Subject :

Year . Month . Date . ( )

تحلیل رگرسیون :

گالتون (۱۸۸۵) در خصوص رابطه بین قد فرزندان و متوسط قد والدین تحقیقی انجام داد و نتیجه گرفت که ~~پسران والدین قد کوتاه~~ بین قد والدین و فرزندان ارتباطی وجود دارد پس وقتی می خواهیم یک متغیر وابسته را از روی یک متغیر مستقل پیش بینی کنیم از تحلیل رگرسیون استفاده می کنیم.

Y معمولاً با $Y|X$ X

قد فرزندان   متوسط قد والدین

وابسته   مستقل

یا فرض کنید می خواهیم معدل کل [illegible] دانشجو در یک نیمسال را از روی نمره درس ریاضیات آنها پیش بینی کنیم

رگرسیون در اقتصاد کاربرد زیادی داریم برای مثال فرض کنید می خواهیم رابطه جمعیت (X) و فروش کالاها (Y) را بررسی کنیم

جمعیت X=x	Y=y
۳۶	۵۴
۲۶	۳۰
۱۲	۲۸
۴۵	۴۸
⋮	⋮

→ پس ارتباط بین X و Y را پیدا می کنیم پس با توجه به معادله که بدست آورده ام به ازاء هر X جدید [illegible] Y را پیش بینی می کنیم.

در این درس مدل رگرسیون خطی ساده مورد بررسی قرار می گیرد.

رگرسیون خطی ساده : Simple linear regression

تحلیل رگرسیونی به مطالعه رابطه ای بین متغیرها مربوط می شود هدف از مطالعه تعیین نوع رابطه برآورد کردن و محاسبه اعتبار رابطه است. پس می توان از رابطه برآورد شده برای پیشگویی مقدار یک متغیر از روی متغیر (متغیرها) دیگر استفاده کرد.

حال اگر بین متغیر مستقل X و متغیر وابسته $Y=Y|X$ یک رابطه خطی برقرار باشد؟

$F_{1,24,0.05}$ $t_{30,0.025}$ $Z_{0.05}$ $Z_{0.025}$ t

Subject:
Year. Month. Date. ( )

گوییم یک مدل رگرسیون ساده خطی بین X و Y برقرار است. برای تشکیل این رابطه

فرض کنید یک نمونه تصادفی ~~(X_1,Y_1) ...... (X_n,Y_n) با متغیرهای ... مشاهده شود~~

~~(x_1,y_1) ...... (x_n,y_n) را مشاهده~~ کنیم n تایی از X و Y داریم.

متغیر کنترل شده

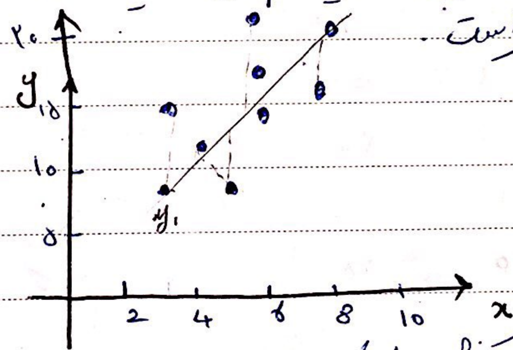
x: متغیر مستقل، متغیر کنترل شده، متغیر پیش‌بینی یا متغیر ورودی (ثبت شده)

y: متغیر پاسخ (متغیر تصادفی)

	پاسخ (وابسته)	مقادیر متغیر مستقل
$y_i = B_0 + B_1 x_i + e_i$	y_1	x_1
	⋮	⋮
	y_n	x_n

باید دقت شود که ابتدا باید بررسی شود که آیا واقعاً بین X و Y ارتباطی به صورت خطی وجود دارد. پس اولین گام برای تحلیل آماری رابطه بین دو متغیر کشیدن نمودار پراکنش دو متغیر است.

دز داروی ضد آلرژی x مستقل	مدت بهبودی (روز) y وابسته
۳	۹
۳	۱۵
۴	۱۲
۵	۹
۶	۱۴
۶	۱۶
۷	۲۲
۸	۱۸
۸	۲۴
۹	۲۳
~~۹~~	~~۲۱~~



نمودار پراکنش scatter plot

پس اگر ببینیم Y و X یک ~~رابطه خطی~~ برقرار است

آن گاه ~~معادله متغیر Y~~ باید خط به صورت زیر برآورد شود.

~~$\hat{Y} = \hat{B}_0 + \hat{B}_1 x$~~

به عبارت دقیق‌تر منظور از رگرسیون خطی این است

$y = B_0 + B_1 x + e$

$E(Y|x) = B_0 + B_1 x$ که میانگین $Y|x$ به طور خطی با x در ارتباط باشد

که در عمل B_0 و B_1 نامعلوم‌اند و به آنها پارامترهای رگرسیونی می‌گوییم

یا ضرایب رگرسیون

فرض‌های اساسی: $E(\varepsilon)=0$

$var(\varepsilon) = var(y) = \sigma^2$

Subject:

Year. Month. Date. ( )

y_1, x_1
y_2, x_2
$\vdots$
y_n, x_n

$$Y_i = B_0 + B_1 x_i + e_i \rightarrow N(0, \sigma^2)$$

$$E(Y \mid x_i) = B_0 + B_1 x_i$$

بنابراین چون B_0 و B_1 نامعلومند از برآورد آنها استفاده می‌شود در پس خط رگرسیون

معمولا با استفاده $\hat{Y}$ برآورد $E(Y \mid x)$ می‌شود. $\hat{Y} = \hat{B}_0 + \hat{B}_1 x$ ~~عرض~~ (عرض از مبدأ یا عرض از مبدأ) B_0 خط رگرسیون

و B_1 شیب خط یا میزان تغییر متغیر پاسخ برحسب هر تغییر یک واحدی متغیر مستقل ~~متغیر~~ را نشان می‌دهد.

دقت کنید چون مقادیر متغیر Y از نقطه‌ای به نقطه‌ای تغییر می‌کند پس مقدار $Y_i = Y \mid x_i$ متغیر برابر با $E(Y \mid x_i)$ نیست مثلاً در مثال ذکر شده متغیر مستقل با ثابت نگه داشتن می‌کنیم دلیل است
میزان هدایت نیز می‌تواند به ازای نقطه‌های مختلف تغییر کند.

