۱۹/۱۰/۹۲

(۱) مروری بر آماره‌های ترتیبی، فاصله اطمینان برای چندک‌ها، فاصله تحمل

(۲) آزمون‌های تک نمونه‌ای: آزمون علامت، رتبه‌دار ویلکاکسون و ...

(۳) آزمون‌های دو نمونه‌ای: آزمون‌های همبسته ویلکاکسون، من-ویتنی ...

- آمار ناپارامتری، دکتر جواد بهبودیان، دانشگاه شیراز ۱۳۸۳

- تحلیل آماری ناپارامتریک با کاربرد آزمون‌های آماری ناپارامتریک در پژوهش‌های علوم مدیریت، عباس طلوعی و محمد سعید صفاکیش، دانشگاه آزاد واحد علوم تحقیقات ۱۳۸۹

- Nonparametrics: statistical methods Based on Ranks, Lehmann(2006)

- Nonparametric statistical methods using R
John kloke, Joseph w. McKean, 2015

(۴) آزمون‌های k نمونه‌ای: آزمون اکروسکال والیس، فریدمن و ...

(۵) معیارهای همبستگی: ضریب همبستگی پیرسن، اسپیرمن، کندال و ...

تاریخچه آمار ناپارامتری:

تاریخ آمار ناپارامتری به اوایل قرن هجدهم میلادی برمی‌گردد. در سال ۱۷۰۰ میلادی مقاله‌ای منتشر شد و طی آن بر اساس آمار نوزادان شهر لندن در فاصله ۱۶۲۰ تا ۱۷۱۰ میلادی ادعا شد که مشیت الهی بر این است که تعداد نوزادان پسر بیشتر از نوزادان دختر باشد. در حقیقت این ادعا یک آزمون ناپارامتری معروف می‌باشد که امروزه به آن آزمون علامت (Sing test) شهرت دارد. آمار ناپارامتری (non parametric) بیش از دو قرن ناشناخته ماند و استفاده از روش‌های توسط ویلکاکسون (۱۹۴۵) و من و ویتنی (Mann and whitney) شهرت آن آغاز شد. البته قبل از این دو مقاله اصطلاح ناپارامتری برای اولین بار در رساله دکترای یک

Date

آماردان بنام (ولفوتیز) wolfwitz (1942) به کار رفته است.

اغلب کارهای آماری و آزمون های آماری با فرض نرمال بودن توزیع صورت می پذیرد اول آزمون در چنین حالتی زمانی معتبر است که فرض نرمال بودن توزیع متغیری مورد اندازه گیری برقرار شود.

برای بررسی نرمال بودن ① هیستوگرام ② QQplot P-Pplot ③ آزمون نکویی برازش

④ آزمون های مختلف

درصورت رد فرض نرمال بودن علاوه بر استفاده از آزمونها ناپارامتری بیشتر می شود

در روش های ناپارامتری بدون در نظر گرفتن فرض خاصی برای توزیع جامعه استنباط آماری می پردازیم.

بنابراین یکی از محاسن اصلی روش های ناپارامتری این است که مستلزم فرض خاصی درباره شکل توزیع جامعه نیستیم و می توان گفت یکی از معایب روش های ناپارامتری این است که منقا از بسیاری از اطلاعات استفاده می شود و تمام اطلاعات استفاده می کنیم. در spss تمام آزمون های ناپارامتری در مسیر زیر است

Analyze / Non parametric Tests / Chi-square
Binomial
Runs
1-sample k-s...
2-Indpendent samples...
k Independent samples
2 Related samples
k Related samples

PAPCO

Subject:
Date

در منی تب هم به همین صورت stat / nonparametric / mann-whitney
1-sample wilcoxon
kruskal-wallis
⋮

در R: توابع مختلف وجود دارد:

wilcox.test(x, conf.int=TRUE)
prop.test
chisq.test
kruskal.test
friedman.test

به روش های ناپارامتری روش هایی آزاد توزیع (distribution free) نیز می گویند
در برخی مقالات پیامدها به صورت رتبه ای واسطی می شوند در این گونه مسائل نیز
علاقه مند به استفاده از روش های ناپارامتری هستیم. برای متغیرهایی که این روش کاربردی
استفاده می شوند، پیش فرض اساسی رعایت توزیع داده ها، هرچند بی تأثیر نباشد.
تذکر: برخی از کارها، چون حجم نمونه در مقاله آن ها بیشتر از عدد ۳۰ است
به اشتباه و با استناد به قضیه حد مرکزی آزمون های پارامتری را به جای ناپارامتری برای
غیرنرمال نیز به کار می برند. درحالی که در قضیه حد مرکزی بیان می شود که در صورت
داشتن نمونه بزرگ، توزیع صابتین به نرمال میل می کند که در این مورد، نویسنده

Date

نکته: اول اینکه عدد ۳۰ همواره کافی بودن نمونه‌ها برای برقراری شرایط قضیه حد مرکزی ممکن است برقرار نباشد و در برخی مسائل عدد بزرگتر از ۳۰ برای برقراری شرایط قضیه لازم است

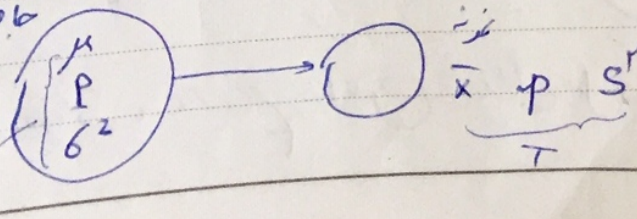
دوم این که شرایط استفاده از این قضیه برای آزمون‌هایی است که شاخص آزمون آن‌ها روی میانگین جامعه تعریف شده باشد و برای آزمون‌هایی مثل تحلیل واریانس با همبستگی کاربرد ندارد.

مثالی از آماره پارامتری و ناپارامتری:

فرض کنید وزن نوزادان متغیر نرمال باشد و هدف برآورد میانه این متغیر است یعنی پارامتری که وزن پنجاه درصد از نوزادان کمتر از آن هستند را برآورد کنیم. اگر تصور کنیم جامعه نرمال است می‌دانیم میانگین و میانه برابرند و در نتیجه میانگین نمونه همان برآورد میانه است.

ولی فرض کنید توزیع متقارن نباشد و میانه را می‌خواهیم برآورد کنیم. ابتدا نمونه را به صورت مرتب شده از کوچک تا بزرگ مرتب کنیم و در صورتی که اندازه نمونه فرد باشد عدد میانی این رشته و در صورتی که اندازه نمونه زوج باشد نصف مجموع دو عدد میانی خواهد بود.

آماره و پارامتر: هر شاخصه‌ای که توصیف کننده جامعه آماری مورد بررسی باشد پارامتری نامند و هر شاخصه‌ای که توصیف کننده نمونه باشد آماره نامیده می‌شود.



PAPCO

Subject:
Date

برآوردگر: فرض کنید θ پارامتر مجهول جامعه و T آماره‌ای مناسب است هر مقدار T را برآوردی برای θ می‌نامند و به صورت $\hat{\theta} = T$ نشان می‌دهند که T برآورد نقطه‌ای θ می‌نامیم و برآوردگر T را خوب می‌نامیم اگر

① از مقدار θ دور نباشد ② نااریب باشد ③ دارای کمترین واریانس باشد و

عبارت دیگر MSE کوچکی داشته باشد ④ سازگار باشد $\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$

پارامترهایی که در آمار ناپارامتری مورد استفاده قرار می‌گیرند اغلب همان پارامترهای هستند که در آمار پارامتری مورد بررسی قرار می‌گیرند ولی تنها پارامترهای نقطه میانه، چندک‌ها و نسبت و تابع توزیع مورد توجه است.

تابع توزیع: فرض کنید X متغیر تصادفی $(X: S \to S_X \subseteq R)$ باشد در اینصورت $F_X(x)$ نشان دهنده تابع توزیع X است با ضابطه:

$$F: R \to [0,1], \quad F_X(x) = P(X \le x)$$

در آمار ناپارامتری اغلب ضابطه تعریف $F_X(x)$ معین نیست و $F_X(x)$ که مقدار تابع در نقطه معین x است، صورت یک کمیت پارامتر مجهول حاصل است.

