# Research Methods @2024/10/24

**一些杂谈：**

* 讲真我觉得这门课很像《概率论与数理统计》，至少这节课的Slides很像
* 自己重新划分了一下章节，见谅
* 可以抓学“实验心理学”的问问，真的很像他们每天对着问卷鼓捣的

## 1 Set Up

只是一些简单的前置知识…

### 1.1 Background

当你试图开展一个 统计向调研 时，你会需要的“三步走”战略：

1. 找到你需要调研的**问题**。

Sample：什么会导致患肺癌？

1. 理解你需要测量的**变量（Variables）**。

Sample：年龄、性别、每日抽🚬数 …

1. 学会**数据分析方法**，给出答案。

Sample：画“每日🚬数-患病率”图，证明每天抽的越多、患病率越高。

### 1.2 Notation 符号[[1]](#footnote-1)

主要用来描述“实验设计”，这里主要就仨：

|  |  |  |
| --- | --- | --- |
| 字符 | 描述 | Sample |
| R | 把调研对象随机分组 | 生物实验中，随机将小鼠分成两组 |
| O | 统计具体数据 | 统计 8:00~10:00 有多少人选了甜豆浆 |
| X | 施加条件 | 在食堂门口放个“施工中”的牌子[[2]](#footnote-2) |

### 1.3 问题定义（我加的）

这一节类似于个人手癖？好吧，只是表述上的癖好。为了您更好的理解我后面BB的“变量”啊、“目标”啊之类的东西，还得污染一下您的眼睛。

对于一个“问题”，我们往往希望探究（一系列）“（自）变量”，和“应变量”Y之间的关系，即找到函数，使得：

还是以“肺癌患病率”为例子，我们可以通过下面的方法描述其中涉及的一些因素：

* **目标**：也就是“应变量”Y —— 肺癌患病率
* **变量**：普遍指“自变量”X —— 年龄、性别、每天抽🚬数量 …

## 2 Types of Studies

统计向调研 视其侧重方向、深入程度不同，可以被分成三类：

1. Descriptive：“观察向”的研究，不会探究变量间的关系/变量-目标间的关系。
2. Relationship：又可以被分为注重？？的 Correlation & 注重？？的 Regression
3. Differences：接受/拒绝假设（再解释一下）

### 2.1 Descriptive Study

嗷，课件只给了两个 Sample（反正看起来就很像思政课问卷的那种）

**例一**：有多少（百分比的）人上了厕所之后不洗手？

|  |  |
| --- | --- |
| Event | 1[[3]](#footnote-3) |
| Group1[[4]](#footnote-4) | O[[5]](#footnote-5) |

如果我们想了解更加详细的信息，我们可以逮着“洗了手”的人问：“你们便后洗手的频次大概是\_\_?”

**例二**：有多少（百分比的）人上了厕所之后不洗手？

|  |  |
| --- | --- |
| Event | 1 |
| Group1(Yes) | O |
| Group2(No) | O |

### 2.2 Relationship Study

#### 2.2.1 Correlation Design

Correlation 的中文是“相关性”，虽然这并不重要。

Correlation注重的是 **变量之间** （而非“变量-目标”间）的**线性相关度**，即：一个变量随另一个变量的增长而增长/减少。

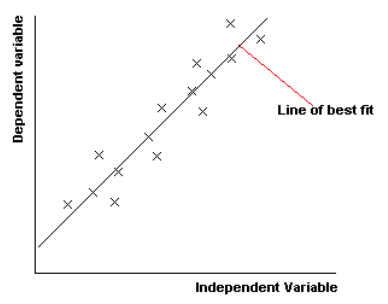
常见的关系有：客厅大小-房屋室内大小（基本高度正相关）、繁华程度-犯罪率（一般负相关）、出生地-生日（基本没关系）。

这个东西的意义在于：你拿两个**高度相关**的特征时，约等于只拿了一个（所以你可以丢掉其中一个、减少计算量）。

Correlation 的实验设计如下表[[6]](#footnote-6)所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject | V1 | V2 | … | VN |
| 1 | 0.0 | 1.0 | . | 0.2 |
| 2 | 0.1 | 0.5 | . | 0.5 |
| … | . | . | . | . |
| M | 0.5 | 0.6 | . | 1.2 |

然后，你从表里抽出**两个变量**（比如 V1，V3），然后画一个散点图（scatterplot）=> 你会得到下图里的 X 号；再拟合一条直线，你就会得到：



当斜率 > 0时，Correlation > 0、变量正相关；反之（略）。

值得注意的是，你需要在给定时刻测定所有的变量值——表中为 Vi（但我平时会表述为Xi），而不是在 t1 只测了 V1，在 t2 只测了 V2+ V3（你需要在每个 t 都把 V1~ VN 全搞定）。

##### 中插小知识：Correlation（相关性） & Causation（因果）

1. **Correlation** 的取值范围为 [-1, +1]，衡量的是“变量之间”的线性相关性。

* 不恰当的例子：“出生年份”+1 => “年龄”-1。
* 值得注意的是，correlation **不能解释“因果”**（具体例子可以看[附录1](#_1_为什么_Correlation)。）

1. **Caution** 好像没啥好说的，举个例子吧：摄入的激素药剂量上升会**导致**病人体重上升。

#### 2.2.2 Regression Design

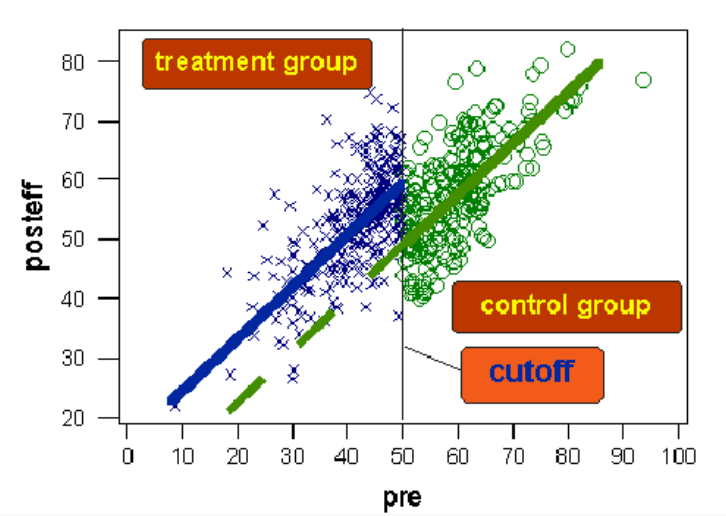
我感觉这东西和我平时学的 Regression 不太一样？

但总的来说就是一个“实验组-对照组”的设计，让我们把表[[7]](#footnote-7)端上来：

|  |  |  |  |
| --- | --- | --- | --- |
| Event | 1 | 2 | 3 |
| R Group 1 | O | X | O |
| R Group 2 | O |  | O |

