

THE THEORY THAT WOULD NOT DIE.

***(TŁUM. „TEORIA, KTÓRA NIE CHCIAŁA UMRZEĆ”, „TEORIA, KTÓRA NIE DAWAŁA SIĘ ZABIĆ”,
„NIEŚMIERTELNA TEORIA” ;)***

ON THE BAYESIAN APPROACH TO STATISTICS

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

THE THEORY THAT WOULD NOT DIE.

***(TŁUM. „TEORIA, KTÓRA NIE CHCIAŁA UMRZEĆ”, „TEORIA, KTÓRA NIE DAWAŁA SIĘ ZABIĆ”,
„NIEŚMIERTELNA TEORIA” ;)***

ON THE BAYESIAN APPROACH TO STATISTICS

(A VERY ROUGH = FAR FROM PERFECT, YET HOPEFULLY INSPIRATIONAL INTRODUCTION... ;)

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

THE THEORY THAT WOULD NOT DIE.

***(TŁUM. „TEORIA, KTÓRA NIE CHCIAŁA UMRZEĆ”, „TEORIA, KTÓRA NIE DAWAŁA SIĘ ZABIĆ”,
„NIEŚMIERTELNA TEORIA” ;)***

ON THE BAYESIAN APPROACH TO STATISTICS

(A VERY ROUGH = FAR FROM PERFECT, YET HOPEFULLY INSPIRATIONAL INTRODUCTION... ;)

→ Czyt.: *Audycja zawiera lokowanie produktu ^^*

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

THE THEORY THAT WOULD NOT DIE.

*(TŁUM. „TEORIA, KTÓRA NIE CHCIAŁA UMRZEĆ”, „TEORIA, KTÓRA NIE DAWAŁA SIĘ ZABIĆ”,
„NIEŚMIERTELNA TEORIA” ;)*

ON THE BAYESIAN APPROACH TO STATISTICS

(A VERY ROUGH = FAR FROM PERFECT, YET HOPEFULLY INSPIRATIONAL INTRODUCTION... ;)

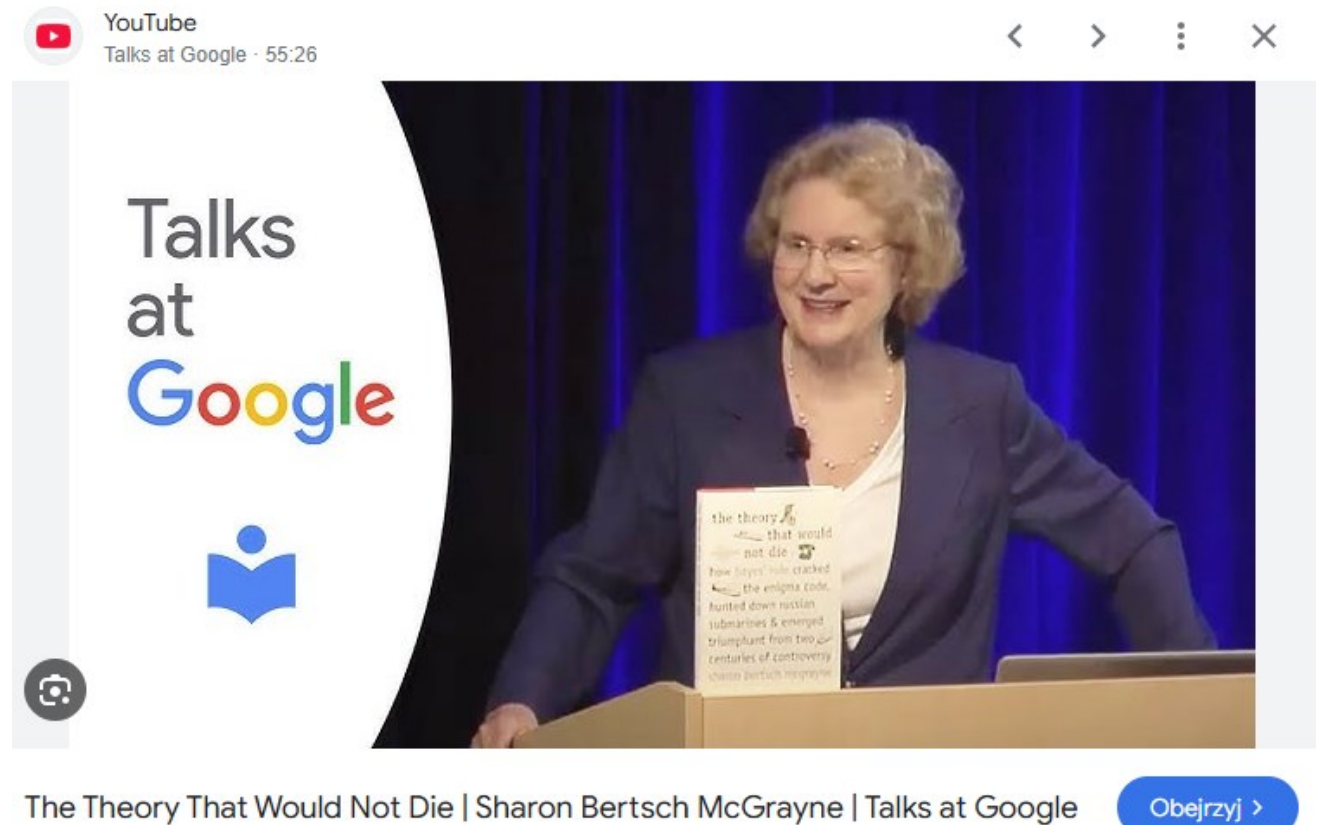
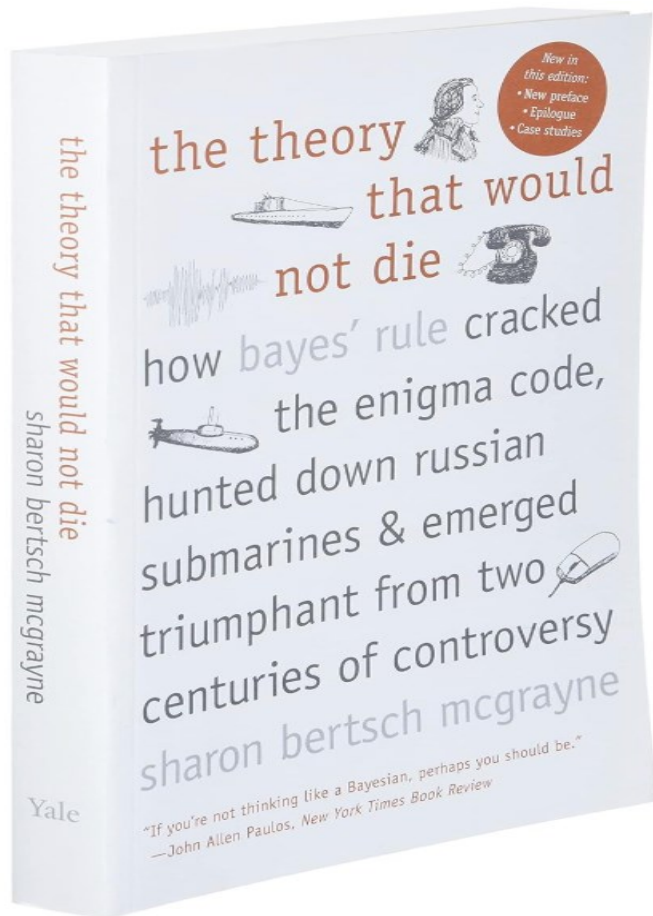
→ To procrastinating perfectionists: *Zrobione jest lepsze od doskonałego ;P*

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

O książce...

Sharon B. McGrayne (2011), *The theory that would not die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, New Haven & London



O książce...

Liczba wyników w przeglądarce internetowej dla słowa „Bayesian”:

- 2001: 100 000
- 2011: 14 000 000

Prawdopodobieństwo – *revisited*

- Pytanie 1: *Jakie jest prawdopodobieństwo, że dziś między 21:00 a 22:00 spadnie deszcz?*

Prawdopodobieństwo – *revisited*

➤ Pytanie 2: *Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?*

Prawdopodobieństwo – *revisited*

➤ Pytanie 2: *Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?*

→ Nieco bardziej poprawnie:

Ile wynosi odsetek wyrzuconych orłów w nieskończonej liczbie rzutów (= parametr $\phi \in [0,1]$)?

Prawdopodobieństwo – *revisited*

➤ Pytanie 2: *Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?*

→ Nieco bardziej poprawnie:

Ile wynosi odsetek wyrzuconych orłów w nieskończonej liczbie rzutów (= parametr $\phi \in [0,1]$)?

→ Mamy wyniki $T = 10$ rzutów: $y_1, y_2, \dots, y_T$, gdzie $y_t = \mathbf{0}$ (reszka) bądź $y_t = \mathbf{1}$ (orzeł) ($t = 1, 2, \dots, T$):

$$y_{(T \times 1)} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

Prawdopodobieństwo – *revisited*

➤ Pytanie 2: Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?

→ Nieco bardziej poprawnie:

Ile wynosi odsetek wyrzuconych orłów w nieskończonej liczbie rzutów (= parametr $\phi \in [0,1]$)?

→ Mamy wyniki $T = 10$ rzutów: $y_1, y_2, \dots, y_T$, gdzie $y_t = \mathbf{0}$ (reszka) bądź $y_t = \mathbf{1}$ (orzeł) ($t = 1, 2, \dots, T$):

$$y_{(T \times 1)} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

⇒ Czy chciał(a)byś zrewidować swoją „wiedzę” o parametrze ϕ ?

Prawdopodobieństwo – *revisited*

➤ Pytanie 2: Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?

→ Nieco bardziej poprawnie:

Ile wynosi odsetek wyrzuconych orłów w nieskończonej liczbie rzutów (= parametr $\phi \in [0,1]$)?

→ Mamy wyniki $T = 10$ rzutów: $y_1, y_2, \dots, y_T$, gdzie $y_t = \mathbf{0}$ (reszka) bądź $y_t = \mathbf{1}$ (orzeł) ($t = 1, 2, \dots, T$):

$$y_{(T \times 1)} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

⇒ Czy chciał(a)byś zrewidować swoją „wiedzę” o parametrze ϕ ?