برای برآورد $\hat{B}_0$ و $\hat{B}_1$ می‌خواهیم مجموع مربعات خطا مینیمم شود $\sum_{i=1}^{n} e_i^2 \rightarrow \min$

روش حداقل مربعات

روش Least squares method
$\underbrace{\hspace{6cm}}_{LS}$

$$\underset{\text{sum of square errors}}{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - B_0 - B_1 x_i)^2$$

مقادیر $\hat{B}_0$ و $\hat{B}_1$ که SSE را مینیمم می‌کنند برآوردهای حداقل مربعات گویند.

$$\begin{cases} \sum -2(y_i - B_0 - \cancel{B_1 x_i}) = 0 \\ \Rightarrow n B_0 = \sum (y_i - B_1 x_i) \Rightarrow \\ -2 \sum x_i (y_i - B_0 - B_1 x_i) = 0 \end{cases}$$

Subject:

Year. Month. Date. ( )

$\sum (y_i^2+\alpha^2+B^2x_i^2-2\alpha y_i-2y_iBx_i+2\alpha Bx_i$

قضیه: در یک مدل رگرسیون خطی ساده مقادیر $\hat{B}_0$ و $\hat{B}_1$ عبارتند از

$$\hat{B}_0=\bar{y}-\hat{B}_1\bar{x}$$

$$\hat{B}_1=\frac{S_{xy}}{S_{xx}}$$

که در آن

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{(\sum x_i)^2}{n}$$

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{(\sum x_i)(\sum y_i)}{n}$$

برهان: با مشتق گیری از SSE نسبت به B_0 و B_1:

$$\begin{cases}2nB_0-2\sum y_i+2B_1\sum x_i=0\\ 2B_1\sum x_i^2-2\sum y_ix_i+2B_0\sum x_i=0\end{cases}\Rightarrow\begin{cases}\sum_{i=1}^{n}y_i-nB_0-B_1\sum_{i=1}^{n}x_i=0\\ \sum_{i=1}^{n}x_iy_i-B_0\sum_{i=1}^{n}x_i-B_1\sum_{i=1}^{n}x_i^2=0\end{cases}$$

$$\Rightarrow\begin{cases}n\bar{y}-B_1n\bar{x}=nB_0\Rightarrow\boxed{\hat{B}_0=\bar{y}-\hat{B}_1\bar{x}}\\ \left[\sum x_iy_i-B_0\sum_{i=1}^{n}x_i\right]/\sum_{i=1}^{n}x_i^2=B_1\end{cases}$$

جایگذاری کنید و ادامه دهید:

$$\frac{\sum x_iy_i-(\sum y_i/n-\hat{B}_1\frac{\sum x_i}{n})(\sum x_i)}{\sum x_i^2}=\hat{B}_1$$

$$\Rightarrow\frac{\sum x_iy_i-\frac{\sum y_i\sum x_i}{n}-\hat{B}_1\frac{(\sum x_i)^2}{n}}{\sum x_i^2}=\hat{B}_1$$

$$\Rightarrow(\sum x_iy_i)(\sum x_i^2)-\left(\frac{\sum y_i\sum x_i}{n}\right)\Big/\sum x_i^2=\hat{B}_1+\hat{B}_1\left(\frac{\frac{(\sum x_i)^2}{n}}{\sum x_i^2}\right)$$

$$\Longrightarrow\hat{B}_1=\frac{S_{xy}}{S_{xx}}=\frac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^2-n\bar{x}^2}$$

(۲۱)

Subject:
Year. Month. Date. ( )

مجموع مربعات رگرسیون :

کمیت زیری $\sum_{i=1}^{n} (y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$ $= S_{yy} = SST$ است که در نهایت

$\cancel{SSR = SSE = S_{yy}} \Rightarrow$ مجموع مربعات رگرسیون

$\boxed{SSR = \hat{\beta}_1 S_{xy}} \Rightarrow \boxed{SSE = \cancel{SSR}\ S_{yy} - SSR}$

مثال : برای داده های جدول زیر برآورد خط رگرسیون را بیابید . خط را در شکل و نقاط را رسم کنید SSR SSE SST را محاسبه کنید

x_i	۱	۳	۴	۶	۸	۹	۱۱	۱۴
y_i	۱	۲	۴	۴	۵	۷	۸	۹

$\sum_{i=1} x_i = ۵۶$ $\quad \sum_{i=1} y_i = ۴۰$ $\quad \sum x_i^2 = ۵۲۴$ $\quad \sum y_i^2 = ۲۵۶$

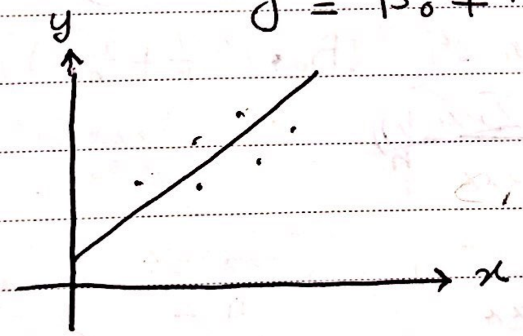
$\sum x_i y_i = ۳۶۴$ $\quad S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = ۱۳۲$

$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = ۳۶۴ - \frac{۵۶ \times ۴۰}{۸} = ۸۴$

$\hat{\beta}_1 = \frac{۸۴}{۱۳۲} = ۰٫۶۳۶$ $\quad SST = S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = ۲۵۶ - \frac{۱۶۰۰}{۸} = ۵۶$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{۴۰}{۸} - ۰٫۶۳۶ \times \frac{۵۶}{۸} = ۰٫۵۴۵$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \Rightarrow \hat{y} = ۰٫۵۴۵ + ۰٫۶۳۶ x$



$SSR = ۰٫۶۳۶ \times ۸۴ = ۵۳٫۴$

$SSE = ۵۶ - ۵۳٫۴ = ۲٫۶$

جدول آنالیز واریانس :

منبع تغییرات	df	SS	MS	F
رگرسیون	۱	SSR	$MSR = SSR/1$	$\frac{MSR}{MSE} > F_{1,n-2,\alpha} \Rightarrow RH_0$
خطا	$n-۲$	SSE	$MSE = SSE/_{n-2}$	
کل	$n-1$	SST		

$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

Subject:
Year. Month. Date. ( )

تمرین : برای داده های جدول زیر برآورد خط رگرسیون را بیابید. نمودار پراکنش و خط رگرسیون برآورد شده را رسم کنید همچنین SSR و SSE را محاسبه کنید

x_i	۰/۵	۰/۷۵	۱	۱/۲۵	۱/۵	۱/۷۵	۲	۲/۲۵
y_i	۱۰	۸	۱۲	۱۲	۱۴	۱۲	۱۶	۱۸
x_i	۲/۵	۲/۷۵	۳	۳/۲۵	۳/۵			
y_i	۱۷	۲۰	۱۸	۲۰	۲۱			

استنباط روی ضرایب رگرسیونی :

دیدیم در مسئله رگرسیونی، معنی دار بودن رگرسیون نیز مورد بررسی قرار گرفت یعنی آیا واقعا بین X و Y ارتباطی وجود دارد.

$$Y = B_0 + B_1 X + \varepsilon$$ خطا

$$E(Y|x) = B_0 + B_1 x \longrightarrow E(\hat{Y}|X) = \hat{B}_0 + \hat{B}_1 X$$

فقط در صورتی که $B_1 = 0$ باشد بین X و Y ارتباطی وجود ندارد پس فرض

$H_0: B_1 = 0$ در برابر فرض مقابل یک طرفه یا دو طرفه $H_1: B_1 \neq 0$ آزمون می شود.

