خلاصه مطالب آمار ۲

موجود در سایت:

www.mbdm.ir

مروری بر

فصل هشتم و نهم

برآورد میانگین، نسبت و واریانس جامعه

ویژگی های برآورد کننده نقطه ای خوب

- ناریب بودن

- کارایی

- سازگاری

- ناریب بودن: آماره $\hat{\boldsymbol{\theta}}$ را یک برآورد کننده ناریب پارامتر θ می گویند هر گاه میانگین توزیع نمونه ای آن برابر پارامتر θ باشد. یعنی

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

آماره $\bar{X}$ (میانگین نمونه ای) یک برآورد کننده ناریب برای μ است زیرا

$$E(\bar{X}) = \mu$$

آماره S^2 (واریانس نمونه ای)، وقتی جامعه نامتناهی باشد، یک برآورد کننده ناریب برای σ^2 است زیرا

$$E(S^2) = \sigma^2$$

نکته : S یک برآورد کننده اریب می باشد زیرا :

$$E(S) = \sigma$$

- **کارایی:** **هر گاه دو** برآورد کننده نا اریب برای پارامتر θ موجود باشد، برآورد کننده ای که که واریانس کوچک تری نسبت به دیگری دارد از کارایی بیشتری برخودار است.

- **سازگاری:** برآورد کننده ای مانند $\hat{\theta}$ را یک برآورد کننده سازگار برای پارامتر θ مینامیم هرگاه با افزایش اندازه نمونه n ، $\hat{\theta}$ با احتمال یشتری به θ نزدیک شود.

برآورد فاصله ای

در برآورد فاصله ای، برای پارامتر جامعه فاصله اطمینانی می سازیم که حد بالا و پایین پارامتر جامعه را مشخص می سازد

$$L < \theta < U$$

حد پایین

پارامتر مجهول

حد بالا

ضریب اطمینان:

ضریب اطمینان یعنی احتمال این که فاصله اطمینان، شامل مقدار واقعی پارامتر برآورد شده باشد.

$$c = P(L < \theta < U)$$

معمولا در برآورد فاصله ای ابتدا ضریب اطمینان را مشخص می کنند و سپس با توجه به آن حد بالا و پایین محاسبه می گردد.

خطای برآورد

بدلیل اینکه در نمونه گیری از بخشی از اطلاعات استفاده می کنیم و نه همه آن، همیشه با درصدی از خطا روبرو هستیم

$$\boldsymbol{d} = \text{خطای برآورد} = \left|\text{مقدار واقعی پارامتر} - \text{مقدار برآورد پارامتر}\right|$$