➔ Wnioskowanie statystyczne o ϕ (a u jego podstaw – pojęcie prawdopodobieństwa)

Prawdopodobieństwo – *revisited*

➤ Dwie (wiodące) interpretacje / sposoby pojmowania prawdopodobieństwa:

1) „**Klasyczna**” – prawdopodobieństwo „obiektywne”, częstościowe (definicja wg Andrieja Kołmogorowa), rozumiane jako **idealizacja częstości względnej**;

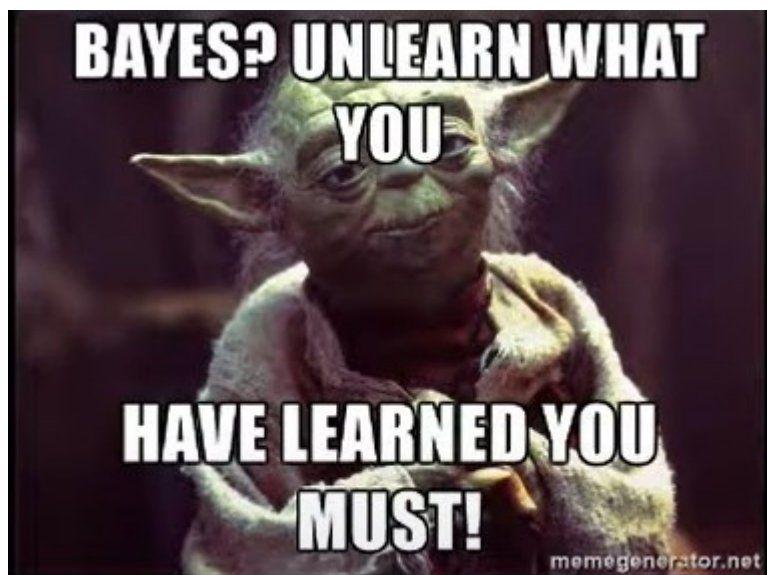
→ **Prawdopodobieństwo „klasyczne”** – zwane także „fizycznym”

→ **ALE** prawdopodobieństwo w istocie **NIE jest jakąś cechą fizyczną, immanentnie** związaną z i określającą jakiś obiekt (np. monetę, urnę czy kurs walutowy), zdarzenie losowe czy zdanie logiczne (hipotezę)

Prawdopodobieństwo – *revisited*

➤ Dwie (wiodące) interpretacje / sposoby pojmowania prawdopodobieństwa:

2) **Bayesowska** – prawdopodobieństwo jest miarą **subiektywnego** stopnia przekonania/pewności co do zajścia jakiegoś zdarzenia (losowego, tj. obarczonego niepewnością) czy też co do prawdziwości pewnej hipotezy (tj. zdania logicznego)



Prawdopodobieństwo – *revisited*

→ Bayesowska = subiektywna interpretacje prawdopodobieństwa:

→ Prawdopodobieństwo jest jak... serce – każdy ma swoje ;)

Frequentists now shouts: *Abomination! Heresy! Blasphemy!* 😂

Prawdopodobieństwo – *revisited*

→ Bayesowska = subiektywna interpretacje prawdopodobieństwa:

→ Prawdopodobieństwo jest jak... serce – każdy ma swoje ;)

Frequentists now shouts: *Abomination! Heresy! Blasphemy!* 😂

→ ALE **nie** oznacza to jakiegoś **dogmatyzmu**

(„*Moje prawdopodobieństwo jest „najmojsze”, i za Chiny Ludowe go nie zmienię!*”)

Wręcz przeciwnie!

Prawdopodobieństwo – *revisited*

→ Prawdopodobieństwo ma dwa „źródła”:

1. wychodzimy od **prawdopodobieństwa *a priori*** (czyli przed wglądem w dane, w empirię, przed doświadczeniem, eksperymentem), a następnie...
2. ...uwzględniamy **informację zawartą w danych** (ang. *data, statistical evidence*)

→ końcowym efektem jest **prawdopodobieństwo *a posteriori*** (czyli po wglądzie w dane):

Prawdopodobieństwo a priori + Dane („evidence”) = Prawdopodobieństwo a posteriori

= **Bayesian learning** = aktualizacja wiedzy – „mojej” wiedzy, rozumianej tu jako „moje” wyobrażenie o kształtowaniu się **niepewności** odnośnie danego zdarzenia losowego/hipotezy/parametru statystycznego

Prawdopodobieństwo – *revisited*

→ **Paradygmat bayesowski** oferuje **w pełni formalny (obiektywny!)**, zgodny z podstawowymi zasadami rachunku prawdopodobieństwa (wychodząc od definicji wg. Pierre Simone de Laplace), wewnętrznie spójny (koherentny) – choć też nie pozbawiony pewnych paradoksów – **sposób modyfikacji mojej wiedzy *a priori* w *a posteriori*** (tj. po doświadczeniu, po wglądzie w dane/obserwacje), czyli sposób wyznaczania mojej końcowej wiedzy (a raczej **niepewności!**), poprzez połączenie wiedzy *a priori* z informacją płynącą z obserwacji

⇒ Zgodne z naszą neurofizjologią procesu uczenia się

→ Prawdopodobieństwo – **subiektywne**, ale sposób przejścia od *a priori* do *a posteriori* – **obiektywny!**

Prawdopodobieństwo – *revisited*

→ Powtórzmy:

a. Prawdopodobieństwo = matematyczny sposób wyrażania „mojego” (zatem subiektywnego) stopnia przekonania (a tym samym stopnia **niepewności**)

→ Zatem na gruncie statystyki bayesowskiej prawdopodobieństwo ma interpretację subiektywistyczną

→ Matematyczny dowód istnienia i jednoznaczności prawdopodobieństwa jako właśnie miary subiektywnego stopnia przekonania nie jest trywialny, ALE JEST :) (XX w.: Morris H. DeGroot, Bruno de Finetti, Leonard Savage)

b. Prawdopodobieństwo podlega modyfikacji (w wyniku procesu „uczenia się”, aktualizacji wiedzy; ang. *Bayesian learning*) – od wiedzy/informacji wstępnej (tzw. *a priori*, czyli przed wglądem w dane/obserwacje) do wiedzy końcowej (tj. *a posteriori*, czyli po uwzględnieniu informacji zawartej w danych; ang. *evidence*)

Dwa (wiodące) podejścia do wnioskowania statystycznego

➤ Bazują na, odpowiednio, „klasycznej” i bayesowskiej interpretacji prawdopodobieństwa:

1) Statystyka „klasyczna”

2) Statystyka bayesowska („bayesizm”)

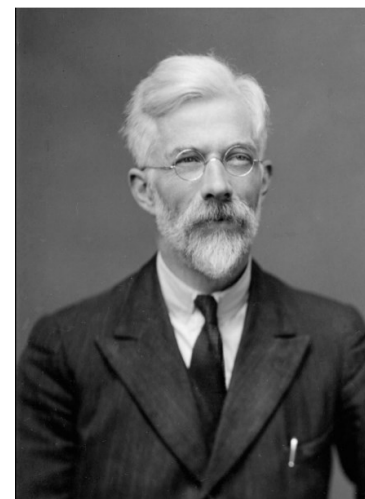
Statystyka „klasyczna”

1. **Statystyka „klasyczna”** – inaczej: niebayesowska, częstościowa (ang. *frequentist statistics/inference*, *frequentist approach to statistics*):



Jerzy Spława Neyman (1894-1984)

→ Przedziały ufności



Ronald Fisher (1890-1962)

→ Metoda Największej Wiarygodności (MNW)

Statystyka „klasyczna”

→ Ad przykład z rzutami monetą:

$\phi \in [0, 1]$ = **parametr statystyczny** – na gruncie **klasycznym** = nieznaną **stałą** wielkość (*nie* zmienna losowa):

$$\phi = \lim_{T \rightarrow \infty} \frac{\text{liczba orłów w } T \text{ rzutach}}{T}$$

Dane: $y_{(T=10 \times 1)} = (0, 1, 1, 1, 0, 1, 1, 1, 0, 1)'$

Model próbkowy: $\forall_{t=1, \dots, T} y_t \sim \text{ii Dwupunktowy}(\phi)$

$$p(y_t; \phi) = \Pr(y_t; \phi) = \phi^{y_t} (1 - \phi)^{1-y_t} = \begin{cases} \phi, & \text{dla } y_t = 1 \\ 1 - \phi, & \text{dla } y_t = 0 \end{cases}$$

gdzie $p(x)$ oznacza, w ogólności, gęstość rozkładu zmiennej losowej x (jakakolwiek by ona nie była)

→ **Model próbkowy = łączny rozkład obserwacji** (zmiennych losowych) przy ustalonej (nieznanej) wartości parametru:

$$p(y; \phi) = p(y_1, y_2, \dots, y_T; \phi) = p(y_1; \phi) \times p(y_2; \phi) \times \dots \times p(y_T; \phi) = \phi^{\sum y_t} (1 - \phi)^{T - \sum y_t}$$

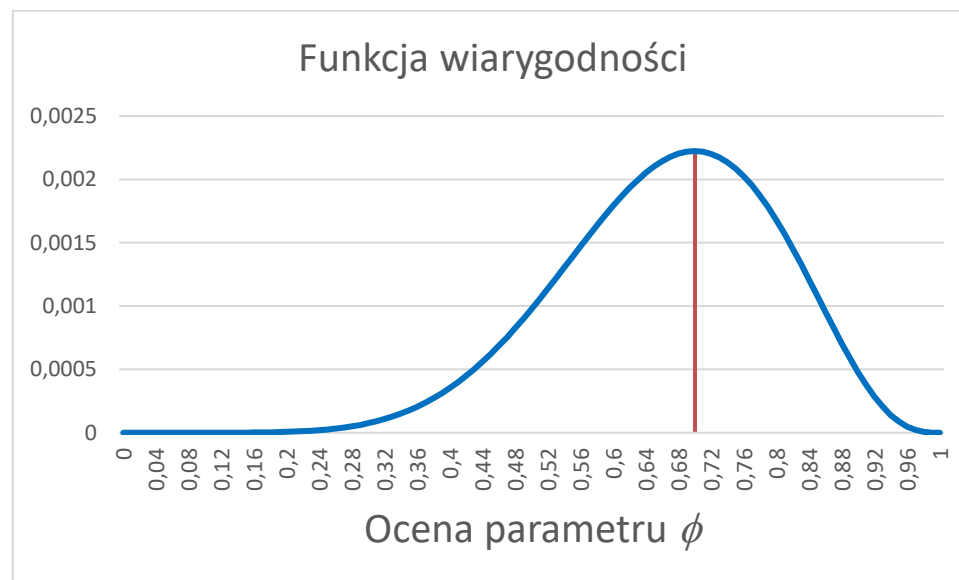
Statystyka „klasyczna”

