BIOSTATYSTYKA

Wykład 2: Populacje i próby danych

dr inż. Bartosz Czech

Katedra Genetyki · Uniwersytet Przyrodniczy we Wrocławiu

bartosz.czech@upwr.edu.pl

Plan wykładu

1. Populacja a próba — o czym naprawdę wnioskujemy
2. Rozkład cechy i statystyki opisowe — położenie, rozproszenie, kształt
3. Rozkład normalny i reguła trzech sigm (68–95–99,7)
4. Błąd próbkowania: SD kontra SE

Cel wykładu

Nauczyć się opisywać zmienność liczbami i rozumieć, dlaczego wniosek z próby zawsze niesie niepewność.

1. Populacja a próba

Populacja i próba

Populacja (*population*)

- Zbiór wszystkich obiektów, o których chcemy wnioskować
- Parametr (*parameter*) — liczba opisująca populację, np. średnia μ lub odchylenie standardowe σ (miara rozrzutu, część 2)
- Jej parametry są zwykle nieznane; całej populacji zwykle nie da się zmierzyć

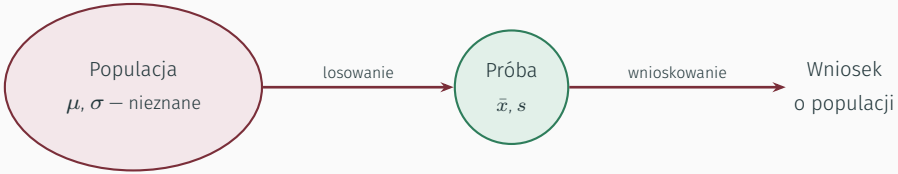
Próba (*sample*)

- Podzbiór populacji, który faktycznie zbadano
- Statystyka (*statistic*) — liczba obliczona z próby: $\bar{x}$, s; służy do oszacowania parametru
- Znana, ale tylko fragment całości

Notacja, która wraca przez cały kurs

Litery greckie = prawda o populacji (nieznana). Litery łacińskie = policzone z próby.

Od próby do wniosku



- Mierzymy **próbę**, ale wniosek dotyczy **populacji**
- Sposób doboru decyduje o braku **systematycznego** przesunięcia wniosku, liczebność — o jego **precyzji**

Ograniczenia pomiaru całej populacji

- **Niemożliwe** — nie zbadamy wszystkich myszy szczepu C57BL/6 na świecie
- **Za drogie i za długie** — pełny pomiar przekracza budżet i czas projektu
- **Niszczące** — pomiar zużywa badany obiekt (preparat histologiczny wymaga pobrania tkanki, test kiełkowania zużywa nasiona); zbadanie całej populacji oznaczałoby jej zniszczenie

> Dlatego mierzymy **fragment** i wnioskujemy o całości — świadomie przyjmując niepewność, którą statystyka pozwala oszacować.

Reprezentatywność a wiarygodność wniosku

Kluczowa zasada

Wniosek jest tak dobry, jak dobór próby. Próba **niereprezentatywna** prowadzi do fałszywego wniosku — niezależnie od tego, jak zaawansowany test zastosujemy.

- Mierząc tylko **największe rośliny z brzegu poletka**, zawyżymy średni plon całego pola
- Systematyczny błąd doboru (*selection bias*; definicja za chwilę) nie znika przy większej próbie ani lepszym teście
- W 1936 r. sondaż *Literary Digest*, oparty na ok. 2,4 mln odpowiedzi, błędnie wskazał zwycięzcę wyborów prezydenckich w USA: próba była liczna, lecz obciążona

Próba reprezentatywna

Próba reprezentatywna (*representative sample*)

Próba, w której struktura cech istotnych dla badania (np. udział płci, grup wieku, szczepów, warunków hodowli) odpowiada strukturze populacji, dzięki czemu statystyki z próby szacują parametry populacji bez systematycznego przesunięcia.

- Reprezentatywność zależy od **sposobu doboru**, a nie od liczebności: duża próba może być niereprezentatywna
- **Losowy dobór** chroni przed **systematycznym** odchyleniem próby od populacji: każdy obiekt ma znaną, niezerową szansę trafienia do próby
- Wymaga **operatu losowania** (*sampling frame*), czyli listy lub procedury wskazującej wszystkie obiekty populacji
- Jest względna: dotyczy konkretnej populacji i cech ważnych dla pytania badawczego

> Losowanie nie gwarantuje, że **każda** próba jest reprezentatywna; gwarantuje brak systematycznego przesunięcia. Mała próba może przypadkiem odbiegać od populacji.

