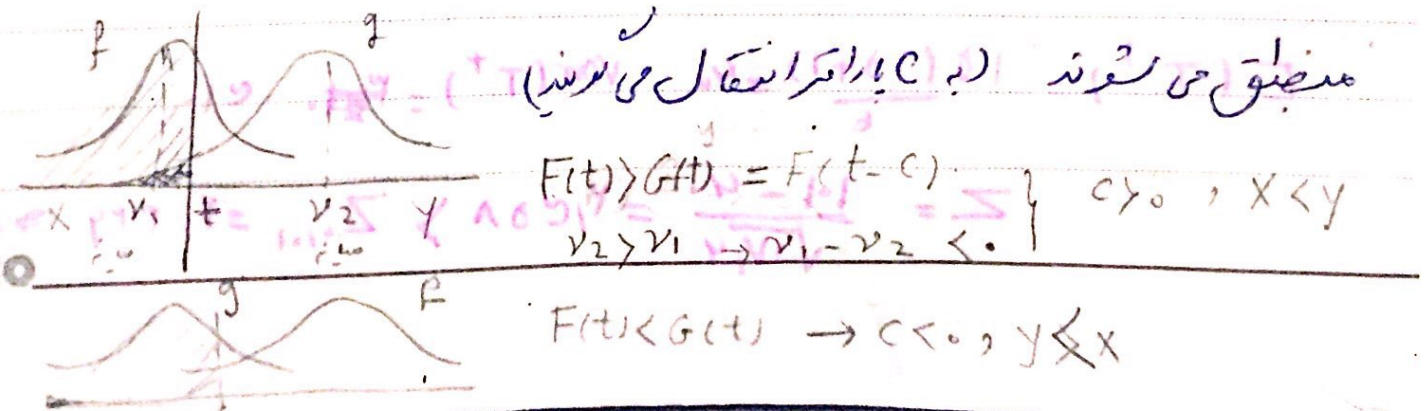
مقایسه دو جامعه با یکدیگر :

فرض کنید دو جامعه داریم که جامعه اول با متغیر تصادفی X از توزیع پیوسته ای مثل $F_X(x)$ توصیف می‌شود و جامعه دوم با متغیر تصادفی Y از توزیع پیوسته ای مثل $G_Y(y)$ توصیف می‌شود. می‌خواهیم فرض $H_0 : F_Z(z) = G_Z(z)$ را آزمون کنیم برای این منظور دو نمونه تصادفی مستقل از دو جامعه اختیار کنیم به صورت $X_1, \dots, X_m$ و $Y_1, \dots, Y_n$ و در پایان نتایج حاصل از این دو نمونه نسبت به فرض H_0 اظهار نظر می‌کنیم. یکی از جمله آزمون‌های ناپارامتری که برای بررسی فرض H_0 استفاده می‌شوند آزمون جمعی رتبه‌ای ویلکاکسون و آزمون من - ویتنی می‌باشند که آن‌ها را مورد مطالعه قرار می‌دهیم.

- آزمون جمعی - رتبه‌ای - ویلکاکسون : Wilcoxon rank-sum test

برای این آزمون ابتدا فرض می‌کنند دو نمونه تصادفی مستقل از دو تابع توزیع پیوسته $F_X(x)$ و $G_Y(y)$ در اختیار است به طوری که $G(z-c) = F(z)$ یعنی نمودار دو تابع F و G منطبق از نقطه نظر ظاهری با یکدیگر اختلاف دارند و با انتقال به اندازه C هم منطبق می‌شوند (به C پارامتر انتقال می‌گوییم)



۲۰

Subject:
Date

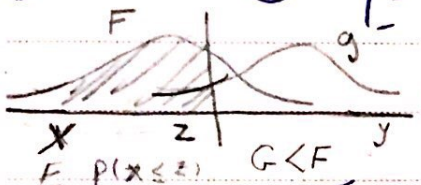
$\begin{cases} H_0: C=0 \\ H_1: C\neq 0 \end{cases}$ یا $\begin{cases} H_0: C=0 \\ H_1: C<0 \end{cases}$ یا $\begin{cases} H_0: C=0 \\ H_1: C>0 \end{cases}$ پس فرضیه به صورت

"less" "greater"

$F \neq G$ $F<G$ ، $X>Y$ به طور تقریبی $F>G$ ، $Y>X$ به طور تقریبی $G(z)=F(z-C)$ ، $C>0$ ، $G<F$

در می آید. ابتدا فرضیه اول $\begin{cases} H_0: C=0 \\ H_1: C>0 \end{cases}$ را در نظر می گیریم برای تعریف آماره آزمون

$F>G$



برای آزمون کردن فرضیه در سطح α مراحل زیر را انجام می دهیم :

① با فرض این که H_0 درست است (یعنی دو نمونه از یک جامعه به دست آمده اند) سپس آن ها را با هم مخلوط می کنیم نمونه ای به حجم $m+n$ از X و Y داریم که این نمونه جدید هر دو جامعه را توصیف می کند

② نمونه مخلوط شده را مرتب کرده و در این ترتیب رتبه X_i را R_i و رتبه Y_j را S_j می نامیم.

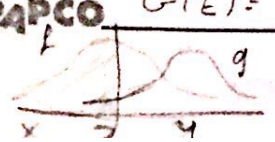
③ آماره ای زیر آماره ای جمعی رتبه ای ویلکاکسون نام دارند و هر یک از دو آماره می توانند برای آزمون فرض H_0 در مقابل H_1 به کار روند.

$$W_X=\sum_{i=1}^{m} R_i \qquad W_Y=\sum_{j=1}^{n} S_j \qquad \text{if } N=n+m \Rightarrow W_X+W_Y=\frac{N(N+1)}{2}$$

④ اکنون فرض H_0 در صورتی رد خواهد شد که H_1 درست باشد و درست بودن H_1 به این معنی است که X گرایش به بزرگ شدن از Y دارند در نتیجه رتبه ای بزرگتری را بدست آورند و W_X بزرگ می گردد در نتیجه H_0 را به ازای مقادیر بزرگ W_X رد خواهیم کرد

PAPCO

$G(z)=F(z-c)$



$C>0 \rightarrow F>G \Rightarrow P(X\leq z) \geq P(Y\leq z)$
$P(X>z)<P(Y>z)$

Subject:
Date

$W_{\alpha(m,n)}$

$W_x \le k_\alpha$ یا $p\text{-value} = p(W_y \ge k_\alpha) \le \alpha \Rightarrow RH_0$

به همین ترتیب برای فرضیه دوم و سوم داریم

$p\text{-value} = p(W_y \le k_\alpha) \le \alpha \Rightarrow RH_0$

$p\text{-value} = 2p(W_y \ge k_\alpha) \le \alpha \Rightarrow RH_0$

توزیع W_x تحت فرض H_0:

از این موضوع که مرتب شدن $N=m+n$ شیء به هر ترتیبی احتمال یکسانی تحت فرض H_0 دارد، پس

$$p(W_x=w) = \frac{\text{تعداد حالات مطلوب}}{\text{تعداد حالات ممکن}} = \frac{\text{تعداد حالات مطلوب}}{C^{n}_{n+m}}$$

مثلاً برای $m=3$ و $n=2$ داریم:

1	2	3	4	5	W_x
y	y	x	x	x	12
y	x	y	x	x	11
x	y	y	x	x	10
y	x	x	y	x	10
x	y	x	y	x	9
y	x	x	x	y	9
x	x	y	y	x	8
x	y	x	x	y	8
x	x	y	x	y	7
x	x	x	y	y	6

$$p(W_x=w) = \begin{cases} \frac{1}{10} & w=12 \\ \frac{1}{10} & w=11 \\ \frac{2}{10} & w=10 \\ \frac{2}{10} & w=9 \\ \frac{2}{10} & w=8 \\ \frac{1}{10} & w=7 \\ \frac{1}{10} & w=6 \end{cases}$$

$$p(W_x=w) = \begin{cases} \frac{1}{10} & w=12,11,7,6 \\ \frac{2}{10} & w=10,9,8 \end{cases}$$

Subject:
Date

محاسبه میانگین و واریانس W_X (تذکر فن ۱۸)

برای هریک از رشته‌های $R_1 \dots R_m$ در نمونه‌گیری $N = m+n$ تایی $X_1 \dots X_m$ $Y_1 \dots Y_n$ ~~وقتی $X_1 \dots X_m$ و $Y_1 \dots Y_n$ داده‌های مربوط به رشته R_i رشته‌ها~~ که $i = 1 \dots m$ یعنی R_i یکی از اعداد $1 \dots N$ را با احتمال مساوی می‌پذیرد. بنابراین برای آماره‌ی رشته‌ای داریم:

$$E(R_i) = \frac{N+1}{2} \qquad V(R_i) = \frac{N^2-1}{12} = \frac{(N+1)(N-1)}{12}$$

$$Cov(R_i, R_j) = -\frac{N+1}{12} \qquad i \neq j$$

$$E(W_X) = E\left(\sum_{i=1}^{m} R_i\right) = \sum_{i=1}^{m} E(R_i) = \frac{m(N+1)}{2} = \frac{m(m+n+1)}{2}$$

$$Var(W_X) = var\left(\sum_{i=1}^{m} R_i\right) = \sum_{i=1}^{m} var(R_i) + \sum_{i \neq j} cov(R_i, R_j)$$

$$= \frac{m(N^2-1)}{12} - \frac{m(m-1)(N+1)}{12} = \frac{mn(m+n+1)}{12}$$

$$\frac{m(N-1)(N+1) - m(m-1)(N+1)}{12} = \frac{m(N+1)(\overbrace{N-1-m+2}^{n})}{12}$$

به همین ترتیب:

$$E(W_Y) = \frac{n(n+m+1)}{2}, \quad var(W_Y) = \frac{mn(m+n+1)}{12}$$

نکته: می‌توان نشان داد

$$P(W_X \leq \omega_k) = P\left(W_X \leq \frac{m(m+1)}{2} + \omega_1\right)$$

$$= P\left(W_Y \leq \frac{n(n+1)}{2} + \omega_1\right)$$

$m = 3, n = 2$

$$P(W_X \leq 10) = P\left(W_X \leq \frac{\overbrace{3 \times 4}^{6}}{2} + 4\right) = P\left(W_Y \leq \frac{\overbrace{2 \times 3}^{4}}{2}(3+1) + 5\right)$$

$$= P(W_Y \leq 7)$$

اثبات:

$$W_X + W_Y = \frac{N(N+1)}{2} = \frac{(m+n)(m+n+1)}{2}$$

$$\frac{m(m+n+1)}{2} + \frac{n(m+n+1)}{2}$$

Subject:
Date

ادامه صفحه ۹۰

$$W_X - E(W_X) \overset{d}{=} E(W_X) - W_X$$

از طرفی با توجه به متقارن بودن W_X حول $E(W_X) = \frac{m(m+n+1)}{2}$:

$$W_X - \frac{m(m+n+1)}{2} = \frac{n(m+n+1)}{2} - W_Y \overset{d}{=} W_Y - \frac{n(n+m+1)}{2}$$

با حذف $\frac{mn}{2}$ از طرفین :

$$W_Y - \frac{n(n+1)}{2} \overset{d}{=} W_X - \frac{m(m+1)}{2}$$

نکته ۲ : از آنجایی که $W_X = \sum_{i=1}^{m} R_i$ وقتی m بزرگ شود می‌توان قضیه حد مرکزی را به کار برد و

$$Z = \frac{W_X - E(W_X)}{\sqrt{Var(W_X)}} \sim N(0,1)$$

مثال ۲۰ : چهار دانش‌آموز اول ابتدایی با روش قدیم و سه نفر با روش جدید، عدد نویسی را آموزش داده می‌شود. پس از مدتی نمره‌ای زیر به دست آورده شد. $\alpha = 0.01$

روش‌های متفاوت

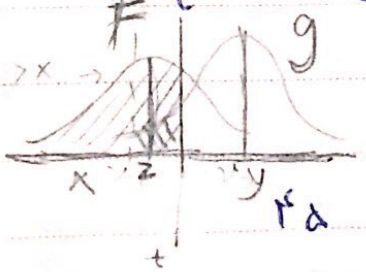
روش قدیم $m=4$: X : ۴۵ ، ۵۹ ، ۶۱ ، ۷۲

روش جدید $n=3$: Y : ۷۰ ، ۷۳ ، ۷۵

آیا روش جدید بهتر است. $\alpha = 0.01$

H_0 : دو روش یکی هستند $H_0: F_X(z) = G_Y(z)$ یا $C = 0$ یا $\theta_1 = \theta_2$

H_1 : روش جدید بهتر است $H_1: F(z) > G(z)$ یا $\theta_2 > \theta_1$ یا $C > 0$ یا $\theta_1 - \theta_2 < 0$



داده‌ها را ردیف می‌کنیم و رتبه می‌دهیم

۴۵	۵۹	۶۱	۷۰	۷۲	۷۳	۷۵
X	X	X	Y	X	Y	Y
۱	۲	۳	۴	۵	۶	۷

$$\begin{cases} W_X = 1+2+3+5 = 11 \\ W_Y = 4+6+7 = 17 \end{cases}$$

(۲۳) pwilcox(11, m, n) = 0.9428

Subject:
Date

جدول m×n دارد

$P(W_x \le 11) = 0.971428 \Rightarrow$ AH.

