ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

2/10/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Statystyczna analiza danych

- Co to są dane
- W jaki sposób zbieramy dane
- W jaki sposób reprezentujemy dane
- W jaki sposób analizujemy dane
- W jaki sposób wyciągamy wnioski

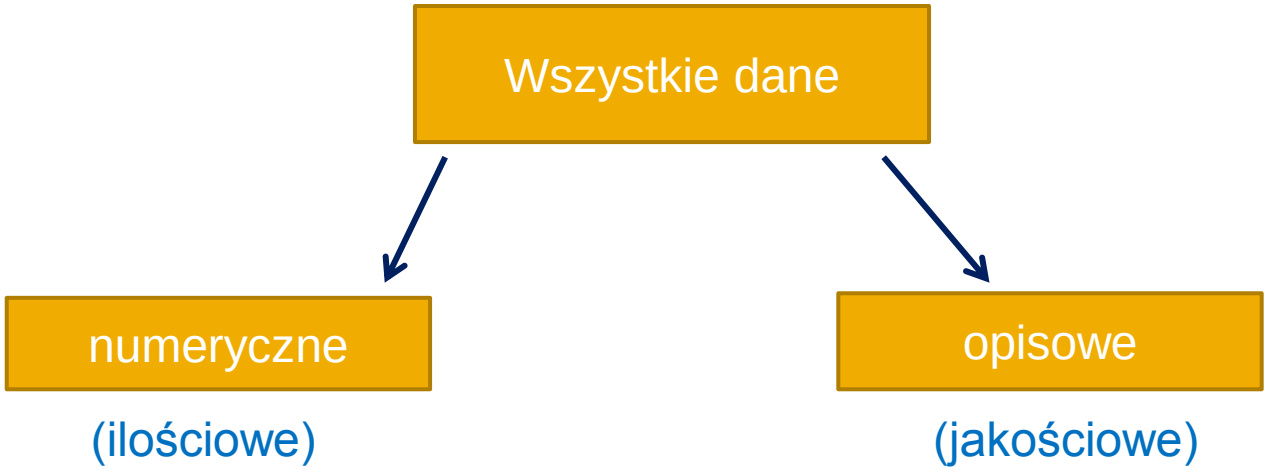
Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

obserwacja

zmienna

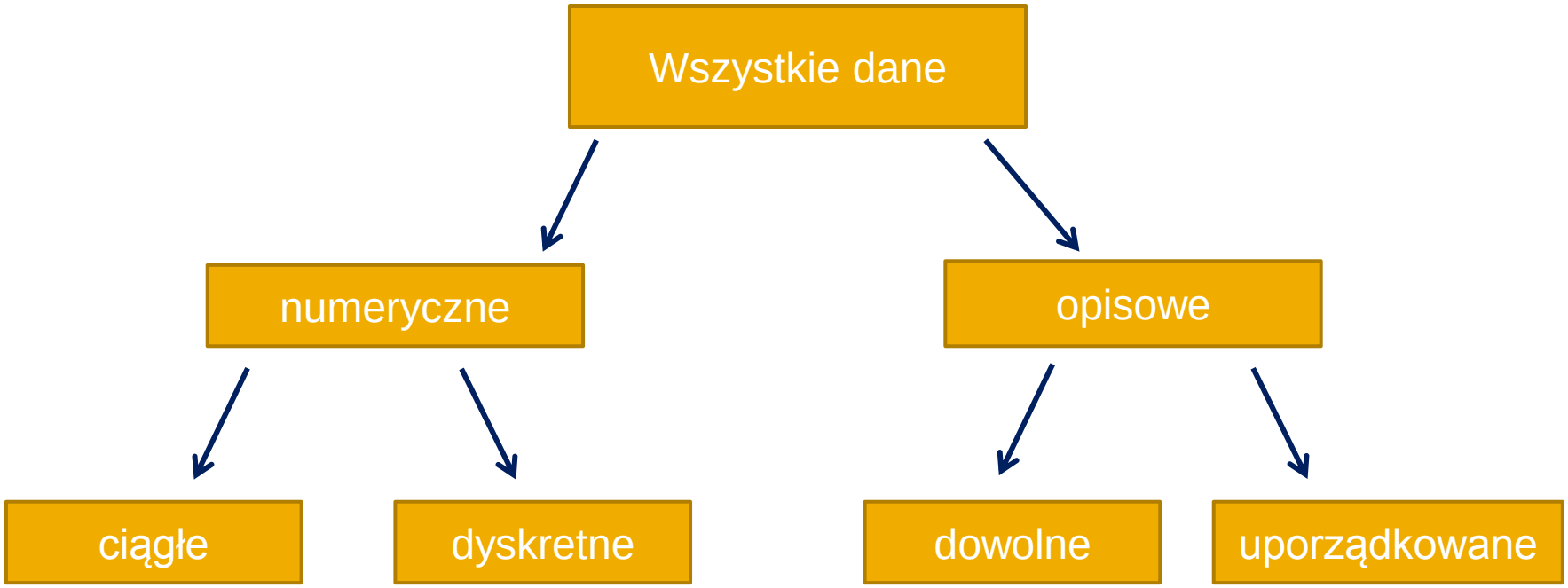
Typy danych



maja wartości liczbowe;
można dodawać, odejmować,
wyliczać średnią, itd.

Skończona ilość kategorii;
mogą być oznaczone
wartościami liczbowymi
ale nie podlegają operacjom
arytmetycznym.

Typy danych



Przyjmują dowolne wartości z jakiegoś przedziału

Przyjmują jedną z określonego zbioru wartości

Poziomy mają wewnątrzną kolejność

Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

country – nazwa kraju

opisowa

Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

↓

cr_req – ilość żądań usunięcia wpisu na www wysłany do Google

dyskretna, numeryczna

Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

cr_comply – procent żądań zaakceptowany przez Google

ciągła, numeryczna

Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

opisowa

hemisphere – na jakiej półkuli jest dany kraj

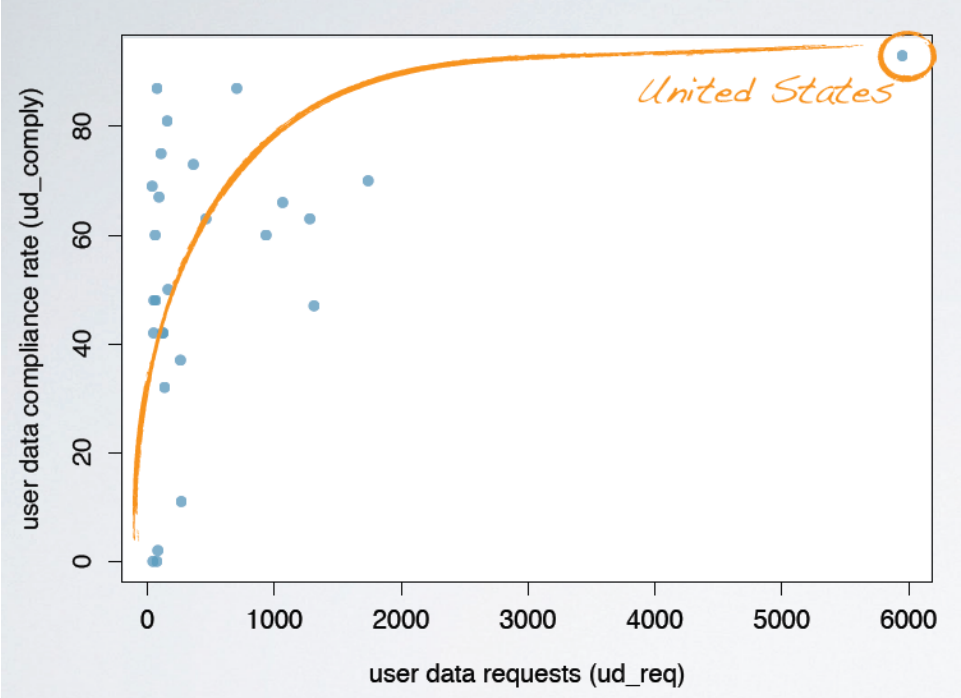
Macierz danych

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

opisowa uporządkowana

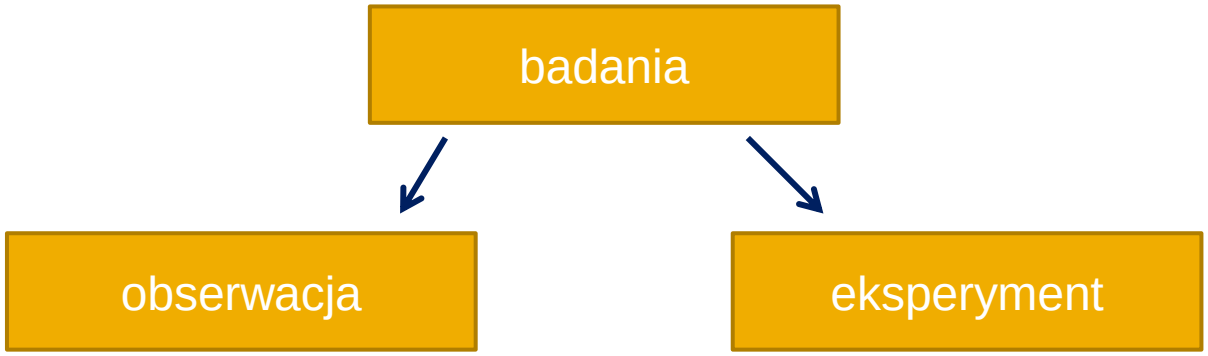
hdi – poziom życia w danym kraju

Relacja pomiędzy danymi



- Te dwie zmienne są ze sobą powiązane (skorelowane)
- Ta korelacja może być dalej sklasyfikowana jako pozytywna lub negatywna
- Jeżeli nie obserwujemy korelacji to mówimy że zmienne są niezależne.

W jaki sposób zbieramy dane?



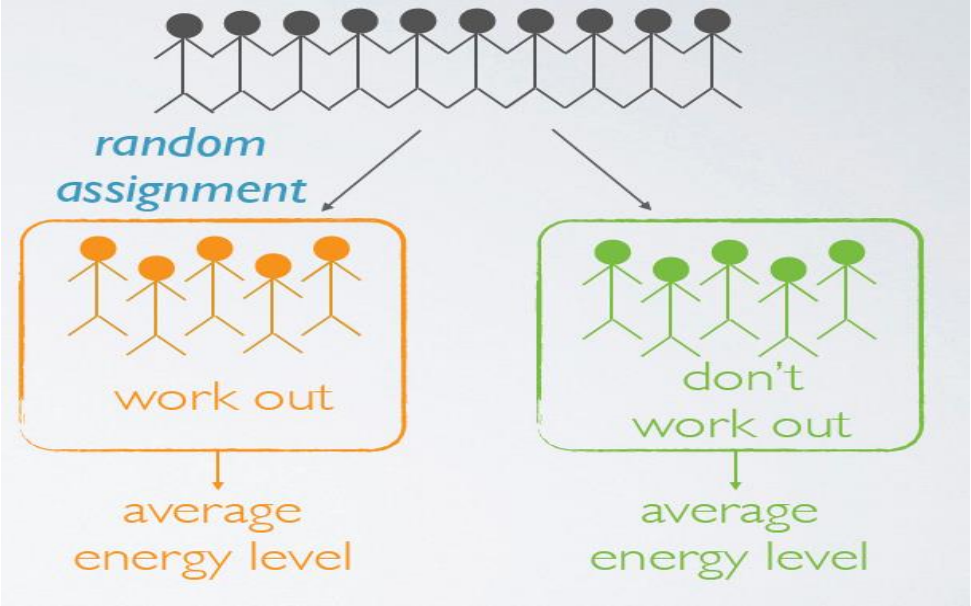
- Nie zaburza procesu w którym dane napływają
- Pozwala tylko na badanie związków pomiędzy danymi
  - Retrospektywne: dot. przeszłości
  - Prognozowane: dot. przyszłości

- Losowe przyporządkowuje podmiot do kategorii
- Pozwala na badanie związków przyczyna-skutek

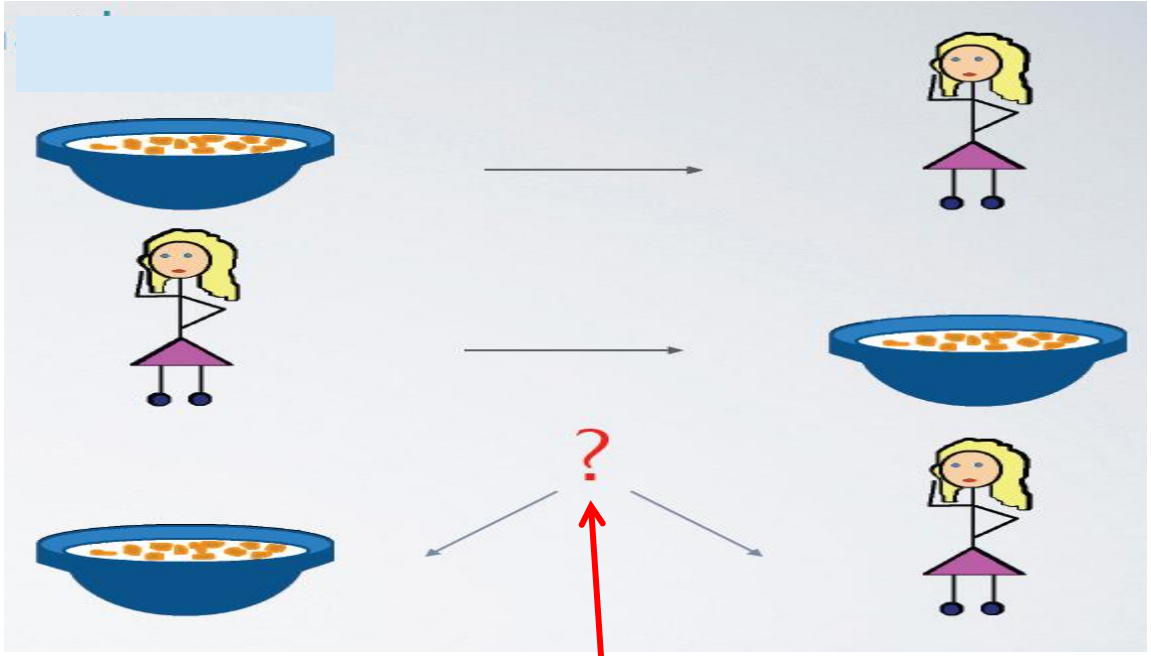
Pytanie: czy zaliczenie kursu jest związane z poziomem aktywności

obserwacja

eksperyment



Czy teza „Jedzenie corn-flakes na śniadania sprzyja prawidłowej wadze” jest prawdziwa?

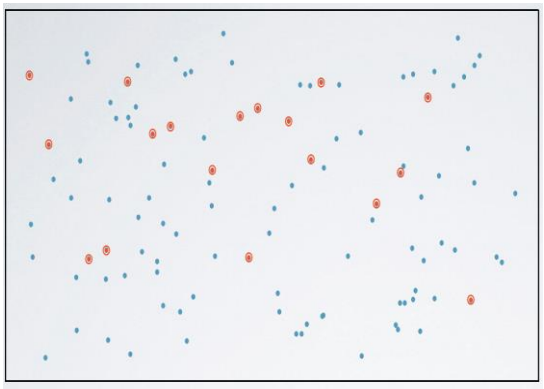


Korelacja nie oznacza wynikania!

W jaki sposób zbierać dane?

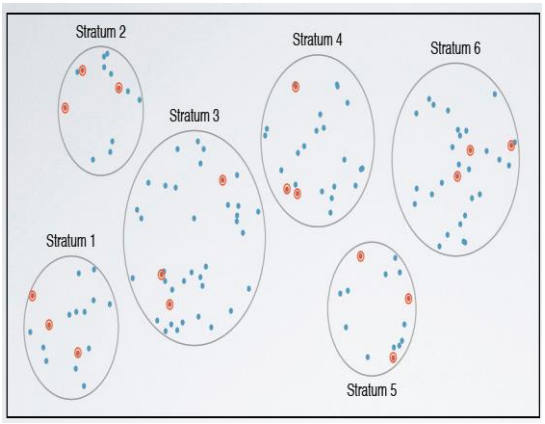
- Spis czyli przebadać cały zbiór ?
  - Niektórzy osobnicy mogą być trudni do zlokalizowania lub przebadania, ale też prawdopodobnie będą inni niż cała reszta osobników
  - Zbiór osobników na ogół nie jest stabilny
- Wybrać tylko reprezentatywną próbkę? tak ale uwaga na możliwość baiasowania:
  - Np. tylko łatwo dostępni osobnicy
  - Np. tylko volontariusze?

W jaki sposób wybierać próbkę?



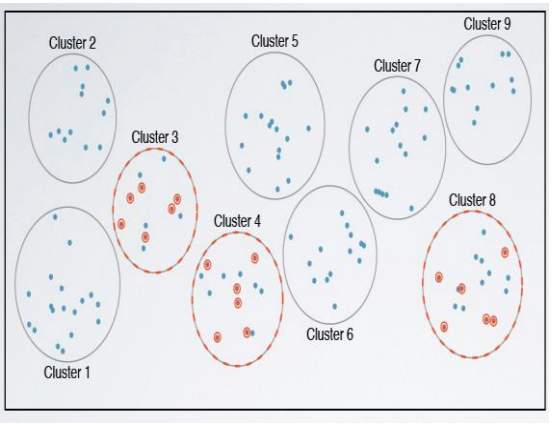
Wybieramy losowo osobników

(random sampling)



Dzielimy na różne kategorie (divide &conquer) , wybieramy losowo kilka kategorii i następnie losowo kilku osobników z każdej z wybranych kategorii.

