به نام خدا

عنوان:

جزو درس مبانی احتمال

این درس دارای سرفصل‌های

- ✓ تعاریف و مقدمات اولیه
- ✓ آمار توصیفی
- ✓ احتمال

می باشد.

منابع مفید پیشنهادی

- ✓ آمار و احتمال مقدماتی (دکتر جواد بهبودیان)
- ✓ آمار و احتمالات مهندسی (دکتر نادر نعمت اللهی)
- ✓ آمار و احتمال (دکتر پرویز نصیری)

تعاریف و مقدمات اولیه

تعریف علم آمار: مجموعه‌ای از ابزارها و روش‌ها بمنظور جمع‌آوری، دسته‌بندی، خلاصه‌سازی، تجزیه و تحلیل، تفسیر و ارائه گزارش از داده‌ها را علم آمار گویند.

جامعه آماری: گردایه‌ای از افراد یا اشیاءِ که حداقل در یک ویژگی مشابه هستند.

نکته ۱: هر جامعه آماری وابسته به نوع تعریف و موضوع مورد تحقیق به واحدهای کوچکی تقسیم بندی می شود که آنها را اعضای جامعه آماری گویند.

مثال: چند مثال از جامعه آماری

۱) دانشجویان رشته عمران دانشگاه خواجه نصیرالدین طوسی

۲) ساکنین شهر تهران

۳) پرندگان مهاجر به تالاب‌های ایران

۴) دانش آموزان مدارس دولتی استان فارس

۵) مشتریان بانک ملی

انواع داده: داده‌ها دارای دو دسته عمده کمی (قابل اندازه‌گیری) و کیفی هستند.

روش‌های جمع‌آوری داده: پس از تعیین حدود و واحدهای جامعه بمنظور کشف ناشناخته‌ها و مجهولات، نیازمند جمع‌آوری داده از آن را داریم. بدین منظور از دو روش

- **سرشماری:** جمع‌آوری داده از تمامی اعضای جامعه را سرشماری گویند.

نکته ۲: در سرشماری نتایج حاصل دقیق و دارای بیشترین اعتبار می باشند ولی در عمل استخراج تمامی اطلاعات از تمامی اعضاءِ جامعه آماری کاری زمانبر، نیازمند نیروی انسانی زیاد، هزینه‌بر، پر دردسر و ... (در بعضی موارد هم غیرممکن) می باشد.

❖ **نمونه‌گیری:** اخذ کسری از اعضای جامعه آماری بعنوان واحدهای مورد مطالعه را نمونه-گیری گویند.

استفاده می گردد.

جامعه آماری همگن: هرگاه موضوع مورد مطالعه بصورت یکسان و مشابه در جامعه آماری توزیع شده باشند، به آن جامعه همگن گویند (در غیر اینصورت به آن ناهمگن گویند).

مثال: هرگاه بخواهیم به موضوع دیابت و عوامل موثر بر آن در بین ساکنین شهر تهران بپردازیم با یک جامعه همگن مواجه هستیم، زیرا امکان ابتلا به دیابت در بین افراد یکسان و مشابه است.

این درحالی هست که اگر بخواهیم به موضوع میزان هزینه کرد ماهیانه خانوارهای ساکن تهران بپردازیم دیگر جامعه آماری (ساکنین شهر تهران) تشکیل یک جامعه ناهمگن را خواهند داد.

نکته: تعداد اعضای جامعه و تعداد نمونه را به ترتیب با N و n نمایش می دهیم.

انواع نمونه‌گیری: وابسته به تعریف جامعه آماری و موضوع مورد تحقیق روش‌های مختلفی و علمی بمنظور انتخاب واحدهای نمونه در علم آمار در حال گسترش می باشد. در اینجا تنها به ارائه ٤ روش پرکاربرد نمونه‌گیری می پردازیم.

❖ **نمونه‌گیری تصادفی ساده:** این روش مناسب جامعه‌ی همگن می باشد که در آن موضوع مورد بررسی در سراسر جامعه بطور یکسان توزیع شده باشند. بمنظور اخذ نمونه تنها کافی است از هر جای جامعه به تعداد مورد نظر نمونه تصادفی انتخاب نمایید.

نمونه تصادفی: هرگاه اعضای انتخابی از یک جامعه مستقل از یکدیگر انتخاب شوند، آن نمونه را نمونه تصادفی گویند.

❖ **نمونه‌گیری طبقه‌بندی:** هرگاه جامعه آماری بطور ذاتی یا مصنوعی به دسته‌هایی تقسیم-بندی شده باشد بطوریکه درون دسته یا طبقات همگنی مشاهده شود و در بین طبقات ناهمگنی، بهترین روش نمونه‌گیری استفاده از روش طبقه‌بندی می باشد. بمنظور بدست آوردن اعضای نمونه کافیست از هر طبقه تعدادی نمونه (حداقل ۳نفر) اخذ شوند. یکی از بهترین روش‌های تعیین حجم نمونه هر طبقه استفاده از قاعده نسبت می باشد.
بمنظور اخذ نمونه در روش طبقه‌بندی فرض کنید حجم طبقه iام برابر N_i باشد و تعداد نمونه کل مدنظر برابر n در این حالت تعداد نمونه طبقه iام برابر $n_i = \frac{n}{N} N_i = \frac{N_i}{N} n$ که در آن $N = \sum_i N_i$ است.

نکات مربوط به نمونه‌گیری طبقه‌بندی

۱- حداقل تعداد نمونه در هر طبقه می بایست ۳ عدد باشد.
۲- اگر عدد بدست آمده از فرمول $n_i = \frac{n}{N} N_i$ اعشار باشد، باید به عدد صحیح بزرگتر از خود گرد شود. بعنوان مثال اگر تعداد حاصل از فرمول $n_i = \frac{n}{N} N_i = 15.12$، آنگاه تعداد نمونه طبقه iام را ۱۶ در نظر بگیرید.

مثال: در نظر داریم عوامل موثر در رده‌های سنی مختلف (نوجوان، جوان، میانسال و افراد مسن) مبتلا به دیابت در شهر تهران را مورد بررسی قرار دهیم. فرض کنید می دانیم به ترتیب ۱۲، ۱۸،

۲۵ و ۳۰ درصد جامعه تهران دارای این رده‌های سنی هستند و همچنین تعداد نمونه مدنظر ۱۸۰ نفر می باشد. تعداد نمونه ممکن از هر رده سنی را مشخص نمایید.

روش نمونه‌گیری مورد استفاده روش طبقه‌بندی می باشد، زیرا می بایست از هر رده سنی در مطالعه نمونه داشته باشیم و بین این رده‌های سنی اختلاف سنی قابل توجه می باشد. نسبت واقعی $\frac{N_i}{N}$ در ٤ رده سنی به ترتیب $0.12 \times \frac{100}{85}$، $0.18 \times \frac{100}{85}$، $0.25 \times \frac{100}{85}$ و $0.3 \times \frac{100}{85}$. بنابراین تعداد نمونه مدنظر از هر رده سنی برابر خواهد بود با عدد بدست آمده از فرمول فوق ضربدر ۱۸۰ که به ترتیب ۲۷، ٤۱، ۵۳ و ٦٤ فرد بدست می آید.

- **نمونه‌گیری سیستماتیک:** هرگاه واحدهای جامعه آماری بر اساس موضوع مورد مطالعه دارای یک نظم زمانی (تقدم و تاخر) باشند همانند مشترکین روزنامه، بانک، مخابرات و غیره. بهترین اخذ واحدهای نمونه در چنین جامعه‌ای نمونه‌گیری سیستماتیک می باشد. نحوه انتخاب واحدهای جامعه دارای گام‌های ادامه می باشد.
  - گام اول: عدد $k = \frac{N}{n}$ محاسبه کنید
  - گام دوم: ایجاد حداقل $n - 1$ دسته‌ی kتایی (عدد بدست آمده در گام اول باید به بالا گرد شود در صورت اعشاری بودن)
  - انتخاب عددی تصادفی همانند r در بین اعداد صحیح $1 \ldots k$
  - انتخاب واحدهای نمونه یعنی rامین، $(r + k)$ام، تا $(r + (n - 1)k)$امین نفر عضو جامعه

مثال: فرض کنید در نظر داریم میزان رضایت مشترکین روزنامه جام جم را مورد بررسی قرار دهیم. با توجه به مدت زمان ۲۰ ساله انتشار این روزنامه و قدمت افراد مختلف بعنوان مشترک نیازمند بهره از روش نمونه‌گیری سیستماتیک می باشیم. فرض کنید تعداد این مشترکین ٤۲۰۰۰ نفر باشد و در نظر داریم یک نمونه ۳۰۰ نفره از این افراد تهیه نماییم.

$k=\frac{42000}{300}=140$ عدد تصادفی r را مقدار ۸۵ انتخاب می نماییم. آنگاه واحدهای منتخب نمونه دارای شماره اشتراک

$$85, \qquad 140+85, \qquad 85+2*140, \ldots, \quad 85+299*140=4145.$$

نکته ۳: اگر مقدار k در گام اول اعشاری باشد وقتی آنرا (به بالا) گرد می کنیم، تعداد واحد در دسته اخر (حاصل از گام دوم) کمتر از k خواهد بود. در این حالت بهتر است در دسته آخر نیز به تصادف یک واحد را انتخاب نماید.

- **نمونه‌گیری خوشه‌ای:** هرگاه دسته یا طبقات ذاتی یا ساخته شده جامعه آماری به گونه‌ای باشند که در بین دسته‌ها همگنی و در درون طبقات امکان وجود ناهمگنی داشته باشیم، از نمونه‌گیری تحت عنوان خوشه‌ای بهره می بریم. به این طبقات همگن خوشه گویند. بعنوان مثال می خواهیم به بررسی میزان رضایت مشتریان بانک ملی از خدمات ارائه شده در شعب این بانک بپردازیم. می دانیم همه شعب موظف به ارائه همه خدمات بانکی مربوطه هستند بنابراین تفاوتی بین شعب وجود ندارد و می تواند مشتریان بسیار متفاوتی از منظر رضایت داشته باشیم. بمنظور نمونه‌گیری در این جوامع ابتدا می بایست از بین خوشه‌ها تعدادی را به تصادف انتخاب و تمامی واحدهای داخل آنرا مورد مطالعه قرار داد.

مثال: در نظر داریم به برررسی میزان رضایت دانش آموزان از محتوای فیزیک سال ۱۱ در دبیرستان A بپردازیم. فرض کنید در این دبیرستان ٤ کلاس سال ۱۱ وجود داشته باشد. بدلیل یکسان بودن کتاب فیزیک تدریسی در تمامی کلاس‌های سال ۱۱ روش نمونه‌گیری مناسب، روش خوشه‌ای است. بدین منظور ۲کلاس از بین ٤ کلاس موجود را به تصادف انتخاب و از دانش آموزان کلاس‌های انتخابی سوالات مدنظر را می پرسیم.

نکته ٤: به این نوع نمونه‌گیری خوشه‌ای تک مرحله‌ای گویند. حال اگر درون هر خوشه انتخاب شده نیز به تصادف تعدادی را انتخاب نماییم به این روش نمونه‌گیری خوشه‌ای دو مرحله‌ای گویند. این نمونه‌گیری می تواند بیش از دو مرحله‌ای نیز اجرا شود.

نکته ۵: با توجه به واحدهای جامعه و موضوع مورد بررسی می توان از ترکیب این نوع نمونه‌گیری-ها بصورت همزمان نیز بهره برد.

انواع مقیاس: ابزارهای آماری بر اساس مقادیر عددی کار میکنند، بنابراین لازم است که داده‌ها تبدیل به عدد شوند. این تبدیل داده به عدد را مقیاس‌بندی گویند و در ادامه انواع مقیاس بندی را ارائه می نماییم.

