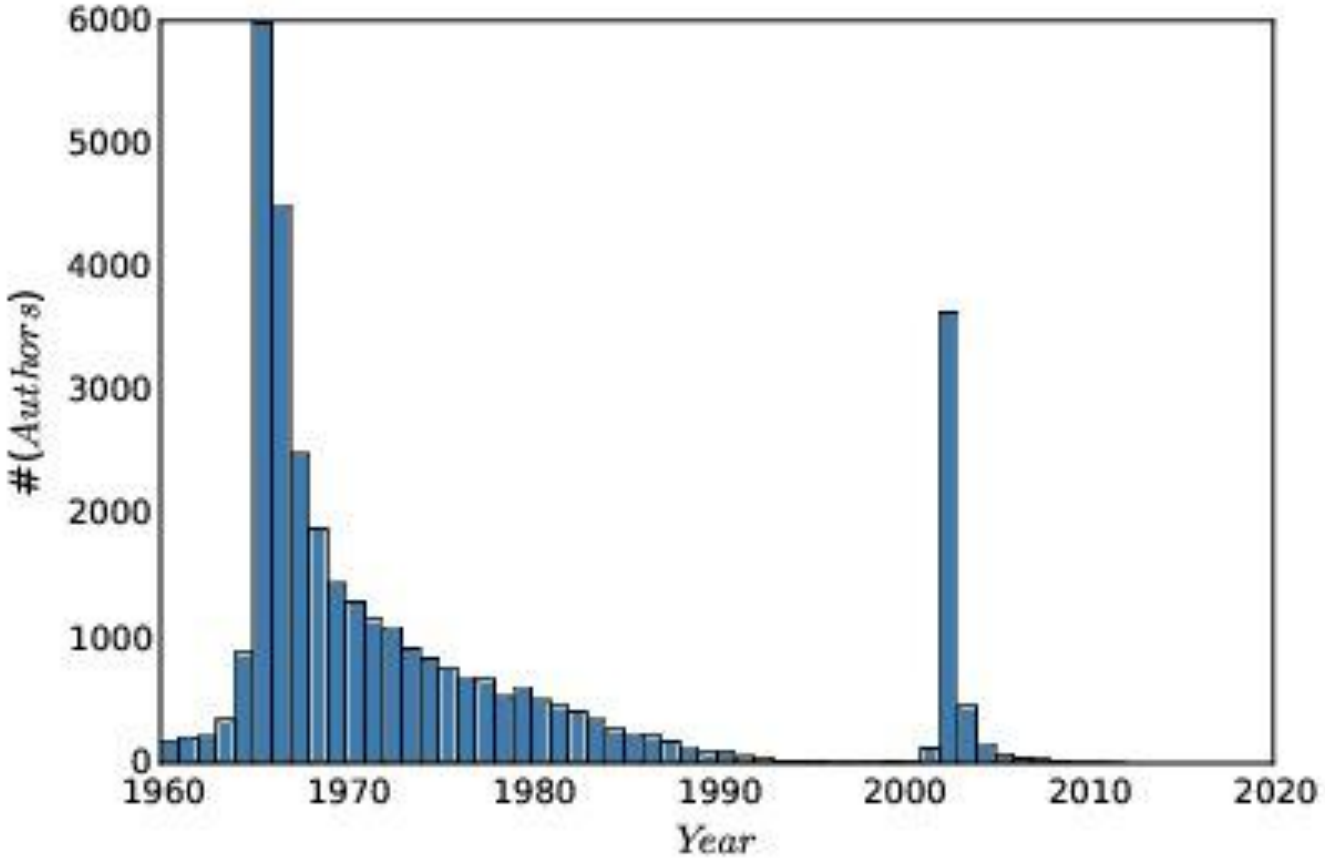
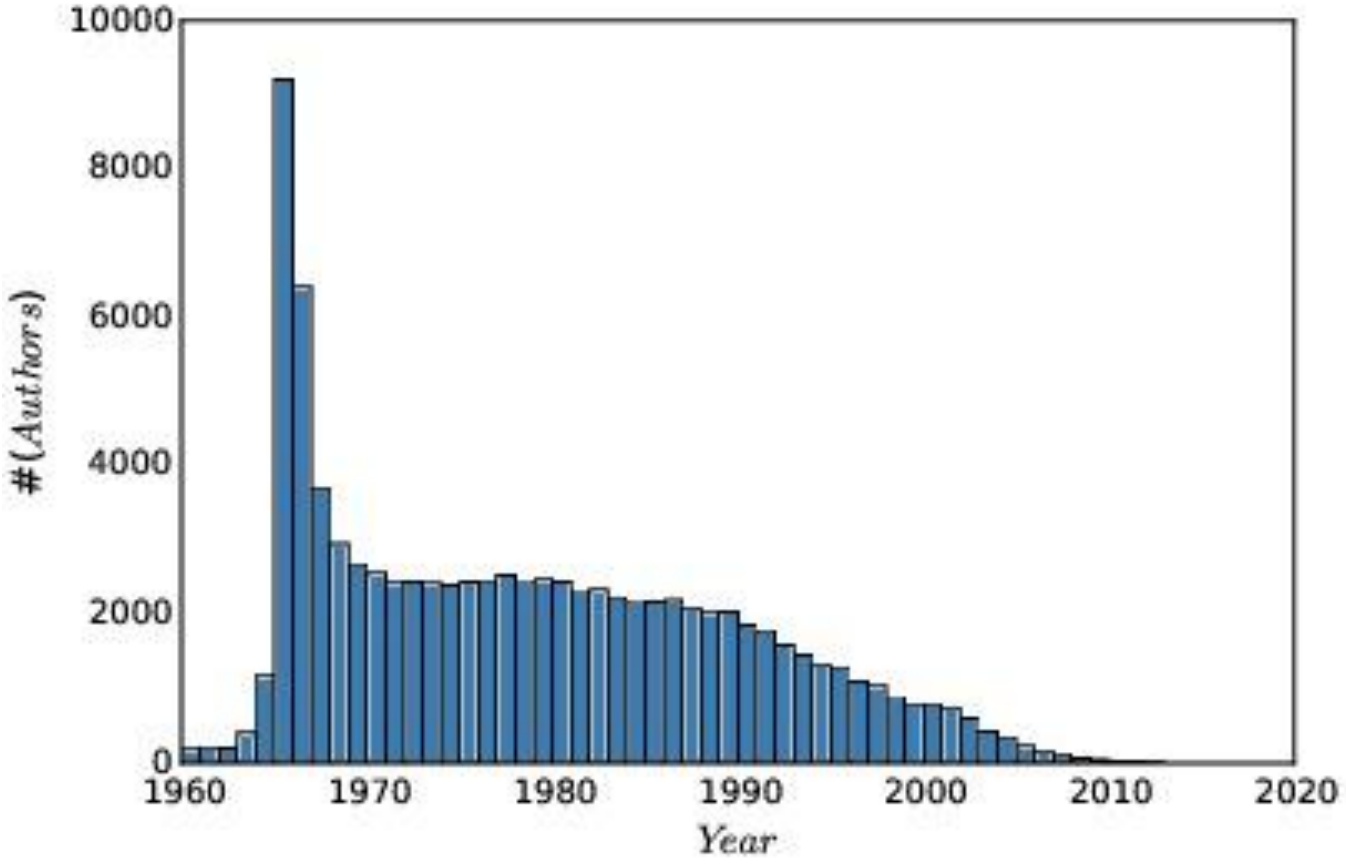
Kedy začalo publikovať 100000 najcitovanejších vedcov?

