Bootstrap

2024-05-22

```
data <- read_csv("data.csv", show_col_types = F)
```

```
vaccine <- data %>%
  filter(Test == "Vaccine")

placebo <- data %>%
  filter(Test == "Placebo")

prop_vaccine <- vaccine$COVID[1] / vaccine$No_COVID[1]
prop_placebo <- placebo$COVID[1] / placebo$No_COVID[1]
observed_diff <- prop_placebo - prop_vaccine

n_vaccine <- vaccine$COVID + vaccine$No_COVID
n_placebo <- placebo$COVID + placebo$No_COVID

observed_diff <- prop_placebo - prop_vaccine
```

```
n_bootstrap <- 10000
bootstrap_diffs <- numeric(n_bootstrap)
set.seed(123)

for (i in 1:n_bootstrap) {
  vaccine_sample <- sample(c(0, 1), size = n_vaccine, replace = TRUE, prob = c(1 - prop_vaccine, prop_va
  placebo_sample <- sample(c(0, 1), size = n_placebo, replace = TRUE, prob = c(1 - prop_placebo, prop_pl

  prop_vaccine_boot <- mean(vaccine_sample)
  prop_placebo_boot <- mean(placebo_sample)

  bootstrap_diffs[i] <- prop_placebo_boot - prop_vaccine_boot
}

# Calculate overall confidence interval
overall_ci <- quantile(bootstrap_diffs, c(0.025, 0.975))

# Calculate confidence intervals for each bootstrap sample
ci_data <- data.frame(
  Iteration = 1:n_bootstrap,
  Lower = numeric(n_bootstrap),
  Upper = numeric(n_bootstrap)
)

for (i in 1:n_bootstrap) {
  sample_diffs <- sample(bootstrap_diffs, n_bootstrap, replace = TRUE)
  ci_data$Lower[i] <- quantile(sample_diffs, 0.025)
  ci_data$Upper[i] <- quantile(sample_diffs, 0.975)
```

```
}
```

```
# Subset for plotting (e.g., plot every 500th iteration)
plot_data <- ci_data[seq(1, n_bootstrap, by = 200), ]

# Plot horizontal segmented line graph with overall confidence interval
ggplot(plot_data, aes(y = Iteration)) +
  geom_vline(xintercept = observed_diff, linetype = "solid", color = "red") +
  geom_segment(aes(yend = Iteration, x = Lower, xend = Upper), color = "blue") +
  geom_vline(xintercept = overall_ci[1], linetype = "dashed", color = "red") +
  geom_vline(xintercept = overall_ci[2], linetype = "dashed", color = "red") +
  labs(title = "Confidence Intervals of Bootstrap Samples",
       y = "Iteration",
       x = "Confidence Interval") +
  scale_x_continuous(limits = c(0.007, 0.011),
                     breaks = seq(0.007, 0.011, by = 0.001)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), # Center the plot title
        axis.text.y = element_text(hjust = 1)) # Angle the y-axis labels for readability
```