- **مقیاس اسمی**: این مقیاس صفر مطلق ندارد و اعداد حاصل از آن هیچ ارزشی عددی ندارد. همانند جنسیت، گروه خونی، محل سکونت و ...
- **مقیاس ترتیبی**: صفر مطلق ندارد و تنها بزرگتری یا کوچکتری در آن بی معناست بدون هیچ ارزش ریاضی برای اعداد حاصل از این نوع مقیاس. همانند میزان تحصیلات، میزان رضایت (خیلی کم، کم، متوسط، زیاد، خیلی زیاد)، رده سنی و غیره
- **مقیاس فاصله‌ای**: این تبدیل عددی بدلیل نداشتن صفر مطلق ارزش عددی (جمع، تفریق، ضرب و تقسیم) ندارد. با در برداشتن خاصیت مقیاس ترتیبی، همچنین ارزش فاصله بین اعداد در آن حفظ می شود. همانند دما، نمره هوش، فشار خون، میزان چربی و غیره
- **مقیاس نسبتی**: کاملترین نوع مقیاس بدلیل داشتن صفر مطلق این نوع مقیاس می باشد. همانند جرم، طول، حجم و ...

متغیر: مقادیر که از هر عضو جامعه به عضو دیگر می تواند تغییر نماید را متغیر گویند.

انواع متغیرها: اعداد حاصل از متغیرها دو حالت گسسته (حاصل از مقیاس‌های اسمی و ترتیبی) و پیوسته (حاصل از مقیاس‌های فاصله‌ای و نسبتی) را شامل می شوند.

ابزارهای آماری: ابزارهای آماری شامل

- **آمار توصیفی**: ارائه نتایج درباره یک متغیر به تنهایی
- **آمار استنباطی**: ارائه نتایج از روابط، مدل‌ها، تفاوت و علت و معلول در بین متغیرها

می شوند.

تعدادی از روابط مفید ریاضی

$$\sum_{j=1}^{n} \mathrm{a_j} = \mathrm{a_1} + \cdots + \mathrm{a_n},$$

$$\sum_{j=1}^{n} \mathrm{a_j} = \sum_{j=1}^{k} \mathrm{a_j} + \sum_{j=k+1}^{n} \mathrm{a_j}; \quad \text{for} \quad k = 1.\ldots.n,$$

$$\sum_{j=1}^{n} \mathrm{b} = \mathrm{nb},$$

$$\sum_{j=1}^{n} \mathrm{ba_j} = \mathrm{b}\sum_{j=1}^{n} \mathrm{a_j}. \quad \sum_{j=1}^{n} \mathrm{b(a_j + c)} = \mathrm{b}\sum_{j=1}^{n} \mathrm{a_j} + \mathrm{nbc},$$

$$\sum_{j=1}^{n} \mathrm{a_j^r} = \mathrm{a_1^r} + \cdots + \mathrm{a_n^r},$$

$$\left(\sum_{j=1}^{n} \mathrm{a_j}\right)^2 = \sum_{j=1}^{n} \mathrm{a_j^2} + 2\sum_{j=1}^{n-1}\sum_{k>j}^{n} \mathrm{a_j a_k},$$

$$\sum_{j=1}^{\mathrm{n}} \mathrm{a_j} = \sum_{j=r}^{\mathrm{n+r}} \mathrm{a_{j-r+1}} = \sum_{j=0}^{\mathrm{n-1}} \mathrm{a_{j+1}}.$$

نکته ۶: از این به بعد مقادیر حاصل از n نمونه ها را با

$$x_1, x_2, \dots, x_n$$

حال اگر نمونه‌ها را از کوچک به بزرگ مرتب کنیم این حالت مرتب شده را با

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

نمایش میدهیم. این بدین معنا است که

$x_{(1)}$ یعنی کوچکترین عدد مشاهده شده یا $x_{(1)} = \min(x_1. \dots . x_n)$

$x_{(n)}$ یعنی بزرگترین عدد مشاهده شده یا $x_{(n)} = \max(x_1. \dots . x_n).$

چه نوع داده‌ای برای تحلیل مناسب است؟

جواب: داده ای برای تحلیل مناسب است که دارای ۳ ویژگی زیر باشد

- وجود داشته باشد
- از کیفیت لازم برخوردار باشد
- از صحت داده ها مطمئن باشیم

آمار توصیفی

ارائه اعداد و ارقام برای متغیرهای مورد مطالعه بدون در نظر گرفتن دلیل، عوامل موثر مدل حاکم و اثراتی که میتواند بر متغیرهای دیگر داشته باشد، را آمار توصیفی گویند.

در ادامه آمار توصیفی و روش های بیان شده غالباً برای متغیرهایی است که پیوسته هستند؛ زیرا متغیرهای گسسته ابزار توصیفی بسیار محدودی دارند.

نکته ۵: به منظور استخراج معیارهای توصیفی از مشاهدات با دو رویکرد مواجه هستیم:

۱- استفاده از مشاهدات خام: مشاهداتی که هیچ تغییر و تبدیلی روی آن ها انجام نشده است؛ این رویکرد بیشتر در تعداد نمونه کم قابل استفاده است.

۲- استفاده از مشاهدات طبقه‌بندی شده: برای تعداد نمونه زیاد استفاده می شود و قبل از انجام محاسبات مربوط به معیارها می بایست آنها را دسته بندی نماییم.

معیارهای توصیفی:

- معیارهای تمرکز (مرکزی)
- معیارهای پراکندگی (توزیع)
- نمودارها (بصری)

انواع معیارهای تمرکز:

- ✓ **میانگین‌ها**
- ✓ **میانه** (M_d): نقطه‌ای است درون مشاهدات که نیمی از آنها قبل از این مقدار قرار دارند.
- ✓ **مد یا نما** (M_o): نقطه یا نقاطی که بیشترین فراوانی را دارند.
- ✓ **چندک** (q_r): نقطه است درون مشاهدات که $100r$ درصد مقادیر قبل آن قرار دارند ($r\epsilon(0,1)$).

انواع میانگین‌ها

۱–میانگین حسابی ($\bar{X}$): برای داده‌هایی از یک جامعه – هم واحد و هم ارزش $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

۲– میانگین وزنی ($\bar{X}_w$): قابل استفاده برای مقادیر حاصل از یک جامعه هم واحد ولی با ارزش‌های عددی متفاوت (همانند معدل ترم)

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}; \quad w_i \geq ۰.$$

۳– میانگین هندسی: بیشترین کاربرد این نوع میانگین برای بدست آوردن متوسط حاصل از نرخ رشد، نسبت و یا درصد که لزوما از یک جامعه نیستند استفاده می شود. این اعداد ($X_1, X_2, \dots, X_n$) لزوما هم ارزش نیستند ($w_1, w_2, \dots, w_n$) اما مثبتند و غیر صفر.

$$G = \sqrt[\sum_{i=1}^{n} w_i]{X_1^{w_1} \times X_2^{w_2} \times \dots \times X_n^{w_n}}.$$

سوال: سرمایه‌ی یک شرکت بر حسب میلیون در ۱٤ سال گذشته به فرم زیر است.

52 30 40 45 ۳۵ ۳۰ ٤۰ ۳۵ ۱۵ ۱۰ ۸ ۲۵ ۳.۷ ۲

متوسط رشد سالیانه‌ی سرمایه این شرکت مقادیر $\frac{52}{30}, \frac{30}{40}, \frac{40}{45}, \frac{45}{35}, \dots, \frac{8}{25}, \frac{25}{3.7}, \frac{3.7}{2}$ می باشند

$$G=\sqrt[13]{\frac{3.7}{2}\times\frac{25}{3.7}\times\frac{8}{25}\times\ldots\times\frac{45}{35}\times\frac{40}{45}\times\frac{30}{40}\times\frac{52}{30}}=\sqrt[13]{\frac{52}{2}}=1.285.$$

تفسیر: این بدان معنا است که متوسط رشد سرمایه این شرکت ۲۸,۵ واحد می باشد. یعنی انتظار داریم که با سرمایه‌گذاری در این شرکت سال آینده حدود ۲۸,۵ واحد سود دریافت نماییم.

متوسط رشد دوسالانه شرکت حاصل از مقادیر $\frac{52}{40},\frac{30}{45},\frac{40}{35},\frac{45}{30}$، $\frac{۳۵}{40}$، ...، $\frac{8}{3.7}$، $\frac{25}{2}$ می باشد که برابر است با

$$G=\sqrt[12]{\frac{25}{2}\times\frac{8}{3.7}\times\ldots\times\frac{45}{30}\times\frac{40}{35}\times\frac{30}{45}\times\frac{52}{40}}=1.562.$$

تفسیر: بطور متوسط انتظار داریم، سرمایه شرکت هر دو سال ۵۶,۲درصد افزایش پیدا نماید.

مثال: کاربرد میانگین هندسی در محاسبه تورم.

پوشاک	غلات	حبوبات	پروتئین	خودرو	مسکن	سوخت	اقلام اساسی
۸۰۰۰۰	۱۵۰۰	۸۰۰۰	٤۰	۱	۱	۱	شهریور۱۳۹۷
۱۸۰۰۰۰	۲۰۰۰	۱۲۰۰۰	۱۲۰	۲	۳	۱	شهریور۱۳۹۸
$\frac{9}{4}$	$\frac{4}{3}$	۱.۵	۳	۲	۳	۱	رشد سالیانه اقلام

$$G=\sqrt[7]{1\times3\times2\times3\times1.5\times\frac{4}{3}\times\frac{9}{4}}=1.873$$

متوسط تورم سالیانه ۸۷,۳ درصد می باشد. همچنین متوسط تورم ماهیانه برابر است با $\sqrt[12]{1.873}=1.054$ یعنی متوسط هر ماه ۵,٤ درصد افزایش قیمت داشته‌ایم.

مثال: بر اساس لینک خبری زیر افزایش قیمت از سوی وزارت صمت از قیمت کالاهای اساسی در دی ماه سال ۱۳۹۹ بشرح جدول است

https://www.iranjib.ir/shownews/80547/%D8%A7%D8%B9%D9%84%D8%A7%D9%85-%D8%AA%D8%BA%DB%8C%DB%8C%D8%B1%D8%A7%D8%AA-%D9%86%D8%B1%D8%AE-%DA%A9%D8%A7%D9%84%D8%A7%D9%87%D8%A7%DB%8C-%D8%A7%D8%B3%D8%A7%D8%B3%DB%8C-%D8%AF%D8%B1-%DB%8C%DA%A9-%D8%B3%D8%A7%D9%84-%DA%AF%D8%B0%D8%B4%D8%AA%D9%87/

اقلام	برنج پاکستانی	برنج تایلندی	برنج طارم	برنج هاشمی	شکر سفید	شکر ۹۰۰ گرمی	گوشت گوساله	گوشت گوسفندی	گوشت مرغ تازه
درصد رشد	۱۲۳	۱۲۹	٤۷	٤۵	٤۷	۵٤	۳۱	۲۷	۶۲

بنابراین مقادیر تورم در این اقلام بصورت متوسط

$$G = \sqrt[9]{223 * 229 * 147 * \dots * 162} = 159.31.$$

یعنی بطور متوسط ۵۹,۳۱درصد در این اقلام رشد قیمت یا همان تورم را تجربه داشته ایم.

٤- میانگین هارمونیک: این نوع میانگین بیشتر برای محاسبه متوسط سرعت زمانی است که سرعت‌ها با وزن‌های مختلف رخ داده‌اند. (تکلیف)

۵- میانگین درجه دوم: (تکلیف)

میانه[1] **(M_d):** مراحل محاسبه میانه: نقطه‌ای است در درون مشاهدات که ۵۰ درصد مقادیر قبل از آن قرار دارد.

[1] Median

۱- مرتب کردن مشاهدات از کوچک به بزرگ

۲- محاسبه
$$M_d = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{تعداد نمونه فرد باشد} \\ \frac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2} & \text{تعداد نمونه زوج باشد} \end{cases}$$

چندک (q_r): نقطه‌ای است در مشاهدات که $100r$ درصد مقادیر قبل آن قرار می گیرد، برای $r \in (0,1)$.

معروف‌ترین چندک‌ها:

۱-**چارک**[2](Q_r) : نقطه ای است در مشاهدات که $25r$ درصد مقادیر قبل از آن است. $r = 1, 2, 3$

۲- **دهک**[3](D_r) : نقطه‌ای است در مشاهدات که ۱۰r درصد مقادیر قبل از آن است. $r = 1, \ldots, 9$

۳- **صدک**[4](P_r) : نقطه‌ای است در مشاهدات که r درصد مقادیر قبل از آن است. $r = 1, \ldots, 99$

پر واضح است که $P_{75} = Q_3$ ، $P_{50} = D_5 = Q_2 = M_d$ و $P_{25} = Q_1$

نکته ۱: میانه یک نوع چندک است.

