**Study Notes**

目录

[统计基础 3](#_Toc499025157)

[16/11/2017 3](#_Toc499025158)

[Convincing Survey 3](#_Toc499025159)

[Define Abstract Concept 3](#_Toc499025160)

[BBC measurement 3](#_Toc499025161)

[Construct构建/抽象概念 3](#_Toc499025162)

[Sampling Error 抽样误差 3](#_Toc499025163)

[用Visualisation预测模型 3](#_Toc499025164)

[Correlation does not imply causation 4](#_Toc499025165)

[Bias Control 4](#_Toc499025166)

[Randomisation 4](#_Toc499025167)

[偏斜分布 vs 正态分布 4](#_Toc499025168)

[Measure of center中心测量值 4](#_Toc499025169)

[Mode 4](#_Toc499025170)

[Mean 5](#_Toc499025171)

[Median 5](#_Toc499025172)

[17/11/2017 5](#_Toc499025173)

[量化数据的分布形态 – Range值域 5](#_Toc499025174)

[异常值Outlier 6](#_Toc499025175)

[箱线图 6](#_Toc499025176)

[衡量差异的最优方法 6](#_Toc499025177)

[为什么标准差重要？ 7](#_Toc499025178)

[Bessel‘s Correction贝塞耳校正 7](#_Toc499025179)

[Standardise归一化 7](#_Toc499025180)

[Z 7](#_Toc499025181)

[标准正态分布 7](#_Toc499025182)

[Probability Density Function概率密度函数 PDF 7](#_Toc499025183)

[Horizontal asymptote 水平渐近线 8](#_Toc499025184)

[P值 - Z表格 8](#_Toc499025185)

[Sampling Distribution 抽样分布 SE 8](#_Toc499025186)

[Central Limit Theorem中心极限定理 8](#_Toc499025187)

[计算特定样本的Z值 9](#_Toc499025188)

[样本均值的百分比 10](#_Toc499025189)

[Margin of Error误差范围 10](#_Toc499025190)

[Confidence Interval置信区间 10](#_Toc499025191)

[精确的Z值 10](#_Toc499025192)

[判定特定因素是否产生干扰的抽样统计流程 10](#_Toc499025193)

[α level 11](#_Toc499025194)

[Z Critical Value 11](#_Toc499025195)

[Significant显著性 11](#_Toc499025196)

[Two-Tails Test双尾临界值 12](#_Toc499025197)

[H0 Null hypothesis 零假设| Ha Alternative hypothesis对立假设 12](#_Toc499025198)

[Reject the null 拒绝零假设 12](#_Toc499025199)

[决策失误 13](#_Toc499025200)

[t-test t检验 13](#_Toc499025201)

[Degree of freedom自由度df 13](#_Toc499025202)

[T表格 13](#_Toc499025203)

[T分布与零假设 14](#_Toc499025204)

[总结：One-sample t-test单样本t检验 14](#_Toc499025205)

[t值对应置信区间和误差界限 15](#_Toc499025206)

[Paired samples & Dependent t-test 15](#_Toc499025207)

[比较两个相依样本间的差异 15](#_Toc499025208)

[Cohen’ d 16](#_Toc499025209)

[T检验 vs Z检验 16](#_Toc499025210)

[Effect Size效应量 16](#_Toc499025211)

[Statistical significance统计显著性 vs 结果的意义 16](#_Toc499025212)

[R2 -确定系数 17](#_Toc499025213)

[Result section有哪些内容？ 17](#_Toc499025214)

[相依样本 优势和劣势 17](#_Toc499025215)

[两个样本集的正态分布差 17](#_Toc499025216)

[合并方差 & 矫正的标准误差 18](#_Toc499025217)

# 统计基础

### 16/11/2017

### Convincing Survey

* How many people I surveyed？我调查了多少人？
* Who I surveyed？ 调查了谁？
* How the survey was conducted？怎么调查的？（Methodology）

### Define Abstract Concept

### BBC measurement

* 你记住了哪些面孔？
* **你从第一部分和第二部分正确识别和放置的百分比？**
* 你是否知道你是第一次还是第二次见到某个面孔？
* 你是否知道某个面孔已存在？
* 你记住的面孔数量？

### Construct构建/抽象概念

如何定义和度量？

* Intelligence: IQ test
* Effort: time, grades, gpa
* Age: maturity, age in years, wisdom
* Hunger: how often tummy grumbles
* Itchiness: How often you scratch/shake, etc.

