

**Testowanie odsetka odrzuceń prawdziwej hipotezy głównej w testach istotności współczynników korelacji**

Autorzy:

Majka Miezianko

Marek Czerwonka

*Projekt zaliczeniowy z przedmiotu:*

*Metody Nieparametryczne w Statystyce*

**Kraków 2021**

**Spis treści**

**Wstęp 3**

**1. Krótka charakterystyka testów oraz danych 3**

**2. Parametry oraz wyniki przeprowadzonych symulacji 5**

**3. Podsumowanie i wnioski oraz obserwacje 28**

**Wstęp**

W ramach projektu zrealizowane zostanie następujące zadanie:

*Za pomocą odpowiednich symulacji zbadać odsetek odrzuceń prawdziwej hipotezy głównej (tzn. błąd I rodzaju) w teście istotności współczynnika korelacji w przypadku, gdy dane pochodzą z rozkładu normalnego o różnych parametrach. W trakcie symulowania danych należy pamiętać, by dane spełniały hipotezę główną. Porównać uzyskane wyniki dla wsp. korelacji Pearsona i Spearmana. Uzyskane wyniki należy opisać w zależności od:*

* *liczby danych;*
* *parametrów rozkładów.*

**1. Krótka charakterystyka testów oraz danych**

W projekcie zostanie wykonana procedura testowania statystycznej istotności współczynników korelacji w układzie hipotez: hipoteza główna obejmuje przypadek współczynnika korelacji równego zeru, zaś hipoteza alternatywna współczynnika korelacji różnego od zera (dwustronny obszar krytyczny testu).

Dla obu współczynników korelacji statystyka testu zbudowana jest w taki sam sposób. Wyznaczamy ją jako:

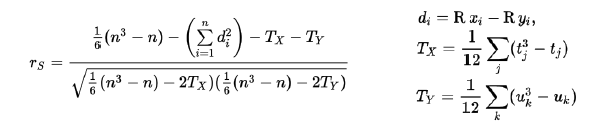


Przy założeniu prawdziwości hipotezy głównej statystyka ta ma rozkład t-Studenta o (n-2) stopniach swobody, gdzie n jest klasycznie liczbą par danych poddawanych analizie.

Zwróćmy uwagę, że dla dwóch analizowanych współczynników korelacji postać statystyki testu jest dokładnie taka sama (istnieją także miary korelacji, dla których postać statystyki byłaby inna, najpowszechniej używaną są odmiany współczynnika korelacji -Kendalla.

Przeglądając – nazwijmy to szumnie – literaturę[[1]](#footnote-1), natknęliśmy się na źródło, które skłoniło nas do drobnego rozszerzenia tematyki projektu, naturalnie przede wszystkim w trosce o danie szansy Recenzentom na ewentualne wytknięcie nam potknięć, a także ponieważ uznaliśmy to za interesujące.

Powszechnie używany bowiem wzór na wartość współczynnika ρ-Spearmana jest prawdziwy tak naprawdę tylko w przypadku braku rang wiązanych. Dla przypadku występowania rang wiązanych prawdziwa jest inna, bardziej rozbudowana formuła:



My wprawdzie generujemy dane z rozkładu ciągłego, więc kwestia rang wiązanych zasadniczo nas nie dotyczy, ale mimo wszystko zbadajmy przynajmniej w takim razie jakość implementacji tych dwóch procedur w pakietach R.

Można także po porangowaniu danych wyznaczyć dla rang współczynnik korelacji r-Pearsona i wynik powinien być dokładnie zgodny z powyżej zacytowanym wzorem.

Jeszcze jedna istotna uwaga metodologiczna: o ile wartość samych współczynników korelacji oraz statystyki testu jesteśmy w stanie wyznaczyć w pełni analitycznie, o tyle oszacowanie wartości p\_value odbywa się numerycznie (i występują różnice w algorytmach prowadzących do tego numerycznego oszacowania). Zatem różne algorytmy mogą dawać dla tej samej wartości statystyki testu nieco inne (choć powinny one być bardzo zbliżone) wartości p\_value. Przykładem testu, w którym wyznaczamy dokładne p\_value jest test dokładny Fishera (niezależności zmiennych dla tablic o rozmiarach 2x2) i jest to sytuacja wyjątkowa (być może jest to jedyny w powszechnym użyciu test z dokładnie wyliczaną wartością p\_value).

Ponieważ w naszym projekcie – jak przypuszczamy – dokonamy oszacowania tych p\_value właśnie dwoma różnymi algorytmami (raz skorzystamy z bezpośrednio zaimplementowanej procedury dla testu istotności współczynnika korelacji ρ-Spearmana, a drugi raz wyznaczymy te wartość jako współczynnik korelacji r-Pearsona dla rang, chcielibyśmy zaznaczyć, że nie byłby do końca prawdziwy potencjalny zarzut, iż porównujemy tu wodę z monotlenkiem diwodoru czy też z kwasem hydroksylowym albo z wodorotlenkiem wodoru i dochodzimy do wniosku, że jest to nadal woda.

Pewną wartość poznawczą jednak takie porównanie naszym zdaniem posiada.

Dane generujemy z rozkładów normalnych (rozkłady dwuparametryczne).

Tak jak w poprzednim projekcie staraliśmy się tak napisać kod, aby ewentualne modyfikacje parametrów były relatywnie łatwe dla recenzentów (o ile będzie im się chciało te parametry modyfikować).

**2. Parametry oraz wyniki przeprowadzonych symulacji**

Załączony do projektu kod R umożliwia dość swobodne manipulowanie parametrami przeprowadzanych symulacji. Parametry prowadzące do uzyskania poniżej zaprezentowanych, przykładowych wyników były następujące:

Liczba symulacji dla każdego pojedynczego zestawu poniższych danych: 1 000

Liczba obserwacji generowanych w pojedynczej symulacji: 500, 1000, 5000

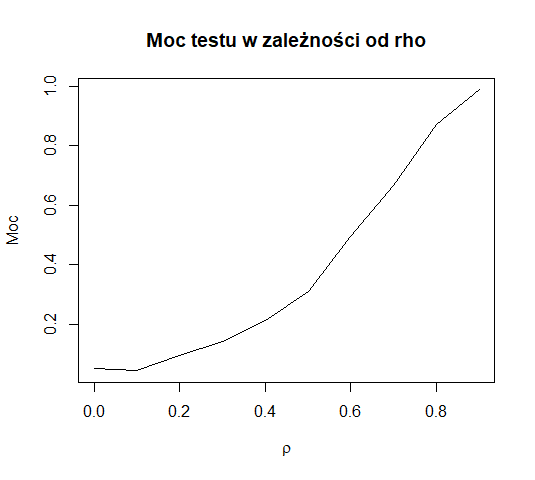
Poziomy istotności: 0,02; 0,05; 0,10 oraz 0,20

Średnie: 1, 6, 10, 15, 20, 25, 30, 40

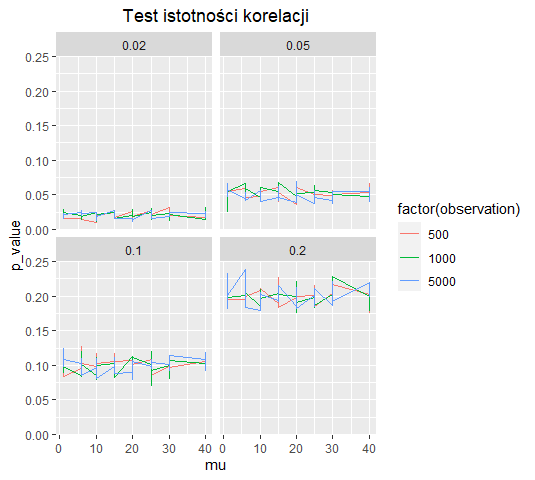
Odchylenia standardowe: 1, 3, 9, 15, 25, 35

Wielkości pojedynczych prób: od 6 do 80 par danych

W zakresie analizy wstępnej, której głównym zadaniem było upewnienie się, że nie mamy jakichś grubych błędów w kodzie, przeprowadziliśmy analizę mocy testu dla istotności statystycznej współczynnika korelacji r-Pearsona. Oczekiwaliśmy monotonicznie rosnącej mocy testu dla zwiększającego się modułu współczynnika korelacji próby i taki też wynik otrzymaliśmy.



