Mýty, omyly, finty a vysvětlení

ve statistice

Martin Sebera

Brno 2020

Obsah

Úvod

Ve statistice, kterou můžeme považovat za velmi přesnou a matematikou nabitou disciplínu, velmi často záleží na detailech. Statistika umí rozhodně počítat velmi přesně, záleží jen na vstupech a na výstupech. Proces vlastního počítání je velmi jasný. Takže se nabízí několik možností: mít přesné výpočty z nepřesných dat nebo nepřesné výpočty z přesných dat? Dostáváme se tak do prostředí filozofických otázek, kde lze hledat odpověď např. zamítnout hypotézu nebo nezamítnout? Přijmout nebo nepřijmout? Je v tom rozdíl? Ano je…

Stejně tak na výstupech, pokud máme výsledek statistické procedury, lze výsledkům udělit hned několik intepretací. A která je správná? Která je méně správná? Která je nesprávná? I za tímto jemným dělení výpovědí lze vysledovat rozdíly.

Tento text jen namátkou chce ukázat některé z těchto rozcestí, na kterém se každý výzkumník zabývající se statistikou někdy ocitne. Asi text nebude pokrývat všechny otázky, snad pomůže ukázat na nejzákladnějších postupech, jaká zákoutí můžeme se statistikou navštívit, a alespoň načrtne správný směr, kudy projít a neudělat velký počet chyb. *Poznámka: tedy ne „mít to správně“, ale mít to „co nejméně nesprávně“* ☺.

1. Jakou střední hodnotu použít, průměr nebo medián?

Rok 2015

- Ve 3. čtvrtletí 2015 byla průměrná měsíční mzda v České republice
  - 26 072,- Kč (vypočítáno pomocí aritmetického průměru) a
  - 22 531,- Kč (vypočítáno pomocí mediánu).
- Zdroj.: *Český statistický úřad*, https://www.czso.cz/csu/czso/cri/prumerne-mzdy-3-ctvrtleti-2015
- Rozdíl 3 541,- Kč představuje 13,6 % rozdíl.

Rok 2020

- Ve 3. čtvrtletí 2020 byla průměrná měsíční mzda v České republice
  - 35 402,- Kč (vypočítáno pomocí aritmetického průměru) a
  - 31 183,- Kč (vypočítáno pomocí mediánu)
- Zdroj.: *Český statistický úřad*, https://www.czso.cz/csu/czso/cri/prumerne-mzdy-3-ctvrtleti-2020
- Rozdíl 4 219,- Kč představuje 11,9 % rozdíl

Závěr:

- Medián představuje střední hodnotu, která není ovlivněna extrémními hodnotami (ať už maximy nebo minimy).
- Použití obou hodnot je správné, avšak pouze za předpokladu, kdy uvedete, kterou střední hodnotu používáte…

2. Když už uvádíte aritmetický průměr, vždy se směrodatnou odchylkou a N

Příklad

Data	Průměr	směrodatná odchylka	N
1; 10; 22	11	10,53	n = 3
11; 11; 11	11	0	n = 3

Na tomto příkladu je zřejmé, jak aritmetický průměr, pokud by byl uveden samostatně, nevyjadřuje přesné informace o původních datech. Navíc pro mnohé další výpočty, např. t-test, potřebujeme znát celou svatou trojici základních statistických charakteristik.

Závěr:

- Pokud se rozhodnete uvést ve své práci hodnotu aritmetického průměru, vždy k ní přidejte hodnotu směrodatné odchylky a počet měření N.
- Uvedením samotného aritmetického průměru ztrácíte velké množství informací o datech

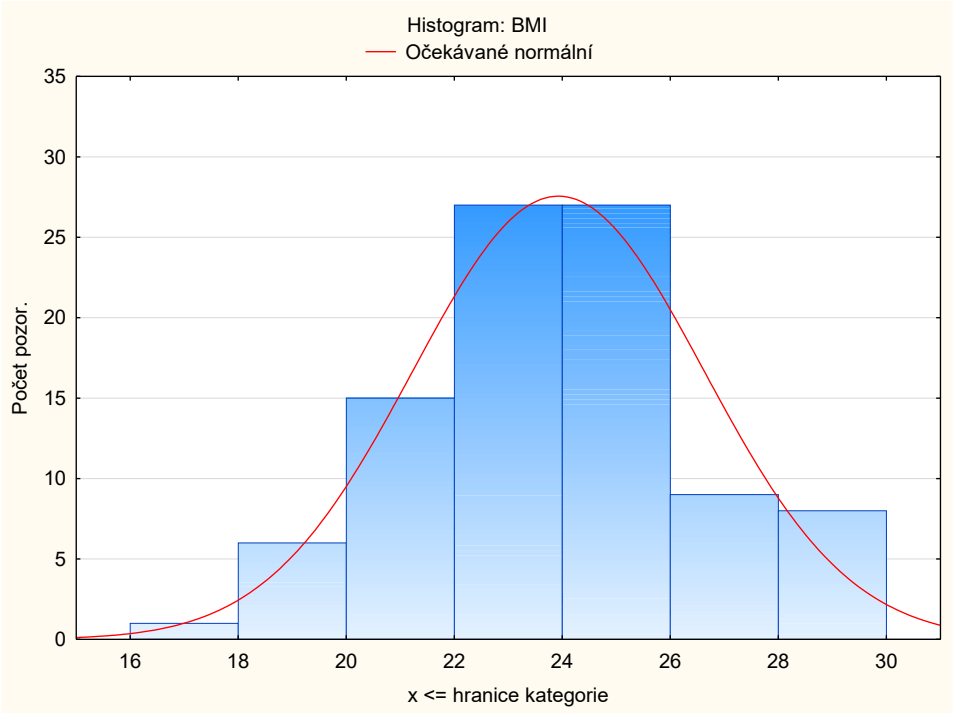
3. Rozložení četností a histogram

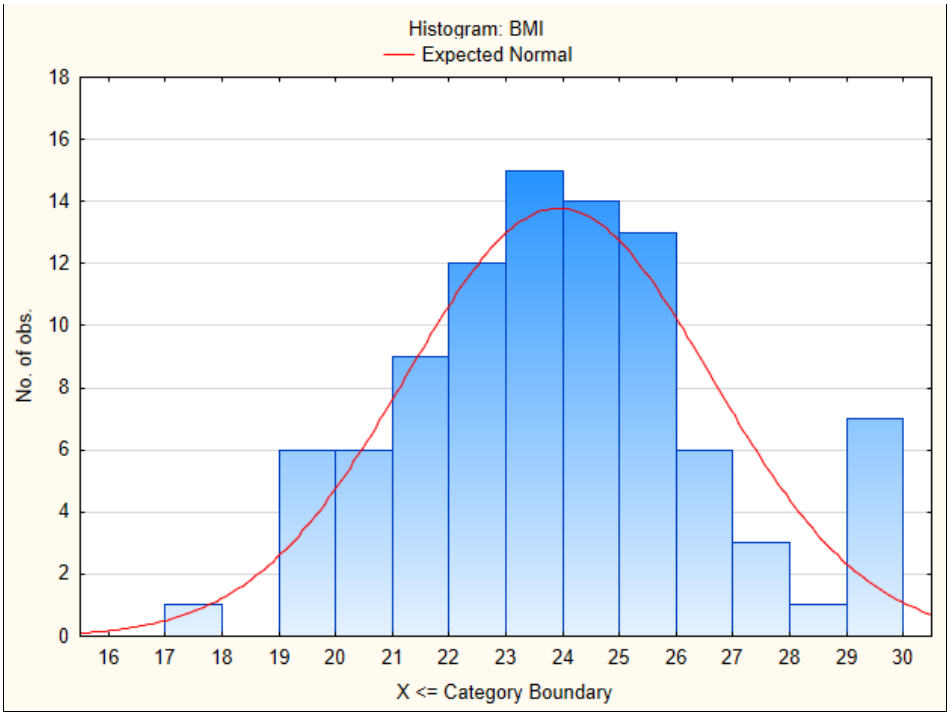
Problém: záleží, kolik bude intervalů a jak budou široké

