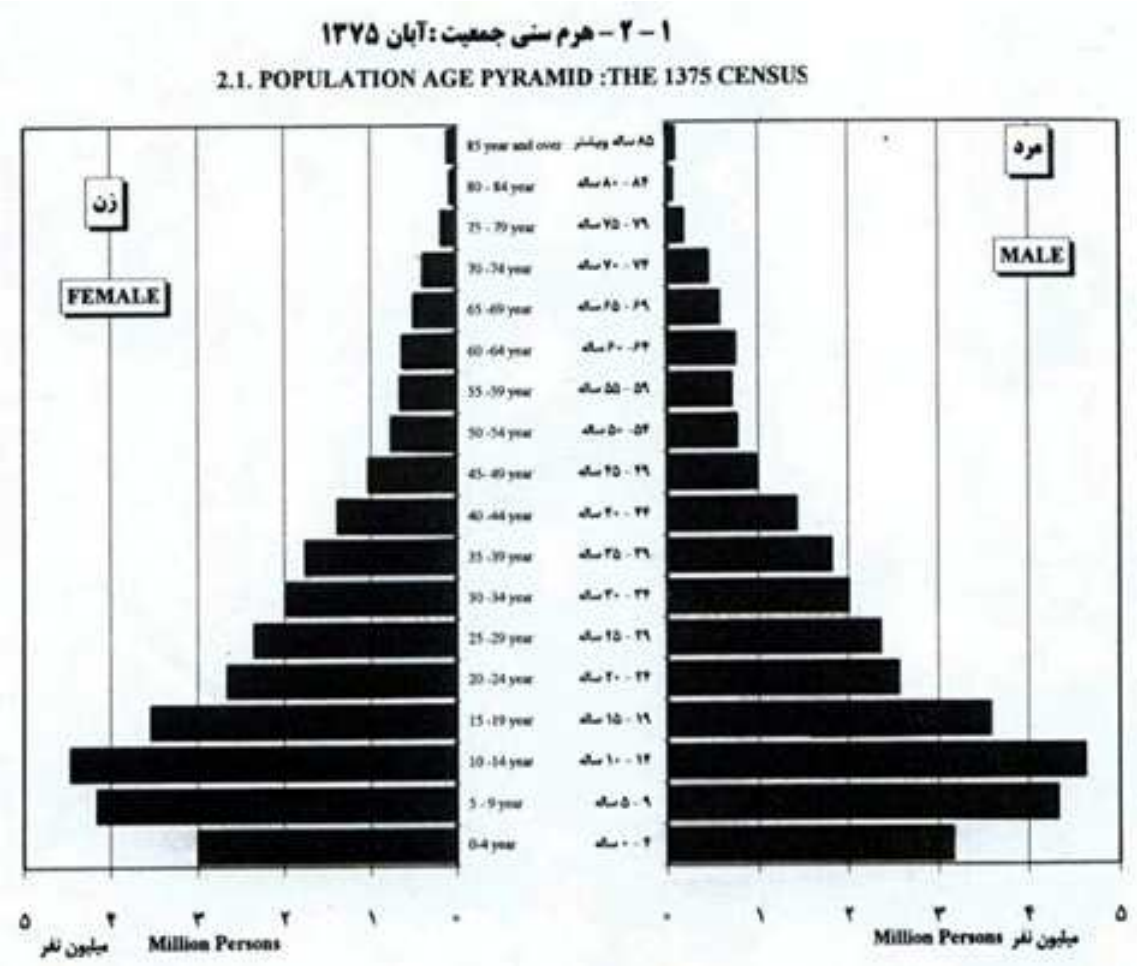
آمار و احتمالات

شاهپور نصرتی



اردیبهشت ۸۶

فصل ۱

آمار

۱.۱ مقدمه

۱.۱.۱ تاریخچه

بشر از زمانی که حکومت کردن را آموخت، لازم دید تا اطلاعاتی را از جمعیت و منابع موجود تحت سلطه خود را برای عملکرد بهتر در دست داشته باشد. این گونه آگاهی از میزان کارگزار و مقدار منابع در تواریخ عهد عتیق، مصریان قدیم، بابلیان و رومیان در هزارهٔ قبل از میلاد دیده می شود. با مرور زمان اهمیت جمع آوری فهرستهای اطلاعاتی بیشتر شده و تهیه فهرست هائی از نیروی انسانی و املاک متعلقه مهم تلقی شده تا جائی که در قرون وسطی این امر را زیر نظر کلیسا انجام می دادند. در سال ۱۳۰۰ در انگلستان بدستور ویلیام فاتح [۱] اولین کتاب اطلاعات از جمعیت و منابع اقتصادی نگاشته شد. بین سالهای ۱۶۶۱ الی ۱۶۶۹ میلادی کاپیتان جان گرانت آماری از زاد و ولد، مرگ و میر تهیه کرد و با ارائهٔ آن در کتاب خود، نشان داد که در انگلیس عدهٔ زنان بیشتر از مردان است.

برای اولین بار کلمه آمار [۲] از کلمه یونانی سیاستمدار [۳] توسط گاتفرید آچنوال [۴] (۱۷۱۹–۱۷۷۲) بکار رفت و در سال ۱۷۸۷ کلمه آمار در انگلستان توسط زیمرمان [۵] بکار گرفته شد و سپس سینکلیر [۶] در کتابش به نام **شمارش آماری اسکاتلند** [۷] آنرا عمومیت داد. در اواسط قرن هفدهم مجلهٔ آماری در فرانسه ایجاد شد و اولین سرشماری در این کشور در ۱۷۹۱م. انجام گرفت، سپس طی بیست سال بعدی سرشماری های دیگری در کشورهای مختلف انجام گرفت و لزوم استفاده از آمار و تحلیلهای عددی را نشان داد. تا آنجا که فرانسیس بیکن (۱۵۶۱–۱۶۲۶) اظهار داشت که تنها با مطالعهٔ داده‌های پدیده‌های تجربی و طبیعی و تحلیل آنها می توان دانش بشری را سامان داد.

استفاده از آمار بعنوان شاخه ای از ریاضیات کاربردی، مختص یک شاخهٔ بخصوص از علوم نبوده و امروزه می توان کاربرد آنرا در بسیاری از علوم مانند اقتصاد، روانشناسی، جامعه شناسی، علوم زیستی، فیزیک، شیمی، پزشکی، کشاورزی و ... یافت.

---

[۱] *William the Conqueror*
[۲] *Statistics*
[۳] *Satista*
[۴] *Gottfried Achenwal*
[۵] *E.A.W.Zimmermann*
[۶] *John Sinclair*
[۷] *Statistical Account of Scotland*

با استفاده از داده های آماری و بهره گیری از علم **احتمالات** می توان نتایج تجربی را تحلیل نمود و قاعده ای کلی را از آنها استنتاج کرد. شناخت آماری مسائل اکنون برای مدیران و صاحبان صنایع و موسّسات از ضروریات بشمار رفته و به آنها کمک می کند که مشکلات را بشناسند و راه مقابله را مشکل را بیابند. آمار بمعنای یک علم تصمیم گیرنده[۸] زیربنای توسعه و پیشرفت بوده و در برنامه ریزی ها و تصمیمات کلان آیندهٔ یک کشور نقش عمده‌ای ایفا می کند.

از پیشگامان آمار لامبرت کتله[۹] (۱۸۷۴–۱۷۹۴) منجم و آمارگر بلژیکی که اولین بار دربارهٔ میانگین و انحراف و خطاهای آماری در زمینه های اجتماعی به بحث پرداخت. کتله قانون اعداد بزرگ را که ژاکوب برنولی (۱۷۰۵–۱۶۵۴) و سیمون پواسون (۱۸۴۰–۱۷۸۱) به آن دست یافته بودند مورد استفاده قرار داد. با مرور زمان آماردانان و ریاضیدانان به وسعت این علم افزودند و راه حلهای آماری و احتمال را در آن وسعت بخشیدند. اکنون محاسبات داده ها توسط رایانه با سرعت و دقت بیشتری انجام می شود و کاوش ماشینی داده ها[۱۰] جای محاسبات دستی را گرفته است. نرم افزارهای آماری اکنون در تحلیل داده ها نقش بسیار مثبتی را عهده دار بوده و تصمیمات نهائی را بهینه می کنند.

۲.۱.۱ تعریف مساله

علم آمار عبارتست از جمع آوری اطلاعات و پردازش آنها و سپس نتیجه گیری و استنباط یک یا چند صفت است. آمار در واقع ابزاری علمی است که اگر آنرا روی مجموعه ای از اطلاعات بکار ببریم آگاهی ما را نسبت به آنها افزایش می دهد. تحقق این امر، در ابتدا منوط به ایجاد ساختاری ریاضی بر روی مجموعهٔ داده ها با **شاخصی معین** می باشد، بنابراین مساله را با زبان ریاضی بیان می کنیم تا بتوانیم از طرفندهای ریاضی برای حل مساله بهره بگیریم. در این حالت مساله را **مدلسازی ریاضی** نموده و از این طریق استنباط روی شاخص موردنظر را انجام می دهیم. مدلسازی **بهینه**، مفاهیم ریاضی بکار برده شده را ساده تر و ابتدائی تر و نتیجهٔ کار را به شخصیت اعضاء مجموعهٔ موردنظر نزدیکتر می سازد. پس از اتخاذ مدلسازی، **روشهای آماری** را برای تجزیه و تحلیل داده ها بکار می بریم.

