Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

PRACA DYPLOMOWA MAGISTERSKA
NA KIERUNKU MATEMATYKA

METODY STATYSTYCZNEJ IDENTYFIKACJI ZMIAN WARIANTÓW SPLICINGOWYCH WRAZ Z PRZYKŁADAMI ZASTOSOWAŃ W ANALIZIE DANYCH RNA-SEQ

Autor:
Paulina Auguścik

Promotor:
dr hab. inż. Przemysław Biecek prof. nadzw.

Warszawa, grudzień 2015

.....

podpis promotora

.....

podpis autora

Streszczenie

Temat pracy magisterskiej:

Metody statystycznej identyfikacji zmian wariantów splicingowych wraz z przykładami zastosowań w analizie danych RNA-Seq

Głównym tematem pracy jest zaproponowanie oraz porównanie metod statystycznych, służących do identyfikacji różnic w dwóch grupach. Istotne jest uwzględnienie zależności między pomiarami. Jest to kluczowa część pracy, która może być zastosowana do różnego rodzaju danych. Niniejsza praca bazuje natomiast na danych biologicznych.

W procesie tworzenia RNA, powstaje prekursorowy RNA, który składa się z eksonów i intronów. Eksony to fragmenty niosące informację biologiczną, a introny to elementy jej nie zawierające. W związku z tym, podczas dojrzewania RNA, introny są usuwane, a eksony łączone ze sobą. Proces ten nazywany jest splicingiem. Jeśli część eksonów zostanie usunięta lub introny pozostaną, mamy doczynienia z procesem alternatywnego splicingu. Na jego skutek powstają różne warianty genów.

Jeśli gen występuje w kilku wariantach splicingowych, to można określić prawdopodobieństwo, że w danej grupie komórek, wystąpi w konkretnej izoformie. Jednym z celów pracy jest przedstawienie metody estymacji prawdopodobieństw występowania wariantów. Zaprezentowany został algorytm Casper, który zakłada, że obserwowane dane pochodzą z mieszaniny dyskretnych rozkładów prawdopodobieństwa. Za pomocą algorytmu EM, estymowane są wagi tej mieszaniny. W pracy przedstawiony jest opis matematyczny modelu w algorytmie Casper. W wyniku jego działania, otrzymuje się prawdopodobieństwa występowania izoform genów, osobno dla każdego pomiaru w danych.

Dwie grupy komórek mogą różnić się, pod względem rozkładu występowania wariantów splicingowych, dla niektórych genów. Kolejnym celem niniejszej pracy jest zaproponowanie metod,

które pozwolą zidentyfikować geny, istotnie różniące się między grupami, pod względem rozkładu występowania izoform. Przedstawiono cztery testy statystyczne, które są wykonywane dla każdego pomiaru oraz wariantu osobno. Trzy z nich uwzględniają występujące w danych zależności między pomiarami. Nie założono, że statystyki wybranych testów mają teoretyczne rozkłady, takie jak dla danych niezależnych. Rozkłady statystyk wyznaczono symulacyjnie, z założeniem różnych rozkładów danych wejściowych oraz różnych zależności w danych. Po zbadaniu wrażliwości tych statystyk na zmiany rozkładów i korelacje, w dalszych analizach stosowane są trzy testy. Ostatecznie porównano je, pod względem mocy statystycznej i stwierdzono, że dwa z nich są lepsze od trzeciego i porównywalne między sobą.

Zgodnie z zaproponowaną procedurą, należy najpierw, dla każdego pomiaru w danych i genu, wyestymować prawdopodobieństwa występowania poszczególnych izoform. Następnie na poziomie wariantu przeprowadzić test, po czym zastosować poprawkę Holma na wielokrotne testowanie, związane z liczbą wariantów genu. Jeśli dla przynajmniej jednej izoformy, test odrzuca hipotezę zerową o równości średnich prawdopodobieństw w dwóch grupach, to gen jest uznawany za istotnie różniący się pod względem rozkładu występowania wariantów.

Ostatnią częścią pracy, jest wykorzystanie przedstawionych metod do rzeczywistych danych. Analiza ta została przygotowana we współpracy z Wielkopolskim Centrum Onkologii. Dane, zawierające 12 pomiarów, pochodzących od 4 pacjentów poddane są badaniom zaprezentowanym w poprzednich częściach. Celem jest zidentyfikowanie genów, które istotnie różnią się pod względem rozkładu wariantów splicingowych, w pierwotnych fibroblastach skóry (PHDF) i indukowanych komórkach pluripotentnych (iPS). Analizie poddane są geny z domeną KRAB-ZNF, ponieważ istnieją przypuszczenia, że domena ta ma wpływ na proces reprogramowania komórek somatycznych do komórek iPS oraz utrzymanie stanu pluripotencji komórek iPS. Analizę przeprowadzono z wykorzystaniem programów TopHat (przygotowanie danych) i R (estymacja prawdopodobieństw za pomocą algorytmu Casper oraz testy statystyczne). Rezultatem jest lista genów, które istotnie różnią się pod względem rozkładu wariantów splicingowych w dwóch grupach. Dodatkowo dla każdego genu przygotowano wizualną prezentację otrzymanych wyników.

Spis treści

Wstęp

Niniejsza praca składa się z dwóch części. Jedna z nich to propozycja oraz opis metod statystycznych, mogących służyć do identyfikacji genów, które istotnie różnią się między dwoma grupami, pod względem rozkładu wariantów splicingowych. Druga część to wykorzystanie przedstawionych metod do danych biologicznych. W związku z tym, należy najpierw nakreślić temat od strony biologicznej.

W procesie powstawania RNA z DNA, tworzony jest prekursorowy RNA (pre-RNA), który składa się fragmentów zawierających informację biologiczną (eksonów) oraz niezawierających informacji (intronów). Następnie introny są usuwane, a eksony łączone ze sobą w procesie splicingu. Może również nastąpić tzw. alternatywny splicing, w którym eksony są ze sobą łączone z pominięciem niektórych z nich lub z pozostawieniem niektórych intronów. W ten sposób z jednego genu może powstać wiele wariantów (izoform), co jest jednym ze źródeł zmienności białek oraz może być przyczyną zmian w organizmie lub chorób. Znane są zarówno geny, dla których nie zidentyfikowano istnienia innych wariantów oraz takie, które występują w wielu izoformach [1], [2].

Pierwszym z celów pracy jest zaprezentowanie metody estymacji prawdopodobieństw występowania wariantów splicingowych genów. Przedstawiony został algorytm Casper, który opiera się na algorytmie EM (Expectation - Maximization), maksymalizującym funkcję a-posteriori.

Drugi cel to zaproponowanie metod, które na podstawie wyestymowanych prawdopodobieństw, pozwolą porównać ze sobą dwie grupy komórek. Grupy te porównujemy pod względem rozkładu występowania poszczególnych wariantów splicingowych. Do porównania dwóch grup komórek zaproponowano cztery testy statystyczne, z których trzy uwzględniają występujące w danych zależności między pomiarami. Bardzo ważnym celem jest również wybór najlepszej z przedstawionych metod.

Trzecim celem jest użycie zaproponowanych metod do rzeczywistych danych oraz zidentyfikowanie genów, które istotnie różnią się pod względem rozkładu wariantów w pierwotnych fibroblastach skóry (PHDF) oraz indukowanych komórkach pluripotentnych (iPS). Więcej na temat analizowanych danych, można przeczytać w rozdziale 4. Spośród 349 genów z domeną

KRAB-ZNF, chcemy wybrać te, które mają istotnie różne rozkłady wariantów w zależności od grupy. Wybrane geny mogą być następnie badane laboratoryjnie. Analiza ta została przygotowana we współpracy z Wielkopolskim Centrum Onkologii.

Trudnością w estymacji prawdopodobieństw występowania wariantów splicingowych jest wielkość danych zawierających sekwencje RNA, a co za tym idzie długi czas działania algorytmu. Dane dla 12 pomiarów są zawarte w 24 plikach. Po wstępnym przetworzeniu danych, czyli mapowaniu, otrzymuje się 12 plików, o wielkości od 2 do 5GB każdy. Problemem w identyfikacji genów, które istotnie różnią się między grupami, pod względem rozkładu wariantów, są zależności występujące w danych. Posiadamy 12 pomiarów, pochodzących od czterech pacjentów, gdzie 4 pomiary powstały z komórek PHDF, a 8 pomiarów z komórek iPS. Oznacza to, że dane są zależne zarówno w obrębie jednej grupy, jak i pomiędzy grupami. W związku z tym, należy zastosować testy, które pozwolą uwzględnić takie zależności w danych.

Rozdział pierwszy niniejszej pracy zawiera wprowadzenie biologiczne do tematu, tzn. wyjaśnienie czym jest DNA oraz RNA, jak powstaje RNA, czym jest splicing, alternatywny splicing oraz warianty genów. Dodatkowo znajdziemy tam intuicyjne wyjaśnienie, jak przebiega estymacja prawdopodobieństw występowania wariantów splicingowych. Rozdział drugi prezentuje algorytm Casper, służący do estymacji prawdopodobieństw występowania wariantów splicingowych, dla każdego genu oraz pomiaru osobno. W rozdziale tym przedstawiono również działanie algorytmu EM, na którym bazuje algorytm Casper oraz teorię na temat rozkładu Dirichleta, również wykorzystywanego w tym algorytmie. W rozdziale trzecim zaprezentowane są cztery testy statystyczne, mogące służyć do porównania między sobą dwóch grup komórek, pod względem rozkładu występowania wariantów splicingowych genów. Metody te mają służyć do identyfikacji genów, które istotnie różnią się pod względem rozkładu wariantów między grupami. W rozdziale trzecim przedstawiono również porównanie zaproponowanych metod, w celu weryfikacji najlepszej z nich. Rozdział czwarty prezentuje wykorzystanie przedstawionych w poprzednich rozdziałach metod do rzeczywistych danych. Znajdziemy tam opis danych, wraz z wyjaśnieniem, dlaczego zajmujemy się ich analizą. Dodatkowo w rozdziale czwartym zaprezentowano wykorzystane programy oraz polecenia, pozwalające przeprowadzić przygotowanie danych oraz estymację prawdopodobieństw algorytmem Casper w R. W ostatnim podrozdziale przedstawione są otrzymane wyniki dla kilku genów oraz porównanie zastosowanych testów dla wszystkich badanych genów.

Dziękuję dr Urszuli Oleksiewicz z Wielkopolskiego Centrum Onkologii za współpracę oraz pomoc w zrozumieniu znaczenia biologicznego przedstawianych rozwiązań.

Rozdział 1

Wprowadzenie biologiczne

Niniejszy rozdział zawiera wprowadzenie biologiczne, które jest potrzebne, do pełnego zrozumienia prezentowanych analiz. W jego dalszej części znajdziemy wyjaśnienie czym są warianty splicingowe genów oraz na czym polega estymacja prawdopodobieństw ich występowania.

Informacja biologiczna jest zapisana w cząsteczce DNA. DNA jest polimerem zbudowanym z połączonych ze sobą, w różnej kolejności, czterech różnych nukleotydów, w ilości od setek do miliardów. W kontekście DNA, nukleotydy te nazywa się skrótowo A, C, G i T. Mówimy, że DNA składa się z nukleotydów lub z zasad azotowych. Informacja niezbędna do reprodukcji organizmu jest zawarta w sekwencji nukleotydów w segmantach DNA stanowiących geny. Wielkość genów jest bardzo różna i waha się w granicach od mniej niż 100 do kilku milionów par zasad. DNA ma postać podwójnej helisy, w której każdy łańcuch składa się z nukleotydów, połączonych ze sobą parami między łańcuchami. Wiązania występują między konkretnymi parami zasad azotowych, tzn. między adeniną i tyminą oraz między cytozyną i guaniną. Informacja ta zostaje udostępniona komórce w procesie określanym jako ekspresja genów. Podczas ekspresji genów informacja zawarta w DNA zostaje przepisana do RNA przez syntezę cząsteczek RNA, których sekwencja zasad jest komplementarna do sekwencji zasad w matrycy DNA. Proces ten nazywamy transkrypcją [1, str. 8,10,12], [2, str. 8].

RNA jest polinukleotydem podobnym do DNA, jednak ma on postać pojedynczego łańcucha. RNA składa się z czterech nukleotydów oznaczanych skrótowo A, C, G i U (podobnie, jak w DNA, jednak U zamiast T) [2, str. 15].

W procesie transkrypcji DNA na RNA, sekwencja nukleotydów w RNA jest komplementarna do sekwencji nukleotydów w DNA. Tworzonych jest wiele cząsteczek RNA, które dzielimy na kodujące i niekodujące. Z punktu widzenia białek, interesuje nas RNA kodujące, które zawiera tylko jedną klasę cząsteczek - RNA informacyjne (mRNA, ang. messenger RNA). mRNA są transkryptami genów kodujących białka, tzn. ulegają translacji na białka w drugim

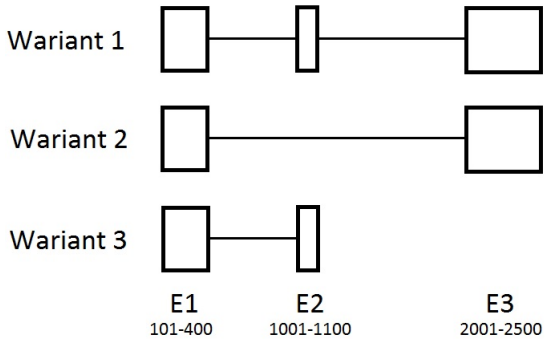
etapie ekspresji genów. Sekwencja zasad RNA determinuje sekwencję aminokwasów w białku, od której uzależniona jest jego przestrzenna struktura, która z kolei dyktuje funkcję białka [2, str. 10, 12, 15-16].

W procesie transkrypcji, przed powstaniem dojrzałej cząsteczki RNA, tworzony jest prekursorowy RNA (pre-RNA), który następnie ulega procesowi dojrzewania. Jednym z etapów dojrzewania jest tzw. splicing, który polega na połączeniu ze sobą fragmentów zawierających informację biologiczną (eksonów). Eksony w pre-mRNA są porozdzielane przez introny, czyli sekwencje nie zawierające użytecznych informacji. Liczba intronów w rożnych genach jest różna i mieści się w granicach od 0 do ponad 50. Także długość eksonów i intronów jest bardzo różna w rożnych genach, a introny są zwykle dłuższe od eksonów i stanowią większość sekwencji genu. Przed wykorzystaniem informacji biologicznej genu do syntezy białka, introny muszą być usunięte z transkryptu, a eksony połączone w jedną ciągłą cząsteczkę RNA. Proces usuwania intronów i równoczesnego łączenia eksonów jest określany jako splicing. Po splicingu dojrzały mRNA jest eksportowany do cytoplazmy, gdzie funkcjonuje jako matryca do syntezy białka [1, str. 11,34], [2, str. 17].

Wiele genów podlega również alternatywnemu splicingowi. Jest to proces biologiczny polegający na połączeniu ze sobą eksonów, z pominięciem niektórych z nich lub z zachowaniem niektórych intronów. W ten sposób z jednego genu może powstać wiele wariantów (izoform), co jest jednym ze źródeł zmienności białek. Np. jeśli pozostawiony w dojrzałej cząsteczce m-RNA intron zawiera kodon stop, wtedy w trakcie translacji powstanie skrócone białko, które może być nieaktywne. Uważa się, że alternatywny splicing przyczynia się do złożoności wyższych organizmów. Dodatkowo wiadomo, że jest on związany z występowaniem wielu chorób [1, str. 295].

Celem badania zaprezentowanego w niniejszej pracy jest identyfikacja genów, dla których rozkłady izoform istotnie różnią się między dwoma grupami komórek. Aby to zrobić, musimy oszacować rozkłady wariantów dla każdego genu i pomiaru w danych (każdego pacjenta, każdej komórki), a następnie w obrębie genu porównać grupy między sobą, pod względem tych rozkładów.

Rozważmy hipotetyczny przykład genu z trzema wariantami splicingowymi (rysunek 1.1). Gen ten składa się z trzech eksonów. Prostokąty na rysunku oznaczają eksony oraz są opisane symbolami E1, E2 i E3. Natomiast przedziały liczbowe znajdujące się pod symbolami oznaczają ich rzeczywiste usytuowanie w DNA. W tym hipotetycznym przykładzie zakładamy, że kiedy gen jest transkryptowany do mRNA mogą powstać trzy izoformy, oznaczone na rysunku jako warianty 1, 2 i 3. Teoretycznie wariantów mogłoby być tyle, ile niepustych podzbiorów eksonów, czyli $2^3 - 1 = 7$, jednak w praktyce geny nie występują we wszystkich możliwych izoformach. Naszym celem jest estymacja proporcji występowania poszczególnych izoform, na



Rysunek 1.1: Przykładowe warianty splicingowe genu. Ten gen składa się z 3 eksonów E1, E2, E3 (prostokąty) oddzielonych dwoma intronami (kreski). Pod nazwami eksonów zapisano ich rzeczywiste usytuowanie w DNA.

podstawie sekwencji nukleotydów występujących w danych. Na przykład wariant 1 składa się na 60% kopii genu, wariant 2 - 30%, a wariant 3 - 10%.

Aby odczytać sekwencje RNA, czyli kolejności nukleotydów w cząsteczce RNA, przeprowadza się proces sekwencjonowania. Najbardziej innowacyjną metodą odczytywania sekwencji jest Sekwencjonowanie Nowej Generacji (NGS, ang. Next Generation Sequencing). Dane, opisane w rozdziale 4, na których przeprowadzamy badania, były sekwencjonowane metodą RNA-Seq, która jest jedną z metod NGS.

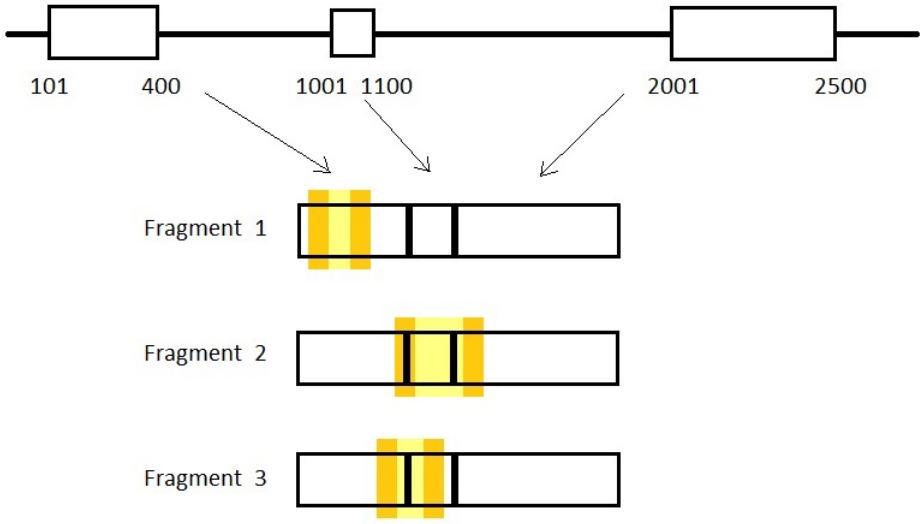
Przebieg sekwencjonowania nowej generacji podzielić można na cztery główne etapy. Pierwszym krokiem procedury jest konwersja RNA na komplementarne DNA (tzw. cDNA), które ma o wiele wyższą stabilność. Drugi krok, to stworzenie biblioteki, czyli bazy cDNA, poprzez losową fragmentację cDNA oraz dołączenie adapterów. Trzeci etap to amplifikacja, a więc powielenie fragmentów. Ostatni krok to masowe, równoległe sekwencjonowanie, co oznacza, że odbywa się ono równocześnie dla wszystkich fragmentów. Identyfikowane są kolejno pojedyncze zasady fragmentu DNA, aż do określonej długości.

Następnie, w celu analiz, odczyty sekwencji są przyrównywane do genomu referencyjnego. Po dopasowaniu możemy zidentyfikować, które fragmenty i w jakiej ilości zostały zsekwencjonowane.

Sekwencjonowanie obustronne (ang. paired-end) polega na sekwencjonowaniu obu końców fragmentów z biblioteki i traktowaniu ich jako pary odczytów. Zaletą tej metody jest fakt, iż w takim samym czasie wytwarzane jest dwa razy więcej odczytów. Dodatkowo, przyrównywanie odczytów jako par, umożliwia dokładniejsze dopasowanie odczytów do genomu referencyjnego, zwłaszcza do trudnych, powtarzalnych regionów w genomie [3] [4].

Rozważmy trzy hipotetyczne, sekwencjonowane obustronnie fragmenty. Na rysunku 1.2 widzimy przedstawiony wcześniej przykładowy gen oraz jego pierwszy wariant, który powstał

z połączenia wszystkich eksonów, po usunięciu intronów. Wariant przedstawiony jest trzy razy, a na każdej kopii zaprezentowano jeden z przykładowych sekwencjonowanych fragmentów. Sekwencjonowanie odbywa się już na mRNA, gdzie występuje konkretna izoforma genu, niekoniecznie jest to wariant pierwszy.



Rysunek 1.2: Przykładowy wariant genu, zawierający trzy eksony, wraz z pozycjami w DNA, oraz trzy przykładowe sekwencjonowane fragmenty genu (kolor żółty z pomarańczowym), wraz z lewymi i prawymi odczytami (kolor pomarańczowy).

Żółtym i pomarańczowym kolorem zaznaczone są sekwencjonowane fragmenty, natomiast kolorem pomarańczowym - lewy oraz prawy odczyt. Przyrównując odczyty do genomu referencyjnego otrzymujemy pozycje nukleotydów, na których znajduje się odczyt w genie. Tablica 1.1 zawiera te pozycje oraz dodatkowo **ścieżkę** fragmentu, czyli dwa podzbiory eksonów, o które zahacza odpowiednio lewy i prawy odczyt danego fragmentu.

	lewy odczyt	prawy odczyt	ścieżka
Fragment 1	110-185	200-274	{1}, {1}
Fragment 2	361-400; 1001-1035	2011-2085	{1,2}, {3}
Fragment 3	301-375	1021-1095	{1}, {2}
...			

Tablica 1.1: Przykładowe odczyty dla trzech sekwencjonowanych fragmentów genu wraz ze ścieżkami, czyli dwoma podzbiorami eksonów, o które zahacza odpowiednio lewy i prawy odczyt danego fragmentu.

Oba końce pierwszego fragmentu należą do eksonu 1. Jako że wszystkie warianty zawierały ekson 1, to ten fragment mógł być wygenerowany przez dowolny wariant. Nie jesteśmy w stanie stwierdzić, czy w izoformie, z której pochodzi wariant występują eksony 2 i 3.

Dla drugiego fragmentu, lewy odczyt należy do eksonów 1 i 2, a prawy do 3, więc ten fragment mógł być wygenerowany tylko z wariantu 1. Wynika to z faktu, iż wiemy, że zahacza on o wszystkie eksony, więc izoforma, z której pochodzi, również musi zawierać wszystkie eksony.

Fragment 3 mógł być wygenerowany z wariantu 1 lub 3, ponieważ izoforma, z której pochodzi musi zawierać eksony 1 i 2.

W praktyce większość genów jest dłuższa i ma bardziej skomplikowane wzorce splicingowe. Idealnie byłoby, gdyby każdy fragment, identyfikowany przez parę odczytów, mógł być unikalnie przydzielony do wariantu. Nie jest to możliwe, ale można określić prawdopodobieństwo, z jakim fragment, z konkretną parą odczytów, pochodzi z danego wariantu. Na przykład odczyty z fragmentu 3 mogą pochodzić z wariantu 1 i 3, ale prawdopodobieństwa, że każdy z wariantów wygenerował fragment z taką parą odczytów, są różne [5].

Pośrednim celem badania jest estymacja tych prawdopodobieństw, tj. poznanie rozkładu wariantów każdego genu, dla każdego pomiaru w danych. Do tego celu posłużymy się algorytmem Casper, przedstawionym w rozdziale drugim. Ostatecznie chcemy zidentyfikować te geny, dla których rozkłady wariantów splicingowych istotnie różnią się między dwoma grupami pomiarów. Przykładowo, może istnieć gen, który w jednej grupie występuje głównie w pierwszym wariancie, natomiast w drugiej grupie pierwszy wariant jest w znikomej ilości. Metodami do identyfikacji takich genów, na podstawie oszacowanych algorytmem Casper rozkładów izoform, zajmiemy się w rozdziale trzecim.

Rozdział 2

Algorytm Casper

Algorytm Casper wykorzystujemy do estymacji prawdopodobieństw występowania wariantów splicingowych. Wykonujemy go dla każdego genu oraz pomiaru osobno. Następnie na podstawie wyników otrzymanych z zastosowania tego algorytmu, porównujemy ze sobą dwie grupy komórek, pod względem rozkładu występowania poszczególnych wariantów.

Podsumujmy dane ze względu na sekwencje eksonów odwiedzanych przez poszczególne fragmenty (tzw. ścieżki, czyli dwa zbiory eksonów, o które zahacza odpowiednio lewy i prawy odczyt fragmentu). Przykład takich danych przedstawiono w tablicy 2.1.

ścieżka	liczność
{1}, {1}	2824
{2}, {2}	105
{3}, {3}	5042
{1}, {2}	27
{1}, {1,2}	423
{1}, {3}	127
{2,3}, {3}	394
{1,2}, {3}	2
{1}, {2,3}	13

Tablica 2.1: Przykładowe liczności fragmentów mających poszczególne ścieżki. Ścieżka fragmentu to dwa zbiory eksonów, o które zahacza odpowiednio lewy i prawy odczyt tego fragmentu. Np.{1}, {1,2} oznacza, że lewy odczyt fragmentu pasuje do eksonu 1, a prawy odczyt kawałkiem do eksonu 1 i kawałkiem do eksonu 2. Ścieżka {1,2}, {3} oznacza, iż lewy odczyt fragmentu został dopasowany częściowo do eksonu 1, a częściowo do eksonu 2 oraz prawy odczyt do eksonu 3.

Tego typu dane, tzn. liczności fragmentów odwiedzających poszczególne sekwencje eksonów, są danymi wejściowymi do algorytmu.

Zanim przejdziemy do sformułowania modelu w algorytmie Casper oraz jego opisu, w kolejnych dwóch podrozdziałach wprowadzamy metodologię, która zostanie wykorzystana w algorytmie.

2.1. Algorytm EM

Do oszacowania parametrów modelu, który zostanie zaproponowany w rozdziale 2.3, stosujemy estymator MAP (ang. Maximum a Posteriori). Jest on szacowany za pomocą algorytmu EM (ang. Expectation Maximization), który jest skuteczną metodą iteracyjnego obliczania estymatorów maksymalnego a-posteriori (MAP) i największej wiarogodności (ML, ang. Maximum Likelihood), przy obecności ukrytych zmiennych (czyli w przypadku niekompletnych danych, między innymi w przypadku mieszanin rozkładów). Bardziej podstawowym i częściej stosowanym jest kryterium ML, natomiast kryterium MAP pozwala na wykorzystanie wiedzy a-priori, dotyczącej rozkładu estymowanych parametrów.

Metodologia algorytmu EM opiera się na przeformułowaniu problemu z niekompletnymi danymi, w terminach problemu z kompletnymi danymi, który jest prostszy do rozwiązania. Należy również ustalić związek pomiędzy funkcją wiarogodności lub a-posteriori tych dwóch problemów i użyć prostszej, pod względem obliczeniowym estymacji, do rozwiązania problemu z danymi kompletnymi.

Algorytm składa się z dwóch kroków: E-Expectation i M-Maximization. Krok E polega na stworzeniu danych dla problemu z danymi kompletnymi, przy użyciu zaobserwowanych danych niekompletnych, tak aby możliwe było wykonanie prostszego kroku M dla danych kompletnych. W kroku E tworzona jest funkcja wiarogodności (lub a-posteriori) dla problemu z danymi kompletnymi. Opiera się ona częściowo na nieobserwowanych danych, więc jest ona zastępowana przez jej warunkową wartość oczekiwaną względem obserwowanych danych. Krok E jest wykonywany przy użyciu bieżących wartości dla nieznanych parametrów. W kroku M szukamy maksimum utworzonej w kroku E funkcji. Rozpoczynając od pewnej wartości początkowej parametrów, powtarzamy kroki E i M, aż do uzyskania zbieżności [7, str. 132].

