定量研究方法  
**Quantitative Research Methods**

**授课教师**彭泽宇 青年副研究员（[zpeng@fudan.edu.cn](mailto:zpeng@fudan.edu.cn)）

**助教**蒋屹阳、周督竣、肖伟林、张睿明、林佳怡

**目录**

[第一讲 绪论：定量研究方法在社科研究中的应用 3](#_Toc180163516)

[一、观察性研究 3](#_Toc180163517)

[（一）观察性研究的案例 3](#_Toc180163518)

[（二）观察性研究的局限性 4](#_Toc180163519)

[二、随机对照试验 4](#_Toc180163520)

[（一）随机对照试验的案例 4](#_Toc180163521)

[（二）随机对照试验的局限性 4](#_Toc180163522)

[第二讲 R语言 4](#_Toc180163523)

[第三讲 概率与条件概率 5](#_Toc180163524)

[一、概率的基本概念 5](#_Toc180163525)

[（一）概率的定义 5](#_Toc180163526)

[（二）概率模型的基本概念 5](#_Toc180163527)

[（三）概率公理 5](#_Toc180163528)

[（四）补集与全概率公式 6](#_Toc180163529)

[二、条件概率 6](#_Toc180163530)

[三、贝叶斯法则 6](#_Toc180163531)

[四、独立性 6](#_Toc180163532)

[第四讲 随机变量和概率分布 7](#_Toc180163533)

[一、随机变量 7](#_Toc180163534)

[（一）随机变量的概念 7](#_Toc180163535)

[（二）随机变量的种类 7](#_Toc180163536)

[二、分布 7](#_Toc180163537)

[（一）伯努利分布 7](#_Toc180163538)

[（二）均匀分布 8](#_Toc180163539)

[（三）二项分布 8](#_Toc180163540)

[（四）正态分布 9](#_Toc180163541)

[三、期望和方差 10](#_Toc180163542)

[（一）期望 10](#_Toc180163543)

[（二）方差 11](#_Toc180163544)

[（三）协方差 11](#_Toc180163545)

[第五讲 点估计 12](#_Toc180163546)

[一、估计 12](#_Toc180163547)

[（一）估计及其相关概念 12](#_Toc180163548)

[（二）估计量的性质 12](#_Toc180163549)

[二、适用于任何样本容量的有限样本特征 12](#_Toc180163550)

[（一）无偏性 12](#_Toc180163551)

[（二）相对有效性 13](#_Toc180163552)

[（三）平均数平方误差 13](#_Toc180163553)

[三、适用于大样本容量的特征 13](#_Toc180163554)

[（一）一致性 13](#_Toc180163555)

第一讲 绪论：定量研究方法在社科研究中的应用

2024.9.6

社会科学量化研究通常分为观察性研究（observational studies）和随机对照试验（randomized controlled trials, RCTs）。

* **观察性研究：**研究人员无法在现实世界中随机分配干预措施，只能观察和收集自然发生的事件中的数据；
* **随机对照试验：**研究人员随机分配干预给研究对象（观察值，observations）。

一、观察性研究

（一）观察性研究的案例

1. 案例背景：无差别炮击是否会引起更强烈的反抗

无差别炮击、轰炸是军事冲突双方常见的策略，同时也是国际安全、军事理论研究中常见的话题。由此催生了一个社会科学问题：在内战中政府军无差别炮击是否会引起民众更强烈的反抗？

这个问题的潜在答案有两个：会引起强烈反抗；或，不会引起强烈反抗，并有效遏制镇压了反抗势力。

我们首先运用推理，提出导向上述答案的可能原因：

* **会引起强烈反抗**
  + 加剧对政府军的仇恨
  + 向叛军/抵抗组织寻求保护
* **有效遏制镇压反抗势力**
  + 减少叛军可获得的资源
  + 威慑恐吓当地民众
  + 打击叛军的后勤网络

