Wstęp do uczenia maszynowego

P-wartości, testowanie wielu hipotez statystycznych

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

4 marca 2024

UNIWERSYTET
WARSZAWSKI

Przykład: testowanie magicznych zdolności

- Gracz twierdzi, że potrafi bez patrzenia dobrze wytypować kolor wylosowanej karty.
- W talii mamy 52 karty, każda ma jeden z czterech kolorów (trefl, karo, kier i pik).
- H_0: ten gracz tylko blefuje
- H_1: ten gracz posiada magiczne zdolności (hipoteza złożona)
- Prosimy gracza aby odgadł bez patrzenia 20 kart wylosowanych ze zwracaniem.
- $T-$ liczba prawidłowo wytypowanych kart z 20
- Jeśli H_0 prawdziwa,
  - $T \sim Binom(20, 1/4)$
  - Duże wartości T nieprawdopodobne
  - Przy ustalonym poziomem istotności α, region krytyczny: $\{t_0, \ldots, 20\}$, gdzie $P(T \geq t_0 | H_0) = \alpha$

Przykład: testowanie magicznych zdolności

- Gracz zgadł poprawnie kolor 9 z 20 wylosowań.
- Dla $\alpha = 0.05$, 9 wpada do regionu krytycznego, hipoteza blefowania byłaby odrzucona - uznalibyśmy, że gracz ma magiczne zdolności
- Dla $\alpha = 0.01$, 9 już nie wpada do rejonu krytycznego, hipoteza blefowania nie byłaby odrzucona - i nie uznalibyśmy, że gracz ma magiczne zdolności

Jak dobrać α?

- Wybór α ma być dokonany z góry, przed testowaniem hipotez, i jest tak naprawdę dowolny.
- Zwyczajowo, wybiera się $\alpha = 0.05$ bądź $\alpha = 0.01$
- Mimo tego, że często w literaturze np. medycznej, wartości te traktowane są jako specjalnie „ważne" poziomy istotności, nie powinniśmy się do nich przesadnie przywiązywać
- Pamiętamy, że odrzucenie H_0 na poziomie α oznacza dokładnie tyle, że prawdopodobieństwo błędu I rodzaju wynosi właśnie α

p-wartość

Aby np. móc porównywać różne próby losowe, które testujemy, czasem wygodniej posługiwać się *p*-**wartościami**. Intuicyjnie, chcielibyśmy tak zdefiniować p-wartość, aby móc oszacować wsparcie danych przeciwko hipotezie H_0. **Im wartość p mniejsza, tym wsparcie silniejsze.**

p-wartość

Najniższy wartość α, dla której dla danej próby hipoteza zerowa zostałaby odrzucona

Dla przykładu z poprzedniego slajdu:

- Gracz zgadł poprawnie kolor 9 z 20 wylosowań.
  p-wartość$=P(T \geq 9|H_0) = 0.041$
- Gracz zgadł poprawnie kolor 10 z 20 wylosowań.
  p-wartość$=P(T \geq 10|H_0) = 0.014$, itd.

Sir Roland Fisher

- brytyjski genetyk, statystyk (1890 - 1962)
- prof. London School of Economics i Uniwersytetu w Cambridge
- autor m.in.
  - podstaw teorii weryfikacji hipotez
  - metody największej wiarogodności
  - analizy wariancji (ANOVA)
  - liniowej analizy dyskryminacyjnej
- Studia (w latach 1911-1941, a więc przed kluczowymi odkryciami w dziedzinie genetyki) nad genetyką populacyjną ukształtowały jego poglądy na tematy rasowe, które - z dzisiejszego punktu widzenia - okazały się zupełnie błędne.

Testowanie hipotez w praktyce

Klasycznie:

- Sprawdzenie, czy statystyka w obszarze krytycznym
- Obszar krytyczny: z tablic rozkładów, dla zadanej hipotezy i poziomu istotności

Teraz (np dzięki obliczeniom w R):

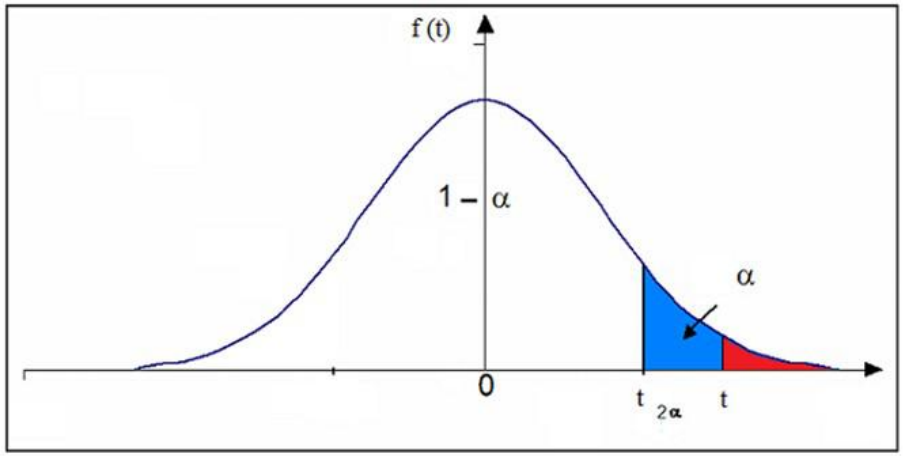
- wartość dystrybuanty dla zadanej wartości statystyki - pozwala wyliczyć p-wartości
- jeśli konieczna jest decyzja, czy odrzucamy H_0, porównujemy otrzymaną p-wartość z przyjętym poziomem istotności.
- Intuicyjnie: p-wartość to prawdopodobieństwo (przy H_0 spełnionej) otrzymania wartości bardziej odległej od oczekiwanej niż wartość zaobserwowana.

p-wartość: przykłady

Wróćmy do testu istotności dla wartości średniej A, rozkład normalny, σ^2 znana.

