Analýza počtu uvězněných občanů obcí pomocí hierarchického beta-binomiálního modelu

Tato analýza navazuje na článek z irozhlas.cz. Namísto percentuálního vyjádření počtu občanů každé z obcí percentuálně tato analýza používá hierarchický bayesovský model v prostředí Stan. Jak ukážeme níže, analýza založená na procentech může být velmi citlivá na statistický šum. Bayesovský model je naopak robustní neboť reprezentuje data pomocí parametrů dvou kategorií: lokální a globální. Lokální parametry reprezentují každou obec pravděpodobností, že náhodně vybraný občan dané obce je ve vězení. Všechny obce jsou pak ještě reprezentovány globálním parametrem, který lze interpretovat jako pravděpodobnost, že náhodně vybraný občan z náhodně vybrané obce je ve vězení. Máme tak dva typy parametrů: globální a lokální pro každou obci. Pro obce s malým počtem obyvatel bude sílit role globálního parametru, protože nemáme dost dat. Pro obce s dostatečným počtem lidí pak bude sílit role lokálního parametru. Model tak automaticky a elegantně využívá všechna dostupná data a je robustní.

Příprava dat

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)

pocob <- read_excel('../data/pocet_obyvatel_obce.xlsx', skip=6,
                    col_names = c("okres_kod", "icob", "obec",
                                  "pocobyv", "muzi", "zeny",
                                  "prum_vek", "prum_vek_muzi",
                                  "prum_vek_zeny")) %>%
  mutate(icob = as.numeric(icob)) %>%
  select(icob, obec)
```

```
## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.
```

```
pocvez <- read_csv('../data/pocvez_okresy.csv') %>%
  inner_join(pocob, by = 'icob')
```

```
## Parsed with column specification:
## cols(
##   veznu = col_double(),
##   icob = col_double(),
##   okres = col_character(),
```

```
##   pocobyv = col_double(),
##   muzi = col_double(),
##   zeny = col_double(),
##   muzu = col_double(),
##   zen = col_double(),
##   trest = col_double(),
##   vazba = col_double(),
##   dete = col_double(),
##   pct = col_double()
## )
```

```
pocvez
```

```
## # A tibble: 6,258 x 13
##    veznu   icob okres pocobyv   muzi   zeny  muzu   zen trest vazba  dete
##    <dbl>  <dbl> <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  1907 554782 Hlav~ 1294513 629550 664963  1685   222  1764   137     6
##  2    21 529303 Bene~   16522   7906   8616    17     4    19     2     0
##  3     0 532568 Bene~     233    113    120     0     0     0     0     0
##  4     0 530743 Bene~     207    104    103     0     0     0     0     0
##  5     0 532380 Bene~     116     54     62     0     0     0     0     0
##  6     0 532096 Bene~      84     40     44     0     0     0     0     0
##  7     0 532924 Bene~     770    382    388     0     0     0     0     0
##  8     4 529451 Bene~    4370   2169   2201     4     0     3     1     0
##  9     1 532690 Bene~     133     68     65     0     1     0     1     0
## 10     0 529478 Bene~     124     63     61     0     0     0     0     0
## # ... with 6,248 more rows, and 2 more variables: pct <dbl>, obec <chr>
```

Hierarchický bayesovský beta-bernoulli model

```
library(rstan)
```

```
## Loading required package: StanHeaders
```

```
## Warning: package 'StanHeaders' was built under R version 3.5.2
```

```
## rstan (Version 2.18.2, GitRev: 2e1f913d3ca3)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
```

```
##
## Attaching package: 'rstan'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
options(mc.cores = 3)

fit_hier <- stan("hier.stan", data=list(N=length(unique(pocvez$icob)),
                                        K=pocvez$pocobyv,
                                        y=pocvez$veznu),
                 chains=3, seed=25, iter = 2000, thin = 1)
```

Analýza výsledků z modelu

Theta jsou parametry modelu vyjadřující pravděpodobnost pro každou obec, že náhodně vybraný její občan je ve vězení. Vezmeme vždy 5, 50, 95-ti procentní percentil pro následující analýzu. Parametr `prob_diff` vyjadřuje rozdíl mezi theta a populačním parametrem phi, který lze interpretovat jako pravděpodobnost, že náhodně vybraný občan z náhodně vybrané obce je ve vězení.

Analýzou `prob_diff` tak můžeme najít obce, které mají výrazně vyšší či nižší podíl svých občanů vě vězení oproti celostátní tendencí vyjádřenou parametrem phi. Bayesovská inference nám dává kompletní aposteriorní distribuci těchto parametrů, takže si můžeme zvolit libovolnou hranici, kterou budeme považovat za významnou.

V následující analýze volím 95% spolehlivost, že daná obec má vyšší nebo nižší pravděpobnost výskytu svých občanů ve vězení oproti "celostátnímu průměru" (vyjádřenou phi). U takovýchobcích v následující analýze řekneme, že mají "výrazně nižší/vyšší podíl vězňů".

