۱-۶ پارامترهای پراکندگی

شاخص های مرکزی تنها یک منطقه را به عنوان محل تمرکز داده ها معرفی می کنند حال آنکه ممکن است دو دسته داده با پراکندگی های متفاوت دارای میانگین برابری باشند به عبارتی نیاز به شاخصی برای نمایش میزان پراکندگی داده ها خواهیم داشت که در ادامه به معرفی این پارامترها می پردازیم.

۱- دامنه داده‌ها: حاصل تفاضل بزرگترین داده از کوچکترین داده را R یا دامنه داده‌ها گویند.

$$R = X_N - X_1$$

که در آن داده‌ها به فرم $X_{(1)} < X_{(2)} < \cdots < X_N$ با فرض اینکه مرتب شده‌اند.

این شاخص معیار خوبی برای محاسبه میزان پراکندگی داده‌ها نیست. زیرا در محاسبه تنها کوچکترین و بزرگترین داده وارد می‌شود.

مثال ۴ : نمرات ۱۰ دانش آموز در دو کلاس متفاوت در درس ریاضی به قرار زیر است:

کلاس A: ۰ , ۴ , ۴ , ۱۲ , ۱۲ , ۱۲ , ۱۴ , ۱۶ , ۱۶ , ۲۰

کلاس B: ۸ , ۸ , ۹ , ۹ , ۱۲ , ۱۲ , ۱۲ , ۱۳ , ۱۳ , ۱۴

با محاسبه مقادیر $\overline{X}$ و m و M دیده می‌شود که برای هر دو کلاس داریم:

$$\overline{X} = ۱۱ \quad , \quad M = ۱۲ \quad , \quad m = ۱۲$$

اما با توجه به مقادیر تک تک نمرات واضح است که میزان پراکندگی نمرات در دو کلاس کاملاً متفاوت است و برای این منظور نیاز به شاخص‌های پراکندگی می‌باشد تا این مطلب را بتوان با مقایسه آنها نشان داد.

برای دو کلاس مقدار دامنه را محاسبه می‌کنیم:

کلاس A : $R = ۲۰ - ۰ = ۲۰$

کلاس B : $R = ۱۴ - ۸ = ۶$

۲- میانگین انحرافات: فاصله داده x_i از میانگین را انحراف از میانگین داده x_i گویند که به صورت $|x_i - \overline{x}|$ محاسبه می شود. اگر این مقدار را برای تمامی داده‌ها محاسبه کنیم و از نتیجه میانگین بگیریم میانگین انحرافات بدست خواهد آمد که عبارتست از : $D = \frac{۱}{N}\sum_{i=۱}^{K} f_i |x_i - \overline{x}|$.

از آنجا که میانگین انحرافات به تمام داده‌ها وابسته است معیار مناسبی برای سنجش پراکندگی داده‌ها محسوب می‌شود اما بدلیل وجود قدر مطلق در فرمول، محاسبه آن مشکل است و نمی‌توان آنرا ساده نمود بنابر این از واریانس و انحراف استاندارد استفاده می‌کنیم.

۳- واریانس و انحراف استاندارد: میانگین مجذور انحرافات را واریانس می‌نامیم و با نماد ∂ یا $S_b^۲$ نمایش می‌دهیم. که عبارتست از:

$$S_b^۲ = \frac{۱}{N}\sum_{i=۱}^{K} f_i (x_i - \overline{x})^۲$$

در مبحث استنباط آماری واریانس را از مجموع مجذور انحرافات داده ها تقسیم بر $N-۱$ بدست می‌آورند و آنرا با $S^۲$ نمایش می‌دهند.

$$S^۲ = \frac{۱}{N-۱}\sum_{i=۱}^{K} f_i (x_i - \overline{x})^۲$$

در مباحث این درس هر جا صحبت از واریانس می‌کنیم منظور $S^۲$ می‌باشد.

اگر از واریانس جذر بگیریم یعنی $S = \sqrt{s^۲}$ در این صورت S را انحراف استاندارد می‌نامیم که مغیار مناسبی برای سنجش پراکندگی می‌باشد.