Estymacja parametru ϕ – **Metoda Największej Wiarygodności:**

Funkcja wiarygodności (ang. *likelihood function*): $L_y(\phi) = p(y; \phi) = \phi^{\sum y_t} (1 - \phi)^{T - \sum y_t}$

Estymator MNW:

$$\hat{\phi} = \hat{\phi}(y) = \arg \max_{\phi \in (0,1)} L_y(\phi) = \text{TU analitycznie} = \frac{\text{liczba orłów w } T \text{ rzutach}}{T} = \frac{\sum y_t}{T}$$



→ $\hat{\phi}$ = funkcja obserwacji $(y_1, \dots, y_T)$ „jeszcze” jako **zmiennych losowych** => $\hat{\phi}$ – **zmienna losowa**

versus

→ $\hat{\phi}$ = funkcja obserwacji $(y_1, \dots, y_T)$ „już” jako **realizacji** zmiennych losowych => $\hat{\phi}$ – realizacja **zmiennej losowej** (czyli pojedyncza liczba) → w rozważanym przykładzie: $\hat{\phi} = 7/10 = 0,7$

Statystyka „klasyczna”

➤ Rozkład estymatora:

$$\hat{\phi} \sim^{T \rightarrow \infty} \text{approx. } N\left(\phi, \frac{\phi(1-\phi)}{T}\right)$$

→ Tu – **przybliżony**, bo przecież $\phi \in [0, 1]$

→ Tu – **asymptotyczny**, tj. przy $T \rightarrow \infty$

⇒ **W ogólności: MNW zazwyczaj tylko asymptotycznie** ma rozkład normalny (i inne pożądane własności), tj. przy $T \rightarrow \infty \Rightarrow$ nie posiada własności małopróbkowych
(Nieliczne **wyjątki**, w tym Klasyczny Model Normalnej Regresji Liniowej, gdzie MNW = MNK)

→ Estymacja punktowa:

- $\hat{\phi} = \frac{\sum y_t}{T} = 0,7 \rightarrow$ ocena punktowa ϕ – pojedynczy punkt w przestrzeni parametrów ($\phi \in [0, 1]$)
- $D(\hat{\phi}) = \sqrt{\frac{\hat{\phi}(1-\hat{\phi})}{T}} \approx 0,145 \rightarrow$ błąd średni szacunku (przybliżony, bo asymptotyczny) – mierzy **niepewność estymacji** punktowej, ale tylko w otoczeniu maximum funkcji wiarygodności

→ Co, gdy funkcja wiarygodności jest *wielomodalna*?

Statystyka „klasyczna”

➤ Rozkład estymatora:

$$\hat{\phi} \sim_{T \rightarrow \infty} \text{approx. } N\left(\phi, \frac{\phi(1-\phi)}{T}\right)$$

→ Wnioskowanie statystyczne o ϕ :

- Testowanie hipotez ϕ , np.: $H_0: \phi = \phi^*$ vs. $H_1: \phi \neq \phi^* \rightarrow p\text{-value}$ jako funkcja oceny ϕ
- $100(1 - \alpha)\%$ -wy przedział ufności: $(\hat{\phi} - z_\alpha D(\hat{\phi}); \hat{\phi} + z_\alpha D(\hat{\phi}))$

$\hat{\phi}$ – zmienna losowa $\Rightarrow \Pr\left\{\phi \in (\hat{\phi} - z_\alpha D(\hat{\phi}); \hat{\phi} + z_\alpha D(\hat{\phi}))\right\} \approx 1 - \alpha$; z_α – wartość kryt.

→ Dla konkretnego wektora liczbowego y (konkretnych danych), przy $1 - \alpha = 0,95$, $z_\alpha = 1,96$:

$$(\hat{\phi} - z_\alpha D(\hat{\phi}); \hat{\phi} + z_\alpha D(\hat{\phi})) = (0,416; 0,984)$$

→ Interpretacja: ...

Statystyka „klasyczna”

$$\left(\hat{\phi} - z_{\alpha} D(\hat{\phi}); \hat{\phi} + z_{\alpha} D(\hat{\phi}) \right) = (0,416; 0,984)$$

→ Interpretacja:

1) Z ufnością / pewnością 95%, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984).

versus

2) Z prawdopodobieństwem $1 - \alpha = 0,95$, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984)

Statystyka „klasyczna”

$$\left(\hat{\phi} - z_{\alpha} D(\hat{\phi}); \hat{\phi} + z_{\alpha} D(\hat{\phi}) \right) = (0,416; 0,984)$$

→ Interpretacja:

1) Z ufnością / pewnością 95%, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984).

versus

2) Z prawdopodobieństwem $1 - \alpha = 0,95$, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984)

versus

3) Przedział (0,416; 0,984) to pojedyncza **realizacja** przedziału losowego, w któryYM (nie w któryEJ) parametr ϕ znajduje się z prawdopodobieństwem $1 - \alpha = 0,95$.

Statystyka „klasyczna”

$$\left(\hat{\phi} - z_{\alpha} D(\hat{\phi}); \hat{\phi} + z_{\alpha} D(\hat{\phi}) \right) = (0,416; 0,984)$$

→ Interpretacja:

1) Z ufnością / pewnością 95%, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984).

versus

2) Z prawdopodobieństwem $1 - \alpha = 0,95$, prawdziwa wartość parametru ϕ mieści się w przedziale (0,416; 0,984)

versus

3) Przedział (0,416; 0,984) to pojedyncza **realizacja** przedziału losowego, w którym (nie który) parametr ϕ znajduje się z prawdopodobieństwem $1 - \alpha = 0,95$.

Ad 1)

Albo $\phi \in (0,416; 0,984)$, *albo* $\phi \notin (0,416; 0,984)$ → nie ma tu już miejsca na ujęcie probabilistyczne!

→ **Założenie o replikowalności obserwacji**: można uzyskać nieskończenie wiele próbek y (o danej T):

dane z eksperymentu kontrolowanego

versus

dane gospodarcze

Statystyka bayesowska

2. Statystyka bayesowska (ang. *Bayesian statistics/inference, Bayesian approach to statistics*):

- i. **(Reverend) Thomas Bayes** (1702-1760) – angielski matematyk i duchowny; twórcą statystyki bayesowskiej jako tako nie był, ale autorem sławetnego **wzoru Bayesa** (na prawdopodobieństwach warunkowe):

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}$$



- ii. Opublikowany pośmiertnie przez przyjaciela wielebnego Bayesa, **Richarda Price'a** (1763) w artykule pt. „*An Essay towards solving a Problem in the Doctrine of Chances*”
- iii. Bazuje na pierwszej formalnej definicji prawdopodobieństwa, autorstwa **Pierre Simone de Laplace** (1749-1827)
 - **Chronologicznie** podejście bayesowskie jest bardziej „klasyczne” niż podejście „klasyczne” ;)
 - **Odnawiciele bayesizmu**: Harold Jeffreys, Bruno de Finetti, Leonard Savage, Arnold Zellner

Statystyka bayesowska

➤ **Wzór Bayesa:**

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}$$

→ Przykładając go do problemu wnioskowania o parametrze $\phi \in [0, 1]$ na podstawie danych y :

$$p(\phi|y) = \frac{p(y, \phi)}{p(y)} = \frac{p(y|\phi)p(\phi)}{p(y)}$$

gdzie $p(x)$ oznacza, w ogólności, gęstość rozkładu zmiennej losowej x

Statystyka bayesowska

➤ Przejdźmy na ogólny **przypadek wielu parametrów**, $\theta = (\theta_1, \theta_2, \dots, \theta_s)' \in \Theta \subseteq \mathbb{R}^s$:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- $p(y|\theta) = L_y(\theta)$ – **model próbkowy** = informacja o parametrach zawarta w danych
→ = **funkcja wiarygodności**
- $p(\theta)$ – **rozkład *a priori*** = nasza wiedza (obraz **niepewności**) o parametrach **PRZED** wglądem w dane
- $p(y, \theta)$ – **bayesowski model statystyczny** = łączny rozkład wszystkich wielkości
- $p(y)$ – brzegowa gęstość wektora obserwacji
- $p(\theta|y)$ – **rozkład *a posteriori***:
= pełny, łączny rozkład prawdopodobieństwa na całej przestrzeni parametrów
= nasza końcowa (tj. **PO** wglądzie w dane) wiedza (obraz **niepewności**) o parametrach

Statystyka bayesowska

➤ Wzór Bayesa – przypomnijmy:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

→ **Bayes w praktyce:** zadajemy $p(y|\theta)$ (czyli **funkcję wiarygodności** – jak w „klasycznym” podejściu) i $p(\theta)$, a następnie wyznaczamy $p(\theta|y)$:

analitycznie *versus* numerycznie (np. metody MCMC = Markov Chain Monte Carlo)

→ **Estymacja bayesowska** = wyznaczenie **rozkładu *a posteriori*** i jego scharakteryzowanie poprzez:

- wykres(y) funkcji gęstości – na tle rozkładu *a priori* (by ocenić, jak dane zmieniły rozkład *a priori*)
- charakterystyki liczbowe:
 - położenia
 - rozproszenia
 - inne

Statystyka bayesowska

→ Zazwyczaj θ jest wektorem...

⇒ Z wykresem „pitnastowymiarowego” $p(\theta|y)$ sobie raczej nie poradzimy ^^'

⇒ Przechodzimy na **brzegowe rozkłady *a posteriori*** – dla $i = 1, 2, \dots, s$:

$$p(\theta_i|y) = \int_{\Theta_{\setminus i}} p(\theta|y) d\theta_{\setminus i}$$

→ $p(\theta_i|y)$ uwzględnia w sobie niepewność związaną z wszystkimi pozostałymi parametrami
(i to w sposób w pełni formalny i zgodny z rachunkiem prawdopodobieństwa!)