```
pocvez$prob_05 <- summary(fit_hier, 'theta', 0.05)$summary[, '5%']
pocvez$prob_50 <- summary(fit_hier, 'theta', 0.5)$summary[, '50%']
pocvez$prob_95 <- summary(fit_hier, 'theta', 0.95)$summary[, '95%']
pocvez$prob_diff_05 <- summary(fit_hier, 'prob_diff', .05)$summary[, '5%']
pocvez$prob_diff_50 <- summary(fit_hier, 'prob_diff', .5)$summary[, '50%']
pocvez$prob_diff_95 <- summary(fit_hier, 'prob_diff', .95)$summary[, '95%']

summary(fit_hier, c('phi', 'kappa'))$summary
```

```
##               mean      se_mean           sd         2.5%          25%
## phi   1.390859e-03 7.227132e-07 2.216278e-05 1.348132e-03 1.375974e-03
## kappa 1.802102e+03 7.614424e+00 9.554226e+01 1.615127e+03 1.738142e+03
##                50%          75%        97.5%    n_eff     Rhat
## phi   1.390358e-03 1.405118e-03 1.435385e-03 940.4088 1.002111
## kappa 1.800780e+03 1.865476e+03 1.989259e+03 157.4407 1.005139
```

Okresy s nějvětším počtem obcí, které mají výrazně vyšší podíl vězňů

```
pocvez %>%
  group_by(okres) %>%
  summarise(sigcount = sum(prob_diff_05 > 0)) %>%
  arrange(desc(sigcount))
```

```
## # A tibble: 77 x 2
##    okres        sigcount
##    <chr>           <int>
##  1 Bruntál            12
##  2 Sokolov            12
##  3 Teplice             9
##  4 Děčín               8
##  5 Litoměřice          8
##  6 Chomutov            7
##  7 Karviná             7
##  8 Most                6
##  9 Karlovy Vary        5
## 10 Kladno              5
## # ... with 67 more rows
```

Okresy s největším počtem obcí, které mají výrazně nižší podíl věznů

```
pocvez %>%
  group_by(okres) %>%
  summarise(sigcount = sum(prob_diff_95 < 0)) %>%
  arrange(desc(sigcount))
```

```
## # A tibble: 77 x 2
##    okres            sigcount
##    <chr>               <int>
##  1 Brno-venkov             6
##  2 Praha-západ             6
##  3 Havlíčkův Brod          2
##  4 Hodonín                 2
##  5 Kladno                  2
##  6 Opava                   2
##  7 Pelhřimov               2
##  8 Praha-východ            2
##  9 Uherské Hradiště        2
## 10 Ústí nad Orlicí         2
## # ... with 67 more rows
```

Obce s největším procentem svých občanů ve vězení

Všimněte si, že tomuto pořadí dominují obce s malým počtem obyvatel, u kterých i pár občanů ve vězení vyústí ve vysoké procento. Jedná se tedy spíše o statistický šum než o indikaci vysoké kriminaliti v obci.

```
pocvez %>%
  select(obec, okres, pct, pocobyv, prob_50, prob_diff_05) %>%
  arrange(desc(pct))
```

```
## # A tibble: 6,258 x 6
##    obec              okres              pct pocobyv prob_50 prob_diff_05
##    <chr>             <chr>            <dbl>   <dbl>   <dbl>        <dbl>
##  1 Županovice        Jindřichův Hradec  3.08      65 0.00227   -0.000473
##  2 Honětice          Kroměříž           2.7       74 0.00224   -0.000496
##  3 Pohorovice        Strakonice         2.63      76 0.00225   -0.000462
##  4 Lužice            Prachatice         2.56      39 0.00172   -0.000812
##  5 Mutkov            Olomouc            1.96      51 0.00171   -0.000809
##  6 Semněvice         Domažlice          1.94     206 0.00304    0.0000361
##  7 Horní Smrčné      Třebíč             1.89      53 0.00172   -0.000841
##  8 Libochovičky      Kladno             1.85      54 0.00172   -0.000799
##  9 Skapce            Tachov             1.83     109 0.00219   -0.000557
## 10 Slezské Pavlovice Bruntál            1.83     218 0.00310    0.0000747
## # ... with 6,248 more rows
```

Obce s nejvýrazněji vyšším počtem občanů ve vězení

Tento seznam už dává více intuitivní smysl. Je na první pohled vidět mnoho obcí, známých vysokou kriminalitou.

