2nd-Year Master of Statistics and Data Science Computer Intensive Methods: Final projects (2024/2025 Project 1

Name: Edmond SACLA AIDE (2159278) Lecturer: Prof. Ziv

2025-01-26

Project 1

In this project, in all questions, we focused on the nassCDS data which is a US data from police-reported car crashes (1997-2002) in which there is a harmful event (people or property). Data are restricted to front-seat occupants, include only a subset of the variables recorded. More information about the dataset can be found using the following link: https://www.rdocumentation.org/packages/DAAG/versions/1.22/topics/nassCDS. The data is a part of the DAAG R package. To get an access to the data you first need to install the package. The list of variables names is shown below.

Part 1

Let Y_i be an indicator variable which takes the value of 1 if an occupant died in an accident (the variable *dead*) and zero otherwise and X_i be the age of occupant in years (the variable *ageOFocc*). We consider the following GLM:

$g(P(Y_i = 1)) = \beta_0 + \beta_1 X_i$

Question 1.1

As we are dealing with binary outcome, the model is estimated by a GLM with a binomial family.

```
##
## Call:
## glm(formula = dead ~ ageOFocc, family = "binomial", data = nassCDS)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.907983   0.072013  -54.27   <2e-16 ***
## ageOFocc     0.021183   0.001484   14.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9610.0  on 26062  degrees of freedom
## Residual deviance: 9418.2  on 26061  degrees of freedom
## AIC: 9422.2
##
## Number of Fisher Scoring iterations: 6
```

The age of the occupant influent significantly the probability of death in an accident at a significant level of 5%.

