جزوه شماره ۴

فاصله اطمینان - تفاضل میانگین دو جامعه :

یکی از مباحثی که در مقایسه دو جمعیت مطرح است، مقایسه میانگین متغیر مورد نظر در دو جمعیت با استفاده از حدود اطمینان و یا فاصله اطمینان است. همچنانکه در بحث مربوط به فاصله اطمینان میانگین جمعیت مورد بحث قرار گرفت، در گام نخست در مقایسه دو جمعیت میانگینهای متغیر مورد نظر با استفاده از نمونه های n_1 و n_2 تایی که از دو جمعیت گرفته شده اند، محاسبه و برآورد نقطه ای تفاضل دو میانگین یعنی $\bar{x}_1 - \bar{x}_2$ برای تخمین تفاضل میانگینهای دو جمعیت یعنی $\mu_1 - \mu_2$ بکار برده می شود. امّا با همان توجیه که در مورد یک میانگین گفته شد، می توان برآورد نقطه ای تفاضل دو میانگین را به برآورد فاصله ای تبدیل کرد. بنابراین لازم قبل از تعیین این حدود اطمینان، توزیع آماری تفاضل دو میانگین (توزیع نمونه برداری تفاضل دو میانگین) مشخص شود.

فرض کنید جامعه اول دارای توزیع نرمال با میانگین μ_1 و انحراف معیار σ_1 و جامعه دوم نیز نرمال با میانگین μ_2 و انحراف معیار σ_2 باشند. برای مقایسه میانگینها، نمونه های n_1 و n_2 تایی از دو جمعیت به صورت تصادفی انتخاب می شوند. اگر $\bar{x}_1$ میانگین نمونه اول و $\bar{x}_2$ میانگین نمونه دوم باشد، $(\bar{x}_1 - \bar{x}_2)$ دارای توزیع نرمال با میانگین $(\mu_1 - \mu_2)$ و انحراف معیار $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ خواهد بود. ذکر این نکته لازم است که حتی اگر توزیع جمعیتهای اول و دوم نرمال نباشد امّا نمونه های اخذ شده از دو جمعیت به اندازه کافی بزرگ باشند (بزرگتر از ۳۰)، $(\bar{x}_1 - \bar{x}_2)$ همچنان بطور تقریبی دارای توزیع نرمال با میانگین و انحراف معیار فوق الذکر می باشد.

مثال - فرض کنید میانگین و انحراف معیار دو جمعیت به ترتیب، $\mu_1 = 50$ و $\sigma_1 = 15$ و $\mu_2 = 30$ با انحراف معیار $\sigma_1 = 15$ باشند. اگر نمونه های ۱۰۰ تایی از این دو جمعیت انتخاب کنیم یعنی $n_1 = 100$ و $n_2 = 100$ (لازم نیست تعداد نمونه اخذ شده از دو نمونه مساوی باشند)

دنباله مثال صفحه قبل:

احتمال بیشتر باشد $\bar{x}_1-\bar{x}_2>25$ چقدر است یعنی $P(\bar{x}_1-\bar{x}_2>25)=?$

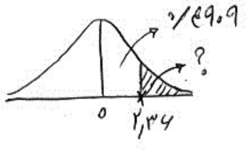
همچنانکه قبلاً گفته شد توزیع نمونه برداری $\bar{x}_1-\bar{x}_2$ نرمال با میانگین و انحراف معیار بشرح زیر است

$$\mu_{(\bar{x}_1-\bar{x}_2)}=\mu_1-\mu_2=50-30=20$$

$$\sigma_{(\bar{x}_1-\bar{x}_2)}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=\sqrt{\frac{15^2}{100}+\frac{15^2}{100}}=\sqrt{4.5}=2.12$$

که برای محاسبه احتمال فوق ما باید با استفاده از استاندارد کردن داده‌های نرمال و استفاده از جدول توزیع نرمال استاندارد، بدینگونه عمل کرد

$$P(\bar{x}_1-\bar{x}_2>25)=P\left(\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}>\frac{25-20}{2.12}\right)=P(Z>2.36)$$



$$P(\bar{x}_1-\bar{x}_2>25)=P(Z>2.36)=\%5-\%4909$$
$$=\%0091$$

فاصله اطمینان تفاضل دو میانگین وقتی انحراف معیار یا واریانس دو جمعیت معلوم هستند:

باتوجه به معلوم بودن σ_1 و σ_2، توزیع نمونه برداری $\bar{x}_1-\bar{x}_2$ نرمال خواهد بود. بنابراین احتمال اینکه مقدار استاندارد شده تفاضل دو میانگین در فاصله مشخصی قرار گیرد و این احتمال برابر با $(1-\alpha)$ باشد را می‌توان به این صورت نوشت.

