hwrk9 - Beimnet Taye

2023-03-22

Worked with Joan Shim and Lucas Yoshida

P1

1.

```
scurvy = medicaldata::scurvy %>%
mutate(
# collapse treatment into two groups
citrus = forcats::fct_collapse(
treatment,
"TRUE" = c("cider", "dilute_sulfuric_acid", "vinegar", "citrus"),
"FALSE" = c("sea_water", "purgative_mixture"),
) %>%
as.logical(),
# turn outcome into a numeric measure
gum_rot = gum_rot_d6 %>%
stringr::str_sub(1, 1) %>%
as.numeric()
) %>%
select(citrus, gum_rot)

naive <- scurvy %>%
  group_by(citrus) %>%
  summarize(mean = mean(gum_rot))

naive_ATE <- naive[2,2] - naive[1,2]

naive
```

```
## # A tibble: 2 x 2
##   citrus  mean
##   <lgl>  <dbl>
## 1 FALSE   3
## 2 TRUE    1.75
```

```
naive_ATE
```

```
##    mean
## 1 -1.25
```

2.

```
scurvy_counter <- scurvy %>%
  mutate(gum_rot_yes_citrus = ifelse(citrus == TRUE, gum_rot, NA),
         gum_rot_no_citrus = ifelse(citrus == FALSE, gum_rot, NA)
         )

scurvy_counter
```

```
## # A tibble: 12 x 4
##    citrus gum_rot gum_rot_yes_citrus gum_rot_no_citrus
##    <lgl>    <dbl>              <dbl>             <dbl>
##  1 TRUE         2                  2                NA
##  2 TRUE         2                  2                NA
##  3 TRUE         1                  1                NA
##  4 TRUE         2                  2                NA
##  5 TRUE         3                  3                NA
##  6 TRUE         3                  3                NA
##  7 FALSE        3                 NA                 3
##  8 FALSE        3                 NA                 3
##  9 TRUE         1                  1                NA
## 10 TRUE         0                  0                NA
## 11 FALSE        3                 NA                 3
## 12 FALSE        3                 NA                 3
```

3.

```
citrus_bad <- scurvy_counter %>%
  mutate(gum_rot_yes_citrus = ifelse(citrus == FALSE, 3, gum_rot),
         gum_rot_no_citrus = ifelse(citrus == TRUE, 0, gum_rot))

citrus_bad %$% {
  mean(gum_rot_yes_citrus) - mean(gum_rot_no_citrus)
}
```

```
## [1] 1.166667
```

```
citrus_bad
```

```
## # A tibble: 12 x 4
##    citrus gum_rot gum_rot_yes_citrus gum_rot_no_citrus
##    <lgl>    <dbl>              <dbl>             <dbl>
##  1 TRUE         2                  2                 0
##  2 TRUE         2                  2                 0
##  3 TRUE         1                  1                 0
##  4 TRUE         2                  2                 0
##  5 TRUE         3                  3                 0
##  6 TRUE         3                  3                 0
##  7 FALSE        3                  3                 3
```

```
##  8 FALSE        3                 3                 3
##  9 TRUE         1                 1                 0
## 10 TRUE         0                 0                 0
## 11 FALSE        3                 3                 3
## 12 FALSE        3                 3                 3
```

4.

```
citrus_good <- scurvy_counter %>%
  mutate(gum_rot_yes_citrus = ifelse(citrus == FALSE, 0, gum_rot),
         gum_rot_no_citrus = ifelse(citrus == TRUE, 3, gum_rot))

citrus_good %$% {
  mean(gum_rot_yes_citrus) - mean(gum_rot_no_citrus)
}
```

```
## [1] -1.833333
```

```
citrus_good
```

```
## # A tibble: 12 x 4
##    citrus gum_rot gum_rot_yes_citrus gum_rot_no_citrus
##    <lgl>    <dbl>              <dbl>             <dbl>
##  1 TRUE         2                  2                 3
##  2 TRUE         2                  2                 3
##  3 TRUE         1                  1                 3
##  4 TRUE         2                  2                 3
##  5 TRUE         3                  3                 3
##  6 TRUE         3                  3                 3
##  7 FALSE        3                  0                 3
##  8 FALSE        3                  0                 3
##  9 TRUE         1                  1                 3
## 10 TRUE         0                  0                 3
## 11 FALSE        3                  0                 3
## 12 FALSE        3                  0                 3
```

5.

```
citrus_equal <- scurvy_counter %>%
  mutate(gum_rot_yes_citrus = ifelse(citrus == FALSE,0, gum_rot),
         gum_rot_no_citrus = ifelse(citrus == TRUE, 3, gum_rot))

citrus_equal[7,3] = 3
citrus_equal[8,3] = 3
citrus_equal[11,3] = 1

citrus_equal %$% {
  mean(gum_rot_yes_citrus) - mean(gum_rot_no_citrus)
}
```

```
## [1] -1.25
```

```
citrus_equal
```

```
## # A tibble: 12 x 4
##    citrus gum_rot gum_rot_yes_citrus gum_rot_no_citrus
##    <lgl>    <dbl>              <dbl>             <dbl>
##  1 TRUE         2                  2                 3
##  2 TRUE         2                  2                 3
##  3 TRUE         1                  1                 3
##  4 TRUE         2                  2                 3
##  5 TRUE         3                  3                 3
##  6 TRUE         3                  3                 3
##  7 FALSE        3                  3                 3
##  8 FALSE        3                  3                 3
##  9 TRUE         1                  1                 3
## 10 TRUE         0                  0                 3
## 11 FALSE        3                  1                 3
## 12 FALSE        3                  0                 3
```