### Sampling Error 抽样误差

µ - 样本相较总体的误差

Population AVG = µ: 整体参数-描述总体的值 （parameter）

Sample AVG = ：样本统计量—估计总体参数 （statistic）

**什么是操作性定义？选择所有适用项**

* □ 我们有兴趣研究的一个抽象概念
* □ 一种将构造（constructs）转变为我们可衡量的变量的方式
* □ 样本统计量和总体参数之间的差异
* □ 一种用我们衡量它的方式描述变量的方式
* □ 调查研究中的一组兴趣个人
* □ 总体参数的估计

### 用Visualisation预测模型

x-predictor variable {independent

y-outcome {dependent

lurking variable { potential factors

### Correlation does not imply causation

* Show relationships → Observational studies，like survey
* Show causation → Controlled experiment

### Bias Control

Blinding单盲，消除被试者bias

Double-blinding双盲，还消除研究者bias

### Randomisation

Random sample vs convenience sample： bias

### 偏斜分布 vs 正态分布

Positively skewed

Negatively skewed

Narmal

### Measure of center中心测量值

Mode 众数：出现频率最高的值

Median中位数： 分布在最中间的值

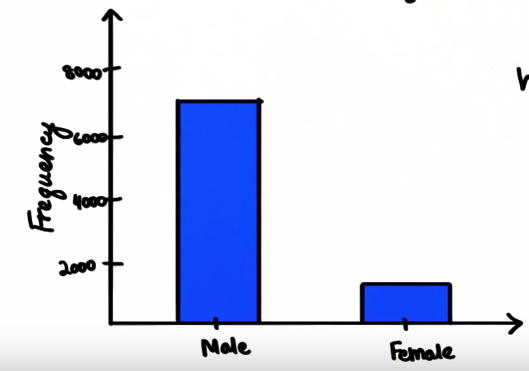
Mean 平均数

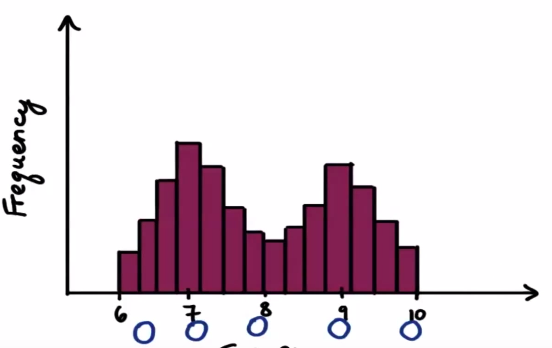
### Mode

出现在x轴，区分frequency

可以numerical或categorical

* Uniform distribution均匀分布 no mode
* Multiple mode like

 bi-mode distribution



多峰分布

### Mean

=

Outlier can make misleading info to audience， that’s why we need robust median

### Median

Odd: number in the middle

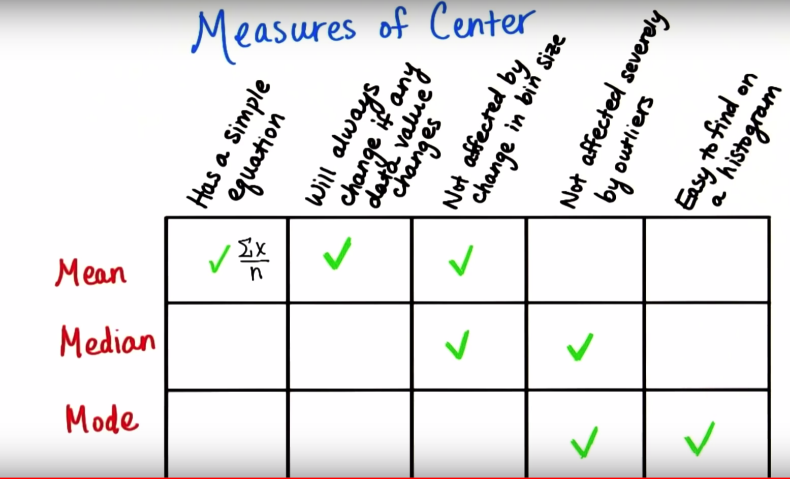
Even: exactly in middle of two numbers

理想正态分布：Mean = Mode = Median

正偏斜分布： Mode < Median < Mean

负偏斜分布： Mean > Median> Mode

因为一般在偏斜分布里，mean明显受outlier影响向一端偏斜，而median受影响较小



### 17/11/2017

### 量化数据的分布形态 – Range值域

Distribution 是相对Y轴来说

Range = Maximum – Minimum （How spread out of data），相对X轴来说

Range受outlier的影响，为了减小outlier的影响，一般统计学上会去掉两端数据的的25%

第一个四分位Q1，和第三个四分位Q3

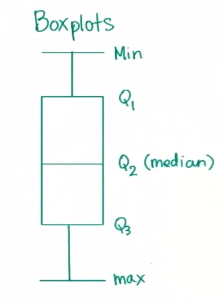
Interquartile range（IQR）= Q3-Q1

### 异常值Outlier

<Q1 -1.5(IQR)

>Q3 + 1.5(IQR)

