CensorCorr: A Tool for Understanding How Data Point Censoring Affects Correlations

Kimberly A. Barchard and Jin Qian
University of Nevada, Las Vegas

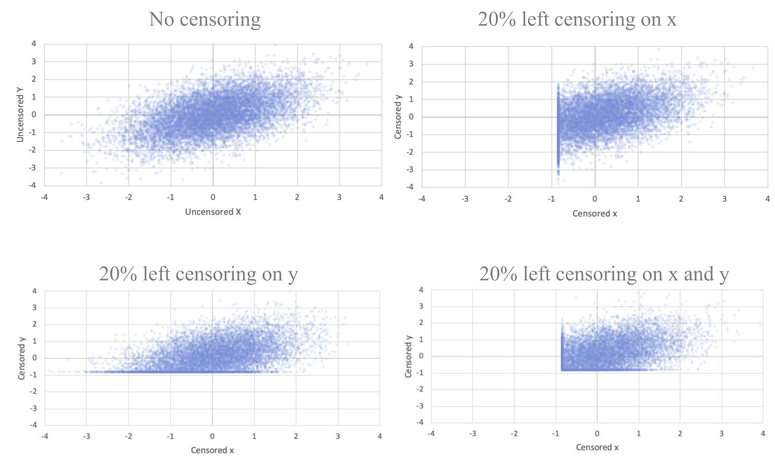
What is Data Censoring?

- Data point censoring occurs when researchers have only partial information about the value of a variable, knowing the value is at least as large as (or no larger than) a given limit of detection, leading to ceiling (or floor) effect.
  - For example, age is 55 or older, income is less than $10,000, or "I feel sad" is True.
- Censoring is common in psychology, but typically unrecognized outside of longitudinal studies.
- CensorCorr in Excel and CensorCorr in R were created to demonstrate the effect of censoring on correlations, histograms, and scatterplots.

What is CensorCorr?

- CensorCorr generates bivariate normal data for X and Y, then censors those values.
  - It allows users to specify the correlation between uncensored variables, the sample size, and the degree of left- and right-censoring for X and Y.
- Let X and Y be two variables covering the whole of the constructs of interest.
- ρ_{XY} is the population correlation between uncensored X and Y; r_{XY} is the sample correlation between X and Y.
- Let x and y be censored versions of these variables.
- ρ_{xy} is the population correlation between censored x and y; and r_{xy} is the sample correlation between censored x and y.

Effect of Censoring on Variables



How Data Censoring Effects Correlation?

The impact censoring has on correlations depends on the original correlation and the degree of censoring.

- Imagine x and y each have .3 left censoring:
  - If $\rho_{XY} = +.8$, then $\rho_{xy} = +.773$,
  - If $\rho_{XY} = -.8$, then $\rho_{xy} = -.632$.
  - Both correlations are affected, but differentially.
- Imagine x has .3 left censoring and y has no censoring:
  - If $\rho_{XY} = +.8$, $\rho_{xy} = +.745$,
  - If $\rho_{XY} = -.8$, $\rho_{xy} = -.745$.
  - Both correlations are affected the same.
  - For more information on how data censoring effects correlation see Barchard and Russell (in press) and Barchard (2024; this session).

UNLV

CensorCorr in Excel

- CensorCorr in Excel (Barchard, 2023) is a 4-tab Excel program that allows the users to visualize the effect of censoring on correlations between bivariate normal variables with pre-determined sample sizes (n = 10,000 and 500,000 versions are available).
  - Instruction tab
    - Explains censoring and how to use CensorCorr
  - Input and Output tab
    - Where users specifies inputs
    - Where output appears
  - Behind the scene tab
    - This tab contains bivariate normally distributed X and Y and the censoring process.
  - License tab

Input and Output

- Input
  - Correlation between uncensored X and Y, ρ_{XY}
  - Left and right censoring on X
  - Left and right censoring on Y
- Output
  - Sample correlation between uncensored X and Y, r_{XY}
  - Sample correlation between censored x and y, r_{xy}
  - The mean and standard deviation of X, Y, x, and y
  - Scatterplot of X and Y; Scatterplot of x and y
  - Histograms of X, Y, x, and y

***Input and Output* tab**

Use CensorCorr to

- Teach your students about ceiling and floor effects
- Teach your students about data point censoring
- Learn to recognize data point censoring in your own datasets and in others'
- Recreate the histograms and scatterplots you see in your observed data (x and y) to estimate the correlation between uncensored variables (X and Y).

CensorCorr in R

- CensorCorr in R (Barchard & Qian, 2023) is a script that allows the users to visualize the effect of censoring on correlations between bivariate normal variables with user-specified sample sizes.
- The script contains 5 parts:
  - User specifications: sample size, correlation, censoring
  - Generate data using MASS package (Venables & Ripley, 2002)
  - Censor the two variables
  - Calculate correlations
  - Create scatterplots and histograms

Input and Output

- Input
  - Sample size
  - Population correlation between uncensored X and Y, ρ_{XY}
  - Left and right censoring on X
  - Left and right censoring on Y
- Output
  - Sample correlation between uncensored X and Y, r_{XY}
  - Sample correlation between censored x and y, r_{xy}
  - The mean and standard deviation of X, Y, x, and y
  - Scatterplot of X and Y; Scatterplot of x and y
  - Histograms of X, Y, x, and y

***User Specifications* section**

User specifies

```
# Sample size.
n <- 1000  # Change 1000 to your desired sample size

# Pearson product-moment correlation between the uncensored variables X and Y
# Must be between -1 and 1.
rhoXY <- -.7  # Change -.7 to your desired correlation.

# The proportion of left censoring for the variable X.
# Must be between 0 and 1.
#The sum of left and right censoring for X cannot be 1 or more.
# Change .1 to the desired degree of left censoring on x.
x_left_censor <- .1

# The proportion of right censoring for the variable X.
# Must be between 0 and 1.
#The sum of left and right censoring for X cannot be 1 or more.
# Change .0 to the desired degree of right censoring on x.
x_right_censor <- .0

# The proportion of left censoring for the variable Y
# Must be between 0 and 1.
#The sum of left and right censoring for Y cannot be 1 or more.
# Change .2 to the desired degree of left censoring on y.
y_left_censor <- .2

# The proportion of right censoring for the variable Y
# Must be between 0 and 1.
# The sum of left and right censoring for Y cannot be 1 or more.
# Change .0 to the desired degree of right censoring on y.
y_right_censor <- .0

# Specifying which graphs should be generated
# If want a scatter plot or histogram, set to TRUE. Otherwise, set to FALSE.
graph_choices <- list(XY_scatter=TRUE,  # Scatter plot for uncensored X and Y
                      xy_scatter=TRUE,  # Scatter plot for censored x and y
                      XY_hist=TRUE,     # Histograms for uncensored X and Y
                      xy_hist=TRUE)     # Histograms for censored x and y

# Do you want to save the graphs on your computer?
save_graphs <- TRUE
```

References

- Barchard. K. A. (2023). *CensorCorr Version 1.28*. [Excel file] https://osf.io/pfqy2/
- Barchard, K. A. (2023). *CensorCorr in R Version 15.* [R script] https://osf.io/pfqy2/
- Barchard, K. A. (2024). *Income < $10,000, Age = 55+, I am sad = True: The effect of censored data on correlations*. [Poster]. Western Psychological Association conference, San Francisco, CA.
- Barchard, K. A., & Russell, J. A. (in press). Distorted correlations among censored data : Causes, effects, and correction. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02086-5
- Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th edition). Springer. https://CRAN.R-project.org/package=MASS