Subject: حسابداری - آمار ۲
Date: ۹۸/۷/۱۸
جلسه اول

Statistics

- آمار: مجموعه‌ای از مفاهیم و روش‌ها است که در زمینه‌های پژوهشی برای گردآوری اطلاعات، تجزیه و تحلیل آنها و نتیجه‌گیری معتبر در مورد جامعه آماری هنگامی که عدم قطعیت وجود دارد می‌باشد.

- جامعه آماری: مجموعه‌ای از افراد یا اشیاء که در یک زمان و مکان حداقل در یک صفت (متغیر) مشترک باشند. برای مثال: کارکنان دانشگاه آزاد اسلامی آباران در سال ۱۳۹۸ تشکیل یک جامعه آماری می‌دهند. متغیرهایی که ممکن است مطالعه کنیم شامل: سن، جنسیت، رضایت شغلی، میزان حقوق، مدرک تحصیلی ...

برای مطالعه یک یا چند صفت در جامعه آماری ۲ راه وجود دارد: الف) سرشماری: اگر همه اعضای جامعه آماری را براساس صفت‌های مورد مطالعه بررسی کنیم سرشماری است. برای مثال: اگر همه کارکنان دانشگاه آزاد آباران را براساس رضایت شغلی مطالعه کنیم سرشماری است.

سرشماری معایبی دارد از جمله: وقت‌گیر است، مشکلات اجرایی و اقتصادی دارد و گاهی اوقات ممکن است باعث از بین رفتن جامعه آماری شود. ب) نمونه‌گیری: اگر تعدادی از اعضای جامعه آماری را به روشی مناسب انتخاب کنیم (تصادفی) و به جای کل جامعه آماری این تعداد را بررسی نماییم، نمونه‌گیری می‌شود. نمونه انتخاب شده از جامعه آماری باید

Subject
Date

نمونه‌ای باشد که ویژگی‌های کلی جامعه را در خود داشته باشد.

- شیوه‌های مهم نمونه‌گیری: ۱- نمونه‌گیری تصادفی ساده: اگر همه اعضای جامعه آماری شانس مساوی برای انتخاب شدن در نمونه را داشته باشند نمونه‌گیری را تصادفی ساده می‌گوییم که می‌تواند با جایگذاری یا بدون جایگذاری فرد انتخاب شده به جامعه آماری باشد.

مثلاً

اندازه جامعه $N = ۴۰$

اندازه نمونه $n = ۴$

ماشین حساب: $\boxed{\text{Shift}}$ $\boxed{\text{Random}}$ $\xrightarrow{\text{عدد تصادفی بین ۰ و ۱}}$ $۰/۳۸۷ \xrightarrow{\times N} ۱۵/۴۸$

$۱۵ + ۱ = ۱۶$

در کانال اینستاگرام
فیلم آن گذاشته شده
است

* فقط مثالی برای درک شکل نمونه‌ی تصادفی است.

۲- نمونه‌گیری سیستماتیک (منظم): هنگامی که جامعه آماری براساس یک لیست نظیر شماره‌های دانشجویی، پلاک منازل لیست شده باشد از این شیوه نمونه‌گیری می‌توان استفاده کرد.

بدین صورت که ابتدا دوره‌ی نمونه‌گیری را مشخص می‌کنیم، سپس

دوره نمونه‌گیری $K = \frac{N}{n}$

از بین K عضو اول جامعه یکی را بطور تصادفی به عنوان مبدأ نمونه‌گیری انتخاب می‌کنیم.

$N = ۴۰$

$n = ۴$

$K = \frac{N}{n} = \frac{۴۰}{۴} = ۱۰$

۵ و ۱۵ و ۲۵ و ۳۵

مبدأ نمونه‌گیری $= ۵$
(بطور تصادفی)

۳- نمونه‌گیری طبقه‌ای: هنگامی که جامعه آماری براساس صفت مورد مطالعه همگون نباشد

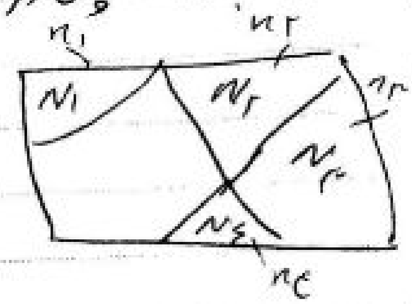
برای افزایش دقت نمونه‌گیری از این شیوه استفاده می‌شود به این صورت که جامعه

Subject
Date

(تفکیک کردن)

جامعه را به تعدادی طبقه (افراز) می کنیم که عناصر درون این طبقات از لحاظ صفت مورد مطالعه همگن باشد. سپس از هر طبقه یک نمونه تصادفی می گیریم

K تعداد طبقات
N اندازه جامعه
n اندازه نمونه
N_i اندازه طبقه i ام
n_i اندازه نمونه i ام



$$\frac{N_i}{N} = \frac{n_i}{n}$$

$$\boxed{n_i = n \times \frac{N_i}{N}}$$

$$N = N_1 + N_2 + \cdots N_K$$

$$n = n_1 + n_2 + \cdots n_k$$

*) مثال: می خواهیم یک نمونه ی تصادفی ۴۰ تایی به روش طبقه ای از کارکنان دانشگاه بگیریم. فرض بر این است که رضایت شغلی بر حسب جنسیت همگن نیست.

$N = 280$

$n = 40$

$$n_i = n \times \frac{N_i}{N}$$

جنسیت	تعداد
مذکر	۱۶۰
مؤنث	۱۲۰

$$n_1 = 40 \times \frac{160}{280} = 22.86 \simeq 23$$

$$n_2 = 40 \times \frac{120}{280} = 17.14 \simeq 17$$

$\Big] 40$

۴- نمونه گیری خوشه ای: هنگامی که جامعه آماری خیلی وسیع یا گسترده باشد برای کاهش مشکلات اجرایی و اقتصادی از این شیوه ی نمونه گیری استفاده می شود. به این صورت که اعضای مجاور در جامعه آماری را به عنوان خوشه ها در نظر می گیریم. سپس تعدادی از خوشه ها را بطور تصادفی انتخاب می کنیم و همه اعضای درون این خوشه ها را بررسی می کنیم.

Subject
Date

- پارامتر: ویژگی مجهول جامعه آماری که گاهی اوقات از طریق سرشماری بدست می‌آید پارامتر گفته می‌شود. در این درس پارامترها را با حروف یونانی نشان می‌دهیم

برای مثال: پارامتر میانگین که نشان دهنده متوسط متغیر مورد مطالعه است در جامعه با پارامتر μ نشان داده می‌شود.

- آماره: ویژگی متناظر با پارامتر جامعه در نمونه را آماره می‌گوییم. آماره تابعی از نمونه تصادفی است بنابراین متغیر بوده و دارای توزیع احتمال است که به آن توزیع نمونه‌گیری آماره گفته می‌شود. در این درس آماره‌ها را با حروف لاتین نشان می‌دهیم.

برای مثال میانگین نمونه‌ای $\bar{X}$ نشان داده می‌شود.

توجه داریم که در آماره پارامتر مجهول وجود ندارد اما توزیع نمونه‌گیری آن وابسته به پارامتر جامعه است.

- برآوردگر نقطه‌ای: از آماره‌ها برای تخمین یا برآورد پارامتر مجهول جامعه استفاده می‌شود. به همین دلیل به آن‌ها برآوردگر گفته می‌شود.

برای مثال ← آماره‌ی $\bar{X}$ میانگین نمونه‌ای برای برآورد میانگین جامعه μ استفاده می‌شود.

البته برای برآورد پارامتر θ برآوردگرهای ... و $\hat{\theta}_1$ ، $\hat{\theta}_2$ ، $\hat{\theta}_3$ ممکن است وجود داشته باشد که از میان آنها باید یکی را انتخاب کنیم. به این منظور خواص زیر را در نظر می‌گیریم.

$E(\hat{\theta})=\theta$

الف - نااریبی: برآوردگر $\hat{\theta}_1$ را برای برآورد پارامتر θ نااریب می‌گوئیم هرگاه مقدار متوسط توزیع نمونه‌گیری آن برابر θ باشد.

برای برآورد یک پارامتر ممکن است برآوردگرهای نااریب متفاوتی وجود داشته باشد.

برای مثال ← در نمونه‌گیری از توزیع نرمال میانگین و میانه‌ی نمونه‌ای برای برآورد میانگین μ نااریب هستند.

ب - کارایی: فرض کنید $\hat{\theta}_1$ و $\hat{\theta}_2$ برای برآورد پارامتر θ نااریب باشند. در این صورت $\hat{\theta}_1$ را کاراتر از $\hat{\theta}_2$ برای برآورد پارامتر θ می‌گوییم. هرگاه $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

- برآوردگرهای نااریب μ و σ^2: فرض کنید x_1 و ... و x_N اطلاعات حاصل از سرشماری جامعه باشند، در این صورت میانگین و واریانس جامعه به صورت زیر محاسبه می‌گردد:

Subject
Date

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$$

حال فرض کنید $x_1, \dots, x_n$ اطلاعات حاصل از نمونه باشد. درانصورت میانگین و واریانس نمونه ای بصورت زیر هستند که برآوردگرهای نااریب μ و σ^2 هستند.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1} \text{ یا } S^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

*) مثال: رضایت شغلی یک نمونه ی تصادفی ۱۰ تایی از کارکنان اداره ی α (آلفا) بصورت زیر است. برآورد نقطه ای پارامترهای میانگین و واریانس جامعه را بدست آورید.

