非参数统计

第二章 相关分析

授课教师：崔畅

2017年9月

学习目标

- 掌握秩相关的基本原理；
- 掌握*Spearman*和*Kendall*相关检验的基本原理和实现计算；
- 掌握列联表分析的基本原理和实现计算。

Spearman秩相关检验

检验问题：

设量为n的样本，$(X,Y)=\{(X_1, Y_1),...,(X_n, Y_n)\}\sim F$（x，y假设检验问题为：

H_0：X与Y不相关 $\leftrightarrow$ H_1：X与Y正相关。秩的简单相关系数定义为：

$$r_S=\frac{\sum_{i=1}^{n}[(R_i-\frac{1}{n}\sum_{i=1}^{n}R_i)(Q_i-\frac{1}{n}\sum_{i=1}^{n}Q_i)]}{\sqrt{\sum_{i=1}^{n}(R_i-\frac{1}{n}\sum_{i=1}^{n}R_i)^2}\sqrt{\sum_{i=1}^{n}(Q_i-\frac{1}{n}\sum_{i=1}^{n}Q_i)^2}}$$

注意到 $\sum_{i=1}^{n}R_i=\sum_{i=1}^{n}Q_i=\frac{n(n+1)}{2}$，$\sum_{i=1}^{n}R_i^2=\sum_{i=1}^{n}Q_i^2=\frac{n(n+1)(2n+1)}{6}$，因此$r_S$可以简化为

$$rS=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i-Q_i)^2$$

Spearman秩相关检验

显著性检验：

(1) 建立零假设： $H_0: r_S = 0 \quad vs \quad H_1: r_S \neq 0$

(2) 构造统计量：

参数统计中用 t 检验来进行相关性检验，在零假设下也可以类似的定义 T 检验统计量：$T = r_S\sqrt{\frac{n-2}{1-r_S^2}}$。该统计量在零假设下服从 $\nu = n-2$ 的 t 分布，当 $T > t_{\alpha,\nu}$ 时，表示两变量有相关关系，反之则无。

若数据中有重复数据，可以采用平均秩法定义秩，当结不多时，仍然可使用 r_S 定义秩相关系数，T 检验仍然可以使用。

Spearman秩相关检验

定理 在零假设之下，*Spearman*秩相关系数分布满足：

(1) $E_{H_0}(r_S)=0$，$\mathrm{var}_{H_0}(r_S)=\dfrac{1}{n-1}$

(2) 关于原点*0*对称。

根据定理可以方便地构造*Spearman*秩相关系数零分部表。如果令 $\alpha(2)$ 表示双边假设“H_0：*X*与*Y*不相关 $\leftrightarrow$ H_1：*X*与*Y*相关”的显著性水平，$\alpha(1)$ 为单边假设“H_0：*X*与*Y*不相关 $\leftrightarrow$ H_1：*X*与*Y*正相关”的显著性水平。

当 $r_S \geq c\alpha_{(1)}$（双边时为 $r_S \geq c\alpha_{(2)}$ 或者 $r_S \leq c\alpha_{(2)}$）时拒绝 H_0。

Spearman秩相关检验

当n较大时，*Hotelling*等人于*1936*年证明，*Spearman*秩相关系数有如下的大样本性质：

当 $n \to \infty$ 时，

$$\sqrt{n-1}r_S \xrightarrow{\ell} N(0,1)$$

因此在大样本时，可用正态近似。

Spearman秩相关检验

当 X 或 Y 样本中有结存在时，可按平均秩法定秩，相应的*Spearman*相关系数：

$$r^* = \frac{\frac{n(n^2-1)}{6} - \frac{1}{12}\left[\sum_i (\tau_i^3(x) - \tau_i(x)) + \sum_j (\tau_j^3(y) - \tau_j(y))\right] - \sum_{i=1}^{n}(R_i - Q_i)^2}{2\sqrt{\left[\frac{n(n^2-1)}{12} - \frac{1}{12}\sum_i (\tau_i^3(x) - \tau_i(x))\right]\left[\frac{n(n^2-1)}{12} - \frac{1}{12}\sum_j (\tau_j^3(y) - \tau_j(y))\right]}}$$