Następnie przeprowadziliśmy zasadniczą część analizy w tym projekcie dla współczynnika korelacji r-Pearsona, czyli zbadaliśmy jak zmiana obu parametrów rozkładu normalnego, z którego generujemy dane wpływa na zmianę odsetka odrzuceń prawdziwej hipotezy głównej dla różnych liczb obserwacji (kolory). Zarówno średnia jak i odchylenie standardowe zmieniają się w zakresie wskazanym w początkowej części projektu.



Zgodnie z oczekiwaniami (dla pewności wykonaliśmy to dla czterech różnych poziomów istotności) oceniamy, że taka zależność nie występuje.

*De facto* na tym wykresie mamy pewne pionowe zakresy (rozstępy), dla ustalonych w kodzie wartości średniej wynikające różnych wyników uzyskanych dla różnych symulacji wykonanych przy różnych wartościach odchylenia standardowego. Kreski „ukośne”, którymi są one połączone nie mają zasadniczo żadnej sensownej interpretacji. Pozostawiliśmy je dlatego, że czynią bardziej czytelnym brak systematycznej zależności. W dalszej części na niektórych wykresach dla porównania zastosowaliśmy inną prezentację danych (w postaci surowych punktów), aby było jasne co oznaczają tak naprawdę te dziwne na pierwszy rzut oka powyższe wykresy.

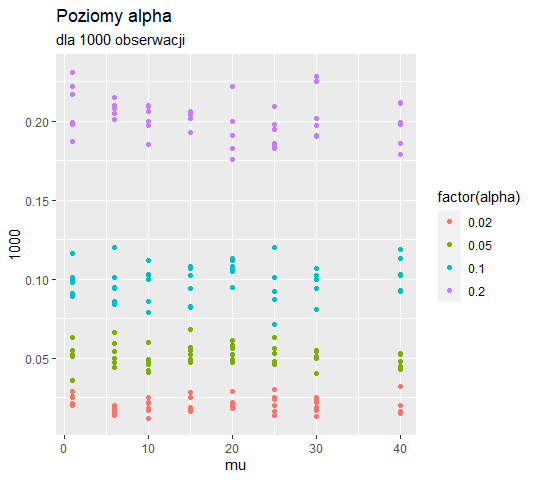
Poniżej prezentujemy rezultat takich samych symulacji jak powyżej dla współczynnika korelacji r-Pearsona, jednakże tym razem oś pozioma reprezentuje wartości odchyleń standardowych, zaś pionowe odcinki reprezentują rozstępy uzyskanych wartości p\_value (dla średniej rozkładu, która również się zmienia), z którego generujemy dane zmieniającej się w przedziale podanym w początkowej części projektu.



Zgodnie z oczekiwaniami (dla pewności wykonaliśmy to dla czterech różnych poziomów istotności) taka zależność nie występuje.

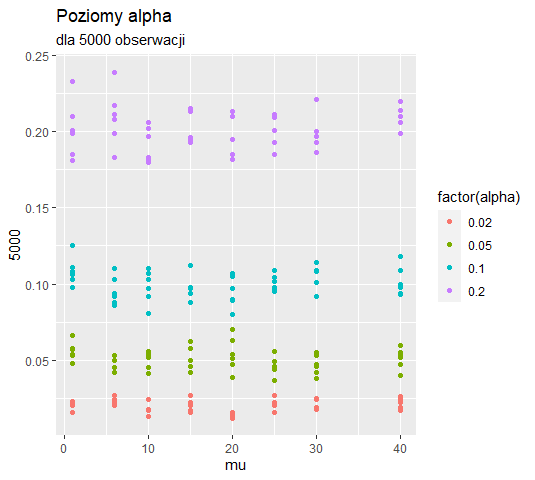
Co do prezentacji danych na wykresie, uwagi analogicznie jak do poprzedniego wykresu.

W kolejnym kroku zbadaliśmy jak dla wielkości próby równej 1000 wyglądają eksperymentalnie uzyskane wartości p\_value wobec ich wartości oczekiwanych przy zmianie obydwu parametrów rozkładu, z którego generujemy dane (dla r-Pearsona).



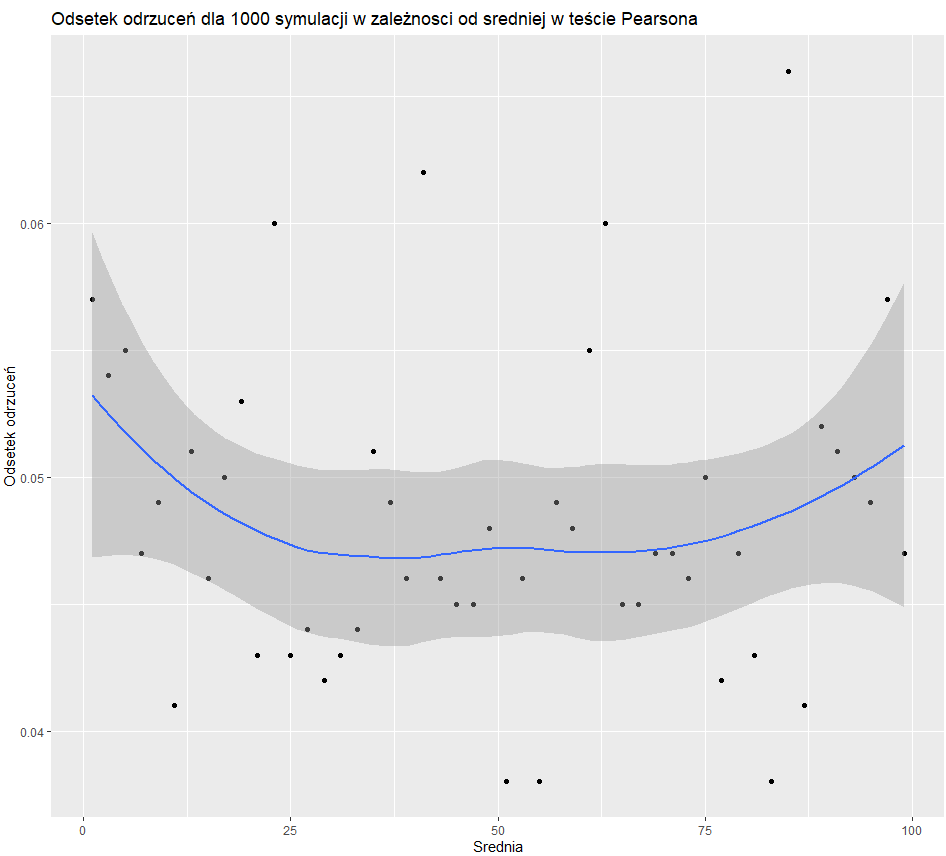
Nie dostrzegamy żadnych wyraźnych tendencji, co jest zgodne z oczekiwaniami. Procentowy rozrzut wokół wartości oczekiwanych jest podobny na każdym ze sprawdzonych poziomów istotności (nominalnie naturalnie te rozrzuty się różnią).

