فصل هشتم: آمار و مدل‌سازی

درسنامه

فصل اول: اندازه‌گیری و مدل‌سازی

علم آمار علم جمع‌آوری اطلاعات و تجزیه و تحلیل آن‌هاست. اطلاعات به دو دسته‌ی «کمّی» و «کیفی یا توصیفی» تقسیم‌بندی می‌شوند. اطلاعات کمّی به آن دسته از اطلاعات می‌گوییم که به صورت عدد و رقم مطرح می‌شوند و قابل تفسیر نیستند. اطلاعات کیفی یا توصیفی به آن دسته از اطلاعات می‌گوییم که قابل تفسیر باشند و ممکن است شنونده‌های مختلف برداشت‌های مختلف از آن داشته باشند. اولین قدم در رسیدن به اطلاعات عددی «اندازه‌گیری» است.

تعریف: **(مدل‌سازی): بیان یک مسأله به زبان ریاضی را «مدل‌سازی» آن مسأله می‌گوییم.**

تفاضل مقدار واقعی و مقدار اندازه‌گیری شده را «خطای اندازه‌گیری» می‌گوییم. معمولاً خطا را با E نمایش می‌دهند، و محاسبه آن به صورت «مقدار تقریبی - مقدار واقعی $=E$» است. توجه کنید که مقدار خطا (E)، می‌تواند مثبت یا منفی باشد و باید از واحد اندازه‌گیری کم‌تر باشد.

مثال: اگر طول ضلع مربعی ۴ سانتی‌متر اندازه‌گیری شده باشد آن‌را به صورت $a=۴+E$ نشان می‌دهند که در آن مقدار خطا در محدوده‌ی $|E|<۱$ سانتی‌متر است.

مثال: فرض کنید وزن فردی ۸۳/۵ کیلو گزارش شده است. در این صورت وزن او را با مدل $W=۸۳/۵+E$ نمایش می‌دهیم که قدرمطلق E کم‌تر از ۰/۵ کیلوگرم یا ۵۰۰ گرم است.

تذکر در مسائل مدل‌سازی از جملاتی که شامل توان دوم خطا یا بالاتر باشند (مانند $E^۲, E^۳, \ldots$) صرف‌نظر می‌کنیم.

مثال: اگر شعاع دایره‌ای به صورت $R=۲+E$ باشد، مدل مساحت این دایره (S) به صورت زیر است:

$$S=\pi R^۲=\pi(۲+E)^۲=\pi(۴+E^۲+۴E)$$

طبق تذکر بالا می‌توان از $E^۲$ صرف‌نظر کرد، پس مدل مساحت این دایره (به صورت تقریبی) برابر است با:

$$S\simeq\pi(۴+۴E)$$

تست ۱: میزان خطای یک ترازوی دیجیتالی -۳ گرم است. اگر جرم جسمی با این ترازو ۱۷۷ گرم نشان داده شود، جرم واقعی این جسم کدام است؟

۱) ۱۸۰ گرم **۲)** ۱۷۴ گرم **۳)** ۱۷۸/۵ گرم **۴)** ۱۷۵/۵ گرم

پاسخ: می‌دانیم «مقدار تقریبی $-$ مقدار واقعی $=$ مقدار خطا»، بنابراین می‌توان نوشت:

$$-۳=\text{مقدار واقعی}-۱۷۷\Rightarrow\text{مقدار واقعی}=۱۷۴$$

بنابراین گزینه‌ی (۲) درست است.

فصل دوم: جامعه و نمونه

تعریف: **(جامعه‌ی آماری): مجموعه‌ای از اشیا یا افراد که درباره‌ی اعضای آن می‌خواهیم موضوع و یا موضوعاتی را مطالعه کنیم، «جامعه‌ی آماری» می‌گوییم. تعداد اعضای جامعه‌ی آماری را «اندازه‌ی جامعه» یا «حجم جامعه» می‌گوییم.**

مثال: فرض کنید بخواهیم نمره‌ی ریاضی ترم اول دانش‌آموزان سال دوم دبیرستان کل کشور را در سال ۱۳۹۴ بررسی کنیم. در این بررسی آماری، جامعه‌ی آماری کل دانش‌آموزان سال دوم دبیرستان در سال ۱۳۹۴ هستند و موضوع مورد مطالعه، نمره‌ی ریاضی ترم اول آن‌هاست.

تعریف: **(سرشـماری):** اگر در یک بررسـی آماری، تمام افراد جامعه‌ی آماری را مورد مطالعه قرار دهیم، می‌گوییم «سرشـماری» کرده‌ایم. معمولاً در سرشـماری با مشـکلاتی مثل، در دسـترس نبودن تمام اعضای جامعه، وقت‌گیر بودن، گران تمام شدن، از بین رفتن جامعه در برخی از مطالعات و ... مواجه هستیم.

(نمونه): زیرمجموعه‌ی یک جامعه‌ی آماری را «نمونه» می‌گوییم. تعداد اعضای نمونه را «اندازه‌ی نمونه» یا «حجم نمونه» می‌گوییم.

تذکـر عمل «نمونه‌گیری» مهم‌ترین بخش آمار است. یک نمونه گروه کوچکی از جامعه‌ی آماری است که به نحوی انتخاب می‌شود که نمایانگر خصوصیات جامعه باشد.

نتایج حاصل از بررسی یا اندازه‌گیری نمونه را «داده» می‌گوییم. داده‌ها را به چهار روش جمع‌آوری می‌کنند که عبارت است از:

۱- داده‌های از پیش تهیه شده
۲- جمع‌آوری داده‌ها از طریق پرسش
۳- جمع‌آوری داده‌ها از طریق آزمایش
۴- جمع‌آوری داده‌ها از طریق مشاهده

فرض کنید بخواهیم عددی را از مجموعه‌ای از اعداد به طور تصادفی انتخاب کنیم. انتخاب تصادفی به معنای آن است که ذهنیت ما هیچ تأثیری در انتخاب عدد تصادفی نداشـته باشـد. برای انتخاب عددی تصادفی در بین تعدادی از اعداد کافی است عدد تصادفی تولید شده توسط ماشین حساب را در اندازه‌ی جامعه ضرب کنیم، سپس قسمت اعشاری عدد به‌دست آمده را حذف کرده و یک واحد به آن اضافه کنیم.

تذکـر اعداد تصادفی که توسط ماشین حساب تولید می‌شوند همواره غیر منفی و کوچک‌تر از ۱ هستند.

مثـال: فرض کنید عدد تصادفی تولید شده توسط ماشین حساب ۰/۲۷۳ باشد و اندازه‌ی جامعه‌ای که می‌خواهیم از آن نمونه را انتخاب کنیم ۲۵۰ باشد، در این صورت خواهیم داشت: ۰/۲۷۳×۲۵۰=۶۸/۲۵

اکنون قسمت اعشاری عدد را حذف کرده و یک واحد به آن اضافه می‌کنیم که به عدد ۶۹ می‌رسیم.

تست ۲: می‌خواهیم از بین اعداد طبیعی ۴۱ تا ۹۰ یک عدد را به تصادف انتخاب کنیم. اگر عدد تصادفی تولید شـده توسـط ماشین حساب برابر ۰/۲۳ باشد، عدد انتخاب شده کدام است؟

۱) ۵۱ **۲)** ۵۳ **۳)** ۵۲ **۴)** ۵۴

پاسخ: ابتدا توجه کنید که تعداد اعداد طبیعی ۴۱ تا ۹۰ برابر است با: ۹۰−۴۱+۱=۵۰

اکنون می‌توان نوشت ۰/۲۳×۵۰=۱۱/۵، پس باید دوازدهمین عدد را با شروع از ۴۱ انتخاب کنیم. بنابراین باید عدد ۵۲ را انتخاب کنیم.

بنابراین گزینه‌ی (۳) درست است.

فصل سوم: متغیرهای تصادفی

موضوع و مشخصه‌ای از اشیاء جامعه که مورد بررسی قرار می‌گیرد، «متغیر تصادفی» نامیده می‌شود. انواع متغیرهای تصادفی به‌صورت زیر است:

متغیرهایی هستند که قابل «اندازه‌گیری» یا «شمارش» هستند و به دو دسته تقسیم می‌شوند:

الف) متغیرهای کمّی گسسته: این متغیرها «قابل شمارش» هستند مانند تعداد نامه‌های یک صندوق، تعداد بیماران مراجعه کننده به یک پزشک در طول روز و ....

ب) متغیرهای کمّی پیوسـته: این متغیرها «قابل اندازه‌گیری» هسـتند مانند وزن نامه‌های موجود در یک صندوق، زمانی که یک بیمار در اتاق انتظار مطب، منتظر است.

این متغیرها قابل اندازه‌گیری و شمارش نیستند و به دو دسته تقسیم می‌شوند:

الف) متغیر کیفی ترتیبی: در این متغیرها نوعی ترتیب طبیعی وجود دارد مانند مراحل تحصیلی که به صورت دبسـتان، راهنمایی، دبیرسـتان و دانشگاه است. یا مراحل زندگی که به صورت نوزادی، کودکی، نوجوانی، جوانی، میان‌سالی و کهنسالی است.