### 箱线图



 点来表示outlier

### 衡量差异的最优方法

Deviation from Mean离均差： 找出每个值和均值之间差的平均值

Absolute deviation绝对偏差

Average absolute deviation平均绝对偏差

∑/N

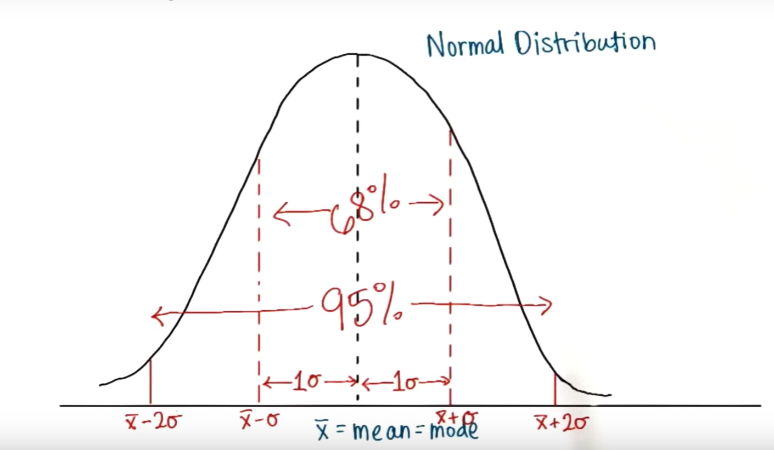
Squared deviation

Sum of Squares（SS）

Variance方差（平均平方偏差）

/N

Standard deviation标准差

σ =

### 为什么标准差重要？

正态分布

68%的数据在 平均数的正负一个标准差内

95%的数据在 平均数的正负两个标准差内

百分比和标准差的数量可以一一对应

### Bessel‘s Correction贝塞耳校正

Ex：在正态分布的数据集中，抽样一般sample都集中在中间部分，这样导致sample的差异性一般会小于population的差异性，即**低估**了总体差异性

