

重温概率统计*

闫钟峰

目录

I	概率论	4
1	概率论基本知识	4
1.1	概率论基本概念	4
1.2	条件概率与独立性	6
1.3	全概率公式与贝叶斯公式	6
2	随机变量及其分布	7
2.1	随机变量及其分布函数	7
2.2	离散型随机变量及其分布	7
2.2.1	0-1 分布 (两点分布)	7
2.2.2	二项分布	7
2.2.3	泊松分布	8
2.2.4	截尾泊松分布	8
2.2.5	超几何分布	8
2.2.6	几何分布	8
2.2.7	负二项分布 (等待时间分布)	8
2.2.8	Zata 分布	8
2.2.9	幂级数分布	8
2.3	连续型随机变量及其分布	8
2.3.1	均匀分布	8
2.3.2	指数分布	8
2.3.3	正态分布	8
2.4	随机变量函数的分布简介	9

*本文最初是在参加 DdataWhale 的统计学时总结的学习笔记, 最初名为《重温统计学》, 完成学习后继续根据吴坚《应用概率统计》补充复习概率论部分 (该工作自 20190520 开始), 并根据该书重新调整了内容结构。下一步准备根据卡塞拉《统计推断》进一步完善。感谢组织者的辛勤付出。

3	多维随机变量及其分布	9
4	随机变量的数字特征	9
5	极限定理	10
5.1	大数定律	10
5.2	中心极限定理	10
II	统计学	12
6	统计学基本知识	12
6.1	统计学基本概念	12
6.2	抽样分布	13
6.2.1	χ^2 分布	14
6.2.2	t 分布	14
6.2.3	F 分布	14
7	参数估计	15
7.1	点估计	15
7.2	区间估计——置信区间	15
7.3	贝叶斯估计	15
8	假设检验	16
8.1	参数假设检验	16
8.1.1	z 检验和 t 检验	16
8.1.2	单样本 t 检验 (z 检验)	17
8.1.3	独立样本 t 检验 (z 检验)	18
8.1.4	相依样本 t 检验	18
8.1.5	χ^2 独立性检验	18
8.1.6	F 检验简介	19
8.2	非参数假设检验	19
8.2.1	χ^2 拟合检验法	20
8.2.2	偏度、峰度检验法	20
8.2.3	秩和检验法	20
8.2.4	其他非参数假设检验法	20
9	方差分析	22
9.1	单因子方差分析	22
9.2	因子方差分析	24
9.3	复测方差分析	24

10 回归分析	25
10.1 相关系数	25
10.1.1 皮尔逊积差相关系数	25
10.1.2 线性相关的统计显著性检测	26
10.1.3 其他几种类型的相关系数	27
10.2 一元线性回归	27
10.2.1 一元线性回归方程	27
10.2.2 最小二乘法	28
10.3 多元线性回归	28

摘要

本次学习内容主要包括：统计学基本知识、二项及泊松分布；大数定律、正态分布；中心极限定理、置信区间；假设检验（一）；假设检验（二）；线性回归；卡方分布与方差分析等。

Part I

概率论

1 概率论基本知识

概率论是一门研究非确定性现象的数学分支。它和测度论有着非常紧密的关系。

1.1 概率论基本概念

确定性现象是指在给定条件下一定会发生或一定不会发生的事件。与之相对的，**随机事件**指的是在给定的条件下可能发生，也可能不发生的事件。

随机事件发生的不确定性，不是事件本身不明确，而是发生的条件不充分，使得在条件与事件之间不能出现确定性的因果关系，从而导致事件的发生与否上表现出不确定的性质，这种不确定性就是**随机性**¹。

在给定的条件下，对于随机事件，如果我们多次进行观察或者试验，会得到不同的结果的不同的出现次数。我们对某一随机现象进行 n 次观察或进行某项试验 n 次，如果事件 A 发生了 n_A 次，则称 A 在给定条件的 n 次实现下的**频数**是 n_A ，而称 $f_n(A) = \frac{n_A}{n}$ 为事件 A 为在给定条件下的 n 次实现之下的**频率**，它反应了事件 A 在给定条件下发生的可能性。当我们进一步考察事件 A 的发生频率 $f_n(A)$ 时，会发现如果实现次数 n 比较大时， $f_n(A)$ 有经常性地接近某一个常数的趋势；如果实现次数 n 不断增大，我们会发现这种接近程度越显著，即在大量的试验和观察下呈现出明显的规律性——这就是**频率稳定性**。

¹对于很多我们通常归属于随机事件的案例，在不同的讨论角度和层面上是有不同看法的。比如说“掷一枚均匀的硬币，出现正面和反面的结果是随机的”，如果我们依据力学原理细究其运动过程，根据初始状态及受力分析其运动过程，只要各种数据量都可以精确测量，理论上是可以精确预测一次投掷的结果是正反面的。但这个分析过程在绝大多数时候是不能也无法进行的，这也就是说，我们把投掷硬币得到正面或反面，在缺乏足够的领域知识和观测数据的情形下，视其为随机事件。

实际上，在很多缺乏领域知识或（和）观测数据的情况下，我们都可以将无法提前预知结果的事件视为随机事件，这也是我们使用统计方法（或者说数据科学）对专业领域的数据进行分析的逻辑前提。

例如，我们使用如下 SQL 代码实现从男女两个性别中随机²抽取一种性别：

```
1 WITH A AS
2   (SELECT XB
3    FROM
4     (SELECT XB, COUNT(*) SL FROM BENKE GROUP BY XB)
5    ORDER BY DBMS_RANDOM.VALUE)
6 SELECT * FROM A WHERE ROWNUM=1 ;
```

如果使用循环语句多次随机抽取，并统计在这这个固定的抽取次数的时候，结果为男所占的比例，就会发现这一比例会比较接近 $\frac{1}{2}$ 。

```
1 --使用for循环，多次选取性别，并计算多次选取结果的男性比例
2 --补充代码！
```

