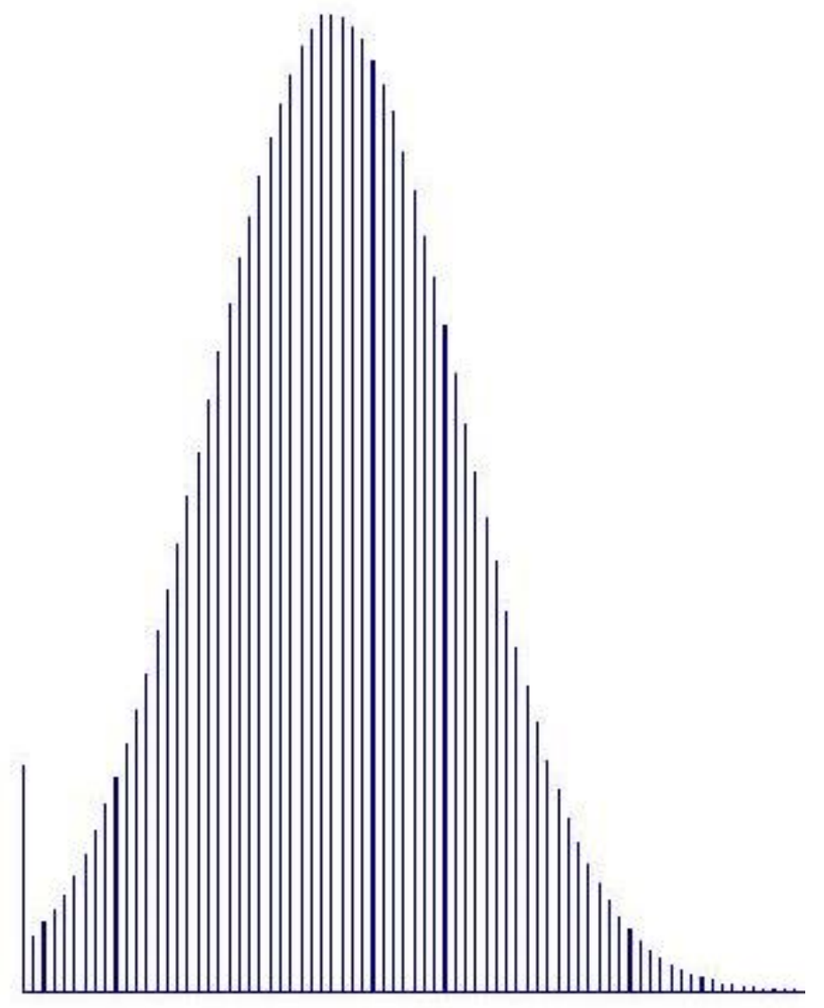
Income < $10,000, Age = 55+, I am sad = True: The Effect of Censored Data on Correlations

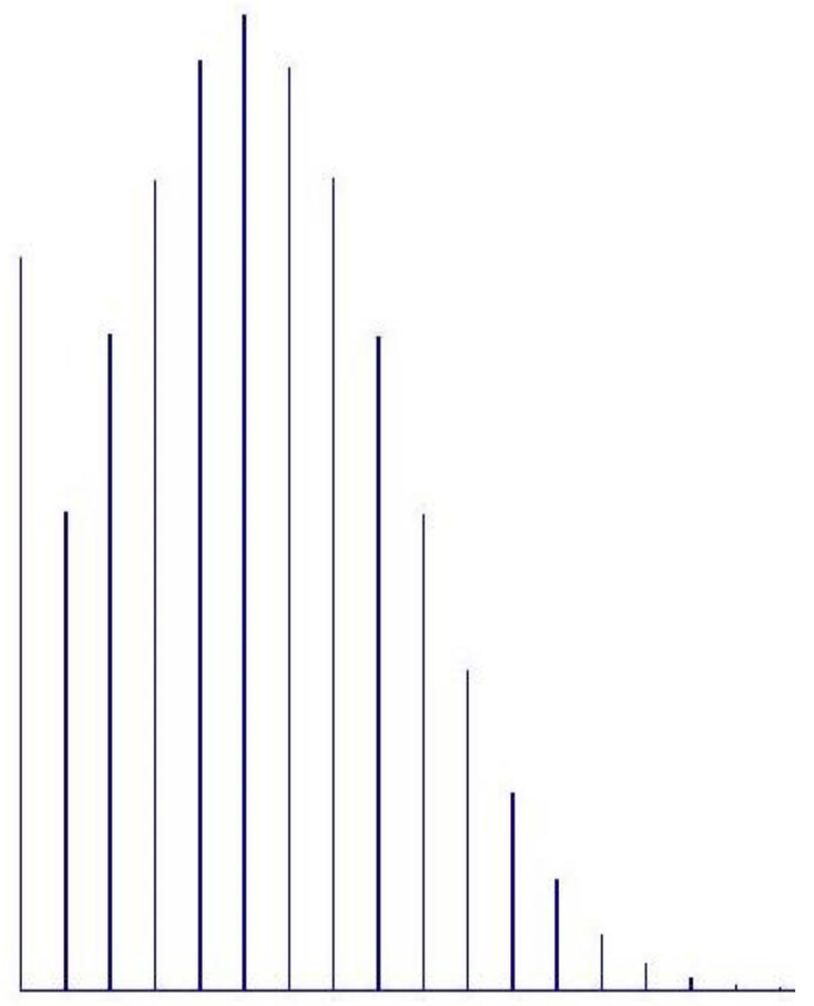
UNLV
Kimberly A. Barchard
University of Nevada, Las Vegas

INTRODCUTION

Data are censored when they provide only partial information about values of variables, indicating values are at least as large as (or no larger than) limits of detection.

LEFT CENSORING: When the left-hand tails of distributions (small values) are obscured.