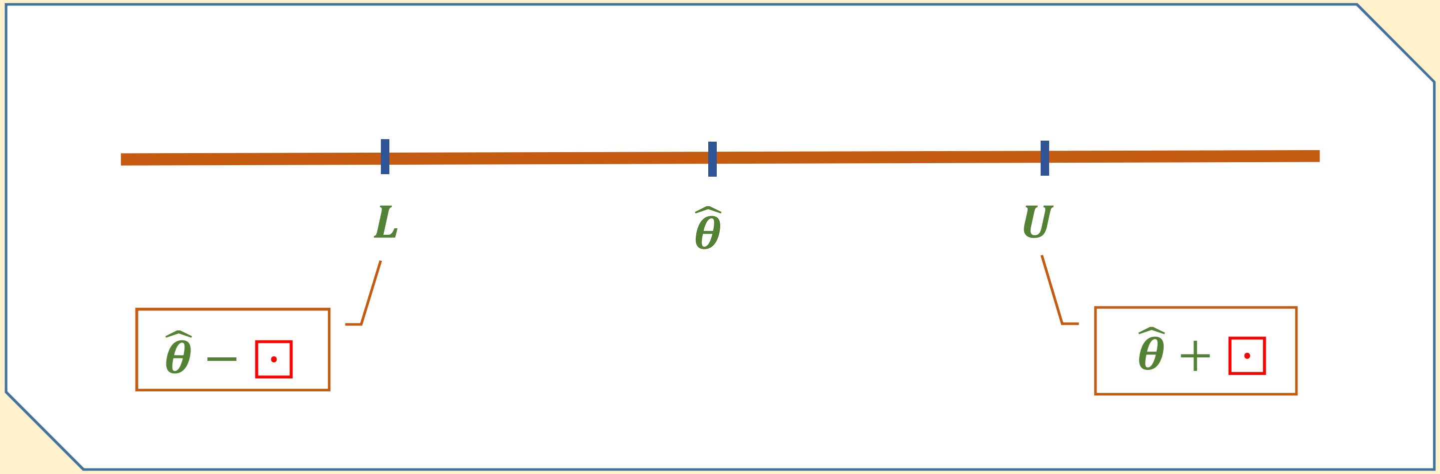
$$P(L < \theta < U) = 1 - \alpha$$

حد پایین

پارامتر جامعه

حد بالا

ضریب اطمینان



$$P(d < \boxed{\cdot}) = 1 - \alpha$$

خطای برآورد

$\bar{X} =$ میانگین نمونه

$S =$ واریانس نمونه

$\mu =$ میانگین جامعه

$\sigma =$ واریانس جامعه

L — $\bar{X}$ — U

$L = \bar{X} - \boxdot$

$U = \bar{X} + \boxdot$

- برآورد فاصله ای میانگین جامعه
  - نمونه بزرگ $(n \geq 30)$
    - σ معلوم: $\boxdot = \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$
    - σ نا معلوم: $\boxdot = \frac{S}{\sqrt{n}} Z_{\frac{\alpha}{2}}$
  - نمونه کوچک $(n < 30)$
    - جامعه نرمال
      - σ معلوم: $\boxdot = \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$
      - σ نا معلوم: $\boxdot = \frac{\sigma}{\sqrt{n}} t_{\left(\frac{\alpha}{2}, n-1\right)}$
    - جامعه غیر نرمال
      - σ معلوم: $\boxdot = \frac{\sigma}{\sqrt{n}} t_{\left(\frac{\alpha}{2}, n-1\right)}$

$$\boxed{\bar{X} = \frac{\sum_{i=۱}^{n} X_i}{n}}$$

میانگین نمونه

$$Variance_{Sample} = \frac{\sum_{i=۱}^{n} \left(X_i - \boxed{\bar{X}}\right)^{۲}}{n-۱}$$

$$Variance_{Population} = \frac{\sum_{i=۱}^{n} \left(X_i - \bar{X}\right)^{۲}}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

برآورد نقطه ای نسبت جامعه

بهترین برآورد کننده نسبت جامعه (P)، نسبت نمونه ($\bar{P} = \frac{x}{n}$) است.

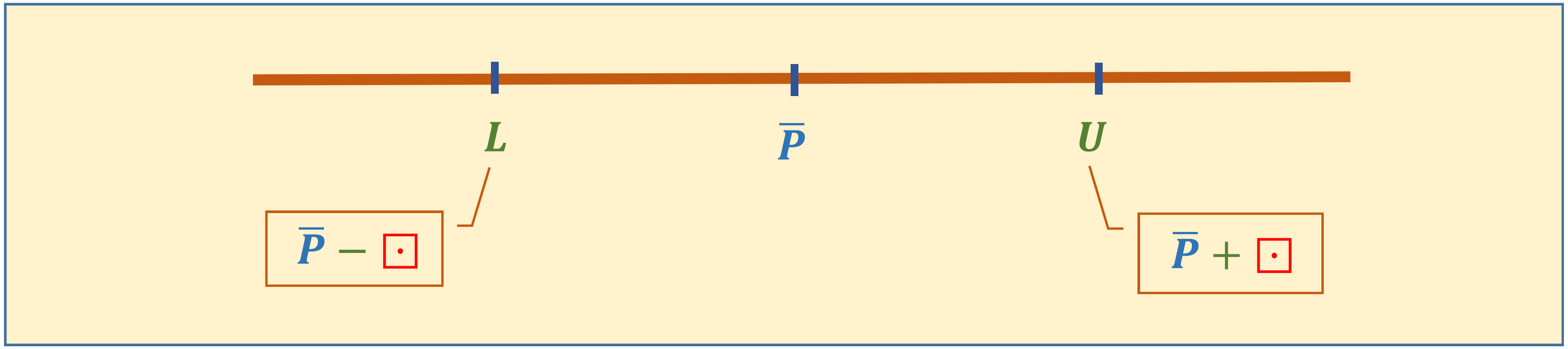
می دانیم که :

