Łukasz Banasiak

Olimpiada Biologiczna
Informator

część statystyczna i filogenetyczna

Warszawa, kwiecień 2017

Drodzy Olimpijczycy!

Zagadnienia dotyczące statystyki nie są objęte programem nauczania biologii w szkołach średnich. Do rzadkości należy także ich omawianie na zajęciach z matematyki. Niemniej jednak dla współczesnego biologa statystyka jest bardzo ważnym narzędziem i to praktycznie niezależnie od specjalności jaką się zajmuje. Z tego powodu na wzór Międzynarodowej Olimpiady Biologicznej zostały wprowadzone do programu Krajowej Olimpiady Biologicznej zadania leżące na pograniczu biologii i matematyki. Kolejną dyscypliną gdzie metody matematyczne w biologii są szczególnie rozwinięte jest filogenetyka, której podstawy także spróbuję Wam nieco przybliżyć.

W ramach przygotowań do zawodów centralnych polecam regularne rozwiązywanie zadań. Nie chodzi tutaj tylko o wykonie rachunków, ale także o zapoznanie się z odpowiednią literaturą. Bardzo liczę na Waszą aktywną pracę i samodzielne zgłębianie coraz bardziej skomplikowanych treści. Materiał jest podzielony na lekcje. Przed rozpoczęciem każdej kolejnej znajdują się opatrzone komentarzem prawidłowe rozwiązania zadań z poprzedniej lekcji, więc będziecie mogli ocenić swoje postępy. Zadania należy rozwiązywać za pomocą kalkulatora, a nie oprogramowania statystycznego, czy arkuszy kalkulacyjnych. Końcowe wyniki podajemy z dokładnością do czterech miejsc po przecinku.

Na końcu informatora znajdują się przykładowe zadania na Olimpiadę Biologiczną. Najlepiej jest je rozwiązać jako próbny arkusz egzaminacyjny po zapoznaniu się ze wszystkimi lekcjami.

v-ce przewodniczący KGOB

Łukasz Banasiak

Lekcja 1

Zadanie 1

Statystyka zajmuje się badaniem dużych populacji na podstawie małych prób. Na przykład chcąc dowiedzieć się czegoś o rozmiarze zooplanktonu *Daphnia* w jeziorach zarybionych i bezrybnych nie sposób jest złapać wszystkie osobniki. Badacz jest zmuszony pobrać próby planktonowe w obydwu jeziorach i na ich podstawie wyciągnąć wnioski o tym, czy są jakieś różnice, np. w długości ciała pomiędzy dwoma dużymi populacjami. Na początku praca statystyka polega na dobrym opisaniu pobranych prób. Poniżej znajdziecie dwa zbiory liczb zawierające pomiary długości ciała *Daphnia* z Czarnego Stawu (bezrybny) i Morskiego Oka (zasiedlony przez pstrąga) dla dwudziestoelementowych prób:

Czarny Staw [mm]	2,1	2,3	2,0	2,8	2,6	2,6	2,2	3,1	2,8	2,2
	2,5	3,6	2,7	1,9	2,4	1,5	1,8	1,5	2,9	2,5

Morskie Oko [mm]	1,7	0,8	1,5	1,8	1,8	2,0	1,0	1,6	2,0	1,5
	1,9	2,1	1,5	1,6	1,7	1,3	1,2	2,7	1,2	2,0

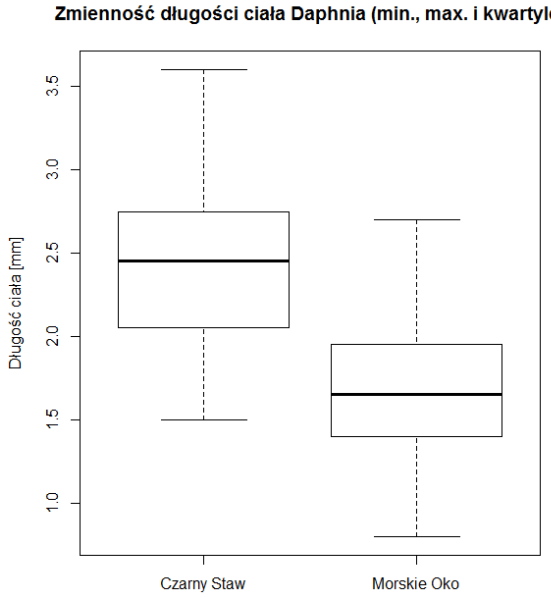
1. Policzcie dla obydwu zbiorów:
    a. średnią artymetyczną
    b. medianę
    c. kwartyle
    d. sumę odchyleń od średniej
    e. sumę kwadratów odchyleń od średniej
    f. wariancję i odchylenie standardowe

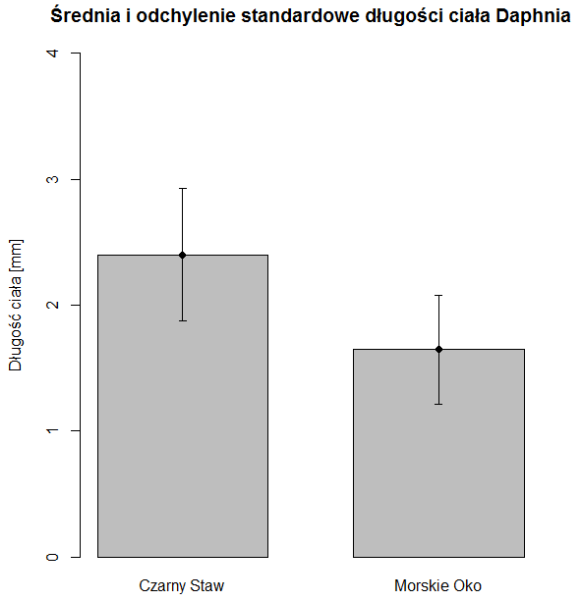
2. Porównajcie dwie grupy za pomocą dwu typów wykresów:
    a. Wykonajcie wykres pudełkowy (ang. *boxplot*) pokazujący w każdej grupie medianę, kwartyle oraz minimum i maksimum.
    b. Porównajcie średnie za pomocą wykresu słupkowego z naniesionymi wartościami odchylenia standardowego.

Jako literaturę polecam wstępne rozdziały podręcznika A. Łomnickiego „Wprowadzenie do statystyki dla przyrodników” oraz zasoby Internetu.

Rozwiązania do Zadania 1

Parametr	Czarny Staw	Morskie Oko
Średnia arytmetyczna	2,4000 mm	1,6450 mm
Mediana	2,4500 mm	1,6500 mm
Dolny kwartyl	2,0500 mm	1,4000 mm
Górny kwartyl	2,7500 mm	1,9500 mm
Suma odchyleń od średniej	0,0000 mm	0,0000 mm
Suma kwadratów odchyleń od średniej	5,2600 mm^2	3,5295 mm^2
Wariancja	0,2768 mm^2	0,1858 mm^2
Odchylenie standardowe	0,5262 mm	0,4310 mm





Komentarz do Zadania 1

Mediana jest nazywana obserwacją środkową. Najłatwiej ją policzyć, jeżeli najpierw uporządkujemy zbiór liczb. W przypadku nieparzystej liczby obserwacji jest to po prostu liczba dzieląca uporządkowany szereg na dwie połowy. Jeżeli mamy do czynienia z parzystą liczbą obserwacji, to medianę liczmy przez wyciągnięcie średniej arytmetycznej z dwu środkowych wartości, czyli w naszym przypadku z 10 i 11 liczby po uporządkowaniu.

Mediana dzieli zbiór obserwacji na dwie połowy, czyli poniżej i powyżej mediany znajduje się po 50% obserwacji. Kwartyle dzielą każdą z tych połówek znowu na pół, czyli ich policzenie sprowadza się do wyliczenia median dla pierwszych 50% i drugich 50% obserwacji. W naszym przypadku będą to średnie arytmetyczne z 5 i 6 oraz 15 i 16 liczby po uporządkowaniu, ponieważ mamy parzystą liczbę obserwacji.

Niestety istnieje kilka algorytmów liczenia kwartyli różniących się w szczegółach i sposób podany wyżej jest tylko jednym z nich, choć jest najczęściej przyjmowany na podstawowych kursach statystyki.

Odchylenie od średniej jest to różnica między wartością konkretniej obserwacji, a wartością średnią wyciągniętą ze wszystkich obserwacji w danej grupie. Odchylenie może mieć wartość ujemną, albo dodatnią, a suma wszystkich odchyleń od średniej jest zawsze równa zero (jeżeli wyszła Wam liczba zbliżona do zera, to znaczy, że musicie popracować nad dokładnością obliczeń na kalkulatorze). Z tego powodu jako miarę zmienności przyjmuje się sumę kwadratów odchyleń od średniej. Im jest ona większa tym generalnie obserwacje są bardziej rozproszone. Niestety ta miara zależy od wielkości próby i żeby móc porównywać zmienność w próbach różniących się liczebnością należy tę sumę kwadratów odchyleń od średniej podzielić przez liczebność próby; w naszym przypadku przez 20. Taka miara zmienności nazywa się wariancją, a jej pierwiastek kwadratowy odchyleniem standardowym. Im większa wariancja i odchylenie standardowe, tym większe rozproszenie obserwacji wokół średniej. Odchylenie standardowe w przeciwieństwie do wariancji ma tę zaletę, że jest podawane w tych samych jednostkach, co mierzona wielkość.

Wariancja i odchylenie standardowe policzone na podstawie n-elementowej próby w sposób podany powyżej to wartości systematycznie zaniżone w porównaniu z rzeczywistymi parametrami oznaczanymi odpowiednio jako σ^2 i σ występującymi w populacji generalnej o liczebności N.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N};\ \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

Innymi słowy spodziewamy się, że liczona w powyższy sposób wariancja z próby jest zaniżona w porównaniu z wariancją występującą w populacji, z której próba została pobrana. Dzieje się tak dlatego, że odchylenia liczymy nie od średniej z populacji generalnej, ale od średniej z próby, która jest tylko przybliżeniem tej pierwszej. Żeby tego uniknąć należy sumę kwadratów odchyleń podzielić przez liczebność próby pomniejszoną o jeden. W ten sposób otrzymujemy nieobciążony estymator wariancji w populacji generalnej, a po wyciągnięciu pierwiastka tzw. skorygowany wzór na odchylenie standardowe z próby:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1};\ s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Dzięki temu powtarzając wielokrotnie próbkowanie z danej populacji i obliczając wariancję wg podanego powyżej wzoru mamy pewność, że otrzymywane oszacowania wariancji oscylują wokół jej faktycznej wartości w populacji generalnej. Niestety nie można tego samego powiedzieć o odchyleniu standardowym, które w dalszym ciągu jest zaniżone, choć mniej niż przed wprowadzeniem korekty.

Zwykłe pokazanie średniej czy mediany oraz podanie liczebności próby nie rozstrzyga o tym, czy wartości średnie badanego parametru w populacji, z której próba została pobrana się różnią. W tym celu należy wykonać odpowiednia testy statystyczne, ale zgrubną ocenę można wykonać przedstawiając na wykresach oprócz mediany i wartości średniej zakresy, bądź miary zmienności. Medianę zwykle pokazuje się w towarzystwie minimum, dolnego kwartyla, górnego kwartyla i maksimum korzystając z wykresu pudełkowego, który dzieli zbiór obserwacji na cztery ćwiartki zawierające po tyle samo obserwacji. Średnią zwykle przedstawia się za pomocą wykresu słupkowego ± odchylenie standardowe pokazywane za pomocą wąsów. Jeżeli rozstępy międzykwartylowe (pudełka bez wąsów na wykresie pudełkowym), bądź średnie razem z odchyleniami standardowymi nie zachodzą na siebie, to można przypuszczać, że istnieje faktyczna różnica miedzy średnimi, albo medianami w populacjach generalnych. W naszym przypadku rozstępy międzykwartylowe mijają się, ale wąsy pokazujące odchylenia standardowe częściowo na siebie zachodzą.

Wykresy pudełkowe i słupkowe można robić w orientacji zarówno poziomej, jak i pionowej.

Jeżeli w zadaniu jest określona liczba miejsc po przecinku z jaką należy podawać wynik, to należy się tego bezwzględnie trzymać. Tzn. należy też podać wszystkie nieznaczące zera. Np. w naszym przypadku liczby całkowite miałyby zapis: 1,0000; 2,0000; 3,0000 … , bo we wstępie napisałem Wam, że wyniki podajemy z dokładnością do czterech miejsc po przecinku. Nie jest też dobrze, jeżeli wynik podajemy dokładniej, niż jest to zawarte w poleceniu do zadania. Należy jednak pamiętać, że obliczenia cząstkowe wykonujemy z jak największą dokładnością (współczesne kalkulatory zwykle liczą do dziewięciu miejsc po przecinku) i zaokrąglamy, stosując się do odpowiednich reguł, dopiero ostateczny wynik. Jeżeli w zadaniu nie ma podanej dokładności obliczeń, to należy przyjąć zasadę, że wyniki podajemy z dokładnością o jedno miejsce po przecinku większą niż były wykonywane pomiary.

