آمار و احتمالات

(آمار زیستی)

تهیه و تنظیم:

صفائی

بخش اول

آمار توصیفی

فصل اول

کلیات و تعاریف

مقدمه:

برای اولین بار در سده هجدهم تدریس آمار به عنوان دولت شناسی در دانشگاه‌ها آغاز شد. علم آمار برای دانشمندان و به ویژه پژوهشگران از سال ۱۹۲۵ شروع شد. در آن سال کتاب معروف فیشر با نام "روش های آماری برای تحقیق" منتشر شد.

برای تبیین ضرورت کاربرد آمار در پژوهش های علمی، لازم است که به مراحل اجرای یک پژوهش علمی نگاهی بیاندازیم. به طور خلاصه در اجرای یک تحقیق، شش مرحله کلی به شرح زیر وجود دارد:

۱- تعیین و تعریف مسئله تحقیق

۲- تدوین فرضیه های تحقیق بر اساس نظریه و پژوهش های گذشته

۳- طرح تحقیق و روش نمونه برداری

۴- جمع آوری داده ها

۵- تجزیه و تحلیل داده ها

۶- تفسیر داده ها بر اساس مسئله تحقیق

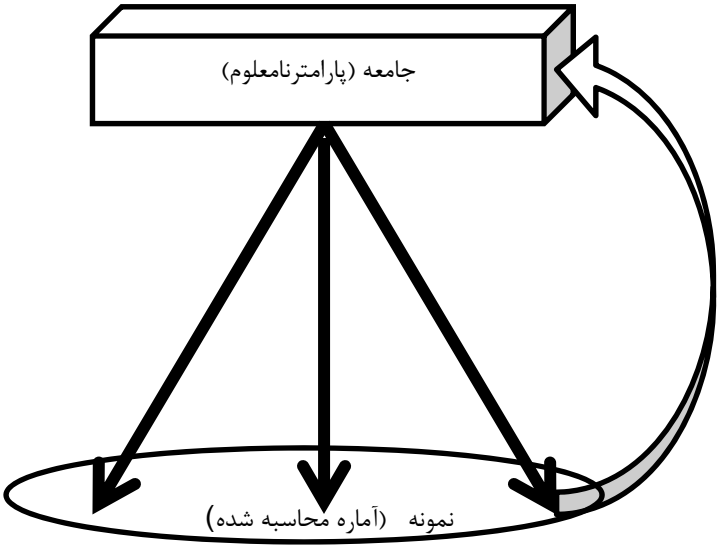
همانطوریکه دیده میشود آماردر بیشتر مراحل اجرای یک تحقیق(۳ الی ۶) به کار گرفته میشود.

توسعه و رشد روزافزون آمار سبب شده است که ارائه تعریف منطقی از آن مشکل شود. تعریف های متعددی برای آمار ارائه شده است. می توان بیان کرد که علم آمار عبارت است از مفاهیم، اصول و روشهایی که برای جمع آوری، تنظیم، خلاصه کردن، تجزیه و تفسیر داده ها به کار می رود. شاخه ای از آمار که در پردازش داده های علوم زیستی و پزشکی به کار می رود را اصطلاحا آمار زیستی می نامیم.

در بررسی های آماری، دو هدف اصلی و مشخص دنبال می شود. هدف اول که در آمار توصیفی منظور می شود، عبارت است از جمع‌آوری، گروه‌بندی، خلاصه کردن، رسم نمودار و تعیین شاخص‌های مرکزی و پراکندگی، مانند میانگین، میانه، نما، دامنه تغییرات، انحراف متوسط از میانگین، انحراف معیار و واریانس. این شاخص‌ها برای محاسبه دیگر شاخص‌های آماری استفاده می شوند.

هدف دوم که در آمار استنباطی گنجانده می شود، عبارت است از تعمیم دادن نمونه به جامعه دراین مفهوم استنباط آماری پلی برای ارتباط بین جامعه و نمونه است (شکل ۱-۱).

از آنجا که اغلب امکان آماربرداری از کل افراد جامعه وجود ندارد، از نمونه برداری استفاده می‌شود. در اینجا احتمال پل ارتباطی بین آمار توصیفی و آمار استنباطی می‌باشد.



شکل ۱-۱- برآورد پارامترهای جامعه از ویژگی های نمونه

جامعه و نمونه

جامعه مجموعه ای از افراد است که حداقل یک ویژگی مشترک داشته باشند. مثل جامعه سلولهای یک بافت زنده، جامعه ماهی های موجود در یک استخر، جامعه درختان یک جنگل ، جامعه بیماران قبلی یک بیمارستان و غیره.

تعداد عناصر موجود در جامعه را حجم جامعه می نامند.

از نظر تعداد افراد، جامعه، محدود یا نامحدود است. وقتی که تعداد افراد یک جامعه محدود و قابل شمارش باشد، آن را جامعه محدود گویند، مثل جامعه سگهای آبی یک برکه. ویژگی این جامعه آن است که اگر از افراد آن برداشته شود(یکی یکی) و جایگذاری نشود، بالاخره تمام می شود. اگر تعداد افراد جامعه نامحدود و غیر قابل شمارش باشد، آن را جامعه نامحدود گویند، مثل جامعه ماسه های کنار دریا. ویژگی این جامعه آن است که هر قدر از افراد آن برداشته شود، تمام نمی شود.

در برابر مفهوم جامعه، نمونه مطرح می شود. در بیشتر موارد مطالعه و بررسی کلیه افراد یک جامعه به علت کمبود وقت و هزینه زیاد امکان‌پذیر نیست، لذا بخشی از جامعه را مورد مطالعه و بررسی قرار می‌دهیم که معرف آن جامعه باشد که به آن نمونه می‌گوییم. نمونه قسمتی از جامعه است که اطلاعات مربوط به آن برای استنتاج درباره جامعه به کار می رود. از ویژگی‌های نمونه مشابهت آن با خصوصیات جامعه است، به طوری که معرف جامعه بوده و زیر مجموعه ای از آن به شمار می رود. برای مثال، جامعه مورد بررسی، گیاهان علفی موجود در مراتع یک منطقه است. هرگاه برای تجزیه و تحلیل، فقط تعدادی از گیاهان آن منطقه انتخاب و مطالعه شود، یک نمونه از جامعه آماری انتخاب شده است.

با این تعریف طبیعی است که از هر جامعه آماری، می توان نمونه های متعددی انتخاب کرد.

انواع متغیرها

ویژگی هایی که پژوهشگران آنها را مشاهده و اندازه می گیرند، متغیر یا صفت نامیده می شوند. در هر بررسی آماری از اندازه گیری صفت مورد مطالعه در عناصر جامعه (یا نمونه) تعدادی عدد به دست می آید (برای مثال، وقتی طول ۵

نهال کاج دو ساله را اندازه بگیریم و اعداد ۱۰۲، ۱۰۵، ۹۶، ۱۰۰ و ۱۰۳ حاصل شود) که هر یک از آنها را داده یا مشاهده می نامند. پردازش روی این داده‌ها منجر به تولید اطلاعات می‌شود که بیشتر قابل فهم و کاربرد است.

متغیرها را می توان به کمی و کیفی تقسیم کرد. متغیرهای کمی، متغیرهای قابل اندازه گیری هستند و به صورت عدد نوشته می شوند، مانند قد، وزن، مساحت، زمان و غیره.

متغیرهای کیفی، متغیرهایی هستند که قابل اندازه گیری کمی نیستند و به صورت غیر عددی نشان داده می شوند، مثل جنس، رنگ، نوع نژاد، نوع بافت، شکل دانه، شدت حساسیت و غیره.

متغیرها بر اساس رابطه ای که با یکدیگر دارند، به دو دسته مستقل و وابسته تقسیم شده اند. متغیر مستقل، نقش علت را به عهده دارد و بر متغیرهای دیگر تاثیر می گذارد و منشا بروز پدیده ها می شود. متغیر وابسته، متغیری است که ارزش آن به متغیر مستقل بستگی دارد. برای مثال، در رابطه بین احتمال سرماخوردگی با دمای هوا، احتمال سرماخوردگی متغیر وابسته و دمای هوا متغیر مستقل است.

متغیرها ممکن است ناپیوسته یا گسسته و پیوسته باشند. متغیر ناپیوسته یا گسسته متغیری است که فقط مجموعه‌ای از ارزش‌های معین به آن اختصاص داده می شود. به عبارت دیگر، در متغیرهای ناپیوسته خرد کردن ارزشها امکان‌پذیر نیست. برای مثال، تعداد گلبرگ‌ها در یک گل، تعداد دانشجویان یک مجموعه و تعداد تخمهای یک پرنده، از متغیرهای نا پیوسته محسوب می‌شوند، زیرا فقط مقادیر ثابتی مثل ۳، ۵ و غیره را به خود اختصاص می‌دهند. در متغیرهای گسسته، ارزش‌های موجود بین دو مقدار یا دو ارزش دارای معنا و مفهوم خاص نیست. برای مثال، در صورتی که تعداد گلبرگهای یک گل ۵ عدد باشد، انتخاب اعداد ۴.۷۵ ، ۴.۲۵ یا ۳.۵ معنا و مفهومی ندارد. از دیگر موارد میتوان به تعداد دندانهای کشیده شده کودکان یک دبستان، تعداد سلولهای یک بافت زنده، گروه خونی افراد یک جامعه، تعداد فرزندان خانواده‌های یک شهرو از این دست اشاره کرد. متغیر ناپیوسته می تواند هم به صورت کیفی باشد، مانند جنس، نوع گونه و هم به صورت کمی مانند تعداد گلبرگ های گل و تعداد گوزنهای موجود در یک گله.

متغیر پیوسته متغیری است که هر ارزش یا مقدار (اعشاری و کسری) را می توان به آن اختصاص داد. به عبارت دیگر هر عددی بین دو واحد آن انتخاب شود، دارای معنا و مفهوم است و یا به عبارت دیگر قابل خورد شدن باشد. وزن یک

موجود زنده، ارتفاع یک درخت و مدت زمان تکثیر یک سلول به تعدادی مشخص، همگی از متغیرهای پیوسته محسوب می شوند. متغیر پیوسته به طور طبیعی کمی است و درجه اندازه‌گیری آن به دقت وسیله اندازه‌گیری بستگی دارند. برای مثال، وزن یک متغیر پیوسته است و می تواند بین صفر و بی‌نهایت باشد.

مقیاس های اندازه گیری

هر بررسی آماری با اندازه‌گیری صفت‌های متغیر سر و کار دارد. اندازه‌گیری فرایندی است که با آن مشاهده‌ها به عدد تبدیل می‌شوند. اندازه‌گیری را در سطح‌های مختلف می‌توان انجام داد که به این سطح‌ها، مقیاس می‌گویند. به عبارت دیگر، می‌توان گفت که اعداد در موقعیت‌های متفاوت معنا و مضمون‌های مختلفی پیدا می‌کنند. به طور کلی متغیرها در چهار مقیاس اندازه‌گیری می‌شوند که بر حسب نوع مقیاس، روش‌های آماری متفاوتی ارائه می‌شود. این مقیاس‌ها عبارتند از:

الف- مقیاس اسمی (Nominal):

این داده‌ها ضعیف‌ترین سطح از اندازه‌گیری را ارائه می‌دهند و فقط می‌توان تعیین کرد که یک شیء متفاوت از شیء دیگر است. بر روی این اعداد نمی توان اعمال ریاضی (جمع، تفریق، ضرب و تقسیم) را انجام داد. تنها ویژگی داده‌های اسمی هم معنا بودن و هم وزنی آنهاست. ارزش ها هیچ رابطه ای با هم ندارند و اغلب به عنوان یک کلاس در نظر گرفته می شوند. مانند شماره شناسنامه افراد، پلاک خانه، شماره ای که به بیماری ها اختصاص می دهیم و ... .

ب- مقیاس رتبه‌ای (Ordinal):

بعضی از متغیرها را نمی توان بر روی مقیاس عددی اندازه گرفت، ولی می توان آنها را در ارتباط با یکدیگر رتبه بندی کرد. در این حالت نیز مانند مقیاس اسمی دارای تعدادی طبقه است، ولی در اینجا، طبقه ها نسبت به هم دارای امتیاز یا رتبه هستند. ارتباط بین گروه‌ها از نوع «بزرگ‌تر»، «سخت‌تر»، «خوب‌تر» و از این قبیل می‌باشد.

دو ویژگی داده‌های رتبه‌ای هم معنا و بیشتر یا کمتر بودن است. برای مثال، اگر طبقات شدت حساسیت به بیماری خاص (ناچیز، کم، متوسط، زیاد و خیلی زیاد)، با ۱، ۲، ۳، ۴ و ۵ رتبه بندی شوند، هر کدام از رتبه‌ها به معنای یک مقدار

مشخص شدت حساسیت است، همچنین رتبه ۲ دارای حساسیت بیشتری از رتبه ۱ بوده و به همین ترتیب رتبه ۳، حساسیت بیشتری از رتبه ۲ دارد. این ارتباطات می توانند از قبیل بزرگتر، شدید تر، بهتر و از این قبیل باشد. اما این بدین معنی نیست که حساسیت ۴ دو برابر حساسیت ۲ می‌باشد. یا مثلا اگر محققی حیوانات را بر حسب میزان هوش با اعداد ۱، ۲، ۳ و ۴ مشخص کند، حیوان شماره ۴ از حیوان شماره ۲ باهوشتر است، اما نمی‌توان گفت که دو برابر او باهوش است.

ج- مقیاس فاصله ای (Interval):

این مقیاس افزون بر دارا بودن ویژگی های دو مقیاس اسمی و رتبه ای (یعنی گروه بندی عناصر جامعه و تعیین رتبه آنها بر اساس میزان صفت مورد نظر)، فواصل واحدهای عددی اندازه گیری یکسانی دارد. به عبارت دیگر اگر یک متغیر عناصر جامعه را گروه‌بندی کند و ارتباط بین گروه‌ها از نوع رتبه‌ای باشدو علاوه بر این دو فاصله بین گروه‌ها عدد ثابتی باشد به این مقیاس اندازه‌گیری مقیاس فاصله‌ای گویند.

در این نوع مقیاس نسبت هر دو فاصله مستقل از واحد اندازه گیری و مستقل از نقطه صفر است. به عنوان مثال دمای بین ۰ تا ۱ درجه سانتی‌گراد و ۱۰۰ درجه تا ۱۰۱ درجه سانتی‌گراد، ۱ درجه می‌باشد. این مسئله در مورد درجه فارنهایت هم صحیح می‌باشد بنابراین ربطی به واحد اندازه‌گیری ندارد.

در مقیاس فاصله ای، نقطه صفر و واحد اندازه گیری اختیاری و قراردادی است. برای مثال، با آنکه واحد اندازه گیری و نقطه صفر در دو مقیاس جداگانه سانتی گراد و فارنهایت با یکدیگر متفاوتند، هر دو برای اندازه گیری درجه حرارت به کار می روند.