مجموعه ای از انسانها یا اشیاء که مورد مطالعهٔ آماری قرار می گیرند و حداقل در یک صفت مشترک می باشند را **جامعهٔ آماری** نامیم. سنجش این صفت مشترک اعم از اینکه صفت کمی یا کیفی باشد، بر روی اعضاء جامعه اعمال می شود و بستگی به تک تک افراد آن دارد. پس جامعهٔ آماری مجموعه‌ای از افراد یا اشیاء است که دربارهٔ اعضای آن می خواهیم موضوع و یا موضوعاتی را مطالعه کنیم. جامعه ممکن است محدود (مثلاً افراد یک کلاس) یا نامحدود (مثلا جمعیت کشور) باشد. اگر تک تک افراد یک جامعه را مورد مطالعه قرار دهیم، در اینصورت **سرشماری** کرده‌ایم و از آنجائیکه سرشماری بدلایل مختلف همیشه امکانپذیر نیست (بخصوص در جوامع نامحدود) یا قادر به سرشماری کل جامعه نیستیم و یا در مواردی ضرورتی به سرشماری نیست. در اینصورت بجای استفاده از کل اعضاء جامعه، تعداد محدودی از اعضاء را در نظر گرفته و شاخص مورد نظر را روی این تعداد **دلخواه** می سنجیم. این تعداد دلخواه از اعضاء جامعه را که بصورت **تصادفی** انتخاب می شوند، **نمونهٔ تصادفی** گوئیم و نمونه زیرمجموعه‌ای از جامعهٔ آماری محسوب می شود.

---

[۸] برخی نیز آمار را شاخه ای از نظریهٔ تصمیم ها *Decision Theory* بشمار می آورند.

[۹] *Lambert A.J.Quetelet*

[۱۰] *Data Mining*

۳.۱.۱ نمونه

کار آمار **نمونه گیری** از تعدادی از افراد جامعه و پس از تصمیم گیری، عمومیت دادن آن به کل جامعه است. اما نمونه گیری از افراد یک جامعه بدون خطا نخواهد بود، جمع آوری داده های مفید، صحیح و سالم و تجزیه و تحلیل دقیق آنها مسلماً نتایج مفید و درخور اعتباری را در بر خواهد داشت. بنابراین می بایست خطاهای مورد نظر در مسیر دستیابی به عمومیت شاخص را در نظر گرفت و تحقق این موضوع به علم **احتمالات** وابسته است. برای برآورد دقیق یک شاخص یا پارامتر روی جامعه بایستی از **نمونهٔ تصادفی** استفاده کنیم. طبعاً شاخص مورد نظر روی اعضاء جامعه (نمونه) متفاوت بوده و صحت برآورد شاخص روی نمونه به تصادفی بودن اعضاء نمونه کاملاً **وابسته** است. اطلاعات حاصله اطلاعات خام محسوب شده و با روشهای آماری نمونه را بررسی می کنیم و نتیجه گیری می نمائیم. اکنون جامعه را کنار گذاشته و به شناخت دقیق یک نمونه می پردازیم. تعداد اعضاء یک جامعه را با N و تعداد اعضاء نمونه را با n نشان خواهیم داد. n را **اندازهٔ نمونه** نیز گویند.

۴.۱.۱ متغیر یک نمونه

گفتیم که اعضاء جامعه دارای صفتی مشترک هستند، شاخصی را که می بایست روی یک نمونه اندازه بگیریم را **متغیر** نامیم. این شاخص می تواند سن یا قد افراد یا دوام محصول یک کارخانه یا ... باشد. متغیر تصادفی به دو دستهٔ **متغیرکمی** و **متغیرکیفی** تقسیم بندی می شود. متغیرهای کمی برخلاف متغیرهای کیفی، متغیرهایی قابل اندازه گیری اند. در اینصورت اگر بخواهیم متغیر کیفی را از قبیل شادی و اندوه و میزان هوش و ... بسنجیم می بایست آن را تبدیل به متغیر کمیّ کنیم. متغیر کمیّ ممکن است پیوسته باشد و **متغیر پیوسته** یک متغیر کمی است که اندازه های ممکن یک فاصله را اختیار کند، بعبارتی دیگر اگر دو مقدار را بتواند اختیار کند، هر مقدار بین آنها را نیز می تواند اختیار نماید. به متغیر کمیّ که پیوسته نباشد **متغیر گسسته** گوئیم. متغیرهای کیفی که در آنها نوعی ترتیب طبیعی وجود دارد **متغیرهای کیفی ترتیبی** می گوئیم و متغیر کیفی که ترتیبی نباشد **متغیر کیفی اسمی** می گوئیم.

به اعضاء یک نمونه **دادهٔ آماری** یا بطور خلاصه **داده** می گوئیم و آن را با x_i نشان می دهیم. در واقع **داده های آماری** بجای متغیر یک نمونه می نشینند. فرض کنید می خواهیم میزان سن دانشجویان یک دانشگاه را مورد سنجش قرار دهیم. در اینجا دانشجویان این دانشگاه جامعهٔ آماری خواهند بود و شاخص مورد سنجش سن دانشجویان است که متغیر نمونه محسوب خواهد شد و داده های حاصل از نمونه برداری تصادفی نیز بجای x_i قرار می گیرند.

۵.۱.۱ جدول فراوانی داده ها

بعد از نمونه گیری و بدست آوردن داده ها، لازم است تا **میانگین** و **انحراف معیار** داده ها — که بعداً راجع به آنها صحبت خواهیم کرد — را بدست آوریم. داده ها را پس از جمع آوری و **مرتب نمودن**، در قالب یک جدول می ریزیم که به آن جدول آماری یا **جدول فراوانی داده‌ها** گوئیم. در ستون اول این جدول x_i (داده‌های نمونه) را جای داده و در ستون دوم **فراوانی مطلق** آنها را یادداشت می کنیم. **فراوانی مطلق** یک دادهٔ x_i برابر با تعداد دفعاتی است که آن داده تکرار شده است و آنرا با f_i نشان می دهیم. اگر f_i فراوانی مطلق دادهٔ iام و تعداد داده های نمونه n باشد، مقدار کسر $\overline{f_i} = \frac{f_i}{n}$ را **فراوانی نسبی** دستهٔ i—ام می

گوئیم و آنرا درستون $\overline{f_i}$ جای می دهیم. **فراوانی تجمعی** F_i هر داده، برابر است با مجموع فراوانی آن داده و داده‌های قبل از آن. فراوانی های نسبی و تجمعی را می توان در انتهای جدول نیز جای داد.

x_i	f_i	F_i	$\overline{f_i}$	$f_i x_i$	$x_i - \bar{x}$	$(xi - \bar{x})^۲$	$f_i(x_i - \bar{x})^۲$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

میانگین داده‌های یک نمونه، همان متوسط داده‌هاست که با $\overline{x}$ نشان می دهیم و برای n دادهٔ $x_۱$ و $x_۲$ و $\cdots$ و x_n برابرست با

$$\overline{x} = \frac{\sum_{i=۱}^{n} x_i}{n}$$

فرمول میانگین برای n دادهٔ $x_۱$ و $x_۲$ و $\cdots$ و x_n با فراوانی های $f_۱$ و $f_۲$ و $\cdots$ و f_n بصورت $\overline{x} = \frac{\sum f_i x_i}{n}$ خلاصه می شود.
واریانس یک نمونه از n دادهٔ $x_۱$ و $x_۲$ و $\cdots$ و x_n با فراوانی های $f_۱$ و $f_۲$ و $\cdots$ و f_n عبارتست از:

$$S_x^۲ = \frac{\sum f_i(x_i - \bar{x})^۲}{n - ۱}$$

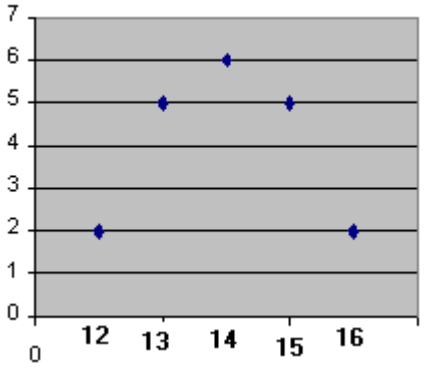
که آنرا **واریانس نمونه ای** نیز گوئیم. برای بدست آوردن این مقدار ستون $x_i - \bar{x}$ و سپس ستون های $(x_i - \bar{x})^۲$ و $f_i(x_i - \bar{x})^۲$ را به جدول اضافه می کنیم تا کار محاسبه آسانتر شود. بعد از بدست آوردن واریانس، مقدار **انحراف معیار** را که جذر آن خواهد بود، بدست می آوریم پس $S_x = \sqrt{S_x^۲}$. به مثال زیر دقت کنید:

مثال ۱.۱.۱ در نمونه گیری از دانشجویان یک دانشگاه، نمرات ۲۰ نفر از آنها در درس آمار بصورت زیر بوده است:
۱۵ ، ۱۴ ، ۱۳ ، ۱۶ ، ۱۵ ، ۱۳ ، ۱۴ ، ۱۵ ، ۱۴ ، ۱۴ ، ۱۴ ، ۱۳ ، ۱۲ ، ۱۳ ، ۱۵ ، ۱۳ ، ۱۴ ، ۱۳ ، ۱۴ ، ۱۵ ، ۱۶ ، ۱۲
مطابق آنچه گفته شد جدول فراوانی داده ها چنین می نویسیم:

x_i	f_i	F_i	$\overline{f_i}$	$f_i x_i$	$x_i - \bar{x}$	$(xi - \bar{x})^۲$	$f_i(x_i - \bar{x})^۲$
۱۲	۲	۲	۰/۱۰	۲۴	-۲	۴	۸
۱۳	۵	۷	۰/۲۵	۶۵	-۱	۱	۵
۱۴	۶	۱۳	۰/۳۰	۸۴	۰	۰	۰
۱۵	۵	۱۸	۰/۲۵	۷۵	-۱	۱	۵
۱۶	۲	۲۰	۰/۱۰	۳۲	-۲	۴	۸
جمع	۲۰		۱	۲۸۰			۲۶

از ستون پنجم می نویسیم $\bar{x} = \frac{\sum f_i x_i}{n} = \frac{۲۸۰}{۲۰} = ۱۴$ و از ستون آخر می توان واریانس را بدست آورد:

$$S^۲ = \frac{\sum f_i(x_i - \bar{x})^۲}{n - ۱} = \frac{۲۶}{۱۹} = ۱/۳۷ \Longrightarrow S_x = ۱/۱۷$$



دقت کنید که سطر آخر جمع فراوانی های مطلق می بایست برابر تعداد داده ها و مجموع فراوانی های نسبی می بایست ۱ شود. همچنین آخرین مقدار فراوانی تجمعی n خواهد بود. نمودار نقطه‌ای این داده ها بصورت مقابل است.