نکته ۲: صدک نسبت به سایر انواع چندک (چارک و دهک) کلی‌تر است. بدین معنا که همه چارک‌ها یا دهک‌ها نوعی صدک می باشند.

---

[2] Quantile
[3] Decade
[4] Percentile

چارک و دهک حالت‌های خاصی از صدک می باشند. $D_i = P_{10i}$ و $Q_i = P_{25i}$

روش محاسبه‌ی صدک rام (P_r):

۱- مرتب کردن مشاهدات (از کوچک به بزرگ)

۲- $k = \frac{n+1}{100} \times r$

۳- $s = [k], \quad w = k - s$

٤- $P_r = (1-w)X_{(s)} + wX_{(s+1)}$

مثال: برای داده‌های زیر که مربوط به معدل دانشجویان کلاس می باشد، دهک دوم را بدست آورید

16.4 – 14.21-12.08- 15.34- 15.86 – 14.96- 15.85- 13.42- 14.85 12.01-17.25

بنابراین $n = 11$ و مرتب شده مقادیر برابر است با

12.01 - 12.08 - 13.42 - 14.21 - 14.85 - 14.96 - 15.12 - 15.34 - 15.85 - 16.4 - 17.25

یعنی $D_2 = P_{20}$ در نتیجه $r = 20$. بنابراین $k = \frac{11+1}{100} \times 20 = 2.4$ و خواهیم داشت

$s = [2.4] = 2$ و $w = 2.4 - 2 = 0.4$ از این مقادیر دهک دوم برابر خواهد بود

$$D_2 = P_{20} = (1-0.4) \times 12.08 + 0.4 \times 13.42 = 12.616.$$

تفسیر: فرض کنید مقادیر مثال مربوط به معدل دانشجویان کلاس باشد، بنابراین می توان گفت ۲۰ درصد معدل کلاس مقداری کمتر یا مساوی ۱۲,۶۲ می باشد.

ب) چارک سوم معدل دانشجویان کلاس را مشخص نمایید.

$$Q_3 = P_{75} \Longrightarrow r = 75.$$

$$k = \frac{11+1}{100}75 = 9 \Longrightarrow s = 9,\ \ w = 0.$$

$$Q_3 = P_{75} = (1-0)15.85 + 0 * 16.4 = 15.85.$$

ج) صدک ۴۳ام را بدست آورید.

$$P_{43} \Longrightarrow r = 43.$$

$$k = \frac{11+1}{100}43 = 5.16 \Longrightarrow s = 5,\ \ w = 0.16.$$

$$P_{43} = (1-0.16)14.85 + 0.16 \times 14.96 = 14.8676.$$

نما یا مد (M_o): مشاهده یا مشاهداتی که بیشترین فراوانی را داشته باشند.

نکته مهم: مد یا نما می تواند وجود نداشته باشد (زمانیکه تعداد مشاهدات یکسان باشد). همچنین در صورت وجود مد می تواند منحصر به فرد نباشد. این درحالی است که میانه، انواع میانگین و چندک‌ها برای مجموعه‌ای از مشاهدات حتما وجود دارند و منحصر به فردند.

معیارهای پراکندگی
- دامنه (R) — $R = X_{(n)} - X_{(1)}$ — بهترین شاخص پراکندگی
- واریانس (S^2)
- انحراف معیار (S)
- انحراف قدر مطلقی (A)
- ضریب تغییرات (CV)

نکته: دامنه بیشترین اختلاف ممکن بین مشاهدات را نمایش می دهد.

واریانس: متوسط، مینیمم انحراف مشاهدات را واریانس گویند.

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}.$$

می دانیم

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \leftrightarrow \sum_{i=1}^{n} X_i = n\overline{X} \leftrightarrow \sum_{i=1}^{n} X_i - n\overline{X} = 0 \leftrightarrow \sum_{i=1}^{n}\left(X_i - \overline{X}\right) = ٠.$$

مثال: فرض کنید میانگین ٤ مشاهده برابر است با ١٥، بطوریکه یکی از اعداد 8، دیگری ١٧، و عدد سوم ١٤ می باشد، مقدار عدد چهارم را مشخص کنید؟

$$\overline{x} = 15, \qquad x_1 = 8, \qquad x_2 = 17, \qquad x_3 = 14 \;\Leftrightarrow x_4 = 21.$$

دلیل استفاده از $n-1$ در مخرج فرمول واریانس S^2، آن است که از $\overline{X}$ در محاسبه‌ی S^2 استفاده کرده‌ایم. هرگاه مقدار میانگین مشاهدات را بدانیم تنها مختاریم $n-1$ مقدار را بدلخواه انتخاب کنیم و nامین مقدار باید بگونه‌ای باشد که در فرمول میانگین صدق کند. بنابراین درجه آزادی واریانس برابر است با $n-1$.

حال نشان می دهیم صورت فرمول واریانس مینیمم انحراف مشاهدات را شامل شده است. بدین منظور برای هر $a \in \mathcal{R}$ تعریف می کنیم

$$l(a) = \sum_{i=1}^{n}(X_i - a)^2.$$

$l(a)$ نشان دهنده میزان انحراف مشاهدات از نقطه a می باشد. از دو روش این مهم را نشان می دهیم.

راه اول: استفاده از مشتق

$$\frac{dl(a)}{da}=\frac{d\sum_{i=1}^{n}(X_i-a)^2}{da}=-2\sum_{i=1}^{n}(X_i-a)=0\Longrightarrow a=\frac{\sum_{i=1}^{n}X_i}{n}=\overline{X}.$$

مشتق دوم را می گیریم و داریم

$$\frac{d^2l(a)}{da^2}=\frac{d(-2\sum_{i=1}^{n}(X_i-a))}{da}=2n\geq 0.$$

راه دوم: اثبات مستقیم

$$\begin{aligned}
l(a)=\sum_{i=1}^{n}(X_i-a)^2&=\sum_{i=1}^{n}\left(X_i-\overline{X}+\overline{X}-a\right)^2=\sum_{i=1}^{n}\left(\left(X_i-\overline{X}\right)+\left(\overline{X}-a\right)\right)^2\\
&=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n}\left(\overline{X}-a\right)^2+2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(\overline{X}-a\right)\\
&=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+n\left(\overline{X}-a\right)^2+2\left(\overline{X}-a\right)\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\\
&=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+n\left(\overline{X}-a\right)^2=l\left(\overline{X}\right)+n\left(\overline{X}-a\right)^2.
\end{aligned}$$

نشان دادیم که $l(a)=l\left(\overline{X}\right)+n\left(\overline{X}-a\right)^2$، این بدان معنا است که مینیمم مقدار تابع $l(a)$ در نقطه $\overline{X}$ ایجاد می شود.

در نگاهی مجدد به فرمول واریانس می بینیم که $S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$ مسیر دشواری برای محاسبه دارد، بنابراین صورت واریانس را براساس مطلب زیر ساده سازی می نماییم.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\overline{X})^2&=\sum_{i=1}^{n}\left(X_i^2-2X_i\overline{X}+\overline{X}^2\right)\\
&=\sum_{i=1}^{n}X_i^2-2\sum_{i=1}^{n}X_i\overline{X}+\sum_{i=1}^{n}\overline{X}^2
\end{aligned}$$

$$= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2$$
$$= \sum_{i=1}^{n} X_i^2 - 2\overline{X} \times n\overline{X} + n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \geq 0.$$

با توجه به مثبت بودن عبارت فوق داریم

$$\sum_{i=1}^{n} X_i^2 \geq n\overline{X}^2 \Longrightarrow \frac{\sum_{i=1}^{n} X_i^2}{n} \geq \overline{X}^2 \Longrightarrow \overline{X^2} \geq \overline{X}^2.$$

بنابراین فرمول واریانس بفرم محاسباتی ساده $S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$ تبدیل می شود.

نکته ۹: واریانس همواره بزرگتر مساوی صفر است و زمانی برابر صفر است که مقادیر همگی برابر باشند.

انحراف معیار: $S = \sqrt{S^2}$

در واریانس واحد مشاهدات به توان ۲ می رسد اما در انحراف معیار این اتفاق نمی افتد. بنابراین واحد اندازگیری انحراف معیار با مشاهدات یکسان می باشد.

انحراف قدر مطلقی:

$$A = \frac{\sum_{i=1}^{n} \left| X_i - \overline{X} \right|}{n},$$

متوسط قدرمطلق اختلاف مشاهدات از میانگین را انحراف قدر مطلقی می نامند.

ضریب تغییرات: $CV = \frac{S}{|\overline{X}|}$

مثال: میخواهیم عملکرد دو مدرس در تدریس یکسری از دانشجویان را بررسی نماییم. میانگین و انحراف معیار مدرس اول در درس با کد ۷۱۰۱ به ترتیب برابر ۱۵٫۳ و ۸ و همین مقادیر برای مدرس دوم در کد درس ۷۱۰۲ برابر است با ۱۳٫۵ و ۳. عملکرد این دو مدرس را مقایسه نمایید.

$$CV_1 = \frac{8}{15.3} = 0.523$$

$$CV_2 = \frac{3}{13.5} = 0.222$$

عملکرد مدرس دوم بهتر بوده است.

$$C.V. = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\bar{X}}\right)^2}{n-1}} \leftarrow \text{فرمول اصلی}$$

به بیان دیگر متوسط انحرافات نسبی مشاهدات از میانگین را ضریب تغییرات گویند.

از عمده ویژگی‌های ضریب تغییرات آن است که به واحد اندازه‌گیری مشاهدات بستگی ندارد.

کاربرد عمده ضریب تغییرات: هرگاه بخواهیم میزان پراکندگی مشاهدات حاصل از دو جامعه با واحدهای متفاوت را مقایسه کنیم از ضریب تغییرات استفاده می کنیم.

مثال: برای مقادیر زیر معیارهای پراکندگی را محاسبه نمایید.

۲۱ ۲۳ ۲۵ ۲۴ ۲۲ ۲۵ ۲۰ ۲۱ ۲۵ ۲۶ ۲۸ ۲۷ ۲۴

$$\overline{X} = 23.92,$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1} = 5.91,$$

$$S = \sqrt{5.91} = 2.43,$$

$$A = \frac{\sum_{i=1}^{n} \left| X_i - \overline{X} \right|}{n} = 1.94,$$

$$C.V. = \frac{S}{|\bar{X}|} = \frac{2.43}{23.92} = 0.101.$$

ویژگی معیارهای مرکزی و پراکندگی:

1: $\forall i \qquad X_i \rightarrow X_i \pm b$

$$\bar{X} \rightarrow \bar{X} \pm b$$

$$M_d \rightarrow M_d \pm b$$

$$M_o \rightarrow M_o \pm b$$

$$q_r \rightarrow q_r \pm b$$

$$R \rightarrow R$$

$$S^2 \rightarrow S^2$$

$$S \rightarrow S$$

$$A \rightarrow A$$

مقدار حاصل از ضریب تغییرات قابل تعیین نمی باشد.

نکته ۸: هیچکدام از معیارهای پراکندگی به جز CV نسبت به انتقال حساسیت ندارند، اما معیارهای مرکزی به مانند مشاهدات تغییر می کنند.

مثال: در مشاهدات مثال قبل ضریب تغییرات را به ازای کاهش ۵ واحد از همه مشاهدات مجدد محاسبه نمایید

بدلیل اینکه از همه مشاهدات ۵ واحد کم شده است، میانگین و انحراف معیار جدید به ترتیب ۱۸,۹۲ و ۲,٤۳ می شوند، بنابراین ضریب تغییرات مشاهدات جدید برابر خواهد شد با $\frac{2.43}{18.92} = 0.128$ که نشان می دهد، ضریب تغییرات افزایش یافته است.

