گزینه دو

مؤسسه آموزشی فرهنگی

آمار و مدل سازی

آمار:

آمار علم جمع‌آوری اطلاعات، سازمان‌دهی آن‌ها و تفصیل و تجزیه و تحلیل و نتیجه‌گیری اطلاعاتی می‌باشد، به عبارت دیگر آمار جمع‌آوری و طبقه‌بندی همراه با استنباط است. هم‌چنین آمار را هنر تصمیم‌گیری در شرایط نامطمئن نیز گفته‌اند.

فصل اول: اندازه‌گیری و مدل‌سازی

بیان مسأله به زبان ریاضی را مدلسازی ریاضی می‌گویند. هر چقدر مفاهیم ریاضی به‌کار برده شده ساده‌تر و ابتدایی‌تر و نتیجه‌ی کار به پدیده مورد نظر نزدیک‌تر باشد، مدل‌سازی با ارزش‌تر است.

اندازه‌گیری عبارت است از تخصیص معیار عددی به یک صفت. اولین قدم برای رسیدن به اطلاعات عددی اندازه‌گیری است.

در مدلسازی ریاضی با اندازه‌گیری سر و کار داریم و در اندازه‌گیری همواره خطا وجود دارد. خطای اندازه‌گیری تفاضل مقدار واقعی داده و مقدار اندازه‌گیری شده می‌باشد. این خطا لزوماً از واحد اندازه‌گیری کمتر است.

☜ مثال: اگر قد شخصی برابر ۱۸۰/۶ اندازه‌گیری شده باشد، مدل قد او را معین کنید.

✍حل:

$$H = ۱۸۰/۶ + E \qquad |E| < ۰/۱$$

چون توانستیم تا ۰/۱ متر را اندازه‌گیری کنیم پس باید خطا از واحد اندازه‌گیری کوچکتر باشد.

☜ مثال: اگر شعاع کره ای به‌صورت $R = ۳ + E$ که در آن E مقدار خطاست مدلسازی شده باشد، مدلی برای حجم کره بنویسید.

✍حل:

$$V = \frac{۴}{۳}\pi(R)^۳ = \frac{۴}{۳}\pi(۳+E)^۳ = \frac{۴}{۳}\pi(۲۷ + ۳ \times ۹E) = ۳۶\pi + ۳۶\pi E = ۳۶\pi + E_۱$$

☜ مثال: در یک استوانه مدل شعاع قاعده $R = ۴ + E_۱$ و مدل ارتفاع $h = ۳ + E_۲$ است. در این صورت حجم استوانه از چه مدلی پیروی می‌کند؟

✍حل:

$$V = \pi R^۲ h = \pi(۴ + E_۱)^۲(۳ + E_۲) = ۴۸\pi + ۲۴\pi E_۱ + ۱۶\pi E_۲$$

فصل دوم: جامعه و نمونه

جامعه‌ی آماری مجموعه‌ای از افراد یا اشیا است که درباره‌ی اعضای آن می‌خواهیم موضوع یا موضوعاتی را مطالعه کنیم. تعداد اعضای جامعه را اندازه‌ی جامعه می‌گوییم. اگر تمام افراد جامعه را مورد مطالعه قرار دهیم می‌گوییم سرشماری کرده‌ایم.

مشکلات سرشماری:

۱- در دسترس نبودن تمام اعضای جامعه

۲- وقت‌گیر بودن دسترسی به تمام اعضای جامعه

۳- گران تمام شدن بررسی تمام اعضای جامعه

۴- از بین رفتن جامعه در برخی از مطالعات

نمونه زیرمجموعه‌ای از جامعه‌ی آماری است.

نمونه تصادفی:

اگر جامعه‌ی آماری را که می‌خواهیم روی موضوعات آن مطالعاتی انجام دهیم کوچک باشد، معمولاً مطالعه را به‌صورت سرشماری انجام می‌دهیم.

نمونه: گاهی به علت وسعت جامعه نمی‌توانیم سرشماری (بررسی کلیه اجزای جامعه) انجام دهیم. در این مواقع اقدام به بررسی تصادفی بخشی از جامعه‌ی آماری می‌کنیم و نتایج حاصل را به‌صورت استقرایی به تمام جامعه تعمیم می‌دهیم. نمونه باید به قسمی انتخاب شود که بتواند بیانگر جامعه باشد.

لذا باید روش انتخاب نمونه به‌گونه‌ای باشد که:

۱- امکان انتخاب هر فرد به‌عنوان عضوی از نمونه امکان‌پذیر باشد.

۲- قبل از انتخاب نمونه، نتوانیم با اطمینان بیشتر درباره‌ی حضور یا عدم حضور عده‌ای در نمونه قضاوت کنیم.

بدیهی است نمونه‌ی مورد نظر باید به‌گونه‌ای انتخاب شود که تعمیم نتایج حاصل از نمونه‌گیری غیرواقعی نباشد. اطلاعات حاصل از نمونه‌گیری را **داده** می‌نامیم و تعداد افراد نمونه را **حجم یا اندازه‌ی نمونه** می‌نامند.

نمونه‌گیری تصادفی ساده به روش‌های مختلف انجام می‌گیرد، یکی از این روش‌ها استفاده از اعداد تصادفی به کمک ماشین حساب است (اعداد تصادفی، اعداد بین صفر و یک هستند که به‌وسیله‌ی کلید RAN تولید می‌شوند.). **پس از تولید عدد تصادفی به‌وسیله‌ی ماشین حساب آن را در حجم جامعه‌ی آماری ضرب کرده و اولین عدد پاسخ بزرگتر از یا مساوی عدد تولید شده را به‌عنوان عدد تصادفی در نظر می‌گیریم. سپس عضوی از جامعه را که عدد به‌دست آمده متناظر با اوست به‌عنوان نمونه در نظر می‌گیریم.**

☜ **مثال: دریک جامعه به حجم ۲۰۰ می‌خواهیم به کمک ماشین حساب نمونه‌گیری انجام دهیم. اعداد تصادفی ۰/۲۹۱ و ۰/۶۵۰ به‌دست آمده است. چه شماره‌هایی متناظر با این اعداد به‌ترتیب باید انتخاب شوند؟**

حل:

$۲۰۰\times۰/۲۹۱=۵۸/۲ \rightarrow$ نمونه $=۵۹$

$۲۰۰\times۰/۶۵۰=۱۳۰ \rightarrow$ نمونه $=۱۳۰$

روش‌های جمع‌آوری داده‌ها:

داده: نتایج حاصل از اندازه‌گیری و یا بررسی نمونه را داده می‌گوییم.

روش‌های جمع‌آوری داده:

۱- استفاده از داده‌های از پیش تهیه شده

۲- از طریق پرسش کتبی یا شفاهی

۳- از طریق مشاهده و ثبت وقایع

۴- از طریق انجام آزمایش

روش طراحی پرسش‌نامه:

۱- قبل از پرسش سؤال، محتوای پرسش‌نامه باید سازماندهی شود.

۲- هدف بررسی باید روشن باشد.

۳- تهیه فهرست از عناوینی که باید راجع به آن اطلاعات جمع‌آوری شود.

۴- خودداری از جمع‌آوری اطلاعات اضافی

۵- سؤالات کوتاه و واضح باشد. نباید از سؤالات چند برداشت شود. حتی‌الامکان جواب‌های تک‌کلمه‌ای یا تک‌رقمی

٦- در سؤالاتی که ممکن است پاسخ‌دهنده نخواهد به آن جواب دقیق بدهد از سؤالات با پاسخ از پیش آماده شده استفاده شود مانند تعیین محدوده

۷- عدم استفاده از سؤالات هدایت‌کننده یعنی سؤالاتی که جواب را به پاسخ‌دهنده القا می‌کند.

۸- حتی‌الامکان از سؤالات با پاسخ چندگزینه‌ای استفاده شود.

۹- برای پرسش عقیده راجع به موضوع یا محصول، پاسخ‌ها به‌صورت گزینه‌های کمی یا کیفی سطح‌بندی شده انتخاب شود. مانند بسیار خوب، خوب، متوسط، ضعیف، بسیار ضعیف.

۱۰- ضمیمه کردن دستورالعمل پاسخگویی (واضح و کامل)

۱۱- تشکر در پایان

فصل سوم: متغیرهای تصادفی

متغیر تصادفی: مشخصه یا صفت ویژه‌ای از افراد جامعه که روی آن مطالعه انجام می‌دهیم.

از دیدگاهی دیگر متغیرها بر دو نوع‌اند:

۱) **کمّی:** متغیری که اندازه و مقدار دارد و دارای واحد اندازه‌گیری می‌باشد، مانند: قد، وزن، عمر

که خود بر دو نوع است
- **پیوسته:** قابل اندازه‌گیری مانند: قد، وزن
- **گسسته:** قابل شمارش مانند: تعداد تصادفات، تعداد کالای خراب، تعداد روزهای بارانی

۲) **کیفی:** متغیری که قابل اندازه‌گیری نیست و عموماً با مقایسه بررسی می‌شود، مانند: هوش، استعداد، درس‌خوان بودن و...

که خود بر دو نوع است
- **ترتیبی:** دارای ترتیب ذاتی هستند مانند حروف الفبای فارسی یا مراحل رشد
- **اسمی:** دارای ترتیب ذاتی نیستند مانند رنگ چشم افراد، گروه‌های خونی

صفت: کمیت یا کیفیتی که متعلق به عناصر جامعه آماری است، صفت نامیده می‌شود که بر دو نوع است:

۱) **صفت ثابت:** صفتی که در بین عناصر جامعه‌ی آماری مشترک است.

