الهی به امید تو

آمار و کاربردهای آن در علوم پزشکی (مقدماتی)

آمار: علم جمع‌آوری، سازماندهی و خلاصه کردن داده‌ها (آمار توصیفی)و استنباط و نتیجه‌گیری از بخشی از داده‌ها و تعمیم آن به مجموعه بزرگتر(آمار استنباطی) است.

آمار با یکسری داده سر و کار دارد.

تفاوت رابطه آماری و رابطه موجود در سایر رشته‌ها مانند ریاضی، فیزیک و...:

تفاوت رابطه آماری و سایر روابط در این است که در رابطه ریاضی مقدار دقیق مشخص است، اما در رابطه آماری مقدار دقیق از قبل از مشخص نیست و همراه با خطاست. خطا ناشی از در نظر نگرفتن یکسری عوامل است و هدف آمار کاهش این خطاها و در نهایت رسیدن به یک رابطه ریاضی با حداقل خطاست.

توجه: در آمار معمولاً قضاوت فردی نیست و روابط آماری روابطی هستند که روی هم رفته درست است.

آمار توصیفی:

جمع‌آوری، سازماندهی و خلاصه کردن داده‌ها با استفاده از جداول و نمودارها و محاسبه شاخص‌های عددی گرایش به مرکز و پراکندگی

مفاهیم:

جامعه: مجموعه‌ای از افراد یا اشیا که حداقل یک ویژگی مشترک دارند و هدف بررسی یک یا چند ویژگی آن است.

- جامعه متناهی
- جامعه نامتناهی

نمونه: قسمتی از جامعه که براساس روش‌های علمی انتخاب می‌شود و بررسی می‌گردد سپس نتایج حاصل از آن به جامعه تعمیم داده می‌شود.

متغیر: ویژگی مورد مطالعه جامعه است که از هر فردی به فردی دیگر می‌تواند تغییر کند.

متغیرتصادفی: اگر مقداری که متغیر می‌تواند اختیار کند تحت تأثیر شانس و به صورت تصادفی باشد.

انواع متغیر:

- کیفی(رسته‌ای): متغیرهایی هستند که با مقدار عددی مشخص نمی‌شوند، بلکه براساس یکسری شاخص افراد در رسته‌های مختلف قرار می‌گیرند.

- اسمی: درصورتی که بین رسته‌های مختلف متغیر هیچ ترتیبی وجود نداشته باشد. (در این حالت برای خلاصه کردن مشاهدات از فراوانی و فروانی نسبی استفاده می‌شود.)
- ترتیبی: درصورتی که بین رسته‌های مختلف متغیر یک ترتیب ذاتی وجود داشته باشد. (در این حالت برای خلاصه کردن مشاهدات علاوه بر استفاده از فراوانی و فروانی نسبی می‌توان از فراوانی نسبی تجمعی نیز استفاده کرد.)

- کمی: متغیرهایی هستند که با مقدار عددی مشخص می‌شوند و معمولاً به صورت اندازه یا تعداد هستند و واحد اندازه‌گیری دارند.
  - گسسته: اگر در دامنه تغییرات متغیر تصادفی نتوان هر عدد را انتخاب کرد. مقادیر ممکن کاملاً متمایز و جدا و اغلب تعداد هستند.
  - پیوسته: اگر در دامنه تغییرات متغیر تصادفی بتوان هر عددی را انتخاب کرد. اندازه‌های پیوسته معمولاً دارای حد بالا و پایین هستند.
- نسبی: درصورتی که متغیر شامل صفر ذاتی باشد یعنی صفر به معنای عدم وجود یک ویژگی باشد. نسبت حفظ می‌شود.
- فاصله‌ای: شامل صفر قراردادی است. فاصله حفظ می‌شود.

پس از جمع‌آوری داده‌ها مرحله بعدی خلاصه کردن داده‌ها است.

- نمودارها
- جداول

نمودار: یکی از روش‌های اصلی نمایش اطلاعات آماری استفاده از نمودار است. نمودارها باید ساده باشند.

اهداف اصلی نمودارها:

۱) ارائه اطلاعات آماری در مقاله‌ها و سایر گزارش‌ها، (زمانی که احساس شود خواننده یک نمایش ساده و جذاب را درک می‌کند)
۲) کمکی برای تحلیل آماری است. آماری‌ها معمولاً از نمودارها برای فهمیدن ساختار داده‌ها و چک کردن برخی پیش‌فرض‌های تحلیل‌ها استفاده می‌کنند.