2: $\forall i \qquad X_i \to CX_i \qquad C \neq 0$

$$\overline{X} \to C\overline{X}$$

$$M_d \to CM_d$$

$$M_o \to CM_o$$

$$q_r \to \begin{cases} Cq_r & C > 0 \\ Cq_{1-r} & C < 0 \end{cases}$$

$$S^2 \to C^2S^2$$

$$S \to |C|\, S$$

$$A \to |C|\, A$$

$$R \to |C|\, R$$

$$CV \to CV$$

مثال: میانگین حاصل از ۱۰ مشاهده برابر ١٤ می باشد. با نگاهی به محاسبات متوجه شدیم ۲ مقدار را به اشتباه وارد محاسبات نموده ایم . یکی از آنها ۱۸ که مقدار واقعی آن ۱۶ و دیگری ۱۳ که مقدار واقعی آن ۱۰ بوده است. میانگین درست مشاهدات را بدست آورید .

$$\sum X_i = 14 \times 10 = 140$$
$$\to 140 - 18 - 13 = 109$$
$$109 + 16 + 10 = 135 \to \sum X_i = 135 \to \overline{X} = \frac{135}{10} = 13.5$$

مثال: معدل دانشجویی برابر ۱۱٫۸ می باشد، اگر وی ترم قبل مشروط شده باشد، حداقل و حداکثر چقدر نمره نیاز دارد تا این ترم مشروط نشود؟

جواب: دانشجو در هر ترم حداقل ۱۲ واحد می بایست اخذ نماید تا بعنوان محصل محسوب شود و زمانیکه ترم سابق مشروط شده باشد، حداکثر ١٤ واحد می توانسته اخذ نماید. بنابراین

حداقل نمره مورد نیاز: ۱۲*۰٫۲=۲٫٤

حداکثر نمره مورد نیاز: ١٤*۰٫۲=۲٫۸

مثال: واریانس حاصل از ٤۰ مشاهده ۱۰۰ می باشد. آیا امکان دارد شخصی مدعی شود یکی از اعداد را ۲۰ واحد زیاد وارد کرده باشد؟ (تکلیف)

آمار توصیفی بر اساس طبقه‌بندی داده‌ها

هرگاه تعداد مشاهدات زیاد باشد یکی از ابزارهای کاهش حجم عملیات محاسباتی، دسته‌بندی یا طبقه‌بندی داده‌ها است. این عمل بر پایه‌ی این دانش است که چه تعداد طبقات (K) نیاز داریم و حدود هر طبقه به چه میزان است؟ در پاسخ به این سوال ۲ رویکرد مختلف وجود دارد:

۱-تعداد طبقات و حدود آنها به صورت علمی یا تجربی وجود دارد. همانند درجه ابتلا به قند خون، دسته‌های هوش، رده‌های سنی، فشار خون، قند خون

۲- در رویکرد دوم طول تمام طبقات را یکسان و برابر h قرار می دهیم که از فرمول $h=\frac{R}{K}$ بدست می آید که در آن K تعداد طبقات می باشد.

نکته: در بیشتر مواقع K بنا به تجربه مشخص می شود ولی در علوم انسانی، اجتماعی و روانشناسی غالب $K\cong\sqrt{n}$ می باشد و در بیشتر علوم تجربه ثابت کرده که $K\cong 1+3.321\times\log_{10}(n)$ می باشد.

نکته: هرگاه مقدار K اعشار گردید، لازم است به اولین عدد صحیح بزرگتر از آن مقدار گرد شود.

مثال: میخواهیم ۸۶ مشاهده را دسته بندی کنیم و میدانیم دامنه این مشاهدات ۵۶ می باشد. طول هر طبقه، h، را مشخص نماید.

$$K\cong 1+3.321\times\log_{10}(n)=1+3.321\times\log_{10}(86)=7.424.$$

تعداد طبقات پیشنهادی بیش از ۷ می باشد، بنابراین $K=8$ در نظر گرفته می شود. طول هر طبقه می بایست برابر $h=\frac{R}{K}=\frac{56}{8}=7$ باشد.

در حالت کلی یک جدول طبقه‌بندی شده دارای فرم کلی زیر است.

X_i	S_i	f_i	F_i	حدود طبقات	دسته‌های موجود
X_1	F_1	$f_1= F_1/n$	F_1	$L_1–U_1$	عنوان دسته اول
X_2	F_1+F_2	$f_2= F_2/n$	F_2	$L_2–U_2$	عنوان دسته دوم
⋮	⋮	⋮	⋮	⋮	⋮
X_k	$F_1+…+F_k=n$	$f_k= F_k/n$	F_k	$L_k–U_k$	عنوان دسته kام

جاییکه در آن L_i و U_i به ترتیب حد پایین و بالای هر طبقه،

F_i فراوانی مطلق طبقه iام،

f_i فراوانی نسبی iامین طبقه،

S_i فراوانی تجمعی طبقه iام

و X_i نماینده آن طبقه می باشند، برای محاسبه این مقدار جمع حدود طبقات را بر دو تقسیم کنید. به بیان دیگر $X_i = \frac{L_i+U_i}{2}$.

مثال برای رویکرد اول: فرض کنید ۲۳۰ نفر در یک آزمون سنجش هوش هیجانی شرکت کرده‌اند و نتیجه به شرح جدول زیر به دست آمده است.

X_i	S_i	f_i	F_i	حدود طبقات	نوع هوش هیجانی
۱۸/۵	۲۰	$\frac{20}{230} = 0.087$	۲۰	۲۵-۱۲	بی بهره
۳۰	۵۷	$\frac{37}{230} = 0.161$	۳۷	۳۵-۲۵	کم بهره

بهره متوسط	۵۰-۳۵	۵۳	$\frac{53}{230} = 0.230$	۱۱۰	٤۲/۵
بهره بالا	۵۶-۵۰	۸۵	$\frac{85}{230} = 0.370$	۱۹۵	۵۳
بهره بسیار بالا	۶۰-۵۶	۳۵	$\frac{35}{230} = 0.152$	۲۳۰	۵۸

نکته: در طبقه‌بندی به روش اول حدود دسته‌ها لزوما یکسان نیست، یعنی طول دسته نابرابرند.

درادامه نحوه محاسبه معیارهای تمرکز، پراکندگی و نمودار را با توجه به مقادیر طبقه‌بندی شده، ارائه می نماییم. در محاسبات بر اساس جدول طبقه‌بندی، مهمترین ستون جدول را ستون میانه طبقات تشکیل می دهد.

مثال (با رویکرد دوم): در یک تحقیق ۳۱۲ نفر شرکت کننده مورد ارزیابی قرار گرفتند. از این افراد خواسته شد نمره خود را درباره محتوای پخش شده از ۱ تا ۲۰ ارائه دهند. با توج به اینکه کمترین و بیشترین نمره داده شده توسط این افراد به ترتیب ۱,۵ و ۱۹,۵ می باشد، جدول زیر نتیجه خواهد شد.

تعداد طبقات از فرمول $K \cong 1 + 3.321 \times \log_{10}(312) = 9.2825$ این یعنی $K = 10$ می باشد. دامنه تغییرات برابر است با $R = 19.5 - 1.5 = 18$، این بدان معنا است که طول هر طبقه برابر است با $h = \frac{18}{10} = 1.8$. جدول بصورت زیر نوشته می شود

شماره طبقه	حدود طبقات	F_i	f_i	S_i	X_i
۱	1.5-3.3	۱۲	۰,۰۳۸	۱۲	۲,٤
۲	3.3-5.1	۳۵	۰,۱۱۲	٤۷	٤,۲
۳	5.1-6.9	۲۵	۰,۰۸۰	۷۲	۶,۰
٤	6.9-8.7	٤۰	۰,۱۲۸	۱۱۲	۷,۸

۵	8.7-10.5	۲۰	۰,۰۶٤	۱۳۲	۹,۶
۶	10.5-12.3	۶۰	۰,۱۹۲	۱۹۲	۱۱,٤
۷	12.3-14.1	٤۰	۰,۱۲۸	۲۳۲	۱۳,۲
۸	14.1-15.9	٤۰	۰,۱۲۸	۲۷۲	۱۵
۹	15.9-17.7	۳۰	۰,۰۹۶	۳۰۲	۱۶,۸
۱۰	17.7-19.5	۱۰	۰,۰۳۲	۳۱۲	۱۸,۶

نکات و نتایج:

- ❖ مشاهده ۵,۱ (حد بالای طبقه دوم) در صورت وجود متعلق به طبقه سوم می باشد. این بدان معنا است که در طبقه بندی کران پایین هر طبقه بسته و کران بالا باز در نظر گرفته می شود. این مطلب برای طبقه آخر استثناءِ می باشد یعنی هر دو کران بسته فرض می شوند.
- ❖ جمع ستون فراوانی مطلق همیشه برابر است با n (تعداد نمونه)
- ❖ جمع ستون فراوانی نسبی برابرا ست با ۱
- ❖ آخرین ردیف فراوانی تجمعی همیشه می بایست برابر با n شود
- ❖ در دسته‌بندی مشاهدات با طول طبقات یکسان، فاصله بین نماینده طبقات متوالی نیز برابر با طول هر دسته می باشد.
- ❖ هرگاه طبقه‌بندی مشاهدات انجام می شود، عملا با داده‌های خام کاری نداریم و تمامی محاسبات بر اساس جدول طبقه‌بندی مشاهدات انجام می شود.

محاسبه معیارهای تمرکز از روی داده‌های طبقه‌بندی شده

میانگین: $\overline{X} = \frac{\sum_{i=1}^{K} F_i X_i}{n} = \sum_{i=1}^{K} f_i X_i \Leftrightarrow \sum_{i=1}^{K} F_i X_i = n\overline{X}$

میانگین مشاهدات جدول هوش هیجانی برابر

$$\overline{X} = \frac{20 \times 18.5 + 37 \times 30 + 53 \times 42.5 + 85 \times 53 + 35 \times 58}{230} = 44.641.$$

بطور متوسط افراد آن جامعه دارای بهره هوشی ٤٤,٦٤١ هستند.

میانگین برای مشاهدات نمره فیلم برابر است با

$$\overline{X} = \frac{12 \times 2.4 + 35 \times 4.2 + 25 \times 6 + \cdots + 10 \times 18.6}{312} = 10.68.$$

این بدان معنا است که متوسط نمره پاسخگویان به سوال درباره کیفیت فیلم ۱۰,۶۸ (یعنی متوسط) بوده است.

میانه: بمنظور محاسبه میانه می بایست،

گام اول: طبقه شامل میانه را مشخص می نماییم. طبقه شامل میانه اولین طبقه‌ای است که فراوانی تجمعی آن بیشتر یا مساوی $\frac{n}{2}$ باشد.

گام دوم: محاسبه مقدار آن از رابطه زیر

$$M_d = L_d + \frac{\frac{n}{2} - S_{d-1}}{F_d} h_d.$$

جاییکه L_d حد پایین طبقه شامل میانه، S_{d-1} فراوانی تجمعی طبقه ماقبل میانه و F_d و h_d به ترتیب فراوانی مطلق و طول طبقه شامل میانه می شود.

بنابر این برای مثال هوش هیجانی، طبقه با بهره هوشی بالا طبقه شامل میانه است زیرا $\frac{n}{2} = 115$ می باشد و طبقه ٤ام اولین طبقه‌ای است که فراوانی تجمعی آن بیشتر از ۱۱۵ می باشد. خواهیم داشت

$$M_d = 50 + \frac{\frac{230}{2} - 110}{85} 6 = 50.353.$$

نیمی از افراد آن جامعه دارای هوش هیجانی کمتر یا مساوی ۵۰,۳۵۳ را دارا هستند.

همچنین برای مثال نمره به فیلم، طبقه ۶ طبقه شامل میانه می باشد و مقدار میانه برابر است با

$$M_d = 10.5 + \frac{\frac{312}{2} - 132}{60} 1.8 = 11.22$$

مد یا **نما**: برای محاسبه مقدار مد می بایست گام‌های ادامه را اجرا نمود

گام اول: تعیین طبقه یا طبقات شامل مد. طبقه یا طبقاتی هستند که بیشترین فراوانی مطلق را دارند. به بیان بهتر نسبت به طبقات قبل و بعد خود فراوانی مطلق قابل توجه بیشتری را دارا است.

