Jak badać rozkład jednej zmiennej?

Wstęp do Eksploracji Danych

Zmienne ilościowe

- height

```
> starwars$height
 [1] 172 167  96 202 150 178 165  97 183 182 188 180 228 180 173 175 170 180  66 170 183 200 190
[24] 177 175 180 150  NA  88 160 193 191 170 196 224 206 183 137 112 183 163 175 180 178  94 122
[47] 163 188 198 196 171 184 188 264 188 196 185 157 183 183 170 166 165 193 191 183 168 198 229
[70] 213 167  79  96 193 191 178 216 234 188 178 206  NA  NA  NA  NA  NA 165
```

- mass

```
> starwars$mass
 [1]   77.0   75.0   32.0  136.0   49.0  120.0   75.0   32.0   84.0   77.0   84.0     NA  112.0
[14]   80.0   74.0 1358.0   77.0  110.0   17.0   75.0   78.2  140.0  113.0   79.0   79.0   83.0
[27]     NA     NA   20.0   68.0   89.0   90.0     NA   66.0   82.0     NA     NA     NA   40.0
[40]     NA     NA   80.0     NA   55.0   45.0     NA   65.0   84.0   82.0   87.0     NA   50.0
[53]     NA     NA   80.0     NA   85.0     NA     NA   80.0   56.2   50.0     NA   80.0     NA
[66]   79.0   55.0  102.0   88.0     NA     NA   15.0     NA   48.0     NA   57.0  159.0  136.0
[79]   79.0   48.0   80.0     NA     NA     NA     NA     NA   45.0
```

Średnia

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

```
> mean(starwars$height, na.rm = TRUE)
[1] 174.358
```

Średnia obcięta/ucięta

```
> mean(starwars$height, trim = 0.2, na.rm = TRUE)
[1] 179.3265
```

Mediana

```
> median(starwars$height, na.rm = TRUE)
[1] 180
```

Odchylenie standardowe

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

```
> sd(starwars$height, na.rm = TRUE)
[1] 34.77043
```

Rozstęp

```
> range(starwars$height, na.rm = TRUE)
[1]  66 264
```

Rozstęp kwartylowy

```
> IQR(starwars$height, na.rm = TRUE)
[1] 24
```

Skośność
Kurtoza, miara spłaszczenia

```
> library(e1071)
> skewness(starwars$height, na.rm = TRUE)
[1] -1.025488
> kurtosis(starwars$height, na.rm = TRUE)
[1] 1.776414
```

Kwantyle

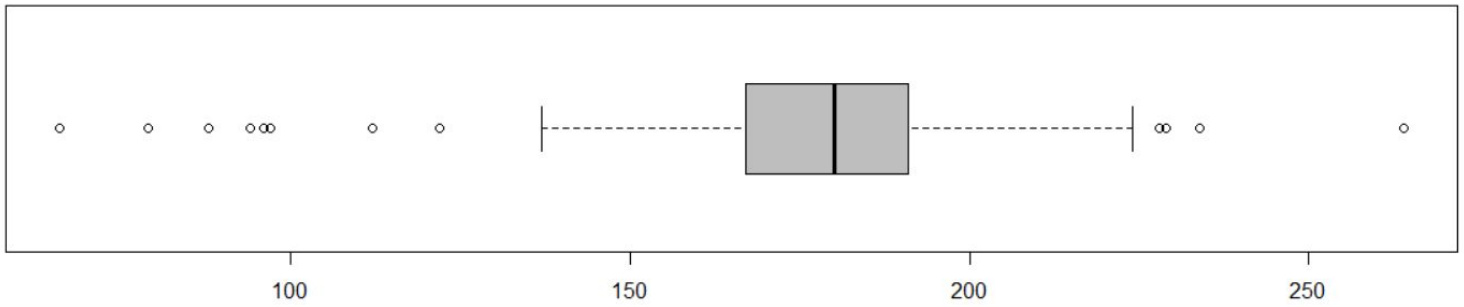
```
> quantile(starwars$height, c(0.1, 0.25, 0.5, 0.75, 0.9), na.rm = TRUE)
10% 25% 50% 75% 90%
122 167 180 191 206
```

Najważniejsze statystyki

```
> summary(starwars$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   66.0   167.0   180.0   174.4   191.0   264.0       6
```

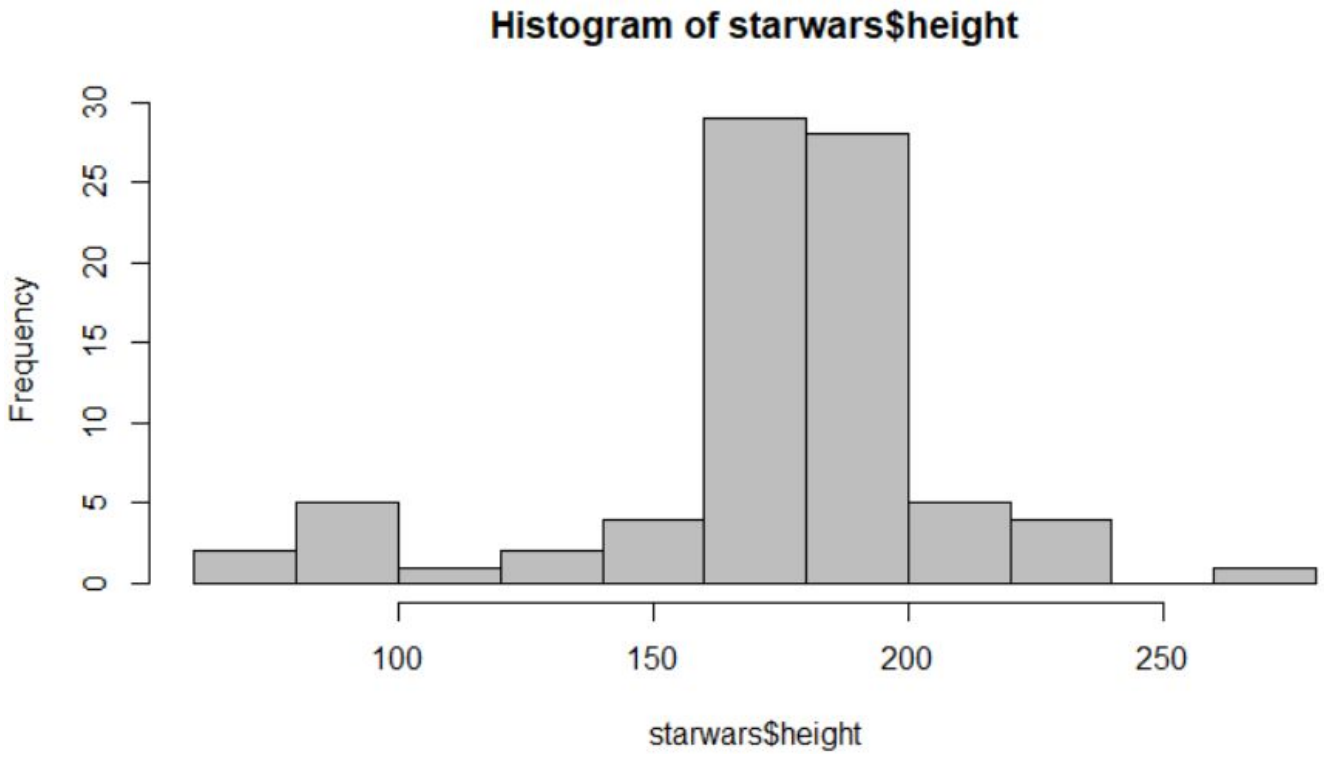
Wykres pudełkowy (boxplot)

```
> boxplot(starwars$height, col = "grey", horizontal = TRUE)
```



