آمار و احتمال مهندسی
فصل اول آمار توصـــیفی

مقدمه

امروزه در علوم پایه و مهندسی به دفعات نـاز به جمع آوری اطلاعات در مورد مجموعه‌های از اشیاء و یا انسانها داریم که این امر بر عهده علم آمار می‌باشد و با کمک آمار توصیفی می‌توان اطلاعات جمع آوری شده را به صورتی منظم گرد آوری نمود بطوریکه بتوان با یک نگاه اجمالی به نتایج بدست آمده یک دید کلی نسبت به کل داده‌ها بدست آورد.

در این فصل به چگونگی جمع آوری، اطلاعات و تجزیه و تحلیل آن با کمک نمودارها و پارامترهای مرکزی و پراکندگی می‌پردازیم.

۱-۱ تعاریف اولیه

جامعه آماری : به مجموعه‌ای از اشیاء یا افراد که حداقل یک ویژگی مشترک آنها مورد مطالعه قرار می‌گـیرد. جامعه آماری می‌گوییم. ویژگی‌های مشترک یک جامعه آماری از عضوی به عضو دیگر تغییر می‌کند. که آنها را متغیر می‌نامند. متغیرها معمولاً به دو نوع تقسیم می‌شوند:

متغیر کمی : متغیرهایی می‌باشند که معمولاً قابل اندازه‌گیری هستند و می‌توان مقدار آنها را به صورت عددی نمایش داد مثل مقدار وزن، قد، حجم و ....

۱- متغیر کیفی : متغیرهایی می‌باشند که مستقیماً توسط اعداد و ارقام قابل اندازه‌گیری نیستند. مثل گروه خونی، شغل، رنگ چشم و ... که برای اندازه‌گیری این متغیرها به آنها عددی نسبت می‌دهیم.

متغیرهای کمی خود بر دو نوع هستند.

۱- گسسته : متغیرهایی که بین دو مقدار متصور آنها هیچ عدد دیگری وجود نداشته باشد.

۲- پیوسته : متغیرهایی که بین هر دو مقدار متصور آنها همواره عددی دیگری وجود دارد. مثل وزن یا طول و قد افراد.

پس از جمع آوری داده‌ها برای رسیدن به اهداف مورد نیاز به بررسی و تجزیه و تحلیل داده‌ها داریم که برای این منظور ابتدا داده‌ها را در یک جدول تنظیم و طبقه‌بندی می‌کنیم و سپس با استفاده از نمودارهای آماری نحوه توزیع داده‌ها را نمایش می‌دهیم و در نهایت داده‌ها را با کمک چند عدد به نام شاخص یا آماره خلاصه می‌کنیم.

۱-۲ جدول آماری

در جداول آماری علاوه بر داده‌ها، تعداد و درصد تکرار آنها نمایش داده می‌شود بطوریکه با یک نگاه به جدول می‌توان اطلاعات مفیدی در مورد پراکندگی و توزیع داده‌ها بدست آورد. یکی از متداول‌ترین جداول آماری جدول فراوانی است در جداول فراوانی همواره موارد زیر را خواهیم داشت:

۱- فراوانی: با شمردن تعداد دفعات تکرار هر داده در میان تمامی داده‌ها فراوانی آن داده بدست می‌آید. فرض کنید N داده داشته باشیم با قرار دادن داده‌های مشابه در یک دسته، در نهایت K طبقه خواهیم داشت $(K \le N)$ که تعداد داده‌ها در هر دسته را فراوانی آن داده می‌نامیم. فراوانی طبقه i ام را با f_i نمایش می‌دهیم که $1 \le i \le k$ و $1 \le f_i \le N$ و $\sum_{i=1}^{k} f_i = N$ (N تعداد کل داده‌ها)

۲- فراوانی نسبی : عبارتست از حاصل تقسیم فراوانی هر طبقه بر تعداد کل داده‌ها که آنرا با $r_i = \frac{f_i}{N}$ ، $i = 1, \cdots k$ نمایش می‌دهیم.

$$\sum_{i=1}^{k} r_i = \sum_{i=1}^{k} \frac{f_i}{N} = \frac{1}{N} \sum_{i=1}^{k} f_i = 1$$

هم چنین اگر فراوانی نسبی را در عدد ۱۰۰ ضرب کنید درصد وقوع هر داده را در میان کل داده‌ها بدست می‌آورید.

