非参数统计

王 成 [1]
http://math.sjtu.edu.cn/faculty/chengwang/
上海交通大学数学系

2016 – 2017 第一学期

[1]讲义中的任何错误或建议请联系chengwang@sjtu.edu.cn. Update time: January 10, 2017

Contents

Chapter 1

课程介绍及预备知识

1.1 课程说明

- 作业用Latex和R完成，可以使用两者的结合体R Markdown, Sweave, Knitr等，作业通过邮件提交 chengwang@sjtu.edu.cn。
- 提交的作业拒绝抄袭，多数作业都是开放式问题，没有统一答案，提交的作业可能会进行交叉比对。
- 课堂积极参与讨论
- 最后成绩： 平时作业+课堂参与+最后考试

1.2 背景介绍

在一个统计问题中，假定总体分布的数学形式已知，仅包括(少数)有限个未知参数，这个问题就是参数统计问题，否则就是非参数性质的。

统计不是数学，不用完全划清参数和非参数之间的界限。非参数方法当然也可以用到参数问题中去，例如我们可以用非参数方法去检验正态分布的均值，反过来非参数方法很多时候本质上又回到参数问题，例如经典的符号检验本质还是把任意分布转化成了二项分布。

几个非参数统计相关的名词：Distribution-free， Nonparametric， Semi-parametric等

假定两个变量(X, Y)之间有如下的回归关系：

$$Y = f(X) + \epsilon,$$

大家知道，统计里关心的就是 $f(x)$，也就是我们要估计$f(x)$。从数学角度，我们知道$f(X)$的形式有无穷多种，线性，多项式，三角函数等等。我们从线性代数的基的角度来看：

- $f(x) = a$， 那么对应的基为 $\{1\}$, 可以认为是1维的。估计$f(x)$也就是在一维空间里面找一个数值；
- $f(x) = a + bx$， 那么对应的基为 $\{1, x\}$, 可以认为是2维的。估计$f(x)$也就是在二维空间里面找一个数值。例如我们的最小二乘不就是在二维平面上找一个点嘛?
- $f(X) = a+bx+cx^2$， 那么对应的基为 $\{1, x, x^2\}$, 可以认为是3维的，
- $f(X) = a+a_1x+\cdots+a_nx^n$， 那么对应的基为 $\{1, x, \ldots, x^n\}$, 可以认为是$(n+1)$维的。
- $f(x) = acos(bx)$???

由Taylor展开，对于一大类性质良好的函数，我们都可以有 $f(X) = a_0 + a_1x + \cdots + a_nx^n + \ldots$， 估计$f(x)$也就是找一列 $a_0, a_1, \ldots$ 这个角度可以把非参数看成无穷维的参数问题，这也是Larry Wasserman书中的观点。

1.3 课程内容

经典非参数内容：

1. 次序统计量
   把一组样本$X_1, \ldots, X_n$按照大小排序后就得到次序统计量$X_{(1)}, \ldots, X_{(n)}$.
   这一部分主要介绍次序统计量的一些相关的分布，渐进分布以及一些统计问题上的应用。

2. U统计量
   U统计量是Hoeffding 1948年引入的一类统计量，在非参数的估计和检验问题中有大量的应用。这一块会介绍这类统计量的定义，分布以及理论渐进分布等。这一块内容最近几年在高维数据的统计分析，尤其是假设检验里有大量的应用，我们会介绍一些相关的工作。

3. 秩统计量
   各个样本在其大小排序中所占的位次，这其中也有绝对秩和符合秩之分等等。这一部分要介绍在独立和不同分布结构情况下的秩统计量的分布等，介绍秩统计量在各种统计问题，估计问题，独立性检验等等

4. 置换检验 稳健估计等

现代非参数部分：

1. 密度函数估计 给定一组iid样本 $X_1, \ldots, X_n$, 估计分布函数 $F(x) = P(X_1 \leq x)$ 和密度函数 $f(x) = F'(x)$.
2. 非参数回归 给定样本 (X_1, Y_1), $\ldots, (X_n, Y_n)$, 估计回归函数$r(x) = E(Y|X = x)$.
3. 其他非参数方法如决策树、交叉验证、 Bootstrap、多元正态分布转换等。

1.4 预备知识

Probability Space $(\Omega, \mathcal{A}, \mathbb{P})$:

- Ω all the possible results of an experiment;
- $\mathcal{A}$ the σ-field based on Ω:
  1. $\emptyset \in \mathcal{A}$;
  2. if $A \in \mathcal{A}$, $A^c \in \mathcal{A}$;
  3. $A_1, A_2 \cdots \in \mathcal{A}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.
- A probability measure $\mathbb{P}$ is defined on σ-filed $\mathcal{A}$ such that:
  1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$;
  2. $\mathbb{P}(\Omega) = 1$;
  3. if $A_1, A_2, \ldots$ are disjoint then

$$\mathbb{P}(\overset{\infty}{\underset{i=1}{\cup}} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a **probability space**.

Random Variable: a random variable is a map $X : \Omega \to \mathbb{R}$ such that , for every real x,

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}.$$

Convergence of Variables: A sequence of random variables X_n converges to a random variable X,

- *in distribution*:

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x), \tag{1.1}$$

  for all points x at which $F(x) = \mathbb{P}(X \leq x)$ is continuous. $X_n \rightsquigarrow X$.

- *in probability*: for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \epsilon) \to 0, \text{ as} n \to \infty. \tag{1.2}$$

  Written as $X_n \xrightarrow{p} X$.

- *almost surely*:

$$\mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \to 0\}) = 1. \tag{1.3}$$

  Denote as $X_n \xrightarrow{a.s.} X$. Noting

$$\{\omega : |X_n(\omega) - X(\omega)| \to 0\} = \cap_{k=1}^{\infty} \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} \{\omega : |X_m(\omega) - X(\omega)| < \frac{1}{k}\}$$

Theorem 1.4.1 (Slutsky's Theorem)**.** *If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for some constant c, then*

- $X_n + Y_n \rightsquigarrow X$;
- $X_n Y_n \rightsquigarrow cX$;
- $X_n / Y_n \rightsquigarrow X/c, \ c \neq 0$.

Theorem 1.4.2 (Continuous Mapping Theorem)**.** *Let g be a continuous function, then*

$$\begin{aligned} &X_n \rightsquigarrow X \Rightarrow g(X_n) \rightsquigarrow g(X); \\ &X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X); \\ &X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X). \end{aligned}$$

Theorem 1.4.3 (Delta Method)**.** 如果 $g(\cdot)$ 在 μ 点可导且$g'(\mu) \neq 0$, 那么

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2) \Rightarrow \sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, (g'(\mu))^2 \sigma^2) \tag{1.4}$$

1.5 参考书目

1. 陈希儒 等. 非参数统计, 中国科学技术出版社， 2012.

2. 孙山泽. 非参数统计讲义, 北京大学出版社， 2000.

3. Gibbons, Jean Dickinson, and Subhabrata Chakraborti. Nonparametric statistical inference. Springer Berlin Heidelberg, 2011. (侧重于经典非参数统计)

4. Wasserman, Larry. All of Nonparametric Statistics. Springer Science & Business Media, 2006. （侧重于现代非参数尤其大样本性质）

5. Fan, Jianqing, and Qiwei Yao. Nonlinear time series: nonparametric and parametric methods. Springer Science & Business Media, 2003.

所有课件会放在 `http://math.sjtu.edu.cn/faculty/chengwang/teach.html`

Chapter 2

次序统计量

2.1 学习目标

1. 了解次序统计量的定义以及极大、极小、中位数、极差等几个重要的次序统计量，
2. 学会计算次序统计量的分布及次序统计量线性函数分布的计算，
3. 学习次序统计量的大样本性质，包括极值分布，中位数渐近分布等，
4. 用次序统计量构造一些统计应用并解释原理。

2.2 背景介绍

2.2.1 定义

给定一组样本 $X_1, \ldots, X_n$, 将其按照大小排序得到的新样本$X_{(1)} \leq \cdots \leq X_{(n)}$ 称为样本 $X_1, \ldots, X_n$的次序统计量。

思考： 次序统计量是随机变量嘛? 给定一组样本，如何计算某一个次序统计量 $X_{(k)}$? 了解数据排序算法。

2.2.2 重要的次序统计量

极大： $X_{(n)} = max(X_1, \ldots, X_n)$　　极小： $X_{(n)} = min(X_1, \ldots, X_n)$,
中位数： n为奇数时$X_{((n+1)/2)}$， 偶数时： $(X_{(n/2)} + X_{(n/2+1)})/2$,

极差： $X_{(n)} - X_{(1)}$.

其中，极值，以及奇数情形下的中位数是单个的次序统计量，而偶数情形的中位数和极差是组合形式的次序统计量。

2.2.3 常见应用

次序统计量在社会生活中经常以各种不同的形式出现，例如"百年不遇的....." "前十大....." "世界首富"等等。下面我们以一个NBA球队工资为例来看下次序统计量的应用：

$2297, 1969, 1641, 964, 899, 677, 500, 495, 210, 150, 150, 128, 118, 115, 98, 98, 84.5, 84.5.$

样本平均： 593. 极大：2297; 极小：84.5; 中位数： 180 极差： 2212.5.

2.3 基本分布

假定$X_1, \ldots, X_n$是从一个分布为F的总体中抽取的iid样本，本节我们考虑次序统计量的相关分布。

2.3.1 $X_{(k)}$ 的分布

记$X_{(k)}$的分布函数为F_k, 我们有

$$\begin{aligned} F_k(x) = P(X_{(k)} \le x) &= P(X_1, \ldots, X_n \text{中至少有k个不大于x}) \\ &= \sum_{r=k}^{n} P(X_1, \ldots, X_n \text{中r个不大于x,n-r个大于x}) \\ &= \sum_{r=k}^{n} C_n^r (P(X_1 \le x))^r (P(X_1 > x))^{n-r} \\ &= \sum_{r=k}^{n} C_n^r F^r(x)(1 - F(x))^{n-r}. \end{aligned}$$

利用恒等式：

$$\sum_{j=r}^{n} C_n^j p^j (1-p)^{n-j} = \frac{n!}{(r-1)!(n-r)!} \int_0^p t^{r-1}(1-t)^{n-r} dt, \ r = 1, \ldots, n, \ 0 \le p \le 1,$$

我们有

$$F_k(x) = P(X_{(k)} \le x) = \frac{n!}{(k-1)!(n-k)!}\int_0^{F(x)} t^{r-1}(1-t)^{n-r}dt.$$

如果$F(x)$的密度函数存在，记为$f(x)$，对应的我们有$X_{(k)}$的密度函数：

$$f_k(x) = F_k'(x) = \frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}(1-F(x))^{n-k}f(x). \tag{2.1}$$

特别的，对于两个极值，我们有

$$X_{(1)} : F_1(x) = 1-(1-F(x))^n,\ f_1(x) = n(1-F(x))^{n-1}f(x);$$
$$X_{(n)} : F_n(x) = F^n(x),\ f_n(x) = nF^{n-1}(x)f(x).$$

Remark 2.3.1. 如果X_1的期望存在，即$\int |x|dF(x) < \infty$, 那么$X_{(k)}$的期望也一定存在。可以从其分布函数出发推导，也可以直接由下面的不等式得到：

$$|X_{(k)}| \le |X_1| + \cdots + |X_n|.$$

当然，我们也可以像一般的随机变量一样，研究次序统计量的期望，方差，特征函数，生成函数等等，稍后我们用均匀分布的例子来看看次序统计量的统计特征。

2.3.2 联合分布

对于任意的$k < j$, 我们研究 $(X_{(k)}, X_{(j)})$的联合分布。对于任意的$x \le y$,

$$\begin{aligned}
F_{kj}(x,y) &= P(X_{(k)} \le x, X_{(j)} \le y)\\
&= P(X_1,\ldots,X_n\text{中至少有k个不大于x,j个不大于y})\\
&= \sum_{r=j}^{n} P(X_1,\ldots,X_n\text{中r个不大于y,n-r个大于y,至少k个不大于x})\\
&= \sum_{r=j}^{n}\sum_{i=k}^{r} P(X_1,\ldots,X_n\text{中i个不大于x,r-i个大于x不大于y，n-r个大于y})\\
&= \sum_{r=j}^{n}\sum_{i=k}^{r} C_n^r C_r^i F(x)^i(F(y)-F(x))^{r-i}(1-F(y))^{n-r}\\
&= \sum_{r=j}^{n}\sum_{i=k}^{r} \frac{n!}{i!(r-i)!(n-r)!}F(x)^i(F(y)-F(x))^{r-i}(1-F(y))^{n-r}.
\end{aligned}$$

如果$F(x)$的密度函数$f(x)$存在，我们有$(X_{(k)}, X_{(j)})$密度函数：

$$f_{kj}(x,y) = \frac{n!}{(k-1)!(j-k-1)!(n-j)!}F(x)^{k-1}(F(y)-F(x))^{j-k-1}(1-F(y))^{n-j}f(x)f(y).$$

对于$x > y$的情形， $P(X_{(k)} \le x, X_{(j)} \le y) = P(X_{(j)} \le y)$ 退化到单个的情形，对应的密度函数为0.

Remark 2.3.2. 从这里我们可以看出，即使对于iid的$X_1, \ldots, X_n$, $X_{(k)}$与$X_{(j)}$也是不独立的.

Example 2.3.1. 对于极值$(X_{(1)}, X_{(n)})$的联合分布，我们有

$$F_{1n}(x,y) = \sum_{i=1}^{n} \frac{n!}{i!(n-i)!}F(x)^i(F(y)-F(x))^{n-i} = F^n(y) - (F(y)-F(x))^n.$$

如果$F(x)$的密度函数$f(x)$存在，我们有$(X_{(1)}, X_{(n)})$密度函数：

$$f_{1n}(x,y) = n(n-1)(F(y)-F(x))^{n-2}f(x)f(y).$$

Remark 2.3.3. 全体次序统计量$(X_{(1)}, \ldots, X_{(n)})$的联合密度函数为：

$$f_{1\ldots n}(x_1, \ldots, x_n) = n!f(x_1)\ldots f(x_n),\ x_1 \le x_2 \le \cdots \le x_n. \tag{2.2}$$

上述提及的所有单个的或者联合的本质上都可以通过对全体的联合密度函数积分得到。

2.3.3 极差和样本分位数的分布

在介绍这部分内容之前，我们先回顾一下随机变量的变换。

对于一个随机变量X, 分布函数为$F(x)$, 密度函数$f(x)$, 我们考虑随机变量$g(X)$. 从微积分的角度，$g(X)$的分布函数为：

$$H(x) = P(g(X) \le x) = \int_{g(t)\le x} f(t)dt = \int_{g_1(x)}^{g_2(x)} f(t)dt = F(g_2(x)) - F(g_1(x)),$$

对应的密度函数应该为： $h(x) = f(g_2(x))g_2'(x) - f(g_1(x))g_1'(x)$. 这里要求$g(x)$函数有很好的光滑性，例如连续等，其中

$$\{t:\ g(t) \le x\} = \{t:\ g_1(x) \le t \le g_2(x)\}.$$

特别的如果$g(x)$是一个严格增函数，那么$g_2(x) = g^{-1}(x)$, $g_1(x) = -\infty$和

$$H(x) = F(g^{-1}(x)),\ h(x) = f(g^{-1}(x))/g'(x). \tag{2.3}$$

Example 2.3.2. 例如，$g(x) = x^2$,即我们取随机变量的平方，那么我们有X^2的分布函数和密度函数分别为：

$$H(x) = F(\sqrt{x}) - F(-\sqrt{x}),\ h(x) = \frac{1}{2}x^{-1/2}(f(\sqrt{x}) + f(-\sqrt{x})).$$

另一个常见变换是$g(x) = \log x$, 对应的我们有$g_1(x) = -\infty,\ g_2(x) = e^x$,

$$H(x) = F(e^x),\ h(x) = f(e^x)e^x.$$

反过来我们去验证所得的密度函数时候，又会用到变量代换把整个积分倒回去。本质上一元的函数型随机变量对应积分里的变量代换。

现在我们考虑二元的情形(X, Y)，密度函数为$F(x, y)$,密度函数$f(x, y)$, 我们考虑变换后的一元随机变量$g(X, Y)$ (例如$X + Y, X - Y$等)。我们先看一个简单问题，如何求$(X - Y, X + Y)$的联合分布?

$$P(X - Y \le u, X + Y \le v) = \int_{x-y\le u, x+y\le v} f(x, y)dxdy = \int_{-\infty}^{(u+v)/2} (\int_{x-u}^{v-x} f(x, y)dy)dx$$

或者在积分过程中，我们做变量代换，$a = x - y, b = x + y$, 那么

$$\int_{x-y\le u, x+y\le v} f(x, y)dxdy = \int_{a\le u, b\le v} f((a + b)/2, (b - a)/2)/2dadb,$$

即$(X - Y, X + Y)$的密度函数为$f((u + v)/2, (v - u)/2)/2$, 其中$1/2$的出现时变换$(x, y) \Rightarrow (u = x - y, v = x + y)$的变换Jacobi，$dxdy = 1/2dudv$.