۲) **صفت متغیر:** صفتی که از هر عضو به عضو دیگر تغییر می‌کند.

مثلاً در جامعه‌ی آماری ایرانی‌ها متغیر تصادفی وزن افراد صفت متغیر و ایرانی بودن صفت ثابت است.

فصل چهارم: دسته‌بندی داده‌ها و جدول فراوانی

فراوانی یک داده:

در داده‌های آماری دسته‌بندی نشده به تعداد دفعاتی که هر داده تکرار می‌شود، فراوانی مطلق آن داده گفته می‌شود و آن را با f_i نمایش می‌دهند و نسبت فراوانی مطلق هر داده به تعداد داده‌ها، فراوانی نسبی نامیده می‌شود.

$$F_i = \frac{f_i}{N} \qquad 0 \le F_i \le 1$$

درصد فراوانی نسبی $= F_i \times 100$

وقتی فراوانی داده‌های آماری که تعداد و پراکندگی‌شان زیاد است مورد مطالعه قرار می‌گیرند استفاده از جدول توزیع فراوانی بسیار دشوار است، لذا داده‌ها را دسته‌بندی می‌کنیم.

٤- دسته‌بندی داده‌ها:

اگر تعداد داده‌ها کم باشد یا داده‌ها خیلی پراکنده و دارای توزیع وسیع نباشند، جدولی متشکل از داده‌ها و تعداد تکرار آن‌ها رسم می‌کنیم که به آن جدول توزیع فراوانی گفته می‌شود.

تعاریف مرتبط:

الف – دامنه‌ی داده‌ها:

تفاضل کمترین از بیشترین داده در یک جامعه‌ی آماری، دامنه جامعه‌ی آماری نامیده می‌شود. $R = X_{\max} - X_{\min}$

ب- حدود دسته‌ها:

به اعدادی که دو طرف یک دسته قرار می‌گیرند، حدود آن دسته و یا کران‌های پایین و بالا گفته می‌شود و معمولاً به‌صورت قراردادی به‌استثنای دسته‌ی آخر حد بالای هر دسته متعلق به آن دسته نیست و جزو دسته‌ی بعد حساب می‌شود.

ج- تعداد دسته‌ها:

دستور خاصی برای انتخاب تعداد دسته‌ها وجود ندارد، اما معمولاً تعداد دسته‌ها را به‌گونه‌ای انتخاب می‌کنند که طول دسته‌ها دچار تعارض نگردد. به این معنا که فراوانی هیچ دسته‌ای صفر نباشد.

د- طول دسته:

تفاضل کران پایین هر دسته از کران بالای آن است که در درس ما معمولاً طول تمام دسته‌ها برابر فرض می‌شود. اگر در مورد تعداد دسته‌ها تصمیم‌گیری کردیم، طول دسته‌ها از رابطه‌ی زیر به‌دست می‌آید:

$$\text{طول دسته} = \frac{\text{دامنه‌ی تغییرات}}{\text{تعداد دسته‌ها}}$$

اگر بر اثر گرد کردن تعداد دسته‌ها، حاصل‌ضرب طول دسته در تعداد دسته از دامنه‌ی تغییرات بزرگ‌تر شد، این مقدار اضافی را بین دسته‌ی اول و دسته‌ی آخر به‌صورت مساوی تقسیم می‌کنیم.

اگر در مورد طول دسته‌ها تصمیم‌گیری کرده باشیم، در این صورت تعداد دسته‌ها از رابطه‌ی زیر به‌دست می‌آید.

$$\text{تعداد دسته‌ها} = \frac{\text{دامنه‌ی تغییرات}}{\text{طول دسته‌ها}}$$

که اگر عدد فوق اعشاری بود **حتماً** از رابطه‌ی $\left[\frac{\text{دامنه‌ی تغییرات}}{\text{طول دسته‌ها}}\right] + ۱$ استفاده می‌شود.

☜ **مثال: در یک آمارگیری بیشترین داده ۷۱ و دامنه‌ی تغییرات ۳۷ می‌باشد. اگر طول دسته‌ها ۵ انتخاب شده باشد و حد پایین اولین دسته را برابر کوچک‌ترین داده انتخاب کنیم، حد بالای آخرین دسته کدام است؟**

✍حل:

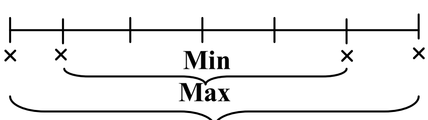
$$R = x_{\max} - x_{\min} = ۳۷ \to ۷۱ - x_{\min} = ۳۷ \to x_{\min} = ۳۴$$

پس اولین داده ۳٤ است. حال باید به اندازه‌ی $n = \left[\frac{۳۷}{۵}\right] + ۱ = ۸$ دسته با طول ۵ جلو برویم.

$۳۴ + ۸ \times ۵ = ۷۴$ = حد بالای آخرین دسته

☜ **مثال: در یک آمارگیری پس از دسته‌بندی داده‌ها ٦ دسته با طول ٤ به‌وجود آمده است، به فرض آن که هیچ دسته‌ای خالی نباشد، حدود تغییرات دامنه این داده‌ها را به‌دست آورید.**

✍حل:



$$۴ \times ۴ < d \le ۶ \times ۴ \to ۱۶ < d \le ۲۴$$

ه- فراوانی یک دسته در داده‌های دسته‌بندی شده:

تعداد داده‌هایی که در هر دسته قرار می‌گیرند را فراوانی مطلق آن دسته می‌گویند.

و- فراوانی نسبی و تجمعی یک دسته در داده‌های دسته‌بندی شده:

فراوانی نسبی دسته i ام **در داده‌های دسته‌بندی شده** برابر است با: $F_i = \dfrac{f_i}{\sum_{i=1}^{n} f_i}$

اگر فراوانی مربوط به هر دسته را با مجموع فراوانی‌های دسته‌های قبل از آن دسته جمع کنیم، فراوانی تجمعی مربوط به آن دسته به‌دست می‌آید: $f_{c_i} = f_1 + f_2 + \dots + f_i$

و فراوانی تجمعی نسبی دسته i ام **در داده‌های دسته‌بندی شده** برابر است با: $F_{c_i} = \dfrac{f_{c_i}}{\sum_{i=1}^{n} f_i}$

و اگر اعداد فوق را در ۱۰۰ ضرب کنیم درصد فراوانی نسبی و فراوانی تجمعی نسبی به‌دست می‌آید.

نکات:

۱) $f_{i+1} = f_{c_{i+1}} - f_{c_i}$

۲) در هر جدول توزیع فراوانی، مجموع فراوانی‌های نسبی برابر ۱ است. $\dfrac{f_1}{\sum f_i} + \dfrac{f_2}{\sum f_i} + \dots + \dfrac{f_n}{\sum f_i} = 1$

۳) فراوانی تجمعی دسته‌ی آخر برابر است با مجموع فراوانی‌ها

۴) در هر جدول توزیع فراوانی، فراوانی تجمعی نسبی دسته‌ی آخر برابر ۱ است.

☜ **مثال: توزیع زیر را در نظر می‌گیریم، درصد فراوانی نسبی متناظر با $x_i = 5$ کدامست؟**

x_i	۱	۲	۳	٤	۵	٦
f_i	٤	٦	۸	۷	۳	۲

✍**حل:**

درصد فراوانی نسبی $= \dfrac{3}{4+6+8+7+3+2} \times 100 = 10\%$

☜ **مثال: کوچک‌ترین و بزرگ‌ترین داده‌های آماری ۳۱ و ۵۲ می‌باشند. این داده‌ها در ۷ دسته، دسته‌بندی شده‌اند. ۳۷ درصد داده‌ها کمتر از ۴۰ و ۴۸ درصد آن‌ها بیشتر یا مساوی ۴۳ می‌باشد. اگر فراوانی کل ۸۰ باشد، فراوانی دسته وسط کدام است؟**

✍**حل:**

پس ۱۵ درصد داده‌ها بین ۴۰ تا ۴۳ هستند. چون دامنه ۲۱ و تعداد دسته‌ها ۷ است پس طول دسته‌ها برابر ۳ است. لـذا دسـته‌ی چهارم همان دسته‌ی ۴۳- ۴۰ خواهد بود. پس فراوانی دسته‌ی وسط برابر است با: $80 \times \dfrac{15}{100} = 12$

ز- نشان دسته (مرکز یا نماینده‌ی دسته):

فلسفه دسته‌بندی کردن داده‌ها در آمار، دادن ارزش یکسان به داده‌هایی است که در یک دسته قرار می‌گیرند، در داده‌های دسته‌بندی شده، نشان دسته (نماینده‌ی دسته) را به‌صورت زیر در نظر می‌گیریم:

$$x_i^* = \frac{\text{کران بالا} + \text{کران پایین}}{2}$$

طول دسته $= x_{i+1}^* - x_i^*$

☜ **مثال: در جدول توزیع فراوانی تعداد طبقات ۸ و طول هر دسته ۳ می‌باشد، در صورتی که نشان دسته‌ی طبقه سوم ۱۸ باشد، نشان دسته‌ی طبقه‌ی اول و آخر به‌ترتیب کدام‌اند؟**

حل:

$$x_1^* = x_3^* - 2c \rightarrow x_1^* = 18 - 2\times 3 = 12$$

$$x_8^* = x_3^* + 5c \rightarrow x_8^* = 18 + 5\times 3 = 33$$

☜ **مثال: در جدول فراوانی داده‌های دسته‌بندی شده، اگر درصد فراوانی نسبی دسته‌ی وسط ۲۴ باشد، فراوانی مطلق دسته‌ی چهارم کدام است؟**

