

CZYTANIE DANYCH I STATYSTYKI

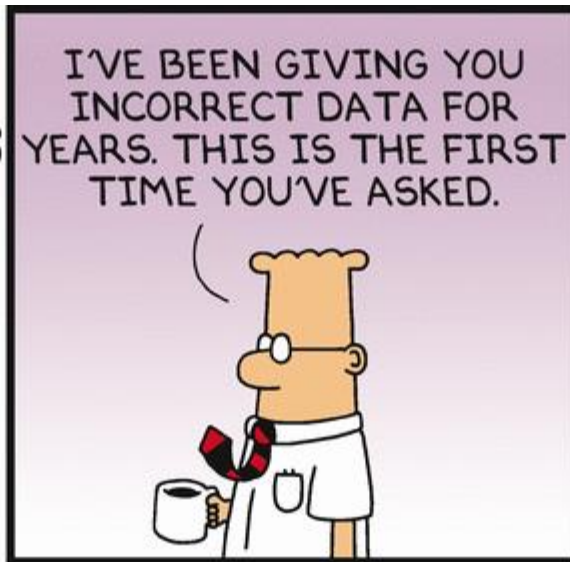
PRACA Z RAPORTAMI

Praca z raportami

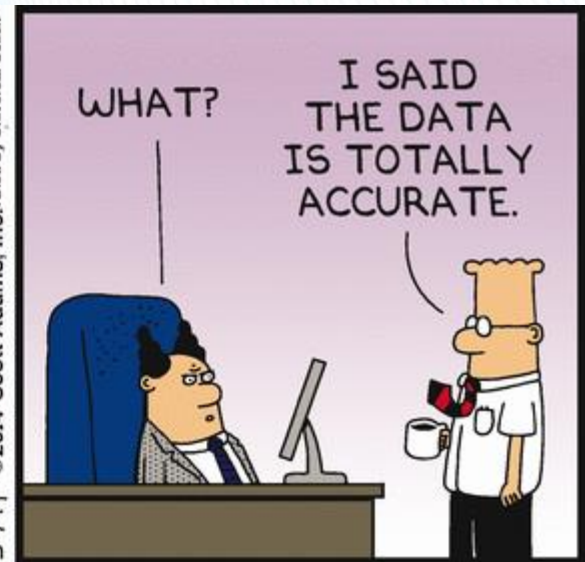
1. Interpretacja kluczowych elementów każdego raportu: dobór próby, operacjonalizacja pojęć teoretycznych, narzędzia badawcze
2. Najczęstsze błędy we wnioskowaniu statystycznym. Jak ich nie popełniać i nie dać się oszukać innym?
3. Data storytelling - jak opisywać wyniki w sposób zrozumiały i ciekawy dla wszystkich



Dilbert.com DilbertCartoonist@gmail.com



5-7-14 © 2014 Scott Adams, Inc./Dist. by Universal Uclick



Interpretacja raportu

Otrzymaliśmy na biurku raport, badanie lub publikację.

Skąd możemy wiedzieć, że analiza została wykonana rzetelnie?

Interpretacja raportu

Czynniki, na które warto zwrócić uwagę:

- Dobór próby
- Czy wyniki odpowiadają hipotezie badawczej?
- Do czego przyrównywane są wyniki?
- W jaki sposób zostały zebrane dane?
- Czy sposób prezentacji danych i wykresów nie wprowadza w błąd?

Rozmiar próby

Rozmiar próby powinien być na tyle duży, by wyniki były istotne statystycznie. Będzie to zależało od badania i wielkości populacji.

Przykładowy kalkulator:

<https://www.naukowiec.org/dobor.html>

Niestety badania bardzo często wykonywane są na próbach, których rozmiar nie jest uzasadniony w badaniu:

<https://bmjopen.bmj.com/content/9/10/e030312>

Istotność statystyczna

Co to znaczy, że wyniki są **istotne statystycznie**?

Oznacza to, że uzyskanie takiego wyniku przez przypadek jest mniej prawdopodobne niż przyjęty przez nas **poziom istotności**.

Często przyjmowanymi poziomami istotności jest 0.1, 0.05, 0.01

Wzrost Polaków

Przeczytałam, że średni wzrost Polaka wynosi 180cm. Chcąc zweryfikować tę hipotezę przeprowadziłam badanie i zmierzyłam 30 osób. Uzyskałam średnią 182cm.

Czy mogę powiedzieć, że średni wzrost Polaków nie wynosi 180cm?

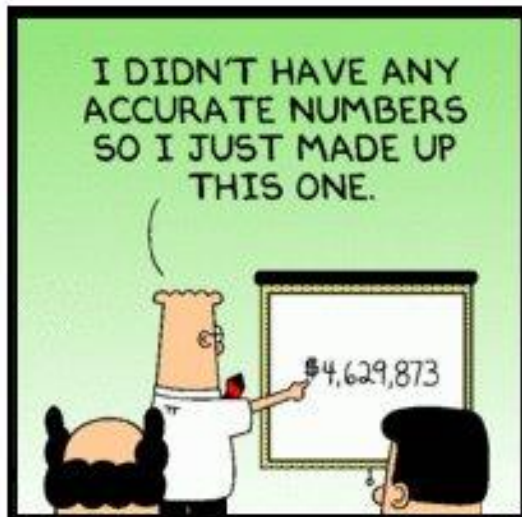
Wzrost Polaków

Po policzeniu odpowiednich statystyk okazuje się, że miałam ponad 14% prawdopodobieństwo na uzyskanie takiego wyniku jeśli Polacy faktycznie mierzą 180cm. W związku z tym nie mogę powiedzieć, że różnica między wynikami jest istotna statystycznie na żadnym z typowo przyjętych poziomów istotności.