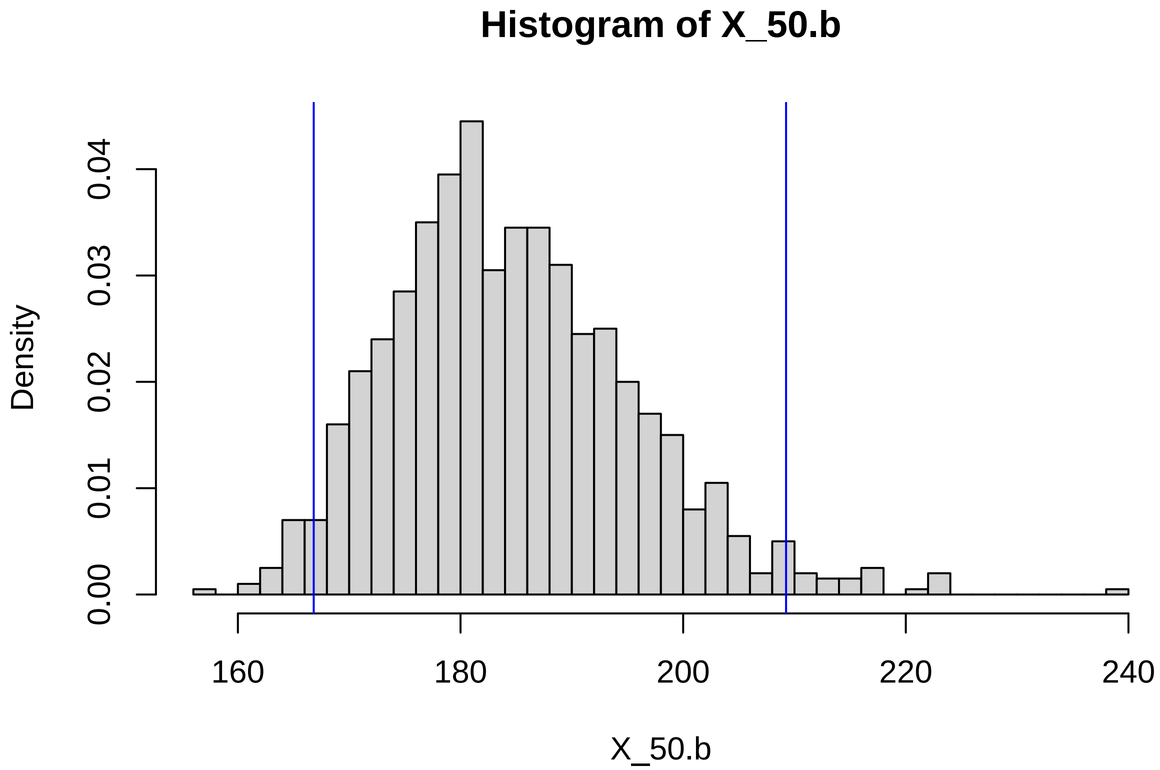
Question 1.2

Estimation of X_50

$P(Y_i = 1)$=$exp(\beta_0 + \beta_1 * X)/(1 + exp(\beta_0 + \beta_1 * X))$ For $P(Y_i = 1) = 0.5$, $exp(\beta_0 + \beta_1 * X_{50})/(1 + exp(\beta_0 + \beta_1 * X_{50}))$ Hence $X_{50} = -\beta_0/\beta_1$

```
## [1] "Median effective level (X_50): 184"
```

Non parametric bootstrap The bootstrap algorithm is as follow: - Resample the data with replacement. - Refit the GLM to each bootstrap sample - Compute $X_50 = -\beta_0/\beta_1$ for each resample - Generate the empirical distribution of X_50 - Compute the mean of X_50 across bootstrap samples - Construct a 95% CI based on the bootstrap distribution. We used percentiles CI.



```
## [1] "Bootstrap CI of the median effective level (X_50): 167"
## [2] "Bootstrap CI of the median effective level (X_50): 209"
```

Question 1.3

OR for a unit increased in age The Or for a unit increased in age is calculated as the exponentiation value of β_1.

```
## [1] "OR for a unit increased in age: 1.02"
```

Non parametric bootstrap to construct a 95% C.I: percentile - Resample the data with replacement. - Fit the GLM to each bootstrap sample. - Extract β_1 from each bootstrap fit and compute the OR for each sample. - Use the 2.5th and 97.5th percentiles of the bootstrap distribution of OR to construct the CI.

Non parametric bootstrap to construct a 95% C.I: t interval methods - Resample the data with replacement. - Fit the GLM to each bootstrap sample. - Extract β_1 and the related standard error from each bootstrap fit for each sample. - Evaluate for each bootstrap the OR as $exp((\beta_{1.b} - \beta_1)/se(\beta_1))$ - Use the 2.5th and 97.5th percentiles of the bootstrap distribution of OR to construct the CI.

The bootstrap t interval is wider compared to the bootstrap percentile interval. We prefer the bootstrap percentile interval. The percentile method is generally preferred for the odds ratio, as the bootstrap-t method assumes symmetry and might not account for the skewness in the OR distribution.

Question 1.4

We focus on the odds ratio (OR) for a unit increased in age. Use parametric bootstrap to test the null hypothesis $H_0 : OR = 1$.

Null hypothesis (H_0): : The odds ratio for a unit increase in age (OR) is 1. This implies that $\beta 1 = 0$ in the logistic regression model.

Alternative Hypothesis (H_1):The odds ratio is not equal to 1 ($\beta 1 \neq 0$).

The bootstrap procedure is as follow: - Fit the GLM to the observed data under H_0, where $\beta 1 = 0$ - Simulate response data using the fitted model under H_0 - Refit the GLM to each simulated dataset and calculate the test statistic ($\beta 1$) - Compare the observed test statistic to the distribution of test statistics from the bootstrap samples.

Question 1.5

The logistic regression model gives:

$$\pi_{33} = \frac{\exp(\beta_0 + \beta_1 \cdot 33)}{1 + \exp(\beta_0 + \beta_1 \cdot 33)}$$

The bootstrap procedure is described as follow: - Fit the GLM to the observed data and estimate β_0 and β_1 - Simulate response data based on the fitted probabilities under the original model. - Refit the GLM to each bootstrap sample. - Calculate π_{33} for each refitted model. - Estimate the standard error (SE) of π_{33} from the bootstrap distribution. - Construct the 90% CI using the bootstrap percentiles

```
## [1] "Standard error for pi_{33} is   : 0.00"
```

```
## [1] "CI for pi_{33} is   : 0.04" "CI for pi_{33} is   : 0.04"
```

Part 2

In this question we fit a robust GLM for the model specified in Q1. Use the R package *glmRob* to fit the model.