نشان دسته	۱۳	۱۵	۱۷	۱۹	۲۱
فراوانی تجمعی	۵	۱۴	a	۴۱	۵۰

حل:

نشان دسته	۱۳	۱۵	۱۷	۱۹	۲۱
فراوانی تجمعی	۵	۱۴	a	۴۱	۵۰
فراوانی مطلق	۵	۹	$a-14$	$41-a$	۹

$$\frac{a-14}{50} = \frac{24}{100} = \frac{12}{50} \rightarrow a = 26 \rightarrow f_4 = 41 - 26 = 15$$

شاخص‌های آماری و تحلیل داده‌ها:

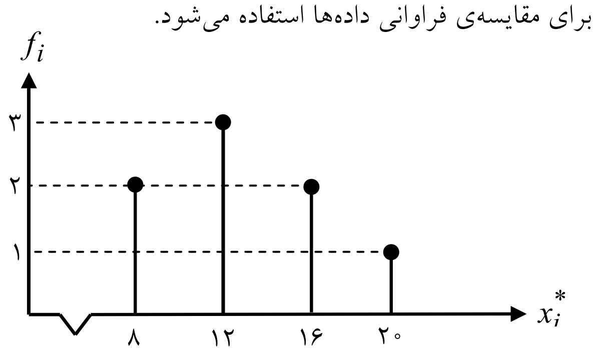
آنچه تاکنون بر روی داده‌ها انجام دادیم، به نوعی سازمان‌دهی آن‌ها برای بهره‌برداری ساده‌تر بود که عمدتاً بر مبنای خلاصه کردن داده‌ها بنا شده بود، اما با تلفیق و به هم آمیختن داده‌ها، سرانجام مقادیر عددی به‌دست خواهد آمد که به نحو شایسته‌ای می‌تواند جامعه را تحلیل کند. این مقادیر عددی را شاخص‌های عددی می‌گوییم، در مقابل برای تجسم جامعه آماری از شاخص‌های هندسی بهره می‌گیریم.

فصل پنجم: نمودارها و تحلیل داده‌ها (شاخص‌های هندسی)

انواع مختلف نمودارهای آماری مفید به قرار زیرند:

۱) نمودار میله‌ای: اگر نقاط متناظر با متغیرها (یا نشان دسته‌ها) را روی محور xها مشخص کنیم و روی هر نقطه پاره‌خطی به ارتفاع فراوانی (مطلق یا نسبی) نظیر آن رسم کنیم، شکل حاصل نمودار میله‌ای داده‌ها می‌باشد.

این نمودار برای متغیرهای کیفی و کمی گسسته مناسب است و بیشتر برای مقایسه‌ی فراوانی داده‌ها استفاده می‌شود.



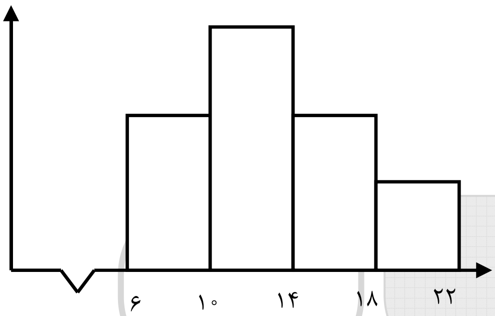
قرارداد: در مواردی که مقیاس بزرگ است و ممکن است تمام نمودار در صفحه قرار نگیرد، نزدیک‌ترین داده به مبدأ را با علامت ⌄ به مبدأ نزدیک می‌کنیم.

۲) نمودار مستطیلی (هیستوگرام):

هنگامی که جدول توزیع فراوانی در دست است از این نمودار استفاده می‌کنیم. در این نمودار مستطیل‌هایی رسم می‌کنیم که یک ضلع آن منطبق بر دسته‌ها و ضلع دیگر، فراوانی (مطلق یا نسبی) دسته متناظر باشد.

نمودار مستطیلی برای داده‌های **کمی پیوسته** مناسب است.

اگر طول دسته‌ها با هم برابر باشند، فراوانی دسته‌ای که مساحت آن بزرگ‌تر است، بیشتر خواهد بود. در واقع فراوانی متناسب با مساحت مستطیل‌هاست.



نمودار مستطیلی= هیستوگرام، بافت نگار

☜ **مثال: جدول زیر جدول فراوانی تجمعی داده‌های آماری دسته‌بندی شده است اگر فراوانی نسبی دسته سوم برابر ۰/۴۵ باشد آن‌گاه مساحت مستطیل مربوط به دسته‌ی چهارم در نمودار مستطیلی این داده‌های آماری کدام است؟**

مرکز دسته	۱۷	۲۰	۲۳	۲۶	۲۹	۳۲
فراوانی تجمعی	۴	۱۳	a	۴۹	۵۸	۶۰

✍حل:

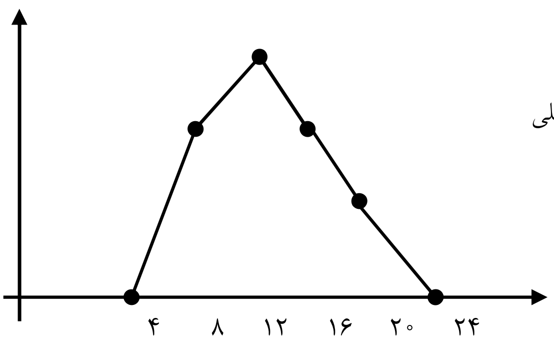
طول دسته $= x_2^* - x_1^* = 20 - 17 = 3$

فراوانی نسبی دسته‌ی سوم $= \frac{a-13}{60} = \frac{45}{100} \rightarrow a - 13 = 27 \rightarrow a = 40$

فراوانی مطلق دسته‌ی چهارم $= 49 - a = 9 \rightarrow$ **مساحت مستطیل دسته‌ی چهارم** $= 9 \times 3 = 27$

۳) چندبر فراوانی:

در این نمودار نقاطی از صفحه را مشخص می‌کنیم که طول آن‌ها برابر مرکز دسته‌ها و عرض آن‌ها، فراوانی (مطلق یا نسبی) همان دسته است. با به هم وصل کردن این نقاط چندبر فراوانی به‌دست می‌آید. در چندبر فراوانی، دو دسته با فراوانی صفر به ابتدا و انتهای داده‌ها اضافه می‌کنیم که به این وسیله مساحت چندبر فراوانی با مساحت نمودار مستطیلی متناظرش برابر می‌شود.



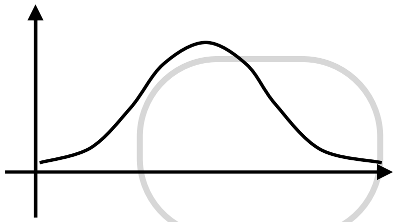
مساحت چندبر = مساحت نمودار مستطیلی

نمودار چندبر فراوانی برای داده‌های **کمی پیوسته** مناسب است.

نمودار چند بر فراوانی را می‌توان با داشتن فراوانی نسبی نیز رسم کرد که آن را چندبر فراونی نسبی می‌گوییم. در این صورت اطلاعات منسجم‌تری در اختیار ما قرار می‌گیرد، چون می‌توان فراوانی را با کل جامعه مقایسه کرد.

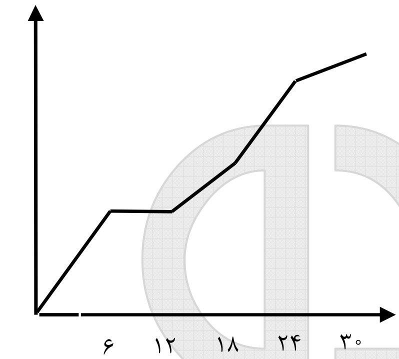
اگر بخواهیم تغییرات متغیر را در فاصله‌ی بین دو دسته یا در خود دسته بهتر نشان دهیم از نمودار چندبر فراوانی استفاده می‌کنیم.

اگر بخواهیم دو جدول توزیع فراوانی را از لحاظ هندسی با هم مقایسه کنیم بهتر است از نمودار چندبر فراوانی استفاده کنیم.
نمودارهای چندبر فراوانی و مستطیلی به منظور نمایش هندسی جدول‌های فراوانی مورد استفاده قرار می‌گیرند، از این‌رو می‌توان به دلخواه یکی از آن دو را انتخاب کرد. ولی هنگامی که قصد داشته باشیم دو یا چند توزیع را به‌صورت هندسی با یکدیگر مقایسه کنیم، مشکل بتوانیم با انطباق مستطیل‌ها بر یکدیگر این مقایسه را انجام دهیم. در چنین مواردی رسم چندبر فراوانی انتخاب مناسب‌تری است زیرا مقایسه‌ی توزیع فراوانی را آسان تر می‌کند.
با افزایش داده‌ها یا کوچکتر شدن حدود دسته‌ها چندبر به یک منحنی شبیه می‌شود از این‌رو این چندبر فراوانی به منحنی فراوانی تغییر نام می‌دهد.
منحنی نرمال: منحنی‌ای است که در اکثر پدیده‌های طبیعی ظاهر می‌شود و شکلی شبیه زنگوله دارد و سطح زیر منحنی آن برابر فراوانی مطلق داده‌هاست. (اگر منحنی را با فراوانی نسبی رسم کنیم سطح زیر منحنی برابر ۱ است)



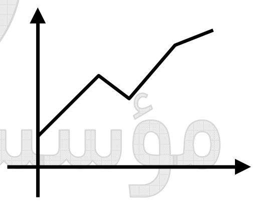
این منحنی دارای ماکسیمم است و فراوانی در دو طرف این ماکسیمم به‌طور یکنواخت به سمت صفر میل می‌کند. این منحنی متقارن است.

