Historia Rachunku Prawdopodobieństwa i Statystyki
WYKŁAD XI: Historia i przykłady błedów statystycznych

MiNI PW, semestr zimowy

Mathematical statistics and probability is hard. It often involves what, at a first glance, involves complicated calculations and the sheer volume of data coming out of some studies can often be hard to interpret, even if you know all of the mathematics behind it. Although it is important to understand the math, it is equally important (or perhaps even more important) to understand what the results mean and don't mean. It is easy to get dazzled by fancy mathematics or over-interpret results to mean something they really do not. Therefore, a basic understanding of statistical fallacies should be a part of every scientific skeptics toolbox or baloney detection kit.

post na stronie Debunking Denialism

Problemy z warunkowaniem

Rak odbytu $P(C) = 0.3\%$ (prewalencja)
Czułość testu diagnostycznego T: 50%
FDR= P(test dodatni$|C^c$)=3%.
Jakie prawdopodobieństwo $P(C = 1|T = +)$?

$$P(T = +) = P(T = +|C)P(C) + P(T = +|C^c)P(C^c) =$$

$$= (1/2)(3/1000) + (3/100)(997/1000) \approx (3/100)$$

Zatem

$$P(C = +1|T = +) = \frac{P(T = +|C = 1)P(C = 1)}{P(T = +)} \approx \frac{(1/2)(3/1000)}{(3/100)} \approx (5/100)$$

pytanie zadane niemieckim lekarzom z 14 letnim stażem.
Odpowiedzi od 1% do 99%, z czego 50%:
odpowiedź albo 50% (czułość) albo 47% (czulość - FDR)
Test musi mieć znacznie mniejszy FDR przy wykrywaniu rzadkich chorób
!!!

Modus tolens: wersja losowa ?

Modus tolens:

$$B \to A^c \equiv A \to B^c$$

nie ma wersji losowej !!

$$P(A^c|B) = P(B^c|A) \equiv P(A|B) = P(B|A)$$

Jeśli człowiek $\to$ nie prezydent USA
Prezydent USA $\to$ nie człowiek ...

Inne problemy z warunkowaniem.. Zdarzenia sprzyjające

A jest sprzyjające zajściu B, jeśli

$$P(B|A) > P(B)$$

Częsty błąd:

$$P(A|B) - \text{duże} \rightarrow P(B|A) > P(B)$$

Typowy przykład
B - wypadek narciarski w Szwajcarii z udziałem narciarza nie-Szwajcara
A - narciarz niemiecki w Szwajcarii
$P(A|B)$ -duże (najwięcej narciarzy nie-Szwajcarów w Szwajcarii to Niemcy), ale z tego nie wynika, że

$$P(\text{wypadek}|\text{narciarz}: \text{ Niemiec}) > P(\text{wypadek})$$

czyli niekoniecznie mamy

$$P(B|A) > P(B)$$

Zdarzenia sprzyjające, ciąg dalszy

Faktycznie czy

$$P(B|A) = \left(\frac{P(A|B)}{P(A)}\right)P(B) > P(B)\,?$$

Tylko wtedy gdy

$$\frac{P(A|B)}{P(A)} > 1.$$

$P(A|B)/P(A)$ nie musi być większe od 1, jeśli $P(A)$ duże (większość narciarzy nie-Szwajcarów w Szwajcarii to Niemcy)

Tytuły gazetowe

- Uwaga na niemieckich narciarzy
- Chłopcy bardziej narażeni na ryzyko na rowerach
- Piłka nożna najniebezpieczniejszym sportem
- Owczarki alzackie to najniebezpieczniejsze psy
- Zacisze domowe czarnym punktem ?