W kolejnym kroku zbadaliśmy jak dla wielkości próby równej 5000 wyglądają eksperymentalnie uzyskane wartości p\_value wobec ich wartości oczekiwanych przy zmianie obydwu parametrów rozkładu, z którego generujemy dane (dla r-Pearsona).



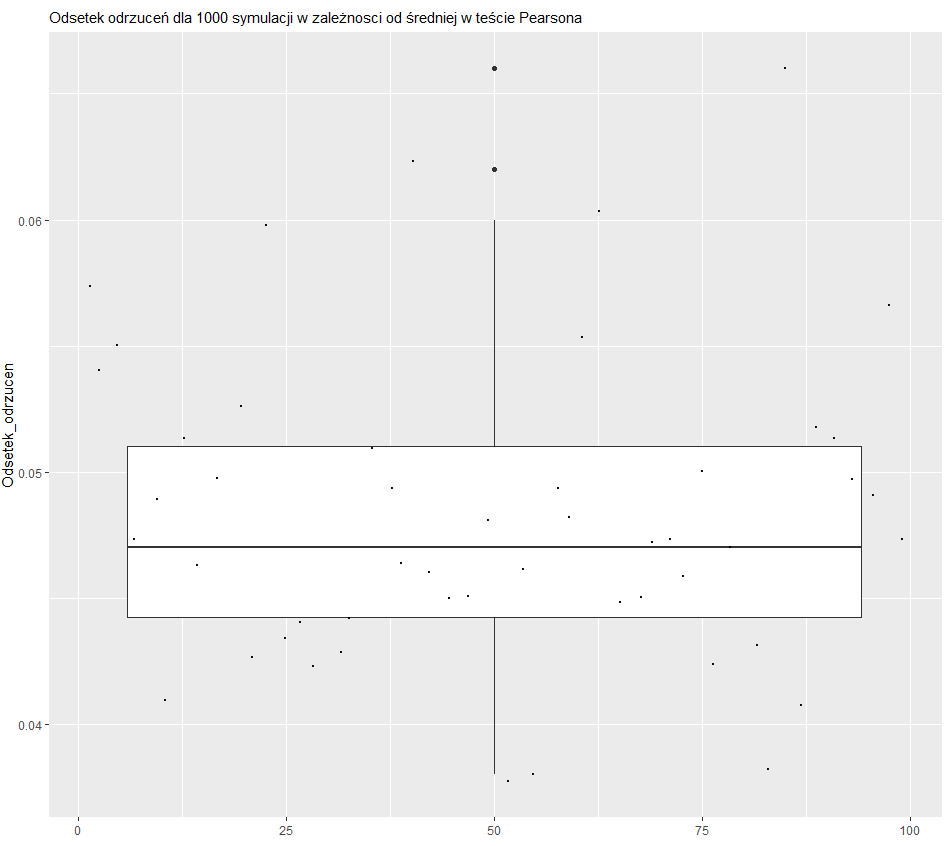
Nie dostrzegamy żadnych wyraźnych tendencji, co jest zgodne z oczekiwaniami. Procentowy rozrzut wokół wartości oczekiwanych jest podobny na każdym ze sprawdzonych poziomów istotności (nominalnie naturalnie te rozrzuty się różnią).

Poniżej prezentujemy eksperymentalną krzywą odsetka odrzuceń hipotezy głównej (przy przyjętym poziomie istotności 0,05) w zależności od zmienianej średniej arytmetycznej przy stałym odchyleniu standardowym.



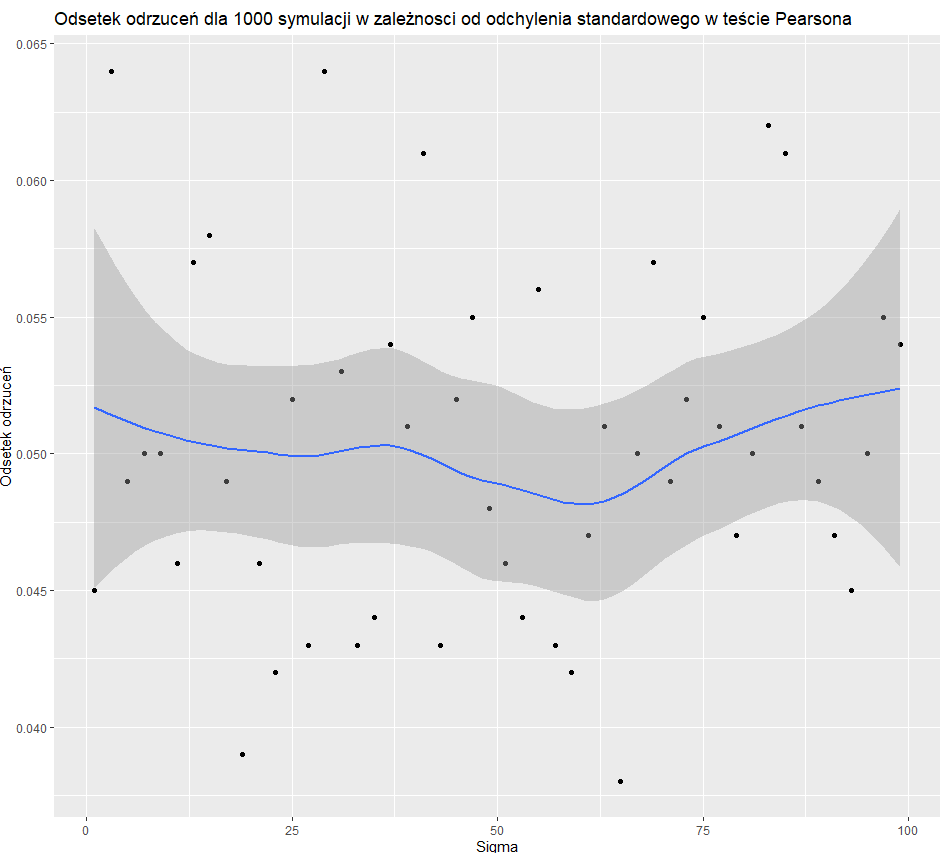
Uzyskaną wygładzoną postać zależności uważamy za artefakt obliczeniowy, nie widzimy podstaw aby rzeczywista zależność nie miała charakteru prostej poziomej (czyli *de facto* braku zależności od średniej).

Poniżej prezentujemy wykres obrazujący eksperymentalnie uzyskany rozkład odsetka odrzuceń.



