Uczenie maszynowe w bioinformatyce

Wykład 9: selekcja cech różnicujących

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Etapy przetwarzania i analizy danych

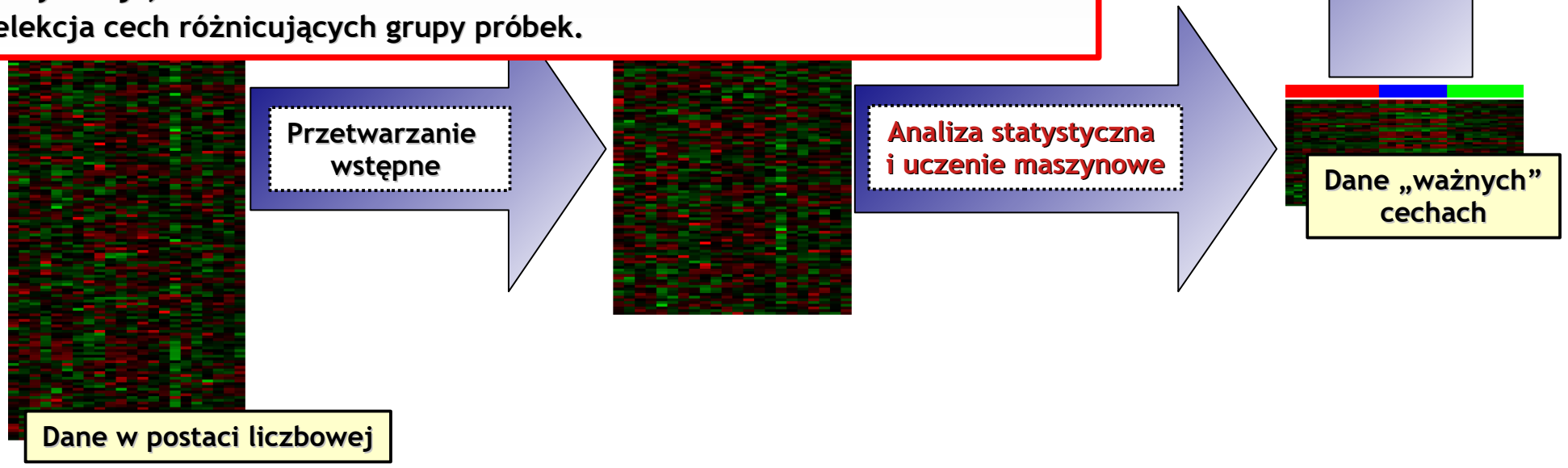
Wybór **metod statystycznych i uczenia się maszyn** używanych na etapie analizy danych jest ściśle uzależniony od celu badań. Stosowane tutaj techniki można w ogólności podzielić na **działające z nadzorem** lub **nienadzorowane**.

Metody działające bez nadzoru stosuje się do wykrywania zależności pomiędzy próbkami i/lub cechami jedynie w oparciu o dane, bez używania dodatkowych informacji (w tym też nie uwzględniając przynależności badanych obiektów do zadanych z góry klas). Do metod nienadzorowanych zaliczają się m. in.:

- ✓ klasteryzacja;
- ✓ techniki redukcji wymiarowości.

Metody nadzorowane korzystają z danych i dodatkowej wiedzy, dotyczącej np. oczekiwanych wartości na wyjściu lub sposobu przypisania próbek do znanych klas. Przykładami zagadnień, w których używa się takich metod są:

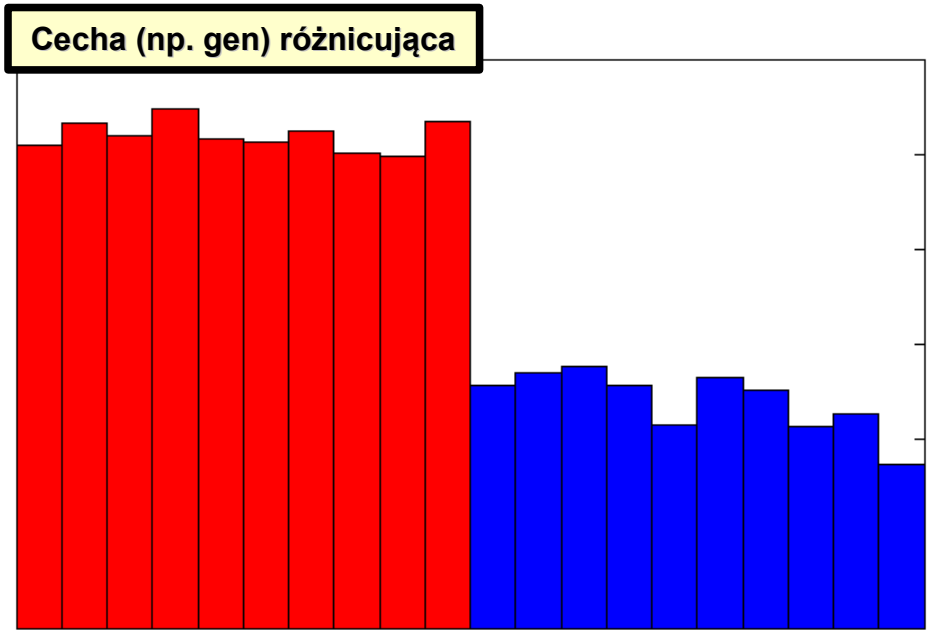
- ✓ klasyfikacja;
- → selekcja cech różnicujących grupy próbek.

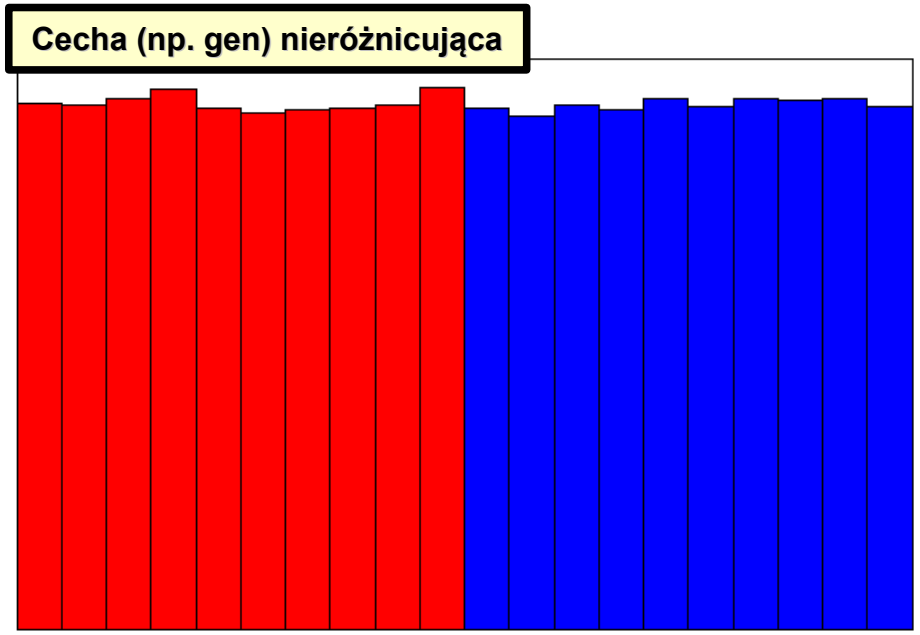


Selekcja cech różnicujących

W szerokim ujęciu, selekcja cech to procedura prowadząca do wybrania zbioru $R < P$ cech, które spełniają pewne zadane z góry kryteria (inaczej mówiąc: mających jakieś „interesujące” z punktu widzenia danego eksperymentu właściwości).

Najczęściej jednak termin ten jest używany w węższym znaczeniu, ograniczonym do selekcji cech różnicujących, czyli takich, których wartości średnie w istotnie różnią się między grupami próbek reprezentującymi rozpatrywany układ eksperymentalny.





Selekcja cech różnicujących

Techniki selekcji cech różnicujących zależą od celu w jakim jest ona prowadzona.

- Jeżeli wykonuje się **selekcję cech na potrzeby klasyfikacji**, to celem jest uzyskanie możliwie małego zbioru cech, które **wspólnie** posłużą do zbudowania klasyfikatora charakteryzującego się minimalnym błędem.

  Taką selekcję można najskuteczniej wykonać za pomocą **metod korzystających z klasyfikatora, którego jakość jest oceniana poprzez walidację krzyżową (CV, *cross validation*)**, aczkolwiek możliwe są również inne podejścia.

- Gdy **selekcja prowadzona jest w celu określania cech typowych dla badanych grup** (np. zbiorów genów o ekspresji zmienionej w wyniku schorzenia) zwykle dąży się do uzyskania jak najdłuższych list cech, z których każda **samodzielnie i niezależnie od pozostałych** wykazuje się znaczącym stopniem różnicowania.

  W tym przypadku typowym podejściem do selekcji cech jest użycie **statystycznych testów istotności** badających różnice rozkładów wartości cech w grupach.

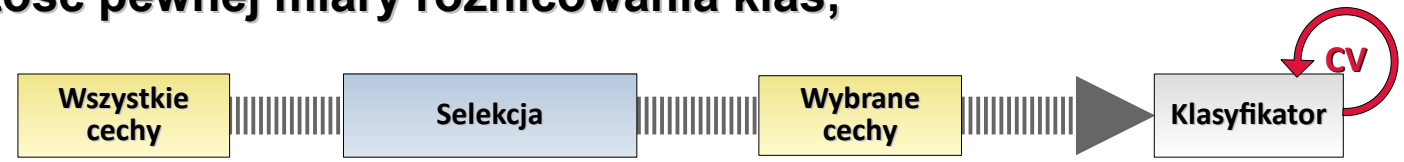
Uczenie maszynowe w bioinformatyce

Wykład 9: selekcja cech pod kątem klasyfikacji

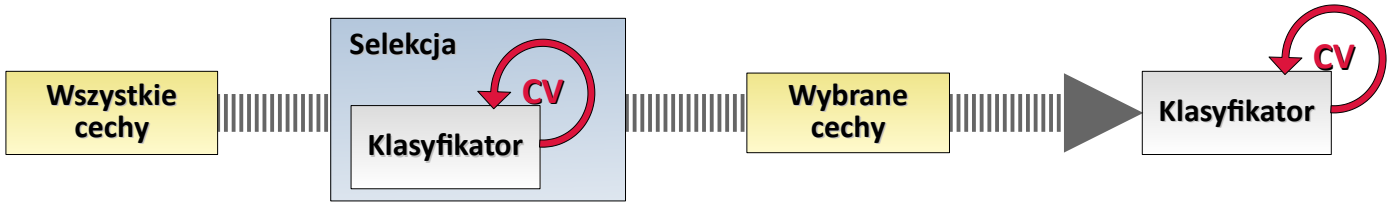
Selekcja cech pod kątem klasyfikacji

Metody selekcji cech na potrzeby klasyfikacji można podzielić na trzy ogólne grupy, w zależności od tego, czy i w jaki sposób jest w nich używany klasyfikator:

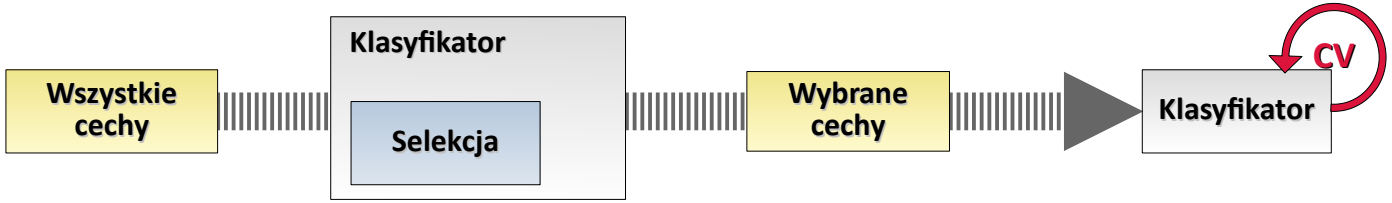
- **metody „filtrujące” (*filter*)** działają niezależnie od klasyfikatora, przypisując każdej z cech wartość pewnej miary różnicowania klas;



- w **metodach „opakowujących” (*wrapper*)** miarą przydatności cechy jest jej wpływ na jakość klasyfikatora, zwykle weryfikowaną na drodze walidacji krzyżowej;



- **metody „wbudowane” (*embedded*)** oceniają cechy w oparciu o ich znaczenie dla modelu klasyfikacyjnego.



Selekcja cech pod kątem klasyfikacji: metody „filtrujące”

W odróżnieniu od dwóch pozostałych podejść, **działanie metod „filtrujących” nie jest w żaden sposób powiązane z konkretnym klasyfikatorem**: są one uruchamiane przed etapem budowania modelu klasyfikacyjnego i nie biorą pod uwagę jego właściwości.

Typowy algorytm „filtrujący” składa się z trzech podstawowych etapów:

- wyznaczenie dla każdej cechy wartości miary f_i, świadczącej o stopniu, w jakim ta cecha różnicuje klasy;
- sortowanie cech zgodnie z rosnącymi wartościami miary f_i;
- wybór zadanej liczby cech o największych wartościach f_i.

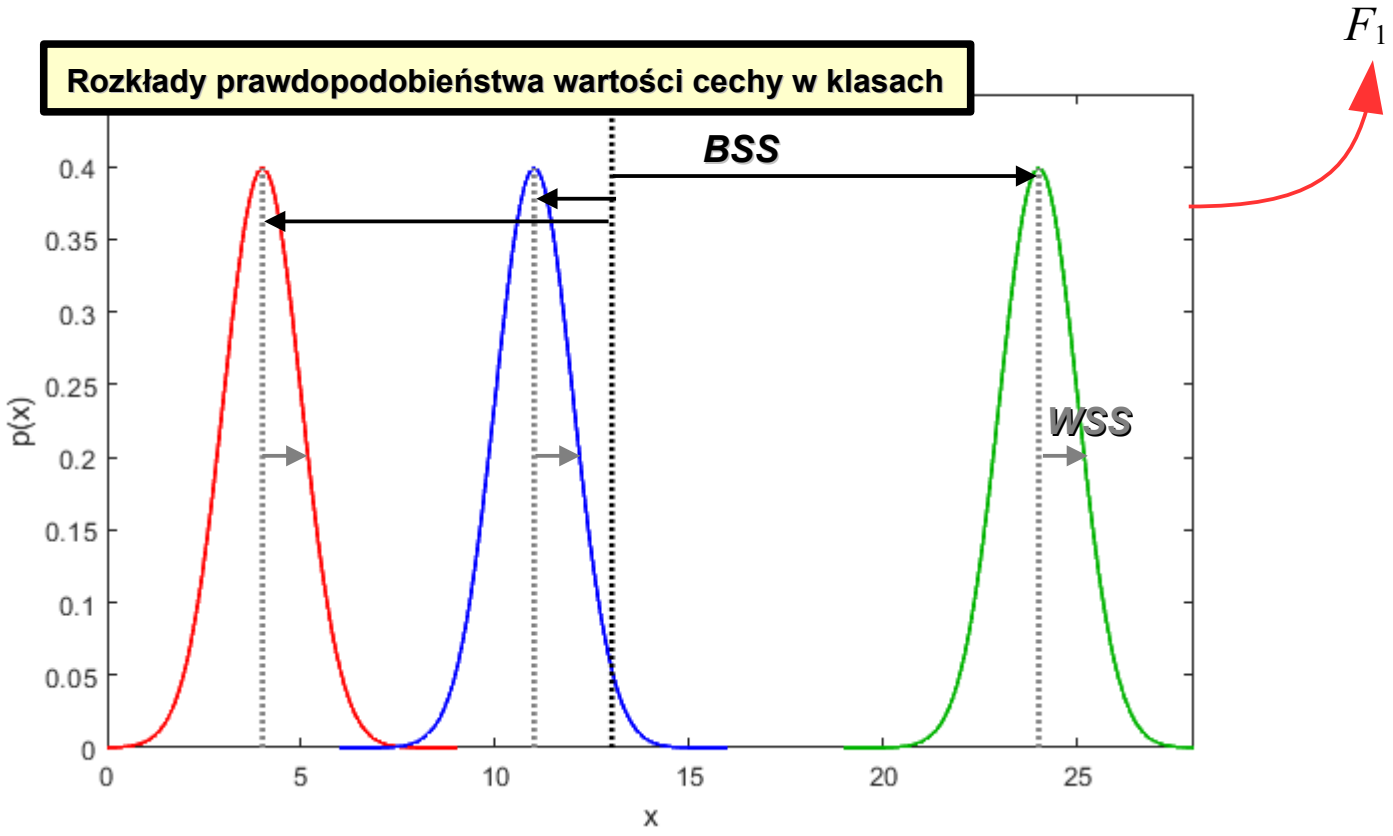
Powyższy schemat jest wspólny dla wszystkich tego typu metod, które jednak różnią się sposobem wyznaczania miary różnicowania cech.