当我们进一步增大随机抽取的次数，并多次观察时，会发现这种接近性会随着随机抽取次数的增多而更加接近。

```
1 --使用for循环，多次选取性别，并计算多次选取结果的男性比例
2 --增大选取的次数
3 --并按照更大的选取次数，重复进行上述选取和计算
4 --补充代码！
```

随机现象有其偶然性，但也有其必然性，这种必然性表现为大量观察或试验中，随机事件发生的频率的稳定性，即随机事件发生的频率在某个定值附近摆动，而且随着试验次数的增多，摆动总体上会逐渐减小，这种规律性我们称之为**统计规律性**。频率稳定的这种统计规律性表明随机事件发生的可能性大小是随机事件本身固有的，是不随人们意志而改变的一种客观属性，因此可以对其进行度量。

对于随机事件 A ，我们依据上述论述，有理由认为在频率稳定性中 $f_n(A)$ 所围绕变动的那个与 A 有关的固定常数刻画了 A 的一个重要特性——事件 A 发生的可能性大小，记该数为 $P(A)$ ，称 $P(A)$ 为**随机事件 A 的概率**。这是一个介于 0 和 1 之间的数。

以上只是概率的统计定义，并不能作为严格的数学定义，严格的概率定义是借助于概率的公理化结构而给出的。有了概率的概念，我们就可以对随机现象进行定量研究了，这就是**概率论**这门数学分支的目的。

一般地，设 E 为一试验，如果不能事先准确地预言它的结果，而且在相同条件下可以重复进行，就称 E 为一次**随机试验**。一次随机试验 E 的可能结果（记为 ω ）称为**基本事件**，也称为样本点。样本点的全体 $\Omega = \{\omega\}$ 构成了**样本空间**。

²注意这里的随机，严格来讲是一种所谓的“伪随机”，关于“伪随机”，可参考网络上的相关文章。

在具体问题中，明确认清基本事件空间 Ω 是由什么构成的，是非常重要的。它是描述随机现象的第一步。**事件**可以定义为样本空间 Ω 的一个子集，称**某事件发生当且仅当其包含的某一基本事件（样本点）发生**。

由于事件定义为集合，因此事件之间的关系和运算满足集合论的结果。例如，事件的包含表示事件 A 的元素（样本点）也是事件 B 的元素；事件的相等表示两个事件作为集合互相包含；事件的和表示两个集合的并集；事件的差表示两个集合的差集，事件的积定义为两个事件同时发生，表示两个集合的交集；事件的互斥表示两个事件的积为空集；事件的互逆表示两个互斥事件的并集是真个样本空间。

以下列出一些常用的事件的运算律：

1. (事件的和与积的) 交换律
2. (事件的和与积的) 结合律
3. 分配率
4. 德·摩根律

概率的公理化定义。

定义 1.1. 设随机试验 E 的基本事件空间为 Ω ，对于 E 的每个事件³ A 赋予一个实数 $P(A)$ ，如果它满足如下三条公理：

1. $0 \leq P(A) \leq 1$;
2. $P(\Omega) = 1$;
3. 对于可列无限个两两互不相容的事件 $A_i (i = 1, 2, \dots)$ ，有

$$P(A_i) = \Sigma P(A_i)$$

则称 $P(A)$ 为事件 A 的概率。

概率具有非负性、规范性和有限可加性。

古典概型

1.2 条件概率与独立性

条件概率是指

³一般来说不把 Ω 的所有子集都视作事件，因为这涉及到了测度及可测集的定义，如果将不可测集视为一个事件，则无法给它赋予一个恰当的实数。详细的论证参见测度论相关书籍。

1.3 全概率公式与贝叶斯公式

全概率公式:

贝叶斯公式:

2 随机变量及其分布

2.1 随机变量及其分布函数

2.2 离散型随机变量及其分布

离散型随机变量

本章的以下小节将分别介绍一些常见的离散型随机变量与连续型随机变量分布族。

首先是离散型随机变量的分布：

2.2.1 0-1 分布（两点分布）

又称为伯努利分布。

2.2.2 二项分布

二项分布是最基本的离散分布。二项分布就是重复 n 次的伯努利试验。所谓的伯努利试验，就是只有两个可能结果的试验。二项分布必须满足如下四个条件：

1. 每次试验的结果都是互斥的两个事件之一。
2. 不同的两次试验是相互独立的，即一次试验的结果并不会对另外一次试验的结果造成任何影响。
3. 每次试验中，每个可能的结果事件的发生概率都是相等的，记为 p 。
4. 试验次数是固定的，记为 n 。

在一个 n 次伯努利试验中，出现 k 次事件 A 的概率为

$$\binom{n}{k} p^k (1-p)^{(n-k)}$$

其中 p 为一次试验当中事件 A 的发生概率。

当我们考察二项分布的概率密度函数时会发现，随着试验次数 n 的增加，相应的概率密度函数逐渐变得形状接近于钟形曲线。事实上，根据中心极限定理，当试验次数 n 趋向于无穷大的时候，二项分布的概率密度函数和正态分布是完全一致的。

期望值 $E(X)$ 是总体均值 (μ) 的概念当总体数量是无穷大时的推广。二项分布的期望值 $E(X) = np$

2.2.3 泊松分布

在上述二项分布中，当 n 很大，而 p 很小，且 np 也是一个比较小的数时，二项分布就近似于泊松分布了。泊松分布是一种较为常见的重要分布，社会科学和物理学中的很多现象都符合泊松分布，泊松分布可以用来描述大量试验当中稀有事件出现次数的概率分布模型。以下给出泊松分布的正式定义：

定义 2.1. 如果一个随机变量的所有取值情况为非负整数，并且取各个值的概率分别为

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \lambda > 0, k = 0, 1, 2, 3, \dots$$

则称随机变量 X 服从参数为 λ 的泊松分布。

显然，泊松分布的概率密度函数即为 $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ ，这里，常数 λ 即为泊松分布的数学期望 $E(X)$

