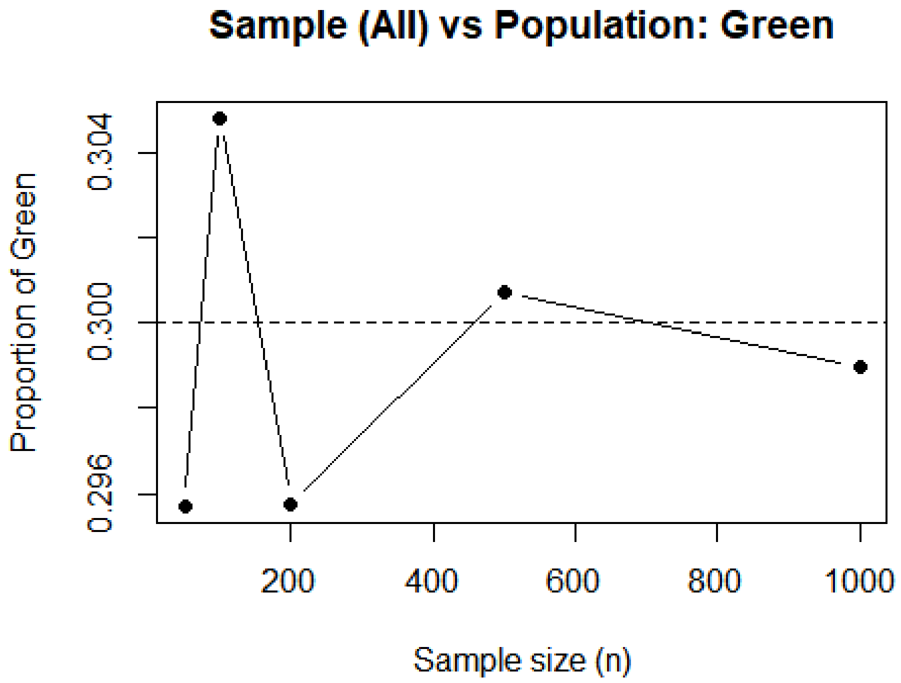
POLS 602 - PS1 - Joseph Kim

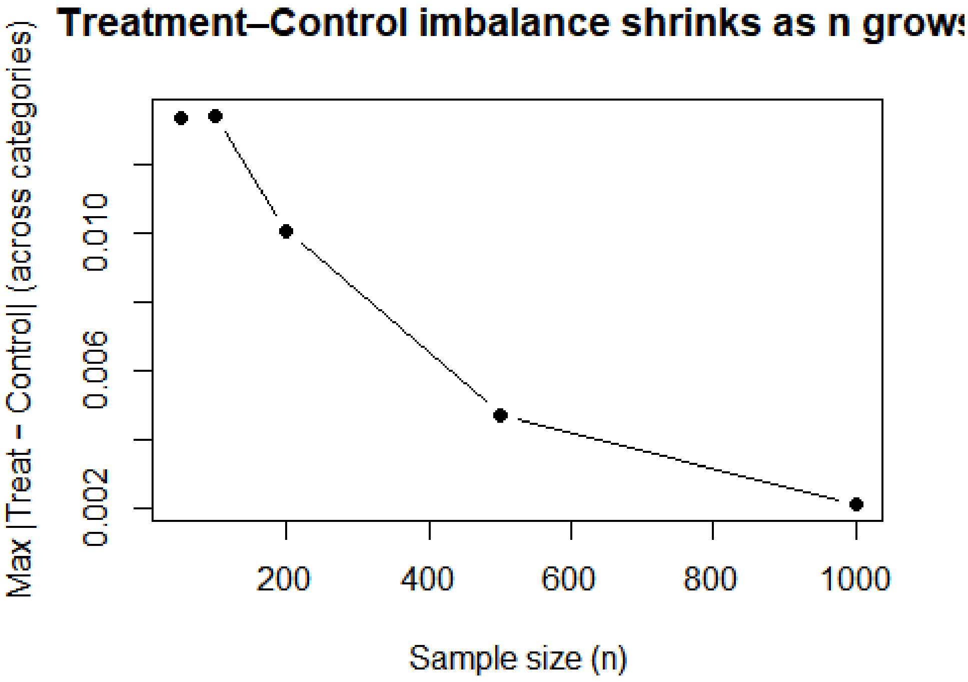
Part 1: Simulation

```r
voting <- read.csv("voting.csv")
set.seed(123)

# Population categories and proportions
colors <- c("Orange","Blue","Pink","Green","Purple")
p_pop  <- c(0.20, 0.10, 0.20, 0.30, 0.20); names(p_pop) <- colors

# Sample sizes and number of repetitions
n_vals <- c(50, 100, 200, 500, 1000)
M <- 200

# Container for results
res <- NULL

for (n in n_vals) {
  for (m in 1:M) {
    trait <- sample(colors, size = n, replace = TRUE, prob = p_pop)
    Z <- rbinom(n, 1, 0.5)  # 50/50 assignment

    # Compute proportions
    p_all <- as.numeric(table(factor(trait, levels = colors))) / n

    n_t <- sum(Z == 1); n_c <- n - n_t
    p_t <- if (n_t == 0) rep(NA_real_, length(colors)) else
      as.numeric(table(factor(trait[Z == 1], levels = colors))) / n_t
    p_c <- if (n_c == 0) rep(NA_real_, length(colors)) else
      as.numeric(table(factor(trait[Z == 0], levels = colors))) / n_c

    res <- rbind(
      res,
      cbind(n = n, group = "All",     cat = colors, prop = p_all),
      cbind(n = n, group = "Treat",   cat = colors, prop = p_t),
      cbind(n = n, group = "Control", cat = colors, prop = p_c)
    )
  }
}

res <- as.data.frame(res, stringsAsFactors = FALSE)
res$n <- as.integer(res$n)
res$prop <- as.numeric(res$prop)

# Average proportions by n, group, category
```

```r
avg <- aggregate(prop ~ n + group + cat, data = res, FUN = mean)

# Treat-Control imbalance summaries
avg_t <- avg[avg$group == "Treat", c("n","cat","prop")]
avg_c <- avg[avg$group == "Control", c("n","cat","prop")]
tc <- merge(avg_t, avg_c, by = c("n","cat"), suffixes = c("_t","_c"))
tc$abs_diff <- abs(tc$prop_t - tc$prop_c)

imbalance_max <- aggregate(abs_diff ~ n, data = tc, FUN = max)
imbalance_L1  <- aggregate(abs_diff ~ n, data = tc, FUN = sum)
names(imbalance_L1)[2] <- "L1_sum_diff"

# Plot 1: Sample (All) → population
cat_to_plot <- "Green"
sub_all <- avg[avg$group=="All" & avg$cat==cat_to_plot, c("n","prop")]
plot(sub_all$n, sub_all$prop, type="b", pch=16,
     xlab="Sample size (n)", ylab=paste0("Proportion of ", cat_to_plot),
     main=paste0("Sample (All) vs Population: ", cat_to_plot))
abline(h = p_pop[cat_to_plot], lty = 2)
```



```r
# Plot 2: Imbalance shrinks with n
plot(imbalance_max$n, imbalance_max$abs_diff, type="b", pch=16,
     xlab="Sample size (n)", ylab="Max |Treat - Control| (across categories)",
     main="Treatment-Control imbalance shrinks as n grows")
abline(h = 0, lty = 2)
```