$$P\left(Z_{1-\frac{\alpha}{2}}\leq\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leq Z_{\frac{\alpha}{2}}\right)=1-\alpha$$

که با اندکی محاسبه جبری رابطه فوق را می‌توان به صورت زیر نوشت.

$$P\left(\bar{x}_1-\bar{x}_2+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leq\mu_1-\mu_2\leq\bar{x}_1-\bar{x}_2+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

که $Z_{\frac{\alpha}{2}}$ و $Z_{1-\frac{\alpha}{2}}$ قرینه همدیگر هستند. بدین ترتیب مقدار خطای برآورد یعنی d برابر است با

$$d=Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

و فاصله اطمینان تفاضل دو میانگین برابر خواهد بود با

$$\bar{x}_1-\bar{x}_2 \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad \text{یا} \quad \left[\bar{x}_1-\bar{x}_2-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \; , \; \bar{x}_1-\bar{x}_2+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

مثال: در مقایسه هزینه ماهیانه مواد غذایی خانوارهای دو شهر A و B، تجربیات گذشته نشان می دهد که انحراف معیار هزینه های ماهیانه مواد غذایی مربوط به این دو شهر به ترتیب $\sigma_A = ۲۷۰۰$ تومان و $\sigma_B = ۳۵۰۰۰$ تومان است. نتایج حاصل از نمونه های ۸۰ و ۱۰۰ تایی از خانوارهای این دو شهر نشان می دهد که در شهر A میانگین هزینه ها برابر ۴۲۵۰۰۰ تومان و در شهر B برابر با ۴۴۰۰۰۰ تومان است. فاصله اطمینان ۹۵ درصد برای تفاضل میانگین هزینه های ماهیانه مواد غذایی خانوارهای این دو شهر چقدر است؟

$$\%۹۵ \Rightarrow \alpha = 1-\%۹۵ \rightarrow \alpha = \%۵ \;,\; \frac{\alpha}{2} = \%۲۵$$

با استفاده از جدول توزیع نرمال استاندارد

$$Z_{\frac{\alpha}{2}} = ۱,۹۶$$

با توجه به اینکه انحراف معیار در جامعه معلوم است، فاصله اطمینان ۹۵٪ برابر است با

$$\bar{x}_1-\bar{x}_2 \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \Rightarrow ۴۲۵-۴۴۰ \pm ۱,۹۶\sqrt{\frac{(۲۷)^2}{۸۰}+\frac{(۳۵)^2}{۱۰۰}}$$

$$۳۵ \pm ۱,۹۶\sqrt{۲۱,۴۶}$$

(از ۳ صفر صرفنظر شده است)

$$۳۵ \pm ۹,۰۶$$

پس فاصله اطمینان ۹۵٪ برابر خواهد بود با $[۲۵,۹۴ \;,\; ۴۴,۰۶]$

به عبارت دیگر با اطمینان ۹۵٪ تفاضل میانگین هزینه های ماهیانه مواد غذایی در این دو شهر بین ۲۵۹۴۰ و ۴۴۰۶۰ تومان خواهد بود.

البته تفسیر فاصله اطمینان تفاضل دو پارامتر، بصورت فوق مفهوم نیست و در عمل کمتر بکار برده می شود. زمانی که فاصله اطمینان برای تفاضل دو پارامتر از یک جامعه محاسبه می شود، فاصله بدست آمده یکی از سه حالت زیر خواهند بود:

۱- اگر هر دو کرانه پائینی و بالایی فاصله بدست آمده مثبت باشند، می‌توان به این صورت نتیجه را تفسیر کرد که با اطمینان $(1-\alpha)$ درصد و یا در سطح خطای α، پارامتر مورد نظر از جمعیت اول از پارامتر مورد نظر از جمعیت دوم بزرگتر است.

۲- اگر کرانه پائینی منفی و کرانه بالایی مثبت باشند، می‌توان به این صورت نتیجه را تفسیر کرد که با اطمینان $(1-\alpha)$ درصد و یا در سطح خطای α، تفاوت معنی‌داری بین پارامتر مورد نظر جمعیت اول و جمعیت دوم وجود ندارد.

۳- اگر کرانه پائینی و کرانه بالایی فاصله اطمینان محاسبه شده، هر دو منفی باشند، می‌توان نتیجه را به این صورت تفسیر کرد که با اطمینان $(1-\alpha)$ درصد و یا در سطح خطای α، پارامتر مورد نظر از جمعیت اول بطور معنی‌داری از پارامتر مورد نظر از جمعیت دوم کوچکتر است.