Kedy začalo publikovať 100000 najcitovanejších vedcov?

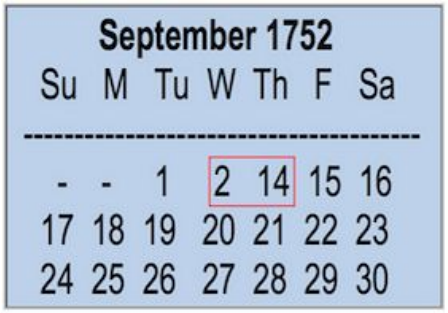


Kedy začalo publikovať 100000 najcitovanejších vedcov?

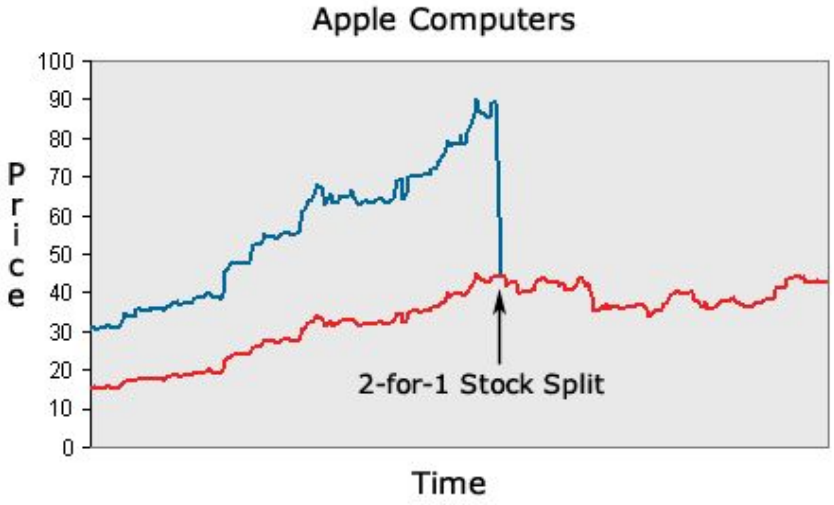


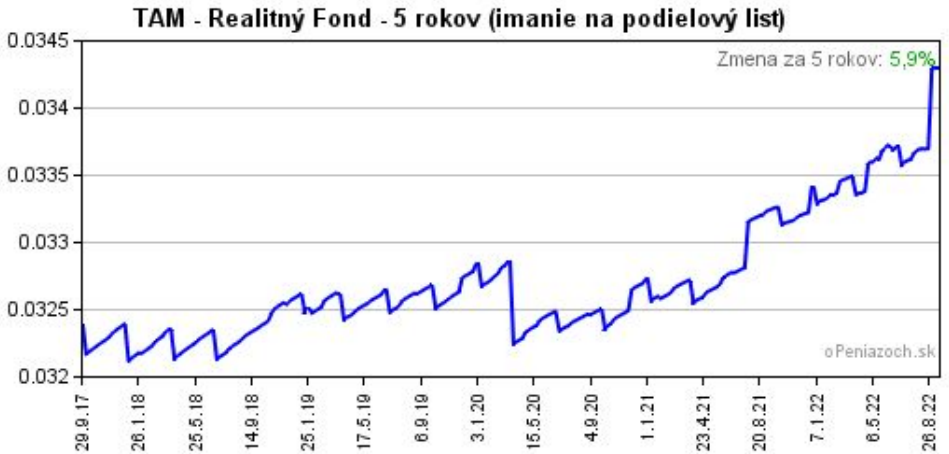
Problémy s konverziami

- Pri meraných hodnotách, čiarka alebo bodka nemusí byť vždy desatinné číslo (napr. feet,in lb,oz a pod.)
- locale pri veľkých hodnotách (čiarky, bodky, medzery…)
- reprezentácia reťazcov (napríklad - má niekoľko rôznych UTF-8 hodnôt)
  - skús soundex
- časové zóny
- dátumy a staré dátumy
- voľné dni (stock prices vs. počasie)
- kurzy a hodnota peňazí



- Česká republika dáva na školstvo oveľa viac peňazí ako Slovensko
- Ceny akcií dlhodobo korelujú s cenami ropy
- Ktorý film zarobil najviac peňazí v histórii? Gone with the Wind $390 mil.
  Harry Potter (8 filmov) $7.7 mld.



