Statistical Inference Note

Statistical Inference Note

Chapter 1 概率论

1.1. 概率论基础

1.1.1. 公理化基础

- 样本空间的一个结果 => 一个事件（结果的集合） => Borel域（事件的集合）
  通过Borel域用公理定义概率，绕开了频率学派和贝叶斯学派关于概率本质的争论，满足三个特性：
    - 在Borel域中任意事件的概率大于零
    - 样本空间总体构成的事件概率为1
    - 对于属于Borel域的不相交的多个事件，事件相与的概率等于单个事件概率求和
- 上面定义比较麻烦，具体判别一个函数是否为概率的方法：样本空间的每个结果都有一个非负值，且和为1，事件A包含包含的结果的非负值累加即为事件A的概率
- 此处定义的概率就是在基于样本空间，满足特定条件的映射函数值
- 用公理定义概率的缺点（暂时）是做不了题，因为它对于系统的知识只有样本空间，比如说掷硬币，样本空间为正面和反面，按照公理，概率可以是0.3/0.7，也可以是0.5/0.5，等等。这样一个样本空间的概率分布情况是无穷的，我们要通过对于系统更加深入的了解来确定具体概率分布是哪种情况。书中的飞镖例子本质是扮演上帝：在完全了解系统的情况下去算概率，当然满足上述概率定义，同时又符合真实情况
- 问题：在定义概率的时候，引入Borel域的必要性是什么？

1.1.2. 计数

1.1.3. 枚举结果

- 1.2.3和1.2.4就是在1.2.1公理化定义概率之后针对一种特定系统（样本空间的每一个事件可以通过计数得到它包含多少的结果，同时包含结果数除以总结果数就是概率），得到符合现实规律的概率值

1.2. 条件概率与独立性

1.2.1. 三个囚犯

- 只要注意一点就行：“看守说B将被处死” 和 “B将被处死” 是两个条件

- B将被处死：这是一个客观条件，在引入这样一个条件之后，样本空间坍缩成A或C被处死（空间变为原来的2/3），概率各为1/2
- 看守说B将被处死：有两种情况，第一种是A被赦免的情况下看守说B被处死，第二种是C被赦免的情况下看守说B被处死，最终计算得到看守说B被处死的概率为1/2（空间坍缩成原来的1/2），然后去算在坍缩的空间里，A被赦免的概率，还是1/3

1.3. 随机变量

大写 X 是变量，小写 x 是固定的取值

Chapter 2 常见分布族

2.1. 引言

分布族就是带参分布，固定一组参数就是一个确定分布，一堆确定分布就是一族

2.2. 离散分布和连续分布

2.3. 指数族

一个分布族可以写成一个特定形式，这个族就可以称为指数族

比一般分布族优秀的点：在求期望和方差的时候，能将积分转化成求导

在说明指数族性质的过程中，只有一个限制：随机变量的取值集不能随着分布函数的参数变化

比如二项分布，p取0到1之间的任何一个值，x的值集都是[0, n]，而一旦p=0, 1，x就只能等于0, n，所以在书中说明二项分布是指数族的时候，剔除了p=0, 1的情况

而在具体的函数表示为如下所示：

$$f(x|\theta) = h(x)\ c(\theta)\ exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

这个函数最重要的含义：在 $f(x|\theta)$ 层面，分布函数只能为干干净净的右边这种形式，而不能是有右边这种形式，最后还带一个对某个变量或参数值的限制。这是因为$h(x)$的作用就是包含 x 的全部信息，而 $c(\theta)$ 就是包含 θ 的全部信息，而自变量和参数之间唯一允许的纠缠只能是在指数函数内相乘的形式。这种结构表明如果对于函数的整理最后一定还要加一个“拖油瓶”，那么就代表着 x 和 θ 存在允许范围之外的纠缠，那么这一族分布就肯定不是指数族

而不在 $f(x|\theta)$ 层面，而在 $h(x)$ ，$c(\theta)$ 的里面，函数形式可以随便变化

104页的示性函数的作用是将只能用区间表示的限制条件表达成函数，然后与 $h(x)$ ，$c(\theta)$ 合并，我觉得有点多余。可以直接说明不仅函数形式必须分离，区间依赖也必须分离（函数=函数具体形式加上自变量区间）

自然参数空间：

参数变化生成一族分布，如果不人为加上限制，对于参数的限制的来源只有两个：x对于参数的作用（x的形式），概率积分要为1。满足这两点的参数构成自然参数空间。

曲指数族：

独立的参数个数小于函数表述中exp内的项数 k

而等于则称为完全指数族，通常情况是等于

2.4. 位置与尺度族

One particular distribution can belong to more than one family. The basic thread of this section is to use mean and variance to construct a family based on one particular distribution. The reason to construct a family this way is mean and variance have physical interpretation which putting a random variable on a random spot cannot achieve.

定理**3.5.7**：使用的时候，一个随机变量 X ，分布具有均值 μ ，方差 σ^2 ，则 $Z=\frac{X-\mu}{\sigma}$ 的分布就满足均值为0，方差为1

2.5. 3.6 Inequalities and Identities

Chebychev Inequality does not confine itself in any particular distribution, it is universal across all kinds of distributions, hence when we discuss a specific problem, the boundry from Chebychev can be rather loose, it can be considered as a baseline.

In identities part, the author introduced two types of relations. One is the relation between two smilar distributions, and the other is the relation of moments of different orders.

Chapter 3 多维随机变量

3.1. Joint and Marginal Distributions

3.2. Conditional Distributions and Independence

Example 4.2.13 is the perfect example to demonstrate that the goal of moment generating function(mgf) is not to calculate moments, but to find the distribution of an variable.

这一小节的关键点在定义域的缺失问题

首先我们明确一点，对于连续分布，讨论某一个点的概率密度值是没有实际意义的，因为我们通过概率密度函数想要刻画的是某件事情发生的概率，这是我们最关心的，所以我们可以认为概率密度只是为了我们能够得到连续变量概率而使用的一个工具。换句话说，只要最终得到的概率是正确的，我就认为使用的概率密度函数就是正确的。在通常情况下，一个问题的概率密度函数只有一个，但是由于积分的特性，导致我可以写出无数个正确的概率密度函数。，比如，分布函数是正态分布：$f(x)=n(\mu,\sigma),\ x\in(-\infty,+\infty)$，那么 $f(x)=n(\mu,\sigma),\ x\in(-\infty,1)\cup(1,\infty)$ 也是对的（x=1处的值可以是0，也就是概率在此处无定义，也可以是任意的正有限值，后面出于说明方便的原因，将两种情况统称为问题），因为在任何区间积分，前者都等于后者，我们甚至可以说两个分布函数是一样的。

在判定独立的时候也是如此，我们最关心的是在定义域叉积内 $p(x,y)=p(x)*p(y)$，只要它成立，我就认为两个分布独立。

- $p(x,y)=p(x)*p(y)$ 成立的特殊情况：

  我们考虑连续分布，x,y 的分布函数为 $f_X(x)$，$f_Y(y)$，它们的定义域为 $A,\ B$，那么分布

  $$f(x,y)=f_X(x)*f_Y(y),\ (x,y)\in A\times B$$

  中的两个变量肯定是独立的，这是最理想的情况。

  接下来我们改变定义域，在 $(x,y)\in A\times B$ 中零散挖点，只遵循一个原则，所有的点最多只能连成线，而不能连成面，那么这些点对于其他的位置没有任何影响。因为概率密度函数满足的最基本要求是积分为1，在二元积分中也就是体积为1，而这些零散点/线的构成的体积为0。

  那么挖点产生的新函数 $f^*(x,y)$ 和 $f(x,y)$ 从函数角度来说，就是两个完全不同的函数了。但是用这个函数求某个事件发生的概率时，得到的结果和原来的一模一样；同时通过 $f^*(x,y)$ 推导边缘概率密度会得到初始的概率密度函数（并且定义域也相同），因此边缘概率也相同，最终结果就是：

  $$p^*(x,y)=p(x)*p(y)$$
  $$but\ sometimes$$
  $$f^*(x,y)\neq f_X(x)*f_Y(y)$$

  这个特殊情况会在，判断变量是否独立，用联合推边缘时出现，我们注意只要密度函数的不等情况不构成面，只构成点或线，概率就会相等，那么两个变量就独立。

  这种现象出现的本质原因：

对于离散分布，我们研究的所谓概率质量函数就是概率，这就和判定 $p(x,y)=p(x)*p(y)$ 直接对接，变量独立和等式成立是等价的。

但是对于连续分布，我们借助的却是概率密度函数（也就是概率的一阶微分），严格来说，我们应该将密度函数积分，然后去比较。可是比较麻烦，所以我们就只比较 $f(x,y)$ 和 $f_X(x)f_Y(y)$，由于两者的等式成立条件并不是完全一样，概率等式相等可能出现密度等式相等，或是大部分相等但散点不等。所以我们判别变量独立的时候，密度等式满足其中一种就可认为变量独立。

这种特殊情况在144页中间说明。

- $p(x,y)=p(x)*p(y)$ 不成立的基础情况：

  不独立，代表变量之间相互影响，函数式子中的相互影响比较难看出来，但定义域中的相互影响却很直观。比如 $0<x<y<+\infty$，x 和 y 的取值相互影响，不用看联合分布也知道两者不独立。

总结判定两变量是否独立的流程：

1. 看定义域，定义域相互缠绕，不满足 $A\times B$，则必定不独立
2. 推导边缘概率质量（密度），或是将联合直接拆分两自变量函数相乘的形式，两种方式得到的函数地位上是等价的。离散必须严格相等，连续全局相等或散点不等。

   直接拆得到的两个函数与边缘概率质量（密度）只会相差一个常数

3.3. Bivariate Transformations

在二元积分换元中会涉及Jacobian矩阵，它的意义就是将一个坐标系中的增量 $dxdy$ 转换成另一个坐标系中的增量 $dudv$，而需要将矩阵行列式取绝对值是因为坐标关系函数可能不是正相关的，而增量一定是正的。

Jocobian矩阵和单变量分布变换时的反函数导数是一个东西，本质就是换元。并且两者的限制类似，即能够用某种形式直接表示出目标变量分布的先决条件是输入输出维数一样，并且函数是一对一的，直观表现就是Jacobian矩阵是方阵。

广泛来说，如果不是方阵，又或者函数不是一对一，那就要从基本定义入手求分布。

3.4. Hierarchical Models and Mixture Distributions

3.5. Covariance and Correlation

两个随机变量的关系有：

- 独立：协方差和相关系数都为0
- 不独立
  - 线性相关：协方差和相关系数不为0

- 非线性相关：协方差和相关系数都为0

Chapter 4 Properties of a Random Sample

4.1. Basic Conceptes of Random Samples

重要的思想：在数学处理上，随机样本只不过是独立同分布随机变量的集合，随机样本从本质上还是分布。从符号上来说，这一章对于样本的讨论用的全是大写字母，可以说讨论的仍是由独立同分布叠加而成的统计量（均值，方差）分布，和具体某次实验没有关系。

这也随机两字的含义，就是不是某次具体实验的具体数据

4.2. Sums of Random Variables from a Random Sample

1. 介绍了样本均值和方差的定义，然后是他们的期望与总体分布特征之间的关系
2. 然后讨论如何求样本均值的分布：
    - 矩母函数：有局限性，当总体分布没有矩母函数或是未知矩母函数的时候不好办
    - 变量替换：也就是4.3节那个东西，这里用柯西分布举例
    - 位置尺度族：在一个特定的样本均值分布已经知道的情况下，可以用位置尺度族内关系造出另外的样本均值分布，也是用柯西分布举例

      这里需要说明一点，柯西分布不存在方差，所以在描述柯西分布的波动情况时无法用variance一词，只能用dispersion（离散度）。在一般情况下，样本均值方差和总体方差的关系是 $Var\overline{X} = \frac{\sigma^2}{n}$ ，即样本越多，均值方差越小，这个关系的限制条件是 $\sigma^2 < +\infty$ ；但是柯西分布不满足这种情况（应该是无穷大），在此分布中，样本平均分布离散度和总体分布的离散度是一样的。
    - 指数分布族：

4.3. Sampling from The Normal Distribution

这一节的处理方法：在定理的证明中，总体正态分布是标准正态分布。这是因为简化计算。同时以其他正态分布为总体分布的采样，任何统计量都和标准情况统计量存在函数关系，也就是属于同一族，比如 $\overline{X} = \sigma * \overline{Z} + \mu$ ， $S_X^2 = \sigma^2 S_Z^2$ ，并且标准情况所有的性质其他情况也都有。那么只要标准情况讨论清楚了，其他情况都能直接写答案。

1. 给出总体为正态分布的样本均值分布和方差的具体分布以及相关性质，相当于介绍5.2的一种特殊情况。
    - 均值和方差是独立随机变量，是通过拆联合概率分布函数证明的（判断独立的标准做法）

另一种做法是通过特殊结论：对于独立的正态随机变量的线性函数构成的随机变量，协方差等于0等价于独立。

- 均值分布通过矩母函数得到
- 方差分布形式复杂，通过归纳法证明

2. 给出总体为正态分布的学生t分布，需要这个分布是因为我们需要推断 μ ，但是一般正态分布中会同时存在 μ 和 σ 两个不确定量，那么就很理所当然将 σ 换成它的无偏估计量 S ，最终 $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ 满足t分布，在均值和方差都已知的情况下就能去估计 μ 。
3. 给出两个总体为正态分布的F分布，它产生的原因是因为我们想要研究两总体方差的比值

