Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
specjalność: Analiza Danych

Kamil Zaborniak

Optymalny próg wykrycia w modelach typu: „igła w stogu siana"

Praca licencjacka
napisana pod opieką
dr. Liudmyly Zaitsevy

Wrocław, 2023

Spis treści

1 Wstęp

Głównym celem niniejszej pracy jest zanlezienie przez nas optymalnych progów wykrycia dla modeli typu „igła w stogu siana". Uzyskane wyniki przeanalizujemy i sprawdzimy jak wpływają na optymalizację procesów testowania wcześniej wspomnianych modeli statystycznych.

Rozważania rozpoczniemy od poznania i przypomnienia podstaw teorii testowania hipotez statystycznych. Rozważymy czym są hipotezy statystyczne oraz w jaki sposób prawidłowo je konstruować. W kolejnym kroku określimy typy rozpatrywanych sądów, ze względu na to do czego się odnoszą, jaki kierunek obierają oraz jak złożone są. Przemyślenia te pozwolą nam poprawnie łączyć je w pary, czyli tworzyć modele statystyczne. Następnym krokiem będzie poznanie przez nas idei ich testowania. Uzyskamy wiedzę na temat sposobów konstrukcji oraz klasyfikacji testów statystycznych ze względu na postać modelu statystyczego. Zapoznamy się z rodzajami błędów, które możemy uzyskać w wyniku weryfikacji hipotez statystycznych, oraz ich prawdopodobieństwami. Po przeanalizowaniu podstaw, poznamy narzędzia testów oraz postaramy się przeprowadzić kilka prostych weryfikacji przykładowych modeli statystycznych, a następnie wyznaczymy ich moce. Po tym, poznamy metodę konstrukcji testu najmocniejszego, opartą o lemat Neymana-Pearsona, oraz rozpatrzymy przykład jej wykorzystania.

W następnym kroku postaramy się odpowiedzieć na pytanie: *„Jaki rozkład przyjmuje p-wartość?"*. Rozpatrzymy przypadki w zależności od rodzaju rozkładu prawdopodobieństwa statystyki testowej. Problem rozwiążemy przy pomocy własności dystrybuanty poszczególnych rozkładów. W rezultacie dowiemy się, że jeśli statystyka testowa pochodzi z rozkładu ciągłego, to rozkład jej przekształcenia jest rozkładem jednostajnym na przedziale $[0, 1]$. Ustalimy również rozkład p-wartości w sytuacji, kiedy statystyka testowa będzie pochodziła z rozkładu Poissona. Wtedy, dowiemy się, że p-wartość przyjmuje asymptotycznie rozkład $\mathcal{U}[0, 1]$.

Następnym zagadnieniem, które poznamy, będzie globalna hipoteza zerowa. Zapoznamy się ze sposobem jej konstrukcji. Dowiemy się jak korekta Bonferroniego pozwala nam sprawdzić prawdziwość globalnej hipotezy zerowej. W konsekwencji przeprowadzonych rozważań na temat rozkładu p-wartości, udowodnimy, że korekta Bonferroniego kontroluje prawdopodobieństwo odrzucenia globalnej hipotezy zerowej na ustalonym poziomie istotności $\alpha \in (0, 1)$.

Konsekwencją przeprowadzonych przez nas rozważań będzie definicja modelu typu „igła w stogu siana". Pierwszym takim modelem, który poddamy analizie będzie model oparty o dane pochodzące z rozkładu normalnego o parametrze średniej — μ. Model będzie się składał z globalnej hipotezy zerowej: $H_0 : \forall i \in \{1, \ldots, n\} : \mu_i = 0$, gdzie n określa liczbę prób, na podstawie, których dokonamy weryfikacji, oraz alternatywy: $H_1 : \exists! i \sim \mathcal{U}\{1, \ldots, n\} : \mu_i > 0$. Wyzanczymy asymptotyczną wartość kwantyla rozkładu normalnego, czyli asymptotycznego progu wykrycia. W oparciu o to wyznaczymy asymptotyczną moc korekty Bonferroniego. Następnie przekształcimy model i wyznaczymy asymptotyczną moc testu opartego o lemat Neymana-Pearsona. Wynikiem przeprowadzonych testów będzie fakt, że będziemy mogli ustalić czy pośród obserwacji z rozkładu $\mathcal{N}(0, 1)$, znajduje się obserwacja z rozkładu $\mathcal{N}(\mu, 1)$, wyłącznie gdy $\mu(1 + \varepsilon)\sqrt{2\ln(n)}$, gdzie $\varepsilon \in (0, 1)$. Sprawdzimy również przypadek, kiedy model typu „igła w stogu siana" będzie oparty o dane pochodzące z rozkładu Poissona o parametrze λ. Analizowany model będzie się składal z hipotez: $H_0 : \forall i \in \{1, \ldots, n\} : \lambda_i = \lambda'$, $H_1 : \exists! i \sim \mathcal{U}\{1, \ldots, n\} : \lambda_i > \lambda'$. Wyaproksymujemy wartość asymptotycznego progu wykrycia, oraz postaramy się sprawdzić jak korekta Bonferroniego pozwala nam wykryć obserwację odstającą oraz jaką moc osiąga, przy liczbie prób: $n \to \infty$. W rezultacie dowiemy się, że korekta Bonferroniego nie wykryje obserwacji odstającej, pochodzącej z rozkładu Poissona, jeśli jej parametr rozkładu będzie równy $(1 - \varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$.

2 Podstawy do teorii testowania hipotez

Jednym z głównych działów statstyki matematycznej jest wnioskowanie statystyczne. Obejmuje ono zasady i metody uogólniania wyników otrzymanych poprzez badanie próby losowej na całą populacje generalną, z której pochodzi dana próba. Głównymi domenami wnioskowania statystycznego są:

Estymacja — szacowanie postaci lub parametrów rozkładu zmiennej losowej w populacji generalnej na podstawie próby badawczej,

Weryfikacja hipotez statystycznych — sprawdzenie przypuszczeń dotyczących rozkładu populacji generalnej poprzez badanie jej części.

W przypadku estymacji, wnioski na temat populacji generalnej stwierdzamy w oparciu o wyniki próby. W odróżnieniu, przeprowadzając testowanie hipotez statystycznych, najpierw wysuwamy określone sądy na temat zbiorowości generalnej, a w następnych krokach sprawdzamy ich prawdziwość. W niniejszym rozdziale, opierając się o wiedzę zawartą w książce pt. „Introduction to Mathematical Statistics (6th Edition)"[1], przeprowadzimy głównie rozważania na temat teorii weryfikacji hipotez.

2.1 Wprowadzenie

Hipotezami statystycznymi nazywamy założenia dotyczące rozkładu prawdopodobieństwa badanej cechy populacji, która jest zmienną losową. Na nich oparty jest proces ich weryfikacji. Hipotezy statystyczne dzielimy na parametryczne i nieparametryczne. **Hipotezy nieparametryczne** dotyczą postaci funkcji gęstości rozkładu parametru populacji, którego badamy. Ich przykładem jest hipoteza: *„Rozkład zmiennych losowych pochodzących z populacji jest rozkładem wykładniczym."*. **Hipotezy parametryczne** dotyczą wartości parametrów rozkładu cechy populacji generalnej. Przykładem takiej hipotezy jest założenie: *„Średnie dwóch populacji o rozkładzie normalnym są równe."*. Przypuszczenia te dzielimy na hipotezy niekierunkowe oraz hipotezy kierunkowe. **Hipoteza niekierunkowa** zakłada tylko różnice wielkości badanego parametru względem konkretnej wartości. Przkładem jest hipoteza: „Odsetek osób palących w populacji generalnej jest różny od 0.2". Natomiast **hipoteza kierunkowa** określa kierunek spodziewanych wyników. Taka hipoteza może mieć formę lewostronną lub prawostronną. **Hipoteza lewostronna** zakłada, że badany parametr jest mniejszy od konkretnej wartości. Jej przykałdem jest hipoteza: „Odsetek osób palących w populacji generalnej jest mniejszy niż 0.2". **Hipoteza prawostronna** stwierdza natomiast, że wartość badanego parametru należy do zbioru wartości większych od konkretnej wartości. Przykładem takiej hiptezy jest założenie: „Odsetek osób palących jest większy od 0.2".Rodzinę rozkładów lub zbiór wszystkich wartości badanego parametru, określanych przez wszystkie możliwe hipotezy statystyczne, dotyczące tego parametru, będziemy oznaczać symbolem Θ.

Każda wiedza dotycząca populacji generalnej pozwala nam ograniczyć zbiór możliwych hipotez do **zbioru hipotez dopuszczalnych**. Przykładowo, jeśli wiemy, że badana cecha populacji — zmienna losowa, pochodzi z rozkładu Poissona, to naszym zbiorem hipotez dopuszczalnych będzie zbiór wszystkich rozkładów Poissona różniących się wyłącznie parametrem λ. Bezcelowe jest w takim przypadku testowanie hipotezy, czy dana cecha ma rozkład inny niż rozkład Poissona.

Jeśli hipoteza ze zbioru hipotez dopuszczalnych jednoznacznie określa rozkład populacji, taką hipotezę nazywamy **hipotezą prostą**. Przykładem jest hipoteza: *„ Populacja pochodzi z rozkładu Poissona o prametrze λ równym 2."*. Jakakolwiek hipoteza, która nie jest hipotezą prostą jest **hipotezą złożoną**, na przykład: *„Populacja pochodzi z rozkładu Poissona o parametrze λ większym od 2."*.

Założenie dotyczące rozkładu zmiennej losowej X w populacji generalnej, którego prawdziwość chcemy sprawdzić nazywamy **hipotezą zerową** i oznaczamy symbolem H_0. Zazwyczaj hipoteza zerowa będzie określała założenie, które wynika z podstawowej wiedzy jaką posiadamy na temat rozkładu badanej populacji. W przypadku badania związku między dwiema populacjami, za hipotezę zerową będziemy stawiać założenie, które nie zakłada żadnych różnic między nimi. Zaprzeczeniem hipotezy zerowej jest **hipoteza alternatywna**, którą oznaczamy symbolem H_1. Określa ona wszystkie lub część hipotez ze zbioru hipotez dopuszczalnych nie będących hipotezą zerową. W wyniku odrzucenia hipotezy zerowej, możemy uznać hipotezę alternatywną za prawdziwą. W przypadku hipotezy zerowej, zbiór wartości parametru lub rodzinę rozkładów, które ona zakłada, będziemy oznaczać symbolem Θ_0. W przypadku hipotezy alternatywnej — symbolem Θ_1.

W podjęciu decyzji o odrzuceniu hipotezy zerowej pomaga nam test statystyczny. **Testem statystycznym** jest ściśle określone postępowanie, w wyniku którego, dla każdej wylosowanej próby z populacji, podejmujemy decyzję o odrzuceniu hipotezy zerowej dotyczącej całej populacji. Test ten oznaczamy symbolem φ. Wyróżniamy dwa rodzaje testów statystycznych:

Testy parametryczne służące do testowania hipotez parametrycznych. W celu ich przeprowadzenia potrzebna jest znajomość rozkładu badanego parametru w populacji generalnej. Przykładami takichh testów są m.in. test Studenta.

Testy nieparametryczne słżące do weryfikacji hipotez nieparametrycznych. Przykładami takich testów są: test Kołmogorowa-Smirnowa oraz test Shapiro-Wilka.

W celu prawidłwego przeprowadzenia testu statystycznego próba musi być losowo wybrana z populacji generalnej, a jej obserwacje nie mogą być od siebie zależne.

Biorąc pod uwagę fakt, że próba jest wybierana w sposób losowy, ten sam test statystyczny może prowadzić do podjęcia dwóch różnych decyzji dla dwóch różnych prób, przez co możemy popełnić jeden z dwóch rodzajów błędów. **Błędem I rodzaju** nazywamy decyzję o odrzuceniu hipotezy zerowej, wtedy kiedy jest ona prawdziwa. **Błąd II rodzaju** popełniamy natomiast, kiedy przyjmujemy za prawdziwą hipotezę zerową, kiedy w rzeczywistości jest ona fałszywa.

Ważnymi elementami przeprowadzenia weryfikacji hipotez są prawdopodobieństwa popełnienia powyższych błędów. Ryzyko popełnienia błędu I rodzaju, nazywamy **poziomem istotności** oraz oznaczamy symbolem α, a prawdopodobieństwo popełnienia błędu II rodzaju oznaczamy symobolem β. Prawdopodobieństwo wystąpienia zdarzenia przeciwnego do sytuacji, w której popełniamy błąd II rodzaju, nazywamy **mocą testu**. Innymi słowy moc testu określa prawdopodobieństwo przyjęcia hipotezy alternatywnej, kiedy jest ona prawdziwa. Współczynnik ten oznaczamy symbolem γ.