对于得到的联合概率密度函数，单个积分即可以得到每一个的密度函数，例如

$$X - Y : \int_{-\infty}^{\infty} f((u + v)/2, (v - u)/2)/2dv,$$
$$X + Y : \int_{-\infty}^{\infty} f((u + v)/2, (v - u)/2)/2du.$$

现在我们可以尝试推导出极差的分布，由前面的结果，我们有$(X_{(1)}, X_{(n)})$的联合分布密度为：

$$f_{1n}(x, y) = n(n - 1)(F(y) - F(x))^{n-2}f(x)f(y)I(y \ge x).$$

类似的, 我们有$X_{(n)} - X_{(1)}$的密度函数为：

$$\int_{-\infty}^{\infty} f_{1n}((v-u)/2,(v+u)/2)/2dv,$$
$$= n(n-1)\int_{-\infty}^{\infty}(F((v+u)/2) - F((v-u)/2))^{n-2}f((v-u)/2)f((v+u)/2)I(u \geq 0)/2dv$$
$$= n(n-1)I(u \geq 0)\int_{-\infty}^{\infty}(F(x+u) - F(x))^{n-2}f(x+u)f(x)dx \quad (x = (v-u)/2).$$

类似的我们也可以考虑偶数情形下的中位数的分布，这里我们考虑更一般的p分位数。

Definition 2.3.1 (p 分位数). 以$[\alpha]$记不超过α的最大的整数。对于任意的$0 < p < 1$, 称

$$\epsilon_{np} = X_{([np])} + (n+1)(p - \frac{[np]}{n+1})(X_{([np]+1)} - X_{[np]}), \tag{2.4}$$

为样本$(X_1, \ldots, X_n)$的p分位数。按照此定义$1/(n+1), \ldots, n/(n+1)$的分位数正好是次序统计量$X_{(1)}, \ldots, X_{(n)}$. 之前定义的样本中位数不论对于奇数还是偶数也符合此处的定义。

注意到我们之前已经求出任意的两个次序统计量的联合分布，而p分位数具有$aX + bY$的形式，所以可以类似极差的方法写出其分布。

2.3.4 条件分布

感兴趣的自己参看相关文献。

2.4 均匀分布$U(0,1)$的情形

对于任意的随机变量X,记其分布函数为F,我们首先研究随机变量$Y := F(X)$的分布,对于任意的$0 \leq y \leq 1$,

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y. \tag{2.5}$$

如果$F(x)$是严格增函数，上述推导没有问题。但是我们知道并不是所有分布函数都是严格增函数的，例如离散分布等。

Example 2.4.1. X是*Bernoulli*分布，我们有

$$F(x)=\begin{cases}0 & if\, x<0;\\ \frac{1}{2} & if\, 0\le x<1;\\ 1 & if\, x\ge 1.\end{cases}$$

而$Y=F(X)$的分布为一个$(\frac{1}{2},1)$的两点分布。

Theorem 2.4.1. 设随机变量X的分布函数F处处连续，则

$$Y:=F(X)\sim U(0,1). \tag{2.6}$$

Proof: 记 $F^{-1}(y)=\sup\{x:F(x)\le y\}$. 对于 $0<y<1$,

$$\begin{aligned}P(Y\le y)=P(F(X)\le y)&=P(X\in\{x:F(x)\le y\})\\&=P(X\le F^{-1}(y))\\&=F(F^{-1}(y))\\&=y,\ \text{这里用到}F(x)\text{的连续性质}\end{aligned}$$

Theorem 2.4.2 (生成任意连续分布)**.** 设分布函数F处处连续，则随机变量

$$X:=F^{-1}(R)\sim F, \tag{2.7}$$

其中$R\sim U(0,1)$.

由此定理，对于处处连续的总体分布函数，研究次序统计量 $X_{(1)},\dots,X_{(n)}$，可以等价的研究

$$(F(X_{(1)}),\dots,F(X_{(n)}))\overset{d}{=}(U_{(1)},\dots,U_{(n)}), \tag{2.8}$$

这里，$U_{(1)},\dots,U_{(n)}$是来自$U(0,1)$的iid样本$U_1,\dots,U_n$的次序统计量。基于之前的结果，我们可以写出$U_{(1)},\dots,U_{(n)}$的相关分布如下：

1. $U(k)$的分布函数：

$$F_k(x)=P(U_{(k)}\le x)=\frac{n!}{(k-1)!(n-k)!}\int_0^x t^{k-1}(1-t)^{n-k}dt,\ 0\le x\le 1.$$

密度函数：

$$f_k(x)=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}I_{(0,1)}(x). \tag{2.9}$$

2. 联合分布函数

$$F_{kj}(x,y)=\sum_{r=j}^{n}\sum_{i=k}^{r}\frac{n!}{i!(r-i)!(n-r)!}x^i(y-x)^{r-i}(1-y)^{n-r}.$$

联合密度函数：

$$f_{kj}(x,y)=\frac{n!}{(k-1)!(j-k-1)!(n-j)!}x^{k-1}(y-x)^{j-k-1}(1-y)^{n-j},\ 0<x<y<1.$$

3. 全体次序统计量的密度函数

$$f_{1\cdots n}(x_1,\ldots,x_n)=n!,\ 0<x_1\le x_2\le\cdots\le x_n<1. \tag{2.10}$$

本节我们主要研究次序统计量的大样本性质，即当样本个数$n\to\infty$时候，次序统计量的极限分布等。为了强调与n的关系，我们记样本$X_1,\ldots,X_n$的次序统计量为

$$X_{n1},X_{n2},\ldots,X_{nn}.$$

对于每个自然数n,选取一个正整数k_n, k_n/n称为次序统计量X_{nk_n}的秩. 若k_n/n在$n\to\infty$时候的极限存在，记为$\lambda\in(0,1)$, 则我们称序列

$$X_{1k_1},\ldots,X_{nk_n},\ldots \tag{2.11}$$

的极限秩为λ. 上述的次序统计量序列的极限分布与λ有关，我们可以按照λ的大小分为以下三种情况：

- $0<\lambda<1$, 2.11 称为 **中心项序列** (central term); 这里特别的当$\sqrt{n}(k_n/n-\lambda)\to 0$, 我们称为正则中心项序列；
- $\lambda=0$ 但$k_n\to\infty$ 或者对称的 $\lambda=1$ 但$n-k_n\to\infty$, 称为**中间项序列** (intermediate term);
- k_n 有界或者$n-k_n$有界，称为 **边项序列**(extreme term)

2.4.1 $U(0,1)$的结果

Theorem 2.4.3 (4.1). 记$U_{(1)},\ldots,U_{(n)}$是来自$U(0,1)$的iid样本$U_1,\ldots,U_n$的次序统计量. $Z_1,\ldots,Z_n$为iid的负指数分布,密度函数为$e^{-x}I(x>0)$. 定义

$$Y_k=\sum_{i=1}^{n+1-k}\frac{Z_i}{n+1-i}, \tag{2.12}$$

则我们有

$$(-\log U_{(1)},\ldots,-\log U_{(n)})\stackrel{d}{=}(Y_1,\ldots,Y_n). \tag{2.13}$$

或

$$(U_{(1)},\ldots,U_{(n)})\stackrel{d}{=}(e^{-Y_1},\ldots,e^{-Y_n}). \tag{2.14}$$

证明思路利用Jacobi直接验证两边的密度函数即可以。由此定理，我们有

$$U_{nk_n}\stackrel{d}{=}e^{-\sum_{i=1}^{n+1-k_n}\frac{Z_i}{n+1-i}}. \tag{2.15}$$

Lemma 2.4.1. 对于 $Z_1,\ldots,Z_n$为iid的负指数分布,密度函数为$e^{-x}I(x>0)$. 记

$$Y_{nk_n}=\sum_{i=1}^{n+1-k_n}\frac{Z_i}{n+1-i}. \tag{2.16}$$

那么当 $k_n/n\to\lambda\in(0,1)$且$\sqrt{n}(k_n/n-\lambda)\to 0$ 时，

$$EY_{nk_n}=\sum_{i=1}^{n+1-k_n}\frac{1}{n+1-i}\to\int_0^{1-\lambda}\frac{1}{1-x}dx=-\log\lambda; \tag{2.17}$$

$$nVar(Y_{nk_n})=n\sum_{i=1}^{n+1-k_n}\frac{1}{(n+1-i)^2}\to\int_0^{1-\lambda}\frac{1}{(1-x)^2}dx=\frac{1}{\lambda}-1; \tag{2.18}$$

$$\tag{2.19}$$

以及根据一般的中心极限定理，

$$\sqrt{n}(Y_{nk_n}-(-\log\lambda))\stackrel{d}{\to}N(0,\frac{1}{\lambda}-1). \tag{2.20}$$

由此结果，在Delta方法中取 $g(x) = e^{-x}$, 我们有下面的一般结果。定理后我们给出一个初等的证明。

Theorem 2.4.4. 对于$U(0,1)$上的次序统计量， 当 $k_n/n \to \lambda \in (0,1)$且$\sqrt{n}(k_n/n - \lambda) \to 0$ 时，

$$\sqrt{n}(U_{nk_n} - \lambda) \xrightarrow{d} N(0, \lambda(1-\lambda)). \tag{2.21}$$

Proof: U_{nk_n} 的分布函数为：

$$F_{k_n}(x) = \frac{n!}{(k_n-1)!(n-k_n)!} \int_0^x t^{k_n-1}(1-t)^{n-k_n} dt,\ 0 \le x \le 1,$$

密度函数：

$$f_{k_n}(x) = \frac{n!}{(k_n-1)!(n-k_n)!} x^{k_n-1}(1-x)^{n-k_n} I_{(0,1)}(x). \tag{2.22}$$

计算得到$\sqrt{n}(U_{uk_n} - \lambda)$的密度函数为

$$\begin{aligned} h_n(x) &= \frac{1}{\sqrt{n}} f_{k_n}(\frac{x}{\sqrt{n}} + \lambda) \\ &= \frac{1}{\sqrt{n}} \frac{n!}{(k_n-1)!(n-k_n)!} (\frac{x}{\sqrt{n}} + \lambda)^{k_n-1} (1 - (\frac{x}{\sqrt{n}} + \lambda))^{n-k_n} I_{(0,1)}(\frac{x}{\sqrt{n}} + \lambda)\Theta \end{aligned}$$

$\forall x \in \mathbb{R}$, 我们有(为了记号简单，$k = k_n$)

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{n!}{(k-1)!(n-k)!} (\frac{x}{\sqrt{n}} + \lambda)^{k-1} (1 - (\frac{x}{\sqrt{n}} + \lambda))^{n-k_n} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\sqrt{2\pi n} n^n e^{-n}}{\sqrt{2\pi(k-1)}(\frac{k-1}{e})^{k-1}\sqrt{2\pi(n-k)}(\frac{n-k}{e})^{n-k}} (\frac{x}{\sqrt{n}} + \lambda)^{k-1} (1 - (\frac{x}{\sqrt{n}} + \lambda))^{n-k} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi e \lambda(1-\lambda)}} (\frac{n}{k-1})^{k-1} (\frac{n}{n-k})^{n-k} (\frac{x}{\sqrt{n}} + \lambda)^{k-1} (1 - (\frac{x}{\sqrt{n}} + \lambda))^{n-k} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} \lambda^{-(k-1)} (1-\lambda)^{-(n-k)} (\frac{x}{\sqrt{n}} + \lambda)^{k-1} (1 - (\frac{x}{\sqrt{n}} + \lambda))^{n-k} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} (1 + \frac{x}{\sqrt{n}\lambda})^{k-1} (1 - \frac{x}{\sqrt{n}(1-\lambda)})^{n-k} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} exp\left((k-1)\log(1 + \frac{x}{\sqrt{n}\lambda}) + (n-k)\log(1 - \frac{x}{\sqrt{n}(1-\lambda)}) \right) \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} exp\left((k-1)(\frac{x}{\sqrt{n}\lambda} - \frac{x^2}{2n\lambda^2}) - (n-k)(\frac{x}{\sqrt{n}(1-\lambda)} + \frac{x^2}{2n(1-\lambda)^2}) \right) \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} exp\left(-\frac{x^2}{2\lambda(1-\lambda)} + \frac{x}{\sqrt{n}}(\frac{k-1}{\lambda} - \frac{n-k}{1-\lambda}) \right) \\
&= \frac{1}{\sqrt{2\pi\lambda(1-\lambda)}} exp\left(-\frac{x^2}{2\lambda(1-\lambda)} \right) = h(x)
\end{aligned}
$$

这里用到 Stirling 公式， $n! = \sqrt{2\pi n}(n/e)^n$ 和 $\log(1-x) = -x - x^2/2 + O(x^3)$.
另外，对于$\forall x \in [-M, M]$，收敛是一致的。即$\forall \epsilon > 0, \exists N$

$$\sup_{x\in[-M,M]} |h_n(x) - h(x)| \le \epsilon,\ n > N.$$

这里$h(x)$是正态分布$N(0, \lambda(1-\lambda))$的密度函数。 对于给定的$x_0 \in \mathbb{R}$ 和 $\forall \epsilon > 0$, 首先存在$M > x_0$,使得

$$1 - \epsilon \le \int_{-M}^{M} h(x)dx \le 1. \tag{2.23}$$

其次，存在N使得$n > N$时：

$$\sup_{x\in[-M,M]} |h_n(x) - h(x)| \le \frac{\epsilon}{2M},\ n > N.$$

所以

$$\int_M^M |h_n(x) - h(x)|dx \leq \epsilon,$$

和

$$1 - 2\epsilon \leq \int_{-M}^{M} h_n(x)dx \leq 1,\ n > N. \tag{2.24}$$

现在我们可以看依分布收敛：

$$\begin{aligned} &|\int_{-\infty}^{x_0} h_n(x)dx - \int_{-\infty}^{x_0} h(x)dx| \\ \leq &\int_{-\infty}^{-M} h_n(x)dx + \int_{-\infty}^{-M} h(x)dx + \int_{-M}^{x_0} |h_n(x) - h(x)|dx \leq 4\epsilon. \end{aligned}$$

因此，我们得到

$$\sqrt{n}(U_{nk_n} - \lambda) \xrightarrow{d} N(0, \lambda(1-\lambda)). \tag{2.25}$$

Corollary 2.4.1. 由定理*2.4.4*和*Slutsky*定理，我们有

$$\frac{U_{nk_n} - EU_{nk_n}}{\sqrt{var(U_{uk_n})}} \xrightarrow{d} N(0,1), \tag{2.26}$$

这里$EU_{uk_n} = \frac{k_n}{n+1}$ 和 $var(U_{uk_n}) = \frac{k_n(n-k_n+1)}{(n+1)^2(n+2)}$.