4.4. Order Statistics

我最开始疑惑的次序统计量的点在于，Casella这本书说样本是分布，不和确定的某次实验挂钩，那么为什么能将样本进行排序呢？

首先要明确一点，统计量都是样本的函数，通过样本分布就能导出统计量分布，前面的样本均值和样本方差都是这样。但并不是只有显式的函数表达式才能是函数，只要有输入有输出就能称之为函数，这里就是后者。

次序统计量的机理如下：

以样本中位数为例，现有总体分布为某种分布的11个样本，他们的序号依次为1-11，在一次特定的实验中，他们的值为：0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1.0，1.1，那么样本中位数为0.6，这个值来自于第六个样本。我再做一次实验，他们的值依次为：0.6，1.0，0.9，0.8，0.7，1.1，0.5，0.4，0.3，0.2，0.1，样本中位数为0.6，这个值来自于第一个样本。经过无数次实验，样本中位数，也就是排序之后位于第6个位置上的数（也就是 $X_{(6)}$ ）会有一个分布，它即为样本中位数分布。

这只是为直观理解而举例，在导出分布的时候遵循的思路是：在依照样本概率取值为a时，有多大的概率会有五个比a大，五个比a小的样本出现。可以看出，在次序统计量中，次序就是作用在样本上的函数，因此 $X_{(i)}$ 和 $X_{(j)}$ 不再是独立同分布。

我感觉不自然的点应该也是函数表现形式发生了变化，在前面我有明确的统计量表达式，对于统计量分布的推导，是将表达式带回样本联合分布，而这里没有统计量表达式，统计量和样本之间的关系是用一段话进行描述的，那么这里求统计量分布则只能用总体分布加上排列组合进行求解。

4.5. Convergence Concepts

在之前，我们如果看到两个大写字母，他们描述的是两个样本空间（尽管可能相互依赖），甚至在讨论随机样本的时候，两个随机样本的独立同分布也是被理解成一模一样但却是两个独立的样本空间。比如一个样本空间事件 $S_i = s_1$ 发生，也就是观测到随机变量 $X_i = x_1$ ，但是这并不会直接让属于另一个样本空间的 X_j 取值确定。

5.5节中引入的新概念是：我们现在讨论的随机变量的渐进性质，渐进性质是样本空间从1到n独立且不变，而样本空间到随机变量之间的映射形式会随着样本序号增大而变化，而收敛则是当样本序号无穷大时，映射形式（函数）是一个与序号无关的稳定结构

讨论收敛的目的：并不是随机样本的每一个 X_n 真的会随着 n 而变化，我们做实验的时候必定是要每一次 X_n 都要是同分布的。这里讨论收敛是想要研究 $\overline{X_n}$ ，它是随着 n 会变化的。

- 引出大数定理，发现在 n 趋近于无穷大时，$\overline{X}_n$ 是以概率1取一个和总体分布有关的常数；换句话说，只要样本足够大，我每次算 $\overline{X}_n$ 就是算这个常数；通过实验对系统进行推断的做法有了理论支持
- 引出中心极限定理

收敛的种类：Convergence in Probability，Almost Sure Convergence，Convergence in Distribution

Convergence in Probability：一个概率为 p_1 的特定的事件 s_1 ，计算用渐进映射形式得到的随机变量值 x_n 以及用稳定映射结构得到的随机变量值 x ，如果两者相等（这里用 ϵ 语言）那么 $p1$ 就累加到满足要求的概率上，最后对 n 取极限。如果概率为1，也就是当 n 趋近于无穷大时，满足两种映射值相同的样本空间中的点的范围趋近于全体样本空间，就称随机样本满足依概率收敛。

$$\lim_{n\to+\infty} P(|X_n(s) - X(s)| < \epsilon) = 1$$

从这里可以导出弱大数定律，将 X 改成常数 μ ，意思就是在 n 趋近无穷大时，随机变量 $\overline{X_n}$ 分布为在 μ 处概率为1，其他地方都为零（样本空间任何事件发生，利用极限情况下的渐进映射形式，得到的随机变量值都为1）

Almost Sure Convergence：“几乎处处收敛”这个名字比较直观，一个概率为 $p1$ 的特定事件 s_1 ，计算用渐进形式得到的随机变量值 x_n 以及用稳定映射结构得到的随机变量值 x ，此时对 n 取极限，如果极限情况下两者相等（同样用 ϵ 语言）那么 $p1$ 就累加到满足要求的概率上。

$$P(\lim_{n\to+\infty} |X_n(s) - X(s| < \epsilon) = 1$$

$$or$$

$$P(\lim_{n\to+\infty} X_n(s) = X(s)) = 1$$

从这里可以导出强大数定律，将 X 改成常数 μ ，意思就是在 n 趋近无穷大时，随机变量 $\overline{X_n}$ 分布为在 μ 处概率为1，其他地方都为零（样本空间任何事件发生，利用极限情况下的渐进映射形式，得到的随机变量值都为1）。由于依概率收敛弱，几乎处处收敛强，对应大数定律也就有强弱之分。

Convergence in Distribution：它与上面两种收敛的区别在于，$F_{X_n}(x)$ 和 $F_X(x)$ 并不需要来自同一样本空间，这种收敛的考量只在 X 的层面，只要两个函数关系式一样，等式就成立。所以它也会比上述两种收敛要弱，因为完全存在累积分布函数一样，但是底层的样本空间不一样的情况。

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

在依分布收敛这里，作者提出了中心极限定理，为什么？$\sqrt{n}(\overline{X}_n - \mu)/\sigma$ 的累积分布函数 $G_n(x)$ 满足标准正态分布：

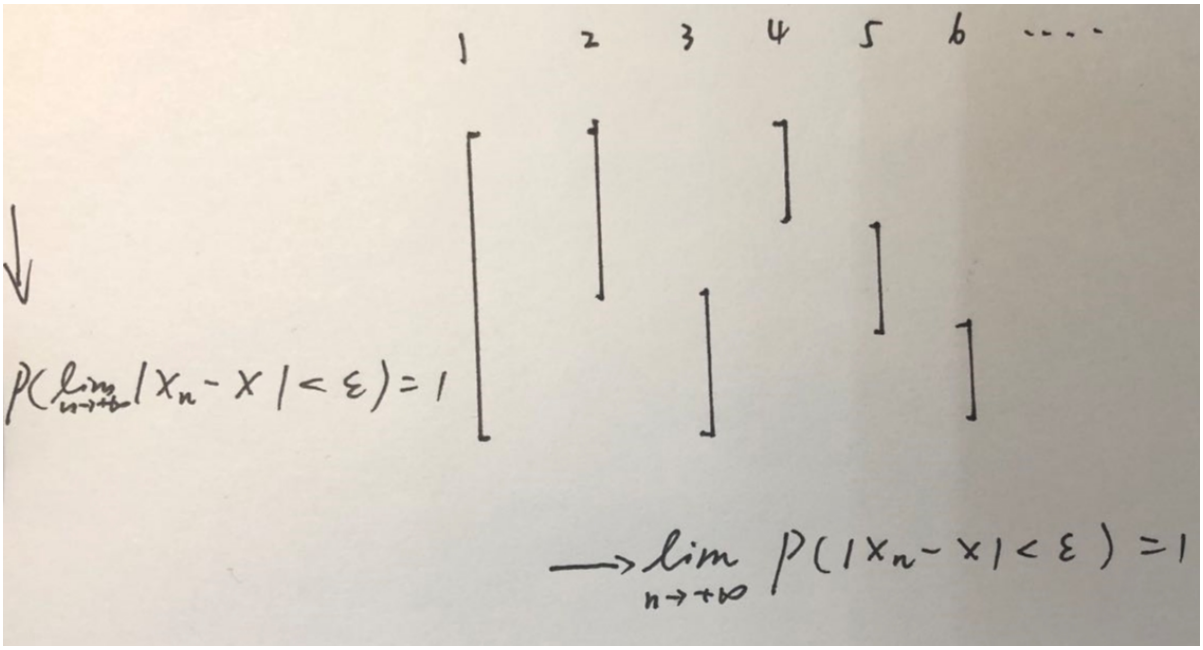
$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

三种分布的关系思考：

依分布收敛是累积分布函数的收敛，依概率收敛和殆必收敛都必定满足依分布收敛

收敛强度：几乎处处收敛 > 依概率收敛 > 依分布收敛

依概率收敛和几乎处处收敛的不同：为了方便想象，将样本空间视作均匀分布，那么样本空间的体积就相当于概率。Convergence in Probability的极限符号是放在 P 前面，那么它关注的是不满足要求的事件构成得体积，只要这个体积数值随着 n 一直在减小并且极限为0，就满足Convergence in Probability，而不关心这个体积所在位置关于 n 的变化情况。Almost Sure Convergence的极限符号放在 X_n 前面，它关注的是一个特定事件 s 的取值 x_n 的收敛情况，如果 $x_1, x_2, \ldots, x_n$ 不是收敛，它就不能被算在符合要求的概率中。从下图来理解（例5.5.8），依概率收敛是横向扫描，殆必收敛是竖向扫描，因此例5.5.8满足前者，不满足后者，也可以看出几乎处处收敛要比依概率收敛要严格。



三种收敛导出三种定理：

- 依概率收敛=>弱大数定理

$\overline{X}_n$ 满足：

$$\lim_{n\to+\infty} P(|\overline{X}_n-\mu|<\epsilon)=1$$

- 几乎处处收敛=>强大数定理

  $\overline{X}_n$ 满足：

$$P(\lim_{n\to+\infty} |\overline{X}_n-\mu|<\epsilon)=1$$

- 依分布收敛=>中心极限定理

  $\sqrt{n}(\overline{X}_n-\mu)/\sigma$ 的累积分布函数 $G_n(x)$ 满足：

$$\lim_{n\to\infty} G_n(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$$

观察到，大数定理是样本均值分布会收敛到一个常数（相当于在 μ 处概率为1，其他地方概率为0的分布），而只要将形式修改成带一个 $\sqrt{n}$，就会收敛到正态分布。

Δ 方法：为中心极限定理的推广，意义在于我们很多时候不仅关心 $\overline{X}_n$ 在极限情况下的分布，关心跟 $\overline{X}_n$ 相关量（函数）的极限分布，Δ 方法就是在已知 $\overline{X}_n$ 极限分布的情况下，让我们能够直接写出相关量的极限分布

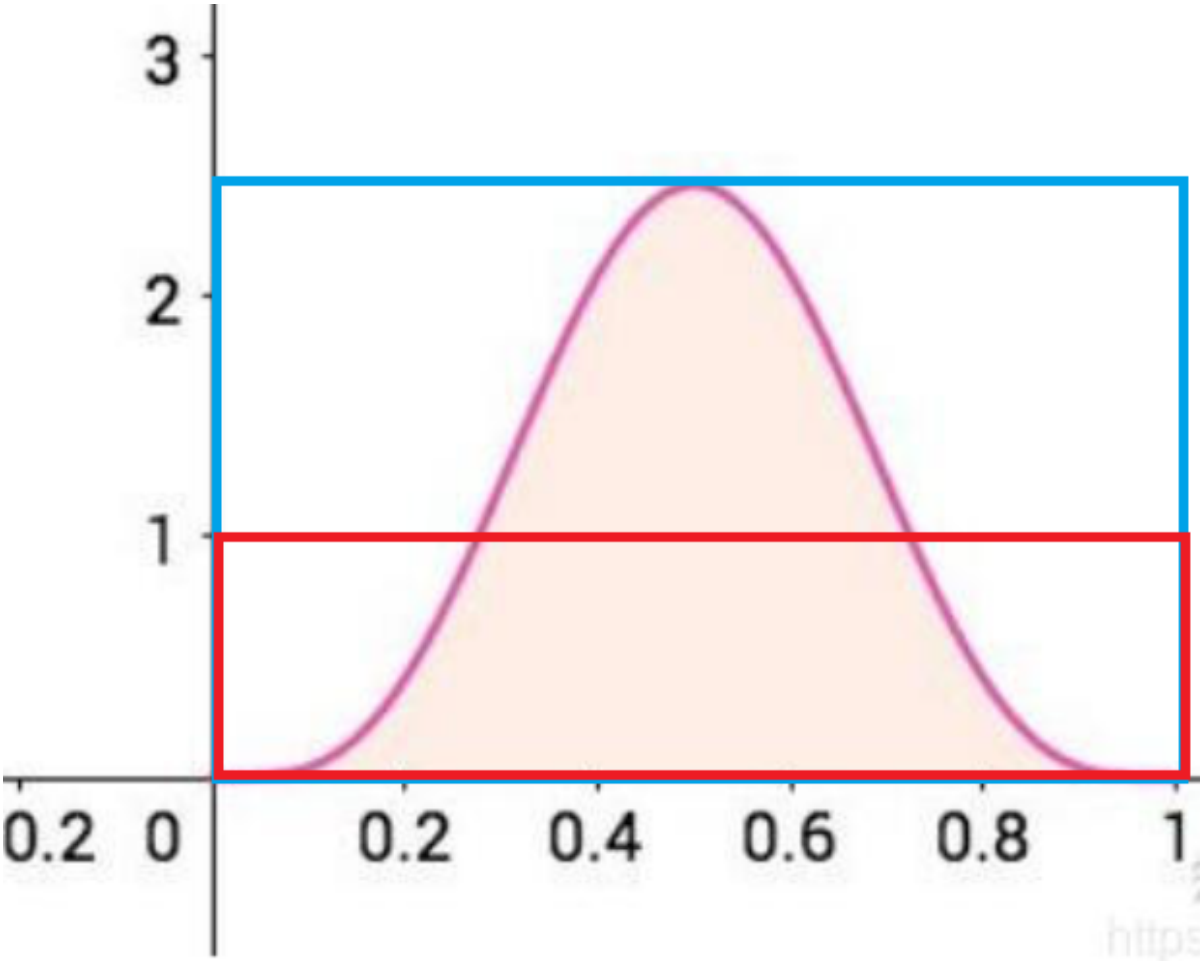
4.6. Generating a Random Sample

本小节是做一次实验，取一组随机样本的观测值

在实际计算过程中，一台计算机底层只能取均匀分布，其他分布都是从均匀分布导出，根据所需分布和均匀分布之间是否有显式的函数关系，我们可将生成随机样本的方法分为直接发和间接法

