

Podstawy statystyki

W opracowaniu statystycznym wyników badań można wyróżnić dwa etapy:

- zbieranie wyników pomiarów,
- stawianie wniosków wynikających z analizy badań.

Poprawnie przeprowadzone opracowanie statystyczne powinno uwzględniać następujące elementy:

- estymację parametrów badanych cech w populacji na podstawie wyników uzyskanych z badanej próby,
- weryfikację czyli testowanie hipotez statystycznych dotyczących rozkładu badanej cechy w analizowanej próbie,
- modelowanie zależności w populacji pomiędzy cechami czyli określenie ich wzajemnych korelacji i regresji.

W literaturze krajowej można znaleźć wiele pozycji dotyczących podstaw rachunku prawdopodobieństwa, wydaje się jednak, że najbardziej przydatne są prace [2], [3], [4], które podają w sposób przystępny podstawy statystyki, przykłady ich zastosowań oraz najczęściej wykorzystywane tablice statystyczne. Do opracowań statystycznych przydatne są również algorytmy obliczeniowe dostępne w eksploatowanych systemach komputerowych. Sposoby zastosowania analizy statystycznej w badaniach ruchu drogowego można znaleźć w [1],[5]. Poniżej podano podstawowe pojęcia statystyki matematycznej i najczęściej stosowane rozkłady zmiennej losowej w badaniach drogowych oraz przykłady ich zastosowania.

LITERATURA

- [1] Benjamin J., Cornel C. A., Rachunek prawdopodobieństwa, statystyka matematyczna I teoria decyzji dla inżynierów, WNT, Warszawa 1977.
- [2] Greń J., Statystyka matematyczna, modele i zadania. PWN, Warszawa 1981.
- [3] Praca zbiorowa, Pomiary i badania ruchu drogowego, WKiŁ, Warszawa 1984.
- [4] Smirnow N. W., Dunin-Borkowski I. W., Kurs rachunku prawdopodobieństwa i statystyki matematycznej, PWN, Warszawa 1969.
- [5] Orczykowska Z., Wykorzystanie wyników pomiarów w projektowaniu dróg. Problemy Projektowania Dróg i Mostów. Rok 1987, Nr 3.

Podstawowe pojęcia statystyki matematycznej i rachunku prawdopodobieństwa

- **Populacja generalna** - zbiór mający przynajmniej jedną właściwość wspólną dla wszystkich jego elementów, kwalifikującą je do tego zbioru oraz przynajmniej jedną właściwość, ze względu na którą mogą różnić się elementy tego zbioru.
- **Próbka** - podzbiór populacji podlegający badaniu.
- **Klasa** - najmniejszy przedział wartości badanej cechy w próbce.
- **Długość klasy** - różnica między górną i dolną granicą klasy.
- **Histogram** - graficzne przedstawienie wyników badania takie, że na osi odciętych odkłada się przedziały klasowe, natomiast na osi rzędnych odkłada się prostokąty o wysokości proporcjonalnej do liczebności badanych klas.
- **Wartość średnią** $\bar{x}$ w próbce – dla rozkładów dyskretnych wyznacza się ze wzoru:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

gdzie: $\bar{x}$ - wartość średnia,

x_k – k-ty pojedynczy wynik badania,

n – liczebność badanej próbki.

Wariancja w próbce s^2 jest średnią z kwadratów odchyłeń wartości zmiennej od wartości średniej w próbce. Dla prób o małej liczności wariancję można wyznaczyć ze wzoru

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$$

Natomiast dla dużych prób($n > 30$) wariancję wyznacza się ze wzoru

$$s^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

Odchylenie standardowe s jest miarą rozrzutu badanej zmiennej od wartości średniej, określa się je jako dodatnią wartość pierwiastka kwadratowego z wariancji w próbie.

