Comparative Analysis of Proportion χ^2 and Logistic Regression Wald Tests

Ephrata Getachew

2024-05-06

Introduction

In statistical analyses in fields such as illness (epidemiology) and many other investigative contexts, comparing two rates is a fundamental problem. This analysis is important for understanding outcome differences across varied conditions or groups. The article Comparing Two Tests for Two Rates by Chunpeng Fan, Lin Wang, and Lynn Wei (2017) offers a comparison between two statistical tests used for this purpose: the proportion χ^2 test and the logistic regression Wald test.

Although these tests are asymptotic equivalent under the null hypothesis that the rates being compared, under alternative hypotheses, they may lead to different inferences. This difference might cause concerns about the power of the tests, which is the ability to correctly reject the null hypothesis when it is false. The article compares these tests under various conditions and proves the proportion χ^2 test is superior to the logistic regression Wald test.

This report will summarize the key findings of the article and connect these findings to the statistical concepts we learned in class on hypothesis testing, error types, and power. Also, expand upon the concepts to enhance our understanding and go beyond to confirm the superior performance of the proportion χ^2 test under different sample sizes and the magnitude of rate differences on the performance of the two tests.

Summary of the Article

Proportion χ^2 Test is a statistical test used to determine if there is a relationship between two categorical variables without the need for adjusting for covariate - variables that could influence the outcome but are not the primary variables of interest (UCLA, n.d).

In class, we discussed calculating Z values for hypothesis testing based on pooled proportions when comparing two independent samples. The Z_{Prop} formula (1) derived in the article uses the pooled proportion $\hat{p}$, which we learned to calculate as $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$.

For the proportion χ^2 test, the test statistic is derived as

$$Z_{\text{Prop}} = \frac{|\hat{p}_2 - \hat{p}_1|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (1)$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions, and $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ is the pooled proportion.

The null hypothesis (H_0) is there is no difference between the two proportions.

$$(H_0: \ p_1 = p_2)$$

The alternative hypothesis (H_a) is there is a difference between the two proportions.

$$(H_a: \ p_1 \neq p_2)$$

The corresponding p-value is then

$$p-value_{Prop} = 2(1 - \Phi(Z_{Prop})$$

A p-value below α level would lead to rejecting the null hypothesis, indicating a statistically significant difference between the two proportions. In contrast, a p-value above α level indicates insufficient evidence to reject the null hypothesis, suggesting there may not be a significant difference between the groups.

Logistic Regression Wald Test: logistic regression is a predictive analysis used to explain the relationship between one dependent binary variable (a variable that can only take on two distinct values), the response and a set of explanatory variable also referred to as an independent variable or predictor (Friendly & Meyer, 2015). It can incorporate both discrete and continuous covariates. The logistic regression yields coefficients, which are then used to compute the Wald test statistic $Z_{Logistic}$ using formula (2).

$$Z_{Logistic} = \frac{|\hat{\beta}_1|}{\sqrt{\text{Var}(\hat{\beta}_1)}},$$

where $\hat{\beta}_1$ is the estimated coefficient for a predictor variable in the logistic regression model. It quantifies the effect of that predictor on the log odds of the dependent variable, adjusting for other predictors in the model. Log odds is the natural logarithm of the odds ratio between the probability p and $(1 - p)$ of the two proportions (Friendly & Meyer, 2015).

This follows the structure of a Z-value calculation we have used for hypothesis testing. In this formula, $\hat{\beta}_1$ acts like the mean difference we see in a Z-test, and $\text{Var}(\hat{\beta}_1)$ represents the estimated variance of this coefficient, similar to how we compute the standard error in a Z-test.

The null hypothesis (H_0) is there is no effect on the group variable (predictor) on the outcome.

$$(H_0: \beta_1 = 0)$$

The alternative hypothesis (H_a) is there is effect on the group variable on the outcome.

$$(H_a: \beta_1 \neq 0)$$

$Z_{Logistic}$ can be explicitly written as

$$Z_{Logistic} = \frac{\left|\log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) - \log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right)\right|}{\sqrt{n_1\hat{p}_1(1-\hat{p}_1)^{-1} + n_2\hat{p}_2(1-\hat{p}_2)^{-1}}}, \qquad (2)$$

The corresponding p-value is then

$$p\text{-value}_{Logistic} = 2\left(1 - \Phi(Z_{Logistic})\right)$$

A p-value below α level would lead to rejecting the null hypothesis, indicating a statistically significant effect of the group variable on the outcome. In contrast, a p-value above α level indicates insufficient evidence to reject the null hypothesis, suggesting that the group variable may not significantly impact the outcome.

