STATYSTYKA MATEMATYCZNA

WYKŁAD 1

Wiadomości wstępne

Statystyka to dyscyplina naukowa, której zadaniem jest wykrywanie, analiza i opis prawidłowości występujących w procesach masowych.

Populacja to zbiorowość podlegająca badaniu statystycznemu.
Aby populację określić jednoznacznie charakteryzujemy ją pod względem:

–rzeczowym

–czasowym

–przestrzennym (terytorialnym).

Cecha to właściwość elementów populacji ze względu na którą prowadzimy badanie statystyczne.

Warianty to wartości cechy (cecha powinna mieć przynajmniej dwa warianty).

Przykład

Populacja:

Studenci II semestru Wydziału Elektroniki WAT, wg stanu na 1.03.2015.

Cechy:

- płeć,
- wzrost,
- kolor oczu,
- ocena na egzaminie z matematyki po I semestrze,
- ulubiony tygodnik,
- wysokość miesięcznych dochodów,
- czas poświęcony na naukę w tygodniu poprzedzającym ostatnią sesję egzaminacyjną.

Przykład

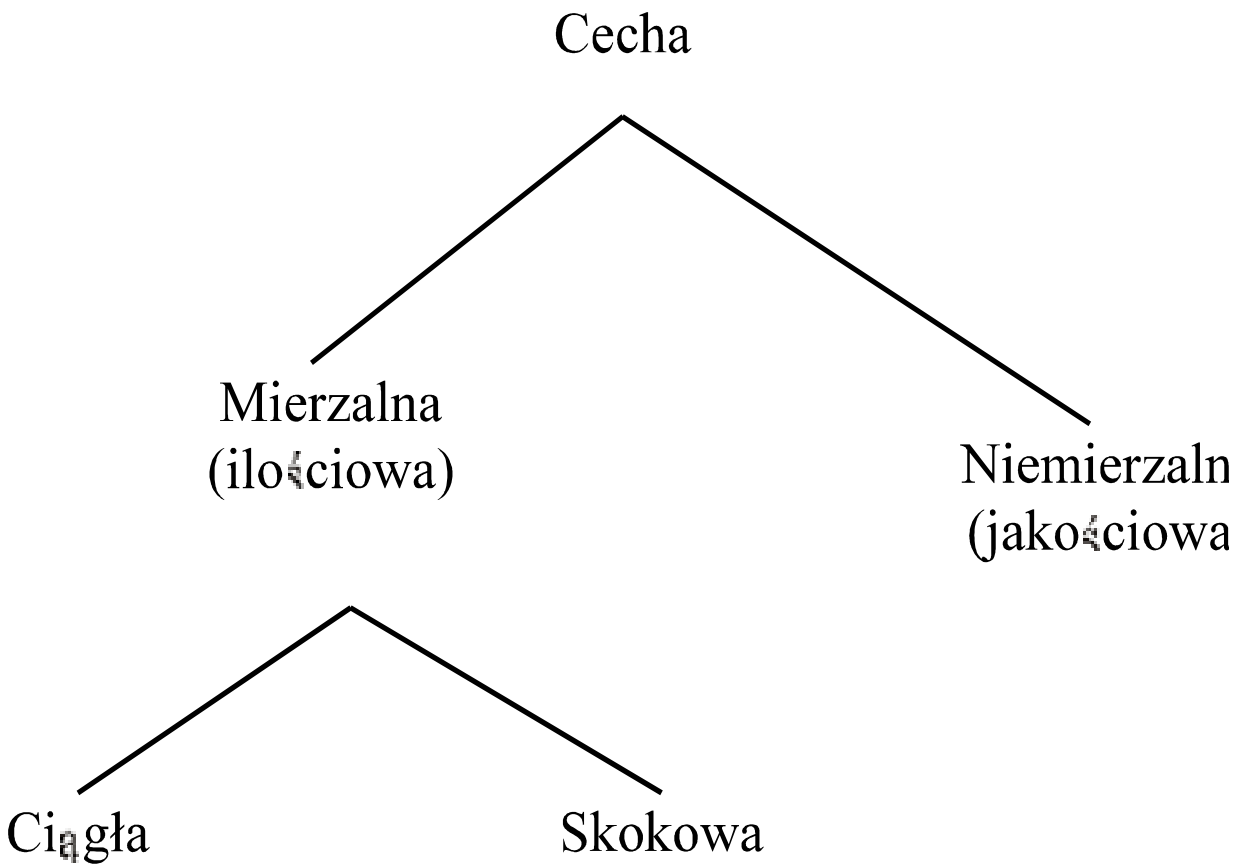
Populacja:

Samochody osobowe zarejestrowane w Warszawie, wg stanu na 1.09.2015.