Nie należy też zapominać o jednostkach. W naszym przypadku średnią i odchylenie standardowe należy podać w mm, ale wariancję w mm^2. Uczniowie bardzo często zapominają w ogóle o jednostkach!

Na koniec dla rozrywki możecie spróbować znaleźć błąd w następującym rozumowaniu: 1 zł × 1 zł = 1 zł, ale 100 gr × 100 gr = 10000 gr = 100 zł.[1]

[1] Mała podpowiedź 1 zł × 1 zł faktycznie równa się nie 1 zł, ale 1 $zł^2$ („złotówka kwatratowa”)

Lekcja 2

Zadanie 2

Praca statystyka nie kończy się na opisaniu prób. Kolejnym krokiem jest zwykle przetestowanie hipotezy, którą trzeba najpierw w formalny sposób postawić. W naszym przypadku chcemy się dowiedzieć, czy średnia długość ciała *Daphnia* jest w obydwu zbiornikach taka sama, czy może odmienna. W statystyce hipotezy, które podlegają testowaniu mówi się, że stawiane są nieco na opak – w taki sposób, żeby nie było żal ich odrzucić. My jesteśmy naturalnie jako biologowie zainteresowani znajdowaniem różnic pomiędzy środowiskami, a więc hipoteza, którą postawimy będzie brzmiała:

„Średnie długości ciała *Daphnia* w obydwu zbiornikach są sobie równe."

Hipotezę tę będziemy nazywać **hipotezą zerową** i oznaczać jako H_0. To, co może zrobić statystyka, to wyłącznie odrzucić tę hipotezę – nigdy nie może jej potwierdzić. Wykonując test statystyczny możemy dojść jedynie do jednego z dwu poniższych wniosków:

1. Są podstawy do odrzucenia hipotezy zerowej i przyjmujemy **hipotezę alternatywną** (H_A): „Średnie długości ciała *Daphnia* w dwu zbiornikach są różne." (Zaprzeczenie H_0)
2. Nie ma podstaw do odrzucenia hipotezy zerowej – zachowujmy *status quo*.

Na podstawie testu statystycznego **nigdy nie można** powiedzieć:

„Przyjmujemy hipotezę zerową, czyli średnie długości ciała *Daphnia* nie różnią się między zbiornikami"

To trochę jak w sądzie: nieskazany z braku dowodów niekoniecznie jest niewinny.

Testy statystyczne, niezależnie od ich typu, od strony obliczeniowej przebiegają zawsze według tego samego schematu. Najpierw należy policzyć tzw. **statystykę testu** po prostu podstawiając do wzoru dane znajdujące się w treści zadania. Następnym krokiem jest porównanie wartości obliczonej statystyki z tablicą, która zawiera **rozkład statystyki** w warunkach spełnionej hipotezy zerowej. Innymi słowy tablica zawiera informacje, jakie wartości statystyki testu są częste, a jakie rzadkie w przypadku kiedy hipoteza zerowa jest prawdziwa. Jeżeli z tablicy wynika, że obliczona wartość statystyki testu jest trudna do otrzymania, kiedy hipoteza zerowa zachodzi, to hipotezę tę odrzucamy.

Tablice *de facto* zawierają wartości prawdopodobieństwa otrzymania takiej bądź większej wartości statystyki testu przy założeniu, że hipoteza zerowa jest prawdziwa. Określają one zatem z jakim prawdopodobieństwem popełniamy **błąd pierwszego rodzaju** (ang. *type I error*), czyli odrzucamy hipotezę zerową, kiedy jest ona prawdziwa. To prawdopodobieństwo nazywamy **p-wartością** (ang. *p-value*). Zwykło się przyjmować, że jeżeli p-wartość jest mniejsza niż 0,05 to uznajemy wynik testu statystycznego za istotny i przyjmujemy hipotezę alternatywną. Tę wartość graniczną nazywamy poziomem istotności i oznaczamy go grecką literą alfa; zatem można powiedzieć, że zwykle przyjmuje się alfa = 0,05. Najczęściej z tablic nie poznamy dokładnej p-wartości, a przedział w jakim się ona znajduje.

Waszym zadaniem będzie przetestowanie postawionej we wstępie hipotezy zerowej za pomocą testu t-studenta dla prób niezależnych.

1. **Oblicz statystykę testu wg poniższego wzoru (dane weź z Zadania 1):**

$$t_0 = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$\bar{x}_i$ średnia z i-tej próby
n_i liczebność i-tej próby
s_i^2 wariancja i-tej próby

Wzór na odchylenie standardowe z próby:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

2. **Oblicz liczbę stopni swobody wg poniższego wzoru:**

$$df = n_1 + n_2 - 2$$

3. **Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej znajdź przedział w jakim znajduje się p-wartość.** Weź pod uwagę wartość bezwzględną statystyki testowej.

4. **Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,01.**

Tabela 1. Wartości krytyczne statystyki testowej.

Liczba stopni swobody	p-wartość w teście dwustronnym							
	0.2	**0.1**	**0.05**	**0.04**	**0.02**	**0.01**	**0.002**	**0.001**
1	3,07768	6,31375	12,7062	15,8945	31,8205	63,6568	318,306	636,627
2	1,88562	2,91999	4,30265	4,84873	6,96456	9,92484	22,3272	31,5990
3	1,63774	2,35336	3,18245	3,48191	4,54070	5,84091	10,2145	12,9240
4	1,53321	2,13185	2,77644	2,99853	3,74695	4,60409	7,17318	8,61031
5	1,47588	2,01505	2,57058	2,75651	3,36493	4,03214	5,89344	6,86884
6	1,43976	1,94318	2,44691	2,61224	3,14267	3,70743	5,20763	5,95880
7	1,41492	1,89458	2,36462	2,51675	2,99795	3,49948	4,78528	5,40787
8	1,39682	1,85955	2,30600	2,44898	2,89646	3,35539	4,50079	5,04130
9	1,38303	1,83311	2,26216	2,39844	2,82144	3,24984	4,29681	4,78092
10	1,37218	1,81246	2,22814	2,35931	2,76377	3,16927	4,14370	4,58691
11	1,36343	1,79588	2,20099	2,32814	2,71808	3,10581	4,02470	4,43697
12	1,35622	1,78229	2,17881	2,30272	2,68100	3,05454	3,92963	4,31779
13	1,35017	1,77093	2,16037	2,28160	2,65031	3,01228	3,85198	4,22083

14	1,34503	1,76131	2,14479	2,26378	2,62449	2,97684	3,78739	4,14045
15	1,34061	1,75305	2,13145	2,24854	2,60248	2,94671	3,73283	4,07276
16	1,33676	1,74588	2,11991	2,23536	2,58349	2,92078	3,68615	4,01500
17	1,33338	1,73961	2,10982	2,22385	2,56693	2,89823	3,64576	3,96512
18	1,33039	1,73406	2,10092	2,21370	2,55238	2,87844	3,61048	3,92164
19	1,32773	1,72913	2,09302	2,20470	2,53948	2,86094	3,57940	3,88341
20	1,32534	1,72472	2,08596	2,19666	2,52798	2,84534	3,55181	3,84952
21	1,32319	1,72074	2,07961	2,18943	2,51765	2,83136	3,52715	3,81927
22	1,32124	1,71714	2,07387	2,18289	2,50832	2,81876	3,50499	3,79214
23	1,31946	1,71387	2,06866	2,17696	2,49987	2,80734	3,48496	3,76762
24	1,31784	1,71088	2,06390	2,17154	2,49216	2,79694	3,46678	3,74539
25	1,31635	1,70814	2,05954	2,16659	2,48511	2,78744	3,45019	3,72514
26	1,31497	1,70562	2,05553	2,16203	2,47863	2,77871	3,43500	3,70660
27	1,31370	1,70329	2,05183	2,15783	2,47266	2,77068	3,42103	3,68959
28	1,31253	1,70113	2,04841	2,15393	2,46714	2,76326	3,40816	3,67391
29	1,31143	1,69913	2,04523	2,15033	2,46202	2,75639	3,39624	3,65941
30	1,31041	1,69726	2,04227	2,14697	2,45726	2,75000	3,38519	3,64596
31	1,30946	1,69552	2,03951	2,14383	2,45282	2,74404	3,37490	3,63345
32	1,30857	1,69389	2,03693	2,14090	2,44868	2,73848	3,36531	3,62180
33	1,30774	1,69236	2,03452	2,13816	2,44479	2,73328	3,35634	3,61091
34	1,30695	1,69092	2,03224	2,13558	2,44115	2,72840	3,34793	3,60072
35	1,30621	1,68957	2,03011	2,13316	2,43772	2,72381	3,34004	3,59115
36	1,30551	1,68830	2,02809	2,13087	2,43449	2,71948	3,33262	3,58215
37	1,30485	1,68709	2,02619	2,12871	2,43145	2,71541	3,32563	3,57367
38	1,30423	1,68595	2,02439	2,12667	2,42857	2,71156	3,31903	3,56568
39	1,30364	1,68488	2,02269	2,12474	2,42584	2,70791	3,31279	3,55811
40	1,30308	1,68385	2,02108	2,12291	2,42326	2,70446	3,30688	3,55096
41	1,30254	1,68288	2,01954	2,12117	2,42080	2,70118	3,30127	3,54418
42	1,30204	1,68195	2,01808	2,11952	2,41847	2,69807	3,29595	3,53774
43	1,30155	1,68107	2,01669	2,11794	2,41625	2,69510	3,29089	3,53162
44	1,30109	1,68023	2,01537	2,11644	2,41413	2,69228	3,28607	3,52580
45	1,30065	1,67943	2,01410	2,11500	2,41212	2,68959	3,28148	3,52025
46	1,30023	1,67866	2,01290	2,11364	2,41019	2,68701	3,27710	3,51496
47	1,29982	1,67793	2,01174	2,11233	2,40835	2,68456	3,27291	3,50990
48	1,29944	1,67722	2,01063	2,11107	2,40658	2,68220	3,26891	3,50507
49	1,29907	1,67655	2,00958	2,10987	2,40489	2,67995	3,26508	3,50045
50	1,29871	1,67590	2,00856	2,10872	2,40327	2,67779	3,26141	3,49601

Zadanie 3

Pewien mężczyzna dosyć często trafiał po pracy do pubu zamiast do domu. Żona oczywiście denerwowała się bardzo i mąż żeby być uczciwym pomyślał, że będzie postępował w następujący sposób. Będzie wychodził z pracy o losowej porze między 15 a 16 i szedł na przystanek, skąd odjeżdżają w tym samym kierunku dwa tramwaje. Jeden skręca po jakimś czasie do domu w lewo, a drugi do pubu w prawo. Po trzech miesiącach okazało się, że pomimo jego szczerych chęci do pubu jeździł trzy razy częściej niż do domu. **Jak to możliwe? Przyjmij założenie, że tramwaje jeżdżą z tą samą częstotliwością.**

Zadanie 4

Pewien uczony wykonał test statystyczny, w którym wykazał działanie przeciwbólowe pocieszaliny. Mówiąc językiem bardziej formalnym odrzucił hipotezę zerową mówiącą o tym, że pocieszalina działa tak samo jak placebo. Wiele lat później uczeni z konkurencyjnych zakładów farmaceutycznych powtórzyli wielokrotnie badania na większej próbie i nie udało im się odrzucić tak postawionej hipotezy zerowej i doszli do wniosku, że w oryginalnych badaniach popełniono błąd. **Którego rodzaju błąd (I, II czy III) popełnił uczony w pierwszym badaniu pocieszaliny? Podaj definicje błędów I, II i III rodzaju.**

Rozwiązania do zadań 2-4

$t_0 = 4{,}9643$
df = 38
p-wartość < 0,001
Odrzucamy H_0 bo p-wartość < alfa

Tramwaje jeżdżą z tą samą częstotliwością, ale ich rozkład jazdy jest przesunięty w fazie o ¼.

Uczony popełnił błąd I rodzaju.