(stratifying)



Dzielimy na identyczne podzbiory (warstwy) i wybieramy losowo osobników z kilku losowo wybranych podzbiorów

(clustering)

W jaki sposób planować eksperyment?

- Próbka kontrolna (control sample): porównuj wynik dla testowanej próbki i dla próbki kontrolnej
- Losowość (randomize): wybierz z próbki losowo osobników którzy są podlegają testowaniu
- Grupowanie (blocking): pogrupuj wpierw osobników względem zmiennej o której wiemy że może wpływać na wynik badania
- Powtarzanie (replicate): powtarzaj testowanie wielokrotnie na różnych próbkach

Przykład: grupowanie

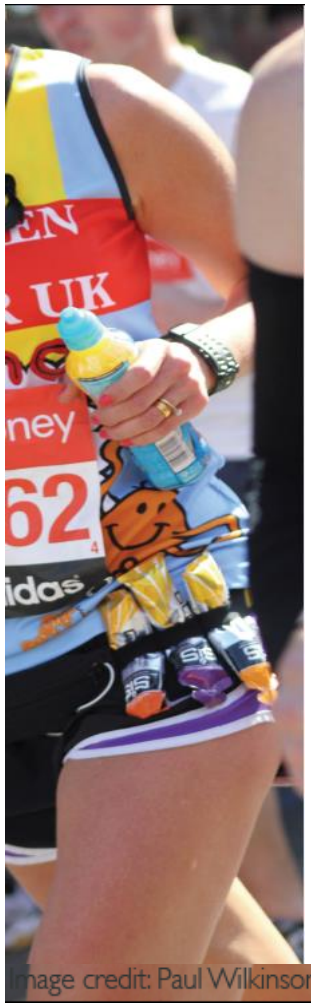


Image credit: Paul Wilkinson CC-BY 2.0 http://www.flickr.com/photos/28477990@N03/8683998728

- Zaplanuj eksperyment badający czy napój energetyzujący pomaga biegać?
  - Próbka testowana: podajemy napój
  - Próbka kontrolna: nie podajemy napoju
- Ale profesjonaliści mogą różnie reagować na napój niż amatorzy
- Grupuj najpierw ze względu na status:
  - Podziel osobników na „pro”, „amator”
  - Podziel każdą grupę na „testowaną” i „kontrolną”
  - „Pro” i „amator” są jednakowo reprezentowani w próbce testowanej i kontrolnej.

Blocking vs stratifying

- Zmienne objaśniające (explanatory): to coś co możemy zażądać od testowanej próbki
- Zmienne charakterystyczne (blocking variables) które służą do grupowania to właściwości naszych osobników nad którymi chcemy mieć kontrolę ale których nie możemy zmienić
- Jeżeli nasza populacja ma pewne własności charakterystyczne, to najpierw grupujemy, a potem losowo wybieramy osobników.

Terminologia

- „Placebo”: „fałszywy test” na grupie kontrolnej często używana technika w medycznych testach
- „Placebo” efekt: pojawia się efekt w grupie kontrolnej której nie poddano działaniu preparatu
- „Blinding”: grupa nie wie do jakiej kategorii należy
- Podwójny „blinding”: eksperymentatorzy też nie wiedzą do jakiej kategorii należy grupa

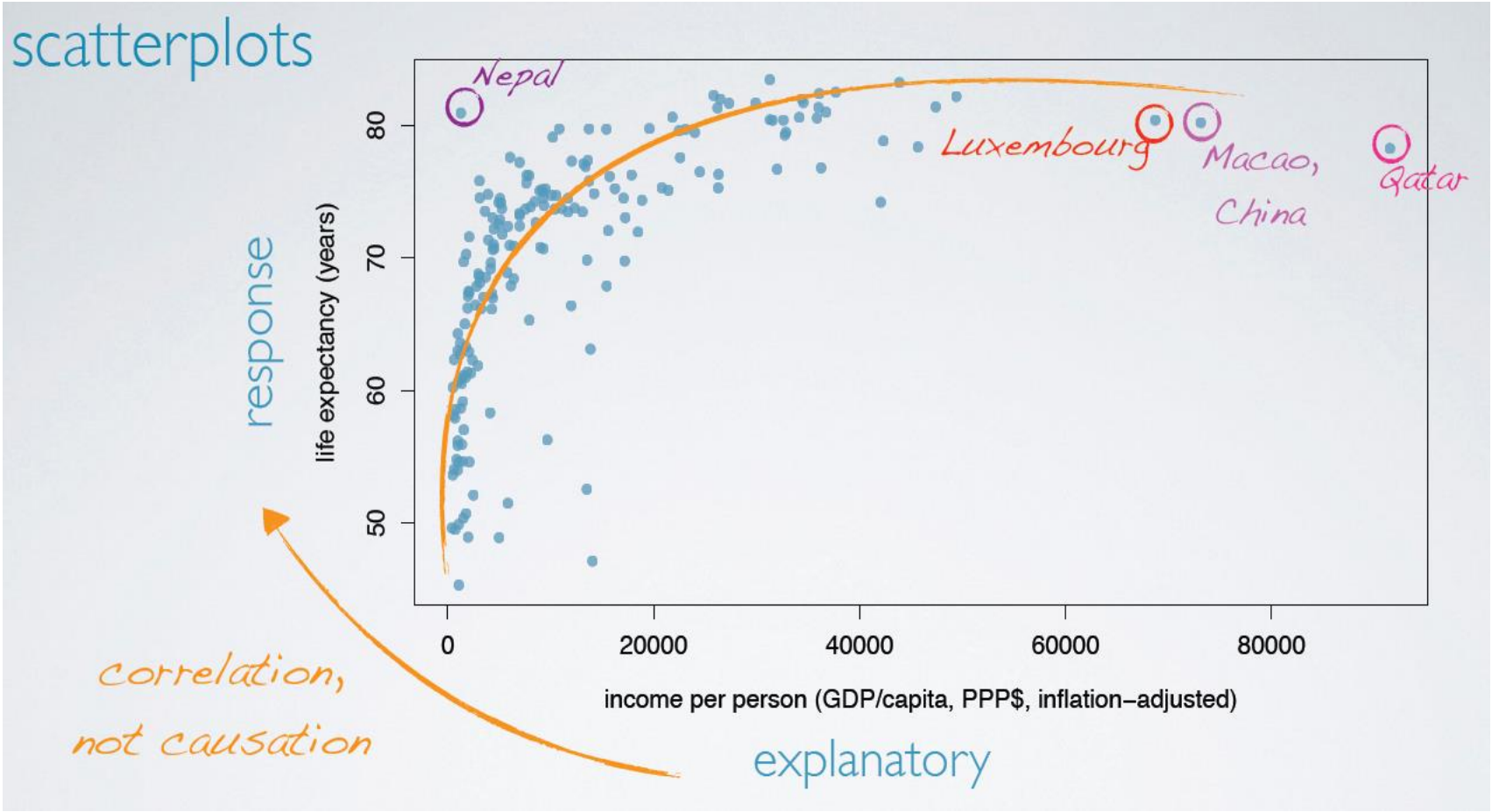
Wizualizacja danych

data	income per person ($, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
…	…	…
Zimbabwe	545.3	58.142