首先，我们把被试者随机分成 Group 1 & 2 两组（这就是第一列的意义），其中 Group 1 为实验组、Group 2 为对照组。

我们在施加条件前（t=1）和施加条件后（t=3）时测量两组的相关数据，在t=2时刻对实验组（Group 1）进行相关处理。



绘制如上的散点图，图中蓝色的X来自“实验组”、绿色的X来自“对照组”。同样的，我们绘制由两组数据拟合的直线，不难看出两条直线间存在距离（蓝色实线x 绿色虚线）=> 这说明我们施加的实验条件确实影响了最终目标。

### 2.3 Differences Study

#### 2.3.1 Experimental Design

P16中的四句话翻译如下：

1. 采用随机分组（Random Assignment）的叫“Experimental Design（实验）”
2. 受条件限制不能随机分组/设立对照组的叫“Quasi-Experimental（类实验）”
3. 独立变量（Independent Variables）是 Categorical[[8]](#footnote-8) 类型的；
4. 非独立变量可能是 Categorical 类型 / Continual[[9]](#footnote-9) 类型的。

相关的实验设计如下表[[10]](#footnote-10)所示：

|  |  |  |  |
| --- | --- | --- | --- |
| Event | 1 | 2 | 3 |
| Group 1 | O | X | O |
| Group 2 | O |  | O |

其中，两组的被试人员**互相独立**。

#### 2.3.2 Factorial Design 析因分析[[11]](#footnote-11)

当**施加的实验条件X**中存在多个独立变量，并且你想探究这些变量之间的关联时，你可以采用如下表[[12]](#footnote-12)所示的设计：

|  |  |  |
| --- | --- | --- |
| Event | 1 | 2 |
| R Group 1 | X1A | O |
| R Group 2 | X1B | O |
| R Group 3 | X2A | O |
| R Group 4 | X2B | O |

表中的 (X1, X2) 是对同一变量的处理、(A,B) 是对另一对 Categorical类型变量的处理。本实验探究了三个问题：

1. 采用实验条件 X1 或 X2 会对实验人员造成什么不同的影响？
2. 组别 A、B 之间有什么不同吗？
3. 不同的实验条件组合会带来什么影响？

#### 2.3.3 Time Series Design

在此类研究中，你将以**固定时间间隔**采集数据（没错你就像一个打点计时器），然后观察数据在时间轴维度上的变化趋势。

课件强调了在：处理前、**处理过程中**、处理后，都需要采集数据；而 Experimental Designs 只在 **处理完成后** 采集数据。

举个不恰当的例子：Time Series Design 会要求你在给小鼠动手术的过程中要全程监控血压、心率啥的，但 Experimental Designs 只要在从手术室推出来之后测一次就可以了。

课件上的例子是：全程监控糖尿病人在实验条件下（我猜是吃饭/打胰岛素）血糖的完整变化曲线。

## 3 Statistical Tests[[13]](#footnote-13) 统计学检验方法

您统计的数据可以模糊的分成以下两个大类：

* 非参数统计 Non Parametric（定性数据），实际上有两种情况：

1. 统计的数据不属于任何一种特定的分布（比如一串完全随机的数）
2. 统计的数据有排名（rank / order），并且不能直接把排名对应到数值

* 参数统计 Parametric（定量数据）：统计的数据符合特定分布（比如“正态分布”）

对于“参数统计”，我们一般采用“假设-验证假设”的方法证明数据是否符合特定分布（误差足够小）。若符合，我们就可以对数据进行更准确的预测 —— 如果嫌疑人在一整片区域内行踪不定，我们就很难抓；如果他有明确的常用行径路线，我们就很方便包抄。

实际上，各种 Tests 只是用来比较采集的数据与标准分布之间差异性的方法。

### 不同的数据类型 Categories of Data

在手搓调查问卷时，我们可以将选项（Data）的数据分成以下四类[[14]](#footnote-14)：

1. Nominal：是一些“离散的”选项，你没办法给选项排序、或者给他们计算有意义的“距离”（比如平面/空间中两点之间的距离）。

Sample：{Male, Female, Else}

1. Ordinal：数据是“有序的”，但也没有“距离”的意义

Sample：请大家给出一个 1~10 的打分（但你不好说明给1分的 & 给5分的具体间的差距）

1. Interval Scale：某取值范围的一串点，所以具有“距离”的概念；但没有用于表示“XX不存在了”的 zero-point（0摄氏度也是有温度的）。

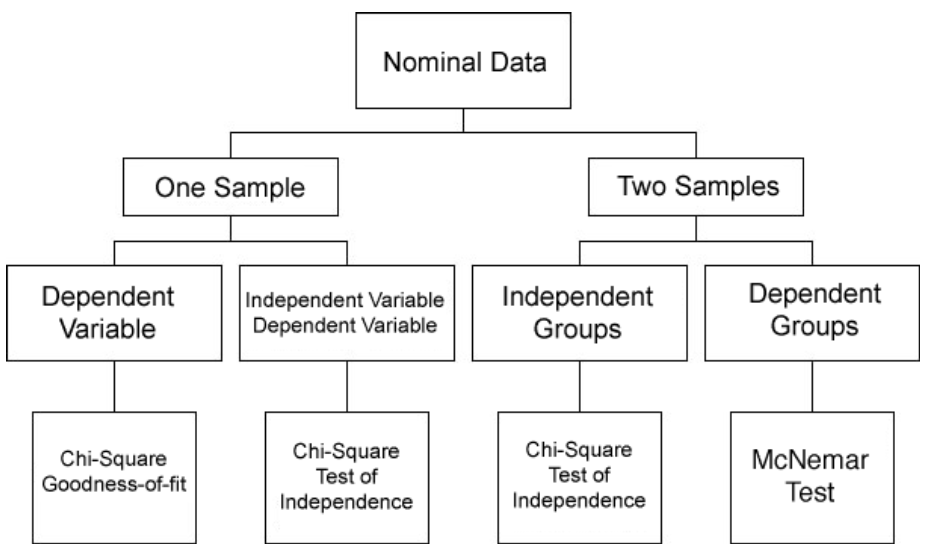
Sample：华氏度 / 摄氏度的取值（你可以具体量化 -1摄氏度 & 23摄氏度之间的 gap）。

