

Deň otvorených dverí

Štatistické paradoxy okolo nás

Lukáš Lafférs

  UMBmath
www.umbmath.sk

Február 13, 2024

Naša intuícia \neq Pravda

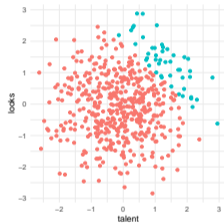
Skríningové Testovanie



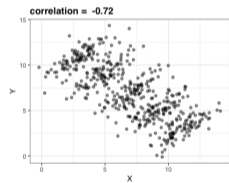
Narodeniny



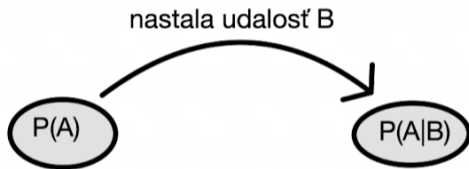
Collider Bias

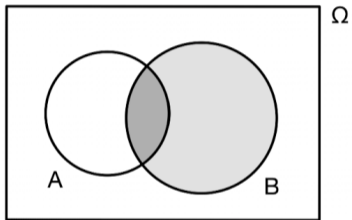


Simpsonov paradox



Skríningové Testovanie





$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{shaded intersection}}{\text{shaded circle B}}$$

- \oplus - pozitívny test
- \ominus - negatívny test
- ☺ - človek je zdravý
- ☹ - človek je chorý
- $P(\text{☹}) = 0.01$
- $P(\oplus|\text{☹}) = 0.9$ (senzitivita)
- $P(\ominus|\text{☺}) = 0.99$ (špecificita)

Náhodne vybraného človeka budeme testovať, test vyjde pozitívny. Aká je pravdepodobnosť, že tento človek je skutočne pozitívny?

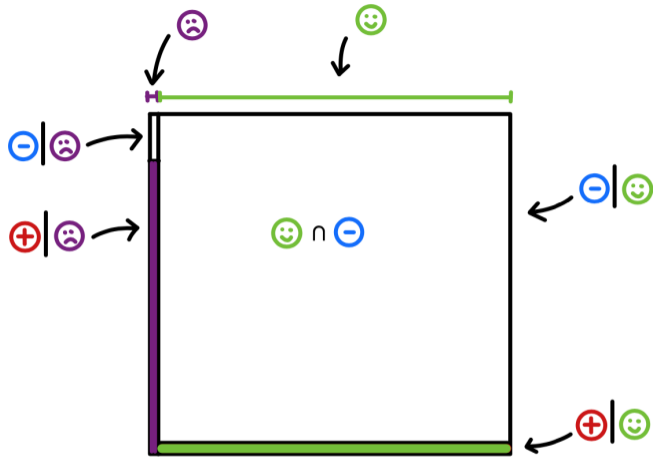
???



$$\begin{aligned} P(\ominus|\oplus) &= \frac{P(\ominus \cap \oplus)}{P(\oplus)} = \frac{P(\oplus|\ominus)P(\ominus)}{P(\oplus|\ominus)P(\ominus) + P(\oplus|\odot)P(\odot)} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.01 \cdot 0.99} = 0.476. \end{aligned}$$

$$\begin{aligned} P(\ominus|\oplus) &= \frac{P(\ominus \cap \oplus)}{P(\oplus)} = \frac{P(\oplus|\ominus)P(\ominus)}{P(\oplus|\ominus)P(\ominus) + P(\oplus|\odot)P(\odot)} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.01 \cdot 0.99} = 0.476. \end{aligned}$$

48% ??



$$P(\text{frowny}|\text{red}) = \frac{P(\text{frowny}n\text{red})}{P(\text{red})} = \frac{P(\text{frowny}n\text{red})}{P(\text{frowny}n\text{red}) + P(\text{smiley}n\text{red})} = \frac{\text{purple bar}}{\text{purple bar} + \text{green bar}}$$

Narodeninový paradox

Na predpandemickej párty (offline) sa stretne k ľudí.

Aké je najmenšie k také, aby bola šanca, že aspoň jedna dvojica má narodeniny v ten istý deň väčšia ako 50%?

Kol'ko ???