校正后的公式：

Variance =

s = （sample standard deviation，样本标准偏差，约等于σ）

校正后使得差异性更大

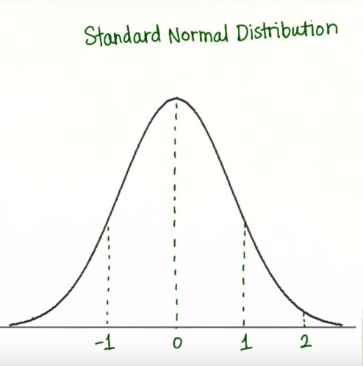
### Standardise归一化

Absolute frequencies vs relative frequencies

绝对频率，一般指具体数值

相对频率，一般指百分比，或比例

### Z

Number of standard deviations away from the mean

### 标准正态分布

数据归一化处理为Z值，

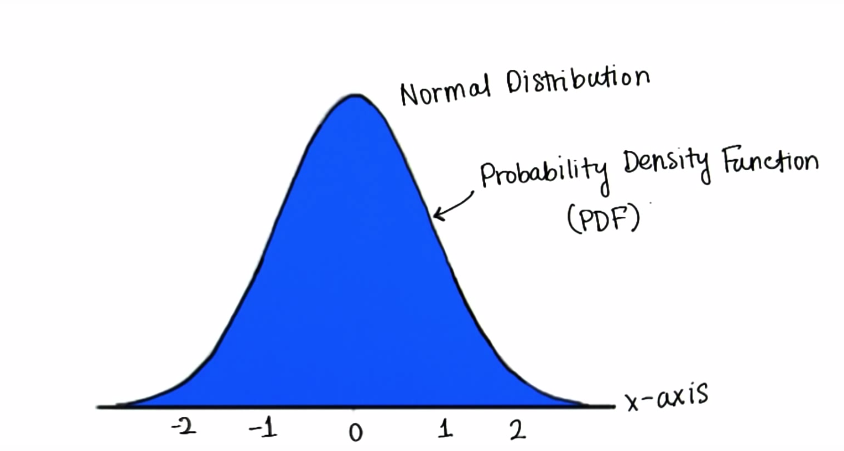
将原有均值移动到0，标准差除以本身变成1：

归一化分布的新标准差是1，平均值是0

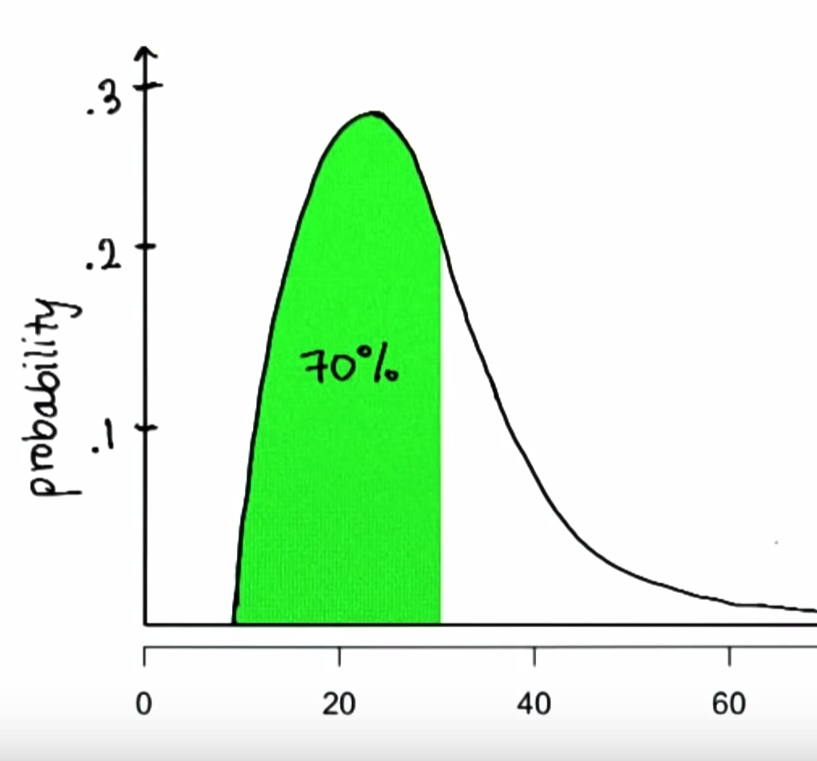
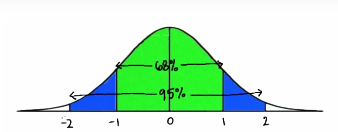
**正态分布**