Question 2.1

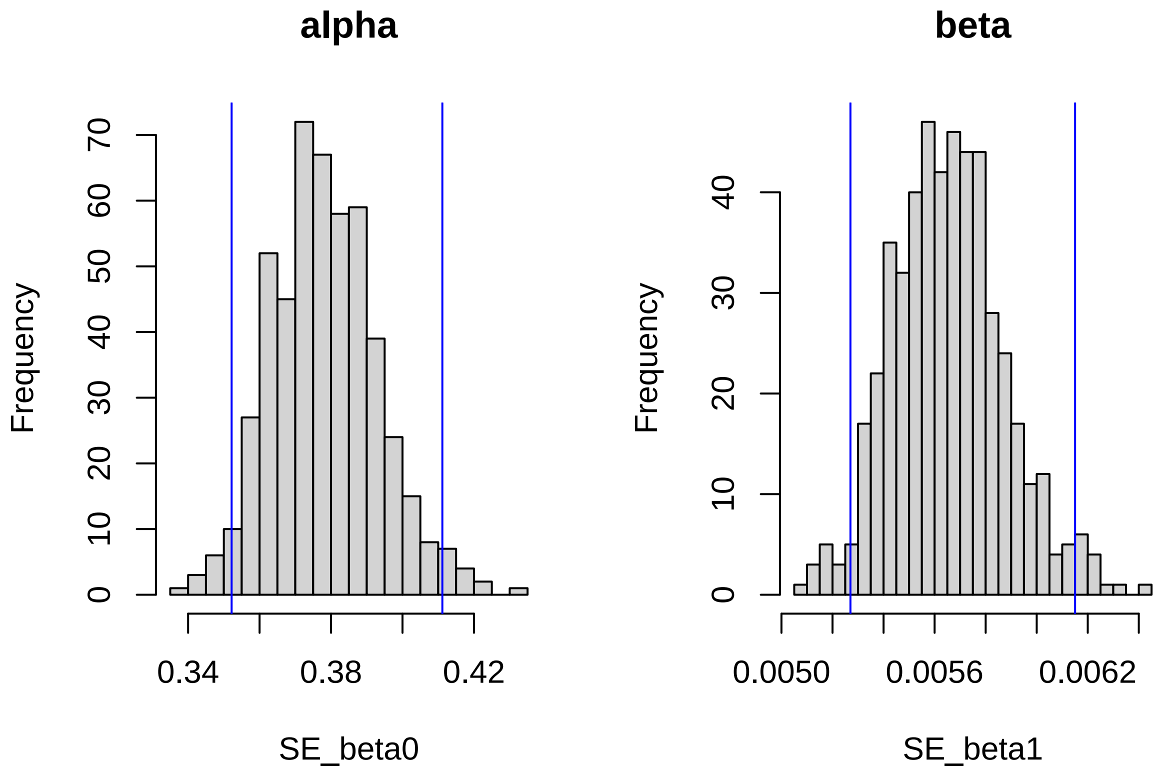
Estimate the model using the R package glmRob.

Question 2.2

Use non parametric bootstrap to estimate the SE for the intercept and slope. The bootstrap algorithm is as follow: - Resample the data with replacement. - Refit the robust GLM to each bootstrap sample - Compute the intercept and the slope for each bootstrap sample - Generate the empirical distribution of the intercept and the slope - Compute the mean of intercept and the slope across bootstrap samples

```
## [1] "SE for intercept   : 0.38"
```

```
## [1] "SE for slope   : 0.01"
```



Question 2.3

Use the jackknife and the bootstrap procedures to estimate the bias and MSE for the intercept and slope estimated by: (1) the GLM model in Q2.1 and (2) the robust GLM model estimated in Q2.2. Which method you prefer to use for the estimation of the intercept and slope?

For each GLM approach, the bias and MSE were evaluated after the sampling procedures: bootstrap and Jackknife.

a.Jackknife: Remove one observation at a time. Refit the model and calculate the estimates. Use jackknife estimates to compute bias and MSE.

