Elementarna statystyka

Alexander Bendikov

Uniwersytet Wrocławski

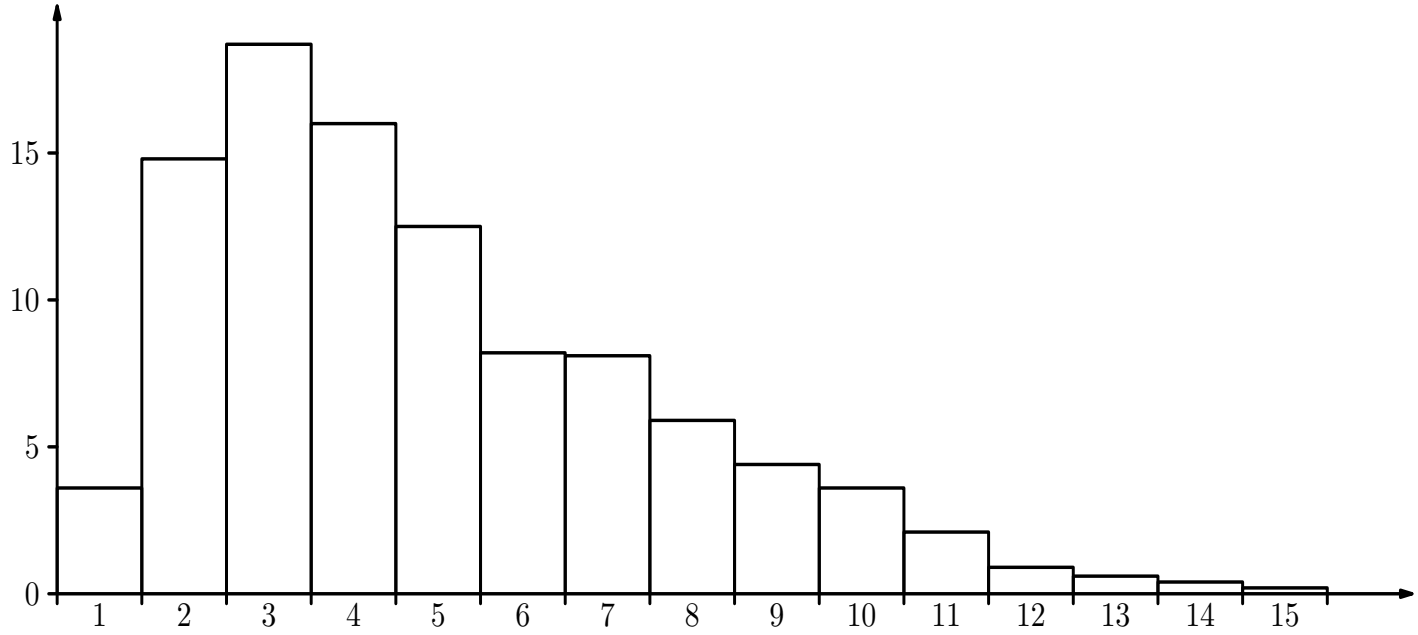
Semestr wiosenny 2017

Elementarna analiza danych

- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.
- *Zmienne kategorialne* zaliczają każdy przypadek do pewnej *kategorii*: mężczyzna/kobieta, status społeczny itp.
- *Zmienne ilościowe* podają pewną charakterystykę numeryczną, wysokość, wagę, pensję itp.
- *Badawcza analiza danych* stosuje wykresy i parametry liczbowe aby opisać zmienne oraz zależności pomiędzy zmiennymi w zbiorze danych.
- *Rozkład* zmiennej ilościowej opisuje wartości zmiennej i ich częstości.

Zmienne ilościowe: histogramy

1. Rozkładamy zakres wartości zmiennej na przedziały równej długości
2. Zliczamy ilość przypadków z wartościami w poszczególnych przedziałach
3. Rysujemy prostokąty: podstawa każdego pokrywa kolejny przedział, wysokość to zliczona ilość przypadków w danym przedziale.



Rysunek: Histogram zmiennej ilościowej „Długość słów”

Długość	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Odsetek	3,6	14,8	18,7	16,0	12,5	8,2	8,1	5,9	4,4	3,6	2,1	0,9	0,6	0,4	0,2

Tablica: Długość słów w czasopiśmie *Popular Science*

Opis rozkładu przy pomocy parametrów

- $X = x_1, x_2, \ldots, x_n$ (uporządkowane rosnąco)
- Średnia $\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots, x_n)$
- Mediana $M = \begin{cases} x_{(n+1)/2} & n \text{ nieparzysta} \\ \frac{1}{2}(x_{1/2} + x_{n/2+1}) & n \text{ parzysta.} \end{cases}$

Uwagi: 1) Dla $n \gg 1$, $\overline{x} \approx E(X)$ z prawa wielkich liczb,
2) Dla niewielkich n średnia $\overline{x}$ istotnie zależy od obserwacji nietypowych (outlier), natomiast mediana M nie zależy.