ب) متغیر کیفی اسمی: در این نوع متغیر کیفی، هیچ‌گونه ترتیبی وجود ندارد مانند گروه خونی، RH خون، رنگ اشیا.

تست ۳: میزان «آلودگی هوا» کدام نوع متغیر است؟

۱) کمّی - گسسته ۲) کمّی - پیوسته ۳) کیفی - ترتیبی ۴) کیفی - اسمی

پاسخ: میزان آلودگی هوا را به کمک دستگاه‌هایی اندازه‌گیری می‌کنند، پس متغیری کمّی و پیوسته است. بنابراین گزینه‌ی (۲) درست است.

فصل چهارم: دسته‌بندی داده‌ها و جدول فراوانی

معمولاً داده‌ها در قالب یک جدول نوشته می‌شوند که به آن «جدول فراوانی» می‌گوییم.

فراوانی مطلق داده‌ی x_i برابر تعداد دفعاتی است که آن داده تکرار شده است. فراوانی مطلق داده‌ی x_i را با نماد f_i نشان می‌دهیم. «فراوانی مطلق» را گاهی به اختصار «فراوانی» می‌گویند.

تذکر مجموع تمام فراوانی‌های مطلق برابر تعداد کل داده‌هاست.

مثال: فرض کنید نمرات یک کلاس ۱۰ نفره به صورت ۱۹,۱۶,۱۶,۱۶,۱۴,۱۴,۱۲,۹,۸,۸ است. در این صورت جدول فراوانی این داده‌ها به صورت مقابل است:

x_i	۸	۹	۱۲	۱۴	۱۶	۱۹
f_i	۲	۱	۱	۲	۳	۱

به عنوان نمونه، طبق جدول بالا، فراوانی مطلق داده‌ی $x_i=۱۶$ برابر ۳ است.

برای بررسی تعدادی داده، گاهی بهتر است آن‌ها را دسته‌بندی کنیم. دسته‌بندی به این صورت انجام می‌گیرد که داده‌هایی که نزدیک به هم هستند در یک دسته قرار می‌گیرند. واضح است که هر چقدر تعداد دسته‌ها بیش‌تر باشد، نتایج حاصل از آن دسته‌بندی دقیق‌تر خواهد بود. قبل از بیان روش دسته‌بندی، به تعریف «دامنه‌ی تغییرات» می‌پردازیم.

اختلاف بین بزرگ‌ترین و کوچک‌ترین داده را «دامنه‌ی تغییرات» می‌گوییم. دامنه‌ی تغییرات را معمولاً با R نمایش می‌دهند.

مثال: در داده‌های ۱۹,۱۸/۵,۱۸,۱۷,۱۷,۱۷,۱۴,۱۳/۵,۱۲,۱۱,۱۱,۱۰ دامنه‌ی تغییرات برابر است با: $R=۱۹-۱۰=۹$

تذکر اگر به همه‌ی داده‌ها عدد ثابت k را اضافه کنیم یا از همه‌ی داده‌ها عدد ثابت k را کم کنیم دامنه‌ی تغییرات داده‌ها تغییری نمی‌کند. به همین ترتیب، اگر همه‌ی داده‌ها را در عدد k ضرب کنیم دامنه‌ی تغییرات در $|k|$ ضرب می‌شود.

مثال: اگر دامنه‌ی تغییرات داده‌های $x_۱, x_۲, \ldots, x_{۱۰}$ برابر ۱۷ باشد، دامنه‌ی تغییرات داده‌های $x_۱+۳, x_۲+۳, \ldots, x_{۱۰}+۳$ نیز برابر ۱۷ خواهد بود.

مثال: اگر دامنه‌ی تغییرات $x_۱, x_۲, \ldots, x_{۲۰}$ برابر ۱۴ باشد، دامنه‌ی تغییرات داده‌های $-۲x_۱, -۲x_۲, \ldots, -۲x_{۲۰}$ برابر است با:

$|-۲|\times ۱۴=۲۸$

تست ۴: اگر دامنه‌ی تغییرات داده‌های $x_۱, x_۲, \ldots, x_۵$ برابر ۱۹ باشد، دامنه‌ی تغییرات داده‌های $۳x_۱+۷, ۳x_۲+۷, \ldots, ۳x_۵+۷$ کدام است؟

۱) ۲۶ ۲) ۵۷ ۳) ۶۴ ۴) ۱۹

پاسخ: همان‌طور که در مطلب بالا ملاحظه کردید اضافه یا کم کردن عددی ثابت به همه‌ی داده‌ها تأثیری در دامنه‌ی تغییرات ندارد. بنابراین می‌توان نتیجه گرفت دامنه‌ی تغییرات داده‌های $۳x_۱+۷, ۳x_۲+۷, \ldots, ۳x_۵+۷$ برابر است با $۳\times ۱۹=۵۷$.

بنابراین گزینه‌ی (۲) درست است.

برای دسته‌بندی داده‌ها، ابتدا دامنه‌ی تغییرات را محاسبه می‌کنیم. اگر دامنه‌ی تغییرات را بر تعداد دسته‌ها تقسیم کنیم، طول دسته‌ها محاسبه می‌شود. به‌جز دسته‌ی آخر، بقیه‌ی دسته‌ها را به صورت $[a_i, b_i)$ در نظر می‌گیریم که در آن‌ها به a_i، کران پایین و به b_i کران بالای دسته می‌گوییم. توجه کنید که دسته‌ی آخر را به صورت $[a_i, b_i]$ در نظر می‌گیریم.

تعریف: (مرکز دسته): **به میانگین کران بالا و پایین هر دسته، «نشان دسته» یا «مرکز دسته» می‌گوییم.**

مثال: فرض کنید نمرات یک کلاس ۱۵ نفره به صورت ۲۰,۱۹,۱۸,۱۷,۱۷,۱۷,۱۷,۱۶,۱۶,۱۴,۱۴,۱۲,۱۲,۱۲,۸ است. اگر بخواهیم این داده‌ها را به ۴ دسته تقسیم کنیم، با توجه به این‌که، دامنه‌ی تغییرات برابر $R=۲۰-۸=۱۲$ است، طول دسته‌ها برابر $\frac{۱۲}{۴}=۳$ است. پس دسته‌ها به صورت زیر خواهند بود:

$[۸,۱۱),[۱۱,۱۴),[۱۴,۱۷),[۱۷,۲۰]$

اکنون جدول فراوانی این داده‌های دسته‌بندی شده به صورت زیر است:

دسته‌ها	مرکز دسته‌ها	فراوانی هر دسته
$[۸,۱۱)$	۹/۵	۱
$[۱۱,۱۴)$	۱۲/۵	۳
$[۱۴,۱۷)$	۱۵/۵	۴
$[۱۷,۲۰]$	۱۸/۵	۷

نکته: ۱- اختلاف مرکز دو دسته‌ی متوالی برابر است با طول دسته

۲- اختلاف کران پایین و کران بالای هر دسته برابر است با طول دسته

۳- فاصله‌ی کران پایین (یا کران بالا) هر دسته تا مرکز آن دسته برابر است با نصف طول آن دسته

مثال: در یک آمارگیری، تعداد دسته‌ها ۵ و دامنه‌ی تغییرات ۵۰ است. اگر کم‌ترین داده ۲ باشد، کران پایین دسته‌ی سوم و مرکز دسته‌ی چهارم را محاسبه می‌کنیم. برای این منظور می‌دانیم طول دسته‌ها برابر است با $\frac{۵۰}{۵}=۱۰$. اکنون با توجه به این‌که طول دسته‌ها برابر ۱۰ و کران پایین دسته‌ی اول برابر ۲ است، این دسته‌ها به صورت زیر هستند:

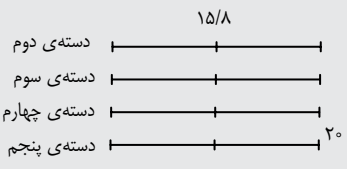
$[۲,۱۲)$: دسته‌ی اول ، $[۱۲,۲۲)$: دسته‌ی دوم ، $[۲۲,۳۲)$: دسته‌ی سوم ، $[۳۲,۴۲),\ldots$: دسته‌ی چهارم

پس کران پایین دسته‌ی سوم برابر ۲۲ و مرکز دسته‌ی چهارم برابر است با: $\frac{۳۲+۴۲}{۲}=۳۷$

تست ۵: یک سری داده‌ی آماری را در دسته‌هایی با طول‌های مساوی، دسته‌بندی کرده‌ایم. اگر مرکز دسته‌ی دوم ۱۵/۸ و کران بالای دسته‌ی پنجم ۲۰ باشد، کران بالای دسته‌ی هشتم کدام است؟

۱) ۲۶/۳ **۲)** ۲۳/۶ **۳)** ۲۲/۶ **۴)** ۲۴/۶

پاسخ: با توجه به شکل مقابل، اگر طول دسته‌ها را x در نظر بگیریم فاصله‌ی مرکز دسته‌ی دوم تا کران بالای دسته‌ی پنجم برابر $۳/۵x$ است پس می‌توان نوشت:



$$۳/۵x=۲۰-۱۵/۸\Rightarrow x=\frac{۶}{۵}=۱/۲$$

اکنون می‌توان نتیجه گرفت کران بالای دسته‌ی هشتم برابر است با: $۲۰+۳x=۲۰+۳/۶=۲۳/۶$

بنابراین گزینه‌ی (۲) درست است.