Selekcja cech pod kątem klasyfikacji: metody „filtrujące” (*Fisher score*)

Przykładem miary różnicowania cechy jest wskaźnik Fishera *(Fisher score, F-score)*:

$$F_i = \frac{BSS_i}{WSS_i} = \frac{\sum_{k=1}^{K} N_k \left(\bar{x}_{ki} - \bar{x}_i \right)^2}{\sum_{k=1}^{K} N_k \cdot s_{ki}^2}$$

gdzie: $\bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_j$; $\bar{x}_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} x_{kij}$; $s_{ki}^2 = \frac{1}{N_k} \sum_{j=1}^{N_k} \left(x_{kij} - \bar{x}_{ki} \right)^2$

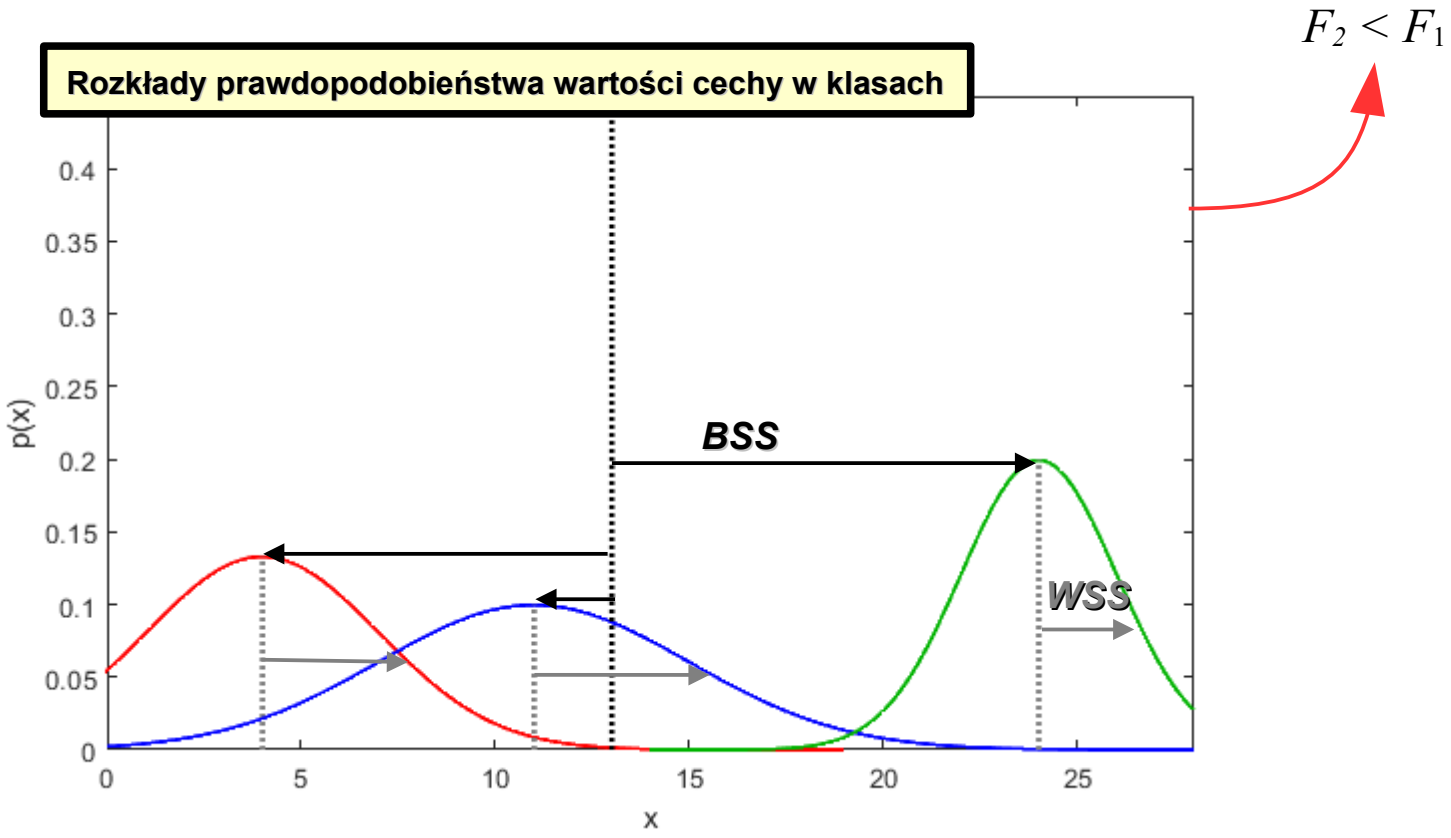


Selekcja cech pod kątem klasyfikacji: metody „filtrujące” (*Fisher score*)

Przykładem miary różnicowania cechy jest wskaźnik Fishera *(Fisher score, F-score)*:

$$F_i = \frac{BSS_i}{WSS_i} = \frac{\sum_{k=1}^{K} N_k \left(\bar{x}_{ki} - \bar{x}_i \right)^2}{\sum_{k=1}^{K} N_k \cdot s_{ki}^2}$$

gdzie: $\bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_j$; $\bar{x}_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} x_{kij}$; $s_{ki}^2 = \frac{1}{N_k} \sum_{j=1}^{N_k} \left(x_{kij} - \bar{x}_{ki} \right)^2$

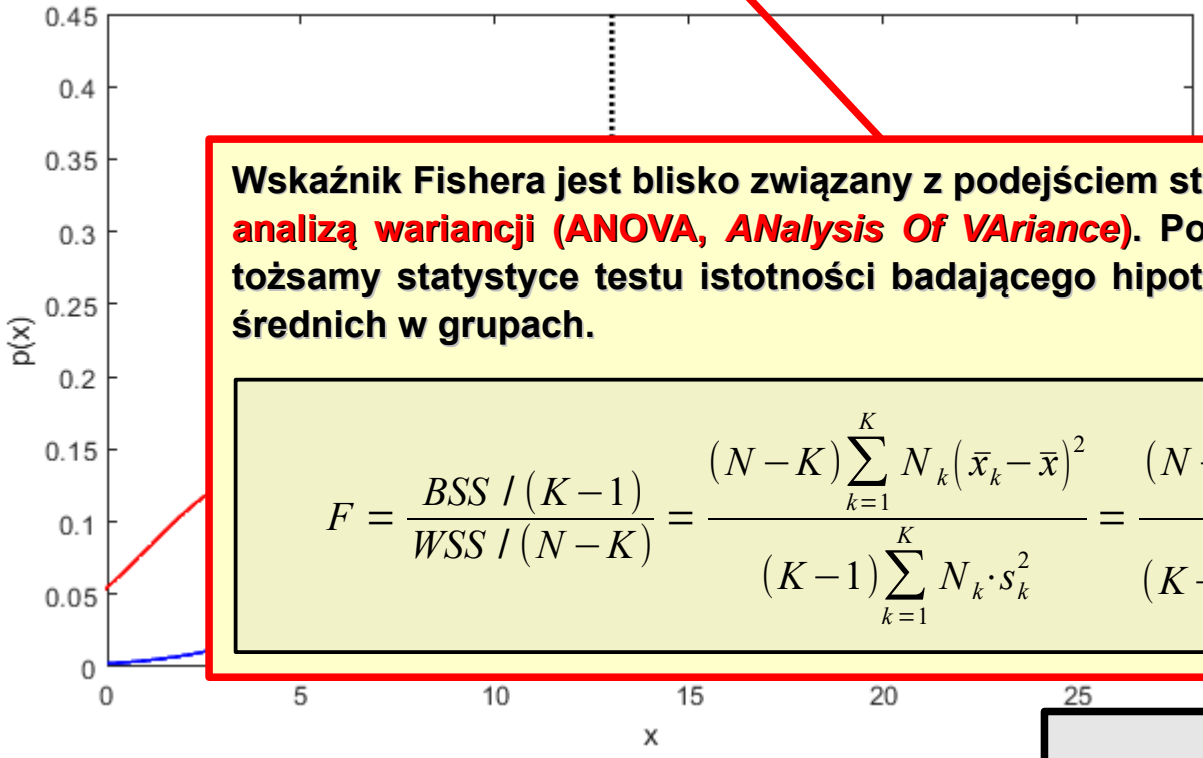


Selekcja cech pod kątem klasyfikacji: metody „filtrujące” (*Fisher score*)

Przykładem miary różnicowania cechy jest wskaźnik Fishera *(Fisher score, F-score)*:

$$F_i = \frac{BSS_i}{WSS_i} = \frac{\sum_{k=1}^{K} N_k \left(\bar{x}_{ki} - \bar{x}_i \right)^2}{\sum_{k=1}^{K} N_k \cdot s_{ki}^2}$$

gdzie: $\bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_j$; $\bar{x}_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} x_{kij}$; $s_{ki}^2 = \frac{1}{N_k} \sum_{j=1}^{N_k} \left(x_{kij} - \bar{x}_{ki} \right)^2$



Wskaźnik Fishera jest blisko związany z podejściem statystycznym nazywanym analizą wariancji (ANOVA, *ANalysis Of VAriance*). Po przeskalowaniu jest on tożsamy statystyce testu istotności badającego hipotezę o równości wartości średnich w grupach.

$$F = \frac{BSS\ /\ (K-1)}{WSS\ /\ (N-K)} = \frac{(N-K)\sum_{k=1}^{K} N_k \left(\bar{x}_k - \bar{x} \right)^2}{(K-1)\sum_{k=1}^{K} N_k \cdot s_k^2} = \frac{(N-K)\sum_{k=1}^{K} N_k \left(\bar{x}_k - \bar{x} \right)^2}{(K-1)\sum_{k=1}^{K} \sum_{j=1}^{N_k} \left(x_{kj} - \bar{x}_k \right)^2}$$

	MATLAB	R
ANOVA:	`anova1`	`oneway.test`

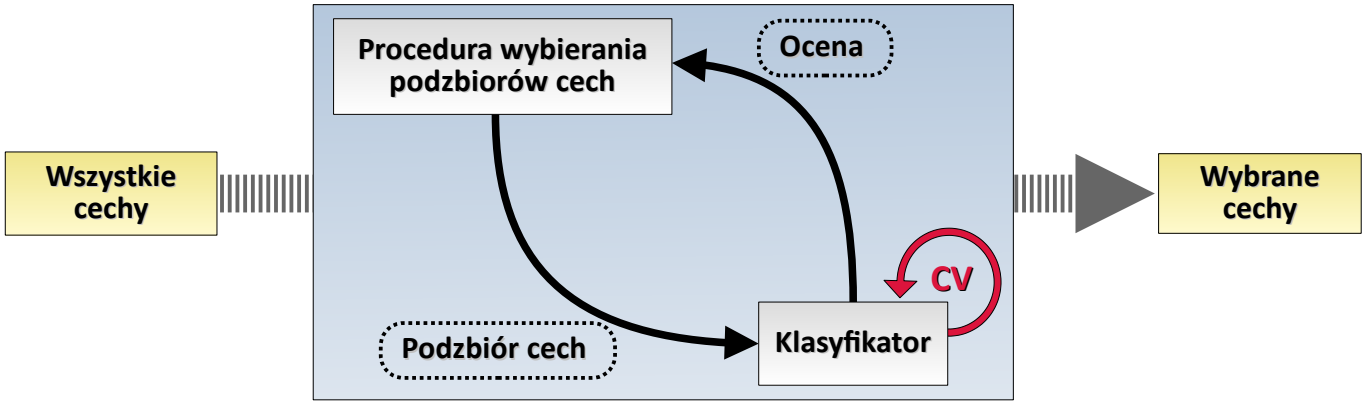
Selekcja cech pod kątem klasyfikacji: metody „opakowujące”

Metody „opakowujące” korzystają z konkretnego klasyfikatora do ustalenia wpływu poszczególnych cech na jakość klasyfikacji, ocenianą za pomocą ustalonej miary.

Przy zadanym rodzaju klasyfikatora metody te iteracyjnie:

- wyznaczają nowy podzbiór cech;
- oceniają aktualny podzbiór cech przez zbudowanie klasyfikatora i określenie jego jakości (np. miary ACC lub AUC), najczęściej przy użyciu walidacji krzyżowej (CV).

W wyniku działania procedury wybierane są te cechy, które w kolejnych iteracjach miały najbardziej korzystny wpływ na wynik klasyfikacji.



Selekcja cech pod kątem klasyfikacji: metody „opakowujące” (SFS)

Przykładem prostej koncepcyjnie metody tego rodzaju jest **procedura sekwencyjnej selekcji cech (SFS, *Sequential Feature Selection*)**. Ma ona dwie odmiany, w których cechy są sekwencyjnie dodawane (*forward*) lub usuwane (*backword*).

Przykładowo, działanie algorytmu typu *forward* zaczyna się od wybrania pojedynczej cechy maksymalizującej zadane kryterium jakości klasyfikatora.

W następnych iteracjach dla każdej jeszcze niewybranej cechy powtarza trzy kroki:

- tymczasowe dodanie nowej cechy do podzbioru już wcześniej wybranych cech;
- zbudowanie klasyfikatora na podstawie powiększonego podzbioru cech;
- wyznaczenie jakości klasyfikatora za pomocą walidacji krzyżowej.

Po zakończeniu iteracji (tzn. po rozpatrzeniu każdej z „wolnych” cech) wybiera się tą cechę, której dołączenie do wcześniejszego podzbioru zapewniło maksymalizację kryterium jakości klasyfikatora.

Selekcja cech pod kątem klasyfikacji: metody „wbudowane” (RFE-SVM)

Metod „wbudowanych” można użyć dla klasyfikatorów, których modele zawierają parametry o wartościach odzwierciedlających istotność poszczególnych cech.

Przykładową metodą z tej grupy jest RFE-SVM (*Recursive Feature Elimination* SVM), który dokonuje selekcji na podstawie wartości wag w_i opisujących granicę decyzyjną liniowego klasyfikatora SVM.

Algorytm RFE-SVM zaczyna pracę od pełnego zestawu, a potem w każdej iteracji:

- **budowany jest liniowy klasyfikator SVM (tzn. obliczane są wagi w_i i wyraz wolny w_0);**
- **usuwana jest cecha, której odpowiada waga o najmniejszej wartości bezwzględnej.**

Uczenie maszynowe w bioinformatyce

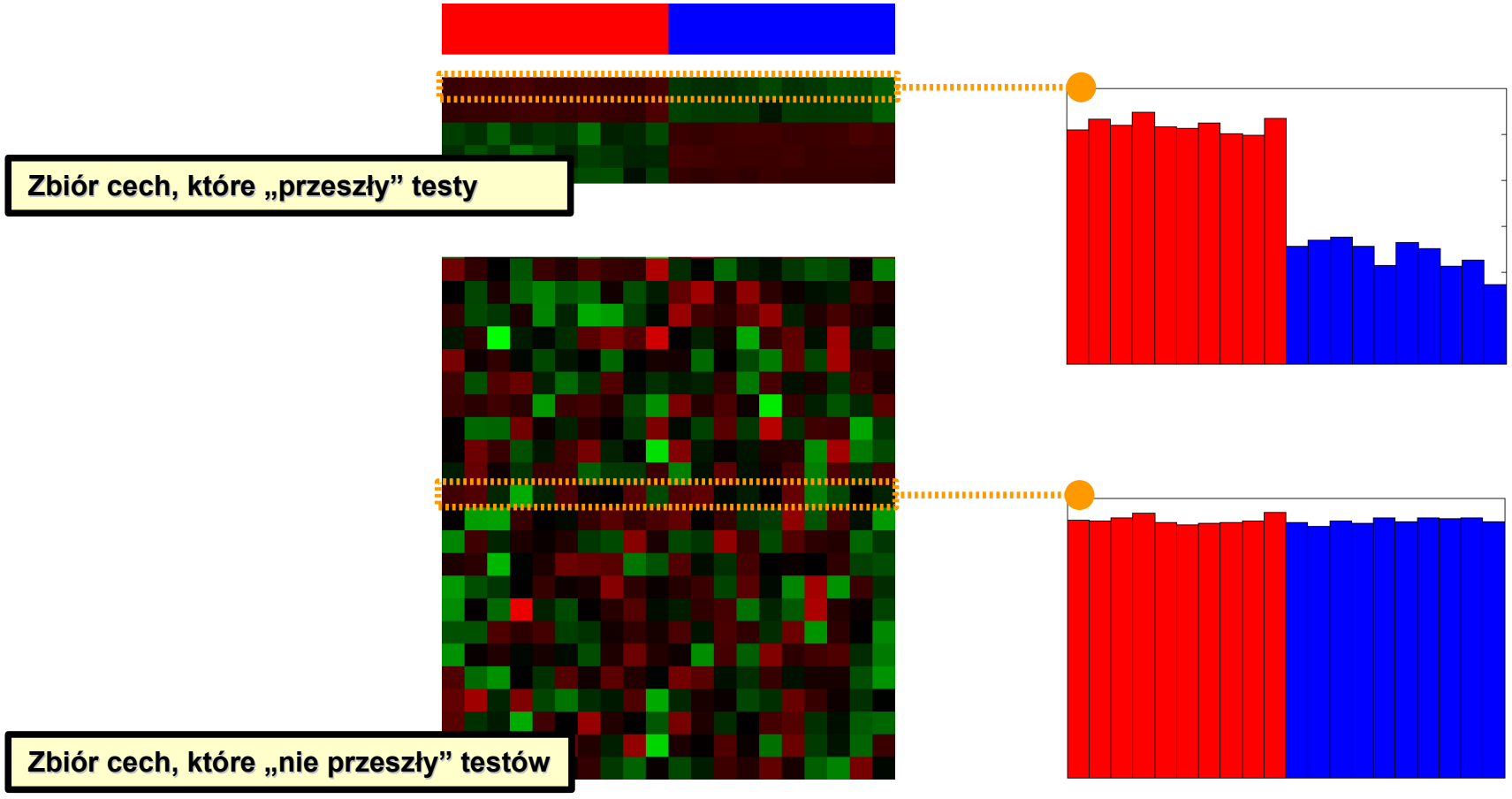
Wykład 9: selekcja cech przy użyciu testów istotności

Selekcja cech różnicujących przy użyciu testów istotności

Najprostszym podejściem do selekcji jest traktowanie każdej z cech jako niezależnej od pozostałych, zaniedbując możliwość występowania korelacji pomiędzy cechami.

W takim przypadku selekcja może odbywać się przez wykonanie osobno dla każdej cechy **statystycznego testu istotności weryfikującego hipotezę o równości wartości średnich w grupach badanych**.

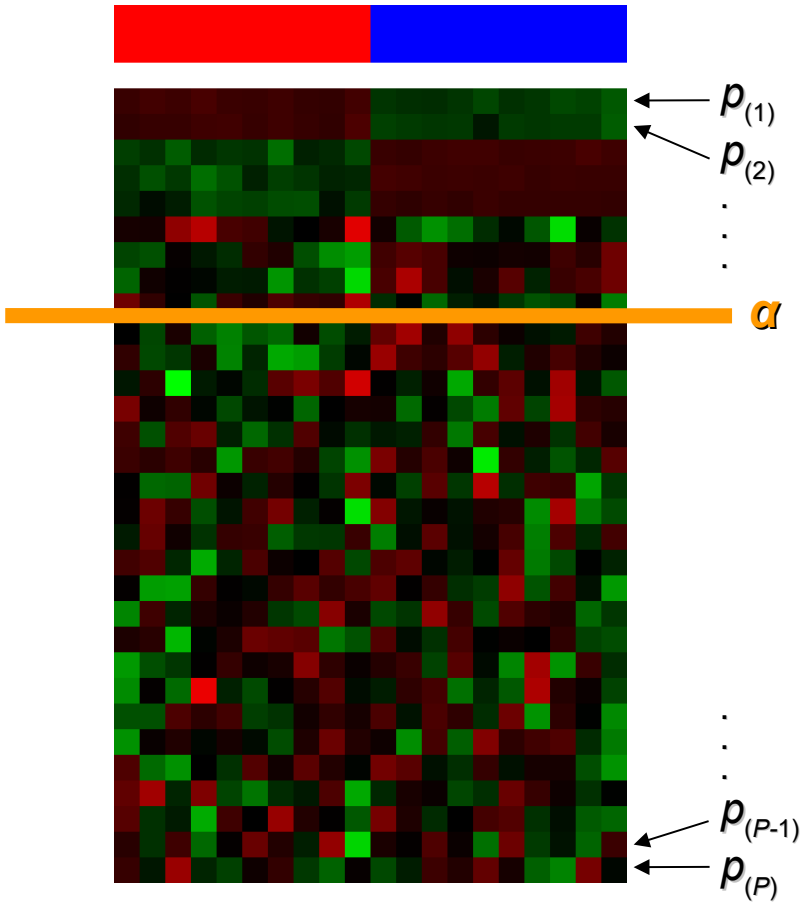
Efektem selekcji jest zbiór takich cech, dla których wyniki testów sugerują istnienie związku pomiędzy średnią ekspresją a przynależnością do grupy.