Schematy losowego doboru próby

Schemat	Zasada	Przykład biologiczny
Prosty losowy (*simple random*)	każdy obiekt i każdy zestaw n obiektów ma taką samą szansę wyboru	20 myszy wylosowanych ze 200 oznaczonych numerami
Warstwowy (*stratified*)	populacja dzielona na rozłączne warstwy według ważnej cechy; losowanie osobno w każdej warstwie	po 10 losowo wybranych samców i samic z kolonii
Systematyczny (*systematic*)	losowy start, a potem co k-ty obiekt uporządkowanej listy	co piąta roślina w rzędzie, start od losowo wybranej z pierwszych pięciu
Zespołowy (*cluster*)	losowanie całych naturalnych grup; w wybranych grupach bada się wszystkie obiekty	4 poletka z 20, wszystkie rośliny z tych poletek
Wielostopniowy (*multistage*)	losowanie w kilku etapach: najpierw grupy, potem obiekty w wylosowanych grupach	5 stawów z 30, w każdym 10 losowo odłowionych ryb

Warstwowy zmniejsza niepewność, gdy cecha silnie wpływa na wynik. Systematyczny zawodzi, gdy lista ma okresowość zgodną z k. Zespołowy jest tańszy, ale daje większą niepewność niż prosty losowy o tej samej liczebności (obiekty w grupie są podobne).

Błędy doboru próby

Błąd doboru (*selection bias*)

Systematyczna różnica między próbą a populacją, wynikająca ze sposobu, w jaki obiekty trafiły do próby.

Rodzaj	Mechanizm	Przykład biologiczny
Dobór wygodny (*convenience*)	badane są obiekty najłatwiej dostępne	liście pobrane tylko z roślin przy ścieżce
Samoselekcja (*volunteer bias*)	zgłaszają się osoby o określonych cechach	ochotnicy do badania diety zainteresowani zdrową żywnością
Błąd przeżywalności (*survivorship bias*)	analizowane są tylko obiekty, które przeszły etap selekcji	ocena leku wyłącznie na zwierzętach żywych w dniu 28
Ubytki (*attrition*)	obiekty odpadają z badania nielosowo	rezygnacja osobników z działaniami niepożądanymi
Brak odpowiedzi (*non-response*)	część wylosowanych nie odpowiada, a odmawiający różnią się od odpowiadających	ankieta o zdrowotności stad odesłana głównie przez hodowców bez problemów

Błąd systematyczny a błąd losowy

Większa liczebność zmniejsza błąd losowy, ale nie usuwa błędu systematycznego doboru. To rozróżnienie jest analogiczne do dokładności i precyzji pomiaru (W1).

Losowy dobór a losowy przydział

Losowy dobór (*random sampling*)

- Jak wybieramy obiekty z populacji do próby
- Umożliwia uogólnienie wniosku na populację: trafność zewnętrzna (*external validity*)

Losowy przydział (*random assignment*)

- Jak rozdzielamy wybrane obiekty między grupy (kontrola, zabieg)
- Pozwala przypisać różnice zabiegowi, a nie czynnikom ubocznym: trafność wewnętrzna (*internal validity*)

> Jedno nie zastępuje drugiego. Zwierzęta z jednej hodowli, losowo przydzielone do grup, dają wiarygodne porównanie grup, lecz uogólnienie na cały gatunek wymaga ostrożności.

Ćwiczenie: rozpoznawanie błędu doboru

Dla każdego opisu wskazać błąd doboru lub schemat losowania:

- Ankietę o nawykach żywieniowych wysłano e-mailem wszystkim studentom; odpowiedziało 18%.
- Efekt leku oceniono wyłącznie na zwierzętach, które dożyły końca 28-dniowego doświadczenia.
- Z kolonii 200 myszy wylosowano 10 klatek po 5 zwierząt i zbadano wszystkie zwierzęta z tych klatek.
- Do pomiaru wzrostu roślin wybrano 20 egzemplarzy rosnących najbliżej ścieżki.

Ćwiczenie: rozpoznawanie błędu doboru

Dla każdego opisu wskazać błąd doboru lub schemat losowania:

- Ankietę o nawykach żywieniowych wysłano e-mailem wszystkim studentom; odpowiedziało 18%. → samoselekcja i brak odpowiedzi
- Efekt leku oceniono wyłącznie na zwierzętach, które dożyły końca 28-dniowego doświadczenia. → błąd przeżywalności
- Z kolonii 200 myszy wylosowano 10 klatek po 5 zwierząt i zbadano wszystkie zwierzęta z tych klatek. → dobór zespołowy, schemat losowy
- Do pomiaru wzrostu roślin wybrano 20 egzemplarzy rosnących najbliżej ścieżki. → dobór wygodny