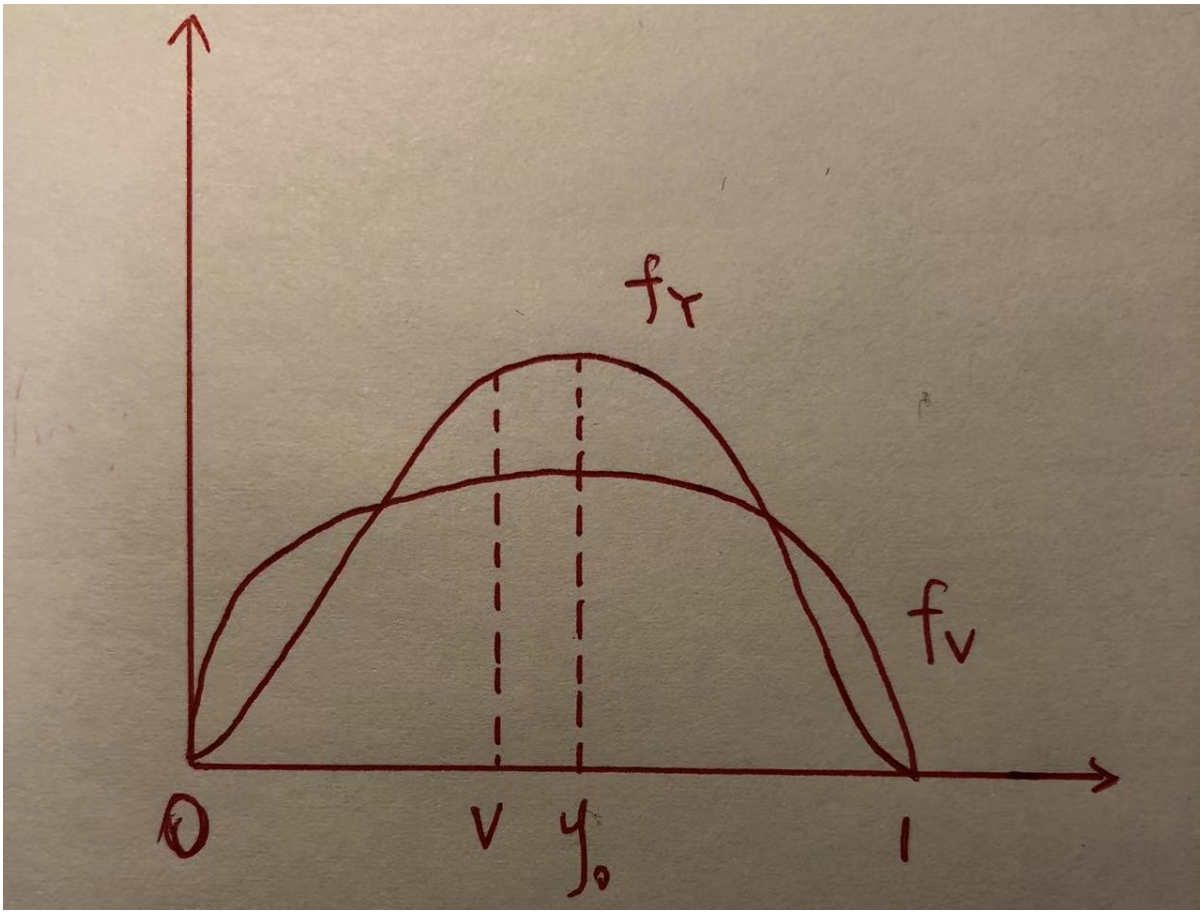
直接法：可以直接构造在 0-1 间均匀分布的 U，以及所研究随机变量 Y 和 U 取值间的函数映射关系，并且两者可以推出 Y 的分布就是我们研究的分布，生成样本时就可以随机在 0-1 间打点，并将通过函数将 U 值转换到 Y 值。所以重点就是找到函数关系，作者给出映射函数就是 Y 的累积分布函数的反函数。

间接法：直接法简单，但是“找到函数关系”这一点其实要求很高，间接法采取另外一种思路：在概率密度函数图像上进行撒点，如下：



先用一个矩形将整个密度函数刚刚好包起来，也就是长为随机变量区间长，宽为概率密度取值极差，在这个区域内均匀撒点（也就是横轴/纵轴都均匀分布撒点），比如 (u, v)，如果这个点在函数与横轴包成的区域内，则 u 为产生的一个随机样本。

舍选法：



在间接法中，本质上是将 V 均匀分布产生的点用 U 分布函数值去挑选，而舍选法则是推广，也就是任意分布的 V 产生的点用 U 分布函数值去挑选。

首先我们思考，舍选法的本质就是在**A**分布下产生点，然后按照**B**分布将一些点去除，从而使得产生的点满足**B**分布。用0-1分布打比方，A分布是 $P_A = 0.4$，B分布是 $P_B = 0.8$，现在我用电脑依照A分布产生了16个1，24个0，如果我要使这些数据满足B分布，我只要去掉20个0，使得总共有16个1，4个0。

首先找到使得如下比值最大的 y_0，这是因为从一个分布里选数据得到另一个分布，最通用的做法其实是每一个取值都裁掉数据点，比如说上面可以裁掉8个1和22个0，剩下8个1和2个0，但这是不明智的（产生的数据点浪费了），所以裁数据的基本原则是保留尽可能多的数据点。在 y_0 处的数据点全部保留，其他点数据进行裁剪

$$y_0 = \{y : \sup_y \frac{f_Y(y)}{f_V(y)}\}$$

在 y_0 , V 处产生的样本点个数分别为 N , Q，满足：

$$\frac{N}{Q} = \frac{f_V(y_0)}{f_V(V)}$$

求 V 处剩下的 Q'，其满足：

$$\frac{N}{Q'} = \frac{f_Y(y_0)}{f_Y(V)}$$

则：

$$
\begin{aligned}
Q^{'} &= \frac{f_Y(V)}{f_Y(y_0)}N = \frac{f_Y(V)}{f_Y(y_0)}\frac{f_V(y_0)}{f_V(V)}Q \\
&= (\frac{f_Y(V)}{f_V(V)}/\frac{f_Y(y_0)}{f_V(y_0)})Q \\
&= (\frac{f_Y(V)}{f_V(V)}/M)Q
\end{aligned}
$$

那么只要借助另一个0-1均匀分布的 U ，当 U 值小于 Q 前系数时保留数据，当大于时裁掉数据即可。

这里我们可以看出舍选法的一个"所谓"缺点，两种分布的支撑集（也就是定义域）一定要相同，“所谓”是因为相同可以用位置-尺度族搞定。

4.7. Miscellanea

4.7.1. MCMC算法概述

MCMC方法，第一个MC是Markov Chain，它可以造出任何分布；第二个MC指代Monte Carlo，其中心思想是“不断抽样，逐渐逼近”。实际操作中，Markov Chain产生我们需要的概率分布采样，再用Monte Carlo进行求解。我们可以看出，MCMC方法的重点是Markov Chain。

在这里论述一下Markov Chain的必要性，采样分布 π 我们现状是

- 计算机能做伪均匀分布
- 有诸如直接法，间接法，舍选法
- 有目标分布函数 π （意味着我们知道每个点 x 的概率函数值）

这种情况会遇到如下问题

- 对于一些二维分布p(x,y)，有时候我们只能得到条件分布p(x|y)和p(y|x)，却很难得到二维分布p(x,y)一般形式，这时我们无法用舍选法得到其样本集。
- 对于一些高维的复杂非常见分布p(x1,x2,...,xn)，我们要找到一个合适的q(x)和k非常困难，也就是无法用舍选法

可以看出，问题关键就是多维和高维使得我们掌握仅仅只有条件分布，而基础方法需要联合分布，所以我们借助另外的逻辑，将问题转化成我们用这些基本方法能做的子问题，这个另外的逻辑就是Markov Chain，详细来说，是Markov Chain中的Gibbs。

从MCMC推导过程来讲，MCMC逻辑=>M-H采样=>Gibbs采样的推导逻辑中，M-H也可以用于采样，但是它局限于一维，而一维问题我们基础方法就能解决，何必用M-H，而Gibbs才是解决我们碰到的多维/高维问题的最终武器。

4.7.2. 离散分布Markov Chain采样

$$P(X_{t+1}|\dots X_{t-2}, X_{t-1}, X_t) = P(X_{t+1}|X_t)$$

当前状态的概率分布，只依赖于上一个状态，而和除它之外的历史状态都没有关系，很自然我们可以将其表示为矩阵

$$P = \begin{pmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

行代表上一时刻状态，列代表当前状态，矩阵的意思是，如果上一时刻状态为 1 ，那么当前时刻状态为 1，2，3的概率分别为0.9，0.075，0.025。引入矩阵乘法

$$\pi P = (0.7,\ 0.1,\ 0.2) \begin{pmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} = (0.695,\ 0.1825,\ 0.1225)$$

它的意思是，上一个时刻的状态也不是确定的，为概率分布 π ，而在此之上，如果要计算当前状态的概率分布，就用 π 乘 P 即可。比如，π 乘上P的第一列，0.7*0.9为在上一时刻为1的情况下当前状态为1的概率；0.1*0.15为在上一时刻为2的情况下当前状态为1的概率；0.2*0.25为在上一时刻为3的情况下当前状态为1的概率，三者相加，就是当前状态为1的概率，这个值放在结果的第一个位置上。可以看到概率逻辑和矩阵运算逻辑是一致的。

至此，我们完成了一件事，那就是将满足马尔可夫特性的条件概率翻译成了矩阵乘法，那么我们之后对于马尔可夫链的研究就可以通过研究矩阵来进行。

这个矩阵被称为状态转移矩阵，它有如下性质

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \dots & \pi(j) & \dots \\ \pi(1) & \pi(2) & \dots & \pi(j) & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \pi(1) & \pi(2) & \dots & \pi(j) & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

可以推出，不管初始概率分布是什么样，最终经过足够多的状态转移后，概率分布都会变成 $(\pi(1)\ \pi(2)\ \dots \pi(j)\ \dots)$

而从实际操作上来说，就是我在时间步一按照任意分布产生A1点，然后在时间步二，按照一已产生A1点的条件下时间步二的条件概率产生A2点，然后产生A3，A4...，在时间步足够大时产生的An本质上就是由 $(\pi(1)\ \pi(2)\ \dots \pi(j)\ \dots)$ 独立分布产生的。

再有，如果我想产生多个点，我没有必要再从头开始，因为在足够大的n之后，每一个时间步分布都收敛于 $(\pi(1)\ \pi(2)\ \dots \pi(j)\ \dots)$，那么我只要输入An，接着产生点就行了。这个结论我觉得不太自然，后一个点的产生依赖前一个点，也就是说时间步上临近的两个点取值存在联系，为什么他们会是独立采样于 $(\pi(1)\ \pi(2)\ \dots \pi(j)\ \dots)$？原因是：条件分布产生点只是过程（形式），我们将n之后的每一个时间步分隔开看，他们独立分布 $(\pi(1)\ \pi(2)\ \dots \pi(j)\ \dots)$ 是结果（本质）

接下来的任务就是给出一个我们需要的特定分布，如何求出状态转移矩阵？

分布和矩阵元素关系如下：

$$\pi(i)P(i,j)=\pi(j)P(j,i)$$

通过方程就能直接算出来矩阵元素具体值，但是我猜这样做工程上是不划算的（解方程要浪费时间和资源），所以先取一个随机的状态转移矩阵 Q，那么很明显 Q 中元素不会满足上述等式

$$\pi(i)Q(i,j)\neq\pi(j)Q(j,i)$$

但是

$$\pi(i)Q(i,j)\pi(j)Q(j,i)=\pi(j)Q(j,i)\pi(i)Q(i,j)$$

是恒成立的，那么

$$P(i,j)=\ Q(i,j)\ \pi(j)\ Q(j,i)=\ Q(i,j)\ \alpha(i,j)$$

实际操作中，先根据 $Q(i,j)$ 产生点 j^*，然后再0-1均匀分布产生数，如果此数小于 $\alpha(i,j)$ 那么当前点为 j^*，如果大于，那就再当前点再来一次（本质是间接法）

到这里Markov Chain采样的整体架构已经论述清楚，这里出现我思考的一个问题：

每一行 $\sum\limits_j Q(i,j)\ \pi(j)\ Q(j,i)<\sum\limits_j Q(i,j)=1$，那么为什么可以将找到的矩阵当作状态转移矩阵？打个比方，我们按照公式找到了如下矩阵 $P_1=\begin{pmatrix}0.1 & 0.3\\0.2 & 0.2\end{pmatrix}$，因为我们在整个过程中从来没有限制过每一行为1，这种情况必会发生。按照严格定义这肯定不是状态转移矩阵，但我们依旧可以按照它产生点，也就是当我处于状态1时，0-1中均匀产生随机数，落在0-0.1中取状态1，落在0.1-0.4中取状态1（第二行同理），那么我们真正的状态转移矩阵是 $P_2=\begin{pmatrix}0.25 & 0.75\\0.5 & 0.5\end{pmatrix}$。