Histogram

```
> hist(starwars$height, col = "grey")
```



Dystrybuanta empiryczna

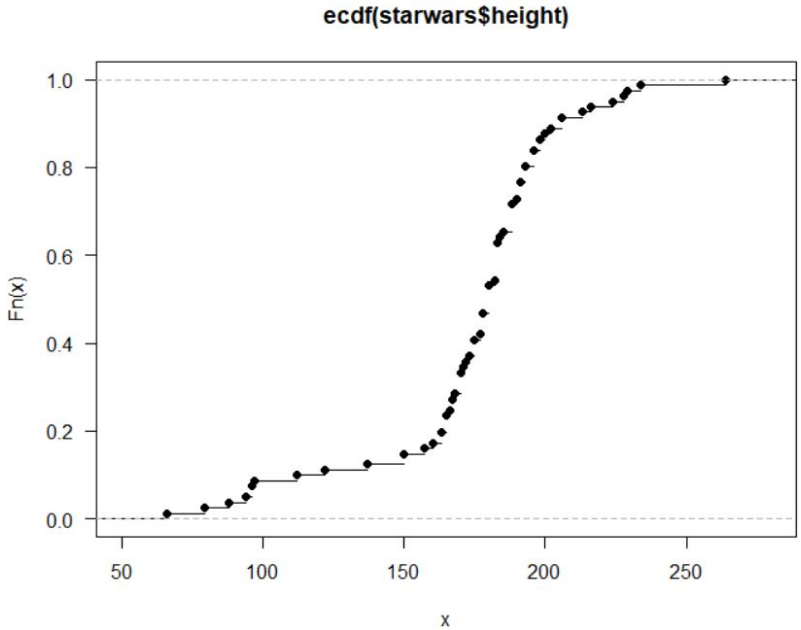
Definicja

Niech $X_1, X_2, \ldots, X_n$ będzie próbką z rozkładu o dystrybuancie F. Funkcję $\hat{F}_n : \mathbf{R} \times \Omega \to [0,1]$ określoną wzorem

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,\infty)}(x) = \frac{\#\{i : X_i \leqslant x\}}{n}$$

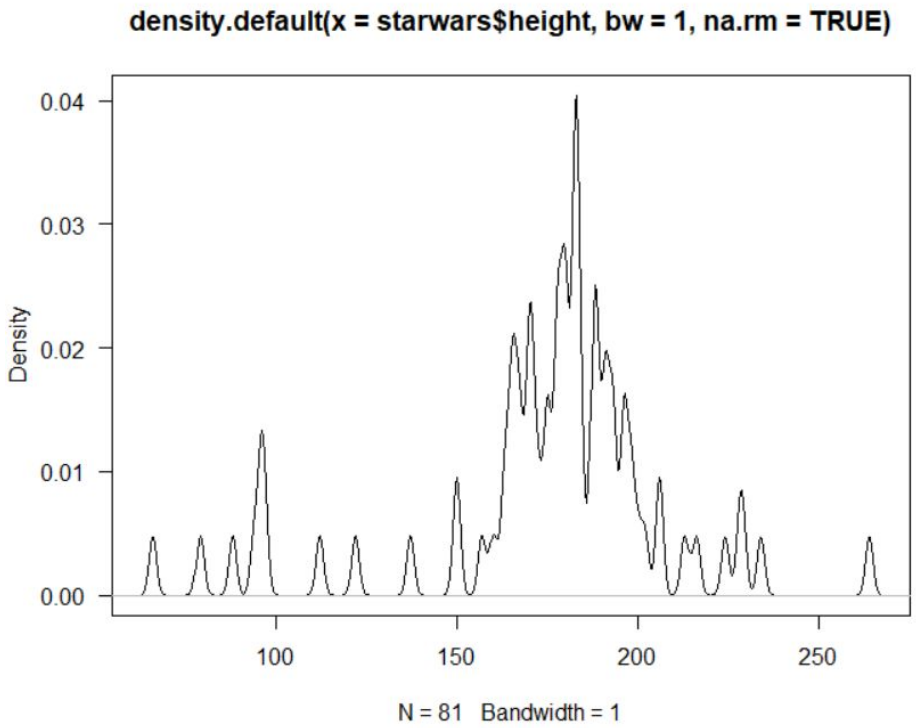
nazywamy *dystrybuantą empiryczną*.

```
> plot(ecdf(starwars$height), las = 1)
```



