1 Różne pojęcia wstępne

Statystyka: analiza struktury, przedziały ufności i weryfikacja hipotez, analiza współzależności.

Etapy analizy statystycznej: – przełóż obserwacje na postać liczbową – wnioskuj (zastosuj odpowiednie statystyki)

Opis statystyczny – liczbowe przedstawienie badanych zbiorowości lub zjawisk w postaci opisu: – tabelarycznego; – graficznego; – parametrycznego

Opis statystyczny może dotyczyć: – struktury zbiorowości; – współzależności; – zmian zjawisk w czasie.

Badanie statystyczne to zespół czynności zmierzających do uzyskania (za pomocą metod statystycznych) informacji charakteryzujących badaną zbiorowość lub zjawisko. Najważniejsze kryteria klasyfikacji badań: – zakres obserwacji badanych jednostek (pełne, częściowe); – częstotliwość: (ciągłe, okresowe, doraźne); – zasięg przestrzenny (międzynarodowe, krajowe, regionalne, środowiskowe, monograficzne); – dziedzina badań (demograficzne, społeczne, ekonomiczne, rolnicze, jakości środowiska naturalnego itp.

Populacja, zbiorowość statystyczna: zbiór obiektów (osób, przedmiotów, zdarzeń) logicznie ze sobą powiązanych (ale nie identycznych), poddany badaniu statystycznemu.

Jednostka statystyczna: jednostki statystyczne w danej populacji różnią się od innych jednostek spoza danej populacji poprzez swoje własności wspólne (cechy stałe), jednocześnie różnią się między sobą cechami (cechy zmienne), które są przedmiotem zainteresowania badacza.

Cechy statystyczne – właściwości jednostek statystycznych Cechy stałe

– jednakowe dla wszystkich jednostek badania: rzeczowa (co? kto? jest badane/y) przestrzenna (gdzie?) czasowa (kiedy?)

Cechy zmienne – różnicujące jednostki, będące przedmiotem zainteresowania: – jakościowe – nominalne lub porządkowe – dwudzielne lub wielodzielne – ilościowe – skokowe lub ciągłe

Cecha statystyczna mierzalna (ilościowa) – określana jest za pomocą liczb np. oceny, płace. Cechy mierzalne dzielą się na skokowe i ciągłe. Skokowe są to cechy, które przyjmują skończoną liczbę wartości, zwykle są to liczby całkowite; Ciągłe są to cechy, które przyjmują dowolne wartości liczbowe z pewnego przedziału liczbowego np. dochody, długość ziarna fasoli. Cecha porządkowa.

Rodzaje skal pomiarowych – nominalna (nominal scale), klasyfikuje: płeć; – porządkowa (ordinal scale), klasyfikuje i porządkuje: zdolność kredytowa firmy, stadia choroby, – Przedziałowa (interval scale), posiada jeszcze stałą jednostkę miary i umowne zero: temperatura – Ilorazowa (rational scale), klasyfikuje, porządkuje od zera: wiek, wzrost, obrót

Pytanie: oceny w szkole to jaka skala?

Szereg rozdzielczy (punktowy, przedziałowy) – jest to prosta tablica statystyczna złożona z dwóch kolumn lub z dwóch wierszy. W pierwszej kolumnie (wierszu) wypisujemy wartości badanej cechy, a w drugiej liczby jednostek, które mają daną cechę.

Szereg prosty powstaje poprzez uporządkowanie notowań według rosnących lub malejących poziomów cech.

2 Opracowanie danych

Klasyfikacja to ustalenie (wyodrębnienie) wariantów cechy.

Grupowanie – podział zbiorowości na jednorodne lub względnie jednorodne podgrupy z punktu widzenia wyróżnionej cechy (cech): – grupowanie typologiczne (cechy jakościowe); – grupowanie wariancyjne (cechy ilościowe).