dane British Home Office
1984-1988: 1221 zabójstw kobiet w GB, z tego 44% zabitych przez mężów/partnerów, 18% przez innych krewnych, 18% przez znajomych i 14% przez nieznajomych.
$A = \{\text{mąż}\}$, $B - \{\text{morderstwo}\}$. $P(A|B)$-duże.
Czy wynika z tego, że

$$P(\text{morderstwo}|\text{mąż}) > P(\text{morderstwo})$$

Prosecutor's fallacy

Twierdzenie Bayesa i prawdopodobieństwo warunkowe jest często źle interpretowane i zrozumiane. Typowy przykład to tzw. błędne przekonanie oskarżyciela: **Prosecutor's fallacy**.
E - dowody (evidence) I - niewinny (innocent)
Powinniśmy wnioskować:
$P(I|E)$ -małe $\rightarrow$: oskarżony może być winny.
Prosecutor's fallacy: Zastąpienie $P(I|E)$ poprzez $P(E|I)$. Mamy

$$P(I|E) = \frac{P(E|I)P(I)}{P(E)}.$$

Jeśli $P(I)/P(E) >> 1$ to $P(I|E) >> P(E|I)$.

Sprawa Sally Clark

Jeden najbardziej znanych (i tragicznych) przykładów: ekspertyza R. Meadowsa na procesie Sally Clark (1999) dotycząca dwóch przypadków nagłego zgonu niemowląt SIS. Medows sformułował tzw **syndrom M*ü*nhausena przez przeniesienie**: opiekunka/matka pozoruje chorobę dziecka, aby zwrócić na siebie uwagę. Dlatego:
unless proven otherwise one cot death is tragic, two is suspicious and three is a murder.
Sally Clark straciła dwóch synów (w wieku 11 i 8 tygodni). Meadows ocenił

$$P(I|E) \approx P(E|I) = (1/8500)^2 = 1/(73 \times 10^6)$$

na podstawie tego orzeczenia została skazana, wypuszczona z więzienia dopiero po 3 latach.

Inny przykład błednego rozumowania sądowego

Próbka DNA z miejsca przestępstwa jest porównanana z bazą danych zawierającą DNA 20 000 osób.

$$P(\text{dwie losowe próby są takie same}) = 1/10000.$$

Znaleziono zgodność z osobą X w bazie.
Czy to oznacza, że

$$P(\text{X niewinny}) = 1/10000$$

Oczywiście nie ! Prawdopodobieństwo, że w bazie znajdziemy przynajmniej jedną przypadkową zgodność

$$1 - (1 - \frac{1}{10000})^{20000} = 86\%$$

Problem wielokrotnego testowania

Paradoks Simpsona

Możliwa jest sytuacja:

$$P(A|B \cap C) > P(A)$$

i

$$P(A|B \cap C^c) > P(A)$$

ale

$$P(A|B) < P(A)$$

Agregacja danych zmienia charakter zależności
Przykłady: prawdopodobieństwo zgonu na raka w Niemczech bylo większe w 2001 niż w 1970.
Rozbicie na kategorie wiekowe i płcie pokazałoo, że w kazdej kategorii prawdopodobieństwo faktycznie spadło

Sprawa O.J. Simpsona

W 1994 O.J. Simpson został oskarżony o zamordowanie swojej byłej żony i jej partnera. W czasie trwania małżeństwa wielokrotnie O.J.S. używał przemocy fizyczniej wobec żony. Oskarżyciel w procesie użył tego jako argumentu wskazującego na winę O.J.S. argumentując, że:

uderzenie jest wstępem do zabójstwa

Obrońca O.J.S, Allan Dershowitz, argumentował, że

$$P(K|przemoc) = 1/2500$$

gdzie K- zdarzenie, że mąż zabił żonę.
Ale roozpatrywana sytuacja wymaga innego warunkowania:

$$A = \{przemoc\ i\ zabicie\ kobiety\}$$

.

$$P(K|A) = 8/9\,!$$

(Good (1996), obliczone na podstawie statystyk policyjnych)

Problem wielokrotnego testowania

Rozpatrzmy nowy płyn do czyszczenia składajacy sie ze 100 składników. Przepisy stwierdzaja, ze mieszanina jest bezpieczna jesli kazdy ze składników jest bezpieczny. Testujemy 100 komponentów na karcinogennosc testem na poziomie 5 procent.Jesli testy sa niezalezne, w 99.4 procentach ($= 100 \times (1 - 0.95^{100})$) przypadków płyn zostanie uznany za niebezpieczny dla zdrowia i niedopuszczony do obrotu .

