Średnia, odchylenie standardowe i odchylenie standardowe średniej. Inne parametry statystyczne.

mgr Maciej Wróbel

Uniwersytet Śląski, Katowice 2010

1 Średnia arytmetyczna

Dla większości najlepszym przybliżeniem wartości prawdziwej jest średnia arytmetyczna otrzymanych pomiarów:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{k=1}^{M} x_k k_k}{N} = \sum_{k=1}^{M} p_k x_k,$$

gdzie x_i to i-ty pomiar, N oznacza ilość otrzymanych wyników, M to liczba *różnych* otrzymanych wartości, n_k to liczba otrzymanych wartości x_k ($\sum n_k = N$), a $p_k = \frac{n_k}{N}$ ($\sum p_k = 1$) to częstość otrzymania wartości x_k. Przykładowo:

W wyniku serii dwudziestu pomiarów otrzymano następujące wartości średnicy drewnianej kulki [mm]: 11, 12, 11, 11, 12, 10, 11, 12, 13, 12, 12, 13, 11, 11, 9, 10, 10, 12, 11, 11. Tabela przedstawia otrzymane wyniki:

k	1	2	3	4	5	**suma**
x_k	**9**	**10**	**11**	**12**	**13**	**225**
n_k	1	3	8	6	2	**20**
$n_k x_k$	9	30	88	72	26	**225**
p_k	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{10}$	**1**
$p_k x_k$	0,45	1,5	4,4	3,6	1,3	**11,25**

Tablica 1: Przykładowe wyniki pomiarów pogrupowane według ilości wystąpień i częstości wystąpień

Widać, że średnie otrzymane z każdej „wersji" równania dają tą samą wartość, jednak ostatnie dwie postacie równania są szczególnie wygodne, gdy posługujemy się dużą liczbą obserwacji.

2 Odchylenie standardowe

Każdy pomiar obarczony jest niepewnością wynikającą z np. niedokładności przyrządów pomiarowych. Możemy jednak otrzymać pewną „średnią" niepewność pomiaru sprawdzając, o ile różnią się wartości otrzymane od wartości średniej. Miara ta jest *niezależna* od przyrządów, którymi się posługujemy. Przyjęło się mierzyć średnią niepewność jako:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum_k n_k (x_k - \bar{x})^2} = \sqrt{\sum_k p_k (x_k - \bar{x})^2}.$$

Wielkość tą nazywamy odchyleniem standardowym. Określa ono *przeciętne odchylenie każdego z pomiarów od wartości prawdziwej.* Często spotkacie się z alternatywną definicją odchylenia standardowego:

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{N-1}\sum_k n_k (x_k - \bar{x})^2}$$

Definicja ta jest lepiej uzasadniona teoretycznie i daje nieco większą wartość σ_x. Dla dużych wartości N wielkości te prawie nie różnią się liczbowo, *należy jednak zawsze* pisać, której definicji używamy.

3 Odchylenie standardowe średniej

Możemy spodziewać się, że niepewność otrzymania każdego z pomiarów będzie większa, niż niepewność wyznaczenia średniej wartości (bo przecież średnia otrzymana z wielu wartości, a błędy przypadkowe powinny się w jakimś stopniu znosić). Rzeczywiście, niepewność otrzymania średniej wynosi:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

i nazywa się odchyleniem standardowym średniej.

4 Interpretacja odchylenia standardowego

Odchylenie standardowe wyznacza ok. 68% przedział ufności otrzymanego wyniku. Oznacza to, że ok. 68% wyników znajduje się w odległości nie większej niż σ_x od wartości średniej przy założeniu, że otrzymane pomiary podlegają rozkładowi normalnemu (o tym na kolejnych zajęciach). Jeżeli chcielibyśmy mieć większą szansę na to, że otrzymany pomiar zmieści się w obszarze błędu, to musielibyśmy przyjąć błąd nie jako $\pm\sigma_x$, ale jako większą wartość. To, jaki będzie przedział ufności (dla rozkładu normalnego) wyznacza tzw. funkcja błędu (oznaczana przez erf(t)). Argument funkcji (t) ma interpretację $\delta x = \frac{t\sigma_x}{\sqrt{N}}$, tzn jeżeli $t > 1$, to przedział ufności jest większy niż dla σ_x (bo wyliczona niepewność jest większa i jest większa szansa, że otrzymany wynik „wpadnie" w przedział), a mniejszy dla $t < 1$. Wartość funkcji erf(t) jest tablicowana, pozwalają ją także wyliczyć bardziej zaawansowane kalkulatory. Warto zapamiętać, że:

1. dla $t = 1$ ufność jest na poziomie 68%
2. dla $t = 2$ ufność jest na poziomie 95%
3. dla $t = 3$ ufność jest na poziomie 99%

Jest to tak zwana reguła trzech sigm.