Zasady logiki formalnej: grupowanie musi być wyczerpujące – każda jednostka zbiorowości musi być sklasyfikowana i włączona do odpowiedniej podgrupy; – grupowanie powinno być rozłączne – wyodrębnione podgrupy muszą się wzajemnie wykluczać; – grupowanie powinno być efektywne – wyróżnione podgrupy powinny być na tyle jednorodne jakościowo, by mogły stanowić podstawę twierdzeń uogólniających

Szeregiem statystycznym nazywamy materiał statystyczny uporządkowany lub uporządkowany i pogrupowany według określonych kryteriów (przyjętych wariantów cechy).

Szereg strukturalny:

Tablica 1. Struktura próby mieszkańców wg wykształcenia

Wykształcenie	liczba osób	odsetek w %
podstawowe i gimnazjalne	130	13,0
zawodowe	272	27,2
średnie	444	44,5
wyższe	153	15,3
Ogółem	999	100,0

Szereg rozdzielczy punktowy:

Struktura gospodarstw domowych wg liczby samochodów

Liczba samochodów	Liczba gospodarstw

0	230
1	280
2	70
3 i więcej	5
Razem	585

Szereg rozdzielczy przedziałowy:

Studenci według czasu wolnego

Czas wolny w min.	Liczba osób
30,1 - 60	3
60,1 - 90	4
90,1 - 120	6
120,1 - 150	5
150,1 - 180	3
180,1 - 210	1
Razem	22

Szereg kumulacyjny, Szereg czasowy, Szereg przestrzenny (geograficzny).

Budowa tablic statystycznych: 1. Część liczbowa (kolumny i wiersze); 2. Część opisowa: – tytuł tablicy; – boczek (nazwy wierszy); – główka (nazwy kolumn); – źródło danych; – ewentualne uwagi odnoszące się do danych liczb.

Wykresy statystyczne są graficzną formą prezentacji materiału statystycznego, są mniej precyzyjne i szczegółowe niż tablice, natomiast bardziej sugestywne.

Rodzaje wykresów: – punktowe (szereg szczegółowy, rozdzielczy punktowy, diagram korelacyjny); – obrazkowe (szereg strukturalny, wykresy popularyzatorskie); – powierzchniowe (prostokąty, kwadraty

i koła) (rozdzielcze strukturalne, rozdzielcze przedziałowe (histogram), czasowe i przestrzenne); – liniowe (szeregi czasowe, rozdzielcze przedziałowe (krzywa liczebności, wielobok liczebności), rozdzielcze punktowe, funkcje regresji); – mapowe (szeregi geograficzne (kartogram lub kartodiagram)); – złożone

3 Analiza struktury

Analiza struktury: badanie budowy wewnętrznej zbiorowości ze względu na obserwowane w badaniu cechy zmienne.

Podstawę do oceny struktury zbiorowości stanowią dane w postaci szeregu szczegółowego, bądź też pogrupowane

Analizę prowadzić można na podstawie wykresów, szeregów rozdzielczych oraz (najczęściej) za pomocą odpowiednio obliczonych charakterystyk, zwanych parametrami (dla populacji) lub statystykami (dla próby).

Tylko szereg rozdzielczy pokaże bezpośrednio rozkład cechy, czyli rozłożenie jednostek zbiorowości do poszczególnych wariantów badanej cechy.

Rozkład cechy: przyporządkowanie liczby wystąpień (liczebności, częstości lub prawdopodobieństwa) odpowiednim wartościom cechy zmiennej.

Analiza struktury obejmuje: określenie tendencji centralnej (wartość przeciętba, mediana, dominanta) zróżnicowanie wartości (rozproszenie) asymetrię koncentrację

Miary położenia

Klasyczne i pozycyjne. Miary przeciętne charakteryzują średni lub typowy poziom wartości cechy. Są to więc takie wartości, wokół

których skupiają się wszystkie pozostałe wartości analizowanej cechy.

Do miar klasycznych zalicza się: średnią arytmetyczną

Miary pozycyjne: mediana, moda (dominanta), kwartyle, kwantyle, decyle

Średnia artmetyczna (*Mean*, *Arithmetic mean*): Oblicznie średniej dla szeregu prostego:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

Średnia arytmetyczna ważona. Oblicznie średniej dla szeregu rodzielczego:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} \tag{2}$$

Mediana (Median, kwartyl drugi) dzieli zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od Me. Stąd też mediana bywa nazywana wartością środkową.