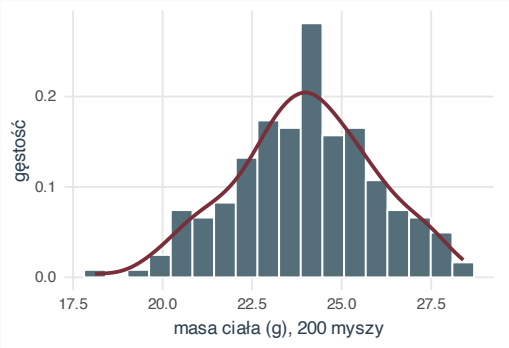
2. Rozkład cechy i statystyki opisowe

Rozkład cechy, histogram i gęstość

Rozkład cechy (*distribution*)

Opis tego, jak często występują poszczególne wartości cechy; dla cechy ciągłej — jak często występują wartości z poszczególnych przedziałów.

- Histogram (*histogram*) — wykres słupkowy liczebności obserwacji w kolejnych przedziałach wartości
- Krzywa gęstości (*density curve*) — wygładzony histogram; pole pod krzywą między dwiema wartościami = odsetek obserwacji w tym zakresie (całe pole = 1)



Położenie, rozproszenie i kształt

Gdzie jest środek?

Położenie (*central tendency*) — typowa, reprezentatywna wartość.

średnia, mediana, dominanta

Jak szeroko rozrzucone?

Rozproszenie (*dispersion*) — jak bardzo osobniki różnią się między sobą.

rozstęp, wariancja, SD, IQR

Jaki kształt?

Kształt (*shape*) — symetria rozkładu i grubość jego ogonów.

skośność, kurtoza

> Sam środek nie wystarcza. Dwie grupy o tej samej średniej mogą być zupełnie różne — o tym decyduje rozproszenie i kształt rozkładu.

Miary położenia

Miara	Co opisuje	Uwaga
Średnia	Suma podzielona przez liczebność	Wrażliwa na wartości skrajne
Mediana	Wartość środkowa po uporządkowaniu	Odporna na skrajne wartości
Dominanta	Wartość najczęstsza	Przydatna dla danych kategorycznych

Wybór miary zależy od skali pomiarowej (nominalna — tylko dominanta; porządkowa — mediana i dominanta) oraz od kształtu rozkładu: przy symetrii średnia i mediana są zbliżone, przy skośności się rozchodzą.

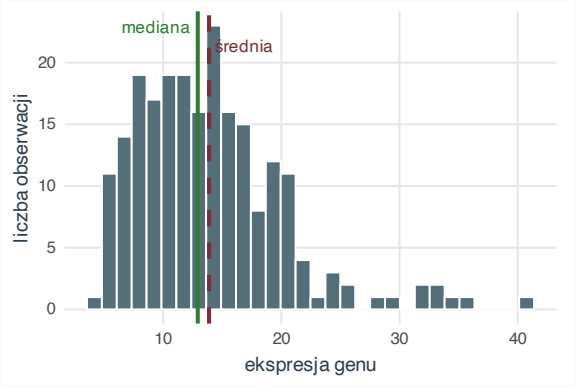
Średnia kontra mediana — wartości odstające i skośność

Ekspresja genu (5 próbek)

10	12	11	13	90

Jedna próbka silnie odstająca (np. artefakt hodowli).

- Średnia = 27,2 — przesunięta przez wartość odstającą
- Mediana = 12 — odporna na nią



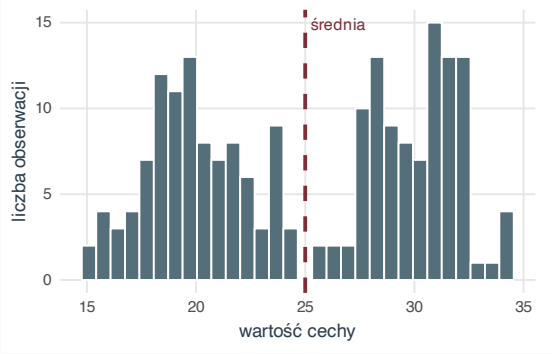
Wartość odstająca **ciągnie** średnią; skośność **przesuwa** ją w stronę długiego ogona.

Dominanta i rozkłady wielomodalne

Dominanta (*mode*)

Wartość występująca najczęściej; w rozkładzie ciągłym wartość, przy której krzywa gęstości ma maksimum (na histogramie: najwyższy przedział).