Under the null hypothesis $H_0: p_1 = p_2$, Z_{Prop} and $Z_{Logistic}$ are asymptotically equivalent because both test statistics measure the difference between two population proportions. By the central limit theorem, their distributions become increasingly normal as the sample size increases. This convergence results in both tests

still appropriately control the Type I error rate at the desired nominal level, and the difference between the test statistics Z_{Prop} and Z_{Logistic} approaches zero:

$$Z_{\mathrm{Prop}} - Z_{\mathrm{Logistic}} \to 0 \text{ under } H_0 : p_1 = p_2,$$

However, the two tests may lead to different inferences under the alternative hypothesis $H_A : p_1 \neq p_2$. This observation is supported by the study's theoretical derivations and numerical simulation.

Theoretical Derivations

"Theorem 1. For any $n_1, n_2 > 0$ and $0 < \hat{p}_1, \hat{p}_2 < 1$, when $\hat{p}_1 \neq \hat{p}_2$" (Fan, Wang, & Wei, 2017),

Assuming that the two groups come from the same population, the proportion χ^2 test is based on the difference between the two proportions and scaled by the variance. So, when the true proportions differ, this test is more sensitive and thus responds more to that difference. However, the logistic regression Wald test estimates parameters in a model that fits the log odds of these proportions (Friendly & Meyer, 2015). The variance is accounted for by the model's overall error term, so it tends to be less sensitive to small changes. Thus,

$$Z_{\mathrm{Prop}} > Z_{\mathrm{Logistic}}$$

A higher Z-score results in a smaller p-value, which is statistical evidence against the null hypothesis ($H_0 : p_1 = p_2$). Since $Z_{\mathrm{Prop}} > Z_{\mathrm{Logistic}}$, the p-value for the proportion χ^2 will be lower, which means it is more likely to indicate a statistically significant difference when one exists.

$$p\text{-value}_{\mathrm{Prop}} < p\text{-value}_{\mathrm{Logistic}}$$

Because the proportion test results in higher Z-score and smaller p-value, it is more likely to exceed the cutoff needed to reject the null hypothesis. This results in the proportion χ^2 test having higher power than the logistic regression Wald test. Thus, the proportion χ^2 test is more reliable for detecting true differences between the proportions, which is a reduced chance of committing Type II errors.

$$Power_{Prop} > Power_{Logistic}$$

Numerical Simulations:

The authors did a simulation study to visually compare the Z-score and Power generated by each test under scenarios where $n_1 = n_2 = 30$, 50, or 100; $\hat{p}_1 = 0.1$, 0.3, or 0.5; and $\hat{p}_2$ between 0 and 1 by 0.01. For each selection of $n_1 = n_2$, $\hat{p}_1$, and $\hat{p}_2$, the proportion χ^2 test statistic Z_{Prop} and the logistic regression Wald test statistic Z_{Logistic} were calculated using formulas (1) and (2), respectively.
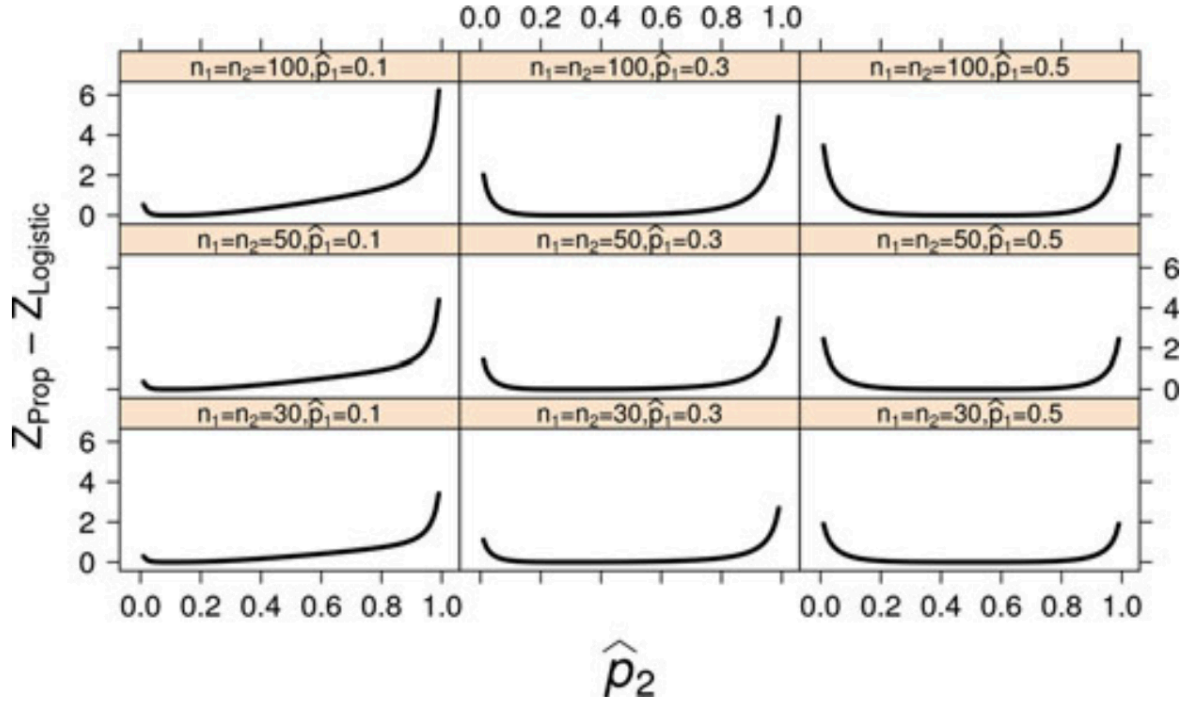
Figure 1 (Fan, Wang, & Wei, 2017) presents a numerical comparison between Z_{Prop} and $Z_{Logistic}$.