2.2.4 截尾泊松分布

2.2.5 超几何分布

2.2.6 几何分布

2.2.7 负二项分布（等待时间分布）

又称为帕斯卡分布。

2.2.8 Zata 分布

2.2.9 幂级数分布

2.3 连续型随机变量及其分布

连续型随机变量

连续型随机变量以下介绍一些常见的连续型随机变量分布族

2.3.1 均匀分布

2.3.2 指数分布

2.3.3 正态分布

正态分布也被称为高斯分布——事实上正态分布有很多种不同的名称。正态分布是自然界中最常见的分布，现实世界中满足正态分布的现象有很多，当然自然界中也有很多不满足正态分布的现象。

当总体满足正态分布，并且总体（或者用以估计总体的样本）的均值和标准差已知时，对于总体中的某个个体数据，我们可以通过计算其对应的 z 分数 ($z = \frac{x-\mu}{\sigma}$) 来快速获取该个体数据所处的百分位，因此正态分布有时也被称为是 z 分布。

除了通过对个体数据计算其 z 分数进而快速获取这个个体数据在总体中所处的百分位之外，我们还可以计算从总体中随机抽取某一个体，这个个体取某一具体值的概率是多少。

正态分布的概率密度函数

$$f(x) = \frac{1}{\sqrt{2\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}$$

在很多初等统计学教材中，正态分布都是占据了主要地位的。关于正态分布的更多信息，见网上广为流传的《正态分布的前世今生》。

下面我们介绍随机变量函数的分布：

2.4 随机变量函数的分布简介

随机变量的函数，顾名思义是指自变量是随机变量的函数，因此随机变量的函数也是一个随机变量。由于随机变量有各种各样的分布，因此随机变量的函数（作为随机变量）也应该会服从某种形式的分布。通过随机变量的分布去求随机变量函数的分布，在统计学中有重要的应用。最常见的随机变量的函数是线性函数，它和原始随机变量服从相同的分布。

当函数形式为二次函数时，情况就比较复杂了。例如对于服从标准正态分布的随机变量，其平方和构成的函数服从所谓的 χ^2 分布。

χ^2 分布是一类重要的分布，它和 t 分布、 F 分布并称为三大抽样分布。 χ^2 检验是一种重要的非参数检验方法。

3 多维随机变量及其分布

4 随机变量的数字特征

5 极限定理

极限定理的内容十分广泛，其中重要的两种形式为大数定律和中心极限定理。

5.1 大数定律

尽管随机事件的结果是无法确切预知的，但随机事件发生的频率是具有稳定性的：当试验的次数不断增加的时候，随机事件发生的频率将逐渐稳定于某一个常数，这个常数就是随机事件发生的概率。

多次观测一个随机变量的取值并将观测结果取均值，随着观测次数的增加，这个均值将趋近于随机变量的数学期望。

大数定律保证了我们在讨论某一事件有多大可能性时，可以使用（多次观测得到的）频率来替代（很难甚至无法确切得知的）概率。因此，大数定律为推断统计学提供了理论保证。

大数定律有多种不同的表现形式，例如切比雪夫大数定律、辛钦大数定律、伯努利大数定律等等，定理的具体形式和证明可见教科书。

例如，使用投针法来计算圆周率，就是利用了大数定律。从直观上看，当我们向一个画出了内切圆的边长为 $2r$ 的正方形内投针时，投入到圆内的概率，就应该等于圆的面积除以正方形的面积（ $\frac{\pi r^2}{4r^2}$ ）。因此，根据大数定律，当我们不断重复投针实验时，随着重复次数的不断增大，观测到的投入圆内的实际次数除以总的投针次数，就会逐渐接近上述理论计算值。当然在实际观测中会发生投针次数增多但比例远离理论计算值的情况，但总体上而言，观测值是会逐渐接近理论值的。

5.2 中心极限定理

前一节提到的通过使用个体数据的 z 分数来快速获取其在总体中所处的百分位的方法，有个前提条件是：所研究的总体符合正态分布。然而现实中很多时候所研究的总体并不满足正态分布。

中心极限定理指出，无论总体的分布如何，只要抽取的样本的容量足够大（例如 $n \geq 30$ ），那么样本的均值的抽样分布就符合正态分布。中心极限定理使得我们能够基于正态分布来进行统计推断，尽管总体并不符合正态分布。

独立同分布的中心极限定理：

定理 5.1 ()。设 X_1, \dots, X_n 是从某个均值为 μ ，方差为 σ^2 的总体中随机抽取的样本。当 n 充分大时， X_1, \dots, X_n 的均值 \bar{X} 满足正态分布

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (1)$$

上述定理是说，独立同分布的随机变量 X_1, \dots, X_n 的均值，当 n 充分大时近似服从于正态分布，而不论 X_1, \dots, X_n 所服从的分布是什么。独立同分布形式的中心极限定理为大样本推断提供了理论基础。

李雅普诺夫形式的中心极限定理进一步去掉了同分布的限制，无论随机变量 X_1, \dots, X_n 分别服从什么分布，只要他们的数学期望和方差满足一定的条件，那么随机变量之和的均值仍满足正态分布。

关于李雅普诺夫定理的具体内容可参考概率统计教材。

中心极限定理保证了抽样分布的均值符合正态分布。在复习正态分布时，我们提到了对于满足正态分布的总体，可以使用 z 分数来快速计算从总体中随机选取一个个体取某一个具体值的概率是多少。和对个体的情形类似，由于抽样分布的均值符合正态分布，因此我们现在也可以在总体不满足正态分布的时候，对于从总体中随机抽取的一个样本，计算其样本均值取某一具体值的概率。类似于计算 z 分数，这时候我们也需要通过计算另外一种类型的 z 分数—— t 值。回忆 z 分数的计算公式，我们发现其计算原理是使用个体值与均值之差除以标准差。具体到抽样分布的 t 值，我们就要使用样本的均值减去总体均值得到的差，除以一个另一种形式的标准差，即均值标准误。