Komentarz do zadań 2-4

Rzadko udaje się od razu uczniom otrzymać wartość statystyki testu z dokładnością do czterech miejsc po przecinku. Jeżeli jesteś w tej grupie to musisz pamiętać, że **nie należy zaokrąglać cząstkowych obliczeń!**

Liczba stopni swobody była bardzo prosta do policzenia, ale samo pojęcie stopni swobody (ang. *degrees of freedom*) jest dość skomplikowane, bo pojawia się w statystyce wielokrotnie, ale w różnych kontekstach. Najlepiej jest wyjaśnić to pojęcie na przykładach:

1. Jeżeli zmierzycie długość i szerokość jakiegoś zwierzęcia to można wyliczyć na tej podstawie stosunek długości do szerokości. Z koli znając ten stosunek i np. długość, można wyliczyć szerokość. Mamy zatem trzy liczby (obserwacje), ale tylko dwa stopnie swobody, bo trzecia obserwacja wynika z dwu pozostałych.
2. Podobnie jest ze zbiorem liczb i wartością średnią. Jeżeli znacie n – 1 liczb i wartość średnią wyciągniętą z n wartości, to ostatnią nieznaną liczbę można wyliczyć. Mówimy, że podczas obliczania średniej z n liczb, mamy n – 1 stopni swobody.

Odnalezienie zakresu w jakim znajduje się p-wartość nie sprawia z reguły problemów, ale zapis odpowiedzi często już tak. Możliwe są trzy notacje:

$$p < 0{,}001$$

$$p \in (0; 0{,}001)$$

$$0 < p < 0{,}001$$

Proszę pamiętajcie, że p-wartość to pewnego rodzaju prawdopodobieństwo, a więc w żadnym przypadku nie może wyjść poza przedział [0;1].

Należy odrzucić rozwiązania mówiące o tym, że częste wizyty w pubie to błąd losowy, bo mężczyzna w ciągu trzech miesięcy miał około 60 prób, więc oszacowanie częstości wizyt w pubie jest rzetelne. Odrzucamy także rozwiązania mówiące o czynniku ludzkim – jest to zagadka matematyczna. Kluczem do rozwiązania zagadki jest rozkład jazdy. Tramwaje jeździły z tą samą częstotliwością, ale ich okresy mijały się w fazie o około jedną czwartą. Np. obydwa jeździły co dwadzieścia minut, ale pierwszy odjeżdżał o 15:00, 15:20, 15:40 …, a drugi o 15:05, 15:25, 15:45 … Morał z tej zagadki jest przede wszystkim taki, że zdarzenia zupełnie losowe nie muszą być równo prawdopodobne!

Błąd I rodzaju to wynik fałszywie dodatni (ang. *False Positive* – litera P ma jedną pionową kreskę), czyli istotny statystycznie wynik testu, kiedy hipoteza zerowa jest faktycznie spełniona. P-wartość wprost określa prawdopodobieństwo popełnienia błędu pierwszego rodzaju, oraz jest to najniższy poziom istotności, na którym można odrzucić hipotezę zerową.

Błąd II rodzaju to wynik fałszywie ujemny (ang. *False Negative* – litera N ma dwie pionowe krestki), czyli nieistotny statystycznie wynik testu, kiedy hipoteza zerowa faktycznie nie jest spełniona. Prawdopodobieństwo popełnienia tego błędu z reguły nie jest dokładnie znane. Testy, które mają małe prawdopodobieństwo popełnienia tego błędu mówimy, że mają dużą moc, lub że są czułe. Testy o małej mocy potrzebują większych liczebności prób, żeby wykryć odstępstwa od hipotezy zerowej (wykluczyć, że różnice między próbami wynikają jedynie z błędu losowego przy próbkowaniu).

Błąd III rodzaju polega na przyjęciu złej hipotezy alternatywnej, jeżeli jest więcej niż jedna, już po odrzuceniu faktycznie nieprawdziwej hipotezy zerowej.

Lekcja 3

Zadanie 5

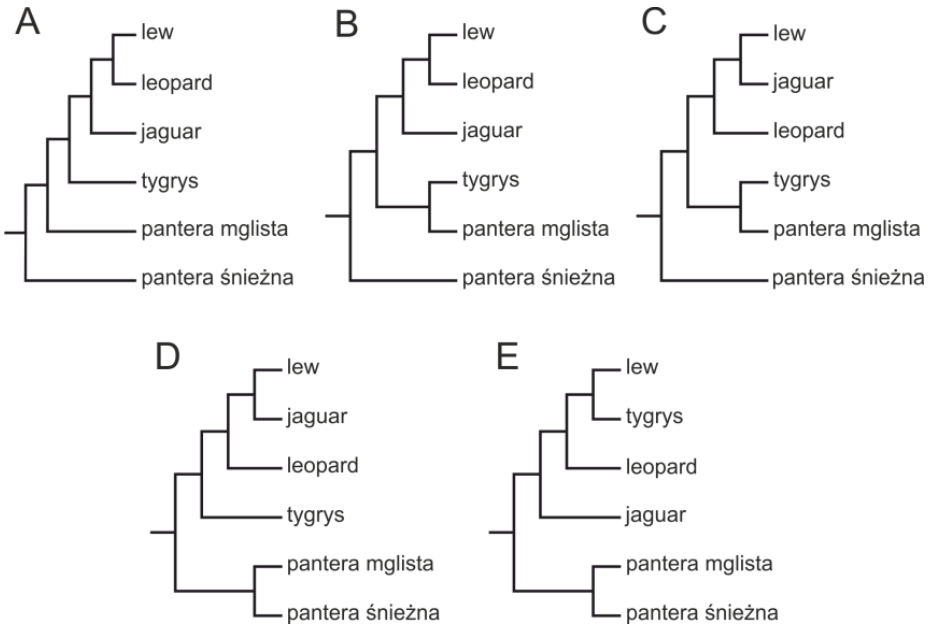
Odrzucenie hipotezy zerowej w Zadaniu 2 kończy pracę statystyka, ale rozpoczyna pracę biologa. Biorąc pod uwagę, że Morskie Oko jest zarybione (pstrąg), a Czarny Staw bezrybny na podstawie własnych przemyśleń i literatury wyjaśnij, skąd bierze się ta różnica.

Zadanie 6

Zanim powrócimy do zadań statystycznych rozpoczniemy wątek filogenetyczny. Na początek chciałbym, żebyście nieco dokładniej rozwiązali jedno z zadań z testu finałowego sprzed lat.

Poniższa tabela zawiera dane o obecności (+), lub braku (-) siedmiu cech morfologicznych dla sześciu gatunków kotów z rodzaju *Panthera*. Dla każdej z siedmiu cech (I-VII) oraz każdej z pięciu hipotez filogenetycznych w postaci drzew (A-E) oblicz minimalną liczbę zmian ewolucyjnych. Wyniki wpisz w odpowiednie miejsce tabeli.

	I	II	III	IV	V	VI	VII
jaguar	-	+	+	-	+	+	-
leopard	+	+	+	-	+	+	+
lew	+	+	+	+	+	+	+
tygrys	+	-	+	-	+	-	-
pantera mglista	-	-	+	-	-	-	-
pantera śnieżna	-	-	-	+	-	-	-



Minimalna liczba zmian dla każdej z cech w zależności od przyjętej hipotezy.

Cecha/drzewo	A	B	C	D	E
I					
II					
III					
IV					
V					
VI					
VII					
Suma					

Zadanie 7

Relacje pokrewieństwa między organizmami z reguły przedstawia się za pomocą dychotomicznych drzew. Mogą one być zakorzenione, bądź nie. Ile jest możliwych niezakorzenionych topologii przedstawiających różne stosunki pokrewieństwa pomiędzy czterema taksonami A-D? Ile jest możliwych zakorzenionych topologii przedstawiających różne stosunki pokrewieństwa pomiędzy trzema taksonami A-C? Narysuj te topologie.

Rozwiązania i komentarz do zadań 5-7

Pstrągi są drapieżnikami polującymi na *Daphnia* za pomocą wzroku. Mniejsze ciało ma swoje oczywiste negatywne skutki choćby w postaci mniejszych możliwości filtracji glonów, którymi się odżywiają, ale pozwala *Daphnia* unikać presji drapieżnika, który poluje za pomocą wzroku.

Cecha/drzewo	A	B	C	D	E
I	2	2	3	2	1
II	1	1	1	1	2
III	1	1	1	1	1
IV	2	2	2	2	2
V	1	2	2	1	1
VI	1	1	1	1	2
VII	1	1	2	2	2
Suma	9	10	12	10	11

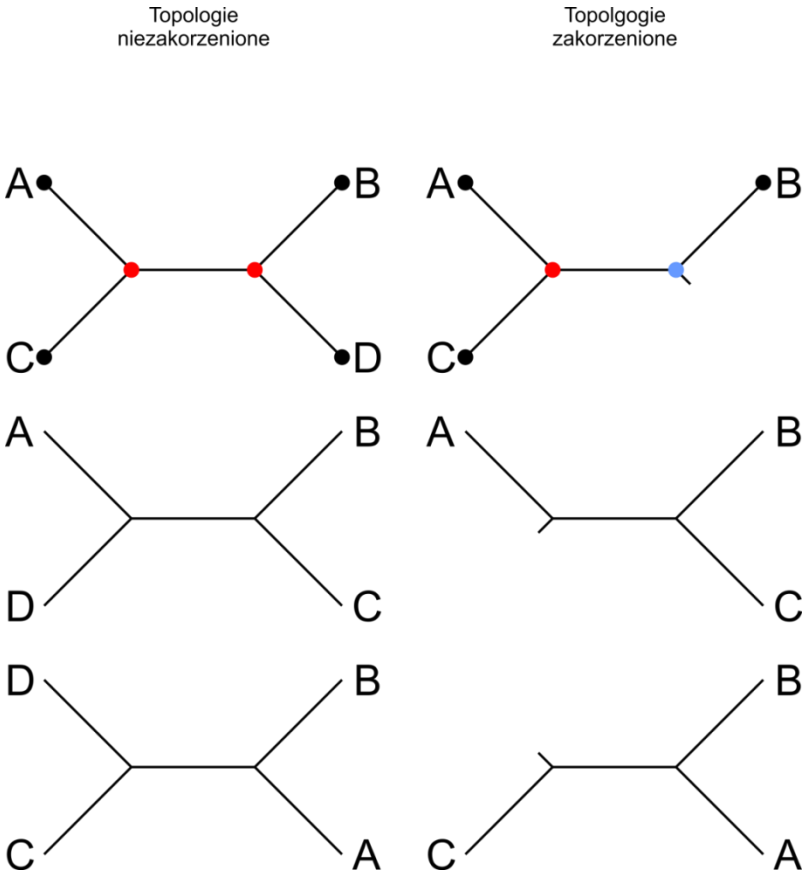
Zadania tego typu są dosyć żmudne, ale pewne obliczenia można skrócić. Cecha I wymaga faktycznie uważnego prześledzenia jej możliwych historii na drzewie i niewiele można ten proces przyspieszyć. Cecha II zaznaczona na zielono jest podobnie trudna do prześledzenia, ale warto zauważyć, że dokładnie w ten sam sposób jest zakodowana cecha VI, co znacznie przyspieszy rozwiązanie zadania. Osoby, które mają rożne wyniki dla cechy II i VI mają zatem oczywisty błąd! Cecha III zaznaczona na żółto jest autapomorficzna, tzn. wszyscy przedstawiciele rodzaju *Panthera* oprócz jednego mają ten sam stan cechy. W takim przypadku niezależnie od topologii **zawsze** na drzewie będzie tylko jedna zmiana. Cechy IV, V i VII (zaznaczone na czerwono) są do siebie podobne z tego powodu, że wszystkie oprócz dwu taksonów mają ten sam stan cechy. W takim przypadku minimalna liczba zmian może przyjmować tylko dwie wartości: 1 jeżeli dwa odmienne taksony są siostrami, lub tworzą dwa kolejne odgałęzienia od głównego pnia, lub 2 jeżeli dwa odmienne taksony nie są ze sobą bezpośrednio spokrewnione. Nie należy czynić żadnych założeń o stanie cechy u korzenia drzewa, jeżeli nie ma o tym mowy wprost w zadaniu. Przy skomplikowanych topologiach i układach cech często się zdarza, że jest kilka sposobów na zrekonstruowanie ewolucji danej cechy, które minimalizują liczbę zmian.

W oryginalnym zadaniu na teście finałowym trzeba było policzyć sumaryczną liczbę zmian na każdym drzewie, żeby wybrać najbardziej oszczędne, bądź mówiąc naukowym żargonem „najkrótsze”. W tym przypadku jest to drzewo oznaczone literą A i jego długość to 9 kroków. Metoda szacowania filogenezy polegająca na wyborze „najkrótszego” drzewa spośród wszystkich możliwych topologii nazywa się metodą największej parsymonii. W naszym przypadku do obliczeń można było z góry pominąć cechy III (autapomorficzna) i VI (taka sama jak II).

