Description of the projects

Guillaume KON KAM KING

Université Paris-Saclay, INRAE, MaIAGE

Contact: guillaume.konkamking@inrae.fr

Slides available at:
`https://sites.google.com/site/guillaumekonkamking/courses`

Outline

100 songs analysed by Spotify and tested by your instructor

- For each song: did your instructor like it or not ?
- Can we generalise: what kind of songs please this instructor ?
- Can you suggest him a new song that he will probably like ?

First look at the data

```
read_csv("data/spotify_preferences.csv")
```

```
## # A tibble: 500 x 14
##    danceability energy   key loudness  mode spee
##           <dbl>  <dbl> <dbl>    <dbl> <dbl>
##  1        0.292  0.246     0    -9.76     1
##  2        0.466  0.657     3    -6.38     0
##  3        0.571  0.479     7    -7.73     1
##  4        0.634  0.674     5    -9.75     0
##  5        0.328  0.607     2   -11.1      1
##  6        0.646  0.665     6    -7.88     1
##  7        0.585  0.717     4   -10.3      0
##  8        0.452  0.179     9   -11.2      1
##  9        0.375  0.199     2   -10.4      1
## 10        0.374  0.943     6    -4.11     0
## # i 490 more rows
## # i 7 more variables: instrumentalness <dbl>, li
```

- **acousticness** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **danceability:** Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **duration_ms** The duration of the track in milliseconds.
- **energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

- **instrumentalness** Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **key** The key the track is in. Integers map to pitches using standard Pitch Class notation.
  `https://en.wikipedia.org/wiki/Pitch_class`
- **liveness** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

- **loudness** The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- **mode** Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **speechiness** Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like

- **tempo** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **time_signature** An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **valence** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Outline

Wage disparities across gender, race, work experience, etc.

Explore, among other things, how hourly wages differ among men and women with similar observed characteristics.

- In the United States, despite the efforts of equality proponents, income inequality persists among races and ethnicities.
- Asian Americans have the highest average income, followed by white Americans, Latino Americans, African Americans, and Native Americans.
- A variety of explanations for these differences have been proposed such as differing access to education, two parent home family structure (70% of African American children are born out of wedlock), high school dropout rates and experience of discrimination
- The topic is highly controversial

The gender pay gap or gender wage gap is the average difference between the remuneration for men and women who are working.

- Women are generally paid less than men.
- There are two distinct numbers regarding the pay gap: unadjusted versus adjusted pay gap.
  - raw difference
  - takes into account differences in hours worked, occupations chosen, education and job experience
- Ex: someone who takes time off (e.g. maternity leave) will likely not earn as much as someone who does not take time off from work
- In the United States, for example the unadjusted average female's annual salary has commonly been cited as being 78% of the average male salary, compared to 80–98% for the adjusted average salary

The latter takes into account differences in hours worked, occupations chosen, education and job experience. For example, someone who takes time off (e.g. maternity leave) will