$n=10$

x_i	۹	۱۲	۱۳	۸	۷	۶	۵	۹	۱۲	۱۵
x_i^2	۸۱	۱۴۴	۱۶۹	۶۴	۴۹	۳۶	۲۵	۸۱	۱۴۴	۲۲۵

$\sum x_i = 96$

$\sum x_i^2 = 1018$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{96}{10} = 9/6$$

$\hat{\mu} = 9/6$

$\hat{\sigma}^2 = 10/71$

$$S^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \Rightarrow S^2 = \frac{1018 - 10(9/6)^2}{9} = 10/71$$

- برآوردگر نااریب نسبت جامعه: فرض کنید جامعه ی آماری براساس متغیر مورد مطالعه به ۲ طبقه ی موفقیت و شکست طبقه بندی شده باشد.
فرض کنید $x_1, \dots, x_N$ اطلاعات حاصل از سرشماری جامعه باشد. درانصورت نسبت در

Subject
Date

جامعه (π) بصورت زیر می باشد.

$$\pi = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{X}{N}$$

$x_i = \begin{cases} 1 & \text{موفقیت} \\ 0 & \text{شکست} \end{cases}$ ، X: تعداد موفقیت ها در جامعه

همچنین اگر $x_1, \ldots, x_n$ اطلاعات حاصل از نمونه باشند. در اینصورت نسبت نمونه ای که برآوردگر نا اریب (π) می باشد بصورت زیر است.

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n}$$

x: تعداد موفقیت ها در نمونه

مثال: در یک نمونه ی تصادفی ۱۲۰ تایی از دانشجویان دانشگاه الفا ۵۰ نفر آنها شاغل بوده اند. اگر تعداد دانشجویان این دانشگاه ۷۰۰۰ نفر باشد، برآورد نقطه ای تعداد دانشجویان شاغل دانشگاه را بدست آورید.

$n = 120$ $\quad p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n}$

$x = 50$

$N = 7000$ $\quad p = \frac{x}{n} = \frac{50}{120} = 0/417$

انتظار می رود ۴۱/۷٪ دانشجویان شاغل باشند

$\hat{\pi} = 0/417 \quad \pi = \frac{X}{N} \rightarrow X = N\pi$

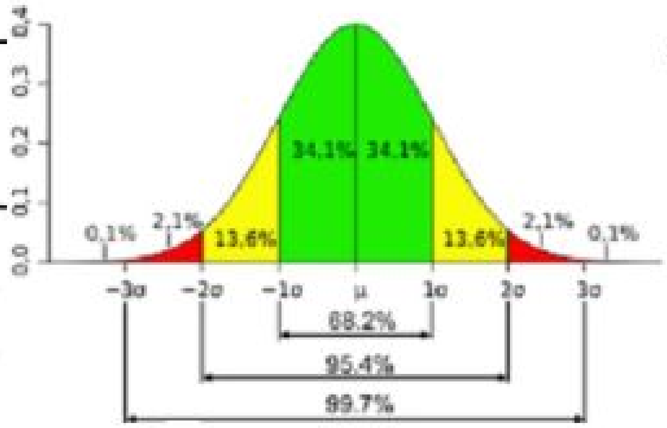
$\hat{X} = N\hat{\pi} = 7000 \times 0/417 = 2919$ انتظار می رود از ۷۰۰۰ نفر ۲۹۱۹ نفر شاغل باشند

- توزیع های حجم نمونه گیری: ۱- توزیع نرمال: متغیر نرمال یک متغیر پیوسته می باشد که هر مقدار حقیقی را می تواند داشته باشد. دارای ۲ پارامتر می باشد که به ترتیب میانگین توزیع μ و واریانس توزیع σ^2 است. استاندارد شده ی متغیر نرمال دارای توزیع نرمال است.

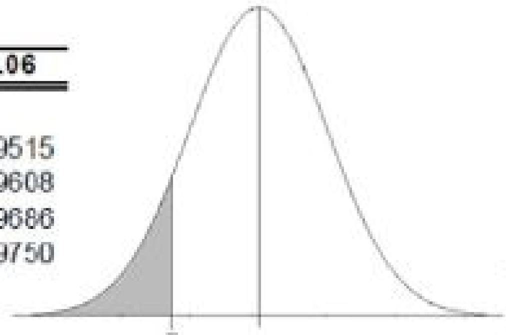
Subject
Date

استاندارد نرمال $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$\mu_Z = 0$, $\sigma^2_Z = 1$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

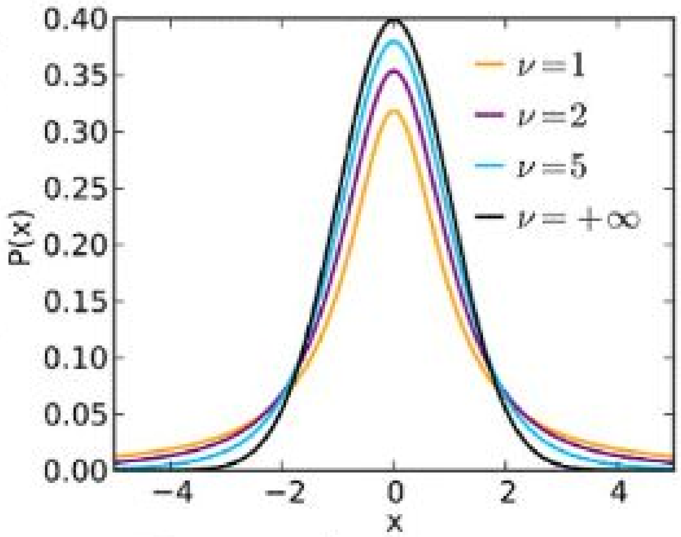


۲- توزیع T : توزیع T مانند توزیع Z (نرمال استاندارد) یک توزیع متقارن با میانگین صفر می باشد. دارای یک پارامتر است که به آن درجه ی آزادی می گوئیم. هنگامی که درجه ی آزادی توزیع به سمت بی نهایت ∞ میل کند، توزیع T به توزیع Z تبدیل می شود. به عنوان یک قاعده ی سرانگشتی اگر درجه ی آزادی حداقل ۳۰ باشد از توزیع Z به جای توزیع T استفاده می کنیم.

نکته: برای پیدا کردن داده های پرت، داده ها را از میانگین کم می کنیم بر انحراف معیار تقسیم می کنیم
داده هایی که از ۳- کمتر و از ۳ بیشتر هستند داده های پرت می باشند.

الف) $t_{0.025}(9) =$ از جدول $=$ ۲/۲

ب) $t_{0.025}(45) \simeq Z_{0.975} = 1.96$



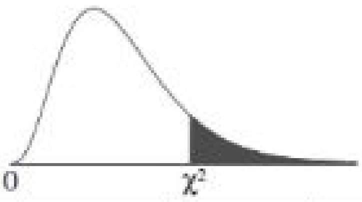
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250

Subject آمار و احتمال

Date ۹۸/۸/۲ ۹۸/۸/۱۶

- متغیر کای دو، یک متغیر پیوسته می باشد که هر مقدار مثبت را در بر می‌گیرد. این توزیع دارای یک پارامتر است، که به آن درجه آزادی گفته می‌شود.

$X \sim \chi^2(v)$



$E(X) = v$

$V(X) = 2v$

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013

الف) $\chi^2_{0.025}(6) = 14.449$

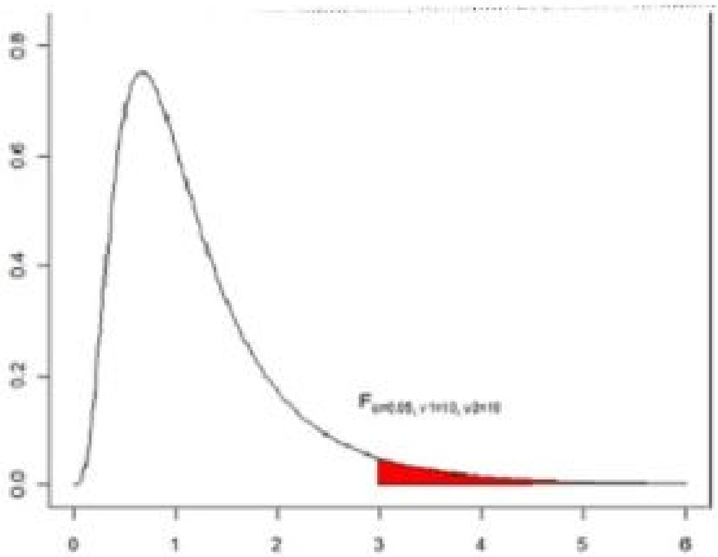
ب) $\chi^2_{0.95}(7) = 2.167$

- توزیع F: فرض کنید $X \sim \chi^2(v_1)$ و $Y \sim \chi^2(v_2)$

درجه آزادی صورت

$$F = \frac{X/v_1}{Y/v_2} \sim F(v_1, v_2)$$

درجه آزادی مخرج



نکته: $F_\alpha(v_1, v_2) = \dfrac{1}{F_{1-\alpha}(v_2, v_1)}$

استفاده از نکته

الف) $F_{0.05}(6, 8) = 3.58$

ب) $F_{0.95}(8, 6) = \dfrac{1}{F_{1-\alpha}(v_2, v_1)} = \dfrac{1}{F_{1-0.95}(v_2, v_1)} = \dfrac{1}{F_{0.05}(6, 8)} = \dfrac{1}{3.58} = 0.27$

	p	1	2	3	4	5	6	7	8
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05

- برآورد فاصله ای : در این قسمت می خواهیم فاصله ای بصورت (L , U) بدست آوریم

با کمک برآورد نقطه ای پارامتر که با اطمینان $(1-\alpha)100\%$ پارامتر جامعه را

در خود داشته باشد.

