

STATYSTYKA OPISOWA

Dr Alina Gleska

Instytut Matematyki WE PP

28 września 2018

- 1 Miary statystyczne
- 2 Miary położenia

Wyróżniamy następujące miary statystyczne:

- POŁOŻENIA, które służą do określenia takiej wartości cechy, wokół której skupiają się wszystkie pozostałe wartości tej cechy;
- ZMIENNOŚCI (ROZPROSZENIA, DYSPERSJI), które określają stopień zróżnicowania wartości badanej cechy;
- ASYMETRII (SKOŚNOŚCI), które są wykorzystywane do badania kierunku i stopnia nierównomierności rozmieszczenia jednostek zbiorowości ze względu na wartości badanej cechy na całym obszarze jej zmienności;
- KONCENTRACJI (KURTOZY), które określają stopień nierównomierności rozłożenia ogólnej sumy wartości cechy pomiędzy poszczególne jednostki zbiorowości lub ukazują stopień skupienia poszczególnych jednostek wokół średniej;

- Wymienione statystyki opisowe pozwalają w sposób syntetyczny określić właściwości badanych rozkładów empirycznych i dokonać ich porównania. W szczególności można wyróżnić dwa typy porównań:

- porównanie rozkładów dla różnych zbiorowości względem tej samej cechy,
- porównanie rozkładów dla różnych cech dotyczących jednej zbiorowości.

Wyróżniamy dwa podstawowe rodzaje miar położenia: KLASYCZNE i POZYCYJNE. Klasyczne miary położenia oblicza się na podstawie wszystkich wartości cechy i zaliczamy do nich: średnią arytmetyczną, średnią harmoniczną i średnią geometryczną.

Pozycyjne miary położenia otrzymuje się w wyniku uznania wartości cechy dla pewnej jednostki za wyróżnioną z punktu widzenia pozycji tej jednostki, jaką zajmuje ona w uporządkowanym szeregu statystycznym. Do pozycyjnych miar położenia zaliczamy dominantę (zwaną też modą) oraz kwantyle, spośród których wyróżniamy: kwartył pierwszy, medianę (czyli kwartył drugi), kwartył trzeci, decyle i centyle.

(I) średnią prostą, którą oblicza się z wartości cechy przedstawionej w postaci szeregu szczegółowego:

gdzie x_i jest wartością cechy w i -tej jednostce, a n - liczebnością próby,

- $$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i,$$

a) dla szeregu rozdzielczego punktowego x_i jest i -tą wartością cechy skokowej o liczebności n_i , która przyjmuje k wartości,

- Zauważmy, że jeżeli nie znamy liczebności klas, a jedynie wskaźniki struktury tych klas, czyli ich udziały procentowe, to stosujemy wzory

- dla szeregów przedziałowych: $\bar{x} = \sum_{i=1}^k x_i^0 \omega_i$ lub $\bar{x} = \frac{\sum_{i=1}^k x_i^0 p_i}{100}$.

UWAGI :

- szereg przedziałowy pozwala tylko na obliczenie przybliżonej wartości średniej, ponieważ cały przedział klasowy reprezentowany jest tylko jedną liczbą - środkiem przedziału, a przecież obserwacje w danym przedziale nie muszą być rozłożone równomiernie,
- środek przedziału NIE JEST średnią wartością cechy - różnice między tymi wartościami mogą być znaczne, tym większe, im bardziej nierównomiernie rozłożone są obserwacje,

- ◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

(III) średnia ważona jako tzw. średnia ze średnich

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^k \bar{x}_i n_i,$$

gdzie $\bar{\bar{x}}$ to średnia arytmetyczna całej zbiorowości (tj. wszystkich grup łącznie), $\bar{x}_i$ to średnia arytmetyczna i-tej grupy, n_i to liczebność i-tej grupy, a k to liczba grup, na jakie podzielono całą zbiorowość;

(IV) średnia dla wielkości stosunkowych, nazywanych wskaźnikami natężenia (prędkość pojazdu, gęstość zaludnienia, plony, ceny, wydajność pracy itp.). Obliczamy ją ze wzoru:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i,$$

przy czym wagami n_i będą te wielkości, których jednostki miary są w mianowniku.