b.Bootstrap: Resample data (with replacement) many times. Refit the model on each bootstrap sample. Use bootstrap estimates to compute bias and MSE.

```
{r, cache=TRUE, include=FALSE} # set.seed(2025) # # Classical
GLM and Robust GLM # classical_model <- glm(dead ~ ageOFocc,
data = nassCDS, family = binomial) # robust_model <- glmRob(dead
~ ageOFocc, data = nassCDS, family = binomial) #  # # Number
of observations # n <- nrow(nassCDS) #  # # Initialize storage
for jackknife and bootstrap # B <- 1000  # Number of bootstrap
samples # coef_jack_glm <- matrix(0, n, 2)  # Store jackknife
estimates for GLM # coef_jack_robust <- matrix(0, n, 2)  #
Store jackknife estimates for robust GLM # coef_boot_glm <-
matrix(0, B, 2)  # Store bootstrap estimates for GLM # coef_boot_robust
<- matrix(0, B, 2)  # Store bootstrap estimates for robust GLM
#  # # Jackknife procedure # for (i in 1:n) { #   # Exclude
one observation #   jack_data <- nassCDS[-i, ] #    #   # Classical
GLM #   jack_model_glm <- glm(dead ~ ageOFocc, data = jack_data,
family = binomial) #   coef_jack_glm[i, ] <- coef(jack_model_glm)
#    #   # Robust GLM #   jack_model_robust <- glmRob(dead ~
ageOFocc, data = jack_data, family = binomial) #   coef_jack_robust[i,
] <- coef(jack_model_robust) # } #  # # Bootstrap procedure #
for (b in 1:B) { #   # Resample data with replacement #   boot_data
<- nassCDS[sample(1:n, replace = TRUE), ] #    #   # Classical
GLM #   boot_model_glm <- glm(dead ~ ageOFocc, data = boot_data,
family = binomial) #   coef_boot_glm[b, ] <- coef(boot_model_glm)
#    #   # Robust GLM #   boot_model_robust <- glmRob(dead ~
ageOFocc, data = boot_data, family = binomial) #   coef_boot_robust[b,
] <- coef(boot_model_robust) # } #  # # Calculate bias and MSE
# # Original coefficients # coef_glm <- coef(classical_model)
# coef_robust <- coef(robust_model) #  # # Bias and MSE for
Jackknife # bias_jack_glm <- colMeans(coef_jack_glm) - coef_glm
# mse_jack_glm <- bias_jack_glm^2 + apply(coef_jack_glm, 2,
var) #  # bias_jack_robust <- colMeans(coef_jack_robust) -
coef_robust # mse_jack_robust <- bias_jack_robust^2 + apply(coef_jack_ro
2, var) #  # # Bias and MSE for Bootstrap # bias_boot_glm <-
colMeans(coef_boot_glm) - coef_glm # mse_boot_glm <- bias_boot_glm^2
+ apply(coef_boot_glm, 2, var) #  # bias_boot_robust <- colMeans(coef_bo
- coef_robust # mse_boot_robust <- bias_boot_robust^2 + apply(coef_boot_
2, var) #  # # Combine results # results <- list( #   Jackknife_GLM
= list(Bias = bias_jack_glm, MSE = mse_jack_glm), #   Jackknife_Robust
= list(Bias = bias_jack_robust, MSE = mse_jack_robust), #   Bootstrap_GL
= list(Bias = bias_boot_glm, MSE = mse_boot_glm), #   Bootstrap_Robust
= list(Bias = bias_boot_robust, MSE = mse_boot_robust) # ) #
# results #
```