Przykład 1

- $H_0 : \mu \leq \mu_0$
- $H_1 : \mu > \mu_0$
- Niech wartość statystyki to t_{obl}
- p-wartość $p = P(T \geq t_{obl} \mid H_0)$
- Gdy $p < \alpha$ odrzucamy H_0



p-wartość: przykłady

Przykład 2

- $H_0 : \mu \geq \mu_0$
- $H_1 : \mu < \mu_0$
- p-wartość $p = P(T \leq t_{obl} \mid H_0)$

Przykład 3

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$
- p-wartość $p = P(|T| \geq |t_{obl}| \mid H_0)$

Mała wartość p ($p < \alpha$): odrzucamy H_0

Ostrożność przy wnioskowaniu z odrzucenia hipotez

- Nieodrzucenie H_0 nie dowodzi, że jest ona prawdziwa
- p-wartość to prawdopodobieństwo odrzucenia H_0 podczas gdy jest ona prawdziwa - zawsze jest ono większe od 0
- przy testowaniu należy zwracać uwagę na:
  - właściwe określenie statystyki testowej i jej rozkładu,
  - spełnienie założeń testu,
  - ustalenie właściwych obszarów krytycznych
  - korektę p-wartości otrzymanych w wielu testach

Winston Churchill (1874-1965):
"Wierzę tylko w te statystyki, które sam sfałszowałem"[1]

[1] Najprawdopodobniej jest to cytat fałszywie przypisywany Churchillowi - https://tinyurl.com/yc776jk3

Jak zachowują się p-wartości dla wielu testów

A Pewna liczba testów gdzie H_0 niespełniona - rozkład jednostajny wzbogacony w okolicach 0

B Przy spełnionej hipotezie zerowej w każdym teście - rozkład jednostajny

C Przy źle dobranym rozkładzie - wzbogacenia daleko od 0

D Przy zastosowaniu rozkładu ciągłego do danych dyskretnych

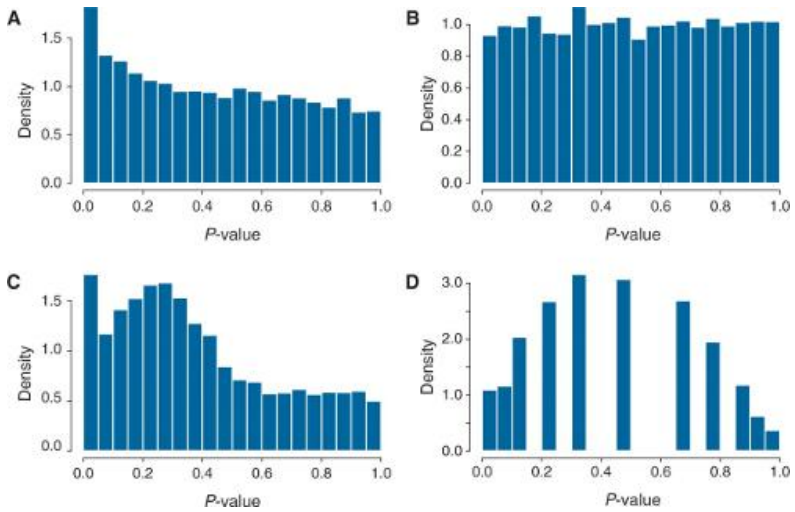
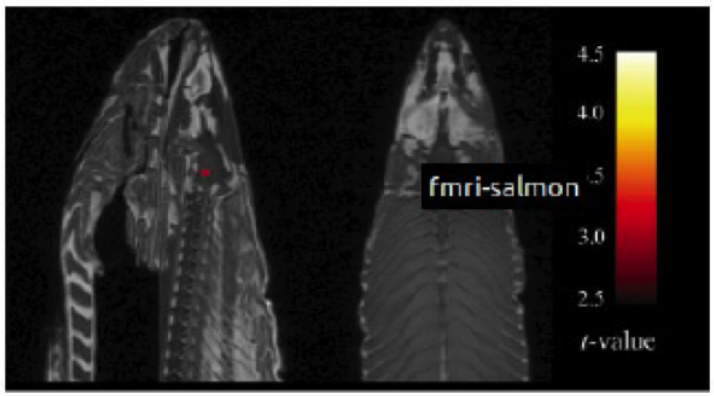


Image source: https://www.researchgate.net/publication/275141231

O czym myśli martwy łosoś?

SCANNING DEAD SALMON IN FMRI MACHINE HIGHLIGHTS RISK OF RED HERRINGS



NEUROSCIENTIST CRAIG BENNETT purchased a whole Atlantic salmon, took it to a lab at Dartmouth, and put it into an fMRI machine used to study the brain. The beautiful fish was to be the lab's test object as they worked out some new methods.