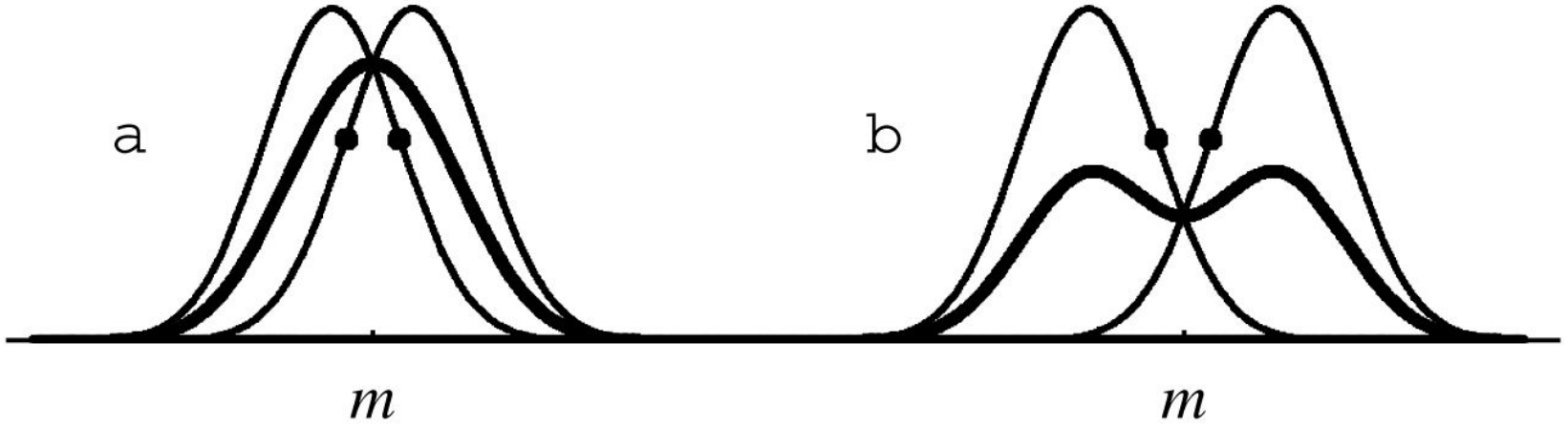


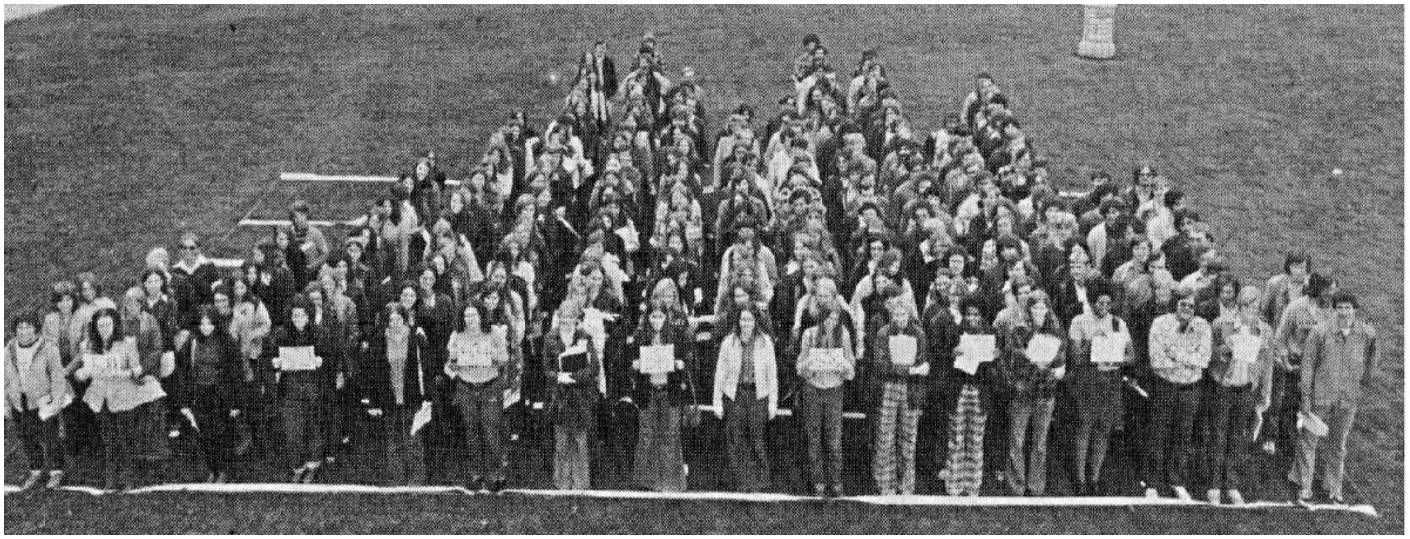
Bimodálne distribúcie

- Bimodálna distribúcia čohokoľvek je podozrivá
- Môže vzniknúť napr. chybným (alebo chýbajúcim) prepočtom jednotiek
- Z-skóre je celkom dobrá “bezrozmerná” veličina

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

Výška ľudí MÔŽE BYŤ bimodálna





Schilling, Watkins, Watkins, 2014

Chýbajúce pozorovania (ascertainment bias)

- Vedľajšie účinky očkovania sú oveľa horšie ako priebeh COVIDu

Science

REPORTS

Cite as: M. Pavelka *et al.*, *Science* 10.1126/science.abf9648 (2021).

The impact of population-wide rapid antigen testing on SARS-CoV-2 prevalence in Slovakia

Martin Pavelka[1,2,3]*, Kevin Van-Zandvoort[4,5], Sam Abbott[4,5], Katharine Sherratt[4,5], Marek Majdan[6], CMMID COVID-19 working group[5], Inštitút Zdravotných Analýz[2], Pavol Jarčuška[7], Marek Krajčí[1], Stefan Flasche[4,5]†, Sebastian Funk[4,5]†

[1]Slovak Ministry of Health, Bratislava, Slovakia. [2]Inštitút Zdravotných Analýz (Institute of Health Analyses), Bratislava, Slovakia. [3]Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK. [4]Department for Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [5]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK. [6]Institute for Global Health and Epidemiology, Faculty of Health Sciences and Social Work, Trnava University, Trnava, Slovakia. [7]Faculty of Medicine, Pavol Jozef Šafárik University, Košice, Slovakia.

