Why use an Average Treatment Effect? How to do statistical inference for average treatmente effects?

Jake Bowers[1,2]

EGAP Learning Days 4: Chile

[1] *Political Science & Statistics & NCSA @ University of Illinois*

[2] *White House Social and Behavioral Sciences Team*

jwbowers@illinois.edu — http://jakebowers.org

Why randomize?

To facilitate interpretable statements about comparisons (i.e. to remove confounds, to decide on intervention).

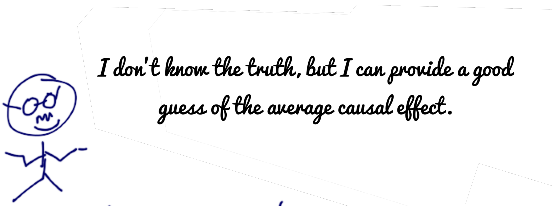
To facilitate interpretable statements about information (i.e. to justify hypothesis tests and estimators.)

What does it mean to say that $E(\widehat{\text{ATE}}) = \text{ATE}$

Confidence Intervals

Recall the ATE

Like hypotheses and imputation, the ATE can help with the fundamental problem of causal inference.



i	Z_i	Y_i	y_{i1}	y_{i0}
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i1}$	$\bar{y}_{i0}$

$$\widehat{ATE} = \bar{Y}_i | Z_i = 1 - \bar{Y}_i | Z_i = 0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

$E(\widehat{\text{ATE}}) = $ ATE means that $\widehat{\text{ATE}}$ is unbiased for ATE.

```r
## Bias refers to a relationship between the repeated operation of a procedure and a truth. So we have to invent a truth.
dat$y0<-dat$rpre ## create true potential outcomes to control
trueATE<-.2 ## posit a true average treatment effect
dat$y1<-dat$y0+trueATE+rnorm(nrow(dat),mean=0,sd=sd(dat$y0)) ## create potential outcomes to treatment
dat$obsy<-with(dat, z*y1+(1-z)*y0 ) ## what we observe
trueATE<-with(dat,mean(y1)-mean(y0))
estATE<-coef(lm(obsy~z,dat))["z"] ## same as a mean difference on obsy
## Define two functions: (1) calc est ATE and (2) re-assign treatment
makeNewObsyAndEst<-function(thez){
    newobsy<-with(dat, thez*y1+(1-thez)*y0 )
    lmATE<-coef(lm(newobsy~thez))[["thez"]]
    return(c(lmATE=lmATE))
}
makeNewZ<-function(thez,theb){
        unsplit(lapply(split(thez,theb),sample),theb)
}
## Does the pair of functions do what we want them to do?
replicate(5,makeNewObsyAndEst(makeNewZ(dat$z,dat$s)))

    lmATE     lmATE     lmATE     lmATE     lmATE
 0.741860 -1.862096 -5.509230  0.545759 -5.990799
```

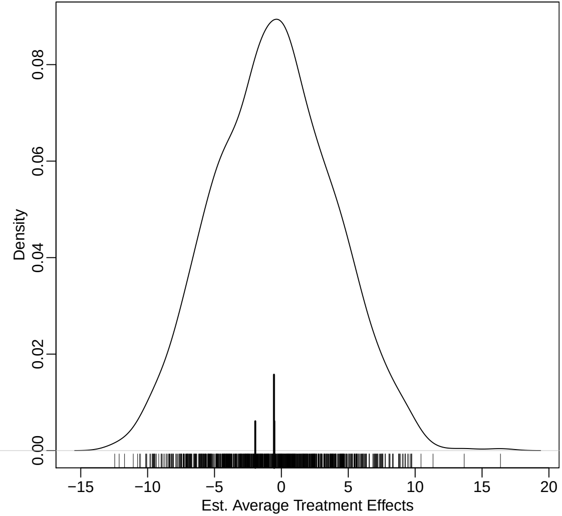
```r
nsims <- 1000
set.seed(20150313)

## For many of the possible ways to run the experiment, calculate this mean difference 
## dist.sample.est<-replicate(nsims,makeNewObsyAndEst(makeNewObsyAndEst(makeNewZ(da
## on your unix-based machine (mac or linux):
require(parallel)
ncores <- detectCores()
dist.sample.est <- simplify2array(mclapply(1:nsims, function(i) {
    makeNewObsyAndEst(makeNewZ(dat$z, dat$s))
}, mc.cores = ncores))

c(EestATE = mean(dist.sample.est), ATE = trueATE, estATE = estATE)
```

```
   EestATE        ATE   estATE.z
-0.5250059 -0.5534720 -1.9511192
```

```r
## And recall that we have simulation error on the order of 1/sqrt(nsims)
SEsims <- sqrt(var(dist.sample.est)/nsims)
SEsims
```

```
[1] 0.1397582
```

What does it mean to say we have an unbiased estimator?