So, as the fish sat in the scanner, they showed it "a series of photographs depicting human individuals in social situations." To maintain the rigor of the protocol (and perhaps because it was hilarious), the salmon, just like a human test subject, "was asked to determine what emotion the individual in the photo must have been experiencing."

http://www.wired.com/2009/09/fmrisalmon/

O czym myśli martwy łosoś?

Journal of Serendipitous and Unexpected Results

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1]*, Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1]Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3]Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of at least one false positive is almost certain. Proper correction for multiple comparisons should be completed during the analysis of these datasets, but is often ignored by investigators. To highlight the danger of this practice we completed an fMRI scanning session with a post-mortem Atlantic Salmon as the subject. The salmon was shown the same social perspective-taking task that was later administered to a group of human subjects. Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon's brain cavity and spinal column. Statistics controlling for the family-

With 130,000 voxels in a single functional neuroimaging volume it is now common practice to do tens of thousands of tests per contrast over multiple contrasts. While this extreme dimensionality offers dramatic new opportunities in terms of analysis it also comes with dramatic new opportunities for false positives in the results. As a result the nagging issue of multiple comparisons has been thrust to the forefront of discussion in a diverse array of scientific fields, including cognitive neuroscience. More and more researchers have realized that correcting for chance discoveries is a necessary part of imaging analysis. This is a positive trend, but it over-

O czym myśli martwy łosoś?

"In fMRI, you have 160,000 darts, and so just by random chance, by the noise that's inherent in the fMRI data, you're going to have some of those darts hit a bull's-eye by accident"

Ostrożnie z interpretacją p-wartości

p-wartości

- zależą od dwóch rzeczy
  1. wielkości statystyki
  2. wielkości próby
- może być duża nawet gdy hipoteza alternatywna prawdziwa, bo próba jest za mała
- może być dowolnie mała gdy próba $\rightarrow \infty$

Dlatego oprócz $p-$wartości, warto też sprawdzic wielkość efektu (ang. effect size)

Przykład. Wielkość efektu dla dwóch sparowanych prób

Efekt:

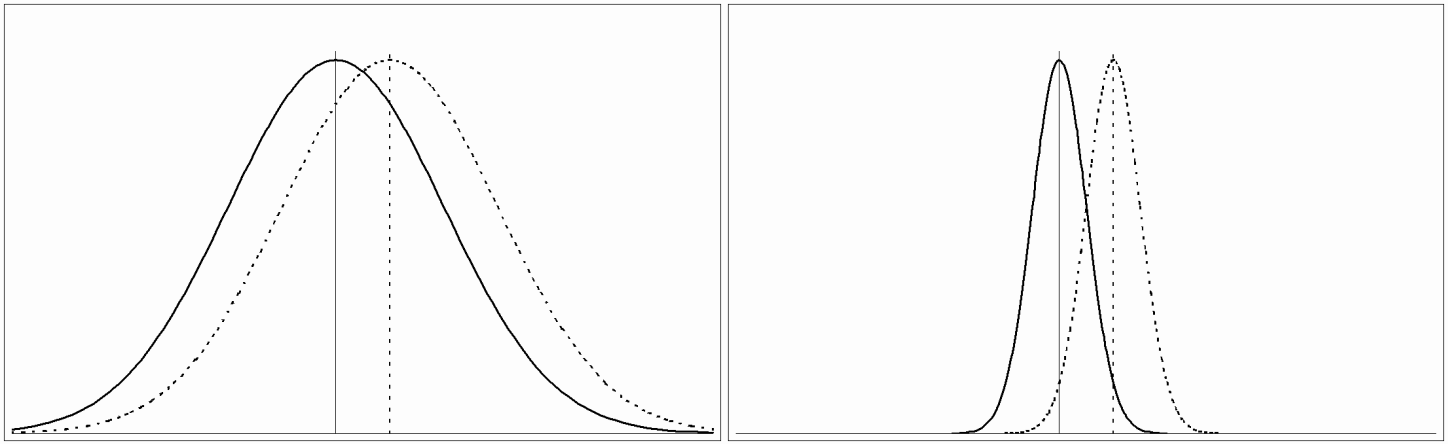
- różnica średnich z dwóch (sparowanych) prób
- tzw fold change dla ekspresji genów

Problemy z oceną wielkości efektu

- czy dwukrotny wzrost ekpresji genów to dużo, czy mało?
- a półtorakrotny?
- a 1.1-krotny?

Ocena zależy od wariancji porównywanych rozkładów.

Miara wielkości efektu d Cohena



$$d = \frac{\mu_{\text{group1}} - \mu_{\text{group2}}}{\hat{S}_{\text{pooled}}}$$

gdzie μ to średnia z próby, a $\hat{S}_{\text{pooled}}$ to odchylenie standardowe z połączonych prób,

$$\hat{S}_{\text{pooled}} = \sqrt{\frac{\hat{S}^2_{\text{group1}} + \hat{S}^2_{\text{group2}}}{2}}$$

Miara efektu Cohena dla prawego obrazka jest większa.

O czym myśli martwy łosoś?

"By complete, random chance, we found some voxels that were significant that just happened to be in the fish's brain," Bennett said. "And if I were a ridiculous researcher, I'd say, 'A dead salmon perceiving humans can tell their emotional state.'"

