Permutation Tests
置换检验

蒋贵豪

南开大学统计与数据科学学院

2021 年 10 月 27 日

1. 置换检验介绍
2. 一元同分布检验
3. 多元同分布检验
4. 应用：距离相关性

前言

- 我们一般平时较为常用的检验为有参检验，但是其要求样本必须满足近似正态，无离群点，数据量大等要求。
- 在非参数检验方法中，Bootstrap 用共享观测数据集的数据进行“实验”进行数据分布的检验。除此之外，置换检验 (Permutation Tests) 也是用已有的观测数据进行分布的检验。
- 置换检验由 Fisher 提出，利用样本数据的随机排列，适合用于总体分布未知的小样本数据。

方法

- $X_1, X_2, ..., X_n$ 和 $Y_1, Y_2, ..., Y_m$ 是分布 F_X 和 F_Y 的独立随机样本。
- $Z = \{X_1, ..., X_n, Y_1, ..., Y_m\} = \{Z_1, ..., Z_n, Z_{n+1}, ..., Z_{n+m}\}$
- $\nu = \{1, ..., n, n+1, ..., n+m\} = \{1, ..., N\}$
- 置换 π：$Z^* = (X^*, Y^*)$, $Z_i^* = Z_{\pi(i)}$
- 所有置换数目：C_N^n, 且为等概率的。

方法

- 原假设 H_0：$F_X = F_Y \quad vs \quad$ 备择假设 H_1：$F_X \neq F_Y$
- 如果 $\widehat{\theta}(X,Y) = \widehat{\theta}(Z,\nu)$ 是一个统计量。则 $\widehat{\theta}^*$ 的置换分布为：

$$\{\widehat{\theta}^*\} = \{\widehat{\theta}^{(j)} = \widehat{\theta}(Z,\pi(\nu)), j = 1, ..., \mathrm{C}_N^n\}$$

- $\widehat{\theta}^*$ 的 CDF 定义为：

$$F_{\theta^*}(t) = P(\widehat{\theta}^* \leq t) = (\mathrm{C}_N^n)^{-1} \sum_{j=1}^{N} I(\widehat{\theta}^{(j)} \leq t)$$

- Achieved significance level(ASL):

$$P(\widehat{\theta}^* \geq \widehat{\theta}) = (\mathrm{C}_N^n)^{-1} \sum_{j=1}^{N} I(\widehat{\theta}^{(j)} \geq \widehat{\theta})$$

类似地，我们可以计算左尾，双边的 ASL.

置换检验流程

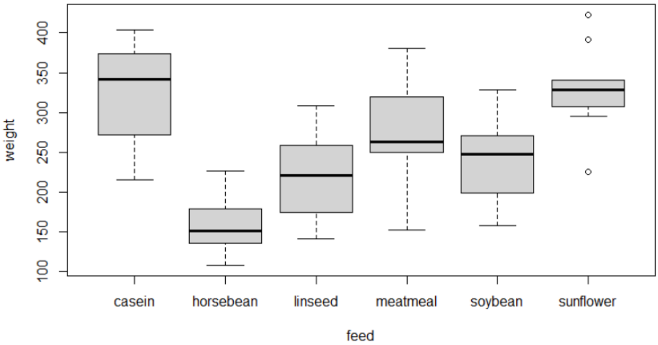
- 计算观测数据的统计量：$\widehat{\theta}(X,Y)=\widehat{\theta}(Z,\nu)$
- 设我们共进行 B 次置换，$b=1,2,...,B$:
  (a) 生成随机排列：$\pi_b=\pi(\nu)$
  (b) 计算统计量：$\widehat{\theta}^{(b)}=\widehat{\theta}^*(Z,\pi_b)$
- 计算 ASL：

$$\widehat{p}=\frac{1+\sum_{b=1}^{B} I(\widehat{\theta}^{(b)}\geq\widehat{\theta})}{B+1}$$

