BIOSTATYSTYKA

Wykład 4: Testy statystyczne I (testy t)

dr inż. Bartosz Czech

Katedra Genetyki · Uniwersytet Przyrodniczy we Wrocławiu

bartosz.czech@upwr.edu.pl

Plan wykładu

1. Idea testu t: sygnał i szum
2. Test t jednej próby
3. Dwie niezależne próby: test t-Studenta i test Welcha
4. Założenia testu i wynik w R

Cel wykładu

Zrozumieć, od czego zależy wynik testu porównującego średnie cechy ciągłej, i nauczyć się czytać ten wynik.

1. Idea testu t

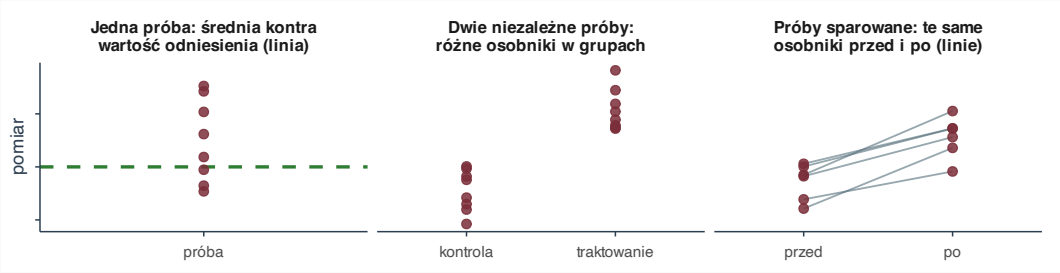
Co porównuje test t

Odpowiada na pytanie

Czy średnie różnią się bardziej, niż wynikałoby z samego przypadku?

Warunki stosowania

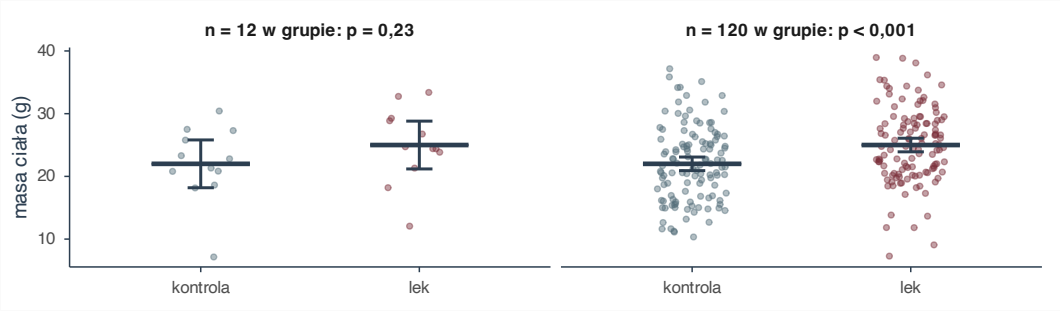
Zmienna ciągła (masa, ekspresja, stężenie), porównywane są jedna lub dwie średnie.



Dla dwóch niezależnych prób są dwie wersje: test t-Studenta i test Welcha; w teście sparowanym porównuje się różnice w parach.

Sygnał kontra szum — wpływ liczebności

Rozwinięcie wykładu 1: tam różnica 3 g ginęła w rozrzucie (SD $= 6$ g, po 12 myszy w grupie). O wiarygodności różnicy decyduje jej błąd standardowy, który rośnie z rozrzutem danych i maleje z liczebnością grup.



Ta sama różnica średnich (3 g) i to samo SD (6 g) co w prawym panelu z wykładu 1; odcinki: 95% przedziały ufności średnich. Zmienia się tylko n.

Statystyka t i jej składniki

Statystyka t (*t statistic*)

Różnica średnich podzielona przez jej błąd standardowy, czyli sygnał podzielony przez szum:

$$t = \frac{\text{różnica średnich}}{\text{błąd standardowy różnicy}}$$

Gdy rośnie...	wartość $\lvert t \rvert$...	bo...
różnica średnich	rośnie	większy sygnał
odchylenie standardowe w grupach	maleje	większy szum
liczebność grup n	rośnie	większa próba daje mniejszy błąd standardowy

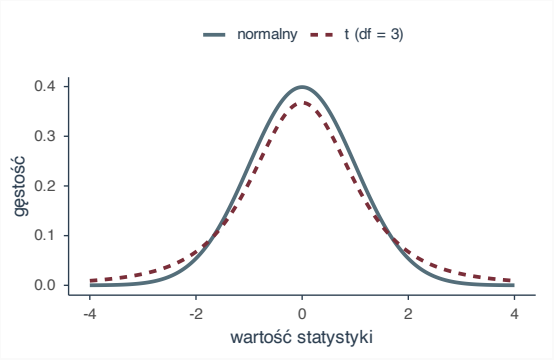
Każdy test t ma tę samą budowę; warianty różnią się tylko sposobem obliczenia szumu. Rachunek wykonuje R.

