Lab 7: Hypothesis Testing

Due: Monday, March 4, 11:59 PM

Name: Your name here

Mac ID: The first half of your Mac email address

1 Hypothesis Testing

The following code simulates a two-arm experiment in which the policy in question has a positive effect

```
N = 300 # sample size


tau = 0.2 # true unobserved treatment effect


# M
model = declare_model(
  N = N,
  U = rnorm(N),
  potential_outcomes(Y ~ tau * Z + U)
)


# I
```

```
inquiry = declare_inquiry(
  ATE = mean(Y_Z_1 - Y_Z_0)
)

# D
assign = declare_assignment(
  Z = complete_ra(N, prob = 0.5)
  )

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

# A
estimator = declare_estimator(
  Y ~ Z, inquiry = "ATE"
  )

rct = model + inquiry + assign + measure + estimator
```

The following code draws data data from a single realization of the experiment:

```
set.seed(123) # replace with student id

observation = simulate_design(rct, sims = 1)

observation$estimate
```

```
[1] 0.1604132
```

Is this estimate enough evidence to claim that the policy works? To know that we need to repeat an exercise similar to the lady tasting tea experiment we discussed in class.

However, we now have 300 observations, so we cannot simply list all the possible ways in which the experiment could be conducted. Instead, we will simulate the same experiment a large number of times assuming the policy does not have an effect.

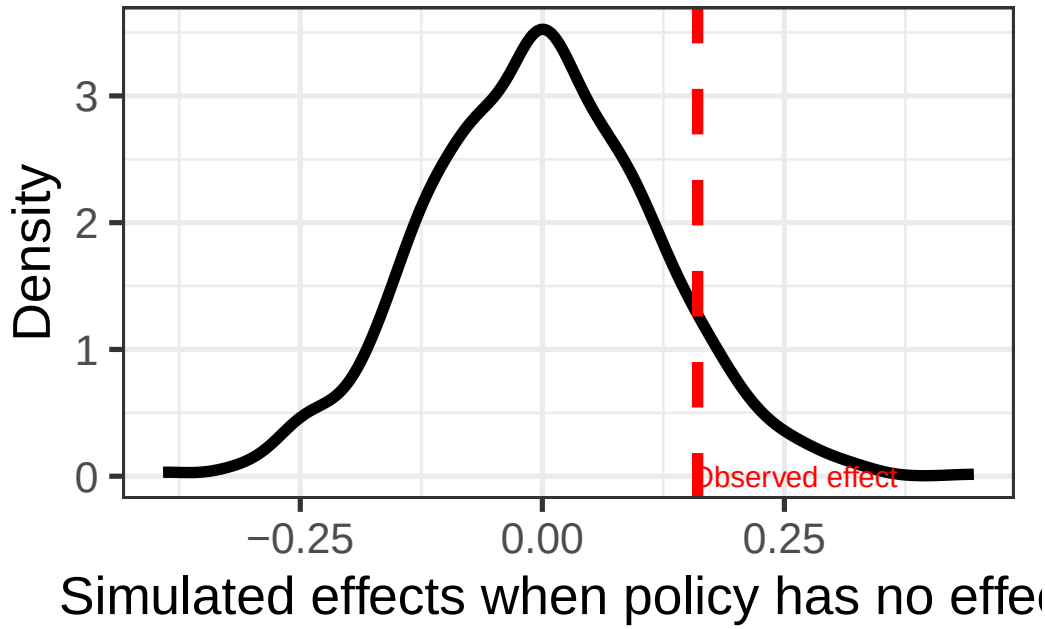
```
# Redesign to make true effect 0
no_effect = redesign(rct, tau = 0)


# Simulate 1k times
set.seed(123) # replace with student id
simulation = simulate_design(no_effect, sims = 1000)
```