Question 3.1

Define the observation unit (X_i, Y_i) for the question

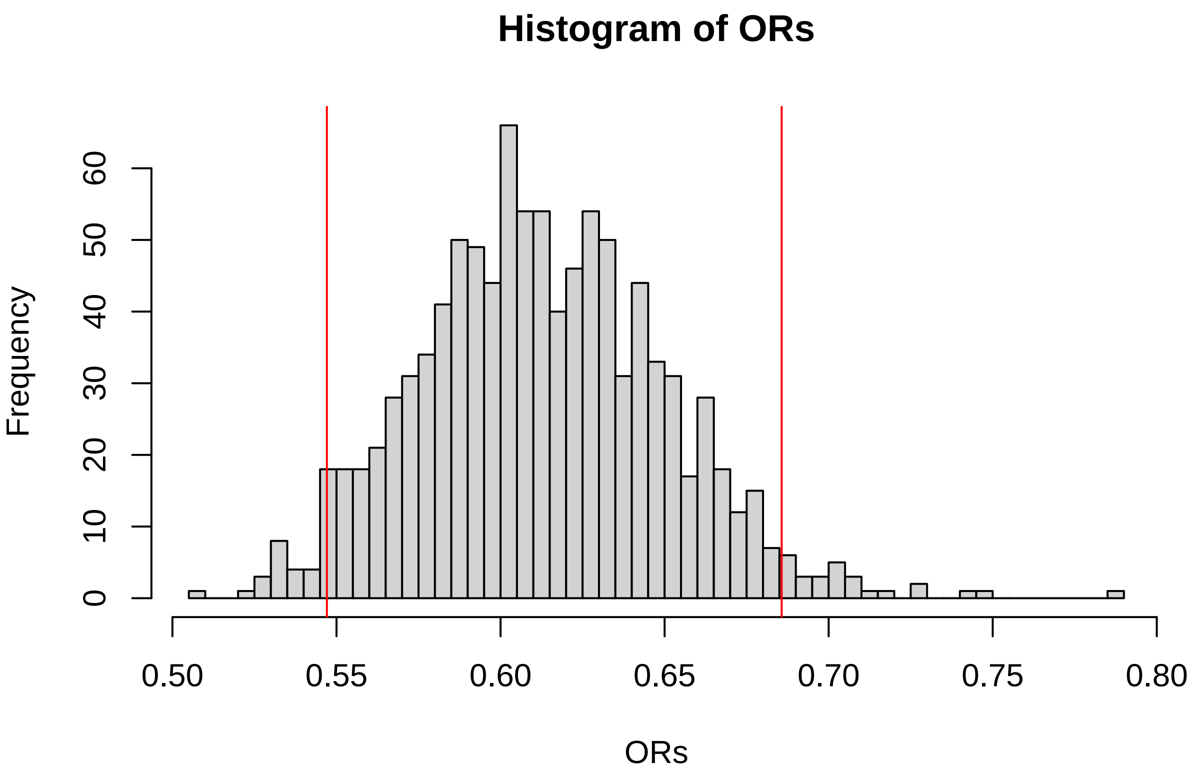
```
##        alive dead   Sum
## none   11058  669 11727
## airbag 13825  511 14336
## Sum    24883 1180 26063
```

Question 3.2

The OR is less than 1, airbags are associated with decreased odds of survival.

Question 3.3: Parametric bootstrap to construct a construct a 95% confidence interval for the OR.

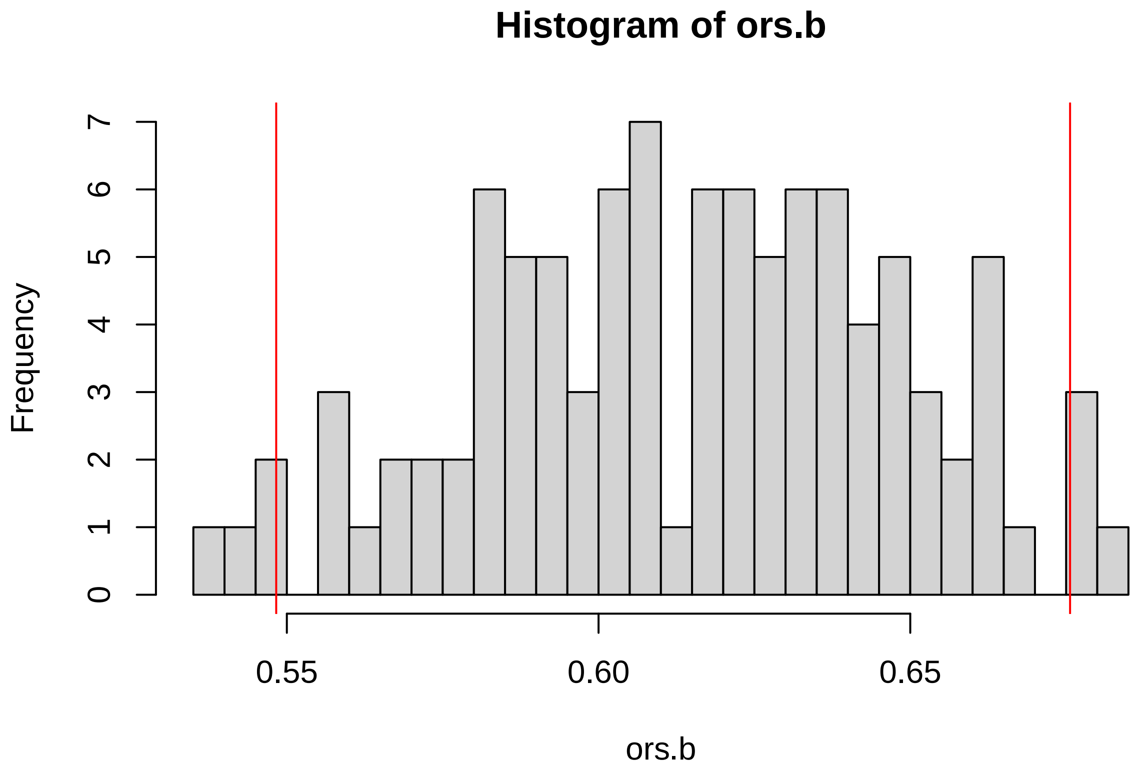
Simulate B samples from the multinomial distribution using the observed cell proportions. For each bootstrap sample: - Reconstruct the contingency table. - Compute the odds ratio



```
## [1] "CI for OR   : 0.55" "CI for OR   : 0.67"
```