- Rozkład **jednomodalny** ma jeden szczyt, **dwumodalny** (*bimodal*) — dwa
- Dwa szczyty zwykle oznaczają **dwie podpopulacje**: płcie, genotypy, podtypy choroby
- Średnia i mediana *mogą* leżeć wtedy między szczytami, tam gdzie prawie nie ma obserwacji



Rodzaje średnich

Średnia arytmetyczna, geometryczna, harmoniczna i ważona

- Arytmetyczna (*arithmetic mean*) — suma wartości podzielona przez ich liczbę; właściwa dla wielkości, które się sumują, np. masy ciała
- Geometryczna (*geometric mean*) — pierwiastek n-tego stopnia z iloczynu wartości (tylko dla wartości dodatnich); właściwa dla ilorazów i przyrostów procentowych
- Harmoniczna (*harmonic mean*) — odwrotność średniej z odwrotności wartości; właściwa dla ilorazów przy stałym liczniku (np. średnia prędkość na odcinkach równej długości), a w genetyce populacji — dla efektywnej wielkości populacji w kolejnych pokoleniach
- Ważona (*weighted mean*) — średnia, w której wartości mają różne wagi, np. liczebności grup

Przykłady

Zmiany ekspresji $0{,}5$ (spadek o połowę) i 2 (wzrost dwukrotny): średnia arytmetyczna to $1{,}25$, a geometryczna 1, czyli brak zmiany netto. Grupa 10 osobników ze średnią 20 i grupa 30 osobników ze średnią 24 dają średnią ważoną $(10 \cdot 20 + 30 \cdot 24)/40 = 23$.

Percentyle i kwantyle

Percentyl (*percentile*) i kwantyl (*quantile*)

p-ty percentyl to wartość, poniżej której leży $p\%$ obserwacji. Kwantyl rzędu q $(0 < q < 1)$ to ta sama idea wyrażona ułamkiem, a nie procentem.

- Kwartyle (*quartiles*) Q1, Q2, Q3 dzielą uporządkowane dane na cztery równe części: poniżej Q1 leży 25% obserwacji, poniżej Q3 — 75%; Q2 to mediana (50. percentyl); decyle dzielą dane na dziesięć części
- Zastosowania: siatki centylowe wzrostu i masy dzieci, przedziały referencyjne wyników laboratoryjnych (zwykle od 2,5. do 97,5. percentyla osób zdrowych)
- Percentyle środkowe (np. kwartyle) są odporne na pojedyncze wartości skrajne, dlatego dobrze opisują rozkłady skośne

Przy małej próbie kwantyl zwykle wypada między dwiema obserwacjami; programy stosują różne reguły interpolacji, więc mogą podawać nieco inne kwartyle — przy dużych próbach różnice znikają.

Miary rozproszenia

- **Rozstęp** (*range*): różnica maksimum i minimum — prosty, ale wrażliwy na skrajne wartości
- **Wariancja** (*variance*): przeciętny kwadrat odchylenia od średniej (jednostki do kwadratu); populacji σ^2 — odchylenia od μ, dzielnik N; z próby s^2 — suma kwadratów odchyleń od $\bar{x}$ podzielona przez $n-1$
- **Odchylenie standardowe** (SD, *standard deviation*): pierwiastek z wariancji (σ lub s) — w tych samych jednostkach co dane
- **Rozstęp międzykwartylowy** (IQR, *interquartile range*): $\mathrm{IQR} = Q3 - Q1$, szerokość środkowych 50% danych — odporny na wartości odstające

Dlaczego $n-1$, a nie n

Średnia z próby leży bliżej obserwacji niż nieznane μ, więc dzielenie przez n zaniżałoby rozrzut. Odchylenia od $\bar{x}$ sumują się do zera, więc po poznaniu $n-1$ z nich ostatnie jest wyznaczone — stąd $n-1$ **stopni swobody** (*degrees of freedom*).

Odchylenie standardowe — obliczenie na przykładzie

Odchylenie standardowe z próby

Pierwiastek ze średniego kwadratu odchyleń od średniej, z dzieleniem przez $n-1$:
$s = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$.

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	20	-4	16
	22	-2	4
	23	-1	1
	25	1	1
	30	6	36
suma 120, $\bar{x} = 24$		0	58

- Wariancja $s^2 = 58/(5-1) = 14{,}5$, odchylenie $s \approx 3{,}81$
- Odchylenia od średniej sumują się do zera, dlatego podnosimy je do kwadratu

Współczynnik zmienności (CV)

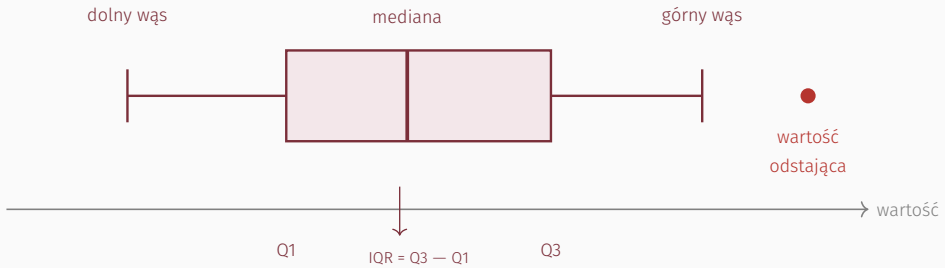
CV = odchylenie standardowe odniesione do średniej (zwykle w %).