1. Ratio Scale：取值范围内的一串点，同样有“距离”概念；但存在一个 zero-point 表示“往下啥也没有了”

Sample：在开尔文（热力学温度）下，不存在低于绝对0度的温度

### 用于 Nominal Data 的检验方法

来吧，上图[[15]](#footnote-15)：



怎么看这张图（我瞎掰扯的）：

* One Sample / Two Samples

这边其实是用的“方法数量”，belike：1-Sample 只要求你捅一个鼻拭子，但 2-Samples 的情况下可能是医生说“捅完再顺便取去拍一下 CT 吧”

* 然后你可能想探究一下 CT/鼻拭子 的监测结果是不是一致的（比如说有个逆天的 CT-鼻拭子 测试结果差不多是全反的）
* Dependent / Independent Variable

Dependent Var是“因变量/结果”Y（一般只有一个）；

Independent Var是“自变量/变量”X（可能会有多个）

好的，我们开始介绍底下4个不完全相同的测试方法。事实上，这四个都是卡方检验（Chi-Square Test）的子类 => 于是先唐突介绍起 Chi-Square Test：

**卡方检验（Chi-Square Test）：**

* 适用条件：所用的数据都是 Nominal 类型的，比如说 {‘阴性’, ‘阳性’}
* 目的：证明实际采集的数据和你的预期间是否存在较大差异

Sample：你估摸着老师是按正态分布给分的（但是你得验证一下）

**卡方分布（Chi-Square Distribution）:**

就不说它的数学定义了（但是求值需要查表），只要知道：越接近卡方分布说明 => 两个变量之间约接近相互独立

Sorry，这边没有按照从左到右的顺序来（把左边倆换了一下位置）。

**（1） 卡方独立性检验（Chi-Square Test for Independence）[[16]](#footnote-16)：**

你可以看到它的分支条件是 Dependence / Independence => 其实是你根本 不知道这俩变量之间有没有关联性，所以拿这个 Test 验算一下。

Sample：你不知道一个人“最喜欢的颜色”和他“最喜欢的运动”之间有没有相关性 —— 于是你开始了奇妙小调研。

一般来说，我们会把统计数据填入一张 X 行 Y 列的表格（其中X，Y分别为两个变量的可能取值数量）。好的，下面是一些你需要知道的数学符号：

1. 假设H，在这个测试里我们规定：

零假设H0： 我们所选的两个变量之间 **是** 独立的

替代假设 H1：我们所选的两个变量之间 **不是** 独立的

1. 自由度 ：在这里等于
2. 理论频数：公式是

其中N为样本总数、sum(x=i) 是第i行数据和、sum(y=j) 是j列数据和

1. 卡方统计量：重要的东西 —— 这个数越小 => 约接近卡方分布（两个变量互相独立），下面是公式：

其中， 是表格中第i行j列中的实际数值。

1. 显著性水平：我们允许统计数据和理想情况之间存在一些差距，这个值控制了允许的最大偏差，一般取0.05/0.01（分别对应95%和99%的置信度）。

好的，下面是激动人心的时刻 —— 上面算的狗屎都有什么用呢？我们什么时候应该接受H0（一口咬定两个变量之间独立），or 拒绝H0（咬定两个变量之间相关）呢？

答案是，符合以下条件时，我们接受H0：

当你选择 时，就是 => 左右两个数字都是查表得到的。

然后就没有然后了，还是康康sample 吧。

1. **卡方拟合优度检验（Chi-Square Goodness of Fit）：**

用于检验当前数据是否符合理论分布（树有没有按书上画的样子长），此时你手上只有一组数据（比如说一组大学牲的学历：{‘高中’,‘本科’, ‘硕士’,‘博士’}）。

* k：我们数据的总类数（在上面的学历sample中，k=4）
* 自由度：这次不用df来表示了，反正值等于 k-1
* 零假设H0：我们将所有标签记为{}，有

H0：理想的各标签占比为{}

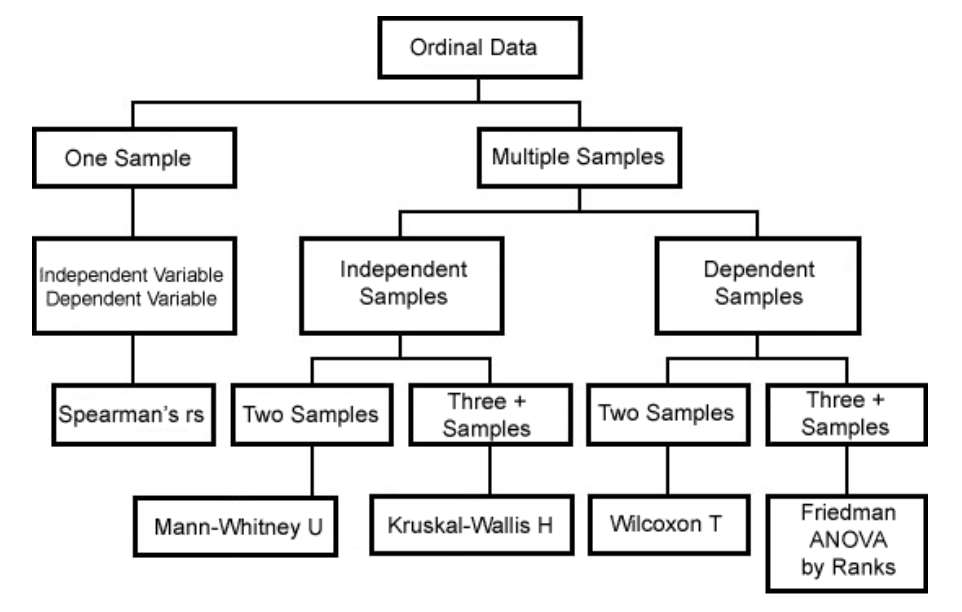
* 卡方统计量：其中N为样本总数、为该类标签的实际样本数量。

好的又到了接受/拒绝条件时间，当满足以下条件时，接受H0：

右侧的数也是查表得到的，注意这里是（不是前面的1-）。

捏妈妈滴剩下两个不是很想写了，先下一趴吧（反正PPT值介绍了Chi-Square Test）

### 用于 Ordinal Data 的检验方法



#### Spearman Rank Sum

* 用于**序列数据**，这里给一个Sample：在校运会中，假设有五名运动员、两名裁判：A裁判给运动员的表现排序是5-1-2-4-3，B裁判排的是4-2-1-3-5
* 我们需要如何衡量两个裁判给出的排序序列之间的相似性？
* 平时会用符号（rho）表示
* 取值范围是 [-1,+1]，其实两个序列之间的**相似程度**应该用的**绝对值**表示。