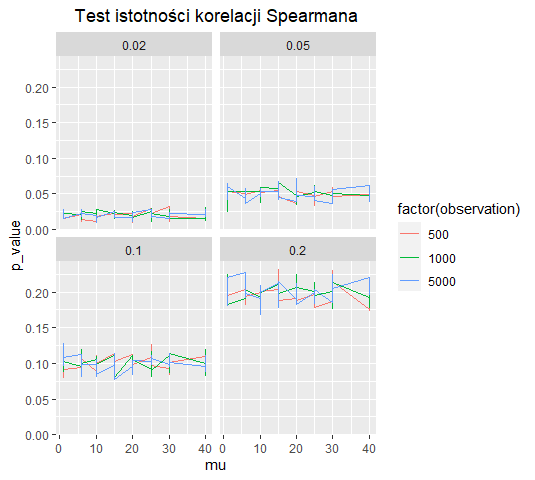
Naszym zdaniem widoczna asymetria (przynajmniej jest ona w oczekiwaną stronę) może być uzasadniona tym, że p\_value z pewnością nie jest zmienną o charakterze addytywnym. Przesunięcie mediany nieco w dół wydaje się być artefaktem obliczeniowym powiązanym ze zbyt niską liczbą symulacji.

Poniżej prezentujemy eksperymentalną krzywą odsetka odrzuceń hipotezy głównej (przy przyjętym poziomie istotności 0,05) w zależności od zmienianego odchylenia standardowego przy stałej średniej.



Uzyskaną wygładzoną postać zależności uważamy za artefakt obliczeniowy, nie widzimy podstaw aby rzeczywista zależność nie miała charakteru prostej poziomej (czyli *de facto* braku zależności od średniej).

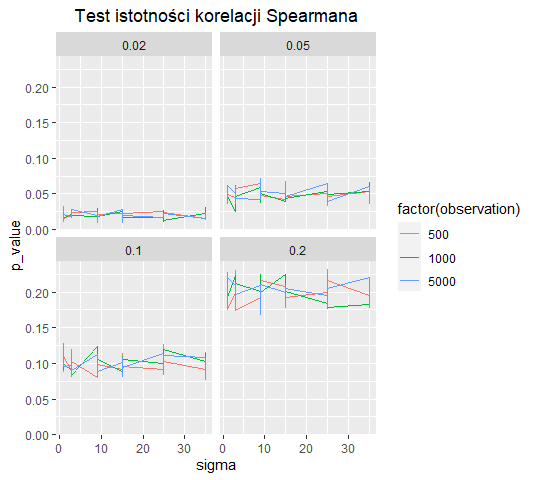
Następnie przeprowadziliśmy zasadniczą część analizy w tym projekcie dla współczynnika korelacji ρ-Spearmana, czyli zbadaliśmy jak zmiana obu parametrów rozkładu normalnego, z którego generujemy dane wpływa na zmianę odsetka odrzuceń prawdziwej hipotezy głównej dla różnych liczb obserwacji (kolory). Zarówno średnia jak i odchylenie standardowe zmieniają się w zakresie wskazanym w początkowej części projektu.



Zgodnie z oczekiwaniami (dla pewności wykonaliśmy to dla czterech różnych poziomów istotności) oceniamy, że taka zależność nie występuje.

*De facto* na tym wykresie mamy pewne pionowe zakresy (rozstępy), dla ustalonych w kodzie wartości średniej wynikające różnych wyników uzyskanych dla różnych symulacji wykonanych przy różnych wartościach odchylenia standardowego. Kreski „ukośne”, którymi są one połączone nie mają zasadniczo żadnej sensownej interpretacji. Pozostawiliśmy je dlatego, że czynią bardziej czytelnym brak systematycznej zależności. W dalszej części na niektórych wykresach dla porównania zastosowaliśmy inną prezentację danych (w postaci surowych punktów), aby było jasne co oznaczają tak naprawdę te dziwne na pierwszy rzut oka powyższe wykresy.

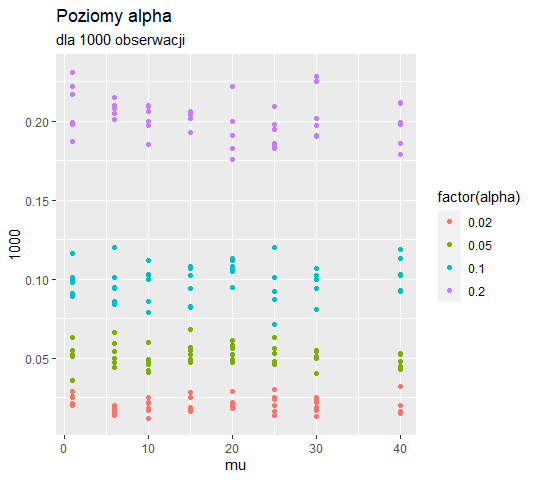
Poniżej prezentujemy rezultat takich samych symulacji jak powyżej dla współczynnika korelacji ρ-Spearmana, jednakże tym razem oś pozioma reprezentuje wartości odchyleń standardowych, zaś pionowe odcinki reprezentują rozstępy uzyskanych wartości p\_value (dla średniej rozkładu, która również się zmienia), z którego generujemy dane zmieniającej się w przedziale podanym w początkowej części projektu.



Zgodnie z oczekiwaniami (dla pewności wykonaliśmy to dla czterech różnych poziomów istotności) taka zależność nie występuje.

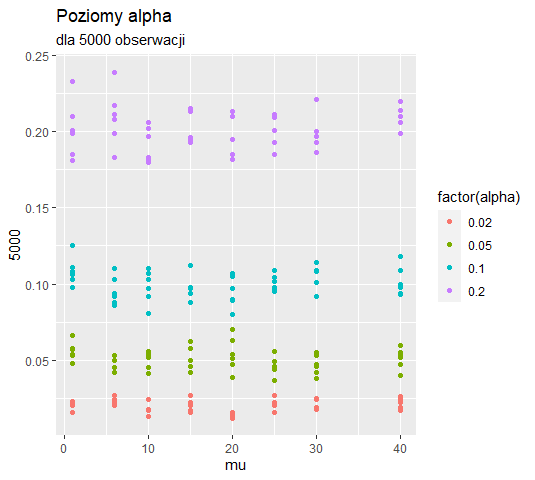
Co do prezentacji danych na wykresie, uwagi tak jak wcześniej.

W kolejnym kroku zbadaliśmy jak dla wielkości próby równej 1000 wyglądają eksperymentalnie uzyskane wartości p\_value wobec ich wartości oczekiwanych przy zmianie obydwu parametrów rozkładu, z którego generujemy dane (dla ρ-Spearmana).