Testowanie wielu hipotez a błędy I rodzaju

Przykład

- Mamy 10000 genów na mikromacierzy
- Załóżmy, że żaden z nich tak naprawdę nie jest różnicowo eksprymowany
- Testujemy dla każdego z nich, czy jest różnicowo eksprymowany (robimy 10000 testów)
- Przyjmimy $\alpha = 0.01$
- Oczekiwać można $10000 \times 0.01 = 100$ genów z p-wartością < 0.01.

Kalkulacje prawdopodobieństw błędu I rodzaju

Z_n - statystyka z próby n elementowej, W- obszar krytyczny, $W' = \mathbb{R} \setminus W$- obszar przyjęć.

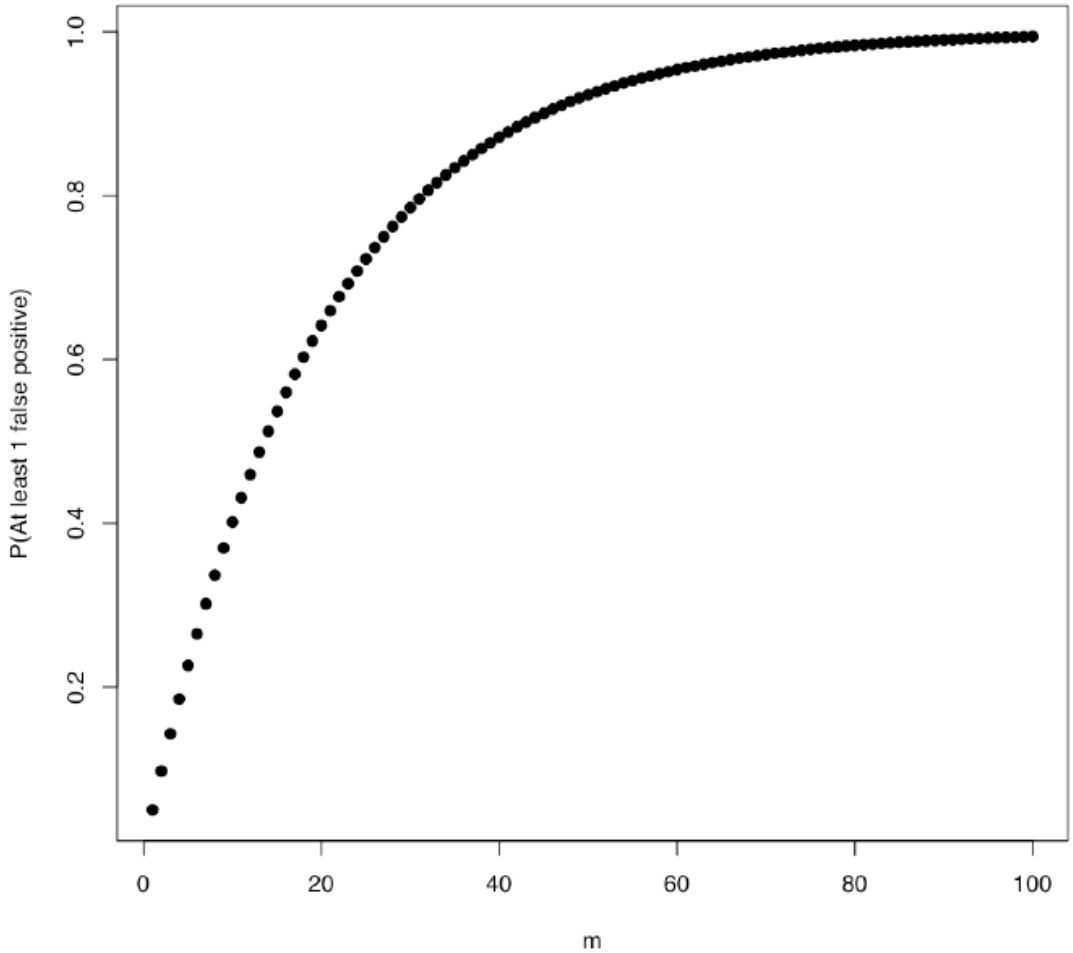
Przy jednym teście:

- Błąd I rodzaju (fałszywy pozytyw, ang. *false positive, FP*):
  - odrzucimy H_0, gdy w istocie jest ona prawdziwa.
  - z prawdopodobieństwem $\alpha(W) = P(Z_n \in W | H_0)$.
- Prawdopodobieństwo niepopełnienia błędu I rodzaju $P(Z_n \in W' | H_0) = 1 - \alpha$

Przy m testach:

- Prawdopodobieństwo niepopełnienia błędu I rodzaju w żadnym z testów $(1 - \alpha)^m$
- **Prawdopodobieństwo popełnienia błędu I rodzaju w conajmniej jednym z testów** $1 - (1 - \alpha)^m$

Prawdopodobienstwo popełnienia conajmniej jednego błędu I rodzaju w funkcji liczby testów



Oznaczenia

- Testujemy m hipotez $H_1, \ldots, H_m$
- m_0: liczba prawdziwych H_0
- R: liczba hipotez zerowych odrzuconych
- FP: liczba błędów I rodzaju (*false positives*)

	H_0 prawdziwa	H_1 prawdziwa	Łącznie
H_0 nie odrzucona	*TN*	*FN*	$m - R$
H_0 odrzucona	*FP*	*TP*	R
	m_0	$m-m_0$	m