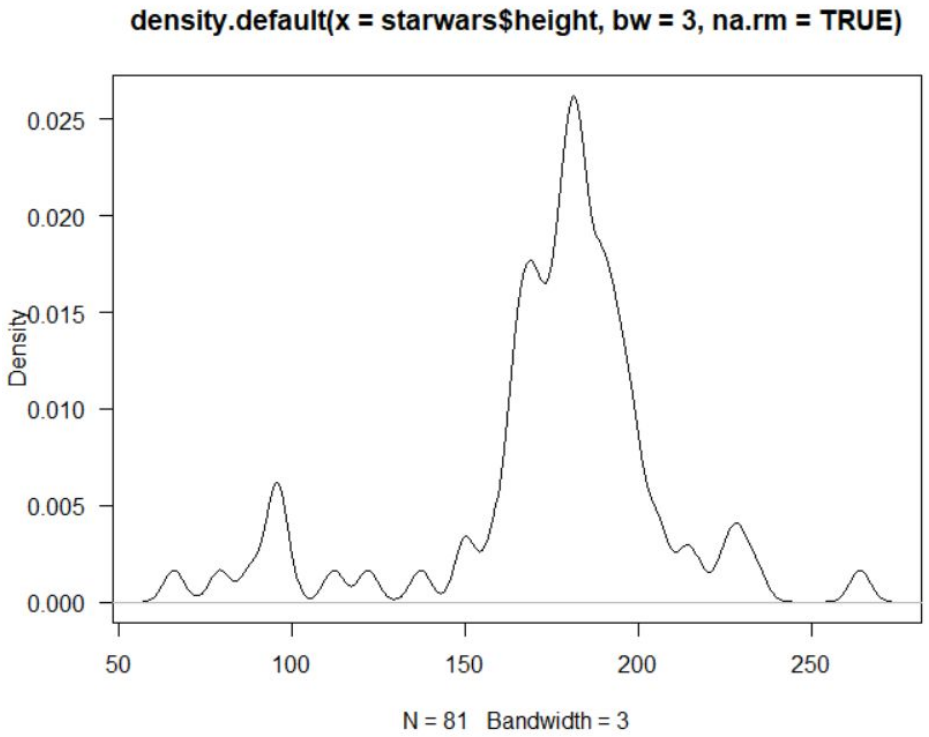
Jądrowy estymator gęstości

```
> plot(density(starwars$height, bw = 1, na.rm = TRUE), las = 1)
```



Jądrowy estymator gęstości

```
> plot(density(starwars$height, bw = 3, na.rm = TRUE), las = 1)
```



Zmienne jakościowe

- eye_color

```
> starwars$eye_color
 [1] "blue"           "yellow"         "red"          "yellow"       "brown"
 [6] "blue"           "blue"           "red"          "brown"        "blue-gray"
[11] "blue"           "blue"           "blue"         "brown"        "black"
[16] "orange"         "hazel"          "blue"         "brown"        "yellow"
[21] "brown"          "red"            "red"          "brown"        "blue"
[26] "orange"         "blue"           "brown"        "brown"        "black"
[31] "blue"           "red"            "blue"         "orange"       "orange"
[36] "orange"         "blue"           "yellow"       "orange"       "brown"
[41] "brown"          "yellow"         "pink"         "hazel"        "yellow"
[46] "black"          "orange"         "brown"        "yellow"       "black"
[51] "brown"          "blue"           "orange"       "yellow"       "black"
[56] "blue"           "brown"          "brown"        "blue"         "yellow"
[61] "blue"           "blue"           "brown"        "brown"        "brown"
[66] "brown"          "yellow"         "yellow"       "black"        "black"
[71] "blue"           "unknown"        "red, blue"    "unknown"      "gold"
[76] "black"          "green, yellow"  "blue"         "brown"        "white"
[81] "black"          "dark"           "hazel"        "brown"        "black"
[86] "unknown"        "brown"
```

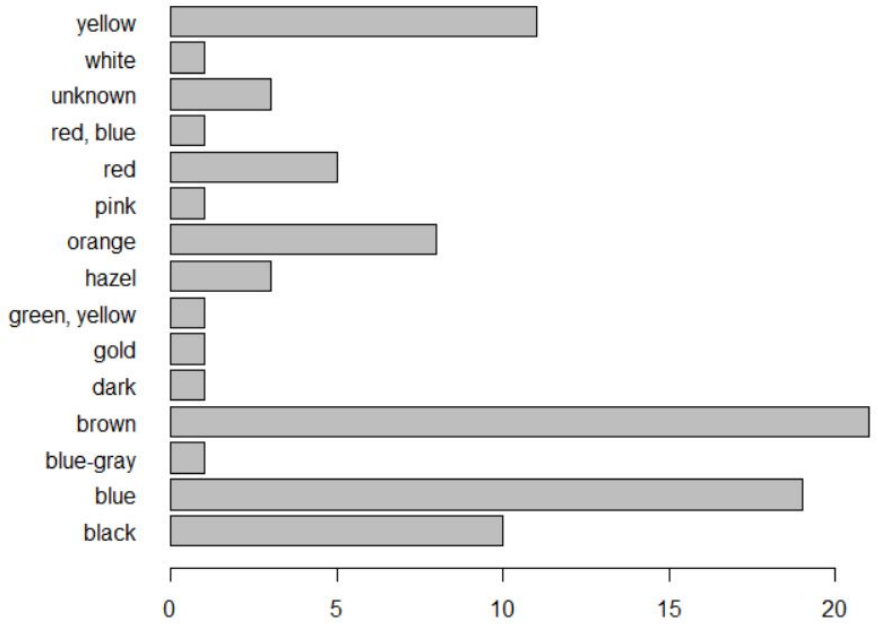
Tabela liczebności

```
> table(starwars$eye_color)

        black          blue     blue-gray         brown          dark          gold
           10            19             1            21             1             1
green, yellow         hazel        orange          pink           red     red, blue
            1             3             8             1             5             1
      unknown         white        yellow
            3             1            11
```

```
> proportions(table(starwars$eye_color))

        black          blue     blue-gray         brown          dark          gold
   0.11494253    0.21839080    0.01149425    0.24137931    0.01149425    0.01149425
green, yellow         hazel        orange          pink           red     red, blue
   0.01149425    0.03448276    0.09195402    0.01149425    0.05747126    0.01149425
      unknown         white        yellow
   0.03448276    0.01149425    0.12643678
```