要解答这个问题，我们只要说明：真状态转移矩阵也是满足要求的。真状态转移矩阵的元素为

$$P_r(i,j)=\frac{Q(i,j)\ \pi(j)\ Q(j,i)}{\sum_k Q(i,k)\ \pi(k)\ Q(k,i)}$$

而

$$\pi(i)P_r(i,j)=\pi(i)\frac{Q(i,j)\,\pi(j)\,Q(j,i)}{\sum_k Q(i,k)\,\pi(k)\,Q(k,i)}$$
$$\pi(j)P_r(j,i)=\pi(j)\frac{Q(j,i)\,\pi(i)\,Q(i,j)}{\sum_k Q(j,k)\,\pi(k)\,Q(k,j)}$$

两者对比我们发现分子是相等的，而分母是对相同的量进行求和，但这里并不满足两者相等

问题没有解决，但是接着往下走，默认是真状态转移矩阵是满足要求的

我们可以发现，上述描述的全是离散情况，连续情况 π 是连续函数，而 P 同样是连续函数，$P(i,j)$ 为带有一个参数的的概率分布函数，当 X_{t-1} 取定，它作为参数带入 $P(X_t|X_{t-1})$（也就是 $P(i,j)$ ），再根据概率函数取 x_t 。

4.7.3. 离散分布M-H采样

α 过小会导致每一个时间步要很多次才能得到点，那我们就调大 α ，

$$\pi(i)Q(i,j)\alpha(j,i)=\pi(j)Q(j,i)\alpha(i,j)$$

既然在我们的操作流程中 α 是拒绝率，那么默认它小于等于 1 ，所以

$$\alpha(i,j)=min\{\frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)},1\}$$

通常取的 Q 对称

$$\alpha(i,j)=min\{\frac{\pi(j)}{\pi(i)},1\}$$

这就是 M-H采样

4.7.4. 离散分布Gibbs采样

它还是在 $\pi(i)P(i,j)=\pi(j)P(j,i)$ 上面做文章

$$\pi(X_{A1},X_{A2})\pi(X_{B2}|X_{A1})=\pi(X_{A1},X_{B2})\pi(X_{A2}|X_{A1})$$
$$\pi(X_{A1},X_{A2})\pi(X_{C1}|X_{A2})=\pi(X_{C1},X_{A2})\pi(X_{A1}|X_{A2})$$

只不过现在状态不是一维，而是多维，将其他维数固定，一维活动，每一维存在一个状态转移矩阵。可以看出，状态转移矩阵都是由条件概率构成，因此它就解决了高维分布采样的问题

4.7.5. 连续分布Markov Chain采样

离散分布中的状态转移矩阵其实就是在前一状态确定的情况，下一状态的分布情况，也就是条件分布

$$\pi(x)P(y|x) = \pi(y)P(x|y)$$

4.7.6. 贝叶斯统计和MCMC方法的结合

贝叶斯公式为

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

因为 $P(y)$ 是常数，我们可以改写成

$$P(\theta|y) = CP(y|\theta)P(\theta)$$

这就是MCMC方法的目标分布。回忆前面寻找状态转移矩阵的公式

$$P(i,j) = \ Q(i,j)\ \pi(j)\ Q(j,i) = \ Q(i,j)\ \alpha(i,j)$$

改写成连续形式

$$P(\theta \to \theta') = Q(\theta \to \theta')\ P(\theta'|y)\ Q(\theta' \to \theta)$$

我一直疑问的是贝叶斯公式中的常数如何影响MCMC公式？答案是不影响，因为之前我们说过

$$\sum_{\theta'} Q(\theta \to \theta')\ P(\theta'|y)\ Q(\theta' \to \theta) < \sum_{\theta'} Q(\theta \to \theta') = 1$$

我们按照公式计算实际上遵循的状态转移矩阵式

$$P(\theta \to \theta') = \frac{Q(\theta \to \theta')\ P(\theta'|y)\ Q(\theta' \to \theta)}{\sum_{\theta'} Q(\theta \to \theta')\ P(\theta'|y)\ Q(\theta' \to \theta)} = \frac{1}{C(\theta)} Q(\theta \to \theta')\ P(\theta'|y)\ Q(\theta' \to \theta)$$

但是我们并不用管 $C(\theta)$，只要根据我们找到的 $P(\theta \to \theta')$ 采样就行，将先验部分带入

$$P(\theta \to \theta') = \frac{C}{C(\theta)} Q(\theta \to \theta')\ P(y|\theta')P(\theta')\ Q(\theta' \to \theta) = \frac{1}{C(\theta)} Q(\theta \to \theta')\ P(y|\theta')P(\theta')\ Q(\theta'$$

也就是贝叶斯公式中的常数被 $C(\theta)$ 吸收掉了，或者说这个常数完全由 $C(\theta)$ 控制，而我们又不管 $C(\theta)$

最终导致要使得MCMC方法奏效，对于封装好的MCMC包，使用者只要给出 $P(y|\theta')P(\theta')$ 的显式表达式即可，编程上就是给出从参数到数据的分布链关系，MCMC包只要严格遵守 $Q(\theta \to \theta')\ P(y|\theta')P(\theta')\ Q(\theta' \to \theta)$ 进行状态转移即可

有些论文里会涉及推导M-H中的拒绝率，我疑惑的点在于：我只要输入从参数到数据的分布链关系，其他的计算机自己算就行了，为什么还要推拒绝率？

这是因为计算机本身是很笨的，它只会按照较简单的分布产生随机数，而软件开发者需要基于计算机开发Domain-Specific的MCMC软件，换句话说，只要问题（模型）一换，市面上就找不到可以实现当前问题的MCMC软件。论文讨论将贝叶斯方法用到新的问题上，实际上也就是在论述新软件的逻辑，因此从先验分布，到如何采样都必须详细论述。

4.8. 总结

本章逻辑：

- 总体分布，特征有均值 μ 和方差 σ^2
- 抽样分布，导出了两个分布：$\overline{X}$ 样本均值分布，S^2 样本方差分布，又因为存在 $E(\overline{X}) = \mu$，$E(S^2) = \sigma^2$，所以等式前者称为等式后者的无偏估计量，为了求后者，我需要将分布的具体形式推导出来
- 抽样分布，还有一类为次序统计量
- 收敛，三个定理，极限均值分布收敛到常数，以及一个极限正态分布，从而有 $\overline{X} = \hat{\mu} = E(\overline{X}) = \mu$
- Δ 方法，讨论极限均值函数的分布

一直到这一章，本书讲的其实都是同一个内容，在已知一个随机变量A的分布情况，它的函数 $B = f(A)$ 具有什么样的性质。这个思路可以有两种表现形式：函数分布推导，期望方差推导

分布函数推导：有矩母函数，变量替换

期望推导：

$$E(aX + bY) = aE(x) + bE(y)$$

X 和 Y 相互独立，有

$$E(XY) = E(X)E(Y)$$

Stein引理：对于正态分布，恒有如下等式，可用来削减所求期望的复杂程度

$$E[g(X)(X - \theta)] = \sigma^2 E g^{'}(X)$$

Jensen不等式：对于凸函数 g (凸函数是向下的包），当函数 g 存在线性区域，并且随机变量取值都处在此区域中，等式成立

$$E(g(X)) \geq g(E(X))$$

方差推导：

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2ab\,Cov(X,Y)$$

基础知识点：

Gamma函数：

$$\Gamma(z+1) = \int_0^\infty x^z e^{-x} dx$$
$$\Gamma(z+1) = z\Gamma(z)$$

4.9. 习题

5.9：“全体样本点位于一条直线上”，这个条件让我疑惑了一会。当时的想法是，每一个样本都是分布，分布怎么能在坐标系中表现出来？

这个条件的正确理解方式：

首先明确，样本是分布这一点是没有问题的，位于直线上是说，在每一次具体实验中收集到的实际观测数据都有“两者位于一条直线上”这个现象，那么只有一个可能：随机变量 X 和 Y 刻画的是同一个系统，他们之间存在函数关系 $(Y-\mu_y) = C(X-\mu_x)$

5.10：Stein Lemma，给目标期望化简

5.11：Jensen不等式

5.13：这道题可看出 S^2 是 σ^2 的无偏估计，但是 S 并不是 σ 的无偏估计

5.17：这道题展示了 F 分布的若干性质：

- F 分布函数从卡方分布导出
- F 分布的期望和方差也可以借助卡方分布求出
- 满足 F 分布的随机变量的倒数仍满足卡方分布
- $X \sim F_{p,q}$ 那么 $\frac{\frac{p}{q}X}{1+\frac{p}{q}X} \sim Beta(\frac{p}{2}, \frac{q}{2})$

5.18：t分布和F分布的若干关系：

$$X \sim t_p \Rightarrow X^2 \sim F_{1,p}$$

$$X \sim t_p \Rightarrow \ X \sim n(0,1)\ when\ p \to \infty$$

$$X^2 \sim F_{1,p} \Rightarrow X^2 \sim \chi_1^2$$

在两种证明极限情况中，都使用了Stirling公式，即在x足够大时，有：

$$\Gamma(x) \approx \sqrt{2\pi} e^{-x} x^{x-\frac{1}{2}}$$

5.20：这道题的解题思想可以被更广泛地阐述

X 是正随机变量，分布函数是 $f_X(x)$，Y 是正随机变量，分布函数是 $f_Y(y)$，如何求 $t, \frac{X}{Y} \leq t$ 的概率分布？

- $X \leq Yt$，固定 t
- 固定 Y 为 y 的时候，收集所有满足条件的 X 构成的概率，也就是 $P(X \leq yt)$
- 这时计算的还只是一个 y 时的概率，还要对所有的 y 也就是 Y 收集概率，形式为积分

$$P(\frac{X}{Y} \leq t) = \int_0^\infty P(X \leq yt)\, f_Y(y)\, dy$$

- 左右是 t 的函数，同时求导，并且右边积分与 t 无关，因此求导可以放在积分号里面

$$f_T t = \int_0^\infty f_X(yt)\, y\, f_y(y)\, dy$$

Chapter 5 Principles of Data Reduction

5.1. The Sufficiency Principle

5.1.1. 函数

函数的定义中，x 和 y 的关系可以是一对一，可以是多对一，但不可以是一对多。从信息流动的角度，如果是一对一，那么两个不同的数 x_1 和 x_2 通过函数转换成了 y_1 和 y_2 两个不同的数，我们在已知 y_1 和 y_2，不知道 x_1 和 x_2 的情况下却仍能分清楚 x_1 和 x_2，那么我们可以认为经过函数处理，信息量没有任何变化。如果是多对一，那么也就是 x_1 和 x_2 对应同一个 y，我们在已知 y 时，是分不清 x_1 和 x_2 的，那么我们可以认为信息量经过函数之后减少了。

从图像上理解，假设自变量是一个二维平面，函数其实就是在这个平面上画圈，并且给不同的圈分配不同的数字，一对多就是一个圈里面有很多点，一对一就是一个圈里面只有一个点。

而如果函数形式上在 y 后面再加一个函数导出 z，就是将已经画好的圈再以几个为一组包到更大的圈中，信息量也就进一步被压缩。

5.1.2. 充分统计量

其核心就是上述函数概念，在概率统计中，X 的具体取值是存在概率的，并且概率函数是存在参数的，也就是每个具体 x 概率值受到参数 θ 控制，那么这么多具体 x 以及对应的概率构成的整体也就能完整反映 θ 的信息（从这里来看，如果样本数为1，那么 X 就是 θ 的充分统计量，按照定义也是对的）。

统计量是基于随机样本的，也就是很多个独立并且包含同一个参数的概率分布函数相乘，那么 $(X_1, X_2, \dots, X_n)$ 的具体取值以及概率分布能够完整刻画 θ 信息。$(X_1, X_2, \dots, X_n)$ 确实是 θ 的充分统计量，可是问题在于它太基础了，包含杂质也太多了，并且也不符合我们对于统计量的一贯印象（总结数据，而不是照搬数据），所以我们希望能够能够在不损失 θ 信息的情况下压缩无关信息，也就是用函数进行映射。

可是又产生的新的问题：我们知道 $(X_1, X_2, \dots, X_n)$ 是充分统计量，但是我们如何判定某一个随机样本的函数 $T(x)$ 为充分统计量呢？书中介绍了两种方法

$$P_\theta(X = x | T(X) = T(x)) = \frac{p(x|\theta)}{q(T(x)|\theta)}$$

以及

$$T(X)\text{为充分统计量} \ \leftrightarrows \ f(x|\theta) = g(T(x)|\theta)h(x)$$

第一种其实是充分统计量的定义式，当 X 在 $T(x)$ 下的条件分布和 θ 没有关系，也就是 X 的概率质量（密度）函数除以 $T(X)$ 的概率质量（密度）函数得到的结果中不含 θ，那就代表 $T(X)$ 是 θ 的充分统计量。直观上来说，就是 $T(X)$ 因为提取了 θ 的全部信息，变成了 θ 的一个化身，那么给定 $T(X)$ 也就意味着给定 θ，最终得到的结果肯定就不含 θ 了。而从图像上来看，就是在一个圈里面，X 的分布固定下来，不再是 θ 的函数。判定流程：写出样本联合概率密度函数 p，求出充分统计量的分布函数 q（有些能直接写，比如正态分布的 $\overline{X}$），两者相除。它的缺点在于需要先找到充分统计量（直觉），然后还要计算它的分布函数（麻烦）。

