BIOSTATYSTYKA

Wykład 1: Wprowadzenie

dr inż. Bartosz Czech

Katedra Genetyki · Uniwersytet Przyrodniczy we Wrocławiu

bartosz.czech@upwr.edu.pl

Plan i cel

Zakres materiału

1. Wprowadzenie do biostatystyki
2. Populacje i próby danych
3. Testowanie hipotez i estymacja
4. Testy statystyczne I: testy t
5. Testy statystyczne II: test t sparowany, dane kategoryczne, testy nieparametryczne
6. Regresja liniowa
7. Regresja nieliniowa
8. Jakość dopasowania regresji
9. Korelacja
10. Analiza wariancji (ANOVA)

Zasady zaliczenia

Zaliczenie

Kurs kończy się zaliczeniem pisemnym — sprawdza rozumienie pojęć oraz umiejętność doboru i interpretacji właściwego testu, nie odtwarzanie wzorów z pamięci.

Nieobecności i konsultacje

Obecność

- dopuszczalne nieobecności: **do 20%** zajęć — przy 10 wykładach **najwyżej 2** (czyli jedno spotkanie)
- przekroczenie limitu rozpatrywane zgodnie z regulaminem studiów
- usprawiedliwienie — niezwłocznie po ustaniu przyczyny

Konsultacje: poniedziałek 14:00–15:30, pokój 306 w Katedrze Genetyki
(proszę o wcześniejszy kontakt mailowy: `bartosz.czech@upwr.edu.pl`)

Literatura

Podręczniki

- Dobek, Szwaczkowski — *Statystyka Matematyczna dla Biologów* (podstawowy)
- Whitlock, Schluter — *The Analysis of Biological Data*
- Beckerman, Childs, Petchey — *Getting Started with R*

Darmowe, dostępne online

- Childs i in. — *Introductory Biostatistics with R*
- Holmes, Huber — *Modern Statistics for Modern Biology*
- Navarro — *Learning Statistics with R*
- *OpenIntro Statistics*
- Wickham i in. — *R for Data Science*

Plan dzisiejszego wykładu

1. Proces naukowy i hipotezy — od obserwacji do wniosku
2. Typy zmiennych i skale pomiarowe
3. Jakość pomiaru: dokładność i precyzja
4. Pierwszy kontakt z R

Po co statystyka w biologii

Czym jest biostatystyka?

Biostatystyka (*biostatistics*)

Biostatystyka to dział statystyki stosowany do danych z badań biologicznych i medycznych — zestaw metod, którymi planujemy doświadczenie, analizujemy wyniki i wyciągamy z nich wnioski.

- Towarzyszy **całemu badaniu** — od zaplanowania doświadczenia, przez analizę danych, po wniosek
- Potrzebna, bo biologia jest z natury **zmienna**: dwie myszy z miotu różnią się masą, dwie kolonie rosną w różnym tempie, dwa powtórzenia qPCR dają inny wynik
- Pozwala oddzielić **prawdziwy sygnał** — efekt leku, mutacji czy nawozu — od **szumu**, czyli przypadkowej zmienności

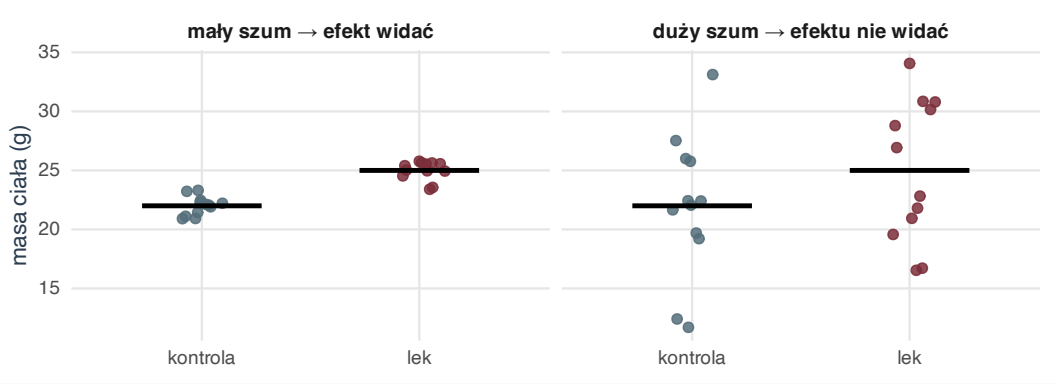
Pytanie, na które odpowiada ten przedmiot

Skąd wiemy, że różnica między grupami jest rzeczywista, a nie przypadkowa?