Po co, skoro mamy SD?

CV jest **bezwymiarowy** — pozwala porównać zmienność cech o różnych jednostkach lub bardzo różnych średnich.

- Czy masa ciała jest *względnie* bardziej zmienna niż długość ogona? SD w gramach i w centymetrach nie da się porównać wprost — CV tak
- Przydatny do oceny **powtarzalności** metody pomiarowej (im niższy CV, tym stabilniejsza)
- Ma sens tylko dla cech na skali **ilorazowej** o średniej wyraźnie większej od zera (nie dla temperatury w °C ani wartości ujemnych)

Boxplot — graficzne podsumowanie rozkładu



- Pudełko = środkowe 50% danych (od Q1 do Q3, długość = IQR); kreska w środku = **mediana**
- Wąsy sięgają najdalszych obserwacji leżących nie dalej niż 1,5 IQR od brzegów pudełka
- Punkty poza wąsami to **kandydaci na wartości odstające**

Wartości odstające — reguła 1,5 IQR

Wartość odstająca na wykresie pudełkowym (reguła Tukeya)

Obserwacja leżąca poniżej $Q1 - 1{,}5 \cdot \mathrm{IQR}$ lub powyżej $Q3 + 1{,}5 \cdot \mathrm{IQR}$.

- Przykład z danymi 10, 11, 12, 13, 90: $Q1 = 11$, $Q3 = 13$, $\mathrm{IQR} = 2$; granice 8 i 16, więc **90** jest odstająca, a 10 nie
- Reguła **wskazuje kandydatów**, nie dowodzi błędu: wartość odstająca bywa pomyłką pomiaru lub zapisu, prawdziwą rzadką obserwacją albo sygnałem innej podpopulacji

Postępowanie

Najpierw sprawdzić źródło (protokół, zapis, kalibrację). Nie usuwać bez uzasadnienia, tylko raportować wynik z nią i bez niej. Gdy odstających jest dużo, stosować miary odporne: medianę i IQR.

Raportowanie statystyk opisowych

Rozkład	Miary	Zapis (przykład)
symetryczny	średnia i SD	$23{,}4 \pm 1{,}8$ g (średnia $\pm$ SD, $n = 40$)
skośny lub z wartościami odstającymi	mediana i kwartyle (Q1–Q3)	$12{,}0$ ($11{,}0$–$13{,}0$) ng/ml ($n = 40$)
dwumodalny	opis osobno w podgrupach	dwa wiersze, po jednym na grupę

- Zawsze podawać liczebność n i jednostkę
- Zaznaczyć, czy po znaku $\pm$ stoi SD, czy SE (błąd standardowy średniej, część 4)
- Liczba miejsc po przecinku nie większa niż precyzja pomiaru

3. Rozkład normalny

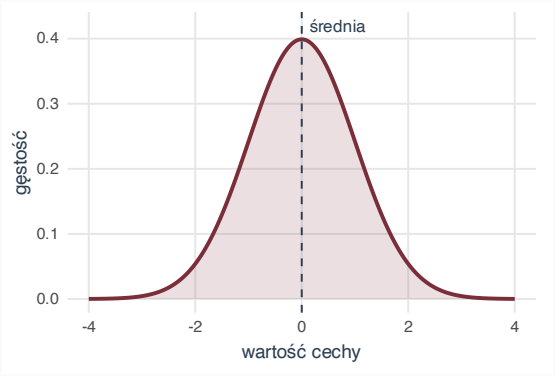
Rozkład normalny

Rozkład normalny (*normal distribution*)

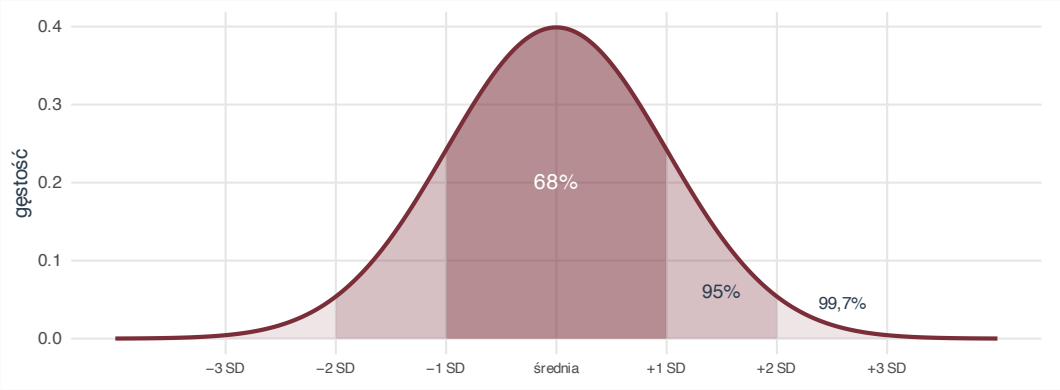
Rozkład cechy ciągłej o symetrycznej, dzwonowej krzywej gęstości, całkowicie określony przez średnią μ i odchylenie standardowe σ.