W najprostszym wariancie metody największej parsymonii kodowanie cech jest arbitralne i nie są one uporządkowane. Tzn. cechy można kodować za pomocą różnych znaków, ale najczęściej są to kolejne liczby całkowite poczynając od zera, kolejne litery alfabetu, bądź znaki „+” i „–”, ale wszystkie symbole są równoważne i oznaczają po prostu odmienny stan cechy. Nie zakładamy też żadnego kierunku ewolucji – każda cecha może zmienić się bezpośrednio w każdą inną i nie stawiamy żadnych założeń o tym, jaki stan cechy występował u wspólnego przodka wszystkich taksonów. Zatem jedynki nie są lepsze od zer, ani plusy od minusów, a symbole użyte do kodowania można zupełnie dowolnie zmieniać i nie ma to żadnego wpływu na analizę.

Drzewo filogenetyczne z formalnego punktu widzenia można rozumieć jako rodzaj grafu złożonego z wierzchołków i krawędzi ich łączących. Krawędzie zwyczajowo nazywa się gałęziami. Wierzchołki z kolei dzielimy ze względu na ich stopień, czyli liczbę krawędzi z nich wychodzących. Wierzchołek n-tego stopnia posiada n wychodzących krawędzi. W ten sposób można dla dychotomicznego drzewa podać definicję liści (taksonów) – wierzchołki o stopniu 1 oraz węzłów wewnętrznych, czyli przodków – wierzchołki o stopniu 3. W drzewie zakorzenionym występuje dodatkowo jeden wierzchołek o stopniu 2, nazywany korzeniem, oznaczany zazwyczaj w postaci krótkiej kreski i wskazujący początek procesu ewolucji.

W Zadaniu 8 możliwe są trzy dychotomiczne topologie niezakorzenione i trzy zakorzenione (dla przykładu dla pierwszych drzew z każdej kategorii na czarno zaznaczono węzły stopnia 1, na czerwono węzły stopnia 3, a na niebiesko węzeł stopnia 2). Jest jeszcze jedna zakorzeniona topologia, gdzie wszystkie trzy taksony A–C + korzeń wychodzą z jednego węzła, ale nie spełnia ona warunku dychotomicznego drzewa – takie drzewo zawiera tzw. politomię. Pamiętajmy, że drzewo korzenimy przez „złamanie” jednej z gałęzi, a nie przez wybór któregoś z taksonów.



Lekcja 4

Zadanie 8

Metoda największej parsymonii dostarcza kryterium wyboru drzewa. Innymi słowy przeszukiwaliśmy pewien zbiór drzew i badaliśmy długość każdego z nich, żeby wybrać najkrótsze. Inną metodą służącą do rekonstrukcji filogenezy jest metoda UPGMA, która ma zupełnie odmienną zasadę działania – opiera się na wykorzystaniu pewnego algorytmu grupowania taksonów. Macierz zakodowanych cech organizmów najpierw przekształcamy w macierz odległości. Na przykład jeżeli badaliśmy DNA to macierzą cech będzie tablica z sekwencjami fragmentu określonego genu (markera molekularnego). Będzie miała ona tyle wierszy, ile taksonów uwzględniliśmy w badaniach i tyle kolumn ile, nukleotydów długości liczył badany odcinek DNA, a dozwolonymi wpisami do tej tablicy będą cztery stany: „A”, „T”, „C” i „G”. Odległości genetyczne między organizmami na podstawie takiej macierzy można policzyć na wiele różnych sposobów, ale najprostszy polega na określeniu liczby różniących się nukleotydów między każdą parą sekwencji. Poniżej znajdziecie przykład macierzy cech molekularnych i odpowiadającą jej macierz odległości:

Takson A	A	T	C	C	A	T	G	G	A	A
Takson B	A	T	A	C	A	T	G	C	C	C
Takson C	C	T	A	C	A	T	G	C	C	C
Takson D	A	T	C	C	A	T	G	G	C	C

	A	B	C	D
A				
B	4			
C	5	1		
D	2	2	3	

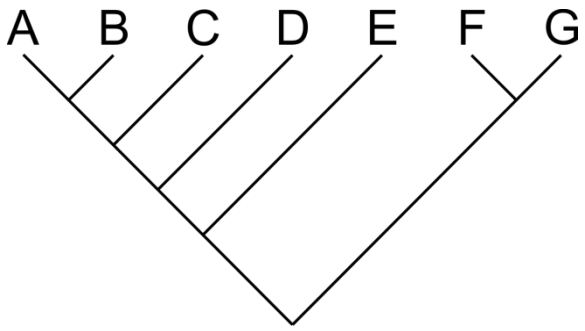
Następnym krokiem jest wykonanie algorytmu, który zbuduje drzewo w oparciu o macierz odległości. Doskonały jego opis (niestety w j. angielskim) znajdziecie na następującej stronie internetowej:

http://homepages.ulb.ac.be/~dgonze/TEACHING/phylo_trees.pdf

Korzystając z tego opisu narysuj drzewo filogenetyczne dla powyższych czterech taksonów uwzględniając oprócz topologii także długości gałęzi. Przedstaw także cząstkowe obliczenia.

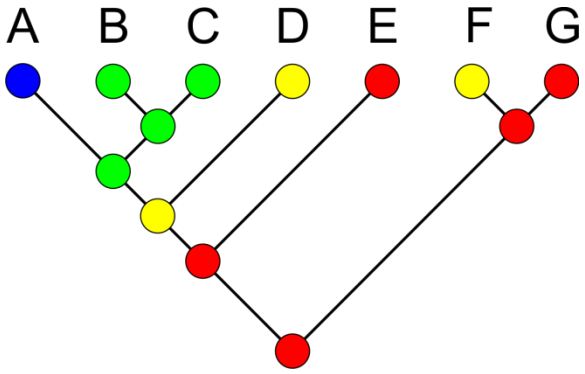
Zadanie 9

Czy grupa obejmująca taksony C i E na przedstawionym poniżej drzewie filogenetycznym jest monofiletyczna, parafiletyczna, czy polifiletyczna? Odpowiedź uzasadnij.



Zadanie 10

Poniżej znajdziesz rekonstrukcję ewolucji czterostanowej cechy zmapowaną na drzewie filogenetycznym. Poszczególne stany tej cechy są zaznaczone kolorami czerwonym, niebieskim, zielonym i żółtym. Wskaż plezjomorfie, autapomorfie, synapomorfie i homoplazje. Odpowiedź uzasadnij.



Zadanie 11

Czas powrócić do statystyki. Tym razem tematem będzie test chi-kwadrat. Służy on do sprawdzania zgodności wartości oczekiwanych z wartościami obserwowanymi. Np. rzucając kostką sześciościenną 600 razy i przyjmując za hipotezę zerową, że kostka jest dobrze wyważona będziemy oczekiwali takiego wyniku, że wypadnie 100 razy jedynka, 100 razy dwójka, itd. Oczywiście jest to bardzo mało prawdopodobne, nawet jeżeli kostka jest wyważona idealnie, że otrzymamy dokładnie taki rezultat i powstaje pytanie, czy różnice miedzy wartościami oczekiwanymi i obserwowanymi są tylko błędem losowym, czy może kostka jednak posiada wadę (hipoteza zerowa nie jest spełniona).

W biologii bardzo często stosuje się jeden z wariantów testu chi-kwadrat, żeby sprawdzić czy jakieś cechy organizmów współwystępują częściej niżby wynikało to z czystego przypadku (hipoteza zerowa). Możemy na przykład zadać pytanie, czy owłosione liście spotykamy istotnie częściej u roślin, które mają parzące łodygi? Hipoteza zerowa formalnie jest skonstruowana w następujący sposób:

Nie ma związku między obecnością cech.

Żeby tę hipotezę przetestować najpierw należy zbadać rośliny, a wyniki wygodnie jest wpisać do tzw. tabeli liczności:

		liście	
		owłosione	nagie
łodygi	parzące	48	36
	nieparzące	53	13

Z tej tabeli można np. odczytać, że spośród 150 przebadanych roślin 48 miało jednocześnie parzące łodygi i owłosione liście i jest to wartość obserwowana. Kolejnym krokiem jest wyliczenie wartości oczekiwanych. W tym celu dobrze jest najpierw policzyć tzw. sumy brzegowe i sumę całkowitą:

		liście		
		owłosione	nagie	suma
łodygi	parzące	48	36	**84**
	nieparzące	53	13	**66**
	suma	**101**	**49**	**150**

Wiemy teraz, że liście owłosione posiada 101 roślin, łodygi parzące 84 rośliny. Prawdopodobieństwo, że wylosujemy z tego zbioru roślinę o liściach owłosionych wynosi więc 101/150, a że wylosujemy roślinę o łodygach parzących 84/150. Przy założeniu, że hipoteza zerowa jest prawdziwa (cechy są niezależne) prawdopodobieństwo obecności tych dwóch cech jednocześnie można wyliczyć mnożąc ich indywidualne prawdopodobieństwa przez siebie, co daje wartość 0,3770667. Oznacza to, iż oczekujemy, że około 38% roślin będzie miało jednocześnie owłosione liście i parzące łodygi. Mając w sumie 150 roślin oznacza to, że oczekujemy 56,56 osobników z taką kombinacją cech:

		liście	
		owłosione	nagie
łodygi	parzące	56,5600	?
	nieparzące	?	?

PRZED ROZPOCZĘCIEM ROZWIĄZYWANIA PODPUNKTÓW ZNAJDUJĄCYCH SIĘ NA NASTĘPNEJ STRONIE WŁĄCZ STOPER! KIEDY SKOŃCZYSZ, ZAPISZ ILE CZASU ZAJĘŁO CI ROZWIĄZANIE ZADANIA.

DO MIERZONEGO CZASU WLICZA SIĘ TAKŻE PRZEPISANIE ODPOWIEDZI NA CZYSTO!

NIE CZYTAJ KOLEJNEJ STRONY DOPÓKI SIĘ NIE UPEWNISZ, ŻE ZROZUMIAŁAŚ/EŚ WSTĘP DO ZADANIA.

1. **Wylicz pozostałe wartości oczekiwane i wpisz je do tabeli.**
2. **Oblicz statystykę testu wg poniższego wzoru:**

$$\chi^2 = \sum_{i=1}^{n} \frac{(E_i - T_i)^2}{T_i}$$

n – liczba pól (grup) w tabeli
E_i – liczebność obserwowana
T_i – liczebność oczekiwana

3. **Oblicz liczbę stopni swobody wg poniższego wzoru:**

$$df = (k-1)(w-1)$$

k – liczba kolumn w tabeli
w – liczba wierszy w tabeli

4. **Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej znajdź przedział w jakim znajduje się p-wartość.**

5. **Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,01.**

6. **Podaj czas w jakim rozwiązałaś/eś zadanie.**

Tabela 1. Wartości krytyczne statystyki testowej.

Liczba stopni swobody	p-wartość							
	0,15	**0,1**	**0,05**	**0,025**	**0,02**	**0,01**	**0,005**	**0,001**
1	2,07225	2,70554	3,84146	5,02389	5,41189	6,63490	7,87944	10,8276
2	3,79424	4,60517	5,99146	7,37776	7,82405	9,21034	10,5966	13,8155
3	5,31705	6,25139	7,81473	9,34840	9,83741	11,3449	12,8382	16,2663
4	6,74488	7,77944	9,48773	11,1433	11,6678	13,2767	14,8603	18,4668
5	8,11520	9,23636	11,0705	12,8325	13,3882	15,0863	16,7496	20,5150
6	9,44610	10,6446	12,5916	14,4494	15,0332	16,8119	18,5476	22,4578
7	10,7479	12,0170	14,0671	16,0128	16,6224	18,4753	20,2777	24,3219
8	12,0271	13,3616	15,5073	17,5345	18,1682	20,0902	21,9550	26,1245
9	13,2880	14,6837	16,9190	19,0228	19,6790	21,6660	23,5893	27,8772
10	14,5339	15,9872	18,3070	20,4832	21,1608	23,2093	25,1882	29,5883
11	15,7671	17,2750	19,6751	21,9201	22,6179	24,7250	26,7569	31,2641
12	16,9893	18,5493	21,0261	23,3367	24,0540	26,2170	28,2995	32,9095
13	18,2020	19,8119	22,3620	24,7356	25,4715	27,6882	29,8195	34,5282
14	19,4062	21,0641	23,6848	26,1189	26,8728	29,1412	31,3194	36,1232
15	20,6030	22,3071	24,9958	27,4884	28,2595	30,5779	32,8013	37,6973
16	21,7931	23,5418	26,2962	28,8454	29,6332	31,9999	34,2672	39,2523
17	22,9770	24,7690	27,5871	30,1910	30,9950	33,4087	35,7185	40,7902

18	24,1555	25,9894	28,8693	31,5264	32,3462	34,8053	37,1565	42,3124
19	25,3289	27,2036	30,1435	32,8523	33,6874	36,1909	38,5823	43,8202
20	26,4976	28,4120	31,4104	34,1696	35,0196	37,5662	39,9968	45,3147
21	27,6620	29,6151	32,6706	35,4789	36,3434	38,9322	41,4011	46,7970
22	28,8225	30,8133	33,9244	36,7807	37,6595	40,2894	42,7957	48,2679
23	29,9792	32,0069	35,1725	38,0756	38,9683	41,6384	44,1813	49,7282
24	31,1325	33,1962	36,4150	39,3641	40,2704	42,9798	45,5585	51,1786
25	32,2825	34,3816	37,6525	40,6465	41,5661	44,3141	46,9279	52,6197
26	33,4295	35,5632	38,8851	41,9232	42,8558	45,6417	48,2899	54,0520
27	34,5736	36,7412	40,1133	43,1945	44,1400	46,9629	49,6449	55,4761
28	35,7150	37,9159	41,3371	44,4608	45,4188	48,2782	50,9934	56,8923
29	36,8538	39,0875	42,5570	45,7223	46,6927	49,5879	52,3356	58,3012
30	37,9903	40,2560	43,7730	46,9792	47,9618	50,8922	53,6720	59,7031

Rozwiązania i komentarz do zadań 8-11

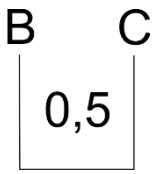
Zadanie 8

Algorytm UPGMA na początku może wydawać się żmudny i na pewno wymaga wielokrotnego przećwiczenia. Poniżej rozwiązanie zadania.