Statystyka wpisana w biologię

- Grzegorz Mendel odkrył prawa dziedziczenia, bo *policzył* tysiące roślin grochu i porównał proporcje z oczekiwaniami
- Ronald Fisher stworzył podstawy planowania doświadczeń, analizując wyniki pól doświadczalnych
- Dziś genomika, ekologia i farmakologia opierają swoje wnioski na analizie danych

Sygnał kontra szum



Ta sama różnica średnich (3 g) i liczebność, różny rozrzut. Wniosek zależy od różnicy *względem* rozrzutu.

Dwa zadania statystyki

Statystyka opisowa (*descriptive statistics*)

Podsumowuje to, co faktycznie zmierzyliśmy: średnie, wykresy, tabele. Nie wychodzi poza próbę.

„Jaka była średnia masa w mojej grupie 20 myszy?”

Statystyka wnioskująca (*inferential statistics*)

Przenosi wniosek z próby na populację (wszystkie obiekty, o których chcemy wnioskować — szczegóły na W2), zawsze z określoną niepewnością.

„Czy podanie leku zmienia *średnią* masę myszy tego szczepu?”

Po co to biologowi?

- **Praca dyplomowa:** „Czy mój związek / nawóz / dieta rzeczywiście działa?” — bez testu to tylko wrażenie
- **Czytanie publikacji:** zapis „p = 0,03, n = 12, $\log_2$FC = 1,8” ($\log_2$ krotności zmiany ekspresji, tu ok. 3,5-krotny wzrost) trzeba rozumieć, by ocenić wiarygodność wniosków
- **Planowanie eksperymentu:** „Ile zwierząt, powtórzeń, prób potrzeba, żeby w ogóle coś wykryć?”
- **Krytyczna ocena cudzej pracy:** „Czy autorzy zastosowali właściwy test, czy naciągnęli wynik?”

Filozofia tego kursu

Zasada przewodnia

Nie uczymy wzorów na pamięć ani liczenia ręcznego. Uczymy się **myśleć o danych**: dobierać właściwy test, rozumieć jego wynik i unikać typowych pułapek.

- Obliczenia wykonuje się w R — na ćwiczeniach
- Na wykładzie budujemy **intuicję**: co liczymy, kiedy i dlaczego oraz jak odczytać wynik z komputera
- Miara sukcesu: **umiejętność odczytania** wyniku testu lub wykresu i oceny, czy wniosek jest uczciwy

1. Proces naukowy

Cykl badania naukowego



nowy wniosek rodzi nowe pytanie

Hipoteza badawcza i test statystyczny

Hipoteza badawcza (*research hypothesis*)

Sprawdzalne przypuszczenie o zależności między cechami lub o różnicy między grupami, sformułowane przed zebraniem danych.

Test statystyczny (*statistical test*)

Procedura, która na podstawie danych z próby ocenia, czy są one zgodne z założeniem braku efektu (hipotezą zerową); wynik to decyzja: odrzucić to założenie albo nie.

> Hipotezę badawczą przekłada się na parę hipotez statystycznych (zerowa i alternatywna), a test sprawdza, czy dane pozwalają odrzucić hipotezę zerową na korzyść alternatywnej. Szczegóły na wykładzie 3.

Cechy dobrej hipotezy

- **Testowalna** — da się ją sprawdzić danymi
- **Falsyfikowalna** — da się wyobrazić wynik, który by ją obalił
- **Konkretna** — mówi o mierzalnej wielkości, nie o wrażeniu

Słaba

„Ten gen jest ważny dla wzrostu bakterii.” „Ważność” nie jest mierzalną wielkością — nie wiadomo, co z czym porównać.