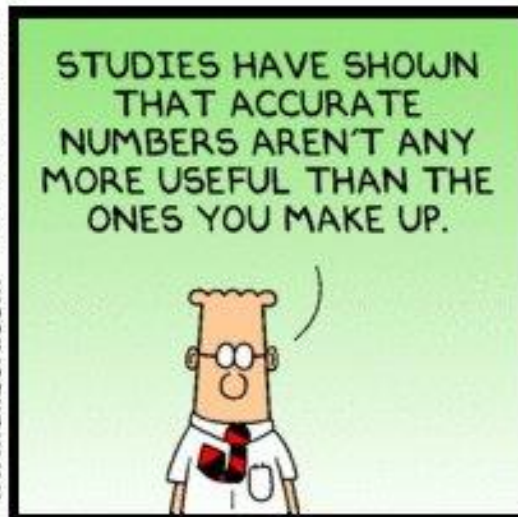
$$0.14 > 0.1$$

$$0.14 > 0.05$$

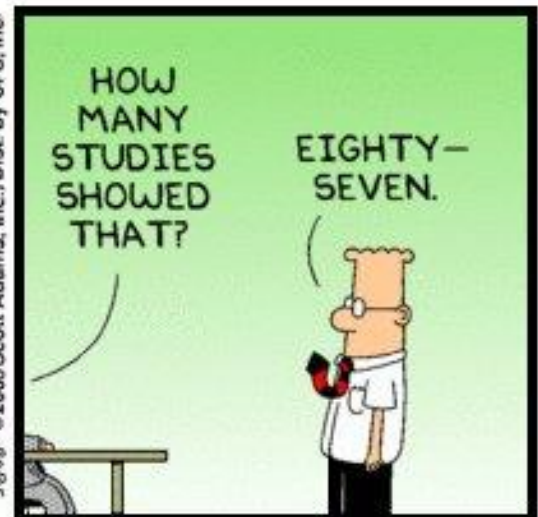
$$0.14 > 0.01$$



www.dilbert.com
scottadams@aol.com



5/8/08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.



© Scott Adams, Inc./Dist. by UFS, Inc.

80% Dentystów poleca pastę X

Takie stwierdzenie sugeruje, że pasta X to zdecydowany lider rynku.

- Czy dentysta mógł polecić jedną czy więcej past?

Odp: Więcej

- W takim razie jak często polecane były pasty konkurencji?

Odp: Tak samo często



Kto raportuje dane?



Proszę samodzielnie zważyć bagaż i podać wynik pracownikowi w stanowisku odprawy....

Błędy we wnioskowaniu statystycznym

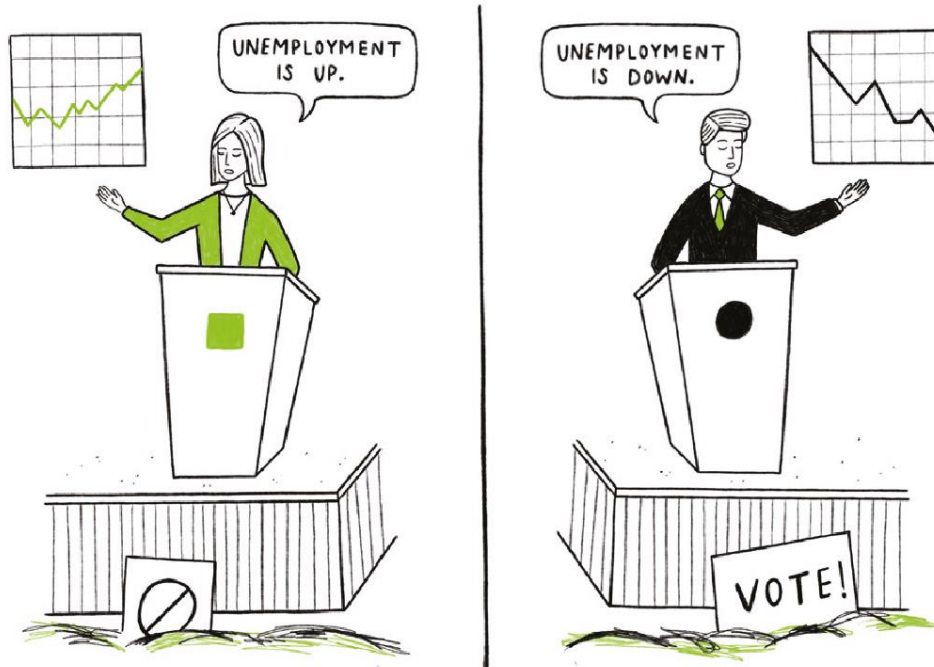
Mówiliśmy już o *oszustwach* na wykresach, zobaczmy, gdzie mogą pojawić się błędy w rozumowaniu.

Celowo lub nieumyślnie możemy wprowadzić w błąd innych (oraz siebie!) jeśli popełnimy jeden z opisanych niżej błędów we wnioskowaniu.

Mając świadomość jakie błędy są często popełniane nieco łatwiej będzie nam ich unikać.