۲.۱ تحلیل داده‌ها

پس از مرتب سازی داده ها و جدول فراوانی، برای تصمیم گیری نهائی لازم است پارامترها و معیارهای مختلفی را روی آنها پیاده نمود تا تصمیم گیری بر اساس این معیارها انجام شود. تحلیل اطلاعات با دو نوع شاخص انجام می شود: **شاخص های عددی** و **شاخص های هندسی**. شاخص عددی مقداریست که از داده ها حاصل شده و خصوصیات کمیّ جامعه براساس آنها تعیین می گردد. انواع شاخص های عددی عبارتند از شاخص های مرکزی، شاخص های پراکندگی، چولگی و کشیدگی. شاخص های هندسی نیز همان نمودارها هستند.

۱.۲.۱ شاخص های مرکزی

شاخصی که تمرکز صفت مورد نظر را در بین داده ها مشخص می کند**شاخص مرکزی** نامیده می شود. با توجه به داده‌های یک نمونهٔ کاملاً تصادفی، دیده می شود که قرارگیری تعدادی از داده‌ها تمرکز و ویژگی خاصی نسبت به بقیه ایجاد نموده است. بنابراین شاخص مرکزی، محل تمرکز اکثر داده‌هاست. مهمترین شاخص های مرکزی عبارتند از میانگین، میانه و مد. **مد یا نما** (M) داده ای است که دارای بیشترین فراوانی است. در مثال ۱.۱.۱ دادهٔ ۱۴ دارای بیشترین فراوانی بوده و بنابراین $M = ۱۴$. در مواردی که داده‌ها دارای دو فراوانی بیشین هستند — وضعیت داده های دو مدی — اگر داده ها در کنار هم واقع شوند میانگین آنها را بعنوان مد در نظر می گیریم و در حالتیکه داده ها در کنار هم نباشند تمام آنها مد محسوب می شوند. اگر داده ها همه فراوانی برابر داشته باشند مد نداریم. منظور از **میانه** (m) عبارتست از دادهٔ وسط. بنابراین برای بدست آوردن میانه، پس از مرتب کردن داده‌ها، مقداری را که تعداد داده‌های بعد از آن با تعداد داده های قبل از آن برابر است را میانه می نامیم. اگر تعداد داده‌ها فرد باشد دادهٔ وسطی میانه خواهد بود $m = x_{\frac{n+۱}{۲}}$ و در صورت زوج بودن تعداد داده‌ها، میانه عبارتست از میانگین دو دادهٔ وسط $m = \frac{۱}{۲}(x_{\frac{n}{۲}} + x_{\frac{n}{۲}+۱})$. در مثال ۱.۱.۱ اگر داده ها را مرتب کنیم:

۱۲ ، ۱۲ ، ۱۳ ، ۱۳ ، ۱۳ ، ۱۳ ، ۱۳ ، ۱۴ ، ۱۴ ، ۱۴ ، ۱۴ ، ۱۴ ، ۱۴ ، ۱۵ ، ۱۵ ، ۱۵ ، ۱۵ ، ۱۶ ، ۱۶

چون تعداد داده ها زوج است بنابراین $m = ۱۴$.

مثال ۱.۲.۱ شاخص های مرکزی را برای داده های زیر بیابید ۵ ، ۳ ، ۴ ، ۲ ، ۳ ، ۴ ، ۳ ، ۲ ، ۶ ، ۵

حل. داده ها را مرتب می کنیم ۶ ، ۵ ، ۵ ، ۴ ، ۴ ، ۳ ، ۳ ، ۳ ، ۲ ، ۲

$$\bar{x} = \frac{\sum_{i=۱}^{n} x_i}{n} = \frac{\sum_{i=۱}^{۱۰} x_i}{۱۰} = \frac{۲+۲+۳+۳+۳+۴+۴+۵+۵+۶}{۱۰} = ۳/۷$$

دادهٔ ۳ دارای بیشترین فراوانی است بنابراین $M = ۳$ و چون تعداد داده ها زوج هستند پس $m = \frac{۳+۴}{۲} = ۳/۵$.

تمرین ۱.۲.۱ **کلاسی.** در یک نمونه ۲۰ نفره دانش آموزان یک مدرسه نمونه سن آنها چنین بدست آمده است:

۸ و ۷ و ۱۰ و ۸ و ۹ و ۶ و ۷ و ۷ و ۶ و ۷ و ۸ و ۱۰ و ۹ و ۸ و ۹ و ۷ و ۶ و ۸ و ۷

پس از رسم جدول فراوانی داده ها، میانگین و واریانس و شاخص های مرکزی را بیابید.

۲.۲.۱ شاخص های پراکندگی

با اینکه میانگین شاخص مهمی در تصمیم گیری است، ولی بالا بودن میانگین در یک نمونه به معنی بالا بودن کل داده ها نیست مثلاً بالا بودن میانگین نمرات یک کلاس ممکن است تنها بخاطر چند نمرهٔ ۲۰ باشد و سایر نمرات نزدیک ۱۰ باشند و بنابراین داده ها می بایست مثلاً با میانه نیز مقایسه شوند. معیاری که می تواند تفاوت داده ها و میزان این تفاوت و بخصوص دوری یا نزدیکی آنها را از میانگین برای ما بیان کند**شاخص پراکندگی** نامیم. مهمترین شاخص های پراکندگی عبارتند از واریانس، انحراف معیار و ضریب تغییرات. بعنوان بهترین شاخص، می توان به **واریانس** اشاره کرد که به معنی تفاوت و تغییر است و عبارتست از میانگین مجذور انحرافات از میانگین. اگر داده ها نزدیک $\bar{x}$ باشند، واریانس کوچک می شود، بنابراین واریانس معیار خوبی برای سنجش پراکندگی داده ها از $\bar{x}$ است. اما چون واریانس واحد داده ها را مربع می کند، از جذر آن که **انحراف معیار** یا **انحراف استاندارد** نامیده می شود استفاده می کنیم و با S_x نشان می دهیم. واحد انحراف معیار همان واحد x_i است و پراکندگی نسبی داده ها را مشخص می کند. در واقع انحراف معیار، انحراف از میانگین $\bar{x}$ است و در حالت خاص اگر داده‌ها همه برابر باشند واریانس آنها صفر است که دارای پراکندگی نیستند.

مثال ۲.۲.۱ از دانشجویان دو کلاس آمار نمره های زیر بدست آمده است:

A : ۱۴ و ۱۸ و ۱۴ و ۱۵ و ۱۳ و ۱۷ و ۱۴

B : ۱۱ و ۱۵ و ۲۰ و ۱۸ و ۱۴ و ۱۷ و ۱۰ و ۱۹ و ۱۱

با اینکه میانگین نمره ها در این هر دو برابر ۱۵ است ولی پراکندگی کلاس B بیشتر است و $S_A = ۱/۸۲$ و $S_B = ۳/۷۴$.

معیار دیگری برای میزان پراکندگی **ضریب تغییرات** V_x است که عبارتست از خارج قسمت انحراف معیار بر میانگین $V_x = \frac{S_x}{\bar{x}}$. بعبارتی دیگر ضریب تغییرات عبارتست از میزان پراکندگی بازای یک واحد از میانگین. این ضریب مستقل از واحد است.

مثال ۳.۲.۱ کارخانه ای دو نوع لامپ ۱۰۰ واتی تولید می کند. نوع A دارای میانگین عمر ۹۰۰۰ ساعت و انحراف استاندارد ۱۸۰۰ ساعت و نوع B دارای میانگین عمر ۱۰۰۰۰ ساعت و انحراف استاندارد ۱۲۰۰ ساعت است. کدام نوع لامپ بهتر است ؟.

حل. ضریب تغییرات را برای هر دو نوع حساب می کنیم:

$$V_A = \frac{۱۸۰۰}{۹۰۰۰} = ۰/۲۰ = \%۲۰ \qquad V_B = \frac{۱۲۰۰}{۱۰۰۰۰} = ۰/۱۲ = \%۱۲$$

نوع B بهتر است زیرا هم میانگین بیشتری دارد و هم دارای ضریب تغییر کمتری است.