In all scenarios, the plots show that $Z_{Prop} - Z_{Logistic}$ is greater than zero, mainly when the difference between $\hat{p}_1$ and $\hat{p}_2$ is relatively large. This suggests that the proportion χ^2 test tends to generate larger test statistics compared to the logistic regression Wald test.
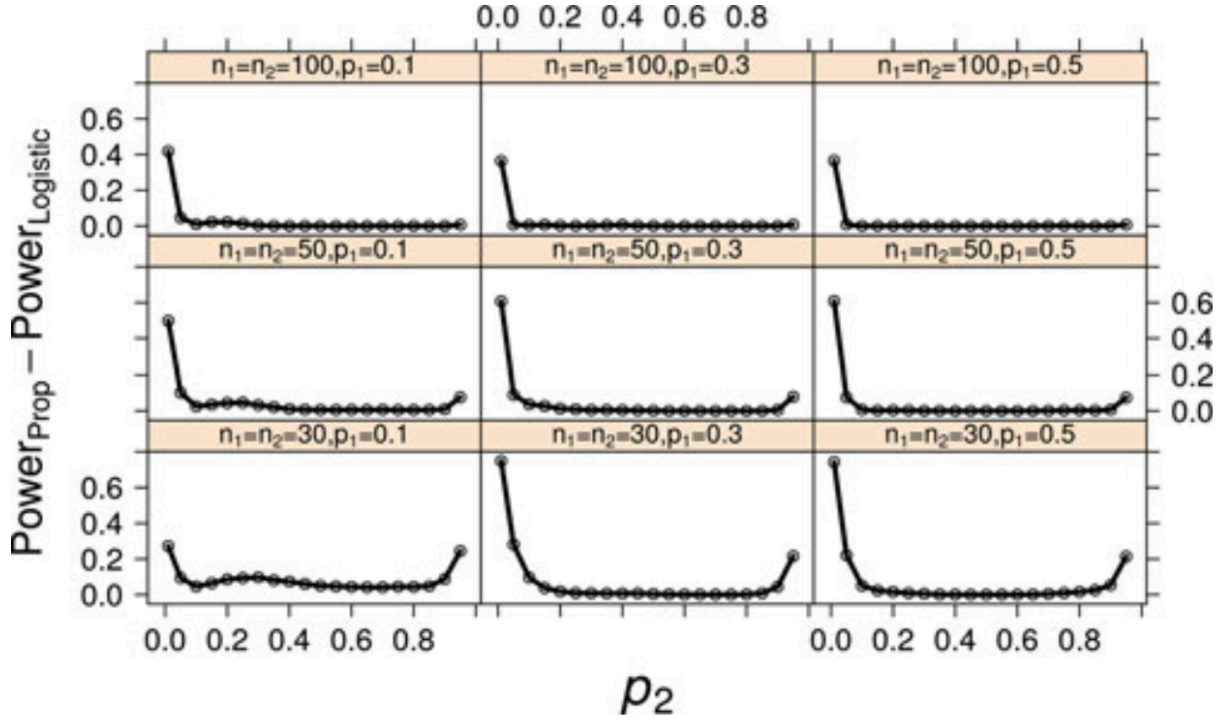
Figure 2 (Fan, Wang, & Wei, 2017) presents a numerical comparison between $Power_{Prop}$ and $Power_{Logistic}$.