اگر فراوانی هر دسته را به تعداد کل داده‌ها تقسیم کنیم، عدد به‌دست آمده را «فراوانی نسبی» آن دسته می‌گویند.

تذکر مجموع فراوانی‌های نسبی همه‌ی دسته‌ها برابر یک است.

تست ۶: در جدول توزیع فراوانی روبه‌رو حاصل $x+y+z$ کدام است؟

دسته‌ها	فراوانی	فراوانی نسبی
۰-۵	۳۰	۰/۲
۵-۱۰	x	۰/۱
۱۰-۱۵	y	z

۱) ۱۲۰/۷
۲) ۱۲۲/۷
۳) ۱۲۵/۷
۴) ۱۲۴/۳

پاسخ: با توجه به این که مجموع فراوانی‌های نسبی برابر ۱ است می‌توان فهمید: $۰/۲+۰/۱+z=۱\Rightarrow z=۰/۷$

اکنون با توجه به این که فراوانی‌های مطلق و نسبی دسته‌ی اول به ترتیب ۳۰ و ۰/۲ است می‌توان تعداد کل داده‌ها را محاسبه کرد. اگر تعداد کل داده‌ها را n فرض کنیم، داریم: $۰/۲=\frac{۳۰}{n}\Rightarrow n=۱۵۰$

از آن‌جا که فراوانی نسبی دسته‌ی دوم و تعداد کل داده‌ها را می‌دانیم می‌توانیم مقدار x و y را محاسبه کنیم:

$$۰/۱=\frac{x}{۱۵۰}\Rightarrow x=۱۵ \quad , \quad y=۱۵۰-(۳۰+۱۵)=۱۰۵$$

پس حاصل $x+y+z$ برابر ۱۲۰/۷ است.

بنابراین گزینه‌ی (۱) درست است.

اگر فراوانی نسبی هر دسته را در عدد ۱۰۰ ضرب کنیم، عدد به‌دست آمده را «درصد فراوانی نسبی» آن دسته می‌گوییم.

تذکر مجموع درصدهای فراوانی نسبی همه‌ی دسته‌ها برابر ۱۰۰ است.

تست ۷: در یک جدول توزیع فراوانی، فراوانی مطلق داده‌ها در دسته‌های اول تا پنجم دارای روندی صعودی است. اگر فراوانی مطلق دسته‌ی سوم برابر ۸ باشد و درصد فراوانی نسبی دسته‌های دوم و چهارم به ترتیب ۲۵ و ۴۰ باشد، آن‌گاه فراوانی کل کدام می‌تواند باشد؟

۱) ۳۴ **۲)** ۳۶ **۳)** ۳۰ **۴)** ۱۸

پاسخ: اگر فراوانی کل داده‌ها را n در نظر بگیریم، درصد فراوانی نسبی دسته‌ی سوم برابر $\frac{۸}{n}\times ۱۰۰$ می‌باشد.

از طرفی، طبق صورت مسأله، فراوانی‌های دسته‌ها روندی صعودی دارند پس می‌توان فهمید درصد فراوانی‌های نسبی هم روندی صعودی دارند پس می‌توان نوشت:

$$۲۵\le\frac{۸۰۰}{n}\le ۴۰\Rightarrow ۲۰\le n\le ۳۲$$

در بین گزینه‌ها، فقط عدد ۳۰ در این محدوده قرار دارد.

بنابراین گزینه‌ی (۳) درست است.

مجموع فراوانی‌های مطلق دسته‌های اول، دوم، . . . تا iاُم را «فراوانی تجمعی دسته‌ی iاُم» می‌گوییم.

تذکر فراوانی تجمعی دسته‌ی اول با فراوانی مطلق آن برابر است. همچنین فراوانی تجمعی دسته‌ی آخر برابر است با تعداد کل داده‌ها.

اگر فراوانی تجمعی هر دسته را بر تعداد کل داده‌ها تقسیم کنیم، به آن «فراوانی تجمعی نسبی» آن دسته می‌گوییم و اگر این عدد را در ۱۰۰ ضرب کنیم، به آن «درصد فراوانی تجمعی نسبی» آن دسته می‌گوییم.

تذکر در جدول فراوانی، فراوانی تجمعی نسبی و درصد فراوانی تجمعی نسبی دسته‌ی آخر به ترتیب برابر ۱ و ۱۰۰ است.

مثـال: برای کلاس ۱۵ نفره با نمرات ۸، ۱۲، ۱۲، ۱۲، ۱۴، ۱۴، ۱۶، ۱۶، ۱۷، ۱۷، ۱۷، ۱۷، ۱۸، ۱۹ و ۲۰ جدول فراوانی را در مبحث روش دسته‌بندی نوشتیم، می‌توانیم جدول را به صورت زیر بنویسیم.

دسته‌ها	مرکز دسته	فراوانی مطلق	فراوانی نسبی	درصد فراوانی نسبی	فراوانی تجمعی	فراوانی تجمعی نسبی	درصد فراوانی تجمعی نسبی
$[8,11)$	۹/۵	۱	$\frac{۱}{۱۵}$	$\frac{۱}{۱۵}\times ۱۰۰$	۱	$\frac{۱}{۱۵}$	$\frac{۱}{۱۵}\times ۱۰۰$
$[11,14)$	۱۲/۵	۳	$\frac{۳}{۱۵}$	$\frac{۳}{۱۵}\times ۱۰۰$	۴	$\frac{۴}{۱۵}$	$\frac{۴}{۱۵}\times ۱۰۰$
$[14,17)$	۱۵/۵	۴	$\frac{۴}{۱۵}$	$\frac{۴}{۱۵}\times ۱۰۰$	۸	$\frac{۸}{۱۵}$	$\frac{۸}{۱۵}\times ۱۰۰$
$[17,20]$	۱۸/۵	۷	$\frac{۷}{۱۵}$	$\frac{۷}{۱۵}\times ۱۰۰$	۱۵	۱	۱۰۰
		۱۵= مجموع	۱= مجموع	۱۰۰= مجموع			

تست ۸: اگر در یک جدول فراوانی، فراوانی تجمعی دسته‌های یازدهم و دوازدهم با هم برابر باشند، درصد فراوانی نسبی دسته‌ی دوازدهم کدام است؟

۱) ۱٪ **۲)** ۱۰٪ **۳)** صفر **۴)** ۱۰۰٪

پاسخ: می‌دانیم «فراوانی مطلق دسته‌ی دوازدهم + فراوانی تجمعی دسته‌ی یازدهم = فراوانی تجمعی دسته‌ی دوازدهم» اکنون با توجه به این که فراوانی تجمعی دسته‌های یازدهم و دوازدهم با هم برابر هستند می‌توان فهمید فراوانی مطلق دسته‌ی دوازدهم برابر صفر است. پس فراوانی نسبی (و درصد فراوانی نسبی) نیز در این دسته برابر صفر است. بنابراین گزینه‌ی (۳) درست است.

فصل پنجم: نمودارها و تحلیل داده‌ها

نمودارها وسیله‌ای سودمند برای به تصویر کشیدن ویژگی‌های یک جامعه‌ی آماری هستند. در ادامه انواع نمودارها را معرفی می‌کنیم:

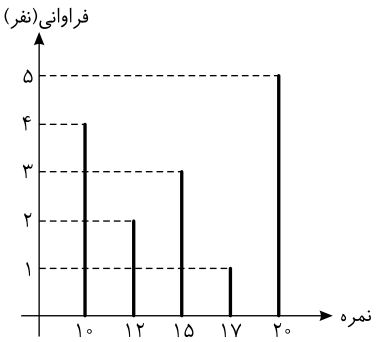
نمودار میله‌ای از سه قسمت اصلی تشکیل شده است.

الف) عنوان: زیر هر نمودار باید موضوع مورد مطالعه به‌طور خلاصه نوشته شود.

ب) برچسب محورها: هر یک از محورها برچسبی دارند که مشخص‌کننده‌ی متغیری است که آن محور نشان می‌دهد.

پ) مقیاس: مقیاس هر محور باید مشخص باشد.