Příklad: Máme 93 hodnot z měření BMI (body mass index)

17,9 19,2 19,3 19,6 19,6 19,7 19,8 20,1 20,3 20,3 20,4 20,9 20,9 21,1 21,1 21,1 21,4 21,6 21,6 21,6 21,8 21,9 22,1 22,2 22,2 22,3 22,3 22,4 22,6 22,7 22,8 22,8 22,9 23,0 23,1 23,1 23,2 23,3 23,3 23,4 23,4 23,4 23,6 23,7 23,8 23,9 23,9 23,9 24,0 24,1 24,1 24,1 24,3 24,4 24,4 24,5 24,5 24,5 24,7 24,8 24,9 24,9 25,0 25,1 25,1 25,1 25,1 25,2 25,3 25,3 25,4 25,5 25,6 25,7 25,8 25,9 26,3 26,3 26,5 26,8 26,9 26,9 27,1 27,7 28,0 28,6 29,2 29,4 29,4 29,4 29,4 29,7 30,0

Sestavíme histogram neboli graf četností:





Závěr: rychlý pohled na výsledné histogramy neukazuje téměř žádné velké rozdíly, detailnější pohled pak detekuje odlišnosti v četnostech jednotlivých intervalů a též ve tvaru proložené křivky normálního rozdělení. Histogramy lze použít pro rychlou kontrolu, zda data pocházejí z normálního rozdělení. Vhodnější postup je testovat hypotézu o normalitě pomocí vybraných testů, např. Kolmogorov-Smirnovův nebo Shapiro-Wilksův test.

4. Testování hypotéz - Čarodějnický soudní proces

Inspirováno a použito z newsleetteru vydaného dne 10. 12. 2012 firmou Statsoft, distributora statistického software Statistica

Tento příběh se podle dostupných zdrojů odehrál již před mnoha lety v místě, myslím, zvaném Středozemě. Jeden z místních obyvatel, který se posléze stane obviněným v tomto příběhu, si všiml podobných znaků mezi synem a otcem, matkou a dcerou. A tak začal tyto znaky předpovídat přímo i u nově narozených dětí. Jeho (přesněji řečeno její) předpovědi se po narození s postupem času, tak jak miminka stárla, vyplňovaly. Po té, co se s odstupem času potvrdilo několik relativně úspěšných předpovědí, byla tato osoba, mladá dáma, obviněna z čarodějnictví. Zajímavé, že její kamarád, se kterým na tomto výzkumu také pracovala, obviněn nebyl. Soud středověkého charakteru, který je v danou dobu považován za jediný spravedlivý, zasedl.

Jde o jednoduchý soudní proces, ve kterém existuje pouze jediný možný trest (upálení). Soud rozhodne, jestli se tomuto trestu obžalovaná podrobí anebo nepodrobí. Na rozdíl od dnešních soudů zde neexistuje žádné odvolání. Obžalovaná je obviněna z čarodějnictví, což je velmi závažné obvinění. Mimochodem, kdyby byla například obviněna z toho, že je pohádková bytost, tak by to ještě šlo, neboť vyhoštění z vesnice do pohádkového lesa se dá přežít, protože v lese důsledkem soudních procesů mezitím vyrostlo vlastní, ve stromech skryté město, ve kterém žijí všichni za pohádkové bytosti prohlášení. Byl zde skoro každý člověk, co v něčem (kromě hrubé síly) výrazně vynikal.

No ale čarodějnictví, to je horší. „Můžou nastat dvě situace, jak to dopadne. Buď prokážeme, na základě důkazů z posledních let, že je to čarodějnice, anebo nám tyto důkazy stačit nebudou, ale čarodějnice to je tak i tak," sdělí žalobce den před procesem svému kolegovi. Pokud je test statistiky významný, pak žalobce prokáže svou tezi (alternativní hypotézu). Zamítnutí nulové hypotézy je ekvivalentní přijetí alternativy. Pokud však test nezamítne, žalobce zkoumanou hypotézu neprokáže, ale to neznamená, že neměl pravdu. Nemyslete si, pan žalobce si bude pořád myslet, že má pravdu, třeba jen neměl dostatek důkazů.

Jak už jsme si naznačili výše, mohou nastat dva případy špatného rozhodnutí:

I. Obžalovaná je nevinná, ale soud jí odsoudí a pošle na hranici – CHYBA I. DRUHU (α)

II. Obžalovaná je vinná, je to skutečně čarodějnice, ale důkazní materiál předložený žalobou na to nestačí a odsouzena a upálena na hranici nebude – CHYBA II. DRUHU (β).

Které z těchto pochybení je horší? Neupálit nebezpečnou pohádkovou bytost oplývající magickými schopnostmi nebo upálit nevinnou ženu, které nemá s kouzly nic společného? Odpověď si řekneme za chvíli, nyní se pojďme podívat, jak tento příběh dopadl.

Soudný den

Pan soudce, který tento proces soudí, je relativně liberální a příčí se mu, že by odsoudil nevinného člověka. Je známý tím, že potřebuje dostatečné množství důkazů na to, aby prokázal tvrzení žalobce, tedy rozhodl o vině. V této historické době nevídaná věc, neboť většina soudců je velmi konzervativních až posedlých a k odsouzení jim stačí i nepřímé a často hloupé důkazy. Je pravda, že skutečný viník jim nikdy neutekl, ale na hranicích a v žalářích končí velké množství populace úplně zbytečně. Potom se není čemu divit, že v pohádkovém lese vyrostlo již zmíněné město odsouzených, složené z převážně nadaných lidí.

Souzená má tedy štěstí, že bude soudit tento soudce (mezi kolegy pro svůj styl nazývaný „Měkouš“). V naší terminologii můžeme tohoto soudce přirovnat k testu s velmi nízkou chybou prvního druhu (chce mít jistotu, že neodsoudí nevinného). U tohoto soudce se ale může stát a často se to i stává, že obvinění zamítne, ve skutečnosti je však viník „zlosyn“. Žalobce prostě nesehnal dostatek důkazního materiálu k přesvědčení soudce. Soudcovi drsnější kolegové bychom poté přirovnali naopak k testům s velkou hodnotou α.