Cechy:

- kolor karoserii,
- przebieg,
- średnie zużycie paliwa na 100 km,
- marka,
- czas osiągania prędkości 100 km/godz.

Uproszczona klasyfikacja cech:



Badanie statystyczne może być:
– **pełne** (obejmuje całą populację),
– **częściowe** (obejmuje część populacji – próbę).

Próba powinna być **reprezentatywna** tzn. rozkład wariantów badanej cechy w próbie powinien być zbliżony do rozkładu w całej populacji.

George Gallup 1901-1984

Pionier w dziedzinie badania opinii publicznej.
Rozwinął technikę doboru grupy reprezentatywnej

Rok 1936 - wybory prezydenckie w USA.
Franklin Delano Roosvelt - Partia Demokratyczna,
Alf Landon - Partia Republikańska.

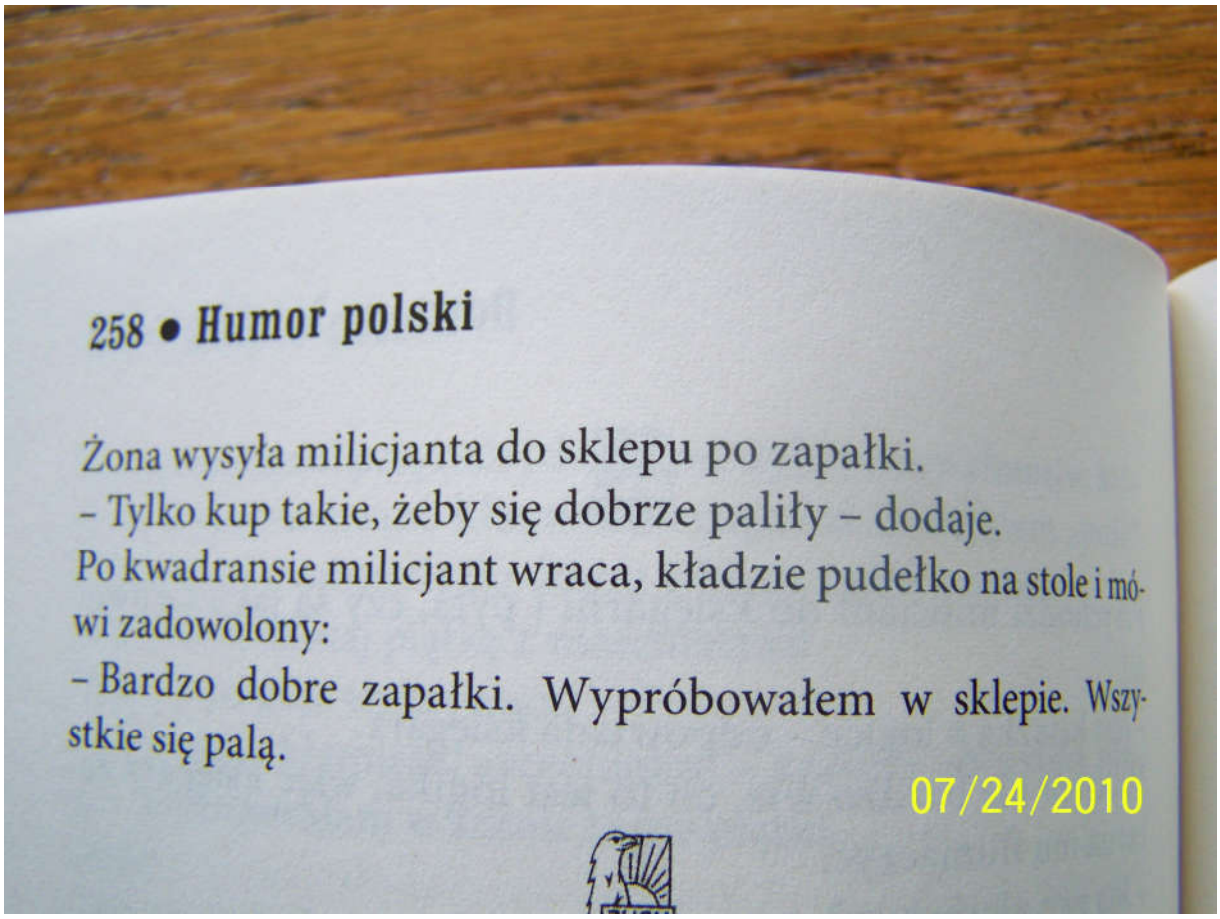
"*Literary Digest*" 10 mln ankiet (zwrot ok. 2mln),
- nieprawidłowa prognoza.