Region X	Region Y
gęstość zal. 650 osób na km ²	gęstość zal. 80 osób na km ²
pow. 3800 km ²	pow. 10500 km ²

$$\bar{X} = \frac{650 \cdot 3800 + 80 \cdot 10500}{3800 + 10500} = \frac{2470000 + 840000}{14300} = \frac{3310000}{14300} \approx 231.5$$

PODSUMOWUJĄC:

- średnia arytmetyczna jest miarą odpowiednią dla oceny przeciętnego poziomu cechy w sytuacjach, gdy:
 - a) zbiorowość jest jednorodna, czyli nie występują wartości skrajne, znacznie różniące się od pozostałych, a rozkład jest symetryczny albo umiarkowanie asymetryczny (tzn. zbliżony do symetrycznego),
 - b) rozkład jest jednomodalny (ma jedno wyraźnie zaznaczone maksimum),
- średnia arytmetyczna nie znajduje zastosowania w przypadku, gdy:
 - a) występuje bardzo silna asymetria,
 - b) rozkład jest dwumodalny lub wielomodalny,
 - c) rozkład jest siodłowy (w kształcie litery U - wówczas bowiem średnia arytmetyczna odpowiada takiej wartości cechy, wokół której skupia się stosunkowo najmniejsza część zbiorowości, podczas gdy my oczekujemy, aby średnia obrazowała taką wartość, którą można uznać za typową, czyli taką, wokół której skupia się duża część jednostek).

ŚREDNIA GEOMETRYCZNA jest miarą klasyczną, wykorzystywaną jednak w zagadnieniach szczególnych, jak np. analiza szeregów czasowych. Jeśli poszczególne wartości cechy oznaczmy jako: $x_1, x_2, \dots, x_n$, a średnią geometryczną przez $\overline{x_g}$, to wyrażamy ją wzorem:

$$\overline{x_g} = \sqrt[n]{x_1 x_2 \dots x_n}.$$

Za pomocą średniej geometrycznej oblicza się średnie tempo zmian zjawisk w czasie.

ŚREDNIA HARMONICZNA - stosuje się ją wówczas, kiedy wartości cechy podane są w jednostkach stosunkowych (tzw. łamanych), czyli jeśli cecha wyrażona jest w przeliczeniu na jednostkę innej cechy, np.

- gęstość zaludnienia (wyrażona jako liczba osób na km^2),
- dochód na osobę (wyrażony w zł na 1 osobę),
- cena (w zł za 1 kg lub w zł za 1 szt),
- prędkość (w km na godzinę).

Średnią harmoniczną oblicza się według wzoru:

$$\overline{x_H} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

Miara ta jest stosowana bardzo rzadko.

Dla dodatnich wartości cechy zachodzą związki między średnimi:

$$\overline{x_H} \leq \overline{x_g} \leq \overline{x}.$$

DOMINANTA (inaczej: MODA lub WARTOŚĆ MODALNA) jest najczęściej stosowaną pozycyjną miarą położenia. Oznaczenia: Mo, Do, D. Jedyna miara położenia, która ma sens w przypadku analizy cech jakościowych. Dominanta jest to taka wartość cechy (lub wariant w przypadku cech jakościowych), która w danej zbiorowości występuje najczęściej (dominuje). UWAGA: dominanta to wartość cechy, a nie jej liczba wystąpień!