Problem wielokrotnego testowania:Ig-Nobel 2009

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

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

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

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 a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

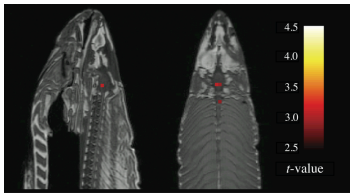
Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

GLM RESULTS

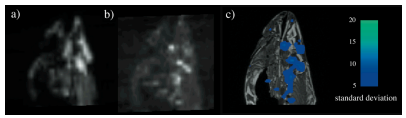


A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

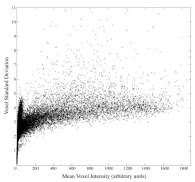


To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.



Outcome switching. Study 329, Paxil

Lek antydepresyjny Paxil (GlaxoSmithKline) często stosowany do leczenia depresji u młodzieży. na początku XXI wieku stwierdzono, że może nasilać tendencje samobójcze. Analiza testu klinicznego tego leku Study 329 wykazała, ze w czasie jego trwania dokonano zmiany celu badania (outcome switching) wobec nieosiągania celów założonych. W efekcie wyniki badania miały wątpliwą wartość. W 2003 GSK zapłaciło 3 mld dolarów za zmiany w różnych testach klinicznych.
COMPare: organizacja analizująca poprawność testów klinicznych.
Zanalizowała 67 testów klinicznych, 9: bez zarzutu, 58 pozostałych
błedy: 300 wyników, które miały być raportowane, nie bylo, 357, które nie miały być raportowane, zostały dodane.
Większość listów do redakcji pozostało bez odpowiedzi..
http://www.economist.com/news/science-and-technology/21695381-too-many-medical-trials-move-their-goalposts-halfway-through-new-initiative

Duży FDR

Klasyfikacja pasażerów metro w Pekinie na podstawie tras (trajektoria dana przez sygnał z karty telefonu komórkowego).
Dwie kategorie: zwykły pasażer i złodzej kieszonkowy operujący w pewnym rejonie.
True selection rate PSR , czułość, **recall** t-zbiór złodziei kieszonkowych)

$$PSR = \frac{|t \cap \hat{t}|}{|t|} = 0.93$$

ale False Discovery Rate (1- **precision**)

$$FDR = \frac{|t^c \cap \hat{t}|}{|\hat{t}|} = 0.92$$

(13-tu na 14-tu zatrzymanych było niewinnych) Ogromny problem czasów Big Data: koszty społeczne metod stosowanych dla dużych danych.
http://www.economist.com/news/science-and-technology/21705296-artful-dodger-your-time-may-be-up-cutpurse-capers

FDR wykrywacza kłamstw

W 2002 roku przeprowadzono badania statystyczne wykrywacza kłamstw. Stwierdzono, że przy czułości 0.8 jego FDR wynosi również około 0.8 i zalecono nie używania wykrywacza kłamstw do rutynowego sprawdzania prawdomówności pracowników.

Liczy się tylko sukces ..

Ioannidis (2005) WHY MOST PUBLISHED RESULTS ARE FALSE ?
Wyniki publikowane są statystycznie obciązone w tym sensie, że ponieważ publikuje się tylko wyniki dotyczące odrzuconych hipotez zerowych (a więc potwierdzających założoną hipotezę badawczą) sztucznie podwyższa się frakcję fałszywych odkryć FDR.
Przykład:
Testujemy 1000 hipotez badawczych, z czego 100 jest prawdziwych, przy użyciu testu o mocy 0.8. Odkryjemy 80 spośród 100 prawdziwych hipotez. Jednocześnie spośród 900 hipotez fałszywych przy użyciu testów na poziomie 5% $900 \times 0,05 = 45$ zostanie fałszywie odkrytych. Zatem, spośrod 125 odkryć ok. $1/3$ jest fałszywych. Jesli moc spadnie do 40%, połowa 'odkryć' będzie fałszywa.
Remedium: Publikowanie wyników dotyczacych hipotez badawczych, które się nie sprawdziły. W ten sposób zwiekszamy potencjalnie liczbę prawdziwych hipotez, które testujemy...

