Biostatistics

Mithat Gonen
Charlie White

Missing Data

- A thorn on our side, really
- A reality for almost all studies
- And a bigger threat to the validity of our conclusions than we often realize

What do we mean by missing data?

- Think of your data in an excel sheet, rows are patients and columns are variables
- Any cell in the sheet that is not filled is missing data
- It is a data point that you should have had, but you don’t

Why worry?

- Missingness might represent selection
  - Low-risk patients are less likely to be tested, scanned etc.
- If that is the case patients with missing data do not represent a random subset and excluding them from the analysis results in a bias
- So, it is all about deciding whether patients with missing data are similar to those patients with no missing data; or are they different in a systematic way?

Types of Missing Data

- Missing by Design
- Missing Completely at Random (MCAR)
  - No rhyme or reason that a given patient has missing data
- Missing at Random (MAR)
  - Certain patients are more likely to have missing data but that can be explained by the values other variables take
- Missing Not At Random (MNAR)
  - Certain patients are more likely to have missing data but that cannot be explained by the values other variables take

Missing by Design

- Sometimes called missing by definition
- One can only have pathologic response if one had neoadjuvant therapy, so in a data set of all gastric cancer patients the column Path_Response will be missing unless the patient had a BMT.
- Nothing to do here, other than being aware of this and realizing that pathological respobse can only be used for the subset of patients who had neoadjuvant treatment
- GVHD and BMT

Types of Missing Data

- Missing by Design
- Missing Completely at Random (MCAR)
  - No rhyme or reason that a given patient has missing data
- Missing at Random (MAR)
  - Certain patients are more likely to have missing data but that can be explained by the values other variables take
- Missing Not At Random (MNAR)
  - Certain patients are more likely to have missing data but we do not have the covariates in the dataset to explain this

Missing Completely at Random (MCAR)

- Some cells in your data set are missing but that it is missing has nothing to do with any of the other variables
- Example CEA vs Age
- Mean Age in those with CEA Available 64.0 vs those with CEA Missing 64.1
  - 95% Confidence Interval: -3.8 to 3.7
  - P = 0.97
- Availability of CEA has nothing to do with Age

Missing Completely at Random (MCAR)

- Some cells in your data set are missing but that it is missing has nothing to do with any of the other variables
- Example CEA availability by T stage
- P = 0.09
- !!!

T Stage	I	II	III	IV
CEA Present	152	180	590	105
CEA Missing	7 (4.4%)	10 (5.2%)	40 (6.4%)	11 (9.5%)

MCAR

- To declare MCAR
  - Take the variable with missing values (e.g. CEA)
  - Analyze missing vs not as a binary outcome against all the other variables you want to include in the analysis
  - If you are convinced that there is no difference between missing vs not with respect to any of the variables, then you can conclude MCAR for CEA
    - Do not rely only on significance
  - Now repeat for all the other variables with missing data
  - If all variables with missing data are MCAR then you can ”assume” MCAR for the analysis, you are about to conduct

Why did I say assume?

- MCAR, MAR, MNAR are all assumptions, but they are (only somewhat) testable
- Testing these assumptions can be tricky, especially with reliance on p-values
  - Small data set → Too little power → Everything looks like MCAR
  - Large data set → Too much power → Nothing looks like MCAR

How do you deal with MCAR

- You can exclude patients with missing data
- No worry for bias
- But you will have a loss of power
- Side Note: Ignoring the missing data problem is equivalent to assuming MCAR → you should never ignore and instead test for it
  - Exception: small amounts of missing data (~5-10%)

What if you conclude it is not MCAR?

- Then you have a set of variables that are associated with missingness
- If you are willing to assume that, among variables not included in your data set, none are associated with missingness then it is MAR
- Which means you know all the variables that “define” when a variable will be missing

How to Deal with MAR

- ”Predict” the missing value from the other covariates
- Which means, set up a regression model for CEA using only the patients who have CEA values
- Then use this regression model to predict the missing value
- Impute this predicted value in place of the missing value
- Do this for all variables with missing values
- You have a complete data set, analyze as such
- Called single imputation and should never be used

Multiple Imputation

- You just filled in data and acted as if you observed it
- To be fair you need to penalize yourself a little bit by recognizing not all data in your imputed are equal
  - Some are observed, some are imputed
- There is a method called multiple imputation that properly recognizes the imputed values and makes you pay a price (standard errors are larger, CI's are wider, p-values are larger compared with single imputation)

What if MNAR?

- Very difficult problem, requires strong untestable assumptions
- Seriously consider abandoning the analysis for not having the appropriate data set
- In my 20+ years I never did an MNAR analysis

Summary of Missing Data

- If by design, no problem
- Otherwise ask if missingness might have anything to do with some factors and ask if you have all those factors in your data set
  - If yes and no, then MNAR
  - If yes and yes, then MAR → go find those factors
  - If no to first question, then MCAR but I would still recommend testing this assumption

What is Boosting?

- Many weak predictors built sequentially.
- Each model focuses on previous mistakes.
- Final prediction = weighted combination.
- Turns weak learners into a strong learner.

Why Might Physicians Care?

- Used in clinical prediction models.
- Often better than single models.
- Captures nonlinear interactions automatically.
- More interpretable than complex ML methods.

AdaBoost: Core Idea

- Assume we are predicting a continuous variable
- Start with equal patient weights
- Fit a simple predictor (eg. regression tree stump).
- Increase weight on misclassified patients, making the next model focus on them
- Repeat for many rounds — combine the models
  - Each model gets a weight based on its performance

Binary variable

- You can either focus on misclassified patients (0-1 weights, 1 for misclassified)
- Or you can use the predicted probabilities minus the 0-1 outcome as weights
- Otherwise same idea

A Conceptual Way to Think About It

- Model keeps focusing on 'hard' patients.
- Like teaching: more attention to struggling students.
- Weak learners accumulate into a strong model.