第二种称为因子分解定理，能够只考察样本分布函数就求得充分统计量。其中 $g(T(x))$ 是 $T(x)$ 的分布函数的核心，$h(x)$ 是在 $T(x)$ 下的 x 分布函数的核心。判断流程：将概率密度函数分解成两个部分，其中一个部分含有参数 θ 以及随机样本和 θ 纠缠不清的部分（无法分开的部分），比如 $I_{(0,\infty)}(x)$，当中的随机样本及其函数部分称为 $T(x)$，而另一个部分只含随机样本及其函数，那么就有 $T(x)$ 就是 θ 的充分统计量。多参数的充分统计量为向量形式，因子分解定理的操作和单参数一样，将参数尽量和样本随机变量分离，得到 $g(T(x)|\theta)$ 和 $h(x)$，不同的是，在 g 中关于样本不同形式的函数构成充分统计量向量 $(T_1(x), T_2(x), \dots)$

5.1.3. 极小充分统计量

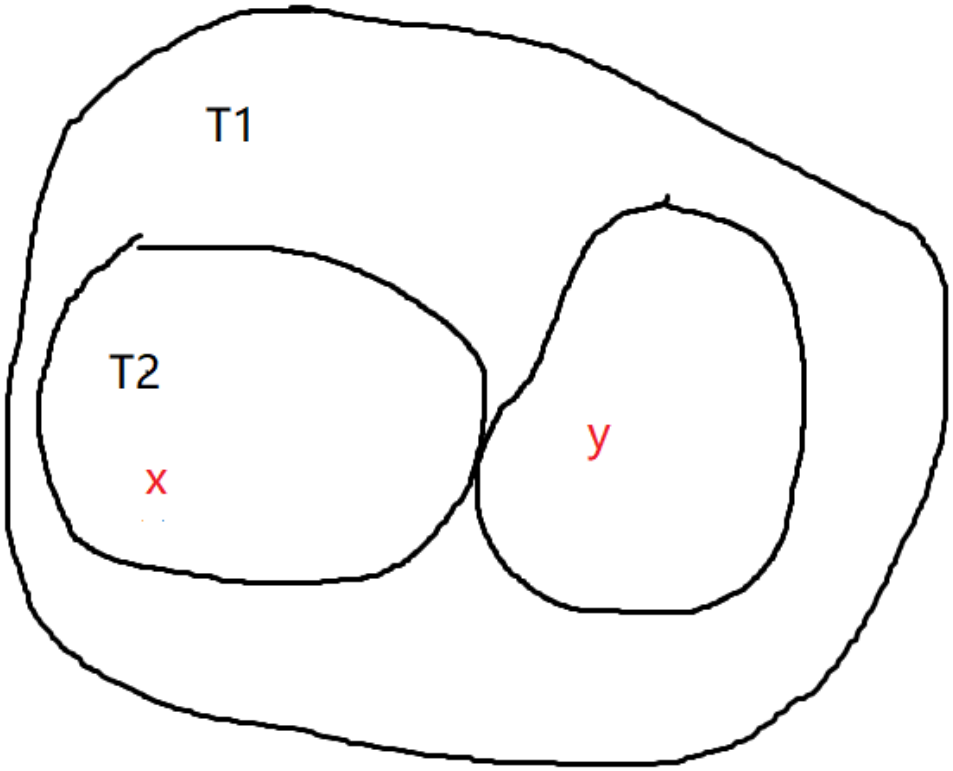
将 θ 信息量压缩到最小的充分统计量，在这个充分统计量中只包含 θ 信息。

第一种判定：对于 θ 的任意充分统计量 $T'(x)$，本身为 θ 充分统计量的 $T(x)$，都有 $T(x)$ 是 $T'(x)$ 的函数

第二种判定：

$$\text{当且仅当 } T(x)=T(y)$$
$$\text{有 } \frac{f(x|\theta)}{f(y|\theta)} \text{是}\theta\text{的常函数}$$

极小充分统计量可以通过下图理解，T_1 和 T_2 都是 θ 的充分统计量



从图像上看，同一个参数的充分统计量一定是包裹的关系，而大圈不能切割小圈；从函数形式上来看 T_1 的具体函数形式总是包含 T_2

这时我们来看第二种极小充分统计量判定思路：

x 与 y 都是随机变量的具体取值，而使用两个字母的作用是说明他们是两个不同的点，首先我们来看判定条件中的 当（**if**），如果 $T_2(x)=T_2(y)$，那么也就是 x 和 y 都被划分到 T_2 划分的同一个小圈里，在同一个小圈的条件下，里面两个点的概率比和 θ 没有关系，那么 T_2 肯定就是一个能够完整收集 θ 信息的正确划分（充分统计量基础判定也就是这个意思，同一个圈里的概率分布和 θ 没有关系，所以第二种极小充分统计量判定去掉“仅当”，可以用作充分统计量判定）。所以当（**if**） 说明当前划分确实是充分统计量。

可是我们看到 T_1 划分也是可以的，所以我们如果关注判定的比值式的时候会发现：$T_2(x)=T_2(y)$ 确实可以使得比值结果为 θ 无关量， 可我们同时还会发现，函数形式中包住 T_2 的 T_1 在 $T_2(x)\neq T_2(y)$ 的时候能够做到 $T_1(x)=T_1(y)$ 并且使得比值也与 θ 无关（空间划分包裹必会函数形式包裹，同时为多对一进行了信息压缩）。所以 T_2 明显不是极小统计量。所以我们要加上 仅当（**only if**），它的意思是从 T_2 往外看，没有任何包住它的函数形式能做到在 T_2 不相等的情况下相等（多对一信息压缩），并且使得比值为 θ 无关量。

所以极小充分统计量第二种判定中， if 保证当前量是充分统计量，only if 保证当前量是极大充分统计量。

5.1.4. 实际情况中充分统计量与极小充分统计量的判定

充分统计量判定步骤：

- 将联合分布函数拆成 $f(x|\theta)=g(T(x)|\theta)h(x)$ ，
- 在 g 内所有随机变量的无 θ 组合（函数）形式，就是 θ 的充分统计量

在题目中我体会到了一个细节：有些题目中， $h(x)$ 中 x 感受到 θ 的作用，是通过 $T(x)$ 作为媒介进行的，也就是 $T(x)$ 会在 g 和 h 中同时出现。这一点重要是因为它能帮我们判断谁是 $T(x)$ 。比如例6.2.7，我们之所以找到了 $\overline{X}$ 为 μ 的充分统计量，是因为只有它作为一个整体出现在了 g 和 h 中，令它为充分统计量是一件很自然的事。但是我们不能把这一点视为理所当然，而应该把 $h(x)$ 正好具有函数形式当作锦上添花。

习题6.5展示了另一种情况，它在构造因子分解的时候，将全体函数（以及用示性函数形式加进来的定义域）都当作 g ，而 $h(x)=1$ ，我们就没有办法从 $h(x)$ 中参考应该取 g 中哪个结构作为充分统计量。

$$f=\frac{1}{2^n\ n!\ \theta^n}I_{(-\infty,1+\theta)}(max(\frac{x_i}{i}))I_{(1-\theta,\infty)}(min(\frac{x_i}{i}))$$

那我们只能从小往大观察，找不含参数的随机变量函数，首先我们能看出 $(\frac{x_1}{1},\frac{x_2}{2},\dots,\frac{x_n}{n})$ 肯定是，再往外扩，发现 $max(\frac{x_i}{i})$, $min(\frac{x_i}{i})$ 也是，但是再往外扩就会发现示性函数区间中存在 θ ，就此打住。因为这道题找的是充分统计量，而不是最小，那么随便选，最自然选法肯定是 min 和 max 。在寻找充分统计量时，因为函数包裹能产生多种选择方案，这也是这类题目中要对充分统计量的维数进行限制的原因（这里说为二维，那么只能写 min 和 max ）

极小充分统计量判定步骤：

- 将函数写成 $\frac{f(x|\theta)}{f(y|\theta)}$
- θ 系数为零的结构

其实单就比值成立，它说明的仍是相关函数结构是充分统计量，而我们使用这个方法，而不是极小充分统计量的定义，优点在于最大的函数结构（也就是书上定义的 仅当 成立情形）可以通过 θ 系数为零直接找到

5.1.5. 次序统计量

在做题的时候，我发现很多时候函数形式不好化简，只能是 $x_i = y_i$ 使得比值为 θ 无关结构，这时往往 $x_{(i)} = y_{(i)}$ 也能使得比值为 θ 无关结构。次序统计量就是在样本随机变量（充分统计量）基础上去掉了一丁点杂质的充分统计量。

充分统计量从大到小

- 样本随机变量=>次序统计量：稍微压缩，维数没有变化
- 次序统计量=>其他统计量：进一步压缩，维数开始减小

很多时候我们只能压缩到次序统计量（这时次序统计量也就是极小充分统计量），而只有指数族分布以及其他一些特定分布能继续往下压缩。这也是在p252中间，说"其他分布中少有维数低于样本大小的充分统计量"的原因

辅助统计量判定：

如果统计量 $S(X)$ 的分布与 θ 无关，则称 $S(X)$ 为辅助统计量（ancillary statistic）

5.1.6. 完全统计量

Jun Shao完全统计量阐述比较清晰，完全统计量的逻辑：

- 现在有一个统计量 $T(x)$（和充不充分无关），并且它的分布函数中有一个参数 $f(x|\theta)$
- 这个世界上有无数个函数，我们希望找到函数 g 满足对任意 θ，都有 $E_\theta(g(T(x))) = 0$，很自然就是用 $E_\theta(g(T(x))) = 0$ 去求这个（或这些）函数 g
- 我们最终一定能够推出满足条件的函数为 $g(T(x)) = 0$（也就是满足 $E_\theta(g(T(x))) = 0$ 的函数形式 g 只可能为全零函数）

注意：

- 这个逻辑是如果具有，令 $E(g(T(x))) = 0$ 为 A，令 $g(T(x)) = 0$ 为 B，一个完全统计量只要能走到 A，那么它一定能顺着 A 走到 B，但是它能不能走到 A 我并不关心（取决于 g 的具体形式）。
- 在定义完全统计量时，虽然名字里有“统计量”三个字，但是所有操作都是围绕着 T 及其分布函数 $f(t|\theta)$ 展开的，也就是“完全”这个概念是着重于一个随机变量（T）以及它的分布，只要它满足我们规定的性质，就被称作完全
- 完全统计量是针对分布族 $f(t|\theta)$，意思是 $E_\theta(g(T(x))) = 0$ 限制的是 g 必须在所有可能的 θ 取值时都做到 $E_\theta(g(T(x))) = 0$，最终得到唯一的 g 函数形式是全零函数。也可以理解成在某个特定的 θ_0 下，我们可以通过限制条件找到一堆不同的 g_0；然后在下一个点 θ_1，我们在 g_0 中选择在当前点满足限制条件的函数形式得到 g_1；然后再下个点，再下个点，最终会发现在每个点都符合限制条件的函数形式只剩下一种，那就是全零函数。

- p261最上面的文字就是第三点的意思，

  “ $X \sim n(0,1)$ 为一个特定分布，我们发现取 $g(x)=x$ 会使得 $E_(g(T(x)))=0$ ，但是 $g(x)=x$ 并不是全零函数，然后我们说 $X \sim n(0,1)$ 不是完全统计量”

  这个说法是错误的，错误原因是关注了一个特定分布，上面的现象确实存在，但是这个现象的功能仅为通过 $\theta=0$ 这个点帮我们过滤了 $g(x)=x$ 这个函数，和最终我们判定 $n(\theta,1)$ 是否为完全分布族（也就是 X 是否为完全统计量）没有关系。

那么证明一个统计量是完全统计量就要证明函数 g ，只要它满足 $E_\theta(g(T(x)))=0$ ，那么一定能够推出 $g(T(x))=0$ （g 为全零函数），那么就不能去用特定的函数结构，只能用符号 g ；而证明一个统计量不是完全统计量，就要说明有一个例外就行，就是拿一个特定的函数结构去所有的 θ 处过滤，比如平方，说明虽然 $E_\theta((T(x))^2)=0$ ，但是 $(T(x))^2$ 确实不是处处为零的。

6.2.24 Basu定理证明卡了很久的也是上述如果具有没有理解清楚，作者构造的是：

$$g(t)=P(S(X)=s|T(X)=t)-P(S(X)=s)$$

并且已经说明 $A=>B$ 走得通，那么任务就是证明 $E_\theta g(T)=0$ （也就是这个函数形式能走到 A ），接着也就能到 B 了。

5.2. The Likelihood Principle

5.3. 习题

6.2：这道题的启示在于，如何参数在定义域中，我们一般是通过示性函数将其放到函数表达式中。

Chapter 6 Point Estimation

6.1. Introduction

6.1.1. 变量与值的关系

只要是字母就可以动，体现在于取值变化或是积分微分

随机变量总共有三层：

- 随机变量 X ：它是一个变量，它具有概率分布，方差等的性质，可动
- 取值 x ：也就是高中学的因变量，因此带入具体值，求导积分等都能执行
- 数：1，2，3...