تمرین ۲.۲.۱ کلاسی. ضریب تغییرات دو کلاس مذکور در مثال ۲.۲.۱ را یافته، این دو کلاس را با هم مقایسه کنید.

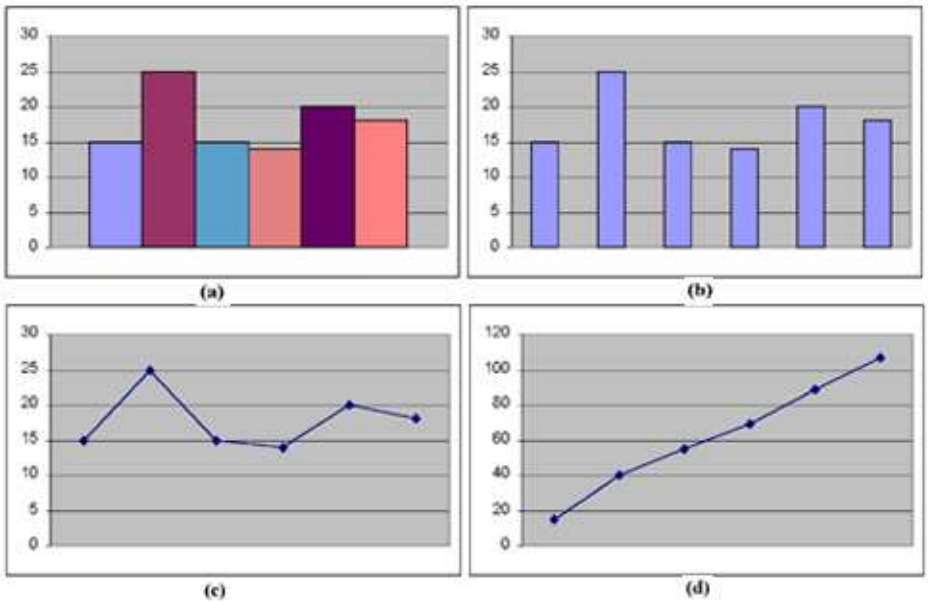
۳.۲.۱ نمودارها

نمودار نمایش هندسی داده ها و روشی برای تحلیل اطلاعاتست. نمودارها شاخص های هندسی و کمکی برای ایجاد تصویر ذهنی از داده ها بوده و برای تصویرسازی از یک نمونه، نمودار تا حدی میزان جایگیری و پراکندگی داده‌ها را مشخص می کند و بدین ترتیب شاخصی برای تصمیم گیری محسوب می گردد. انواع مختلفی از نمودارها وجود دارد که مهمترین آنها را بیان می

کنیم. البته تنوع نمودار برای ارائه شکل قرارگیری داده‌ها زیاد است (تمرین۲۱) که تنها چند نوع آن دارای کاربرد بیشتری است. از انواع مهم نمودار می توان به موارد ذیل اشاره نمود:

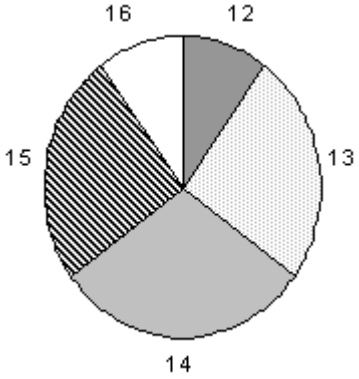
الف) **نمودار بافتی یا هیستوگرام یا نمودار ستونی** که بیشتر برای نشان دادن داده های پیوسته و داده‌های دسته بندی شده بکار می رود و از انواع دیگر ساده تر است، در این نوع نمودار، دسته ها عرض مستطیل ها بوده و ارتفاع مستطیل ها را متناسب با فراوانی دسته‌ها انتخاب می کنیم (تصویر a).

ب) **نمودار میله‌ای** مانند نمودار بافتی ولی با ستونهای مجزاست و عرض میله ها نیز برابر است (تصویر b).



ج) **نمودار چندبرفراوانی یا نمودار خطی (چندضلعی)** که درآن نقاط داده ها را بهم متصل می کنیم و این نوع نمودار علمی تر بوده و می توان آنرا با نمودار نرمال که بعداً خواهیم آورد مقایسه نمود، نقاط نمودار مانند نقاط صفحهٔ مختصات ترسیم می شوند که در آنها x_i طول نقاط و f_i ها عرض نقاط را تشکیل خواهند داد.در این نوع نمودار، اگر وسط بالای مستطیل ها را بصورت نقطه‌ای در نظر بگیریم و آنها را بهم وصل کنیم، دراینصورت آنرا نمودار **چندبر فراوانی** نامیم (تصویر c).

د) **نمودار فراوانی تجمعی** که بجای فراوانی مطلق درستون عرضها از فراوانی تجمعی استفاده می کنیم، این نمودار صعودی بوده در این حالت اگر متغیر پیوسته باشد توزیعی تراکمی یا اوجیو بوجود می آید و این نمودار در مواردی مختلفی مانند پزشکی کاربرد دارد (تصویر d).



ه) **چندبرفراوانی نسبی** در این نوع بجای فراوانی مطلق، نمودار فراوانی نسبی را رسم کنیم، چون اطلاعات با کل نمونه مقایسه می شود این نمودار رسم شده اطلاعات جالبی را بیان خواهد داشت.

و) **نمودار دایره ای (کلوچه‌ای)** که برای تجسم داده ها و مقایسه آنهاست. در این نوع تقسیم بندی که در خلاف جهت عقربه های ساعت انجام می شود، ترسیم فراوانی نسبی داده ها بر روی یک دایره خواهد بود. مقدار زاویهٔ در نظر گرفته شده برابر $۲\pi\bar{f} = ۳۶۰ \times \bar{f}$ (درجه) خواهد بود. نمودار دایره ای برای داده‌های مثال ۱.۱.۱ نشان داده شده است.

۴.۲.۱ دسته‌بندی داده‌ها

اگر تنوع داده‌ها زیاد بوده و فراوانی هر کدام از آنها کم باشد، برای جلوگیری از حجم زیاد محاسبات، داده ها را **دسته بندی** می کنیم. در این روش ابتدا **دامنهٔ تغییرات** R را بدست آورده و سپس تعداد دسته‌ها و طول هر دسته را معین می کنیم و بعد جدول را مطابق قبل رسم می کنیم. **دامنهٔ تغییرات** R یک متغیر عبارتست از طول بازه‌ای که متغیر در آن تغییر می کند و آن عبارتست از تفاضل بیشترین داده و کمترین داده، می نویسیم $R = x_{max} - x_{min} + ۱$. تعیین تعداد دسته‌ها T قانون مشخصی ندارد و نظرات متفاوتی درباره آن ارائه شده، ولی غالباً ما آنرا بین ۵ تا ۲۰ دسته انتخاب می کنیم. یک فرمول از استارجیس[۱۱] بصورت $T = ۱ + ۳/۳۲۲ log^{n}_{۱۰}$ پیشنهاد شده است. بهرطریق پس از انتخاب T، طول دسته ها از رابطهٔ زیر بدست می آید:

$$\text{طول دسته } c = \frac{R}{T}$$

مجموع طول دسته‌ها نباید از دامنه‌ی تغییرات کمتر باشد. بهترین حالت انتخاب هر دسته بصورت نیم بازهٔ $[a_i, b_i)$ است که a_i را کران پائین دسته و b_i را کران بالای دسته می نامیم. **نمایندهٔ دسته** نیز عبارتست از $x_i = \frac{a_i + b_i}{۲}$ و مسلماً تفاضل دو کران پائین متوالی یا دو کران بالای متوالی طول دسته c خواهد بود. نکته ای که در انتخاب دسته ها باید دقت شود این است که کران بالا و پائین هر دسته می بایست بگونه ای انتخاب شود که هر داده تنها در یک دسته شمارش شود. بفرض اگر داده ای در دو دسته قرار گرفت می بایست از تمامی کرانها مقداری (مثلاً ۰/۵) کم شود تا داده تنها در یک دسته قرار گیرد، یا اینکه دسته بعدی بفاصله کمی از کران بالای دسته قبلی شروع شود $b_i \lneqq a_{i+۱}$ بدین ترتیب بین طبقات تداخل داده ایجاد نخواهد شد. مثال زیر را ببینید.