- Zapis $N(\mu, \sigma)$; w wielu podręcznikach $N(\mu, \sigma^2)$ – z wariancją
- Średnia = mediana = dominanta
- Pole pod krzywą w danym zakresie = odsetek obserwacji w tym zakresie

Wiele testów zakłada, że dane (lub ich średnie) mają rozkład w przybliżeniu normalny.



Reguła trzech sigm (68–95–99,7)



W rozkładzie normalnym ok. 68% obserwacji mieści się w odległości jednego σ od średniej, 95% — dwóch, 99,7% — trzech (stąd „trzy sigmy”). Przy rozkładach skośnych reguła nie obowiązuje.

Standaryzacja i rozkład standardowy

Standaryzacja (*standardization*)

Przekształcenie wartości x na liczbę odchyleń standardowych od średniej; wynik to wartość standaryzowana (*z-score*): $z = (x - \mu)/\sigma$, a dla próby $z = (x - \bar{x})/s$.

- $z = 2$: obserwacja leży 2 SD powyżej średniej; $z = -1$: 1 SD poniżej
- Standardowy rozkład normalny $N(0, 1)$ ma średnią 0 i SD równe 1; każdy rozkład normalny sprowadza się do niego przez standaryzację
- Reguła 68–95–99,7 w skali z: $|z| \leq 1$, $|z| \leq 2$, $|z| \leq 3$

Przykład

Masa myszy 27 g przy średniej 24 g i SD równym 2 g daje $z = (27 - 24)/2 = 1{,}5$: mysz jest o 1,5 SD cięższa od przeciętnej.

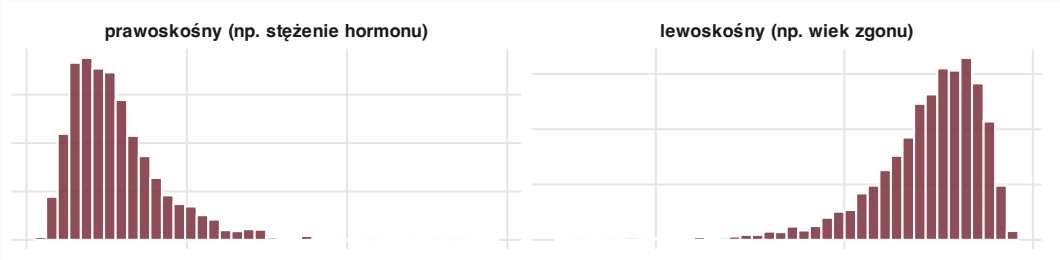
Pochodzenie rozkładu normalnego cech

- *Jeśli* cecha powstaje jako suma wielu drobnych, niezależnych wpływów (wielu genów i wielu czynników środowiskowych), jej rozkład zbliża się do normalnego
- Gdy wpływy nie sumują się, lecz mnożą, lub dominuje jeden czynnik, rozkład bywa skośny
- Stąd wzrost, masa ciała czy ciśnienie krwi bywają w przybliżeniu normalne

Intuicja

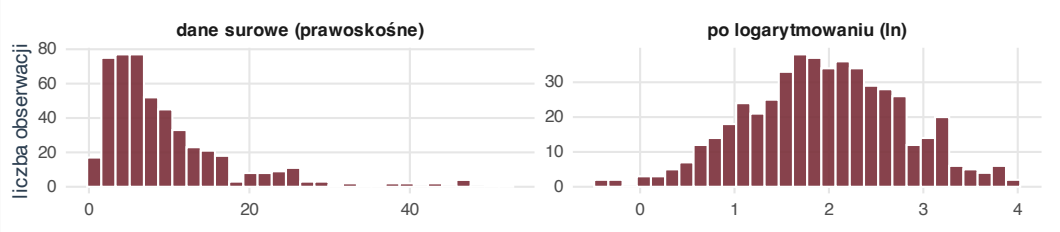
Jeden gen daje efekt skokowy. Setki genów plus środowisko dają płynne, symetryczne „rozmycie” wokół typowej wartości.

