Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
AGH University of Science
and Technology

AGH

Testowanie odsetka odrzuceń prawdziwej hipotezy głównej w teście Kołmogorowa i teście χ^2

Autorzy:

Majka Miezianko

Marek Czerwonka

Projekt zaliczeniowy z przedmiotu:

Metody Nieparametryczne w Statystyce

Kraków 2021

Spis treści

Wstęp

W ramach projektu zrealizowane zostanie następujące zadanie:

Za pomocą odpowiednich symulacji zbadać odsetek odrzuceń prawdziwej hipotezy głównej (tzn. błąd I rodzaju) w teście Kołmogorowa i teście χ^2 w przypadku weryfikacji zgodności z rozkładem t-Studenta o różnych stopniach swobody, tzn. generować dane pochodzące z rozkładu t-Studenta o k stopniach swobody i weryfikować hipotezę, że dane mają rozkład t-Studenta o k st. swobody. Należy porównać wyniki klasycznego sposobu testowania i testowania z wykorzystaniem PIT (probability integral transform). Uzyskane wyniki należy przedstawić na odpowiednich wykresach ilustrujących rozmiary testów z uwzględnieniem:

- liczby danych,

- liczby stopni swobody generowanego rozkładu.

1. Test Kołmogorowa zgodności z rozkładem

1.1. Krótka charakterystyka testu

Test Kołmogorowa jest testem zgodności rozkładu, który może być stosowany w przypadku rozkładów ciągłych (a takim rozkładem jest rozkład t-Studenta). Opiera się na porównaniu dystrybuanty empirycznej oraz teoretycznej (odpowiadającej rozkładowi zgodność, z którym podlega testowaniu).

Układ hipotez jest klasyczny: hipoteza główna obejmuje przypadek równości dwóch powyżej wymienionych dystrybuant, zaś hipoteza alternatywna ich różności. Statystyka testu oparta jest o moduł różnicy dystrubuant empirycznej oraz teoretycznej.

Dla zbiorów danych o niskiej liczności można skorzystać ze stablicowanego rozkładu statystyki D, gdzie:

$$D_n = \max_{1 \le i \le n} \left| K(x_i) - F(x_i) \right|$$

W przypadku populacji o dużej liczebności, korzystany raczej z rozkładu granicznego statystyki λ, opisanej wzorem:

$$\lambda = D_n \sqrt{n}$$

Obszar krytyczny (odrzucenia hipotezy głównej) w tym teście jest prawostronny.

1.2. Parametry oraz wyniki przeprowadzonej 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: 5 000

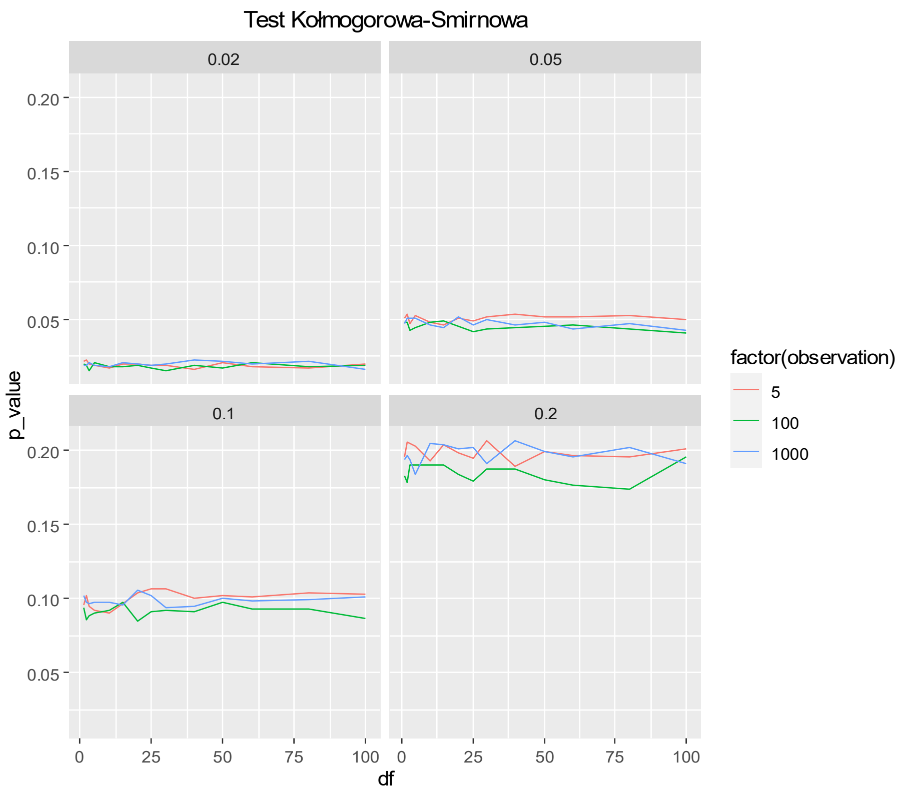
Liczba obserwacji generowanych w pojedynczej symulacji: 5, 100, 1000

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

Liczba stopni swobody: 1, 2, 3, 5, 10, 15, 20, 25, 30, 40, 50, 60, 80, 100.

Rozkład t-Studenta dość szybko zbiega do rozkładu normalnego standaryzowanego. Dlatego wybraliśmy do przykładowej symulacji niskie wartości liczby stopni swobody. Jeżeli można by się spodziewać jakichś nieoczekiwanych efektów to zapewne właśnie od strony niskiej liczby stopni swobody, szczególnie interesujący mógłby być przypadek 1 stopnia swobody, dla której to wartości nawet wartość oczekiwana rozkładu t nie jest dobrze określona.