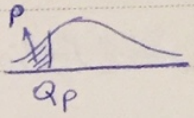
Subject:
Date

میانه: با فرض پیوسته بودن X میانه مقداری است که تابع توزیع در آن $\frac{1}{2}$ است

$$F_X(m) = \frac{1}{2} \quad ; \quad \int_{-\infty}^{m} f_X(x)\,dx = \frac{1}{2}$$

چندک p: Q_p چندک مرتبه p ام است، در صورت پیوسته بودن X،

$$F_X(Q_p) = \int_{-\infty}^{Q_p} f_X(x)\,dx = p$$



در نمونه، اگر تعداد داده فرد باشد پس از مرتب کردن داده‌ها، عدد وسط میانه است و اگر زوج باشد جمع دو عدد وسط میانه است.

$$X_1, \ldots, X_n \longrightarrow Y_1 = \min(X_1 \cdots X_n),\ Y_2 \cdots,\ Y_n = \max(X_1 \cdots X_n)$$

m = آماره ترتیبی وسط $= Y_{k+1}$ if $n = 2k+1$

m = نصف مجموع دو آماره ترتیبی وسط $= \frac{Y_k + Y_{k+1}}{2}$ if $n = 2k$ or $m = aY_k + (1-a)Y_{k+1}$ $0 < a < 1$

و برای محاسبه چندک‌ها از روی نمونه:

$$q_p = \begin{cases} Y_{[np]} & \\ \text{or} & \\ Y_{[np]+1} & \\ \text{or} & \\ aY_{[np]} + (1-a)Y_{[np]+1} & \end{cases}$$

بزرگترین عدد صحیح کمتر از np

تابع توزیع نمونه‌ای تجربی:

تابع توزیع تجربی با $F_n(x)$ نشان داده می‌شود و به صورت ~~$F_n(x) = \frac{1}{n}$~~

$$F_n(x) = \frac{\#(x_i \le x)}{n} \quad \text{تعداد} \rightarrow$$

تعریف می‌شود تابع توزیع تجربی تجمعی

مثال: فرض کنید از توزیع $F_x(x)$ نمونه‌ای ده‌تایی به صورت $x_1, x_2 \dots x_{10}$ در اختیار داریم

$x_1 = 1,2 \quad x_2 = 0,95 \quad x_3 = 1,15 \quad x_4 = 1,7$

$x_5 = 2,15 \quad x_6 = 2,18 \quad x_7 = 1,9 \quad x_8 = 2,2 \quad x_9 = 2,4 \quad x_{10} = 1,8$

برای کل این یافته‌ها $F_n(x)$ و $y_1 \dots y_{10}$ و m و $q_{0,25}$ را بیابید.

x	$F_n(x)$
0,95	1/10
1,15	2/10
1,2	3/10
1,4	4/10
1,7	5/10
1,8	6/10
2,2	7/10
2,4	8/10
2,15	9/10
2,18	1

$y_1 = 0,95 \quad y_2 = 1,15 \quad \dots \quad y_{10} = 2,18$

$10 = 2 \times 5$

$m = \frac{y_5 + y_6}{2} = \frac{1,7 + 1,8}{2} = 1,75$

$q_{0,25} = ? \qquad np = 10 \times 0,25 = 2,5$

$[np] = 2 \qquad q_{0,25} = \begin{cases} y_2 = 1,15 \\ \text{or} \\ y_3 = 1,2 \\ \text{or} \\ \frac{y_2 + y_3}{2} = 1,17 \end{cases} \leftarrow a = \frac{1}{2}$

PAPCO

آماره های ترتیبی

order statistics

نمونه تصادفی $X_1 \dots X_n$ را بدون داشتن یافته ای آن یعنی نمی دانیم کدام یک از X_i ها بزرگتر یا کوچکتر می باشد به صورت غیر نزولی به صورت $Y_1 \le Y_2 \le \dots \le Y_n$ فرض کنید. Y_1 را کوچکترین و Y_n را بزرگترین و به طور کلی Y_r را rامین آماره ترتیبی برای نمونه $X_1 \dots X_n$ می نامند. گاهی اوقات آماره ای ترتیبی به صورت $X_{(1)} \dots X_{(n)}$ نیز نمایش داده می شوند.

می دانیم X_i ها چون نمونه تصادفی هستند هم توزیع و مستقل هستند اما وقتی یک Y_r همراه دارای توزیع ای متفاوت و نامستقل هستند.

فرض کنید G_r تابع توزیع Y_r است پس:

$$G_r(y) = P(Y_r \le y)$$

تعریف می کنیم

$$Z_i = \begin{cases} 1 & X_i \le y \quad \text{موفقیت} \\ 0 & X_i > y \quad \text{شکست} \end{cases}$$

$$B = \sum_{i=1}^{n} Z_i \rightarrow$$ داریم تعداد X_i ها کمتر یا برابر y (یکبه یک)

PAPCO

Subject:
Date

$Z_i \sim Ber(p)$

موفقیت: $p = P(X_i \le y) = F_X(y) \rightarrow \beta \sim Bin(n, p)$

X_i کوچکتر یا مساوی y از r تا $r = 1, \dots, n$

$(Y_r \le y)$ ، $(\beta \ge r)$ معادلند داریم

حداقل r تا از X_i ها کوچکتر یا مساوی y پس موفقیت ما میشه

$$G_r(y) = P(Y_r \le y) = P(\beta \ge r)$$

$$= \sum_{k=r}^{n} \binom{n}{k} (F_X(y))^k (1 - F_X(y))^{n-k}$$

تابع چگالی:

$$(F_X(y) + (1 - F_X(y))^n - \underbrace{\sum_{k=0}^{r-1}}_{\text{k=0 تا r-1}} (1 - F_X(y))^n \quad (*)$$

اگر X گسسته باشد

$$g_r(y) = P(Y_r = y) = G_r(y) - G_r(y^-)$$

اگر X پیوسته باشد $\rightarrow$ g مشتق پذیر است: (اثبات صفحه ۷ منبع ۱ بهبودی)

(اثبات حذف) $$g_r(y) = r C_n^r (F_X(y))^{r-1} (1 - F_X(y))^{n-r} f_X(y)$$

$$= \frac{n!}{(r-1)!(n-r)!} [F_X(y)]^{r-1} [1 - F_X(y)]^{n-r} f_X(y)$$

توزیع بتا درجه r

نتایج: (X پیوسته)

if $r = 1$

$$\begin{cases} g_1(y) = n(1 - F_X(y))^{n-1} f_X(y) \rightarrow \text{چگالی آماره مینیم} \\ G_1(y) = 1 - (1 - F_X(y))^n \rightarrow \text{توزیع آماره مینیم} \end{cases}$$

(*)

if $r = n$

$$\begin{cases} g_n(y) = n f_X(y) [F_X(y)]^{n-1} \rightarrow \text{چگالی آماره ماکسیم} \\ G_n(y) = [F_X(y)]^n \rightarrow \text{چگالی توزیع آماره ماکسیم} \end{cases}$$

PAPCO

X گسسته

$f_1(y) = P(Y_1 = y) = (1 - F_X(y^-))^n - (1 - F_X(y))^n$

$f_n(y) = P(Y_n = y) = (F_X(y))^n - (F_X(y^-))^n$

مثال ۱: تاس تاریبی را به تصادف و مستقل سه بار می اندازیم اگر X_1, X_2, X_3

تنها راه حل آن است که تابع تصادفی سه تایی از توزیع زیر است

x	۱	۲	۳	۴	۵	۶
$P(X=x)$	۱/۶	۱/۶	۱/۶	۱/۶	۱/۶	۱/۶

توزیع و چگالی Y_3 ؟

$$F_X(x)=\begin{cases}0 & x<1\\ x/6 & 1\le x<6\\ 1 & x\ge 6\end{cases} \qquad P_X(x)=\begin{cases}1/6 & x=1,\dots 6\\ 0 & o.w\end{cases}$$

$$G_3(y)=\sum_{k=3}^{3} C_3^k \left(\frac{y}{6}\right)^k \left(1-\frac{y}{6}\right)^{3-k}=\left(\frac{y}{6}\right)^3$$

$P(Y_3\le y+1)$ $P(Y_3\le y)$ $P(Y_3<y)$

$$G_3(y)=\begin{cases}0 & y<1\\ (y/6)^3 & 1\le y<6\\ 1 & y>6\end{cases} \qquad g_3(y)=G_3(y)-G_3(y^-)$$

$$=\left(\frac{y}{6}\right)^3-\left(\frac{y-1}{6}\right)^3$$

مثال ۲: یک تابع تصادفی به صورت X_1, X_2, X_3, X_4 از توزیع

$$f_X(x)=\theta e^{-\theta x} \qquad x>0 \quad \theta>0$$

داریم توزیع Y_1 را بیابید.

