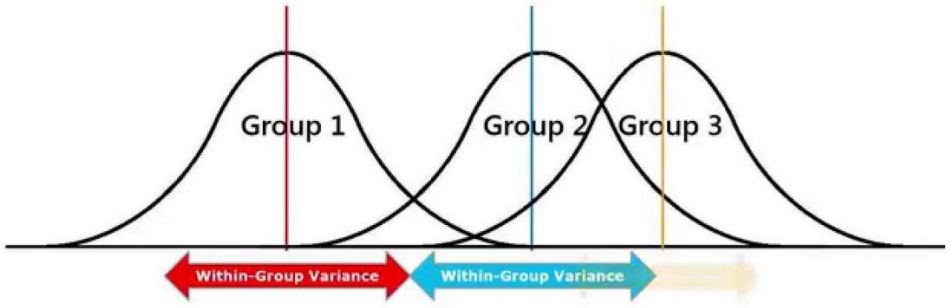
ضروریات آمار برای علم داده

تحلیل واریانس

تحلیل واریانس



- واریانس کل داده‌ها براساس یک یا چند متغیر عامل به دو یا چند بخش تفکیک شده است.
- براساس آزمون‌های مربوط به واریانس می‌توان همگون یا ناهمگون بودن گروه‌ها را آزمود.

شرایط :

- نمونه‌های حاصل از هر گروه یا جامعه، تصادفی و از توزیع نرمال باشند.
- واریانس‌های گروه‌ها برابر ولی نامشخص هستند. (آزمون فلینگر)
- گروه‌ها از یکدیگر مستقل هستند.

تحلیل واریانس

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: & \sim H_0 \end{cases}$$

H_0 : میانگین همه گروه‌ها یکسان است

H_1 : حداقل میانگین یکی از گروه‌ها با بقیه متفاوت است

$$F = \frac{MS_{between}}{MS_{within}} \sim F_{a-1,\ a(n-1)}$$

تحلیل واریانس یک طرفه

تنها یک عامل در گروه‌های مختلف بررسی می‌شود.

مثال:

یک مدیر ارشد فروشگاه‌های زنجیره‌ای ادعا نموده است که خلاقیت‌های متفاوت مدیران، اثرهای متفاوتی بر متوسط میزان فروش می‌گذارد. وی برای تعیین صحت و سقم ادعای خود، ۵ تا از فروشگاه‌هایی که به جز مدیر آن دارای شرایط یکسانی بوده‌اند انتخاب و فروش (به میلیون تومان) هفت روز هفته‌ای را که به تصادف انتخاب کرده است ثبت نموده است. چنانچه داده‌ها از توزیع نرمال پیروی کنند، آن‌ها را در سطح معنی داری ۵ درصد تحلیل نمایید.

مدیر منابع انسانی کارخانه‌ای ادعا کرده است که کارایی سبک اجرای قوانین کار در کارخانه‌ها به طور متوسط با یکدیگر متفاوت است. وی برای اثبات ادعای خود نمونه‌ای از سه نوع سبک اجرای مختلف از یک قانون کار را مشخص و کارایی آن‌ها را در موارد متعددی که در کارخانه‌های مختلف اجرا شده است به وسیله تخصیص نمره به آن‌ها اندازه گیری نموده است. چنانچه نمره‌ها از توزیع نرمال پیروی کنند، داده‌ها را در سطح معنی داری ۵ درصد تحلیل کنید.

		میزان فروش روزانه‌ی یک هفته							
		۱	۲	۳	۴	۵	۶	۷	جمع $(y_{i.})$	میانگین $(\bar{y}_{i.})$
خلاقیت مدیران	۱	۵	۹	۷	۳	۴	۵	۵	$y_{1.} = ۳۸$	$\bar{y}_{1.} = ۵/۴۳$
	۲	۳	۲	۳	۴	۳	۲	۳	$y_{2.} = ۲۰$	$\bar{y}_{2.} = ۲/۸۶$
	۳	۹	۸	۶	۷	۹	۷	۷	$y_{3.} = ۵۳$	$\bar{y}_{3.} = ۷/۵۷$
	۴	۴	۵	۵	۶	۴	۵	۴	$y_{4.} = ۳۳$	$\bar{y}_{4.} = ۴/۷۱$
	۵	۷	۶	۶	۷	۵	۷	۶	$y_{5.} = ۴۴$	$\bar{y}_{5.} = ۶/۲۹$
جمع کل									$y_{..} = ۱۸۸$	
میانگین کل									$\bar{y}_{..} = ۵/۳۷$	