First look at the data

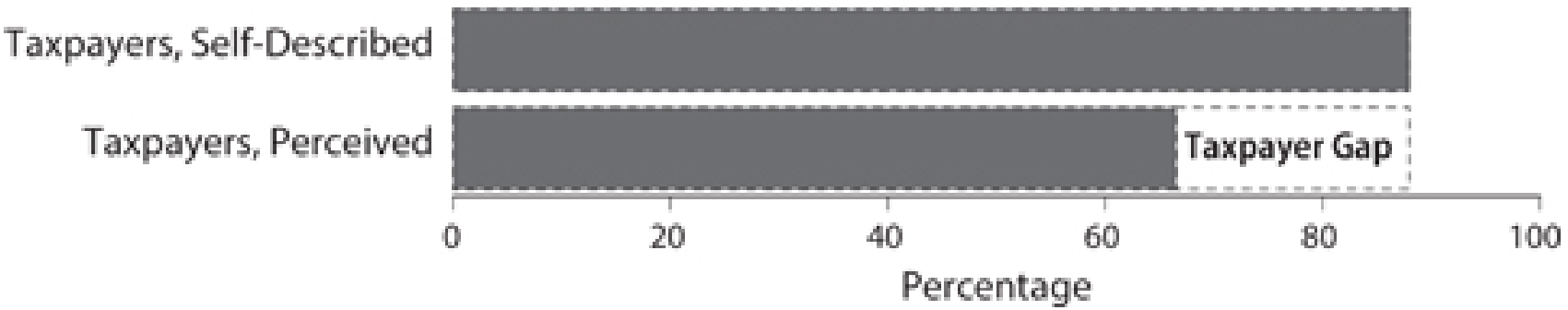
```
read_csv('data/wage_data.csv')
```

```
## # A tibble: 534 x 11
##    education south female workexp unionmember wa
##        <dbl> <dbl>  <dbl>   <dbl>       <dbl> <d
##  1         8     0      1      21           0   5
##  2         9     0      1      42           0   4
##  3        12     0      0       1           0   6
##  4        12     0      0       4           0   4
##  5        12     0      0      17           0   7
##  6        13     0      0       9           1  13
##  7        10     1      0      27           0   4
##  8        12     0      0       9           0  19
##  9        16     0      0      11           0  13
## 10        12     0      0       9           0   8
## # i 524 more rows
## # i 3 more variables: sector <chr>, married <dbl>
```

Outline

V. Williamson "Read My Lips: Why Americans are Proud to pay Taxes." Princeton University Press, 2017.

Nearly every American adult pays some taxes, whether at the local, state or federal level (think V.A.T.). Yet:



- People are angry at supposed non-taxpayers !
- Are people aware of taxes ?
  - income taxes
  - V.A.T.
  - property tax

First look at the data

```
read_csv('data/read_my_lips/Q14 survey_for datavers
```

```
## # A tibble: 1,000 x 42
##    statename  gender  educ partyid firstthought
##    <chr>       <dbl> <dbl>   <dbl> <chr>
##  1 Wisconsin       2     2       4 Paying mone~
##  2 Kansas          2     5       4 fees that y~
##  3 Ohio            2     2       2 An expense ~
##  4 Texas           1     5       5 Internal re~
##  5 New York        2     2      NA money that ~
##  6 California      1     6       7 Government ~
##  7 North Car~      2     2       1 i think tax~
##  8 Massachus~      2     4       5 money out o~
##  9 New York        1     3       4 that I will~
## 10 Texas           2     2       7 Paying out ~
## # i 990 more rows
## # i 33 more variables: biggest <chr>, wagegal <d
```

- **gender** What gender do you identify with? Select the one option that you most strongly identify with.
- **educ** What is the highest level of education that you have completed? Elementary, middle, or junior high school (1), High school (2), etc.
- **firstthought** When you hear the word "taxes," what comes to mind? Several open-ended questions, need some language processing
- **taxpayer** Are you a taxpayer?
- **percenttp** What percentage of adults in the United States do you think are taxpayers?

Who pays taxes: taxpayer perception

- **taxpayer** Are you a taxpayer?
- **percenttp** What percentage of adults in the United States do you think are taxpayers?

These two variables tell us: 90 % of the population pay taxes, many people have a wrong perception of this, **taxpayer gap**.

- Who makes the largest error ?
- Is it related to political affiliation, to income, to watching the news, etc. ?

Is the money used wisely?

- **wastecent** How many cents out of every tax dollar do you think the government wastes?
- **wastethink** When you were thinking of government waste, what specifically came to mind?

These two variables tell us: do people think tax money is spent wisely ?

- What determines this amount ?
- Is it related to political affiliation, to income, to watching the news, actually paying taxes, taxpayer gap, etc. ?