۳- فراوانی تجمعی نسبی : حاصل جمع فراوانی هر طبقه با طبقات قبل از آنرا فراوانی تجمعی می‌نامیم و به g_i $(1 \le i \le k)$ نمایش می‌دهیم.

فراوانی تجمعی طبقه iام $= g_i = f_1 + f_2 + \cdots + f_1 = \sum_{j=1}^{i} f_j$

به همین ترتیب حاصل جمع فراوانی نسبی هر طبقه با طبقات قبل آنرا فراوانی تجمعی نسبی می‌نامیم و به s_i $(1 \le i \le k)$ نمایش می‌دهیم

فراوانی تجمعی طبقه iام $= s_i = r_1 + r_2 + \cdots + r_1 = \sum_{j=1}^{i} s_j = s_i = \frac{1}{N} g_i$

توجه : در محاسبه g_i و s_i اگر x_i نماینده طبقه iام باشد می‌بایستی داده‌ها به صورت مرتب و از کوچک به بزرگ در جدول قرار داده شوند بطوریکه داشته باشیم $x_1 < x_2 < \cdots x_n$.

۱-۲-۱ جدول فراوانی برای داده های گسسته

در مثال زیر چگونگی تشکیل جدول فراوانی برای داده‌های گسسته را بیان می‌کنیم.

مثال ۱ : داده‌های زیر تعداد فرزندان تحت پوشش بیمه در ۳۰ خانواده را نشان می‌دهد.

5 , 4 , 2 , 1 , 1 , 1 , 3 , 3 , 3 , 2 , 0 , 5 , 0 , 2 , 2 , 2 ,2 , 4 , 4 , 1 , 0 , 1 , 1 , 3 , 2 , 5 , 2 , 1 , 0 , 3

ابتدا داده‌ها را از کوچک به بزرگ برای تشکیل جداول فراوانی مرتب می‌کنیم:

0 , 0 , 0 ,0 , 1 , 1 , 1 , 1 , 1 , 1 , 1 , 2 , 2 , 2 , 2 , 2 ,2 , 2 , 2 , 3 , 3 , 3 , 3 , 4 , 4 , 4 , 4 , 5 , 5 , 5

در این صورت ۶ طبقه بدست خواهد آمد. نماینده هر طبقه در ستون x_i و فراوانی ان در ستون f_i نوشته می‌شود به این ترتیب جدول زیر را خواهیم داشت:

x_i	f_i	r_i	g_i	s_i
۰	۴	۰/۱۳	۴	۰/۱۳
۱	۷	۰/۲۳	۱۱	۰/۳۶
۲	۸	۰/۲۶	۱۹	۰/۶۲
۳	۴	۰/۱۳	۲۳	۰/۷۵
۴	۴	۰/۱۳	۲۷	۰/۸۸
۵	۳	۰/۱	۳۰	۱
جمع	۳۰	۱		

با توجه به جدول می‌توان به نتایج زیر رسید:

۱- با توجه به عدد ۰/۲۶ در ستون فراوانی نسبی می‌توان نتیجه گرفت که ۲۶ درصد از خانوارها دارای ۲ فرزند می‌باشند.

۲- با توجه به عدد ۰/۷۵ در ستون فراوانی تجمعی نسبی می‌توان نتیجه گرفت که ۷۵ درصد خانوارها حداکثر دارای ۳ فرزند می‌باشند.

۱-۲-۲ جدول فراوانی برای داده‌ها پیوسته

مثال ۲ : داده‌های زیر طول عمر ۴۰ عدد لامپ را نشان می‌دهند که به نزدیک‌ترین عدد صحیح گرد شده‌اند.