Przy definiowaniu dominanty należy wprowadzić następujące rozróżnienie:

- dla cechy skokowej dominanta jest to taka wartość cechy, która w danej zbiorowości występuje najczęściej,
- dla cechy ciągłej dominanta to taka wartość cechy, wokół której oscyluje (jest zgrupowanych) najwięcej pomiarów. W szeregach rozdzielczych przedziałowych można wskazać przedział, w którym dominanta występuje, zaś jej wartość przybliżoną obliczyć za pomocą wzoru interpolacyjnego

$$Do = Mo = x_{ld} + \frac{(n_s - n_{s-1})}{(n_s - n_{s-1}) + (n_s - n_{s+1})} \cdot d,$$

gdzie:

- s - numer przedziału klasowego o największej liczebności,
- x_{ld} - dolna granica s -tego przedziału klasowego,
- d - długość przedziału klasowego,
- n_s - liczebność s -tego przedziału klasowego.

Przy obliczaniu dominanty na podstawie szeregu rozdzielczego przedziałowego należy przestrzegać następujących zasad praktycznych:

- wyznaczanie dominanty ma sens wtedy, kiedy rozkład empiryczny jest jednomodalny, czyli kiedy w szeregu występuje jedno wyraźnie zaznaczone maksimum,
- przedział, w którym występuje dominanta (przedział o największej liczebności) oraz dwa sąsiadujące z nim przedziały muszą mieć takie same rozpiętości; w przeciwnym razie dominanty nie obliczamy, gdyż w szeregach o różnej długości przedziałów klasowych duża liczebność może być spowodowana większą rozpiętością przedziału w stosunku do innych przedziałów,

- obliczanie dominanty jest nieuzasadnione, gdy rozkład empiryczny jest wielomodalny,
- wartość dominanty nie zmienia się, gdy we wzorze interpolacyjnym zamiast liczebności wprowadzimy częstości.

Graficzny sposób wyznaczania dominanty - po sporządzeniu odpowiedniego histogramu (lub części histogramu dla przedziału dominanty i dwóch przedziałów sąsiadujących z nim) postępujemy następująco:

- z górnych wierzchołków najwyższego prostokąta wykreślamy dwa odcinki łączące bliższe wierzchołki sąsiednich prostokątów,
- zaznaczamy punkt przecięcia tych odcinków i rzutujemy ten punkt na oś X (oś wartości cechy),
- rzut ten wskazuje nam wartość dominanty, którą odczytujemy na osi X z pewnym przybliżeniem.

MEDIANA (ozn. Me) - tę miarę można wyznaczać dla cech wyrażonych w dowolnej skali z wyjątkiem nominalnej. Mediana jest wartością cechy, jaką ma jednostka leżąca w środku uporządkowanego ciągu obserwacji. Z tego względu mediana nazywana jest często WARTOŚCIĄ ŚRODKOWĄ. Numer jednostki środkowej nazywa się pozycją mediany.

Upraszczając zagadnienie można powiedzieć, że mediana dzieli zbiorowość statystyczną na dwie równe części w ten sposób, że połowa jednostek zbiorowości przyjmuje wartości mniejsze od mediany, natomiast druga połowa - wartości większe od mediany. W przypadku cech skokowych należałoby stwierdzić, że przynajmniej połowa jednostek przyjmuje wartości MNIEJSZE lub RÓWNE medianie i przynajmniej połowa jednostek przyjmuje wartości WIĘKSZE lub RÓWNE medianie.

JAK ZNALEŹĆ MEDIANĘ?

Schemat postępowania w przypadku szeregu szczegółowego

A) Przypadek nieparzystej liczby obserwacji.

- 1) Porządkujemy wartości cechy (od najmniejszej do największej lub odwrotnie).
- 2) Wyznaczamy pozycję mediany według wzoru

$$\text{poz}(Me) = \frac{n+1}{2},$$

gdzie n oznacza liczbę obserwacji (czyli liczebność zbiorowości),

- 3) Odczytujemy z szeregu, jaką wartość cechy ma jednostka na wyznaczonej przez nas pozycji. Ta wartość jest medianą.

$$Me = x_{\frac{n+1}{2}}$$

B) Przypadek parzystej liczby obserwacji.

