جزوه شماره ۱۵

آزمون تفاضل دو میانگین

در بسیاری از مطالعات، یکی از اهداف مطالعه، مقایسه دو جمعیت مستقل از طریق مقایسه میانگین آنهاست. بدین منظور انجام آزمون آماری جهت تائید یا رد ادعای محقق در قالب فرضیه های آماری اجتناب ناپذیر است. مراحل انجام آزمون مقایسه دو میانگین مانند آزمون میانگین با بیان فرضیه های آماری و سپس انتخاب نمونه های n_1 و n_2 تایی از دو جمعیت مورد مطالعه آغاز می شود. در این آزمونها نیز مانند آزمون میانگین وضعیت واریانسهای دو جمعیت و تعداد نمونه اخذ شده از دو جمعیت می توانند در توزیع آماری شاخص آماری آزمون موثر باشند.

گام اول در انجام آزمون مقایسه دو میانگین، مانند سایر آزمونها بیان فرضیه های پژوهشی است. فرضیه های پژوهشی می توانند به یکی از صورتهای زیر باشند.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

فرضیه دو طرفه یا دو سویه یا دو دنباله

$$H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_1 > \mu_2$$

$$H_0: \mu_1 \geq \mu_2$$
$$H_1: \mu_1 < \mu_2$$

فرضیه های یکطرفه یا یک سویه یا یک دنباله

حالت اول:

واریانس متغیر مورد مطالعه در دو جمعیت معلوم است

در شرایط خاص بر اساس مطالعات گذشته و اطلاعات موجود از جمعیتها ممکن است واریانس و به دنبال آن انحراف معیار متغیر مورد مطالعه در دو جامعه معلوم باشد یعنی σ_1^2 و σ_2^2 معلوم باشند. در این حالت بدون توجه به تعداد نمونه، توزیع آماری شاخص آماری آزمون، توزیع نرمال استاندارد خواهد بود و فرمول

این شاخص برابر است با :

$$Z=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)_{H_0}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

که در این فرمول $\bar{x}_1$ و $\bar{x}_2$ میانگین متغیر مورد مطالعه براساس نمونه‌های n_1 و n_2 تایی که به صورت تصادفی از دو جمعیت مورد مطالعه اخذ شده اند می باشند. σ_1^2 و σ_2^2 به ترتیب واریانس‌های متغیر مورد نظر در دو جمعیت و n_1 و n_2 نمونه‌های گرفته شده از دو جمعیت هستند. در این فرمول $(\mu_1-\mu_2)_{H_0}$، تفاضل دو میانگین براساس آنچه که در قالب فرض صفر (H_0) بیان شده می باشد که این مقدار در فرضیه‌های بیان شده در صفحه قبل برابر با صفر خواهد بود. اما در شرایط کلی‌تر فرض صفر و فرض محقق می تواند به صورت زیر باشد

$$H_0:\ \mu_1-\mu_2=a$$
$$H_1:\ \mu_1-\mu_2\neq a$$

$$H_0:\ \mu_1-\mu_2\leq a$$
$$H_1:\ \mu_1-\mu_2> a$$

$$H_0:\ \mu_1-\mu_2\geq a$$
$$H_1:\ \mu_1-\mu_2< a$$

در این فرضیه‌ها a یک مقدار ثابت می باشد که در صورتی که فرضیه‌های آماری به صورت‌های رو برو باشند، $(\mu_1-\mu_2)_{H_0}=a$ خواهد بود.

بقیه مراحل انجام آزمون آماری مانند مراحل گفته شده در صفحات قبل می باشد

مثال : نتایج حاصل از نمونه‌های ۵۰ و ۶۰ تایی از خانمها و آقایان کارمند که به صورت تصادفی از کارمندان یک سازمان بزرگ انتخاب شده اند نشان می دهد که میانگین سنوات خدمت خانمها و آقایان به ترتیب ۱۶ و ۱۸ سال است. لازم به توضیح است که براساس تجربیات و مطالعات گذشته واریانس

سنوات خدمت کارمندان زن و مرد در یک سازمان $\sigma_1^2=\sigma_2^2=25$ می باشد

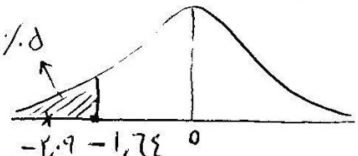
آیا بر اساس این نتایج می توان در سطح خطای $\alpha=5\%$ ادعا کرد میانگین سنوات خدمت کارمندان زن از میانگین سنوات خدمت مردان کارمند کمتر است؟