因为 -1 表示完全负相关（排名刚好倒过来），+1 表示完全正相关（完全一致）；但是0表示完全**无关**

* 计算：Python 里面有现成的函数，直接丢进去算就好了（手算好麻烦）

#### Mann-Whitney

* 对于**不符合正态分布的小样本**，我们要如何判断**两个分布**是否相似？
* 甚至可以用在两组大小不同的数据上（比如A组有20个人，B组有30个）
* 具体操作可以看 [这里](https://zhuanlan.zhihu.com/p/179273307)（反正Python也有参数）

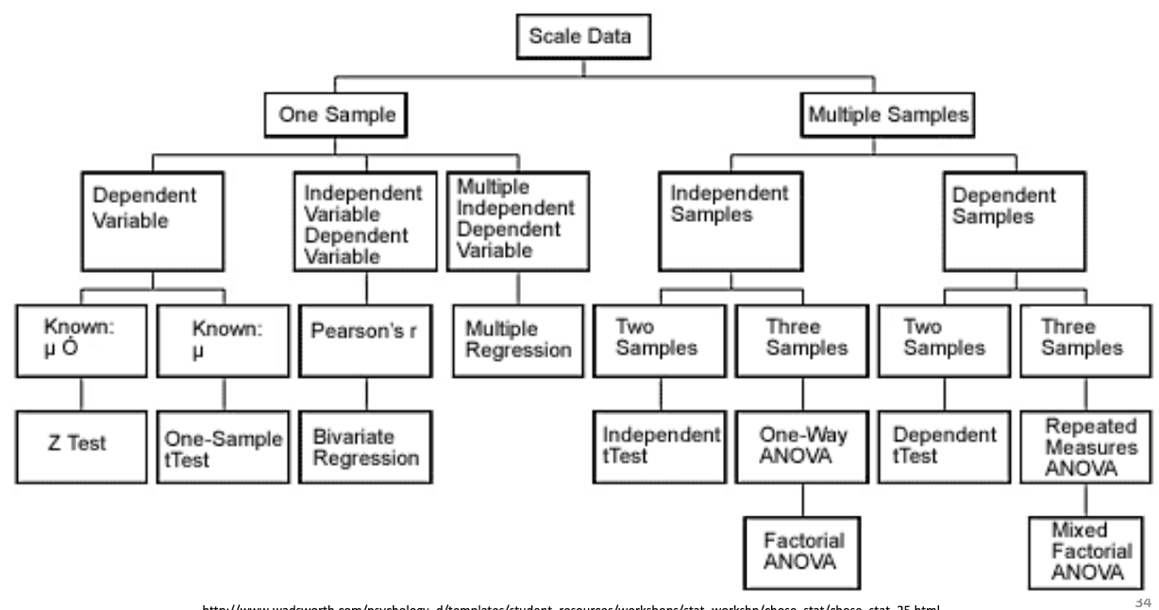
#### Kruskal-Wallis

* 对于**不符合正态分布**的样本，我们要如何判断**两个/多个分布**之间是否相似？
* 虽然不要求每组数据都长的和正态分布曲线一样，但要求各组之间的分布曲线比较像（不要求大家长得夺标志，但要丑的比较统一）

#### Wilcoxon T

* 适用于**长度相等、非正态分布**的两组数据
* 一般用于证明两组数据的**中位数**是否相等

### 用于 Interval/Ratio Data 的检验方法



#### Pearson’s r

* 用于度量两个**变量**之间的相关程度
* 类似的，这东西的取值是 [-1，+1]

0 表示两者完全无关，-1表示完全负相关，+1表示完全正相关

* 反映的是两个变量之间的**线性**关系 & 相关性的方向（其实就是散点图是否接近“直线”和直线的斜率）
* 关于 P36 中图的解释可以看 [百度百科](https://baike.baidu.com/item/%E7%9A%AE%E5%B0%94%E9%80%8A%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0/12712835)（笑）

#### T-test

* T检测有很多种，这里浅说一下单总体检验：
* 给定一个较小的、**符合正态分布**的数据集（数量<30），比较数据集的标准差（类似于平均数）和总体样本空间的均值之间的差异。

Sample：已知全体女性的平均寿命为A，你抓了一群人数<30、年龄符合正态分布的女性 => 比较你抓的女性的平均寿命和A之间的差异

* 其实也可以看 [百度百科](https://baike.baidu.com/item/t%E6%A3%80%E9%AA%8C/9910799) 上面的sample

#### Analyze of Variance (ANOVA) 方差分析[[17]](#footnote-17)

算了举抽象例子吧：你想知道一个人是否酗酒，然后你手机了他的年龄、性别、出生地、学历、职业等等一堆信息。显然，后面这一坨对“是否酗酒”都有影响。但是，你现在想搞明白**这一坨乱七八糟的东西之间**是否存在关联？

* 出生地会决定学历吗？学历会影响职业吗？把三个因素揉在一起考虑呢？

显然，一对对枚举会让人原地去世，此时就需要我们的 ANOVA 分析了。

1. One Way ANOVA 单因素分析

只关注一个变量（因素），可以用于比较2或更多组的数据

Sample：你想关注UI设计因素对浏览量的影响，然后你chuachua 搞出了ABC三版设计（活着更多） => 然后观察每两个组之间的差异性

1. Factorial ANOVA（其实就是Two Way）双因素分析

关心俩变量（因素），每个因素都可以拥有 >= 2种取值

Sample：“商品销量”可能同时受 价格&广告投放 两种因素的影响

* 价格/广告投放 **各自** 对商品销量的影响
* 不同价格 对商品销量的影响
* 不同广告投放 对商品销量的影响

1. Repeated Measured ANOVA（重复测量方差分析）[[18]](#footnote-18)

同一研究对象，在不同条件下测量相同指标，检验不同测量（如不同时点间的结果、不同方法的结果）之间的差异进行检验。

Sample：患者每天要做血常规，同一份血样会分成N份用不同方法测

## 4 Data Collection 数据收集

**如何测量变量：**

* 对于耗时、错误率、温度、距离等直观因素 => 直接测就行了
* 对于可用性、生活质量、压力水平等抽象因素 => 需要找一些表征旁敲侧击

**什么是这里的 Theoretical Construct：**

Theoretical Construct是指正在研究或检验的**抽象概念或想**法。它代表无法直接观察或测量的理论或假设实体，通常用于解释和理解各个研究领域中的复杂现象或关系。

**Construct 的有效性（Validity）和可靠性（Reliability）：**

* 有效性：你这个理论有多大程度上真的是真的呢（或许你只是抓住了谬误）
* 可靠性：在不同的背景条件下，这个理论的一致性如何呢（可以是恒真的吗）

**背景调查：**

* 有什么别人已经开发好的、用于测定 Construct 的工具吗？
* 这些工具在类似情况下是否被广泛使用（有没有啥约定俗尘的方法）
* 这些工具是否已经被学界公认为有效的、可靠的
* 如果你自己从头搓了一个工具，你得自己证明这为啥能work