Niestety, nie możemy skonstruować testu, który jednocześnie minimalizowałby współczynniki α i β. W celu udowodnienia tego, załóżmy, że przyjmujemy $\alpha = 0$ i chcemy kontrolować wpółczynnik β na poziomie 0. Z czego wynika, że zawsze przyjmujemy hipotezę zerową, nawet kiedy jest ona fałszywa. przez co ryzyko popełnienia błędu II rodzaju jest większe od 0 i przeczy wcześniejszemu założeniu, że $\beta = 0$.

W celu prawidłowego preprowadzenia weryfikacji hipotez statystycznych, powinniśmy najpierw ustalić poziom istotności. Zazwyczaj, będziemy ustalać jego wartość na poziomie 0.10, 0.05. Bezpieczniejsze jest przeprowadzenie przez nas testu hipotez statystycznych w oparciu o ryzyko pepełnienia błędu I rodzaju, niż w oparciu o ryzyko popełnienia błędu II rodzaju. Rozpatrzmy sytuację, w której chcemy sprawdzić skuteczność nowo opracowanego leku zmniejszającego ciśnienie krwi, przeznaczonego dla osób chorych na nadciśnienie. Za hipotezę zerową przyjmujemy założenie, że lek nie zmniejsza ciśnienia krwi, zaś za alternatywę — założenie, że lek zmniejsza ciśnienie krwi. Skutkiem podjęcia decyzji o odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej może być dopuszczenie leku do sprzedaży. Korzystniejsze jest w tym przypadku popełnienie błędu II rodzaju, czyli

stwierdzenia, że lek nie działa, kiedy w rzeczywistości jest on skuteczny, czego konsekwencją jest niedopuszczenie leku do dystrybucji. Z kolei przyjęcie skuteczności leku, kiedy w rzeczywistości on nie działa, czyli popełnienie błędu I rodzaju, wiąże się z ryzykiem podania tego leku osobie chorej w sytuacji zagrażającej jej życiu. Sytuacja ta pokazuje, że bezpieczniejsze jest, podczas przeprowadzania testowania hipotez statystycznych, kontrolowanie ryzyka popełnienia błędu I rodzaju.

Jako że dla jednej pary hipotez oraz ustalonego poziomu istotności możemy skonstruować wiele różnych testów statystycznych, decyzję o odrzuceniu hipotezy zerowej podejmujemy w oparciu o test, który ma największą moc. Test, który spośród innych testów na tym samym poziomie istotności α, przyjmuje największą moc, nazywamy **testem jednostajnie najmocniejszym** na poziomie istotności α. Niestety, nie zawsze taki test istnieje.

2.2 Procedura weryfikacji hipotez statystycznych

Proces testowania hipotez statystycznych rozpoczynamy od określenia hipotezy zerowej, dotyczącej rozkładu badanego parametru populacji generalnej, oraz jej alternatywy. Układ tych hipotez nazywamy modelem statystycznym. W kolejnym kroku przyjmujemy poziom istotności na podstawie, którego testowanie ma przebiegać. Następnie losujemy n-elementową próbę obserwacji z populacji generalnej. Zazwyczaj taką próbę oznaczamy symbolem $\underline{X}$, a i-tą obserwcje symbolem x_i. W celu poznania kolejnych kroków procedury weryfikacji hipotez, spróbujmy przeprowadzić ją na poniższym przykładzie.

Przyjmijmy, że dysponujemy próbą $\underline{X}$ złożoną z 16 niezależnych zmiennych losowych, pochodzących z rozkładu normalnego, o nieznanej średniej μ oraz wariancji $\sigma^2 = 1$. Wiemy, że średnia arytmetyczna tej próby — $\hat{\mu}$ wynosi 0.4375. Chcemy sprawdzić, na poziomie istotności $\alpha = 0.05$, czy parametr μ jest równy 0. W tym celu tworzymy układ hipotez statystycznych, które chcemy przetestować:

$$\begin{aligned} H_0 &: \quad \mu = 0, \\ H_1 &: \quad \mu \neq 0. \end{aligned} \tag{A}$$

Kolejnym krokiem, w celu przeprowadzenia testu, jest wyznaczenie wartości statystyki testowej. Taką statystyką nazywamy statystykę pozwalającą orzekać w powyższym i jakimkolwiek innym problemie testowania hipotez, oraz oznaczamy:

$$T\left(\underline{X}\right) = T((x_1, x_2, \ldots, x_n)).$$

W zależności od przeprowadzanego eksperymentu statystyka testowa ma ściśle określony wzór, w tym przypadku ma postać:

$$T\left(\underline{X}\right) = \frac{\hat{\mu} - \mu_{H_0}}{\sigma}\sqrt{n} = \frac{0.4375 - 0}{1} \cdot \sqrt{16} = 1.75.$$

Widzimy, że statystyka testowa zależy od hipotezy zerowej, a także od wyników próby, z czego wynika, że każda statystyka testowa jest zmienną losową, której rozkład określamy na podstawie założenia prawdziwości hipotezy zerowej. W tym przypadku statystyka T ma rozkład standardowy normalny.

Znając rozkład statystyki testowej możemy przejść do kolejnej czynności, czyli wyznaczenia obszaru krytycznego. Zbiór ten najczęściej oznaczamy symbolem C. Tworzą go wartości, które, przy założeniu prawdziwości hipotezy zerowej, są małoprawdopodobne do osiągnięcia przez statystykę testową. Wartości te jednocześnie zaprzeczają hipotezie zerowej. W przypadku zaobserwowania, że wartość naszej statystyki testowej należy do obszaru krytycznego, możemy odrzucić hipotezę zerową. Na położenie obszaru krytycznego ma wpływ hipoteza alternatywna. W zależności od jej formy wyróżniamy trzy rodzaje obszaru krytycznego:

- Dwustronny obszar krytyczny:

$$C = \left(-\infty, t_{\frac{\alpha}{2}}\right) \cup \left(t_{1-\frac{\alpha}{2}}, \infty\right),$$

gdzie $t_{\frac{\alpha}{2}}$, $t_{1-\frac{\alpha}{2}}$ są kwantylami rozkładu statystyki testowej, rzędów odpowiednio: $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$. Obszar ten wyznaczamy na podstawie hipotezy alternatywnej niekierunkowej.

- Lewostronny obszar krytyczny:

$$C = (-\infty, t_\alpha),$$

gdzie t_α jest kwantylem rzędu α rozkładu statystyki T. Przedział ten wyznaczamy, kiedy alternatywa jest lewostronna.

- Prawostronny obszar krytyczny:

$$C = (t_{1-\alpha}, \infty),$$

gdzie $t_{1-\alpha}$ jest kwantylem rzędu $1-\alpha$ rozkładu statystyki testowej. Powyższy obszar wyznacza hipoteza alternatywna prawostronna.

Wartości brzegowe obszaru krytycznego nazywamy wartościami krytycznymi. Dopełnienie obszaru krytycznego stanowią wartości statystyki testowej, które potwierdzają prawdziwość hipotezy zerowej. Zbiór ten nazywamy obszarem przyjęcia hipotezy zerowej.

Kolejnym sposobem, który pozwala nam zweryfikować, czy hipoteza zerowa jest prawdziwa, jest wyznaczenie ***p*-wartości**. Nazywamy w ten sposób przekształcenie statystyki testowej. Tak samo jak w przypadku obszaru krytycznego, wzór p-wartości zależy od kierunku, wyznaczanego przez hipotezę alternatywną. Jeśli alternatywa jest niekierunkowa, p-wartość definiujemy jako:

$$p := 2\min\left\{\mathbb{F}_{H_0}\left(T\left(\underline{X}\right)\right), 1-\mathbb{F}_{H_0}\left(T\left(\underline{X}\right)\right)\right\},$$

gdzie $\mathbb{F}_{H_0}$ jest dystrybuantą rozkładu statystyki tesowej, przy założeniu prawdziwości hipotezy zerowej. W sytuacji gdzie hipoteza alternatywna jest lewostronna, p-wartość ma wzór:

$$p := \mathbb{F}_{H_0}\left(T\left(\underline{X}\right)\right),$$

a w sytuacji kiedy alternatywa jest prawostronna:

$$p := 1-\mathbb{F}_{H_0}\left(T\left(\underline{X}\right)\right).$$

Hipotezę zerową odrzucamy, gdy wartość p jest mniejsza od poziomu istotności testu.

W naszym modelu statystycznym (A) hipoteza alternatywna jest niekierunkowa, przez co wyznacza dwustronny obszar krytyczy. Wiedząc, że statystyka testowa pochodzi z rozkładu $N(0,\,1)$, a poziom istotności równy jest 0.05, obliczmy wartości brzegowe:

$$t_{\frac{\alpha}{2}} = t_{0.025} = -1.96,$$

$$t_{1-\frac{\alpha}{2}} = t_{0.975} = 1.96.$$

Z powższego wynika, że obszar krytyczny ma postać:

$$C_A = (-\infty, -1.96) \cup (1.96, \infty).$$

Obliczmy jeszcze p-wwartość:

$$p_A = 2\min\left\{\Phi(1.75), 1-\Phi(1.75)\right\} = 2\cdot(1-\Phi(1.75)) = 2\cdot 0.04 = 0.08$$

Widzimy, że wartość satystyki testowej — $T(\underline{X}) = 1.75$, nie należy do obszaru krytycznego oraz p-wartość jest większa od poziomu istotności. Możemy przez to odrzucić prawdziwość hipotezy zerowej — $H_0 : \mu = 0$.

W kolejnym przypadku, bazując ciągle na danych z modelu (A) oraz zakładając ten sam poziom istotności, rozpatrzmy modele:

$$\begin{aligned} H_0 : &\quad \mu = 0, \\ H_1 : &\quad \mu < 0, \end{aligned} \tag{B}$$

oraz:

$$\begin{aligned} H_0 : &\quad \mu = 0, \\ H_1 : &\quad \mu > 0. \end{aligned} \tag{C}$$

Dostrzegamy, że powyższe modele różnią się od modelu (A) tylko formami hipotez alternatywnych, przez co statystyka testowa przyjmuje tą samą wartość dla wszystkich tych modeli. W przypadku modelu (B) wyznaczamy lewostronny obszar krytyczny, którego wartością brzegową jest kwantyl rzędu 0.05 rozkładu $N(0,\, 1)$.

$$t_\alpha = t_{0.05} = -1.64.$$

$$C_B = (-\infty, t_\alpha) = (-\infty, -1.64)\,.$$

Wyznaczmy jeszcze wartość p:

$$p = \Phi(1.75) = 0.96.$$

Ponieważ $T(\underline{X}) \notin C_B$ i $p > 0.05$, uznajemy hipotezę zerową modelu (B) za prawdziwą. Dla modelu (C) wyznaczamy prawostronny obszar krytyczny, którego wartość krytyczna równa jest kwanntylowi rzędu 0.95 standardowego rozkładu normalnego.

$$t_{1-\alpha} = t_{0.95} = 1.64.$$

$$C_C = (t_\alpha, \infty) = (1.64, \infty)\,.$$

P-wartość w tym przypadku liczymy jako:

$$p = 1 - \Phi(1.75) = 0.04.$$

W tym modelu wartość statystyki testowej należy do obszaru krytycznego, a p-wartość jest mniejsza od poziomu istotności. Odrzucamy więc hipotezę zerową modelu (C) na rzecz jej alternatywy.

Widzimy, że test i-tego modelu, gdzie $i \in \{(\mathrm{A}), (\mathrm{B}), (\mathrm{C})\}$, możemy utożsamić z funkcją:

$$\varphi_i : \quad \underline{X} \to \{0, 1\}\,,$$

$$\varphi_i(\underline{X}) = \mathbb{1}_{C_i}\left(T\left(\underline{X}\right)\right).$$

Uzyskaną przez funkcję γ_i wartość „1" interpretujemy jako decyzję o odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej, a wartość „0" jako decyzję o przyjęciu hipotezy zerowej. Moc testu i-tego modelu wyznaczamy funkcją:

$$\gamma_i : \quad \mathbb{R} \to [0, 1],$$

$$\gamma_i(\mu) = \mathbb{E}_\mu[\varphi_i\left(\underline{X}\right)],$$

gdzie $\mu \in \Theta_{1i}$.

2.3 Lemat Neymana-Pearsona

Jeden ze skutecznych sposobów przeprowadzenia weryfikacji hipotez statystycznych podaje nam Lemat Neymana-Pearsona. Wykorzystanie tego twierdzenia pozwala nam skonstruować test jednostajnie najmocniejszy na poziomie istotności α. W przypadku, kiedy nasz model statystyczny opiera się na parze hipotez prostych, a rozmiar próby, w oparciu, o którą test ma zostać przeprowadzony, jest znany. Test ten opiera się na ilorazie wiarygodności.

