Statystyczna analiza danych (molekularnych) – podstawowe testy statystyczne

Anna Gambin

18 marca 2012

Spis treści

1 Testowanie hipotez statystycznych

Testowanie hipotez pozwala nam z wykorzystaniem narzędzi statystyki testować hipotezy dotyczące analizowanych danych molekularnych. Możemy np szukać odpowiedzi na pytania w rodzaju:

1. Czy średnia ekspresja danego genu u pacjentów cierpiących na białaczkę typu AML jest różna od średniej ekspresji tego samego genu u grupy chorych na białaczkę typy ALL ?

2. Czy średnia ekspresja badanego genu jest niezerowa ?

3. Do jakiego stopnia ma rozkład normalny ?

4. Czy w badanej próbce znajdują się elementy odstające (*ang. outliers*)?

5. W jaki sposób wybrać cechy najlepiej dyskryminujące dwie badane populacje ?

6. Jal zbadać, czy częstości występowania danego motywu w różnych sekwencjach DNA jest taka sama?

Klasyczne (nie bayesowskie) podejście do testowania hipotez sprowadza się do następujących kroków:

1. Sformułuj **hipotezę zerową** H_0 oraz **hipotezę alternatywną** H_1. Najczęściej hipoteza zerowa odpowiada sytuacji nieciekawej, średniej, nie wyróżniającej się cechy. Natomiast odrzucenie hipotezy zerowej, równoznaczne przyjęciu hipotezy alternatywnej sugeruje, że rozważana cecha w istotny sposób dyskryminuje dwie populacje. Ponieważ decyzję o przyjęciu lub odrzucecniu hipotezy podejmujemy na podstawie danych, które traktujemy jako próbę losową, czyli realizację pewnego procesu losowego mamy niezerową szansę pomyłki. Odrzucenie poprawnej hipotezy zerowej określamy jako **błąd typu I**, natomiast przyjęcie fałszywej hipotezy zerowej nazywamy **błędem typu II**. Łatwo dostrzec, że obydwie te wielkości są ze sobą powiązane i dlatego musimy wybrać, którą z nich chcemy kontrolować. W zastosowaniach najczęściej konsekwencję różnych typów błędów są niesymetryczne i zazwyczaj przyjmuje się, że interesuje nas utrzymanie błędów typu I na odpowiednio niskim poziomie α (np. $\alpha = 1\%$ lub $\alpha = 5\%$).

2. Ustal poziom α dla błędów typu I.

3. Sformułuj odpowienią statystykę testową, czyli obliczaną na podstawie danych wielkość, której wartość będzie odpowiadała za przyjęcie bądź odrzucenie hipotezy zerowej. Jest to bardzo ważny krok i łatwo się zgodzić, że wybór kiepskiej statystyki zaważy na jakości testu.

4. Określ, które wartości statystyki testowej prowadzą do odrzucenia hipotezy zerowej. Wybór tych wartości jest taki, aby kontrolować poziom błędów α założony w kroku 2. W tym celu przydatne okazuje się pojęcie **p-wartości** (*ang. p-value*). Dla danej wartości statystyki testowej p-wartość jest zdefiniowana, jako prawdopodobieństwo uzyskania tej lub bardziej ekstremalnej wartości przy założeniu hipotezy zerowej. Jeśli tak policzona p-wartość jest mniejsza niż zakładany poziom błędów α, hipoteza zerowa zostaje odrzucona.

5. W ostatnim kroku analizujemy dostępne dane i sprawdzamy, czy wartość statystyki testowej odpowiada p-wartości pozwalającej nam odrzucić hipotezę zerową.

Omówimy teraz kilka najbardziej popularnych testów statystycznych stosowanych w bioinformatyce.

Z test

Zakładamy, że badane przez nas obserwacje $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego o nieznanej średniej i znanej wariancji σ. Jako hipotezę zerową przyjmujemy, że średnia wynosi μ_0, natomiast hipoteza alternatywna H_1 stwierdza, że $\mu \neq \mu_0$ (hipoteza złożona) lub $\mu > \mu_0$ (hipoteza prosta).

Liczymy następnie statystykę testową:

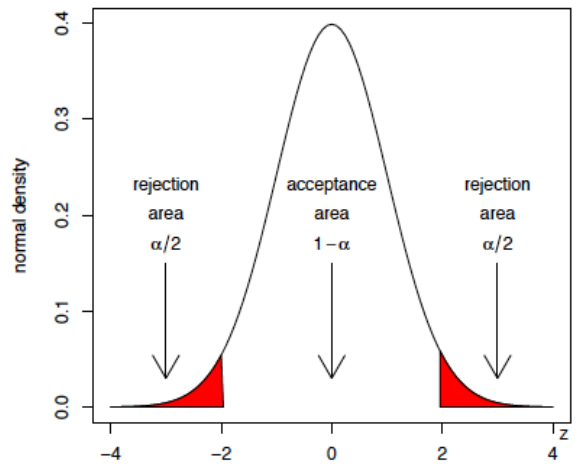
$$Z = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma}$$

Zmienna losowa Z ma standardowy rozkład normalny, a więc potrafimy policzyć P-wartość, czyli prawdopodobieństwo, że zmienna$Z$ przyjmie wartość bardziej ekstremalną niż $|z|$ – porównaj rysunek 1. Odrzucamy hipotezę zerową, jeżeli:

$$P(Z \leq -|z|) + P(Z \geq |z|) = 2P(Z \leq -|z|) \leq \alpha = 0.05$$

jednopróbkowy T-test

Jeżeli wariancja badanej populacji nie jest znana, to do przetestowania hipotezy dotyczącej średnich używamy jednopróbkowego T-testu.