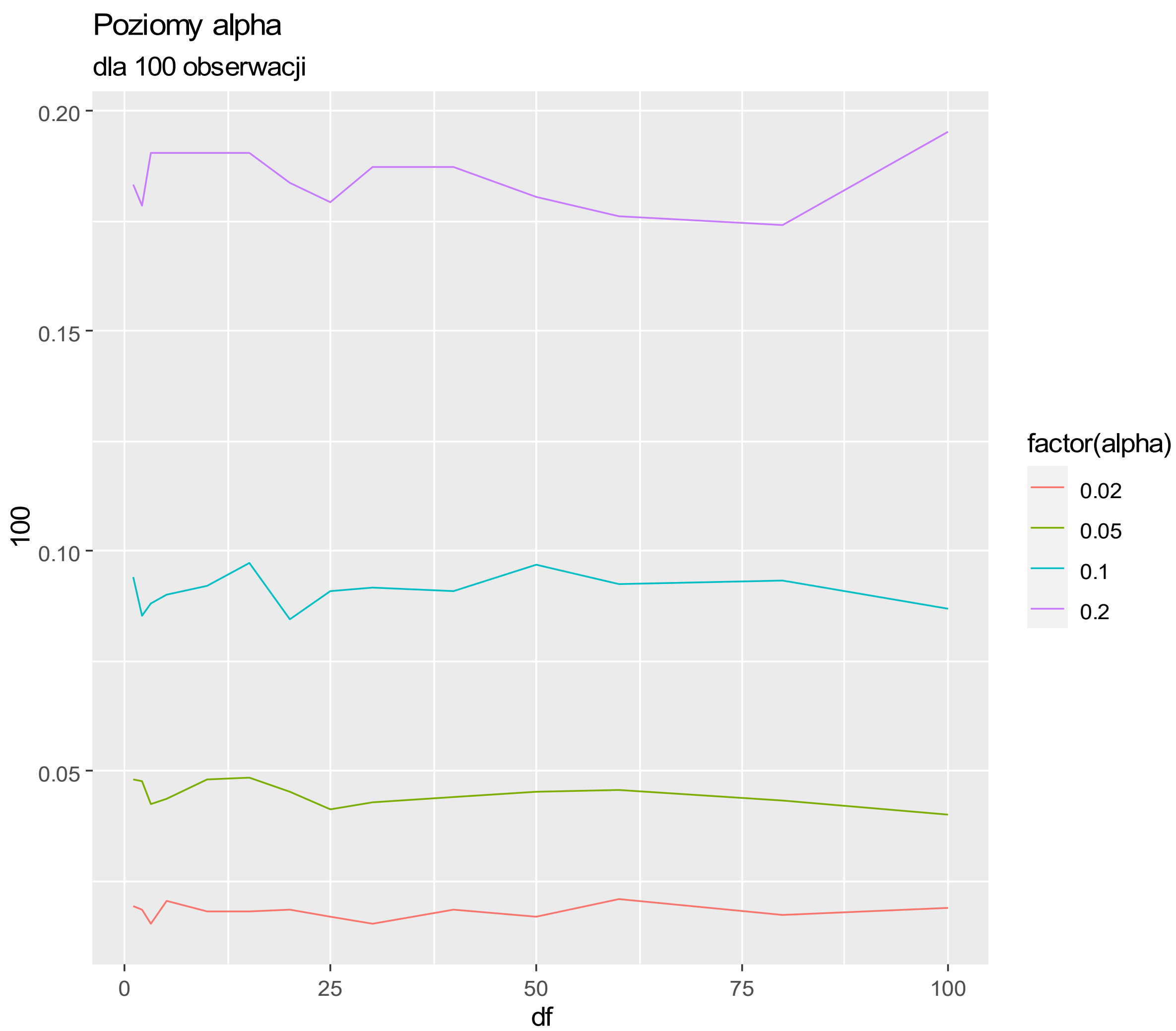
Podobnie w przypadku liczby danych, o ile można oczekiwać jakichś interesujących i zaskakujących wyników to raczej przy niskich liczebnościach, stąd ta dość patologiczna liczba generowanych danych równa 5.



Z ciekawości przeprowadziliśmy symulacje dla kilku wartości alfa. Na powyższym wykresie widzimy brak zależności od liczby stopni swobody. Jest to oczekiwany wynik i świadczy o tym, że procedura testu Kołomogorowa jest dobrze zaimplementowana w odpowiednim pakiecie R, a także że wartości krytyczne statystyki Kołmogorowa są wyznaczone poprawnie. Interpretując te wykresy należy pamiętać, że w istocie są to tylko punkty dla wybranej liczby stopni swobody (która została wysymulowana) połączone odcinkami.

Należałoby na podstawie twierdzeń granicznych oczekiwać, że linia niebieska będzie najbardziej stabilna z trzech narysowanych i to w zasadzie jest widoczne. Co na tych wykresach może być nieco zaskakujące to fakt, że zielona linia (odpowiadająca liczebności próbki równej 100, regularnie znajduje się „najniżej" oraz wydaje się faktycznie systematycznie zaniżać średnią uzyskiwaną wartość p_value wobec tego czego należałoby oczekiwać.