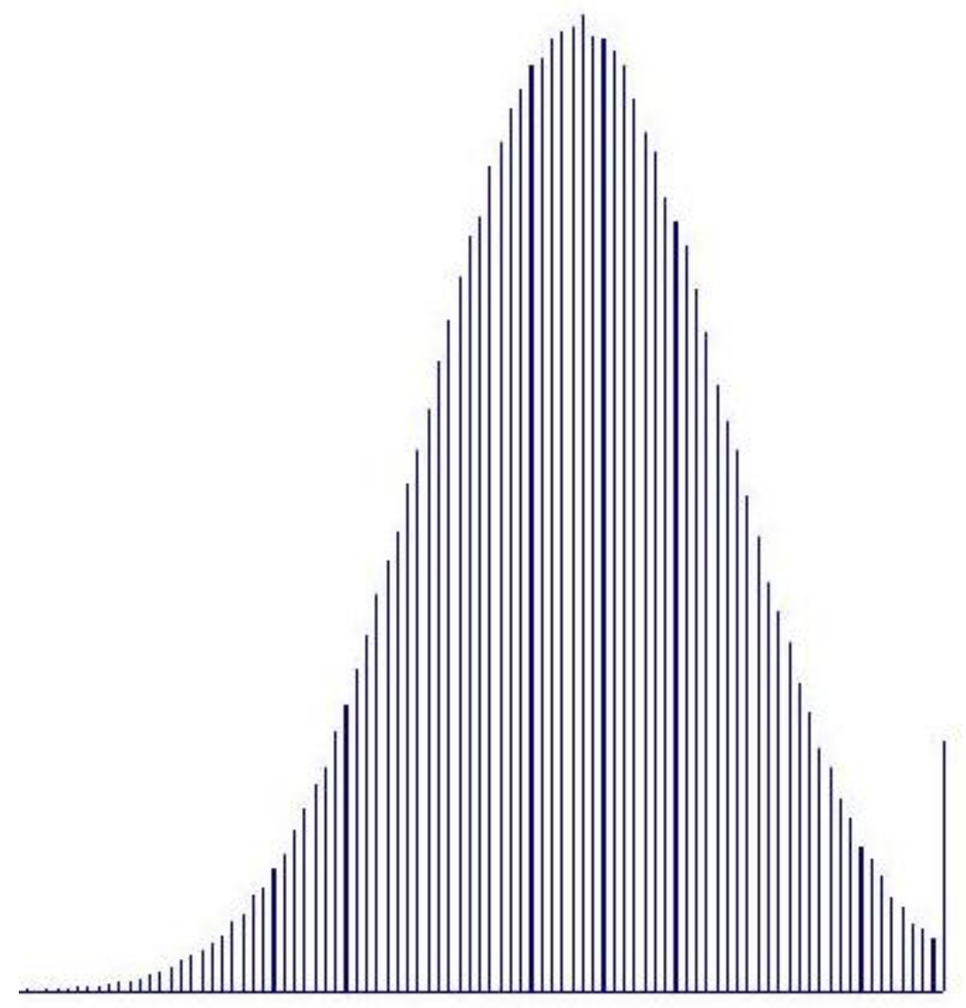
Income:
< $10,000
$10,000 - $19,999
$20,000 – $29,999
Etc.

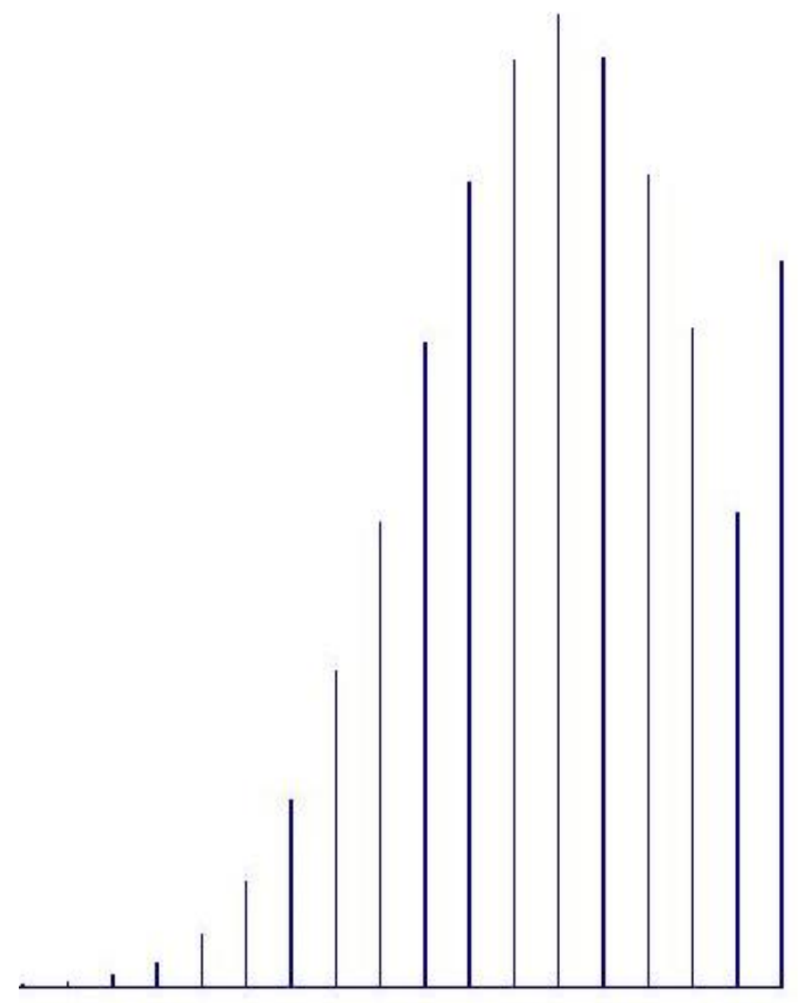
I enjoy wild parties:
Disagree
Neutral
Agree

Limit of detection
0.50 ng/dl

Difficult items

RIGHT CENSORING: When the right-hand tails of distributions (large values) are obscured.





I feel sad:
True
False

Age:
< 18
19
20
21
.
.
> 55

How long have you been married?
1 year
2 years
3 years
.
.
20+ years

Easy items

METHOD

R package mvtnorm (function rmvnorm) generated bivariate normal data for 10,000 cases, where ρ_{XY} is the population correlation between X and Y, r_{XY} is the sample correlation, and r_{XY} is within .005 of ρ_{XY}.

R censored the data at the desired percentiles. For example, if x has 30% left censoring, $x = X$ if X is greater than its 30th percentile, otherwise x = the 30th percentile of X.

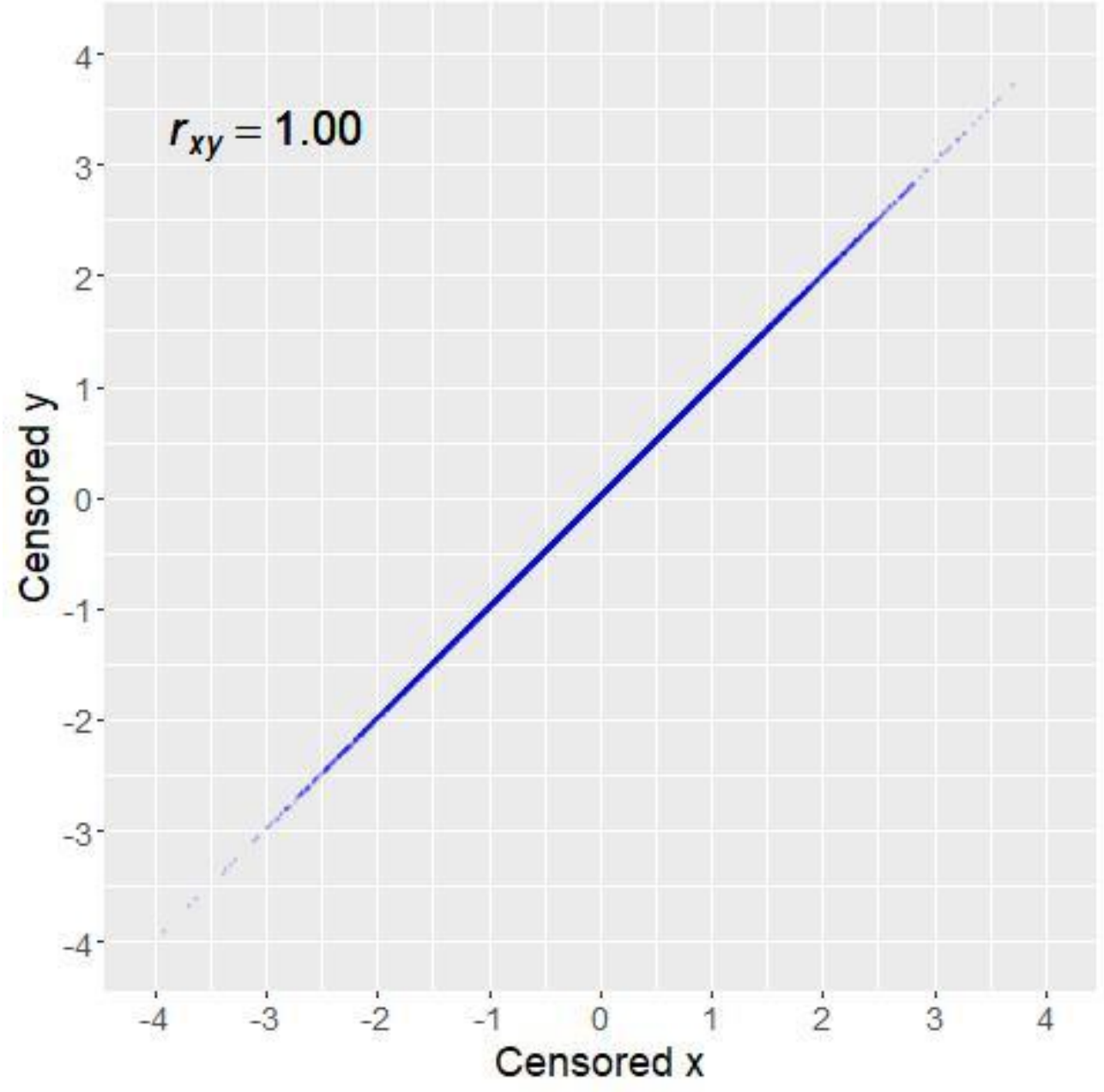
R calculated the correlation between x and y, r_{xy}.