1 – $P(\text{každý má narodeniny v iný deň})$

$$\frac{\# \text{ vhodné možnosti}}{\# \text{ všetky možnosti}}$$

$1 - P(\text{každý má narozeniny v iný deň})$

$$\frac{\# \text{ vhodné možnosti}}{\# \text{ všetky možnosti}}$$

$$\# \text{ vhodné možnosti} = \underbrace{365 \cdot 364 \cdot \dots \cdot (365 - k + 1)}_k = \frac{365!}{(365 - k)!}$$

$1 - P(\text{každý má narozeniny v iný deň})$

$$\frac{\# \text{ vhodné možnosti}}{\# \text{ všetky možnosti}}$$

$$\# \text{ vhodné možnosti} = \underbrace{365 \cdot 364 \cdot \dots \cdot (365 - k + 1)}_k = \frac{365!}{(365 - k)!}$$

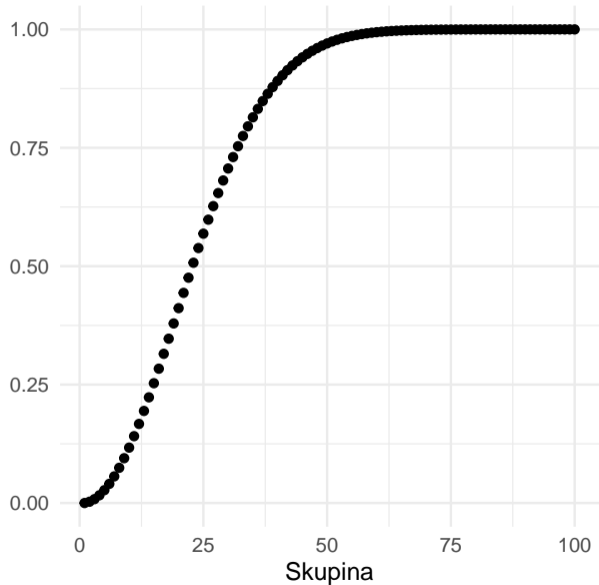
$$\# \text{ všetky možnosti} = \underbrace{365 \cdot 365 \cdot \dots \cdot 365}_k = 365^k$$

$$1 - \frac{\# \text{ vhodné možnosti}}{\# \text{ všechny možnosti}} = 1 - \frac{365!}{(365-23)! 365^{23}} = 0.5073.$$

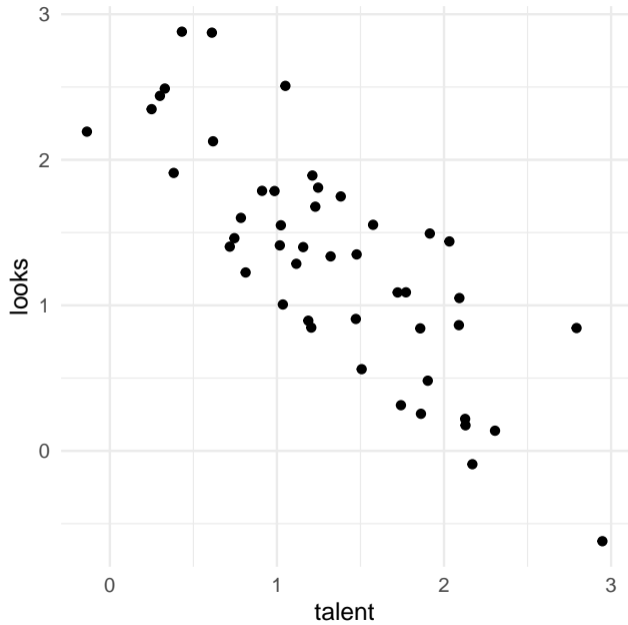
$$1 - \frac{\# \text{ vhodné možnosti}}{\# \text{ všechny možnosti}} = 1 - \frac{365!}{(365-23)! 365^{23}} = 0.5073.$$

23 lidí ??