→ Podsumowanie wiedzy o θ_i zawartej w $p(\theta_i|y)$:

- wykres funkcji gęstości $p(\theta_i|y)$ na tle $p(\theta_i)$ (by ocenić, jak dane zmieniły rozkład *a priori*)
- charakterystyki liczbowe $p(\theta_i|y)$:
 - położenia: wartość oczekiwana $E(\theta_i|y)$, modalna $Mo(\theta_i|y)$, mediana $Me(\theta_i|y)$
 - rozproszenia: odchylenie standardowe $D(\theta_i|y)$
 - inne (np. kwantylowy przedziały wiarygodności, obszar najwyższej gęstości *a posteriori*)

Statystyka bayesowska

➤ Zauważmy, że:

- **Nie ma tu estymatorów** = „pośredników” we wnioskowaniu o θ – na gruncie bayesowskim wnioskowanie dotyczy bezpośrednio parametrów θ
- Podstawą i punktem wyjścia estymacji bayesowskiej jest od razu cały rozkład prawdopodobieństwa parametrów = **pełny obraz niepewności** odnośnie ich wartości
 - ⇒ Jeśli jednak chcemy jakieś (bayesowskie) **oceny punktowe**, to i owszem – statystyczna teoria decyzji podpowiada, że mogą nimi być wartości oczekiwane, modalne, mediany i inne.
- **Nie ma tu asymptotyki** – wnioskowanie bayesowskie daje poprawne wyniki dla dowolnej liczby obserwacji:
 - aczkolwiek metody MCMC coś tam z asymptotyki korzystają ^^

Statystyka bayesowska

➤ Ponadto nr 1:

→ Interpretacja kwantylowego przedziału wiarygodności

→ Przy zadanym prawdopodobieństwie *a posteriori* $1 - \alpha$, to stosowny przedział LICZBOWY w $p(\theta_i|y)$:

$$\left(Q_{\frac{\alpha}{2}}(\theta_i|y); Q_{1-\frac{\alpha}{2}}(\theta_i|y) \right)$$

Przykład: przy $1 - \alpha = 0,95$, otrzymano dla parametru θ_1 przedział (2; 3):

$$\Pr(\theta_1 \in (2; 3)|y) = 0,95$$

→ Interpretacja – w pełni probabilistyczna i naturalna: ...

⇒ Bez założenia o replikowalności danych!

⇒ Całe **wnioskowanie bayesowskie** pozostaje **w pełni probabilistyczne** przy danym, konkretnym wektorze obserwacji

Statystyka bayesowska

➤ Ponadto nr 2:

→ **Bayesowska predykcja** – też w pełni probabilistyczna, bo oparta na **całym rozkładzie *a posteriori* prognozowanej wielkości, y^f** (tzw. rozkładzie predyktywnym):

$$p(y^f | y) = \int_{\Theta} p(y^f, \theta | y) d\theta = \int_{\Theta} p(y^f | \theta, y) p(\theta | y) d\theta$$

⇒ Naturalnie wpisuje się w obecnie **zyskujący na popularności nurt** wyznaczania całych **rozkładów prawdopodobieństwa prognozowanej wielkości** (ang. ***density forecasts***)

→ to znacznie więcej niż:

- prognoza punktowa (z błędem średnim predykcji *ex ante*)
- prognoza przedziałowa

⇒ Rozkład predyktywny $p(y^f | y)$ formalnie uwzględnia niepewność o θ

versus

„Klasyczna” prognoza = charakterystyki rozkładu $p(y^f | \theta = \hat{\theta}, y)$

Statystyka bayesowska

➤ Ponadto nr 3:

→ Bayesowskie porównywanie modeli i łączenie wiedzy z alternatywnych modeli

Statystyka bayesowska

➤ Tak, znalazłoby się jeszcze parę innych „ponadto” 😊

Statystyka bayesowska

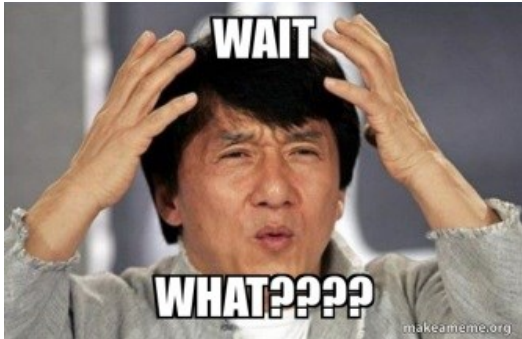
➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

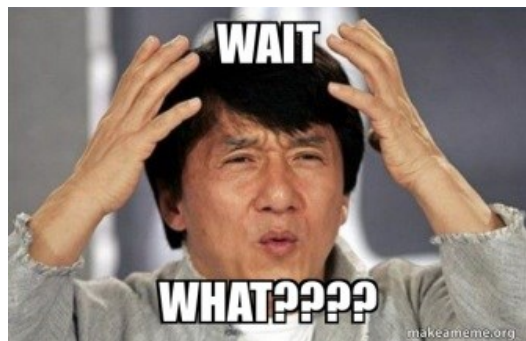


→ „Łejdamynył”! Czy to oznacza, że parametry (θ) są „jednak” **zmiennymi losowymi**?! (Przypomnijmy: na gruncie „klasycznym” są **nieznanymi stałymi**)

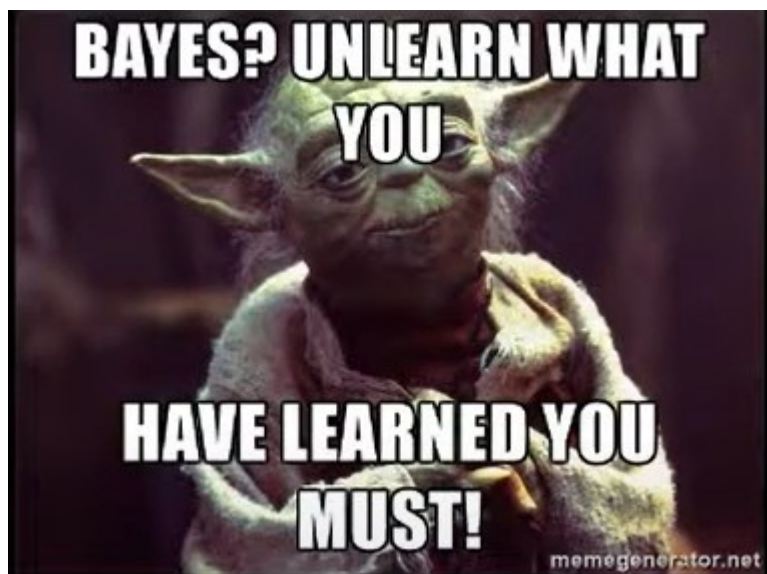
Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$



→ „Łejdamynył”! Czy to oznacza, że parametry (θ) są „jednak” **zmiennymi losowymi**?!

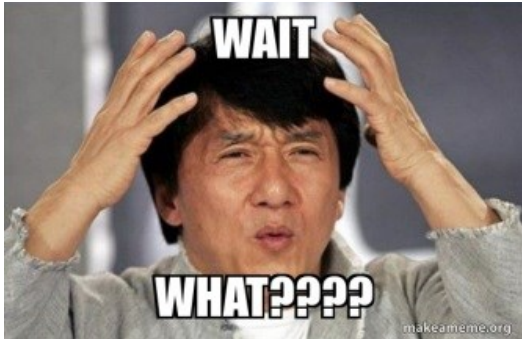


→ „Yep”...

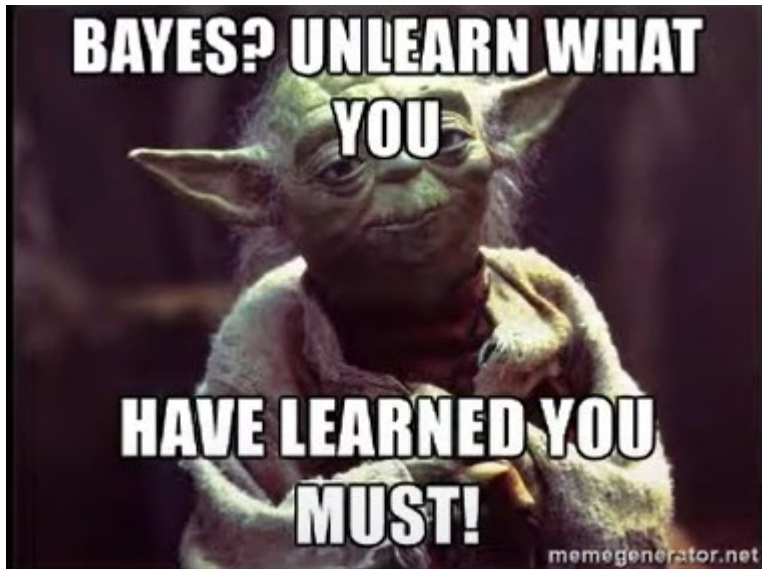
Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$



→ „Łejdamynył”! Czy to oznacza, że parametry (θ) są „jednak” **zmiennymi losowymi**?!

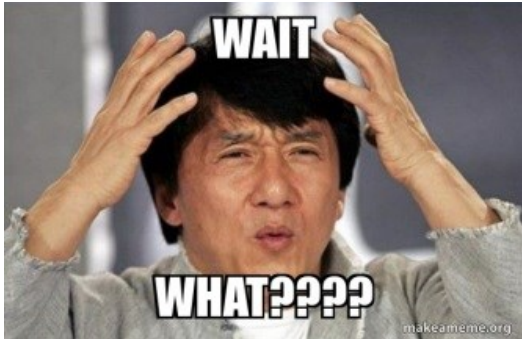


→ „Yep”... i jakoby zarazem „Not yep” 😂🧐

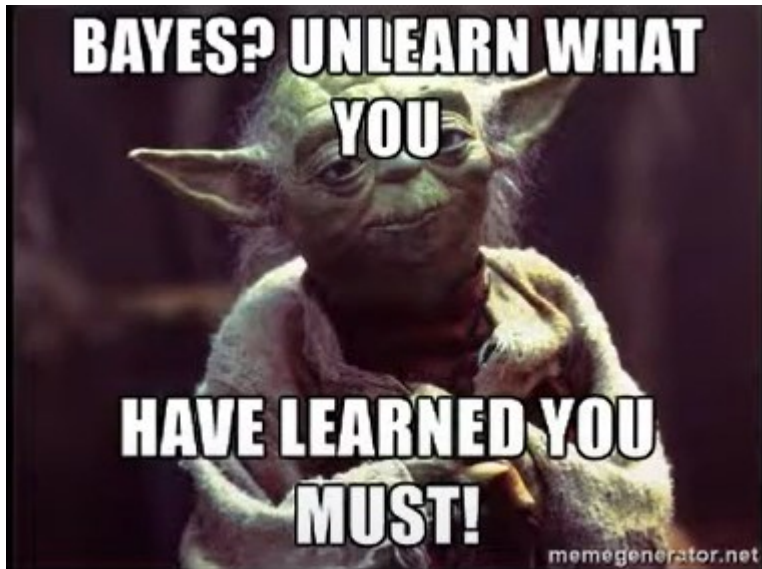
Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$



→ „Łejdamynył”! Czy to oznacza, że parametry (θ) są „jednak” **zmiennymi losowymi**?!



