METODY STATYSTYCZNE W BIOLOGII

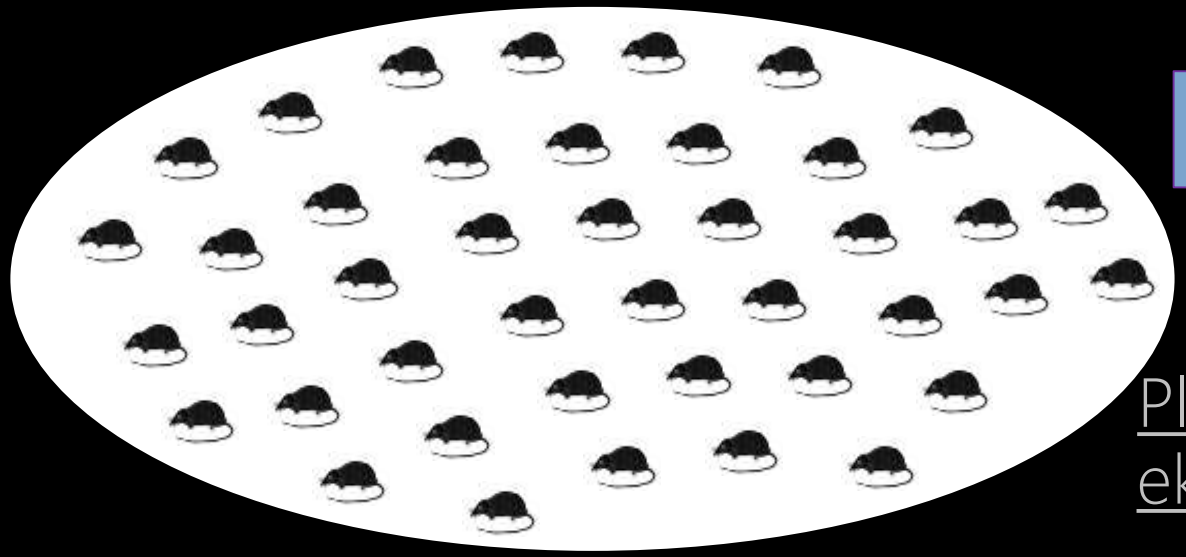
1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne I
6. Najczęściej wykorzystywane testy statystyczne II
7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
10. Korelacja
11. Elementy statystycznego modelowania danych
12. Porównywanie modeli
13. Analiza wariancji
14. Analiza kowariancji
15. Podsumowanie materiału, wspólna analiza przykładów, dyskusja

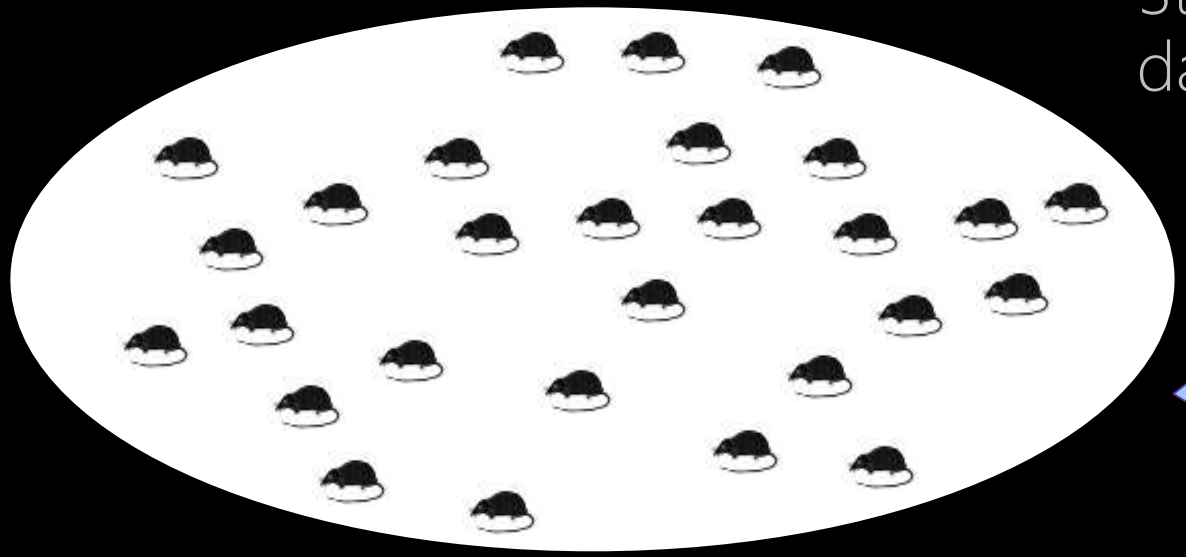
Wstęp

1. Po co planować eksperyment i jakie są etapy planowania eksperymentu ?
2. Moc testu
   - określanie mocy
   - czynniki wpływające na moc
3. Rodzaje prób danych
   - losowe
   - wybór wg określonego kryterium
   - badawczo-kontrolna
   - próby zblokowane
   - cross-over
   - split plot
4. Wykonywanie pomiarów
   - kalibracja
   - niedokładność
   - wpływ obserwatora
   - przykłady cech

Wstęp

Po co planujemy eksperymenty biologiczne ???





Planowanie eksperymentu

Strategia wyboru próby danych z populacji

Wstęp

Po co planujemy eksperymenty biologiczne ???