Selekcja cech różnicujących przy użyciu testów istotności

Niezależnie od rodzaju stosowanego testu badającego równość wartości średnich w grupach, procedura selekcji cech różnicujących wygląda tak samo:

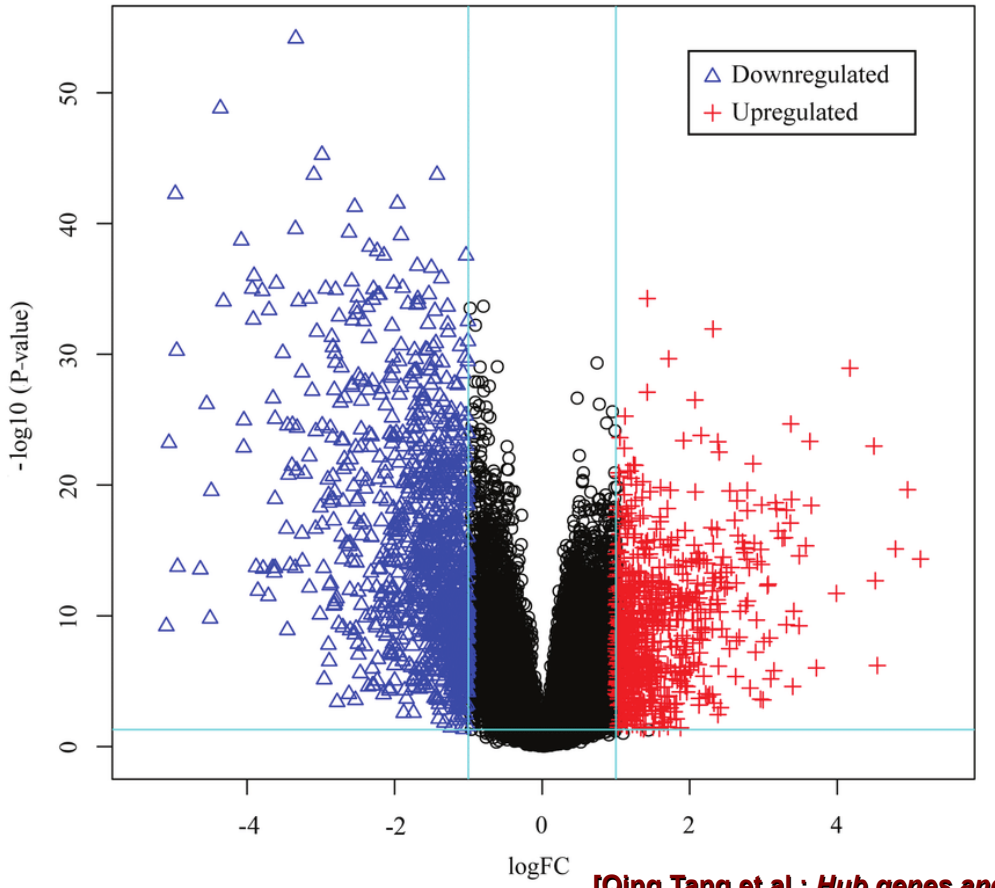
- **dla każdej z cech wykonuje się test istotności i zapamiętuje uzyskaną *p*-wartość;**
- **cechy sortowane są zgodnie z niemalejącymi *p*-wartości:** $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(P)}$;
- **pozostawia się te cechy, których *p*-wartości nie są większe od pewnego progu** α.



Selekcja cech różnicujących przy użyciu testów istotności

Zazwyczaj drugim czynnikiem uwzględnianym podczas selekcji cech jest stosunek wartości średnich w porównywanych grupach próbek (*Fold Change*, FC). Również w jego przypadku zakłada się pewne wartości progowe, po przekroczeniu których dana cecha jest uznawana za różnicującą (np. $FC \geq 2.0$ i $FC \leq 0.5$).

Popularnym sposobem graficznej prezentacji efektów działania procedury selekcji cech różnicujących z zadanymi progami *p*-wartości i *FC* jest *volcano plot*.



[Qing Tang et al.: *Hub genes and key pathways of non-small lung cancer identified using bioinformatics*. Oncology Letters, 16(2), 2018]

Uczenie maszynowe w bioinformatyce

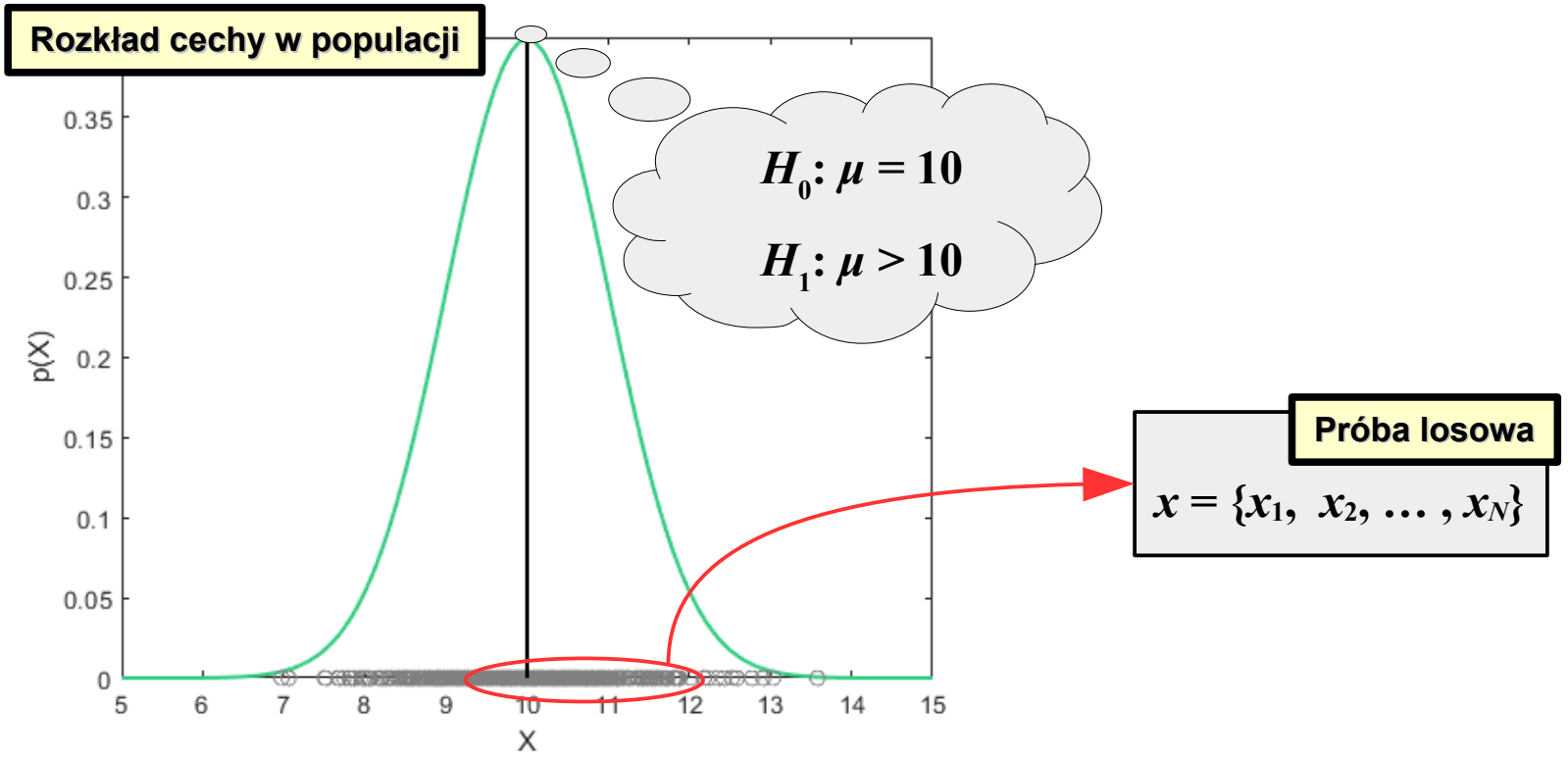
Wykład 9: selekcja cech przy użyciu testów istotności (ogólny schemat testowania hipotez statystycznych)

Weryfikacja hipotez statystycznych: definicje pojęć

Weryfikacja hipotez statystycznych jest ogólną metodologią sprawdzania hipotez dotyczących populacji na podstawie pobranej próby losowej.

Hipotezą statystyczną nazywamy dowolne przypuszczenie dotyczące właściwości rozkładu prawdopodobieństwa badanej cechy w całej populacji (dotyczące np. jego postaci funkcyjnej lub wartości jego parametrów).

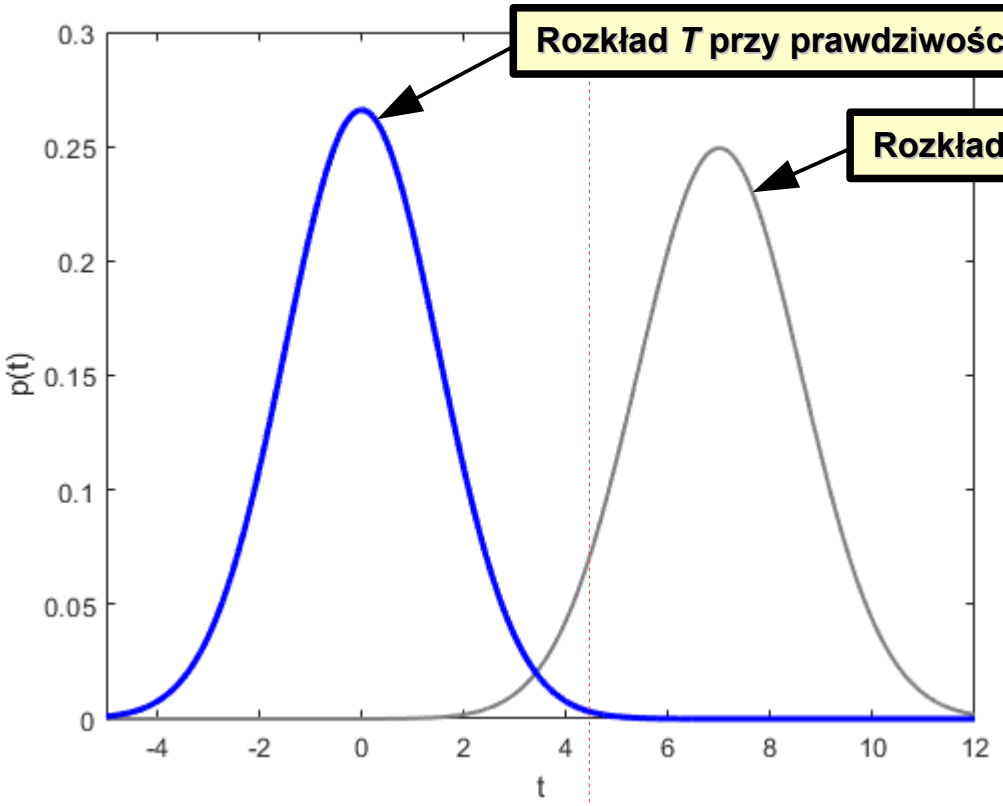
Test istotności jest postępowaniem mającym na celu weryfikację czy hipoteza statystyczna nie jest fałszywa. W toku testu weryfikacji poddawana jest hipoteza **zerowa** H_0 wobec hipotezy **alternatywnej** H_1. Na podstawie wyniku testu decydujemy o odrzuceniu hipotezy zerowej lub stwierdzamy, że nie ma do tego podstaw.



Weryfikacja hipotez statystycznych: definicje pojęć

O odrzuceniu H_0 (bądź stwierdzeniu, że nie ma do tego podstaw) wnioskujemy na podstawie wartości statystyki testowej T, wyznaczonej dla danej próby losowej.

Statystyka testowa jest funkcją elementów próby losowej, o postaci adekwatnej do weryfikowanej hipotezy. Przeprowadzenie testu wymaga znajomości rozkładu prawdopodobieństwa statystyki testowej przy prawdziwości hipotezy H_0.



Próba losowa

$$x = \{x_1, x_2, \ldots, x_N\}$$

Statystyka testowa T

$$t = T(x)$$

Weryfikacja hipotez statystycznych: definicje pojęć

Z wnioskowaniem na podstawie testu mogą być związane dwa rodzaje błędów:

- **błąd I rodzaju (wynik fałszywie pozytywny) – błąd wnioskowania polegający na odrzuceniu hipotezy H_0, gdy w rzeczywistości jest ona prawdziwa;**
- **błąd II rodzaju (wynik fałszywie negatywny) – błąd wnioskowania polegający na nieodrzuceniu hipotezy H_0, gdy w rzeczywistości jest ona fałszywa.**

Błąd I rodzaju: odrzucenie H_0 gdy jest prawdziwa

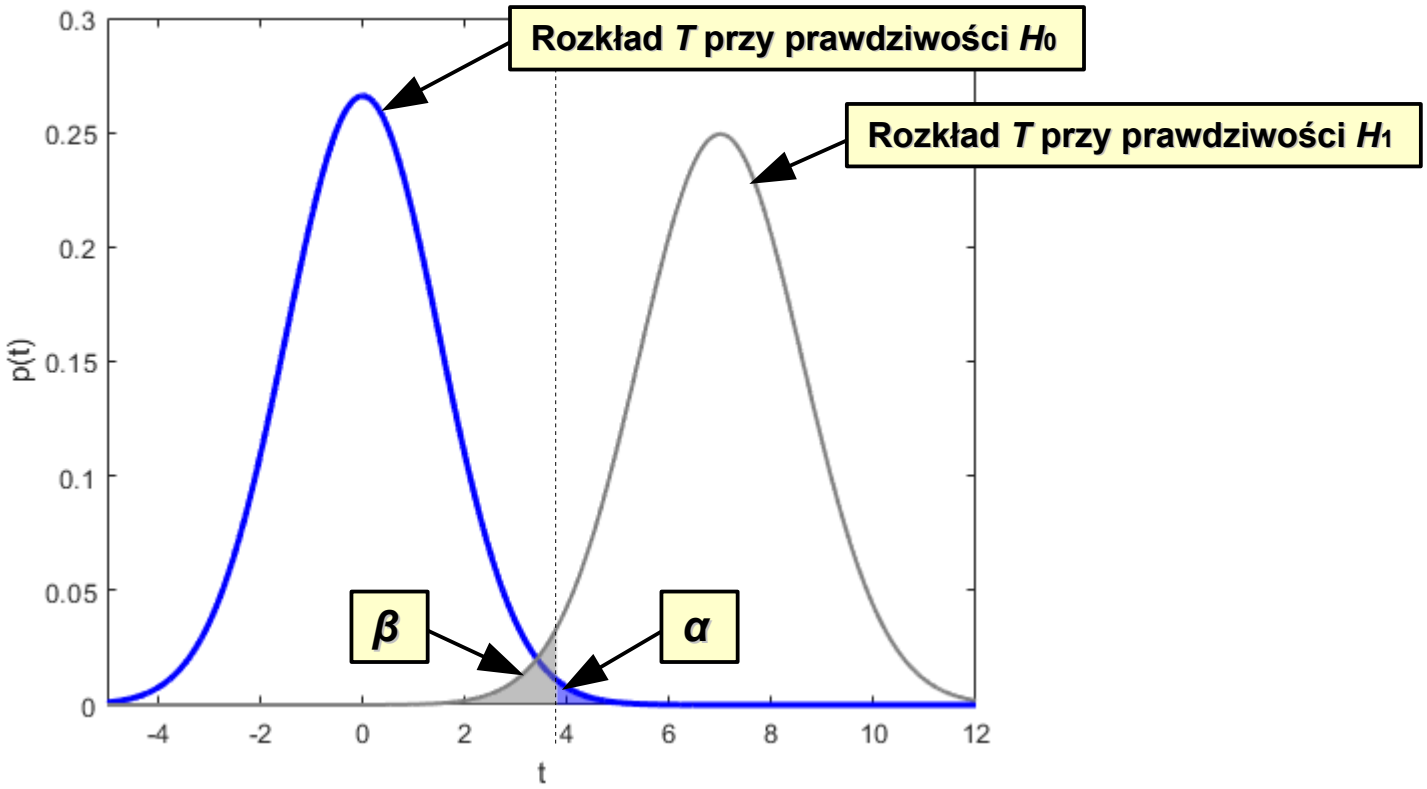
Rzeczywistość: hipoteza H_0	Wniosek o hipotezie H_0	
	nie odrzucać	odrzucić
prawdziwa	prawidłowy	nieprawidłowy
nieprawdziwa	nieprawidłowy	prawidłowy

Błąd II rodzaju: nieodrzucenie H_0 gdy jest fałszywa

Weryfikacja hipotez statystycznych: definicje pojęć

α (poziom istotności): prawdopodobieństwo popełnienia błędu I rodzaju. Przed wykonaniem testu zgadzamy się na pewną wartość α, decydującą o tym jak często test da wynik fałszywie pozytywny (np. raz na 20 przypadków gdy $\alpha = 0.05$).

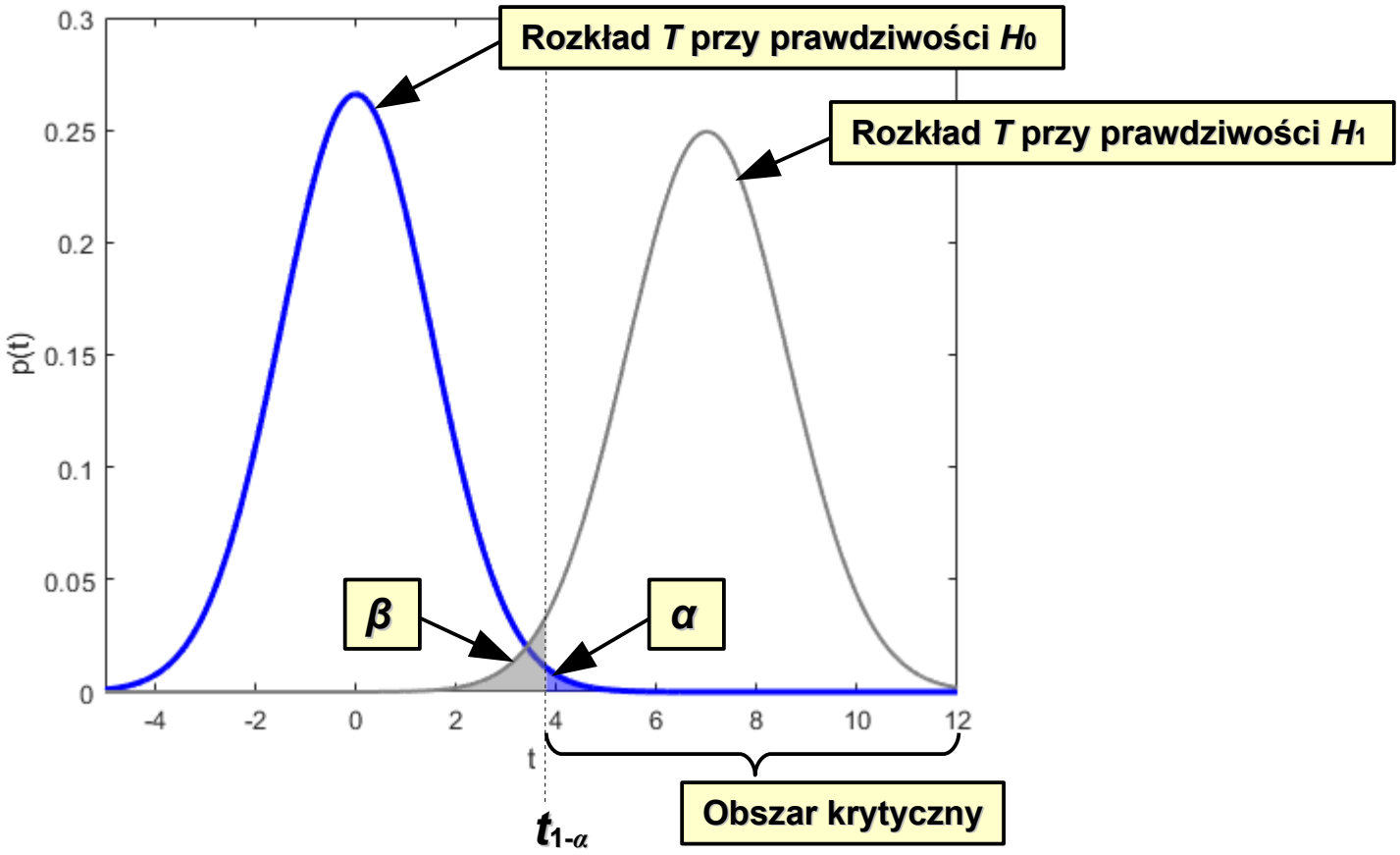
β: prawdopodobieństwo popełnienia błędu II rodzaju. Wielkość $1-\beta$ nazywana jest mocą testu. W trakcie testu nie mamy bezpośredniego wpływu na moc, ale możemy ją minimalizować przez stosowanie odpowiednich statystyk testowych i zwiększenie liczebności próby losowej.