همچنین $S^۲$ را می‌توان از فرمول زیر محاسبه نمود:

$$S^۲ = \frac{۱}{N-۱}\left[\sum_{i=۱}^{K} f_i x_i^۲ - \frac{۱}{n}\sum_{i=۱}^{K} f_i x_i)^۲\right] = \frac{۱}{n-۱}\left[\sum_{i=۱}^{K} f_i x_i^۲ - n\overline{x}^۲\right]$$

اثبات:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} f_i (x_i - \bar{x})^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i (x_i^2 - 2\bar{x}_i \bar{x} + \bar{x})$$

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

مثال ۵: مقدار واریانس را برای مثال ۲ محاسبه کنید:

$$\overline{X} = ۱۴/۹ \qquad N = ۴۰$$

$$S^2 = \frac{1}{۴۰-۱}\left[(۵۰\times ۶۴ + ۸\times ۱۲۱ + ۹\times ۱۹۶ + ۹\times (۱۷)^2 + ۶\times ۴۰ + ۳\times (۲۳)^2 - ۴۰\times ۱۴۹)^2\right]$$

۴- ضریب تغییرات : در محاسبه میزان پراکندگی داده‌ها همواره با داده‌هایی سروکار داریم که با مقیاس‌های مختلفی اندازه‌گیری شده‌اند بنابر این برای مقایسه میزان پراکندگی داده‌های بدست آمده از دو جامعه آماری که با مقیاس‌های مختلفی اندازه‌گیری شده‌اند استفاده از واریانس مناسب نمی‌باشد زیرا واریانس به مقیاس اندازه‌گیری وابسته می‌باشد بنابر این از مقیاس مناسبتری به نام ضریب تغییرات استفاده می‌کنیم که از رابطه $CV = \frac{S}{\overline{X}}$ بدست می‌آید و معمولاً با ضریب آن در عدد صد بر حسب درصد بیان می‌شود.

مثال ۶: یک کارخانه تولید لاستیک دو نوع محصول A و B تولید می‌کند. لاستیک نوع A دارای میانگین طول عمر ۲۰۰۰۰ کیلو متر و انحراف استاندارد ۲۰۰۰ کیلو متر می‌باشد و لاستیک نوع B دارای میانگین طول عمر ۱۸۰۰۰ کیلو متر و انحراف استاندارد ۲۰۰ کیلو متر می‌باشد، کدام نوع لاستیک برای خرید مناسب‌تر می‌باشد؟

$$X_A = ۲۰۰۰۰ \qquad \overline{X}_B = ۱۸۰۰۰ \quad \Rightarrow \quad CV_A = \frac{۲۰۰۰}{۲۰۰۰۰} = ۰/۱$$

$$S_A = ۲۰۰۰ \qquad S_B = ۲۰۰ \quad \Rightarrow \quad CV_B = \frac{۲۰۰}{۱۸۰۰۰} = ۰/۰۱$$

$$CV_A = ۰/۱ \times ۱۰۰ = \%۱۰$$

$$CV_A = ۰/۰۱ \times ۱۰۰ = \%۱$$

همانطور که ملاحظه می‌کنید میانگین طول عمر لاستیک دوم از لاستیک اول کمتر است ولی با توجه به اینکه ضریب تغییرات لاستیک دوم کمتر از لاستیک اول است خرید لاستیک دوم به صرفه‌تر می‌باشد.

۱-۷ تغییر مقیاس و مبدأ

داده‌ها را با واحدهای متفاوتی می توان از جامعه آماری جمع آوری نمود به عنوان مثال فرض کنید داده‌های مربوط به وزن ۴۰ نفر از دانشجویان یک کلاس را با واحد کیلوگرم جمع آوری کرده باشید و بخواهید مقادیر میانگین و واریانس را بر حسب پاوند بدست بیاورید برای این منظور نیازی به محاسبه مجدد میانگین و واریانس نمی‌باشد بلکه کافیست از روش تغییر مقیاس و مبدأ استفاده کنید.