Poniżej przedstawiamy działanie algorytmu EM dla kryterium ML, a później przekształcimy kroki dla kryterium MAP. Podrozdziały te opierają się na [6] i [7].

2.1.1. Algorytm EM maksymalizujący funkcję wiarogodności

Rozważmy sytuację, w której mamy zbiór obserwowanych wartości x. Dane te możemy traktować jako realizację zmiennej losowej X, o funkcji gęstości $f(x|\theta)$, gdzie θ jest parametrem modelu, należącym do przestrzeni Θ, który chcemy estymować.

Estymator największej wiarogodności (ML) definiuje się poprzez brzegową funkcję wiarogodności, jako

$$\hat{\theta}_{ML} = \text{argmax}_\theta\Big(f(x|\theta)\Big) = \text{argmax}_\theta\Big(\log f(x|\theta)\Big). \tag{2.1}$$

Rozważamy również ukrytą zmienną losową Z, czyli taką, która nie jest obserwowana. Zmienną losową, której realizacją są kompletne dane oznaczmy przez $Y = (X, Z)$. Funkcję gęstości Y oznaczmy przez $g(y|\theta) = g(x, z|\theta)$. $f(x|\theta)$ nazywamy brzegową funkcją wiarogodności, ponieważ jest zdefiniowana dla niekompletnych danych, natomiast $g(y|\theta)$ jest pełną funkcją wiarogodności.

Logarytm funkcji wiarogodności danych kompletnych ma postać

$$\log g(y|\theta) = \log g(x, z|\theta). \tag{2.2}$$

Związek między funkcjami gęstości danych kompletnych i niekompletnych ma postać

$$f(x|\theta) = \int g(x, z|\theta)dz. \tag{2.3}$$

Jeśli przestrzeń, z której wartości przyjmuje Z, jest dyskretna, całkę należy zastąpić sumą po wszystkich możliwych wartościach.

Algorytm EM znajduje estymator największej wiarogodności, zdefiniowany dla niekompletnych danych w (2.1), pośrednio, wykorzystując funkcję log-wiarogodności kompletnych danych (2.2). Funkcja ta jest nieobserwowalna, więc jest zastępowana przez jej warunkową wartość oczekiwaną względem wektora X, przy użyciu bieżącej wartości parametru θ (ozn. $\theta^{(t)}$).

Niech

$$Q(\theta, \theta_0) = E\Big(\log g(Y|\theta)\Big|X = x, \theta = \theta_0\Big). \tag{2.4}$$

Niech $\theta^{(0)}$ będzie wartością inicjalizacyjną θ. Dla t=1,2,3... kroki algorytmu EM są następujące:

1. **Krok E:** Obliczamy

$$Q(\theta, \theta^{(t)}) = E\Big(\log g(Y|\theta)\Big|X = x, \theta = \theta^{(t)}\Big) \tag{2.5}$$

2. **Krok M:** Maksymalizujemy $Q(\theta, \theta^{(t)})$ względem θ, po całej przestrzeni parametrów Θ, tzn. wybieramy $\theta^{(t+1)}$ takie, że

$$Q(\theta^{(t+1)}, \theta^{(t)}) \geqslant Q(\theta, \theta^{(t)}) \quad \forall\theta \in \Theta. \tag{2.6}$$

Dla zadanego $\epsilon > 0$ powtarzamy kroki, aż do

$$\log f(x|\theta^{(t+1)}) - \log f(x|\theta^{(t)}) < \epsilon$$

lub gdy liczba iteracji przekroczy maksymalną założoną wartość.

2.1.2. Algorytm EM maksymalizujący funkcję a-posteriori

Załóżmy dodatkowo rozkład a-priori parametru θ: $p(\theta)$.

Estymator maksymalnego a-posteriori (MAP) definiuje się, analogicznie jak poprzednio, poprzez brzegową funkcję a-posteriori, jako

$$\hat{\theta}_{MAP} = \text{argmax}_\theta\Big(p(\theta|x)\Big) = \text{argmax}_\theta\Big(\log p(\theta|x)\Big). \tag{2.7}$$

Rozkład a-posteriori $p(\theta|x)$ z twierdzenia Bayes'a możemy zapisać jako

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int_{\mu\in\Theta} f(x|\mu)p(\mu)d\mu}. \tag{2.8}$$

Mianownik w (2.8) jest niezależny od θ, więc maksymalizacja $p(\theta|x)$ jest równoważna maksymalizacji $f(x|\theta)p(\theta)$. Oznacza to, iż estymator MAP w (2.7) możemy równoważnie zapisać jako

$$\hat{\theta}_{MAP} = \text{argmax}_\theta\Big(\log f(x|\theta) + \log p(\theta)\Big). \tag{2.9}$$

Algorytm EM w prezentowanej wersji znajduje estymator MAP, zdefiniowany dla niekompletnych danych w (2.7), również pośrednio, wykorzystując funkcję a-posteriori kompletnych danych $p_c(z, \theta|x)$. Funkcja ta jest nieobserwowalna, więc jest zastępowana przez jej warunkową wartość oczekiwaną względem wektora X, przy użyciu bieżącej wartości parametru θ (oznaczmy przez $\theta^{(t)}$), tzn.

$$E\Big(\log p_c(Z, \theta|X)\Big|X = x, \theta = \theta^{(t)}\Big) =$$

$$= E\Big(\log g(Z, X|\theta)\Big|X = x, \theta = \theta^{(t)}\Big) + \log p(\theta). \tag{2.10}$$

Analogicznie, jak w przypadku algorytmu EM maksymalizującego funkcję wiarogodności (patrz (2.5),(2.6)), korzystając z (2.10), kroki algorytmu są następujące:

1. **Krok E:** Obliczamy

$$R(\theta, \theta^{(t)}) = E\Big(\log g(Y|\theta)\Big|X = x, \theta = \theta^{(t)}\Big) + \log p(\theta) = \tag{2.11}$$

$$= Q(\theta, \theta^{(t)}) + \log p(\theta)$$

2. **Krok M:** Maksymalizujemy $R(\theta, \theta^{(t)})$ względem θ, po całej przestrzeni parametrów Θ, tzn. wybieramy $\theta^{(t+1)}$ takie, że

$$R(\theta^{(t+1)}, \theta^{(t)}) \geqslant R(\theta, \theta^{(t)}) \quad \forall\theta \in \Theta. \tag{2.12}$$

Dla zadanego $\epsilon > 0$ powtarzamy kroki, aż do

$$\log p(\theta^{(t+1)}|x) - \log p(\theta^{(t)}|x) < \epsilon$$

lub gdy liczba iteracji przekroczy maksymalną założoną wartość.

2.2. Rozkład Dirichleta

Rozkład Dirichleta jest użyty w algorytmie Casper, jako rozkład a-priori estymowanych parametrów. Poniżej przedstawione są podstawowe informacje o tym rozkładzie. Podrozdział opiera się na [8].

Rozkład Beta

Ciągła zmienna losowa X ma **rozkład Beta** z dodatnimi parametrami α i β, jeżeli jej gęstość wyraża się przez

$$f(x) = \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0;1)}(x),$$

gdzie $\beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, a $\Gamma()$ jest funkcją gamma.

Wartość oczekiwana i wariancja rozkładu Beta wynoszą odpowiednio

$$E(X) = \frac{\alpha}{\alpha+\beta}, \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Gęstość rozkładu Beta jest dodatnia jedynie dla przedziału $(0;1)$, więc rozkład ten jest używany do modelowania danych w postaci proporcji.

Rozkład Dirichleta

Rozkład Dirichleta jest wielowymiarowym uogólnieniem rozkładu Beta. k-wymiarowy, ciągły wektor losowy $\mathbf{X} = (X_1, X_2, ..., X_k)$ ma rozkład Dirichleta z dodatnimi parametrami $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$, jeżeli jego gęstość jest postaci

$$f(\mathbf{x}) = \begin{cases} \frac{1}{\beta(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, & \text{jeżeli } \sum_{i=1}^{k} x_i = 1, x_i > 0 \quad \forall i, \\ 0, & \text{wpp}, \end{cases} \tag{2.13}$$

gdzie $\beta(\alpha) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$ oraz $\mathbf{x} = (x_1, x_2, ..., x_k)$.

Gęstość jest dodatnia na (k-1)-wymiarowym sympleksie punktów $\mathbf{x} = (x_1, x_2, ..., x_k)$ w przestrzeni k-wymiarowej takich, że $x_i > 0$ dla $i = 1, 2, ..., k$ oraz $\sum_{i=1}^{k} x_i = 1$.

Wartość oczekiwana, wariancja oraz kowariancja rozkładu Dirichleta wynoszą odpowiednio:

$$E(X_i) = \frac{\alpha_i}{\sum_{l=1}^{k} \alpha_l}, \quad i = 1, 2, ..., k, \tag{2.14}$$

$$Var(X_i) = \frac{(\sum_{l=1}^{k} \alpha_l)(\sum_{l=1}^{k} \alpha_l - \alpha_i)}{(\sum_{l=1}^{k} \alpha_l)^2 (1 + \sum_{l=1}^{k} \alpha_l)}, \quad i = 1, 2, ..., k,$$

$$Cov(X_i, X_j) = \frac{-\alpha_i \alpha_j}{(\sum_{l=1}^{k} \alpha_l)^2 (1 + \sum_{l=1}^{k} \alpha_l)}, \quad i, j = 1, 2, ..., k; i \neq j.$$

Jeśli zmienna losowa X ma rozkład Beta z parametrami α, β, to wektor $\mathbf{Y} = (X, 1 - X)$ ma rozkład Dirichleta z parametrami α, β. Dodatkowo, jeżeli $\mathbf{X} = (X_1, X_2, ..., X_k)$ ma rozkład Dirichleta z parametrami $\alpha_1, \alpha_2, ..., \alpha_k$, to rozkład brzegowy i-tej składowej X_i jest rozkładem Beta z parametrami $\alpha_i, \sum_{l=1}^{k} \alpha_l - \alpha_i$. Z tego powodu rozkład Dirichleta nazywamy wielowymiarowym rozkładem Beta.

Rozkład Dirichleta jest często używany wraz z twierdzeniem Bayesa, jako rozkład a-priori dla parametrów dyskretnych rozkładów prawdopodobieństwa. Skorzystamy z niego w kolejnych rozdziałach.

2.3. Model w algorytmie Casper

Formułujemy model na poziomie genu. Wnioskowanie przeprowadzamy osobno dla każdego genu.

Rozważmy gen z E eksonami, gdzie $E \in \{1, 2, ..., \infty\}$. Zbiór wariantów splicingowych oznaczamy $\nu \in 2^{\{1,2,...,E\}}$, a jego liczność $|\nu|$. Każdy wariant (ozn. d) jest charakteryzowany przez rosnącą sekwencję liczb naturalnych $i_1, i_2, ...$, które oznaczają eksony należące do niego. Przykładowo $d = \{1, 2, 4\}$ oznacza, że wariant d zawiera tylko eksony oznaczone numerami 1, 2 i 4.

Rozważmy również sekencjonowany obustronnie fragment. Niech k będzie liczbą eksonów odwiedzanych przez lewy odczyt, a k' - przez prawy odczyt fragmentu. Oznaczamy ścieżkę, czyli sekwencję eksonów odwiedzanych przez fragment, jako $\iota = (\iota_l; \iota_r)$, gdzie $\iota_l = \{i_j, ..., i_{j+k}\}$, $\iota_r = \{i_{j'}, ..., i_{j'+k'}\}$. ι_l oznacza eksony, o które zahacza lewy odczyt, natomiast ι_r eksony, które zawiera prawy odczyt. Niech P^* będzie zbiorem wszystkich możliwych ścieżek, a P jego podzbiorem - zbiorem obserwowanych ścieżek.

Obserwowane dane są realizacją zmiennej losowej $Y = (Y_1, ..., Y_N)$, gdzie N jest liczbą sparowanych odczytów, a $Y_i \in \{1, ..., |P^*|\}$ oznacza ścieżkę, odpowiadającą i-tej parze odczytów. Formalnie, Y_i pochodzi z mieszaniny $|\nu|$ dyskretnych rozkładów prawdopodobieństwa, gdzie każda składowa odpowiada innemu wariantowi splicingowemu. i-ty dyskretny rozkład, wchodzący w skład mieszaniny, jest określony na przestrzeni $P_i \subset \{1, ...|P^*|\}$, zawierającej ścieżki, które mogły być wygenerowane przez i-ty wariant. Wagi mieszaniny $\pi = (\pi_1, ...\pi_{|\nu|})$ dają proporcję odczytów generowanych przez każdy wariant. Dla i-tej pary odczytów prawdopodobieństwo ścieżki y_i pod warunkiem wag wynosi

$$P(Y_i = y_i|\pi) = \sum_{d=1}^{|\nu|} p_{y_i d}\pi_d,$$

gdzie $p_{y_i d} = P(Y_i = y_i|\delta_i = d)$ jest prawdopodobieństwem ścieżki y_i pod warunkiem wariantu d, a δ_i jest nieobserwowaną zmienną oznaczającą wariant, z którego pochodzi Y_i.

Zakładając, że każdy sparowany fragment jest obserwowany niezależnie, mamy:

$$P(Y = y|\pi) = \prod_{i=1}^{N} \sum_{d=1}^{|\nu|} p_{y_i d}\pi_d = \prod_{k=1}^{|P|} \Big(\sum_{d=1}^{|\nu|} p_{kd}\pi_d \Big)^{x_k}, \tag{2.15}$$

gdzie $x_k = \sum_{i=1}^{N} I(y_i = k)$ oznacza liczbę odczytów z k-tą ścieżką.

Ostatnia równość w równaniu (2.15) wynika z tego, że $p_{y_i d}\pi_d$ jest takie samo dla wszystkich odczytów z tą samą ścieżką. W związku z tym, mnożenie możemy wykonywać po wszystkich

ścieżkach, natomiast wewnątrz zastosować potęgowanie, gdzie wykładnik odpowiada liczbie odczytów z daną ścieżką.

W celu estymacji parametrów modelu, tzn. wag mieszaniny $\pi = (\pi_1, ...\pi_{|\nu|})$, używamy algorytmu EM maksymalizującego funkcję a-posteriori parametrów (patrz rozdział 2.1.2.). Algorytm ten wymaga założenia prawdopodobieństwa a-priori estymowanych parametrów. W algorytmie Casper zakładamy, prawdopodobieństwo a-priori Dirichleta (patrz rozdział 2.2.)

$$\pi \sim Dir(q_1, ...q_{|\nu|}). \tag{2.16}$$

Szczegółowo opisane kroki algorytmu EM w kontekście prezentowanych danych, wraz z ich wyprowadzeniem, przedstawiamy poniżej.

1. *krok E*

Przypomnijmy, że Y_i jest obserwowaną zmienną oznaczającą ścieżkę odpowiadającą i-tej parze odczytów, a $\delta_i \in \{1, ..., |\nu|\}$ jest nieobserwowaną zmienną oznaczającą wariant, z którego pochodzi i-ta para odczytów, $i = 1, ..., N$.

Rozkład a-posteriori $p(\pi|y, \delta)$ z twierdzenia Bayes'a może być przedstawiony jako

$$p(\pi|y, \delta) = \frac{P(\pi)P(y, \delta|\pi)}{P(y, \delta)}.$$

$P(y, \delta)$ nie zależy od π, więc do maksymalizacji funkcji wiarogodności wystarczy nam licznik wyrażenia. Funkcja log-a-posteriori $\log(p(\pi|y, \delta))$ jest proporcjonalna do

$$p_0(\pi|y, \delta) = \log P(\pi) + \log P(y, \delta|\pi). \tag{2.17}$$

Przyjmując rozkład a-priori π, jako rozkład Dirichleta z parametrami $q = (q_1, ..., q_{|\nu|})$ (patrz (2.13) i (2.16)), mamy

$$P(\pi) = \frac{1}{\beta(q)} \prod_{d=1}^{|\nu|} \pi_d^{q_d - 1}, \tag{2.18}$$

a co za tym idzie

$$\log P(\pi) = \sum_{d=1}^{|\nu|} (q_d - 1) \log(\pi_d) - \log \beta(q). \tag{2.19}$$

Natomiast $\log P(y, \delta|\pi)$ możemy zapisać jako

$$\log P(y, \delta|\pi) = \log \Big(P(y|\pi, \delta)P(\delta|\pi) \Big) =$$

$$= \log\Big(\prod_{i=1}^{N} P(y_i|\pi,\delta_i)P(\delta_i|\pi)\Big) =$$

$$= \sum_{i=1}^{N}\Big(\log(P(y_i|\pi,\delta_i)) + \log(P(\delta_i|\pi))\Big) = \tag{2.20}$$

$$= \sum_{i=1}^{N}\Bigg(\sum_{d=1}^{|\nu|}\log(p_{y_i d})I(\delta_i = d) + \sum_{d=1}^{|\nu|}\log(\pi_d)I(\delta_i = d)\Bigg) =$$

$$= \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} I(\delta_i = d)\Big(\log(p_{y_i d}) + \log(\pi_d)\Big).$$

Z (2.17), (2.19) i (2.20) mamy

$$p_0(\pi|y,\delta) = \sum_{d=1}^{|\nu|}(q_d - 1)\log(\pi_d) - \log\beta(q) + \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} I(\delta_i = d)\Big(\log(p_{y_i d}) + \log(\pi_d)\Big).$$

Wartość oczekiwana $p_0(\pi|y,\delta)$ pod warunkiem y i $\pi = \pi^{(j)}$ wynosi

$$E\Big(p_0(\pi|y,\delta)\Big|y,\pi^{(j)}\Big) = \tag{2.21}$$

$$= \sum_{d=1}^{|\nu|}(q_d - 1)\log(\pi_d) - \log\beta(q) + \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} P(\delta_i = d|y_i,\pi^{(j)})\Big(\log(p_{y_i d}) + \log(\pi_d)\Big).$$

Czynnik $\log\beta(q)$ można pominąć przy maksymalizacji, więc

$$E\Big(p_0(\pi|y,\delta)\Big|y,\pi^{(j)}\Big) \propto \sum_{d=1}^{|\nu|}(q_d - 1)\log(\pi_d) + \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} P(\delta_i = d|y_i,\pi^{(j)})\Big(\log(p_{y_i d}) + \log(\pi_d)\Big). \tag{2.22}$$

Ponieważ $\pi_{|\nu|} = 1 - \sum_{d=1}^{|\nu|-1}\pi_d$, otrzymujemy (2.22) równe

$$\sum_{d=1}^{|\nu|-1}(q_d - 1)\log(\pi_d) + (q_{|\nu|} - 1)\log(1 - \sum_{d=1}^{|\nu|-1}\pi_d) +$$

$$+ \sum_{i=1}^{N}\Bigg(\sum_{d=1}^{|\nu|-1} P(\delta_i = d|y_i,\pi^{(j)})\Big(\log(p_{y_i d}) + \log(\pi_d)\Big) + \tag{2.23}$$

$$+ P(\delta_i = |\nu||y_i,\pi^{(j)})\Big(\log(p_{y_i|\nu|}) + \log(1 - \sum_{d=1}^{|\nu|-1}\pi_d)\Big)\Bigg).$$

2. *krok M*

Chcemy zmaksymalizować wyrażenie (2.23) ze względu na π.

Niech $\gamma_{id} = P(\delta_i = d|y_i, \pi^{(j)})$. Pochodna cząstkowa wyrażenia (2.23) względem π_d wynosi

$$(q_d - 1)\frac{1}{\pi_d} - (q_{|\nu|} - 1)\frac{1}{1 - \sum_{d'=1}^{|\nu|-1} \pi_{d'}} + \sum_{i=1}^{N} \gamma_{id}\frac{1}{\pi_d} - \sum_{i=1}^{N} \gamma_{i|\nu|}\frac{1}{1 - \sum_{d'=1}^{|\nu|-1} \pi_{d'}}. \tag{2.24}$$

Przyrównując pochodne cząstkowe z (2.24) do zera, dla każdego d otrzymujemy

$$\frac{\pi_d}{1 - \sum_{d'=1}^{|\nu|-1} \pi_{d'}} = \frac{q_d - 1 + \sum_{i=1}^{N} \gamma_{id}}{q_{|\nu|} - 1 + \sum_{i=1}^{N} \gamma_{i|\nu|}},$$

z czego wnioskujemy, że

$$\pi_d \propto q_d - 1 + \sum_{i=1}^{N} \gamma_{id}. \tag{2.25}$$

γ_{id} możemy inaczej zapisać jako

$$\begin{gathered} \gamma_{id} = P(\delta_i = d|y_i, \pi^{(j)}) = \\ = P(Y_i = y_i|\delta_i = d, \pi^{(j)})P(\delta_i = d|\pi^{(j)})/P(Y_i = y_i|\pi^{(j)}) = \\ = p_{y_i d}\pi_d^{(j)} / \sum_{d'=1}^{|\nu|} p_{y_i d'}\pi_{d'}^{(j)}. \end{gathered} \tag{2.26}$$

Z (2.25) i (2.26) otrzymujemy

$$\pi_d \propto q_d - 1 + \sum_{i=1}^{N} \frac{p_{y_i d}\pi_d^{(j)}}{\sum_{d'=1}^{|\nu|} p_{y_i d'}\pi_{d'}^{(j)}}.$$

Biorąc $x_k = \sum_{i=1}^{N} I(y_i = k)$, możemy pogrupować wszystkie y_i przyjmujące tą samą wartość i otrzymujemy

$$\pi'_d \propto q_d - 1 + \sum_{k=1}^{|P|} x_k \frac{p_{kd}\pi_d^{(j)}}{\sum_{d'=1}^{|\nu|} p_{kd'}\pi_{d'}^{(j)}}.$$

Na koniec skalujemy π tak, aby $\sum_{d=1}^{|\nu|} \pi_d = 1$.

Za wartość początkową w algorytmie przyjmujemy wartość oczekiwaną rozkładu a-priori (patrz wzór (2.14)).

Kroki algorytmu Casper są więc następujące:

1. inicjalizacja prawdopodobieństw: $\pi_d^{(0)} = \frac{q_d}{\sum_{k=1}^{|\nu|} q_k}$,
2. aktualizacja prawdopodobieństw: $\pi_d^{(j+1)} = q_d - 1 + \sum_{k=1}^{|P|} x_k \frac{p_{kd}\pi_d^{(j)}}{\sum_{d'=1}^{|\nu|} p_{kd'}\pi_{d'}^{(j)}}$,
3. skalowanie prawdopodobieństw, aby $\sum_{d=1}^{|\nu|} \pi_d^{(j+1)} = 1$.

Powtarzamy krok 2. i 3. do stabilizacji, np. $\pi_d^{(j+1)} - \pi_d^{(j)} < \epsilon \ \forall d$. Zauważmy, że p_{kd} i x_k pozostają stałe we wszystkich krokach, więc wystarczy, że są policzone tylko raz.

Rozdział 3

Metody wykrywania różnic w dwóch grupach

Przypomnijmy, że estymacja prawdopodobieństw występowania wariantów splicingowych genów jest tylko pośrednim celem przeprowadzanego badania. Ostatecznym celem jest porównanie ze sobą dwóch grup pomiarów oraz identyfikacja genów, dla których rozkłady wariantów istotnie różnią się między grupami. Algorytm Casper wykonujemy dla każdego genu oraz próbki osobno. Następnie wyniki dla wszystkich próbek łączymy tak, aby móc badać różnice w występowaniu wariantów w dwóch grupach komórek. Przyjmijmy, że dla każdego genu posiadamy macierz wyestymowanych prawdopodobieństw $\hat{\pi}$ występowania wariantów splicingowych, gdzie w wierszach umieszczone są wszystkie możliwe warianty genu, a w kolumnach próbki. Wartości w każdej kolumnie sumują się więc do 1.

Przykładowe macierze dla dwóch hipotetycznych genów, o liczbie wariantów 2 i 4 oraz dla dwóch pomiarów w każdej z grup zostały przedstawione w tablicy 3.1.

gen 1	grupa 1		grupa 2	
	P1	P2	P3	P4
wariant 1.1	0,3	0,2	0,5	0,85
wariant 1.2	0,7	0,8	0,5	0,15

gen 2	grupa 1		grupa 2	
	P1	P2	P3	P4
wariant 2.1	0,3	0,2	0,1	0,15
wariant 2.2	0,2	0,2	0,6	0,15
wariant 2.3	0,15	0,35	0,05	0,1
wariant 2.4	0,35	0,25	0,25	0,6

Tablica 3.1: Przykładowe macierze estymowanych prawdopodobieństw występowania wariantów dla czterech pacjentów w dwóch grupach.

Mając dane macierze, jak w tablicy 3.1, chcemy zidentyfikować geny, które różnią się istotnie rozkładem wariantów między grupami. Proponujemy następujący algorytm wykonywany dla każdego genu:

1. przeprowadź test statystyczny dla każdego wariantu wybranego genu,
2. przeskaluj p-wartości testu, uwzględniając poprawkę Holma na liczbę wariantów genu,
3. jeżeli choć jeden wariant statystycznie różni się między grupami, pod względem wyestymowanego prawdopodobieństwa, to gen uznajemy za różniący się między grupami, pod względem rozkładu wariantów.

Poprawka Holma została wprowadzona w celu kontroli błędu FWER (ang. familly-wise error rate), czyli prawdopodobieństwa odrzucenia co najmniej jednej prawdziwej hipotezy zerowej, ze zbioru testowanych hipotez.

Rozpatrzmy następujące testy statystyczne, mogące służyć do zidentyfikowania genów istotnie różniących się rozkładem wariantów między grupami:

1. klasyczny test t-studenta dla dwóch prób (rozdział 3.1),
2. zmodyfikowany test t-studenta dla dwóch prób zależnych (rozdział 3.2),
3. test Walda oraz test ilorazu wiarogodności, badające istotność efektów stałych w modelach liniowych z efektami losowymi (rozdział 3.3).

Testy z punktów 2 i 3 zostały zaproponowane w celu uwzględniania zależności między pacjentami w danych, tzw. podatności (w obu grupach posiadamy pomiary od tych samych pacjentów - dokładny opis danych znajduje się w rozdziale 4.1). W rozdziale 3.4 przedstawiono porównanie wymienionych testów, w celu wyboru najlepszego.

3.1. Klasyczny test dla dwóch prób

Niech $(X_1, ..., X_{n_X})$ i $(Y_1, ..., Y_{n_Y})$ będą dwiema niezależnymi próbami losowymi, których cechy X i Y mają rozkład normalny, odpowiednio $N(\mu_X, \sigma^2)$, $N(\mu_Y, \sigma^2)$, gdzie wariancja σ^2 jest nieznana.

Weryfikujemy hipotezę zerową

$$H_0 : \quad \mu_X = \mu_Y,$$

względem hipotezy alternatywnej

$$H_1 : \quad \mu_X \neq \mu_Y.$$

Statystyką testową testu t-studenta jest

$$T = \frac{\overline{Y} - \overline{X}}{\sqrt{S^2(\frac{1}{n_X} + \frac{1}{n_Y})}} = \tag{3.1}$$

$$= \frac{\overline{Y} - \overline{X}}{\sqrt{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}} \sqrt{\frac{n_X n_Y (n_X + n_Y - 2)}{n_X + n_Y}},$$

gdzie $\overline{X}$ i $\overline{Y}$ są średnimi, a S_X^2 i S_Y^2 wariancjami próbkowymi z prób X i Y, tzn.

$$\overline{X} = \frac{1}{n_X} \sum_{i=1}^{n_X} X_i, \quad \overline{Y} = \frac{1}{n_Y} \sum_{i=1}^{n_Y} Y_i,$$

$$S_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} (X_i - \overline{X})^2, \quad S_Y^2 = \frac{1}{n_Y - 1} \sum_{i=1}^{n_Y} (Y_i - \overline{Y})^2$$

oraz

$$S^2 = \frac{\sum_{i=1}^{n_X} (X_i - \overline{X})^2 + \sum_{i=1}^{n_Y} (Y_i - \overline{Y})^2}{n_X + n_Y - 2} = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{(n_X - 1) + (n_Y - 1)}.$$

Przy prawdziwości hipotezy zerowej H_0, statystyka T ma rozkład t-studenta $T\,(n_X + n_Y - 2)$.

Obszarem krytycznym testu jest $K = \{t : |t| > t_{n_X+n_Y-2}(1 - \alpha/2)\}$, gdzie t jest realizacją statystyki T, $t_{n_X+n_Y-2}(1 - \alpha/2)$ jest kwantylem rzędu $1 - \alpha/2$ rozkładu $T(n_X + n_Y - 2)$, a α poziomem istotności.