بر اساس ادعای بیان شده فرضیه های این مسئله عبارتند از

$$H_0: \mu_1 \geq \mu_2$$
$$H_1: \mu_1 < \mu_2$$

مقدار شاخص آماری که توزیع آن Z می باشد برابر است با

$$Z=\frac{16-18}{\sqrt{\frac{25}{50}+\frac{25}{60}}}=\frac{-2}{\sqrt{0.9166}}=\frac{-2}{0.957}=-2.09$$



با توجه به جدول توزیع نرمال استاندارد در سطح خطای $\alpha=0.05$، عدد بحرانی -1.64 خواهد بود که چون $-2.09<-1.64$ است، آزمون معنی دار و در نتیجه بر اساس نتایج حاصله می توان ادعای محقق را پذیرفت یعنی H_0 رد و H_1 تأیید می شود.

حالت دوم:

واریانس متغیر مورد مطالعه در دو جمعیت معلوم نیست، اما نمونه ها بزرگ هستند. پارامترهای جمعیت مانند واریانس جمعیت معمولاً نامعلوم است ولی می توان با استفاده از نمونه n تایی که از جمعیت بطور تصادفی انتخاب می شود، این پارامتر را برآورد کرد یعنی با استفاده از رابطه $S^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ ، S^2 را به عنوان برآوردی از σ^2 ، بدست آورد و در محاسبه شاخص آماری آزمون از آن استفاده کرد. در این حالت نیز σ_1^2 و σ_2^2 معلوم نیستند، اما می توان با استفاده

از S_1^2 و S_2^2، این دو پارامتر را تخمین زد. البته ذکر این نکته ضروریست که علی‌رغم آنکه واریانس‌های دو جمعیت معلوم نیستند، بالتبع فرض کرد و حتی در مواردی با استفاده از آزمون تساوی واریانس‌ها، همگن بودن دو جمعیت را تأیید کرد یعنی باید فرض کرد و یا ثابت کرد که $\sigma_1^2=\sigma_2^2$ می‌باشد. از طرف دیگر با توجه به اینکه این آزمون یک آزمون پارامتریک محسوب می‌شود، توزیع آماری داده‌های جمع‌آوری شده از دو جمعیت، بالتبع از توزیع آماری مشخصی (مشخصاً در این آزمون توزیع نرمال) برخوردار باشد که برای اثبات این موضوع نیز بالتبع با استفاده از آزمونهای مناسب و یا روشهای دیگر آماری نرمال بودن توزیع داده‌ها را نشان داد.

در هر حال، پس از تأیید موارد فوق، با توجه به اینکه واریانس‌های دو جمعیت معلوم نیستند توزیع آماری شاخص آماری آزمون، توزیع t با (n_1+n_2-2) درجه آزادی است یعنی

$$t=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)_{H_0}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

که در این حالت با توجه به بزرگ بودن تعداد نمونه‌ها (بالتبع هر دو نمونه از ۳۰ بزرگتر باشند) و همچنین خاصیت توزیع t، می‌توان برای مشخص کردن عدد بحرانی، از جدول توزیع Z (نرمال استاندارد) استفاده کرد.

مثال: نتایج حاصل از نمونه‌های ۸۰ و ۱۰۰ تایی از خانوارهای دو شهر A و B در رابطه با هزینه‌های ماه فرهنگی در دوران سالی آنها به شرح زیر است.

	شهر A	شهر B
n	۸۰	۱۰۰
$\bar{x}$	۱۳۸	۱۳۶/۵
S	۵	۸

آیا می‌توان براساس این نتایج در سطح خطای $\alpha=0.01$ ادعا کرد، هزینه‌های فرهنگی دوران سالی خانوارهای شهر A بیشتر از این هزینه‌ها در شهر B است؟