1. Wyjściowa macierz odległości:

	A	B	C	D
A				
B	4			
C	5	1		
D	2	2	3	

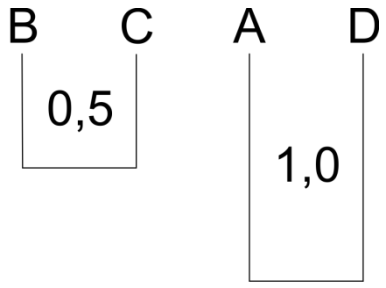
2. Najbliżej siebie są taksony B i C, które łączymy gałęzią o długości 1, którą łamiemy dokładnie w środku jej długości:



3. Następnie na podstawie wyjściowej macierzy przeliczamy odległości między taksonami A, D i klastrem BC: (2 + 3) / 2 oraz (4 + 5) / 2

	A	BC	D
A			
BC	4,5		
D	2	2,5	

4. Teraz najbliżej siebie są taksony A i D więc łączymy je gałęzią o długości 2, którą znów łamiemy w połowie długości:



5. Następnie znów na podstawie wyjściowej macierzy musimy przeliczyć odległości miedzy klastrami BC i AD: (2 + 3 + 4 + 5) / 4

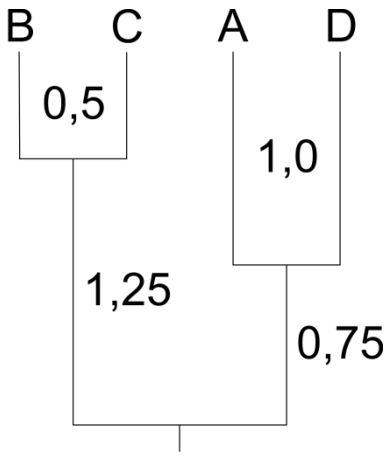
	AD	BC
AD		
BC	3,5	

Prostsze algorytmy w tym miejscu nie odwoływały by się do wyjściowej macierzy, ale do tej z punktu 3 i obliczenia wyglądałyby następująco:

$$(4,5 + 2,5) / 2 = 3,5$$

W tym szczególnym wypadku wynik jest ten sam, ale w ogólności ta równość nie zachodzi!

6. Łączymy klaster AD i BC gałęzią o długości 3,5 złamaną w połowie, co daje 1,75 licząc od liści do korzenia:



Zwróćcie uwagę na dwie istotne charakterystyki drzewa UPGMA. Po pierwsze zawsze jest to drzewo zakorzenione, tzn. miejsce złamania gałęzi w ostatnim kroku wskazuje pozycję korzenia i trzeba go koniecznie zaznaczyć na rysunku. Po drugie jest to zawsze drzewo ultrametryczne, czyli odległość od korzenia do każdego z liści jest taka sama – w naszym przypadku 1,75. Jeżeli podczas obliczeń wyjdzie wam drzewo pozbawione tych charakterystyk, to jest to na pewno wynik błędny.

Zadanie 9

To zadanie nie ma prostej odpowiedzi. Grupa złożona z taksonów C i E na pewno nie jest monofiletyczna. Taksony w takim przypadku musiałyby pochodzić od jednego wspólnego przodka i obejmować jego wszystkich współczesnych potomków. Grupa polifiletyczna to każda grupa nie będąca grupą monofiletyczną, a więc w ogólności tak można nazwać grupę złożoną z taksonów C i E. Problem polega na tym, że grupa parafiletyczna to szczególny przypadek grupy polifiletycznej, który nie ma do końca jasnej definicji. Mówi się, że to grupa pochodząca od jednego wspólnego przodka, ale nieobejmująca wszystkich jego współczesnych potomków. W tym kontekście grupa polifiletyczna zawiera organizmy pochodzące od różnych przodków, czyli wywodzące się z różnych linii filogenetycznych. Jednak wszystkie organizmy mają jednego wspólnego pradawnego przodka i rozumując w ten sposób każda grupa polifiletyczna staje się grupą parafiletyczną. Z tego powodu narzuca się ograniczenie, że grupa parafiletyczna ma obejmować wszystkich potomków wspólnego przodka z wyłączeniem małej liczby niewielkich grup monofiletycznych. Nie wiadomo jednak jak małe i jak mało ma być tych wyłączonych grup. Dlatego można uznać, że wyłączenie taksonów A i B oraz D z grupy ABCDE tworzy grupę parafiletyczną, lub uznać grupę CE za polifiletyczną w zupełnie ogólnym znaczeniu.

Zadanie 10

Stan cechy zaznaczony na czerwono jest stanem wyjściowym (ancestralnym) i dlatego nazywamy go plezjomorfią. Stan cechy zaznaczony na żółto powstał dwa razy niezależnie – jest to konwergencja. Jeżeli stan danej cechy ulega konwergencji lub rewersji mówimy o homoplazji. Stan cechy zaznaczony na zielono to synapomorfia, czyli cecha ewolucyjnie nowa, ale charakterystyczna dla monofiletycznego kladu – w tym przypadku BC. Stan cechy zaznaczony na niebiesko to autapomorfia, czyli cecha ewolucyjnie nowa, która jest obecna tylko u jednego z przedstawicieli.

Synapomorfie są bardzo pożądane podczas odtwarzania filogenezy i są dobrymi cechami diagnostycznymi podczas opracowywania naturalnych kluczy. Homoplazje to inaczej szum filogenetyczny, który zaciera relacje pokrewieństwa podczas szacowania relacji pokrewieństwa.

Zadanie 11

1. Wartości oczekiwane:

		liście	
		owłosione	nagie
łodygi	parzące	56,5600	**27,4400**
	nieparzące	**44,4400**	**21,5600**

2. Statystyka testu chi-kwadrat – **9,0132**
3. Liczba stopni swobody – **1**
4. **0,001 < p < 0,005** dokładna p-wartość wyliczona przez odpowiednie oprogramowanie wynosi 0,00268
5. **Odrzucamy hipotezę zerową** bo p < alfa
6. Rozwiązanie zadania nie powinno trwać więcej niż **10 min**.

PRZED ROZPOCZĘCIEM ROZWIĄZYWANIA ZADAŃ **12** i **16** ZNAJDUJĄCYCH SIĘ NA NASTĘPNYCH STRONACH WŁĄCZ STOPER! KIEDY SKOŃCZYSZ, ZAPISZ ILE CZASU ZAJĘŁO CI ROZWIĄZANIE KAŻDEGO ZADANIA.

DO MIERZONEGO CZASU WLICZA SIĘ TAKŻE PRZEPISANIE ODPOWIEDZI NA CZYSTO ORAZ ZAPOZNANIE SIĘ ZE WSTĘPEM!

Lekcja 5

Zadanie 12

Wróćmy na chwilę do Zadania 2. Naszym celem było przetestowanie hipotezy zakładającej, że *Daphnia* w obydwu zbiornikach mają tę samą średnią długość ciała. W tym układzie doświadczalnym mieliśmy dwie niezależne grupy organizmów, które mierzyliśmy. Ten schemat postępowania sprawdził się w przypadku *Daphnia*, ale na przykład podczas testowania działania leków na nadciśnienie tętnicze byłby niezbyt dobry. Zamiast uwzględnić w badaniach dwie grupy ludzi, z których jedna bierze lek, a druga placebo, dużo lepiej jest mieć jedną grupę pacjentów, która będzie miała zmierzony poziom ciśnienia tętniczego przed i po kuracji badanym lekiem. Zebrane dane wyglądają na pierwszy rzut oka całkiem podobnie jak w przypadku *Daphnia*, bo są to dwa zbiory liczb, ale należy pamiętać, że są one powiązane w pary. Z tego powodu, żeby sprawdzić, czy lek działa, czy nie należy zastosować nieco inną wersję testu t-studenta dla prób zależnych (dla par). Bada on średnią wartość przyrostu/spadku wartości badanego parametru po zastosowaniu zabiegu.

Twoim zadaniem będzie sprawdzenie czy badany farmaceutyk skutecznie obniża ciśnienie tętnicze krwi.

1. Poniższa tabela zawiera informacje o ciśnieniu skurczowym pacjentów przed i po trzytygodniowej kuracji badanym lekiem. **Oblicz różnicę wartości skurczowego ciśnienia tętniczego krwi po i przed kuracją dla każdego pacjenta. Wyniki zapisz w tabeli.**

ID pacjenta	Ciśnienie skurczowe przed kuracją a [mm Hg]	Ciśnienie skurczowe po kuracji b [mm Hg]	Różnica $x = b - a$
1	160	159	
2	151	160	
3	158	156	
4	161	155	
5	161	170	
6	163	155	
7	153	163	
8	159	158	
9	163	160	
10	158	157	
11	162	165	
12	164	163	
13	158	163	

2. **Sformułuj hipotezę zerową.**

3. **Oblicz statystykę testu wg poniższego wzoru:**

$$t_0 = \frac{\bar{x}}{s/\sqrt{n}}$$

$\bar{x}$ – średnia różnica między wartością ciśnienia po i przed kuracją
n – liczba pacjentów

s – odchylenie standardowe różnicy miedzy wartością ciśnienia po i przed kuracją
Wzór na odchylenie standardowe z próby:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

4. **Oblicz liczbę stopni swobody wg poniższego wzoru:**

$$df = n - 1$$

5. **Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej znajdź przedział w jakim znajduje się p-wartość.** Weź pod uwagę wartość bezwzględną statystyki testowej.

6. **Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,05.**

7. **Zinterpretuj wynik testu statystycznego – określ, czy można powiedzieć, że badany lek skutecznie obniża ciśnienie krwi.**

Tabela 1. Wartości krytyczne statystyki testowej.