Rysunek 1: Obszar krytyczny dla testu Z.

Niech $H_0 : \mu = \mu_0$ oraz $H_1 : \mu \neq \mu_0$. Statystyka testowa jest zmienną losową:

$$T = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

gdzie s^2 jest wariancją z próby, czyli:

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

.

Ponieważ statystyka testowa ma rozkład T-Studenta o $n-1$ stopniach swobody potrafimy policzyć P-wartość testu:

$$\text{P-val} = 2P(T_{n-1} \leq -|t|)$$

Odrzucamy H_0 jeśli jest ona mniejsza niż ustalony poziom istotność testu α.

dwupróbkowy T-test (Welch'a)

Załóżmy, że podobnie jak w dwóch poprzednich sytuacjach obserwacje pochodzą z rozkładu normalnego, tylko dysponujemy dwiema populacjami (np. próbki pobrane od osób zdrowych i chorych). Testujemy hipotezę $H_0 : \mu_x = \mu_y$ przeciwko $H_1 : \mu_x \neq \mu_y$. Załóżmy, że obserwacje w grupach wynoszą odpowiednio: $x_1, x_2, \ldots x_n$ oraz $y_1, y_2, \ldots y_m$. Niech $\bar{x}$ będzie

średnią dla pierwszej grupy, a $\bar{y}$ dla drugiej. Podobnie oznaczmy przez s_x^2 oraz s_y^2 wariancje z próby. T-statystykę obliczamy następująco:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

.

Przykład 1 Hipoteza o zróżnicowanej ekspresji: *Załóżmy, że porównujemy poziom ekspresji pewnego genu w dwóch populacjach komórek (np. pochodzących od m zdrowych i n chorych dawców). Niech zmienne $X_1, X_2, \ldots, X_n$ oznaczają poziom ekspresji w zdrowych komórkach natomiast $Y_1, Y_2, \ldots, Y_m$ opisują populację chorych komórek. Zakładamy, że pomiary dotyczące poziomów ekspresji są niezależne i pochodzą z rozkładu normalnego o nieznanej wariancji σ^2 identycznej w obydwu grupach oraz nieznanych wartościach oczekiwanych μ_x oraz μ_y. Naturalna hipoteza zerowa mówi, że obydwie rozważane wartości oczekiwane są idenyczne, czyli $\mu_x = \mu_y = \mu$. Natomiast hipoteza alternatywna mówi, że są różne $\mu_x \neq \mu_y$, czyli badany gen w istotny sposób różnicuje dwie populacje. Okazuje, się że adekwatną statystyką w tym zadaniu jest statystyka t omówiona powyżej.*

dwupróbkowy T-test (przypadek równych wariancji)

Powróćmy do poprzedniej sytuacji, kiedy testujemy równość średnich w dwóch populacjach pochodzących z rozkładu normalnego przy założeniu, że wariancje obydwu rozkładów są równe: $\sigma_x^2 = \sigma_y^2$. Podobnie jak poprzednio testujemy hipotezę $H_0 : \mu_x = \mu_y$ przeciwko $H_1 : \mu_x \neq \mu_y$. Zdefiniujmy **łączną wariancję** jako:

$$s_{xy}^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Wtedy następująca zmienna losowa ma rozkład T-Studenta o $m + n - 2$ stopniach swobody:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_{xy}^2 \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

F-test

Do tej pory zajmowaliśmy się testowaniem hipotez dotyczących średniej, czasem zakładaliśmy też (tak jak w poprzednim teście) że wariancje w dwóch badanych populacjach są równe.

F-test testuje tą właśnie własność: hipoteza zerowa zakłada, że $\sigma_x^2 = \sigma_y^2$, natomiast hipoteza alternatywna $H_1 : \sigma_x^2 \neq \sigma_y^2$. Statystyka testowa jest równa:

$$f = \frac{s_x^2}{s_y^2}$$

i ma rozkład F o $(n-1, m-1)$ stopniach swobody. s_x^2 oraz s_y^2 oznaczają wariancje z próby. Nie odrzucimy hipotezy zerowej jeśli:

$$P(F_{n-1,m-1} < f) \geq \frac{\alpha}{2} \text{ dla } f < 1 \text{ lub } P(F_{n-1,m-1} > f) \geq \frac{\alpha}{2} \text{ dla } f > 1$$