$\frac{n(m+n+1)}{2}$:

$\frac{m(m+n+1)}{2} = 14$ $E(W_Y) = \frac{n(n+m+1)}{2} = 12$ $V(W_Y) = \frac{12(8)}{2} = 48$

X <- c(59, 61, 72, 60)
Y <- c(70, 90, 76, 70)

$P(W_Y \le 14) = P(Z \le \frac{14-12}{\sqrt{48}}) = P(Z \le 0.2887)$

$P(W_X \le 11)$

مثال ۲۱: دو نوع پلاستیک که هرکدام توسط یک فرآیند متفاوت تولید شده است برای استحکام (دو نوع پلاستیک (دو نمونه مستقل))
نمونه‌ی مورد آزمایش قرار گرفته اندازه‌ای استحکام در جدول زیر نمایش داده شده آیا این اطلاعات

به این است آیا این داده‌ها می‌توانند تفاوتی بین توزیع استحکام دو پلاستیک را نشان دهد $\alpha = 0.05$

plastic 1	15.3 (2)	18.7 (6)	22.3 (10)	17.6 (4)	19.1 (7)	14.8 (1)
plastic 2	21.2 (9)	22.4 (11)	18.3 (5)	19.3 (8)	17.1 (3)	27.7 (12)

$\begin{cases} H_0: C = 0 \\ H_1: C \neq 0 \end{cases}$ $n = 6$ $m = 6$ $N = 12$

$W_X = \sum_{i=1}^{m} R_i = 9 + 11 + 5 + 8 + 3 + 12 = 48$

$W_X = \sum_{i=1}^{m} R_i = 2 + 6 + 10 + 4 + 7 + 1 = 30$

$(W_X \ge 30) = 1 - P(W_X \le 29) = 0.0032$ $P(W_X \le 30) = 0.9794$ $0.0032 \times 2 = 0.0064 < 0.05$

$E(W_X) = 6 \times 13/12 = 39$ $var(W_X) = 39$ $\Rightarrow$ AH

$P(W_X \ge 30) = 1 - P(W_X \le 29) = 1 - P(Z \le \frac{29 + 0.5 - 39}{\sqrt{39}}) = 1 - P(Z \le -1.44)$

$1 - [1 + \Phi(1.44)]$

$= 1 - \Phi(1.36) = 0.086$ $= 0.0764$

p-value $= 2 \times 0.086 = 0.173 \Rightarrow$ AH. p-val $= 2 \times 0.0764$ $= 0.1528 \rightarrow$ AH

PAPCO

Subject: _____
Date

استفاده از آزمون رتبه علامت ویلکاکسون برای نمونه‌های وابسته :

مثال ۲۲ : از هشت نفر خواسته شده که در دو حالت عادی و استرس زا تکلیف یادگیری ساده را بچینند. در حالت استرس زا به آن گفته شده است که سه دقیقه شوک خفیف پس از آغاز آزمایش و ۳۰ ثانیه پس از آن یک وارد خواهد شد و زمانی که به پایان رسیده مشخص خواهد شد. آن اندازه‌گیری می‌شود داده‌ها در جدول زیر ارائه شده است

آیا شواهد کافی در اطلاعات موجود است که نشان دهد تحت شرایط استرس زا تکمیل تکلیف سریعتر است. از آزمون رتبه علامت ویلکاکسون استفاده کنید

نرمال A	استرس B	D_i	$\|D_i\|$	R_i	Z_i	$R_i Z_i$
۱۲۶	۱۳۰	۴	۵	۵	۱	۵
۱۱۷	۱۱۸	۱	۱	۱	۱	۱
۱۱۵	۱۲۵	۱۰	۴	۴	۱	۴
۱۱۸	۱۲۰	۲	۳	۳/۵	۱	۳/۵
۱۱۸	۱۲۱	۳	۳	۳/۵	-۱	-۳/۵
۱۲۸	۱۲۵	-۳	۳	۳/۵	-۱	-۳/۵
۱۲۵	۱۲۰	۵	۵	۶	۱	۶
۱۲۵	۱۲۰	۵	۵	۶	۱	۶
۱۲۰	۱۲۰	۰	۰	۷	۰	۰
						۲۱

$$\begin{cases} H_0: B - A = 0 \rightarrow H_0: M_D = 0 \\ H_1: B - A > 0 \quad H_1: M_D > 0 \end{cases}$$

نکته ۱ : در مطالعات زوجی (نمونه دوتایی یا نمونه وابسته) بردار Z_i به جای ۰ و ۱ باید ۱- و ۱ در نظر بگیرید در این صورت با تعریف $Z_i = \begin{cases} -1 & \text{با احتمال } \frac{1}{2} \\ 1 & \text{با احتمال } \frac{1}{2} \end{cases}$ و

$T = \sum R_i Z_i$ $E(T^+) = 0$ $V(T^+) = \frac{n(n+1)(2n+1)}{6}$

مجموع رتبه‌های مثبت + مجموع رتبه‌های منفی −

داریم $T = T^+ + T^- = \frac{n(n+1)}{2}$

$$\begin{cases} T^- = \frac{n(n+1)}{4} - \frac{T}{2} \\ T^+ = \frac{n(n+1)}{4} + T/2 \end{cases} \qquad \begin{cases} T = \sum R_i Z_i \\ (T^+) \end{cases}$$

PAPCO

Subject:
Date

$$\Rightarrow P(T \le c) = P\left(\frac{T}{2} + \frac{n(n+1)}{4} \le \frac{c}{2} + \frac{n(n+1)}{4}\right) = P\left(T^{+} \le \frac{c}{2} + \frac{n(n+1)}{4}\right)$$

نکته ۱: درصورتی که داده‌ها گره داشته باشند یعنی به هر دلیلی داده‌ها یکسان گزارش شده باشند یعنی نمونه مستقل $x_1 \dots x_m$ و $y_1 \dots y_m$ برخی داده‌ها یکسان باشند در این صورت باید متوسط مرتبه‌ها به دست آورده شود. برای مثال سه تا کمترین داده اول یکسان باشند $(1+2+3)/3 = 2$ به هر سه رتبه ۲ می‌دهیم. در این صورت

$$E(W_X^*) = \frac{m(N+1)}{2}$$

$$Var(W_X^*) = \frac{mn(N+1)}{12} - \frac{mn\sum_{i=1}^{k}(t_i^3 - t_i)}{12N(N-1)}$$

$E(T^*) = 0, V(T^*) = \frac{n(n+1)(2n+1)}{6} - \frac{\sum(t_i^3 - t_i)}{12}$ و برای نمونه دوم (t_i تعداد یافته‌ها، k تعداد دسته‌های گره‌دار)

نکته ۲: اگر در دو رصفت مقدار ثابتی داشته باشیم، این دو صفت نمونه حذف می‌شود

۱۲۰ بعد ۱۲۰ قبل

نکته ۳: تعداد افرادی که نمره ثابت در دو صفت داشته باشند از نمونه حذف می‌شوند

ادامه مثال: $n=7 \quad k=1 \quad t_1=2$ (تعداد یافته‌ها) $\quad E(T^*)=0$

$$Var(T^*) = \frac{7(7+1)(2\times 7+1)}{6} - \frac{2^3-2}{12} = 19.5$$ دو نمونه داریم

$$p\text{-value} = P(T \ge 21) = 1 - P(T \le 21)$$

$$= 1 - P\left(T \le \frac{21 + 0.5}{\sqrt{19.5}}\right) = 1 - \Phi(1.72) = 0.0413 < 0.05$$

$\Rightarrow RH_0$

PAPCO

x <- c(120, ... 120) y <- c(130, 90, ... 120) wilcox.test(y, x, correct=T, paired=T, alternative="greater")

Subject:
Date

در نمونه دابته

مثال ۲۲. دو روش A و B برای کنترل ترافیک در ۱۲ تقاطع برای مدت یک هفته به کار گرفته شده است. تعداد حوادثی که در این دوره زمانی اتفاق افتاده در جدول زیر ثبت شده است. با ترتیب: کارایی این دو روش با یکدیگر متفاوت است؟

- با استفاده از آزمون علامت بررسی کنید دو روش تفاوت معنی داری دارد؟
- $\alpha = 0.1$ " " رتبه ای - شبه نرمال و ملاحظات کمی

A	B	D_i	$\lvert D_i \rvert$	R_i	Z_i	$R_i Z_i$
5	4	-1	0	—	—	—
6	4	-2	1	3.5	-1	-3.5
8	9	+1	1	3.5	+1	+3.5
3	2	-1	1	3.5	-1	-3.5
6	3	-3	1	3.5	-1	-3.5
1	0	-1	1	3.5	+1	+3.5
2	3	+1	1	3.5	-1	-3.5
4	1	-3	2	7.5	-1	-7.5
7	9	+2	2	7.5	+1	+7.5
5	2	-3	3	10	-1	-10
6	5	-1	3	10	-1	-10
1	1	0	3	10	-1	-10
						-37

$\frac{11}{4} = 3.5$

$$\begin{cases} H_0: M_D = 0 \\ H_1: M_D \neq 0 \end{cases} \qquad B = 8$$

$$p\text{-value} = 2\min\{P(B \le 8), P(B \ge 8)\}$$

$$= 0.2266 \rightarrow A H_0$$

$$> 0.1$$

PqPCO

Subject:
Date

$n=11 \quad k=3 \quad t_1=4 \quad t_2=2 \quad t_3=3$

$$var(T) = \frac{n(n+1)(2n+1)}{4} - \frac{\sum(t_i^3 - t_i)}{12}$$

$$= \frac{11(11+1)(2\times11+1)}{4} - \frac{(2^3-2)+(4^3-4)+(5^3-3)}{12} = 484$$

$$E(T) = 0$$

$$2\min\{p(T \geq 27), p(T \leq -27)\} = 2\times 0.0154 = 0.0308$$

$$p(T \leq -27) \neq p\left(Z \leq \frac{-24.5}{\sqrt{484}}\right) = 1-\Phi(1.295) = 0.0484$$

$$\cdot = 2\times 0.0484 = 0.0968 < 0.1$$

چون خیلی کوچک است دنبال نیست $\rightarrow RH_0$

$n=11, \quad T=-27$

$$p(T \leq -27) = p\left(\frac{T}{2} + \frac{n(n+1)}{4} \leq \frac{-27}{2} + \underbrace{\frac{11(11+1)}{4}}_{15}\right)$$

$$= p(T^+ \leq 15.5) = 0.0581$$

R: psignrank(15, 11)

$$p\text{-value} = 2\times 0.0581 = 0.1162 > 0.1$$

$\Rightarrow AH_0$

Subject:
Date

مثال ۲۴ . از ده بیمارستان داده های زیر مربوط به اثربخشی دو داروی درمان
کبد بیماران جمع آوری شده است . آیا اطلاعات تفاوت یا اختلاف
در میزان بازیابی دو گروه را نشان می دهد . (از آزمون علامت)

بیمارستان	A	B	Di	
1	75	85.4	10.4	+
2	69.8	83.1	13.3	+
3	85.7	80.2	-5.5	۰
4	74	74.5	۰.۵	+
5	69	70	1	+
6	83.3	81.5	-1.8	۰
7	68.9	75.4	6.5	+
8	77.8	7.2	1.4	+
9	72.2	85.4	13.2	+
10	77.4	80.4	3	+

$B = 8$

$$\begin{cases} H_0 : M_D = 0 \\ H_1 : M_D \neq 0 \end{cases}$$

$$p_value = 2\min(P(B \geq 8), P(B \leq 8))$$

$$= 2\min\{1 - P(B \leq 7), P(B \leq 8)\}$$

$$= 2\min\{1 - 0.9453, 0.9893\} =$$

$$= 2 \times 0.0547 = 0.1094 \rightarrow AH_0$$

از آزمون t نمی توان استفاده کرد چون نرمال بودن داده ها رد می شود و
نمی توان t زوجی بکار گرفت)

Subject: (۲۵)

Date

```
x<-c(75,69.....,77.4)
y<-c(85.4..... 80.4)
d=y-x
d=d[d=!0]
n=length(d)
B=length(d[d>0])
p-value=2*min(pbinom(B,n,1/2),1-pbinom(B-1,n,1/2))
p-value
```

مثال: امتیازهای دو گروه A و B از دانشجویان که به ترتیب شامل ۱۰ و ۱۰ نفری باشند

در سطح ۵٪ مقایسه کنید. نتایج یک آزمون برای این دانشجویان به صورت امتیازی عالی، متوسط و ضعیف

می باشد. نتایج آزمون به تفکیک گروه ها در جدول زیر آمده است در سطح معنی دار

۵٪ با استفاده از آزمون جمعی رتبه ای ویلکاکسون بررسی کنید آیا دو گروه در این

آزمون یکسان اند.