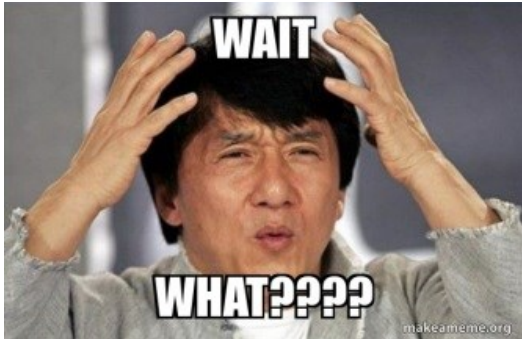
→ „Yep”... i jakoby zarazem „Not yep” 😂😂

Uuu... „szredingerowsko”! 😂

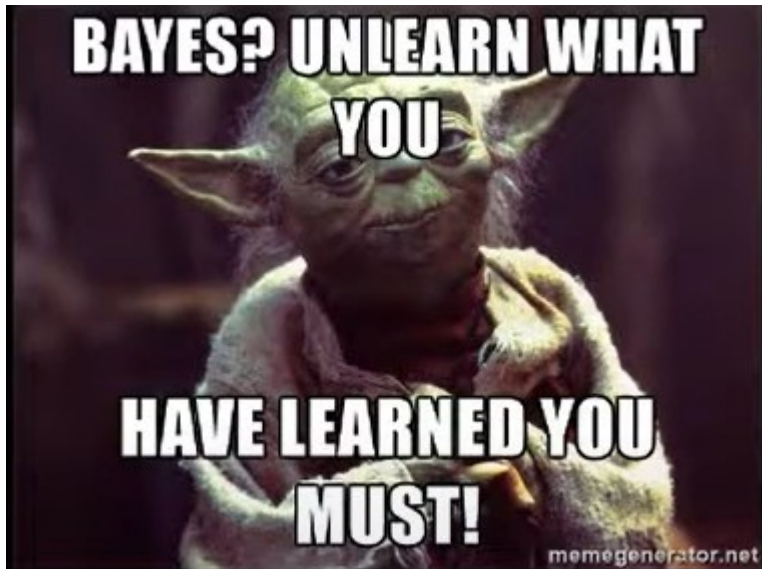
Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$



→ „Łejdamynył”! Czy to oznacza, że parametry (θ) są „jednak” **zmiennymi losowymi**?!



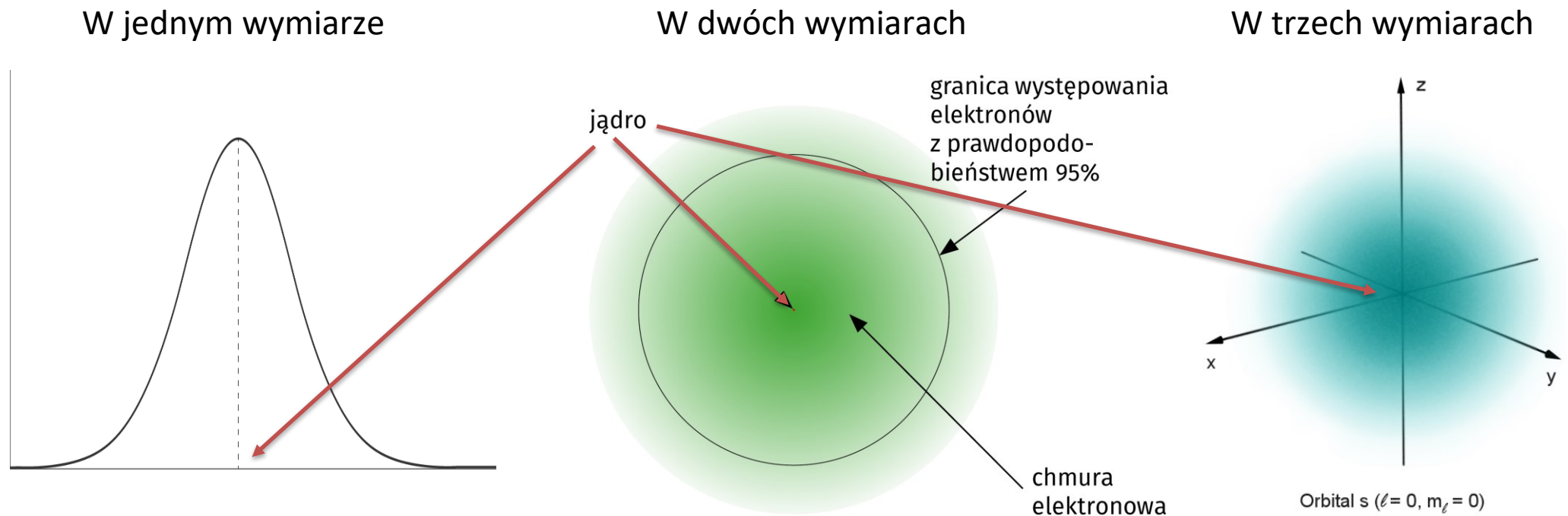
→ „Yep”... i jakoby zarazem „Not yep” 😊😊

- Parametry – w swej epistemologicznej istocie – pozostają *nieznanymi (i niepoznawalnymi, nieobserwowalnymi) stałymi*
- ALE na poziomie *modelu statystycznego* (= jedynego „narzędzia obserwacji” parametrów, *wnioskowania* o nich) parametry faktycznie są **zmiennymi losowymi**

Statystyka bayesowska

→ Najbardziej adekwatną metodą **ujmowania niepewności o wielkościach nieobserwowalnych** okazują się właśnie rozkłady prawdopodobieństwa, co wymaga traktowania ich jako zmienne losowe

→ Wartość parametru w (bayesowskim) modelu statystycznym $\approx$ **położenie elektronu** w fizyce kwantowej



Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

→ **Rozkład *a priori*, $p(\theta)$**

Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

→ **Rozkład *a priori*, $p(\theta)$** – czy to faktycznie „pięta Achille...” znaczy się „pięta Bayesa”??? 😄

Statystyka bayesowska

➤ ALE wróćmy do wzoru Bayesa:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

→ **Rozkład *a priori*, $p(\theta)$** – czy to faktycznie „pięta Achille...” znaczy się „pięta Bayesa”??? 😄

➤ Statystyka bayesowska = w pełni probabilistyczny obraz niepewności co do parametrów

→ Aby jednak otrzymać $p(\theta|y)$ (tj. aby w ogóle móc skorzystać ze wzoru Bayesa) MUSIMY zadać jakiś rozkład *a priori*, $p(\theta)$, choćbyśmy nawet nie chcieli „zanieczyścić” przezeń tej informacji o parametrach, którą niosą ze sobą dane (a która „wchodzi” do modelu poprzez rozkład próbkowy, $p(y|\theta)$, czyli funkcję wiarygodności)

➤ Skoro MUSIMY tu (czyli na gruncie statystyki bayesowskiej) wprowadzić jakąś (nie)wiedzę o parametrach *a priori* do modelu statystycznego (czego nie ma na gruncie „klasycznym”), to można zadać pytanie: *Czy to dobrze, czy niedobrze?*

Statystyka bayesowska

➤ Różne aspekty **obecności rozkładu *a priori*** w modelowaniu bayesowskim:

- **Zalety/korzyści:**

- 1) Ta FUNDAMENTALNA: **umożliwia skorzystanie ze wzoru Bayesa**, w wyniku którego uzyskujemy cały rozkład prawdopodobieństwa parametrów (czyli $p(\theta)$ jako „cena” za tę możliwość)
- 2) Umożliwia wprowadzenie do modelu statystycznego **dodatkowej (poza samymi tylko danymi) informacji/wiedzy o parametrach** = „naturze” modelowanego zjawiska, w tym wiedzy eksperckiej czy też wynikającej ze znajomości podobnych badań

- **Wady/ograniczenia:**

- 1) Sam fakt **konieczności** wprowadzenia informacji *a priori* do modelu (a my może byśmy właśnie tego nie chcieli?)

→ ALE: nieinformacyjne rozkłady *a priori* (w przybliżeniu takowe), obiektywne rozkłady *a priori*

- 2) Rozkład *a priori* stwarza możliwość manipulacji („forsowania”) wyników *a posteriori* – poprzez odpowiednie zacieśnienie rozkładu *a priori* nie pozwalamy danym „przemówić” zbyt swobodnie

→ ALE:

- taki **dogmatyzm** jest brakiem uczciwości badawczej :/
- możemy przeprowadzić **analizę wrażliwości** rozkładu *a posteriori* na różną informację *a priori*
- wpływ rozkładu *a priori* na *a posteriori* **maleje wraz ze wzrostem liczebności próby**

Statystyka bayesowska

➤ Wróćmy do naszego przykładu z rzutami monetą – tym razem **po bayesowsku**:

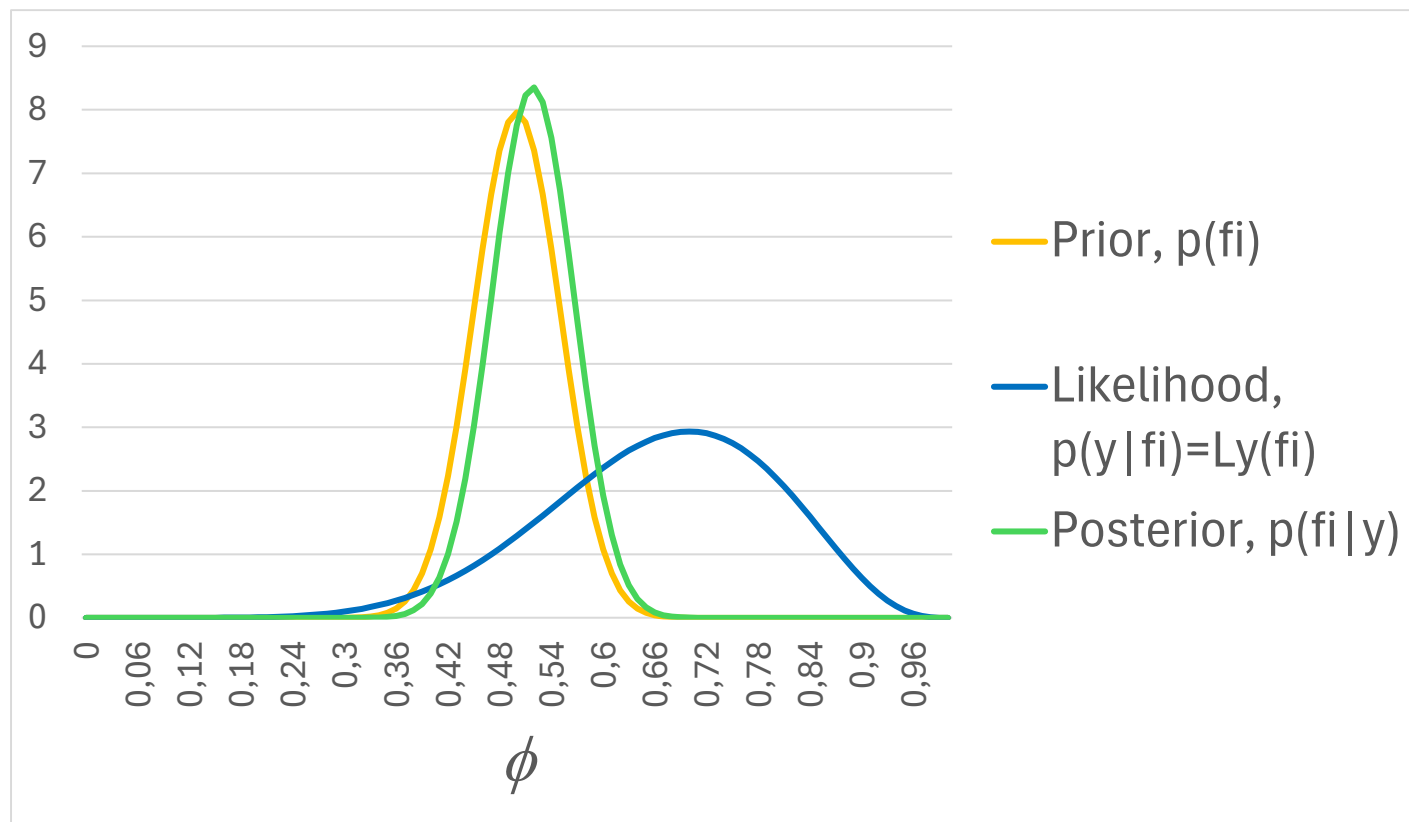
→ Zadajemy **rozkład próbkowy**, $p(y|\phi)$ (= funkcja wiarygodności) – ten sam, co w podejściu „klasycznym”

→ Dobieramy **rozkład a priori**, $p(\phi)$ – możemy **zadać** go na różne sposoby, byle by był określony na stosownym nośniku, tu – przedział $[0, 1]$

⇒ Otrzymamy różne **rozkłady a posteriori**, $p(\phi|y)$ (pomijam tu wszelkie obliczenia/wyprowadzenia)

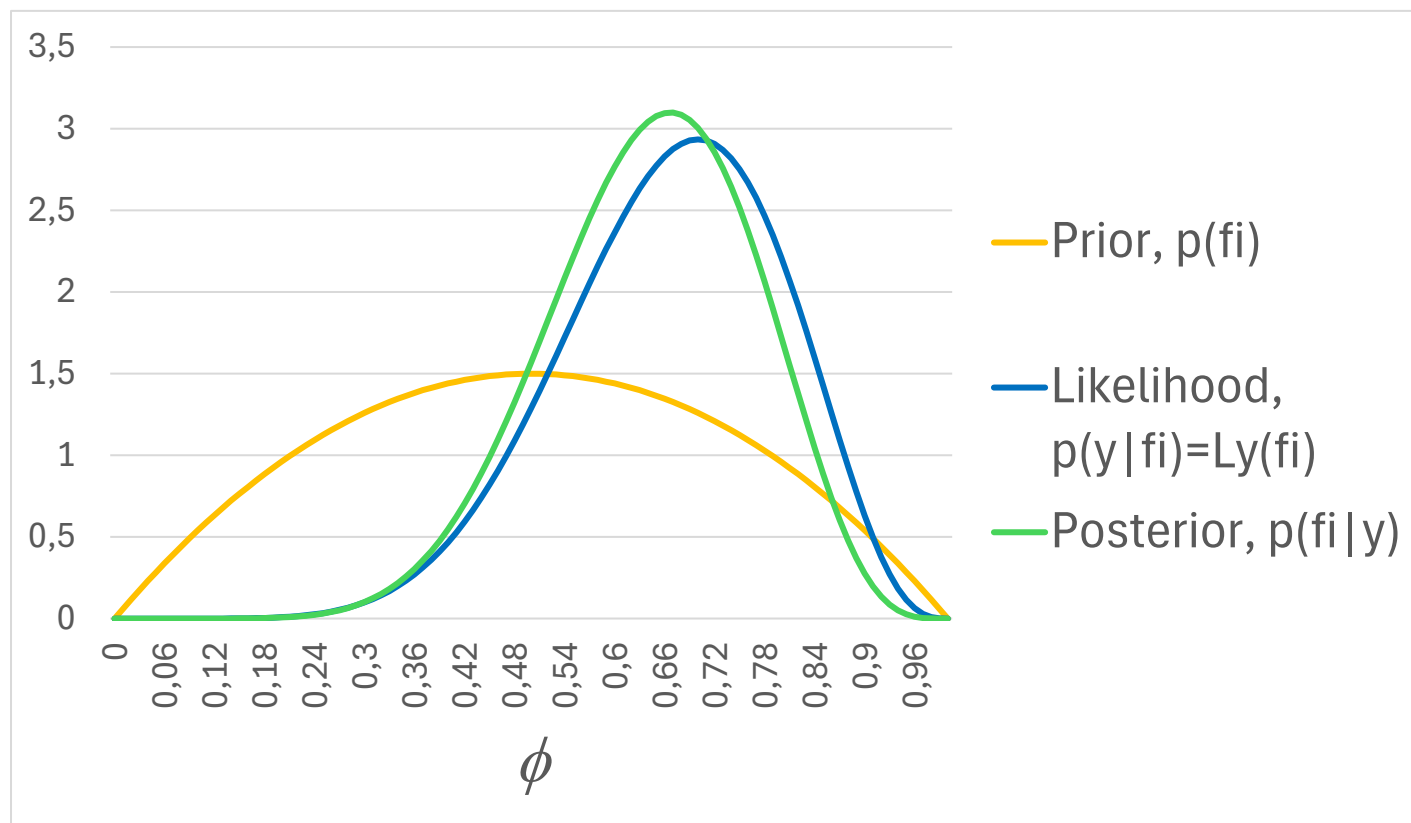
Statystyka bayesowska

- Dogmatyczny rozkład *a priori* – silnie preferujemy wartości $\phi \approx 0,5$:



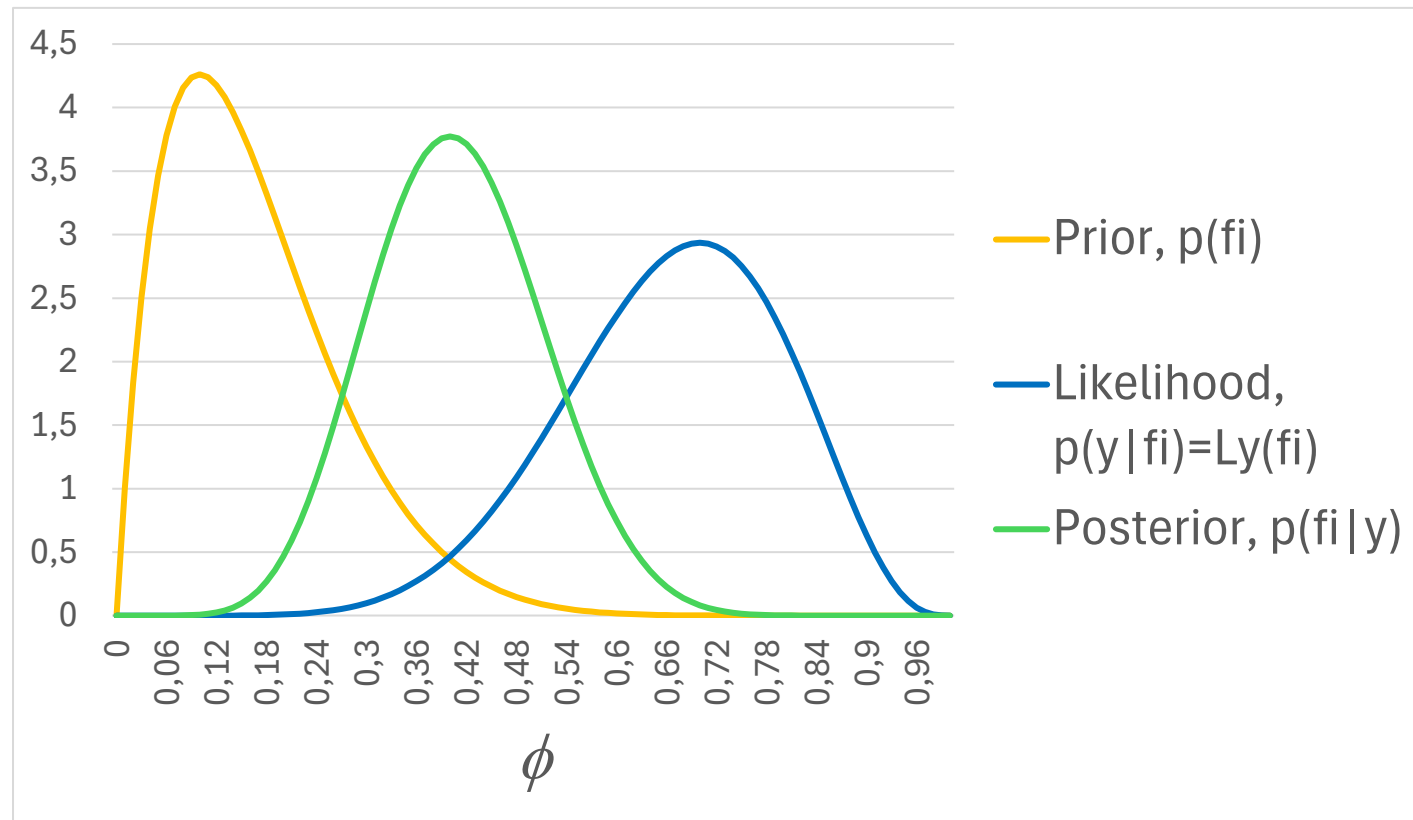
Statystyka bayesowska

➤ *A priori* tylko luźno preferujemy wartości $\phi \approx 0,5$:



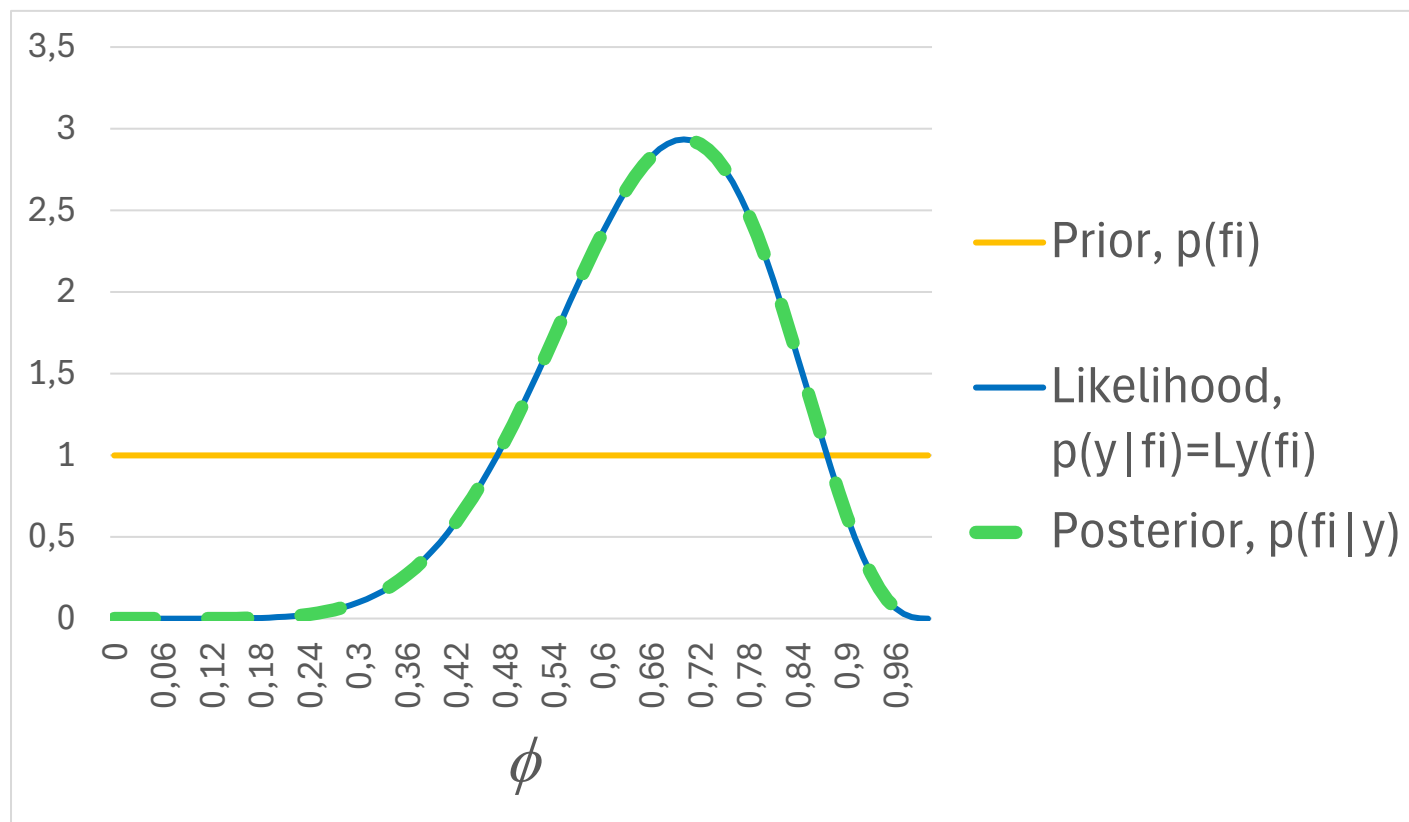
Statystyka bayesowska

➤ Rozkład *a priori* „stojący w poprzek danych”



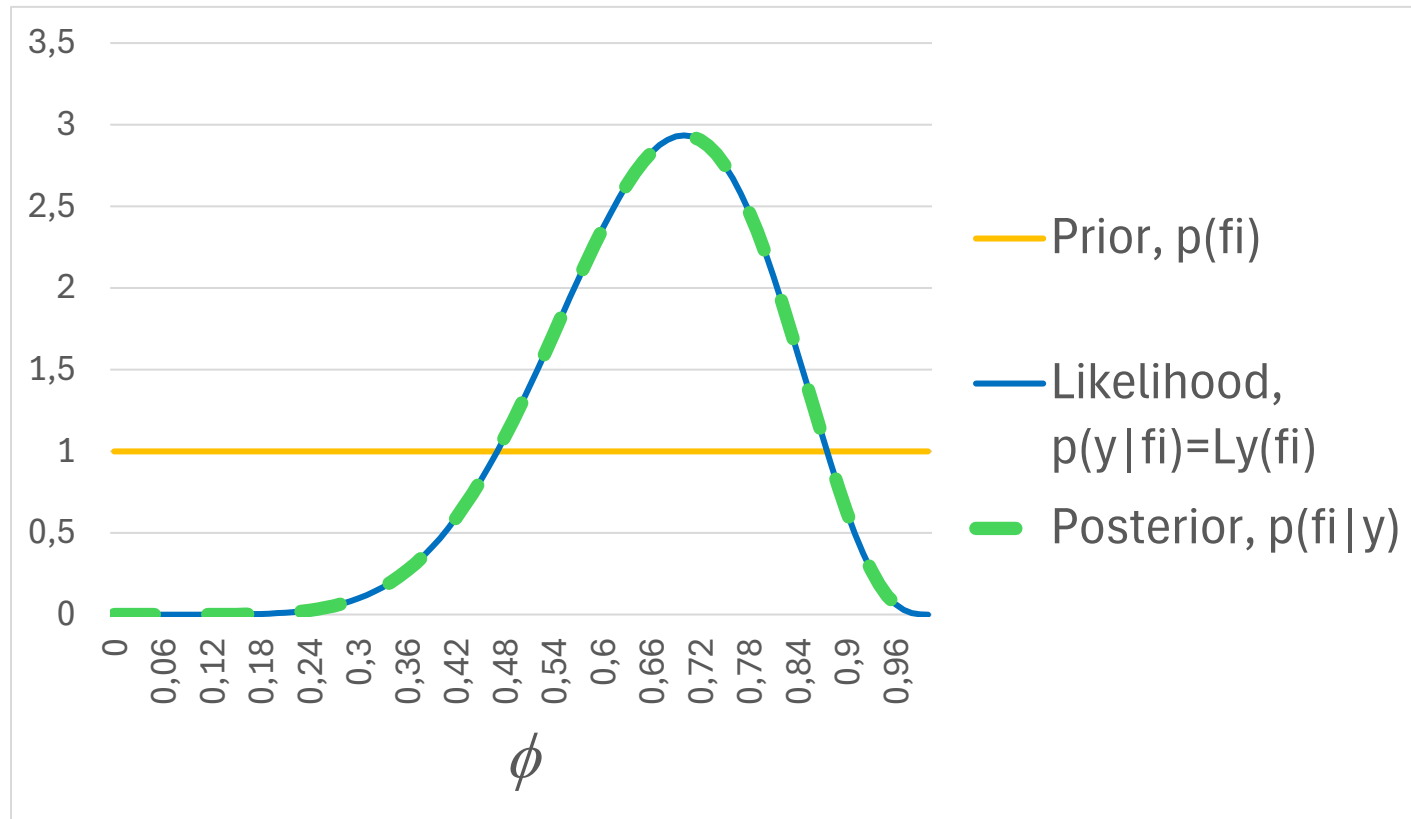
Statystyka bayesowska

- Płaski rozkład *a priori* = „nieinformacyjny” rozkład *a priori* (przynajmniej „na chłopski rozum”)



Statystyka bayesowska

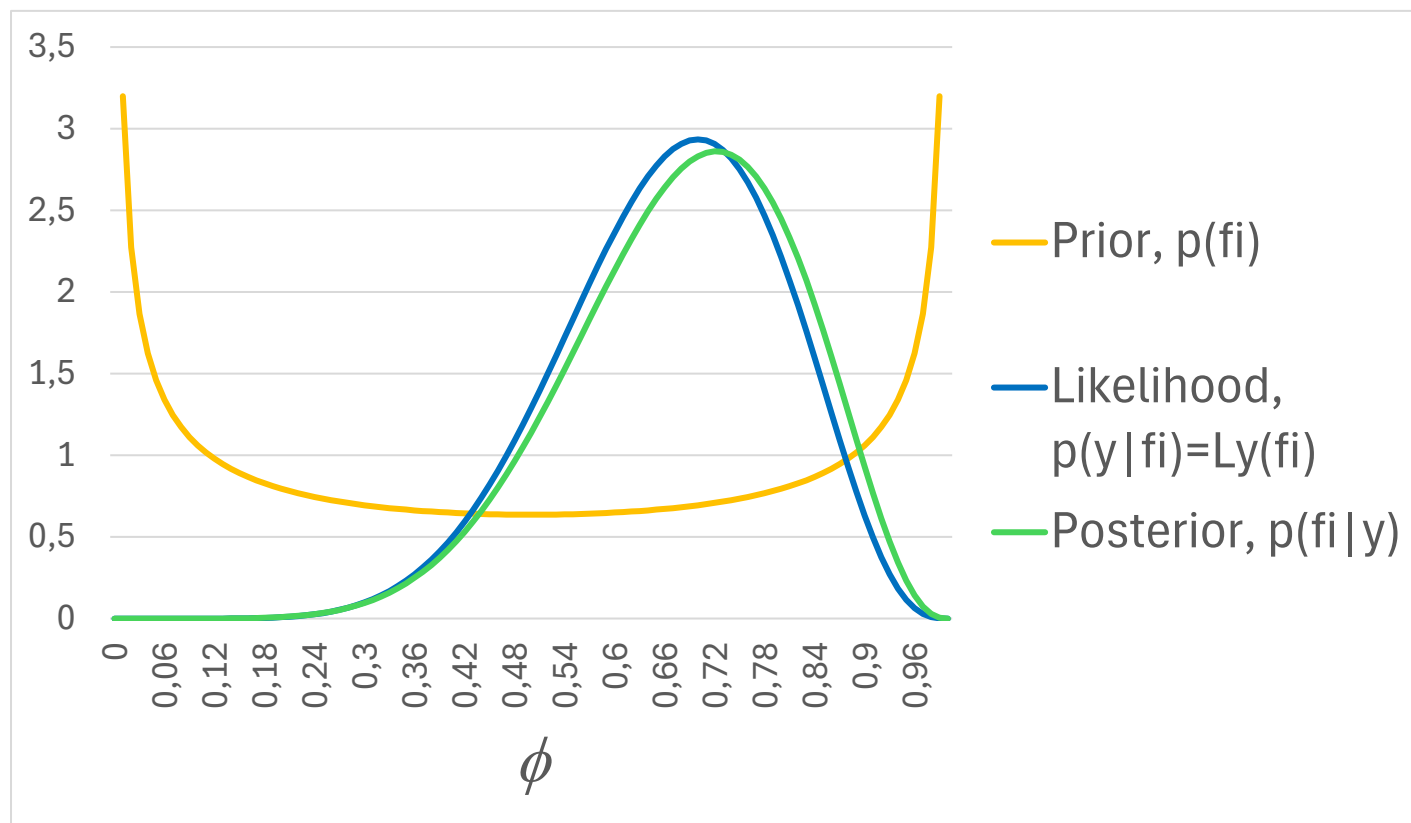
➤ Płaski rozkład *a priori* = „nieinformacyjny” rozkład *a priori* (przynajmniej „na chłopski rozum”)