Test dwumianowy

Załóżmy, że badamy sekwencję mikroRNA i sformułowaliśmy hipotezę zerową mówiącą, że prawdopodobieństwo występowania puryny na danej pozycji w sekwencji jest równe $p = p_0$. Hipoteza alternatywna postuluje, że to prawdopodobieństwo jest większe $H_1 : p > p_0$. Po zsekwencjonowaniu sekwencji długości n okazało się, że występuje w niej k puryn. Zakładając rozkład dwumianowy dla H_0, możemy policzyć P-wartość, czyli prawdopodobieństwo zaobserwowania k lub więcej puryn przy założeniu hipotezy zerowej:

$$\text{P-val} = P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

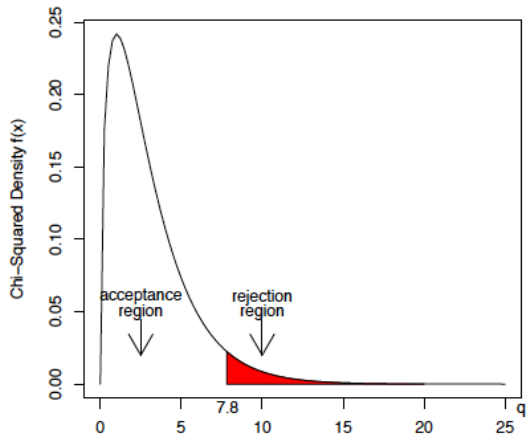
. Jeżeli dla otrzymanych danych k, n i przyjętego p_0, P-wartość jest odpowiednio mała odrzucamy hipotezę zerową. Bardziej ambitny przykład testu dwumianowego był wspomniany przy okazji rozkładu dwumianowego i dotyczył badania wzbogacenia adnotacji genów bliskich badanym obszarom genomowym.

Test chi kwadrat (Pearson)

Będziemy teraz testować hipotezę, która dotyczy więcej niż jednego parametru rozkładu, np niech $H_0 : (\pi_1, \pi_2 \ldots \pi_n) = (p_1, p_2, \ldots p_n)$ oraz $H_1 : (\pi_1, \pi_2 \ldots \pi_n) \neq (p_1, p_2, \ldots p_n)$. Jeśli badane parametry opisują prawdopodobieństwo uzyskania obserwacji danego rodzaju, to możemy policzyć oczekiwaną liczbę obserwacji i-tego rodzaju jako $e_i = np_i$, gdzie n jest rozmiarem badanej próbki. Niech o_i oznacza zaobserwowaną w próbce liczbę wyników i-tego rodzaju. Statystyka:

$$q = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

ma rozkład chi kwadrat o $n-1$ stopniach swobody. Obszar krytyczny dla testu chi kwadrat o 3 stopniach swobody jest zilustrowany na rysunku 2.



Rysunek 2: Obszar krytyczny dla testu χ^2_3

Przykład 2 *Jako przykład zastosowania testu Pearsona rozważmy białko Zyksynę (składnik macierzy pozakomórkowej) i wysuńmy hipotezę zerową, że nukleotydy występują w tym białku z równymi częstościami:* $H_0 : (\pi_1, \pi_2, \pi_3, \pi_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. *Zyksyna składa się z* $n = 2166$ *nukleotydów, i przy założeniu hipotezy zerowej* $e_i = 541,5$ *dla* $i = 1, 2, 3, 4$. *Jeśli policzymy statystykę* q *dla* $o_i \in \{410, 789, 573, 394\}$, *przekonamy się, że* $q \approx 187$, *co odpowiada P-wartości: P-val*≈ 0, *czyli możemy z czystym sumieniem odrzucić hipotezę zerową.*

2 Testy asocjacyjne

Test asocjacyjny chi kwadrat

Często dane, które badamy mają postać tabeli, której każda komórka jest zmienną losową:

	1	2	3	...	c	$\sum$
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1c}	$y_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2c}	$y_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	Y_{r1}	Y_{r2}	Y_{r3}	...	Y_{rc}	$y_{r.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	...	$y_{.c}$	y

Dla wyrobienia intuicji załóżmy, że powyższa tabela posiada jedynie dwa wiersze i dwie

kolumny. Wiersze odpowiadają podziałowi według płci, a kolumny według ręczności (praworęczność vs leworęczność). Hipoteza zerowa zakłada, że nie istnieje żadna zależność pomiędzy płcią a leworęcznością, czyli prawdopodobieństwo spotkania leworęcznego mężczyzny jest takie samo jak spotkania leworęcznej kobiety.

Często sumy komórek w poszczególnych wierszach i kolumnach mogą być ustalone przed etapem testowania hipotezy (dlatego są oznaczane małymi literkami, jako że nie odpowiadają zmiennym losowym). Dodatkowo, żeby opisywany test był poprawny musimy założyć, że wszystkie obserwacje, które odpowiadają zmiennym zliczającym w komórkach tabeli są od siebie niezależne. W przypadku badania leworęczności dwóch bliźniaków jednojajowych ten warunek nie będzie spełniony. Z tego powodu testy asocjacyjne dla sekwencji DNA powinny być używane bardzo ostrożnie, ponieważ często rozważane organizmy są blisko spokrewnione i badane obserwacje nie są niezależne.

Przy założeniu, że hipoteza zerowa jest poprawna, czyli kategorie wierszy i kolumn są od siebie niezależne, możemy obliczyć oczekiwaną liczbę obserwacji w komórce (j, k), czyli wartość oczekiwaną zmiennej losowej Y_{jk}:

$$E_{jk} = E(Y_{jk}) = \frac{y_{j.}y_{.k}}{y}$$

Jeśli hipoteza zerowa jest prawdziwa, to obserwowane wartości zmiennych Y_{jk} powinny być bliskie oczekiwanym, czyli znowu liczymy statystykę chi-kwadrat:

$$\sum_{jk} \frac{(Y_{jk} - E_{jk})^2}{E_{jk}}$$

która ma asymptotycznie rozkład chi kwadrat o $\nu = (r-1)(c-1)$ stopniach swobody.