Obžalovaná povstaňte

Podle dochovaných záznamů se dochovala část přepisu tohoto dějství:

- Soudce: „Stojíte před vážným obviněním, jste obviněna z čarodějnictví. Prosím pana žalobce o předložení důkazů.“
- Žalobce: „Důkazy jsou zde,“ ukáže na syna s blonďatými vlasy a modrýma očima, které jsou stejné, jako má jeho otec.
- „Obžalovaná rodině sama řekla, že syn bude mít tyto rysy a přitom se ještě nenarodil. Není to náhoda, stejné je to u těchto dalších 4 dětí“.
- Soudce: „Tvrdíte, že předpověděla barvu očí a vlasů u všech těchto dětí, ještě před jejich narozením?“
- Žalobce: „Ano, a u těchto dvou dokonce jejich výšku,“ řekl žalobce a ukázal na enormně vysokého chlapce, jehož otec mezitím zemřel, ale patřil k nejvyšším dlouhánům ve vesnici.
- Soudce: „Hmm, tyto důkazy jsou celkem závažné, nechť promluví obžalovaná.“
- Obžalovaná: „ Pane soudce, všimla jsem si, že těmito znaky existuje souvislost, dědičnost, tedy, že

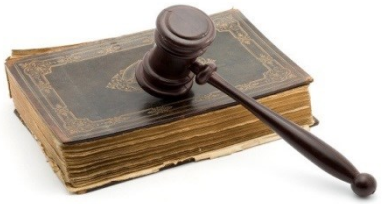
existuje závislost mezi výškou otce a výškou jeho syna.“ Matka obžalované po tomto výroku omdlela.

- Soudce: „V kolika případech se obžalovaná spletla při svých předpovědích?“
- Obhájce: „Celkem u zhruba dalších pěti dětí, kde předpověděla úplně jiné znaky, než má jejich otec. Je to tedy celé pouhá spekulace. Trváme na tom, že obžalovaná je člověk a vždy jím byla.“
- Soudce: „Aha, takže ono to zas s tou předpovědí není tak horké, s podobným úspěchem předpovídáme počasí, vezmu si čas na rozmyšlenou…“

Po chvíli soudce vyřkl rozsudek.

- Soudce: „Obžalovaná je zproštěna obvinění, tvrzení, že je čarodějnice, se neprokázalo. Stále platí, že je člověk jako každý z nás v této místnosti a tak s ní budeme i nakládat.“

Obžalovaná byla tedy pro nedostatek důkazů propuštěna, ale bylo to správně? Zachránilo jí, že její předpovědi měly přibližně 50% úspěšnost, ale nebylo to pouze nevhodným vzorkem dat? Tento vzorek byl velmi malý. Zasvěcení navíc věděli, že obžalovaná měla pravdu i v mnoha dalších případech, ale protože si opravdu mysleli, že je čarodějnice, tak ze strachu raději mlčeli a o skutečném původu některých dětí se raději nezmiňovali. V městečku tedy panovala podivná atmosféra a zhruba půlka stále věřila, že se o bytost s nadpřirozenými schopnostmi opravdu jedná. Proto někteří začínali pomalu shánět další důkazy pro nový soudní proces, tedy nový vzorek dat.



Shrnutí

Ať už se soudce mýlil nebo ne, v důkazním materiálu bylo příliš mnoho pozorování, která hovořila v neprospěch alternativy, kterou chtěla obžaloba prokázat.

V úvodu jsme vymezili dva případy, jak se může soudce zmýlit. Která z těchto pochybení je závažnější? Všeobecně se domníváme, že mnohem závažnější je odsoudit nevinného člověka než neodsoudit skutečného zločince (navíc, upálíme-li někoho, jistě s tím později již nebudeme moci nic dělat) a takto se chovají i testy. Proto se také běžně stanovuje α pevně na nějakou hodnotu – chceme mít jistotu pro velikost chyby, že odsoudíme nevinného.

A co soudcovo pojmenování „Měkouš“? Byl tento soudce špatný? Kvalita soudce by měla být zkoumána nejen podle toho, jakou chce mít jistotu neodsouzení nevinného, ale také podle toho, jak správně odsoudí skutečně vinného (což určuje v terminologii testování hypotéz vlastně síla testu). Žádné z těchto pochybení není možné úplně vyrušit, protože snížím-li možnost výskytu jednoho typu

pochybení soudce, enormně vzroste výskyt druhého pochybení. Chyby prvního a druhého druhu jdou proti sobě.

Jak tedy poznat kvalitního soudce? V teorii testování hypotéz se to řeší následovně: stanoví se pevná hladina α a z testů, které ji dosahují, se vybere ten, který má nejmenší chybu druhého druhu, tedy největší sílu testu. Rozhodnutí o pojmenování „Měkouš“ bychom tedy museli rozhodnout na základě porovnání s jinými soudci, kteří se chovají stejně liberálně jako on. Různé možnosti vztahu mezi skutečností a rozhodnutím soudu (výsledkem testu) ukazuje následující tabulka:

		rozhodnutí	
		obžalovaný je nevinen (H_0 platí)	obžalovaný je vinen (H_1 platí)
skutečnost	obžalovaný je nevinen	správné rozhodnutí	CHYBA I. DRUHU
	obžalovaný je vinen	CHYBA II. DRUHU	správné rozhodnutí

Pravděpodobnost chyby prvního druhu (α) nazýváme hladina významnosti testu. Pravděpodobnost správného zamítnutí neboli síla testu je 1-β.

Závěrem

Ať už se příběh stal nebo nestal, osobně jsme s rozsudkem soudce spokojeni. Přeci jenom by nám bylo líto odsouzení nadějného mladého statistika z čarodějnictví…

