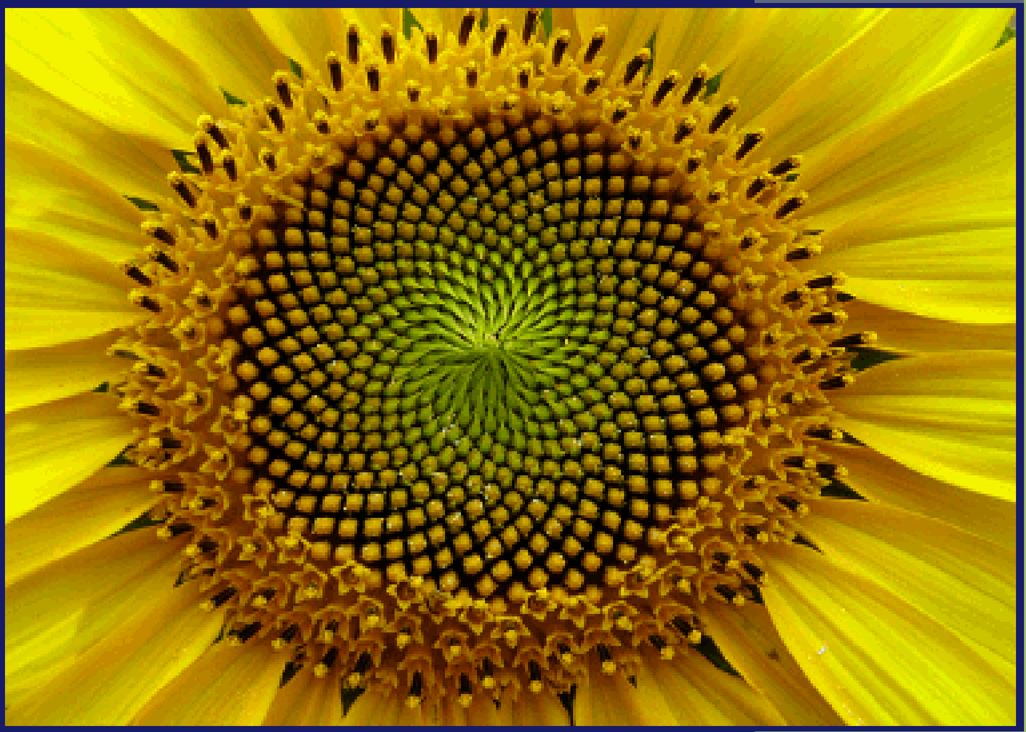
انتشارات دانشگاه کردستان

مقدمه‌ای بر

آمار و احتمال

چاپ دوم



قادر میرزاقادری

محمد مرادی

افشین فلاح

شامل حل مسائل منتخب با نرم‌افزار Minitab

بسم الله الرحمن الرحیم

مقدمه‌ای بر

آمار و احتمال

تألیف:

دکتر قادر میرزاقادری

عضو هیئت علمی دانشگاه کردستان

دکتر محمد مرادی

عضو هیئت علمی دانشگاه رازی

دکتر افشین فلاح

عضو هیئت علمی دانشگاه بین‌المللی امام خمینی

سرشناسه	:	میرزاقادری، قادر، ۱۳۵۳-
عنوان و نام پدیدآور	:	مقدمه‌ای بر آمار و احتمال/ تألیف قادر میرزاقادری، محمد مرادی، افشین فلاح.
مشخصات نشر	:	سنندج، دانشگاه کردستان، ۱۳۸۹.
مشخصات ظاهری	:	۳۴۳ ص.: مصور، جدول، نمودار.
شابک	:	۹۷۸-۹۶۴-۲۷۹۷-۲۷-۱
وضعیت فهرست‌نویسی	:	فیپا
یادداشت	:	کتابنامه.
موضوع	:	آمار ریاضی -- راهنمای آموزشی (عالی)
موضوع	:	احتمالات -- راهنمای آموزشی (عالی)
موضوع	:	آمار ریاضی -- مسائل، تمرین‌ها و غیره (عالی)
موضوع	:	احتمالات -- مسائل، تمرین‌ها و غیره (عالی)
شناسه افزوده	:	مرادی، محمد، ۱۳۵۵ -
شناسه افزوده	:	فلاح، افشین، ۱۳۵۴ -
شناسه افزوده	:	دانشگاه کردستان
رده‌بندی کنگره	:	QA۲۷۶/۲//م۹۳م۷ ۱۳۸۹
رده‌بندی دیویی	:	۵۱۹/۵۰۷۶
شماره کتابشناسی ملی	:	۲۱۵۵۲۶۰

انتشارات دانشگاه کردستان

نام کتاب: مقدمه‌ای بر آمار و احتمال

مؤلفین: میرزاقادری، مرادی، فلاح

چاپ دوم

ناشر: دانشگاه کردستان

تیراژ: ۲۰۰۰ نسخه

قطع: وزیری

فهرست مطالب

فصل ۱ آمار توصیفی

فصل ۲ احتمال

فصل ۳ متغیرهای تصادفی گسسته

فصل ٤ متغیرهای تصادفی پیوسته

فصل ۵ توزیع‌های نمونه‌ای

فصل ٦ استنباط آماری: فاصله اطمینان

فصل ۷ استنباط آماری: آزمون فرض

فصل ۸ آزمون کای‌دو و کاربردهای آن

فصل ۹ استنباط آماری در مورد دو جامعه

فصل ۱۰ مقدمه‌ای بر تحلیل واریانس

فصل ۱۱ تحلیل رگرسیون خطی

فصل ۱۲ آمار ناپارامتری

جداول ضمیمه

فصل ۱ آمار توصیفی

۱-۱ مقدمه

علم آمار به عنوان مجموعه‌ای از روش‌های جمع‌آوری، پاک‌سازی، تلخیص، تحلیل و نتیجه‌گیری قابل اعتماد از داده‌ها، تعریف می‌شود. پس از جمع‌آوری و پاک‌سازی داده‌ها می‌توان کارهای آماری بروی آنها را به دو بخش کلی توصیف و استنباط تقسیم کرد. بخش اول شامل تلخیص داده‌های پاک‌سازی شده در قالب جداول فراوانی و نمایش آنها در نمودارها و محاسبه ویژگی‌هایی از داده‌ها مانند مرکزیت، پراکندگی، تقارن و عدم تقارن می‌باشد، که به آن آمار توصیفی می‌گویند. در واقع هدف آمار توصیفی، نمایش و توصیف داده‌ها به صورتی راحت و قابل فهم است. بویژه در صورتی که حجم داده‌ها زیاد باشد بررسی صرفاً چشمی آنها کاری دشوار خواهد بود. آمار فقط به توصیف داده‌ها ختم نمی‌شود، بلکه اغلب اوقات محقق نمونه‌ای از یک جامعه گرفته و با توجه به اطلاعات موجود در نمونه می‌خواهد در مورد جامعه اظهار نظر کند. تعمیم نتایج حاصل از نمونه به جامعه را با استفاده از روش‌های آماری، آمار استنباطی می‌نامند.

۱-۲ جامعه آماری

جامعه آماری مجموعه عناصری است که دست کم یک صفت مشترک دارند. جوامع آماری به دو نوع متناهی و نامتناهی تقسیم می‌شوند. جامعه متناهی به جامعه‌ای گفته می‌شود که تعداد عناصر آن محدود است و جامعه نامتناهی نیز جامعه‌ای است که تعداد عناصر آن بینهایت باشد. در عمل گاهی اوقات که تعداد افراد جامعه بسیار بزرگ باشد، جامعه را نامتناهی در نظر می‌گیرند مثل ماهی‌های یک دریا یا باکتری‌های موجود در یک متر مربع از زمین زراعی.

مجموعه افراد یک شهر یا مجموعه ارقام زراعی گندم در ایران مواردی از جامعه متناهی هستند ولی جامعه تعداد پرتاب‌های لازم یک سکه برای مشاهده اولین شیر ({یک پرتاب، دو پرتاب،...}) نامتناهی است. اگر تعداد اعضای جامعه اندک باشد می‌توان نسبت به بررسی همه افراد آن اقدام کرد. بررسی آماری همه واحدهای جامعه را سرشماری گویند. به عنوان مثال در جامعه دانشجویان یک کلاس می‌توان نمرات همه دانشجویان را در یک درس خاص ثبت کرده و میانگین آنها را به دست آورد. گاهی در آمار با جوامع آنقدر بزرگی روبرو هستیم که بررسی همه اعضای آن غیرممکن، یا از لحاظ زمان یا هزینه سرشماری مقرون به صرفه نیست. برای مثال یک مزرعه بزرگ گندم را در نظر بگیرید. برای تعیین میانگین تعداد دانه در خوشه در این مزرعه نمی‌توان تعداد دانه را در همه خوشه‌ها شمارش و میانگین کل را به دست آورد. روش منطقی، انتخاب نمونه‌ای تصادفی از خوشه‌های گندم در این مزرعه و محاسبه میانگین تعداد دانه در خوشه‌های منتخب است.

۱-۳ نمونه

هر زیر مجموعه‌ای از واحدهای جامعه را نمونه گویند. نمونه ممکن است به صورت شانسی انتخاب شود که اصطلاحاً به آن نمونه احتمالی می‌گویند. در مقابل، نمونه غیراحتمالی وجود دارد که بر اساس سلیقه افراد انتخاب می‌شود. در آمار نمونه احتمالی به دو دلیل بر نمونه غیر احتمالی ترجیح داده می‌شود. اول اینکه قوانین احتمال در مورد آن صادق است و استنباط آماری نیز بر اساس همین قوانین ساخته می‌شود، دوم اینکه نمونه‌های غیراحتمالی به دلیل سلیقه‌ای بودن همواره در معرض این اتهام هستند که بیانگر کل جامعه نمی‌باشند. به عنوان مثال نمونه‌ای از محصول که توسط کشاورز برای فروش به خریدار عرضه می‌شود، ممکن است از طرف خریدار مورد شک گل‌چین بودن واقع شود. شعار نمونه‌گیری این است که مشت باید نمونه خروار باشد و این شعار در مورد نمونه‌های احتمالی صادق است. ساده‌ترین نمونه احتمالی نمونه تصادفی ساده (یا به طور خلاصه نمونه تصادفی) است که در آن هر فرد جامعه شانس مساوی برای انتخاب شدن در نمونه دارد. یک نمونه تصادفی را از راه قرعه‌کشی، جدول اعداد تصادفی یا به کمک نرم‌افزارهای آماری می‌توان به دست آورد. نمونه‌ای می‌تواند مشخصات جامعه را با دقت

برآورد کند که حجم بهینه‌ای داشته باشد و نمونه‌های خیلی کوچک، برآوردهای دقیقی از جامعه به دست نخواهند داد.

۱-٤ متغیر و داده

متغیر، مشخصه یا صفتی است که از عنصری به عنصر دیگر تغییر می‌کند. مانند نژاد، رتبه کنکور، دمای بدن و وزن افراد. متغیرها را می‌توان به چهار نوع اسمی، رتبه‌ای، فاصله‌ای و نسبتی یا با تقسیم بندی کلی‌تری به کیفی و کمی تقسیم‌بندی نمود.
داده یا مشاهده نیز اطلاعاتی است که از اندازه‌گیری یا تعیین مقدار متغیر مربوط به عناصر مورد مطالعه حاصل می‌شود مانند اندازه قد هر فرد، میزان محصول یک درخت یا رنگ گل‌های رز در یک باغچه.

۱-۵ مقیاس‌های استیونز

استیونز در سال ۱۹۵۱ چهار نوع مقیاس معرفی نمود که می‌توان همه متغیرهای ممکن را با این مقیاس‌ها اندازه‌گیری کرد.

۱-۵-۱ مقیاس اسمی

این مقیاس تنها برای شناسایی عناصر به کار می‌رود. مثلاً ممکن است کارگران کارخانه‌ای که متولد شهرهای تهران، تبریز، شیراز و اصفهان هستند با اعداد ۱، ۲، ۳ و ٤ مشخص شوند. این اعداد ارزش مقایسه نداشته و صرفاً بیانگر اسم شهرها هستند. لذا نمی‌توان روی آنها اعمال ریاضی مانند جمع، تفریق و تقسیم انجام است.