我们可以运用定量研究方法验证上述假说。当然，对于这种社会科学问题（尤其是与国际经济军事相关的问题），观察性研究的数据采集难度是很大的。

2. 案例研究：自变量、因变量与干扰因素

杰森·力欧（Jason Lyall, 2009）收集了2000年至2005年俄罗斯联邦军队对车臣村庄实施炮击的数据。数据中存在自变量和因变量之分：

* **自变量（解释变量、预测变量）：**相对稳定且能够决定其他变量的数值的变量；
* **因变量（被解释变量、响应变量）：**数值取决于自变量或其他变量的变量。

我们需要以实证方式验证自变量与因变量之间的关系。在该案例中，自变量是该村是否受到过俄军的无差别炮击；因变量是炮击前后俄军和亲政府武装在该村遭受车臣反政府武装袭击的次数。

当然，存在影响自变量的干扰因素，如人口、贫困率、地形高度（海拔）、是否有俄军进驻、控制该村的车臣领导人派系等。干扰因素又称控制变量。

（二）观察性研究的局限性

观察性研究可能存在选择性偏误（selection bias），其内在效度（internal validity）可能会受到影响，难以证实因果关系。例如，在上面的例子中，研究者可能没有观测到一个同时影响了俄军炮击某一村庄的决定以及当地袭击俄军及亲俄武装的频率的干扰因素。

因此，我们可以在随机对照试验当中研究者可以将干预变量进行随机分配，从而解决选择性偏误问题。

二、随机对照试验

（一）随机对照试验的案例

1. 案例背景：劳动力市场是否存在种族歧视

种族歧视是以美国为代表的西方国家中普遍存在的社会问题。西方国家的劳动力市场是否真的存在种族歧视是社会科学界常年争论的学术话题——例如，不同种族之间的失业率差异是否是由于其他因素（比如教育程度差异）造成的？

在实验中，研究者很难操纵或改变一个人的种族。因此，一个重要的问题就是：如何设计一个随机试验来检视雇主是否会区别对待不同种族的求职者？

2. 案例研究：随机对照试验的设计

经济学家玛丽安娜·伯特兰德（Mariannne Bertrand）和赛德希尔·穆来纳森（Sendhil Muillainathan）进行了以下试验来考察美国劳动力市场的种族歧视问题：研究人员向在报纸上登广告的潜在雇主发去虚构的求职简历；只改变求职者的名字，而将其他信息保持不变。一些求职者用了非常典型的非洲裔美国人的名字（如Lakisha），而其他求职者用了典型的白人名字（如Emily）；比较这两个群体之间的电话回复率。

（二）随机对照试验的局限性

随机对照试验的内在效度普遍高于观察性研究，通常被认为是建立许多科学学科的因果关系的黄金标准。然而，出于道德和操作的种种因素，在很多情况下研究者无法在现实世界随机分配干预措施。例如，在上文提及的“无差别炮击”案例中，社会科学家就不可能将每次炮击的目标随机分配；在本案例中，寄出大量虚构的求职简历也可能会扰乱用人单位人事部门的正常运转。

随机对照试验可能缺乏“外在效度（external validity）”。如在本案例中，研究对象是否包含一个国家/社会中所有所有族群？在美国，简历上一般不贴照片；那么这个实验是否符合其他国家的社会文化背景？

第二讲 R语言

2024.9.13 / 2024.9.20 / 2024.9.27

见“R语言语法速查”。

第三讲 概率与条件概率

2024.9.27 / 2024.9.29

一、概率的基本概念

概率论是解决社会科学研究中不确定性的重要工具。社会科学研究一般由两步构成：

* 构建理论：在所学的基础上建立一个概率模型；
* 运用数据和其他经验性证据证明概率模型从而支持你的理论。

（一）概率的定义

**概率**是一系列在现实世界中用来测量随机性并对随机性建模的数学工具。

频率统计学派（frequentist）认为，概率是对应频率的极限，是事件发生次数与同样条件下重复试验次数之比，即当在相同条件下反复进行的实验次数接近无穷次是某事件发生频率的极限。

