Statystyka opisowa

Stawia się pytania: pytanie „**co**?” poprzedza pytanie „**jak**?”. Najpierw potrzebna jest miara, potem można badać zmiany tej miary.

Potrzebne są miary zbiorcze, charakteryzujące zbiorowości (populacje).

Miarą zbiorczą może być histogram albo wykres słupkowy, czyli rozkład częstości, ale bardziej przydatne są miary liczbowe (numeryczne).

Są dwa podstawowe rodzaje miar statystyki opisowej: **statystyki położenia** oraz **statystyki rozrzutu**.

Statystyka położenia opisuje miejsce próby na osi liczbowej, na skali. Statystyka taka musi być reprezentantem dla więcej niż jednej obserwacji. Nie opisuje ona rozkładu częstości ani jego nachylenia. Rozkład opisywany przez taką statystykę może być U-kształtny, może być rozciągnięty lub bardzo ostry (zależy to także od przyjętej skali), może posiadać dwa lub więcej szczytów, może być nawet bardzo niesymetryczny. Potrzebne są zatem także miary rozproszenia wyników.

Średnia arytmetyczna

Najbardziej popularna miara statystyki położenia znana jest wszystkim. Nazywana jest **wartością średnią**, **średnią** lub **wartością przeciętną**. Obliczana jest przez sumowanie poszczególnych obserwacji w próbie i podzielenie tej sumy przez ilość wyników.

Notacja: $X_1, X_2, ..., X_n$ – poszczególne wyniki pomiarów.

Suma wyników pomiarów: $\sum_{i=1}^{i=n} X_i = X_1 + X_2 + ... + X_n$.

Duża litera grecka Σ (sigma) oznacza sumę wszystkich wskazanych wielkości; należy ją czytać jako **suma**. Wyrażenie i = 1 oznacza, że sumowanie ma rozpocząć się od wartości pierwszej; sumowanie ma zakończyć się na liczbie ostatniej - i = n. Indeksy dolny i górny wskazują granice sumowania, przy czym często zamiast i = n pisze się n.

Różne notacje sumowania, od najbardziej złożonej do najbardziej prostej:

$$\sum_{i=1}^{i=n} X_i = \sum_{i=1}^{n} X_i = \sum_{i} X_i = \sum X.$$

Symbol wartości średniej arytmetycznej to przykładowo $\overline{X}$, ale stosowane są również inne, np. $\overline{x}$, $\overline{Y}$ czy $\overline{y}$. Zatem:

$$\overline{X} = \frac{\sum X}{n} \text{ lub } \overline{X} = \frac{\sum X}{N}.$$

Czyta się: sumę wszystkich (n) wartości podzielić przez ich ilość (n).

Wartość średnia dla próby reprezentuje środek obserwacji w próbie.

Często konieczne jest uśrednianie średnich lub innych statystyk, które mogą różnić się pewnością; statystyki takie mogą reprezentować próby o różnej wielkości lub mogą być zróżnicowane w inny sposób. W takich przypadkach należy obliczać **średnie ważone**. Ogólny wzór obliczeniowy średniej ważonej dla zbioru wartości X_i można przedstawić jako

$$\overline{X}_w = \frac{\sum_i w_i X_i}{\sum_i w_i},$$

gdzie uśrednia się n liczb, każda ważona przez współczynnik w_i. Średnie ważona i nieważona nie muszą dawać tych samych wartości.

Niekiedy zmienne są **transformowane** i interpretowane są logarytmy lub odwrotności wartości mierzonych. Jeżeli oblicza się wartość średnią dla wartości transformowanych i na powrót zmienia się skalę na oryginalną, to otrzymana liczba nie będzie taka sama jak wartość średnia obliczona z wyników wyjściowych. Odwrotnie transformowana średnia obliczona dla wartości transformowanych logarytmicznie nazywa się **średnią geometryczną**. Oblicza się ją jako

$$\overline{X}_G = \text{anty}\log\left(\frac{1}{n}\sum \log X\right),$$

średnia geometryczna jest więc antylogarytmem średniej wartości obliczonej z logarytmów wartości **X**. Ponieważ dodawanie logarytmów jest równoważne mnożeniu ich antylogarytmów, więc inną metodą obliczania tej wielkości jest

$$\overline{X}_G = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n}.$$

Podobnie jak w przypadku obliczania średniej arytmetycznej według zapisu symbolicznego działanie takie można zapisać (Π - pi – duża litera w alfabecie greckim) jako