Korekta p-wartości przy testowaniu wielu hipotez

To tak naprawdę kontrolowanie błędów I rodzaju przy testowaniu wielu hipotez:

- **Per comparison error rate (PCER)** Wartość oczekiwana liczby błędów I rodzaju na liczbę testowanych hipotez

$$PCER = \frac{\mathbb{E}[FP]}{m}$$

- **Per-family error rate (PFER):** Wartość oczekiwana liczby błędów I rodzaju

$$PFER = \mathbb{E}[FP]$$

- **Family-wise error rate (FWER):** Prawdopodobieństwo conajmniej jednego błędu I rodzaju

$$FWER = P(FP \geq 1)$$

Procedury kontrolowania błędów I rodzaju przy wielokrotnym testowaniu hipotez

Podział procedur

- **Jednokrokowe (single step):** Każda p-wartość jest dopasowywana tak samo
- **Wielokrokowe (sequential):** Różne dopasowanie dla każdej p-wartości, bierze pod uwagę rozkład p-wartości

Procedury kontrolowania FWER: Korekta Bonferroniego

Korekta Bonferroniego

- Jednokrokowa procedura testowania z poziomem istotności $\frac{\alpha}{m}$, kontroluje $FWER = P(FP \geq 1)$ na poziomie α.
- Najprostsza metoda kontroli FWER
- p-wartości po korekcie:

$$\tilde{p}_i = \min[mp_i, 1]$$

$$FWER = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha.$$

(Nierówność Boole'a)

Przykład: wiele testów ekpresji różnicowej genów

- Chcemy mieć FWER 0.05
- Wykonujemy 10000 testów
- Potrzebujemy p-wartości rzędu $0.05/10000 = 5x10^{-6}$ aby odrzucić hipotezę zerową

Krytyka procedury Bonferoniego kontrolowania FWER

- "Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" Perneger (1998)
- Bardzo konserwatywna, duże prawdopodobieństwo błędu II rodzaju

Procedury kontrolowania FWER: metoda Holma

Wielokrokowa korekta Holma

- Uporządkuj p-wartości rosnąco $p_1 \leq p_2 \leq p_3 \ldots \leq p_m$
- Aby uzyskać kontrolę FWER na poziomie α, metoda Holma koryguje p-wartości następująco

$$\tilde{p}_i = \min[(m - i + 1)p_i, 1]$$

- Czyli nie mnożymy wszystkich p-wartości przez tę samą wartość.
- Na przykład, chcąc mieć FWER 0.05 i wykonując 10,000 testów, korygujemy

$$\tilde{p}_1 = 10000 \cdot p_1, \tilde{p}_2 = 9999 \cdot p_1, \ldots \tilde{p}_m = p_m$$

Krytyka kontrolowania FWER

- Rzadko kiedy boimy się błędów I rodzaju aż tak, że nie chcemy dopuścić do żadnego takiego błędu
- Często możemy zgodzić się, żeby wśród wszystkich z wielu odrzuconych hipotez zerowych znalazło się kilka fałszywych pozytywów.
- Wówczas lepiej kontrolować False discovery rate (FDR)

FDR a FPR

	H_0 prawdziwa	H_1 prawdziwa	Łącznie
H_0 nie odrzucona	*TN*	*FN*	***m - R***
H_0 odrzucona	*FP*	*TP*	***R***
	m_0	$m-m_0$	***m***

- **False discovery rate (FDR):**

$$FDR = \frac{FP}{R}.$$

- **False positive rate (FPR):**

$$FPR = \frac{FP}{m_0}$$

Procedury kontrolowania FDR: metoda Benjamini i Hochberga

Aby kontrolować FDR na poziomie δ

- Uporządkuj p-wartości rosnąco $p_1 \leq p_2 \leq p_3 \ldots \leq p_m$
- Znajdz test z najwyzszą rangą j, dla której zachodzi

$$p_j \leq \delta \frac{j}{m}$$

- Uznaj wszystkie testy o rangach $1, 2, \ldots, j$ za istotne (dla nich odrzucamy H_0)
  - Wówczas odrzucamy $\frac{j}{m}$-tą część hipotez zerowych.
  - δ-część z nich, która średnio oczekujemy, że jest fałszywa (z definicji p_j)
- Rownoważnie, zdefiniuj korektę

$$\tilde{p}_j = p_j \frac{m}{j}$$

Procedury kontrolowania FDR: metoda Benjamini i Hochberga

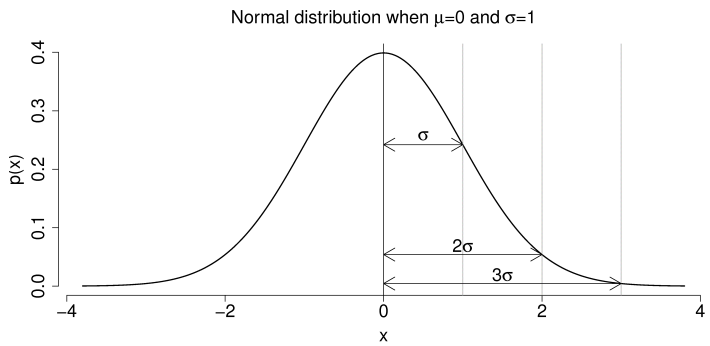
Przykład, korekta Benjamini i Hochberga z $\delta = 0.05$