۱-۷-۱ تغییر مقیاس

اگر تمامی داده‌ها در عدد a ضرب شوند در این صورت داریم:

$$x_۱, x_۲, \ldots, x_n \quad \rightarrow \quad ax_۱, ax_۲, \ldots, ax_n$$

$$\Rightarrow \overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i \quad \rightarrow \quad \overline{X}_a = \frac{1}{N}\sum_{i=1}^{n} ax_i = a\left(\frac{1}{N}\sum_{i=1}^{n} x_i\right) = a\overline{X} \quad \Rightarrow \quad \overline{X}_a = a\overline{X}$$

به همین ترتیب برای محاسبه واریانس بدست می‌آید:

تمرین ۱ :

$$S_a^2 = a^2 S^2 \rightarrow S_a = |a| S$$

$$S_a = \frac{1}{n-1}\sum_1^n (a x_i - a\bar{x})^2 = a^2 \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2 = a^2 S^2$$

$$\Rightarrow S_a = \sqrt{a^2 S^2} = |a| S$$

تمرین ۲ :

$$X \rightarrow X + b$$

$$\bar{X}_b = \bar{X} + b$$

$$\bar{X}_b = \frac{1}{n}\sum_1^n (x_i + b) = \frac{1}{n}\left[\sum_1^n x_i + \sum_1^n b\right] = \frac{1}{n}\sum_1^n x_i + \frac{b}{n}\sum_1^n (1) = \bar{X} + b\frac{n}{n} = \bar{X} + b$$

$$S_b^2 = S^2$$

$$S_b^2 = \frac{1}{n-1}\sum_1^n (x_i + b - \bar{X} - b)^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2 = S^2$$

تمرین ۳ –

$$X \rightarrow Y = aX + b$$

$$\bar{Y} = a\bar{X} + b \quad , \quad S_Y^2 = a^2 S^2$$

اثبات : $\bar{Y} = \frac{1}{n}\sum_1^n (ax_i + b) = \frac{a}{n}\sum_1^n x_i + \frac{bn}{n} = a\bar{X} + b$

$$S_Y^2 = \frac{1}{n-1}\sum_1^n (ax_i + b - a\bar{x} - b)^2 = \frac{a^2}{n-1}\sum_1^n (x_i - \bar{x})^2 = a^2 S^2$$

استاندارد سازی

مثال : نمره علی از امتحان فیزیک و ریاضی به ترتیب برابر ۴۰ و ۶۰ شده است اگر میانگین نمرات امتحان فیزیک و ریاضی به ترتیب برابر ۲۰ و ۵۰ باشد و انحراف معیار امتحان فیزیک و ریاضی به ترتیب برابر ۱ و ۲ باشد علی کدام درس را بهتر امتحان داده است.

حل: برای اینکه بتوان نمرات دو درس را با یکدیگر مقایسه نمود می‌بایستی ابتدا نمرات را استاندارد سازی نمود و سپس آنها را با یکدیگر مقایسه نمود.

نمرات استاندارد شده علی در درس ریاضی $= \frac{۶۰ - ۵۰}{۲} = ۵$

نمرات استاندارد شده علی در درس فیزیک $= \frac{۴۰ - ۲۰}{۱} = ۲۰$

با وجود اینکه نمره علی در درس فیزیک کمتر از ریاضی می‌باشد اما با استاندارد نمودن نمره دو درس مشاهده می‌کنیم که نمره وی در درس فیزیک بالاتر از درس ریاضی می‌باشد به عبارتی علی درس فیزیک را بهتر از درس ریاضی امتحان داده است.

۱-۷-۲ تغییر مبدأ

در صورتی که به تمام داده‌ها مقدار b را اضافه یا کم کنیم می‌توان نشان داد که مقادیر $\overline{X}$ و S^2 جدید از روابط زیر محاسبه می‌شوند:

$$\overline{X}_b = \overline{X} + b$$

تغییر مبدأ روی واریانس بی تأثیر است

$$S_b^2 = S^2$$

با اعمال همزمان تغییر مبدأ و مقیاس خواهیم داشت:

$$\overline{Y} = a\,\overline{X} + b$$

$$S_a^2 = a^2 S^2$$

مطالب فوق برای میانه و مد نیز صادق می‌باشند و داریم:

$$m' = am + b$$

$$M' = a\,m + b$$

۱-۱۳ استاندارد سازی

از یک جامعه آماری n نمونه $x_1, x_2, \ldots x_n$ بصورت تصادفی انتخاب می‌کنیم بطوریکه میانگین و واریانس نمونه‌ها بترتیب $\overline{x}$ و S_x^2 می‌باشد. با توجه به تغییر مبدأ و مقیاس مقدار هر نمونه را از میانگین نمونه‌ها کم می‌کنیم و حاصل را بر S_x تقسیم می‌کنیم بنابر این داده‌های $y_1 = \frac{x_1 - \overline{x}}{S_x}$, $y_2 = \frac{x_2 - \overline{x}}{S_x}$, $y_n = \frac{x_n - \overline{x}}{S_x}$ را خواهیم داشت.