Przykład 3 *Jako przykład rozważmy test statystyczny na niezależność Markowa, który sprowadzi się do testu asocjacyjnego w tabeli* 4 *x* 4 *i będzie badał, czy częstość występowania danego nukleotydu na danej pozycji zależy od rodzaju nukleotydu na pozycji sąsiedniej.*

Przeprowadza się taki test, żeby ocenić czy można modelować sekwencję DNA jako ciąg prób Bernoulliego, czy raczej z użyciem łańcucha Markowa, który uwzględni taką zależność. Tabela asocjacyjna wygląda w naszym przypadku następująco:

	a	c	g	t	$\sum$
a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$y_{1.}$
c	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$y_{2.}$
g	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$y_{3.}$
t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	$y_{4.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{.4}$	y

Kategoria wiersza określa nukleotyd na pozycji i-tej, kategoria kolumny nukleotyd na pozycji $(i+1)$-szej, np: Zmienna losowa Y_{11} zlicza występowanie dinukleotydów aa w badanej sekwencji DNA. Hipoteza zerowa o niezależności Markowa odpowiada hipotezie zerowej o niezależności wierszy i kolumn. Dla większości sekwencji DNA będziemy zmuszeni odrzucić hipotezę zerową, ponieważ łańcuch Markowa lepiej modeluje sekwencję. Okazuje się też, że jeszcze lepiej modelują łańcuchy Markowa wyższego rzędu, a w niektórych przypadkach nawet niehomogeniczne łańcuchy Markowa.

Test dokładny Fishera

Test dokładny Fishera stosujemy do tablic wymiaru 2 x 2, w których komórkach zmienne losowe przyjmują niewielkie wartości. Dla dużych próbek możemy stosować omówiony powyżej test chi kwadrat, ale jeśli liczby w tablicy są mniejsze bądź równe 5, to postępujemy odmiennie.

	K	M	$\sum$
dieta	$a = 9$	$b = 1$	$a + b = 10$
brak diety	$c = 3$	$d = 11$	$c + d = 14$
$\sum$	$a + c = 12$	$b + d = 12$	$n = 24$

Załóżmy, że znamy wartości brzegowe, czyli $a+b$, $c+d$, $a+c$ oraz $b+d$ w powyższej tabeli. Dodatkowo przyjmijmy, że przedstawione liczby zliczają osoby stosujące i nie stosujące dietę w losowej próbie 12 kobiet i 12 mężczyzn. Hipoteza zerowa stwierdza, że płeć i decyzja o stosowaniu diety są zmiennymi niezależnymi, czyli kobiety tak samo często jak mężczyźni przechodzą na dietę. Przy założeniu hipotezy zerowej, P-wartość dla naszej tabeli możemy policzyć stosując rozkład hipergeometryczny:

$$\text{P-val} = P(Y_{11} = a, Y_{12} = b, Y_{21} = c, Y_{22} = d) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

3 Testy na normalność

Test Shapiro-Wilka

Załóżmy, że chcemy sprawdzić, czy nasze obserwacje: $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego. W wielu metodach statystycznej analizy danych czyni się takowe założenie, dlatego bardzo ważne jest sprawdzenie czy jest ono prawdziwe, albo chociaż, czy nie będziemy zmuszeni odrzucić hipotezy zerowej mówiącej o normalności rozkładu. Popularnym testem adekwatnym w tym przypadku jest **test Shapiro-Wilka** o następującej statystyce:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie $x_{(I)}$ jest i-tym co do wielkości elementem spośród $x_1, x_2, \ldots x_n$, natomiast stałe a_i dla $i = 1, \ldots n$ wynoszą:

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

Wektor $m = (m_1, m_2, \ldots m_n)^T$ odpowiada wartościom oczekiwanym **statystyk porządkowych**[1] dla niezależnych zmiennych losowych o standardowym rozkładzie normalnym, a macierz V jest ich macierzą kowariancji.

4 Testy na elementy odstające

W przypadku kiedy nasze dane nie pochodzą z rozkładu normalnego z dużym prawdopodobieństwem napotkamy w nich elementy odstające (*ang. outliers*). Uwzględnienie takich obserwacji w statystyce testowej może znacząco zaburzyć jej wartość. Z tego powodu konstruuje się statystyki odporne na elementy odstające (*ang. robust*). Przykładem takim jest mediana jako odporna wersja średniej.

Omówimy teraz testy pozwalające sprawdzić, czy wśród badanych obserwacji są elementy odstające. Hipoteza zerowa twierdzi, że takowych nie ma, natomiast hipoteza alternatywna stawia tezę, że wśród naszych danych jest co najmniej jeden taki odstający element. Przy założeniu, że dane $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego (warto to przetestować), hipotezę tą testuje **test Grubbsa**. Statystyka testowa jest równa:

$$G = \frac{\max_{i=1\ldots n} |X_i - \bar{X}|}{s}$$

Odrzucimy hipotezę zerową na poziomie istotności α jeżeli

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{n-2}(\frac{\alpha}{2n})}{n-2+t^2_{n-2}(\frac{\alpha}{2n})}}$$

gdzie $t^2_{n-2}(\frac{\alpha}{2n})$ jest górną wartością krytyczną w rozkładzie T o $n-2$ stopniach swobody, przy poziomie istotności $\frac{\alpha}{2n}$. Czyli taką wartością, że całka z górnego ogona tego rozkładu liczonego od tej wartości wynosi $\frac{\alpha}{2n}$.