In all scenarios, the plots show that the proportion χ^2 test has higher power than the logistic regression test, mainly when p_1 or p_2 approaches 0 or 1. This suggests that the proportion χ^2 test tends to have larger power compared to the logistic regression Wald test.

Thus, both figure 1 and 2 further support Theorem 1 that the proportion χ^2 test is always superior to the logistic regression Wald test.

Example

An example from a 12-week clinical trial is used to illustrate the applications of the two tests, focusing on comparing the response rate of an active treatment and a placebo. Both tests are employed to determine whether there are statistically significant differences between the treatment and placebo groups, using test statistics and p-values derived from formulas (1) and (2) and software such as R and SAS. The results of all the three methods are reported in Table 1 (Fan, Wang, & Wei, 2017).

In this illustrative study,

The null hypothesis (H_0) is there is no difference in the treatment effects between the active treatment group (p_1) and the placebo group (p_2).

$$(H_0 : p_1 = p_2)$$

The alternative hypothesis (H_a) is there is a difference in the treatment effects between the active treatment group (p_1) and the placebo group (p_2).

$$(H_a : p_1 \neq p_2)$$

5

	Proportion χ^2 test			Logistic regression Wald test		
	Method	Z_{Prop}	p-value$_{\text{Prop}}$	Method	Z_{Logistic}	p-value$_{\text{Logistic}}$
Equation	Equation (1)	1.9712	0.0487	Equation (2)	1.9490	0.0513
R	prop.test	1.9712	0.0487	glm	1.9490	0.0513
SAS®	PROC FREQ	1.9712	0.0487	PROC GENMOD	1.9490	0.0513

Table 1 (Fan, Wang, & Wei, 2017), shows the test statistics and the corresponding p-value by the proportion χ^2 test and the logistic regression Wald test.

The proportion χ^2 test and the logistic regression Wald test yields different p-values of 0.0487 and 0.0513, respectively. The proportion χ^2 test has a p-value below the cutoff ($\alpha = 0.05$), suggesting a significant treatment effect. This will allow further investigation of additional endpoints in the study. In contrast, the logistic regression Wald test has a p-value above the cutoff, suggesting insufficient evidence to conclude the treatment effect at the significant $\alpha = 0.05$ level, which could end further exploration. This scenario shows how minor differences can lead to different decision-making, emphasizing the importance of selecting appropriate statistical tests. Thus, in such borderline cases, the superiority of the proportion χ^2 test may have a greater advantage.

Go beyond

To extend the analysis presented in the article, we conducted a simulation study to assess the performance of the proportion χ^2 test and the logistic regression Wald test under unequal sample sizes $n_1 \neq n_2$ and rate conditions.