Now we can compare our `observation` with the `simulation`. The following code makes a figure with the distribution of effects when the **null hypothesis of no effect** is true and compares it with the observed effect.

```
ggplot(simulation) +
  aes(x = estimate) +
  geom_density(linewidth = 2) +
  geom_vline(xintercept = observation$estimate,
             linetype = "dashed",
             linewidth = 2,
             color = "red") +
  annotate("text",
           x =  observation$estimate + 0.1,
           y = 0,
           label = "Observed effect",
           size = 4,
```

```
           color = "red") +
  labs(x = "Simulated effects when policy has no effect",
       y = "Density")
```



The visual illustration is helpful, but perhaps we want a one-number summary of the probability of observing something equal or more extreme than what we observed when the policy has no effect.

```
# Proportion of simulated estimates
# more extreme than the observed estimate
# either more positive or more negative
mean(abs(simulation$estimate) >= abs(observation$estimate))
```

```
[1] 0.166
```

This is our **p-value**, the proportion of simulated estimates that are equal or more extreme than what we observed.

Fortunately, we do not need to calculate this by hand for most research designs. The

`DeclareDesign` package takes care of that and we already stored the p-value when we saved `observation`.

```
observation$p.value
```

```
[1] 0.1435053
```

You may get slightly different numbers when you calculate p-values by hand and use the one generated by `DeclareDesign`. This is normal, most packages use an off-the-shelf formula to calculate p-values to approximate simulated p-values while saving computing time.

Task 1

- In the previous exercise , how do you interpret `observation$estimate` in words?
- How do you interpret `observation$p-value` in words?
- Based on these two pieces of information, would you feel confident claiming that the policy works? Why?

2 Statistical Power

Experiments can always give small p-values by chance, even if the policy has no effect. If our p-value gives a lot of evidence against the null hypothesis of no effect but there is no effect in reality, we call this a **false positive**. Our data strategy suggests there is an effect when there is none.

How can we minimize the possibility of false positives **before** conducting an experiment? Our research design should meet two criteria:

1. The chances of false positives are low
2. The chances of detecting a true effect when it exists are high

We meet the first criterion by requiring p-values to be very small before claiming we have enough evidence against the null hypothesis. A common number in the social sciences

is $p < 0.05$. This number made sense when researchers needed to calculate p-values by hand. It does not make much sense nowadays, but it's still a useful rule of thumb for most applications.

For the second criterion, we can calculate the **statistical power** of an experimental answer strategy by simulating many realizations of the same experiment, assuming there is an effect, and calculating how often p-values surpass an arbitrary threshold.

For example, the following simulates many realizations of the `rct` design and calculates power under $p < 0.05$.

```
set.seed(123) # replace with student id
# simulate design with effect many times
rct_sim = simulate_design(rct, sims = 1000)


# calculate proportion of p.values smaller than 0.05
power = mean(rct_sim$p.value < 0.05)


power
```

```
[1] 0.39
```

This proportion represents how often we would be willing to claim there is enough evidence against the null hypothesis of no effect. This is for a research design for which **we already know there is an effect** because we declared it so in our `model` step.

Once again, this is not something that you need to calculate by hand. You obtain this when you diagnose a research design.

```
set.seed(123) # replace with student id


diag = diagnose_design(rct)
```

```
diag
```

```
Research design diagnosis based on 500 simulations. Diagnosis completed in 6 secs. Diagn
```

Design	Inquiry	Estimator	Outcome	Term	N	Sims	Mean Estimand	Mean Estimate
rct	ATE	estimator	Y	Z	500		0.20	0.20
							(0.00)	(0.01)

Bias	SD Estimate	RMSE	Power	Coverage
0.00	0.11	0.11	0.40	0.95
(0.01)	(0.00)	(0.00)	(0.02)	(0.01)

This output has a column named `power` that should be similar to how we calculate it by hand, discounting simulation variability and rounding.

Task 2

- Based on the statistical power that you saw above, would you say this is a good research design to evaluate the policy in question?
- What happens to the power of the** `rct` design when you increase `tau`? Why?
- What happens to power when you make** `tau` equal zero? Why?
- What happens to power when you increase or decrease (pick one) `N`? Why?

3 Answers

3.1 Task 1

Work on your answers here.

3.2 Task 2

Work on your answers here.