作为检验统计量，其中 $\tau_i(x)$，$\tau_j(y)$ 分别表示 X，Y 样本中的结统计量。

当结的长度较小时，关于 r^* 的零分布仍可用无结时的零分布近似，当 n 较大时，也可用如下的极限分布：

$$r^* \sqrt{n-1} \xrightarrow{\ell} N(0,1)$$

进行大样本检验。

Spearman秩相关检验举例1

例2.1 为了研究品牌知名度和售后服务质量之间的关系，随机抽取10个品牌的产品，其知名度和售后服务质量排序结果如下：

表 2-1 10个品牌知名度和售后服务质量排序

编号	1	2	3	4	5	6	7	8	9	10
知名度	9	4	3	6	5	8	1	7	10	2
售后	8	2	5	4	7	9	1	6	10	3

分析品牌知名度和售后服务质量之间是否存在显著相关性。

Spearman秩相关检验举例1

解： (1)建立假设 $H_0:\gamma_s=0;$

$H_1:\gamma_s\neq 0.$

(2)计算*Spearman*相关系数 $\gamma_s=0.879.$

(3)给定显著性水平0.05，否定域为

$$\Theta=\{\gamma_s \,|\, |\gamma_s|>0.648\}.$$

(4)拒绝零假设，品牌知名度和售后服务质量的*Spearman*相关系数显著。

Spearman秩相关检验举例2

例2.2　为了研究客户和公司对员工服务态度评价之间的关系，随机抽取12名员工，客户和公司对其服务态度的评价分数如下表

表2-2　客户和公司对12名员工评价分数

编号	1	2	3	4	5	6	7	8	9	10	11	12
客户	83	95	76	89	97	87	92	89	79	92	95	90
公司	87	94	80	92	94	89	91	88	74	96	94	89

分析客户和公司对员工服务态度的评价是否存在显著相关性。

Spearman秩相关检验举例2

解： (1)建立假设：$H_0: \gamma_s = 0;$

$H_1: \gamma_s \neq 0.$

(2)计算检验统计量

$$\gamma_s = 0.879, t = 5.830$$

(3)给定显著性水平0.05，否定域为

$$\Theta = \{t \,|\, |t| > t_{0.025}(10) = 2.228\}$$

(4)拒绝零假设，即客户和公司对员工服务态度评价存在显著的正相关关系。

Kendall相关检验

双变量Kendall相关检验：

*Kendall*提出一种类似于*Spearman*秩相关的检验方法，从两变量*X*和*Y*是否协同一致的角度出发来检验变量之间的相关性。首先引入协同的概念：假设有n对观测值 $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$

若$(x_j-x_i)(y_j-y_i)>0,\ \forall j>i,i,j=1,2,\ldots,n$，则称数对$(x_i,y_i)$与$(x_j,y_j)$协同。

若$(x_j-x_i)(y_j-y_i)<0,\ \forall j>i,i,j=1,2,\ldots,n$，则称数对$(x_i,y_i)$与$(x_j,y_j)$不协同。

协同性测量了前后两个数对秩的大小变化同向还是反向，若前一对均比后一对秩小，则前后数对具有同向性；反之则前后两对数对反向。

Kendall相关检验

全部数据所有可能前后数对共有$\binom{n}{2}=n(n-1)/2$对，如果用N_C表示同向数对的数目，N_{d}表示反向数对的数目，则$N_{\mathrm{c}}+N_d=n(n-1)/2$。*Kendall*相关系数统计量由二者的平均差定义，如下所示：

$$\tau=\frac{N_{\mathrm{c}}-N_d}{n(n-1)/2}=\frac{2S}{n(n-1)}$$

式中$S=N_c-N_d$。若所有数对协同一致，则$N_{\mathrm{c}}=n(n-1)/2$，$N_d=0$，$\tau=1$，表示两组数据正相关；若所有数对全反向，则$N_{\mathrm{c}}=0$，$N_d=n(n-1)/2$，$\tau=-1$，表示两组数据负相关；Kendall τ为零，表示数据中同向和反向的数对势力均衡，没有明显的趋势，这与相关性的含义是一致的。

Kendall相关检验

另外，我们注意到，如果定义：

$$sign((X_1-X_2)(Y_1-Y_2))=\begin{cases}1,(X_1-X_2)(Y_1-Y_2)>0,\\0,(X_1-X_2)(Y_1-Y_2)=0,\\-1,(X_1-X_2)(Y_1-Y_2)<0;\end{cases}$$

则

$$\tau=\frac{2}{n(n-1)}\sum_{1\le i<j\le n}sign((x_i-x_j)(y_i-y_j))$$

式中 $sign((X_1-X_2)(Y_1-Y_2))$ 是 $P((X_1-X_2)(Y_1-Y_2)>0)$ 的核估计量，因而 τ 是U 统计量。用U 统计量的方法，可以证明下面的定理。