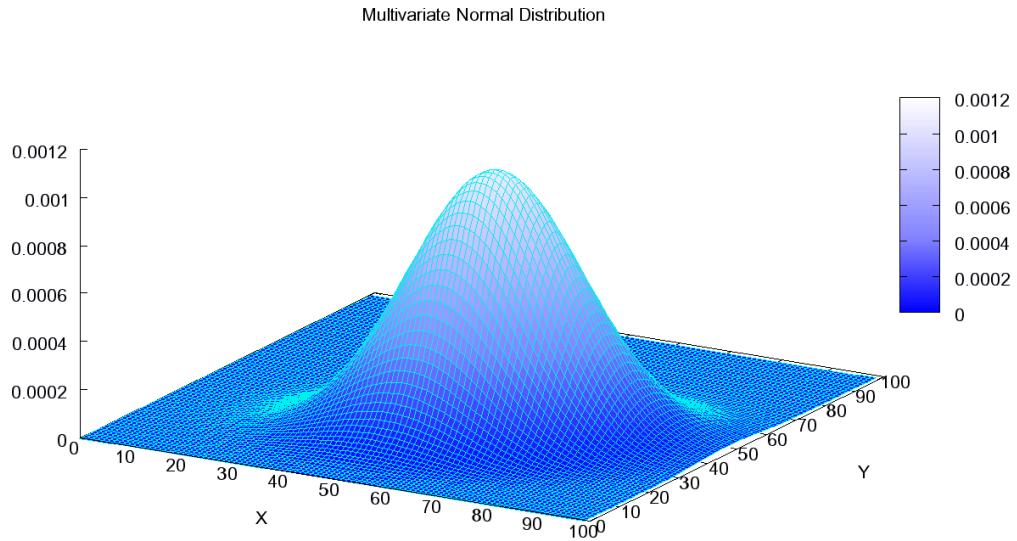
Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Referencje

- http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.pdf
- http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2900.html
- http://www.stat.berkeley.edu/~hhuang/STAT141/Lecture-FDR.pdf
- http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf

Rozkład dwóch zmiennych - słowo o zależnościach





- Zmienność rozkładu dwóch zmiennych: kowariancja

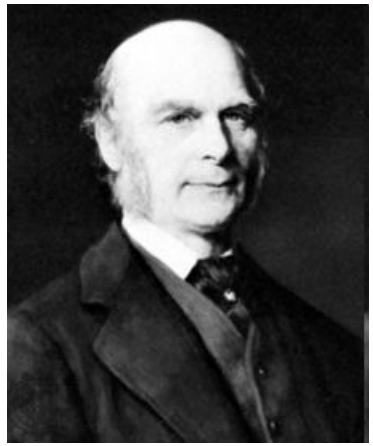
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]E[Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- Wzór podobny do wzoru na wariancję jednej zmiennej.
- Wartości kowariancji jest trudno zinterpretować.
- Wygodniejsza jest **korelacja**. A co to takiego?

Francis Galton

Koncept korelacji pochodzi od Sir Francisa Galtona, który również

- wprowadził pojęcie regresji
- jest ojcem dziedziny psychometryki
- spopularyzował użycie ankiet dla zbierania danych
- jest autorem frazy "nature versus nurture"
- Podobnie jak Fisher i wielu im współczesnych, był propagatorem "eugeniki" jako dyscypliny naukowej



Karl Pearson

Karl Pearson

- Wprowadził pojęcia
  - testu chi-kwadrat
  - odchylenia standardowego
  - współczynnika korelacji (sformalizował pojęcie korelacji Galtona).

Korelacja Pearsona

- Dla populacji

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]}{\sigma_X \sigma_Y}$$

- Dla próby

$$r = r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{nS_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