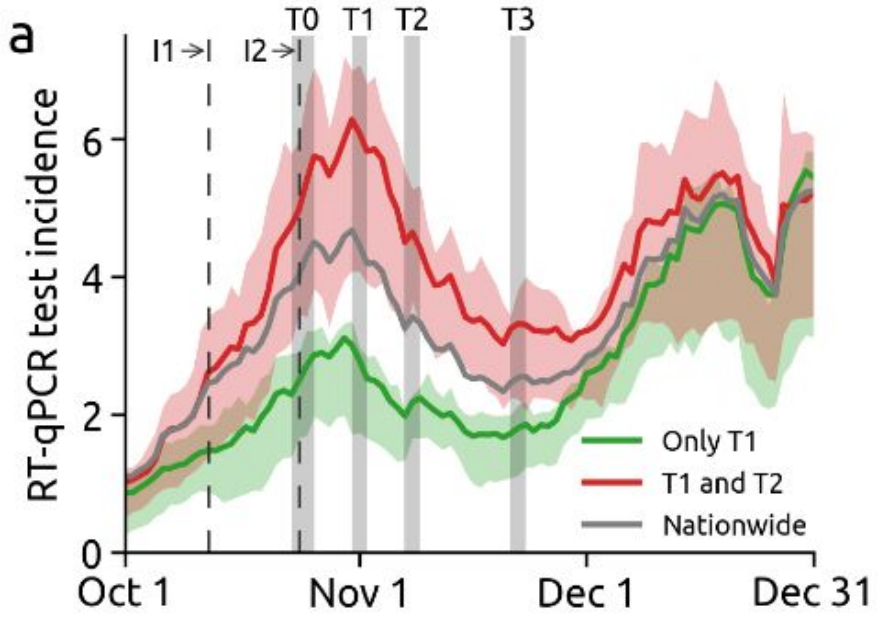
***Corresponding author. Email: martin.pavelka@health.gov.sk**

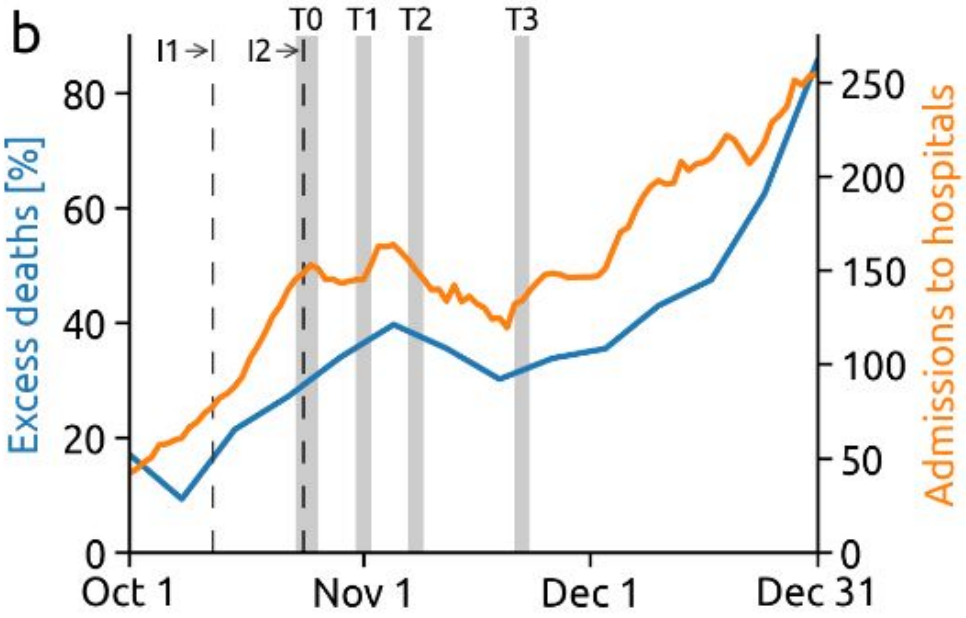
†These authors contributed equally to this work.