Gallup 4000 ankiet (w 1935 założył pierwszy na świecie instytut badania opinii publicznej) - **prawidłowa prognoza**.

Wyniki:
Roosvelt - 60,8%,
Landon - 36,5%.

Uwaga

Badania pełne nie zawsze są możliwe lub celowe (badania niszczące, duża próba, wysokie koszty).



258 • Humor polski

Żona wysyła milicjanta do sklepu po zapałki.
– Tylko kup takie, żeby się dobrze paliły – dodaje.
Po kwadransie milicjant wraca, kładzie pudełko na stole i mówi zadowolony:
– Bardzo dobre zapałki. Wypróbowałem w sklepie. Wszystkie się palą.

07/24/2010

„Humor Polski” – lata 80-te

Liczebność próby.
Dla reprezentatywnej próby dorosłej liczebności Polski zwykle 1000 – 1300 osób.

Jerzy Spława-Neyman (1894 - 1981)
polski i amerykański matematyk i statystyk.
Wprowadził pojęcie przedziału ufności.

ROZKŁADY PODSTAWOWYCH STATYSTYK

X – zmienna losowa – odpowiednik badanej cechy,

$(X_1, X_2, ..., X_n)$ – próba losowa (zmienna losowa n wymiarowa,

X_i – niezależne zmienne losowe o takim samym rozkładzie jak X (taką próbę nazywamy **próbą prostą**).

Jeśli x_i jest wartością zmiennej X_i ($i = 1, 2, ..., n$) to ciąg $(x_1, x_2, ..., x_n)$ nazywamy realizacją próby (są to dane statystyczne).

Statystyka to praktycznie dowolna funkcja od próby

$$Y = g(X_1, X_2, \ldots, X_n)$$

Statystyka przekształca informację zawartą w próbie czyniąc **prostszym** wnioskowanie o rozkładzie cechy w populacji.

Statystyka jako funkcja od zmiennej losowej jest też zmienną losową i możemy mówić o jej rozkładzie.

Statystyka ma rozkład **dokładny,** jeśli jest spełniony dla każdego *n*.
Statystyka ma rozkład **asymptotyczny,** jeśli jest spełniony, gdy *n* dąży do nieskończoności.

Statystyki podstawowe:

$$\overline{X} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$ średnia z próby

Gdy X_i mają rozkład zerojedynkowy (1 – sukces, 0 – porażka) to średnią możemy zapisać w postaci

$$W = \frac{Y_n}{n}$$

gdzie Y_n jest liczbą sukcesów w próbie

Ten szczególny przypadek średniej nazywamy średnią częstością sukcesu.

$$S^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

wariancja z próby

Uwaga.
$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} \; - \left(\overline{X}_n\right)^2$$

$$S = S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2} = \sqrt{S_n^2}$$

odchylenie standardowe z próby

$$V = V_n = \frac{S_n}{\overline{X}_n}$$

Współczynnik zmienności (dla cech o wariantach dodatnich)

$$\hat{S}^2 = \hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

wariancja z próby – **nieobciążona**

$$S^{02} = S_n^{02} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2$$

wariancja z próby dla danej wartości oczekiwanej m.

Uwaga

$$\hat{S}_n^2 = \frac{n}{n-1} S_n^2 \qquad S_n^2 = \frac{n-1}{n} \hat{S}_n^2$$

zatem dla dużych n $\quad \hat{S}_n^2 \approx S_n^2$

Momenty zwykłe,

$M_k = \frac{1}{n}\sum X_i^k$ – moment rzędu k cechy X ($M_1 = \overline{X}_n$).

$M_{kl} = \frac{1}{n}\sum X_i^k \cdot Y_i^l$ – moment rzędu k, l jednocześnie badanych cech (X, Y).

Momenty centralne,

$\tilde{M}_k = \frac{1}{n}\sum (X_i - \overline{X})^k$ – moment rzędu k cechy X .

$\tilde{M}_{kl} = \frac{1}{n}\sum (X_i - \overline{X})^k (Y_i - \overline{Y})^l$ – moment rzędu k, l jednocześnie badanych cech (X, Y).

