Iwona Mróz,

Instytut Fizyki Doświadczalnej,

Uniwersytet Wrocławski

**„Podstawy statystyki i analizy danych” – materiały do wykładu nr 2 z dnia 15.10.2021**

Niniejsze materiały mają charakter roboczy. Bardzo proszę o zgłaszanie zauważonych błędów, braków, niedociągnięć i niejasności. Prośba dotyczy też przypisów. Z góry dziękuję za pomoc☺.

**Opis zbiorowości statystycznej (statystyka opisowa) – podstawy**

„*Szeregiem statystycznym* nazywamy ciąg wielkości statystycznych pogrupowany według określonych kryteriów.”[[1]](#footnote-1)

Najbardziej podstawowym szeregiem statystycznym jest tzw. szereg szczegółowy, którego tworzenie polega na uporządkowaniu wartości obserwowanych cech statystycznych rosnąco lub malejąco.[[2]](#footnote-2) Zauważmy, że nie zawsze możemy uporządkować zebrane dane, nie jest to możliwe gdy cecha statystyczna jest wyrażona w skali nominalnej. Kiedyś ze szeregów szczegółowych korzystano opisując niewielkie zbiorowości, obecnie są one wykorzystywane znacznie częściej. O tzw. szeregach rozdzielczych opowiem Państwu na kolejnym wykładzie.

Opis zbiorowości statystycznej, którego celem jest wspomniane powyżej poznanie jej struktury, rozpoczynamy najczęściej od poznania skali pomiarowej badanej cechy statystycznej, graficznej prezentacji danych oraz, jeśli to możliwe, wyznaczenia pewnych miar ilościowych charakteryzujących zbiorowość. Z powodów, które będziemy stopniowo poznawać w przyszłości, wyznaczamy tzw. miary położenia (tendencji centralnej), miary zmienności (rozproszenia, dyspersji), miary asymetrii oraz miary koncentracji.[[3]](#footnote-3) Pamiętajmy, że wspomniane miary mają nam dostarczyć informacji o zbiorowości. Jeśli jakaś miara tych informacji nie dostarcza, czyli np. trudno jest ją zinterpretować, poszukajmy innej, lepszej miary. Informacje podane poniżej przytaczam za podręcznikami S. Ostasiewicz i wsp.[[4]](#footnote-4) oraz A. Stanisza (tom 1)

UWAGA! W poniższych rozważaniach nie różnicujemy zbiorowości na zbiorowość generalną i próbę. Opisujemy po prostu zbiorowość!

1. **Miary tendencji centralnej**

Mówiąc obrazowo, miary tendencji centralnej maja dostarczyć nam informacji na temat „środka” zbiorowości. Do miar położenia zaliczamy: miary klasyczne - średnie (arytmetyczną, geometryczną, harmoniczną) oraz miary pozycyjne - modalną (wartość najczęstszą) oraz kwantyle (najczęściej podaje się tzw. kwartyl pierwszy Q1 nazywany też dolnym, medianę Me, czyli kwartyl drugi, kwartyl trzeci Q3 nazywany górnym) oraz decyle. W niektórych zastosowaniach stosujemy centyle.

Dla danych ilościowych możemy wyliczyć klasyczne miary zmienności: średnią arytmetyczną (ang. mean), ważoną lub nieważoną, średnią harmoniczną i średnią geometryczną. Dla szeregu szczegółowego średnia arytmetyczna nieważona jest obliczana według powszechnie znanego wzoru:

(1.1)

gdzie *n* jest liczebnością zbiorowości, a oznacza wartości cechy statystycznej otrzymaną dla *i*-tej jednostki statystycznej.

Jeżeli rozpatrywana zbiorowość jest zbiorowością generalną, w literaturze jako oznaczenie średniej arytmetycznej często spotkamy *m* lub µ.

Pamiętajmy, że obliczyć średnią arytmetyczną dla szeregu szczegółowego jest łatwo, jednak jej interpretacja może być kłopotliwa. Sytuacje, w których podawanie średniej arytmetycznej jako miary tendencji centralnej w zbiorowości nie jest najlepszym pomysłem będziemy poznawać stopniowo. O pozostałych średnich porozmawiamy na drugim lub trzecim wykładzie.

Niejednokrotnie lepszą miarą tendencji centralnej niż średnia arytmetyczna jest mediana (ang. median) (Me), zaliczana do miar pozycyjnych. Mediana, nazywana też drugim kwartylem, dzieli badaną zbiorowość na dwie równe części. Połowa jednostek statystycznych wykazuje wartości cechy mniejsze lub równe medianie, a druga połowa – większe lub równe medianie. Medianę wyznaczamy następująco:

Jeżeli *n* (liczebność zbiorowości) jest nieparzyste:

(1.2a)

Jeżeli *n* jest parzyste:

(1.2b)

Mediana należy do tzw. kwantyli. Oprócz niej, najczęściej wyznacza się pierwszy kwartyl (Q1), dla którego 25% zbiorowości wykazuje wartości cechy mniejsze lub równe wartości Q1 oraz trzeci kwartyl (Q3), dla którego 25% zbiorowości wykazuje wartości cechy większe lub równe wartości Q3. W podobny sposób wyznaczamy decyle (dzieląc populację na grupy co 10%) oraz centyle (co 1%).