Wykres słupkowy (paskowy)

```
> barplot(table(starwars$eye_color), horiz = TRUE, las = 1)
```



Kolory i skale

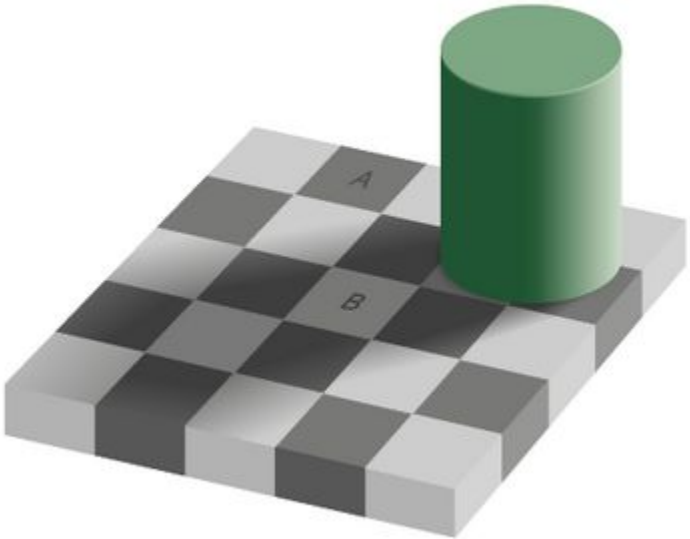
Kolory

Dlaczego dobór koloru jest ważny?

Podświadomie przypisujemy barwom znaczenia.

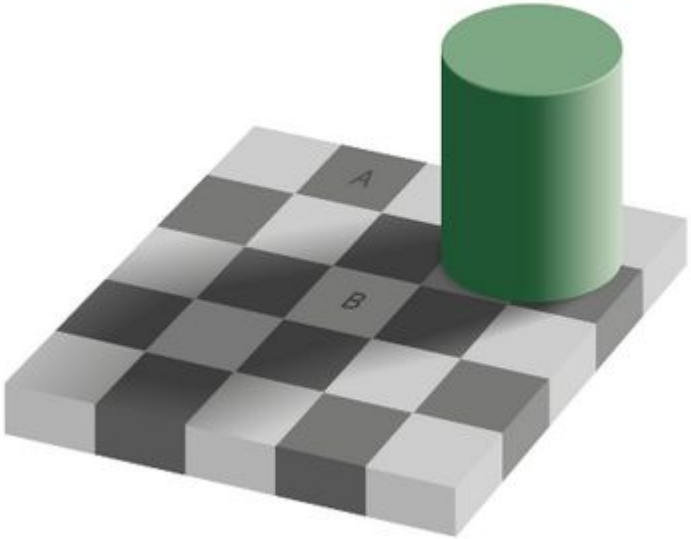
Postrzeganie barw różni się w zależności od oświetlenia jakości wydruku, ekranu lub projektora.