- 1) Porządkujemy wartości cechy (od najmniejszej do największej lub odwrotnie).
- 2) Wyznaczamy pozycję mediany, która w przypadku, gdy n jest parzyste znajduje się między dwiema środkowymi obserwacjami. Oznaczmy te obserwacje $L1$ i $L2$. Ich pozycje określimy według następujących wzorów:

$$poz(L1) = \frac{n}{2}, \quad poz(L2) = \frac{n}{2} + 1.$$

- 3) Odczytujemy z szeregu wartości $L1$ i $L2$, a następnie obliczamy wartość mediany jako średnią arytmetyczną wskazanych liczb, czyli:

$$Me = \frac{L1 + L2}{2}.$$

Tak więc w przypadku parzystej liczby obserwacji mediana jest średnią arytmetyczną dwóch środkowych wartości cechy. Można to również zapisać tak:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}.$$

UWAGA 1: częstym błędem jest mylenie mediany z jej pozycją. Należy pamiętać, że mediana jest wartością cechy środkowej jednostki, a nie pozycją tej jednostki.

UWAGA 2: zaletą mediany jest to, że jest nieczuła na zmianę wartości ekstremalnych (a właśnie wartości ekstremalne uzyskane z pomiaru są najbardziej niepewne). Mediana, w odróżnieniu od średniej arytmetycznej, nie zmienia się, gdy wartości ekstremalne zmieniają się lub znikną.

UWAGA 3: w przypadku rozkładów asymetrycznych mediana dobrze informuje o przeciętnym poziomie cechy. Jej wartość, w odróżnieniu od średniej arytmetycznej, będzie zbliżona do większości pomiarów.

UWAGA 4: jak już zostało powiedziane, pole powierzchni pod krzywą liczebności równe jest n , czyli liczebności zbiorowości. Ponieważ mediana dzieli zbiorowość na dwie części o takich samych liczebnościach, więc graficznie medianie odpowiada taka wartość cechy (na osi X), dla której pole powierzchni pod krzywą (liczebności lub częstości) jest podzielone na połowy (rys).

Schemat postępowania w przypadku szeregu punktowego
Podobnie jak poprzednio:

- 1) wyznaczamy numer jednostki środkowej jako $\frac{n}{2}$ (lub $\frac{n+1}{2}$ dla szeregów o nieparzystej liczbie obserwacji),
- 2) ustalamy, w której klasie szeregu punktowego znajduje się jednostka środkowa; będzie to ta klasa, w której po raz pierwszy liczebność skumulowana osiągnie lub przekroczy $\frac{n}{2}$,
- 3) odczytujemy bezpośrednio z szeregu punktowego wartość cechy w ustalonej klasie. Jest to wartość mediany. Jeżeli zamiast liczebności mamy wskaźniki struktury, to szukamy klasy, w której po raz pierwszy skumulowany wskaźnik struktury przekroczy 0,5.

lub

$$Me = x_{lm} + \frac{\frac{n}{2} - n_{m-1}^{skum}}{n_m} \cdot d_m$$

lub

$$Me = x_{lm} + \frac{0.5 - \omega_{m-1}^{skum}}{\omega_m} \cdot d_m$$

$$Me = x_{lm} + \frac{50 - p_{m-1}^{skum}}{p_m} \cdot d_m,$$

gdzie Me - mediana, x_{lm} - lewy koniec przedziału mediany, n - liczebność zbiorowości, n_m - liczebność przedziału mediany, n_{m-1}^{skum} - liczebność skumulowana przedziału poprzedzającego przedział mediany, d_m - długość przedziału mediany.