Weryfikacja hipotez statystycznych: definicje pojęć

Wybór poziomu istotności α jednocześnie określa obszar krytyczny testu, czyli zakres wartości statystyki T, dla których odrzucamy H_0. Obszar ten reprezentuje przypadki, dla których pomimo prawdziwości H_0 obserwuje się duże wartości statystyki testowej. Przypadki te staną się fałszywie pozytywnymi wynikami testu: odrzucimy hitpotezę zerową podczas gdy w rzeczywistości jest ona prawdziwa.

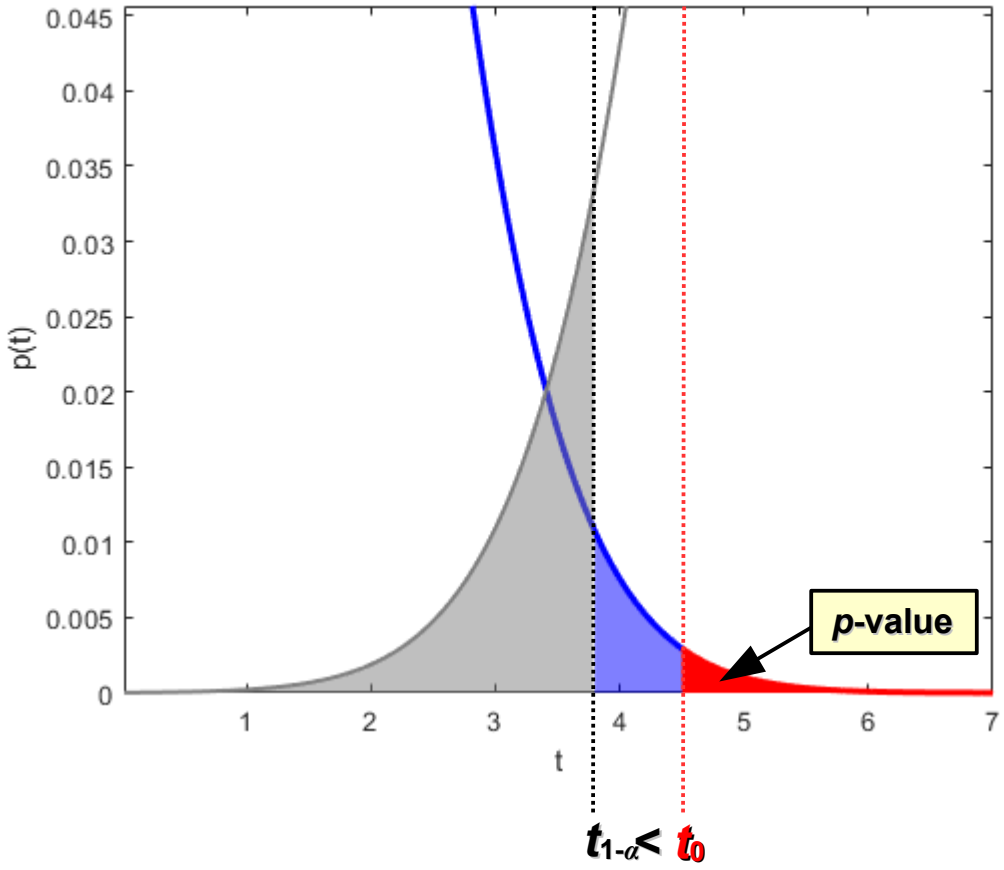
Granice obszaru krytycznego dane są przez kwantyle rozkładu statystyki testowej. Dla testu jednostronnego z tego przykładu będzie to kwantyl rzędu $1\text{-}\alpha$.



Weryfikacja hipotez statystycznych: definicje pojęć

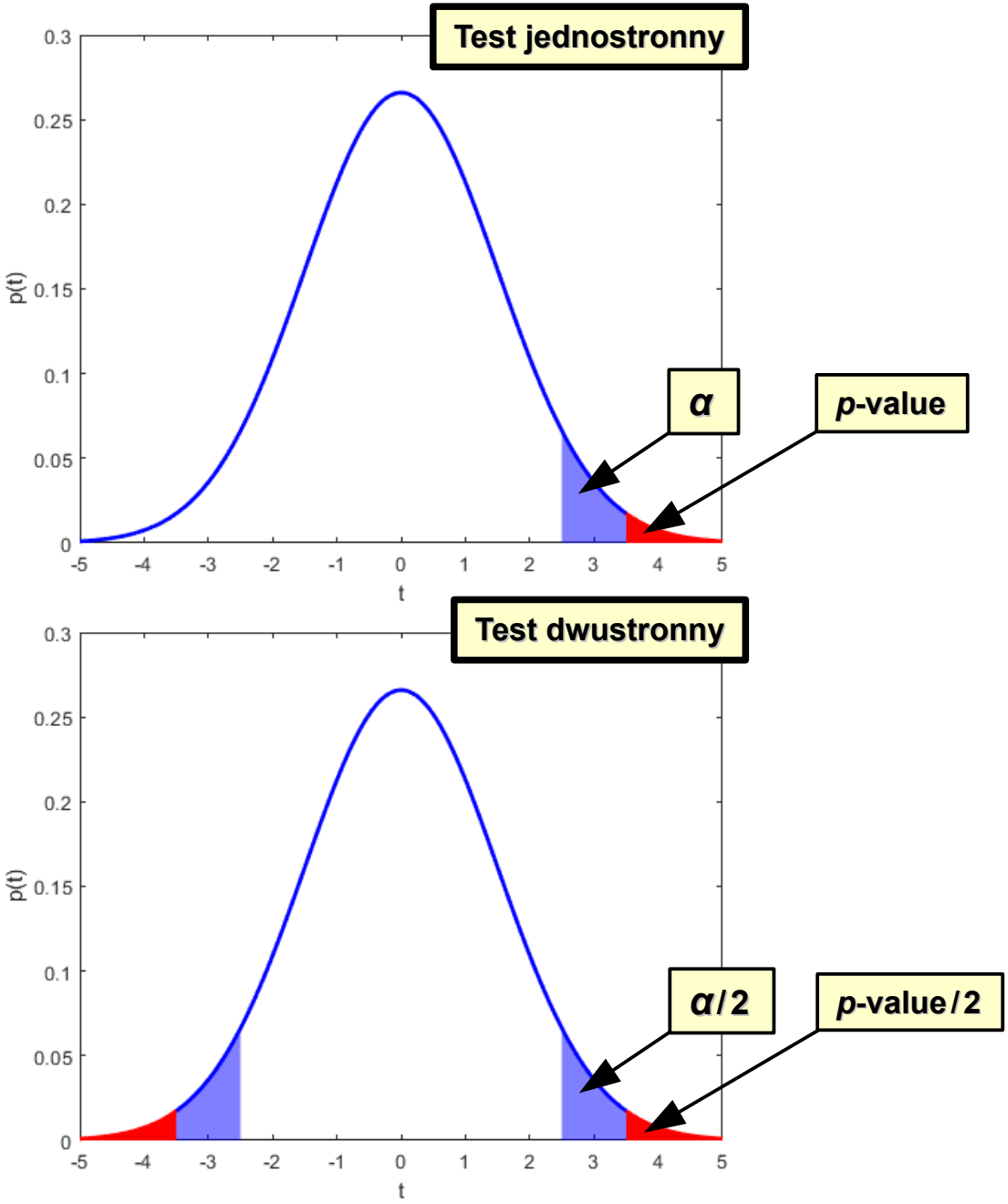
Jeżeli wartość t_0 znajdzie się w obszarze krytycznym, to odrzucamy hipotezę zerową, bo prawdopodobieństwo obserwacji tak dużej wartości przy prawdziwości H_0 jest nie większe od zadanego α. W przeciwnym razie stwierdzamy, że na poziomie istotności α nie ma podstaw do odrzucenia H_0.

Miarą odstępstwa od prawdziwości hipotezy zerowej jest *p*-wartość (*p-value*), czyli prawdopodobieństwo uzyskania wartości statystyki testowej nie mniejszej od tej wyznaczonej dla danej próby losowej przy założeniu prawdziwości H_0.



Weryfikacja hipotez statystycznych: definicje pojęć

Położenie obszaru krytycznego i znaczenie *p*-wartości są zależne od sposobu, w jaki zdefiniowana jest statystyka testowa oraz sformułowane są hipotezy.



Weryfikacja hipotez statystycznych: ogólna procedura postępowania

Standardowy przebieg testu istotności

1. **Formułujemy dwie hipotezy: zerową H_0 oraz alternatywną H_1;**

2. **Wybieramy statystykę testową T odpowiednią dla testowanej hipotezy;**

3. **Ustalamy poziomu istotności α (typowe wartości to: 0.001, 0.01, 0.05);**

4. **Liczymy wartość t_0 statystyki testowej T dla próby losowej;**

5. **Porównujemy uzyskaną wartość t_0 ze znanym rozkładem prawdopodobieństwa, który przyjmuje statystyka testowa jeżeli H_0 jest prawdziwa;**

6. **Odrzucamy H_0 jeżeli wartość t_0 znajduje się w obszarze krytycznym, wyznaczonym przez odpowiednie kwantyle rozkładu statystyki testowej przy prawdziwości hipotezy zerowej. Alternatywnie: odrzucamy H_0 jeżeli *p*-wartość jest większa od progu istotności α.**

Weryfikacja hipotez statystycznych: ogólna procedura postępowania

Standardowy przebieg testu istotności

1. Formułujemy dwie hipotezy: zerową H_0 o
2. Wybieramy statystykę testową T odpowi
3. Ustalamy poziomu istotności α (typowe
4. Liczymy wartość t_0 statystyki testowej T
5. Porównujemy uzyskaną wartość t_0 ze
   który przyjmuje statystyka testowa jeżeli
6. Odrzucamy H_0 jeżeli wartość t_0 znajduje
   przez odpowiednie kwantyle rozkładu
   hipotezy zerowej. Alternatywnie: odrzuc
   progu istotności α.

IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

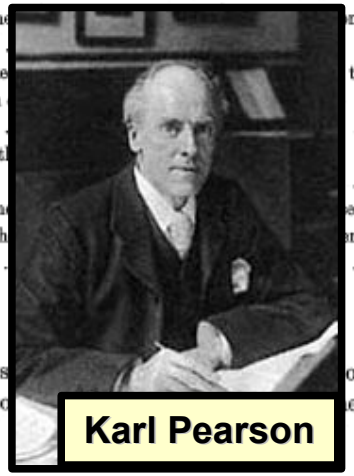
By J. NEYMAN, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. PEARSON, *Department of Applied Statistics, University College, London.*

(*Communicated by* K. PEARSON, *F.R.S.*)

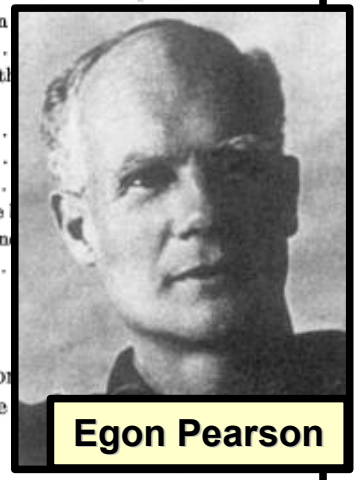
(Received August 31, 1932.—Read November 10, 1932.)

CONTENTS.

Jerzy Neyman



Karl Pearson



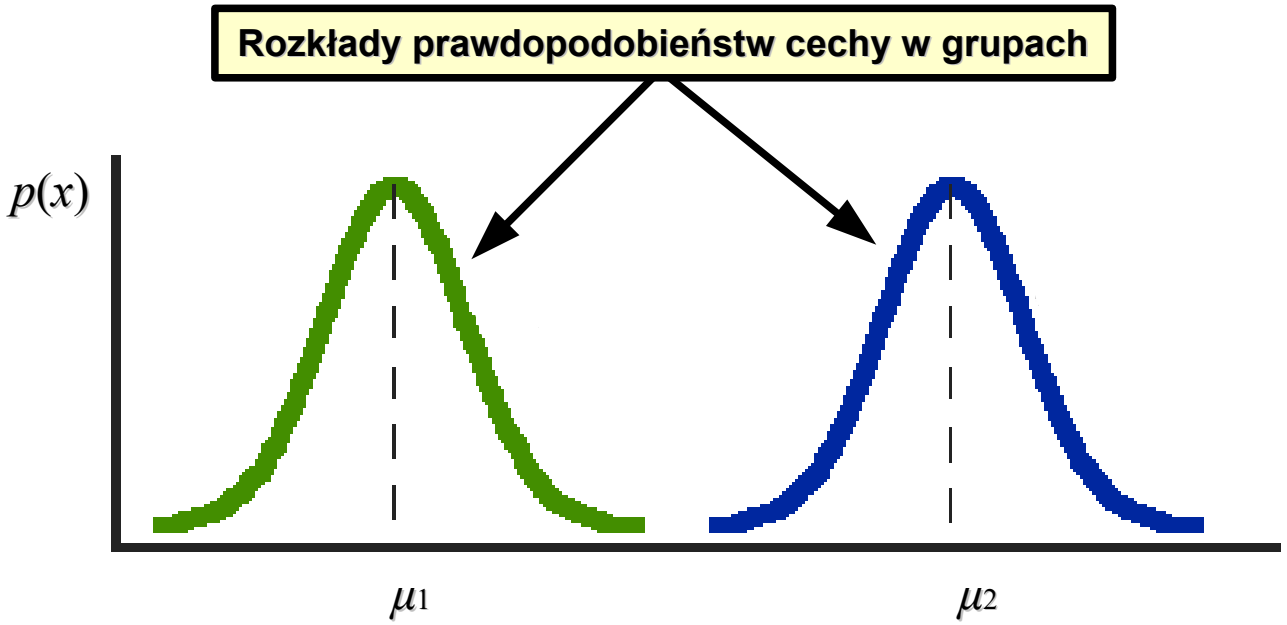
Egon Pearson

Uczenie maszynowe w bioinformatyce

Wykład 9: selekcja cech przy użyciu testów istotności (testy równości wartości średnich)

Weryfikacja hipotez statystycznych: idea testu równości średnich

Celem testu jest wnioskowanie o relacji pomiędzy wartościami średnimi μ_1 i μ_2 cechy w dwóch grupach badanych:



Wnioskowanie o nieznanych parametrach rozkładów prawdopodobieństwa cech w grupach badanych przeprowadzamy na podstawie pobranych z nich prób losowych x_1 i x_2 o liczebnościach N_1 i N_2:

$$x_1 = \{x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,N1}\}$$

$$x_2 = \{x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,N2}\}$$

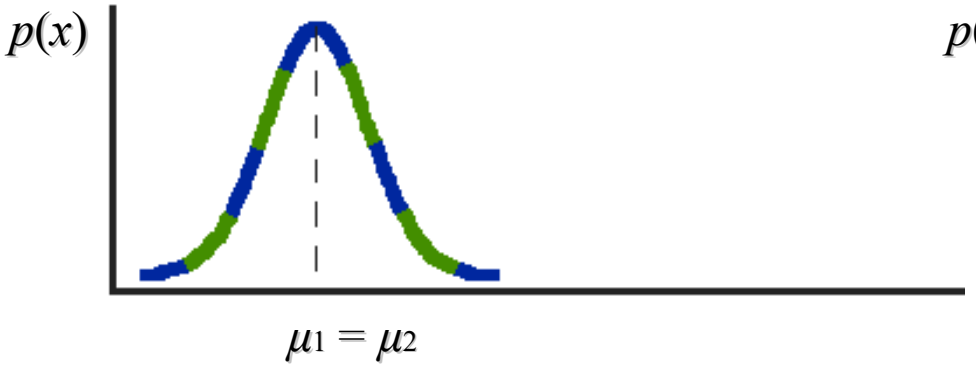
Weryfikacja hipotez statystycznych: idea testu równości średnich

Hipoteza zerowa H_0 jest tą, którą poddajemy weryfikacji i w toku testu może zostać odrzucona. Tym samym powinna ona być tak sformułowana, aby jej odrzucenie prowadziło do interesującego nas przypadku (tutaj: genów różnicujących, czyli tych o różnych średnich ekspresjach).

Dlatego też w teście badamy hipotezę zerową o równości średnich w grupach, wobec alternatywnej, mówiącej że średnie nie są równe. Należy przy tym pamiętać, że hipoteza zawsze dotyczy wartości parametrów populacji (tutaj: wartości średnich μ_1 i μ_2) a nie wartości ich estymatorów z próby – te są znane bez testowania.