Outline

Mayfly sensitivity to salt

Perils of salinity increase for freshwater species:

- Road salts used to avoid ice during winter wash out in rivers after rain.
- Mayfly larvae are particularly sensitive species
- We could be interested in assessing if changing the type of salt may be less damaging to Mayfly larvae

Bioassays

Assessment of a substance toxicity to a species is usually done performing a bioassay:

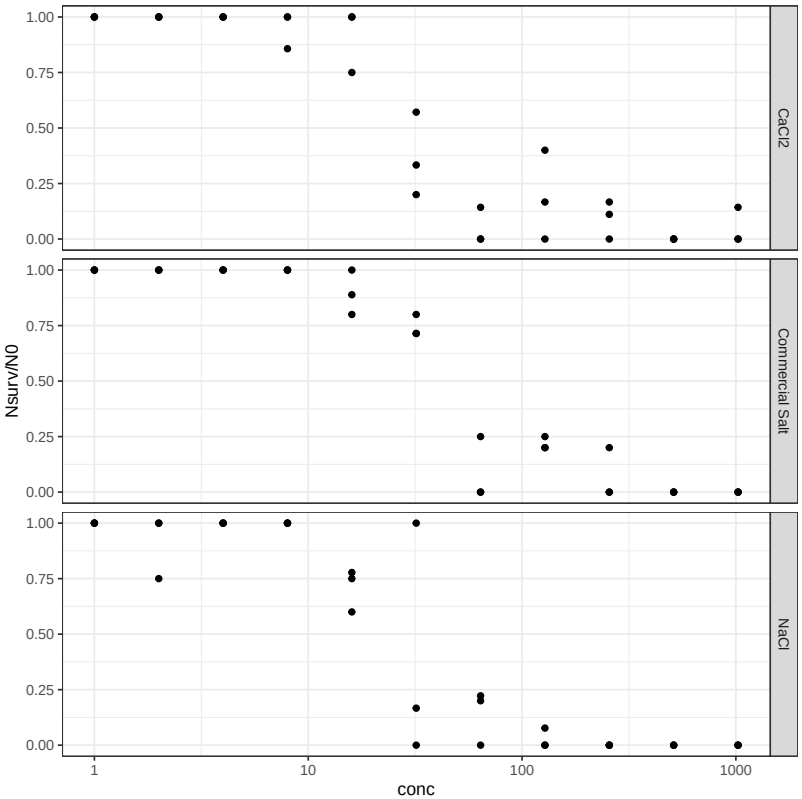
- Collect some mayflies in the wild
- Distribute them in several water tanks
- Pour a controlled quantity of salt in each tank, so that the concentration is known
- Measure survival after a fixed period (say 96h)

First look at the data

```
read_csv('data/mayflies_salt_survival.csv')
```

```
## # A tibble: 99 x 4
##       N0 Nsurv  conc Salt
##    <dbl> <dbl> <dbl> <chr>
##  1     9     9     1 NaCl
##  2     4     4     2 NaCl
##  3     5     5     4 NaCl
##  4     6     6     8 NaCl
##  5     9     7    16 NaCl
##  6     4     4    32 NaCl
##  7     9     2    64 NaCl
##  8    10     0   128 NaCl
##  9     7     0   256 NaCl
## 10     7     0   512 NaCl
## # i 89 more rows
```

First look at the data



Log-logistic model:

$$f(conc) = \frac{d - c}{1 + \left(\frac{conc}{e}\right)^b} + c$$

- d = survival probability at 0 conc
- c = survival probability at ∞ conc
- b = related to slope
- e = inflexion point