Rozkłady niektórych statystyk (n>1):

Jeśli cecha X ma rozkład N(m, σ), to:

a) statystyka $\overline{X}_n$ ma rozkład $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$,

b) statystyka $\frac{\overline{X}_n - m}{S_n}\sqrt{n-1}$ ma rozkład Studenta z n - 1 stopniami swobody,

c) statystyka $\frac{nS_n^{0\,2}}{\sigma^2}$ ma rozkład chi kwadrat z n stopniami swobody,

d) statystyka $\frac{nS_n^{\,2}}{\sigma^2}$ ma rozkład chi kwadrat z n - 1 stopniami swobody,

d') statystyki $\overline{X}_n$ i S_n^2 są zmiennymi losowymi niezależnymi (zachodzi też własność odwrotna),

Jeśli cecha X ma rozkład $N(m_1, \sigma_1)$ a cecha Y ma rozkład $N(m_2, \sigma_2)$, (próby niezależne odpowiednio n_1 i n_2 elementowe) to:

e) statystyka $\overline{X}_{n_1} - \overline{Y}_{n_2}$ ma rozkład $N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$,

gdy X ma rozkład $N(m, \sigma)$, Y ma rozkład $N(m, \sigma)$, to

e') statystyka $\frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{n_1 S_{n_1}^2 + n_2 S_{n_2}^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$

ma rozkład Studenta z $n_1 + n_2$ - 2 stopniami swobody,

f) statystyka $\frac{\frac{\hat{S}_{n_1}^2(X)}{\sigma_1^2}}{\frac{\hat{S}_{n_2}^2(Y)}{\sigma_2^2}}$ ma rozkład Snedecora F_{n_1-1, n_2-1},

Ad. a) Zmienna losowa $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ jako suma niezależnych zmiennych losowych o rozkładach normalnych pomnożona przez stałą ma rozkład normalny.

Obliczymy jej parametry korzystając z własności wartości oczekiwanej i wariancji.

$$E\left(\overline{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\cdot n \cdot m = m$$

$$D^2\left(\overline{X}_n\right) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

zatem $D\left(\overline{X}_n\right) = \frac{\sigma}{\sqrt{n}}$

Ad. b) wykorzystamy a), d), d'),

Ponieważ $\frac{\overline{X}_n - m}{S_n}\sqrt{n-1} = \frac{\frac{\overline{X}_n - m}{\sigma}\sqrt{n}}{\sqrt{\frac{nS_n^2}{\sigma^2(n-1)}}}$

licznik ma rozkład N(0, 1)

$\frac{nS_n^{\,2}}{\sigma^2}$ ma rozkład chi kwadrat z n - 1 stopniami swobody,

Statystyki te są niezależne.

Zatem (z definicji) statystyka $\frac{\overline{X}_n - m}{S_n}\sqrt{n-1}$ ma rozkład Studenta z n - 1 stopniami swobody,

Ad. a), d, d')

Niech ($Y_1, Y_2, ..., Y_n$) – próba losowa dla cechy o rozkładzie $N(0, 1)$.

Niech $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ i $K = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$

Aby wykazać a), d, d') wystarczy pokazać, że te statystyki są niezależne i mają rozkłady:

$\overline{Y}$ ma rozkład $N\left(0, \frac{1}{\sqrt{n}}\right)$

K ma rozkład chi kwadrat z n - 1 stopniami swobody

bo $\overline{X}$ ma rozkład taki jak $\sigma\overline{Y} + \mu$

a K ma rozkład $\frac{nS_n^{\,2}}{\sigma^2}$

$$\left(S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\left(\frac{X_i - m}{\sigma}\right) - \left(\frac{\overline{X}_n - m}{\sigma}\right)\right)^2 \approx \frac{\sigma^2}{n}K \right)$$

1. Określamy zmienne losowe

$$Z_k = \sum_{i=1}^{n} c_{ki} Y_i , \quad k = 1, ..., n$$

za pomocą ortonormalnej macierzy $C = [c_{ki}]$.

Pierwszy wiersz ma jednakowe elementy równe $\frac{1}{\sqrt{n}}$

(taka macierz zawsze istnieje).

Zmienne Z_k mają rozkład normalny.

2. $m_k = E(Z_k) = 0$,

$cov(Z_k, Z_j) = 0$ dla $k \neq j$ (z niezależności $Y_1, Y_2, ..., Y_n$ i ortogonalności C)

Zatem $Z_1, Z_2, ..., Z_n$ są niezależne (funkcje mierzalne niezależnych zmiennych losowych są niezależne) o rozkładzie $N(0, 1)$.