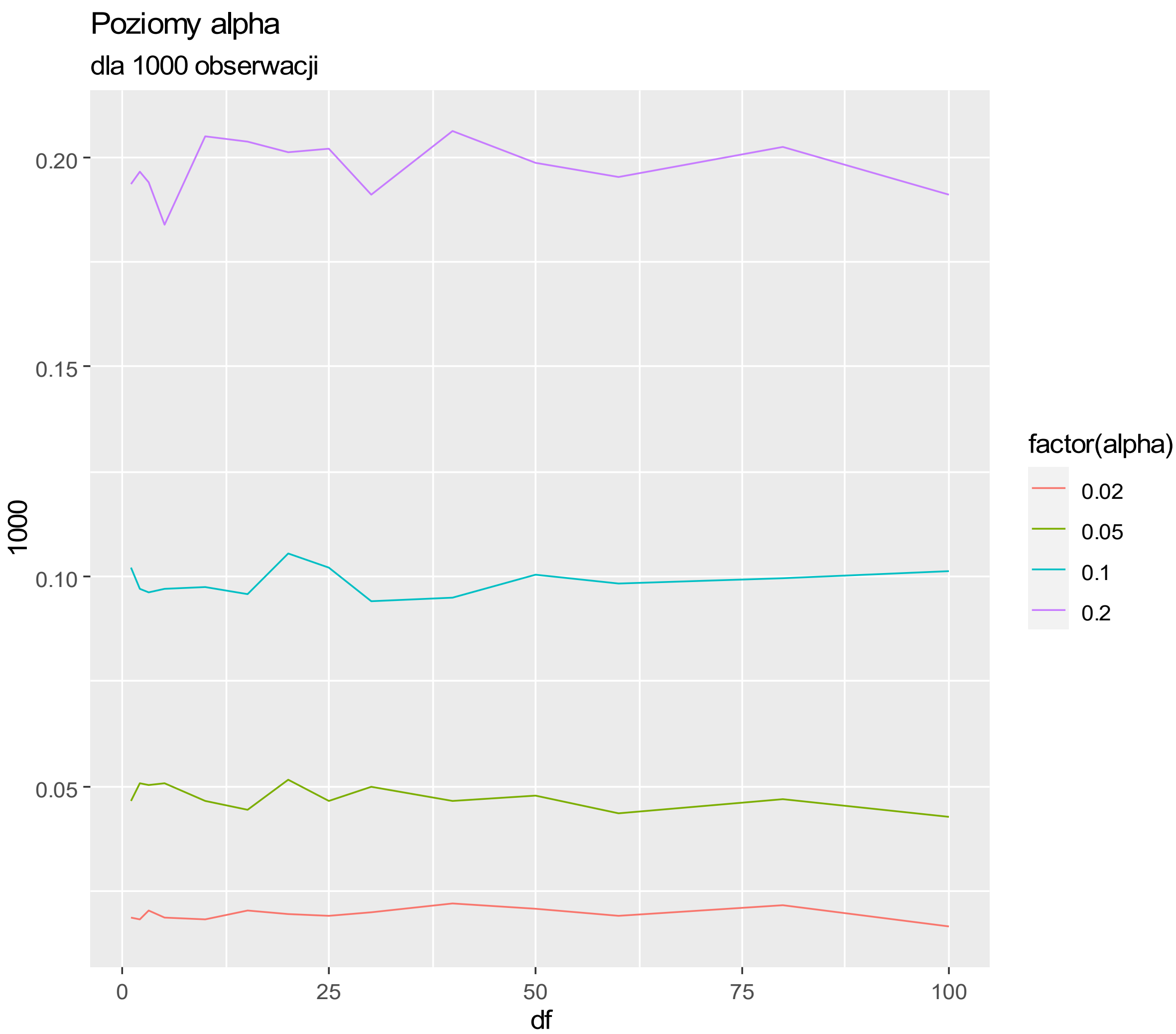
Przyjrzyjmy się zatem bliżej temu zagadnieniu na dwóch kolejnych wykresach.



Jeśli przyjrzymy się bliżej widzimy, że nasze symulacje – przynajmniej dla wielkości próby równej 100 – generalnie dla każdego przyjętego poziomu istotności alfa generują niższe faktyczne poziomy odrzuceń hipotezy zerowej. Możliwa interpretacja tego faktu to – zakładając prawidłowe przeprowadzenie symulacji oraz dobre działanie generatora liczb pseudolosowych – korzystanie przez procedurę testową z jakiegoś dodatkowego zabezpieczenia przed nadmiernie częstym popełnianiem błędu pierwszego rodzaju, które być może daje taki efekt, że test jest bardziej restrykcyjny niż teoretycznie powinien być.

Zagadnienie to mogłoby być przedmiotem analizy poprzez przeprowadzenie symulacji przy użyciu innego generatora liczb pseudolosowych.

Sprawdźmy jeszcze czy to zjawisko będzie powtarzalne w przypadku bardziej licznej próby.



Zatem tym razem efektu takiego nie obserwujemy lub jest on w znacznie niższym natężeniu. Pozwala to przypuszczać, że poprzedni wynik był jednak artefaktem obliczeniowym lub immanentną wadą generatora liczb pseudolosowych dla niskiej liczby generowanych przypadków. Ze zwiększaniem się liczebności próby powinna zatem rosnąc precyzja z jaką średnie generowane p_value trafia w jego wartość oczekiwaną.

2. Test zgodności z rozkładem χ^2

2.1. Krótka charakterystyka testu

Test zgodności χ^2 jest testem zgodności rozkładu, który może być stosowany zarówno w przypadku rozkładów ciągłych jak i dyskretnych. Opiera się na porównaniu dystrybuanty empirycznej oraz teoretycznej (odpowiadającej rozkładowi, zgodność z którym podlega testowaniu).