Simulation Approach

Using R, we did graphical analyses for three unequal different sample sizes: (n_1=40, n_2=60), (n_1=80, n_2=150), and (n_1=120, n_2=200). Within each scenario, $\hat{p}_1$ was set to 0.1, 0.3, and 0.5. For each combination of n_1, $\hat{p}_1$, and a range of values for $\hat{p}_2$ from 0 to 1, $Z_{\text{Prop}} - Z_{\text{Logistic}}$ was computed using formula (1) and (2) respectively. Figure 3 illustrates how the difference in test statistics varied with the $\hat{p}_2$ change.
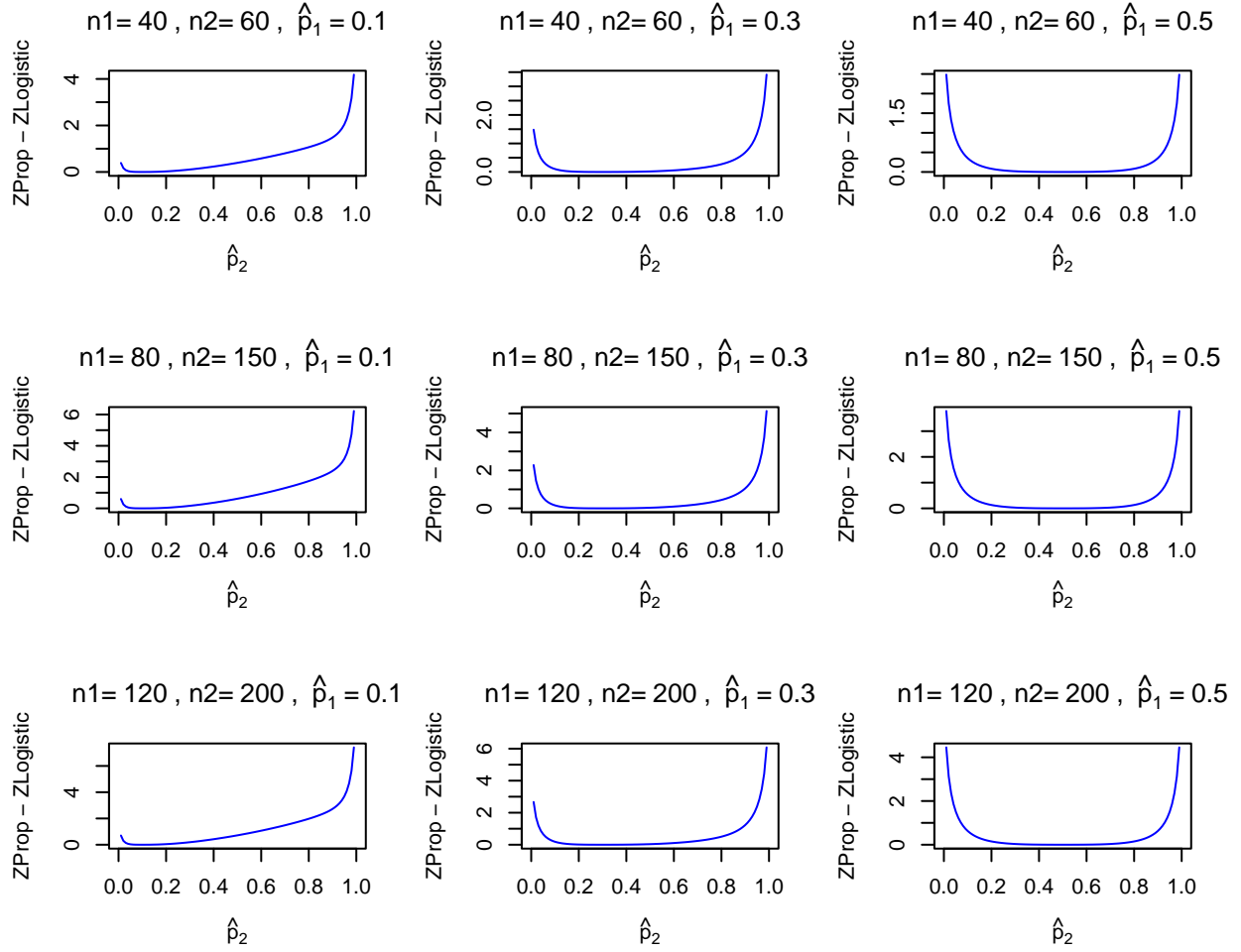
Figure 3: Numerical comparisons between Z_{Prop} and Z_{Logistic}: $Z_{\text{Prop}} - Z_{\text{Logistic}} > 0$ in all cases, and the difference is more visible when the difference between $\hat{p}_1$ and $\hat{p}_2$ is relatively large.

Across all graphs, the curve representing the Z-value difference is greater than zero, which means that the $Z_{\text{Prop}} > Z_{\text{Logistic}}$, suggesting that it is more sensitive in detecting differences between two proportions. The difference is more visible when $\hat{p}_1$ and $\hat{p}_2$ are relatively large, noticeable as $\hat{p}_2$ approaches 0 or 1. This is likely due to the property of logistic regression Wald test. It may not perform well because it involves transformations (log-odds) under extreme probability values.

An additional simulation was conducted using R, a statistical computing environment(R Core Team, 2022) to calculate the power of the proportion χ^2 test and the logistic regression Wald test, aiming to go beyond the existing findings by exploring scenarios with uneven sample sizes and fixed rates, ($n_1 = 40, 70, 90$), ($n_2 = 60, 100, 150$), ($\hat{p}_1 = 0.1$), and ($\hat{p}_2 = 0.4$). We run 10,000 simulations to estimate the rejection rates for both tests at a significance level of α=0.05.

