Delta Method for Testing Relative Differences

2023-08-17

Typical T-Test

The typical t-test finds the statistical significance of *absolute* differences of random variables.

$$T = \frac{\bar{X_2} - \bar{X_1}}{\sqrt{Var(\bar{X_1}) + Var(\bar{X_2})}}$$

$$Var(X) = \frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2$$

Relative Differences

For relative differences ($\frac{\bar{X_2} - \bar{X_1}}{\bar{X_1}}$), the variance needs to be adjusted. One way to see what adjustment needs to be made is to look at the relative difference as a ratio metric (e.g. clicks per pageviews). The numerator and denominator are both random variables. Thus, it's incorrect to simply divide the variance in the above equation by $\bar{X_2}$. The correct way to do it is using the Delta Method.

Delta Method Equation

The variance of the ratio of two random variables can be estimated by:

$$Var(\frac{\bar{Y}}{\bar{X}}) \approx \frac{\mu_y^2}{\mu_x^2}\left(\frac{Var(Y)}{\mu_y^2} - 2\frac{Cov(X,Y)}{\mu_x \mu_y} + \frac{Var(X)}{\mu_x^2}\right)$$

For the relative difference between two independent random variables, this equation can be simplified to

$$Var(\frac{\bar{X_2} - \bar{X_1}}{\bar{X_1}}) \approx \frac{Var(\bar{X_2})}{\bar{X_1}^2} + \frac{Var(\bar{X_1})\bar{X_2}^2}{\bar{X_1}^4}$$

Examples

Helper functions:

```r
library(scales)
library(ggplot2)
library(dplyr)

# Function to run t-test given a metric, variance, alpha and degrees of freedom
#' @param metric test metric for significance test (e.g. absolute or relative difference between means)
#' @param var variance of test metric
#' @param alpha false positive rate
#' @param dof degrees of freedom
#' @returns results of t tests with p-value, lower and upper confidence intervals
t_test <- function(metric, var, alpha, dof) {
  test_stat <- metric / sqrt(var)
```

```r
  p_value <- 2 * (1 - pt(abs(test_stat), dof))
  width <- qt(1 - alpha/2, df = dof) * sqrt(var)
  confidence_interval <- c(metric - width, metric + width)
  results <- list(
    p_value = p_value,
    LCL = metric - width,
    UCL = metric + width)
  return(results)
}

# Function to run t-tests on relative and absolute difference of test metric
#' @param experiment list of experiment parameters including number of observations, baseline metric an
#' @param alpha false positive rate
#' @returns list of results for t tests of absolute difference (with unadjusted variance) and relative
compare_tests <- function(experiment, alpha) {
  mu_1 <- experiment$mu_1
  mu_2 <- mu_1*(1+experiment$lift)
  n_1 <- experiment$n_1
  n_2 <- experiment$n_2

  # If variance is not already given, assume binomial distribution and
  # calculate variance. Otherwise, use stored variance
  if (is.null(experiment$var_1)) {
    var_1 <- mu_1*(1-mu_1)
    var_2 <- mu_2*(1-mu_2)
  } else {
    var_1 <- experiment$var_1
    var_2 <- experiment$var_2
  }

  # Get degrees of freedom
  dof <- ((var_2 / n_2) + (var_1/n_1))**2 / (var_2**2/(n_2**2*(n_2-1)) + var_1**2/(n_1**2*(n_1-1)))

  # Variance of absolute difference
  unadjusted_var <- (var_1/n_1) + (var_2/n_2)

  # Variance of relative difference
  adjusted_var <- (var_2 / (mu_1**2*n_2)) + (var_1 * mu_2**2 / (mu_1**4*n_1))

  # Run t-tests
  absolute_results <- t_test(metric = mu_2 - mu_1, unadjusted_var, alpha, dof)
  relative_results <- t_test(metric = (mu_2/mu_1) - 1, adjusted_var, alpha, dof)

  results <- list(absolute = absolute_results, relative = relative_results)
  return(results)
}
```

Simulations for two AB tests:

Test 1

```r
test_1_lifts <- seq(.065, .075, .001)
test_1_absolute_p_values <- c()
```