که دیدیم آماره آزمون $F = \frac{MSR}{MSE}$ است در اینجا با فرض نرمال بودن خطاها آماره دیگری معرفی می کنیم

قضیه: در مدل رگرسیون خطی ساده با فرض اینکه $\varepsilon_i \sim N(0, \sigma^2)$ ، $i = 1, \dots, n$

داریم: $\left(B_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

$\cancel{E(\bar{Y} - \hat{B}_1 \bar{X})}$ $\cancel{Y_i = B_0 + BX_i + \varepsilon_i \to Y_i \sim N(B_0 + B_1 X_i, \sigma^2)}$

$$\hat{B}_0 \sim N\left(B_0, \frac{\sigma^2 \sum x_i^2}{n S_{xx}}\right) \qquad \left(\frac{S_{xx} + \sum x_i^2}{n S_{xx}}\right)$$

$\varepsilon \sim N(0, \sigma^2) \Rightarrow Y \sim N(B_0 + B_1 X, \sigma^2)$

$$\hat{B}_1 \sim N\left(B_1, \frac{\sigma^2}{S_{xx}}\right), \quad S^2 = \frac{SSE}{n-2}$$ که در آن

نکته: در عمل معمولا σ^2 مجهول است و یک برآورد خوب ~~برآوردی برای σ^2 است~~

$MSE = \frac{SSE}{n-2}$ برای σ^2 عبارت است از

$$\frac{SSE}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$ می توان نشان داد

PAPCO

$$E(\bar{Y} - \hat{B}_1 \bar{X}) = B_0 + B_1 \bar{X} - \hat{B}_1 \bar{X} \qquad var(\bar{Y} - \hat{B}_1 \bar{X}) = \frac{\sigma^2}{n} + \bar{x}^2 var(\hat{B}_1)$$

Subject:

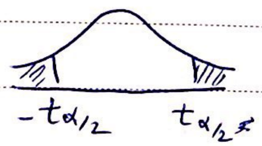
Year. Month. Date. ( )

پس اگر σ^2 مجهول

$$\frac{\hat{B}_1 - B_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

تحت H_0، آماره آزمون : $B_1 = 0$

$$T_0 = \frac{\hat{B}_1}{S/\sqrt{S_{xx}}}$$



$$|T_0| > t_{n-2,\alpha/2}$$

$$\Rightarrow \text{if } T_0 > t_{n-2,\alpha/2} \text{ یا } T_0 < -t_{n-2,\alpha/2} \Rightarrow RH_0$$

آزمون فرض برای B_0 :

$$\frac{\hat{B}_0 - B_0}{S\sqrt{\frac{\sum x^2}{n S_{xx}}}} \sim t_{n-2}, \qquad \frac{\sum x^2}{n S_{xx}} \rightarrow \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}$$

$$\begin{cases} H_0: B_0 = 0 \\ H_1: B_0 \neq 0 \end{cases}$$

$$T_0 = \frac{\hat{B}_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

$$\Rightarrow \text{if } T_0 > t_{n-2,\alpha/2} \text{ یا } T_0 < -t_{n-2,\alpha/2} \Rightarrow RH_0$$

فاصله اطمینان برای $\hat{B}_0$ و $\hat{B}_1$:

$$B_0 : \left(\hat{B}_0 \pm t_{n-2,\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right) \quad \%100(1-\alpha)$$

$$B_1 : \left(\hat{B}_1 \pm t_{n-2,\alpha/2}\, \frac{S}{\sqrt{S_{xx}}}\right)$$

نکته:

تمرین : برای داده‌های جدول زیر معادله خط رگرسیون را بیابید. یک مطالعه توسط یک خرده فروش در ارتباط با رابطه بین مخارج تبلیغ و میزان فروش به طور هفتگی صورت پذیرفته است و داده‌ها به صورت زیر به دست آمده است معادله خط رگرسیون را جهت پیش‌بینی فروش هفتگی از روی هزینه تبلیغات بیابید.

ب) فاصله اطمینان ۹۰ درصد را برای B به دست آورید.

ج) معنی داری آزمون را بررسی کنید.

د) صفر بودن شیب خط رگرسیون را آزمون کنید

PAPCO

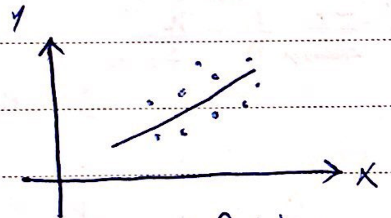
Subject:
Year. Month. Date. ( )

مخارج تبلیغ	۴۰	۲۰	۲۵	۲۰	۳۰	۵۰	۴۰	۲۰	۵۰	۴۰	۲۵	۵۰
فروش	۳۸۵	۴۰۰	۳۹۵	۳۶۵	۴۷۵	۴۴۰	۴۹۰	۴۲۰	۵۶۰	۵۲۵	۴۸۰	۵۱۰

ضریب همبستگی خطی

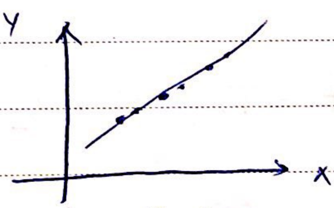
تاکنون فرض براین بوده متغیر مستقل X یک متغیر کنترل شده است ~~و~~ به عبارتی یک متغیر تصادفی نیست حال فرض کنید X و Y هر دو تصادفی باشند روش زیر سودی در مواردی مناسب است که متغیر تصادفی Y به متغیر X که اسب کنترل می شود بستگی داشته باشد. برای اندازه گیری همبستگی دو متغیر تصادفی از معیاری به نام ضریب همبستگی خطی استفاده می شود میزان که به صورت زیر تعریف می شود

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{var(X)\,var(Y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$



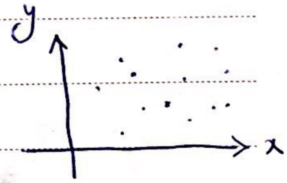
$0 < \rho < 1$

همبستگی مثبت

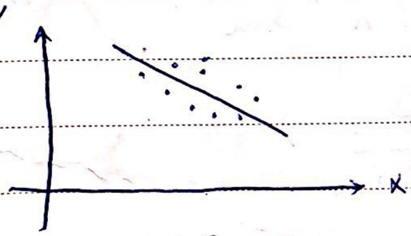


$\rho = 1$

همبستگی کامل مثبت

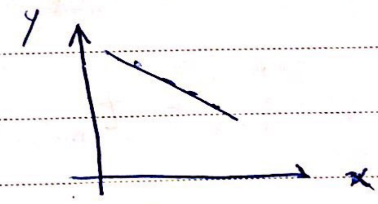


$\rho = 0$



$-1 < \rho < 0$

همبستگی منفی



$\rho = -1$

همبستگی کامل منفی

برای برآورد ضریب همبستگی یک نمونه تصادفی (X_1,Y_1) ــــ (X_n,Y_n) از (X,Y) انتخاب می کنیم واریانسها X و Y و کواریانس X و Y را بدست می آوریم