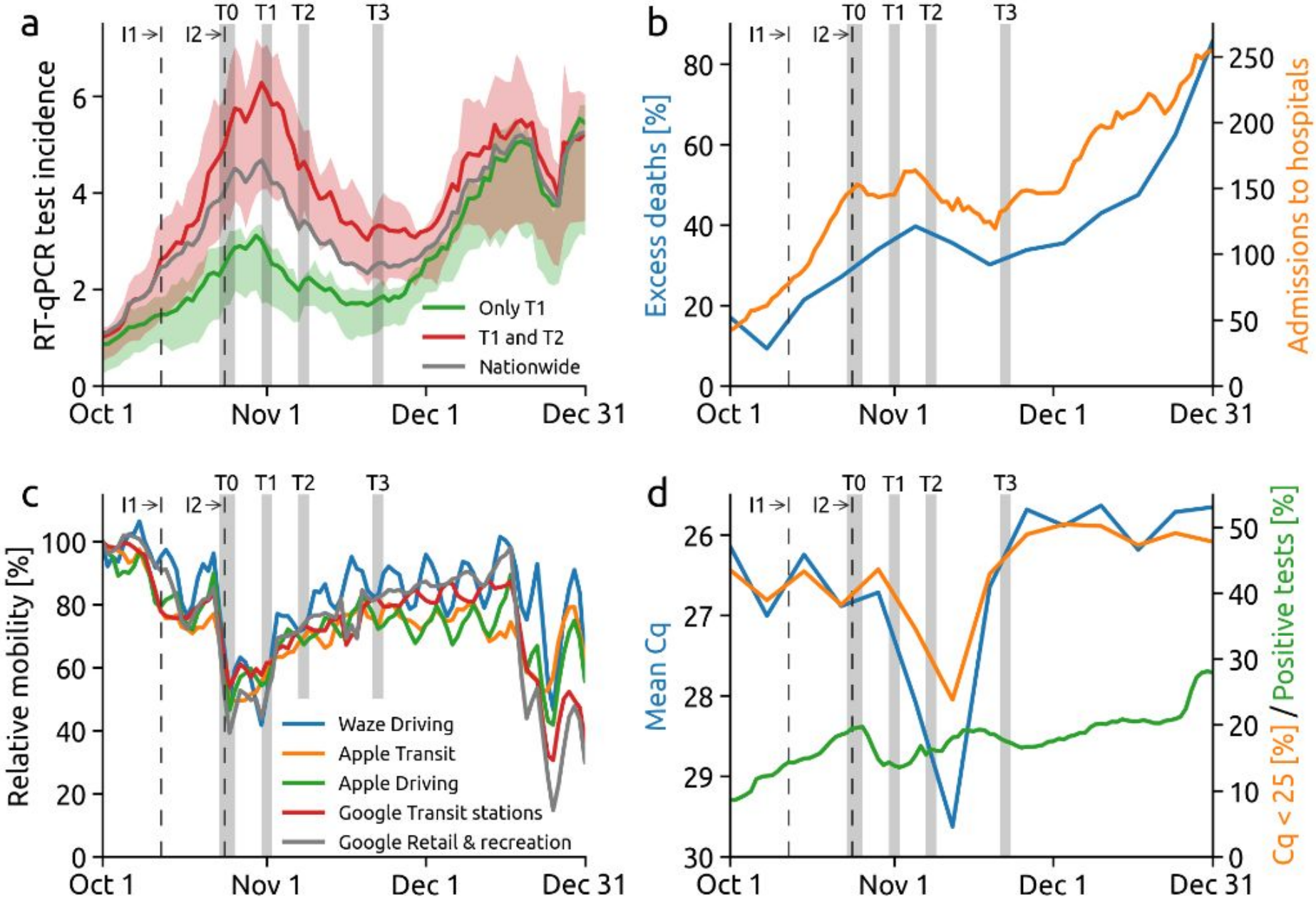
Slovakia conducted multiple rounds of population-wide rapid antigen testing for SARS-CoV-2 in late 2020, combined with a period of additional contact restrictions. Observed prevalence decreased by 58% (95% CI: 57-58%) within one week in the 45 counties that were subject to two rounds of mass testing, an estimate that remained robust when adjusting for multiple potential confounders. Adjusting for epidemic growth of 4.4% (1.1-6.9%) per day preceding the mass testing campaign, the estimated decrease in prevalence compared to a scenario of unmitigated growth was 70% (67-73%). Modelling indicated that this decrease could not be explained solely by infection control measures, but required the additional impact of isolation and quarantine of household members of those testing positive.

Výskyt ochorenia sa znížil o 58% mezi prvým (T1) a druhým kolom (T2) plošného testovania

- V prvom kole boli testovaní všetci, ktorí prišli
- V druhom kole (o týždeň neskôr) boli testovaní všetci, ktorí prišli, **okrem:**
  - tých, čo boli pozitívni v prvom kole
  - ich rodinných príslušníkov a úzkych kontaktov (už boli v karanténe)
  - tých, čo nechceli riskovať dvojtýždňové “domáce väzenie” pre celú rodinu