با محاسبه مقایر میانگین و واریانس داده‌های جدید داریم: $\overline{y} = \frac{\overline{x} - \overline{x}}{S_x} = 0$; $S_y^2 = (\frac{1}{S_x^2})\, S_x^2 = 1$

همانطور که ملاحظه می‌کنید داده‌های جدید دارای میانگین صفر و واریانس ۱ می‌باشند که به آنها داده‌های استاندارد شده می‌گوییم. همینطور اگر داده‌های x_1 تا x_n را با متغییر تصادفی X نمایش دهیم در این صورت $Y = \frac{X - \overline{x}}{S_x}$ فرم استاندارد شده یا صورت معیاری متغیر تصادفی X می‌باشد.

مسائل فصل اول :

۱- دو جامعه با اندازه‌های میانگین $\overline{X}_1$, $\overline{X}_2$ و انحراف معیار S_1 , S_2 را با یکدیگر ادغام می‌کنیم ثابت کنید میانگین و انحراف معیار جدید از روابط زیر بدست می‌آید:

$$\overline{X} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2}$$

$$S^2 = \frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2} + \frac{N_1 N_2}{(N_1 + N_2)^2}\ (\overline{X}_1 - \overline{X}_2)$$

۲- میانگین و واریانس ۲ داده به ترتیب ۱۵ و ۵ می‌باشد. اگر به جای عدد ۲۵ اشتباهاً عدد ۱۵ را در محاسبات اعمال کرده باشیم میانگین و واریانس جدید را بدست بیاورید.

۳- نشان دهید تغیر مقیاس داده بر روی مقدار ضریب تغیرات $CV = \frac{S}{X}$ بی‌اثر می‌باشد، آیا این مطلب در مورد تغیر مبدأ نیز صادق است؟

۴- ثابت کنید برای میانگین حسابی، هندسی و همساز رابطه زیر برقرار است.

$$\overline{X}_H \le \overline{X}_G \le \overline{X}$$

۵- اگر میانگین را از داده‌های یک خامعه آماری کم کنیم و نتیجه را بر انحراف معیار تقسیم کنیم (یعنی $y_i = \frac{x_i - \overline{X}}{S}$) نشان دهید میانگین و انحراف معیار جدید به ترتیب صفر و یک می‌باشد.

۶- برای بدست آوردن یک معیار پراکندگی جدید داده‌ها را دو به دو با یکدیگر مقایسه می‌کنیم و میانگین $n^۲$ داده جدید $(X_i - X_j)$ را با $S_{ij}^۲$ نمایش می‌دهیم:

$$S_{ij}^۲ = \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{j=۱}^{n} (X_i - X_j)^۲$$

نشان دهید $S_{ij}^۲ = ۲S^۲$ (راهنمایی: داخل پرانتز مقدار $\overline{X}$ را اضافه و کم کنید)

۷- نشان دهید میانگین حسابی و واریانس نخستین n عدد طبیعی به ترتیب $\frac{n+۱}{۲}$ و $\frac{n^۲-۱}{۱۲}$ می‌باشد.

۸- جدول زیر را برای داده‌ها و فراوانی آنها در نظر بگیرید:

x	۰	۱	۲	...	n
f	$\binom{n}{۰}$	$\binom{n}{۱}$	$\binom{n}{۲}$	...	$\binom{n}{n}$

نشان دهید میانگین و واریانس این داده‌ها به ترتیب عبارتست از:

$$\overline{X} = \frac{n}{۲} \quad , \quad S^۲ = \frac{n}{۴}$$

۹- اگر متحرکی مسافت $x_۱$ را با سرعت $v_۱$ و ... و مسافت x_n را با سرعت v_n طی کنید ثابت کنید سرعت متوسط این متحرک با استفاده از رابطه میانگین همساز یا هارمونیک بدست می‌آید که در آن v_i معادل مقادیر داده‌ها و x_i معادل فراوانی آنهاست.