Układ hipotez jest klasyczny: hipoteza główna obejmuje przypadek równości dwóch powyżej wymienionych dystrybuant, zaś hipoteza alternatywna ich różności. Statystyka testu wyznaczana jest wg wzoru:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i}$$

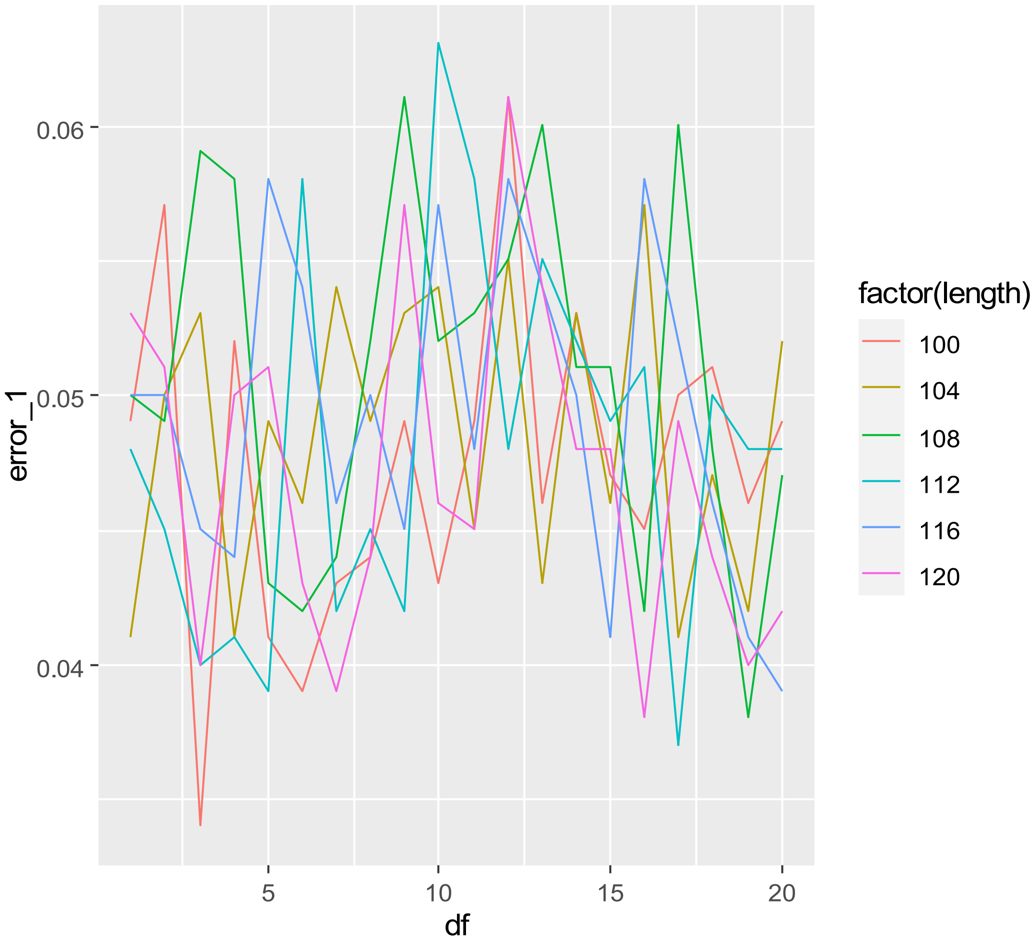
Statystyka ta dla dostatecznie dużych n ma asymptotyczny rozkład χ^2 o k-m-1 stopniach swobody, gdzie m jest liczbą parametrów rozkładu, zgodność z którym testujemy. Rozkład t jest rozkładem jednoparametrycznym.

Obszar krytyczny (odrzucenia hipotezy głównej) w tym teście jest prawostronny. Dodatkowym ograniczeniem stosowalności tego testu jest to, że liczebność oczekiwana dla każdej wydzielonej klasy powinna być równa co najmniej 5.

Sposób wydzielenia przedziałów może być różny. W naszej symulacji, aby każdemu z przedziałów przypisać tę samą „wagę" udziału w statystyce testowej dokonaliśmy podziału wyznaczając odpowiednie kwantyle rozkładu wygenerowanych danych i tworząc przedziały tak, aby w każdym z nich znajdowało się w przybliżeniu tyle samo danych.