Dobra

„Odsetek kolonii opornych na ampicylinę jest wyższy u szczepu z plazmidem niosącym gen *bla* (β-laktamazy) niż bez niego.”

Hipoteza PRZED zebraniem danych

HARKing (Hypothesizing After the Results are Known)

Dopasowanie hipotezy do już zobaczonych danych. Przegląda się dziesiątki genów, jeden okazuje się istotny przez przypadek i dopiero wtedy powstaje hipoteza, że to właśnie on był przedmiotem badania.

- Przy wielu porównaniach bardzo prawdopodobne jest, że **któreś** okaże się „istotne” przez przypadek
- Hipoteza dobrana po fakcie nie została przetestowana — wymaga potwierdzenia na nowych danych
- Dlatego rzetelne badania rejestruje się z góry (*preregistration*)

Dwie hipotezy: zerowa i alternatywna

H_0 — zerowa

- „Nic się nie dzieje”, brak efektu, status quo
- To ją test bierze pod lupę
- Zawsze zawiera równość: *brak różnicy*

H_1 — alternatywna

- „Coś się dzieje”, jest efekt
- To zwykle nasza hipoteza badawcza
- Zawiera różnicę / zależność

Przykład: nokaut genu a masa ciała myszy

H_0: średnia masa myszy typu dzikiego = średnia masa myszy z nokautem.

H_1: średnie masy się różnią.

Analogia sądowa

Sąd

- Domniemanie niewinności = punkt wyjścia
- Oskarżyciel musi **udowodnić** winę
- Brak dowodów $\rightarrow$ „niewinny”
- „niewinny” $\neq$ „na pewno nie zrobił”

Test statystyczny

- H_0 (brak efektu) = punkt wyjścia
- Dane muszą **wykazać** efekt
- Słabe dowody $\rightarrow$ „nie odrzucamy H_0”
- to nie znaczy „efektu nie ma”

Nie odrzucić $H_0 \neq$ udowodnić H_0. Brak dowodu efektu to nie dowód braku efektu.

Zasady planowania doświadczenia

Zasady Fishera

- **Replikacja** (*replication*) — wiele niezależnych osobników w każdej grupie (powtórzenia biologiczne); bez niej nie da się oddzielić efektu od zmienności między osobnikami
- **Randomizacja** (*randomization*) — losowy przydział osobników do grup; wyrównuje między grupami czynniki uboczne (np. wiek, masę), także nieznane
- **Kontrola lokalna** (*local control*) — łączenie podobnych osobników w bloki (np. miot, szklarnia) i porównywanie grup w obrębie bloku

Dobre praktyki

- **Grupa kontrolna** (*control group*) — punkt odniesienia bez badanego czynnika (np. sam nośnik zamiast leku)
- **Ślepa próba** (*blinding*) — oceniający wynik nie wie, która grupa jest która

Bez randomizacji badacz może — nawet nieświadomie — umieścić w grupie leczonej zwierzęta młodsze lub zdrowsze. Bez grupy kontrolnej nie odróżni się efektu leku od naturalnego przebiegu choroby.

2. Zmienne i skale pomiarowe

Czym jest zmienna?

Zmienna (*variable*)

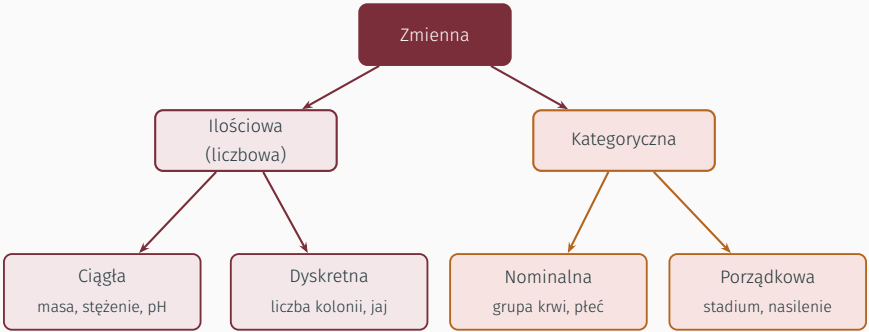
Cecha mierzona lub obserwowana, której wartości różnią się między jednostkami badania (osobnikami, próbkami, poletkami).