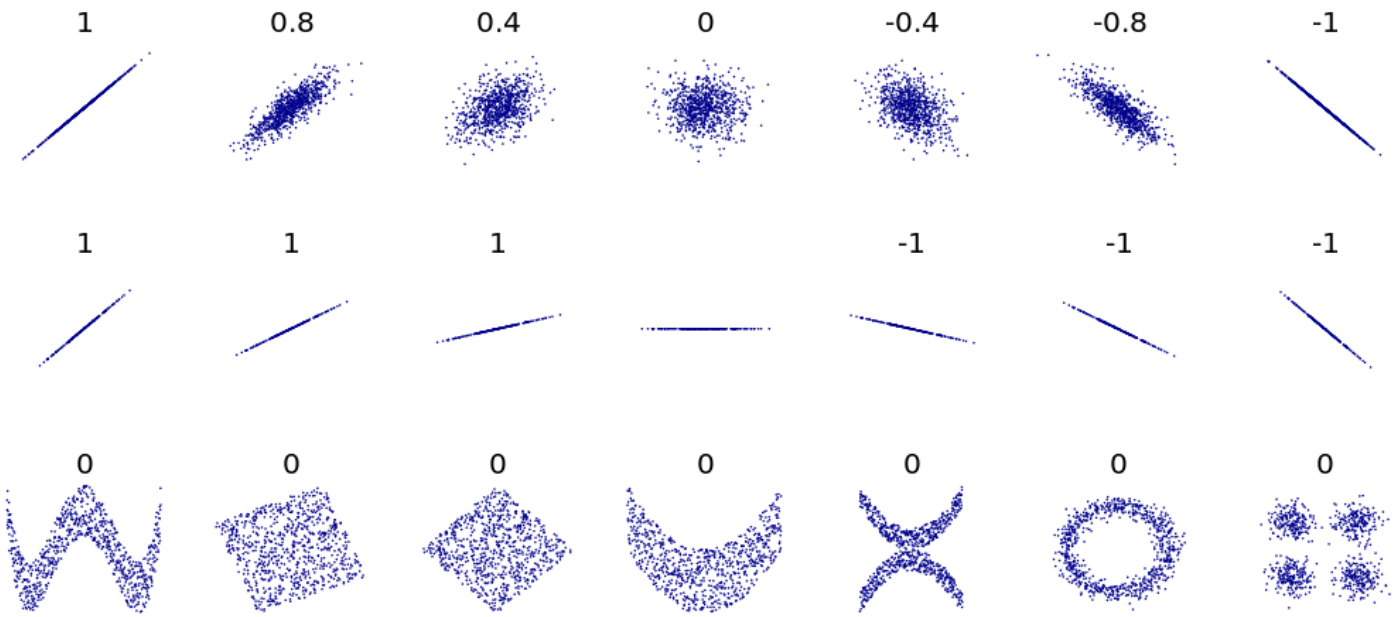
- Dla próby upraszczając

$$r = r_{XY} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\ \sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}.$$

Korelacja Pearsona-własności

- Symetryczność $\rho_{X,Y} = \rho_{Y,X}$.
- $\rho_{X,Y} = +1$ w przypadku idealnej zależności liniowej pomiędzy X i Y
- $\rho_{X,Y} = -1$ w przypadku idealnej odwrotnej zależności liniowej pomiędzy X i Y (antykorelacja)
- $\rho_{X,Y} \in [-1, 1]$ wskazuje na stopień zależności liniowej
- Gdy X i Y niezależne to $\mathbb{E}\,\rho_{X,Y} = 0$
- $\rho_{X,Y} = 0$ nie oznacza niezależności zmiennych, tylko brak zależności liniowej
- W naszym przykładzie, $r_{X,Y} = 0.125$, nieduże.

Ciekawe przykłady zależności w danych



Wartości: korelacja Pearsona.

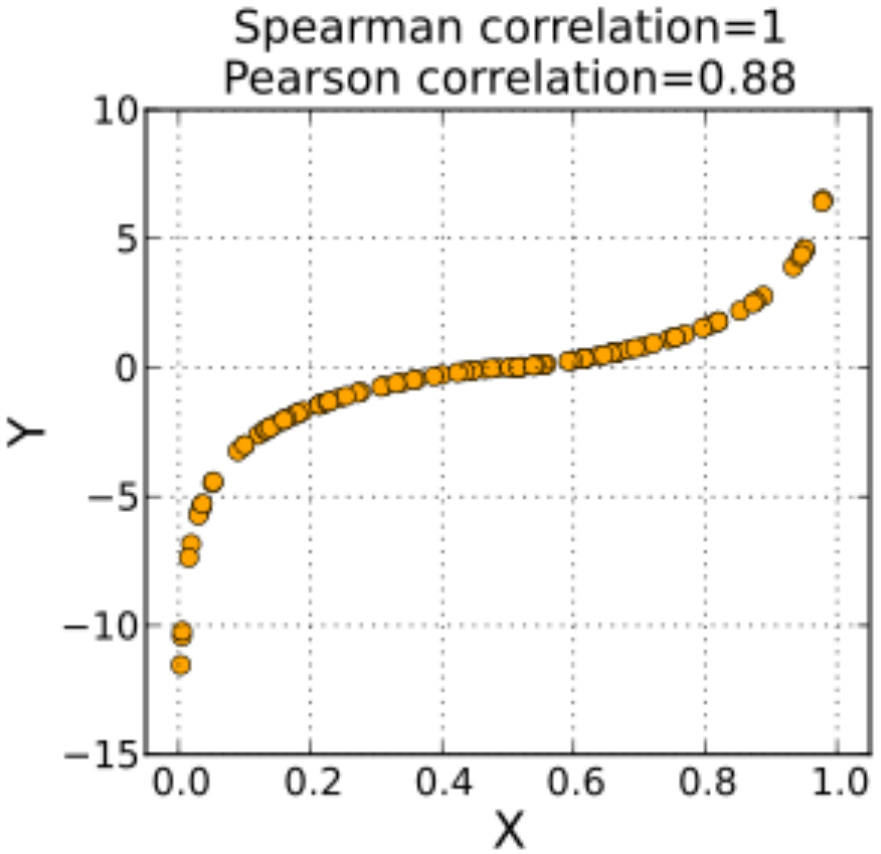
- $\rho_{X,Y}$ zależy od szumu w danych i kierunku zależności (1szy rząd)
- $\rho_{X,Y}$ nie zależy od nachylenia (2gi rząd) linii zależności
- $\rho_{X,Y}$ oddaje tylko zależność liniową
- Dla zależności po środku $\rho_{X,Y}$ jest niezdefiniowane, bo $\mathrm{Var}(Y) = 0$.