- Kwartyle Q_1 i Q_3:

$$Q_1 = M(x_1, \ldots, x_k;\ k = \lfloor \tfrac{n+1}{2} \rfloor),$$
$$Q_3 = M(x_m, \ldots, x_n;\ m = \lceil \tfrac{n+1}{2} \rceil).$$

Przykłady:

- $9 \quad 9 \quad 22 \quad 32 \quad 33 \quad 39 \quad 40 \quad 42 \quad 49 \quad 52 \quad 52 \quad 70$

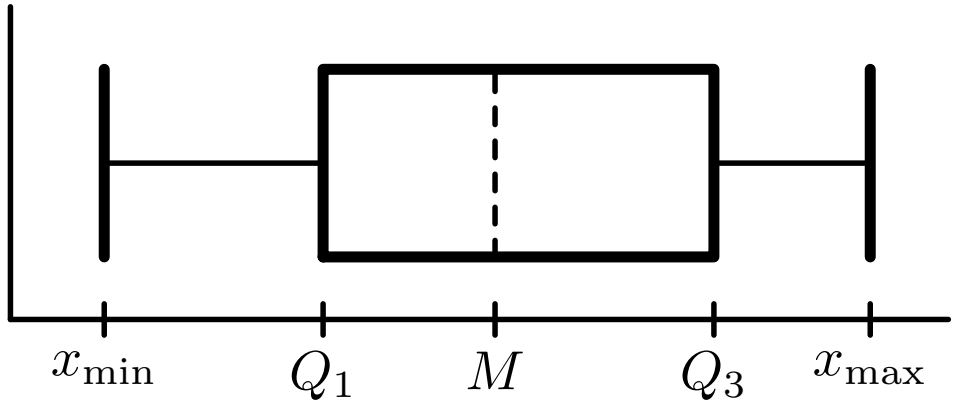
  $Q_1 = 27 \quad M = 39,5 \quad Q_3 = 50,5$

- $22 \quad 25 \quad 34 \quad 35 \quad 41 \quad 41 \quad 46 \quad 46 \quad 46 \quad 47 \quad 49 \quad 54 \quad 54 \quad 55 \quad 60$

  $Q_1 \qquad\qquad M \qquad\qquad Q_3$

- 5 - liczbowe podsumowanie: $x_{\min} - Q_1 - M - Q_3 - x_{\max}$.

- Wykres pudełkowy:



- Rozrzut: odchylenie standardowe próbki s:

$$X \quad : \quad x_1, x_2, \ldots, x_n,$$

$$s^2 = \frac{1}{n-1}\left((x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right) = \frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}\overline{x}^2.$$

Własności:

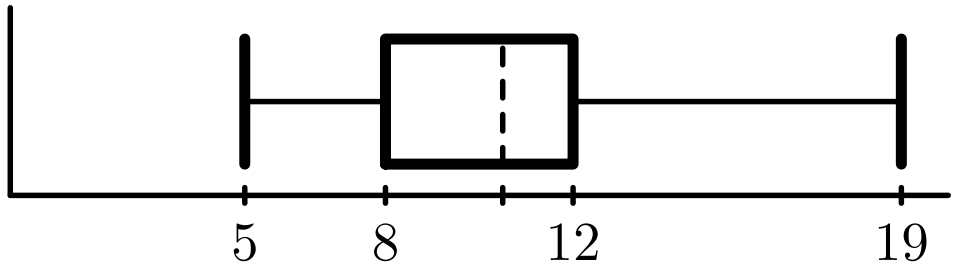
- $E(\overline{x}) = E(X)$,
- $E(s^2) = Var(X)$.

Przykład: $X: \ 5 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 12 \quad 15 \quad 19$

$Q_1 \qquad M \qquad Q_3$

1) $\min = 5, \ \max = 19, \ M = \frac{1}{2}(10+11) = 10,5, \ Q_1 = 8, \ Q_3 = 12$

2)