Dla próby $\underline{X} = (x_1, \ldots, x_n)$, z rozkładu o funkcji prawdopodobieństwa $f(x, \theta)$, gdzie parametr $\theta \in \Theta$, ilorazem wiarygodności nazywamy wyrażenie:

$$r(H_0, H_1) = \frac{L\left(\theta_1, \underline{X}\right)}{L\left(\theta_0, \underline{X}\right)},$$

gdzie $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. Natomiast funkcję:

$$L : \Theta \to \mathbb{R},$$

$$L(\theta, \underline{X}) = \prod_{i=1}^{n} f(x_i),$$

nazywamy funkcją wiarygodności.

Lemat 2.1 (Neymana-Pearsona)**.** *Niech $\underline{X} = (x_1, \ldots, x_n)$, przy $n \geqslant 1$, będzie próbą niezależnych zmiennych z rozkładu prawdopodobieństwa F o funkcji gęstości lub funkcji masy prawdopodobieństwa $f(x, \theta)$. Model statystyczny niech będzie oparty na prostych hipotezach:*

$$\begin{aligned} H_0 : &\quad \theta = \theta_0, \\ H_1 : &\quad \theta = \theta_1. \end{aligned}$$

Wtedy, test φ jednostajnie najmocniejszy na poziomie istotności α określa funkcja:

$$\varphi\left(\underline{X}\right) = \begin{cases} 1 & gdy \quad \prod\limits_{i=1}^{n} f_0(x_i) < k \prod\limits_{i=1}^{n} f_1(x_i) \\ \delta & gdy \quad \prod\limits_{i=1}^{n} f_0(x_i) = k \prod\limits_{i=1}^{n} f_1(x_i) \\ 0 & gdy \quad \prod\limits_{i=1}^{n} f_0(x_i) > k \prod\limits_{i=1}^{n} f_1(x_i) \end{cases},$$

gdzie $k \geqslant 0$ oraz $\delta \in [0, 1]$, są stałymi wyznaczonymi z warunku:

$$\mathbb{E}_{H_0}[\varphi(\underline{X})] = \alpha.$$

Obszar krytyczny testu φ ma postać:

$$C = \left\{ (x_1, \ldots, x_n) : \frac{L\left(\theta_1, \underline{X}\right)}{L\left(\theta_0, \underline{X}\right)} > k \right\}.$$

W celu lepszego zrozumienia Lematu Neymana-Pearsona rozpatrzmy poniższy przykład.

Przykład 2.1. Załóżmy, próbę niezależnych zmiennych losowych: $\underline{X} = \{-0.50, 0.13, -0.08, 0.89, 0.12, 0.32, -0.58, 0.71\}$, pochodzi z rozkładu normalnego, o nieznanym parametrze średniej μ oraz wariancji równej 1. Przyjmijmy poziom istotności $\alpha = 0.05$. Przetestujmy, przy pomocy lematu Neymana-Pearsona, poniższy model statystyczny:

$$\begin{aligned} H_0 : &\quad \mu = 0, \\ H_1 : &\quad \mu = 4. \end{aligned}$$

Na początku wyznaczmy funkcje wiarygodności dla powyższych hipotez:

$$L_0\left(\theta_0, \underline{X}\right) = \prod_{i=1}^{8} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_i^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^8 \exp\left(\frac{-\sum\limits_{i=1}^{8} x_i^2}{2}\right),$$

$$L_1\left(\theta_1, \underline{X}\right) = \prod_{i=1}^{8} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-4)^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^8 \exp\left(\frac{-\sum\limits_{i=1}^{8} (x_i-4)^2}{2}\right).$$

W kolejnym kroku wyznaczmy iloraz wiarygodności:

$$\begin{aligned} r(H_0, H_1) = \frac{L_1\left(\theta_1, \underline{X}\right)}{L_0\left(\theta_0, \underline{X}\right)} &= \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{8}(x_i-4)^2 - \sum_{i=1}^{8} x_i^2\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(-8\sum_{i=1}^{8} x_i + 128\right)\right) \\ &= \exp\left(4\sum_{i=1}^{8} x_i - 64\right), \end{aligned}$$

gdzie:

$$\begin{aligned} \exp\left(4\sum_{i=1}^{8} x_i - 64\right) > k_1 \quad &/ \ln \\ 4\sum_{i=1}^{8} x_i - 64 > \ln k_1 \\ \sum_{i=1}^{8} x_i > \frac{\ln k_1 - 64}{4} = k. \end{aligned}$$

Zatem obszar krytyczny jednostajnie najmocniejszego testu φ ma postać:

$$C = \left\{(x_1, \ldots, x_8) : \exp\left(4\sum_{i=1}^{8} x_i - 64\right) > k_1\right\} = \left\{(x_1, \ldots, x_8) : \sum_{i=1}^{8} x_i > k\right\}$$

Obliczmy wartość k z warunku:

$$P\left(\underline{X} \in C \mid H_0\right) = \alpha.$$

Wiemy, że dla niezależnych zmiennych losowych $y_1, \ldots, y_n$ ze standardowego rozkładu nor-

malnego , $\sum_{i=1}^{n} y_i \sim N(0,n)$. Z tego wynika:

$$\begin{aligned} P\left(\underline{X} \in C \mid H_0\right) &= P\left(\sum_{i=1}^{8} x_i > k \mid H_0\right) \\ &= P\left(\frac{1}{\sqrt{8}} \sum_{i=1}^{8} x_i > \frac{k}{\sqrt{8}} \mid H_0\right) \\ &= 1 - \Phi\left(\frac{k}{\sqrt{8}}\right). \end{aligned}$$

Wobec tego wartość k wynosi:

$$\begin{aligned} \frac{k}{\sqrt{8}} &= \Phi^{-1}(1 - 0.05) \\ k &= \sqrt{8}\Phi^{-1}(0.95) \approx 4.6523. \end{aligned}$$

Widzimy, że:

$$\sum_{i=1}^{8} x_i = 1.01 < k.$$

W skutek tego możemy przyjąć hipotezę zerową.

3 Rozkład p-wartości przy założeniu hipotezy zerowej

Z poprzedniego rozdziału wiemy, że p-wartość jest przekształceniem statystki testowej. Inaczej, p-wartość określa prawdopodobieństwo, przy założeniu prawdziwości hipotezy zerowej, przyjęcia przez statystykę testową wartości „ekstremalnych", sygerujących, aby w wyniku testu odrzucić hipotezę zerową. Z tego wynika, że czynnikiem, który wpływa na rozkład p-wartości, jest rozkład statystyki testowej. Spróbujmy określić rozkład p-wartości w dwóch przypadkach: kiedy statystyka testowa przyjmuje rozkład ciągły oraz, kiedy statystyka testowa przyjmuje rozkład dyskretny.

3.1 Rozkład p-wartości opartej na ciągłej statystyce testowej

Rozpatrzmy sytuację, kiedy statystyka testowa pochodzi z ciągłego rozkładu prawdopodobieństwa. Rozwiązanie problemu określenia rozkładu p-wartości w tej sytuacji podaje nam poniższe twierdzenie.

Twierdzenie 3.1. *Kiedy statystyka testowa pochodzi z rozkładu bezwzględnie ciągłego, wyznaczona na jej podstawie p-wartość przyjmuje rozkład jednostajny na przedziale* $[0,1]$.

Dowód. Załóżmy, że w wyniku lewostronnego testu prostej hipotezy zerowej, opartego na próbie niezależnych zmiennych losowych: $\underline{X} = (x_1, \ldots, x_n)$, gdzie $n \geqslant 1$, pochodzącej z ciągłego rozkładu prawdopodobieństwa $\mathcal{G}$, otrzymaliśmy statystykę testową: $T\left(\underline{X}\right)$. Niech przy założeniu hipotezy zerowej:

$$T\left(\underline{X}\right) \sim \mathcal{F}.$$

Jako że próba $\underline{X}$ pochodzi z rozkładu ciągłego, bez straty ogólności rozważań, załóżmy, że $\mathcal{F}$ jest ciągłym rozkładem prawdopodobieństwa, którego dystrybuanta $\mathbb{F}$ jest ściśle rosnąca dla nośnika rozkładu $\mathcal{F}$. Następstwem tego jest fakt, że istnieje funkcja kwantylowa $\mathbb{F}^{-1}$ na zbiórze wartości, dla których gęstość rozładu $\mathcal{F}$ przyjmuje wartości niezerowe, czyli nośnika rozkładu $\mathcal{F}$. Przyjmijmy zmienną losową $z \sim \mathcal{F}$. Wyznaczmy p-wartość:

$$p = P\left(z < T\left(\underline{X}\right) \mid H_0\right) = \mathbb{F}\left(T\left(\underline{X}\right)\right).$$

Z tego wynika, że $p \in (0,1)$. Przyjmijmy zmienną $x \in (0,1)$, oraz obliczmy prawdopodobieństwo, przy założeniu prostej hipotezy zerowej, tego że $p \leqslant x$:

$$\begin{aligned} P(p \leqslant x \mid H_0) &= P\left(\mathbb{F}\left(T\left(\underline{X}\right)\right) \leqslant x \mid H_0\right) \\ &= P\left(T\left(\underline{X}\right) \leqslant \mathbb{F}^{-1}(x) \mid H_0\right) \\ &= \mathbb{F}\left(\mathbb{F}^{-1}(x)\right) \\ &= x. \end{aligned}$$

Wiedząc, że dla każdej zmiennej losowej $y \sim \mathcal{U}[0,1]$, spełniony jest warunek:

$$P(y \leqslant x) = x,$$

równość:

$$P(p \leqslant x \mid H_0) = x$$

dowodzi, że p-wartość pochodzi z rozkładu $\mathcal{U}[0,1]$.

■

Wykorzystanie testu prawostronnego lub testu niekierunkowego nie zmienia faktu, że wartość p, wyznaczona na podstawie statystyki o rozkładze ciągłym, ma rozkład $\mathcal{U}[0,1]$. Dowód oparty na teście lewostronnym jest najmniej skomplikowany i najłatwiejszy do przeprowadzenia.

3.2 Rozkład p-wartości opartej na dyskretnej statystyce testowej

Rozpatrzmy przypadek, kiedy statystyka testowa pochodzi z rozkładu dyskretnego. W tej sytuacji, rozkładu p-wartości nie można określić jednoznacznie, jednak w niektórych sytuacjach, rozważane przekształcenie statystyki testowej przyjmuje rozkład asymptotycznie jednostajny na przedziale $[0,1]$. Oznacza to, że kiedy rozmiar próby, na podstawie której określona jest wartość statystyki testowej, dąży do nieskończoności, to rozkład p-wartości dąży do rozkładu $\mathcal{U}[0,1]$. Pniższa uwaga wynika z Prawa Wielkich Liczb oraz Twierdzenia Gliwienki-Cantellego.

Uwaga 3.1. *Kiedy statystyka testowa pochodzi z rozkładu dyskretnego, p-wartość wyznaczona na jej podstawie może przyjąć rozkład asymptotycznie jednostajny na przedziale* $[0,1]$.

Aby sprawdzić prawdziwość powyższej uwagi, rozważmy poniższy przykład.