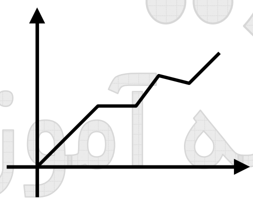
٤) نمودار تجمعی:

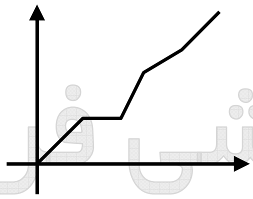


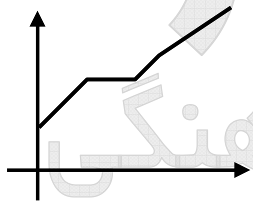
هنگامی که جدول توزیع فراوانی در دست است از این نمودار استفاده می‌کنیم.
در این نمودار حدود دسته‌ها منطبق بر محور xهاست و در انتهای هر دسته فراوانی تجمعی (مطلق یا نسبی) دسته متناظر قرار می‌گیرد.
این نمودار همواره به‌صورت صعودی است.

☜ **مثال: کدام شکل نمودار فراوانی تجمعی است؟**

۱)


۲)


۳)


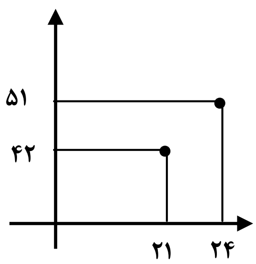
٤)


✍حل:
نمودار تجمعی صعودی است و حتماً از مبدأ شروع می‌شود، لذا فقط گزینه‌ی ۳ می‌تواند جواب صحیح باشد.

☜ **مثال: اگر در جدول فراوانی داده‌های پیوسته و طبقه‌بندی شده دو نقطه‌ی (۲۱,۴۲) و (۲۴,۵۱) دو نقطه‌ی متوالی از نمودار فراوانی تجمعی باشند، کدام نقطه‌ی زیر روی چندبر فراوانی قرار دارد؟**

۱) (۲۱,۵۱)   ۲) (۲۲/۵,۹)   ۳) (۲۴,۹)   ٤) (۲۲/۵,۴۲)

✍حل:



دو نقطه‌ی داده شده مشخص می‌کند فراوانی مطلق دسته‌ی ۲٤– ۲۱ برابر ۹ است، پس نقطه‌ی **(۲۲/۵,۹)** روی چندبر فراوانی قرار می‌گیرد.

۵) نمودار دایره‌ای:

برای رسم نمودار دایره‌ای، مساحت دایره را به‌وسیله‌ی قطاع‌هایی به نسبت فراوانی‌ها تقسیم می‌کنیم. اگر فراوانی دسته‌ای معلوم باشد، زاویه‌ی هر قطاع از رابطه‌ی $\theta_i = \frac{f_i}{N} \times ۳۶۰^\circ$ به‌دست می‌آید. نمودار دایره‌ای امکان مقایسه بین فراوانی‌ها را با سرعت بیشتر فراهم می‌کند.

☜ **مثال: در انتخابات یک شهر ۵۴۰۰۰۰ نفر شرکت کرده‌اند. اگر آنان را به ۵ گروه سنی تقسیم نموده و با نمودار دایره‌ای نشان دهیم، درصد شرکت‌کنندگان در یک گروه سنی با زاویه‌ی قطاع ۶۳ درجه نشان داده می‌شود. تعداد اعضای این گروه کدام است؟**

✍**حل:**

$$\theta_i = \frac{f_i}{N} \times ۳۶۰ = ۶۳^\circ \rightarrow f_i = \frac{۶۳}{۳۶۰} \times N = \frac{۷}{۴۰} \times ۵۴۰۰۰۰ = ۹۴۵۰۰$$

☜ **مثال: جدول مقابل درصد فراوانی نسبی گروه خونی افراد یک جامعه است. در نمودار دایره‌ای، زاویه‌ی سطح مربوط به گروه خونی O چند درجه است؟**

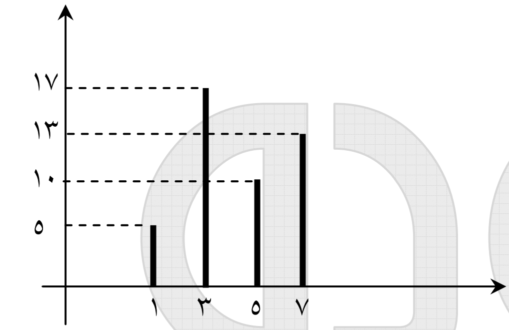
گروه خونی	A	B	AB	O
درصد فراوانی نسبی	۲۴	۲۲/۵	۳۶	α

✍**حل:**

$$\alpha = ۱۷/۵$$

$$\theta = \frac{۱۷/۵}{۱۰۰} \times ۳۶۰ = ۶۳^\circ$$

☜ **مثال: نمودار دایره‌ای متناظر با نمودار میله‌ای روبه‌رو را رسم کنید.**



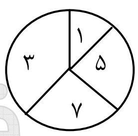
✍**حل:**

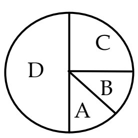
$$\theta_۱ = \frac{۵}{۴۵} \times ۳۶۰ = ۴۰^\circ$$

$$\theta_۳ = \frac{۱۷}{۴۵} \times ۳۶۰ = ۱۳۶^\circ$$

$$\theta_۵ = \frac{۱۰}{۴۵} \times ۳۶۰ = ۸۰^\circ$$

$$\theta_۷ = \frac{۱۳}{۴۵} \times ۳۶۰ = ۱۰۴^\circ$$





☜ **مثال: در نمودار زیر تعداد فراوانی افرادی که در دسته‌های B , C , D قرار دارند، به‌ترتیب ۲، ۳ و ٦ برابر تعداد افرادی است که در دسته‌ی A قرار دارند. زاویه‌ی A کدام است؟**

✍**حل:**

$$\theta_A + \theta_B + \theta_C + \theta_D = \theta_A + ۲\theta_A + ۳\theta_A + ۶\theta_A = ۱۲\theta_A = ۳۶۰ \rightarrow \theta_A = ۳۰^\circ$$

☜ **مثال: نمودار دایره‌ای داده‌های $x_۱, ..., x_n$ را با نمودار دایره‌ای داده‌های $x_۱, x_۱, ..., x_n, x_n$ مقایسه کنید.**

✍**حل:**

چون فراوانی نسبی داده‌ها تغییر نکرده پس زاویه نمودار تغییر نمی‌کند.

☜ مثال: داده‌های آماری در ۹ طبقه دسته‌بندی شده‌اند. فراوانی تجمعی نسبی در دسته‌ی چهارم و پنجم به‌ترتیب ۰/۲۸ و ۰/٤۰ است. در نمودار دایره‌ای زاویه‌ی مربوط به دسته‌ی پنجم چند درجه است؟

✍حل:

فراوانی نسبی دسته‌ی پنجم $= ۰/۴ - ۰/۲۸ = ۰/۱۲ \rightarrow \theta = ۰/۱۲ \times ۳۶۰ = ۴۳/۲^{\circ}$

٦) نمودار ساقه و برگ:

برای آن‌که اعداد را به‌گونه‌ای مرتب کنیم که چیزی شبیه نمودار میله‌ای به‌دست آید، از نمودار ساقه و برگ استفاده می‌کنیم. برای تهیه‌ی نمودار ساقه و برگ داده‌ها را به دو بخش تقسیم می‌کنیم. ساقه شامل یک یا چند رقم اولیه و برگ که با فاصله‌ی کمی از ساقه جلوی آن به‌صورت صعودی نوشته می‌شود، شامل ارقام باقی‌مانده است. اگر هم عددی تکرار شود به دفعات تکرار آن را می‌نویسیم. برای فهم آن‌که چند رقم ساقه و چند رقم برگ است گاهی کلید نمودار را به‌صورت زیر مشخص می‌کنند:

ساقه	برگ
۳۲	١ ٤ ٥ ٧ ٨ ٨ ٩
۳۳	٠ ٠ ٤ ٥ ٥ ٦
۳٤	٢ ٣ ٦ ٦ ٧

کلید: ۹ ۲۳

این نمودار به خوبی فاصله‌ی کمترین و بیشترین داده را نشان می‌دهد.

☜ مثال: در نمودار ساقه و برگ زیر چه اعدادی به‌جای ؟ می‌توانند قرار گیرند؟

ساقه	برگ
١	١ ٤ ؟ ٧ ٨ ٨ ٩
٢	٠ ٠ ٤ ٥ ؟ ٦
٣	٢ ٣ ٦ ٦ ٧

✍حل:

$$\left.\begin{array}{l} ۴ \le ? \le ۷ \\ ۵ \le ? \le ۶ \end{array}\right\} \Rightarrow ? = ۵ \text{ یا } ۶$$

☜ مثال: در نمودار ساقه و برگ، ۵۰ داده بین ۱۰ و ۲۰ می‌باشد، که برخی داده‌ها شامل یک رقم اعشار بوده، برگ‌های مربوط به اتصال ١٤ روی این ساقه به‌صورت ۰۳۵۵۵۵٦٧٧٩، ١٤ نمایش داده شده «چند درصد داده‌ها» درست اعداد ١٤/٥ بوده‌اند؟

✍حل:

این داده‌ها عبارتند از: ۱۴/۹, ۱۴/۷, ۱۴/۷, ۱۴/۶, ۱۴/۵, ۱۴/۵, ۱۴/۵, ۱۴/۵, ۱۴/۳, ۱۴

پس $\frac{۴}{۵۰}$ یا ۸% داده‌ها عدد ۱۴/۵ بوده است.