Korelacja rangowa Spearmana

Zamiast wartości liczbowych X i Y rozważamy rangi obserwacji x,y

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

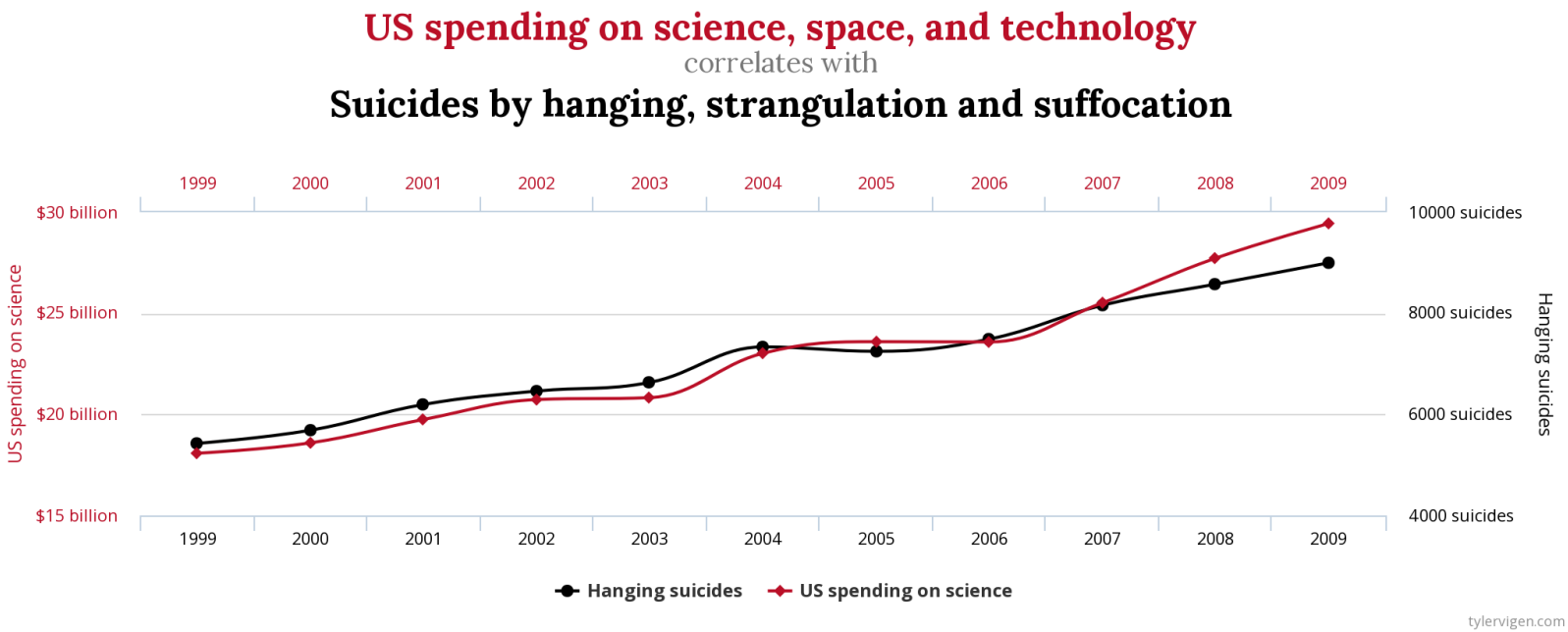
$$d = X_i - Y_i$$

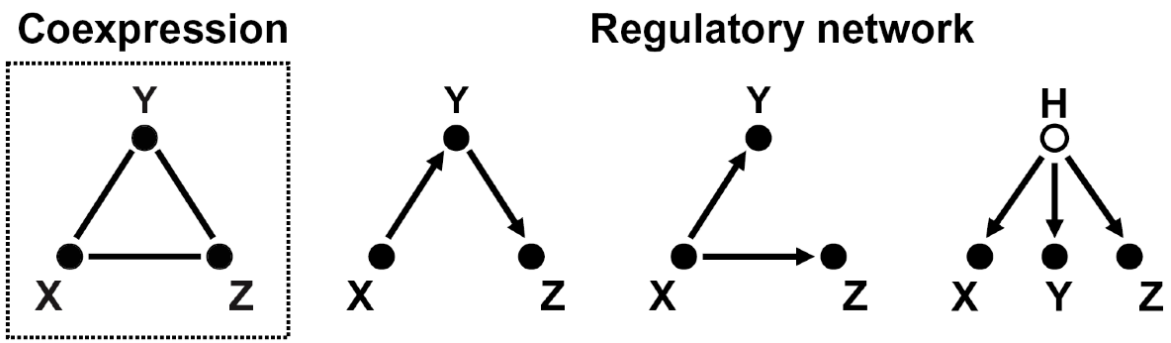


Interpretacja korelacji

- Korelacja liniowa: mierzy, ile zmienności jednej zmiennej może być wytłumaczone przez liniową zależność od drugiej zmiennej
- Korelacja rangowa: mierzy, w jakim stopniu, gdy jedna zmienna rośnie, to druga też wzrasta, bez konieczności by wzrost ten był wyrażony zależnością liniową
- Korelacja to nie to samo co
  - zależność zmiennych losowych (pojęcie ogólniejsze)
  - związek przyczynowo-skutkowy (inne pojęcie)

Correlation does not mean causation





---

http://www.tylervigen.com/spurious-correlations

Markowetz and Spang, Inferring cellular networks – a review (2007)