11	9	12	15	20	13	14	17	23	22
8	16	17	21	11	18	21	12	11	10
14	13	19	16	15	17	20	8	7	13
15	17	16	14	22	12	11	9	18	19

زمانی که با داده‌های پیوسته کار می‌کنیم داده‌ها را به رده‌های (فاصله‌ها) با طول مساوی تقسیم می‌کنیم و فراوانی داده‌ها را در هر رده بدست می‌آوریم. در این حالت نیاز به ثبت تمامی داده‌ها در جدول نمی‌باشد بلکه هر رده را به صورت مرتب از کوچک به بزرگ در جدول ثبت می‌کنیم.

در این حالت روند به این صورت است که :

۱- از آنجا که هر عدد به نزدیکترین عدد صحیح گرد شده است پس مثلاً عدد ۱۲ در واقع عددی بین (۱۲/۵ و ۱۱/۵) بوده است برای اینکه این مقدار نیز در عملیات وارد شود از مفهوم میزان تغیر پذیری داده‌ها که با s نمایش می‌دهیم استفاده می‌کنیم.

$$S=\frac{\text{واحد گرد شده داده‌ها}}{2}=\frac{1}{2}=0/5$$

۲- دامنه واقعی داده‌ها و کوچکترین و بزرگترین داده را به این ترتیب محاسبه می‌کنیم.

$\min=$ کوچکترین داده $\quad -S=7-0/5=6/5$

$\max=$ بزرگترین داده $\quad +S=23+0/5=23/5$

دامنه تغیرات داده‌ها $R=\max-\min=23/5-6/5=11$

۳- با توجه به دامنه داده‌ها می‌توان طول هر رده را معین نمود. برای این می‌بایستی دامنه را بر تعداد رده‌ها (که به صورت دلخواه قابل انتخاب است) تقسیم نمود.

$$\text{طول رده}=\omega=\frac{R}{K} \qquad \text{تعداد رده‌ها}=K$$

معمولاً تعداد رده‌ها طوری انتخاب می‌شوند که هر رده حداقل پنج داده را در برداشته باشد می‌توان از فرمول $K=1+3/322\log_{10}^{N}$ برای تعیین تعداد رده استفاده نمود بنابر این:

$$K=1+3/3222\log_{10}^{40}=6/322\cong 7 \Rightarrow \omega=\frac{17}{7}=2/4\cong 3$$

۴- حالا کافیست ۷ رده به طول ۳ در ستون اول جدول فراوانی ایجاد کنیم و مجدداً مطابق مثال قبل جدول را کامل کنیم. اولین رده از کوچکترین عنصر آغاز می‌شود.

رده‌ها	خط و نشان	x_i	fi	r_i	g_i	s_i
6/5−9/5	𝍸	۸	۵	0/125	۵	0/125
9/5−12/5	𝍸 |||	۱۱	۸	0/2	۱۳	0/325
12/5−15/5	𝍸 ||||	۱۴	۹	0/225	۲۲	0/55
15/5−18/5	𝍸 ||||	۱۷	۹	0/225	۳۱	0/775
18/5−21/5	𝍸		۲۰	۶	0/15	۳۷	0/925
21/5−24/5	|||	۲۳	۳	0/075	۴۰	1/0
جمع				۴۰			

توجه شماره ۱ :

- x_i معرف نماینده هر رده است که عضو میانی رده می‌باشد.
- چون آخرین رده شامل هیچ عضوی نبود آنرا حذف می‌کنیم مثلاً در این مثال رده ۲۴/۵-۲۷/۵ شامل هیچ عضوی نیست بنابر این حذف می‌شود.

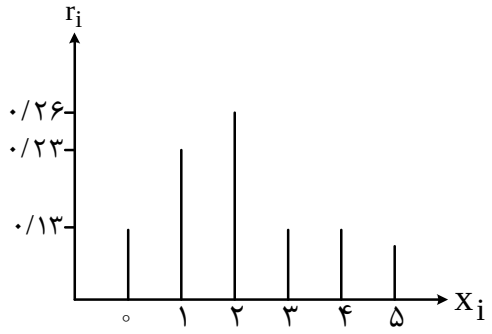
۱-۳ نمودارهای آماری

یکی دیگر از روشهای مناسب برای خلاصه نمودن داده‌های آماری استفاده از نمودار می‌باشد. نمودارها بهترین ابزار برای نمایش نحوه توزیع داده‌ها می‌باشند که در ذیل معروفترین آنها را معرفی می‌کنیم.