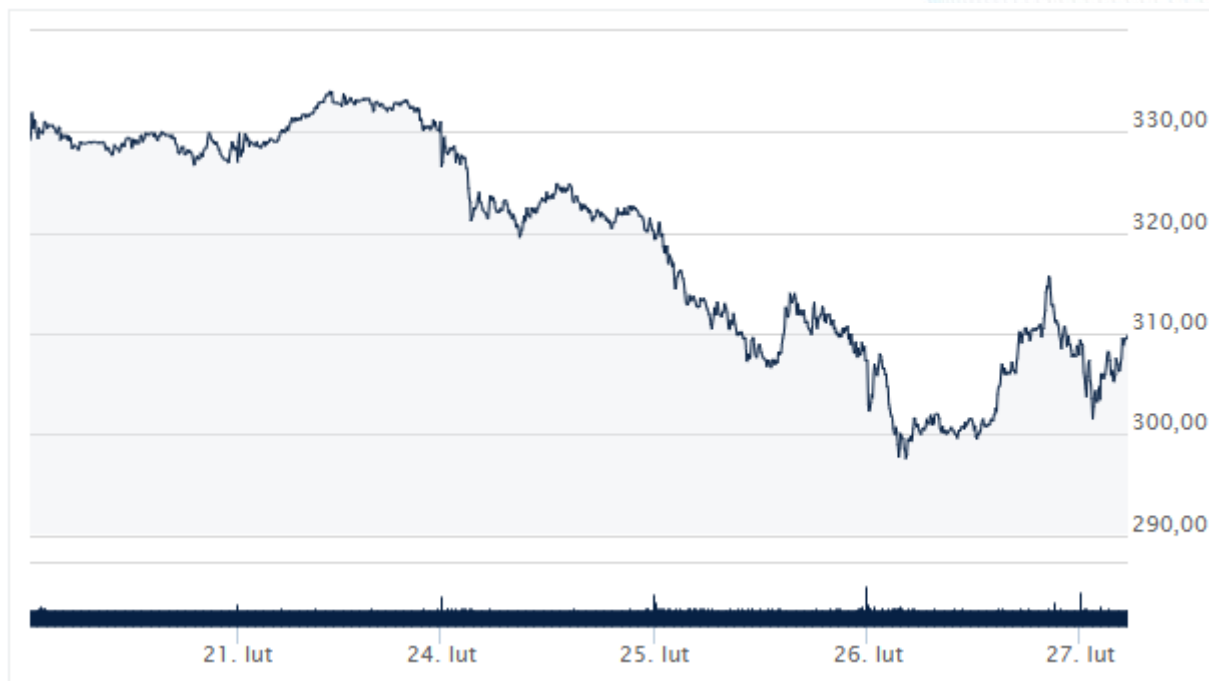
Źródło obrazków: <https://www.geckoboard.com/best-practice/statistical-fallacies/>

Cherry picking



Celowe lub przypadkowe dzielenie się wycinkiem danych potwierdzającym naszą hipotezę