  类似地，我们可以计算左尾，双边的 ASL.
- 对于置信水平 α, 如果 $\widehat{p}\leq\alpha$, 拒绝原假设 H_0

例：t 统计量的置换分布

```r
data("chickwts")
plot(weight ~ feed, chickwts)
```



例：t 统计量的置换分布

```
attach(chickwts)
x <- sort(weight[feed == 'soybean'])
y <- sort(weight[feed == 'linseed'])
detach(chickwts)
cat('X:',x)
## X: 158 171 193 199 230 243 248 248 250 267 271 316 327 329
cat('Y:',y)
## Y: 141 148 169 181 203 213 229 244 257 260 271 309
```

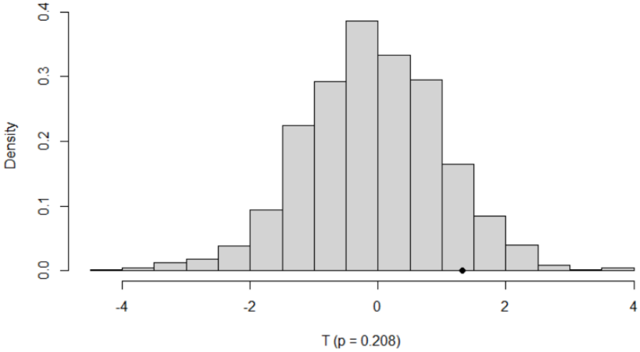
例：t 统计量的置换分布

```r
R <- 999
z <- c(x, y)
K <- 1:26
reps <- numeric(R)
t_0 <- t.test(x, y)$statistic

for (i in 1:R) {
    k <- sample(K, size = 14, replace = FALSE)
    x_1 <- z[k]
    y_1 <- z[-k]
    reps[i] <- t.test(x_1, y_1)$statistic
}
p = mean(c(t_0, reps) >= t_0)
cat('p=', p)
## p= 0.104
```

例：t 统计量的置换分布

```r
hist(reps, main = '', freq = FALSE, xlab = "T (p = 0.208)",
    breaks = 'scott')
points(t_0, 0, cex = 1, pch = 16)
```



例：K-S 统计量的置换分布

- 原假设 H_0：$F = G$ vs 备择假设 H_1：$F \neq G$
- $$D = \sup_{1 \leq i \leq N} |F_n(z_i) - G_m(z_i)|$$

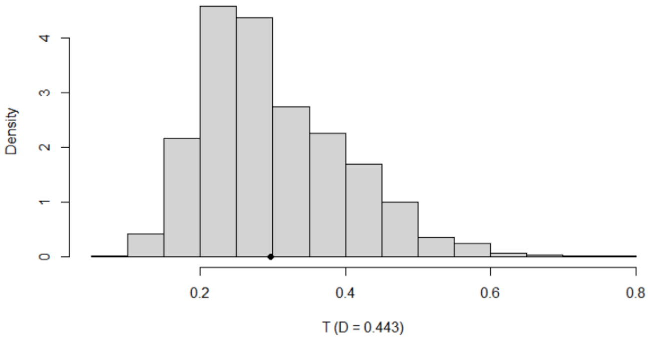
例：K-S 统计量的置换分布

```r
R <- 999
z <- c(x, y)
K <- 1:26
D <- numeric(R)
options(warn = -1)
D_0 <- ks.test(x, y, exact = FALSE)$statistic

for (i in 1:R) {
    k <- sample(K, size = 14, replace = FALSE)
    x_1 <- z[k]
    y_1 <- z[-k]
    D[i] <- ks.test(x_1, y_1, exact = FALSE)$statistic
}
p = mean(c(D_0, D) >= D_0)
options(warn = 0)
cat('p=', p)
## p= 0.443
```

例：K-S 统计量的置换分布

```r
hist(D, main = '', freq = FALSE, xlab = "T (D = 0.443)",
    breaks = 'scott')
points(D_0, 0, cex = 1, pch = 16)
```



例： K-S 统计量的置换分布

```
attach(chickwts)
x <- sort(weight[feed == 'sunflower'])
y <- sort(weight[feed == 'linseed'])
detach(chickwts)
summary(cbind(x,y))
##         x               y
##  Min.   :226.0   Min.   :141.0
##  1st Qu.:312.8   1st Qu.:178.0
##  Median :328.0   Median :221.0
##  Mean   :328.9   Mean   :218.8
##  3rd Qu.:340.2   3rd Qu.:257.8
##  Max.   :423.0   Max.   :309.0
```