Dla dużych prób, $n > 30$, można skorzystać z procedury upraszczającej obliczenia, jednak kosztem dokładności. Wyniki grupuje się w szereg rozdzielczy o r klasach i wtedy wartość średnią i wariancję wyznacza się z zależności:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^r \bar{x}_k n_k$$

$$s^2 = \frac{1}{n} \sum_{k=1}^r (\bar{x}_k - \bar{x})^2 n_k$$

gdzie: $\bar{x}_k$ środek k -tego przedziału klasowego,

n_k liczebność k -tego przedziału klasowego

r – liczba klas.

Liczbę klas można wyznaczyć ze wzoru $r = (0,5 - 1)\sqrt{n}$, gdzie stosuje się 0,5 dla małej liczby wartości w próbie, a 1 dla dużej liczbie elementów w próbie. W klasie powinno być minimum 5 – 7 wartości.

Współczynnik zmienności v jest względną miarą rozproszenia wyników, można go wyznaczyć z zależności

$$v = \frac{s}{\bar{x}} \cdot 100\%$$

gdzie oznaczenia są takie jak we wzorach powyżej.

Mediana jest to środkowa co do wielkości wartość zmiennej uporządkowanej według wartości niemalejąco w próbie, czyli wartość, która w 50% jest przekraczana.

Wartość modalna (moda lub dominanta) jest to najczęściej powtarzająca się wartość w badanej próbie.

Hipoteza statystyczna jest to przypuszczenie dotyczące rozkładu cechy w populacji generalnej.

Test statystyczny jest to jednoznacznie określony sposób postępowania, służący do weryfikacji hipotez statystycznych.

Błąd pierwszego rodzaju - błąd polegający na odrzuceniu hipotezy prawdziwej.

Błąd drugiego rodzaju – błąd polegający na przyjęciu hipotezy fałszywej.

Poziom istotności α - prawdopodobieństwo popełnienia błędu pierwszego rodzaju. Przyjmuje się zwykle jako bardzo małe (0,1; 0,05; 0,01; 0,001). Odrzucenie sprawdzanej hipotezy na poziomie istotności $\alpha = 0,05$ oznacza, że ryzyko popełnienia błędu pierwszego rodzaju jest równe 5%.

Przedział ufności u - zbudowany na podstawie wyników badania próbki przedział, który z określonym prawdopodobieństwem określa nieznaną wartość średnią m szacowanego parametru populacji. W rozważaniach technicznych przyjmuje się najczęściej $u = 0,9; 0,95; 0,99$.

Dla wartości średniej w populacji przedział ufności wyznacza się ze wzorów:

dla małej próby, $n < 30$ przy założeniu, że rozkład cechy w populacji jest normalny

$$\bar{x} - t_{\alpha} \frac{s}{\sqrt{n-1}} < m < \bar{x} + t_{\alpha} \frac{s}{\sqrt{n-1}}$$

dla dużej próby, $n > 30$ przy założeniu, że rozkład cechy w populacji jest normalny

$$\bar{x} - F_u \frac{s}{\sqrt{n}} < m < \bar{x} + F_u \frac{s}{\sqrt{n}}$$

gdzie:

F_u, t_{α} - wartość współczynników ufności dla określonego poziomu ufności u , F_u z tablic rozkładu normalnego standaryzowanego, t_{α} z tablic rozkładu t-Studenta, dla $n-1$ stopni swobody. Wybrane wartości współczynników podano w tabeli 1 i tabeli 2.

$\bar{x}$ - wartość średnia w próbie; m – wartość średnia w populacji; n – liczebność próbki

Tabela 1. Wybrane wartości F_u

Poziom ufności u	0,99	0,95	0,90
Wartość F_u	2,58	1,96	1,64

Tabela 2. Wybrane wartości t_α

Liczba stopni swobody u	1	5	10	15	20	30	60	∞
0,95	12,71	2,57	2,23	2,13	2,09	2,04	2,00	1,96
0,90	6,31	2,02	1,81	1,75	1,72	1,70	1,67	1,64

Funkcja rozkładu prawdopodobieństwa – funkcja określająca prawdopodobieństwo zdarzenia, że zmienna losowa przyjmie określoną wartość.