$$F_X(x)=1-e^{-\theta x} \qquad x>0 \quad \theta>0$$

$$G_1(y)=1-(1-F_X(y_1))^n=1-(1-1+e^{-\theta y})^n$$

$$=1-e^{-4\theta y}$$

$$g_1(y)=4\theta e^{-4\theta y} \qquad y>0$$

$$\theta>0$$

PAPCO

- توزیع توأم Y_r و Y_s :

در حالت پیوسته اگر Y_r و Y_s دو آماره ترتیبی از توزیع پیوسته $F_X(x)$ باشند آن‌گاه می‌توان نوشت:

$$g_{r,s}(y,z) = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} (F_X(y))^{r-1}$$

$$(F_X(z) - F_X(y))^{s-r-1} (1 - F_X(z))^{n-s} f_X(y) f_X(z)$$

- حالت کلی توأم تمام آماره‌های ترتیبی یک نمونه تصادفی n تایی در حالت پیوسته:

$$g_{1,2,\dots,n}(y_1 \dots y_n) = \begin{cases} n! f_X(y_1) f_X(y_2) \dots f_X(y_n) & y_1 < y_2 < \dots < y_n \\ 0 & \text{o.w} \end{cases}$$

مثال ۳: نمونه تصادفی X_1, X_2, X_3, X_4, X_5 از توزیع نمایی در نظر گیرید. حالت توأم (Y_2, Y_4) را بیابید. حالت توأم $(Y_1, Y_2, Y_3, Y_4, Y_5)$ را بیابید

$$g_{2,4}(y,z) = \frac{5!}{1!\,1!\,1!} (1 - e^{-y})(e^{-y} - e^{-z}) e^{-z} e^{-y} e^{-z}$$

$$= 120 (1 - e^{-y})(e^{-y} - e^{-z}) e^{-y} e^{-2z} \qquad 0 < y < z$$

$$g_{1,2,3,4,5}(y_1, y_2, y_3, y_4, y_5) = 120\, e^{-(y_1 + \dots + y_5)}$$

$$y_1 < y_2 < \dots < y_5$$

مثال ۴: نمونه تصادفی X_1, X_2, X_3 از تابع توزیع زیر گرفته شده است

$$F_X(x)=\begin{cases}0 & x<0\\ x^2 & 0\le x<1\\ 1 & x\ge 1\end{cases} \rightarrow f_X(x)=2x \quad 0<x<1$$

چگالی آماره برد را بیابید

$$R = Y_3 - Y_1 \qquad y=y_2 \qquad z=y_3$$

$$g_{1,3}(y,z)=\begin{cases}24yz(z^2-y^2) & y<z\\ 0 & y\ge z\end{cases}$$

$$\frac{3!}{0!\,1!\,0!}\cdot[y^2]^0[z^2-y^2]^1[1-z^2]^0\,2y2z$$

$$= 24\,yz(z^2-y^2)$$

$$\begin{cases} r=z-y \\ w=z \end{cases} \rightarrow \begin{cases} y=w-r \\ z=w \end{cases}$$

$$J=\begin{vmatrix}\frac{\partial y}{\partial r} & \frac{\partial y}{\partial w}\\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial w}\end{vmatrix}=\begin{vmatrix}-1 & 1\\ 0 & 1\end{vmatrix}=-1$$

$$f_{R,W}(r,w)=\frac{1}{|J|}\times g_{1,2}(w-r,w)$$

$$= 24r^3w+48r w^3-72r^2w^2$$

$$0\le r\le w\le 1$$

$$f_R=\int_r^1 f_{R,W}(r,w)\,dw = 12r-24r^2+12r^3 \quad 0<r<1$$

PqPCO

$$24r\cdot\frac{w^2}{2}+48r\frac{w^4}{4}-72r^2\frac{w^3}{3}$$

ـ آماره‌های ترتیبی توزیع یکنواخت و خواص آن ؟

فرض کنید U متغیر تصادفی یکنواخت در فاصله $(0,1)$ است لذا

$$f_U(u)=\begin{cases}1 & 0<u<1\\ 0 & o.w\end{cases} \qquad F_U(u)=\begin{cases}0 & u<0\\ u & 0<u<1\\ 1 & u>1\end{cases}$$

اگر $U_1 \dots U_r \dots U_n$ نمونه‌ای تصادفی از این توزیع و $V_1 \dots V_r \dots V_n$ آماره‌های ترتیبی آن باشند توزیع این آماره‌ها به صورت زیر است

$$G_r(v)=\sum_{k=r}^{n} C_n^k\, v^k\,(1-v)^{n-k} \qquad 0<v<1 \quad (1)$$

(تعداد حالات k تا از n تا) $\sum C_n^k [F_X(y)]^k (1-F_X(y))^{n-k}$

$$g_r(v)=r C_n^r\, v^{r-1}\,(1-v)^{n-r}\cdot 1 \qquad 0<v<1 \quad (2)$$

$r C_n^r (F_X(y))^{r-1}(1-F_X(y))^{n-r} f_X(y)$

$$r C_n^r = r\cdot\frac{n!}{r!(n-r)!}=\frac{n!}{(n-r)!(r-1)!}$$

از طرفی می دانیم توزیع بتا:

$$\frac{1}{B(\alpha,\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}, \qquad B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$\frac{(r-1)!\,(n-r)!}{n!}$ $\quad$ $\frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}$

پس (۳) توزیع بتا با پارامترهای $\alpha=r$ و $\beta=n-r+1$ است

و می دانیم

$$E(X)=\int_0^1 x\,\frac{1}{B(\alpha,\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\alpha}{\alpha+\beta}$$

$$E(V_r)=\frac{\alpha}{\alpha+\beta}=\frac{r}{n+1}$$

$$E(V_k-V_r)=\frac{k-r}{n+1} \qquad E(V_{r+1}-V_r)=\frac{1}{n+1}$$

$\frac{k}{n+1}-\frac{r}{n+1}=\frac{k-r}{n+1}$ $\qquad$ $\frac{r+1}{n+1}-\frac{r}{n+1}=\frac{1}{n+1}$

PAPCO

Subject:
Date

نتیجه: اگر V_r و V_k دو آماره ترتیبی از آماره مرتب نمونه‌ای تصادفی با توزیع یکنواخت $(0,1)$ باشند

$$V_k - V_r \overset{d}{=} V_{k-r} \rightarrow \begin{cases} Y = V_k - V_r \\ Z = V_r \end{cases}$$

و

$$Var(V_r) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{r(n-r+1)}{(n+1)^2(n+2)}$$

پس آماره V_r، $1 \le r \le n$، دارای توزیع بتا با پارامترهای $\alpha = r$ و $\beta = n-r+1$ است.

قضیه تبدیل انتگرال (Probability Integral Transformation)

قضیه: اگر متغیر تصادفی X در طول فاصله (a,b) دارای توزیع پیوسته $F_X(x)$ باشد آن‌گاه متغیر تصادفی $U = F_X(X)$ (تابعی از X) دارای توزیع $U(0,1)$ می‌باشد.

اثبات: $U = F_X(x)$ $\quad g_U(u) = f_X(x)\,|J| = f_X(x)\cdot f_X^{-1}(x) = 1$

$$\Downarrow$$

$$J = \frac{1}{f_X(x)}$$

نتیجه: فرض کنید X متغیر تصادفی با تابع توزیع پیوسته $F_X(x)$ است و $X_1, \dots, X_n$ نمونه‌ای از این توزیع و $Y_1, \dots, Y_n$ آماره ترتیبی نمونه مذکور است با تغییر متغیر زیر نتایج جالبی به دست می‌آید:

① در تبدیل متغیر تصادفی $F_X(X_1), \dots, F_X(X_n)$ که اگر $U_i = F_X(X_i)$، نمونه $U_1, \dots, U_n$ نمونه‌ای تصادفی از توزیع یکنواخت $(0,1)$ است.

② می‌دانیم $Y_1 < \dots < Y_n$ از طرفی F تابع صعودی است پس

$$F(Y_1) < F(Y_2) < \dots < F(Y_n)$$

PAPCO

که اثر تراز دهم $V_r = F(Y_r)$ آن‌گاه $V_1 \dots V_n$ آماره‌ی ترتیبی توزیع یکنواخت هستند که دارای توزیع بتا می‌باشند.