د- مقیاس نسبتی (Ratio):

مقیاس نسبتی مثل مقیاس فاصله‌ای است، اما با یک نقطه صفر واقعی، بنابراین محاسبه نسبت‌ها اهمیت دارد. اگر داده‌ها واحد اندازه گیری و صفر واقعی داشته باشند، در این صورت دارای مقیاس نسبتی هستند. برای مثال، مقیاس هایی چون پوند و گرم نقطه صفر واقعی دارند و نسبت بین هر دو وزن دلخواه، مستقل از واحد اندازه گیری است. وزن دو شیء

مختلف را می توان هم با گرم و هم با پوند اندازه گیری کرد. در این صورت نسبت دو وزن بر حسب پوند، برابر با نسبت همان دو وزن بر حسب گرم است. بدان جهت نسبتی گفته می‌شود که نسبتها در مقیاسهای اندازه‌گیری مختلف یکسان هستند. به عنوان نمونه اگر فرد اول ۵۰ کیلوگرم وزن داشته باشد، این میزان برابر با ۱۱۰ پوند خواهد بود. حال اگر شخص دوم ۶۰ کیلوگرم وزن داشته باشد، این میزان برابر ۱۳۲ پوند می‌باشد. لذا نسبت وزن فرد دوم به اول چه بر حسب پوند و چه بر حسب کیلوگرم برابر خواهد بود با:

$$\frac{۱۳۲}{۱۱۰} = \frac{۶}{۵} \quad و \quad \frac{۶۰}{۵۰} = \frac{۶}{۵}$$

تمارین:

۱- پل ارتباطی بین نمونه و جامعه و بین آمار توصیفی و آمار استنباطی چیست؟

۲- تفاوت بین داده و اطلاعات چیست؟

۳- مثالی از یک تحقیق بزنید که در آن متغیرهای مستقل و وابسته وجود دارد.

٤- مقیاس‌های اندازه‌گیری مناسب و گسستگی یا پیوستگی متغیرهای زیر را مشخص کنید.

الف) گروه خونی ب) ضریب هوشی ج) دما د) بیومس

فصل دوم

تنظیم و خلاصه‌کردن داده‌ها

مقدمه:

از آنجا که در علوم زیستی متغیرهای متعددی جمع آوری می شوند، لازم است داده‌ها با روش صحیحی خلاصه شوند، در ادامه به روشهای آماری مورد استفاده برای خلاصه کردن و توصیف آنچه که در یک پژوهش جمع آوری می‌شود، ارائه خواهد شد. این روشها داده‌های به دست آمده از هر متغیر را خلاصه و توصیف می‌کنند.

توصیف داده‌ها ممکن است آخرین مرحله یک تجزیه و تحلیل ساده آماری باشد یا این که مرحله مقدماتی تجزیه و تحلیل‌های بعدی به شمار رود. برای توصیف داده‌ها از ابزارهایی مانند ۱)جدول، ۲)نمودار، ۳)محاسبه شاخص‌های مرکزی و ۴)پراکندگی استفاده می‌کنند.

۱) ابزار جدول برای طبقه‌بندی داده‌ها

بعد از جمع‌آوری و ویرایش داده‌ها، طبقه‌بندی اولین گامی است که باید برداشته شود. مرتب کردن داده‌ها بر اساس ویژگی‌های مشترک را طبقه‌بندی گویند. با توجه به نوع داده‌ها (کیفی و کمی) می‌توان طبقه‌بندی مناسبی را انجام داد.

طبقه‌بندی عبارت است از تهیه جدولی که مقادیر متغیر مورد نظر را به تعداد افرادی که این کمیت‌ها یا کیفیت‌ها را معرفی کردند، مربوط می‌کند. هر کدام از این تعداد را، فراوانی می‌نامند. جدول به‌دست آمده را، جدول توزیع آماری یا جدول توزیع فراوانی می‌گویند. بسته به نوع متغیرها (گسسته یا پیوسته) و اهداف تحقیق، به دو روش زیر می‌توان جدول توزیع فراوانی آنها را ترسیم کرد.

جدول توزیع فراوانی برای داده‌های گسسته

اگر هدف این باشد تا اندازه‌های یک متغیر را در K طبقه، نمایش دهند، مقادیر آنها را در ستونی با عنوان X_i (اندازه ویژگی مورد مطالعه) مشخص کرده و در ستون دیگر فراوانی مطلق آنها را با F_i (که F_i تعداد X_iها در بین کل داده‌هاست) نشان می‌دهند. همچنین اگر فراوانی مطلق بر تعداد کل داده‌ها تقسیم شود، فراوانی نسبی (P_i) به‌دست می‌آید. جمع فراوانی‌های مطلق طبقات پشت سر هم، ستون فراوانی تجمعی مطلق (F_c) و جمع فراوانی نسبی، طبقات پشت سر هم فراوانی تجمعی نسبی (P_c) را تشکیل می‌دهند. (مطابق شکل زیر)

مثلا جدول فراوانی مربوط به گروه های خونی یک جامعه میتواند به شکل زیر باشد.

X_i	F_i	Fc	Pi	Pc
A	1	1	0.09	0.09
B	3	4	0.27	0.36
O	2	6	0.18	0.55
AB	5	11	0.45	1.00
مجموع	11		1.00	

از خواص جداول فراوانی این است که مجموع فراوانیهای مطلق برابر با تعداد کل داده هاست، فراوانی نسبی بین ۰ تا ۱ (یا بین ۰ تا ۱۰۰ در مقیاس درصد) تغییر می‌کند، مجموع فراوانی نسبی برابر ۱ (یا ۱۰۰ درصد در مقیاس درصد) است. اگر اندازه‌های یک متغیر گسسته زیاد باشد مقادیر آنها را طبقه‌بندی کرده و سپس جدول فراوانی را تهیه می‌کنیم که در ادامه توضیح داده خواهد شد.

جدول توزیع فراوانی برای داده‌های پیوسته

برای تشکیل جدول توزیع فراوانی داده های پیوسته، موارد زیر انجام می‌شود:

الف- تعیین تعداد دسته‌ها: ابتدا تعداد دسته‌ها از فرمول‌های تجربی، یول یعنی $(n)^{1/2}$ * ۲.۵ یا از فرمول استورجس $\mathrm{K} = 1 + 3.3\log(n)$ یا از فرمول تقریبی دیگری، که عبارت است از: $n = 2^k$ تعیین می‌شود که در آن $k = \frac{ln_n}{ln_2}$ می باشد. در فرمول‌های بالا n: تعداد مشاهده‌ها یا داده‌ها و K: تعداد دسته‌هاست.

مثال: اگر تعداد کل مشاهده‌ها برابر ۶۰ باشد، تعداد دسته‌ها با استفاده از سه فرمول ذکر شده در قبل برابر است با:

فرمول استورجس: $K = ۱ + ۳.۳ \log ۶۰ = ۱ + ۳.۳ \log ۱.۷۷۸۲ = ۶.۸۷$

فرمول یول: $K = ۲.۵ * n^{\frac{۱}{۴}} = ۲.۵ * ۶۰^{\frac{۱}{۴}} = ۶.۹۶$

فرمول تقریبی: $۵۰ = ۲^k \rightarrow K = \frac{LN(n)}{LN(۲)} = \frac{۴.۰۹}{۰.۶۹} = ۵.۹۲$

تعداد دسته‌ها باید عددی رند باشد بنابراین تعداد دسته ها با استفاده از این سه فرمول به ترتیب ۷،۷ و ۶ بدست می‌آید.

لازم به ذکر است که هیچ یک از فرمول‌های بالا، بر دیگری برتری ندارند و انتخاب تعداد دسته‌ها اختیاری بوده و به تجربه پژوهشگر بستگی دارد. به طور معمول تعداد دسته‌ها ۲۰-۱۰ عدد در نظر گرفته می‌شود. درصورتی که تعداد دسته‌ها کمتر از ۱۰ باشد، اندازه دسته‌ها بزرگ می‌شود و همین امر سبب خواهد شد اطلاعات بیشتری از دست برود. چنانچه تعداد دسته‌ها بیشتر از ۲۰ شود، تهیه و تنظیم جدول مشکل خواهد بود. بعضی معتقدند که بهتر است دسته‌ها کمتر از ۵ و بیشتر از ۲۰ نباشد. اما در نهایت این متخصص است که تعداد دسته ها را با توجه به دقت مورد نیاز تعیین می کند.

ب- تعیین فاصله دسته‌ها: فاصله دسته‌ها با استفاده از رابطه زیر تعیین می‌شود:

$$I = \frac{R}{K} = \frac{Max(X_i) - Min(X_i)}{K}$$

که در آن I: فاصله بین دو دسته، X_i: مقادیر داده‌ها، R: دامنه مقادیر و K: تعداد دسته‌هاست.

برای پرهیز از مشکلات بعدی، بهتر است که فاصله برای تمام دسته‌ها یکسان و عدد رندی در نظر گرفته شود.

ج- مرتب کردن صعودی یا نزولی داده‌ها: داده‌های کمی را باید به صورت صعودی یا نزولی مرتب کرد. در مورد داده‌های کیفی، باید آنها را به ترتیب حروف الفبا یا موارد دیگر مرتب کرد.

د- نوشتن دسته‌ها: چهارمین مرحله، نوشتن دسته‌هاست. اغلب نوشتن دسته‌ها را از کوچکترین عددی که اندازه یا فاصله دسته‌ها مضربی از آن است، آغاز می‌کنند. برای مثال، وقتی که فاصله دسته‌ها برابر سه باشد، بهتر است که دسته‌ها با یکی از اعداد ۶، ۹، ۱۲، ۱۵ یا ۱۸ شروع شود. پس از تعیین نقطه شروع دسته‌ها، یعنی کوچکترین عدد دسته که حد پایین دسته اول است، بزرگترین عدد یا حد بالای این دسته را محاسبه می‌کنند.

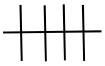
دسته‌بندی صریح داده‌های گسسته و پیوسته :

داده‌های گسسته اعدادی کامل می‌باشند و نیاز به گرد کردن ندارند و هنگام دسته بندی نباید یک عدد در دو دسته متوالی نوشته شود. مثلا اگر مقدار حد اقل داده ۳ و عرض دسته ها ۴ باشد، دسته اول ۳-۶ دسته دوم ۷-۱۰ دسته سوم ۱۱-۱۴ و به همین ترتیب خواهد بود. اما داده‌های پیوسته دارای مقدار دقیقی نیستند. به عنوان مثال اگر طبقه‌بندی دسته اول و دوم قد دانشجویان یک دانشگاه به ترتیب ۶۲-۶۰ و ۶۴-۶۲ سانتی‌متر باشد، فرد نمی‌تواند قضاوت کند که محقق مورد نظر قد ۶۲.۲ را جزو طبقه اول در نظر گرفته با طبقه دوم. اما اگر حدود طبقات را به صورت ۶۱.۵-۵۲.۵ و ۷۲.۵-۶۱.۵در نظر بگیریم مشخص است که محقق قد ۶۱.۲ را در طبقه اول گنجانده است.

توجه: اگر دقت اندازه‌گیری داده‌های پیوسته یک رقم اعشار باشند برای نوشتن حدود صریح طبقات مثلا ۰.۵ واحد از حد پایین و بالا کم می‌شود. اگر دو رقم اعشار باشد ۰.۰۵ واحد از پایین و بالا کم می‌شود و ... .

باید بدانیم که دقت اندازه‌گیری در هر مطالعه‌ای مشخص می‌باشد که می‌توان آن را از مطالعات قبلی بدست آورد، لذا مثلا اگر در مورد یک متغیری دقت اندازه‌گیری یک رقم اعشار باشد اما اندازه‌گیری‌های ما با توجه به دقت پایین دستگاه با تصادفا اعشاری نداشته باشند در این حالت نباید از اعداد کامل استفاده گردد و حتما باید از حدود صریح طبقات استفاده شود تا شبهه‌ای در طبقه‌بندی‌ها به وجود نیاید. نکته آخر اینکه برای جلوگیری از به هدر رفتن دقت اطلاعات بهتر است به جای اینکه داده‌ها را اول گرد کنیم و بعد دسته‌بندی کنیم، اجازه دهیم جداول تعیین کنند که هر داده در کدام دسته قرار می‌گیرند.

ه- خط و نشان: در پایان، خط و نشان‌های هر دسته را کشیده و بعد شمارش در مقابل همان دسته، به صورت عددی می‌نویسند تا فراوانی‌های مطلق هر دسته به‌دست آید.

انواع خط و نشان:  یا 

و- محاسبه فراوانی تجمعی مطلق، فراوانی نسبی و فراوانی تجمعی نسبی.

مثلا جدول فراوانی وزن دانشجویان یک کلاس ۱۳ نفره که نمونه‌ای از داده‌های پیوسته می‌باشد، به شکل زیر است.

دسته ها	F_i	F_c	P_i	P_c
50.5-58.5	3	3	0.23	0.23
58.5-66.5	4	7	0.31	0.54
66.5-73.5	1	8	0.08	0.62
73.5-81.5	5	13	0.38	1.00
مجموع	13		1.00	

جدول فراوانی مصرف قرص استامینوفن ۲۰ خانواده که نمونه‌ای از داده‌های گسسته می‌باشد، به شرح زیر است:

دسته‌ها	F_i	F_c	P_i	P_c
۰-٤	٦	٦	۰.۳	۰.۳
٥-۹	۹	۱٥	۰.٤٥	۰.۷٥
۱۰-۱٤	۳	۱۸	۰.۱٥	۰.۹
۱٥-۱۹	۲	۲۰	۰.۱	۱.۰۰
مجموع	۲۰		۱.۰۰	

کاربرد جداول فراوانی:

جداول فراوانی از مهمترین ابزارهای تنظیم و خلاصه‌سازی داده‌ها می‌باشند. برای نمونه از جدول فراوانی گروه خونی می‌توان برخی نتایج را به شرح زیر اعلام کرد. بیشتر افراد گروه خونی AB و کمتر آنها گروه خونی A دارند (از روی F_i). ٦ نفر از افراد گروه خونی A، B و O دارند (از روی F_c). ۰.۱۸ یا ۱۸ درصد افراد گروه خونی O دارند (از روی P_i). ۰.۵۵ یا ۵۵ درصد افراد گروه خونی A، B و O دارند.

یکی دیگر از کاربردهای F_c تعیین داده شماره nام می‌باشد. برای مثال در جدول گروه خونی، داده شماره سوم دارای گروه خونی B می‌باشد که در داده‌های بازنویسی شده زیر نیز قابل اثبات است.

شماره داده	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱
داده	A	B	B	B	O	O	AB	AB	AB	AB	AB

ذکر این نکته ضروری است که در دسته بندی بهتر است از دسته بندیهای متعارف در مطالعات مشابه استفاده گردد تا مقایسه مطالعات مختلف به آسانی امکانپذیر باشد.

خصوصیات یک جدول مناسب

۱) تا حد ممکن ساده باشد تا به راحتی بتوان از آن اطلاعات مورد نیاز را استخراج کرد. دو یا سه جدول کوچک ساده بهتر از جداول بزرگ ترکیبی متقاطع از چندین متغیر می‌باشد.

۲) کدها و علائم اختصاری و نشانه‌ها می‌بایست در پانویس جدول آورده شود.

۳) در ترسیم جداول باید هر ستون و ردیف را کاملا بتوان از هم تشخیص داد.

۴) واحد داده‌ها در هر ستون باید مشخص باشد.

۵) عنوان جداول باید دقیق و کامل باشد.

۶) عنوان جدول معمولا جدا از بدنه جدول در بالای آن نوشته شود.