5 Przypadek małej liczby pomiarów

Jeżeli otrzymamy małą ilość wyników pomiarów (tak, definicja ta nie jest zbyt ścisła), to założenie o normalności rozkładu nie jest poprawne. Aby otrzymać podobny przedział ufności, jak w przypadku rozkładu normalnego należy przyjąć błąd większy, niż wynika z wyliczenia σ_x. Są przesłanki teoretyczne wskazujące, że lepiej pasującym rozkładem jest rozkład Studenta. Jeżeli obliczymy odchylenie standardowe, to niepewność otrzymanej średniej przyjmujemy większą zgodnie ze wzorem:

$$\delta x = \pm\frac{t_{P,N-1}\sigma_x}{\sqrt{N}},$$

gdzie $t_{P,N-1}$ jest odczytywaną wartością tablicową kwantyla t rozkładu Studenta, P[1] jest zadanym prawdopodobieństwem znalezienia wyniku w przedziale $\bar{x} \pm \delta x$ a N jest ilością otrzymanych wyników.

6 Inne parametry statystyczne

Często chcemy opisać otrzymane wyniki w sposób bardziej pełny niż przy pomocy średniej i jej odchylenia standardowego. Możemy chcieć np. ocenić asymetrię rozkładu wyników lub też ich koncentrację. Pomocne są tym parametry statystyczne. Ze względu na zastosowanie można je podzielić na:

6.1 Miary położenia

Miary położenia pozwalają nam porównywać rozkłady podobne do siebie, przesunięte jednak względem osi odciętych (OX). Poznaną już miarą położenia jest średnia arytmetyczna. Inne przykłady to

1. Mediana m - jest to taka liczba, że połowa otrzymanych wyników ma wartość mniejszą lub równą medianie, a reszta - wartość większą ($P(x \leq m) = P(x \geq m) = \frac{1}{2}$). Zaletą (i wadą) mediany jest odporność na elementy odstające, tj. takie, które nie pasują do modelu (np. błędy grube).

[1] Uwaga: często w tablicach zamiast t_P podaje się t_α, gdzie $\alpha = 1 - P$

2. Moda d - inaczej dominanta, to wartość występująca najczęściej (dla rozkładów dyskretnych) lub o największej wartości funkcji gęstości prawdopodobieństwa (dla rozkładów ciągłych). Zaletą mody jest to, że można ją zastosować także do wartości innych niż liczbowe.

3. Kwantyle rzędu p x_p - takie liczby, że prawdopodobieństwo otrzymania wartości mniejszej niż x_p jest większe lub równe p ($P(x \leq x_p) = p$. Podając kilka kwantyli możemy w sposób pełniejszy opisać otrzymany rozkład wyników.

6.2 Miary zróżnicowania

Miary zróżnicowania określają, jak poszczególne otrzymane wartości różnią się od wartości centralnych (np. od średniej). Przykładem jest odchylenie standardowe średniej. Inne popularne miary to:

1. Średnie odchylenie bezwzględne - obliczane jako $D = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$.

2. Rozstęp - to odległość między wartością największą i najmniejszą.

3. Rozstęp ćwiartkowy - jest to różnica pomiędzy kwantylem 0.75 i kwantylem 0.25 ($IQR = x_{\frac{3}{4}} - x_{\frac{1}{4}}$), tj. określa, jaka w jakim przedziale leży 50% otrzymanych obserwacji. Jest odporny na wartości odstające.

6.3 Miary asymetrii

Często istotną informacją o wynikach jest asymetria ich rozkładu. Asymetrię tą najczęściej opisuje się przez:

1. Trzeci moment centralny (obliczany jako $M_3 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N}$). Jeżeli $M_3 < 0$, to rozkład jest lewostronnie asymetryczny (tzn. więcej wyników jest mniejsza od wartości przeciętnej). $M_3 = 0$ dla rozkładu symetrycznego, a $M_3 > 0$ dla rozkładu prawostronnie asymetrycznego.

2. Współczynnik asymetrii $A = \frac{M_3}{\sigma^3}$. Ma podobne własności do trzeciego momentu centralnego, można jednak przy jego pomocy porównywać różne rozkłady.

3. Współczynnik skośności - proporcjonalny do różnicy pomiędzy dominantą i średnią lub medianą i średnią ($A_d = \frac{\mu - d}{\sigma}$ lub $A_m = 3\frac{\mu - m}{\sigma}$).

6.4 Miary koncentracji

Miary koncentracji określają, jak bardzo wyniki skupione są wokół wartości centralnych. Najpopularniejszą miarą koncentracji jest kurtoza $k = \frac{\frac{1}{N}\sum (x_i - \mu)^4}{\sigma^4} - 3$. k przyjmuje wartość 0 dla rozkładu normalnego, wartości większe niż 0, gdy wartości są silnie skoncentrowane wokół wartości średniej i mniejszą niż zero, gdy rozkład jest „płaski”.