مثال: فرض کنید X یک متغیر تصادفی با تابع توزیع پیوسته $F_X(x)$ است حاصل $E(F_X(X))$ ، $Var(F_X(X))$ چیست؟

$$E(F_X(X)) = \frac{1}{2} \qquad Var(F_X(X)) = \frac{1}{12}$$

مثال: فرض کنید $F_X(x)$ تابع توزیع پیوسته‌ای برای متغیر تصادفی X است اگر $X_1 \dots X_4$ نمونه‌ای از این توزیع، و $Y_1 \dots Y_4$ آماره‌ی ترتیبی این نمونه باشند

$E(X) = a$ فرض:

$$E(Y_1 + Y_2 + Y_3 + Y_4) = ?$$

$$E(F(Y_1) + F(Y_2) + F(Y_3) + F(Y_4)) = ?$$

$$E(Y_1 Y_2 Y_3 Y_4) = ?$$

$$Y_1 + Y_2 + Y_3 + Y_4 = X_1 + X_2 + X_3 + X_4 \Rightarrow E\left(\sum_{i=1}^{4} Y_i\right) = 4a$$

$$Y_1 Y_2 Y_3 Y_4 = X_1 X_2 X_3 X_4 \rightarrow E\left(\prod_{i=1}^{4} Y_i\right) = a^4$$

$$V_r = F(Y_r) \qquad E(V_1 + V_2 + V_3 + V_4) = \frac{1}{n+1} + \frac{2}{n+1}$$

$$E\,V_r = \frac{r}{n+1} \qquad + \frac{3}{n+1} + \frac{4}{n+1} = \frac{10}{n+1} = \frac{10}{5} = 2$$

Subject:
Date

دلیل ناشناختن قضیه انتگرال تبدیل به صورت انتگرال این است که با استفاده از تبدیل $U = F(X)$ می‌توان انتگرال زیر را که در آن چگالی X به کار رفته است

$$P(X \le x) = \int_{-\infty}^{x} f_X(x)\,dx$$

به انتگرال ساده زیر تبدیل کرد

$$P(X \le x) = P(F(X) \le F(x)) = P(U \le u) = \int_0^u du = u$$

قضیه ۲: اگر $X \overset{d}{=} Y$ و $g(t)$ تابعی دلخواه باشد آن‌گاه $g(X) \overset{d}{=} g(Y)$

فرض کنید g تابعی اکیداً صعودی است و داریم

$$F_{g(X)}(z) = P(g(X) \le z) = P(X \le g^{-1}(z))$$

$$F_{g(Y)}(z) = P(g(Y) \le z) = P(Y \le g^{-1}(z))$$

از طرفی $X \overset{d}{=} Y$ پس $F_{g(X)}(z) = F_{g(Y)}(z)$ $\forall z \in R$ پس $g(X) \overset{d}{=} g(Y)$

- به طور کلی اگر $(X_1 \dots X_n) \overset{d}{=} (Y_1 \dots Y_n)$ آنگاه $g(X_1 \dots X_n) \overset{d}{=} g(Y_1 \dots Y_n)$

(اثبات به راحتی با همان ایده در سطح این درس ممکن است)

مثال: فرض کنید $X \overset{d}{=} Y$ نشان دهید که

$$2X^2 - X \overset{d}{=} 2Y^2 - Y$$

$$\sin 2X \overset{d}{=} \sin 2Y$$

از قضیه بالا

PAPCO

قضیه: برای توزیع یکنواخت روی فاصله (۰،۱) با تعریف V_r آماره ترتیبی r ام نمونه

$$\int_0^p g_r(v)dv = P(V_r \le p) = G_r(p)$$

$$= \sum_{k=r}^{n} \binom{n}{k} p^k (1-p)^{n-k} \qquad 0<p<1$$

$= 1 - Bin(n,p,r-1)$ (تقریبا با ارتباط بین توزیع دوجمله‌ای و دنباله نامعلوم است)

اثبات دیگر: $V_r \sim Beta(r, n-r+1)$ پس

$$P(V_r \le p) = \int_0^p r\, C_n^r\, t^{r-1}(1-t)^{n-r}\,dt$$

از طرف دیگر $P(V_r \le p) = \sum_{k=r}^{n} C_n^k p^k (1-p)^{n-k}$

$G_r(v) = \sum C_n^k v^k (1-v)^{n-k}$

پس $\int_0^p r\, C_n^r\, t^{r-1}(1-t)^{n-r}\,dt = \sum_{k=r}^{n} C_n^k p^k (1-p)^{n-k}$

$$\int_0^p \underbrace{t^{r-1}(1-t)^{n-r}}_{\text{انتگرال بتای ناقص}}\,dt = \frac{1}{r\,C_n^r} \underbrace{\sum_{k=r}^{n} C_n^k p^k (1-p)^{n-k}}_{1 - B_{in}(n,p,r-1)}$$

$B \sim Bin(p,p)$ $\quad P(B \le b) = \sum_{k=0}^{b} \binom{n}{k} p^k q^{n-k} = B_{in}(n,p;b)$

$$\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = 1 \Rightarrow \sum_{k=0}^{r-1} C_r^k p^k (1-p)^{n-k} + \sum_{r}^{n} C_n^k p^k (1-p)^{n-k}$$

$$\Rightarrow \sum_{r}^{n} C_n^k p^k (1-p)^{n-k} = 1 - \underbrace{\sum_{k=0}^{r-1} C_r^k p^k (1-p)^{n-k}}_{B_{in}(n,p,r-1)}$$

Subject:
Date

مثال ۲ : انتگرال زیر را حساب کنید

$$\int_0^{1/4} v^4 (1-v)^{15}\, dv$$

$$\begin{cases} r-1=4 \\ n-r=15 \end{cases} \rightarrow r=5 \,,\; n=20$$

$$= \frac{1}{5C_{20}^5} \sum_{k=5}^{20} C_{20}^k \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{20-k}$$

$$= \frac{1}{5C_{20}^5} \left(1 - \underbrace{Bin(20, \tfrac{1}{4}, 4)}_{\text{از جدول } 0/4148}\right) \quad \text{جدول ۲۷۴ ص}$$

$$= 7/849 \times 10^{-9}$$

نکته ۱ : فرض کنید Q_p چندک مرتبه p ام توزیع نیوسته $F_X(x)$ باشد اگر $X_1 \dots X_n$ نمونه تصادفی از توزیع مذکور و $Y_1 \dots Y_n$ آماره‌های ترتیبی آن باشند آنگاه

$$P(Y_r \le Q_p) = G_r(Q_p)$$

$$= \sum_{k=r}^{n} C_n^k \underbrace{[F_X(Q_p)]^k}_{P} \underbrace{[1-F_X(Q_p)]^{n-k}}_{P}$$

$P(X \le Q_p) = p$ ← تعریف چندک

$$= \sum_{k=r}^{n} C_n^k\, p^k (1-p)^{n-k} = 1 - Bin(n, p, r-1)$$

(به قضیه قبل مراجعه کنید)

نکته ۲ : مشابه نکته ۱ داریم

$$P(Y_r \le Q_p \le Y_s) = G_r(Q_p) - G_s(Q_p)$$

$P(Y_r \le Q_p) = P(Y_s \le p)$

Subject:
Date

$= (1 - Bin(n, p, \cancel{r}-1) - (1 - Bin(n, p, s-1))$

$= Bin(n, p, s-1) - Bin(n, p, r-1)$

مثال: در یک توزیع پیوسته اگر $Y_{(5)}$ آماره ترتیبی مرتبه پنجم نمونه‌ای به حجم ۱۰ باشد

$P(Y_5 \le Q_{0.125}) = ?$ $\quad P(Y_{\frac{40}{8}} < Q_{0.5} < Y_{\frac{40}{8}+1})$

$P(Y_{15} \le Q_{0.125}) = 1 - Bin(10, 0.25, 4) = 1 - 0.9219$

$P(Y_5 < Q_{\frac{1}{2}} < Y_7) = Bin(10, 0.75, 6) - Bin(10, 0.75, 4)$

$\frac{0.1209 + 0.2503}{2} \qquad \frac{0.1743 + 0.0949}{2}$

مثال: برای متغیر تصادفی $f_X(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & o.w \end{cases}$ چارک اول را بیابید

بهترین فاصله اطمینان ناپارامتری برای Q_p وقتی $n \ge 30$:

فرض کنید $X_1, \dots, X_n$ نمونه‌ای از تابع توزیع پیوسته $F_X(x)$ باشد. فاصله اطمینان ناپارامتری برای Q_p عبارت است از

$P(Y_r \le Q_p \le Y_s) =$

$Bin(n, p; s-1) - Bin(n, p; r-1)$

$\cong \gamma$

تعریف می‌شود. باید بهترین r و s را پیدا کنیم طوری که ضریب اطمینان حداقل $\gamma = 0.95$ صدق کند. اگر بیش از دو r عبارت فوق بود 0.95 که حداقل یا آزمون رضا

این کار احمقانه است.