۷) اگر داده‌های جدول از منابع دیگری جمع‌آوری شده‌اند، حتما باید ماخذ داده‌ها در پانویس ذکر گردد.

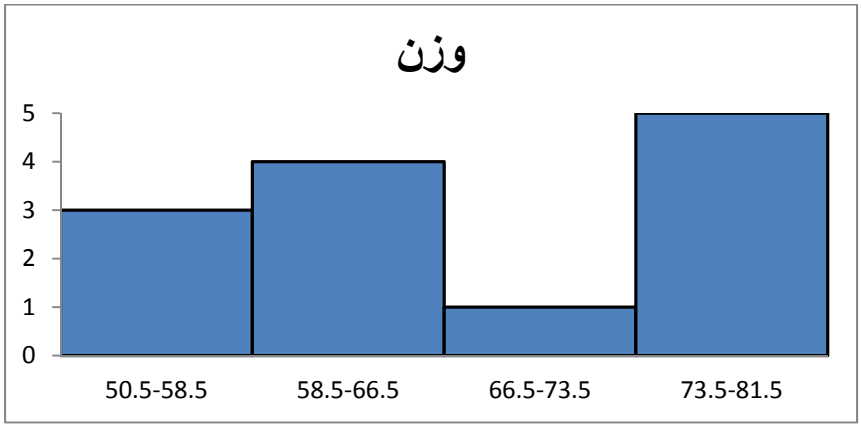
۲) ابزار نمونه‌برداری نمایش هندسی داده‌ها

برای نمایش توزیع‌های فراوانی، اغلب از نمودار استفاده می‌شود. نمودارها کمک می‌کنند تا تصویر توزیع به سهولت دیده شود. روش‌های نمایش داده‌ها، افزون بر استفاده نظری، برای استفاده‌های علمی و کاربردی نیز مورد توجه می‌باشند. به‌کارگیری این روش‌ها در گزارش نویسی، یکی از راه‌های مفید برای انتقال موضوع گزارش به خواننده در کمترین زمان ممکن و با ساده‌ترین بیان است. روش نمایش داده‌ها، با توجه به نوع مقیاس اندازه‌گیری آنها انتخاب می‌شود. چنانچه مقیاس داده‌ها از نوع فاصله‌ای و نسبتی باشد، از نمودارهای کمی استفاده می‌شود. داده‌های با مقیاس اسمی یا رتبه‌ای نیز به کمک نمودارهای توصیفی قابل نمایش هستند.

نمودارهای کمی

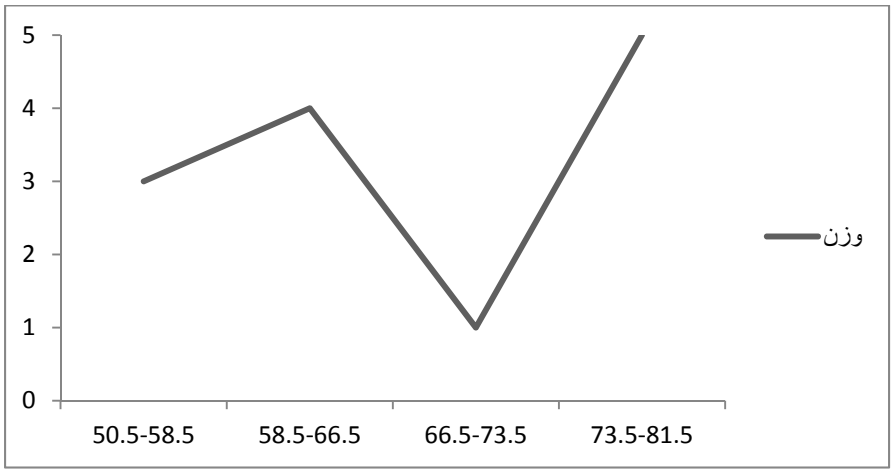
نمودار مستطیلی

در مواردی که داده‌ها، اندازه‌های صفت کمی پیوسته و ناشی از اندازه‌گیری باشند، توزیع فراوانی گروه‌بندی شده برای تنظیم آنها و نمودار مستطیلی برای نمایش هندسی آنها، مناسب است. اگر حد بالا و پایین دسته‌ها، به طور مناسب روی محور X قرار گیرند و بر روی هر فاصله، مستطیلی ساخته شود که قاعده‌اش فاصله آن دسته و ارتفاع آن فراوانی نظیر دسته مورد نظر باشد، مجموعه مستطیل‌هایی که به ترتیب حاصل می‌شوند، هیستوگرام یا نمودار مستطیلی نامیده می‌شود. شکل زیریک نمودار مستطیلی را نشان می دهد.



نمودار چند ضلعی

نوع داده‌های لازم برای نمودار چندضلعی شبیه نمودار مستطیلی است . برای رسم نمودار چندضلعی، ابتدا فراوانی دسته‌ها با نقطه مشخص می‌شود. سپس در مقابل آنها نقطه وسط هر دسته را به دست می‌آورند و نقاط مجاور هم را با خط‌های مستقیمی به یکدیگر متصل می‌کنند. نمودار حاصل را چندضلعی فراوانی می‌گویند. یک نمونه از نمودار چند ضلعی در زیر آمده است.

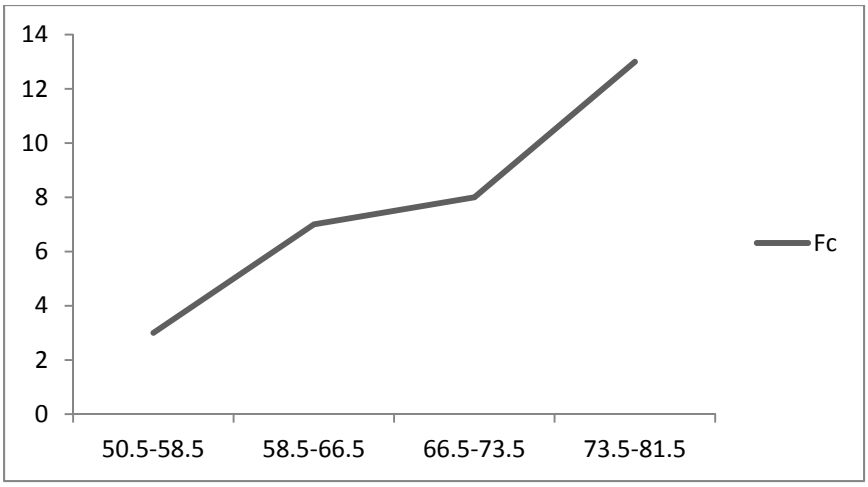


که به عنوان مثال نشان می‌دهد یک نفر از افراد جامعه وزن ۶۶.۵-۷۳.۵ دارند.

نمودار فراوانی تجمعی

توزیع فراوانی تجمعی، توزیعی است که تعداد مشاهده‌ها را قبل از یک نقطه معین در مقایسه با دیگر مشاهده‌ها نشان می‌دهد. نمودار فراوانی تجمعی بر اساس توزیع فراوانی تجمعی رسم می‌شود که می‌تواند مستطیلی یا چند ضلعی باشد.

مثلا فراوانی تجمعی مطلق مثال قبل به شکل زیر خواهد بود:



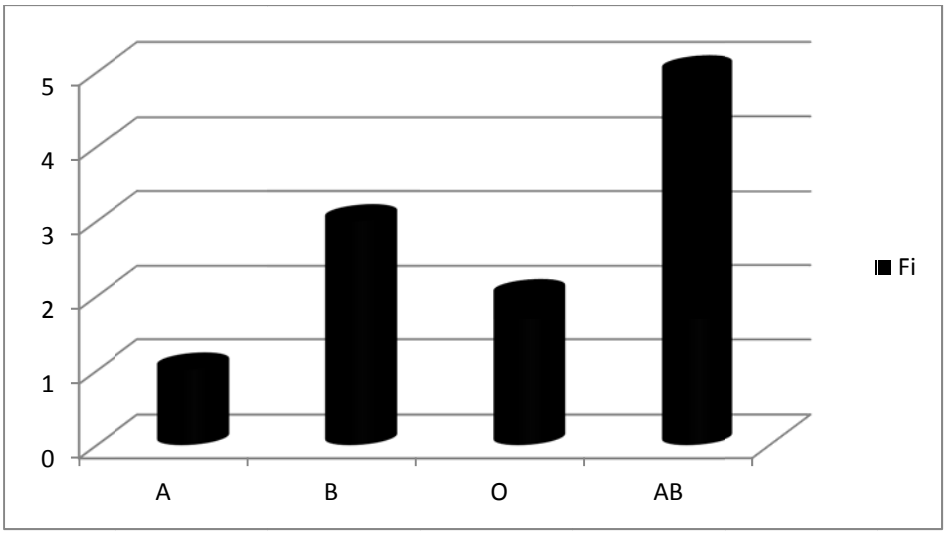
که به عنوان مثال نشان می‌دهد که ۸ نفر از افراد دارای وزن بین ۵۰.۵-۷۳.۵ دارند.

نمودارهای توصیفی

اگر داده‌های جمع‌آوری شده از نوع کیفی باشند، برای نمایش هندسی آنها از نمودارهای توصیفی استفاده می‌کنند. مشاهده‌هایی از نوع خوب، بد، متوسط و گروه‌های A، B، C را می‌توان از نوع مشاهده‌های کیفی دانست. طبیعی است که برای این نوع مشاهده‌ها، تعریف فاصله، چنانچه در نمودارهای مستطیلی گفته شد، مفهومی ندارد. انواع نمودارهای توصیفی به شرح زیر است:

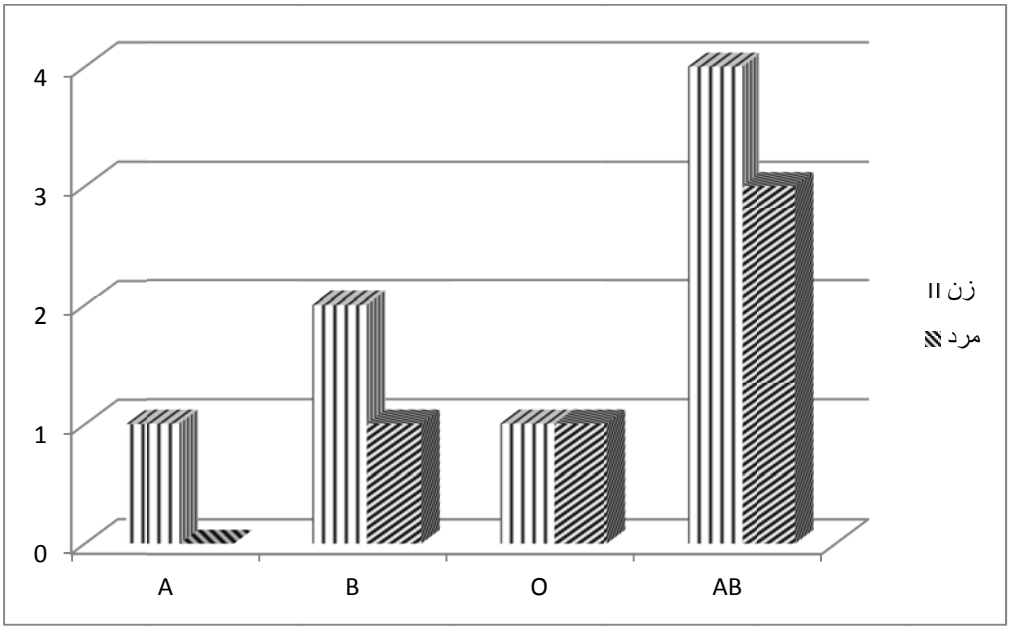
نمودار ستونی یا میله ای ساده

نمودار ستونی ساده را در یک دستگاه مختصات که محور افقی آن نشان‌دهنده کیفیت مشاهده‌ها و محور عمودی آن نشان‌دهنده فراوانی مطلق یا نسبی هر گروه است، رسم می‌کنند. برای مثال به نمودار زیر توجه کنید:



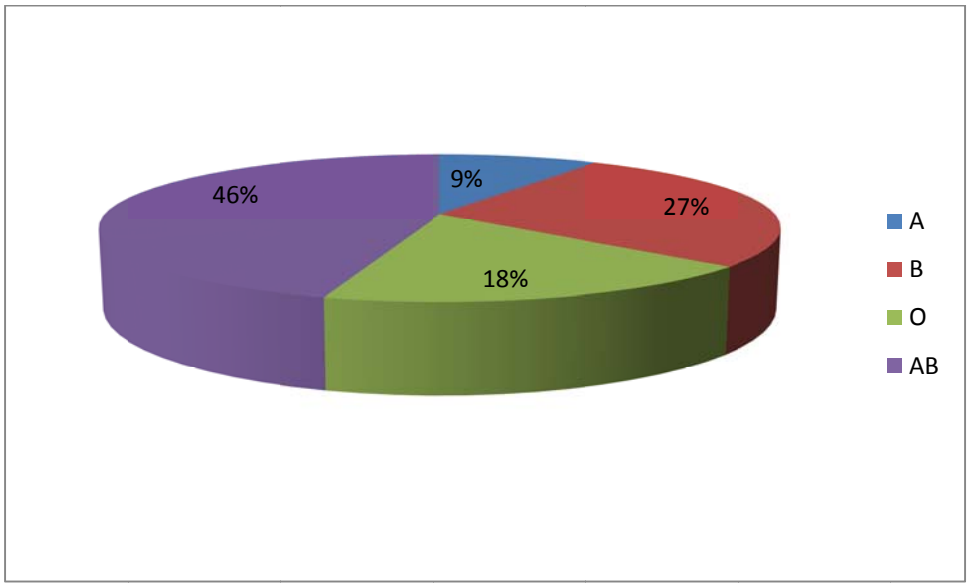
نمودار میله‌ای چندتایی

اگر پژوهشگر به دنبال بررسی بیش از یک صفت کیفی باشد، برای مقایسه آنها کاربرد نمودار میله‌ای ساده وقت‌گیر بوده و مقایسه‌ها را هم به وضوح نشان نمی‌دهد. در این صورت برای مقایسه، از نمودار میله‌ای چندتایی استفاده می‌کنند.



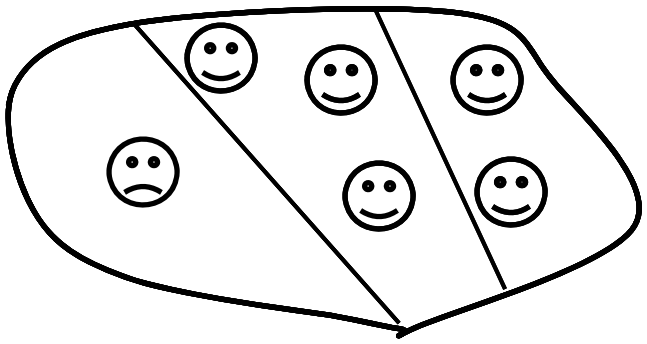
نمودار دایره‌ای یا کلوچه‌ای

به جای نمودار میله‌ای می‌توان از نمودار دایره‌ای یا کلوچه‌ای نیز استفاده کرد.