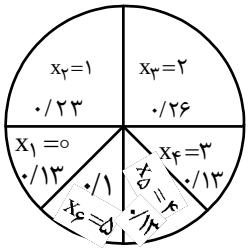
۱-۳-۱ نمودارهای آماری برای داده‌های گسسته

۱- نمودارهای میله‌ای:

در این نمودار محور x ها نمایش دهنده مقادیر داده‌ها و محور y ها نمایش دهنده فراوانی نسبی می‌باشد. چون فراوانی نسبی داده‌های هر جامعه آماری مقادیر بین صفر و یک را می‌گیرد بنابر این می‌توان چندین نمودار را با یکدیگر مقایسه نمود. شکل نمودار میله‌ای را برای مثال ۱ نمایش می‌دهد.



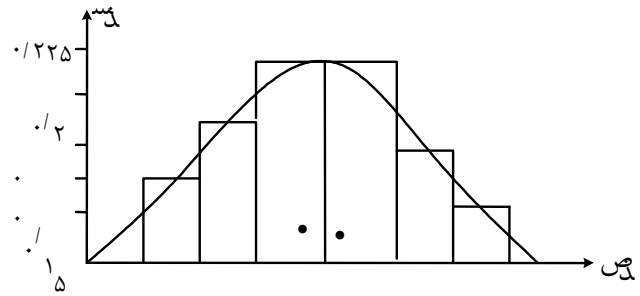
۳- نمودار دایره‌ای : این نمودار درون دایره‌ای رسم می‌شود که به قطاع‌هایی تقسیم شده است و هر قطاع متناسب با مساحت خود معرف یکی از فراوانی‌های نسبی در جدول فراوانی می‌باشد. شکل زیر نمایش دهنده نمودار دایره‌ای برای مثال ۱ می‌باشد.



۱-۳-۲ نمودارهای آماری برای داده‌های پیوسته

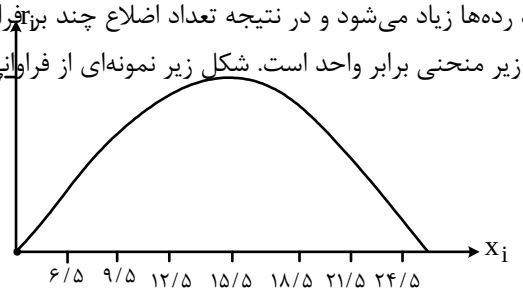
همانطور که در جدول فراوانی مشاهده کردید تفوت عمده رده‌های پیوسته با گسسته در بکارگیری رده‌ها می‌باشد این امر در نمودارها نیز کاملاً صادق است. در این قسمت سه نمودار را که برای نمایش داده‌های پیوسته بکار می‌روند معرفی می‌کنیم.

۱- هیستوگرام (نمودار ستونی) : این نمودار کاملاً مشابه نمودار میله‌ای می‌باشد با این تفاوت که محور xها نمایش دهنده رده‌های جدول فراوانی است و به ازای هر رده مستطیلی که عرض آن یک واحد و ارتفاع آن معادل فراوانی نسبی آن رده است رسم می‌شود. بنابر این مجموع مساحتهای مستطیل‌های رسم شده برابر واحد می‌باشد که همان مجموع کل فراوانی نسبی‌ها است. شکل زیر نمودار هستیوگرام برای مثال ۲ می‌باشد.



۲- چند برابر فراوانی: اگر در نمودار هستیوگرام وسط قاعده‌های بالای مستطیل‌ها را به یکدیگر توسط خطوطی به صورت متوالی متصل کنیم نمودار چند برابر فراوانی بدست می‌آید که در شکل بالا نیز نشان داده شده است. ابتدای خطوط به وسط رده ماقبل و انتهای خطوط به وسط رده مابعد مستطیلهای هستیوگرام متصل می‌شوند به این ترتیب مساحت زیر نمودار چند بر فراوانی مجدداً برابر واحد خواهد بود.