مثال ۹: برای یک نمونه ۱۹ تایی از توزیع پیوسته و صعودی $F_X(x)$ یک فاصله اطمینان

نود درصدی برای حد یک چارک هایی مثل Q_3 پیدا کنید

باید r و s را طوری پیدا کنیم که

$$P(Y_r \le Q_p \le Y_s) = B(19, 0.75, s-1) - B(19, 0.75, r-1) \simeq 0.9$$

از روی جدول

$B(19, 0.75, 14) = 0.9465$

$B(19, 0.75, 8) = 0.0271$

$\rightarrow 0.9465 - 0.0271 \simeq 0.9194$

نکته: اگر n به قدر کافی بزرگ باشد می توان از تقریب نرمال استفاده کرد. $(np \ge 15)$

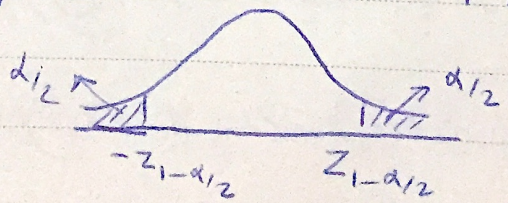
می دانیم توزیع دوجمله ای $B \sim Bin(n,p)$ (تعداد موفقیت) برای n های بزرگ تقریباً با

متغیر نقدی $X \sim N(np, npq)$ هم توزیع است. $B \overset{d}{\simeq} \sum_{i=1}^{n} X_i$

$$Bin(n,p,r-1) - Bin(n,p,s-1)$$

$$= P(r-1 \le B \le s-1) \overset{\text{تصحیح پیوستگی}}{\Longrightarrow} P(r-0.5 \le B \le s-0.5)$$

$$\overset{\text{استفاده از تقریب نرمال}}{\Longrightarrow} P\left(\frac{r-0.5-np}{\sqrt{npq}} \le Z \le \frac{s-0.5-np}{\sqrt{npq}}\right) = \gamma$$



Subject:
Date

$$\Rightarrow \begin{cases} s = np + 0.5 + Z_{1-\alpha/2}\sqrt{np(1-p)} \\ r = np + 0.5 - Z_{1-\alpha/2}\sqrt{np(1-p)} \end{cases}$$

اگر $p = 0.5$ فاصله اطمینانی برای $p = 0.25$ ، فاصله اطمینانی برای چارک اول و اگر $p = 0.75$ فاصله اطمینانی برای چارک سوم به دست می آید.

تذکر: دیدیم که درتعیین فاصله اطمینان برای پارامتر Q_p توزیع آماره ای ترتیبی Y_r و Y_s درنخستین تقریب حاصله مورد بررسی نمی باشند و تنها توزیع آماره B (تعداد موفقیت) در این محاسبه به کار رفته است. از این رو توزیع آماره B را یک توزیع آزاد و فاصله اطمینان به دست آمده را یک فاصله اطمینانی ناپارامتری یا آزاد-توزیع می گویند.

مثال: در یک نمونه تصادفی ۱۰ تایی از توزیع پیوسته $F_X(x)$ ، فاصله (Y_2, Y_9) را به عنوان یک فاصله اطمینانی برای میانه توزیع در تقریبی نزدیک ضریب اطمینانی را بدست آورید.

$$n = 10 \quad p = 0.5 \quad r = 2 \quad s = 9$$

$$\alpha = B(10, 0.5, 8) - B(10, 0.5, 1) = 0.9785 - 0.0107 = 0.9678$$

اگر یافته این نمونه ۱۰ تایی به صورت زیر باشد

Subject:
Date

۴٫۵ ۴٫۶ ۴٫۳ ۵٫۱ ۴٫۳ ۴٫۱ ۳٫۵ ۲٫۱
۳٫۸ ۳٫۹

آن گاه $(Y_2, Y_9) = (4.3, 4.9)$

پس (Y_2, Y_9) یک فاصله اطمینان ۹۸٪ برای میانه است. اگر نمونه را یک نمونه تصادفی از توزیع نرمال در نظر بگیریم میانه همان میانگین توزیع است و فاصله اطمینان آن برابر با

$$\bar{X} \pm t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} \qquad 1-\alpha = 0.98$$

که از روی نمونه برآورد می‌کنیم.

مثال: برای مثال ۹ با استفاده از تقریب نرمال برای Q_3 بازه اطمینان ۹۵٪ به دست آورید.

$$\gamma = np + 0.5 \pm Z_{1-\alpha/2}\sqrt{npq} = 9.4507 \simeq 10$$

$$\delta = 19 * 0.75 + 0.5 + 1.645\sqrt{19 \times 0.75 \times 0.25} = 15.3490 \simeq 15$$

الف)

پس $P(Y_{10} < Q_3 < Y_{15}) = B(19, 0.75, 14)$

$$- B(19, 0.75, 9) \simeq 0.86$$

که از نود و صد کمتر است چون $n = 19$ عدد کمی بزرگ است

مثال ۲: فرض کنید نمونه 30 تایی از توزیعی پیوسته و کندی صعودی استخراج شده است

۰/۱۵ ۰/۱۳۸ ۰/۱۳۳ ۰/۱۲۸ ۰/۱۲۶ ۰/۱۹ ۰/۱۶ ۰/۱۱۴ ۰/۱۱
۰/۹۲ ۰/۹۱ ۰/۱۸۸ ۰/۱۸۷ ۰/۱۸۶ ۰/۱۷۶ ۰/۱۶۳ ۰/۱۴۲ ۰/۱۵۸
۱/۷۲ ۱/۴۶ ۱/۲۱ ۱/۱۹ ۱/۱۵ ۱/۱۵ ۱/۰۱ ۰/۹۵ ۰/۹۴
۱/۷۳ ۱/۹۹

دهک هفتم توزیع را با استفاده از نمونه فوق برآورد کنید و یک فاصله اطمینان ۹۵ درصدی برای میانه توزیع بیابید.

$$a\,y_{[np]} + (1-a)\,y_{[np]+1}$$

$30 \times 0/7 = 21$ $\qquad a\,y_{21} + (1-a)\,y_{22}$

$(n+1)p = 31 \times 0/7 = 21/7 \longrightarrow \begin{matrix} a = 1 - 0/7 = 0/3 \\ 1-a = 0/7 \end{matrix}$

$$0/3\,y_{21} + 0/7\,y_{22} = 0/3 \times 0/95 + 0/7 \times 1/01 = 0/992$$

$$P(y_r < Q_{0/5} < y_s) \underset{\text{تقریبا}}{\geq} 0/95$$

با استفاده از نرم افزار

$= Bi(30, 0/5, 19) - Bi(30, 0/5, 8)$

یا $Bi(30, 0/5, 24) - Bi(30, 0/5, 6)$ $\Big\} \simeq 0/95$

با تقریب نرمال

$$r = np + 0/5 - z_{1-\alpha/2}\sqrt{npq} = 30 \times 0/5 + 0/5 - 1/96\sqrt{30 \times 0/5 \times 0/5}$$
$$= 10/12 \simeq 10$$

$$s = 30 \times 0/5 + 0/5 + 1/96\sqrt{30 \times 0/5 \times 0/5} = 20/87 \simeq 21$$

(y_{10}, y_{21})

Subject:
Date

(الف

```
X <- c(0.11, 0.12, ..... , 1.99)

X <- sort(X)
p <- 0.7
n <- length(X)
r <- floor((n+1)*p)
w <- (n+1)*p - r
Qp <- (1-w)*X[r] + w*X[r+1]
Qp
```