Test t jest testem jednostajnie najmocniejszym do sprawdzania hipotezy $H_0 : \mu_X = \mu_Y$, przeciwko hipotezie $H_1 : \mu_X \neq \mu_Y$ [8, str. 243-245].

Próby losowe w badaniach, które przeprowadzamy, nie są niezależne oraz nie muszą pochodzić z rozkładu normalnego. Dokładniej mówiąc, zarówno wewnątrz jednej z grup, jak i pomiędzy grupami występują zmienne zależne, pochodzące od tego samego pacjenta (szczegółowy opis danych znajduje się w rozdziale 4.1). Dodatkowo, rozkład a-priori obserwacji jest rozkładem Beta (brzegowo, ponieważ w obrębie genu zakładaliśmy wielowymiarowy rozkład Dirichleta), co sugeruje, że wynikowe obserwacje nie pochodzą z rozkładu normalnego. Testy statystyczne badające, czy obserwacje pochodzą z rozkładu normalnego, nie odrzucają hipotezy zerowej. Nieodrzucanie hipotezy zerowej może być jednak spowodowane zbyt małą licznością próby (w grupach mamy 8 i 4 obserwacje). W związku z tym, nie chcemy zakładać normalności rozkładu danych. Ostatecznie, nie zakładamy więc, że statystyka T (3.1) ma rozkład t-studenta, jak by było, gdyby próby pochodziły z rozkładów normalnych i były niezależne. Rozkład statystyki T wyznaczamy symulacyjnie, przyjmując różne założenia, co do rozkładów oraz korelacji w danych. Poniżej znajduje się opis metodologii przeprowadzania symulacji, który zostanie powielony również dla statystyk w rozdziałach 3.2. i 3.3.

Generujemy dane z zatosowaniem dwóch rozkładów: normalnego i Beta. Każdy zestaw danych zawiera 12 liczb, które są podzielone na grupy w stosunku 8:4 oraz przyjmujemy, że pochodzą od 4 pacjentów (są zależne, jak opisano poniżej), w sposób przedstawony w tablicy 3.2, który jest zgodny z rzeczywistym podziałem występującym w analizowanych danych (patrz rozdział 4.1).

numer pomiaru	1	2	3	4	5	6	7	8	9	10	11	12
grupa	1	1	1	1	2	2	2	2	2	2	2	2
pacjent	1	2	3	4	1	1	2	2	3	3	3	4

Tablica 3.2: Podział symulowanych pomiarów na 2 grupy oraz 4 pacjentów.

Dla rozkładu normalnego przyjmujemy wartość oczekiwaną równą $0,5$ oraz wariancję $0,1$ i dodatkowo stosujemy 2 sposoby uwzględniania zależności danych w obrębie tego samego pacjenta:

- po wygenerowaniu danych z rozkładu normalnego, dla każdego pacjenta generujemy liczbę z rozkładu jednostajnego na przedziale

  1) nie generujemy, przyjmujemy 0,
  2) $[-0,2; 0,2]$,
  3) $[-0,4; 0,4]$,
  4) $[-0,6; 0,6]$,
  5) $[-0,8; 0,8]$.

  Następnie wszystkie dane pochodzące od tego pacjenta zmieniamy o wygenerowaną liczbę. Proponujemy taki sposób, ponieważ spodziewamy się, że mogą istnieć pacjenci, u których pewien wariant genu występuje częściej, zarówno w jego komórkach z grupy 1 jak i 2. Chcemy, aby test uwzględniał takie dane i żeby nie zaburzały one analizy.

- generujemy dane z rozkładu normalnego skorelowanego, gdzie nie ma zależności między różnymi pacjentami, natomiast korelacja między danymi w obrębie tego samego pacjenta jest równa

  1) 0,
  2) $0,2$,
  3) $0,4$,
  4) $0,6$,
  5) $0,8$.

Ta metoda również pozwala na uwzględnienie, że dane pochodzące od tego samego pacjenta mogą być zależne.

Dane z rozkładu Beta, również generujemy na dwa sposoby:

- z przyjęciem następujących parametrów:

  1) $\alpha = \beta = 2$,

  2) $\alpha = \beta = 4$,

  3) $\alpha = \beta = 6$,

  4) $\alpha = \beta = 8$,

  5) $\alpha = \beta = 10$.

  Stosujemy różne parametry rozkładu Beta, aby sprawdzić, czy rozkład statystki nie będzie od nich zależny.

- po wygenerowaniu danych z rozkładu Beta z parametrami, jak powyżej, dla każdego pacjenta generujemy liczbę z rozkładu jednostajnego na przedziale

  1) nie generujemy, przyjmujemy 0,

  2) $[-0,2; 0,2]$,

  3) $[-0,4; 0,4]$,

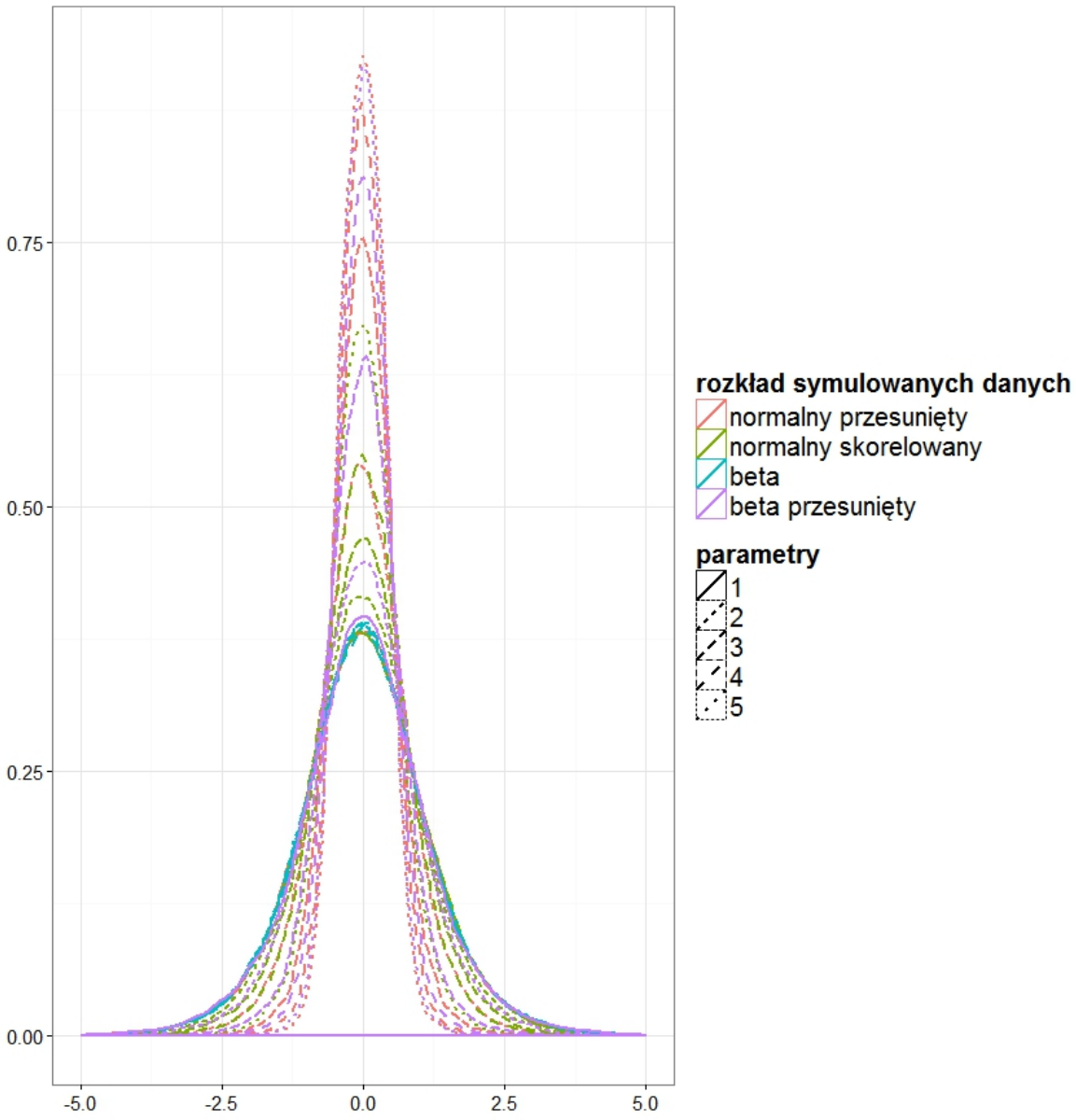
  4) $[-0,6; 0,6]$,

  5) $[-0,8; 0,8]$.

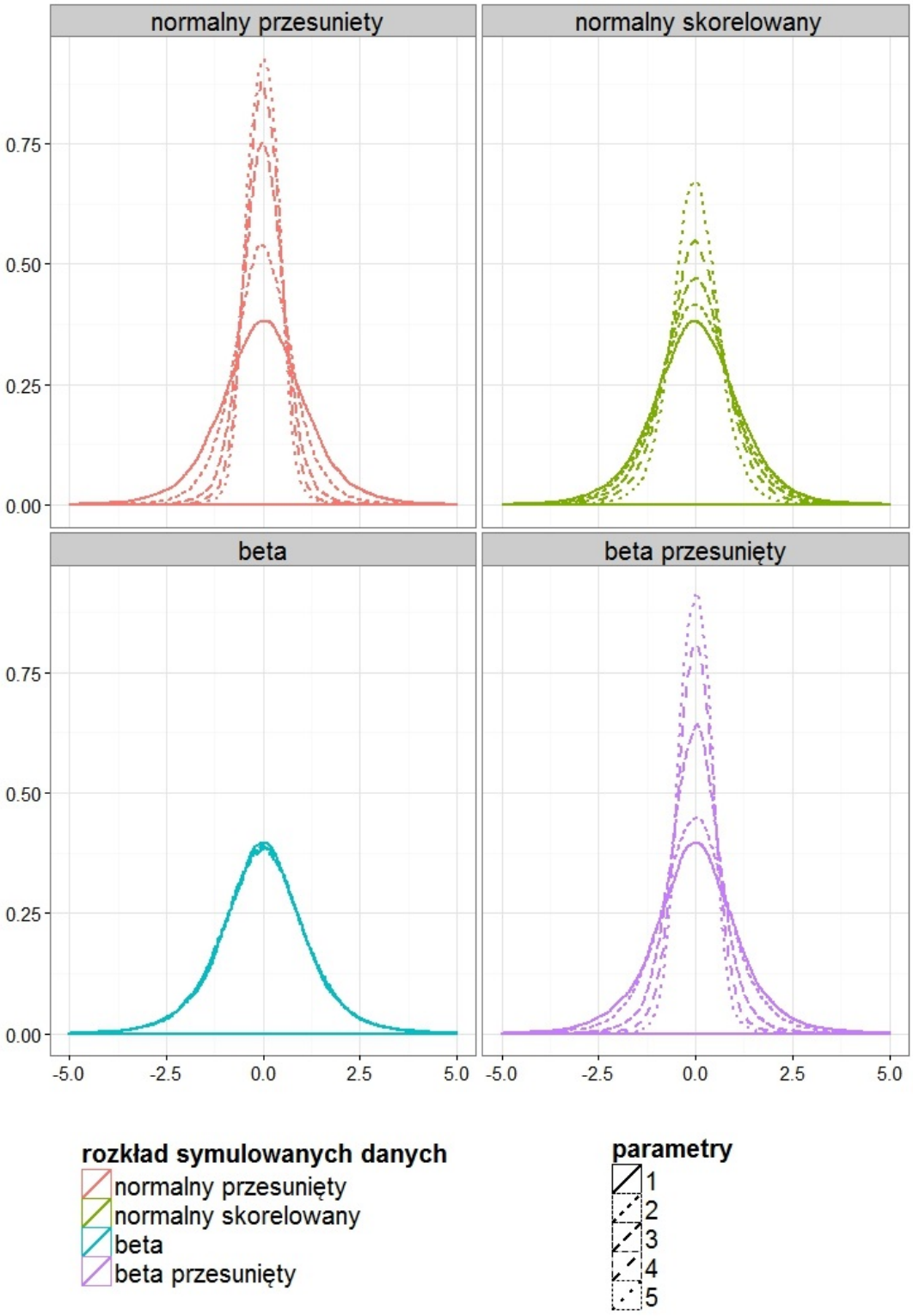
  Następnie wszystkie dane pochodzące od tego pacjenta zmieniamy o wygenerowaną liczbę (uwzględnienie zależności danych w obrębie tego samego pacjenta tak, jak w punkcie pierwszym dla rozkładu normalnego).

Uwzględnienie różnych sił zależności oraz parametrów daje $4 \cdot 5 = 20$ kombinacji dla danych wejściowych. Dla każdego przypadku, dane zawierające 12 liczb generujemy 50 000 razy oraz na podstawie każdego zestawu wyliczamy wartość statystyki T. Pozwala to, dla pojedynczego przypadku generowania danych, otrzymać 50 000 wartości statystyki T przy założeniu, że hipoteza zerowa jest prawdziwa. Na podstawie tych wartości statystyki, wyznaczamy rozkład statystyki T przy założeniu hipotezy zerowej.

Gęstości rozkładu statystyki T przy założeniu hipotezy zerowej, dla wszystkich 20 metod generowania danych, można zobaczyć na rysunku 3.2. Dodatkowo, w celu dokładniejszej obserwacji różnic, wykres rozdzielono na cztery wykresy, w podziale na rozkład generowanych danych (rysunek 3.2).



Rysunek 3.1: Wykres gęstości dla statystyki T testu t-studenta dla dwóch prób (3.1) przy hipotezie zerowej dla różnych metod generowania danych. Próby losowane z 4 rozkładów (na wykresie: rozkład symulowanych danych):
- rozkładu normalnego o średniej 0,5 i odchyleniu standardowym 0,1, dla którego w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry) - kolor czerwony,
- rozkładu normalnego o średniej 0.5 i odchyleniu standardowym 0.1, dla którego w obrębie tego samego pacjenta zmienne są skorelowane z korelacją równą: 1) 0, 2) 0,2, 3) 0,4, 4) 0,6, 5) 0,8 (na wykresie: parametry) - kolor zielony,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry) - kolor niebieski,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry), dla którego dodatkowo w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry)) - kolor fioletowy.



Rysunek 3.2: Wykres gęstości dla statystyki T testu t-studenta dla dwóch prób (3.1) przy hipotezie zerowej dla różnych metod generowania danych. Dokładniejszy opis pod rysunkiem 3.1.

Na podstawie wyznaczonych rozkładów statystyki T, budujemy obszary krytyczne testu do weryfikacji hipotezy $H_0 : \mu_X = \mu_Y$, na poziomie istotności 10%. Przy wyznaczaniu obszarów krytycznych zakładamy symetrię rozkładów statystyki testowej. Obszary krytyczne można zobaczyć w tablicy 3.3.

rozkład symulowanych danych	parametry (możliwe przesunięcie, korelacja, α i β)	obszar krytyczny dla poziomu istotności 10%
normalny przesunięty	1) bez przesunięcia	$(-\infty; -1,91) \cup (1,91; +\infty)$
	2) $[-0,2; 0,2]$	$(-\infty; -1,36) \cup (1,36; +\infty)$
	3) $[-0,4; 0,4]$	$(-\infty; -0,94) \cup (0,94; +\infty)$
	4) $[-0,6; 0,6]$	$(-\infty; -0,75) \cup (0,75; +\infty)$
	5) $[-0,8; 0,8]$	$(-\infty; -0,66) \cup (0,66; +\infty)$
normalny skorelowany	1) $\rho = 0$	$(-\infty; -1,90) \cup (1,90; +\infty)$
	2) $\rho = 0,2$	$(-\infty; -1,75) \cup (1,75; +\infty)$
	3) $\rho = 0,4$	$(-\infty; -1,57) \cup (1,57; +\infty)$
	4) $\rho = 0,6$	$(-\infty; -1,35) \cup (1,35; +\infty)$
	5) $\rho = 0,8$	$(-\infty; -1,06) \cup (1,06; +\infty)$
Beta	1) $\alpha = \beta = 2$	$(-\infty; -1,94) \cup (1,94; +\infty)$
	2) $\alpha = \beta = 4$	$(-\infty; -1,93) \cup (1,93; +\infty)$
	3) $\alpha = \beta = 6$	$(-\infty; -1,92) \cup (1,92; +\infty)$
	4) $\alpha = \beta = 8$	$(-\infty; -1,92) \cup (1,91; +\infty)$
	5) $\alpha = \beta = 10$	$(-\infty; -1,91) \cup (1,91; +\infty)$
Beta przesunięty	1) $\alpha = \beta = 2$, bez przesunięcia	$(-\infty; -1,94) \cup (1,94; +\infty)$
	2) $\alpha = \beta = 4$, $[-0,2; 0,2]$	$(-\infty; -1,64) \cup (1,64; +\infty)$
	3) $\alpha = \beta = 6$, $[-0,4; 0,4]$	$(-\infty; -1,12) \cup (1,12; +\infty)$
	4) $\alpha = \beta = 8$, $[-0,6; 0,6]$	$(-\infty; -0,84) \cup (0,84; +\infty)$
	5) $\alpha = \beta = 10$, $[-0,8; 0,8]$	$(-\infty; -0,69) \cup (0,69; +\infty)$

Tablica 3.3: Obszary krytyczne testów, na poziomie istotności 10%, dla różnych rozkładów danych wejściowych i rożnych parametrów.

Rozkłady statystyki T przy założeniu hipotezy zerowej różnią się, w zależności od sposobu generowania danych wejściowych. W szczególności różnice są widoczne dla danych z korelacją lub przesunięciem w obrębie pacjenta. Im większa korelacja pomiarów dla tego samego pacjenta, tym obszar krytyczny jest większy, czyli częściej odrzucamy hipotezę zerową.

Warto zauważyć, że przy klasycznym teście t-studenta, gdzie zakładamy niezależność obserwacji (czyli statystyka T ma rozkład t-studenta z 10 stopniami swobody), analogiczny obszar krytyczny jest postaci $(-\infty; -1,81) \cup (1,81; +\infty)$.

Dane, które chcemy analizować są zależne. Pomiary od tego samego pacjenta znajdują się zarówno w jednej, jak i w drugiej grupie. Rozkład statystyki T jest natomiast silnie zależny od korelacji między pomiarami. W związku z tym nie możemy użyć tej statystyki do weryfikacji hipotezy o równości średnich.

3.2. Test uwzględniający podatność pacjentów

Zaprezentujemy najpierw teorię dla testu t-studenta dla jednej próby. Niech $(X_1, ..., X_n)$ będzie próbą losową z populacji, której cecha X ma rozkład normalny $N(\mu, \sigma^2)$, gdzie wariancja σ^2 jest nieznana. Weryfikujemy hipotezę zerową

$$H_0: \quad \mu = \mu_0,$$

względem hipotezy alternatywnej

$$H_1: \quad \mu \neq \mu_0.$$

Statystyką testową testu t-studenta dla jednej próby jest

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{S^2}}, \tag{3.2}$$

gdzie $\overline{X}$ jest średnią próbkową, a S^2 wariancją próbkową z próby X.

Przy prawdziwości hipotezy zerowej H_0, statystyka T ma rozkład t-studenta $T(n-1)$. Obszarem krytycznym testu jest $K = \{t : |t| > t_{n-1}(1 - \alpha/2)\}$, gdzie t jest wartością statystyki T, $t_{n-1}(1 - \alpha/2)$ jest kwantylem rzędu $1 - \alpha/2$ rozkładu $T(n-1)$, a α jest poziomem istotności.

Test t jest testem jednostajnie najmocniejszym do sprawdzania hipotezy $H_0 : \mu = \mu_0$, przeciwko hipotezie $H_1 : \mu \neq \mu_0$ [8, str. 243-244].

Poniżej przedstawiamy teorię dotyczącą testu t-studenta dla prób sparowanych. Niech (X_i, Y_i), $i = 1, 2, ..., n$, będą niezależnymi obserwacjami zmiennej losowej (X, Y). Niech $D = Y - X$. Zakłada się, że D jest zmienną losową o rozkładzie normalnym $N(\mu_D, \sigma_D^2)$, przy czym wariancja σ_D^2 jest nieznana. Weryfikujemy hipotezę zerową

$$H_0: \quad \mu_X = \mu_Y,$$

względem hipotezy alternatywnej

$$H_1: \quad \mu_X \neq \mu_Y,$$

gdzie μ_X i μ_Y są wartościami oczekiwanymi populacji o cechach odpowiednio X i Y. Hipotezy te są równoważne hipotezom

$$H_0': \quad \mu_D = 0,$$

$$H_1': \quad \mu_D \neq 0.$$

Statystyką testową testu t-studenta dla prób zależnych jest

$$T = \frac{\sqrt{n}\overline{D}}{\sqrt{S_D^2}}, \tag{3.3}$$

gdzie $\overline{D}$ jest średnią próbkową, a S_D^2 wariancją próbkową zmiennej D. Przy prawdziwości hipotezy zerowej H_0, statystyka T ma rozkład t-studenta $T(n-1)$. Obszarem krytycznym testu jest $K = \{t : |t| > t_{n-1}(1-\alpha/2)\}$, gdzie t jest wartością statystyki T, $t_{n-1}(1-\alpha/2)$ jest kwantylem rzędu $1-\alpha/2$ rozkładu $T(n-1)$, a α jest poziomem istotności [8, str. 246].

W danych, które badamy, występują dwie próby zależne, jednak nie w sposób, jaki jest uwzględniony w teście t-studenta dla prób sparowanych. Nie mamy bowiem par pomiarów, gdzie każdy element pary należy do innej grupy. W naszych danych występują za to pary, trójki lub czwórki pomiarów zależnych, pochodzących od tego samego pacjenta, przy czym zawsze jeden pomiar z tego zbioru należy do grupy pierwszej, a pozostałe do grupy drugiej (patrz tablica 3.2). Użyjemy więc testu, który konstruujemy, jako modyfikację testu t-studenta dla prób sparowanych (którego nie możemy zastosować wprost) z testem t-studenta dla jednej próby.

Niech $(X_1, X_2, ..., X_n)$ i $(Y_{1.1}, ..., Y_{1.n_1}, Y_{2.1}, ..., Y_{2.n_2}, ..., Y_{n.1}, ..., Y_{n.n_n})$ będą zależnymi próbami losowymi, gdzie $(X_1, Y_{1.1}, ..., Y_{1.n_1}), ..., (X_n, Y_{n.1}, ..., Y_{n.n_n})$ są wspomnianymi wcześniej zbiorami pomiarów zależnych. Zmienne pochodzące z tego samego zbioru są zależne, a zmienne pochodzące z dwóch innych zbiorów są niezależne. Niech $D_{ij} = X_i - Y_{i.j}$ dla $i = 1, ...n$, $j = 1, ..., n_i$.

Do zweryfikowania hipotezy zerowej

$$H_0: \quad \mu_D = 0,$$

względem hipotezy alternatywnej

$$H_1: \quad \mu_D \neq 0,$$

gdzie μ_D jest wartością oczekiwaną zmiennej D, używamy statystyki testowej testu t-studenta dla jednej próby, zastosowanego do zmiennej D, tzn.

$$T = \frac{\sqrt{n}\overline{D}}{\sqrt{S_D^2}} \tag{3.4}$$

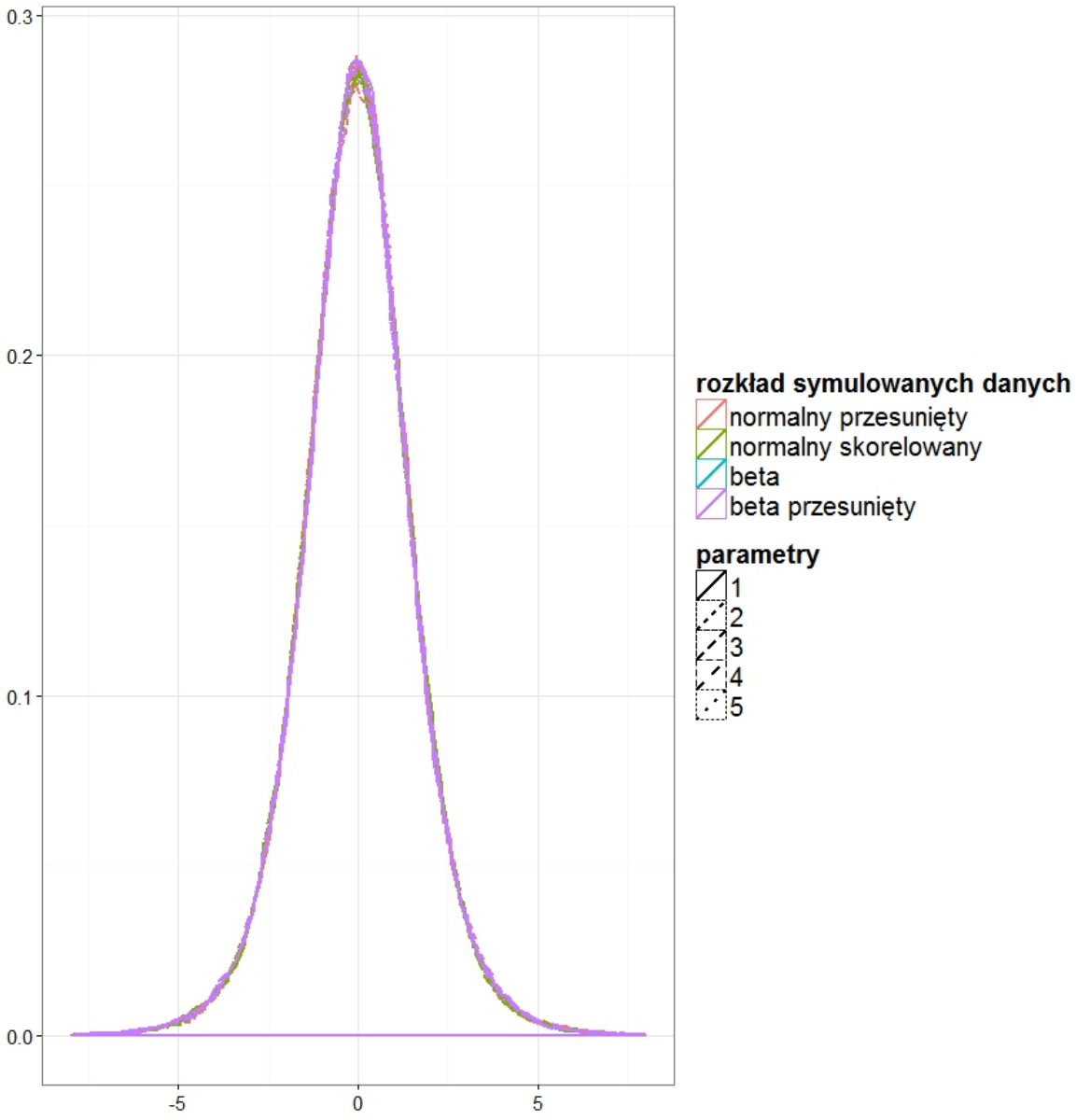
której wzór jest postaci jak (3.3), ale zmienna D jest wyliczana w inny sposób.

Nie możemy przyjąć, że rozkład statystyki T (3.4) jest rozkładem t-studenta, zgodnie z tym, co podaliśmy w opisie testu t-studenta dla jednej próby, ponieważ zmienne D nie są niezależne. $D_{i,j}$ i $D_{k,l}$ są bowiem zależne dla $i = k$. W związku z tym, podobnie jak w rozdziale 3.1, rozkład statystyki T wyznaczamy symulacyjnie.