### Probability Density Function概率密度函数 PDF

面积为1



可以用面积表示分布概率

### Horizontal asymptote 水平渐近线

正态分布中，概率密度函数x轴作为曲线末端的水平渐近线

5个标准差作为临界值，大于5个标准差的概率趋近于0

### P值 - Z表格

查询对应的P值和对应的Z值

链接比Z表方便：

http://sampson.byu.edu/courses/z2p2z-calculator.html

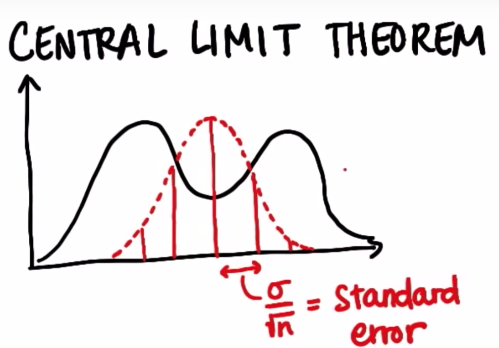
### Sampling Distribution 抽样分布 SE

If sample =n

σ/SE =

### Central Limit Theorem中心极限定理

* 分布的预测形状
* 分布的预期均值
* 分布的标准偏差（标准误差）



对于均值分布，其中每个均值都是样本量为 n 的均值，该分布的标准偏差就等于总体标准偏差除以平方根 n，这就叫做中心极限定理，即：

SE = σ/

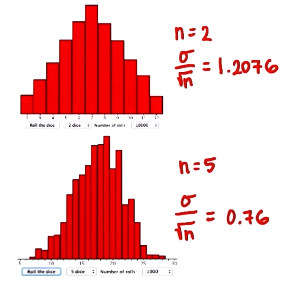
总体可以是任何形状的，即各种分布类型。

**SE本质上描述了抽样分布的平均数，相较总体平均值，偏离了多少**

所以SE就是Standard Error标准误差

随着样本量的增加，标准误差减小

N=1时，均值分布将和总体分布形状一样



Mean of sampling distribution M = population mean µ

### 计算特定样本的Z值

Z= or Z=

σ = SD（population）

n= size of sample

M = mean of population

### 样本均值的百分比

对于抽样分布，大概的误差范围：

68%的样本均值会落在个µ内

95%的的样本均值会落在个µ内

### Margin of Error误差范围

Margin of error =

， ） 大约95% of samples within this interval

: 为受到intervention干扰后的均值，不是μ，也不是M。

“干预因素”是指，我们预计会改变总体参数的因素。示例中，“干预因素”为 Bieber Tweeter

### Confidence Interval置信区间

近似的均值95%的置信区间为： ()

**样本量越大，总体参数在干预后的置信区间就越小**

“处理效应”会在干预因素影响到总体均值时显现出来。当样本均值位于抽样分布尾侧的远端，因而不可能是偶然事件时，这表明存在处理效应的证据。

### 精确的Z值

在干涉情况下，总体参数特定a%置信区间的计算方法:

p=(1-0.0a)/2;

再用P值找Z值

所以a%对应的精确置信区间范围为：

()

精确的均值95%的置信区间，Z值为（-1.96， 1.96）

所以准确的95%的置信区间为()

以此类推，98%的置信区间对应的Z值范围就是（-2.33, 2.33）

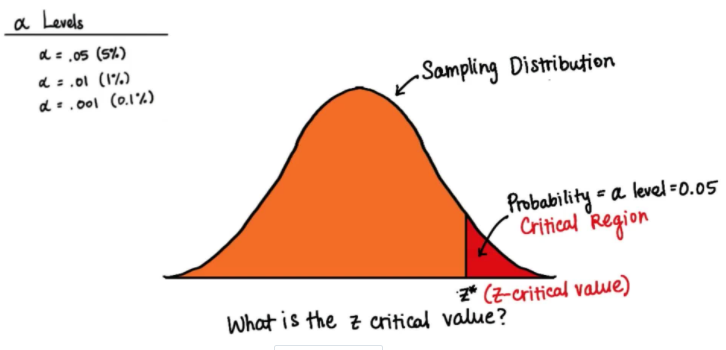
判定特定因素是否产生干扰的抽样统计流程

1. 找出总体均值M
2. 选定特定的干涉因素，并进行抽样
3. 计算抽样均值 和标准差 σ,
4. 根据σ和抽样数量n，计算抽样的抽样误差 sampling error SE
5. 根据情况计算特定样本均值的Zx值，Zx= 和对应概率P
6. 根据情况选定置信区间的%，找到对应Z值，计算抽样的置信区间
7. 如果M在置信区间内，说明因素未产生影响，如果在置信区间外，说明产生了影响
8. \*或者根据拒绝零假设的方法判断（见后文）