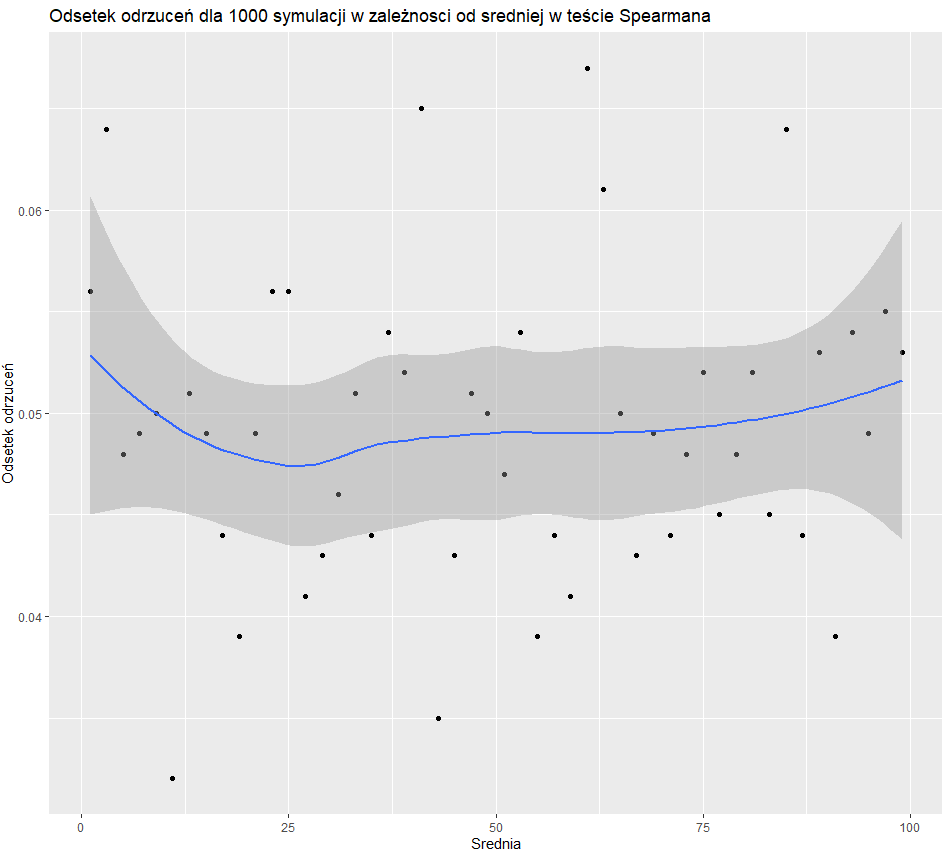
Nie dostrzegamy żadnych wyraźnych tendencji, co jest zgodne z oczekiwaniami. Procentowy rozrzut wokół wartości oczekiwanych jest podobny na każdym ze sprawdzonych poziomów istotności (nominalnie naturalnie te rozrzuty się różnią).

W kolejnym kroku zbadaliśmy jak dla wielkości próby równej 5000 wyglądają eksperymentalnie uzyskane wartości p\_value wobec ich wartości oczekiwanych przy zmianie obydwu parametrów rozkładu, z którego generujemy dane (dla ρ-Spearmana).



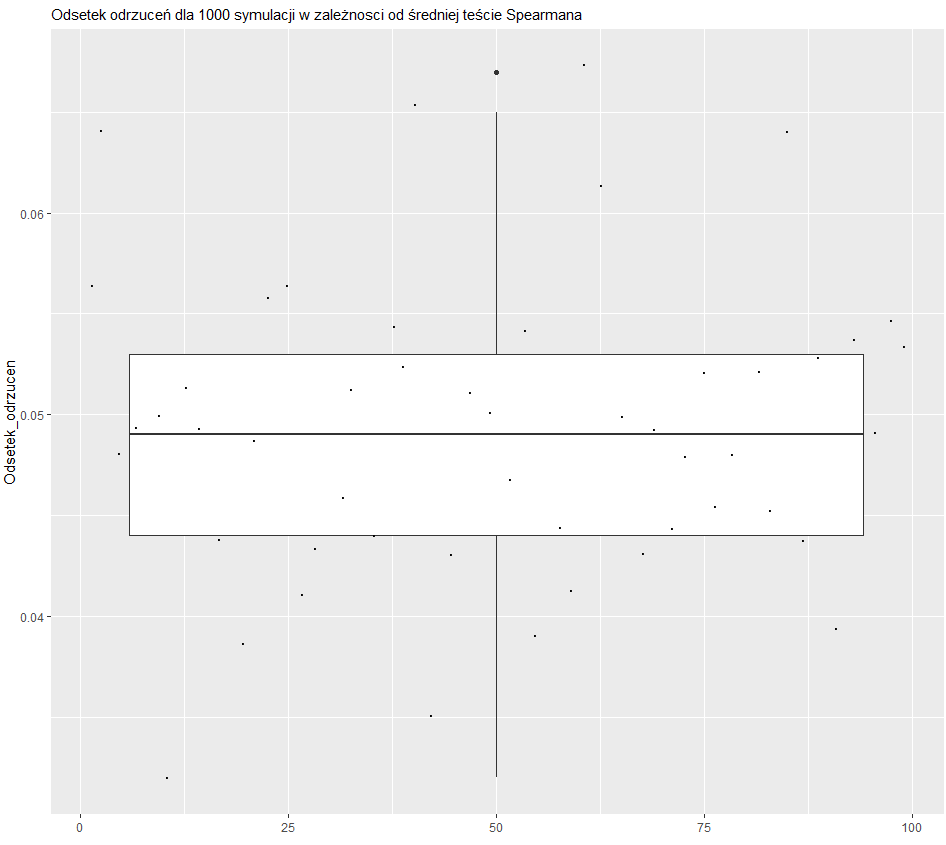
Nie dostrzegamy żadnych wyraźnych tendencji, co jest zgodne z oczekiwaniami. Procentowy rozrzut wokół wartości oczekiwanych jest podobny na każdym ze sprawdzonych poziomów istotności (nominalnie naturalnie te rozrzuty się różnią).

Poniżej prezentujemy eksperymentalną krzywą odsetka odrzuceń hipotezy głównej (przy przyjętym poziomie istotności 0,05) w zależności od zmienianej średniej arytmetycznej przy stałym odchyleniu standardowym.



Uzyskaną wygładzoną postać zależności uważamy za artefakt obliczeniowy, nie widzimy podstaw aby rzeczywista zależność nie miała charakteru prostej poziomej (czyli *de facto* braku zależności od średniej).

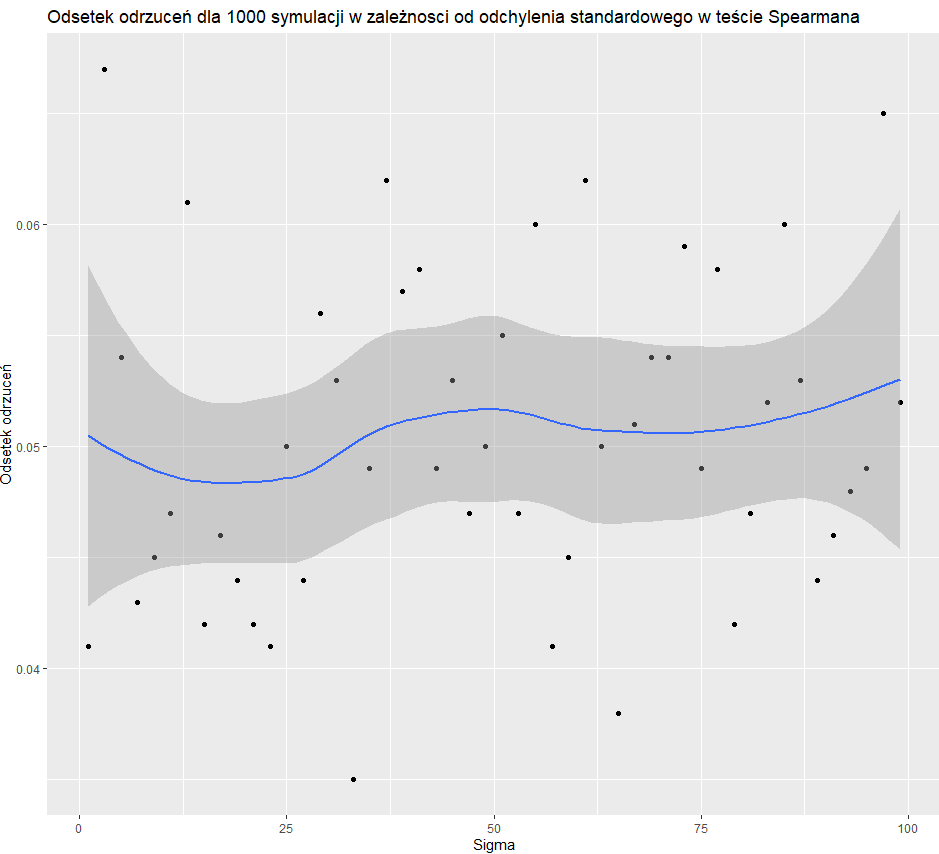
Poniżej prezentujemy wykres obrazujący eksperymentalnie uzyskany rozkład odsetka odrzuceń.