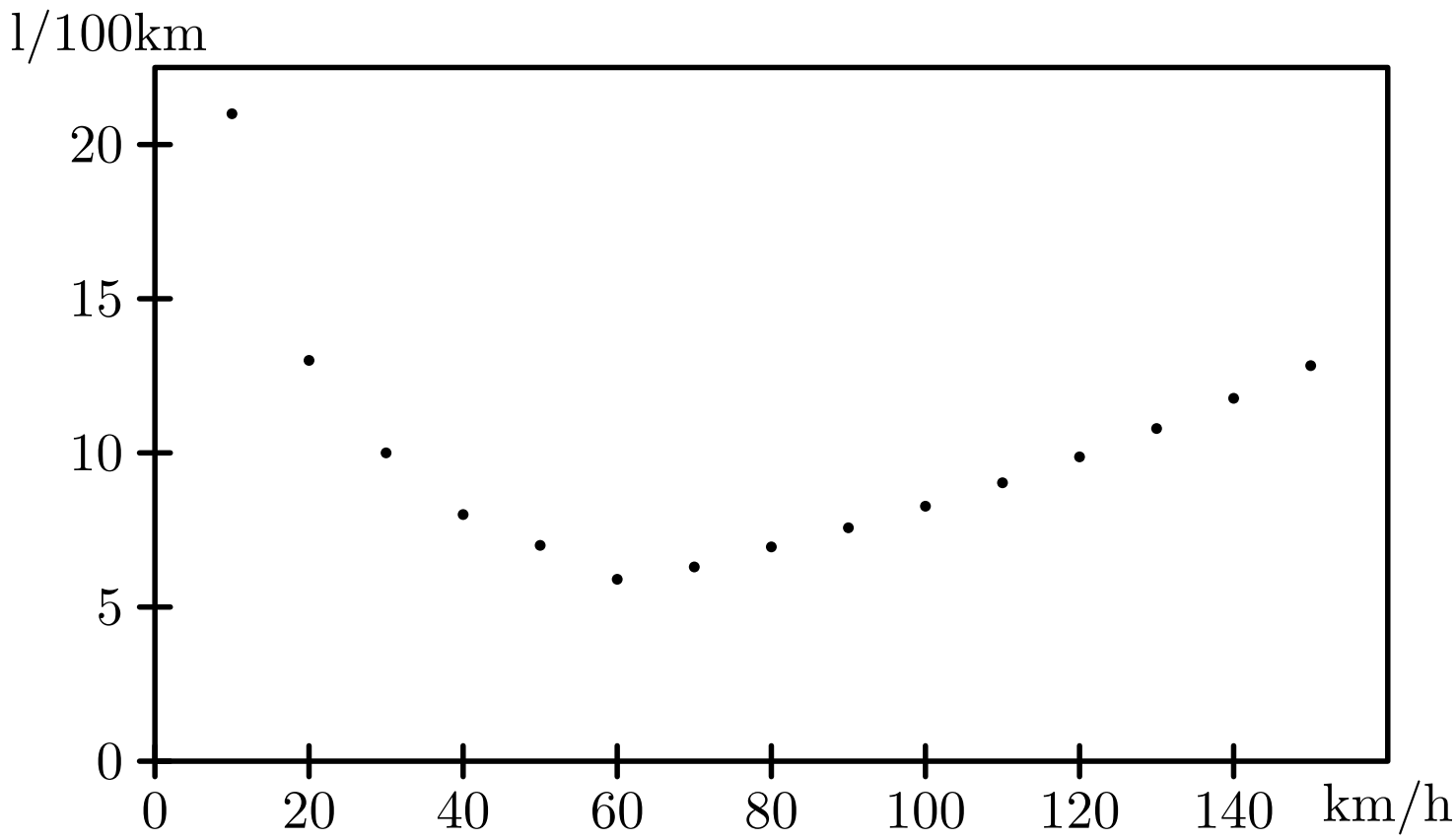


3) $\overline{x} = \frac{1}{10}(5+7+\cdots+19) = 10,8 > M, \ s^2 = 16,4, \ s = 4,06.$

Elementarna analiza korelacji

Czy jadąc szybko marnujemy paliwo? W tabeli są dane dotyczące zużycia paliwa (brytyjska wersja Forda Escorta)

Prędkość (km/h)	Zużycie paliwa (l/100 km)	Prędkość (km/h)	Zużycie paliwa (l/100 km)
10	21,00	90	7,57
20	13,00	100	8,27
30	10,00	110	9,03
40	8,00	120	9,87
50	7,00	130	10,79
60	5,90	140	11,77
70	6,30	150	12,83
80	6,95		



Rysunek: Wykres punktowy zmiennych „Prędkość” i „Zużycie paliwa”

W przykładzie są dwie zmienne zależne:
X (prędkość), *zmienna objaśniająca*, która jest zmienną decydującą w tej zależności,
Y (zużycie paliwa), *zmienna zależna*, która jest zmienną reagującą.
Główne zadanie to objaśnienie rodzaju zależności

$$X \longleftrightarrow Y.$$

Wykres punktowy pokazuje zależność pomiędzy dwoma zmiennymi ilościowymi X i Y. Poszczególne obserwacje zbioru danych odpowiadają punktom wykresu.