对于任意的$0 < p < 1$, 考虑样本$(U_1, \ldots, U_n)$的p分位数

$$\epsilon_{np} = U_{([np])} + (n+1)(p - \frac{[np]}{n+1})(U_{([np]+1)} - U_{[np]}), \tag{2.27}$$

我们有下面结果。

Corollary 2.4.2. 对于任意的$0 < p < 1$,

$$\sqrt{n}(\epsilon_{np} - p) \xrightarrow{d} N(0, p(1-p)). \tag{2.28}$$

Proof: 在定理2.4.4取$k_n = [np]$和$k_n = [np]+1$，得到

$$\frac{\sqrt{n}(U_{([np])} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1), \quad \frac{\sqrt{n}(U_{([np]+1)} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

由分位数的定义：

$$\frac{\sqrt{n}(U_{([np])} - p)}{\sqrt{p(1-p)}} \leq \frac{\sqrt{n}(\epsilon_{np} - p)}{\sqrt{p(1-p)}} \leq \frac{\sqrt{n}(U_{([np]+1)} - p)}{\sqrt{p(1-p)}}$$

对于任意的$x_0 \in \mathbb{R}$,

$$P(\frac{\sqrt{n}(U_{([np]+1)} - p)}{\sqrt{p(1-p)}} \leq x_0) \leq P(\frac{\sqrt{n}(\epsilon_{np} - p)}{\sqrt{p(1-p)}} \leq x_0) \leq P(\frac{\sqrt{n}(U_{([np])} - p)}{\sqrt{p(1-p)}} \leq x_0),$$

不等式的两端都收敛到$\Phi(x_0)$,其中$\Phi(\cdot)$为标准正态分布的分布函数，所以

$$\sqrt{n}(\epsilon_{np} - p) \xrightarrow{d} N(0, p(1-p)), \tag{2.29}$$

或

$$\frac{\sqrt{n}(\epsilon_{np} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

Remark 2.4.1. 对于$U(0,1)$上得*iid*样本$U_1, \ldots, U_n$, 对于任意$p \in (0,1)$, 我们考虑三个统计量

$$\begin{aligned}
W_n^1 &= U_{[np]}, \\
W_n^2 &= \epsilon_n = U_{([np])} + (n+1)(p - \frac{[np]}{n+1})(U_{([np]+1)} - U_{[np]}), \\
W_n^3 &= \frac{1}{n}\sum_{k=1}^{n} I(U_k \leq p),
\end{aligned}$$

我们知道，对于这三个统计量，都有

$$\sqrt{n}(W_n^i - p) \xrightarrow{d} N(0, p(1-p)), i = 1,2,3. \tag{2.30}$$

其中

$$\{W_n^1 \leq x\} = \{U_1, \ldots, U_n \text{中至少} [np] \text{个不大于} x\}, \tag{2.31}$$
$$\{W_n^3 \leq x\} = \{U_1, \ldots, U_n \text{中至少} [nx] \text{个不大于} p\}. \tag{2.32}$$

2.5 多元联合分布

我们可以考虑多元的情形，有如下的结果：

Theorem 2.5.1. 对于$U(0,1)$上的次序统计量，当 $a_{nj}/n \to p_j \in (0,1)$且$\sqrt{n}(a_{nj}/n - p_j) \to 0$ 时，

$$\sqrt{n}\begin{pmatrix} U_{na_{n1}} - p_1 \\ U_{na_{n2}} - p_2 \\ \vdots \\ U_{na_{nk}} - p_k \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) & \dots & p_1(1-p_k) \\ p_1(1-p_2) & p_2(1-p_2) & \dots & p_2(1-p_k) \\ \vdots & \vdots & \vdots & \vdots \\ p_1(1-p_k) & p_2(1-p_k) & \dots & p_k(1-p_k) \end{pmatrix}\right). \tag{2.33}$$

这里 $0 < p_1 < p_2 < \cdots < p_k < 1$. 注意这里协方差结构里没有 $p_2(1-p_1)$项。

Theorem 2.5.2. 对于$U(0,1)$上的次序统计量，对于任意$0 < p_1 < p_2 < \cdots < p_k < 1$分位点,

$$\sqrt{n}\begin{pmatrix} \epsilon_{np_1} - p_1 \\ \epsilon_{np_2} - p_2 \\ \vdots \\ \epsilon_{np_k} - p_k \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) & \dots & p_1(1-p_k) \\ p_1(1-p_2) & p_2(1-p_2) & \dots & p_2(1-p_k) \\ \vdots & \vdots & \vdots & \vdots \\ p_1(1-p_k) & p_2(1-p_k) & \dots & p_k(1-p_k) \end{pmatrix}\right). \tag{2.34}$$

2.5.1 一般结果

有了这些结果，对于一般的次序统计量，注意到 $X_{nk} \overset{d}{=} F^{-1}(U_{nk})$, 我们可以再次用Delta 方法得到对应的结果。

$$\sqrt{n}\begin{pmatrix} X_{ua_{n1}} - \epsilon_1 \\ X_{ua_{n2}} - \epsilon_2 \\ \vdots \\ X_{ua_{nk}} - \epsilon_k \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\epsilon_1(1-\epsilon_1)}{f(\epsilon_1)f(\epsilon_1)} & \frac{\epsilon_1(1-\epsilon_2)}{f(\epsilon_1)f(\epsilon_2)} & \cdots & \frac{\epsilon_1(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} \\ \frac{\epsilon_1(1-\epsilon_2)}{f(\epsilon_1)f(\epsilon_2)} & \frac{\epsilon_2(1-\epsilon_2)}{f(\epsilon_2)f(\epsilon_2)} & \cdots & \frac{\epsilon_2(1-\epsilon_k)}{f(\epsilon_2)f(\epsilon_k)} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\epsilon_1(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} & \frac{\epsilon_2(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} & \cdots & \frac{\epsilon_k(1-\epsilon_k)}{f(\epsilon_k)f(\epsilon_k)} \end{pmatrix}\right).$$

其中 $\epsilon_j = F^{-1}(p_j)$ 即分布对应的分位点，而$f(\epsilon_i)f(\epsilon_j)$的出现是因为Delta方法中的偏导数部分得到的。

对于样本分位数，任意的$0 < p < 1$,

$$\epsilon_{np} = X_{([np])} + (n+1)(p - \frac{[np]}{n+1})(X_{([np]+1)} - X_{[np]}), \tag{2.35}$$

介于 $\sqrt{n}(X_{([np])} - (X_{([np]+1)}) = o_p(1)$ ，而$X_{([np])}$ 或者 $(X_{([np]+1)}$ 满足上面定理的条件，也有和(28)一样的结果。

Theorem 2.5.3. 对于任意 $0 < p_1 < \cdots < p_k < 1$, 考虑样本分位数

$$\epsilon_n = (\epsilon_{np_1}, \ldots, \epsilon_{np_k})', \tag{2.36}$$

和总体分位数 $\epsilon = (\epsilon_1, \ldots, \epsilon_k) = (F^{-1}(p_1), \ldots, F^{-n}(p_k))'$. 我们有

$$\sqrt{n}(\epsilon_n - \epsilon) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\epsilon_1(1-\epsilon_1)}{f(\epsilon_1)f(\epsilon_1)} & \frac{\epsilon_1(1-\epsilon_2)}{f(\epsilon_1)f(\epsilon_2)} & \cdots & \frac{\epsilon_1(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} \\ \frac{\epsilon_1(1-\epsilon_2)}{f(\epsilon_1)f(\epsilon_2)} & \frac{\epsilon_2(1-\epsilon_2)}{f(\epsilon_2)f(\epsilon_2)} & \cdots & \frac{\epsilon_2(1-\epsilon_k)}{f(\epsilon_2)f(\epsilon_k)} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\epsilon_1(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} & \frac{\epsilon_2(1-\epsilon_k)}{f(\epsilon_1)f(\epsilon_k)} & \cdots & \frac{\epsilon_k(1-\epsilon_k)}{f(\epsilon_k)f(\epsilon_k)} \end{pmatrix} \right). \tag{2.37}$$

对于某一个样本分位数：

$$\sqrt{n}(\epsilon_{np} - F^{-1}(p)) \xrightarrow{d} N(0, p(1-p)/(f(F^{-1}(p)))^2). \tag{2.38}$$

例如中位数， $p = 1/2$,

$$\sqrt{n}(\epsilon_{1/2} - F^{-1}(1/2)) \xrightarrow{d} N(0, 1/4(f(F^{-1}(1/2)))^2). \tag{2.39}$$

Remark 2.5.1. 我们考虑经验分布函数，

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I(X_n \le x),$$

注意到这里的统计量$Z_k := I(X_k \le x)$ 是iid的，且

$$EZ_k = E(I(X_1 \le x)) = P(X_1 \le x) = F(x),\ Var(Z_k) = F(x)(1 - F(x)).$$

由经典中心极限定理，如果$F(x)(1 - F(x)) > 0$,

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x)). \tag{2.40}$$

对任意$0 < p < 1$, 取x 为$p-$分位数 ϵ_p，即$x = \epsilon_p = F^{-1}(p)$, 我们有

$$\sqrt{n}(F_n(\epsilon_p) - p) \xrightarrow{d} N(0, p(1-p)). \tag{2.41}$$

再进一步，如果考虑统计量$W_n = F^{-1}(F_n(\epsilon_p))$ (这个统计量也非常类似于样本p分位数)， 由$Delta$方法，我们也有

$$\sqrt{n}(W_n - F^{-1}(p)) \xrightarrow{d} N(0, p(1-p)/(f(F^{-1}(p)))^2). \tag{2.42}$$

注意这里的条件(最后的$Delta$方法需要用到)，即 $F^{-1}(\cdot)$在 $F_n(\epsilon_p)$可导且导数不为0.

2.5.2 极值的极限分布

对于次序统计量的正则项，我们知道渐近分布为正态分布，下面我们考虑次序统计量的两端即极值统计量的渐近分布。

Definition 2.5.1 (渐近分布). 对一列随机变量$\{X_n\}$, 存在数列$\{a_n > 0\}$ 和$\{b_n\}$,使得$a_n(X_n - b_n)$ 依分布收敛到一个分布函数G.

Remark 2.5.2. 渐近分布中的数列不唯一，例如在极值

$$n(U_{(n)} - 1) \xrightarrow{d} Exp(1). \tag{2.43}$$

中$a_n = n,\ b_n = 1$, 我们可以用任意的$a_n = cn,\ b_n = 1 + c/n^2$都可以的。由*Slutsky*定理，$a_n/n \to C_1,\ a_n b_n \to C_2$即可。

先考虑$U_(0,1)$的特殊情况，考虑$U_{(1)}$和$U_{(n)}$的渐近分布. 对于U_1， 我们有其密度函数，

$$h_n(x) = n(1-x)^{n-1} I_{(0,1)(x)}, \tag{2.44}$$

我们现在来看 $nU_{(1)}$的分布， 其密度函数为

$$\begin{aligned} h_n(x) &= \frac{1}{n} h_n(x/n) \\ &= (1 - \frac{x}{n})^{n-1} \to e^{-x}. \end{aligned}$$

即

$$nU_{(1)} \xrightarrow{d} Exp(-1). \tag{2.45}$$

注意到$U_{(n)} \overset{d}{=} 1 - U_{(1)}$, 对于极大值，

$$n(1 - U_{(n)}) \xrightarrow{d} Exp(-1). \tag{2.46}$$

或者

$$n(U_{(n)} - 1) \xrightarrow{d} Exp(1). \tag{2.47}$$

Remark 2.5.3. 如果样本$X_1, \ldots, X_n$ *iid* 来自均匀分布$U(a,b)$, 那么对应的极值统计量

$$\begin{aligned} &n\frac{X_{(1)} - a}{b-a} = \frac{n}{b-a}(X_{(1)} - a) \xrightarrow{d} Exp(-1); \\ &n(1 - \frac{X_{(n)} - a}{b-a}) = \frac{-n}{b-a}(X_{(1)} - b) \xrightarrow{d} Exp(-1); \\ &\text{或 } \frac{n}{b-a}(X_{(1)} - b) \xrightarrow{d} Exp(1). \end{aligned}$$

现在我们来看一般的极值分布，由于$X_{(1)}$可以转化为样本$-X_1,\ldots,-X_n$中的极大值取负号，故我们只考虑极大值的渐近分布。对于一般的分布函数$F(x)$, 我们知道$X_{(n)}$的分布函数：

$$F_n(x) = F^n(x).$$

那么$a_n(X_{(n)} - b_n)$的分布函数和密度函数为：

$$G_n(x) = F^n(b_n + x/a_n).$$

- $F(x)$为均匀分布$U(0,1)$的分布函数时候，取$a_n = n, b_n = 1$,

$$(b_n + x/a_n)^n \to e^x, x < 0. \tag{2.48}$$

即这里的极限分布函数为：

$$G(x) = \begin{cases} e^x & \text{if } x \le 0; \\ 1 & \text{if } x > 0. \end{cases}$$

- $F(x)$为指数分布$Exp(-1)$的分布函数：

$$F(x) = \begin{cases} 0 & \text{if } x \le 0; \\ 1 - e^{-x} & \text{if } x > 0. \end{cases}$$

我们要考虑序列：

$$(1 - e^{-b_n - x/a_n})^n \tag{2.49}$$

或者等价的

$$\begin{aligned} &n\log(1 - e^{-b_n - x/a_n}) = -ne^{-b_n - x/a_n} + o(1) \\ &\log n - b_n - x/a_n \end{aligned}$$

取$b_n = \log n, a_n = 1$, 我们有任意x,

$$(1 - e^{-b_n - x/a_n})^n \to e^{-e^{-x}},$$

即$X_{(n)} - \log n$的极限分布为Gumbel分布：

$$G(x) = e^{-e^{-x}}, -\infty < x < \infty.$$

- $F(x)$ 为上述的Gumbel分布， 找序列极限：

$$(e^{-e^{-(b_n+x/a_n)}})^n$$

等价的

$$-ne^{-(b_n+x/a_n)}$$
$$\log n-(b_n+x/a_n)$$

所以极限分布还是Gambel分布。

Theorem 2.5.4 (极值分布的三大类型). *若$G(x)$为一个连续的极大值分布，则其必与下列三个类型的分布函数同类，这里同类的意思是存在a,b使得$G(ax+b)$具有下列形式：*

*Gumbel*分布： $G_1(x)=e^{-e^{-x}}$;

*Weibull*分布： $G_2(x)=e^{-x^{-\alpha}}I(x>0),\ \alpha>0$;

*Frechet*分布：

$$G_3(x)=\begin{cases} e^{-|x|^\alpha} & if\ x\le 0; \\ 1 & if\ x>0. \end{cases},\ \alpha>0.$$

文献中有这三种分布对应的总体分布F需要满足的充要条件：

- 存在$c<1$, 使得当$F(x)>c$时，F 在x点连续，则F属于 Gumbel型极大值分布的充要条件为：对于任意x,

$$\lim_{n\to\infty} n(1-F(b_n+\frac{x}{a_n}))=e^{-x} \tag{2.50}$$

其中a_n,b_n满足

$$F(b_n)=1-\frac{1}{n},\ F(b_n+1/a_n)=1-\frac{1}{ne}. \tag{2.51}$$

例如 指数分布的分布函数。

- 对一切x, $F(x) < 1$,且存在$\alpha > 0$, 对于任意的$c > 0$,

$$\lim_{x\to\infty} \frac{1 - F(x)}{1 - F(cx)} = c^{\alpha}.$$

例如 密度函数为$f(x) = C_1 x^{-(1+\alpha)} I(x > 0)$的分布函数。

- 存在有限的ω, $F(\omega) = 1$ 且对$\omega_1 < \omega$, $F(\omega_1) < 1$,并存在$\alpha > 0$, 对于任何的$c > 0$,

$$\lim_{x\to 0^-} \frac{1 - F(cx + \omega)}{1 - F(x + \omega)} = c^{\alpha}.$$

例如 均匀分布的分布函数或者密度函数为$f(x) = C_2(1-x)^{(\alpha-1)} I(0 < x < 1)$的分布函数。

思考： 正态分布属于的极值是哪一个类别呢?

实际应用中需要一定的专业知识去决定选取何种分布，Gumbel分布使用的偏多一点。

2.5.3 次序统计量的线性函数

基于次序统计量，我们可以构造统计量

$$T_n = \sum_{i=1}^{n} \omega_{ni} X_{(ni)}. \tag{2.52}$$

前面提到的所有次序统计量都是这种形式，但都是有限的几个组成的，这里我们考虑一类特殊的：

$$\omega_{ni} = J(\frac{i}{n+1})/n,\ i = 1, \ldots, n.$$

Theorem 2.5.5 (Moore,1968)**.** 设分布F处处连续且期望存在，即

$$\int |x| dF(x) < \infty. \tag{2.53}$$

函数$J(\cdot)$满足条件：

1. J在$[0, 1]$上，除可能有限个第一类间断点，处处连续；

2. 除有限个点之外，J'处处连续；在例外点上指定$J'=0$,J'在$[0,1]$上为有界变差。

3. 记$G(x)=F^{-1}(x)=\inf\{y:F(y)\geq x\}$, 有

$$\sigma^2=\int_0^1\int_0^1 J(s)J(t)min(s,t)(1-max(s,t))dG(s)dG(t)<\infty. \tag{2.54}$$

那么，

$$\sqrt{n}(T_n-\int xJ(F(x))dF(x))\xrightarrow{d}N(0,\sigma^2). \tag{2.55}$$

2.6 次序统计量的应用

1. 总体分位数的估计

$$\sqrt{n}(X_{[np]}-F^{-1}(p))\xrightarrow{d}N(0,p(1-p)/(f(F^{-1}(p)))^2). \tag{2.56}$$

2. 位置与刻度参数: $X_1,\ldots,X_{n_1}iid\sim F(x)$ 和 $Y_1,\ldots,Y_{n_2}iidF(x-\theta)$，即$Y_1\overset{d}{=}X_1+\theta$. 考虑$\theta$的置信区间：

$$[Y_{(s')}-X_{(r')},Y_{(s)-X_{(r)}}]$$

3. 截尾数据 一般的，如果在次序统计量$X_{(1)},\ldots,X_{(n)}$中只观察了$X_{(r)},\ldots,X_{(s)}$而其余的没有观察或者不能被观察，则称$X_{(r)},\ldots,X_{(s)}$为截尾样本或者截尾数据(Censored Data).

4. 极值统计量

5. $\cdots$.