Inną ważną miarą tendencji centralnej jest modalna Mo (ang. mode), która ma sens wartości najczęściej występującej w zbiorowości.

Pomiędzy średnią, medianą i modalną istniej ważne zależności😊.

1. **Miary zmienności (rozproszenia, dyspersji)**

Miary zmienności pokazują nam rozrzut wartości badanej cechy Statystycznej. W ujęciu klasycznym są to to m.in: rozstęp, wariancja i odchylenie standardowe, przy czym dwie ostatnie używane są gdy posługujemy się klasycznymi miarami położenia.

Rozstęp (ang. range) pokazuje nam różnicę pomiędzy największą i najmniejszą wartością badanej cechy statystycznej w zbiorowości:

(1.3)

Wariancję dla zbiorowości (po prostu dla zbiorowości, w tym dla zbiorowości generalnej) obliczamy jako:

(1.4)

W rozpatrywanym przypadku jest to średnia arytmetyczna kwadratów odchyleń wyników od średniej (obliczonej według wzoru (1)) w zbiorowości. W literaturze wariancję wyliczoną dla zbiorowości generalnej oznacza się często jako . Ponadto, jak zobaczymy później, w zależności od specyfiki rozważanego problemu statystycznego, wyrażające średnią w populacji generalnej, może być oznaczane jako µ lub *m.*

Odchylenie standardowe to pierwiastek z wariancji:

(1.5)

Dla odchylenia standardowego w populacji generalnej spotkamy oznaczenie σ.

Klasyczne miary zmienności mogą sprawiać trudności interpretacyjne, tak jak średnia arytmetyczna. Porozmawiamy o nich w przyszłości.

Jeżeli stosujemy miary pozycyjne, to jako miarę zmienności przyjmujemy rozstęp międzykwartylowy (można też spotkać nazwę odstęp międzykwartlowy) IQR = Q3-Q1 (ang. interquartile range) lub odchylenie ćwiartkowe zdefiniowane jako:

(1.6)

gdzie Q1 i Q3 oznaczają odpowiednio pierwszy i trzeci kwartyl.

1. **Miary asymetrii**

Miary asymetrii pozwalają oszacować tzw. skośność (ang. skewness) rozkładu wartości badanej cechy statystycznej. Mówiąc obrazowo, rozkład jest skośny jeśli z jednej strony posiada „ogon”☺ inny niż z drugiej strony. Znany Państwu rozkład normalny ma dwa jednakowe „ogony”. Rozkładem skośnym jest np. rozkład log-normalny, któremu podlehają płace. Omówimy go dokładniej w przyszłości.

Wstępnej oceny asymetrii rozkładu można dokonać porównując średnią arytmetyczną z medianą i modalną, co zrobimy na następnym wykładzie. Można tez wyliczyć tzw. współczynniki asymetrii, klasyczne i pozycyjne. Jeżeli współczynnik asymetrii jest równy zero – rozkład jest symetryczny. Jeśli wartość współczynnika jest większa od zera, rozkład jest prawoskośny, a jeżeli mniejsza od zera – rozkład jest lewoskośny.

Klasyczny współczynnik skośności obliczamy według wzoru:

(1.7a)

gdzie

(1.7b)

Inny współczynnik skośności ma postać:

(1.8)

gdzie oznacza średnią arytmetyczną, Mo - modalną, a *s* - odchylenie standardowe.

Jeżeli korzystamy wyłącznie z parametrów pozycyjnych, współczynnik skośności wyliczymy jako:

(1.9),

oznaczenia jak powyżej.

1. **Miary koncentracji (skupienia)**

Miary koncentracji (skupienia) opisują koncentrację wartości danej cechy wokół średniej. W przypadku najpopularniejszej miary koncentracji, tzw. kurtozy (ang. kurtosis), Kurtoza obliczona dla wartości cech z badanej zbiorowości jest porównywana do kurtozy rozkładu normalnego. Jak zobaczymy w przyszłości, z teorii wynika, że wartość kurtozy dla rozkładu normalnego wynosi 3. Jeżeli wartość kurtozy obliczona dla badanej zbiorowości jest większa od 3, rozkład cechy w zbiorowości jest bardziej wysmukły niż rozkład normalny. Jeżeli obliczona wartość kurt ozy jest mniejsza od 3, rozkład badanej cechy jest bardziej spłaszczony niż rozkład normalny. Kurtozę obliczamy według wzoru:

(1.10a),

gdzie (1.10b).