نمودارهای تصویری

طریقه دیگر نمایش داده‌ها استفاده از نمودارهای تصویری می‌باشد که در آن اندازه یا تعداد تصاویر نشان‌دهنده فراوانی داده‌ها می‌باشد. نمودارهای تصویری را ترجیحا با افزایش تعداد نمایش می‌دهیم تا از طریق بزرگ کردن تصویر. بیشترین کاربرد این نمودارها روی نقشه می باشد. مثلا برای نمایش جمعیت بیشتر یک شهر از تعداد بیشتری شکل شماتیک آدم استفاده میشود.



خصوصیات یک نمودار مناسب

۱) نمودار باید خود کاملا گویا بوده و بدون داده‌های جانبی قابل تفسیر باشد.

۲) عنوان نمودار باید کاملا واضح و روشن بوده و معمولا در پایین نمودار نوشته می‌شود.

۳) واحدهای اندازه‌گیری محورهای افقی و عمودی باید در کنار محورها با علائم خودشان ذکر شوند.

۴) نباید اطلاعات زیاد را در یک نمودار گنجاند.

۵) نمایش اطلاعات تفصیلی و جزئیات را باید روی جداول نشان داد نه نمودار.

۶) بهتر است تا جای ممکن از به کار بردن متن در نمودارها خودداری شود.

۳) ابزار شاخص‌های تمایل مرکزی

اگر تعداد عناصر تشکیل دهنده یک جامعه آماری کم باشد ، در آن صورت پژوهشگر همه اطلاعات موجود در واحدهای جامعه را جمع‌آوری کرده و بر اساس اطلاعات جمع‌آوری شده، ویژگی‌های مورد مطالعه را توصیف می‌کند. اما در پژوهش‌های علوم زیستی تعداد عناصر تشکیل دهنده جامعه بسیار زیاد بوده و لازم است قسمتی از جامعه، به عنوان نمونه انتخاب شود. به هر حال در تمام موارد، اگر پژوهشگر بتواند به جای استفاده از اطلاعات موجود در تک‌تک داده‌های جمع‌آوری شده، شاخص یا شاخص‌هایی که قادرند اطلاعات موجود در مجموعه داده‌ها را به خوبی توصیف کنند، معرفی نماید، سرعت انتقال اطلاعات بیشتر می‌شود. از این‌رو، شاخص‌های تمایل مرکزی معرفی می‌شوند.

مهمترین موضوع در مطالعه هر جامعه آماری، تعیین مقدار مرکزی است، یعنی تعیین مقداری که داده‌ها در اطراف آن توزیع شده باشند. یک شاخص مرکزی خوب، باید بر مبنای تمامی داده‌ها محاسبه شود ویا به وسیله نمونه‌گیری تصادفی دستخوش تغییرات زیادی نشود. شاخص‌های مرکزی معروف در آمار توصیفی، عبارتند از میانگین، میانه و نما.

میانگین

متداول‌ترین شاخص مرکزی که در آمار به کار می‌رود، میانگین است. اگر داده‌ها بر روی یک محور به صورت منظم ردیف شوند، مقدار میانگین، به طور دقیق در نقطه تعادل یا ثقل توزیع قرار می‌گیرد. میانگین با توجه به نوع متغیر اندازه‌گیری شده، دارای انواعی به شرح زیر است:

میانگین حسابی

یکی از مهمترین معیارهای تمایل مرکزی، میانگین حسابی است که به وسیله دو علامت μ برای میانگین جامعه و $\bar{X}$ برای میانگین نمونه، نشان داده می‌شود. مقدار میانگین حقیقی جامعه، ثابت بوده و تابع تغییرات نیست، در حالی که میانگین نمونه، متغیر و مقدار آن در نمونه‌گیری‌های مختلف از یک جامعه از نمونه‌ای به نمونه دیگر تغییر می‌کند. اگر میانگین، از جامعه بدست آید، پارامتر و اگر از نمونه به‌دست آید، به آن آماره می‌گویند. میانگین حسابی از فرمول زیر تعیین می‌شود:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

که در آن X_i: مقادیر هر مشاهده و N: تعداد مشاهده‌هاست.

میانگین وزنی

میانگین ساده حسابی وقتی استفاده می‌شود که داده‌ها اهمیت یکسانی داشته باشند، اما اگر داده‌ها دارای اهمیت مساوی نباشند، به هر یک از داده‌ها، وزنی به تناسب اهمیت آنها اختصاص می‌دهند و میانگین وزنی را حساب می‌کنند. اگر N مشاهده X_1، X_2، X_3 ... و X_n دارای وزن‌های W_1، W_2، W_3 ... و W_n باشند، در آن صورت میانگین وزنی به صورت زیر محاسبه خواهد شد:

$$\bar{X} = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

که در آن X_i: مقدار هر مشاهده و W_i: وزن هر یک از مقادیر است.

مثال: یک جیره غذایی به وزن ۵۰۰ گرم، مخلوطی از چهار ماده A، B، C و D است که در جدول زیر مقادیر هر یک و درصد پروتئین آنها آمده است، متوسط پروتئین این جیره غذایی را محاسبه کنید؟

ماده غذایی	وزن (گرم)	پروتئین (درصد)	W_iX_i
A	۱۵۰	۸	۱۲۰۰
B	۵۰	۱۲	۶۰۰
C	۲۰۰	۹	۱۸۰۰
D	۱۰۰	۱۰	۱۰۰۰
مجموع	۵۰۰	-	۴۶۰۰

$$\bar{X} = \frac{۱۲۰۰ + ۶۰۰ + ۱۸۰۰ + ۱۰۰۰}{۱۵۰ + ۵۰ + ۲۰۰ + ۱۰۰} = \frac{۴۶۰۰}{۵۰۰} = ۹.۲$$

با استفاده از فرمول زیر میانگین را در جدول فراوانی محاسبه می‌کنیم

$$\bar{X} = \frac{\sum_{i=۱}^{k} f_i x_i}{\sum_{i=۱}^{k} f_i}$$

میتوان برای سهولت در محاسبات ستون جدید تحت عنوان $f_i x_i$ در جدول فراوانی ایجاد شود. فراوانی مربوط به هر طبقه را در داده آن طبقه ضرب نموده و سپس حاصل جمع این ستون را به دست آورده و بر تعداد کل داده‌ها تقسیم می‌کنیم تا میانگین حاصل شود.

مثال: در مثال تعداد فرزندان ۱۵۰ خانواده در یکی از شهرهای ایران میانگین را حساب می‌کنیم.

x_i	f_i	$f_i x_i$
۰	۱۵	۰
۱	۳۰	۳۰
۲	۳۵	۷۰
۳	۳۰	۹۰
٤	۲۰	۸۰
٥	۱۵	۷۵
٦	٥	۳۰
مجموع	۱۵۰	۳۷۵

$$\sum_{i=۱}^{۷} x_i f_i = ۳۷۵$$

$$\sum_{i=۱}^{۷} f_i = ۱۵۰$$

$$\bar{X} = \frac{۳۷۵}{۱۵۰} = ۲.۵$$

در صورتی که داده‌ها در یک جدول فراوانی طبقه‌بندی شده باشند و داده‌های خام در دسترس نباشد برای به دست آوردن میانگین داده‌ها ابتدا نماینده هر طبقه که عبارت است از معدل هر طبقه، را به دست می‌آوریم (مثلا برای دسته اول $\frac{۵۰.۵+۵۸.۵}{۲} = ۵۴.۵$) سپس از فرمول فوق میانگین را محاسبه می‌کنیم.

مثال: برای ۱۳ دانشجو میانگین وزن را محاسبه می‌کنیم.

دسته ها	Xi	Fi	XiFi
50.5-58.5	54.5	3	163.5
58.5-66.5	62.5	4	250
66.5-73.5	70.5	1	70.7
73.5-81.5	77.5	5	387.5
مجموع		13	871.5

$$\bar{X} = \frac{\sum_{i=۱}^{۷} f_i x_i}{۱۳}$$

$$\bar{X} = \frac{۸۷۱.۵}{۱۳} = ٦۷.۰۳$$

میانه

اگر داده‌ها را به صورت صعودی یا نزولی مرتب کنند، میانه، عددی خواهد بود که داده‌ها را به دو قسمت مساوی تقسیم می‌کند، به‌طوری که ۵۰ درصد داده‌ها بزرگتر از آن و ۵۰ درصد کوچکتر از آن باشند. اگر تمامی اعداد در مجموعه داده‌ها از ۱ تا n برحسب بزرگی رتبه‌بندی شوند، اگر مجموعه آنها فرد باشد میانه (Md)، عبارت از مقداری است که رتبه آن $\frac{n+1}{2}$ باشد. اگر مجموعه داده‌ها زوج باشد، در این صورت میانه، مساوی متوسط دو مقدار میانی، یعنی مساوی میانگین مقادیری است که رتبه آنها $\frac{n}{2}$ و $\frac{n}{2}+1$ است.

مثال: میانه اعداد ۱۴، ۱۷، ۱۳-، ۴۱ و ۱۲ را مشخص کنید.

با مرتب کردن اعداد به ترتیب صعودی و دادن رتبه به آنها، خواهیم داشت:

مقدار:	-۱۳	۱۲	۱۴	۱۷	۴۱
رتبه:	۱	۲	۳	۴	۵

میانه عددی با رتبه $\frac{5+1}{2} = \frac{n+1}{2} =$ ۳ است، در نتیجه Md=۱۴ خواهد بود.

به طور کلی از میانه به عنوان اندازه تمایل به مرکز توزیع‌هایی استفاده می‌شودکه شامل مقادیر انتهایی، یعنی مقادیری در حد بسیار بالا یا بسیار پایین هستند. از نظر علمی نیز میانه کمتر تحت تاثیرداد های واقع شده در انتهای توزیع قرار می‌گیرد. اگر صفت مورد مطالعه، کیفی باشد، به‌دست آوردن میانه به عنوان یک پارامتر مرکزی مفید خواهد بود. در هنگام میانگین گیری دو عدد میانی در مورد تعیین میانه مجموعه داده های کیفی با تعداد زوج، اگر مقدار میانگین عدد کاملی نبود آنرا به نزدیکترین عدد نسبت می‌دهیم.

برای محاسبه میانه در جداول فراوانی از ستون فراوانی تجمعی استفاده می‌کنیم. اولین طبقه‌ای که فراوانی تجمعی آن نصف تعداد داده‌ها و یا بیشتر از آن باشد، طبقه‌ای است که میانه در آن قرار دارد. میانه را در داده های طبقه بندی شده با استفاده از رابطه زیر بدست می‌آوریم:

$$Md = L_{۰.۵} + \frac{\frac{n}{۲} - fC_{i-۱}}{f_i} R$$

که در آن:

$L_{۰.۵}$: حد پایینی طبقه‌ای است که میانه در آن قرار دارد. (اول باید طبقه میانه را مشخص کنیم)

$fC_{i-۱}$: فراوانی تجمعی مطلق یک طبقه قبل از طبقه‌ای که میانه در آن قرار دارد.

f_i : فراوانی مطلق طبقه‌ای که میانه در آن قرار دارد.

R : فاصله طبقات

مثال: میانه را برای داده‌های زیر بدست می‌آوریم:

دسته ها	Fi	Fc
50.5-58.5	3	3
58.5-66.5	4	7
66.5-73.5	5	12
73.5-81.5	1	13
مجموع	13	

چون تعداد داده‌ها فرد (۱۳) می‌باشد لذا داده میانه برابر خواهد بود با ($\frac{۱۳+۱}{۲} = ۷$) داده شماره ۷ که از روی F_c قابل تشخیص می‌باشد، یعنی طبقه میانه طبقه دوم خواهد بود (۵۸.۵-۶۶.۵) (برای توضیح بیشتر به تیتر "کاربرد جداول فراوانی" مراجعه شود)

$$Md = ۵۸.۵ + \left(\frac{\left(\frac{۱۳}{۲}\right) - ۳}{٤} \times ۸ \right) = ٦٥.٥$$

نما

نما (Mo) در یک توزیع آماری مقداری است که بیشترین تکرار را در میان مشاهده‌ها داشته باشد. از این شاخص برای نمایش داده‌هایی استفاده می‌کنند که نسبت به دیگر مقادیر، تکرار بیشتری دارند. نما یک شاخص مرکزی بسیار متغیر و ناپایدار است که عملیات ریاضی را بر روی آن نمی‌توان انجام داد. در پژوهش‌هایی که هدف تشخیص برتری یک داده یا مشاهده باشد، نما مفیدتر از میانگین و میانه است.

مثال: برای مشاهده‌های ۱۵، ۷، ۲، ۱۵، ۱۰، ۸، ۵ و ۲ نما را تعیین کنید.

با توجه به این که ۲ و ۱۵ نسبت به دیگر مشاهده‌ها، بیشتر تکرار شده‌اند، نماهای مشاهده‌ها ۲ و ۱۵ خواهند بود، یعنی دو نما برای مشاهده‌ها وجود خواهد داشت. داده‌ها می‌توانند نما نداشته باشند یا چند نما داشته باشند.

برای محاسبه نما در جدول فراوانی نیز ابتدا طبقه‌ای که دارای بیشترین فراوانی است در نظر می‌گیریم سپس با استفاده از فرمول زیر، نما را بدست می‌آوریم. توجه شود که تشخیص طبقه مد ساده می‌باشد اما نمی‌دانیم که دقیقا نما مربوط به کدام داده است.

$$Mo = L_{mo} + \frac{d_1}{d_1 + d_2} R$$

که در آن:

L_{mo} : حد پایینی طبقه‌ای که دارای بیشترین فراوانی است .(حد پایین طبقه مد)

d_1 : تفاوت فراوانی مطلق طبقه‌ای که دارای بیشترین فراوانی است با طبقه قبل.

d_2 : تفاوت فراوانی مطلق طبقه‌ای که دارای بیشترین فراوانی است با طبقه بعد.

R : فاصله طبقات

مثال: در مثال قبل طبقه‌ای که دارای بیشترین فراوانی است طبقه ۷۳.۵-۸۱.۵ است بنابراین داریم:

$$Mo = ٦٦.٥ + \left(\frac{١}{١ + ٤} * ٨\right) = ٦٨.١$$

مقایسه میانگین، میانه و نما

از بین شاخص‌های تمایل مرکزی، اهمیت میانگین در بررسی‌های آماری بیشتر است. میانگین برای توصیف متغیرهای نسبتی و فاصله‌ای که از پراکندگی متعادلی برخوردار باشند، مناسب‌ترین شاخص مرکزی خواهد بود، اما برای توصیف متغیرهای رتبه‌ای و اسمی کارایی ندارد. برای توصیف متغیرهای رتبه‌ای، میانه و نما را می‌توان به‌کار برد. برای متغیرهای اسمی تنها از نما استفاده می‌کنند. وقتی گروه‌های اول و آخر باز باشند نیز میانه معیار مرکزی خوبی می‌باشد.

۴) ابزار شاخص‌های پراکندگی

شاخص‌های تمایل مرکزی، فقط بخشی از اطلاعات لازم را برای مجموعه داده‌ها ارائه می‌دهند.

مثال:فرض کنید اعداد زیر نمره‌های دو کلاس در یک آزمون آمار زیستی باشند:

کلاس یک: ۲و ۳و ۵و ۶و ۱۴و ۱۶و ۱۷و ۱۸و ۱۹و ۲۰

کلاس دو: ۹و ۱۰و ۱۰و ۱۱و ۱۱و ۱۲و ۱۳و ۱۴و ۱۵و ۱۵

با وجود اینکه میانگین برای هر دو کلاس ۱۲ می‌باشد نحوه پراکندگی داده‌ها در اطراف ۱۲ متفاوت است در کلاس یک پراکندگی داده‌ها زیاد و اغلب داده‌ها از ۱۲ دور می‌باشند در حالی که در کلاس دو پراکندگی داده‌ها کم و داده‌ها نزدیک به ۱۲ می‌باشند.