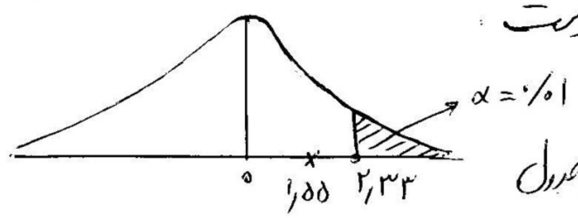
با توجه به ادعای مطرح شده، فرضیه‌ها عبارتند از:

$$H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_1 > \mu_2$$

مقدار شاخص آماری آزمون برابر است با:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{138 - 136.5}{\sqrt{\frac{5^2}{80} + \frac{8^2}{100}}} = \frac{1.5}{\sqrt{0.31 + 0.64}} = \frac{1.5}{\sqrt{0.95}} = 1.55$$

که با توجه به بزرگ بودن تعداد نمونه‌ها، می‌توان از جدول توزیع نرمال استاندارد برای تعیین عدد بحرانی استفاده کرد. به عبارت دیگر می‌توان فرض کرد توزیع شاخص آماری تحت نرمال استاندارد (Z) است:



با توجه به عدد بحرانی بدست آمده از جدول توزیع نرمال استاندارد (2.33)، چون مقدار شاخص آماری آزمون در ناحیه بحرانی (ناحیه رد H_0) قرار ندارد، می‌توان نتیجه گرفت که در سطح خطای $\alpha = 0.01$ آزمون معنی‌دار نیست. به عبارت دیگر در این سطح خطا نمی‌توان ادعای مطرح شده مبنی بر بیشتر بودن هزینه‌های مسافرتی درون‌شهری خانواده‌های شهر A، نسبت به شهر B را پذیرفت.

حالت سوم:

واریانس متغیر مورد مطالعه در دو جمعیت معلوم نیست، اما نمونه‌ها کوچک هستند. در این حالت با توجه به کوچک بودن تعداد نمونه‌ها ($n_1 < 30$ و $n_2 < 30$)، شاخص آماری آزمون دارای توزیع t با $df = n_1 + n_2 - 2$ درجه آزادی است

شاخص آماری آزمون در این حالت برابر است با:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \quad \text{و} \quad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

واریانس ادغام شده
(Pooled Variance)

در حالت نیز فرض نرمال بودن متغیر در جمعیت‌های مورد مطالعه و همچنین همگن بودن دو جمعیت، به عبارت دیگر مساوی بودن واریانس‌های متغیر مورد مطالعه در دو جامعه، باید برقرار باشد.

مثال: نتایج حاصل از نمونه‌های ۱۰ و ۱۲ تایی از مدیران بانکهای دولتی و غیر دولتی در رابطه با میزان آشنایی آنها با روشهای نوین بانکداری الکترونیکی براساس یک روش سنجش کمی استاندارد به شرح زیر است.

	مدیران دولتی	مدیران غیردولتی
n	۱۰	۱۲
$\bar{x}$	۸۹٫۵	۹۱
S	۷	۵

فرض کنید توزیع آماری امتیازات کسب شده نرمال و واریانس‌های متغیر در دو جمعیت با هم مساوی هستند.

آیا در سطح خطای $\alpha = 0.05$ می‌توان ادعا کرد تفاوت معنی‌داری بین میانگین امتیاز مربوط به میزان آشنایی مدیران بانکهای دولتی و غیر دولتی از روشهای نوین بانکداری وجود دارد؟

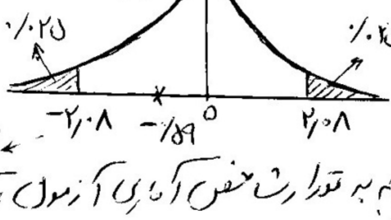
با توجه به ادعای مطرح شده، فرضیه‌های تحقیق عبارتند از

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

با توجه به تعداد نمونه‌ها ابتدا واریانس ادغام شده محاسبه می‌شود

$$S_p^2 = \frac{(10-1)7^2 + (12-1)5^2}{10+12-2} \Rightarrow S_p^2 = \frac{9 \times 49 + 11 \times 25}{20} = 35.8$$

$$t = \frac{89.5 - 91}{\sqrt{\frac{35.8}{10} + \frac{35.8}{12}}} = \frac{-1.5}{\sqrt{6.56}} = -0.59$$



$df = 10 + 12 - 2 = 20$

از جدول توزیع t با ۲۰ درجه آزادی

با توجه به مقدار شاخص آماری آزمون، آزمون معنی‌دار نیست یعنی H_0 رد نمی‌شود و در نتیجه میانگین‌ها با هم مساوی هستند.