Współczynnik korelacji $R_{X,Y}$

Niech X i Y będą zmiennymi losowymi, ze średnimi i odchyleniami standardowymi odpowiednio $\mu_X, \sigma_X, \mu_Y, \sigma_Y$. Jeżeli X i Y są *niezależne* to

$$Var(X+Y) = Var(X) + Var(y) = \sigma_X^2 + \sigma_Y^2.$$

Jeżeli X i Y nie są niezależne, to

$$Var(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2 \cdot \sigma_X \cdot \sigma_Y \cdot R_{X,Y},$$

gdzie

$$R_{X,Y} = E\Big(\frac{X-\mu_X}{\sigma_X}\Big)\Big(\frac{Y-\mu_Y}{\sigma_Y}\Big).$$

Wielkość $R_{X,Y}$ nazywamy *współczynnikiem korelacji* zmiennych X i Y

Własności $R_{X,Y}$

1. $-1 \leq R_{X,Y} \leq 1$,
2. $R_{X,Y} = \pm 1 \Leftrightarrow X, Y$ są *liniowo zależne*, to znaczy

$$Y = k\,X + b, \quad \text{lub} \quad X = k\,Y + b.$$

   W takim przypadku mamy dodatkowo $R_{X,Y} = 1$ jeżeli $k > 0$ i $R_{X,Y} = -1$ jeżeli $k < 0$.
3. Współczynnik korelacji mierzy siłę współzależności typu *liniowego*. Nie opisuje dobrze zależności krzywoliniowych.

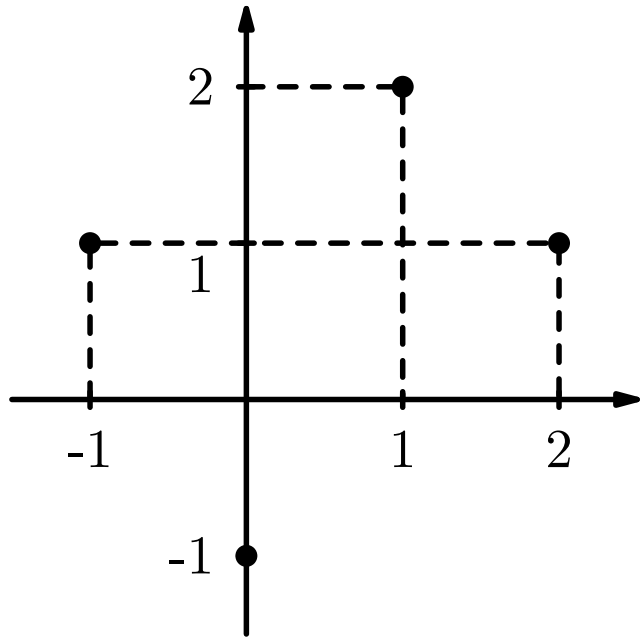
Współczynnik korelacji w próbie $r_{X,Y}$

Załóżmy, że mamy próbki $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ pobrane z populacji o rozkładach X i Y odpowiednio. Możemy korzystać z przybliżeń $\overline{x} \approx \mu_x$, $s_x \approx \sigma_x$, $\overline{y} \approx \mu_y$, $s_y \approx \sigma_y$.
A w jaki sposób możemy przybliżyć współczynnik korelacji $R_{X,Y}$?

- $r_{X,Y} = \frac{1}{n-1} \sum_i (\frac{x_i - \overline{x}}{s_x})(\frac{y_i - \overline{y}}{s_y})$,
- $r_{X,Y} \approx R_{X,Y}$ dla $n \gg 1$.

Przykład:

$x.$	$y.$	$x. - \overline{x}$	$y. - \overline{y}$
-1	1	-1,5	0,25
0	-1	-0,5	-1,75
1	2	0,5	1,25
2	1	1,5	0,25
$\overline{x} = 0,5$	$\overline{y} = 0,75$	$s_x = 1,3$	$s_y = 1,2$