→ Czasem pewne wyniki „klasyczne” dają się zreinterpretować na gruncie bayesowskim

Statystyka bayesowska

- Nieinformacyjny rozkład *a priori* (wyznaczony formalnie, wg reguły Jeffreysa)



Zakończenie: wnioskowanie bayesowskie – to „lepiej” ;)

➤ Statystyka bayesowska:

- a. Daje **pełny obraz niepewności** co do możliwych wartości parametrów (BEZ estymatorów, BEZ asymptotyki, a za to w PEŁNI PROBABILISTYCZNY)
 - a o to właśnie chodzi w statystyce (= nauka o **niepewności**, a nie – by tak rzec – o ocenach punktowych)
- b. **Wewnętrznie spójny paradygmat** (najbardziej ze wszystkich podejść do wnioskowania statystycznego, choć w dalszym ciągu nie jest wolne od pewnych paradoksów)

Zakończenie: wnioskowanie bayesowskie – to „lepsze” ;)

➤ Statystyka bayesowska:

a. Daje **pełny obraz niepewności** co do możliwych wartości parametrów (BEZ estymatorów, BEZ asymptotyki, a za to w PEŁNI PROBABILISTYCZNY)

→ a o to właśnie chodzi w statystyce (= nauka o **niepewności**, a nie – by tak rzec – o ocenach punktowych)

b. **Wewnętrznie spójny paradygmat** (najbardziej ze wszystkich podejść do wnioskowania statystycznego, choć w dalszym ciągu nie jest wolne od pewnych paradoksów)

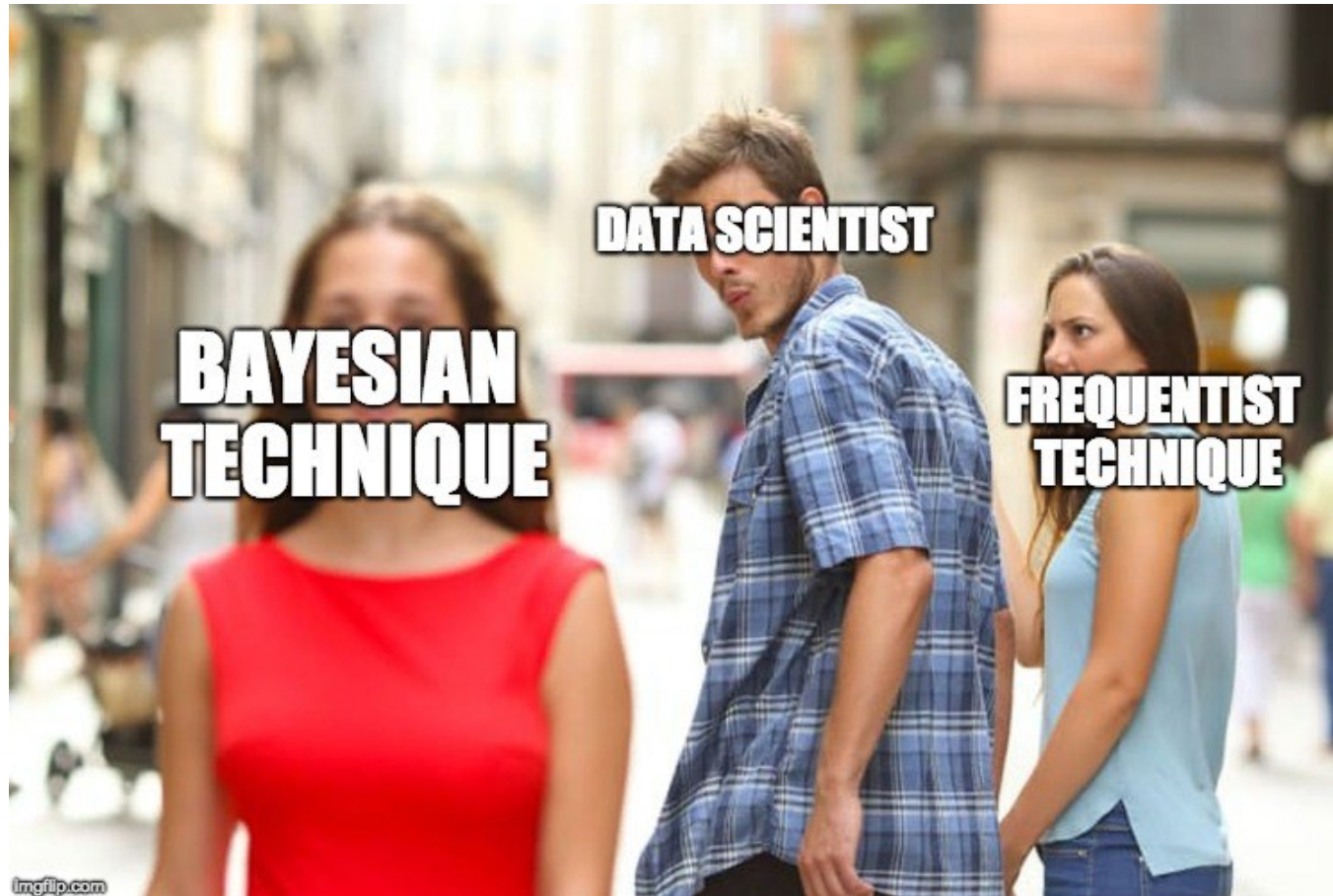


➤ Choć „lepsze”, to jednak jego zwolennikom/twórcom/propagatorom nie szczędzono batów na przestrzeni dziejów. Te przeplatająco-ścierające się dzieje statystyki bayesowskiej i „klasycznej” to istna „*Moda na sukces*”...

→ Sharon B. McGrayne, (2011), *The Theory That Would Not Die: ...*

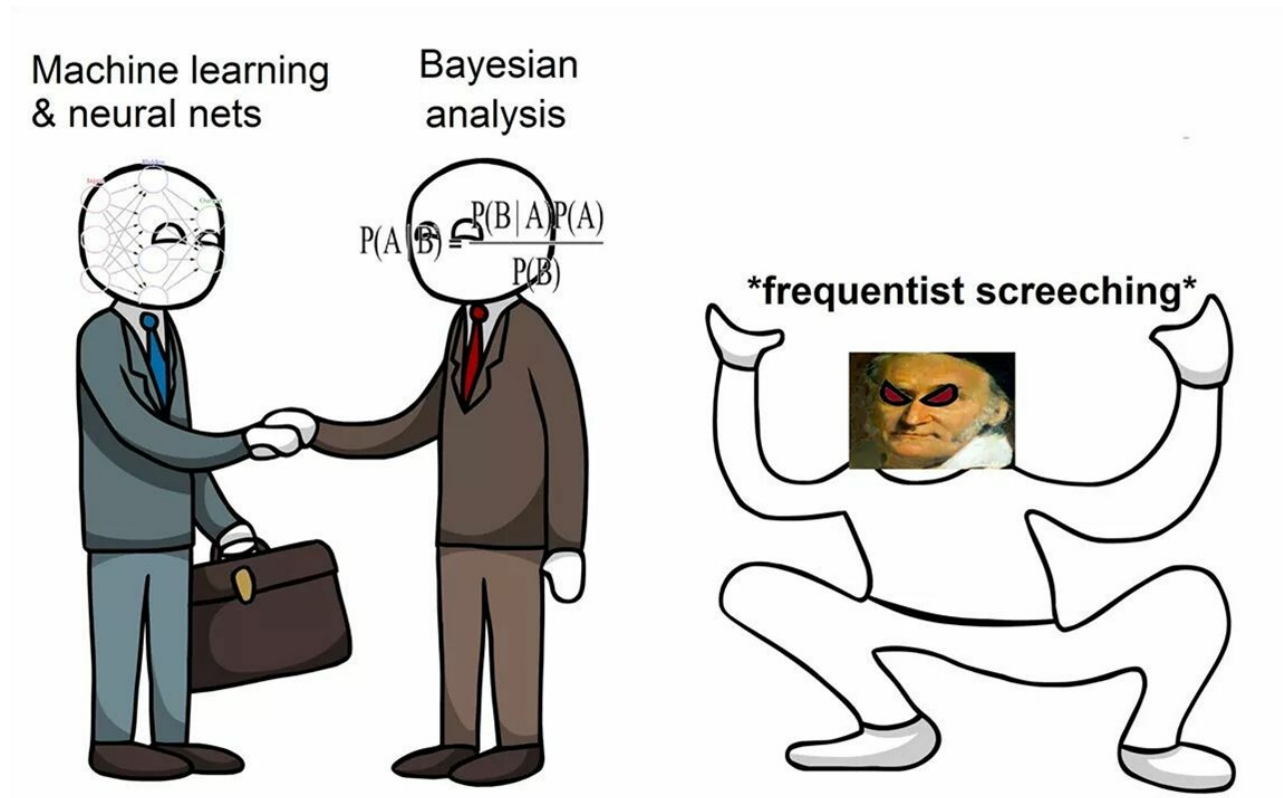
Zakończenie: wnioskowanie bayesowskie – to „lepiej” ;)

➤ Więcej niż tysiąc słów...



Zakończenie: wnioskowanie bayesowskie – to „lepiej” ;)

➤ *Więcej niż tysiąc słów...*



➔ **Bayesowskie sieci neuronowe!**

Dziękuję za uwagę! :)

Uwagi:

- 1) Bayesowskie testowanie hipotez
- 2) Aktualizacja wiedzy: yesterday's posteriori = today's prior
- 3) Zakończenie: przyszliście tu dzisiaj z Waszym dotychczasowym wyobrażeniem o p-stwie i statystyce (Wasze a priori) + dr Kwiatkowski (jako likelihood/data/evidence) = Wasza stopniowe przejście na Bayesa (a posteriori) :D