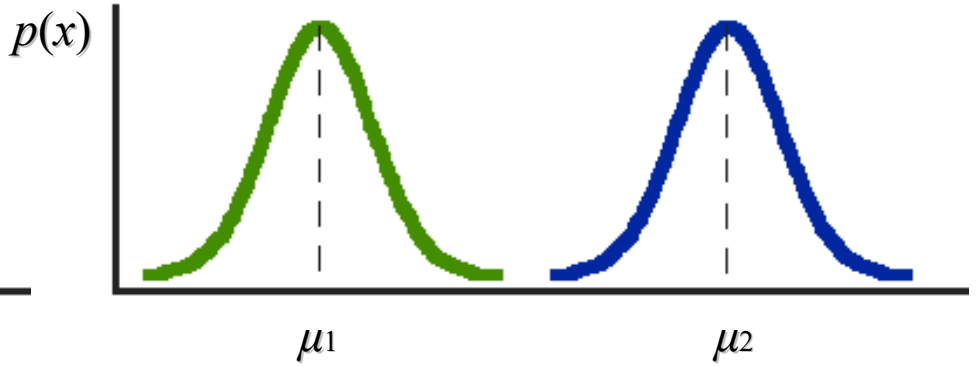
HIPOTEZA ZEROWA

$$H_0: \mu_1 = \mu_2$$



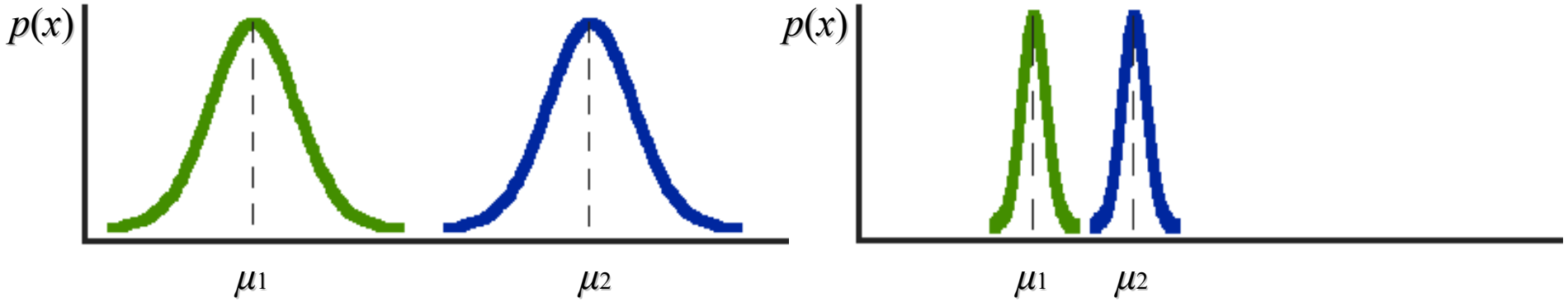
HIPOTEZA ALTERNATYWNA

$$H_1: \mu_1 \neq \mu_2$$



Weryfikacja hipotez statystycznych: idea testu równości średnich

Test dotyczy wartości średnich, jednak przy wyborze statystyki testowej nie możemy ograniczyć się jedynie do uwzględniania odstępu pomiędzy nimi, bo ten może mieć różne znaczenie w zależności od rozrzut wartości cechy w grupach:



W uproszczeniu można powiedzieć, że wartość statystyki statystyki testowej powinna być proporcjonalna do różnicy estymatorów wartości średnich, odniesionej do estymatora sumarycznego odchylenia standardowego w grupach:

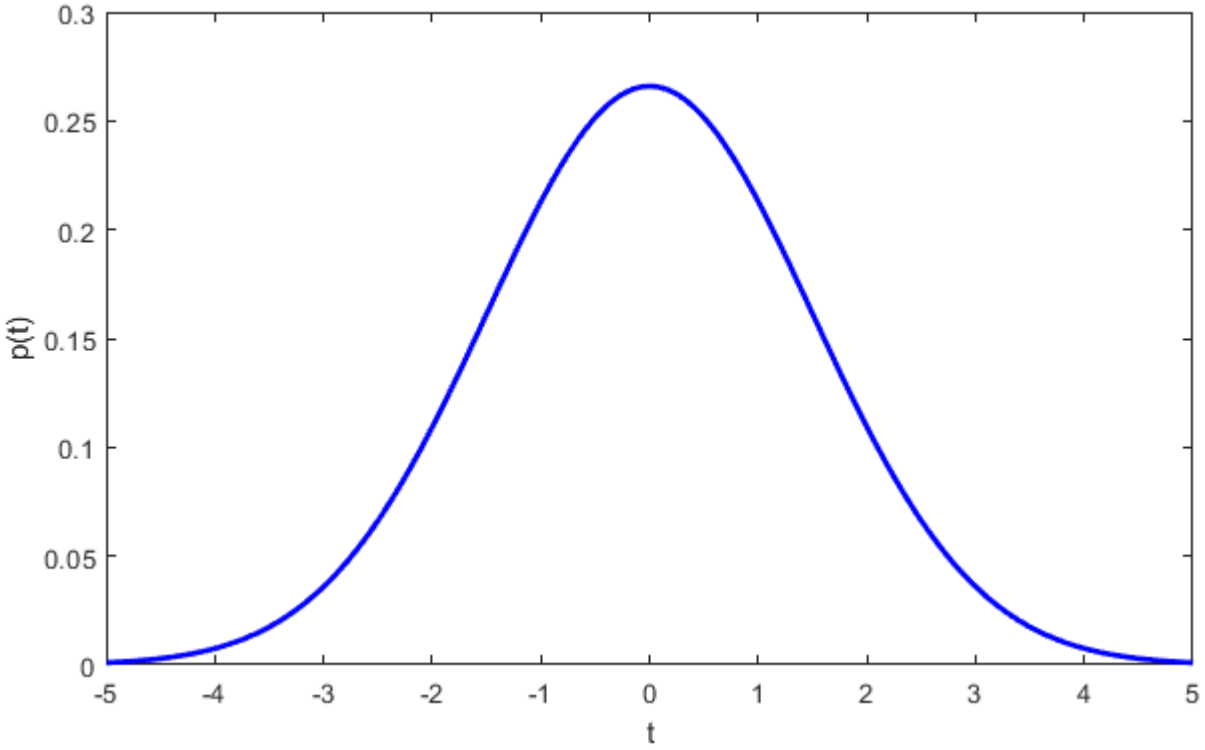
$$t \sim \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\hat{\sigma}_1^{\,2} + \hat{\sigma}_2^{\,2}}}$$

Przy tak zdefiniowanej funkcji, interesujące nas przypadki charakteryzować się będą dużymi co do modułu wartościami statystyki testowej (znak będzie zależny od relacji pomiędzy średnimi w grupach).

Weryfikacja hipotez statystycznych: idea testu równości średnich

Przeprowadzenie testu wymaga znajomości rozkładu statystyki testowej gdy H_0 jest prawdziwa. Rozkład ten zależy od dokładnej definicji statystyki, ale już na podstawie jej ogólnej postaci można wnioskować, że powinien on:

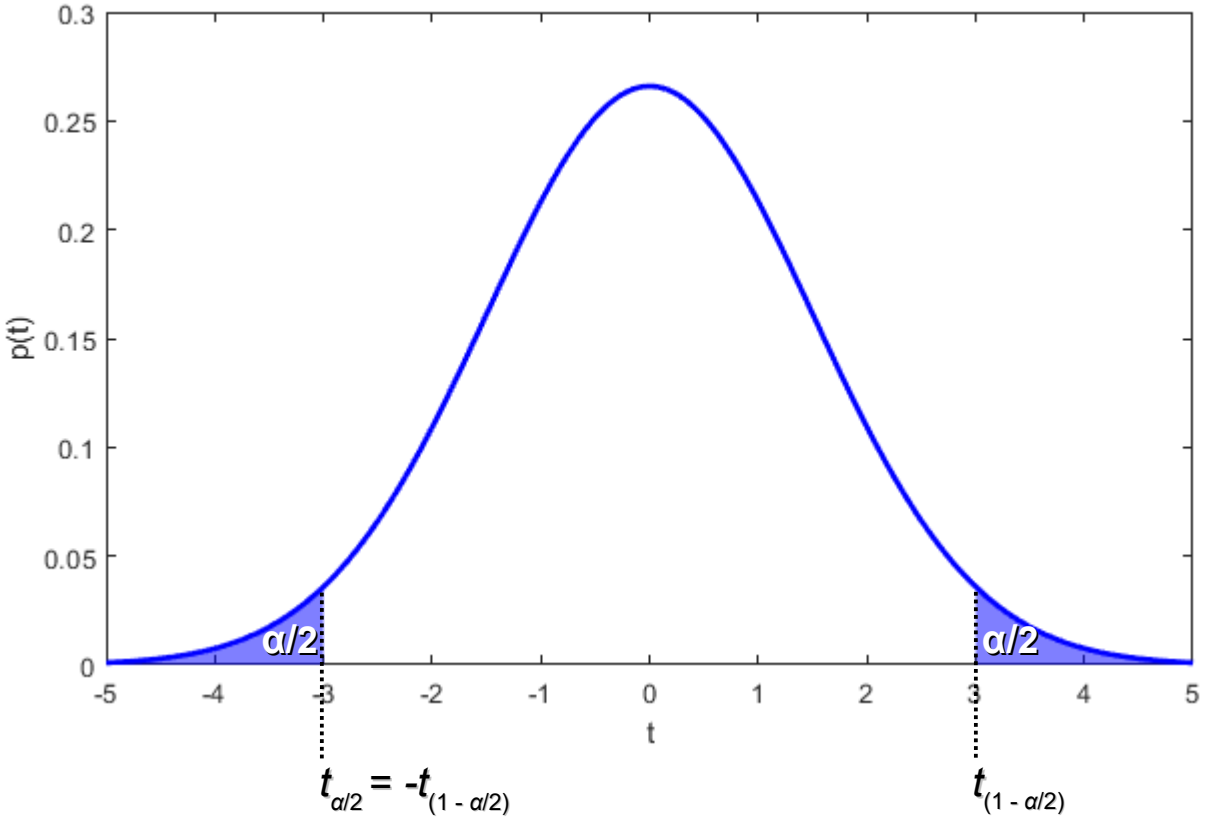
- **mieć jedno maksimum w punkcie 0, bo jeżeli średnie są sobie równe, to najbardziej prawdopodobną wartością ich różnicy będzie właśnie 0;**
- **być symetryczny, bo dodatnie i ujemne wartości różnicy średnich powinny być równie prawdopodobne.**



Weryfikacja hipotez statystycznych: idea testu równości średnich

Sformułowane tutaj hipotezy zerowa (H_0: $\mu_1 = \mu_2$) i alternatywna (H_1: $\mu_1 \neq \mu_2$) oraz symetryczna postać statystyki prowadzą do testu dwustronnego, w przypadku którego obszar krytyczny składa z dwóch przedziałów położonych symetrycznie w obu „ogonach" rozkładu statystyki testowej (dla testu jednostronnego byłby to przedział umiejscowiony tylko w jednym z „ogonów").

Pole powierzchni pod krzywą w zakresie obszaru krytycznego równe jest poziomowi istotności α. Granice przedziałów wyznaczane są przez kwantyle rzędu $\alpha/2$ oraz $1-\alpha/2$ rozkładu statystyki.



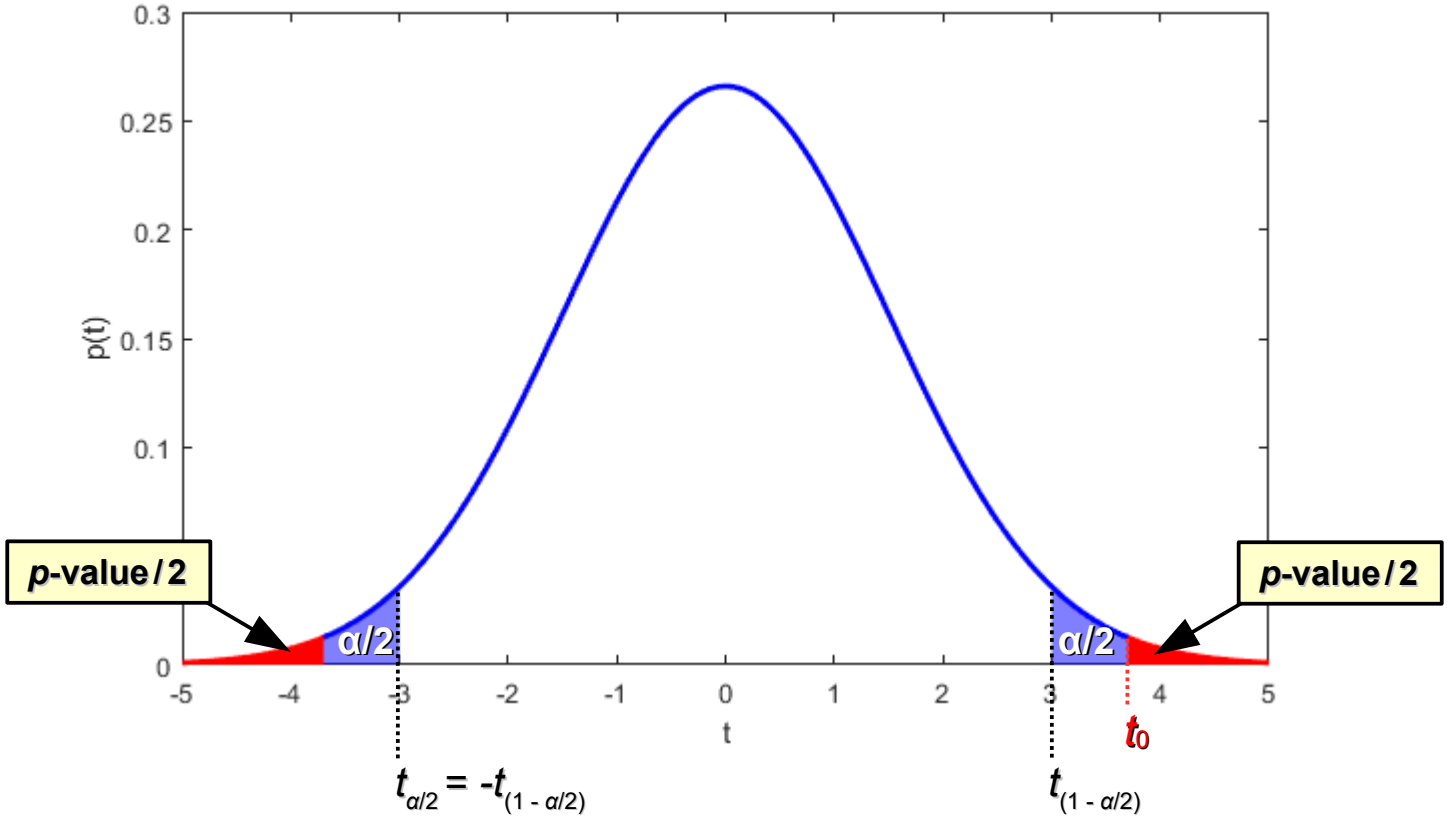
Weryfikacja hipotez statystycznych: idea testu równości średnich

Po wyznaczeniu wartości t_0 statystyki testowej T i porównujemy ją z rozkładem statystyki przy prawdziwości hipotezy H_0:

- **jeżeli $|t_0| \geq t_{(1-\alpha/2)}$, to na poziomie istotności α odrzucamy H_0**
- **jeżeli $|t_0| < t_{(1-\alpha/2)}$, to na poziomie istotności α nie ma podstaw do odrzucenia H_0**

Alternatywnie:

- **jeżeli p-value $\leq \alpha$, to na poziomie istotności α odrzucamy H_0**
- **jeżeli p-value $> \alpha$, to na poziomie istotności α nie ma podstaw do odrzucenia H_0**



Uczenie maszynowe w bioinformatyce

Wykład 9: selekcja cech przy użyciu testów istotności (testy równości wartości średnich: test *t*)

Weryfikacja hipotez statystycznych: test *t*

Przykładem testu weryfikującego równość wartości średnich w dwóch grupach jest test *t* dla prób niezależnych (*two-sample t-test*).

Założenie: próby losowe x_1 i x_2 o rozmiarach N_1 i N_2 (rozmiary prób mogą być różne) pobrane zostały niezależnie z populacji o rozkładach normalnych i nieznanych, ale równych sobie odchyleniach standardowych $\sigma_1 = \sigma_2$

Hipotezy: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Statystyka testowa:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{x_1 x_2}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad s_{x_1 x_2} = \sqrt{\frac{(N_1 - 1)s_{x_1}^2 + (N_2 - 1)s_{x_2}^2}{N_1 + N_2 - 2}}$$

gdzie:

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} \qquad s_{x_i}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(x_{ij} - \bar{x}_i)^2$$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *t*-Studenta o $df = N_1 + N_2 - 2$ stopniach swobody.

Obszar krytyczny: $K = \left(-\infty, -t_{1-\alpha/2}\right] \cup \left[t_{1-\alpha/2}, +\infty\right)$

	MATLAB	R
two-sample t-test:	`ttest2`	`t.test`

Weryfikacja hipotez statystycznych: test *t*

Przykładem testu weryfikującego równo
test *t* dla prób niezależnych (*two-sample*

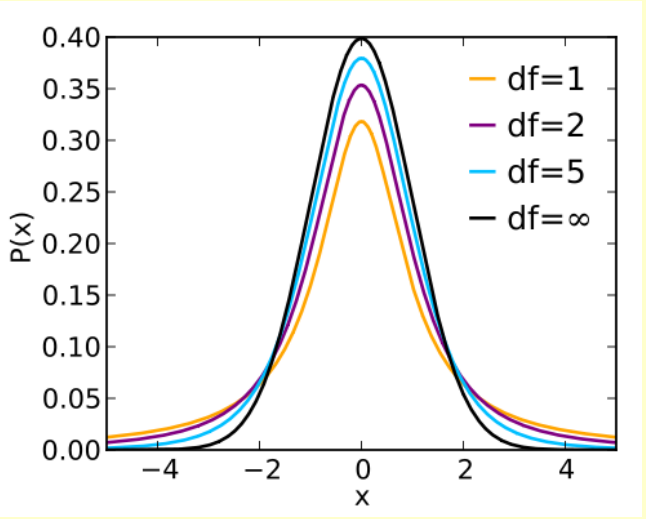
Założenie: próby losowe x_1 i x_2 o rozmia
pobrane zostały niezależnie
nieznanych, ale równych sob

Hipotezy: $H_0: \mu_1 = \mu_2$

Statystyka testowa:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{x_1 x_2}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

Unimodalny, symetryczny rozkład posiadający jeden parametr, nazywanym stopniem swobody (*df*).



Kształt funkcji gęstości prawdopodobieństwa podobny do standardowego rozkładu normalnego $N(0, 1)$, ale z grubszymi „ogonami". Dla dużych stopni swobody może być przybliżany rozkładem $N(0, 1)$.

gdzie:

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} \qquad s_{x_i}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(x_{ij} - \bar{x}_i)^2$$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *t*-Studenta o $df = N_1 + N_2 - 2$ stopniach swobody.

Obszar krytyczny:

$$K = \left(-\infty, -t_{1-\alpha/2}\right] \cup \left[t_{1-\alpha/2}, +\infty\right)$$

	MATLAB	R
two-sample t-test:	`ttest2`	`t.test`

Weryfikacja hipotez statystycznych: test *t*

Przykładem testu weryfikującego równość wartości śred
test *t* dla prób niezależnych (*two-sample t-test*).

Założenie: próby losowe x_1 i x_2 o rozmiarach N_1 i N_2 (rozm
pobrane zostały niezależnie z populacji o
nieznanych, ale równych sobie odchyleniach s

Hipotezy: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Statystyka testowa:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{x_1 x_2}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$$s_{x_1 x_2} = \sqrt{($$

gdzie:

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$$

$$s_{x_i}^2 = \frac{1}{N_i -}$$

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

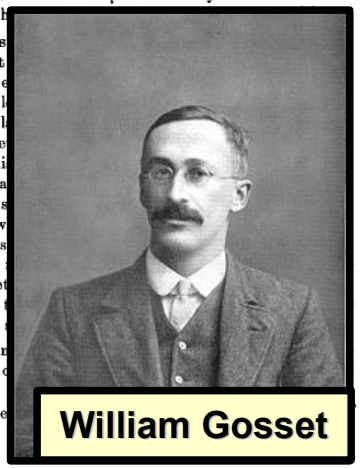
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as th

If the number of experiments as to the value of the mean, but uncertainty:—(1) owing to the "e of experiments deviates more or l (2) the sample is not sufficiently l of individuals. It is usual, howe a very large number of cases, thi sample will give no real informa deviates from normality: since s better to work with a curve w properties are well known. This paper, so that its conclusions are to be normally distributed; yet normality must be very extreme solely with the first of these two

The usual method of determin tion lies within a given distance distribution about the mean of $s/\sqrt{n}$, where s is the standard de the probability integral.

Biometrika VI



William Gosset

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *t*-Studenta o $df = N_1 + N_2 - 2$ stopniach swobody.

Obszar krytyczny: $K = \left(-\infty, -t_{1-\alpha/2}\right] \cup \left[t_{1-\alpha/2}, +\infty\right)$

	MATLAB	R
two-sample t-test:	`ttest2`	`t.test`

Weryfikacja hipotez statystycznych: test *t*

Modyfikacją testu *t* dla dwóch prób niezależnych jest test *t* z poprawką Welcha (czasem zwany testem Satterthwaite'a), w którym nie zakłada się równości wariancji.

Założenie: próby losowe x_1 i x_2 o rozmiarach N_1 i N_2 (rozmiary prób mogą być różne) pobrane zostały niezależnie z populacji o rozkładach normalnych

Hipotezy: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Statystyka testowa:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_{x_1}^2}{N_1} + \frac{s_{x_2}^2}{N_2}}}$$

gdzie: $\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$ $s_{x_i}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(x_{ij} - \bar{x}_i)^2$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma w przybliżeniu rozkład *t*-Studenta o stopniu swobody równym:

$$df = \frac{(s_{x_1}^2/N_1 + s_{x_2}^2/N_2)^2}{(s_{x_1}^2/N_1)^2/(N_1 - 1) + (s_{x_2}^2/N_2)^2/(N_2 - 1)}$$

Obszar krytyczny: $K = \left(-\infty, -t_{1-\alpha/2}\right] \cup \left[t_{1-\alpha/2}, +\infty\right)$

	MATLAB	R
two-sample t-test:	`ttest2`	`t.test`

Weryfikacja hipotez statystycznych: test *t*

W przypadku prób zależnych (np. gdy ta sama grupa pacjentów badana jest przed i po leczeniu) stosuje się test *t* dla prób zależnych (*paired samples t-test*).