$$\mu_{\bar{P}} = E(\bar{P}) = P$$

$$\sigma_{\bar{P}}^2 = \frac{P(1-P)}{n}$$

در صورت مشخص نبودن P ، باید انحراف معیار نسبت نمونه ای رابرآورد نمود

$$S_{\bar{P}}^2 = \frac{\bar{P}(1-\bar{P})}{n}$$



نسبت نمونه $= \bar{P}$

نسبت جامعه $= P$

برآورد فاصله ای نسبت جامعه

نمونه بزرگ
($n \geq 30$)

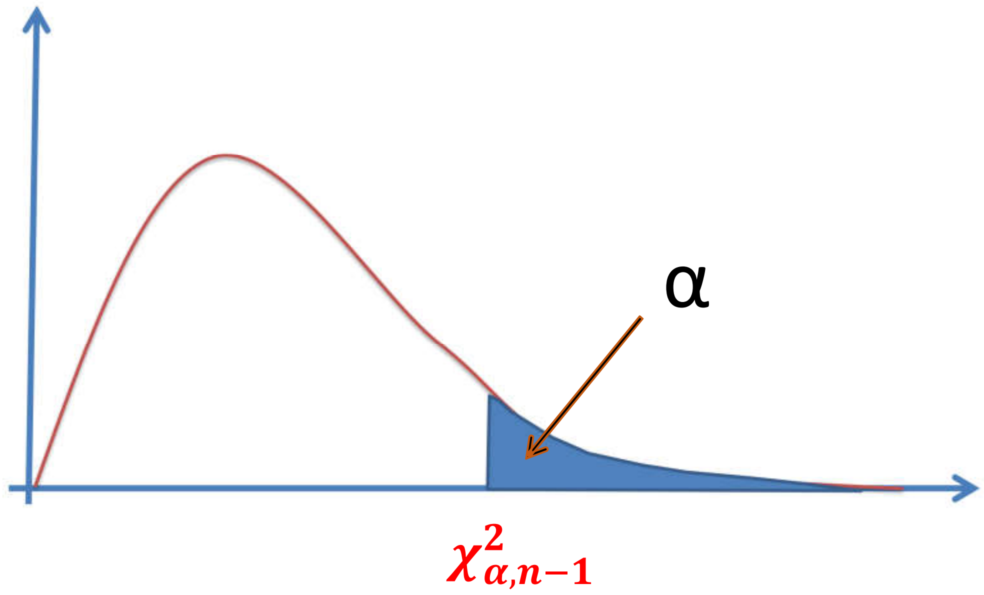
$$\boxdot = \sqrt{\frac{\bar{P}(1-\bar{P})}{n}}\, Z_{\frac{\alpha}{2}}$$