均值标准误的计算公式

$$\sigma_X = \frac{\sigma}{\sqrt{n}} \quad \text{当总体标准差已知时}$$

$$\sigma_X = \frac{s}{\sqrt{n}} \quad \text{当总体标准差未知时}$$

Part II

统计学

6 统计学基本知识

统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议。

也可以说，统计学是收集和分析数据的艺术。

6.1 统计学基本概念

这里我们不加说明地罗列一些统计学里的基本概念，有些概念是如此直观，以至于不需要用更多的语言去解释和说明他们。

1. 总体 (population) 就是研究对象的全部。
2. 样本 (sample) 是从总体中选出来的一部分。
总体相当于讨论问题的基础集。而样本则是总体的一个子集，样本所包括的元素个数称为样本容量。从总体中选择样本的方法有很多种，比如随机抽样、典型抽样等等，选择样本的出发点是要保证样本具有代表性。
3. 参数 (parameter) 是用来描述总体特征的量。
4. 统计量 (statistic) 是指从样本计算而来的主要用于推断总体参数的量。
5. 均值、中位数、众数是用来描述数据的集中趋势的测度。

总体均值用 μ 表示，样本均值用 \bar{x} 表示。

当我们能够掌握总体的所有我们关心的数据的时候，我们就可以直接使用描述统计学来研究总体。当我们只能获取总体的样本，希望能够使用样本数据来推断总体的某些特征的时候，我们就需要使用推断统计学了。

极差、方差和标准差这三个量是用来描述数据的离散程度的测度。稍微详细说明一下方差和标准差的概念与区别：

1. 极差是指所有数据中，最大值和最小值的差。它给出了所讨论数据集合的总体离散程度。

2. 方差和标准差是用来描述数据的平均离散程度的量。需要注意的是，总体的方差、标准差与样本的方差、标准差的计算方式是不一样的。下面分别给出这两个量的定义。

总体的方差

$$\sigma^2 := \frac{\sum (x - \mu)^2}{N}$$

总体的标准差

$$\sigma := \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

样本的方差

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

样本的标准差

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

之所以样本的方差与标准差的计算公式中，分母使用的是样本数减一，是为了能够更好地使用从样本计算出来的方差去估计总体的方差。关于这个选择的更加详细的解释，可以参考常见的概率统计教科书，如较常见的浙大版的概率统计。方差多用于方差分析，作为分子出现。

下边是两个稍微抽象一点的概念，其完整的定义形式可参考概率统计教科书：

定义 6.1. 随机变量是表示随机现象各种结果的变量。随机变量可以分为离散型随机变量（可能的取值结果只有有限个或稀疏的可数无穷多）和连续型随机变量（可能的取值结果构成某个实数区间）。

定义 6.2. 连续型随机变量的概率密度函数是描述这个随机变量的在某个确定的取值点附近的可能性的函数。

以上快速回顾了统计学中的基本知识和定义。接下来复习了主要的抽样分布。

6.2 抽样分布

统计量的分布又称**抽样分布**。 χ^2 分布、 t 分布、 F 分布是最重要的三种抽样分布，后续的假设检验与方差分析部分，都是以这三种分布为基础的。

6.2.1 χ^2 分布

χ^2 分布的定义：

定义 6.3. 若 n 个相互独立的随机变量都服从标准正态分布，则这 n 个随机变量的平方和构成的新的随机变量的分布规律称为 χ^2 分布，其中， n 称为自由度，同时也是 χ^2 分布的均值。

根据中心极限定理，当 χ^2 分布的自由度特别大时， χ^2 分布就会近似于正态分布。

6.2.2 t 分布

6.2.3 F 分布

7 参数估计

7.1 点估计

7.2 区间估计——置信区间

当我们从样本计算出来的统计量并猜测它就是总体参数时，我们就是在做点估计——统计量的是数轴上的一个点。尽管事实上用从样本得到的统计量来推测总体参数差不多是我们所能想到的最好的办法，但通常而言，样本统计量并不会和总体参数一致。我们知道，当我们再次选择一个不同的样本时，那么从这个样本得到的统计量势必有很大概率和第一个样本得到的统计量不一致。每次选择不同的样本（尽管样本容量一致），对于同一种统计量（例如均值），我们得到的统计量的数值总会多少有些差异。于是我们希望能够知道，点估计的这种因偶然因素导致的误差应该怎么衡量？

区间估计能够很好的解决上述疑问。置信区间是一种常用的区间估计方法。所谓的置信区间就是说总体参数在一定的百分比（例如 95%）落在了我们通过计算得到的区间内。假如说我们通过样本计算出了某一参数的 95% 置信区间是 (a, b) ，这也就是说，我们相信总体的参数有 95% 的概率落在这一区间——等价于说总体参数只有 5% 的可能不在这个区间。这个参数不在区间内的百分比，称之为显著性水平，用 α 表示。通常在社会科学领域， α 多取 0.05。显著性水平表示我们愿意以多大的比例做出（因样本的随机性而导致的）错误判断。这种因样本的随机性和我们选择的显著性水平 α 而发生的错误，称之为第一类错误。虽然 0.05 的显著性水平广为应用，但也受到了很多诟病。

置信区间给出了关于点估计的精确度的信息。置信区间的具体计算方法取决于所讨论的统计量。一般来说，置信区间由样本统计量 \bar{X} 和给定显著性水平的误差范围确定，例如 t 检验的 95% 置信区间的计算公式是：

$$CI_{95} = \bar{X} \pm (t_{95})(s_{\bar{X}})$$

其中 $s_{\bar{X}}$ 是标准误。计算置信区间的过程是：先选定显著性水平，通过显著性水平和自由度（大致相当于样本容量减一）通过反向查表等方式获取相应的 t 值，然后由样本来计算均值和均值的标准误，最终根据上述公式计算出给定显著性水平下的置信区间。

