Statystyka matematyczna

Elementy statystyki opisowej

Średnia arytmetyczna – inaczej średnia lub przeciętna z n liczb x_1, x_2, … x_n, dla n $\geq$ 1 jest to liczba $x_{śr}$ określona wzorem:

$x_s = (x_1 + x_2 + \dots + x_n) / n$

Najczęściej oznacza się średnią arytmetyczną jako x nadkreślone:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\bar{x}$ - średnia arytmetyczna

$\sum_{i=1}^{n} x_i$ - suma liczb x dla i od 1 do n, n ≥ 1

Przykład:
Średnia arytmetyczna liczba: -1, 0, 1, 4, 6:
x_s = (-1 + 0 + 1 + 4 + 6) : 5 = 10 : 5 = 2

Średnia arytmetyczna z ciągu złożonego z tej samej liczby jest równa tej liczbie.
Przykład:
Średnia z liczb: 12, 12, 12, 12 wynosi 12, bo:
(12+ 12 + 12 + 12) / 4 = 4*12 : 4 = 12

W przypadku podanych **liczebności** kolejnych danych x_i, średnia x_s wyraża się wzorem:

$Xs = (n_1*x_1 + n_2*x_2 + \dots + n_k*x_k) / (n_1 + n_2 + \dots + n_k)$

$$x_s = \frac{x_1^* n_1 + x_2^* n_2 + \dots + x_n^* n_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^{k} x_i^* n_i}{\sum_{i=1}^{k} n_i} \quad k \in \mathrm{N}$$

Liczebności można potraktować jak wagi i wzór jest analogiczny jak dla średniej ważonej

Przykład

Wartość $\mathbf{x_i}$	0	1	2	3
Liczebność $\mathbf{n_i}$	4	2	3	1

x_s = (n1*x1 + n2 *x2 + n3*x3 + n4*x4)/ (n1 + n2 + n3 + n4)
x_s = (0*4 + 1*2 + 2*3 + 3*1) / (4 + 2 + 3 + 1) = 1.1

W przypadku podanych **częstości** (gdzie suma częstości równa się 1)
$x_s = x_1*c_1 + x_2 * c_2 + \dots + x_n*c_n$

Wartość $\mathbf{x_i}$	1	2	3	4	5	6
Częstość C_i	0	0,09	0,41	0,35	0,10	0,05

x_s = 1*0 + 2*0,09 + 3*0,41 + 4*0,35 + 5*0,10 + 6*0,05 = 3,61

Średnia ważona liczba $\mathbf{x_1, x_2, \dots x_n}$, z wagami równymi odpowiednio $\mathbf{p_1, p_2, \dots p_n}$
Jest to liczba $\mathbf{x_p}$ określona wzorem:
$\mathbf{x_p = (x_1 * p_1 + x_2 * p_2 + \dots x_n * p_n) / (p_1 + p_2 + \dots p_n)}$
$\mathbf{p_i}$ – waga (*power*) - liczba dodatnia
wagi - liczby dodatnie określające ważność poszczególnych składników sumy
(znaczenie, ważność, częstość występowania.

$$x_p = \frac{x_1 * p_1 + x_2 * p_2 + \dots x_n * p_n}{p_1 + p_2 + \dots p_n} = \frac{\sum_{i=1}^{n} x_i * p_i}{\sum_{i=1}^{n} p_i} \qquad k \in \mathrm{N}$$

Przykład 1:
Średnia ważona liczb -1, 0, 1, 4, 6 z wagami odpowiednio: 1, 2, 3, 4, 5
x_p = (-1*1 + 0*2 + 1 * 3 + 4*4 + 6*5) / (1 + 2 + 3 + 4 + 5) = 3,2

Przykład 2:
W pewnej klasie uzyskano następujące oceny ze sprawdzianu:
6 – 1 osoba, 5 - 2 osoby, 4 - 3 osoby, 3 - 5 osób, 2 – 6 osób, 1 - 4 osoby.
Oblicz średnią ocenę ze sprawdzianu.
x_p = (6*1 + 5*2 + 4*3 + 3*5 + 2*6 + 1*4) / (1 + 2 + 3 + 5 + 6 + 4) = (6 + 10 + 12 + 15 + 12 + 4)/49 = 59/15= **3,93**
Odpowiedź: Średnia ocena ze sprawdzianu w klasie wynosi 3,93.