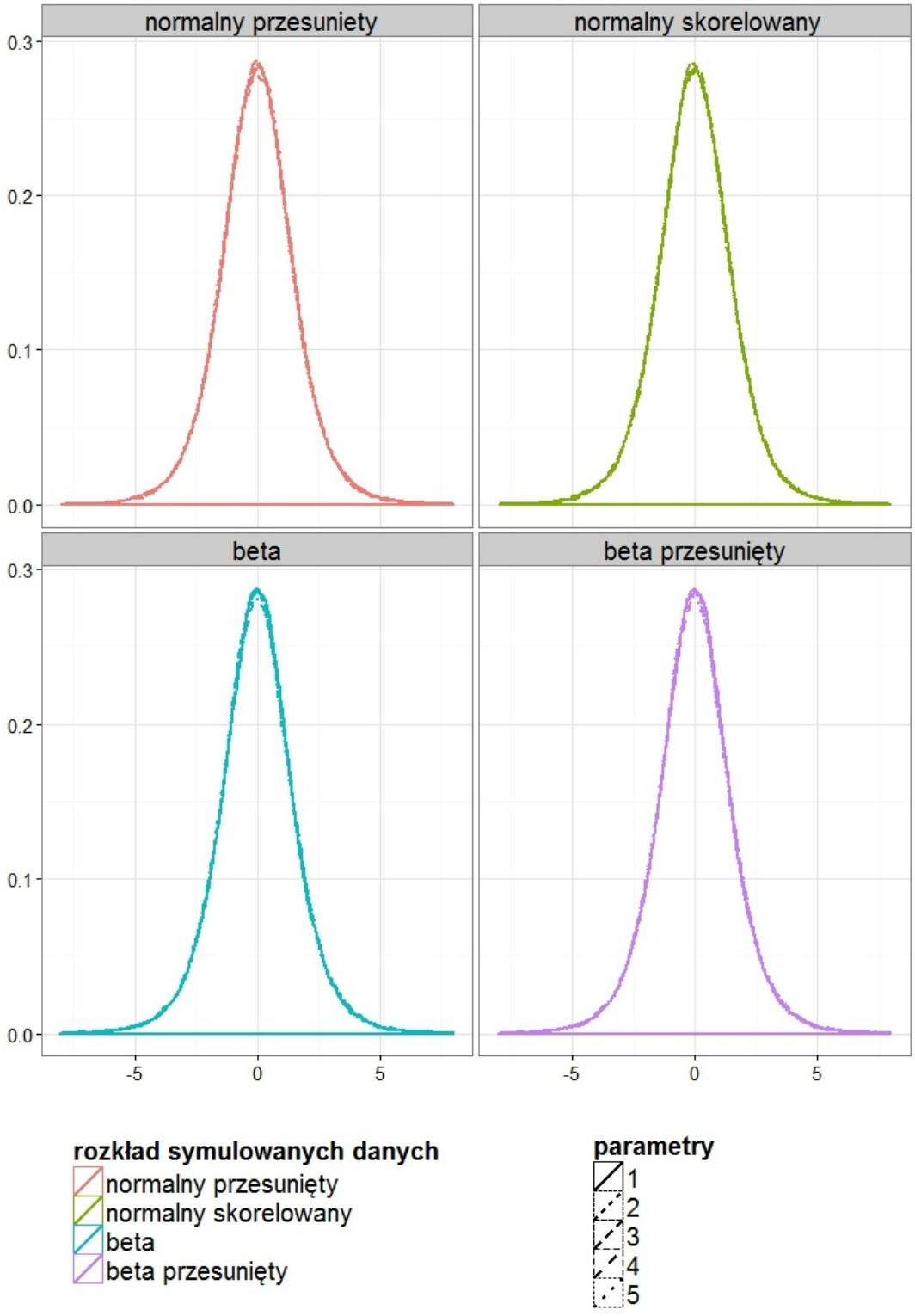
Generujemy dane z zastosowaniem dwóch rozkładów: normalnego i Beta. Każdy zestaw danych zawiera 12 liczb, które są podzielone na grupy w stosunku 8:4 oraz pochodzą od 4 pacjentów (zgodnie z tabelą 3.2). Dla rozkładu normalnego przyjmujemy wartość oczekiwaną równą $0,5$ oraz wariancję $0,1$ i dodatkowo stosujemy 2 sposoby uwzględniania zależności danych w obrębie tego samego pacjenta. Dane z rozkładu Beta również generujemy na dwa sposoby: z zastosowaniem różnych parametrów oraz dodatkowo z uwzględnianiem zależności danych w obrębie tego samego pacjenta. Symulacje przeprowadzamy w taki sam sposób, jak w poprzednim rozdziale. Szczegółowy opis generowania danych do symulacji można znaleźć na stronie 30.

Jedyną różnicą, w stosunku do symulacji z rozdziału 3.1, jest postać statystyki T. Dla każdego z 20 metod generowania danych, otrzymujemy 50 000 wartości statystyki T przy założeniu, że hipoteza zerowa jest prawdziwa. Na podstawie tych wartości statystyki, wyznaczamy rozkład statystyki T przy założeniu hipotezy zerowej.

Gęstości rozkładu statystyki T przy założeniu hipotezy zerowej, dla wszystkich 20 metod generowania danych, można zobaczyć na rysunku 3.3. Dodatkowo w celu dokładniejszej obserwacji różnic, wykres rozdzielono na cztery wykresy, w podziale na rozkład generowanych danych (rysunek 3.4).



Rysunek 3.3: Wykres gęstości dla statystyki T testu t-studenta dla różnic (3.4) przy hipotezie zerowej dla różnych metod generowania danych. Próby losowane z 4 rozkładów (na wykresie: rozkład symulowanych danych):
- rozkładu normalnego o średniej 0,5 i odchyleniu standardowym 0,1, dla którego w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry) - kolor czerwony,
- rozkładu normalnego o średniej 0.5 i odchyleniu standardowym 0.1, dla którego w obrębie tego samego pacjenta zmienne są skorelowane z korelacją równą: 1) 0, 2) 0,2, 3) 0,4, 4) 0,6, 5) 0,8 (na wykresie: parametry) - kolor zielony,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry) - kolor niebieski,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry), dla którego dodatkowo w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry)) - kolor fioletowy.



Rysunek 3.4: Wykres gęstości dla statystyki T testu t-studenta dla różnic (3.4) przy hipotezie zerowej dla różnych metod generowania danych. Dokładniejszy opis pod rysunkiem 3.3.

Na podstawie wyznaczonych rozkładów statystyki T, budujemy obszary krytyczne testu do weryfikacji hipotezy $H_0 : \mu_D = 0$, na poziomie istotności 10%. Przy wyznaczaniu obszarów krytycznych zakładamy symetrię rozkładów statystyki testowej. Obszary krytyczne znajdują się w tablicy 3.4.

rozkład symulowanych danych	parametry (możliwe przesunięcie, korelacja, α i β)	obszar krytyczny dla poziomu istotności 10%
normalny przesunięty	1) bez przesunięcia	$(-\infty; -2,60) \cup (2,60; +\infty)$
	2) $[-0,2; 0,2]$	$(-\infty; -2,59) \cup (2,59; +\infty)$
	3) $[-0,4; 0,4]$	$(-\infty; -2,59) \cup (2,59; +\infty)$
	4) $[-0,6; 0,6]$	$(-\infty; -2,58) \cup (2,58; +\infty)$
	5) $[-0,8; 0,8]$	$(-\infty; -2,59) \cup (2,59; +\infty)$
normalny skorelowany	1) $\rho = 0$	$(-\infty; -2,58) \cup (2,58; +\infty)$
	2) $\rho = 0,2$	$(-\infty; -2,59) \cup (2,59; +\infty)$
	3) $\rho = 0,4$	$(-\infty; -2,58) \cup (2,58; +\infty)$
	4) $\rho = 0,6$	$(-\infty; -2,60) \cup (2,60; +\infty)$
	5) $\rho = 0,8$	$(-\infty; -2,58) \cup (2,58; +\infty)$
Beta	1) $\alpha = \beta = 2$	$(-\infty; -2,62) \cup (2,62; +\infty)$
	2) $\alpha = \beta = 4$	$(-\infty; -2,62) \cup (2,62; +\infty)$
	3) $\alpha = \beta = 6$	$(-\infty; -2,61) \cup (2,61; +\infty)$
	4) $\alpha = \beta = 8$	$(-\infty; -2,61) \cup (2,61; +\infty)$
	5) $\alpha = \beta = 10$	$(-\infty; -2,59) \cup (2,59; +\infty)$
Beta przesunięty	1) $\alpha = \beta = 2$, bez przesunięcia	$(-\infty; -2,62) \cup (2,62; +\infty)$
	2) $\alpha = \beta = 4$, $[-0,2; 0,2]$	$(-\infty; -2,62) \cup (2,62; +\infty)$
	3) $\alpha = \beta = 6$, $[-0,4; 0,4]$	$(-\infty; -2,61) \cup (2,61; +\infty)$
	4) $\alpha = \beta = 8$, $[-0,6; 0,6]$	$(-\infty; -2,61) \cup (2,61; +\infty)$
	5) $\alpha = \beta = 10$, $[-0,8; 0,8]$	$(-\infty; -2,59) \cup (2,59; +\infty)$

Tablica 3.4: Obszary krytyczne testów, na poziomie istotności 10%, dla różnych rozkładów danych wejściowych i rożnych parametrów.

W przeciwieństwie do wyników otrzymanych w rozdziale 3.1, rozkłady statystyki T dla różnych metod generowania danych, nie różnią się znacznie. Wszystkie obszary krytyczne zawierają się między $(-\infty; -2,62) \cup (2,62; +\infty)$ a $(-\infty; -2,58) \cup (2,58; +\infty)$. Na tej podstawie wnioskujemy, że zaprezentowany test nie jest wrażliwy na zmiany rozkładów oraz występowanie korelacji w danych. Będziemy go więc używać do weryfikacji hipotezy o równości średnich dwóch grup w rozdziale 4.

Biorąc wszystkie dane z symulacji, przyjmujemy obszar krytyczny testu postaci $(-\infty; -2,60) \cup (2,60; +\infty)$.

Warto zauważyć, że test t-studenta dla jednej próby o 8 obserwacjach (tyle mamy różnic D), ma obszar krytyczny postaci $(-\infty; -1,89) \cup (1,89; +\infty)$. Zawiera on obszar krytyczny, jaki otrzymujemy symulacyjnie. W związku z tym, zastosowanie testu t-studenta dla różnic, z przyjęciem teoretycznego rozkładu statystyki testowej (czyli zakładając, że różnice są niezależne, gdy w rzeczywistości są zależne), spowodowałoby, że częściej odrzucalibyśmy prawdziwą hipotezę zerową.

3.3. Model liniowy z efektami losowymi uwzględniający podatność pacjentów

W celu sprawdzenia, czy wartości różnią się między grupami, możemy również zastosować model liniowy.

Niech

$$y = X\beta + \epsilon, \tag{3.5}$$

gdzie $y = (y_1, ..., y_n)$ jest wektorem zmiennych objaśnianych, $X_{n\times p}$ jest macierzą kodującą zmienne objaśniające, w której pierwsza kolumna zawiera same jedynki, $\beta = (\beta_1, ..., \beta_p)$ jest wektorem nieznanych parametrów modelu, a $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim N(0, \sigma_\epsilon^2\ \mathrm{I}_{n\times n})$ jest wektorem niezależnych zmiennych losowych o rozkładzie normalnym o średniej 0 i wariancji σ_ϵ^2 (ϵ nazywamy zakłóceniem losowym) [9, str. 3-6].

W naszym zagadnieniu, model konstruujemy na poziomie każdego genu i wariantu osobno, więc $n = 12$ - tyle pomiarów mamy dla każdego przypadku, y jest wektorem wyestymowanych prawdopodobieństw metodą Casper dla omawianego genu i wariantu ($\hat{\pi}$), $p = 2$, X jest macierzą wymiaru 12×2, której pierwsza kolumna zawiera same jedynki, a druga opisuje przynależność do grupy (przyjmijmy 1-grupa 1, 0-grupa 2), $\beta = (\beta_1, \beta_2)$.

Model 3.5 w tym przypadku możemy zapisać jako

$$\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\pi}_3 \\ \hat{\pi}_4 \\ \hat{\pi}_5 \\ \hat{\pi}_6 \\ \hat{\pi}_7 \\ \hat{\pi}_8 \\ \hat{\pi}_9 \\ \hat{\pi}_{10} \\ \hat{\pi}_{11} \\ \hat{\pi}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}, \tag{3.6}$$

W celu sprawdzenia, czy wartości różnią się między grupami, możemy dopasować powyższy model, a następnie przetestować, czy parametr β_2 istotnie różni się od zera.

Przypomnijmy, że nasze dane pochodzą również od czterech różnych pacjentów, przy czym część danych pochodzi od tego samego pacjenta. Należałoby więc uwzględnić dodatkową zmienną objaśniającą w postaci pacjenta, a dokładniej trzy zmienne objaśniające, oznaczające indykator pochodzenia od każdego z trzech pacjentów. Pochodzenie od czwartego pacjenta należy pominąć, ponieważ niepochodzenie od żadnego z trzech pacjentów jest równoważne z pochodzeniem od czwartego.

Z uwzględnieniem zmiennych oznaczających pochodzenie od pacjentów, model jest postaci

$$y = X\beta + \epsilon, \tag{3.7}$$

gdzie $p = 5$, $X_{12\times 5}$ jest macierzą zmiennych objaśniających, przy czym kolumna X_1 zawiera same jedynki, kolumna X_2 określa przynależność do grupy (1-grupa 1, 0-grupa 2), kolumna X_3 określa, czy pomiary pochodzą od pierwszego pacjenta (1-tak, 0-nie), kolumna X_4 - czy pomiary pochodzą od drugiego pacjenta, kolumna X_5 - czy pomiary pochodzą od trzeciego pacjenta oraz $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

Zapisując y jako $\hat{\pi}$ oraz rozpisując macierze, model 3.7 możemy przedstawić jako

$$\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\pi}_3 \\ \hat{\pi}_4 \\ \hat{\pi}_5 \\ \hat{\pi}_6 \\ \hat{\pi}_7 \\ \hat{\pi}_8 \\ \hat{\pi}_9 \\ \hat{\pi}_{10} \\ \hat{\pi}_{11} \\ \hat{\pi}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}, \tag{3.8}$$

Interesuje nas, czy pomiary różnią się między grupami, więc należy testować, czy parametr β_2 różni się istotnie od 0, natomiast parametry $\beta_1, \beta_3, \beta_4, \beta_5$ mogą być dowolne.

W modelu liniowym (3.7) parametry β traktujemy jako nieznane stałe, charakterystyczne dla całej badanej populacji. Jeśli związane są one z kodowaniem zmiennych jakościowych (pacjent), to zakładamy, że obserwujemy wszystkie możliwe poziomy zmiennej. Takie współczynniki nazywamy efektami stałymi.

Jeśli poziomów pewnej zmiennej jest bardzo dużo lub gdy nie obserwujemy wszystkich możliwych poziomów tej zmiennej, możemy zamiast traktować te współczynniki, jako parametry modelu, przyjąć, że są to zmienne losowe z rozkładu, którego parametry będą parametrami modelu. Efekt obserwowanych poziomów zmiennej objaśniającej traktujemy wtedy jako realizację zmiennej losowej, opisującej efekty w całej populacji. Realizacje te nazywamy efektami losowymi. Taką metodologię można zastosować, gdy nie jesteśmy zainteresowani bezpośrednią oceną wartości efektów, a jedynie chcemy uwzględnić ich wpływ lub zbadać zmienność tych efektów w populacji.

Model z efektami losowymi konstruujemy podobnie, jak model liniowy, jednak część efektów zmiennych objaśniających traktujemy jako efekty stałe, a część jako efekty losowe. Model jest postaci

$$y = X\beta + Zu + \epsilon, \tag{3.9}$$

gdzie $y = (y_1, ..., y_n)$ jest wektorem zmiennych objaśnianych, $X_{n\times p}$ jest macierzą kodującą zmienne objaśniające będące efektami stałymi, $Z_{n\times q}$ jest macierzą kodującą zmienne objaśniające odpowiadające efektom losowym, $\beta = (\beta_1, ..., \beta_p)$ jest wektorem nieznanych efektów stałych, $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim N(0, \sigma_\epsilon^2\ \mathrm{I}_{n\times n})$ jest zakłóceniem losowym, a $u = (u_1, ..., u_q) \sim$

$N(0, \sigma^2 D)$ - wektorem zmiennych losowych odpowiadającym efektom losowym, gdzie D jest macierzą wymiaru $q \times q$. Jeśli znamy D, to możemy estymować β, używając metody uogólnionych najmniejszych kwadratów, a jeżeli D jest nieznana, to np. metodą największej wiarogodności [9, str. 144-147].

W naszym zagadnieniu $n = 12$, y jest wektorem wyestymowanych metodą Casper prawdopodobieństw dla omawianego genu i wariantu ($\hat{\pi}$), $p = 2$, X jest macierzą wymiaru 12×2, której pierwsza kolumna zawiera same jedynki, a druga opisuje przynależność do grupy (przyjmujemy 1-grupa 1, 0-grupa 2), $\beta = (\beta_1, \beta_2)$, $q = 4$, Z jest macierzą wymiaru 12×4, w której kolejne kolumny oznaczają pochodzenie pomiarów od kolejnych pacjentów, $u = (u_1, u_2, u_3, u_4) \sim N(0, \sigma^2 D)$, a $\epsilon = (\epsilon_1, ..., \epsilon_{12}) \sim N(0, \sigma_\epsilon^2 \, \mathrm{I}_{12\times 12})$.

Model 3.9 w tym przypadku możemy zapisać jako

$$\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\pi}_3 \\ \hat{\pi}_4 \\ \hat{\pi}_5 \\ \hat{\pi}_6 \\ \hat{\pi}_7 \\ \hat{\pi}_8 \\ \hat{\pi}_9 \\ \hat{\pi}_{10} \\ \hat{\pi}_{11} \\ \hat{\pi}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}, \tag{3.10}$$

W celu sprawdzenia, czy wartości różnią się między grupami, możemy dopasować powyższy model, a następnie przetestować, czy parametr β_2 istotnie rożni się od 0. Interesuje nas tylko wpływ efektu stałego β_2, a nie efektów losowych.

Istnieje kilka rodzajów testów do weryfikacji hipotez dotyczących efektów stałych. Poniżej przedstawimy 2 z nich, służące do testowania hipotezy zerowej

$$H_0 : \quad \beta_i = 0,$$

przeciwko hipotezie alternatywnej

$$H_1 : \quad \beta_i \neq 0.$$

Testy te wykonujemy w celu zbadania istotności parametru β_i dla dowolnego $i = 1, ...p$. W naszym przypadku $p = 2$, a testowanie istotności β_1 nas nie interesuje, więc będzie to tylko test dla jednego parametru.

1) Jeśli $n >> p$, stosuje się test Walda. Statystyką testową tego testu jest

$$T = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}, \tag{3.11}$$

gdzie $\hat{\beta}_i$ jest estymatorem parametru β_i, a $se(\hat{\beta}_i)$ błędem standardowym tego estymatora. Sposób wyznaczania estymatora $\hat{\beta}_i$ oraz jego błędu standardowego można znaleźć w [9, str. 149-159].

Statystyka T, przy prawdziwości hipotezy zerowej H_0, ma asymptotycznie rozkład normalny, a dla mniejszych n jest przybliżana rozkładem t-studenta $T(n - p)$. Rozkład t-studenta nie jest jednak dokładnym rozkładem statystyki testowej, jeżeli nie znamy macierzy D (dla modeli liniowych bez efektów losowych jest dokładnym rozkładem statystyki).

2) Jeśli n nie jest bardzo duże, stosuje się test ilorazu wiarogodności. Statystyką testową tego testu jest

$$T = l - l_{-\beta_i} = \log(L) - \log(L_{-\beta_i}), \tag{3.12}$$

gdzie L i $L_{-\beta_i}$ są wartościami funkcji wiarogodności dla modelu odpowiednio z i-tym efektem i bez niego, a l i $l_{-\beta_i}$ są logarytmami tych wartości.

Statystyka T, przy prawdziwości hipotezy zerowej H_0, ma asymptotycznie rozkład χ^2 z jednym stopniem swobody (ogólnie, z liczbą stopni swobody odpowiadającą różnicy w liczbie parametrów).

Test ilorazu wiarogodności może być stosowany tylko wtedy, gdy jeden model jest zagnieżdżony w drugim. Należy pamiętać o tym przy używaniu testu ilorazu wiarogodności.

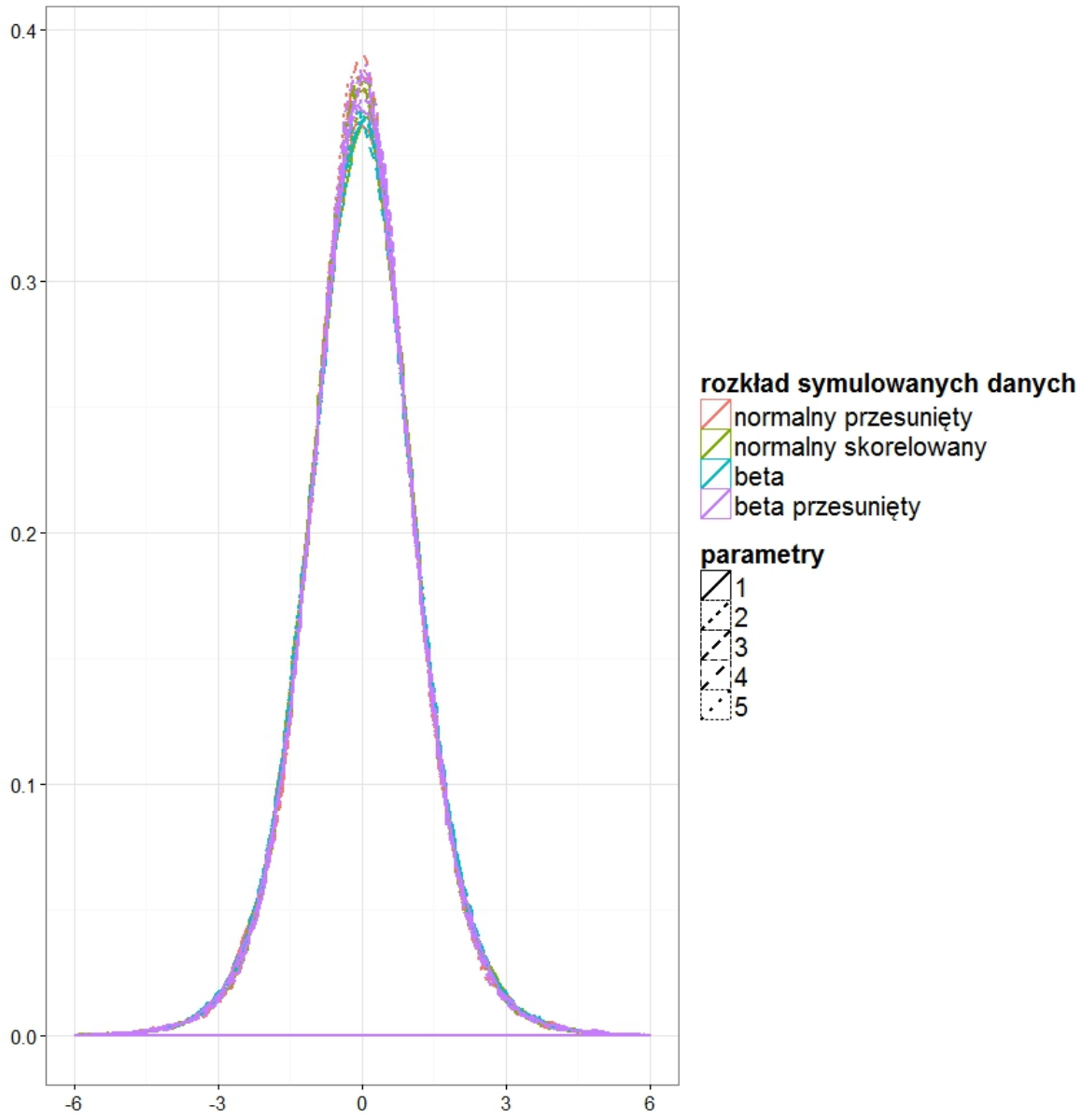
Powyższe dwa testy wykorzystujemy do wykrycia różnic między grupami, jednak w związku z małą liczbą obserwacji oraz skomplikowanymi zależnościami między pomiarami, podobnie jak w rozdziałach 3.1 i 3.2, nie przyjmujemy podanych rozkładów statystyk testowych, a wyznaczamy te rozkłady symulacyjnie.

Generujemy dane z zastosowaniem dwóch rozkładów: normalnego i Beta. Każdy zestaw danych zawiera 12 liczb, które są podzielone na grupy w stosunku 8:4 oraz pochodzą od 4 pacjentów (zgodnie z tabelą 3.2). Dla rozkładu normalnego przyjmujemy wartość oczekiwaną równą $0,5$ oraz wariancję $0,1$ i dodatkowo stosujemy 2 sposoby uwzględniania zależności danych w obrębie tego samego pacjenta. Dane z rozkładu Beta, również generujemy na dwa

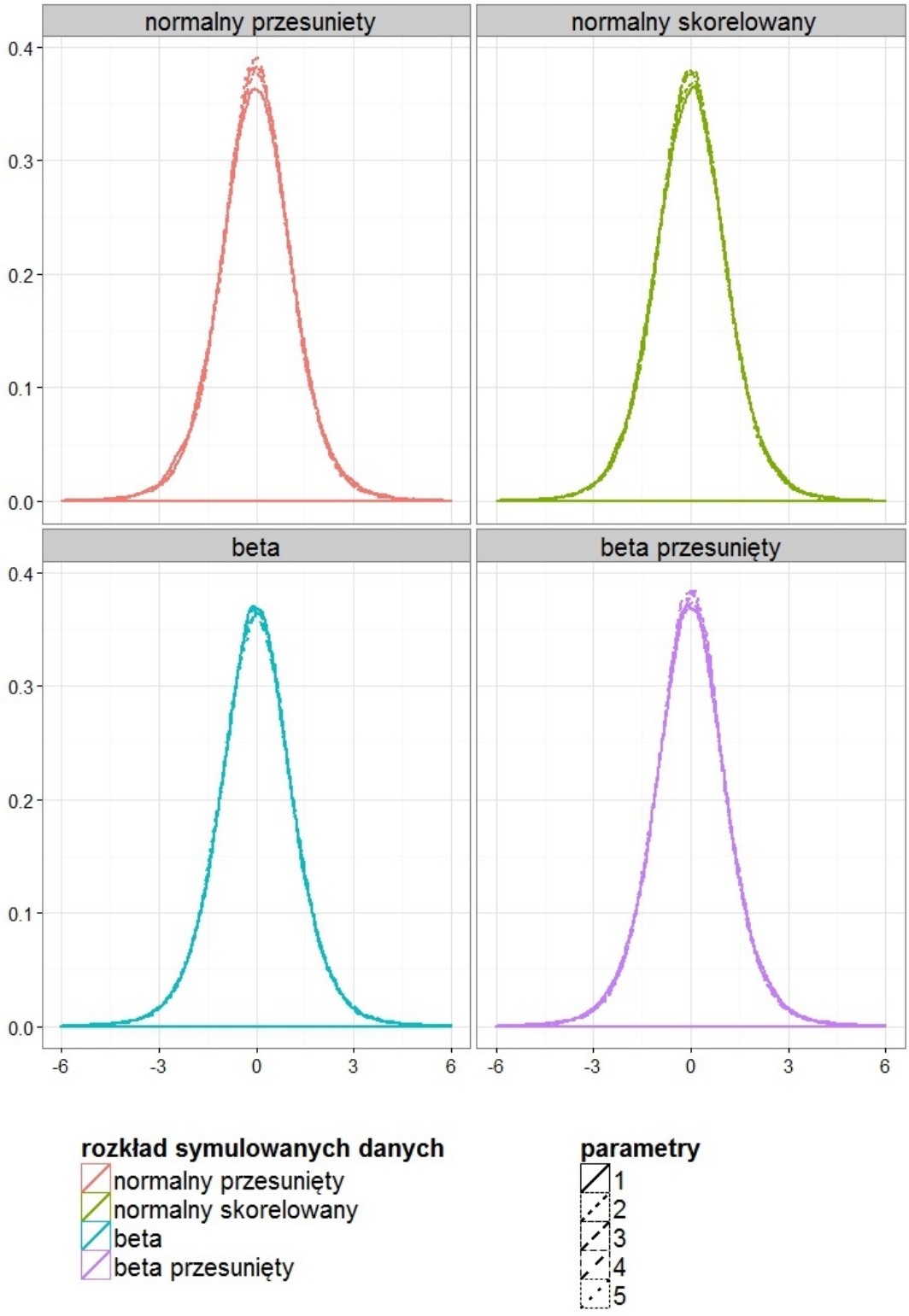
sposoby: z zastosowaniem różnych parametrów oraz dodatkowo z uwzględnianiem zależności danych w obrębie tego samego pacjenta. Symulacje przeprowadzamy w taki sam sposób, jak w rozdziale 3.1. Szczegółowy opis generowania danych do symulacji można znaleźć na stronie 30.