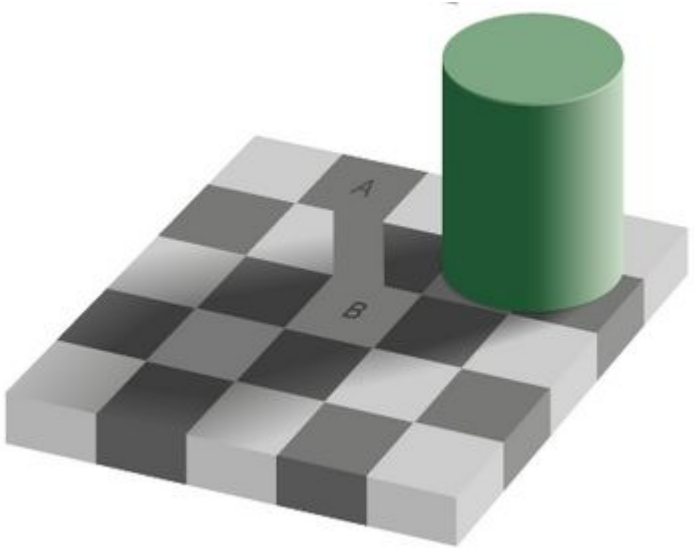
Wszystko jest względne



http://www.biecek.pl/Eseje/indexKolory.html

Wszystko jest względne

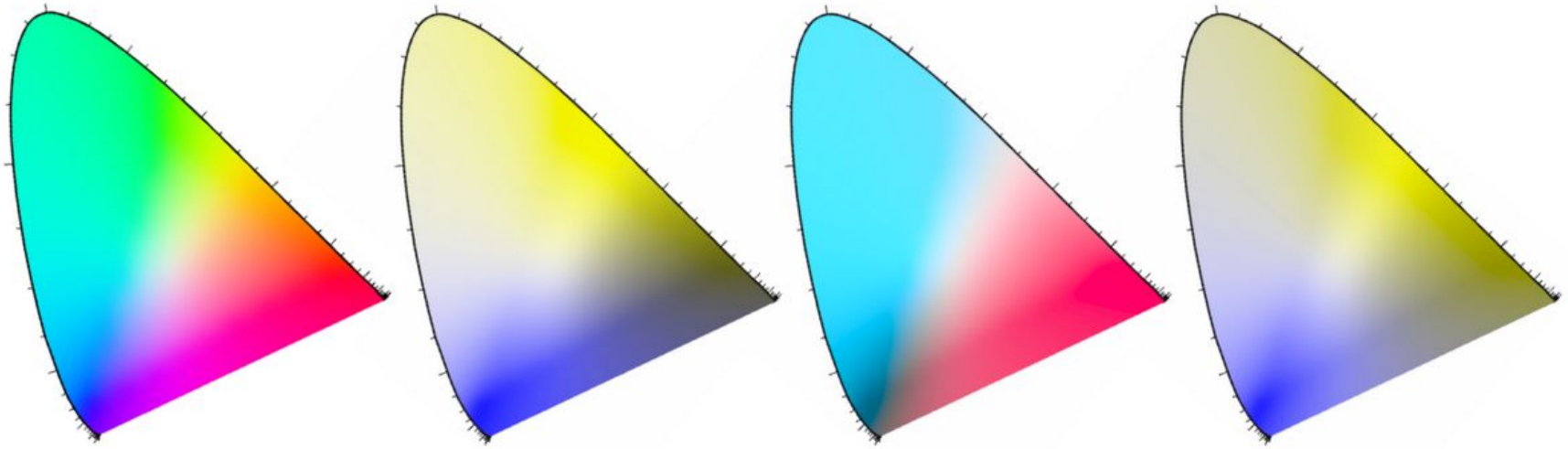




http://www.biecek.pl/Eseje/indexKolory.html

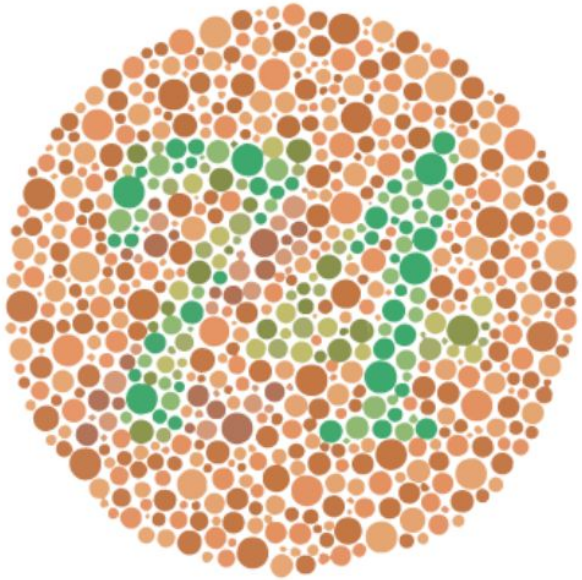
Wszystko jest względne

Zaburzenia w postrzeganiu barw



http://www.biecek.pl/Eseje/indexKolory.html

Zaburzenia w postrzeganiu barw



Wikipedia, domena publiczna

Zaburzenia w postrzeganiu barw

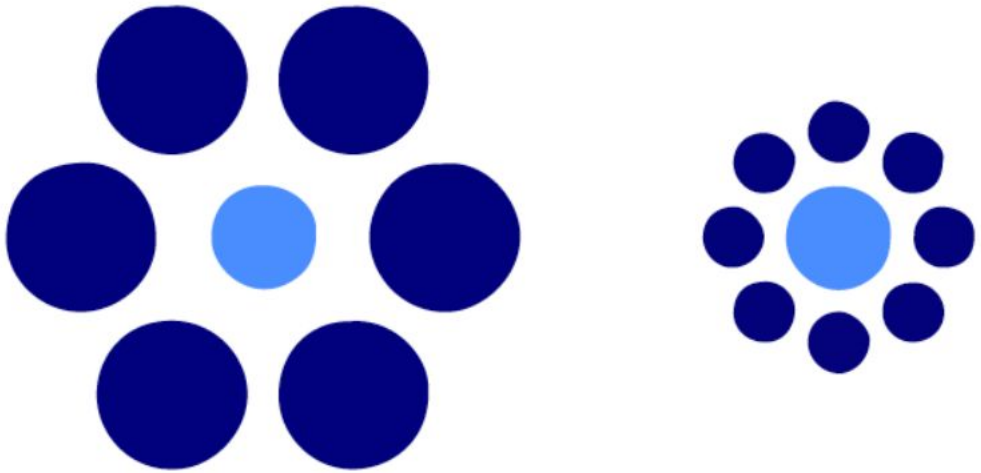
Jeżeli w naszej wizualizacji kolory pełnią kluczową funkcję to warto upewnić się, że przynajmniej osoby z typowymi dysfunkcjami widzenia kolorów będą w stanie odczytać informacje.

https://projects.susielu.com/viz-palette

Skale

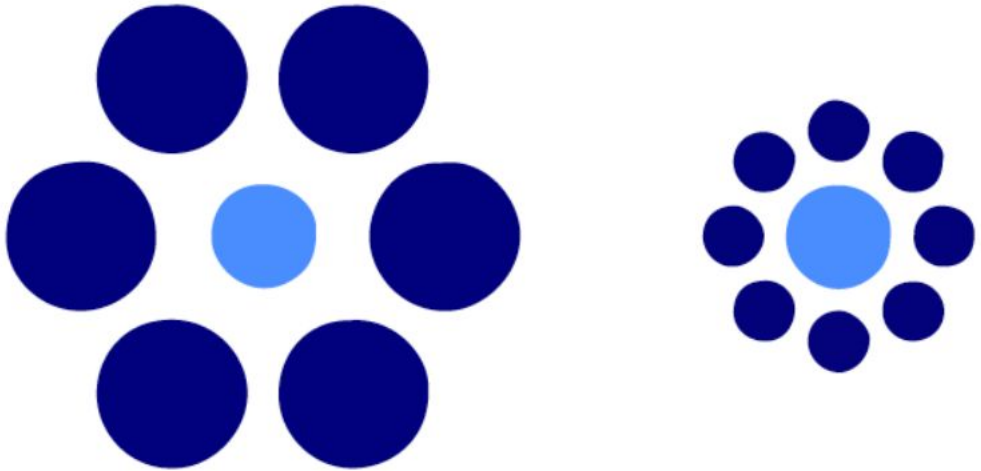
Pasma Macha

http://www.biecek.pl/Eseje/indexObraz.html



http://www.biecek.pl/Eseje/indexObraz.html

Iluzja Titchenera - Zniekształcenie postrzegania wielkości



http://www.biecek.pl/Eseje/indexObraz.html

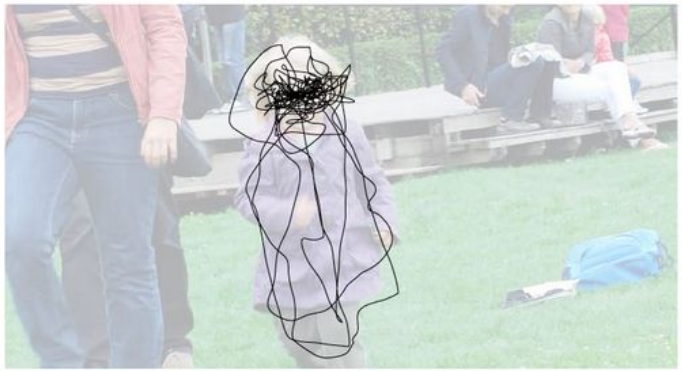
Obszar widzenia

http://www.biecek.pl/Eseje/indexObraz.html

Obszar widzenia

http://www.biecek.pl/Eseje/indexObraz.html

Fiksacje i sakady



http://www.biecek.pl/Eseje/indexObraz.html

Jak wiedza o strategiach przeglądania obrazu może pomóc w przygotowaniu lepszej prezentacji danych?