例：K-S 统计量的置换分布

```r
R <- 999
z <- c(x, y)
K <- 1:26
D <- numeric(R)
options(warn = -1)
D_0 <- ks.test(x, y, exact = FALSE)$statistic

for (i in 1:R) {
    k <- sample(K, size = 14, replace = FALSE)
    x_1 <- z[k]
    y_1 <- z[-k]
    D[i] <- ks.test(x_1, y_1, exact = FALSE)$statistic
}
p = mean(c(D_0, D) >= D_0)
options(warn = 0)
cat('p=', p)
## p= 0.001
```

一元情况下的双样本同分布检验方法

- Kolmogorov-Smirnov (K-S) 统计量：

$$D = \sup_{1 \le i \le N} |F_n(z_i) - G_m(z_i)|$$

- Cram'er-von Mises 统计量：

$$W_2 = \frac{mn}{(m+n)^2}\left[\sum_{i=1}^{n}\left(F_n\left(x_i\right) - G_m\left(x_i\right)\right)^2 + \sum_{j=1}^{m}\left(F_n\left(y_j\right) - G_m\left(y_j\right)\right)^2\right]$$

- 这些统计量无法直接在多元情况下进行自然延拓
- $$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(X_i \le x\right)$$

多元样本表示

- $\boldsymbol{X} = \{X_1, \cdots, X_{n_1}\} \in \mathbf{R}^d, \quad \boldsymbol{Y} = \{Y_1, \cdots, Y_{n_2}\} \in \mathbf{R}^d$
- $$\boldsymbol{Z}_{n\times d} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & & \vdots \\ x_{n_1,1} & x_{n_1,2} & \cdots & x_{n_1,d} \\ y_{1,1} & y_{1,2} & \cdots & y_{1,d} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,d} \\ \vdots & \vdots & & \vdots \\ y_{n_2,1} & y_{n_2,2} & \cdots & y_{n_2,d} \end{pmatrix}$$
- 最近邻检验的基础是：第 1 最近邻到第 r 最近邻的耦合情况

最近邻

- 在二范数距离下：Z_i 的第 r 最近邻为 Z_j, Z_j 需要满足以下条件：
- $\|Z_i - Z_l\| \leq \|Z_i - Z_j\|$ 对恰好 $r-1$ 个下标成立 $(1 \leq l \leq n$ 且 $l \neq i)$
- 用 $NN_r(Z_i)$ 表示 Z_i 的第 r 最近邻

示例

```r
attach(chickwts)
x <- sort(weight[feed == 'sunflower'])
y <- sort(weight[feed == 'linseed'])
detach(chickwts)
cat('X:',x)
## X: 226 295 297 318 320 322 334 339 340 341 392 423
cat('Y:',y)
## Y: 141 148 169 181 203 213 229 244 257 260 271 309
```

- 一般来说，如果抽样分布相等，那么样本的近邻相对于备择假设 (分布不同的情况) 会更分散于混合样本中

耦合和耦合比例

- 样本 Z_i 与其第 r 近邻耦合：Z_i 的第 r 近邻 $NN_r(Z_i)$ 在同一个样本集中
- 计算混合样本的第 r 近邻耦合比率:

$$T_{n,1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_i(1), \quad T_{n,2} = \frac{1}{2n}\sum_{i=1}^{n} (\mathbb{1}_i(1) + \mathbb{1}_i(2))$$

- $T_{n,1}$ 较大支持了分布不同的备择假设
- 置换检验的不放回采样保证了不会存在并列值

第 J 最近邻统计量

- 基于最近邻和耦合比例的概念，引入第 J 最近邻统计量：
  原假设 H_0：$X = Y \quad vs \quad$ 备择假设 H_1：$X \neq Y$
- $$T_{n,J} = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{r=1}^{J} I_i(r)$$
- 第 J 最近邻统计量是用来度量第 1 最近邻到第 J 最近邻耦合的比例
- 最近邻统计量是有序样本元素间距离的函数，由于假设抽样分布是连续的，所以不会存在并列值