Przykład 3.1. Niech $\underline{X} = (x_1, \ldots, x_n)$, gdzie $n \geqslant 1$, będzie próbą niezależnych zmiennych losowych z rozkładu Poissona o parametrze λ. Rozważmy, na poziomie istotności $\alpha \in [0,1]$, prawostronny test hipotez:

$$\begin{aligned} H_0: &\quad \lambda = \lambda_0, \\ H_1: &\quad \lambda > \lambda_0, \end{aligned}$$

gdzie $\lambda_0 \in (0, \infty)$. Wyznaczmy statystykę testową:

$$T\left(\underline{X}\right) = \sum_{i=1}^{n} x_i.$$

Widzimy, że przy założeniu hipotezy zerowej, statystyka $T\left(\underline{X}\right)$ przyjmuje rozkład Poissona o parametrze $n\lambda_0$. Załóżmy, że zmienna losowa z pochodzi z rozkładu Poissona o parametrze $n\lambda_0$. Wyznaczmy p-wartość:

$$p = P\left(z > T\left(\underline{X}\right) \mid H_0\right) = 1 - \mathbb{F}\left(T\left(\underline{X}\right)\right),$$

gdzie $\mathbb{F}(\bullet)$ jest dystrybuantą rozkładu Poissona o parametrze $n\lambda_0$. Niech $x \in [0,1]$. Obliczmy prawdopodobieństwo, przy założeniu prawdziwości hipotezy zerowej, że $p \leqslant x$:

$$\begin{aligned} P(p \leqslant x \mid H_0) &= P\left(1 - \mathbb{F}\left(T\left(\underline{X}\right)\right) \leqslant x \mid H_0\right) \\ &= P\left(\mathbb{F}\left(T\left(\underline{X}\right)\right) \geqslant 1 - x \mid H_0\right). \end{aligned} \tag{1}$$

Niech

$$k = \min\left\{m : \quad \sum_{i=0}^{m} \frac{(n\lambda_0)^i e^{-n\lambda_0}}{i!} \geqslant 1 - x\right\}.$$

Z tego wynika, że:

$$\begin{aligned} P(p \leqslant x \mid H_0) &= P\left(T\left(\underline{X}\right) \geqslant k \mid H_0\right) \\ &= 1 - \mathbb{F}(k), \end{aligned} \tag{2}$$

gdzie:

$$1 - x \leqslant \mathbb{F}(k) \leqslant 1 - x + \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!}. \tag{3}$$

Oznaczmy $L = n\lambda_0$, $a_k = \frac{L^k e^{-L}}{k!}$. Wykażmy, że przy $n \to \infty$, $\mathbb{F}(k) \to 1 - x$. Ponieważ $L > 0$ oraz $k \geqslant 0$, to $\forall k : a_k > 0$. Sprawdźmy monotoniczność ciągu a_k. Ciąg jest rosnący, kiedy:

$$a_k < a_{k+1} \iff \frac{a_{k+1}}{a_k} > 1.$$

Wyznaczmy:

$$\frac{a_{k+1}}{a_k} = \frac{\frac{L^{k+1}e^{-L}}{(k+1)!}}{\frac{L^k e^{-L}}{k!}} = \frac{L}{k+1}.$$

Z tego wynika, że:

$$\frac{a_{k+1}}{a_k} > 1 \iff k < \lfloor L \rfloor,$$

gdzie $\lfloor r \rfloor$ oznacza część całkowitą liczby rzeczywistej r. Natomiast ciąg a_k jest malejący, kiedy:

$$a_k > a_{k+1} \iff \frac{a_{k+1}}{a_k} < 1 \iff k > \lfloor L \rfloor.$$

Z tego wynika, że:

$$\sup_{k \geqslant 0}\left(\frac{L^k e^{-L}}{k!}\right) = \frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor!}.$$

Wobec tego:

$$\begin{aligned} \lim_{n\to\infty}\left(\sup_{k\geqslant 0}\left(\frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!}\right)\right) &= \lim_{L\to\infty}\left(\sup_{k\geqslant 0}\left(\frac{L^k e^{-L}}{k!}\right)\right) \\ &= \lim_{L\to\infty}\left(\frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor!}\right). \end{aligned}$$

Oszacujmy:

$$0 \leqslant \frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor!} \leqslant \frac{(\lfloor L \rfloor + 1)^{\lfloor L \rfloor} e^{-\lfloor L \rfloor}}{\lfloor L \rfloor!}. \tag{4}$$

Korzystając ze wzoru Stirlinga, dla $i \in \mathbb{N}_+$:

$$i! \approx \left(\frac{i}{e}\right)^i \sqrt{2\pi i}, \qquad \text{gdzie } \lim_{i\to\infty} \frac{i!}{\left(\frac{i}{e}\right)^i \sqrt{2\pi i}} = 1, \tag{5}$$

obliczmy:

$$\begin{aligned}
\lim_{L\to\infty}\left(\frac{(\lfloor L\rfloor+1)^{\lfloor L\rfloor}\,e^{-\lfloor L\rfloor}}{\lfloor L\rfloor!}\right) &= \lim_{L\to\infty}\left(\frac{(\lfloor L\rfloor+1)^{\lfloor L\rfloor}\,e^{-\lfloor L\rfloor}}{\lfloor L\rfloor!}\cdot\frac{\lfloor L\rfloor+1}{\lfloor L\rfloor+1}\right)\\
&= \lim_{L\to\infty}\left(\frac{(\lfloor L\rfloor+1)^{\lfloor L\rfloor+1}\,e}{(\lfloor L\rfloor+1)!\cdot e^{\lfloor L\rfloor+1}}\right)\\
&\overset{(5)}{=} \lim_{L\to\infty}\left(\frac{(\lfloor L\rfloor+1)^{\lfloor L\rfloor+1}\,e}{\left(\frac{\lfloor L\rfloor+1}{e}\right)^{\lfloor L\rfloor+1}\sqrt{2\pi\,(\lfloor L\rfloor+1)}\cdot e^{\lfloor L\rfloor+1}}\right)\\
&= \lim_{L\to\infty}\left(\frac{e}{\sqrt{2\pi\,(\lfloor L\rfloor+1)}}\right)\\
&= 0. \qquad (6)
\end{aligned}$$

Z Twierdzenia o trzech ciągach, (4) oraz (6) wynika, że:

$$\lim_{n\to\infty}\left(\sup_{k\geqslant 0}\left(\frac{(n\lambda_0)^k\,e^{-n\lambda_0}}{k!}\right)\right)=0. \qquad (7)$$

Poprzez (3), (2) pokazaliśmy, że:

$$\mathbb{F}(k)\xrightarrow{n\to\infty}1-x. \qquad (8)$$

Z (1), (2), (8) wynika, że:

$$P(p\leqslant x\mid H_0)\xrightarrow{n\to\infty}1-(1-x)=x.$$

W ten sposób zareprezentowaliśmy, że rozkład p-wartości, dla tego przykładu, jest asymptotycznie jednostajny na przedziale $[0,1]$.

4 Globalna hipoteza zerowa oraz sposób jej weryfikacji

W celu znalezienia w zbiorze danych próby, która w stosunku do innych prób, różni się rozkładem prawdopodobieństwa, możemy posłużyć się wielokrotnym testowaniem hipotez. Sposób ten polega na przeprowadzeniu jednakowego testu hipotez każdej próby należącej do badanego zbioru danych. W sytuacji, kiedy liczba prób jest duża, sposób ten jest nieoptymalny. Wynika to z tego, że przy pomocy wielokrotnej weryfikacji hipotez możemy znaleźć nie tylko jedną próbę odstającą, ale również czas przeprowadzenia testu, może być zbyt długi. Bardziej optymalnym rozwiązaniem powyższego problemu jest weryfikacja modelu typu „igła w stogu siana", który jest oparty o globalną hipotezę zerową.

4.1 Globalna hipoteza zerowa

Globalna hipoteza zerowa jest pojęciem związanym z wielokrotnym testowaniem hipotez statystycznych. Załóżmy, że dysponujemy zbiorem danych złożonym z prób: $\underline{X_1},\ldots,\underline{X_n}$, gdzie $n>1$. Niech każda próba $\underline{X_i}$, gdzie $i\in\{1,\ldots,n\}$, pochodzi z rozkładu $\mathcal{F}$ o parametrze $\theta\in\mathbb{R}$. Dla każdej próby $\underline{X_i}$ chcemy przeprowadzić test oparty na modelu:

$$\begin{aligned}
H_{0_i}:&\quad \theta_i=0,\\
H_{1_i}:&\quad \theta_i\neq 0.
\end{aligned}$$

Globalną hipotezą zerową nazywamy wyrażenie:

$$H_0 = \bigcap_{i=1}^{n} H_{0_i},$$

z tego wynika, że globalna hipoteza zerowa, w tym przypadku, ma postać:

$$H_0 : \quad \theta_1 = \ldots = \theta_n = 0.$$

Globalną hipotezę zerową przyjmujemy za prawdziwą wyłącznie wtedy, gdy każda z hipotez: $H_{0_1}, \ldots, H_{0_n}$ jest prawdziwa. Alternatywa powyższej, globalnej hipotezy zerowej ma postać:

$$H_1 : \quad \exists i \in \{1, \ldots, n\} : \theta_i \neq 0.$$

Zakłada ona, że przynajmniej jedna próba pochodzi z rozkładu $\mathcal{F}$ o parametrze $\theta \neq 0$, ale nie określa ile i które to są konkretnie próby.

4.2 Sposób testowania globalnej hipotezy zerowej

Jeden ze sposobów przeprowadzenia wielokrotnego testowania hipotez statystycznych i weryfikacji prawdziwości globalnej hipotezy zerowej, podaje nam **korekta Bonferroniego**. Procedura ta odrzuca globalną hipotezę zerową na poziomie istotności α, gdy:

$$\min_{i \in \{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n},$$

gdzie p_i jest p-wartością wyznaczoną na podstawie statystyki testowej: $T(\underline{X_i})$, uzyskanej w wyniku testu i-tego modelu statystycznego: (H_{0_i}, H_{1_i}). Uzyskanie prawidłowych wyników przez przeprowadzenie przez nas testu opartego na powyższej procedurze, potwierdza następujące twierdzenie.

Twierdzenie 4.1. *Korekta Bonferroniego kontroluje na poziomie α prawdopodobieństwo odrzucenia globalnej hipotezy zerowej, kiedy jest ona prawdziwa.*

Dowód. Niech $p_1, \ldots, p_n \sim \mathcal{U}[0, 1]$, gdzie $n \geqslant 2$, będą p-wartościami uzyskanymi odpowiednio w wyniku testów par hipotez: $(H_{0_1}, H_{1_1}), \ldots, (H_{0_n}, H_{1_n})$. Globalna hipoteza zerowa ma postać:

$$H_0 = \bigcap_{i=1}^{n} H_{0_i}.$$

Niech poziom istotności testu globalnej hipotezy zerowej wynosi α. Obliczmy prawdopodobieństwo popełnienia błędu I rodzaju, w wyniku testu globalnej hipotezy zerowej, opartego na korekcie Bonferroniego:

$$\begin{aligned} P\left(\min_{i \in \{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_0\right) &= P\left(\bigcup_{i=1}^{n} \left\{p_i \leqslant \frac{\alpha}{n}\right\} \mid H_0\right) \\ &\leqslant \sum_{i=1}^{n} P\left(p_i \leqslant \frac{\alpha}{n} \mid H_0\right) \\ &\leqslant \sum_{i=1}^{n} \frac{\alpha}{n} \\ &= \alpha. \end{aligned}$$

W sytuacji, kiedy wartości: $p_1, \ldots, p_n$ są niezależnymi zmiennymi losowymi, pochodzącymi z rozkładu $\mathcal{U}[0,1]$, powyższe prawdopodobieństwo obliczamy jako:

$$\begin{aligned} P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_0\right) &= 1 - P\left(\bigcap_{i=1}^{n}\left\{p_i \geqslant \frac{\alpha}{n}\right\} \mid H_0\right) \\ &= 1 - \prod_{i=1}^{n} P\left(p_i \geqslant \frac{\alpha}{n} \mid H_0\right) \\ &= 1 - \left(1 - \frac{\alpha}{n}\right)^n \\ &\approx 1 - e^{-\alpha} \\ &\leqslant \alpha. \end{aligned}$$

Z powyższego wynika, że korekta Bonferroniego kontroluje poziom istotności testu globalnej hipotezy zerowej.

■

Korekta Bonferroniego pozwala prawidłowo zweryfikować globalną hipotezę zerową. Oznacza to również, że test oparty na tej procedurze jest jednym z bardziej optymalnych sposobów testowania modeli typu „igła w stogu siana". W celu uzyskania większej wiedzy na temat sposobów weryfikacji globalnej hipotezy zerowej, możemy sięgnąć po artykół pt. „An omnibus test for the global null hypothesis"[3].

5 Modele typu „igła w stogu siana" oraz optymalne progi ich wykrycia

Definiując globalną hipotezę zerową oraz objaśniając sposób jej weryfikacji, możemy przejść do rozważań na temat modeli typu „igła w stogu siana". Załóżmy, że próby: $\underline{X_1}, \ldots, \underline{X_n}$, gdzie $n \geqslant 2$, pochodzą z rozkładu $\mathcal{F}$ o parametrze $\theta \in \mathbb{R}$. Niech hipoteza zerowa: $H_{0_i} : \theta_i = 0$, gdzie $i \in \{1, \ldots, n\}$, odpowiada próbie $\underline{X_i}$. Model statystyczny:

$$\begin{aligned} H_0 : &\quad \forall\ i \in \{1, \ldots, n\} : \theta_i = k, \\ H_1 : &\quad \exists!\ i \sim \mathcal{U}\{1, \ldots, n\} : \theta_i \neq k, \end{aligned}$$

gdzie $k \in \mathbb{R}$, nazywamy **modelem typu „igła w stogu siana"**. Aby skonstruować model tego typu, musimy sformułować hipotezę alternatywną do globalnej hipotezy zerowej, tak aby zakładała, że tylko jedna spośród badanych prób różni się rozkładem w stosunku do reszty. Powodem takiej konstrukcji jest fakt, że hipoteza alternatywna:

$$H_1' : \quad \exists\ i \in \{1, \ldots, n\} : \theta_i \neq k,$$

jest zbyt złożona, aby korekta Bonferroniego mogła odrzucić hipotezę H_0 na rzecz hipotezy H_1'. Procedura ta odrzuca hipotezę H_0 w oparciu o najmniejszą wartość p_i, więc nie możemy stwierdzić przy jej pomocy, czy pośród innych prób znajdują się kolejne, których parametr rozkładu jest różny od 0. Przykładem użycia modelu typu „igła w stogu siana" jest próba wykrycia, czy w zbiorze operacji finansowych znajduje się transakcja, która jest nieuczciwa. Jednym z rozwiązań tego problemu jest weryfikacja odpowiednio skonstruowanego modelu typu „igła w stogu siana". Aby znaleźć odpowiedni sposób weryfikacji rozważanych modeli, spróbujmy w następnych podrozdziałach przeanalizować modele oparte na danych pochodzących z rozkładów normalnego oraz Poissona.