[1] statystyka porządkowa rzędu k dla próby losowej to k-ta co do wielkości wartość tej próby, przykładem statystyki porządkowej jest mediana.

5 Testy nieparametryczne

Do tej pory testowaliśmy hipotezy dotyczące parametrów znanych rozkładów prawdopodobieństwa, czyli zakładaliśmy, że dane pochodzą z rozkładu dwumianowego, normalnego, etc. W przypadku, kiedy nie możemy poczynić takich założeń powinniśmy skorzystać z testów nieparametrycznych. Opiszemy tu dwa nieparametryczne zamienniki dla dwupróbkowego T-testu, czyli **test Manna-Whitneya** oraz **test permutacyjny**.

Test Manna-Whitneya

Załóżmy, że obserwowane wartości zmiennych losowych $X_1, \dots X_n$ oraz $Y_1, \dots Y_m$, czyli $x_1, \dots x_n$ oraz $y_1, \dots y_m$ są podane w kolejności rosnącej. Każdej obserwacji przypisujemy jej pozycję rankingową (*rangę*) na wspólnej posortowanej liście, czyli jedną z liczb $1, 2, \dots m + n$. Oczywiście jest to możliwe tylko wtedy, gdy wszystkie obserwacje są różne co niniejszym dla uproszczenia założymy (oczywiście istnieje też ogólna postać tego testu dopuszczająca powtórzenia). Statystyką testową jest suma *rang* obserwacji z pierwszej grupy. Łatwo policzyć, że suma rang wszystkich obserwacji (z obydwu grup) wynosi $R = (m+n)(m+n+1)/2$. Przy założeniu hipotezy zerowej (rozkłady z jakich pochodzą obserwacje są takie same) suma rang obserwacji w pierwszej grupie powinna stanowić $n/(m + n)$ część sumy wszystkich rankg R, czyli jej wartość oczekiwana wynosi:

$$\frac{n}{n+m}\frac{(n+m)(n+m+1)}{2} = \frac{n(n+m+1)}{2}$$

Można też pokazać, że suma rang w pierwszej grupie ma rozkład bliski normalnemu, natomiast wariancja tej sumy wynosi:

$$\frac{nm(n+m+1)}{12}$$

Podsumowując zredukowaliśmy zadanie do testowania wartości średniej rozkładu normalnego o znanej wariancji, co już potrafimy zrobić.

Test permutacyjny

Przy założeniu hipotezy zerowej (dane w próbkach pochodzą z jednakowych rozkładów), wszystkie $\binom{n+m}{n}$ permutacje przypisujące pierwsze n elementów do pierwszej grupy i kolejne m elementów do drugiej grupy są jednakowo prawdopodobne. Dla każdej takiej permutacji liczymy wartość pewnej statystyki testowej (nie jest jednoznacznie powiedziane jaką wybrać może to być np. statystyka dla dwupróbkowego T testu). Jedna z wartości statystyki będzie

tą, która pojawia się dla permutacji prawdziwych obserwowanych danych. Jeśli hipoteza alternatywna twierdzi, że średnia z pierwszej grupy jest większa od średniej z drugiej grupy, to używając parametru α (poziom błędów typu I) odrzucamy hipotezę zerową jeśli obserwowana wartość statystyki jest pośród górnych $\alpha 100\%$ wartości.

Zauważmy, że przy tym podejściu nie potrzebujemy znać rozkładu prawdopodobieństwa badanej statystyki testowej. Dodatkowo jeśli używamy T statystyki nie musimy jej obliczać dla każdej permutacji. Wystarczy jedynie policzyć różnicę średnich w dwóch grupach dla każdej permutacji, a tak naprawdę to wystarczy tylko policzyć średnią obserwacji w pierwszej grupie[2]. Ostatnią miłą własnością testu permutacyjnego jest fakt, że dla danych pochodzących z rozkładu normalnego otrzymane wyniki są bardzo zbliżone do tych z T testu.

Niemiłą z kolei własnością jest złożoność obliczeniowa tego testu (nawet dla średnich wartość n i m liczba permutacji rośnie bardzo szybko). W sytuacji kiedy nie możemy policzyć statystyki testowej dla wszystkich permutacji, losujemy odpowiednio dużą próbkę permutacji i odrzucamy hipotezę zerową jeśli obserwowana wartość statystyki testowej znajdzie się pośród $\alpha 100\%$ rozważanych dodatnich wartości.

Obydwa opisane testy nieparametryczne mają podstawową wadę: hipoteza zerowa, mówiąca o równości rozkładów z jakich pochodzą próbki jest mocniejsza niż hipoteza zerowa którą chcielibyśmy rozważać, czyli że średnie w obydwu grupach są różne.

6 Redukcja wymiaru dla danych wielowymiarowych

Popularnym podejściem do zadania redukcji wymiaru jest **selekcja cech**, polegająca na wyborze pewnego podzbioru spośród zbioru oryginalnych cech. W przypadku selekcji cech stosowanej w zadaniu klasyfikacji dokonujemy wyboru cech najlepiej dyskryminujących dwie populacje. Podsumujemy teraz korzyści, na jakie możemy liczyć stosując redukcję wymiaru danych.