R 语言实现寻找最近邻

```
library('boot')
library('yaImpute') #knnfinder 包已经不用了
# 计算 k 近邻矩阵
attach(chickwts)
x <- as.vector(weight[feed == "sunflower"])
y <- as.vector(weight[feed == "linseed"])
detach(chickwts)
z = c(x,y)
z = as.matrix(z)
#ann 是计算两个样本矩阵之间的 k 近邻，所以在一组样本中选择后面 k-1 近邻，输入为矩阵
NN = ann(z,z,k=4)
NN$knnIndexDist[,2:4]
```

R 语言实现寻找最近邻

```
NN$knnIndexDist[,2:4]
##       [,1] [,2] [,3]      ##       [,1] [,2] [,3]
##  [1,]    3    5    2      ## [13,]   12    7   11
##  [2,]    4    5    9      ## [14,]    6   23   21
##  [3,]    1    5    2      ## [15,]   20   18   21
##  [4,]    2    5    9      ## [16,]   19   20   15
##  [5,]    2    4    9      ## [17,]   22   24   23
##  [6,]   14   21   23      ## [18,]   21   15    6
##  [7,]   12   10   13      ## [19,]   16   20   15
##  [8,]   11   13   12      ## [20,]   15   19   16
##  [9,]    4    2    5      ## [21,]   18    6   14
## [10,]    7   12    9      ## [22,]   17   23   24
## [11,]    8   13   12      ## [23,]   22   14   17
## [12,]    7   10   13      ## [24,]   17   22    8
```

$T_{n,3}$ 近邻检验

```
library('boot')
library('yaImpute') #knnfinder 包已经不用了
# 使用置换检验计算统计量分布
Tn3 = function(z,ix,sizes)
{
  n1 = sizes[1]
  n2 = sizes[2]
  n = n1 + n2
  z = z[ix,]
  z = as.matrix(z)
  NN = ann(z,z,k=4)
  block1 = NN$knnIndexDist[1:n1,2:4]# 选择 2:4 列为 i 的 3 个近邻
  block2 = NN$knnIndexDist[(n1+1):n,2:4]
```

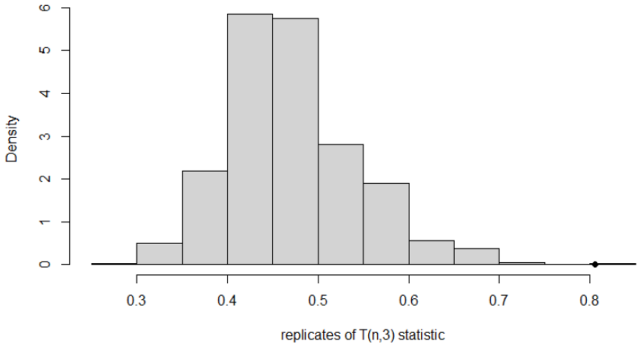
$T_{n,3}$ 近邻检验

```r
  i1 = sum(block1 < n1+0.5)
  i2 = sum(block2 > n1+0.5)
  return((i1 + i2) / (3 * n))
}
N = c(12, 12)
z = as.data.frame(z)
boot.obj = boot(data = z, statistic = Tn3, sim = "permutation",
                R = 999, sizes = N)
tb = c(boot.obj$t, boot.obj$t0)
mean(tb >= boot.obj$t0)
## [1] 0.002
```

- 除此之外也可以使用 coin 包和 lmPerm 包进行置换检验，置换检验在这里本质就是对混合样本对标签进行不放回抽样

绘制直方图

```
# 绘制直方图
hist(tb, freq = FALSE, main = '',
     xlab = "replicates of T(n,3) statistic")
points(boot.obj$t0, 0, cex=1, pch=16)
```



能量同分布检验

- 能量距离或 e 距离统计量 $\mathcal{E}_n$：

$$\mathcal{E}_n = e(\boldsymbol{X}, \boldsymbol{Y}) = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|X_i - Y_j\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|X_i - X_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|Y_i - Y_j\|\right)$$

能量同分布检验

- X, X', Y, Y' 是 $\mathcal{R}_d$ 中具有有限期望且互相独立的随机变量。 $\boldsymbol{X} \overset{D}{=} \boldsymbol{X}', \quad \boldsymbol{Y} \overset{D}{=} \boldsymbol{Y}'$, 那么：

$$2E\|\boldsymbol{X}-\boldsymbol{Y}\| - E\left\|\boldsymbol{X}-\boldsymbol{X}'\right\| - E\left\|\boldsymbol{Y}-\boldsymbol{Y}'\right\| \geqslant 0$$