P2

1.

```
causal_dgp_1 = function(n=100) {
tibble(
X = runif(n,-1,1), # a covariate
A = rbern(n, prob=1/2),
Y0 = rnorm(n, mean=sin(pi*X)+1/3),
Y1 = rnorm(n, mean=cos(pi*X)),
)
}
observable_dgp_1 = function(n=100) {
causal_dgp_1(n) %>%
mutate(Y = ifelse(A, Y1, Y0)) %>%
select(-Y0, -Y1)
}
logistic = function(x) 1 / (1 + exp(-x))

causal_dgp_2 = function(n=100) {
tibble(
X = runif(n,-1,1), # a covariate
A = rbern(n, prob=logistic(X)),
Y0 = rnorm(n, mean=sin(pi*X)+1/3),
Y1 = rnorm(n, mean=cos(pi*X)),
)
}

observable_dgp_2 = function(n=100) {
causal_dgp_2(n) %>%
mutate(Y = ifelse(A, Y1, Y0)) %>%
```

```
select(-Y0, -Y1)
}


naive_ate_estimator = function(data) {
data %$%
{ mean(Y[A==1]) - mean(Y[A==0]) }
}
```

```
true_ate <- causal_dgp_1(1000000) %$% {
  mean(Y1) - mean(Y0)
}
true_ate
```

```
## [1] -0.3303849
```

2.

```
bias_ate <- function(data,true_data, estimator, rep = 10, n = 100){
  true <- true_data(1000000) %$% {
  mean(Y1) - mean(Y0)
  }
  map_df(1:rep,function(.x){
  return(tibble(estimate = estimator(data(n)))
      )
    }
  ) %>%
    summarize(
      bias = mean(estimate) - true,
      variance = var(estimate)
    )
}

estimate_eval <- bias_ate(observable_dgp_1,causal_dgp_1, naive_ate_estimator,rep = 1000)
estimate_eval
```

```
## # A tibble: 1 x 2
##       bias variance
##      <dbl>    <dbl>
## 1 -0.00668   0.0644
```

3

```
estimate_eval2 <- bias_ate(observable_dgp_2,causal_dgp_2, naive_ate_estimator, rep =1000)
estimate_eval2
```

```
## # A tibble: 1 x 2
##    bias variance
##   <dbl>    <dbl>
## 1 0.151   0.0646
```

4.

- The bias is higher in dgp2 but the variance is the same between the two dgps. This makes sense since looking at DGP2 covariate X affects both the outcome Y and exposure A while in DGP 1 covariate X only affects the outcome but not the exposure A. This makes X a confounder in DGP 2 but not in DGP 1 and since we are not controlling for X in either scenario the bias would be higher in the DGP where X is a confounder, in this case DGP2.

P3

1.

- We can't directly calculate the ATE since we are missing data the captures each individuals counterfactual outcome. In this case, we cannot have everyone do both the particular stretch and not doing said stretch. As such we can't measure every individuals outcomes under both conditions. The ATE is defined as the difference in outcomes between when everyone is doing the stretch of interest and when no one is doing the stretch of interest.

2.

- There could be unmeasured characteristics that affect the likelihood of doing that particular stretch and the occurrence and/or self-reporting of pain.For instance the prevalence of that stretch could be higher in younger people than it is in older people. Younger people typically report and/or are in less physical pain than older people usually. Age could then confound the relationship between the stretch and pain by making it seem like the stretch is helpful when it is in fact the age of individuals driving observed effect.

P4.

1.

Statistics and Causal Inference talks about the necessary components of a valid causal question under the SUTVA framework. Specifically under SUTVA a causal question can be broken up into indexed units and treatments with outcomes measured given a particular combination of units and treatments. It argues for thinking of such a framework when evaluating causal questions for further analysis. Does Water Kill attempts to address several criticisms levied against the potential outcomes approach for causal inference. The first issue addressed is the issue of vagueness inherent to causal questions to which the author mentions how the vagueness is a built in component of the potential outcomes approach. The second issue addressed is that the potential outcomes approach does not need to provide a definite yes or no to a causal question. The third issue is addressed by the author stating how the potential outcomes approach could be in fact used in non feasible interventions. Race and Sex are causes argues how race and sex could theoretically be variables that could be intervened on and that causal claims don't have to be tied to potential intervention effects.

On Causes, Causal Inference, and Potential Outcomes defends the potential outcomes approach as a basis of defining causal estimands and linking data with causal estimands. Causal Inference for Social Exposures attempts to argue that quantitative causal inference should work in tandem with more qualitative methods of causal inference in order to describe causal effects as they are in the real world. On the Interpretation of do(x) emphasizes the importance of studying the causal effects of non manipulable variables and the role they play in causal inference.

2.

I mostly agree with the Does Water Kill article since it best synthesizes the issues at hand and coherently addresses each issue clearly. I believe a robust quantitative inferential framework is a beneficial thing since it forces investigators to be specific in regards to their scientific question and hypothesis.The limitation of the scope of the potential outcomes approach is a built in feature not a flaw that makes transparent the limitations and interpretability of a given finding, all of which is critical information needed when making resource allocation decisions.