5.1 Model oparty o dane pochodzące z rozkładu normalnego

Załóżmy, że zbiór danych, który chcemy poddać analizie, składa się z jednoelementowych prób: $\underline{X_1}, \ldots, \underline{X_n}$, gdzie $n \geqslant 2$. Niech, dla $i \in \{1, \ldots, n\}$, próba $\underline{X_i}$ pochodzi z rozkładu $\mathcal{N}(\mu_i, 1)$. Niech każdej próbie $\underline{X_i}$ odpowiada model statystyczny:

$$\begin{aligned} H_{0_i}: &\quad \mu_i = 0 \\ H_{1_i}: &\quad \mu_i > 0. \end{aligned}$$

Z tego wynika, że statystykę testową i-tej hipotezy zerowej obliczamy w następujący sposób:

$$T\left(\underline{X_i}\right) = \frac{\underline{X_i} - 0}{1} = \underline{X_i}.$$

Widzimy, że przy założeniu prawdziwości i-tej hipotezy zerowej, statystyka testowa $T\left(\underline{X_i}\right)$ pochodzi ze standardowego rozkładu normalnego. Wyznaczmy p-wartość opartą na i-tej statystyce testowej:

$$p_i = 1 - \Phi\left(T\left(\underline{X_i}\right)\right) = \left|\Phi\left(T\left(\underline{X_i}\right)\right)\right|$$

Głównym problemem, o który oprzemy dalsze rozważania, jest sprawdzenie, czy pośród prób: $\underline{X_1}, \ldots, \underline{X_n}$, znajduje się wyłącznie jedna, której parametr rozkładu μ jest większy od 0. Opierając się na tym problemie, skonstruujmy model typu „igła w stogu siana":

$$\begin{aligned} H_0: &\quad \forall\ i \in \{1, \ldots, n\} : \mu_i = 0, \\ H_1: &\quad \exists!\ i \sim \mathcal{U}\{1, \ldots, n\} : \mu_i > 0. \end{aligned} \tag{D1}$$

Wiemy, że Korekta Bonferroniego odrzuca globalną hipotezę zerową H_0 na poziomie istotnośc α, kiedy:

$$\min_{i \in \{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n},$$

co jest równozaczne, że H_0 zostanie odrzucona, gdy:

$$\max_{i \in \{1,\ldots,n\}} T\left(\underline{X_i}\right) \geqslant \left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|,$$

gdzie $\Phi^{-1}(\cdot)$ jest funkcją kwantylową standardowego rozkładu normalnego.

5.1.1 Optymalny próg wykrycia

Znalezienie dokładnej wartości kwantyla: $\left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$, stanowi utrudnienie, w przypadku kiedy liczba prób n jest zbyt duża, a wartość poziomu istotności α jest bliska zeru. W rozwiązaniu tego problemu pomaga nam przybliżenie wartości powyższego kwantyla. Jednak w sytuacji, kiedy różnica między rzeczywistą wartościa, a przybliżeniem liczby $\left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$ jest dostatecznie duża, skontruowany przez nas test statystyczny, oparty na przybliżeniu, może doprowadzić do uzykania błędnych rezultatów. Jedną ze skutecznych metod uzyskania odpowiedniej aproksymacji wyrażenia: $\left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$ podaje nam Nierówność Markowa.

Twierdzenie 5.1. *Niech, w przypadku testu jednostronneg, $t = \left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$, lub $t = \left|\Phi^{-1}\left(\frac{\alpha}{2n}\right)\right|$, w przypadku testu niekierunkowego. Dla zmiennej losowej Z, pochodzącej z rozkładu $\mathcal{N}(0, 1)$, prawdziwa jest nierówność:*

$$\frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) \leqslant P(Z > t) \leqslant \frac{\phi(t)}{t},$$

gdzie $\phi(\cdot)$ jest funkcją gęstości standardowego rozkładu normalnego.

Dowód. Niech Z będzie zmienną losową, pochodzącą z rokzładu $\mathcal{N}(0,1)$, o funkcji gęstości: $\phi(\cdot)$, oraz dystrybuancie: $\Phi(\cdot)$. Bez straty ogólności załóżmy, że $t = \left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$, gdzie $\alpha \in (0,1)$ oraz $n \in \mathbb{N}_+$.

(i) Udowodnijmy nierówność: $P(Z > t) \geqslant \frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right)$, która jest jednoznaczna z nierównością: $P(Z > t) - \frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) \geqslant 0$. Wiemy, że $t \in (0, \infty)$. Przeprowadźmy proste badanie przebiegu zmienności funkcji:

$$f: \quad (0,\infty) \to \mathbb{R},$$
$$f(t) = P(Z > t) - \frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) = 1 - \Phi(t) - \frac{\phi(t)}{t} + \frac{\phi(t)}{t^3}.$$

Wyzanczmy granice powyższej funkcji:

$$\begin{aligned}\lim_{t\to 0^+} f(t) &= \lim_{t\to 0^+}(1-\Phi(t)) - \lim_{t\to 0^+}(\phi(t)) \cdot \lim_{t\to 0^+}\left(\frac{t^2-1}{t^3}\right) \\ &= \left[\frac{1}{2} - \phi(0)\cdot(-\infty)\right] = \infty, \qquad (1)\\ \lim_{t\to \infty} f(t) &= \lim_{t\to \infty}(1-\Phi(t)) - \lim_{t\to \infty}(\phi(t)) \cdot \lim_{t\to \infty}\left(\frac{t^2-1}{t^3}\right) \\ &= [0 - 0\cdot 0] = 0. \qquad (2)\end{aligned}$$

Obliczmy pochodne pierwszego rzędu funkcji $\phi(t)$ oraz $f(t)$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\phi(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\right) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \cdot \left(-\frac{1}{2}\cdot 2t\right) \\ &= -t\cdot\phi(t), \qquad (3)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f(t) &= 0 - \frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\phi(t)\left(\frac{t^2-1}{t^3}\right)\right) \\ &= -\phi(t) - \left(\left(\frac{t^2-1}{t^3}\right)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\phi(t) + \phi(t)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t^2-1}{t^3}\right)\right) \\ &\overset{(3)}{=} -\phi(t) - \left(\frac{t^2-1}{t^3}\right)\cdot(-t\cdot\phi(t)) - \phi(t)\left(\frac{-t^2+3}{t^4}\right) \\ &= -\phi(t)\left(1 + \frac{t^2-t^4}{t^4} + \frac{3-t^2}{t^4}\right) \\ &= -\frac{3}{t^4}\phi(t).\end{aligned}$$

Widzimy, że:

$$\forall t \in (0,\infty): \frac{\mathrm{d}}{\mathrm{d}t}f(t) < 0, \qquad (4)$$

więc $f(t)$ jest funkcją malejącą, dla $t \in (0,\infty)$. Z (1), (2), (4) wynika, że prawdziwa jest nierówność:

$$P(Z > t) \geqslant \frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right).$$

(ii) Wykażmy, że $P(Z > t) \leqslant \frac{\phi(t)}{t}$. Niech $k > t$, wtedy:

$$\begin{aligned} P(Z > t) = \int_t^\infty \phi(u)\,\mathrm{d}u &\overset{(u \geqslant t)}{\leqslant} \int_t^\infty \frac{u}{t}\phi(u)\,\mathrm{d}u = \frac{1}{t}\int_t^\infty u \cdot \phi(u)\,\mathrm{d}u \\ &\overset{(3)}{=} \frac{1}{t}\int_t^\infty \frac{\mathrm{d}}{\mathrm{d}u}\left(-\phi(u)\right)\,\mathrm{d}u = -\frac{1}{t}\phi(t)\Big|_t^\infty \\ &= \frac{\phi(t)}{t}. \end{aligned}$$

Z rozważań zawartych w podpunktach (i) oraz (ii) wynika, że prawdziwa jest nierówność:

$$\frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) \leqslant P(Z > t) \leqslant \frac{\phi(t)}{t}.$$

■

Za pomocą powyższego twierdzenia, możemy wyaproksymować wartość: $t = \left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right|$. Wykazaliśmy w ten sposób, że dla $n \in \mathbb{N}_+$ oraz ustalonej wartości $\alpha \in (0, 1)$:

$$1 - \Phi(t) = \frac{\alpha}{n} \iff \frac{\phi(t)}{t} \approx \frac{\alpha}{n},$$

kiedy $t \xrightarrow{n\to\infty} \infty$. Rozpatrzmy następstwo tego faktu:

$$\begin{aligned} \frac{\phi(t)}{t} &\approx \frac{\alpha}{n} \\ \frac{1}{t\sqrt{2\pi}} e^{\frac{-t^2}{2}} &\approx \frac{\alpha}{n} \\ \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2} - \ln(t)} &\approx \frac{\alpha}{n} && \backslash\backslash\, \ln(\cdot) \\ -\ln\left(\sqrt{2\pi}\right) - \frac{t^2}{2} - \ln(t) &\approx \ln(\alpha) - \ln(n) \\ -\frac{t^2}{2} - \ln(t) &\approx \ln(\alpha) - \ln(n) + \ln\left(\sqrt{2\pi}\right) \end{aligned}$$

co, dla $n \to \infty$, jest równoważne z:

$$\begin{aligned} \frac{t^2}{2} &\approx \ln(n) \\ t &\approx \sqrt{2\ln(n)}. \end{aligned}$$

Więc:

$$\frac{\phi(t)}{t} \approx \frac{\alpha}{n} \iff t = \left|\Phi^{-1}\left(\frac{\alpha}{n}\right)\right| \approx \sqrt{2\ln(n)}.$$

Konsekwentnie, w przypadku modelu statystycznego (D1), stosując korektę Bonferronieo, odrzucamy globalną hipotezę zerową H_0, dla $n \to \infty$, gdy:

$$\max_{i \in \{1,\ldots,n\}} \left|\underline{X_i}\right| > \sqrt{2\ln(n)}.$$

Zauważmy, że wynik tak skonstruowanego testu nie zależy od poziomu istotności α dla dużych n. Niezależnie od ustalonego poziomu istotności α testu modelu (D1), próg odrzucenia

dla wartości: $\max_{i\in\{1,\ldots,n\}} \left|\underline{X_i}\right|$, będzie asymptotycznie równy $\sqrt{2\ln(n)}$. Potwierdza to fakt, że przy założeniu prawdziwości globalnej hipotezy zerowej:

$$\frac{\max\limits_{i\in\{1,\ldots,n\}} \left|\underline{X_i}\right|}{\sqrt{2\ln(n)}} \xrightarrow{D} 1.$$

5.1.2 Asymptotyczna moc korekty Bonferroniego

Sprawdźmy, jaką moc osiąga korekta Bonferroniego, w zależności od wielkości parametru średniej rozkładu normalnego, z którego pochodzą próby. Kolejne rozważania oprzyjmy o test modelu (D1) oraz założenia związane z tym modelem.