مثـال: اگر نمرات امتحان نهایی درس ریاضی یک کلاس به صورت ۲۰,۲۰,۲۰,۲۰,۲۰,۱۷,۱۵,۱۵,۱۵,۱۲,۱۲,۱۰,۱۰,۱۰,۱۰ باشد، نمودار میله‌ای این نمرات به صورت زیر است:

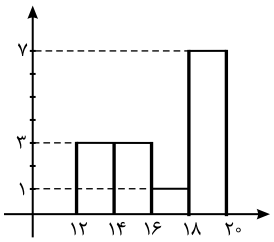


نمودار نمرات امتحان نهایی درس ریاضی کلاس (الف)

در مثال بالا داده‌ها دسته‌بندی نشده بودند. در رسم نمودار میله‌ای برای داده‌های دسته‌بندی شده مرکز دسته‌ها را روی محور افقی و فراوانی هر دسته را روی محور عمودی (محور yها) در نظر می‌گیریم. در این حالت واضح است که فاصله‌ی هر دو میله‌ی متوالی برابر طول دسته‌ها است. نمودار میله‌ای بیش‌تر برای متغیرهای کمّی گسسته و متغیرهای کیفی مناسب است.

نمودار مستطیلی برای متغیرهای کمّی پیوسته مناسب است. این نمودار، نمایشی از داده‌های دسته‌بندی شده است که در آن سطح مستطیل‌ها متناسب با فراوانی دسته‌ها است. در این نمودار هر دسته مستطیلی دارد که اگر طول همه‌ی دسته‌ها یکسان باشد یکی از اضلاع آن طول دسته و ضلع دیگرش فراوانی آن دسته است.

مثال: اگر نمرات یک کلاس ۱۴ نفره به صورت ۱۲/۷۵، ۱۲،۱۳،۱۴/۵، ۱۴،۱۵،۱۷،۱۸/۲۵، ۱۸/۵، ۱۸،۱۹/۵، ۱۹،۲۰،۲۰ باشد و بخواهیم آن‌ها را به ۴ دسته تقسیم کنیم، با توجه به این که دامنه‌ی تغییرات برابر $R=۲۰-۱۲=۸$ است، طول دسته‌ها برابر ۲ خواهد بود و نمودار مستطیلی آن به صورت مقابل است:



تذکر با توجه به نمودار مستطیلی می‌توان فراوانی نسبی هر دسته را از طریق تقسیم «مساحت مربوط به آن دسته» بر «مجموع مساحت همه‌ی مستطیل‌ها» محاسبه کرد. به عنوان مثال در نمودار مستطیلی مثال قبل، فراوانی نسبی دسته‌ی $[۱۲,۱۴)$ برابر است با: $\frac{۶}{۲۸}$

تذکر مساحت زیر یک نمودار مستطیلی وقتی طول همه‌ی دسته‌ها یکسان باشند برابر است با «تعداد کل داده‌ها × طول دسته».

تست ۹: مساحت زیر نمودار مستطیلی n داده که در دسته‌هایی با طول‌های مساوی دسته‌بندی شده‌اند ۵۴۰ است. اگر تعداد کل داده‌ها $n=۹۰$ باشد، طول هر دسته کدام است؟

۱) ۴ ۲) ۶ ۳) ۹ ۴) ۱۰

پاسخ: همان‌طور که گفته شد مساحت زیر یک نمودار مستطیلی در چنین شرایطی برابر است با «طول دسته × تعداد کل داده‌ها». پس می‌توان نوشت: (طول دسته‌ها را x در نظر می‌گیریم)

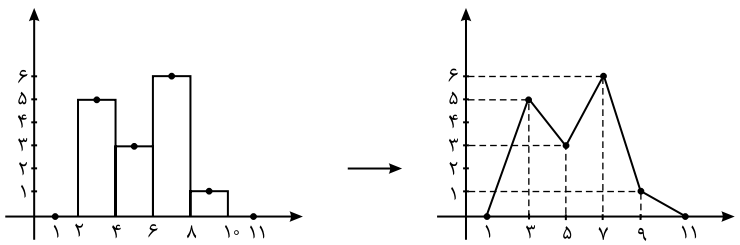
$$۵۴۰=۹۰x \Rightarrow x=۶$$

بنابراین گزینه‌ی (۲) درست است.

نمودار چندبر فراوانی تغییرات یک متغیر پیوسته (مانند وزن، قد، دما و . . .) را بهتر از نمودار مستطیلی نمایش می‌دهد. برای رسم نمودار «چندبر فراوانی»، نقاطی از صفحه را در نظر می‌گیریم که طول آن‌ها مرکز دسته‌ها و عرض هر یک از آن‌ها فراوانی مربوط به آن دسته باشد. سپس آن نقاط را به ترتیب به یک‌دیگر متصل می‌کنیم. در انتها برای آن‌که منحنی «چندبر فراوانی» به شکل بسته ایجاد شود و سطح زیر چندبر با سطح زیر نمودار مستطیلی یکسان شود، دو دسته با فراوانی صفر در ابتدا و انتهای نمودار در نظر می‌گیریم.

تذکر نمودار چندبر فراوانی را می‌توان با داشتن فراوانی نسبی نیز رسم کرد که آن‌را «چندبر فراوانی نسبی» می‌گوییم.

مثال: در شکل زیر از روی نمودار مستطیلی (نمودار سمت چپ) و به کمک نقاط مشخص شده، نمودار «چندبر فراوانی» (سمت راست) رسم شده است. همان‌طور که ملاحظه می‌کنید برای رسم «چندبر فراوانی» از نقاط $(۱,۰),(۳,۵),(۵,۳),(۷,۶),(۹,۱),(۱۱,۰)$ استفاده شده است. یادآوری می‌کنیم که مساحت زیر نمودار «مستطیلی» با مساحت زیر نمودار «چندبر فراوانی» برابر است.



تست ۱۰: هرگاه دسته‌ی اول یک جدول فراوانی به‌صورت $[۹,۱۵)$ و فراوانی آن ۵ باشد، طول‌های نقاط اول و دوم نمودار چندبر فراوانی تشکیل شده به ترتیب چه اعدادی هستند؟

۱) ۶ , ۱۲ **۲)** صفر , ۵ **۳)** ۵ , ۶ **۴)** صفر , ۱۲

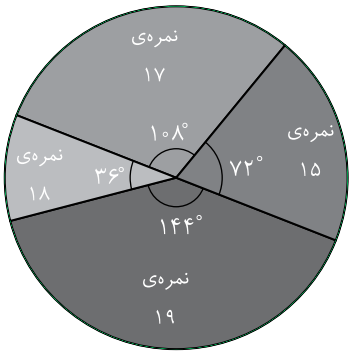
پاسخ: توجه کنید که در نمودار چندبر فراوانی مرکز هر دسته را بر روی محور xها قرار می‌دهیم و قبل از دسته‌ی اول یک دسته با فراوانی صفر به نمودار اضافه می‌کنیم. توجه کنید که فاصله‌ی بین نقاط متوالی برابر طول دسته‌ها است. پس می‌توان فهمید نقطه‌ی دوم چندبر فراوانی برابر مرکز دسته‌ی $[۹,۱۵)$ است که برابر با $\frac{۹+۱۵}{۲}=۱۲$ و طول نقطه‌ی اول برابر است با $۶=۶-۱۲$.

بنابراین گزینه‌ی (۱) درست است.

یکی دیگر از نمودارهایی که می‌تواند اطلاعات موجود در داده‌ها را به سرعت در معرض دید قرار دهد «نمودار دایره‌ای» است. برای رسم این نمودار چنین عمل می‌کنیم:

اگر متغیر تصادفی مورد مطالعه‌ی ما دارای k حالت باشد که فراوانی‌های آن‌ها به ترتیب $f_۱, f_۲, \ldots, f_k$ باشد و $f_۱+f_۲+\ldots+f_k=n$، ابتدا دایره‌ای به شعاع دلخواه رسم می‌کنیم. سپس این دایره را به کمک زاویه‌های مرکزی به k قسمت تقسیم می‌کنیم به‌طوری که زاویه‌ی مرکزی دسته‌ی iاُم (با فراوانی f_i)، برابر است با: $\frac{f_i}{n}\times ۳۶۰$ زاویه‌ی مرکزی دسته‌ی iام

مثال: اگر نمرات یک کلاس ۱۰ نفره به صورت $۱۹,۱۹,۱۹,۱۹,۱۸,۱۷,۱۷,۱۷,۱۵,۱۵$ باشد، نمودار دایره‌ای به صورت زیر است:



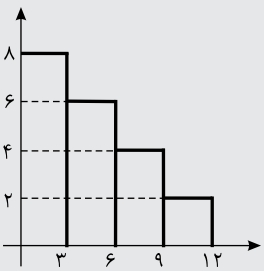
زاویه‌ی مربوط به نمره‌ی ۱۵ $=\frac{۲}{۱۰}\times ۳۶۰=۷۲^\circ$

زاویه‌ی مربوط به نمره‌ی ۱۷ $=\frac{۳}{۱۰}\times ۳۶۰=۱۰۸^\circ$

زاویه‌ی مربوط به نمره‌ی ۱۸ $=\frac{۱}{۱۰}\times ۳۶۰=۳۶^\circ$

زاویه‌ی مربوط به نمره‌ی ۱۹ $=\frac{۴}{۱۰}\times ۳۶۰=۱۴۴^\circ$

تست ۱۱: در نمودار مستطیلی مقابل، زاویه‌ی مرکزی مربوط به دسته‌ای با مرکز ۷/۵ در نمودار دایره‌ای کدام است؟



۱) ۶۰°

۲) ۷۲°

۳) ۸۰°

۴) ۷۵°

پاسخ: در دسته‌ی $[۶,۹)$، مرکز دسته برابر ۷/۵ است. بنابراین برای تعیین زاویه‌ی مرکزی این دسته در نمودار دایره‌ای باید فراوانی نسبی این دسته را در ۳۶۰ ضرب کنیم:

زاویه‌ی مرکزی مربوط به دسته‌ای به مرکز ۷/۵ $=\frac{۴}{۲+۴+۶+۸}\times ۳۶۰^\circ=\frac{۴}{۲۰}\times ۳۶۰^\circ=۷۲^\circ$

بنابراین گزینه‌ی (۲) درست است.