Odstępstwa od rozkładu normalnego



- Skośność (*skewness*) — miara asymetrii rozkładu: dodatnia, gdy dłuższy jest ogon po prawej; ujemna, gdy po lewej
- Kurtoza (*kurtosis*) — miara grubości ogonów w porównaniu z rozkładem normalnym: ogony cięższe oznaczają więcej wartości skrajnych
- Rozkłady skośne są w biologii częste (stężenia, ekspresja, czas przeżycia); mediana i IQR opisują je lepiej niż średnia i SD

Skośność a przekształcenie logarytmiczne

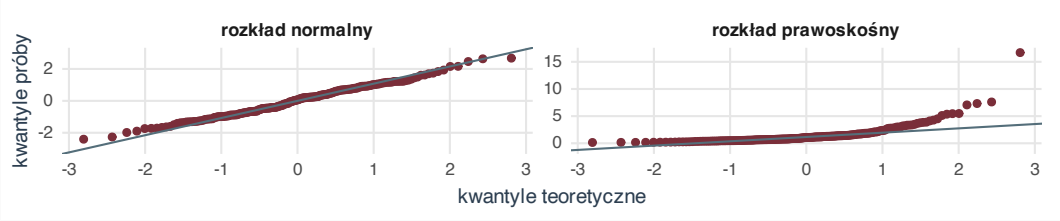


Rozkład log-normalny (*log-normal distribution*)

Rozkład zmiennej, której logarytm ma rozkład normalny; typowy dla stężeń, ekspresji genów i liczebności.

- Analizujemy $y = \ln x$; średnią $\bar{y}$ przekształcamy z powrotem przez $e^{\bar{y}}$ — wynik to średnia geometryczna
- Logarytmu nie stosuje się do zer i wartości ujemnych

Wykres kwantylowy (QQ-plot)



Wykres kwantylowy (*quantile–quantile plot*)

Wykres, w którym kwantyle próby zestawia się z kwantylami standardowego rozkładu normalnego; punkty leżą na prostej, gdy rozkład próby jest bliski normalnemu.

- Łuk wygięty w górę po prawej: dłuższy prawy ogon; kształt litery S: ogony cięższe lub lżejsze

Normalność oceniamy tu wzrokowo: testy normalności (np. Shapiro–Wilka) przy dużym n wskazują odstępstwa bez znaczenia praktycznego, a przy małym często nie wykrywają nawet wyraźnych.

4. Błąd próbkowania: SD a SE

Źródło błędu próbkowania

- Gdyby 100 osób wylosowało niezależnie po 10 myszy **z tej samej populacji**, każda dostałaby nieco inną średnią
- To **nie jest błąd pomiaru** — to naturalna zmienność samego losowania

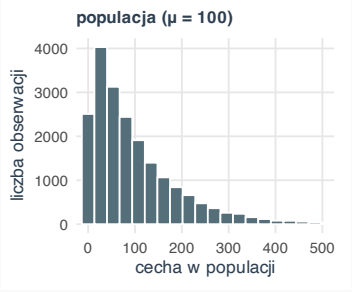
Rozkład z próbkowania (*sampling distribution*)

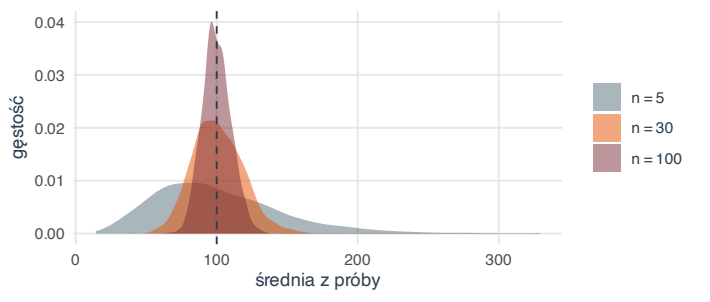
Rozkład wartości statystyki (tu: średniej) we wszystkich możliwych próbach o tej samej liczebności n z danej populacji. Pokazuje, jak bardzo średnia z próby może się wahać wokół prawdziwej średniej μ.

Błąd standardowy średniej (*standard error*, SE)

Odchylenie standardowe rozkładu z próbkowania średniej; teoretycznie $\sigma/\sqrt{n}$, a w praktyce szacowane z próby jako $s/\sqrt{n}$.

Wielkość próby a rozkład średniej





Centralne twierdzenie graniczne (CTG, *central limit theorem*)

Dla niezależnych obserwacji z populacji o skończonej wariancji rozkład średniej z próby zbliża się wraz ze wzrostem n do rozkładu normalnego o średniej μ i odchyleniu standardowym $\sigma/\sqrt{n}$ — nawet gdy sama cecha nie ma rozkładu normalnego.

Precyzja a liczebność próby

n	SE przy $\sigma = 10$
10	$3{,}16$
40	$1{,}58$
160	$0{,}79$
640	$0{,}40$

- Czterokrotny wzrost n zmniejsza SE dwukrotnie
- Od 10 do 40 osobników SE maleje o ok. 1,6; od 160 do 640 tylko o ok. 0,4

Malejące korzyści

Koszt badania rośnie proporcjonalnie do n, a zysk w precyzji maleje, dlatego liczebność planuje się z góry.