5. Bland Altmanův graf

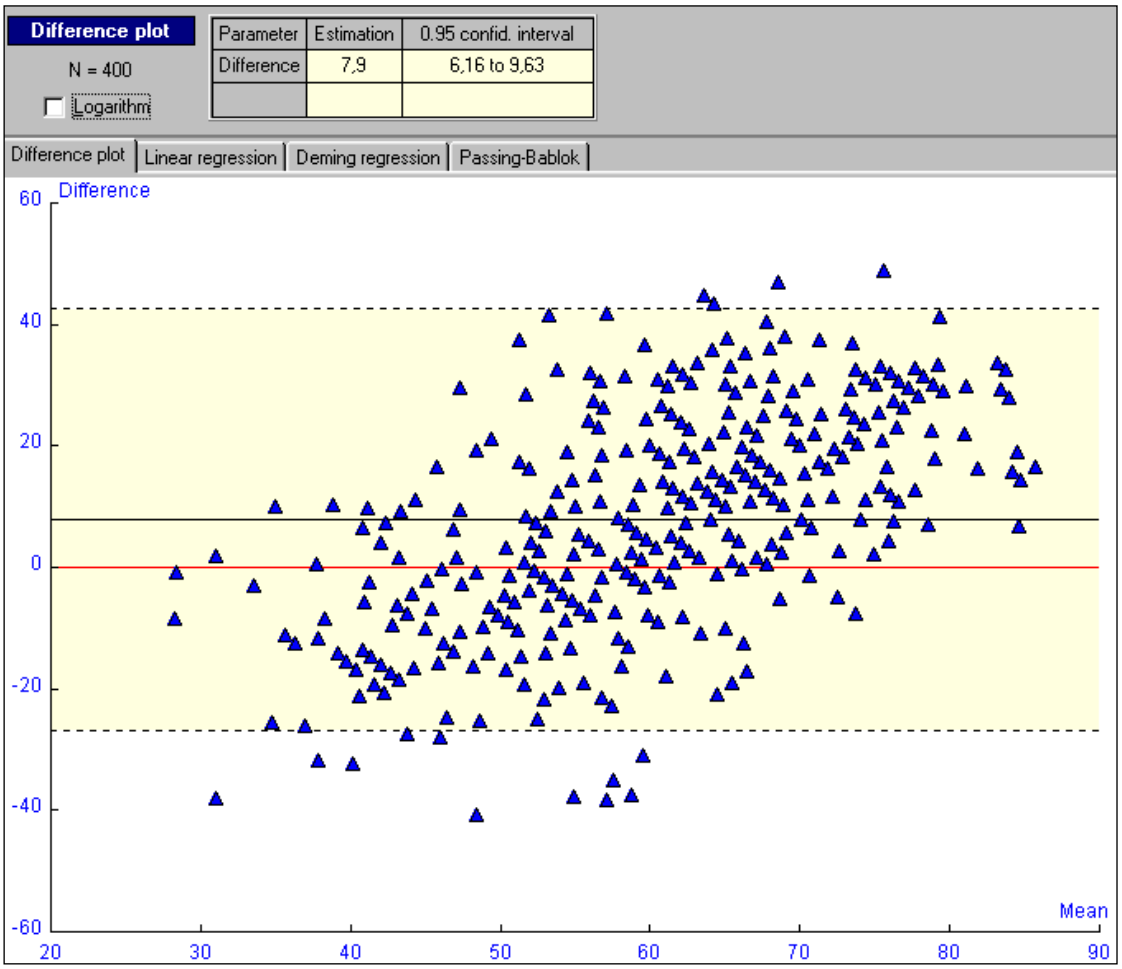
Je tradicí používat jako kritérium srovnatelnosti metod či měření korelační koeficient. Korelace ale posuzuje pouze míru závislosti mezi metodami, není kritériem pro posouzení shody. Korelační koeficient indikuje pouze, jak těsně se v grafu závislosti výsledků obou metod přimykají experimentální body k regresní čáře, nejčastěji přímce. Korelační koeficient nevypovídá ale nic o shodě regresní přímky s přímkou ideální závislosti obou metod. Pokud tedy například metoda 1 dává vždy dvakrát větší výsledek než metoda 2, potom se jedná o velice těsnou závislost (s korelačním koeficientem blízkým jedničce), avšak o shodě výsledků nemůže být řeč. Další problém korelačního koeficientu spočívá v tom, že jeho hodnota je závislá na rozmezí, v němž získáváme výsledky měření. Širší rozmezí dat má vždy tendenci vykazovat větší korelační koeficient (lepší korelaci). Při porovnávání výsledků dvou metod se pak často jeví dobře míněná snaha zařazovat hlavně extrémní data z důvodu zlepšit korelační koeficient.

Tyto problémy jsou už delší dobu popisovány ve statistické literatuře. Altman a Bland (1983) publikovali práci, která poukazuje na jejich důležitost a navrhuje alternativní metodu analýzy. Tato práce byla publikována v časopise Statistician, který je zaměřen na profesionální statistiky. Všeobecně bylo uznáno, že přístup navržený Blandem a Altmanem je technika vhodná k analyzování dat ve studiích porovnávajících.

Základem přístupu Blanda a Altmana je posouzení diferencí mezi oběma metodami. Do grafu vyneseme rozdíly mezi oběma metodami v závislosti na průměrech těchto dvou metod. Tento rozdílový graf umožňuje posouzení, zda rozdíly mají systematický charakter (zda se diference systematicky liší od nuly) a jak moc diference kolísají (jejich rozptyl či směrodatná odchylka). Na obr. 1 je rozdílový graf vstupních dat. Je zřejmé, že mezi metodami je systematický rozdíl (rozdíly nejsou rozloženy symetricky kolem nuly). Většina rozdílů leží mezi -20 a +40 a v rozdílech se projevuje trend, t.j. rozdíly jsou závislé na průměru.

	Průměr	**Interval spolehlivost**	**Limit shody**	**Std.Dev.**	**Standard Error**
rvs	63,93	< 62,14 ; 65,73 >		18,23	0,91
tsp	56,04	< 55,08 ; 57,00 >		9,76	0,49
rvs-tsp	7,89	< 6,13 ; 9,66 >	< -27,40 ; 43,19 >	17,65	0,88
log(rvs)-log(tsp)	0,038	< 0,029 ; 0,057>	< -0,23 ; 0,32 >	0,138	0,007

Jestliže se ukáže, že existuje závislost mezi rozdíly metod a průměry metod, pokusíme se nejprve použít logaritmické transformace dat. Tato transformace často závislost potlačí, takže limity shody mohou být odlogaritmovány a poskytnou rozmezí v násobcích průměru, nikoliv v absolutních hodnotách. Jiný druh transformace nelze doporučit, poněvadž při tomto způsobu je rozumné zpětně transformovat pouze logaritmy.



Obr. 1

Na obr. 2 je rozdílový graf dvou stanovení a jsou v něm zakresleny průměrný rozdíl a limity shody. Se zvětšujícím se průměrem vykazují rozdíly tendenci k menšímu rozptýlení. To je případ, kdy limity shody získané bez transformace budou pro vysoké hodnoty podhodnoceny (příliš široké) a pro nízké nadhodnoceny (příliš úzké). Po logaritmické transformaci vidíme závislost logaritmů rozdílů na průměrech. Distribuce logaritmů rozdílů je zřetelně rovnoměrnější než rozdílů samotných. Limity shody logaritmovaných dat -0,23 a +0,32 ukazují, že asi u 95 % případů je hodnota stanovení č. 2 v mezích daných 0,58 násobkem a 2,09 násobkem hodnoty stanovení č. 1. Uvedená čísla jsou získána odlogaritmováním čísel -0,23 a +0,31.