- Im więcej informacji, tym ważniejsze jest, by informacja była przedstawiana warstwowo.
- Jeżeli wykresowi towarzyszy słowna prezentacja, to warto powiedzieć, gdzie są interesujące elementy, ułatwi to ich lokalizację.
- Tytuł wykresu bardzo pomaga, ponieważ wstępnie informuje percepcję, czego wykres dotyczy i ułatwia wybór elementów wykresu do obserwacji.
- Wiele też zależy od tego, ile czasu odbiorca poświęci na analizę wykresu.

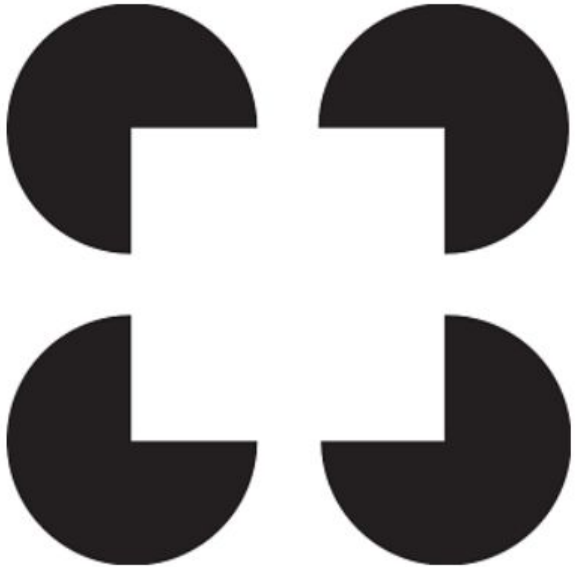
http://www.biecek.pl/Eseje/indexObraz.html

Iluzje

Pierwsza grupa iluzji dotyczy widzenia tego, czego nie ma. Im bardziej skomplikowany wykres, tym większa szansa, że coś przypadkowego zostanie uznane za ten “istotny” wzorzec.

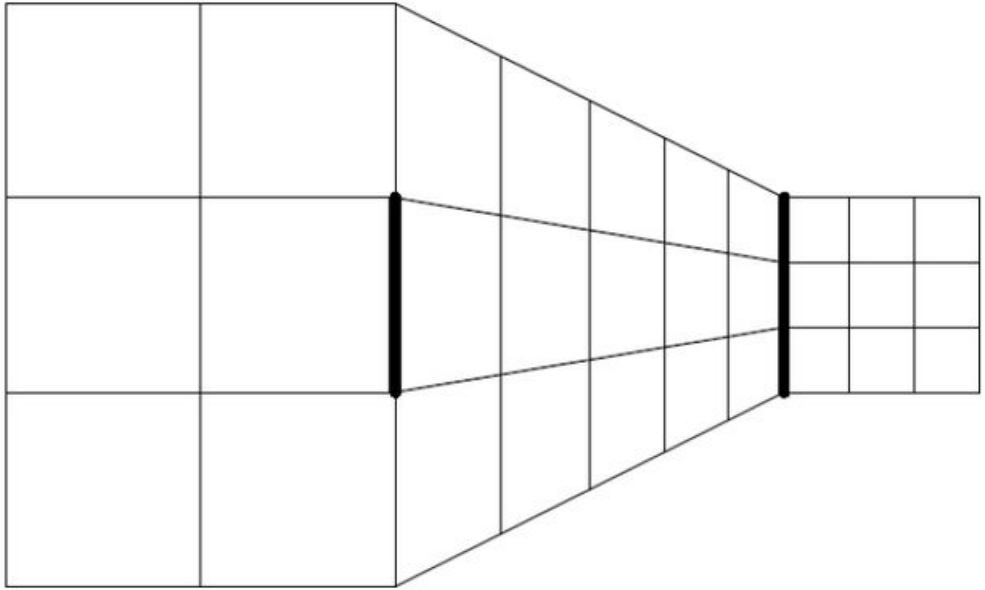
http://www.biecek.pl/Eseje/indexObraz.html

Widzenie tego, czego nie ma



http://www.biecek.pl/Eseje/indexObraz.html

Pseudo perspektywa



http://www.biecek.pl/Eseje/indexObraz.html

Pseudo perspektywa

Jeżeli na wykresie znajdzie się cokolwiek, co może sugerować perspektywę, to zostanie dostrzeżona przez mózg. Automatycznie wpłynie to na zniekształconą ocenę wielkości.

Dlatego wszelkim trójwymiarowym wykresom, czy to kołowym, słupkowym czy piramidowym powinniśmy zdecydowanie powiedzieć: **NIE**.

http://www.biecek.pl/Eseje/indexObraz.html

Problemy z kątami

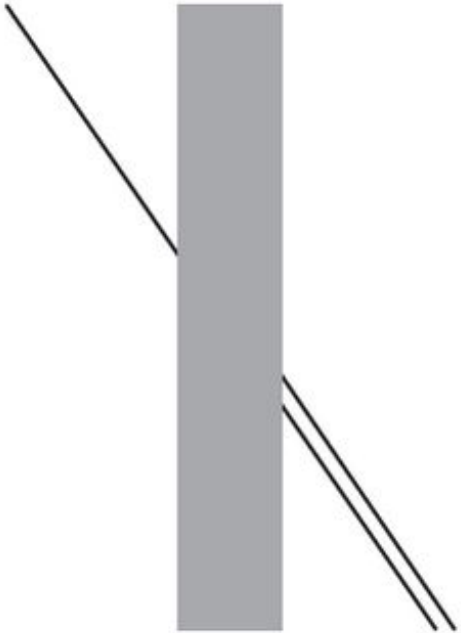
Pewnych charakterystyk mózg nie jest w stanie dobrze ocenić. Dobrym przykładem są kąty.