مثال ۴.۲.۱ در نمونه گیری از سن دانشجویان یک دانشگاه، سن ۵۰ نفر از آنها بصورت زیر بوده است:

۲۱	۱۸	۲۲	۱۹	۲۱	۳۰	۲۰	۱۸	۳۵	۳۰
۲۰	۲۵	۱۸	۱۷	۲۵	۲۷	۲۲	۲۲	۲۰	۲۱
۲۳	۲۰	۲۱	۱۸	۱۹	۲۱	۲۰	۱۹	۳۳	۲۹
۲۰	۱۸	۳۵	۲۲	۲۰	۳۰	۲۵	۲۴	۲۲	۱۷
۱۹	۱۸	۲۰	۱۹	۲۵	۲۶	۲۰	۲۱	۱۸	۲۰

مطابق این نمونه ۵۰ تائی داریم $R = x_{max} - x_{min} + ۱ = ۳۵ - ۱۷ + ۱ = ۱۹$ اگر تعداد دسته را ۵ انتخاب کنیم پس

$$\text{طول دسته} = \frac{۱۹}{۵} \sim ۴$$

با انتخاب دسته‌ها بصورت ۲۱ – ۱۷ و ۲۵ – ۲۱ و ... انتخاب کنیم، دادهٔ ۲۱ متعلق به دو دسته خواهد شد برای رفع این دودستگی، دسته ها را با اختلاف ۰/۵ انتخاب می کنیم و جدول فراوانی داده‌ها بصورت زیر درمی آید:

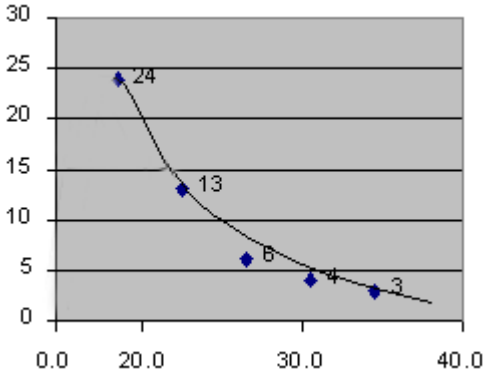
دسته‌ها	دسته‌نماینده x_i	f_i	F_i	$\overline{f_i}$	$f_i x_i$	$x_i - \bar{x}$	$(xi - \bar{x})^۲$	$f_i(x_i - \bar{x})^۲$
۱۶/۵ – ۲۰/۵	۱۸/۵	۲۴	۲۴	۰/۴۸	۴۴۴	−۳/۹۲	۱۵/۳۷	۳۶۸/۷۹
۲۰/۵ – ۲۴/۵	۲۲/۵	۱۳	۳۷	۰/۲۶	۲۹۳	−۰/۰۸	۰/۰۱	۰/۰۸۳۲
۲۴/۵ – ۲۸/۵	۲۶/۵	۶	۴۳	۰/۱۲	۱۵۹	۴/۰۸	۱۶/۶۵	۹۹/۸۷۸
۲۸/۵ – ۳۲/۵	۳۰/۵	۴	۴۷	۰/۰۸	۱۲۲	۸/۰۸	۶۵/۲۹	۲۶۱/۱۵
۳۲/۵ – ۳۶/۵	۳۴/۵	۳	۵۰	۰/۰۶	۱۰۴	۱۲/۰۸	۱۴۵/۹۳	۴۳۷/۷۸
جمع		۵۰		۱	۱۱۲۱			۱۱۶۷/۷

[۱۱] *Sturges*، این فرمول از $n = ۲^{T-۱}$ نتیجه شده است.

از جدول بالا می توان مقادیر زیر را محاسبه نمود:

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{۱۱۲۱}{۵۰} = ۲۲/۴۲ \quad , \quad S_x^۲ = \frac{\sum f_i (x_i - \bar{x})^۲}{n-۱} = \frac{۱۱۶۷/۷}{۴۹} = ۲۳/۸۳ \Longrightarrow S_x = ۴/۸۸$$

$$V_x = \frac{S_x}{\bar{x}} = \frac{۴/۸۸}{۲۲/۴۲} = ۰/۲۱۸ = \%۲۲$$



نمودار داده‌ها بصورت مقابل است.

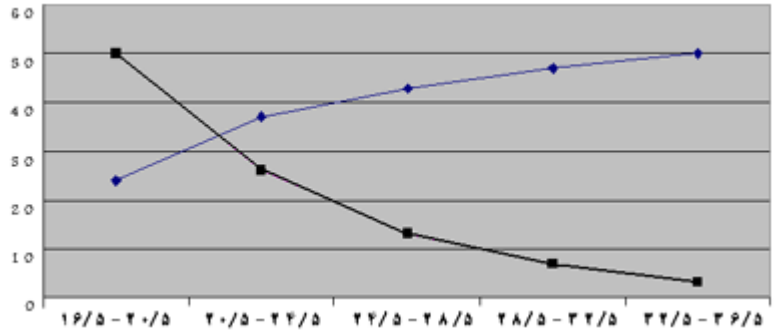
میانه در داده های دسته بندی شده:

برای بدست آوردن میانه با استفاده از جدول فراوانی داده ها اگر n تعداد داده ها و a_i حد پائین دسته‌ای که فراوانی تجمعی آن بلافاصله از $\frac{n}{۲}$ بیشتر یا مساوی باشد و f_i فراوانی مطلق دسته i—ام و $F_{i-۱}$ فراوانی تجمعی دسته ماقبل دسته i—ام باشد سپس میانه عبارتست از [۱۲]

$$m = a_i + \frac{\frac{n}{۲} - F_{i-۱}}{f_i} c$$

که c فاصله طبقات است ($c = a_i - a_{i-۱}$). در مثال قبل چون $\frac{n}{۲} = ۲۵$ پس فراوانی تجمعی $F_۲ = ۳۷$ از آن بیشتر یا مساوی است و $i = ۲$ و $۴ = ۱۶/۵ - ۲۰/۵$ طبق فرمول بالا

$$m = ۲۰/۵ + \frac{۲۵ - ۲۴}{۱۳} ۴ = ۲۰/۸ \in [۲۰/۵, ۲۴/۵]$$



اگر داده ها را مرتب کنید و از روش اصلی دادهٔ وسط را بیابید میانه واقعی برابر $m = ۲۱$ خواهد بود. برای بدست آوردن میانه با استفاده از نمودار کافیست فراوانی تجمعی صعودی و نزولی را در یک صفحه مختصات رسم کنیم، تقاطع آنها میانه خواهد بود.

مد در داده های دسته بندی شده: برای بدست آوردن مددر داده های دسته بندی از فرمول زیر استفاده می کنیم:

$$M = a_i + \frac{d_۱}{d_۱ + d_۲} c$$

با یافتن دسته ای که دارای بیشترین فراوانی است (دسته یا ردهٔ نمائی)، اگر a_i کران پائین دسته و c فاصلهٔ طبقات باشد، $d_۱$ تفاوت فراوانی دسته با دستهٔ قبلی و $d_۲$ تفاوت فراوانی دسته با دستهٔ بعدی است. برای مثال بالا

$$M = a_i + \frac{d_۱}{d_۱ + d_۲} c = ۱۶/۵ + \frac{۲۴}{۲۴ + ۱۱} ۴ = ۱۹/۲۴$$

و اگر داده ها را مرتب کنیم $M = ۲۰$ خواهد بود (البته این مقدار را تنها برای مقایسه و درستی مطلب بدست آورده ایم).

[۱۲] کندال *kendall* آماردان انگلیسی می گوید در حالت n زوج بجای $\frac{n}{۲}$ از $\frac{n+۱}{۲}$ استفاده کنید.

۵.۲.۱ گشتاورها

برای n دادهٔ x_1 و x_2 و $\cdots$ و x_n با فراوانی های f_1 و f_2 و $\cdots$ و f_n، مقدار m_r گشتاور مرکزی مرتبهٔ r و m'_r گشتاور مرتبهٔ r و چنین تعریف می شوند:

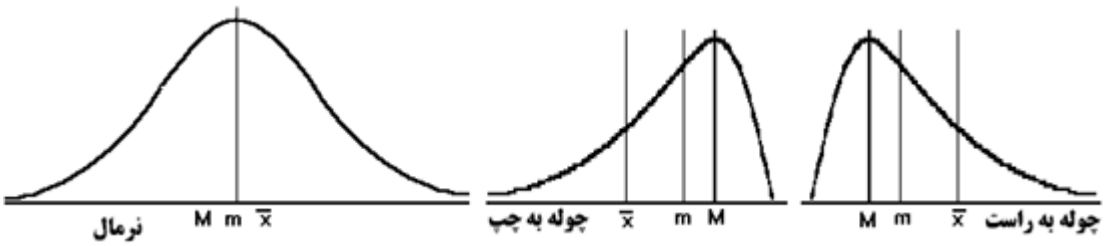
$$m_r = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r \quad , \quad m'_r = \frac{1}{n}\sum_{i=1}^{n} f_i x_i^r$$

می توان دید که $m'_0 = m_0 = 1$ و $m_1 = 0$ ، $m'_1 = \bar{x}$ و همچنین $m_2 \simeq S_x^2$.

تمرین ۳.۲.۱ کلاسی. با استفاده از جدول فراوانی مثال ۴.۲.۱، دو ستون جدید بعنوان ستون های $f_i x_i^3$ و $f_i(x_i - \bar{x})^3$ اضافه کرده و با محاسبهٔ مجموع این ستونها مقادیر گشتاورهای m_3 و m'_3 را بدست آورید.

۶.۲.۱ چولگی و کشیدگی

در یک **جامعهٔ نرمال** ، با بررسی داده ها و رسم شاخص هندسی دیده می شود که منحنی جامعه بصورت منحنی زیر است. این منحنی به منحنی زنگوله‌ای یا **منحنی نرمال** مشهور بوده و شکل جامعهٔ کاملاً نرمال را نشان می دهد. در این توزیع متقارن، مقدار مد، میانه و میانگین برابر بوده و روی محور تقارن نمودار واقع می شوند. از آنجاکه **عملاً** هیچ نمونهٔ کاملاً نرمالی یافت نمی شود، با رسم نمودار داده های یک نمونه و مقایسهٔ آن با نمودار نرمال، میزان انحراف در داده‌ها را می توان دریافت. بنابراین پس از ترسیم نمودار چندبرفراوانی، معمولاً شکل آن شبیه منحنی نرمال، چاوله، J— شکل (مانند مثال ۴.۲.۱) یا در برخی موارد U—شکل خواهد بود. اگر معیار ما نرمال بودن جامعهٔ آماری از لحاظ صفت مورد نظر باشد، پس می توانیم منحنی داده ها را با منحنی نرمال سنجیده و مقدار انحراف داده ها را بدست آوریم. این مقدار انحراف که مقدار دوری داده ها را از توزیع متقارن نشان می دهد تحت عنوان **چاولگی** شناخته می شود.