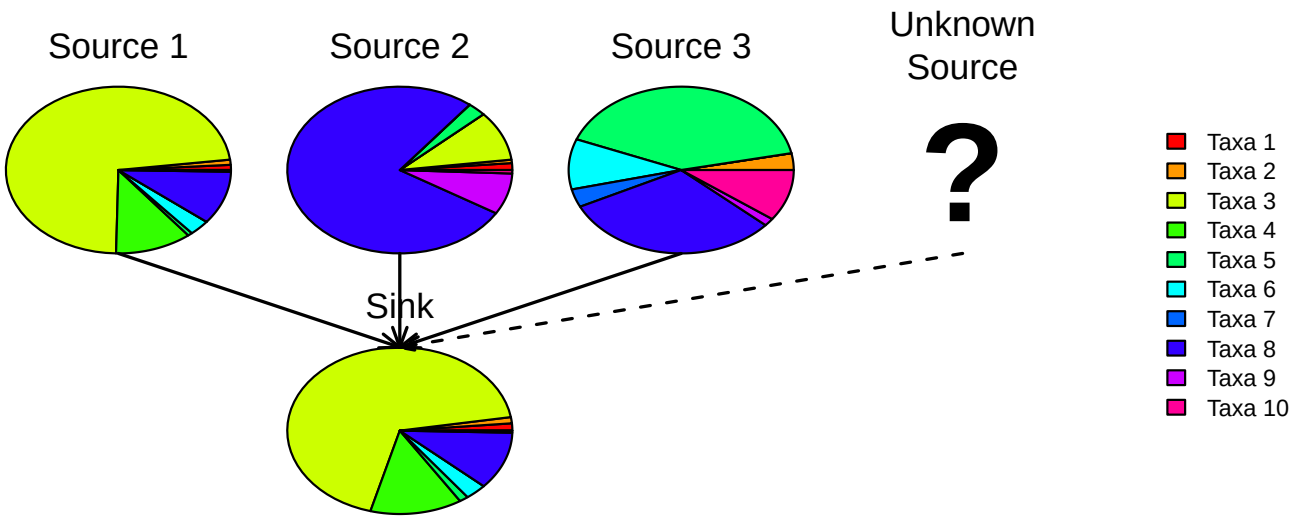
So which salt is more toxic ? Are all parameters necessary ?

Pay attention to log-scale x axis !

Outline

Microbial source tracking

Aim: investigate contamination in settings such as office buildings, hospitals and research laboratories, from several possible sources.



Sink proportions look like those of Source 1 $\Longrightarrow$ most probable source of contamination is Source 1

Dataset

The data is available at `https://github.com/cozygene/FEAST/tree/FEAST_beta/Data_files`.

Table: Metadata for the samples used in the study.

SampleID	Environment	SourceSink
ERR525698	Infant gut 1	Sink
ERR525693	Infant gut 2	Source
ERR525688	Adult gut 1	Source
ERR525699	Adult gut 2	Source
ERR525909	Adult skin 1	Source
ERR525908	Adult skin 2	Source
ERR525954	Soil 1	Source
ERR525949	Soil 2	Source

Taxa (bacteria) from any of the sources could have contaminated the sink.

Dataset

Table: Taxon counts across different samples.

Taxon					
	ERR525698	ERR525693	ERR525688	ERR525699	ERR
taxa_1	0	15	0	4	
taxa_2	5	5	13	5	
taxa_3	0	0	200	0	
taxa_4	20	0	0	20	
taxa_5	0	0	0	0	
taxa_6	0	0	0	0	
taxa_7	0	0	0	0	
taxa_8	0	0	0	0	
taxa_9	0	0	0	0	

Each sample contains certain proportions of different taxa. The sink (ERR525698) contains taxa whose proportions inform about which source is the most likely.

A model [1] for K known sources, one unknown source and one sink is:

$$\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij}$$

$$\boldsymbol{y}_i \sim \text{Multinomial}\left(C_i, (\gamma_{i1}, \ldots, \gamma_{iN})\right)$$

$$\boldsymbol{x} \sim \text{Multinomial}(C, (\beta_1, \ldots, \beta_N))$$

where source i contains taxa j with a relative proportion γ_{ij}, and the sink contains taxa j with a relative proportion β_j.

[1] Based on Shenhav, L., Thompson, M., Joseph, T.A. et al. FEAST: fast expectation-maximization for microbial source tracking. Nat Methods 16, 627–632 (2019). https://doi.org/10.1038/s41592-019-0431-x

Detailed explanations

- C_i taxa have been sampled from source i, each taxa has a probability γ_{ij} to be sampled so the observed data y_i is a **multinomial distribution** with C_i trials and probabilities $(\gamma_{i1}, \ldots, \gamma_{iN})$.
- The same holds for the sink, with C taxa sampled and probabilities $(\beta_1, \ldots, \beta_N)$.
- No taxa have been sampled from the unknown source ($C_{K+1} = 0$), which can explain why the sink may differ from the known sources.
- Now if there is a contamination, the sink taxa are a mixture of the source taxa, with α_i the proportion of source i in the sink. This translates directly to $\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij}$
- Consequence: $\alpha_i \approx 0 \implies$ no visible contamination from source i.