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} X_i}.$$

Odwrotność średniej arytmetycznej obliczonej dla odwrotności wyników pomiarów nazywana jest **średnią harmoniczną**:

$$\frac{1}{\overline{X}_H} = \frac{1}{n}\sum_i \frac{1}{X_i}.$$

Dla tego samego zbioru wartości liczbowych obliczona średnia geometryczna jest zawsze mniejsza od średniej arytmetycznej, a średnia harmoniczna jest zawsze mniejsza od średniej geometrycznej. Zastosowanie tych miar położenia albo tendencji centralnych związane jest z przedstawianiem wyników w postaci graficznej w różnych **układach współrzędnych**: odpowiednio liniowym, logarytmicznym lub odwrotnościowym.

Mediana

Mediana jest statystyką położenia przydatną w przedstawianiu niektórych wyników badań biologicznych. Jest ona definiowana jako wartość zmiennej w szeregu uporządkowanym, która posiada taką samą ilość liczb w obu kierunkach. W ten sposób mediana dzieli rozkład częstości na dwie połowy. Wielkość tę łatwo wyznacza się w przypadku nieparzystej ilości pomiarów; jeżeli ilość wyników jest parzysta, to zwyczajowo oblicza się ją jako wartość środkową pomiędzy dwiema: $n/2$-ą oraz (n/2+1)-ą. Ogólnie można ją obliczyć jako $\mathrm{Me} = X_{(n+1)/2}$.

Jeżeli wyniki pomiarów znane są jedynie w postaci **rozkładu częstości** (**szeregu rozdzielczego**), obliczenie mediany jest nieco bardziej złożone, ale także jednoznaczne.

Wartość modalna

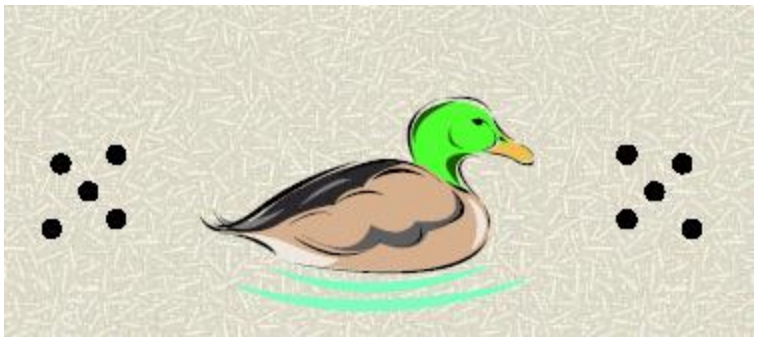
Wartość modalna (moda) jest wartością o najwyższej frekwencji w rozkładzie częstości - maksimum częstości. Rozkłady o dwóch pikach (maksimach) o takiej samej lub o różnej wysokości nazywane są rozkładami dwumodalnymi; jeżeli występuje więcej pików, rozkład jest multimodalny.

Preferowaną wartością charakteryzującą położenie liczb jest średnia arytmetyczna, ponieważ charakteryzuje się ona mniejszą wartością **błędu standardowego**. Wartość średnia ma także dodatkową korzystną cechę: ma tendencję do zbliżania się do rozkładu normalnego w sytuacji, gdy wyjściowe wyniki nie mają rozkładu normalnego. Na średnią w znacznie wpływają wartości odstające, a na medianę i modę nie. Średnia jest z reguły bardziej wrażliwa na kształt rozkładu częstości.

Dla symetrycznego rozkładu jednomodalnego średnia, mediana i moda są identyczne.

Proste statystyki rozproszenia

Identycznymi wartościami średniej arytmetycznej mogą charakteryzować się skrajnie zróżnicowane rozkłady (w przedstawieniu graficznym bardzo zróżnicowane histogramy).



Średnio nieźle, ale po co?

Prostą miarą rozrzutu jest **rozstęp**. Jest to różnica pomiędzy wynikiem największym a najmniejszym. Rozstęp wyrażony jest w takiej samej jednostce, jak pomiary wyjściowe. Na rozstęp ma oczywiście znaczący wpływ nawet pojedyncza wartość odstająca i z tego powodu jest on jedynie zgrubnym oszacowaniem wszystkich wyników w próbie. Rozstęp zależny jest także od wielkości próby; im większa próba, tym większy rozstęp.