دو نمودار فوق که با کمی مچاله شدن داده ها به نمودار نرمال شبیهند، تحت عنوان منحنی **چوله به راست** و **چوله به چپ** مطرحند. وقتی در این منحنی ها دامنه به سمت چپ کشیده تر باشد آنرا چوله به چپ و اگر دامنه به سمت راست کشیده تر باشد آنرا چوله براست گوئیم. پس چوله براست یعنی اکثر داده ها در طرف چپ جمع شده اند. میزان چولگی یک نمودار را توسط مقیاسهای چولگی اول پیرسن b_1 و چولگی دوم پیرسن b_2 و ضریب گشتاوری چولگی g مطابق فرمول های زیر بدست می آوریم.

$$b_1 = \frac{\bar{x} - M}{S_x} \quad , \quad b_2 = 3\frac{\bar{x} - m}{S_x} \quad , \quad g = \frac{m_3}{S_x^3}$$

برای نمودار نرمال $m_۳ = ۰$ و درنتیجه g صفر است. در نمودارهای دیگر اگر این سه معیار چولگی مثبت باشند نمودار چوله به راست و اگر منفی باشند نمودار چوله به چپ خواهد بود. از آنجائیکه مد در برخی مسائل بدقت محاسبه نمی شود (مثل جداول چندمدی)، بجای ضریب $b_۱$، می توان از ضریب $b_۲$ استفاده کرد که از میانه بهره می برد. ایندو کمابیش با هم برابرند و از نظر تجربی دیده شده که $-۱ < b_۱, b_۲ < ۱$ است.

تمرین ۴.۲.۱ در مثال۴.۲.۱ که نمودار چوله به راست است، ضرایب چولگی را محاسبه کنید.

همچنین ممکن است داده های مختلف از لحاظ میانگین و پراکندگی و حتی چاولگی مساوی باشند ولی از لحاظ شکل ظاهری متفاوت باشند مثلاً بسمت بالا **کشیده تر** باشد، این مقدار کشیدگی (برجستگی یا پخی) با یک ضریب بصورت $\beta_۲ = \frac{m_۴}{S_x^۴}$ بیان می شود که آنرا **ضریب کشیدگی پیرسن** نامند. مقدار کشیدگی برای منحنی نرمال محاسبه شده و برابر $\beta_۲ = ۳$ است. بنابراین اگر $\beta_۲ > ۳$ توزیع داده ها از منحنی نرمال بلندتر و اگر $\beta_۲ < ۳$ از منحنی نرمال کوتاهتر و پخ تر است. **فیشر** برای ضریب کشیدگی مقدار $\gamma = \beta_۲ - ۳$ را پیشنهاد می کند پس اگر $\gamma > ۰$ منحنی داده ها از منحنی نرمال بلندتر بطرف بالا کشیده است و اگر $\gamma < ۰$ از منحنی نرمال کوتاهتر و بطرف کناره ها متمایل تر است.

مثال ۵.۲.۱ در یک نمونه‌گیری از درختان یک باغ ۳۰ هکتاری، نمونهٔ قطر ۲۵ درخت برحسب سانتیمتر بصورت زیر بدست آمده است:

۳۰	۲۵	۲۳	۲۰	۳۹
۱۸	۳۵	۳۴	۲۵	۳۲
۲۷	۲۳	۳۱	۳۷	۲۶
۲۵	۴۲	۳۷	۲۴	۳۰
۲۸	۲۲	۳۵	۳۱	۲۹

با رسم جدول توزیع فراوانی، شاخص های مرکزی و پراکندگی را یافته و گشتاورهای مرتبه اول، دوم و سوم را بدست آورید. با رسم نمودارچندبرفراوانی و میله ای، چولگی داده ها را مشخص نموده و ضرایب چولگی را بنویسید.

حل. در این نمونهٔ ۲۵تائی داریم $R = x_{max} - x_{min} + ۱ = ۴۲ - ۱۸ + ۱ = ۲۵$ اگر تعداد دسته را ۵ انتخاب کنیم پس $T = \frac{۲۵}{۵} = ۵$ و طول دسته را برابر ۵ می گیریم. مانند آنچه ذکر شد برای اینکه هر داده در یک دسته قرار گیرد، داده ها را از ۱۷/۵ آغاز می کنیم و جدول فراوانی بصورت زیر خواهد بود:

دسته‌ها	x_i	f_i	F_i	$\overline{f_i}$	$f_i x_i$	$x_i - \bar{x}$	$f_i(x_i - \bar{x})^۲$	$x_i^۲$	$f_i x_i^۲$	$f_i x_i^۳$	$f_i(xi - \bar{x})^۳$
۱۷/۵ – ۲۲/۵	۲۰	۳	۳	۰/۱۲	۶۰	−۹	۲۴۳	۴۰۰	۱۲۰۰	۲۴۰۰۰	−۲۱۸۷
۲۲/۵ – ۲۷/۵	۲۵	۸	۱۱	۰/۳۲	۲۰۰	−۴	۱۲۸	۶۲۵	۵۰۰۰	۱۲۵۰۰۰	−۵۱۲
۲۷/۵ – ۳۲/۵	۳۰	۷	۱۸	۰/۲۸	۲۱۰	۱	۷	۹۰۰	۶۳۰۰	۱۸۹۰۰۰	۷
۳۲/۵ – ۳۷/۵	۳۵	۵	۲۳	۰/۲۰	۱۷۵	۶	۱۸۰	۱۲۲۵	۶۱۲۵	۲۱۴۳۷۵	۱۰۸۰
۳۷/۵ – ۴۲/۵	۴۰	۲	۲۵	۰/۰۸	۸۰	۱۱	۲۴۲	۱۶۰۰	۳۲۰۰	۱۲۸۰۰۰	۲۶۶۲
		۲۵		۱	۷۲۵		۸۰۰	۴۷۵۰	۲۱۸۲۵	۶۸۰۳۷۵	۱۰۵۰

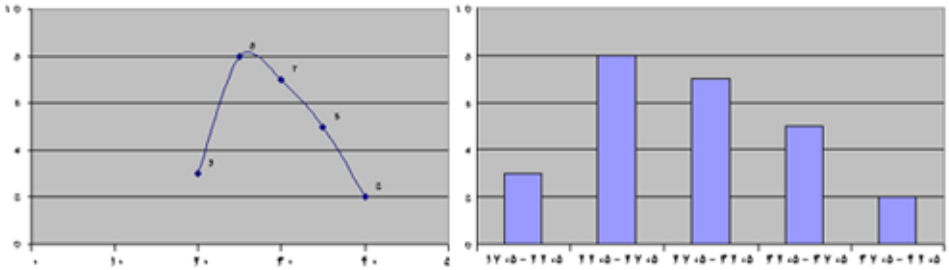
$\bar{x} = ۲۹$, $S_x^۲ = ۳۳/۳۴$, $S_x = ۵/۷۷$, $V_x = ۰/۲ = \%۲۰$

$m = a_i + \frac{\frac{n}{۲} - F_{i-۱}}{f_i} c = ۲۷/۵ + \frac{۱۲/۵ - ۱۱}{۷} ۵ = ۲۸/۵۷$, $M = a_i + \frac{d_۱}{d_۱ + d_۲} c = ۲۲/۵ + \frac{۵}{۵+۱} ۵ = ۲۶/۶۷$

$m_۱ = ۰$, $m_۲ = ۳۲$, $m_۳ = ۴۲$, $m'_۱ = ۲۹$, $m'_۲ = ۸۷۳$, $m'_۳ = ۲۷۲۱۵$

$b_۱ = ۰/۴$, $b_۲ = ۰/۲۲$, $g = ۰/۲۲$, $\gamma = \frac{m_۴}{S_x^۴} - ۳ = \frac{۲۳۰۰}{۱۰۲۴} - ۳ = -۰/۷۵$

نمودار داده ها بصورت ذیل بوده و از نمودار چندبرفراوانی دیده می شود که نمودار داده ها چوله به راست بوده و مقادیر b_1 و b_2 و g این را بخوبی تایید می کنند. مقدار ضریب کشیدگی فیشر $\gamma = -0/75$ مشخص می کند که نمودار داده ها پخ است.



۷.۲.۱ تغییر داده‌ها

اگر n دادهٔ x_1 و x_2 و $\cdots$ و x_n داده‌های یک نمونهٔ مفروض باشند و n دادهٔ y_1 و y_2 و $\cdots$ و y_n چنان باشند که $y_i = ax_i + b$ در آنصورت با داشتن $\bar{x}$ و S_x می توان $\bar{y}$ و S_y را بصورت زیر بدست آورد

$$\bar{y} = a\bar{x} + b \quad , \quad S_y^2 = a^2.S_x^2 \quad , \quad S_y = |a|S_x$$

برای مثال داده های ۴۳ و ۴۱ و ۴۲ و ۴۴ و ۴۰ را در نظر گرفته و آنها را متغیر x_i می نامیم. فرض کنید که $y_i = x_i - 40$ پس y_i—ها بترتیب عبارتند از ۳ و ۱ و ۲ و ۴ و ۰ و از آنجاکه $\bar{y} = 2$ و $\bar{y} = \bar{x} - 40$ نتیجه می گیریم که $\bar{x} = 42$ است. بعلاوه

$$S_y^2 = \frac{(3-2)^2 + (1-2)^2 + (2-2)^2 + (4-2)^2 + (0-2)^2}{4} = \frac{10}{4} = 2/5 \implies S_y = |1|S_x \implies S_x = 2/5$$

در برخی موارد ساده تر است برای بدست آوردن میانگین داده های دسته بندی شده از روش **کدگذاری** استفاده کنیم. اگر x_i نمایندهٔ دسته های موجود در جدول فراوانی بوده و c فاصلهٔ دو دسته و A داده دسته وسطی باشد. با تغییر داده ها بصورت بالا، فرض می گیریم که $y_i = \frac{x_i - A}{c}$ با اضافه کردن ستون y_i به جدول و یافتن $\bar{y}$ از فرمول $\bar{x} = c\bar{y} + A$ مقدار $\bar{x}$ را پیدا می کنیم.