Dominanta (moda) - wartość występująca najczęściej.
W niektórych przypadkach istnieje kilka dominant.
http://pl.wikipedia.org/wiki/Dominanta_(statystyka)
Dominanta *(wartość modalna, moda, wartość najczęstsza) to jedna z miar tendencji centralnej, statystyka dla zmiennych o rozkładzie dyskretnym, wskazująca na wartość o największym prawdopodobieństwie wystąpienia,*
lub wartość najczęściej występująca w próbie.
Dla zmiennej losowej o rozkładzie ciągłym jest to wartość, dla której funkcja gęstości prawdopodobieństwa ma wartość największą.

Przykład 1
Podaj dominantę z liczb: **6**, 2, 4, **6**, 1, **6**, 4
Częstość występowania: **6** – 3 razy, 4 – 2 razy, pozostałe liczby 1 raz
Rozwiązanie: Najczęściej występuje liczba **6**, która jest dominantą zestawu liczb.

Przykład 2

Liczba	**3**	4	**6**	**7**	10
Częstość występowania	4	3	4	4	2

Trzy liczby powtarzają się tak samo często, zatem są trzy dominanty.
Są to: 3, 6 , 7.
Odpowiedź: Są trzy dominanty: 3, 6 i 7.

Przykład 3

Wartość	Prawdopodobieństwo
1	0,3
2	0,25
3	0,35
4	0,15
5	0,30

Wartość modalna dla tego rozkładu wynosi **3**, bo tam jest największe prawdopodobieństwo.

Zastosowanie dominanty

http://pl.wikipedia.org/wiki/Dominanta_(statystyka)

Dominanta może być szczególnie użyteczna, gdy wartości zmiennej obserwowanej nie są liczbowe a opisowe - co uniemożliwia (bez przypisania wartości liczbowych) zastosowania m.in. mediany czy średniej arytmetycznej.

*W zbiorze {**jabłko, gruszka, jabłko, pomarańcza, gruszka, banan, jabłko**} dominantą jest **jabłko**; w klasie jest 5 brunetek, 3 blondynki i 4 szatynki - dominantą jest **brunetka**.*

Dominanta jest niedoceniana w zagadnieniach społecznych czy ekonomicznych np. przy analizowaniu zagadnień płacowych gdyż lepiej od powszechnie stosowanego średniego wynagrodzenia oddaje strukturę wynagrodzeń.

Np. w sklepie pracuje 5 osób: kierownik z wynagrodzeniem 10000 zł; zastępca 7000 zł i trzech sprzedawców po 1000 zł

- średnie wynagrodzenie to 4 tysiące a najczęstsze (dominanta) to 1000 złotych.

Mediana – wartość środkowa n liczb $x_1, x_2, \dots x_n$ uporządkowanych niemalejąco to liczba dzieląca ten zestaw na 2 części o równej liczebności.

Mediana $\mathbf{m_e}$ dzieli zbiór danych wartości na dwie części tak, że liczba danych, których wartości zmiennej są mniejsze od mediany, jest równa liczbie danych, których wartości zmiennej są większe od mediany.

Mediana (zwana też wartością środkową, wartością przeciętną lub drugim kwartylem) – w statystyce wartość cechy w szeregu uporządkowanym, powyżej i poniżej której znajduje się jednakowa liczba obserwacji.

$$m_e = x_{\frac{n+1}{2}} \quad \text{gdy n jest liczbą nieparzystą}$$

$$m_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \quad \text{gdy n jest liczbą parzystą}$$

Mediana m_e

$m_e = x_{(n+1)/2}$ gdy n jest liczbą nieparzystą

$m_e = (x_{n/2} + x_{(n/2+1)}) / 2$ gdy n jest liczbą parzystą

Przykłady – liczby uporządkowane rosnąco

1)

n jest liczbą nieparzystą – mediana jest liczbą środkową

-3 0 **1** 5 9

n = 5, (n+1)/2 = (5+1)/2 = 6/2 =3
$m_e = x_3 = 1$
2) n jest liczbą parzystą – mediana jest średnią arytmetyczną 2 liczb środkowych
-2 -1 **1** **3** 5 10
n/2 = 6/2= 3
n/2 + 1 = 3+1 = 4
$m_e = (x_3 + x_4) / 2 = (1 + 3) / 2 = 2$

Przykłady – liczby nieuporządkowane

Przykład 1
Podaj medianę z liczb: 2, 7, 12, 3, 9.
Najpierw porządkujemy liczby: 2, 3, **7**, 9, 12
n = 5, (n+1)/2 = 6/2 = 3
$m_e = x_3 =$ **7**

Przykład 2
Oblicz medianę z liczb: 7, 4, 6, 4, 7, 1, 9.
Po uporządkowaniu rosnąco: 1, 4, 4, **6,** 7, 7, 9
$m_e =$ 6.