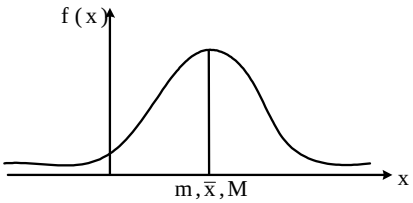
۳- منحنی فراوانی: در صورتی که تعداد داده‌ها زیاد باشد و طول رده‌ها کوچک، تعداد رده‌ها زیاد می‌شود و در نتیجه تعداد اضلاع چند برفراوانی افزایش یافته و در نهایت مشابه یک منحنی خواهد شد که در این حالت نیز مساحت زیر منحنی برابر واحد است. شکل زیر نمونه‌ای از فراوانی نسبت.

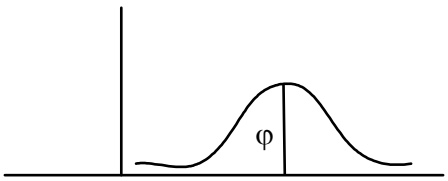


متغیر تصادفی نرمال: اگر مجموعه داده‌های X دارای میانگین $\bar{x}$ و واریانس σ^2 باشند و نمودار آنها از تابع

$$(-\infty < x < +\infty) \qquad f_{(x)} = \frac{1}{\sqrt{2\pi\sigma}}\, e^{\frac{-(x-\bar{x})^2}{2^2}}$$

تبعیت کند می‌گوییم متغیر تصادفی X نرمال می‌باشد. نمودار متغیر تصادفی نرمال نسبت به خط $x=\bar{x}$ متقارن می‌باشد. به شکل زیر توجه کنید.





همچنین با افزایش مقدار δ نمودار در راستای محور xها کشیده تر می‌شود.
با افزایش مقدار $\bar{x}$ نمودار به سمت راست و با کاهش آن به سمت چپ جابجا می‌شود.

۱-۴ پارامترهای مرکزی و پراکندگی

از آنجا که با مطالعه یک جامعه آماری تعداد زیادی داده بدست می‌آوریم و مطالعه روی تک تک یا قسمتی از این داده‌ها مشکل و حتی غیر ممکن است همواره علاقه داریم این داده‌ها را با کمک شاخص‌ها و پارامترهایی خلاصه کنیم تا بتوان با یک نگاه اجمالی به آن یک دید کلی نسبت به کل داده‌ها و جامعه آماری بدست آورد. بنابر این پارامترهای مرکزی و پراکندگی را معرفی می‌کنیم.

۱-۴-۱ پارامترهای مرکزی

عموماً داده‌ها در جامعه آماری یکنوع تجمع .و فشردگی حول یک مقدار خاص از صفت مورد مطالعه را بوجود می‌آورند که این مقدار خاص به عنوان یک پارامتر مرکزی معرفی می‌شود. مهمترین پارامترهای مرکزی عبارتند از: میانگین، میانه و مد (نما)

۱- میانگین: میانگین بر چند نوع است که معروفترین آنها میانگین حسابی، خندسی، همساز (هارمونیک) و درجه دوم می‌باشد.

الف) میانگین حسابی: اگر داده های $x_1, x_2, \ldots, x_k$ به ترتیب دارای فراوانی $f_1, f_2, \ldots, f_k$ باشند و تعداد کل این داده‌ها N باشد $\bar{X}$ یا میانگین حسابی به صورت زیر تعریف می‌شود:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{K} f_i\, x_i$$

برای داده‌های پیوسته x_i را نماینده هر رده در نظر می‌گیریم. به عنوان مثال میانگین حسابی برای مثال ۱ و ۲ عبارتست از:

$$\text{مثال 1} \quad : \bar{X} = \frac{1}{30}\,(4\times 0 + 7\times 1 + 8\times 2 + 4\times 3 + 4\times 4 + 3\times 5) = 2/2$$

$$\text{مثال 2} \quad : \bar{X} = \frac{1}{40}\,(8\times 5 + 11\times 8 + 14\times 9 + 17\times 9 + 20\times 6 + 23\times 3) = 14/9$$