### α level

判定事件发生的可能性

如果概率小于α，说明不太可能发生

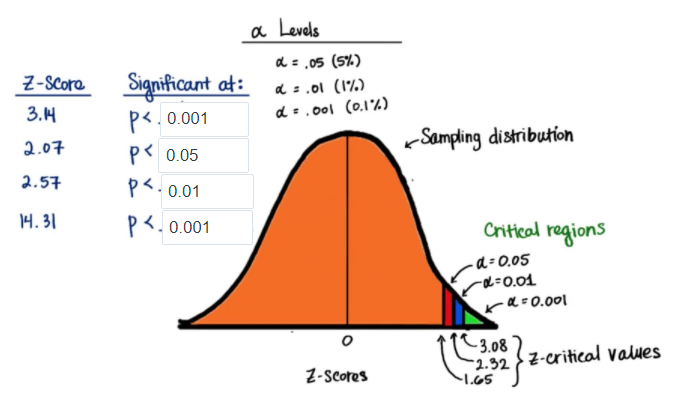


### Z Critical Value

如果样本均值对应的Z值大于/或小于α对应的Z值（趋于一端），说明不太可能发生

这个趋于一段的Z值，称为Z Critical Value，Z临界值

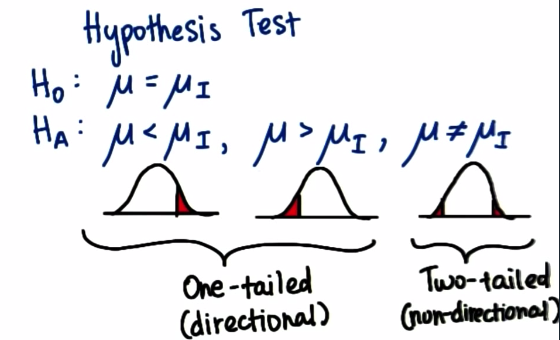
### Significant显著性



显著性是相对于随机发生的，即随机出现某事件的可能性。

所以p小于α level的事件，可以理解为**非偶然性事件**，即通过某种特定条件才发生的事件

### Two-Tails Test双尾临界值



单尾：向性检验

双尾：非向性检验

### H0 Null hypothesis 零假设| Ha Alternative hypothesis对立假设

H0：样本要么位于临界区外，或内，并规定总体参数μ = 干涉后总体参数μI

这里的= 表示**无显著差异**

Ha：μ <,>, ≠ μI，即总体参数和干涉后的参数间**存在显著差异**

### Reject the null 拒绝零假设

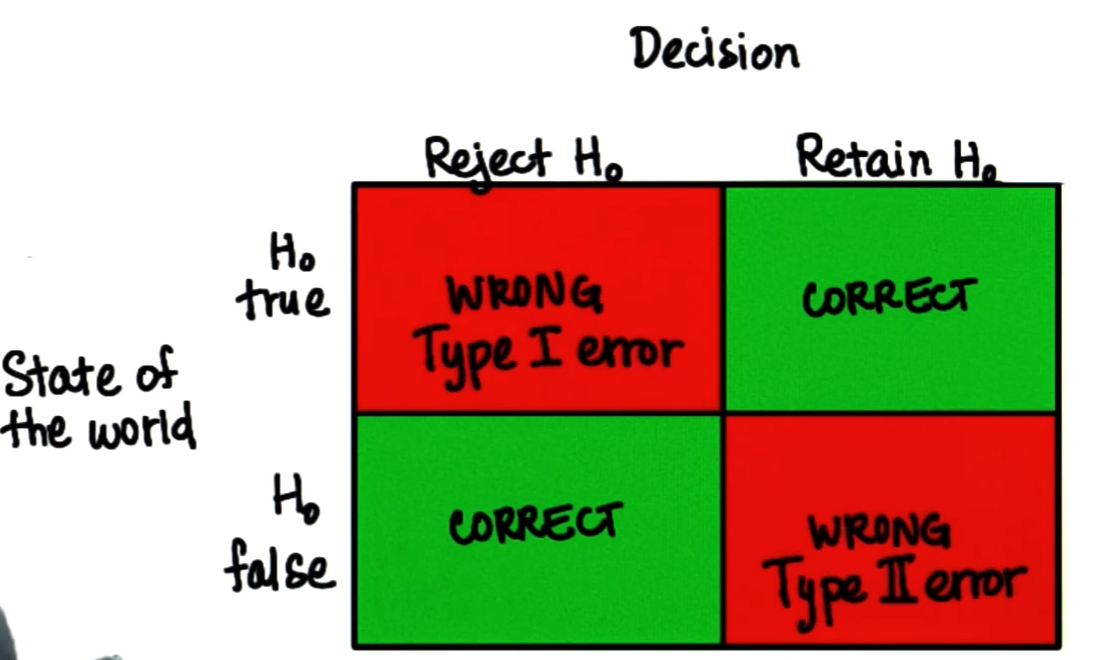
我们无法证明零假设，只能通过试图拒绝零假设，来看结果是倾向零假设还是对立假设

拒绝判定：

1. 样本均值在临界区内；
2. 样本均值的Z值大于Z临界值
3. 得到样本均值的概率小于α水平

Ex: Reject the null， when P < 0.05

### 决策失误

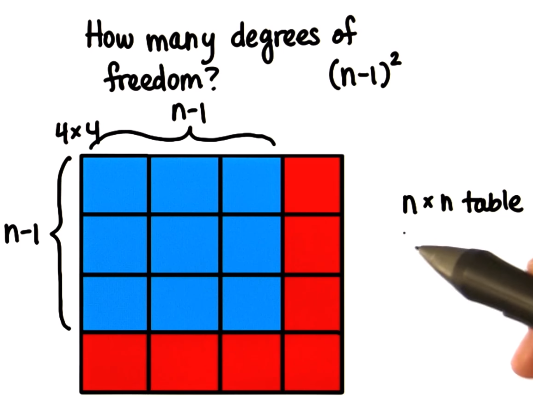


### t-test t检验

T = mean difference/SE

### Degree of freedom自由度df

T分布由自由度来定义



在贝塞耳校准公式中，

σ ≈ s =