1. Struktura próby danych musi umożliwiać przetestowanie założonych hipotez
2. Liczebność próby danych musi gwarantować uzyskanie zadowalającej mocy testowania
3. Liczebność próby danych nie może być zbyt duża: kwestie etyczne, koszty, czas

Etapy planowania eksperymentu

Sformułowanie celu badawczego

np. analiza wpływu powietrza atmosferycznego na koncentrację lipidów w tkankach omułków

Sformułowanie hipotez: H_0 i H_1

np. $H_0: k_1 = k_2$ $H_1: k_1 \neq k_2$

Określenie liczebności i struktury próby danych

Zebranie próby danych → Test → Decyzja nt hipotezy

Moc testu

Moc testu

błąd II-go rodzaju

błędy	**prawdziwa** hipoteza	
	H_0	H_1
przyjęta hipoteza H_0		
przyjęta hipoteza H_1		

 : prawdopodobieństwo odrzucenia prawdziwej H_1

1-  : moc testu = prawdopodobieństwo wykrycia rzeczywistej różnicy

Moc testu

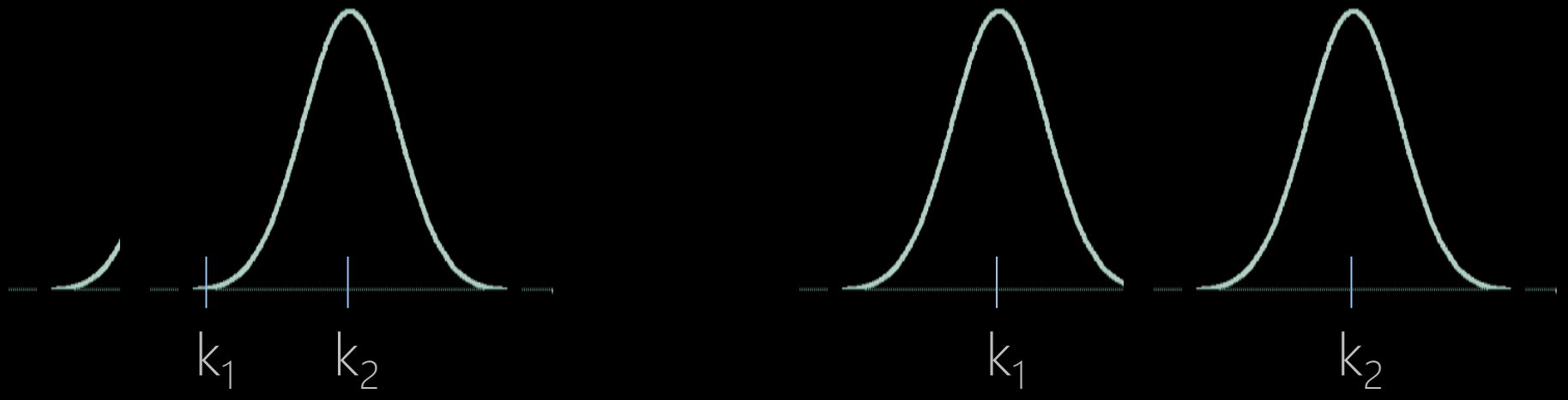
analiza mocy powinna poprzedzać wykonanie każdego eksperymentu

Moc testu - czynniki wpływające na moc

1. Liczba obserwacji w próbie danych
    - łatwiej wykryć efekt w dużych próbach danych
    - więcej informacji
    - mniejszy wpływ błędu próbkowania