ب) میانگین هندسی: در صورتی که همگی داده‌ها مثبت باشند میانگین هندسی به صورت زیر محاسبه می‌شود.

$$G=\left(\prod_{i=1}^{N} x_i^{f_i}\right)^{\frac{1}{N}}$$

ج: میانگین همساز: (همنوا یا هارمونیک) : اگر هیچکدام از داده ها صفر نباشد می‌توان میانگین همساز را از طریق فرمول محاسبه نمود.

$$H=\frac{N}{\sum_{i=1}^{n}\frac{f_i}{x_i}}$$

در حالت کلی رابطه $H \le G \le \overline{X}$ بین این سه میانگین برقرار است. که حالت تساوی زمانی رخ می دهد که همگی داده‌ها برابر باشند.

۲- میانه: با مرتب نمودن داده‌ها به صورت غیر نزولی عدد m را بعنوان میانه آنها در نظر می‌گیریم اگر تقریباً نیمی از داده‌ها سمت چپ آن و نیمی دیگر در سمت راست آن قرار داشته باشند.

همچنین با توجه به جدول فراوانی، میانه کوچکترین مقدار x است که فراوانی تجمعی آن بیشتر یا مساوی با $\frac{N}{۲}$ باشد.

- برای داده‌های گسسته در صورتی که داده‌ها را به صورت غیر نزولی مرتب کنیم اگر تعداد داده‌ها فرد باشد میانه داده وسطی یا به عبارتی $m=x_{(\frac{n+1}{2})}$ خواهد بود و اگر تعداد داده‌ها زوج باشد، میانه میانگین دو داده وسطی می‌باشد

$$m=\frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2} \quad ; m=\frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}=\frac{2+2}{2}=4$$

- برای داده‌های پیوسته ابتدا می‌بایستی رده میانه دار را پیدا نمود. رده میانه دار اولین رده‌ای است که فراوانی تجمعی نسبی آن از ۰/۵ بیشتر است. میانه عددیست که در این رده قرار دارد. حال برای محاسبه آن از فرمول زیر استفاده می‌کنیم:

$$m=L_{0/5}+\frac{(0/5N-g_{0/5})\omega}{f_{0/5}}$$

که در آن :

$L_{0/5}$: کران پایین رده میانه دار

N : تعداد داده ها

$g_{0/5}$: فراوانی تجمعی رده قبل از رده میانه دار

$f_{0/5}$: فراوانی رده میانه دار

ω : طول رده

به عنوان مثال میانه برای مثال ۱ با توجه به تعداد زوج داده‌ها عبارتست از:

برای مثال ۲ نیز میانه برابر است با:

$$m=12/5+\frac{(0/5\times 40-13)3}{9}=14/8$$

۳- مد یا نما: داده‌ای که فراوانی آن از سایر داده‌ها بیشتر باشد مد یا نما گفته می‌شود و با M نمایش داده می‌شود.

الف) روش محاسبه مد برای داده‌های گسسته: پس از بدست آوردن فراوانی داده، داده‌های که فراوانی آن از بقیه بیشتر باشد مد می‌باشد در این صورت سه حالت زیر پیش می‌آید: (که در مثال زیر به آن می‌پردازیم)

مثال ۳ : برای داده‌های مثال ۱ تنها عدد ۲ دارای بیشترین فراوانی می‌باشد. پس عدد ۲ مقدار مد خواهد بود. اما برای داده‌های ۵ و ۴ و ۴ و ۲ و ۲ و ۲ و۱ و ۱ و ۱ چون فراوانی ۱ و ۲ برابر است و این دو عدد به صورت متوالی قرار گرفته‌اند در این حالت مد برابر است با میانگین این دو عدد:

$$M=\frac{1+2}{2}=\frac{3}{2}$$

اما برای داده‌های ۴ و ۴ و۲ و ۲ و ۱ و ۱، فراوانی ۱ و ۴ مساوی می‌باشد ولی مجاور یکدیگر نیستند دو مقدار برای مد خواهیم داشت که عبارتند از:

$M_1 = 1 \qquad M_2 = 4$

توجه کنید در صورتی که فراوانی همه داده‌ها برابر باشد داده‌ها بدون مد در نظر گرفته می‌شوند.