- Redukcja wymiaru pozwala skupić się na cechach istotnych z punktu widzenia zadania klasyfikacji co pozwala z kolei zredukować efekt **przeuczenia**, polegający na zbytnim dopasowaniu klasyfikatora do danych treningowych.
- Poprzez eliminację szumu, czyli cech nieistotnych dla zadania klasyfikacji, zwiększamy skuteczność klasyfikatora.
- Dzięki redukcji wymiaru usuwamy problemy związane z wysokim wymiarem danych.

[2] uzasadnij dlaczego ?

- Redukcja wymiaru danych pozwala w przypadku każdego klasyfikatora, na znaczną oszczędność czasową i pamięciową.
- Umożliwia lepszą interpretację sygnałów zidentyfikowanych w danych.
- W konkretnym przykładzie analizy danych mikromacierzowych, badanie wybranych 50 genów jest wykonalne (czasowo, finansowo, merytorycznie), natomiast analiza wszystkich (nawet 40000) próbek jest po prostu niemożliwa.

7 Metody wyboru cech dyskryminujących

Omówimy teraz popularne metody **selekcji cech**. Będziemy wybierać te spośród oryginalnych zmiennych, które najlepiej (według pewnego testu statystycznego) różnicują obserwacje z różnych klas. Umożliwia to jednocześnie bezpośrednią identyfikację potencjalnych biomarkerów.

Test t Dwupróbkowy T Test może być użyty do wyboru najlepiej dyskryminujących cech np w widmach spektrometrycznych, lub w analizie danych z macierzy transkryptomicznych. Wybieramy cechy o największej wartości statystyki.

Algorytm PPC Algorytm PPC (*ang. Peak Probability Contrasts* (por. [4]) stosuje inne podejście do wyboru dyskryminujących cech. Polega ono na wybraniu osobno dla każdej cechy punktu przecięcia najlepiej różnicującego dwie grupy (np. chorych i zdrowych). Przecięć szuka się wśród ustalonej liczby kwantyli. Przypomnijmy, że kwantylem rzędu p ciągłej zmiennej losowej X nazwiemy liczbę q_p spełniającą warunek:

$$P(X < q_p) = p$$

Wybór przecięcia wśród kwantyli jest uzasadniony przez fakt, że wysokości wierzchołków rozkładają się bardzo nierównomiernie. Kwantyl najlepiej rozdzielający wysokości wierzchołków z różnych klas jest wybrany następująco:

- Niech $q_{\alpha i}$ będzie kwantylem rzędu α zmiennej oznaczającej wysokość wierzchołka na i-tej pozycji w spektrum.
- Dla danych dwóch klas G_1, G_2, o licznościami n_1,n_2, niech $p_{il}(\alpha)$ będzie proporcją obserwacji w klasie l, w których objętość na pozycji i-tej jest większa niż $q_{\alpha i}$:

  $$p_{il}(\alpha) = \sum_{j \in G_l} I[x_{ij} > q_{\alpha i}]/n_l \quad l = 1, 2,$$

  gdzie x_{ij} oznacza wysokość wierzchołka na i-tej pozycji widma u j-tego pacjenta, a $I[\cdot]$ jest indykatorem, równym 1 jeżeli wyrażenie jest prawdziwe i 0 w.p.p..

- Dla każdego i spośród rozpatrywanych rzędów kwantyli $\alpha_1 \dots \alpha_h$ wybierz rząd $\hat{\alpha}_i$ maksymalizujący po α wyrażenie $|p_{i2}(\alpha) - p_{i1}(\alpha)|$. Optymalny kwantyl (punkt przecięcia) dla i-tej cechy oznaczymy przez $\hat{q}_i = q_{\hat{\alpha}_i i}$. Częstości występowania wierzchołków o wysokości większej niż $\hat{q}_i$ w poszczególnych klasach oznaczymy przez $\hat{p}_{i1} = p_{i1}(\hat{\alpha}_i)$ oraz $\hat{p}_{i2} = p_{i2}(\hat{\alpha}_i)$.

W ten sposób możemy ustalić ranking cech według malejących wartości $|\hat{p}_{i2} - \hat{p}_{i1}|$.

7.1 Miary zróżnicowania między rozkładami prawdopodobieństwa

Algorytm PPC nie bierze pod uwagę różnicy w wysokościach wierzchołków, które znajdują się powyżej lub poniżej ustalonego przecięcia. Zamiast wyznaczać pojedynczy punkt przecięcia, proponujemy porównywanie rozkładów wysokości wierzchołków w zależności od klasy. Służą do tego miary zróżnicowania rozkładów prawdopodobieństwa. Jedną z najbardziej znanych takich miar jest odległość Kullbacka-Leiblera[3] zdefiniowana (w przypadku dyskretnym) następująco:

$$D_{KL}(p_1||p_2) = \sum_x p_1(x) log_2 \frac{p_1(x)}{p_2(x)},$$

gdzie p_1 i p_2 są rozkładami zmiennej losowej X, a x przebiega zbiór wartości tej zmiennej. Odległość Kullbacka-Leiblera jest nieujemna i równa 0 wtedy i tylko wtedy, gdy $p_1 \equiv p_2$. Miara ta jest jednak niesymetryczna. Chcąc jednakowo traktować obydwa rozkłady, możemy zastosować miarę *resistor-average* zaproponowaną przez D.H. Johnsona i S. Sinanovića. Jest to średnia harmoniczna odległości Kullbacka-Leiblera:

$$\frac{1}{D_{RA}(p_1, p_2)} = \frac{1}{D_{KL}(p1||p2)} + \frac{1}{D_{KL}(p2||p1)}.$$

Jeżeli $p_1 \equiv p_2$, to przyjmujemy $D_{RA}(p_1, p_2) = 0$.