****

Naszym zdaniem widoczna asymetria (przynajmniej jest ona w oczekiwaną stronę) może być uzasadniona tym, że p\_value z pewnością nie jest zmienną o charakterze addytywnym. Przesunięcie mediany nieco w dół wydaje się być artefaktem obliczeniowym powiązanym ze zbyt niską liczbą symulacji.

Niemniej jednak analogiczny efekt otrzymaliśmy także dla r-Pearsona, więc pojawiły się u nas pierwsze przedwstępne wątpliwości czy to rzeczywiście przypadkowy artefakt czy ślad jakiejś realnej własności procedury testowej.

Poniżej prezentujemy eksperymentalną krzywą odsetka odrzuceń hipotezy głównej (przy przyjętym poziomie istotności 0,05) w zależności od zmienianego odchylenia standardowego przy stałej średniej.

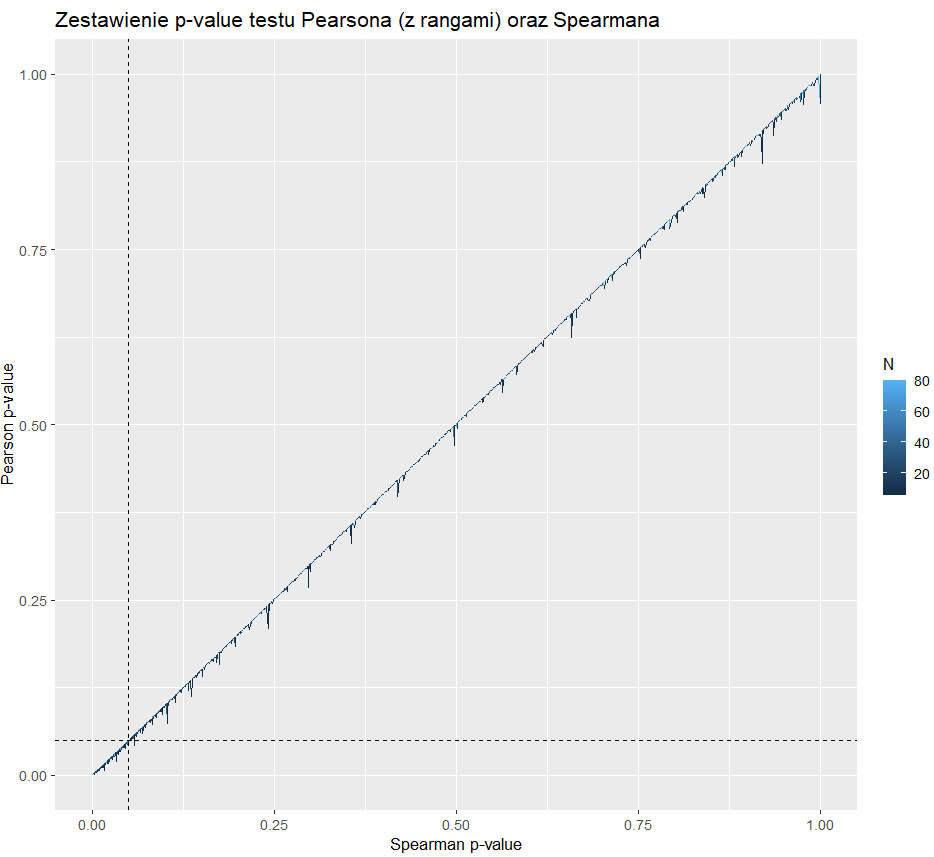


Uzyskaną wygładzoną postać zależności uważamy za artefakt obliczeniowy, nie widzimy podstaw aby rzeczywista zależność nie miała charakteru prostej poziomej (czyli *de facto* braku zależności od średniej).

To w zasadzie jest już koniec głównej części naszego projektu.

W dalszej części przeprowadziliśmy jeszcze kilka dodatkowych analiz, które uznaliśmy za interesujące, a jednocześnie nienadmiernie kłopotliwe.

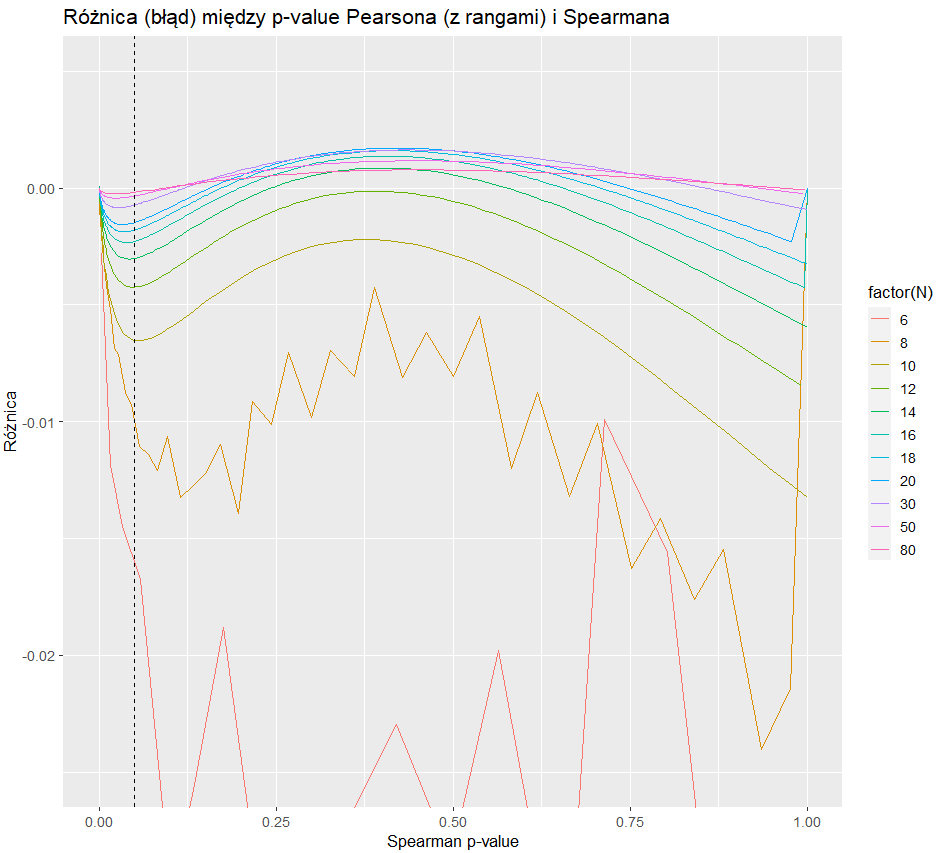
Poniżej prezentujemy porównanie uzyskanych p\_value w teście Pearsona (z rangami) wobec bezpośredniego testu dla ρ-Spearmana.