W celu wyznaczenia asymptotycznej mocy korekty Bonferroniego, bez straty ogólności, załóżmy, że $\mu_1 > 0$. Rozpatrzmy dwa przypadki:

1. Niech $\mu_1 = (1-\varepsilon)\sqrt{2\ln(n)}$, gdzie $\varepsilon \in (0,1)$. Wiedząc, że $\underline{X_1} = z+\mu_1$, gdzie $z \sim \mathcal{N}(0,1)$, obliczmy asymptotyczną moc korekty Bonferroniego:

$$\begin{aligned}
P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) &\leqslant P\left(p_1 \leqslant \frac{\alpha}{n} \mid H_1\right) + P\left(\min_{i\in\{2,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) \\
&\leqslant P\left(\underline{X_1} > \left|\Phi^{-1}(\alpha/n)\right| \mid H_1\right) + 1 - \left(1-\frac{\alpha}{n}\right)^{n-1} \\
&\leqslant P\left(z > \left|\Phi^{-1}(\alpha/n)\right| - \mu_1 \mid H_1\right) + 1 - \left(1-\frac{\alpha}{n}\right)^{n-1} \\
&= P\left(z > \varepsilon\sqrt{2\ln(n)} \mid H_1\right) + 1 - \left(1-\frac{\alpha}{n}\right)^{n-1} \\
&\xrightarrow{n\to\infty} 0 + 1 - e^{-\alpha} \approx \alpha.
\end{aligned}$$

2. Niech $\mu_1 = (1+\varepsilon)\sqrt{2\ln(n)}$. Moc korekty Bonferroniego w tym przypadku wynosi:

$$\begin{aligned}
P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) &\geqslant P\left(p_1 \leqslant \frac{\alpha}{n} \mid H_1\right) \\
&= P\left(\underline{X_1} > \left|\Phi^{-1}(\alpha/n)\right| \mid H_1\right) \\
&= P\left(z > \left|\Phi^{-1}(\alpha/n)\right| - \mu_1 \mid H_1\right) \\
&= P\left(z > -\varepsilon\sqrt{2\ln(n)} \mid H_1\right) \\
&\xrightarrow{n\to\infty} 1
\end{aligned}$$

Z powyższych rozważań wynika, że korekta Bonferroniego odrzuci globalną hipotezę zerową H_0, gdy parametr μ odstającej obserwacji będzie równy $(1+\varepsilon)\sqrt{2\ln(n)}$. W przypadku, kiedy $\mu = (1-\varepsilon)\sqrt{2\ln(n)}$, korekta Bonferroniego nie wykryje obserwacji. Wtedy test oparty na tej procedurze osiąga asymptotycznie moc równą α. Oznacza to, że korekta Bonferroniego nie jest dobrą metodą by przeprowadzać weryfikacje modeli takich jak model (D1), w przypadku, kiedy parametr średniej rozkładu obserwacji odstającej jest równy $(1-\varepsilon)\sqrt{2\ln(n)}$, gdzie $\varepsilon \in (0,1)$.

5.1.3 Asymptotyczna moc testu opartego o lemat Neymana-Pearsona

Określmy, jaką moc osiąga test oparty o lemat Neymana-Pearsona.W tym celu, musimy dokonać zmian w modelu (D1), tak by hipoteza alternatywna była hipotezą prostą. Test oprzyjmy o dane związane z modelem (D1). Skonstruujmy model:

$$\begin{aligned}
H_0 :&\quad \forall\ i \in \{1,\ldots,n\} : \mu_i = 0, \\
H_1 :&\quad \exists!\ i \sim \mathcal{U}\{1,\ldots,n\} : \mu_i = \widehat{\mu},
\end{aligned} \tag{D2}$$

gdzie $\widehat{\mu} = (1-\varepsilon)\sqrt{2\ln(n)}$. Ustalmy poziom istotności $\alpha \in (0,1)$. Oznaczmy $T_n := \sqrt{2\ln(n)}$. Wyznaczmy funkje wiarygodności w oparciu o powyższe hipotezy:

$$\begin{aligned}
L_0\left(0, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\underline{X_i}^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\underline{X_i}^2\right), \\
L_1\left(\widehat{\mu}, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\underline{X_i} - \widehat{\mu}\right)^2\right) \prod_{j:j\neq i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\underline{X_j}^2\right) \\
&= \frac{1}{n}\left(\frac{1}{\sqrt{2\pi}}\right)^n \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\underline{X_i} - \widehat{\mu}\right)^2\right) \prod_{j:j\neq i} \exp\left(-\frac{1}{2}\underline{X_j}^2\right).
\end{aligned}$$

Wyznaczmy iloraz wiarygodności:

$$r\left(H_0, H_1\right) = \frac{L_1\left(\widehat{\mu}, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right)}{L_1\left(\widehat{\mu}, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right)} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(\underline{X_i}\widehat{\mu} - \frac{1}{2}\widehat{\mu}^2\right).$$

Oznaczmy powyższy iloraz jako:

$$R = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

gdzie $Y_i = \exp\left(\underline{X_i}\widehat{\mu} - \frac{1}{2}\widehat{\mu}^2\right)$. Prawdziwe jest stwierdzenie, że jeśli zmienna losowa: z, pochodząca z rozkładu $\mathcal{N}(\mu_z, \sigma_z^2)$, to e^z jest zmienną losową, pochodzącą z rozkładu logarytmicznie normalnego: $\mathcal{LN}(\mu_z, \sigma_z)$. Z tego wynika, że $Y_i \sim \mathcal{LN}\left(-\frac{1}{2}\widehat{\mu}^2, \widehat{\mu}^2\right)$.

W celu wyznaczenia asymptotycznej mocy rozważanego testu, nie możemy skorzystać z Centralnego Twierdzenia Granicznego, ponieważ $\widehat{\mu}$ zależy od n. Musimy wykazać, że $R \xrightarrow{D} 1$, gdy hipoteza H_0 jest prawdziwa. Przyjmijmy zmienną:

$$R' = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{\underline{X_i} \leqslant T_n\}}.$$

Pawdopodobieństwo, że $R \neq R'$ wynosi:

$$\begin{aligned}
P\left(R \neq R'\right) &\leqslant P\left(\max_{i\in\{1,\ldots,n\}} \underline{X_i} \geqslant T_n\right) \\
&\leqslant \sum_{i=1}^{n} P\left(\underline{X_i} \geqslant T_n\right) \\
&\leqslant \sum_{i=1}^{n} \frac{1}{T_n} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}T_n^2\right) \\
&\leqslant \sum_{i=1}^{n} \frac{1}{T_n} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sqrt{2\ln(n)}^2\right) \\
&\leqslant n \cdot \frac{1}{T_n} \cdot \frac{\frac{1}{n}}{\sqrt{2\pi}} \\
&\xrightarrow{n\to\infty} 0.
\end{aligned}$$

Obliczmy wartość oczekiwaną oraz variancję zmiennej R':

$$\begin{aligned}
\mathbb{E}_{H_0}[R'] = \frac{1}{n}\mathbb{E}_{H_0}\left[\sum_{i=1}^{n} Y_i \mathbb{1}_{\{X_i \leqslant T_n\}}\right] &= \mathbb{E}_{H_0}\left[Y_1 \mathbb{1}_{\{X_1 \leqslant T_n\}}\right] \\
&= \int_{-\infty}^{T_n} \exp\left(z\widehat{\mu} - \frac{1}{2}\widehat{\mu}^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}z \\
&= \int_{-\infty}^{T_n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z - \widehat{\mu}\right)^2\right) \mathrm{d}z \\
&= \int_{-\infty}^{T_n - \widehat{\mu}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}z \\
&= \int_{-\infty}^{\varepsilon\sqrt{2\ln(n)}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}z \\
&= \Phi\left(\varepsilon\sqrt{2\ln(n)}\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}_{H_0}[R'] = \frac{1}{n^2}\mathrm{Var}_{H_0}\left[\sum_{i=1}^{n} Y_i \mathbb{1}_{\{X_i \leqslant T_n\}}\right] &= \frac{1}{n}\mathrm{Var}_{H_0}\left[Y_1 \mathbb{1}_{\{X_1 \leqslant T_n\}}\right] \\
&\leqslant \frac{1}{n}\mathbb{E}_{H_0}\left[Y_1^2 \mathbb{1}_{\{X_1 \leqslant T_n\}}\right] \\
&= \frac{1}{n}\int_{-\infty}^{T_n} \exp\left(2z\widehat{\mu} - \widehat{\mu}^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}z \\
&= \frac{1}{n}e^{\widehat{\mu}^2} \int_{-\infty}^{T_n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z - 2\widehat{\mu}\right)^2\right) \mathrm{d}z \\
&= \frac{1}{n}e^{\widehat{\mu}^2} \int_{-\infty}^{T_n - 2\widehat{\mu}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z\right)^2\right) \mathrm{d}z \\
&= \frac{1}{n}e^{\widehat{\mu}^2}\Phi\left(T_n - 2\widehat{\mu}\right).
\end{aligned}$$

Przyjmijmy, że $T_n - 2\widehat{\mu} < 0$, czyli $\varepsilon \in \left(0, \frac{1}{2}\right)$. Wtedy:

$$\begin{aligned}
\frac{1}{n}e^{\widehat{\mu}^2}\Phi\left(T_n - 2\widehat{\mu}\right) = \frac{1}{n}e^{\widehat{\mu}^2}\left(1 - \Phi\left(-T_n + 2\widehat{\mu}\right)\right) &\leqslant \frac{1}{n}e^{\widehat{\mu}^2} \cdot \frac{\phi\left(2\widehat{\mu} - T_n\right)}{T_n - 2\widehat{\mu}} \\
&= \frac{1}{n} \cdot \frac{\exp\left(\widehat{\mu}^2 - \frac{1}{2}\left(2\widehat{\mu} - T_n\right)^2\right)}{\left(2\widehat{\mu} - T_n\right)\sqrt{2\pi}} \\
&= \frac{1}{n} \cdot \frac{\exp\left(-2\varepsilon^2\ln(n) + \ln(n)\right)}{\left(2\widehat{\mu} - T_n\right)\sqrt{2\pi}} \\
&= \frac{n^{-2\varepsilon^2}}{\left(2\widehat{\mu} - T_n\right)\sqrt{2\pi}} \\
&\xrightarrow{n\to\infty} 0
\end{aligned}$$

Kiedy $T_n - 2\widehat{\mu} > 0$, czyli $\varepsilon \in \left(\frac{1}{2}, 1\right)$, wtedy:

$$\frac{1}{n} e^{\widehat{\mu}^2} = \frac{1}{n} \exp\left((1-\varepsilon)^2 \cdot 2\ln(n)\right) = \frac{1}{n} n^{2(1-\varepsilon)^2} \xrightarrow{n\to\infty} 0.$$

Z powyższego wynika, że:

$$\begin{aligned} \mathbb{E}_{H_0}[R'] &= \Phi\left(\varepsilon\sqrt{2\ln(n)}\right), \\ \mathrm{Var}_{H_0}[R'] &= o(1). \end{aligned}$$

Z nierówności Czebyszewa następuje:

$$R' = \Phi\left(\varepsilon\sqrt{2\ln(n)}\right) + o(1).$$

Dowodzi to, że:

$$R \xrightarrow{D} 1.$$

Z tego wynika, że $L_0\left(0, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right) \approx L_1\left(\widehat{\mu}, \left(\underline{X_1}, \ldots, \underline{X_n}\right)\right)$, przy $n \to \infty$. Oznacza to, że H_0 i H_1 są takie same.

Oznaczmy zmienną G, taką, że $\lim\limits_{n\to\infty} P(R \geqslant G \mid H_0) = \alpha$. Obliczmy asymptotyczną moc testu opartego o lemat Neymana-Pearsona:

$$\begin{aligned} P(R > G \mid H_1) &= 1 - P(R \leqslant G \mid H_1) \\ &= 1 - \int \mathbb{1}_{\{R\leqslant G\}} \,\mathrm{d}P_{H_1} \\ &= \left|R = \frac{\mathrm{d}P_{H_1}}{\mathrm{d}P_{H_0}}\right| = 1 - \int R\mathbb{1}_{\{R\leqslant G\}} \,\mathrm{d}P_{H_0} \\ &= 1 - \int \mathbb{1}_{\{R\leqslant G\}} \,\mathrm{d}P_{H_0} - \int (R-1)\mathbb{1}_{\{R\leqslant G\}} \,\mathrm{d}P_{H_0} \\ &\xrightarrow{n\to\infty} = 1 - (1-\alpha) - 0 = \alpha. \end{aligned}$$

Test modelu typu „igła w stogu siana", oparty o lemat Neymana-Pearsona osiąga asymptotycznie moc równą α. Z faktu, że lemat Neymana-Pearsona pozwala nam skonstruować test jednostajnie najmocniejszy na poziomie istotność α, wynika, że nie możemy wyznaczyć testu o mocy większej niż α w przypadku kiedy, $\widehat{\mu} = (1-\varepsilon)\sqrt{2\ln(n)}$. Konsekwencją tego jest fakt, że jeżeli dysponujemy zbiorem danych pochodzących z rozkładu $\mathcal{N}(0,1)$ i chcemy sprawdzić, czy w zbiorze danych znajduje się obserwacja o rozkładzie $\mathcal{N}(\mu, 1)$, to wykryjemy ją tylko wtedy, kiedy $\mu = (1+\varepsilon)\sqrt{2\ln(n)}$.

