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

Project 1

Mikita Bisliuk (2364811), Edmond Sacla Aide (2159278)

26 January, 2025

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.

```
data(nassCDS)
nassCDS <- na.omit(nassCDS)
names(nassCDS)
```

```
##  [1] "dvcat"       "weight"      "dead"        "airbag"      "seatbelt"
##  [6] "frontal"     "sex"         "ageOFocc"    "yearacc"     "yearVeh"
## [11] "abcat"       "occRole"     "deploy"      "injSeverity" "caseid"
```

```
nassCDS$Dead<- ifelse(nassCDS$dead=="dead", 1,0)
```

```
dim(nassCDS)
```

```
## [1] 26063    16
```

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

Estimate the model using a classical GLM approach.

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

```
glm.daag <- glm(dead ~ ageOFocc, data = nassCDS, family = "binomial")
summary(glm.daag)
```

```
##
## 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 coefficient estimate for `ageOFocc` is 0.021183. This means that for each additional year of age, the log-odds of the outcome (`dead`) increase by 0.021183 or alternatively the odds to die in a car accident increase by approximately 2.14%. The p-value for `ageOFocc` is less than 2e-16, indicating that the effect of age on the outcome is significant at 5% level of significance.

Question 1.2

Let X_{50} be the age of occupant for which the probability to die is 0.5, i.e., $P(Y_i = 1) = 0.5$. Estimate X_{50}. Use non parametric bootstrap to estimate the distribution of X_{50} and to construct a 95% C.I. for the X_{50}

Estimation of X_{50}

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$$
$$For \pi_i = 0.5, 0 = \beta_0 + \beta_1 X_{50}$$
$$X_{50} = -\frac{\beta_0}{\beta_1}$$

```
## [1] "Median effective level: 184.49"
```

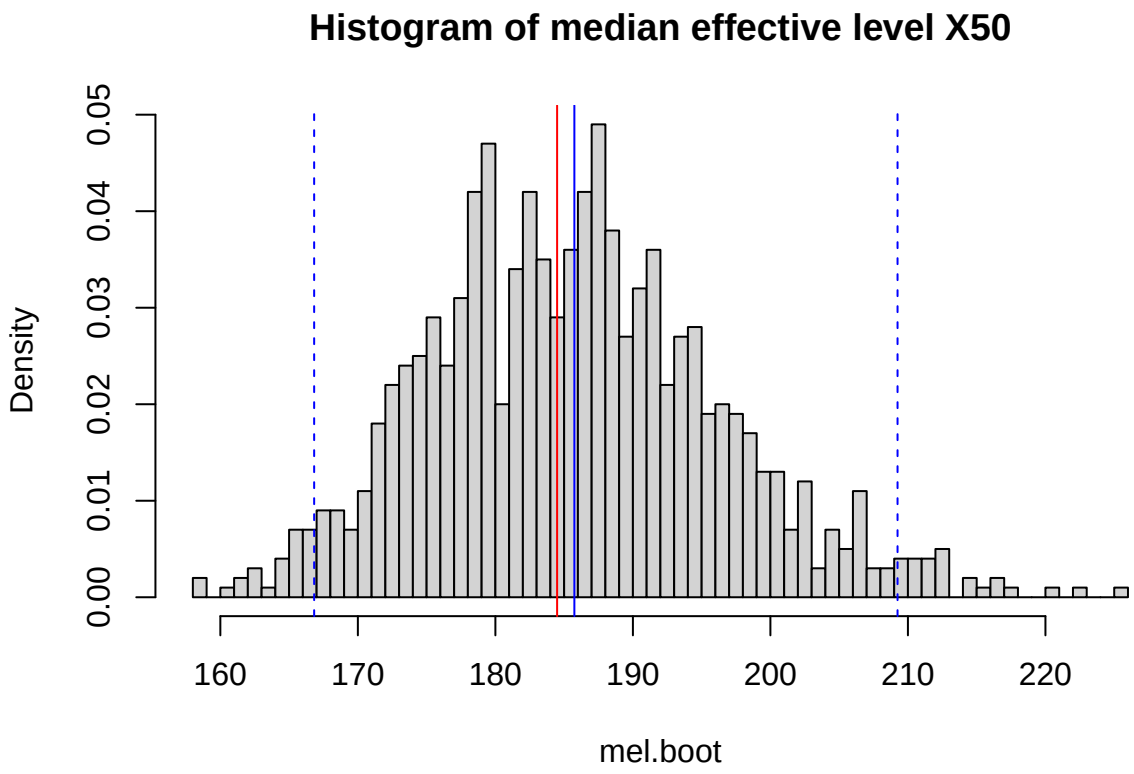
Non parametric bootstrap

The bootstrap algorithm is as follow:

- For each iteration resample the data with replacement.
- Refit the GLM to each bootstrap sample
- Compute $X_{50} = -\frac{\beta_0}{\beta_1}$

Finally, we compute the mean of X_{50} across bootstrap samples and construct a 95% C.I. using percentiles method.

The histogram illustrates the distribution of X_{50}, 95% confidence interval, observed and estimated via bootstrap median effective levels. The estimated bootstrap value for X_{50} is 185.74 which is very close to the observed value of 184.49. To find 95% CI we calculate 2.5% and 97.5% quantiles from bootstrap results giving [166.82, 209.25].



Question 1.3

For the model formulated above, estimate the OR (for a unit increased in age). Use non parametric bootstrap to construct a 95% C.I. for the OR (for a unit increased in age) using the percentile and bootstrap t interval methods, which one do you prefer for the parameter OR ?

OR for a unit increased in age

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

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$$

$$odds_i = \frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 x_i)$$