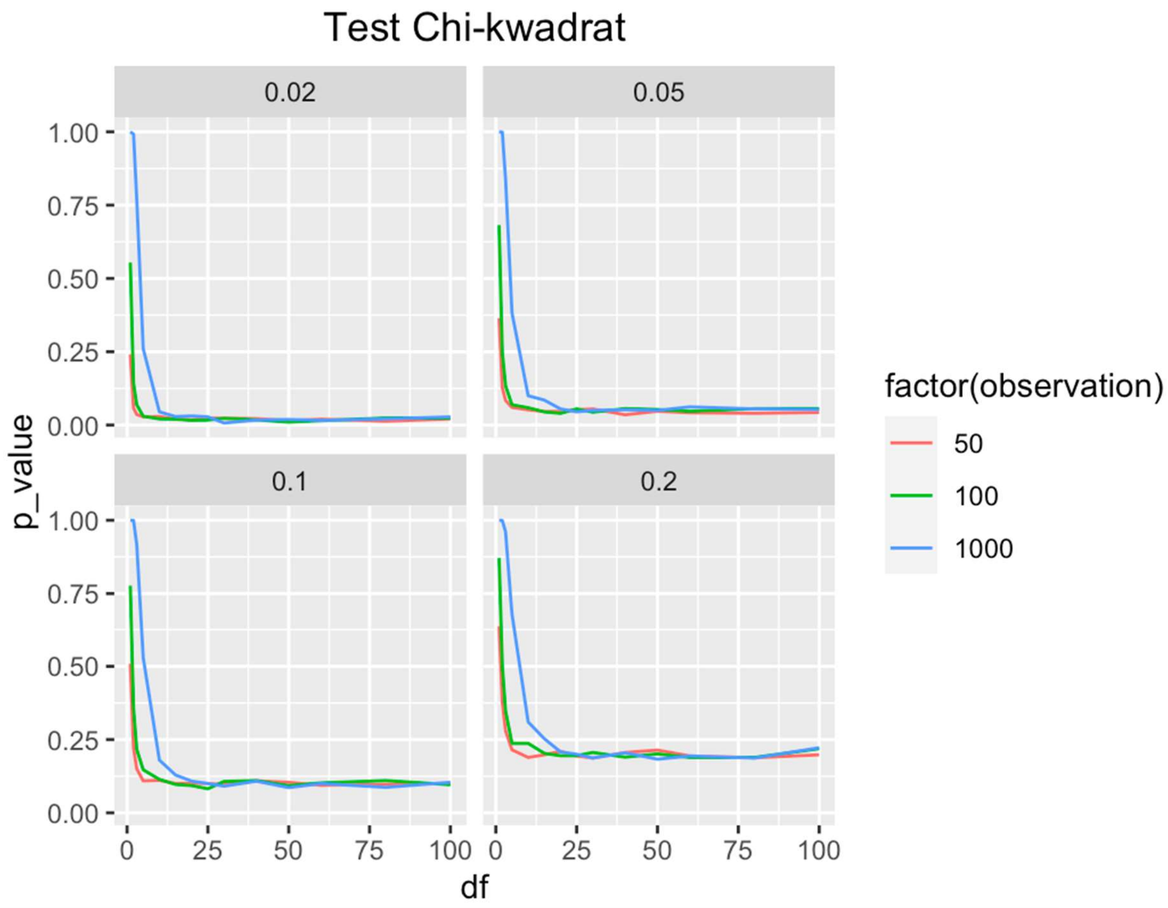
Jako że test ten opiera się na porównywaniu dystrybuant w kilku punktach, należałoby oczekiwać, że będzie testem ogólnie lepszym niż test Kołmogorowa. Nie ma za bardzo możliwości, aby przy prawidłowej implementacji dawał wyniki błędne, co więcej intuicyjnie wydaje się, iż powinien charakteryzować się lepszą precyzją.

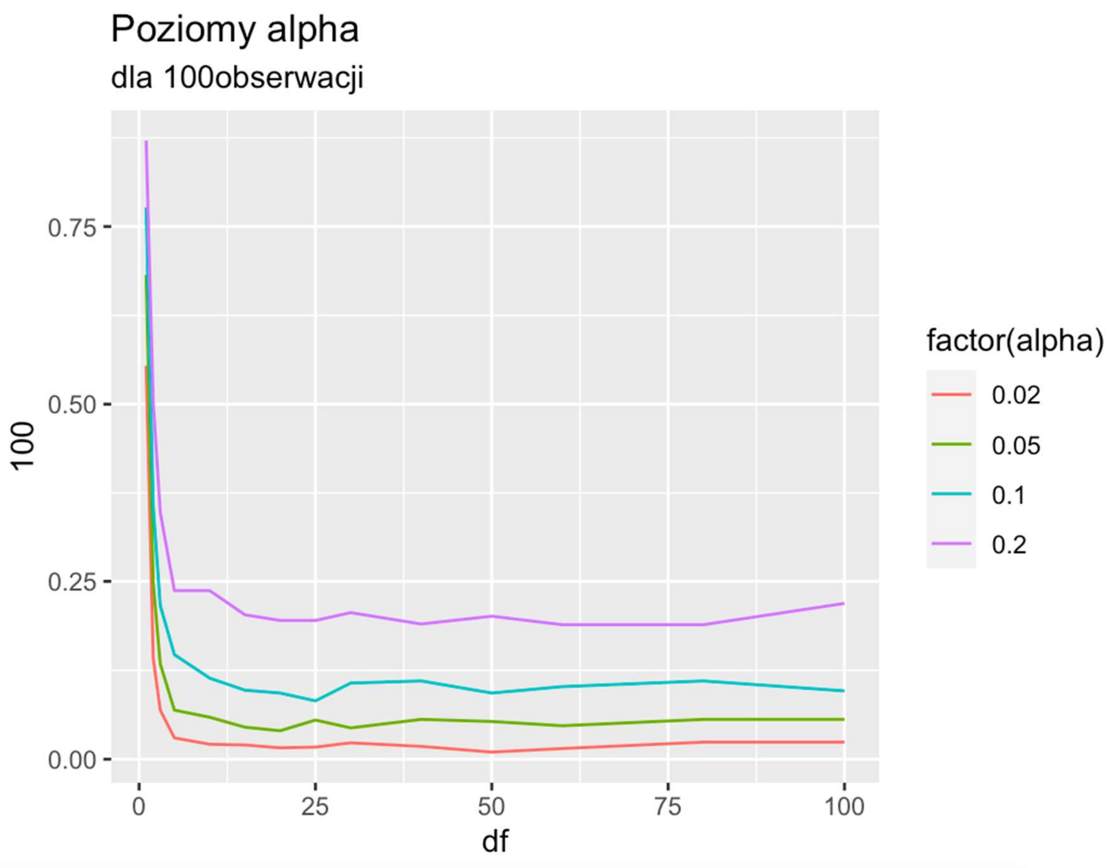
Tym razem ocenimy zatem na ile stabilne są eksperymentalnie uzyskiwane odsetki odrzuceń prawdziwej hipotezy zerowej przy niewielkich zmianach liczby stopni swobody rozkładu oraz liczebności próby.