这里要提一下参数 θ，它是取值，但和 x 不同的是，它不存在上一层（随机变量），因为在频率学派中，它不是概率分布的。在求估计量时会导出 $\theta = g(x)$，我们会说 $\tilde{\theta} = g(X)$ 和 $\hat{\theta} = g(X)$ 是估计量，它的分布如何，估计值为 $\tilde{\theta} = g(x)$ 和 $\hat{\theta} = g(x)$，但是不会说 θ 的分布如何。而在贝叶斯学派中，就会说 θ 的分布如何。（注意：θ 不区分大小写）

6.1.2. 前六章逻辑

统计当中与具体实验不相关内容就一个：X 的分布已知，$Y = g(X)$，研究 Y 分布性质。它的逻辑是随机变量满足 $Y = f(X)$，那么具体值也有 $y = g(x)$，就可以通过这个关系导出 Y 的分布以及性质。

- 第二章研究的是一维随机变量
- 第三章介绍一些一维随机变量常见分布族（也就是概率函数形式一样，θ 取不同值）
- 第四章介绍多维随机变量
- 第五章介绍由很多个独立同分布的随机变量组成的随机变量的性质，也就是 $g(X_1, \dots, X_n)$ 的性质
- 第六章主要介绍的还是 $g(X_1, \dots, X_n)$ 的性质，只不过加上一个考量，它与 X 中参数 θ 的关系，这时 g 被称作统计量

6.1.3. 第七章逻辑

参数 θ 是未知具体值：

我们要去估计这个具体值，逻辑是：在某一次实验中我们得到了 $(x_1, \dots, x_n)$，我们认为这一系列观测到的具体值能够反映 θ，我们的目的就是要从这些值中算出 θ 估计值 $h(x_1, \dots, x_n)$。有两种思路：

- 样本矩等于总体矩：$\frac{1}{n}\sum x_i^j = E(X^j)$，等式构造是数等于数
- 极大似然估计：$\underset{\theta}{max}\ f(x_1, \dots, x_n|\theta)$，一个值 θ 使得 $(x_1, \dots, x_n)$ 数的概率值最大

也就能够得到 $\theta = h(x_1, \dots, x_n)$。正确形式应该确实应该这样写，只不过我们需要在实验之前（不知道 $x_1, \dots, x_n$）给出一个指导性框架，那么也就是 $\theta = h(X_1, \dots, X_n)$。所以我的理解是估计量的形式（具体值 θ 等于随机变量的函数）实属无奈之举，它的侧重点在于我一旦知道 x 就能根据 X 和 θ 构建的等式算出 θ 的具体值，而不是 θ 真的是一个随机变量。根据用的方法不同，在符号上做区别：矩法 $\tilde{\theta} = h(X_1, \dots, X_n)$；极大似然法 $\hat{\theta} = h(X_1, \dots, X_n)$。

它叫做估计量，形式上和统计量一样，都是随机变量的函数，书中明确指出了：**Any statistic is a point estimator.** 同样一个函数，在统计推断的不同阶段被使用，就会被叫做不同的名字。

参数 θ 是一个随机变量：

逻辑是先验 θ 是一个遵从 $\pi(\theta)$ 概率密度函数的具体值，我们做一次实验，那么 $f(x|\theta)$ 就是在具体 θ 的条件下发生 x 的概率，我们通过这个得到修正的具体 θ 出现的概率

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

6.2. Methods of Finding Estimators

6.2.1. 矩法

矩的用法有两种：

第一种：在利用样本构造总体 k 个参数估计量的时候，用前 k 阶样本矩（注意它的形式）分别等于前 k 阶总体矩，构造 k 个方程，也就能解出每一个 θ 的表达式了

第二种：得到统计量分布的近似，或称“矩匹配”。两个相近分布 A 和 B，A 参数已知，B 中有未知参数，如果想用 B 来拟合 A（也就是求 B 中参数的估计量，那么就用 A 的各阶矩等于 B 的各阶矩。

在阐述第二种用法的例子中，我的一点疑惑在于例7.2.3中在倒数第二步，求得的估计量是：

$$\hat{\nu} = \frac{(\sum a_i EY_i)^2}{\sum \frac{a_i^2}{r_i} EY_i^2}$$

但之后就可以直接将期望符号去掉，变成

$$\hat{\nu} = \frac{(\sum a_i Y_i)^2}{\sum \frac{a_i^2}{r_i} Y_i^2}$$

说这个就是 ν 的估计量，为什么期望符号可以直接去掉？

6.2.2. 极大似然法

现实当中观测到的结果应该是概率函数中概率最大的点

参数为连续值：直接求导，导数为零的点逐个排查

参数为离散值：$L(\theta-1) \leq L(\theta),\ L(\theta) < L(\theta-1)$

似然函数不变性的用法：要求一个复杂结构的 MLE ，只要找到它里面基础结构的 MLE 就行

6.2.3. EM算法

极大似然法是有局限性的，比如存在隐变量，也就是概率分布模型中存在我们无法获取观测值的变量。具体来说，可以是由于此变量的属性导致我们无法观测，那么数据中所有的此变量观测值都会缺失 $((x_1),(x_2),\dots,(x_n))$；也可以是由于疏忽导致数据遗漏，那么我们掌握的数据中对应每一个变量都会有值，但零星几条数据会缺失部分 $((x_1),(x_2,y_2),\dots,(x_n,y_n))$。

这个时候基本的极大似然法就失效了，但我们还是想接着用，而EM算法（全称Expectation Maximization算法）就是对于极大似然法的改进，使其能被用在缺失数据上。

整体思路讲完，我们来讲技术细节

观测数据用 Y 表示，缺失数据用 Z 表示，极大似然法需要极大化 $L(\theta|Y,Z)=log\ P(Y,Z|\theta)$，EM算法需要极大化 $L(\theta|Y)=log\ P(Y|\theta)$（这是很自然的，我实际使用的似然函数肯定是基于我有的数据），那如何将前者转化成后者呢？答案是把 Z 积掉，也就是 $log\ P(Y|\theta)=log(\sum\limits_{Z}P(Y,Z|\theta))$ 或 $\log\ P(Y|\theta)=log(\int\limits_{Z}P(Y,Z|\theta))$（取决于Z是连续还是离散变量），李航书中说这因为这个形式包含和的对数，所以难，所以要转换成别的形式，不太明白。但总的来说，就是这个基本形式不好。

和的对数不好，对数的和好，如何将前者转换成后者？我们回忆Jensen不等式

当函数 g 为下凸函数时有

$$E(g(X))\geq g(E(X))$$
$$or$$
$$\sum_X P(X)g(X)\geq g(\sum_X P(X)X)$$

当函数 g 为上凸函数时有

$$E(g(X))\leq g(E(X))$$
$$or$$
$$\sum_X P(X)g(X)\leq g(\sum_X P(X)X)$$

而 log 为上凸函数，g 用 log 取代，就满足我们的目标

同时我们加上启发式算法的设计思路，也就是迭代，我们当前步骤算出的 θ 只要与当前已有的 $\theta^{(i)}$ 相比似然函数值更大就行

$$
\begin{aligned}
L(\theta) - L(\theta^{(i)}) &= \log\left(\sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})}\right) - \log P(Y|\theta^{(i)}) \\
&\geqslant \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - \log P(Y|\theta^{(i)}) \\
&= \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}
\end{aligned}
$$

这里减去当前似然函数不影响，因为 $\theta^{(i)}$ 是已知，整个似然函数都相当常数

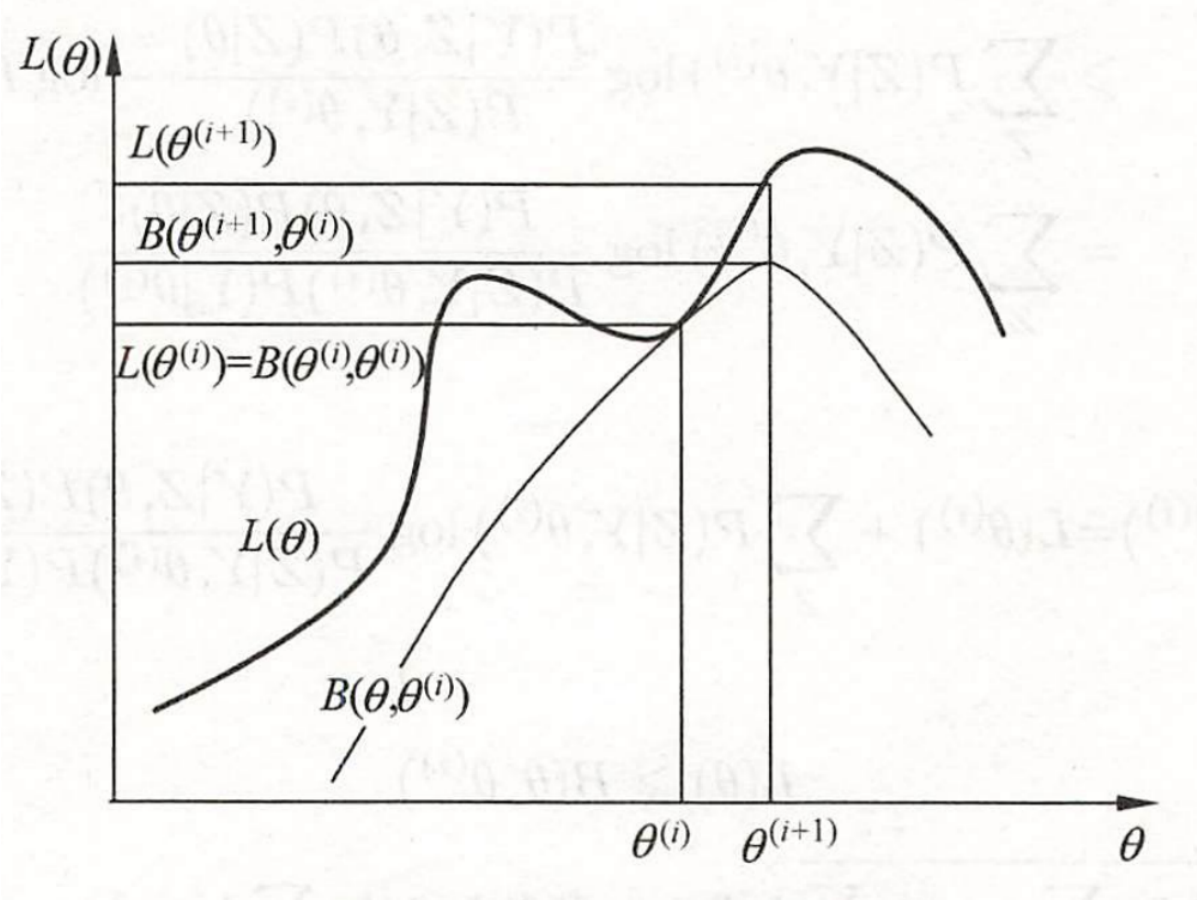
$$
B(\theta,\theta^{(i)}) \hat{=} L(\theta^{(i)}) + \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}
$$

那么

$$
L(\theta) \geqslant B(\theta,\theta^{(i)})
$$

并且等号在 $\theta = \theta^{(i)}$ 处成立

上述运算的逻辑是，我通过Jensen不等式，为比较难求的 $L(\theta|Y)$ 找到了一个比较容易求的下界，我虽然不知道 $L(\theta|Y)$ 的最大值点在哪里，但是我只要找到下界的最大值点，肯定也是比当前的 $\theta^{(i)}$ 要更加优秀的，那也算勉强完成了当前这一步的任务，尽管下界的最大值点可能不是 L 的最大值点



最终得到

$$arg\ \max_{\theta}(\sum_{Z} P(Z|Y,\theta^{(i)})\ log\ P(Y,Z|\theta))$$

所以算法步骤就是

- 初始化 $\theta^{(0)}$
- 计算 $\sum_{Z} P(Z|Y,\theta^{(i)})\ log\ P(Y,Z|\theta)$，求极大值点，循环直至收敛

EM中的E就是计算 $\sum_{Z} P(Z|Y,\theta^{(i)})\ log\ P(Y,Z|\theta)$，因为它的形式是 $log\ P(Y,Z|\theta)$ 在 $P(Z|Y,\theta^{(i)})$ 分布下的期望；M为求其极大值点。并且我们看到，这个算法不能保证找到全局最优，并且算法依赖于初始值

我们也可以注意到，Jensen不等式能够将函数和期望（概率求和）换位置，这个关系在统计学习中应该很重要

6.3. Methods of Evaluating Estimators

6.3.1. 均方误差（**mean squared error**）

$$E_\theta(W-\theta)^2=Var_\theta W+(E_\theta W-\theta)^2$$

这个衡量标准拆开之后很直观，我们建立估计量目的就是只要做一次实验，得到 W 的观测值，它大概就是 θ 值，所以 W 的分布应该是紧密地围绕在 θ 周围，“围绕在 θ 周围”代表第二项接近0，“紧密地”代表第一项接近0

- 在特定情况中会方差和偏差中会存在trade-off关系，即升高偏差会降低方差，使得总体MSE下降，那么当前估计量则虽然不是无偏估计量，但是会比无偏估计量有更小的MSE，如何选择那就要进一步研究。
- 两个估计量的MSE会出现交错情况，比如在样本数 n 较小时，第一个的MSE会大于第二个的MSE，而在较大时反过来
- 第三点涉及同变性原理，暂时没搞懂

最佳无偏估计的逻辑：

首先作者推广了无偏估计量的定义，之前是 $E_\theta(W)=\theta$，将 W 称为 θ 的无偏估计量，这里 $E_\theta(W)=\tau(\theta)$ 虽然不是 θ 的无偏估计量了，但我们将它称作 $\tau(\theta)$ 无偏估计量。在此基础上，我们如果固定一个 $E_\theta(W)=\tau(\theta)$ 那么就会有无数个 W 满足条件，最佳无偏估计量就是在这些 W 中找到 $Var(W)$ 或 MSE 一致最小的那个。这个过程是很复杂的，而 $Cramer-Rao$ 不等式就是确定一个这些 W 能够达到的 Var 下界，那么如果找到一个估计量的 Var 正好等于下界，那么就不用继续找了，当前形式就是 $E_\theta(W)=\tau(\theta)$ 时的最优估计量，如果大于，就知道还有改善的空间。