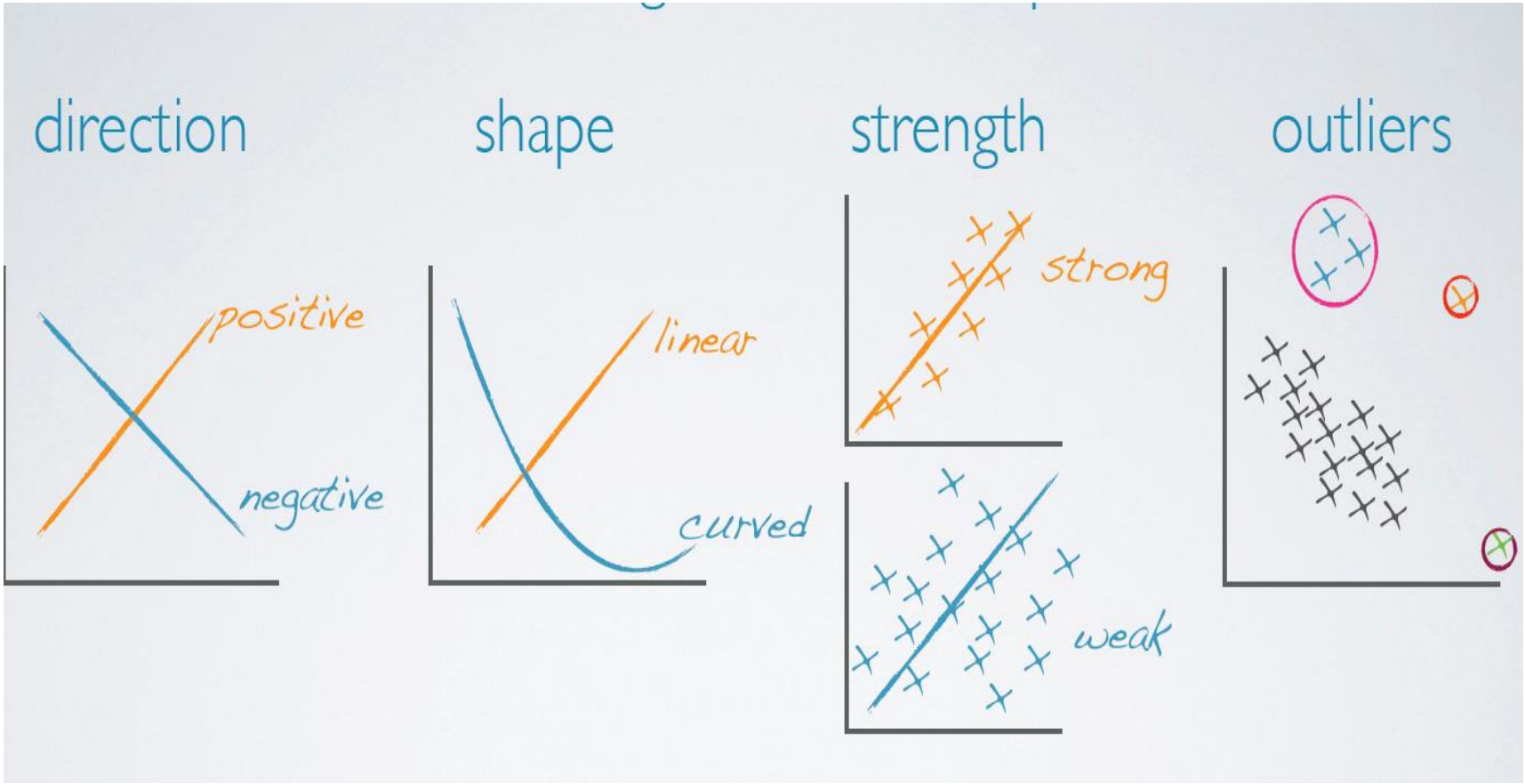
Source: gapminder.com

Wizualizacja danych

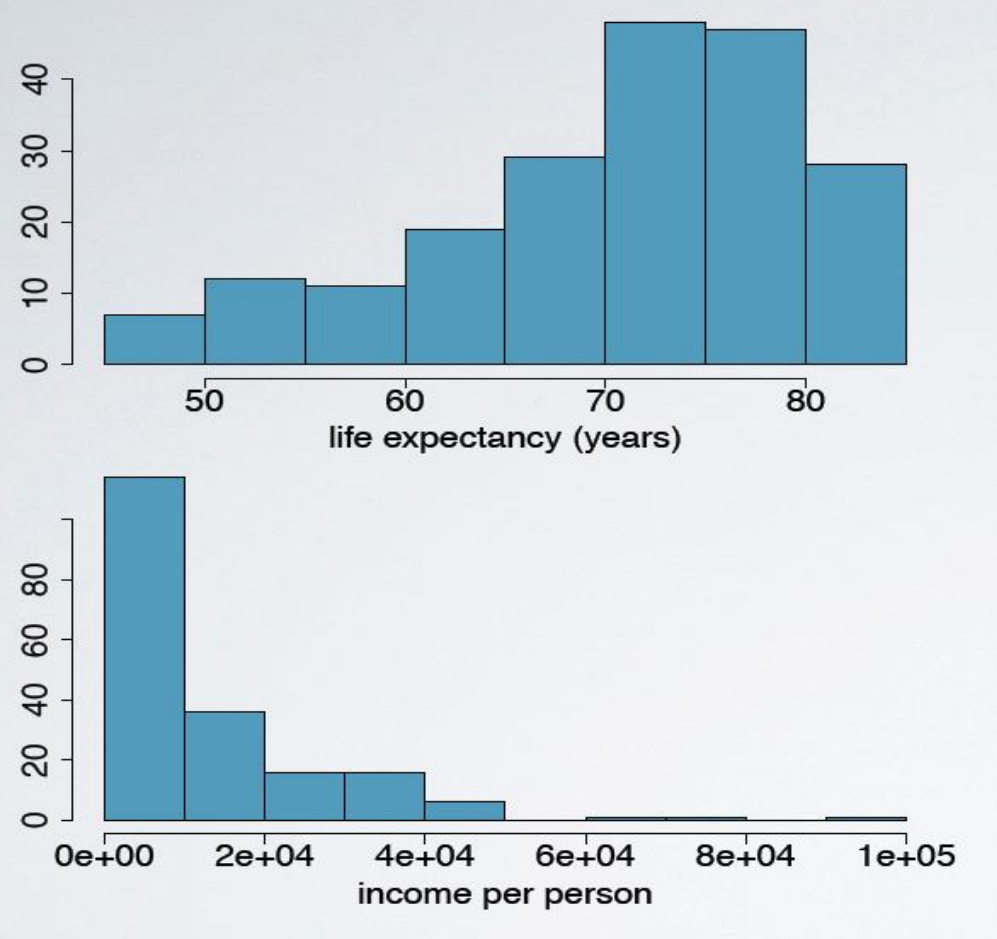
Wizualizacja danych



Związki pomiędzy zmiennymi



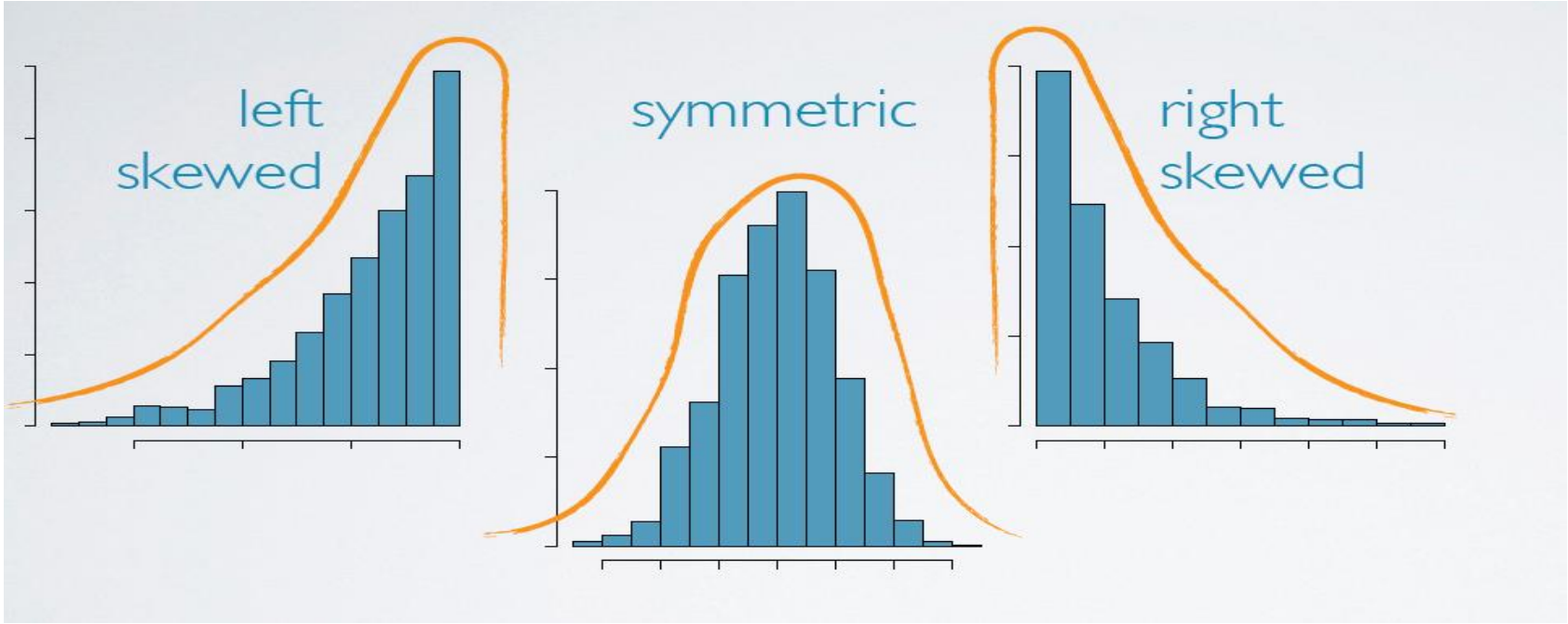
Histogram



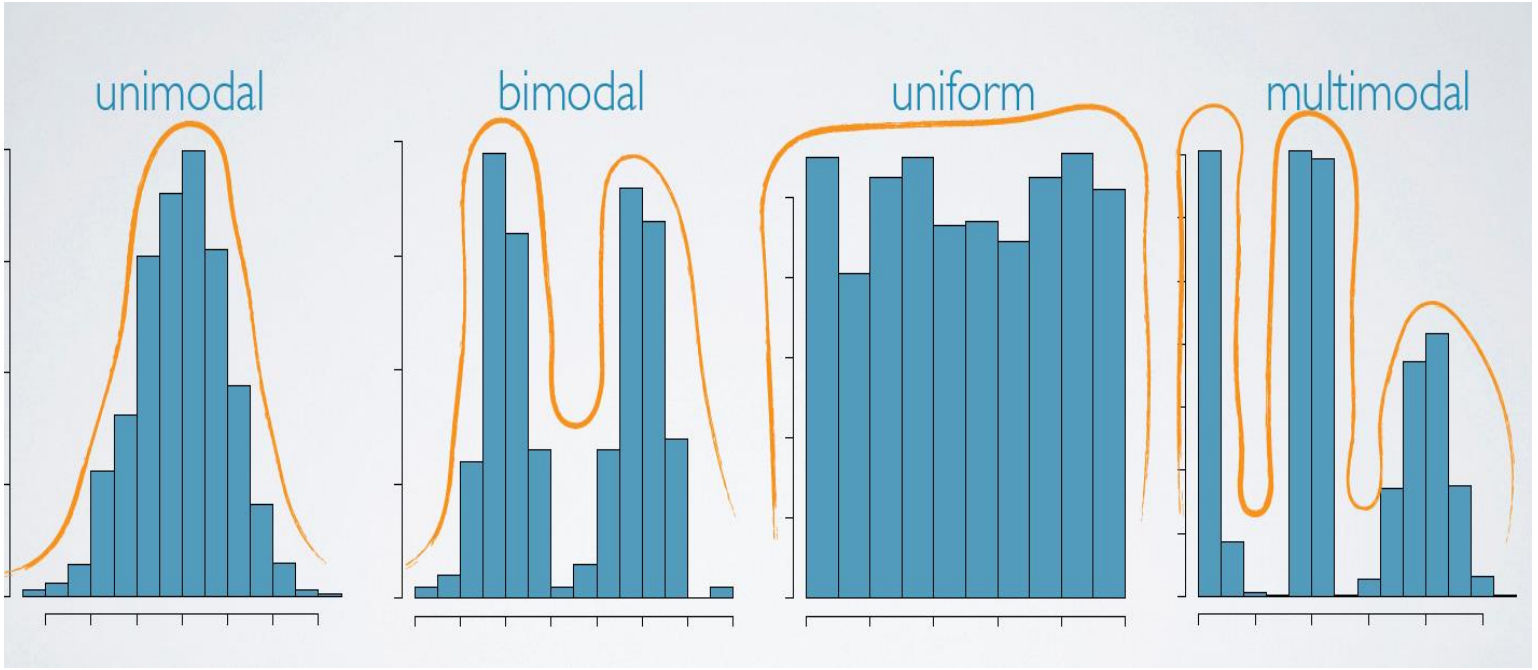
- Pozwala obejrzeć jaki jest rozkład funkcji gęstości
- Specjalnie dogodny aby zilustrować kształt rozkładu

Przekrzywienia (skewed)