## 5 Samples & Activities

### 5.1 一个实操Sample

来自Slides 的 P41～45 ，展示了一个数据软件的实操Sample：

**实验目的：**

探究新设计的仪器（Group2）是否能比原仪器（Group1）更快的完成任务。

**实验数据：**

用新旧仪器跑了很多遍任务，然后测量他们的每次耗时。

**操作：**

1. P42 选中了数据源区域 A1:B26 并新建了一个task – Descriptive Statics

作用是生成 P43 中这两列数据的相关统计数据

设置唯一参数Confidence Level for Mean（置信区间）= 95%[[19]](#footnote-19)

1. 进行一次 T-test：首先数据集比较小（只有25个数据，<30），其次测得的时间基本围着真正的平均值上下浮动（所以认为两个都是正态分布）

这里用的是 Paired Two Sample for Means 类型[[20]](#footnote-20)的 t-test（配对样本T检测，上面没有提到），目的是 **检测两个组的平均值是否相等（相对大小）**

**补充知识：这里的 One-tail 和 Two-tail 是什么**

* 单尾检验 One-tail：验证是否偏高/偏低、**只验证单一方向**。

假设本身是 x1 < x2，活着 x1 > x2

* 双尾检验 Two-tai：

检验抽样的样本统计量与假设参数的**差是否过**大（无论差值是正是负）

假设本身是 x1 != x2

我们可以看到他这里写的是 P(T<=t)，所以做的是 **单尾假设**

**那么，如何看这个单位假设对不对呢？**

**Case1:** 你**猜对了**方向时，one-tail 约为 0.5\*two-tail

**Case2:** 你**猜反了**方向时，one-tail 约为 1 – 0.5\*two-tail

此处，0.5\*two-tail = 6.93e-7（与one-tail相近），而1-0.5\*two-tail = 0.99。

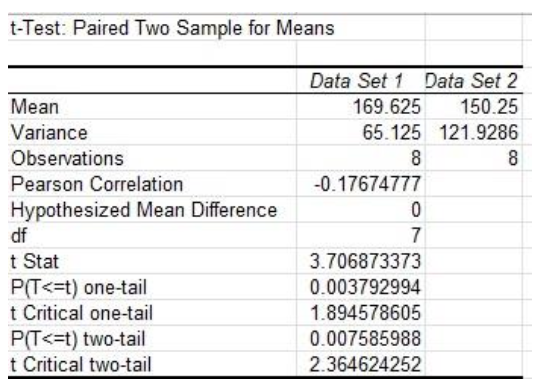
因此，假设 T<=t 是正确的。

### 5.2 Analyzing Data

* **实验要求：**对 P47 左侧提供的两列数据（Data Set 1 & 2）执行以下操作。

1. 生成描述数据（结果如P48左图所示）
2. 对两组数据做一个t-Test，求证两组数据的**均值**之间是否存在较大差异（结果如P48右图所示）

* **操作流程：**略，反正和 P42～45 一样。
* **结果分析：**此处仅对P48的右图（截过来放下面了）进行分析。



1. 简单的部分：
2. Mean 平均数：

从第一行的数据可以看出，两组的平均值分别为169.625 & 150.25，存在一定差距。**我们不能一口咬定Set1的均值（t）一定大于Set2（T）**，但我们不妨假设 t>=T，随后借助统计学知识进行检验

1. Variance 方差：

显然Set2具有更大的方差（121.9286），但这只能说明我们抽到的这些Sample的分布比较散；而Set1的分布比较集中。

1. T-Test（Paired 2 Sample for Means）结果分析：

让我们看倒数四行，显然这边也假设“T<=t”。和上一个文档易纲的，我们需要关注one-tail（0.037）& two-tail（0.0075）的数值。

显然此处 one-tail = 0.5\*two-tail，我们猜对了方向

* 这意味着 Set2 代表的整体所具有的“均值”确实小于Set1

### 5.3 Study Setups

看起来P49是课堂讨论的内容，这里就不放了。下面会对给的三个Topic按顺序介绍一下。

#### Topic 1

**问题描述：**

你设计了一个新的“洗涤-烘干”洗衣机套组，并且希望证明：

1. 你设计的洗衣机在工作上更高效。

=> 你的洗衣机洗+烘的比别人更快、平均耗时更少。

1. 你设计的洗衣机可以降低用户的劳累程度（Physical Stress）。

=> P51中说明用“下背部疼痛（lower back stress）”来衡量“劳累程度”。

**实验设计：**

这里选用的是Differences研究（但这不重要）。

总之就是采集一堆数据，然后运用统计学方法比较“原设计-你的洗衣机”在洗衣耗时、用户的下背部疼痛程度的“均值”上是否存在显著差异。

进一步的，你可以用第1节中的 t-Test 证明你的均值确实比原设计低。

**涉及的变量：**

1. Washer Design洗衣机设计：

Nominal 类型的变量，用于区分该条数据来自原设计 or 你的设计。

取值范围：{‘original’, ‘mine’}，你只能二选一

1. Transfer Time 洗衣用时：

Nominal 类型的变量，因为洗衣时间不可能小于0。

取值范围：(0, +inf)

