Wprowadzenie

Statystyka opisowa to dział statystyki zajmujący się metodami opisu danych statystycznych (np. środowiskowych) uzyskanych podczas **badania statystycznego (**np. badań terenowych, laboratoryjnych**)**.
Badanie statystyczne to proces pozyskiwania danych na temat rozkładu cechy statystycznej w populacji.
Populacja statystyczna – zbiór elementów, podlegających badaniu statystycznemu. Elementy populacji są do siebie podobne pod względem badanej cechy, ale nie są identyczne.
Przykład: Wszyscy studenci na Wydziale Geologii posiadają cechę wzrostu - są pod tym względem podobni, ale nie identyczni: są studenci wysocy i niscy. Populacją w badaniu statystycznym wzrostu studentów Geologii będą wszyscy studenci na Wydziale Geologii.
Próba losowa – wybrane losowo elementy populacji. Np. my na zajęciach.
Nie wszystkie populacje muszą istnieć w rzeczywistości, niektóre z nich mają charakter wyłącznie hipotetyczny.
Elementy populacji statystycznej nazywamy jednostkami statystycznymi, zaś badana cecha to cecha statystyczna.
Ze względu na liczebność zbioru, populacje można podzielić na:
populacje skończone - np. populacja studentów Geologii
populacje nieskończone - np. czas
Ponieważ często badanie statystyczne **całej populacji** jest nieuzasadnione lub niemożliwe (przyczyny: patrz badanie statystyczne), dlatego zwykle bada się jedynie **wybrane losowo elementy populacji, czyli próbę losową**, a następnie wnioskuje na podstawie obserwacji cechy w próbie o możliwych wartościach cechy w populacji. Dlatego właśnie niektóre pojęcia statystyczne mogą odnosić się zarówno do populacji, jak i do próby (są to tzw. wielkości empiryczne).
Estymacja - szacowanie wartości nieznanych parametrów rozkładu na podstawie znanych wyników próby (może być stosowane jedynie w przypadku regularnych rozkładów parametrów cechy w populacji).
Badanie statystyczne może mieć charakter:
pełny - badanie obejmuje całą populację
częściowy - odbywa się na pewnych (zazwyczaj losowo) wybranych elementach populacji, czyli próbie losowej, zazwyczaj reprezentatywnej dla populacji
W ramach badania statystycznego zbierane są wartości określonej **cechy statystycznej** nazywane **wartościami zaobserwowanymi cechy statystycznej lub danymi statystycznymi**. Zróżnicowanie wartości cechy statystycznej powoduje, że można mówić o jej **rozkładzie w populacji** (próbie losowej).
Próba reprezentatywna – część populacji, wybrana do badania metodami statystycznymi, w założeniu badacza, zachowująca strukturę wyróżnionych cech populacji przy założonym **poziomie istotności**.
Poziom istotności – α – określa maksymalne ryzyko błędu, jakie badacz jest skłonny zaakceptować. Wybór wartości α zależy od badacza, natury problemu i od tego jak dokładnie chce on weryfikować swoje hipotezy, najczęściej przyjmuje się $\alpha = 0{,}05$ lub 0,01 czyli obliczamy coś z 95 lub 99% prawdopodobieństwem.
Celem stosowania **metod statystyki opisowej** jest podsumowanie zbioru danych (np. **próby losowej**) i wyciągnięcie pewnych podstawowych wniosków i uogólnień na temat zbioru.
Statystykę opisową stosuje się zazwyczaj jako pierwszy i podstawowy krok w analizie zebranych danych.

Do technik statystyki opisowej można zaliczyć:

1. Opis tabelaryczny.

Dane przedstawiane są w postaci tabel. Dla małych zbiorów danych tabele mogą prezentować wszystkie dane, w przeciwnym przypadku tworzy się różnego rodzaju podsumowania, jak np. szereg rozdzielczy.
Szereg rozdzielczy – uzyskuje się go dzieląc dane statystyczne na pewne kategorie i podając liczebność lub częstość zbiorów danych przypadających na każdą z tych kategorii.
Przykład: Licząc punkty zdobyte przez studentów na kolokwium uzyskałem następujące wyniki:
0, 0, 1, 1, 0, 1, 2, 4, 1, 0, 2, 1, 0, 1, 2, 1, 2, 2, 1, 5