Rysunek: Wykres punktowy zmiennych X i Y

$$\sum (x_\cdot - \overline{x})(y_\cdot - \overline{y}) = 1,5$$

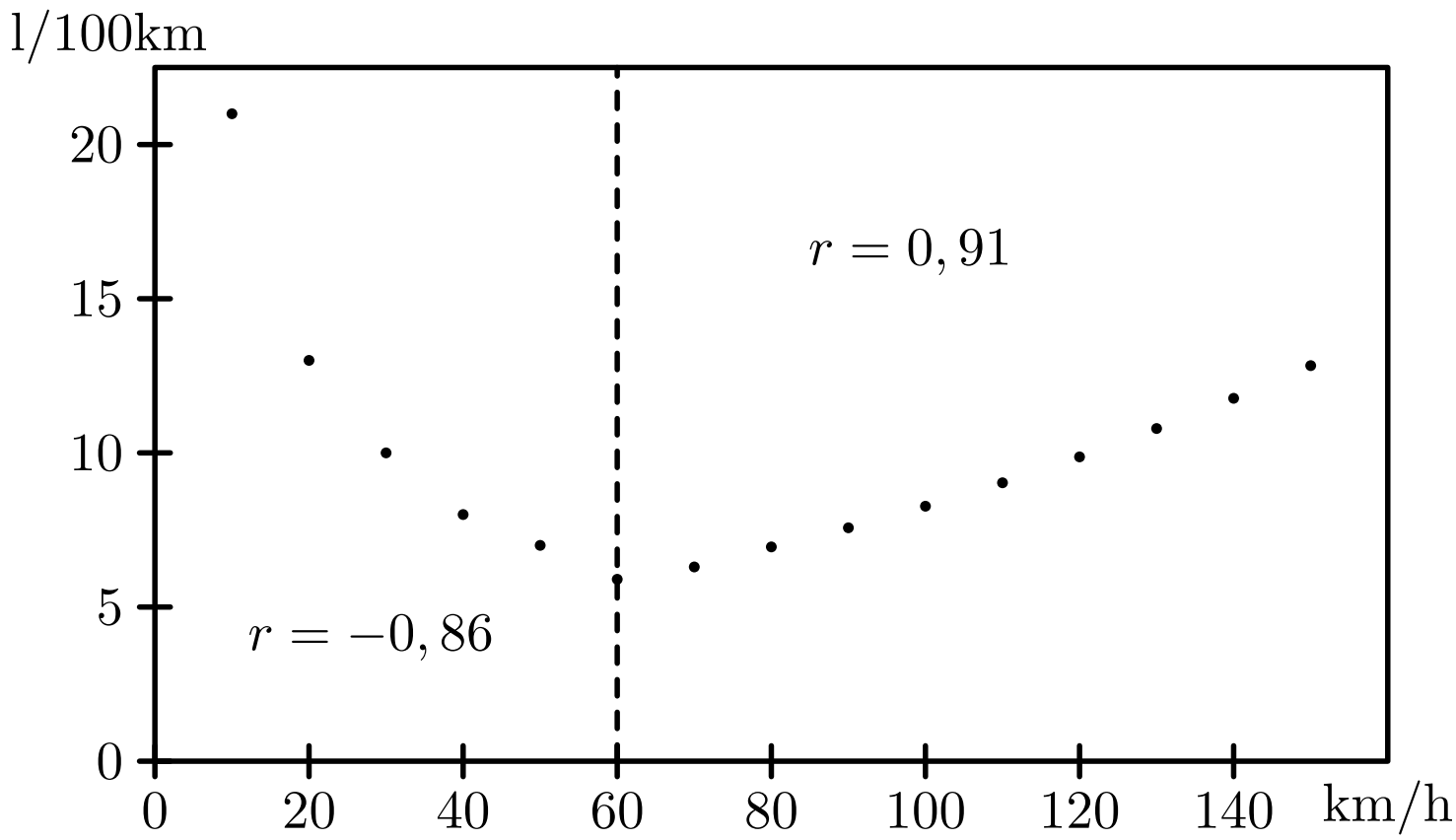
$$r_{X,Y} = 0,32.$$

Własności $r_{X,Y}$

1. $-1 \leq r \leq 1$,
2. $r = \pm 1$ wtedy i tylko wtedy, gdy wszystkie obserwacje leżą na jednej prostej. Czyli $r = \pm 1$ tylko w przypadku idealnie liniowej zależności.
3. $r \approx 0$ oznacza bardzo słaba zależność liniową.

Przykład: W przypadku zużycia paliwa mamy:

- 1 zakres, $10 - 60$ km/h $r = -0,86$
- 2 zakres, $60 = 150$ km/h $r = 0,91$
- W całym zakresie prędkości $10 - 150$km/h mamy $r = -0,15$ - bardzo słaba zależność liniowa



Rysunek: Wykres punktowy, 2 zakresy

Przykłady związane z korelacją

1. Galton (1857) 1078 par pomiarów wzrostów:
   - Ojcowie i synowie: $r \approx 0,5$,
   - Matki i synowie: $r \approx 0,494$.
2. Badania związane z ochroną zdrowia (1960-62)
   - Wzrosty i wagi 411 mężczyzn w wieku 18-24 lat:

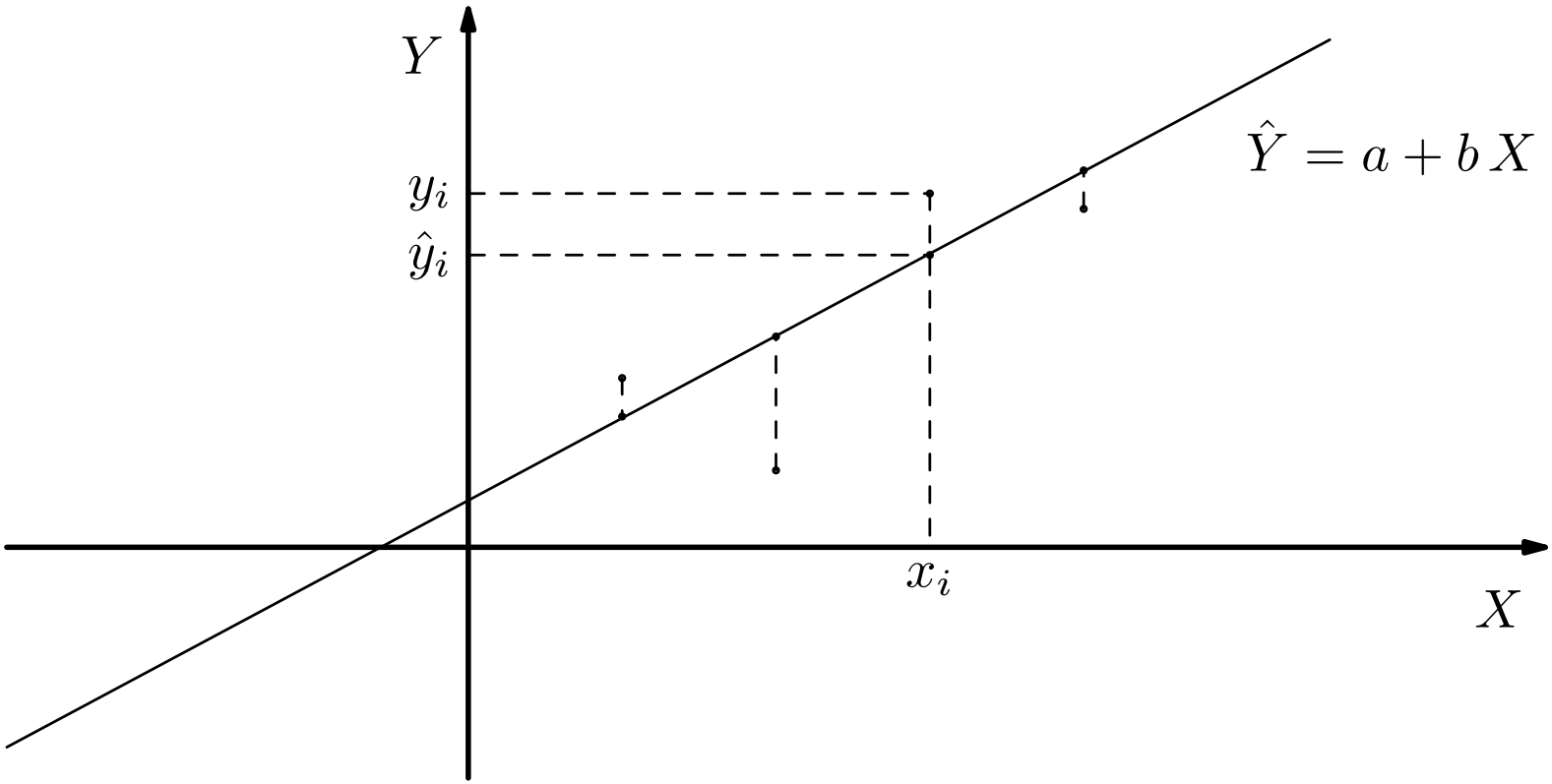
$$r \approx 0,36.$$

   - Wykształcenie i dochód:
     (a) dla mężczyzn w wieku 25-34: $r \approx 0,4$,
     (b) dla mężczyzn w wieku 35-44: $r \approx 0,6$.
3. Iloraz inteligencji identycznych bliźniaków:

$$r \approx 0,95.$$

Linia regresji najmniejszych kwadratów

Zasada: linia regresji najmniejszych kwadratów zmiennych X i Y jest prostą o równaniu $\hat{Y} = a + b\,X$ dla której suma kwadratów $\sum(y_\cdot - \hat{y}_\cdot)^2$ jest najmniejsza



Rysunek: Linia regresji najmniejszych kwadratów

Musimy więc znaleźć a, b takie, że

$$\sum (y_i - \hat{y}_i)^2 \longrightarrow \min,$$

gdzie

- y_i jest obserwacją zmiennej Y,
- $\hat{y}_i = a + b\, x_i$ jest przewidywaną wartością zmiennej Y, odpowiadającą obserwacji x_i zmiennej X
- $y_i - \hat{y}_i$ jest resztą.