Ludzki mózg jest w stanie z dużą dokładnością ocenić, czy kąt jest bliski kątowi prostemu, ale ma duże problemy z oceną wielkości kątów ostrych i rozwartych.

Mózg ma skłonność do zawyżania wielkości kątów ostrych i zaniżania kątów rozwartych.

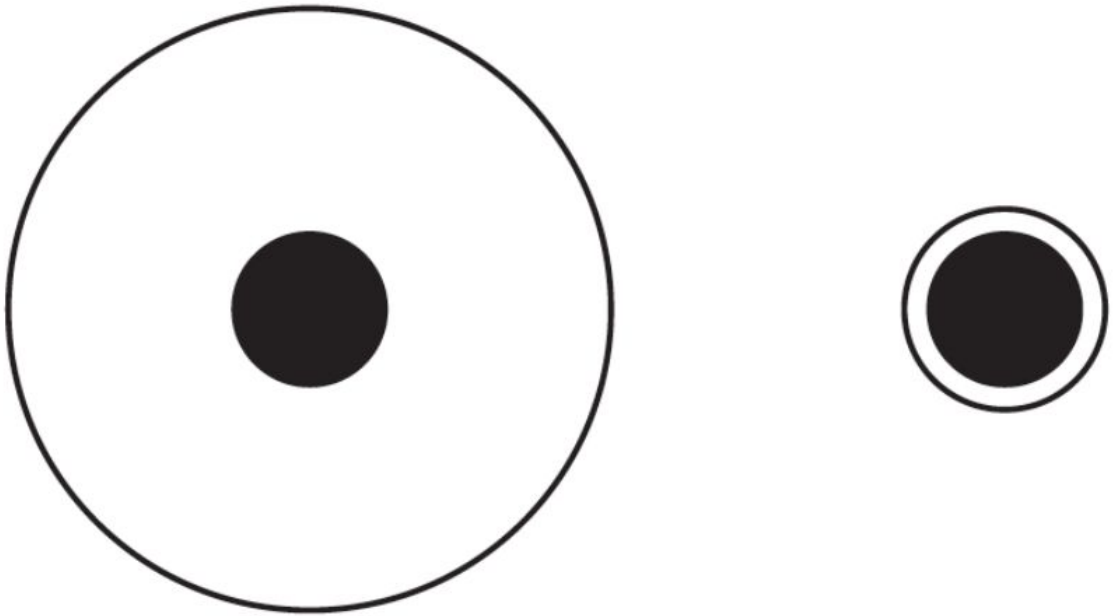
http://www.biecek.pl/Eseje/indexObraz.html

Problemy z kątami



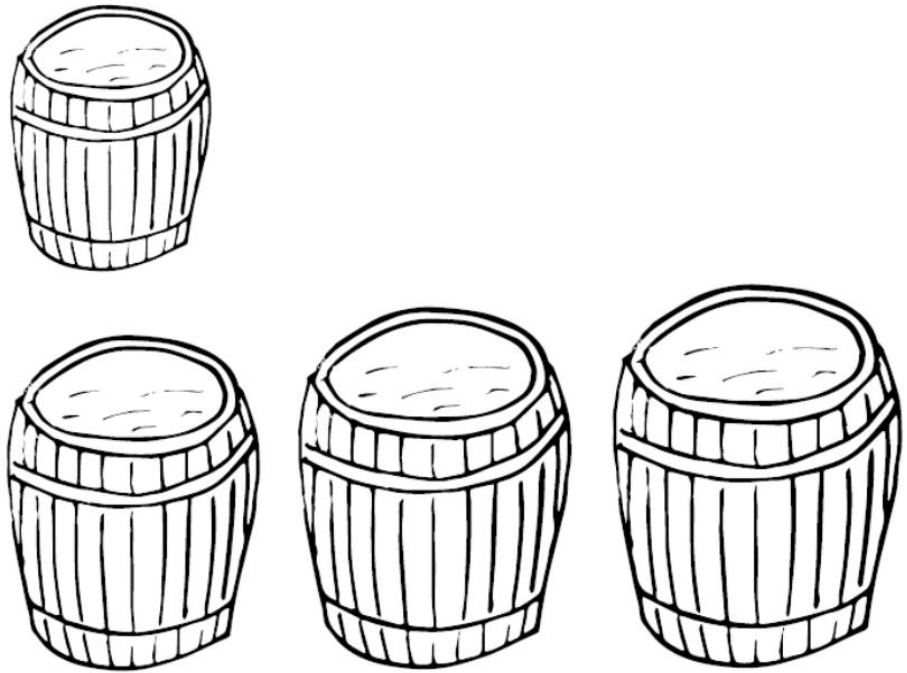
http://www.biecek.pl/Eseje/indexObraz.html

Co łączy dentystę i dietetyka?



http://www.biecek.pl/Eseje/indexObraz.html

Ocena wielkości



http://www.biecek.pl/Eseje/indexObraz.html

Które województwo jest większe?

http://www.biecek.pl/Eseje/indexObraz.html

Które województwo jest większe?

warmińsko-mazurskie

lubelskie

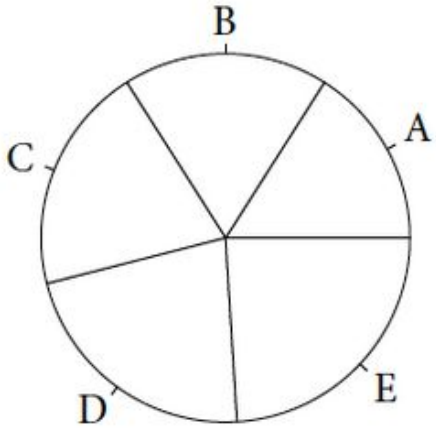
http://www.biecek.pl/Eseje/indexObraz.html

Hierarchia odczytywania charakterystyk

1. pozycje obiektów rozmieszczonych wzdłuż wspólnej skali (przykładowo wykres punktowy),
2. pozycje obiektów wzdłuż takiej samej, ale nie wspólnej skali (przykładowo sąsiadujące wykresy punktowe),
3. długości odcinków rozmieszczonych wzdłuż wspólnej skali,
4. długości odcinków wzdłuż takiej samej, ale nie wspólnej skali (o różnych punktach zaczepienia),
5. wielkości kątów i nachylenia (przy ocenie tempa wzrostu w wykresach liniowych),
6. powierzchnie,
7. objętości, gęstości, natężenia koloru,
8. sama barwa koloru.