(ب

```
X <- c(0.11 , ... , 1.99)
X <- sort(X)
a <- 0.95
n <- length(X)
p <- 0.5
i <- { }

M <- matrix(0, n*(n-1)/2 , 3)
    for(j in 1:(n-1)){
     for(k in (j+1):n){
      i = c(i,k)
      l = length(i)
      M[l,1] = j
      M[l,2] = k
      M[l,3] = pbinom(k-1,n,p) - pbinom(j-1,n,p)
      }
     }
```



```
M
P=M[,3]   a1 = min(p[p >= a])   C=M[M[,3]==a1,]
dimnames(C) = list(NULL, c('j', 'k', '1-alpha'))   C
```

Subject:
Date

```
M
p = M[,3]
a1 = min(p[p>=a])
C = M[M[,3] == a1,]
dimnames(C) = list(NULL, c('j', 'k', '1-alpha'))

C
```

صـ۱۰۳

- آزمون فرض در خصوص چندک p:

فرض کنید X یک متغیر تصادفی پیوسته با تابع توزیع پیوسته $F_X(x)$ است اگر Q_p چندک مرتبه p ام توزیع باشد می‌خواهیم که بتوان برای آن برآورد نقطه‌ای به طریقه بدست آورده. اکنون فرض کنید a یک عدد معلوم و داده شده است می‌خواهیم این فرض را که $a = Q_p$ است آزمون کنیم.

- آزمون علامت (نشانه) (دوجمله‌ای)

برای این یک فرض $Q_p = a$ ، $0<p<1$ آزمون کنیم مراحل زیر را دنبال می‌کنیم. ① فرض‌های آماری عبارتند از

$$\begin{cases} H_0: Q_p = a \\ H_1: Q_p > a \end{cases} \qquad \begin{cases} H_0: Q_p = a \\ H_1: Q_p < a \end{cases} \qquad \begin{cases} H_0: Q_p = a \\ H_1: Q_p \neq a \end{cases}$$

Subject:
Date

② سطح آزمون را α در نظر می‌گیریم یعنی $P(\text{RH}_0 \mid H_0 \text{ درست}) = \alpha$

③ نمونه تصادفی $X_1, \dots, X_n$ از توزیع $F_X(x)$ اختیار می‌کنیم

④ آماره آزمون را تعریف می‌کنیم.

آماره‌ای را که برای آزمون فرض H_0 در مقابل H_1 اختیار می‌کنیم عبارت است از:

$B =$ تعداد X_iای که از a بیشترند

که اگر RR را از هر X_i کم کنیم داریم

$$X_1 - a, X_2 - a, \dots, X_n - a$$

لذا B برابر با تعداد علامت‌های مثبت در عبارت بالا که توزیع این آماره دوجمله‌ای با پارامترهای n و p.

p مرضی $p = P_{H_0}(X_i > a) = 1 - p$

$B \sim Bin(n, 1-p)$

⑤ تعریف ناحیه بحرانی آزمون فرض مقابل $H_1: Q_p > a$ ، $Q_p = a$

I) if $B \geq K_\alpha \Rightarrow \text{RH}_0$ بنابراین آن‌گاه

II) $H_1: Q_p < a$ ، if $B \leq K_\alpha \Rightarrow \text{RH}_0$ همین‌طور برعکس.

III) $H_1: Q_p \neq a$ if $B \geq K_{\alpha_1}$ or $B \leq K_{\alpha_2} \Rightarrow \text{RH}_0$

که در آن‌ها K_α کوچکترین عددی است که

① $P(B \geq K_\alpha) \leq \alpha$

② $P(B \leq K_\alpha) \leq \alpha$

③ $P(B \geq K_{\alpha_1}) + P(B \leq K_{\alpha_2}) \leq \alpha$ $\Rightarrow \text{RH}_0$

P-value

PAPCO

p-value $\begin{cases} P(B \geq b) \leq \alpha \\ P(B \leq b) \leq \alpha \\ 2\min(P(B \geq b), P(B \leq b)) \leq \alpha \end{cases}$

مثال ۳ : برای مثال ۱۲، داده‌های زیر آزمون $\alpha = 0.05$ را در سطح $\begin{cases} H_0: Q_{0.10} = 1 \\ H_1: Q_{0.10} < 1 \end{cases}$ رد می‌کنند؟

$$B = \sum_{i=1}^{30} I_{\{x_i > 1\}} = 9 \qquad B \underset{H_0}{\sim} Bin(n, 0.10)$$

$$p\text{-value} = p_{H_0}(B \leq 9) = 0.0429 < 0.05 \Rightarrow RH_0$$

pbinom(9, 30, 1-p)

X <- c(0.11, ..., 1.099)

library(BSDA)

SIGN.test(X, md=1, alternative='less')

greater
two.side

مثال ۱۴ : داده‌های زیر آخرین مقاومت برحسب کیلوگرم برای یک نوع مفتول است که در سیب آزمایش تهیه شده‌اند

163	165	162	189	161					
171	158	151	169	182					
163	139	172	165	148	166	172	163	187	173

stat / Nonparametrics $\begin{cases} H_0: Q_{0.10} = 14 \\ H_1: Q_{0.10} > 14 \end{cases}$ آزمون

1- sample Sign / Test median ☐ آزمون alternative

✓ confidence نمایش

+ + + + + + − − +
+ + − + + − + + + + +

$$B = \sum_{i=1}^{20} I_{\{x_i > 14\}} = 14 \qquad p(B \geq 13) \leq 0.05$$

چون $14 \geq k_\alpha = 13 \Rightarrow RH_0$

$$p\text{-value} = \underset{H_0}{P}(B \geq 14) = 1 - P(B \leq 14) = 0.0039 < 0.05$$

pbinom(13, 20, 0.5)

X<-c(163, ..., 173) library(BSDA) SIGN.test(X, md=140, alternative='greater')

Subject: p=0.75 A=X[X!=140] n=length(A) b=length(A[A<160]) pvalue=pbinom(b,n,1-p)

Date

مثال 10:

b	pvalue
89 90 86 80 97 81 94 82	
87 93 94 84 83 78 98	n=15

$$\begin{cases} H_0: Q_{3/4} = 85 \\ H_1: Q_{3/4} \neq 85 \end{cases} \qquad \alpha = 0.1$$

+ + + − + − + − + + + − − − +

$$B_1 = \sum_{i=1}^{15} I_{\{X_i > 85\}} = 9$$

$$2 \min\{P(B \geq 9), P(B \leq 9)\} = 2 \times 0.0042 = 0.008 \Rightarrow RH.$$

$P(B \geq 9) = 1 - pbinom(8, 15, 0.25) = 0.0042$

$K_{\alpha_1} = 1 \quad K_{\alpha_2} = 9 \rightarrow$ کوچکترین اعداد ک

$$P(B \geq K_{\alpha_1}) + P(B \leq K_{\alpha_2}) \leq 0.1$$

$$\begin{cases} P(B \leq 1) = 0.0802 \\ P(B \geq 9) = 1 - P(B \leq 8) = 0.042 \end{cases}$$

$$B = 9 \geq K_{\alpha_1} = 1 \rightarrow RH.$$

y<-c(89, 90, ..., 98) p=0.75 A=y[y!=85]

n=length(A) b=length(A[A>85])

d<-pbinom(b, n, 1-p) a<-1-pbinom(b-1, n, 1-p)

P value<-2*min(a, d) pvalue

PAPCO

Subject:
Date

"آماره‌های مرتبه‌ای"

فرض کنید $X_1 \dots X_n$ یک نمونه تصادفی از توزیع پیوسته $F_X(x)$ است و $Y_1 < \dots < Y_n$ آماره‌های ترتیبی نمونه مذکور است. در این نمونه مرتب شده رتبه X_i را با R_i نشان می‌دهیم.