- برآوردگر فاصله ای میانگین جامعه: فرض کنید یک نمونه ی تصادفی n تایی از یک

جامعه ی نرمال با میانگین و واریانس مجهول μ و σ^2 داشته باشیم. در این صورت

با استفاده از توزیع t یک فاصله اطمینان $(1-\alpha)100\%$ برای μ بصورت زیر است

$$\mu \in (\bar{X} - E , \bar{X} + E)$$

$$E = t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}$$

$$S = \sqrt{S^2}$$

واریانس نمونه ای

* مثال) در یک نمونه تصادفی ۱۰ تایی از کارکنان یک اداره میانگین و واریانس نمره ای رضایت شغلی به ترتیب ۹٫۶ و ۱۰٫۷۱ بوده است. مطلوبست محاسبه ی فاصله ی اطمینان ۹۵٪ برای متوسط واقعی رضایت شغلی کارکنان این اداره

$n = 10$

$\bar{X} = 9.6$

$S^2 = 10.71$

$S = \sqrt{10.71} = 3.27$

$1-\alpha = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$

$n-1 = 10-1 = 9$

$t_{0.025}(9) = 2.262$

$$E = t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}$$

$$E = \frac{2.262 \times 3.27}{\sqrt{10}} = \frac{7.39}{3.16} = 2.33$$

$\mu \in (\bar{X}-E,\ \bar{X}+E) \to \mu \in (9.6-2.33,\ 9.6+2.33)$

$\to \mu \in (7.27,\ 11.93)$ با ۹۸٪ اطمینان متوسط رضایت شغلی بین ۷٫۲۷ تا ۱۱٫۹۳ است.

- فاصله اطمینان برای واریانس جامعه: فرض کنید یک نمونه تصادفی n تایی از یک جامعه نرمال داشته باشیم که میانگین و واریانس آن جامعه مجهول باشد.

در اینصورت با استفاده از توزیع کای دو χ^2 یک فاصله اطمینان $(1-\alpha)100\%$ برای σ^2 بصورت زیر است.

$$\sigma^2 \in \left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)},\ \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)$$

مثال: در مثال قبل یک فاصله اطمینان ۹۵٪ برای پراکندگی واقعی رضایت شغلی بسازید.

$$1-\alpha = 0.95 \qquad \chi^2_{0.025}(9) = 19.028$$
$$\frac{\alpha}{2} = 0.025$$
$$1-\frac{\alpha}{2} = 0.975 \qquad \chi^2_{0.975}(9) = 2.7039$$
$$n-1 = 9$$

$$\sigma^2 \in \left(\frac{9\times 10.71}{19.028},\ \frac{9\times 10.71}{2.7039}\right) \to \sigma^2 \in (5.07,\ 35.69)$$

- فاصله اطمینان برای نسبت جامعه: فرض کنید یک نمونه تصادفی n تایی ($n\geq 30$) از یک جامعه‌ی موفقیت و شکست داشته باشیم. در اینصورت یک فاصله اطمینان $(1-\alpha)100\%$

برای نسبت جامعه (P) بصورت زیر است.

Subject
Date

* مثال) در یک نمونه ی تصادفی ۱۲۰ تایی از شرکت های بورس ۲۵ تای آنها ورشکسته بوده اند. یک فاصله اطمینان ۹۵٪ برای نسبت واقعی شرکت های ورشکسته بیابید.

$$\pi \in (p - E , p + E)$$

$$E = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$n = 120$

$x = 25$

$p = \frac{x}{n} = \frac{25}{120} = 0.21$

$1 - \alpha = 0.95$

$Z_{0.975} = 1.96$

$$\pi \in (p - E , p + E)$$

$$E = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 1.96 \sqrt{\frac{0.21 \times 0.79}{120}}$$

$$1.96 \times 0.037 = 0.073$$

$$\pi \in (0.21 - 0.073 , 0.21 + 0.073)$$

$$\pi \in (0.137 , 0.283)$$

با ۹۵٪ اطمینان نسبت شرکت های ورشکسته بین ۱۳٫۷٪ تا ۲۸٫۳٪ است.

$\boxed{\text{تا اینجا میان ترم}}$

Subject
Date ۹۸/۸/۹

آزمون فرض آماری:

در این قسمت می‌خواهیم ادعایی را که در مورد پارامتر مجهول جامعه بیان می‌شود با استفاده از نمونه‌ی تصادفی جمع‌آوری شده از جامعه تأیید یا رد کنیم.

- فرضیه صفر:

بیانگر وضعیت موجود در جامعه آماری است که از طریق نمونه آن را تأیید یا رد می‌کنیم و آن را با نماد H_0 نشان می‌دهیم و می‌تواند مقدار عددی پارامتر جامعه را مشخص کند. (شامل مساوی است)

فرضیه مقابل را با H_1 نشان می‌دهیم.

* مثال: ادعا شده است که متوسط رضایت شغلی کارکنان دانشگاه بیش از ۲۵ می‌باشد.

$$\begin{cases} H_0: & \mu \leq 25 \\ H_1: & \mu > 25 \end{cases}$$

ادعا ← $\mu > 25$

* مثال: ادعا شده است که نسبت برخی منطقه (الف) و منطقه (ب) متفاوت است.

$$\begin{cases} H_0: & \pi = \pi_B \\ H_1: & \pi \neq \pi_B \end{cases}$$

$\pi_A \neq \pi_B$
ادعا نسبت جامعه آماری

هنگامی که فرضیه H_0 را با استفاده از نمونه تأیید یا رد می‌کنیم دو نوع خطا ممکن است رخ دهد:

۱) خطای نوع اول که آن را با α نشان می‌دهیم احتمال رد H_0 به غلط می‌باشد.

۲) خطای نوع دوم که آن را با β نشان می‌دهیم قبول H_0 به غلط می‌باشد.
(تأیید)

Date

آزمونی مناسب است که برای آن خطای نوع اول و دوم کمترین باشند.

ولی هنگامی که یکی از خطاها کاهش یابد خطای دیگر افزایش می یابد.

به همین دلیل چون خطای نوع اول α، خطای مهمتری می باشد آن را در یک سطح کوچک برای مثال

۰/۰۱ یا ۰/۰۵ کنترل کرده و از میان آزمون های موجود، آزمونی را انتخاب می کنیم که β آن

کمترین باشد.

در آزمون های معنی دار فرض H_0 اگر داده های نمونه تصادفی بیان گر رد H_0 باشند می گوییم

آزمون در سطح خطای α معنی دار است.

* تعریف مقدار احتمال:

در آزمون های معنی دار فرض H_0 در کنار مقدار آماری آزمون مقدار احتمال گزارش می شود.

تا میزان شواهد موجود در داده های نمونه در علیه فرض H_0 مشخص باشد.

به عبارت دیگر مقدار احتمال سازگاری داده های نمونه تصادفی را با فرض H_0 بیان می کند.

فرض H_0 در سطح خطای α رد می شود $\longleftrightarrow$ Sing ۲۲

فرض H_0 در سطح خطای α معنی دار می شود یا

* آزمون میانگین یک جامعه:

فرض کنید یک نمونه تصادفی n تایی از جامعه نرمال داشته باشیم که میانگین و واریانس آن

Subject
Date

مجهول است. می خواهیم یکی از آزمون های زیر را انجام دهیم.

مرحله اول:

(۱)
$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$
یک دنباله ای یا یک طرفه

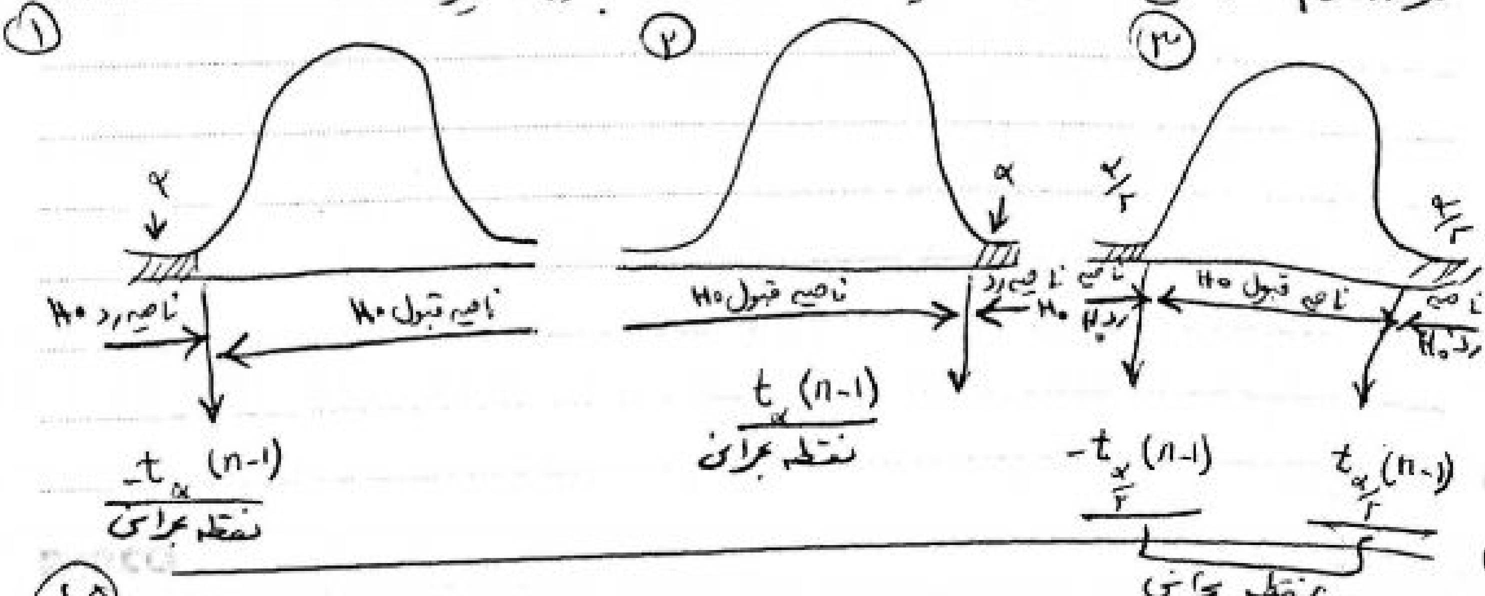
(۲)
$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$
یک دنباله ای یا یک طرفه

(۳)
$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$
دو دنباله ای یا دو طرفه

مرحله دوم: طبق فرض آزمون آماره آزمون بصورت زیر است:

$$t_0 = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t(n-1) , S = \sqrt{S^2}$$

مرحله سوم: در سطح خطای α ناحیه رد فرض H_0 بصورت زیر است.