$\begin{cases} H_0: C=0 \\ H_1: C\neq 0 \end{cases}$

	عالی	متوسط	ضعیف
A	4	3	3
B	3	5	2

۲۸/۷ = ۱۲ ۱۱ ۱۰ ۹ ۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱

A A A A B B B A A A B B B B B A A A B B

۱۴ ۱۳ ۱۲ ۱۱ ۱۰ ۹ ۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱

K = 1 , 2 , 3

$t_1=7$ $t_2=8$ $t_3=5$ $m=10$ $n=10$

Rank: ۴ ۴ ۴ ۴ (۴ ۴ ۴) ۱۱/۵ ۱۱/۵ ۱۱/۵ (۱۱/۵ ۱۱/۵ ۱۱/۵ ۱۱/۵ ۱۱/۵)

۱۸ ۱۸ ۱۸ (۱۸ ۱۸)

$W_B = 3\times 4 + 5\times 11/5 + 2\times 18 = 105/5$

$$p\text{-value} = 2\min\{p(W_B \le 10.5, 5), p(W_B \ge 10.5, 5)\}$$

$$> 0.05 \Rightarrow AH_0$$

از تقریب نرمال:

$$p(W_B \le 10.5, 5) \simeq p\left(Z \le \frac{10.5 + 0.5 - 10.5}{\sqrt{18.9}}\right) \simeq 0.5$$

$$p\text{-value} = 2 \times 0.5 = 1 > 0.05 \Rightarrow AH_0$$

در حالت نمونه:

$$\begin{cases} E(W_B) = n \cdot \frac{(n+m+1)}{2} = 10.5 \\ V(W_B) = nm\frac{(n+m+1)}{12} - \frac{nm\sum(t_i^3 - t_i)}{12N(N-1)} \end{cases}$$

$$= 10 \times 10 \times \frac{21}{12} - \frac{10 \times 10 \left[(7^3-7) + (8^3-8) + (5^3-5)\right]}{12 \times 20 \times 19}$$

$$\simeq 18.9 \qquad 21.0845$$

```
B <- C(rep(1,3), rep(2,5), rep(3,2))
A <- C(rep(4,4), rep(2,3), rep(3,3))
wilcox.test(B, A, alternative = "two.sided", correct=T)
```

$$p(Z \le 0.12\varepsilon)$$

$$\frac{21 - 18}{\sqrt{9.2}}$$

Subject :
Date

۱- آزمون من-ویتنی Mann-whitney

آزمون دیگری که براساس آن می‌توان دو جامعه مستقل از هم را با یکدیگر مقایسه کرد آزمون من ویتنی است که به آزمون U معروف است.

اگر $F(x)$ و $G(y)$ به ترتیب توابع توزیع دو جامعه مذکور و $\underbrace{X_1 \dots X_m}_{F}$ و $Y_1 \dots Y_n$ دو نمونه تصادفی باشند به طوری که F و G پیوسته و نمونه‌ها مستقل فرضیات به صورت

$$\begin{cases} H_0: F(z) = G(z) \\ H_1: F(z) > G(z) \end{cases} = \begin{cases} c=0 \\ c>0 \end{cases} \begin{cases} H_0: F(z) = G(z) \quad c=0 \\ H_1: F(z) < G(z) \quad c<0 \end{cases} \begin{cases} c \le 0 \\ G(z) \neq F(z) \quad c \neq 0 \end{cases}$$

برای آزمون فرضیات بالا در سطح α مراحل زیر را انجام می‌دهیم

① ابتدا دو نمونه را در هم ادغام و مرتب می‌کنیم

مثلاً X X Y Y X Y

② متغیر تصادفی زیر را در نظر می‌گیریم

$$D_{ij} = \begin{cases} 1 & \text{if } x \text{ سمت چپ } y \quad y_j < x_i \quad i = 1 \sim m \\ 0 & \text{if } x \text{ سمت راست } y \quad y_j > x_i \quad j = 1 \sim n \end{cases}$$

آماره آزمون را به صورت

تعداد xهایی که در سمت راست yها هستند

$$U = W_{YX} = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

or

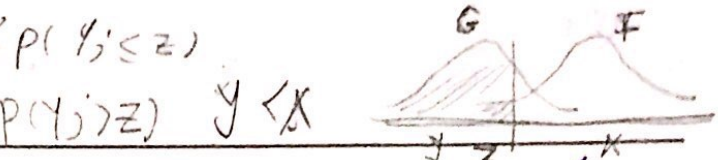
تعداد yهایی که در سمت راست xها هستند

$$U = W_{XY} = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

آماره آزمون من ویتنی

$c<0$ $F<G \Rightarrow P(X_i \le z) \le P(Y_j \le z)$

$\Rightarrow P(X_i > z) > P(Y_j > z)$ $Y < X$



فرض کنید $\begin{cases} H_0: F(z) = G(z) \\ H_1: G(z) > F(z) \end{cases}$ راه آزمون کمپین روشن است که H_0 را وقتی H_1 درست باشد روی کمپین نفی وقتی $Y < X$ باشد و به عبارت دیگر X گرایش به راست دارند (X گرایش بزرگ شدن از Y دارند) پس بدید u و

$F < G \Rightarrow c<0$

$u = W_{YX}$ زیاد شود لذا فرض H_0 را برای مقادیر بزرگ $u = W_{YX}$ یا برای مقادیر کوچک W_{XY} پس

رد می کنیم یعنی if $u = W_{YX} \ge K_\alpha$ رد می شود H_0 و برعکس $\Rightarrow RH_0$

$W_{XY} \le K_\alpha$

ترتیب برای فرض مقابل $G(z) < F(z)$ و $u = W_{YX} \le K_\alpha \Rightarrow RH_0$

$W_{XY} \ge K_\alpha \Rightarrow RH_0$ $F(z) > G(z)$

و $G(z) \ne F(z)$.

$u = W_{YX} \le K_{1\alpha}$ یا $W_{YX} \ge K_{2\alpha}$

- توزیع $u = W_{YX}$:

اگر m و n مشخص برشمرده $m = 2$ ، $n = 2$

	$u = W_{YX}$	$P(u=u)$
x x y y y	0	1/10
x y x y y	1	0/1
x y y x y	2	0/2
x y y y x	3	0/2
y x x y y	4	0/2
y x y x y	5	—
y x y y x	6	0/2
y y x x y	7	—
y y x y x	8	0/1
y y y x x	9	0/1

$$P_{H_0}(u=u) = \begin{cases} 0/1 & u = 0,1,8,9 \\ 0/2 & u = 2,3,4,6 \end{cases}$$

PAPCO

Subject:
Date

- محاسبه صدک‌ها و واریانس U (تست H.)

برای دو متغیر تصادفی X و Y که مستقل از یکدیگر هستند می‌توان نوشت

$$P(Y<X) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f(x)g(y)\,dy\,dx$$

$$= \int_{-\infty}^{\infty} f(x)\,dx \int_{-\infty}^{x} g(y)\,dy = \int_{-\infty}^{\infty} G(x)f(x)\,dx$$

اگر H. در شرط باشد $F(x)=G(x)$

$$\int_{-\infty}^{\infty} F(x)f(x)\,dx = \frac{1}{2}F^2(x)\Big|_{-\infty}^{\infty} = \frac{1}{2}(1-0) = \frac{1}{2}$$

$$\int \frac{2}{2}F(x)f(x)\,dx$$

از اینجا می‌بینیم $E(D_{ij}) = \frac{1}{2}$ و $E(D_{ij}^2) = \frac{1}{4}$ پس $Var(D_{ij}) = \frac{1}{4}$

$$U = \sum\sum D_{ij} \qquad E(U) = \frac{mn}{2}$$

پس

$$Var(U) = \frac{mn(m+n+1)}{12}$$

و به طور مجانبی

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0,1)$$

نکته: رابطه بین W_{xy} و W_y به صورت زیر است

می‌توانند نشان داد

$$W_{xy} = W_y - \frac{n(n+1)}{2}$$

$$W_{yx} = W_x - \frac{m(m+1)}{2}$$

4PCO

Subject:
Date

$$P_{H_0}(W_{YX} \le c) = P_{H_0}\left(W_X - \frac{m(m+1)}{2} \le c\right)$$

$$= P_{H_0}\left(W_X \le c + \frac{m(m+1)}{2}\right)$$

$$E(W_{YX}) = E(W_X) - \frac{m(m+1)}{2} = \frac{mn}{2}$$

$$var(W_{YX}) = var(W_X) = \frac{mn(m+n+1)}{12}$$

پس دقیقاً کمون جمعی و من رتبه‌ی معادلند.

مثال: پانزده باتری تجربی به طور تصادفی از کارخانه A و پانزده باتری استاندارد به طور تصادفی از کارخانه B انتخاب شده است و به طور همزمان این ۳۰ باتری تحت بار الکتریکی یکسان قرار داده شد.

اولین باتری خراب شده از کارخانه‌ی A، دومی از کارخانه B، سومی از کارخانه B... و به ترتیب زیر:

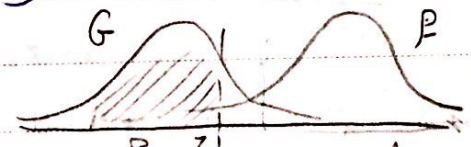
A BBB AB AA BBBB A BA BB B BB A A B A A A B A A A A

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۱۰ ۱۱ ۱۲ ۱۳ ۱۴ ۱۵ ۱۶ ۱۷ ۱۸ ۱۹ ۲۰ ۲۱ ۲۲

با استفاده از آزمون من ویتنی تعیین کنید آیا شواهد کافی برای تأیید این نتیجه که طول عمر باتری تجربی بیشتر از باتری استاندارد است وجود دارد $\alpha = 0.05$

H_0: طول عمر A = طول عمر B

H_1: طول عمر A > طول عمر B



$A > B \rightarrow F_A < G_B \rightarrow c < 0$

$n = 15 \quad m = 15 \quad N = 30$

$W_{AB} \le K_\alpha \Rightarrow RH_0$

$W_B \le K_\alpha$

$$\begin{cases} W_B = 2 + 3 + 4 + 9 + 9 + \dots + 29 = 189 \\ W_{AB} = W_B - \frac{n(n+1)}{2} = 189 - 120 = 69 \end{cases}$$

$E(W_B) = \frac{n(n+m+1)}{2} = 247.5$

$var(W_B) = \frac{nm(n+m+1)}{12} = 618.75$

$p\text{-value} = P(W_B \le 189) = P\left(Z \le \frac{189 + 0.5 - 247.5}{\sqrt{618.75}}\right) = P(Z \le -1.181) = 0.0375$

ولتاژ حدود ولتاژ سول

$C + 120$

PAPCO

$p\text{-value} = P(W_{AB} \le 69) = 0.0375 < 0.05 \rightarrow RH_0$

طول عمر باتری در تجربی بیشتر از استاندارد است

۲۸

Subject:
Date

در R جهت دیدن معادل با من دستی است از wilcox.test

می‌توان استفاده کرد در غیر اینصورت برنامه زیر را اجرا کنید

```
A <- c(1,5,7,8,14,15,20,21,23,24,25,27,28,29,30)
B <- c(2,3,4,6,9,10,11,12,14,16,17,18,19,22,26)
n <- length(A)
m <- length(B)
N <- n+m
C <- c(B,A)
R = rank(C)
WB = sum(R[1:n]
WAB = WB - n*(n+1)/2
WAB
EWB = n*(N+1)/2
VWB = n*m*(N+1)/12
pvalue = pnorm((WB+1/2 - EWB)/sqrt(VWB)
pvalue = pwilcox(WAB, n, m)
pvalue
```