1. Low back stress 用户的下背部疼痛：

Interval 类型的变量，代小数点的数

取值范围：不太清楚。老师用的NIOSH lifting function，大概用的LI指标。

**数据采集方式：**

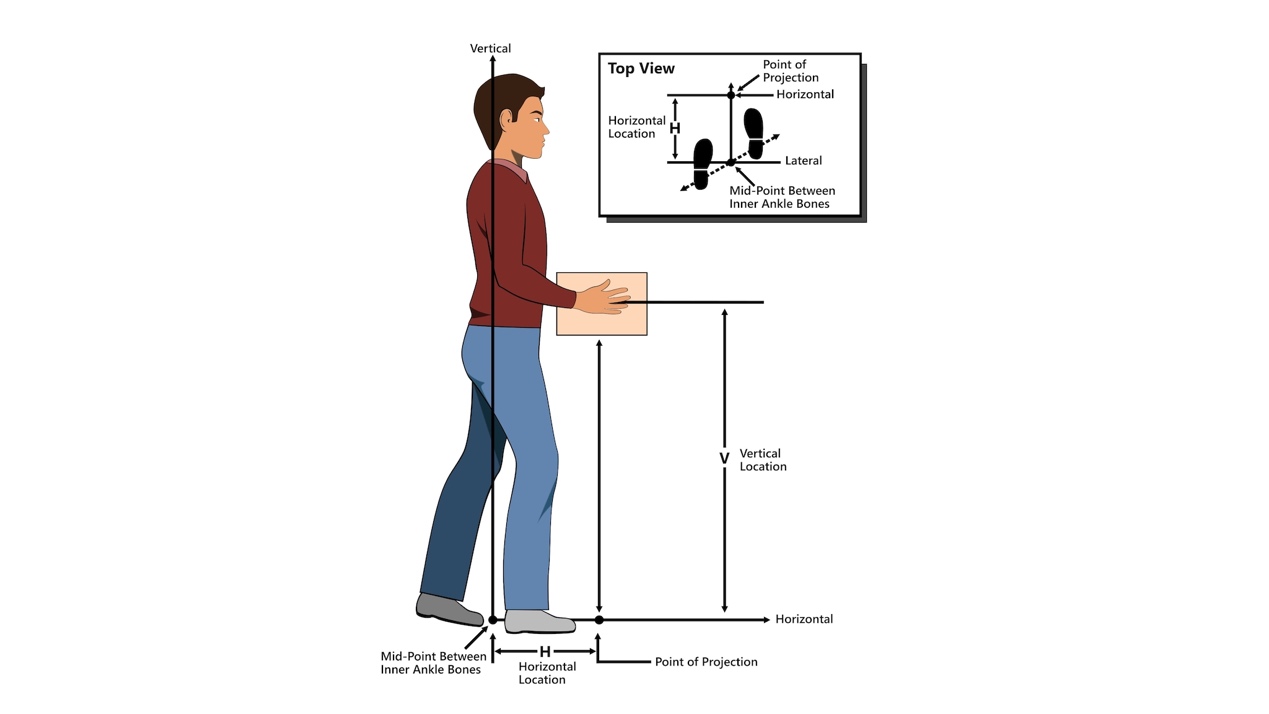
* 对于Transfer Time：按秒计时。

你需要把**单位质量**的衣服从洗到烘用的整个时长测出来。

* 对于Low back stress：用NOISH lifting equation计算LI。

算这个指标需要测量：衣物和身体的水平间隔H，物体相对地板的垂直高度（V），物体的垂直位移（D），人体旋转角度（A），抬举活动的频率（F）。

然后把这些数据填入在线计算工具就可以算出LI（lifting index），这个指标用于评估用户在抬举重物时受伤的受伤风险：LI<1时，受伤风险较小；1 < LI < 3时，有一定受伤风险；LI>3时，受伤风险很大、不可接受。



用于统计实验数据的Excel表格大概长这样：

|  |  |  |  |
| --- | --- | --- | --- |
| 实验编号 | Washer Design | Transfer Times(s) | Low back stress |
| 1 | original | 24.6 | 1.1 |
| 2 | mine | 27.1 | 0.5 |

又或者你可以按 washer design取值的不同，把数据拆分成Group1（original）和Group2（mine），变成P41左图那种形式。

**数据分析：Independent t-Test**

鉴于现有的数学分析工具（比如SPSS）已经提供了大多数检验的自动计算方法，所以=>不需要知道具体怎么算的，只要选对种类、看得懂蹦出来的数就好。

Independent t-test 允许两组的样本数量不同（比如说你只在原设计上洗了5次，但是在新设计上洗了20次）；但之前在用的paired-two sample t-test里，两者的样本数则应该严格相等。

Independent t-test的H0假设是，即两个序列均值相等（不存在显著差异）。

与One-Sample t-test相比，Independent t-test需要增加一步——比较两组数据的方差。可以看一下[这个例子](https://blog.csdn.net/m0_64456571/article/details/124329420)。

#### Topic 2

**问题描述：**新的操作面板布局能减少误操作。

**实验设计：**这里选的也是Difference类型

**涉及变量：**

1. Control Panel Design 控制面板上的按键布局设计：

Nominal 类型，{‘original’, ‘mine’}

1. Number of Errors 发生误操作的次数

Ratio 类型，[0, +inf)

**数据采集：**

* 对于 Number of Errs，操作的同时一个个去数就行

**数据分析：Independent t-Test**

和Topic 1本质上没有区别 —— 比较“原设计/新设计在Number of Errs的均值上是否存在差异”就好了。

Alternative

**问题描述：**新的操作面板布局在进行**多种不同操作任务**时，均能减少误操作。

**实验设计：**还是Difference类型

**涉及变量：**

1. Control Panel Design 控制面板上的按键布局设计：Nominal 类型
2. Number of Errors 发生误操作的次数：Ratio 类型
3. The Assigned Task 操作任务：

Nominal类型，{‘起飞’, ‘巡航’, ‘降落’}

**数据采集：**略，数数就行。

**数据分析：Factorial ANOVA**

显然，此处有两个变量：面板布局设计&任务类型。我们不仅希望探讨“面板布局”对误操作次数的影响，还希望探究“相同面板布局下、不同任务”的误操作次数均值是否不同；“不同面板布局下、相同任务”的误操作次数均值是否不同。

因此，这边要用到ANOVA，来探究两个变量独立作用下/协同作用下对目标（误操作次数）产生的影响。

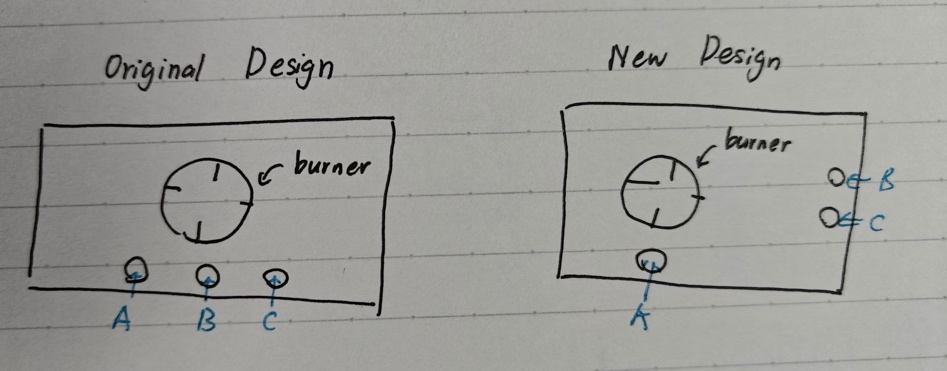
#### Topic 3

**问题描述：**

Users more intuitively associate the correct control to burner functionality of one of the new cooktop designs compared to the old?