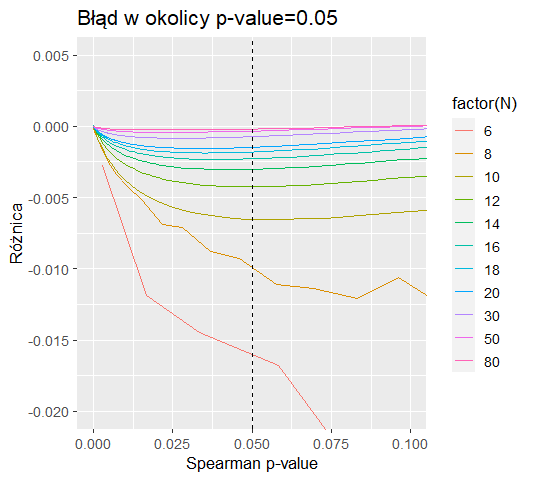
Na wykresie nie mamy sztucznie narysowanej prostej, to są wyłącznie punkty danych. Widzimy, że zależność ta jest wprost proporcjonalna, niemniej jednak zwraca uwagę, że procedura dla ρ-Spearmana czasami daje takie same wartości p\_value jak dla r-Pearsona, czasami niższe, ale nigdy wyższe. To jest być może obserwacja o jakiejś wartości merytorycznej.

Poniższym wykres obrazuje nieco bardziej subtelną kwestię niż poprzedni, o której już wspominaliśmy w części teoretycznej, czyli różnicę w numerycznie oszacowanych wartościach p\_value dla różnych procedur testu dla ρ-Spearmana.

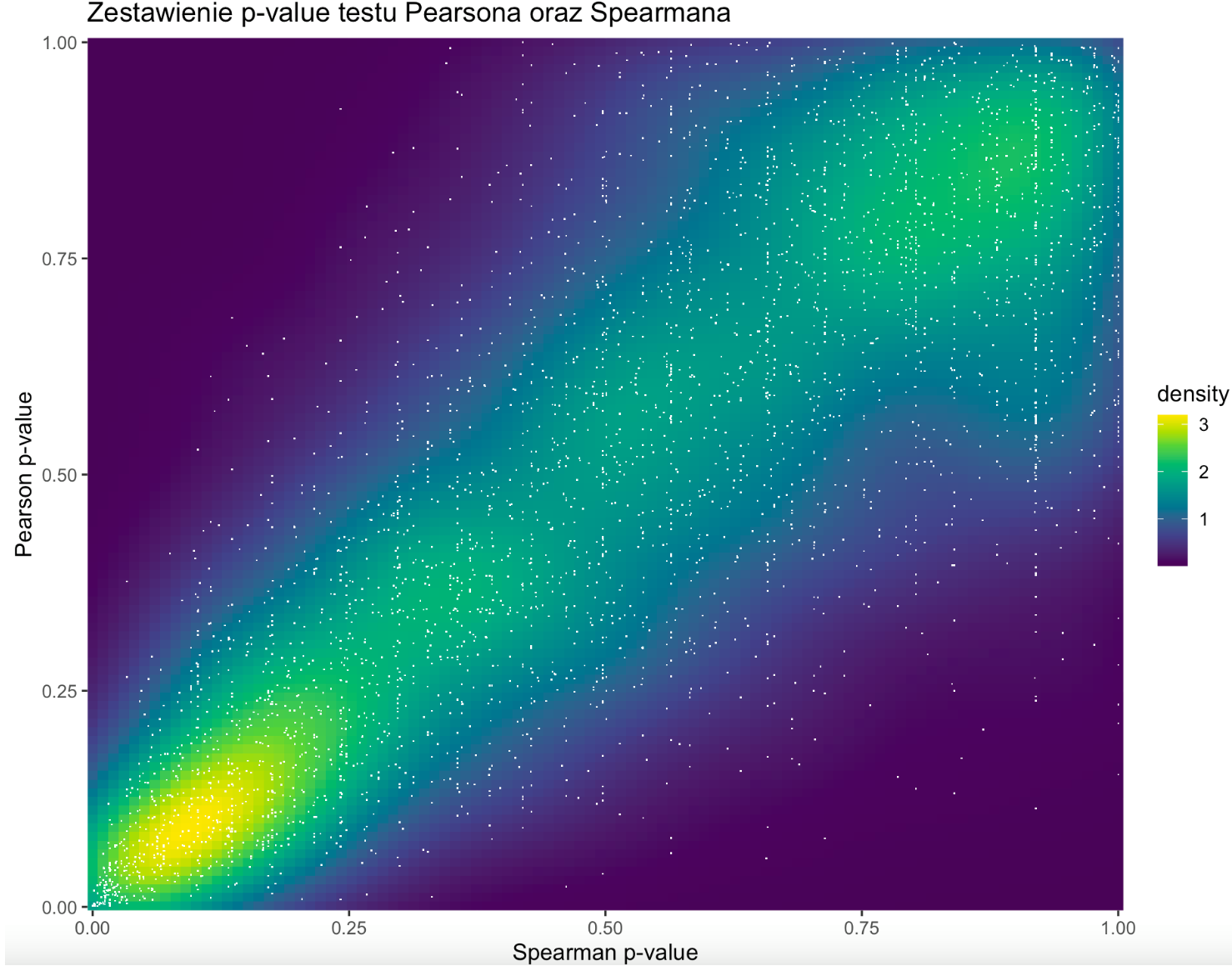
Współczynniki korelacji zawsze są identyczne. Szacowana wartość p jest dostatecznie dokładna (nie różni się dla obu procedur wyraźnie), gdy N > 10 , jednak naprawdę dobrą zgodność osiągamy dopiero dla N > 20. **Jeśli przeprowadzamy badania korelacyjne na próbach o mniejszych rozmiarach możemy mieć zatem dodatkowe problemy z wnioskowaniem (które wynikają wyłącznie z algorytmów używanych do szacowania p\_value).**



Tutaj prezentujemy „w zbliżeniu” fragment powyższego wykresu.



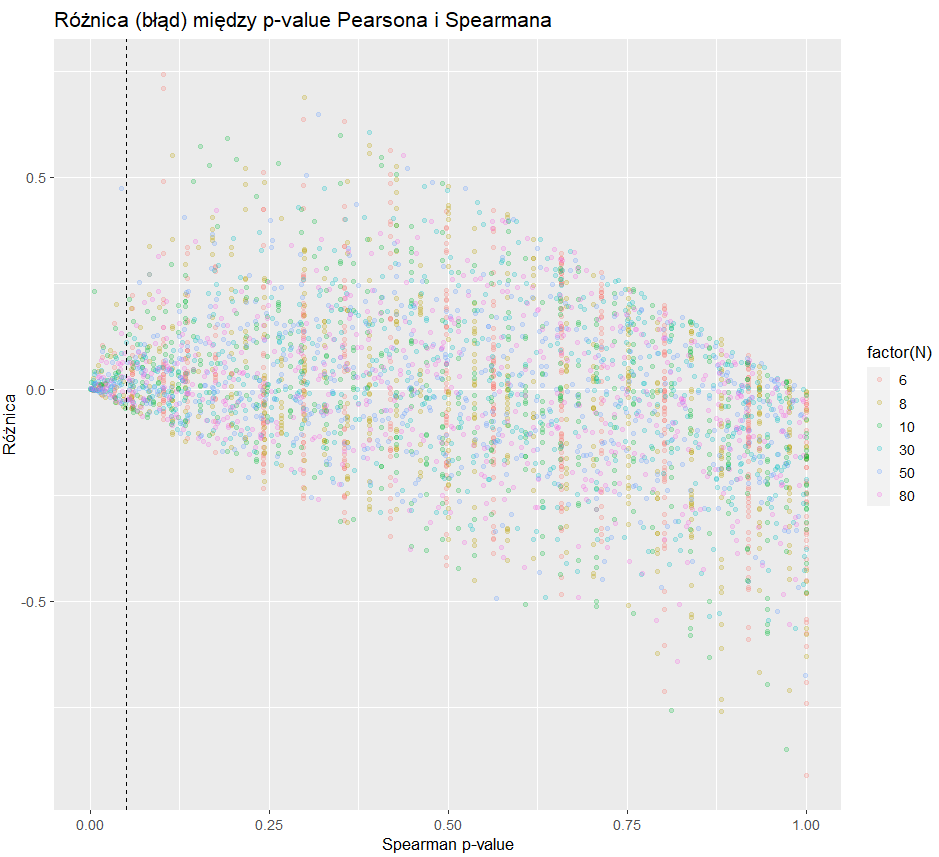
Poniżej z kolei prezentujemy p\_value otrzymywane dla danych surowych (nierangowy, klasyczny Pearson) porównany z procedurą dla ρ-Spearmana. Dalej dostrzegamy zależność monotoniczną, jednak jest ona dużo bardziej „rozmyta”.