For each iteration, random samples were generated for each group using a binomial distribution, reflecting the expected number of successes based on their respective rates. For the proportion χ^2, we used the prop.test function to compute the test statistic based on the difference between the observed proportions of the two groups. Then, we fitted a logistic regression model using the glm (generalized linear model) function, where the response was the binary outcome of success or failure, and the predictor was group membership (indicating whether individuals belonged to a specific group or category). This model provides a logistic regression Wald test statistic from the model's coefficients, testing the null hypothesis about the effect of

group membership on the probability of success (Friendly & Meyer, 2015).

Table 2: Power Comparison of Proportion χ^2 and Logistic Regression Wald Test

	$Power_{Prop}$	$Power_{Logistic}$	$Power_{Prop} - Power_{Logistic}$
n1= 40 n2= 60	0.9392	0.9236	0.0156
n1= 40 n2= 100	0.9698	0.9533	0.0165
n1= 40 n2= 150	0.9854	0.9690	0.0164
n1= 70 n2= 60	0.9843	0.9803	0.0040
n1= 70 n2= 100	0.9973	0.9963	0.0010
n1= 70 n2= 150	0.9978	0.9974	0.0004
n1= 90 n2= 60	0.9914	0.9907	0.0007
n1= 90 n2= 100	0.9993	0.9992	0.0001
n1= 90 n2= 150	0.9999	0.9997	0.0002

Results in Table 2 confirm that $Power_{Prop} - Power_{Logistic} > 0$ in all cases. As the sample size increases, the tests' power increases, and the difference between the two tests decreases. This is consistent with our class material, which states that larger samples provide a more reliable estimate and are more likely to detect true effects if they exist. Again, from the results we can confirm that the proportion χ^2 test has higher power than the logistic regression Wald test.

Conclusion

In conclusion, this comparative analysis showed how the proportion χ^2 test and the logistic regression Wald test handle the test of two proportions under different scenarios. Even though the logistic regression Wald Test has advantage in incorporating discrete and continuous covariates, it does not always yield the power necessary to distinguish between two proportions effectively. As detailed in the report, the proportion χ^2 test has superior power to the logistic regression Wald test, particularly under the alternative hypothesis $(H_a : p_1 \neq p_2)$. This is supported by theoretical conclusions and numerical simulations conducted in this project.

Apendix

Z Statistics Comparison for Proportion χ^2 and Logistic Regression Wald Tests (Figure 3)

```r
# Set seed for reproducibility
set.seed(123)

# Define the number of simulations
num_simulations <- 10000

# Define sample sizes and proportion values for the simulation
n1_values <- c(40, 80, 120)
n2_values <- c(60, 150, 200)
p1_values <- c(0.1, 0.3, 0.5)
p2_values <- seq(0, 1, length.out = 100)

# Setup a plotting layout
layout(matrix(1:9, nrow = 3, byrow = TRUE))

# Define a function to compute Z statistics for comparing two proportions
compute_z_stats <- function(p1, p2, n1, n2, num_simulations) {
  # Calculate pooled proportion
  p <- (p1 * n1 + p2 * n2) / (n1 + n2)
  # Calculate Z statistic for proportion test (formula(1))
  z_prop <- (abs(p2 - p1)) / sqrt(p * (1 - p) * (1/n1 + 1/n2))
  # Calculate Z statistic for logistic regression test (formula(2))
  z_logistic <- (abs(log(p2 / (1 - p2)) - log(p1 / (1 - p1)))) /
              sqrt(1/(n1 * p1 * (1 - p1)) + 1/(n2 * p2 * (1 - p2)))
  # Return the difference between the two Z statistics
  return(z_prop - z_logistic)
}

# Loop over each combination to perform simulations
for (i in seq_along(n1_values)) {
  for (j in seq_along(p1_values)) {
    n1 <- n1_values[i]
    n2 <- n2_values[i]
    p1 <- p1_values[j]

    # Calculate Z statistics difference for each p2
    z_diff <- sapply(p2_values, compute_z_stats, p1 = p1, n1 = n1, n2 = n2)

    # Plot the differences in Z statistics
    plot(p2_values, z_diff, type = 'l', col = 'blue',
         xlab = expression(paste(hat(p)[2])),
         ylab = "ZProp - ZLogistic",
         main = bquote("n1=" ~ .(n1) ~ ", n2=" ~ .(n2) ~ ", " ~
                       hat(p)[1] ~ "=" ~ .(p1)))
  }
}
```