7.3 贝叶斯估计

8 假设检验

在总体的分布函数完全未知，或者仅知道其形式但不知道具体参数等情况下，为了从样本数据推断出总体的某些特性，就需要先提出关于总体的一个假设（零假设）或两个假设（零假设和它的否定：备择假设），然后使用样本数据对这个假设进行检验：假设零假设正确，那么我们得到当前样本的概率是多少（这个概率称为 p 值），如果概率非常小（通常会取 p 值小于 0.05），我们就有理由做出拒绝零假设、选择备择假设的决策。假设检验就是这一做出假设并进行验证最终做出决策的整个过程。

假设检验的核心思想是“小概率事件在一次试验中是几乎不可能发生的”，先假设零假设成立，然后运用统计分析方法进行推理，然后得到该零假设发生的概率是一个极其小的数值⁴，从而推翻最开始的零假设，假设检验使用了反证法的逻辑推理过程。

假设检验根据其检验的对象可分为参数假设检验和非参数假设检验：当总体的分布类型已知，仅对未知的参数提出假设进行检验，称为参数假设检验；除了参数假设检验之外的假设检验都称之为非参数假设检验。本章主要复习参数假设检验方法。

假设检验的一般步骤：

1. 根据要研究的问题，提出零假设 H_0 和备择假设 H_1 ；
2. 选择适当的统计量，使其在 H_0 成立的条件下服从某种确定的分布；
3. 依据实际问题确定某种显著性水平 α ，确定用于作出决策的拒绝域；
4. 根据统计量的分布和显著性水平，确定拒绝域，即确定临界值（过去多是用现成的分布函数表查表获取临界值）；
5. 计算统计量的观测值，若其落入拒绝域则拒绝 H_0 ，否则接受 H_0 ；

8.1 参数假设检验

8.1.1 z 检验和 t 检验

z 检验和 t 检验的过程本质上是一致的，差别主要在两种检验所适用的场合（主要包括总体的是否符合分布、标准差是否已知、特别是样本容量的大小）。当我们事先已知总体满足正态分布时，或者尽管不知道总体是否满足正态分布，但样本容量比较大（ $n > 30$ ）时，根据中心极限定理，大样本的抽样分布也符合正态分布，这时候就可以使用 z 来进行检验。但当样本容量很小时（一般指小于 30），样本均值的抽样分布与正态分布有较

⁴事实上常用的 $p=0.05$ 的显著性水平在某些场合实际是不够严格的，因为这表示我们有 5% 的可能犯第一类错误，即：零假设为真但我们的检验拒绝了零假设

大的差异（实际上它满足 t 分布， t 分布是形状类似于正态分布的一种分布，这里不做详细介绍，可参阅统计学教材），这时就不适合使用正态分布和 z 检验，而应该使用 t 检验。

实际上用于支撑 t 检验的 t 分布，当样本容量逐渐增大时，会逐渐趋于正态分布。

8.1.2 单样本 t 检验 (z 检验)

单样本的 t 检验可以用来根据从符合正态分布的总体中抽取的样本数据来检验总体均值与给定值的大小关系。见下例：

例 1. 已知某厂生产的灯泡的使用寿命符合正态分布，方差 $\sigma^2 = 196$ 。某次检测抽取了 6 个产品做使用寿命检测，测得使用寿命分别为：3236, 2918, 3192, 3189, 3207, 3120（单位：小时）。问该厂灯泡的寿命是否可以认为是 $\mu = 3200$ 小时？

由于已知总体满足正态分布，并且已知总体方差，因此可以根据正态分布，使用 z 检验进行假设检验。以下按照假设检验的一般步骤来进行 z 检验：

1. 作出零假设， H_0 ：该厂灯泡的寿命为 3200 小时，即 $\mu = 3200$ 。那么相应的备择假设就是 H_1 ： $\mu \neq 3200$ 。
2. 计算样本均值 $\bar{X} = \frac{3236+2918+3192+3189+3207+3120}{6} = 3143.67$ 。

虽然这时已经知道样本均值并不等于题设中给定的 3200，但仍然存在的问题是，他们之间的差异是否是统计显著的。而这正是我们需要利用统计分析来给出的判断。

3. 计算 z 分数

$$z = \frac{\bar{X} - \mu_0}{s_{\bar{x}}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} =$$

- 4.

- 5.

使用单侧检验还是双侧检验取决于我们事先设定的零假设的否定是否包含了两种情况。例如零假设是药物无效，则这个命题的否定（也就是备择假设）是药物有效。但药物有效包括有正的作用和负的作用，因此这时候就适用双侧检验。但如果已经先验地知道了检验的拒绝域只会位于分布的一侧，则此时就需要使用单侧检验。

8.1.3 独立样本 t 检验 (z 检验)

独立样本 t 检验用来考察两个相互独立样本在给定变量上的均值是否有显著的差异。

使用独立样本 t 检验的前提条件

1. 独立性：两个样本相互独立，即从一总体中抽取一批样本对从另一总体中抽取一批样本没有任何影响，两组样本个案数可以不同。
2. 正态性：样本来自的两个总体服从正态分布。在样本的总体不满足正态条件时，如果两个样本的分布形状相似，他们的样本量相差不大并且样本量较大时，仍可用 t 检验。
3. 待比较的两个样本方差相同。如果两个样本的样本量大致相等，略微偏离了方差齐性对检验结果的精度影响不大。

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

其中样本均值的标准误的计算公式为

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}$$

如果两个样本的方差不是大致相等，样本容量也非常不同，同时（或者）数据不服从正态分布，那么独立样本 t 检验就不再适用了，这时候就应该使用另外一种替代性的非参数检验方法：曼-惠特尼 U 检验。关于曼-惠特尼 U 检验可参考非参数检验的相关教材。