مرحله چهارم: تصمیم گیری ← اگر مقدار آماره ی آزمون در ناحیه رد H_0 قرار گرفت فرض H_0 را در سطح خطای α رد می کنیم.

* مثال: در یک نمونه تصادفی ۱۰ تایی از کارکنان یک اداره میانگین و واریانس نمونه ای به ترتیب ۹٫۶ و ۱۰٫۷۱ بوده است. درصد خطای ۰٫۰۵ آیا می توان ادعا کرد که متوسط رضایت شغلی کارکنان اداره بیشتر از ۸ باشد؟

$\mu > 8$

$n = 10$

$\bar{X} = 9.6$

$s^2 = 10.71$

$\alpha = 0.05$

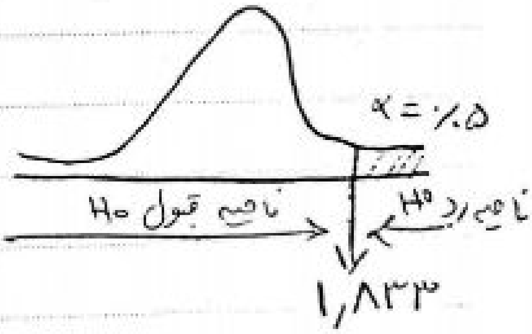
جدول ۲
$$\begin{cases} H_0: \mu \leq 8 \\ H_1: \mu > 8 \end{cases}$$

$$t_0 = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{9.6 - 8}{\frac{3.27}{\sqrt{10}}} = 1.55$$

$$S = \sqrt{s^2} = \sqrt{10.71} = 3.27$$

$$t_{\alpha}(n-1)$$

نقطه بحرانی $t_{0.05}(9) = 1.833$



- چون مقدار آماره آزمون در ناحیه قبول H_0 قرار گرفته است بنابراین در خطای ۰٫۰۵ فرض H_0 را رد نمی کنیم. در سطح خطای ۰٫۰۵ آزمون معنی دار است.

Subject:
Date

مثال: حسابرسی ادعا کرده که میانگین مانده حساب های بدهکاران شرکت ۴۵۰ هزار تومان است.
به منظور بررسی این ادعا یک نمونه ی ۳۶ تایی از حساب های بدهکاران این شرکت انتخاب شده که میانگین آن ۴۰۰ هزار تومان است، و انحراف معیار ۲۰ هزار تومان است. محاسبه کنید آیا می توان ادعا را در سطح خطای ٪۵ پذیرفت.

$n = 36$

$\mu = 450{,}000$

$\bar{X} = 400{,}000$

$S = 20{,}000$

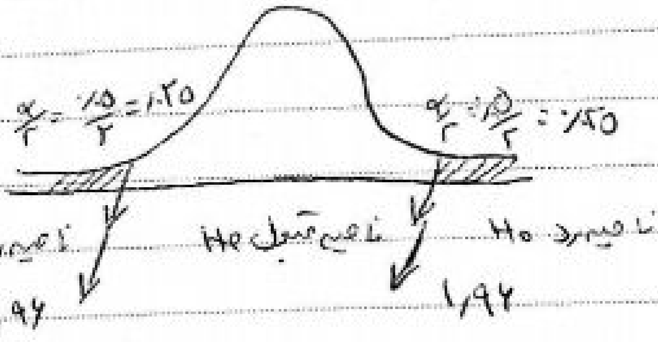
$\alpha = \%5 \rightarrow \frac{\alpha}{2} = .025$

$$t_0 = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{400{,}000 - 450{,}000}{\frac{20{,}000}{\sqrt{36}}} = \frac{-50{,}000}{3333{,}3} = -15{,}000$$

$S = 20{,}000$

$t_{\alpha/2}(n-1) = t_{.025}(35) = 1.96 \longrightarrow$ پایین صفحه

(۳) $\begin{cases} H_0 : \mu = 450{,}000 \\ H_1 : \mu \neq 450{,}000 \end{cases}$



چون مقدار آماره آزمون در ناحیه رد H_0 قرار گرفته است بنابراین فرض H_0 را در سطح خطای ٪۵ رد می کنیم.

نقاط بحرانی $\begin{cases} -t_{\alpha/2}(n-1) = -t_{.025}(35) = -Z_{.975} = -1.96 \\ t_{\alpha/2}(n-1) = t_{.025}(35) = Z_{.975} = 1.96 \end{cases}$

Subject:
Date ۹۸/۸/۲۳

- آزمون مقایسه میانگین های دو جامعه

۱- فرض کنید یک نمونه ی تصادفی m تایی از گروه اول و یک نمونه تصادفی n تایی از گروه دوم داشته باشیم و دو گروه از هم مستقل باشند. همچنین توزیع متغیر مورد مطالعه در دو گروه نرمال باشد، و بخواهیم یکی از آزمون های زیر را انجام دهیم.

$$① \begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

۲- اگر واریانس دو گروه با هم برابر باشند

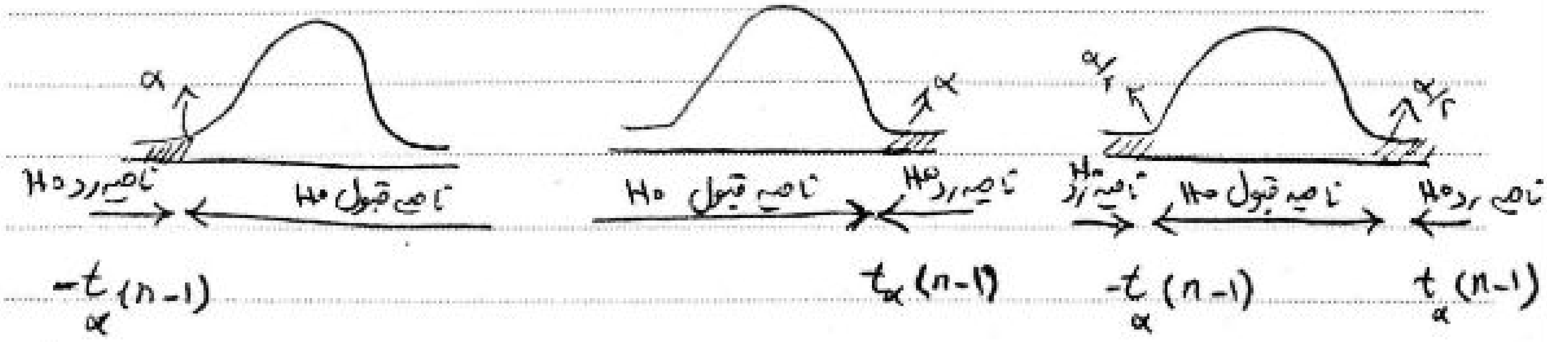
آماره ی آزمون دارای توزیع t است

$$② \begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} \qquad t_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

$$③ \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \qquad S_p = \sqrt{S_p^2}$$

واریانس نمونه ای مشترک $S_p^2 = \dfrac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$

۳- در سطح خطای آلفا ناحیه رد فرض H0 بصورت زیر است:



۴- تصمیم گیری ← اگر مقدار آماره آزمون در ناحیه رد H0 قرار گرفت فرض H0 در سطح خطای α رد می کنیم.

Subject:
Date:

* مثال : فرضیه‌ای به این صورت بیان شده است ، میانگین پرداخت به مدیران در سازمان‌های خصوصی کمتر از سازمان های دولتی است . تحلیل گری برای بررسی فرضیه فوق از هر گروه نمونه‌هایی انتخاب کرده که در جدول زیر آمده است . (داده ها بر حسب ۱۰۰۰ تومان می باشد)

صحت فرضیه را در سطح اطمینان ۹۵٪ آزمون کنید.

س خ	س د
$m = 20$	$n = 25$
$\bar{x}_1 = 95$	$\bar{x}_2 = 100$
$S_1 = 10$	$S_2 = 8$
$S_1^2 = 100$	$S_2^2 = 64$

- نکته : از فرمول ① استفاده شود.