۱-۵-۲ مقیاس ترتیبی

اعداد تشکیل دهنده این مقیاس صرفاً برتری را نشان می‌دهند. مثلاً اگر مدیر یک کارخانه کارگران را از نظر مهارت با اعداد ۱، ۲، ۳، و ٤ مشخص کند، کارگر ٤ از کارگر ۲ ماهرتر است ولی نمی‌توان گفت که دو برابر او مهارت دارد. این گونه اعداد را می‌توان تنها برای مقایسه به کار برد و نمی‌توان با آنها چهار عمل اصلی ریاضی را انجام داد.

۱-۵-۳ مقیاس فاصله‌ای

اگر اعداد تشکیل دهنده مقیاسی، نسبت دو تفاضل (یا دو فاصله) را حفظ کند، آن را مقیاس فاصله‌ای گویند. مثلاً اگر دمای چهار جسم در مقیاس سانتیگراد (x) بصورت

$$x_۱ = ۱۰, \qquad x_۲ = ۱۵, \qquad x_۳ = ۲۰, \qquad x_٤ = ٤۵, \qquad \frac{x_٤ - x_۳}{x_۲ - x_۱} = ۵$$

باشد، همین دماها در مقیاس فارانهایت (y) طبق رابطه $y = \frac{۹x}{۵} + ۳۲$ برابر

$$y_۱ = ۵۰ \qquad y_۲ = ۵۹ \qquad y_۳ = ٦۸ \qquad y_٤ = ۱۱۳ \qquad \frac{y_٤ - y_۳}{y_۲ - y_۱} = ۵$$

هستند. اگر درجه حرارت در مقیاس سانتیگراد صفر باشد در مقیاس فارانهایت درجه حرارت ۳۲ است. در مقیاس فاصله‌ای صفر معنی هیچ نداده و صرفاً قراردادی است. در مقیاس سانتیگراد $x_۳$ دو برابر $x_۱$ است ولی در مقیاس فارانهایت چنین نیست. بنابراین در مقیاس فاصله‌ای نسبت اعداد محفوظ نمی‌ماند پس نمی‌توان گفت که جسم ۳ دو برابر گرمتر از جسم ۱ است

۱-۵-٤ مقیاس نسبی

مقیاس نسبی کامل‌ترین نوع مقیاس است و برای بیشتر متغیرهای قابل شمارش یا اندازه‌گیری مانند تعداد، وزن، طول، سطح و غیره به کار می‌رود. مثلاً اگر وزن دو جسم ٦۰۰۰ و ۲۰۰۰ کیلوگرم باشد، وزن آنها در مقیاس تن به ترتیب ٦ و ۲ می‌باشد. با اعداد حاصل از این نوع مقیاس می‌توان چهار عمل اصلی را انجام داد. صفر در این نوع مقیاس واقعی است.

متغیر کیفی

اگر متغیر با عدد قابل بیان نباشد، به آن متغیر کیفی گویند مانند متغیرهایی که با مقیاس‌های اسمی یا رتبه‌ای بیان می‌شوند. وجود یا عدم وجود کُرک بر روی برگ، رنگ گلبرگ و مقاومت گیاه گندم به بیماری زنگ مثال‌هایی از متغیرهای کیفی هستند.

متغیر کمی

متغیر کمی متغیری است قابل اندازه‌گیری یا شمارش، مانند متغیرهایی که با مقیاس‌های فاصله‌ای یا نسبی بیان می‌شوند. متغیرهای کمی خود به دو نوع پیوسته و گسسته تقسیم می‌شوند. بسیاری از متغیرهای فیزیکی قابل اندازه‌گیری، پیوسته هستند و می‌توانند هر مقدار ممکن در دامنه تغییرات (فاصله بین بیشترین و کمترین مقدار) موجود را به خود اختصاص دهند. متغیرهایی چون وزن دانه، ارتفاع گیاه، درصد روغن بذر، درصد چربی شیر و میزان تولید شیر در دام‌ها، پیوسته هستند. متغیر گسسته متغیری است که مجموعه مقادیر ممکن آن شمارا باشد. مانند تعداد گلبرگ‌های یک گل، تعداد دانه‌ها در یک سنبله و تعداد لاروهای موجود بر روی یک گیاه.

۱-۶ پارامتر

پارامتر کمیتی است که یک ویژگی جامعه را نشان می‌دهد. پارامتر جامعه کمیتی است ثابت ولی ناشناخته و عمدتاً از طریق نمونه‌گیری برآورد می‌شود. البته می‌توان مقدار دقیق آن را از طریق سرشماری به دست آورد. پارامتر با حروف کوچک یونانی بیان می‌شود. مثلاً میانگین جامعه را با μ (مو)، میانه را با η (اِتا) و انحراف معیار جامعه را با σ (سیگما) نشان می‌دهند.

۱-۷ آماره

هر آماره تابعی است که بر اساس مشاهدات حاصل از نمونه تصادفی محاسبه می‌شود و ویژگی خاصی از نمونه را نشان می‌دهد. مقدار هر آماره به دلیل وابستگی آن به مشاهدات نمونه، از نمونه‌ای به نمونه دیگر متغیر است و از این رو تصادفی است. آماره‌ها با حروف انگلیسی نشان

داده می‌شوند.

۱-۸ جدول‌ها و نمودارهای فراوانی

داده‌های جمع‌آوری شده در اولین مرحله به صورت توده‌ای از اعداد خام هستند و قابلیت تعبیر و تفسیر جامعه را ندارند. پس ابتدا باید در یک جدول خلاصه شده و یا در نمودار مناسبی ترسیم شوند تا قابلیت توصیف جامعه را پیدا کنند. برای رسم جدول فراوانی مربوط به داده‌های کیفی یا گسسته، مقادیر مختلف متغیر در یک ستون نوشته شده و مواردی مثل فراوانی (فراوانی مطلق)، فراوانی نسبی و فراوانی تجمعی در ستون‌های دیگری قید می‌شود. فراوانی، تعداد دفعات تکرار هرداده بوده و با f_i نشان داده می‌شود که در آن i شماره داده یا شماره دسته است. فراوانی نسبی که با p_i نشان داده می‌شود برابر است با فراوانی هر داده یا دسته تقسیم بر تعداد کل (N). فراوانی تجمعی نیز با Fc_i نشان داده می‌شود و برابر است با فراوانی هر دسته به اضافه فراوانی‌های طبقات قبل از آن.

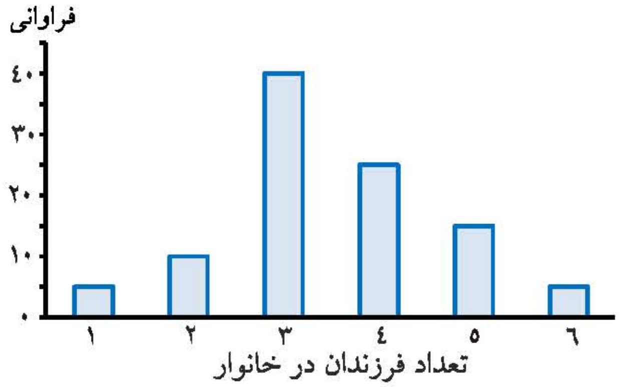
مثال ۱-۱تعداد خانواده‌های ۱، ۲، ۳، ٤، ۵ و ٦ فرزندی در یک روستا به ترتیب برابر با ۵، ۱۰، ٤۰، ۲۵، ۱۵ و ۵ بوده است. جدول فراوانی تعداد فرزندان می‌تواند به صورت زیر باشد.

جدول ۱-۱. جدول تعداد فرزندان در خانواده‌های یک روستا.

جمعیت خانواده	فراوانی (f_i)	فراوانی نسبی (p_i)	فراوانی تجمعی (Fc_i)
۱	۵	۰/۰۵	۵
۲	۱۰	۰/۱۰	۱۵
۳	٤۰	۰/٤۰	۵۵
٤	۲۵	۰/۲۵	۸۰
۵	۱۵	۰/۱۵	۹۵
٦	۵	۰/۰۵	۱۰۰
جمع	$N=$۱۰۰	۱	

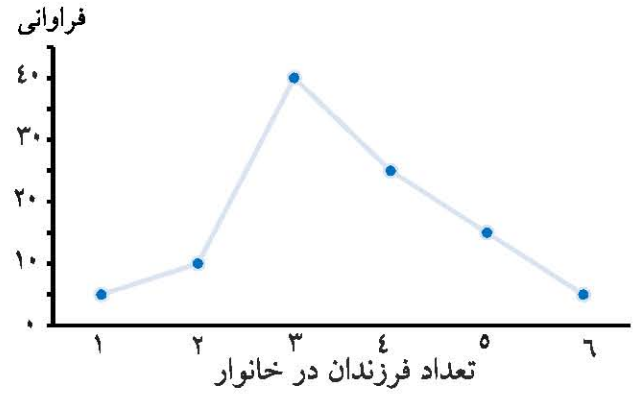
علاوه بر جدول فراوانی، برای توصیف بهتر داده‌ها می‌توان از نمودارهای فراوانی نیز استفاده کرد. در این نمودارها معمولاً مقادیر مشاهده شده متغیر بر روی محور افقی و فراوانی، فراوانی

نسبی یا فراوانی نسبی تجمعی مشاهدات نیز بر روی محور عمودی، مشخص مـی‌شـود (نمـودار ۱-۱، نمودار ۱-۲ و نمودار ۱-۴).

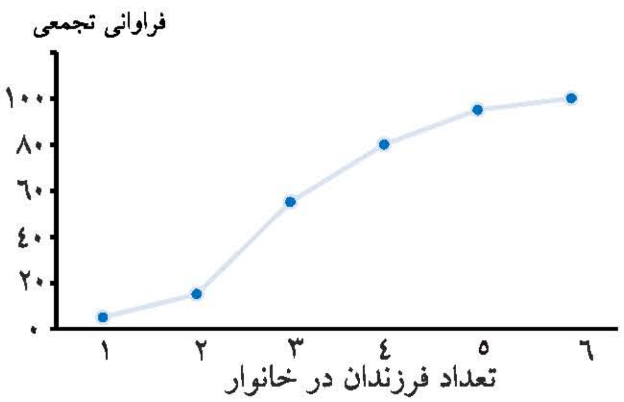


نمودار ۱-۱. نمودار فراوانی ستونی مربوط به داده‌های جدول ۱-۱.

هرگاه فراوانی هر مقدار متغیر با یک نقطه مشخص و سپس نقاط به هم وصل شوند، نمودار چند ضلعی حاصل می‌شود.

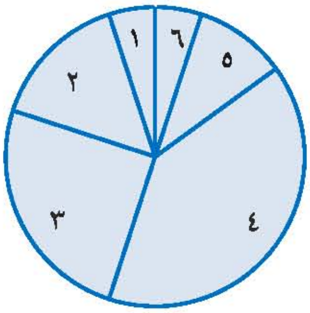


نمودار ۱-۲. نمودار فراوانی چندضلعی مربوط به داده‌های جدول ۱-۱.



نمودار ۱-۳. نمودار فراوانی چندضلعی تجمعی مربوط به داده‌های جدول ۱-۱.

نمودار دایره‌ای (یا کلوچه‌ای) را نیز می‌توان برای داده‌های مثال ۱-۱ رسم نمود. در این نمودار دایره متناسب با فراوانی مقادیر متمایز متغیر تقسیم‌بندی می‌شود. زاویه هر قاچ برابر حاصلضرب فراوانی نسبی در عدد ۳٦۰ است.



نمودار ۱-٤. نمودار دایره‌ای مربوط به داده‌های جدول ۱-۱ که در آن اعداد هر قسمت، تعداد فرزندان و مساحت هر قسمت برابر با فراوانی می‌باشد.

به وسیله نمودارهای فراوانی می‌توان بسیاری از ویژگی‌های مشاهدات را براحتی بیان نمود. مثلاً دامنه تغییرات برابر با ۱۸ (۲۰−۲ = $x_{max} - x_{min}$) و بیشترین فراوانی مربوط به خانواده‌های سه فرزندی است. با استفاده از نمودار چند ضلعی نسبی تجمعی به آسانی می‌توان عنصری را که ۲۰٪ عناصر مقداری کم‌تر از آن دارند، پیدا کرد.