مثال بالا نشان می‌دهد که معیارهای تمایل مرکزی به تنهایی نمی‌توانند معرف کاملی از داده‌ها باشند و لازم است معیارهایی برای سنجش پراکندگی نیز تعریف شوند.

شاخص‌های پراکندگی برای توصیف بهتر متغیرها به کار می‌روند. شاخصی برای تعیین تغییرات مطلوب است که نخست در آن از تمام مشاهده‌ها استفاده شود و دوم این که اندازه تغییرات مشاهده‌ها را بر حسب دوری یا نزدیکی به میانگین نشان دهد. شاخص‌های پراکندگی، انواعی دارند که به ترتیب معرفی می‌شوند:

دامنه تغیرات

یکی از ساده‌ترین شاخص‌های پراکندگی، دامنه تغییرات است. این شاخص از تفاضل کوچکترین مشاهده از بزرگترین مشاهده، محاسبه می‌شود. تعریف دامنه تغییرات(که با R نشان داده می‌شود) چنین است:

$$R= Max(X_i) - Min(X_i)$$

به عنوان مثال ممکن است بخواهیم میانگین وزن دانشجویان دو کلاس ۱ و ۲ را بدست آوریم. مثلا در کلاس ۱ ، میانگین ٦٥ کیلوگرم بدست آمده و رنج حداقل و حداکثر وزن دانشجویان این کلاس ٥٥ و ٧٥ کیلوگرم می‌باشد، میانگین وزن دانشجویان کلاس ۲ نیز ٦٥ کیلوگرم محاسبه شده در حالیکه رنج حداقل و حداکثر وزن این کلاس ٤٥ و ٨٥ کیلوگرم می‌باشد. همانطوریکه دیده می‌شود با توجه به اینکه میانگین وزن دو کلاس با هم برابر است اما دامنه تغییرات دو کلاس با هم متفاوت است.

دامنه تغییرات با وجود سادگی در محاسبه، از ثبات چندانی برخوردار نیست، زیرا در تعیین این شاخص، از مجموع مشاهده‌ها، فقط از کوچکترین و بزرگترین آنها استفاده می‌شود. از این شاخص برای مواردی که به شاخص فوری نیاز باشد یا هنگامی که تعیین بزرگترین و کوچکترین مقدار داده‌ها و فاصله بین آنها مورد نظر باشد، استفاده می‌شود.

مثال: برای داده‌های زیر دامنه تغییرات را مشخص کنید.

$$X_i = ١٥، ١٦، ١٧، ١٧، ٢٠، ٢٠، ٢٢، ٦٠$$

دامنه تغییرات در این مثال، مساوی ۴۵ است که بسیار بزرگتر از پراکندگی واقعی داده‌هاست. دلیل این است که مقدار بزرگترین داده با دیگر داده‌ها، تفاوت چشمگیری دارد و سبب می‌شود پراکندگی بزرگتر از معمول جلوه کند.

متوسط انحراف از میانگین

شاخص دامنه تغییرات قادر به بیان تمامی تغییرات نیست، پس باید به دنبال شاخصی بود که تغییرات تمامی داده‌ها را اندازه‌گیری کند. طبیعی است تغییر، زمانی مفهوم پیدا می‌کند که هر یک از داده‌ها نسبت به یک مبدا مقایسه شوند. به جرات می‌توان گفت که بهترین مرکز برای داده‌ها، میانگین است. از ویژگیهای میانگین آن است که مجموع انحراف داده‌ها از میانگینشان همیشه مساوی صفر است. واضح است که آنچه موجب صفر شدن انحراف‌ها می‌شود، خنثی شدن انحراف‌های منفی با انحراف‌های مثبت است. از آنجا که انحراف‌های مثبت و منفی از نظر کاربردی یکسان هستند، پس می‌توان از قدرمطلق انحراف‌ها به‌صورت زیر استفاده کرد:

$$|e_i| = |x_i - \mu|$$

در محاسبه قدرمطلق توجه کنید که اگر علامت منفی یک عدد منفی برداشته شود، قدرمطلق آن حاصل می‌شود و قدرمطلق عدد مثبت، خود عدد می‌شود مثلا $|-۲| = ۲$ و $|۴| = ۴$.

بر این اساس می‌توان به شاخص جدیدی از پراکندگی، با عنوان متوسط انحراف از میانگین رسید، یعنی:

$$A.D = \frac{\sum|x_i - \mu|}{n}$$

که در آن A.D: انحراف متوسط از میانگین، x_i: مقادیر مشاهده‌ها، μ: میانگین و n:تعداد مشاهده‌هاست.

برای مثال اگر وزن تک‌تک دانشجویان کلاس ۱ را از ٦٥ (میانگین وزن کلاس ۱) کم کنیم، در حقیقت انحراف تک‌تک داده‌ها از میانگین را بدست آورده‌ایم که مقادیر مثبت و منفی می‌باشند، حال اگر قدرمطلق این انحرافها را با هم جمع کرده و بر تعداد انحرافها (تعداد کل داده‌ها) تقسیم کنیم، متوسط انحراف از میانگین کلاس ۱ بدست می‌آید. اگر متوسط انحراف از میانگین کلاس ۲ بیشتر از کلاس ۱ باشد یعنی پراکندگی وزن دانشجویان کلاس ۲ از کلاس ۱ بیشتر است.

برای محاسبه میانگین انحرافها در جدول فراوانی از فرمول زیر استفاده می‌کنیم :

$$M.D. = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{x}|}{\sum_{i=1}^{k} f_i}$$

از این شاخص، هنگامی که در محاسبه انحراف‌ها، تعداد اندکی انحراف زیاد و در مقابل تعداد زیادی انحراف اندک موجود باشد، استفاده می‌شود، زیرا اگر از انحراف معیار استفاده شود، با توجه به این که انحراف میانگین مجذور می‌شود، شاخص انحراف معیار، انحراف‌ها بیش از مقدار واقعی نشان می‌دهد.

انحراف معیار و واریانس

انحراف معیار امروزه به عنوان معتبرترین و پر استفاده‌ترین شاخص پراکندگی به کار می‌رود. این شاخص افزون بر این که از درجه اعتبار بالایی برخوردار است، بسیاری از معایبی را که برای شاخص‌های قبلی برشمرده شد، ندارد. مفاهیم آماری دیگر بر اساس این شاخص، پایه‌گذاری شده است. فرمول محاسبه انحراف معیار جامعه که آن را با σ نشان می‌دهند چنین است:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)}{n}}$$

فرمول محاسبه واریانس جامعه (σ^2) به‌صورت زیر است:

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

برای محاسبه انحراف معیار (S) و واریانس (S^2) نمونه در مخرج کسر فرمول‌های بالا n-۱ قرار می‌گیرد.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

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

در فرمول‌های بالا، X_i: مقدار هر کدام از مشاهده‌ها، $\bar{X}$: میانگین نمونه، μ: میانگین جامعه و n: تعداد مشاهده‌هاست.

اگر تعداد مشاهدات زیاد باشد استفاده از دستور بالا خسته کننده خواهد بود با استفاده از جبر مقدماتی فرمول واریانس به صورت زیر که از نظر محاسبه ساده‌تر است درمی‌آید:

$$S^{۲} = \frac{n\sum_{i=۱}^{n} x_i^{۲} - [\sum_{i=۱}^{n} x_i]^{۲}}{n(n-۱)}$$

مثال: واریانس داده‌های زیر را با استفاده از فرمول اصلی ساده شده واریانس بدست آورید:

x_i	۱	۷	۰	۳	۹
$x_i - \bar{x}$	۱-٤=-۳	۷-٤=۳	۰-٤=-٤	۳-٤=-۱	۹-٤=٥
$(x_i - \bar{x})^{۲}$	$(-۳)^{۲} = ۹$	$۳^{۲} = ۹$	$(-٤)^{۲} = ۱٦$	$(-۱)^{۲} = ۱$	$٥^{۲} = ۲٥$

$$\bar{x} = ٤$$

با استفاده از فرمول اصلی:

$$S = \sqrt{\frac{٦۰}{٤}} = \sqrt{۱٥} = ۳.۸۷ \qquad S^{۲} = (۳.۸۷)^{۲} = ۱٥$$

با استفاده از فرمول ساده شده:

$$\sum_{i=۱}^{n} x_i = ۱ + ۷ + ۰ + ۳ + ۹ = ۲۰$$

$$\sum_{i=۱}^{n} x_i^{۲} = ۱^{۲} + ۷^{۲} + ۰^{۲} + ۳^{۲} + ۹^{۲} = ۱٤۰$$

$$S^{۲} = \frac{n\sum_{i=۱}^{n} x_i^{۲} - [\sum_{i=۱}^{n} x_i]^{۲}}{n(n-۱)} = \frac{۵(۱٤۰) - (۲۰)^{۲}}{۵ \times ٤} = \frac{۷۰۰ - ٤۰۰}{۲۰} = ۱۵$$

S=۳.۸۷۳

برای محاسبه واریانس در جدول فراوانی از فرمول زیر استفاده می‌کنیم :

$$S^{۲} = \frac{\sum_{i=۱}^{k} f_i[x_i - \bar{x}]^{۲}}{\sum_{i=۱}^{k} f_i - ۱}$$

با استفاده از بسط $(x_i - \bar{x})^{۲}$ می‌توان فرمول واریانس را به صورت ساده‌تر نوشت که محاسبات آنرا آسان‌تر می‌کند :

$$S^{۲} = \frac{\sum_{i=۱}^{k} f_i x_i^{۲} - (\bar{x})^{۲}\sum_{i=۱}^{k} f_i}{\sum_{i=۱}^{k} f_i - ۱}$$

ضریب تغییرات

انحراف معیار به عنوان یک شاخص پراکندگی مناسب در آمار به کار برده می‌شود، ولی از آنجا که این شاخص مرکزی نیز، متاثر از طبیعت داده‌ها و به ویژه واحدهای اندازه‌گیری است، متاسفانه اگر پژوهشگر هوشیارانه عمل نکند، می‌تواند تصمیم‌گیری را دچار اشتباه کند. یکی از دلایل این امر بدان خاطر صورت می‌گیرد که مثلا دو نوع متغیری که وارد محاسبه می‌شوند با واحد مختلف اندازه‌گیری شوند. برای مثال ممکن است برای جمعیت معین بررسی کنیم که آیا تغییرات قد یک جامعه متغیرتر از وزن آنها است، که یکی بر حسب متر و دیگری بر حسب کیلوگرم اندازه‌گیری شده است.

علاوه بر آن ممکن است واحدهای اندازه‌گیری یکسان بکار می‌روند اما میانگینها کاملا متفاوت باشند. اگر انحراف معیار وزن کودکان کلاس اول دبستان را با انحراف استاندارد وزن دانش‌آموزان سال اول دبیرستان مقایسه کنیم، ممکن است انحراف معیار دومی از نظر عددی بزرگتر از اولی باشد زیرا وزنها خود بزرگتر هستند، نه به خاطر آنکه تغییرات بیشتر است.

برای دوری از هر گونه خطا، شاخص پراکندگی نسبی در آمار وجود دارد که به آن ضریب تغییرات (CV) می‌گویند. این شاخص نشان می‌دهد که میِزان پراکندگی، چند درصد میانگین است. از این ضریب در مواردی که بخواهند تغییرپذیری دو یا چند گروه را با هم مقایسه کنند، استفاده می‌شود. اگر انحراف معیار هر توزیع (σ) به میانگین آن (μ) تقسیم شود، انحراف معیار نسبی با ضریب تغییرات (CV) به دست می‌آید.

$$CV = \left(\frac{\sigma}{\mu}\right)۱۰۰$$

از آنجایی که میانگین و انحراف معیار هر دو با یک واحد بیان می‌شوند، در نتیجه ضریب تغییرات به واحد اندازه‌گیری بستگی ندارد.

مثال: در جدول زیر میانگین و انحراف معیار وزن یک گرو ه از جوانان ۲۰ ساله و یک گروه از نوجوانان ۱۴ ساله گردآوری شده است، تغییرات وزن دو گرون را مقایسه نمایید.

سن	۲۰ سال	۱۴ سال
میانگین وزن	۱۴۵ پوند	۳۲ کیلوگرم
انحراف استاندارد	۱۰ پوند	۴ کیلوگرم

جوانان ۲۰ ساله $C.V. = \frac{۱۰}{۱۴۵} \times ۱۰۰ = ۶.۹$

جوانان ۱۴ ساله $C.V. = \frac{۴}{۳۲} \times ۱۰۰ = ۱۲.۵$

بنابراین تغییرات وزن در نوجوانان ۱۴ ساله بیشتر از تغییرات وزن در جوانان ۲۰ ساله است.

تمارین:

۱- یک جدول فراوانی حداقل دارای چه ستونهایی می‌باشد؟

۲- یک جدول فراوانی فرضی ساخته و در مورد هر کدام از ستونهای F_i ، F_c ، P_i و P_c یک نتیجه بنویسید.

۳- در مورد جدول فراوانی مسئله ۲ نمودارهای مناسب F_i و F_c آن را رسم کنید.

۴- در مورد جدول فراوانی مسئله ۲ یکی از شاخصهای مناسب تمایل مرکزی و پراکندگی را محاسبه کنید.

۵- ضریب تغییرات چه استفاده‌ای در آمار دارد.مثالی بزنید.

فصل سوم

احتمال

مقدمه

واژه احتمال بر "عدم اطمینان" دلالت می‌کند. احتمالات مبحثی از ریاضیات است که اصول نظریه آمار بر مبنای آن بنا شده است. افزون بر این از احتمالات در پیش‌بینی وقایع آینده بر مبنای اطلاعات و شواهد موجود استفاده می‌شود. برای مثال، می‌توان احتمال وقوع بارندگی، حضور یک گونه گیاهی و ... را تعیین کرد. آمار توصیفی معمولا به توصیف داده‌ها می‌پردازد که نمونه‌هایی از یک جامعه آماری هستند. تعمیم نمونه‌ها به جامعه که مربوط به آمار استنباطی می‌باشد همواره دارای مقداری خطا می‌باشد که برای اندازه‌گیری آن از تئوری احتمالات استفاده می‌شود.