P(existuje pár s rovnakými narodeninami)

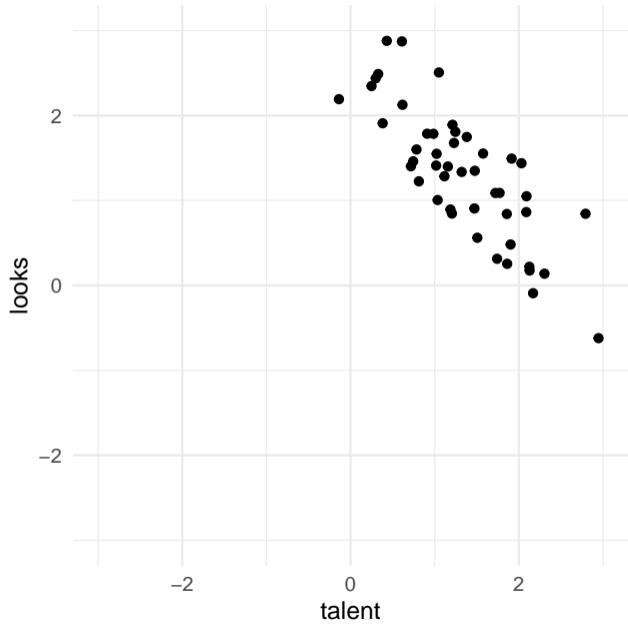


Collider Bias



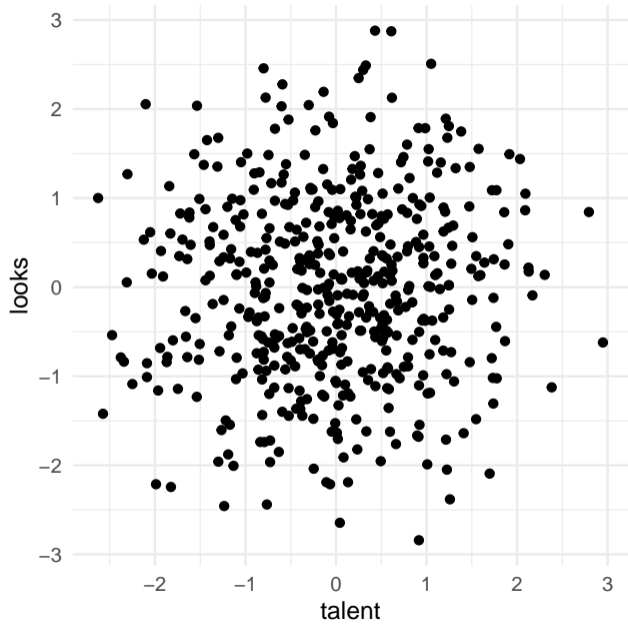
Prečo ???