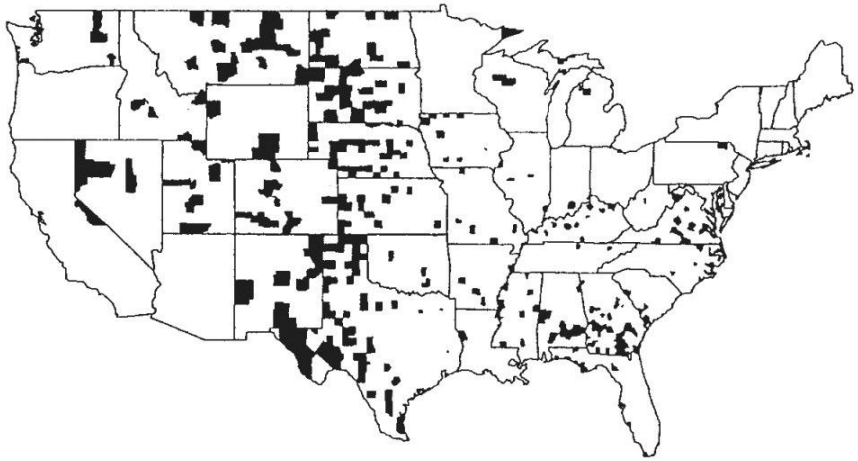




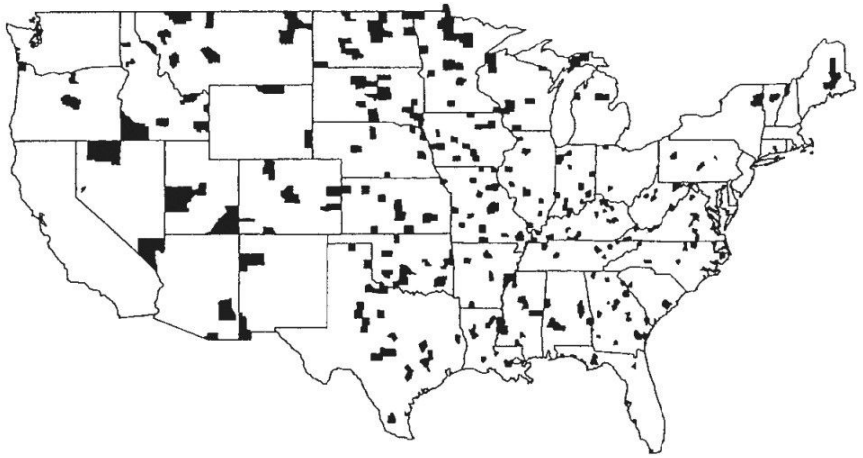

a
I1
I2
T0
T1
T2
T3
RT-qPCR test incidence
0
2
4
6
Only T1
T1 and T2
Nationwide
Oct 1
Nov 1
Dec 1
Dec 31
b
Excess deaths [%]
0
20
40
60
80
Admissions to hospitals
0
50
100
150
200
250
c
Relative mobility [%]
0
20
40
60
80
100
Waze Driving
Apple Transit
Apple Driving
Google Transit stations
Google Retail & recreation
d
Mean Cq
26
27
28
29
30
Cq < 25 [%] / Positive tests [%]
0
10
20
30
40
50

Odľahlé hodnoty (outliers)

- Najväčší evidovaný stavec dinosaura je o 50% väčší
  ako všetky ostatné nájdené stavce dinosaurov
  - skutočnosť alebo artefakt?
  - BTW, príslušné múzeum už nevie lokalizovať príslušný exponát
- Na minimá, maximá, hodnoty s príliš veľa štandardnými odchýlkami od priemeru sa treba pozerať kriticky
- Zistiť **prečo** máme odľahlé hodnoty, nie ich jednoducho mazať
  - Zmazanie odľahlých hodnôt môže viesť k lepšiemu modelu
    (napr. neobvyklé chyby merania)
  - Zmazanie odľahlých hodnôt môže viesť k horšiemu modelu
    (napr. príliš jednoduchý model dát)



Counties with the LOWEST kidney cancer death rates (1980-1989)



Counties with the HIGHEST kidney cancer death rates (1980-1989)