گام دوم: محاسبه نما براساس فرمول

$$M_o = L_o + \frac{d_1}{d_1 + d_2} h_o.$$

جاییکه در آن L_o و h_o به ترتیب حد پایین و طول طبقه شامل مد هستند و همچنین d_1 و d_2 به ترتیب تفاضل فراوانی مطلق طبقه شامل مد از طبقه قبل و بعد تعریف می شوند.

برای مثال هوش هیجانی طبقه با بهره هوشی بالا طبقه شامل مد نیز بشمار می رود، طبقه با بهره هوشی بالا یا همان طبقه چهارم طبقه شامل مد می باشد همچنین

$d_1 = 85 - 53 = 32$ و $d_2 = 85 - 35 = 50$. بنابر این مقادیر خواهیم داشت

$$M_o = 50 + \frac{32}{32 + 50} \times 6 = 52.341.$$

این بدان معنا است که بیشترین فراوانی مشاهدات هوش هیجانی مربوط به مقدارهای حول و حوش ۵۲,۳٤۱ می باشد.

در مثال نمره دهی به فیلم نمایش داده شده، طبقات ۲، ٤ و ۶ شامل مد می باشند، زیرا فراوانی مطلق آن طبقات از طبقات مجاور قابل توجه بیشتر است.

مد طبقه دوم:

$$M_{o1} = 3.3 + \frac{23}{23 + 10} 1.8 = 4.554,$$

مد طبقه چهارم:

$$M_{o2} = 6.9 + \frac{15}{15 + 20} 1.8 = 7.67,$$

همچنین مد طبقه ششم بصورت زیر بدست می آید. مقدار $d_1 = 60 - 20 = 40$ و $d_2 = 60 - 20 = 40$ می باشند. همچنین، خواهیم داشت

$$M_{o3} = 10.5 + \frac{40}{40+20} 1.8 = 11.7.$$

نکته: مد می تواند وجود نداشته باشد و یا حتی منحصر به فرد نباشد.

نکته: اگر طبقه شامل مد طبقه اول یا انتهایی باشد فراوانی مطلق طبقه قبل یا بعد از آنها برابر صفر می باشند.

چندک در داده‌های طبقه‌بندی شده (q_r): از آنجایی که می توان با استفاده از محاسبه‌ی صدک به چارک و دهک رسید، نحوه محاسبه صدک (P_r) را در ادامه بیان می کنیم.

گام اول) تشخیص طبقه‌ی شامل صدک rام. اولین طبقه‌ای که فراوانی تجمعی آن بیشتر یا مساوی $\frac{nr}{100}$ باشد.

گام دوم) آنگاه محاسبه صدک بر اساس فرمول

$$P_r = L_r + \frac{\frac{nr}{100} - S_{r-1}}{F_r} h_r.$$

جاییکه در آن L_r، F_r و h_r به ترتیب حد پایین، فراوانی مطلق و طول طبقه شامل صدک rام و S_{r-1} فراوانی تجمی طبقه ماقبل تعریف می باشند.

مثال: برای داده‌های طبقه‌بندی شده بهره هوشی صدک ۳۶ام، یعنی P_{36}، را بدست آورید.

گام اول: یافتن طبقه شامل صدک ۳۶ام: می دانیم $r = 36$ و محاسبه می کنیم $230 * \frac{36}{100} = 82.8$ اولین طبقه‌ای که فراوانی تجمعی آن بیشتر یا مساوی ۸۲٫۸ باشد، طبقه سوم است

گام دوم: محاسبه فرمول

$$P_{36} = 35 + \frac{82.8 - 57}{53} 15 = 42.302.$$

بدین معنا است که حدود ۳۶درصد جامعه دارای نمره بهره هوش هیجانی کمتر یا مساوی ۴۲٫۳۰۲ می باشند.

مثال: برای داده های بدست آمده از نمره محتوای فیلم، دهک ۹ام و چارک اول را بدست آورید.

گام اول: یافتن طبقه شامل دهک ۹ام: می دانیم دهک ۹ام همان صدک ۹۰ام می باشد، یعنی $r = 90$، بنابراین $90 * \frac{312}{100} = 280.8$. بنابر این طبقه شامل دهک ۹ام طبقه ۹می باشد.

گام دوم: محاسبه دهک ۹ام

$$D_9 = P_{90} = 15.9 + \frac{280.8 - 272}{30} 1.8 = 16.428.$$

محاسبه چارک اول:

گام اول: یافتن طبقه شامل چارک اول: یعنی $r = 25$، بنابراین $78 = \frac{312}{100} * 25$.طبقه شامل چارک اول، طبقه ٤ می باشد.

گام دوم: محاسبه مقدار چارک اول

$$Q_1 = P_{25} = 6.9 + \frac{78-72}{40}1.8 = 7.17.$$

تفسیر: نمره ۹۰درصد مردم به پخش فیلم نمره ای کمتر از ۱۶,٤۲ و ۲۵درصد نیز نمره ای زیر۷,۱۷ به این فیلم داده‌اند. ۶۵درصد افراد نمره‌ای بین ۷,۱۷ تا ۱۶,٤۲ به این فیلم داده اند.

معیارهای پراکندگی

در این قسمت در نظر داریم فرمول‌های محاسبه معیارهای پراکندگی براساس جدول طبقه‌بندی مشاهدات بهره ببریم.

محاسبه واریانس بر اساس مشاهدات جدول فراوانی:

$$S^2 = \frac{\sum_{i=1}^{K} F_i (X_i - \bar{X})^2}{n-1}.$$

مثال: مقدار واریانس جدول بهره هوشی را بدست آورید.

$$S^2 = \frac{20(18.5-44.641)^2 + 37(30-44.641)^2 + 53(42.5-44.641)^2 + 85(53-44.641)^2 + 3}{229} = 148.588.$$

انحراف معیار مشاهدات بهره هوشی برابر است با $S = \sqrt{S^2} = \sqrt{148.588} = 12.189$

مثال: میزان پراکندگی نمرات داده شده به فیلم را بدست آورید.

$$S^2 = \frac{\sum_{i=1}^{k} F_i (X_i - \bar{X})^2}{n-1}$$
$$= \frac{12(2.4 - 10.68)^2 + 35(4.2 - 10.68)^2 + \cdots + 10(18.6 - 10.68)^2}{312 - 1}$$
$$= 19.220.$$

تفسیر: پراکندگی مشاهدات در نمرات داده شده به فیلم قابل توجه زیاد می باشد.

بمنظور ساده‌سازی فرمول واریانس، بصورت زیر اقدام می نماییم

$$\sum_{i=1}^{k} F_i (X_i - \bar{X})^2 = \sum_{i=1}^{k} F_i \left({X_i}^2 + \bar{X}^2 - 2X_i\bar{X}\right)$$
$$= \sum_{i=1}^{k} \left(F_i{X_i}^2 + F_i\bar{X}^2 - 2F_iX_i\bar{X}\right)$$
$$= \sum_{i=1}^{k} F_i{X_i}^2 + \sum_{i=1}^{k} F_i\bar{X}^2 - 2\sum_{i=1}^{k} F_iX_i\bar{X}$$
$$= \sum_{i=1}^{k} F_i{X_i}^2 + \bar{X}^2\sum_{i=1}^{k} F_i - 2\bar{X}\sum_{i=1}^{k} F_iX_i$$
$$= \sum_{i=1}^{k} F_i{X_i}^2 + n\bar{X}^2 - 2\bar{X} \times n\bar{X}$$
$$= \sum_{i=1}^{k} F_i{X_i}^2 - n\bar{X}^2,$$
$$\Rightarrow S^2 = \frac{\sum_{i=1}^{k} F_i{X_i}^2 - n\bar{X}^2}{n-1}.$$

برای مثال بهره هوشی خواهیم داشت

$$\sum_{i=1}^{k} F_i{X_i}^2 = 20 * 18.5^2 + 37 * 30^2 + 53 * 42.5^2 + 85 * 53^2 + 35 * 58^2$$
$$= 492381.25.$$

$$S^2 = \frac{492381.25 - 230 * (44.641)^2}{229} = 148.615.$$

برای مثال نمره داده شده به فیلم داریم

$$\sum_{i=1}^{k} F_i {X_i}^2 = 12 * 2.4^2 + 35 * 4.2^2 + \cdots + 10 * 18.6^2 = 41557.32.$$

$$S^2 = \frac{41557.32 - 312 * (10.68)^2}{311} = 19.195.$$

نکته: این اختلاف بین دو مقدار بدست آمده از دو فرمول مربوط به واریانس، از گردکردن مقادیر در محاسبات مختلف ناشی می شود.

انحراف قدرمطلقی:

$$A = \sum_{i=1}^{K} \frac{F_i |X_i - \bar{X}|}{n} = \sum_{i=1}^{K} f_i |X_i - \bar{X}|.$$

برای جدول هوش هیجانی داریم

$$A = \frac{20}{230}|18.5 - 44.641| + \frac{37}{230}|30 - 44.641| + \frac{53}{230}|42.5 - 44.641| + \frac{85}{230}|53 - 44.641| + \frac{35}{230}|58 - 44.641| = 10.243.$$

تکلیف: برای مشاهدات طبقه‌بندی شده نمره به فیلم، انحراف قدرمطلقی را محاسبه نمایید.

تکلیف*:** انواع نمودارها را بصورت یک گزارش تحویل نمایید. (نتایج تحلیل در امتحان‌ها خواهد آمد).

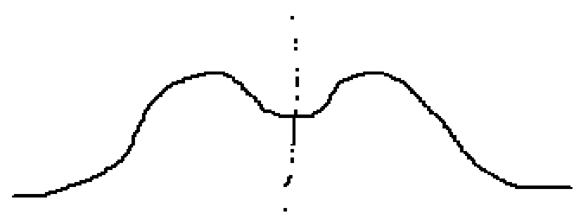
انواع نمودار

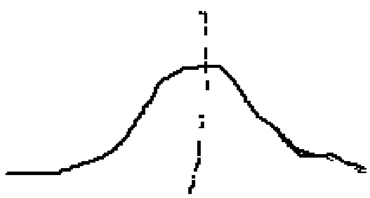
نمایش بصری مشاهدات را با کمک نمودارهای مختلف می توان انجام داد. در نمودار نحوه توزیع مشاهدات در جامعه واضح بیان خواهد شد. همچنین از جمله خصوصیات داده‌های طبقه‌بندی شده آن است که کمک شایانی در رسم انواع نمودار می نماید.

ویژگی‌های توزیعی مشاهدات

1- **چولگی[1] یا عدم متقارن**: هرگاه مشاهدات رفتاری غیر متقارن نمایش دهند، گوییم توزیع داده‌ها دارای چولگی می باشد یا به بیان بهتر داده‌ها را چوله گوییم.

توزیع متقارن: هرگاه توزیع داده‌ها نسبت به یک نقطه رفتاری یکسان و مشابه داشته باشند، آن مشاهدات را متقارن گویند. بعنوان مثال





چوله به راست: هرگاه تقارن بدلیل وجود مشاهداتی در سمت راست نمودار بر هم خورده باشد، گوییم مشاهدات دارای چولگی از نوع راست می باشند.

$$M_o < M_d < \bar{X}$$

چوله به چپ: هرگاه تقارن بدلیل وجود مشاهداتی در سمت چپ نمودار بر هم خورده باشد، گوییم مشاهدات دارای چولگی از نوع چپ می باشند.

---
[1] Skewness

$$\bar{X} < M_d < M_o$$

نکته (پیرسون): انتظار داریم رابطه زیر بین 3شاخص تمرکز میانگین، میانه و مد وجود داشته باشد

$$3(\bar{X} - M_d) \cong (\bar{X} - M_o)$$

نکته: چولگی مبحثی است مربوط به توزیع مشاهدات در جوامع تک مد می باشد.

نکته: عدم تقارن انواع مختلف دارد که چوله به راست یا چپ دو نوع از آن را شامل می شود.

شاخص‌های چولگی: هر ابزار آماری که کمک می نماید تا به کشف عدم تقارن در مشاهدات بپردازیم را شاخص چولگی گویند.