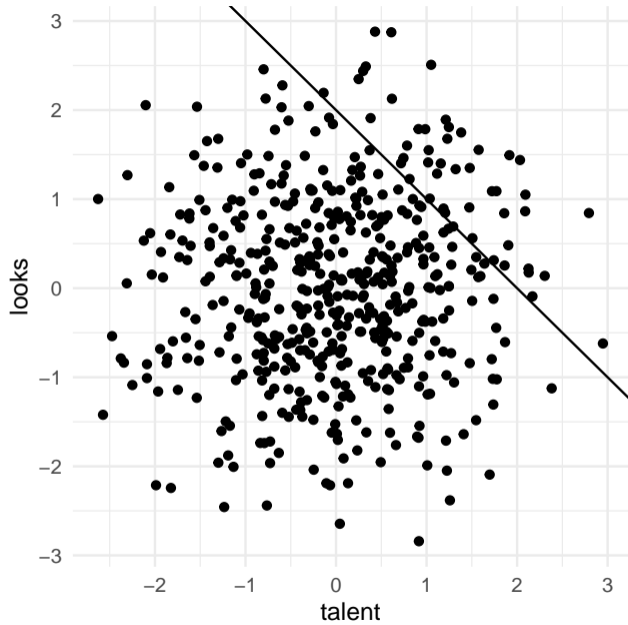


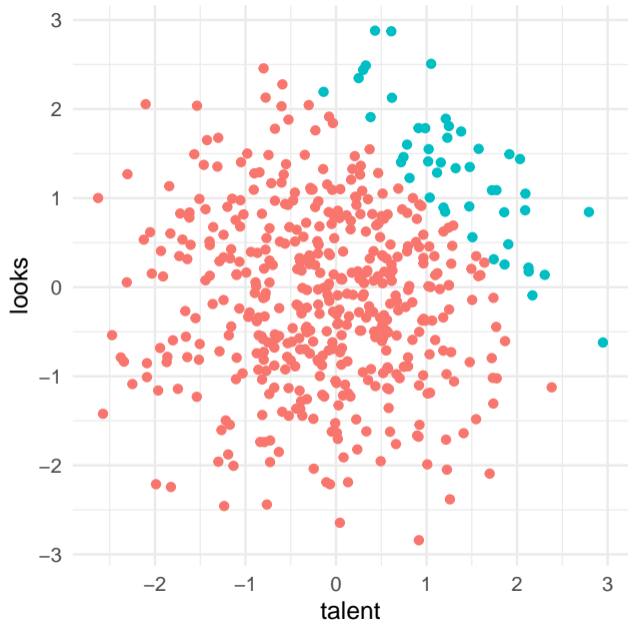


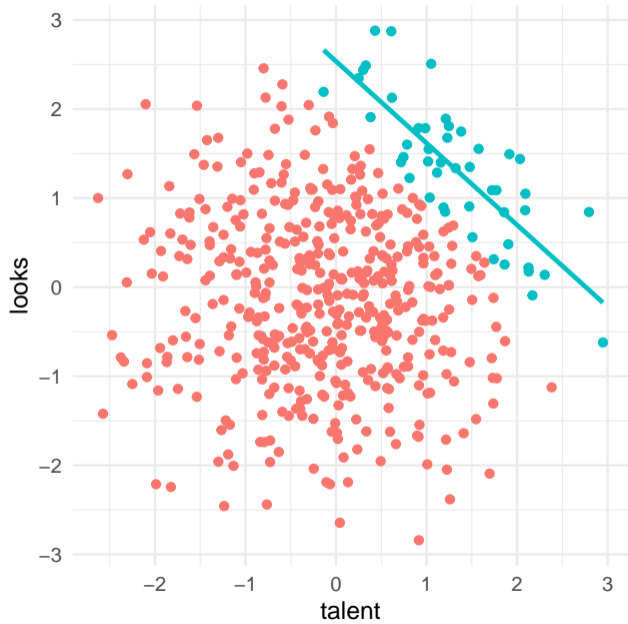
Pomohlo ???

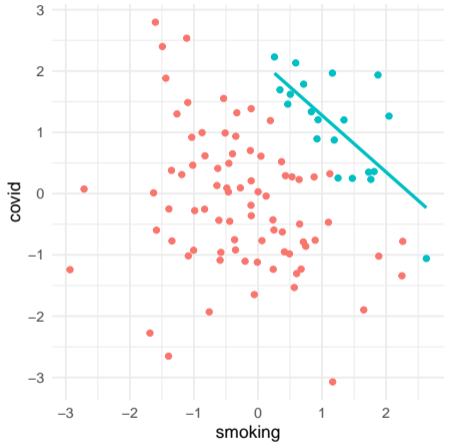
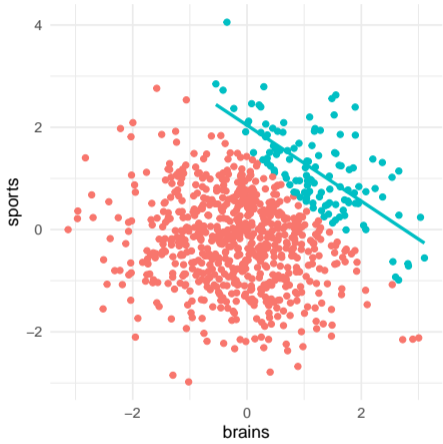












1973 UC Berkeley

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

Zdroj:

<https://en.wikipedia.org/wiki/Simpson27sparadox>

Diskriminácia ???