Rozwiązanie problemu minimalizacji: Linia regresji najmniejszych kwadratów ma równanie

$$\hat{Y} = a + bX,$$

gdzie:

1. Współczynnik kierunkowy $b = r \cdot \frac{s_Y}{s_X}$
2. Odsunięcie $a = \overline{y} - b\overline{x}$

Równoważnie:

$$\frac{\hat{Y} - \overline{y}}{s_Y} = r \cdot \frac{X - \overline{x}}{s_X}.$$

Przykład: Dla naszego Forda Escorta mamy:

1. W zakresie $10 - 60$ km/h mamy

$$\hat{Y} = -0,3X + 21,5,$$

a więc następujące prognozy: $x = 25, \hat{y} = 14$, $x = 40, \hat{y} = 9,5$, $x = 50, \hat{y} = 6,5$, $x = 70, \hat{y} = 0,5$?!

2. W całym zakresie $10 - 150$ km/h mamy

$$\hat{y} = -0,01466\ X + 11,058,$$

a więc następujące prognozy: $x = 25, \hat{y} = 10,65$, $x = 40, \hat{y} = 9,32$, $x = 70, \hat{y} = 10,03$!

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

Powyższy wzór łatwo jest uzasadnić korzystając z równania regresji

$$\frac{\hat{y}_\cdot - \overline{y}}{s_Y} = r \cdot \frac{\hat{x}_\cdot - \overline{x}}{s_X}.$$

1) $\sum \frac{\hat{y}_\cdot - \overline{y}}{s_Y} = r \cdot \sum \frac{\hat{x}_\cdot - \overline{x}}{s_X} = 0 \;\Rightarrow\; \overline{\hat{y}} = \overline{y}$

2) $\frac{1}{n-1} \sum \left(\frac{\hat{y}_\cdot - \overline{\hat{y}}}{s_Y} \right)^2 = r^2 \frac{1}{n-1} \sum \left(\frac{\hat{x}_\cdot - \overline{x}}{s_X} \right)^2 = r^2$

W końcu,

$$r^2 = \frac{\sum (\hat{y}_\cdot - \overline{\hat{y}})^2}{\sum (y_\cdot - \overline{y})^2}.$$

Innymi słowy, r^2 to procent zmienności Y, który można uzasadnić linią regresji.

Planowanie próby

1. **Populacja:** to cała zbiorowość przypadków (osobników), którą badamy.
2. **Próba:** to część populacji, którą wyselekcjonowaliśmy do rzeczywistych badań, pomiarów.
3. **Planowanie próby:** to metoda, którą wybraliśmy w celu wyselekcjonowania próby z populacji.

Przykład 1. Znana dziennikarka Ann Launders, prowadząca działy z poradami kiedyś zapytała czytelników: „Czy gdybyś miał/miała decydować ponownie, czy zdecydowałbyś/zdecydowałabyś się na dzieci?" Po kilku tygodniach znała już odpowiedź czytelników-rodziców, i swój kolejny felieton zatytułowała: „70% rodziców uważa, że nie warto „wchodzić" w dzieci". Istotnie, 70% z prawie 10 000 rodziców, którzy wysłali swoje odpowiedzi stwierdziło, że gdyby mieli możliwość ponownie zadecydować o posiadaniu dzieci, nie zdecydowaliby się.

Z punktu widzenia statystyki tak zebrane dane są bezwartościowe jako wskazówka na temat opinii ogółu amerykańskich rodziców. Ci, którzy odpowiedzieli na pytanie Ann Launders reprezentowali część populacji rodziców, która miała na ten temat silne opinie, do tego stopnia, że zadała sobie trud napisania odpowiedzi. Reprezentowali tą część populacji, która była zła na swoje dzieci. Ta część populacji nie jest (w kwestii zadowolenia z posiadania dzieci) reprezentatywna dla całej populacji amerykańskich rodziców!!!

Porządne badanie

Właściwie statystycznie zaplanowane badanie opinii w tej samej sprawie kilka miesięcy później pokazało, że 91% rodziców zdecydowałoby się ponownie na dzieci.

Wniosek: Ann Launders użyła spontanicznej odpowiedzi, jako metody budowy próby. Doprowadziło to do błędnych wniosków.

Przykład 2. Producenci i agencje reklamowe często przeprowadzają ankiety w centrach handlowych, w których pytają o zwyczaje kupujących, i starają się określić skuteczność reklam. Wybór próby spośród klientów centrum handlowego jest szybki i tani. Ale klienci przepytywanie w centrach handlowych nie stanowią reprezentatywnej próbki całej populacji konsumentów. W zależności od czasu badania można, na przykład, otrzymać nadreprezentację emerytów i nastolatków. Taki wybór próby nazywa się „próbą dogodną" która, podobnie jak próba spontanicznej odpowiedzi z reguły wykazuje obciążenie (tendencyjność), czyli wbudowany, systematyczny błąd. Taki wybór próby systematycznie promuje pewien typ odpowiedzi.