به داده‌های آماری مقابل دقت کنید: ۴۰,۴۰,۴۰,۴۰,۴۱,۴۱,۴۲,۴۳,۵۰,۵۲,۵۴,۵۶,۵۶,۵۶,۵۷

این داده‌ها را می‌توان به صورت نمودار زیر که «نمودار ساقه و برگ» نامیده می‌شود نمایش داد:

ساقه	برگ							
۴	۰	۰	۰	۰	۱	۱	۲	۳
۵	۰	۲	۴	۶	۶	۶	۷	

کلید نمودار: ۴ ۱=۴۱

در نمودار بالا، ستون سمت چپ را که بیانگر قسمت مشترک تعدادی عدد است «ساقه» و اعداد روبه‌روی این ستون را «برگ» می‌نامند.

مثال: داده‌های اعشاری زیر را با نمودار ساقه و برگ نمایش می‌دهیم:

۴۱,۴۱/۱,۴۱/۲,۴۱/۲,۴۲,۴۲,۴۲/۳,۴۲/۳,۴۲/۳,۴۲/۵,۴۳,۴۳,۴۳,۴۳/۹

ساقه	برگ					
۴۱	۰	۱	۲	۲		
۴۲	۰	۰	۳	۳	۳	۵
۴۳	۰	۰	۰	۹		

کلید نمودار: ۴۱ ۱=۴۱/۱

همان‌طور که ملاحظه می‌کنید برای نمایش اعداد اعشاری که قسمت صحیح آن‌ها شبیه هم است می‌توان از نمودار ساقه و برگ استفاده کرد. نمودار جعبه‌ای را بعد از معرفی میانه توضیح خواهیم داد.

فصل ششم: شاخص‌های مرکزی

شاخص‌های مرکزی به شاخص‌هایی گفته می‌شود که محل تمرکز داده‌ها را معرفی می‌کند. شاخص‌های مرکزی عبارت هستند از: مُد (نَما)، میانه و میانگین که به بررسی هر یک می‌پردازیم.

مُد داده‌ای است که بیش‌ترین فراوانی را دارد.

مثال: در داده‌های ۱۵,۱۷,۱۷,۱۸,۱۸,۱۸,۱۹,۲۰,۲۰ عدد ۱۸ مُد است، چون بیش‌ترین فراوانی را دارد.

تذکر مُد ممکن است منحصر به فرد نباشد به عنوان مثال داده‌های ۱۵,۱۶,۱۶,۱۷,۱۸,۱۹,۱۹,۲۰ دارای دو مُد ۱۶ و ۱۹ هستند به این نوع جوامع، ۲ مُدی می‌گویند. جامعه ممکن است «چند مُدی» هم باشد. مُد در این قبیل جامعه‌ها شاخص معتبری نیست.

تست ۱۲: در نمودار ساقه و برگ مقابل مُد کدام است؟

ساقه	برگ					
۱	۰	۱	۲	۳	۳	
۲	۱	۲	۳	۴	۴	۴

۱) ۱۴ **۲)** ۲۴

۳) ۱۰ **۴)** ۲۱

پاسخ: با توجه به نمودار، داده‌ی ۲۴ دارای بیش‌ترین فراوانی است، پس مقدار مُد برابر ۲۴ است. بنابراین گزینه‌ی (۲) درست است.

پس از مرتب کردن داده‌ها، داده‌ای که وسط همه‌ی داده‌ها قرار می‌گیرد را «میانه» می‌گویند. (مشابه تعریف میانه‌ی مثلث در هندسه!!) اگر تعداد داده‌ها زوج باشد، میانگین دو داده‌ی وسط را میانه در نظر می‌گیریم.

مثال: میانه‌ی داده‌های ۱۷,۱۵,۱۴,۱۸,۱۹,۲۰,۱۲، پس از مرتب کردن داده‌ها، برابر ۱۷ است.

مثال: میانه‌ی داده‌های ۱۲,۱۴,۱۵,۱۶,۱۷,۲۰، برابر است با: $\frac{۱۵+۱۶}{۲}=۱۵/۵$

تذکر همان‌طور که ملاحظه می‌کنید میانه، لزوماً عضوی از داده‌ها نیست.

تذکر میانه و مد هیچ‌کدام نسبت به اندازه‌ی داده‌ها حساسیتی نشان نمی‌دهند. به عنوان مثال در هر دو جامعه‌ی $(۰,۲,۳)$ و $(-۲۰۰,۲,۷۰۰)$ میانه عدد ۲ است در حالی‌که اندازه‌ی داده‌ها در این دو جامعه خیلی متفاوت است. همین مطلب در مورد مُد نیز برقرار است.

تست ۱۳: هر یک از داده‌های زیر را با چه عددی جمع کنیم تا میانه‌ی داده‌ها برابر ۱۵ باشد؟

$۸,۱۰,۱۱,۱۷,۱۸,۱۹$

۱) ۲ ۲) ۱/۵ ۳) ۲/۵ ۴) ۱

پاسخ: فرض کنید که به همه‌ی داده‌ها x واحد اضافه کرده‌ایم. در این صورت میانه‌ی داده‌ها به‌صورت زیر خواهد بود:

$$\text{میانه}=\frac{(۱۱+x)+(۱۷+x)}{۲}\Rightarrow ۱۵=\frac{۲۸+۲x}{۲}\Rightarrow ۳۰=۲۸+۲x\Rightarrow x=۱$$

بنابراین گزینه‌ی (۴) درست است.

نمودارهایی که تاکنون شناخته‌ایم (میله‌ای، مستطیلی، چندبر، دایره‌ای و ساقه‌وبرگ) همگی، برای مقایسه‌ی داده‌ها بسیار مفیدند ولی هیچ‌کدام به سؤالاتی از قبیل «آیا داده‌ها به هم نزدیک هستند؟» و «داده‌ها بیش‌تر در اطراف کدام داده متمرکزند؟ میانگین، بیش‌ترین داده یا کم‌ترین داده؟» پاسخ نمی‌دهند. برای پاسخ به این گونه سؤالات نمودار جعبه‌ای کاربرد دارد که به روش رسم آن می‌پردازیم.

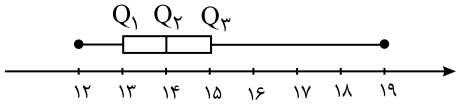
چارک اول: به میانه‌ی نیمه‌ی اول داده‌ها (داده‌هایی که قبل از میانه قرار دارند)، چارک اول می‌گوییم و آن‌را با $Q_۱$ نمایش می‌دهیم.

چارک دوم: به میانه‌ی همه‌ی داده‌ها چارک دوم می‌گوییم و آن‌را با $Q_۲$ نمایش می‌دهیم. پس چارک دوم همان میانه است.

چارک سوم: به میانه‌ی نیمه‌ی دوم داده‌ها (داده‌هایی که بعد از میانه قرار دارند)، چارک سوم می‌گوییم و آن‌را با $Q_۳$ نمایش می‌دهیم.

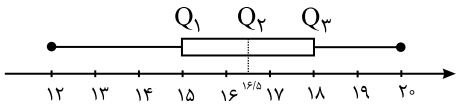
تعریف: نمودار جعبه‌ای نموداری تصویری است که داده‌ها را بر اساس پنج مقدار کوچک‌ترین داده، چارک اول، چارک دوم (میانه)، چارک سوم و بزرگ‌ترین داده نمایش می‌دهد.