```
beta_1 <- coef(glm.daag)[2]
OR <- exp(beta_1)
sprintf("OR for a unit increased in age: %.2f", OR)
```

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

Non parametric bootstrap to construct a 95% C.I using percentile method

- Resample the data with replacement.
- Fit the GLM to each bootstrap sample.

- Extract β_1 estimate 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 t statistic as $(\beta_{1.b} - \beta_1)/se(\beta_1)$
- Calculate the 2.5th and 97.5th percentiles of the bootstrap distribution of t statistics.
- Use found percentiles to construct C.I of logOR, take exp() to find C.I. for OR.

```
## [1] "95% C.I. using bootstrap t-interval method:"
```

```
## [1] 1.018319 1.024627
```

```
## [1] "95% C.I. using bootstrap percentile method:"
```

```
## [1] 1.018282 1.024623
```

The bootstrap t interval is slightly 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. In general, `Studentized intervals` do not respect transformation of the form $\phi = m(\theta)$

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:

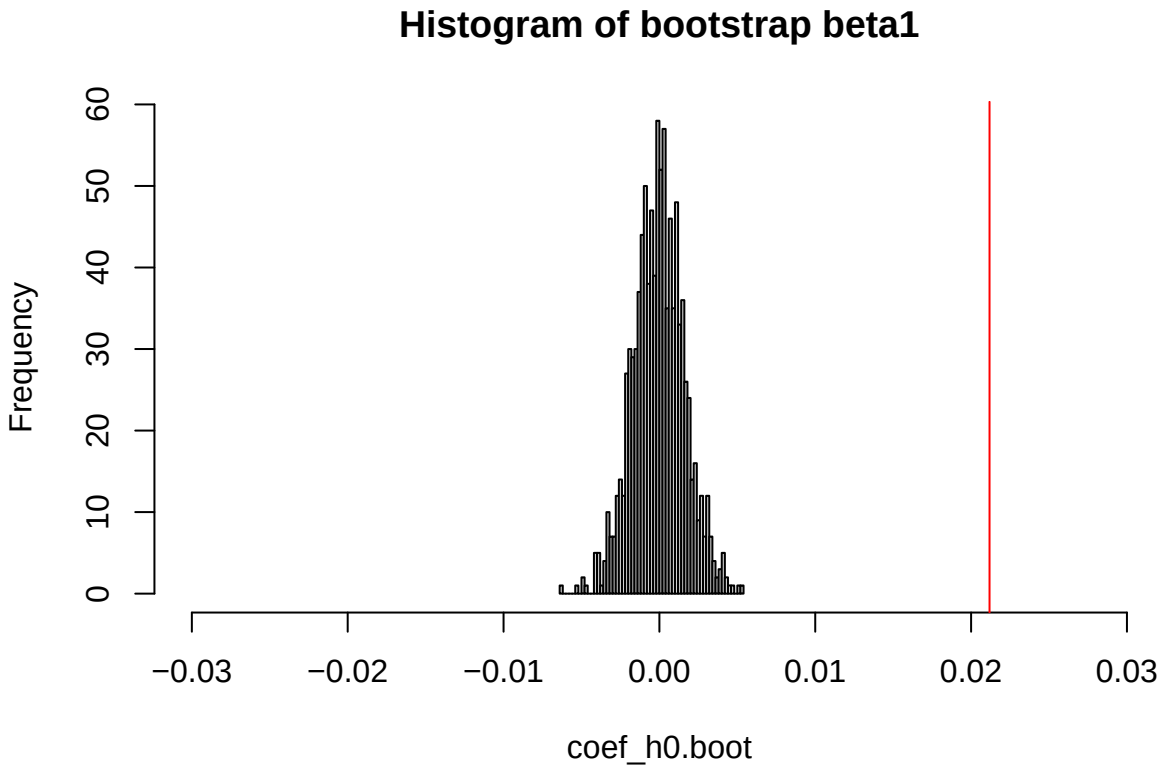
- Fit the GLM to the observed data under H_0, where $\beta 1 = 0$
- Simulate response data using the fitted model under H_0
- Refit the GLM to each simulated dataset and calculate the test statistic for ($\beta 1$)
- Compare the observed test statistic to the distribution of test statistics from the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

The summary of null model (of no age effect) is given below:

```
##
## Call:
## glm(formula = dead ~ 1, family = binomial(link = "logit"), data = nassCDS)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.04867    0.02979  -102.3   <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  on 26062  degrees of freedom
## Residual deviance: 9610  on 26062  degrees of freedom
## AIC: 9612
##
## Number of Fisher Scoring iterations: 5
```

The probability of death under H_0 is equal to 0.0453 at each age level (i.e., the overall prevalence in the sample). Use it for resampling step in parametric bootstrap.

The distributions of bootstrap replicates for $\hat{\beta}_b$ is shown below. The Monte-Carlo p-value is equal to $\frac{1}{B+1}$ and, hence, we reject the null hypothesis.



Question 1.5

Let π_{33} be the probability of death for an occupant at age 33. Use parametric bootstrap to calculate the standard error for π_{33} and construct a 90% C.I. for π_{33}.

The logistic regression model gives:

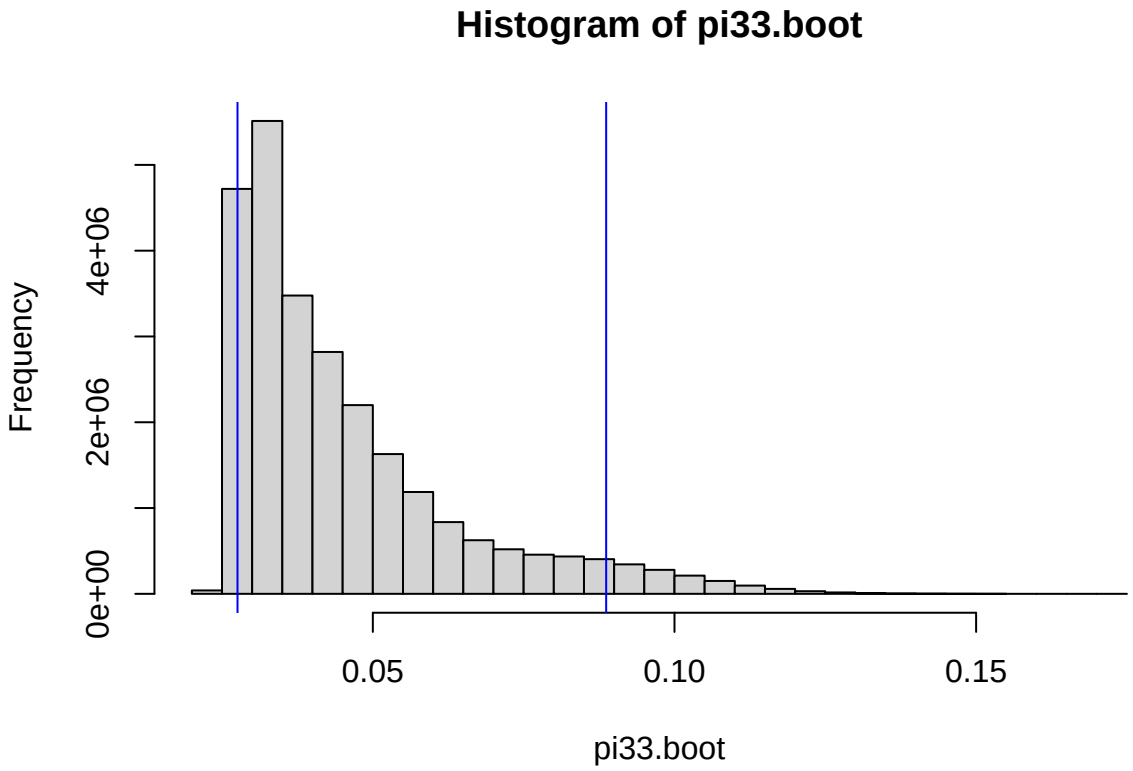
$$\pi_{33} = \frac{\exp(\beta_0 + \beta_1 \cdot 33)}{1 + \exp(\beta_0 + \beta_1 \cdot 33)}$$

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: 0.02"
```

```
## [1] "90% CI for pi_33 is: [0.028, 0.089]"
```