Rozkład t-Studenta

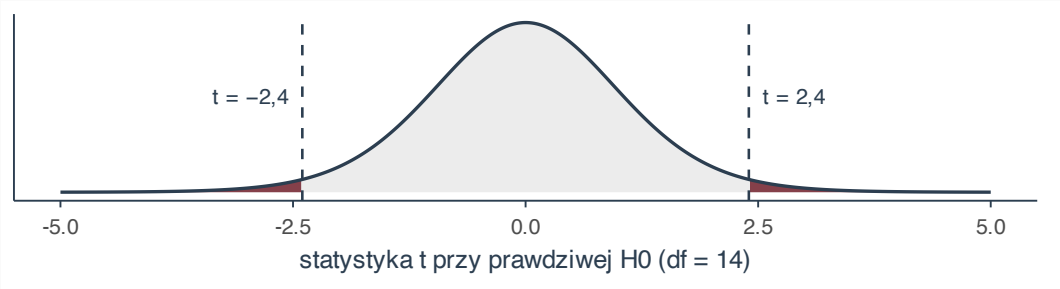
Rozkład t-Studenta (*Student's t distribution*)

Rozkład statystyki t przy prawdziwej H_0, gdy odchylenie standardowe populacji zastępuje się odchyleniem z próby.

- Symetryczny wokół zera, grubsze ogony niż rozkład normalny
- Kształt zależy od wielkości próby: im mniejsza próba, tym grubsze ogony; w wynikach wielkość tę opisuje liczba stopni swobody (df)
- Przy dużej próbie zbliża się do rozkładu normalnego



Wartość p w teście t



- Wartość p = pole w obu ogonach rozkładu t poza obserwowaną wartością $\pm t$: jak często przy braku różnicy wynik byłby co najmniej tak skrajny
- Na wykresie $t = 2{,}4$, więc $p = 0{,}031$; większe $|t|$ daje mniejsze pole i mniejsze p
- Przy większych próbach $p < 0{,}05$ odpowiada w przybliżeniu $|t| > 2$; dokładne pole oblicza R

2. Test t jednej próby

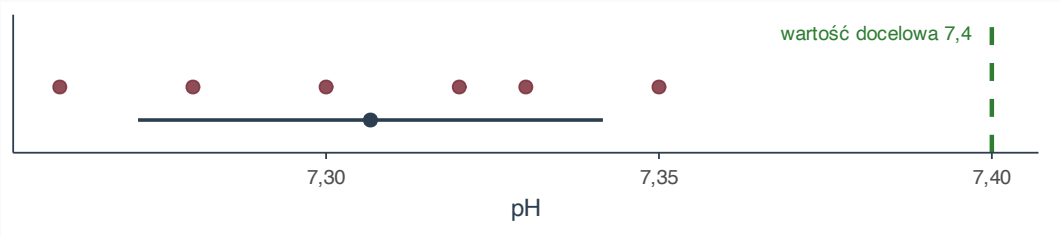
Test t jednej próby

Test t jednej próby (*one-sample t-test*)

Test hipotezy $H_0 : \mu = \mu_0$, w którym średnią jednej próby porównuje się z wartością odniesienia μ_0 ustaloną niezależnie od danych.

- Wartość odniesienia z normy, deklaracji lub celu technologicznego, nie z danych
- Przykłady: masa ciała myszy z hodowli kontra norma referencyjna szczepu; pH hodowli komórkowej kontra wartość docelowa 7,4; zawartość białka w paszy kontra wartość deklarowana przez producenta
- Wynik: istotne odchylenie od wartości odniesienia albo brak podstaw do stwierdzenia odchylenia (co nie dowodzi zgodności)

Test jednej próby — przykład pH hodowli



- Sześć hodowli (punkty); średnia 7,31 i jej 95% przedział ufności (odcinek pod punktami)
- Sygnał: średnia niższa od 7,4 o 0,09; szum: rozrzut bardzo mały (SD 0,03)
- Średnia leży około 7 błędów standardowych poniżej 7,4: $t = -6{,}9$, df $= 5$, $p < 0{,}001$

Wniosek

Średnie pH hodowli jest istotnie niższe od wartości docelowej 7,4.

