Problem Set 1

Ellie Choe

population

ages <- c(“Teens”, “20s”, “30s”, “40s”, “50s”) p_pop <- c(0.15, 0.30, 0.25, 0.20, 0.10) names(p_pop) <- ages

sample sizes

n_vals <- c(50, 100, 200, 500, 1000)

container

res <- data.frame(n = integer(), group = character(), age = character(), prop = numeric())

—- simulation —-

for (n in n_vals) {

sample traits

trait <- sample(ages, size = n, replace = TRUE, prob = p_pop)

Treatment assignment

Z <- rbinom(n, 1, 0.5)

proportions for All / Treat / Control

prop_all <- as.numeric(table(trait)) / n n_treat <- sum(Z ==1) n_ctrl <- n - n_treat prop_t <- as.numeric(table(factor(trait[Z==1], levels = ages))) / n_treat prop_c <- as.numeric(table(factor(trait[Z==0], levels = ages))) / n_ctrl

res <- bind_rows(res, data.frame(n = n, group = “All”, age = ages , prop = prop_all), data.frame(n = n, group = “Treat”, age = ages, prop = prop_t), data.frame(n =n, group = “Control”, age = ages, prop = prop_c)) }

join population props

res_full <- res %>% left_join(data.frame(age = ages, pop_prop = p_pop), by = "age") %>% mutate(imbalance = pop_prop - prop)

—- quick check table —-

res_wide <- res_full %>% select(-imbalance) %>% pivot_wider(names_from = group, values_from = prop) %>% rename(population = pop_prop) %>% arrange(n, age)

—- plot 1: convergence to population —-

pop_tbl <- tibble(age = ages, pop_prop = p_pop)

print(ggplot(res, aes(x = n, y = prop, color = group)) + geom_point() + geom_line() + geom_hline(data = pop_tbl, aes(yintercept = pop_prop), linetype = 2) + facet_wrap(~ age, nrow = 2) + labs(x = "Sample size (n)", y = "Proportion", title = "Bigger Sample → More Balance Across Age Groups", subtitle = "Dashed Line = Population Proportion") + theme_minimal())

—- plot 2: imbalance measure —-

imbalance <- res %>% filter(group %in% c("Treat","Control")) %>% select(n, age, group, prop) %>% pivot_wider(names_from = group, values_from = prop) %>% mutate(abs_diff = abs(Treat - Control)) %>% group_by(n) %>% summarise(max_abs_diff = max(abs_diff), l1_sum_diff = sum(abs_diff), .groups = "drop")

print(ggplot(imbalance, aes(x = n, y = max_abs_diff)) + geom_hline(yintercept = 0, linetype = 2) + geom_point(size = 2) + geom_line() + labs(x = "Sample size (n)", y = "Max |Treat − Control| across age groups", title = "Bigger Sample → Smaller Worst-Case Imbalance") + theme_minimal())

—- data analysis —-

df <- read.csv("voting.csv")

1. Treatment variavle

variable: message

type: discrete

data type: character

2. create a binary variable

df$treat <- ifelse(df$message == "yes", 1, 0)

3. Compute the average outcome for the treatement and control groups

avg_treat <- mean(df$voted[df$treat == 1]) # average voting rate among treated voters avg_ctrl <- mean(df$voted[df$treat == 0]) # average voting rate among control voters. If avg_treat > avg_ctrl, the message increased turnout.

4. Subset the data frame

treat_df <- df[df$treat == 1,]$ctrl_df <- df[df$treat == 0,]

5. Average birth year

mean(treat_df$birth)$mean(ctrl_df$birth)

6. Estimated Average Causal Effect

ate <- avg_treat - avg_ctrl ate # it means treated voters were 8.1 percentage points more likely to vote.

7. Assumption for generalization

The sample must be representative of the US population,

and the treatment effect must be homogeneous across subgroups.