3. Skoro $Z_1 = \sum_{i=1}^{n} c_{1i} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ to $\overline{Y} = \frac{Z_1}{\sqrt{n}}$, zatem $\overline{Y}$ ma rozkład $N\left(0, \frac{1}{\sqrt{n}}\right)$

4. Liniowe przekształcenie ortonormalne zachowuje normę zatem $\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} Y_i^2$.

Zatem $\frac{K}{n} = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 - \frac{Z_1^2}{n} = \frac{1}{n} \sum_{i=2}^{n} Z_i^2$

co oznacza z definicji rozkładu chi kwadrat, że K ma rozkład chi kwadrat z n - 1 stopniami swobody.

5. $\overline{Y}$ i K jako funkcje mierzalne niezależnych zmiennych losowych są niezależne.

Ad. e) Zmienna losowa $\overline{X}_{n_1} - \overline{Y}_{n_2}$ jako różnica niezależnych zmiennych losowych o rozkładach normalnych (punkt a)) ma rozkład normalny. Obliczymy jej parametry korzystając z własności wartości oczekiwanej i wariancji.

$\overline{X}_{n_1}$ ma rozkład $N\left(m_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$,

$\overline{Y}_{n_2}$ ma rozkład $N\left(m_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$,

$$E\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) = E\left(\overline{X}_{n_1}\right) - E\left(\overline{Y}_{n_2}\right) = m_1 - m_2$$

$$D^2\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) = D^2\left(\overline{X}_{n_1}\right) + D^2\left(\overline{Y}_{n_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

zatem $$D\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ .$$

Ad. f) korzystając z d) mamy

$$\frac{\frac{\hat{S}^2_{n_1}(X)}{\sigma_1^2}}{\frac{\hat{S}^2_{n_2}(Y)}{\sigma_2^2}} = \frac{\frac{\frac{n_1}{n_1-1}S^2_{n_1}(X)}{\sigma_1^2}}{\frac{\frac{n_2}{n_2-1}S^2_{n_2}(Y)}{\sigma_2^2}} = \frac{\frac{1}{n_1-1}\cdot\frac{n_1 S^2_{n_1}(X)}{\sigma_1^2}}{\frac{1}{n_2-1}\cdot\frac{n_2 S^2_{n_2}(Y)}{\sigma_2^2}} = \frac{\frac{1}{n_1-1}\cdot Y_{n_1-1}}{\frac{1}{n_2-1}\cdot Y_{n_2-1}} = F_{n_1-1,n_2-1}$$

Uwaga.

1) Ciąg średnich z próby jest zbieżny (wg prawdopodobieństwa) do wartości oczekiwanej m rozpatrywanej cechy
(zakładamy, że EX = m istnieje),
2) Ciąg wariancji z próby jest zbieżny (wg prawdopodobieństwa) do wariancji σ^2 rozpatrywanej cechy
(zakładamy, że $D^2X = \sigma^2 > 0$ istnieje),
3) Gdy spełnione są założenia punktu 1) i 2) to średnia ma dla dużych *n* w przybliżeniu rozkład $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ (rozkład asymptotyczny)

W szczególności średnia częstość sukcesu $W = \frac{Y_n}{n}$ ma rozkład asymptotyczny $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$,
gdzie *p* – prawdopodobieństwo sukcesu.

Uogólnienie

Jeśli cecha X ma momenty odpowiednio wysokiego rzędu to momenty te mają rozkłady asymptotyczne normalne.

Moment M_k ma asymptotyczny rozkład

$$N\left(m_k, \sqrt{\frac{m_{2k} - m_k^2}{n}}\right)$$

Moment $\tilde{M}_k$ ma asymptotyczny rozkład

$$N\left(\mu_k, \sqrt{\frac{\mu_{2k} - 2k\mu_{k-1}\mu_{k+1} - \mu_k^2 + k^2\mu_2\mu_{k-1}^2}{n}}\right)$$

Przykład

Dochód miesięczny (zł) w pewnej populacji osób ma rozkład normalny $N(1600; 300)$.

a) Jakie jest prawdopodobieństwo, że średni miesięczny dochód 25 osób z tej populacji wynosi mniej niż 1500 zł?

b) Jakie jest prawdopodobieństwo, że miesięczny dochód osób z tej populacji wynosi mniej niż 1500 zł?