در نهایت:

برآورد ضریب همبستگی $\hat{\rho} = \dfrac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_{xx}\hat{\sigma}_{yy}}}$

$$S_x = \sqrt{\underbrace{\frac{1}{n-1}\sum(x_i-\bar{x})^2}_{S_{xx}}}$$

$$S_y = \sqrt{\underbrace{\frac{1}{n-1}\sum(y_i-\bar{y})^2}_{S_{yy}}}$$

$$S_{xy} = \underbrace{\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})}_{S_{xy}}$$

دقت کنید همو تکلیف

ضریب نمونه ای $\Longrightarrow r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

وقتی مقادیر مشاهده شده این نمونه در دست باشد (x_1, y_1) ... (x_n, y_n) آن گاه:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

آزمون برای ρ:

برای نمونه های بزرگ تقریباً دارای توزیع $W = \frac{1}{2} \ln \frac{1+R}{1-R}$ فیشر ثابت کرده که

نرمال با میانگین تقریبی $\mu \simeq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ و واریانس تقریبی

$$z = \frac{\ln\sqrt{\frac{1+r}{1-r}} - \ln\sqrt{\frac{1+\rho_0}{1-\rho_0}}}{\frac{1}{\sqrt{n-3}}} = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1-\rho_0}{1+\rho_0} \Big/ \frac{1-r}{1+r}\right)$$

$$\Rightarrow z = \frac{\sqrt{n-3}}{2} \ln\left[\left(\frac{1-\rho_0}{1+\rho_0}\right) \Big/ \left(\frac{1-r}{1+r}\right)\right]$$

می باشد پس داریم $\sigma_w^2 \simeq \frac{1}{n-3}$

or

$$z = (\text{Arctanh}(r) - \text{Arctanh}(\rho_0))\sqrt{n-3}$$

$$\boxed{\text{Arctanh}\, x = \tfrac{1}{2} \ln \frac{1+x}{1-x}}$$

$$Z = \frac{W - \mu_w}{\sigma_w} \sim N(0,1)$$

اکنون فرض ها

$$\begin{cases} H_0: \rho = \rho_0 \\ H_1: \rho \neq \rho_0 \end{cases} \qquad \begin{cases} H_0: \rho = \rho_0 \\ H_1: \rho > \rho_0 \end{cases} \qquad \begin{cases} H_0: \rho = \rho_0 \\ H_1: \rho < \rho_0 \end{cases}$$

$$|Z| > Z_{\alpha/2} \Rightarrow R.H_0 \qquad Z > Z_{\alpha} \Rightarrow R.H_0 \qquad Z < -Z_{\alpha} \Rightarrow R.H_0$$

ρ : ۱

مثال: داده های زیر مربوط به مقاومت برحسب دقیقه ترانزیستورها و زمان شکست برحسب هم و زمان شکست ... با بار اضافی می باشد ضریب همبستگی بین X و Y را ...؟

مقاومت X	۴۳	۲۹	۴۴	۳۳	۳۳	۴۷	۳۴	۳۱	۴۸	۳۴	۴۶	۳۷
زمان شکست Y	۳۲	۲۰	۴۵	۳۵	۲۲	۴۶	۲۸	۲۶	۲۷	۴۴	۴۷	۳۰

$\sum x_i = ۴۵۹ \qquad \sum x_i^2 = ۱۸۰۷۵ \qquad \sum y_i = ۴۰۱ \qquad \sum y_i^2 = ۱۴۳۰۱$

$\sum x_i y_i = ۱۵۹۰۷$

$$S_{xx} = ۱۸۰۷۵ - \frac{۴۵۹^2}{۱۲} = ۵۱۸/۲۵$$

$$S_{xy} = ۱۵۹۰۷ - \frac{(۴۵۹)(۴۰۱)}{۱۲} = ۵۶۸/۷۵$$

$$S_{yy} = ۱۴۳۰۱ - \frac{(۴۰۱)^2}{۱۲} = ۹۰۰/۹۱۷$$

$$r = \frac{۵۶۸/۷۵}{\sqrt{۵۱۸/۲۵ \times ۹۰۰/۹۱۷}} = ۰/۸۳۲ \leftarrow$$

$$\rho : \left[\tanh\left(\text{arctanh}\, r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right), \tanh\left(\text{arctan}\, r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)\right]$$

فاصله اطمینان برای ρ

Subject:
Year. Month. Date. ( )

همچون مقدار r به یک نزدیک است پس رابطه خطی قوی بین X و Y برقرار است.

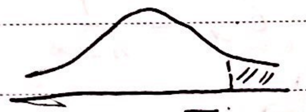
آزمون فرض $\begin{cases} H_0: \rho = 0/5 \\ H_1: \rho > 0/5 \end{cases}$ را آزمون کنید

$$w = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1/824}{0/168} = 1/195$$

$$\mu_{w_0} = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} = \frac{1}{2} \ln \frac{1/5}{0/5} = 0/549$$

$$\sigma_w = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{9}} = 0/333$$

$$Z_0 = \frac{w - \mu_{w_0}}{\sigma_w} = \frac{1/195 - 0/549}{0/333} = 1/94$$



$Z_\alpha = 1/64$

$Z_0 = 1/94 > 1/64 \Rightarrow RH_0$

بعد از نکته

تمرین: در جدول زیر نمرات امتحان پایان ترم درس روشهای آماری و رگرسیون 10 دانشجو که به طور تصادفی انتخاب شده اند درج گردیده است

روشهای آماری	12	13	10	13	15	11	12	16	17	14
رگرسیون	12	14	13	15	11	14	11	17	15	13

الف) ضریب همبستگی نمونه ای را بدست آورید و آن را تفسیر کنید

ب) فرضیه ناهمبسته بودن داده ها یعنی $\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$ را در سطح معنی داری 0.05 آزمون کنید

~~جهت یک فاصله اطمینان 95% برای ضریب همبستگی بدست آورید~~

نکته: برای آزمون فرضیه صفر بودن ضریب همبستگی

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

آماره مناسب $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ است

if $|t_0| > t_{n-2, \alpha/2} \Rightarrow RH_0$

PAPCO

Scanned with CamScanner

Subject: برای آزمون کردن دو فرض ① فرض در مورد توزیع جامعه ② فرض استقلال از استفاده می‌شود. مورد ترین

Year. Month. Date. ( )

آزمون نیکویی برازش : آزمون

تاکنون در یک آزمون فرض‌ها، فرض کردیم که جامعه دارای توزیع خاص مثل نرمال است و در مورد پارامترهای جامعه مثل μ و σ^2 آزمون‌هایی را انجام دادیم. در این بخش آزمونی مورد بررسی قرار می‌گیرد که عکس می‌کند آیا جمعیت دارای یک توزیع مشخص است یا نه. به عبارتی توزیع احتمال جمعیت خود مجهول است و فرم جمعیت توزیع مشخصی را است داده و آن را مورد آزمون قرار می‌دهیم. این هدف اصلی ما آزمون این مسئله است که آیا مدل ارائه شده در فرض صفر، برازنده داده‌ها است؟