Rozkłady są przekrzywione w stronę długich ogonów rozkładu



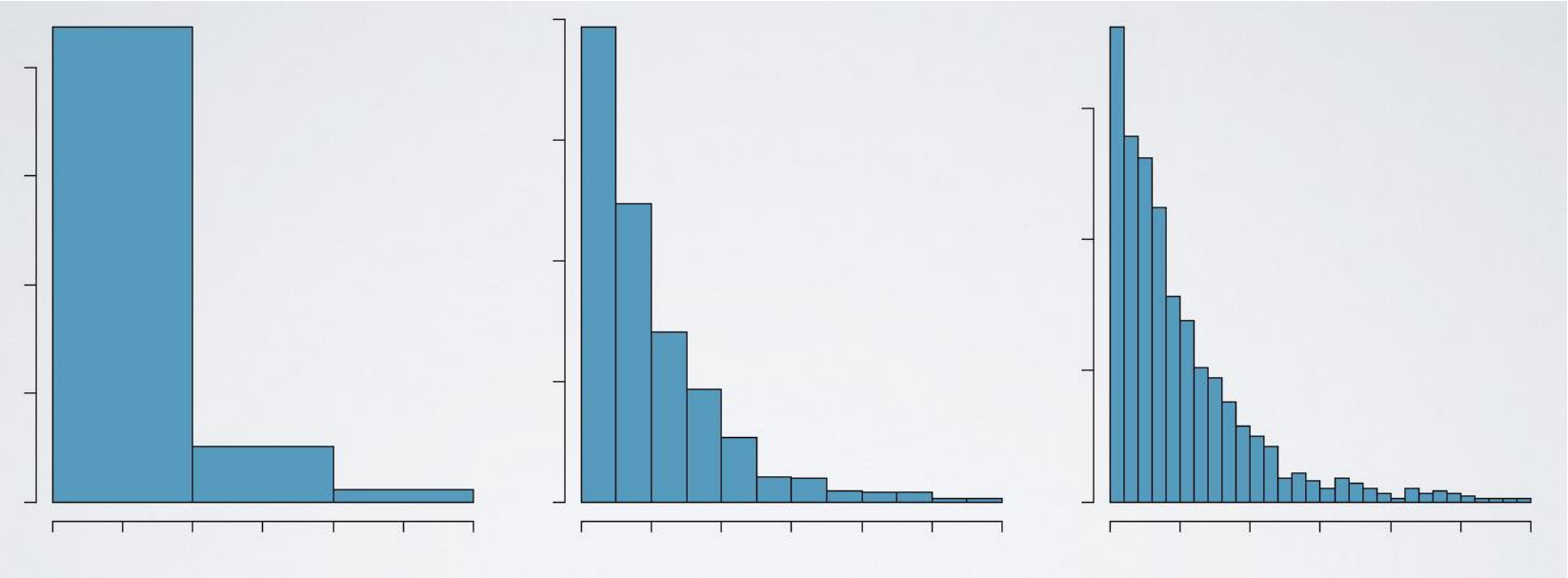
Modalność rozkładu



Modalność rozkładu

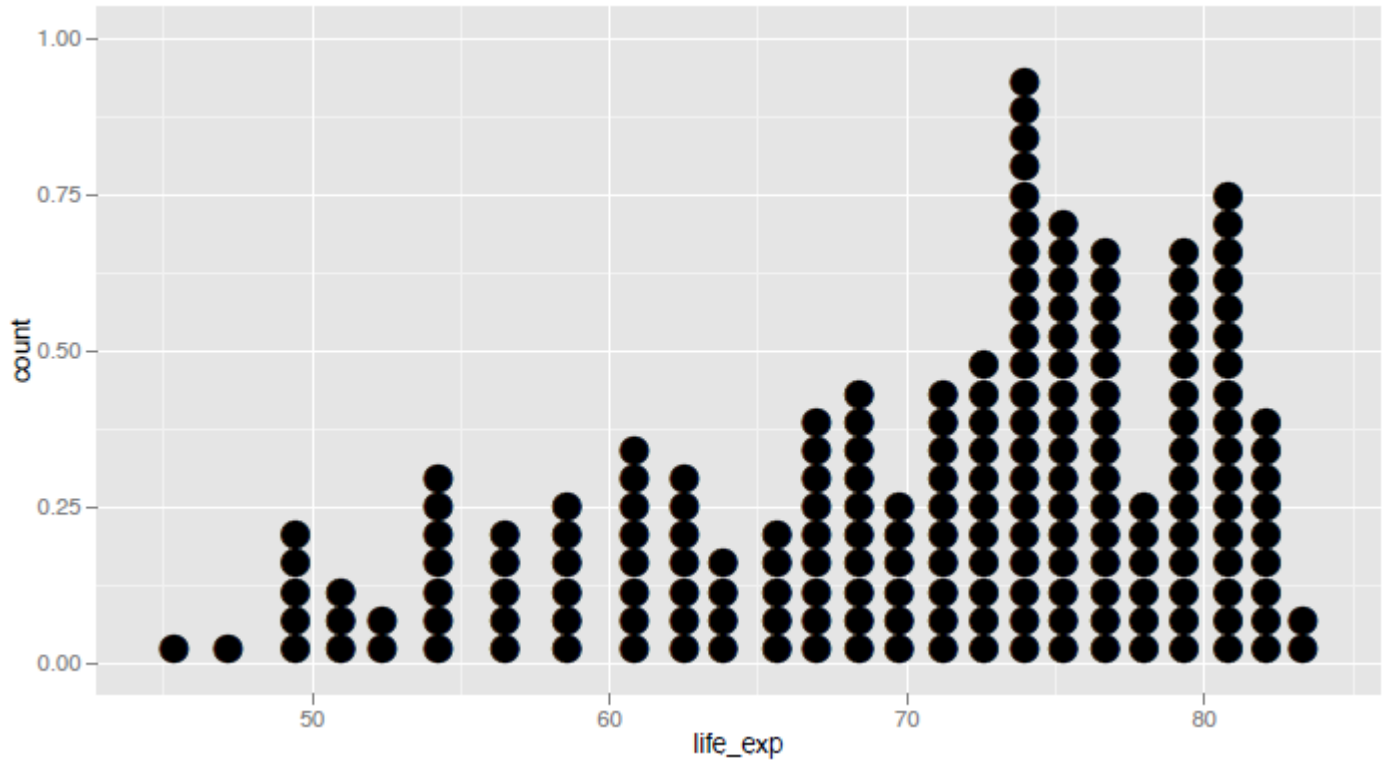
Histogram i szerokość binu

Wybór szerokości binu może ukryć informację



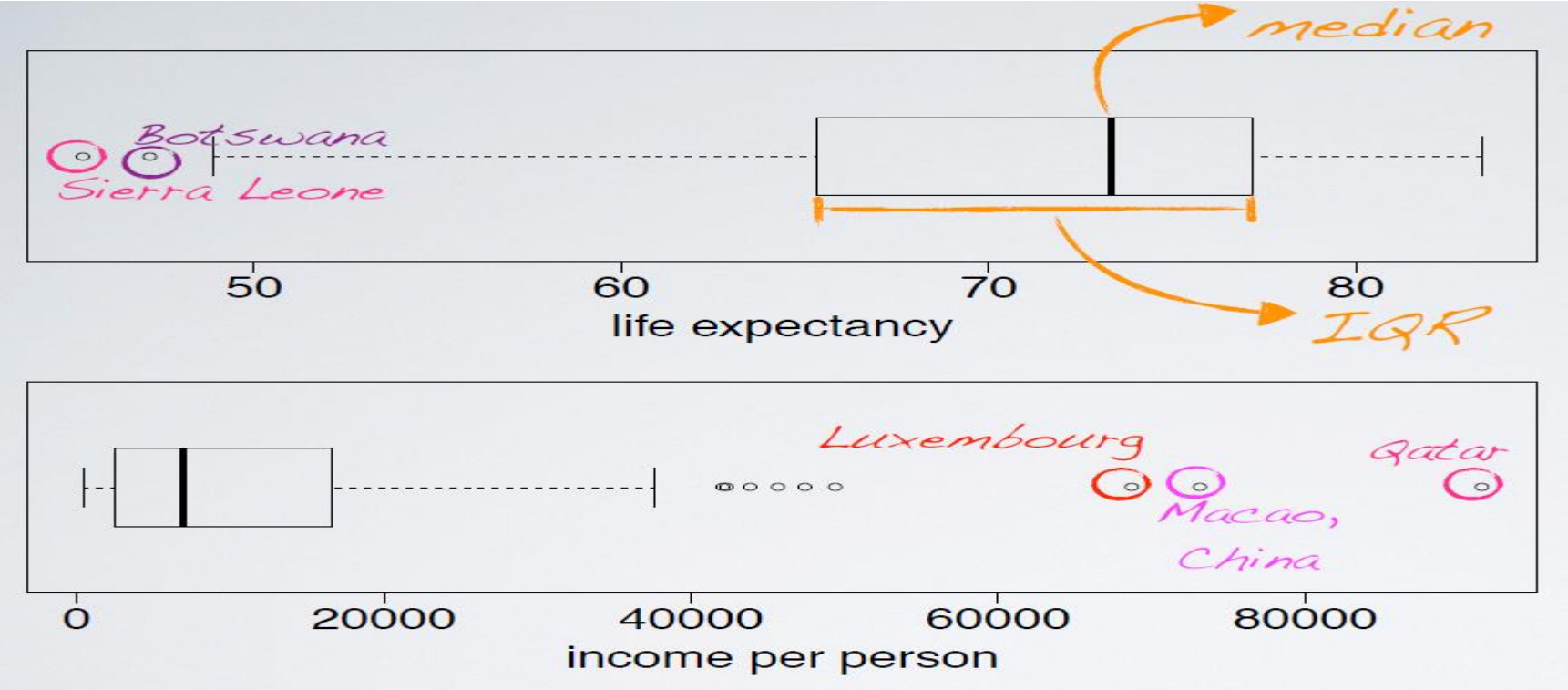
Punktowe ploty

Też wygodny, ale przy niedużej statystyce

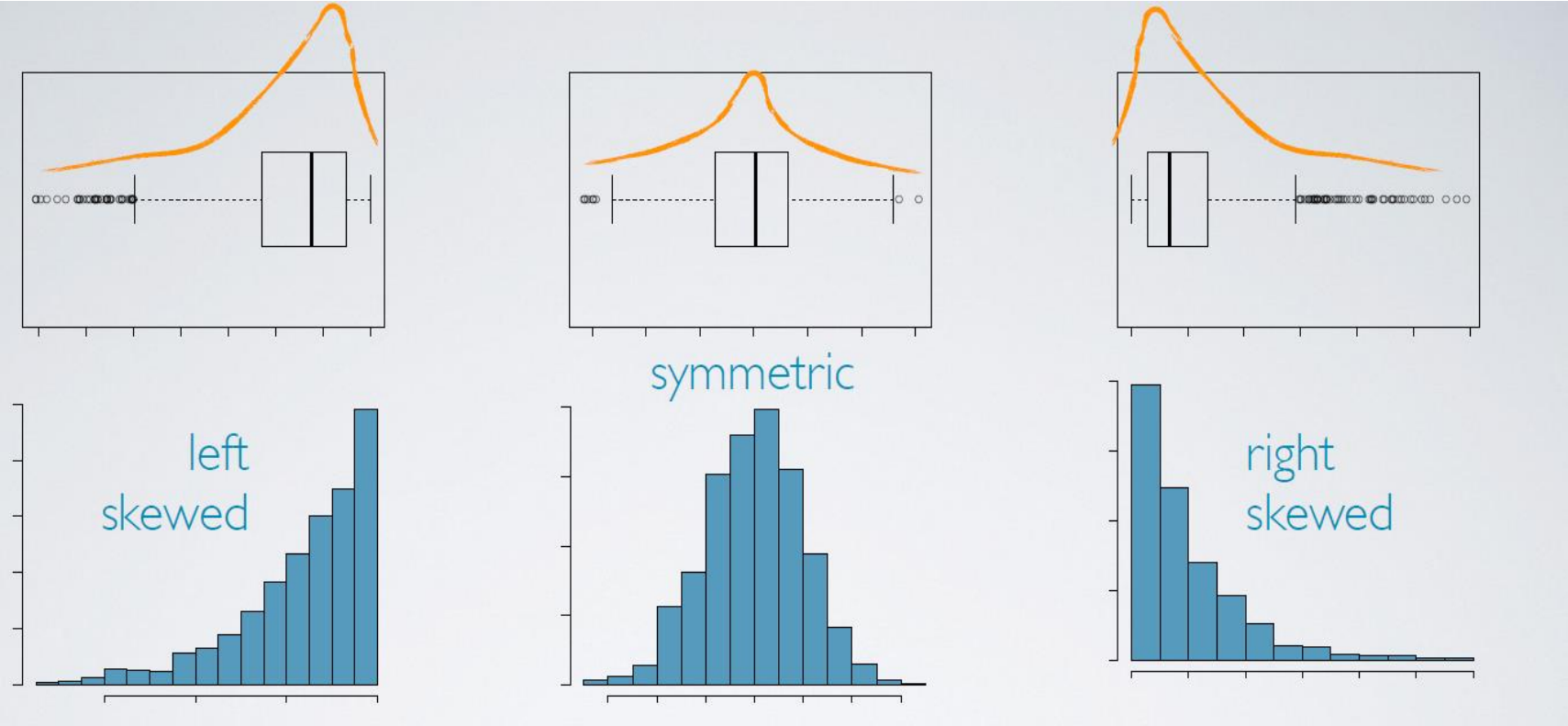


Box-plot

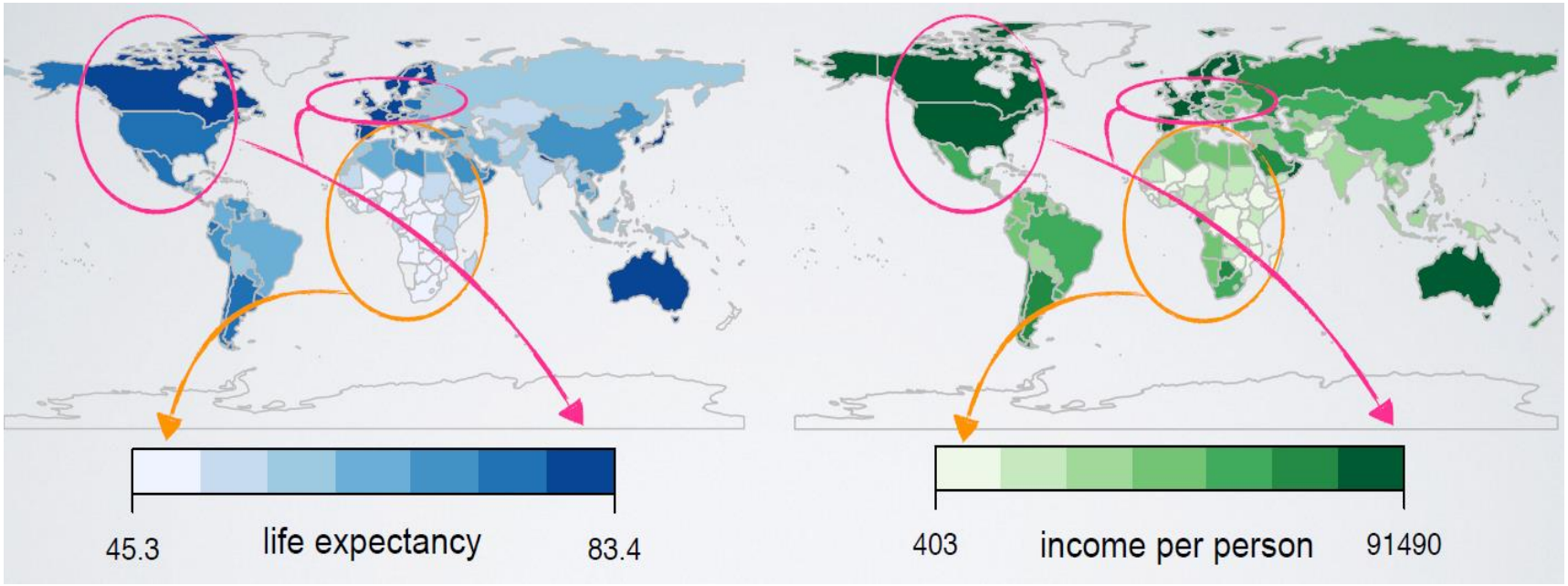
- Wygodny aby zaznaczyć medianę (< 50% rozkładu), zakres międzykwartylowy (IQR), outliers



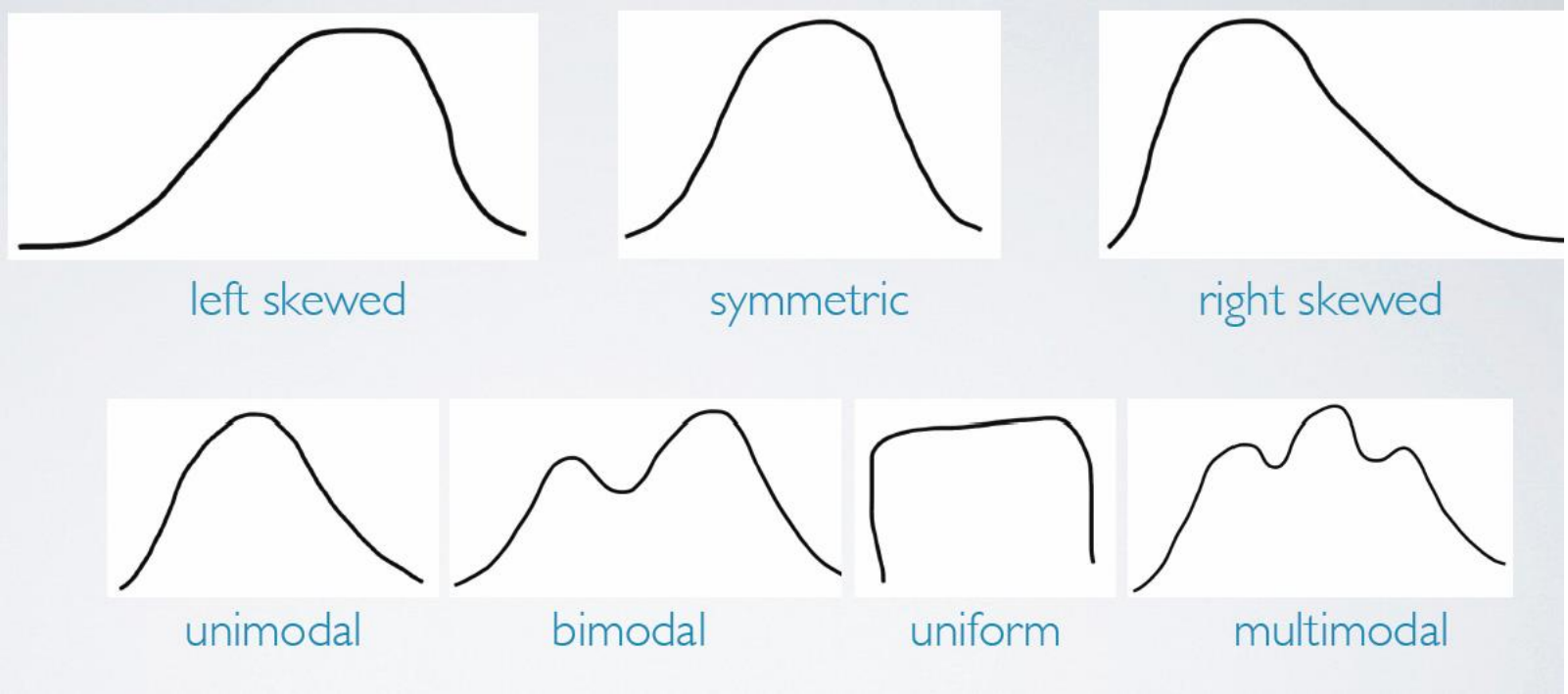
Wizualizacja danych



Mapa gęstości



Kształt rozkładu



Jeden parameter: środek rozkładu

- Mean: średnia arytmetyczna
  - $\bar{x}$ średnia z próbki
  - μ średnia całej populacji
- Mode: wartość o największym prawdopodobieństwie
- Mediana: środek rozkładu (poniżej 50%)

Przykład: wynik egzaminu 9-ciu studentów

75, 69, 88, 93, 95, 54, 87, 88, 27

mean: $\frac{75+69+88+93+95+54+87+88+27}{9} = 75.11$

mode: 88

median: 27, 54, 69, 75, 87, 88, 88, 93, 95

A gdyby było 10-ciu

27, 54, 69, 75, 87, 88, 88, 93, 95, 100

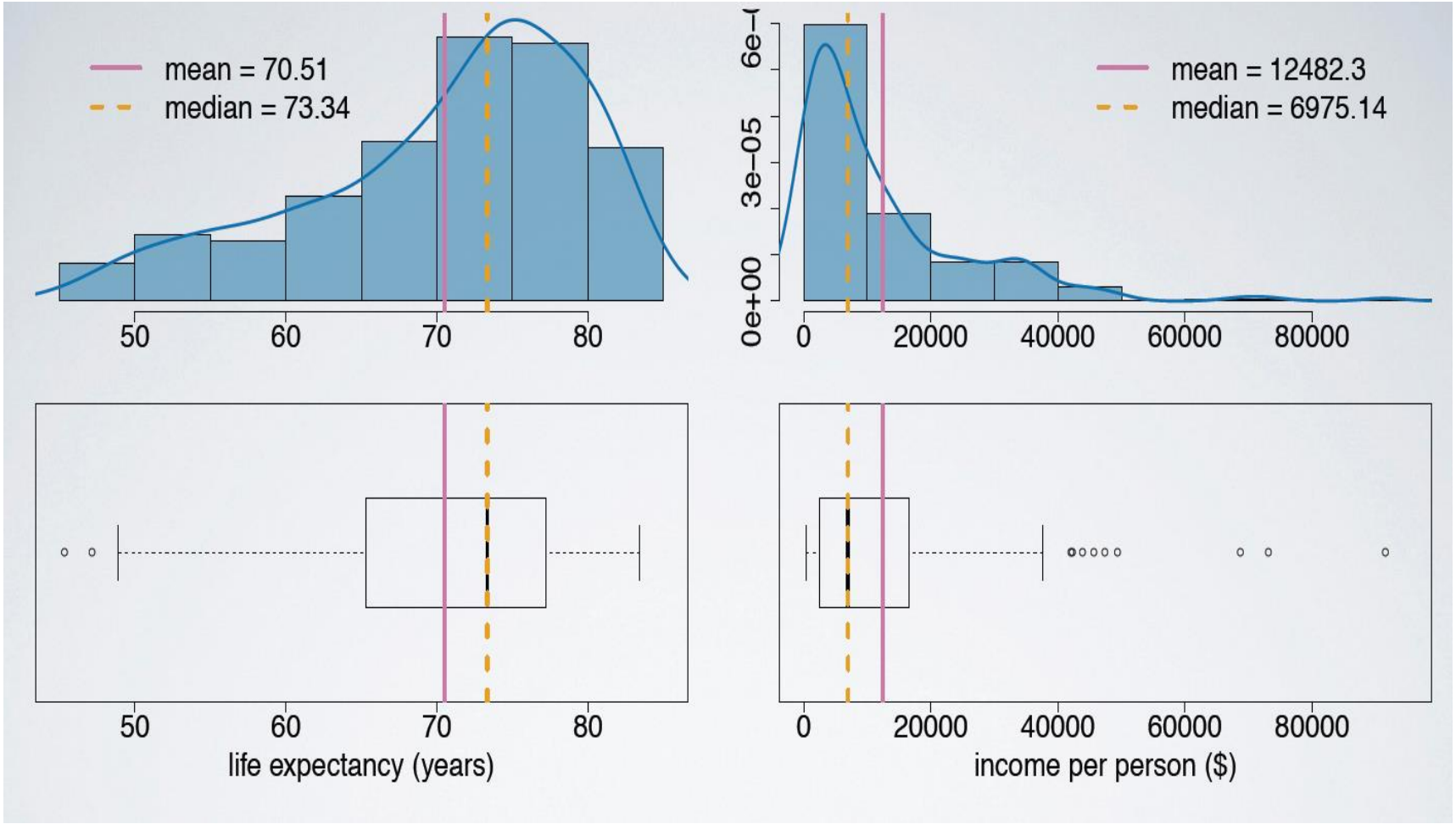
$\frac{87+88}{2} = 87.5$

Wracamy do tabelki

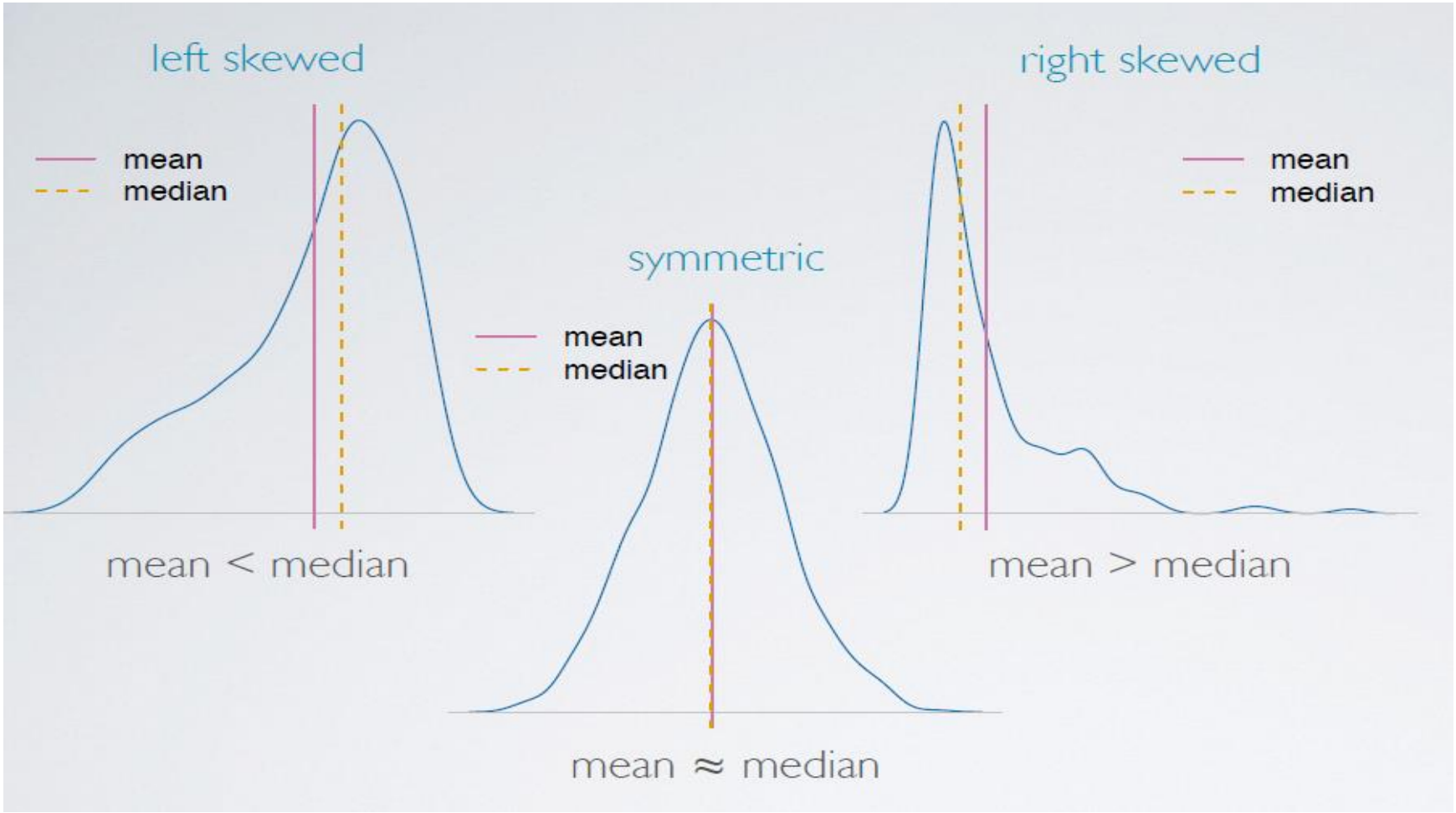
data	income per person ($, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
…	…	…
Zimbabwe	545.3	58.142

Source: gapminder.com

Wizualizacja danych



Skrzywienie vs środek



Wariancja

sample variance s^2

population variance σ^2

w przybliżeniu średni kwadrat odchylenia od średniej

$$s^2 = \frac{\sum_1^{n=1}(x_i - \bar{x})^2}{n-1}$$

Przykład: zakładając że średni czas życia jest 70.5 lat i w tabelce mamy 201 krajów

$$s^2 = \frac{(60.3 - 70.5)^2 + (77.2 - 70.5)^2 + \ldots + (58.1 - 70.5)^2}{201 - 1}$$

$$= 83.06 \text{ years}^2$$

	country	life exp
1	Afghanistan	60.3
2	Albania	77.2
3	Algeria	70.9
	...	...
201	Zimbabwe	58.1

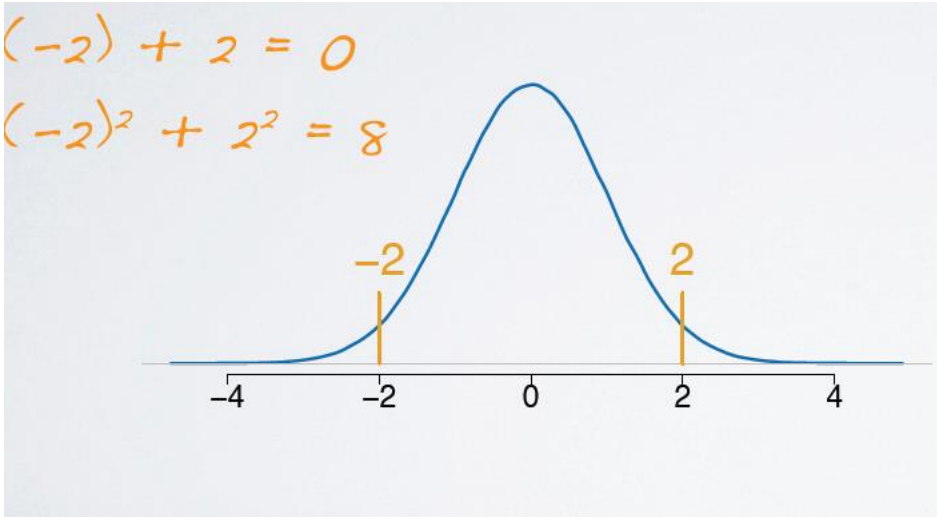
Wariancja

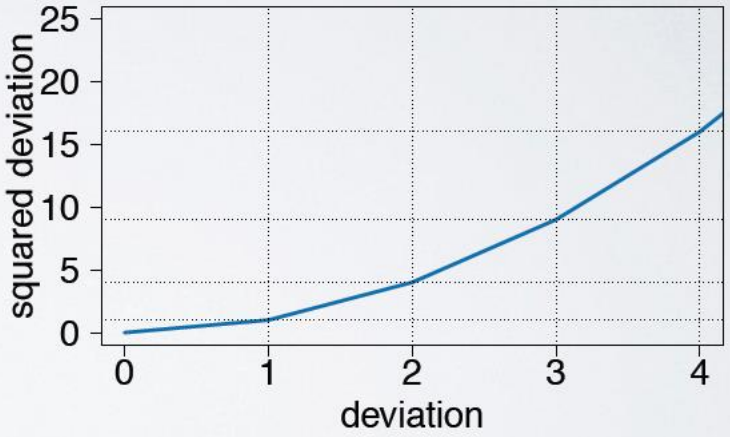
$$s^2 = \frac{\sum_1^{n=1} (x_i - \bar{x})^2}{n-1}$$

- Dlaczego różnicę podnosimy do kwadratu?