در حالتی که متغیر پیوسته یا گسسته دارای تعداد مقادیر متمایز زیادی باشد، معمولاً داده‌ها به چند دسته با فاصله مساوی تقسیم شده و فراوانی برای دسته‌ها مشخص می‌شود. در ستون اول جدول، حدود دسته‌ها نوشته می‌شود. یک ستون جداگانه نیز به نام نشان دسته اضافه می‌شود، که مقدار آن برابر نقطه وسط هر دسته است. برای مثال فرض کنید ارتفاع ۲۰ بوته از یک گونه گیاهی به صورت زیر بوده است.

جدول ۱-۲. ارتفاع ۲۰ بوته از یک گونه گیاهی بر حسب سانتی‌متر.

۲۳/۵	۱۹/٦	۱۸/٤	۱۸/۲	۲۲/۵	۱۳/۳	۲۳/۵	۲۰/۵	۲۳/۲	۱۹/۵
۲۷/۲	۲۲/۵	۲۲/۹	۲۱/۸	۲۰/٤	۱۷/۵	۲٤/۳	۲۲/۲	۱۵/۵	۱٦/۷

برای بررسی راحت‌تر بهتر است این داده‌ها، که مثالی از مقادیر یک متغیر پیوسته هستند، به چند دسته با فاصله مساوی تقسیم شده و فراوانی یا فراوانی نسبی داده‌ها در هر دسته مشخص شود. معمولاً بسته به تعداد کل داده‌ها، تعداد دسته‌ها را از ٦ تا ۲۵ در نظر می‌گیرند. برای مثال داده‌های

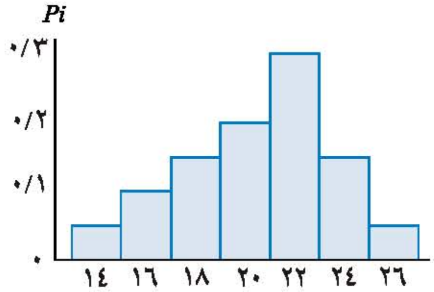
جدول ۱-۲ در نمودار ۱-۳ به ۷ دسته مجزا تقسیم شده و فراوانی و فراوانی نسبی افراد در هر دسته به دست آمده است. در این جدول دامنه هر دسته برابر با ۲ در نظر گرفته شده است. با توجه به اینکه بزرگترین و کوچکترین داده به ترتیب برابر با ۱۳/۳ و ۲۷/۲ می‌باشند، دامنه کلی تغییرات برابر با ۱۳/۹ است. اگر هدف دسته بندی داده‌ها به ۷ دسته برابر باشد دامنه تغییرات به ۷ تقسیم می‌شود که برابر با ۱/۹۸ خواهد بود. چون انجام محاسبات با این رقم مشکل است، معمولاً آن را گرد کرده و برابر با ۲ در نظر می‌گیرند. با وجود ۷ دسته با فواصل مساوی ۲، دامنه کلی تغییرات از ۱۳/۹ به ۱٤ افزایش می‌یابد. لذا دسته اول از ۱۳/۲۵ شروع و دسته آخر تا ۲۷/۲۵ ادامه می‌یابد.

جدول ۱-۳. تقسیم‌بندی داده‌های جدول ۱-۲ در هفت دسته مجزا.

حدود دسته	فراوانی (f_i)	فراوانی نسبی (p_i)	نشان دسته
۱۳/۲۵-۱۵/۲۵	۱	۰/۰۵	۱٤/۲۵
۱۵/۲۵-۱۷/۲۵	۲	۰/۱۰	۱٦/۲۵
۱۷/۲۵-۱۹/۲۵	۳	۰/۱۵	۱۸/۲۵
۱۹/۲۵-۲۱/۲۵	٤	۰/۲۰	۲۰/۲۵
۲۱/۲۵-۲۳/۲۵	٦	۰/۳۰	۲۲/۲۵
۲۳/۲۵-۲۵/۲۵	۳	۰/۱۵	۲٤/۲۵
۲۵/۲۵-۲۷/۲۵	۱	۰/۰۵	۲٦/۲۵
	۲۰	۱	

بافت‌نگار

در بافت‌نگار یا نمودار مستطیلی برای هر دسته یک مستطیل رسم می‌شود که مساحت آن فراوانی نسبی دسته و عرض آن برابر با حدود دسته است بعلاوه مساحت تمام مستطیل‌ها برابر یک واحد است. بافت‌نگار برای داده‌های جدول ۱-۳ در زیر نشان داده شده است.



نمودار ۱-۵. بافت‌نگار مربوط به داده‌های جدول ۱-۳.

نمودار شاخه و برگ

نمودار شاخه و برگ، داده‌ها را به صورتی فشرده طوری نمایش می‌دهد که تمام داده‌ها را می‌توان در آن دید. در این نمودار ارقام هر داده به دو بخش تقسیم می‌شود. یک یا چند رقم اولیه در ستونی در سمت چپ (ستون شاخه) از کوچک به بزرگ و ارقام باقی‌مانده در ستونی دیگر در سمت راست (ستون برگ) نمایش داده می‌شود. نمودار ۱–٦ نمودار شاخه و برگ را برای داده‌های جدول ۱–۲، نشان می‌دهد. در این مورد، هر داده به دو بخش عدد صحیح و عدد اعشاری تقسیم شده، اعداد صحیح در ستون مربوط به شاخه (ستون دوم)، به ترتیب صعودی مرتب شده و در ستون مربوط به برگ (ستون سوم) قسمت اعشاری آنها نمایش داده شده است. اعدادی در شاخه که هیچ رقم اعشاری (حتی صفر) برای آنها در برگ ثبت نشده است، در واقع در مجموعه داده‌ها وجود ندارند. همانگونه که در نمودار زیر دیده می‌شود، برخی نرم‌افزارهای آماری در ستون اول نیز از بالا و پایین به طرف مرکز، فراوانی تجمعی را نشان می‌دهند. به عنوان مثال به پنجمین داده از بالا (۱۸/۲) در نمودار ۱–٦ دقت کنید.

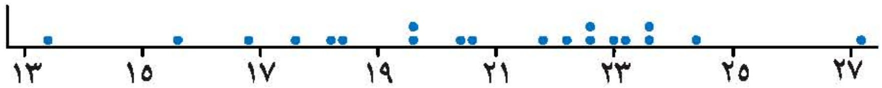
```
Stem-and-leaf of C1  N  = 20
Leaf Unit = 0.10

 1   13  3
 1   14
 2   15  5
 3   16  7
 4   17  5
 6   18  24
 8   19  56
 10  20  45
 10  21  8
 9   22  2559
 5   23  255
 2   24  3
 1   25
 1   26
 1   27  2
```

نمودار ۱–٦. نمودار شاخه و برگ مربوط به داده‌های جدول ۱–۲

نمودار نقطه‌ای

نمودار نقطه‌ای مربوط به داه‌های جدول ۱–۲ در زیر نشان داده شده است. در این نمودار، هر داده به صورت نقطه‌ای در بالای خط افقی نشان داده می‌شود. اگر داده‌ای دو بار تکرار شده باشد، نقاط مربوطه روی یکدیگر قرار می‌گیرند.



نمودار ۱-۷. نمودار نقطه‌ای مربوط به داده‌های جدول ۱-۲

۱-۹ علامت مجموع و حاصلضرب

فرض کنید مجموعه داده‌های $x_1, x_2, x_3 \cdots, x_i, \cdots, x_n$ موجودند. در این مجموعه، x_1 اولین داده، x_2 دومین داده، x_3 سومین داده، x_i، iامین داده و بالاخره x_n، nامین داده یا آخرین داده می‌باشد. در بیشتر محاسبات آماری لازم است که یک مجموعه از داده‌ها با هم جمع شوند. مجموع داده‌های مذکور ($x_1 + x_2 + x_3 + \cdots + x_i + \cdots + x_n$) به صورت

$$x_1 + x_2 + x_3 + \cdots + x_i + \cdots + x_n = \sum_{i=1}^{n} x_i$$

نمایش داده می‌شود، که در آن علامت $\sum$ (سیگما) برای کوتاه‌تر کردن عبارت بالا به کار رفته و بیانگر مجموع xها از اولین تا آخرین داده می‌باشد. به همین ترتیب حاصلضرب داده‌ها یعنی $x_1 \cdot x_2 \cdots\cdots x_N$ با استفاده از علامت Π (پی) خلاصه شده و بصورت زیر نشان داده می‌شود.

$$x_1 \cdot x_2 \cdots\cdots x_N = \prod_{i=1}^{n} x_i$$

در ادامه مثال‌های بیشتری با استفاده از مجموعه داده $x_1 = 3,\ x_2 = 9,\ x_3 = 11$ در مورد کاربرد علامت جمع آورده شده است.

$$\sum_{i=1}^{3} x_i^2 = x_1^2 + x_2^2 + x_3^2 = 3^2 + 9^2 + 11^2 = 9 + 81 + 121 = 211$$

$$\sum_{i=1}^{3} (3x_i + 1) = (3 \times 3 + 1) + (3 \times 9 + 1) + (3 \times 11 + 1) = 3 \times (3 + 9 + 11) + (3 \times 1) = 72$$

در موارد زیر a، b و c اعداد ثابت هستند.

$$\sum_{i=1}^{n} i = 1+2+\cdots+n$$
$$\sum_{i=1}^{n} c = nc$$
$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$
$$\sum_{i=1}^{n} (ax_i+b) = a\sum_{i=1}^{n} x_i + nb$$
$$\sum_{i=1}^{n} (x_i+y_i)^2 = \sum_{i=1}^{n} (x_i^2+y_i^2+2x_iy_i) = \sum x_i^2 + \sum y_i^2 + 2\sum x_iy_i$$
$$\sum_{i=1}^{n} (x_i+a)^2 = \sum_{i=1}^{n} (x_i^2+a^2+2ax_i) = \sum x_i^2 + na^2 + 2a\sum x_i$$

۱-۱۰ شاخص‌های مرکزی

در توصیف مجموعه‌ای از داده‌ها دانستن محل تمرکز آنها اهمیت زیادی دارد. البته ممکن است محل تمرکز داده‌ها یک نقطه نباشد. توابع آماری مثل میانگین‌ها، میانه و نما (مد) از شاخص‌های مرکزی مشهور هستند.

۱-۱۰-۱ میانه

برای یک سری n عددی $x_1, x_2, \cdots, x_n$ که به ترتیب صعودی نوشته شده‌اند، اگر n فرد باشد میانه عبارت است از داده وسطی و اگر n زوج باشد میانه برابر است با میانگین دو داده وسطی.

مثال ۱-۲میانه اعداد ٦، ۷، ۹، ۱۲، ١٦ و ۲۰ را پیدا کنید.

$$m = \frac{9+12}{2} = 10/5$$

مثال ۱-۳ میانه اعداد ٦، ۷، ۹، ۱۲، ١٦، ۲۰، ۲۲ را پیدا کنید. $m = 12$

محاسبه میانه در جدول فراوانی

برای محاسبه میانه لازم است که دسته‌ها از کوچک به بزرگ در جدول فراوانی مرتب شوند. در

این حالت، میانه معمولاً مقدار $n/۲$اُمین داده است که به صورت زیر محاسبه می‌شود.

نحوه محاسبه میانه در جداول فراوانی

$$m = L_m + \frac{(\frac{n}{۲} - g)}{f_m} w$$

L_m، مرز پایین واقعی دسته میانه

g، فراوانی تجمعی دسته ماقبل دسته میانه

f_m، فراوانی دسته‌ای که میانه در آن قرار دارد و

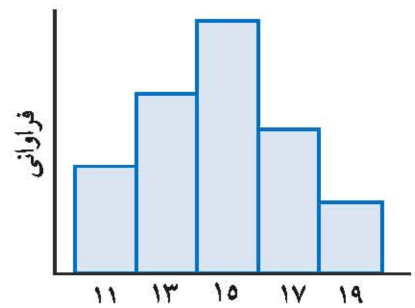
w، فاصله دسته‌ها

مثال ۱-٤. میانه را در جدول فراوانی زیر حساب کنید.