نکته: $p_{H_0}(W_{XY}(m,n) \le c) = p_{H_0}(W_{xy}(n,m) \le c)$

$$W_{XY}(m,n) = W_y(m,n) - \frac{n(n+1)}{2} \overset{d}{=} W_x(m,n) - \frac{m(m+1)}{2}$$

$$\overset{d}{=} W_y(n,m) - \frac{m(m+1)}{2} \overset{d}{=} W_{XY}(n,m)$$

Subject:
Date

مثال: مثال ۴ راستۀ آموزشی من ـ متنی $\alpha = 0/01$

x	x	x	y	x	y	y
۴۵	۵۹	۶۱	۷۰	۷۲	۷۴	۷۵

$W_{YX} = 1$

$$\begin{cases} H_0: C = 0 \\ H_1: C > 0 \end{cases}$$

$$P(W_{YX} \leq k_\alpha) = P(W_{YX} \leq 1) = 0.057$$

$\Rightarrow$ AH۰

in R: pwilcox(1, 3, 4) = 0.057

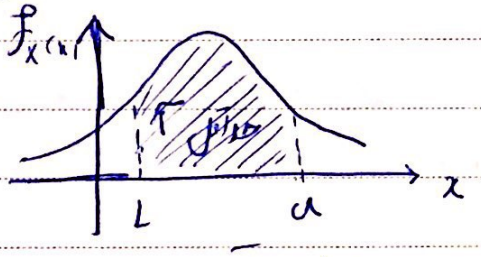
فاصله تحمل:

تعریف: فاصله تصادفی (L,U) یک فاصله تحمل به میزان τ $(0<\tau<1)$ ضریب

$0<\gamma<1$ ی نامند هرگاه

$$(*)\quad P\big(P(L<X<U)\geq\tau\big)=P\big(F_X(U)-F_X(L)\geq\tau\big)=\gamma$$

یعنی فاصله تصادفی (L,U) حداقل $\tau\cdot 100$ درصد از کل احتمال یا سطح زیر منحنی تابع $f_X(x)$ را با $\gamma\cdot 100$ درصد اطمینان تحمل می‌کند.



(فاصله تحمل)

در این درس فاصله تحمل آماره‌های ترتیبی مورد نظر است یعنی

$$P\big(F(Y_k)-F(Y_j)\geq\tau\big)=\gamma$$

$$P\big(F(Y_k)-F(Y_j)\geq\tau\big)=P(V_k-V_j\geq\tau)=P(V_{k-j}\geq R)$$

$$=1-\sum_{r=k-j}^{n}\binom{n}{r}\tau^r(1-\tau)^{n-r}$$

$$=\sum_{r=0}^{k-j-1}\binom{n}{r}\tau^r(1-\tau)^{n-r}$$

Subject:
Date

$$P(F(Y_k) - F(Y_j) \geq \tau) = \gamma = Bin(n, \tau; k-j-1)$$

مثال: نمونه تصادفی $X_1 \dots X_5$ را به آن ترتیب $Y_1 \dots Y_5$

از یک توزیع پیوسته و صعودی $F(x)$ در نظر می‌گیریم احتمال اینکه فاصله (Y_1, Y_5) لااقل ۷۵٪ از کل احتمال را در بر گیرد محاسبه کنید

$$P(F(Y_5) - F(Y_1) \geq 0.75) = B(5, 0.75, 3) = 0.6328$$

نکته ۱: برای انتخاب j و k به طوری که $1-j-k$ تا حداقل

به اندازه γ نزدیک شود باید از جدول توزیع دوجمله‌ای استفاده کرد و به طوری که

$$Bin(n, \tau; k-j-1) \geq \gamma$$

نکته ۲: اگر $n > 25$ از تقریب نرمال استفاده می‌کنیم

$$B(n, \tau, k-j-1) = P(0 \leq B \leq k-j-1)$$

$$= P(-0.5 \leq B_c \leq k-j-0.5)$$

$$= P\left(\frac{-0.5 - n\tau}{\sqrt{n\tau(1-\tau)}} \leq Z \leq \frac{k-j-0.5-n\tau}{\sqrt{n\tau(1-\tau)}}\right)$$

$$= \Phi\left(\frac{k-j-0.5-n\tau}{\sqrt{n\tau(1-\tau)}}\right) - \Phi\left(\frac{-0.5-n\tau}{\sqrt{n\tau(1-\tau)}}\right)$$

$$= z_\gamma$$

معمولا وقتی n به اندازه کافی بزرگ باشد این حالت صحیح است $\tau \to 1/2$ هرچه

PAPCO

گشت یا رُن (Run)

فرض کنید m شی همانند مانند A و n شی شبیه ولی متفاوت مانند B داریم که به صورت

هم گشت و به تفاوف ردیف شده اند مثلاً برای $m=5$ و $n=2$

A A B A A B B A A B B

در این صورت هر گروه از اشیاء شبیه که کنار هم باشند یک گشت یا گروش

(Run) نامیده می شوند. مثلاً در مثال فوق به ترتیب گشت های از A ها ابتدا سپس

B و باز A ... و بالاخره B داریم که تعداد این گشتها روی هم ۶ است. به طور کلی

تعداد گشتهای مربوط به شی اول مثلاً A را با R_1 و تعداد گشتهای مربوط به شی دوم (B) را با R_2

و کل گشتها را با R نشان می دهیم. در مثال $R_1=3$ ، $R_2=3$ ، $R=6$

$$R = R_1 + R_2$$

رابطه بین گشتها

از این که گشتهای دو شی یک در میان قرار می گیرند رابطه زیر را بین R_1 و R_2 داریم

الف) اگر گشتها به ترتیب A شروع و به گشت مربوط به A ختم شوند آن گاه

$$R_1 = R_2 + 1$$

ب) اگر گشتها با B شروع و به گشت B ختم شوند آن گاه $R_2 = R_1 + 1$

ج) اگر با گشت کدام شروع و به یکدیگر ختم شوند آن گاه $R_1 = R_2$

Subject:
Date

توزیع R_1 ، R_2 و R وقتی جایگشت‌ها اتفاقی هستند

در صورتی که m و n داده شود و فرض جایگشت‌ها هم به تصادف توزیع شده باشد می‌توان با استفاده از احتمال متغیرهای R_1 و R_2 و R را به دست آورد

مثال: اگر $m=2$ و $n=3$ باشد

تعداد جایگشت‌های ممکن $= \frac{(m+n)!}{m!\,n!} = \frac{5!}{2!3!} = 10$

که شانس هرکدام از جایگشت‌ها $\frac{1}{10}$ است داریم:

جایگشت	R_1	R_2	R
AA BBB	۱	۱	۲
BAABB	۱	۲	۳
BBAAB	۱	۲	۳
BBBAA	۱	۱	۲
ABABB	۲	۲	۴
ABBAB	۲	۲	۴
ABBBA	۲	۱	۳
BABAB	۲	۳	۵
BABBA	۲	۲	۴
BBAAB	۲	۲	۴

R_1	۱	۲
احتمال	$\frac{4}{10}$	$\frac{6}{10}$

R_2	۱	۲	۳
احتمال	$\frac{3}{10}$	$\frac{6}{10}$	$\frac{1}{10}$

R	۲	۳	۴	۵
احتمال	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{1}{10}$

توزیع مشترک R_1 و R_2:

R_2 \ R_1	۱	۲	حاشیه R_1
۱	۰/۲	۰/۱	۰/۳
۲	۰/۲	۰/۴	۰/۶
۳	۰	۰/۱	۰/۱
حاشیه R_2	۰/۴	۰/۶	۱

۳۰

Subject:

Date

قضیه: اگر m مقدار A و n مقدار B و حالتهای تصادفی با شند تابع احتمال متغیر تصادفی R به صورت زیر است

$$P(R=r)=\begin{cases}\dfrac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{n}} & 2\le r=2k\le n+m \\ \dfrac{\binom{m-1}{k-1}\binom{n-1}{k}+\binom{m-1}{k}\binom{n-1}{k-1}}{\binom{m+n}{n}} & 2\le r=2k+1\le n+m\end{cases}$$

(اثبات کتاب) ←

و خصوصیات آن از امید ریاضی و واریانس زیاد داریم:

$$E(R)=\mu_R=1+2\frac{mn}{m+n}$$

$$V(R)=\sigma_R^2=\frac{(\mu_R-1)(\mu_R-2)}{m+n-1}$$

برای R جدول تهیه شود

- توزیع مجانبی R اگر $n_2\uparrow$ ، $n_1\uparrow$

$$Z=\frac{R-E(R)}{\sqrt{Var(R)}}\sim N(0,1)$$

به همین ترتیب ثابت می شود

$$P(R_1=r_1,R_2=r_2)=\begin{cases}\dfrac{\binom{n-1}{r_1-1}\binom{m-1}{r_2-1}}{\binom{n+m}{n}} & r_1=r_2+1 \\ \dfrac{2\binom{n-1}{r_1-1}\binom{m-1}{r_2-1}}{\binom{n+m}{n}} & r_1=r_2\end{cases}$$

$r_1=1,\dots,n$

$r_2=1,\dots,m$

فرض کنید R_2 مهره از نوع B دارد r_1 ی گذاریم به طوری که هیچکدام از خانه ها خالی نماند به $\binom{n-1}{r_1-1}$ طریق امکان پذیر

m مهره از نوع A دارد

ظرف سیاه

$\binom{m-1}{r_2-1}$ طریق که هیچکدام خالی نماند

PqPCO

Subject:
Date

$$P(R_1 = r_1) = \frac{\binom{n-1}{r_1-1}\binom{m+1}{r_2}}{\binom{n+m}{n}} \qquad r_1 = 1, \dots m$$

$$P(R_2 = r_2) = \frac{\binom{m-1}{r_2-1}\binom{n+1}{r_2}}{\binom{n+m}{n}} \qquad r_2 = 1, 2, \dots m$$

توضیحات بیشتر کتاب صفحه ۱۷۶ تا صفحه ۱۷۷

$n_1 = m$ $n_2 = n$

$$E(R_1) = \frac{m(n+1)}{2} \qquad E(R_2) = \frac{n(m+1)}{2}$$

$$Var(R_1) = \frac{mn(m-1)(n+1)}{(n+m)^2(n+m-1)} \qquad Var(R_2) = \frac{mn(n-1)(m+1)}{(n+m)^2(n+m-1)}$$

کاربرد R :

تا کنون کلی از فرضهای اساسی استنباط آماری این بود که نمونه $X_1, \dots X_n$ یک بخش تقادی از $F_x(x)$ است. اکنون این مسئله را زیر سؤال برده می‌ی تقادی بودن مشاهدات شک می‌کنیم و تقادی بودن مشاهدات را فرض آماری در نظر گرفته و به استفاده R آن را آزمون می‌کنیم.

الف) مشاهدات به صورت کلی هستند

فرض کنید m مشاهده راجع به ویژگی A و n مشاهده راجع به ویژگی B جمع آوری شده‌اند. این اطلاعات همگی نمونه جمع آوری شده‌اند به صورت ردیف

Subject:
Date

صف شوانده :

A A B B B A B A . . . . A B

فرض های آماری عبارتند از :

$$\begin{cases} H_0: \text{مشاهدات تصادفی اند} \\ H_1: \text{مشاهدات تصادفی نیستند} \end{cases}$$

برای یک سطح دلخواه فرض کنید $m = 5$ و $n = 3$ و ثبت اطلاعات به صورت زیر

A A A A A B B B $R = 2$

روشن است که

تصادفی بودن مشاهدات مشکوک است و مایل به رد H_0 هستیم در چنین

حالتی می گوییم بین مشاهدات دنباله روی یا روند وجود دارد . در این حالت

A B A B A B A B A A $R = 7$

$R = 2$ کوچک است اکنون فرض کنید اطلاعات به صورت

ثبت شوانده در این حالت نیز تصادفی بودن یکسب سنت و

بین مشاهدات دوره یا تناوب وجود دارد و در این وضعیت $R = 7$ بزرگ است

به طور کلی فرض H_0 به ازای مقادیر بزرگ و کوچک A رد می شود

$$R \geq k_{1\alpha} \qquad R \leq k_{2\alpha}$$

که $k_{1\alpha}$ و $k_{2\alpha}$ از جدول توزیع R و با استفاده از معادله

$$P_{H_0}(R > k_{1\alpha}) + P_{H_0}(R < k_{2\alpha}) = \alpha$$

قابل حل به است.