Spadek akcji CDPROJEKTu

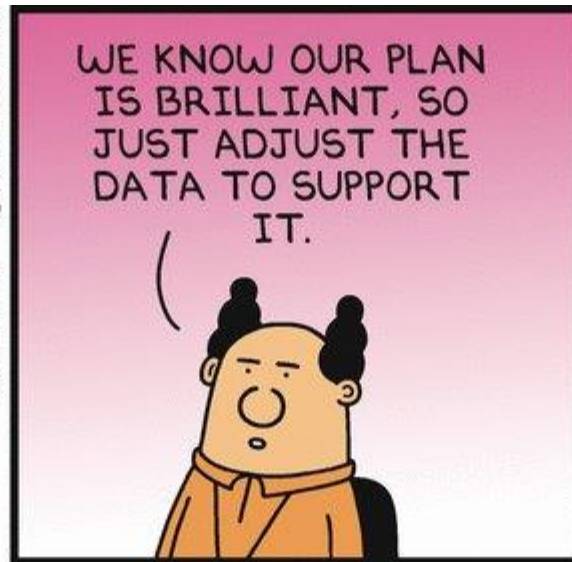


CDPROJEKT na fali wzrostów

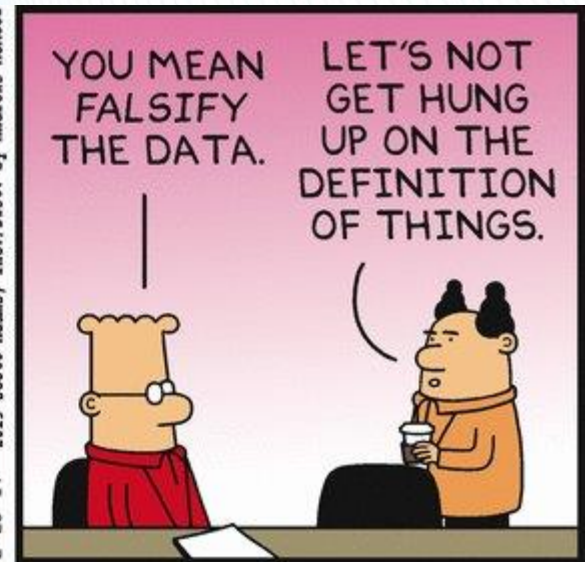




DILBERT.COM @SCOTTADAMSSAYS



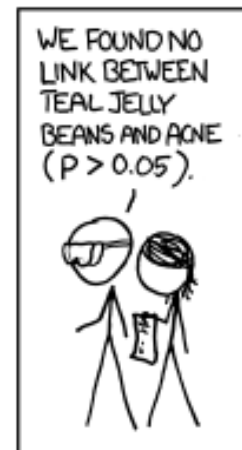
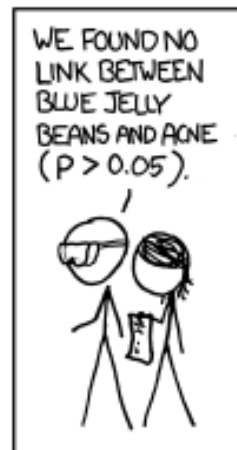
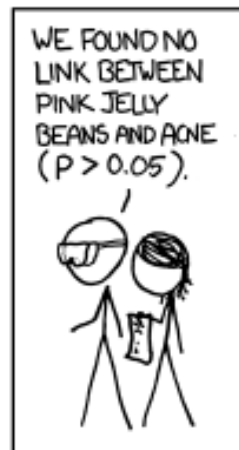
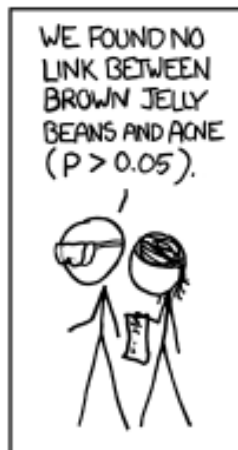
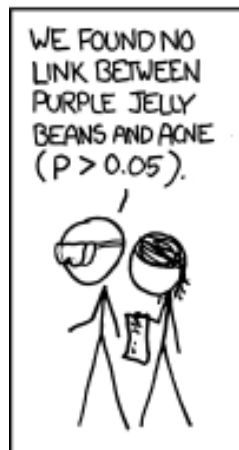
2-20-19 2019 Scott Adams, Inc./Dist. by Andrews McMeel



Data Dredging

Testowanie
różnych hipotez
tak długo, aż
znajdziemy
potwierdzenie
któreś z nich





WE FOUND NO
LINK BETWEEN
BEIGE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
LILAC JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BLACK JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
PEACH JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
ORANGE JELLY
BEANS AND ACNE
($P > 0.05$).




News

**GREEN JELLY
BEANS LINKED
TO ACNE!**

95% CONFIDENCE

**ONLY 5% CHANCE
OF COINCIDENCE!**

SCIENTISTS...

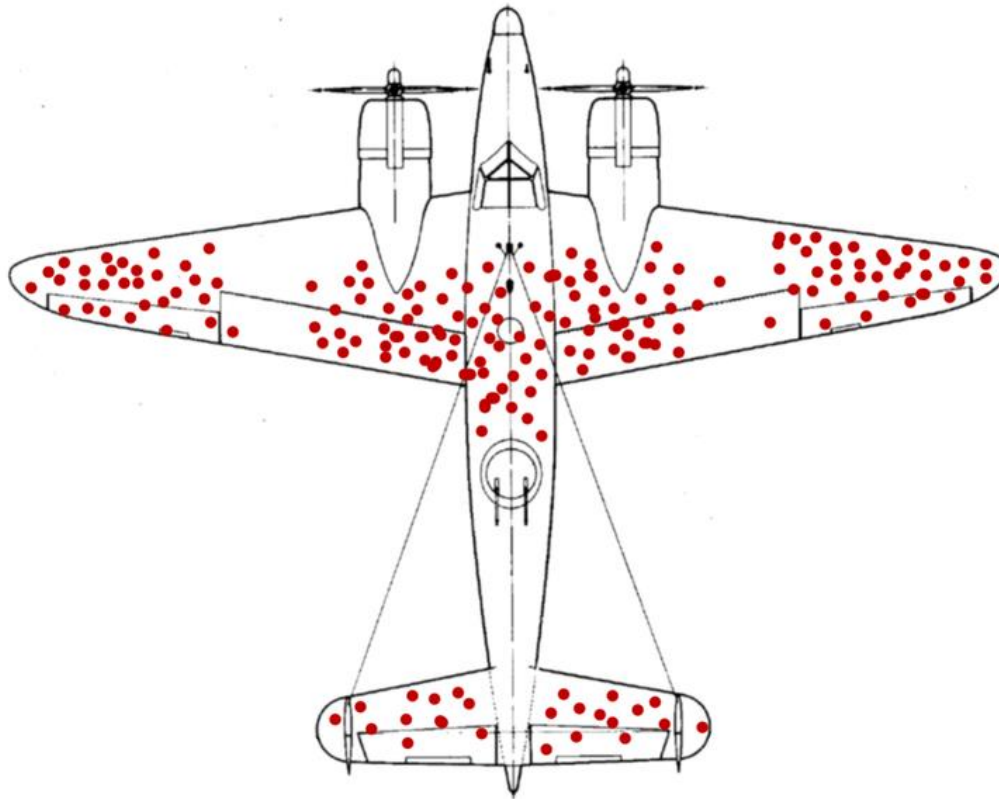


Survivorship bias

Błąd przeżywalności -
Podejmujemy
wnioski na
podstawie wcześniej
ocenzurowanych
danych

