simulation-notes

Tony Ni

10/5/2020

Feedback (on generating-missing-data.Rmd)

- want to preserve the truth, dont want to overwrite value column
- BUILD UP dataset instead of making new one each time so we can compare in the end
- want to repeat this process multiple times when i set up simulation study (do this at a larger scale)
- read up on how to do simulation studies (prof baileys github) here: https://github.com/bebailey/research-tutorials/tree/master/04_coding-simulation
- what do we the structure to be like in the end?
- THINKING OF EXPLCIIT NAMING SCHEMES INSTEAD OF JUST MY_DF# AND 'value' (think of what the names should be)
- when we are generating data, the ideal simulation set up will be setting up a df, and then that df should be able to be put into the functions w/o any changes
- rerun these functions MULTIPLE times repeatedly with new datasets (does it do better if we assume lognormal dist vs. normal dist? how closely do these methods get to the actual statistics?)
- create a function to apply all of these functions to the generated data
- FOR NEXT WEEK: focus on understanding what a simulation study IS – put on pause the codework
- read the morris paper on simulation studies and then thinking about what it is i want to show with this simulation study (there is the big picture: comparison of how these methods perform vs the truth BUT what are we making these comparisons based off of, how do we know substitution method is better/worse etc. HOW DO WE COMPARE THE ACTUAL DISTRIBUTION OF VALUES – usually just summary statistics… we might look at overall mean and median of values or summary statistics based off on whether or not if they are LOD values – how does avg values compare btwn methods)
- after getting better sense of larger simulation study might look like – will have to go back into code (go back into lnorm and making sure that what you're inputting is what the function is asking for, the parameters that r takes in is not always what wikipedia takes in – verify parameters)
- verifying what the functions are outputting, can you get the rawdata, if so how? if not, are there anyways where we can?

ADEMPS

Aims

- Q: what are the aims of the study?
- A: (1) investigate effectiveness (accuracy?) of methods designed to estimate summary statistics (mean and sd) of left-censored data

Data-Generating Mechanisms

- Q: resampling vs. simulation from parametric dist?
- Q: how simple/complex should the model be?
- Q: should it be based on real data?
- Q: which factors to vary? which levels of factors to use?
- A: resampling from a simple lognormal distribution (not based on real data, for ease of calculations/comprehension)
- A: data generated from $n_{obs} = 1000$ wells that represent a possible collection of chemical concentration measurements (unrealistic, but simple – i personally think drawing it from the actual data itself might be not as useful? we could also vary the sample size to see how the methods perform when different % of the data is censored (refer to Bolks2014 – "different methods work better depending on small/large values of n and small/large percent of data censored))
- A: values talked about in paper are $n < 50$ compared to $n \geq 50$ and $< 50\%$ data censored compared to $50 - 80\%$ censored ($> 80\%$ is too high to compute summary statistics)
- A: let X_i be the concentration of an arbitrary chemical at a well with $X_i \in \text{Lognormal}(\mu, \sigma^2)$
- A: factors we might consider varying is the sample size n and the percent of data censored
- A: $\mu = 1$ and $\sigma^2 = 1$

Estimand

- Q: define estimands of the simulation study
- A: the estimands (variable which is to be estimated in a statistical analysis) of our study are the mean, and variance from which we generated our lognormal distribution form
- A: we want to observe the mean estimator $\bar{X}$, which is the estimator of μ
- A: we want to observe the variance estimator s^2, which is the estimator of σ^2

Methods

- Q: identify methods to be evaluated; are they appropriate?
- A: each simulated dataset(s) (with left censored data) is to be used in the following 5 methods (substitution, MLE, kaplan-meier. imputation with MLE, imputation with kaplan-meier) in order to compute the estimators s^2 and $\bar{X}$

Performance Measures

- Q: list all performance measures to be estimated
- Q: talk about how relevant they are to the estimands?
- Q: choose value of n_sim which achieves acceptable Monte Carlo SE (?)
- A: we will assess the following two performance measures for each method...
- A: root mean square error (RMSE): measures the difference in sample and population values predicted by an estiamtor/model, RMSE $= \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n}}$ where $\hat{y_i}$
- A: bias which is defined as: Bias $= E[\hat{\theta}] - \theta$ where θ is the estimator and $\hat{\theta}$ is the estimated value