انواع شاخص چولگی یا عدم تقارن:

- ✓ شاخص چولگی گشتاوری: برای مشاهدات با یک مد این شاخص توسط دانشمند مطرح علم آمار آقای پیرسون پیشنهاد شده است. بفرم زیر محاسبه می شود. به این شاخص ضریب چولگی پیرسون نیز گفته می شود.

$$\rho_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(S^2)^{\frac{3}{2}}} = \frac{\sum_{i=1}^{n} X_i^3 - 3\bar{X}\sum_{i=1}^{n} X_i^2 + 2n\bar{X}^3}{nS^3}.$$

زیرا برای داده‌های خام داریم

$$\sum_{i=1}^{n}(X_i-\bar{X})^3=\sum_{i=1}^{n}(X_i^3-3X_i^2\bar{X}+3X_i\bar{X}^2-\bar{X}^3)$$
$$=\sum_{i=1}^{n}X_i^3-3\bar{X}\sum_{i=1}^{n}X_i^2+3\bar{X}^2\sum_{i=1}^{n}X_i-n\bar{X}^3$$
$$=\sum_{i=1}^{n}X_i^3-3\bar{X}\sum_{i=1}^{n}X_i^2+3n\bar{X}^3-n\bar{X}^3=\sum_{i=1}^{n}X_i^3-3\bar{X}\sum_{i=1}^{n}X_i^2+2n\bar{X}^3.$$

برای داده‌های طبقه بندی شده، این فرمول بفرم زیر می باشد

$$\rho_S=\frac{\sum_{i=1}^{K}F_i(X_i-\bar{X})^3}{n(S^2)^{\frac{3}{2}}}=\frac{\sum_{i=1}^{k}F_iX_i^3-3\bar{X}\sum_{i=1}^{k}F_iX_i^2+2n\bar{X}^3}{nS^3}.$$

مثال: برای داده‌های هوش هیجانی، ضریب چولگی پیرسون را محاسبه نمایید.

برای محاسبه واریانس مقادیر $\sum_{i=1}^{k}F_iX_i^2$ و $\bar{X}$ به ترتیب 492381.35 و 44.641 بدست آمده‌اند، همچنین S^2 برابر 148.615 می باشد.

$$\sum_{i=1}^{5}F_iX_i^3=20\times 18.5^3+37\times 30^3+\cdots+35\times 58^3=24677675.625.$$

$$\rho_S=\frac{\sum_{i=1}^{k}F_iX_i^3-3\bar{X}\sum_{i=1}^{k}F_iX_i^2+2n\bar{X}^3}{nS^3}$$
$$=\frac{24677675.625-3\times 44.641\times 49238.35\ +2\times 230\times 44.641^3}{230\times(148.615)^{\frac{3}{2}}}$$
$$=-0.81891$$

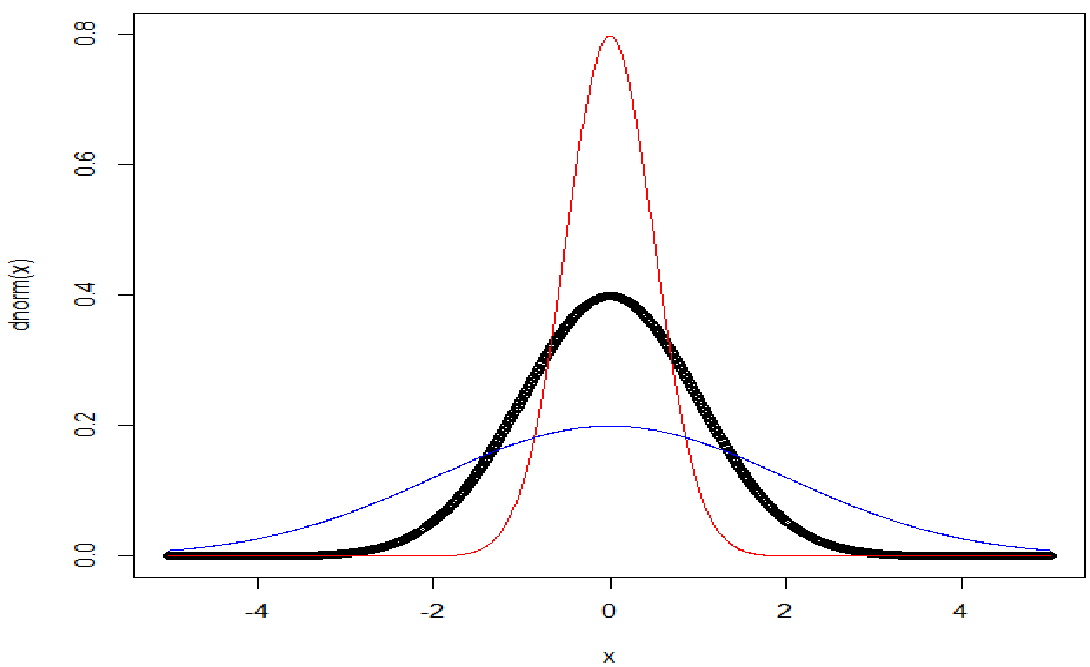
تفسیر: این ضریب همبستگی نشاندهنده چوله به چپ بودن مشاهدات یا تمایل به گرفتن یا داشتن مقادیر کوچک در دامنه مشاهدات می باشد. بنابراین برای بهبود وضعیت هوش هیجانی جامعه، یعنی متقارن یا طبیعی شدن هوش هیجانی در بین اقشار مردم، می بایست نمرات پایین هوش هیجانی بهبود یابد. بدین منظور می توان از روشهای فرهنگ سازی یا آموزش همگانی بهره برد.

تکلیف (دانشجو): انواع ضرایب چولگی را تحقیق همراه با یک مثال ذکر نمایید.

نحوه تفسیر اعداد حاصل از ضرایب چولگی (ρ)

- برای اعداد منفی ($\rho \ll -1$) گوییم مشاهدات (شدیدا) چوله به چپ هستند
- برای اعداد منفی ($\rho \cong -1$) گوییم مشاهدات چوله به چپ هستند
- برای اعداد تقریبا صفر ($\rho \cong 0$ یا $-0.5 \le \rho \le 0.5$) گوییم مشاهدات متقارن هستند
- برای اعداد مثبت ($\rho \cong 1$) گوییم مشاهدات چوله به راست هستند
- برای اعداد مثبت ($\rho \gg 1$) گوییم مشاهدات (شدیدا) چوله به راست هستند

2- **کشیدگی**[۲]: بمنظور بررسی نحوه توزیع مشاهدات نسبت به یک جامعه نرمال، ابزاری طراحی شده که کمک شایانی در درک نحوه توزیع می نماید، این ابزار تحت عنوان کشیدگی مطرح می شود.



[2] Kurtosis

ضریب کشیدگی گشتاوری (فیشر): این ابزار بمنظور درک تفاوت یا وجود کشیدگی در داده‌ها نسبت به توزیع نرمال طراحی شده است. به این ضریب کشیدگی، ضریب کشیدگی گشتاوری نیز گویند.

$$\rho_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{(S^2)^2} - 3 = \frac{\sum_{i=1}^{n} X_i^4 - 4\bar{X}\sum_{i=1}^{n} X_i^3 + 6\bar{X}^2\sum_{i=1}^{n} X_i^2 - 3n\bar{X}^4}{nS^4} - 3.$$

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^4 &= \sum_{i=1}^{n}(X_i^4 - 4X_i^3\bar{X} + 6X_i^2\bar{X}^2 - 4X_i\bar{X}^3 + \bar{X}^4) \\
&= \sum_{i=1}^{n} X_i^4 - 4\sum_{i=1}^{N} X_i^3\bar{X} + 6\sum_{i=1}^{n} X_i^2\bar{X}^2 - 4\sum_{i=1}^{n} X_i\bar{X}^3 + \sum_{i=1}^{n}\bar{X}^4 \\
&= \sum_{i=1}^{n} X_i^4 - 4\bar{X}\sum_{i=1}^{N} X_i^3 + 6\bar{X}^2\sum_{i=1}^{n} X_i^2 - 4\bar{X}^3\sum_{i=1}^{n} X_i + n\bar{X}^4 \\
&= \sum_{i=1}^{n} X_i^4 - 4\bar{X}\sum_{i=1}^{n} X_i^3 + 6\bar{X}^2\sum_{i=1}^{n} X_i^2 - 3n\bar{X}^4.
\end{aligned}$$

تکلیف: انواع ضریب کشیدگی را با مرجع ذکر نمایید.

نحوه تفسیر اعداد حاصل از ضرایب کشیدگی (ρ_k)

- برای اعداد منفی ($\rho_k \ll -1$) گوییم مشاهدات کشیدگی شدیدا کمتری نسبت به توزیع نرمال دارند یا به بیان دیگر دارای پخی شدید هستند.
- برای اعداد منفی ($\rho_k \cong -1$) گوییم مشاهدات کشیدگی تقریبا کمتری نسبت به توزیع نرمال دارند. به بیان دیگر دارای پخی هستند
- برای اعداد تقریبا صفر ($\rho_k \cong 0$ یا $0.5 \le \rho \le -0.5$) گوییم توزیع مشاهدات مشابه توزیع نرمال است.
- برای اعداد مثبت ($\rho_k \cong 1$) گوییم مشاهدات کشیدگی تقریبا بیشتری نسبت به توزیع نرمال دارند.
- برای اعداد مثبت ($\rho_k \gg 1$) گوییم مشاهدات کشیدگی شدیدا بیشتری نسبت به توزیع نرمال دارند.

انواع نمودار

در ادامه به معرفی تعدادی از پرکاربردترین نمودارهای علم آمار را مطرح می نماییم. در رسم هر نمودار باید توجه نمود که

- برای چه داده‌هایی مفید است
- بر حسب چه مشخصه‌ای (فراوانی نسبی، مطلق، تجمعی یا ...) از مشاهدات رسم

می شود.

1- هیستوگرام[3] یا نمودار مستطیلی

این نمودار برای متغیرهای پیوسته طبقه‌بندی شده مفید می باشد. همچنین برای رسم هیستوگرام از فراوانی نسبی یا فراوانی مطلق استفاده می شود. عرض هر ستون از هیستوگرام به اندازه طول طبقه می باشد و ارتفاع آن متناسب است با فراوانی نسبی یا مطلق.

نکته: زمانی که طول طبقات یا دسته‌های ایجاد شده برای داده‌ها یکسان نیست، می بایست یک دسته را بعنوان طبقه مرجع در نظر بگیریم و فراوانی مدنظر مابقی طبقات بر حسب طبقه مرجع از فرمول

$$F_i^* = \frac{h_M}{h_i} F_i, \qquad\qquad f_i^* = \frac{h_M}{h_i} f_i$$

محاسبه شود. جاییکه h_M طول طبقه مرجع می باشد. ستون مربوط به طبقه مرجع حتما می بایست با رنگی متفاوت مشخص گردد.

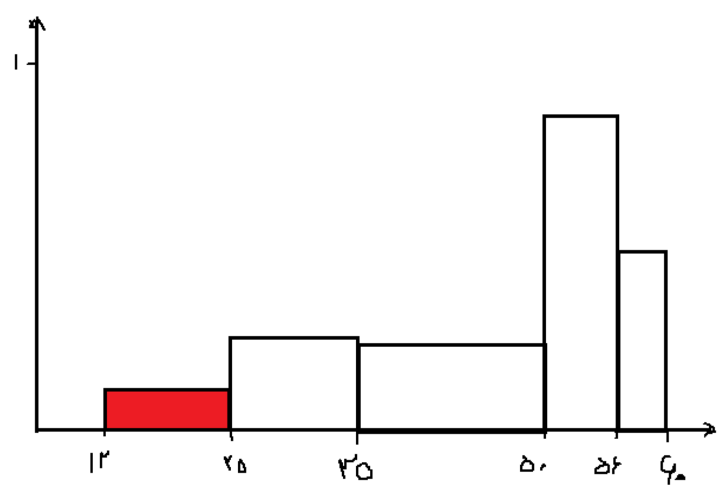
مثال: برای داده‌های هوش هیجانی نمودار هیستوگرام برحسب فراوانی نسبی را رسم نمایید.