۲- شاخص‌های عددی:

معمولاً علاقمندیم به کمک مقادیری تمرکز داده‌ها را نشان دهیم. برای این منظور باید بدانیم، داده‌ها حول چه نقطه‌ای تجمع پیدا کرده‌اند و ضمناً چگونه تجمع پیدا کرده‌اند. آیا داده‌ها به هم نزدیکند یا از هم دورند. برای منظور اول از شاخص‌های مرکزی استفاده می‌کنیم که نشان‌دهنده‌ی مرکز تجمع داده‌ها می‌باشند و برای منظور دوم از شاخص‌های پراکندگی استفاده می‌کنیم که نشان‌دهنده‌ی میزان تجمع داده‌ها حول مرکز است.

فصل ششم: شاخص‌های مرکزی

شاخص‌های مرکزی عبارتند از: میانگین، میانه، مد

۱) میانگین:

میانگین n داده‌ی آماری $x_1, x_2, ..., x_n$ برابر است با:

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \sum_{i=1}^{n} x_i = n\overline{X}$$

☜ **مثال: اگر میانگین داده‌های x_1, x_2, x_3, x_4 برابر $\overline{X}$ باشد، میانگین داده‌های $2x_1 + x_2, 2x_2 + x_3, 2x_3 + x_4, 2x_4 + x_1$ کدام است؟**

✍**حل:**

$$\overline{x} = \frac{2x_1 + x_2 + 2x_2 + x_3 + 2x_3 + x_4 + 2x_4 + x_1}{4} = \frac{3(x_1 + x_2 + x_3 + x_4)}{4} = 3\overline{x}_{\text{قدیم}}$$

نکات:

۱) در یک جامعه‌ی آماری، میانگین عددی منحصر به فرد است که همواره بین بزرگ‌ترین و کوچک‌ترین داده قرار دارد.

۲) اگر داده‌های آماری تشکیل تصاعد حسابی بدهند، میانگین آن‌ها عبارت است از:

$$\overline{X} = \frac{\overset{\text{اولین داده}}{x_1} + \overset{\text{آخرین داده}}{x_n}}{2}$$

۳) مفیدترین خاصیت میانگین آن است که اگر به‌جای تک‌تک داده‌ها، میانگین آن را قرار دهیم، مجموع و میانگین داده‌ها تغییر نخواهد کرد.

$$\sum_{i=1}^{n} x_i = n\overline{X} \qquad \sum_{i=1}^{n} \overline{X} = n\overline{X}$$

٤) اگر میانگین m داده‌ی آماری $\overline{X}$ و میانگین n داده‌ی آماری $\overline{Y}$ باشد، میانگین کلیه داده‌ها برابر است با: $\dfrac{m\overline{X} + n\overline{Y}}{m+n}$

و از همین‌جا می‌توان نتیجه گرفت اگر میانگین m_1 داده‌ی آماری $\overline{X}_1$ و میانگین m_2 داده‌ی آماری $\overline{X}_2$ و... و میانگین m_n داده‌ی آماری $\overline{X}_n$ باشد، میانگین کل داده‌ها برابر است با:

$$\overline{X} = \frac{\sum_{i=1}^{n} m_i \overline{X}_i}{\sum_{i=1}^{n} m_i}$$

☜ **مثال: میانگین داده‌های $x_1, x_2, \ldots, x_{10}$ برابر ۵ است. میانگین مقادیر $x_1, x_2, \ldots, x_{10}, 16$ کدام است؟**

✍**حل:**

$$\frac{\sum_{1}^{10} x_i}{10} = 5 \rightarrow \sum_{1}^{10} x_i = 50 \rightarrow \bar{x} = \frac{50+16}{11} = \frac{66}{11} = 6$$

☜ **مثال: اگر میانگین نمرات یک کلاس ۳۰ نفری ۱۲ و میانگین یک کلاس ۲۰ نفری ۱۶ باشد، میانگین نمرات این ۵۰ دانش‌آموز چقدر است؟**

✍**حل:**

$$\bar{x} = \frac{30\times12+20\times16}{30+20} = \frac{360+320}{50} = 13/6$$

۵) اگر میانگین داده‌های $x_1, x_2, \ldots, x_n$ ، $\bar{X}$ باشد، میانگین داده‌های $ax_1+b, ax_2+b, \ldots, ax_n+b$ برابر است با: $a\bar{X}+b$

۶) مجموع تفاضل داده‌ها از میانگین برابر صفر است. $\sum_{i=1}^{n}(x_i-\bar{X})=0$

محاسبه‌ی میانگین در جدول فراوانی: (میانگین وزن دار)

میانگین n داده‌ی آماری که در آن داده‌ها دارای فراوانی باشند برابر است با:

$$\bar{X} = \frac{\overbrace{\sum f_i x_i}^{\text{مجموع داده‌ها}}}{\underbrace{\sum f_i}_{\text{تعداد داده‌ها}}}$$

که در این صورت: $\sum f_i(x_i-\bar{X})=0$

اگر جدول طبقه‌بندی شده داده‌ها در دسترسی باشد، برای محاسبه‌ی میانگین به‌جای داده‌ها در رابطه‌ی فوق از نشان دسته بهره می‌بریم.

$$\bar{X} = \frac{\sum_{1}^{n} f_i x_i^*}{\sum_{1}^{n} f_i}$$

که در این صورت: $\sum f_i(x_i^*-\bar{X})=0$

☜ **مثال: میانگین داده‌های جدول زیر را به‌دست آورید.**

دسته	۲-۴	۴-۶	۶-۸	۸-۱۰
فراوانی	۲	۴	۳	۱

✍**حل:**

$$\bar{x} = \frac{3\times2+5\times4+7\times3+9\times1}{2+4+3+1} = \frac{6+20+21+9}{10} = 5/6$$

۲) میانه:

اگر داده‌های آماری را به‌صورت صعودی یا نزولی مرتب کنیم، در صورتی که تعداد داده‌ها فرد باشد، عددی که در وسط قرار می‌گیرد و اگر تعداد داده‌ها زوج باشد، نصف مجموع دو عددی که در وسط قرار گرفته، میانه نام دارد. معمولاً میانه مقداری است که تعداد اعضایی از جامعه که از آن بیشترند برابر تعداد اعضایی است که از آن کمترند. (نماد میانه: M و $\tilde{x}$)

نکته: اگر میانه داده‌های $x_1, x_2, ..., x_n$ برابر M باشد، میانه‌ی داده‌های $ax_1 + b, ..., ax_n + b$ برابر است با: $aM+b$

☜ **مثال: در یک امتحان ریاضی نمرات ۱۵ نفر به‌صورت زیر است:**

۵، ۹، ۱۰، ۱٤، ۱۹، ۱۷، ۱٤، ۱۵، ۱۷، ۱۱، ۱۲، ۳، ۷، ۷، ٤

میانه این نمرات کدام است؟

✍**حل:**

۳,۴,۵,۷,۷,۹,۱۰,۱۱,۱۲,۱۴,۱۴,۱۵,۱۷,۱۹

چون ۱۵ داده داریم، داده‌ی ۸ ام میانه است. پس میانه ۱۱ است.

برای n داده‌ی فرد، میانه $\frac{n+1}{2}$مین داده است.

نکته: اگر جدول توزیع فراوانی برای داده‌های دسته‌بندی نشده در دسترس باشد، برای پیدا کردن میانه، ابتدا فراوانی تجمعی را به‌دست آورده و سپس اولین ردیفی که فراوانی تجمعی آن بزرگ‌تر از یا مساوی $\frac{n+1}{2}$ می‌باشد، ردیف میانه خواهد بود.

☜ **مثال: در جدول زیر میانه را به‌دست آورید.**

x_i	۲	۳	٤	۵	٦	۷
f_i	۲	۲	۳	۵	٤	٤

✍**حل: ابتدا در جدول فراوانی تجمعی را به‌دست می‌آوریم:**

x_i	۲	۳	٤	۵	٦	۷
f_i	۲	۲	۳	۳	٦	٤
f_{c_i}	۲	٤	۷	۱۲	۱٦	۲۰

چون ۲۰ داده داریم، میانه جمع داده‌ی ۱۰م و ۱۱ام است که هر ۲ در دسته‌ی $x_i = 5$ قرار دارند (داده‌های ۷ تا ۱۲ همگی ۵اند)

پس میانه برابر $\frac{5+5}{2} = 5$ است.

چارک‌ها:

چارک اول: در واقع عددی است که از $\frac{1}{4}$ داده‌ها بزرگ‌تر و از $\frac{3}{4}$ داده‌ها کوچک‌تر است و با Q_1 نمایش داده می‌شود.

چارک دوم: که با Q_2 نمایش داده می‌شود همان میانه است.

چارک سوم: عددی است که از $\frac{3}{4}$ داده‌ها بزرگ‌تر و از $\frac{1}{4}$ داده‌ها کوچک‌تر است و با Q_3 نمایش داده می‌شود.

برای یافتن چارک اول و چارک سوم، پس از مرتب کردن داده‌ها، میانه را پیدا می‌کنیم و سپس مجدداً برای داده‌های قبل و بعد از میانه، میانه پیدا می‌کنیم.

☜ **مثال: چارک‌های اول و دوم و سوم داده‌های زیر را بیابید.**

۹، ٦، ۳، ۱۲، ۷، ۵، ۱۰

✍**حل:**

۳,۵,۶,۷,۹,۱۰,۱۲

پس میانه یا چارک دوم ۷ است، چارک اول ۵ و چارک سوم ۱۰ است.