8.1.4 相依样本 t 检验

相依样本 t 检验又叫做配对 t 检验，用于比较单个样本在两个不同时间点的取值之间的差异。

8.1.5 χ^2 独立性检验

χ^2 独立性检验适用于考察样本的两种分类变量之间的关系。 χ^2 独立性检验可以确定样本对象落入各类别的比例是否与随机期望比例相等，也就是说，考察样本对象是否均匀地分布于不同组别。

例如，当我们考察大学生的专业选择是否受性别影响时，我们可以把大学生按照专业进行划分，结合性别对学生再次进行划分，这样就得到了一个列联表。

表 1: χ^2 独立性检验的性别与专业数据					
性别	专业	心理学	数学	...	英语
男生		p_{11}, q_{11}	p_{12}, q_{12}	\cdots	p_{1j}, q_{1j}
女生		p_{21}, q_{21}	p_{22}, q_{22}	\cdots	p_{2j}, q_{2j}

我们可以把每一个细分的组的学生们的实际数量，填写到列连表里的 p_{ij} 的位置，这就是**观测频数**，然后我们还可以计算**预期频数**（纯粹因随机性而应该落入表格相应位置的预期数量）填写到 q_{ij} 的位置。

接下来，我们使用如下公式计算所谓的 χ^2 值。

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

上式中 O 表示各个单元格的观测值（观测频数）， E 表示各个单元格的预期值（预期频数）。

此外，我们还需要一个称为自由度的值。在 χ^2 独立性检验中，自由度的计算公式为

$$df = (R - 1)(C - 1)$$

上式中 R 和 C 分别是列连表的行数和列数。

计算出自由度之后，结合设定的显著性水平通过查表就可以得到相应的临界值，通过比较 χ^2 值和临界值的大小，就可以确定这种差异是否统计显著。

8.1.6 F 检验简介

从两个符合正态分布的总体中选取的相互独立的两个样本组，如果两个总体的总体方差相差不大（称为满足方差齐性），需要检验两个总体的方差的大小，这时就适用 F 检验。 F 检验中最重要的一种形式就是方差分析，关于不同类型的方差分析，留待最后一章再复习。

8.2 非参数假设检验

前一节已经说过，非参数假设检验是除了参数假设检验法之外的假设检验方法，因此非参数假设检验包括了很多种检验方法。有效使用参数假设检验需要满足若干假设条件，但在很多时候这些条件无法满足，因此非参数假设检验方法是在实践中。笼统地讲，非参数假设检验就是在推断过程中不涉及有关总体分布的参数，仅仅利用样本数据来对总体的分布形态等等信息进行假设检验。前一章在复习独立样本 t 检验时提到过的曼-惠特尼 U 检验就是一种典型的非参数假设检验方法。

8.2.1 χ^2 拟合检验法

χ^2 拟合检验法用于在总体分布未知时，根据样本数据来检验关于总体 X 的分布的假设

H_0 : 总体 X 的分布函数是 $F(X)$

H_1 : 总体 X 的分布函数不是 $F(X)$

的方法。

8.2.2 偏度、峰度检验法

偏度、峰度检验法是用来检验分布是否为正态分布的一种检验法。

8.2.3 秩和检验法

秩和检验法用于对服从相同分布（但分布的形态未知）、概率密度函数仅相差一个平移的两个连续型总体的均值是否相等进行比较。

8.2.4 其他非参数假设检验法

以下是另外一些来自于百度百科的非参数检验方法，留待后续学习：

1. 两独立样本的非参数检验
2. 曼-惠特尼 U 检验
3. K-S 检验
4. 游程检验
5. 极端反应检验
6. 多独立样本的非参数检验
7. 中位数检验
8. Kruskal-Wallis 检验
9. Jonckheere-Terpstra 检验
10. 两配对样本的非参数检验
11. McNemar 检验
12. 符号检验

13. Wilcoxon 符号秩检验
14. Friedman 检验
15. 多配对样本的非参数检验
16. Cochran Q 检验
17. Kendall 协同系数检验

9 方差分析

一个变量的方差有多少可以与另一个变量共享，或由另一个变量所解释？这个问题就是方差分析的核心问题。我们在假设检验章节中复习的 F 检验方法时提到，方差分析是 F 检验的一种重要形式，它是通过计算 F 值来进行的假设检验。

方差分析的前提条件是：

1. 所有样本都是相互独立的；
2. 所有样本来源的总体都服从正态分布；
3. 所有总体的方差都相等（即所谓的**方差齐性**），但方差的值是未知的；

9.1 单因子方差分析

单因子方差分析要解决的问题类似于独立样本 t 检验，它们都是想要得到不同组的均值之间的平均差异相对于各组内部平均差异而言是否统计显著，从而得出结论认为因为分组变量的取值不同，导致另外一个变量的均值产生了显著（或不显著）的差异。

但他们之间也存在着差别：

1. 支持比较的组数不同。独立样本 t 检验仅仅能够对两个独立样本进行检验，而单因子方差分析则可以对两个或更多的独立样本进行比较；
2. 所使用的比值不同。独立样本 t 检验用的是 t 值，而单因子方差分析使用的是 F 值（通过使用某种方差计算出来的值）。
 - (a) t 值的分子是两个样本均值之间的简单差异 ($\bar{X}_1 - \bar{X}_2$)；而 F 值的分子是使用称为组间均方 (MS_b) 的量来计算三个及以上样本均值的平均差异的。
 - (b) t 值的分母是均值之差的标准误，本质上是某种形式的标准差，而 F 值的分母是使用称为组内均方 (MS_w ，也称均方误) 的量来计算不同组的均值差异的。