Chapter 3

U统计量

3.1 学习目标

1. 了解V，U统计量及其定义；
2. 理解V,U统计量的原理
3. 学习U统计量的大样本性质
4. 可以用U统计量构造相关的统计应用。

3.2 引例

给定iid的样本$X_1, \ldots, X_n$, 记期望为$\theta = EX_i$, 如何构造θ的估计?

一般的，由距估计， 用样本均值去估计θ, 即

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

这里的下标n是为了强调这里的估计与样本个数n相关，突出其大样本性质。由强大数定律，

$$\bar{X}_n \xrightarrow{a.s.} \theta,$$

进一步的如果方差$\sigma^2 = Var(X_1) < \infty$存在，经典中心定理：

$$\sqrt{n}(\bar{X}_n - \theta) \rightsquigarrow N(0, \sigma^2).$$

考虑一个具体的假设检验问题：

$$H_0:\ \theta=\theta_0\ \ vs\ \ H_1:\ \theta\neq\theta_0. \tag{3.1}$$

由CLT结果，

$$\frac{\sqrt{n}(\bar{X}_n-\theta)}{\sigma}\ \rightsquigarrow N(0,1).$$

当然，一般我们不知道方差σ^2,由Slutsky Theorem, 我们可以用其相合估计来代替，即我们有

$$s_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\xrightarrow{a.s.}\sigma^2,$$
$$\frac{\sqrt{n}(\bar{X}_n-\theta)}{s_n}\ \rightsquigarrow N(0,1),$$

所以检验统计量和拒绝域为：

$$T_n=\frac{\sqrt{n}(\bar{X}_n-\theta_0)}{s_n},\ \{|T_n|\geq z_{1-\alpha/2}\}.$$

注意这里和传统数理统计结果的区别：1. 完全不基于正态分布(只需要二阶距有限)；2. 渐近分布为正态.

进一步的我们可以计算检验统计量的功效Power如何：

$$\begin{aligned}
&P(T_n|\geq z_{1-\alpha/2})\\
&=P(|\frac{\sqrt{n}(\bar{X}_n-\theta_0)}{s_n}|\geq z_{1-\alpha/2})\\
&=P(|\frac{\sqrt{n}(\bar{X}_n-\theta)}{s_n}+\frac{\sqrt{n}(\theta-\theta_0)}{s_n}|\geq z_{1-\alpha/2})\\
&=P(\frac{\sqrt{n}(\bar{X}_n-\theta)}{s_n}\geq z_{1-\alpha/2}-\frac{\sqrt{n}(\theta-\theta_0)}{s_n})+P(\frac{\sqrt{n}(\bar{X}_n-\theta)}{s_n}\leq -z_{1-\alpha/2}-\frac{\sqrt{n}(\theta-\theta_0)}{s_n})\\
&\approx 1-\Phi(z_{1-\alpha/2}-\frac{\sqrt{n}(\theta-\theta_0)}{s_n})+\Phi(-z_{1-\alpha/2}-\frac{\sqrt{n}(\theta-\theta_0)}{s_n})
\end{aligned}$$

所以检验的功效基于$\frac{\sqrt{n}(\theta-\theta_0)}{\sigma}$. 我们可以看到备择假设$\theta$和数据的标准差$\sigma$对于检验效果的影响.

思考： 如何构造θ^2的估计？这一估计在构造损失函数时候会用到。例如上面例子构造$(\theta-\theta_0)^2$的估计，即θ和θ^2的估计。

当然我们可以用$\bar{X}_n^2$去估计θ^2, 这也是数理统计中经常用的结果，直接计算得到：

$$\begin{aligned} E\bar{X}_n^2 &= E\frac{1}{n^2}\sum_{ij} X_iX_j \\ &= E\frac{1}{n^2}\sum_{i\neq j} X_iX_j + E\frac{1}{n^2}\sum_k X_k^2 \\ &= \frac{n(n-1)\theta^2}{n^2} + \frac{\theta^2+Var(X_1)}{n} \\ &= \theta^2 + \frac{1}{n}Var(X_1). \end{aligned}$$

所以这一估计是系统偏大的，而且如果数据本身方差不存在(即二阶矩不存在), 系统偏差会非常大。

从刚刚的计算中可以发现平方项是不需要的，也就是我们可以直接用下面的估计：

$$T_n = \frac{1}{n(n-1)}\sum_{i\neq j} X_iX_j, \tag{3.2}$$

其是θ^2的一个无偏估计。这里T_n就是一个U统计量。

3.3 V统计量和U统计量

介绍U统计量之前，我们先介绍V统计量。

对于一个iid样本$X_1,\ldots,X_n$,其分布函数为F, 统计中感兴趣的的参数θ是分布函数的某个函数，即$\theta=T(F)$，例如

- $T(F)=F(c)=\int I(x\leq c)dF(x)$,某一点的概率；
- $T(F)=F^{-1}(p)$,某一个分位点；
- $T(F)=\int xdF(x)$,即我们常见的期望，类似的也可以定义方差等。

对于F的估计，经验分布函数(Empirical Distribution Function),

$$F_n(x) = \frac{1}{n}\sum_{k=1}^n I(X_k\leq x). \tag{3.3}$$

是$F(x)$的无偏估计. 此外，EDF还有如下很好的性质：

1. For any fixed x,

$$E(\hat{F}_n(x)) = F(x), \text{ and } Var(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}. \quad (3.4)$$

Further, if $0 < F(x) < 1$, we have

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \rightsquigarrow N(0, F(x)(1-F(x)). \quad (3.5)$$

2. Gilvenko-Cantelli Theorem:

$$\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0. \quad (3.6)$$

3. Dvoretzky–Kiefer–Wolfowitz (DKW) inequality: For any $\epsilon > 0$,

$$P(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}. \quad (3.7)$$

Definition 3.3.1 (V-Statistics, Richard von Mises,1947). 对于参数$\theta = T(F)$,我们可以用经验分布函数对应的$\hat{\theta} = T(F_n)$去估计。

例如前面提到的参数：

- 某一点概率 $T(F) = F(c) = \int I(x \leq c) dF(x)$,

$$\begin{aligned} T(F_n) &= \int I(x \leq c) dF_n(x) \\ &= \sum_{k=1}^{n} \frac{I(X_k \leq c)}{n} = \frac{1}{n} \sum_{k=1}^{n} I(X_k \leq c) = F_n(c); \end{aligned}$$

- 分位点 $T(F) = F^{-1}(p)$： $T(F_n)$对应的就是F_n的分位点，但是由于F_n^{-1}的不唯一性，定义$F_n^{-1}(p) = \inf_x\{x : F(x) \geq p\}$;

- $T(F) = \int x dF(x)$,即我们常见的期望，类似的也可以定义方差等。

$$T(F_n) = \int x dF_n(x) = \frac{1}{n} \sum_{k=1}^{n} X_k,$$

方差：

$$T(F) = \int x^2 dF(x) - (\int x dF(x))^2;$$
$$T(F_n) = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - (\frac{1}{n}\sum_{k=1}^{n} X_k)^2.$$

如果$T(F)$是线性函数形式，即$T(F) = \int h(x)dF(x)$, 我们有

$$T(F_n) = \frac{1}{n}\sum_{k=1}^{n} h(X_k),$$

类似样本均值，我们也有强大数定律和中心极限定理等。

对于方差，我们有

$$\begin{aligned} T(F) &= \int x^2 dF(x) - (\int x dF(x))^2 \\ &= \frac{1}{2}(\int x^2 dF(x) + \int y^2 dF(y) - \int x dF(x) \int y dF(y)) \\ &= \frac{1}{2}\int\int (x-y)^2 dF(x) dF(y). \end{aligned}$$

即一般的

$$T(F) = \int \cdots \int h(x_1, \ldots, x_m) dF(x_1) \cdot dF(x_m). \tag{3.8}$$

那么对应的V统计量为：

$$T(F_n) = \frac{1}{n^m}\sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}) = \frac{1}{n^m} \sum_{(i_1,\ldots,i_m)\in(1,2,\ldots,n)} h(X_{i_1}, \ldots, X_{i_m})$$

Example 3.3.1. 我们现在来看$\theta = E|X_1 - X_2| = \int\int |x-y| dF(x)dF(y)$, 对应的$V$统计量就为：

$$T(F_n) = \frac{1}{n^2}\sum_{i,j=1}^{n} |X_i - X_j| = \frac{2}{n^2}\sum_{i<j} |X_i - X_j|.$$

求期望：

$$ET(F_n) = \frac{2}{n^2}\sum_{i<j} E|X_i - X_j| = \frac{n-1}{n}\theta,$$

即V统计量是有系统偏差的。

仔细检查会发现 3.8也是有系统偏差的。这里一般的

$$Eh(X_1, \ldots, X_1) = T(F)$$

不一定成立，类似的$Eh(X_{i_1}, \ldots, X_{i_m}) = T(F)$ 当下标$i_1, \cdots, i_m$有重复时候都不一定成立。

Definition 3.3.2 (U Statistics)**.**

$$W_n = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{(i_1,\ldots,i_m)}^{*} h(X_{i_1}, \ldots, X_{i_m})$$

这里的$\sum^*$对所有不相等的下标求和，即$i_1, \ldots, i_n$中任意两个都不相等。

这里的$h(x_1, \ldots, x_n)$ 可以假定是对称的，否则，我们可以重新定义，

$$\phi(x) = \frac{1}{m!} \sum_{(i_1,\ldots,i_m)=\pi(1,\ldots,m)} h(x_{i_1}, \ldots, x_{i_m}) \tag{3.9}$$

这里$\pi(1, \ldots, m)$ 表示$(1, \ldots, m)$的所有排列组合。

Definition 3.3.3 (Wassily Hoeffding， 1948)**.**

$$U_n = \frac{1}{C_n^m} \sum_{1 \le i_1 < \cdots < i_m \le n} \phi(X_{i_1}, \ldots, X_{i_m})$$

这里的ϕ称为U统计量的核函数，m称为阶。

3.4 U 统计量的渐近性质

对于U统计量：

$$U_n = \frac{1}{C_n^m} \sum_{1 \le i_1 < \cdots < i_m \le n} \phi(X_{i_1}, \ldots, X_{i_m}),$$

我们考虑其大样本性质。 对于其均值，我们知道：

$$E(U_n) = E\phi(X_1, \ldots, X_m).$$

现在我们计算其方差：

$$\begin{aligned} Var(U_n) &= Var(\frac{1}{C_n^m} \sum_{1\le i_1<\cdots<i_m\le n} \phi(X_{i_1},\ldots,X_{i_m})) \\ &= \frac{1}{(C_n^m)^2} \sum_{i_1<\cdots<i_m} \sum_{j_1<\cdots<j_m} Cov(\phi(X_{i_1},\ldots,X_{i_m}),\phi(X_{j_1},\ldots,X_{j_m})). \end{aligned}$$

Definition 3.4.1. 基于核函数$\phi(x_1,\ldots,x_m)$,我们定义一列新的函数：

$$\phi_k(x_1,\ldots,x_k) = E\phi(x_1,\ldots,x_k,X_{k+1},\ldots,X_m),\ 0\le k\le m,$$

特别的

$$\phi_0 = E\phi(X_1,\ldots,X_n),\ \phi_m = \phi(x_1,\ldots,x_m) = \phi.$$

对于任意的k, $E\phi_k(X_1,\ldots,X_k) = E\phi(X_1,\ldots,X_m)$. 对于其方差，我们定义

$$\sigma_k^2 = Var(\phi_k(X_1,\ldots,X_k)).$$

特别的，$\sigma_0^2 = 0$ 和 $\sigma_m^2 = Var(\phi(X_1,\ldots,X_m))$. 进一步的，我们有

$$0 = \sigma_0^2 \le \sigma_1^2 \le \cdots \le \sigma_m^2 = Var(\phi(X_1,\ldots,X_m)).$$

所以为了控制这列函数的方差，对于核函数我们需要条件：

$$\sigma^2 = Var(\phi(X_1,\ldots,X_m)) < \infty.$$

所以，对于U_n的方差，我们有

$$Cov(\phi(X_{i_1},\ldots,X_{i_m}),\phi(X_{j_1},\ldots,X_{j_m})) = \sigma_k^2,$$

其中k是$(i_1,\ldots,i_m)$与$(j_1,\ldots,j_m)$中相等的对数。

Lemma 3.4.1. 记$\sigma_k^2 = Var(\phi_k(X_1,\ldots,X_k))$,我们有

$$\begin{aligned} Var(U_n) &= Var(\frac{1}{C_n^m} \sum_{1\le i_1<\cdots<i_m\le n} \phi(X_{i_1},\ldots,X_{i_m})) \\ &= \frac{1}{(C_n^m)^2} \sum_{i_1<\cdots<i_m} \sum_{j_1<\cdots<j_m} Cov(\phi(X_{i_1},\ldots,X_{i_m}),\phi(X_{j_1},\ldots,X_{j_m})) \\ &= \frac{1}{C_n^m} \sum_{j_1<\cdots<j_m} Cov(\phi(X_1,\ldots,X_m),\phi(X_{j_1},\ldots,X_{j_m})) \\ &= \frac{1}{C_n^m} \sum_{k=1}^{m} C_m^k C_{n-m}^{m-k} \sigma_k^2. \end{aligned}$$

对于很大的n, 我们有

$$Var(U_n) = \frac{m^2}{n}\sigma_1^2(1 + o(1)). \tag{3.10}$$

Theorem 3.4.1. 假定$\sigma^2 = Var(\phi(X_1, \ldots, X_m)) < \infty$ 和$\sigma_1^2 = Var(\phi_1(X_1)) > 0$, 我们有

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, m^2\sigma_1^2), \tag{3.11}$$

其中$\theta = E\phi(X_1, \ldots, X_m)$.

Proof: 我们首先构造一个新的"U"估计，

$$U_n^* = \frac{1}{n}\sum_{k=1}^{n}\phi_1(X_k),$$

注意这里的ϕ_1并不能成为一个真正的核函数，因为其本身一般要基于分布函数。 我们有

$$\phi_1(X_1), \cdots, \phi_1(X_n),$$

iid的服从均值为θ，方差为$\sigma_1^2 \le \sigma^2 < \infty$.由标准的CLT，

$$\sqrt{n}(U_n^* - \theta) \rightsquigarrow N(0, \sigma_1^2).$$

下面我们证明：

$$\sqrt{n}(U_n - \theta) = m\sqrt{n}(U_n^* - \theta) + o_p(1).$$

记

$$T_n = \sqrt{n}(U_n - \theta) - m\sqrt{n}(U_n^* - \theta).$$

直接计算可以得到 $ET_n = 0$,

$$\begin{aligned} Var(T_n) &= nVar(U_n) + m^2 nVar(U_n^*) - 2mnCov(U_n, U_n^*) \\ &= \frac{n}{C_n^m}\sum_{k=1}^{m} C_m^k C_{n-m}^{m-k}\sigma_k^2 + m^2\sigma_1^2 - 2m^2\sigma_1^2 \\ &= \sum_{k=1}^{m}\frac{m!m!}{k!(m-k)!(m-k)!}\frac{(n-m)!(n-m)!}{(n-1)!(n-2m+k)!}\sigma_k^2 - m^2\sigma_1^2 \\ &= O(\frac{1}{n}). \end{aligned}$$

由Markov不等式(或者Chebyshev不等式)，我们有

$$T_n \xrightarrow{p} 0.$$

由Slutsky定理，我们有

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, m^2\sigma_1^2). \tag{3.12}$$

3.4.1 检验总体均值

对于iid样本$X_1, \ldots, X_n$, 其中$\mu = E(X_1)$, $\sigma^2 = Var(X_1)$. 对于参数$\theta = \mu^2 = (E(X_1))^2$, 我们知道其U统计量为：

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} X_i X_j.$$

这里

$\phi(x_1, x_2) = x_1 x_2$;
$\phi_1(x) = E\phi(xX_1) = E(xX_1) = \mu x$;
$\sigma_2^2 = Var(\phi(X_1, X_2)) = E(X_1X_2 - \theta)^2 = (EX_1^2)^2 - \theta^2 = (\theta + \sigma^2)^2 - \theta^2$,
$\sigma_1^2 = Var(\phi_1(X_1)) = \theta\sigma^2$.