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} = \frac{(19\times 100)+(24\times 64)}{43} = 79.91$$

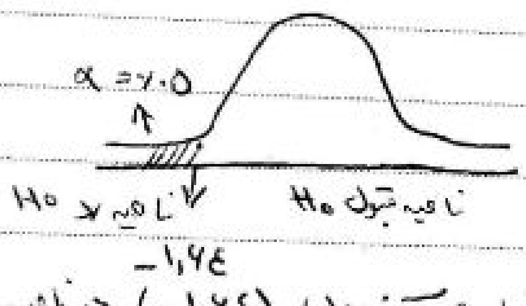
$$S_p = \sqrt{79.91} = 8.94$$

$$\begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

$$1-\alpha = 0.95 \Rightarrow \alpha = 0.05$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{95-100}{8.94\sqrt{\frac{1}{20}+\frac{1}{25}}} = \frac{-5}{2.682} = -1.86$$

$$t_{0.05}(43) = Z_{0.95} = 1.64$$

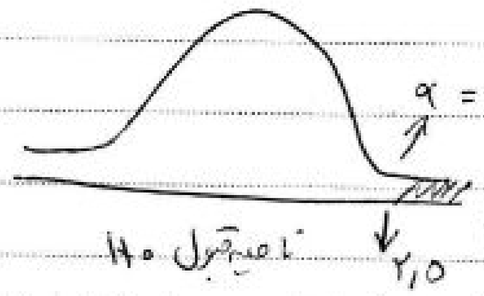


- چون مقدار آماره ی آزمون (-1.64) در ناحیه رد H_0 قرار گرفته است بنابراین در سطح خطای ۰.۰۵ فرض H_0 رد می شود و می پذیریم که میانگین پرداخت مدیران در سازمان های خصوصی کمتر از سازمان های دولتی است.

* مثال: فرضیه این صورت بیان شده است ، روش آموزش متمرکز برای مدیران بهتر از روش غیر متمرکز است . با توجه به جدول زیر صحت فرضیه را در سطح خطای ۰/۰۵ (۰/۰۱) بررسی کنید

n	x	s	t	t_2
۱۵	۴۵	۸	۱/۷۶	±۲/۵
۱۰	۵۲	۱۲		

$$\begin{cases} H_0: \mu_1 \le \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$$



$t_0 = ۱/۷۶$

$t_{0/05}(۲۳) = ۲/۵$

چون مقدار آماره آزمون ۱/۷۶ را در ناحیه قبول H_0 است بنابراین در سطح خطای پنج درصد H_0 رد نمی‌شود و ادعا را نمی‌پذیریم.

۹۸/۸/۳۰

- آزمون t زوجی:

فرض کنید بخواهیم میانگین دو گروه وابسته را آزمون کنیم، برای مثال بخواهیم استعداد دوقلوهای همسان را باهم مقایسه کنیم یا پیش آزمون و پس آزمون روی یک سری شیء یا افراد داشته باشیم. برای مثال نگرش دانشجویان به درس آمار قبل و بعد از آموزش برای هر فرد، باهم مقایسه شود.

زوج	قبل x	بعد y	$D = x - y$
۱	x_1	y_1	D_1
۲	x_2	y_2	D_2
⋮	⋮	⋮	⋮
n	x_n	y_n	D_n

Subject
Date: ۹۸/۸/۳۰

الف - روی زوج ها وابستگی وجود دارد.

ب - تفاوت های زوج از هم مستقل اند.

ج - تفاوت زوجی دارای توزیع نرمال است.

① $\begin{cases} H_0: \mu_D \geq 0 \\ H_1: \mu_D < 0 \end{cases}$

② $\begin{cases} H_0: \mu_D \leq 0 \\ H_1: \mu_D > 0 \end{cases}$

③ $\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D \neq 0 \end{cases}$

آماره آزمون تحت فرض H_0 دارای توزیع t می باشد.

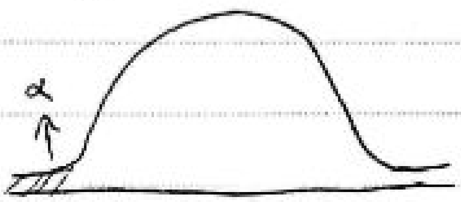
$$t_0 = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} = \frac{\sqrt{n}\bar{D}}{S_D} \sim t(n-1)$$

$$\frac{\sqrt{n}\,\bar{D}}{S_D}$$

$$S_D^2 = \frac{\Sigma D^2 - n\bar{D}^2}{n-1}$$

ناحیه رد فرض H_0 در سطح خطای α به صورت زیر است.

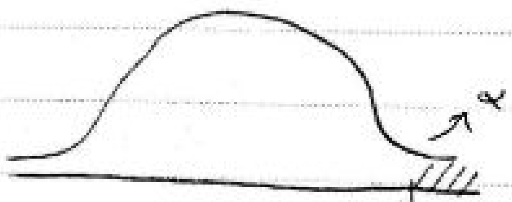
①



ناحیه رد   ناحیه قبول H_0

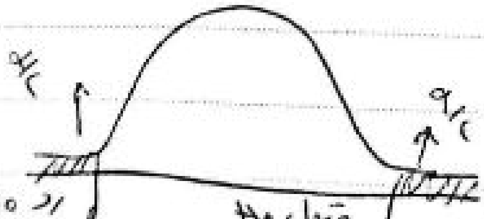
$-t_{\alpha}(n-1)$

②



ناحیه قبول H_0 ناحیه رد H_0

$t_{\alpha}(n-1)$



رد H_0 قبول H_0 رد H_0

$-t_{\frac{\alpha}{2}}(n-1)$ $t_{\frac{\alpha}{2}}(n-1)$

- تصمیم گیری: اگر مقدار آماره ی آزمون در ناحیه ی رد H_0 قرار گرفت فرض H_0 را در سطح خطای آلفا α رد می کنیم

یا می گوییم آ.ت. آزمون در سطح خطای α معنی دار است.

Subject
Date

* مثال: در نظریه‌های مدیریت بیان شده که مهمترین انگیزه پول است. یکی از دانشجویان مدیریت میزان رضایت شغلی ۷ کارمند را قبل از افزایش حقوق ماهانه و بعد از آن اندازه گیری و در این جدول خلاصه کرده است.

کارمند	قبل از افزایش	بعد از افزایش	$D=x-y$	D^2
۱	۵۰	۶۰	-۱۰	۱۰۰
۲	۷۰	۷۵	-۵	۲۵
۳	۸۰	۸۰	۰	۰
۴	۶۰	۷۰	-۱۰	۱۰۰
۵	۵۵	۶۵	-۱۰	۱۰۰
۶	۵۰	۶۲	-۱۲	۱۴۴
۷	۶۸	۷۰	-۲	۴

ادعا $H_0: \mu_D < 0$

$$\begin{cases} H_0: \mu_D \geq 0 \\ H_1: \mu_D < 0 \end{cases}$$

$$\Sigma D = -49 \qquad \Sigma D^2 = 473$$

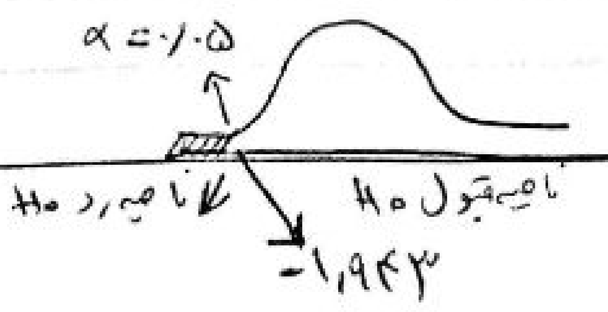
$$\bar{D} = \frac{\Sigma D}{n} = \frac{-49}{7} = -7 \qquad S_D = \sqrt{21.66} = 4.65$$

$$S_D^2 = \frac{\Sigma D^2 - n\bar{D}^2}{n-1} = \frac{473 - 7(7)^2}{7-1} = \frac{473-343}{6} = \frac{130}{6} = 21.66$$

$$t_0 = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} = \frac{\sqrt{n}\,\bar{D}}{S_D} = \frac{(\sqrt{7})(-7)}{4.65} = -3.98$$

$$\alpha = 0.05 \qquad t_{0.05}(6) = 1.943 \qquad df = n-1 = 6$$

درجه آزادی



- در سطح خطای ۰/۰۵ آزمون معنی دار است و ادعا را می پذیریم.

* اگر قدر مطلق آماره ی آزمون بیشتر نقطه بحرانی باشد فرض H_0 رد می شود.

- آنالیز واریانس یک طرفه:

فرض کنید بخواهیم میانگین های بیش از دو گروه را باهم مقایسه کنیم که گروه ها از هم مستقل باشند و متغیر مورد مطالعه هر گروه دارای توزیع نرمال (ح) با واریانس ثابت باشند.

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_k \\ H_1 : \text{حداقل دوتا از میانگین ها متفاوتند} \end{cases}$$

فرض کنید از هر گروه یک نمونه تصادفی به حجم n_i داشته باشیم. مدل طرح به صورت زیر است:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, \dots, k \qquad j = 1, \dots, n_i$$

در اینجا ε_{ij} ها، خطای مدل فرض می شود و دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 است. بنابراین متغیر وابسته y دارای توزیع نرمال خواهد بود. در روش آنالیز واریانس یک طرفه تغییرات کلی (SST) به تغییرات مربوط به گروه بندی یا T_R (SSG) و تغییرات مربوط به خطا (SSE) تفکیک می شود.

گروه اول	گروه دوم	گروه k
y_{11}	y_{21}	y_{k1}
y_{12}	y_{22}	y_{k2}
y_{1n_1}	y_{2n_2}	y_{kn_k}
$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{k.}$

میانگین گروه i ام: $\bar{y}_{i.}$

میانگین کلی: $\bar{y}_{..}$

$$\bar{x}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}}{n} \qquad n = \sum_{i=1}^{k} n_i$$

$$\boxed{SST = SSG + SSE}$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2$$

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

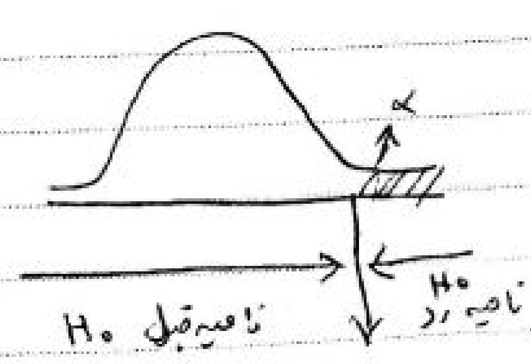
$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

جدول آنالیز واریانس ANOVA

منبع تغییرات	مجموع مربعات	درجه آزادی df	میانگین مربعات MS	F
گروه	SSG	$K-1$	$MSG = \frac{SSG}{K-1}$	$F_0 = \frac{MSG}{MSE}$
خطا	SSE	$n-K$	$MSE = \frac{SSE}{n-K}$	
کل	SST	$n-1$		

اگر $F_0 > F_{ج}$ آنگاه فرض H_0 در سطح خطای α رد می‌شود.