What does it mean to say that $E(\widehat{\text{ATE}}) = \text{ATE}$

Confidence Intervals

Confidence Interval Ingredients

$$\text{CI}(\text{ATE}) = \widehat{\text{ATE}} \pm z_{\alpha/2}\text{SE}(\widehat{\text{ATE}})$$

where, $z_{\alpha/2}$ for $\alpha = .05$ is 1.96.

Standard Errors for the Estimated ATE

What is a standard error in the context of a randomized experiment?

Here, pretending that the randomization was simple and not blocked.

```
## See the Dunning / Freedman, Pisani, Purves derivation
y0 <- dat$y0
y1 <- dat$y1
Z <- dat$z
Y <- Z * y1 + (1 - Z) * y0
V <- var(cbind(y0, y1))
varc <- V[1, 1]
vart <- V[2, 2]
covtc <- V[1, 2]
N <- length(y0)
n <- sum(Z)
m <- N - n
varestATE <- ((N - n)/(N - 1)) * (vart/n) + ((N - m)/(N - 1)) * (varc/m) + (2/(N - 1)) * c
## And the *feasible* version (where we do not observe the potential outcomes)
varYc <- var(Y[Z == 0])
varYt <- var(Y[Z == 1])
fvarestATE <- (N/(N - 1)) * ((varYt/n) + (varYc/m))
```

```r
lm1 <- lm(Y ~ Z)
iidSElm1 <- sqrt(diag(vcov(lm1)))[["Z"]]

c(trueSE = sqrt(varestATE), feasible = sqrt(fvarestATE), iid = iidSElm1, simSE = sd(di

  trueSE feasible      iid    simSE
5.635336 6.550205 6.481614 4.419543
```

Different Confidence Intervals

```
theiidci <- confint(lm1, level = 0.95, parm = "Z")
feasCI <- estATE + c(1, -1) * qnorm(0.05/2) * sqrt(fvarestATE)
bestCI <- estATE + c(1, -1) * qnorm(0.05/2) * sqrt(varestATE)

rbind(feasCI, theiidci, bestCI)

            2.5 %    97.5 %
feasCI -14.78928 10.887046
Z      -14.99793 11.095694
bestCI -12.99617  9.093936
```

Which is better?

A good test casts doubt on the truth rarely.

A good confidence interval contains the truth at least 100α % of the time. (Because a confidence interval is a collection of hypotheses against which we have little information to argue. A confidence interval is collection of unsurprising hypotheses.)

Checking Coverage

```
makeFeasibleSE <- function(y, z) {
    varYc <- var(y[z == 0])
    varYt <- var(y[z == 1])
    N <- length(y)
    stopifnot(N == length(z))  ## a test of the code
    fvarestATE <- (N/(N - 1)) * ((varYt/n) + (varYc/m))
    return(fvarestATE)
}

makeCIs <- function(y, thez) {
    lm1 <- lm(y ~ thez)
    estATE <- coef(lm1)["thez"]
    theiidci <- confint(lm1, level = 0.95, parm = "thez")
    fvarestATE <- makeFeasibleSE(y = y, z = thez)
    thefeasci <- estATE + c(1, -1) * qnorm(0.05/2) * sqrt(fvarestATE)
    truthinIIDci <- 0 >= min(theiidci) & 0 <= max(theiidci)
    truthinFeasci <- 0 >= min(thefeasci) & 0 <= max(thefeasci)
    return(c(truthinIIDci = truthinIIDci, truthinFeasci = truthinFeasci))
}

makeCIs(y = Y, thez = sample(Z))

 truthinIIDci truthinFeasci
         TRUE          TRUE
```

Checking Coverage

```
set.seed(20160509)
nsims <- 10000
coverageCheck <- simplify2array(mclapply(1:nsims, function(i) {
    makeCIs(y = Y, thez = sample(Z))
}, mc.cores = ncores))
## coverageCheck<-replicate(10000, makeCIs(y=Y,thez=sample(Z))) ##makeNewZ(Z,Y)))
apply(coverageCheck, 1, mean)

 truthinIIDci truthinFeasci
       0.9434        0.9398
```

Review

What is unbiasedness? Why do we care? How would we assess bias?

What is a confidence interval? How would we assess coverage?