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.

```
##
## Call: glmRob(formula = dead ~ ageOFocc, family = "binomial", data = nassCDS)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5396 -0.3220 -0.2757 -0.2484  2.6821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.90798   0.072014  -54.27 0.00e+00
## ageOFocc     0.02118   0.001484   14.27 3.39e-46
##
## (Dispersion Parameter for binomial family taken to be 1 )
##
##     Null Deviance: 36131 on 26062 degrees of freedom
##
## Residual Deviance: 9418.158 on 26061 degrees of freedom
##
## Number of Iterations: 2
##
## Correlation of Coefficients:
##          (Intercept)
## ageOFocc -0.9096
```

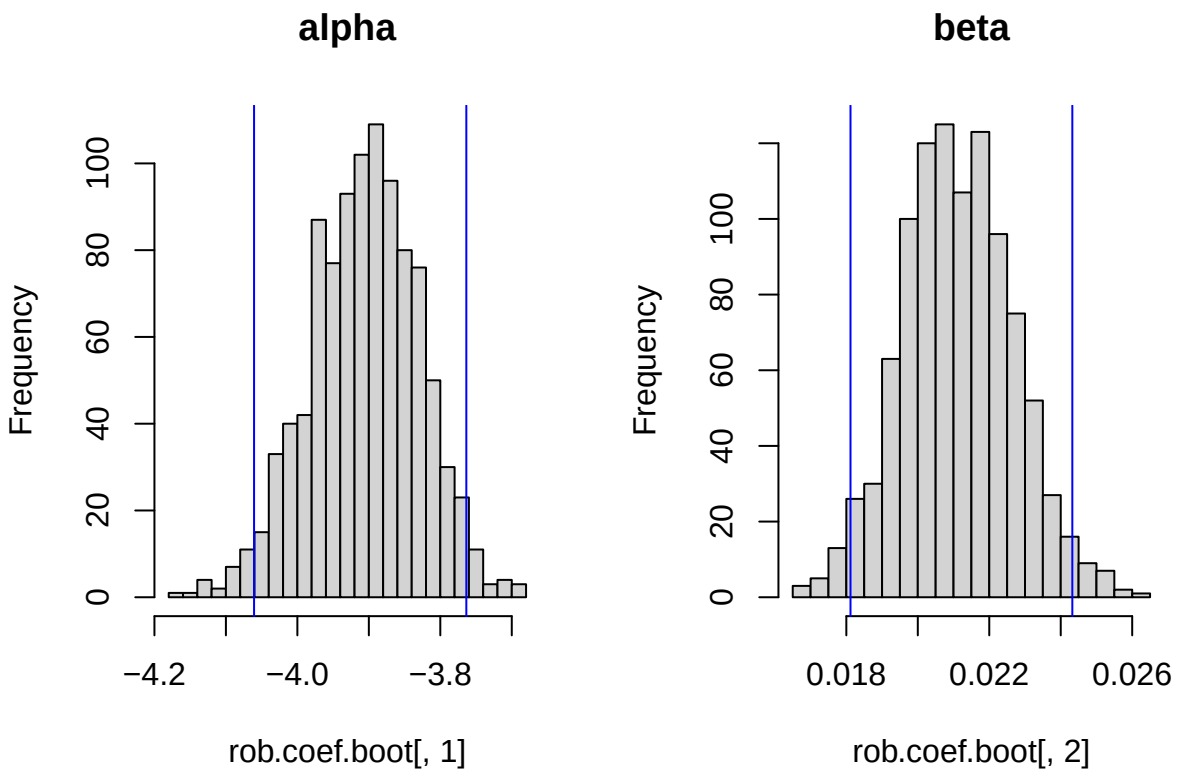
Question 2.2

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: 0.07616"
```

```
## [1] "SE for slope: 0.00156"
```



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.

Jackknife:

- Remove one observation at a time.
- Refit the model and calculate the estimates.
- Use jackknife estimates to compute bias and MSE. Also account for inflation factor.

Bootstrap:

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