← جدول



$F_{ج} = F_{\alpha}(K-1, n-K)$

← جدول

Subject:
Date

* مثال: محصولات تولیدی یک کارخانه از ۳ ماشین می‌باشد تعداد ضایعات آن‌ها در ۵ روز به صورت زیر است:

	ماشین ۱	ماشین ۲	ماشین ۳	$(x_{ij}-\bar{x}_{..})^2$
	۸۶	۸۹	۸۲	۳۶
	۷۹	۸۳	۶۸	۱
	۸۱	۸۸	۷۳	۱
	۷۰	۷۶	۷۱	۱۰۰
	۸۴	۹۰	۸۱	۱۶
				۸۱
جمع	۴۰۰	۴۲۵	۳۷۵	۴
				۶۴
$\bar{x}_{i.}$	۸۰	۸۵	۷۵	۱۶
				۱۰۰
				۴
				۱۴۴
				۴۹
				۸۱
				۱

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = 698$$

آیا تفاوت معنی‌داری بین تعداد ضایعات ماشین‌ها وجود دارد؟

$$\bar{X}_{..} = \frac{400 + 425 + 375}{15} = 80$$

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \Rightarrow SSG = 5(80-80)^2 + 5(85-80)^2 + 5(75-80)^2$$

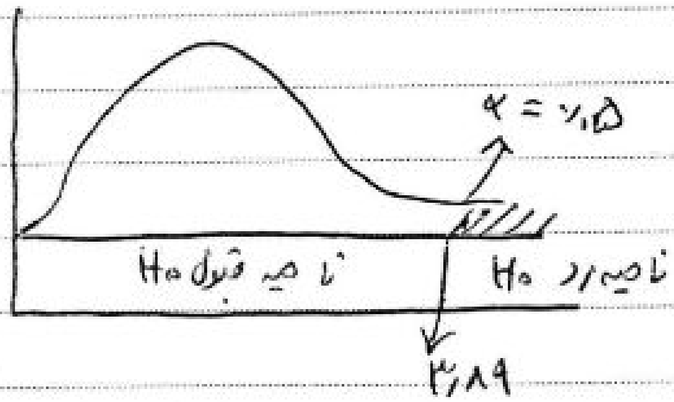
$$= 0 + 125 + 125 = 250$$

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{حداقل ۲ تا از میانگین‌ها متفاوت اند} \end{cases}$$

منبع تغییرات	SS	df	MS	F
گروه	۲۵۰	۲	۱۲۵	$F_0 = \frac{۱۲۵}{۳۷٫۳۳} = ۳٫۳۵$
خطا	۴۴۸	۱۲	۳۷٫۳۳	
کل	۶۹۸	۱۴		

$$SSE = SST - SSG = ۶۹۸ - ۲۵۰ = ۴۴۸$$

$$F_{\%۵}(۲,۱۲) = ۳٫۸۹$$



چون مقدار آماره‌ی آزمون در ناحیه قبول H_0 قرار گرفته است بنابراین آزمون در سطح خطای ۰٫۵٪ معنادار نیست و فرض H_0 را رد نمی‌کنیم یعنی مقدار ضایعات در سه ماشین در سطح خطای ۰٫۵٪ یکسان است.

Subject: ضریب همبستگی
Date: ۹۷/۹/۷

ضریب همبستگی:

فرض کنید بخواهیم رابطه بین ۲ متغیر را مشخص کنیم. و هر دو متغیر مورد بررسی از نوع متغیرهای کمی باشند.

در این صورت روش پارامتری برای این بررسی استفاده از ضریب همبستگی خطی «پیرسون» است.

فرض کنید یک نمونه‌ی تصادفی n تایی بصورت $(x_1,y_1), \ldots, (x_n,y_n)$

از جفت متغیرهای x و y داشته باشیم. در این صورت به عبارت دیگر ضریب همبستگی خطی (ρ) بصورت زیر است:

$$R = \frac{S_{xy}}{S_x S_y} \qquad -1 \le R \le 1$$

S_{xy}: کوواریانس بین x و y

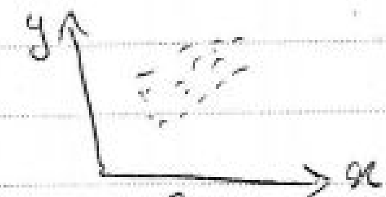
S_x: انحراف معیار x

S_y: انحراف معیار y

در محاسبات دستی

$$R = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$$

الف- اگر $R > 0$ باشد در این صورت

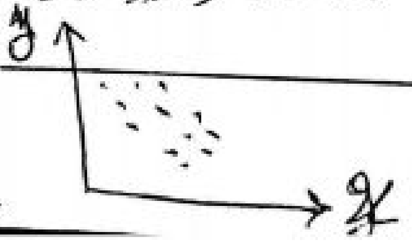


$R>0$

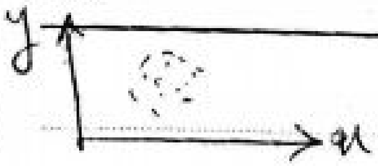
بین متغیرهای x و y رابطه‌ی خطی مستقیم وجود دارد. یعنی با افزایش یک متغیر، متغیر دیگر نیز افزایش می‌یابد. مثل رابطه‌ی بین حقوق و رضایت شغلی.

ب) اگر $R < 0$ باشد، در اینصورت می‌گوییم بین متغیرهای x و y رابطه‌ی خطی معکوس وجود دارد.

یعنی با افزایش یک متغیر، متغیر دیگر کاهش می‌یابد. مثلاً سالهای کارکرد اتومبیل و قیمتش.



Subject
Date



ج- اگر $R=0$ باشد، درانصورت رابطه‌ی خطی بین متغیرها وجود ندارد.

- قبل از محاسبه‌ی ضریب همبستگی خطی، نمودار پراکندگی بین دو متغیر را رسم می‌کنیم که رابطه‌ی خطی بین ۲ متغیر را مشاهده کرده و درصورت وجود داده‌های پرت آنها را مشخص می‌کنیم.

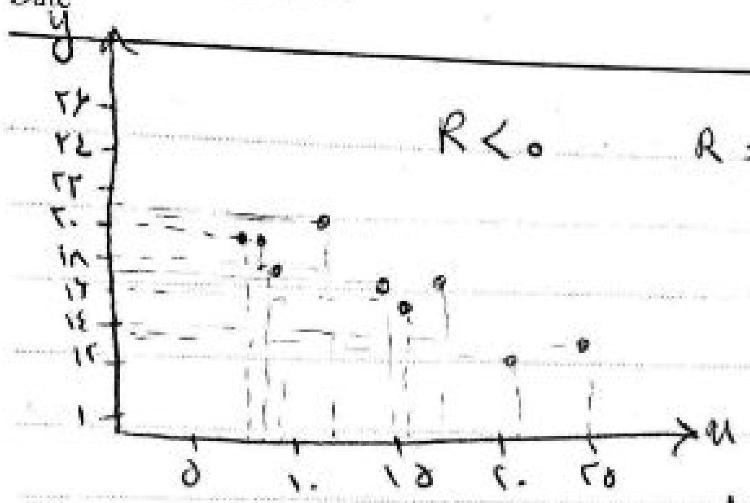
* مثال: در یک نمونه‌ی تصادفی ۹ تایی از کارکنان یک اداره میزان انگیزه و مرخصی سالانه‌ی استفاده شده بصورت زیر است:

مقدار همبستگی بین این دو متغیر را محاسبه کرده و آن را تفسیر کنید.

x تعداد روز مرخصی	y انگیزه شغلی	x^2	y^2	xy
۲۰	۱۵	۴۰۰	۲۲۵	۳۰۰
۹	۱۸	۸۱	۳۲۴	۱۶۲
۱۶	۱۶	۲۵۶	۲۵۶	۲۵۶
۸	۲۰	۶۴	۴۰۰	۱۶۰
۲۵	۱۴	۶۲۵	۱۹۶	۳۵۰
۲۲	۱۳	۴۸۴	۱۶۹	۲۸۶
۱۷	۱۸	۲۸۹	۳۲۴	۳۰۶
۱۲	۲۱	۱۴۴	۴۴۱	۲۵۲
۱۵	۱۷	۲۲۵	۲۸۹	۲۵۵
Σx ۱۴۶	Σy ۱۵۲	Σx^2 ۲,۵۶۸	Σy^2 ۲,۶۲۴	Σxy ۲,۳۲۷

$$\bar{x} = \frac{\Sigma x}{n} = \frac{144}{9} = 16$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{152}{9} = 16{,}88$$



$R < 0$

$$R = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\,\sqrt{\sum y^2 - n\bar{y}^2}}$$

$$= \frac{2327 - (9 \times 16 \times 16.88)}{\sqrt{2568 - 9(16)^2}\,\sqrt{2624 - 9(16.88)^2}}$$

$$= \frac{-103.72}{(16.25)(7.72)} = -\frac{103.72}{125.44} = -0.83$$

رابطه بین انگیزه شغلی و استفاده از مرخصی سالانه معکوس می‌باشد یعنی در صورت افزایش یکی دیگری کاهش می‌یابد.