☜ **مثال: در داده‌های آماری با نمودار ساقه و برگ، داده‌های کمتر از چارک اول و بیشتر از چارک سوم را حذف می‌کنیم. میانگین داده‌های باقی‌مانده کدام است؟**

ساقه	برگ
۳	۱ ٤ ۵ ۷ ۸ ۸ ۹
٤	۰ ۰ ٤ ۵ ۵ ٦
۵	۲ ۳ ٦ ٦ ۷

✍**حل:**

چون ۱۸ داده داریم، میانه‌ی ۹ داده‌ی اول داده‌ی ۵ام و ۹ داده‌ی دوم، داده‌ی ١٤ام است، یعنی ۳۸ چارک اول و ۵۲ چارک سوم است. پس میانگین داده‌های باقی‌مانده عبارت است از: ٤۲/۷

$$\overline{X}=\frac{(۳۸+۳۸+۳۹+۴۰+۴۰+۴۴+۴۵+۴۵+۴۶+۵۲)}{۱۰}=۴۲/۷$$

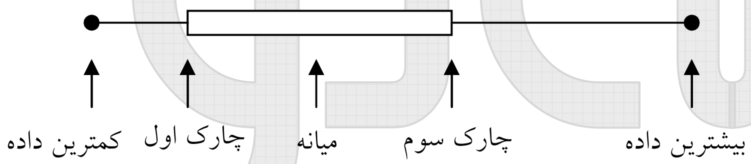
☜ **مثال: داده‌های آماری در ۹ طبقه با طول دسته‌ی ٤، دسته‌بندی شده‌اند. اگر ۸ داده بین چارک اول و سوم به آن اضافه شود و یک واحد از طول دسته کم کنیم، در دسته‌بندی جدید تعداد دسته‌ها کدام است؟**

✍**حل:**

دامنه‌ی داده‌ها $۳۶=۴\times۹$ است. حال اگر طول دسته‌ها برابر ۳ باشد، تعداد دسته‌ها $۱۲=\frac{۳۶}{۳}$ می‌شود. اضافه شدن تعدادی داده داخل دامنه، تأثیری بر حدود دسته نمی‌گذارد.

نمودار جعبه‌ای:

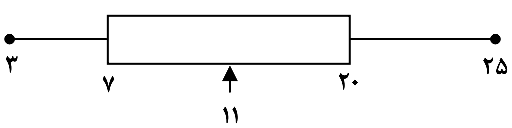
برای نشان دادن بهتر پراکندگی از این نمودار استفاده می‌کنیم. برای تهیه‌ی این نمودار کوچکترین و بزرگترین داده را پیدا می‌کنیم. چارک اول و سوم را می‌یابیم سپس با رسم جعبه‌ای بین چارک اول و سوم و وصل کردن دو سر جعبه به بزرگترین و کوچکترین داده، نمودار را رسم می‌کنیم.



☜ **مثال: نمودار جعبه‌ای داده‌های زیر را رسم کنید: ۱۰،۳،۲۰،۱۳،۷،۱۱،۲۵**

✍**حل:**

۳ , ۷ , ۱۰ , ۱۱ , ۱۳ , ۲۰ , ۲۵



۳) مد (نما):

مقدار یا مقادیری از متغیر که در آن‌ها فراوانی ماکزیمم مطلق باشد، مد نامیده می‌شود $(\hat{x})$. در واقع مد داده‌ای است که بیشترین تکرار را در میان داده‌ها داشته باشد و توجه کنید که مد **منحصر به فرد نیست.**

معمولاً مد برای تحلیل متغیرهای کیفی به‌کار می‌رود.

اگر $\hat{x}$ مد داده‌های $x_1,...,x_n$ باشد، مد داده‌های $ax_1+b,...,ax_n+b$ برابر است با $a\hat{x}+b$. اگر جدول داده‌های دسته‌بندی شده در دسترس باشد، نشان دسته طبقه‌ای که بیشترین فراوانی را دارد، مد می‌باشد.

فصل هفتم: شاخص‌های پراکندگی:

شاخص‌های مرکزی در مورد پراکندگی داده‌ها اطلاعاتی به ما نمی‌دهند لذا به تعریف شاخص‌هایی برای پراکندگی می‌پردازیم.

۱) دامنه‌ی تغییرات:

با تعریف دامنه‌ی تغییرات آشنا شدیم. $R = x_{max} - x_{min}$

دامنه‌ی تغییرات برای بیان پراکندگی اطلاعات زیادی به ما نمی‌دهد زیرا فقط بزرگ‌ترین و کوچک‌ترین داده در آن مؤثرند.

مثال: اگر کلیه‌ی داده‌های آماری در عدد ثابتی ضرب و با عدد ثابتی جمع شوند، دامنه‌ی آن چه تغییری می‌کند؟

حل:

$$\mathbf{R} = |\mathbf{x}_{max} - \mathbf{x}_{min}|$$

$$\mathbf{R'} = |(\mathbf{a}\mathbf{x}_{max} + \mathbf{b}) - (\mathbf{a}\mathbf{x}_{min} + \mathbf{b})| = |\mathbf{a}|\mathbf{R}$$

اگر $a < 0$ باشد، جای داده ماکزیمم و مینیمم برعکس می‌شود.

مثال: چرا دامنه‌ی تغییرات به‌عنوان معیار پراکندگی مناسب نیست؟

۱) با ازدیاد داده‌ها فقط می‌تواند کم شود. **۲) مقدار آن اغلب بزرگ است.**

۳) مقدار آن از یک نمونه به نمونه دیگری تغییر می‌کند. **۴) در محاسبه‌ی آن فقط از دو اندازه استفاده می‌شود.**

حل: دامنه فقط تابع بزرگ‌ترین و کوچک‌ترین داده است.

۲) واریانس (پراش):

بهترین پارامتر پراکندگی واریانس است.

برای نشان دادن پراکندگی باید از مفهوم انحراف از میانگین بهره بگیریم، اما چون مجموع انحراف از میانگین‌ها در یک جامعه‌ی آماری صفر می‌باشد، مجموع مربع‌های آن‌ها را در نظر می‌گیریم. اگر این مجموع را بر تعداد داده‌ها تقسیم کنیم واریانس (پراش) داده‌ها حاصل می‌شود.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n}$$

اگر جدول داده‌های طبقه‌بندی نشده در دسترس باشد، از رابطه‌ی زیر استفاده می‌کنیم.

x_i	x_1	x_2	....	x_n
f_i	f_1	f_2	....	f_n

$$\sigma^2 = \frac{f_1(x_1 - \overline{X})^2 + f_2(x_2 - \overline{X})^2 + \dots + f_n(x_n - \overline{X})^2}{f_1 + f_2 + \dots + f_n}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \overline{X})^2}{\sum_{i=1}^{n} f_i}$$

اگر جدول داده‌های دسته‌بندی شده در دسترس باشد از نشان هر دسته به‌عنوان نماینده‌ی آن استفاده می‌کنیم تا مسأله به حالت قبلی تبدیل شود.

حدود	$\overset{x_1^*}{\uparrow}$ $x_1 - x_2$	$\overset{x_2^*}{\uparrow}$ $x_2 - x_3$	...	$x_{n-1} - x_n$
فراوانی	f_1	f_2	...	f_n

$$\sigma^2 = \frac{\sum_{1}^{n} f_i(x_i^* - \overline{X})^2}{\sum_{1}^{n} f_i}$$

☜ مثال: واریانس داده‌های جدول زیر کدام است؟

x_i	۲	۳	٤	۵	٦	۷
f_i	۱	۱	۳	۵	٤	۲

✍حل:

$$\bar{x}=\frac{۲\times۱+۳\times۱+۴\times۳+۵\times۵+۶\times۴+۷\times۲}{۱+۱+۳+۵+۴+۲}=۵$$

$$\sigma^{۲}=\frac{۱(۵-۲)^{۲}+۱(۳-۵)^{۲}+۳(۴-۵)^{۲}+۵(۵-۵)^{۲}+۴(۶-۵)^{۲}+۲(۷-۵)^{۲}}{۱۶}=۱/۷۵$$

☜ مثال: واریانس داده‌های آماری جدول زیر کدام است؟

x_i	۰-۲	۲-٤	٤-٦	٦-۸
f_i	۱	۲	۹	٤

✍حل:

$$\bar{x}=\frac{۱\times۱+۲\times۳+۹\times۵+۴\times۷}{۱+۲+۹+۴}=۵$$

$$\to\sigma^{۲}=\frac{(۱-۵)^{۲}+۲(۳-۵)^{۲}+۹(۵-۵)^{۲}+۵(۵-۵)^{۲}+۴(۶-۵)^{۲}+۲(۷-۵)^{۲}}{۱۶}=۱/۷۵$$

☜ مثال: در بین داده‌های آماری زیر در کدام گزینه واریانس بزرگ‌تر است؟

۱) ۷، ۷، ۱۳، ۱۳ ۲) ٦، ١٦، ١٦، ١٠، ١٠ ۳) ۷، ۸، ۱۲، ۱۳ ٤) ١٠، ١٠، ١٠، ١٠، ١٠

✍حل: تقریباً میانگین همه‌ی گزینه‌ها برابر ۱۰ است. پس گزینه‌ی ۲ دارای پراکندگی بیش‌تری است، چون فاصله‌ی داده‌ها از میانگین بزرگ‌تر است.