Jedyną różnicą, w stosunku do symulacji z rozdziału 3.1, jest postać statystyki T. Dla każdego z 20 metod generowania danych, otrzymujemy 50 000 wartości statystyki T testu Walda oraz statystyki T testu ilorazu wiarogodności przy założeniu, że hipoteza zerowa jest prawdziwa. Na podstawie tych wartości, wyznaczamy rozkłady statystyk testu Walda i testu ilorazu wiarogodności przy założeniu hipotezy zerowej.

Gęstości rozkładu statystyki T testu Walda przy założeniu hipotezy zerowej, dla wszystkich 20 metod generowania danych, można zobaczyć na rysunku 3.5. Dodatkowo, w celu możliwości dokładniejszej obserwacji różnic, wykres rozdzielono na trzy wykresy, w podziale na rozkład generowanych danych (rysunek 3.6).



Rysunek 3.5: Wykres gęstości dla statystyki T testu Walda (3.11) przy hipotezie zerowej dla różnych metod generowania danych. Próby losowane z 4 rozkładów (na wykresie: rozkład symulowanych danych):
- rozkładu normalnego o średniej 0,5 i odchyleniu standardowym 0,1, dla którego w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry) - kolor czerwony,
- rozkładu normalnego o średniej 0.5 i odchyleniu standardowym 0.1, dla którego w obrębie tego samego pacjenta zmienne są skorelowane z korelacją równą: 1) 0, 2) 0,2, 3) 0,4, 4) 0,6, 5) 0,8 (na wykresie: parametry) - kolor zielony,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry) - kolor niebieski,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry), dla którego dodatkowo w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry)) - kolor fioletowy.



Rysunek 3.6: Wykres gęstości dla statystyki T testu Walda (3.11) przy hipotezie zerowej dla różnych metod generowania danych. Dokładniejszy opis pod rysunkiem 3.5.

Na podstawie wyznaczonych rozkładów statystyki T testu Walda, budujemy obszary krytyczne testu do weryfikacji hipotezy $H_0 : \mu_X = \mu_Y$, na poziomie istotności 10%. Przy wyznaczaniu obszarów krytycznych zakładamy symetrię rozkładów statystyki testowej. Obszary krytyczne można zobaczyć w tablicy 3.5.

rozkład symulowanych danych	parametry (możliwe przesunięcie, korelacja, α i β)	obszar krytyczny dla poziomu istotności 10%
normalny przesunięty	1) bez przesunięcia	$(-\infty; -1,97) \cup (1,97; +\infty)$
	2) $[-0,2; 0,2]$	$(-\infty; -1,90) \cup (1,90; +\infty)$
	3) $[-0,4; 0,4]$	$(-\infty; -1,90) \cup (1,90; +\infty)$
	4) $[-0,6; 0,6]$	$(-\infty; -1,89) \cup (1,89; +\infty)$
	5) $[-0,8; 0,8]$	$(-\infty; -1,89) \cup (1,89; +\infty)$
normalny skorelowany	1) $\rho = 0$	$(-\infty; -1,97) \cup (1,97; +\infty)$
	2) $\rho = 0,2$	$(-\infty; -1,95) \cup (1,95; +\infty)$
	3) $\rho = 0,4$	$(-\infty; -1,92) \cup (1,92; +\infty)$
	4) $\rho = 0,6$	$(-\infty; -1,91) \cup (1,91; +\infty)$
	5) $\rho = 0,8$	$(-\infty; -1,89) \cup (1,89; +\infty)$
Beta	1) $\alpha = \beta = 2$	$(-\infty; -1,97) \cup (1,97; +\infty)$
	2) $\alpha = \beta = 4$	$(-\infty; -1,98) \cup (1,98; +\infty)$
	3) $\alpha = \beta = 6$	$(-\infty; -1,97) \cup (1,97; +\infty)$
	4) $\alpha = \beta = 8$	$(-\infty; -1,98) \cup (1,98; +\infty)$
	5) $\alpha = \beta = 10$	$(-\infty; -1,97) \cup (1,97; +\infty)$
Beta przesunięty	1) $\alpha = \beta = 2$, bez przesunięcia	$(-\infty; -1,97) \cup (1,97; +\infty)$
	2) $\alpha = \beta = 4$, $[-0,2; 0,2]$	$(-\infty; -1,93) \cup (1,93; +\infty)$
	3) $\alpha = \beta = 6$, $[-0,4; 0,4]$	$(-\infty; -1,90) \cup (1,90; +\infty)$
	4) $\alpha = \beta = 8$, $[-0,6; 0,6]$	$(-\infty; -1,90) \cup (1,90; +\infty)$
	5) $\alpha = \beta = 10$, $[-0,8; 0,8]$	$(-\infty; -1,89) \cup (1,89; +\infty)$

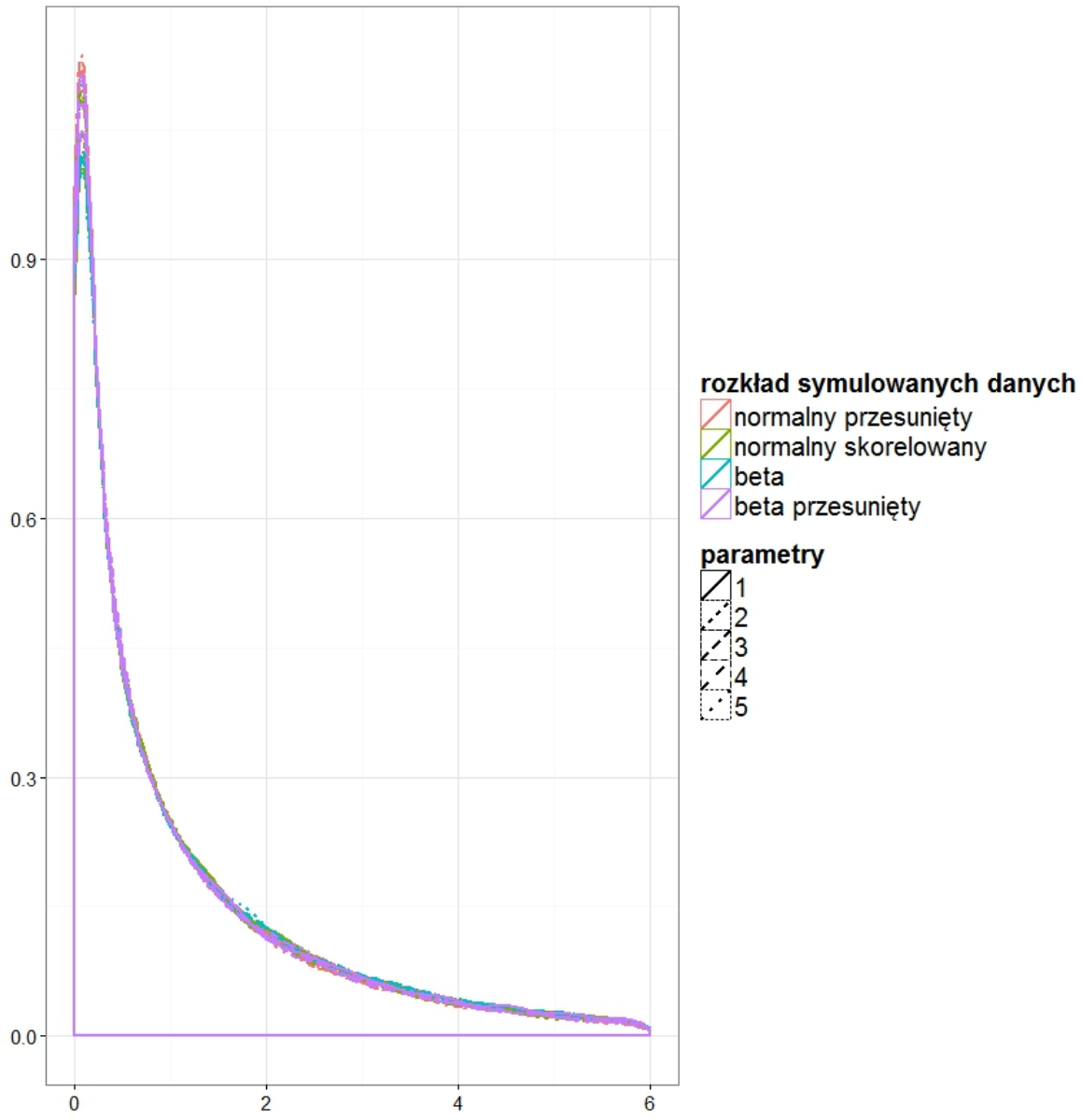
Tablica 3.5: Obszary krytyczne testów, na poziomie istotności 10%, dla różnych rozkładów danych wejściowych i rożnych parametrów.

Rozkłady statystyki T testu Walda, dla różnych metod generowania danych, nie różnią się znacznie. Wszystkie obszary krytyczne zawierają się między $(-\infty; -1,98) \cup (1,98; +\infty)$ a $(-\infty; -1,89) \cup (1,89; +\infty)$. Warto jednak zauważyć, że różnice są większe, niż w przypadku testu t-studenta dla różnic (patrz rozdział 3.2), a konkretnie dla danych niezależnych obszary krytyczne są węższe, niż dla danych zależnych. Nie są to jednak duże różnice, więc przyjmujemy, że zaprezentowany test nie jest wrażliwy na występowanie korelacji w danych. Dodatkowo nie jest on też wrażliwy na zmiany rozkładów. Będziemy go więc używać do weryfikacji hipotezy o równości średnich dwóch grup w rozdziale 4.

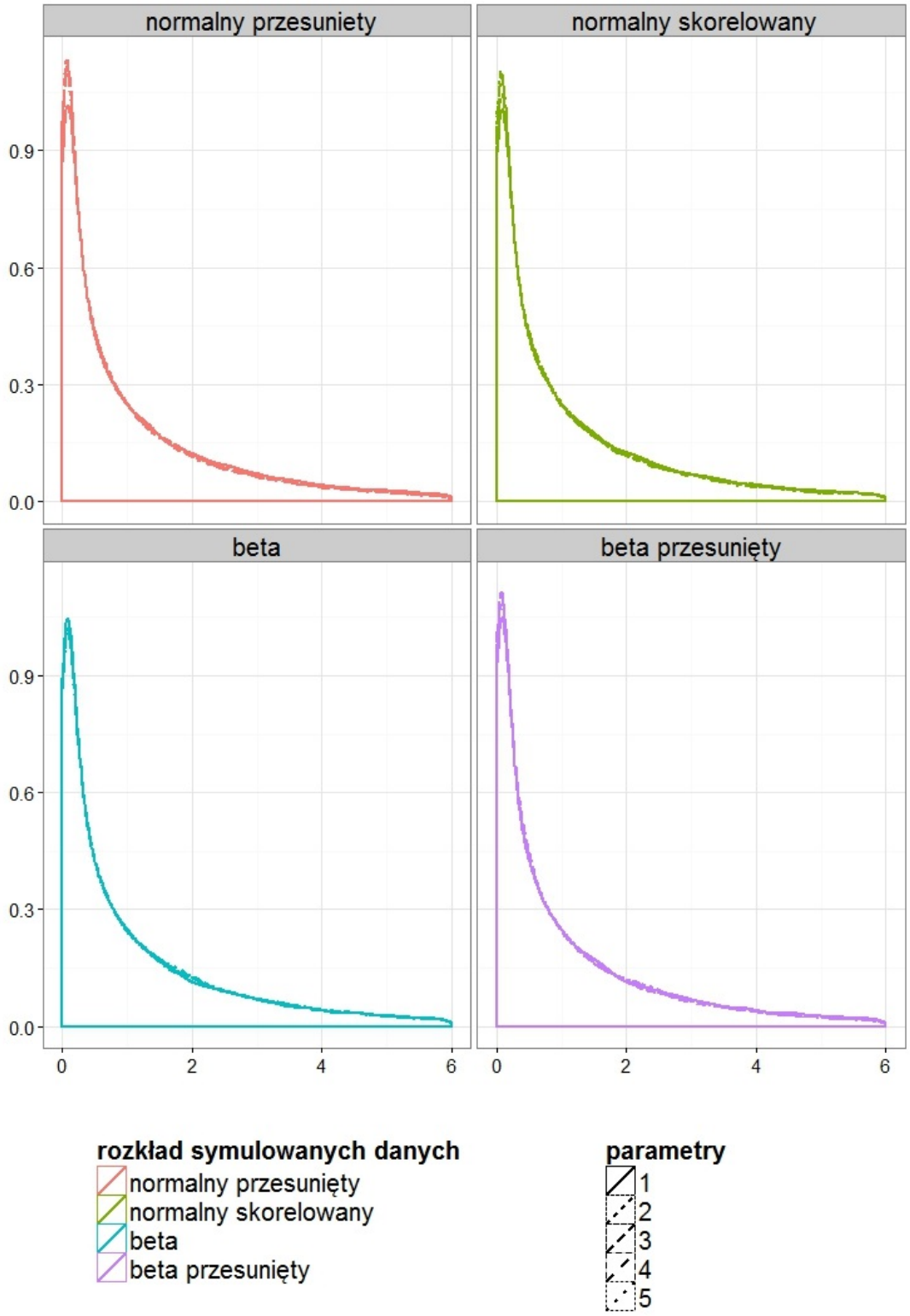
Biorąc wszystkie dane z symulacji, przyjmujemy obszar krytyczny testu postaci $(-\infty; -1,93) \cup (1,93; +\infty)$.

Warto zauważyć, że gdybyśmy przyjęli rozkład statystyki testowej testu Walda, jako $T(n-p)$, tzn. $T(10)$, to obszar krytyczny byłby postaci $(-\infty; -1,81) \cup (1,81; +\infty)$. Zawiera on obszar krytyczny, jaki otrzymujemy symulacyjnie. W związku z tym, zastosowanie testu Walda, z przyjęciem teoretycznego rozkładu statystyki testowej, spowodowałoby, że częściej odrzucalibyśmy prawdziwą hipotezę zerową.

Przejdźmy teraz do omówienia rozkładu statystyki T testu ilorazu wiarogodności. Gęstości rozkładu tej statystyki przy założeniu hipotezy zerowej dla wszystkich 20 metod generowania danych, można zobaczyć na rysunkach 3.5 i 3.6.



Rysunek 3.7: Wykres gęstości dla statystyki T testu ilorazu wiarogodności (3.12) przy hipotezie zerowej dla różnych metod generowania danych. Próby losowane z 4 rozkładów (na wykresie: rozkład symulowanych danych):
- rozkładu normalnego o średniej 0,5 i odchyleniu standardowym 0,1, dla którego w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry) - kolor czerwony,
- rozkładu normalnego o średniej 0.5 i odchyleniu standardowym 0.1, dla którego w obrębie tego samego pacjenta zmienne są skorelowane z korelacją równą: 1) 0, 2) 0,2, 3) 0,4, 4) 0,6, 5) 0,8 (na wykresie: parametry) - kolor zielony,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry) - kolor niebieski,
- rozkładu Beta z parametrami α i β równymi: 1) 2, 2) 4, 3) 6, 4) 8, 5) 10 (na wykresie: parametry), dla którego dodatkowo w obrębie tego samego pacjenta zastosowano przesunięcie wyniku o zmienną z rozkładu jednostajnego na przedziale: 1) bez przesunięcia, 2) [-0,2;0,2], 3) [-0,4;0,4], 4) [-0,6;0,6], 5) [-0,8;0,8] (na wykresie: parametry)) - kolor fioletowy.



Rysunek 3.8: Wykres gęstości dla statystyki T testu ilorazu wiarogodności (3.12) przy hipotezie zerowej dla różnych metod generowania danych. Dokładniejszy opis pod rysunkiem 3.7.

Na podstawie wyznaczonych rozkładów statystyki T testu ilorazu wiarogodności, budujemy obszary krytyczne testu do weryfikacji hipotezy $H_0 : \mu_X = \mu_Y$, na poziomie istotności 10%. Znajdują się one w tablicy 3.6.

rozkład symulowanych danych	parametry (możliwe przesunięcie, korelacja, α i β)	obszar krytyczny dla poziomu istotności 10%
normalny przesunięty	1) bez przesunięcia	$(3,72;+\infty)$
	2) $[-0,2;0,2]$	$(3,35;+\infty)$
	3) $[-0,4;0,4]$	$(3,33;+\infty)$
	4) $[-0,6;0,6]$	$(3,29;+\infty)$
	5) $[-0,8;0,8]$	$(3,30;+\infty)$
normalny skorelowany	1) $\rho = 0$	$(3,71;+\infty)$
	2) $\rho = 0,2$	$(3,57;+\infty)$
	3) $\rho = 0,4$	$(3,44;+\infty)$
	4) $\rho = 0,6$	$(3,39;+\infty)$
	5) $\rho = 0,8$	$(3,30;+\infty)$
Beta	1) $\alpha = \beta = 2$	$(3,75;+\infty)$
	2) $\alpha = \beta = 4$	$(3,74;+\infty)$
	3) $\alpha = \beta = 6$	$(3,73;+\infty)$
	4) $\alpha = \beta = 8$	$(3,74;+\infty)$
	5) $\alpha = \beta = 10$	$(3,74;+\infty)$
Beta przesunięty	1) $\alpha = \beta = 2$, bez przesunięcia	$(3,75;+\infty)$
	2) $\alpha = \beta = 4$, $[-0,2;0,2]$	$(3,47;+\infty)$
	3) $\alpha = \beta = 6$, $[-0,4;0,4]$	$(3,34;+\infty)$
	4) $\alpha = \beta = 8$, $[-0,6;0,6]$	$(3,34;+\infty)$
	5) $\alpha = \beta = 10$, $[-0,8;0,8]$	$(3,30;+\infty)$

Tablica 3.6: Obszary krytyczne testów, na poziomie istotności 10%, dla różnych rozkładów danych wejściowych i rożnych parametrów.

Z wykresów wynika, że rozkłady statystyki T testu ilorazu wiarogodności, dla różnych metod generowania danych, nie różnią się znacznie. Natomiast patrząc na obszary krytyczne, widzimy występowanie różnic. Są to węższe obszary dla danych niezależnych, niż dla danych zależnych. Obszary krytyczne zawierają się między $(3,29;+\infty)$ a $(3,75;+\infty)$. Przyjmujemy, że zaprezentowany test nie jest wrażliwy na zmiany rozkładów oraz występowanie korelacji w danych. Będziemy używać tego testu do weryfikacji hipotezy o równości średnich dwóch grup w rozdziale 4.

Biorąc wszystkie dane z symulacji, przyjmujemy obszar krytyczny testu postaci $(3,51;+\infty)$.

Warto zauważyć, że gdybyśmy przyjęli rozkład statystyki testowej testu ilorazu wiarogodności, jako χ^2 z jednym stopniem swobody, to obszar krytyczny byłby postaci $(2,71; +\infty)$. Zawiera on obszar krytyczny, jaki otrzymujemy symulacyjnie. W związku z tym, zastosowanie testu ilorazu wiarogodności, z przyjęciem teoretycznego rozkładu statystyki testowej, spowodowałoby, że częściej odrzucalibyśmy prawdziwą hipotezę zerową.

3.4. Porównanie metod

W rozdziale 3.1 stwierdziliśmy, że nie możemy użyć statystyki T testu t-studenta dla dwóch prób, do weryfikacji hipotezy o równości średnich dla danych takiego typu, jakie posiadamy (gdzie pomiary od tego samego pacjenta znajdują się zarówno w jednej, jak i w drugiej grupie). Dzieje się tak, ponieważ istnieje silna zależność rozkładu tej statystyki od korelacji między pomiarami. Pozostałe 3 testy, przedstawione w rozdziałach 3.2 i 3.3, tj.

- test t-studenta dla różnic, a dokładniej modyfikacja testu t-studenta dla prób sparowanych z testem t-studenta dla jednej próby,
- test Walda,
- test ilorazu wiarogodności,

mogą być używane, do testowania równości średnich w dwóch grupach pomiarów.

Chcemy wybrać najlepszy z zaprezentowanych testów. W tym celu wykonujemy symulacje badające błąd pierwszego rodzaju oraz moc trzech powyższych testów.

W symulacjach 3000 razy powtarzamy następującą procedurę:
Dla każdego $r \in \{0; 0,05; 0,1; 0,15; 0,2; 0,25; 0,3\}$, $a \in \{0; 0,2; 0,4; 0,6; 0,8\}$ oraz dla każdej z dwóch metod generujemy zestaw 12 liczb, w których przyjmujemy zależności, jak w symulacjach opisanych w rozdziale 3.1., tzn.

numer pomiaru	1	2	3	4	5	6	7	8	9	10	11	12
grupa	1	1	1	1	2	2	2	2	2	2	2	2
pacjent	1	2	3	4	1	1	2	2	3	3	3	4

Tablica 3.7: Podział symulowanych pomiarów na 2 grupy oraz 4 pacjentów.

Wspomniane 2 metody są następujące:

- Każdy zestaw 12 liczb generujemy z rozkładu normalnego, gdzie wartość oczekiwana elementów z grupy pierwszej wynosi $0,5-r/2$, a drugiej grupy $0,5+r/2$, tak aby różnica wartości oczekiwanych wynosiła r. Wariancja wynosi 0,1. Następnie wartości oznaczone tym samym pacjentem zmieniamy, o liczbę wygenerowaną z rozkładu jednostajnego na przedziale $[-a, a]$.

- Każdy zestaw 12 liczb generujemy z rozkładu normalnego, gdzie wartość oczekiwana elementów z grupy pierwszej wynosi $0,5-r/2$, a drugiej grupy $0,5+r/2$, wariancja jest równa 0,1 oraz korelacja danych oznaczonych tym samym pacjentem wynosi a.

Dla każdego zestawu 12 liczb wykonujemy trzy rozważane testy statystyczne oraz notujemy, czy hipoteza zerowa o równości średnich jest odrzucana na poziomie istotności 10%.

W ten sposób dla obydwu metod, każdej wartości r i a oraz każdego z 3 testów otrzymujemy ilość odrzuceń hipotezy zerowej, na 3000 możliwych.

W tablicy 3.8 (str. 57) przedstawiono jak często hipoteza zerowa jest odrzucana przez poszczególne testy, dla danych generowanych pierwszą metodą. Trzy tabele pokazują wartości dla trzech analizowanych testów. W kolumnach znajdują się różne wartości maksymalnego przesunięcia danych w obrębie jednego pacjenta (a), natomiast w wierszach różnica wartości oczekiwanych generowanych danych (r). W związku z tym, pierwszy wiersz, dla różnicy wartości oczekiwanych równej zero, pokazuje nam błąd pierwszego rodzaju testu, natomiast pozostałe wiersze moc statystyczną. Najlepszy test powinien mieć wartości w pierwszym wierszu równe 10% oraz jak największe wartości w pozostałych wierszach. Tablica 3.9 (str. 58) przedstawia analogiczne wyniki dla drugiej metody generowania danych, tzn. zamiast maksymalnego przesunięcia w obrębie tego samego pacjenta, określamy korelacje danych pochodzących od tego samego pacjenta. W kolumnach znajdują się różne wartości korelacji (a).

Dane z tablic 3.8 i 3.9 są dodatkowo przedstawione na rysunkach 3.9 i 3.10. W przypadku pierwszej metody generowania danych, dla testu t-studenta błąd pierwszego rodzaju oraz moc statystyczna testu nie zależą od przesunięcia. Natomiast dla pozostałych dwóch testów, dla konkretnej różnicy pomiędzy wartościami oczekiwanymi, częściej odrzucamy hipotezę zerową gdy dane są niezależne, niż gdy są zależne. Inaczej sytuacja wygląda dla danych generowanych z rozkładu normalnego skorelowanego, gdzie skorelowane są dane pochodzące od tego samego pacjenta. Tutaj, dla konkretnej, niezerowej różnicy między wartościami oczekiwanymi, im większa korelacja, tym częściej odrzucamy hipotezę zerową. Moc statystyczna testów rośnie więc wraz ze wzrostem korelacji. Dla zerowej różnicy między wartościami oczekiwanymi, tzn. dla prawdziwej hipotezy zerowej, częstość odrzuceń jest mniejsza dla danych skorelowanych. Oznacza to, że błąd pierwszego rodzaju jest mniejszy dla tych danych. Wszystkie trzy testy wykazują podobną zależność, tzn. wraz ze wzrostem korelacji moc statystyczna testu rośnie oraz błąd pierwszego rodzaju maleje. Podsumowując, wszystkie trzy testy działają lepiej dla danych skorelowanych niż nieskorelowanych. Natomiast dla danych przesuniętych w obrębie tego samego pacjenta, test Walda i ilorazu wiarogodności działają gorzej, niż dla danych niezależnych, a wyniki testu t-studenta nie zależą od przesunięcia w danych.

W celu porównania ze sobą trzech zaprezentowanych testów, na rysunkach 3.11 - 3.16 przedstawiono różnicę procentu odrzuconych hipotez przez pierwszy test oraz drugi test, dla każdej

pary testów. Wykresy pokazane są w podziale na metodę generowania danych (przesunięcie w obrębie tego samego pacjenta lub korelacja danych pochodzących od tego samego pacjenta), różnice wartości oczekiwanych oraz wartość przesunięcia lub korelację. Najlepszy test powinien dla każdej metody generowania danych oraz dla każdego przesunięcia lub korelacji, dawać większe wartości odrzuceń dla niezerowej różnicy wartości oczekiwanych (tzn. większą moc statystyczną).

Rysunki 3.11 i 3.12 pokazują różnicę częstości odrzucania hipotez dla testu Walda i testu t-studenta. Obserwujemy większą moc statystyczną testu Walda, w porównaniu z testem t-studenta. Warto zauważyć, że różnice w częstości odrzuceń hipotezy zerowej maleją wraz ze wzrostem różnicy wartości oczekiwanych. Należy jednak wziąć pod uwagę, że częstość odrzuceń hipotezy zerowej jest wtedy dla obu testów wysoka i działają one wtedy bardzo dobrze. Błąd pierwszego rodzaju przyjmuje dla nich podobne wartości. Na tej podstawie można stwierdzić, że test Walda jest lepszy niż test t-studenta, zwłaszcza dla danych zależnych.

Podobne wyniki możemy zaobserwować na rysunkach 3.13 i 3.14, gdzie przedstawiona jest różnica procentu odrzuconych hipotez dla testu ilorazu wiarogodności i testu t-studenta. Dla danych z przesunięciem w obrębie tego samego pacjenta oraz niedużej różnicy wartości oczekiwanych, częstości odrzuceń są porównywalne, a w niektórych przypadkach nawet większe dla testu t-studenta. Dla większych różnic między wartościami oczekiwanymi odrzuceń jest więcej w przypadku testu ilorazu wiarogodności. Dla danych niezależnych oraz skorelowanych moc jest większa dla testu ilorazu wiarogodności. Podobnie jak wcześniej, różnice w częstości odrzuceń hipotezy zerowej maleją wraz ze wzrostem różnicy wartości oczekiwanych, jednak oba testy działają wtedy bardzo dobrze. Błąd pierwszego rodzaju testu ilorazu wiarogodności i t-studenta jest podobny. Na tej podstawie można stwierdzić, że test ilorazu wiarogodności jest lepszy niż test t-studenta, zwłaszcza dla danych skorelowanych.

Wiemy już, że biorąc pod uwagę moc statystyczną testów, test t-studenta jest gorszy, niż pozostałe dwa. Na rysunkach 3.15 i 3.16 przedstawiono różnicę procentu odrzuconych hipotez dla testu Walda i ilorazu wiarogodności. Dla danych niezależnych, hipoteza zerowa jest częściej odrzucana przez test ilorazu wiarogodności, niż przez test Walda, niezależnie od tego, czy jest ona prawdziwa, czy fałszywa. Natomiast dla danych zależnych (zarówno przesuniętych, jak i skorelowanych) sytuacja wygląda odwrotnie, tzn. hipoteza zerowa jest częściej odrzucana przez test Walda, niż przez test ilorazu wiarogodności, niezależnie od tego, czy jest prawdziwa. Nie pozwala to jednoznacznie stwierdzić, który z nich jest lepszy. Zauważmy również, że różnice w mocy tych dwóch testów nie przekraczają 2 punktów procentowych, co pozwala stwierdzić, że są one porównywalne.