مثال: نمودار جعبه‌ای داده‌های مرتب شده‌ی $۱۲,۱۳,۱۴,۱۴,۱۵,۱۵,۱۹$، به صورت زیر تعیین می‌شود:



$$\underbrace{۱۲,۱۳,۱۴}_{Q_۱=۱۳},\ \underbrace{۱۴}_{Q_۲=۱۴},\ \underbrace{,۱۵,۱۵,۱۹}_{Q_۳=۱۵}$$

مثال: نمودار جعبه‌ای داده‌های مرتب‌شده‌ی $۱۲,۱۵,۱۶,۱۷,۱۸,۲۰$ به صورت زیر تعیین می‌شود:



$$\underbrace{\overbrace{۱۲,۱۵,۱۶}^{Q_۱=۱۵},\overbrace{۱۷,۱۸,۲۰}^{Q_۳=۱۸}}_{Q_۲=\frac{۱۶+۱۷}{۲}=۱۶/۵}$$

تست ۱۴: چه تعدادی از داده‌های یک مجموعه‌ی ۱۲ عضوی بین چارک اول و سوم قرار دارند؟

۱) ۵ ۲) ۷ ۳) ۶ ۴) ۳

پاسخ: اگر داده‌های مورد نظر را به‌صورت مرتب $a_۱,a_۲,\ldots,a_{۱۲}$ در نظر بگیریم (توجه کنید که داده‌ها متمایز هستند.)می‌توان فهمید میانه‌ی نیمه‌ی اول داده‌ها برابر $Q_۱=\frac{a_۳+a_۴}{۲}$ و میانه‌ی نیمه‌ی دوم داده‌ها برابر $Q_۳=\frac{a_۹+a_{۱۰}}{۲}$ است. پس داده‌های $a_۴,a_۵,\ldots,a_۹$ بین چارک اول و چارک سوم هستند که تعداد آن‌ها ۶ تا است. بنابراین گزینه‌ی (۳) درست است.

میانگین به معنای معدل داده‌ها است و شاخص خوبی برای نشان دادن مرکزیت داده‌ها است. میانگین داده‌های $x_۱,x_۲,\ldots,x_n$ را با $\overline{x}$ نمایش می‌دهیم و به صورت زیر تعریف می‌شود:

$$\overline{x}=\frac{x_۱+x_۲+\ldots+x_n}{n}=\frac{\sum_{i=۱}^{n}x_i}{n}$$

مثال: میانگین داده‌های ۱۲,۱۳,۱۴,۱۵,۱۵,۱۵,۱۶,۳۶ برابر است با:

$$\overline{x}=\frac{۱۲+۱۳+۱۴+۱۵+۱۵+۱۵+۱۶+۳۶}{۸}=\frac{۱۳۶}{۸}=۱۷$$

میانگین وزن‌دار (وزنی)

فرض کنید داده‌های $x_۱, x_۲, \ldots, x_n$ به ترتیب دارای ضریب‌های $f_۱, f_۲, \ldots, f_n$ باشند، این داده‌ها را برای سادگی می‌توانیم در جدول زیر خلاصه کنیم:

داده‌ها	$x_۱$	$x_۲$	...	x_n
ضریب (وزن)	$f_۱$	$f_۲$	...	f_n

در این صورت میانگین داده‌های بالا با احتساب ضرایب مربوط، به صورت زیر محاسبه می‌شود:

$$\text{میانگین وزنی}=\frac{f_۱.x_۱+f_۲.x_۲+\ldots+f_n.x_n}{f_۱+f_۲+\ldots+f_n}=\frac{\sum_{i=۱}^{n} f_i.x_i}{\sum_{i=۱}^{n} f_i}$$

مثال: نمرات دانش‌آموزی به صورت مقابل است. میانگین وزن‌دار نمرات او چقدر است؟

نمره	۱۹	۱۸	۱۵	۲۰
ضریب	۳	۲	۱	۲

$$\text{میانگین وزنی}=\frac{۳\times۱۹+۲\times۱۸+۱\times۱۵+۲\times۲۰}{۳+۲+۱+۲}=\frac{۱۴۸}{۸}=۱۸/۵$$

تذکر میانگین وزن‌دار را نیز با نماد $\overline{x}$ نشان می‌دهند و اگر بخواهند تمایزی بین میانگین معمولی و میانگین وزن‌دار قائل شوند، میانگین وزن‌دار را با نماد $\overline{x}_w$ نشان می‌دهند.

تذکر با توجه به فرمول میانگین در مواردی که با داده‌های دسته‌بندی شده مواجه هستیم، به جای x_iها از مرکز هر دسته استفاده می‌کنیم.

مثال: به جدول مقابل توجه کنید، اکنون میانگین داده‌های این جدول را به‌دست می‌آوریم:

دسته	$[۰,۵)$	$[۵,۱۰)$	$[۱۰,۱۵)$	$[۱۵,۲۰]$
فراوانی	۲	۱	۷	۶

می‌دانیم مرکز دسته‌ها به ترتیب، ۲/۵, ۷/۵, ۱۲/۵ و ۱۷/۵ است، پس میانگین وزنی این داده‌ها برابر است با:

$$\overline{x}=\frac{۲\times۲/۵+۱\times۷/۵+۷\times۱۲/۵+۶\times۱۷/۵}{۲+۱+۷+۶}=\frac{۲۰۵}{۱۶}\simeq۱۲/۸$$

تست ۱۵: در جدول مقابل درصد نمرات داوطلبی نمایش داده شده است. اگر حداقل میانگین برای پذیرش ۷۵ باشد، حداقل نمره‌ی ادبیات وی برای پذیرش کدام است؟

درس	ادبیات	معارف	زبان	اختصاصی
درصد	؟	۹۰	۸۱	۷۰
ضریب	۴	۲	۳	۸

۱) ۷۱ **۲)** ۷۲

۳) ۷۳ **۴)** ۷۴

پاسخ: ابتدا به کمک فرمول میانگین وزن‌دار، میانگین این فرد را محاسبه می‌کنیم: (درصد درس ادبیات را a در نظر می‌گیریم.)

$$\overline{x}=\frac{۴a+۲(۹۰)+۳(۸۱)+۸(۷۰)}{۴+۲+۳+۸}=\frac{۴a+۹۸۳}{۱۷}$$

اکنون می‌توان نوشت:

$$\frac{۴a+۹۸۳}{۱۷}\geq۷۵\Rightarrow۴a+۹۸۳\geq۱۲۷۵\Rightarrow a\geq۷۳$$

پس حداقل باید درس ادبیات را ۷۳ درصد بزند. بنابراین گزینه‌ی (۳) درست است.

ویژگی‌های میانگین

۱- اگر $\bar{x}$ میانگین داده‌های $x_1, x_2, \ldots, x_n$ باشد در این صورت میانگین داده‌های $ax_1 \pm b, ax_2 \pm b, \ldots, ax_n \pm b$ را با نماد $\overline{ax \pm b}$ نشان می‌دهیم و داریم: $\overline{ax \pm b} = a\bar{x} \pm b$

تساوی بالا به معنای آن است که اگر همه‌ی داده‌ها را با یک عدد ثابت جمع کنیم، میانگین اولیه نیز با همان عدد جمع می‌شود. همچنین اگر همه‌ی داده‌ها را در یک عدد ثابت ضرب کنیم، میانگین اولیه نیز در همان عدد ضرب می‌شود.

۲- اگر $\bar{x}$ میانگین داده‌های $x_1, x_2, \ldots, x_n$ باشد، آن‌گاه مجموع اختلاف داده‌ها از میانگین برابر صفر است. یعنی:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

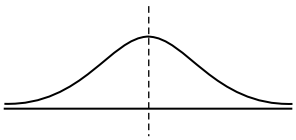
مثال: میانگین داده‌های $1, 4, 17, 18, 20$ برابر ۱۲ است. مجموع اختلاف داده‌ها از میانگین برابر است با:

$$(1-12) + (4-12) + (17-12) + (18-12) + (20-12) = -11 - 8 + 5 + 6 + 8 = 0$$

مقایسه‌ی میانگین و میانه

اگر میانه از میانگین کوچک‌تر باشد، به معنای آن است که اکثریت داده‌ها نسبت به میانگین کوچک‌تر هستند. اگر میانه از میانگین بزرگ‌تر باشد، به معنای آن است که اکثریت داده‌ها از میانگین بزرگ‌تر هستند.

در منحنی نرمال (شکل مقابل) مقادیر میانگین، میانه و مد با هم برابر هستند.



میانه، مد، میانگین

روش سریع محاسبه‌ی میانگین

فرض کنید می‌خواهیم میانگین اعداد $20, 20, 20, 20, 19, 18, 17, 17, 16, 15, 15, 15$ را محاسبه کنیم. ابتدا، مقداری حدسی برای میانگین در نظر می‌گیریم، ما در این‌جا مقدار ۱۸ را در نظر می‌گیریم. اکنون کافی است $x_i - \bar{x}$ (تفاضل داده با میانگین) هر یک از داده‌ها را محاسبه کرده و میانگین آن‌ها را محاسبه کنیم و آن را به میانگین حدس زده شده اضافه کنیم.

$$\frac{4(20-18) + (19-18) + (18-18) + 2(17-18) + (16-18) + 3(15-18)}{12} = \frac{-4}{12} = -\frac{1}{3}$$

پس میانگین واقعی برابر است با: $18 - \frac{1}{3} = \frac{53}{3}$

نکته: اگر داده‌ها تعدادی از جملات متوالی از یک دنباله‌ی حسابی باشند، میانگین داده‌ها برابر جمله‌ی وسط (یا میانگین دو جمله‌ی وسط) است.