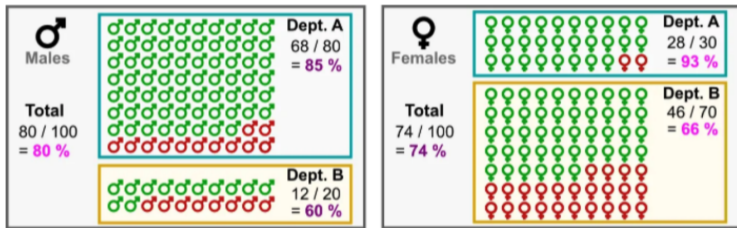
1973 UC Berkeley

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%

Zdroj:

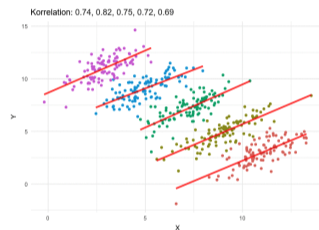
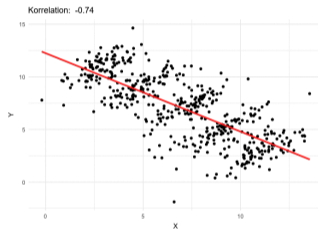
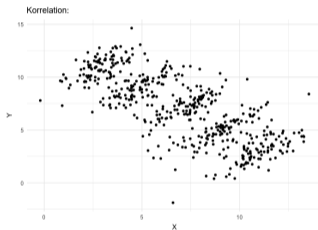
<https://en.wikipedia.org/wiki/Simpson27sparadox>

Zjednodušené vysvetlenie



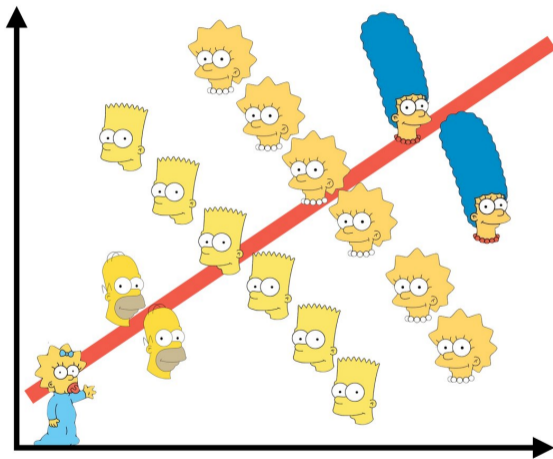
Zdroj: [<https://towardsdatascience.com/top-3-statistical-paradoxes-in-data-science-e2dc37535d99>]

Simpson's Simpson's paradox



Zdroj: <https://en.wikipedia.org/wiki/Simpson27sparadox>

Simpson's Simpson's paradox



Zdroj: <https://twitter.com/infowetrust/status/984536880199876608>

Fajčenie a úmrtnosť.

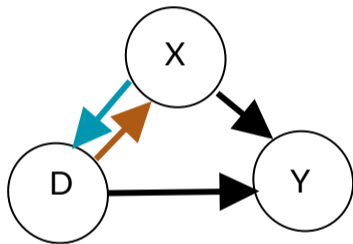


Vakcinácia a úmrtnosť.



Bonus

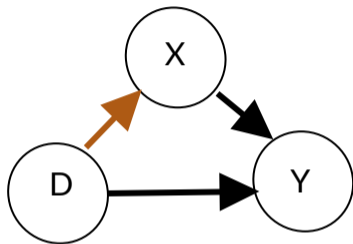
Rôzne závery z tých istých dát



- X - manažérska pozícia
- D - **pohlavie** alebo **životný štýl**
- Y - mzda

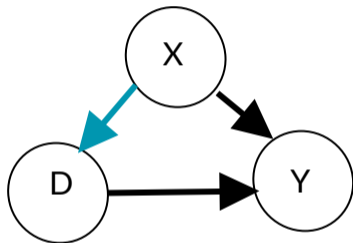
Kauzálna štruktúra je dôležitá.

Paul Hunermund's course: <https://www.udemy.com/course/causal-data-science/>



- X - manažérska pozícia
- D - **pohlavie**
- Y - mzda

Paul Hunermund's course: <https://www.udemy.com/course/causal-data-science/>



- X - manažérska pozícia
- D - životný štýl
- Y - mzda

Paul Hunermund's course: <https://www.udemy.com/course/causal-data-science/>

	♀	♂
Not manager	3163 (87)	3015 (59)
Manager	5592 (13)	5319 (41)

- X - manažérska pozícia
- D - pohlavie
- Y - mzda

$$E[Y(\text{♀}) - Y(\text{♂})] = \dots = 3478.77 - 3959.64 = -480.87$$

Paul Hunermund's course: <https://www.udemy.com/course/causal-data-science/>

	😊	☹️
Not manager	3163 (87)	3015 (59)
Manager	5592 (13)	5319 (41)