کاربردهای اصلی نمودارهای آماری:

۱) برای مقایسه دو یا چند عدد
۲) برای بیان توزیع مشاهدات یا اندازه‌های رده‌های مختلف (نمودار هیستوگرام)
۳) برای بیان تغییر در برخی مقادیر در طول یک دوره زمانی (اتصال نقاط مربوط به اندازه‌ها و نمودار خطی)
۴) برای بیان ارتباط بین دو اندازه (نمودار پراکنش)

انواع نمودار:

۱) هیستوگرام: نموداری مرکب از چند مستطیل که ارتفاع آن فراوانی یا فراوانی نسبی یک متغیر کمی پیوسته است. مساحت مستطیل‌ها با جمع فراوانی‌ها برابر است.
۲) چندبر فراوانی: نموداری که از اتصال نقاط میانی مستطیل‌های هیستوگرام حاصل می‌شود.
۳) ساقه و برگ: نمودار مفیدی است که به طور همزمان داده‌ها را جدول‌بندی می‌کند و شکل و توزیع فراوانی داده‌ها را نشان می‌دهد.
۴) جعبه‌ای: این نمودار برای نشان دادن نحوه پراکندگی داده‌ها مفید است.
۵) دایره‌ای: برای نمایش داده‌های رتبه‌ای و اسمی و کمی گسسته به کار برده می‌شود. برمبنای فراوانی نسبی داده‌ها رسم می‌شود. درصورتی که فراوانی نسبی رده $\frac{f_i}{n}$ باشد، نسبت $\frac{f_i}{n}\times 360$ از دایره به آن رده اختصاص داده می‌شود.
۶) خطی: برای نشان دادن روند وقایع در طول زمان مفید است.
۷) پراکنش: برای بررسی ارتباط بین دو متغیر کمی مفید است.
۸) نقطه‌ای: زمانی که تعداد داده‌ها کم باشد استفاده می‌شود.
۹) میله‌ای

جدول: روش دیگری برای خلاصه کردن و ارائه برخی از ویژگی‌های مهم یک مجموعه داده استفاده از جدول است.

- خانه‌ای از جدول نباید صفر باشد(در این حالت برخی رده‌ها ادغام می‌شوند).

- رده‌بندی متغیرهای یک جدول می‌تواند براساس مطالعات قبلی انجام شده یا معمولا بین ۵تا ۱۰ رده درنظر گرفته می‌شود.
- عنوان جدول باید در بالای جدول نوشته شود.
- جدول باید طوری باشد که بتوان اطلاعات مناسبی را از آن استخراج کرد.

پس از جمع‌آوری و خلاصه کردن داده‌ها مرحله بعد مشخص کردن توزیع داده‌ها با استفاده از شاخص‌های مرکزی و پراکندگی است.

شاخص‌های مرکزی: اندازه‌های مکانی یا اندازه‌هایی که گرایش به مرکز را نشان می‌دهد:

۱) میانگین
- میانگین حسابی
- میانگین وزنی
- میانگین هندسی(هرگاه داده‌ها شامل درصد، نسبت و نرخ رشد باشد همچنین برای اعداد صفر و مقادیر منفی تعریف نمی‌شود)
- میانگین هارمونیک(اگر مقیاس سنجش داده‌ها ترکیبی باشد مثلاً کیلومتر برساعت)
- میانگین اصلاح شده(میانگین نمونه پیراسته، میانگین نمونه وینزوری)

۲) میانه: مقداری که ۵۰ درصد داده‌ها بیشتر از آن و ۵۰ درصد کمتر از آن هستند.

۳) چندک‌ها:
- چارک: چارک i ام داده ایست که i/4 ام داده‌ها کمتر از آن و i/4-۱ داده‌ها بیشتر از آن باشد.
- صدک: صدکi ام داده ایست که i/100 ام داده‌ها کمتر از آن و i/100-۱ داده‌ها بیشتر از آن باشد.
- دهک: چارک i ام داده ایست که i/10 ام داده‌ها کمتر از آن و i/10-۱ داده‌ها بیشتر از آن باشد.

۴) مد: داده‌ای که فراوانی آن نسبت به سایر داده‌ها بیشتر باشد.

مزایا و معایب شاخص‌های مرکزی:

۱) در محاسبه میانگین از اطلاعات تمام داده‌ها استفاده می‌شود، بنابراین در آمار اهمیت زیادی دارد.

۲) در محاسبه میانه از اطلاعات تمام افراد استفاده نمی‌شود، بنابراین کارایی آن نسبت به میانگین کمتر است، چرا که برخی از اطلاعات از بین می‌رود.

۳) اگر دو گروه داشته باشیم و مشاهدات دو گروه را ترکیب کنیم. میانه گروه آمیخته بیانگر میانه دو گروه نیست. اما این موضوع خیلی برای میانگین برقرار نیست. چرا که میانگین آمیخته میانگین وزنی دو گروه است ($(n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1+n_2)$).