Właśności mediany: – może być obliczana w tych przypadkach, w których obliczenie średniej arytmetycznej (np. szeregi o otwartych przedziałach klasowych), a także modalnej (różne rozpiętości przedziałów klasowych) jest niemożliwe; – mediana nie reaguje na zmiany cech skrajnych jednostek, czyli na tzw. obserwacje nietypowe (przypadkowe); – jeżeli rozkład danych jest symetryczny, wówczas Me = D = $\bar{x}$.

parzysta liczba jednostek w wielkość zbiorowości:

$$\mathrm{Me} = \frac{x_{n/2} + x_{n/2+1}}{2} \tag{3}$$

lub (nieparzysta liczba jednostek w wielkość zbiorowości):

$$\mathrm{Me} = x_{(n+1)/2} \tag{4}$$

lub (szereg rozdzielczy, przedziałowy):

$$\mathrm{Me} = x_0 + (\frac{N}{2} - \mathrm{cum}_{n-1})\frac{c_0}{n_0} \tag{5}$$

gdzie x_0 – dolna granica przedziału mediany n_0 – liczebność przedziału mediany cum_{n-1} – liczebność przedziału poprzedzającego przedział mediany c_0 – rozpiętość przedziału mediany.

wysokość płac	l.pracowników	
800-1000	300	300
1000-1600	2400	2700
1600-2000	1200	3900
2000-3000	2500	6400
3000 i więcej	1000	7400

Dominanta (*Mode*, Moda, wartość modalna, wartość najczęstsza) jest to wartość cechy statystycznej, która w szeregu empirycznym występuje najczęściej. W szeregach prostych i rozdzielczych jest to wartość cechy, której odpowiada największa liczebność (częstość).

szereg rozdzielczy, przedziałowy:

$$\mathrm{D} = x_0 + c_0\frac{n_0 - n_{-1}}{(n_0 - n_{-1}) + (n_0 - n_{+1})} \tag{6}$$

gdzie x_0 – dolna granica przedziału najliczniejszego n_0 – liczebność (gęstość) przedziału najliczniejszego c_0 – rozpiętość przedziału najliczniejszego

Jeżeli przedziały mają różną rozpiętość, to można posługiwać się pojęciem *gęstości*.

Kwartyle (Q, *quartile*, Q_1, Q_3), kwantyle (D, wartości dziesiętne), centyle (P, wartości setne)

$$\text{Poz}_{Qr,v} = (N+1)\frac{r}{v} \qquad \text{lub} \qquad \text{Poz}_{Qr,v} = N\frac{r}{v} \tag{7}$$

lub

$$Q_{r,v} = x_0 + (\text{Poz}_{Qr,v} - \text{cum}_{n-1})\frac{c_0}{n_0} \tag{8}$$

gdzie x_0 – dolna granica przedziału mediany n_0 – liczebność przedziału $Q_{r,v}$ cum_{n-1} – liczebność przedziału poprzedzającego przedział $Q_{r,v}$ c_0 – rozpiętość przedziału $Q_{r,v}$.

4 Miary zmienności

Wariancja, odchylenie standardowe, odchylenie przeciętne, współczynnik zmienności (Pearsona)

Wariancja (*variance*) jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej zbiorowości.

Oblicznie wariancji dla szeregu prostego:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{9}$$

często zamiast dzielenie przez N dzielimy przez $N-1$.

Oblicznie wariancji dla szeregu rodzielczego:

$$S^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \tag{10}$$

lub (prościej):

$$S^2 = \frac{1}{N}\sum_{i=1}^{k}x_i^2 n_i - \bar{x}^2 \tag{11}$$

Odchylenie standardowe (*standard deviation*, sd) jest pierwiastkiem kwadratowym z wariancji. Parametr ten określa przeciętne

zróżnicowanie poszczególnych wartości cechy od średniej arytmetycznej.