در یک توزیعی را که حدس می‌زنیم داده‌ها از آن باشند، توزیع برازنده بر داده‌ها، این آزمون لازم را آزمون برازندگی یا نیکویی برازش گوییم. ~~یکی از آزمون‌های برازندگی~~ آزمون نیکویی برازش χ^2 پیرسون است که به شرح زیر می‌باشد. معروف‌ترین

فرض کنید که نتیجه یک آزمایش تصادفی به k طبقه دو به دو مجزای $C_1, C_2, \ldots, C_k$ متعلق باشد به طوری که احتمال متعلق بودن به طبقه C_i برابر مقدار $0<p_i<1$ باشد به طوری که $\sum_{i=1}^{k} p_i = 1$ اگر این آزمایش تصادفی n مرتبه به طور مستقل انجام می‌شود و تعریف می‌کنیم:

C_1	...	C_k
p_1	...	p_k
n_1	...	n_k

$O_i =$ تعداد دفعاتی از این آزمایش که نتیجه آزمایش به طبقه C_i متعلق است، $i=1,\ldots,k$

به عبارتی یعنی فراوانی مشاهده شده در هر طبقه Observed Frequency

در این صورت می‌دانیم $O_i \sim Bin(n, p_i)$ $\quad i=1,\ldots,k$

در n آزمایش با احتمال p_i، شیء در C_i قرار می‌گیرد.

پس انتظار داریم که در این n آزمایش به طور متوسط np_i بار نتیجه آزمایش را در C_i مشاهده کنیم. عدد np_i را با e_i نشان می‌دهیم و آن را Expected Frequency می‌نامیم. است درنهایت:

$$\underbrace{\sum_{i=1}^{k} O_i}_{\text{مشاهده کردیم}} = \underbrace{\sum_{i=1}^{k} e_i}_{\text{انتظار داریم}} = n$$

H_0: آزمایش تصادفی با k دارای مدل احتمال بصورت ← است.

مجموع	C_1	...	C_k
احتمال	p_1		p_k

$H_0: p_1 = p_{10}, \ldots, p_k = p_{k0}$ یا به عبارتی می توان گفت

که $p_{10}, \ldots, p_{k0}$ مقادیر عددی ارائه شده اند به طوری که $\underbrace{p_{10} + \cdots + p_{k0}}= 1$

$(H_0: p_1 = p_{10} \ldots p_k = p_{k0})$

برای انجام آزمون اگر فراوانیهای مشاهده شده (o_i) و فراوانیهای مورد انتظار (e_i) بیش از حد تفاوت داشته باشند نسبت به فرض H_0 بدگمان می شویم یعنی مدل احتمال ارائه شده در فرض صفر صحیح نمی باشد، به عبارتی اگر

$$\sum_{i=1}^{n} (o_i - e_i)^2$$

بیش از حد بزرگ شد فرض H_0 رد می شود. اما چون عبارت فوق توزیع خاصی ندارد

پیرسن آماره

$$\chi^2_0 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \leftarrow \text{آماره } \chi^2 \text{ پیرسن}$$

را معرفی کرد و نشان داد اگر n بزرگ باشد این آماره دارای توزیع χ^2_{k-1} تحت H_0 است. (در رابطه بین فراوانیها)

$$\text{if } \chi^2_0 > \chi^2_{k-1,\alpha} \Rightarrow R H_0$$

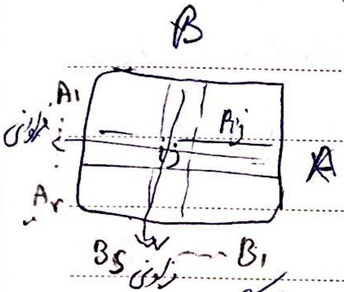
نکته ۱: معمولاً تعداد طبقات یعنی k طوری انتخاب می شوند که فراوانی داده ها در هر طبقه حداقل ۵ باشد. اگر طبقاتی دارای فراوانی کمتر از ۵ باشد باید در طبقات دیگر ادغام شوند.

نکته ۲: به ازای هر برآورد پارامتری که در مسئله انجام می دهیم از درجه آزادی χ^2 کاسته می شود.

نکته ۳: از آزمون χ^2 برای آزمون برازش استقلال دو متغیر تصادفی X و Y نیز استفاده می شود. به طور مثال فرض کنید دو ویژگی قد و وزن ($\underset{A}{\text{قد}}$، $\underset{B}{\text{وزن}}$) از جمعیتی مد نظر است بطوریکه A در ۳ طبقه و ویژگی B در ۵ طبقه بررسی گردد.

Subject :

Year . Month . Date . ( )



پس اگر دو ویژگی A و B بخواهند مستقل باشند احتمال توأم برای طبقه C_{ij} :

$$p_{ij}=P(C_{ij})=P(A_i)P(B_j)\qquad i=1,\dots,r \quad j=1,\dots,s$$

و برای اینکه این احتمال را تحت فرض H_0 برآورد کنیم نمونهٔ اتفاقی n تایی در نظر بگیرید

$$\hat{p}_{ij}=\underbrace{\frac{\text{فراوانی } A_i}{n}}_{\hat{p}(A_i)}\times\underbrace{\frac{\text{فراوانی } B_j}{n}}_{\hat{p}(B_j)}$$

از طرفی تحت H_0 ، فراوانی مورد انتظار برای C_{ij} : $e_{ij}=n\times\hat{p}_{ij}=\frac{A_iB_j}{n}$ می‌شود

و اگر O_{ij} فراوانی مشاهده شده برای C_{ij} باشد آماره χ^2 به صورت زیر تعریف می‌شود

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

تعداد پارامترهایی که برآورد شده‌اند

if $\chi^2>\chi^2_{(r-1)(s-1),\alpha}\Rightarrow RH_0$

نکته ۳: به ازای هر برآورد پارامتری که ... یک درجه آزادی از آماره χ^2 کاسته می‌شود. χ^2_{k-t}

مثال ۱) یک تاس را ۱۲۰ بار می‌اندازیم و مشاهدات به صورت زیر بدست آمده است، در سطح $\alpha=0.05$ آزمون کنید آیا تاس سالم است.

$$\begin{cases}H_0: p_1=\dots=p_6=\frac{1}{6}\\ H_1: p_i\neq\frac{1}{6}\quad i=1,\dots,6\end{cases}$$