پیش از تعریف احتمال, مثالی را در نظر می گیریم. در مطالعه تاثیر جوانه زدن بذر پنبه، لازم است بدانیم که درصد جوانه زدن یا قابلیت زنده ماندن بذر پنبه چقدر است. برای پی بردن به این نکته می توانیم با کاشت ۱۰۰ بذر و شمارش تعداد بذرهای جوانه زده، شروع نمائیم. اگر ۴۹ بذر جوانه بزند، یعنی اگر در بین ۱۰۰ حادثه ۴۹ موفقیت داشته باشیم(کلمه موفقیت در آمار برای وقوع حادثه مورد بحث به کار می رود)، می گوئیم که فراوانی نسبی موفقیت (تعداد موفقیت بخش بر تعداد آزمایش)برابر ۰.۴۹ است. حال اگر ۲۰۰ بذر بکاریم و از این تعداد بفرض ۱۰۳ بذر جوانه بزند، فراوانی نسبی برابر است با $\frac{۱۰۳}{۲۰۰} = ۰.۵۱۵$ خواهد بود. اگر ما این آزمایش را بارها تکرار کنیم و هر بار تعداد بیشتری بذر بکاریم، یک سری مقادیر از فراوانی ها ی نسبی متناظر به دست خواهد آمد که در جدول زیر آمده است. اگر این فراوانی های نسبی به یک حد نزدیک شوند، این مقدار حدی را احتمال موفقیت در یک مرتبه آزمایش می نامیم. از داده های جدول زیر چنین بر می آید که فراوانی های نسبی به طرف مقدار ۰.۵۲ میل می نماید که این مقدار را به نام احتمال جوانه زدن یک بذر پنبه تیمار نشده خواهیم نامید.

تعداد آزمایش (n)	تعداد موفقیت (s)	فراوانی نسبی (s/n)
۱۰۰	٤۹	۰.٤۹
۲۰۰	۱۰۳	۰.۵۱۵
۵۰۰	۲٦۲	۰.۵۲٤
۱۰۰۰	۵۱۷	۰.۵۱۷
۱۰۰۰۰	۵۱۹۳	۰.۵۱۹۳

اکنون می توان احتمال را به صورت زیر در نظر گرفت.

تعریف ۱ : اگر تعداد موفقیت در n آزمایش را به وسیله s نشان دهیم و اگر سری فراوانی های نسبی s/n که برای مقادیر بیشتر و بیشتر (تعداد آزمایش‌های بیشتر) n به دست می آید به سمت یک حد میل کند، این حد را "احتمال موفقیت در یک آزمایش" می نامند. به این مفهوم قانون اعداد بزرگ می‌گویند، اگر تعداد آزمایشها به سمت بی‌نهایت میل کند فراوانی نسبی به حدی نزدیک می‌شود که همان احتمال رخ دادن پیشامد مورد نظر است یعنی :

$$P = \lim_{n\to\infty} \frac{F_i}{\sum_{i=1}^{n} F_i}$$

در عمل، احتمال را برابر فراوانی نسبی که برای بیشترین تعداد آزمایش بدست آمده باشد در نظر می‌گیرند. بنابراین، برای اعداد جدول فوق اگر آزمایش برای ۱۰۰۰۰ بذر پنبه به کار رفته باشد. در این صورت فراوانی نسبی ۰.۵۱۹۳ را احتمال جوانه‌زدن یک بذر پنبه می‌توان در نظر گرفت.

نکته جالب این است که بدانید این حد بدست آمده در تکرار آزمایشات زیاد هیچ‌گاه کاملا ثابت نمی‌شود و همواره نوساناتی دارد. بدین دلیل محاسبات احتمال با فرضیات ریاضی کاملا سازگار نیست.

این تعریف احتمال، یعنی در نظر گرفتن احتمال به عنوان فراوانی نسبی موفقیت در تعداد زیاد ولی محدود آزمایش‌ها، برای مقاصد دیگر مثلا برای فهم و اثبات قضایای احتمالات مناسب خواهد بود. به این نوع احتمال که مبنای مشاهداتی دارد احتمال عینی گفته میشود.

احتمال عینی را برای هر ویژگی خاصی می‌توان محاسبه نمود ولی باید توجه داشت که مقدار این احتمال نسبت به شرایط مختلف زمانی و مکانی فرق می‌کند و در نتیجه میزان استفاده از آن بسیار محدود است. به عبارت دیگر این روش احتمال وقوع ویژگی خاصی را در یک زمان و مکان معین (شرایط خاص) محاسبه می‌نماید از طرف دیگر این پیشامدها هم ارز نیستند یعنی امکان وقوع پیشامد اول با امکان وقوع پیشامدهای دیگر برابر نیست و تحت تاثیر عوامل مختلفی قرار دارند. ناگفته نماند چنین احتمالی در مواردی خاص دارای ارزش بسیار فراوانی است برای مثال اگر احتمال عینی وقوع یک بیماری در یک زمان در دو نقطه مختلف کشور فرق داشته باشد. چون احتمال وقوع موفقیت قبل از انجام آزمایش مشخص نیست به این نوع احتمالات، احتمال پسین می‌گویند.

اگر قادر باشیم احتمال وقوع یک ویژگی خاص را قبل از آنکه آن حادثه اتفاق بیفتد محاسبه نماییم آن را احتمال اولیه، پیشین یا کلاسیک گوییم. برای مثال می‌دانیم که اگر یک سکه سالم را پرتاب نماییم احتمال شیر آمدن برابر است با $\frac{۱}{۲}$ یا اگر یک تاس را پرتاب نماییم احتمال اینکه عدد ۳ بدست آید برابر است با $\frac{۱}{۶}$. علت این نتیجه‌گیری آن است که در پرتاب یک سکه کاملا متعادل فقط امکان رخ دادن دو حادثه، یعنی شیر یا خط وجود دارد و هر دو واقعه دارای احتمال وقوع مساوی (هم شانس یا هم تراز، یعنی امکان وقوع پیشامد اول برابر امکان وقوع پیشامدهای دیگر است) هستند. نکته بسیار مهم آن است که در تعریف احتمال عینی هر چقدر تعداد دفعات آزمایش را بیشتر کنیم مقدار احتمال عینی (نسبت پیشامد ویژگی) به احتمال کلاسیک یا اولیه نزدیکتر می‌گردد.

برای چنین حالاتی که در آنها موفقیت‌ها یا عدم موفقیت‌ها فقط به صورت چند حادثه هم‌تراز اتفاق می‌افتند، تعریف زیر مناسب است:

تعریف ۲ : اگر حادثه‌ای بتواند به S صورت رخ دهد و به f صورت رخ ندهد، و اگر احتمال وقوع هر یک از این S+f صورت‌ها مساوی باشد، "احتمال موفقیت (پیشامد حادثه) در یک بار آزمایش" برابر است با:

$$P = \frac{S}{S+f}$$

و احتمال عدم وقوع حادثه برابر است با:

$$q = \frac{f}{S + f}$$

واضح است که تعریف ۲ حالتی خاص از تعریف ۱ است. بنابراین، احتمال آمدن یک شیر در پرتاب یک سکه کاملا متعادل را می‌توان از هر یک از این تعاریف به دست آورد. بدیهی است که به دست آوردن احتمال از تعریف دوم خیلی آسانتر است، چون در یک مرتبه پرتاب یک سکه، یک شیر می‌تواند فقط به S=۱ صورت رخ دهد و یک خط می‌تواند فقط به f=۱ صورت اتفاق افتد و هر دو صورت دارای احتمال وقوع مساوی می‌باشند. بنابراین، داریم:

$$P = \frac{۱}{۱ + ۱} = \frac{۱}{۲}$$

اگر تعریف ۱ را به کار ببریم، ناچار خواهیم بود که سکه را به دفعات زیاد آزمایش نموده و تعداد شیرها را یادداشت نمائیم. مسلم است که نسبت تعداد شیرها به تعداد دفعات پرتاب سکه برای تعداد بیشتر و بیشتر آزمایش به عدد $\frac{۱}{۲}$ نزدیک می‌شود.

با این مقدمه در ادامه مفاهیم اساسی در احتمالات را تعریف می‌کنیم.

آزمایش: در نظریه احتمال فعالیتی که نتیجه آن از قبل مشخص نباشد، به "آزمایش" معروف است. برای مثال، با پرتاب سکه، شیر یا خط ظاهر می‌شود. انجام این کار یک آزمایش است.

فضای نمونه:

مجموعه پیامدهای ممکن یک آزمایش را فضای نمونه آن آزمایش گویند که آن را با S نمایش می‌دهند.

مثال: فضای نمونه پرتاب یک سکه را مشخص کنید.

اگر ظاهر شدن شیر را با H و خط را با T نشان دهند، در این صورت: S={H,T}

مثال: فضای نمونه پرتاب یک تاس را مشخص کنید. S={۱,۲,۳.۴,۵,۶}

پیشامد : در نظریه احتمال، پیشامد، یکی از زیرمجموعه‌های فضای نمونه است. در پرتاب یک سکه، خط آمدن یک پیشامد و شیر آمدن پیشامد دیگری است.

احتمال یک پیشامد: فرض کنید فضای نمونه‌ای وجود دارد که در آن، همه پیامدهای مقدماتی، از شانس مساوی برای انتخاب شدن برخوردارند(هم‌تراز). در این صورت احتمال وقوع پیشامد خاصی مانند A عبارت است از تعداد عضوهای پیشامد A به تعداد عضوهای فضای نمونه. اگر تعداد عضوهای فضای نمونه را با n(s) و تعداد عضوهای پیشامد A را با n(A) نشان دهند، احتمال A، یعنی P(A) به صورت زیر خواهد بود:

$$P(A) = \frac{n(A)}{n(S)}$$

مثال: احتمال بیرون آوردن مهره‌ای سفید، از ظرفی که محتوی ۵ مهره قرمز، ۴ مهره سفید و ۳ مهره سیاه است، را محاسبه کنید.

تعداد ۱۲ مهره در ظرف وجود دارد که ۴ عد سفید است. اگر پیشامد مورد نظر را با B نمایش دهند، آنگاه:

$$P(B) = \frac{٤}{١٢} = \frac{١}{٣}$$

در ادامه به برخی "اصول اساسی احتمالات" اشاره می‌شود.

الف- احتمال وقوع پیشامدی همانند A همواره بزرگتر یا مساوی صفر و کوچکتر یا مساوی یک است، ($٠ \leq P(A) \leq ١$) (از روی فراوانی نسبی P_i قابل اثبات است)

ب- مجموع احتمال تمام پیشامدهای ممکن فضای نمونه، همواره برابر یک است (P(S)=۱) (از روی فراوانی نسبی تجمعی P_c قابل اثبات است)

با توجه با اینکه در محاسبات احتمالات همواره در حال شمارش تعداد اعضای مجموعه فضای نمونه و مجموعه پیشامد (به عنوان زیرمجموعه‌ای از فضای نمونه) هستیم، لذا اطلاع کافی در زمینه نظریه مجموعه‌ها می‌تواند درک قوانین

احتمالات را ساده کند. بدین منظور جدول متقاطع زیر که گروه خونی افراد را بر حسب جنسیت تفکیک نموده است ارائه می‌شود.

گروه خونی	زن:D۱	مرد:D۲	مجموع
c۱:A	۱	۰	۱
c۲:B	۲	۱	۳
c۳:O	۰	۲	۲
c۴:AB	۳	۳	۶
مجموع	۶	۶	۱۲

در جدول بالا مجموعه‌های c۱ الی c۴ از افرادی تشکیل شده که به ترتیب دارای گروه های خونی AB,O,B,A می‌باشند و مجموعه‌های D۱ (زن) وD۲ (مرد) آنها را از لحاظ جنسیتی تفکیک می‌کند. با استفاده از مفاهیم اشتراک, اجتماع و مکمل در مورد مجموعه‌ها می‌توان مجموعه‌های دیگری را از مجموعه اصلی (مرجع) مشخص ساخت. برای مثال، مجموعه $C_۱ \cap D_۱$ (اشتراک) از زنهایی که گروه خونی A دارند به وجود آمده (یا به عبارت دیگر تعداد افراد گروه خونی A به شرطی که از زنها باشند) و $n(C_۱ \cap D_۱) = ۱$ است. مجموعه $C_۴ \cup D_۲$ (اجتماع) از افراد دارای گروه خونی AB یا افراد مرد یا هر دو تشکیل گردیده است و برابر با $n(C_۴ \cup D_۲) = ۶ + ۶ - ۳ = ۹$ می‌باشد. در محاسبه $n(C_۴ \cup D_۲)$ عدد ۳ را از مجموع دو عدد دیگر کم می‌کنیم، چون در شمارش تعداد عناصر موجود در اجتماع مجموعه‌های $A_۲$ و $B_۲$ دو بار شمارش شده است. مکمل $D_۲$ ، یعنی $\overline{D}_۲$ از افرادی که شامل مردان نمی‌شوند تشکیل گردیده که همان زنها می‌باشند و برابر با $n(\overline{D}_۲) = ۱۲ - ۶ = ۶$ خواهد بود. لازم به ذکر است که برای درک بهتر موارد مذکور از نمودار "ون" نیز می‌توان استفاده نمود که یک نمایش دو بعدی از مجموعه‌ها روی کاغذ به صورت دو بعدی می‌باشد.

ج- قانون جمع

قانون جمع زمانی به کار می‌رود که یکی از چند پیشامد اتفاق بیفتد، یعنی حادثه اول یا حادثه دوم یا ... . قبل از بیان قانون جمع، لازم است که پیشامدهای سازگار و ناسازگار تعریف شوند.

دو پیشامد A و B ناسازگارند، هرگاه برآمد مشترکی نداشته باشند $(A \cap B = \emptyset)$ و در نتیجه، وقوع یکی از آنها مانع رخ دادن دیگری باشد.یا اگر دو یا چند پیشامد چنان باشند که در هر یک آزمایش منفرد، بیش از یکی از آنها نتواند رخ بدهد، آنها را نسبت به هم ناسازگار یا"مانعه الجمع" می‌نامند. مانند پیشامد گروه خونی A یا مرد بودن که اشتراکی ندارند.

دو پیشامد A و B سازگارند، هرگاه برآمد مشترک داشته باشند $(A \cap B \neq \emptyset)$ و در نتیجه همزمان با یکدیگر هم بتوانند رخ دهند. مانند پیشامد گروه خونی AB با زن بودن که اشتراک دارند.

قانون جمع در مورد دو پیشامد ناسازگار A و B عبارتست از : $P(A \cup B) = P(A) + P(B)$

مثلا احتمال اینکه در جدول بالا فرد انتخاب شده مرد یا گروه خونی A باشد:

$$P(A \cup B) = \frac{٦}{١٢} + \frac{١}{١٢} = \frac{٧}{١٢} = ٠.٥٨$$

قانون جمع در مورد دو پیشامد سازگار A و B عبارتست از:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

مثلا احتمال این را حساب کنید که فرد انتخابی زن یا گروه خونی AB باشد:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B\ )$$

$$P(A\text{یا}B) = \frac{٦}{١٢} + \frac{٦}{١٢} - \frac{٣}{١٢} = \frac{٩}{١٢}$$

قانون جمع را به بیش از دو پیشامد نیز، می‌توان تعمیم داد. برای مثال، در مورد سه پیشامد ناسازگار A، B و C:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

در مورد سه پیشامد سازگار A، B و C:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) + P(A \cap B \cap C)$$

امید ریاضی

اگر p احتمال وقوع یک پیشامد در یک آزمایش منفرد باشد، "تعداد دفعات مورد انتظار وقوع" با "امید ریاضی" آن پیشامد در n آزمایش برابر np خواهد بود. همانطوریکه در قبل دیدیم احتمال سبز کردن بذر پنبه برابر با ۰.۵۱۳۹ محاسبه شده است. اگر ۱۰۰۰۰ بذر پنبه بکاریم تعداد بذرهای سبز شده برابر با ۱۰۰۰۰ × ۰.۵۱۹۳ یعنی ۰.۵۱۹۳ بذر خواهد بود. یا اینکه در جدول قبل دیدیم که احتمال گروه خونی AB برابر با $\frac{۶}{۱۲}$ می‌باشد. اگر تعداد کل افراد ۱۲ نفر باشند، ۱۲ × $\frac{۶}{۱۲}$ یا ۶ نفر آنها گروه خونی AB خواهند داشت.