Założenie: badamy dwukrotnie próbę losową x o rozmiarze N pobraną populacji o rozkładzie normalnym

Hipotezy: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Statystyka testowa:

$$t = \frac{\bar{x}_d}{s_d / \sqrt{N}}$$

gdzie:

$$\bar{x}_d = \frac{1}{N}\sum_{j=1}^{N} x_{dj} = \frac{1}{N}\sum_{j=1}^{N} \left(x_{1j} - x_{2j}\right)$$

$$s_d = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N} \left(x_d - \bar{x}_d\right)^2} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N} \left(\left(x_{1j} - x_{2j}\right) - \bar{x}_d\right)^2}$$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *t*-Studenta o $df = N - 1$ stopniach swobody.

Obszar krytyczny:

$$K = \left(-\infty, -t_{1-\alpha/2}\right] \cup \left[t_{1-\alpha/2}, +\infty\right)$$

	MATLAB	R
paired t-test:	`ttest`	`t.test`

Uczenie maszynowe w bioinformatyce

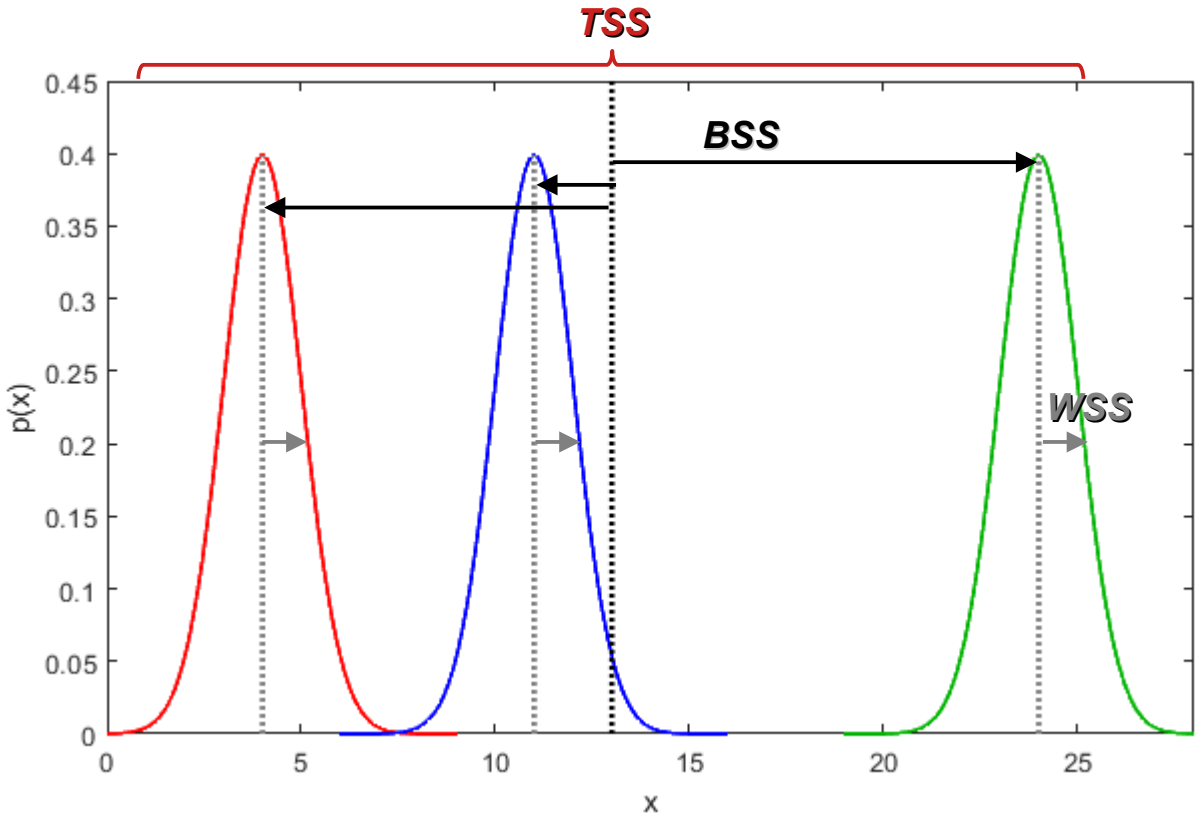
Wykład 9: selekcja cech przy użyciu testów istotności (testy równości wartości średnich: ANOVA)

Weryfikacja hipotez statystycznych: test ze statystyką *F*

Test *t* ograniczony jest do porównywania dwóch grup. W przypadku J grup ($J \geq 2$), reprezentowanych przez J prób losowych o liczebnościach N_j ($N = N_1 + N_2 + \ldots + N_J$) można zastosować test ze statystyką *F*, wywodzący się z analizy wariancji (ANOVA).

Całkowitą zmienność próby losowej (TSS, *Total Sum of Squares*), zdefiniowaną jako suma kwadratów różnic między wartościami poszczególnych elementów x_{ji} a średnią ogólną całej próby, można zapisać jako sumę dwóch składników:

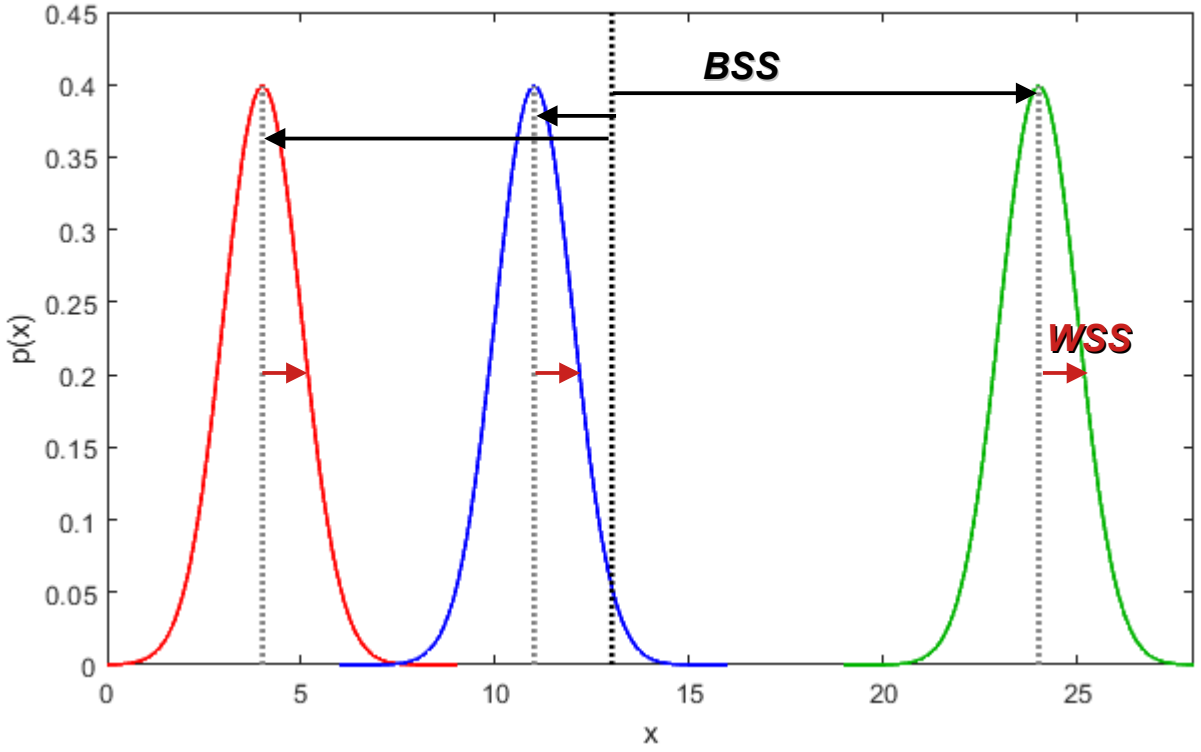
$$TSS = \sum_{j=1}^{J}\sum_{i=1}^{N_j}\left(x_{ji} - \bar{x}\right)^2 = \sum_{j=1}^{J}\sum_{i=1}^{N_j}\left(x_{ji} - \bar{x}_j\right)^2 + \sum_{j=1}^{J} N_j\left(\bar{x}_j - \bar{x}\right)^2 = WSS + BSS$$



Weryfikacja hipotez statystycznych: test ze statystyką *F*

Suma kwadratów odchyleń wartości od średnich wewnątrz grup (WSS, *Within Sum of Squares*) jest miarą zmienności w obrębie grup badanych:

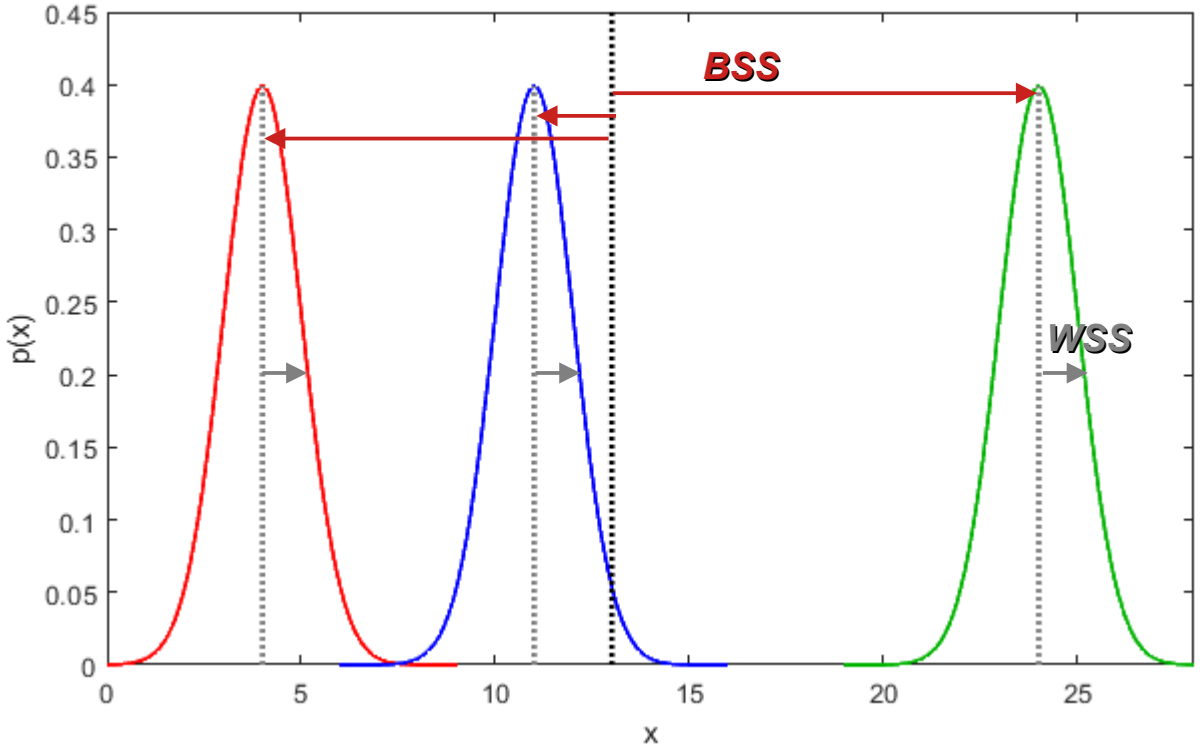
$$WSS = \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(x_{ji} - \bar{x}_j \right)^2$$



Weryfikacja hipotez statystycznych: test ze statystyką *F*

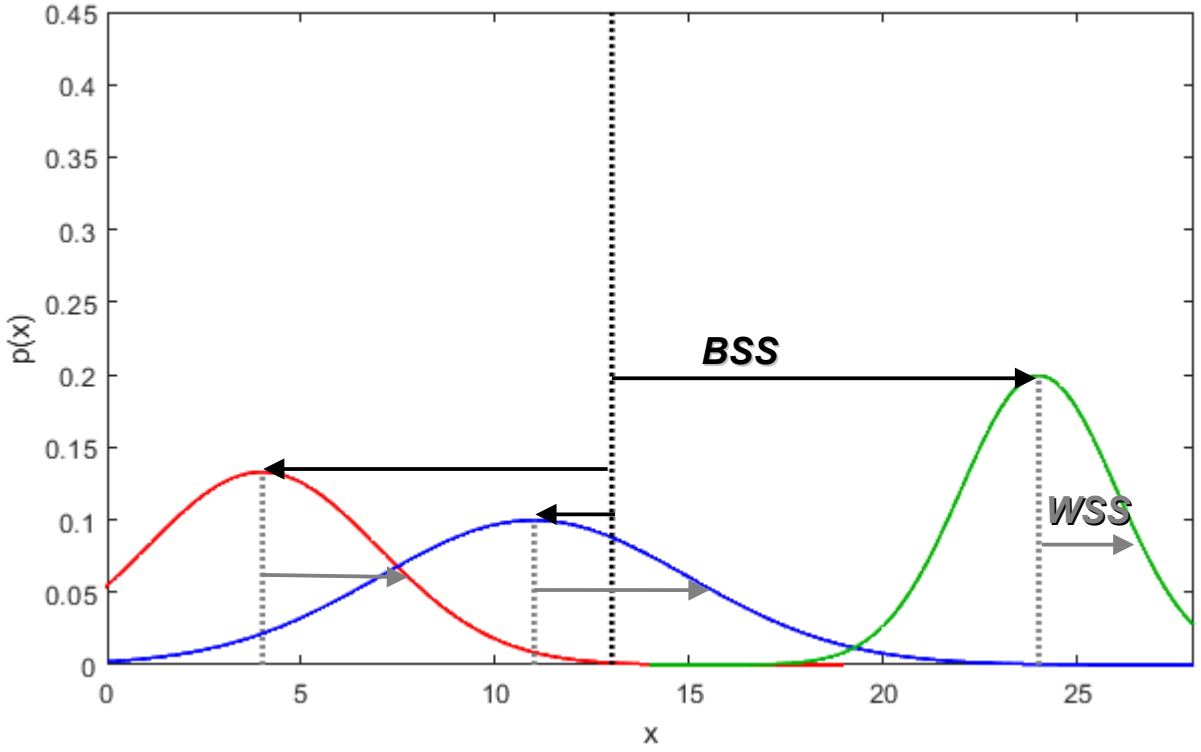
Miarą zmienności pomiędzy grupami jest suma kwadratów różnic wartości średnich grup i ogólnej wartości średniej próby (BSS, *Between Sum of Squares*):

$$BSS = \sum_{j=1}^{J} N_j \left(\bar{x}_j - \bar{x}\right)^2$$



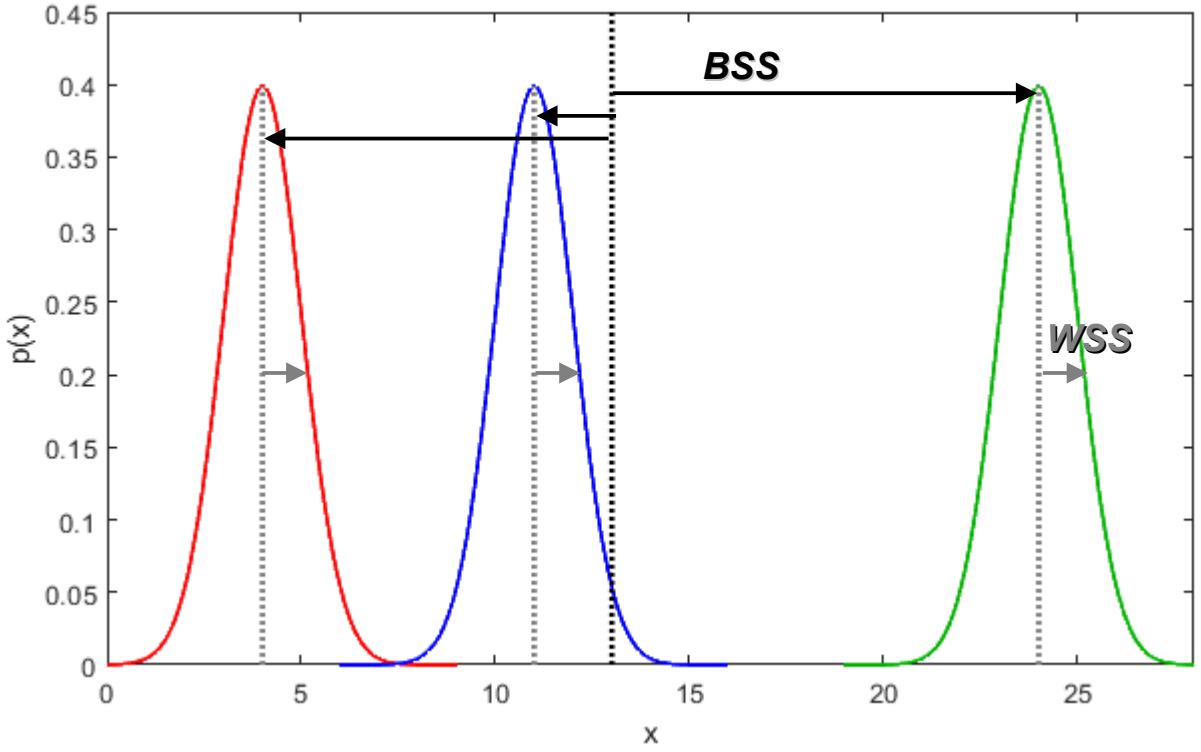
Weryfikacja hipotez statystycznych: test ze statystyką *F*

Podstawą testu jest założenie, że jeżeli próby zostały pobrane z tej samej populacji, stosunek rozrzutu średnich grup (*BSS*) do rozrzutu próbek w grupach (*WSS*) powinien być mały. Natomiast gdy *BSS* / *WSS* jest duży, to próby pobrano z różnych populacji.



Weryfikacja hipotez statystycznych: test ze statystyką *F*

Podstawą testu jest założenie, że jeżeli próby zostały pobrane z tej samej populacji, stosunek rozrzutu średnich grup (*BSS*) do rozrzutu próbek w grupach (*WSS*) powinien być mały. Natomiast gdy *BSS* / *WSS* jest duży, to próby pobrano z różnych populacji.



Weryfikacja hipotez statystycznych: test ze statystyką *F*

Założenie: próby losowe x_j o rozmiarach N_j (rozmiary prób mogą być różne) pobrane zostały niezależnie z populacji o rozkładach normalnych i nieznanych, ale równych sobie odchyleniach standardowych $\sigma_1 = \sigma_2 = \ldots = \sigma_J$

Hipotezy:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_J \qquad H_1 : \bigvee_{i,j:i\neq j} \mu_i \neq \mu_j$$

Statystyka testowa:

$$F = \frac{BSS/(J-1)}{WSS/(N-J)} = \frac{(N-J)\sum_{j=1}^{J} N_j \left(\bar{x}_j - \bar{x}\right)^2}{(J-1)\sum_{j=1}^{J}\sum_{i=1}^{N_j}\left(x_{ji} - \bar{x}_j\right)^2}$$

gdzie:

$$\bar{x}_j = \frac{1}{N_j}\sum_{j=1}^{N_j} x_{ji} \qquad \bar{x} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} x_{ji} = \frac{1}{N}\sum_{j=1}^{J} N_j \bar{x}_j$$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *F*-Snedecora o stopniach swobody $f_1 = J - 1$ oraz $f_2 = N - J$.