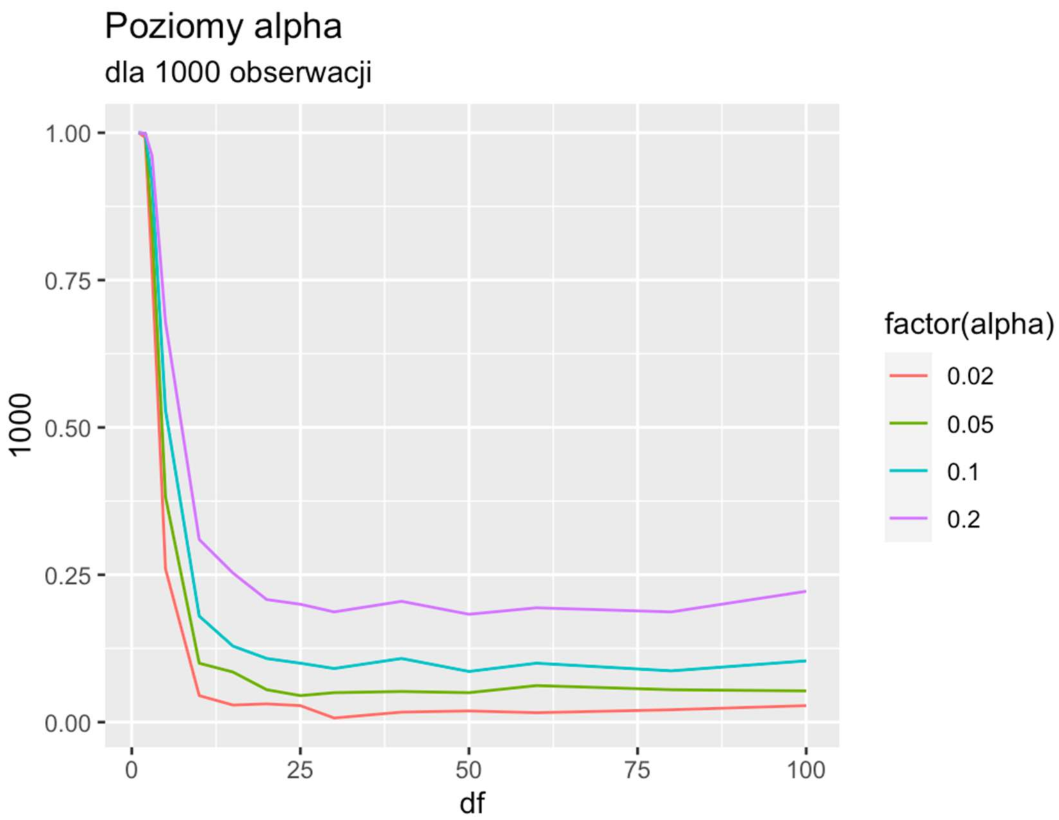


Uzyskany wykres potwierdza intuicyjnie oczekiwaną niezależność od liczby stopni swobody (gdyby miała występować to po stronie niskiej liczby stopni swobody), a także niezależność średniego p_value od wielkości próby. Natomiast wraz ze wzrostem próby powinna maleć zmienność uzyskiwanych p_value, co zostało pokazane już w przypadku testu Kołmogorowa.

Wygenerowaliśmy też wykresy analogicznej do naszej analizy testu Kołmogorowa z drugiego wariantu tego kawałka kodu (dotyczącego testu chi-2).


Test Chi-kwadrat
0.02
0.05
0.1
0.2
p_value
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
df
factor(observation)
50
100
1000


Poziomy alpha
dla 100obserwacji
100
0.75
0.50
0.25
0.00
0
25
50
75
100
df
factor(alpha)
0.02
0.05
0.1
0.2



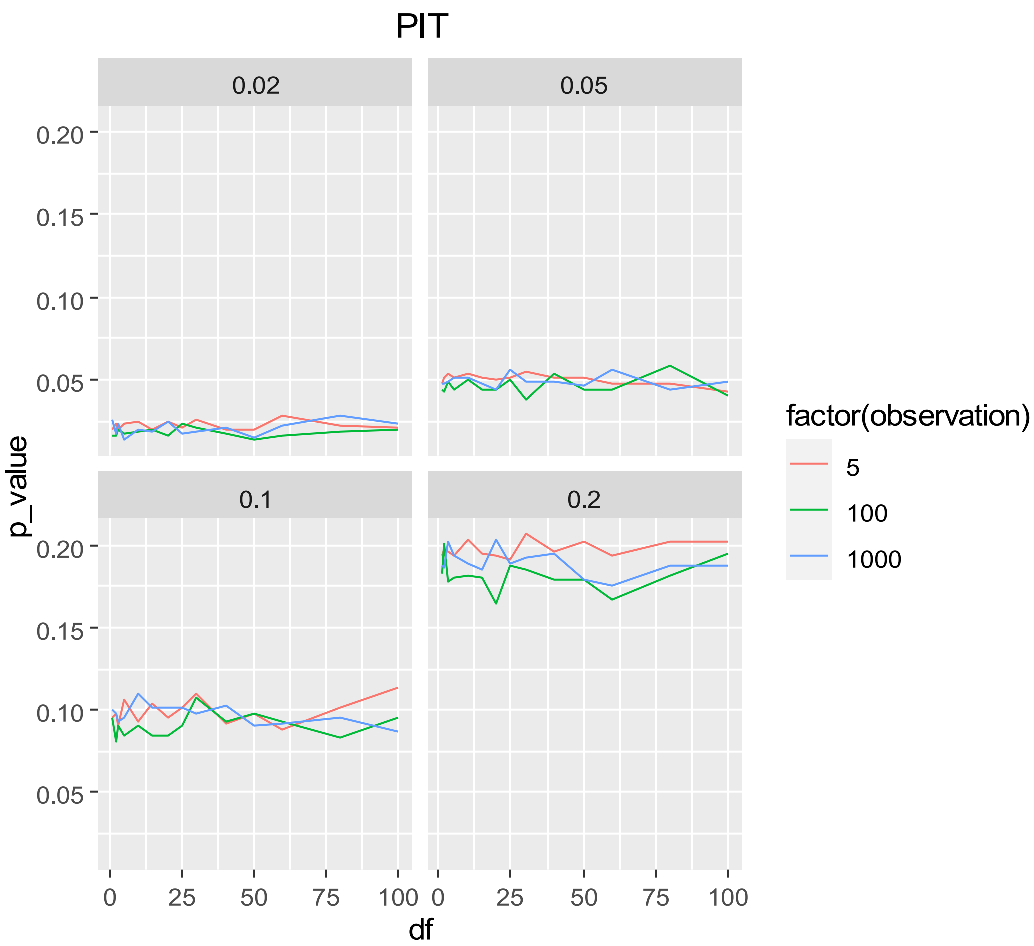
Z pewnością warte są one szerszego opisu bo wskazują być może na jakiś błąd w kodzie (taki efekt dla niskiej liczby stopni swobody wg nas nie powinien występować), ale łatwo przewidzieć, o której godzinie rozpoczęliśmy wykonywanie tego projektu w niedzielę i reszta niech będzie milczeniem.