其中N-1 被称为effective sample size有效取样数

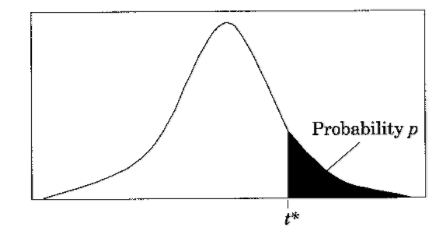
随着自由度df增大，t分布更趋于正态分布

### T表格

用t tale查t临界值

<https://www.graphpad.com/quickcalcs/pValue1/>

https://surfstat.anu.edu.au/surfstat-home/tables/t.php

单尾t检验

当我们期望结果是 < 0时，用negative的单尾检验，

当我们期望的结果是> 0时，用positive的单尾检验

双尾t检验： 自由度=n-1

### T分布与零假设

T统计量在任意方向远离0，即来自一个总体的某个样本均值，远离总体的均值，就拒绝零假设

T= ，S≈σ；为样本均值，μ0为总体均值，μ为所来自的总体的样本均值

会出现下面三种情况：

越大，越能证明μ大于μ0；

越小，越能证明μ小于μ0；

离μ0越远，越能证明μ≠μ0

所以越靠近两端，样本均值与总体均值μ0越有可能有显著不同

反过来推，T统计量绝对值，表面T统计量越远离μ0

### 总结：One-sample t-test单样本t检验

总结：

T=

因为我们只有一个sample，均值为

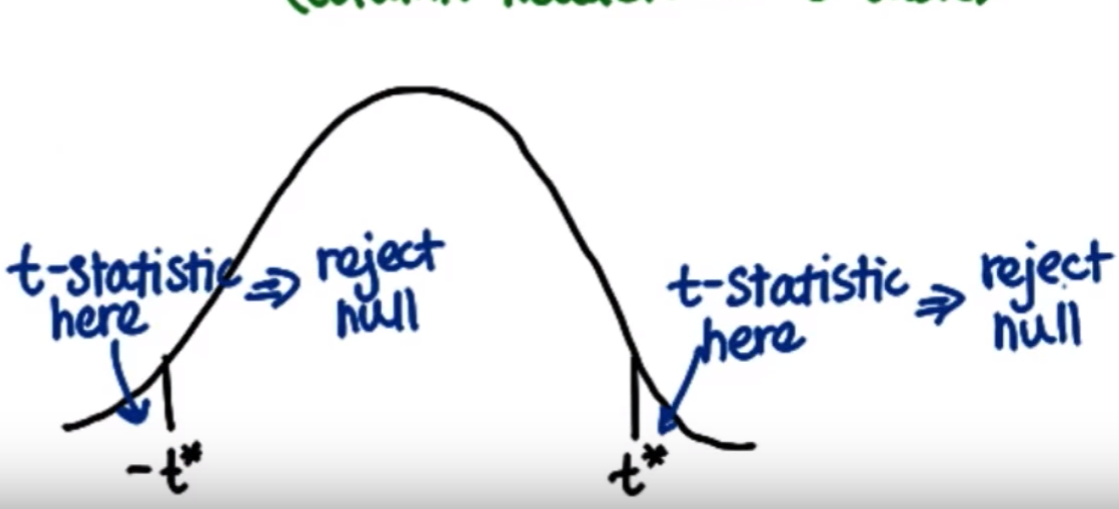
我们想知道该sample来自的总体，是否与具有μ0均值的总体显著不同？

所以，H0： μ = μ0

Ha: μ <,>,≠ μ0

α levels（column header of t table）

t表显示了划分α-level的t临界值，所以可以用来查找对应概率α-level的t临界值



然后，根据样本的t统计值和自由度，我们可以得到从μ中得到的概率P，进而和α-level对应的临界概率进行比较，如果P小于临界概率，拒绝零假设

### t值对应置信区间和误差界限

置信区间：

()