Podsumowując, zarówno test Walda, jak i ilorazu wiarogodności dają lepsze rezultaty niż test t-studenta. Nie możemy natomiast jednoznacznie stwierdzić który z tych dwóch testów jest lepszy. Warto zwrócić uwagę na fakt, iż badając rozkłady statystyk testowych wymienionych

testów, dla każdego z nich stwierdziliśmy brak wrażliwości na zależności w danych, jednak testy Walda i ilorazu wiarogodności były bardziej wrażliwe. To sugerowałoby, że test t-studenta jest lepszy niż pozostałe dwa. W związku z brakiem silnej wyższości jednego testu nad pozostałymi, do danych w rozdziale 4 zastosujemy wszystkie 3 wymienione testy.

test t-studenta dla różnic		maksymalne przesunięcie				
		0	0,2	0,4	0,6	0,8
różnica wartości oczekiwanych	0	10,2%	10,2%	9,2%	9,9%	10,0%
	0,05	18,5%	18,2%	19,3%	18,8%	17,6%
	0,1	41,9%	40,4%	41,1%	39,7%	39,9%
	0,15	68,1%	67,1%	69,5%	66,7%	66,9%
	0,2	86,7%	86,8%	87,2%	86,8%	88,4%
	0,25	96,2%	96,0%	95,6%	95,5%	96,7%
	0,3	99,2%	99,5%	99,0%	99,1%	98,9%

test Walda		maksymalne przesunięcie				
		0	0,2	0,4	0,6	0,8
różnica wartości oczekiwanych	0	11,5%	9,4%	9,4%	9,1%	9,5%
	0,05	19,9%	18,1%	18,5%	18,3%	17,6%
	0,1	46,0%	41,0%	42,1%	41,0%	41,1%
	0,15	74,2%	69,8%	70,3%	68,9%	67,9%
	0,2	91,3%	89,0%	89,3%	88,7%	89,4%
	0,25	98,3%	97,4%	97,3%	96,7%	97,5%
	0,3	99,8%	99,8%	99,4%	99,6%	99,4%

test ilorazu wiarogodności		maksymalne przesunięcie				
		0	0,2	0,4	0,6	0,8
różnica wartości oczekiwanych	0	12,1%	9,1%	8,9%	8,5%	9,2%
	0,05	20,9%	17,5%	17,7%	17,6%	17,1%
	0,1	47,7%	40,2%	41,0%	39,7%	39,8%
	0,15	75,4%	69,1%	69,3%	67,5%	66,7%
	0,2	92,1%	88,4%	88,2%	88,0%	88,9%
	0,25	98,6%	97,3%	97,0%	96,2%	97,2%
	0,3	99,8%	99,8%	99,1%	99,5%	99,4%

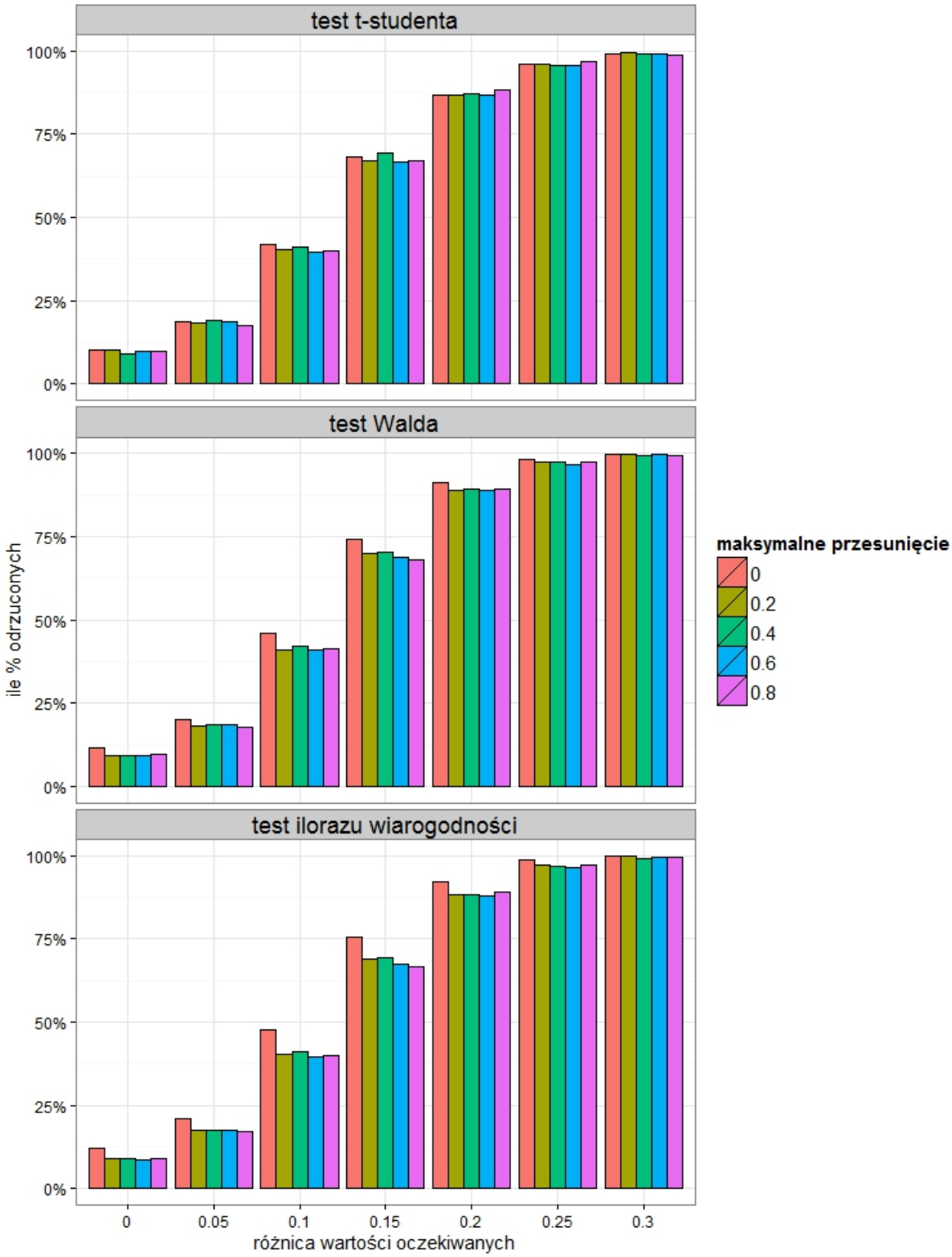
Tablica 3.8: Tabele dla testu t-studenta, Walda i ilorazu wiarogodności. Przedstawiają procent odrzuconych hipotez zerowych dla danych generowanych z rozkładu normalnego przesuniętego. Zastosowano podział ze względu na różnice wartości oczekiwanych w dwóch grupach oraz wartości maksymalnego przesunięcia danych w obrębie tego samego pacjenta.

test t-studenta dla różnic		korelacja				
		0	0,2	0,4	0,6	0,8
różnica wartości oczekiwanych	0	10,6%	9,9%	9,6%	10,2%	9,7%
	0.05	18.4%	20.6%	23.6%	31.1%	46.8%
	0.1	41.7%	45.6%	56.1%	70.8%	92.2%
	0.15	66.6%	74.5%	85.2%	93.3%	99.8%
	0.2	86.3%	92.3%	96.5%	99.4%	100%
	0.25	95.9%	98.2%	99.7%	100%	100%
	0.3	99.1%	99.8%	100%	100%	100%

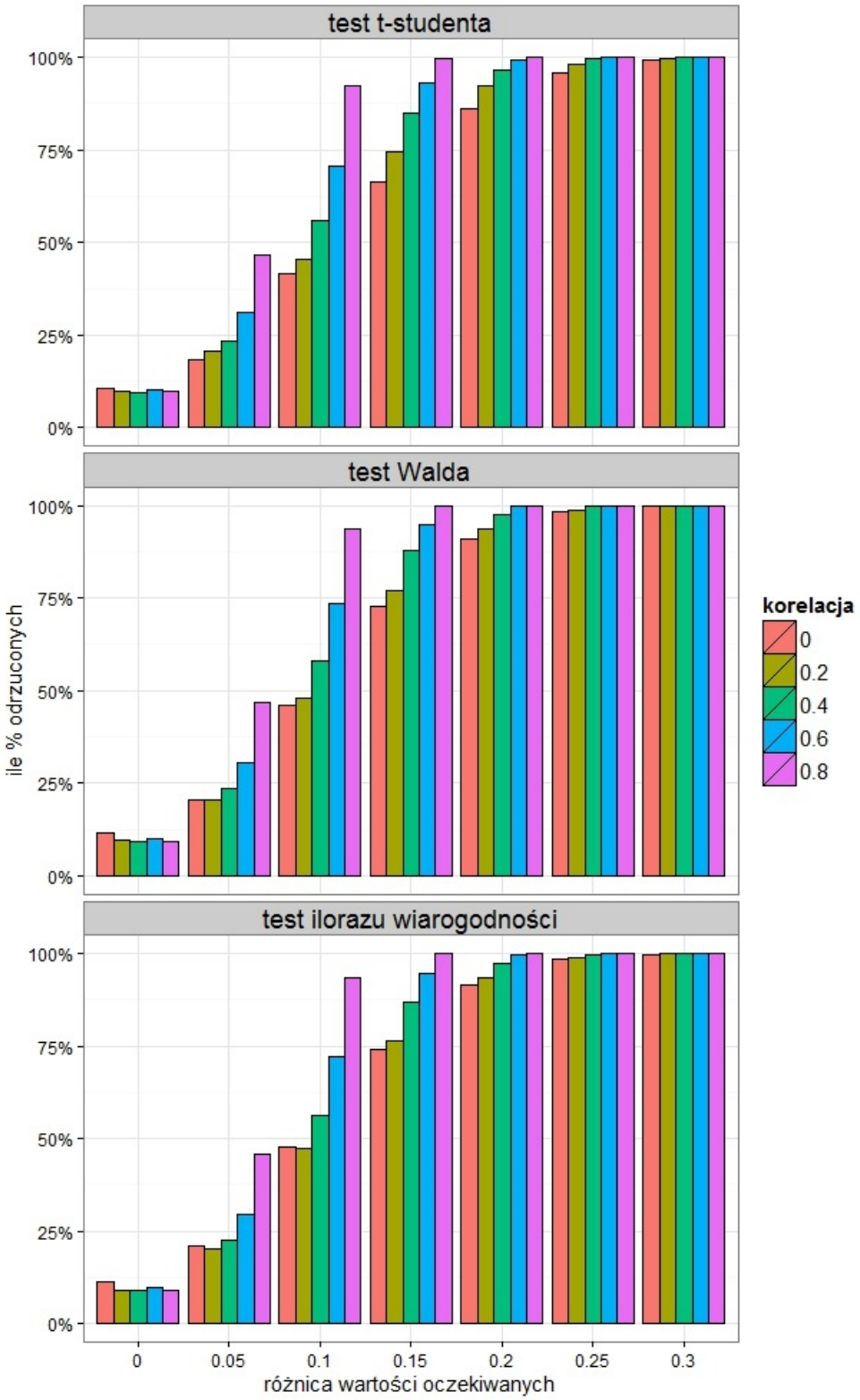
test Walda		korelacja				
		0	0.2	0.4	0.6	0.8
różnica wartości oczekiwanych	0	11.4%	9.8%	9.3%	10.1%	9.2%
	0.05	20.5%	20.5%	23.5%	30.5%	46.9%
	0.1	45.9%	48.1%	58.1%	73.4%	93.6%
	0.15	72.8%	77.2%	87.7%	95.0%	100%
	0.2	91.1%	93.7%	97.7%	99.7%	100%
	0.25	98.5%	98.9%	99.8%	100%	100%
	0.3	99.8%	99.9%	100%	100%	100%

test ilorazu wiarogodności		korelacja				
		0	0,2	0,4	0,6	0,8
różnica wartości oczekiwanych	0	11,5%	9,2%	9,0%	9,7%	8,9%
	0,05	21,0%	20,1%	22,7%	29,7%	45,9%
	0,1	47,9%	47,2%	56,2%	72,0%	93,3%
	0,15	74,2%	76,4%	86,9%	94,5%	100%
	0,2	91,5%	93,4%	97,4%	99,5%	100%
	0,25	98,6%	98,8%	99,8%	100%	100%
	0,3	99,8%	99,9%	100%	100%	100%

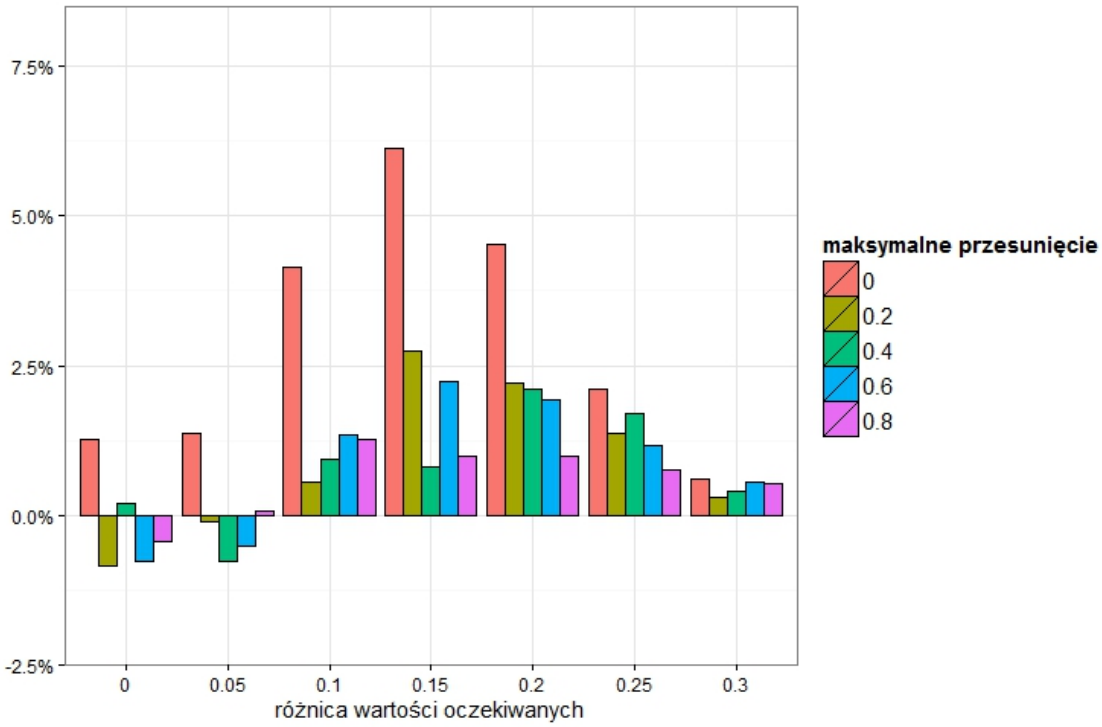
Tablica 3.9: Tabele dla testu t-studenta, Walda i ilorazu wiarogodności. Przedstawiają procent odrzuconych hipotez zerowych dla danych generowanych z rozkładu normalnego skorelowanego. Zastosowano podział ze względu na różnice wartości oczekiwanych w dwóch grupach oraz wartości korelacji.



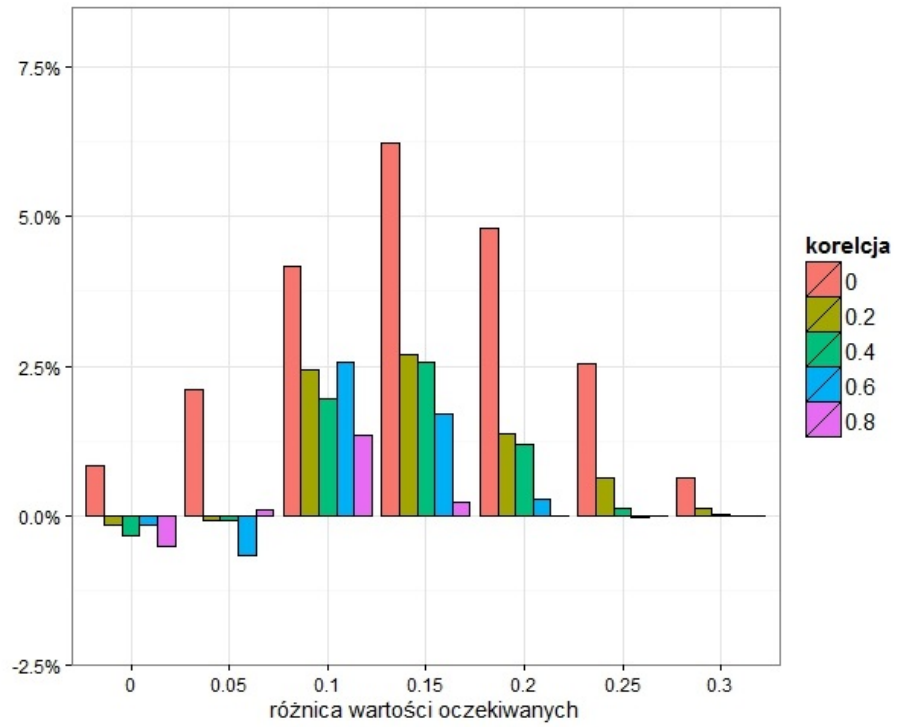
Rysunek 3.9: Dane z tabeli 3.8, przedstawiające w ilu procentach przypadków poszczególne testy odrzuciły hipotezę zerową o równości średnich (moce testów), dla różnic między średnimi oraz różnych przesunięć w danych.



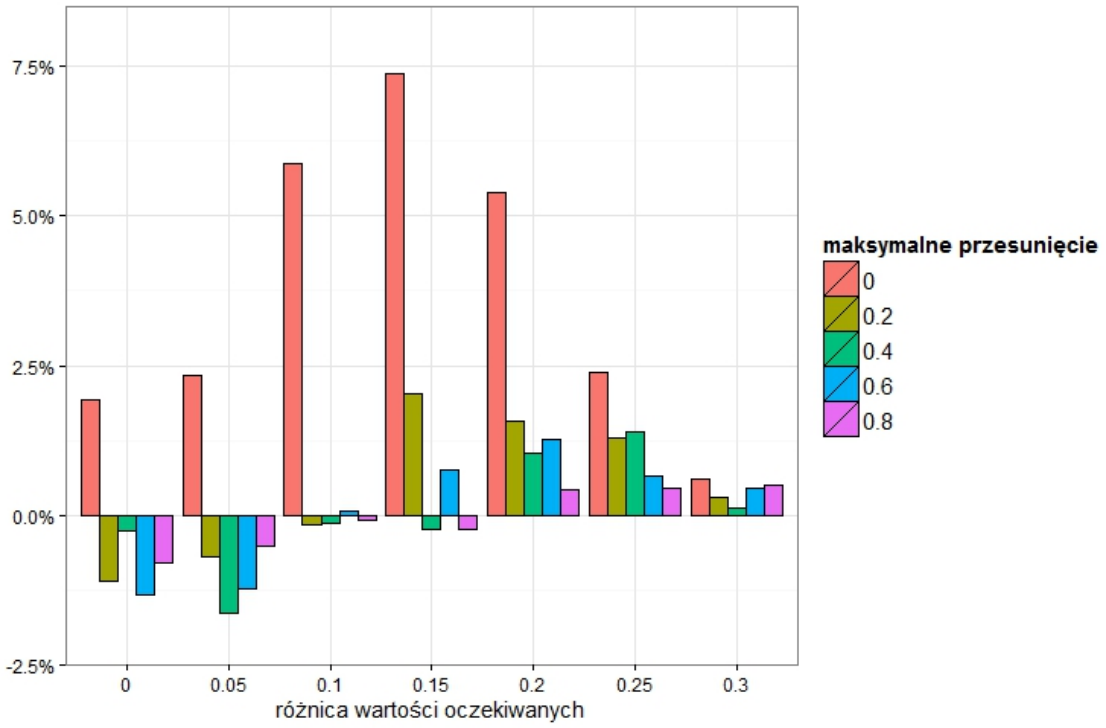
Rysunek 3.10: Dane z tabeli 3.9, przedstawiające w ilu procentach przypadków poszczególne testy odrzuciły hipotezę zerową o równości średnich (moce testów), dla różnic między średnimi oraz różnych korelacji w danych.



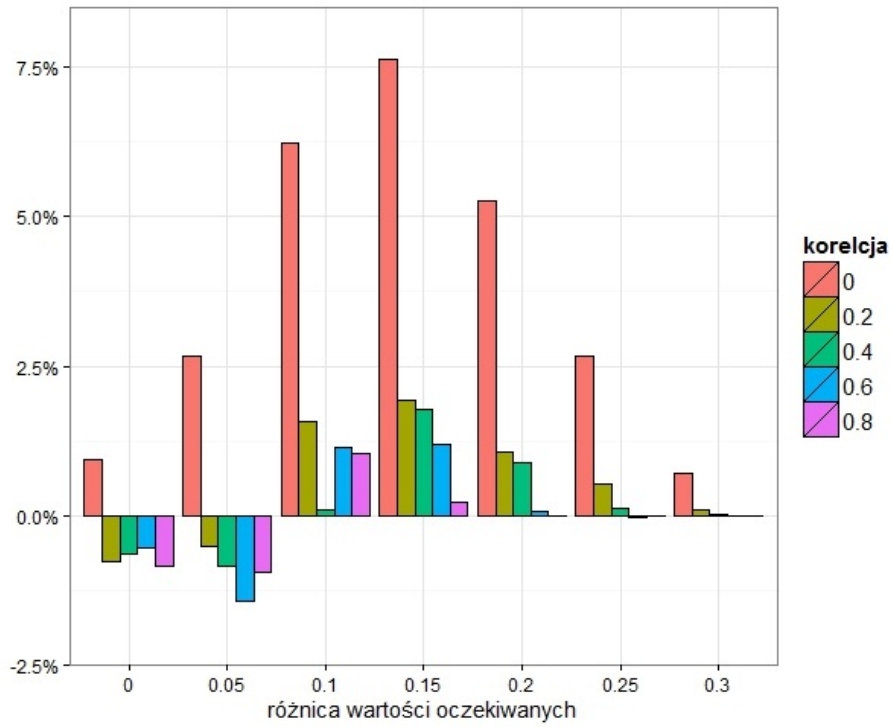
Rysunek 3.11: Różnica mocy testu Walda i testu t-studenta, dla różnic między średnimi oraz różnych przesunięć w danych.



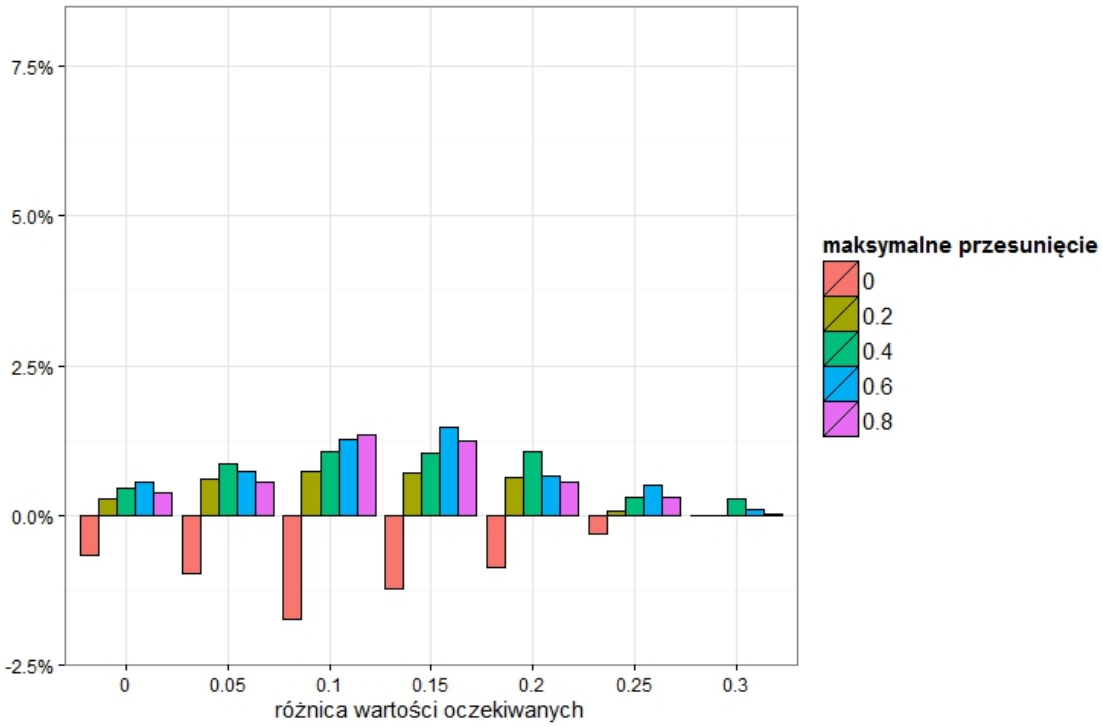
Rysunek 3.12: Różnica mocy testu Walda i testu t-studenta, dla różnic między średnimi oraz różnych korelacji w danych.



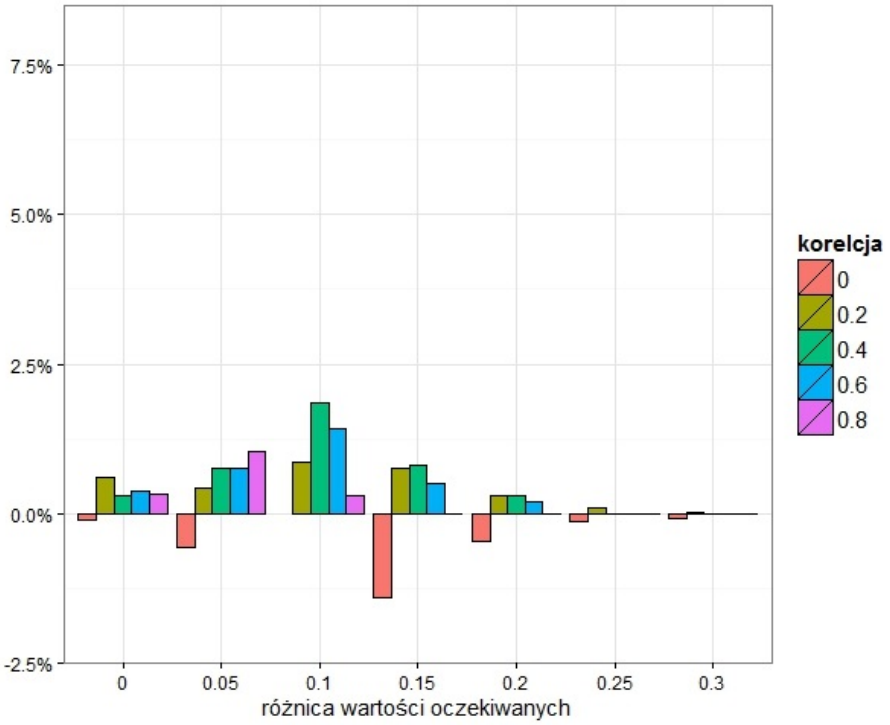
Rysunek 3.13: Różnica mocy testu ilorazu wiarogodności i testu t-studenta, dla różnic między średnimi oraz różnych przesunięć w danych.



Rysunek 3.14: Różnica mocyu test ilorazu wiarogodności i testu t-studenta, dla różnic między średnimi oraz różnych korelacji w danych.



Rysunek 3.15: Różnica mocy testu Walda i testu ilorazu wiarogodności, dla różnic między średnimi oraz różnych przesunięć w danych.



Rysunek 3.16: Różnica mocy testu Walda i testu ilorazu wiarogodności, dla różnic między średnimi oraz różnych korelacji w danych.

Rozdział 4

Analiza wyników dla danych KRAB-ZNF

4.1. Dane

Dane, które analizujemy zawierają 12 próbek, które możemy podzielić na dwie grupy:

1. **PHDF** (Primary Human Dermal Fibroblasts), czyli ludzkie pierwotne fibroblasty skóry (komórki wyspecjalizowane) - 4 próbki

2. **iPS** (induced Pluripotent Stem cells), czyli indukowane komórki pluripotentne (komórki macierzyste) - 8 próbek.