Graficzny sposób wyznaczania mediany

W tym celu należy sporządzić łamaną liczebności skumulowanej (lub łamaną częstości skumulowanej), a następnie:

- 1) na osi Y zaznaczyć punkt określający pozycję mediany, czyli $\frac{n}{2}$ (w przypadku łamanej częstości skumulowanej będzie to punkt 0,5 czyli 50%),
- 2) na sporządzonej łamanej znajdujemy punkt odpowiadający wyznaczonej pozycji, czyli prowadzimy linię poziomą przecinającą oś Y w punkcie $\frac{n}{2}$,
- 3) punkt znaleziony na łamanej rzutujemy na oś X (oś wartości cechy),
- 4) rzut ten wskazuje nam wartość mediany, którą odczytujemy na osi X z pewnym przybliżeniem (rys).

Pomiędzy średnią arytmetyczną, dominantą i medianą istnieje ustalony związek w zależności od typu rozkładu.

A) W rozkładzie symetrycznym jednomodalnym (z jednym ekstremum) wszystkie trzy miary: średnia arytmetyczna, mediana i dominanta, są równe.

$$\bar{x} = Do = Me.$$

B) W rozkładzie o asymetrii prawostronnej (dodatniej) dominanta przyjmuje wartość najmniejszą, większa od niej jest mediana, a największa jest średnia arytmetyczna.

$$Do < Me < \bar{x}.$$

C) W rozkładzie o asymetrii lewostronnej (ujemnej) najniższą wartość przyjmuje średnia arytmetyczna, większa od niej jest mediana, a największa dominanta.

$$\bar{x} < Me < Do.$$

Obok powyższych zależności między średnią arytmetyczną, medianą i dominantą istnieje zależność zwana równaniem Pearsona: $Do = 3Me - 2\bar{x}$. Zależność ta jest prawdziwa tylko w przypadku, gdy rozkład badanej cechy jest symetryczny, słabo asymetryczny, ewentualnie umiarkowanie asymetryczny. Na wykładzie o miarach asymetrii rozkładu będzie wyjaśnione, kiedy rozkład jest słabo, umiarkowanie czy silnie asymetryczny.

KWANTYLE są pozycyjnymi miarami położenia. Są takimi wartościami cechy, które dzielą zbiorowość na określone części pod względem liczby jednostek. Części te pozostają względem siebie w określonych proporcjach. Najczęściej używanymi kwantylami są:

- kwartyle, które dzielą zbiorowość na cztery liczebnie równe części (po 25% jednostek każda),
- decyle, które dzielą zbiorowość na dziesięć liczebnie równych części (po 10% jednostek każda),
- percentyle (inaczej: centyle), które dzielą zbiorowość na sto liczebnie równych części (po 1% jednostek każda).

Należy podkreślić, że kwantyle, a w szczególności decyle i centyle, stosuje się tylko w przypadku, gdy liczebność zbiorowości jest dostatecznie duża.

KWARTYLE są takimi wartościami cechy, które dzielą uporządkowany ciąg obserwacji na cztery liczebnie równe części zwane grupami kwartylowymi. Podział na cztery części daje nam trzy kwartyle:

- kwartyl pierwszy Q_1 (tzw. kwartyl dolny) dzieli zbiorowość na dwie części - w pierwszej znajduje się 25% jednostek o wartościach mniejszych od niego, a w drugiej 75% jednostek o wartościach większych od niego,
- kwartyl drugi to mediana,
- kwartyl trzeci Q_3 (tzw. kwartyl górny) dzieli zbiorowość na dwie części - w pierwszej znajduje się 75% jednostek o wartościach mniejszych od niego, a w drugiej 25% jednostek o wartościach większych od niego.