由之前的渐近结果，如果$\sigma^2 < \infty$:

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, 4\theta\sigma^2).$$

基于此，我们可以构造一个新的检验统计量:

$$H_0 : \mu = \mu_0 \ v.s. \ H_1 : \mu \neq \mu_0. \tag{3.13}$$

即我们用U统计量去估计参数$\theta = (\mu - \mu_0)^2$,所用的核函数为：

$$\phi(x_1, x_2) = x_1 x_2 - (x_1 + x_2)\mu_0 + \mu_0^2 = (x_1 - \mu_0)(x_2 - \mu_0);$$

对应的U统计量：

$$T_n = U_n = \frac{2}{n(n-1)} \sum_{i<j} (X_i X_j - (X_i + X_j)\mu_0 + \mu_0^2) = \frac{2}{n(n-1)} \sum_{i<j} (X_i - \mu_0)(X_j - \mu_0).$$

计算可以得到：

$$\begin{aligned}
&Var(\phi(X_1, X_2)) = \dots \\
&\phi_1(x) = E\phi(x, X_1) = \mu x - (\mu + x)\mu_0 + \mu_0^2 = (\mu - \mu_0)(x - \mu_0), \\
&\sigma_1^2 = Var(\phi_1(X_1)) = (\mu - \mu_0)^2\sigma^2
\end{aligned}$$

所以如果$\sigma^2 < 0$:

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, 4\theta\sigma^2) \tag{3.14}$$

然而对于原假设情形H_0成立时候，我们知道$\theta = 0$,所以上面的渐近结果是不对的，并不能给出我们需要的置信区间或者分位点。

当$\mu = \mu_0$时候，这时候我们需要高一阶的方差：

$$Var((X_1 - \mu)(X_2 - \mu)) = \sigma^4.$$

所以$Var(T_n) = C_n^2\sigma^4 = \frac{n(n-1)}{2}\sigma^4$,这是否意味着我们有

$$\frac{T_n}{\sqrt{\frac{n(n-1)}{2}\sigma^4}} \rightsquigarrow N(0, 1)?$$

此时，实际上

$$T_n = \frac{2}{n(n-1)}\sum_{i<j}(X_i - \mu)(X_j - \mu) = \frac{1}{n-1}(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) - \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2)$$

由标准的CLT和强大数定律，

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) \rightsquigarrow N(0, \sigma^2), \\
&\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 \xrightarrow{a.s.} \sigma^2
\end{aligned}$$

所以： $nT_n \rightsquigarrow (Z^2 - 1)\sigma^2$,其中$Z$是标准正态分布.基于此，我们可以给出$H_0$下的$T_n$的分位点等。注意这时候$\sigma^2 = Var(X_1)$未知，我们还需要用样本方差构造出其的相合估计，最后可以得到上述问题的置信区间等，结果非常类似于t分布的平方。

3.5 两样本U统计量

考虑两个分布函数(F,G),统计上我们关心参数$\theta = \theta(F,G)$，类似的对于iid样本 $X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2}$ 我们也有U统计量:

$$U_{n_1,n_2} = \frac{1}{C_{n_1}^{m_1}C_{n_2}^{m_2}} \sum_{1\le i_1<\cdots<i_{m_1}\le n_1} \sum_{1\le j_1<\cdots<j_{m_2}\le n_2} \phi(X_{i_1},\ldots,X_{i_{m_1}},Y_{j_1},\ldots,Y_{j_{m_2}}) \tag{3.15}$$

注意这里的对称化,所有X_i地位是一样的，所有Y_j地位是一样的，但是两者不一定可以交换。

$$U_{n_1,n_2} = \frac{1}{m_1!m_2!C_{n_1}^{m_1}C_{n_2}^{m_2}} \sum_{i_1,\ldots,i_{m_1}} \sum_{j_1,\ldots,j_{m_2}} h(X_{i_1},\ldots,X_{i_{m_1}},Y_{j_1},\ldots,Y_{j_{m_2}}) \tag{3.16}$$

类似于一元的情形，我们也可以定义：

$$\begin{aligned}\sigma_{ij}^2 =& Cov(\phi(X_1,\ldots,X_i,X_{i+1},\ldots,X_{m_1},Y_1,\ldots,Y_j,Y_{j+1},\ldots,Y_{m_2}),\\ &\phi(X_1,\ldots,X_i,X_{i+1}^*,\ldots,X_{m_1}^*,Y_1,\ldots,Y_j,Y_{j+1}^*,\ldots,Y_{m_2}^*))\end{aligned}$$

这里$X_1,\ldots,X_i,X_{i+1},\ldots,X_{m_1},X_{i+1}^*,\ldots,X_{m_1}^*$ 独立同分布来自于F, Y_j的定义类似。 和一元情形一样，我们有

Theorem 3.5.1. 对于U_{n_1,n_2}的方差，我们有

$$Var(U_{n_1,n_2}) = \sum_{i=0}^{m_1}\sum_{j=0}^{m_2} \frac{C_{m_1}^i C_{n_1-m_1}^{m_1-i}}{C_{n_1}^{m_1}} \frac{C_{m_2}^j C_{n_2-m_2}^{m_2-j}}{C_{n_2}^{m_2}} \sigma_{ij}^2. \tag{3.17}$$

进一步的，如果$\sigma_{m_1,m_2}^2 < \infty$, $n_1+n_2\to\infty$且 $n_1/(n_1+n_2)\to p\in(0,1)$,

$$\sqrt{n_1+n_2}(U_{n_1,n_2}-\theta) \rightsquigarrow N(0,\sigma^2),\ \sigma^2 = \frac{m_1^2}{p}\sigma_{10}^2 + \frac{m_2^2}{1-p}\sigma_{01}^2. \tag{3.18}$$

3.6 U统计量的应用

U统计量为我们提供了一种新的无偏估计方法，而且方法不拘泥于特别的分布，具有很好的稳健性。结合大样本结果，我们看一些U统计量应用的例子，包括参数估计、置信区间、假设检验等。

3.6.1 方差相关项

对于iid样本$X_1, \ldots, X_n$, 考虑参数$\theta = E|X_1 - X_2|$。参数θ是评价数据离散程度的一个度量，类似于我们常用的总体标准差。 其U统计量为：

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} |X_1 - X_2|$$

我们有

$$\begin{aligned} \sigma^2 &= Var(|X_1 - X_2|), \\ \phi_1(x) &= E|x - X_1|, \\ \sigma_1^2 &= Var(\phi_1(X_1)), \end{aligned}$$

和

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, 4\sigma_1^2).$$

这里的方差σ_1^2是未知的，我们需要从样本出发得到其相合的估计,

$$\begin{aligned} \sigma_1^2 &= Var(\phi_1(X_1)) = Var(E[|X_1 - X_2||X_1]) \\ &= E|X_1 - X_2||X_1 - X_3| - (E|X_1 - X_2|)^2. \end{aligned}$$

所以我们可以再用U统计量构造出σ_1^2的估计， 这里有两种方式。第一种，因为已经有了θ的相合估计U_n, 所以U_n^2是θ^2的相合估计(连续映射定理)，我们只需要构造出$E|X_1 - X_2||X_1 - X_3|$的相合估计。 第二种是把σ_1^2看成一个新的参数，完全基于U统计量来构造。具体的核函数为：

$$\begin{aligned} h_1(x_1, x_2, x_3) &= (|x_1 - x_2||x_1 - x_3| + |x_2 - x_1||x_2 - x_3| + |x_3 - x_1||x_3 - x_2|)/3, \\ h_2(x_1, x_2, x_3, x_4) &= \frac{1}{4}(h_1(x_1, x_2, x_3) + h_1(x_1, x_2, x_4) + h_1(x_1, x_3, x_4) + h_1(x_2, x_3, x_4)) \\ &\quad + \frac{1}{3}(|x_1 - x_2||x_3 - x_4| + |x_1 - x_3||x_2 - x_4| + |x_1 - x_4||x_2 - x_3|) \end{aligned}$$

相应的两个估计为：

$$H_{1n} = \frac{1}{C_n^3} \sum_{i<j<k} h_1(X_i, X_j, X_k) - U_n^2; \tag{3.19}$$

$$H_{2n} = \frac{1}{C_n^4} \sum_{i<j<k<l} h_2(X_i, X_j, X_k, X_l). \tag{3.20}$$

在正常的条件下如$Var(h_1(X_i, X_j, X_k) < \infty, Var(h_2(X_i, X_j, X_k, X_l) < \infty$下，由前面的计算U统计量方差的表达式，我们知道两个都是σ_1^2的相合估计。结合Slutsky定理，我们可以构造出置信区间、假设检验、检验功效函数等。

3.6.2 检验对称性

在很多统计问题中，都会假设分布函数对称，这里我们介绍一个如何检验分布函数是否对称的方法。我们考虑连续函数，简单起见我们假定关于0点对称，构造参数：

$$\theta = \int (F(x) - (1 - F(-x)))^2 dF(x) = \int (F(x) + F(-x) - 1)^2 dF(x).$$

整理后，只需要构造

$$\int F(-x)dF(x), \int F(-x)^2 dF(x)$$

的估计, 我们可以用

$$I(X_1 + X_2 \leq 0), I(X_1 + X_2 \leq 0, X_1 + X_3 \leq 0)$$

分别估计。如果对称点未知的情形，我们可以用$X_1 - X_2, X_3 - X_4, X_5 - X_6$去替代上述的$X_1, X_2, X_3$. 类似的，我们可以构造统计量直接去检验分布函数：

$$H_0 : \ F(x) = F_0(x), \ \ vs \ \ H_1 : F(x) \neq F_0(x). \tag{3.21}$$

3.6.3 检验相关性

对于iid样本$(X_1, Y_1), \ldots, (X_n, Y_n)$;考虑

$$\theta = P(X_1 < X_2, Y_1 < Y_2) + P(X_2 < X_1, Y_2 < X_1) \tag{3.22}$$

如果$\theta > 1/2$,说明正相关；否则负相关；特别的如果$\theta = 1/2$,说明两者没有线性相关关系。

3.6.4 两样本的例子

检验两个分布函数F, G是否相等,我们可以构造参数

$$\theta = \int (F(x) - G(x))^2 dF(x) + \int (F(x) - G(x))^2 dG(x)$$

的估计，注意到

$$\begin{aligned}
&\int F(x)^2 dF(x) = \frac{1}{3}, \ \int G(x)^2 dG(x) = \frac{1}{3}, \\
&2\int F(x)G(x)dF(x) = \int G(x)dF(x)^2 = 1 - \int F(x)^2 dG(x) \\
&2\int F(x)G(x)dG(x) = \int F(x)dG(x)^2 = 1 - \int G(x)^2 dF(x)
\end{aligned}$$

所以

$$\theta = 2\int F(x)^2 dG(x) + 2\int G(x)^2 dF(x) - \frac{4}{3}$$

问题是构造$\int F(x)^2 dG(x)$的估计，对应的我们有：

$$EI(X_1 \le Y, X_2 \le Y) = \int F(x)^2 dG(x).$$

所以最后的核函数为：

$$\begin{aligned}
\phi(x_1, x_2, y_1, y_2) =& I(x_1 \le y_1, x_2 \le y_1) + I(y_1 \le x_1, y_2 \le x_1) \\
&+ I(x_1 \le y_2, x_2 \le y_2) + I(y_1 \le x_2, y_2 \le x_2) - \frac{4}{3},
\end{aligned}$$

对应的统计量为：

$$U_{n_1,n_2} = \frac{1}{C_{n_1}^2 C_{n_2}^2} \sum_{i_1<i_2} \sum_{j_1<j_2} \phi(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}).$$

相关的大样本性质：

$$\begin{aligned}
\sigma_{10}^2 &= Cov(\phi(X_1, X_2, Y_1, Y_2), \phi(X_1, X_3, Y_3, Y_4)) \\
\sigma_{01}^2 &= Cov(\phi(X_1, X_2, Y_1, Y_2), \phi(X_3, X_4, Y_1, Y_3)).
\end{aligned}$$

如果$Var(\phi(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2})) < \infty$, $n_1 + n_2 \to \infty$且 $n_1/(n_1 + n_2) \to p \in (0, 1)$,

$$\sqrt{n_1 + n_2}(U_{n_1,n_2} - \theta) \rightsquigarrow N(0, \sigma^2), \ \sigma^2 = \frac{m_1^2}{p}\sigma_{10}^2 + \frac{m_2^2}{1-p}\sigma_{01}^2. \tag{3.23}$$

3.6.5 多元情形

类似的，我们可以针对单样本情形下的μ, Σ和两样本的$\mu_1, \mu_2, \Sigma_1, \Sigma_2$，甚至进一步的分布函数相关的参数构造出相关的估计、假设检验等。

一般的，在假设检验中

$$H_0: \ \theta = \theta_0 \ \ vs \ \ H_1 : \theta \neq \theta_0. \tag{3.24}$$

我们构造损失函数$L(\theta, \theta_0) = (\theta - \theta_0)^2$的估计。我们可以由U统计量得到这个参数估计，进一步如果有渐近分布，那么就可以得到检验统计量等。我们可以看一个检验多元协方差矩阵的例子：

假定p维$X_1, \cdots, X_n$ 独立同分布来自于一个总体均值为0，协方差矩阵为Σ的分布，对于假设检验问题：

$$H_0 : \Sigma = \Sigma_0 \ \ vs \ \ H_1 : \Sigma \neq \Sigma_0. \tag{3.25}$$

我们可以构造$\theta = tr(\Sigma - \Sigma_0)^2$的U统计量估计

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} [(X_i' X_j)^2 - X_i' \Sigma_0 X_i - X_j' \Sigma_0 X_j + tr(\Sigma_0^2)]$$

Chapter 4

秩统计量与秩方法

4.1 秩(Rank)的定义

Definition 4.1.1 (秩). 对于互不相等的一组实数$x_1, \ldots, x_n$,x_k在从小到大的次序$x_{(1)} < \cdots < x_{(n)}$中所在位置$r_k$称为其秩；对应的对于样本$X_1, \ldots, X_n$,

$$R = (R_1, \ldots, R_n) \tag{4.1}$$

称为$(X_1, \ldots, X_n)$的秩统计量。

Remark 4.1.1. 对于单个秩统计量，我们有

$$R_i = \sum_{k=1}^{n} I(X_k \leq X_i),\ i = 1, \ldots, n.$$

对于一列连续独立样本$X_1, \ldots, X_N$, 其分布函数为F, 记$(R_1, \ldots, R_n)$为其秩统计量。这里我们考虑连续iid样本$X_1, \ldots, X_n$,其秩统计量的$R = (R_1, \ldots, R_n)$的分布及相关指标。

Theorem 4.1.1. 假定$X_1, \ldots, X_n$ *iid*来自于一个连续分布，以$R = (R_1, \ldots, R_n)$记样本$(X_1, \ldots, X_n)$的秩，则有对称性质，对于$(1, \ldots, n)$的任意一个置换$\pi(1, \ldots, n)$,有

$$P(R = \pi(1, \ldots, n)) = \frac{1}{n!}.$$

思考： 如果不是连续分布，定理结果会如何?