فراوانی تجمعی صعودی	فراوانی	حدود دسته
۳	۳	۱۰-۱۲
۸	۵	۱۲-۱٤
۱۵	۷	۱٤-۱٦
۱۹	٤	۱٦-۱۸
۲۱	۲	۱۸-۲۰
	۲۱	

نقطه $\frac{n}{۲} = ۱۰/۵$ به عنوان عضو مرکزی داده‌ها در نظر گرفته می‌شود. ایـن عضـو در دسـته سـوم قرار گرفته است. از آنجا که حد پایین دسته سوم ۱٤ است، میانه به صورت زیر به دست می‌آید.

$$m = ۱٤ + (\frac{۱۰/۵ - ۸}{۷})۲ = ۱٤/۷۱$$



نمودار ۱-۸ نمودار مستطیلی برای داده‌های مثال ۱-٤.

در مواری که داده‌ها گسسته بوده و به صورت گسسته دسته‌بندی شده‌اند، برای محاسبه صحیح میانه لازم است که دسته‌ها به صورت پیوسته فرض شده و نقطه میانی دو دسته به عنوان مرز واقعی دسته‌ها در نظر گرفته شود.

مثال ۱-۵. میانه را برای مشاهدات زیر به دست آورید.

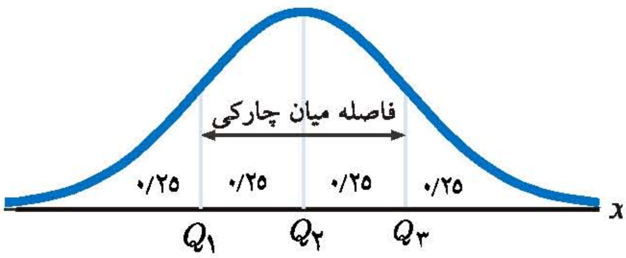
حدود دسته	فراوانی	فراوانی تجمعی صعودی
۹-۱۳	۲	۲
۱۴-۱۸	۹	۱۱
۱۹-۲۳	۱۰	۲۱
۲۴-۲۸	۹	۳۰
	۳۰	

با توجه به اینکه در اینجا دسته‌ها به صورت گسسته هستند باید ابتدا آنها را پیوسته فرض کرد. برای این منظور در هر دسته، نیم واحد از طرف چپ کم کرده و نیم واحد به طرف راست اضافه می‌کنیم. بدین ترتیب فاصله دسته‌ها از ۴ به ۵ افزایش می‌یابد.

$$m = ۱۸/۵ + (\frac{۱۵-۱۱}{۱۰})۵ = ۲۰/۵$$

۱-۱۰-۲ چارک‌ها، دهک‌ها و صدک‌ها

چارک اول بر روی منحنی توزیع داده‌ای است که ۲۵٪ داده‌ها از آن کوچکتر و ۷۵٪ داده‌ها از آن بزرگتر هستند. به همین ترتیب چارک سوم داده‌ای است که ۷۵٪ از افراد جامعه مورد نظر از آن کوچکتر و ۲۵٪ افراد از آن بزرگتر هستند. چارک دوم نیز همان میانه می‌باشد. دهک‌ها و صدک‌ها نیز مفهومی چون چارک داشته و کل جامعه را به ترتیب به ۱۰ و ۱۰۰ قسمت مساوی تقسیم می‌کنند. به عنوان مثال دهک سوم (یا صدک سی‌ام) جامعه، داده‌ای است که ۳۰٪ داده‌ها از آن کوچکتر و ۷۰٪ داده‌ها یا افراد از آن بزرگتر هستند. روش محاسبه چارک‌ها، دهک‌ها و صدک‌ها نیز مانند میانه می‌باشد.



نمودار ۱-۹. نمودار یک توزیع متقارن که در آن مناطق چارکی نشان داده شده است.

مثال ۱-۶. در مورد داده‌های مثال ۱-۵، برای محاسبه چارک اول ابتدا باید فراوانی در $\frac{1\times n}{4}$ ضرب تا دسته چارک اول با توجه به فراوانی‌ها مشخص شود ($\frac{1\times 30}{4}$ = ۷/۵). در این مورد، ۷/۵ امین عدد در دسته دوم قرار دارد. از آنجا که حد پایین واقعی این دسته، ۱۴/۵ است، چارک اول بصورت $Q_۱ = ۱۴/۵ + (\frac{۷/۵-۲}{۹})۵ \; = \; ۱۷/۵۵$ محاسبه می‌شود. به همین ترتیب سایر چارک‌ها، دهک‌ها یا صدک‌ها را می‌توان بصورت زیر محاسبه کرد.

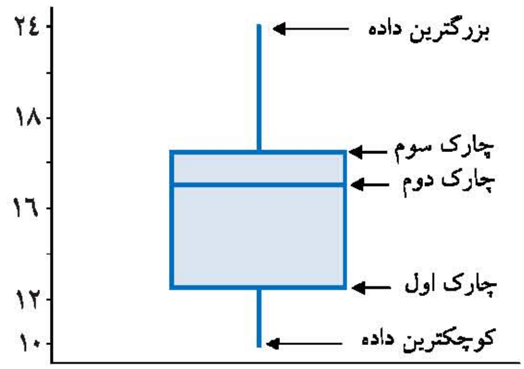
محاسبه چارک سوم : $\frac{۳\times n}{۴} = \frac{۳\times ۳۰}{۴} = ۲۲/۵ \;\Rightarrow\; Q_۳ = ۲۴/۵ + (\frac{۲۲/۵-۲۱}{۹})۵ \; = \; ۲۵/۳۳$

محاسبه چارک دوم : $\frac{۲\times n}{۴} = \frac{۲\times ۳۰}{۴} = ۱۵ \;\Rightarrow\; Q_۲ = ۱۹/۵ + (\frac{۱۵-۱۱}{۱۰})۵ \; = ۲۱/۵$

محاسبه دهک ششم : $\frac{۶\times n}{۱۰} = \frac{۶\times ۳۰}{۱۰} = ۱۸ \;\Rightarrow\; D_۶ = ۱۹/۵ + (\frac{۱۸-۱۱}{۱۰})۵ \; = \; ۲۳$

محاسبه صدک سی‌ام : $\frac{۳۰\times n}{۱۰۰} = \frac{۳۰\times ۳۰}{۱۰۰} = ۹ \;\Rightarrow\; P_۳ = ۱۳/۵ + (\frac{۹-۲}{۹})۵ \; = \; ۱۶/۸۹$

یکی از روش‌های نشان دادن وضعیت پراکندگی داده‌ها استفاده از نمودار جعبه‌ای است. این نمودار چارک‌ها و کوچکترین و بزرگترین داده را نشان می‌دهد. نمودار جعبه‌ای در زیر برای مجموعه‌ای از داده‌ها نشان داده شده است و بیان می‌کند که چارک اول، میانه و چارک سوم در این مجموعه داده به ترتیب ۱۲/۵، ۱۷ و ۱۸/۵ می‌باشد. همچنین نمودار نشان می‌دهد که کوچکترین داده ۱۰ و بزرگترین داده ۲۴ است.



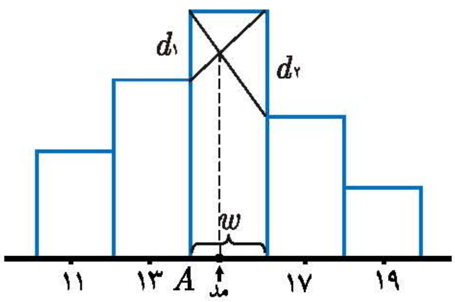
نمودار ۱-۱۰. نمودار جعبه‌ای، کمترین داده، چارک اول، میانه، چارک سوم و بیشترین داده را به ترتیب از پایین به بالا نشان می‌دهد.

۱-۱۰-۳ نما

نما یا مد داده‌ای است که بیشترین فراوانی را دارد. بـرای مثـال در داده‌هـای ۹، ۹، ٦، ۷، ٦، ٦، ۵، ٤، ٤، ۳، ۲ نما برابر با ٦ است. حال به محاسبه نما در جدول فراوانی زیر پرداخته می‌شود.

۱۸–۲۰	۱٦–۱۸	۱٤–۱٦	۱۲–۱٤	۱۰–۱۲	دسته
۲	٤	۷	۵	۳	فراوانی (f_i)

اگر بافت‌نگار توزیع فراوانی در این جدول رسم شود، با رسم دو خط از گوشـه‌هـای بلنـدتـرین مستطیل به گوشه مستطیل‌های مجاور، می‌توان موقعیت مد را بر روی بافت‌نگار به صـورت زیـر مشخص کرد.



نمودار ۱-۱۱. بافت‌نگار توزیع فراوانی داده‌های اخیر که در آن موقعیت نما (مد) نشان داده شده است.

روش بالا یک روش ترسیمی است که موقعیت نما را بر روی بافت‌نگـار نشـان مـی‌دهـد. بـرای محاسبه مقدار عددی نما به صورت زیر عمل می‌شود.

نحوه محاسبه نما در جدول فراوانی

$$M = A + (\frac{d_۱}{d_۱ + d_۲})w$$

A، حد پایین دسته نما

$d_۱$، فراوانی دسته نما منهای فراوانی دسته قبل از آن

$d_۲$، فراوانی دسته نما منهای فراوانی دسته بعد از آن و

w، فاصله دسته‌ها

نما برای داده‌های جدول فراوانی صفحه قبل برابر است با:

$$M = ۱٤ + [\frac{۷ - ۵}{(۷ - ۵) + (۷ - ٤)} \times ۲] = ۱٤ + [\frac{۲}{۲ + ۳} \times ۲] = ۱٤/۸$$

۱-۱۰-٤ میانگین حسابی

میانگین حسابی که از این پس آن را میانگین می‌نامیم، عبارت است از حاصل‌جمع داده‌ها تقسیم بر تعداد آنها. میانگین جامعه‌ای از داده‌ها را با حرف یونانی μ (مو)، هر داده را با x_i و تعداد کل داده‌ها یا مشاهدات را با N نشان می‌دهند. میانگین نمونه نیز با $\bar{x}$ نشان داده می‌شود.

$$\mu = \frac{۱}{N}\sum_{i=۱}^{N} x_i$$

اگر داده‌ها در جدول فراوانی تنظیم شده باشند، میانگین از طریق رابطه $\mu_x = \sum x_i f_i \Big/ \sum f_i$ یا $\mu_x = \sum x_i p_i$ محاسبه می‌شود. مثالی از محاسبه میانگین در جدول زیر آورده شده است.

جدول ۱-٤. داده‌های فرضی و فراوانی آنها.

نسبت یا فراوانی نسبی (p_i)	فراوانی (f_i)	مشاهدات (x_i)
۰/۰۵	۱	۱
۰/۳	٦	۳
۰/۳۵	۷	٤
۰/۲	٤	۲
۰/۱	۲	٦

$$\mu_x = \frac{\sum x_i f_i}{\sum f_i} = \frac{(۱\times۱)+(۳\times۶)+(٤\times۷)+(۲\times٤)+(۶\times۲)}{۱+۶+۷+٤+۲} = ۳/۳۵$$

$$\mu_x = \sum x_i p_i = (۱\times۰/۰۵)+(۳\times۰/۳)+(٤\times۰/۳۵)+(۲\times۰/۲)+(۶\times۰/۱) = ۳/۳۵$$

در مـواردی کـه داده‌هـا در جـدول فراوانـی دسـته بنـدی شـده باشـند، از همـان روابـط $\mu_x = \sum x_i f_i \Big/ \sum f_i$ یا $\mu_x = \sum x_i p(x_i)$ برای محاسبه میانگین استفاده مـی‌شـود کـه در آن x_i، نشان دسته است (جدول ۱-۶، صفحه ۲۷).