3. PIT (universality of the uniform, probability internal transform)

Ze względu na teorię stojącą za tym sposobem (opartym o odwrotną dystrybuantę), który jest odpowiedni wyłącznie dla zmiennych o rozkładzie ciągłym i sprowadza się do takiego przekształcenia znanego rozkładu aby w efekcie był porównywany z rozkładem jednostajnym na ustalonym przedziale, należy oczekiwać dokładnie takich samych wyników jak w przypadku braku takiego przekształcenia.

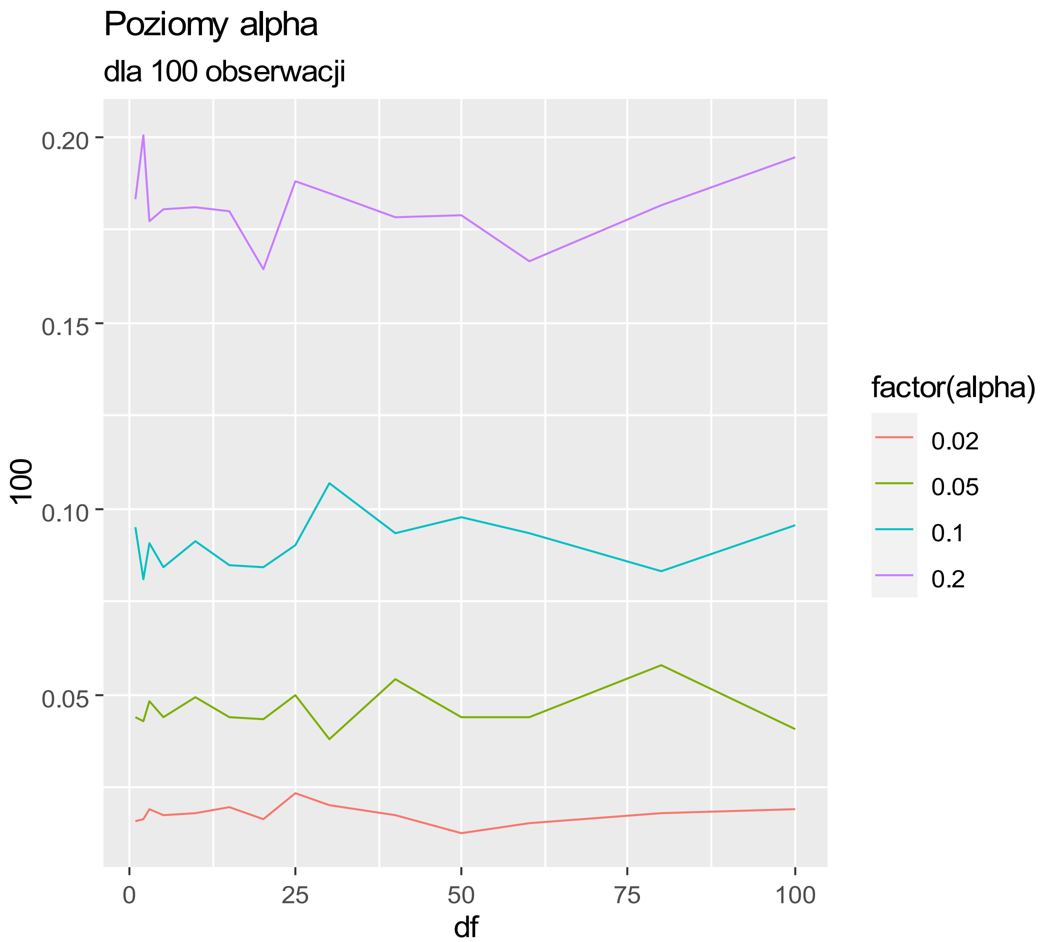
Zakładając, że przekształcenie to wykonywane jest poprawnie, mamy w istocie do czynienia z testowaniem przetransformowanej hipotezy, po transformacji danych. Niemniej jednak w tym przypadku wydaje nam się interesujące wykonanie analogicznych symulacji jak w przypadku testu Kołmogorowa.

Wygenerujmy zatem analogiczne wykresy jak w przypadku testu Kołmogorowa.