ب) محاسبه مد برای داده‌های پیوسته: رده‌ای که فراوانی آن از سایر رده‌ها بیشتر است را به عنوان رده نمایی انتخاب می‌کنیم در این حالت می‌توان نماینده رده را به عنوان مد یا نما انتخاب کرد. اما برای محاسبه دقیق تر از فرمول زیر نیز می‌توان استفاده نمود:

$$M = L_M + \left(\frac{D_1}{D_1 + D_2} \right) \omega$$

L_M: کران پایین رده نمایی.

D_1: اختلاف فراوانی‌های نسبی رده نمایی و رده قبل از آن.

D_2: اختلاف فراوانی‌های نسبی رده نمایی و رده بعد از آن.

ω: طول رده.

در مثال ۲ داریم:

$M_1 = 14 \quad ; \quad M_2 = 17$

$$M = \frac{14+17}{2} = 15/5$$

و یا بصورت دقیق تر:

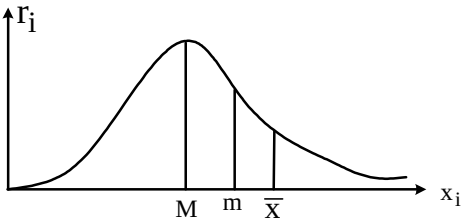
$$M_1 = 12/5 + \left(\frac{0/025}{0/025+0}\right) \times 3 = 15/5$$

$$M_2 = 15/5 + \left(\frac{0}{0+0/075}\right) \times 3 = 15/5$$

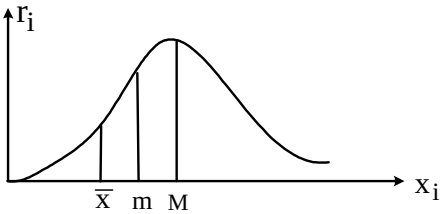
$$\Rightarrow M = 15/5$$

۱-۵ چولگی توزیع فراوانی

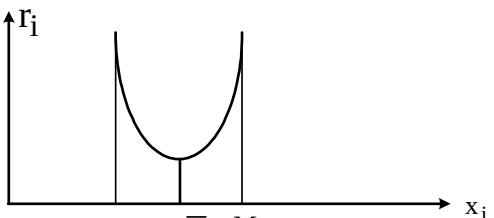
با رسم نمودار فراوانی داده‌های پیوسته عموماً یکی از اشکال زیر بدست می‌آید. که چگونگی قرار گرفتن میانگین و میانه و مد را در اشکال زیر مشاهده می‌شود.



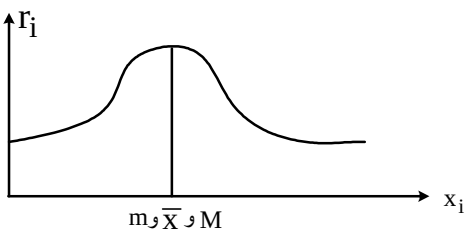
$M < m < \overline{X}$ چوله به راست



$\overline{X} < m < M$ چوله به چپ



$m = \overline{X} = M$ متقارن



$M = \overline{X} = m$ متقارن

برای محاسبه میزان عدم تقارن منحنی، شاخصی به نام B_1 (ضریب چولگی) وجود دارد که در فصل‌های بعدی با آن آشنا می‌شوید.
در توزیع‌هایی که چولگی زیاد نباشد رابطه تجربی زیر که به رابطه پیرسن معروف است برقرار می‌باشد.

$$\overline{X} - M \cong 3(\overline{X} - m)$$

مقدار عدم تقارن منحنی که با ضریب چاولگی سنجیده می‌شود را با نمودار نرمال مقایسه می‌کنیم برای بدست آوردن ضریب چاولگی از مفهوم گشتاور مرتبه kام طول میانگین (گشتاور مرکزی مرتبه kام) استفاده می‌کنیم. که عبارتست از :