☜ مثال: دو نفر در یک آزمایشگاه در ۵ روز متوالی همزمان شروع به کار کرده‌اند. امتیازات دقت کاری آن‌ها مطابق جدول زیر است. دقت کاری کدام بیشتر است؟

نفر اول	۷	۹	۸	۹	۷
نفر دوم	۱۰	۸	٦	۷	۹

۱) نفر اول ۲) نفر دوم

۳) یکسان ٤) نیاز به اطلاعات بیشتر

✍حل:

دقت کاری کسی بیش‌تر است که پراکندگی امتیازش، کم‌تر باشد که نفر اول پراکندگی‌اش کم‌تر است. (هر ۲ نفر میانگین یکسانی می‌دارند)

نکات:

۱) واریانس را می‌توان از دستور زیر نیز محاسبه کرد:

$$\sigma^{۲}=\frac{\sum_{۱}^{n}x_i^{۲}}{n}-\left(\frac{\sum_{۱}^{n}x_i}{n}\right)^{۲}=\frac{\sum_{۱}^{n}x_i^{۲}}{n}-(\bar{X})^{۲}$$

نکته: واریانس را می‌توان در جدول فراوانی طبقه‌بندی نشده از رابطه‌ی زیر نیز به‌دست آورد.

$$\sigma^{۲}=\frac{f_{۱}(x_{۱}-\bar{X})^{۲}+f_{۲}(x_{۲}-\bar{X})^{۲}+\dots+f_n(x_n-\bar{X})^{۲}}{f_{۱}+f_{۲}+\dots+f_n}$$

x_i	$x_{۱}$	....	x_n
f_i	$f_{۱}$	....	f_n

$$\sigma^{۲}=\frac{\sum f_i x_i^{۲}}{\sum f_i}-\left(\frac{\sum f_i x_i}{\sum f_i}\right)^{۲}$$

مثال: اگر $\sum x_i^2 = ۱۰۰$ و $\sum x_i = ۲۰$ و تعداد داده‌ها برابر ۱۰ باشد، واریانس داده‌ها کدام است؟

حل:

$$\sigma^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \frac{100}{10} - \left(\frac{20}{10}\right)^2 = 6$$

۲) اگر واریانس داده‌های آماری $x_1,...,x_n$، σ^2 باشد، واریانس داده‌های $ax_1+b,...,ax_n+b$ برابر است با $a^2\sigma^2$.

$$\sigma^2 = \frac{\sum\left((ax_i+b)-(a\bar{X}+b)\right)^2}{n} = a^2\underbrace{\frac{(x_i-\bar{X})^2}{n}}_{\sigma^2}$$

مثال: اگر واریانس در سال گذشته ۱۰۰۰ ریال باشد و امسال ۱۰٪ به قیمت‌ها افزوده شده باشد، واریانس قیمت‌های جدید چقدر است؟

حل:

$$x_{\text{جدید}} = x_{\text{قدیم}} + \frac{1}{10}x_{\text{قدیم}} = \frac{11}{10}x_{\text{قدیم}}$$

$$\sigma^2_{\text{جدید}} = |a|^2\sigma^2_{\text{قدیم}} = \left(\frac{11}{10}\right)^2 \times 1000 = 1210$$

مثال: اگر واریانس داده‌های $x_1,...,x_n$ برابر صفر باشد میانگین داده‌های $x_1-2x_n, x_2-2x_n,...,x_n-2x_n$ چقدر است؟

حل:

اگر یکی از پارامترهای پراکندگی برابر صفر باشد. یعنی همه‌ی داده‌ها برابرند. پس: $x_1 = .. = x_n$

پس داده‌ها عبارتند از: $x_n-2x_n, x_n-2x_n,...,x_n-2x_n$ میانگین همان است. $-x_n$

مثال: اگر واریانس x_1, x_2, x_3, x_4, x_5 برابر صفر باشد، واریانس داده‌های $x_1, x_2+1, x_3+2, x_4+3, x_5+4$ چقدر است؟

حل:

$$x_1 = x_2 = ... = x_5$$

$$\to x_1, x_1+1, x_1+2, x_1+3, x_1+4 \to \bar{x} = x_1+2 \to \sigma^2 = \frac{2^2+1^2+1^2+2^2}{5} = 2$$

مثال: واریانس داده‌های جدول زیر را به‌دست آورید.

x_i	۳۲	۴۰	۵۶	۷۲
f_i	۲	۳	۱	۲

حل: ابتدا داده‌ها را بر ۸ تقسیم می‌کنیم:

x_i	۴	۵	۷	۹
f_i	۲	۳	۱	۲

$$\to \bar{x} = \frac{4\times2+5\times3+7\times1+9\times2}{2+3+1+2} = \frac{48}{8} \to \sigma^2 = \frac{2(2)^2+3(1)^2+1(1)+2(3)^2}{2+3+1+2} = \frac{30}{8} = \frac{15}{4} \to \sigma^2_{\text{قدیم}} = \left(8^2\times\frac{15}{4}\right) = 240$$

۳) در حالت کلی واریانس داده‌هایی که تصاعد حسابی با قدر نسبت d برابر است با: $\sigma^2 = \underset{\text{قدر نسبت}}{d^2}\,\frac{n^2-1}{12}$

واریانس داده‌های $1,2,...,n$ برابر است با:

$$\sigma^2 = \frac{n^2-1}{12}$$

$$\sigma^2 = \frac{\sum x_i^2}{n} - (\bar{X})^2 = \frac{1^2+...+n^2}{n} - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{n}\right)^2 = \frac{n^2-1}{12}$$

مثال: واریانس داده‌های آماری ۴۰۱، ۴۰۳، ۴۰۵، ۴۰۷ را به‌دست آورید.

حل:

ابتدا از هر داده ۴۰۰ واحد کم می‌کنیم و می‌دانیم واریانس آن تغییر نمی‌کند. حال داده‌های جدید تصاعد حسابی با قدر نسبت ۲ می‌سازند:

۱,۳,۵, ۷

$$\sigma^{۲} = ۲^{۲} \times \frac{۴^{۲}-۱}{۱۲} = ۵$$

۳) انحراف معیار:

از جذر مثبت واریانس، انحراف معیار به‌دست می‌آید که تفاوت آن با واریانس در این است که انحراف معیار از جنس داده‌ها می‌باشد **(هم‌واحد با داده‌هاست)** ولی واریانس از جنس مربع داده‌ها می‌باشد. مثلاً اگر داده‌های ما طول‌هایی برحسب سانتی‌متر باشند، انحراف معیار دارای واحد سانتی‌متر و واریانس سانتی‌متر مربع می‌باشد. برای داده‌های $x_1,\dots,x_n$ انحراف معیار از دستور زیر به‌دست می‌آید.

$$\sigma = \sqrt{\frac{\sum_{۱}^{n} (x_i - \overline{X})^{۲}}{n}}$$

در جدول داده‌های دسته‌بندی نشده:

$$\sigma = \sqrt{\frac{\sum_{۱}^{n} f_i (x_i - \overline{X})^{۲}}{n}}$$

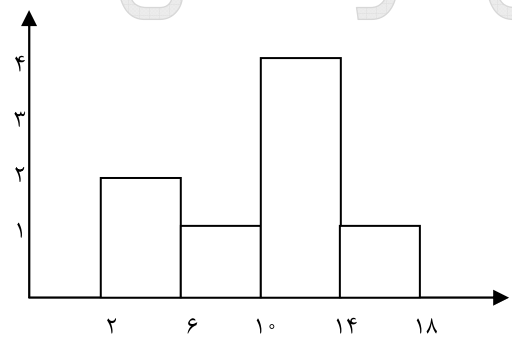
و در جدول داده‌های دسته‌بندی شده:

$$\sigma = \sqrt{\frac{\sum_{۱}^{n} f_i (x_i^{*} - \overline{X})^{۲}}{n}}$$

همچنین از رابطه زیر نیز می‌توان استفاده کرد:

$$\sigma = \sqrt{\frac{\sum_{۱}^{n} x_i^{۲}}{n} - (\overline{X})^{۲}} = \sqrt{\frac{\sum_{۱}^{n} x_i^{۲}}{n} - \left(\frac{\sum x_i}{n}\right)^{۲}} \quad \text{یا} \quad \sigma = \sqrt{\frac{\sum_{۱}^{n} f_i\,(x_i)^{۲}}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^{۲}}$$

مثال: با توجه به نمودار زیر انحراف معیار داده‌ها را به‌دست آورید.



حل:

$$\overline{x} = \frac{۴\times۲+۸\times۱+۱۲\times۴+۱۶\times۱}{۲+۱+۴+۱} = \frac{۸۰}{۸} = ۱۰$$

$$\sigma = \sqrt{\frac{۲\times۴^{۲}+۱\times۸^{۲}+۴\times۱۲^{۲}+۱\times۱۶^{۲}}{۸} - (۱۰)^{۲}} = \sqrt{۱۶} = ۴$$

مثال: مجموع مربع ده داده برابر ۲۲۵۰ و میانگین این ده داده برابر $۱۰\sqrt{۲}$ می‌باشد. انحراف معیار این داده‌ها کدام است؟

حل:

$$\sigma = \sqrt{\frac{۲۲۵۰}{۱۰} - (۱۰\sqrt{۲})^{۲}} = \sqrt{۲۲۵ - ۲۰۰} = ۵$$

☜ مثال: کدام مقدار را نمی‌توان از روی نمودار میله‌ای فراوانی داده‌های دسته‌بندی شده به‌صورت قطعی به‌دست آورد؟

۱) میانگین دسته‌ها ۲) انحراف معیار دسته‌ها ۳) فراوانی نسبی دسته‌ها ٤) دامنه‌ی تغییرات

حل:

میانگین و انحراف معیار و فراوانی از روی خود داده‌ها و یا از روی نشان دسته‌ها قابل حصول است، اما دامنه‌ی تغییرات دقیقاً با خود داده‌ها سر و کار دارد، لذا از نشان دسته نمی‌توان پی به دامنه برد.