مثال: ۱۳ تا سی (B) و ۱۲ اتومبیل شخصی (A) به صورت زیر در جلوی اداره راهنمایی شهر توقف کرده اند تصادفی بودن توقف را در سطح $\alpha = 0.05$ آزمون کنید

AABBBABABAAABBBAABBBBBBABAABA

$\frac{\text{تردید}}{\text{دارد}}$ $12.5 \longrightarrow 11$

$R = 11 < \mu_R = 12.5$ $\mu_R = 1 + \frac{2mn}{m+n} = 1 + 2\frac{13 \times 12}{13+12} = 13.5$ $\frac{156}{25}$

$$\begin{cases} P(R \le 11) = p(z \le -) \\ p(R > 11) = 1 - p(R \le 10) \end{cases} \quad VAR(R) = \frac{(12.5-1)(13.5-2)}{25-1} = 6$$

$$\frac{11 - 12.5}{\sqrt{2.44}}$$

ب) مشاهدات مستقل نیستند

فرض کنید $X_1 \dots X_n$ نمونه تصادفی از توزیع پیوسته $F_X(x)$ دست می خواهیم براساس آزمون ران تست

$$\begin{cases} H_0: \text{داده ها تصادفی هستند} \\ H_1: \text{داده ها تصادفی نیستند} \end{cases}$$

را آزمون کنیم.

ابتدا میانه داده ها را محاسبه کرده و به داده ای بزرگتر از میانه $+$ یا و به داده ای کوچکتر $-$ یا می دهیم در این صورت براساس تعداد حاصل از این نمادگذاری آماره آزمون عبارتی R محاسبه می شود سپس $|Z| > Z_{\alpha/2} \Rightarrow RH_0$

Subject:
Date

مثال: یک نمونه ۲۰ تایی وسیله فلزی انتخاب کرده و مقاومت هر یک را تا لحظه شکست برحسب کیلوگرم به صورت زیر داریم

۱۲۵٫۷ ۹۱۰٫۴ ۸۹٫۴ ۸۶۷٫۹ ۲۴۶٫۷ ۱۳۲ ۲۵۵٫۵ ۱۵۹

۱۵۳٫۱ ۵۸۵٫۸ ۱۸۹٫۸ ۲۰۷٫۲ ۲۰۱٫۹ ۱۱۷٫۶ ۱۵۰٫۴

ص۱۸۳ آتاب ۱۴۱٫۵ ۲۲۲٫۳ ۵۶۷ ۵۱۱ ۵۹۵٫۴

متقاضی بودن این نمونه را در سطح ۰/۰۵ بیازمایید (از توزیع نرمال)

ابتدا مرتب کنید ۸۹٫۴ ۱۱۷٫۶ .....

$$m = \frac{۲۰۱٫۹ + ۲۰۷٫۲}{۲} = ۲۰۴٫۶$$

- + - + + - + - - - - + - + - + + + + -

$R = ۱۳$ $\mu_R = ۱ + \frac{۲ n_1 n_2}{n_1 + n_2} = ۱ + \frac{۲ \times ۱۰ \times ۱۰}{۲۰} = ۱۱$

$R > \mu_R$ $P(R \geq ۱۳)$ $\sigma_R^۲ = \frac{(\mu_R - ۱)(\mu_R - ۲)}{n_1 + n_2 - ۱} = \frac{۹۰}{۱۹}$

$$Z_R = \frac{R - E(R)}{\sqrt{Var(R)}} = \frac{۱۳ - ۱۱}{۲٫۱۷۶}$$

$$P(R \geq ۱۳) = P(Z \geq \frac{۱۲٫۵ - ۱۱}{۲٫۱۷۶}) = P(Z \geq ۰٫۴۸۹)$$

$$= ۱ - ۰٫۶۸۸ = ۰٫۳۱۲ \Rightarrow$$

Subject:
Date

- کاربرد آزمون ران سام در مقایسه دو جامعه با یکدیگر

فرض کنید متغیر تصادفی X صرف جامعه اول با تابع توزیع پیوسته $F_X(x)$ و متغیر تصادفی Y صرف جامعه دوم و با تابع توزیع $G_Y(y)$ است

$$\begin{cases} H_0: F_X(z) = G_Y(z) & \forall z \\ H_1: F_X(z) \neq G_Y(z) & \exists z \end{cases}$$

برای بررسی فرض فوق ابتدا نمونه تصادفی $X_1 \dots X_m$ و $Y_1 \dots Y_n$ را استقلال استخراج می‌کنیم آن‌ها را در هم ادغام و مرتب می‌کنیم. اگر فرض H_0 درست باشد ردیفی ساخته شده انتظار داریم توزیع تصادفی ساخته شده باشد و انتظار داریم X و Y از نظر کمیتی در آن داشته باشند و جابجایی حاصل شان نباید زیاد داشته باشند در غیر این صورت اگر مقادیر بزرگ R می‌تواند دلیلی بر درست بودن H_0 باشد لذا اگر $R \leq k_\alpha \Rightarrow$ رد H_0

مثال: ۴ دانش آموز اول ابتدایی با روش قدیم و ۳ نفر با روش جدید عدد نویسی می‌آموزیم نمرات زیر بدست آمده است

Subject:
Date

روش قدیم	۵۹	۶۱	۷۲	۴۵	۸
روش جدید	۷۸	۷۳	۷۰		۷

دو روش یکی هستند

$$\begin{cases} H_0 : F(z) = G(z) \\ H_1 : f(z) \neq G(z) \end{cases}$$

۴۵	۵۹	۶۱	۷۰	۷۲	۷۳	۷۵
x	x	x	y	x	y	y

$$R = ۴ \qquad \mu_R = 1 + \frac{۲mn}{m+n} = 1 + ۲\,\frac{۴ \times ۳}{۴+۳} = ۴,۴۵$$

$$P(R \leq ۴) = P\left(Z \leq \frac{۴ - ۴,۴۵}{۱,۱۸}\right) \qquad Var(R) = \sigma_R^2 = \sqrt{\frac{(\mu_R - 1)(\mu_R - ۲)}{m+n-1}} = ۱,۱۸$$

~~$P(R > ۴)$~~

$$|Z| = \frac{|۴ - ۴,۴۵|}{۱,۱۸} = ۰,۹۴ \not> ۱,۹۶ \Rightarrow AH_0$$

- آزمون نیکویی برازش $P(Z \leq ۰,۹۴) = ۰,۸۲۷۹ \rightarrow AH_0$

فرض کنید X یک متغیر تصادفی با تابع توزیع $F_x(x)$ است و $F_0(x)$ یک تابع توزیع داده شده معین است. می خواهیم فرض H_0 را آزمون کنیم

$$H_0 : F(x) = F_0(x) \quad \forall x$$

برای آزمون این فرض "آزمون نیکویی برازش" معرفی شده است که شرح ذیل می باشد.

Subject:
Date

کای دو پیرسون

الف) آزمون χ^2 : فرض کنید جامعه‌ای که با متغیر X توصیف می‌شود به k گروه یا دسته‌ای C_1 و ... و C_k افراز شده است که در این صورت می‌توانیم

$$p_i = P_{H_0}(X \in C_i) \quad F(x) \text{ تابع توزیع}$$

$$= \begin{cases} \sum_{C_i} f_0(x) \\ \int_{C_i} f_0(x)\,dx \end{cases}$$

اگر نمونه‌ای تصادفی و به حجم n داشته باشیم انتظار داریم که $e_i = np_i$ از اعضای نمونه در C_i واقع باشد. در واقعیت تعداد X_j ای ($j = 1, \dots, n$) که در C_i قرار گرفته‌اند با e_i تفاوت دارد. در صورتی که H_0 درست باشد انتظار داریم که تفاوت $|O_i - e_i|$ تا حد ممکن کم شود. پیرسون آماره زیر را که برای آن توزیع مجذور χ^2 پیدا کرد و برای $|O_i - e_i|$ های بزرگ معرفی نمود

$$\chi^2_{obs_0} = V = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

که برای n های به قدر کافی بزرگ ثابت می‌شود که توزیع مجانبی آن χ^2 با $k-1$ درجه آزادی است و فرض H_0 در سطح α رد می‌شود اگر

$$\chi^2_{obs_0} \geq \chi^2_{k-1,\alpha}$$

Subject:
Date

دقت کنید: $\sum_{i=1}^{k} O_i = n$ و $\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} np_i = n$

O_i: وابسته — e_i: مستقل

نکته: اگر e_i کمتر از ۵ شد با e_i بعدی c_i را هم ادغام کنید.

مثال: یک تاس ۱۲۰ بار انداخته می‌شود آیا تاس سالم است

$$\begin{cases} H_0: \text{تاس سالم} & H_0: p_i = 1/6 \\ H_1: \text{تاس ناسالم} & H_1: p_i \neq 1/6 \end{cases}$$

c_i	1	2	3	4	5	6
H_0 تحت p_i :	1/6	1/6	1/6	1/6	1/6	1/6
$e_i = np_i$:	۲۰	۲۰	۲۰	۲۰	۲۰	۲۰

پس از یک تاس ۱۲۰ مرتبه

	۱	۲	۳	۴	۵	۶
O_i :	۱۸	۲۳	۱۹	۲۱	۱۸	۲۱

$$\chi_0^2 = \sum \frac{(O_i - e_i)^2}{e_i} = ۲/۵ \qquad \chi^2_{۵,۰/۹۵} = 11$$

$$۲/۵ \not> 11 \Rightarrow AH_0$$

Subject:
Date

مثال: از ۲۰۰ نفر آزمون هوش به عمل آمده و نتایج زیر به دست آمده است

نمره ۲۴ نفر بین ۱ و ۱- و بقیه زیر ۱- و نمره زیر ۱- کمتر از ۳۶ نفر است آیا می‌توان ادعا کرد توزیع نمرات این افراد نرمال استاندارد است (نمرات استاندارد شده)

$$\begin{cases} H_0: Z \sim N(0,1) \\ H_1: Z \not\sim N(0,1) \end{cases} \qquad Z = \frac{X-\mu}{\sigma} \text{ نمره استاندارد هوش}$$

$$P(X \in C_1) = P_1 = \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0,16 = P(Z \le -1)$$

$$P(X \in C_2) = P_2 = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0,68 = P(-1 \le Z \le 1)$$

$$P(X \in C_3) = P_3 = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0,16 = P(Z \ge 1)$$

$$e_1 = np_1 = 200 \times 0,16 = 32 \qquad e_2 = np_2 = 200 \times 0,68 = 136$$

$$e_3 = np_3 = 200 \times 0,16 = 32$$

	$+\infty$ تا 1	1 تا -1	-1 تا $-\infty$
O_i :	۲۴	~~۱۴۰~~ ۱۳۶	~~۳۶~~ ۳۲
e_i :	۳۲	۱۳۶	۳۲

$$\chi^2_{2,0.05} = 5,99$$

$$\chi^2_0 = \sum \frac{(O_i - e_i)^2}{e_i} = 2,94 < 5,99 \Rightarrow H_0$$

Subject:
Date

نکته: درصورتی که تقریب داده شده $F_0(x)$ در فرض به پارامترای بستگی داشته باشد و پارامتر از روی نمونه برآورد شوند به تعداد پارامترای که برآورد می‌شوند از درجه آزادی کمای دو کم می‌شود.

$$k-1-m$$

m: تعداد پارامترای برآورد شده

کاربردهای دیگر χ^2:

- آزمون استقلال: فرض کنید دو ویژگی A و B داشته مثال تحصیلات پدر خانواده و تعداد فرزندان خانواده و یا قد و وزن افراد در جامعه‌ای مورد بررسی است می‌خواهیم این فرض را بسنجیم دو ویژگی A و B استقلال وجود دارد آزمون

B \ A	A_1	A_2	...	A_r
B_1				
⋮				
B_s				

لنکنیم. تحت استقلال

P_{ij} = احتمال قرارگرفتن در سطح A_i و سطح B_j

پس P_{ij} احتمال قرارگرفتن در سطح i ویژگی A و سطح j ویژگی B.