Prosta próba losowa (SRS)

- W próbie spontanicznej odpowiedzi ludzie sami wybierają, czy odpowiedzieć czy nie.
- W próbie dogodnej pytający dokonuje wyboru, kiedy dokonać badania, kogo pytać.

W obu przypadkach arbitralny wybór powoduje obciążenie, systematyczny błąd.

- Prosta próba losowa (SRS) o rozmiarze n składa się z n osobników wybranych z populacji w taki sposób, że każdy możliwy układ n osobników ma taką samą szansę bycia wybranym.
- Ideę SRS można zrealizować przy użyciu *tabeli cyfr losowych*

Cyfry losowe

- Tablica cyfr losowych jest ciągiem (z reguły długim) cyfr $0, 1, 2, \ldots, 9$ o następujących własnościach:
  1) Każda pozycja w ciągu może być każdą z 10 cyfr z jednakowym prawdopodobieństwem,
  2) Poszczególne elementy ciągu są wzajemnie niezależne. Znajomość dowolnego fragmentu ciągu nie zawiera żadnej informacji o reszcie.
  3) Każda *para* kolejnych elementów ciągu z może mieć postać $00, 01, 02, \ldots, 99$ z jednakowym prawdopodobieństwem
  4) Każda *trójka* kolejnych elementów ciągu z może mieć postać $000, 001, 002, \ldots, 999$ z jednakowym prawdopodobieństwem
  5) .....

Wyselekcjonowanie prostej próby losowej sprowadza się do 2 kroków: oznakowania całej populacji cyframi i wyboru cyfr z tabeli.

Linia								
	.	.	.					
101	19223	95034	05756	28713	96409	12531	42544	07511
102	73676	47150	99400	01927	27754	42648	82425	75592
103	45467	71709	77558	00095	32863	29485	82226	47052
104	52711	38889	93074	60227	40011	85848	48767	94322
105	95592	94007	69971	91481	60779	53791	17297	65561
106	68417	35013	15529	72765	85089	57067	50211	71035
107	82739	57890	20807	47511	81676	55300	94383	43367
108	60940	72024	17868	24943	61790	90656	87964	90597
109	36009	19365	15412	39638	85453	46816	83485	53946
110	38448	48789	18338	24697	39364	42006	76688	13258
	.	.	.					

Tablica: Tabela cyfr losowych

Przykład: Lista studentów zapisanych na wykład prof. Bendikova ze statystyki elementarnej zawiera 18 nazwisk. Prof. Bendikov postanowił porozmawiać z wybraną grupą 3 studentów, żeby zasięgnąć ich opinii o wykładzie. Żeby uniknąć ryzyka, że nieświadomie wybierze swoich ulubionych studentów i w ten sposób wpłynie na wyniki ankiety, postanowił skonstruować prostą próbę losową SRS. Grupa studentów posiada już swoje numery, mogą to być numery na liście studentów. Każdemu studentowi przyporządkowana jest więc para cyfr: $01, 02, \ldots, 18$.
Aby wykorzystać tabelę cyfr losowych postępujemy następująco:
Krok 1: Każdemu studentowi przyporządkowujemy niepowtarzalny, liczbowy identyfikator, używając jak najmniejszej ilości cyfr. W tym przykładzie każdemu studentowi przyporządkowujemy dwucyfrowy numer, który ma na liście studentów zapisanych na wykład:

$$01, 02, 03, \ldots, 16, 17, 18.$$

Krok 2: W tabeli cyfr losowych wybieramy losowo miejsce. Następnie odczytujemy kolejne cyfry począwszy od tego miejsca, i dzielimy je na kolejne grupy 2 cyfr. Powiedzmy, że wybraliśmy drugą kolumnę linii 105:

$$94\ 00\ 76\ 99\ 71\ 91\ 48\ 16\ 07\ 79\ 53\ 79\ 11\ 72\ 97\ 65\ 56\ \ldots$$

Wybiera z nich pierwszą parę, która odpowiada identyfikatorowi któregoś studenta (czyli jest postaci $01, 02, 03, \ldots, 18$. Nastepnie kontynuujemy, wybierając kolejne pary. Jeżeli wybrana para została już wcześniej wybrana, lub jeżeli nie odpowiada żadnemu studentowi, odrzucamy ją, i kontynuujemy. W tym przykładzie otrzymujemy studentów:

$$16, \quad 07, \quad 11.$$

Każdy student miał jednakową szansę bycia wybranym.