7.2 Wzajemna informacja dwóch źródeł

Inną metodą wyboru istotnych cech jest zbadanie jak dobrze każda z nich objaśnia zmienną decyzyjną czyli wektor $\mathbf{y}$. Można w tym celu zastosować współczynnik wzajemnej informacji wywodzący się z teorii informacji Shanona.

Niech źródło A nadaje komunikaty ze zbioru $\Omega_1 = \{x_1, \dots, x_m\}$, a źródło B ze zbioru $\Omega_2 = \{y_1, \dots, y_n\}$. Niech $p_1(x)$ oznacza prawdopodobieństwo otrzymania komunikatu x ze

[3] Odległość Kullbacka-Leiblera mimo przyjętej nazwy nie jest oczywiście metryką.

źródła A, a $p_2(y)$ prawdopodobieństwo otrzymania komunikatu y ze źródła B. Dalej, niech $p(x, y)$ oznacza prawdopodobieństwo otrzymania pary komunikatów (x, y) odpowiednio ze źródeł A i B. Wtedy współczynnik wzajemnej informacji źródeł A i B zdefiniowany jest następująco:

$$I(A, B) = \sum_x \sum_y p(x, y) \log_2 \frac{p(x, y)}{p_1(x)p_2(y)}$$

Współczynnik wzajemnej informacji jest nieujemny i równy 0 wtedy i tylko wtedy, gdy źródła nadają komunikaty niezależnie od siebie tzn. $p(x, y) = p_1(x)p_2(y)$. Wzajemna informacja jest także nie większa niż entropia każdego z dwóch źródeł.

W naszym przypadku będziemy badać wzajemną informację wektora wartości danej cechy w różnych próbkach oraz wektora wyznaczającego klasyfikację tych próbek.

7.3 Współczynnik Goodmana-Kruskalla

Wzajemna informacja jest miarą zależności między zmiennymi wykorzystującą entropię jako miarę zmienności. Inną często wykorzystywaną miarą zmienności jest współczynnik Gini zdefiniowany dla dyskretnej zmiennej losowej X w następujący sposób:

$$V(X) = 1 - \sum_x p(x)^2,$$

gdzie x przebiega zbiór wartości zmiennej losowej X. Współczynnik osiąga najmniejszą wartość (0), gdy zmienna X może przyjmować tylko jedną wartość, a największą, gdy rozkład zmiennej X jest jednostajny. Korzystając z miary zmienności V, średnią zmienność warunkową zmiennej losowej Y pod warunkiem X opiszemy przez:

$$E[V(Y|X)] = \sum_x p(x)V(Y|x),$$

gdzie $V(Y|x)$ wyraża zmienność Y pod warunkiem, że $X = x$:

$$V(Y|x) = 1 - \sum_y p(y|x)^2.$$

W związku z powyższym stopień zmniejszenia zmienności zmiennej Y przy znajomości zmiennej X nazywany współczynnikiem τ Goodmana-Kruskala zdefiniujemy następująco:

$$\tau_{GK} = \frac{V(Y) - E[V(Y|X)]}{V(Y)}$$

Współczynnik τ_{GK} osiąga wartość najmniejszą (równą 0) wtedy i tylko wtedy, gdy zmienne X i Y są niezależne, natomiast wartość największą (równą 1), gdy dla każdego x przebiegającego zbiór wartości zmiennej X istnieje y (wartość zmiennej Y) taki, że $p(y|x) = 1$. W praktyce będziemy za zmienną Y przyjmować zmienną decyzyjną, czyli zmienną oznaczającą klasę obserwacji, natomiast za zmienną X badaną cechę.

7.4 Korelacja

Ostatnią zastosowaną przez nas metodą jest ustalenie rankingu cech według wartości bezwzględnej korelacji między badanymi cechami, a zmienną decyzyjną. Współczynnik korelacji próbkowej τ_P ma postać:

$$\tau_P = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}},$$

gdzie $\overline{x}$ i $\overline{y}$ oznaczają średnie w próbach $x_1, \ldots, x_n$ oraz $y_1, \ldots, y_n$.

8 Jednoczesne testowanie wielu hipotez

Wróćmy teraz do naszych wielowymiarowych danych. Filtrując cechy np. przy pomocy testu t chcielibyśmy zredukować wymiar danych poprzez wybór cech które w statystycznie istotny sposób różnicują dwie populacje. Zauważmy, że nasze zadanie jest równoważne problemowi jednoczesnego testowania wielu tysięcy hipotez zerowych: $H_1, H_2, \ldots, H_m$. Oznaczmy przez R liczbę odrzuconych hipotez (czyli np. w przypadku mikromacierzy liczbę genów o różnicującej ekspresji). Mamy następującą sytuację:

	# przyjętych H_0	# odrzuconych H_0	$\sum$
# prawdziwych H_0	U	V	m_0
# fałszywych H_0	T	S	m_1
$\sum$	$m - R$	R	m

gdzie R jest obserwowaną zmienną losową, m_0 oraz m_1 nieznanymi parametrami, podobnie jak U, V, T oraz S są nieobserwowanymi zmiennymi losowymi.

