Modeling Censored Immunological Data

Zane Billings, Andreas Handel

https://tinyurl.com/hg-cens

handelgroup.uga.edu

Acknowledgements

Censored data

- Values are "partially known": we know a range where the value lies, but not the exact value.

Censored data

- Values are "partially known": we know a range where the value lies, but not the exact value.
- Like missing data, but we know a bit more about the true value.

Censored data

- Values are "partially known": we know a range where the value lies, but not the exact value.

- Like missing data, but we know a bit more about the true value.
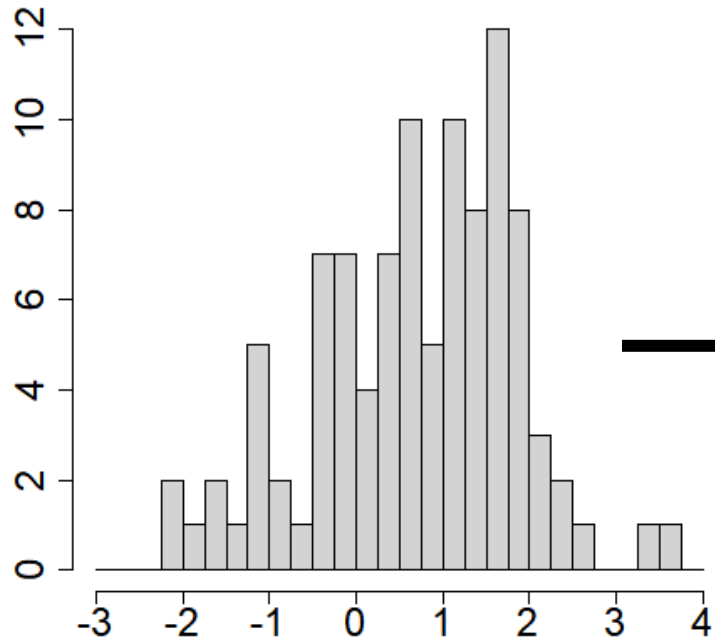
- Limits of Detection/Quantification (LoD/LoQ)

Censored data

- Values are "partially known": we know a range where the value lies, but not the exact value.
- Like missing data, but we know a bit more about the true value.
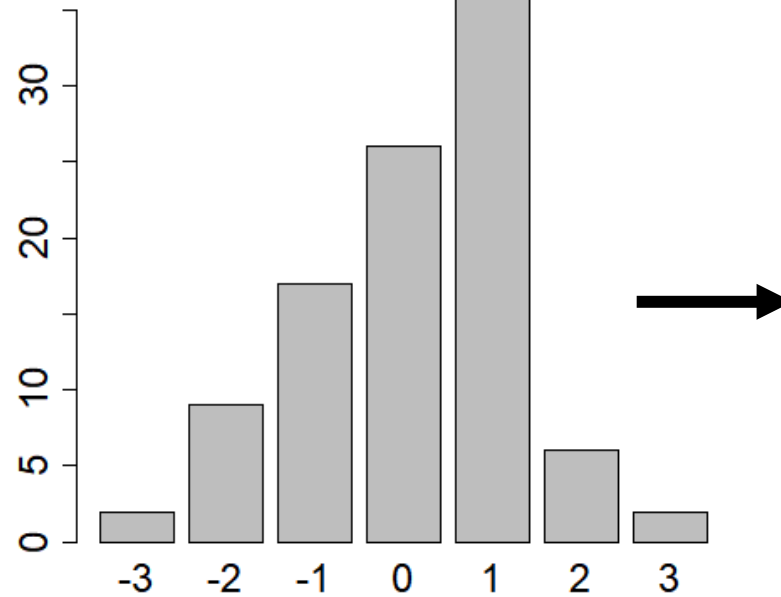- Limits of Detection/Quantification (LoD/LoQ)
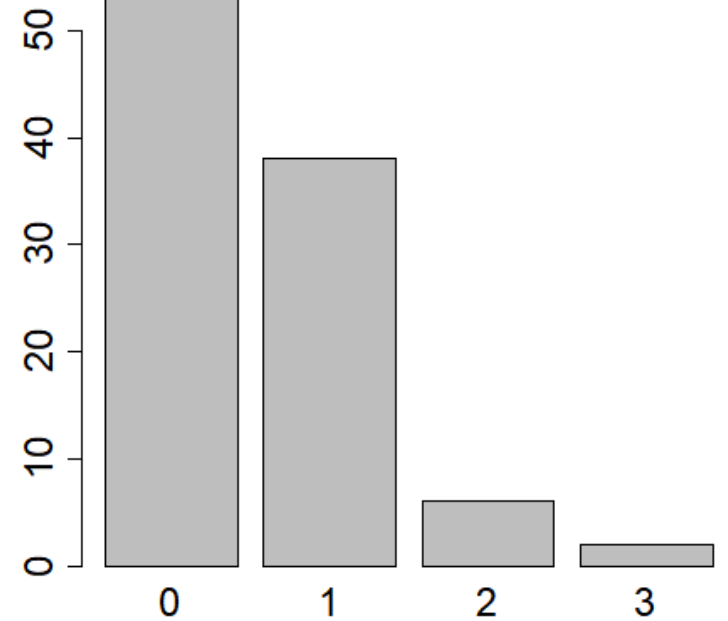- Binning/discretizing/rounding/etc. a continuous measurement

E.g., HAI titer data for flu research

Continuous "latent" titer — Floored titer (rounded down) — Floored titer with LoD

Censored data and statistics

- **Coding censored data as a constant will bias analyses! (E.g. setting HAI titer to 5.)**

Censored data and statistics

- **Coding censored data as a constant will bias analyses! (E.g. setting HAI titer to 5.)**
- Dealing with censored outcomes is (usually) not so bad. There are methods and implementations!

Censored data and statistics

- **Coding censored data as a constant will bias analyses! (E.g. setting HAI titer to 5.)**
- Dealing with censored outcomes is (usually) not so bad. There are methods and implementations!
  - Frequentist: "tobit" model, `survival` models in R
  - Bayesian: `brms` R package makes it easy

Censored data and statistics

- **Coding censored data as a constant will bias analyses! (E.g. setting HAI titer to 5.)**
- Dealing with censored outcomes is (usually) not so bad. There are methods and implementations!
  - Frequentist: "tobit" model, `survival` models in R
  - Bayesian: `brms` R package makes it easy
- Censored predictors are harder.

Problem: reading stats papers and package documentation is hard/annoying.

(Even for mathematicians/statisticians!)

(Partial) Solution: we're writing a series of tutorials with examples.

https://tinyurl.com/hg-cens

https://tinyurl.com/hg-cens

Modeling with Censored Data

AUTHOR
handelgroup

PUBLISHED
2024-07-20

Preface

This book contains our notes on dealing with censored data, including methods for dealing with both censored outcomes and predictors. Where possible, we try to include both frequentist and Bayesian models.

- The short link to this website is: https://tinyurl.com/hg-cens.
- The full link is currently: https://wzbillings.github.io/Censored-Data-Tutorials/.
- For source code, go here.
- Please submit bug reports, typos, and comments here.

Contributors:

- Zane Billings (https://wzbillings.com/)

https://tinyurl.com/hg-cens

2.3 Integration method for censored data 🔗

We can regain some measure of our lost diginity using the integration trick we discussed in the introduction. Of course, in the introduction, we only talked about adjusting for censoring in the univariate case, but fortunately we are modeling the **conditional distribution of** y so we can use the same trick:

$$\mathcal{L}(\theta \mid y_i, x_i) = f_{Y_i \mid X_i = x_i}(y_i \mid \theta, x_i)$$

$$= \left(f(y_i \mid \theta, x)\right)^{1-c_i}\left(P(Y_i = y_i \mid x)\right)^{c_i}$$

$$= \left(f(y_i \mid \theta, x_i)\right)^{1-c_i}\left(\int_{-\infty}^{y_{\min}} f(y_i \mid \theta, x_i) \, dy_i\right)^{c_i}$$

$$= \left(f(y_i \mid \theta, x_i)\right)^{1-c_i}\left(F(y_{\min} \mid \theta, x_i)\right)^{c_i}.$$

The likelihood for y_i is easy to write out here since the censoring structure is (relatively) simple. This gives rise to the likelihood of the sample, which (under the assumption of mutual conditional independence) is

$$\mathcal{L}(\theta \mid x, y) = \prod_{i=1}^{n} \mathcal{L}(\theta \mid y_i, x_i).$$

```r
# First we have to transform the outcome
surv_model <- survival::survreg(
    # Creating the "survival time" outcome
    survival::Surv(
        # If the value is lower than LoD, replace it w/ LoD, then take th
        pmax(y, LoD) |> log(),
        # The censoring indicator needs to be the opposite of what makes
        # zero for censored, one for uncensored -- it's actually an indic
        # an "event" occurring, for us this is the event
        # "getting a reliable measurement."
        !c,
        # Specify left censoring
        type = 'left'
    ) ~
        # All the other linear model stuff as usual
        dist_from_farm + personal_gly_use + water_filter_use,
    data = gly_data_obs,
    dist = "gaussian"
)
```

```r
tobit_model <- AER::tobit(
    log(y) ~ dist_from_farm + personal_gly_use + water_filter_use,
    data = gly_data_obs,
    left = log(LoD),
    right = Inf,
```

Explanation + code examples!

Conclusions and future work

- Our tutorials for censored data analysis are available free and online.

Conclusions and future work

- Our tutorials for censored data analysis are available free and online.

- They're still under construction though!

Conclusions and future work

- Our tutorials for censored data analysis are available free and online.

- They're still under construction though!

- Models for censored outcomes are more straightforward, but we're still experimenting with the best methods for censored predictors.

Thank you!
https://tinyurl.com/hg-cens