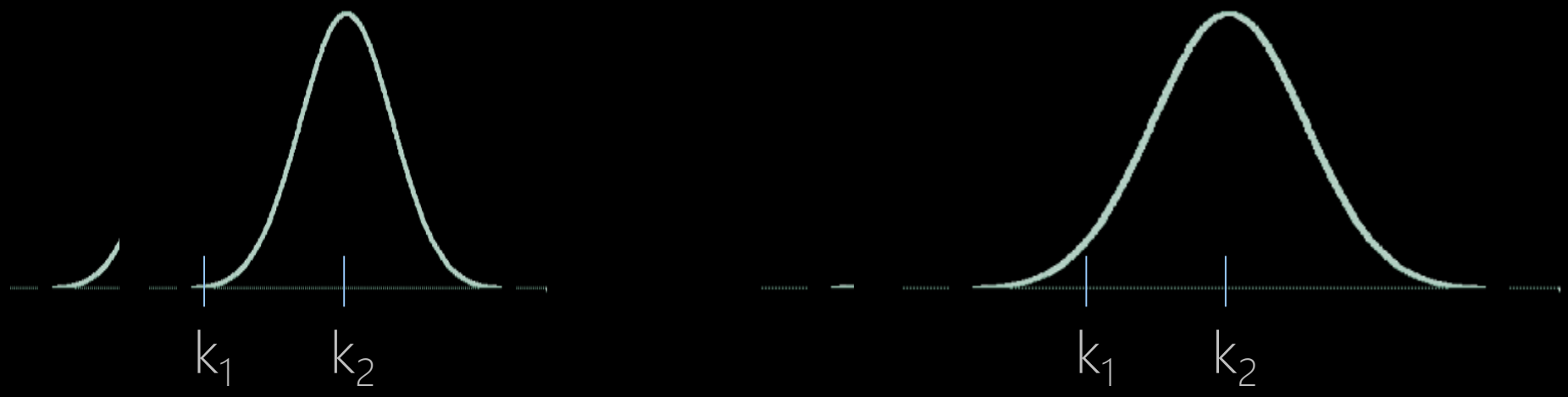
Moc testu - czynniki wpływające na moc

2. Siła wpływu testowanego efektu H_0: $k_1=k_2$ H_1: $k_1\neq k_2$
   - łatwiej wykryć efekt o dużym wpływie



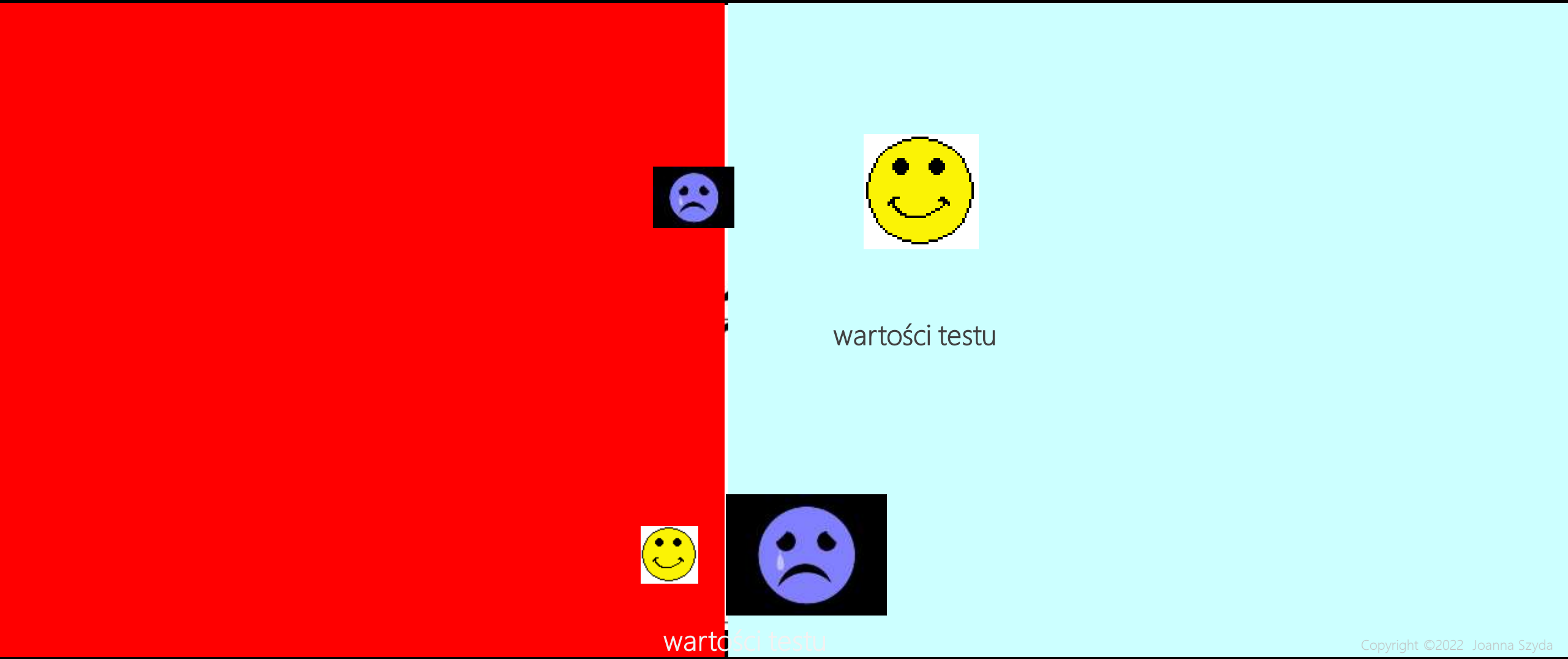
Moc testu - czynniki wpływające na moc

3. Zmienność pomiarów w próbie danych (wariancja próby)
   - Łatwiej wykryć różnice w homogennych próbach danych



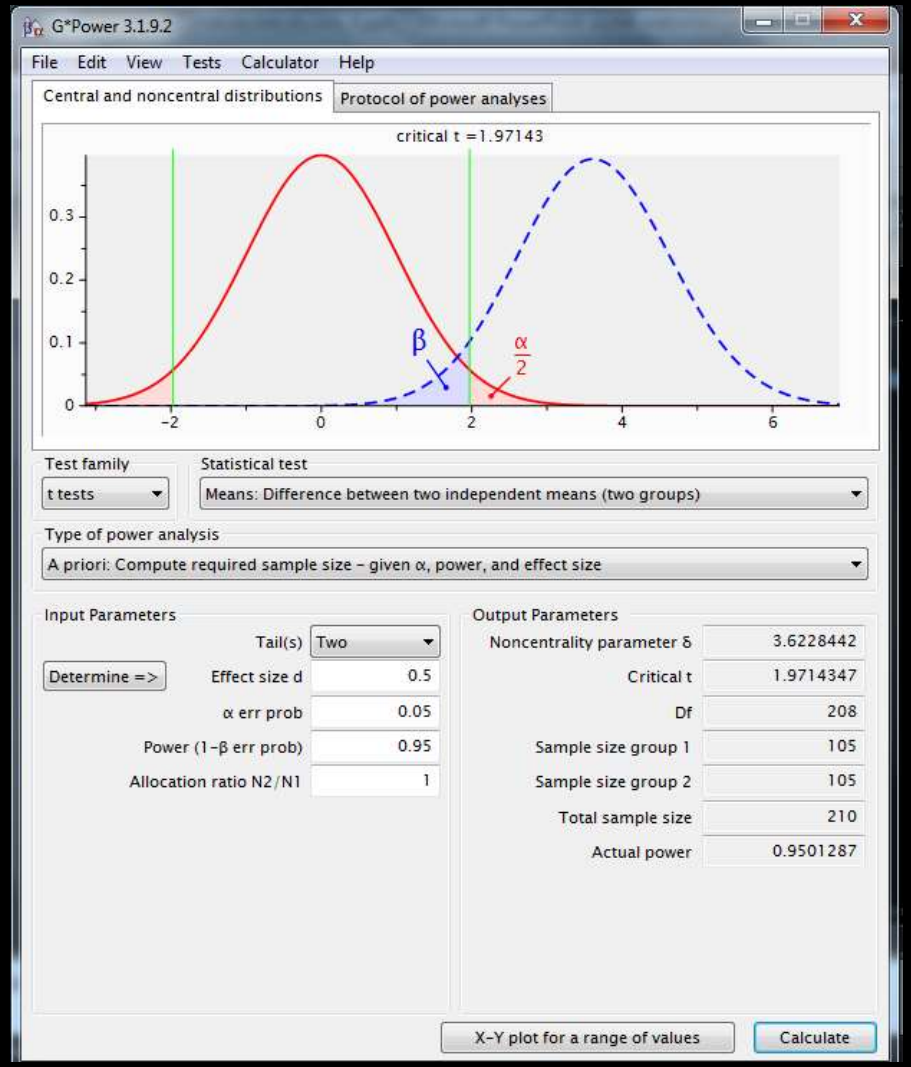
Moc testu - czynniki wpływające na moc

4. Założony poziom błędu I-go rodzaju α
   - Łatwiej wykryć różnice gdy założymy większe α
   - czyli gdy pozwalamy na wyższe prawdopodobieństwo odrzucenia prawdziwej H_0



Moc testu - przykład

1. Obliczenie liczebności próby danych wymaganej dla założonych α , 1-β i wielkości efektu
2. Test t
3. G-power software www.gpower.hhu.de



Moc testu - symulacje

Obliczenie mocy testu w zależności od:

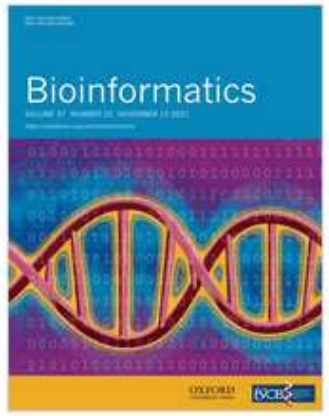
- wpływu testowanego efektu
- zmienności w próbie danych
- N
- α=0.05

```
## define values
mean1=10
mean2=10.5
n=100
nsim=10
alpha_max=0.05

# generate samples and count decisions
H0=0
H1=0
for (i in 1:nsim) {
x=c(rnorm(n,mean=mean1,sd=1))
y=c(rnorm(n,mean=mean2,sd=1))
alpha_t<-t.test(x,y)$p.value
if(alpha_t>alpha_max) {H0=H0+1} else {H1=H1+1}
cat('type 1 error=',alpha_t,"\n")}

# calculate the number of wrong decisions
beta=H0/nsim
power=H1/nsim
cat('H0:',H0,'    beta=',beta)
cat('H1:',H1,'    power=',power)
```

Moc testu - określanie liczebności próby dla danej mocy

Informacje z literatury – symulacje Monte Carlo

- wielokrotne (1 000) tworzenie sztucznych zbiorów danych generowanych wg określonych założeń
- znane wartości prawdziwe
- testowanie
- określenie liczby błędów I- i II-go rodzaju



Volume 37, Issue 22
15 November 2021