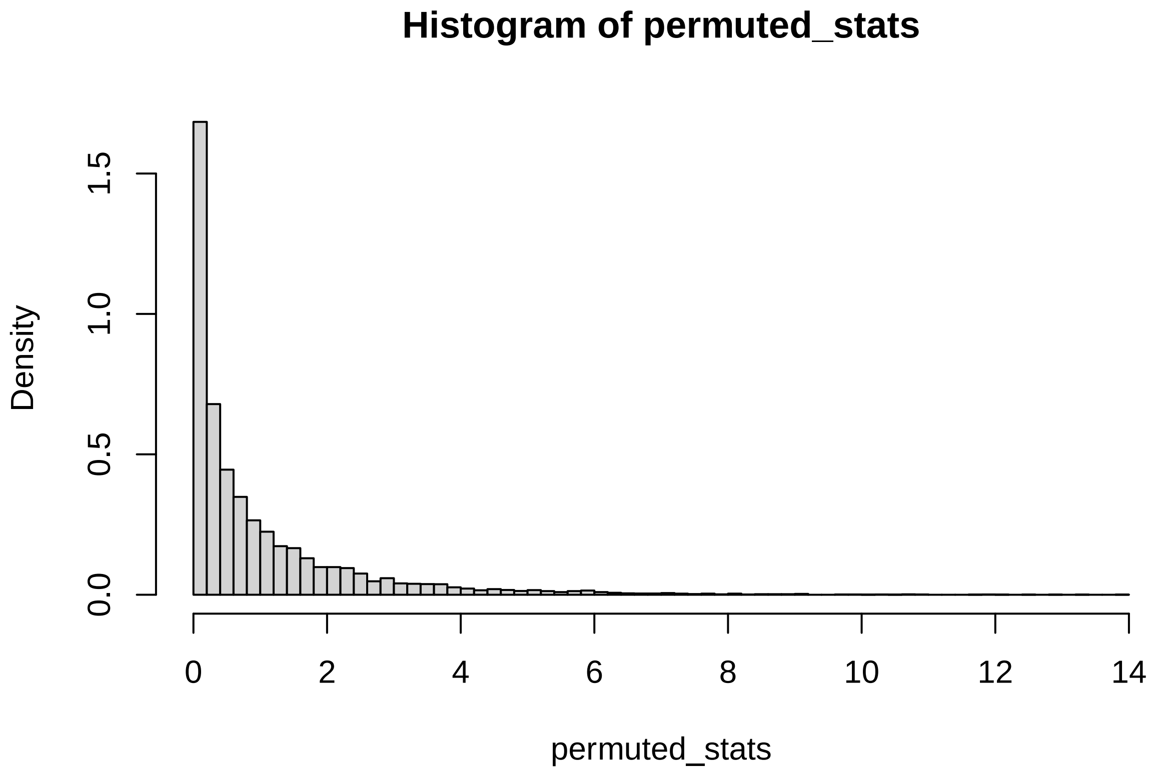
```
## [1] 0.6146497
```

```
##      2.5%     97.5%
## 0.5482931 0.6756358
```