Współczynnik jednorodności k – określa związek między wartością przyjętą za minimalną na określonym poziomie ufności a wartością średnią badanej zmiennej, może być wyznaczony z zależności:

$$k = \frac{x_{\min}}{\bar{x}}$$

gdzie : $\bar{x}$ - wartość średnia

$x_{\min}$ - dolna granica zmienności badanej cechy na przyjętym poziomie ufności, którą można wyznaczyć ze wzoru

$$x_{\min} = \bar{x} - t_\alpha s$$

gdzie : t_α - wartość współczynnika ufności z tabeli rozkładu t-Studenta dla $n-1$ stopni swobody.

s - odchylenie standardowe

Badając populację ze względu na dwie lub więcej cech stosuje się rachunek korelacyjny i regresję.

Korelacja – określa siłę zależności między badanymi zmiennymi, natomiast **regresja** zajmuje się określeniem kształtu tej zależności. Po ustaleniu, że między badanymi cechami istnieje silna korelacja, przystępuje się do określenia funkcji regresji, pozwalającej przewidywać wartość jednej cechy, przy założeniu, że druga cecha przyjęła określoną wartość.

Miernikiem korelacji jest **współczynnik korelacji R** , który może przyjmować wartości z przedziału $[-1, 1]$., Gdy $|R| = 1$, to między zmiennymi istnieje zależność w

postaci funkcji liniowej.. Gdy $R = 0$, wtedy badane zmienne są nieskorelowane. Im $|R|$ jest bliższe jedności, tym zależność między badanymi zmiennymi jest mocniejsza.

Regresja liniowa – występuje w przypadku liniowej zależności między badanymi zmiennymi i opisuje ją funkcja liniowa np.

$$y = ax + b$$

gdzie: x, y - wartości pomierzone

a, b - współczynniki regresji

Regresja krzywoliniowa – występuje w przypadku nieliniowej zależności między badanymi zmiennymi i opisuje ją funkcja nieliniowa np.

$$y = ax^b$$

gdzie oznaczenia jak we wzorze powyżej.

Określenie liczebności badanej próby

Minimalna liczebność próby zależy od:

- dopuszczalnego błędu oceny badanej cechy, np. wartości średniej,
- żądanego stopnia pewności, że dopuszczalny błąd nie zostanie przekroczony.

Wyznaczoną minimalną wielkość próby należy porównać z możliwościami praktycznymi. Duża wymagana liczebność próby może się wiązać z kosztem lub z czasem jej pozyskania. Można wtedy zwiększyć dopuszczalny błąd oceny lub zmniejszyć poziom ufności.

Aby wyznaczyć liczebność próby przy założeniu, że cecha w populacji generalnej ma rozkład normalny losuje się niewielką próbę o liczności n_0 i wyznacza się dla niej wariancję ze wzoru:

$$s^2 = \frac{1}{n_0 - 1} \sum_{k=1}^{n_0} (x_k - \bar{x})^2$$

oraz minimalną liczebność próby n na wymaganym poziomie ufności u ze wzoru:

$$n = \frac{t_{\alpha}^2 s^2}{b^2}$$

gdzie: oznaczenia jak we wzorach powyżej a b jest dopuszczalnym błędem oceny. Zwykle przyjmuje się wartość dopuszczalnego błędu μ jako 1, 5 lub 10 % wartości szacowanej. Czyli $b = \mu \bar{x}$.

Przykład 1.

Na podstawie 10 wyników wstępnych pomiarów rozkładu czasu postoju autobusów na przystanku, określić potrzebną liczebność pomiarów, zakładając 90% pewności, że błąd oceny wartości średniej nie przekroczy $\mu = 5\%$. Pomierzone czasy postoju autobusów na przystanku: $t_i = 25, 11, 16, 13, 17, 19, 16, 16, 12, 15$ sekund, $n_0 = 10$.