مثال: میانگین داده‌های ۱۳۱، ۱۲۹، ۱۲۷، ۱۲۵ و ۱۲۳ برابر ۱۲۷ است.

مثال: اگر داده‌های b ، a ، ۱۹ ، ۸ ، y و x جملات متوالی یک دنباله‌ی حسابی باشند میانگین آن‌ها برابر $\frac{8+19}{2} = 13/5$ است.

فصل هفتم: شاخص‌های پراکندگی

فرض کنید نمرات دو کلاس A و B به صورت مقابل باشند:

$$\begin{cases} \text{نمرات کلاس A}: 10, 10, 10, 20, 20, 20 \\ \text{نمرات کلاس B}: 14, 14, 14, 16, 16, 16 \end{cases}$$

به راحتی می‌توان فهمید که میانگین نمرات هر دو کلاس برابر ۱۵ است. ولی این دو کلاس از جهتی با هم تفاوت دارند و آن «پراکندگی نمرات» است. توجه کنید پراکندگی نمرات در کلاس A خیلی بیش‌تر از کلاس B است. به عبارت دیگر در کلاس B نمرات به هم نزدیک‌تر هستند. شاخص‌هایی که میزان پراکندگی داده‌ها را در یک جامعه‌ی آماری مشخص می‌کنند، «شاخص‌های پراکندگی» نام دارند. این شاخص‌ها عبارت‌اند از دامنه‌ی تغییرات، واریانس، انحراف معیار و ضریب تغییرات که به بررسی آن‌ها می‌پردازیم.

همان‌طـور کـه قبلاً گفتیم اختلاف بین بزرگ‌ترین و کوچک‌ترین داده را «دامنـه‌ی تغییرات» می‌گوییم. به عنوان نمونه، دامنه‌ی تغییرات نمرات کلاس A، برابر ۱۰ و دامنه‌ی تغییرات نمرات کلاس B، برابر ۲ است. همان طور که ملاحظه می‌کنید، دامنه‌ی تغییرات کلاس A بیش‌تر از دامنه‌ی تغییرات کلاس B است.

واریانس داده‌های آماری $x_1, x_2, \ldots, x_n$ را با نماد σ^2 نمایش می‌دهند و به یکی از دو صورت زیر محاسـبه می‌شـود: (σ از حروف کوچک یونانی است و سیگما خوانده می‌شود. حرف بزرگ آن $\sum$ است.)

$$\begin{cases} \sigma^2 = \dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \dfrac{\sum(x_i-\bar{x})^2}{n} \\ \sigma^2 = \dfrac{x_1^2+x_2^2+\ldots+x_n^2}{n} - (\bar{x})^2 = \dfrac{\sum x_i^2}{n} - (\bar{x})^2 \end{cases}$$

مثـال: برای محاسبه‌ی واریانس داده‌های $1, 2, 3, 4, 5$ با هر دو فرمول ابتدا میانگین را محاسبه می‌کنیم: $\bar{x} = \dfrac{1+2+3+4+5}{5} = 3$

اکنون طبق فرمول اول واریانس می‌توان نوشت:

$$\sigma^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5} = \frac{4+1+0+1+4}{5} = 2$$

حال از فرمول دوم واریانس این داده‌ها را محاسبه می‌کنیم:

$$\sigma^2 = \frac{1^2+2^2+3^2+4^2+5^2}{5} - (3)^2 = \frac{55}{5} - 9 = 2$$

تست ۱۶: اگر تفاضل داده‌ها از میانگین داده‌ها به‌صورت $-4, 0, 2, 9$ و a باشد، واریانس داده‌ها کدام است؟

۱) ۲۰ **۲)** ۲۵ **۳)** ۳۰ **۴)** ۳۵

پاسخ: می‌دانیم مجموع انحراف از میانگین داده‌ها همواره برابر صفر است، پس می‌توان نوشت:

$$a-4+0+2+9=0 \Rightarrow a=-7$$

اکنون می‌توان واریانس را به‌صورت زیر محاسبه کرد:

$$\sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{(-7)^2+(-4)^2+0^2+2^2+9^2}{5} = \frac{150}{5} = 30$$

بنابراین گزینه‌ی (۳) درست است.

تذکـر همان‌طور که ملاحظه کردید فرقی نمی‌کند از کدام فرمول واریانس استفاده شود. فقط در مواقعی که مجموع مربعات داده‌ها به نحوی در مسأله مطرح باشد، بهتر است از فرمول دوم واریانس استفاده کنیم.

مثـال: اگر مجموع ۲۰ داده‌ی آماری برابر ۱۴۰ و مجموع مربعات آن‌ها برابر ۱۹۰۰ باشد، برای محاسبه‌ی واریانس داده‌ها می‌توان نوشت:

$$\bar{x} = \frac{140}{20} = 7$$

اکنون می‌توان واریانس را از فرمول دوم محاسبه کرد: (توجه کنید که $x_1^2+x_2^2+\ldots+x_{20}^2 = 1900$)

$$\sigma^2 = \frac{1900}{20} - 7^2 = 95 - 49 = 46$$

تذکـر اگر داده‌ها دسته‌بندی شده باشند و نشان دسته‌های آن‌ها به‌ترتیب x_1، x_2، ... و x_n باشد، واریانس آن‌ها را به‌صورت مقابل محاسبه می‌کنیم:

$$\sigma^2 = \frac{\sum f_i(x_i-\bar{x})^2}{\sum f_i}$$

تست ۱۷: در جدول فراوانی تجمعی داده‌های دسته‌بندی شده‌ی مقابل، واریانس چقدر است؟

حدود دسته	۰-۲	۲-۴	۴-۶	۶-۸
فراوانی تجمعی	۱	۳	۱۲	۱۶

۱) ۱/۷۵ ۲) ۲

۳) ۲/۲۵ ۴) ۲/۵

پاسخ: جدول فراوانی مطلق داده‌های مورد نظر به‌صورت مقابل است:

حدود دسته	۰-۲	۲-۴	۴-۶	۶-۸
فراوانی مطلق	۱	۲	۹	۴

بنابراین میانگین و واریانس این داده‌ها به‌صورت زیر به‌دست می‌آید:

$$\bar{x}=\frac{۱\times۱+۲\times۳+۹\times۵+۴\times۷}{۱+۲+۹+۴}=\frac{۸۰}{۱۶}=۵ \quad , \quad \sigma^۲=\frac{۱(۱-۵)^۲+۲(۳-۵)^۲+۹(۵-۵)^۲+۴(۷-۵)^۲}{۱۶}=۲/۵$$

بنابراین گزینه‌ی (۴) درست است.

نکاتی درباره‌ی واریانس

۱- اگر داده‌ها برابر باشند، واریانس آن‌ها صفر است و برعکس، یعنی اگر واریانس صفر باشد، در این صورت همه‌ی داده‌ها با هم برابرند.

مثال: واریانس داده‌های ۱۳۹۳,۱۳۹۳,۱۳۹۳,۱۳۹۳ برابر صفر است.

مثال: اگر واریانس داده‌های $b-۲, a-۱, ۷$ برابر صفر باشد می‌توان دریافت همه‌ی داده‌ها با هم برابرند. پس می‌توان نوشت:

$$a-۱=۷ \Rightarrow a=۸ \quad , \quad b-۲=۷ \Rightarrow b=۹$$

۲- واحد واریانس مجذور واحد داده‌هاست. به عنوان مثال اگر داده‌ها برحسب متر باشند، واریانس برحسب مترمربع است. (این موضوع، یکی از معایب استفاده از واریانس است.)

۳- اگر داده‌های $x_۱, x_۲, \ldots, x_n$ را با عدد ثابت b جمع (یا تفریق) کنیم، واریانس داده‌ها تغییری نمی‌کند. ولی اگر این داده‌ها را در عدد حقیقی a ضرب کنیم، واریانس اولیه در $a^۲$ ضرب می‌شود. این مطلب را به صورت زیر می‌توان نوشت:

$$\sigma^۲_{ax\pm b}=a^۲\sigma^۲_x$$

مثال: اگر واریانس داده‌های $x_۱, x_۲, \ldots, x_n$ برابر ۸ باشد، واریانس داده‌های $x_۱+۳, x_۲+۳, \ldots, x_n+۳$ نیز برابر ۸ است.