- 当且仅当 X, Y 同分布时等号成立
- X, Y 之间的 $\mathcal{E}$ 距离为：

$$\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = 2E\|\boldsymbol{X}-\boldsymbol{Y}\| - E\left\|\boldsymbol{X}-\boldsymbol{X}'\right\| - E\left\|\boldsymbol{Y}-\boldsymbol{Y}'\right\|$$

性质

- 较大的 e 距离对应着不同的分布
- 相较于经验分布函数统计量而言，e 距离并不依赖于分布表的概念，并且是对多元的分布间距离度量
- 如果 X, Y 不是同分布的，且 $n = n_1 + n_2$, 那么 $E[\mathcal{E}_n]$ 逼近于一个常数乘以 n。由于样本大小 n 趋于无穷，在零假设下 $E[\mathcal{E}_n]$ 趋于一个常数（此时 $\mathcal{E}_n$ 本身收敛），在备择假设下则趋于无穷。
- 基于 $\mathcal{E}_n$ 的同分布检验对所有具有有限一阶距的备择假设来说都是一致的，$\mathcal{E}_n$ 的渐进分布都是中心化的高斯随机变量的二次型，其系数依赖于 X 和 Y 的分布

双样本能量统计量

```r
edist.2 <- function(x, ix, sizes) {
  dst <- x
  n1 <- sizes[1]
  n2 <- sizes[2]
  ii <- ix[1:n1]
  jj <- ix[(n1 + 1):(n1 + n2)]
  w <- n1 * n2 / (n1 + n2)
  m11 <- sum(dst[ii, ii]) / (n1 * n1)
  m22 <- sum(dst[jj, jj]) / (n2 * n2)
  m12 <- sum(dst[ii, jj]) / (n1 * n2)
  e <- w * ((m12 + m12) - (m11 + m22))
  return (e)
}
```

双样本能量统计量

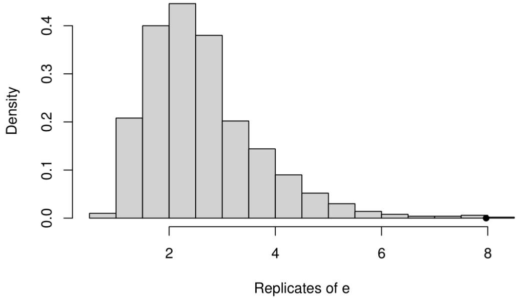
```r
d <- 3
a <- 2/sqrt(d)
x <- matrix(rnorm(20 * d), nrow = 20, ncol = d)
y <- matrix(rnorm(10 * d, a, 1), nrow = 10, ncol = d)
z <- rbind(x, y)
dst <- as.matrix(dist(z))
edist.2(dst, 1:30, sizes = c(20, 10))
## [1] 7.968832
```

双样本能量检验

```
library(boot)
N <- c(20, 10)
dst <- as.matrix(dist(z))
boot.obj <- boot(data = dst, statistic = edist.2,
                 sim = "permutation", R = 999, sizes = N)
boot.obj
##
## DATA PERMUTATION
##
##
## Call:
## boot(data = dst, statistic = edist.2, R = 999, sim = "permutation",
##     sizes = N)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* 7.968832 -5.344543    1.070509
```

计算显著性水平和绘制直方图

```
e <- boot.obj$t0
tb <- c(e, boot.obj$t)
mean(tb >= e)
## [1] 0.002
hist(tb, main = "", breaks="scott", freq=FALSE,
     xlab="Replicates of e")
points(e, 0, cex=1, pch=16)
```



验证

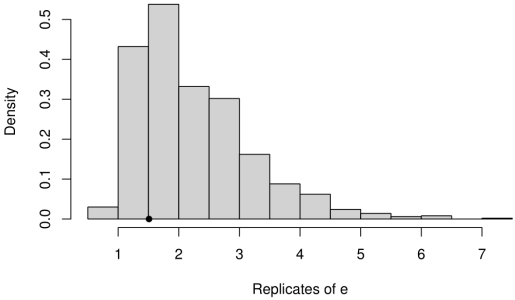
```
d <- 3
a <- 0
x <- matrix(rnorm(20 * d), nrow = 20, ncol = d)
y <- matrix(rnorm(10 * d, a, 1), nrow = 10, ncol = d)
z <- rbind(x, y)
dst <- as.matrix(dist(z))
```