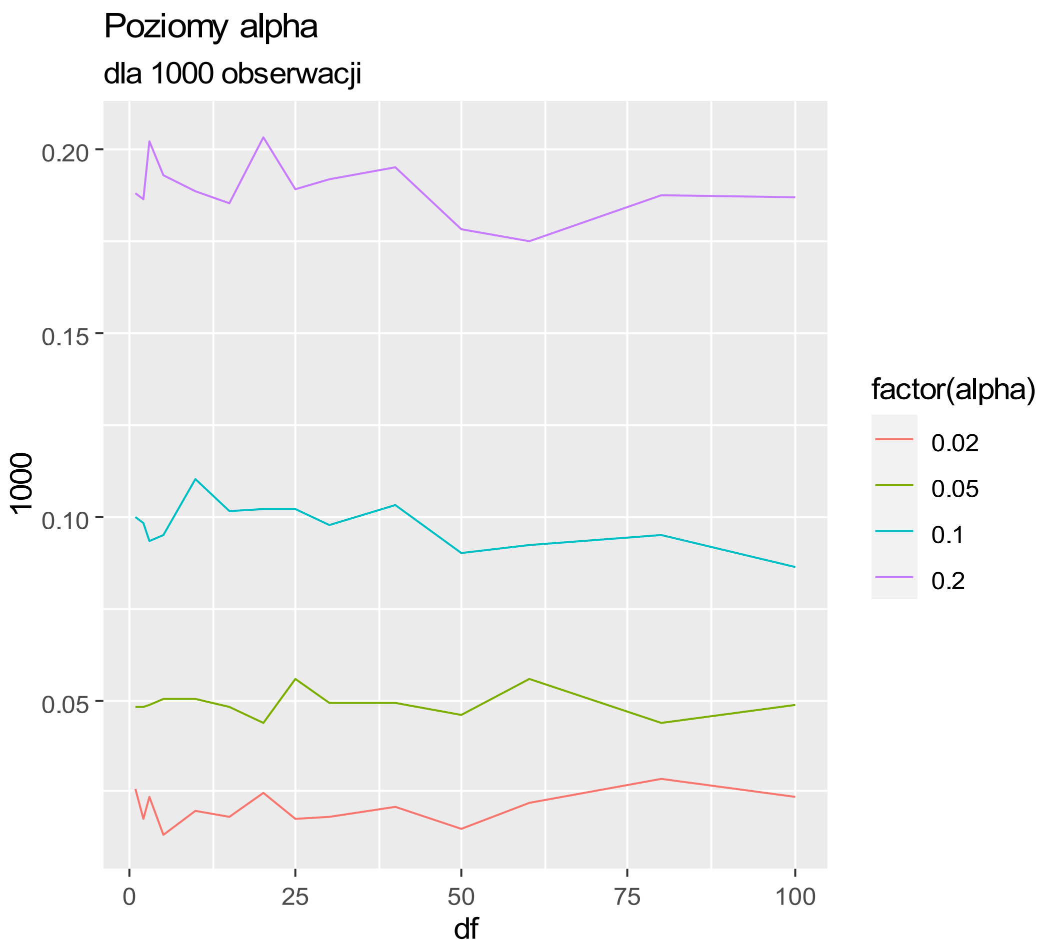


Jak widzimy uzyskujemy wyniki zgodne z uzyskanymi poprzednio. Niebieska linia zgodnie z oczekiwaniami „trzyma się" najbliżej wartości oczekiwanej, poza przypadkiem dla alfa = 0,20.

Ponownie zwraca uwagę to, że linia zielona przynajmniej wizualnie podejrzanie częściej niż w 1/3 przypadków znajduje się najniżej. Ponownie nie mamy na to racjonalnego wytłumaczenia (wprawdzie tutaj zmniejszyliśmy liczbę generowanych zestawów danych do 2000 z 5000, niemniej jednak co do zasady nie powinien taki efekt występować).



Jak widać ponownie obserwujemy efekt pewnego systematycznego zaniżania uzyskiwanych p_value wobec tych jakich należałoby oczekiwać. Wygląda więc na to, że ten sposób testowania także jest bardziej restrykcyjny niż wynikałoby to z nominalnych wartości krytycznych.



Przy tysiącu generowanych obserwacji uzyskiwane przeciętne p_value jest bardziej zbliżone do oczekiwanych, choć wydaje się że nadal jest nieco poniżej wartości oczekiwanej.

3. Podsumowanie i wnioski oraz obserwacje

A. Co do zasady zgodnie z oczekiwaniami rozmiar rozważonych testów nie zależy od liczby stopni swobody rozkładu t-Studenta, z którego były generowane dane.

B. Co do zasady zgodnie z oczekiwaniami rozmiar rozważonych testów nie zależy od liczby danych generowanych w pojedynczej symulacji.

C. Zgodnie z oczekiwaniami na podstawie twierdzeń granicznych wraz ze wzrostem liczby symulacji maleje zróżnicowanie uzyskiwanych wyników w rozumieniu odchylenia standardowego otrzymywanej eksperymentalnej wartości p_value.

D. Przeciwnie do oczekiwań w przypadku testu Kołomogorowa przypadek z wielkością próby równą 100 najsilniej niedoszacowuje p_value wobec wartości oczekiwanej (test wydaje się być bardziej restrykcyjny).

E. Ogólnie wyniki symulacji sugerują, że uzyskiwane w symulacjach wyniki są nieco niższe niż można by oczekiwać (test jest bardziej restrykcyjny niż należałoby oczekiwać). Na chwile obecną nie mamy logicznego uzasadnienia do tej obserwacji, co nie oznacza jednak, że nie będziemy jej mieli podczas obrony projektu.