（Maybe老师想说的是）用户是否主观认为：相比于原本的灶台布局设计，新的设计能更加正确的控制灶台（burner）？

**实验设计：**还是Differences



超劣质Sample：左右分别是新/旧灶台设计（那个长得跟准星似大圈是灶子）

在A，B，C三个旋钮中：一个用来控制灶子火的大小，另外两个用于控制油烟机和其他沟槽功能

**涉及变量：**

1. Cooktop/Control Layout 灶台布局设计：

Nominal类型，{‘original’, ‘mine’}

1. User Assumption用户假设：

Nominal类型，{‘A’, ‘B’, ‘C’}

**数据采集：**

搞一个问卷，让用户做单选题“你觉得哪个旋钮用于控制灶子”。

在有N个burner的情况下，你就让用户做N个单选就好了。

**数据分析：卡方独立性检验 Chi-Square Test of Independence**

之前解释过这个检测用于验证——两个Nominal类型变量之间是否存在（显著）关联？

但是，这个东西只能说明“灶台布局设计-用户选择”间**是否相关**。

**不能说明**原设计是否让人更倾向于选择A旋钮，新设计是否让用户更倾向于选择C旋钮（可行的佐证方案：用饼图可视化不同选项的占比）。

### 5.4 [课后论文](https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Choi_August2019.pdf)

**标题：**

**Applying Tangible Augmented Reality for Product Usability Assessment**

有形的现实增强技术（AR）在产品可用性评估上的应用

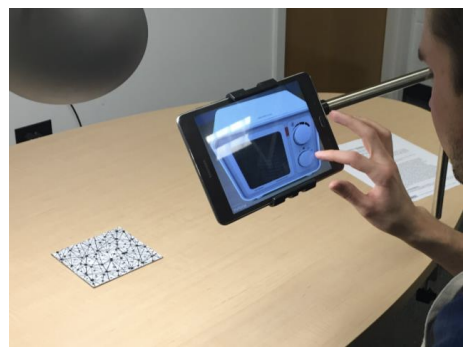
**内容：**

本文需要下面这个简单的加热器模型操作面板的用户可用性进行评估：



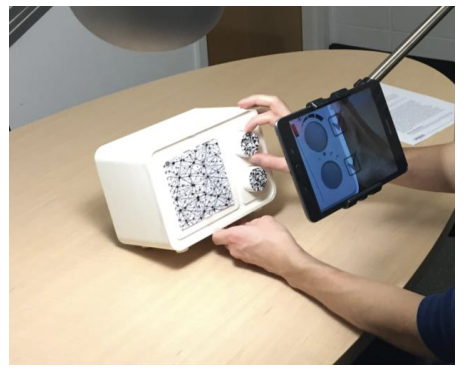
1. 传统的AR技术：

传统的AR识别一张Marker Card（桌面上那个花里吧啦），然后在显示屏上展示对应的3D建模。



1. 本文提出的Tangible AR技术：

就是在一个3D打印的白模上贴Marker Card（来可视化面板上的一些文字信息）。好消息是用户可以真的对白模上下其手，拧拧旋钮、开开柜门（然后平板会给出一些声音、数显上的反馈）。



**实验设计：**从论文P6的Evaluation这一节开始

1. 问卷设计

用户需要完成一份USE调研问卷（包含了30道题，每道题需要从1～7七个选项中选择最符合的一项），问卷对Usefulness（8道）、Ease of Use。（11道）、Ese of Learning（4道）、Satisfaction（7道）四个方面进行评估。最后的Usability Score由所有题目得分相加得到。

1. 数据分析

实验分了三个实验组（AR、有型AR、真实模型）。另外，每个问卷的选项其实是Ordinal类型的数据。

1. Kruskal-Wallis测试

用于确定**三个实验组**数据是否是**相似的**（并不影响用户的观感）。

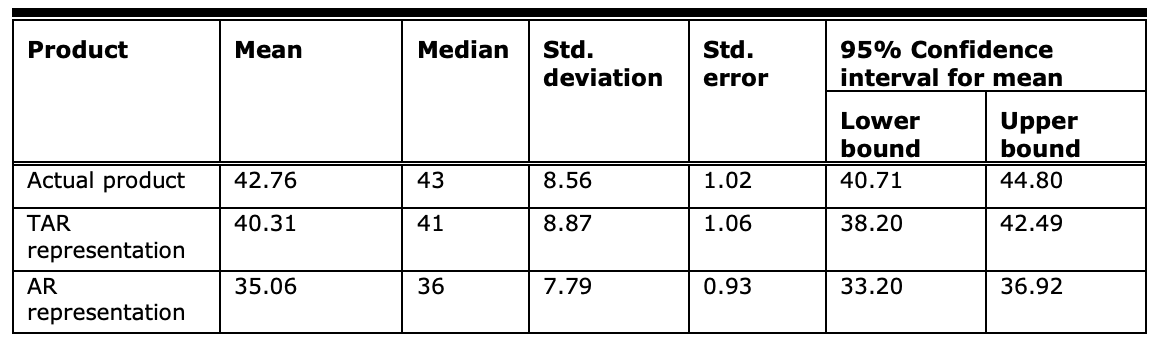
1. Mann-Whitney测试

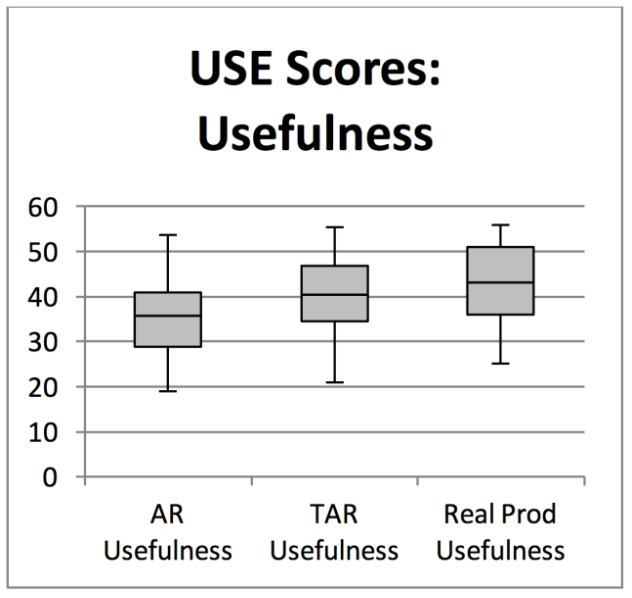
用于**两两比较**实验组数据之间是否存在差异（AR-TAR，AR-真实，TAR-真实）。

1. 分析结果

文章里实际上对总分（Usability Score）和四个方面的得分分别做了研究，下面以对总分的评估文段为例（因为剩下四段文字都是一个模子里刻出来的）。

首先让我把对应的两张图表丢上来，其实光从统计数据和箱线图已经差不多可以看出TAR技术比纯AR技术效果要好了（但我们还是需要统计学知识来严格证明：是真的好）。





A Kruskal-Wallis test was conducted to evaluate differences among the three groups (真实产品, AR产品模型,TAR产品模型)for the total USE questionnaire scores.