Wiedza ogólna i wspólna

Wiedza ogólna (**mutual knowledge**) każda osoba z grupy, wie, że A.
Wiedza wspólna (**common knowledge**) każda osoba z grupy, wie, że każda osoba z grupy wie, że A.
Często potrzebujemy wiedzy wspólnej, a nie ogólnej:
Każdy kierowca wie, że nie wolno przejeżdzać przejścia przy czerwonym świetle, ale
do prowadzenia potrzeba mu świadomości, że każdy inny kierowca wie, że nie przejeżdża się przejścia na czerwonym świetle.

Zmiana statusu wiedzy

Red hat problem 3 osoby w czerwonych kapeluszach widzą kapelusze innych, ale nie swój. Kapelusze są przydzielane losowo, niebieskie i czerwone.
Z wybiciem każdej następnej godziny, jeśli wiedzą, jaki kapelusz mają, wychodzą.
Nikt nie wychodzi.
Dodatkowa wiedza: ' Co najmniej jedna osoba ma czerwony kapelusz '
Po trzecim uderzeniu zegara wszyscy wychodzą.

Inny wariant

Grupa 3 osób ma zgadnąć kolory swoich kapeluszy: każdy ma podać kolor swego, lub spasować.
Grupa wygrywa: co najmniej jedna osoba zdoła podać kolor prawidłowo (reszta może spasować)
Strategia losowego zgadywania: czerwony, niebieski:
wygrana 1/8.
Strategia losowego zgadywania: czerwony, niebieski, pas.
wygrana: 0.259.
Strategia dająca największe prawdopodobieństwo wygranej:

Inny wariant

Grupa 3 osób ma zgadnąć kolory swoich kapeluszy: każdy ma podać kolor swego, lub spasować.
Grupa wygrywa: co najmniej jedna osoba zdoła podać kolor prawidłowo (reszta może spasować)
Strategia losowego zgadywania: czerwony, niebieski:
wygrana 1/8.
Strategia losowego zgadywania: czerwony, niebieski, pas.
wygrana: 0.259.
Strategia dająca największe prawdopodobieństwo wygranej:
Dwie osoby, które widzę, mają kapelusze w tym samym kolorze, zgaduję, że mój ma kolor przeciwny.
Dwie osoby, które widzę, mają kapelusze w różnych kolorach, pasuję.
Prawdopodobieństwo wygranej: 6/8=3/4

3 osoby pomyliły się łacznie 6 razy i 6 razy miały łacznie rację. Ale pomylily sie wspólnie, co wyeliminowało tylko dwa przypadki z ośmiu !

Table 6.1.

A	B	C	A	B	C	Outcome
Red	Red	Red	GBW	GBW	GBW	Lose
Red	Red	Blue	Pass	Pass	GBC	Win
Red	Blue	Red	Pass	GBC	Pass	Win
Red	Blue	Blue	GRC	Pass	Pass	Win
Blue	Red	Red	GBC	Pass	Pass	Win
Blue	Red	Blue	Pass	GRC	Pass	Win
Blue	Blue	Red	Pass	Pass	GRC	Win
Blue	Blue	Blue	GRW	GRW	GRW	Lose

GBW, guess blue: wrong; GBC, guess blue: correct; GRW, guess red: wrong; GRC, guess red: correct.

Dla grupy n osobowej:
Prawdopodobieństwo wygrania dla optymalnej strategii: $n/(n+1)$.

Pytanie:
W czasie drugiej wojny światowej najwięcej bombowców amerykańskich wracających do baz miało przestrzelone skrzydła. Które części bombowców należalo wzmocnić ?