Pro data znázorněná na obr. 2 je průměr rozdílů 0,038 a směrodatná odchylka 0,138. Směrodatná odchylka průměru rozdílů (standard error of mean) SEM se vypočte dělením směrodatné odchylky rozdílů druhou odmocninou počtu pozorování n, v našem případě

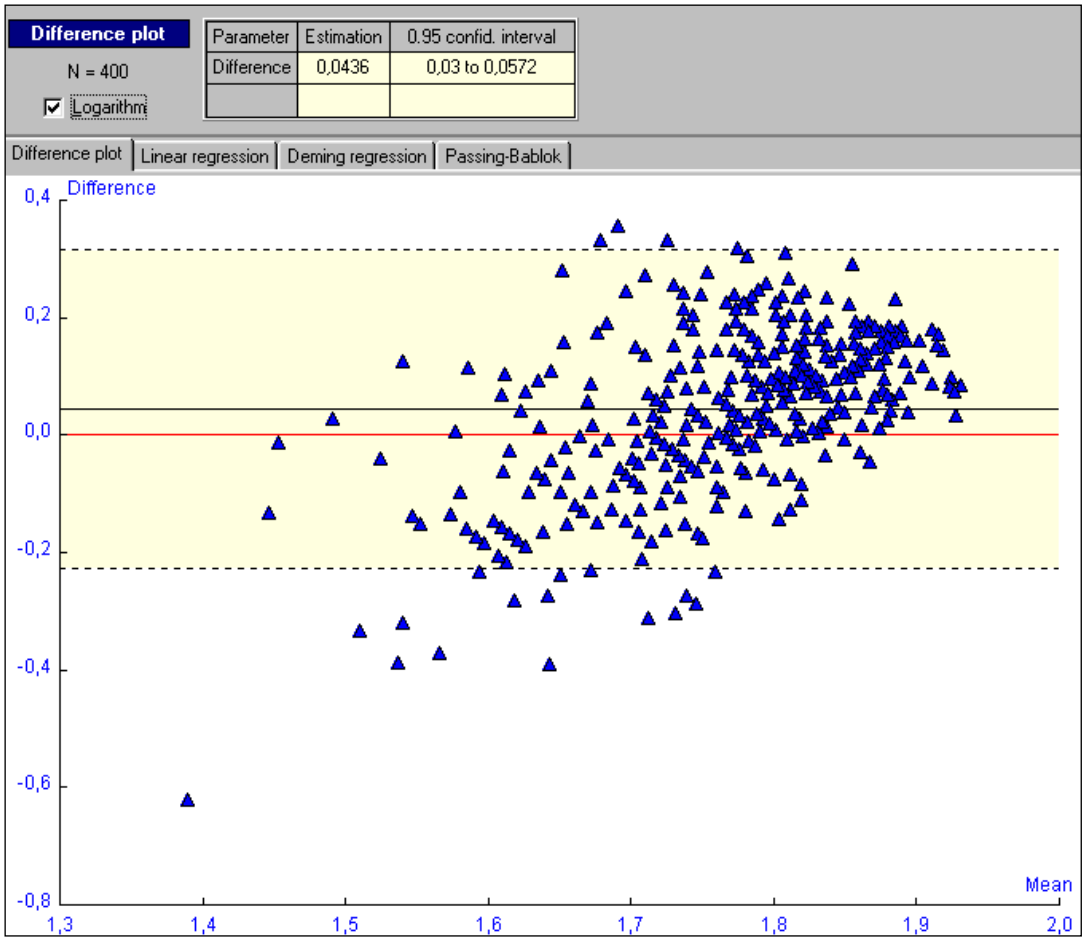
SEM = 0,138 / SQRT(400) = 0,0069

95 % interval spolehlivosti průměru rozdílů IS (udává neurčitost průměru) je potom

IS = 0,038 ± 2 * 0,0069 = <0,029 ; 0,057>

LS = 0,038 ± 2 * 0,138 = <-0,23 ; 0,32>

Neexistuje definitivní pravidlo, kdy má být použita logaritmická transformace. Je třeba mít na paměti, že cílem je určit limity shody, které jsou validní pro celé rozmezí hodnot. Někdy bude nemožné stanovit jednoduché limity shody a to tehdy, když logaritmování závislost neodstraní nebo když jsou v datech odlehlé body, které výrazně ovlivňují hodnotu směrodatné odchylky. V těchto případech se má rovněž zkonstruovat rozdílový graf jak je popsáno výše, ale limity shody by měly být určeny jinak než statisticky.



Obr. 2

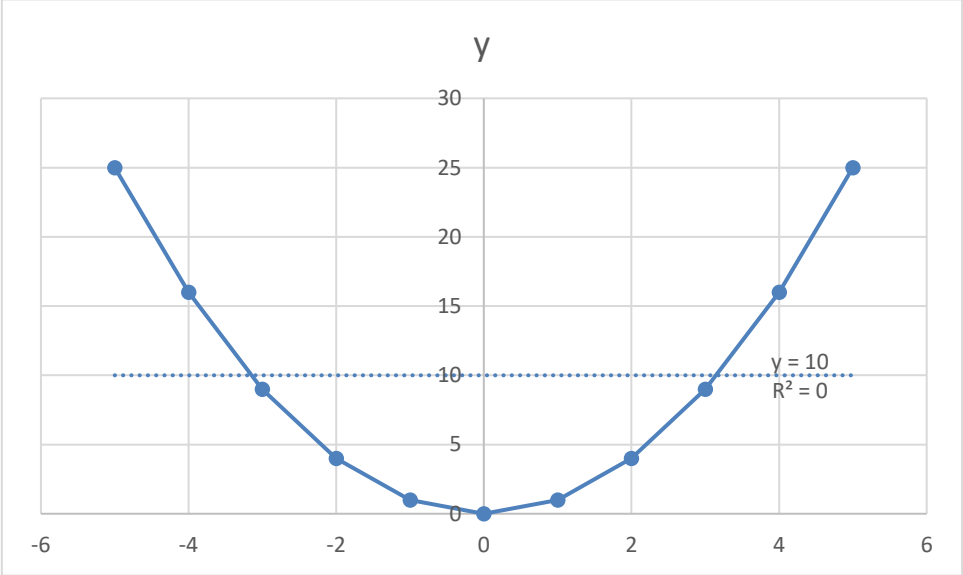
- Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, (3). 307.
- Dohnal, L. (2002). *Porovnání výsledků dvou metod stanovení, lineární regresí nebo jinak?* Retrieved form http://www1.lf1.cuni.cz/~ldohna/vhodregr/obsah.htm.

6. Korelace = 0 nemusí vždy znamenat nezávislost

Mějme následující data

x	y
-5	25
-4	16
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9
4	16
5	25

Pokud změříme závislost pomocí korelačního koeficientu, vyjde nám jeho hodnota rovna 0, což značí, že zde není **LINEÁRNÍ ZÁVISLOST**. A to je pravdivé tvrzení. Závislost mezi dvěma proměnnými zde ale je, vyjádřena polynomem 2. stupně neboli parabolou.



Závěr: Korelační koeficient testuje lineární závislost. Grafické posouzení experimentálních dat by mělo vždy předcházet výpočtům. Rychlá orientace v grafickém vyjádření pomůže správnému pochopní modelovaných závislostí.

7. Standardizované z-skóre mimo interval -3 a 3

Standardizované skóre (též standardní skóre) je ve statistice označení pro čísla, vzniklá lineární transformací z původně naměřených či jinak zjištěných hodnot (označovaných jako hrubé skóre) tak, aby výsledné rozložení mělo předem dané vlastnosti. Nejčastějším příkladem standardizovaného skóre je z-skóre s průměrem 0 a směrodatnou odchylkou 1. Předpokladem pro použití standardizovaných skóre je normální rozdělení původních hodnot.