Kendall相关检验

定理 在零假设 H_0：X 与 Y 不相关成立时，

（1） $E_{H_0}(\tau)=0, var_{H_0}(\tau)=\dfrac{n(n-1)(2n+5)}{18}$

（2）关于原点 0 对称。

Kendall相关检验

实际中，不失一般性，假定 x_i 已从小到大或从大到小排序，因此协同性问题就转化为 y_i 秩的变化。令 $d_1,d_2,..,d_n$ 为 $y_1,y_2,...,y_n$ 的秩，因而 x,y 的秩形成 $(1,d_1),(2,d_2),\ ...,(n,d_n);\ \forall 1\le i\le n$。记

$$p_i=\sum_{j>i} I(d_j>d_i),i=1,2,...,n;\quad q_i=\sum_{j>i} I(d_j<d_i),i=1,2,...,n.$$

令 $P=\sum_{i=1}^{n} p_i, Q=\sum_{i=1}^{n} q_i$，则 Kendall τ 统计量的值为 $K=\frac{P-Q}{n(n-1)/2}$，即对每个 y_i 求当前位置后比 y_i 大的数据的个数，将这些数相加所得就是 N_c，同理可计算 N_d。

Kendall相关检验

若 x_i 或 y_i 有相等秩时，用平均秩计算各自的秩，Kendall的τ公式校正如下：

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T_x}\sqrt{n(n-1)/2 - T_y}}$$

式中，$T_x = \frac{1}{2}\sum^{g_x}(\tau_x^3 - \tau_x), T_y = \frac{1}{2}\sum^{g_y}(\tau_y^3 - \tau_y)$，$\tau_x, \tau_y$分别为$\{x_i, y_i\}$的结长，$g_x, g_y$ 分别为两变量中结的个数。

Kendall相关检验举例1

例2.3　为了研究两类消费者对某种产品的评价标准是否一致，在众多品牌中，随机抽取8个不同品牌的产品。针对两类消费者各举办一次焦点座谈会，两类消费者对随机选择的产品A, B, C, D, E, F, G和H的排序结果如下表

表2-3　两类消费者对8个品牌产品的排序

编号	A	B	C	D	E	F	G	H
类型1	8	1	3	5	7	2	4	6
类型2	7	3	4	2	6	1	8	5

分析两类消费者的评价标准是否存在显著差异。

Kendall相关检验举例1

解：（1）建立假设 $H_0: \tau = 0;$

$H_1: \tau \neq 0.$

（2）计算检验统计量：$\tau = 0.5$

（3）给定显著性水平 $\alpha = 0.05,$ 否定域为

$$\Theta = \{\tau \mid \tau > 0.571\}$$

（4）**接受零假设**，即两类消费者对产品评价标准的*Kendall*秩相关系数不显著。

Kendall相关检验举例2

例2.4　通过深入访谈，得到12家企业近5年新产品数量和新产品开发人员的数据如下表：

表 2-4　12家企业新产品开发人员和新产品数量

编号	1	2	3	4	5	6	7	8	9	10	11	12
新产品/件	4	7	13	2	2	10	1	8	4	3	9	12
开发人员/人	5	12	18	8	6	23	8	9	6	10	14	31

分析新产品开发人员和新产品数量之间是否存在显著相关性。

Kendall相关检验举例2

解： （1）建立假设：$H_0:\tau=0;$

$H_1:\tau\neq 0.$

（2）计算检验统计量：$Z=2.687$

（3）给定显著性水平$\alpha=0.05,$ 否定域为

$$\Theta=\{Z\,\|\,Z\,|>z_{0.025}=1.96\}.$$

（4）**拒绝零假设**，即新产品数量同开发人员之间存在正相关关系。

列联表分析

- 分析按两个或多个特征分类的频数数据，这种数据通常称为交叉分类数据，它们一般都以表格的形式给出，称为列联表。
- 例如，在考察色盲与性别有无关联时，随机抽取1000人按性别（男或女）及色觉（正常或色盲）两个属性分类，得到如下二维列联表，又称2×2表或四格表。

性别	视觉	
	正常	色盲
男	535	65
女	382	18

列联表分析

一般，若总体中的个体可按两个属性A与B分类，A有r个类$A_1,\ldots,A_r$，B有c个类，$B_1,\ldots,B_c$，从总体中抽取大小为n的样本，设其中有n_{ij}个个体既属于类A_i又属于类B_j，n_{ij}称为频数，将$r\times c$个n_{ij}排列为一个r行c列的二维列联表，简称$r\times c$列联表。

列联表分析

r×c 列联表

A\B	1	…	j	…	c	和
1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1\cdot}$
…	…		…		…	…
i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i\cdot}$
…	…		…		…	…
r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r\cdot}$
列和	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot c}$	n