Rozwiązanie

a) $\overline{X}_{25}$ – średni miesięczny dochód 25 osób,

$$\overline{X}_{25} - N\left(1600, \frac{300}{\sqrt{25}}\right) = N(1600,60)$$

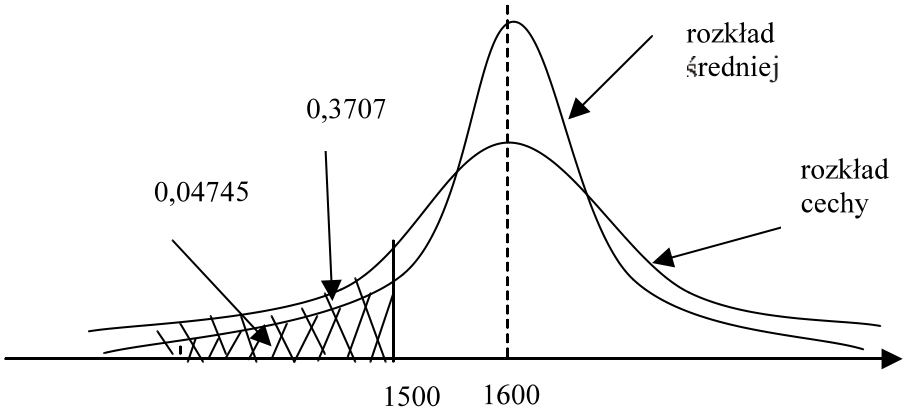
$$P(\overline{X}_{25} < 1500) = P\left(\frac{\overline{X}_{25} - 1600}{60} < \frac{1500 - 1600}{60}\right) = P(Y < -1{,}67) =$$

$$= \Phi(-1{,}67) = 1 - \Phi(1{,}67) = 1 - 0{,}95254 = 0{,}04745$$

b) X – wysokość miesięcznego dochodu, $X - N(1600,300)$

$$P(X < 1500) = P\left(\frac{X - 1600}{300} < \frac{1500 - 1600}{300}\right) = P(Y < -0{,}33) =$$

$$= \Phi(-0{,}33) = 1 - \Phi(0{,}33) = 1 - 0{,}6293 = 0{,}3707$$



Wniosek

Rozkład średniej charakteryzuje się mniejszym odchyleniem standardowym niż rozkład badanej cechy.

Przykład

Błędy pomiarów wykonywanych dalmierzem mają rozkład normalny o odchyleniu standardowym 0,1 m. Dokonano 15 pomiarów odległości tym dalmierzem. Jakie jest prawdopodobieństwo, że odchylenie standardowe z tych pomiarów będzie większe niż 0,07 m?

Rozwiązanie

Statystyka: $\frac{15S^2}{0{,}1^2}$ ma rozkład chi kwadrat z

15 – 1 = 14 stopniami swobody

Zatem

$$P(S > 0{,}07) = P(S^2 > 0{,}0049) = P\left(\frac{15S^2}{0{,}1^2} > \frac{15 \cdot 0{,}0049}{0{,}1^2}\right) =$$

$$= P\left(Y_{14} > 7{,}35\right) \approx 0{,}91$$

Przykład

X, Y dochody (setki zł) pracowników w firmach *A* i *B*. Zakładamy, że $X - N(23,4)$, $Y - N(25, 3)$. Oblicz prawdopodobieństwo, że średni dochód 64 wylosowanych pracowników firmy *A* jest większy niż średni dochód 36 wylosowanych pracowników firmy *B*.

Rozwiązanie

Statystyka: $\overline{X}_{64} - \overline{Y}_{36}$ ma rozkład $N\left(23-25, \sqrt{\frac{4^2}{64}+\frac{3^2}{36}}\right)$,

zatem

$$P(\overline{X}_{64} > \overline{Y}_{36}) = P(\overline{X}_{64} - \overline{Y}_{36} > 0) = P\left(\frac{\overline{X}_{64} - \overline{Y}_{36} - (23-25)}{\sqrt{\frac{16}{64}+\frac{9}{36}}} > \frac{-(23-25)}{\sqrt{\frac{16}{64}+\frac{9}{36}}}\right) =$$

$$= 1 - P(Y \le 2{,}86) = 1 - \Phi(2{,}86) = 1 - 0{,}9979 = 0{,}002$$

Zatem szansa, że średni dochód 64 wylosowanych pracowników firmy A jest większy niż średni dochód 36 wylosowanych pracowników firmy B jest znikomo mała.

L.Kowalski 28.09.2017