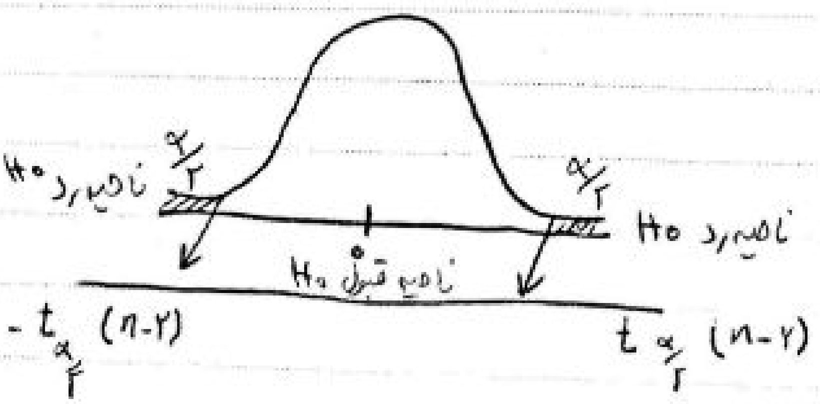
۲- آزمون صفر بودن ضریب همبستگی: فرض کنید بخواهیم آزمون زیر را انجام دهیم:

$$\begin{cases} H_0: \rho = 0 & \text{بین x و y رابطه وجود ندارد} \\ H_1: \rho \neq 0 & \text{بین x و y رابطه وجود دارد} \end{cases}$$

تحت فرض اینکه توزیع توأم (x, y) نرمال باشد، و تحت فرض H_0 آماره‌ی آزمون بصورت زیر است:

$$t_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{(n-2)}$$

در سطح خطای آلفا ناحیه رد فرض H_0 بصورت زیر است.



اگر مقدار آماره آزمون در ناحیه‌ی رد H_0 قرار گیرد فرض H_0 را در سطح خطای آلفا رد می‌کنیم.

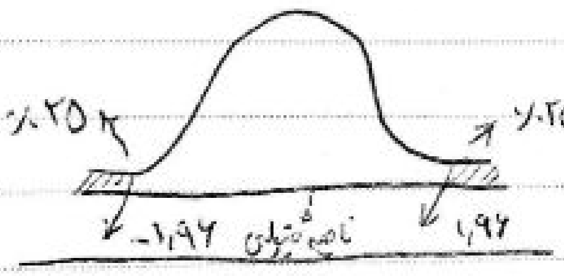
* مثال: در یک نمونه ۳۸ تایی از کارکنان یک اداره مقدار ضریب همبستگی بین انگیزه شغلی و تعداد روزهای مرخصی سالانه ۰٫۸۲- بوده است. در سطح خطای ۵٪ (درصد) آیا می توان ادعا کرد که بین این دو متغیر رابطه‌ی معناداری وجود دارد؟

$$t_0 = \frac{-.82\sqrt{38-2}}{\sqrt{1-(-.82)^2}} = \frac{-4.92}{.57} = -8.63$$

$\alpha = .05$

$\alpha/2 = .025 \qquad t_{.025}(36) \simeq Z_{.975} = 1.96$

چون مقدار آماره آزمون در ناحیه رد H_0 قرار گرفته است.



در سطح خطای ۵٪ بین انگیزه شغلی و استفاده از مرخصی سالانه رابطه‌ی معنی دار وجود دارد.

۹۸/۹/۱۴

رگرسیون خطی ساده: هدف اصلی در تحلیل رگرسیون، پیش بینی یک متغیر از روی یک یا چند متغیر دیگر است. - متغیری که می خواهیم آن را پیش بینی کنیم، متغیر وابسته بوده و آن را با y نشان می دهیم. - متغیرها که در کنترل بوده و برای پیش بینی به کار می روند، متغیرهای مستقل یا پیش بین گفته شده و آنها را $x_1, x_2, \dots, x_p$ نشان می دهیم.

یک مدل رگرسیونی خطی ساده به این صورت می باشد

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Subject ۹۸,۹,۱۴
Date ۹۷,۹,۱۴

در این جا β_0 عرض از مبدأ خط رگرسیونی β_1 شیب خط رگرسیونی و ε خطای تابع رگرسیونی

$f(x) = \beta_0 + \beta_1 x$ در پیش‌بینی y می‌باشد. همچنین فرض می‌کنیم خطاها دارای توزیع نرمال

با میانگین ۰ و واریانس ثابت σ^2 می‌باشند. $\varepsilon \sim N(0, \sigma^2)$

همچنین بین خطاهای مدل خودهمبستگی وجود ندارد. در یک مدل رگرسیونی خطی ساده β_0 و β_1

و σ^2 پارامترهای مدل هستند که آنها را باید برآورد کنیم.

فرض کنید یک نمونه n تایی از جفت متغیرهای x و y داشته باشیم. در این صورت برآوردگرهای

کمترین مربعات معمولی β_0 و β_1 بصورت زیر می‌باشند.

$b_1 = \dfrac{S_{xy}}{S^2_x}$ برآوردگر شیب خط رگرسیون

یا $b_1 = \dfrac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$

$b_0 = \bar{y} - b_1\bar{x}$ برآوردگر عرض از مبدأ

با استفاده از b_0 و b_1 برآوردگر خط رگرسیونی که برای پیش بینی در آینده استفاده می‌شود.

بصورت زیر است: $\hat{y} = b_0 + b_1 x$

اختلاف مقدار مشاهده شده و مقدار پیش بینی شده را باقیمانده‌های مدل می‌گوییم که از آنجا برای

چک کردن شرایط مدل استفاده می‌شود $e = y - \hat{y}$ باقیمانده

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum (y-\hat{y})^2}{n-2}$$

Subject
Date

* مثال: در یک نمونه ی ۸ تایی از فروشگاه های زنجیره ای ۲ آنها میزان تبلیغات و فروش بصورت

e^2 زیر بوده است.

میزان تبلیغات x	فروش y	x^2	y^2	xy	$\hat{y}=7.2+1.65x$	$e=y-\hat{y}$	e^2
4	15	16	225	60	13.8	1.2	1.44
3	12	9	144	36	12.15	-0.15	0.0225
4	14	16	196	56	13.8	0.2	0.04
5	16	25	256	80	15.45	0.55	0.3025
6	18	36	324	108	17.1	0.9	0.81
2	10	4	100	20	10.5	-0.5	0.25
1	9	1	81	9	8.85	0.15	0.0225
7	17	49	289	119	18.75	-1.75	3.0625
$\Sigma x=32$	$\Sigma y=111$	$\Sigma x^2=156$	$\Sigma y^2=1615$	$\Sigma xy=488$			5.95

اگر میزان تبلیغات شرکتی ۴٫۵ میلیون تومان باشد میزان فروش فروشگاه را پیش بینی کنید.

$$\bar{x}=\frac{\Sigma x}{n}=\frac{32}{8}=4$$

$$\bar{y}=\frac{\Sigma y}{n}=\frac{111}{8}=13.8$$

$$b_1=\frac{488-(8\times 4\times 13.8)}{156-8(4)^2}=\frac{46.4}{28}=1.65$$

$\boxed{\text{علامت } b_1 \text{ و } R \text{ یکی است}}$ نکته مهم

$$b_0=\bar{y}-b_1\bar{x}=13.8-(1.65\times 4)=7.2$$

$$\hat{y}=7.2+1.65x$$

$$\hat{y}(x=4.5)=7.2+(1.65\times 4.5)=14.62$$

$$\sigma^2=\frac{\sum_{i=1}^{n} e_i^2}{n-2}=\frac{5.95}{6}=0.99$$

$$\hat{\sigma}=\sqrt{0.99}=0.994$$

Subject
Date

*آزمون صفر بودن ضرایب: با توجه به فرض نرمال بودن خطا

$$\text{①}\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

بنابراین متغیر وابسته y نیز دارای توزیع نرمال است.

$$\text{②}\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

درنتیجه چون b_0 و b_1 ترکیب خطی از y می باشند،

بنابراین آنها نیز دارای توزیع نرمال هستند:

$$b_0 \sim N(\beta_0, V(b_0))$$

$$b_1 \sim N(\beta_1, V(b_1))$$

$$V(b_1) = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

خطای معیار $SE(b_0) = \sqrt{\hat{V}(b_0)} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x-\bar{x})^2}}$

$$SE(b_1) = \sqrt{\hat{V}(b_1)} = \frac{\hat{\sigma}}{\sqrt{\sum (x-\bar{x})^2}}$$

$$\sum (x - \bar{x})^2 = \sum x^2 - n\bar{x}^2$$

تحت فرض H_0 آماره‌ی آزمون دارای توزیع t می باشد.

$$t_0 = \frac{b}{SE(b)} \sim t(n-2)$$

و فرض H_0 را در سطح خطای آلفا رد می کنیم

اگر $|t_0| \geq t_{\frac{\alpha}{2}}(n-2)$ باشد.