۱۰- عدد QP که $۰ < P < ۱$ را چندک Pام داده‌ها تعریف می‌کنیم هر گاه فراوانی تجمعی نسبی (r_i) آن بزرگتر یا مساوی با عدد P باشد. به عبارت دیگر هر گاه XP٪ ۱۰۰ از داده‌ها قبل از آن قرار گیرند. به عنوان مثال $Q_{./۵}$ که به آن چارک دوم می‌گوییم همان میانه می‌باشد چرا که ۵۰٪ داده‌ها قبل از آن قرار دارند. در حالت کلی $Q_{./۲۵}$ و $Q_{./۵}$ و $Q_{./۷۵}$ را به ترتیب با $Q_۱, Q_۲, Q_۳$ نمایش می‌دهیم و به آنها چارکهای اول و دوم و سوم می‌گوییم.

الف) اگر برای داده‌های گسسته x_i داشته باشیم $x_۱ < x_۲ < \cdots < x_i < \cdots x_N$ انگاه نشان دهید $Q_P = (۱-\omega)\, x_r + \omega x_{r+۱}$ که در آن
$r = (n+۱)P \quad , \quad \omega = (n+۱)P - r$.

ب) برای داده‌های پیوسته رده‌ای که فراوانی تجمعی آن بزرگتر یا مساوی با عدد P باشد را رده QP می‌نامیم
نشان دهید چندک Pام برای داده‌های پیوسته از رابطه زیر بدست می‌آید:

$$Q_P = L_P + \frac{(n\,p - g_p)}{f_p}$$

که در آن :
L_P : کران پایین رده QP
g_P : فراوانی تجمعی رده قبل از رده QP
f_P : فراوانی رده QP

ω : طول رده QP

۱۱- جدول زیر تعداد کتب فروخته شده توسط کتابفروشی را در طول ۳۰ روز نمای‌ش می‌دهد

۱۵	۱۰	۷	۲۰	۱۱	۱۳	۱۸	۶	۵	۴
۱۱	۱۹	۱۲	۱۶	۹	۱۰	۲۱	۱۳	۸	۱۴
۱۷	۲۰	۱۰	۱۲	۱۶	۱۳	۱۱	۱۲	۷	۱۱

برای داده‌های فوق :

الف: یک جدول فراوانی تشکیل دهید و نمودار میله‌ای داده‌ها را رسم کنید.

ب: میانگین داده‌ها $\overline{X}$ ، مد M و میانه m را بدست آورید.

ج: دامنه داده‌ها R، میانگین انحرافات، واریانس S^2 و ضریب تغیر را بدست آورید.

د: اگر به بزرگترین داده مقدار x $[x \geq ۰]$ واحد اضافه کنیم کدامیک از مقادیر مد یا میانه بدون تغیر باقی می‌مانند.

ه: چارک اول و سوم را بدست آورید و دامنه چارکها را محاسبه کنید .(دامنه چارکها : $Q_۳ - Q_۱$)

و: مقدار دهک چهارم را محاسبه کنید . (دهک چهارم همان $Q_{.۴}$ می‌باشد)

۱۲- در یک شهر میزان درجه حرارت در طول ۳۰ روز به قرار زیر است:

۵	۷	۸/۳	۱۰	۱۱	۱۲/۵	۱۳/۸	۱۳	۱۲	۱۳/۱
۱۴	۱۵/۲	۱۵/۶	۱۶	۱۵/۴	۱۵	۱۶/۵	۱۷	۱۹	۱۹/۷
۲۰/۶	۲۱	۲۱/۳	۲۰/۵	۲۲	۲۲/۸	۲۱/۷	۲۳	۲۴/۱	۲۵

الف: برای داده‌های فوق یک جدول فراوانی تشکیل دهید و هستیوگرام و چند بر فراوانی را رسم کنید

ب: مقادیر میانگین، میانه و مد را محاسبه کنید.

ج: مقادیر واریانس و ضریب تغیر را محاسبه کنید.

د: چند درصد داده‌ها در فاصله $(\overline{X}-S\ ,\ \overline{X}+S)$ و چند درصد داده‌ها در فاصله $(\overline{X}-۲S\ ,\ \overline{X}+۲S)$ قرار دارند؟

ه: چند درصد داده‌ها بین چارک اول و سوم قرار دارند؟