验证抽样分布相同时的检验结果

```
library(boot)
N <- c(20, 10)
dst <- as.matrix(dist(z))
boot.obj <- boot(data = dst, statistic = edist.2,
                 sim = "permutation", R = 999, sizes = N)
boot.obj
##
## DATA PERMUTATION
##
##
## Call:
## boot(data = dst, statistic = edist.2, R = 999, sim = "permutation",
##     sizes = N)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* 1.509022 0.730646   0.9506633
```

计算显著性水平和绘制直方图

```
e <- boot.obj$t0
tb <- c(e, boot.obj$t)
mean(tb >= e)
## [1] 0.766
hist(tb, main = "", breaks="scott", freq=FALSE,
     xlab="Replicates of e")
points(e, 0, cex=1, pch=16)
```



最近邻检验和能量检验比较

二元正态位置选择$F_1 = N_2((0,0)^{\mathrm{T}}, \boldsymbol{I}_2)$和$F_2 = N_2((0,\delta)^{\mathrm{T}}, \boldsymbol{I}_2)$的显著性检验（在$\alpha = 0.1, se \leqslant 0.5\%$ 下的最接近整体百分比）

		$\delta=0$		$\delta=0.5$		$\delta=0.75$		$\delta=1$	
n_1	n_2	ε_n	$T_{n,3}$	ε_n	$T_{n,3}$	ε_n	$T_{n,3}$	ε_n	$T_{n,3}$
10	10	10	12	23	19	40	29	58	42
15	15	9	11	30	21	53	34	75	52
20	20	10	12	37	23	64	38	86	58
25	25	10	11	43	25	73	42	93	65
30	30	10	11	48	25	81	47	96	70
40	40	11	10	59	28	90	52	99	78
50	50	10	11	69	29	95	58	100	82
75	75	10	11	85	37	99	69	100	93
100	100	10	10	92	40	100	79	100	100

1. 置换检验介绍
2. 一元同分布检验
3. 多元同分布检验
4. 应用：距离相关性

距离相关性

- 为说明随机向量的独立性提供了一种新的方法

$$\|\gamma(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 w(t,s) \mathrm{d}t\mathrm{d}s$$

- 权重函数 $w(t,s)$ 使得上述积分存在的恒大于 0 的任意函数
- 定义 X, Y 的距离协方差：

$$\begin{aligned}\mathcal{V}^2(X,Y;w) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)\mathrm{d}t\mathrm{d}s\end{aligned}$$

距离相关性

- 定义 X, Y 的距离协方差：

$$\begin{aligned}\mathcal{V}^2(X,Y;w) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)\mathrm{d}t\mathrm{d}s\end{aligned}$$

- 考虑到距离相关系数对随机变量线性变换的不变性 $\{(X,Y) \mapsto (\epsilon X, \epsilon Y), \text{ for } \epsilon > 0\}$，和运算的简便。
  这里取 $w(t,s) = \left(c_p c_q |t|_p^{1+p} |s|_q^{1+q}\right)^{-1}$
  其中 $c_d = C(d,1) = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$

距离相关性

- 设 X, Y 别为 $\mathbb{R}^p$,$\mathbb{R}^q$ 上随机变量，其特征函数分别为 $f_X(t), f_Y(t)$ 它们的距离协方差定义为：

$$\mathcal{V}^2(X,Y) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2$$
$$= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} \mathrm{d}t\mathrm{d}s$$

- X 的方差可以写成：

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2$$

距离相关性

- X, Y 的距离相关系数就可以写成：

$$\mathcal{R}^2(X, Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \nu^2(X)\mathcal{V}^2(Y) = 0 \end{cases}$$

- 当 $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ 离相关系数就可以写成：$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \left\| f_{X,Y}^n(t,s) - f_X^n(t) f_Y^n(s) \right\|_w^2$
- 其中

$$f_{X,Y}^n(t,s) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle t, X_k\rangle + i\langle s, Y_k\rangle\}$$

$$f_X^n(t) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle t, X_k\rangle\}$$

$$f_Y^n(s) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle s, Y_k\rangle\}$$