Przykład 3
Oblicz medianę z liczb 12, 6, 24, 18, 40, 8
Po uporządkowaniu: 6, 8, **12, 18**, 24, 40
$m_e = (12 + 18)/ 2= 15$
Rozstęp R - różnica między największą i najmniejszą z liczb w zbiorze
(lub różnica między najwyższą i najniższą zaobserwowaną wartością zmiennej).
Jest to jedna z prostszych miar rozrzutu.
Rozstęp pokazuje zatem jedynie jaki jest zakres obserwacji nie informuje w żaden sposób co dzieje się "w środku" tego zakresu
np. jaka wartość występowała najczęściej, czy jaka jest średnia dla tego zbioru obserwacji.

http://pl.wikipedia.org/wiki/Rozst%C4%99p
Rozstęp jest najprostszą z miar rozrzutu, mało precyzyjną,
gdyż opiera się tylko na dwu zaobserwowanych wartościach zmiennej, a pozostałe wartości nie mają wpływu na jej wielkość.
Przykład zastosowania:
w pedagogice w analizie ilościowej wyników egzaminowania rozstęp bywa obliczany dla uzyskania wstępnej orientacji
co do rezultatów egzaminowania albo wtedy,
gdy chodzi wyłącznie o krańcowe wyniki.

http://www.naukowiec.org/wiedza/statystyka/rozstep_711.html
Dwie grupy osób poproszono o ocenę na dziesięciostopniowej skali.
W jednej grupie najniższy wynik wynosił 3, najwyższy 9.
W drugiej grupie natomiast wynik najniższy to 1 a najwyższy 9.
R1 = X1 max - X1 min = 9 - 3 = 6
R2 = X2 max - X2 min = 9 - 1 = 8
Widzimy zatem, że w grupie 2 zakres odpowiedzi był większy niż w grupie 1.
Oznacza to, że w grupie 2 udzielano bardziej skrajnych odpowiedzi niż w grupie 1.

Miarę tą często stosuje się dla takich zmiennych jak wiek osób badanych czy uzyskiwany dochód. ***Rozstęp*** *informuje nas wtedy, jak duża była dysproporcja pomiędzy osobą najmłodszą a najstarszą, pomiędzy osobą o najniższych dochodach a najwyższych. Określa jak duża jest ta różnica.*

Odchylenie wartości x od średniej arytmetycznej

$d = x_i - x_s$

Przykład

$x1 = 4, \; x2 = -3, \; x3 = 2$

$x_s = x = (4 - 3 + 2)/3 = 1$

$x_1 - x = 4 - 1 = 3$

$x_2 - x = -3 - 1 = -4$

$x_3 - x = 2 - 1 = 1$

Odchylenie bezwzględne elementu zbioru danych statystycznych to wartość bezwzględna różnicy wartości

dla tego elementu x_i i pewnego ustalonego punktu x.

$d = |x_i - x|$

Za ustalony punkt, od którego liczone jest odchylenie bezwzględne,

przyjmuje się zazwyczaj medianę lub średnią arytmetyczną.

Dla przykładu wyżej:

$d1 = |x_1 - x| = |4 - 1| = 3$

$d2 = |x_2 - x| = |-3 - 1| = 4$

$d3 = |x_3 - x| = |2 - 1| = 1$

Odchylenia bezwzględne są wykorzystywane przy obliczaniu średniego odchylenia bezwzględnego.

Odchylenie średnie jest to średnia arytmetyczna odchyleń od wartości średniej.

Średnie odchylenie bezwzględne (inaczej: odchylenie **przeciętne**)

to średnia arytmetyczna z odchyleń bezwzględnych dla wszystkich elementów zbioru danych statystycznych.

$D = \sum(x_i - x) / n$, i = 1 .. n   n – liczebność zbioru danych

Dla przykładu wyżej

$D = (3 + |-4| + 1) / 3 = 8/3 = 2,7$

Wariancja to w statystyce klasyczna miara zmienności.