3. Test t dwóch niezależnych prób

Test t dla prób niezależnych

Test t dla prób niezależnych (*independent two-sample t-test*)

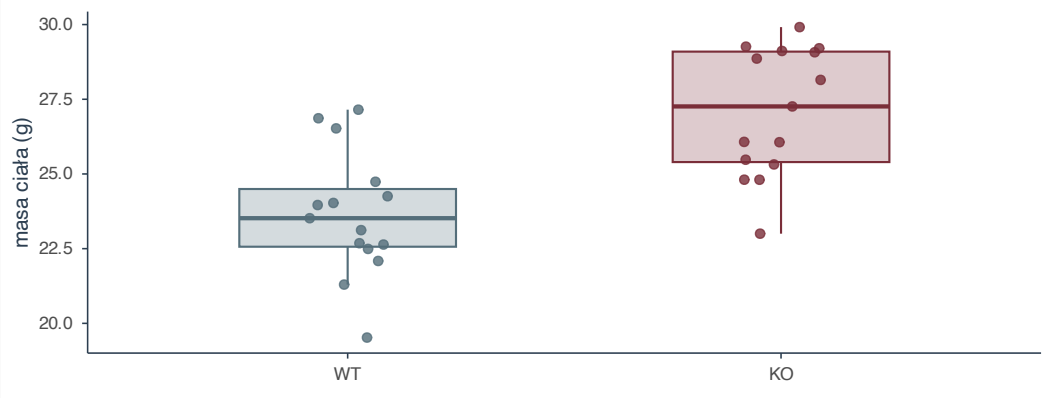
Test hipotezy $H_0 : \mu_1 = \mu_2$ dla dwóch grup złożonych z różnych jednostek, w których przynależność obiektu do jednej grupy nie wpływa na obiekty drugiej.

- kontrola kontra traktowanie
- typ dziki kontra mutant
- osobniki zdrowe kontra chore

Przykład prowadzący

Masa ciała myszy typu dzikiego (WT) kontra nokaut genu (KO), po 15 osobników w grupie.

Kontrola kontra nokaut — dane



Grupa KO wygląda na cięższą — czy różnica jest **wiarygodna**?

Test t-Studenta i test Welcha

	Test t-Studenta (klasyczny)	**Test Welcha**
Założenie	równe wariancje w obu grupach	wariancje mogą się różnić
Szum (błąd standardowy różnicy)	z jednego, wspólnego rozrzutu obu grup	z osobnego rozrzutu każdej grupy
W R	`t.test(..., var.equal = TRUE)`	`t.test(...)` (domyślnie)

Zalecenie praktyczne

Domyślnie stosować test Welcha. Gdy wariancje i liczebności grup są podobne, oba testy dają prawie ten sam wynik.

Dlaczego test Welcha jest domyślny

Grupy (średnie równe, H_0 prawdziwa)	t-Studenta	Welch
po 15 osobników, równy rozrzut	4,2%	4,1%
po 15 osobników, rozrzut 3 razy większy w jednej grupie	5,0%	4,3%
25 i 8 osobników, większy rozrzut w **mniejszej** grupie	22,6%	5,6%
25 i 8 osobników, większy rozrzut w **większej** grupie	0,3%	5,0%

- Odsetek fałszywych alarmów w 4000 symulowanych doświadczeń; powinien wynosić ok. 5%
- Test t-Studenta zawodzi przy różnych wariancjach **i** różnych liczebnościach; Welch trzyma 5%

Wynik testu dla grup WT i KO

	n	średnia (g)	SD (g)
WT	15	23,7	2,1
KO	15	27,1	2,1

- **Sygnał**: myszy KO są cięższe średnio o 3,4 g
- **Szum**: błąd standardowy różnicy 0,8 g; zależy od SD obu grup i ich liczebności
- **Statystyka**: $t \approx 3{,}4 \,/\, 0{,}8 \approx 4{,}4$, df ≈ 28, $p < 0{,}001$ (test Welcha)
- Różnica wielokrotnie przekracza swój błąd standardowy; przy braku efektu taki wynik byłby bardzo rzadki

Przedział ufności dla różnicy średnich

Przedział ufności dla różnicy średnich

Zakres wartości różnicy średnich w populacjach zgodnych z danymi: różnica z próby plus minus mniej więcej dwa błędy standardowe (dokładny mnożnik z rozkładu t).