مثال: اگر واریانس داده‌های $x_۱, x_۲, \ldots, x_n$ برابر ۵ باشد، واریانس داده‌های $۲x_۱, ۲x_۲, \ldots, ۲x_n$ برابر است با: $۴\times۵=۲۰$

مثال: اگر واریانس داده‌های $x_۱, x_۲, \ldots, x_n$ برابر ۷ باشد، واریانس داده‌های $۳x_۱+۱۳, ۳x_۲+۱۳, \ldots, ۳x_n+۱۳$ برابر است با: $۹\times۷=۶۳$

تست ۱۸: واریانس ۲۵ داده‌ی آماری برابر ۴ است. اگر ۵ تا از داده‌ها را که با میانگین برابر هستند از بین داده‌ها حذف کنیم، واریانس ۲۰ داده‌ی باقی‌مانده کدام است؟

۱) ۴ ۲) ۵ ۳) ۹ ۴) ۸

پاسخ: از آن‌جا که ۵ تا از ۲۵ داده‌ی اولیه برابر با میانگین است، می‌توان داده‌های اولیه را به صورت زیر درنظر گرفت:

$$\bar{x}, \bar{x}, \bar{x}, \bar{x}, \bar{x}, x_۱, x_۲, \ldots, x_{۲۰}$$

حال می‌توان نوشت:

$$\sigma^۲=\frac{\sum(x_i-\bar{x})^۲}{n} \Rightarrow ۴=\frac{\sum(x_i-\bar{x})^۲}{۲۵} \Rightarrow \sum(x_i-\bar{x})^۲=۱۰۰$$

با حذف ۵ داده‌ای که با میانگین برابر هستند، تغییری در میانگین و حاصل عبارت $\sum(x_i-\bar{x})^۲$ رخ نمی‌دهد. پس واریانس ۲۰ داده‌ی باقی‌مانده برابر است با:

$$\sigma^۲=\frac{\sum(x_i-\bar{x})^۲}{۲۰}=\frac{۱۰۰}{۲۰}=۵$$

بنابراین گزینه‌ی (۲) درست است.

همان‌طور که ملاحظه کردید واحد واریانس مجذور واحد داده‌هاست. برای رفع این مشکل، معمولاً در بررسی‌های آماری از جذر واریانس استفاده می‌کنند. جذر واریانس را «انحراف معیار» می‌گویند که آن‌را با σ نمایش می‌دهند و به کمک یکی از دو فرمول زیر محاسبه می‌شود:

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

تذکر همان‌طور که از فرمول نیز مشخص است، واحد انحراف معیار با واحد داده‌ها یکسان است.

مثال: واریانس داده‌های $1,2,3,4,5$ برابر ۲ است، پس انحراف معیار آن‌ها برابر $\sqrt{2}$ است.

ویژگی‌های انحراف معیار

۱- اگر همه‌ی داده‌ها با هم برابر باشند، انحراف معیار برابر صفر است و برعکس.

۲- اگر به داده‌ها عدد ثابت b را اضافه (یا کم) کنیم، انحراف معیار آن‌ها تغییری نمی‌کند. ولی اگر همه‌ی داده‌ها را در عدد a ضرب کنیم، انحراف معیار آن‌ها در $|a|$ ضرب می‌شود. به‌طور خلاصه می‌توان نوشت: $\sigma_{ax\pm b}=|a|\sigma_x$

مثال: اگر انحراف معیار داده‌های $x_1, x_2, \ldots, x_n$ برابر ۴ باشد، انحراف معیار داده‌های $-2x_1+3, -2x_2+3, \ldots, -2x_n+3$ برابر است با:

$|-2|\times 4=8$

تست ۱۹: اگر انحراف معیار داده‌های مجموعه‌ی $A=\{2,4,6,8,10,12,14\}$ برابر a باشد، آن‌گاه انحراف معیار داده‌های مجموعه‌ی $B=\{5,6,7,8,9,10,11\}$ کدام است؟

۱) $2a$ ۲) $\frac{a}{2}$ ۳) $\frac{a}{4}$ ۴) $4a$

پاسخ: توجه کنید که اگر داده‌های مجموعه‌ی A را نصف کنیم و به هر کدام از آن‌ها ۴ واحد اضافه کنیم داده‌های مجموعه‌ی B به دست می‌آیند. بنابراین اگر داده‌های مجموعه‌ی A را با x_i نمایش دهیم، داده‌های مجموعه‌ی B به صورت $\frac{1}{2}x_i+4$ هستند پس انحراف معیار داده‌های مجموعه‌ی B برابر است با $\frac{1}{2}a$. بنابراین گزینه‌ی (۲) درست است.

نکته: اگر داده‌های x_1، x_2، ... و x_n دنباله‌ای حسابی با قدر نسبت d باشند، واریانس و انحراف معیار آن‌ها را می‌توان از فرمول زیر محاسبه کرد:

$$\sigma^2=\left(\frac{n^2-1}{12}\right)d^2 \quad , \quad \sigma=|d|\sqrt{\frac{n^2-1}{12}}$$

مثال: مقادیر واریانس و انحراف معیار داده‌های ۱، ۴، ۷، ۱۰ و ۱۳ را محاسبه می‌کنیم. توجه کنید که این داده‌ها دنباله‌ای حسابی با قدر نسبت $d=3$ هستند و تعداد آن‌ها $n=5$ است. پس می‌توان نوشت:

$$\sigma^2=\left(\frac{25-1}{12}\right)\times 9=18 \quad , \quad \sigma=\sqrt{18}=3\sqrt{2}$$

همان‌طور که ملاحظه کردید انحراف معیار با داده‌ها «هم‌واحد» است. به عنوان مثال انحراف معیار قد برحسب «متر» و انحراف معیار وزن برحسب «کیلوگرم» است. برای آن‌که بتوانیم پراکندگی‌های دو نوع داده، با واحدهای متفاوت را با هم مقایسه کنیم از شاخصی «بدون واحد» که به آن «ضریب تغییرات» گفته می‌شود، استفاده می‌کنیم. ضریب تغییرات را با CV نمایش می‌دهند و به صورت زیر تعریف می‌شود:

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

پس ضریب تغییرات از تقسیم انحراف معیار بر میانگین محاسبه می‌شود.

مثـال: انحراف معیار و میانگین داده‌های $۱,۲,۳,۴,۵$ به ترتیب برابر $\sqrt{۲}$ و ۳ است. پس ضریب تغییرات این داده‌ها برابر است با:

$$CV=\frac{\sigma}{\overline{x}}=\frac{\sqrt{۲}}{۳}$$

تذکـر از آن‌جایی که ضریب تغییرات، معیاری برای میزان پراکندگی است، باید مثبت باشد. لذا ضریب تغییرات را فقط برای داده‌های مثبت تعریف می‌کنیم.

تست ۲۰: اگر $x_۱, x_۲, \ldots$ و x_n داده‌هایی با میانگین بزرگ‌تر از ۳ باشند، آن‌گاه ضریب تغییرات کدام دسته از داده‌های زیر بیش‌تر است؟

۱) $۳x_۱+۲, ۳x_۲+۲, \ldots, ۳x_n+۲$

۲) $۳x_۱+۱, ۳x_۲+۱, \ldots, ۳x_n+۱$

۳) $x_۱+۴, x_۲+۴, \ldots, x_n+۴$

۴) $۳x_۱-۱, ۳x_۲-۱, \ldots, ۳x_n-۱$

پاسخ: اگـر میانگیـن و انحراف معیار داده‌های $x_۱, x_۲, \ldots$ و x_n را به ترتیب با $\overline{x}$ و σ نمایش دهیـم، می‌توانیم گزینه‌ها را به صورت زیر بررسی کنیم.

گزینه‌ی (۱): انحراف معیار داده‌های $۳x_۱+۲, ۳x_۲+۲, \ldots$ و $۳x_n+۲$ برابر است با: $CV_۱=\frac{۳\sigma}{۳\overline{x}+۲}$

گزینه‌ی (۲): انحراف معیار داده‌های $۳x_۱+۱, ۳x_۲+۱, \ldots$ و $۳x_n+۱$ برابر است با: $CV_۲=\frac{۳\sigma}{۳\overline{x}+۱}$

گزینه‌ی (۳): انحراف معیار داده‌های $x_۱+۴, x_۲+۴, \ldots$ و $x_n+۴$ برابر است با: $CV_۳=\frac{\sigma}{\overline{x}+۴}=\frac{۳\sigma}{۳\overline{x}+۱۲}$

گزینه‌ی (۴): انحراف معیار داده‌های $۳x_۱-۱, ۳x_۲-۱, \ldots$ و $۳x_n-۱$ برابر است با: $CV_۴=\frac{۳\sigma}{۳\overline{x}-۱}$

اکنون با مقایسـه‌ی مقادیـر $CV_۱, CV_۲, CV_۳$ و $CV_۴$، مقدار $CV_۴$ از همه‌ی مقادیر بزرگ‌تر اسـت (زیرا مخرج عبارت $CV_۴$ از بقیه‌ی عبارت‌ها کوچک‌تر است!)

بنابراین گزینه‌ی (۴) درست است.

پراکندگی در منحنی نرمال

منحنی نرمال مقابل را در نظر بگیرید. تقریباً ٪۶۸ داده‌ها در بازه‌ای به مرکز میانگین و شعاع انحراف معیار هستند.

بـه همین ترتیب تقریبـاً ٪۹۶ از داده‌ها در بازه‌ای به مرکز میانگین و شـعاع دو برابر انحراف معیار هستند.

و سـرانجام تقریباً ٪۱۰۰ از داده‌ها در بازه‌ای به مرکز میانگین و شـعاع سـه برابر انحراف معیار قرار دارند.