۴) میانه نسبت به میانگین قابلیت کمتری دارد. بنابر ویژگی‌های میانگین در آمار استنباطی استفاده از میانگین بیشتر است(مثلاً قضیه حد مرکزی)

۵) میانگین و میانه منحصر به فرد هستند درحالی که مد منحصر به فرد نیست (یعنی جامعه می‌تواند تک مدی یا چند مدی باشد).

۶) مد عددی را گزارش می‌کند که در واقعیت وجود دارد اما میانگین و میانه اینطور نیست.

۷) در توزیع‌های چوله، میانگین به شدت تحت تأثیر داده‌های دور افتاده است، بنابراین در این شرایط میانه پایدار تر از میانگین است.

۸) درصورتی که توزیع متقارن باشد، میانگین و میانه برابرند. اگر توزیع چولگی مثبت داشته باشد میانگین بیشتر از میانه و اگر چولگی منفی داشته باشد، مانگین کمتر از میانه خواهد بود.

حال فرض کنید میانگین جوامع یکسان باشد آیا برابری میانگین‌ها نشان‌دهنده یکسان بودن جوامع است؟ خیر، زیرا ممکن است پراکندگی متفاوتی داشته باشند.

بنابراین شاخصی که کمک می‌کند توزیع داده‌ها را بفهمیم شاخص پراکندگی است.

شاخص‌های پراکندگی: تغییرپذیری حول مرکز داده‌ها

۱) برد: اغلب در مواردی که حجم نمونه کمتر از ۸باشد به کار می‌رود. این شاخص وسعت پراکندگی را منعکس می‌کند اما بیانگر خوبی برای تغییرپذیری داده‌ها نیست.

۲) میانگین انحرافات از میانگین

۳) واریانس

۴) انحراف استاندارد:[1] جذر واریانس

۵) خطای استاندارد:[2] نسبت انحراف معیار بر جذر نمونه

۶) ضریب تغییر: نسبت انحراف استاندارد به میانگین، این ضریب به واحد اندازه‌گیری بستگی ندارد، برای مقایسه به کار می‌رود و هرچه کوچکتر باشد بهتر است.

---

[1] Standard deviation

[2] Standard erorr

روش‌های تشخیص خطاهای بزرگ

۱) بررسی منطقی
۲) بررسی آماری
۳) برآورهای رباست

احتمال:

آزمایش تصادفی: آزمایشی که تمام نتایج ممکن آن از قبل معلوم باشد ولی نتیجه آزمایش، قبل از انجام آزمایش معلوم نباشد.

فضای نمونه: مجموعه نتایج ممکن یک آزمایش تصادفی

پیشامد: هر زیر مجموعه از فضای نمونه

یکسری قواعد مربوط به پیشامدها:

۱. قانون جابه جایی $A \cup B = B \cup A, \quad A \cap B = B \cap A$
۲. قانون شرکت‌پذیری $(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C)$
۳. قانون توزیع‌پذیری
$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$
۴. قانون دمورگان $(\bigcup_{i=1}^{n} A_i)' = \bigcap_{i=1}^{n} A_i'$

احتمال: اندازه امکان وقوع یک پیشامد که سه شرط دارد:

- برای هر پیشامد مثلاً A: $P(A) \geq 0$
- برای هر پیشامد حتمی S: $P(S)=1 \leftarrow 0 \leq P(A) \leq 1$
- برای هر دو پیشامد ناسازگار A و B: $P(A \cup B) = P(A) + P(B)$

توجه: دو پیشامد ناسازگارند اگر: $\left.\begin{matrix} P(A \cap B) = 0 \\ A \cap B = \phi \end{matrix}\right\}$

احتمال را براساس شاخص‌های مختلف مانند نسبت حالات مساعد به حالات ممکن، مساحت، تجربی و ... می‌توان اندازه گرفت

- **همشانسی:** هرگاه فضای نمونه یک آزمایش تصادفی شامل n پیشامد ساده باشد که از نظر رخ دادن هیچ یک بر دیگری برتری نداشته باشد پیشامدها همشانس هستند. بنابراین احتمال وقوع هر یک از پیشامدها $\frac{1}{n}$ است.