- KO — WT: 3,4 g, 95% CI od 1,8 do 5,0 g
- Przedział nie zawiera zera, co jest równoważne $p < 0{,}05$
- Środek przedziału to najlepsze oszacowanie różnicy, a szerokość opisuje jego precyzję

Wielkość efektu w przykładzie WT i KO

- d Cohena (wykład 3): różnica średnich wyrażona w odchyleniach standardowych
- W przykładzie: różnica 3,4 g przy typowym rozrzucie około 2,1 g, więc $d \approx 1{,}6$ (efekt duży)
- Odczyt względem progów umownych 0,2 (mały); 0,5 (średni); 0,8 (duży)
- d nie zależy od liczebności, w odróżnieniu od p

> Raport: różnica średnich w gramach, przedział ufności, p oraz d lub inna miara wielkości efektu.

Pełny przebieg analizy: WT i KO

- **Hipotezy**: H_0 — średnia masa myszy WT i KO jest taka sama; H_1 — średnie się różnią
- **Poziom istotności**: $\alpha = 0{,}05$, ustalony przed analizą
- **Test**: test Welcha dla dwóch niezależnych grup po 15 myszy; założenia spełnione: zwierzęta niezależne, rozkłady bez silnej skośności
- **Wynik**: $t \approx 4{,}4$, df ≈ 28, $p < 0{,}001$
- **Decyzja i wniosek**: $p < \alpha$, więc odrzucamy H_0; myszy KO są cięższe od WT średnio o 3,4 g (95% CI od 1,8 do 5,0 g)

4. Założenia i czytanie wyniku

Założenia testu t

Założenie	Co znaczy	Gdy złamane
Skala ilościowa	cecha mierzona na skali interwałowej (przedziałowej) lub ilorazowej	dla skali porządkowej — test rangowy (porównuje kolejność obserwacji, nie ich wartości)
Losowość	obiekty wybrane losowo z populacji lub losowo przydzielone do grup	wniosku nie można uogólnić na populację
Niezależność	każda obserwacja z innej jednostki (inny osobnik, klatka, miot, płytka); obie grupy niezależne od siebie	najgroźniejsze — unieważnia wynik
Normalność	rozkład cechy w **każdej grupie** zbliżony do normalnego; przy większych grupach wystarczy brak silnej skośności i wartości odstających	przy małym n: transformacja lub test rangowy
Równość wariancji	tylko w teście klasycznym; test Welcha jej nie zakłada	stosować wariant Welcha

Hierarchia: **losowość i niezależność** wynikają z planu doświadczenia i są kluczowe; normalność i równość wariancji są znacznie mniej krytyczne.

Pseudoreplikacja (*pseudoreplication*)

Traktowanie obserwacji, które nie są niezależne (powtórzenia techniczne, kilka zwierząt z jednej klatki, kilka studzienek z jednej hodowli), jak niezależnych replik zabiegu.

- Powtórzenia techniczne mówią o precyzji pomiaru, nie o zmienności biologicznej
- Prawdziwe n to liczba niezależnych jednostek, którym przydzielono zabieg (osobników, klatek, hodowli)
- Licząc $n = 15$ zamiast $n = 3$, zaniża się błąd standardowy, więc p wychodzi zbyt małe



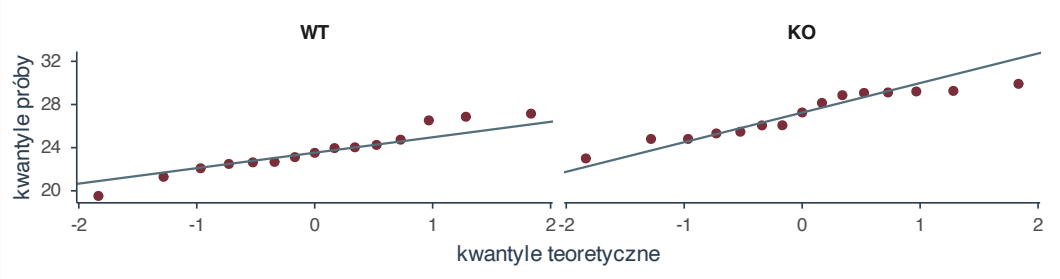
3 myszy
po 5 pomiarów
$n = 3$



15 myszy
po 1 pomiarze
$n = 15$

Kropki: pomiary; kółka: osobniki.