اگر استقلال وجود داشته باشد پس

$$P_{ij} = P_{i\cdot} \cdot P_{\cdot j}$$

تعداد هم برابر حاصلضرب حاشیه‌ای

Subject :
Date

از طرفی می دانیم $\hat{P}_{ij} = \frac{O_{ij}}{n}$ که O_{ij} فراوانی مشاهده شده خانه i و j ام است و $\hat{P}_{i.} = \frac{O_{i.}}{n}$ و $\hat{P}_{.j} = \frac{O_{.j}}{n}$ بین دو صفت مستقل

$$\hat{e}_{ij} = n\hat{P}_{ij} = \frac{O_{i.}}{n} \cdot \frac{O_{.j}}{n} \cdot n$$

و آماره χ^2 :

$$\chi^2_0 = V = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$\text{if } \chi^2_0 \geq \chi^2_{(r-1)(s-1),\alpha} \Rightarrow RH.$$

مثال : در شهری بزرگ ۱۰۰۰ خانواده به تصادف انتخاب می شوند و از هر خانواده درباره تعداد فرزندان و سواد خانواده می پرسیم. می خواهیم بدانیم که آیا این دو صفت مستقلند. بین از سه فرزند دو یا سه فرزند هیچ و یک فرزند

A : A_1 A_2 A_3

بی سواد / با سواد

B : B_1 B_2

B \ A	A_1	A_2	A_3	$O_{.j}$
B_1	۱۸۲ / ۲۰۱٫۶	۲۱۵ / ۲۰۴٫۶	۲۰۳ / ۱۸۷٫۸	۶۰۰
B_2	۱۵۴ / ۱۳۴٫۴	۱۳۶ / ۱۴۹٫۴	۱۱۰ / ۱۲۵٫۲	۴۰۰
$O_{i.}$	۳۳۶	۳۵۱	۳۳۱	۱۰۰۰

$$\chi^2_0 = \frac{(۱۸۲-۲۰۱,۴)^۲}{۲۰۱,۴} + \cdots = ۸,۰۴۸$$

$$\chi^2_{۲,۰۰۵} = ۵,۹۹ \qquad \chi^2_0 > \chi^2_{۲,۰۰۵} \Rightarrow RH_0$$

استقلال برقرار نمی‌شود

- کاربرد آزمون χ^2 در مقایسه دو جامعه با هم

فرض کنید متغیر تصادفی X توصیف کننده جامعه اول با تابع توزیع $F(x)$ و Y متغیر تصادفی توصیف کننده جامعه دوم و با توزیع $G(y)$ است می‌خواهیم این $H_0: F(z) = G(z)$ را در مقابل $H_1: F(z) \neq G(z)$ بررسی کنیم. آزمایک تجمیع‌شده‌ای جامعه اول به k گروه $C_1, \ldots, C_k$ افراز شده و همچنین برهمین اساس جامعه دوم به k گروه $D_1, \ldots, D_k$ افراز گردد تعریف می‌کنیم

$$\left.\begin{aligned} p_i &= P(X \in C_i) \\ q_{۰i} &= P(Y \in D_i) \end{aligned}\right\} \quad H_0: p_i = q_{۰i}$$

اگر $X_1 \ldots X_n$ نمونه‌ای از جامعه اول و متقلّ از نمونه $Y_1 \ldots Y_m$ از جامعه دوم در نظر گیریم و همین تعریف کنیم:

Subject:
Date

فراوانی مورد انتظار: e_i — فراوانی مشاهده شده از جامعه اول: O_i

از جامعه اول $n p_i$ — (تعداد x_i واقع در گروه C_i)

فراوانی مورد انتظار از: f_i — فراوانی مشاهده شده جامعه دوم: M_i

از جامعه دوم $m q_i$ — تعداد y_i واقع در گروه D_i

$$V_1 = \sum \frac{(O_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$$

$$V_2 = \sum \frac{(M_i - f_i)^2}{f_i} \sim \chi^2_{k-1}$$

که چون نمونه‌ها مستقلند طبق خواص کای دو:

$$V = V_1 + V_2 \sim \chi^2_{2(k-1)}$$

در اینصورت برای H_0 به ازای مقادیر بزرگ V رد می‌شود.

نکته: معمولاً برای برآورد p_i و q_i از روابط زیر استفاده می‌کنند

$$Z_i = \frac{O_i + M_i}{n + m} \qquad i = 1, \ldots, k-1 \qquad \sum_{i=1}^{k} Z_i = 1 \text{ (بجای } k-1\text{)}$$

در اینصورت $k-1$ از رده‌ها آزاد

مثال: دو شهر ... گواهی می‌شود

از ۱۰۰۰ دانش‌آموز که به تصادف انتخاب شده‌اند ۵۰۰ نفر در دوره ابتدایی، ۴۰۰ نفر در دوره راهنمایی و ۱۰۰ نفر در دوره دبیرستان می‌باشند در شهری دیگر از بین ۱۵۰۰ دانش‌آموز ۵۰۰ ابتدایی، ۶۰۰ راهنمایی و ۴۰۰ دبیرستانی می‌باشند

Subject:
Date

ما میزان اثر آموزش هستیم که دو شهر از نظر ترکیب دانش آموزی هم توزیع می‌باشند

	ابتدایی	راهنمایی	دبیرستان	
O_i	۵۰۰	۴۰۰	۱۰۰	۱۰۰۰
e_i	۰/۴×۱۰۰۰=۴۰۰	۰/۴×۱۰۰۰=۴۰۰	۰/۲×۱۰۰۰=۲۰۰	۱۰۰۰
m_i	۵۰۰	۶۰۰	۴۰۰	۱۵۰۰
f_i	۰/۴×۱۵۰۰=۶۰۰	۹۰۰	۳۰۰	۱۵۰۰

$$Z_1 = \frac{۵۰۰+۵۰۰}{۲۵۰۰} = ۰/۴ \qquad Z_2 = \frac{۴۰۰+۶۰۰}{۲۵۰۰} = ۰/۴ \rightarrow Z_3 = ۰/۲$$

$$V_1 = \frac{(۵۰۰-۴۰۰)^2}{۴۰۰} + \frac{(۴۰۰-۴۰۰)^2}{۴۰۰} + \frac{(۱۰۰-۲۰۰)^2}{۲۰۰}$$

$$V_2 = \frac{(۵۰۰-۶۰۰)^2}{۶۰۰} + \frac{(۶۰۰-۶۰۰)^2}{۶۰۰} + \frac{(۴۰۰-۳۰۰)^2}{۳۰۰}$$

$$V = V_1 + V_2 = ۱۲۵ \qquad \chi^2_{(K-1)(r-1)\,0.05} = ۹/۲$$

۱۲۵ > ۹/۲ → H_0 رد می‌شود

Subject:
Date

ضرایب همبستگی

- ضریب همبستگی پیرسون

اگر $f(x,y)$ تابع احتمال مشترک X و Y باشد می‌دانیم

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$

ضریب همبستگی خطی پیرسون $\rightarrow$ $$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

که $-1 \leq \rho \leq 1$

و از روی نمونه برآورد می‌شود: $$\hat{\rho}(X,Y) = R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

$$S_{xy} = \sum (x_i-\bar{x})(y_i-\bar{y}) \qquad S_{xx} = \sum (x_i-\bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum (y_i-\bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

که $-1 \leq R \leq 1$ و دقت کنید، ضریب همبستگی خطی پیرسون که ارتباط خطی بین دو متغیر X و Y را اندازه‌گیری می‌کند مقداری برای متغیرهایی است که کمی هستند و برای متغیرهای کیفی استفاده نمی‌شود.

در صفحه بعد ضریب همبستگی اسپیرمن معرفی می‌شود که برای متغیرهای رتبه‌ای دارند استفاده می‌گردد.

ضریب همبستگی رتبه‌ای اسپیرمن :

وقتی اندازه‌گیری‌ها رتبه‌ای است ضریب همبستگی اسپیرمن استفاده می‌شود در این حالت

نمونه (X_1, Y_1) و ... (X_n, Y_n) را بصورت (R_1, S_1) ... (R_n, S_n)

در نظر می‌گیریم که R_i رتبه X_i ام از مرتب شدن آن‌ها است و S_i رتبه Y_i ام از مرتب شدن آن‌ها است.

$$X_{(1)} < X_{(2)} < \dots < X_{(m)}$$
$$Y_{(1)} < Y_{(2)} < \dots < Y_{(n)}$$

که اگر همان روش محاسبه ضریب همبستگی پیرسن را به کار بریم برای رتبه‌ها، ضریب همبستگی رتبه‌ای اسپیرمن به دست می‌آید.

$$R_S = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}}$$

که در آن ۲ فرمول دیگر نیز نوشته می‌شود داریم:

$$\bar{R} = \frac{1}{n} \underbrace{\sum_{i=1}^{n} R_i}_{\frac{n(n+1)}{2}} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\bar{S} = \frac{n+1}{2} \qquad \sum (R_i - \bar{R})^2 = \sum R_i^2 - n\bar{R}^2$$

$$= \frac{n(n+1)(2n+1)}{6} - n\frac{(n+1)^2}{4} = \frac{n(n+1)}{2}\left(\frac{2n+1}{3} - \frac{n+1}{2}\right)$$

Subject:
Date

$$\sum_i (R_i - \bar{R})^2 = \sum_i (S_i - \bar{S})^2 = \frac{n(n^2-1)}{12}$$ که

$$R_S = \frac{\sum R_i S_i - n\bar{R}\bar{S}}{\frac{n(n^2-1)}{12}} = \frac{\sum R_i S_i - \frac{n}{4}(n+1)^2}{\frac{n(n^2-1)}{12}}$$

$$= \frac{12\sum R_i S_i - 3n(n+1)^2}{n(n^2-1)} = \frac{12\sum R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

یا اگر بردارد هم $D_i = R_i - S_i$ ، $D_i^2 = R_i^2 + S_i^2 - 2R_i S_i$

$$\sum D_i^2 = \frac{n(n+1)(2n+1)}{6} - 2\sum R_i S_i$$ که

$$12\sum R_i S_i = n(n+1)(2n+1) - 6\sum D_i^2$$

$$R_S = \frac{2n(n+1)(2n+1) - 6\sum D_i^2}{n(n^2-1)} - 3\frac{n+1}{n-1}$$

$$= 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

$$\frac{2n^3+2n^2+n}{2n(n+1)(2n+1) - 6\sum D_i^2 - 3n(n+1)^2}{n(n^2-1)}$$

$$= \frac{4n^3+2n^2+2n^2+n-3n^3-6n^2-3n}{n(n^2-1)}$$

$$= \frac{n^3-n-6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

۳- آزمون همبستگی: با فرض این که ρ ضریب همبستگی خطی بین دو متغیر X و Y است می‌توان برای کل آماره‌ای R و R_S فرض‌های زیر را آزمون کرد.

$$\begin{cases} H_0: \rho = 0 & \text{X و Y همبستگی ندارند} \\ H_1: \rho \neq 0 & \text{" همبستگی دارند} \end{cases}$$

Subject :
Date

وقتی از آماره R_s استفاده می‌کنیم توزیع R_s به توزیع جامعه بستگی ندارد و از رتبه‌ها استفاده می‌کند و آزاد توزیع است زیرا از آن مقادیر مختلف n جدول‌هایی وجود دارد که در ازای آن‌ها وقتی فرض H_0 می‌توان چندک‌های توزیع را محاسبه کرد.)

ثابت می‌شود $E(R_s)=0$ و $Var(R_s)=\frac{1}{n-1}$ بنابراین

$$Z=\frac{R_s-E(R_s)}{\sqrt{Var(R_s)}}\sim N(0,1)$$

$$\sqrt{Var(R_s)}=\frac{1}{R\sqrt{n-1}}$$

نکته: رتبه‌بندی داده‌ای یکسان

می‌دانیم وقتی با متغیرهای تصادفی پیوسته سروکار داریم احتمال برابر بودن آن‌ها صفر است ولی در عمل به دلیل دقیق نبودن وسایل اندازه‌گیری و یا چشم‌پوشی از تفاوت‌های جزئی برخی مقادیر ممکن است که بعضی از یافته‌ها برابر ظاهر شوند. در این خصوص دو روش پیشنهاد شده است یک روش این است که داده تکرار شده را حذف و از حجم نمونه کاهیم و روش دیگر پیش‌نهاد رتبه متوسط برای چنین داده‌هایی است که روش مفیدتری است.