مثالی دیگر: در ۹۰۰ مرتبه آزمایش پرتاب دو تاس، تعداد مورد انتظار برای حالتی که مجموع کمتر از ۵ باشد، چقدر است؟

برای پیدا کردن احتمال p برای به دست آوردن مجموعی کمتر از ۵ در یک مرتبه ریختن دو تاس

چون برای مجموع کمتر از ۵ حالتهای زیر میتواند برای دو تاس رخ دهد :

۱-۳یا ۱-۲ یا ۲-۳ یا ۴-۱ یا ۱-۱ یا۲-۲ لذا:

$p = \frac{۶}{۳۶} = \frac{۱}{۶}$ پس S+f=۳۶ ،S=۶

تعداد دفعات مورد انتظار در بین ۹۰۰ آزمایش برای حالتی که مجموع کمتر از ۵ باشد، عبارت است از ۱۵۰ = ۹۰۰ × $\frac{۱}{۶}$

به آسانی می‌توان دید که امید ریاضی محتمل‌ترین تعداد موفقیت برای وقوع در بین n آزمایش است. به این ترتیب، در مثال قبل، محتمل‌ترین پیشامد عبارت از آمدن ۱۵۰ مرتبه مجموع کمتر از ۵ می‌باشد. مع‌هذا، این مفهوم بدین معنا نیست که در ۹۰۰ مرتبه پرتاب دو تاس حتما باید این پیشامد اتفاق بیفتد. در واقع، حتی به این معنی هم نیست که این پیشامد با احتمال زیاد اتفاق می‌افتد. به طور مسلم، معنی آن این است که احتمال وقوع این پیشامد (آمدن ۱۵۰ مرتبه مجموع کمتر از ۵) از پیشامدهای دیگر بیشتر است. اگر احتمال p به صورت فراوانی نسبی در n مرتبه آزمایش تعیین شده باشد، به آسانی دیده می‌شود که در این n آزمایش امید ریاضی با مقدار واقعی موفقیت برابر است. به این ترتیب، اگر احتمال جوانه‌زنی بذرهای تیمار نشده پنبه، همان‌طور که قبلا دیده شد برابر ۰.۵۱۹۳ باشد، امید ریاضی جوانه‌زنی در ۱۰۰۰۰ بذر برابر با ۵۱۹۳ = ۱۰۰۰۰ × ۰.۵۱۹۳ می‌باشد که با تعداد واقعی جوانه‌های ظاهر شده برابر است.

د- احتمال شرطی:

گاهگاهی امکان دارد مجموعه همه نتایج ممکن زیر مجموعه‌ای از مجموعه مرجع را تشکیل دهد. به بیان دیگر، می‌توان جمعیت مساعد را با اعمال مجموعه شرایطی که بر کل جمعیت قابل اعمال نیست، کاهش داد. اگر مخرج کسر احتمالات زیرمجموعه‌ای از مجموعه مرجع باشد، حاصل را احتمال شرطی می‌نامند. مفهوم احتمال شرطی را با مراجعه مجدد به جدول گروه خونی و جنس می‌توان شرح داد. فرض کنید بخواهیم احتمال آنکه یک نفر انتخاب شده گروه خونیAB داشته باشد را محاسبه کنیم. احتمال بالا احتمال غیرشرطی است، چه بر مجموعه همه نتایج ممکن هیچ شرطی بار نکرده‌ایم.

این احتمال را به صورت زیر محاسبه می‌نماییم:

$$P(C_٤) = \frac{n(C_٤)}{n(U)} = \frac{٦}{١٢} = ٠.٥$$

حال اگر مجموعه همه نتایج ممکن را به مجموعه افراد مرد (مجموعه $D_٢$) کاهش دهیم،حال حساب کنید احتمال آن که یکی از افراد انتخاب شده گروه خونی AB داشته باشد به شرط آنکه وی از میان مردها انتخاب شده باشد، چقدر است؟

در جدول مربوطه مشاهده می‌کنیم که مجموعه $D_۲$ از ۶ عضو، یعنی از افراد مرد درست شده است. در همین جدول می‌بینیم که ۳ نفر گروه خونی AB دارند که معادل با تعداد اعضای مجموعه $C_۴ \cap D_۲$ می‌باشد. از این‌رو، پاسخ سوال بالا به صورت زیر است:

$$P(C_٤|D_۲) = \frac{n(C_٤ \cap D_۲)}{n(D_۲)} = \frac{۳}{٦} = ۰.۵$$

خط عمودی در $P(C_۴|D_۲)$ خوانده می‌شود "به شرطی که ... داده شده باشد" پس، $P(C_۴|D_۲)$ چنین خوانده می‌شود: "احتمال پیشامد $C_۴$ به شرطی که پیشامد $D_۲$ داده شده باشد" برای حل مسائل شرطی همیشه اول شرط را اجرا می‌کنیم که بعد از خط عمودی (|) می‌آید. اگر صورت و مخرج کسر بالا را بر تعداد اعضای مجموعه مرجع n(U) ، یعنی تعداد کل افراد تقسیم کنیم خواهیم داشت:

$$P(C_٤|D_۲) = \frac{\frac{n(C_٤ \cap D_۲)}{n(U)}}{\frac{n(D_۲)}{n(U)}}$$

$$P(C_٤|D_۲) = \frac{P(C_٤ \cap D_۲)}{P(D_۲)} = \frac{\frac{۳}{۱۲}}{\frac{٦}{۱۲}} = ۰.۵$$

نتیجه حاصله ما را به تعریف کلی "احتمال شرطی" به صورت زیر رهنمون می‌سازد.

تعریف : احتمال شرطی A به شرطی که پیشامد B داده شده باشد، برابر است با احتمال $A \cap B$ تقسیم بر احتمال پیشامد B به شرط آن که احتمال پیشامد B مخالف صفر باشد.

یعنی:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq ۰$$

ه - قانون ضرب

قانون ضرب زمانی به کار می‌رود که چند پیشامد با هم اتفاق بیفتند یعنی حادثه اول و حادثه دوم و ... . قبل از استفاده از این قانون لازم است پیشامدهای مستقل و وابسته تعریف شوند.

دو یا چند پیشامد را موقعی "مستقل" می‌نامند که احتمال وقوع هر یک از آنها تحت تاثیر وقوع دیگری قرار نگیرد. در غیر این صورت، پیشامد را "غیر مستقل یا وابسته" می‌نامند.

اگر تاسی را دو بار بریزیم، وقوع ۳ در ریزش اول و وقوع ۵ در ریزش دوم وقایع مستقل می‌باشند یا اگر دو تاس را یکمرتبه پرت کنیم، احتمال وقوع ۳ در تاس اول و ۵ در تاس دوم مستقل از هم هستند. حالا مثلا اگر گروه خونی افراد یک جامعه آماری خاص را روی کارتهای مستقل بنویسیم به ترتیبی که در آن ۴ کارت گروه خونی A، ۴ کارت B، ۴کارت O و ۱۲ کارت AB و در مجموع ۲۶ کارت داشته باشیم. اگر دو کارت به طور متوالی از یک دسته کارت، بدون جایگذاری بیرون کشیده شوند، ظهور یک A در کشیدن مرتبه اول و ظهور یک B در دفعه دوم وقایع مستقل نیستند. زیرا احتمال به دست آمدن B در کشیدن دفعه دوم بستگی به حادثه‌ای دارد که در کشیدن ورق اول اتفاق افتاده است. اگر بعد از انتخاب ورق آن را جایگذاری کنیم، در این صورت، نتیجه کشیدن ورق دوم از کشیدن ورق اول مستقل خواهد بود. برای تشخیص دو رویداد وابسته یا مستقل از روی جدول گروه خونی می‌توان به این صورت نیز عمل کرد که اگر اعمال شرط تغییری در میزان احتمال بوجود نیاورد آن دو پیشامد را مستقل و در صورتی که اعمال شرط احتمال محاسبه شده را تغییر دهد آن دو پیشامد را وابسته می‌نامیم. مثلا در بحث احتمال شرطی احتمال گروه خونی AB به شرط مرد بودن با احتمال گروه خونی AB بدون هیچ شرطی برابر ۰.۵ محاسبه شده است، بنابراین دو پیشامد گروه خونی AB و مرد بودن مستقل هستند. اما دو پیشامد گروه خونی B و مرد بودن وابسته هستند (چگونه؟).

در صورتی که دو پیشامد، مستقل از یکدیگر باشند، احتمال این که هر دوی آنها اتفاق بیفتند، مساوی حاصل ضرب احتمال وقوع هر یک از آنها خواهد بود: $P(A \cap B) = P(A).P(B)$

اگر دو پیشامد وابسته باشند، یعنی احتمال وقوع یکی از آنها به احتمال وقوع با عدم وقوع دیگری بستگی داشته باشد، احتمال آن که هر دوی آنها اتفاق بیفتند، عبارت است از:

$$P(A \cap B) = P(A).P(B|A) = P(B).P(A|B)$$

همانطوریکه دیده میشود این قانون از روی طرفین وسطین کردن فرمول احتمال شرطی به راحتی اثبات میشود.

مثال: از همان دسته ۲۶ تایی کارت گروه خونی، یک ورق به طور تصادفی کشیده می‌شود (بدون جایگذاری). سپس یک ورق دیگر نیز کشیده می‌شود. احتمال آنکه ورق اول A و ورق دوم B باشد چقدر است؟

احتمال کشیدن یک A در دفعه اول برابر $p_1 = \frac{۴}{۲۶}$ و احتمال کشیدن یک B در بار دوم پس از آنکه یک A از بین برگها بیرون کشیده شده باشد، برابر $p_2 = \frac{۴}{۲۵}$ است. بنابراین احتمال مورد نظر برابر است با:

$$P = \frac{٤}{٢٦} \times \frac{٤}{٢٥} = \frac{١٦}{٦٥٠}$$

یعنی: $P(B|$مابقی کارتها$)$

این بدان معنی است که اگر عمل کشیدن دو کارت به طور متوالی از بین ۲۶ برگ را به دفعات زیاد انجام دهیم (و همیشه به همین صورت و بدون جایگذاری)، فقط در ۱۶ بار از ۶۵۰ بار یا حدود ۱ بار در هر ۴۰ آزمایش، ابتدا یک A و بعد از آن یک B بیرون کشیده می‌شوند.

قانون ضرب را به بیش از دو پیشامد نیز، می‌توان تعمیم داد. برای مثال در مورد سه پیشامد مستقل می‌توان نوشت:

$$P(A \cap B \cap C) = P(A).P(B).P(C)$$

و در مورد سه پیشامد وابسته خواهیم داشت:

$$P(A \cap B \cap C) = P[(A \cap B) \cap C] = P(A \cap B).P(C|A \cap B) = P(A).P(B|A).P(C|A \cap B)$$

مثال: در یک کیسه ۴ مهره سفید، ۵ مهره سیاه و ۸ مهره قرمز وجود دارد. از این کیسه ۳ مهره، به طور تصادفی انتخاب می‌شود. احتمال این که اولی سفید، دومی سیاه و سومی قرمز باشد، چقدر است؟

$$P\left(\text{سومی قرمز ، دومی سیاه ، اولی سفید}\right) = \frac{٤}{١٦} * \frac{٥}{١٥} * \frac{٨}{١٤} = ٠.٠٤٨$$

توجه کنید که اختلاف عمده بین قانون جمع و ضرب در این است که در اولی وقوع حادثه اول یا دوم یا سوم، الی آخر، مدنظر است، در حالی که در مورد دوم وقوع حادثه اول و دوم و سوم و غیره مورد نظر می‌باشد. بنابراین، غالبا در یک

مسئله احتمال اگر "یا" در برداشته باشد، باید احتمال‌های وقایع را با هم جمع کرد (البته به شرطی که آنها مانعه الجمع یا ناسازگار باشند) و از طرف دیگر، اگر مسئله احتمال "و" دربرداشته باشد، لازمه‌اش این است که احتمالهای وقایع را در هم ضرب نمود (به شرطی که وقایع از هم مستقل باشند). گاهی اوقات در حل بعضی از مسائل هر دو قانون ضرب و جمع با هم استفاده می‌شود:

مثال: احتمال این را حساب کنید که در یک خانواده از سه بچه دو تا پسر و یکی دختر باشد.

حالات مطلوب ما عبارتند از:

$$A = \text{دختر ، پسر ، پسر}$$

$$B = \text{پسر ، دختر ، پسر}$$

$$C = \text{پسر ، پسر ، دختر}$$

$$P(A) = P(\text{پسر}) \times P(\text{پسر}) \times P(\text{دختر}) \frac{۱}{۸}$$

$$P(B) = P(\text{پسر}) \times P(\text{دختر}) \times P(\text{پسر}) \frac{۱}{۸}$$

$$P(C) = P(\text{دختر}) \times P(\text{پسر}) \times P(\text{پسر}) \frac{۱}{۸}$$

$$P(A \text{یا} B \text{یا} C) = P(A) + P(B) + P(C) = \frac{۱}{۸} + \frac{۱}{۸} + \frac{۱}{۸} = \frac{۳}{۸}$$

احتمال عدم رخداد

احتمال عدم رخداد پیشامد A یا مکمل پیشامد A برابر است با:

$$P(\bar{A}) = ۱ - P(A)$$

مثلا در جدول گروه خونی احتمال مکمل پیشامد $D_۲$ چقدر است: $P(\overline{D_۲}) = ۱ - \frac{۶}{۱۲} = \frac{۶}{۱۲}$ که همان احتمال زن بودن است.

مثال: اگر ۳۰ درصد جامعه مستعد تصادف باشند (*P(A)*)، چند درصد جامعه مستعد تصادف نیستند ($P(\bar{A})$)؟

$$P(\bar{A}) = ۱ - P(A) = ۱ - \frac{۳۰}{۱۰۰} = \frac{۷۰}{۱۰۰}$$

مثال: اگر *P(A)*=۰.۳ و *P(B)*=۰.۲ و *A* و *B* از هم مستقل باشند $P(\bar{A} \cap \bar{B})$ چقدر است؟

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = \big(۱ - P(A)\big) \times \big(۱ - P(B)\big) = (۱ - ۰.۳) \times (۱ - ۰.۲) = (۰.۷) \times (۰.۸) = ۰.۵۶$$

مثال: سه نفر A و B و C به ترتیب احتمال ۰.۴ و ۰.۷ و ۰.۵ یک مسئله را حل می‌کنند. مطلوب است:

الف) احتمال آن که فقط یکی مسائل را حل کند.

ب) احتمال آن که مسائل حل شود.

