Lab 6: External Validity

Due: Monday, February 26, 11:59 PM

Name: Your name here

Mac ID: The first half of your Mac email address

1 External Validity

In the readings for this week, Coppock et al (2018) mention how the correspondence in effects between representative and convenience samples depends on the distribution of individual treatment effects.

The following design simulates a model with **heterogeneous treatment effects**, and compares the result of survey experiments conducted with a representative and convenience sample.

```
# Parameters

N = 1000 # population


n = 100 # sample


effect = 0.5


# Model
```

```
model = declare_model(

  N = N,

  U = rnorm(N),

  X = runif(N), # observed covariate

  potential_outcomes(

    # interaction represents heterogeneity

    Y ~ Z * effect * X + U

    )

)
```

We are also specifying X as an observed covariate that moderates the treatment effect, some-
thing like digital literacy. It's generated by random draws of a uniform distribution between
0 and 1 (hence the runif function). If X is 1, the unit experiences the full effect. If it is zero,
the effect disappears. The numbers in between scale the treatment effect accordingly. This
is a way to simulate heterogeneous treatment effects.

The inquiry is standard fare:

```
# Inquiry

inquiry = declare_inquiry(

  ATE = mean(Y_Z_1 - Y_Z_0)

)
```

Then we have to compare two data strategies, the survey experiment with a random sample
and the one with a convenience sample. At this point our research design branches into two
paths, since each data strategy will also have its own analogous answer strategy. They are
essentially two different designs, but we can recycle some components.

This is how it looks for the representative sample:

```
# Data strategy

r_sampling = declare_sampling(S = complete_rs(N, n = n))


assignment = declare_assignment(Z = complete_ra(N))


# Answer strategy

measurement = declare_measurement(Y = reveal_outcomes(Y ~ Z))


estimator = declare_estimator(

  Y ~ Z,

  inquiry = "ATE"

  )
```

Then put everything together:

```
r_design = model + inquiry +

  r_sampling + assignment +

  measurement + estimator
```

To create our convenience sample, we need a custom sampling function that makes it so that units with higher X are more likely to be drawn.

```
convenience_sampling = function(data){


  id = sample(data$ID, size = n, prob = data$X)


  data$S = with(

    data,

    ifelse(
```

```
     data$ID %in% id, 1, 0
     )

   )


  data[data$S == 1, ]
}
```

Then we pass this custom function to `declare_sampling`.

```
c_sampling = declare_sampling(handler = convenience_sampling)
```

And now we can create a separate design for our convenience sample

```
c_design = model + inquiry +

  c_sampling + assignment +

  measurement + estimator
```

Then we can diagnose both designs at once:

```
# remember to replace with student number
set.seed(123)


r_diag = diagnose_design(r_design)


c_diag = diagnose_design(c_design)
```

And we can use the following function to fetch the bias and RMSE of each design.

```
diagnosands = rbind(

  r_diag$diagnosands_df %>%

  select(design, bias, rmse),

  c_diag$diagnosands_df %>%
```

```
    select(design, bias, rmse)
)



diagnosands
```

```
    design         bias       rmse

1 r_design -0.003418615 0.2066810

2 c_design  0.067789877 0.2099169
```

> **i Task 1**
>
> Which design is better in terms of bias and RMSE? What explains this?

> **i Task 2**
>
> What happens to the bias and RMSE of both designs as the sample size `n` *increases* but the population `N` remains constant? What happens when the sample size decreases? What explains this?
>
> *Hint: Show two results, one with a larger sample and one with smaller sample.*

> **i Task 3**
>
> What happens to the bias and RMSE when the population and sample sizes are the same? What explains this?
>
> *Hint: It may be faster to compute this by choosing a number in between the original population and sample sizes rather than making both equal to 1,000.*

2 Answers

2.1 Task 1

Work on your answer here.

2.2 Task 2

Work on your answer here.

2.3 Task 3

Work on your answer here.