Chapter 7 Hypothesis Testing

7.1. Introduction

与点估计一致，在假设检验也是通过研究统计量进行的。

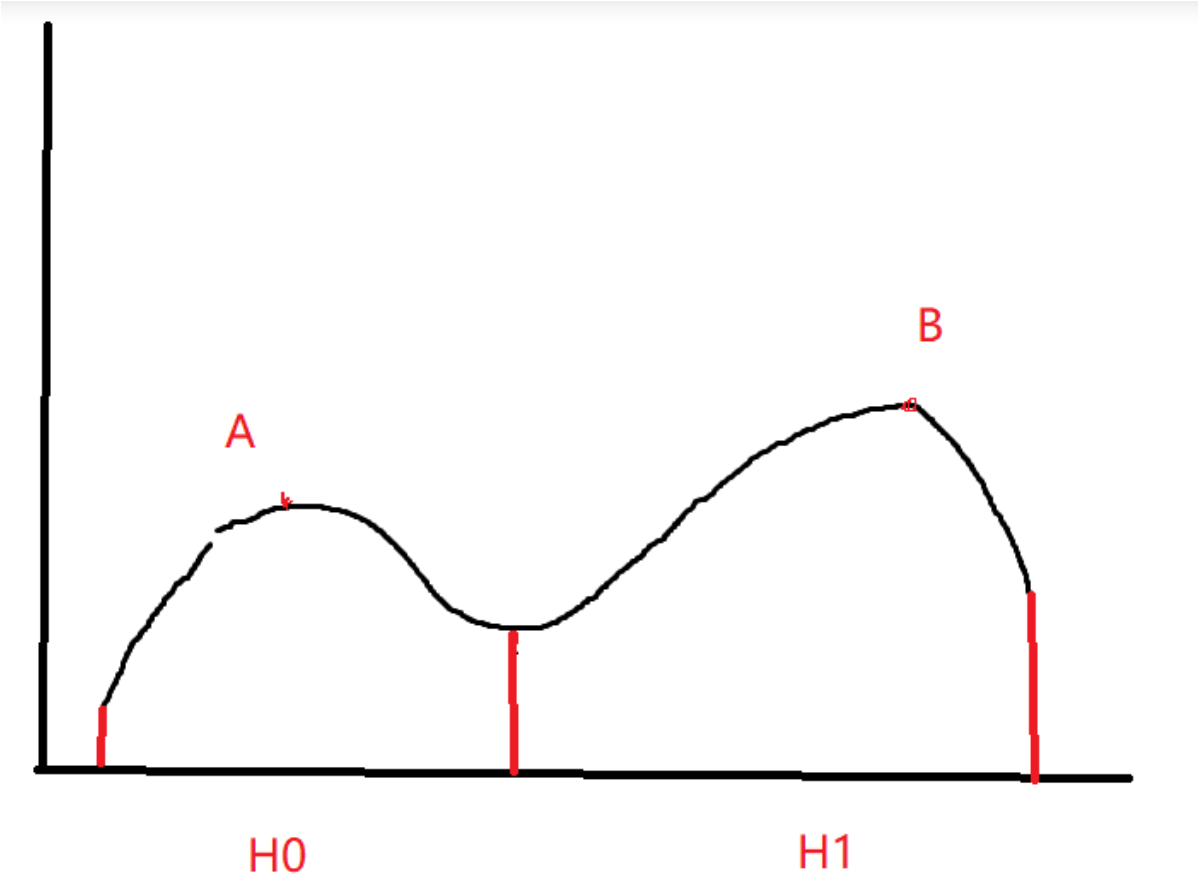
还可以更大胆说，对于一个分布的各种研究，都是通过统计量，提取实验数据中所蕴含的参数信息完成的，这包括点估计，区间估计，假设检验

茆诗松概率论和数理统计教程种，对于假设检验有类型总结

7.2. Methods of Finding Tests

7.2.1. 似然比检验

似然比检验的图像理解：



似然比分子是在 H_0 去搜寻使得函数极大的点，找到了 A，而分子则是在 H_0+H_1 去搜寻点，而这会有两种情况，第一种是找到的仍是 A，就意味着在 H_1 中没有找到比 A 更大的点，可以说我们认为 θ 具体值为 A，又因为 A 在 H_0 中，所以 H_0 为真；第二种是找到了更大的点 B，但这里需要注意并不是只要找到更大的点就抛弃原假设，我们需要 B 比 A 要大过一定程度，也就是比值小过一个人为规定值 c。

似然比检验我不太熟练的地方在于，在似然的时候我们认为 θ 是变量，紧接着在组合生成 $\lambda(x)$ 时我们认为 x 是变量，解出假设拒绝区域

定理8.2.4的直观理解：似然检验比求得拒绝区域的逻辑是，观测到的随机样本具体值 $(x_1,x_2,\ldots,x_n)$ 到底是多少时我能借助从里面提取的 θ 信息去判断 $theta$ 在或者不在 H_0，而从前面我们知道 $(x_1,x_2,\ldots,x_n)$ 中包含的 θ 信息可以精炼成 $T(x)$，我们观察原始样本其实是观察里面的 $T(x)$，那么从数学形式上很合理的推论是我们依据 $(x_1,x_2,\ldots,x_n)$ 造出来的 $\lambda(x)$ 和 $\lambda^*(T(x))$ 处处相等，只有这样我们通过两种思路解出来的拒绝区域才能相等。并且如果最终的两者拒绝区域形式一样，而 $T(x)$ 又能体现样本中 θ 信息，那么很合理说形式应该是 $\phi(T(x))>0$，换句话说：通过 $\lambda(x)$ 解出来的拒绝区域不等式中，和 $(x_1,x_2,\ldots,x_n)$ 有关的形式一定是 θ 的充分统计量 $T(x)$

注意定理8.2.4只在似然比检验中成立。

冗余参数问题：在求解似然比时，也需考虑冗余参数，它能在分布所给的参数定义域中自由取值，最后导致似然比一般是分段函数（例8.2.6）

$$\lambda(x)=\begin{cases}1 & \hat{\mu}\leq\mu_0\\ g(x) & \hat{\mu}>\mu_0\end{cases}$$

分段函数的理解是，当观测数据位于 $\hat{\mu}(x)\leq\mu_0$ 时，肯定不拒绝原假设，只有在 $\hat{\mu}(x)>\mu_0$ 时，才有可能拒绝原假设

7.2.2. 贝叶斯检验

得到后验分布 $\pi(\theta|x)$，得到 H_0 对应的概率，同个这个概率值判断是否接受 H_0

7.2.3. 并-交检验与交-并检验

是基础检验的组合

7.3. Methods of Evaluating Tests

7.3.1. 功效函数

从输入输出来看，我们就是输入 Θ_0，通过一些运算，输出了 R，他们两个之间关系为如果 θ 在 Θ_0 中，观测到的值出现在 R 中的概率很低，所以只要观测到的 x 在 R 中，我们就认为 θ 不在 Θ_0 中，R 也就是 Θ_0 的拒绝区域。

可是就算 H_0 成立，也是有可能观测到 R 的，只不过概率较小而已，这种情况如果真的发生就会导致我们犯第一类错误：把 H_0 误判成假，犯错概率为 $P_\theta(X\in R)\ when\ \theta\in\Theta_0$。我们当然是不想犯错的，因此优化方向为：当 $\theta\in\Theta_0$ 时，$P_\theta(X\in R)$ 越小越好，这里的概率是关于 θ 的函数。

那么第二类错误也就很好理解，θ 滑动到 Θ_0^c，这个时候观测到 $X\in R$ 我们的判定是对的，而只有观测到 $X\in R^c$ 才会误导我们的判断，那么犯错概率为 $P_\theta(X\in R^c)$，为了形式统一，我们仍研究 $P(X\in R)=1-P(X\in R^c)$，我们这时希望它越大越好，它仍是一个关于 θ 的函数。

那么功效函数 $P_\theta(X\in R)$ 的理想图形也就确定了，在 $\theta\in\Theta_0$ 时，函数值全为零，在 $\theta\in\Theta_0^c$ 时，函数值全为1，具体问题中的假设检验方案就要朝这个方向优化。注意：举例 Θ_0 都是 $\theta<somthing$ 的形式，所以书中图像都是左低右高，但是不要形成固定印象，因为完全可能 Θ_0 在右边。

7.3.2. 最大功效检验

逻辑是：首先控制犯第一类错误的概率（水平为 α ），在此基础上去选择犯第二类错误一致最小的检验

控制第一类的关键词：水平，真实水平

小节论述逻辑：

- NP定理，将如何在简答假设（假设都是一个点）中识别 α 水平UMP假设
- 将NP定理与充分统计量结合
- Karlin-Rubin定理，利用充分统计量，MLR，将NP定理推广到复杂假设（假设是区间）

Neyman-Pearson引理：

结合其他参考书，我们对这个引理的陈述做了一点修改

考虑检验 $H_0:\theta=\theta_0$ ，对 $H_1:\theta=\theta_1$ ，其中相对于 θ_i 的概率密度函数或概率质量函数是 $f(x|\theta_i)\ i=0,1$

$$
\begin{gathered}
x \in R \ \ if \ \ f(x|\theta_1) > k\, f(x|\theta_0) \\
and \\
x \in R^c \ \ if \ \ f(x|\theta_1) < k\, f(x|\theta_0) \\
k \geq 0
\end{gathered}
$$

以及

$$
\begin{gathered}
\alpha = P_{\theta_0}(X \in R) \\
0 \ < \ a < \ 1
\end{gathered}
$$

有如下结论：

- （充分性）在水平为 α 的检验中，通过条件一和条件二一定能找到满足条件的检验，并且能够推出
    - 此检验是一个UMP检验
- （必要性）在水平为 α 的检验中，找到一个UMP检验推出
    - 此检验满足条件二
    - 此检验满足条件一

Neyman-Pearson引理的解释：

首先我们有一个形式固定的函数 $f(x|\theta_i)$ 以及一个数 α，它有两个备选的参数，现在我们提要求：希望能够找到一个拒绝区域使得 $\alpha = P_{\theta_0}(X \in R)$。这个区域其实是随便就能找到的，并且能够找到无数个，因为我只要在随机变量的取值空间中找很多点，他们在 θ_0 取值下概率累积为 α 就可以了。而称这些我们随便找的区域为拒绝区域也是没有问题的，因为只要是区域就可以是拒绝区域，效果好不好另说。

在所有水平为 α 的检验中

- 顺方向，似然比检验挑选一个比值 k，作为区域划分的标准，那么这里的作用机制为：固定 α，去搜寻一组 k, R 满足 $\alpha = P_{\theta_0}(X \in R)$，也就是 α, k, R 一一对应，并且总是能找到的。
- 逆方向，在这些检验中一顿乱选，挑选UMP检验

结论一说明，顺到逆，似然比检验就是UMP检验

结论二说明，逆到顺，UMP检验就是似然比检验

两者能够互相转换，原因在于假设都是点（简单假设），而不是区间

定理**8.3.17**（**Karlin-Rubin**）：

- T是 θ 的充分统计量
- T的分布函数有MLR
- $H_0 : \theta \leq \theta_0$，$H_1 : \theta > \theta$

有，$T > t_0$ 拒绝 H_0 是 $\alpha = P_{\theta_0}(T > t_0)$ 水平的UMP检验

证明第一步：$\beta(\theta)$ 是一个非递减函数

证明第二步：对于 $H_0^{'} : \theta = \theta_0$ $H_1^{'} : \theta = \theta^{'} \ where \ \theta^{'} > \theta_0$ 来说，$T > t_0$ 是 α 水平UMP检验

证明第三步（重点）：

- $\beta(\theta)$ 对于原检验也是 α 水平的
- 我们在原假设检验中找，找到的所有 α 水平检验肯定都会满足在 θ_0 点 $\beta^*(\theta_0) \leq \alpha$，那么对于第二步中两个点的假设检验他们也是满足水平为 α 的。
- 如果我们用这种方式找到了一个检验，对于原假设检验，它在 $\theta > \theta_0$ 处存在比 $T > t_0$ 功效大的点，那么我们就可以提供这个检验给第二步的假设检验去用，从而证明 $T > t_0$ 不是第二步假设检验的 α 水平UMP。
- 可是我们已经证明了 $T > t_0$ 是第二步假设检验的 α 水平UMP，也就是我们在原假设检验中根本找不到这样的检验。
- 那么既然我们已经证明了 $\beta(\theta)$ 对原检验也是 α 水平的，并且在 $\theta > \theta_0$ 处原假设检验搜寻到的检验都打不过 $\beta(\theta)$，那么 $\beta(\theta)$ 也是原假设检验的 α 水平UMP

- 我理解的难点在于：区间型假设检验的检验能够直接中间取点，给点型假设检验作为检验，点型检验的 $\beta(\theta)$ 能直接推广给区间型检验用（这不是普遍结论）
- 解释为：不要管区间型还是点型，重点是一个检验本质就是一个拒绝区域，这个区域当然可以同时给区间型和点型假设检验同时使用

例**8.3.19**：

第一步是直接从例8.3.18中的区间型假设检验中直接截取点，说既然在上例中 $\beta(\theta)$ 是UMP，那么这里对于点来说，它也是UMP。这是因为例8.3.18满足Karlin-Rubin定理，那么这个具体例子一定满足 Karlin-Rubin 证明中的第二步，也就是看上去是用区间型检验打点得到点型检验，而实际上是先满足点型检验才能证出例8.3.18中的区间检验，我们只不过是将它中间步骤的小结论拿出来用而已。