Ocena normalności w grupach



Test Shapiro–Wilka (*Shapiro–Wilk test*)

Test hipotezy zerowej, że próba pochodzi z rozkładu normalnego; mała wartość p wskazuje odstępstwo od normalności.

- Przy małym n test ma niską moc, przy dużym n wykrywa odstępstwa bez znaczenia praktycznego
- Ocenę normalności prowadzi się osobno w każdej grupie, głównie wzrokowo

Testy wstępne założeń

Schemat spotykany w podręcznikach

Test Shapiro–Wilka w każdej grupie → gdy obie normalne: test F równości wariancji (w R `var.test()`) → test t-Studenta albo test Welcha; gdy któraś nienormalna: test rangowy (Manna–Whitneya).

Grupy (dane normalne, średnie równe)	schemat	zawsze Welch
po 15 osobników, równy rozrzut	5,5%	5,2%
25 i 8 osobników, SD 1 i 1,5	8,8%	5,0%
10 i 5 osobników, SD 1 i 2	8,8%	5,4%
20 i 8 osobników, SD 1 i 2	8,4%	5,5%

- Odsetek fałszywych alarmów w 4000 symulowanych doświadczeń; powinien wynosić ok. 5%
- Testy wstępne przy małych próbach często się mylą, a wybór testu na podstawie ich wyniku zwiększa odsetek fałszywych alarmów
- Zalecenie: założenia oceniać **wykresami**, test wybrać **przed analizą** na podstawie planu doświadczenia i rodzaju danych; dla dwóch grup domyślnie **test Welcha**

Gdy założenia nie są spełnione

Założenie	Objaw	Postępowanie
Niezależność	powtórzenia techniczne, te same osobniki	zmiana planu lub analiza uwzględniająca strukturę danych (np. uśrednienie powtórzeń technicznych dla każdego osobnika)
Normalność	silna skośność przy małym n	logarytm danych (zmniejsza skośność np. stężeń) lub test rangowy
Równość wariancji	różny rozrzut w grupach	wariant Welcha (domyślny w R)
Wartości odstające	pojedyncze skrajne obserwacje	sprawdzić źródło; analiza z nimi i bez nich

Zależności nie usunie transformacja ani zmiana testu; można ją jedynie uwzględnić, np. uśredniając pomiary do poziomu niezależnej jednostki (osobnika, klatki). Gdy zabieg przydzielono jednej jednostce na grupę, żadna analiza tego nie naprawi.

Najczęstsze błędy przy teście t

- Wiele porównań parami zamiast jednej analizy dla większej liczby grup (rosnące ryzyko fałszywego alarmu)
- Test niezależnych prób dla danych sparowanych i odwrotnie
- Pseudoreplikacja: obserwacje zależne (powtórzenia techniczne, zwierzęta z jednej klatki) jako niezależne
- Test jednostronny wybrany po obejrzeniu danych
- Usuwanie wartości odstających bez uzasadnienia
- Raport samej wartości p bez różnicy średnich i przedziału ufności

Ćwiczenie: interpretacja wyników testu t

Dla każdego wyniku sformułować wniosek:

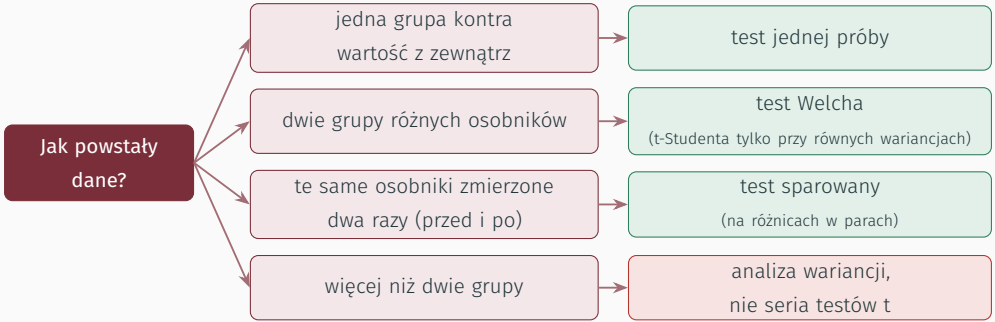
- $p = 0{,}033$; 95% CI różnicy od $0{,}2$ do $4{,}6$ g.
- $p = 0{,}30$; 95% CI od $-1{,}5$ do $4{,}3$ g.
- Trzy myszy po pięć pomiarów, test na 15 obserwacjach daje $p = 0{,}001$.