5.2 Model oparty o dane pochodzące z rozkładu Poissona

Postarajmy się przeprowadzić weryfikację modelu typu „igła w stogu siana", opartego o dane pochodzące z rozkładu Poissona. Niech zbiór danych, o który oprzemy późniejsze rozważania, składa się z niezależnych zmiennych losowych: $\underline{X_1}, \ldots, \underline{X_n}$, gdzie $n \geqslant 2$. Przyjmijmy, że $\underline{X_i}$, dla $i \in \{1, \ldots, n\}$, pochodzi z rozkładu Poissona o parametrze λ_i oraz, że statystyka testowa:

$$T\left(\underline{X_i}\right) = \underline{X_i},$$

odpowiada i-tej hipotezie zerowej: $H_{0_i} : \lambda_i = \widetilde{\lambda}$, gdzie $\widetilde{\lambda} \in (0, \infty)$, jest ustaloną wartością. Wiemy, że przy założeniu prawdziwości hipotezy H_{0_i}, $T\left(\underline{X_i}\right) \sim Poisson\left(\widetilde{\lambda}\right)$. Wyznaczmy model typu „igła w stogu siana" dla tego przypadku:

$$\begin{aligned} H_0 :&\quad \forall\, i \in \{1, \ldots, n\} : \lambda_i = \widetilde{\lambda}, \\ H_1 :&\quad \exists!\, i \sim \mathcal{U}\{1, \ldots, n\} : \lambda_i > \widetilde{\lambda}. \end{aligned} \tag{E1}$$

Odrzucamy globalną hipotezę zerową H_0, na usalonym poziomie istotności α, przy pomocy korekty Bonferroniego, gdy:

$$\max_{i\in\{1,\ldots,n\}} T\left(\underline{X_i}\right) \geqslant k,$$

gdzie wartość k definiujemy jako:

$$k := \min\left\{m : \quad \sum_{i=m}^{\infty} \frac{\widetilde{\lambda}^i e^{-\widetilde{\lambda}}}{i!} \leqslant \frac{\alpha}{n}\right\}.$$

Ponieważ rozkład Poissona jest rozkładem dyskretnym, określenie wartości k w taki sposób pozwala nam na znalezienie kwantyla rzędu $1-\frac{\alpha}{n}$, lub wyższego, rozkładu $Poisson\left(\widetilde{\lambda}\right)$. W skutek tego, kontrolujemy prawdopodobieństwo popelnienia błędu I rodzaju na ustalonym poziomie α.

5.2.1 Asymptotyczna wartość kwantyla

Aby rozpocząć rozważania na temat efektywności korekty Bonferroniego oraz testu opartego o lemat Neymana-Pearsona, dla wyżej skonstruowanego modelu typu „igła w stogu siana", musimy wyznaczyć optymalny próg wykrycia. Wiemy, że dystrybuanta rozkładu Poissona o parametrze λ jest określona i ma postać:

$$\mathbb{F}_\lambda(T') = P(t \leqslant T') = \sum_{i=0}^{T'} \frac{\lambda^i e^{-\lambda}}{i!},$$

dla określonej liczby $T' \in \{0,1,2,\ldots\}$, oraz zmiennej losowej $t \sim Poisson(\lambda)$. Z tego wynika, że:

$$1-\mathbb{F}_\lambda(T') = P(t > T') = \sum_{i=T'+1}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!}.$$

Znajomość dystrybuanty rozkładu Poissona pozwala nam, w stosunku do rozkładu standardowego normalnego, szybciej i łatwiej znaleźć przybliżenie kwantyla rzędu $1-\frac{\alpha}{n}$ rozkładu $Poisson\left(\widetilde{\lambda}\right)$, gdzie $\alpha \in (0,1)$, $n \in \mathbb{N}_+$. Niech $T' \in \{0,1,2,\ldots\}$ i $t \sim Poisson\left(\widetilde{\lambda}\right)$, $T = T'+1$. Przy pomocy wzoru Stirlinga wyznaczmy:

$$\begin{aligned}
1-\mathbb{F}_{\widetilde{\lambda}}(T') = P(t \geqslant T) &= \sum_{i=T}^{\infty} \frac{\widetilde{\lambda}^i e^{-\widetilde{\lambda}}}{i!} \\
&= \frac{\widetilde{\lambda}^T e^{-\widetilde{\lambda}}}{T!}\left(1+\frac{\widetilde{\lambda}}{T+1}+\frac{\widetilde{\lambda}^2}{(T+1)(T+2)}+\ldots\right) \\
&\approx \widetilde{\lambda}^T e^{-\widetilde{\lambda}}\frac{e^T}{T^T\sqrt{2\pi T}}\left(1+\frac{\widetilde{\lambda}}{T+1}+\frac{\widetilde{\lambda}^2}{(T+1)(T+2)}+\ldots\right).
\end{aligned}$$

Oznaczmy $A_{T,\widetilde{\lambda}} = \left(1+\frac{\widetilde{\lambda}}{T+1}+\frac{\widetilde{\lambda}^2}{(T+1)(T+2)}+\ldots\right)$. Oszacujmy $A_{T,\widetilde{\lambda}}$, wiedząc, że:

$$1 \leqslant A_{T,\widetilde{\lambda}} \leqslant \sum_{i=0}^{\infty} \frac{\widetilde{\lambda}^i}{i!}.$$

Z tego wynika, że $A_{T,\widetilde{\lambda}} \in [1, e^{\widetilde{\lambda}}]$. Dla $n \to \infty$, wyznaczmy przybliżenie wartości T:

$$\begin{aligned} P(t \geqslant T) &\approx \frac{\alpha}{n} \\ \widetilde{\lambda}^T e^{-\widetilde{\lambda}} \frac{e^T}{T^T \sqrt{2\pi T}} A_{T,\widetilde{\lambda}} &\approx \frac{\alpha}{n} \qquad \backslash\backslash \ln(\cdot) \\ T \ln\left(\widetilde{\lambda}\right) - \widetilde{\lambda} + T - \left(T + \frac{1}{2}\right) \ln(T) - \frac{1}{2} \ln(2\pi) + \ln(A_{T,\widetilde{\lambda}}) &\approx \ln(\alpha) - \ln(n), \end{aligned}$$

dla $n \to \infty$, jest to równoważne z:

$$\begin{aligned} T \ln\left(\widetilde{\lambda}\right) + T - \left(T + \frac{1}{2}\right) \ln(T) &\approx -\ln(n) \\ T \left(\ln\left(\widetilde{\lambda}\right) + 1\right) - \left(T + \frac{1}{2}\right) \ln(T) &\approx -\ln(n). \end{aligned}$$

Niech:

$$T \approx b_n \ln(n).$$

Z powyższego wynika, że:

$$\begin{aligned} -b_n \ln(n) \left(\ln\left(\widetilde{\lambda}\right) + 1\right) + \left(b_n \ln(n) + \frac{1}{2}\right) \ln\left(b_n \ln(n)\right) &\approx \ln(n) \\ -b_n \ln\left(\widetilde{\lambda}\right) - b_n + \left(b_n + \frac{1}{2\ln(n)}\right) \left(\ln(b_n) + \ln\left(\ln(n)\right)\right) &\approx 1 \\ -b_n \ln\left(\widetilde{\lambda}\right) - b_n + b_n \ln(b_n) + b_n \ln\left(\ln(n)\right) + \frac{\ln(b_n)}{2\ln(n)} + \frac{\ln\left(\ln(n)\right)}{2\ln(n)} &\approx 1 \\ b_n \left(-\ln\left(\widetilde{\lambda}\right) + \ln\left(\ln(n)\right)\right) - b_n + b_n \ln(b_n) + \frac{\ln(b_n)}{2\ln(n)} + \frac{\ln\left(\ln(n)\right)}{2\ln(n)} &\approx 1. \end{aligned}$$

Posługując się następującym faktem z analizy matematycznej:

$$d_n \approx e_n \iff \frac{d_n}{e_n} \xrightarrow{n \to \infty} 1,$$

Zdefiniujmy:

$$b_n := \frac{1}{\ln\left(\ln(n)\right)}.$$

W związku z czym:

$$\begin{aligned} \frac{-\ln\left(\widetilde{\lambda}\right) + \ln\left(\ln(n)\right)}{\ln\left(\ln(n)\right)} - \frac{1}{\ln\left(\ln(n)\right)} - \frac{\ln\left(\ln\left(\ln(n)\right)\right)}{\ln\left(\ln(n)\right)} \\ - \frac{\ln\left(\ln\left(\ln(n)\right)\right)}{2\ln(n)} + \frac{\ln\left(\ln(n)\right)}{2\ln(n)} \approx 1. \end{aligned}$$

Przy czym:

$$\begin{aligned}
\frac{-\ln\left(\widetilde{\lambda}\right)+\ln\left(\ln(n)\right)}{\ln\left(\ln(n)\right)} &\xrightarrow{n\to\infty} 1,\\
-\frac{1}{\ln\left(\ln(n)\right)} &\xrightarrow{n\to\infty} 0,\\
-\frac{\ln\left(\ln\left(\ln(n)\right)\right)}{\ln\left(\ln(n)\right)} &\xrightarrow{n\to\infty} 0,\\
-\frac{\ln\left(\ln\left(\ln(n)\right)\right)}{2\ln(n)} &\xrightarrow{n\to\infty} 0,\\
\frac{\ln\left(\ln(n)\right)}{2\ln(n)} &\xrightarrow{n\to\infty} 0.
\end{aligned}$$

Z powyższych rozważań wynika, że przybliżenie wartości T jest postaci:

$$T \approx \frac{\ln(n)}{\ln\left(\ln(n)\right)}.$$

W porównaniu do progu wykrycia z poprzedniego podrozdziału, również w tym przypadku przybliżenie kwantyla rzędu $1-\frac{\alpha}{n}$, dla dużych n, nie zależy od wartości ustalonego wcześniej poziomu istotności α oraz parametru $\widetilde{\lambda}$.

5.2.2 Asymptotyczna moc korekty Bonferroniego

Znając aproksymację kwantyla rzędu $1-\frac{\alpha}{n}$ rozkładu Poissona o parametrze: $\lambda=\widetilde{\lambda}$, wyznaczmy asymptotyczną moc korekty Bonferroniego w oparciu o weryfikację modelu typu „igła w stogu siana". Zanim do tego przejdziemy, sprawdźmy, czy korekta Bonferroniego kontroluje ryzyko błędu I rodzaju na ustalonym poziomie istotności $\alpha\in(0,1)$. Kolejne rozważania oprzyjmy o model (E1). Przyjmując zmienną losową $t\sim Poisson\left(\widetilde{\lambda}\right)$, wyznaczmy p-wartość odpowiadającą statystyce $T\left(\underline{X_i}\right)$:

$$p_i = 1-P\left(t\leqslant T\left(\underline{X_i}\right)\right) = 1-\sum_{i=0}^{\lfloor T\left(\underline{X_i}\right)\rfloor}\frac{\widetilde{\lambda}^{-i}e^{-\widetilde{\lambda}}}{i!}.$$

Z rozważań przeprowadzonych w rozdziale trzecim wynika, że p_i przyjmuje asymptotycznie rozkład $\mathcal{U}[0,1]$. Oznaczmy:

$$\begin{aligned}
k &:= \max\left\{m:\quad \sum_{i=m+1}^{\infty}\frac{\widetilde{\lambda}^{i}e^{-\widetilde{\lambda}}}{i!}>\frac{\alpha}{n}\right\},\\
T_n &:= \frac{\ln(n)}{\ln\left(\ln(n)\right)},
\end{aligned}$$

takie, że:

$$k\approx T_n.$$

Rozpatrzmy i udowodnijmy poniższe twierdzenie.