۸.۲.۱ تصحیح شپارد برای واریانس

در دسته بندی داده ها بعلت داخل شدن خطا در دسته بندی داده ها، مقدار $\frac{c^2}{12}$ را از واریانس S_x^2 کم می کنیم تا جبران خطا شود این مقدار به **تصحیح شپارد** معروف است. بنابراین واریانس تصحیح شده $\hat{S}_x^2$ از رابطهٔ زیر بدست می آید:

$$\hat{S}_x^2 = S_x^2 - \frac{c^2}{12}$$

۹.۲.۱ میانگین اصلاح شده

در نمونه هائی که یک یا چند دادهٔ پرت دارند چون این تعداد کم بهرحال روی میانگین تاثیر می گذارند در این موارد بهتر است بجای میانگین از میانه استفاده کنیم، زیرا میانه تحت تاثیر داده پرت قرار نمی گیرد. روش دیگر حذف داده های پرت از فهرست داده هاست. در این روش پس از مرتب کردن داده ها، به تعداد داده های پرت از ابتدا و انتهای فهرست حذف نموده و میانگین

مابقی داده ها را بعنوان میانگین کل داده ها بکار می گیریم. برای مثال در داده های ۱ و ۲ و ۳ و ۴ و ۵ و ۴۵ مقدار دادهٔ ۴۵ مقداری پرت بوده و میانگین ۱۰ معیار تمرکز مفیدی نیست. در این حالت می توان میانهٔ $m = ۳/۵$ را بجای میانگین بکار برد و یا با استفاده از روش دوم میانگین ۲ و ۳ و ۴ و ۵ را که برابر ۳/۵ است، برای نمونه انتخاب نمود. اگر تعداد داده های پرت نمونه مرتب شده برابر r باشد، میانگین اصلاح شده مرتبهٔ r عبارتست از

$$\hat{\bar{x}} = \frac{۱}{n - ۲r} \sum_{i=r+۱}^{n-r} x_i$$

۱۰.۲.۱ چندک ها

دیدیم که میانه داده ها را به دو دسته تقسیم می کرد و در حالتیکه داده ها دارای فراوانی بودند، میانه فراوانی را دو قسمت کرده و دادهٔ «وسط» را انتخاب می نمود. اگر در یک نمونه، فراوانی داده‌ها را به چهار دسته تقسیم کنیم، داده های **چارکی** Q_j بدست می آیند. بطورکلی ما تنها دارای سه چارک $Q_۱$ و $Q_۲$ و $Q_۳$ هستیم، که چارک $Q_۲$ همان میانه خواهد بود. در داده های فاقد دسته بندی — پس از مرتب کردن داده ها— برای محاسبهٔ چارک Q_j عبارت $\frac{j}{۴}(n+۱)$ را محاسبه کرده و مقدار صحیح آن را با k و اعشار آنرا با ω نشان می دهیم. سپس فرمول $Q_j = x_k + \omega(x_{k+۱} - x_k)$ مقدار چارک را بدست می دهد.

مثال ۶.۲.۱ چارکها را در داده های ۱۴ و ۱۷ و ۲۴ و ۱۸ و ۲۱ بدست آورید.

با مرتب کردن داده ها داریم: ۲۴ و ۲۱ و ۱۸ و ۱۷ و ۱۴ از آنجاکه $n = ۵$

$\frac{۱}{۴}(۵+۱) = \frac{۳}{۲} = ۱/۵ \Rightarrow k = ۱, \omega = ۰/۵ \Rightarrow Q_۱ = x_۱ + ۰/۵(x_۲ - x_۱) = ۱۵/۵$ چارک اول

$\frac{۲}{۴}(۵+۱) = \frac{۶}{۴} = ۳ \Rightarrow k = ۳, \omega = ۰ \Rightarrow Q_۲ = x_۳ + ۰(x_۴ - x_۳) = ۱۸$ چارک دوم

$\frac{۳}{۴}(۵+۱) = \frac{۱۸}{۴} = ۴/۵ \Rightarrow k = ۴, \omega = ۰/۵ \Rightarrow Q_۳ = x_۴ + ۰/۵(x_۵ - x_۴) = ۲۲/۵$ چارک سوم

تمرین ۵.۲.۱ **کلاسی.** چارک ها را در داده های زیر محاسبه کنید.

$$۵, ۷, ۹, ۱۰, ۱۱, ۱۵$$

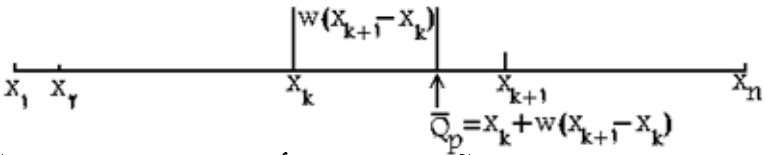
در داده های دسته بندی شده مانند فرمول میانه، دستهٔ i که فراوانی تجمعی آن بیشتر یا مساوی $j\frac{n}{۴}$ باشد را یافته و از فرمول

$$Q_j = a_i + \frac{j\frac{n}{۴} - F_{i-۱}}{f_i} c \quad , \quad j = ۱, ۲, ۳$$

برای محاسبه چارک ها استفاده می کنیم. مانند چارک، اگر دامنه به ۱۰ بخش تقسیم شود، هر قسمت را **دهک** گفته و با $D_۱, D_۲, \cdots, D_۹$ نشان می دهیم، بعلاوه دهک پنجم همان میانه است. اگر دامنه ۱۰۰ قسمت شود آنرا **صدک** گوئیم و با $P_۱, P_۲, \cdots, P_{۹۹}$ نشان داده و صدک پنجاهم همان میانه خواهد بود.

عدد $\overline{Q}_p$ را بازاء $۰ < p < ۱$ حقیقی، **چندک مرتبه** p نامیم، هرگاه تقریباً $p۱۰۰\%$ داده ها کوچکتر از آن باشند. مثلاً صدک دوم ($p = \frac{۲}{۱۰۰}$) را با $\overline{Q}_{\frac{۲}{۱۰۰}}$ نشان داده و مقداری است که دو درصد داده ها کمتر از آن قرار می گیرند. بهمین ترتیب دهک هفتم ($p = \frac{۷}{۱۰}$) را با $\overline{Q}_{\frac{۷}{۱۰}}$ و چارک سوم ($p = \frac{۳}{۴}$) را با $\overline{Q}_{\frac{۳}{۴}}$ نشان می دهیم، روشن است که $m = \overline{Q}_{\frac{۱}{۲}} = Q_۲ = D_۵ = P_{۵۰}$.

در داده های فاقد دسته‌بندی – پس از مرتب کردن داده ها– برای محاسبهٔ چندک $\overline{Q}_p$ عبارت $p(n+۱)$ را محاسبه کرده، مقدار صحیح آن را با k و اعشار آنرا با ω نشان می دهیم. سپس $\overline{Q}_p = x_k + \omega(x_{k+۱} - x_k)$ مقدار چندک را بدست می دهد.



برای n دادهٔ دسته بندی شده، مقادیر چندکی چنین بدست می آیند که ابتدا دستهٔ i–ام که فراوانی تجمعی آن بیشتر یا مساوی pn باشد را یافته و فرمول زیر را بکار می گیریم:

$$\overline{Q}_p = a_i + \frac{pn - F_{i-۱}}{f_i}\, c$$

تمرین ۶.۲.۱ منزل.

۱) سرشماری چیست؟ تفاوت آن را با نمونه گیری بیان کنید.

۲) ویژگیهای یک نمونه برداری سالم و کارآمد چیست؟ مثال بزنید.