مثال صفحه ۱۴۷

Subject:
Date

ضریب همبستگی کندال:

فرض کنید دو متغیر X و Y در جامعه‌ای مورد بررسی است و فرض کنید این جامعه دارای n عضو است. زوج (X,Y) اندازه دو متغیر را برای یک فرد نشان می‌دهند. فرض کنید فرد i ام و j ام انتخاب شوند پس اندازه دو متغیر X و Y به ترتیب این دو فرد (X_i, Y_i) و (X_j, Y_j) است. اگر رابطه $(X_j - X_i)(Y_j - Y_i) > 0$ برقرار باشد می‌گوییم زوج‌ها هماهنگ هستند و اگر $(X_j - X_i)(Y_j - Y_i) < 0$ می‌گوییم زوج‌ها هماهنگ هستند. چون (X_i, Y_i) و (X_j, Y_j) از جهتی تصادفی هستند در عبارت فوق دو پیشامد هستند. اگر احتمال رخ دادن پیشامد اول را با π_c نشان داده و آن را احتمال هماهنگی بگوییم

$$\pi_c = P((X_j - X_i)(Y_j - Y_i) > 0)$$

و احتمال رخ دادن پیشامد دوم را با π_d:

$$\pi_d = P((X_j - X_i)(Y_j - Y_i) < 0)$$

روشن است که

$$\pi_d + \pi_c = 1$$

با استفاده از π_c و π_d ضریب همبستگی کندال توسط کندال به صورت زیر

Subject:
Date

تعریف می شود:

$$\tau_b = \pi_c - \pi_d = 2\pi_c - 1$$

if $\pi_c = 1 \Rightarrow \tau = 1$ دو متغیر در یک جهت تغییر می کنند

if $\pi_c = 0 \Rightarrow \tau = -1$ دو متغیر در دو جهت مخالف تغییر می کنند

if $\pi_c = 1/2 \Rightarrow \tau = 0$ دو متغیر نامستقل اند

$$-1 \le \tau \le 1$$ به طور کلی

- برآورد τ : ابتدا x_i ها را به ترتیب صعودی در نظر می گیریم یعنی $x_j - x_i > 0$

و متناظر علامت $y_j - y_i$ را به دست می آوریم، قراردهیم A_{ij} را:

$$A_{ij} = \begin{cases} 1 & (x_j - x_i)(y_j - y_i) > 0 \\ 0 & (x_j - x_i)(y_j - y_i) = 0 \\ -1 & (x_j - x_i)(y_j - y_i) < 0 \end{cases}$$

P: تعداد علامت‌های مثبت A_{ij}

N: تعداد علامت‌های منفی A_{ij}

و ثابت می شود

$$T = \hat{\tau} = \frac{P - N}{P + N}$$

مثال. در یک امتحان شفاهی ۲ معلم a, b و ۸ دانشجوی $A B \dots H$

صحت جدول زیر برعهده گذاری می کنند ضریب τ را به دست آورید.

	A	B	C	D	E	F	G	H
a	۲	۱	۳	۵	۶	۸	۷	۴
b	۱	۲	۴	۵	۷	۶	۸	۳

P4PCO

Subject:
Date

	B	A	C	E	D	H	G	F
X	۱	۲	۳	۴	۵	۶	۷	۸
Y	۲	۱	۴	۷	۵	۳	۸	۶

A_{ij} — ۰ — ۰ — صفر ۱ ۰ ۱ ۳ ۱ ۱ → $N = V$

(۱-۲) (۴-۳) (۵-۳) (۶-۳) (۷-۵) (۸-۷)

مثبت ۶ ۶ ۴ ۱ ۲ ۲ → $P = ۲۱$

(۲-۱)(۱-۲) (۳-۱)(۴-۲)

$$\hat{\tau} = \frac{۲۱-۷}{۲۱+۷} = ۱/۲$$

معمولاً به ازای مقادیر مختلف n و p، برای τ حدودی تعیین شده است اما معمولاً

از توزیع مجانبی T استفاده می‌شود ثابت می‌شود

$$E(T) = ۰ \qquad Var(T) = \frac{۲(۲n+۵)}{۹n(n+۱)}$$

$$Z = \frac{T - E(T)}{\sqrt{Var(T)}} \sim N(۰,۱)$$

Subject:
Date

آنالیز واریانس یکطرفه:

فرض کنید $X_1 \dots X_k$ متغیرهای تصادفی مستقل هستند که هرکدام توصیف کننده جامعه‌ای با توابع توزیع $F_1(x) \dots F_k(x)$ هستند می‌خواهیم فرض

$$\begin{cases} H_0: F_1(x) = \dots = F_k(x) \quad \forall x \\ H_1: \text{حداقل ۲ تابع توزیع یکی نیستند} \end{cases}$$

را آزمون نیم، اگر $\mu_1 \dots \mu_k$ میانگین‌های توزیع‌ها وجود داشته باشد

$$H_0: \mu_1 = \dots = \mu_k$$

- آزمون کروسکال - والیس

ابتدا از هر یک از توزیع‌ها نمونه‌ای تصادفی اختیار می‌کنیم

X_1	X_2	$\dots$	X_k
x_{11}	x_{21}	$\dots$	x_{k1}
$\vdots$			
x_{1n_1}	x_{2n_2}	$\dots$	x_{kn_k}

سپس با توجه به فرض H_0 که می‌گوید توزیع‌ها یکسان هستند همه k نمونه را درهم ادغام می‌کنیم تا نمونه‌ای به حجم $n = \sum n_i$ داشته باشیم این نمونه را مرتب کرده و رتبه‌ها را در جدول مشابه زیر جایگزین x_{ij} ها می‌کنیم

Subject:
Date

	۱	۲	...	i	...	k
	R_{11}	R_{21}		R_{i1}		R_{k1}
	R_{12}	R_{22}		R_{i2}		R_{k2}
	⋮	⋮		⋮		⋮
	R_{1n_1}	R_{2n_2}		R_{in_i}		R_{kn_k}
جمع	R_1	R_2	...	R_i	...	R_k
متوسط	$\bar{R}_1$	$\bar{R}_2$	...	$\bar{R}_i$	...	$\bar{R}_k$

$$\bar{R} = \frac{1}{n}\sum R_{ij} = \frac{1}{n}\cdot\frac{n(n+1)}{2} = \frac{n+1}{2}$$

- آماره آزمون: مشابه آنچه در تحلیل واریانس یک طرفه دیدیم می‌توان آماره زیر را در نظر گرفت

$$V = \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2$$

و حتی فرض H_0 درست است انتظار داریم V کوچک باشد هرچه V بزرگتر شد H_0 را رد می‌کنیم.

چون V توزیع خاصی ندارد کروسکال و والس آماره زیر را تعریف کرد

$$H = \frac{12V}{n(n+1)}$$

که تحت H_0 به طور تقریبی دارای χ^2_{k-1} است پس اگر $H > \chi^2_{k-1,\alpha} \Rightarrow RH_0$

پس از جایگذاری V داریم:

$$H = \frac{12}{n(n+1)}\sum_i \frac{R_i^2}{n_i} - 3(n+1)$$

Subject:
Date

مثال: در جدول زیر نمرات آزمون ورودی دانشجویان کدام دوره دکترای سه دانشگاه A و B و C داده شده است می خواهیم با $\alpha = 0.05$ بیازماییم که دانشجویان هر سه دانشگاه به طور متوسط یکسان هستند

A	۱۹	۱۱٫۷	۱۷٫۸	۱۴٫۸	۱۳٫۹				
B	۱۸٫۲	۱۴٫۸	۱۳٫۱	۱۲٫۶	۱۵٫۲	۱۲٫۸			
C	۱۳٫۴	۱۴٫۱	۱۲٫۳	۱۲٫۳	۱۴٫۷	۱۳٫۹	۱۵٫۸	۱۴٫۳	

A	۱۱٫۷	—	—	—	—	—	—	—	—	۱۳٫۹	—	—	—	—	۱۴٫۸
B	—	—	—	۱۲٫۶	۱۲٫۸	۱۳٫۱	—	—	—	—	—	—	—	۱۴٫۸	—
C	—	۱۲٫۳	۱۲٫۳	—	—	—	۱۳٫۴	۱۳٫۸	۱۳٫۹	—	۱۴٫۱	۱۴٫۳	۱۴٫۷	—	—

A	—	۱۷٫۸	—	۱۹
B	۱۵٫۲	—	۱۸٫۲	—
C	—	—	—	—

A	۱	۱۰	۱۵	۱۷	۱۹
B	۴	۵	۶	۱۴	۱۶	۱۸
C	۲	۳	۷	۸	۹	۱۱	۱۲	۱۳

۹٫۱۰ → ۹٫۵

۱۴٫۱۵ → ۱۴٫۵

۲٫۳ → ۲٫۵

درنهایت

A	۱	۹٫۵	۱۴٫۵	۱۷	۱۹			
B	۴	۵	۶	۱۴٫۵	۱۶	۱۸		
C	۲٫۵	۲٫۵	۷	۸	۹٫۵	۱۱	۱۲	۱۳

→

PARCO

Subject:
Date

آزمون کلموگروف - اسمیرنوف

تابع توزیع تجربی: فرض کنید $X_1, \dots, X_n$ نمونه‌ای تصادفی از توزیع $F_X(x)$ باشد برازای هر x داده شده تابع توزیع نمونه‌ای یا تجربی به صورت زیر تعریف می‌شود

$$F_n(x) = \frac{\#(X_i \le x)}{n} = \frac{B}{n}$$

که در آن متغیر تصادفی B دارای توزیع دو جمله‌ای با پارامترهای n و $F(x)$ است بنابراین داریم:

$$E(B) = nF(x) \qquad Var(B) = nF(x)[1-F(x)]$$

$$E(F_n(x)) = F(x) \qquad Var(F_n(x)) = \frac{F(x)[1-F(x)]}{n}$$

یعنی $F_n(x)$ یک برآورد نااریب برای $F(x)$ است و علاوه بر این سازگار نیز می‌باشد پس $F_n(x)$ تقریبی برای $F(x)$ است.

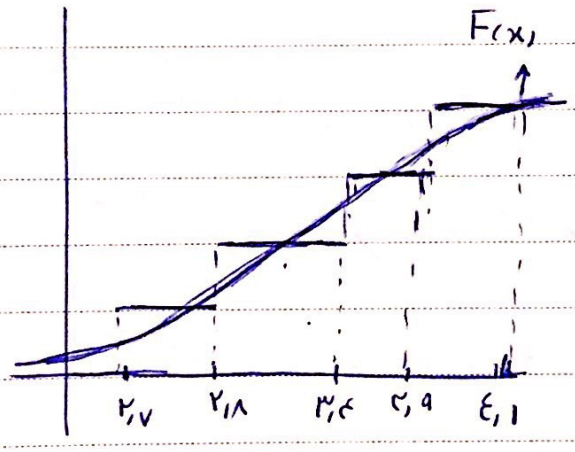
مثلاً اگر در یک نمونه تصادفی داشته باشیم

$$x_1 = 3.4 \quad x_2 = 2.8 \quad x_3 = 3.9 \quad x_4 = 2.7 \quad x_5 = 4.1$$

آنگاه داریم

$$F_n(x)=\begin{cases}0 & x<2.7\\ \frac{1}{5} & 2.7\le x<2.8\\ \frac{2}{5} & 2.8\le x<3.2\\ \frac{3}{5} & 3.4\le x<3.9\\ \frac{4}{5} & 3.9\le x<4.1\\ 1 & x\ge 4.1\end{cases}\longrightarrow \hat{F}(x)=\begin{cases}\\ F_n(x)\\ \\ \\ \\ \\ \end{cases}$$

به طور کلی اگر $y_1 \dots y_n$ یافته‌ارِ آماره‌ای ترتیبی نمونه $X_1 \dots X_n$ باشند داریم



$$F_n(x)=\begin{cases}0 & x<y_1\\ 1/n & y_1\le x<y_2\\ \vdots & \\ i/n & y_i\le x<y_{(i+1)}\\ \vdots & \\ 1 & x\ge y_n\end{cases}$$

PAPCO

Subject:
Date

آزمون برازندگی کلموگروف - اسمیرنف K-S

Kolmogorov - Smirnor statistic

فرض کنید $X_1 \dots X_n$ نمونه‌ای تصادفی از توزیع پیوسته $F(x)$ است همچنین

فرض کنید $F_0(x)$ توزیع شناخته شده معلومی است می‌خواهیم فرضیه

$$\begin{cases} H_0: F_n(x) = F_0(x) & \forall x \\ H_1: F_n(x) \neq F_0(x) & \exists x \end{cases}$$

آزمون کنیم. قبلاً دیدیم که برای کای χ^2 پیرسن این فرضیه قابل آزمون است

آزمون دیگر، آزمون برازندگی کلموگروف اسمیرنف است.