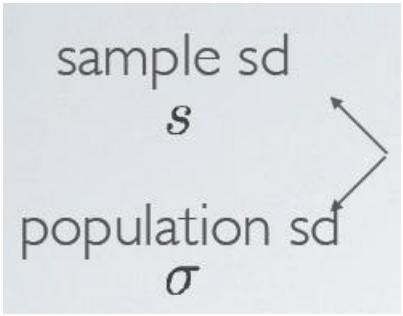
Aby dodatnie i ujemne różnice nawzajem się nie znosiły

Aby zwiększyć wkład od dużych różnic bardziej niż od małych





Odchylenie standardowe



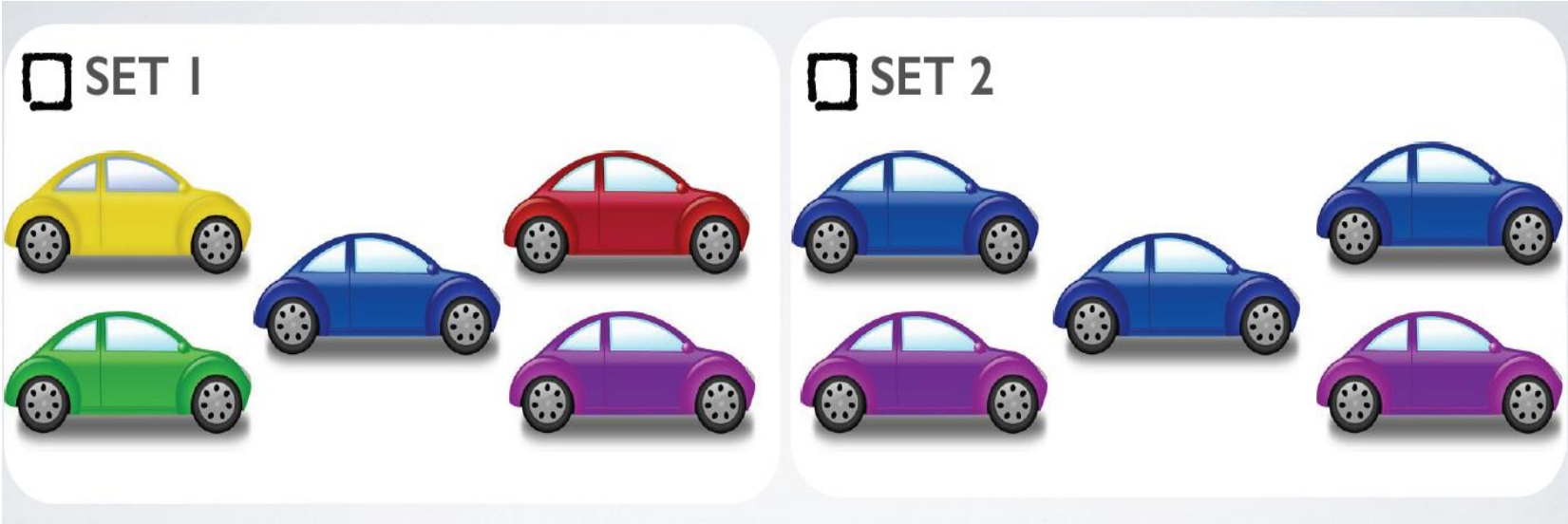
W przybliżeniu średnie odchylenie od wartości średniej i jest w tych samych jednostkach co dane

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_1^{n=1}(x_i - \bar{x})^2}{n-1}}$$

square root of the variance

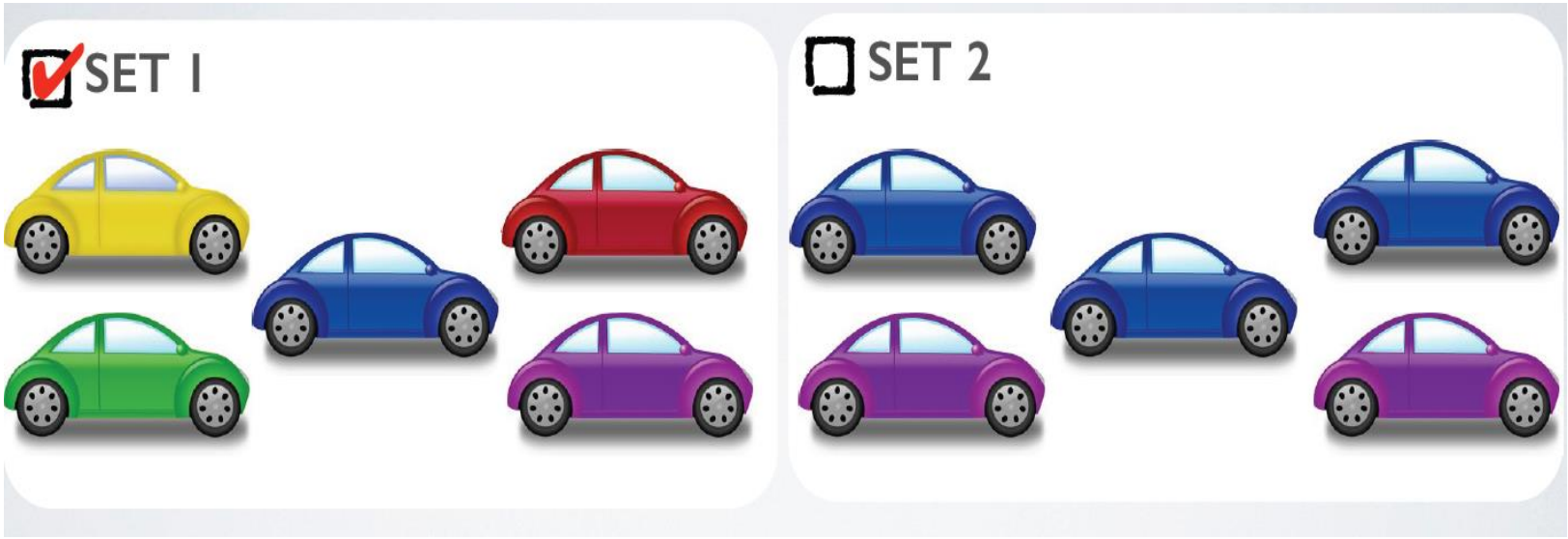
Zmienność vs różnorodność

Który z zestawów ma większą różnorodność kolorów



Zmienność vs różnorodność

Który z zestawów ma większą różnorodność kolorów

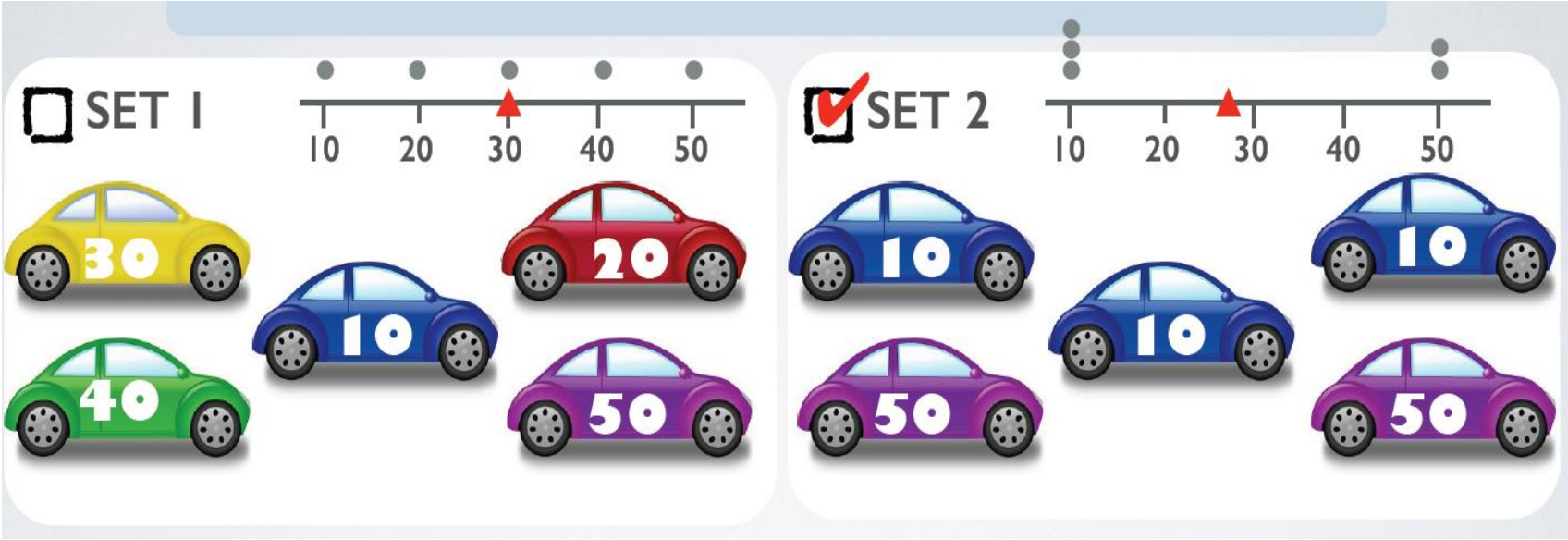


Zmienność vs różnorodność

Który z zestawów ma większą zmienność zużycia benzyny?

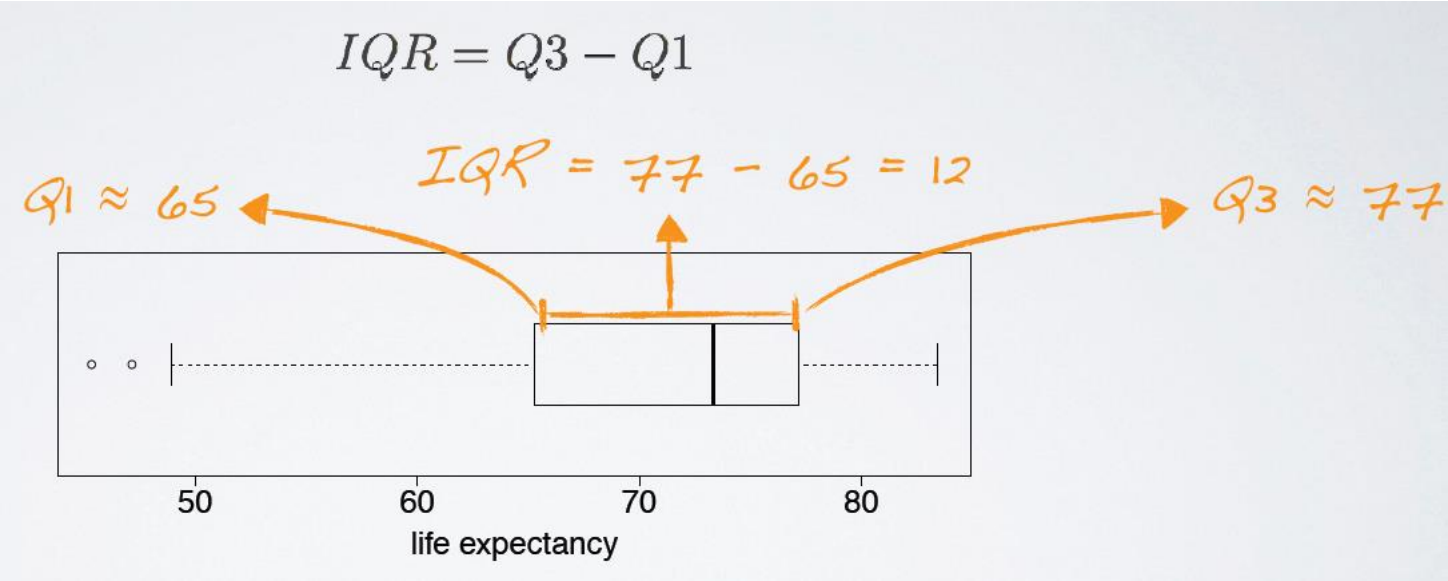
Zmienność vs różnorodność

Który z zestawów ma większą zmienność zużycia benzyny? (zmienność = rozrzut)



Zakres międzykwartylowy (interquartile range)

- Zakres w którym mieści się 50% danych, dystans od pierwszej (25% percentyla) do trzeciej (75% percentyla)



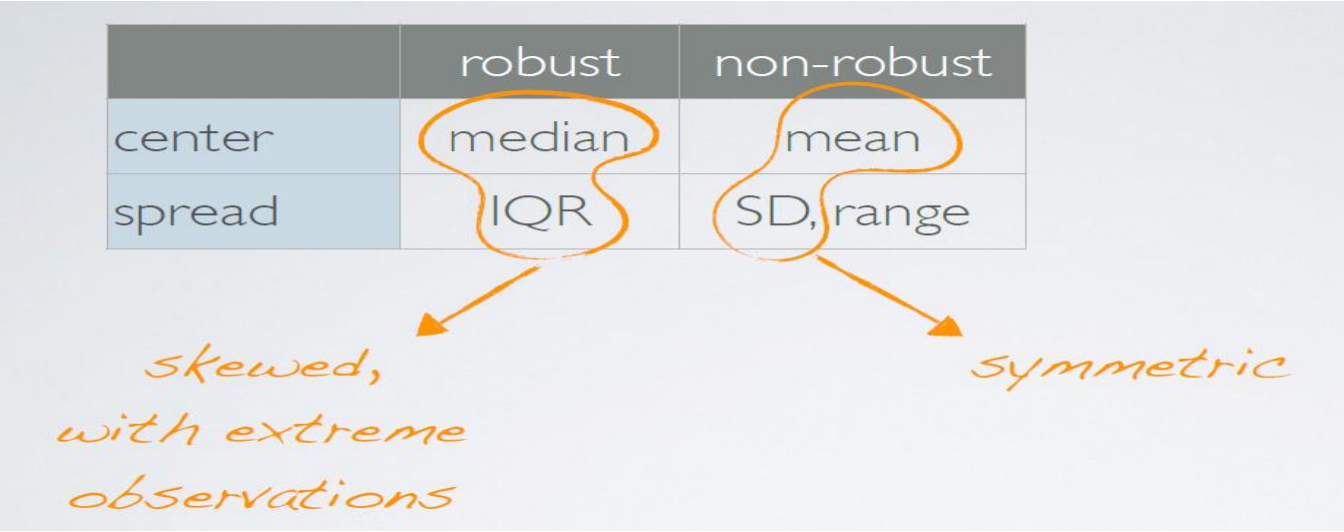
„Odporna” statystyka

- Czyli taka aby ekstremalne wartości miały mały wpływ na parametry populacji
- Przykład:

data	mean	median
1, 2, 3, 4, 5, 6	3.5	3.5
1, 2, 3, 4, 5, 1000	169	3.5

„Odporna” statystyka

- Czyli taka aby ekstremalne wartości miały mały wpływ na wartość badaną

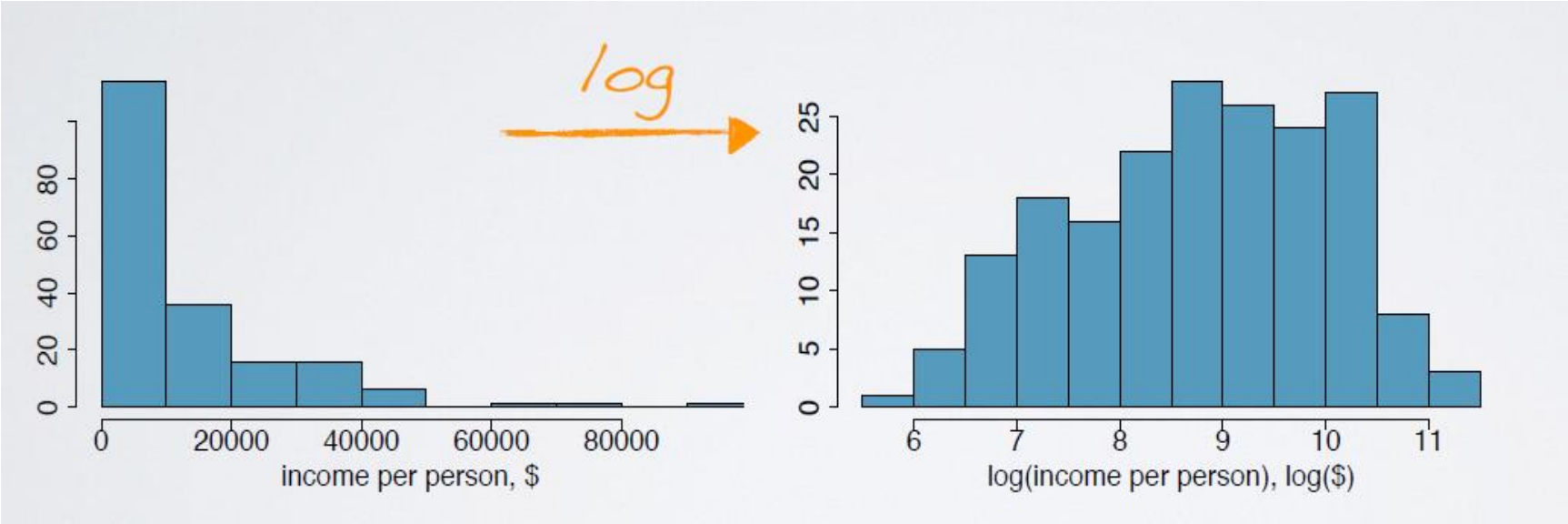


Transformacja danych

- To jest operacja na danych przy pomocy jakiejś funkcji, np. logarytm
- Jeżeli rozkład w danych jest bardzo przekrzywiony (ma długi ogon) używamy transformacji aby go było łatwiej modelować
- Inne przykłady:
  - Zobaczyć dane w innej reprezentacji
  - Zredukować skrzywienia rozkładu
  - Wyprostować zależność na scatter plocie