powerEQTL: an R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis

Xianjun Dong ✉, Xiaoqi Li, Tzuu-Wang Chang, Clemens R Scherzer, Scott T Weiss, Weiliang Qiu ✉ Author Notes

Bioinformatics, Volume 37, Issue 22, 15 November 2021, Pages 4269–4271, https://doi.org/10.1093/bioinformatics/btab385

Published: 19 May 2021 **Article history** ▾

Moc testu - określanie liczebności próby dla danej mocy

Informacje z literatury – symulacje Monte Carlo

SCIENTIF

OPEN

Power Cal
Combinec
with Appli
Associatic

Received: 12 January 2016
Accepted: 28 April 2016
Published: 18 May 2016

Zhengbang Li[1,*], Wei Zhar

40.5 cmType	ρ	β_1	p	q	HT	oPC (0.8)	SSU	SKAT	mCPC ($k_{0.8}$)	tCPC
40.1 cmI 40.5 cm error	0.20	0	0.20	0.20	0.047	0.047	0.044	0.048	0.044	0.048
	0.50	0	0.20	0.20	0.051	0.052	0.043	0.049	0.051	0.045
	0.80	0	0.20	0.20	0.053	0.049	0.051	0.052	0.050	0.051
	0.95	0	0.20	0.20	0.046	0.053	0.049	0.052	0.047	0.047
	0.20	ln 1.6	0.05	0.30	0.448	0.054	0.140	0.446	0.610	0.404
	0.20	ln 1.4	0.20	0.20	0.754	0.736	0.772	0.762	0.700	0.766
	0.20	ln 1.4	0.30	0.05	0.884	0.922	0.996	0.886	0.856	0.996
	0.50	ln 1.6	0.05	0.30	0.450	0.106	0.288	0.430	0.524	0.496
40.8 cm Power	0.50	ln 1.4	0.20	0.20	0.724	0.756	0.810	0.816	0.702	0.856
	0.50	ln 1.4	0.30	0.05	0.886	0.928	0.996	0.894	0.874	0.988
	0.80	ln 1.6	0.05	0.30	0.450	0.144	0.352	0.440	0.456	0.528
	0.80	ln 1.4	0.20	0.20	0.740	0.778	0.918	0.918	0.754	0.924
	0.80	ln 1.4	0.30	0.05	0.872	0.942	0.984	0.818	0.910	0.964
	0.95	ln 1.6	0.05	0.30	0.484	0.246	0.300	0.368	0.500	0.520
	0.95	ln 1.4	0.20	0.20	0.736	0.940	0.972	0.972	0.924	0.948
	0.95	ln 1.4	0.30	0.05	0.880	0.996	0.962	0.684	0.988	0.920

Table 1. Empirical type I error rates and powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$) and tCPC for Constant correlations.