Wartości cechy	Liczebność	Częstość
0	5	0.25
1	8	0.40
2	5	0.25
4	1	0.05
5	1	0.05

2. Graficzna prezentacja wyników.

Dane prezentowane są w formie graficznej. Podstawowymi narzędziami są tutaj: histogram i krzywa liczebności, które wykreślane są bezpośrednio na podstawie danych z szeregu rozdzielczego.
Histogram to jeden z graficznych sposobów przedstawiania rozkładu cechy. Składa się z szeregu prostokątów umieszczonych na osi współrzędnych. Prostokąty te są z jednej strony wyznaczone przez przedziały klasowe (patrz: Szereg rozdzielczy) wartości cechy, natomiast ich wysokość jest określona przez liczebności (lub częstości)
Liczba przedziałów powinna wynosić od 5 do 15, w przeciwnym wypadku przestaje być on czytelny. Szerokości przedziałów histogramu powinny być równe.

Przykładowa interpretacja histogramu:
Luka w histogramie
podejrzenie nieprawidłowego odczytu (brak danych)
podejrzenie błędu urządzenia pomiarowego
Histogram z dwoma wierzchołkami
Tzw. **rozkład dwumodalny** (bimodalny). Powstaje często, gdy badana populacja jest połączeniem dwóch odrębnych populacji,

Zasady tworzenia przedziałów:
Liczba klas powinna być równa (w przybliżeniu) **pierwiastkowi z liczebności próby**.

$$k \approx \sqrt{n}$$

Liczba obserwacji n	Liczba zalecanych klas k
40-60	6-8
60-100	7-10
100-200	9-12
200-500	11-17

UWAGA:
W klasie nie powinno być mniej niż 5% wszystkich obserwacji.

Rozpiętość przedziału (równe przedziały):

$$h \approx \frac{x_{\max} - x_{\min}}{k} = \frac{R}{k}$$

3. Wyznaczanie miar rozkładu.

Do opisu służą **miary rozkładu** - różnego rodzaju wielkości obliczane na podstawie uzyskanych danych. Interpretacja wartości tych miar dostarcza informacji na temat charakteru rozkładu cechy.

Miary można podzielić na kilka podstawowych kategorii:

- miary położenia, np. kwantyl
Miara położenia rozkładu to taka miara rozkładu, która określa relację między dwoma identycznymi rozkładami, ale przesuniętymi względem osi odciętych układu współrzędnych.

Kwantyle rzędu 1/4, 1/2, 3/4 są inaczej nazywane **kwartylami.**
Kwantyle rzędu 1/5, 2/5, 3/5, 4/5 to inaczej kwintyle.
Kwantyle rzędu 1/10, 2/10,..., 9/10 to inaczej decyle.
Kwantyle rzędu 1/100, 2/100,..., 99/100 to inaczej **percentyle.**

Kwartyl pierwszy Q1 (dolny) dzieli uporządkowaną niemalejąco zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości zmiennej mniejsze lub równe kwartylowi pierwszemu Q1, a 75% równe lub większe od tego kwartyla.
Kwartyl drugi Q2 (mediana, wartość środkowa) dzieli uporządkowaną niemalejąco zbiorowość na dwie części w ten sposób, że połowa jednostek zbiorowości ma wartości zmiennej równe lub większe od mediany, stąd też mediana bywa nazywana wartością środkową.
Kwartyl trzecie Q3 (górny) dzieli uporządkowaną niemalejąco zbiorowość na dwie części w ten sposób, że 75% jednostek zbiorowości ma wartości zmiennej mniejsze lub równe kwartylowi trzeciemu Q3, a 25% równe lub większe od tego kwartyla.

Gdy liczba obserwacji jest nieparzysta, wówczas medianą jest wartość środkowa. Jeżeli liczebność zbiorowości jest liczba parzystą, przyjmuje się, że mediana jest średnią arytmetyczną dwóch środkowych wartości zmiennej.

Kwartyl pierwszy i trzeci z szeregu szczegółowego wyznacza się w sposób analogiczny jak medianę. Zbiorowość dzieli się na dwie rozłączne części: pierwszą, której jednostki przyjmują wartości nie większe od mediany i drugą, złożoną z pozostałych jednostek. Dla każdej z tych części można wyznaczyć ponownie medianę według zamieszczonego wyżej wzoru.
Dla pierwszej części wartość mediany będzie odpowiadała kwartylowi pierwszemu (Q1), a dla drugiej – kwartylowi trzeciemu (Q3).

w tym miary tendencji centralnej

Miara tendencji centralnej rozkładu - taka miara rozkładu, która określa położenie wartości centralnych rozkładu (wartości przeciętnych, średnich). Istnieje wiele definicji co tak naprawdę określić jako wartości przeciętne i każda z tych definicji to dana miara tendencji centralnej.

np. średnia arytmetyczna, średnia geometryczna, średnia harmoniczna, średnia kwadratowa, mediana, moda

Średnia ważona niepustej listy danych

$$[x_1, x_2, ..., x_n],$$

z odnoszącymi się do nich nieujemnymi wagami

$$[w_1, w_2, ..., w_n],$$

z których co najmniej jedna jest dodatnia, jest określona przez:

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

co oznacza:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}.$$

W ten sposób dane którym przypisano większe wagi mają większy udział w określeniu średniej ważonej niż dane, którym przypisano mniejsze wagi.