Typ zmiennej wyznacza dozwolone metody — najważniejsza decyzja w całej analizie.

Najczęstszy błąd początkującego

Dobór testu do „co chcę udowodnić” zamiast do „jaki typ danych mam”.

Drzewo typów zmiennych



Przykłady z laboratorium

Typ	Cecha biologiczna	Co można policzyć
Ciągła	Masa guza, stężenie białka, pH, długość korzenia	Średnia, SD, test t
Dyskretna	Liczba kolonii, liczba jaj, liczba mutacji	Średnia; testy dla zliczeń
Nominalna	Genotyp (WT/mutant), płeć, przeżył/zginął	Proporcje, chi-kwadrat
Porządkowa	Stadium (I–IV), skala nasilenia objawów	Mediana, testy rangowe

Porządkowa ma kolejność, ale odstępy nie są określone: różnica między stadium I a II nie musi być „taka sama” jak między III a IV.

Skale pomiarowe (Stevens)

Skala	Co ma sens	Przykład
Nominalna	tylko $=$, $\neq$ (rozróżnienie kategorii)	grupa krwi, gatunek
Porządkowa	dodatkowo $<$, $>$ (kolejność)	stadium nowotworu
Interwałowa	dodatkowo różnice	temperatura w °C
Ilorazowa	dodatkowo ilorazy („dwa razy więcej”)	masa, stężenie, liczba kolonii

- Każda kolejna skala dopuszcza operacje poprzedniej i jedną nową
- Podział na ciągłe i dyskretne jest **niezależny** od skali: liczba kolonii jest dyskretna i ilorazowa
- **Zmienna dychotomiczna** (*dichotomous*, binarna) — nominalna o dwóch kategoriach: przeżył/zginął, WT/mutant

Skala interwałowa a ilorazowa

Interwałowa (przedziałowa)

Równe odstępy, ale **zero jest umowne**. Nie wolno mówić „dwa razy więcej”.

Temperatura w °C: 20° nie jest „dwa razy cieplejsze” niż 10°.

Ilorazowa

Zero jest **bezwzględne** — oznacza całkowity brak wielkości (0 mg = brak masy). Dlatego stosunki mają sens.

Masa, stężenie, długość: 10 mg to naprawdę dwa razy więcej niż 5 mg.

Większość pomiarów biologicznych jest ilorazowa.

Pułapka: liczba to nie zawsze zmienna ilościowa

To, że coś jest zapisane cyfrą, nie znaczy, że wolno liczyć z tego średnią.

- Numer klatki, numer próbki, kod pacjenta — to etykiety, nie wielkości; średnia nie ma sensu
- Skala bólu 1–10 — porządkowa; różnica „3→4” nie musi odpowiadać różnicy „7→8”
- Genotyp zapisany jako **liczba kopii allelu** (0 / 1 / 2) — zliczenie, czyli **ilościowa dyskretna**; tak traktuje się go w modelu addytywnym. Ten sam genotyp zapisany jako AA / Aa / aa — kategorie

Test kontrolny

„Czy średnia z tych liczb ma sens biologiczny?” Jeśli nie — to nie jest zmienna ilościowa.

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	
Liczba młodych w miocie	
Grupa krwi (A, B, AB, 0)	
Stopień złośliwości (G1–G4)	
Genotyp 0 / 1 / 2 (liczba kopii allelu)	

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	ilościowa ciągła
Liczba młodych w miocie	
Grupa krwi (A, B, AB, 0)	
Stopień złośliwości (G1–G4)	
Genotyp 0 / 1 / 2 (liczba kopii allelu)	

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	ilościowa ciągła
Liczba młodych w miocie	ilościowa dyskretna
Grupa krwi (A, B, AB, 0)	
Stopień złośliwości (G1–G4)	
Genotyp 0 / 1 / 2 (liczba kopii allelu)	

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	ilościowa ciągła
Liczba młodych w miocie	ilościowa dyskretna
Grupa krwi (A, B, AB, 0)	nominalna
Stopień złośliwości (G1–G4)	
Genotyp 0 / 1 / 2 (liczba kopii allelu)	