Výpočet standardního skóre se řídí podle následujícího vzorce:

$$x' = \mu' + \sigma' \frac{(x - \mu)}{\sigma}$$

kde x' je standardizovaný skór, x původní hrubý skór (který chceme standardizovat), σ původní směrodatná odchylka, μ původní průměrná hodnota, σ' požadovaná směrodatná odchylka standardizovaných skóre a μ' požadovaná průměrná hodnota standardizovaných skóre. V případě nejčastějšího z-skóre (z), které má průměr 0 a směrodatnou odchylku 1, lze vzorec zjednodušit:

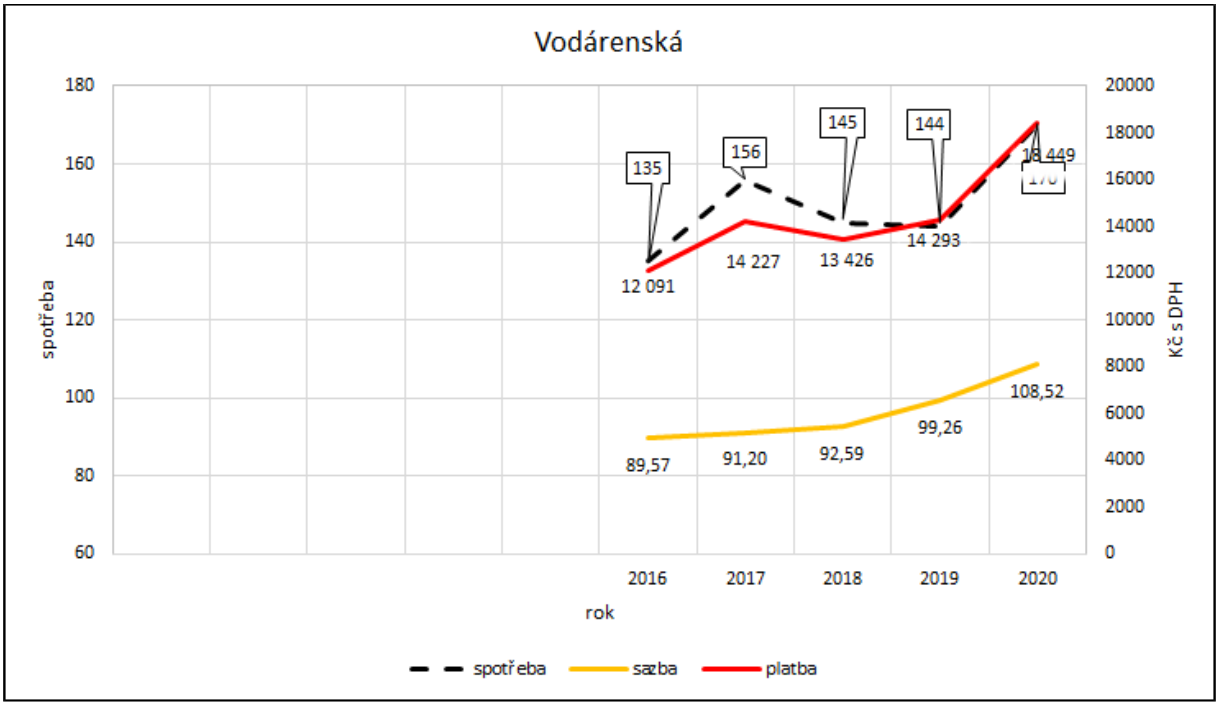
$$z = \frac{x - \mu}{\sigma}$$

Velmi často se uvádí, že výsledné standardizované z-skóre je z intervalu (-3; 3). Data s extrémními hodnotami, mohou ve výsledků být i mimo interval. Např.

x	z
1	-0,35
2	-0,35
5	-0,35
2	-0,35
1	-0,35
4	-0,35
5	-0,35
2	-0,35
50000000000	3,02
průměr	5555555558,00
sm. odch.	15713484025,50

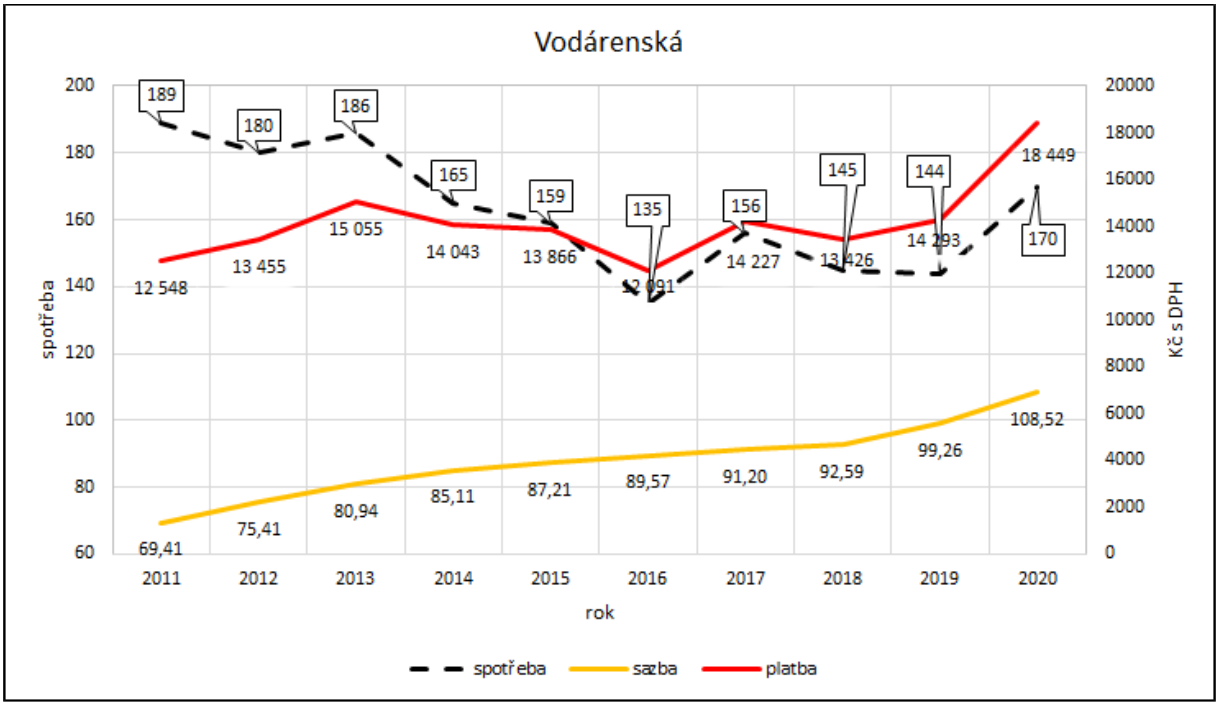
Závěr: Je to jen hypotetické porušení obecně používaného pravidla u z-skóre.