作者使用KW测试来对比三组在总问卷得分上存在的差异。

The test, which was corrected for tied ranks, was significant H-stat = 36.08, p < .001. (df = 2, N = 70)

通过tied rank方法对数据进行替换（看批注），KW计算结果如下：H-stat=36.08，p<0.001。

Follow-up comparisons using Mann-Whitney tests were conducted to evaluate pairwise differences among the three groups, controlling for Type I error across tests by using the Bonferroni approach.

此外，作者使用MW测试对数据组进行两两对比。同时通过Bonferroni法控制I类错误。

The results of these tests showed that the AR representation (M = 141.7, SD = 26.76) was rated significantly lower (p < .001) than the actual product (M = 169.79, SD = 27.14) for the overall total USE score. No significant differences were found between the total USE scores for the TAR representation and the actual product.

结果表明，AR模型组得分明显低于真实模型组的得分，TAR模型得分差异则与真实模型组无明显差距。

These results support the hypothesis that usability evaluations of a TAR representation of a product will be similar to evaluations of the product itself. The results did not support the hypothesis that usability evaluations of an AR representation would also be similar.

这些统计结果支持了我们的假设：使用TAR模型进行可用性评估的结果与使用真实模型的结果相近。同时也拒绝了假设：AR模型的效果也与真实模型相近。

## Appendix

### 1 为什么 Correlation != Caution[[21]](#footnote-21)

这边有三种潜在假设：

**（1）第三变量问题**

第三变量会在两个变量之间产生虚假关系。这取决于你所考虑的两个变量与第三个变量之间的相关模式。

**你知道冰淇淋销量和鲨鱼袭击之间存在正相关关系吗？**

现在，冰淇淋销量并不会导致鲨鱼袭击事件的增加。那么，这是怎么一回事呢？

原来，室外温度与冰淇淋销量和鲨鱼袭击机会呈正相关（因为更多的人去海边）。因此，当气温升高时，销售量和袭击次数都会同步增加，从而在两者之间产生虚假的相关性。

在这种情况下，我们称温度为潜伏变量或混杂因素。

**（2）因果关系的方向**

有时，两个变量可能存在因果关系，但**不清楚哪个变量是因，哪个变量是果**。

**研究人员发现，学生花在社交媒体上的时间与他们的学习成绩之间存在相关性。**

**情况1：**可能是花更多时间在社交媒体上分散了学生的学习注意力，导致成绩下降。

**情况2：**相反，学习困难的学生可能会转向社交媒体来逃避现实，这意味着成绩下降会导致社交媒体使用量增加。

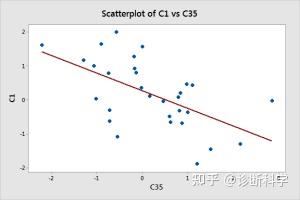
在这种情况下，究竟是使用社交媒体导致成绩下降，还是相反，尚不明确。

**（3）纯属偶然**

样本数据中的随机偶然性会在变量之间产生明显的关系。如果你收集了足够多的随机样本，随机性偶尔会在实际上不存在相关性的地方产生相关性的表象。

我曾在一次关于数据挖掘的座谈中谈到过这种现象：当你筛选足够多的数据时，完全随机生成的不应该存在任何相关性的数据是如何产生相关性的。

下图显示的似乎是一种负相关性，但我让计算机软件随机生成了许多变量，然后系统性地疏浚数据，找到了相关性。



**图1：显示两个变量之间明显关系的散点图**

由于这些原因，你可能会在数据中**看到相关性**，但其实**并没有因果关系**。或者至少你可能无法确定这一点。

1. Tic-tac-toe 就是“井字棋”游戏，这里只是借用一下符号 [↑](#footnote-ref-1)
2. 这里**不统计数据**，只是施加实验条件（对实验组） [↑](#footnote-ref-2)
3. （俺觉得）这表示“在 t 时刻执行的操作”：统计数据（O）/ 施加条件（X） [↑](#footnote-ref-3)
4. （俺觉得）此处的 Group1 表示：上厕所之后不洗手的人群 [↑](#footnote-ref-4)
5. （俺觉得）此处的“O”表示：填入“上厕所后不洗手人群”占总人群的百分比（%） [↑](#footnote-ref-5)
6. 就是 Slide P13 里的那个实验 [↑](#footnote-ref-6)
7. 是 Slide 里 P14 的那个实验 [↑](#footnote-ref-7)
8. Categorical：在给定的离散值中选一个，例如：“性别”的取值是 {男，女，其他} 中的任一种 [↑](#footnote-ref-8)
9. Continual：连续型变量，比如能任取 [0,1] 区间内的任何数 —— 0.1，0.11，0.9999999 [↑](#footnote-ref-9)
10. 是 Slides 里 P17 的那一张 [↑](#footnote-ref-10)
11. 可以看一下 [WIKI 中的介绍](https://zh.wikipedia.org/wiki/%E6%9E%90%E5%9B%A0%E5%AE%9E%E9%AA%8C) [↑](#footnote-ref-11)
12. 是 Slides 里 P18 的那一张 [↑](#footnote-ref-12)
13. 可以看一下这篇 [知乎文章](https://zhuanlan.zhihu.com/p/334574642) [↑](#footnote-ref-13)
14. 这边也有一篇 [知乎](https://zhuanlan.zhihu.com/p/144895107) 的解释，可以康康？ [↑](#footnote-ref-14)
15. Slides 里 P27 那张 [↑](#footnote-ref-15)
16. 这边有一个对于“卡方独立性测试”应该如何计算的 [详细例子](https://statorials.org/cn/%e5%8d%a1%e6%96%b9%e7%8b%ac%e7%ab%8b%e6%80%a7%e6%a3%80%e9%aa%8c/) [↑](#footnote-ref-16)
17. 让我们看 [知乎](https://zhuanlan.zhihu.com/p/454192298) [↑](#footnote-ref-17)
18. 来看看 [Sample](https://www.trialstats.com/SPSS/Repeated-Measures-ANOVA/) [↑](#footnote-ref-18)
19. 可以看看这里对 [95%置信区间含义](https://blog.csdn.net/pannn0504/article/details/82455934) 的解释 [↑](#footnote-ref-19)
20. 关于[“配对双样本T-test”](https://www.jmp.com/zh_cn/statistics-knowledge-portal/t-test/two-sample-t-test.html) 的具体例子 [↑](#footnote-ref-20)
21. 来自知乎：https://zhuanlan.zhihu.com/p/669044454 [↑](#footnote-ref-21)