Odchylenie przeciętne (*average absolute deviation*, d):

$$d = \frac{1}{N}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{12}$$

miara bardzo rzadko używana.

Współczynniki pozycyjne. Odchylenie ćwiartkowe (Q, *midhinge*):

$$Q = \frac{Q_3 - Q_1}{2} \tag{13}$$

i rozstęp ćwiartkowy (*interquartile range*):

$$R_Q = \frac{Q_3 - Q_1}{2} \tag{14}$$

Współczynnik zmienności jest ilorazem bezwzględnej miary zmienności cechy i średniej wartości tej cechy. W analizie struktury korzysta się z różnych miar położenia i zmienności, dlatego są współczynniki zmienności klasyczne i pozycyjne.

Współczynniki klasyczne:

$$V_s = \frac{s}{\bar{x}} \qquad \text{lub} \qquad V_d = \frac{d}{\bar{x}} \tag{15}$$

pozycyjne

$$V_Q = \frac{Q_3 - Q_1}{\mathrm{Me}} \tag{16}$$

albo (*Quartile coefficient of dispersion*):

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{17}$$

Współczynnik zmienności jest wartością niemianowaną. Wartości liczbowe współczynników zmienności najczęściej są podawane

w procentach. Przyjmuje się, że jeżeli współczynnik zmienności jest poniżej 10%, to cechy wykazują zróżnicowanie statystycznie nieistotne. Duże wartości tego współczynnika świadczą o dużym zróżnicowaniu, a więc niejednorodności zbiorowości.

Współczynnik zmienności stosuje się zwykle w porównaniach, gdy chce się ocenić zróżnicowanie: kilku zbiorowości pod względem tej samej cechy, tej samej zbiorowości pod względem kilku różnych cech.

5 Momenty

Uogólniając klasyczne miary położenia i zmienności można zdefiniować następującą funkcję zwaną momentem (rzędu r):

$$M_r = \frac{1}{N} \sum_{i=1}^{N} (x_i - p)^r \tag{18}$$

jeżeli $p = 0$ to moment nazywamy zwykłym, jeżeli $p = \bar{x}$ centralnym. Momenty centralne zwykle oznacza się grecką literą μ (mju jak seria wodoszczelnych aparatów Olympusa). Momenty zwykłe są rzadziej używane (za wyjątkiem pierwszego).

Jak widać średnia to pierwszy moment zwykły a wariancja to drugi moment centralny. „Uproszczony" wzór na wariancję podany wyżej sprowadza się zatem do (m_2 to drugi moment zwykły):

$$\mu_2 = m_2 - \mu_1^2 \tag{19}$$

6 Miary asymetrii

Asymetria (*skewness*), to odwrotność symetrii. Szereg jest symetryczny jeżeli jednostki są rozłożone „równomiernie" wokół wartości średniej:

$$\bar{x} = \mathrm{Me} = D \tag{20}$$

Asymetria prawostronna, lewostronna; wskaźnik asymetrii (skośności), współczynniki asymetrii (skośności).

Moment trzeci centralny – średnia arytmetyczna z podniesionych do potęgi trzeciej odchyleń wartości cechy od średniej arytmetycznej

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3 \tag{21}$$

albo (za pomocą momentów zwykłych):

$$\mu_3 = m_3 - 3m_2\bar{x} + 2\bar{x}^3 \tag{22}$$

jeżeli $\mu_3 = 0$ szereg symetryczny, $\mu_3 > 0$ asymetria dodatnia (prawostronna), $\mu_3 < 0$ asymetria ujemna (lewostronna)

Moment trzeci względny określa siłę i kierunek asymetrii:

$$g_1 = \frac{\mu_3}{s^3} \tag{23}$$

Na podstawie badań empirycznych: $-2 < g_1 < 2$, w skrajnych przypadkach może przekraczać ten przedział.