8. Vysoká hra grafů

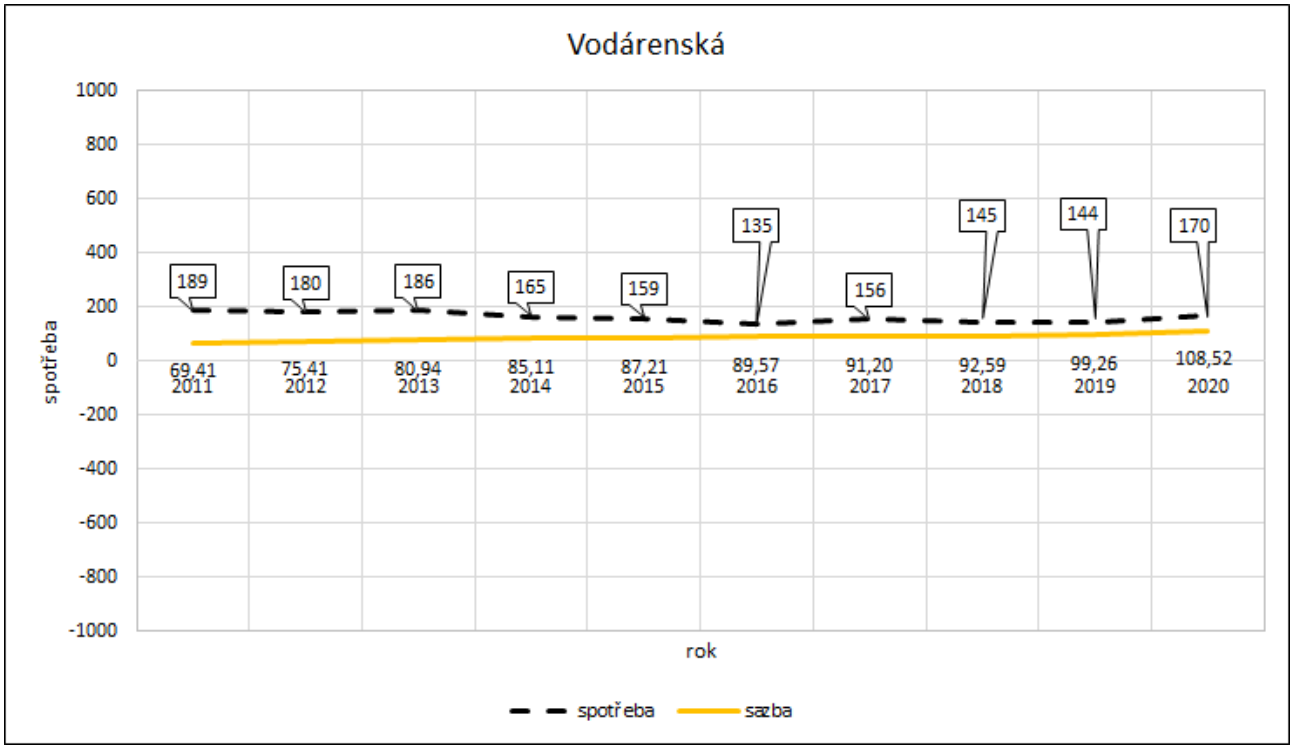


Co řekneme o trendu vývoje spotřeby vody v domácnost ze znalosti roku 2016-2020?

Pozn., prosím odmyslete si tu levou prázdnou část ☺



A nyní?



A nyní? Poté, co jsme změnili jednotky na ose Y?

9. Co je a co není statistická významnost α a věcná významnost

výsledek je statisticky významný:

- NEZNAMENÁ, že je „významný“ ve smyslu vědeckého důkazu
- NEZNAMENÁ, že je „důležitý nebo podstatný“

Statisticky významný znamená jen a jen, že výsledek je „statisticky zobecnitelný“ z reprezentativního-randomizovaného výběru na základní soubor a to se zvoleným rizikem α

Příklad 1

- ✓ 6 měsíční tréninková intervence
- ✓ skupina 10 sprinterů na 100 m s velmi slabou výkonností (cca 20 s).
- ✓ došlo k průměrnému zlepšení o 0,05 s.

Jak se na toto zlepšení můžeme dívat?

Diskuse: Rozdíl není statisticky významný (viz tab.). Ke zlepšení de facto vůbec nedošlo. Rozdíl 0,05 s totiž mohl být způsoben mnoha faktory. Přiznejme, že jedním faktorem mohl být i trénink ☺.

Příklad 2

Vše zůstává, jen těchto slaboučkých sprinterů je 1000.

Diskuse: Rozdíl je statisticky významný!

Variable	T-test for Dependent Samples (statistika-myty a omyly 2020) Marked differences are significant at p < ,05000								
	Mean	Std.Dv.	N	Diff.		t	df	p	Cohen d
Př. 1: výkon 100 m (1)	20,59348	0,282457							
Př. 1: výkon 100 m (2)	20,36741	0,349495	10	0,226067		1,343253	9	0,212078	**0,68**
Př. 2: výkon 100 m (1)	20,50885	0,283752							
Př. 2: výkon 100 m (2)	20,40449	0,289174	1000	0,104359		8,007576	999	0,000000	**0,36**

Příklad 3

Vše zůstává stejně jako v př. 1.

Nyní se jedná o skupinu elitních světových sprinterů (časy 10 s na 100 m).

Jak se na toto zlepšení můžeme dívat?

Diskuse: Zlepšení o 0,05 s? Naprosto nevídaném zlepšení a sen všech těchto sprinterů. A ani na to nepotřebuji statistiku (věcnou významnost jsem stanovil pomocí absolutního rozdílu).

10. Čísla nelžou, ale… Čelíme nákaze opravdu úspěšněji než Německo?

Data k 30. 7. 2020

Ano, čísla nelžou, ale jejich výklad může svést z cesty.

Covid-19 a vývoj za poslední týden nebo den. Za úterek u nás bylo na covid pozitivně testováno 275 lidí, v Německu 684. Přepočteme-li to na lidnatost (poměr 1 : 8), zjistíme, že třikrát lépe na tom není Česko, nýbrž právě Německo.

Žádná "tvrdá" čísla nejsou. Liší se metody vyhodnocování, spolehlivost testů, počty (absolutní i relativní) testovaných osob, složení testovaných skupin není reprezentativní atd. atd. Závěry z toho může dělat leda tak věštkyně na nejmenované TV a bulvarizovaná média, která prodávají špatné až děsivé zprávy…

Zdroj: https://www.lidovky.cz/nazory/petracek-cisla-nelzou-ale-celime-nakaze-opravdu-uspesneji-nez-nemecko.A200730_210717_ln_nazory_sei

11. Koronavirus: exponenciální nebo logistický trend?

Nejprve si na dávné legendě o vzniku šachů ukážeme sílu exponenciálního rozdělení. Kdosi kdysi vymyslel hru šachy, a když ji představil svému králi, ten by nadšený a vynálezce se rozhodl odměnit. Autor byl znalý exponenciálního rozdělení a tak požádal o 1 zrnko pšenice na 1. políčko. Na další políčko pak dvojnásobek předchozího. Tedy políčko č. 2 mělo 2 zrnka, políčko č. 3 4 zrnka pšenice atd. Král se pousmál, že s takovou žádostí nemá problém… A to do chvíle než zjistil vlastnost exponenciálního rozdělení… ☺