۱-۱۰-٤-۲ ویژگی‌های میانگین حسابی

- مجموع انحرافات مجموعه‌ای از مشاهدات از میانگین حسابی همیشه برابر با صفر است.

$$\sum_{i=۱}^{N}(x_i-\mu) = \sum x_i - N\mu = N\mu - N\mu = ۰$$

- اگر کلیه داده‌ها در عددی مانند a ضرب و با عددی مانند b جمع شوند، میانگین نیز بـا همـین ضابطه تغییر می‌کند یعنی رابطه $\mu_{ax+b} = a\mu_x + b$ برقرار است.
- مقدار تک تک داده‌ها روی میانگین مؤثر است، طوری که یک داده پرت (خیلی بزرگ یا خیلی کوچک) می‌تواند مقدار میانگین را خیلی جا به جا کند. این ویژگی یـک نقیصـه بـرای میـانگین محسوب می‌شود به طوری که میانگین در برابر داده‌های پرت، استوار نیست.
- مجموع توان‌دوم انحراف مقادیر از میانگین، کوچکتر از مجموع توان‌دوم انحراف مقادیر از هـر عدد دیگری است. یعنی

$$\sum_{i=۱}^{N}(x_i-\mu)^۲ = \min \sum_{i=۱}^{N}(x_i-a)^۲$$

۱-۱۰-۵ میانگین حسابی وزنی

در برخی از داده ها برای محاسبه میانگین حسابی، به دلیل اینکه مقادیر مشاهده شده ارزش هـای متفاوت دارند لازم است به هر مشاهده وزنی را اختصاص داده و سپس میـانگین داده هـای وزن داده شده را محاسبه نمود. میانگین وزنی از فرمول زیر محاسبه می شود.

$$\overline{x}_w=\frac{\sum w_i x_i}{\sum w_i}=\frac{w_1x_1+w_2x_2+\ldots+w_nx_n}{w_1+w_2+\ldots+w_n}$$

مثال ۱-۷. کارنامه دو داوطلب شرکت کننده در یک آزمون استخدامی به صورت زیر است. میانگین کدامیک بیشتر است؟

	زبان (ضریب ۳)	اطلاعات عمومی (ضریب ۱)	پرسش های تخصصی (ضریب ۳)
داوطلب اول	۱۹	۱٤	۱۷
داوطلب دوم	۱۷	۲۰	۱٦

میانگین اولی $\overline{x}_w=\frac{\sum w_i x_i}{\sum w_i}=\frac{۳(۱۹)+۱(۱٤)+۳(۱۷)}{۷}=۱۷/٤۳$

میانگین دومی $\overline{x}_w=\frac{۳(۱۷)+۱(۲۰)+۳(۱٦)}{۷}=۱۷$

۱-۱۰-٦ میانگین پیراسته

برای محاسبه میانگین پیراسته تمام داده های بزرگتر از چارک سوم و کوچکتر از چارک اول را کنار گذاشته و سپس میانگین حسابی بقیه محاسبه می‌شود.

۱-۱۰-۷ میانگین هندسی

میانگین هندسی (g) برای N عدد غیر منفی $(x_1\cdot x_2\cdot\cdots\cdot x_N)$ برابر با

$$g=\sqrt[n]{\prod_{i=1}^{n}x_i}=\sqrt[N]{x_1\cdot x_2\cdot\cdots\cdot x_N}=(x_1\cdot x_2\cdot\cdots\cdot x_N)^{\frac{1}{N}}$$

است. در صورتی که N بزرگ باشد، گرفتن ریشه Nام از آنها سخت است، در این موارد می‌توان لگاریتم داده‌ها را در مبنای ۱۰ بدست آورده و میانگین حسابی آن‌ها را محاسبه کرد. در این صورت، میانگین هندسی برابر با عدد ۱۰ به توان میانگین حسابی محاسبه شده خواهد بود.

مثال ۱-۸. میانگین هندسی اعداد ۳، ٤ و ۵ را به دست آورید.

$$\log g = \frac{۱}{۳}(\log ۳ + \log ٤ + \log ٥) = \frac{۱}{۳}(۰/٤۷۷ + ۰/٦۰۲ + ۰/٦۹۹) = ۰/۵۹۲$$
$$g = anti\log ۰/۵۹۲ = ۱۰^{۰/۵۹۲} = ۳/۹۱$$

میانگین هندسی برای داده‌هایی که به صورت نسبت یا درصد هستند مورد استفاده قرار می‌گیرد.

مثال ۱-۹. قیمت کالای A در تیرماه دو برابر و در مردادماه ۳ برابر کالای B بوده است. میانگین نسبت کالای A به B را در این دو ماه حساب کنید.

$$g = \sqrt{۲ \times ۳} = ۲/٤٥$$

مثال ۱-۱۰. میانگین نسبت کالای B به A در این دو ماه در مثال بالا چقدر بوده است؟

$$g = \sqrt{\frac{۱}{۲} \times \frac{۱}{۳}} = \frac{۱}{\sqrt{٦}}$$

مثال ۱-۱۱. میزان خسارت ناشی از یک آفت در ٤ بوته مختلف ۲٥، ۳۰، ۲٤ و ٤۰ درصد بوده است. متوسط خسارت بر حسب درصد چقدر است؟

$$g = \sqrt[٤]{۲٥ \times ۳۰ \times ۲٤ \times ٤۰} = ۲۹/۱۳$$

میانگین هندسی در جدول فراوانی با k دسته بصورت زیر محاسبه می‌شود که در آن N تعداد داده‌ها و x_i و f_i به ترتیب نشان و فراوانی دسته iام می‌باشند.

$$g = \sqrt[N]{\prod x_i^{f_i}} = \sqrt[N]{x_۱^{f_۱} x_۲^{f_۲} \ldots x_k^{f_k}}$$

۱-۱۰-۸ میانگین همساز

میانگین همساز (h) اغلب برای اندازه‌گیری متوسط سرعت‌ها وقتی که فواصل مربوط به آنها مساوی باشد به کار می‌رود و بصورت زیر محاسبه می‌شود.

$$h = \frac{N}{\sum \frac{۱}{x_i}} = \frac{N}{\frac{۱}{x_۱} + \frac{۱}{x_۲} + \cdots + \frac{۱}{x_N}}$$

مثال ۱-۱۲. یک موتورسوار یک سوم فاصله‌ای را با سرعت ۳۰ کیلومتر در ساعت، یک سوم دیگر فاصله را با سرعت ۴۰ کیلومتر در ساعت و یک سوم آخر را با سرعت ۶۰ کیلومتر در ساعت پیموده است. متوسط سرعت او در این مسیر چقدر بوده است؟

$$h = \frac{۳}{\frac{۱}{۳۰} + \frac{۱}{۴۰} + \frac{۱}{۶۰}} = ۴۰$$

مثال ۱-۱۳. شخصی طی چهار سال متوالی کالایی را به قیمت کیلویی ۱۶، ۱۸، ۲۱ و ۳۵ ریال خریداری نموده است. متوسط قیمت پرداختی برای این کالا را در دو حالت زیر محاسبه کنید: الف) مقدار خرید در هر سال ۱۰ کیلوگرم بوده است. ب) میزان پول پرداختی در هر سال ۲۰۰۰ ریال بوده است.

الف) توجه داشته باشید که در حالت اول میانگین به میزان خرید بستگی ندارد، بنابراین از میانگین حسابی برای محاسبه مقدار متوسط استفاده می‌شود.

$$\mu = \frac{\text{قیمت کل}}{\text{مقدار کل}} = \frac{۱۶۰ + ۱۸۰ + ۲۱۰ + ۳۵۰}{۴۰} = ۲۲/۵$$

ب) در صورت پرداخت یک مقدار ثابت پول در سال، هر چه کالا گران‌تر شود مقدار خرید کمتر خواهد شد. بنابراین از میانگین همساز برای محاسبه مقدار متوسط استفاده می‌شود.

$$h = \frac{۴}{\frac{۱}{۱۶} + \frac{۱}{۱۸} + \frac{۱}{۲۱} + \frac{۱}{۳۵}} = ۲۰/۵۹$$

در کل برای یک گروه از اعداد مثبت رابطه $h \leq g \leq \mu$ برقرار است. البته در صورتی که اعداد با هم برابر باشند، هر سه نوع میانگین، برابر خواهند بود.

۱-۱۱ مقایسه میانگین، میانه و نما

میانگین حسابی در واقع نقطه ثقل داده‌ها است و متداول‌ترین شاخص مرکزی می‌باشد. میانگین حسابی دارای خواص احتمالی خوبی است بدین معنی که می‌توان میانگین جامعه را با اطمینان

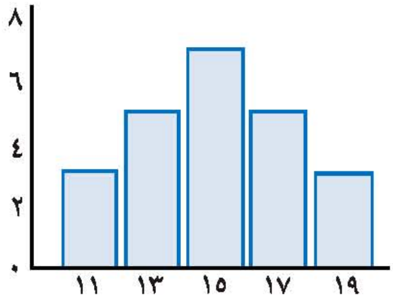
بالایی به کمک میانگین نمونه برآورد کرد. اگر از جامعه‌ای که دنباله‌های توزیع آن خیلی به سمت چپ یا راست کشیده نیست، نمونه‌های مختلف انتخاب شوند، میانگین حسابی نمونه‌ها به میانگین جامعه نزدیکتر خواهند بود تا میانه آنها. میانه به اندازه میانگین تحت تأثیر داده‌های پرت قرار نمی‌گیرد و این موضوع به ویژه در مورد توزیع‌هایی که دنباله‌های کشیده دارند، اهمیت زیادی دارد. میانگین برای متغیرهای ترتیبی و اسمی کارایی ندارد. برای توصیف متغیرهای ترتیبی، میانه و نما و برای توصیف متغیرهای اسمی، تنها نما کاربرد دارد. برای توصیف مرکزیت داده‌های کمی، میانگین و میانه از اهمیت بیشتری برخوردارند. یک مجموعه داده ممکن است نما نداشته باشد یا بیش از یک نما داشته باشد. در توزیع‌های متقارن میانگین، میانه و نما با هم برابرند، ولی در توزیع‌های چوله (نامتقارن) چنین نیست. در مثال‌های زیر رابطه شاخص‌های مرکزی بیشتر مورد بررسی قرار می‌گیرند.

۱-۱۲ چولگی (عدم قرینگی)

داده‌های ۲۲جدول ۱-۵ را در نظر بگیرید. نمودار ۱-۱۲ که مربوط به داده‌های این جدول است، نشان می‌دهد که شکل توزیع داده‌ها متقارن و تا حدودی زنگوله‌ای شکل است. همانگونه که مقادیر شاخص‌های مرکزی نشان می‌دهد، در چنین توزیع‌های متقارنی، میانگین، میانه و نما با هم برابر هستند.

جدول ۱-۵. داده‌های دسته‌بندی شده و فراوانی آنها.

حدود دسته	۱۰-۱۲	۱۲-۱۴	۱۴-۱۶	۱۶-۱۸	۱۸-۲۰
فراوانی (f_i)	۳	۵	۷	۵	۳



نمودار ۱-۱۲. نمودار داده‌های جدول ۱-۵ که نشان دهنده متقارن بودن شکل توزیع است.