Wyznaczenie wartości średniej:

$$\bar{t} = \frac{1}{n_0} \sum_{i=1}^{n_0} t_i = \frac{160}{10} = 16s$$

Wyznaczenie wariancji:

$$s^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (t_i - \bar{t})^2 = \frac{142}{9} = 15,78s^2$$

Wyznaczenie minimalnej liczebności próby, zapewniającej prawdopodobieństwo równe 0,9, że błąd oceny wartości średniej nie przekroczy 5%.

Dla 9 stopni swobody (szacujemy jeden parametr a liczebność próby wynosi 10) i poziomu ufności równego 0,9 odczytany współczynnik ufności $t_\alpha = 1,83$ (stosujemy interpolację liniową).

Dopuszczalny błąd oceny wartości średniej jest równy

$$b = \mu \bar{t} = 0,05 \cdot 16 = 0,8s$$

Zatem minimalna liczebność próby powinna wynosić:

$$n = \frac{t_\alpha^2 s^2}{b^2} = \frac{1,83^2 \cdot 15,78}{0,8^2} = 82,57$$

Należy przyjąć liczebność próby na $n = 83$ pomiary.

Zatem należy dodatkowo określić czas postoju dla 73 autobusów i powtórnie wyznaczyć szukane parametry.

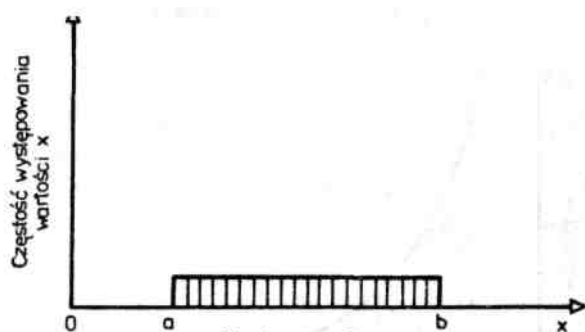
Przykłady rozkładów zmiennych losowych

Poniżej podano charakterystyki rozkładów zmiennych losowych wykorzystywane w badaniach prowadzonych w inżynierii ruchu.

Rozkład równomierny - powszechnie stosowany w modelowaniu - jest to rozkład prawdopodobieństwa zmiennej losowej ciągłej o funkcji gęstości określonej wzorem

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{dla } a \leq x \leq b \\ 0 & \text{dla } x < a \text{ oraz } x > b \end{cases}$$

Na rysunku zobrazowano graficznie funkcję gęstości prawdopodobieństwa rozkładu równomiernego.