Komórki iPS, są komórkami pluripotentnymi, tzn. że mogą z nich powstać wszystkie tkanki budujące organizm. Mogą one być bardzo przydatne, np. w medycynie regeneracyjnej. Komórki iPS powstały poprzez zaindukowanie w nich ekspresji czterech genów. Geny te powodują reprogramowanie (tzn. przekształcenie) fibroblastów do komórek pluripotentnych, które morfologicznie i funkcjonalnie przypominają zarodkowe komórki macierzyste [10]. Komórki iPS wygenerowano z linii komórkowych PHDF. Cztery linie komórkowe PHDF pochodzą od czterech różnych pacjentów, a osiem linii komórkowych iPS od tych samych czterech pacjentów, zgodnie z rozkładem przedstawionym w tablicy 4.1.

numer pomiaru	1	2	3	4	5	6	7	8	9	10	11	12
grupa	PHDF	PHDF	PHDF	PHDF	iPS	iPS	iPS	iPS	iPS	iPS	iPS	iPS
pacjent	1	2	3	4	1	1	2	2	3	3	3	4

Tablica 4.1: Podział próbek na 2 grupy (PHDF i iPS) oraz 4 pacjentów.

Dane wejściowe są zawarte w formacie FASTAQ (rysunek 4.1). Dla każdej próbki dane znajdują się w dwóch plikach FASTAQ, ponieważ próby były sekwencjonowane obustronnie (tzn. zarówno lewy, jak i prawy koniec fragmentu był sekwencjonowany).

Rysunek 4.1: Fragment danych w formacie FASTAQ, zawierający: identyfikator sekwencji (rodzaj maszyny, w którym miejscu fizycznie była próba, itp.), odczyt sekwencji, informację o jakości sekwencjonowania.

Celem badania jest zidentyfikowanie genów, w których rozkłady wariantów splicingowych istotnie różnią się między grupą PHDF, a grupą iPS.

Zajmujemy się analizą genów KRAB-ZNF, które zawierają domenę KRAB (Kruppel - Associated Box) i ZNF (Zinc Finger). Domeny to fragmenty białka, które spełniają w nim konkretną funkcję. Spodziewamy się, że czynniki KRAB-ZNF będą miały wpływ na proces reprogramowania komórek somatycznych do komórek iPS oraz na utrzymanie stanu pluripotencji komórkek iPS. Czynniki KRAB-ZNF wiążą się bezpośrednio do specyficznych sekwencji DNA poprzez domenę palców cynkowych (ZNF). Następnie białka KRAB-ZNF oddziałują z białkiem TRIM28/KAP1, które pomaga w heterochromatyzacji (wyłączenie ekspresji genów). Brak białka TRIM28/KAP1 w komórkach pluripotentnych powoduje utratę macierzystości - komórki zaczynają się różnicować do innych, bardziej wyspecjalizowanych typów. Natomiast w komórkach somatycznych, brak TRIM28/KAP1 promuje reprogramowanie do iPS. Ponieważ TRIM28/KAP1 ma wpływ na stan pluripotencji, można również założyć, że dzieje się tak właśnie poprzez białka KRAB-ZNF. Chcemy zidentyfikować te czynniki KRAB-ZNF, które ulegają nadekspresji w komórkach iPS. Ponieważ mogą one występować w wielu izoformach, chcemy sprawdzić, które izoformy są specyficzne dla komórek pluripotentnych. Wiedząc to, można prowadzić dalsze badania w laboratorium. Z wcześniejszych badań wynika np., że gen ZNF620 w komórkach iPS występuje przede wszystkim w wariancie NM_001256167, gdzie nie ma części domeny KRAB. Może to prawdopodobnie powodować jego nieprawidłowe oddziaływanie z białkiem TRIM28/KAP1, przez co nie dochodzi do heterochromatyzacji.

4.2. Wykorzystane programy: TopHat i pakiet Casper w R

Zanim użyjemy przedstawionego w rozdziale drugim algorytmu Casper, musimy najpierw przekształcić dane w formacie FASTAQ, tak, aby wiedzieć, w których miejscach zaczynają się kolejne eksony.

W tym celu używamy programu TopHat, który mapuje odczyty RNA-seq do genomu referencyjnego, w celu zidentyfikowania węzłów splicingowych, czyli połączeń między eksonami (miejsc, gdzie nastąpiło wycięcie intronu i połączenie eksonów).

TopHat wywołujemy dla każdej próbki osobno, tzn. dla każdej pary odczytów, w następujący sposób:

```
tophat - r 200 <genome_index_base> read_1.fastaq, read_2.fastaq,
```

gdzie `read_1.fastaq` jest plikiem w formacie FASTAQ zawierającym lewe odczyty sekencjonowanych fragmentów, a `read_2.fastaq` - prawe odczyty, `<genome_index_base>` jest ścieżką do indeksu Bowtie, a `- r` jest opcją wymaganą, gdy analizujemy sparowane odczyty i oznacza średnią długość sekwencjonowanego fragmentu. Długość jest zazwyczaj między 200 a 300, więc dowolna wartość `- r` w tym przedziale, powinna być odpowiednia. Po wprowadzeniu danych do R, możemy użyć funkcji `getDistrs` z pakietu `casper`, do estymacji długości fragmentów, aby sprawdzić, czy podana przez nas wartość była prawidłowa. TopHat posiada też inne, niewymagane opcje. Więcej informacji można znaleźć w [11].

Aby znaleźć węzły splicingowe poprzez wywołanie TopHat, musimy posiadać, wspomniany wyżej, indeks Bowtie dla organizmu, którego dotyczy badanie. Jest to plik ze wszystkimi sekwencjami genomu danego organizmu przyporządkowanymi do konkretnych miejsc w genomie. Dla człowieka i wielu innych organizmów dostępne są gotowe indeksy. Używamy indeksu hg19, do którego mapowane są sekwencje z plików FASTAQ.

TopHat zwraca nam listę dopasowanych odczytów w formacie SAM/BAM. Plik wynikowy ma nazwę `accepted_hits.bam`.

Przejdźmy teraz do estymacji prawdopodobieństw występowania wariantów splicingowych poprzez pakiet `casper` w R. Funkcja `wrapKnown` z tego pakietu wywołuje całą analizę dla jednej próbki danych w formacie BAM. Plik wejściowy musi być posortowany i indeksowany, a indeks musi znajdować się w tym samym katalogu, co plik. Możemy go uzyskać poprzez wywołanie

```
idx <- indexBam(”sciezka_do_pliku/accepted_hits.bam”).
```

Zanim wywołamy funkcję `wrapKnown` potrzebujemy jeszcze genomu referencyjnego, który może być wygenerowany z bazy UCSC poprzez poniższy kod.

```
genDB <- makeTranscriptDbFromUCSC(genome='hg19',tablename=''refGene'')
hg19DB <- procGenome(genDB=genDB, genome='hg19')
```

Funkcje pochodzą z pakietu `GenomicFeatures`. Obiekt `hg19DB` jest klasy `annotatedGenome` i zawiera informacje dotyczące genów, transkryptów i eksonów.

Za pomocą funkcji `wrapKnown` wykonujemy całą analizę opisaną w rozdziale 2.3, możemy to zrobić przez wywołanie poniższego kodu.

```
ans <- wrapKnown(bamFile=''sciezka_do_pliku/accepted_hits.bam'', genomeDB=hg19DB,
readLength=75, rpkm=FALSE),
```

gdzie `readLength` jest długością odczytów, a parametr `rpkm` ustawiony na wartość `FALSE` powoduje, że estymowane są prawdopodobieństwa dla każdego wariantu, sumujące się do 1 dla każdej wyspy. W celu otrzymania wyników sumujących się do 1 w obrębie genu, wywołujemy następnie funkcję

```
ans_g <- relexprByGene(ans$exp).
```

Funkcja `wrapKnown` najpierw dokonuje wstępnego przetworzenia danych, a następnie określa eksony odwiedzane przez każdy odczyt, które nazywamy ścieżką oraz liczbę odczytów mających tę samą ścieżkę, po czym dokonuje estymacji metodą opisaną w rozdziale 2.3.

Otrzymane poprzez wywołanie `wrapKnown` (i ewentualnie `relexprByGene`) obiekty klasy `ExpressionSet` dla wszystkich próbek łączymy poprzez zastosowanie funkcji `mergeExp`

```
ans_merge <- mergeExp(ans_g_1, ans_g_2, ans_g_3, ...).
```

Wywołanie `fData(ans_merge)` daje tablicę z wyestymowanymi prawdopodobieństwami, którą możemy następnie analizować metodami przedstawionymi w rozdziale 3. Szczegółowy opis działania wymienionych funkcji z pakietu `casper` oraz wszystkich funkcji, które działają wewnątrz wywołania funkcji `wrapKnown` można znaleźć w [12].

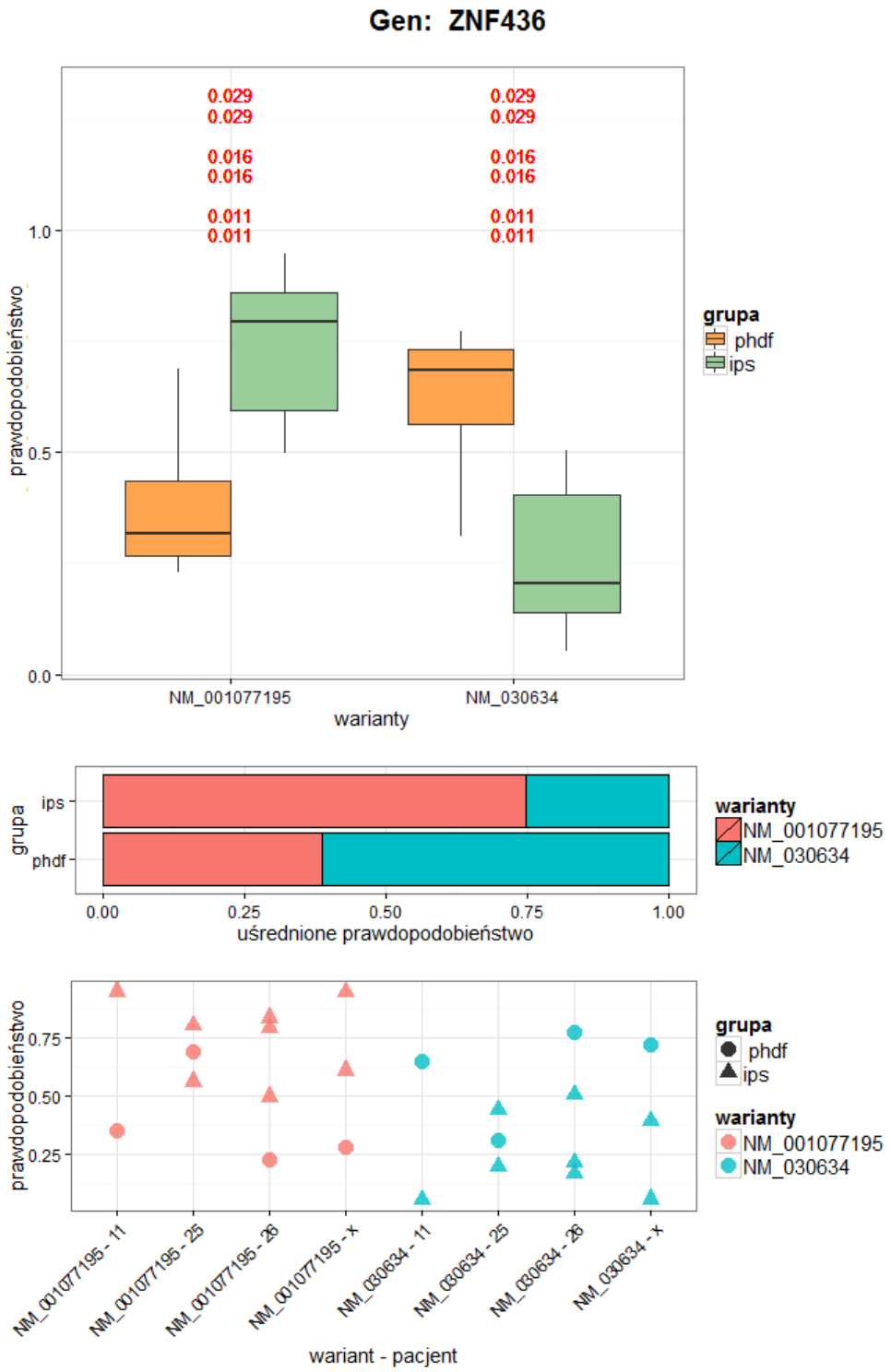
4.3. Wyniki

Otrzymane prawdopodobieństwa wyestymowanych wariantów splicingowych genów, z zastosowaniem pakietu casper, poddajemy analizie opisanej w rozdziale 3, w celu identyfikacji genów, w których rozkłady występowania wariantów istotnie różnią się między grupą PHDF a iPS. Analizie poddajemy 349 genów z domeną KRAB-ZNF, z których każdy ma między 1 a kilkanaście wariantów, łącznie 847 wariantów. Po usunięciu genów posiadających tylko jeden wariant, pozostaje 179 genów z 677 wariantami.

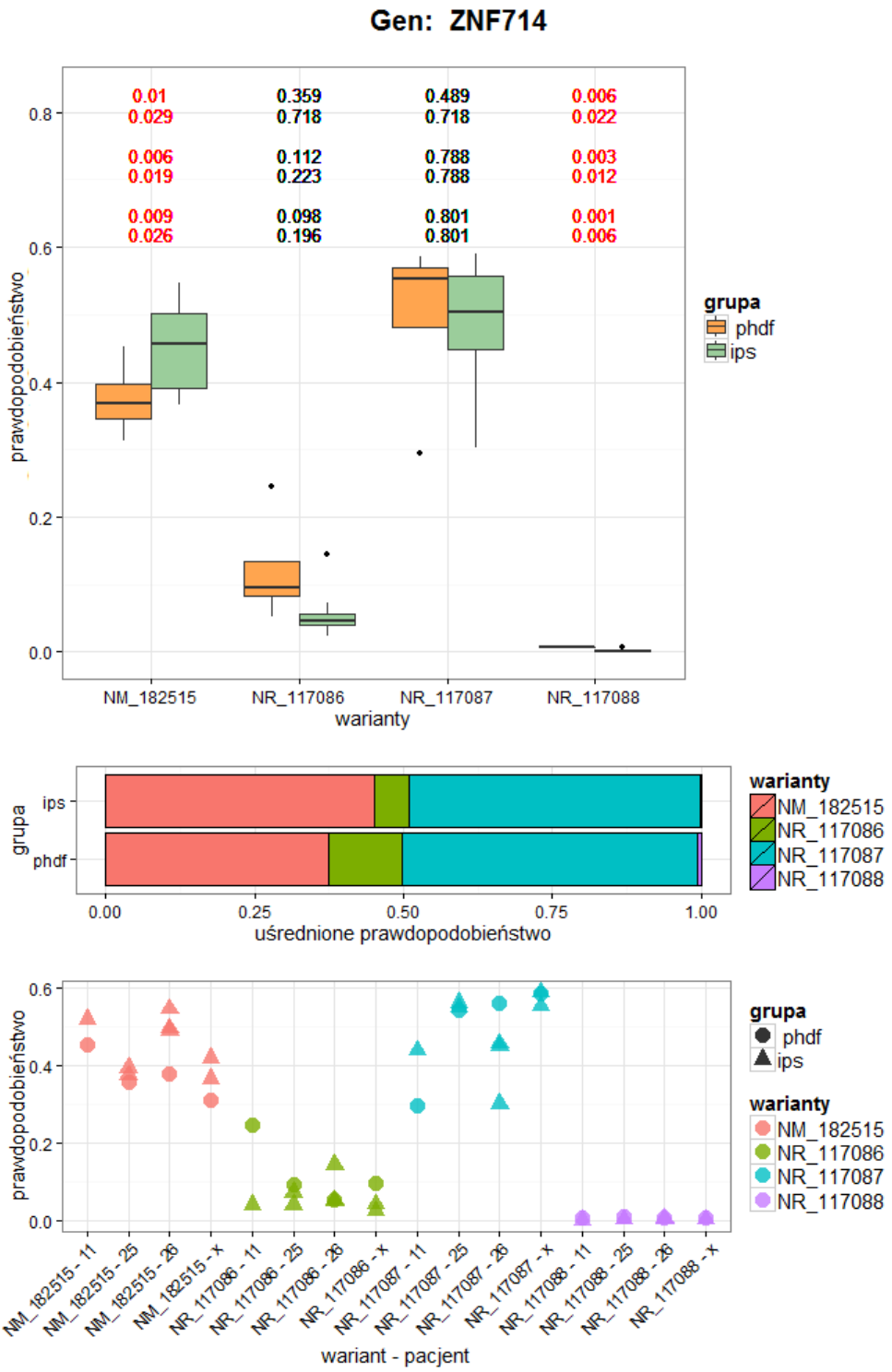
Dla każdego wariantu stosujemy trzy testy statystyczne: przedstawiony w rozdziale 3.2 test t-studenta dla różnic oraz zaprezentowane w rozdziale 3.3 testy Walda i ilorazu wiarogodności, zastosowane po dopasowaniu modelu liniowego z efektami losowymi.

W obrębie każdego genu stosujemy dodatkowo poprawkę Holma na wielokrotne testowanie, związane z liczbą wariantów genu. Każdy test klasyfikuje gen, jako istotnie różniący się pod względem rozkładu wariantów, jeśli po uwzględnieniu poprawki Holma, choć dla jednego wariantu, na poziomie istotności 10%, odrzucana jest hipoteza o tym, że prawdopodobieństwo dla tego wariantu jest takie samo w obu grupach.

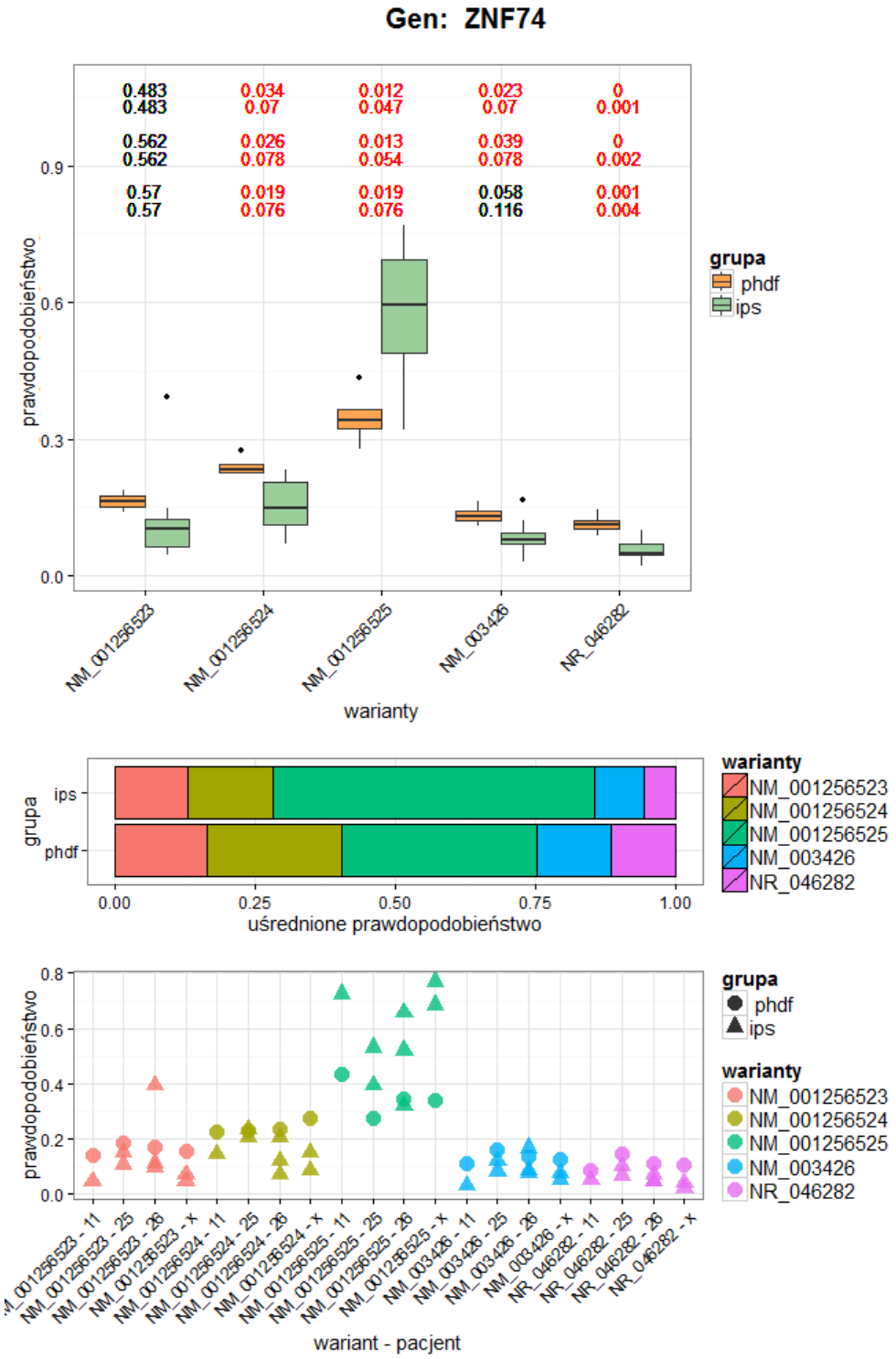
Prezentację wyników dla kilku ciekawych genów można zobaczyć na rysunkach 4.2 - 4.7. Każdy rysunek zawiera trzy wykresy wizualizujące dane. Pierwszy z nich przedstawia boxploty dla wyestymowanych prawdopodobieństw, w podziale na warianty genu i grupy oraz dodatkowo p-wartości i skorygowane (poprawką Holma) p-wartości dla trzech testów. Na czerwono zaznaczono p-wartości i skorygowane p-wartości testów, które wskazują istotność statystyczną na poziomie 10%. Drugi wykres prezentuje uśrednione prawdopodobieństwa występowania wariantów w podziale na grupy. Jest to wykres przydatny do oceny skali różnicy, gdy testy wskażą istotność statystyczną. Trzeci wykres pokazuje wszystkie wyestymowane prawdopodobieństwa, w podziale na warianty oraz pacjentów. Na wykresie tym możemy zaobserwować zależności w pomiarach pochodzących od tego samego pacjenta, czego nie da się zrobić patrząc tylko na dwa pierwsze wykresy. Dodatkowo wykres pierwszy oraz trzeci mogą być przydatne do zrozumienia, czemu testy nie odrzucają hipotezy o równości średnich, mimo że uśrednione wartości wydają się znacznie różnić. Może tak się dziać, np. gdy wariancje w danych są bardzo duże. Opisy otrzymanych wyników znajdują się pod rysunkami.



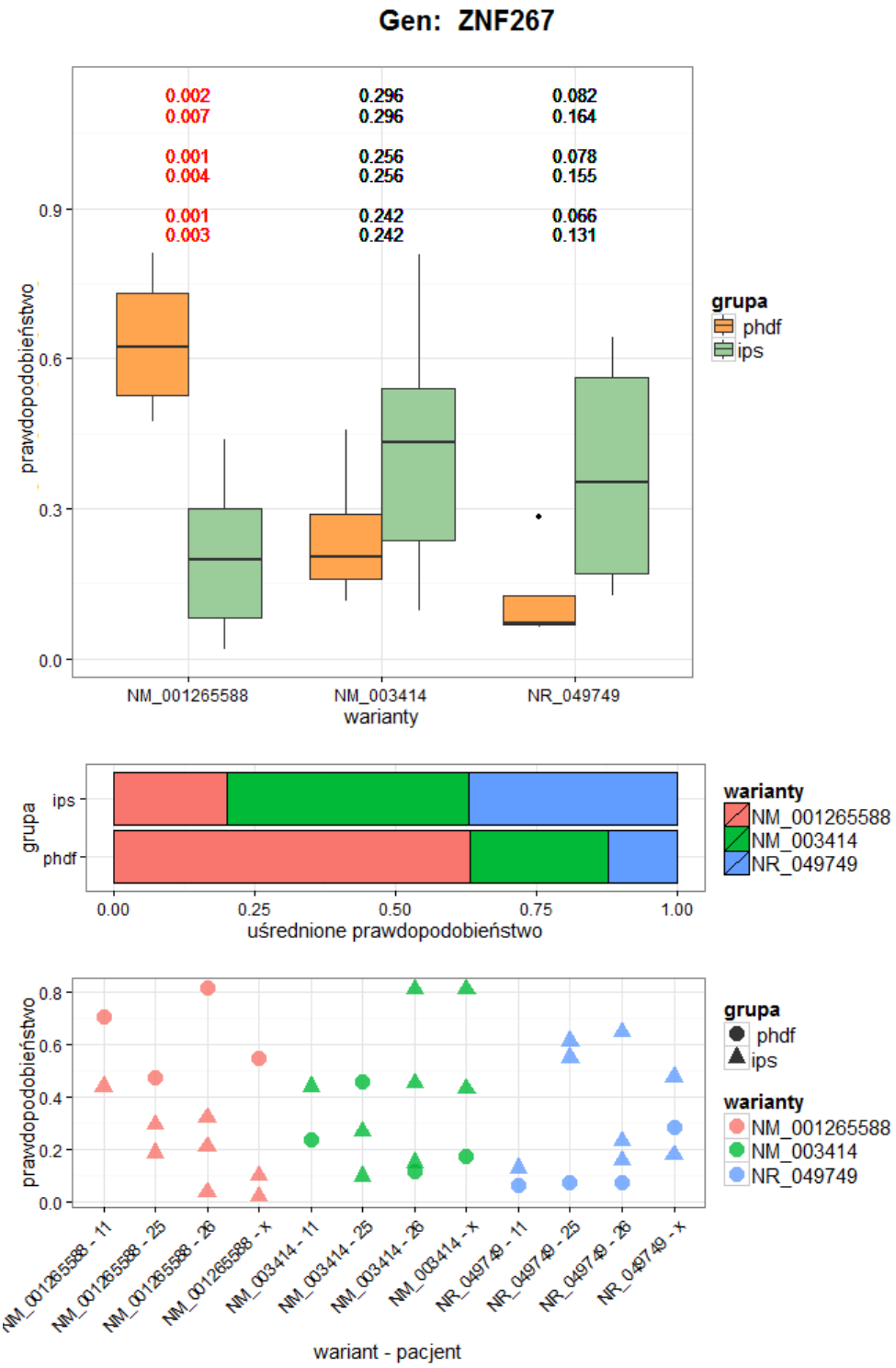
Rysunek 4.2: Gen ZNF436, występujący w dwóch wariantach. Test dla każdego z wariantów, to tak naprawdę ten sam test, ponieważ wyestymowane prawdopodobieństwa dla każdej próbki muszą sumować się do 1. W związku z tym, p-wartości testów są takie same oraz nie jest stosowana poprawka na wielokrotne testowanie. W tym genie, wariant NM_001077195 występuje średnio w 75% przypadków w grupie iPS oraz 38% przypadków w grupie PHDF. Biorąc pod uwagę możliwość występowania zależności między próbkami pochodzącymi od tego samego pacjenta, różnica ta jest istotna statystycznie na poziomie 10%. Każdy z testów odrzuca hipotezę o równości średnich w grupach.