距离相关性

- 利用文献[1]的定理一，可以证明：

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}$$

- $$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$$
$$B_{kl} = b_{kl} - \bar{b}_{k} - \bar{b}_{.l} + \bar{b}_{..}$$

$a_{kl} = \|X_k - X_l\|_p, b_{kl} = \|Y_k - Y_l\|_q, k, l = 1, \ldots, n$

- 那么，距离相关性为：

$$\mathcal{R}_n^2(\mathbf{X},\mathbf{Y}) = \begin{cases} \frac{V_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{v_n^2(\mathbf{X})V_n^2(\mathbf{Y})}}, & V_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0 \\ 0, & V_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) = 0 \end{cases}$$

[1] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. Annals of Statistics, 35(6), December 2007.

距离协方差统计量

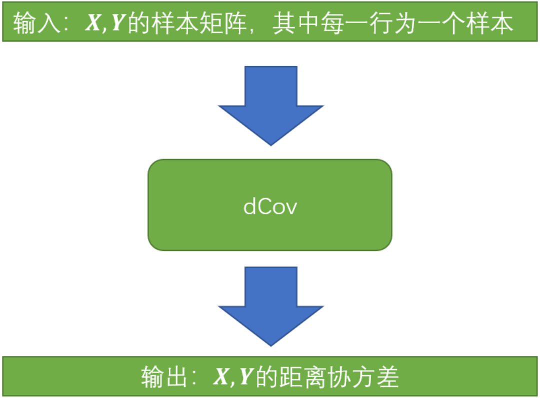
```r
dCov <- function(x, y) {
x <- as.matrix(x)
y <- as.matrix(y)
n <- nrow(x)
m <- nrow(y)
if (n != m || n < 2) stop("Sample sizes must agree")
if (! (all(is.finite(c(x, y)))))
stop("Data contains missing or infinite values")
```

距离协方差统计量

```r
Akl <- function(x) {
d <- as.matrix(dist(x))
m <- rowMeans(d)
M <- mean(d)
a <- sweep(d, 1, m)
b <- sweep(a, 2, m)
return(b + M)
}

A <- Akl(x)
B <- Akl(y)
dCov <- sqrt(mean(A * B))
dCov
```

距离协方差统计量



样本独立性检验

- H_0：$F_{XY} = F_X F_Y \quad vs \quad$ 备择假设 H_1：$F_{XY} \neq F_X F_Y$
  其中，$X \in \mathbb{R}^p, Y \in \mathbb{R}^q$
- 检验统计量：$n\mathcal{V}_n^2(X, Y)$
- 性质：当 X, Y 互独立，且 $E\left(|X|_p + |Y|_q\right) < \infty$ 时，
  $n\mathcal{V}_n^2 \xrightarrow[n\to\infty]{D} \|\zeta(t,s)\|^2$，其中 $\zeta(t,s)$ 为均值 0，方差为
  $R\left(u, u_0\right) =$
  $\left(f_X\left(t-t_0\right) - f_X(t)\overline{f_X\left(t_0\right)}\right)\left(f_Y\left(s-s_0\right) - f_Y(s)\overline{f_Y\left(s_0\right)}\right)$
  的复高斯随机过程。(定理五 [1])

操作步骤

- 根据原始样本，计算检验统计量 $n\mathcal{V}_n^2(X,Y)$
- 对 Y 生成一个置换 Y^* 计算 $n\mathcal{V}_n^2(X,Y)^{(b)}$
- 重复步骤 2，直到达到次数上限
- 计算 p 值，其中：

$$\hat{p} = \frac{\left\{1 + \sum_{b=1}^{B} I\left(nv_n^2\left(X, Y^*\right)^{(b)} \geq nv_n^2(X,Y)\right)\right\}}{B+1}$$