شماره تاس	1	2	3	4	5	6	جمع
فراوانی مشاهده شده O_i	18	23	16	21	18	24	120
فراوانی مورد انتظار	۲۰	۲۰	۲۰	۲۰	۲۰	۲۰	۱۲۰

$$\chi^2=\sum_{i=1}^{6}\frac{(O_i-e_i)^2}{e_i}=\frac{(18-20)^2}{20}+\dots+\frac{(24-20)^2}{20}=2.5$$

$$\chi^2_{\underset{k-1}{5},\,0.05}=11.1 \qquad 2.5<11.1\Rightarrow AH_0$$

پس تاس سالم است

مثال ۲ : تعداد غلط‌های چاپی در ۱۰۰ صفحه یک کتاب را شمرده‌ایم و مشاهدات در جدول زیر آورده شده است.
آیا توزیع پواسن در سطح معنی‌داری ۰/۰۵ برای داده‌ها برازنده است.

i تعداد غلط‌ها	۰	۱	۲	۳	۴	۵	۶
O_i تعداد صفحات	۳۶	۴۰	۱۹	۲	۰	۲	۱

$$\begin{cases} H_0: X \sim P(\mu) \\ H_1: X \nsim P(\mu) \end{cases}$$

μ را با میانگین داده‌ها $\hat{\mu}$ برآورد کنیم
μ میانگین X است

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{K} f_i x_i = \frac{1}{100}(0\times 36 + \dots + 6\times 1) = 1$$

تحت H_0:

$$\underset{P(X=x)}{f_X(x)} = \frac{e^{-1}\,1^x}{x!} = \frac{1}{x!}e^{-1} \qquad x = 0,1,\dots$$

پس:

$$\hat{P}(X=0) = e^{-1} \quad \dots \quad \hat{P}(X=6) = 0/0005$$

i	۰	۱	۲	۳	۴	۵	۶	
O_i	۳۶	۴۰	۱۹	۲	۰	۲	۱	۱۰۰
e_i	۳۶٫۷۹	۳۶٫۷۹	۱۸٫۳۹	۶٫۱۳	۱٫۵۳	۰٫۳۱	۰٫۰۵	۱۰۰

$\underbrace{36/79}_{n\hat{p}_i = 100\times e^{-1}}$

< 5

i	۰	۱	۲	> 2	جمع
O_i	۳۶	۴۰	۱۹	۵	۱۰۰
e_i	۳۶٫۷۹	۳۶٫۷۹	۱۸٫۳۹	۸٫۰۲	۱۰۰

$$\chi^2 = 1/654 \qquad k-1-1 = 4-1-1 = 2 \qquad \chi^2_{0/05} = 5/99$$

$$\Rightarrow AH_0$$

مثال ۳) جدول زیر شماره تصادف‌ها برای ۵۰ هفته است

شماره تصادف‌ها در هفته	۰	۱	۲	≥ 3
فراوانی هفته‌ها	۳۲	۱۲	۶	۰

فرض کنید Y شماره تصادف‌ها در هر هفته باشد. با سطح ۵ درصد بیازمائید Y دارای توزیع پواسن است.

مثال ۴) در درس آمار و احتمال جدول فراوانی نمره‌ها چنین است. در سطح ۱ درصد بیازمائید، توزیع نمره‌ها یکنواخت است. $\alpha = 0/01$

نمره	الف	ب	ج	د	ه
فراوانی	۴۰	۱۸	۳۲	۲۰	۱۶
مورد انتظار	۲۰	۲۰	۲۰	۲۰	۲۰

ubject:
Year. Month. Date. ( )

با استفاده از جدول فراوانی زیر، مستقل بودن مصرف سیگار و عصبی بودن را با سطح ۵ درصد بیازمایید

	A_1 سیگاری	A_2 غیرسیگاری	جمع
عصبی	۲۱	A_{12} ۳۶	۵۷
غیرعصبی	۴۸	۲۵	۷۳
جمع	۶۹	۶۱	۱۳۰

$n \times A_i$ فراوانی

$$e_{11} = n \times \hat{p}_{11} = \frac{69}{130} \times \frac{57}{130} \times 130 \qquad \hat{p}_{12} = \frac{61}{130} \times \frac{57}{130} \times 130 \qquad \hat{p}_{21} = \frac{69}{130} \times \frac{73}{130} \times 130$$

$= 0.23$ ۱۶۹۰۰ $= 0.21$ $= 0.3$

$$\hat{p}_{22} = \frac{61}{130} \times \frac{73}{130} \times 130 = 0.26$$

~~۲۶٫۷۴~~ ~~۳۴٫۲۵~~

جدول e_i

۳۰٫۲۵	۲۶٫۷۴
۳۸٫۷۴	۳۴٫۲۵

$$\chi^2 = 10.76 \not< \chi^2_{1, 0.05} \Rightarrow RH_0$$

مقایسه چندین جامعه - (تیمار شده)، - طرح تصادفی ساده :

مقایسه چندین تیمار (جامعه)

تاکنون در مورد مقایسه میانگین دو جامعه بحث کردیم. اکنون فرض کنید هدف مقایسه چندین میانگین است. برای مقایسه چندین تیمار از تحلیل واریانس

Analysis of variance (ANOVA)

استفاده می‌شود در این درس طرح تک عاملی کاملاً تصادفی شده مطرح می‌شود

طرح تک عاملی کاملاً تصادفی شده:

فرض کنید یک کشاورز می‌خواهد میزان بازدهی محصولات خود را تحت تاثیر چند نوع کود مقایسه کند یعنی هدف او مقایسه چندین کود است برای مثال ۵ نوع کود مد نظر کشاورز می‌باشد برای این کار ابتدا زمین خود را به ۲۵ قطعه تقسیم می‌کند

۳	۴	۵	۳	۲
۲	۳	۱	۳	۱
۴	۳	۴	۲	۲
۵	۱	۱	۵	۲
۴	۵	۱	۴	۵

و بطور کاملاً تصادفی این ۵ نوع کود در کرت‌ها (قطعات زمین) تقسیم می‌شود. مثلاً بین ۱-۲۵ عددی انتخاب و اولین نوع کود (اولین تیمار) از آن کرت استفاده می‌کند و همین ترتیب ۱-۲۵ انتخاب و دومین تیمار استفاده می‌شود ... سپس میزان محصول برداشتی از هر کرت را اندازه‌گیری می‌کنند و جدول زیر تشکیل می‌شود

تیمار	تکرار ۱	۲	۳	۴	۵	مجموع $y_{i.}$	میانگین $\bar{y}_{i.}$
۱	۷	۷	۱۵	۱۱	۹	۴۹	۹/۸
۲	۱۲	۱۷	۱۲	۱۸	۱۸	۷۷	۱۵/۴
۳	۱۴	۱۸	۱۸	۱۹	۱۹	۸۸	۱۷/۶
۴	۱۹	۲۵	۲۲	۱۹	۲۳	۱۰۷	۲۱/۶
۵	۷	۱۰	۱۱	۱۵	۱۱	۵۴	۱۰/۸
						مجموع کل ۳۷۶	۱۵/۰۴

Subject:
Year. Month. Date. ( )

پس به طور کلی سه حالت را داریم: برای یک طرح کاملاً تصادفی شده با a تیمار به صورت زیر می‌باشد