贝叶斯统计学派（Bayesian）认为，概率是用来测量一个人认为某件事发生可能的主观信念。

（二）概率模型的基本概念

**试验（experiment）**是一个或一组产生与某一问题相关的随机事件的行动。

**样本空间（sample space）**是试验所有可能的一组结果，通常用Ω表示。

**事件（event）**是样本空间的一个子集。

（三）概率公理

第一，任一事件*A*的概率非负，即：

第二，样本空间中所有结果中任一发生的概率是1，即：

第三，概率遵循加法法则，即：

对于任意一给定的事件*A*和*B*：

如果事件*A*和*B*是互斥的：

（四）补集与全概率公式

**补集**是所有不属于给定的事件的结果的集合，通常用上标*c*表示。因为*A*和*Ac*是互斥的，且它们一起构成了整个样本空间，故根据加法法则，对于任一给定的事件*A*，有：

同时，我们可以使用韦恩图得出全概率公式：

二、条件概率

事件*A*的**条件概率**指的是事件*A*随着事件*B*的发生而发生，用*P*(*A*|*B*)表示：

在这一公式中，*P*(*A和B*)是两个事件都发生的概率即**联合概率（joint probability）**，其中，*P*(*B*)是事件*B*的**边际概率**。整理公式，我们可以得到乘法法则（multiplication rule）：

综合之前提到的全概率公式，可得出全概率公式的另一种表达方式：

三、贝叶斯法则

贝叶斯法则如下：

其中，*P*(*A*)被称为先验概率（prior probability），反映了一个人关于事件*A*发生可能性的初始信念。在观测到事件*B*代表的数据后，我们更新了自己的信念并得到*P*(*A*|*B*)，*P*(*A*|*B*)被称为后验概率（posterior probability）。

贝叶斯法则的公式是基于条件概率公式推导出的，故可看作是条件概率公式的一种转写。

四、独立性

在直觉上，两个事件的**独立性（independence）**指的是一个事件的信息不会给我们更多关于另一个事件是否发生的信息。根据乘法法则，当且仅当事件*A*和*B*的联合概率等于边际概率的乘积，那么*A*和*B*互相独立的，即：

给定事件*C*，*A*和*B*的**条件独立**意味着给定*C*时*A*和*B*的联合概率等于两个条件概率的乘积：

同时，这也意味着：

第四讲 随机变量和概率分布

2024.9.29 / 2024.10.11 / 2024.10.18

一、随机变量

（一）随机变量的概念

随机变量*X*指的是将样本空间中的每个事件赋予一个实数的函数（即赋予每一个事件一个数）。

随机变量的数值必须表示互斥（mutually exclusive），不同数值不能表示同一事件（如事件*A*不能同时用1和2来表示）。

随机变量的数值必须是完备的（exhaustive），全部事件应该有一些数值来表示。

（二）随机变量的种类

* **离散随机变量：**数值为有限个，如家庭成员个数、态度（支持/反对）等。其概率分布函数包括概率质量函数（probability mass function, PMF）、累积分布函数（cumulative distribution function, CDF）。
* **连续随机变量：**在某个实线区间内有无限个取值，如长度、GDP等。其概率分布函数包括概率密度函数（probability density function, PDF）、累积分布函数。

二、分布

（一）伯努利分布

1. 伯努利随机变量的定义

我们先定义一个二元随机变量（binary random variable），该二元随机变量的数值为两个不同的数。我们将之称为**伯努利随机变量（Benoulli random variable）**。我们一般认为，事件*X*=1表示成功，*X*=0表示失败；我们用*p*表示成功的概率。

伯努利随机变量可以用**概率质量函数（PMF）**表示。一个随机变量*X*的PMF的定义是：在随机变量取某一值*x*时，*f*(*x*)=*P*(*X*=*x*)的概率。以伯努利随机变量为例，PMF在*x*=1（成功）时取*p*值，在*x*=0（失败）时取值为1−*p*，除此之外函数*f*(*x*)的值对应其他任何*x*值都是0。

离散随机变量的**累积分布函数（CDF）***F*(*x*)表示随机变量*X*取值小于或等于某一特定值*x*时的累积概率，即*F*(*x*)=*P*(*X*≤*x*)。因此CDF表示PMF *f*(*x*)对于所有*x*值的和。任何离散随机变量PMF *f*(*x*)和CDF *F*(*x*)之间的关系可以写为：