Gelman, Nolan: Teaching Statistics

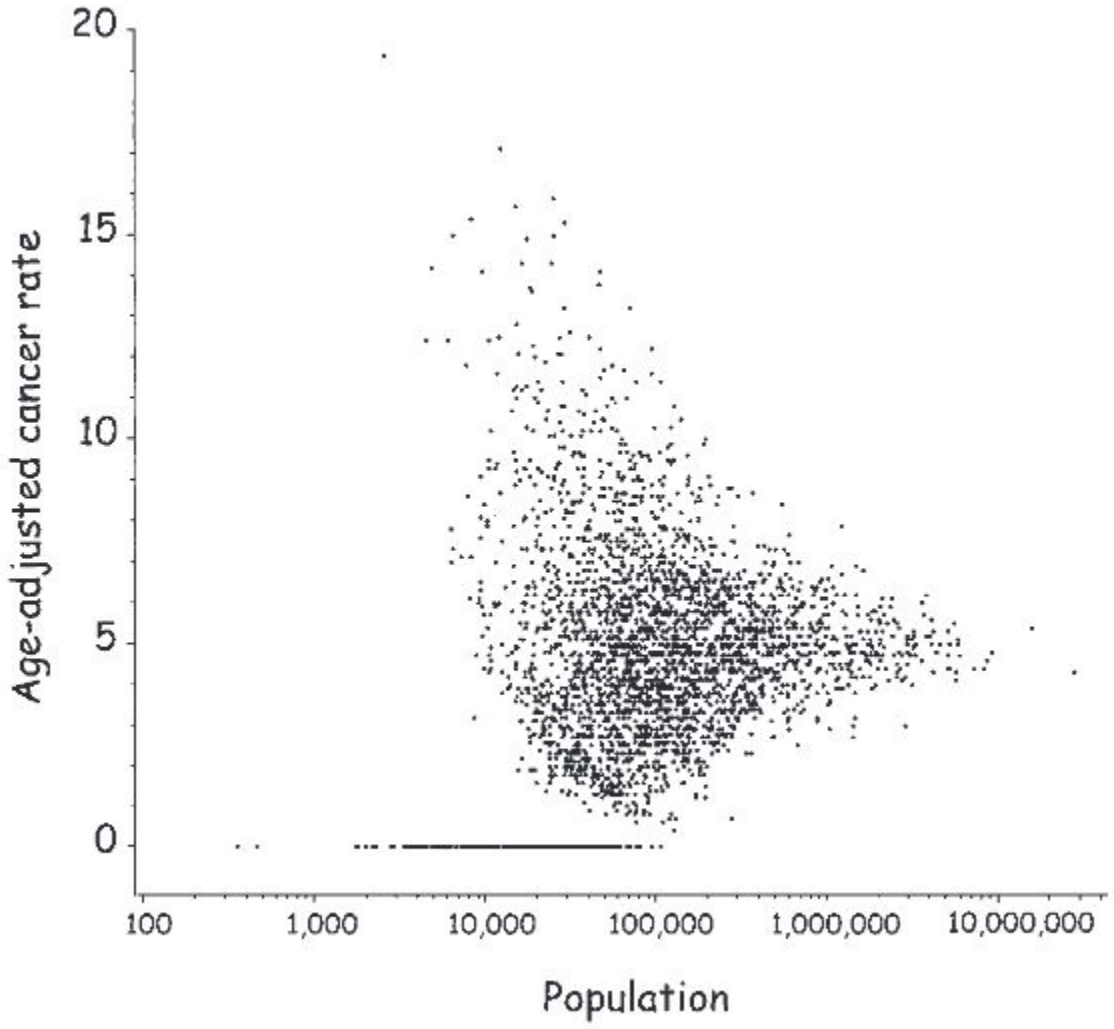


Figure 1.3.
Age-adjusted kidney cancer rates for all U.S. counties in 1980–1989 shown as a function of the log of the county population.

Howard Wainer: Picturing the Uncertain World, 2009

Chýbajúce dáta

- Väčšina metód sa nedokáže vyrovnať s chýbajúcimi dátami
  (napr. rok úmrtia osoby, ktorá ešte žije, chýbajúce meranie, zjavne nezmyselná hodnota, …)
- **Riešenie 1:** zahoď záznamy, ktoré majú chýbajúce dáta
  ak máme veľa atribútov, môže sa stať, že takmer v každom zázname bude niečo chýbať
- **Riešenie 2: imputácia chýbajúcich hodnôt**
  - priemer hodnôt daného atribútu
  - náhodná existujúca hodnota daného atribútu
  - hodnota získaná predikciou z iných atribútov (napr. lineárnou regresiou)

  Ovplyňuje zvolený spôsob imputácie výsledky analýzy?

Zhrnutie: Čistenie dát je veľmi dôležitým (a zložitým) krokom pri ich analýze

- Chyby v dátach často príčinou chybných analýz
- Náhodné chyby (napr. chyby merania) sú obvykle OK
  - pozor na početnosť hodnôt / štandardné odchýlky!
- Artefakty (systematické chyby) robia problémy:
  - chyby unifikácie dát
  - nepozorovateľné hodnoty (ascertainment bias)
  - odľahlé hodnoty (outliers)
- Chýbajúce hodnoty