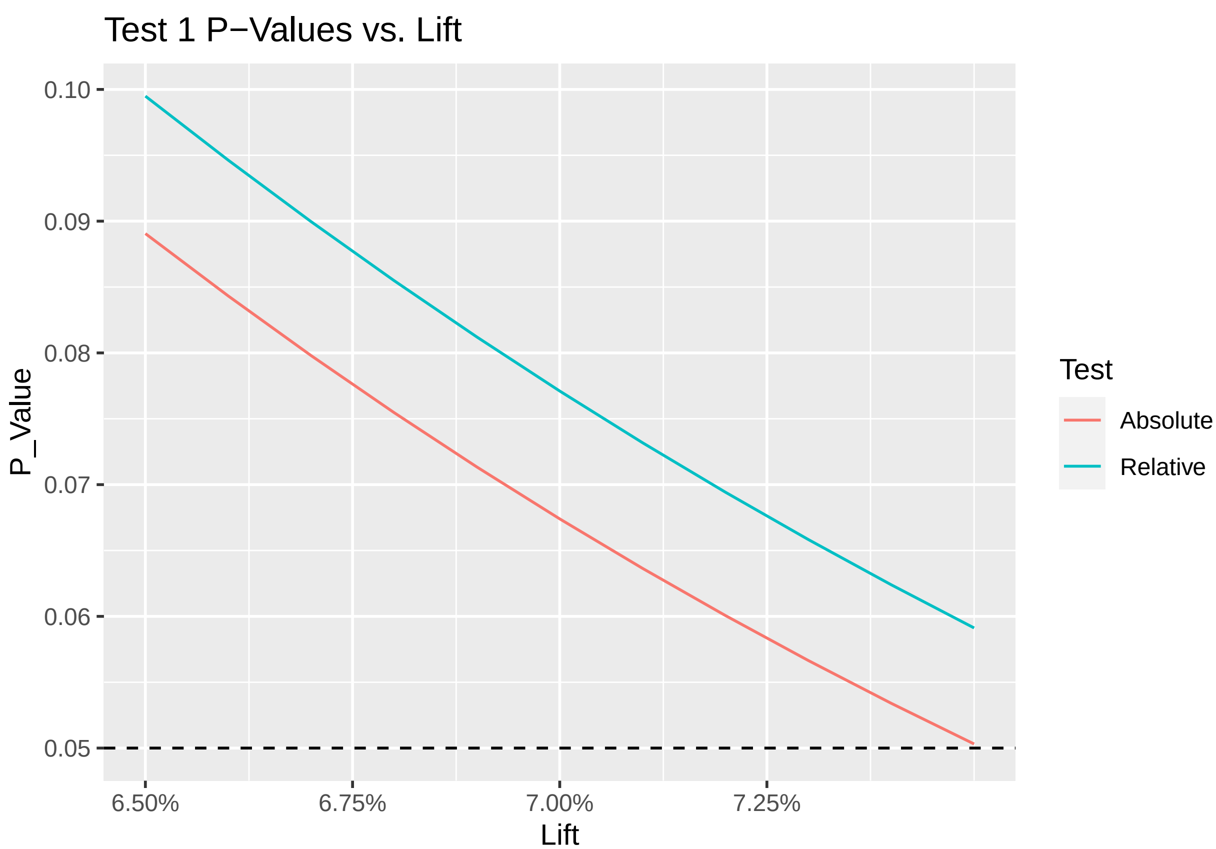
```r
test_1_relative_p_values <- c()

# Get results from t-tests for absolute and relative differences
for (lift in test_1_lifts) {
  test_1 <- list(n_1 = 700000, n_2 = 710000, mu_1 = 0.002, lift = lift)
  results <- compare_tests(test_1, alpha)
  test_1_absolute_p_values <- c(test_1_absolute_p_values, results$absolute$p_value)
  test_1_relative_p_values <- c(test_1_relative_p_values, results$relative$p_value)

}

# Prepare data for plotting
test_1_plot_data <- data.frame("Test" = c(rep("Absolute", length(test_1_lifts)),
                                          rep("Relative", length(test_1_lifts))),
                        "Lift"= rep(test_1_lifts, 2),
                        "P_Value" = c(test_1_absolute_p_values,
                                      test_1_relative_p_values))

# Plot p-value vs. lifts for relative and absolute tests
ggplot(test_1_plot_data, aes(x = Lift, y = P_Value, color = Test)) + geom_line() +
  geom_hline(yintercept = alpha, linetype = "dashed") +
  scale_x_continuous(labels = scales::percent,
                     breaks = c(.065, .0675, .07, .0725)) +
  ggtitle("Test 1 P-Values vs. Lift")
```