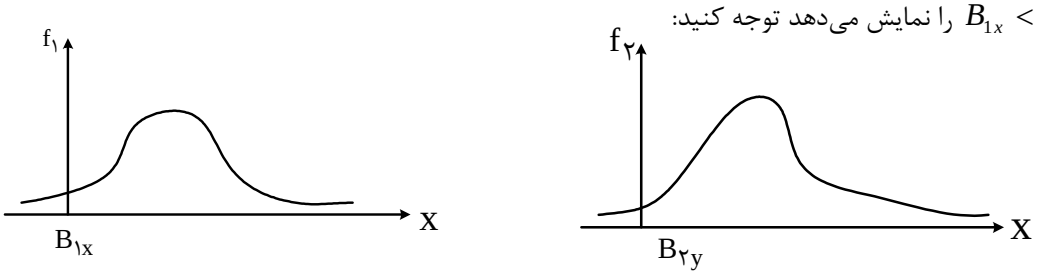
$$\mu_k = \frac{1}{N} \sum_{i=1}^{n} f_i \ (x_i - \overline{x})^k$$

ضریب چاولگی که با B_1 نمایش داده می‌شود میزان عدم تقارن منحنی را نسبت به منحنی نرمال نمایش می‌دهد که به شکل زیر تعریف می‌شود:

$$B_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\delta^3}$$

در صورتی که منحنی متقارن باشد مقدار B_1 برابر صفر می‌باشد و هر چه B_1 مقدار بزرگتری داشته باشد نشان دهنده افزایش عدم تقارن می‌باشد.

به شکل زیر که دو مقدار $B_{1x} < B_{1y}$ را نمایش می‌دهد توجه کنید:

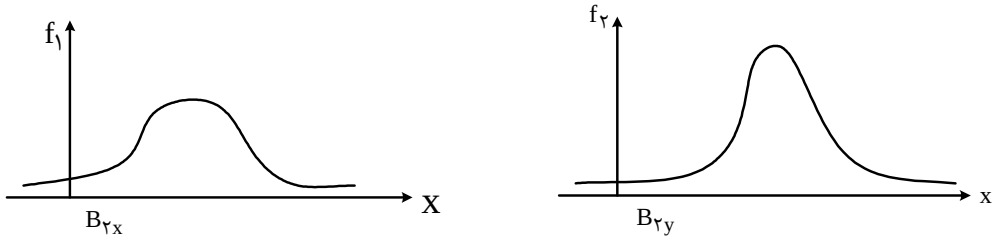


ضریب کشیدگی : نشان دهنده میزان کشیدگی منحنی می‌باشد و به صورت زیر تعریف می‌شود.

$$B_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\delta^4}$$

دو منحنی زیر متقارن می‌باشند اما ضریب کشیدگی در آنها متفاوت است.

$B_{2x} < B_{2y}$



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:

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

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

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

$$\overline{X} = 14/9 \qquad N = 40$$

$$S^2 = \frac{1}{40-1}\left[(50\times 64 + 8\times 121 + 9\times 196 + 9\times (17)^2 + 6\times 40 + 3\times (23)^2 - 40\times 149)^2\right]$$

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؟

$$X_A = 20000 \qquad \overline{X}_B = 18000 \quad \Rightarrow \quad CV_A = \frac{2000}{20000} = 0/1$$

$$S_A = 2000 \qquad S_B = 200 \quad \Rightarrow \quad CV_B = \frac{200}{180000} = 0/01$$

$$CV_A = 0/1 \times 100 = \%10$$

$$CV_A = 0/01 \times 100 = \%1$$

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:

$$x_1, x_2, \ldots, x_n \quad \rightarrow \quad ax_1, ax_2, \ldots, ax_n$$

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

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:

$$m' = am + b$$

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

۱۳-۱ 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: $S_y^2 = (\frac{1}{S_x^2})\, S_x^2 = 1$; $\overline{y} = \frac{\overline{x} - \overline{x}}{S_x} = 0$

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 عددP 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؟