Odchylenie standardowe bierze pod uwagę wszystkie pomiary, a do każdego przykłada wagę, którą jest jego odległość od środka rozkładu. Poszczególne odległości, czyli odchylenia, oblicza się raczej jako $X - \overline{X}$, a nie jako $\overline{X} - X$. Suma takich odchyleń dla dowolnych zbiorów liczb jest zawsze równa zero. Aby uniknąć tego rodzaju niewygodnych zależności, zamiast odchyleń dodaje się ich kwadraty. W ten sposób otrzymuje się ważną wielkość - **sumę kwadratów odchyleń** lub krócej **sumę kwadratów**. Po podzieleniu tej sumy przez ilość wyników otrzymuje się inną ważną wielkość – **wariancję** (**s^2**), która jest kwadratem średniego odchylenia wyników od średniej i wyraża się ją w jednostkach kwadratowych. Dodatni pierwiastek kwadratowy z wariancji nosi nazwę **odchylenia standardowego** (**s**); jest ono wyrażane w jednostkach oryginalnych i jest miarą średniego odchylenia wyników od wartości średniej.

Statystyki z próby a parametry.

Prawidłowo obliczone wartości średniej arytmetycznej i odchylenia standardowego są zawsze prawdziwymi miarami położenia i rozrzutu dla prób, z których są obliczane. Jednak dla badacza rzadko jest interesujące obliczanie **statystyk** charakteryzujących próby, natomiast interesujące są wielkości charakteryzujące **populacje**, z których **próby** zostały pobrane. Nie jest interesująca wartość średnia obliczona z (przykładowo) bardziej czy mniej przypadkowych wyników czterech pomiarów, lecz prawdziwa, rzeczywista wartość - **średnia dla populacji**. Statystyki charakteryzujące populacje są nieznane i najczęściej nawet niepoznawalne. Zatem statystyki obliczone dla prób są oszacowaniami **parametrów**, czyli statystyk charakteryzujących populacje.

Umownie litery greckie używane są jako symbole parametrów populacji, a litery alfabetu łacińskiego jako symbole statystyk prób. Tak więc $\overline{X}$ jest oszacowaniem μ (litera alfabetu greckiego mi) - wartości średniej charakteryzującej populację, a wariancja (s^2) oszacowaniem σ^2 (litera alfabetu greckiego sigma) – wariancji charakteryzującej populację. Takie estymatory powinny być nieobciążone, czyli próby (bez względu na ich wielkość) pobrane z populacji o znanym parametrze powinny dawać statystyki prób, które po uśrednieniu będą dawać wartość parametru. Estymator, który tego nie daje, nazywany jest obciążonym. Średnia próby $\overline{X}$ jest nieobciążonym estymatorem średniej dla populacji μ. Jednakże wariancja dla próby nie jest nieobciążona, lecz średnio niedoszacowuje wielkość wariancji dla populacji σ^2. Wykazano, że jeżeli sumę kwadratów podzieli się przez $n - 1$, to otrzymana w efekcie wartość wariancji z próby jest nieobciążonym estymatorem wariancji dla populacji. Zatem przyjmuje się, że wariancję dla próby oblicza się dzieląc sumę kwadratów przez $n - 1$. Im większa jest próba, tym różnica pomiędzy estymatorami obciążonym i nieobciążonym jest mniejsza. Wielkość z mianownika, $n - 1$, nazywana jest **ilością stopni swobody**. Jeżeli badacz zainteresowany jest tylko interpretacją swoich wyników lub bada całą populację, wtedy - i tylko wtedy - w mianowniku wyrażenia na wariancję podaje się n.

Niekiedy interesujące jest znalezienie odpowiedzi na pytanie, czy populacje są tak samo zmienne. Jednak w przypadku, gdy wartości średnie są mocno zróżnicowane, porównanie wariancji lub odchyleń standardowych jest mało przydatne, ponieważ większe organizmy są z reguły bardziej zmienne niż mniejsze. W celu porównania względnych wielkości zmienności populacji oblicza się **współczynnik zmienności** ν, czyli iloraz odchylenia standardowego i wartości średniej arytmetycznej. Wartość współczynnika zmienności jest niezależna od jednostek pomiarowych. Obliczany ze wzoru

$$\nu = \frac{s \cdot 100}{\overline{X}}$$

jest estymatorem obciążonym. Estymator nieobciążony ν^* można obliczyć stosując wzór $\nu^* = \left(1 + \frac{1}{4n}\right) \cdot \nu$; dla prób o niewielu wynikach poprawka może być znacząca.