Causal Inference I Assignment 1

Tyler Rongxuan Chen, Alice Malmberg

2025-06-24

```
options(scipen = 999)
acorn <- read.csv("acorn03.csv")
```

This answer key make references to Thomas Leavitt's original answer key.

Question 1

First, define variables according to the prompt.

```
Z <- acorn$z
y <- acorn$vote03
n <- length(Z)
n1 <- sum(Z)
n0 <- n - n1
```

Compute the observed statistic.

```
obs_stat <- mean(y[Z == 1])
```

Simulate the null distribution.

```
set.seed(123) # for reproducibility
n_sim <- 10000
sim_stats <- numeric(n_sim)

for (i in 1:n_sim) {
  Z_perm <- rep(0, n)
  Z_perm[sample(1:n, n1)] <- 1
  sim_stats[i] <- mean(y[Z_perm == 1])
}
```

Compute p-value, null-mean, and null-variance.

The upper, one-sided p-value is simply the proportion of null test statistics greater than or equal to the observed test statistic (the dashed line in the plot above).

To calculate the null expected value of our test statistic, simply take the mean of our simulations of the test statistic under the null.

```r
# (a) Two-sided p-value
p_val_2 <- mean(abs(sim_stats - mean(sim_stats))
                >= abs(obs_stat - mean(sim_stats)))

# One-sided p-value
p_val_1 <- mean(sim_stats >= obs_stat)

# (b) Null expected value
null_mean <- mean(sim_stats)

# (c) Null variance
null_var <- var(sim_stats)
```

Output the results.

```r
cat("Observed test statistic is", obs_stat, "\n")
```

```
## Observed test statistic is 0.3248104
```

```r
cat("Simulation two-sided p-value is", p_val_2, "\n")
```

```
## Simulation two-sided p-value is 0.1529
```

```r
cat("Simulation one-sided p-value is", p_val_1, "\n")
```

```
## Simulation one-sided p-value is 0.0759
```

```r
cat("Null expected value is", null_mean, "\n")
```

```
## Null expected value is 0.3067747
```

```r
cat("Null variance is", null_var, "\n")
```

```
## Null variance is 0.0001564705
```

Question 2

Under the strict null hypothesis (no effect), the treatment assignment Z is independent of the outcome variable y. In this case, treatment is assigned randomly: selecting n_1 units from the total n units.

$$
\begin{aligned}
&\mathbb{E}\left[n_1^{-1}\mathbf{Z}'\mathbf{y}\right] \\
&\quad = n_1^{-1}\mathbb{E}\left[\mathbf{Z}'\mathbf{y}\right] && \text{Since } \mathbb{E}\left[c\right] = c \\
&\quad = n_1^{-1}\mathbb{E}\left[\sum_{n=1}^{n} Z_i y_i\right] && \text{By the definition of matrix multiplication} \\
&\quad = n_1^{-1}\sum_{n=1}^{n}\mathbb{E}\left[Z_i y_i\right] && \text{By the linearity of expectations} \\
&\quad = n_1^{-1}\sum_{n=1}^{n} y_i\mathbb{E}\left[Z_i\right] && \text{Since } y_i \text{ is a constant} \\
&\quad = n_1^{-1}\sum_{n=1}^{n} y_i\frac{n_1}{n} && \text{By the random assignment process of the experiment} \\
&\quad = n_1^{-1}\left(y_1\frac{n_1}{n}\right) + \cdots + \left(y_n\frac{n_1}{n}\right) && \text{By the definition of the summation operator} \\
&\quad = n_1^{-1}\frac{n_1}{n}\left(y_1 + \cdots + y_n\right) && \text{By the distributive property } (ab)+(ac) = a(b+c) \\
&\quad = \frac{1}{n_1}\frac{n_1}{n}\left(y_1 + \cdots + y_n\right) && \text{Since } n_1^{-1} = \frac{1}{n_1} \\
&\quad = \frac{1}{n}\left(y_1 + \cdots + y_n\right) && \text{Since } \frac{n_1}{nn_1} = \frac{1}{n} \\
&\quad = \frac{\left(y_1 + \cdots + y_n\right)}{n} \\
&\quad = \overline{y}
\end{aligned}
$$

The expected value of the mean outcome in the treated group is equal to the mean of all outcomes:

$$\mathbb{E}[\bar{y}_{\text{treat}}] = \bar{y}$$

```
ybar <- mean(y)
cat("The expected value is", ybar, "\n")
```

```
## The expected value is 0.3066632
```

Question 3

Let's break this down, and note that this is close to our in-class exercise.