دیدیم که با ازای هر x داده شده $F_n(x)$ برآوردی نااریب و سازگار برای $F(x)$

است آماره زیر به عنوان آماره K-S عرضه شده است

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

که ثابت می‌شود تحت فرض H_0 توزیع آن به توزیع جامعه بستگی نداشته

و آزاد توزیع است و جداولی برای آن تهیه شده است.

مثال: فرض کنید نمونه $x_1 = 0.98$ $x_2 = 0.17$ $x_3 = 0.48$
$x_4 = 0.1$
از توزیع $F(x)$ در اختیار است اگر $F_0(x)$ توزیع یکنواخت در فاصله $(0,1)$ باشد

x	$F_4(x)$	$F_0(x)$	$F_n(x) - F(x)$
0.1	1/4	0.1	0.15
0.45	2/4	0.45	0.15
0.47	3/4	0.47	0
0.98	1	0.98	0.02

$D_4 = \sup_x |F_n(x) - F_0(x)|$
$\Rightarrow D_4 = 0.15$

جدول در توزیع یکنواخت $F_0(x) = x$

① اگر $\begin{cases} H_0: F_n(x) = F_0(x) \\ H_1: F_n(x) \neq F_0(x) \end{cases}$ فرض H_0 به ازای مقادیر بزرگ D_n رد می‌شود

یعنی $D_n \geq d_\alpha \Rightarrow RH_0$

② اگر $\begin{cases} H_0: F_n(x) = F_0(x) \\ H_1: F_n(x) > F_0(x) \end{cases}$ ، آماره زیر تعریف می‌شود

$D_n^+ = \sup (F_n(x) - F_0(x))$ if $D_n^+ \geq d_\alpha \Rightarrow RH_0$

اثبات ص ۲۰۸ کتاب $= \max(0, \max(\frac{i}{n} - F(x_{(i)}))$

③ اگر $\begin{cases} H_0: F_n(x) = F_0(x) \\ H_1: F_n(x) < F_0(x) \end{cases}$

$D_n^- = \sup_x (F_0(x) - F_n(x))$: آماره

اثبات ص ۲۰۹ کتاب $= \max(0, \max(F(x_{(i)}) - \frac{i-1}{n}))$

if $D_n^- \geq d_\alpha \Rightarrow RH_0$

نکته ضمناً داریم $D_n = \max(D_n^+, D_n^-)$

نکته ۲:

$P(D_n \geq z) \approx p D_n^+ \Rightarrow$

- توزیع مجانبی D_n :

با فرض این که $X_1 \dots X_n$ نمونه‌ای تصادفی از توزیع $F(x)$ است ثابت می‌شود که توزیع مجانبی $x_n = 4n D_n^{+2}$ که χ^2_2 است و داریم

$$P(D_n \geq d) \approx P(D_n^+ \geq d) + P(D_n^- \geq d)$$

$$\simeq 2P(\chi^2_2 \geq 4nd^2)$$

مثال: نمونه تصادفی ۱۰ تایی داریم

۰/۱۴۸۰ ۰/۱۴۲۹ ۰/۱۵۸۱ ۰/۱۷۱۰ ۰/۲۴۷۷ ۰/۲۰۰۳ ۰/۱۵۰۳ ۰/۱۴۲۱
۰/۱۵۵۴ ۰/۱۳۸۲

مثال: $\alpha = 0/05$ این فرض را بیازمایید که این نمونه از توزیع یکنواخت روی فاصله $(0,1)$ ساخته شده است

x	$F_0(x)$	$F_{10}(x) = i/n$	$F_{10}(x) - F_0(x) = D^+$	$D^- = F_0(x) - \frac{i-1}{n}$
۰/۱۲۰۳	۰/۱۲۰۳	۱/۱۰	-۰/۱۰۳	۰/۱۲۰۳
۰/۱۳۲۹	۰/۱۳۲۹	۲/۱۰	-۰/۱۲۹	max (۰/۱۲۹)
۰/۱۳۸۲	۰/۱۳۸۲	۳/۱۰	-۰/۰۸۲	۰/۰۸۲
		۴/۱۰	-۰/۰۷۷	
۰/۱۷۱۰	۰/۱۷۱۰	۱	(۰/۱۲۹) max	-۰/۱۹۰

$$\sup |F_n(x) - F_0(x)| = D^+_{10} = 0/129$$

$$\Rightarrow D_n = 0/129$$

$$\sup (F_0(x) - F_n(x)) = D^-_{10} = 0/129$$

Subject:
Date

H_0 رد می‌شود اگر $D_{10} \geq d_{0.05}$ از جدول

$d_{10,0.05} = 0.409$

$0.129 < 0.409 \Rightarrow$ H_0 قبول

با استفاده از تقریب χ^2:

$$p(D_{10} \geq d) = 2p(\chi^2_2 \geq 40d^2) = 0.05 \Rightarrow 40d^2 = 5.99 \Rightarrow d = \sqrt{\frac{5.99}{40}}$$

$$p\text{-value} = p(D_{10} \geq 0.129) \approx 2p(\chi^2_2 \geq 40(0.1295)) = 2 \times 0.1 = 0.3$$

(برای آزمون‌های نظیر χ^2 که مثبت هستند مقدار p-value در دم راست می‌شود)

مثال: آیا فرض $\begin{cases} H_0: F(x) = F_0(x) \\ H_1: F(x) \neq F_0(x) \end{cases}$ $f_0(x) = \begin{cases} 3/4 & 0 < x < 1/2 \\ 1/4 & 1/2 < x < 1 \\ 0 & \text{o.w} \end{cases}$

برای داده‌های زیر برقرار است

$x =$	0.01	0.06	0.08	0.09	0.11	0.14	0.22	0.24	0.29
	0.3	0.38	0.4	0.41	0.42	0.46	0.47	0.49	0.71
	0.78	0.79	0.82	0.88					

حل:

$$F_0(x) = \int_{-\infty}^{x} f_0(t)dt = \int_0^x 3/4\, dt = \frac{3}{4}x$$

$$F_0(x) = \int_0^{1/2} 3/4\, dt + \int_{1/2}^{x} 1/4\, dt = \frac{x+1}{2}$$

$$F_0(x) = \begin{cases} 3/4\, x & 0 \leq x < 1/2 \\ x + 1/4 & 1/2 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

۴۵

Subject:

Date

x	$F_0(x)$	$F_{25}(x) = i/n$	$\lvert i/25 - F_0(x)\rvert$
۰/۰۱	۰/۰۱۵	۱/۲۵	۰/۰۲۵
۰/۰۶	۰/۰۹	۲/۲۵	۰/۰۱
۰/۰۸	۰/۱۲	۳/۲۵	۰
۰/۰۹	۰/۱۳۵	۴/۲۵	۰/۰۲۵
۰/۱۱	۰/۱۹۵	۵/۲۵	۰/۰۵۵
۰/۱۴	۰/۲۴	۶/۲۵	۰
۰/۲۲	۰/۳۳	۷/۲۵	۰/۰۵
۰/۲۳	۰/۳۴۵	۸/۲۵	۰/۰۲۵
۰/۲۹	۰/۴۳۵	۹/۲۵	۰/۰۷۵
۰/۳	۰/۴۵	۱۰/۲۵	۰/۰۵
۰/۳۵	۰/۵۲۵	۱۱/۲۵	۰/۰۷۵
۰/۳۸	۰/۵۷	۱۲/۲۵	۰/۰۹
۰/۴	۰/۶	۱۳/۲۵	۰/۰۸
۰/۴۱	۰/۶۱۵	۱۴/۲۵	۰/۰۵۵
۰/۴۲	۰/۶۳	۱۵/۲۵	۰/۰۳
۰/۴۸	۰/۷۲	۱۶/۲۵	۰/۰۸
۰/۴۷	۰/۷۰۵	۱۷/۲۵	۰/۰۶
۰/۴۹	۰/۷۳۵	۱۸/۲۵	۰/۱۱
۰/۵۱	۰/۷۶۵	۱۹/۲۵	۰/۰۹۵
۰/۵۲	۰/۷۸	۲۰/۲۵	۰/۰۵۲
۰/۵۸	۰/۸۱۰	۲۱/۲۵	۰/۰۶
۰/۵۹	۰/۸۶۵	۲۲/۲۵	۰/۰۱۵
۰/۶۲	۰/۹۱	۲۳/۲۵	۰/۰۱
۰/۶۸	۰/۹۴	۲۴/۲۵	۰/۰۴
۰/۶۹	۰/۹۵	۲۵/۲۵	۰/۰۶

$D_{25} = 0/11$

جدول ۲۸ صفحه

$d_{25,0.05} = 0/24$

$0/11 < 0/24 \Rightarrow AH_0$

$2p(\chi^2_2 \geq 100 \times (0/11)^2) = 2p(\chi^2_2 \geq 1/21) \geq 0/05$

$\Rightarrow AH_0$

Subject:
Date

مقایسه دو جامعه با استفاده از K-S

فرض کنید جامعه اول با متغیر تصادفی X توصیف می‌شود که تابع توزیع پیوسته $F(x)$ دارد و همچنین فرض کنید جامعه دوم متغیر تصادفی Y را توصیف می‌کند، توزیع پیوسته $G(y)$ دارد. هدف بررسی فرضیه

$$\begin{cases} H_0: F(z) = G(z) \\ H_1: F(z) \neq G(z) \end{cases}$$

است.

برای این منظور نمونه تصادفی $X_1 \dots X_n$ را از جامعه اول می‌گیریم و داریم $F_n(x)$ برآورد ناریب $F(x)$ است. همچنین مستقل نمونه تصادفی $Y_1 \dots Y_m$ را از جامعه دوم اختیار می‌کنیم که $G_m(y)$ برآورد ناریب $G(y)$ است. آنگاه با استفاده از مکاره K-S این دو برآورد هم مقایسه می‌شوند

$$D_{n,m} = \sup_z |F_n(z) - G_m(z)|$$

که از آن مقادیر بزرگ $D_{n,m}$ فرض H_0 رد می‌شود

$$D_{n,m} \geq d_\alpha \iff RH_0$$

مثال. در یک آزمون ریاضی نمره‌ی ۶ دختر و پسر عبارتند از

پسرها: ۱۰/۳ ۱۱/۲ ۱۱/۵ ۱۱/۹ ۱۲/۸

دخترها: ۱۰/۴ ۱۱/۸ ۱۲/۰ ۱۲/۶ ۱۲/۸ ۱۲/۹

Subject :
Date

آیا با سطح ۰.۰۵ این نمونه‌ها هم‌توزیع‌اند

x	y	$F_{\Delta}(x)$	$G_{4}(y)$	$\lvert F_{\Delta}(x) - G_{4}(y) \rvert$
۱۰٫۲	—	۱/۵ $x \le 10.2$	۰	۰/۲
—	۱۰.۱۶	۱/۵ $x \le 10.16$	۱/۴	۰٫۰۳۳
۱۱٫۲	—	۲/۵ $x \le 11.2$	۱/۴	۰٫۲۵۲
۱۱٫۵	—	۳/۵ $x \le 11.5$	۱/۴	۰٫۴۵۲
—	۱۱٫۸	۳/۵ $x \le 11.8$	۲/۴	۰٫۲۴۴
۱۱٫۹	—	۴/۵ $x \le 11.9$	۲/۴	۰٫۴۴۴ → max
—	۱۲٫۵	۴/۵ $x \le 12.5$	۳/۴	۰/۳
—	۱۲٫۴	۴/۵ $x \le 12.4$	۴/۴	۰٫۱۵۲
۱۲٫۸	—	۱	۴/۴	۰٫۵۵۲
—	۱۳٫۸	۱	۵/۴	۰٫۱۴۴
—	۱۳٫۹	۱	۱	۰

$D_{5,4} = 0.444 \not> d_{5,4,0.05} = 0.4444 \Rightarrow AH_0$

PARCO