حدود طبقات	f_i	f_i^*
12-25	$\frac{20}{230} = 0.087$	0.087
25-35	$\frac{37}{230} = 0.161$	$\frac{13}{10}0.161 = 0.2093$

[3] Histogram

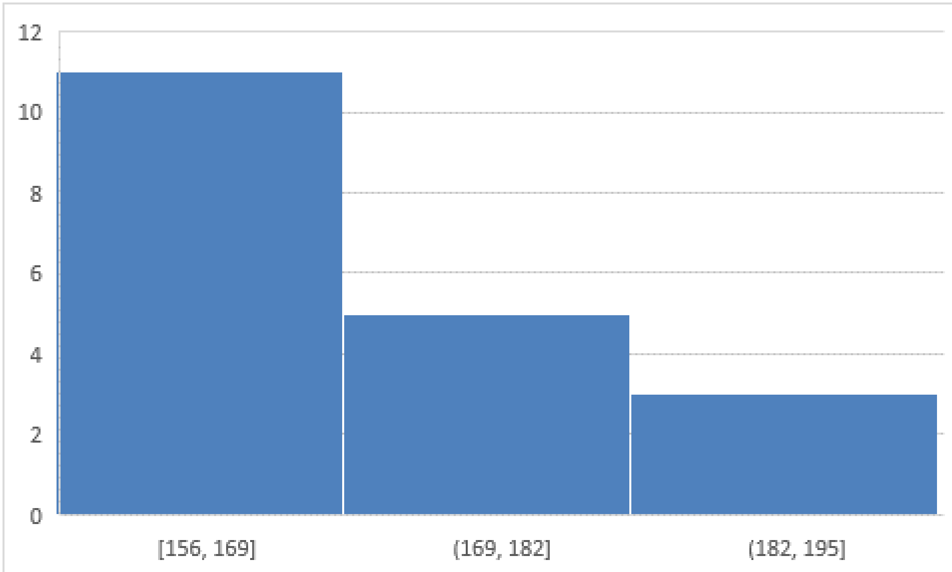
35–50	$\frac{53}{230} = 0.230$	$\frac{13}{15}0.23 = 0.19933$
50–56	$\frac{85}{230} = 0.370$	$\frac{13}{6}0.370 = 0.8016$
56–60	$\frac{35}{230} = 0.152$	$\frac{13}{4}0.152 = 0.494$

در این مثال، طبقه مرجع را طبقه اول فرض نمودیم. بنابراین هستوگرام بفرم زیر ایجاد می شود

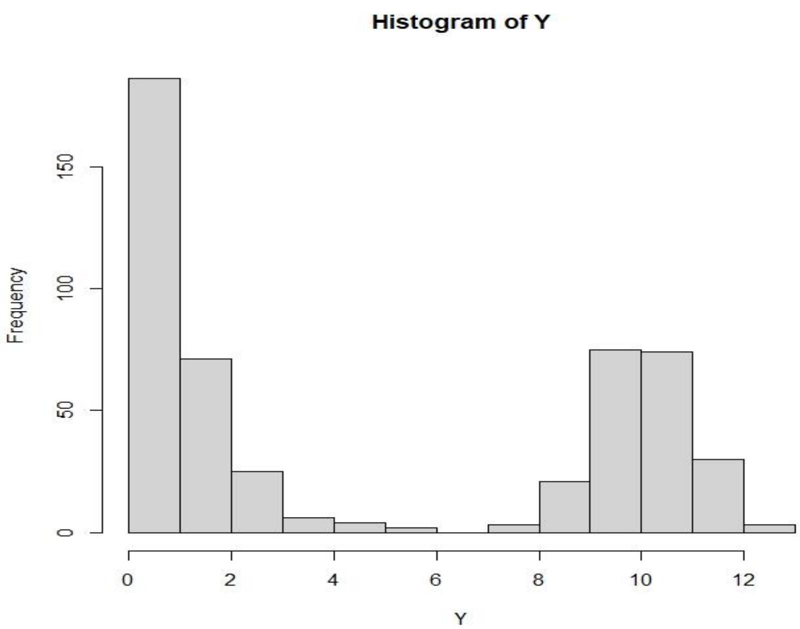


نکته: در نرم افزارهای متداول، طول دسته‌ها بصورت پیش‌فرض مساوی در نظر گرفته شده است تا مسائل فوق پیشامد ننماید.

مثال: هیستوگرام قد دانشجویان بر حسب فراوانی نسبی بفرم زیر می باشد.



مثال: 500 مشاهده از نرم افزار آماری R تولید و هیستوگرام آنها را در 10 دسته بصورت زیر نمایش می دهیم.

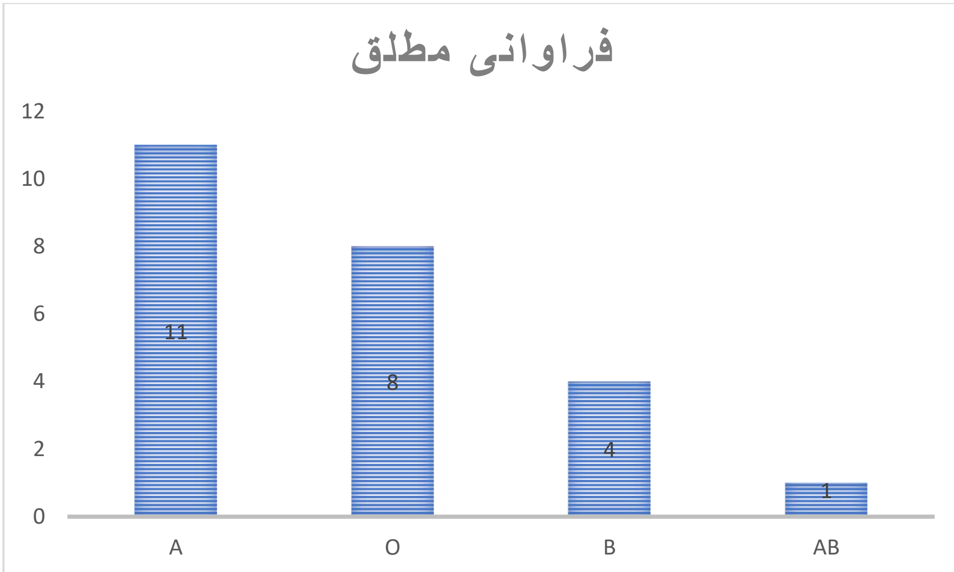


2- **نمودار میله‌ای**[4]: بطور استاندارد نمودار میله‌ای برای رسم مشاهدات کیفی یا متغیرهای گسسته مورد استفاده قرار می گیرد. این نمودار بر حسب فراوانی مطلق یا نسبی قابل رسم است.

---

[4] Bar chart

مثال: نمودار گروه خونی دانشجویان بصورت زیر می باشد

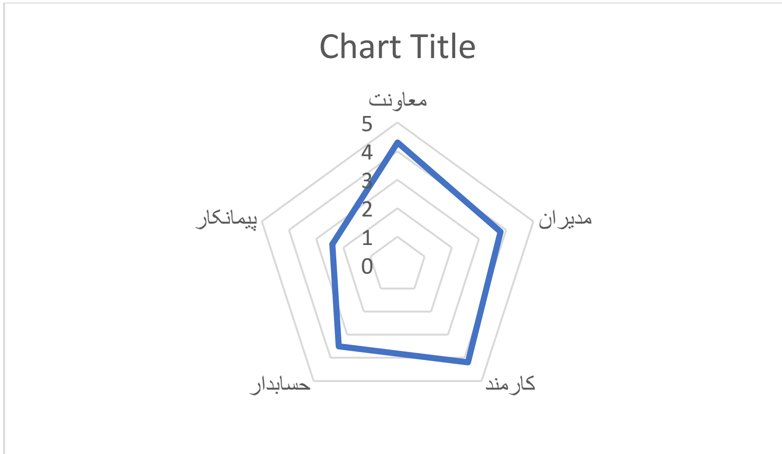


نمودار عنکبوتی[5]: این نمودار زمانی مناسب است که بخواهیم متوسط رفتار یک متغیر پیوسته را در سطوح یک متغیر گسسته مشاهده نماییم.

مثال: در یک پرسشنامه موضوع مورد علاقه میزان رضایت شغلی افراد در سمت‌های مختلف یک سازمان می باشد. نتایج در جدول زیر خلاصه شده است.

سمت سازمانی	معاونت	مدیران	کارمند	حسابدار	پیمانکار
متوسط رضایت شغلی	4.3	3.8	4.2	3.5	2.4

[5] Radar chart



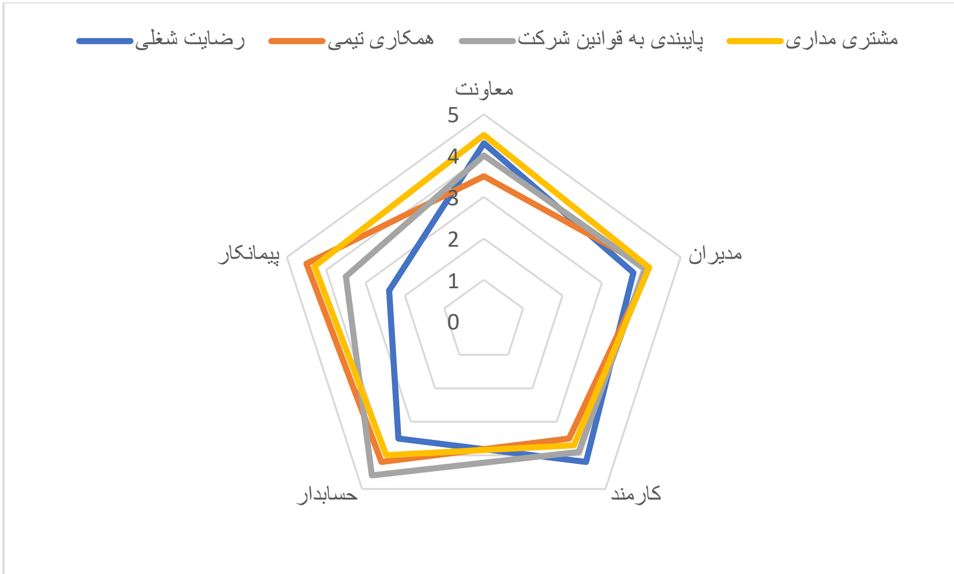
رفتار هر سمت در شاخص‌های رضایت شغلی بفرم زیر است

مثال: فرض کنید می خواهیم رفتار پست‌های سازمانی مختلف در شاخص‌های گوناگون را مشاهده نماییم.

سمت سازمانی	معاونت	مدیران	کارمند	حسابدار	پیمانکار
رضایت شغلی	4.3	3.8	4.2	3.5	2.4
همکاری تیمی	3.5	4.2	3.5	4.2	4.5
پایبندی به قوانین شرکت	4	4.1	3.9	4.6	3.5
مشتری مداری	4.5	4.2	3.7	4	4.3

نکته: همه سطوح مدنظر در نمودار عنکبوتی می بایست نمره‌ای متناسب و در یک مجموعه از قبل مشخص شده، قرار گیرند. بعنوان مثال فرض کنید k سطح داشته باشیم که می خواهیم همه سطوح مقادیرشان در بازه (a,b) باشد. با فرض آنکه یکی از سطوح مقدارش می تواد در بازه (c,d) باشد می بایست اعداد بدست آمده (I) در آن سطح را در فرمول زیر قرار داد.

$$I_{new} = a + (b-a)\frac{I-a}{d-c}.$$



3- نمودار چندبر فراوانی:

این نمودار برای نشان دادن وضعیت توزیع مقادیر مختلف یک متغیر پیوسته بسیار پرکاربرد می باشد. این نمودار بر حسب فراوانی نسبی رسم می شود. بمنظور رسم این نمودار می بایست مراحل زیر را طی نمود.