Question 3.4

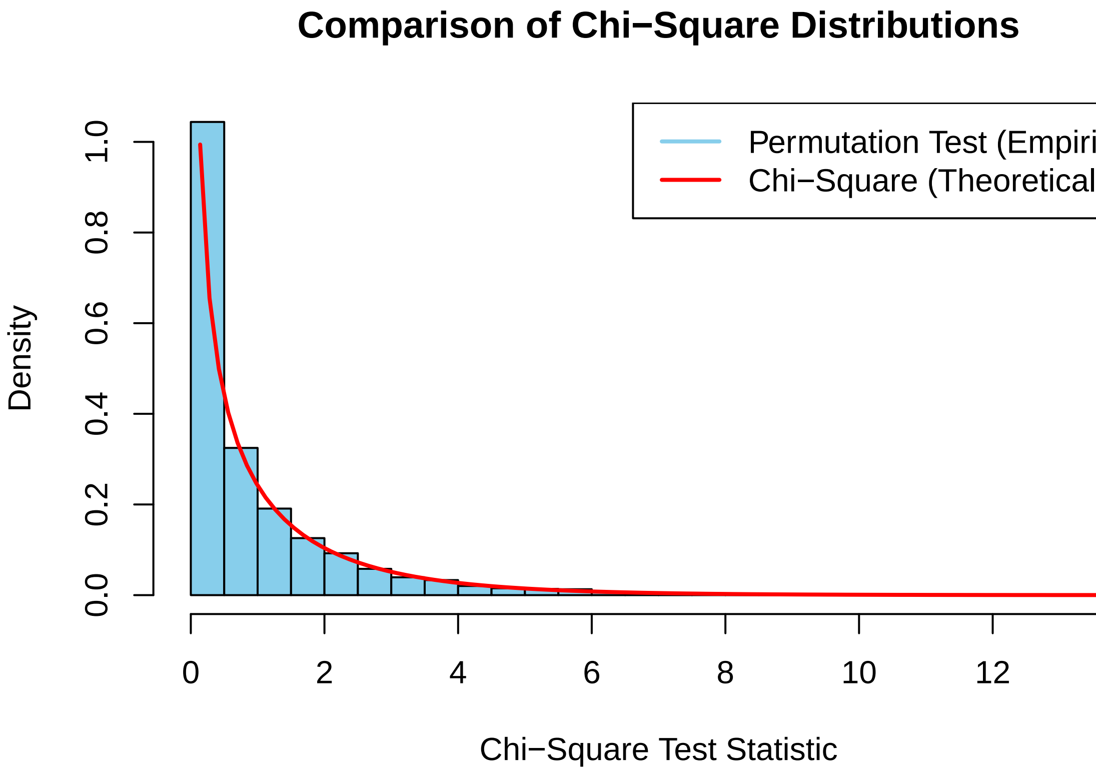
Null Hypothesis (H_0):Airbags have no effect on the accident outcome (i.e., the rows and columns of the table are independent). The permutation test est performed as follow: - Use the observed 2X2 table, table to calculate the test statistic - Permutations Under the Null: –> Shuffle the outcome labels ("Alive" or "Dead") randomly, keeping the marginal totals fixed. –> For each permutation, create a new 2X2 table and compute the test statistic - Calculate the p-value: –> Compare the observed test statistic to the distribution of test statistics from the permutations.



```
## [1] 9.999e-05
```

p-value=9.999e-05. We reject the H_0 => airbags in the car influences significantly the accident outcome.

The distributions align well, the theoretical chi-square distribution is a good approximation for the test statis-



tic under the null hypothesis.

Part 4

If this question we focus on the variables dead (the outcome of the accident) and the gender (the variable sex).

Question 4.1

To estimate the proportion of males ( ) and female ( ) that died in the accidents, we calculated the proportion of deaths within each gender category. The formula for each proportion is:

$$\pi_M = \frac{\text{Number of males who died}}{\text{Total number of males}}$$

$$\pi_F = \frac{\text{Number of females who died}}{\text{Total number of females}}$$

```
## [1] "The proportion of male  that died in the accidents is  : 0.051"
```

```
## [1] "The proportion of female that died in the accidents is  : 0.038"
```

Question 4.2

Test the hypothesis that the proportion of male and female that died in an accident are equal using a classical two-samples test (use a two sided test).

Null hypothesis (H_0): $\pi_M = \pi_F$ Alternative hypothesis (H_A): $\pi_M \neq \pi_F$

```
## Reject the null hypothesis: The proportions are significantly different.
```

Question 4.3

Use parametric bootstrap to test the hypothesis that the proportion of male and female that died in an accidents are equal against a two sided alternative.