1. 对于所有的秩统计量：

$$P(R = \pi(1, \ldots, n)) = \frac{1}{n!}.$$

2. 对于单个的秩统计量，$R_1, \ldots, R_n$具有相同的分布，且

$$P(R_i = k) = (n-1)!/n! = \frac{1}{n},\ k = 1, \ldots, n.$$

3. 对于两个秩统计量 (R_i, R_j):

$$P(R_i = r, R_j = s) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

4. 对于期望，方差，协方差等：

$$\begin{aligned} &ER_i = \frac{n+1}{2}; \\ &Var(R_i) = \frac{(n+1)(n-1)}{12}; \\ &Cov(R_i, R_j) = \sum_{r \neq s} \frac{rs}{n(n-1)} - \frac{(n+1)^2}{4} = -\frac{(n+1)}{12}. \end{aligned}$$

Definition 4.1.2 (符号秩). 对于一组实数$x_1, \ldots, x_n$, 假定$|x_1|, \ldots, |x_n|$互不相等，记$\phi_i = I(x_i > 0)$, R_i^+ 为 $|x_i|$在$|x_1|, \ldots, |x_n|$中的秩， 则

$$R^+ = (\phi_1 R_1^+, \ldots, \phi_n R_n^+) \tag{4.2}$$

称为$(x_1, \ldots, x_n)$的符号秩。

对于符号秩，我们可以想象因为其涉及到符号，对于不同的分布F, ϕ取0或者1的情况完全不同，所以一般的符号秩统计量应该与F密切相关，下面定理考虑了一个特别的分布族。

Theorem 4.1.2. 若F连续且关于0对称，则

$$\phi_1, |X_1|, \ldots, \phi_n, |X_n|,$$

相互独立。进一步的$\phi_1, \ldots, \phi_n, (R_1^+, \ldots, R_n^+)$相互独立，且

$$\begin{aligned} &P(\phi_k = 0) = P(\phi_k = 1) = 1/2; \\ &P((R_1^+, \ldots, R_n^+) = \pi(1, \ldots, n)) = 1/n!. \end{aligned}$$

4.2 引例

Example 4.2.1 (Mann-Whitney 检验). 两个连续随机变量X, Y，其分布函数为F_1, F_2, 我们感兴趣的问题是检验参数$\theta = P(X < Y) = 1/2$. 其中观察样本为iid样本$X_1, \ldots, X_m$和$Y_1, \ldots, Y_n$. 从U统计量的角度，构造参数θ的U统计量：

$$U_n = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j).$$

这里核函数为：$\phi(x, y) = I(x < y)$. 对应的我们可以计算：

$$\begin{aligned}
\sigma_{11}^2 &= Var(I(X < Y)) = P(X < Y)(1 - P(X < Y)) = \theta(1 - \theta) < \infty^2; \\
\sigma_{01}^2 &= cov(I(X < Y), I(X < Y_1)) = \int (1 - F_2(x))^2 dF_1(x) - \theta^2; \\
\sigma_{10}^2 &= cov(I(X < Y), I(X_1 < Y)) = \int F_1(x)^2 dF_2(x) - \theta^2.
\end{aligned}$$

当$m + n \to \infty$ 且$m/(m + n) \to p \in (0, 1)$时候，由两样本$U$统计量的渐近结果，我们有：

$$\sqrt{m + n}(U_n - \theta) \rightsquigarrow N(0, \sigma^2), \ \sigma^2 = \frac{1}{p}\sigma_{10}^2 + \frac{1}{1 - p}\sigma_{01}^2.$$

所以再用一次U统计量的方法，可以构造出σ^2的相合估计$\hat{\sigma}^2$，然后由$Slutsky$定理，我们就有了

$$\frac{\sqrt{m + n}(U_n - \theta)}{\hat{\sigma}^2} \rightsquigarrow N(0, 1).$$

据此可以得到检验的接受域：$A = \{Z_{\alpha/2} \leq \frac{\sqrt{m+n}(U_n - 1/2)}{\hat{\sigma}^2} \leq z_{1-\alpha/2}\}$. 当然对于更平凡的$F_1 = F_2$,可以直接计算出$\sigma^2$,代入即可，这也是原始的$Mann$-$Whitney$检验的结果。

如果利用我们这里学习的秩统计量，如果θ太大或者太小，我们都有理由认为：

$$W_n = \sum_{k=1}^{m} R_k,$$

很大或者很小，其中$(R_1,\ldots,R_{m+n})$为样本$(X_1,\ldots,X_m,Y_1,\ldots,Y_n)$的秩统计量，注意这里我们只把$X_i$对应的秩做了相加(等价的也可以考虑把$Y_j$对应的秩相加)。对于秩统计量，我们有：

$$R_i=\sum_{k=1}^{m}I(X_k\le X_i)+\sum_{j=1}^{n}I(Y_j\le X_i),\ i=1,\ldots,m.$$

那么

$$\begin{aligned}W_n=\sum_{i=1}^{m}R_i&=\sum_{i=1}^{m}(\sum_{k=1}^{m}I(X_k\le X_i)+\sum_{j=1}^{n}I(Y_j\le X_i))\\&=\sum_{i=1}^{m}\sum_{j=1}^{m}I(X_i\le X_j)+\sum_{i=1}^{m}\sum_{j=1}^{n}I(Y_j\le X_i)\\&=\frac{m(m+1)}{2}+mn-\sum_{i=1}^{m}\sum_{j=1}^{n}I(X_i<Y_j).\end{aligned}$$

所以我们有：$W_n=\frac{m(m+1)}{2}+mn(1-U_n)$.

当然，由U_n的中心极限定理，我们有：

$$\sqrt{m+n}(1+\frac{m+1}{2n}-\frac{1}{mn}W_n-\theta)\rightsquigarrow N(0,\sigma^2),$$

这也就是经典的**Wilcoxon两样本秩和检验**。当然这里的渐近分布是由U统计量导出的，我们也可以直接考虑秩统计量的渐近中心极限定理。

4.3 同分布下的线性秩统计量

对于来自连续分布F的iid样本$(X_1,\cdots,X_N)$的Rank统计量$(R_1,\cdots,R_N)$,我们考虑一般的线性秩统计量：

$$S_N=\sum_{i=1}^{N}c_ia_N(R_i).$$

其中$c_1,\ldots,c_N$为任意的回归系数，"Regression Constants", a_N为得分函数，通常由一个定义在$(0,1))$上的函数$\phi(t)$生成，一般有下面两种形式：

1. $a_N(i)=\phi(\frac{i}{N+1}),\ i=1,\ldots,N.$

2. $a_N(i) = E\phi(U_{Ni})$, 其中U_{Ni}为均匀分布的N个iid样本的第i个次序统计量。

我们首先计算这个统计量的期望，方差:

Theorem 4.3.1. 记

$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i,\ \bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_N(i)$$

我们有

$$ES_n = \sum_{i=1}^{N} c_i E a_N(R_i) = N\bar{c}\bar{a};$$
$$Var(S_n) = \frac{1}{N-1}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2 \sum_{k=1}^{N}(c_i-\bar{c})^2.$$

Proof: 我们直接计算$a_N(R_i)$的期望，协方差：

$$Ea_N(R_i) = \frac{1}{N}\sum_{k=1}^{N} a_N(k) = \bar{a};$$
$$Var(a_N(R_i)) = \frac{1}{N}\sum_{k=1}^{N} a_N(k)^2 - \bar{a}^2 = \frac{1}{N}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2;$$
$$cov(a_N(R_i), a_N(R_j)) = \frac{1}{N(N-1)}\sum_{i\neq j} a_N(i)a_N(j) - \bar{a}^2 = -\frac{1}{N(N-1)}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2;$$

所以

$$\begin{aligned} ES_n &= \sum_{i=1}^{N} c_i E a_N(R_i) = N\bar{c}\bar{a}; \\ Var(S_n) &= \sum_{i,j} c_i c_j cov(a_N(R_i), a_N(R_j)) \\ &= \sum_{i=1}^{N} c_i^2 Var(a_N(R_i)) + \sum_{i\neq j} c_i c_j cov(a_N(R_i), a_N(R_j)) \\ &= \frac{1}{N}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2 \sum_{i=1}^{N} c_i^2 - \frac{1}{N(N-1)}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2 \sum_{i\neq j} c_i c_j \\ &= \frac{1}{N}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2 [\sum_{i=1}^{N} c_i^2 - \frac{1}{N-1}\sum_{i\neq j} c_i c_j] \\ &= \frac{1}{N-1}\sum_{k=1}^{N}(a_N(k)-\bar{a})^2 \sum_{k=1}^{N}(c_i-\bar{c})^2. \end{aligned}$$

Theorem 4.3.2 (Hajek, 1968)**.** *若c_i和$a(i)$满足条件M,N(类似无穷小条件),则*

$$\frac{S_n - ES_n}{\sqrt{Var(S_n)}} \rightsquigarrow N(0,1).$$

4.4 Examples

这里我们一般只考虑同分布情形下的线性秩统计量，所以其针对的一般就是原假设两个分布相同的情形。不同分布情形下的线性秩统计量可以参看Chernoff和Savage的工作。这部分内容对于考虑一个检验的功效或者在局部对立假设下的第二类错误等很适用。这里我们只看几个例子：

Example 4.4.1. *回到之前的Wilcolon-Mann-Whitney例子，如果$F_1 = F_2 = F$,我们可以计算出$\theta = 1/2$以及*

$$\begin{aligned} \sigma_{01}^2 &= \int (1-F(x))^2 dF(x) - \theta^2 = \frac{1}{12}; \\ \sigma_{10}^2 &= \int F_1(x)^2 dF_2(x) - \theta^2 = \frac{1}{12}. \end{aligned}$$

当$m+n\to\infty$ 且$m/(m+n)\to p\in(0,1)$时候，我们有：

$$\frac{m+n}{12}(\frac{1}{m}+\frac{1}{n})\to\sigma^2.$$

对于W_n的中心极限定理，

$$\sqrt{m+n}(1+\frac{m+1}{2n}-\frac{1}{mn}W_n-1/2)/\sqrt{\frac{m+n}{12}(\frac{1}{m}+\frac{1}{n})}\rightsquigarrow N(0,1),$$

我们可以整理得到：

$$\frac{W_n-\frac{m(N+1)}{2}}{\sqrt{\frac{mnN}{12}}}\rightsquigarrow N(0,1).$$

另一方面，从刚刚定义的线性秩统计量出发，我们有：$a(i)=i$和$c_i=I(i\le m)$,

$$\begin{aligned}&\bar{a}=\frac{N+1}{2},\bar{c}=\frac{m}{N};\\&\sum_{k=1}^{N}(a(k)-\bar{a})^2=\frac{N(N^2-1)}{12};\\&\sum_{k=1}^{N}(c_i-\bar{c})^2=\frac{mn}{N}.\end{aligned}$$

所以

$$\begin{aligned}&ES_n=N\frac{N+1}{2}\frac{m}{N}=\frac{m(N+1)}{2};\\&Var(S_n)=\frac{mn(N+1)}{12}\end{aligned}$$

代入同分布情形下的线性秩统计量，我们有

$$\frac{W_n-\frac{m(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}}\rightsquigarrow N(0,1).$$

与从U统计量出发推导得到的中心极限定理一致的。

当然，我们也没有必要直接用秩本身，而是用其函数形式来替代，后者直观上会更具有灵活性，这也是引入线性秩统计量的原因。这里我们给几个检验的例子：

1. Fisher-Yates 检验：

$$a_N(i) = EX_{Ni},$$

其中X_{Ni}为从$N(0,1)$中抽取的iid样本$X_1, \ldots, X_N$的次序统计量$X_{N1} \leq \ldots X_{NN}$.

2. Van Der Waerden检验：

$$a_N(i) = \Phi^{-1}(\frac{i}{N+1}),$$

其中Φ为标准正态分布的分布函数，所以这里其实取的是其分位点。

一般的，对于一个严格增的连续分布函数Q,我们都可以考虑这样两种计分函数：

$$a_N(i) = Q^{-1}(\frac{i}{N+1}),$$

和

$$a_N(i) = EX_{Ni},$$

其中X_{Ni}为从Q中抽取的N个iid样本的次序统计量.

问题： 取什么样的Q适合呢?

Example 4.4.2. 我们考虑刻度检验问题：

$$X_1, \ldots, X_m \sim F(x),\ Y_1, \ldots, Y_n \sim F(x/\theta). \tag{4.3}$$

检验 $H_0 : \theta = 1$.

注意Y分布上等价于θX, 所以联合样本等价于：

$$(X_1, \ldots, X_m, \theta X_{m+1}, \ldots, \theta X_{m+n}).$$

当$\theta > 1$时候，Y_j对应的秩统计量会集中在两端。进一步如果F不是0左右对称的，如$X_1, \ldots, X_m$中正的比较多，那么Y_j的秩都会比较大。 反之如

果$\theta < 1$,那么X_i的都会在两段，Y_j的秩集中到中间一段。综上，检验统计量可以设置为：

$$W_n = \sum_{k=1}^{m} a_N(R_k),$$

其中$a_N(k)$应该满足两段比较大，中间比较小的特点。几个相关的检验统计量：

- Mood检验：

$$a_N(i) = (i - \frac{N+1}{2})^2;$$

- Ansary-Bradly 检验：

$$a_N(i) = |i - \frac{N+1}{2}|;$$

- Copan 检验和Klotz检验：

$$a_N(i) = EX_{Ni}^2 \ or \ a_N(i) = [Q^{-1}(\frac{i}{N+1})]^2;$$

- Siegel-Turkey 检验：

$$N, N-3, N-4, \ldots, N-2, N-1.$$

讨论： 什么样检验统计量好?

Chapter 5

•

3.1第四章 核估计

5.1 目标

在统计问题中，如何估计一条曲线$h(x)$?

- 例如 $X_1, \cdots, X_n$ iid $\sim F$, 我们如何估计其分布函数$F(x)$和密度函数$f(x) = F'(x)$?
- 对于回归问题(X, Y), 在均方损失下，我们有

$$\begin{aligned} E(Y - f(X))^2 &= E_x E[Y^2 - 2Y f(X) + f^2(X)|X] \\ &E_x(E[Y^2|X] - (E[Y|X])^2 + (E[Y|X])^2 - 2E[Y|X]f(X) + f^2(X)) \\ &= E_x(E[Y^2|X] - (E[Y|X])^2 + E_x(E[Y|X] - f(X))^2 \end{aligned}$$

所以回归函数为

$$r(x) = E[Y|X = x]. \tag{5.1}$$

给定iid样本$(X_1, Y_1), \cdots, (X_n, Y_n)$时候如何估计回归函数$r(x)$?

进一步的，如果我们得到了一个估计函数$\hat{h}(x)$,如何衡量所得到估计的好坏?

5.2 介绍

给定一组iid的样本$X_1, \ldots, X_n$, 我们可以用经验分布函数去刻画其分布函数$F(x)$:

Definition 5.2.1 (Empirical distribution function)**.** 我们可以定义经验分布函数*(Empirical distribution function),*

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I(X_k \le x). \tag{5.2}$$

作为$F(x)$的无偏估计，EDF有如下的性质：

1. For any fixed x,

$$E(\hat{F}_n(x)) = F(x), \text{ and } Var(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}. \tag{5.3}$$

Further, if $0 < F(x) < 1$, we have

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \rightsquigarrow N(0, F(x)(1-F(x)). \tag{5.4}$$

2. Gilvenko-Cantelli Theorem:

$$\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0. \tag{5.5}$$

3. Dvoretzky–Kiefer–Wolfowitz (DKW) inequality: For any $\epsilon > 0$,

$$P(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \le 2e^{-2n\epsilon^2}. \tag{5.6}$$

Theorem 5.2.1. 记

$$L(x) = \max\{F_n(x) - \epsilon_n, 0\},\ R(x) = \min\{F_n(x) + \epsilon_n, 1\},$$

其中 $\epsilon_n = \sqrt{\log(2/\alpha)/2n}$, 我们有

$$P(L(x) \le F(x) \le R(x),\ \forall\ x) \ge \alpha. \tag{5.7}$$

5.2.1 密度函数的估计

我们知道对于很多随机变量，我们更多时候是以其密度函数来判断或者刻画分布的。那么，如何通过iid样本来估计或者刻画这组数据的密度函数$f(x)$?

- 我们可以仿照之前的V统计量，现在要估计$F'(x)$,那么我们用经验分布函数的导数$F'_n(x)$来估计，我们有：

$$F'_n(x) = \frac{1}{n}\sum_{k=1}^{n}\delta_{X_k}(x)$$

这里的δ是 Dirac delta 函数：

$$\delta_x(t) = \begin{cases} \infty & \text{if } t = x; \\ 0 & \text{if } t \neq x. \end{cases}$$

并且满足$\int \delta_x(t)dt = 1$.

- 直方图.如果确定直方图的结点$t_0 < t_1 < \cdots < t_{m-1} < t_m$, 定义一个新的函数

$$\gamma_x(t) = \sum_{k=1}^{m} I(t_{k-1} < x, t \leq t_k)$$

那么计数直方图可以表示为：

$$H_n(x) = \sum_{i=1}^{n}\gamma_{X_i}(x). \tag{5.8}$$

如果我们想展示密度函数，可以定义

$$\gamma_x(t) = \sum_{k=1}^{m} I(t_{k-1} < x, t \leq t_k)/(t_k - t_{k-1})$$

然后直方图表示为：

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n}\gamma_{X_i}(x). \tag{5.9}$$

实际上，这里的

$$\gamma_x(t) = \begin{cases} \frac{1}{t_k - t_{k-1}} & t_{k-1} < t \le t_k; \\ 0 & \text{others.} \end{cases}$$

这里对于某个k, $t_{k-1} < x \le t_k$, 且满足$\int \gamma_x(t)dt = 1$. 对于一般直方图，我们会选取一个起点h_0及固定的宽度δ,而实际问题中这两个参数的选取对于最后的结果影响也是很大的。

5.3 核密度估计 Kernel Density Estimation

5.3.1 Motivations

仔细检查直方图，我们会发现直方图对于所有发生在$(t_{k-1}, t_k]$中的样本都同等的对待。简单起见我们考虑$[0,1]$区间，如果有两个样本取值$0^+, 1^-$,那么我们应该认为前者表达的信息是分布在0周围有密度，而后者应该是在1周围有密度，而不能把两个完全同等的看待。

因此，对于直方图，给定宽度δ, 我们认为样本X_i应该反应区间$[X_i - \delta/2, X_i + \delta/2]$的信息，例如我们定义

$$K_\Delta(x) = I(-\delta/2 \le x \le \delta/2)/\delta.$$

那么对应的估计为：

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_\delta(X_i - x). \tag{5.10}$$

再进一步我们把宽度δ看成参数，直接定义

$$K(x) = I(|x| \le 1)/2.$$

然后估计为：

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \tag{5.11}$$

当然这里我们取的是同等权重，如果考虑一般的权，那么我们也就得到了一般的核密度估计形式：

Definition 5.3.1 (Kernel density Estimate)**.**

$$\hat{f}_h(x) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{h} K(\frac{X_t - x}{h}) = \int K_h(\mu - x)d\hat{F}(\mu), \tag{5.12}$$

这里$K_h(\cdot) = K(\cdot/h)/h$称为核函数, h 是 窗宽***(Bandwidth)*** 参数。

5.3.2 核函数 Kernel function

What is the condition that K should satisfy?