Ten wynik jest zastanawiający. Dla nas aż takie „rozmycie” oraz lekkie odchylenie od liniowości (wklęsłość) jest nieoczekiwane.

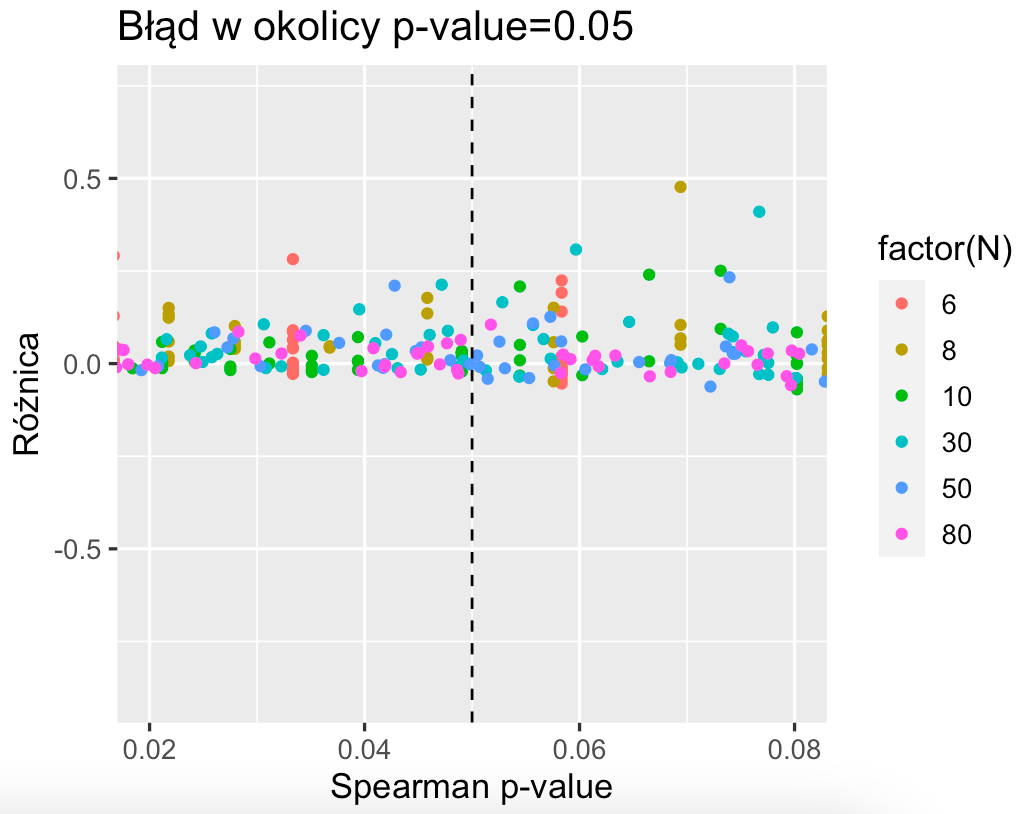
Gradient koloru opisany w legendzie wykresu jako *density* jest istotnym nośnikiem informacji, ponieważ wskazuje gęstość wartości p\_value w punkcie o danych współrzędnych.

Ze względu na wątpliwości związane z poprzednim uzyskanym wykresem, (różnica pomiędzy pearson\_p.value, a spearman\_p.value) przyglądamy się bliżej sąsiedztwu zerowych różnic pomiędzy uzyskanymi wartościami p\_value.



Oczywistych zależności nie dostrzegamy.

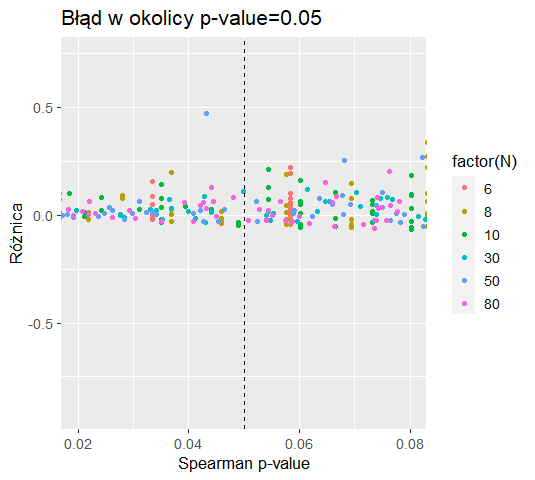
Wygenerujemy jeszcze jeden wykres w dwóch odsłonach (na jednym efekt będzie silniej widoczny, a na drugim słabiej).



Jak możemy zauważyć najmniejsze różnice występują przy 80 obserwacjach, to nam sugeruje, że te różnice będą dążyły do zera lub przynajmniej w okolice zera wraz ze zwiększającym się rozmiarem próby.

Być może zatem któryś z testów nie jest dobrze wyskalowany w całym przedziale możliwych liczebności.

Tutaj wspomniany efekt także można uznać za widoczny, choć jest nieco trudniejszy do dostrzeżenia.



1. **Podsumowanie i wnioski oraz obserwacje**
2. **Zarówno w przypadku współczynnika korelacji r-Pearsona jak i ρ-Spearmana prawdopodobieństwa popełnienia błędu pierwszego rodzaju nie zależą od parametrów rozkładu normalnego, z którego generowane są dane spełniające hipotezę główną.**
3. **Pewne interesujące efekty można ewentualnie w przeprowadzonych przez nas dodatkowych analizach obserwować dla niskich liczebności prób.**
4. **Przeprowadzone dodatkowe analizy wstępnie dały pewne lekko interesujące wyniki dotyczące zaimplementowanych procedur testów dla hipotez dotyczących współczynników korelacji, jednak to czy jest to jakiś rzeczywisty efekt (czyli coś co wymagałoby poprawienia procedur znajdujących się w pakietach) czy też artefakt obliczeniowy albo niedoskonałości w kodzie pozostaje nierozstrzygnięte.**

1. https://lindeloev.github.io/tests-as-linear/simulations/simulate\_spearman.html [↑](#footnote-ref-1)