列联表分析

列联表分析的基本问题是，考察各属性之间有无关联，即判别两属性是否独立。在$r\times c$列联表中，若以 $p_{i\cdot}$,$p_{\cdot j}$和p_{ij}分别表示总体中的个体仅属于A_i，仅属于B_j，同时属于A_i与B_j的概率，可得一个二维离散分布表，则“A、B两属性独立”的假设可以表述为：

$$H_0:\ p_{ij}=p_{i\cdot}p_{\cdot j},\ \ i=1,\ldots,r,\ \ j=1,\ldots,c$$

列联表分析

r×c 列联表

A\B	1	…	j	…	c	和
1	p_{11}	…	p_{1j}	…	p_{1c}	$p_{1\cdot}$
…	…		…		…	…
i	p_{i1}	…	p_{ij}	…	p_{ic}	$p_{i\cdot}$
…	…		…		…	…
r	p_{r1}	…	p_{rj}	…	p_{rc}	$p_{r\cdot}$
列和	$p_{\cdot 1}$	…	$p_{\cdot j}$	…	$p_{\cdot c}$	1

列联表分析

这就变为诸 p_{ij} 不完全已知时的分布拟合检验。这里诸 p_{ij} 共有rc个参数，在原假设H_0 成立时，这rc个参数 p_{ij}由$r+c$个参数 $p_{1\cdot},\ldots,p_{r\cdot}$和$p_{\cdot 1},\ldots,p_{\cdot c}$ 决定，在这后$r+c$个参数中存在两个约束条件：$\sum_{i=1}^{r} p_{i\cdot}=1$，$\sum_{j=1}^{c} p_{\cdot j}=1$，所以此时实际上 p_{ij} 由$r+c-2$个独立参数所确定。据此，检验统计量为

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n\hat{p}_{ij})^2}{np_{ij}}$$

在 H_0成立时，上式近似服从自由度为 $rc-(r+c-2)-1=(r-1)(c-1)$的χ^2分布。其中诸 p_{ij} 是在 H_0成立下得到的 p_{ij}的最大似然估计，其表达式为

$$\hat{p}_{ij}=\hat{p}_{i\cdot}\hat{p}_{\cdot j}=\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}$$

对给定的显著性水平$\alpha(0<\alpha<1)$，检验的拒绝域为 $W=\{\chi^2\geq\chi^2_{1-\alpha}((r-1)(c-1))\}$

列联表分析举例1

例2.5 随机抽取50名消费者，出示三种由红、黄和蓝颜色包装的同样产品各一件，让其从中选出最喜欢的包装颜色。50名消费者的性别构成其挑选的颜色如下表

表2-5 颜色和性别关系数据

	红	黄	蓝
男	3	15	8
女	14	7	3

列联表分析举例1

解： （1）**建立假设**

H_0：消费者性别及其喜欢的颜色相互独立

H_1：消费者性别及其喜欢的颜色不相互独立

（2）**构造和计算检验统计量**

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}} = 12.239.$$

在零假设下，统计量服从自由度为(2-1)(3-1)=2的 χ^2 分布。

列联表分析举例1

（3）**设定显著性水平和确定否定域**

给定显著性水平 $\alpha = 0.05$, 否定域为

$$\Theta = \{\chi^2 \mid \chi^2 > \chi^2_{0.05}(2) = 5.99\}.$$

（4）**做出统计决策**

由于 $\chi^2 = 12.239$, 落在否定域 Θ 中，从而拒绝零假设，即性别同颜色偏好之间存在相关性。

（5）**计算列联系数**

$$C = \sqrt{\frac{12.24}{12.24 + 50}} = 0.443.$$

列联表分析举例2

例2.6 某公司的工业设计部门为了检验目标市场对三种设计好的手机款式的偏好是否相同，随机从目标市场中抽取36名消费者进行调研，得到他们对三种手机款式的偏好数据如下表：

表 2-6 消费者对三款手机的偏好数据

	喜欢	一般	不喜欢
款式1	1	8	3
款式2	4	5	2
款式3	6	2	5

试对不同手机款式的消费者偏好度是否相同作出评价。

列联表分析举例2

解： （1）建立假设

H_0：手机款式和消费者偏好度相互独立

H_1：手机款式和消费者偏好度不相互独立

（2）计算检验统计量 $\chi^2 = 8.021.$

在零假设下，统计量服从自由度为(3-1)(3-1)=4的 χ^2 分布。

（3）给定显著性水平 $\alpha = 0.05,$ 否定域为 $\Theta = \{\chi^2 \mid \chi^2 > \chi^2_{0.05}(4) = 9.49\}$

（4）由于 $\chi^2 = 8.021,$ 没有落在否定域中，从而接受零假设，

即消费者对不同款式的手机偏好不存在差异性。

（5）计算列联系数 $C = \sqrt{\dfrac{8.021}{8.021+36}} = 0.427.$

Thank You !