1. $K(x) \geq 0$ for all x;
2. $\int K(x)dx = 1$.

Therefore, $K(\cdot)$ is a density function and any density function can act as a kernel function in theory.

Actually, a kernel function is usually a nonnegative symmetric, unimodal probability density function.

Commonly used kernel functions:

- Uniform

$$K(\mu) = \frac{1}{2}I(|\mu| \leq 1); \tag{5.13}$$

- Triangular

$$K(\mu) = (1 - |\mu|)I(|\mu| \leq 1); \tag{5.14}$$

- Epanechnikov:

$$K(\mu) = \frac{3}{4}(1 - \mu^2)I(|\mu| \leq 1); \tag{5.15}$$

- Quartic (biweight):

$$K(\mu) = \frac{15}{16}(1 - \mu^2)^2 I(|\mu| \leq 1); \tag{5.16}$$

- Triweight:

$$K(\mu) = \frac{35}{32}(1 - \mu^2)^2 I(|\mu| \leq 1); \tag{5.17}$$

- Tricube:

$$K(\mu) = \frac{70}{81}(1 - |\mu|^3)^3 I(|\mu| \leq 1); \tag{5.18}$$

- Gaussian:

$$K(\mu) = \frac{1}{\sqrt{2\pi}} exp(-\mu^2/2); \tag{5.19}$$

- Cosine:

$$K(\mu) = \frac{\pi}{4} cos(\frac{\pi}{2}\mu) I(|\mu| \leq 1); \tag{5.20}$$

- Logistic:

$$K(\mu) = \frac{1}{e^{\mu} + 2 + e^{-\mu}} \tag{5.21}$$

Q: how to calculate the constant part such as $\frac{35}{32}$?

5.3.3 Bandwidth

Q: Why do we use the thresholding 1 not 2 or other parameters?

Remark 5.3.1. *If $K(\mu)$ is a kernel function, for any $h > 0$,*

$$\frac{1}{h} K(\frac{\mu}{h}) \quad or \quad hK(h\mu) \tag{5.22}$$

can both serve as the kernel functions.

Notes on the bandwidth

- A large bandwidth h-bias, over-smooth.
- A small bandwidth h-large variance.
- Optimal bandwidth should achieve a trade-off between bias and variance.

Conclusion: "It is well-known both empirically and theoretically that the choice of kernel functions is not very important to the kernel density estimator. As long as they are symmetric and unimodal, the resulting kernel density estimator performs nearly the same when the bandwidth h is optimally chosen. "

5.4 窗宽选择Bandwidth Selection

为了估计样本的密度函数$f(x)$, 我们构造了核估计

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_i - x}{h}) = \int K_h(\mu - x)d\hat{F}(\mu), \tag{5.23}$$

作为一个估计，我们可以考虑估计相关的性质。

5.4.1 期望 Expectation

我们直接计算核估计的期望：

$$\begin{aligned} E\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}E\frac{1}{h}K(\frac{X_i - x}{h}) &= E\frac{1}{h}K(\frac{X_1 - x}{h}) \\ &= \int \frac{1}{h}K(\frac{t-x}{h})f(t)dt \\ &= \int K(t)f(x+ht)dt. \end{aligned}$$

所以，一般来说，核估计是有偏的，因为

$$\int K(\mu)f(x+h\mu)d\mu \neq f(x).$$

我们可以看下不同核下的期望：

1. Uniform $K(\mu) = \frac{1}{2}I(|\mu| \leq 1)$, 我们有

   $$E\hat{f}_h(x) = \int K(t)f(x+ht)dt = \frac{1}{2}\int I(|t| \leq 1)f(x+ht)dt = \frac{1}{2}\int_{-1}^{1} f(x+ht)dt.$$

   即从$[x-h, x+h]$上的平均密度。窗宽度h决定选取的区间宽度，直观上窗宽越小误差越小，但是方差越大。反之选取的区间越大，估计的偏差越大，但是波动性越小。

2. Gaussian: $K(\mu) = \frac{1}{\sqrt{2\pi}}exp(-\mu^2/2)$,

   $$E\hat{f}_h(x) = \int K(t)f(x+ht)dt = \frac{1}{2\pi}\int exp(-t^2/2)f(x+ht)dt = Ef(x+hY).$$

   这里的$Y \sim N(0,1)$. 即整个实数区间上的加权平均，窗宽也有类似现象。

5.4.2 方差 Variance

注意到核估计是独立同分布随机变量的样本平均，所以

$$\begin{aligned} nVar(\hat{f}_h(x)) &= Var(\frac{1}{h}K(\frac{X_1 - x}{h})) \\ &= \int \frac{1}{h^2}K^2(\frac{t-x}{h})f(t)dt - (\int K(t)f(x+ht)dt)^2 \\ &= \frac{1}{h}\int K^2(t)f(x+ht)dt - (\int K(t)f(x+ht)dt)^2 \end{aligned}$$

所以估计的方差为：

$$Var(\hat{f}_h(x)) = \frac{1}{nh}\int K^2(t)f(x+ht)dt - \frac{1}{n}(\int K(t)f(x+ht)dt)^2$$

类似的我们可以看下Uniform核的情况：当$K(\mu) = \frac{1}{2}I(|\mu| \leq 1)$, 我们有

$$Var(\hat{f}_h(x)) = \frac{1}{2nh}\int_{-1}^{1} f(x+ht)dt - \frac{1}{n}(\int_{-1}^{1} f(x+ht)dt)^2$$

5.4.3 均方损失 Mean Square Error

因为核估计是有偏差的，所以我们考虑均方损失来衡量估计的好坏：

$$\begin{aligned} MSE(x) &= E(\hat{f}_h(x) - f(x))^2 \\ &= Var(\hat{f}_h(x)) + (f(x) - E(\hat{f}_h(x)))^2 \\ &= \frac{1}{nh}\int K^2(t)f(x+ht)dt - \frac{1}{n}(\int K(t)f(x+ht)dt)^2 \\ &\quad + (f(x) - \int K(t)f(x+ht)dt)^2 \end{aligned}$$

5.4.4 窗宽选取

当样本$n \to \infty$时候，注意到核估计是iid的平均，所以经典CLT可以给出其渐近分布，这里我们具体来看均方损失的极限状况。

一般窗宽$h \to 0$, 把$f(x+ht)$在x处做Taylor展开：

$$f(x+ht) = f(x) + f'(x)ht + \frac{1}{2}f''(x)h^2t^2 + O(h^3t^3), \tag{5.24}$$

我们有：

$$\frac{1}{h}\int K^2(t)f(x+ht)dt = \frac{1}{h}\int K^2(t)(f(x)+f'(x)ht+\frac{1}{2}f''(x)h^2t^2+O(h^3t^3))dt$$
$$= \frac{f(x)}{h}\int K^2(t)dt + \frac{h}{2}f''(x)\int K^2(t)t^2dt + \cdot$$

类似的

$$\int K(t)f(x+ht)dt = \int K(t)(f(x)+f'(x)ht+\frac{1}{2}f''(x)h^2t^2+O(h^3t^3))dt$$
$$\approx f(x)+\frac{h^2}{2}f''(x)\int K(t)t^2dt$$

所以

$$MSE(x) \approx \frac{f(x)}{nh}\int K^2(t)dt + \frac{1}{4}(f''(x))^2h^4(\int t^2K(t)dt)^2,$$

这里$h \to 0$, 所以只有主项。如果我们为了让在x处核估计的MSE最小，优化上面的MSE即可。但是一般的这里我们考虑的是一个密度函数的估计，所以不能仅仅从某一点x来判断h的选取。为此，我们引入 **Mean Integrated Square Error** (MISE):

$$\begin{aligned}\text{MISE} &= E\int(\hat{f}_h(x)-f(x))^2dx\\ &\approx \int[\frac{f(x)}{nh}\int K^2(t)dt+\frac{1}{4}(f''(x))^2h^4(\int t^2K(t)dt)^2]dx\\ &\approx \underline{\frac{1}{nh}\int K^2(t)dt}+\frac{h^4}{4}\int(f''(x))^2dx(\int t^2K(t)dt)^2.\end{aligned}$$

最小化MISE，我们得到最优窗宽(实际上是渐近最优窗宽)

$$h_{opt} = \arg\min_h \text{MISE} = n^{-1/5}(\int(f''(x))^2dx)^{-1/5}(\int K^2(t)dt)^{1/5}(\int t^2K(t)dt)^{-2/5}. \tag{5.25}$$

记

$$\|g\|_2^2 = \int g^2(x)dx, \quad \mu_2(K) = \int t^2K(t)dt, \tag{5.26}$$

最优窗宽为：

$$h_{opt} = (\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2})^{1/5} \tag{5.27}$$

此时最小的MISE为：

$$\begin{aligned} MISE_{opt} &= \frac{5}{4}\frac{\|K\|_2^2}{n}(\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2})^{-1/5} \\ &= \frac{5}{4}(\frac{\|K\|_2^2}{n})^{4/5}(\mu_2^2(K)\|f''\|_2^2)^{1/5}. \end{aligned} \tag{5.28}$$

5.4.5 窗宽的经验选取

对于正态核函数：

$$h_{opt} \approx 1.06\hat{\sigma}n^{-1/5}.$$

对于Epanechnikov核：

$$h_{opt} \approx 2.34\hat{\sigma}n^{-1/5}.$$

这里的样本协方差是从$\int(f''(x))^2dx$项得到的，对于正态分布，两次导数得到的是方差。

5.4.6 最优核 Optimal Kernel function

这一节我们比较核函数对于核估计的影响。从最优MISE的结果，我们发现：

$$\begin{aligned} MISE_{opt} &= \frac{5}{4}\frac{\|K\|_2^2}{T}(\frac{\|K\|_2^2}{T\mu_2^2(K)\|f''\|_2^2})^{-1/5} \\ &= \frac{5}{4}(\frac{\|K\|_2^2}{T})^{4/5}(\mu_2^2(K)\|f''\|_2^2)^{1/5}. \end{aligned} \tag{5.29}$$

所以最好的核函数应该是：

$$K_{opt} \in \arg\min(\|K\|_2^2)^2(\mu_2(K)) = (\int K^2(\mu)d\mu)^2(\int \mu^2K(\mu)d\mu) := \beta(K). \tag{5.30}$$

In addition, K_{opt} should satisfy

$$\int K(\mu)d\mu = 1, \quad \int \mu K(\mu)d\mu = 0. \tag{5.31}$$

Note the fact

$$\beta(K(x)) = \beta(\frac{1}{h}K(\frac{x}{h})). \tag{5.32}$$

Therefore, if $K_{opt}(\cdot)$ is an optimal kernel, $\frac{1}{h}K(\frac{\cdot}{h})$ is also an optimal kernel. 所以我们只需要最小化

$$\int K^2(t)dt,$$

$K(t)$需要满足条件

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2K(t)dt = 1. \tag{5.33}$$

记Epanechnikov核

$$K_0(t) = \frac{3}{4}(1-t^2)I(|t| \le 1); \tag{5.34}$$

我们考虑函数$\delta(t) = K(t) - K_0(t)$,我们有

$$\int t^m\delta(t)dt = 0, \ m = 0, 1, 2.$$

所以

$$\int (1-t^2)\delta(t)dt = 0.$$

现在考虑：

$$\begin{aligned}\frac{4}{3}\int \delta(t)K_0(t)dt &= \int_{|t|\le 1} \delta(t)(1-t^2)dt \\ &= -\int_{|t|>1} \delta(t)(1-t^2)dt \\ &= \int K(t)(t^2-1)dt \ge 0.\end{aligned}$$

所以

$$\int K(t)^2dt = \int K_0^2(t)dt + 2\int \delta(t)K_0(t)dt + \int \delta^2(t)dt \geq \int K_0^2(t)dt.$$

因此Epanechnikov核为最优核，或者更一般的，最优核函数为：

$$K_{opt}(t) = \frac{3}{4\alpha}(1 - t^2/\alpha^2)I(|t| \leq \alpha); \tag{5.35}$$

5.5 密度函数相合估计的应用

对于一般分布样本分位数的渐近分布，由Delta方法，会出现密度函数的估计： 对于某一个样本分位数：

$$\sqrt{n}(\epsilon_{np} - F^{-1}(p)) \xrightarrow{d} N(0, p(1-p)/(f(F^{-1}(p)))^2). \tag{5.36}$$

例如中位数， $p = 1/2,$

$$\sqrt{n}(\epsilon_{1/2} - F^{-1}(1/2)) \xrightarrow{d} N(0, 1/4(f(F^{-1}(1/2)))^2). \tag{5.37}$$

从假设检验或者置信区间的角度，我们需要密度函数的相合估计，这里我们可以采用核估计$\hat{f}(x)$.

5.6 核估计的延伸

- 边界效应
- 多维核估计

Chapter 6

非参数回归

给定一组iid样本$(X_i, Y_i),\ i = 1, \cdots, n$, 我们考虑如何基于$X_i$解释应变量$Y_i$. 一般的, 如果有一个新的观测值$X_0$, 如何预测$\hat{Y}_0$? 可以考虑多维的情形$X_i \in \mathbb{R}^p$, 应变量$Y_i$理论上也可以是多维的，这里简单起见我们假定为一维的$Y_i \in \mathbb{R}$.

从统计模型的角度来说，我们希望找一个映射函数$r(\cdot) : \mathbb{R}^p \to \mathbb{R}$, 使得

$$Y_i \approx r(X_i).$$

可以考虑非参数回归模型 :

$$Y_i = r(X_i) + \epsilon_i,$$

其中ϵ_i为独立同分布的误差项，均值为0，方差为未知的σ^2 (方差的大小代表了模型的噪声大小，或者模型的解释性大小)。

如果我们从均方损失的角度来考虑，可以得到

$$r(x) = \arg\min_f E(Y_0 - f(X_0))^2 = E(Y|X = x).$$

一般称 $r(x) = E(Y|X = x)$为回归函数。本章主要讨论估计回归函数的一般方法。

6.1 线性模型回顾

把回归模型用矩阵形式表示：

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}, X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}_{n\times p}. \tag{6.1}$$

为了表达出常数项，可以把X的第一列设定为1，即$X_{i1} = 1$. 记误差项为$\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$,回归系数为 $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$‘, 那么回归模型可以表达为：

$$Y = X\beta + \epsilon.$$

对于任意的回归系数，可以用均方损失(Mean Squared Error, MSE) 来衡量模型拟合的好坏：

$$MSE(\beta) = \frac{1}{n}(Y - X\beta)'(Y - X\beta), \tag{6.2}$$

通过优化MSE, 我们可以得到最小二乘估计(Least Square Estimation，LSE)

$$\hat{\beta} = \arg\min_{\beta} MSE(\beta) = (X'X)^{-1}X'Y, \tag{6.3}$$

这里的前提是$X'X$满秩，否则最优解从代数的角度来说应该是一个线性空间。

注意LSE是基于观察样本得到的，等价的来说就是对于观察样本拟合的最好的线性模型。如果基于所得的LSE，我们再计算MSE，可以得到

$$MSE(\hat{\beta}) = \frac{1}{n}Y'(I_n - X(X'X)^{-1}X')Y = \frac{1}{n}\epsilon'(I_n - X(X'X)^{-1}X')\epsilon.$$

这里I_n为$n \times n$ 的单位矩阵。 如果把X看作固定的(例如在实验设计中，在指定条件下观察被解释变量), 对于上式最小MSE求期望，可以得到

$$E\frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \frac{tr(I_n - X(X'X)^{-1}X')}{n}\sigma^2 = \frac{n-p}{n}\sigma^2. \tag{6.4}$$

Remark 6.1.1. 这里的结果直接反映了在回归模型中样本个数n和解释变量的维度p对于回归模型的影响. 总的来说，样本越多，模型拟合的越好。

对于得到的估计$\hat{\beta}$, 理论上我们应该在新的样本集上来看估计的好坏，即我们应该考虑

$$E(Y_0 - X_0'\hat{\beta})^2. \tag{6.5}$$

类似的，如果把X当成非随机的，我们有

$$\begin{aligned} E(Y_0 - X_0'\hat{\beta})^2 &= E(X_0'\beta + \epsilon_0 - X_0'\hat{\beta})^2 \\ &= \sigma^2 + X_0' E(\beta - \hat{\beta})(\beta - \hat{\beta})' X_0 \\ &= \sigma^2 + \frac{1}{n} X`_0 (X'X)^{-1} X_0. \end{aligned}$$

Remark 6.1.2. 仔细比较*6.4* 和 *6.5,* 前者一般称为基于训练数据集的损失 *training error*，后者是基于测试集的损失*(*也是模型的真实损失*) test error*。直观上，因为后者不可得，可以把前者当成后者的一个样本估计。但是因为模型本身也是基于训练数据得到的，所以前者一般都是比真实的损失要小的。例如这里的结果，*6.4*的结果小于σ^2甚至比最优的知道真实β还要好。对于*6.5,* 除了固定项σ^2, 剩下的正好是我们用样本估计真实系数β带来的损失。

Remark 6.1.3. 损失函数是统计学的核心。对于任何的统计决策，都应该有对应的损失函数来衡量决策的好坏。真实的损失很难去计算，基于训练数据集本身得到的损失函数一定要慎用！实际问题中，一般是对训练数据再做一次重抽样*(Resample)*来构造出真实损失的估计, 例如*Leave-One-Out, Bootstrapping*和分层*k-Folds*等方法。

6.2 引例-逻辑回归

接下来，我们通过一个稍微复杂的逻辑回归来阐述一般的回归函数。对于线性回归模型，一般适用于被解释变量是连续的情形。如果Y的取值本身是离散的，简单的线性回归模型就不再适用(因为其给出的估计一般不是离散的), 这时我们用逻辑回归。

假定$Y \in \{0, 1\}$, 即有两个类别， 考虑回归函数

$$r(x) = E\{I(Y = 1)|X = x\} = P(Y = 1|X = x) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}, \tag{6.6}$$

回归函数对应的是一个0-1之间的数，反映了应变量取1的概率大小。当然也可以考虑其他形式的函数，可以参看logit或者probit模型。

下面我们看下均方损失函数，

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \hat{Y}_i),$$

恰好就是分类问题中错分的比例，即错分率。

如果我们设定预测为$\hat{Y} = I(r(X) > 0.5) = I(X'\beta > 0)$, 那么MSE损失为：

$$MSE = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq I(X_i'\beta > 0) = \frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i'\beta \leq 0),$$

这里$2Y_i - 1$把原来的$\{0, 1\}$转化为了$\{+1, -1\}$.