```
## [1] "Bootstrap intercept bias: 2.03e-03"
```

```
## [1] "Bootstrap intercept MSE: 4.10e-06"

## [1] "Bootstrap slope bias: -6.69e-05"

## [1] "Bootstrap slope MSE: 4.48e-09"

## [1] "Jackknife intercept bias: 1.20e-02"

## [1] "Jackknife intercept MSE: 1.43e-07"

## [1] "Jackknife slope bias: -2.17e-04"

## [1] "Jackknife slope bias: 4.73e-11"

## [1] "Bootstrap intercept bias: 2.03e-03"

## [1] "Bootstrap intercept MSE: 4.10e-06"

## [1] "Bootstrap slope bias: -6.69e-05"

## [1] "Bootstrap slope MSE: 4.48e-09"

## [1] "Jackknife intercept bias: 1.20e-02"

## [1] "Jackknife intercept MSE: 1.43e-07"

## [1] "Jackknife slope bias: -2.17e-04"

## [1] "Jackknife slope bias: 4.73e-11"
```

Part 3

In this question we focus of the following 2×2 table (for the complete case analysis) for the variables airbag and dead.

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

Calculate the odds ratio for usage of airbag and the accident outcome (dead/alive) and construct 95% confidence interval. You can use the R function oddsratio. What is your conclusions? Do you think that airbags in the car influence the accident outcome ?

```
## $data
##        alive dead Total
## none   11058  669 11727
## airbag 13825  511 14336
## Total  24883 1180 26063
##
## $measure
##                          NA
## odds ratio with 95% C.I.  estimate     lower     upper
##                    none   1.0000000        NA        NA
##                    airbag 0.6109508 0.5430396 0.6873547
##
## $p.value
##          NA
## two-sided   midp.exact fisher.exact   chi.square
##    none             NA           NA           NA
##    airbag 1.110223e-16 1.752803e-16 1.360575e-16
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

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

Question 3.3

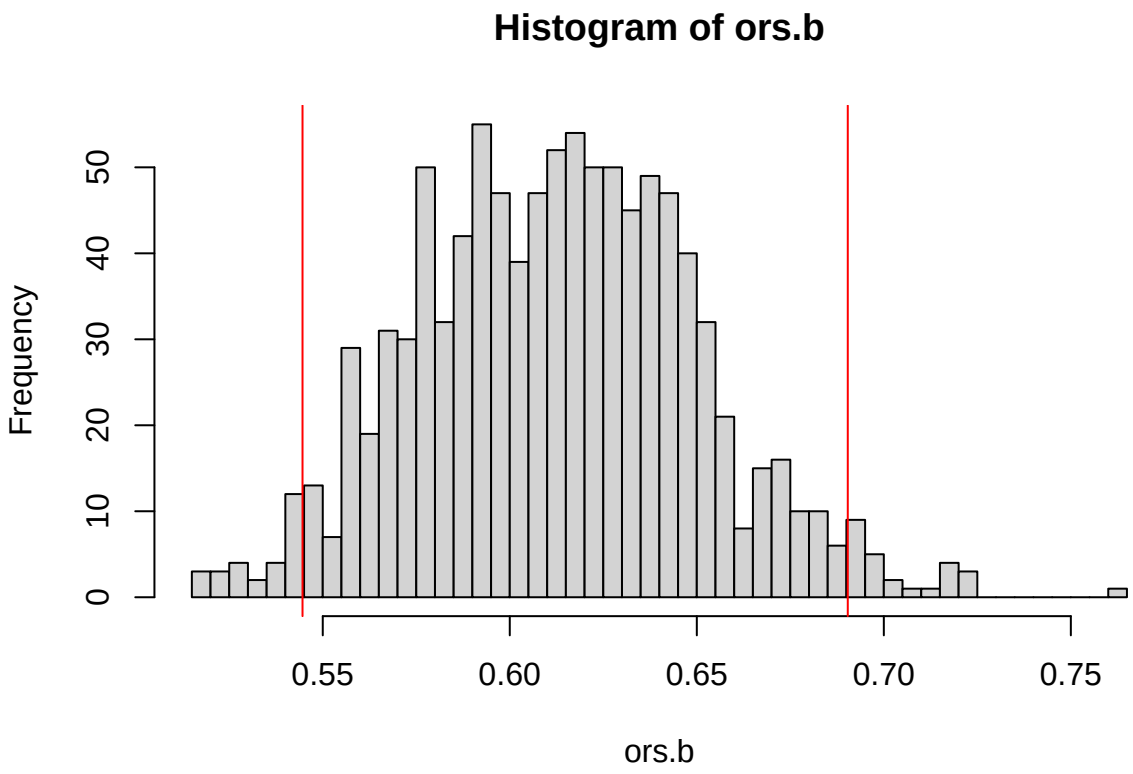
Use parametric bootstrap to construct a construct a 95% confidence interval for the OR.