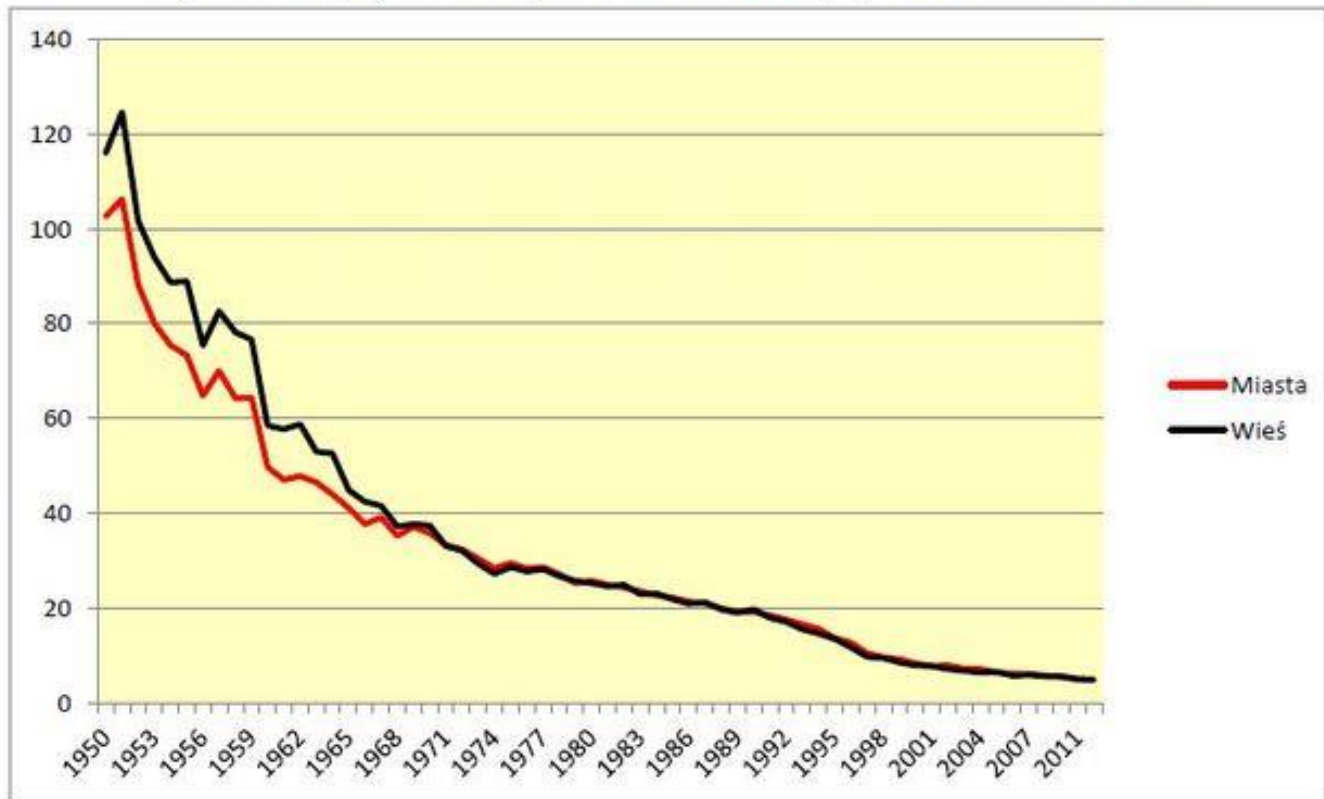


Statystyk Abraham Wald zwrócił w trakcie drugiej wojny światowej uwagę, że planowanie dodatkowych wzmocnień samolotów według uszkodzeń tych, które wróciły z lotów bojowych, jest błędem – najbardziej wrażliwe mogą być właśnie obszary, których uszkodzeń nie stwierdzono, ponieważ nie pozwoliły na powrót samolotu.



Za naszych czasów nie było X/Y/Z i żyjemy

Wykres 9. Zgony niemowląt na 1000 urodzeń żywych w latach 1950-2011



Inne przykłady?

- Słyszymy tylko o biznesmenach, aktorach i piosenkarzach, którzy odnieśli sukces
- „Już nie produkują takich [zegarków/pralek/wiertarek/...] jak kiedyś...” – porównujemy wszystkie współczesne produkty tylko z najlepszymi produktami z przeszłości, które przetrwały do dziś

Powiązane błędy

- Efekt potwierdzenia – szukamy tylko danych potwierdzających naszą tezę, stosujemy dowody anegdodyczne
- Błąd projekcji – nasze własne doświadczenia i preferencje przenosimy na całość populacji
- Rhyme-as-reason effect – stwierdzenia, które się rymują wydają nam się prawdziwe

Cobra Effect

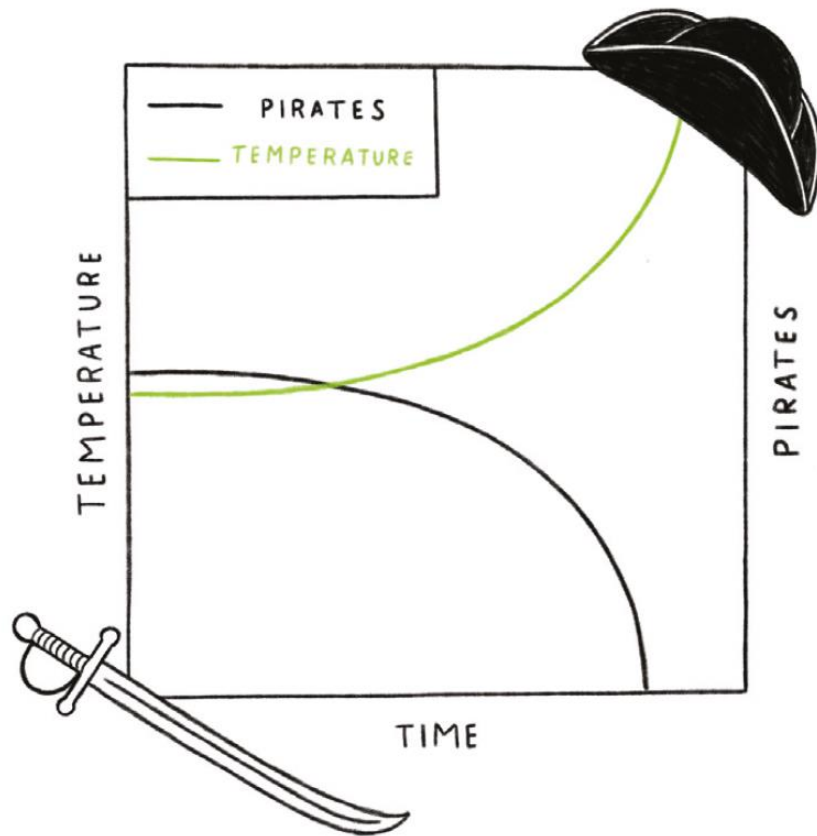
Nagroda za rozwiązanie problemu powoduje zwiększenie skali problemu