RESULTS

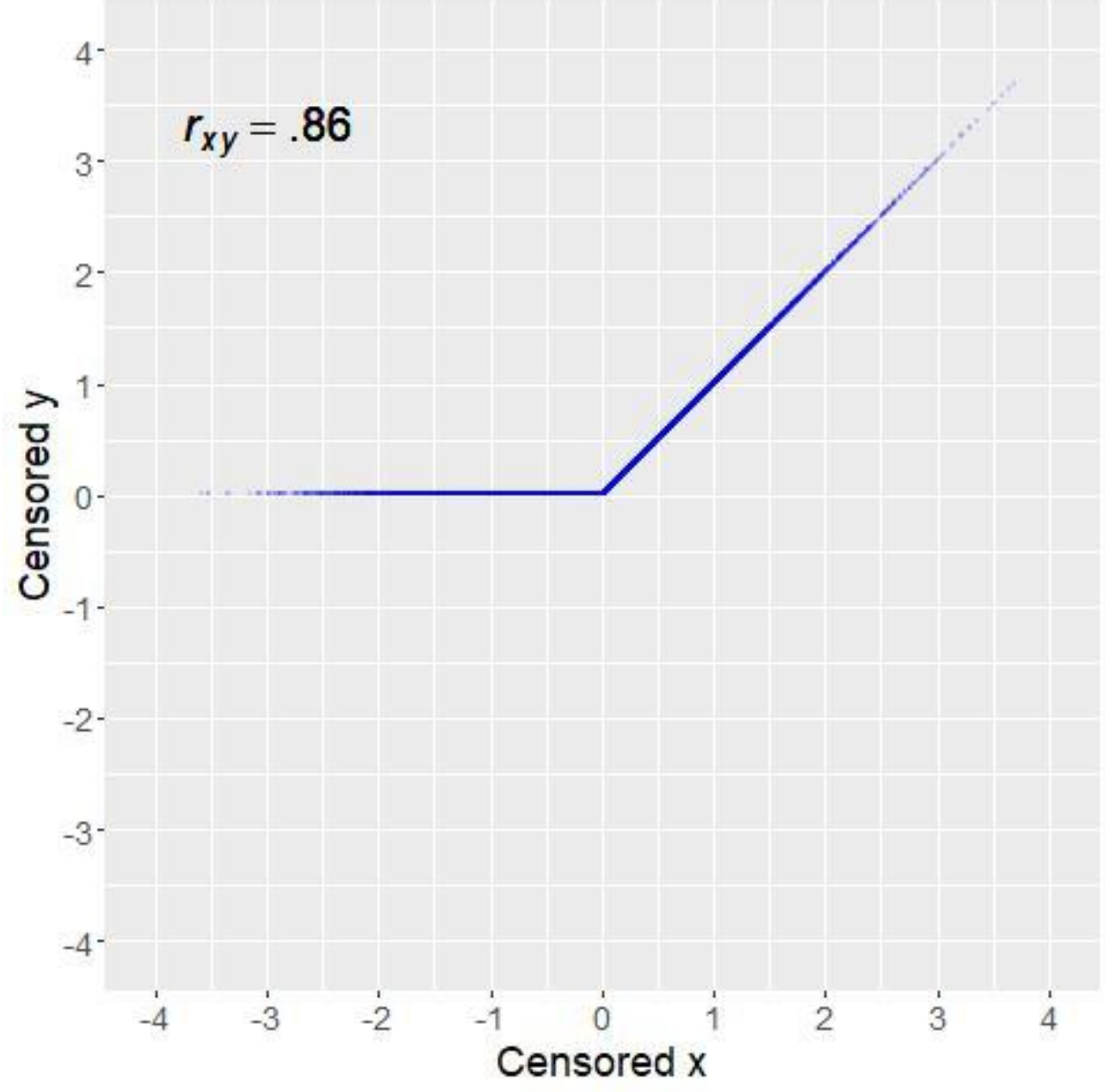
Censoring changes correlations.

Usually, r_{xy} is closer to 0 than r_{XY}, misrepresenting the relationships between the variables.

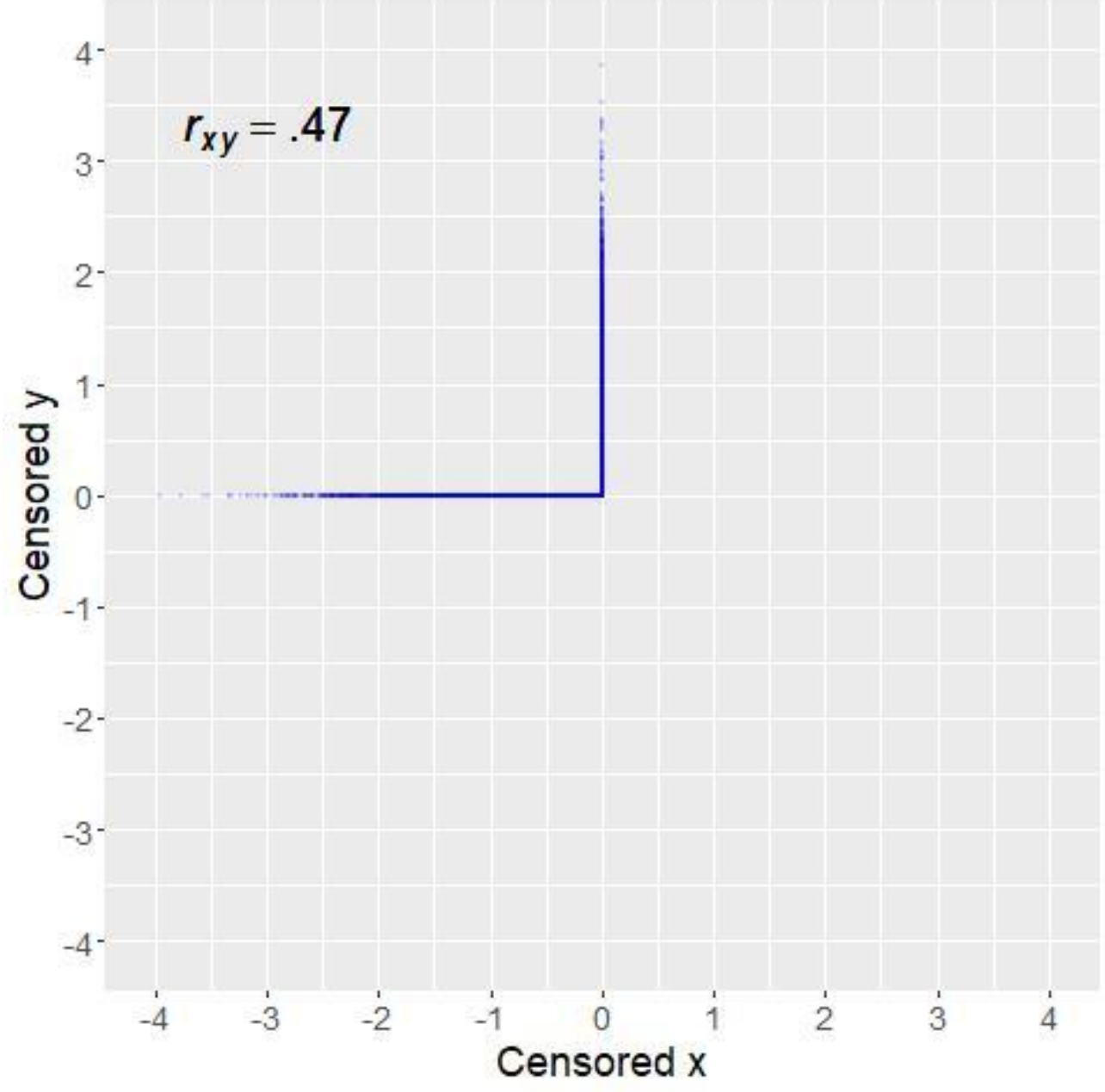
Panel A
No censoring
r_{XY} **= 1**
r_{xy} **= 1**



Panel B
One variable censored
r_{XY} **= 1**
r_{xy} **= .86**
0% left censoring on x
50% left censoring on y



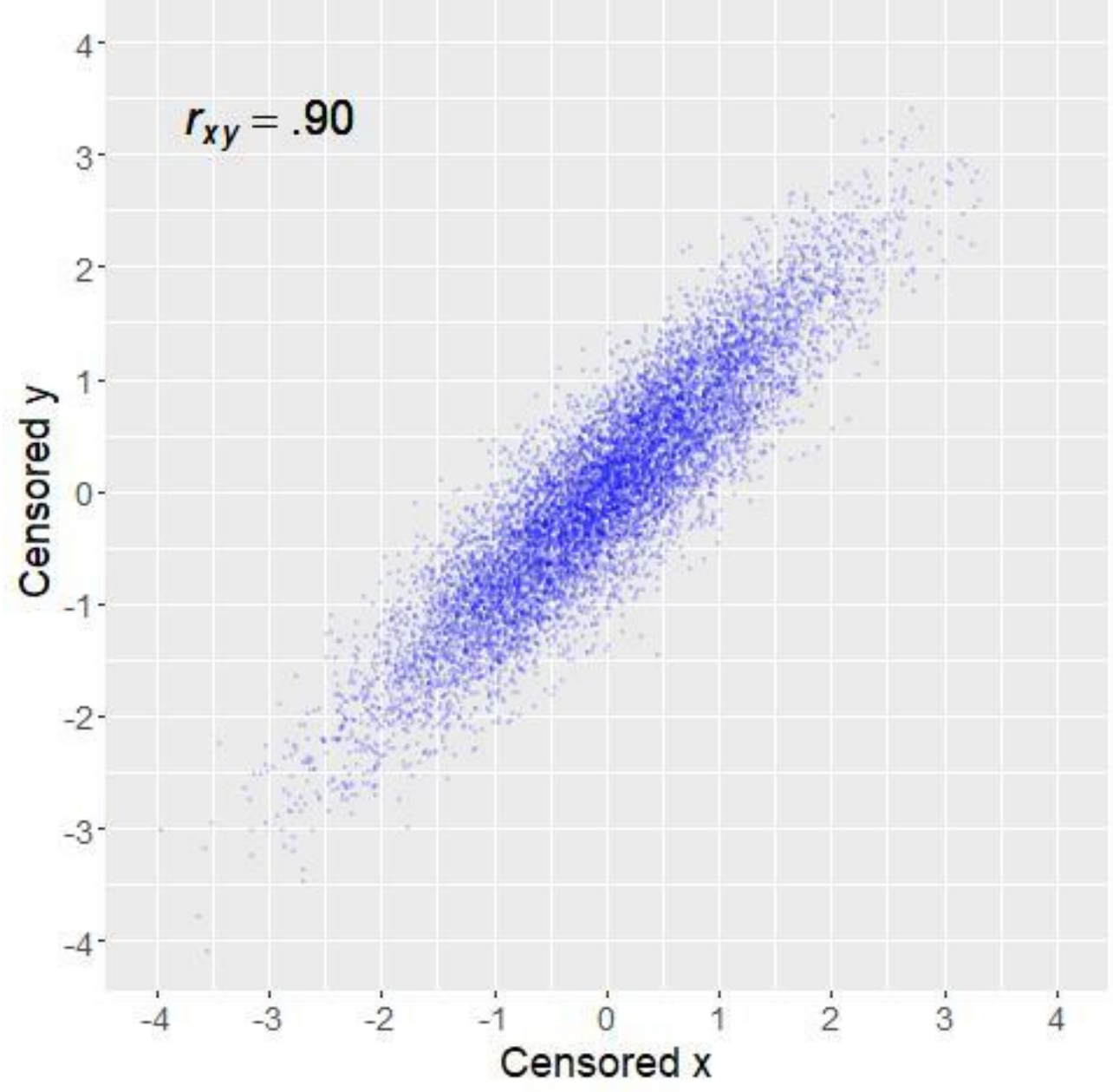
Panel C
Both variables censored
r_{XY} **= 1**
r_{xy} **= .47**
50% right censoring on x
50% right censoring on y



Uneven censoring has bigger effects on correlations than equal censoring.

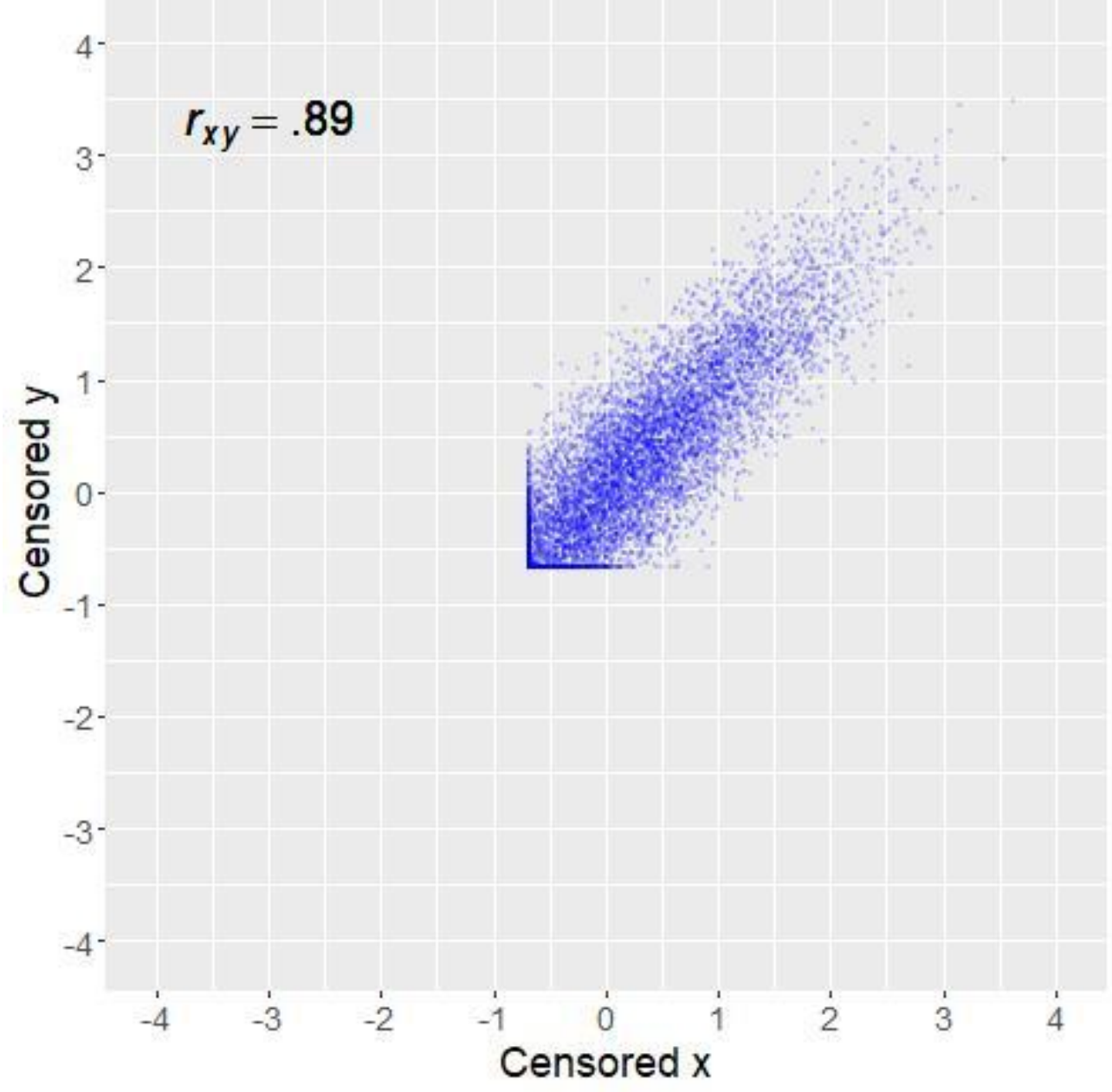
Panel A
No censoring
r_{XY} **= .90**
r_{xy} **= .90**
0% censoring on x
0% censoring on y

No censoring on bottom left



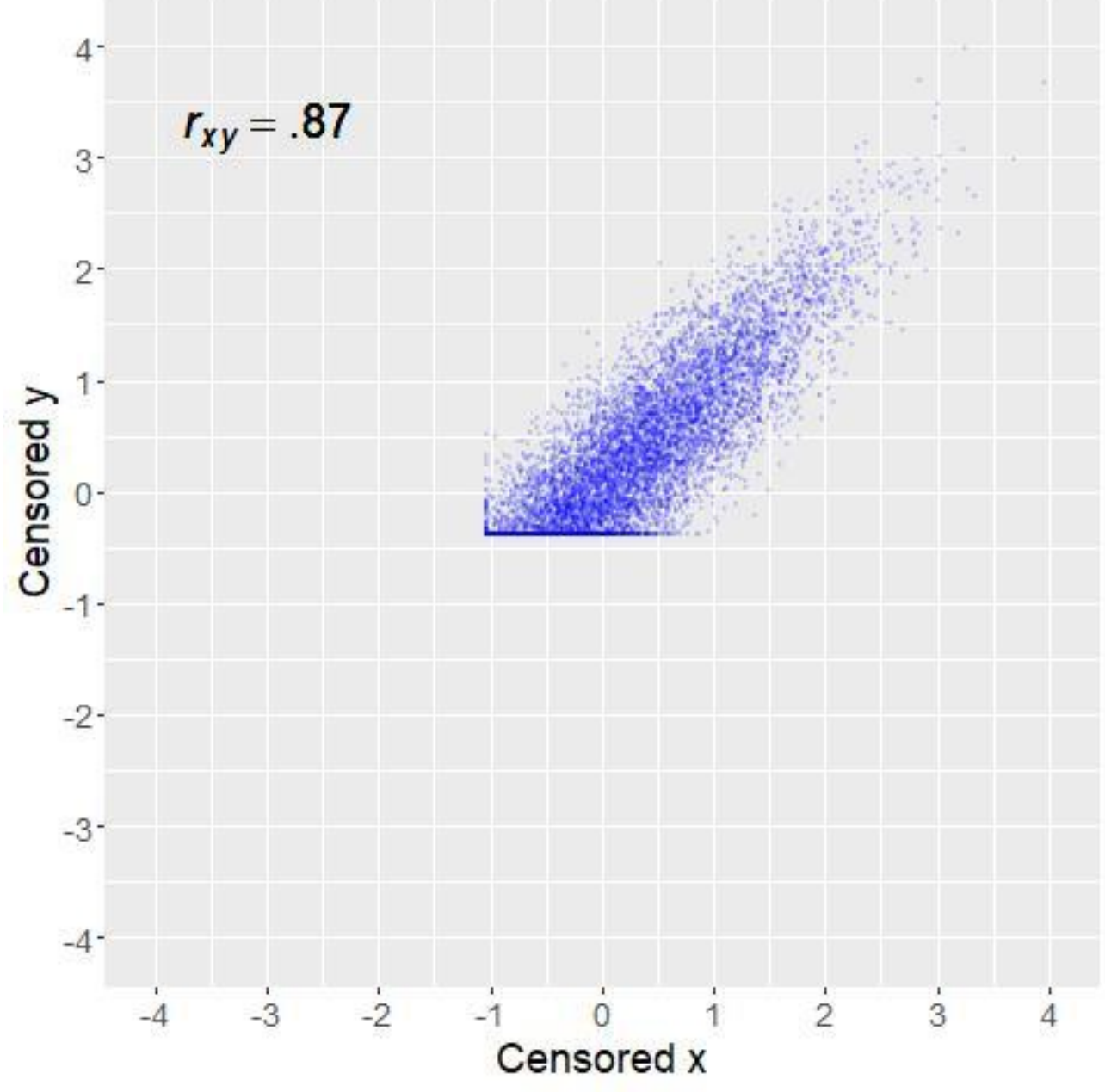
Panel B
Even censoring
r_{XY} **= .90**
r_{xy} **= .89**
25% left censoring on x
25% left censoring on y

Even censoring on bottom left:
- Medium-long horizontal line
- Medium-long vertical line



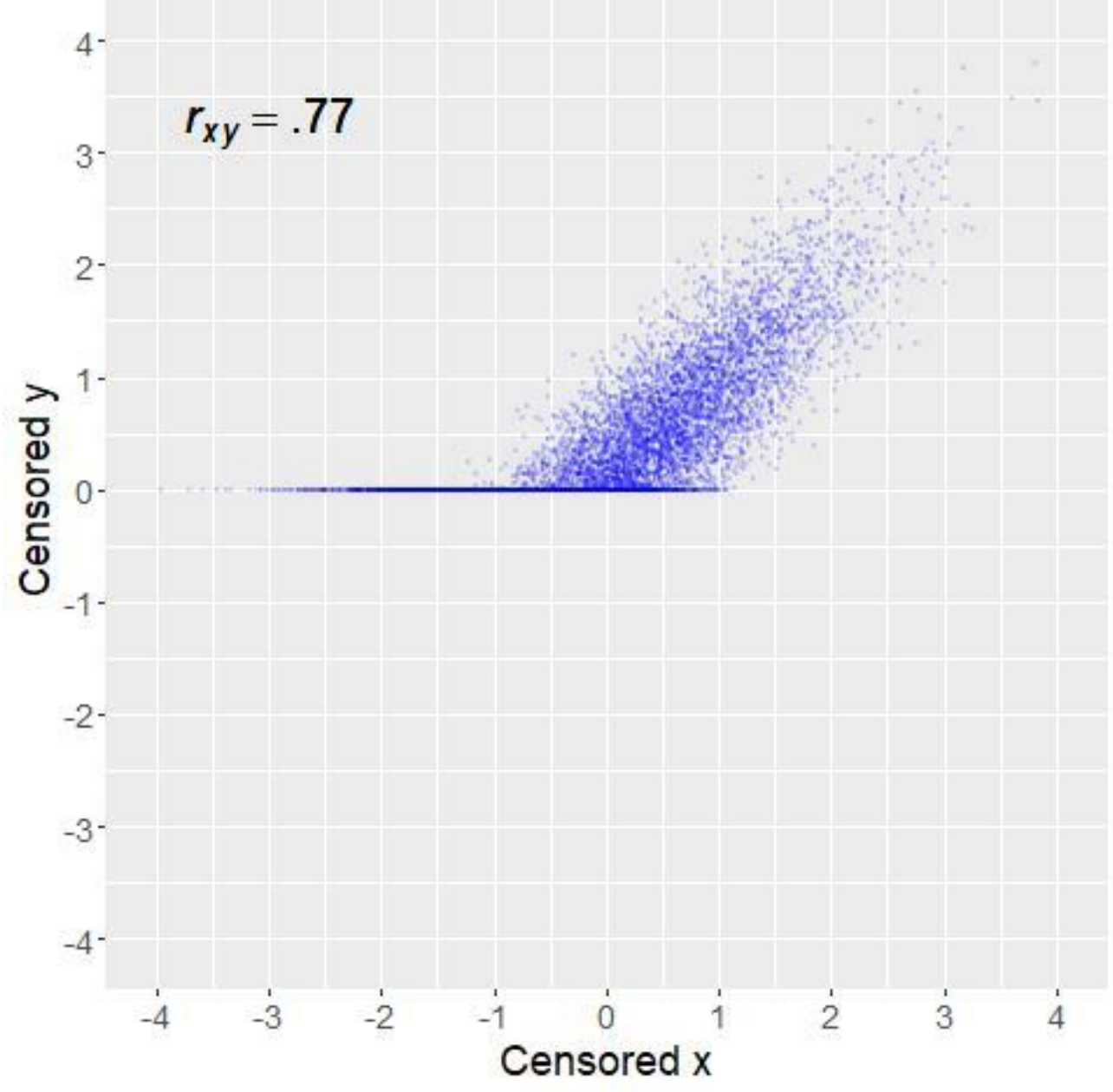
Panel C
Uneven censoring
r_{XY} **= .90**
r_{xy} **= .87**
15% left censoring on x
35% left censoring on y

Uneven censoring on bottom left:
- Long horizontal line
- Short vertical line



Panel D
Uneven censoring
r_{XY} **= .90**
r_{xy} **= .77**
0% left censoring on x
50% left censoring on y

Uneven censoring on bottom left:
- Long horizontal line
- No vertical line



DISCUSSION

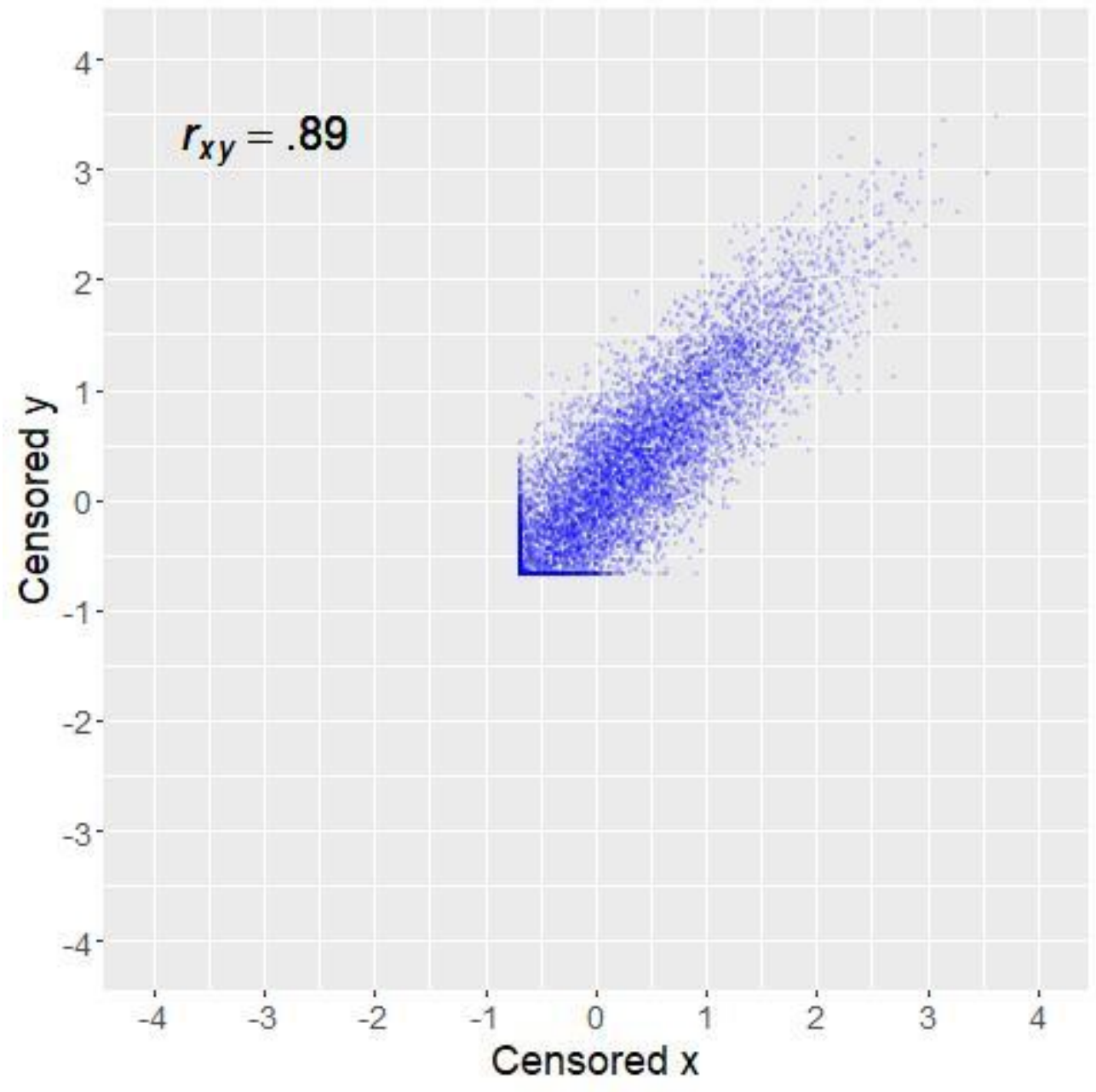
Censoring distorts the relations between variables. As a consequence, it also misrepresents the relations among sets of variables used in SEM, factor analyses, and longitudinal analyses. Censoring can therefore lead to difficulty factors and poor item statistics.

Researchers should account for censoring in their analyses. Mplus and R package *lavaan* can treat censored variables as ordinal, which improves estimation of the relations among uncensored variables. Even better, Monte Carlo studies show that R package *lava* accurately models bivariate (Barchard & Russell, in press) and multivariate relations (Holst et al., 2015) among uncensored variables.

For negative correlations, uneven censoring (and stronger effects on correlations) are more likely to occur when both variables are left censored or both right (e.g., both items are easy or both difficult). For positive correlations, uneven censoring (and stronger effects on correlations) are more likely to occur when one variable is left censored, the other right (e.g., one easy item, one hard item).

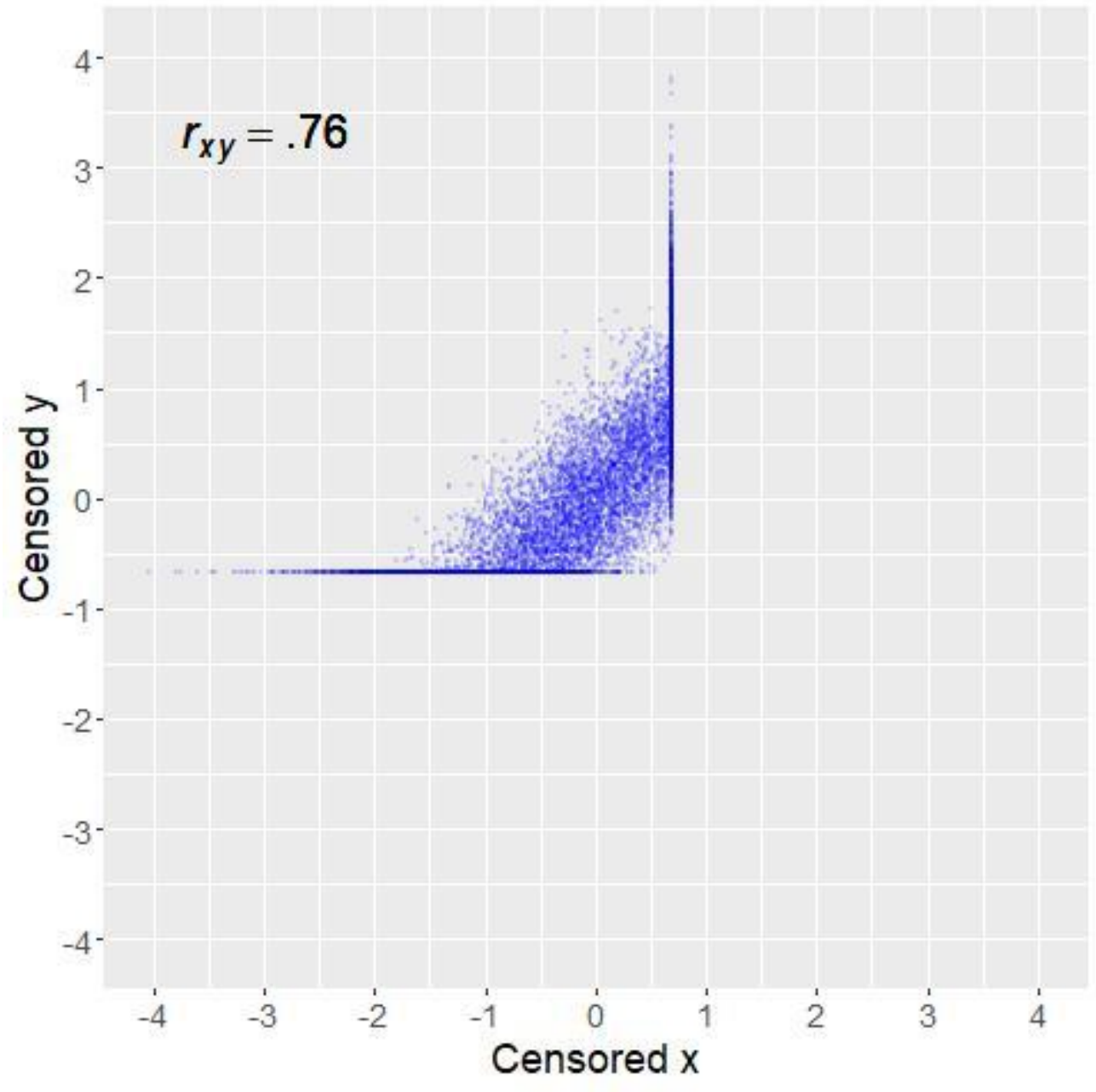
Panel A
Positive correlation
r_{XY} **= .90**
r_{xy} **= .89**
25% left censoring on x
25% left censoring on y

Even censoring on bottom left
No censoring on top right



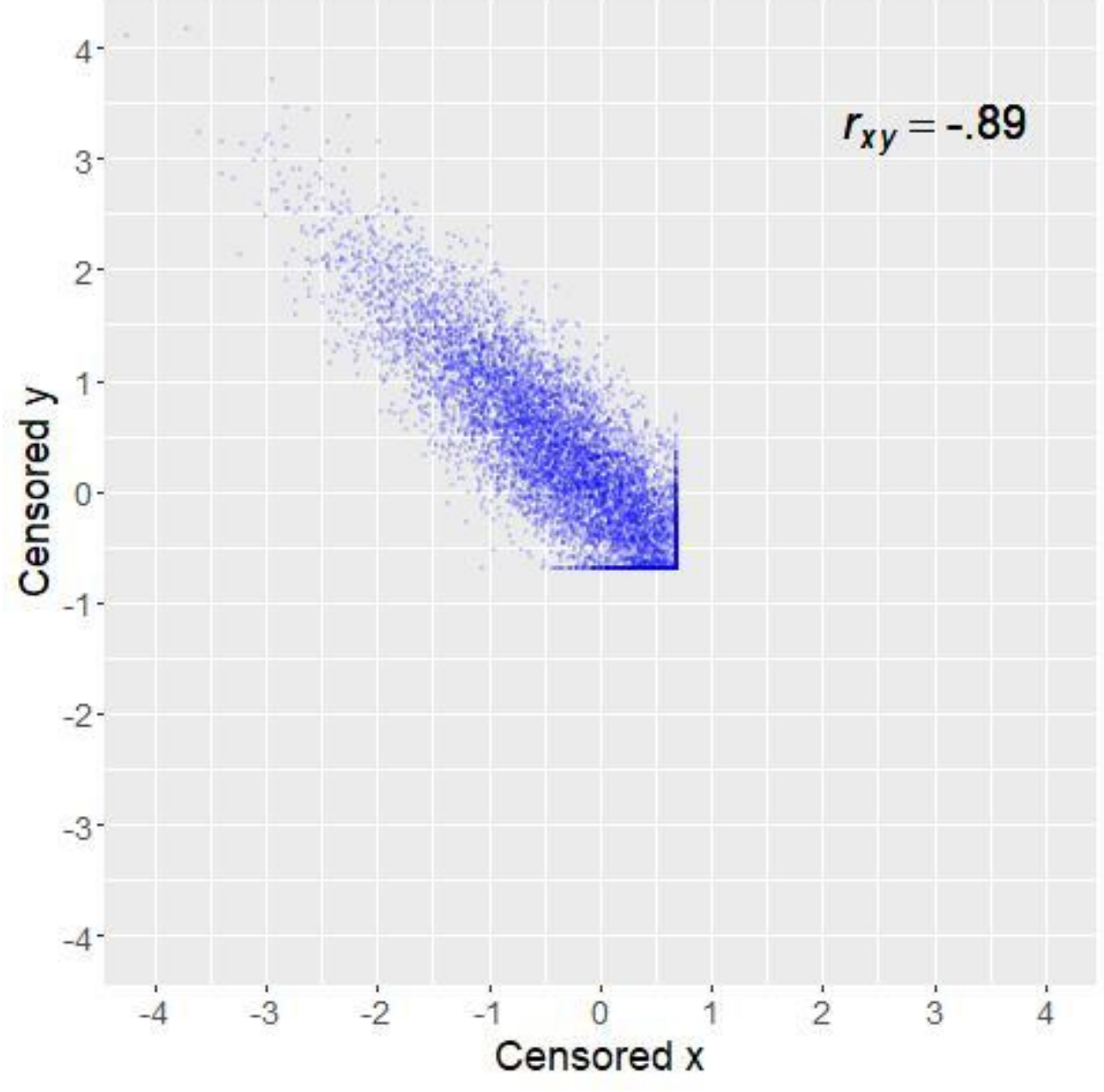
Panel B
Positive correlation
r_{XY} **= .90**
r_{xy} **= .76**
25% right censoring on x
25% left censoring on y

Uneven censoring on bottom left
Uneven censoring on top right



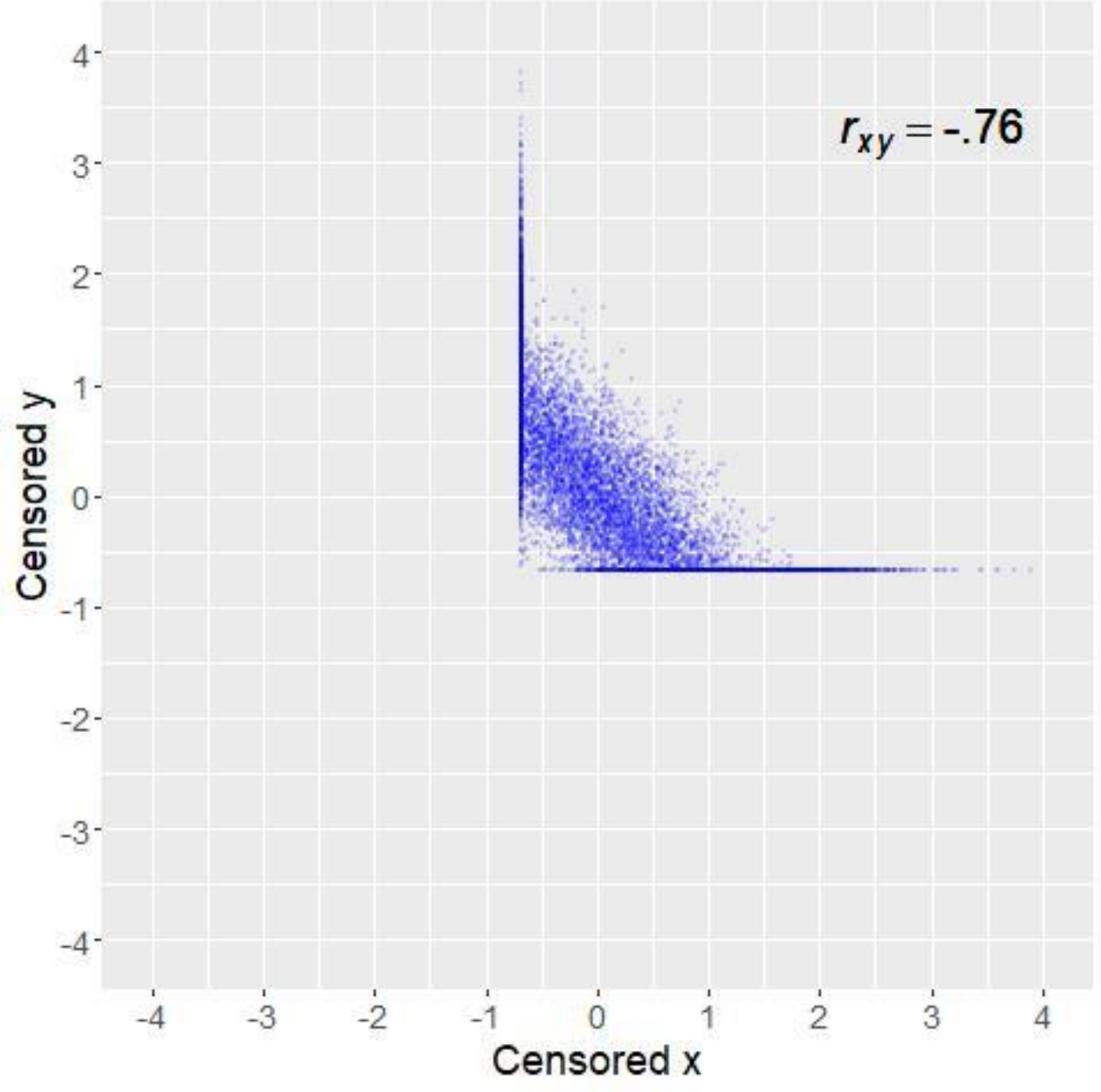
Panel C
Negative correlation
r_{XY} **= -.90**
r_{xy} **= -.89**
25% left censoring on x
25% right censoring on y

No censoring on top left
Even censoring on bottom right



Panel D
Negative correlation
r_{XY} **= -.90**
r_{xy} **= -.76**
25% left censoring on x
25% left censoring on y

Uneven censoring on top left
Uneven censoring on bottom right



REFERENCES

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

Holst, K. K., Budtz-Jørgensen, E., & Knudsen, G. M. (2015). *A latent variable model with mixed binary and continuous response variables.* https://arxiv.org/pdf/1507.01182.pdf