تیمار	مشاهدات	مجموع	متوسط
۱	$y_{11}, y_{12}, \ldots\ldots y_{1n_1}$	$y_{1\cdot}=\sum_{j=1}^{n_1} y_{1j}$	$\bar{y}_1 \to y_{1\cdot}/n_1$
۲	$y_{21}, y_{22}, \ldots y_{2n_2}$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$y_{a1}, y_{a2}, \ldots y_{an_a}$	$y_{a\cdot}=\sum y_{aj}$	$\bar{y}_a \to y_{a\cdot}/n_a$

$$y_{i\cdot}=\sum_{j=1}^{n_i} y_{ij}$$

$$y_{\cdot\cdot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}$$

$\bar{y}$ کل با $y_{\cdot\cdot}$ هم نمایش می‌دهند

$$\bar{y}=\frac{\text{حاصل جمع همه مشاهدات}}{n_1+n_2+\cdots+n_a}=\frac{y_{\cdot\cdot}}{\underbrace{\sum_{i=1}^{a} n_i}_{n}}$$

$$=\frac{n_1\bar{y}_1+\cdots+n_a\bar{y}_a}{n_1+\cdots+n_a}$$

اگر فرض کنید μ_i میانگین متغیر پاسخ در جامعه به ازای تیمار i و ε_{ij} را انحراف پاسخ از مقدار میانگین یا خطا فرض کنیم در این صورت مدل آماری به صورت زیر تعریف می‌شود

$$① \quad y_{ij}=\mu_i+\varepsilon_{ij} \qquad i=1,\ldots,a \quad j=1,\ldots,n_i$$

و با در نظر گرفتن μ به عنوان میانگین کل $\mu=\frac{1}{a}\sum_{i=1}^{a}\mu_i$ می‌توان اثر iامین تیمار را به صورت زیر تعریف کرد:

$$② \quad \tau_i=\mu_i-\mu \Leftarrow \text{اثر iامین تیمار}$$

نمی‌توان تعیین کرد

در نهایت مدل به صورت:

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij}$$

خواهد شد.

هدف آزمون:

$$\begin{cases} H_0: \mu_1=\cdots=\mu_a \\ H_1: \mu_i\neq\mu_j \end{cases}$$

حداقل برای یک جفت i و j

$\mu_1-\mu=\mu_2-\mu$

از ② معادل است با آزمون:

$$\star\begin{cases} H_0: \tau_1=\cdots=\tau_a=0 \\ H_1: \tau_i\neq 0 \end{cases}$$

حداقل برای یک i

چون μ_i ها و μ مجهولند در عمل داریم :

$$Y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_i)$$

در نهایت برای آزمون کردن فرض * باید جدول زیر را تشکیل دهید :

منبع تغییر	SS مجموع مربعات	d.f	MS میانگین مربع	F_0
تیمار	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$F_0 = \frac{MSA}{MSE}$
خطا	SSE	$n-a$ $\sum_{i=1}(n_i-1)$	$MSE = \frac{SSE}{n-a}$	$\sim F_{a-1,n-a,\alpha}$
کل	SST	$n-1$		

$$\text{if} \quad F_0 > F_{a-1,n-a,\alpha} \Rightarrow RH_0$$

که در جدول :

$$SSA = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{a} \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n}$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{a} \frac{Y_{i.}^2}{n}$$

$$= (n_1-1)s_1^2 + \cdots + (n_a-1)s_a^2$$

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{..}^2}{n}$$

و داریم :

$$SST = SSE + SSA$$

مثال : در آزمایشی برای مطالعه رشد یک گیاه در اثر میزان فسفر موجود در خاک پنج سطح مختلف فسفر به میزان ۰، ۲۵، ۵۰، ۷۵ و ۱۰۰ سنت در میلیون را در نظر گرفته سپس خاکی مناسب را اختیار کرده و بعد از مخلوط کردن آن خاک با فسفر به نسبتهای فوق، مخلوط را به مقدار مساوی درون گلدانهای ریخته و گیاه مورد نظر را در آن خاک رویش داده ایم بعد از گذشت سه ماه از رویش گیاه، گلدانها را دقیقا وزن و مشاهدات تبدیل یافته را به صورت جدول زیر بدست آورده ایم

Subject:

Year. Month. Date. ( )

تیمار (مقدار فسفر)	مشاهدات				مجموع	میانگین
۰	۱۲	۱۰	۱۱	۱۴	$y_{1.}$ ۴۷	۱۱,۷۵
۲۵	۲۱	۱۹	۲۰	۲۳	$y_{2.}$ ۸۳	۲۰,۷۵
۵۰	۲۲	۲۵	۲۴	۲۵	۹۶	۲۴
۷۵	۲۷	۲۶	۲۵	۲۷	۱۰۵	۲۶,۲۵
۱۰۰	۲۹	۲۶	۲۵	۲۷	۱۰۷	۲۶,۷۵
					۴۳۸ $y_{..}$	$\bar{y} = \frac{۴۳۸}{۲۰}$

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{n}$$

$$= (۱۲^۲ + ۱۰^۲ + \dots + ۲۷^۲) - \frac{۴۳۸^۲}{۲۰} = ۶۳۹,۸$$

$$SSA = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n} = \frac{۴۷^۲ + ۸۳^۲ + \dots + ۱۰۷^۲}{۴} - \frac{۴۳۸^۲}{۲۰} = ۶۰۴,۸$$

$$SSE = SST - SSA = ۳۵$$

منبع	SS	df	MS
تیمار (درصد فسفر)	۶۰۴,۸	۵−۱=۴	$\frac{۶۰۴,۸}{۴} = ۱۵۱,۲$
خطا	۳۵	۱۵	۳۵/۱۵ = ۲,۳۳
کل	۶۳۹,۸	۲۰−۱=۱۹	

$$F_0 = \frac{۱۵۱,۲}{۲,۳۳} = ۶۴,۸۹$$

$$F_{0/01, ۴, ۱۵} = ۴,۸۹ \qquad F_0 > F_\alpha \Rightarrow RH_0$$

یعنی با اطمینان ۹۵٪ وزن ساقه گیاهچه تحت تاثیر مقدار فسفر موجود در خاک است.