其中，*k*表示所有小于或等于*x*的随机变量*X*的值。

2. 伯努利随机变量的概率分布函数

一个成功概率为*p*的伯努利随机变量的概率质量函数（PMF）为：

其中*f*(1)和*f*(0)分别代表成功和失败的概率。

一个成功概率为*p*的伯努利随机变量的累积分布函数（CDF）为：

若*x*≥1，函数值为1-*p*+*p*=1。

（二）均匀分布

1. 均匀随机变量的定义

**均匀随机变量（uniform random variable）**是连续随机变量的一种简单类型。均匀随机变量在给定区间[*a*, *b*]内取一值的可能性相同。

均匀随机变量可用概率密度函数（PDF）或者累积分布函数（CDF）来表示。

概率密度函数的数值非负且可以大于1，但是PDF下方的区域面积必须等于1。需要注意的是，与概率质量函数（PMF）不同，PMF的数值不能大于1，而PDF的取值可以大于1。**均匀分布的PDF**是一条由定义的水平线，因为PDF下方区域的面积等于1。

**均匀分布的CDF** *F*(*x*)表示随机变量*X*取值小于或等于给定值*x*的概率，即*P*(*X*≤*x*)。如果*a*≤*x*< *b*，PDF下方与*x*对应的面积就是CDF的值。数理上，我们可以用积分来表示这个概念：

2. 均匀随机变量的概率分布函数

在区间[*a*, *b*]的均匀随机变量的概率密度函数为：

在区间[*a*, *b*]的均匀随机变量的累积分布函数为：

（三）二项分布

1. 二项随机变量的定义

**二项分布（binomial distribution）**是对伯努利分布的一般化。二项随机变量*X*记录了*n*次独立且相同的、成功概率为*p*的试验中的成功次数。

一个**二项随机变量**是*n*个独立并且相同分布（independently and identically distributed, 或者简称为i.i.d.）的伯努利随机变量之和。在二项分布中，*X*可以取0到*n*的整数。

2. 二项随机变量的概率分布函数

试验次数*n*，成功率为*p*的二项随机变量的**概率质量函数**公式为：

其中，

即*n*选*x*。

对于*x*=0, 1, ..., *n*, **累积分布函数**可以写为：

（四）正态分布

1. 正态随机变量的定义

一个**正态随机变量（normal random variable）**可以取实线(−∞, ∞)上任意一点。一个正态分布有两个重要参数：平均数*µ*，标准差*σ*。一个正态随机变量*X*可以写作*X*∼*N*(*µ*, *σ*2)，其中*σ*2表示方差。

2. 正态随机变量的分布函数

一个正态随机变量的**概率密度函数（PDF）**为：

对于实线上任意一个*x*，正态随机变量的**累积分布函数（CDF）**为：

CDF代表PDF下方从-∞到*x*所形成的面积。

正态分布PDF形状为钟型，中心为平均数，分布范围由标准差大小决定。正态分布PDF下的面积为1。不同的平均数只移动PDF和CDF，但不改变形状。标准差增加可造成PDF更平坦，CDF增长趋势更平缓。

3. z分数

假设随机变量*X*∼*N*(*µ*, *σ*2)。令*c*为任一常数，下列性质成立：

* 定义为*Z*=*X*+*c*的随机变量服从*Z*∼*N*(*µ*+*c*, *σ*2)的正态分布；
* 定义为*Z*=*cX*的随机变量服从*Z*∼*N*(*cµ*, (*cσ*)2)的正态分布。

根据以上性质，我们可以推出：

我们称为正态随机变量的***z*分数**，即：

*z*分数也可以理解为随机变量*X*的某一个数值与平均值*µ*之间有多少个标准差*σ*的距离。如果数据按照正态分布，则约2/3的数值在平均值的一个标准差之内，约95%的数值在平均值的两个标准差之内。