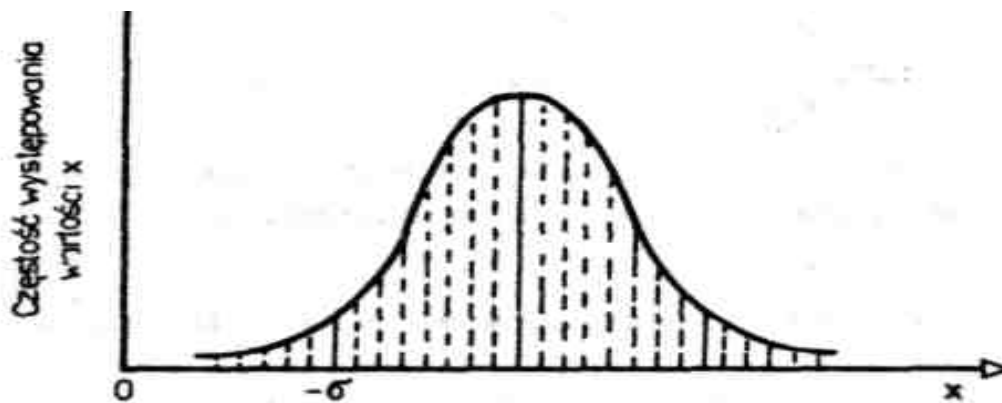
Wynik pomiarów

Wykres funkcji gęstości prawdopodobieństwa rozkładu równomiernego

Rozkład normalny jest ciągłym i symetrycznym rozkładem, często stosowanym w modelowaniu i symulacji komputerowej do opisu prędkości pojazdów. Funkcja gęstości prawdopodobieństwa rozkładu normalnego o wartości średniej m i wariancji σ^2 dana jest wzorem

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(x-m)^2}{2\sigma^2} \right], \text{ dla } -\infty < x < +\infty$$

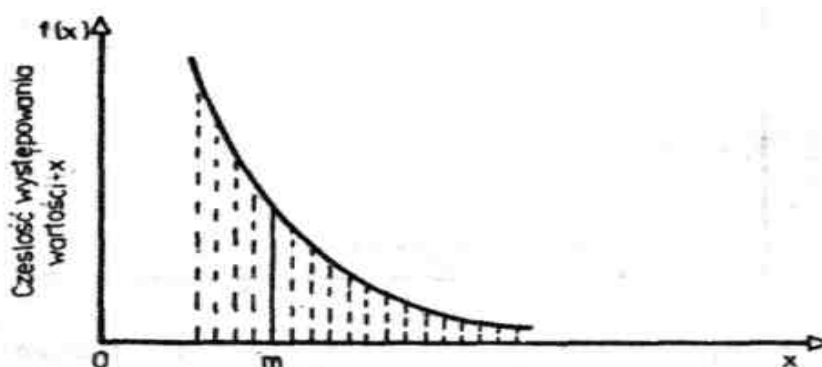
Wykres funkcji gęstości prawdopodobieństwa rozkładu normalnego przedstawiono na rysunku poniżej.



Wykres funkcji gęstości prawdopodobieństwa rozkładu normalnego

Rozkład wykładniczy jest rozkładem ciągłym, często stosowanym do aproksymacji odstępów czasu między zgłoszeniami pojazdów. Funkcja gęstości rozkładu wykładniczego o wartości średniej m dana jest wzorem:

$$f(x) = \begin{cases} \frac{1}{m} \exp\left(-\frac{x}{m}\right) & \text{dla } x \geq 0 \\ 0 & \text{dla } x < 0 \end{cases}$$

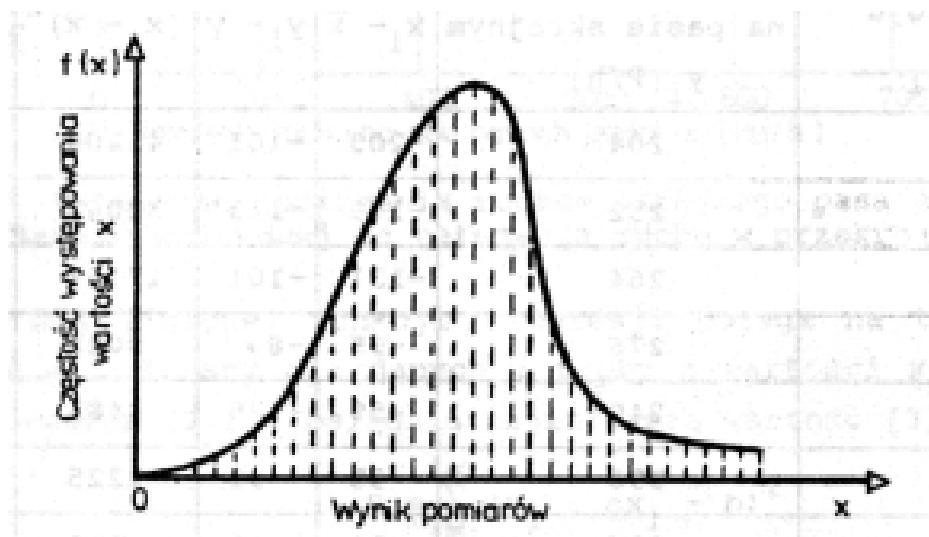


Wykres funkcji gęstości rozkładu wykładniczego

Rozkład gamma jest rozkładem ciągłym o funkcji gęstości prawdopodobieństwa określonej zależnością:

$$f(x) = \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}, \text{ dla } 0 < x < \infty, \lambda > 0$$