Współczynnik asymetrii (skośności) oparty na odległościach między średnimi (K. Pearson).

$$W_s = \frac{\bar{x} - D}{s} \tag{24}$$

rzadziej używa się:

$$W_s = \frac{\bar{x} - \mathrm{Me}}{s} \tag{25}$$

Współczynnik asymetrii (skośności) oparty na odległościach między kwartylami lub decylami:

$$W_{sq} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \tag{26}$$

7 Miary koncentracji

Koncentracja: rozłożenie wartości cechy pomiędzy poszczególne jednostki populacji. Brak koncentracji: wszystkie jednostki mają różne wartości; maksymalna koncentracja: wszystkie jednostki mają tą samą wartość.

$$K = \frac{\mu_4}{s^4} - 3 \tag{27}$$

wartość K wynosi 0 dla rozkładu normalnego; jeżeli $K < 0$ rozkład jest spłaszczony, jeżeli $K > 0$ rozkład jest bardziej skoncentrowany niż rozkład normalny

Współczynnik Giniego i Krzywa Lorenza. Krzywa Lorenza jest kreślona w układzie współrzędnych XY. Oś OX reprezentuje kumulowaną liczebność (wyrażoną najczęściej jako udział w całkowitej liczebności populacji). Oś OY reprezentuje kumulowaną wartość cechy (także w %%). Każdy punkt na krzywej Lorenza reprezentuje wtedy stwierdzenie typu: 20% jednostek ma 5% łącznej wartości cechy. Np. 25% ludności posiada 8% łącznych dochodów, albo 50% rolników posiada 15% łacznych areałów, itp.

Przy założeniu, że dane są w postaci szeregu rozdzielczego a wartości kumulowane liczebności populacji oraz wartość cechy są wyrażone w %%. Współczynnik Giniego można obliczyć jako:

$$G = 1 - \left(\sum_{i=1}^{k} \frac{x_{cum_i} + x_{cum_{i-1}}}{2} (n_{cum_i} + n_{cum_{i-1}}) \right) / 5000 \tag{28}$$

przy czym $x_{cum_0} = 0$ oraz $n_{cum_0} = 0$; x_{cum_i} – wartość kumulowana cechy dla przedziału i; n_{cum_i} – wartość kumulowana liczebności dla przedziału i;

Powyższe to po prostu obliczenie pola **pod linią łamaną** w kwadracie 100×100. Połowa pola tego kwadratu wynosi 5000. Jeżeli krzywa Lorenza pokrywa się z przekątną kwadratu, to $G = 0$ (brak koncentracji). Wartości większ od zera wskazują na koncentrację wartości cechy.

http://www.fao.org/ag/AGP/agpc/doc/Counprof/uruguay/uruguay.htm

	Farms Area		Farm size (hectares)	
	Number	%%	Hectares	%%
1 to 4	6,260	10.9	16,516	0.1
5 to 9	7,086	12.4	47,611	0.3
10 to 19	7,118	12.5	97,841	0.8
20 to 49	8,934	15.6	285,254	1.7
50 to 99	6,647	11.6	472,928	2.9
100 to 199	6,382	11.2	910,286	5.5
200 to 499	6,783	11.9	2,162,836	13.2
500 to 999	3,687	6.8	2,725,637	16.6
1000 to 2499	2,912	5.1	4,441,627	27.0
2500 to 4999	838	1.5	2,837,134	17.3
5000 to 9999	228	0.4	1,504,482	9.2
10000 and more	56	0.1	917,531	5.6
TOTAL	57,131	100.0	16,419,683	100.0

Zadanie 1:

Trasa wyścigu dookoła Flandrii liczyła w 2007 roku 259 km. Od 125 km

do mety zaczęły się słynne flandryjskie pagórki, których wykaz jest na stronie `http://pl.wikipedia.org/wiki/Dookoła_Flandrii_2007`.

1. Wyznaczyć średnią wysokość wzniesień oraz przeprowadzić wszechstronną analizę szeregu szczegółowego wykorzystując pozycyjne miary przeciętnego poziomu (dominanta, mediana, kwartyle)

2. Przeprowadzić szczegółową analizę wykorzystując miary dyspersji i asymetrii.

3. Zbudować szereg rozdzielczy.