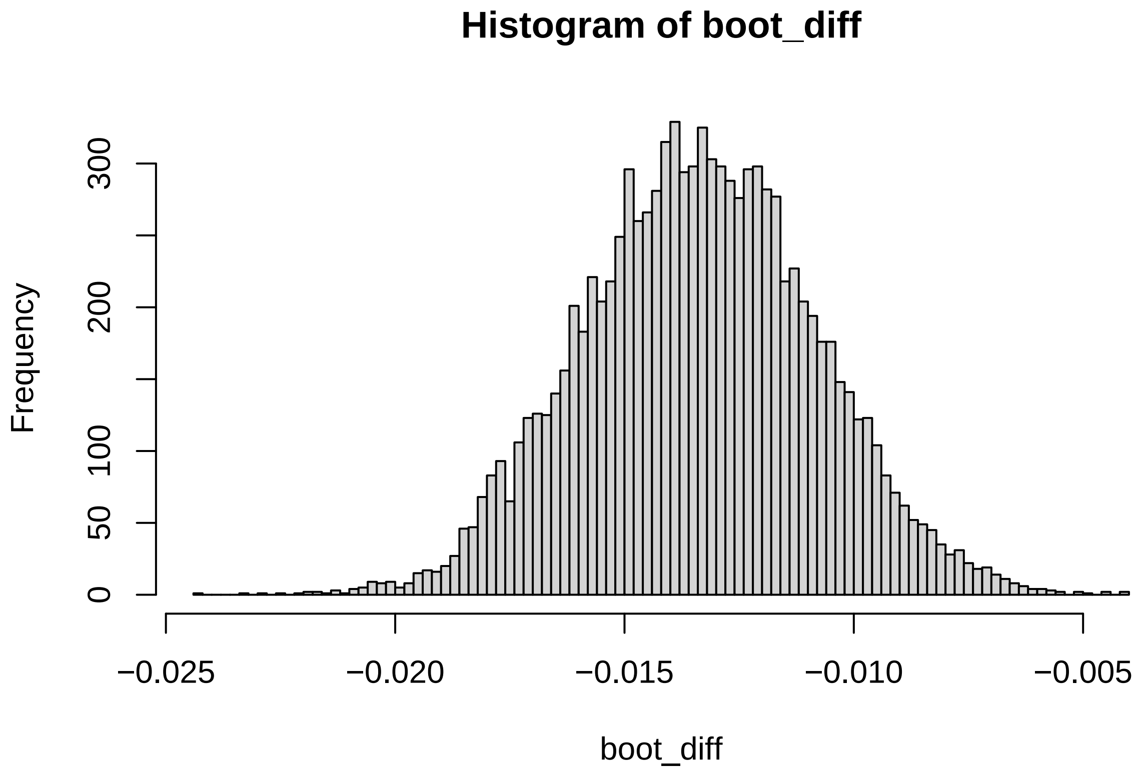
Under H_0, the proportion of males and females who died are the same. The pooled proportion π^{hat} is the same for both groups. The bootstrap procedure is as follow: - Under the null, generate new datasets by randomly sampling deaths for both males and females based on the pooled proportion π^{hat}, while keeping the sample sizes fixed (the number of males and females). - For each bootstrap sample, calculate the difference in proportions between the males and females who died. - Compute the p-value by comparing the observed test statistic to the distribution of the bootstrapped test statistics and make decision.

```
## Reject the null hypothesis: The proportions are significantly different.
```

p-value=9.999e-05. The p-value is less than 0.05, you would reject the null hypothesis and conclude that there is a statistically significant difference in the proportions of males and females who died in accidents.

Question 4.4

The non-parametric bootstrap procedure to construct a 95% confident interval for $\pi M - \pi F$ is described as follow: - calculate the observed difference in proportions $\pi M - \pi F$ from the original data. - Create bootstrap samples by resampling with replacement from the observed data for both males and females. - For each bootstrap sample, calculate the difference in proportions between males and females. - Construct the 95% percentiles Confidence Interval.



```
## [1] "The difference in the proportion is  : -0.018"
## [2] "The difference in the proportion is  : -0.008"
```

The CI 95% CI: [-0.018, 0.008]. The CI does not contain zero: Reject the null hypothesis, suggesting a significant difference between the proportions of males and females who died.