☜ مثال: اگر انحراف معیار داده‌های $x_1, x_2, ..., x_n$، ۱۶ برابر صفر باشد، میانگین داده‌ها کدام است؟

حل:

چون یکی از شاخص‌های پراکندگی صفر است پس:

$x_1 = x_2 = ...x_n = 16$

پس میانگین نیز برابر است با یکی از داده‌ها:

$\bar{x} = 16$

نکات:

۱) اگر انحراف معیار داده‌های $x_1, ..., x_n$، σ باشد، انحراف معیار داده‌های $ax_1 + b, ..., ax_n + b$ برابر است با:

$\sigma^2_{\text{جدید}} = a^2\sigma^2_{\text{قدیم}} \Rightarrow \sigma_{\text{جدید}} = |a|\sigma_{\text{قدیم}}$

☜ مثال: اگر انحراف معیار داده‌های ۱، ۳، ۵، ۷، σ باشد، انحراف معیار داده‌های ۳، ۷، ۱۱، ۱۵ کدام است؟

حل:

چون داده‌ها در ۲ ضرب شده و با ۱ جمع شده‌اند، انحراف معیار در ۲ ضرب می‌شود.

☜ مثال: اگر انحراف معیار $x_1, ..., x_n$، برابر ٤ باشد، واریانس داده‌های $-2x_1 + 3, ..., -2x_n + 3$ کدام است؟

حل:

$\sigma^2 = (-2)^2 \times 16 = 64$

☜ مثال: اگر داده‌های آماری مفروضی را در اختیار داشته باشیم و $\bar{X}$ میانگین این داده‌ها باشد با توجه به جدول مقابل، انحراف معیار داده‌ها، تقریباً برابر کدام است؟

حل:

f_i	۳	۱	۲	۵	۱۱	۵
$x_i - \bar{X}$	۵	۲	۱	۲-	t	٤-

$\sum f_i(x_i - \bar{x}) = 0 \to 3(5) + 1(2) + 2(1) + 5(-2) + 11(t) + 5(-4) = 0 \to t = 1$

$\to \sigma = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i}} = \sqrt{\frac{3\times5^2 + 1\times2^2 + 2\times1^2 + 5\times(-2)^2 + 11(1)^2 + 5(4)^2}{27}} = \sqrt{\frac{64}{9}} = \frac{8}{3}$

۲) در حالت کلی انحراف معیار داده‌هایی که تصاعد حسابی با قدر نسبت d برابر است با:

$\sigma = \underset{\text{قدر نسبت}}{\underset{\downarrow}{d}}\frac{\sqrt{n^2 - 1}}{12}$

انحراف معیار داده‌های $1, 2, ..., n$ برابر است با:

$\sigma = \sqrt{\frac{n^2 - 1}{12}}$

☜ **مثال: انحراف معیار ۴۰۱، ۴۰۲، ۴۰۳، ۴۰۴، ۴۰۵ را به‌دست بیاورید.**

✍**حل:**

ابتدا از تمام داده‌ها ۴۰۰ واحد کم می‌کنیم، بدون آن‌که انحراف معیار تغییر کند انحراف معیار جدید برابر است با:

$$\sigma=\sqrt{\frac{۵^۲-۱}{۱۲}}=\sqrt{۲}$$

۳) اگر در داده‌های آماری $x_n,\dots,x_1$ میانگین $\overline{X}$ و انحراف معیار σ باشد، در داده‌های $Y_i=\frac{x_i-\overline{X}}{\sigma}$ میانگین صفر و واریانس ۱ است.

$$\sigma^۲_{\text{جدید}}=a^۲\sigma^۲_{\text{قدیم}}=(\frac{۱}{\sigma})^۲\sigma^۲=۱$$

این کار را استانداردسازی داده‌ها می‌نامند و داده‌های جدید، داده‌های استاندارد نام دارند.

۴) ضریب تغییرات:

اگر بخواهیم معیاری را معرفی کنیم که در آن پراکندگی به نسبت میانگین (که با بزرگی داده‌ها در ارتباط است) سنجیده شود، از معیار جدیدی به‌نام ضریب تغییرات استفاده می‌کنیم.

$$CV=\frac{\sigma}{\overline{X}}$$

ضریب تغییرات بدون واحد است که آن را معمولاً به‌صورت درصد بیان می‌کنند.

ضریب تغییرات از آن جهت حائز اهمیت است که امکان مقایسه داده‌های غیر هم‌واحد را فراهم می‌سازد.

☜ **مثال: ضریب تغییرات داده‌های ۸، ۱۰، ۱۲، ۱۴ کدام است؟**

✍**حل:**

$$\left.\begin{array}{l}\bar{x}=\frac{۱۴+۱۲+۱۰+۸}{۴}=\frac{۴۴}{۴}=۱۱\\ \to\sigma=\sqrt{\frac{۳^۲+۱^۲+۱^۲+۳^۲}{۴}}=\sqrt{۵}\end{array}\right\}\Rightarrow c.v=\frac{\sqrt{۵}}{۱۱}$$

☜ **مثال: در ۵۰ داده‌ی آماری مجموع تمام داده‌ها ۱۰۰ و مجموع مربعات داده‌ها ۲۷۲ می‌باشد. ضریب تغییرات کدام است؟**

✍**حل:**

$$\sigma=\sqrt{\frac{۲۷۲}{۵۰}-(\frac{۱۰۰}{۵۰})^۲}=\sqrt{\frac{۷۲}{۵۰}}=\sqrt{\frac{۳۶}{۲۵}}=\frac{۶}{۵}\to CV=\frac{\sigma}{\bar{x}}=\frac{\frac{۶}{۵}}{۲}=\frac{۳}{۵}$$

نکته: اگر همه‌ی داده‌های آماری در عدد ثابت مثبتی ضرب شوند، ضریب تغییرات ثابت می‌ماند.

اما اگر همه‌ی داده‌های آماری با عدد ثابت مثبتی جمع شوند، ضریب تغییرات جدید از ضریب تغییرات اولیه کوچک‌تر است. و اگر همه‌ی داده‌های آماری با عدد ثابت منفی‌ای جمع شوند، ضریب تغییرات جدید از ضریب تغییرات اولیه بزرگتر است. (به شرط آن‌که علامت ضریب تغییرات تغییر نکند.)

$$CV=\frac{\sigma}{\overline{X}}$$

☜ **مثال: اگر ضریب تغییرات داده‌هایی برابر ٤ باشد و داده‌ها را دو برابر کنیم، ضریب تغییرات جدید کدام است؟**

✍**حل:**

هم میانگین، هم ضریب تغییرات دو برابر می‌شود، پس ضریب تغییرات، تغییر نمی‌کند.

☜ **مثال: در ۶۰ داده‌ی آماری میانگین ۳ و انحراف معیار ۱/۲ محاسبه شده است. اگر به تمام داده‌ها ۹ واحد اضافه شود، ضریب تغییرات داده‌های جدید کدام است؟**

✍**حل:**

$$\left.\begin{array}{l} \text{میانگین جدید} = 3+9 = 12 \\ \text{انحراف معیار جدید} = 1/2 \end{array}\right\} \to C.V = \frac{1/2}{12} = \frac{1}{10}$$

☜ **مثال: در داده‌های آماری با میانگین $\overline{X}$ و انحراف معیار σ اگر به هریک از داده‌ها، مقدار $\overline{X}$ را اضافه کنیم تا داده‌های جدید حاصل شود، ضریب تغییرات داده‌های جدید چند برابر ضریب تغییرات داده‌های قبلی است؟**

✍**حل:**

$$\left.\begin{array}{l} \text{میانگین جدید} = \overline{x}+\overline{x} = 2\overline{x} \\ \text{میانگین معیار جدید} = \sigma \end{array}\right\} \to CV = \frac{\sigma}{2\overline{x}} = \frac{1}{2} CV_{\text{قدیم}}$$

☜ **مثال: اگر ۲۰ داده‌ی آماری را دو برابر کرده و سپس ۷ واحد از هرکدام کم کنیم، ضریب تغییرات داده‌های جدید، ۱/۵ برابر ضریب تغییرات داده‌های قبلی می‌شود. مجموع داده‌های قبلی کدام است؟**

✍**حل:**

$$\left.\begin{array}{l} \overline{x}_{\text{جدید}} = 2\overline{x}_{\text{قدیم}} - 7 \\ \sigma_{\text{جدید}} = 2\sigma_{\text{قدیم}} \end{array}\right\} \to CV_{\text{جدید}} = \frac{2\sigma}{2\overline{x}-7} = \frac{3}{2}\frac{\sigma}{\overline{x}} \to 4\sigma\overline{x} = 6\sigma\overline{x} - 21\sigma \to 2\sigma\overline{x} = 21\sigma \to \overline{x} = \frac{21}{2}$$

$$\overline{x} = \frac{\sum x_i}{20} = \frac{21}{2} \to \sum x_i = 210$$

دو نکته‌ی کلی در مورد پارامترهای پراکندگی:

۱) شرط لازم و کافی برای آن‌که در یک نمونه‌ی آماری تمام اعضا با هم برابر باشند آن است که تمام پارامترهای پراکندگی برابر صفر باشد.

$$R = \sigma^2 = \sigma = CV = 0$$

که در این صورت میانگین داده‌ها با خود داده‌ها برابر است.

۲) دقت پارامترهای واریانس، انحراف معیار و دامنه‌ی تغییرات به‌صورت زیر است:

۱) واریانس　　۲) انحراف معیار　　۳) دامنه‌ی تغییرات

هرچه دامنه‌ی تغییرات بزرگ‌تر باشد، باعث بزرگ‌تر شدن انحراف معیار و واریانس می‌شود و برعکس.