Obszar krytyczny: $K = \left[f_{1-\alpha}, +\infty\right)$

	MATLAB	R
ANOVA:	`anova1`	`oneway.test`

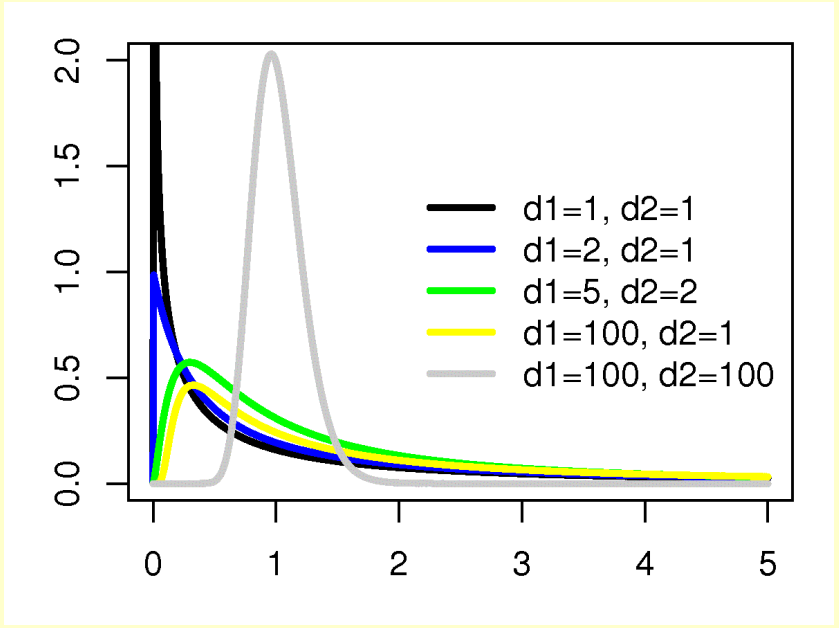
Weryfikacja hipotez statystycznych: test ze statystyką *F*

Założenie: próby losowe x_j o rozmiarac
zostały niezależnie z popula
równych sobie odchyleniac

Hipotezy: $H_0 : \mu_1 = \mu_2 =$

Statystyka testowa: $F = \dfrac{BSS/(J-1}{WSS/(N-J}$

Unimodalny, niesymetryczny rozkład o dwóch parametrach, nazywanych stopniami swobody (f_1 i f_2).



gdzie: $\bar{x}_j = \dfrac{1}{N_j}\sum_{j=1} x_{ji}$ $\quad \bar{x} = \dfrac{1}{N}\sum_{j=1}\sum_{i=1} x_{ji} = \dfrac{1}{N}\sum_{j=1} N_j \bar{x}_j$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *F*-Snedecora o stopniach swobody $f_1 = J - 1$ oraz $f_2 = N - J$.

Obszar krytyczny: $K = \left[f_{1-\alpha}, +\infty \right)$

	MATLAB	R
ANOVA:	`anova1`	`oneway.test`

Weryfikacja hipotez statystycznych: test ze statystyką *F*

Założenie: próby losowe x_j o rozmiarac
zostały niezależnie z popula
równych sobie odchyleniac

UWAGA: w przypadku porównywania wartości średnich dwóch grup ($J = 2$), analiza wariancji prowadzi do takich samych wyników, jak *t*-test (z obu testów uzyskuje się jednakowe *p*-wartości).

Hipotezy: $$H_0 : \mu_1 = \mu_2 = \ldots = \mu_J$$

Statystyka testowa: $$F = \frac{BSS/(J-1)}{WSS/(N-J)} = \frac{(N-J)\sum_{j=1}^{J} N_j \left(\bar{x}_j - \bar{x}\right)^2}{(J-1)\sum_{j=1}^{J}\sum_{i=1}^{N_j}\left(x_{ji} - \bar{x}_j\right)^2}$$

gdzie: $$\bar{x}_j = \frac{1}{N_j}\sum_{j=1}^{N_j} x_{ji} \qquad \bar{x} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} x_{ji} = \frac{1}{N}\sum_{j=1}^{J} N_j \bar{x}_j$$

Przy prawdziwości hipotezy zerowej H_0 statystyka testowa ma rozkład *F*-Snedecora o stopniach swobody $f_1 = J - 1$ oraz $f_2 = N - J$.

Obszar krytyczny: $K = \left[f_{1-\alpha}, +\infty \right)$

	MATLAB	R
ANOVA:	`anova1`	`oneway.test`

Uczenie maszynowe w bioinformatyce

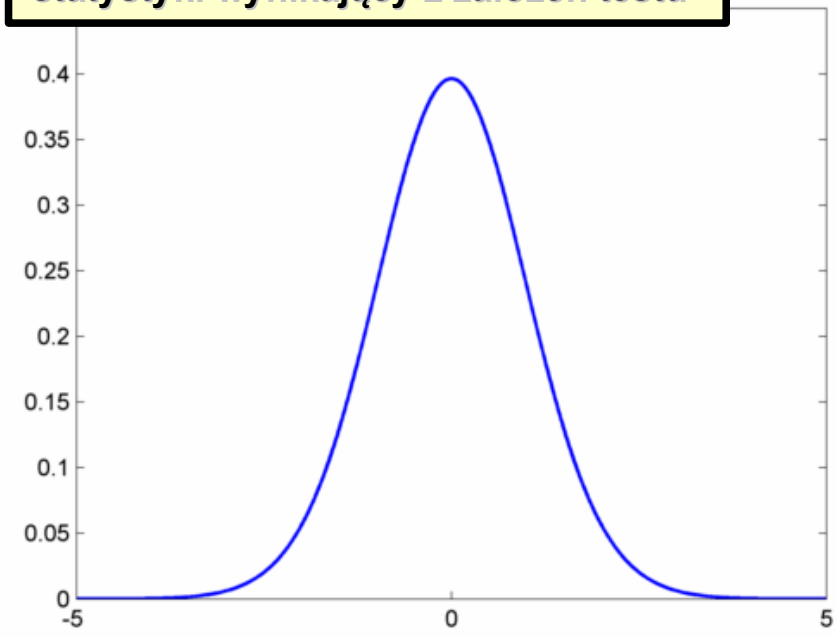
Wykład 9: selekcja cech przy użyciu testów istotności (testy równości wartości średnich: resampling)

Weryfikacja hipotez statystycznych: resamplingowe testy istotności

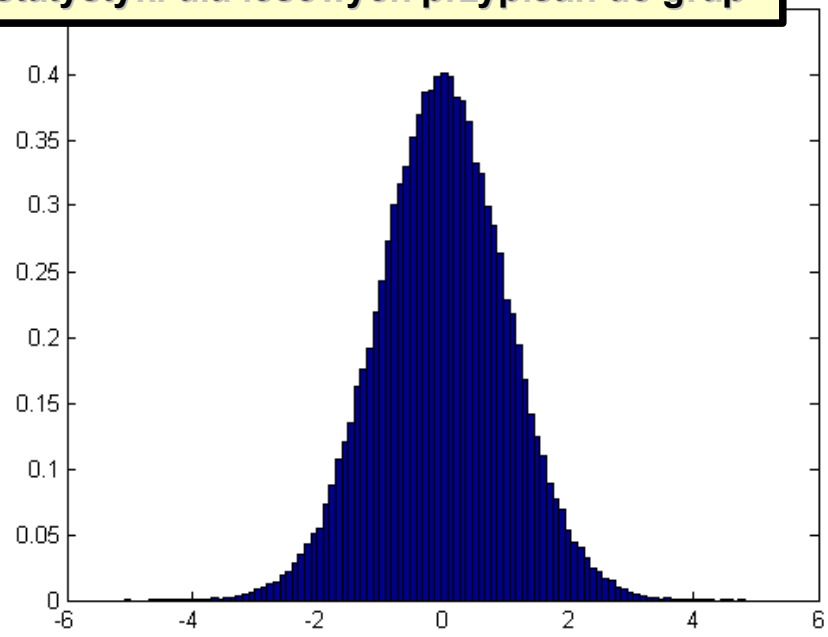
Wadą klasycznych testów istotności jest wymóg znajomości rozkładu statystyki testowej przy prawdziwości hipotezy zerowej H_0. Rozkłady te zostały określone dla wielu rodzajów statystyk, jednak ich stosowalność jest zwykle uzależniona od spełnienia pewnych założeń dotyczących próby losowej.

W ogólnym przypadku założenia leżące u podstaw klasycznych testów nie zawsze muszą być spełnione. Dlatego też często spotykanym podejściem jest użycie testów resamplingowych, w których *p*-wartości określane są w oparciu o empiryczne rozkłady statystyk testowych. Rozkłady te budowane są na podstawie wartości statystyki testowej obliczanych po wielokrotnych losowych zmianach sposobu przypisania elementów próby do grup badanych.

Klasyczne testy: teoretyczny rozkład statystyki wynikający z założeń testu



Testy resamplingowe: estymowany rozkład statystyki dla losowych przypisań do grup



Weryfikacja hipotez statystycznych: resamplingowe testy istotności

Idea testów resamplingowych wynika bezpośrednio z definicji *p*-wartości (czyli prawdopodobieństwa uzyskania przy prawdziwości H_0 wartości statystyki testowej nie mniejszej od obserwowanej dla próby losowej), przy czym prawdopodobieństwo zastąpione jest w tym przypadku częstością.

Wyznaczanie rozkładu statystyki T przy prawdziwości H_0 polega na wykonaniu B losowych „przetasowań" sposobu przypisania elementów próby do grup badanych. Każdemu przetasowaniu towarzyszy wyznaczenie wartości t_i statystyki testowej. Postać samej statystyki jest w zasadzie dowolna, ale często stosuje się funkcje podobne do tych używanych w klasycznych testach istotności, np. zaczerpnięte z testu *t*-Studenta (dla dwóch grup) lub analizy wariancji (dla większej liczby grup).

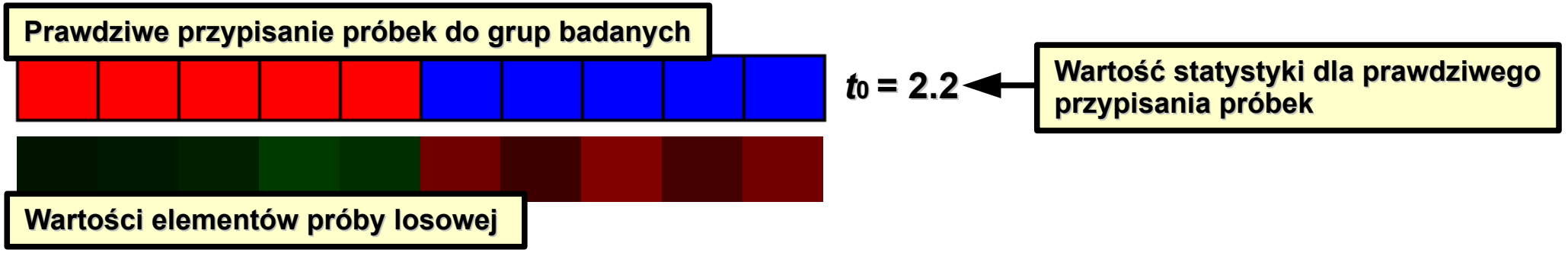
W najprostszym przypadku, estymatę *p*-wartości testu można wyznaczyć jako częstość przypadków, dla których wartości bezwzględne t_i są nie mniejsze od wartości bezwzględnej t_0 obliczonej dla prawidłowego przypisania próbek do grup:

$$p = \frac{\sum_{i=1}^{B} I\left(|t_i| \geq |t_0|\right)}{B}$$

gdzie $I(\cdot)$ jest funkcją przyjmująca wartości: 1 gdy spełniony jest warunek w nawiasie oraz wartość 0 w przeciwnym wypadku.

	MATLAB	R
Testy resamplingowe:	`mafdr`	`perm`

Weryfikacja hipotez statystycznych: resamplingowe testy istotności



Weryfikacja hipotez statystycznych: resamplingowe testy istotności

$t_0 = 2.2$ ← Wartość statystyki dla prawdziwego przypisania próbek

$t_1 = 0.4$ ← Wartości statystyki dla losowych przypisań próbek

$t_i = \{ 0.4 \}$

Weryfikacja hipotez statystycznych: resamplingowe testy istotności

$t_0 = 2.2$ ← Wartość statystyki dla prawdziwego przypisania próbek

$t_2 = 1.1$ ← Wartości statystyki dla losowych przypisań próbek

$t_i = \{ 0.4, 1.1 \}$

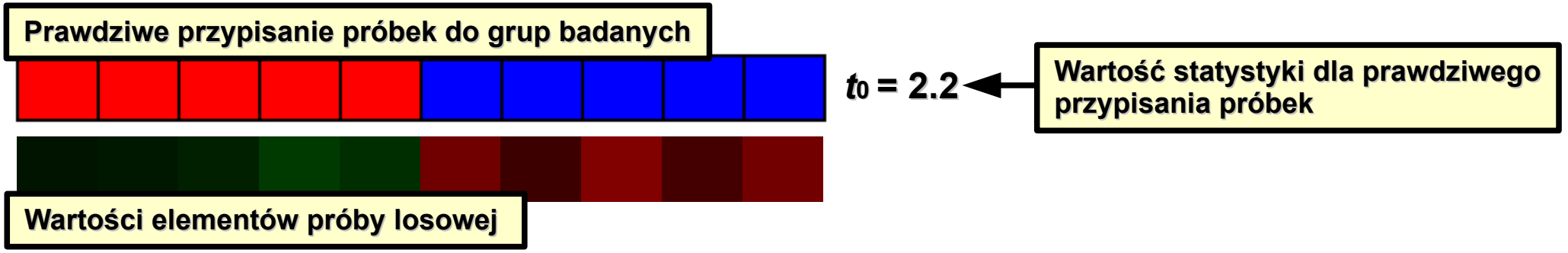
Weryfikacja hipotez statystycznych: resamplingowe testy istotności

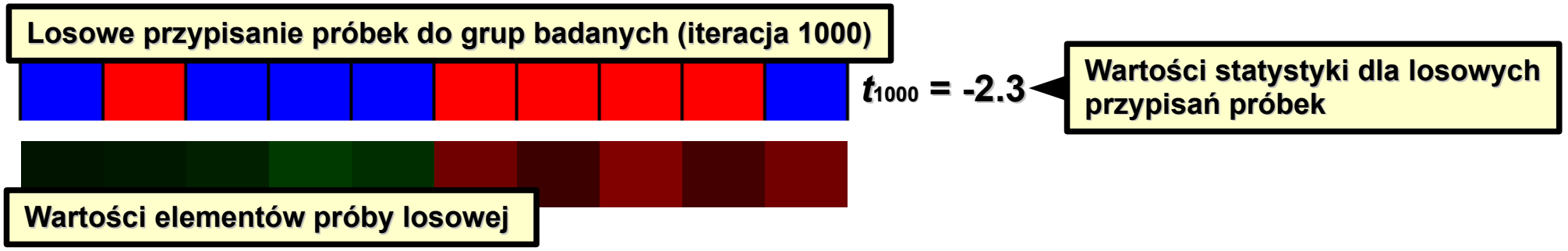
$t_0 = 2.2$ ← Wartość statystyki dla prawdziwego przypisania próbek

$t_3 = 2.4$ ← Wartości statystyki dla losowych przypisań próbek

$t_i = \{ 0.4, 1.1, 2.4\}$

Weryfikacja hipotez statystycznych: resamplingowe testy istotności

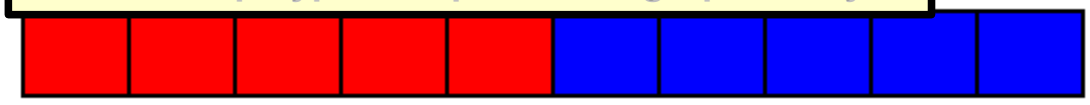




t_i = {-2.8, -2.6, -2.3, -2.0, -1.9, ..., 1.6, 1.8, 2.1, 2.2, 3.8}

Weryfikacja hipotez statystycznych: resamplingowe testy istotności

Prawdziwe przypisanie próbek do grup badanych



$t_0 = 2.2$

Wartość statystyki dla prawdziwego przypisania próbek

Wartości elementów próby losowej

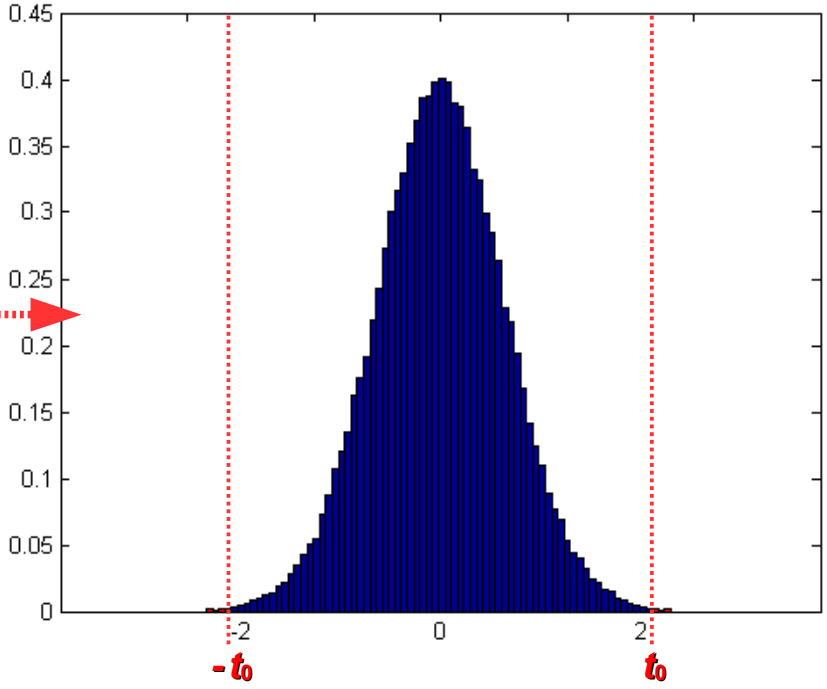
Losowe przypisanie próbek do grup badanych (iteracja 1000)

$t_{1000} = -2.3$

Wartości statystyki dla losowych przypisań próbek

Wartości elementów próby losowej

$t_i = \{-2.8, -2.6, -2.3, -2.0, -1.9, ..., 1.6, 1.8, 2.1, 2.2, 3.8\}$



Weryfikacja hipotez statystycznych: resamplingowe testy istotności

$t_0 = 2.2$

Wartość statystyki dla prawdziwego przypisania próbek

$t_{1000} = -2.3$

Wartości statystyki dla losowych przypisań próbek

$t_i = \{-2.8, -2.6, -2.3, -2.0, -1.9, ..., 1.6, 1.8, 2.1, 2.2, 3.8\}$

$$p = \frac{\sum_{i=1}^{B} I\left(|t_i| \geq |t_0|\right)}{B} = \frac{5}{1000} = 0.005$$

Liczba przypadków, dla których wartości bezwzględne t_i są nie mniejsze od wartości bezwzględnej t_0 obliczonej dla prawidłowego przypisania próbek do grup

Liczba iteracji

p-wartość

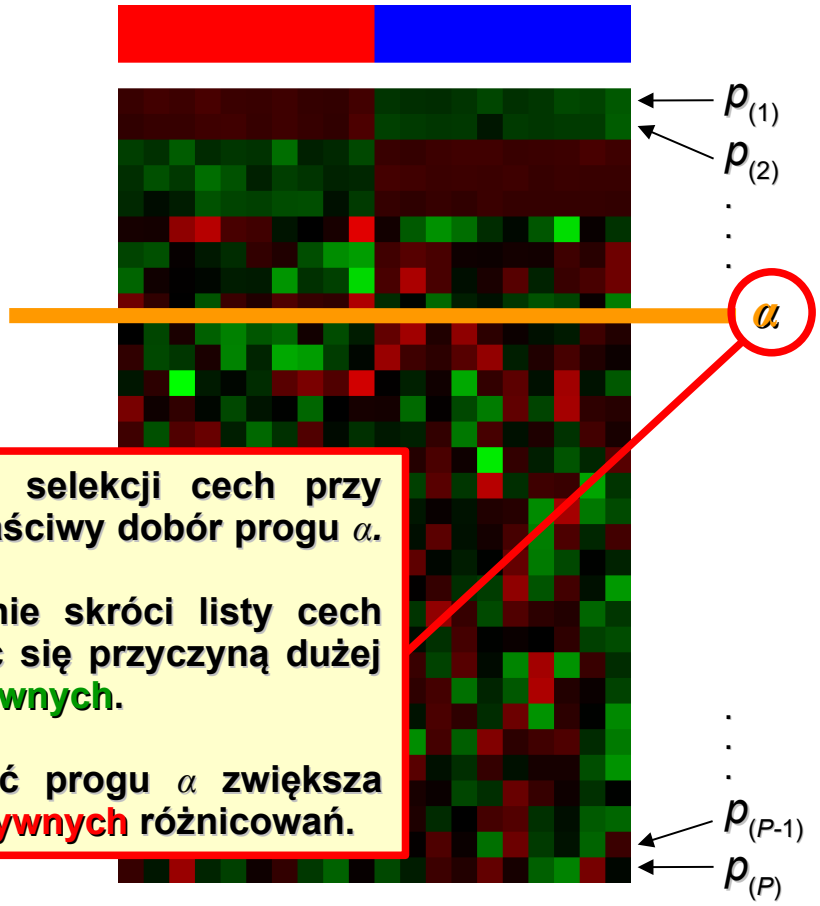
Uczenie maszynowe w bioinformatyce

Wykład 9: selekcja cech przy użyciu testów istotności (problem wielokrotnego testowania)

Problem jednoczesnego testowania wielu hipotez

Niezależnie od rodzaju stosowanego testu badającego równość wartości średnich w grupach, procedura selekcji cech różnicujących wygląda tak samo:

- **dla każdej z cech wykonuje się test istotności i zapamiętuje uzyskaną *p*-wartość;**
- **cechy sortowane są zgodnie z niemalejącymi *p*-wartości: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(P)}$;**
- **pozostawia się te cechy, których *p*-wartości nie są większe od pewnego progu α.**



Głównym problemem podczas selekcji cech przy użyciu testów istotności jest właściwy dobór progu α.

