差未知，故选择t检验。

计算统计量，并通过统计量获取P值。

Ttest_1sampResult(statistic=1.7599751687043313, pvalue=0.07546856490783705)

通过假设检验可以计算如下题目：

某车间用一台机器制作袋装糖，袋装糖的净重是一个随机变量，服从正态分布。机器运行正常时，其均值为0.5kg，标准差为0.015kg。某日工作后，检验包装机是否正常，随机抽取9袋糖，称得净重为 (kg)：0.497、0.506、0.518、0.524、0.498、0.511、0.520、0.515、0.512，请问机器是否正常？

设置原假设与备择假设：

原假设： $\mu = \mu_0 = 0.5\text{kg}$ （机器正常）

备择假设： $\mu \neq \mu_0 \neq 0.5\text{kg}$ （机器不正常）

设置显著性水平：

$\alpha = 0.05$

根据问题选择假设检验的方式：

根据题意已知糖的净重呈正态分布且总体标准差已知，故选择Z检验。

计算统计量，并通过统计量获取P值。

统计量Z： 2.24444444444471

P_value值： 0.02460081900325589

通过结果可知，P值小于0.05，则拒绝原假设，接受备择假设，我们可以认为机器运作不正常。

AB测试

1) ABtest的选择场景

我们做AB Test，“如果样本量足够大，那么Z检验和t检验将得出相同的结果。对于大样本，样本方差是对总体方差的较好估计，因此即使总体方差未知，我们也可以使用样本方差的Z检验”。[6]，但正常来说，除非是长期的实验(0.5-1年)，例如算法，会选择Z检验。正常的短期AB Test基本是实验1个月内甚至说1-2周，那么此时建议选择T检验。

版权声明：本文为CSDN博主「画扇落汗」的原创文章，遵循CC 4.0 BY-SA版权协议，转载请附上原文出处链接及本声明。

原文链接：<https://blog.csdn.net/garbageSystem/article/details/122603832>