Comparison of Power of Proportion χ^2 and Logistic Regression Wald Tests (Table 2)

```r
#Load library
library(knitr)

# Set seed for reproducibility
set.seed(123)

# Define the number of simulations
num_simulations <- 10000

# Define sample sizes and proportion values for the simulation
n1_values <- c(40, 70, 90)
n2_values <- c(60, 100, 150)
p1_values <- 0.1
p2_values <- 0.4
alpha <- 0.05

# Define a function to simulate one iteration of the experiment
simulate_one <- function(n1, n2, p1, p2) {
  # Generate random data based on binomial distributions for both groups
  r1 <- rbinom(1, n1, p1)  # Number of successes in the first group
  r2 <- rbinom(1, n2, p2)  # Number of successes in the second group

  # Method adapted from R Core Team. (2022). R Documentation: prop.test function
  prop_test_result <- prop.test(c(r1, r2), c(n1, n2), correct = FALSE)
  # Check if the null hypothesis is rejected
  prop_reject <- prop_test_result$p.value < alpha

  # Prepare data for logistic regression
  data <- data.frame(success = c(rep(1, r1), rep(0, n1 - r1),
                                 rep(1, r2), rep(0, n2 - r2)),
                    group = factor(c(rep(1, n1), rep(2, n2))))
  # Method adapted from (Friendly & Meyer, 2015)
  # Fit a logistic regression model
  glm_fit <- glm(success ~ group, family = binomial, data = data)
  #  Check if the null hypothesis is rejected
  logistic_reject <- summary(glm_fit)$coefficients[2,4] < alpha

  # Return the results of both tests
  return(c(prop_reject, logistic_reject))
}
# Method adapted from (DataMentor., n.d.)
# Initialize a matrix to store power
results_matrix <- matrix(0, nrow = length(n1_values) * length(n2_values),
                        ncol = 3,
                        dimnames = list(
                            paste("n1=", rep(n1_values, each=length(n2_values))
                            , "n2=", rep(n2_values, times=length(n1_values))),
                             c("$Power_{Prop}$", "$Power_{Logistic}$",
                                "$Power_{Prop} - Power_{Logistic}$")
                        ))
```

```r
# Loop over each combination of n1 and n2
for (index in 1:(length(n1_values) * length(n2_values))) {
# Each value of n1 is paired with every value of n2 before moving to the next n1
    n1_index <- ((index - 1) %/% length(n2_values)) + 1
    # Cycles through n2_values for each value of n1
    n2_index <- ((index - 1) %% length(n2_values)) + 1
    # Select the n1 and n2 values based on the indices calculated above
    n1 <- n1_values[n1_index]
    n2 <- n2_values[n2_index]

    # Run simulation for the current combination of n1 and n2
    results <- replicate(num_simulations, simulate_one(n1, n2, p1_values[1],
                                            p2_values[1]))
    # Calculate the power of each test
    power_prop <- mean(results[1,])
    power_logistic <- mean(results[2,])
    difference <- power_prop - power_logistic

    # Store results
    results_matrix[index, ] <- c(power_prop, power_logistic, difference)
}

# Convert results_matrix to a data frame
results_df <- as.data.frame(results_matrix)

kable(results_df, caption = "Table 2. Power Comparison of Proportion $\\chi^2$
      Test and Logistic Regression Wald Test", align = 'c')
```

References

DataMentor. (n.d.). *R matrix (with examples)*. Retrieved May 4, 2024, from https://www.datamentor.io/r-programming/matrix/

Fan, C., Wang, L., & Wei, L. (2017). Comparing two tests for two rates. *The American Statistician*, 71(3), 275-281. https://doi.org/10.1080/00031305.2016.1246263

Friendly, M., & Meyer, D. (2015). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. CRC Press.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

R Core Team. (2022). R documentation: prop.test function. Retrieved from https://www.rdocumentation.org/packages/stats/versions/4.2.1/topics/prop.test

UCLA: Statistical Consulting Group. (n.d.). What statistical analysis should I use? Statistical analyses using SPSS. Retrieved May 4, 2024, from https://stats.oarc.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-use-statistical-analyses-using-spss/