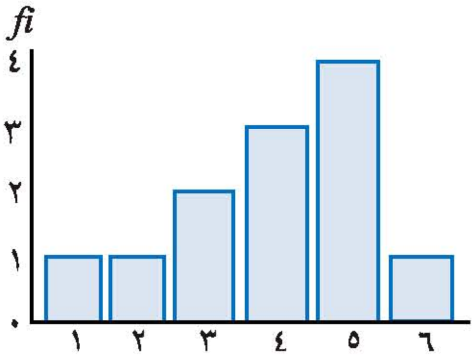
میانگین، میانه و نما برای این توزیع متقارن بصورت زیر محاسبه می‌شود.

$$\mu = \frac{۱۱(۳) + ۱۳(٥) + ۱٥(۷) + ۱۷(٥) + ۱۹(۳)}{۲۳} = ۱٥$$

$$m = ۱٤ + (\frac{۱۱/٥ - ۸}{۷})۲ = ۱٥$$

$$M = ۱٤ + (\frac{۲}{۲ + ۲})۲ = ۱٥$$

مثال ۱-۱٤. تعداد تصادفات رانندگی در ۱۲ شهر در طول یک ماه ۳، ٤، ٥، ۲، ۳، ٥، ٤، ٥، ۱، ٤، ٦، ٥ بوده است. نمودار میله‌ای داده‌ها را رسم کرده و میانگین، میانه و نما را برای داده‌ها به دست آورید. چه ارتباطی بین میانگین، میانه و نما با شکل توزیع وجود دارد؟



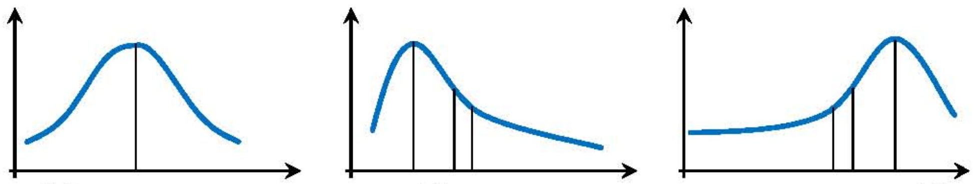
نمودار ۱-۱۳. نمودار توزیع فراوانی داده‌های مثال ۱-۱٤ که نشاندهنده چولگی (عدم تقارن) توزیع به سمت چپ است.

$$\mu = \frac{\sum x_i}{N} = \frac{٤۷}{۱۲} = ۳/۹۱$$

$$m = ۳/٥ + (\frac{٦ - ٤}{۳})۱ = ٤/۱٦$$

$$M = ٤/٥ + (\frac{۱}{۱ + ۳})۱ = ٤/۷٥$$

ملاحظه می‌شود در مورد یک توزیع چوله به چپ میانگین کوچکتر از میانه و نما بوده و نما بزرگترین است. به طور کلی همانگونه که در نمودار زیر مشاهده می‌شود، در توزیع‌های دارای چولگی مثبت، رابطه $M < m < \mu$ و در توزیع‌های دارای چولگی منفی رابطه $\mu < m < M$ وجود دارد.



$\mu < m < M$	$M < m < \mu$	$M = m = \mu$
توزیع چوله به چپ	توزیع چوله به راست	توزیع متقارن
(چولگی منفی)	(چولگی مثبت)	(چولگی صفر)

اگر n داده $x_1, x_2, x_3 \cdots, x_i, \cdots, x_n$ با فراوانی های $f_1, f_2, f_3 \cdots, f_i, \cdots, f_n$ موجود باشد، میانگین توان rام $x_i - \mu$ را گشتاور مرتبه rام پیرامون μ نامیده و صورت زیر نشان می‌دهند.

$$\mu_r = \frac{\sum (x_i - \mu)^r}{N}$$

واضح است که گشتاور مرتبه دوم پیرامون μ همان واریانس جامعه است ($\mu_2 = \sigma^2$). پارامتر $\frac{\mu_3}{\sigma^3}$ که ضریب چولگی گشتاوری نام دارد، برابر با

$$SK = \frac{\mu_3}{\sigma^3} = \frac{\frac{\sum (x_i - \mu)^3}{N}}{\left(\sqrt{\frac{\sum (x_i - \mu)^3}{N}}\right)^3}$$

بوده و به عنوان مهمترین شاخص چولگی (عدم تقارن) مورد استفاده قرار می‌گیرد. از دیگر پارامترهایی که به عنوان شاخص چولگی به کار می‌روند می‌توان به ضریب چولگی اول پیرسون، $(\mu\text{-}M)/\sigma$ و ضریب چولگی دوم پیرسون، $[3(\mu\text{-}m)]/\sigma$ اشاره کرد که در آن M و m به ترتیب مد و میانه هستند. در صورتیکه داده‌ها نسبت به میانگین متقارن باشند، همه ضرایب چولگی برابر صفر خواهند بود.

گرچه روابط مختلفی برای محاسبه مقدار چولگی وجود دارد ولی در نرم‌افزارهای آماری از جمله $Minitab$، $Excel$، SAS و $SPSS$ برای برآورد چولگی از ضریب چولگی گشتاوری، $\frac{\mu_3}{\sigma^3}$ استفاده شده و مقدار آن از رابطه زیر با استفاده نمونه برآورد می‌شود.

$$SK = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{s}\right)^3$$

۱-۱۳ برجستگی

میزان برجستگی یا تیزی منحنی در نقطه ماکزیمم منحنی فراوانی را نسبت به منحنی نرمال استاندارد، برجستگی می نامند. فرض کنید μ_4 گشتاور مرکزی چهارم و σ انحراف استاندارد باشد. در این صورت معیار برجستگی را از رابطهٔ $K = \frac{\mu_4}{\sigma^4} - 3$ محاسبه می‌شود. مقدار K برای منحنی نرمال نزدیک صفر است. بر حسب آنکه K مثبت یا منفی باشد منحنی فراوانی نسبت به منحنی نرمال استاندارد مرتفع‌تر یا مسطح‌تر می‌باشد. همانگونه که در مورد چولگی گفته شد، رابطه‌ای که برای برآورد ضریب کشیدگی در اغلب نرم‌افزارهای آماری از جمله $Minitab$، $Excel$، SAS و $SPSS$ برای برآورد برجستگی در نمونه مورد استفاده قرار می‌گیرد، مقداری با رابطه $\frac{\mu_4}{\sigma^4} - 3$ متفاوت است.

۱-۱۴ شاخص‌های پراکندگی

شاخص‌های پراکندگی میزان پراکندگی داده‌ها را نشان می‌دهند. دو مجموعه داده ممکن است میانگین یکسانی داشته باشند، ولی پراکندگی آنها با هم متفاوت باشد به عنوان مثال میانگین مجموعه داده ۲، ۳ و ٤ برابر با میانگین ۱، ۳ و ۵ است، ولی میزان پراکندگی داده‌های مجموعه دوم بیشتر است.

۱-۱٤-۱ دامنه

ساده‌ترین معیار پراکندگی دامنه (R) است که تفاوت بین بزرگترین داده و کوچکترین داده را نشان می‌دهد. دامنه اغلب معیار خوبی برای پراکندگی نیست. زیرا مفهوم واقعی پراکندگی به فاصله کلیه داده‌ها از هم بستگی دارد، در حالی که فقط بزرگترین و کوچکترین داده در محاسبه دامنه مورد استفاده قرار می‌گیرند. برای مثال دو مجموعه متفاوت از داده‌ها را در نظر بگیرید. با فرض اینکه مجموعه اول داده‌های ۱۰، ۰، ۸، ۱۰ و ۹ و مجموعه دوم متشکل از داده‌های ۰، ۵، ۷، ۳ و ۱۰ باشند، دامنه برای هر دو مجموعه برابر با ۱۰ خواهد بود.

۱-۱٤-۲ میانگین انحرافات

انتظار می‌رود میانگین فاصله مشاهدات از میانگین خود، شاخص خوبی برای بیان پراکندگی باشد، اما با توجه به اینکه مجموع انحرافات یک سری داده از میانگین آنها یعنی $\sum(x_i - \mu)$ همیشه برابر صفر است، در نتیجه مقدار $\sum(x_i - \mu)$ در اندازه‌گیری پراکندگی فاقد ارزش است. یک راه حل، استفاده از مجموع قدر مطلق انحرافات یعنی $\sum|x_i - \mu|$ است. اگر مجموع قدر مطلق انحرافات بر تعداد داده‌ها تقسیم شود، میانگین انحرافات از میانگین (MD) به دست می‌آید که می‌تواند به عنوان شاخصی در اندازه‌گیری پراکندگی داده‌ها به کار رود. به عنوان مثال MD برای دو مجموعه از داده‌های مذکور در بند قبل به صورت زیر محاسبه می‌شود.

$$MD_۱ = \frac{\sum|x_i - \mu|}{N} = \frac{|۲/٦| + |-۷/٤| + |۰/٦| + |۲/٦| + |۱/٦|}{۵} = ۲/۹٦$$

$$MD_۳ = ۲/۸$$

۱-۱٤-۳ واریانس

می‌توان به جای قدر مطلق انحرافات از میانگین، از توان‌دوم آنها استفاده نمود. میانگین توان‌دوم انحرافات، شاخص پراکندگی دیگری است که واریانس نام دارد.

واریانس (میانگین توان‌دوم انحرافات از میانگین) یک مجموعه داده به صورت زیر محاسبه می‌شود:

$$\sigma_x^۲ = \frac{\sum(x_i - \mu)^۲}{N} = \frac{\sum x_i^۲ - \frac{(\sum x_i)^۲}{N}}{N}$$

محاسبه واریانس با استفاده از اعداد خام

به عنوان مثال واریانس اعداد ۳، ۵، ٦، ۳ و ۳ به صورت زیر به دست می‌آید.

$$\sigma^{۲}=\frac{(۳-٤)^{۲}+(٥-٤)^{۲}+(٦-٤)^{۲}+(۳-٤)^{۲}+(۳-٤)^{۲}}{٥}$$
$$=\frac{(۹+۲٥+۳٦+۹+۹)-\frac{(۳+٥+٦+۳+۳)^{۲}}{٥}}{٥}=۱/٦$$

محاسبه واریانس با استفاده از جدول فراوانی

اگر در جدول فراوانی، میان‌دسته‌ها $x_۱ \cdot x_۲ \cdots\cdots x_N$ و فراوانی متناظر با آنها $f_۱ \cdot f_۲ \cdots\cdots f_N$ باشد، واریانس به صورت زیر به دست می‌آید.

$$\sigma_x^{۲}=\frac{\sum(x_i-\mu)^{۲}f_i}{N}=\frac{\sum x_i^{۲}f_i-\frac{(\sum x_i f_i)^{۲}}{N}}{N}$$

مثال ۱-۱٥. در جدول زیر طول بلال در ۲۰ بوته ذرت به سانتی‌متر داده شده است. میانگین و واریانس را با استفاده از داده‌های اصلی و جدول فراوانی به دست آورده و با هم مقایسه کنید.

۲۳/٥	۱۹/٦	۱۸/٤	۱۸/۲	۲۲/٥	۱۳/۳	۲۳/٥	۲۰/٥	۲۳/۲	۱۹/٥
۲۷/۲	۲۲/٥	۲۲/۹	۲۱/۸	۲۰/٤	۱۷/٥	۲٤/۳	۲۲/۲	۱٥/٥	۱٦/۷

جدول ۱-٦. تقسیم داده‌های جدول بالا در هفت دسته مجزا.

فاصله دسته	فراوانی (f_i)	نسبت یا فراوانی نسبی (p_i)	نشان دسته
۱۳/۲٥-۱٥/۲٥	۱	۰/۰٥	۱٤/۲٥
۱٥/۲٥-۱۷/۲٥	۲	۰/۱۰	۱٦/۲٥
۱۷/۲٥-۱۹/۲٥	۳	۰/۱٥	۱۸/۲٥
۱۹/۲٥-۲۱/۲٥	٤	۰/۲۰	۲۰/۲٥
۲۱/۲٥-۲۳/۲٥	٦	۰/۳۰	۲۲/۲٥
۲۳/۲٥-۲٥/۲٥	۳	۰/۱٥	۲٤/۲٥
۲٥/۲٥-۲۷/۲٥	۱	۰/۰٥	۲٦/۲٥
	۲۰	۱	

در ابتدا میانگین با استفاده از جدول توزیع فراوانی محاسبه می‌شود.

$$\mu=\frac{\sum x_i f_i}{\sum f_i}=\frac{(۱۴/۲۵\times ۱)+(۱۶/۲۵\times ۲)+\cdots+(۲۶/۲۵\times ۱)}{۱+۲+۳+۴+۶+۳+۱}=۲۰/۷۵\ \ cm$$

سپس واریانس با استفاده از جدول توزیع فراوانی بصورت زیر بدست می‌آید.

$$\begin{aligned}\sigma^۲&=\frac{[(۱۴/۲۵-۲۰/۷۵)^۲ ۱]+[(۱۶/۲۵-۲۰/۷۵)^۲ ۲]+\cdots+[(۲۶/۲۵-۲۰/۷۵)^۲ ۱]}{۲۰}\\&=۹/۱۵\ \ cm^۲\end{aligned}$$

اگر واریانس به دست آمده با استفاده از جدول فراوانی با واریانس به دست آمده از داده‌های خام مقایسه شود، مشاهده می‌شود که این دو کاملاً برابر نیستند. واضح است که واریانس حاصل از داده‌های خام دقیق‌تر است زیرا در جدول فراوانی، داده‌ها دسته‌بندی شده و در هر دسته یک داده به عنوان نماینده در نظر گرفته می‌شود. لذا واریانس به دست آمده در این حالت (۹/۱۵) اندکی از مقدار واقعی (۱۰/۷۱) انحراف نشان داده است.

$$\mu=\frac{\sum x_i}{N}=\frac{۱۹/۵+\cdots+۲۷/۲}{۲۰}=۲۰/۶۶\ \ cm$$

$$\sigma^۲=\frac{(۱۹/۵-۲۰/۶۶)^۲+\cdots+(۲۷/۲-۲۰/۶۶)^۲}{۲۰}=۱۰/۷۱\ \ cm^۲$$

۱-۱۴-۳-۲ واریانس نمونه

روابطی که تا اینجا ذکر شد برای محاسبه واریانس جامعه بود. غالباً کل داده‌های جامعه در اختیار نیست و هدف، برآورد واریانس جامعه از طریق نمونه‌ای n عضوی است. در این حالت واریانس باید به گونه‌ای محاسبه شود که تقریباً برابر با واریانس جامعه باشد.