Ćwiczenie: interpretacja wyników testu t

Dla każdego wyniku sformułować wniosek:

- $p = 0{,}033$; 95% CI różnicy od $0{,}2$ do $4{,}6$ g. $\rightarrow$ **różnica istotna**, ale jej wielkość niepewna (od 0,2 do 4,6 g)
- $p = 0{,}30$; 95% CI od $-1{,}5$ do $4{,}3$ g. $\rightarrow$ **brak podstaw do odrzucenia** H_0; szeroki przedział, mała próba
- Trzy myszy po pięć pomiarów, test na 15 obserwacjach daje $p = 0{,}001$. $\rightarrow$ **pseudoreplikacja**; prawdziwe $n = 3$, wynik niewiarygodny

Który wariant testu t



Silnie skośne dane przy małym n: transformacja (np. logarytm) lub test rangowy.

Ćwiczenie: dobór wariantu testu

Dla każdej sytuacji wskazać właściwą analizę:

- Masa 12 myszy przed i po czterech tygodniach diety.
- Ekspresja genu w 10 guzach i w 10 zdrowych tkankach od innych pacjentów.

- Zawartość wody w 8 partiach nasion kontra wartość deklarowana 12%.
- Plon czterech odmian pszenicy.

Ćwiczenie: dobór wariantu testu

Dla każdej sytuacji wskazać właściwą analizę:

- Masa 12 myszy przed i po czterech tygodniach diety. $\longrightarrow$ **test sparowany**
- Ekspresja genu w 10 guzach i w 10 zdrowych tkankach od innych pacjentów. $\longrightarrow$ **test Welcha**, próby niezależne
- Zawartość wody w 8 partiach nasion kontra wartość deklarowana 12%. $\longrightarrow$ **test jednej próby**
- Plon czterech odmian pszenicy. $\longrightarrow$ **analiza wariancji**, nie seria testów t

5. W praktyce: R

Test t jednej próby w R

```r
ph <- c(7.32, 7.28, 7.35, 7.30, 7.26, 7.33)
t.test(ph, mu = 7.4)
```

```
    One Sample t-test

data:  ph
t = -6.8723, df = 5, p-value = 0.0009977
alternative hypothesis: true mean is not equal
  to 7.4
95 percent confidence interval:
 7.271756 7.341578
sample estimates:
mean of x
 7.306667
```

- Argument **mu = 7,4** to wartość odniesienia z H_0
- $p < 0{,}001$ → średnie pH hodowli istotnie niższe niż 7,4
- CI **[7,27; 7,34]** dotyczy średniej populacji i nie zawiera 7,4 — ten sam wniosek co z p

Test t dwóch prób (Welch) w R

```r
t.test(masa ~ genotyp, data = myszy)
```

```
    Welch Two Sample t-test

data:  masa by genotyp
t = -4.4376, df = 27.982, p-value = 0.0001289
alternative hypothesis: true difference in means between
  group WT and group KO is not equal to 0
95 percent confidence interval:
 -5.018243 -1.848465
sample estimates:
mean in group WT mean in group KO
        23.65918         27.09254
```

- masa ~ genotyp: „masa w zależności od genotypu”
- Układ wydruku jak na wykładzie 3; tu $p < 0{,}001$, CI [-5,0; -1,8] nie zawiera zera
- Różnica i CI w kolejności WT — KO (wiersz `alternative hypothesis`)
- Ułamkowe df = 27,98 to cecha testu Welcha
- `var.equal = TRUE`: test t-Studenta; tu $p < 0{,}001$, prawie jak w teście Welcha, bo rozrzut w grupach jest podobny

Podsumowanie

Do zapamiętania

- Test t porównuje **średnie** cechy **ciągłej**
- Statystyka t to **sygnał przez szum**: rośnie z różnicą średnich i liczebnością, maleje z odchyleniem standardowym
- Wartość p to **pole ogonów** rozkładu t poza obserwowaną statystyką
- Dwie grupy: **test t-Studenta** zakłada równe wariancje, **test Welcha** nie; domyślnie Welch
- **Niezależność** to najważniejsze założenie — należy unikać **pseudoreplikacji**
- Raportować kierunek, wielkość efektu i **przedział ufności**, nie samo p

Na następny wykład

Wykład 5: Testy statystyczne II

- Test t **sparowany**; bloki i jednostka doświadczalna
- Dane **kategoryczne**: chi-kwadrat, test Fishera, test McNemara
- Testy **nieparametryczne**
- Drzewo wyboru właściwego testu

Dziękuję za uwagę

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