مثلاً

$X_1 = 3.4 \quad X_2 = 2.8 \quad X_3 = 5.2 \quad X_4 = 1.9$

$Y_1 = X_4 = 1.9 \rightarrow R_4 = 1$

$Y_2 = X_2 = 2.8 \rightarrow R_2 = 2$

$Y_3 = X_1 = 3.4 \rightarrow R_1 = 3$

$Y_4 = X_3 = 5.2 \Rightarrow R_3 = 4$

بردار R_i یک متغیر تصادفی گسسته است

$R_i = k$ یعنی X_i در جایگاه k ام مرتب

n جایگاه داریم

k ام $\quad n-1$ بقیه جایگاه‌ها

که اعداد ۱ و ۲ و ... n را می‌توانند پر کنند داریم:

$$P(R_i = k) = \frac{\text{تعداد حالات مساعد}}{\text{تعداد حالات ممکن}} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

$r_1 \; r_2 \dots r_n \quad n!$ کل جایگشت

پس توزیع R_i یکنواخت گسسته است و

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

توزیع توأم R_i و R_j:

در این حالت داریم

X_i در جایگاه r ام و X_j در جایگاه s ام بقیه $(n-2)!$

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

$$Cov(R_i, R_j) = E(R_i R_j) - E(R_i)E(R_j)$$

Subject:
Date

$$\sum_{i=1}^{n} R_i = \frac{n(n+1)}{2}$$ می دانیم

$$var(R_1 + \dots R_n) = var\left(\frac{n(n+1)}{2}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} var(R_i) + \sum\sum_{i \neq j} cov(R_i, R_j) = 0$$

$$\frac{n(n^2-1)}{12} + n(n-1)\,cov(R_i, R_j) = 0$$

$R_1\ R_2\ R_3\ R_4$

$4(4-1) = 12$ / 4 جایگشت

$$\Rightarrow Cov(R_i, R_j) = \frac{-n(n^2-1)/12}{n(n-1)} = -\frac{n+1}{12}$$

پس R_i ها هم توزیع ولی غیر مستقل هستند.

- توزیع توأم R_i ها:

$$P(\underbrace{R_1 = r_1, \dots, R_n = r_n}_{\text{حالت مساعد} = 1}) = \begin{cases} \frac{1}{n!} & r \in R \\ 0 & \text{o.w} \end{cases}$$

کل $= n!$

- تعریف آماره رتبه ای:

فرض کنید $X_1 \dots X_n$ نمونه ای تصادفی از توزیع پیوسته $F_{X(x)}$ و $R = (R_1 \dots R_n)$ بردار رتبه ها باشد. هر تابعی از R مانند $T(R)$ را یک آماره رتبه ای می نامند و چون توزیع R به $F_{X(x)}$ بستگی ندارد پس آماره $T(R)$ از آماره رتبه ای توزیع آزاد یا ناپارامتری می باشد.

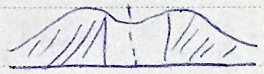
آزمون رتبه‌ای علامت‌دار ویلکاکسون (آزمون رتبه-علامت ویلکاکسون)

Wilcoxon Signed-rank test

فرانک ویلکاکسون (1945)

در آزمون فرض معمول $H_0: \theta = \theta_0$ دیدیم، آماره آزمون براساس علامت به صورت $X_1 - \theta_0 \dots X_n - \theta_0$ ساخته می‌شود و آماره آزمون براساس تعداد مثبت‌ها محاسبه و مقدار احتمال به دست آورده می‌شد، آزمون علامت بدین نحو تهیه شد. اکنون می‌خواهیم علاوه بر علامت این تفاوت‌ها به نحوی از مقادیر شان نیز استفاده کنیم

فرض کنید $X_1 \dots X_n$ نمونه‌ای تصادفی از توزیعی پیوسته است و توزیع نسبت به θ متقارن است یعنی $P(X \le \theta - x) = P(X \ge \theta + x)$ یا $f(\theta + x) = f(\theta - x)$



با قبول این فرض آماره رتبه علامت‌دار ویلکاکسون به صورت زیر تعریف می‌گردد

۱- قدر مطلق $D_i = X_i - \theta_0$

۲- $|D_i|$ها را مرتب کرده و رتبه آن‌ها را به اندازه می‌آوریم و فرض می‌کنیم R_i رتبه $|D_i|$ پس از مرتب‌سازی

$$Z_i = \begin{cases} 1 & D_i > 0 \\ 0 & D_i < 0 \end{cases}$$

Subject:
Date

۳- آماره مورد نظر به صورت زیر تعریف می‌شود

$$T^+ = W = \sum_{i=1}^{n} R_i Z_i$$

(جمع رتبه‌هایی که علامت مثبت دارند)

مثال: فرض کنید $x_1 = 2.2$ $x_2 = 2.8$ $x_3 = 1.3$ $x_4 = 4.3$

$H_0 : \nu = 1.5$

	D_i	$\lvert D_i \rvert$	R_i	Z_i	$R_i Z_i$
$4.2 - 1.5$	1.7	0.2	1	0	0
$2.8 - 1.5$	1.3	1.3	2	1	2
	-0.2	1.7	3	1	3
	2.8	2.8	4	1	4

$W = 9$

- توزیع W (T^+)

با استفاده از احتمال و معلوم بودن n و تحت فرض H_0 حالتی T^+ را می‌سنجیم

$\min T^+_{min} = 0$ و این است که همه D_i ها منفی

$\max T^+ = \frac{n(n+1)}{2}$ همه D_i ها مثبت

$1 + 2 + \cdots + n$

اگر $n = 3$ باشد با توجه به حالات مختلفی که D_1، D_2 و D_3 دارند توان نوشت

1	2	3	Z_i	T^+	
−	−	−	0 0 0	0	$\leftarrow \sum R_i Z_i = 0$
−	−	+	0 0 1	3	$\leftarrow 0 + 3 \times 1 = 3$
−	+	−	0 1 0	2	$\leftarrow 0 + 2 \times 1 = 2$
+	−	−	1 0 0	1	$\leftarrow 1 \times 1 + 0 = 1$
−	+	+	0 1 1	5	$\leftarrow 0 + 2 \times 1 + 3 \times 1 = 5$
+	+	−	1 1 0	3	$\leftarrow 1 \times 1 + 2 \times 1 + 0 = 3$
+	−	+	1 0 1	4	$\leftarrow 1 + 0 + 3 = 4$
+	+	+	1 1 1	6	$1 + 2 + 3 = 6$

PqPCO

Subject:
Date

$$g(t_1) = \frac{\text{تعداد حالات مساعد}}{\text{تعداد حالات ممکن}} = \begin{cases} \frac{1}{8} & t=0 \\ \frac{1}{8} & t=1 \\ \frac{1}{8} & t=2 \\ \frac{1}{4} & t=3 \\ \frac{1}{8} & t=4 \\ \frac{1}{8} & t=5 \\ \frac{1}{8} & t=6 \end{cases}$$

$$g(t) = \begin{cases} \frac{1}{8} & t=0,1,2,3,4,5,6 \\ \frac{2}{8} & t=2 \end{cases} \Leftarrow$$

به طور کلی داریم $g(t) = \frac{k_t}{2^n}$ که k_t تعداد حالاتی است که در آن $T^+ = t$

توزیع T^* به توزیع حاصل بستگی نداشته و آزاد توزیع است. جدول‌های وجود دارد که از آن مقادیر مختلف n و توزیع T^+ را به ما می‌دهند. اکنون با استفاده از آماره T^+ فرض B را آزمون می‌کنیم. در آزمون علامت دیدیم هرگاه توزیع

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

دلیل نادرست بودن فرض H_0 است در این وضعیت نیز به طور شهودی مقادیر T^+ که نشان‌دهنده نادرستی H_0 هستند بزرگ‌اند لذا اگر $T^+ > k_\alpha$ ، H_0 رد می‌شود. پس اگر t مقدار مشاهده شده T^+ باشد (از روی نمونه):

$$p\text{-value} = P_{H_0}(T^+ \geq t) \quad \text{if } p\text{-value} > \alpha \Rightarrow AH_0$$
$$p\text{-value} \leq \alpha \Rightarrow RH_0$$

Subject:
Date

به همین ترتیب برای فرضیه $\begin{cases} H_0: \nu = \nu_0 \\ H_1: \nu < \nu_0 \end{cases}$

$$p\text{-value} = P(T^+ \leq t) \leq \alpha \Rightarrow RH_0$$

و برای فرضیه $\begin{cases} H_0: \nu = \nu_0 \\ H_1: \nu \neq \nu_0 \end{cases}$

$$p\text{-value} = 2P(T^+ \geq t) \leq \alpha \Rightarrow RH_0$$

نکته: آزمون ویلکاکسون رتبه علامت شانه دقیق‌تر از آزمون علامت است و چون علاوه بر توجه به علامت مقدار تفاوت‌ها را نیز در نظر می‌گیرد