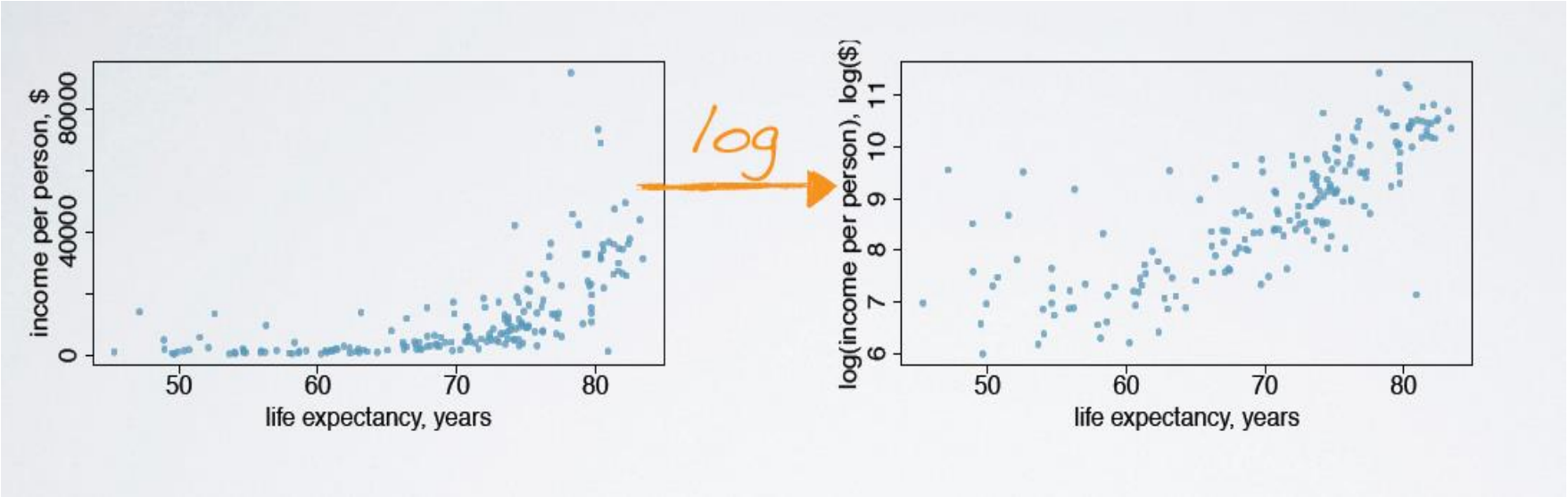
Transformacja przy pomocy $\log_e$

Naturalny log czyli przy podstawie e.
Wygodny jeżeli dane sklastrowane koło zera.

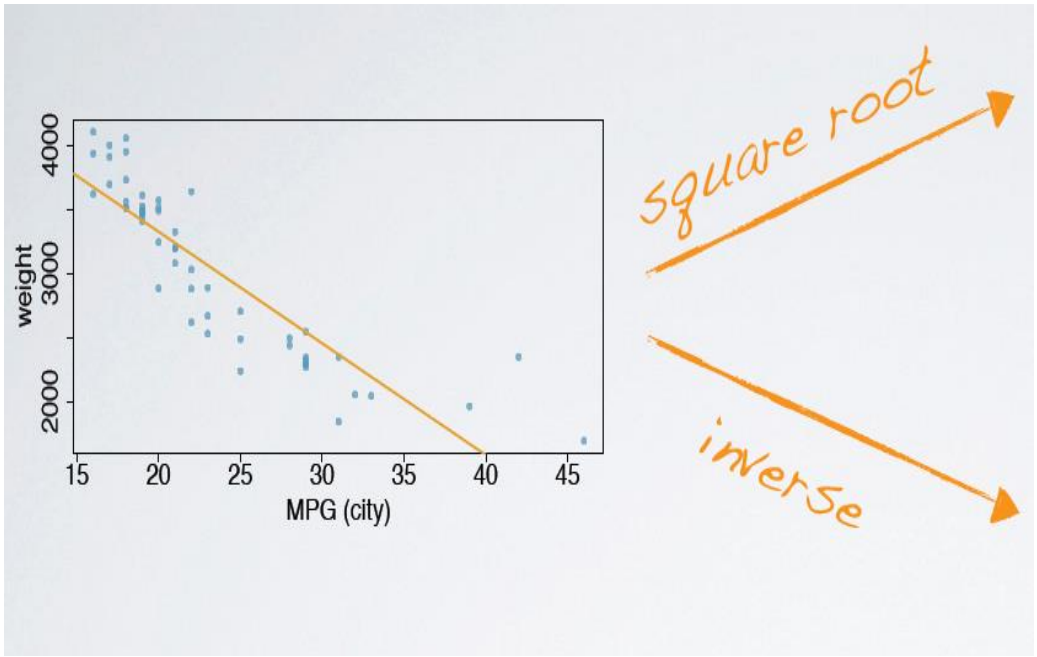


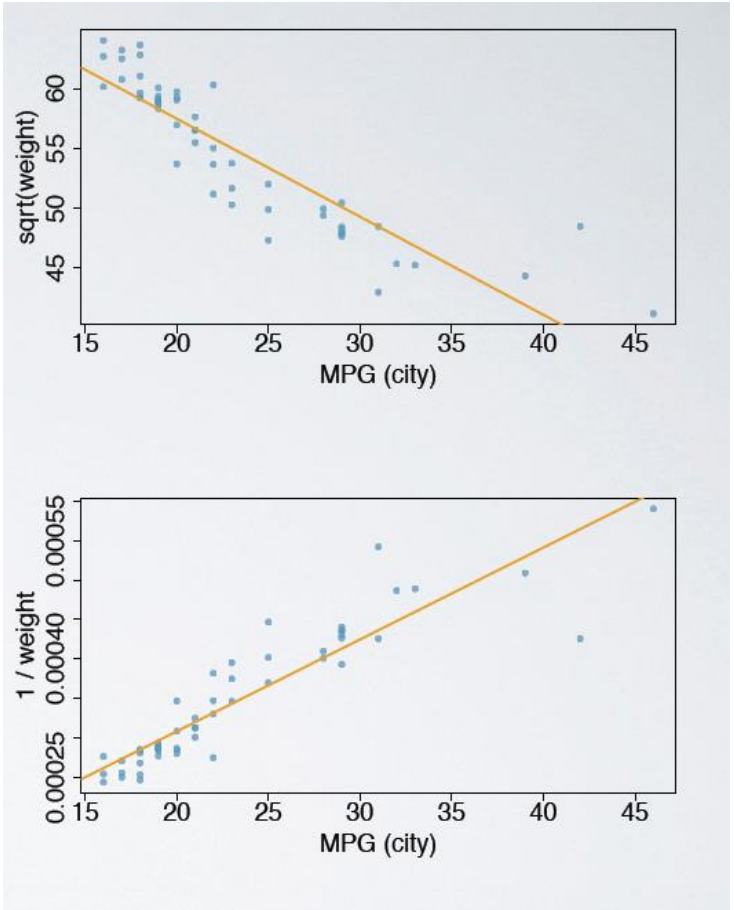
Przekształcenie przy pomocy logarytmu

Po przekształceniu zależność pomiędzy zmiennymi bardziej liniowa, łatwiejsza do modelowania.



Inne transformacje



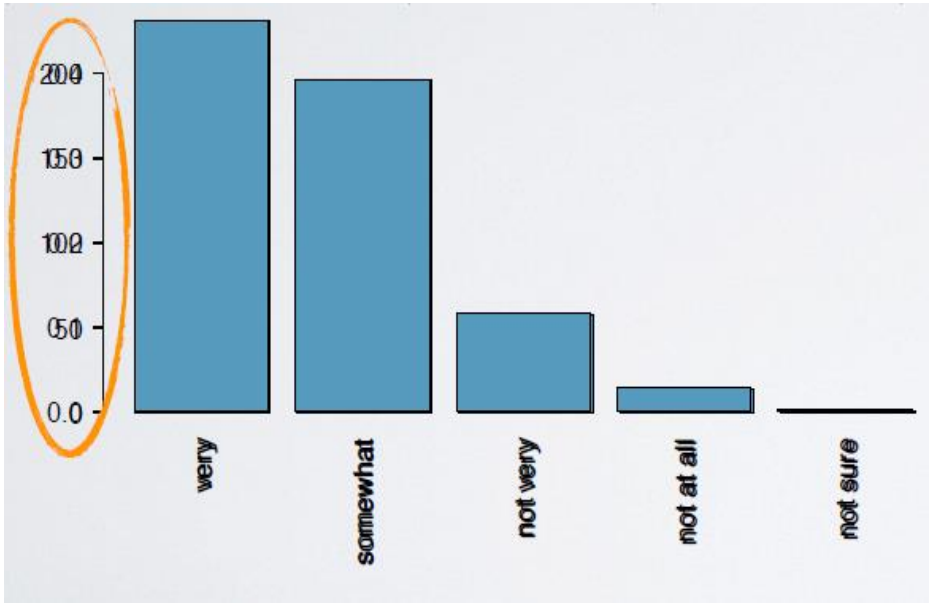


Zmienne opisowe

Tabela

Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%

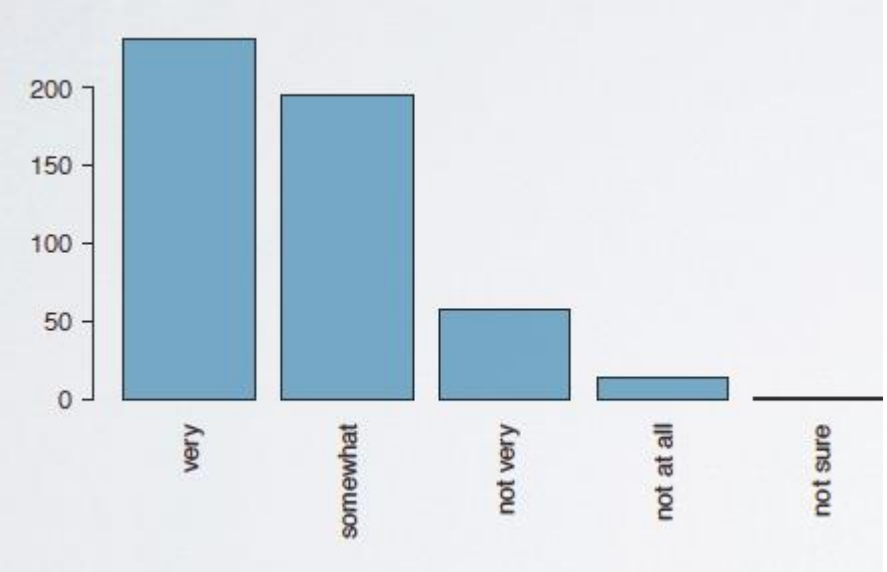
Wykres słupkowy (bar plot)

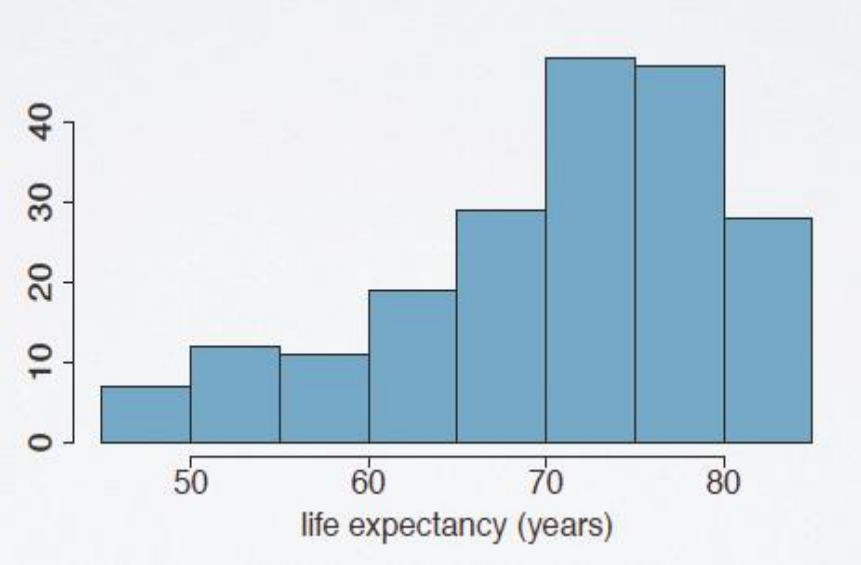


Wykres słupkowy a histogram

Wykres słupkowy używamy do wizualizacji zmiennych opisowych
Kolejność słupków może być zmieniana

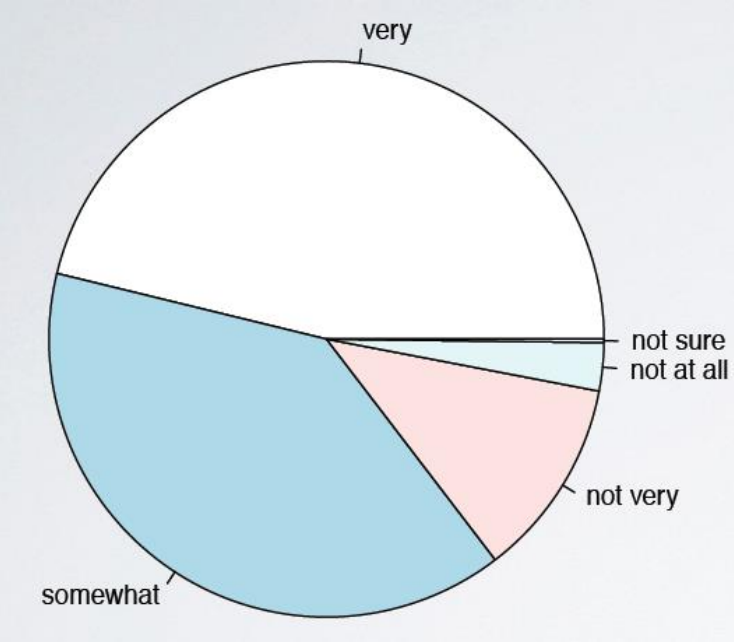
Histogram używamy do wizualizacji zmiennych numerycznych
oś-x jest zmienna numeryczna, ustalona kolejność binów