Intuicyjnie utożsamiana ze zróżnicowaniem zbiorowości;

Wariancja jest średnią arytmetyczną kwadratów odchyleń (różnic) poszczególnych wartości cechy od wartości oczekiwanej.

Wariancja liczb x1, x2, … xn jest to średnia arytmetyczna z kwadratów odchyleń liczb od ich wartości średniej

$$\boxed{\sigma^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}$$

$\sigma^2 = \sum(x_i - x)^2 / n = (x_1 - x)^2 + (x_2 - x)^2 + \ldots (x_n - x)^2) / n$ *dla* $i = 1$ *do* n
gdzie **x** = średnia arytmetyczna wartości $\mathbf{x_i}$

$\mathbf{x = \sum(x_i) / n = (x_1 + x_2 + \ldots x_n) / n}$

$\sigma^2 = \sum(x_i)^2 / n - x = (x_1^2 + x_2^2 + \ldots + x_n^2) / n - x^2$

<u>Inna postać wzoru:</u>
$\sigma^2 = \sum x_i^2 / n - x$, gdzie $x = x_s$ - średnia arytmetyczna

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

Odchylenie standardowe liczb od średnie arytmetycznej – pierwiastek kwadratowy z wariancji

$\sigma = \sqrt{\sigma^2} = \sqrt{}(x_i - x)^2 / n = \sqrt{}(\sum(x_i)^2 / n - x)$ gdzie x – średnia arytmetyczna wartości x_i

W przypadku występowania **liczebności** danych wzory uwzględniające **liczebność**:

$\sigma^2 = \sum(n_i * x_i - x)^2 / n = n_1 * (x_1 - x)^2 + n_2 * (x_2 - x)^2 + \ldots n_k * (x_n - x)^2) / (n_1 + n_2 + \ldots n_k)$

$$\sigma^2 = \frac{n_1 (x_1 - \bar{x})^2 + n_2 (x_2 - \bar{x})^2 + \ldots + n_k (x_n - \bar{x})^2}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{\sum_{i=1}^{k} n_i}$$

$\sigma = \sqrt{\sigma^2}$

<u>Analiza dokładności</u>

Odchylenie wartości od średniej arytmetycznej
$\mathbf{v_i = x_i - x_s}$ – odchylenie wartości xi od średniej arytmetycznej xs

Odchylenie przeciętne
$d = (| x_1 - x_s | + | x_2 - x_s | + \ldots + | x_n - x_s |) / n = (| v_1 | + | v_2 | + \ldots + | v_n |) / n = (\Sigma | x_i - x_s |) / n$

Wariancja
Wariancja to w statystyce klasyczna miara zmienności.
Wariancja informuje o tym, <u>jak duże jest zróżnicowanie wyników w danym zbiorze wyników</u>
$\sigma^2 = [(x_1 - x_s)^2 + (x_2 - x_s)^2 + \ldots + (x_n - x_s)^2] / n = (\Sigma(x_i - x_s)^2) / n$

Odchylenie standardowe

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

Odchylenie standardowe jest pierwiastkiem kwadratowym z wariancji.

$\sigma = \surd\,([\,(x_1 - x_s)^2 + (x_2 - x_s)^2 + \ldots + (x_n - x_s)^2\,] / n) = \surd\,(\,\Sigma(x_i - x_s)^2\,) / n\,)$

$\boldsymbol{\sigma = \surd\,\sigma^2}$

Współczynnik zmienności v – odchylenie względne

$\boldsymbol{v = \sigma / x_s * 100\%}$

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.

Przykład

Obliczenie średniej ocen z analizą dokładności

Obliczenie sredniej ocen, odchylenia przeciętnego, wariancji i odchylenia standardowego				
Lp	**x(i)**	**v(i) = x(i) - x**	**| x(i) - x|**	**v(i)^2**
1	5	2	2	4
2	4	1	1	1
3	3	0	0	0
4	3	0	0	0
5	3	0	0	0
6	2	-1	1	1
7	1	-2	2	4
Suma	21	0	6	10
średnia	**3**		0,86	
Srednia	=21/7		**3**	
Odchylenie przeciętne d	= 2/7		**0,86**	
Wariancja sigma^2	= 2/7		**1,43**	
Odchylenie stand. sigma	Pierwiastek(2/7)		**1,20**	
Współczynnik zmienności v	=1,20 / 3 * 100%		**17,07%**	