SD a SE — trzy rozkłady

Rozkład	Co opisuje	Miara rozrzutu	Zależy od n?	Kiedy raportować
Cechy w populacji	jak cecha różni się między wszystkimi osobnikami; nieznany	σ	—	—
Cechy w próbie	histogram n zmierzonych obserwacji; obserwujemy go	s (SD)	nie — szacuje σ	zmienność osobników w grupie
Średnich z próbkowania	rozkład średnich ze wszystkich prób o tym samym n; teoretyczny	$\sigma/\sqrt{n}$ (SE), szacowany przez $s/\sqrt{n}$	tak — maleje z n	niepewność oszacowania średniej

- Średnia rozkładu z próbkowania równa się μ: średnia z próby nie jest systematycznie przesunięta w żadną stronę
- SD opisuje rozrzut osobników, SE — precyzję średniej; to miary dwóch różnych rozkładów

Pułapka publikacyjna: SD czy SE?

Uwaga przy czytaniu wykresów

Słupki błędów bywają podawane jako SE, bo są krótsze i wykres wygląda korzystniej — sugerując mniejszą zmienność, niż jest naprawdę.

- Zawsze sprawdzać w podpisie, czy to SD, SE, czy przedział ufności (wykład 3)
- Te same dane z SE wyglądają na „mocniejszy” wynik niż z SD
- Bez tej informacji słupki błędów są nieczytelne

5. W praktyce w R

Statystyki opisowe w R

Ramka `myszy_dieta` (dane przykładowe): masa ciała (g) 40 myszy, po 20 w grupie kontrolnej i na diecie. Opis zaczyna się od jednej grupy:

```
kontrola <- subset(myszy_dieta, grupa == "kontrola")
summary(kontrola$masa)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.20   21.18   22.00   22.20   22.80   25.10
```

- **Mean** (22,20) i **Median** (22,00): porównanie średniej z medianą wskazuje na symetrię lub skośność
- **1st/3rd Qu.** wyznaczają **IQR** = 22,8 — 21,2 = 1,6 g (środkowe 50% zwierząt); zakres **Min–Max**: 20,2–25,1 g
- Statystyki liczy się **osobno w każdej grupie**: rozrzut wszystkich 40 myszy obejmowałby też różnicę między dietami
- Przed liczeniem średniej obejrzeć histogram (`hist`); `NA` w danych wymaga `na.rm = TRUE`

Statystyki w rozbiciu na grupy

```
tapply(myszy_dieta$masa, myszy_dieta$grupa, mean)
tapply(myszy_dieta$masa, myszy_dieta$grupa, sd)
tapply(myszy_dieta$masa, myszy_dieta$grupa,
       function(x) sd(x) / sqrt(length(x)))   # SE
```

```
kontrola    dieta
  22.205   24.600
kontrola    dieta
1.447130 1.220009
 kontrola     dieta
0.3235880 0.2728022
```

- `tapply(wartość, grupa, funkcja)` — statystyka **osobno w każdej grupie**; funkcją może być własna reguła, tu SE $= s/\sqrt{n}$
- Średnia: **dieta** 24,6 g, kontrola 22,2 g; **SD** (1,45 i 1,22 g) — rozrzut osobników, **SE** (0,32 i 0,27 g) — precyzja średniej
- To **opis**, nie test — czy różnica jest istotna, rozstrzygnie test (W3–W4)

Podsumowanie

Do zapamiętania

- Próba $\neq$ populacja: **statystyka** ($\bar{x}$, s) szacuje **parametr** (μ, σ)
- O wniosku decyduje **sposób doboru**: losowanie chroni przed błędem systematycznym przy wyborze obiektów (nie przed brakiem odpowiedzi i ubytkami), a większa liczebność zmniejsza tylko błąd losowy
- Losowy **dobór** umożliwia uogólnienie wniosku na populację, losowy **przydział** — przypisanie różnic zabiegowi
- Przy skośności **mediana i IQR** są bezpieczniejsze niż średnia i SD
- **SD** opisuje rozrzut osobników, **SE** — rozrzut średniej z próby (maleje z n); średnie mają w przybliżeniu rozkład normalny (CTG)
- Nie wszystko jest normalne — rozkład oceniamy wzrokowo (histogram, QQ-plot)

Na następny wykład

Wykład 3: Testowanie hipotez i estymacja

- Schemat testu statystycznego
- Jak czytać **wartość p**
- Błędy I i II rodzaju
- Estymacja przez **przedział ufności**

Dziękuję za uwagę

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