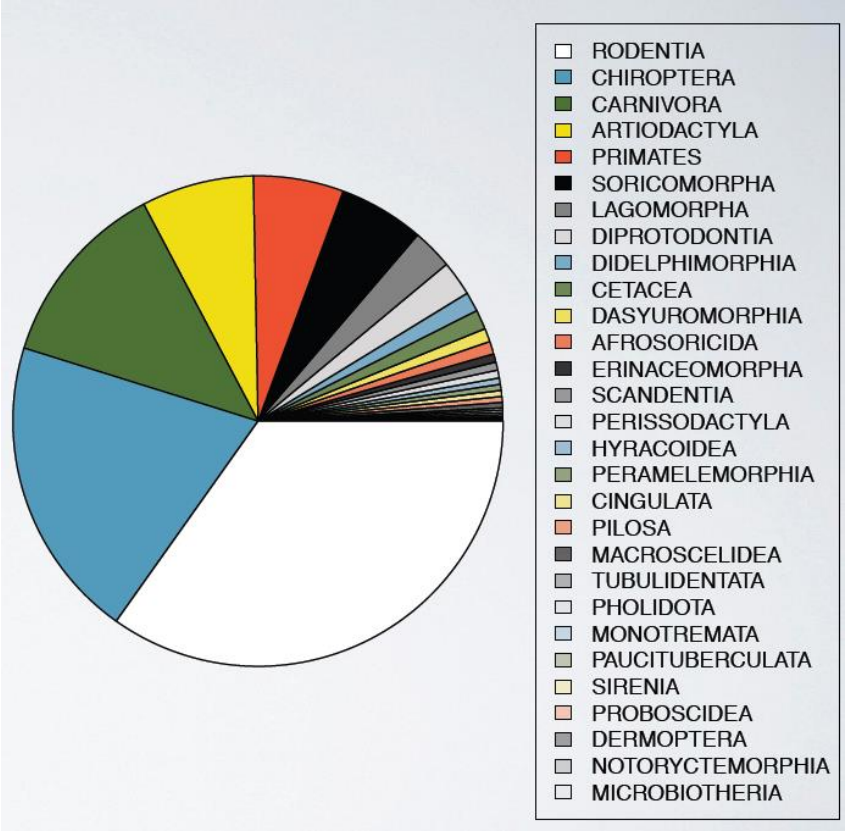




Wykres kołowy (pie chart)

- Raczej bezużyteczny





Przykład: czy łatwo jest oszczędzać

		Income				
		< $40K	$40-80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

Względna częstość

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

<$40K: 128 / 202 = 63% find it very difficult to save

$40K-$80K: 63 / 148 = 43%

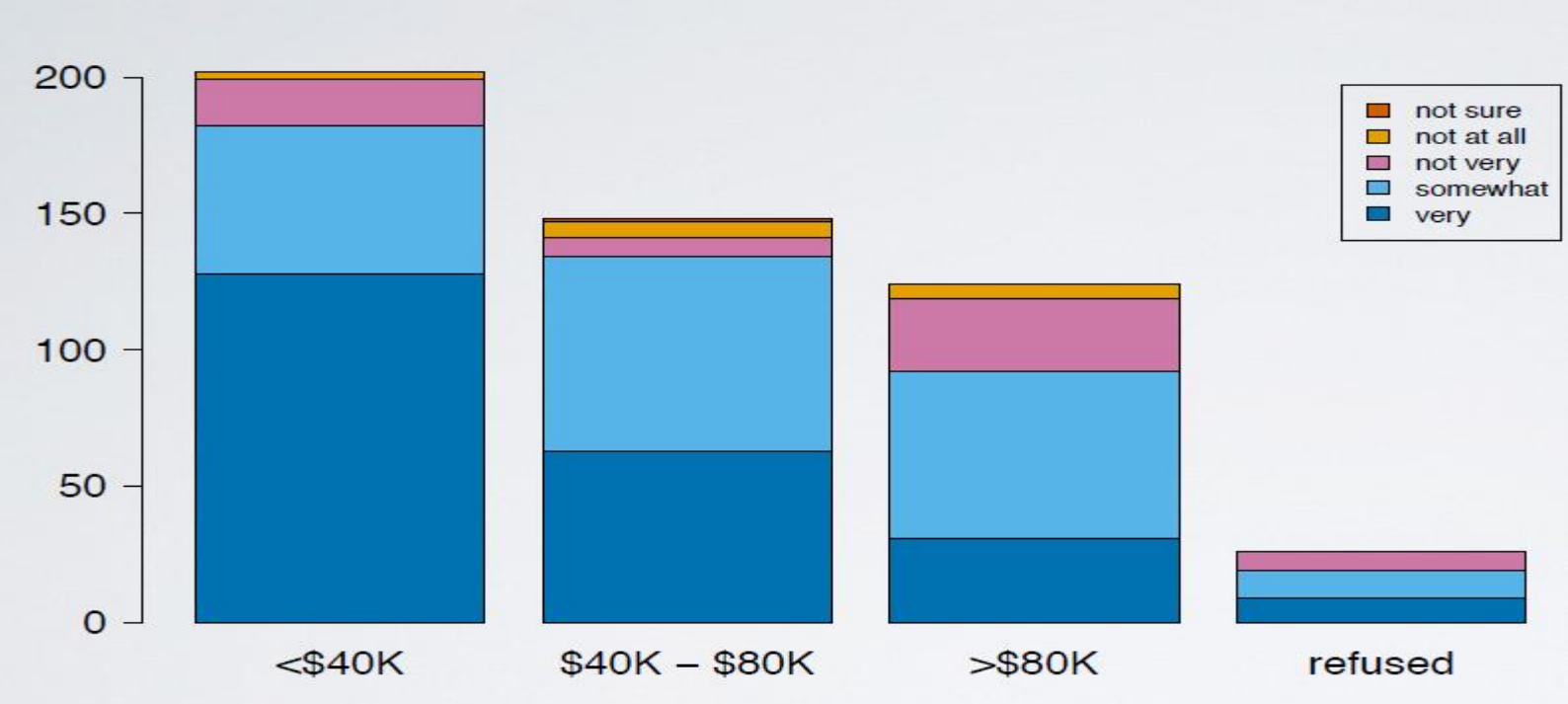
>$80K: 31 / 124 = 25%

Refused: 9 / 26 = 35%

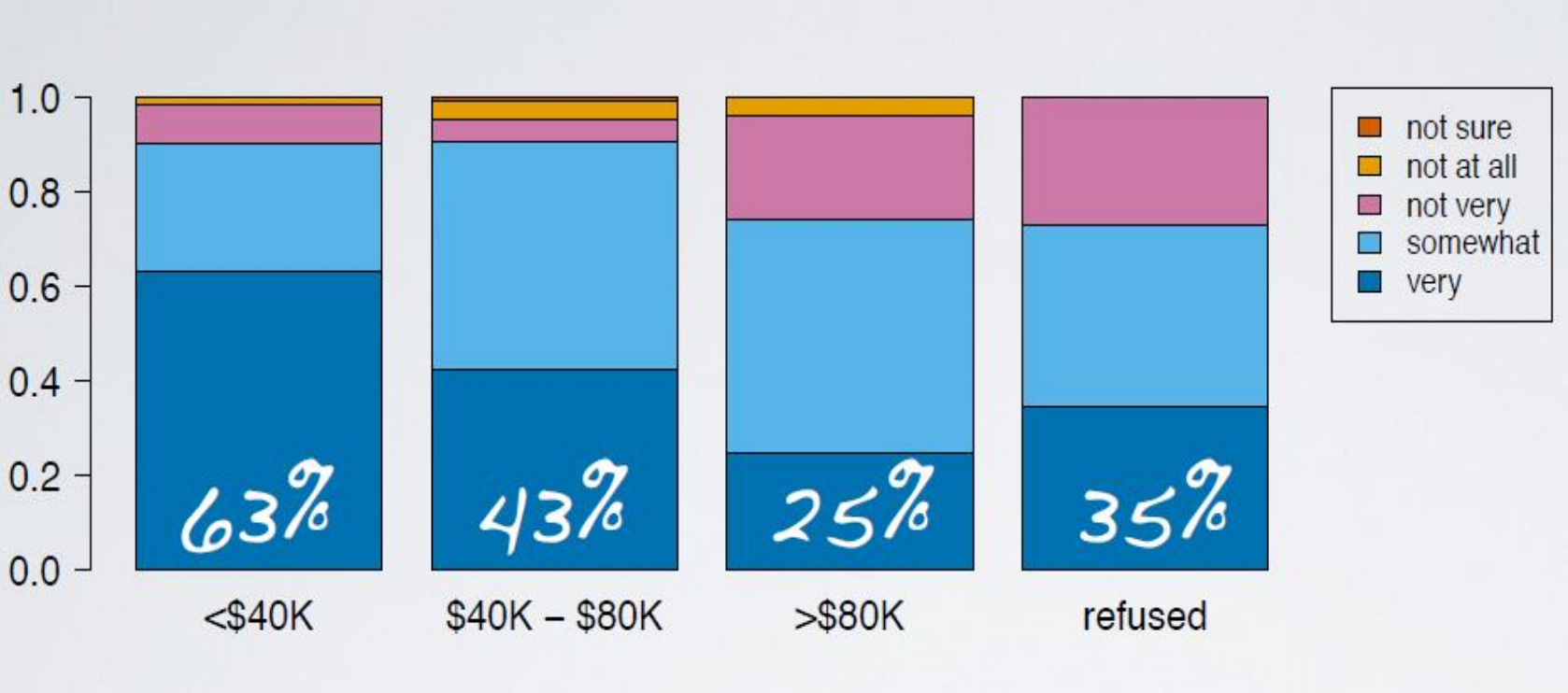
Względna ocena „jak trudno jest oszczędzać" i „dochód" są od siebie zależne

Segmentowany wykres słupkowy

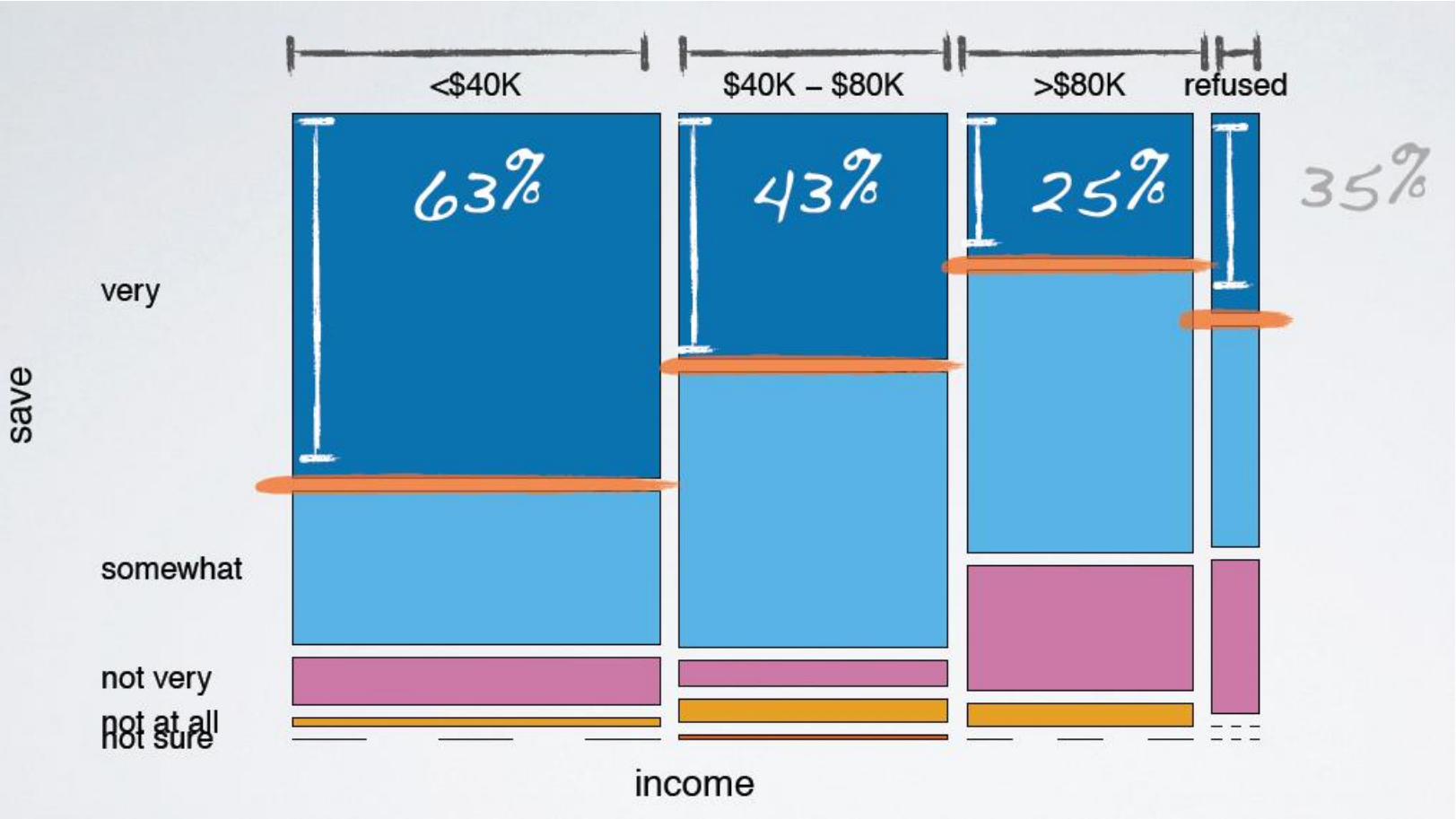
- Wygodny aby wizualizować względna częstości i je ze sobą porównywać



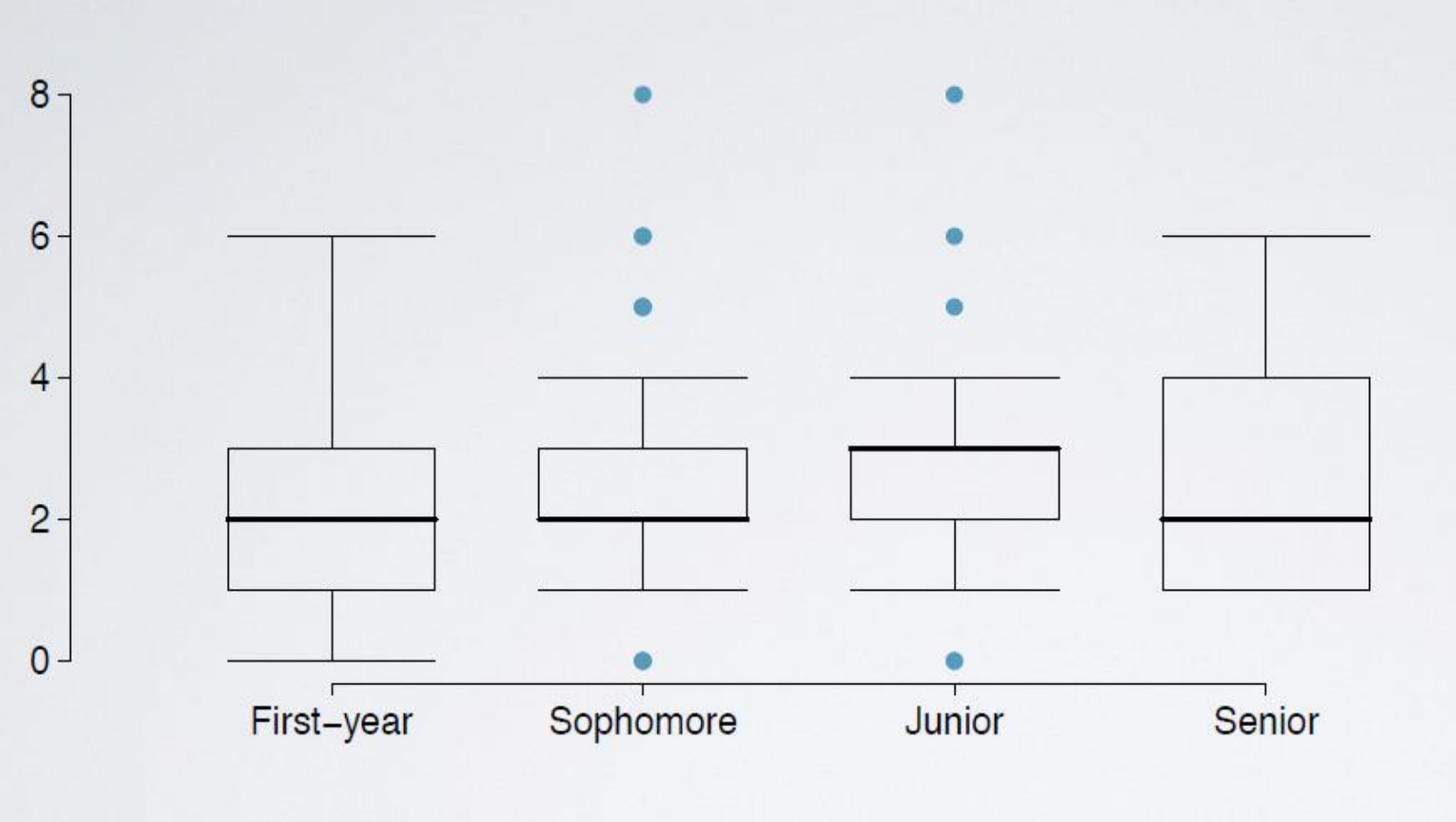
Wykres słupkowy względnej częstości



Wykres mozaikowy



Kilka wykresów pudełkowych



Wstęp do statystyki

- Przykład: Badanie hipotezy o dyskryminacji kobiet
  - 48 mężczyzn-kierowników analizowało te same akta personalne aby podjęć decyzję o promocji
  - Akta były identyczne, różna była tylko płeć
  - Losowo rozdzielono akta pomiędzy kierowników
  - 35/48 akt uzyskało promocje

  - Czy kobiety były dyskryminowane?