نسبت حالات مساعد به حالات ممکن: اگر یک آزمایش تصادفی انجام شود و نتایج آن همگن و دو به دو ناسازگار باشند احتمال وقوع پیشامدی مانند A:

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

قوانین:

$$P(A')=1-P(A)$$
$$if \quad A\subset B \quad \rightarrow \quad P(A)\le P(B)$$
$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$
$$P(\bigcup_{i=1}^{n}A_i)\le\sum_{i=1}^{n}P(A_i)$$
$$P(A)+P(B)-1\le P(A\cap B)\le min(P(A),P(B))$$
$$P(A_1)+P(A_2)+...+P(A_n)-(n-1)\le P(A_1\cap A_2,...,A_n)$$
$$P(A\cap B')=P(A)-P(A\cap B)$$
$$P(A'\cap B')=1-P(A\cup B)=1-P(A)-P(B)-P(A\cap B)$$
$$P(A\,|\,B)=\frac{P(A\cap B)}{P(B)}$$
$$P(A_1\times A_2\times...\times A_n)=P(A_1)\times P(A_2\,|\,A_1)\times...\times P(A_n\,|\,A_1...A_{n-1})$$
$$P(A)=P(A\cap B)+P(A\cap B')=P(A\,|\,B)\,P(B)+P(A\,|\,B')\,P(B')$$
$$P(B_j\,|\,A)=\frac{P(A\,|\,B_j)\,P(B_j)}{\sum_{i=1}^{n}P(A\,|\,B_i)\,P(B_i)}$$

- **تعریف حدی احتمال:** فراووانی وقوع پیشامد را وقتی که آزمایش تصادفی تحت شرایط یکسان مکرراً انجام شود.

$$P(A)=\lim_{n\to\infty}\frac{n(A)}{n}$$

- **تعریف تجربی احتمال:** میزان اطمینان یک فرد براساس اطلاعات و تجربیات او نسبت به وقوع یک پیشامد

- **توزیع‌ها**

انواع توزیع‌ها:

- توزیع‌های گسسته

۱) توزیع برنولی: یک آزمایش تصادفی که فضای نمونه تنها دو حالت(پیروزی، شکست) دارد.

$$P(X=x)=p^x(1-p)^{1-x} \qquad x=0,1 \qquad 0\le p\le 1$$

$$\left.\begin{matrix}P(X=1)=p\\P(X=0)=q\end{matrix}\right\} p+q=1$$

$$E(X)=p$$

$$Var(X)=pq$$

$$M_x(t)=E(e^{tx})=p\,e^t+q$$

$$if \quad X_i\sim Ber(p) \rightarrow \begin{cases}Y=\sum X_i\sim B(n,p)\\ Y=X_i^k\sim Ber(p)\\ Y=min(X_i)\sim Ber(p^n)\end{cases}$$

۲) توزیع دوجمله‌ای: تکرار n بار آزمایش برنولی به صورت مستقل از هم

x نشان‌دهنده تعداد موفقیت‌ها در n آزمایش برنولی

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \qquad x=0,1,\dots,n \qquad 0\le p\le 1$$

$$\left.\begin{matrix}E(X)=np\\Var(X)=npq\end{matrix}\right\} Var(X)<E(X)$$

$$M_x(t)=E(e^{tx})=(p\,e^t+q)^n$$

$$if \quad X_i\sim B(n,p) \rightarrow \{X_i\,|Y\sim HG(n,1,y)$$

در این توزیع اگر p=0.5 باشد توزیع متقارن، اگر p<0.5 باشد توزیع چوله به راست و اگر p>0.5 باشد توزیه چوله به چپ خواهد بود.

۳) توزیع پواسون: این توزیع برای بسیاری از پدیده‌های تصادفی که در یک دوره زمانی رخ می‌دهند به کار می‌رود.

x نشان‌دهنده تعداد موفقیت‌ها در یک بازه زمانی است.

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x=0,1,\ldots, \qquad \lambda>0$$

$$E(X)=\lambda$$

$$Var(X)=\lambda$$

$$M_x(t)=E(e^{tx})=exp(\lambda(e^{t}-1))$$

$$if \quad X_i \sim P(\lambda) \rightarrow \begin{cases} Y=\sum X_i \sim P(n\lambda) \\ X_i | Y \sim B(y,\frac{1}{n}) \end{cases}$$

این توزیع چوله به راست است و با افزایش λ بیشتر حالت تقارن پیدا می‌کند.

- توزیع‌های پیوسته

۱) توزیع نرمال

$$X \sim N(\mu,\sigma^2)$$

$$f(X=x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty \le x \le \infty$$

$$E(X)=\mu$$

$$Var(X)=\sigma^2$$

$$M_x(t)=exp(\mu t-\frac{1}{2}\sigma^2 t^2)$$

$$if \ Z_1,\ldots,Z_n \sim N(0,1) \rightarrow \begin{cases} Y=\sum Z_i \sim N(0,n) \\ Y=\sum Z_i^2 \sim \chi_n^2 \\ U=\frac{Z_1}{Z_2} \sim C(0,1) \\ V=\frac{Z_1}{|Z_2|} \sim C(0,1) \end{cases}$$