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	ilościowa ciągła
Liczba młodych w miocie	ilościowa dyskretna
Grupa krwi (A, B, AB, 0)	nominalna
Stopień złośliwości (G1–G4)	porządkowa
Genotyp 0 / 1 / 2 (liczba kopii allelu)	

Ćwiczenie: rozpoznawanie typu zmiennej

Cecha	Typ
Stężenie glukozy we krwi [mg/dl]	ilościowa ciągła
Liczba młodych w miocie	ilościowa dyskretna
Grupa krwi (A, B, AB, 0)	nominalna
Stopień złośliwości (G1–G4)	porządkowa
Genotyp 0 / 1 / 2 (liczba kopii allelu)	ilościowa dyskretna (model addytywny)

Zmienna zależna i niezależna

Niezależna (X) — objaśniająca

To, co zmieniamy lub według czego dzielimy grupy: dawka, genotyp, traktowanie, temperatura.

Zależna (Y) — objaśniana

Odpowiedź, którą mierzymy: masa, ekspresja, przeżycie, plon.

Pytanie badawcze zwykle brzmi: „Czy Y zależy od X?”

o wpływie przyczynowym można mówić dopiero w eksperymencie z losowym przydziałem

Dawka leku (X) $\longrightarrow$ wielkość guza (Y). Genotyp (X) $\longrightarrow$ tempo wzrostu (Y).

Zmienna zakłócająca

Zmienna zakłócająca (*confounder*)

Czynnik, który wpływa na Y i jest powiązany z X, ale nie jest skutkiem X; jeśli go pominąć, zależność X–Y jest zniekształcona.



Jeśli myszy jednego genotypu dobrano starsze, „efekt genotypu” na masę może być w rzeczywistości efektem wieku.

3. Jakość pomiaru

Dokładność a precyzja

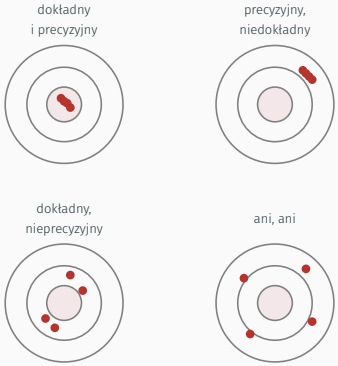
Dokładność (*accuracy*)

Zgodność średniej z pomiarów z wartością prawdziwą (ściślej: poprawność, *trueness*).

Precyzja (*precision*)

Zgodność wyników kolejnych pomiarów między sobą.

To dwie **różne** rzeczy. Pomiar może być precyzyjny, a mimo to systematycznie zły.



Przykłady z pracowni

Sytuacja	Problem	Objaw
Pipeta źle skalibrowana (95 zamiast 100 µl)	niedokładność	wyniki systematycznie zaniżone
Pipeta zużyta, tłok „skacze”	brak precyzji	duży rozrzut powtórzeń
Spektrofotometr bez próby zerowej	niedokładność	stałe przesunięcie odczytów
Zmienne ustawienie oka przy odczycie menisku	brak precyzji	losowe wahania odczytu

Błąd systematyczny a losowy

Systematyczny (bias)

- Przesuwa wyniki zawsze w tę samą stronę
- Psuje **dokładność**
- **Nie znika** przy uśrednianiu powtórzeń

Losowy (random)

- Rozrzuca wyniki w obie strony
- Psuje **precyzję**
- **Uśrednia się** — więcej powtórzeń go redukuje

Reguła praktyczna

Powtarzanie pomiaru redukuje błąd losowy, ale NIE naprawia błędu systematycznego. Ten trzeba usunąć u źródła — kalibracja, próba zerowa (*blank*), kontrole. Sto pomiarów źle skalibrowaną pipetą to sto zaniżonych wyników.