Inference goal

- The goal is to estimate the α_i to identify where the contamination came from.
- The γ_{ij} are estimated with some degree of uncertainty from the y_{ij}, you must take into account this uncertainty in the estimation of α_i.
- Therefore you want to obtain the posterior distribution of α_i given:
  - the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ and $\boldsymbol{x}$
  - reasonable prior distributions on the unknown parameters

Advice:

- α, γ_i are probability vectors, they live on the **simplex**, you will want to use a prior distribution that respects this constraint.
- Think about whether a given dataset $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K, \boldsymbol{x}$ can allow you to estimate the α_i with enough precision (*practical identifiability*).
- Skim through the Shenhav et al. paper for a deeper understanding[2].
- If you decide to tackle the multi sink version of the data, think about what is independent and could be analysed in parallel.

[2] but they do not use Bayesian inference, so you cannot use their method directly, they do not perform uncertainty quantification.

Final comments

For guidance, before starting please give a look to "Regression and Other Stories, Gelman, Hill, Vehtari (2020)", available at `https://avehtari.github.io/ROS-Examples/`

Your report must include:

- A brief introduction to the problem and data
- A description of your models, with discussion of the priors
- Implementation in Stan/Rstanarm/brms/JAGS, with reproducible code in attachment
- MCMC convergence checks
- Fake data check (read section 16.6 of “Regression and Other Stories”)
- Evaluation/validation/performance for each model
- Conclusions
- References and sources used

Advice !

- Finish the practical exercise available at: `https://sites.google.com/site/guillaumekonkamking/courses` to make sure you know how to use probabilistic programming software such as JAGS, Stan, etc. At the end you have regression examples
- JAGS examples: `https://github.com/andrewcparnell/jags_examples/tree/master`
- Stan examples: `https://github.com/stan-dev/example-models/wiki`
- Split the work: one should read "Regression and Other Stories, Gelman, Hill, Vehtari (2020)" and decide on the model, one should simulate fake data to test the code, one should program the MCMC algorithm and prepare the reproducible code, one should think about model checking, etc.

Final comments

Declaration of generative AI in scientific writing[3]:

"Where authors use generative artificial intelligence (AI) and AI-assisted technologies in the writing process, authors should only use these technologies to improve readability and language. Applying the technology should be done with human oversight and control, and authors should carefully review and edit the result, as AI can generate authoritative-sounding output that can be incorrect, incomplete or biased. AI and AI-assisted technologies should not be listed as an author or co-author, or be cited as an author. Authorship implies responsibilities and tasks that can only be attributed to and performed by humans."

[3] from: `https://www.sciencedirect.com/journal/stochastic-processes-and-their-applications/publish/guide-for-authors`

Final comments

Need guidance ?

- You can come to ask me (*mail me first*)
- A. G. Gelman and J. Hill, Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, 2007.
- Chapter 9 of P. D. Hoff, A first course in Bayesian statistical methods, vol. 580. Springer, 2009.
- A question about a specific Bayesian issue? Check *Bayesian Data Analysis*, Gelman, Carlin, Stern, Dunson, Vehtari and Rubin.
  `http://www.stat.columbia.edu/~gelman/book/`

Thanks for your attention !

contact: guillaume.kon-kam-king@inrae.fr

References I

- A. Gelman and C. R. Shalizi, “Philosophy and the practice of Bayesian statistics,” Br. J. Math. Stat. Psychol., vol. 66, no. 1, pp. 8–38, 2013.
- P. D. Hoff, A first course in Bayesian statistical methods, vol. 580. Springer, 2009.

Image credits: I

- https://en.wikipedia.org/wiki/Monte_Carlo_integration
- P. D. Hoff, A first course in Bayesian statistical methods, vol. 580. Springer, 2009.
- A. G. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Bayesian Data Analysis, Third edit. CRC press, 2014.