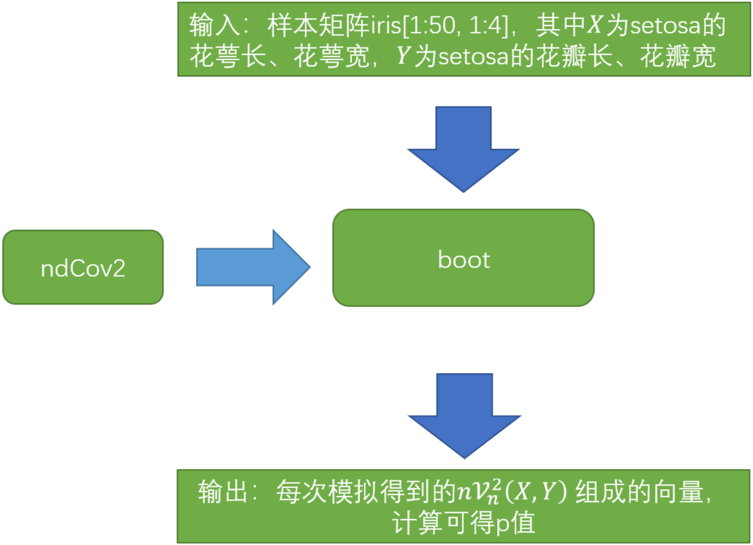
- 当 p 值小于显著性水平 α 时，拒绝原假设；否则接受原假设

距离协方差检验

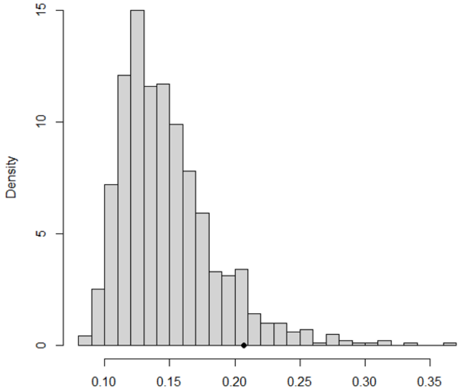
```
ndCov2 <- function(z, ix, dims) {
p <- dims[1]
q1 <- p + 1
d <- p + dims[2]
x <- z[ , 1:p]
y <- z[ix, q1:d]
return(nrow(z) * dCov(x, y)^2)
}


library(boot)
z <- as.matrix(iris[1:50, 1:4])
boot.obj <- boot(data = z, statistic = ndCov2,
R = 999,sim = "permutation", dims = c(2, 2))
tb <- c(boot.obj$t0, boot.obj$t)
mean(tb >= boot.obj$t0)
## [1] 0.059
```

距离协方差检验


输入：样本矩阵iris[1:50, 1:4]，其中X为setosa的花萼长、花萼宽，Y为setosa的花瓣长、花瓣宽
ndCov2
boot
输出：每次模拟得到的$n\mathcal{V}_n^2(X, Y)$组成的向量，计算可得p值

dCov 统计量重复实验直方图



- p 值为 0.059

和 Wilks Lambda 检验作比较

- Wilks Lambda 检验：$\underline{X} = \left[\underline{X}_1', \underline{X}_2'\right]' \sim N_p(\underline{\mu}, \Sigma)$，其中：

$$\underline{\mu} = \left[\underline{\mu}_1', \underline{\mu}_2'\right]', \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- 检验统计量：

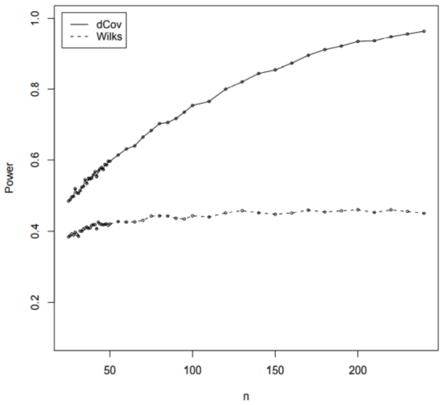
$$\Lambda = \frac{|V|}{|V_{11}||V_{22}|}$$

其中 V_{ii} 为 Σii 的极大似然估计，V 为 Σ 的极大似然估计[2]

[2]Luís Miguel Grilo, Coelho C A . The Exact and Near-Exact Distributions for the Wilks Lambda Statistic Used in the Test of Independence of Two Sets of Variables[J]. American Journal of Mathematical & Management Sciences, 2010, 30(1-2):111-145.

dCov 的功效

Statistical Computing with R



Thanks!