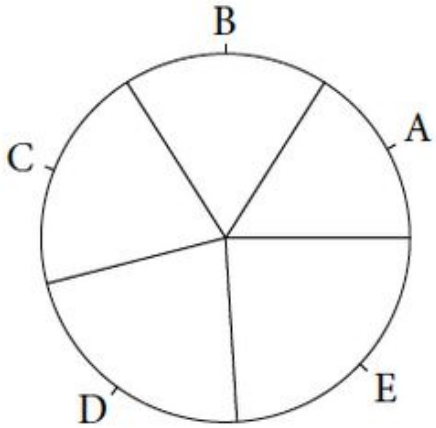
Na bazie The Visual Decoding of Quantitative Information on Statistical Graphs [Journal of the Royal Statistical Society Series A, 150:192–229, 1987]

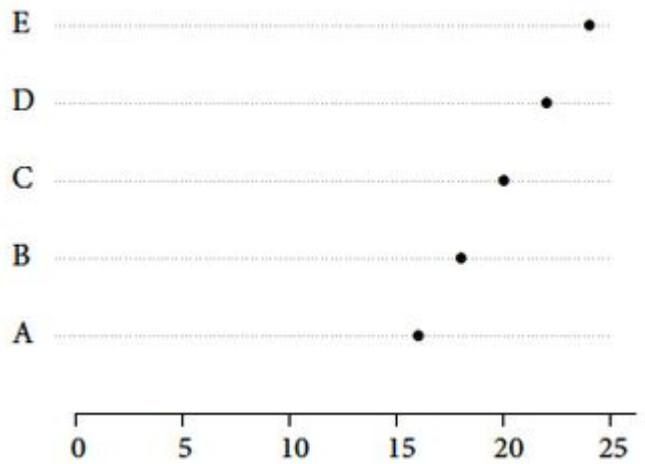
Porównanie



http://www.biecek.pl/Eseje/indexObraz.html

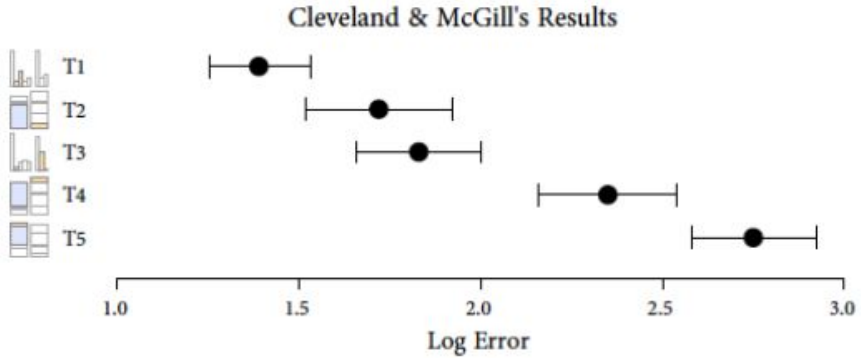
Badania nad percepcją

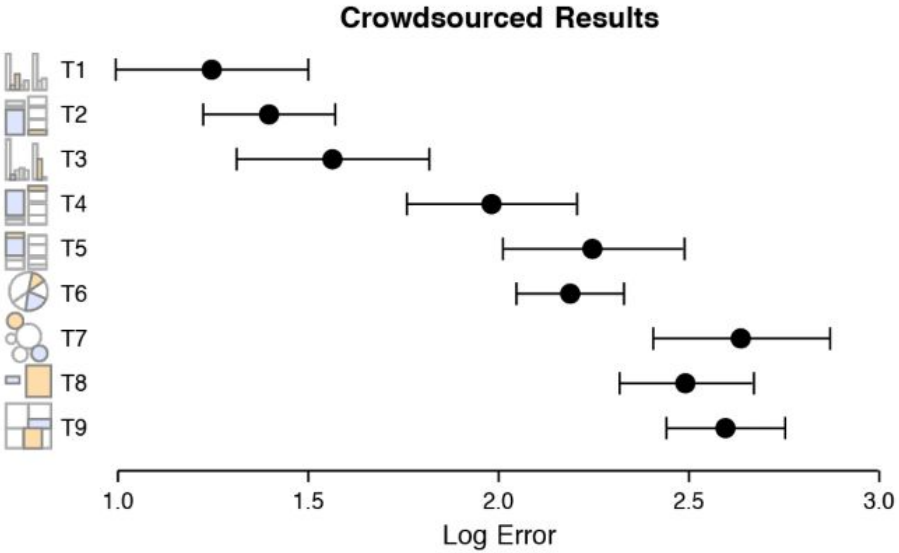




http://www.biecek.pl/Eseje/indexObraz.html

Badania nad percepcją





T1 – wysokości sąsiednich pasków,
T2 – pola prostokątów o wspólnej podstawie,
T3 – wysokości odległych pasków,
T4, T5 – pola prostokątów bez wspólnej podstawy,
T6 – pola wycinków koła,
T7 – pola kół,
T8, T9 – pola niewyrównanych prostokątów.

Do poczytania

Percepcja danych:
http://www.biecek.pl/Eseje/indexDane.html

Percepcja kolorów:
http://www.biecek.pl/Eseje/indexKolory.html

Skale dzielimy są na trzy grupy:

- skale sekwencyjne uporządkowane rozpinają się równomiernie pomiędzy dwoma kolorami,
- skale rozbieżnych kolorów pozwalają na określenie wartości odniesienia (wartości neutralnej) najczęściej za pomocą koloru białego lub żółtego, skrajne wartości odpowiadają ciemnym kolorom z kontrastującą barwą,
- skale jakościowe (czyli nieuporządkowane) pozwalają na możliwie największą rozróżnialność określonej liczby wartości, które jednak nie mają żadnego naturalnego porządku.