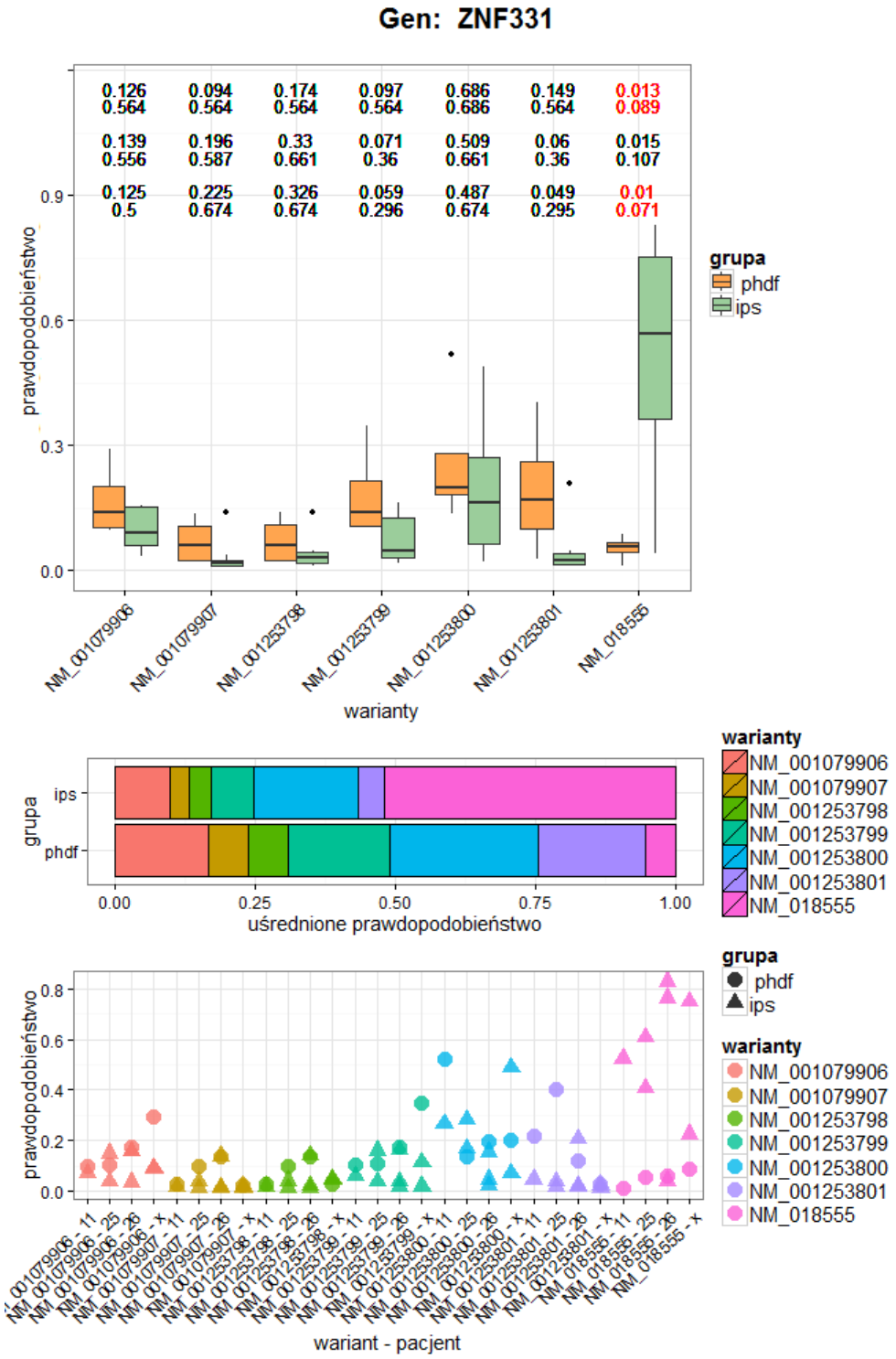
Rysunek 4.3: Gen ZNF714, występujący w 4 izoformach. Dwie z nich (NM_182515 i NR_117088) są uznane przez testy za istotnie różniące się między grupami, pod względem częstości występowania. Wariant NR_117088 występuje w znikomej ilości w każdej z grup, jednak grupy różnią się istotnie. Nawet taka niewielka różnica może mieć jakieś skutki biologiczne, więc nie należy jej lekceważyć. Zwróćmy jeszcze uwagę na wariant NR_117086, dla którego uśrednione prawdopodobieństwa występowania wynoszą ok. 5% w grupie iPS i 12% w grupie PHDH. Różnica jest ponad dwukrotna, mogłoby się wydawać, że grupy się różnią, jednak testy nie odrzucają hipotezy zerowej. Na trzecim wykresie widzimy, że duża wartość średniego prawdopodobieństwa w grupie PHDF była spowodowana dużą wartością dla pacjenta z numerem 11.



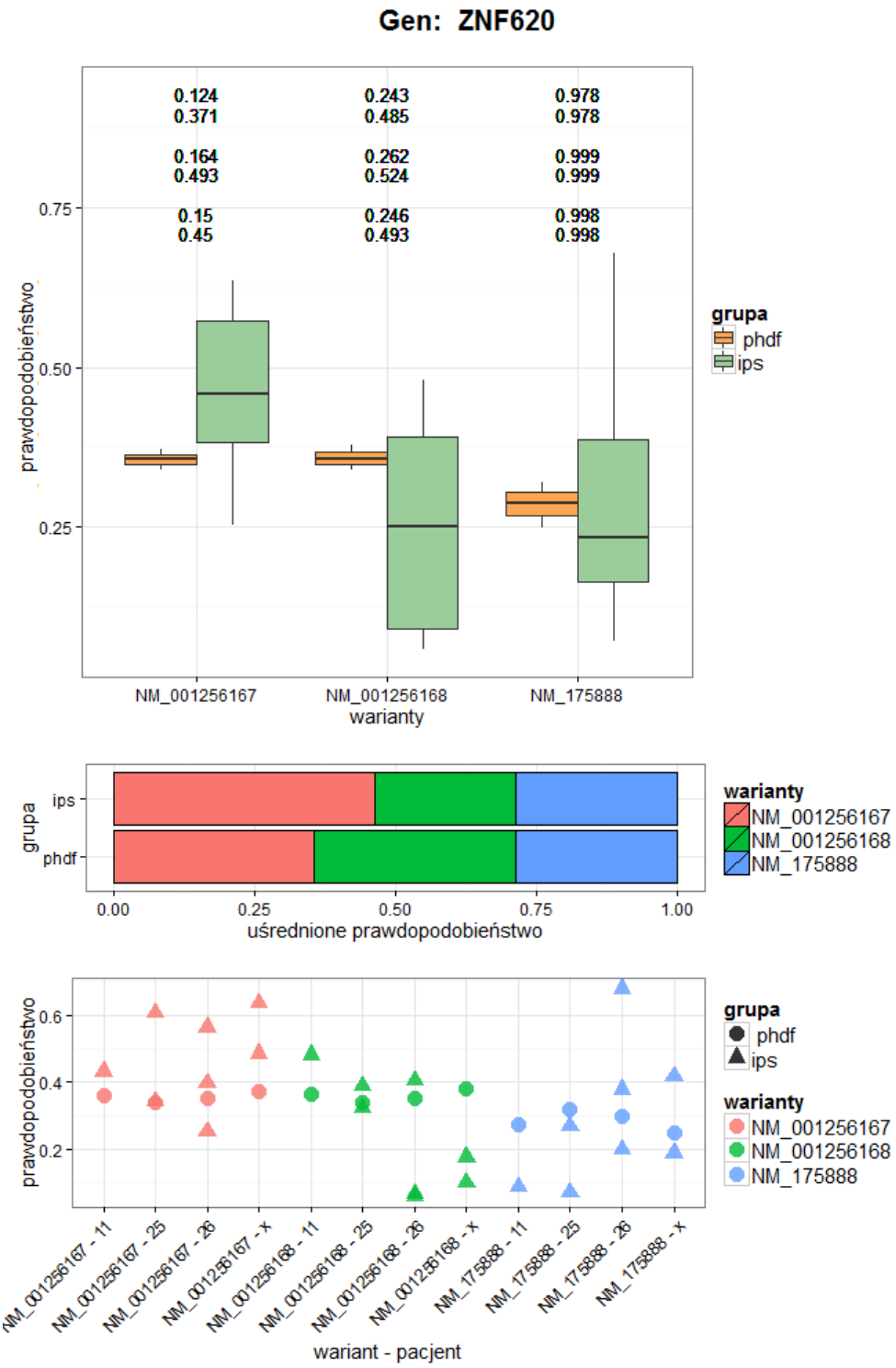
Rysunek 4.4: Gen ZNF74, mający pięć wariantów. Cztery z nich są uznane przez testy za istotnie różniące się między grupami, pod względem prawdopodobieństwa występowania (wariant NM_003426 ma skorygowaną p-wartość większą od 0.1 dla testu ilorazu wiarogodności, ale pozostałe dwa testy odrzucają hipotezę zerową). Warto zwrócić uwagę na to, że uśrednione prawdopodobieństwa dla wariantu NR_046282 są mniej więcej równe prawdopodobieństwom wspomnianego wcześniej wariantu NR_117086 genu ZNF714, dla którego testy nie odrzucają hipotezy o równości średnich prawdopodobieństw, natomiast tutaj hipoteza jest odrzucana.



Rysunek 4.5: Gen ZNF267, występujący w trzech wariantach. Jeden z nich (NM_001265588) uznany jest za istotnie różniący się, pod względem prawdopodobieństwa między grupami. Izoforma ta występuje w ok. 20% przypadków ekspresji genu w grupie iPS i 63% przypadków w grupie PHDF.

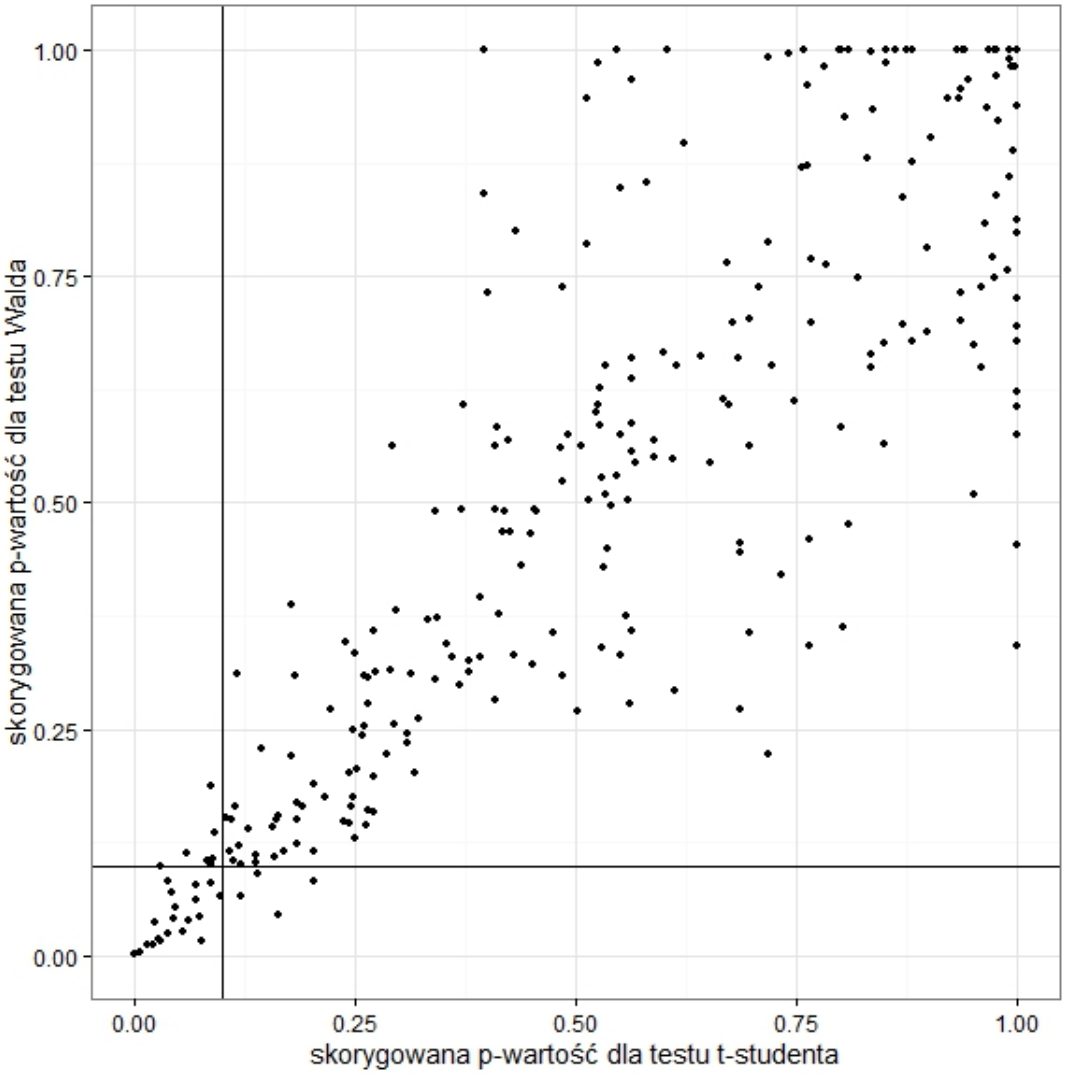


Rysunek 4.6: Gen ZNF331, występujący aż w 7 izoformach. Jedynie jedna z nich, NM_018555, jest wskazana przez dwa testy, jako różniąca się, pod względem prawdopodobieństwa między grupami. Uśrednione prawdopodobieństwa różnią się znacznie między grupami- 52%w grupie iPS i 6% w grupie PHDF, a mimo to test Walda nie odrzuca hipotezy zerowej o równości prawdopodobieństw. Może to powodować jedna odstająca obserwacja pacjenta z numerem 26 w grupie iPS.

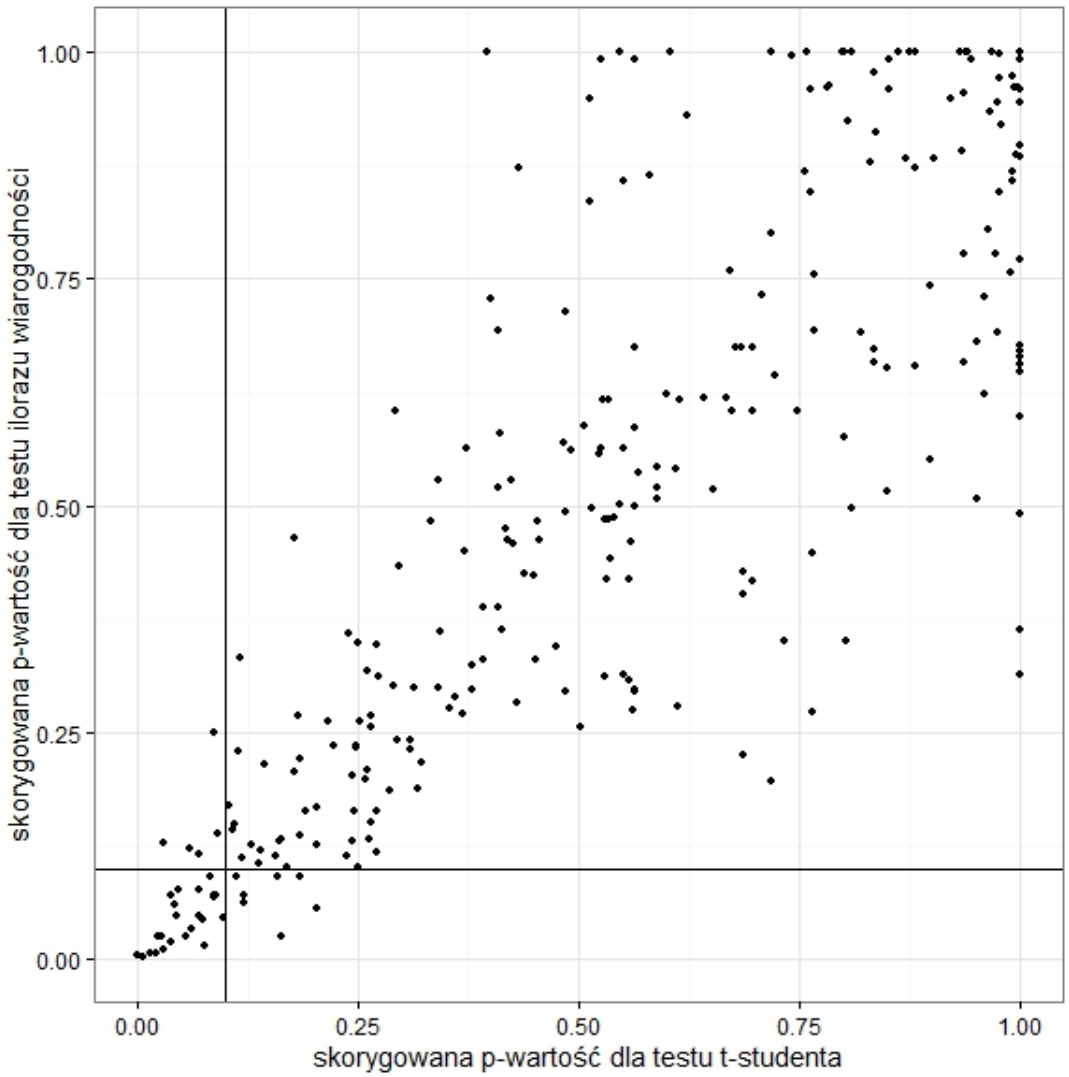


Rysunek 4.7: Gen ZNF620, o którym wspomnieliśmy wcześniej, iż spodziewano się, że w komórkach iPS będzie on występował głównie w wariancie NM_001256167. Testy statystyczne nie wykazują natomiast różnic w rozkładach wariantów, a żadna z izoform silnie nie dominuje.

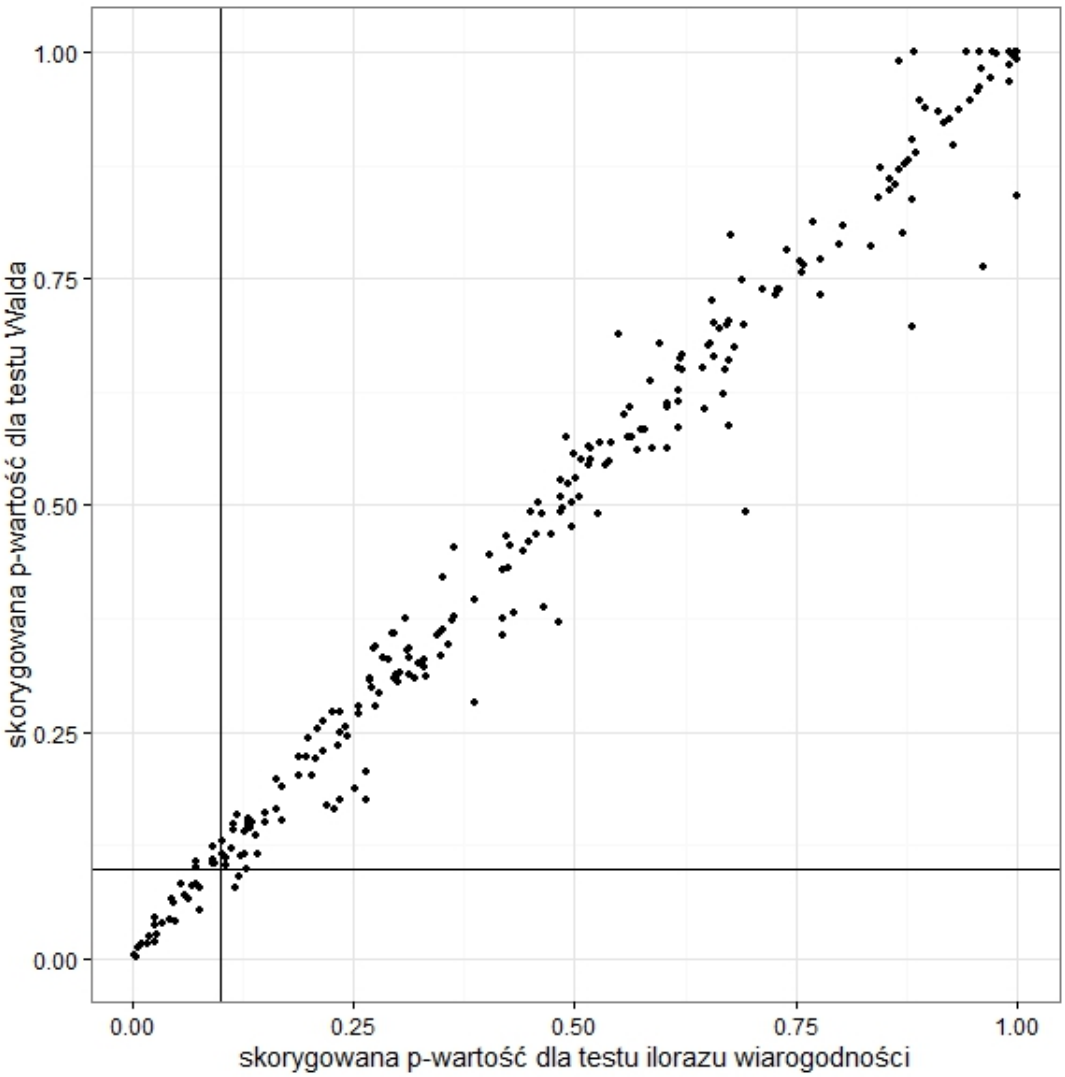
Na rysunkach 4.8 - 4.10 przedstawiono skorygowane p-wartości dla każdej pary, z trzech zastosowanych testów. Dodatkowo na rysunkach zaznaczono linie dla wartości 0,1, aby można było zaobserwować, ile hipotez jest odrzucanych przez każdy z testów.



Rysunek 4.8: Skorygowane p-wartości testu t-studenta oraz testu Walda dla wszystkich analizowanych wariantów genów.



Rysunek 4.9: Skorygowane p-wartości testu t-studenta oraz testu ilorazu wiarogodności dla wszystkich analizowanych wariantów genów.



Rysunek 4.10: Skorygowane p-wartości testu ilorazu wiarogodności oraz testu Walda dla wszystkich analizowanych wariantów genów.

Widzimy, że dla każdej pary testów, istnieje kilka wariantów, dla których hipoteza zerowa jest odrzucana przez jeden z nich oraz nieodrzucana przez drugi. Obserwujemy również, że test Walda oraz test ilorazu wiarogodności dają zbliżone p-wartości, podczas gdy mogą one się znacznie różnić od p-wartości testu t-studenta. W rozdziale 3.4 pokazaliśmy, że testy Walda oraz ilorazu wiarogodności są lepsze, niż test t-studenta, jednak żaden z tych testów nie okazał się najlepszym z wszystkich trzech. W związku ze zbliżonymi p-wartościami testów Walda i ilorazu wiarogodności, geny statystycznie różniące się pod względem rozkładu wariantów splicingowych, możemy badać na podstawie dowolnego z tych dwóch testów, wybranego apriori.

Zakończenie

Pierwszym z celów pracy było zaprezentowanie metody estymacji prawdopodobieństw występowania wariantów splicingowych genów. Cel został zrealizowany. W rozdziale 2 przedstawiony został algorytm Casper, stosowany dla każdego genu oraz pomiaru osobno. Metoda ta zakłada, że obserwowane dane pochodzą z mieszaniny dyskretnych rozkładów prawdopodobieństwa. Estymujemy wagi tej mieszaniny. Używamy w tym celu algorytmu EM (Expectation - Maximization), maksymalizującego funkcję a-posteriori parametrów. Wynikiem działania algorytmu Casper są wyestymowane prawdopodobieństwa występowania wariantów splicingowych genu, dla każdego pomiaru osobno.

Drugim celem pracy było zaproponowanie metod, które pozwolą porównać ze sobą dwie grupy komórek. Grupy te porównujemy pod względem rozkładu występowania poszczególnych wariantów splicingowych, na podstawie oszacowanych prawdopodobieństw. Zaproponowano porównanie wyestymowanych prawdopodobieństw w dwóch grupach dla każdego wariantu osobno. Następnie uwzględniano poprawkę na wielokrotne testowanie, związane z liczbą wariantów genu. Gen był uznawany za istotnie różniący się pod względem rozkładu wariantów, jeśli dla chociaż jednego wariantu genu, test odrzucał hipotezę zerową o równości średnich prawdopodobieństw w dwóch grupach. Do porównania wyestymowanych prawdopodobieństw w dwóch grupach komórek zaproponowane zostały cztery testy statystyczne:

- klasyczny test t-studenta dla dwóch prób,
- modyfikacja testu t-studenta dla prób sparowanych,
- test Walda, weryfikujący istotność efektu stałego, w modelu liniowym z efektami losowymi,
- test ilorazu wiarogodności, weryfikujący istotność efektu stałego, w modelu liniowym z efektami losowymi.

W związku z występowaniem zależności w analizowanych danych, nie założono, że statystyki wymienionych testów mają teoretyczne rozkłady, takie jak dla danych niezależnych. Był to bardzo ważny element pracy. Rozkłady statystyk testowych wyznaczono symulacyjnie. Zbadano również wrażliwość tych statystyk na zmiany rozkładów i korelacje w danych.

Pierwszy z zaproponowanych testów okazał się być wrażliwy na występowanie zależności w danych, w związku z czym nie był uwzględniany w dalszych analizach. Statystyki pozostałych trzech testów nie wykazały wrażliwości na zmiany rozkładów oraz występowanie korelacji w danych.

W celu wyboru najlepszego z zaproponowanych testów, porównano je ze względu na moc statystyczną. Testy Walda oraz ilorazu wiarogodności okazały się być według tego kryterium lepsze, niż test t-studenta. Żaden z tych dwóch testów nie był jednak lepszy od wszystkich pozostałych. Dalsze analizy pokazały, że oba testy dają podobne p-wartości, więc uznano, że są porównywalne i można używać dowolnego z nich.

Ostatnim celem pracy było użycie zaproponowanych metod do rzeczywistych danych oraz zidentyfikowanie genów, które istotnie różnią się pod względem rozkładu wariantów w pierwotnych fibroblastach skóry (PHDF) oraz indukowanych komórkach pluripotentnych (iPS). Analizie poddano 349 genów z domeną KRAB-ZNF. Wynikiem analizy miał być zbiór genów, które mają istotnie różne rozkłady wariantów w zależności od grupy. Do analizy użyto wszystkich trzech testów, tzn. zmodyfikowanego testu t-studenta dla prób sparowanych, testu Walda i testu ilorazu wiarogodności. Jedynie kilka wariantów zostało zakwalifikowanych sprzecznie przez różne testy. Geny, dla których rozkład wariantów różnił się istotnie między grupami, przynajmniej przy zastosowaniu jednego testu, zostały przekazane do Wielkopolskiego Centrum Onkologii i poddane dalszej weryfikacji biologicznej.

Bibliografia

[1] H. Fletcher, I. Hickey, P. Winter, *Krótkie wykłady GENETYKA*, Wydawnictwo Naukowe PWN, Warszawa, 2010,

[2] T.A. Brown, *genomy*, Wydawnictwo Naukowe PWN, Warszawa, 2009,

[3] *An Introduction to Next-Generation Sequencing Technology* `http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`,

[4] *Sekwencjonowanie nowej generacji coraz częstszym gościem w laboratoriach* `http://biotechnologia.pl/biotechnologia/artykuly/sekwencjonowanie-nowej-generacji-coraz-czestszym-gosciem-w-laboratoriach,14846.html`,

[5] D. Rossell, C.S.O. Attolini, M. Kroiss, A. Stöcker, *Quantifying alternative splicing from paired-end RNA-sequencing data*, The Annals of Applied Statistics, Volume 8, Number 1 (2014), str. 309-330,

[6] M. A.T. Figueiredo, *Lecture Notes on the EM Algorithm*, Portugal, 2004,

[7] R. Huptas, *Zastosowanie algorytmu EM do estymacji parametrów rozkładu na podstawie danych pogrupowanych*, Zeszyty Naukowe nr 740 Akademii Ekonomicznej w Krakowie, 2007, str. 131-145,

[8] R. Magiera, *Modele i Metody Satystyki Matematycznej, Część I, Rozkłady i Symulacja Stochastyczna*, Wydanie drugie rozszerzone, Oficyna Wydawnicza GiS, Wrocław, 2007,

[9] P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Wydawnictwo Naukowe PWN, Warszawa, 2013,

[10] K. Archacka, I. Grabowska, M.A. Ciemerych, *Indukowane komórki pluripotentne - nadzieje, obawy i perspektywy*, Postępy biologii komórki, Tom 37, 2010 nr 1, str. 41-62,

[11] *TopHat, A spliced read mapper for RNA-Seq, Manual,* `https://ccb.jhu.edu/software/tophat/manual.shtml`,

[12] *Manual for the R casper package* http://www.bioconductor.org/packages/release/bioc/vignettes/casper/inst/doc/casper.pdf.

Paulina Auguścik
Nr albumu 237478

Warszawa, 1 grudnia 2015

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Metody statystycznej identyfikacji zmian wariantów splicingowych wraz z przykładami zastosowań w analizie danych RNA-Seq", której promotorem jest dr hab. inż. Przemysław Biecek prof. nadzw., wykonałam samodzielnie, co poświadczam własnoręcznym podpisem.

.....

Paulina Auguścik