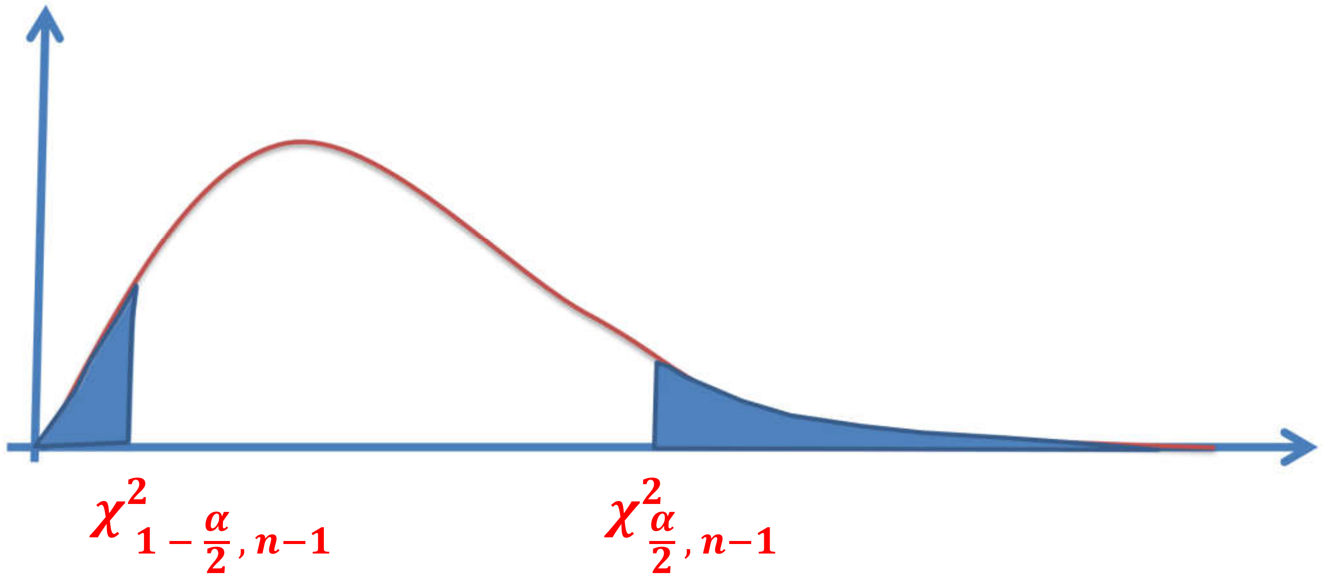
$$P(d < \boxdot) = 1 - \alpha$$

خطای برآورد

برآورد واریانس جامعه

- ✓ روش های استنباط در مورد واریانس و انحراف معیار جامعه با فرض <u>توزیع جامعه نرمال</u> بررسی می شوند.
- ✓ برای پیدا کردن حدود اطمینان σ^2 از نقاط درصد توزیع χ^2 استفاده می شود. شکل این توزیع وابسته به درجه ازادی توزیع است و منحنی نامتقارنی است که چوله به راست می باشد





$$P\left(\underbrace{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}}_{L} < \sigma^2 < \underbrace{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}_{U}\right) = 1-\alpha$$

واریانس جامعه

ضریب اطمینان

حد پایین

حد بالا

مروری بر

فصل دهم و یازدهم

آزمونهایی برای میانگین، نسبت و واریانس جامعه

مراحل عمومی آزمون فرض آماری

۱- تعریف فرضیه های آماری H_0 و H_1

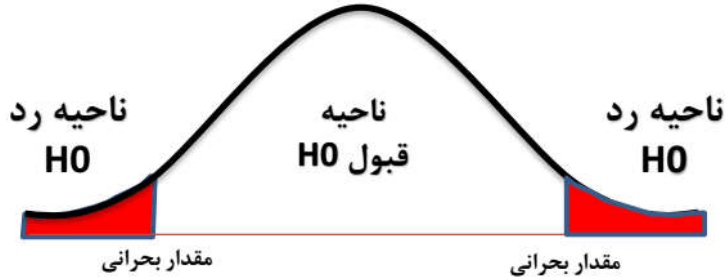
۲- تعیین نوع آماره آزمون

۳- تعیین مقادیر بحرانی

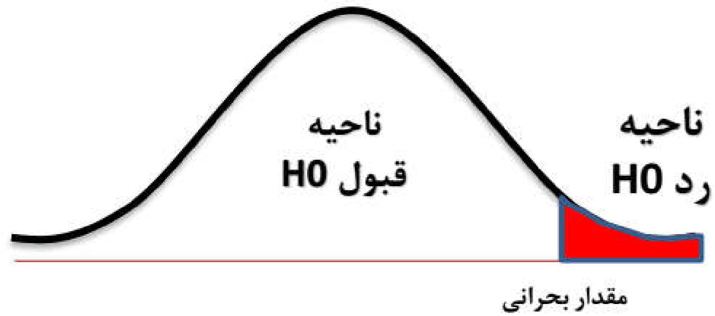
۴- تصمیم گیری

تعریف فرضیه های آماری H_0 و H_1

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$



$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$



$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

تعیین نوع آماره آزمون

- برآورد فاصله ای میانگین جامعه
  - نمونه بزرگ ($n \geq 30$)
    - σ معلوم — ۱
    - σ نا معلوم — ۲
  - نمونه کوچک ($n < 30$)
    - جامعه نرمال
      - σ معلوم — ۳
      - σ نا معلوم — ۴
    - جامعه غیر نرمال
      - σ معلوم — ۵

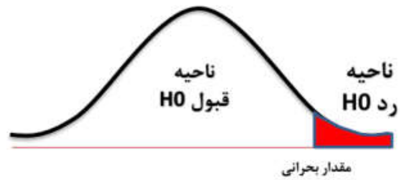
آماره آزمون در حالت‌های ۱ و ۳ عبارتست از

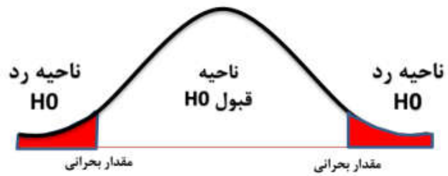
$$\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

آماره آزمون در حالت‌های ۲، ۴ و ۵ عبارتست از

$$\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$$

تعیین مقادیر بحرانی



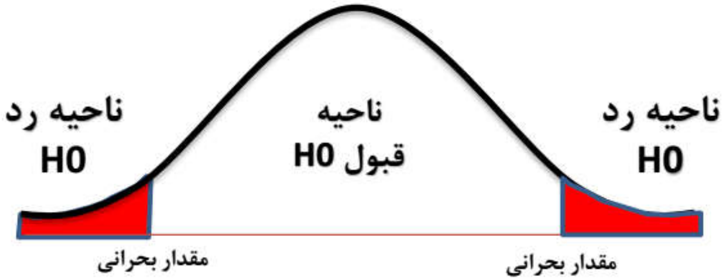


مدل ۱	Z_{α}	$Z_{\frac{\alpha}{2}}$	$-Z_{\frac{\alpha}{2}}$	$-Z_{\alpha}$
مدل ۲	$t_{(\alpha,n-1)}$	$t_{\left(\frac{\alpha}{2},n-1\right)}$	$-t_{\left(\frac{\alpha}{2},n-1\right)}$	$-t_{(\alpha,n-1)}$

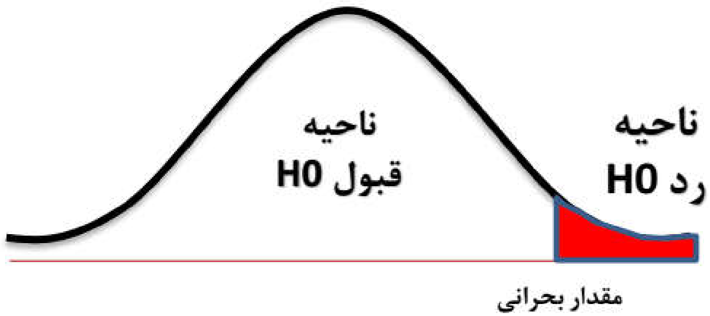
- برآورد فاصله ای میانگین جامعه
  - نمونه بزرگ ($n \geq 30$)
    - σ معلوم → مدل ۱
    - σ نا معلوم → مدل ۱
  - نمونه کوچک ($n < 30$)
    - جامعه نرمال
      - σ معلوم → مدل ۱
      - σ نا معلوم → مدل ۲
    - جامعه غیر نرمال
      - σ معلوم → مدل ۲

تعریف فرضیه های آماری H_0 و H_1

$$H_0 : P = P_0$$
$$H_1 : P \neq P_0$$



$$H_0 : P \leq P_0$$
$$H_1 : P > P_0$$



$$H_0 : P \geq P_0$$
$$H_1 : P < P_0$$

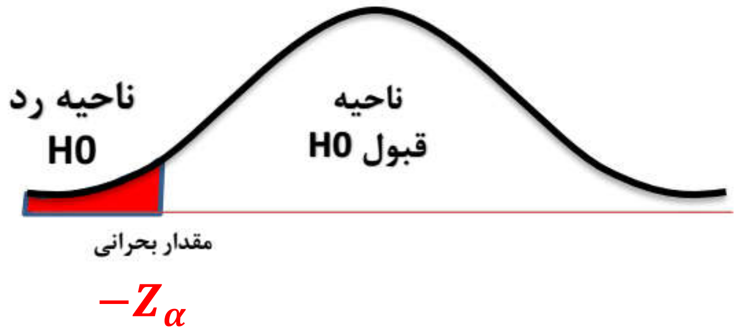
تعیین نوع آماره آزمون

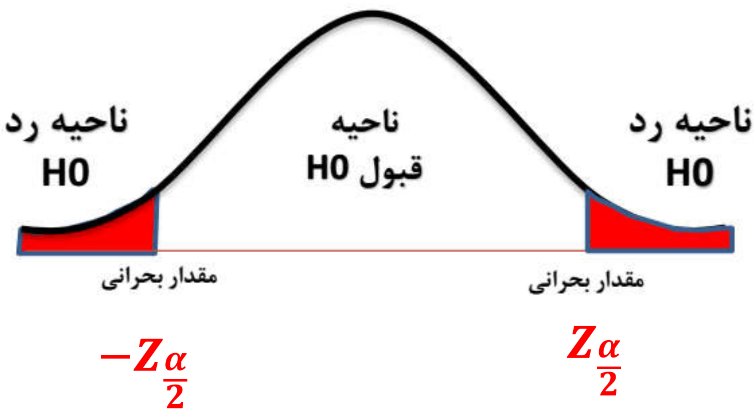
نمونه بزرگ
($n \geq 30$)

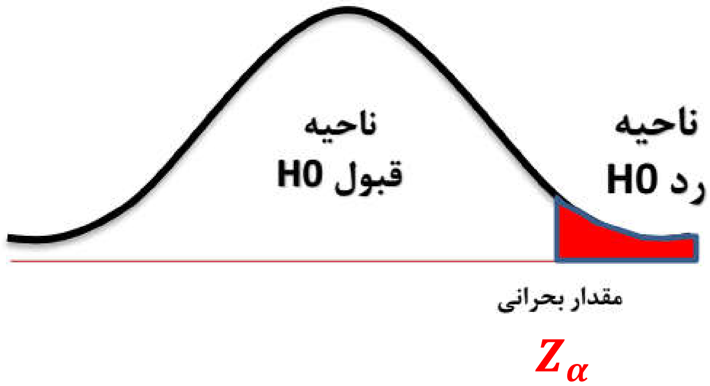
آماره آزمون عبارتست از

$$Z = \frac{\overline{P} - P_0}{\sigma_{\overline{P}}} = \frac{\overline{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$$

تعیین مقادیر بحرانی







آزمون‌های فرض آماری برای واریانس جامعه:

آزمون‌های مربوط به واریانس جامعه براساس آماره‌ی آزمون زیر انجام می‌گیرند:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

قاعده‌ی تصمیم‌گیری:

وقتی جامعه‌ی مورد نظر نرمال یا تقریباً نرمال باشد، آماره‌ی χ^2 دارای توزیع χ^2 با $(n-1)$ درجه‌ی آزادی است و بنابراین ناحیه‌ی رد آزمون‌های مطرح شده در سطح α عبارتند از:

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases} \Rightarrow \text{آزمون دوطرفه} \xrightarrow{\text{ناحیه رد}} \chi^2 > \chi^2_{(\frac{\alpha}{2}, n-1)} \qquad \chi^2 < \chi^2_{(1-\frac{\alpha}{2}, n-1)}$$

$$\begin{cases} H_0: \sigma^2 \leq \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases} \Rightarrow \text{آزمون یک طرفه راست} \xrightarrow{\text{ناحیه رد}} \chi^2 > \chi^2_{(\alpha, n-1)}$$

$$\begin{cases} H_0: \sigma^2 \geq \sigma_0^2 \\ \end{cases} \Rightarrow \text{آزمون یک طرفه چپ} \xrightarrow{\text{ناحیه رد}} \chi^2 < \chi^2_{(1-\alpha, n-1)}$$

فصل دوازدهم

تحلیل (آنالیز) واریانس

www.mbdm.ir

نحوه تصمیم گیری

آماره F با مقدار Fجدول (مقدار بحرانی) مقایسه می‌گردد و چنانچه مقدار محاسبه شده آماره کوچکتر از Fجدول (مقدار بحرانی) گردد، فرض H0 مبنی بر برابر بودن میانگین جامعه‌ها پذیرفته می‌شود.

به منظور سهولت به کارگیری تحلیل واریانس از جدولی که بدین منظور طراحی گردیده استفاده می‌گردد.

F	واریانس میانگین انحرافات MS	درجه آزادی d.f	مجموع توان دوم انحرافات SS	منبع انحرافات
$F = \frac{MSR}{MSE}$	MSR	K-1	SSR	بین نمونه‌ها
	MSE	N-k	SSE	درون نمونه‌ها
		N-1	SST	کل

مثال

می‌خواهیم تعداد ضایعات سه ماشین را با هم مقایسه کنیم و بدانیم که آیا تفاوت معنی‌داری بین میانگین ضایعات آن‌ها وجود دارد یا خیر. و یا این‌که اختلاف معلول تصادف است(در سطح معنی‌دار ۵٪). نمونه‌های زیر، از این ماشین‌ها جهت بررسی جمع‌آوری گردیده است.

ماشین‌ها	تعداد ضایعات				
	روز ۱	روز ۲	روز ۳	روز ۴	روز ۵
۱	۸۶	۷۹	۸۱	۷۰	۸۴
۲	۸۹	۸۲	۸۸	۷۶	۹۰
۳	۸۲	۶۸	۷۳	۷۱	۸۱

منبع انحرافات	مجموع توان دوم انحرافات SS	درجه آزادی d.f	واریانس میانگین انحرافات MS	F
بین نمونه‌ها	۲۵۰	K-1=۲	۱۲۵	۳/۳۵
درون نمونه‌ها	۴۴۸	N-k=۱۲	۳۷/۳۳	
کل	۶۹۸	N-1=۱۴		

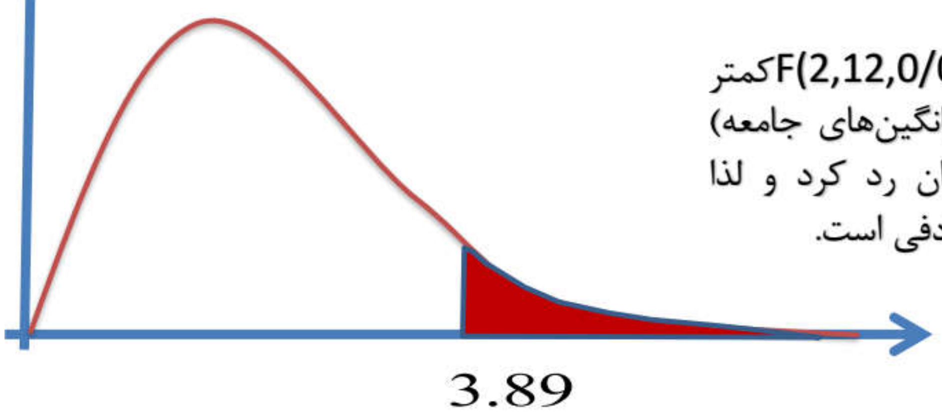
$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_1 : حداقل یکی از μها با بقیه اختلاف دارد.

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{T^2}{N} = 96698 - \frac{(1200)^2}{15} = 698$$

$$SSR = \sum_{i=1}^{k}\frac{T_i^2}{n} - \frac{T^2}{N} = (\frac{400^2}{5} + \frac{425^2}{5} + \frac{375^2}{5}) - \frac{(1200)^2}{15} = 250$$

$$SSE = SST - SSR = 698 - 250 = 448$$



چون F=3/35 از F(2,12,0/05)=3/89 کمتر است. فرض صفر (فرض تساوی میانگین‌های جامعه) در سطح معنی دار ۵/. را نمی‌توان رد کرد و لذا اختلاف در میانگین‌ها شانسی و تصادفی است.

$$F_{2,12,0.05} = 3.89$$

فصل سیزدهم

ضریب همبستگی و رگریسون

www.mbdm.ir

مثال

داده‌های زیر هزینه تبلیغات شرکتی را همراه با تعداد فروش محصولش را در ۹سال مختلف نشان می‌دهد. خط رگرسیون آن را پیدا کنید.

سال	x	y	X^2	y^2	xy
۸۶	۳	۱۱	۹	۱۲۱	۳۳
۸۷	۵	۲۰	۲۵	۴۰۰	۱۰۰
۸۸	۴	۱۶	۱۶	۲۵۶	۶۴
۸۹	۷	۲۴	۴۹	۵۷۶	۱۶۸
۹۰	۹	۲۶	۸۱	۶۷۶	۲۳۴
۹۱	۶	۱۵	۳۶	۲۲۵	۹۰
۹۲	۵	۲۱	۲۵	۴۴۱	۱۰۵
۹۳	۴	۱۸	۱۶	۳۲۴	۷۲
۹۴	۸	۲۷	۶۴	۷۲۹	۲۱۶
جمع کل	۵۱	۱۷۸	۳۲۱	۳۷۴۸	۱۰۸۲

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{51}{9} = 5.667 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{178}{9} = 19.778$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 321 - 9(5.667)^2 = 321 - 289.034 = 31.966$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 3748 - 9(19.778)^2 = 227.476$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = 1082 - 9(5.667)(19.778) = 73.263$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}} = \frac{73.263}{31.966} = 2.292$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 19.778 - 2.292(5.667) = 6.789$$

$$\hat{y} = 6.789 + 2.282x$$

$$SSE = S_{yy} - \hat{\beta}^2 S_{xx} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 227.476 - \frac{(73.263)^2}{31.966} = 59.564$$

پیش‌بینی امید ریاضی پاسخ

خط رگرسیون برای هر سطح معینی از متغیر مستقل، امید ریاضی متغیر پاسخ (وابسته) را ارائه می‌دهد.

$$E(y/x^*) = \alpha + \beta x^*$$

از این‌رو یک برآورد کننده ناااریب برای $E(y/x^*)$ عبارتست از: $y^* = \hat{\alpha} + \hat{\beta} x^*$

مثال: معادله رگرسیونی به صورت $\hat{y} = 2 + 0.387x$ محاسبه شده است. امید ریاضی **y** زمانی که **x** برابر با ۴ می‌باشد را به دست آورید.

$$y^* = 2 + 0.387(4) = 3.548$$

ضریب همبستگی

روش رگرسیون خطی، زمانی مناسب و قابل کاربرد است که بین متغیرهای تصادفی x و y یک رابطه خطی و قوی برقرار باشد.

برای تعیین شدت رابطه خطی بین x و y از یک معیار عددی به نام ضریب همبستگی استفاده می‌شود.

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y} \qquad Cov(x, y) = E\left[(x - \mu_x)(y - \mu_y)\right]$$

به منظور محاسبه ρ، توزیع جامعه باید معلوم باشد. در صورتی که توزیع جامعه مشخص نباشد، ρ با استفاده از نمونه‌ای تصادفی که از جامعه استخراج شده، برآورد می‌گردد.

ضریب همبستگی نمونه‌ای:

آماره‌ای که جهت برآورد ضریب همبستگی خطی x و y به کار گرفته می‌شود با r نشان داده می‌شود و آن را ضریب همبستگی نمونه‌ای یا ضریب همبستگی پیرسون می‌نامیم.

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum x_i\sum y_i}{\sqrt{\left[n\sum x_i^2-(\sum x_i)^2\right]\left[n\sum y_i^2-(\sum y_i)^2\right]}}$$

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

تعبیر و تفسیر مقدار r

ضریب همبستگی شدت رابطه و همچنین نوع رابطه (مستقیم یا معکوس) را نشان می‌دهد.

با توجه به فرمول ارائه شده برای محاسبه r می‌توان نتیجه گرفت که مقدار r می‌تواند مقادیری بین ۱- تا ۱+ را اختیار کند.

$$-1 \leq r \leq +1$$

همبستگی مستقیم ($0 < r \leq +1$): در اینصورت x و y دارای همبستگی مستقیم هستند، یعنی با افزایش x، y نیز افزایش می‌یابد. به عبارتی شیب خط رگرسیون مثبت است.

همبستگی معکوس ($-1 \leq r < 0$): در اینصورت x و y دارای همبستگی منفی هستند، یعنی با کاهش x، y افزایش می‌یابد. به عبارتی شیب خط رگرسیون منفی است.

ناهمبستگی (r=0): در اینصورت x و y دارای همبستگی خطی ندارند، و نمی‌توان یک رابطه خطی بین x و y به دست آورد.