Jak podają Ostasiewicz, Rusnak i Siedlecka[[5]](#footnote-5), współczynnik koncentracji jest także określany jako:



Wówczas jego wartość dla rozkładu normalnego wynosi zero. Wartości dodatnie K’ wskazują na rozkład cechy bardziej wysmukły niż rozkład normalny, a ujemne wartości K’ świadczą o rozkładzie bardziej spłaszczonym niż rozkład normalny. W literaturze K’ nazywa się też współczynnikiem ekscesu (ekscesem).

Proszę zauważyć, że wariancja, klasyczny współczynnik skośności i kurtoza są wyliczane z wykorzystaniem kolejnych potęg odchyleń wartości cechy od średniej (odpowiednio: drugiej, trzeciej i czwartej). Jak zobaczymy na jednym z następnych wykładów, fakt ten ma głębokie uzasadnienie teoretyczne i wiąże się z pojęciem momentów.

**Typowy obszar zmienności cechy**

W przypadku parametrów klasycznych typowy obszar zmienności cechy wynosi:

, gdzie jest wartością średnią, a *s* – odchyleniem standardowym obliczonymi dla badanej zbiorowości.[[6]](#footnote-6)

Dla parametrów pozycyjnych przyjmujemy:

, gdzie e oznacza medianę, a Q – odchylenie ćwiartkowe.

W typowym obszarze zmienności wartości cechy mieszczą się wartości cechy około 2/3 jednostek badanej zbiorowości.[[7]](#footnote-7)

**Ilustracja graficzna wartości badanej cechy statystycznej – wykres ramka-wąsy**

Bardzo użyteczna jest ilustracja graficzna rozkładu badanych statystyk przy użyciu tzw, **wykresu ramka-wąsy**.[[8]](#footnote-8) W centralnym punkcie wykresu leży średnia lub mediana, granice ramki oznaczają odpowiednio odchylenie standardowe lub pierwszy i trzeci kwantyl, wąsy mogą określać rozstęp lub inny zakres wartości, o którym piszemy poniżej. Wykres ten, szczególnie przy zastosowaniu miar pozycyjnych, pozwala w łatwy sposób zauważyć asymetrię rozkładu. Z tego względu parametry pozycyjne są częściej wykorzystywane do sporządzania wykresu dla danych, z którymi dopiero zaczynamy się zaznajamiać.

Wykres ramka-wąsy pozwala też łatwo zidentyfikować tzw. obserwacje nietypowe (odstające i ekstremalne). Obserwacje takie należy zawsze starannie przeanalizować od strony merytorycznej, gdyż mogą one znacznie zaburzyć wyniki analizy. W tym miejscu podkreślmy bardzo mocno, iż nie można usuwać obserwacji nietypowych z analizy w sposób automatyczny.

Do identyfikacji obserwacji nietypowych (odstających i ekstremalnych) często wykorzystuje się kryterium Tukeya[[9]](#footnote-9). Według tego kryterium, za obserwacje nietypowe (odstające, ang, outliers) uważa się te wartości badanej cechy (oznaczone jako *x*), które spełniają warunek:

 lub 

Wartości, dla których zachodzi:

 lub 

są traktowane jako obserwacje ekstremalne.

Identyfikacja obserwacji nietypowych (odstających i ekstremalnych) jest konieczna ze względu na poprawność prowadzonej analizy.



Rys. 2.1 Wykres ramka-wąsy*. Źródło: opracowanie własne*

Wąsy mogą rozciągać się od wartości minimalnej do wartości maksymalnej badanej cechy, a zatem obejmować cały zakres wartości. Mogą też obejmować zakres od najmniejszej wartości uznanej za nieodstającą, do największej wartości uznanej za nieodstającą. Obserwacje odstające, jeżeli istnieją, są na wykresie oznaczane punktami. W przykładzie pokazanym na Rys. 2.1 nie ma obserwacji odstających. Wąsy obejmują zakres obserwacji nieodstających.

1. Ostasiewicz, S., …,op. cit. s. 16. [↑](#footnote-ref-1)
2. Ibidem, s. 16. [↑](#footnote-ref-2)
3. Ibidem, s. 46 oraz Stanisz, A., op. cit, s. 115 . [↑](#footnote-ref-3)
4. Ostasiewicz, S.,…, op. cit., s. 47-74. [↑](#footnote-ref-4)
5. Ibidem, s. 74. [↑](#footnote-ref-5)
6. Ibidem, s. 62. [↑](#footnote-ref-6)
7. Ibidem, s. 67. [↑](#footnote-ref-7)
8. Stanisz, A., op. cit., s. 240. [↑](#footnote-ref-8)
9. Ibidem, s. 240. [↑](#footnote-ref-9)