نقطه بحرانی

Subject

Date

* مثال: در مثال قبل صفر بودن شیب خط رگرسیون را آزمون کنید:

$$\begin{cases} H_0: B_1 = 0 \\ H_1: B_1 \neq 0 \end{cases}$$

$$\sum (x-\bar{x})^2 = \sum x^2 - n\bar{x}^2 = 156 - 8(4)^2 = 28$$

$$SE(b_1) = \sqrt{\hat{V}(b_1)} = \frac{\hat{\sigma}}{\sqrt{\sum (x-\bar{x})^2}} = \frac{0.994}{\sqrt{28}} = \frac{0.994}{5.29} = 0.18$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025 \rightarrow t_{0.025}(6) = 2.447$$

$$t_0 = \frac{b_1}{SE(b_1)} = \frac{1.65}{0.18} = 9.16$$

چون $|t_0| > 2.447$ بنابراین در سطح خطای ۵ درصد صفر بودن شیب خط رگرسیون رد می‌شود

تمرین: صفر بودن عرض از مبدأ را در مثال قبل آزمون کنید:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

$$SE\,b_0 = \sqrt{\hat{V}(b_0)} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x-\bar{x})^2}}$$

$$= 0.994\sqrt{\frac{1}{8} + \frac{16}{28}} =$$

$$0.994 \times 0.83 = 0.829$$

$$t_0 = \frac{b_0}{SE(b_0)} = \frac{7.2}{0.829} = 8.68$$

$$|t_0| > 2.447$$

می‌باشد بنابراین در سطح خطای ۵ درصد صفر بودن عرض از مبدأ خط رگرسیون را رد می‌کنیم

$$|t_0| > t_{\frac{\alpha}{2}}(n-2)$$

Subject
Date ۹۸/۹/۲۱

- ضریب تعیین: مربع ضریب همبستگی را (R^2) ضریب تعیین می‌گوییم.

درصد تغییرات متغیر وابسته می‌باشد که توسط مدل رگرسیونی با متغیر مستقل x بیان می‌شود.

هر چه R^2 به یک نزدیک‌تر باشد قدرت مدل برای پیش‌بینی بیشتر است. $0 \leq R^2 \leq 1$

★ مثال: مدل فرضی زیر را در نظر بگیرید.

$\hat{y} = ۳ - ۱/۲ x$

$(۱/۱۵) \quad (۰/۴۱) \leftarrow$ اعداد درون پرانتز خطای معیار

$R^2 = ۰/۶۴$

$n = ۴۸$

- x نسبت بدهی شرکت

- y سود خالص شرکت

الف: معنادار بودن شیب خط رگرسیونی را آزمون کنید.

ب: ضریب همبستگی را بدست آورده آن را تفسیر کنید.

ج: اگر بدهی شرکتی ۰/۲۹ باشد سود آن را پیش‌بینی کنید.

ضریب	b	SE
ثابت	۳	۱/۱۵
x	$-۱/۲$	۰/۴۱

$R^2 = ۰/۶۴ \quad n = ۴۸$

$S \leftarrow$ انحراف معیار x $\quad S$

$SE \leftarrow$ خطای معیار $\bar{x}$ $\quad SE(\bar{x}) = \frac{S}{\sqrt{n}}$

سؤال: تفاوت انحراف معیار با خطای معیار چیست.

$\alpha = ۰/۵$ اگر α را در مسئله نگفت باید بگیریم $\Rightarrow \frac{\alpha}{2} = ۰/۰۲۵$

الف)

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

آماره آزمون $t_0 = \frac{b_1}{SE(b_1)} = \frac{-۱/۲}{۰/۴۱} = -۲/۹۲$

نقطه بحرانی $t_{(۴۶)} = ۱/۹۶ \qquad |t_0| > ۱/۹۶$

بنابراین در سطح خطای ۵٪ شیب خط رگرسیون

مخالف صفر است. بنابراین نسبت بدهی اثر معنی‌دار بر سود خالص دارد.

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \qquad t_0 = \frac{b_0}{SE(b_0)} = \frac{۳}{۱٫۱۵} = ۲٫۶$$

$$|t| > ۱٫۹۶$$

در سطح خطای ۵٪ عدد ثابت (عرض از مبدأ) مخالف صفر است.

ب) $R = \sqrt{۰٫۶۴} = \pm ۰٫۸ \qquad R = -۰٫۸$

چون $b_1 = -۱٫۲$ می‌باشد برای R ، $-۰٫۸$ را قبول می‌کنیم.

بین بدهی و سود خالص رابطه معکوس وجود دارد.

ج) $\hat{y}(x = ۰٫۲۹) = ۳ - ۱٫۲(۰٫۲۹)$

$$\Rightarrow ۲٫۶۵۲$$

در امتحان پایان ترم از رگرسیون ۲ سؤال - از کای دو ۱ سؤال.

- آزمون استقلال کای دو:

فرض کنید متغیرهای x و y از نوع متغیرهای کیفی باشند و بخواهیم آزمون زیر را انجام دهیم.

$$\begin{cases} H_0 : \text{متغیرهای } x \text{ و } y \text{ رابطه ندارند} \\ H_1 : \text{متغیرهای } x \text{ و } y \text{ رابطه دارند} \end{cases}$$

همچنین متغیر x دارای k طبقه و متغیر y دارای c طبقه باشند و یک نمونه n تایی از جفت متغیرهای x و y داشته باشیم و آنها را در جدول توافقی زیر طبقه بندی کرده باشیم:

$x \backslash y$	B_1	B_2	B_c	جمع
A_1	O_{11}	O_{12}	O_{1c}	$O_{1.}$
A_2	O_{21}	O_{22}	O_{2c}	$O_{2.}$
A_R	O_{R1}	O_{R2}	O_{R3}	O_{R}
جمع	$O_{.1}$	$O_{.2}$	$O_{.3}$	n

O_{ij} : تعداد مشاهدات در سلول (i,j) ام

$O_{i.}$: جمع سطر i ام

$O_{.j}$: جمع ستون j ام

تحت فرض H_0 مقدار مورد انتظار از رابطه زیر محاسبه می‌شود:

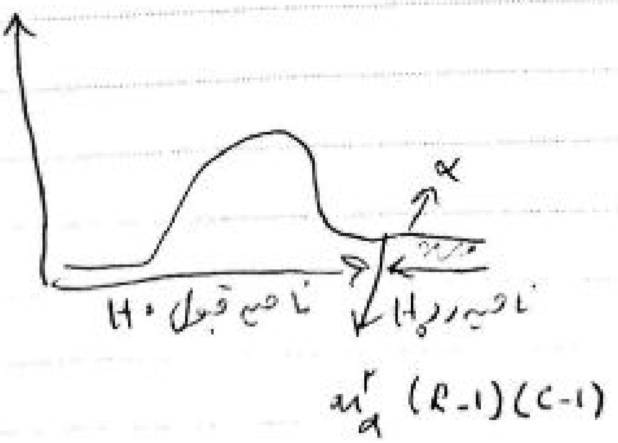
$$e_{ij} = \frac{O_{i.} \times O_{.j}}{n}$$

و آماره آزمون دارای توزیع کای دو می‌باشد.

$$\chi_0^2 = \sum_{i=1}^{R} \sum_{j=1}^{c} \frac{(e_{ij} - O_{ij})^2}{e_{ij}} \sim \chi^2_{(R-1)(C-1)}$$

فرض H_0 در سطح خطای آلفا رد می‌شود اگر کای دو صفر بیشتر از فرض جدول باشد .... H_0 رد می‌شود.

$$\chi_0^2 > \chi^2_{\alpha\,(R-1)(C-1)}$$



مثال: جدول زیر نشان دهنده میزان علاقه به رشته‌ی تحصیلی در میان دانشجویان رشته‌ی مدیریت می‌باشد:

جنسیت \ علاقه	ضعیف	متوسط	زیاد	جمع
مرد	(۳۹) ۴۵	(۵۱) ۵۰	(۳۰) ۲۵	۱۲۰
زن	(۲۶) ۲۰	(۳۴) ۳۵	(۲۰) ۲۵	۸۰
جمع	۶۵	۸۵	۵۰	n=۲۰۰

آیا در سطح خطای ۵٪ می‌توان ادعا کرد که بین جنسیت و علاقه به رشته‌ی تحصیلی رابطه وجود دارد؟

H_0: جنسیت و علاقه به رشته رابطه ندارند

H_1: جنسیت و علاقه به رشته رابطه دارند

$$e_{ij} = \frac{o_{i.} \times o_{.j}}{n}$$

$$e_{11} = \frac{120 \times 65}{200} = 39 \qquad e_{21} = \frac{80 \times 65}{200} = 26$$

$$e_{12} = \frac{120 \times 85}{200} = 51 \qquad e_{22} = \frac{80 \times 85}{200} = 34$$

$$e_{13} = \frac{120 \times 50}{200} = 30 \qquad e_{23} = \frac{80 \times 50}{200} = 20$$

e_{ij}	o_{ij}	$\frac{(e_{ij} - o_{ij})^2}{e_{ij}}$
۳۹	۴۵	$\frac{(39-45)^2}{39} = 0/92$
۵۱	۵۰	$\frac{(51-50)^2}{51} = 0/01$
۳۰	۲۵	$= 0/83$
۲۶	۲۰	$= 1/38$
۳۴	۳۵	$= 0/02$
۲۰	۲۵	$= 1/25$

جمع ستون آماره‌ی آزمون $\chi^2 = 3/41$

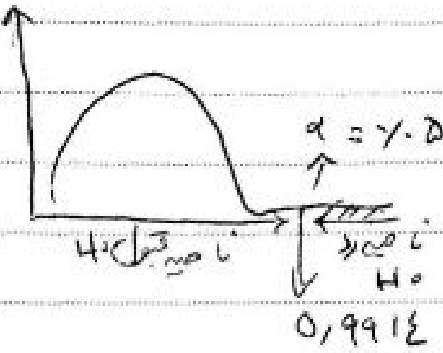
Subject:
Date

$R=2$

$C=3$

$df = (2-1)(3-1) = 2$

نقطه بحرانی $\chi^2_{\%5}(2) = 5,99147$



چون مقدار آماره آزمون (41,41) در ناحیه قبول H_0 است

بنابراین در سطح خطای ۵ درصد

H_0 رد می‌شود.