Part 2: Data Analysis

1. The treatment variable is the ‘message’ column of the dataset. This is a discrete variable. The data type is binomial dummy variable comprised of “Yes” and “No” (but not a numeric variable)

2. New binomial numeric treatment variable (‘treat_bin’ column)

R codes for creating the column

```
voting$treat_clean <- tolower(trimws(as.character(voting$message)))
voting$treat_bin <- NA_integer_
voting$treat_bin[voting$treat_clean == "yes"] <- 1L
voting$treat_bin[voting$treat_clean == "no"]  <- 0L
```

3. R codes

```
treat_mean   <- mean(voting$voted[voting$treat_bin == 1], na.rm = TRUE)
control_mean <- mean(voting$voted[voting$treat_bin == 0], na.rm = TRUE)
cat("\nQ3: Mean (voted)\n")

##
## Q3: Mean (voted)

cat("  Treated  =", round(treat_mean, 4), "\n")

##   Treated  = 0.3779
```

```
cat("  Control  =", round(control_mean, 4), "\n")

##   Control  = 0.2966
```

Treated = 0.3779 Control = 0.2966

The results indicate that voters who received the message were more likely to vote on the election day than voters who did not receive them. In other words, the treatment group and the control group showed noticeable differences in the outcome variable, which measured whether they voted.

4. R codes

```
treat_group   <- voting[voting$treat_bin == 1, ]
control_group <- voting[voting$treat_bin == 0, ]
```

5. R codes for this process:

```
mean_treat_birth   <- mean(treat_group$birth,   na.rm = TRUE)
mean_control_birth <- mean(control_group$birth, na.rm = TRUE)

cat("\nQ5: Mean(birth year)\n")

## 
## Q5: Mean(birth year)

cat("  Treated  =", round(mean_treat_birth,   2), "\n")

##   Treated  = 1956.15

cat("  Control  =", round(mean_control_birth, 2), "\n")

##   Control  = 1956.19
```

Average birth year for the treatment and control group respectively:

Treated = 1956.15 Control = 1956.19

6. R codes for calculating estimated average treatement effect:

```
ATE <- treat_mean - control_mean

cat("\nQ6: Estimated ATE (Treated - Control)\n")

## 
## Q6: Estimated ATE (Treated - Control)

cat("  ATE =", round(ATE, 4), "\n")

##   ATE = 0.0813
```

ATE = 0.0813

The derivative of the estimated average treatment effect indicates that the treatment group was approximately 8% more likely to cast a ballot on the election day than the control group.

7, To claim that the findings can be generalized to the entire U.S. population, the sample must be representative of that population—ideally through random selection of individuals from the national population. In addition, the treatment and control groups should be comparable so that differences in outcomes can be attributed to the treatment rather than to pre-existing differences. The nearly identical average birth years of the two groups suggest that this comparability holds, at least in terms of age.