محاسبه واریانس جامعه و برآورد نمونه‌ای آن

واریانس جامعه:

$$\sigma_x^2 = \frac{\sum (x_i - \mu_x)^2}{N}$$

برآورد واریانس جامعه هنگامی که فقط نمونه‌ای از جامعه موجود است:

$$s_x^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

ثابت می‌شود که این رابطه، واریانس جامعه را به طور نااُریب برآورد می‌کند، به این معنی که اگر همه نمونه‌های n عضوی ممکن از جامعه‌ای استخراج شود، میانگین واریانس نمونه‌ها برابر با واریانس جامعه خواهد بود.

مثال ۱-۱۶. نمونه‌ای از داده‌های ۰، ۱، ۲، ۳ و ٤ به ترتیب با فراوانی‌های ۱۸، ۲۳، ٦، ۲ و ۱ تشکیل شده است. واریانس این نمونه را محاسبه کنید.

$$s_x^2 = \frac{\sum (x_i - \overline{x})^2 f_i}{n-1} = \frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}}{n-1} = \frac{81 - \frac{45^2}{50}}{49} = \frac{81 - 40.5}{49} = 0.82$$

۱-۱٤-۳-۳ ویژگی‌های واریانس

اگر همه داده‌ها در عددی غیر صفر مانند a ضرب و با عددی مانند b جمع شوند، واریانس داده‌های جدید a^2 برابر خواهد شد. لذا اگر مقدار ثابتی به همه داده‌ها افزوده شود واریانس بدون تغییر باقی خواهد ماند، یعنی رابطه $\sigma_{ax+b}^2 = a^2\sigma_x^2$ برقرار است.

مثال ۱-۱۷. اگر $x_1 = 3$، $x_2 = 5$ و $x_3 = 7$ و همچنین $y_i = 2x_i + 3$ باشد، واریانس x و y را حساب کرده و روابط آنها را بدست آورید.

$$y_۱ = ۹,\ \ y_۲ = ۱۳,\ \ y_۳ = ۱۷$$

$$\sigma_x^۲ = \frac{\sum (x_i - \mu_x)^۲}{N} = ۲/۶۶$$

$$\sigma_y^۲ \ = \frac{\sum (x_i - \mu_y)^۲}{N} = ۱۰/۶۴$$

$$\sigma_y^۲ = ۲^۲ \times \sigma_x^۲ = ۴ \times ۲/۶۶ = ۱۰/۶۴$$

مثال ۱–۱۸. واریانس جامعه داده‌های زیر را به دست آورید.

دسته	۱۵–۲۵	۲۵–۳۵	۳۵–۴۵	۴۵–۵۵	۵۵–۶۵
فراوانی	۳	۵	۲	۵	۵

در ابتدا نشان هر دسته را به عنوان نماینده آن دسته در نظر می‌گیریم:

$$\sigma_x^۲ = \frac{\sum x_i^۲ f_i - \frac{(\sum x_i f_i)^۲}{N}}{N} = \frac{(۲۰^۲ \times ۳) + \cdots + (۶۰^۲ \times ۵) - \frac{[(۲۰ \times ۳) + \cdots + (۶۰ \times ۵)]^۲}{۲۰}}{۲۰} = ۲۰۶$$

یک روش ساده کردن محاسبات این است که با یک تبدیل خطی، مقادیر نشان دسته به اعداد ساده‌تری تبدیل و میانگین و واریانس آنها محاسبه شوند. سپس میانگین و واریانس متغیر اصلی به کمک روابط موجود به دست آیند. اگر نشان دسته‌ها در مثال ۱–۱۸ بر ۱۰ تقسیم شوند، محاسبه واریانس داده‌های حاصل ساده‌تر است.

x_i	۲	۳	۴	۵	۶
فراوانی	۳	۵	۲	۵	۵

$$\sigma_x^۲ = \frac{(۲^۲ \times ۳) + \cdots + (۶^۲ \times ۵) - \frac{[(۲ \times ۳) + (۳ \times ۵) + \cdots + (۶ \times ۵)]^۲}{۲۰}}{۲۰} = ۲/۰۶$$

اگر واریانس حاصل (۲/۰۶) در $۱۰^۲$ ضرب شود واریانس داده‌های اصلی (۲۰۶) به دست می‌آید. روش دیگر ساده کردن این است که ابتدا میانگین از همه داده‌ها کم شده و سپس واریانس داده‌های جدید به دست آیند.

مثال ۱-۱۹. ارتفاع بوته‌ها در جامعه‌ای متشکل از شش بوته، ۲، ۳، ۹، ۵ و ٦ سانتی‌متر بوده است. واریانس را به دست آورید.

اگر میانگین (یعنی ۵) از هر کدام از این داده‌ها کم شود، اعداد ۳-، ۲-، ٤، ۰ و ۱ به دست خواهند آمد که میانگین آنها برابر با صفر است. واریانس این داده‌ها برابر با واریانس داده‌های اولیه خواهد بود.

$$\sigma^{۲}_{x'} = \frac{\sum(x'_i-\mu)^{۲}}{N} = \frac{\sum(x'_i-۰)^{۲}}{N} = \frac{\sum x'^{۲}_i}{N} = \frac{۹+٤+۱٦+۰+۱}{۵} = ٦ \ cm^{۲}$$

اگر مجموع داده‌ها صفر باشد، میانگین نیز برابر با صفر شده و واریانس برابر با مجموع توان‌دوم داده‌ها تقسیم بر N است.

۱-۱٤-٤ انحراف معیار

واحد واریانس توان‌دوم واحد اصلی داده‌هاست در حالی که اغلب بیان واحد اندازه‌گیری در همان واحد اصلی داده‌ها مورد نظر است. انحراف معیار، چنین معیاری از پراکندگی است که با جذر گرفتن از واریانس به دست می‌آید. انحراف معیار در مورد مثال ۱-۱۹ برابر است با:

$$\sigma = \sqrt{\sigma^{۲}} = \sqrt{٦} = ۲/٤۵ \ cm$$

در اینجا انحراف معیار با ارائه یک مثال دیگر بیشتر بررسی می‌شود. فرض کنید ۳۰ دانش‌آموز در مسابقه‌ای، مسافتی را در فواصل زمانی مختلف پیموده و نتایج بر حسب دقیقه مطابق جدول ۱-۷ بوده است.

جدول ۱-۷. فاصله‌های زمانی (بر حسب دقیقه) که طی آن ۳۰ دانش‌آموز یک مسافت مشخص را طی کردند.

۳/٦۵	۳/۳۷	۲/۰۲	٤/٤٤	٦/۰٦	۱/۱۵	۳/۲۰	٤/۰۲	۰/٦۰	۱/۹۷
۳/۱٦	۷/٦۰	٤/۵۵	۵/۲۱	۵/٦۳	۸/۲۹	۹/۷۰	۳/۸۱	۲/۷۵	۱/۷٤
۱/٦۵	۲/۰٦	۵/۱۵	۱/۹۳	٤/۲۵	۲/٤۷	۱/۷۱	۱/۰٦	۵/۳٦	۳/۷۷

در این مجموعه داده چه تعداد از داده‌ها در فاصله یک انحراف معیار از میانگین قرار دارند؟ چه

تعداد از داده‌ها در فاصله دو انحراف معیار از میانگین قرار دارند؟ پس از یک محاسبه ساده $\overline{x} = ۳/۷۴۴$ و $s = ۲/۱۹۸$ به دست می‌آید که در خروجی نرم‌افزار Minitab نیز نشان داده شده است.

جدول ۱–۸ محاسبه میانگین و انحراف معیار با نرم‌افزار Minitab.

مراحل:

۱. داده‌ها را در ستون اول صفحه داده‌ها وارد کنید و وارد مسیر زیر شوید:

Stat > Basic Statistics > Display Descriptive Statistics...

۳. در پنجره مربوطه در قسمت Variables ستون اول یا C1 را انتخاب کنید (همچنین می‌توانید بر روی ...Statistics کلیک کرده و در پنجره حاصل آماره‌های مورد نظر را برای محاسبه انتخاب نمایید).

۴. گزینه OK را انتخاب نمایید.

Variable	N*	Mean	SE Mean	StDev	Minimum	Median	Maximum
C1	0	3.744	0.401	2.198	0.600	3.510	9.700

در این مثال:

$$\overline{x} + s = ۳/۷٤ + ۲/۲۰ = ۵/۹٤ \qquad \overline{x} - s = ۳/۷٤ - ۲/۲۰ = ۱/۵٤$$

$$\overline{x} + ۲s = ۳/۷٤ + ٤/٤۰ = ۸/۱٤ \qquad \overline{x} - ۲s = ۳/۷٤ - ٤/٤۰ = -۰/٦٦$$

با بررسی داده‌ها متوجه می‌شویم که ۲۳ داده از ۳۰ داده (تقریباً ٪۷۷ داده‌ها) در فاصله $\overline{x} - s$ تا $\overline{x} + s$ یعنی ۱/۵٤ تا ۵/۹٤ و ۲۸ داده (تقریباً ٪۹۳ آنها) در فاصله $\overline{x} - ۲s$ تا $\overline{x} + ۲s$ یعنی ۰/٦٦- تا ۸/۱٤ قرار گرفته‌اند. غالب داده‌های مربوط به پدیده‌های طبیعی دارای چنین ویژگی بوده و حداقل ٪۷۵ آنها در فاصله دو انحراف معیار از میانگین ($\overline{x} - ۲s$ تا $\overline{x} + ۲s$) قرار می‌گیرند. این ویژگی برای اولین بار توسط چِبیشِف ریاضی‌دان روسی به صورت زیر بیان شد.

قاعده چبیشف

در یک مجموعه N تایی از مشاهدات حداقل $(1-\frac{1}{K^2})$ درصد از مشاهدات در دامنه K انحراف معیار از میانگین قرار می‌گیرند. K عددی مثبت است که مقادیر بزرگتر از یک را شامل می‌شود. به عنوان مثال اگر K برابر با ۲ باشد، حداقل $\frac{3}{4}$ یعنی ٪۷۵ داده‌ها در فاصله دو انحراف معیار از میانگین ($\overline{x}-2s$ تا $\overline{x}+2s$)، و اگر K برابر با ۳ باشد، حداقل $\frac{8}{9}$ یعنی ٪۸۹ داده‌ها در فاصله $\overline{x}-3s$ تا $\overline{x}+3s$ قرار دارند.

قاعده دومی نیز در این مورد وجود دارد که به قاعده تجربی معروف است. این قاعده نیز همانند قاعده چبیشف یک حساب سرانگشتی از پراکندگی داده‌ها نسبت به میانگین ارائه می‌دهد. قاعده تجربی که به صورت زیر تعریف می‌شود، غالباً در مورد توزیع‌های متقارن صدق می‌کند.