(1) The Variance Under Complete Randomization (Finite Population)

$$\mathbb{V}\left[\bar{y}_1\right] = \frac{1}{n_1}\frac{n_0}{n}\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

where:

- n_1: number assigned to treatment,
- n_0: number assigned to control,

- $n = n_1 + n_0$,
- $\bar{y}$: mean of all y.

we can further reexpress the variance of the test statistic as follows:

$$\begin{aligned}
\mathbb{V}\left[n_1^{-1}\mathbf{Z}'\mathbf{y}\right] &= \frac{1}{n_1}\frac{n_0}{n}\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}{n-1} \\
&= \frac{1}{n_1}\frac{n_0}{n}\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)}{n-1} \\
&= \frac{n_0\left(\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\right)}{(n-1)\,(n)\,(n_1)} \\
&= \frac{n_0}{n-1}\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)}{n}\frac{1}{n_1} \\
&= \frac{n_0}{n-1}\sigma_y^2\frac{1}{n_1} \\
&= \frac{n-n_1}{n-1}\frac{\sigma_y^2}{n_1}
\end{aligned}$$

which is the expression for the variance of the sample mean given in the question.

(2) The Usual (i.i.d, Infinite Population) Variance Formulas

- **Sample-based** (sample variance of treated group):

$$\widehat{\mathbb{V}}[\bar{y}_1] = \frac{s_y^2}{n_1}$$

`where $s_y^2$ is the variance of the treated group.`

- **Population-based**:

$$\mathbb{V}[\bar{y}_1] = \frac{\sigma_y^2}{n_1}$$

`where $\sigma_y^2$ is the population variance.`

(3) Which Formula is Larger?

The **usual i.i.d formula** ($\frac{\sigma_y^2}{n_1}$) gives a **larger value** than the randomization formula.

Why?

- The randomization formula multiplies by $\frac{n_0}{n}$, which is **less than 1** unless all units are treated ($n_0 = n_1$).
- The usual i.i.d formula assumes infinite population sampling (sampling with replacement), so each draw is independent.
- **Randomization variance** reflects sampling **without replacement** from a finite population (once a unit is assigned, it can't be assigned again), which **reduces variance**.
- In fact, the **randomization formula is the same as the classical formula times the finite population correction (FPC)**:

$$\text{FPC} = \frac{n_0}{n}$$

So,

$$\text{Randomization variance} = \text{IID variance} \times \text{FPC}$$

(4) Why Does This Make Sense?

- **Sampling without replacement (finite population, randomization):** There's **less variability** because once you assign a unit to treatment, you can't assign it again—so the samples are less variable.
- **Sampling with replacement (infinite population, IID):** Each sample is independent, so variability is higher.

In summary: The **randomization variance is smaller** because it's based on a finite, fixed set of units, and every assignment is "without replacement," so each treated group's mean is less variable than if you were sampling from an infinite population.

Question 4

Compute the variance using the formula:

```
# Sample variance of y
s2 <- sum((y - ybar)^2) / (n - 1)

# Variance under randomization formula
var_formula <- (1 / n1) * (n0 / n) * s2
cat("Variance under randomization formula is", var_formula, "\n")
```

```
## Variance under randomization formula is 0.0001553611
```

```
cat("Variance using simulation is", null_var, "\n")
```

```
## Variance using simulation is 0.0001564705
```

```r
# Absolute error
# Because simulation results are different, this number can vary a lot.
abs_error <- abs(null_var - var_formula)

# Percent error (as a percent of formula value)
# Because simulation results are different, this number can vary a lot.
# But you get the idea that the two numbers are close to each other.
percent_error <- abs_error / var_formula * 100

# Print all
cat("The variance from the formula is", var_formula, "\n")
```

```
## The variance from the formula is 0.0001553611
```

```r
cat("The simulated variance is", null_var, "\n")
```

```
## The simulated variance is 0.0001564705
```

```r
cat("The absolute error is", abs_error, "\n")
```

```
## The absolute error is 0.000001109395
```

```r
cat("The percent error is", percent_error, "%", "\n")
```

```
## The percent error is 0.7140751 %
```

Question 5

In the case of the Acorn experiment, we can calculate the Z-score as follows:

$$\begin{aligned}\text{Z-score} &= \frac{n_1^{-1}\mathbf{Z}'\mathbf{y} - \mathbb{E}\left[n_1^{-1}\mathbf{Z}'\mathbf{y}\right]}{\sqrt{\text{Var}\left[n_1^{-1}\mathbf{Z}'\mathbf{y}\right]}} \\ &\approx \frac{(0.3248 - .3067)}{0.0125}\end{aligned}$$

Let's do this in R:

```r
# z-score
z <- (obs_stat - null_mean) / sqrt(var_formula)

# Normal approximation p-value (right tail)
pval_normal <- 1 - pnorm(z)
cat("The approximated p-value is", pval_normal, "\n")
```

```
## The approximated p-value is 0.07395193
```

Note that the p-value is close to what we get from Question 1.

Question 6

First, a quick summary of the three methods:

- **Method I:** Toss a coin for each subject (fully independent randomization).
- **Method C:** Three "30" cards and three "60" cards, shuffle, assign.
- **Method P:** Pair up subjects, toss a coin for each pair, one gets 30, one gets 60.

Each subject gets $Z = 0$ or $Z = 1$ (30 or 60 min), for a total of 6 subjects.