Liczba stopni swobody	p-wartość w teście dwustronnym							
	0.2	**0.1**	**0.05**	**0.04**	**0.02**	**0.01**	**0.002**	**0.001**
1	3,07768	6,31375	12,7062	15,8945	31,8205	63,6568	318,306	636,627
2	1,88562	2,91999	4,30265	4,84873	6,96456	9,92484	22,3272	31,5990
3	1,63774	2,35336	3,18245	3,48191	4,54070	5,84091	10,2145	12,9240
4	1,53321	2,13185	2,77644	2,99853	3,74695	4,60409	7,17318	8,61031
5	1,47588	2,01505	2,57058	2,75651	3,36493	4,03214	5,89344	6,86884
6	1,43976	1,94318	2,44691	2,61224	3,14267	3,70743	5,20763	5,95880
7	1,41492	1,89458	2,36462	2,51675	2,99795	3,49948	4,78528	5,40787
8	1,39682	1,85955	2,30600	2,44898	2,89646	3,35539	4,50079	5,04130
9	1,38303	1,83311	2,26216	2,39844	2,82144	3,24984	4,29681	4,78092
10	1,37218	1,81246	2,22814	2,35931	2,76377	3,16927	4,14370	4,58691
11	1,36343	1,79588	2,20099	2,32814	2,71808	3,10581	4,02470	4,43697
12	1,35622	1,78229	2,17881	2,30272	2,68100	3,05454	3,92963	4,31779
13	1,35017	1,77093	2,16037	2,28160	2,65031	3,01228	3,85198	4,22083
14	1,34503	1,76131	2,14479	2,26378	2,62449	2,97684	3,78739	4,14045
15	1,34061	1,75305	2,13145	2,24854	2,60248	2,94671	3,73283	4,07276
16	1,33676	1,74588	2,11991	2,23536	2,58349	2,92078	3,68615	4,01500
17	1,33338	1,73961	2,10982	2,22385	2,56693	2,89823	3,64576	3,96512
18	1,33039	1,73406	2,10092	2,21370	2,55238	2,87844	3,61048	3,92164
19	1,32773	1,72913	2,09302	2,20470	2,53948	2,86094	3,57940	3,88341
20	1,32534	1,72472	2,08596	2,19666	2,52798	2,84534	3,55181	3,84952
21	1,32319	1,72074	2,07961	2,18943	2,51765	2,83136	3,52715	3,81927
22	1,32124	1,71714	2,07387	2,18289	2,50832	2,81876	3,50499	3,79214
23	1,31946	1,71387	2,06866	2,17696	2,49987	2,80734	3,48496	3,76762

24	1,31784	1,71088	2,06390	2,17154	2,49216	2,79694	3,46678	3,74539
25	1,31635	1,70814	2,05954	2,16659	2,48511	2,78744	3,45019	3,72514
26	1,31497	1,70562	2,05553	2,16203	2,47863	2,77871	3,43500	3,70660
27	1,31370	1,70329	2,05183	2,15783	2,47266	2,77068	3,42103	3,68959
28	1,31253	1,70113	2,04841	2,15393	2,46714	2,76326	3,40816	3,67391
29	1,31143	1,69913	2,04523	2,15033	2,46202	2,75639	3,39624	3,65941
30	1,31041	1,69726	2,04227	2,14697	2,45726	2,75000	3,38519	3,64596
31	1,30946	1,69552	2,03951	2,14383	2,45282	2,74404	3,37490	3,63345
32	1,30857	1,69389	2,03693	2,14090	2,44868	2,73848	3,36531	3,62180
33	1,30774	1,69236	2,03452	2,13816	2,44479	2,73328	3,35634	3,61091
34	1,30695	1,69092	2,03224	2,13558	2,44115	2,72840	3,34793	3,60072
35	1,30621	1,68957	2,03011	2,13316	2,43772	2,72381	3,34004	3,59115
36	1,30551	1,68830	2,02809	2,13087	2,43449	2,71948	3,33262	3,58215
37	1,30485	1,68709	2,02619	2,12871	2,43145	2,71541	3,32563	3,57367
38	1,30423	1,68595	2,02439	2,12667	2,42857	2,71156	3,31903	3,56568
39	1,30364	1,68488	2,02269	2,12474	2,42584	2,70791	3,31279	3,55811
40	1,30308	1,68385	2,02108	2,12291	2,42326	2,70446	3,30688	3,55096
41	1,30254	1,68288	2,01954	2,12117	2,42080	2,70118	3,30127	3,54418
42	1,30204	1,68195	2,01808	2,11952	2,41847	2,69807	3,29595	3,53774
43	1,30155	1,68107	2,01669	2,11794	2,41625	2,69510	3,29089	3,53162
44	1,30109	1,68023	2,01537	2,11644	2,41413	2,69228	3,28607	3,52580
45	1,30065	1,67943	2,01410	2,11500	2,41212	2,68959	3,28148	3,52025
46	1,30023	1,67866	2,01290	2,11364	2,41019	2,68701	3,27710	3,51496
47	1,29982	1,67793	2,01174	2,11233	2,40835	2,68456	3,27291	3,50990
48	1,29944	1,67722	2,01063	2,11107	2,40658	2,68220	3,26891	3,50507
49	1,29907	1,67655	2,00958	2,10987	2,40489	2,67995	3,26508	3,50045
50	1,29871	1,67590	2,00856	2,10872	2,40327	2,67779	3,26141	3,49601

Zadanie 13

Wartości z jakiego przedziału mogą przyjmować p-wartość i poziom istotności? Uzasadnij z jakiego powodu najczęściej przyjmuje się poziom istotności alfa = 0,05. Jaki jest związek wybranej wielkości poziomu istotności z trudnością odrzucenia hipotezy zerowej?

Zadanie 14

Jaki jest związek liczebności próby z prawdopodobieństwem popełnienia błędu II rodzaju? Odpowiedź uzasadnij.

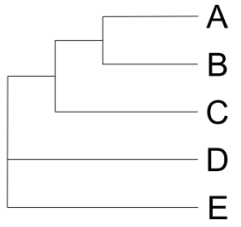
Zadanie 15

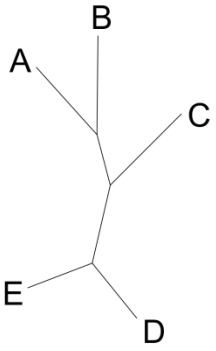
Wartości z jakiego przedziału mogą przyjmować 1) statystyka testu t-studenta dla prób niezależnych, 2) statystyka testu t-studenta dla prób zależnych (dla par) oraz 3) statystka testu chi-kwadrat?

Zadanie 16

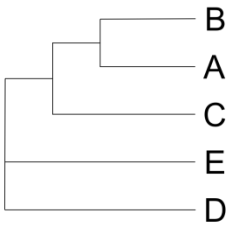
Które z przedstawionych par dychotomicznych drzew filogenetycznych przedstawiają te same topologie?

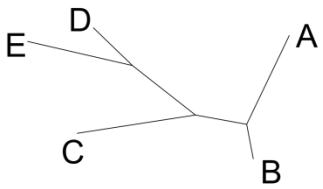
Para I



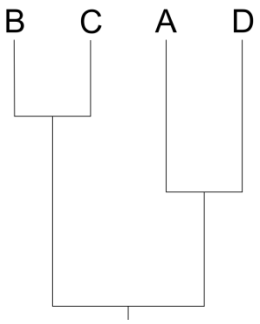


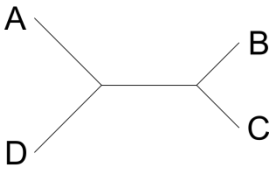
Para II





Para III





Odpowiedzi i komentarz do zadań 12-16

1. Różnice wartości skurczowego ciśnienia tętniczego krwi po i przed kuracją dla każdego pacjenta.

ID pacjenta	Ciśnienie skurczowe przed kuracją a [mm Hg]	Ciśnienie skurczowe po kuracji b [mm Hg]	Różnica $x = b - a$
1	160	159	**-1**
2	151	160	**9**
3	158	156	**-2**
4	161	155	**-6**
5	161	170	**9**
6	163	155	**-8**
7	153	163	**10**
8	159	158	**-1**
9	163	160	**-3**
10	158	157	**-1**
11	162	165	**3**
12	164	163	**-1**
13	158	163	**5**

2. Hipoteza zerowa: **Średnia różnica ciśnienia skurczowego krwi tętniczej po i przed terapii wynosi zero ($\hat{x} = 0$)** – daszek oznacza wartość średnią w populacji generalnej.

3. Statystyka testu: $\boldsymbol{t_0 = 0,6245}$.
4. Liczba stopni swobody: $\boldsymbol{df = 12}$

5. P-wartość: $\boldsymbol{p > 0,2}$ (dokładna wartość wyliczona przez odpowiednie oprogramowanie to 0,544)

6. **Nie odrzucamy hipotezy zerowej** bo $p > \alpha$

7. **Nie udało się odrzucić hipotezy zerowej, więc nie została wykazana skuteczność leku w walce z nadciśnieniem tętniczym.**

8. Zadanie powinno zostać rozwiązane poprawnie w czasie poniżej **15 min**.

P-wartość to prawdopodobieństwo popełnienia błędu pierwszego rodzaju, pod warunkiem, że hipoteza zerowa jest spełniona. Poziom istotności jest to graniczna (największa) wartość prawdopodobieństwa popełnienia błędu pierwszego rodzaju na jaką się godzimy. Z definicji prawdopodobieństwo zawiera się w przedziale $[0; 1]$. Między innymi ze względu na to, że dysponujemy zawsze skończoną próbą to p-wartość nigdy nie osiąga zera – zawsze będzie brakowało dowodów, żeby na 100% odrzucić hipotezę zerową. Podobnie nigdy na podstawie próby nie powiemy, że hipoteza zerowa jest prawdziwa, a p-wartość praktycznie nie może osiągnąć wartości jeden. Zatem p-wartość nie przyjmuje wartości skrajnych i zawiera się w przedziale $(0; 1)$ podobnie jak poziom istotności. Jeżeli wyliczone wartości wychodzą poza ten zakres, to znaczy, że gdzieś w obliczeniach jest błąd!

Im poziom istotności jest mniejszy tym trudniej jest odrzucić hipotezę zerową bo godzimy się na mniejsze prawdopodobieństwo popełnienia błędu pierwszego rodzaju

(rzadziej dostajemy wynik fałszywie dodatni). Niestety poziom istotności jest wartością zupełnie arbitralną. To od nas zależy na jakie prawdopodobieństwo błędu się godzimy. Oczywiście w badaniach ważnych leków ten poziom będziemy zwykle ustawiać mniejszy, a gdy badania będę dotyczyły np. kremu na zmarszczki, który chcemy sprzedać nie bardzo się przejmując jego faktycznym działaniem, to będziemy skłonni do mniej rygorystycznych badań.

Im większa jest próba badana, tym łatwiej jest wykazać odstępstwa od hipotezy zerowej, więc prawdopodobieństwo popełnienia błędu drugiego rodzaju maleje. Z drugiej jednak strony przy bardzo licznych próbach praktycznie zawsze hipoteza zerowa jest odrzucana – test staje się wtedy bardzo czuły. Weźmy na przykład badania, w których testowano skuteczność dwóch leków obniżających ciśnienie. Jest to zupełnie nieprawdopodobne, żeby obydwa leki dawały zupełnie ten sam efekt. Przy mało licznej próbie nie uda się znaleźć odstępstw od hipotezy zerowej, bo różnica w ich działaniu jest zbyta mała i popełnimy błąd II rodzaju. Z kolej jeżeli przebadamy kilka tysięcy pacjentów znajdziemy tę różnicę i określimy ją jako istotną statystycznie, a nasze badania będą pozbawione błędu. Niemniej jednak jeżeli faktyczna różnica pomiędzy lekami polega na tym, że grupa pacjentów przyjmująca jeden z nich ma skurczowe ciśnienie krwi tętniczej obniżone o 3 mm Hg w porównaniu z drugą grupą to jest wynik nieistotny klinicznie. Dojdziemy do wniosku, że różnica jaka faktycznie występuje między lekami nie ma znaczenia praktycznego, a leki można stosować zamiennie zważając raczej na koszty terapii i jej skutki uboczne, a nie skuteczność obniżania ciśnienia krwi.

Statystyka testu t-studenta (niezależnie którego) zawiera się w przedziale $(-\infty; +\infty)$, statystyka testu chi-kwadrat w przedziale $[0; +\infty)$.

W pierwszych dwóch parach drzew jest wszędzie ta sama topologia (cztery drzewa). Trzecia para zawiera dwie różne topologie – pierwsze drzewo III pary jest zakorzenione, a drugie nie. Pierwsze drzewa z I i II pary nie są zakorzenione – posiadają tzw. bazalną trichotomię, czyli wierzchołek o stopniu 3. W skład topologii wchodzi ułożenie gałęzi oraz miejsce zakorzenienia. Bez znaczenia są długości gałęzi, czy forma rysunku.

PRZED ROZPOCZĘCIEM ROZWIĄZYWANIA ZADAŃ **17** i **19** ZNAJDUJĄCYCH SIĘ NA NASTĘPNYCH STRONACH WŁĄCZ STOPER! KIEDY SKOŃCZYSZ, ZAPISZ ILE CZASU ZAJĘŁO CI ROZWIĄZANIE KAŻDEGO ZADANIA.

DO MIERZONEGO CZASU WLICZA SIĘ TAKŻE PRZEPISANIE ODPOWIEDZI NA CZYSTO ORAZ ZAPOZNANIE SIĘ ZE WSTĘPEM!