Rozkład gamma można zastosować np. do aproksymacji miejsc zjazdu pojazdów z drogi głównej na pas wyłączenia.



Wykres funkcji gęstości prawdopodobieństwa rozkładu Gamma

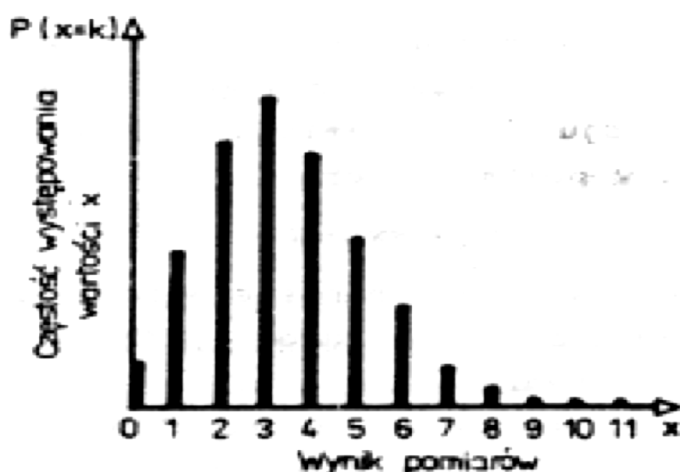
Rozkład Poissona jest rozkładem dyskretnym o funkcji gęstości prawdopodobieństwa określonej zależnością:

$$P(x=k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ dla } \lambda > 0, k = 0, 1, 2, \dots$$

gdzie: $P(x=k)$ - prawdopodobieństwo, że zmienna losowa x przyjmie wartość k ,

λ - wartość średnia w rozkładzie.

Na poniższym rysunku przedstawiono wykres funkcji gęstości rozkładu Poissona.



Wykres funkcji gęstości rozkładu Poissona.