如果把Y_i生成机制设定为概率为$r(X_i)$的二项分布，我们可以写出似然函数

$$L(\beta) = \prod_{i=1}^{n} r(X_i)^{Y_i}(1 - r(X_i))^{1-Y_i} = \prod_{i=1}^{n} \frac{e^{Y_i X_i'\beta}}{1 + e^{X_i'\beta}},$$

逻辑回归估计为

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} \frac{e^{Y_i X_i'\beta}}{1 + e^{X_i'\beta}} = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} \{Y_i X_i'\beta - \log(1 + e^{X_i'\beta})\}, \quad (6.7)$$

优化过程没有显示解，一般通过迭代的过程来完成。

相比于线性回归函数，这里的回归函数形式就要复杂很多

$$r(x) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}. \quad (6.8)$$

Remark 6.2.1. 对于真实数据，两个损失函数很难比较，因为不知道真实的数据生成机制。

如果仿照线性回归，优化MSE，我们也可以得到一个估计：

$$\hat{\beta}_1 = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i'\beta \leq 0) = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i'\beta > 0)$$

与逻辑回归相比，这里的损失函数不连续，即$\hat{\beta}_1$一般不唯一。直观上，所得的估计平面$\hat{\beta}_1$有微小变动，损失函数应该是不变的。从这个角度来说逻辑回归更合适一点。

6.3 局部光滑-Local Smooth

前面介绍的线性回归和逻辑回归，都是把回归曲线的估计转化为参数估计的问题，下面介绍两种可以处理一般曲线的局部光滑方法。

- Regressongram: 类似于密度函数中的直方图Histogram, 在回归函数估计问题中我们也可以把解释变量$X_1, \cdots, X_n$所在的全部空间分成m组$B_1, \cdots, B_m$. 对于任意的x, 一定存在一个组使得$x \in B_j$, 定义回归函数为

$$\hat{r}(x) = \frac{\sum_{i:\ X_i \in B_j} Y_i}{\#\{i:\ X_i \in B_j\}}\text{fi} \tag{6.9}$$

  即把每一组内的Y_i的平均作为这一组内回归函数的取值。

- Local Average: 在Regressongram中，我们需要确定组数，而且组数对结果影响应该很大。另外确定好组数之后，如何分组也是一个问题。我们可以换一种方式来估计回归函数。对于任意的x, 我们选择一个邻域$B_x = \{y : \|y - x\| \le h\}$，定义回归函数

$$\hat{r}(x) = \frac{\sum_{i:\ X_i \in B_x} Y_i}{\#\{i:\ X_i \in B_x\}}. \tag{6.10}$$

  如果选择一般的距离函数，理论上我们可以处理多维的情形，这里的h可以视为窗宽，其大小决定我们用x周围多少的邻居来估计回归函数$r(x)$.

6.4 非参数核估计

在刚刚介绍的几种回归模型中，对于新的X_0, 预测$\hat{Y}_0$ 都是当前已有观察值$Y_1, \cdots, Y_n$的一个线性组合，即我们有linear smooth的概念：

Definition 6.4.1. 基于当前样本$(X_1, Y_1), \cdots, (X_n, Y_n)$, 对于新观察值$X_0$的预测为：

$$\hat{Y}_0 = L(X_0, X_1, \cdots, X_n)Y. \tag{6.11}$$

其中$L(X_0, X_1, \cdots, X_n) = (w_1, \cdots, w_n)'$.

例如，

- 线性模型：

$$\hat{Y_0} = X_0'(X'X)^{-1}X'Y. \tag{6.12}$$

这里$L(X_0, X_1, \cdots, X_n) = X_0'(X'X)^{-1}X'$. 这里的一个有趣现象是，如果对于一元样本$(X_1, Y_1),\ (X_n, Y_n))$, 我们考虑一元回归

$$Y = \alpha + \beta X + \epsilon, \tag{6.13}$$

新的预测为当前$Y_1, \cdots, Y_n$的线性组合。我们如果考虑更复杂的多项式回归

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon, \tag{6.14}$$

预测还是当前$Y_1, \cdots, Y_n$的线性组合，只是权重计算方式不同。

- Regressongram回归： 如果把X_i做一个排序，那么L的形式为

$$L(X_0, X_1, \cdots, X_n) = (0, \cdots, 0, \frac{1}{k}, \cdots, \frac{1}{k}, 0, \cdots, 0).$$

- Local Average回归: L的形式同上式，只是计算方式不同.

注意在后两者的形式中，所得权重都有一般的$\sum_{i=1}^n w_i = 1,\ w_i \geq 0$， 但是线性模型没有这一约束。

下面我们介绍一般的Non-Parametric Kernel Regression

Definition 6.4.2 (Nadaraya–Watson kernel estimator)**.** 对于样本$(X_1, Y_1), \cdots, (X_n, Y_n)$, 定义回归函数为：

$$\hat{r}(x) = \frac{\sum_{i=1}^n K(\frac{x - X_i}{h}) Y_i}{\sum_{i=1}^n K(\frac{x - X_i}{h})}, \tag{6.15}$$

这里的$K(x), h$是核函数和窗宽*(bandwidth)*. 如果记

$$\omega_i = \frac{K(\frac{x - X_i}{h})}{\sum_{j=1}^n K(\frac{x - X_j}{h})},$$

那么回归函数估计为$\hat{r}(x) = \sum_{i=1}^n \omega_i Y_i$, 而且权重$\sum_{i=1}^n \omega_i = 1,\ \omega_i \geq 0$.

和密度函数的核估计一样，相对于窗宽这里的核函数选取没有那么的重要。理论上，我们也可以在模型假设下去计算MSE、MISE，然后优化损失函数得到最优窗宽表达式。对于回归函数来说，我们可以通过重抽样等交叉验证的方式得到损失函数形式，进而选择窗宽和核函数。

6.5 Local Polynomial Regression

我们知道样本均值

$$\bar{Y} = \arg\min_a \frac{1}{n}\sum_{i=1}^{n}(Y_i - a)^2, \tag{6.16}$$

这意味着如果没有任何的解释变量，我们预测下一个观察值，那么就应该是样本均值$\bar{Y}$. 我们考虑一个加权的MSE

$$\frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_j}{h})(Y_i - a)^2, \tag{6.17}$$

这里核函数的引入是为了突出我们希望在x周围损失尽量小一点，得到的估计正好就是Nadaraya-Watson估计

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})} = \arg\min \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h})(Y_i - a)^2. \tag{6.18}$$

我们可以发现，核估计是在局部用一个常数去拟合样本。当然，我们可以考虑稍微复杂一点的，例如

$$\arg\min \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h})(Y_i - a - bX_i)^2 \tag{6.19}$$

或者更一步的，一般的局部多项式

$$\arg\min \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h})(Y_i - a - b_1X_i - \cdots - b_kX_i^k)^2. \tag{6.20}$$

记

$$\omega_i = \frac{K(\frac{x-X_i}{h})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})},$$

和权重矩阵$W_x = diag\{\omega_1, \cdots, \omega_n\}$. 根据多项式形式写出对应的设计矩阵$X$，那么我们有均方损失表达式

$$\frac{1}{n}(Y - X\beta_x)'W_x(Y - X\beta_x), \tag{6.21}$$

进而得到加权最小二乘估计

$$\hat{\beta}_x = (X'W_xX)^{-1}X'W_xY.$$

注意这里的权重矩阵W_x和回归系数β_x都是和关注的点x相关的，最后回归函数为

$$\hat{r}(x) = \hat{\beta}_x' l(1, x, x^2, \cdots, x^k).$$

回归函数中除了窗宽h, 还有一个多项式阶数k需要确定。一般的，这些都可以通过交叉验证得到. 实际中$k \leq 2$, 即局部二项式。

6.6 Penalized Regression

核估计的思想是，我们利用周围邻居的信息做加权平均作为当前点的估计。如果我们从全局的角度来看，理论上最好的回归函数为

$$\hat{r}(x) = \arg\min_r \frac{1}{n}\sum_{i=1}^{n}(Y_i - r(X_i)), \tag{6.22}$$

得到的估计就为经过每一个观察值的任意函数曲线(如果出现X_i有重合的情形，估计函数应该经过平均值点)。这样得到的估计函数很多时候会非常扭曲，严重的过度拟合(over-fitting, 在训练集上表现完美，但是测试集上可能让人失望). 出现这个问题的原因是，我们让回归函数过于随意了，如果对所得函数加上一定的限制条件，得到的就是 Penalized Regression.

一个常用的形式为：

$$\hat{r}(x) = \arg\min_r \frac{1}{n}\sum_{i=1}^{n}(Y_i - r(X_i)) + \lambda\int(f''(x))^2dx. \tag{6.23}$$

这里$f''(x)$的作用就是尽量让函数光滑，特别的$f''(x) = 0$ 对应的就是线性函数。而λ为调节参数，大小体现了我们希望所得函数的光滑程度。 值得注意的是， 调节参数选取和样本个数有很大关系。例如样本不是很多，这时候λ选取的应该大一点，希望得到的回归函数尽量简单一点。反之，如果有足够多的样本，调节参数就可以小一点，我们可以训练出复杂的回归函数。两个极端情况是：

- $\lambda = 0$, 即我们不考虑惩罚，得到的函数是经过所有点(或平均)的函数

- $\lambda = \infty$, 对应$f''(x) = 0$,得到的估计就是最小二乘估计。

解优化问题6.23, 得到的就是一个 Cubic Spline.

Definition 6.6.1. 给定$x_1 < x_2 < \cdots < x_m$, M阶样条*(Spline)*函数为：

- 在每一个区间$(-\infty, x_1], [x_1, x_2], \cdots, [x_{m-1}, x_m],\ [x_m, \infty)$上是一个$M$阶多项式.
- 在每一个节点x_i*(knots)*处, 函数有$M-1$阶导数.

M阶样条函数的参数个数为： $(M+1)(m+1) - mM = m + M + 1$.

常用的是3阶样条，即Cubic Spline. 直观上，Spline的基为：

$$B_0(x) = 1, B_2(x) = x, \cdots, B_M(x) = x^M, B_{M+j}(x) = (x - x_j)_+^M,\ j = 1, \cdots, m. \tag{6.24}$$

这是一组非常自然的数学基，从计算或者形式上有更为简单的B-Spline.

给定节点$x_1, \cdots, x_m$, 基于spline基，回归函数可以表达为：

$$r(x) = \sum_{j=0}^{m+M} \beta_j B_j(x). \tag{6.25}$$

然后从回归的角度，我们需要估计系数β_j。记

$$G_{ij} = B_j(x_i),\ i = 1, \cdots, n, j = 0, \cdots, M + m. \tag{6.26}$$

最小二乘估计为：

$$\hat{\beta} = \arg\min_{\beta} \|Y - G\beta\|^2 = (G'G)^{-1}G'Y, \tag{6.27}$$

得到的回归函数为

$$\hat{r}(x) = \sum_{j=0}^{m+M+1} \hat{\beta}_j B_j(x). \tag{6.28}$$

考虑惩罚函数

$$\begin{aligned}\hat{r}(x) &= \arg\min_r \sum_{i=1}^{n}(Y_i - r(X_i)) + \lambda \int (f''(x))^2 dx \\ &= \arg\min_r \frac{1}{n}\|Y - G\beta\|^2 + \lambda\beta'\Omega\beta \\ &= (G'G + \lambda\Omega)^{-1}G'Y,\end{aligned}$$

综上，对于样条函数回归，我们需要确定节点, 然后基于数据得到回归函数。不管是直接的样条基最小二乘还是惩罚的形式，预测结果都具有linear smooth的形式，即对于新的观察值，利用当前已有的加权平均作为估计。

从线性代数函数基的角度，我们也可以考虑其他形式的基。例如数学中通用的Fourier变换, 时间序列中常用的小波变换等。所有这些非参数回归都可以用R中现有的package完成，所以多练习这些方法。

6.7 Multivariate Non-parametric Regression

对于多元的解释变量X,

- Additive Model: 假定回归函数具有可加的形式

$$r(X) = E[Y|X] = \beta_0 + \sum_{i=1}^{p} f_i(X_i). \tag{6.29}$$

- Partial Linear Model

$$r(X) = X_1'\beta + g(X_2) \tag{6.30}$$

即回归函数是一部分解释变量的线性形式加上另外一部分解释变量的一般形式。

- Index Model:

$$r(X) = f(X'\beta_0) \tag{6.31}$$

回归函数是当前解释变量的线性组合的函数。

- 其他Semi-parametric Model, Multivariate Index Model 等.

6.8 高维数据回归

从函数基的角度出发，所有模型都是线性回归。对于样本$(X_1, Y_1), \cdots, (X_n, Y_n)$, 确定基函数

$$B_1(x), \cdots, B_p(x), \tag{6.32}$$

和回归函数

$$r(x) = \sum_{i=1}^{p} \beta_i B_i(x) \tag{6.33}$$

最小二乘回归为：

$$\sum_{i=1}^{n} \|Y_i - r(X_i)\|^2 \tag{6.34}$$

类似的，写出设计矩阵

$$X = (B_j(X_i))_{n \times p}$$

残差平方和为：

$$\|Y - X\beta\|^2. \tag{6.35}$$

这里可以发现，即使原始的解释变量不是高维的，回归模型仍然可能是高维的。例如对于一元的(X, Y)， 我们考虑p阶多项式回归：

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_p X^p + \epsilon. \tag{6.36}$$

针对高维回归模型的方法：

- 惩罚：我们希望所得的估计是一个稀疏的，可以给系数β 加上惩罚，例如Lasso

$$\arg\min \|Y - X\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|. \tag{6.37}$$

其他的很多例如SCAD, Adaptive Lasso, Group Lasso等。 另一个特别的是 Dantzig Selector:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{p} |\beta_i| \text{ subject to } |X'(Y - X\beta)|_\infty \leq \lambda_n.$$

- 变量选择 Variable Selection， 对于高维回归，直观的认为很多变量都是noise， 我们只需要选择部分需要的即可。 所以可以用一些Independent Screening， 例如选择相关系数比较大的解释变量。也有一些相对一般的方法来选择变量。通常这一步做完之后，可以直接用传统的最小二乘或者第一步中的 penalty方法。

- Feature Selection 和 Dimension Reduction， 把整个解释变量看成一个高维空间，假定回归函数只是其中几个方向的函数，即Index Model:

$$r(x) = f(B'x),\ B_{p\times r}. \tag{6.38}$$

  对应 Dimension Reduction， 包括SIR，PHD等方法。另一个角度可以从主成分分析PCA，典则相关系数CCA，因子模型等进行 Feature Selection.