Twierdzenie 5.2. *Korekta Bonferroniego kontroluje prawdopodobieństwo popełnienia błędu I rodzaju na poziomie $\alpha\in(0,1)$, dla modelu typu „igła w stogu siana", opartego o dane pochodzące z rozkładu Poissona.*

Dowód. Następujący dowód oprzyjmy o model (E1) oraz dane z nim związane. Przyjmijmy poziom istotności $\alpha \in (0,1)$. Niech statystyka testowa: $T\left(\underline{X_i}\right) \sim Poisson\left(\widetilde{\lambda}\right)$, gdzie $i \in \{1,\ldots,n\}$. Prawdopodobieństwo popełnienia błędu I rodzaju wynosi:

$$\begin{aligned} P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_0\right) &= 1 - P\left(\bigcap_{i=1}^{n}\left\{p_i \geqslant \frac{\alpha}{n}\right\} \mid H_0\right) \\ &= 1 - \prod_{i=1}^{n} P\left(p_i \geqslant \frac{\alpha}{n} \mid H_0\right). \end{aligned}$$

Przy czym:

$$\prod_{i=1}^{n} P\left(p_i \geqslant \frac{\alpha}{n} \mid H_0\right) = \prod_{i=1}^{n} P\left(T\left(\underline{X_i}\right) < k \mid H_0\right) = \prod_{i=1}^{n} \mathbb{F}_{\widetilde{\lambda}}(k) = \left(\mathbb{F}_{\widetilde{\lambda}}(k)\right)^n,$$

gdzie:

$$\begin{aligned} 1 - \frac{\alpha}{n} \leqslant \quad \mathbb{F}_{\widetilde{\lambda}}(k) \quad &\leqslant 1 - \frac{\alpha}{n} + \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!} \\ \left(1 - \frac{\alpha}{n}\right)^n \leqslant \left(\mathbb{F}_{\widetilde{\lambda}}(k)\right)^n &\leqslant \left(1 - \frac{\alpha}{n} + \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!}\right)^n. \end{aligned}$$

Przy $n \to \infty$:

$$\exp(-\alpha) \leqslant \left(\mathbb{F}_{\widetilde{\lambda}}(k)\right)^n \leqslant \exp\left(n\left(-\frac{\alpha}{n} + \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!}\right)\right),$$

gdzie:

$$\exp\left(n\left(-\frac{\alpha}{n} + \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!}\right)\right) = \exp(-\alpha)\exp\left(\frac{n\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!}\right) \xrightarrow{n\to\infty} \infty,$$

ponieważ:

$$\begin{aligned} \frac{n\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!} &\xrightarrow{n\to\infty} n\widetilde{\lambda}^k e^{-\widetilde{\lambda}} e^k k^{-k} \left(2\pi k\right)^{-\frac{1}{2}} \\ &\xrightarrow{n\to\infty} \frac{e^{-\widetilde{\lambda}}}{\sqrt{2\pi}} \exp\left(\ln(n) + k\ln(\widetilde{\lambda}) + k - k\ln(k) - \frac{1}{2}\ln(k)\right) \\ &\xrightarrow{n\to\infty} \frac{e^{-\widetilde{\lambda}}}{\sqrt{2\pi}} \exp\left(\ln(n) + T_n\ln(\widetilde{\lambda}) + T_n - T_n\ln(T_n) - \frac{1}{2}\ln(T_n)\right) \\ &\xrightarrow{n\to\infty} \frac{e^{-\widetilde{\lambda}}}{\sqrt{2\pi}} \exp\left(-T_n\left(-\frac{\ln(n)}{T_n} - \ln(\widetilde{\lambda}) - 1 + \ln(T_n) + \frac{\ln(T_n)}{2T_n}\right)\right) \\ &\xrightarrow{n\to\infty} \frac{e^{-\widetilde{\lambda}}}{\sqrt{2\pi}} \exp\left(\frac{\ln(n)}{\ln(\ln(n))}\ln(\ln(\ln(n)))\right) \xrightarrow{n\to\infty} \infty. \end{aligned}$$

Z powyższego wynika, że, przy $n \to \infty$:

$$\begin{aligned} P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_0\right) &= 1 - \prod_{i=1}^{n} P\left(T\left(\underline{X_i}\right) < k \mid H_0\right) \\ &\leqslant 1 - \left(1 - \frac{\alpha}{n}\right)^n \approx 1 - e^{-\alpha} \\ &\leqslant \alpha. \end{aligned}$$

■

Uwaga 5.1. *Porównując powyższe modele, ryzyko popełnienia błędu I rodzaju, dla modelu opartego o dane pochodzące z rozkładu ciągłego, równe jest* $1-e^{-\alpha}$*, natomiast, w przypadku modelu opartego o dane pochodzące z rozkładu dyskretnego, nie możemy określić dokładnej wartości prawdopodobieństwa popełnienia błędu I rodzaju (wiemy, że jest ono mniejsze lub równe* $1-e^{-\alpha}$*).*

Wykazaliśmy, że dla modelu (E1), korekta Bonferroniego także kontroluje błąd I rodzaju. Postarajmy się wyznaczyć asymptotyczną moc rozważanej procedury. Ustalmy poziom istotności $\alpha \in (0,1)$. Załóżmy, bez straty ogólności rozważań, że $\lambda_1 > \widetilde{\lambda}$. Tak jak w poprzednim podrozdziale, rozważmy dwa przypadki:

1. Niech $\lambda_1 = (1-\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$, gdzie $\varepsilon \in (0,1)$. Wyzanczmy asymptotyczną moc korekty Bonferroniego, w oparciu o powyższy model:

$$\begin{aligned} P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) &\leqslant P\left(p_1 \leqslant \frac{\alpha}{n} \mid H_1\right) + P\left(\min_{i\in\{2,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) \\ &= P\left(\underline{X_1} > k \mid H_1\right) + 1 - \prod_{i=2}^{n} P\left(p_i \geqslant \frac{\alpha}{n} \mid H_1\right) \\ &\leqslant P\left(\underline{X_1} > T_n \mid H_1\right) + 1 - \left(1-\frac{\alpha}{n}\right)^{n-1}. \end{aligned}$$

   Przy czym:

$$\begin{aligned} P\left(\underline{X_1} > k \mid H_1\right) &= P\left(\underline{X_1} > T_n \mid H_1\right) \\ &= P\left(\frac{\underline{X_1}-\lambda_1}{\sqrt{\lambda_1}} > \frac{T_n-\lambda_1}{\sqrt{\lambda_1}} \mid H_1\right) \\ &= P\left(\frac{\underline{X_1}-\lambda_1}{\sqrt{\lambda_1}} > \frac{\varepsilon}{\sqrt{1-\varepsilon}}\sqrt{T_n} \mid H_1\right). \end{aligned}$$

   W oparciu o fakt, że rozkład $\mathcal{N}(\lambda,\lambda)$ jest doskonałym przybliżeniem rozkładu $Poisson(\lambda)$, gdy $\lambda \to \infty$, wynika, że:

$$\frac{\underline{X_1}-\lambda_1}{\sqrt{\lambda_1}} \xrightarrow{D} \mathcal{N}(0,1).$$

   W związku z:

$$\frac{\varepsilon}{\sqrt{1-\varepsilon}}\sqrt{T_n} \xrightarrow{n\to\infty} \infty,$$

   dochodzimy do:

$$P\left(\frac{\underline{X_1}-\lambda_1}{\sqrt{\lambda_1}} > \frac{\varepsilon}{\sqrt{1-\varepsilon}}\sqrt{T_n} \mid H_1\right) \xrightarrow{n\to\infty} 0.$$

   Wobec powyższego, asymptotyczna moc korekty Bonferroniego wynosi w tym przypadku:

$$P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) \leqslant 0 + 1 - e^{-\alpha} \leqslant \alpha.$$

2. Niech $\lambda_1 = (1+\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$, gdzie $\varepsilon \in (0,1)$. Wyzanczmy asymptotyczną moc korekty

Bonferroniego:

$$\begin{aligned} P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) &\geqslant P(p_1 \geqslant \frac{\alpha}{n} \mid H_1) \\ &= P\left(\underline{X_1} > k \mid H_1\right) \\ &= P\left(\underline{X_1} > T_n \mid H_1\right) \\ &= P\left(\frac{\underline{X_1} - \lambda_1}{\sqrt{\lambda_1}} > \frac{T_n - \lambda_1}{\sqrt{\lambda_1}} \mid H_1\right) \\ &= P\left(\frac{\underline{X_1} - \lambda_1}{\sqrt{\lambda_1}} > \frac{-\varepsilon}{\sqrt{1+\varepsilon}}\sqrt{T_n} \mid H_1\right). \end{aligned}$$

Na mocy faktów:

$$\begin{aligned} \frac{\underline{X_1} - \lambda_1}{\sqrt{\lambda_1}} &\xrightarrow{D} \mathcal{N}(0,1), \\ \frac{-\varepsilon}{\sqrt{1+\varepsilon}}\sqrt{T_n} &\xrightarrow{n\to\infty} -\infty, \end{aligned}$$

asymptotyczna moc procedury wynosi:

$$P\left(\min_{i\in\{1,\ldots,n\}} p_i \leqslant \frac{\alpha}{n} \mid H_1\right) \geqslant P\left(\frac{\underline{X_1} - \lambda_1}{\sqrt{\lambda_1}} > \frac{-\varepsilon}{\sqrt{1+\varepsilon}}\sqrt{T_n} \mid H_1\right) \xrightarrow{n\to\infty} 1.$$

Z powyższych podpunktów wynika, że korekta Bonfferoniego wykryje obserwację pochodzącą z rozkładu $Poisson(\lambda)$ pośród obserwacji z rozkładu $Poisson\,(\lambda')$, jeśli λ będzie równa $(1+\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$. W przypadku, kiedy $\lambda = (1-\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$, test oparty o rozważaną procedurę osiąga asymptotycznie moc mniejszą niż α. W rezultacie, ryzyko popełnienia błędu II rodzaju jest zbyt duże, żeby test uznać za skuteczny. Dla modelów typu „igła w stogu siana", opartych na danych pochodzących z rozkładu Poissona, korekta Bonferroniego nie jest dobrą procedurą weryfikacji. Przyczyną tego stwierdzenia jest fakt, że nie zawsze procedura ta osiąga wysoką moc.

6 Wnioski

Podsumujmy wszytskie rozważania jakie przeprowadziliśmy we wcześniejszych rozdziałach. Podstawy teorii testowania hipotez statystycznych pomogły nam zrozumieć ideę tworzenia testów, oraz nauczyły nas ich konstrukcji. Dzięki przeprowadzonym testom udowodniliśmy, że p-wartość pochodzi z rozkładu $\mathcal{U}[0,1]$, kiedy statystyka testowa pochodzi z rozkładu ciągłego. Wykazaliśmy również, że jeśli statystyka testowa pochodzi z rozkładu Poissona, to jej przekształcenie osiąga asymptotycznie rozkład jednostajny na przedziale $[0,1]$. Rozważania na temat wartości p pozwoliły nam wykazać, że korekta Bonferroniego kontroluje błąd polegający na odrzuceniu globalnej hipotezy zerowej w sytuacji, kiedy jest ona prawdziwa.

W konsekwencji, skonstruowaliśmy modele typu „igła w stogu siana" dla danych pochodzących z rozkładów normalnego oraz Poissona. Dla rozkładu normalnego wyznaczyliśmy asymptotyczną wartość kwantyla rzędu $1-\frac{\alpha}{n}$ (gdzie: $\alpha \in (0,1)$, $n \to \infty$), czyli asymptotyczny próg wykrycia, równą $\sqrt{2\ln(n)}$. Wykazaliśmy, że korekta Bonferroniego osiąga asymptotycznie moc równą 1, w sytuacji, gdy parametr średniej, obserwacji różniącej się rozkładem, jest równy: $\mu = (1+\varepsilon)\sqrt{2\ln(n)}$, gdzie $\varepsilon \in (0,1)$. W przypadku kiedy ten sam parametr jest równy $(1-\varepsilon)\sqrt{2\ln(n)}$, pokazaliśmy, że korekta Bonferroniego osiąga moc bliską α. Za pomocą lematu Neymana-Pearsona pokazaliśmy, że nie możemy wyznaczyć testu o mocy większej niż α, gdy hipoteza alternatywna zakłada, że pośród $n-1$ obserwacji o rozkładzie $\mathcal{N}(0,1)$ znajduje się obserwacja o rozkładzie $\mathcal{N}((1-\varepsilon)\sqrt{2\ln(n)},1)$, gdzie $\varepsilon \in (0,1)$.

Dla modelu opartego o n obserwacji z rozkładu Poissona, wyznaczyliśmy próg wykrycia równy $\frac{\ln(n)}{\ln(\ln(n))}$. Przeprowadziliśmy test modelu i wykazaliśmy, że, tak jak w poprzednim przypadku, korekta Bonferroniego przyjmuje asymptotycznie moc równą 1, kiedy parametr λ rozkładu odstającej obserwacji jest równy $(1+\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}$. Podobnie wykazaliśmy, że moc korekty Bonferoniego jest mniejsza od poziomu istotności α, gdy obserwacja różniąca się rozkładem, pochodzi z rozkładu $Poisson\left((1-\varepsilon)\frac{\ln(n)}{\ln(\ln(n))}\right)$.

Literatura

[1] R. V. Hogg, J. W. McKean, A. T. Craig. *Introduction to Mathematical Statistics (6th Edition)*. Pearson Education, 2005.

[2] N. L. Johnson, A. W. Kemp, S. Kotz. *Univariate Discrete Distributions(Third Edition)*. John Wiley & Sons, Inc. Hoboken, New Jersey. 2005.

[3] A. Futschik, T. Taus, S. Zehetmayer. *An omnibus test for the global null hypothesis*. `https://journals.sagepub.com/doi/pdf/10.1177/0962280218768326`. Dostęp: 30.01.2023.

[4] E. Candes. *STATS 300C: Theory of Statistics, Spring 2022. Lecture 2*. 31.03.2022. `https://candes.su.domains/teaching/stats300c/Lectures/Lecture02.pdf`. Dostęp: 30.01.2023.