W literaturze nie ma zgodności co do sposobu wyznaczania i interpretacji kwartyli. Najczęściej przyjmuje się następujące schematy

- dla szeregu szczegółowego Q_1 to wartość cechy na pozycji $n/4$, a Q_3 to wartość cechy na pozycji $3n/4$ (jeżeli któraś z tych pozycji przypada między obserwacjami, to bierzemy średnią arytmetyczną tych dwóch sąsiednich obserwacji),
- dla szeregu rozdzielczego punktowego postępujemy podobnie jak w przypadku mediany - określamy klasy, w których liczebność skumulowana po raz pierwszy osiągnie lub przekroczy $n/4$ i $3n/4$, a potem odczytujemy kwartyle,
- dla szeregu rozdzielczego przedziałowego korzystamy ze wzorów interpolacyjnych:

$$Q_1 = x_{/q_1} + \frac{\frac{n}{4} - n_{q_1-1}^{skum}}{n_{q_1}} \cdot d_{q_1}$$

lub

$$Q_1 = x_{/q_1} + \frac{0.25 - w_{q_1-1}^{skum}}{w_{q_1}} \cdot d_{q_1}$$

lub

$$Q_1 = x_{l_{q_1}} + \frac{25 - p_{q_1-1}^{skum}}{p_{q_1}} \cdot d_{q_1},$$

gdzie Q_1 - kwartyl dolny, $x_{l_{q_1}}$ - lewy koniec przedziału pierwszego kwartyla, n - liczebność zbiorowości, n_{q_1} - liczebność przedziału pierwszego kwartyla, $n_{q_1-1}^{skum}$ - liczebność skumulowana przedziału poprzedzającego przedział pierwszego kwartyla, d_{q_1} - długość przedziału pierwszego kwartyla;

$$Q_3 = x_{lq_3} + \frac{\frac{3n}{4} - n_{q_3-1}^{skum}}{n_{q_3}} \cdot d_{q_3}$$

lub

$$Q_3 = x_{lq_3} + \frac{0.75 - \omega_{q_3-1}^{skum}}{\omega_{q_3}} \cdot d_{q_3}$$

lub

$$Q_3 = x_{lq_3} + \frac{75 - p_{q_3-1}^{skum}}{p_{q_3}} \cdot d_{q_3},$$

gdzie Q_3 - kwartył górny, x_{lq_3} - lewy koniec przedziału trzeciego kwartyła, n - liczebność zbiorowości, n_{q_3} - liczebność przedziału trzeciego kwartyła, $n_{q_3-1}^{skum}$ - liczebność skumulowana przedziału poprzedzającego przedział trzeciego kwartyła, d_{q_3} - długość przedziału trzeciego kwartyła.

GRAFICZNY sposób wyznaczania kwartyli jest taki sam, jak w przypadku mediany.

DECYLE to takie wartości cechy, które dzielą uporządkowany ciąg obserwacji na dziesięć liczebnie równych części, po 10% jednostek zbiorowości każda. Podział ten daje dziewięć decyli. Podział decylowy wykorzystywany jest m.in. w badaniach poziomu życia ludności i w analizie wynagrodzeń.

CENTYLE (inaczej: percentyle) to wartości cechy uzyskiwane przy podziale zbiorowości na sto liczebnie równych części, po 1% jednostek zbiorowości każda. Podział taki daje 99 centyli. Podział centylowy stosowany jest przy analizie wielu kwestii społecznych, ale jego najbardziej znanym zastosowaniem są siatki centylowe obrazujące wysokość ciała, masę ciała i obwód głowy dzieci (oddzielnie chłopców i dziewcząt).

The figure consists of two line graphs showing growth curves for children. Both graphs plot weight gain in kilograms (Masa ciała w kilogramach) on the y-axis against time in months (Miesiące życia) on the x-axis.

Top Graph: The y-axis ranges from 3 to 12 kg, and the x-axis ranges from 1 to 12 months. It displays nine curves representing different centiles: 3, 10, 25, 50, 75, and 97. The curves show a rapid initial weight gain that gradually levels off over the 12-month period.

Bottom Graph: The y-axis ranges from 10 to 50 kg, and the x-axis ranges from 1 to 18 months. It displays nine curves representing different centiles: 3, 10, 25, 50, 75, and 97. The curves show a steady, nearly linear increase in weight gain over the 18-month period.