```
pocvez %>%
  select(obec, okres, pct, pocobyv, prob_50, prob_diff_05) %>%
  arrange(desc(prob_diff_05))
```

```
## # A tibble: 6,258 x 6
##    obec           okres            pct pocobyv prob_50 prob_diff_05
##    <chr>          <chr>          <dbl>   <dbl>   <dbl>        <dbl>
##  1 Trmice         Ústí nad Labem 1.17     3339 0.00796      0.00471
##  2 Ústí nad Labem Ústí nad Labem 0.580   93040 0.00569      0.00391
##  3 Bílina         Teplice        0.63    17203 0.00580      0.00357
##  4 Teplice        Teplice        0.53    49563 0.00512      0.00323
##  5 Šluknov        Děčín          0.74     5563 0.00583      0.00312
##  6 Duchcov        Teplice        0.62     8398 0.00534      0.00286
##  7 Karviná        Karviná        0.48    53522 0.00469      0.00282
##  8 Jirkov         Chomutov       0.51    19466 0.00480      0.00265
##  9 Chomutov       Chomutov       0.45    48666 0.00435      0.00248
## 10 Varnsdorf      Děčín          0.51    15429 0.00465      0.00247
## # ... with 6,248 more rows
```

Obce s procentuálně nejmenším počtem obyvatel ve vězení

Tak jako v předešlém případě, tomuto seznamu dominují obce s malým počtem obyvatel, u kterých lze pochybovat o statistické průkaznosti. Nelze tak například věrohodně tvrdit, že jsou tyto obce výrazně bezpečnější, než jiné.

```
pocvez %>%
  select(obec, okres, pct, pocobyv, prob_50, prob_diff_95) %>%
  arrange(pct)
```

```
## # A tibble: 6,258 x 6
##    obec            okres     pct pocobyv  prob_50 prob_diff_95
##    <chr>           <chr>   <dbl>   <dbl>    <dbl>        <dbl>
##  1 Bernartice      Benešov     0     233 0.00109      0.00129
##  2 Bílkovice       Benešov     0     207 0.00108      0.00136
##  3 Blažejovice     Benešov     0     116 0.00112      0.00152
##  4 Borovnice       Benešov     0      84 0.00113      0.00157
##  5 Bukovany        Benešov     0     770 0.000872     0.000750
##  6 Čakov           Benešov     0     124 0.00111      0.00149
##  7 Červený Újezd   Benešov     0     329 0.00103      0.00133
##  8 Český Šternberk Benešov     0     155 0.00112      0.00141
##  9 Drahňovice      Benešov     0      98 0.00114      0.00153
## 10 Dunice          Benešov     0      65 0.00116      0.00155
## # ... with 6,248 more rows
```

Obce s výrazně nižším počtem občanů ve vězení

V následujícím seznamu lze vidět mnoho obcí, které, vzhledem ke své velikosti, mají statistiky výrazně menzí počet obyvatel ve vězení.

```
pocvez %>%
  select(obec, okres, pct, pocobyv, prob_50, prob_diff_95) %>%
  arrange(prob_diff_95)
```

```
## # A tibble: 6,258 x 6
##    obec            okres             pct pocobyv  prob_50 prob_diff_95
##    <chr>           <chr>           <dbl>   <dbl>    <dbl>        <dbl>
##  1 Bystřice        Frýdek-Místek    0       5281 0.000311    -0.000587
##  2 Bolatice        Opava            0       4462 0.000359    -0.000511
##  3 Letohrad        Ústí nad Orlicí  0.02    6315 0.000386    -0.000509
##  4 Dolní Břežany   Praha-západ      0       3993 0.000380    -0.000434
##  5 Šlapanice       Brno-venkov      0.04    7486 0.000556    -0.000332
##  6 Humpolec        Pelhřimov        0.06   10835 0.000644    -0.000299
##  7 Velvary         Kladno           0       3031 0.000445    -0.000294
##  8 Poděbrady       Nymburk          0.06   14111 0.000705    -0.000278
##  9 Jesenice        Praha-západ      0.05    9132 0.000646    -0.000272
## 10 Mníšek pod Brdy Praha-západ      0.04    5712 0.000552    -0.000247
## # ... with 6,248 more rows
```

Ulož data

```
write_csv(pocvez, '../data/results.csv')
```