- X - manažérska pozícia
- D - životný štýl
- Y - mzda

$$E[Y(\text{😊}) - Y(\text{☹️})] = \dots = 3818.83 - 3637.08 = 181.75$$

Paul Hunermund's course: <https://www.udemy.com/course/causal-data-science/>

Nejaké detaily...

$$E[Y(d)] = E[Y|D = d] = \sum_{x \in \{0,1\}} E[Y|D = d, X = x]Pr(X = x|D = d)$$

$$E[Y(\text{♀})] = \sum_{x \in \{0,1\}} E[Y|D = \text{♀}, X = x]Pr(X = x|D = \text{♀}) = 3163 \cdot 0.87 + 5592 \cdot 0.13 = 3478.77$$

$$E[Y(\text{♂})] = \sum_{x \in \{0,1\}} E[Y|D = \text{♂}, X = x]Pr(X = x|D = \text{♂}) = 3015 \cdot 0.59 + 5319 \cdot 0.41 = 3959.64$$

$$E[Y(d)] = E[E[Y|D = d, X]] = \sum_{x \in \{0,1\}} E[Y|D = d, X = x]Pr(X = x)$$

$$E[Y(\text{☺})] = \sum_{x \in \{0,1\}} E[Y|D = \text{☺}, X = x]Pr(X = x) = 3163 \cdot \frac{87 + 59}{200} + 5592 \cdot \frac{13 + 41}{200} = 3818.83$$

$$E[Y(\text{☹})] = \sum_{x \in \{0,1\}} E[Y|D = \text{☹}, X = x]Pr(X = x) = 3015 \cdot \frac{87 + 59}{200} + 5319 \cdot \frac{13 + 41}{200} = 3637.08$$

Skríningové Testovanie



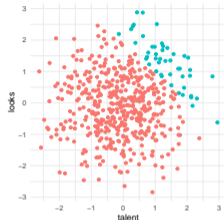
Zdravých ľudí je
veľa.

Narodeniny



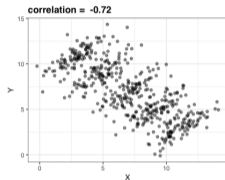
Rôznych dvojíc je
veľa.

Collider Bias



Holyvúdske ★-y
sú iné ako my.

Simpsonov paradox



Ak niečo platí pre
každú podskupinu,
neznamená to, že
to platí aj celkovo.

Bonus: Na kontexte záleží.



Otázky?

- lukas.laffers@umb.sk
- Zoom/Teams
- osobne [FPV UMB], F226A

Táto prezentácia bude na [<https://www.lukaslaffers.com/misc>].

štúdijný program: **Matematika (Bc)**

štúdijný program: **Matematika v analýze dát a vo financiách (Mgr)**

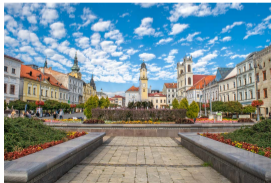
Prečo matematika?

- → **uplatnenie**
- analytické myslenie
- precíznosť
- jazykom prír. vied
- zábava



Prečo na UMB?

- → **moderný program**
- práca so softvérom
- špičková veda
- rodinné prostredie
- Banská Bystrica



Uplatnenie

- data science
- škola/veda
- fin. matematika
- štatistika
- programovanie



Čo sa naučíte?

- rozmanité oblasti matematiky
- pracovať analyticky
- dátovú analýzu
- základy programovania



Katedra matematiky

- v akreditácii hodnotenie A
- špičkový vedecký tím
- spolupráca so zahr. inštitúciami (Oxford, Londýn, Sydney)

Naši absolventi

- Národná Banka Slovenska
- analytické útvary (ČSOB, VÚB)
- poisťovne
- softvérové firmy
- technologické firmy (Dell, IBM)
- vládne think-tanky (Inštitút sociálnej politiky)

Ďakujem za pozornosť

- www.umbmath.sk
- lukas.laffers@umb.sk
-   umbmath

Ďakujem za pozornosť!



KM

KATEDRA MATEMATIKY
UNIVERZITA MATEJA BELA



UMBmath



umbmath