而此例中UMP不存在的原因是，我们之前只对检验加了一个限制： α 水平，满足条件的检验称为A，而 $\beta(\theta)$ 在备选假设上一致最大是一件很奢侈的事情，这里的情况是，在A中，我们找到的功效函数总是按下葫芦起了瓢（备选假设的一个子区域上一致最大），这种情况就称为UMP不存在。

解决方案则是再加一个限制：无偏检验，将满足条件的检验集合称为B，它包含的检验数量进一步缩小，在这些检验里面找备选假设上一致最大的，也许找到的功效函数在备选假设上会比A中的某些功效函数低，可是它是B中在备选假设上一致最高的。称之为无偏检验类中的 α 水平UMP

7.3.3. 并交检验与交并检验的真实水平

定理**8.3.21**和定理**8.3.23**的对比：

他们都给组合出来的检验的 α 算出了一个上界，但是8.3.23比8.3.21要好一些，因为定理8.3.21的缺点在于它涉及到的每个检验都是似然比检验，它成立必须依赖这样一个具体形式，并且似然比检验算上界可能比较复杂。

可是8.3.23的缺点是，得到的水平可能太糙了，因为 $R=\underset{\gamma\in\Gamma}{\cap}R_\gamma$ 交出来的 R 完全可能是一个点，那么就通过定理8.3.24算出真实水平

定理8.3.24的逻辑就是在8.2.23的基础上再满足一个小特征，那么子检验中的某个水平 α 就可以变成IUT的真实水平

7.3.4. p-值

假设检验的过程：通过特定方法求得一个检验统计量 $W(X)$，并且设定一个条件，之后做实验得到观测值 x，接着判断是否满足条件，得到结论。

p值和真实水平 α 分属于两派，这也是陈希孺在概率论和数理统计中讲了 α 就没有讲p值的原因，但是两者也是有一定联系的，p值的直观理解为：当前观测到的数据的极端程度。

p值只基于原假设，而和备选假设没有关系

首先我们来看简单假设：

$H_0:\ \theta=\theta_0$

如果原假设成立，那么 $W(X)$ 是一个参数确定的分布，然后观测到 $W(x)$，然后去考察 $W(x)$ 在分布下的极端程度。而这个极端程度不能是 $W(x)$ 的概率，因为单个情况的概率没有办法体现极端程度（极端需要对比），并且连续情况下概率还是0，那么我把极端程度就规定成 $W(x)$ 和比 $W(x)$ 还极端的所有情况的概率之和，那么这就是p值。

要判定这个极端程度是否可以接受，肯定要设一个标准，通常是0.01，0.05等，这个数称为显著性水平。拿0.01举例，如果极端程度比0.01还小，也就是在 H_0 成立时，观测到 x 的概率小到我们不能接受，但是它却发生了，那么我们就拒绝 H_0。

越极端，**p**值越小，越要丢弃原假设

其次看复杂假设：

$H_0:\ \theta\in\Theta_0$

我们可以看到求p值一定需要 $W(X)$ 的分布确定，那么如果 H_0 有一堆值，p值该怎么算呢？Casella给出的解决办法如下：

$$p(x)=\underset{\theta\in\Theta_0}{sup}(W(x)\text{的极端程度})_\theta$$

也就是找到极端程度的上确界，这个方法蕴含着对原假设宽容的倾向。上面我们看到p值越小就越要丢掉原假设，参数区间的p值是每一个参数p值的比较出来的最大值，就意味着不管各个p值有多小，只要存在一个参数点的p值大于0.01，我就认为原假设成立。

这个思想和假设检验的假设设立原则是一致的，原假设一般是没有充分理由不能否定的情况，那么我们肯定也就对原假设比较宽容了。

还有一个问题，我们如何判定哪些取值比 x 更加极端？这个需要具体问题具体分析，比如当原假设是分布均值 $\mu=\mu_0$，而得到的统计量是 $\overline{X}$，那么它的分布均值也是 μ_0，那么极端就代表远离均值，也就是从 $\overline{x}$ 往远离 μ_0 的方向走都是更极端的值。

7.3.5. p值和真实水平 α 的关系

我们可以借助p值来判断我们得到的实验数据是否落在拒绝区域中

我们需要明确，p值研究中离谱的那些数据，和真实水平 α 研究的拒绝域数据，当中一个肯定包含另一个，这是因为两者本质都是检验统计量 $W(X)$ 分布中那些不太可能发生观测到的数据（低概率数据）

在简单假设中，p值为 $W(x)$ 以及比 $W(x)$ 更极端数据的概率和，那么如果我们现在设置 $\alpha < p$ ，也就是拒绝区域的概率比极端数据的概率和要小，那么也就是拒绝区域比极端数据区域要小，后者将前者包住，而极端数据区域边界是 $W(x)$ ，那么 $W(x)$ 对应的 x 没有落在拒绝区域内，那么在当前 x 被观测到时就不能拒绝原假设。

同理，如果 $\alpha \geq p$ ，那么此时有拒绝区域包住了极端数据，那么在当前处于极端数据边界 x 被观测到的情况下，肯定能拒绝 H_0 。换句话说，p是能拒绝 H_0 的最小 α 。

在复杂假设中，p值定义是极端数据和的概率上确界，而真实水平 α 定义是 H_0 拒绝区域的概率上确界，那么**p**值是能拒绝 H_0 的最小 α 还成立吗？这种情况暂时想不清楚，但是应该是成立的。

所以我们看 α 从大到小，我们会经历拒绝 H_0 到 接受 H_0 ，以p值作为转变点，并且他们的关系和前面的显著性水平是一样的：当显著性水平给定，p小于它拒绝 H_0，当 α 给定，p小于它拒绝 H_0 ，所以在 α 名字也经常加上“显著性”三字混着用。

7.4. Exercises

8.5

$$T = log[\frac{\prod_{i=1}^{n} X_i}{(min\ X_i)^n}]$$

我们之前求分布函数的时候，随机变量间的函数关系只有两种，第一种是具有显式函数关系，解法就是直接求，第二种就是类似次序统计量，那就用计数原理做。

但是这里特殊的是，既有原随机变量 X_i 的组合，又有非显式函数关系导出的随机变量 $min\ X_i$ ，两者混在一起。它的思路需要从最终看如何导致这样的结果，n 个iid随机变量同时产生当然可以得到 T ，但是还可以先产生一个 $X_{(1)}$ ，然后在这个条件下产生其他的 $(n-1)$ 个 X_i 。

这个题巧合的是，我们算出的 $X_{(1)}$ 条件下的 T 的分布不包含 $X_{(1)}$ ，也就是 T 和 $X_{(1)}$ 独立，那么此分布也就是 T 的无条件分布

Chapter 8 Interval Estimation

8.1. Introduction

类似点估计，对于观测样本，我们设计两个函数：$L(x)$ 和 $U(x)$ ，做出推断 $L(x) \leq \theta \leq U(x)$ ，并把 $[L(X), U(X)]$ 称为区间估计量。

它的逻辑是： $L(X)$ 和 $U(X)$ 是随机变量，那么他们产生的区间也是随机变量，也就是做很多次实验会产生很多个区间的具体值，这些具体区间有很多都能包住 θ 。所以要衡量区间估计量的效果，只要算出有多大的概率 $[L(X), U(X)]$ 能包住 θ 就行。这样阐述是因为只要不涉及贝叶斯，θ 就是一个未知的值，它不可能是随机变量，所以概率分布只可能 X 以及 X 的函数。

包住的概率和置信系数之间并不严格相等，当算出的概率不含有 θ 时，很显然概率就是一个常数，那么此概率就是置信系数；而当算出的概率包含 θ 时，那么此概率的最小值是置信系数。

但是这个解释只是对最终得到的 $[L(X), U(X)]$ 合理解释，它暗含的是 θ 是固定值，我通过 x 的变化去适应它。可在推导当中会出现 x 固定，变化 θ 的情况。

第七章Introduction中写的变量与值的关系在这里得到了重要体现：

只要是字母就可以动，体现在于取值变化或是积分微分

随机变量总共有三层：

- 随机变量 X ：它是一个变量，它具有概率分布，方差等的性质，可动
- 取值 x ：也就是高中学的因变量，因此带入具体值（或称为固定 x ，向下趋势），求导积分等（向上趋势）都能执行
- 数：**1**，**2**，**3...**

8.2. Methods of Finding Interval Estimators

8.2.1. 反转检验统计量

在假设检验中，$H_0 : \mu = \mu_0$ ， μ_0 是固定值，为向下的属性，得到接受区域为 $A(\mu_0) < x < B(\mu_0)$（在具体实验中，我们只要将比如 $\mu_0 = 1$ ，就能在实际情况中进行检验）

紧接着对于同意式子，我们固定 x ，而认为 μ_0 是变化的，得到区间 $C(x) < \mu_0 < D(x)$

最后由于 μ_0 仅是一个字母（代号），没有被带入数字，也就是是任何数字都可以，得到区间 $C(x) < \mu < D(x)$

其中的关键就是动静属性的转变

8.2.2. 枢轴量（**Pivotal Quantities**）

来自样本的随机变量函数 $Q(X_1, \ldots, X_n, \theta)$ 的分布独立于所有参数，也就是它的分布函数里没有参数 θ，那么这个随机变量称为枢轴量

- 充分统计量：样本分布在充分统计量条件下，和参数无关
- 枢轴量：枢轴量本身分布和参数无关

原理：在分布函数中，参数会给随机变量的分布带来某种影响，我只要在样本函数中消除这种影响就能得到一个分布不含参数的随机变量，比如总体分布 $f(x-\mu)$，那么很明显 μ 影响是 X 的平均水平增大 μ，那么我要消除这种影响，只需将平均水平向减小 $\overline{X}-\mu$，这个随机变量分布就和 μ 没有关系了。而没有了参数，分布就是某种标准分布，求置信系数变得方便。形式上，我们构造的枢轴量总是含有参数，这是因为参数的效果肯定需要参数去消除。

疑问：在获取枢轴量后，为什么区间一定是 $P_\theta(a \ \leq Q(X,\theta) \leq b) \geq 1-\alpha$ 这种夹中间的形式？
回答：我们需要 θ 的上下界，也就是 θ 形式需要夹中间，只有初始夹中间，推到后面才能夹中间

8.2.3. 枢轴化累积分布函数

累积分布函数天然就是一个枢轴量

这里涉及到了随机递增（递减） 的概念，它的特殊之处在于，对于每个 t，$F(t|\theta)$ 是一个关于 θ 的递减函数，则一族分布函数 $F(t|\theta)$ 关于 θ 随机递增

8.2.4. Bayes区间

Chapter 9 Asymptotic Evaluations

还是点估计，假设检验和区间估计，只不过此时样本数 n 趋近于无穷大，会导致一些新的现象，简化计算

Chapter 10 Analysis of Variance and Regression

10.1. 11.1 Introduction

陈希孺书中的分析总共三种：回归分析，相关分析，方差分析，这里只阐述了两种

从第七章到第九章，不管怎么变化，研究的问题只有一个：有一个形式已知，参数未知并且我们能够观测的随机变量分布，我想要收集实验数据去提取参数的信息，我应该怎么办？我们在 X 的基础上，对随机变量纵向进行提炼，设计各种新的随机变量 $Y=g(X)$，尽管我们能够观测的还是只有 X 这一个量，但是我们可以算出 Y，最终得到相关结果。我们可以看出 X 和 Y 之间是一个确定的依赖关系，也就是 x 被观测到，相应的 Y 只可能是 $g(x)$，条件概率是 $P(Y=g(x)|x)=1$ 而 $P(Y\neq g(x)|x)=0$

而分析研究的重点不再是单个分布，而是实验中可以观测或控制的多个随机变量。X 和 Y 之间是不确定关系，比如线性回归中的 $Y=\alpha+\beta x+\epsilon$（其中 ϵ 是随机变量），在 x 被观测到的情况下，Y 仍是一个分布。条件概率是 $P(Y=y|x)$。任何一本统计书中引入的联合分布和条件分布讨论的都是这种不确定关系。（注意大多数时候，两个随机变量之间的关系不太可能用线性函数表示，线性回归是一种简化情况）

从 Y 的角度来看，条件概率中不管是 x 还是真正参数 θ 都是影响分布的因素，那么 x 和 θ 对于 Y 来说都是参数，所以将条件概率写成 $P(Y=y|x,\theta)$ 也是没问题的。

从图像的角度，前面几章的工作为针对一个可观测随机变量的纵向提炼，而这一章的工作为针对多个可观测随机变量的横向扩展。

10.2. ANOVA

方差分析研究的是一个因子对于另一个随机变量的影响，这里用因子（factor），而不是随机变量称呼。

并且从这里来看，假设检验是一个很广义的话题，并不局限于前面第八章的单可观测随机变量的参数假设。只要能基于实验给我一个量以及一个衡量标准，那么这能够是一个假设检验。

10.3. Regression

回归分析研究是多个可观测的随机变量，他们之间具有带参的函数关系，我们研究重点不再是每一个随机变量他们本身的分布，而是他们之间的函数关系 $Y = f(X_1, \ldots, X_n) + \epsilon$ 。它的现实意义在于，很多时候我们并不在意分布函数，而是变量之间的关系，比如研究青少年身高体重关系，最终目的在于给我一个身高，我就能算出一个正常体重（范围），而青少年全体人口的身高分布，体重分布在这个问题中是次要问题。（关键词：变量关系）

机器学习中的回归和统计中的回归是同一个东西