۳) نماد سیگما $\sum$ به معنای جمع چند عدد است، بدینصورت

$$\sum_{i=m}^{n} x_i = x_m + x_{m+۱} + x_{m+۲} + \cdots + x_n$$

مثلاً

$$\sum_{i=۱}^{۴} \frac{i}{i+۵} = \frac{۱}{۱+۵} + \frac{۲}{۲+۵} + \frac{۳}{۳+۵} + \frac{۴}{۴+۵}$$

با استفاده از این تعریف مقادیر زیر را بدست آورید:

$$(a)\ \sum_{i=۱}^{۸} \frac{۲i+۶}{۵} \quad , \quad (b)\ \sum_{i=۱}^{۱۰} ۶i-۳ \quad , \quad (c)\ \sum_{i=۱}^{۴} \frac{i}{۴} \quad , \quad (d)\ \sum_{i=۱}^{۴} (۲i-۶) \quad , \quad (e)\ \sum_{i=۱}^{۱۰} \frac{i+۲}{۳} \quad , \quad (f)\ \sum_{i=۱}^{۵} i^۲+i+۱$$

۴) نماد سیگمای تعریف شده در تمرین قبل دارای خواصی بصورت زیر است که از آنها بهره می گیریم:

$$\sum_{i=۱}^{n} C = nC \quad , \quad \sum_{i=۱}^{n} a x_i = a \sum_{i=۱}^{n} x_i \quad , \quad \sum_{i=۱}^{n} (x_i + y_i) = \sum_{i=۱}^{n} x_i + \sum_{i=۱}^{n} y_i$$

اگر $\sum_{۱}^{۵} x_i = ۴$ و $\sum_{۱}^{۵} x_i^۲ = ۱۳$ باشد مطلوبست مقادیر

$$(a)\ \sum_{۱}^{۵} (۲x_i + ۳) \quad , \quad (b)\ \sum_{۱}^{۵} x_i(x_i + ۲) \quad , \quad (c)\ \sum_{۱}^{۵} (x_i + ۳)^۲ \quad , \quad (d)\ \sum_{۱}^{۵} \frac{x_i^۳ - ۱}{x_i - ۱}$$

۵) در یک نمونه گیری از ۵۰ خانوادهٔ ایرانی تعداد فرزندان این خانواده ها بصورت زیر بدست آمده است:

۲ و ۱ و ۳ و ٦ و ۱ و ۲ و ۰ و ۱ و ۵ و ۲ و ۴ و ٦ و ۲ و ۳ و ۰ و ۱ و ۲ و ۵ و ۱ و ۴ و ۲ و ۳ و ۱ و ۱ و ۰ و ۱ و ۲ و ۳ و ۳ و ۰ و ۰ و ۲ و ۱ و ۲ و ۳ و ۱ و ۵ و ۴ و ۲ و ۱ و ۲ و ۰ و ۱ و ۲ و ۱ و ۱ و ۳ و ٦ و ۲ و ۴

با رسم جدول توزیع فراوانی داده ها، شاخص های مرکزی و پراکندگی را یافته و گشتاورهای مرتبه اول، دوم و سوم و ضریب کشیدگی را بدست آورید. با رسم نمودار میله ای و چند بر فراوانی، چولگی داده ها را مشخص نموده و ضرایب چولگی را بنویسید.

٦) بصورت تجربی مشخص شده که بین میانگین، میانه و مد رابطه تقریبی زیر برقرار است $\bar{x} - M \simeq ۳(\bar{x} - m)$ صحت این قاعده را برای مثال های ۱.۲.۱ و ۴.۲.۱ و تمرین۵ بررسی نمائید. (دقت کنید که این رابطه برای بدست آوردن یک پارامتر از دو پارامتر دیگر نبوده و تنها روشی برای کنترل اطلاعات و صحت ارقام بدست آمده محسوب می شود).

۷) یک کارخانه کنسروسازی، قوطی هائی با وزن ۲۰۰ گرم عرضه می کند. در یک نمونه گیری از ۴۰ قوطی عرضه شده به بازار وزن آنها برحسب گرم بصورت زیر اندازه گیری شده است:

۱۹۸ و ۱۸۱ و ۱۹۲ و ۱۹۹ و ۲۰۳ و ۲۰۵ و ۱۹۴ و ۱۸۰ و ۲۰۵ و ۲۰۷ و ۱۹۹ و ۲۰۰ و ۲۰۷ و ۲۱۸ و ۱۹۰ و ۱۸۸ و ۱۹۱ و ۲۱۰ و ۱۹۹ و ۲۰۱ و ۲۰۷ و ۲۱۴ و ۲۰٦ و ۲۰۴ و ۱۹۹ و ۱۹٦ و ۱۹۴ و ۲۰۱ و ۱۸۴ و ۲۰۲ و ۱۹۸ و ۲۰۰ و ۲۰۲ و ۱۸۲ و ۱۹۵ و ۱۹۴ و ۱۸۲ و ۲۰۴ و ۱۹۹ و ۲۰۳

با رسم جدول توزیع فراوانی داده ها، شاخص های مرکزی و پراکندگی را یافته و گشتاورهای مرتبه اول، دوم و سوم را بدست آورید. با رسم هیستوگرام و نمودار داده ها، چولگی داده ها را مشخص نموده و ضرایب چولگی را بنویسید. نمودار چندبرفراوانی صعودی و نزولی را رسم و میانه را با استفاده از آن بیابید.

۸) از نظر هندسی و نمودارها، چندک $\overline{Q}_p$ چه معنی دارد؟ چارکها، دهک سوم $D_۳$ و دهک هفتم $D_۷$ تمرین۷ را حساب کنید.

۹) برای داده های ٦ و ۲ و ۱۲ و ۹ و ۴ و ۸ و ۵ مقادیر چندکی $\overline{Q}_{\frac{۱۲}{۲۰}}$, $\overline{Q}_{\frac{۹}{۱۰۰}}$, $\overline{Q}_{\frac{۴}{۵}}$, $P_{۱۰}$, $D_۲$, $Q_۱$ را حساب کنید.

۱۰) اگر میانگین داده های $x_۱, x_۲, \cdots, x_n$ برابر با $\bar{x}$ باشد ثابت کنید $\sum_۱^n (x_i - \bar{x}) = ۰$

۱۱) برای داده های مثبت $x_۱, x_۲, \cdots, x_n$ میانگین هندسی G و میانگین توافقی H را چنین تعریف می کنیم:

$$G = \sqrt[n]{x_۱ x_۲ \cdots x_n} \quad , \quad H = \frac{n}{\sum \frac{۱}{x_i}}$$

رابطه ای که بین $\bar{x}$ و G و H برقرار است بصورت $H \leq G \leq \bar{x}$ این رابطه را برای داده های مثال ۱.۲.۱ بررسی نمائید.

۱۲) گشتاورهای مرتبهٔ اول تا چهارم را برای داده های ۱ و ۳ و ۵ و ۷ و ۹ حساب کنید.

۱۳) گاهی دامنه تغییرات R را بعنوان شاخص پراکندگی محسوب می کنند، بنظر شما آیا می توان از این شاخص نتیجه‌ای آماری گرفت؟ میانبرد که با $R' = \frac{۱}{۲}(x_{max} + x_{min})$ تعریف می شود چطور؟.

۱۴) میانه داده های ۵ و ٦ و ۷ و ۸ با فراوانی های بترتیب برابر ۴ و ۵ و ۳ و ۵ چند است. مد آنها چند است؟

۱۵) ثابت کنید که در فرمول واریانس اگر بجای $\bar{x}$ هر عدد دیگری قرار دهیم، مقدار بدست آمده از واریانس بیشتر خواهد شد. بعبارتی مقدار واریانس کوچکترین عددی است که با مقدار میانگین از فرمول حاصل می شود.

۱۶) یکی از مقیاس های سنجش چولگی در نمودارها با استفاده از چارکها و بررسی علامت $Q_1 + Q_3 - 2Q_2$ است. اگر این مقدار مثبت باشد منحنی چوله به راست و اگر منفی چوله به چپ است و در حالت نرمال صفر خواهد بود. این مقدار را برای مثالهای۴.۲.۱ و ۵.۲.۱ بررسی و چولگی آنها را بسنجید.

۱۷) برای تعیین میانگین اعداد وزن دار مانند نمرات دارای ضریب، از فرمول میانگین وزن دار بصورت زیر استفاده می کنیم:

$$\text{میانگین وزنی} = \overline{x}_w = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n}$$

ضرایب فردی در کنکور در درس ریاضی، فیزیک، شیمی و ادبیات بترتیب ۴۰، ۲۸، ۳۲ و ۵۰ درصد است اگر این دروس بترتیب دارای ضرایب ۴،۳،۳و۴ باشند رتبه فرد را محاسبه کنید.

۱۸) ثابت کنید اگر به تمام داده ها مقدار ثابت a اضافه شود، گشتاورهای مرکزی، معیارهای چولگی و برجستگی داده ها تغییری نخواهند کرد.

۱۹) کالفورد[۱۳] آمار شناس معروف نشان داد که اگر یک **توزیع متقارن** یا **خفیفاً چوله** باشد و اندازه نمونه نیز بزرگ باشد انحراف معیار یک ششم دامنه تغییرات است یعنی $R = 6S_x$. درستی این رابطه را برای مثال های۱.۲.۱، ۴.۲.۱ و۵.۲.۱ بررسی نمائید. (این رابطه صرفاً برای بررسی صحت عملیات است).

۲۰) برای استاندارد کردن داده‌های x_i از $z_i = \frac{x_i-\bar{x}}{S_x}$ استفاده کرده و z_i را متغیر استاندارد شده گوئیم. ثابت کنید $\bar{z} = 0$ و $S_z = 1$.

۲۱) **انواع نمودار در اکسل.** در برنامهٔ اکسل[۱۴] کامپیوتر، انواع متنوع و مختلفی از نمودارها معرفی شده‌اند. برای آشنائی بیشتر با این نمودارها، شما بایستی با در نظرگرفتن داده‌های مثال های۱.۲.۱ و ۵.۲.۱ نمایش مختلف داده ها را بصورت نمودار رسم نمائید: ستونی، میله‌ای، خطی، کلوچه‌ای، $XY-$ پراکندگی، مسطح، دوناتی، راداری، سطحی، حبابی، انباشته، استوانه‌ای،مخروطی و هرمی.

---

[۱۳] $Cullford$
[۱۴] $Excel$