مثال ۱۲: مقادیر مشاهده شده یک نمونه تصادفی ده تایی از توزیعی با میانه ν عبارتند از

۱۱٫۸ ۹٫۵ ۹٫۸ ۷٫۲ ۱۱٫۳ ۹٫۲ ۱۴٫۱ ۱۰٫۲
۱۰٫۸ ۸٫۷

در سطح $\alpha = 0.05$ فرضیه $\begin{cases} H_0: \nu = 8 \\ H_1: \nu > 8 \end{cases}$ را آزمون کنید

$X_i - 8 = D_i$	۳٫۸	۱٫۵	۱٫۲	۳٫۳	-۰٫۸	۱٫۸	-۱٫۵	۳٫۸	۰٫۷	۲٫۸
$\lvert D_i \rvert$ جدا از علامت	۰٫۷	۰٫۸	۱٫۲	۱٫۵	۱٫۸	۲٫۲	۲٫۸	۳٫۳	۳٫۸	۹٫۱
R_i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Z_i	۱	۰	۱	۰	۱	۱	۱	۱	۱	۱
t $R_i Z_i$	۱	۰	۳	۰	۵	۶	۷	۸	۹	۱۰

$$T^+ = \sum R_i Z_i = 49$$

$$g(t) = \begin{cases} 1/10 & t = 1,3,5,6,7,8,9,10 \\ 1/5 & t = 0 \end{cases}$$

```
x <- c(10.2, ... 10.8)   y <- 8
wilcox.test(x-y, alternative = "greater")
```

PAPCO

(۱۸)

R: 1 - psignrank(۴۸, ۱۰)

Subject:
Date

$P(T^+ \le 6)$ جدول کتاب با $1-p(T^+ \le 48) = 1 - 0.9863 \simeq 0.014$

$p(T^+ \ge 49) = 0.014 \le 0.05 \Rightarrow RH_0$

$p(T^* \ge 49) = 0.042$: حداقل معناداری H_0 رد می‌شود بیابید

$t > 45 \rightarrow RH_0$ $(49) > 45$

مثال ۱۸: فرض کنید X_i طول یک‌نوع ماهی پرورشی برای ده نمونه از این نوع ماهی اعداد زیر

به دست آمده است فرضیه $\begin{cases} H_0: \mu = 3.7 \\ H_1: \mu \neq 3.7 \end{cases}$

(میانگین طول ماهی از ۳٫۷ متفاوت است)

5 3.9 5.2 5.5 2.8 6.1
6.4 2.6 1.7 4.3

X_i	5	3.9	5.2	5.5	2.8	6.1	6.4	2.6	1.7	4.3
D_i	1.3	0.2	1.5	1.8	-0.9	2.4	2.7	-1.1	-2	0.6
$\lvert D_i \rvert$	1.3	0.2	1.5	1.8	0.9	2.4	2.7	1.1	2	0.6
رتب $\lvert D_i \rvert$	0.2	0.6	0.9	1.1	1.3	1.5	1.8	2	2.4	2.7
R_i	1	2	3	4	5	6	7	8	9	10
Z_i	1	1	0	0	1	1	1	0	1	1

$T^+ = 1 + 2 + 5 + 6 + 7 + 9 + 10 = 40$

$p\text{-value} = 2p(T^+ \ge t) = 2 \times 0.116 = 0.232 \Rightarrow AH_0$

نکته: در کتاب منفی مقادیر عکس داده یعنی ۰ و ۱ را جا به جا کنید مثلاً در این مثال

$0 + 0 + 3 + 4 + 0 + 0 + 0 + 8 + 0 + 0 = 15$

$2p(T^+ \le 15) = 2 \times 0.116 = 0.1232 \Rightarrow A H_0$

P4PCO

Subject:
Date

میانگین و واریانس T^+ تحت فرض H_0:

با توجه به تعریف $T^+ = \sum_{i=1}^{n} R_i Z_i$ و

$$E(T^+) = \frac{n(n+1)}{4}$$

چون $T^+ = \sum_{i=1}^{n} R_i Z_i$ هم توزیع با متغیر U است که به صورت زیر تعریف می‌شود: مجموعی از اعداد ۱ تا n (۱۰۰۰۰) با احتمال $1/2$ و $1/2$

$$U = \sum_{i=1}^{n} u_i \qquad u_i = \begin{cases} 0 & 1/2 \\ i & 1/2 \end{cases}$$

u_i ها تحت فرض H_0 یعنی تحت تقارن پس یعنی شانس علامت هر i مثبت و منفی در T^+ مساوی است پس هم توزیع با U است

$$E(U) = \sum_{i=1}^{n} E(u_i) = \sum_{i=1}^{n} \left[0 \times \frac{1}{2} + i \times \frac{1}{2}\right] = \frac{1}{2} \sum_{i=1}^{n} i$$

$$= \frac{1}{2} \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

$$var(T^+) = \sum_{i=1}^{n} var(u_i) = \sum_{i=1}^{n} \left[E(u_i^2) - E^2(u_i)\right]$$

(u_i ها مستقلند)

$$= \sum_{i=1}^{n} \left\{\left[0^2 \frac{1}{2} + i^2 \frac{1}{2}\right] - \left(\frac{i}{2}\right)^2\right\} = \sum \left(\frac{i^2}{2} - \frac{i^2}{4}\right) = \sum \frac{i^2}{4}$$

$$= \frac{1}{4} \sum_{i=1}^{n} i = \frac{1}{4} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$$

نکته ۱- وقتی n بزرگ باشد توزیع مجانبی T^+ نرمال است و داریم

$$Z = \frac{T^+ - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

نکته ۲: ثابت می‌شود $P(T^+ \geq c) = P(T^+ \leq c^*)$ طوری که

$$c + c^* = \frac{2n(n+1)}{4}$$

PAPCO

(۱۹)

Subject:
Date

$$P(T^+ \geq c) = P\left(T^+ \leq \frac{n(n+1)}{2} - c\right)$$ و

مثال ۱۹: فرض کنید θ میانه توزیعی متقارن از نوع پیوسته است. برای آزمون فرض

$\begin{cases} H_0: \theta = 25 \\ H_1: \theta < 25 \end{cases}$ در سطح $\alpha = 0.011$ یک نمونه تصادفی به حجم $n = 18$ اختیار و مشاهده

کردیم که انحرافات این مشاهدات از ۲۵ اعداد زیر است

۲٫۵ ۱٫۲ -۰٫۱ ۰٫۹ -۰٫۲ -۰٫۸ ۰٫۴ ۱٫۶ -۰٫۵ ۱٫۵

۰٫۶ ۱٫۹ -۰٫۳ ۰٫۷ ۱٫۳

به طور مستقیم و با استفاده از ترتیب زمال آزمون انجام دهید.

$\lvert D_i \rvert$	۰٫۱	۰٫۱	۰٫۲	۰٫۳	۰٫۴	۰٫۵	۰٫۶	۰٫۷	۰٫۸	۰٫۹	۱٫۲	۱٫۴	۱٫۵	۱٫۶	۱٫۹	۲٫۵
R_i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷
Z_i	۰	۰	۰	۱	۰	۱	۱	۰	۱	۱	۱	۱	۱	۱	۱	۱	۱
$R_i Z_i$	۰	۰	۰	۴	۰	۶	۷	۰	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷

$۱+۲+۳+۵+۸ = 19$

$$T^+ = \sum R_i Z_i = 101$$

$$P(T^+ \geq 101) = P(T^+ \geq 19) = 1 - \underbrace{0.1009}_{P(T^+ \leq 19)} \geq 0.011 \Rightarrow AH_0$$

$$\frac{n(n+1)}{2} - c = 101 \rightarrow \frac{18 \times 19}{2} - 101 = c \rightarrow c = 19$$

$$E(T^+) = \frac{18(18+1)}{4} = 90 \quad Var(T^+) = \frac{18(18+1)(2\times18+1)}{24} = 1710$$

$$Z = \frac{101 - 90}{\sqrt{1710}} = 2.658 > Z_{0.01} = 2.326 \Rightarrow AH_0$$

(دنباله دارد)

PAPCO

- رتبه - علامت ویلکاکسون برای بررسی یک متغیر پیوسته یا ترتیبی (یک جامعه) و یا یک متغیر با دو نمونه وابسته مثل قبل و بدون نیاز به فرض دارد. (معادل آزمون تی جامعه وزوجی. میانگین در پارامتری)