Przykład?

- W programie Medicare w USA lekarze przepisujący droższe leki otrzymują wyższe wynagrodzenie. W skutek tego lekarze zaczęli przepisywać droższe produkty nawet, gdy tańsza terapia była wystarczająca.

False casualty

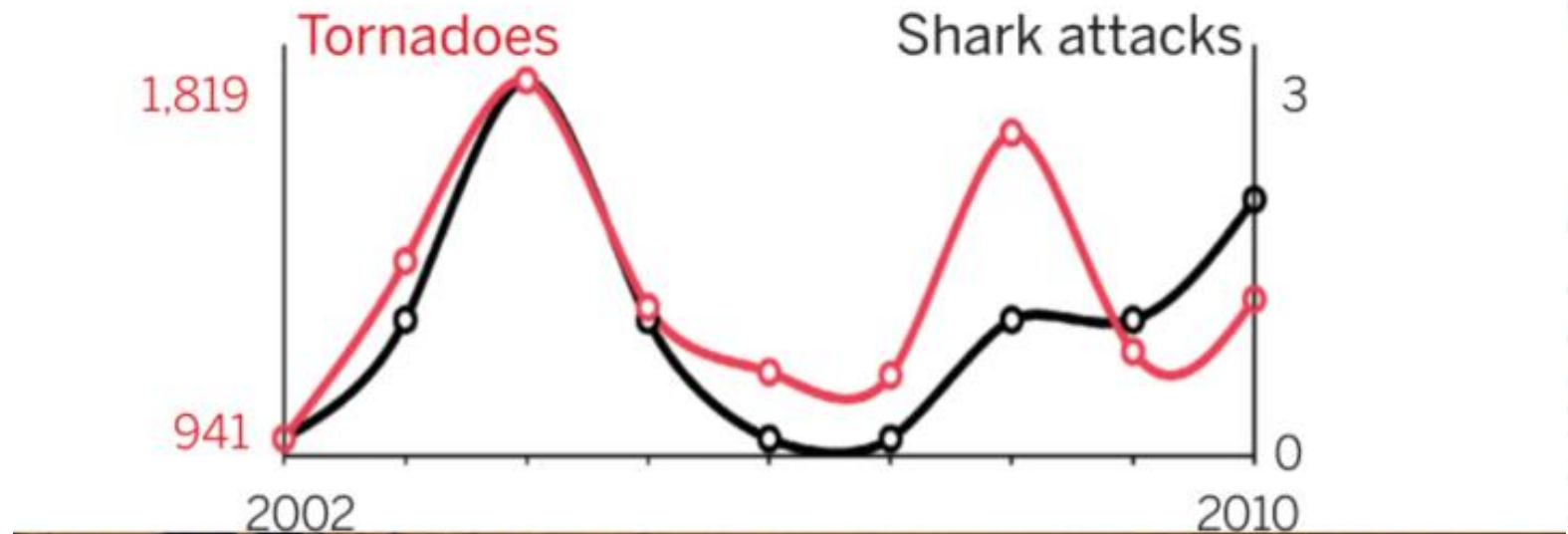


Nieprawidłowe przypisywanie zależności przyczynowo-skutkowej nie powiązanym czynnikom.

Film oparty na faktach?



Tornada a ataki rekinów w latach 2002-2010



Sampling bias

Próba nie
reprezentuje całej
populacji



Przykład

Chcemy zbadać nastawienie osób starszych do technologii. W celu przeprowadzenia badania rozsyłamy mailem ankietę internetową.