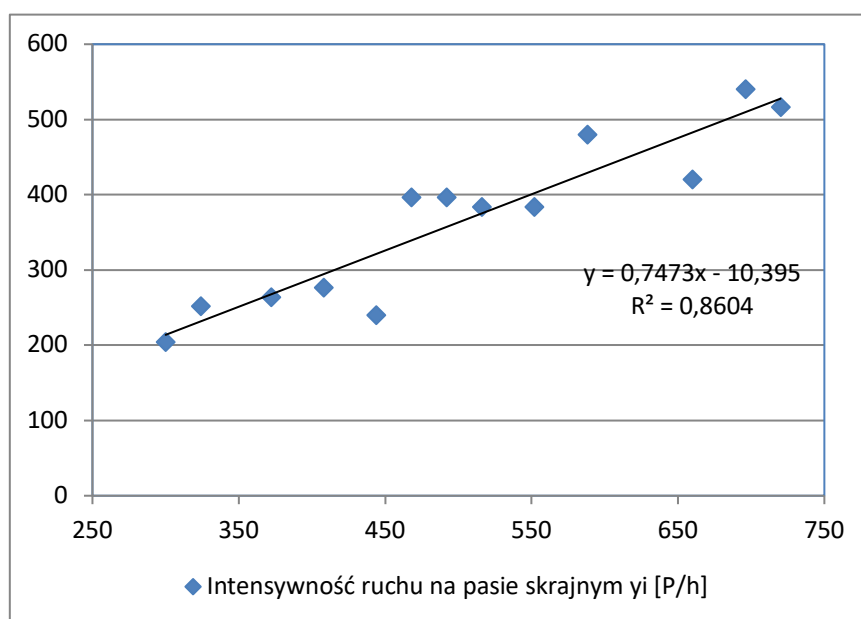
Przykład 2

W celu określenia zależności obciążenia skrajnego pasa ruchu drogi jednokierunkowej dwupasowej od intensywności ruchu, w przekroju drogi przeprowadzono pomiar intensywności ruchu w przedziałach 5 minutowych na poszczególnych pasach jezdni. Otrzymane wyniki przedstawiono w tabeli 3.

Tabela 3.

Lp	Intensywność ruchu w przekroju drogi x_i [P/h]	Intensywność ruchu na pasie skrajnym y_i [P/h]	$x_i - x_{\text{śr}}$	$y_i - y_{\text{śr}}$	$(x_i - x_{\text{śr}})^2$	$(x_i - x_{\text{śr}})(y_i - y_{\text{śr}})$
1	300	204	-203,08	-161,54	41240,24	32804,73
2	324	252	-179,08	-113,54	32068,54	20332,12
3	372	264	-131,08	-101,54	17181,16	13309,35
4	408	276	-95,08	-89,54	9039,62	8513,04
5	444	240	-59,08	-125,54	3490,08	7416,43
6	468	396	-35,08	30,46	1230,39	-1068,50
7	492	396	-11,08	30,46	122,70	-337,42
8	516	384	12,92	18,46	167,01	238,58
9	552	384	48,92	18,46	2393,47	903,20
10	588	480	84,92	114,46	7211,93	9720,43
11	660	420	156,92	54,46	24624,85	8546,27
12	696	540	192,92	174,46	37219,31	33657,66
13	720	516	216,92	150,46	47055,62	32638,58
Średnia	503,08	365,54		Suma	223044,92	166674,46

Na rysunku przedstawiono graficznie uzyskane wyniki pomiarów.



Zależność intensywności ruchu na pasie skrajnym od intensywności ruchu w przekroju drogi.

Z rysunku wynika, że w badanym zakresie można przyjąć zależność liniową między intensywnością ruchu w przekroju drogi i na pasie skrajnym drogi. Przyjęto zatem funkcję regresji w postaci:

$$y = \alpha x + \beta$$

gdzie:

y – intensywność ruchu na pasie skrajnym

x – intensywność ruchu w przekroju drogi

α, β współczynniki regresji.

Aby wyznaczyć współczynniki regresji określono ich estymatory a i b z równania

$$S = \sum_{i=1}^k (y_i - ax_i - b)^2$$

Metoda najmniejszych kwadratów dla liniowej funkcji regresji polega na takim oszacowaniu parametrów a i b , aby dla danych z próby n wartości x_i oraz y_i funkcja określona powyższym wzorem osiągała najmniejszą wartość. Funkcja S jest funkcją dwóch zmiennych a i b . Stosując warunki istnienia ekstremum funkcji dwóch zmiennych:

$$\frac{\partial S}{\partial a} = 0 \rightarrow a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^k (x_i - \bar{x})^2}$$

$$\frac{\partial S}{\partial b} = 0 \rightarrow b = \bar{y} - a\bar{x}$$

Gdzie $\bar{x}$ z kreską i $\bar{y}$ z kreską oznaczają odpowiednio średnie z próby dla serii x i Y .

Po odpowiednich obliczeniach otrzymuje się wartości:

$$a = 0,7473 \text{ i } b = -10,395$$

Zatem poszukiwana funkcja regresji jest postaci:

$$y = 0,7473x - 10,395$$

Gdzie y intensywność ruchu w pasie skrajnym, a x intensywność ruchu na obu pasach jezdni. Uzyskana zależność jest prawidłowa w badanym zakresie intensywności ruchu, tzn. od 300 do 700 P/h.