Stochastic sample sizes

Felix Thoemmes

2019-04-01

Consider that you are conducting a randomized experiment - let's say a type of intervention for school children that tries to improve academic achievement. The program is administered to students who can be classified into three racial groups. For purposes of this example, we may call the three groups "White", "African-American", and "Asian". We further assume that the treatment effect is not constant across the three races. Despite this effect modification, we still might be interested in the average effect of the treamtent. You might object to taking an average in the presence of effect modification, but bear with me for now, and assume that it can in some circumstances still be interesting to look at an average.

Assume that the data at hand look like this:

Table 1: Hypothetical data from an education intervention. Numbers show means, and number in parantheses are observed proportions.

Condition	White	African-American	Asian	Proportion
Treatment	40 (25%)	30 (5%)	60 (20%)	(50%)
Control	50 (25%)	20 (5%)	50 (20%)	(50%)
Total	45 (50%)	25 (10%)	55 (40%)	(100%)

paper

R code to replicate results and graphs

```
library(lavaan)
library(ggplot2)
library(ggthemes)
library(ggdag)
library(dagitty)


x <- dagitty('dag {
             bb="0,0,1,1"
             M [pos="0.5,0"]
             U [pos="0.5,0.5"]
             X [pos="0,0"]
             Y [pos="1,0"]
             M -> Y
             U -> X
             U -> Y
             X -> M
             }')

ggdag(x) + theme_economist() + theme(axis.line=element_blank(),axis.text.x=element_blank(),
                                     axis.text.y=element_blank(),axis.ticks=element_blank(),
                                     axis.title.x=element_blank(),
```

```
                                    axis.title.y=element_blank(),legend.position="none",
                                    panel.background=element_blank(),panel.border=element_blank(),panel


fd <- function(n,a,b,u1,u2) {
  U <- rnorm(n,0,1)
  e_x <- rnorm(n,0,1)
  e_m <- rnorm(n,0,1)
  e_y <- rnorm(n,0,1)
  X <- u1*U + e_x
  M <- a*X + e_m
  Y <- b*M + u2*U + e_y

  df1 <- data.frame(X,M,Y,U)

  fdmodel <- "M ~ a*X
              Y ~ b*M
              X ~~ Y
              FD := a*b"


  fd <- sem(fdmodel,df1)@ParTable$est[7]
  unadj <- coef(lm(Y~X))[2]
  adj <- coef(lm(Y~X+U))[2]
  return(c(fd,unadj,adj))
}

res <- data.frame(t(replicate(5000,fd(100,.5,.5,.5,.5))))
names(res) <- c("fd","unadj","adj")

cols <- viridis(3)

ggplot(res,aes(x=fd)) + geom_density(alpha=.3,fill=cols[1]) + theme_economist() +
  geom_vline(xintercept = .25,col=cols[2],size=1.25) +
  geom_vline(xintercept = mean(res$fd),col=cols[3],size=1.25) + xlab("Front-door") + ylab("Density") +
  coord_cartesian(xlim=c(.1,.8)) + ggtitle("Parameter estimates using front-door criterion")

ggplot(res,aes(x=adj)) + geom_density(alpha=.3,fill=cols[1]) + theme_economist() +
  geom_vline(xintercept = .25,col=cols[2],size=1.25) +
  geom_vline(xintercept = mean(res$adj),col=cols[3],size=1.25) + xlab("Adjusted") + ylab("Density") +
  coord_cartesian(xlim=c(.1,.8)) + ggtitle("Parameter estimates using adjustment")

ggplot(res,aes(x=unadj)) + geom_density(alpha=.3,fill=cols[1]) + theme_economist() +
  geom_vline(xintercept = .25,col=cols[2],size=1.25) +
  geom_vline(xintercept = mean(res$unadj),col=cols[3],size=1.25) + xlab("Unadjusted") + ylab("Density")
  coord_cartesian(xlim=c(.1,.8)) + ggtitle("Unadjusted parameter estimates")
```