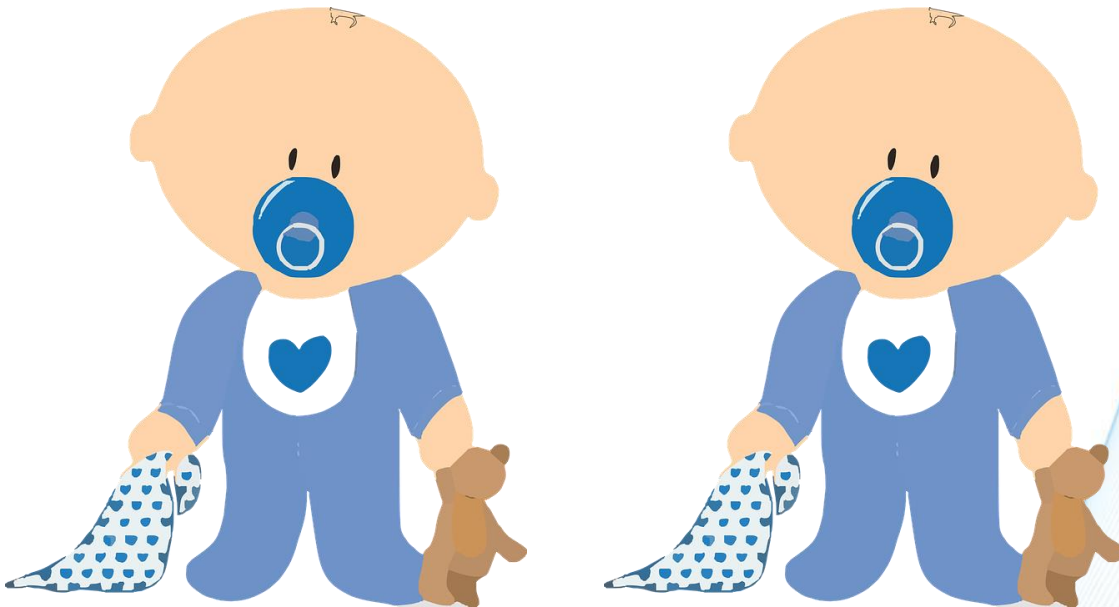
Gambler's fallacy



Paradoks hazardzisty –
traktowanie
niezależnych zdarzeń
losowych jako zdarzenia
zależne

Przykład

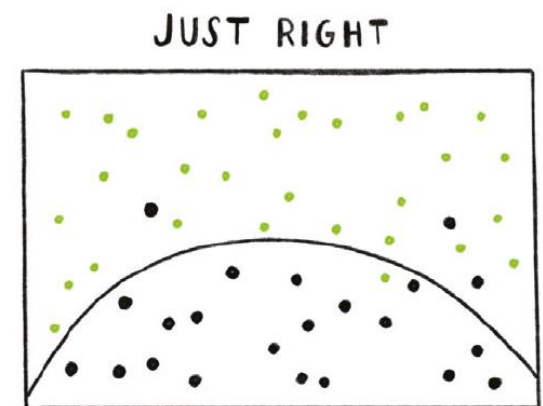
Sąsiedzi mają już dwóch synów, więc teraz urodzi im się córka.
Nieprawda, szanse są nadal 50/50



Overfitting

Nadmierne
dopasowanie modelu

Kiedy model jest zbyt
złożony może
wychwytywać
zależności, które
wystąpiły tylko w
próbie, ale nie w
populacji



Publication bias



Publikujemy
tylko wyniki
ciekawe, istotne
statystycznie

Przykład

Firma A: marketing internetowy okazał się u nas nieskuteczny

Firma B: marketing internetowy okazał się u nas nieskuteczny

Firma C: marketing internetowy okazał się u nas nieskuteczny

**Firma D: Marketing internetowy jest
kluczem do sukcesu!!!**

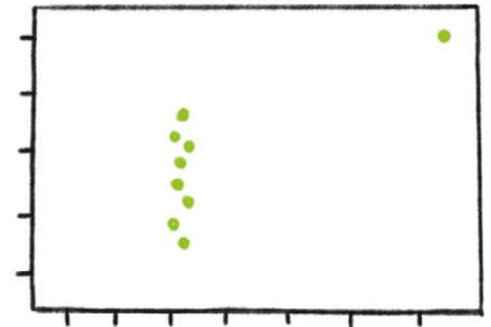
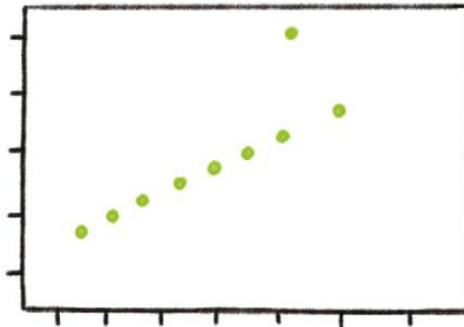
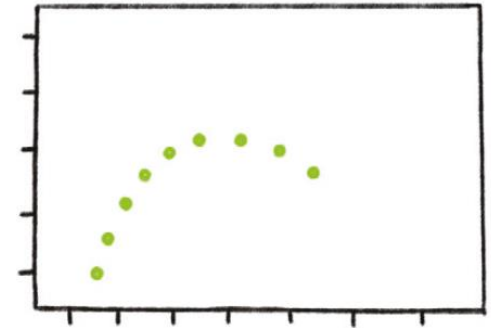
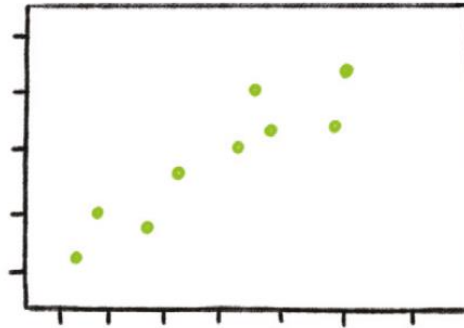


Anscombe's Quartet

Błąd polegający na patrzeniu tylko na statystyki opisowe

Te 4 zestawy danych mają identyczne:

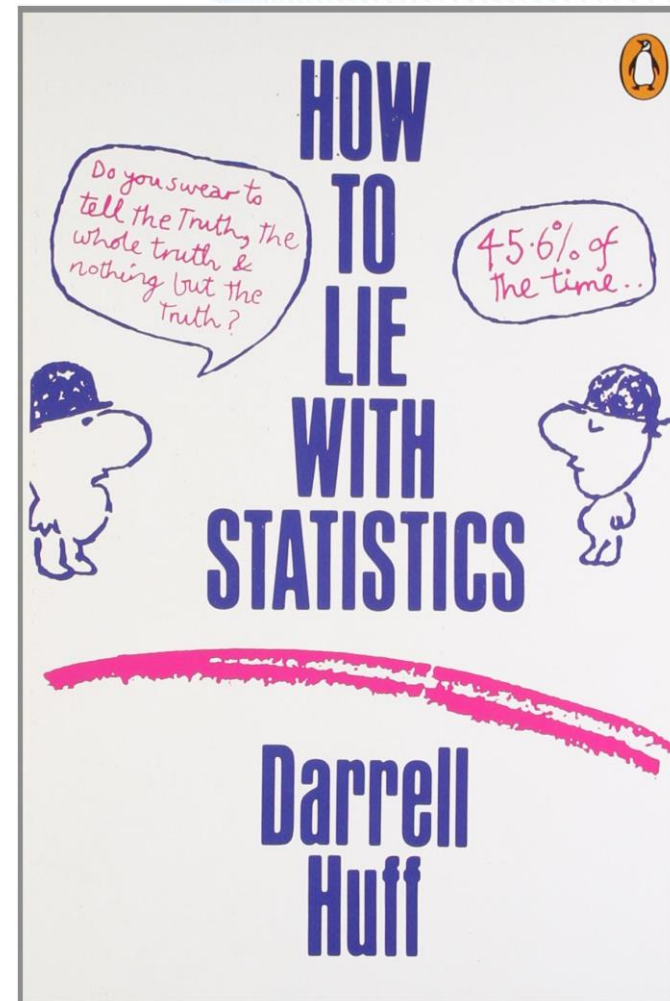
- Średnią
- Wariancję
- Regresję liniową



Więcej na ten temat

How to Lie with Statistics

Autor [Darrell Huff](#)



Data storytelling

Niezwykle ważne jest przedstawienie naszej idei w taki sposób, aby została wysłuchana a nasze rekomendacje wdrożone w życie.

Spójrzmy na kilka kroków efektywnej prezentacji raportu.

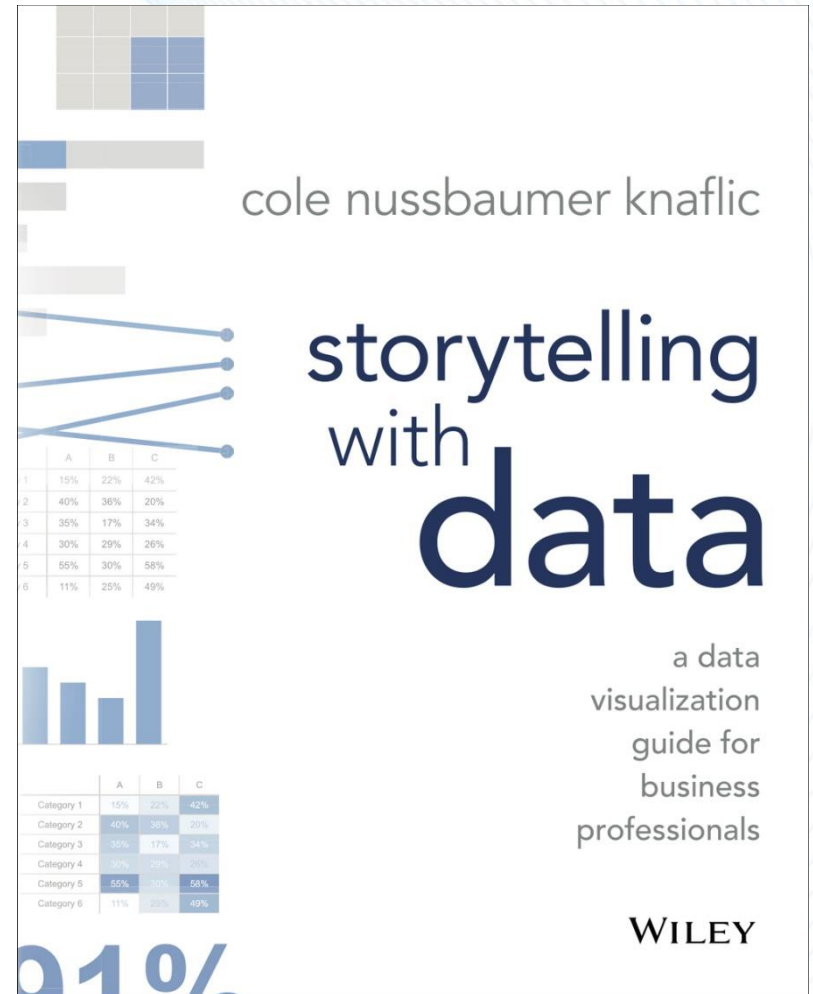
Data storytelling

1. Zdefiniuj cel swojej prezentacji
2. Przygotuj odpowiednie wizualizacje i tabele
3. Usuń zbędne elementy, które nie dodają wartości
4. Przyciągnij uwagę słuchaczy do informacji, które chcesz im przekazać
5. Dopracuj stronę estetyczną prezentacji
6. Ułóż kolejność historii, którą zamierzasz opowiedzieć

Więcej na ten temat

Storytelling with data

Autor Cole Nussbaumer Knaflic



Ćwiczenie

- Przygotujcie w grupach raport końcowy rekomendujący wybraną dzielnicę
- Dane znajdziecie w folderze Exercise