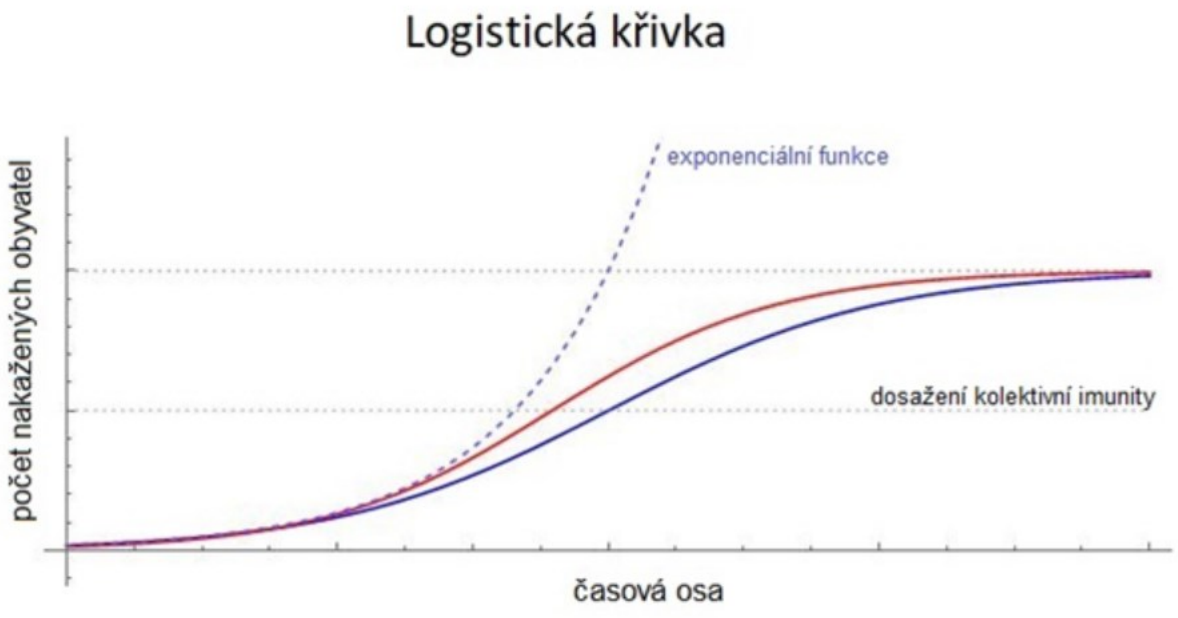
Pole	Zrnek
1	1
2	2
3	4
4	8
5	16
6	32
7	64
8	128
9	256
10	512
11	1 024
12	2 048
13	4 096
14	8 192
15	16 384
16	32 768
17	65 536
18	131 072
19	262 144
20	524 288
21	1 048 576
22	2 097 152
23	4 194 304
24	8 388 608
25	16 777 216
26	33 554 432
27	67 108 864
28	134 217 728
29	268 435 456
30	536 870 912
31	1 073 741 824
32	2 147 483 648
33	4 294 967 296

Pole	Zrnek
34	8 589 934 592
35	17 179 869 184
36	34 359 738 368
37	68 719 476 736
38	137 438 953 472
39	274 877 906 944
40	549 755 813 888
41	1 099 511 627 776
42	2 199 023 255 552
43	4 398 046 511 104
44	8 796 093 022 208
45	17 592 186 044 416
46	35 184 372 088 832
47	70 368 744 177 664
48	140 737 488 355 328
49	281 474 976 710 656
50	562 949 953 421 312
51	1 125 899 906 842 620
52	2 251 799 813 685 250
53	4 503 599 627 370 500
54	9 007 199 254 740 990
55	18 014 398 509 482 000
56	36 028 797 018 964 000
57	72 057 594 037 927 900
58	144 115 188 075 856 000
59	288 230 376 151 712 000
60	576 460 752 303 423 000
61	1 152 921 504 606 850 000
62	2 305 843 009 213 690 000
63	4 611 686 018 427 390 000
64	9 223 372 036 854 780 000
Celkem	**18 446 744 073 709 600 000**

Celkem se jedná o 18 446 744 073 709 600 000, tedy o $1{,}8^{19}$ zrnek (rychlý součet geometrické posloupnosti je 2^{64}-1), což je více, než je aktuální celosvětová produkce.

Epidemie koronaviru se může v počátcích též šířit exponenciálně. A takový trend by byl zničující a devastující. Pokud bychom Českou republiku s cca 10 mil obyvateli nasadili na trend šíření „šachové“ exponenciály, tak už v 34. dni máme cca 8,6 mil nakažených obyvatel (viz předchozí tabulka).

Naštěstí realita není takto přísná, vstupuje zde mnoho různých faktorů, které ovlivňují šíření koronaviru. Ve skutečnosti je trend spíše podobný logistické křivce. Začátek je exponenciální, pak dochází ke zlomu v rychlosti šíření (inflexní body logistické křivky), pak následuje zpomalení šíření, a následně zastavení a pokles. Nebezpečí spočívá v rychlosti nárůstu v počátcích epidemie, kdy může být zahlcen zdravotnický systém se svými defacto fixními kapacitami. Proto se zavádějí opatření ke zmírnění šíření…



Použito z https://www.matfyz.cz/clanky/matematika-koronaviru-exponenciala-vs-logisticka-krivka
Autor: Tereza Bártlová, citováno 12. 12. 2020

12. Hesla, mota, citáty

- Generování náhody je příliš důležité, než abychom ji mohli ponechat náhodě
- Jestliže má jednotlivec rád čísla, pokládá se to za neurózu. Celá společnost se ale sklání před statistickými čísly. Alfred Paul Schmidt
- Když má hlavu v sauně a nohy v ledničce, hovoří statistik o příjemné průměrné teplotě. Franz Josef Straus
- Máloco je statisticky tak dobře doloženo jako lidská smrtelnost ve 100 %. Stanislav Komárek
- ~~Statistika je děvka, nechá se od každého znásilňovat. Jaromír Korčák~~
- Statistika je jako bikini. Co odhaluje, je zajímavé, co skrývá, je podstatné. Aaron Levenstein
- Statistika je jako naivní stará dáma. Podle toho jak se jí otážeme, tak odpoví. Helmut Müller
- Když lovec mine zajíce jednou zleva a podruhé zprava, je zajíc v průměru mrtvý.
- Mnozí (politici, manažeři, atd...) používají statistiku jako opilec pouliční lampu - k udržení rovnováhy a ne k osvětlení!
- Nedůvěřuj statistice, kterou jsi sám nezfalšoval!
- Statistika je metoda, jak vyjádřit nejistá data s přesností na setinu procenta.
- Statistika je přesný součet nepřesných čísel... Oldřich Fišer
- „Podle statistiky připadá u nás na každou rodinu 4,1 osoby. Ta jedna desetina je otec.“ Josef Lukl
- Statistiky dokazují, že manželství je nejlepší prevence proti sebevraždám. Ovšem sebevražda je nejlepší prevencí proti manželství