margin of error：

### Paired samples & Dependent t-test

相依样本和依赖性T检验

针对同一个被试做的不同实验，即同一个被试得到的不同样本，例如

1.同一个实验的两种条件；Repeated measure design

H0：μ1 = μ2

2. pre-test & post-test

H0：μpre = μpost

3. 或随着时间变化的同一个测试的不同取样（Longitudinal）

H0: μtime1 = μtime2

|  |  |  |
| --- | --- | --- |
| xi | yi | Di = xi - yi |
| x1  x2  x3  …  xn | y1  y2  y3  …  yn | D1 = x1 - y1  D2 = x2 - y2  D3 = x3 - y3  …  Dn = xn - yn |

### 比较两个相依样本间的差异

t =

S=

### Cohen’ d

标准化的均值差。即均值之间相差多少个标准差（或以标准差为单位，均值间相差多少）

d = 或d = MD/SD,  MD, or

CI: MD ± Tcritical (SD/), SD =

T =

### T检验 vs Z检验

当知道总体参数时，使用Z检验

当不知道总体参数时，使用T检验

### Effect Size效应量

实验性研究：指处理效应的大小

非实验性研究：变量之间的关系强度

在t/z检验中，最简单的EZ指标是mean difference均值差

Effect Size类型

1. Difference Measuress：

Mean difference 均值差异

Standardised differences 标准化差异

Cohen‘s d – SD units本质还是均值差异，用标准偏差做单位

2. Correlation measures:

**R2 -** 某个变量的变化比例（或百分比），与另一个变量的关系，即通过知道另一变量的值，来知道对应变量的变化比例

### Statistical significance统计显著性 vs 结果的意义

意味着：

1. 拒绝零假设
2. 结果不是偶然发生（或抽样错误导致的）

不是统计结果很重要的意思

所以meaningfulness of results，即统计结果是否重要，主要看：

1. 我们测量的变量是否具有实践、社会或理论研究价值；
2. Effect size效应量
3. 能否排除抽样错误（random chance）
4. 能否排除结果的alternative explanations（lurking variables）

### R2 -确定系数

一个比例，范围（0，1）

0表示两个变量没关系，1表示完全相关

t = t统计值

df = 自由度

### Result section有哪些内容？

1. Descriptive statistics，描述统计量，如mean，SD
2. Inferential statisctics，推论统计量，如假设检验，置信区间

* CI Eg: confidence interval on the mean difference: 95% CI=(4,6)
* 检验类型，Eg one sample test等；
* 统计数据，如t值，df，
* P值
* 检验方向，双尾还是单尾

用APA格式：

t(df) = x.xx, P = .xx, direction

1. Effect size measures

- r2  r2 = .xx

- d d = x.xx

### 相依样本 优势和劣势

优势：

控制个体差异

- use fewer subjects 被试更少

- 成本低

- 耗时短

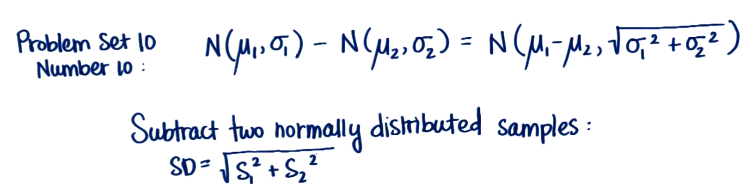
劣势：

- carry-over effects 残留效应

- 第二次measurement会受到第一次的treatment处理的影响

- 受试顺序会影响结果

### 两个样本集的正态分布差



SE= 假定两个样本大小差不多

Df = n1 + n2 – 2

Tstatistic =

CI: （) - tcritical · SE， ) + tcritical · SE）

### 合并方差 & 矫正的标准误差

Pooled Variance ：

SSx=

所以 校准的SE =

独立样本的完整t统计量计算公式

Tstatistic =

H0：μx -μy = a

A： expected difference

Assumptions：

1. X和y都是从各自的总体中随机抽取的
2. 总体基本是正态分布
3. 样本数据可以用来估计总体方差population variance
4. Population variance应该大体相同，使得我们能够使用pooled variance计算校正的标准误差