计算正态随机变量位于平均值*k* (*k* > 0)个标准差之间的概率可用：

其中，*P*(-*k* ≤ *Z* ≤ *k*)也可以被转写成CDF在-*k*和*k*取值时的差，即：

三、期望和方差

（一）期望

1. 期望的基本概念

**期望（expectation）**代表一个概率分布下的平均值，也称为总体平均值（population mean）。随机变量的期望在给定的概率模型中是固定的。我们通常用*E*(*X*)来表示随机变量*X*的期望。

用*E*(*X*)表示的随机变量的期望值定义为：

其中，*f*(*x*)是离散变量*X*的概率质量函数（PMF）或连续变量*X*的概率密度函数（PDF）。

2. 期望的性质

令*X*和*Y*为随机变量，*a*和*b*是任意常数。期望具有如下线性性质：

如果*X*和*Y*独立，则有：

应注意，*E*(*XY*)=*E*(*X*)*E*(*Y*)是*X*和*Y*独立的必要非充分条件。

（二）方差

1. 方差的基本概念

随机变量的标准差（standard deviation）的平方即**方差（variance）**。我们通常用*V*(*X*)来表示随机变量*X*的方差。

随机变量*X*的方差被定义为：

上式可被转写为：

2. 方差的性质

令*X*和*Y*为随机变量，*a*和*b*是任意常数。方差具有如下性质：

如果*X*和*Y*独立，则有：

（三）协方差

协方差用于度量两个随机变量*X*和*Y*的相关变化。我们通常用Cov(*X*, *Y*)来表示随机变量*X*和*Y*的协方差。

随机变量*X*和*Y*的协方差被定义为：

若*X*=*Y*，则Cov(*X*, *Y*)=*V*(*X*)。

如果*X*和*Y*独立，则有：

应注意，Cov(*X*, *Y*)=0是*X*和*Y*独立的必要非充分条件。

第五讲 点估计

2024.10.18

一、估计

（一）估计及其相关概念

**统计推断**是利用来自某总体的一个样本从而对总体有所了解。统计推断的方法大致可归类为**估计（estimation）**和假设检验（hypothesis testing）；估计又分为点估计（point estimate） 和区间估计（interval estimate）。

与估计相关的概念有估计目标、估计量与估计值。

**估计目标（estimand）**是我们需要估计的总体的参数（parameter）。通常用希腊字母表示（如*μ*、*θ*）。

**估计量（estimator）**是我们估算估计目标的方法。通常用希腊字母加hat表示（如）。

**估计值（estimate）**是我们运用估计量从所得样本中进行估算所得的具体数值。

（二）估计量的性质

估计量是随机变量，其随机性来源于重复抽样。由重复随机抽样而得来的估计量的分布常被称为抽样分布（sampling distribution）。估计量的性质指的是其抽样分布的特征。

适用于任何样本容量的有限样本特征有：

* **无偏性（unbiasedness）：**估计量的抽样分布是否围绕在估计目标参数附近（）；
* **相对有效性（efficiency）：**抽样分布的方差是否较小（）。

适用于大样本容量的特征有：

* **一致性（consistency）：**当我们的样本容量（n）足够大的时候，估计量的抽样分布是否趋近于估计目标；
* **渐进正态性（asymptotic normality）：**当我们的样本容量（n）足够大的时候，估计量的抽样样本是否呈正态分布。

二、适用于任何样本容量的有限样本特征

（一）无偏性

**偏误（bias）**指的是估计量的期望与估计目标的期望值的差。

我们称一个估计量的**无偏的**，如果

如果我们能够从总体中抽取关于*X*的无限多个样本，并且每次都计算一个估计值，那么将所有随机样本的这些估计值平均起来，我们便得到*θ*。

（二）相对有效性

相对有效性是比较不同无偏估计量的方法之一。在多个无偏误的估计量当中，我们通常选取方差最小、相对有效性最高的那一个。

**相对有效性**就是：比较和，如果的方差小于的方差，那么更有效。

（三）平均数平方误差

一般而言，我们倾向选择无偏并且相对有效的估计量。如果估计量都存在偏误，我们该如何选取合适的估计方案？此时就需要引入平均数平方误差，即均方误（mean-squared error, MSE）：

可见，均方误实际上是估计量的方差与偏误平方之和。

三、适用于大样本容量的特征

（一）一致性

如果一个估计量的抽样分布随着样本容量*n*的增加越来越集中于估计目标*θ*，那么就是**一致的**。

即

一致性体现在估计量的期望值越来越接近估计目标以及估计量的方差逐渐趋近于0。一致的估计量不一定无偏，无偏误的估计量不一定具有一致性。