Dane

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

% of males promoted = 21/24 ≈ 88%

% of females promoted = 14/24 ≈ 58%

Dwa możliwe wnioski

- „Hipoteza O” (null hypothesis).
  - Dyskryminacja nie występuje, jest to losowa fluktuacja
- „Hipoteza A” (alternative hypothesis)
  - Tak kobiety były dyskryminowane, wynik jest statystycznie znaczący
- Test statystyczny: może stwierdzić że nie ma podstaw aby wyeliminować H_0 lub stwierdzić że są podstawy i przyjąć H_A

Jak testujemy hipotezę

- Startujemy z H_0 przyjmując że reprezentuje „status quo”
- Formułujemy hipotezę H_A w postaci pytania na które chcemy odpowiedzieć
- Przeprowadzamy test, zakładając że H_0 jest prawdziwe, albo przy pomocy symulacji lub rozważań teoretycznych
  - Jeżeli przeprowadzona symulacja nie daje przekonywującej ewidencji na H_A, przyjmujemy że odpowiedź jest H_0
  - Jeżeli daje, odrzucamy H_0 i przyjmujemy że odpowiedź jest H_A

Symulacja: weźmy talię kart

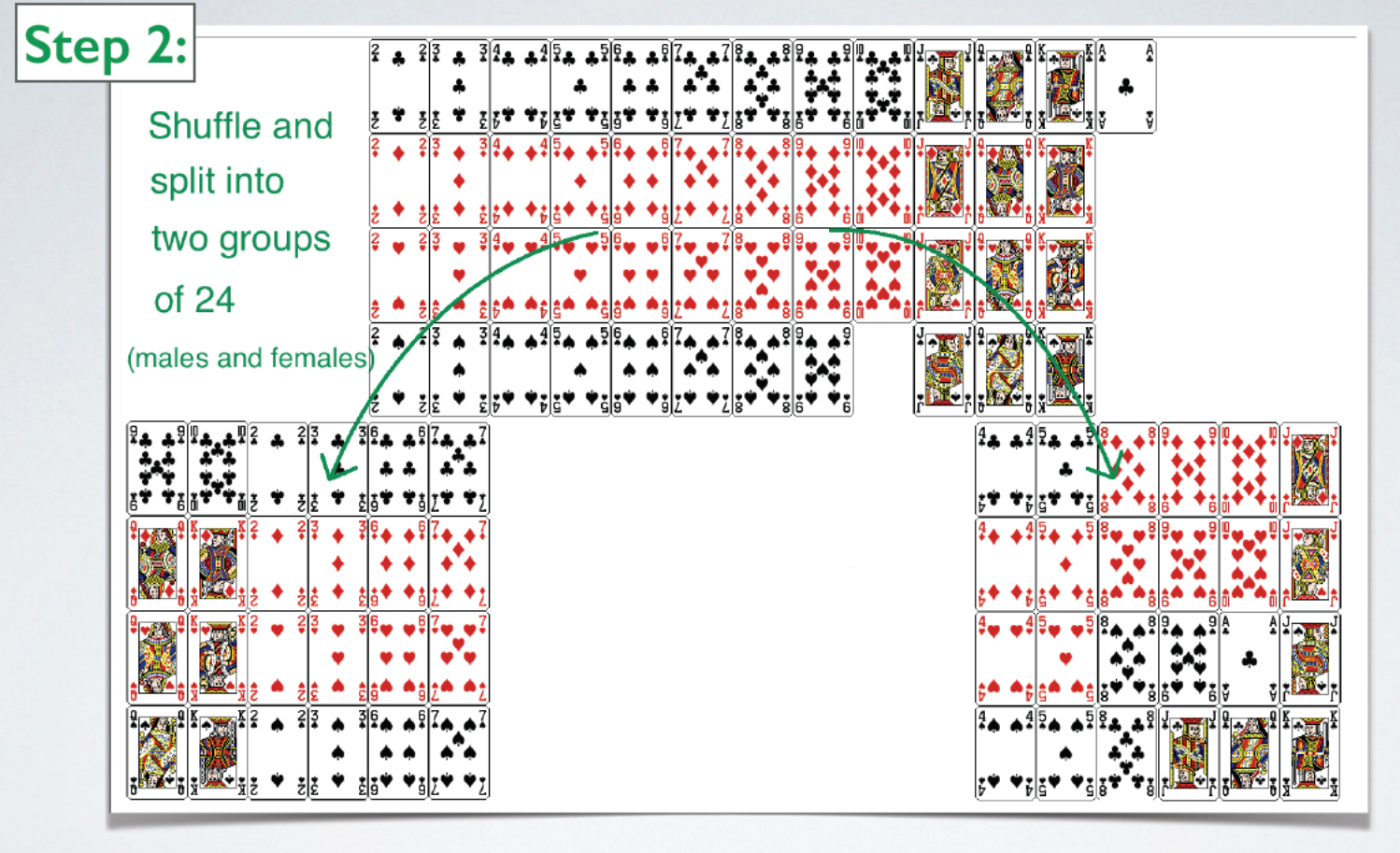
- „twarze” – reprezentują nie-promowanych, „liczby” – reprezentują promowanych
  - Odrzucamy jokery
  - Odrzucamy 3 Asy – zostaje 13 „twarzy” (A, K, D, W)
  - Odrzucamy 1 kartę z liczbą – zostaje 35 „liczb”

Przygotowujemy symulację

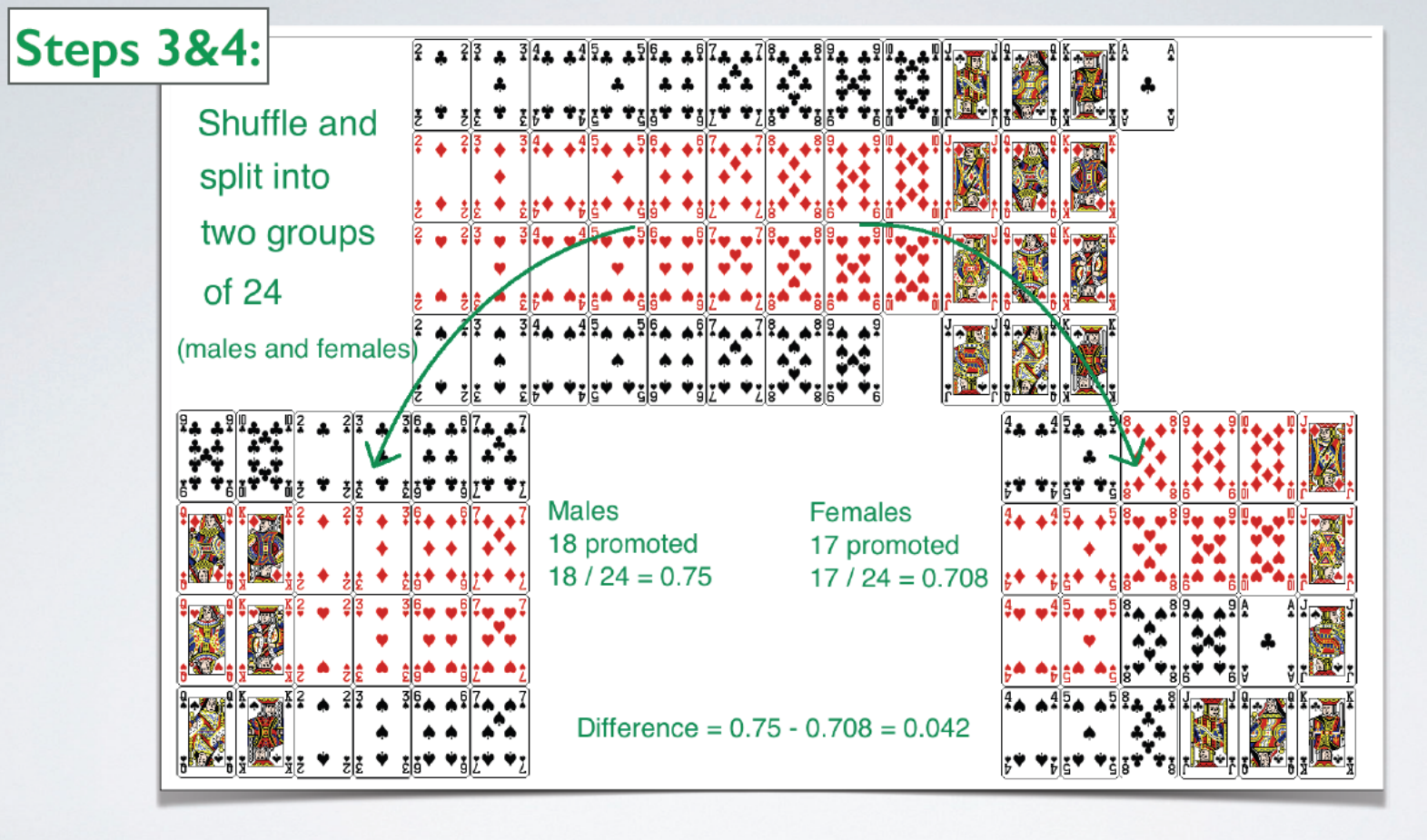
Kolejne kroki

1) Tasujemy karty
2) Dzielimy na dwie grupy (losowo) reprezentujące „mężczyzn” i „kobiety”
3) Liczymy ilość „liczbowych kart” w każdej grupie, reprezentuje to promocje
4) Liczmy proporcje promocji w każdej grupie, zapisujemy wynik różnicy w proporcji
5) Powtarzamy 1-4 wiele razy

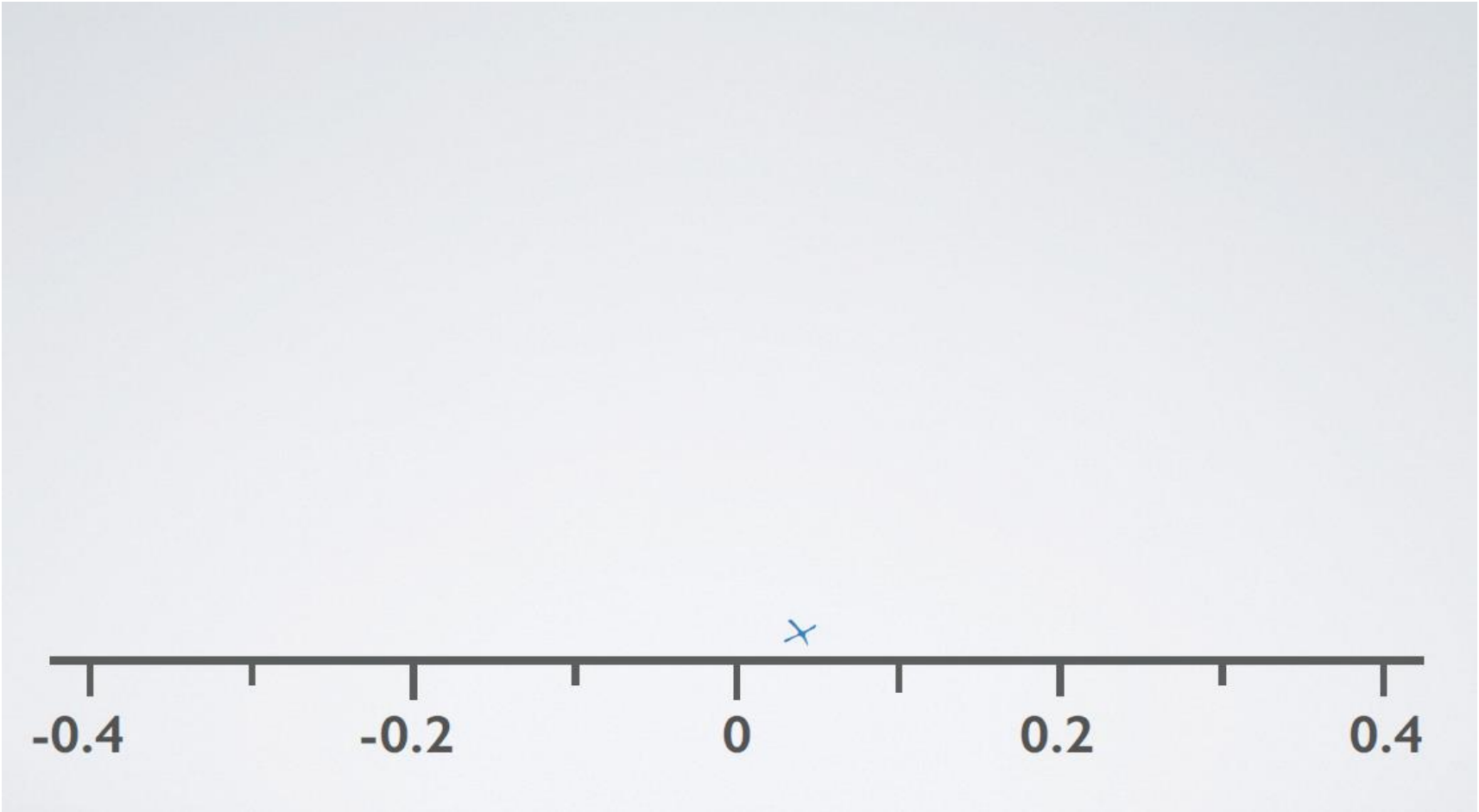
Tasujemy, dzielimy na grupy



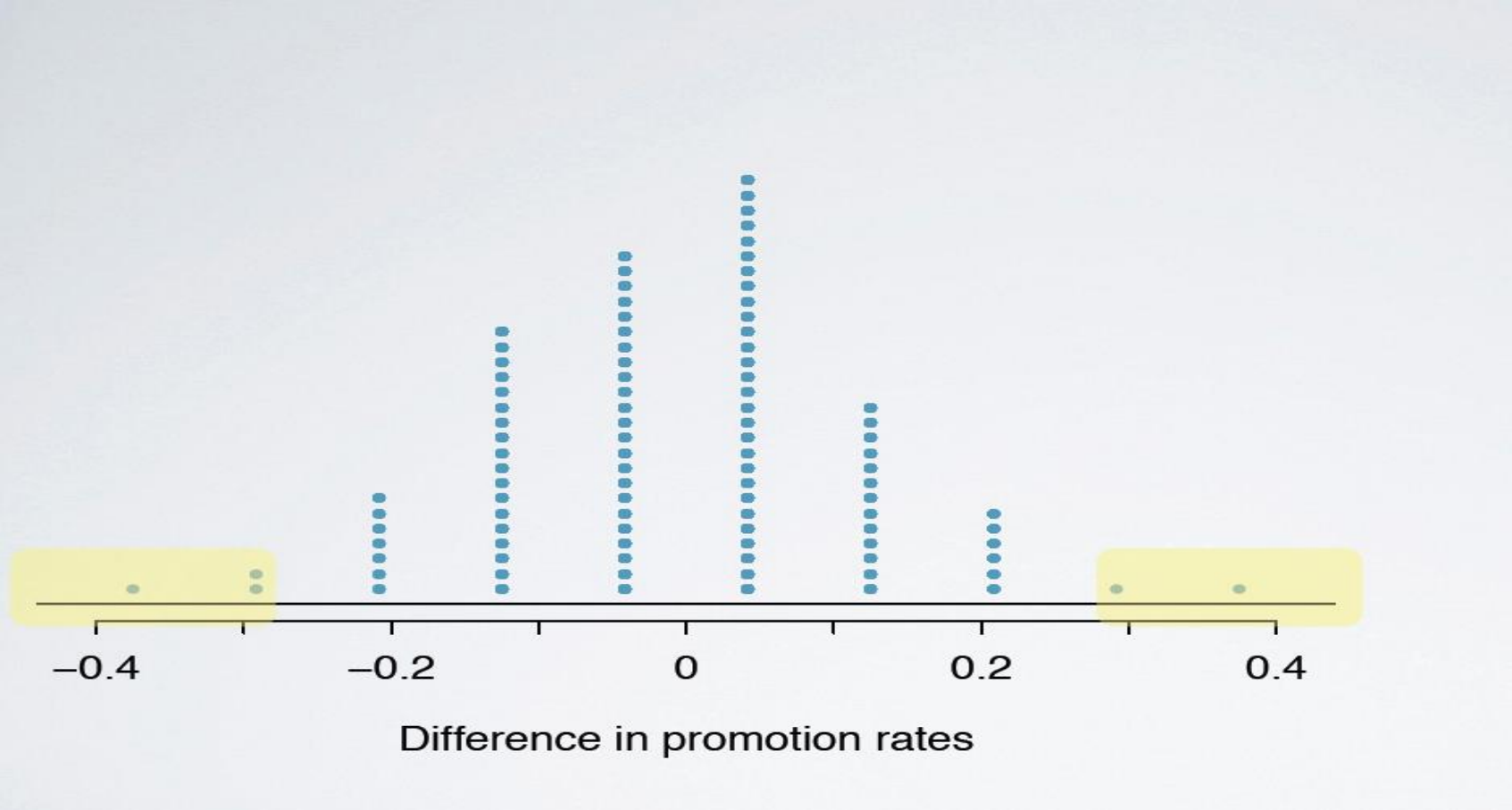
Liczymy proporcje promocji w każdej grupie



Zaznaczamy wynik na wykresie



Powtarzamy wiele razy



Interpretacja wyników

- Wyniki z symulacji nie wyglądają jak dane, przyjmujemy za prawdę hipotezę H_A
- Różnica 0.30 to daleki ogon rozkładu, możemy policzyć jej **p-value** czyli prawdopodobieństwo że wynik symulacji byłby taki jak dane lub bardziej ekstremalny.