در مثال فوق با توجه به اینکه هر دو کرانه مربوط به فاصله اطمینان بدست آمده مثبت هستند، می‌توان چنین نتیجه گرفت که با پذیرش ۵٪ خطا و یا در سطح اطمینان ۹۵٪ میانگین هزینه‌های ماهیانه مواد غذایی در شهر A از این مقدار در شهر B بیشتر است.

فاصله اطمینان تفاضل دو میانگین، وقتی انحراف معیار یا واریانس دو جمعیت معلوم نیستند:

در اکثر مواقع پارامترهای جمعیت نامعلوم هستند و این پارامترها با استفاده از نمونه‌های n تایی برآورد می‌شوند. بنابراین دور از انتظار نیست که واریانس‌های دو جمعیت و در نتیجه انحراف معیارهای مربوط به دو جمعیت معلوم نباشند. که در این حالت با استفاده از نمونه‌های n_1 و n_2 تایی واریانس‌های دو جمعیت با استفاده از رابطه زیر برآورد می‌شوند

$$S_1^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

البته علی رغم اینکه واریانس‌های دو جمعیت معلوم نیستند، امّا فرض می‌کنیم این دو واریانس با هم مساوی هستند یعنی $\sigma_1^2=\sigma_2^2$ ، بعبارت دیگر دو جمعیت همگن هستند. در این صورت با دو حالت مواجه خواهیم بود.

حالت اول : تعداد نمونه‌های اخذ شده از دو جمعیت بزرگ هستند $n_1 , n_2 > 30$

در این حالت توزیع آماری مقدار استاندارد شده تفاضل دو میانگین، توزیع t با n_1+n_2-2 درجه آزادی خواهد بود یعنی

$$\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

که در این صورت با توجه به آنچه در قسمت قبل بیان شد، فاصله اطمینان برای تفاضل دو میانگین برابر است با :

$$\bar{x}_1-\bar{x}_2 \pm t_{\frac{\alpha}{2}} \times \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

که s_1^2 و s_2^2 ، واریانس‌های بدست آمده از نمونه‌های n_1 و n_2 تایی است و $t_{\frac{\alpha}{2}}$ با استفاده از جدول توزیع t بدست می‌آید. امّا این نکته را باید توجه داشت که با توجه به خاصیت توزیع t ، چون تعداد نمونه‌ها بزرگ هستند، می‌توان بجای استفاده از جدول توزیع t از جدول توزیع نرمال استاندارد استفاده کرد. یعنی در این حالت $t_{\frac{\alpha}{2}} = Z_{\frac{\alpha}{2}}$.

مثال : نتایج حاصل از نمونه‌های ۵۰ و ۶۰ تایی از کارمندان زن و کارمندان مرد یک سازمان بشرح زیر است.

	زن	مرد
n	۵۰	۶۰
$\bar{x}$	۱۲	۱۵
s	۳	۴

متغیر مورد نظر سنوات خدمت این کارمندان است. یک فاصله اطمینان ۹۹%

دنباله مثال صفحه قبل:

برای تفاضل سنوات خدمت کارکنان زن و مرد این سازمان یعنی برای $\mu_1-\mu_2$ را حساب کنید. در این مثال μ_1، میانگین سنوات کارکنان زن و μ_2 میانگین سنوات خدمت مرد می‌باشد. با توجه به اینکه انحراف معیارهای بدست آمده با استفاده از نمونه‌های ۵۰ و ۶۰ تایی بدست آمده‌اند (نمونه‌ها بزرگ هستند) فاصله اطمینان را می‌توان بدین صورت حساب کرد

$$\bar{x}_1-\bar{x}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_2}+\frac{S_2^2}{n_2}} \Rightarrow 12-15 \pm 2.58\sqrt{\frac{9}{50}+\frac{16}{60}}$$

$$-3 \pm 2.58\sqrt{0.4466}$$

$$-3 \pm 2.58 \times 0.67$$

$$-3 \pm 1.73$$

بنابراین فاصله اطمینان بدست آمده برابر است با $[-4.73, -1.27]$

یعنی بر اساس آنچه قبلاً گفته شد، با اطمینان ۹۹٪ یا در سطح خطای ۱٪ می‌توان نتیجه گرفت که میانگین سنوات خدمت کارکنان زن از میانگین سنوات خدمت کارکنان مرد بطور معنی‌داری کمتر است.