要计算组间均方 MS_b ，需要先计算组间平方和 SS_b

$$SS_b = \Sigma[n(\bar{X} - \bar{X}_T)^2]$$

然后用组间平方和除以 SS_b 的自由度 ($K - 1$) 就得到了组间均方

$$MS_b = \frac{SS_b}{K - 1}$$

要计算组内均方 MS_w ，需要先计算误差平方和 SS_e

$$SS_e = \sum \sum (X - \bar{X}_i)^2$$

然后用组内均方除以 SS_e 的自由度 ($N - K$) 就得到了组内均方，也就是均方误

$$MS_e = \frac{SS_e}{N - K}$$

上式中， \bar{X} 表示各组的均值， \bar{X}_T 表示所有组的样本合并之后计算出的均值， n 表示各组样本的对象数。

最后我们给出 F 值的计算公式

$$F = \frac{MS_b}{MS_e}$$

从上边各式的意义可以看出来， F 值实际上就是在比较，相对于均方误，也就是组内的平均差异 (MS_e) 而言，组间的平均差异 (MS_b) 是否足够大？

当样本组数是两个的时候， F 值近似等于 t 值的平方乘以一个与自由度有关的量。特别的，当使用单因子方差分析对两个独立样本进行比较时，所得结果是和独立样本 t 检验完全一致的。

使用单因子方差分析进行假设检验，和之前的假设检验过程是完全一致的。但是当分组变量的取值多于两个的时候，我们从分析结果只能得出结论：可能存在某两组使得这两组之间的均值差异是统计显著的，但是我们仍然不知道是哪两组之间存在显著的差异。因此我们还需要做**事后检验**，以期进一步确定到底是哪两组或哪几组之间存在着统计显著的差异。

事后检验的方法有很多，有些比较保守（判定组间差异统计显著的标准比较严格），有些则比较宽松。所有事后检验的比较原则都是在控制比较的组数的条件下，对所有组的均值进行两两比较，然后确定其是否显著不同。常用的比较宽松的事后检验方法是 *TukeyHSD* 事后检验，此外还有相对严格的 *Scheffe* 事后检验。以下是进行 *TukeyHSD* 事后检验时需要计算的统计量

$$TukeyHSD = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}}}$$

式中的 $s_{\bar{X}}$ 表示某种标准误

$$s_{\bar{X}} = \sqrt{\frac{MS_e}{n_g}}$$

其中 n_g 表示各组的对象数（各组对象数相等是使用 *TukeyHSD* 事后检验的必要的前提条件）。

在计算出了 *TukeyHSD* 值之后，还是要通过查表来确定差异是否是统计显著的。

9.2 因子方差分析

因子方差分析进一步对单因子方差分析进行了推广。单因子方差分析只能考虑一个因素（即一个分组变量），而因子方差分析则可以同时考虑多个影响因素（多个分组变量），既能够检测每种影响因素的主效应，又能考虑不同因素之间的交互效应。

9.3 复测方差分析

复测方差分析是对相依样本 t 检验的推广。就像单因子分析是对独立样本 t 检验在分组变量上从两个分组推广到多个分组类似，复测方差分析是把相依样本 t 检验在两个时间点的取值推广到了多个时间点。此外，还可以同时进行因子方差分析，将分组变量从一种推广到多种。我们甚至还可以考虑协变量的因素，进行协方差分析。限于时间关系，暂时不做展开。

以下给出复测方差分析的适用条件：

1. 正态性。各组数据服从正态分布。
2. 方差齐性。各组方差相等。
3. 球对称假设。对于自变量的各取值水平组合而言（对于被试内因素的各个水平组合而言），因变量的协方差矩阵相等。

10 回归分析

回归和相关有着紧密的联系，在复习回归之前，我们先复习相关性的概念。

在统计学基本知识章节里，我们复习的统计量和参数都是一次描述一个变量，尽管单个变量的统计很重要，但很多时候我们会对两个及以上变量之间的关系更感兴趣：学生的考试成绩和他的备考时间有关系吗？日平均气温和冷饮的销量有关系吗？等等诸如此类的问题需要引入变量之间的关联程度的统计量：相关系数。

10.1 相关系数

现实世界中，不同的变量之间经常表现出某种相关关系。例如，一般来说，身高较高的人，体重也相对较重，等等。然而这种相关关系很多时候并不能一眼就看出来，这就需要我们使用统计学工具来根据已有的样本数据去考察两个变量之间是否真的有某种相关性。

根据两个变量的类型等不同因素，相关系数也有多种类型。以下我们以皮尔逊积差相关系数为例进行复习。

10.1.1 皮尔逊积差相关系数

皮尔逊积差相关系数考察的是两个定比或定距变量之间的相关性。我们关心两个变量之间的相关性，实际上主要关心的是两个方面的因素：

1. 相关性的方向。

正相关意味着我们分析的两个变量的取值平均而言会同时增大或减小。而负相关则意味着两个变量的取值反方向变化：平均而言，一个变量增大时，另一个变量会减小。

注意上述陈述中的“平均而言”四个字表明，可能存在和总的趋势相悖的例外值。

2. 相关性的强度或量级。


相关性的强度的取值范围为 $[-1, 1]$ ，特别的，当两个变量的相关系数为 0 时，表示两个变量没有直线相关关系，这时候在两个线性代数里可以看作样本数据在两个变量上的取值所组成的两个向量是线性无关的。

一般而言，实际的社会科学研究中，完全正（负）相关都是极其罕见的。一些前人研究的经验法则告诉我们，相关系数绝对值小于 0.2 时，可以视作两个变量弱相关，相关系数绝对值位于 0.2~0.5 时，称两个变量中等程度的相关，当相关系数的绝对值大于 0.5 时，表示强相关。

皮尔逊积差相关系数 r 的计算公式:

$$r = \frac{\Sigma(z_x z_y)}{N}$$

其中 z_x, z_y 分别为变量 X, Y 的 z 分数, N 表示 X, Y 的配对个数, 如果是在考察一个样本中对象的两个变量之间的关系, 则这个数值就是样本所含的对象数。我们注意到上式实际上是两个 z 分数的交叉乘积的平均值, 它相当于把两个变量的协方差进行了标准化。