Moc testu - określanie liczebności próby dla danej mocy

Informacje z literatury – określenie mocy numerycznie

Annals of human genetics

doi: 10.1111/ahg.12147

On Sample Size and Power Calculation for Variant Set-Based Association Tests

Baolin Wu[1]* and James S. Pankow[2]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA*
[2]*Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA*

Summary

Sample size and power calculations are an important part of designing new sequence-based association studies. The recently developed SEQPower and SPS programs adopted computationally intensive Monte Carlo simulations to empirically estimate power for a series of variant set association (VSA) test methods including the sequence kernel association test (SKAT). It is desirable to develop methods that can quickly and accurately compute power without intensive Monte Carlo simulations. We will show that the computed power for SKAT based on the existing analytical approach could be inflated especially for small significance levels, which are often of primary interest for large-scale whole genome and exome sequencing projects. We propose a new χ^2-approximation-based approach to accurately and efficiently compute sample size and power. In addition, we propose and implement a more accurate "exact" method to compute power, which is more efficient than the Monte Carlo approach though generally involves more computations than the χ^2 approximation method. The exact approach could produce very accurate results and be used to verify alternative approximation approaches. We implement the proposed methods in publicly available R programs that can be readily adapted when planning sequencing projects.

Rodzaje prób danych

Rodzaje prób danych - losowa

1. Stosunkowo łatwa do zebrania
   - brak konieczności prowadzenia eksperymentu
   - łatwo uzyskać dużą liczebność próby danych
2. Wyniki można odnieść do całej populacji

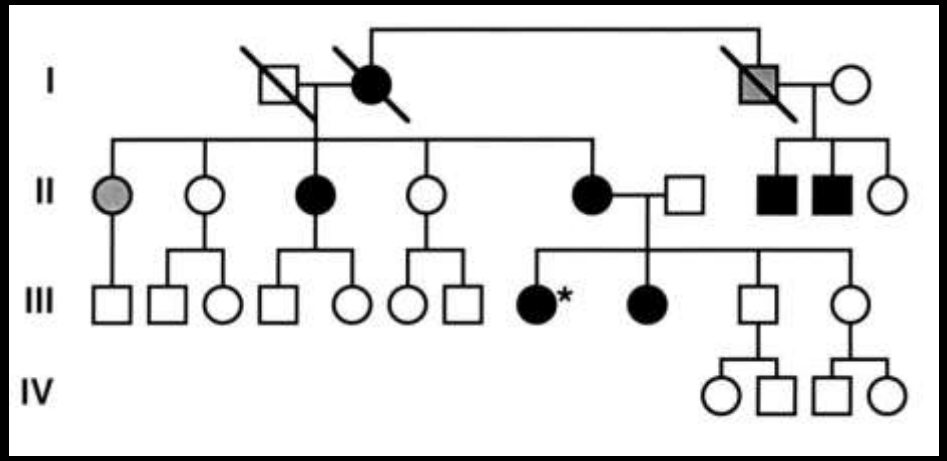
Rodzaje prób danych - wybór wg określonego kryterium