منبع تغییرپذیری SV	مجموع توان‌های دوم SS	درجه‌ی آزادی DF	میانگین توان‌های دوم MS	مقدار آماره‌ی آزمون F_0
تیمار (A)	$SSA=\sum_{i=1}^{a=3}\frac{y_{i.}^2}{n_i}-\frac{y_{..}^2}{N}$ $=(\frac{17^2}{3}+\frac{18^2}{2}+\frac{14^2}{4})-\frac{49^2}{9}$ $=40/55$	$a-1=3-1$ $=2$	$MSA=\frac{SSA}{a-1}$ $=\frac{40/55}{2}=20/275$	$F_0=\frac{MSA}{MSE}$ $=\frac{20/275}{2/61}=7/77$
خطا (E)	$SSE=SST-SSA$ $=56/23-40/55=15/68$	$N-a=9-3$ $=6$	$MSE=\frac{SSE}{N-a}$ $=\frac{15/68}{6}=2/61$	
کل (T)	$SST=\sum_{i=1}^{a=3}\sum_{j=1}^{n_i}y_{ij}^2-\frac{y_{..}^2}{N}$ $=(8^2+5^2+\cdots+5^2)-\frac{49^2}{9}$ $=323-266/77=56/23$	$N-1=9-1$ $=8$		

		تکرار				y
		۴	۳	۲	۱	
سبک اجرای کار	۱		۴	۵	۸	$y_{۱.} = ۱۷$
	۲			۸	۱۰	$y_{۲.} = ۱۸$
	۳	۵	۲	۳	۴	$y_{۳.} = ۱۴$
						$y_{..} = ۴۹$

منبع تغییرپذیری (SV)	مجموع توان‌های دوم (SS)	درجه‌ی آزادی (DF)	میانگین توان‌های دوم (MS)	مقدار آماره‌ی آزمون F_0
خلاقیت مدیران (A)	$SSA=\frac{1}{n}\sum_{i=1}^{a=5}y_{i.}^2-\frac{y_{..}^2}{an}$ $=\frac{1}{7}\left(\begin{matrix}38^2+20^2+53^2\\+33^2+44^2\end{matrix}\right)$ $-\frac{188^2}{5\times7}=87/03$	$a-1$ $=5-1$ $=4$	$MSA=\frac{SSA}{a-1}$ $=\frac{87/03}{4}$ $=21/76$	$F_0=\frac{MSA}{MSE}$ $=\frac{21/76}{1/37}$ $=15/88$
خطا (E)	$SSE=SST-SSA$ $=128/17-87/03$ $=41/14$	$a(n-1)$ $=5(7-1)$ $=30$	$MSE=\frac{SSE}{a(n-1)}$ $=\frac{41/14}{30}$ $=1/37$	
کل (T)	$SST=\sum_{i=1}^{a=5}\sum_{j=1}^{n=7}y_{ij}^2-\frac{y_{..}^2}{an}$ $=\left(5^2+9^2+\cdots+6^2\right)$ $-\frac{188^2}{5\times7}=128/17$	$an-1$ $=5\times7-1$ $=34$		

تحلیل واریانس دو طرفه

اثر دو عامل و همچنین اثر متقابل آن‌ها در گروه‌های مختلف آزمون می‌شود.

مثال

فرض کنید با یک مجموعه داده با ۴۸ سطر و ۳ ستون (متغیر) مواجه هستیم که مربوط به زمان اثر سم به یک نوع خوک با توجه به نوع سم و نحوه درمان خوک‌ها است. می‌خواهیم بدانیم که آیا نوع سم در کاهش طول عمر خوک‌ها موثر است یا خیر.

اگر شیوه درمان هم به عامل نوع سم اضافه کنیم یعنی می‌خواهیم تحت نوع سم، شیوه درمان و اثر متقابل شیوه درمان و نوع سم، میانگین طول عمر خوک‌ها را تحت تاثیر قرار می‌دهد.

آزمون تعقیبی (Post Hock)

تحلیل واریانس نشان می‌دهد که آیا نمونه‌ها متعلق به جامعه هستند یا خیر. در صورتی که فرض صفر رد شود، معلوم نیست که کدام یک از نمونه‌ها در کدام جامعه قرار دارند. به عبارت دیگر، معنی‌دار شدن نسبت Fبه ما نمی‌گوید که اختلاف بین کدام جفت از میانگین‌ها معنی‌دار است. بلکه با آماره F، تنها می‌توانیم پی ببریم که اختلاف بین میانگین گروه‌ها معنی‌دار است.

برابری واریانس‌ها

آزمون لون

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2} \qquad W > F(\alpha;\ k-1, N-k), \quad Reject\ \ H_0$$

ناپارامتری

آزمون کروسکال والیس، آزمونی ناپارامتری برای تحلیل واریانس یک طرفه است

فریدمن، آزمونی ناپارامتری برای تحلیل واریانس دو طرفه است