حل الف)

$$P = P\big(A \text{ حل کند و } B \text{ حل نکند و } C \text{ حل نکند}\big) + P\big(A \text{ حل نکند و } B \text{ حل کند و } C \text{ حل نکند}\big) + P\big(A \text{ حل نکند و } B \text{ حل نکند و } C \text{ حل کند}\big)$$

$$P = A\acute{B}\acute{C} + \acute{A}B\acute{C} + \acute{A}\acute{B}C = (۰.۴ \times ۰.۳ \times ۰.۵) + (۰.۶ \times ۰.۷ \times ۰.۵) + (۰.۶ \times ۰.۳ \times ۰.۵) = ۰.۳۶$$

حل ب)

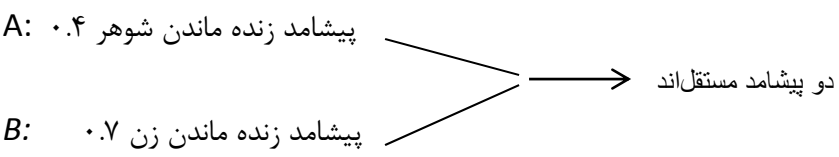
حل شدن مسئله بدین معنا است که حداقل یکی مسائل را حل کند. بنابراین از مکمل آن استفاده می‌کنیم.

$$= ۱ - \acute{A}\acute{B}\acute{C} = ۱ - (۱ - ۰.۵)(۱ - ۰.۷)(۱ - ۰.۴) = ۱ - (۰.۵)(۰.۳)(۰.۶) = ۰.۹۱$$

مثال:فرض کنید احتمال آنکه در ۲۰ سال آینده زن و شوهری زنده بمانند به ترتیب ۰.۷ و ۰.۴ باشد، مطلوبست محاسبه احتمال:

الف) هر دو زنده بمانند. ب) هیچکدام زنده نمانند ج) احتمال آنکه حداقل یکی زنده بماند(در ۲۰ سال آینده شخصی زنده بماند) د) فقط یکی زنده بماند ه) فقط زن زنده بماند و)زن زنده بماند

حل الف)



$$P(A \cap B) = P(A) \times P(B) = ۰.۴ \times ۰.۷ = ۰.۲۸$$

حل ب)

$$P(\acute{A} \cap \acute{B}) = P(\acute{A}) \times P(\acute{B}) = ۰.۶ \times ۰.۳ = ۰.۱۸$$

حل ج)

راه حل اول: $P(A \cup B) = ۱ - P(\acute{A} \cap \acute{B}) = ۱ - (\text{هیچکدام زنده نمانند}) = ۱ - ۰.۱۸ = ۰.۸۲$

راه حل دوم: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = ۰.۴ + ۰.۷ - (۰.۴)(۰.۷) = ۰.۸۲$

حل د)

$$P(\acute{A} \cap B) + P(A \cap \acute{B}) = P(\acute{A}).P(B) + P(A).P(\acute{B}) = (۰.۶)(۰.۷) + (۰.۴)(۰.۳) = ۰.۵۴$$

حل ه)

$$P(\acute{A} \cap B) = P(\acute{A}).P(B) = (۰.۶)(۰.۷) = ۰.۴۲$$

حل و)

$$P(B) = ۰.۷$$

نکته: مکمل احتمال وقوع پیشامد A به شرط وقوع حادثه B عبارتند از: $P(\bar{A}|B) = ۱ - P(A|B)$

مثال: اگر P(A)=۰.٤ و P(B)=۰,٦ و $P(B|A) = ۰.۱$ باشد، آنگاه $P(\bar{A}|B)$ را حساب کنید.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = ۰.۱ \times ۰.٤ = ۰.۰٤$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{۰.۰٤}{۰.٦} = ۰/۰٦۷$$

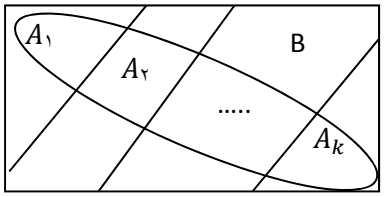
$$P(\bar{A}|B) = ۱ - P(A|B) = ۱ - ۰.۰٦۷ = ۰.۹۳$$

احتمال متوسط

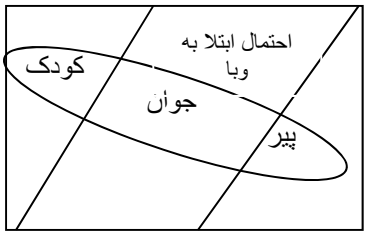
هرگاه پیشامد B به شرط وقوع هر یک از پیشامدهای ناسازگار $A_۱$، $A_۲$ ... A_k بتواندرخ دهد (وابسته به پیشامدهای A)، آنگاه احتمال وقوع پیشامد B به طور متوسط برابر است با:

$$P(B) = \sum_{i=۱}^{k} P(A_i)P(B|A_i)$$

نمایش گرافیکی این مفهوم به شکل زیر است:



به عنوان مثال اگر افراد یک جامعه به سه طبقه کودک، جوان و پیر تقسیم شده باشند و بیماری وبا در این جامعه شایع شده باشد، احتمال ابتلا به میکروب وبا به طور متوسط در این جامعه برابر است با:



$$P\left(\text{ابتلا به وبا}\right) = P\left(\text{ابتلا به وبا و کودک}\right) + P\left(\text{ابتلا به وبا و جوان}\right) + P(\text{ابتلا به وبا و پیر})$$

از آنجا که پیشامد ابتلا به وبا (B) وابسته به پیشامدهای کودک، جوان و پیر بودن (A_1, A_2, A_3) است لذا فرمول ضرب پیشامدهای وابسته استفاده می‌شود.

مثال: اگر یک فرد حواس‌پرت در یک سال به احتمال ۴۰٪ تصادف کند و این احتمال برای فرد حواس‌جمع ۲۰٪ باشد. اگر فرض کنیم ۳۰٪ افراد جامعه حواس‌پرت باشند، احتمال اینکه یک فرد از این جامعه در یک سال تصادف داشته باشد چقدر است؟

اگر A احتمال تصادف و B احتمال حواس‌پرتی باشد خواهیم داشت:

$P(B)$=۳۰٪ $\quad P(\bar{B}) = ۷۰٪ \quad P(A|B) = ٤۰٪ \quad P(A|\bar{B}) = ۲۰٪$

$$= P(B).P(A|B) + P(\bar{B}).P(A|\bar{B}) = \left(\frac{۳۰}{۱۰۰} \times \frac{٤۰}{۱۰۰}\right) + \left(\frac{۷۰}{۱۰۰} \times \frac{۲۰}{۱۰۰}\right) = \frac{۲٦}{۱۰۰}$$

قضیه بیز

ممکن است احتمال وقوع پیشامدی را در شرایط معمول بدانید، ولی اگر اطلاعات جدیدی به دست آوردید، در احتمال وقوع پیشامد اولیه، تجدید نظر کنید. به احتمال وقوع پیشامد قبل از اطلاعات جدید "احتمال پیشین" و به اااحتمال وقوع آن پیشامد، بعد از کسب اطلاعات جدید، "احتمال پسین" گویند. این دو گونه احتمال در نظریه تصمیم‌گیری کاربرد فراوان دارد. قضیه بیز ما را برای تجدیدنظر در احتمالات در صورت دسترسی به اطلاعات جدید، کمک می‌کند.

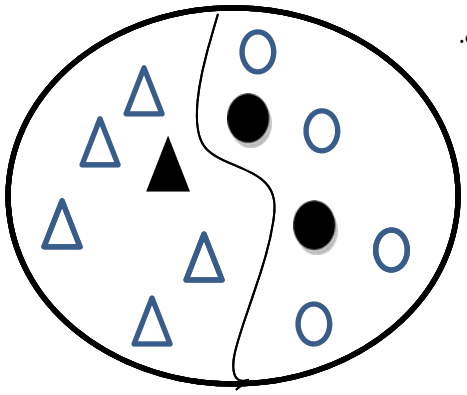
فرض کنید پیشامد B به همراه یکی و تنها یکی از پیشامدهای ناسازگار A_1، A_2 ... A_k رخ دهد، آنگاه:

$P(Ai|B)=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}$ یعنی همان احتمال متوسط P(B)

یا $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$ ← یعنی همان ضرب احتمالات $P(A \cap B)$

که در آن $P(A_i)$ را احتمال پیشین و $P(A_i|B)$ را احتمال پسین پیشامد A_i می‌نامند.

برای روشن شدن مطلب فرض کنید در شکل مقابل از جمعیتی ۱۲ نفره شامل ۶ نفر زن (O) و ۶ نفر مرد (Δ) تعداد ۱ زن و ۲ مرد بیمار تشخیص داده شده است.



احتمال آنکه یک مریض انتخاب شود به شرط آنکه زن باشد برابر خواهد بود با:

$$P\left(\text{مریض}\middle|\text{زن}\right)=\frac{\frac{۳}{۱۲}\times\frac{۲}{۳}}{\frac{۶}{۱۲}}=\frac{۲}{۶}$$

احتمال آنکه یک زن انتخاب شود به شرط آنکه مریض باشد برابر است با:

$$P\left(\text{زن}\middle|\text{مریض}\right)=\frac{\frac{۶}{۱۲}\times\frac{۲}{۶}}{\frac{۳}{۱۲}}=\frac{۲}{۳}$$

هر دو این احتمالات به راحتی از روی شکل نیز قابل محاسبه هستند. همانطوریکه دیده می‌شود این دو مفهوم با هم فرق می‌کنند و احتمالات متفاوتی دارند.

مثال: با مطالعه ژنوم بیماران یک بیماری خاص ۴ ژن $A_۱$، $A_۲$، $A_۳$ و $A_۴$ به ترتیب با فراوانی نسبی ۳۰٪ ، ۲۰٪، ٤۰٪ و ۱۰٪ مشاهده شده است و اگر میزان توانایی تشخیص هر یک از ژنهای مذکور با ابزارهای موجود به ترتیب ۳٪، ۲٪، ۱٪ و ٤٪ باشد، پیدا کنید.

الف) احتمال تشخیص بیماری خاص در فرد مشکوک *P(A)* از روی این ژنها

	$A_۱$	$A_۲$	$A_۳$	$A_٤$
فراوانی نسبی	٪۳۰	٪۲۰	٪٤۰	٪۱۰
درصد توانایی تشخیص	٪۳	٪۲	٪۱	٪٤

$$P(A) = \left(\frac{۳۰}{۱۰۰} \times \frac{۳}{۱۰۰}\right) + \left(\frac{۲۰}{۱۰۰} \times \frac{۲}{۱۰۰}\right) + \left(\frac{٤۰}{۱۰۰} \times \frac{۱}{۱۰۰}\right) + \left(\frac{۱۰}{۱۰۰} \times \frac{٤}{۱۰۰}\right) = ۰.۰۲۱$$

ب) اگر فردی بیماری خاصی داشته باشد احتمال اینکه از روی بررسی ژن‌های $A_۱$ یا $A_٤$ بتوان آن را تشخیص داد چقدر است.

$$P(B) = P(A_۱|A) + P(A_٤|A)$$

$$P(A_۱|A) = \frac{\frac{۳۰}{۱۰۰} \times \frac{۳}{۱۰۰}}{۰.۰۲۱} = ۰.٤۲۸$$

$$P(A_٤|A) = \frac{\frac{۱۰}{۱۰۰} \times \frac{٤}{۱۰۰}}{۰.۰۲۱} = ۰.۱۹۰٤$$

P(B)= ۰.٤۲۸ + ۰.۱۹۰٤ = ۰.٦۱۸

مثال: پنجاه و دو درصد از جمعیت شهری را زنان تشکیل می‌دهند، ۰.۰۰۴ از زنان و ۰.۰۰۳ از مردان مبتلا به آسم هستند: الف- چه نسبتی از اهالی شهر مبتلا به آسم هستند؟

ب- اگر شخصی انتخاب شود و مبتلا به آسم باشد،احتمال اینکه زن باشد چقدر است؟

ابتدا B را پیشامد زنان (در نتیجه $\bar{B}$پیشامد مردان) و A را پیشامد مبتلایان به آسم تعریف می‌کنیم

$$P(B) = ۰.۵۲ \quad P(\bar{B}) = ۰.۴۸ \qquad P(A/B) = ۰.۰۰۴ \qquad P(A/\bar{B}) = ۰.۰۰۳$$

الف- $P(A)=?$

$$P(A) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B}) = (۰.۰۰۴ \times ۰.۵۲) + (۰.۰۰۳ \times ۰.۴۸) = ۰.۰۰۳۵۲$$

ب- $P(B/A) =?$

$$P(B/A) = \frac{P(A/B)P(B)}{P(A/B)P(B) + P(A/\bar{B})P(\bar{B})} = \frac{۰.۰۰۴ \times ۰.۵۲}{(۰.۰۰۴ \times ۰.۵۲) + (۰.۰۰۳ \times ۰.۴۸)} = \frac{۰.۰۰۲۰۸}{۰.۰۰۳۵۲} = ۰.۵۹۱$$

مثال: الف) از روی جدول گروه خونی $P(C_۴|D_۲)$ را حساب کنید.

$$P(C_۴|D_۲) = \frac{P(C_۴ \cap D_۲)}{P(D_۲)} = \frac{\frac{۳}{۱۲}}{\frac{۶}{۱۲}} = \frac{۳}{۶}$$

ب) بدون استفاده از جدول گروه خونی $P(C_۴|D_۲)$ را حساب کنید.

$$P(C_۴|D_۲) = \frac{P(C_۴).P(D_۲|C_۴)}{P(D_۲)} = \frac{\frac{۶}{۱۲} \times \frac{۳}{۶}}{\frac{۶}{۱۲}} = \frac{۳}{۶}$$

تمارین:

۱- فرض کنیم تست توبرکولین برای افراد مسلول با احتمال ۹۹ درصد و برای افراد غیرمسلول با احتمال ۱۰ درصد مثبت است. چنانچه ۵ درصد افراد جمعیتی مسلول باشند. اگر از بین این جمعیت یک نفر را به طور تصادفی برگرفته مورد تست توبرکولین قرار دهیم.

اولا این احتمال را حساب کنید که جواب تست او مثبت باشد.

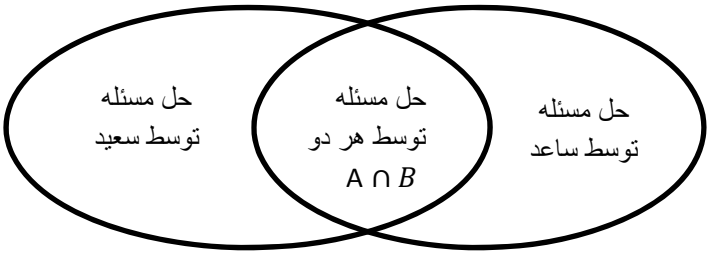
ثانیا اگر بدانیم که جواب تست او مثبت است، این احتمال را حساب کنید که او مسلول باشد.

۲- سعید و ساعد، جدا از همدیگر، برای حل مساله‌ای می‌کوشند. احتمال حل مساله توسط سعید برابر با $\frac{۱}{۳}$ و توسط ساعد برابر با $\frac{۱}{۴}$ است. احتمال هر یک از پیشامدهای زیر را به دست آورید.

اولا حل مساله توسط هر دو نفر

ثانیا حل مساله توسط فقط یکی از دو نفر

ثالثا احتمال حل شدن مساله



۳- ۳٪ از انسانها به یک بیماری خاص مبتلا می‌شوند. در یک گروه ۱۰۰ نفری از مردان، احتمالات مربوط به موارد زیر چقدر است؟

اولا دقیقا ۳ نفر به این بیماری مبتلا باشند.

ثانیا ۲، ۳ یا ٤ نفر به این بیماری مبتلا باشند.