قاعده تجربی

در صورتی که توزیع داده‌ها متقارن باشد (میانگین، میانه و نما تقریباً برابر باشند) تقریباً ٪۶۸ داده‌ها در فاصله یک انحراف معیار از میانگین قرار دارند. همچنین تقریباً ٪۹۵ داده‌ها در فاصله دو انحراف معیار از میانگین و ٪۹۹/۷ داده‌ها (تقریباً تمام آنها) در فاصله سه انحراف معیار از میانگین قرار دارند.

۱-۱٤-۵ ضریب تغییرات

ضریب تغییرات از تقسیم انحراف معیار بر میانگین به دست آمده و معمولاً بر حسب درصد بیان می‌شود. در واقع ضریب تغییرات پراکندگی داده‌ها را نسبت به میانگین نشان می‌دهد.

ضریب تغییرات جامعه: $$CV=\frac{\sigma}{\mu}\times 100\%$$

ضریب تغییرات نمونه: $$CV=\frac{s}{\overline{x}}\times 100\%$$

برای مثال دو نمونه: الف) ۲/۱، ۳/۵، ٤/۷، ۵/۲ و ٦/٤ و ب) ۱۰۲/۱، ۱۰۳/۵، ۱۰٤/۷، ۱۰۵/۲ و

۱۰٦/٤ را در نظر بگیرید. این دو مجموعه پراکندگی یکسانی دارند و انحراف معیار در هر دو برابر با ۱/٦٤٥ است. با این حال پراکندگی در مجموعه "الف" نسبت به میانگین آن (٤/۳۸) خیلی بیشتر بوده و ضریب تغییرات در آن برابر با ۳۷/٥٦٪ است. این در حالی است که داده‌های مجموعه دوم نسبت به میانگین خود خیلی متغیر نبوده و ضریب تغییرات آن فقط ۱/٥۸٪ است.

اگر متغیر تعداد دانه در خوشه گندم دارای انحراف معیار ۱/٤ و میانگین ٥۰ و متغیر وزن دانه بر حسب گرم دارای انحراف معیار ۰/۱٤ و میانگین ٥ باشد، مقایسه پراکندگی تعداد دانه و وزن دانه از طریق مقایسه ضریب تغییرات مشخص می‌شود. ضریب تغییرات نشان می‌دهد که میزان پراکندگی نسبت به میانگین در هر دو صفت یکسان است:

$$CV_۱ = \frac{۱/٤}{٥۰} \times ۱۰۰\% = ۲/۸\%$$

$$CV_۲ = \frac{۰/۱٤}{٥} \times ۱۰۰\% = ۲/۸\%$$

۱-۱٤-٦ استاندارد کردن داده‌ها

مقایسه داده‌های مربوط به جوامع مختلف، به این دلیل که ممکن است داده‌ها دارای واحدهای اندازه‌گیری متفاوت و میانگین یا انحراف معیار متفاوتی باشند بی‌فایده است. برای حذف اثر واحد اندازه‌گیری و یکسان نمودن میانگین و انحراف معیار، داده‌های جوامع مختلف را با استفاده از رابطه زیر استاندارد می‌کنند. مقایسه نمره‌ها پس از تبدیل آنها به داده‌های استاندارد در شرایط عادلانه صورت می‌گیرد. نمره‌های استاندارد حاصل از هر جامعه‌ای بدون واحد بوده و توزیعی با میانگین و پراکندگی یکسان تشکیل می‌دهند. میانگین و واریانس چنین جامعه‌ای که به جامعه استاندارد معروف است، به ترتیب برابر با صفر و یک می‌باشد.

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \quad \text{(نمره استاندارد)}$$

مثال ۱-۲۰. نمرات دانشجویان دانشکده‌ای در درس آمار دارای توزیعی متقارن با میانگین ۱٤/٥ و انحراف معیار ٤ و در درس ریاضیات دارای توزیعی متقارن با میانگین ۱٤/٤ و انحراف معیار ۲ می‌باشند. دانشجویی در درس آمار ۱۷/٥ و در درس ریاضیات ۱٦ گرفته است. او در کدام درس

به طور نسبی قوی‌تر است؟

این دو نمره را نمی‌توان مستقیماً با هم مقایسه کرد، زیرا نمونه‌های مربوط به آنها میانگین و انحراف معیار متفاوت دارند. لذا برای مقایسه آنها لازم است که ابتدا نمره‌ها استاندارد شوند.

$$z_۱ = \frac{x_۱ - \mu_۱}{\sigma_۱} = \frac{۱۷/۵ - ۱٤/۵}{٤} = ۰/۷۵$$

$$z_۲ = \frac{x_۲ - \mu_۲}{\sigma_۲} = \frac{۱٦ - ۱٤/٤}{۲} = ۰/۸۰$$

مقایسه نمره‌های استاندارد نشان می‌دهد که با وجود اینکه دانشجوی مذکور در درس ریاضیات نمره کمتری گرفته ولی نسبت به همکلاسی‌هایش در این درس قوی‌تر است.

جدول ۱-۹. محاسبه نمرات استاندارد در نرم‌افزار Minitab.

مراحل:

۱. داده‌هایی را که می‌خواهید استاندارد شوند در یک ستون وارد کرده و سپس وارد مسیر زیر شوید.

Calc > Standardize...

۲. در مقابل Input column(s) ستونی را که داده‌ها در آن قرار دارند انتخاب و در مقابل Store results in ستونی را که می‌خواهد داده‌های استاندارد شده در آنجا ذخیره شوند (مثلاً C2)، بنویسید.

۳. گزینه‌های مختلفی در این پنجره وجود دارد. می‌توانید گزینه Subtract first value and divide by second را انتخاب و به ترتیب میانگین و انحراف معیار را وارد کنید. سپس بر روی OK کلیک کنید.

تمرین‌ها

۱-۱. میانه و نما را برای داده‌های زیر به دست آورید.

الف) ۱, ۲, ۳, ٤, ۵, ۵, ۵, ۷, ۷, ۹
ب) ۱, ۲, ۳, ٤, ۵, ۵, ۵, ٦, ۸

۱-۲. نمودار بافت‌نگار را برای ۵۰۰ داده جدول زیر رسم کنید به طوری که محور عمودی، فراوانی نسبی باشد.

فراوانی نسبی	حدود دسته
۰/۱۰	۰/۵–۲/۵
۰/۱۵	۲/۵–٤/۵
۰/۲۵	٤/۵–٦/۵
۰/۲۰	٦/۵–۸/۵
۰/۰۵	۸/۵–۱۰/۵
۰/۱۰	۱۰/۵–۱۲/۵
۰/۱۰	۱۲/۵–۱٤/۵
۰/۰۵	۱٤/۵–۱٦/۵

۱–۳. در جامعه داده‌های تمرین ۱–۲، **الف)** تعداد افرادی که در هر کلاس قرار می‌گیرند را به دست آورید. **ب)** میانگین حسابی، میانگین هندسی، میانه، نما و انحراف معیار را به دست آورید.

۱–۴. جدول زیر را که نمرات دانشجویان در درس آمار است، تکمیل کنید.

فراوانی نسبی	فراوانی	نمرات
۰/۰۸		۹۰–۱۰۰
	۳٦	۸۰–۸۹
	۹۰	٦۵–۷۹
	۳۰	۵۰–٦٤
	۲۸	کمتر از ۵۰
۱	۲۰۰	جمع

۱–۵. سه ماشین به تولید یک کالا مشغولند. اولی کالا را در ۲، دومی در ۳ و سومی در ٦ دقیقه تولید می‌کند. اگر این سه ماشین با هم کار کنند به طور متوسط یک کالا در چند دقیقه تولید می‌شود؟

۱–٦. به سؤالات مطرح شده در ارتباط با تغییرات داده‌ها پاسخ دهید. **الف)** نشان دادن پراکندگی داده‌ها بر حسب دامنه تغییرات چه عیبی دارد؟ **ب)** واریانس نمونه را در جمله‌ای تعریف کنید. **ج)** واریانس جامعه را در جمله‌ای تعریف کنید. **د)** آیا واریانس می‌تواند منفی باشد؟ چرا؟ **ه)** آیا واریانس می‌تواند کوچکتر از انحراف معیار باشد؟ چرا؟

۱–۷. واریانس و انحراف معیار را در موارد زیر به دست آورید.

الف) $n = ۱۰, \sum x^۲ = ۸٤, \sum x = ۲۰$ ب) $n = ٤۰, \sum x^۲ = ۳۸۰, \sum x = ۱۰۰$

ج) $n = ۲۰, \sum x^۲ = ۱۸, \sum x = ۱۷$

۱–۸. دامنه، واریانس و انحراف معیار را برای نمونه‌های زیر به دست آورید. **الف)** ۱,۰,۱,٤,۲ ب) ۳,۰,۳,۲,۲,٦,۱ ج) ۳,۳,۱,٤,٤,۵,۳,۱,۲–,۸ د) ۱,۰,۱–,۲–,۳–,۰,۱–,۱,۱–,۰,۱,۲–,۱,۱–,۰,۰,۲,۰

۱–۹. مقادیر $\overline{x}$، $s^۲$ و s را برای داده‌های زیر به دست آورده و در هر مورد در صورت وجود واحد را نیز ذکر

کنید.

الف) ۳,۱,۱۰,۱۰,٤

ب) $\text{۳}m, \text{۱۰}m, \text{۳۲}m, \text{٥}m$

ج) $-\text{۳}, -\text{۲}, -\text{٤}, -\text{٥}, -\text{٤}$

د) $\text{۱/٥}kg, \text{۱/٥}kg, \text{۲/٥}kg, \text{٤/٥}kg$

۱–۱۰. با استفاده از داده‌های ۱ تا ۱۰ دو مجموعه داده بنویسید که میانگین یکسان ولی واریانس‌های متفاوت داشته باشند.

۱–۱۱. نمونه مقابل را که متشکل از ٥ داده می‌باشد در نظر بگیرید: ۲، ۱، ۱، ۰، ۳. **الف)** دامنه، $s^{\text{۲}}$ و s را به دست آورید. **ب)** به هر کدام از داده‌ها ۳ را اضافه کرده و محاسبات را تکرار کنید. **ج)** عدد ٤ را از هر کدام کم کرده و محاسبات خواسته شده در قسمت الف را تکرار کنید. **د)** جواب‌ها را مقایسه کرده و اثر افزودن یا کاستن یک مقدار ثابت از داده‌ها را بر روی دامنه، $s^{\text{۲}}$ و s بیان کنید.

۱–۱۲. جوامع زیر را در نظر بگیرید. **الف)** حدس می‌زنید چه رابطه‌ای بین واریانس جوامع a با b و همچنین c با d وجود داشته باشد؟ **ب)** واریانس‌ها را به دست آورده و مقایسه کنید. **ج)** نمودار نقطه‌ای را برای جوامع a و b رسم کنید.

الف) ۲, ۳, ٥, ۷

ب) ۲, ۲, ۲, ۳, ۳, ۳, ٥, ٥, ٥, ۷, ۷, ۷

ج) ٦, ۹, ۱٥, ۲۱

د) ۱۰, ۱۳, ۱۹, ۲٥

۱–۱۳. داده‌های تمرین ۱–۱ را به عنوان یک جامعه در نظر بگیرید. **الف)** میانه را به دست آورید. **ب)** واریانس و انحراف معیار داده‌ها را به دست آورید.

۱–۱۴. در یک نمونه ٦ عضوی میانگین ۲ و مجموع توان دوم‌ها ۳۰ می‌باشد. واریانس نمونه را به دست آورید.

۱–۱۵. الف) ضریب تغییرات متغیر x، ۲۰٪ است. اگر داده‌های مربوطه بر ٥ تقسیم شوند، ضریب تغییرات چقدر خواهد شد؟

۱–۱۶. اگر N=۱۰، $\sum x_i$ =۲۰۰ و $\sum x_i^{\text{۲}}$ = ٥۰۰ باشد، مقدار ضریب تغییرات را به دست آورید.

۱–۱۷. در چه صورتی هر سه نوع میانگین μ، g و h با یکدیگر برابر خواهند بود؟

۱–۱۸. کدام‌یک از پارامترهای واریانس، میانه، دامنه میانچارکی و میانگین انحرافات بیشتر تحت تأثیر داده‌های پرت قرار می‌گیرند؟