1. Osobniki wybierane nielosowo
2. Muszą spełniać określone kryteria

rodziny jednopokoleniowe

np.
- rodzice + chore dziecko

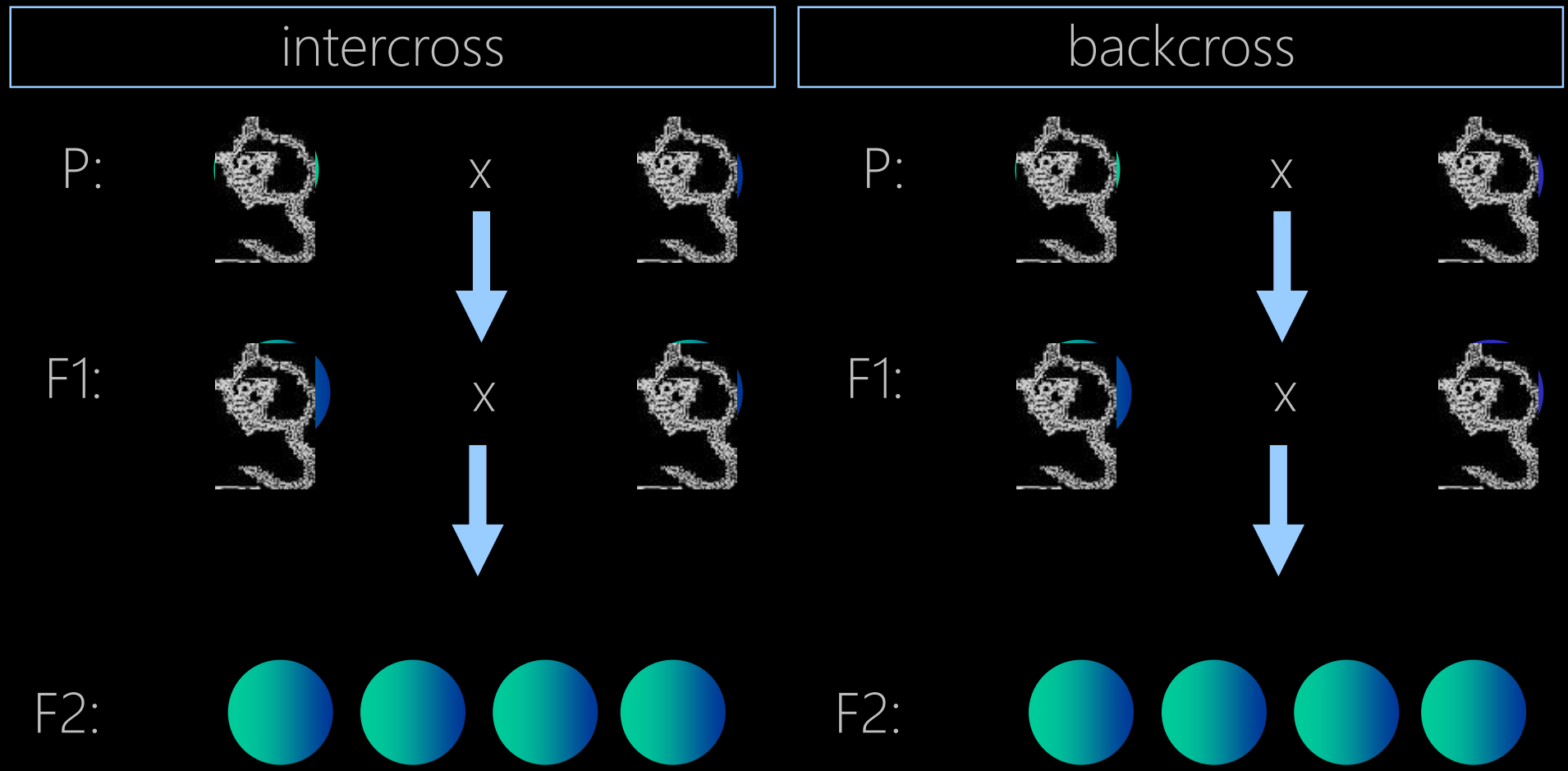
rodziny wielopokoleniowe



np.
- skomplikowana struktura spokrewnienia, osoby wybrane na podstawie wystąpienia choroby

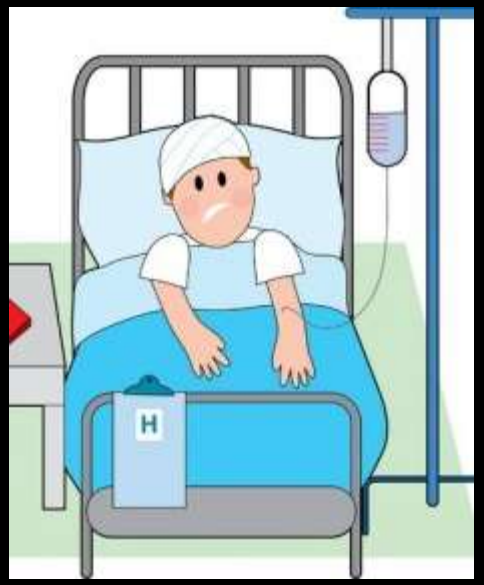
Rodzaje prób danych - wybór wg określonego kryterium

1. Sztuczna struktura spokrewnienia
2. Krzyżowanie linii zinbredowanych (gatunki laboratoryjne)



Rodzaje prób danych - badawczo-kontrolna

1. Osobniki w obrębie każdej kategorii wybierane losowo
2. Grupa badawcza - poddana czynnikowi eksperymentalnemu
3. Grupa kontrolna - referencyjna dla porównania z grupą badawczą

badawcza	kontrolna
	
np. • osoby chore • osoby pobierające leki	np. • osoby zdrowe • osoby pobierające placebo

Rodzaje prób danych - próby zblokowane

badawcza	kontrolna
np. dieta A	np. dieta B

1. Osobniki w obrębie obu kategorii są podobne pod względem jednej lub kilku cech, które mogą mieć potencjalny wpływ na wynik testu - blokowanie
2. Blokowanie zmniejsza wpływ zmienności indywidualnej wewnątrz grupy - większa moc testowania
3. Próby danych często trudne do zebrania
4. Często nie wiemy na podstawie jakich kryteriów przeprowadzić blokowanie

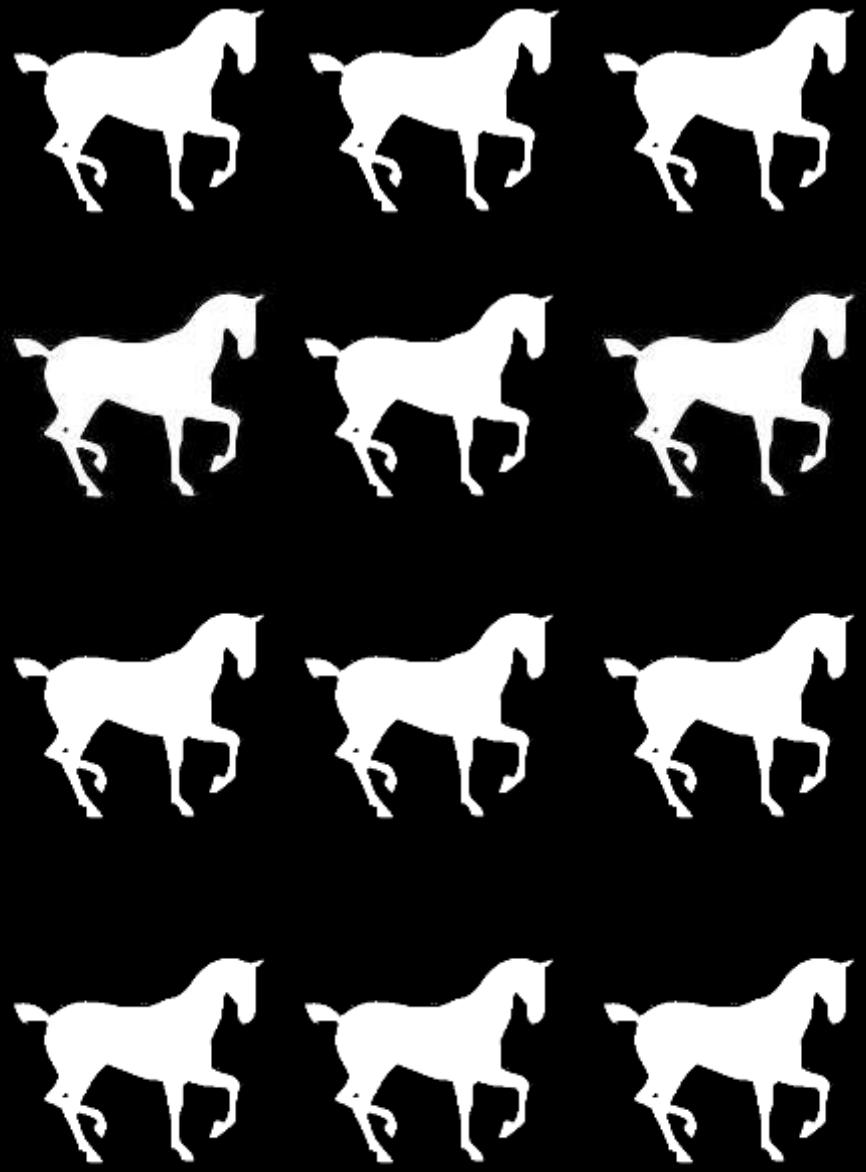
Rodzaje prób danych - cross over

dieta A — Etap 1 — dieta B

dieta A — Etap 2 — dieta B

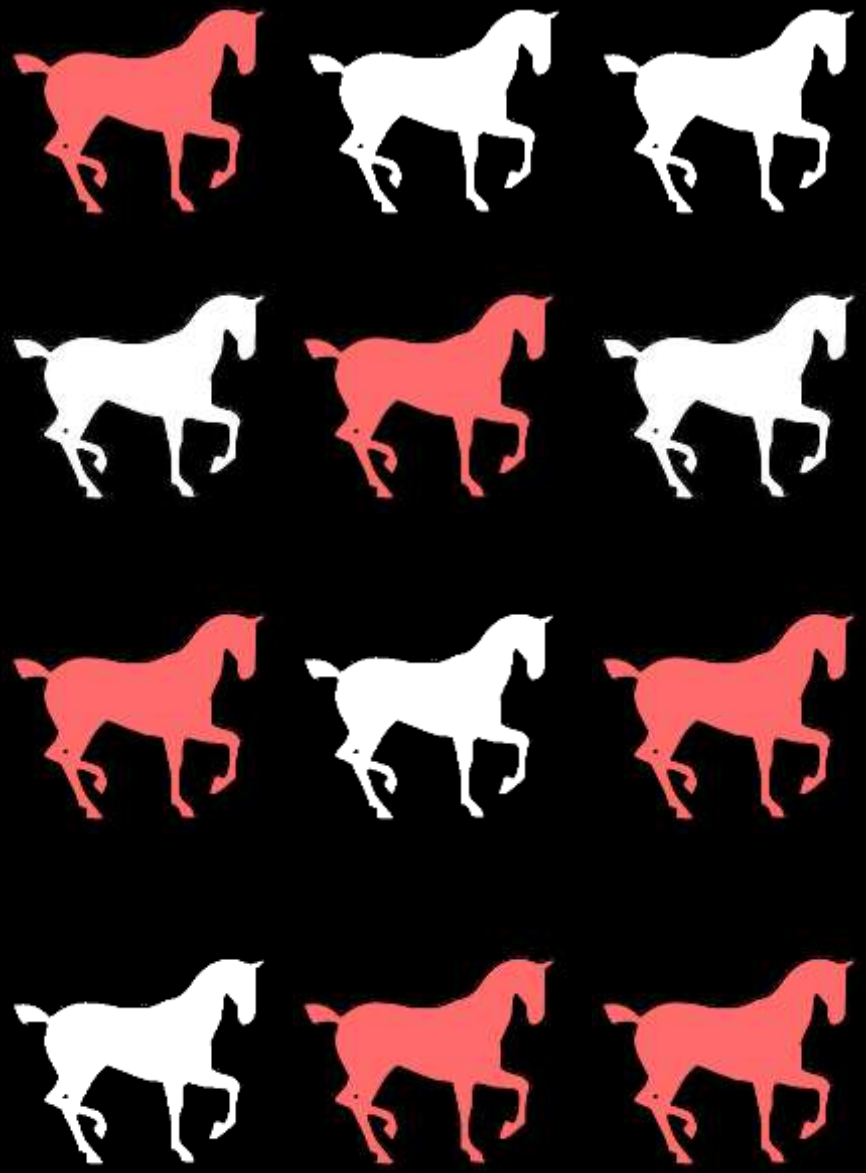
1. Te same osobniki występują w obu grupach
2. Eliminacja zmienności wewnątrzosobniczej
3. Możliwe tylko dla niektórych rodzajów badań (np. pomiar cechy tani, przyżyciowy, proste warunki utrzymania osobników)

Rodzaje prób danych - split plot



2 czynniki eksperymentalne

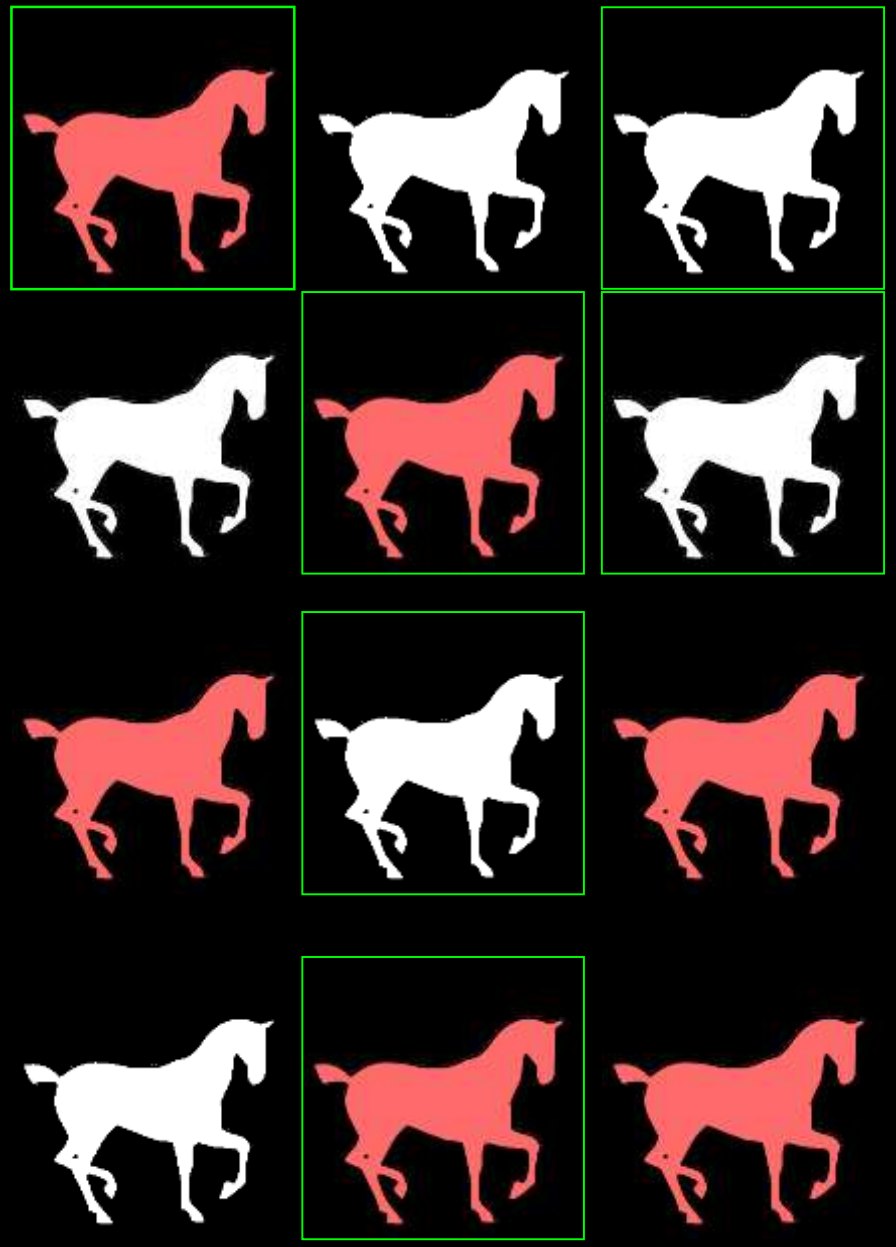
Rodzaje prób danych - split plot



dieta A, dieta B

2 czynniki eksperymentalne

Rodzaje prób danych - split plot



dieta A, dieta B

trening A, trening B

2 czynniki eksperymentalne

Wykonywanie pomiarów

Pomiary - niedokładność

1. Kalibrowanie urządzeń pomiarowych
   - ustawianie / kontrola urządzeń pomiarowych na podstawie analizy próbek o znanych wartościach
   - wielokrotne kalibrowanie w czasie wykonywania pomiarów

2. Niedokładność pomiarów
   - nie jesteśmy w stanie dokonać pomiaru z nieskończoną dokładnością
   - precyzja powinna być jednakowa dla wszystkich próbek w danym badaniu

Pomiary - wpływ obserwatora

wewnątrz obserwatora

np.
- zmęczenie
- zmiana oceny subiektywnej

pomiędzy obserwatorami

np.
- różnice w subiektywnych ocenach

Zasady prowadzenia obserwacji:
- Nie wykonywać zbyt wielu obserwacji na raz
- Nie stosować uproszczonych skrótów
- Tworzyć zapasowe kopie danych
- Tworzyć protokoły przebiegu eksperymentu
- Wykorzystywać elektroniczne formularze bazy danych

Pomiary - przykłady cech

Łatwe do skwantyfikowania	trudne do skwantyfikowania
np. wzrost	np. obserwacje behawioralne

1. Po co planować eksperyment i jakie są etapy planowania eksperymentu ?
2. Moc testu
   - określanie mocy
   - czynniki wpływające na moc
3. Rodzaje prób danych
   - losowe
   - wybór wg określonego kryterium
   - badawczo-kontrolna
   - próby zblokowane
   - cross-over
   - split plot
4. Wykonywanie pomiarów
   - kalibracja
   - niedokładność
   - wpływ obserwatora
   - przykłady cech