مثال ۲ : آزمون فرض $H_0: \mu_{B_1}=\mu_{B_2}=\mu_{B_3}=\mu_{B_4}$ را در سطح ٪۰٫۵ برای داده‌های زیر انجام دهید
تیمار

		مجموع	میانگین
B_1	۱۰، ۱۵، ۸ ۱۲ ۱۵	$y_{1.}=۶۰$	$\bar{y}_1=۱۲$
$B_۲$	۱۴ ۱۸ ۲۱ ۱۵	$y_{2.}=۶۸$	$\bar{y}_2=۱۷$
$B_۳$	۱۷ ۱۶ ۱۴ ۱۵ ۱۷ ۱۵ ۱۸	$y_{3.}=۱۱۲$	$\bar{y}_3=۱۶$
B_4	۱۴ ۱۵ ۱۷ ۱۵ ۱۶ ۱۵	$y_{4.}=۹۰$	$\bar{y}_4=۱۵$
		۳۳۰	

$$\bar{y} = \frac{۱۰+۱۵+\dots+۱۵}{۲۲} = ۱۵$$

$$SST_{\text{کل}} = \sum_{i=1}^{a=۴}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{۲۲} = ۱۰^۲+\dots+۱۵^۲ - \frac{۳۳۰^۲}{۲۲} = ۱۶۲$$

$$SSA_{\text{تیمار}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{۲۲} = \underbrace{\frac{۶۰^۲}{۵}}_{۷۲۰} + \underbrace{\frac{۶۸^۲}{۴}}_{۱۱۵۶} + \underbrace{\frac{۱۱۲^۲}{۷}}_{۱۷۹۲} + \underbrace{\frac{۹۰^۲}{۶}}_{۱۳۵۰} - \underbrace{\frac{۳۳۰^۲}{۲۲}}_{۴۹۵۰} = ۶۸$$

$$SSE = SST - SSA = ۹۴$$

منبع تغییرات	SS	df	MS	F_0
تیمار	۶۸	۳	$\frac{۶۸}{۳}=۲۲٫۶۷$	۴٫۳۴
خطا	۹۴	۱۸	$\frac{۹۴}{۱۸}=۵٫۲۲$	
کل	۱۶۲	۲۱		

$$F_{۰٫۰۵,\,۳,\,۱۸} = ۳٫۱۶ \qquad F_0 > F_\alpha$$

$\Rightarrow RH_0$

پس تفاوت معنی‌داری بین چهار تیمار وجود دارد

مثال: فرض کنید $X \sim N(\mu, 4)$ و برای آزمون

$$\begin{cases} H_0: \mu = 0 \\ H_1: \mu = 1 \end{cases}$$

یک نمونه تصادفی ۲۵تایی از X در نظر گرفته‌ایم. اگر $\bar{X} \geq 0/4$ ناحیه بحرانی باشد (ناحیه رد فرض صفر) خطاها و توان آزمون را محاسبه کنید.

$$\alpha = P(RH_0 \mid H_0 \text{ درست}) = P_{\mu=0}(\bar{X} > 0/4)$$

$$= P\left(Z > \frac{0/4 - 0}{\frac{2}{5}}\right) = P(Z > 1) = 1 - \underbrace{\Phi(Z < 1)}_{0/8413}$$

$$= 0/1587$$

$$\beta = P(RH_1 \mid H_1 \text{ درست}) = P_{\mu=1}(\bar{X} \leq 0/4)$$

احتمال خطای نوع دوم

پذیرفتن H_0 به شرط H_0 غلط

$AH_0 \mid H_0$ غلط

$RH_1 \mid H_1$ درست

$$\Rightarrow P\left(Z \leq \frac{0/4 - 1}{2/5}\right) = P(Z \leq -1/5)$$

$$= 0/0668$$

$$\text{توان آزمون} = 1 - \beta = 0/9332$$

(۲) در ظرف سربسته‌ای هفت مهره می‌باشد که θ از آنها سفید و بقیه سیاه هستند. می‌خواهیم آزمون زیر را انجام دهیم.

$$\begin{cases} H_0: \theta = 2 \\ H_1: \theta = 4 \end{cases}$$

شخصی دو مهره با هم از ظرف بیرون می‌آورد. اگر هر دو سفید باشد H_0 را رد می‌کند. خطای نوع اول و دوم و توان را حساب کنید.

$$\theta = \text{تعداد مهره‌های سفید} \qquad 7 - \theta = \text{تعداد مهره‌های سیاه}$$

$$\alpha = P(RH_0 \mid H_0 \text{ درست}) = P(\text{هر دو سفید} \mid \theta = 2) = \frac{\binom{2}{2}}{\binom{7}{2}} = \frac{1}{21}$$

$$\beta = P(RH_1 \mid \overline{H_1} \text{ درست}) = P(\text{حداقل یک سیاه} \mid \theta = 4)$$

$$= \frac{\binom{4}{1}\binom{3}{1} + \binom{4}{0}\binom{3}{2}}{\binom{7}{2}} = \frac{5}{7}$$

$$\pi \text{ توان} = 1 - \beta = \frac{2}{7}$$

۲. فرض کنید $X \sim N(\mu, 4)$ و یک نمونه ۱۰۰تایی در اختیار داریم به طوریکه $\bar{x} = 4/2$

آزمون زیر را با روش p-مقدار انجام دهیم

در سطح ۰/۱ و ۰/۰۵

$$\begin{cases} H_0: \mu = 3/9 \\ H_1: \mu > 3/9 \end{cases}$$

$$P(\underbrace{\bar{X} \geq 4/2}_{RH_0} \mid \underbrace{\mu = 3/9}_{H_0 \text{ درست}}) = P(Z \geq 1/5)$$

$$= 1 - P(Z \leq 1/5)$$

$$= P(Z \leq -1/5)$$

$$= 0/0668$$

مقدار عددی آماره ~~...~~ بزرگتر

مقدار با شرط محقق شده به شرط H_0 درست باشد

(یا ناامیدی بزرگی)

$0/0668 \cancel{<} 0/05$ برای $\alpha = 0/05$، H_0 رد نمی‌شود

$0/0668 < 0/1$ برای $\alpha = 0/1$، H_0 رد می‌شود

تمرین قبل را داریم

دقت کنید برای آزمون‌های دوطرفه:

$$\begin{cases} H_0: \mu = 3/9 \\ H_1: \mu \neq 3/9 \end{cases}$$

$$\text{مقدار-}p = 2 \min\{P(\bar{X} \geq 4/2 \mid \mu = 3/9),$$

$$P(\bar{X} \leq 4/2 \mid \mu = 3/9)\} = 2\min\{P(Z \geq 1/5),$$

$$P(Z \leq 1/5)\} = 2 \times 0/0668$$

$$= 0/1336$$

برای هر α، AH_0

۴- مقدار کوچکترین مقداری که برای α می‌توان در نظر گرفت تا H_0 رد شود.

و به طور کلی فرض کنید T آماره آزمون تحت H_0 و t مقدار آن باشد به طور کلی داریم:

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

$p\text{-value} = p(T_0 \geq t)$, $p.v = p(T_0 \leq t)$, $p\text{-val} = 2\min\{p(T_0 \leq t), p(T_0 \geq t)\}$

T_0 آماره تحت H_0 یعنی H_0 درست باشد

سؤالها پایان ترم:

سؤال طرح: سه نوع باطری تحت مطالعه اند. گمان می‌رود که طول عمر باطریها متفاوت باشد نتایج آزمون پنج باطری از هر نوع به صورت زیر است

طول عمر بر حسب هفته

نوع ۱	نوع ۲	نوع ۳
۱۰۰	۷۶	۱۰۸
۹۶	۸۰	۱۰۰
۹۲	۷۵	۹۶
۹۶	۸۴	۹۸
۹۲	۸۲	۱۰۰

آیا این باطریها طول عمر متفاوت دارند؟