Test 2

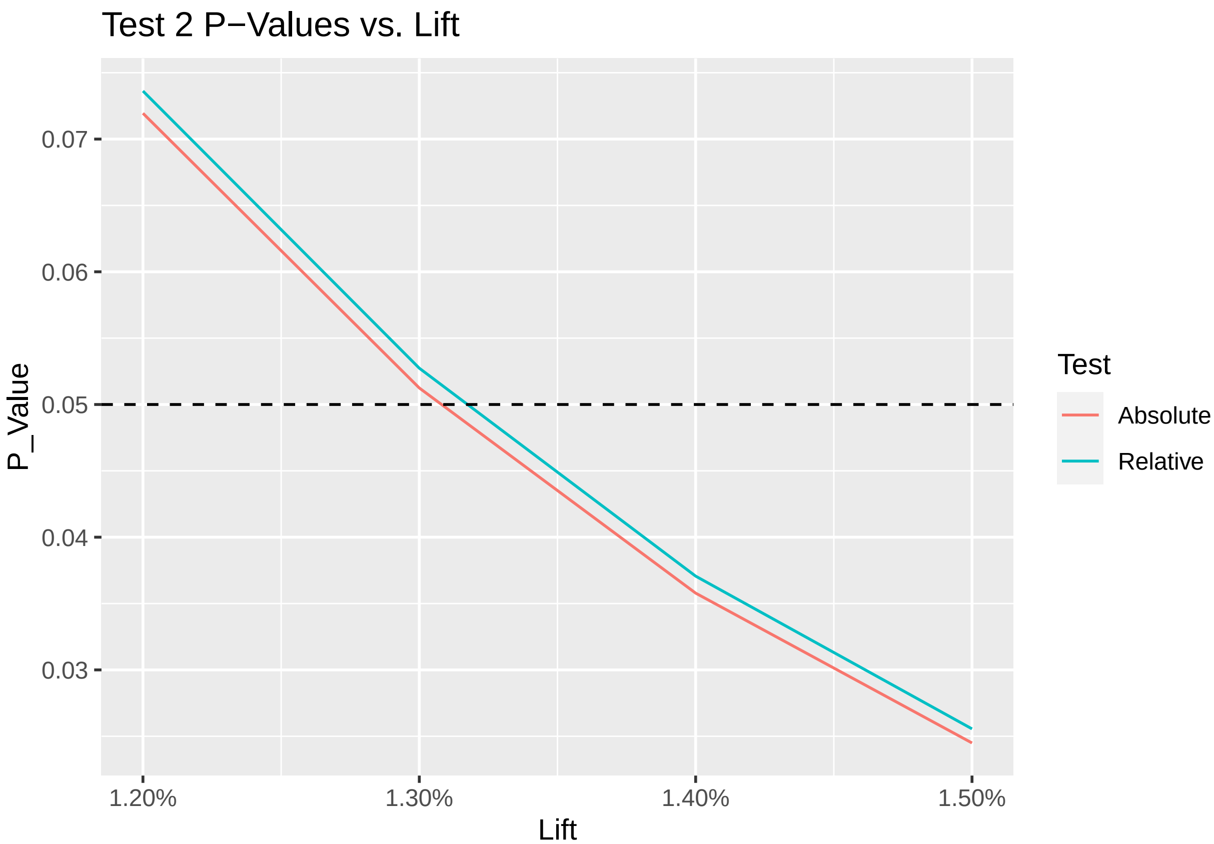
```r
test_2_lifts <- seq(.012, .015, .001)
test_2_absolute_p_values <- c()
test_2_relative_p_values <- c()

# Get results from t-tests for absolute and relative differences
for (lift in test_2_lifts) {
  test_2 <- list(n_1 = 46000, n_2 = 46700, mu_1 = 0.33, lift = lift, var_1 = 0.33^2,
                 var_2 = 0.34^2)
  results <- compare_tests(test_2, alpha)
  test_2_absolute_p_values <- c(test_2_absolute_p_values, results$absolute$p_value)
  test_2_relative_p_values <- c(test_2_relative_p_values, results$relative$p_value)

}

# Prepare data for plotting
test_2_plot_data <- data.frame("Test" = c(rep("Absolute", length(test_2_lifts)),
                                          rep("Relative", length(test_2_lifts))),
                               "Lift"= rep(test_2_lifts, 2),
                               "P_Value" = c(test_2_absolute_p_values,
                                             test_2_relative_p_values))

# Plot p-value vs. lifts for relative and absolute tests
ggplot(test_2_plot_data, aes(x = Lift, y = P_Value, color = Test)) + geom_line() +
  geom_hline(yintercept = alpha, linetype = "dashed") +
  scale_x_continuous(labels = scales::percent) +
  ggtitle("Test 2 P-Values vs. Lift")
```



Further Discussion Topics

- Confidence intervals - the Delta Method widens the confidence intervals more for bigger lifts
- Negative lifts - smaller variance with the Delta Method
- How do different adjustment methods for multiple comparisons affect the difference in p-values?