لازم به توضیح است که در این مسئله با توجه به بزرگ بودن تعداد نمونه‌ها به جای تعیین مقدار $t_{\frac{\alpha}{2}}$ از جدول t با $50+60-2 = 108$ درجه آزادی، می‌توان از جدول توزیع نرمال استاندارد استفاده کرد

حالت دوم: تعداد نمونه‌های اخذ شده از دو جمعیت کوچک هستند: $n_1, n_2 < 30$

در این حالت نیز توزیع آماری مقدار استاندارد شده تفاضل دو میانگین، توزیع t با n_1+n_2-2 درجه آزادی خواهد بود، اما با توجه به اینکه تعداد نمونه‌ها کوچک هستند، در فرمول مربوطه بجای استفاده از واریانس‌های بدست آمده از نمونه‌های n_1 و n_2 تایی، از ادغام این دو واریانس استفاده می‌شود. به عبارت دیگر ابتدا واریانس ادغام شده مربوطه

دو جمعیت با استفاده از رابطه زیر محاسبه می‌شود:

$$s_p^۲ = \frac{(n_۱-۱)\, s_۱^۲+(n_۲-۱)\, s_۲^۲}{n_۱+n_۲-۲}$$

Pooled Variance

که در این فرمول $s_۱^۲$ و $s_۲^۲$ واریانس‌های بدست آمده از نمونه‌های $n_۱$ و $n_۲$ تایی هستند.

با توجه به آنچه گفته شد، فاصله اطمینان برای تفاضل دو میانگین برابر خواهد بود با:

$$\bar{x}_۱ - \bar{x}_۲ \pm t_{\frac{\alpha}{۲}} \sqrt{\frac{s_p^۲}{n_۱}+\frac{s_p^۲}{n_۲}}$$

درجه آزادی توزیع t در این رابطه $df = n_۱+n_۲-۲$ خواهد بود.

که می‌توان از $s_p^۲$ فاکتور گرفت و رابطه را به صورت زیر نوشت

$$\bar{x}_۱ - \bar{x}_۲ \pm t_{\frac{\alpha}{۲}} \times s_p \sqrt{\frac{۱}{n_۱}+\frac{۱}{n_۲}}$$

مثال: نتایج حاصل از نمونه‌های ۱۲ و ۱۵ تایی از مدیران بانکهای خصوصی و دولتی در رابطه با میزان اطلاعات آنها از بانکداری الکترونیکی بر اساس یک روش استاندارد سنجش مدیران بانکها، بشرح زیر است.

	خصوصی	دولتی
n	۱۲	۱۵
$\bar{x}$	۱۸۴	۱۷۹
s	۵	۴

فاصله اطمینان ۹۵ و ۹۹ درصدی برای تفاضل میانگین میزان اطلاعات مدیران بانکهای دولتی و خصوصی از بانکداری الکترونیک را حساب کنند.

(در این مثال فرض می‌کنیم $\sigma_۱^۲=\sigma_۲^۲$)

با توجه به کوچک بودن تعداد نمونه‌ها ابتدا $s_p^۲$ محاسبه می‌شود

$$s_p^۲ = \frac{(۱۲-۱)۲۵+(۱۵-۱)۱۶}{۱۲+۱۵-۲} = \frac{۲۷۵+۲۲۴}{۲۵} = \frac{۴۹۹}{۲۵} = ۱۹{,}۹۶ \Rightarrow s_p = ۴{,}۴۷ \,,\quad df = ۲۵$$

$$\%۹۵ \Rightarrow ۱۸۴-۱۷۹ \pm ۲{,}۰۶ \times ۴{,}۴۷\sqrt{\frac{۱}{۱۲}+\frac{۱}{۱۵}} \Rightarrow ۵ \pm ۳{,}۵۶ \Rightarrow [۱{,}۴۴ , ۸{,}۵۶]$$

با پذیرش ۵٪ خطا میانگین اطلاعات مدیران بانکهای خصوصی از میانگین اطلاعات مدیران دولتی بیشتر است

$$\%۹۹ \Rightarrow ۱۸۴-۱۷۹ \pm ۲{,}۷۸ \times ۴{,}۴۷\sqrt{\frac{۱}{۱۲}+\frac{۱}{۱۵}} \Rightarrow ۵ \pm ۴{,}۸۱ \Rightarrow [۰/۰۹ , ۹{,}۸۱]$$

با پذیرش ۱٪ خطا نیز میانگین اطلاعات مدیران بانکهای خصوصی از میانگین اطلاعات مدیران دولتی بیشتر است