Lekcja 6

Zadanie 17

W Zadaniu 13 wszystkie obliczenia prowadzące do określenia wartości statystyki testu wykonywaliśmy na zbiorze liczb określających różnicę pomiędzy ciśnieniem skurczowym po i przed kuracją. Tak naprawdę testowaliśmy czy średnia wartość tych różnic jest równa zero w populacji generalnej. W ogólności możemy średnią dowolnego zbioru liczb porównać z jakąkolwiek wartością teoretyczną. Wzór na obliczenie statystyki testowej jest wtedy następujący:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\bar{x}$ – wartość średnia z próby
μ_0 – wartość teoretyczna
n – liczebność próby
s – odchylenie standardowe próby wg wzoru:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Jak widzicie w przypadku hipotezy zerowej mówiącej o tym, że średnia w populacji generalnej jest równa zero ($\mu_0 = 0$) wzór ten redukuje się zupełnie do postaci z Zadania 13.

W pewnym mieście zmierzono wszystkich mężczyzn i okazało się, że średnia ich wzrostu podana z dokładnością do 1 cm wynosi 180 cm. Twoim zadaniem będzie odpowiedź na pytanie: **Czy w drugim mieście średnia wzrostu mężczyzn jest taka sama?**

Poniższa tabela zawiera pomiary wzrostu 10 mężczyzn z drugiego miasta wyrażone w centymetrach:

180	174	179	181	181	182	176	179	182	179

1. **Sformułuj hipotezę zerową.**

2. **Sformułuj hipotezę alternatywną.**

3. **Oblicz statystykę testu wg poniższego wzoru:**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\bar{x}$ – wartość średnia z próby
μ_0 – wartość teoretyczna
n – liczebność próby
s – odchylenie standardowe próby wg wzoru:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

4. **Oblicz liczbę stopni swobody wg poniższego wzoru:**

$$df = n - 1$$

5. **Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej znajdź przedział w jakim znajduje się p-wartość.** Weź pod uwagę wartość bezwzględną statystyki testowej.

6. **Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,05.**

7. **Zinterpretuj wynik testu statystycznego – określ, czy można powiedzieć, że średni wzrost mężczyzn w obydwu miastach jest taki sam?**

Tabela 1. Wartości krytyczne statystyki testowej.

Liczba stopni swobody	p-wartość w teście dwustronnym							
	0.2	**0.1**	**0.05**	**0.04**	**0.02**	**0.01**	**0.002**	**0.001**
1	3,07768	6,31375	12,7062	15,8945	31,8205	63,6568	318,306	636,627
2	1,88562	2,91999	4,30265	4,84873	6,96456	9,92484	22,3272	31,5990
3	1,63774	2,35336	3,18245	3,48191	4,54070	5,84091	10,2145	12,9240
4	1,53321	2,13185	2,77644	2,99853	3,74695	4,60409	7,17318	8,61031
5	1,47588	2,01505	2,57058	2,75651	3,36493	4,03214	5,89344	6,86884
6	1,43976	1,94318	2,44691	2,61224	3,14267	3,70743	5,20763	5,95880
7	1,41492	1,89458	2,36462	2,51675	2,99795	3,49948	4,78528	5,40787
8	1,39682	1,85955	2,30600	2,44898	2,89646	3,35539	4,50079	5,04130
9	1,38303	1,83311	2,26216	2,39844	2,82144	3,24984	4,29681	4,78092
10	1,37218	1,81246	2,22814	2,35931	2,76377	3,16927	4,14370	4,58691
11	1,36343	1,79588	2,20099	2,32814	2,71808	3,10581	4,02470	4,43697
12	1,35622	1,78229	2,17881	2,30272	2,68100	3,05454	3,92963	4,31779
13	1,35017	1,77093	2,16037	2,28160	2,65031	3,01228	3,85198	4,22083
14	1,34503	1,76131	2,14479	2,26378	2,62449	2,97684	3,78739	4,14045
15	1,34061	1,75305	2,13145	2,24854	2,60248	2,94671	3,73283	4,07276
16	1,33676	1,74588	2,11991	2,23536	2,58349	2,92078	3,68615	4,01500
17	1,33338	1,73961	2,10982	2,22385	2,56693	2,89823	3,64576	3,96512
18	1,33039	1,73406	2,10092	2,21370	2,55238	2,87844	3,61048	3,92164
19	1,32773	1,72913	2,09302	2,20470	2,53948	2,86094	3,57940	3,88341
20	1,32534	1,72472	2,08596	2,19666	2,52798	2,84534	3,55181	3,84952
21	1,32319	1,72074	2,07961	2,18943	2,51765	2,83136	3,52715	3,81927
22	1,32124	1,71714	2,07387	2,18289	2,50832	2,81876	3,50499	3,79214
23	1,31946	1,71387	2,06866	2,17696	2,49987	2,80734	3,48496	3,76762
24	1,31784	1,71088	2,06390	2,17154	2,49216	2,79694	3,46678	3,74539
25	1,31635	1,70814	2,05954	2,16659	2,48511	2,78744	3,45019	3,72514

26	1,31497	1,70562	2,05553	2,16203	2,47863	2,77871	3,43500	3,70660
27	1,31370	1,70329	2,05183	2,15783	2,47266	2,77068	3,42103	3,68959
28	1,31253	1,70113	2,04841	2,15393	2,46714	2,76326	3,40816	3,67391
29	1,31143	1,69913	2,04523	2,15033	2,46202	2,75639	3,39624	3,65941
30	1,31041	1,69726	2,04227	2,14697	2,45726	2,75000	3,38519	3,64596
31	1,30946	1,69552	2,03951	2,14383	2,45282	2,74404	3,37490	3,63345
32	1,30857	1,69389	2,03693	2,14090	2,44868	2,73848	3,36531	3,62180
33	1,30774	1,69236	2,03452	2,13816	2,44479	2,73328	3,35634	3,61091
34	1,30695	1,69092	2,03224	2,13558	2,44115	2,72840	3,34793	3,60072
35	1,30621	1,68957	2,03011	2,13316	2,43772	2,72381	3,34004	3,59115
36	1,30551	1,68830	2,02809	2,13087	2,43449	2,71948	3,33262	3,58215
37	1,30485	1,68709	2,02619	2,12871	2,43145	2,71541	3,32563	3,57367
38	1,30423	1,68595	2,02439	2,12667	2,42857	2,71156	3,31903	3,56568
39	1,30364	1,68488	2,02269	2,12474	2,42584	2,70791	3,31279	3,55811
40	1,30308	1,68385	2,02108	2,12291	2,42326	2,70446	3,30688	3,55096
41	1,30254	1,68288	2,01954	2,12117	2,42080	2,70118	3,30127	3,54418
42	1,30204	1,68195	2,01808	2,11952	2,41847	2,69807	3,29595	3,53774
43	1,30155	1,68107	2,01669	2,11794	2,41625	2,69510	3,29089	3,53162
44	1,30109	1,68023	2,01537	2,11644	2,41413	2,69228	3,28607	3,52580
45	1,30065	1,67943	2,01410	2,11500	2,41212	2,68959	3,28148	3,52025
46	1,30023	1,67866	2,01290	2,11364	2,41019	2,68701	3,27710	3,51496
47	1,29982	1,67793	2,01174	2,11233	2,40835	2,68456	3,27291	3,50990
48	1,29944	1,67722	2,01063	2,11107	2,40658	2,68220	3,26891	3,50507
49	1,29907	1,67655	2,00958	2,10987	2,40489	2,67995	3,26508	3,50045
50	1,29871	1,67590	2,00856	2,10872	2,40327	2,67779	3,26141	3,49601

Zadanie 18

Korzystając z danych i obliczeń wykonanych do Zadania 18 oraz przekształcając podane tam wzory znajdź przedział w jakim znajdują się wszystkie wartości teoretyczne μ_0, dla których nie można odrzucić hipotezy zerowej na poziomie istotności $\alpha = 0{,}01$.

Zadanie 19

Posługując się macierzą danych molekularnych z zadania 9 oszacuj relacje pokrewieństwa za pomocą metody największej parsymonii dla taksonów A-D. Oblicz długość najkrótszego drzewa (lub drzew) i określi minimalną liczbę zmian dla każdego miejsca w przyrównaniu na tym drzewie.

Odpowiedzi i komentarz do Zadań 17-19

1. Hipoteza zerowa: **1) Średnia wzrostu mężczyzn w pierwszym i drugim mieście jest sobie równa,** lub **2) Średnia wzrostu mężczyzn w drugim mieście jest równa 180 cm.**

2. Hipoteza alternatywna: **1) Średnia wzrostu mężczyzn w pierwszym i drugim mieście nie jest sobie równa,** lub **2) Średnia wzrostu mężczyzn w drugim mieście nie jest równa 180 cm.**

3. Statystyka testu: $\boldsymbol{t_0 = -0,8566}$

4. Liczba stopni swobody: $\boldsymbol{df = 9}$

5. P-wartość: $\boldsymbol{p > 0,2}$ (dokładna wartość obliczona przez odpowiednie oprogramowanie to 0,4139)

6. **Nie odrzucamy hipotezy zerowej** bo $p > \alpha$

7. **Nie udało się odrzucić hipotezy zerowej, więc zachowujemy *status quo*. Możemy jedynie powiedzieć, że nie udało się znaleźć dowodów na różnice we wzroście mężczyzn w tych dwu miastach.** Błędem byłoby powiedzieć, że tych różnic nie ma. Ze względu na niewielką liczebność próby mamy prawdopodobnie do czynienia z dużym błędem II rodzaju.

W Zadaniu 19 problem jest postawiony nieco na opak. Szukamy takich wartości μ_0, które na poziomie istotności $\alpha = 0{,}01$ dałyby wynik na granicy odrzucenia i zachowania hipotezy zerowej. Innymi słowy szukamy takich wartości teoretycznych μ_0, dla których wartość statystyki testowej równa byłaby jej wartości krytycznej dla danego poziomu istotności $\alpha = 0{,}01$. Z tego powodu należy dokonać przekształcenia wzoru z Zadania 18:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$\mu_0 = \bar{x} + t_0 \frac{s}{\sqrt{n}}$$

Wartości n, $\bar{x}$ oraz s zostały już obliczone w Zadaniu 18 bezpośrednio na podstawie danych o wzroście mężczyzn. Następnym krokiem jest odnalezienie w tabeli wartości krytycznych statystyki testowej wartości, która odpowiada poziomowi istotności $\alpha = 0{,}01$ dla 9 stopni swobody. Jest to t_α= 3,24984, ale musimy pamiętać, że statystyka testowa może przyjmować wartości ujemne lub dodatnie, a podczas obliczeń posługiwaliśmy się do tej pory jej wartością bezwzględną – to dlatego w tabeli znajdują się wartości wyłącznie dodatnie. Z tego powodu górną i dolną granicę przedziału, w którym znajdują się wartości μ_0 znajdziemy ze wzoru:

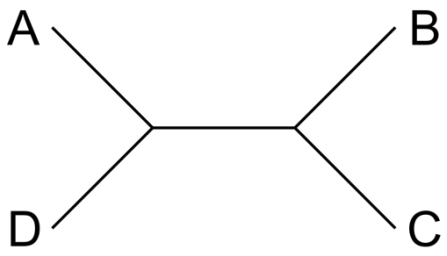
$$\mu_0 = \bar{x} \pm t_{\alpha=0,01} \frac{s}{\sqrt{n}}$$

Co daje wynik w postaci $\boldsymbol{\mu_0 \in (176,6443; 181,9557)}$.

Obliczony przedział nazywa się w statystyce 99% przedziałem ufności dla średniej. Najczęściej jest interpretowany **mylnie** jako przedział, w którym z 99% prawdopodobieństwem znajduje się faktyczna wartość średnia w populacji, z której pobierane były próby. Jedyną słuszną interpretacją tego przedziału jest to, że obejmuje on wszystkie hipotezy zerowe o wartości średniej populacyjnej, których nie da się odrzucić na poziomie istotności $\alpha = 0{,}01$. Im większa liczebność próby, tym przedział ufności staje się węższy.

Odpowiedzi i komentarz do Zadania 19

Metoda największej parsymonii przeszukuje przestrzeń rozwiązań w postaci wszystkich możliwych niezakorzenionych topologii i dla każdej z nich oblicza długość drzewa, czyli minimalną liczbę zmian ewolucyjnych jakie muszą zajść na tym drzewie. W naszym przypadku są możliwe tylko trzy topologie, z których jedna w sposób najbardziej oszczędny tłumaczy ewolucję cech w postaci pięciu zmian:



Takson A	A	T	C	C	A	T	G	G	A	A
Takson B	A	T	A	C	A	T	G	C	C	C
Takson C	C	T	A	C	A	T	G	C	C	C
Takson D	A	T	C	C	A	T	G	G	C	C
Liczba zmian	1	0	**1**	0	0	0	0	**1**	1	1

Tylko trzecie i ósme miejsce w przyrównaniu należało uwzględnić w obliczeniach. Pozostałe miejsca są niezmienne (zero zmian), lub autapomorficzne (zawsze jedna zmiana niezależnie od topologii). Trzecia i ósma cecha tworzą zgodne synapomorfie wyróżniające klady AD oraz BC. Na drzewie nie ma żadnych homoplazji.