Jednostki i cyfry znaczące

- Każda liczba w biologii ma **jednostkę** — bez niej „stężenie 5” nic nie znaczy
- Liczba **cyfr znaczących** mówi, jak precyzyjny był pomiar — zapis „12,7384 g” obiecuje rozdzielczość (najmniejszą rozróżnialną różnicę) do dziesiątych części miligrama, której zwykła waga laboratoryjna nie ma
- Nie raportować więcej cyfr, niż naprawdę daje metoda pomiaru

4. Pierwszy kontakt z R

Dane w R: ramka danych

Obserwacje = **wiersze**, zmienne = **kolumny**; taka tabela to **ramka danych** (*data frame*).

```r
myszy <- data.frame(
  genotyp = c("WT","WT","WT","KO","KO","KO"),
  masa_g  = c(22.1, 23.4, 21.8, 25.6, 26.9, 24.8)
)
str(myszy)
tapply(myszy$masa_g, myszy$genotyp, mean)
tapply(myszy$masa_g, myszy$genotyp, sd)
```

```
'data.frame':   6 obs. of  2 variables:
 $ genotyp: chr  "WT" "WT" "WT" "KO" ...
 $ masa_g : num  22.1 23.4 21.8 25.6 26.9 24.8
      KO       WT
25.76667 22.43333
       KO        WT
1.0598742 0.8504901
```

Jak to czytać

`str()` podaje typ kolumny: `chr` — tekst (zmienna nominalna), `num` — liczby (zmienna ilościowa). `tapply()` liczy średnią i odchylenie standardowe (SD — miara rozrzutu wokół średniej, W2) osobno w grupach.

Sygnał czy szum?

Średnie grup różnią się o ponad 3 g, a rozrzut *wewnątrz* grup wynosi ok. 1 g. Czy to efekt nokautu, czy przypadek, rozstrzyga dopiero test statystyczny (W3–W4).

Wykres jednym poleceniem

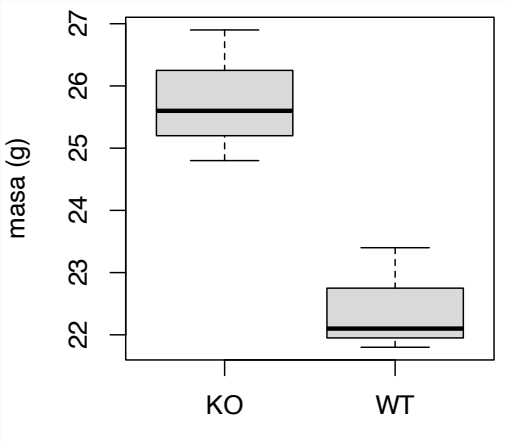
```
boxplot(masa_g ~ genotyp, data = myszy,
        ylab = "masa (g)", col = "grey85")
```

Zapis y ~ x

„masa w zależności od genotypu”; `data = myszy` wskazuje, gdzie szukać kolumn. Ten sam zapis wróci przy testach i regresji.

Jak czytać pudełko

Linia w środku to mediana (wartość środkowa), brzegi pudełka to kwartyle — między nimi leży środkowa połowa danych. Wąsy sięgają najdalszych obserwacji leżących nie dalej niż 1,5 długości pudełka od jego brzegów; punkty dalej to kandydaci na wartości odstające (W2).



Podsumowanie

Do zapamiętania

- Biologia jest zmienna — zadaniem statystyki jest oddzielić **sygnał** od **szumu**
- **Hipotezę formułujemy przed danymi**, nie po zobaczeniu wyniku; plan doświadczenia opiera się na replikacji, randomizacji i kontroli lokalnej
- **Typ zmiennej** i **skala pomiarowa** (nominalna, porządkowa, interwałowa, ilorazowa) rządzą wyborem testu; liczba $\neq$ zmienna ilościowa
- Dokładność $\neq$ precyzja — **błąd systematyczny się nie uśrednia**, trzeba go usunąć u źródła

Na następny wykład

Wykład 2: Populacje i próby danych

Nauczymy się **opisywać zmienność liczbami**: średnia i mediana, odchylenie standardowe, rozkład normalny oraz kluczowe rozróżnienie SD kontra SE.

Dziękuję za uwagę

dr inż. Bartosz Czech · bartosz.czech@upwr.edu.pl