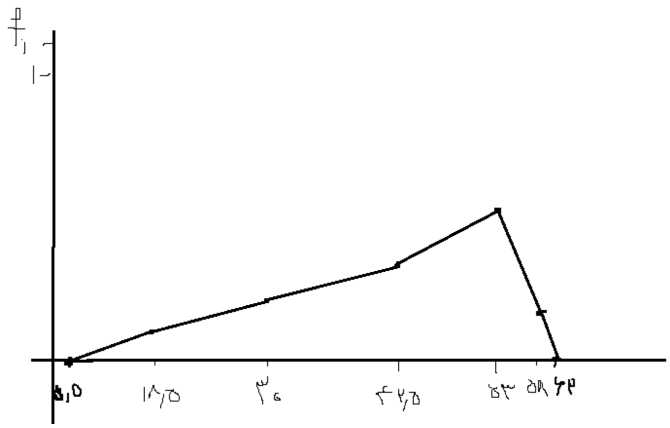
گام اول: طبقه‌بندی مشاهدات

گام دوم: رسم نمودار دکارتی با محور افقی شامل نماینده طبقات و محور عمودی فراوانی نسبی. به ازای هر طبقه (نماینده طبقه) و مقدار فراوانی نسبی آن یک نقطه رسم می نماییم.

گام سوم: اتصال نقاط بدست آمده در مرحله دوم و همچنین اتصال نقاط ابتدا و انتها به محور افقی.

نکته: بمنظور رسم نقطه ابتدا یا انتها به محور افقی می بایست به اندازه طول طبقه اول یا طبقه آخر از نماینده این طبقات فاصله گرفت و نقاط ابتدا و انتها را به این نقاط جدید متصل نمود.

مثال: نمودار چندبر فراوانی برای نمره هوش هیجانی بفرم زیر خواهد بود



توجه: نمودار چندبر فراوانی یکی از ابزارهای تشخیص چولگی می باشد.

تکلیف: در یک گزارش مفصل (فایل *Word*) نحوه رسم نمودارهای

- دایره‌ای[4]
- جعبه‌ای[5]
- پراکندگی[6]

هر مطلب می بایست شامل مواردی همانند اینکه این نمودار مناسب چه نوع داده‌های است؟، بر اساس چه مقداری از مشاهدات قابل رسم است؟، نحوه رسم و مرجع یا منبع ذکر گردند.

ضریب همبستگی[7]

میزان تغییرپذیری متغیرها نسبت به یکدیگر را همبستگی گویند. هرگاه بخواهیم میزان تاثیر تغییرات یک متغیر بر متغیر دیگر را در نظر بگیریم با مفهوم وابستگی سروکار داریم. ضریب همبستگی ابزاری برای تعیین نوع و درجه *رابطه خطی* یک متغیر پیوسته با متغیر پیوسته دیگر است. فرض کنید از هر واحد نمونه زوج مقادیر (X_i, Y_i) برای $i = 1. \dots . n$ را بدست آورده باشیم.

---

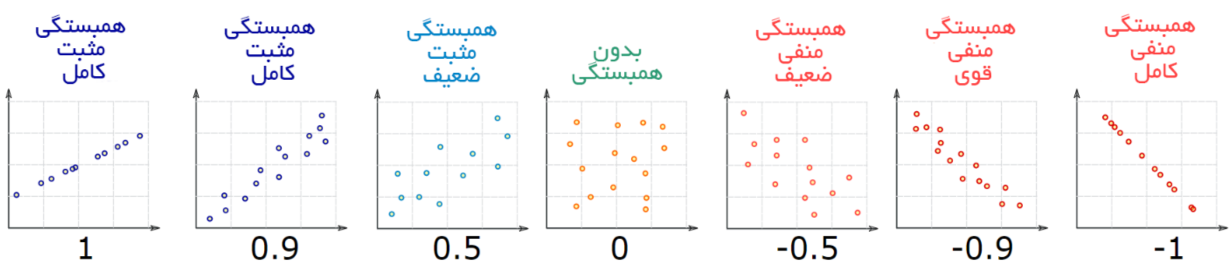
[4] Pie chart
[5] Box plot
[6] Scatter plot
[7] correlation coefficient

مثال: متوسط میزان مصرف (نوشیدن) روزانه آب توسط هر شخص حاضر در بین نمونه ها (متغیر X_i) و مقدار شاخص سلامت فرد (متغیر Y_i) در نظر می گیریم. میخواهیم بدانیم وابستگی شاخص سلامت فرد به متوسط مصرف آب روزانه را مشخص نماییم. بدین منظور از هر شخص حاضر در نمونه یک زوج مرتب شامل این دو متغیر را بدست می آوریم.

مثال: در نظر داریم بررسی نماییم که مقدار سیمان بکاررفته در ساخت بتن به چه میزان باعث افزایش استحکام آن می شود. به بیان دیگر سوالی که مطرح می شود آن است که میزان تاثیر مقدار سیمان و استحکام بتن چه میزان رابطه یا همبستگی با هم دارند؟

نکته: ضریب همبستگی، تنها به بررسی میزان وابستگی خطی بین دو متغیر پیوسته می پردازد.

بدین منظور از ابزار آماری ضریب همبستگی بهره می بریم. رابطه بین این دو متغیر می تواند یکی از حالات زیر را شامل شود



اشکال زیر را نمودار پراکندگی[8] گویند. بمنظور ایجاد این نمودار کافی است مقادیر سیمان را روی محور X و مقادیر بدست آمده از آزمایش استحکام را بر روی محور Y در نظر بگیریم و نقاط مشاهده شده از نمونه ها را رسم نماییم. نمودار حاصل را نمودار پراکندگی گویند.

با نگاه کلی به شکل نمودار پراکندگی می توان به وجود رابطه و یا عدم وجود آن بین دو متغیر پی برد. در صورتیکه با افزایش یک متغیر دیگری نیز افزایش یابد همبستگی مثبت بین دو متغیر برقرار است و در صورتیکه با افزایش یکی، دیگری کاهش یابد، همبستگی بین دو متغیر منفی خواهد بود.

---

[8] Scatter Plot

ضریب همبستگی، یکی از معیارهای مورد استفاده در تعیین همبستگی دو متغیر است. ضریب همبستگی شدت رابطه خطی و همچنین نوع رابطه (مستقیم یا معکوس) را نشان می‌دهد. فرمول ضریب همبستگی (پیرسون[9]) بفرم

$$r_{x,y}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\frac{\sum_{i=1}^{n}x_iy_i\ -n\bar{y}\bar{x}}{(n-1)\sqrt{s_x^2\ \ s_y^2}}.$$

در این فرمول s_x^2 و s_y^2 به ترتیب واریانس‌های بدست آمده از مقادیر مشاهده شده $x_1.\dots.x_n$ و مقادیر مشاهده شده $y.\dots.y_n$ می باشند. زیرا

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})&=\sum_{i=1}^{n}(y_ix_i-x_i\bar{y}-y_i\bar{x}+\bar{y}\bar{x})\\&=\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}x_i\bar{y}-\sum_{i=1}^{n}y_i\bar{x}+\sum_{i=1}^{n}\bar{y}\bar{x}\\&=\sum_{i=1}^{n}y_ix_i-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+\bar{y}\bar{x}\sum_{i=1}^{n}1\\&=\sum_{i=1}^{n}y_ix_i-\bar{y}n\bar{x}-\bar{x}n\bar{y}+\bar{y}\bar{x}n=\sum_{i=1}^{n}y_ix_i-n\bar{y}\bar{x}.\end{aligned}$$

مخرج

$$\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sqrt{\left((n-1)s_y^2\right)\left((n-1)s_x^2\right)}=(n-1)\sqrt{s_y^2\ s_x^2}$$

$r_{x.y}$ در بازه $[-1,1]$ مقدار می‌پذیرد و در صورت عدم وجود رابطه بین دو متغیر، برابر صفر (یا نزدیک صفر) است. لازم به توضیح است که همبستگی اشاره شده از نوع خطی است و چنانچه رابطه غیر خطی بین دو متغیر برقرار باشد همبستگی خطی آنها نیز صفر نتیجه خواهد شد.

نکته: اگر مقدار ضریب همبستگی پیرسون نزدیک صفر نتیجه دهد بدین معنا است که همبستگی بین دو متغیر مدنظر وجود ندارد.

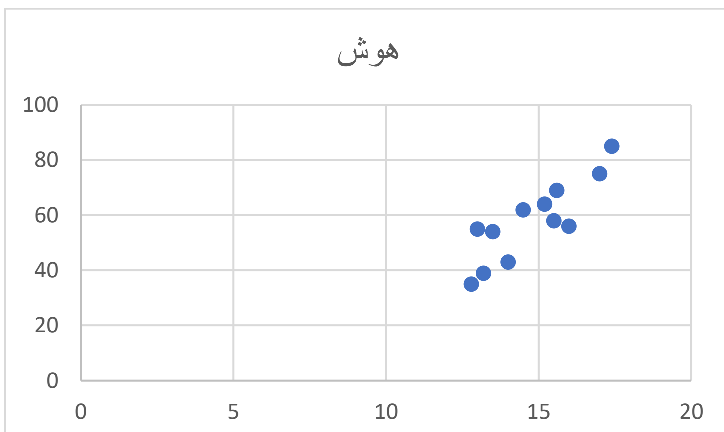
[9] Pearson

نکته: چنانچه رابطه بین دو متغیر کاملا خطی باشد بسته به شیب نمودار که مثبت است یا منفی، ضریب همبستگی ۱ و یا -۱ خواهد بود.

مثال: فرض کنید جدول مشاهدات زیر از ۱۲ دانشجویان مهندسی نقشه برداری (ورودی ۹۸) بدست آمده باشد.

شماره	۱	۲	۳	٤	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
معدل :Y	۱۶	۱۵.۵	۱٤	۱۳.۲	۱۲.۸	۱۵.۲	۱۷	۱۷.٤	۱۵.۶	۱۳.۵	۱۳	۱٤.۵
نمره :X هوش	۵۶	۵۸	٤۳	۳۹	۳۵	۶٤	۷۵	۸۵	۶۹	۵٤	۵۵	۶۲

ابتدا نمودار پراکندگی را رسم می نماییم



نمودار پراکندگی که رابطه مستقیم و خوبی را نشان می دهد. بمنظور بدست آوردن مقدار ضریب همبستگی بصورت زیر اقدام می نماییم.

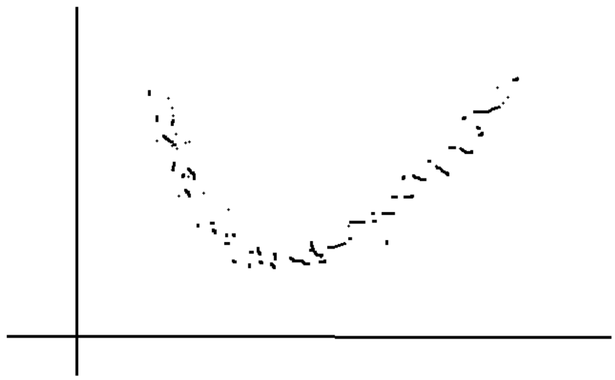
$$\bar{y} = 14.808. \quad S_y^2 = 2.413 \quad \bar{x} = 57.917. \quad S_x^2 = 212.265. \quad \sum_{i=1}^{12} x_i y_i = 10506$$

$$r_{x.y} = \frac{\sum_{i=1}^{n} x_i y_i \ - n\bar{y}\bar{x}}{(n-1)\sqrt{s_x^2 \ s_y^2}} = \frac{10506 - 12 \times 57.917 \times 14.808}{11\sqrt{212.265 \times 2.413}} = \ 0.861$$

این مقدار ضریب همبستگی نشان می دهد که رابطه مستقیم (صعودی) و قوی بین معدل اکتسابی شخص و نمره هوش وی وجود دارد. بطوریکه این میزان همبستگی مقدار ۰.۸۶۱ می باشد.

تکلیف: مثالی از داده‌های واقعی (باذکر مرجع) برای محاسبه ضریب همبستگی ذکر نمایید. (حداقل نمونه ۵۰ عدد)

نکته: به نمودار پراکندگی بین دو متغیر فرضی در شکل زیر توجه کنید



بسیار واضح است که بین این دو متغیر یک رابطه خطی (از درجه دوم) وجود دارد. ولی بطور حتم اگر ضریب همبستگی بین این دو متغیر را بدست آوریم عددی نزدیک صفر بدست خواهد آمد، زیرا ضریب همبستگی تنها عددی را ارائه می