---

a) Strengths and Weaknesses of Each Method

Method I (independent coins):

Method I independently assigns each subject to treatment ($Z_i = 1$) with probability 0.5. Under simple random assignment all subjects are assigned to groups without regard to the assignments of other subjects in the study; this assignment process is especially simple to implement. With a small n, however, this method may result in no subjects in one of the two conditions. If $n = 6$, then, under simple random assignment (method I), the probability that all units are assigned to the treatment condition is $0.5^6 \approx 0.0156$ and the probability that all units are assigned to the control condition is also $0.5^6 \approx 0.0156$. Although small, the probability of these two outcomes taken together is $0.5^6 + 0.5^6 \approx 0.0312$.

- **Strength:** Simple, truly independent, easy to implement.
- **Weakness:** Number assigned to 60 (Z=1) could vary from 0 to 6; imbalance between treatment and control is possible.

Method C (shuffled cards):

Method C has the benefit of enabling the researcher to assign a predetermined number of subjects to treatment and control such that there is a fixed number of participants in each condition. Method P assigns units to treatment and control within blocked pairs, which (if covariates are predictive of potential outcomes) decreases the variance of the randomization distribution.

- **Strength:** Guarantees exactly 3 will be assigned 60 and 3 assigned 30.
- **Weakness:** Not independent; assignment to one affects assignment to others.

Method P (pairs):

- **Strength:** Guarantees balance *within* each pair and exactly 3 will be 60, 3 will be 30.
- **Weakness:** Possible correlation within pairs, not independent. Also, less flexibility if pairs are not meaningful.

---

b) How do answers change for $n = 600$?

- **Method I:** With 600 subjects, law of large numbers would mean that groups would likely be balanced (~300 in each), so weakness is less important.
- **Method C:** Still exact balance.
- **Method P:** Still exact balance, but pairing 600 subjects may be impractical unless pairs are meaningful.

---

c) $\mathbb{E}[Z_1]$ under each method

- **Method I:** Each coin is fair, so $\mathbb{E}[Z_1] = 0.5$.
- **Method C:** Of 6 shuffled cards, 3 are "1", so $\mathbb{E}[Z_1] = 3/6 = 0.5$.
- **Method P:** Each pair, one gets 1, one gets 0, so $\mathbb{E}[Z_1] = 0.5$.

---

d) $\mathbb{E}[Z_1 + Z_2 + \cdots + Z_6]$ under each method

- **Method I:** $6 \times 0.5 = 3$.
- **Method C:** Always exactly 3 are "1", so expectation is 3.
- **Method P:** Always exactly 3 are "1", so expectation is 3.

---

e) $\mathbb{E}[\mathbf{Z}'\mathbf{1}]$ under each method

This is the same as part (d), because $\mathbf{1}$ is a vector of all 1's, so $\mathbf{Z}'\mathbf{1} = Z_1 + Z_2 + \cdots + Z_6$.

So: Answer is 3 for all three methods.

---

f) For which methods does $\mathbb{E}[(\mathbf{Z}'\mathbf{1} - \mathbb{E}[\mathbf{Z}'\mathbf{1}])^2] = 0$?

- **Method I:** No; sum can vary from 0 to 6.
- **Method C:** Yes; always exactly 3, so variance is 0.
- **Method P:** Yes; always exactly 3, so variance is 0.

So, **Methods C and P only.**

---

g) (somehow there is no g here)

---

h) For which methods does linearity of expectation property hold?

- The property: $\mathbb{E}[\frac{Z_1 x_1 + \cdots + Z_6 x_6}{Z_1 + \cdots + Z_6}] = (x_1 + \cdots + x_6)/6$.

What does this property imply?

If you randomly assign "treatment" (here, asking to donate more time) to subjects, and then look at the average value among those who are treated, the expected value of that average is just the average of everyone, regardless of how the treatment was assigned (as long as the method gives everyone an equal chance).

In other words:

On average, the "treated" group looks like a random sample from the whole group.

This property entails the following strengths:

Unbiasedness: If you pick a random subset of people, their average trait/value is an unbiased estimator of the overall group average.

Randomization Justification: This property justifies why random assignment is so valuable in experiments: it means (in expectation) your treated and untreated groups are "representative" of the whole.

Simplicity: You don't need to worry about complicated weighting or adjustment (unless treatment assignment isn't equally likely for all, or the denominator can be zero).

Then when would this property break?

If assignment isn't random or isn't independent, or if there's some constraint that makes not all individuals equally likely to be treated (as in Method P).

Or, for small samples and Method I, you technically have to deal with the chance that nobody is treated.

- **Method I:** it's possible (with probability $(\frac{1}{2})^6 = \frac{1}{64}$) that all Z's are 0, so no one is treated. In that case, the denominator is zero.
- **Method C:** For Method C, exactly 3 are treated (denominator always 3).
- **Method P:** For Method P, always 3 treated as well (denominator always 3).

The argument fails for Method I.