Opiszemy teraz jak uogólnia się zadanie kontroli błędów typu I w problemie testowania wielu hipotez. W przypadku pojedynczego testu (hipotezę zerową oznaczamy tutaj dość myląco przez H_1) potrafiliśmy policzyć wielkość c_α, taką że:

$$\Pr(|T_1| \geq c_\alpha | H_1) \leq \alpha$$

gdzie T_1 jest wartością statystyki testowej. Odrzucaliśmy hipotezę H_1 jesli $|T_1| \geq c_\alpha$. Najczęsciej stosowane uogólnienia tego podejścia są następujące (por. [1]):

- **PCER** (*ang. Per-comparison error rate*), miara jest zdefiniowana jako średnia z wartości oczekiwanej błędów typu I, czyli:

$$\text{PCER} = \frac{E(V)}{m}$$

- **PFER** (*ang. Per-family error rate*), odpowiada oczekiwanej liczbie błędów typu I:

$$\text{PFER} = E(V)$$

- **FWER** (*ang. Family-wise error rate*), jest zdefiniowana jako prawdopodobieństwo co najmniej jednego błędu typu I:

$$\text{FWER} = \Pr(V \geq 1)$$

- **FDR** (*ang. False discovery rate*), definiujemy jako oczekiwaną proporcję błędów typu I pomiędzy odrzuconymi hipotezami zerowymi (jest to procent fałszywych pozytywów, czyli cech uznanych niesłusznie za istotne):

$$\text{FDR} = E(Q)$$

$$Q = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}$$

Omówmy bardziej szczegółowo tylko dwie ostatnie miary jako najczęściej stosowane w bioinformatyce. W przypadku FWER stosujemy tzw **poprawkę Bonferroniego** i odrzucamy hipotezę zerową H_j ($j = 1, 2, \ldots, m$) jeśli odpowiednia p-wartość jest mniejsza bądź równa $\frac{\alpha}{m}$ (gdzie α jest dopuszczalnym procentem błędów I typu w pojedynczym teście).

Sensowność poprawki uzasadnia następujące rozumowanie: załóżmy, że poprawne hipotezy zerowe to $H_1, \ldots H_{m_0}$. Niech P_j będzie zmienną losową opisującą p-wartość dla hipotezy H_j, natomiast $\tilde{P}_j = \min(mP_j, 1)$ jej „poprawionym” odpowiednikiem. Mamy:

$$\text{FWER} = \Pr(V \geq 1) = \Pr(\cup_{j=1}^{m_0} \{\tilde{P}_j \leq \alpha\}) \leq \sum_{j=1}^{m_0} \Pr(\tilde{P}_j \leq \alpha) \leq$$

$$\leq \sum_{j=1}^{m_0} \Pr(P_j \leq \frac{\alpha}{m}) = \frac{m_0 \alpha}{m}$$

Ostatnia równość wynika z obserwacji, że P_j mają rozkład jednostajny na odcinku $(0, 1)$ przy założeniu hipotezy zerowej.

Problem związany z poprawką Bonferroniego jest taki, że ograniczając liczbę fałszywych pozytywów przez obniżenie progu p-wartości do $\frac{\alpha}{m}$ drastycznie zwiększamy próg statystyki testowej, co z kolei prowadzi do wielu fałszywych negatywów, czyli przyjęcia niepoprawnej hipotezy zerowej, a więc nie wykrycia istotnych cech.

8.1 Ocena istotności cech przy użyciu testu FDR

Opiszemy teraz praktyczną implementację metody FDR (False Discovery Rates). Dla danego poziomu t oznaczymy przez T liczbę cech, dla których wartość badanego testu przekracza t. Następnie wielokrotnie (R razy) permutujemy etykiety klas obserwacji, za każdym razem obliczając wartość testu dla każdej cechy oraz oznaczając liczbę cech, dla których wartość testu przekracza t przez T_i^* ($i = 1, \ldots, R$). Wartość FDR dla poziomu t będziemy estymować wzorem:

$$\widehat{FDR}(t) = \frac{\sum_{i=1}^{R} T_i^* / R}{T}.$$

Wyznaczamy taki poziom $\hat{t}$, dla którego FDR jest niskie (np. 0.05). Cechy, dla których wartość testu przewyższa $\hat{t}$ będziemy traktować jako istotne.

Literatura

[1] Dudoit,S., Shaffer,J.P. and Boldrick,J.C. *Multiple Hypothesis Testing in Microarray Experiments*, Statistical Science, Vol. 18, No. 1, p. 71-103.

[2] Hastie,T., Tibshirani,R. and Friedman,J-H. *The Elements of Statistical Learning*, Springer Verlag, 2001.

[3] J. Koronacki, J. Ćwik *Statystyczne systemy uczące się*, WNT 2005.

[4] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, Q. Le *Sample classification from protein mass spectrometry, by "peak probability contrasts"*, Bioinformatics Advance Access, June 28, 2004.

[5] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28(5):495-50, 2010.

[6] K. Seefeld, E. Linder, Statistics using R with biological examples, 2007.

[7] W.P. Krijnen Applied statistics for bioinformatics using R, 2009.

[8] S.K. Mathur, Statistical Bioinformatics with R, Elsevier Academic Press, 2010.

[9] W. J. Ewens, G. R. Grant, Statistical Methods in Bioinformatics Springer-Verlag, 2001.

[10] Lawrence Leemis, Relationships Among Common Univariate Distributions, American Statistician 40:143-146. 1986