 相关性仅仅意味着一个变量取值的变动**对应于**另一个变量取值的变动, 除此之外, 没有告诉我们任何其他的事情。绝不能从相关系数的计算得出两个变量之间的因果关系。作为对比, 我们需要明确一下: 因果关系意味着一个变量取值的变动**导致了**另一个变量取值的变动。因果关系通常是借助于领域知识做出的决定, 而不应该是相关性的计算。

关于相关和因果的关系, 可以参阅《别拿相关当因果》这本有趣的书。

此外, 我们还需注意, 相关性仅仅考察了两个变量之间是否具有线性相关的关系, 但很多时候两个变量之间的关系并不仅仅是简单的线性相关, 比如, 有很多变量之间的关系是曲线而不是直线。这个问题留待以后进一步讨论。

10.1.2 线性相关的统计显著性检测

计算出了两个变量之间的相关系数之后, 我们还需要进一步判断, 这个**从样本中发现的相关性是否代表了抽样总体中两个变量之间的关系**? 这个问题恰好就是我们前两章复习的假设检验所能够解决的。

和其他的假设检验一样, 相关性的假设检验也是先提出一个零假设: 两个总体的变量之间完全无关, 然后我们进行单样本的 t 检验, 利用样本数据进行统计分析, 计算出在零假设成立的前提下, 仅仅由于随机误差得到当前样本的概率是多大, 进而根据一定的显著性水平来拒绝或接受零假设。

在这个过程中, 关键是计算出 t 值。我们给出如下的计算公式

$$t = \frac{r - \rho}{s_r}$$

其中 r 即为我们从样本数据计算出的皮尔逊积差相关系数, ρ 为我们的零假设给出的总体的两个变量之间的相关系数 (零假设是不相关, ρ 即为 0), 而 s_r 则代表相关系数的标准误

$$s_r = \sqrt{\frac{1 - r^2}{N - 2}}$$

上式中的 r^2 为相关系数的平方, 我们把它称为**决定系数**, 这一统计量和后边的方差分析中的其他统计量都是使用可释方差百分比来测度两个变量之间的相关强度。

10.1.3 其他几种类型的相关系数

1. 考察一个定比或定距变量与一个分类变量之间的相关性的点二列相关系数；
2. 考察两个二分变量之间的相关性的 ϕ 相关系数；
3. 考察两个定序变量之间的相关性的斯皮尔曼 ρ 相关系数，这是皮尔逊 r 相关系数的一种特殊情形。

10.2 一元线性回归

当我们在考察两个变量之间的相关系数的时候，我们并不明确区分谁是因变量谁是自变量。这么做的不便之处在于，例如，我们想要通过一个变量的改变程度对另一个变量的改变进行预测，光有相关系数就不够用了，这时候就需要使用线性回归来确定自变量和因变量之间的函数关系，以便进行预测。回归和相关的区别是：相关分析不区分自变量和因变量，但回归一定区分自变量和因变量。

线性回归根据所涉及的变量个数而分为一元线性回归（或者称为简单回归）和多元线性回归。一元线性回归涉及一个自变量（或者称为预测变量）和一个因变量（也称为结果变量），我们可以根据预测变量的给定值来对结果变量的取值进行预测。而多元回归则涉及多个自变量和一个因变量，多元回归分析使得我们能够考察多个自变量和因变量之间的关系的性质与强度、若干自变量对因变量的相对预测能力，以及在控制了一个或多个协变量的情况下，一个或多个自变量的独特贡献，此外还能检验交互效应。

我们这里讨论的回归所使用的自变量与因变量都必须是定距或定比变量。对于二分变量作为预测变量的回归，可以参考 Logit 回归（也称为 Logistic regression, “逻辑回归”）。

10.2.1 一元线性回归方程

一元线性回归的方程为

$$\hat{Y} = bX + a$$

其中， \hat{Y} 为变量 Y 的预测值， b 为未标准化的回归系数（或斜率）， a 为截距， X 为自变量。这里重点说明一下 b 和相关系数的关系。我们说“ b 为未标准化的回归系数”，其实就是说它经过标准化后，就得到了皮尔逊相关系数 r 。以下不加证明地给出二者之间的关系：

$$b = r \frac{s_Y}{s_X}$$

上式中， s_Y, s_X 分别为因变量 Y 和自变量 X 的标准差。

我们可以从量纲角度来考察上式。注意到相关系数 r 是一个比率，它是无量纲的，两个变量标准差的量纲分别和变量的量纲一致，标准差之比实现了将量纲从自变量转换为了因变量。因此在一元线性回归方程中， b 实现了量纲从自变量到因变量的转换（当然事实上它还实现了将自变量的变化幅度转化为因变量的变化幅度）。

需要注意的是，上述一元线性回归方程只是预测了因变量的变化，而不是因变量的实际值，如果想要把上式方程中的预测值替换为实际值，需要增加一个误差项 e ，因此我们可以给出如下的另外一个线性回归方程

$$Y = bX + a + e$$

这个方程将误差项考虑在内，使得我们能够把实际的因变量观测值和自变量的值联系起来。

10.2.2 最小二乘法

如果我们把自变量和因变量的值组成的有序偶视为直角坐标系中的点的坐标，则我们可以在坐标系中绘出 **散点图**，然后我们求回归方程的目的就变成了寻求穿过这些数据点的“最好”的直线。这里的“最好”，意味着尽管这条直线不一定穿过了最多的点，但是当我们计算所有点到这条之间的距离之和（注意这个和其实主体部分也是某种形式的平方和）的时候，我们希望这条“最好”的直线是所有直线中使得这个距离之和最小的直线。这就是最小二乘法的思想。

10.3 多元线性回归

多元线性回归使得我们可以讨论的自变量个数不再是一个，但它的基本形式是一致的，下边我们给出多元线性回归方程的具体形式：

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

关于多元线性回归的具体陈述，限于时间关系，这里暂时不做展开，可以参考相关的统计学教材。后续将逐步完善。

参考文献