Zakorzenienie w dowolnym miejscu (przez złamanie dowolnej z pięciu gałęzi) nie zmienia długości drzewa. Metoda największej parsymonii w przeciwieństwie do UPGMA nie odpowie nam na pytanie, jak zakorzenić drzewo.

Przykładowe zadania na Olimpiadę Biologiczną

Statystyka i filogenetyka

Czas 90 min.
Łączna liczba punktów do zdobycia 12

Odpowiedzi zapisz w miejscu na to przeznaczonym przy każdym z zadań.
UWAGA! Wszystkie wyniki w postaci ułamków należy podawać z dokładnością do czterech miejsc po przecinku!

Zadanie 1

Dla poniższego zbioru liczb reprezentującego pomiary wysokości populacji roślin wyrażone w cm znajdź średnią arytmetyczną, wariancję, oraz odchylenie standardowe. Wyniki wpisz w tabelę.

$$A = \{11{,}86; 0{,}62; 9{,}32; 13{,}10; 13{,}63; 15{,}40; 4{,}09; 10{,}48; 15{,}02; 9{,}18; 14{,}43\}$$

Wzór na wariancję w populacji:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Wzór na odchylenie standardowe w populacji:

$$\sigma = \sqrt{\sigma^2}$$

Parametr	
Średnia arytmetyczna	
Wariancja	
Odchylenie standardowe	

Zadanie 2

Za pomocą testu t-studenta dla par sprawdź czy testowany herbicyd zmniejsza liczbę chwastów rosnących na polu.

1. W poniższej tabeli znajdziesz dane zebrane z 14 poletek badawczych przedstawiające liczbę obcych roślin przed i tydzień po zastosowaniu oprysku. **Oblicz różnicę w liczbie chwastów przed i po zastosowaniu oprysku. Wyniki wpisz w tabelę.**

ID poletka	Liczba chwastów przed opryskiem a	Liczba chwastów po oprysku b	Różnica $x = b - a$
1	245	230	
2	207	213	
3	236	210	
4	249	271	
5	250	208	
6	256	242	
7	219	221	
8	240	231	
9	255	220	

10	236	251	
11	253	242	
12	260	241	
13	236	226	
14	242	264	

2. Sformułuj hipotezę zerową.

.....

.....

.....

3. Oblicz statystykę testu wg poniższego wzoru:

$$t_0 = \frac{\bar{x}}{s/\sqrt{n}}$$

$\bar{x}$ – średnia różnica między liczbą chwastów przed i po oprysku
n – liczba poletek
s – odchylenie standardowe różnicy miedzy liczbą chwastów przed i po oprysku

Wzór na odchylenie standardowe w próbie:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$t_0 =$.....

4. Oblicz liczbę stopni swobody wg poniższego wzoru:

$$df = n - 1$$

$df =$.....

5. Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej znajdź przedział w jakim znajduje się p-wartość. Weź pod uwagę wartość bezwzględną statystyki testowej.

……. $< p <$ …….

6. **Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,05. Odpowiedź uzasadnij.**

.....

.....

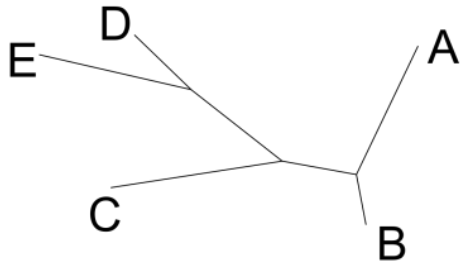
.....

Wartości krytyczne statystyki testowej.

Liczba stopni swobody	**p-wartość w teście dwustronnym**							
	0.2	**0.1**	**0.05**	**0.04**	**0.02**	**0.01**	**0.002**	**0.001**
1	3,07768	6,31375	12,7062	15,8945	31,8205	63,6568	318,306	636,627
2	1,88562	2,91999	4,30265	4,84873	6,96456	9,92484	22,3272	31,5990
3	1,63774	2,35336	3,18245	3,48191	4,54070	5,84091	10,2145	12,9240
4	1,53321	2,13185	2,77644	2,99853	3,74695	4,60409	7,17318	8,61031
5	1,47588	2,01505	2,57058	2,75651	3,36493	4,03214	5,89344	6,86884
6	1,43976	1,94318	2,44691	2,61224	3,14267	3,70743	5,20763	5,95880
7	1,41492	1,89458	2,36462	2,51675	2,99795	3,49948	4,78528	5,40787
8	1,39682	1,85955	2,30600	2,44898	2,89646	3,35539	4,50079	5,04130
9	1,38303	1,83311	2,26216	2,39844	2,82144	3,24984	4,29681	4,78092
10	1,37218	1,81246	2,22814	2,35931	2,76377	3,16927	4,14370	4,58691
11	1,36343	1,79588	2,20099	2,32814	2,71808	3,10581	4,02470	4,43697
12	1,35622	1,78229	2,17881	2,30272	2,68100	3,05454	3,92963	4,31779
13	1,35017	1,77093	2,16037	2,28160	2,65031	3,01228	3,85198	4,22083
14	1,34503	1,76131	2,14479	2,26378	2,62449	2,97684	3,78739	4,14045
15	1,34061	1,75305	2,13145	2,24854	2,60248	2,94671	3,73283	4,07276
16	1,33676	1,74588	2,11991	2,23536	2,58349	2,92078	3,68615	4,01500
17	1,33338	1,73961	2,10982	2,22385	2,56693	2,89823	3,64576	3,96512
18	1,33039	1,73406	2,10092	2,21370	2,55238	2,87844	3,61048	3,92164
19	1,32773	1,72913	2,09302	2,20470	2,53948	2,86094	3,57940	3,88341
20	1,32534	1,72472	2,08596	2,19666	2,52798	2,84534	3,55181	3,84952
21	1,32319	1,72074	2,07961	2,18943	2,51765	2,83136	3,52715	3,81927
22	1,32124	1,71714	2,07387	2,18289	2,50832	2,81876	3,50499	3,79214
23	1,31946	1,71387	2,06866	2,17696	2,49987	2,80734	3,48496	3,76762
24	1,31784	1,71088	2,06390	2,17154	2,49216	2,79694	3,46678	3,74539
25	1,31635	1,70814	2,05954	2,16659	2,48511	2,78744	3,45019	3,72514
26	1,31497	1,70562	2,05553	2,16203	2,47863	2,77871	3,43500	3,70660
27	1,31370	1,70329	2,05183	2,15783	2,47266	2,77068	3,42103	3,68959
28	1,31253	1,70113	2,04841	2,15393	2,46714	2,76326	3,40816	3,67391
29	1,31143	1,69913	2,04523	2,15033	2,46202	2,75639	3,39624	3,65941
30	1,31041	1,69726	2,04227	2,14697	2,45726	2,75000	3,38519	3,64596

Zadanie 3

Dla poniższego drzewa filogenetycznego i podanej macierzy danych molekularnych oblicz minimalną liczbę zmian ewolucyjnych dla każdego miejsca w przyrównaniu. Wyniki wpisz w tabeli.



Takson A	A	T	C	C	G	T	G	G	A	A
Takson B	A	T	A	C	T	T	G	C	T	C
Takson C	C	T	A	C	A	T	G	C	C	C
Takson D	A	T	C	C	G	T	G	G	G	C
Takson E	A	T	C	C	A	T	G	G	C	C
Liczba zmian										

Zadanie 4

Określ, które z miejsc w przyrównaniu z Zadania 3 są nieprzydatne do szacowania relacji pokrewieństwa metodą największej parsymonii. Odpowiedź uzasadnij.

.....

.....

.....

.....

.....

Zadanie 5

Drzewo z Zadania 3 nie jest drzewem optymalnym wg kryterium największej parsymonii. **Narysuj najkrótsze dychotomiczne niezakorzenione drzewo. Podpowiedź: Jest tylko jedno takie drzewo i jego długość wynosi 9 kroków.**

Schemat punktowania

Zadanie 1

2 pkt. – za poprawne obliczenie wszystkich trzech parametrów.
1 pkt. – za błędne obliczenie średniej arytmetycznej, ale właściwe obliczenie na jej podstawie wariancji i odchylenia standardowego.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Rozwiązanie:

Parametr	
Średnia arytmetyczna	**10,6482 cm**
Wariancja	**19,9898 cm^2**
Odchylenie standardowe	**4,4710 cm**

Zadanie 2.1

1 pkt. – za prawidłowe obliczenie wszystkich różnic.
0 pkt. – za każdą inną odpowiedź lub brak odpowiedzi.

Rozwiązanie:

ID poletka	Liczba chwastów przed opryskiem a	Liczba chwastów po oprysku b	Różnica $x = b - a$
1	245	230	**-15**
2	207	213	**6**
3	236	210	**-26**
4	249	271	**22**
5	250	208	**-42**
6	256	242	**-14**
7	219	221	**2**
8	240	231	**-9**
9	255	220	**-35**
10	236	251	**15**
11	253	242	**-11**
12	260	241	**-19**
13	236	226	**-10**
14	242	264	**22**

Zadanie 2.2

1 pkt. – za prawidłową hipotezę zerową sformułowaną w sposób opisowy lub jednoznacznego wyrażenia matematycznego.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Przykładowe rozwiązania:
- Oprysk nie zmniejsza liczby chwastów na polu.
- Średnia różnica w liczbie chwastów na polu przed i po oprysku jest równa zero.
- $\mu_x = 0$

Zadanie 2.3

1 pkt. – za prawidłowe obliczenie statystyki testu na podstawie obliczonych różnic.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Rozwiązanie: **1,5453** (przy założeniu obliczenia prawidłowych różnic w zadaniu 2.1)

Zadanie 2.4

1 pkt. – za prawidłowe obliczenie liczby stopni swobody.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Rozwiązanie: **13**

Zadanie 2.5

1 pkt. – za określenie obydwu granic przedziału, w którym znajduje się p-wartość na podstawie prawidłowego porównania obliczonej wartości statystyki testu z tabelą wartości krytycznych.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Rozwiązanie: $\mathbf{0,1 < p < 0,2}$ (przy założeniu obliczenia prawidłowej wartości statystki testu w zadaniu 2.3)

Zadanie 2.6

1 pkt. – za określenie, czy należy odrzucić hipotezę zerową na podstawie prawidłowego porównania obliczonego zakresu, w którym znajduje się p-wartość z zadanym poziomem istotności.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Przykładowe rozwiązania (przy założeniu obliczenia prawidłowej p-wartości):

- Nie odrzucamy hipotezy zerowej, bo $p > \alpha$.
- Nie należy odrzucić hipotezy zerowej, ponieważ p-wartość jest większa od 0,05.

Zadanie 3

2 pkt. – za prawidłowe określenie liczby zmian dla wszystkich 10 miejsc w przyrównaniu.
1 pkt. – za prawidłowe określenie liczby zmian dla 8 lub 9 miejsc w przyrównaniu.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Rozwiązanie:

Takson A	A	T	C	C	G	T	G	G	A	A
Takson B	A	T	A	C	T	T	G	C	T	C
Takson C	C	T	A	C	A	T	G	C	C	C
Takson D	A	T	C	C	G	T	G	G	G	C
Takson E	A	T	C	C	A	T	G	G	C	C
Liczba zmian	**1**	**0**	**2**	**0**	**3**	**0**	**0**	**2**	**3**	**1**

Zadanie 4

1 pkt. – za wskazanie miejsc 1, 2, 4, 6, 7, 10 z prawidłowym uzasadnieniem odnoszącym się do ich zmienności.
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Przykładowe rozwiązania:

- Nieprzydatne są miejsca 1, 2, 4, 6, 7, 10 ponieważ są niezmienne lub autapomorficzne.
- Nieprzydatne do szacowania filogenezy są miejsca w przyrównaniu, które mają wszystkie takie same nukleotydy, lub takie, które mają jeden nukleotyd inny od pozostałych, które są takie same. W tym przypadku są to wiec miejsca 1, 2, 4, 6, 7 i 10.

Zadanie 5

1 pkt. – za prawidłowe narysowanie topologii drzewa (długości gałęzi nie mają znaczenia).
0 pkt. – za odpowiedź niespełniającą powyższych kryteriów lub brak odpowiedzi.

Uwaga: Nie zalicza się odpowiedzi, w których drzewo jest zakorzenione (zawiera węzeł o stopniu równym dwa).

Przykładowe rozwiązanie: