# Simulating Data in R

Heide Jackson
heidej@umd.edu

University of Maryland Population Research Center

August 2019

# High Level Things to Know

- R makes creating "fake" data very easy.
- Setting a seed makes findings reproducible
- Simulated data allows us to test model inferences and the sensitivity of results to violations of our assumptions.

# Simulating Different Types of Distributions

- Different functions can easily generate different types of variables.

```
#series of integers
id<-seq(1:100)
# 100 cases, mean 0, sd 1 normally distributed
error<-rnorm(n=100, mean=0, sd=1)
#binomial sample, values 0 and 1, probability of 1, .3
missing<-rbinom(n=100,1,.3)
#simulate a four category var, different probabilities
x3<-sample(x=0:3, size=100, replace=TRUE,
prob=c(.2,.1,.4,.3))
```

# Simulating Relationships

- Objects can also be created relative to other simulated data.

```
y2<-.5*x1+-.1*x2+error+.3*missing
y=ifelse(missing==1,NA,y2)
```

For analysis, it is helpful to wrap this up into a data frame.

```
fakedata<- data.frame(y2, x1, x2, missing)
```

# Running Simulated Data within a Function

- Simulating data within a broader function can be helpful for power calculations and modifying distributional assumptions.