Jeśli wszystkie wagi są równe, wówczas średnia ważona jest równa średniej arytmetycznej.

Średnią geometryczną $n$dodatnich liczb $a_1, a_2, ..., a_n$nazywamy liczbę

$$\sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n}.$$

Na przykład średnią geometryczną liczb 2, 2, 5 i 7 jest

$$\sqrt[4]{2 \cdot 2 \cdot 5 \cdot 7} \approx 3.44.$$

Średnia ta jest stosowana, gdy zmienna ma rozkład logarytmiczno-normalny.

- miary zróżnicowania

np. **odchylenie standardowe**, wariancja, rozstęp, rozstęp ćwiartkowy, średnie odchylenie bezwzględne, odchylenie ćwiartkowe**, współczynnik zmienności**

Odchylenie standardowe – klasyczna miara zmienności, obok średniej arytmetycznej najczęściej stosowane pojęcie statystyczne.

Intuicyjnie rzecz ujmując, odchylenie standardowe mówi, **jak szeroko wartości jakiejś wielkości są rozrzucone wokół jej średniej**. Im mniejsza wartość odchylenia tym obserwacje są bardziej skupione wokół średniej.

Współczynnik zmienności to klasyczna miara zróżnicowania rozkładu cechy. W odróżnieniu od odchylenia standardowego, które określa bezwzględne zróżnicowanie cechy, współczynnik zmienności jest miarą *względną*, czyli zależną od wielkości średniej arytmetycznej. Definiowany jest wzorem:

$$V = \frac{s}{\overline{x}}, \quad \overline{x} \neq 0$$

gdzie

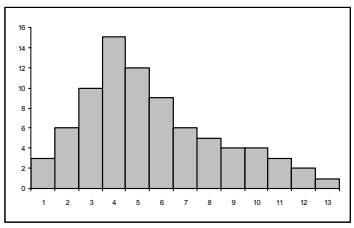
s to odchylenie standardowe z próby,
$\overline{x}$ to średnia arytmetyczna z próby.

- miary asymetrii

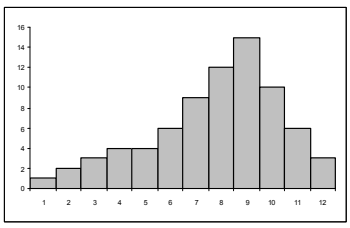
np. współczynnik skośności,

Współczynnik skośności przyjmuje wartość zero dla rozkładu symetrycznego, **wartości ujemne dla rozkładów o lewostronnej asymetrii** (wydłużone lewe ramię rozkładu) i wartości **dodatnie dla rozkładów o prawostronnej asymetrii** (wydłużone prawe ramię rozkładu). Dla rozkładu normalnego = 0.

\+



\-

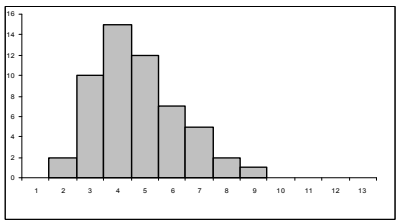


- miary koncentracji

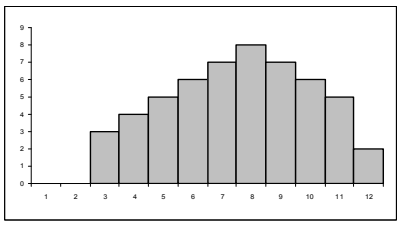
np. współczynnik Giniego, Kurtoza g2

Kurtoza to jedna z miar koncentracji rozkładu wartości cechy. Miara **kształtu rozkładu (wysmukły (szpiczasty), spłaszczony).**

\+



\-



- kurtoza rozkładu normalnego wynosi 0

Rozkłady prawdopodobieństwa można podzielić ze względu na wartość kurtozy na rozkłady:

- **mezokurtyczne** - wartość kurtozy wynosi 0, spłaszczenie rozkładu jest podobne do spłaszczenia rozkładu normalnego (dla którego kurtoza wynosi dokładnie 0)
- **leptokurtyczne** - kurtoza jest dodatnia, wartości cechy bardziej skoncentrowane niż przy rozkładzie normalnym
- **platykurtyczne** - kurtoza jest ujemna, wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym

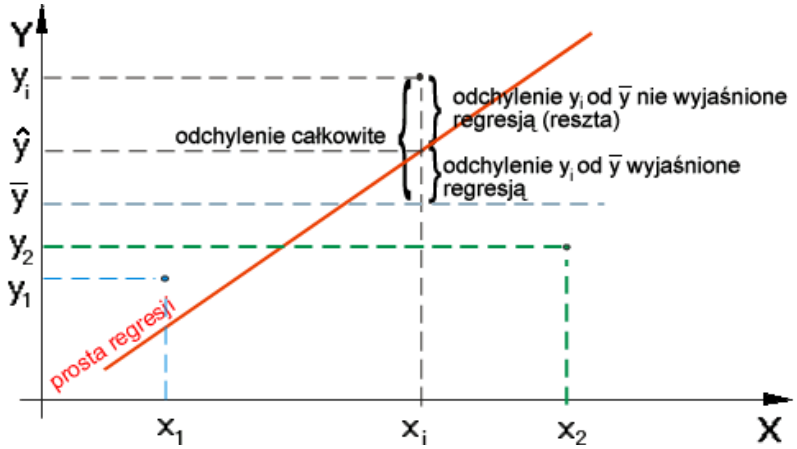
Do określenia zależności pomiędzy dwoma cechami można zastosować **wykres rozrzutu** (rozproszenia).
Trzeba ustalić charakter zależności:

1. czy zależność jest liniowa?
2. dodatnia czy ujemna?
3. silna czy słaba?

Siłę zależności wyznaczyć możemy za pomocą **współczynnika determinacji R^2 (0-1)**. Informuje on o tym, jaka część zmienności całkowitej zmiennej losowej Y została wyjaśniona regresją liniową względem X. On zawsze dodatni – nie mówi o kierunku zależności.

Nieznane parametry modelu
$y = ax + b$
muszą być estymowane na podstawie odpowiedniej próby losowej.
Zagadnienie estymacji parametrów modelu sprowadza się do takiego dobrania parametrów aby suma kwadratów odległości każdego punktu empirycznego od prostej regresji była jak najmniejsza – **metoda najmniejszych kwadratów.**

Współczynnik determinacji (dopasowania)



Jeżeli X będzie zmienną niezależną (objaśniającą), a Y zmienną zależną (objaśnianą), powiązaną z X zależnością korelacyjną, to odchylenie całkowite (patrz rys powyżej) punktu o wartości zmiennnej $Y=y_i$ od wartości średniej (y) można przedstawić następująco:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

W analogiczny sposób, podnosząc do kwadratu obie strony równości i sumując po i = 1, 2,..., n, całkowitą zmienność wszystkich wartości yi możemy określić jako sumę kwadratów. Podniesienie do kwadratu jest konieczne ponmieważ część wartości Y=yi odchyla się od wartości średniej in plus, a część in minus. Tak więc dostaniemy:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Równość ta wyraża podział całkowitej sumy kwadratów odchyleń dla zmiennej y na dwa składniki:
- sumę kwadratów odchyleń wyjaśnioną efektem regresji (EFEKT),
- resztową sumę kwadratów odchyleń (nie wyjaśnioną regresją) (RESZA).

Dla uproszczenia zapisu równanie powyższe często jest przedstawiane w postaci:

$SS_T = SS_E + SS_R$

gdzie:
SS_T - zmienność całkowita zmiennej zależnej, ***całkowita suma kwadratów***
SS_E - część zmienności wyjaśniona modelem regresji,
SS_R - zmienność przypadkowa (losowa) - suma odchyleń wartości yi od prostej regresji.

SSE (Sum of Squares of Errors) – suma kwadratów błędów

Współczynnikiem dopasowania nazywamy:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 jest częścią zmienności, wyjaśniona przez regresję. Zazwyczaj wyraża się w procentach. Im większe R^2, tym lepiej (estymowana) prosta regresji „pasuje” do punktów doświadczalnych.

Informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model.

Współczynnik determinacji jest opisową miarą dopasowania modelu regresji do danych, czyli **miarą siły liniowego związku między danymi**. Mierzy on część zmienności zmiennej objaśnianej Y, która została wyjaśniona liniowym oddziaływaniem zmiennej objaśniającej X