Strengths and Weaknesses

- Strengths:
  - Improved accuracy
  - Automated implementation with minimal input
- Weaknesses:
  - less interpretable (like random forests or neural nets)
  - sensitive to noise/overifts (focusing on a few hard cases makes you fit to noise)
- Popularity somewhat waned due to emergence of neural networks

Bayesian Analysis

- Traditional statistical analysis ignores context.
- Example: binary outcome in 12 patients, all events.
- Point estimate: 100%, 95% CI 74%–100%.
- Suppose you were not expecting this level of activity; should that be part of your analysis?
- Bayesian statistics formally incorporates context.

The NEW ENGLAND JOURNAL of MEDICINE

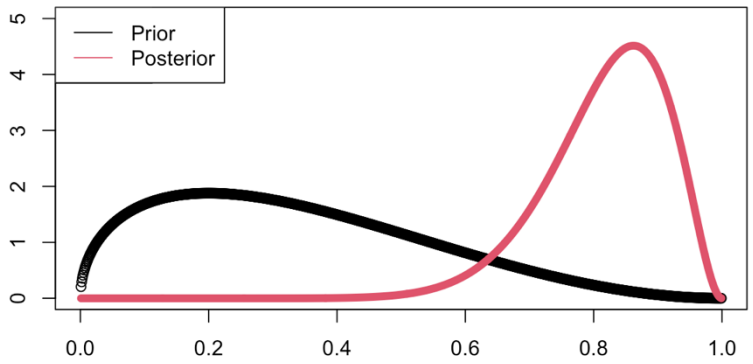
ESTABLISHED IN 1812 JUNE 23, 2022 VOL. 386 NO. 25

PD-1 Blockade in Mismatch Repair–Deficient, Locally Advanced Rectal Cancer

A. Cercek, M. Lumish, J. Sinopoli, J. Weiss, J. Shia, M. Lamendola-Essel, I.H. El Dika, N. Segal, M. Shcherba, R. Sugarman, Z. Stadler, R. Yaeger, J.J. Smith, B. Rousseau, G. Argiles, M. Patel, A. Desai, L.B. Saltz, M. Widmar, K. Iyer, J. Zhang, N. Gianino, C. Crane, P.B. Romesser, E.P. Pappou, P. Paty, J. Garcia-Aguilar, M. Gonen, M. Gollub, M.R. Weiser, K.A. Schalper, and L.A. Diaz, Jr.

Motivating Example

- Suppose prior belief: response rate around 30%.
- Represent prior as Beta(1.5,3) → prior mean ~ 30%.
- Observe 12/12 responses.
- Posterior = Beta(1.5+12, 3+0) = Beta(13.5,3).



Prior → Data → Posterior

- Prior distribution encodes contextual knowledge.
- Data update the prior using Bayes' Rule.
- Posterior represents updated belief after seeing data.

Posterior Distribution

- Posterior for response rate p = Beta(13.5,3).
- Posterior mean = 0.82.
- 95% credible interval easy to compute from distribution
  - 63% – 97%
- Contrast with 100% (74% - 100%)

We had more patients

- 84/103 complete responses when the trial was extended
- Classical estimates: 81% (73%-89%)
  - Note the new classical estimate is very close to the old Bayesian estimate
- New Bayesian estimate: 81% (74%-88%)
- With a lot of data classical and Bayesian results agree

All You Need is the Posterior

- Posterior mean or mode as point estimate.
- Credible interval from area under posterior curve.
- Probability statements: $P(RR > 0.5)$, $P(RR < 0.3)$, etc.
- Example:
  - $P(RR > 0.9)$ for $Beta(13.5,3) \approx 0.20$.

Easy Interpretation

- Posterior is a probability distribution.
- “Probability $p > 0.3$ is 0.9.”
- 95% credible interval: 0.32–0.78 means 95% probability p is in this range.
- Interpretation aligns with clinical intuition.

So Why is not everyone a Bayesian?

B. Efron

The American Statistician
Vol. 40, No. 1 (Feb., 1986), pp. 1-5 (5 pages)

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Why Isn't Everyone a Bayesian?

B. EFRON*

Originally a talk delivered at a conference on Bayesian statistics, this article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends

3. FISHERIAN STATISTICS

In its inferential aspects Fisherian statistics lies closer to Bayes than to NPW in one crucial way: the assumption that there is a *correct* inference in any given situation. For example, if $x_1, x_2, \ldots, x_{20}$ is a random sample from a Cauchy distribution with unknown center θ,

$$f_\theta(x_i) = \frac{1}{\pi[1 + (x_i - \theta)^2]},$$

then in the absence of prior knowledge about θ the correct

Why not indeed?

- Prior, prior, prior
- Prior makes it possible
  - To incorporate context
  - To make these probability statements
- But context might mean subjectivity
- Two people looking at the same data can come to different conclusions
- Bayesians say it happens anyway, we are just quantifying it
- It is a sharp divide in statistics, as bad as Yankees vs Red Sox

Clinical Trial Application

- Bayesian analyses handle sequential data naturally.
- Today's posterior becomes tomorrow's prior.
- Two-stage design example:
- Stage I posterior → Stage II prior.
- Enables adaptive decision making.

Summary

- Everything flows from the prior.
- If you accept the prior, you gain:
  - Intuitive interpretation
  - Easy adaptation to new data
- Sensitivity analysis and simulation essential.