Zbyt mała wartość α nadmiernie skróci listy cech uznanych za różnicujące, stając się przyczyną dużej liczby wyników fałszywie negatywnych.

Z drugiej strony, duża wartość progu α zwiększa ryzyko wykrycia fałszywie pozytywnych różnicowań.

Problem jednoczesnego testowania wielu hipotez

Testowaniu poddawany jest zbiór hipotez $H_0{}^i$, gdzie $i = 1, ..., P$.

W dalszej części stosowane będą oznaczenia zgromadzone w poniższej tabeli:

Liczba wyników fałszywie pozytywnych (cech niesłusznie uznanych za różnicujące)

	Liczba nieodrzuconych H_0	Liczba odrzuconych H_0	Suma
Liczba prawdziwych H_0	U	V	P_0
Liczba fałszywych H_0	T	S	$P - P_0$
Suma	$P - R$	R	P

Liczba wyników fałszywie negatywnych (cech niesłusznie uznanych za nieróżnicujące)

Problem jednoczesnego testowania wielu hipotez

Gdy wykonujemy test dla pojedynczej cechy akceptowalne prawdopodobieństwo popełnienia błędu pierwszego rodzaju (odrzucenia hipotezy H_0, gdy w rzeczywistości jest ona prawdziwa) określane jest przez wybrany z góry poziomu istotności α. Typowo stosuje się wartości α równe 0.05, 0.01 lub 0.001.

Należy jednak pamiętać, że w ramach procedury selekcji cech różnicujących dla danych z technik wielkoskalowych (np. z mikromacierzy DNA lub spektrometrii mas) wykonywanych jednocześnie P testów (po jednym dla każdej cechy).

Jeżeli dla każdego z testów ustalony zostanie poziom istotności równy α, wówczas – przy założeniu niezależności testów – prawdopodobieństwo popełnienia co najmniej jednego błędu typu I dla całego zbioru P jednocześnie testowanych hipotez wynosi:

$$\Pr(V \geq 1) = 1 - (1 - \alpha)^P$$

Przykładowo:

$$\alpha = 0.01,\ P = 10 \quad \rightarrow \quad \Pr(V \geq 1) = 1 - (1 - 0.01)^{10} \quad \approx 0.0956$$

$$\alpha = 0.01,\ P = 100 \quad \rightarrow \quad \Pr(V \geq 1) = 1 - (1 - 0.01)^{100} \quad \approx 0.6340$$

$$\alpha = 0.01,\ P = 1000 \quad \rightarrow \quad \Pr(V \geq 1) = 1 - (1 - 0.01)^{1000} \quad \approx 1.0000$$

Problem jednoczesnego testowania wielu hipotez

Zaprezentowana na poprzednim slajdzie zależność (zwana równaniem Šidáka) uświadamia konieczność rozróżnienia prawdopodobieństwa α popełnienia błędu typu I w pojedynczym teście od prawdopodobieństwa A („duże” α) popełnienia błędu typu I w całej rodzinie testów.

W przypadku dużej liczby jednocześnie badanych cech stosowanie typowych wartości progów istotności dla poszczególnych testów nie jest wystarczające do utrzymania na akceptowalnym poziomie liczby fałszywie pozytywnych wyników w rodzinie testów (czyli tutaj: liczby cech fałszywie uznanych za różnicujące).

Prowadzi to do konieczności stosowania technik korekcji pod kątem jednoczesnego testowania wielu hipotez. Korekcja ta realizowana jest przez kontrolę wybranej miary częstości występowania wyników fałszywie pozytywnych i może być zrealizowana na dwa równoważne sposoby:

- **określenie skorygowanych (obniżonych) poziomów istotności dla kolejnych testów;**
- **wyznaczenie skorygowanych (podwyższonych) *p*-wartości poszczególnych testów.**

Problem jednoczesnego testowania wielu hipotez: FWER

Klasycznym podejściem do problemu jednoczesnego testowania wielu hipotez jest takie skonstruowanie procedury testowej, aby kontroli na zadanym poziomie poddawana była miara FWER (*Family-Wise Error Rate*), czyli prawdopodobieństwo odrzucenia w całym zbiorze P testów co najmniej jednej prawdziwej hipotezy zerowej H_0 (popełnienia co najmniej jednego błędu typu I):

$$FWER = \Pr(V \geq 1) = 1 - \Pr(V = 0) = 1 - (1 - \alpha)^P$$

Przez procedurę kontrolującą FWER na ustalonym poziomie A rozumie się procedurę zapewniającą spełnienie warunku:

$$FWER \leqslant A$$

Warunek ten jest równoważny zapewnieniu, że prawdopodobieństwo, że w rodzinie testów żadna hipoteza zerowa nie zostanie fałszywie odrzucona (tzn. żadna cecha, która w rzeczywistości nie różnicuje grup nie zostanie uznana za różnicującą) będzie nie mniejsze od $1 - A$.

Problem jednoczesnego testowania wielu hipotez: FWER

Najbardziej oczywista procedura kontroli FWER rodziny P testów na poziomie A wynika bezpośrednio z równania Šidáka. Polega ona na odrzuceniu wszystkich hipotez zerowych, którym odpowiadają *p*-wartości nie większe od progu:

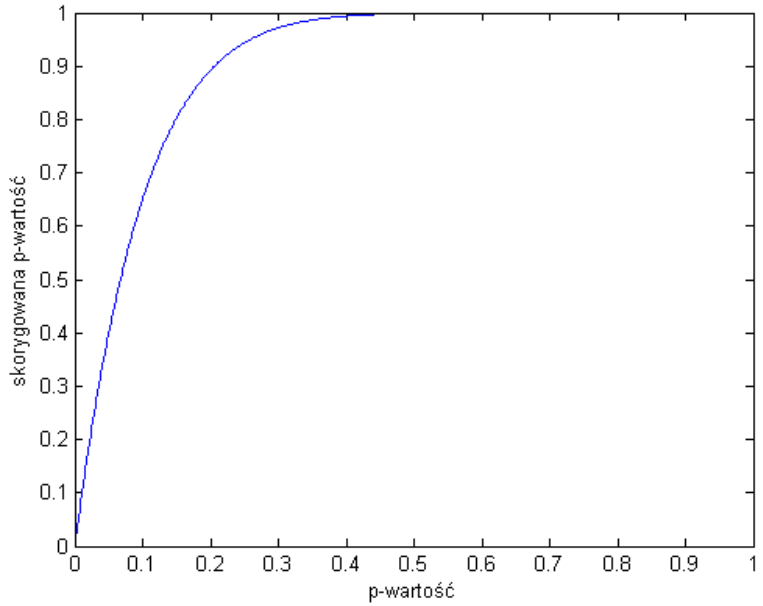
$$\alpha_{Šidák} = 1 - (1 - A)^{1/P}$$

Przykładowo, chcąc kontrolować FWER rodziny 10 testów na poziomie A=0.05 należy ich *p*-wartości porównać z progiem $\alpha_{Šidàk}$=0.0051.

Równoważnym podejściem jest wyznaczenie skorygowanych *p*-wartości jako:

$$p_{Šidák} = 1 - (1 - p)^{P}$$

i odrzucenie hipotez o skorygowanych *p*-wartościach nie większych od progu α=A.



	MATLAB	R
FWER:	-	`p.adjust`

Problem jednoczesnego testowania wielu hipotez: FWER

Liniowym przybliżeniem procedury Šidáka jest procedura Bonferroniego, w której odrzucane są wszystkie hipotezy o *p*-wartościach nie większych od progu:

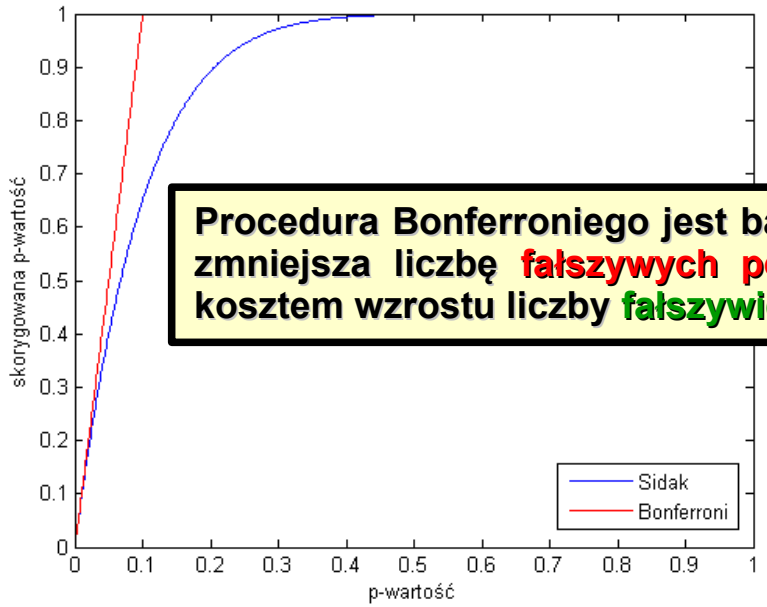
$$\alpha_{Bonferroni} = A/P$$

Przykładowo, chcąc kontrolować FWER rodziny 10 testów na poziomie $A=0.05$ należy ich *p*-wartości porównać z progiem $\alpha_{Bonferroni}=0.0050$.

Podobnie jak poprzednio możliwe jest wyznaczenie skorygowanych *p*-wartości jako:

$$p_{Bonferroni} = \min(p \cdot P, 1)$$

i odrzucenie hipotez o skorygowanych *p*-wartościach nie większych od progu $\alpha=A$.



Procedura Bonferroniego jest bardziej zachowawcza: zmniejsza liczbę fałszywych pozytywnych wyników kosztem wzrostu liczby fałszywie negatywnych.

	MATLAB	R
FWER:	-	`p.adjust`

Problem jednoczesnego testowania wielu hipotez: FDR

Alternatywnym podejściem do problemu wielokrotnego testowania w jest kontrola miary FDR (*False Discovery Rate*), czyli wartości oczekiwanej frakcji fałszywie odrzuconych hipotez zerowych w zbiorze wszystkich odrzuconych:

$$FDR = E(Q)$$

$$Q = \begin{cases} V / R, & dla\ R > 0 \\ 0, & dla\ R = 0 \end{cases}$$

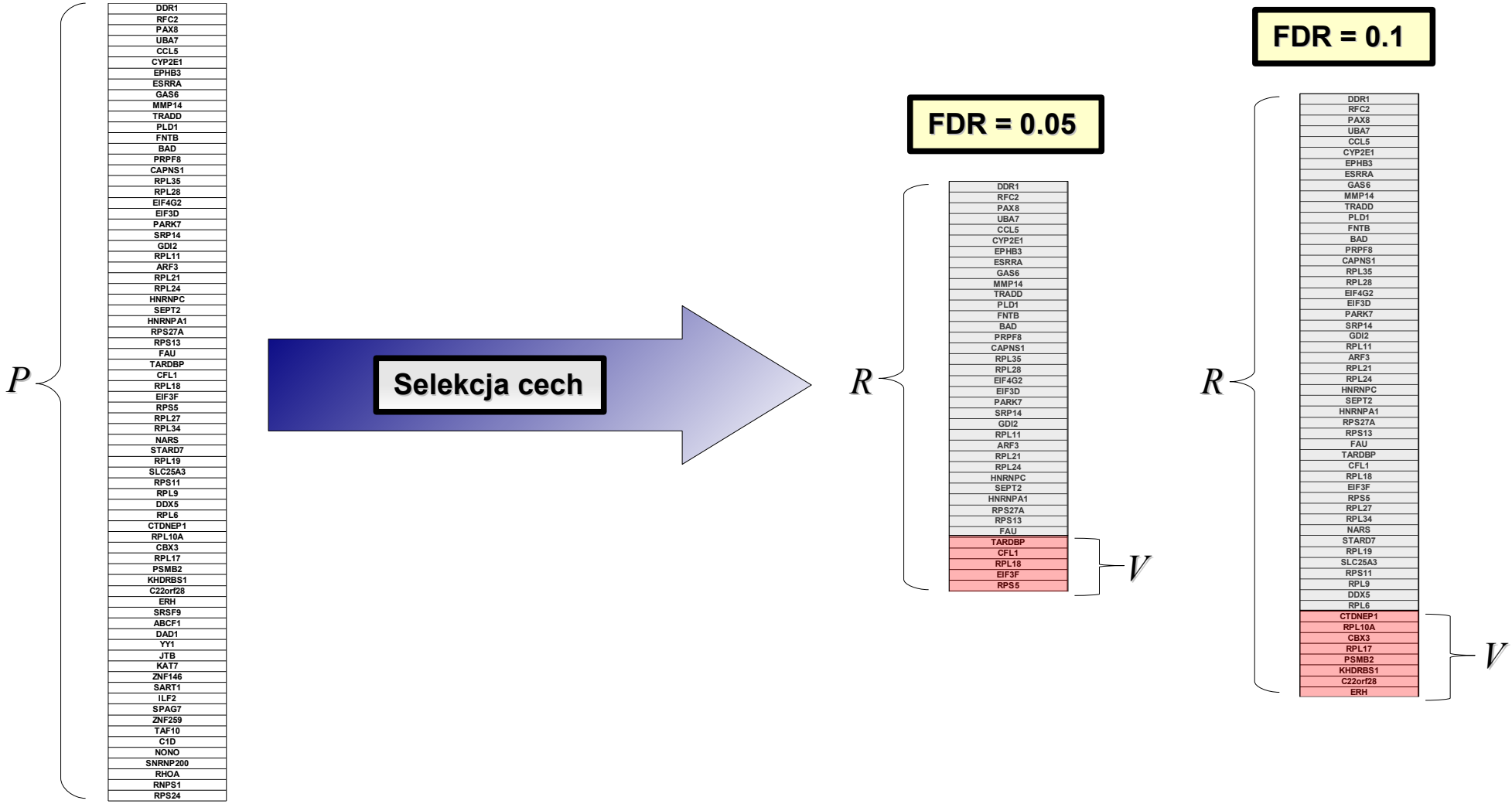
Jeżeli wszystkie hipotezy zerowe są prawdziwe (tzn. w zbiorze danych nie ma żadnych cech różnicujących), to FDR jest równy FWER. Jednak w przypadku, gdy nie wszystkie hipotezy są prawdziwe zawsze zachodzi zależność:

$$FDR \leq FWER$$

W kontekście danych z mikromacierzy główną zaletą procedur kontroli FDR jest mniejszy spadek mocy (rozumianej jako zdolność do wykrywania fałszywych hipotez zerowych) wraz ze wzrostem liczby testów.

Problem jednoczesnego testowania wielu hipotez: FDR

Istotną zaletą FDR jest również intuicyjna interpretacja tej miary: kontrola FDR na poziomie 0.05 oznacza, że w zbiorze cech, które uznamy za różnicujące możemy oczekiwać występowania 5% fałszywie pozytywnych wyników. W praktycznych zastosowaniach jest to zwykle bardziej użyteczna informacja niż ta dostępna przy kontroli FWER, gdyż daje badającemu możliwość ustalenia kompromisu pomiędzy długością list cech różnicujących a liczba fałszywie pozytywnych wyników.



Problem jednoczesnego testowania wielu hipotez: FDR

Historycznie pierwszym i jednocześnie najczęściej stosowanym sposobem kontroli wartości FDR dla P testów jest procedura Benjaminiego-Hochberga (B-H). Jest to sekwencyjna procedura, w której każda z *p*-wartości jest porównywana jest z inną wartością progową. O wartości progu (czyli o stopniu korekcji) dla danej *p*-wartości decyduje jej pozycja w posortowanym ciągu wszystkich *p*-wartości rodziny testów.

Przebieg procedury B-H:

- ***p*-wartości wszystkich testów sortowane są w kolejności niemalejącej:**

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(P)}$$

- **jeżeli dla hipotezy o największej *p*-wartości zachodzi $p_{(P)} \leq A$, to nie przyjmujemy żadnej hipotezy (uznajemy wszystkie cechy za różnicujące), w przeciwnym razie przyjmujemy wszystkie hipotezy, których *p*-wartości są większe od progów:**

$$\alpha_{B\text{-}H\,(i)} = A \cdot i / P$$

Najmniejsza *p*-wartość zostanie porównana z progiem A/P (czyli takim samym jak przy kontroli FWER procedurą Bonferroniego), dla kolejnych *p*-wartości progi będą jednak coraz wyższe.

	MATLAB	R
FDR:	`mafdr`	`p.adjust`

Problem jednoczesnego testowania wielu hipotez

W przypadku danych z mikromacierzy najczęściej stosowana jest kontrola FDR, jako że stanowi ona kompromis pomiędzy liczbą wyników fałszywie negatywnych (decydującą o wielkości zbioru genów uznanych za różnicujące) a liczbą wyników fałszywie pozytywnych (decydującą o poziomie błędu w tym zbiorze).