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] 0.6127328
```

```
##      2.5%     97.5%
## 0.5445870 0.6903582
```



Question 3.4

Use permutations test to test the hypothesis that airbags in the car DO NOT influence the accident outcome using a chi-square test for a 2 × 2 table. Compare the distribution of the chi-square test statistic in this question to the theoretical distribution of the test statistic.

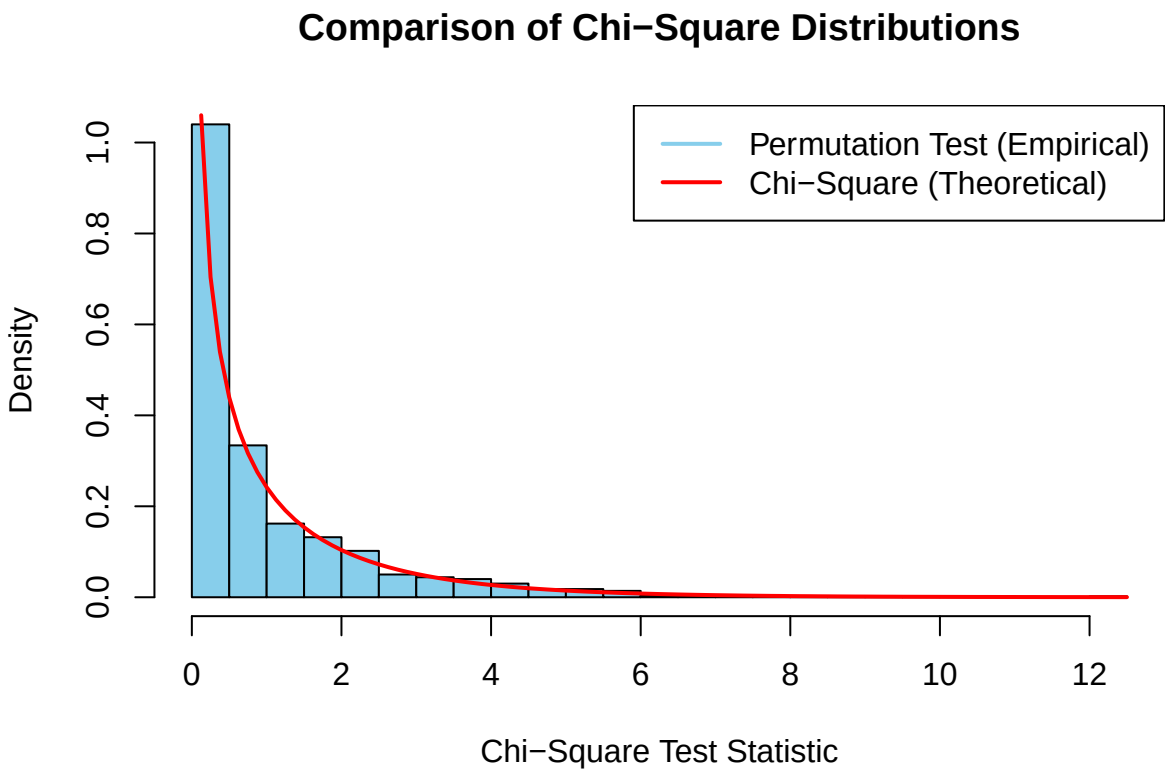
Null Hypothesis (H_0): Airbags have no effect on the accident outcome (i.e., the rows and columns of the table are independent). The permutation test est performed as follow:

- 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] 0.000999001
```

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 statistic 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

Estimate the proportion of male ( ) and female ( ) that died in the accidents.

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.

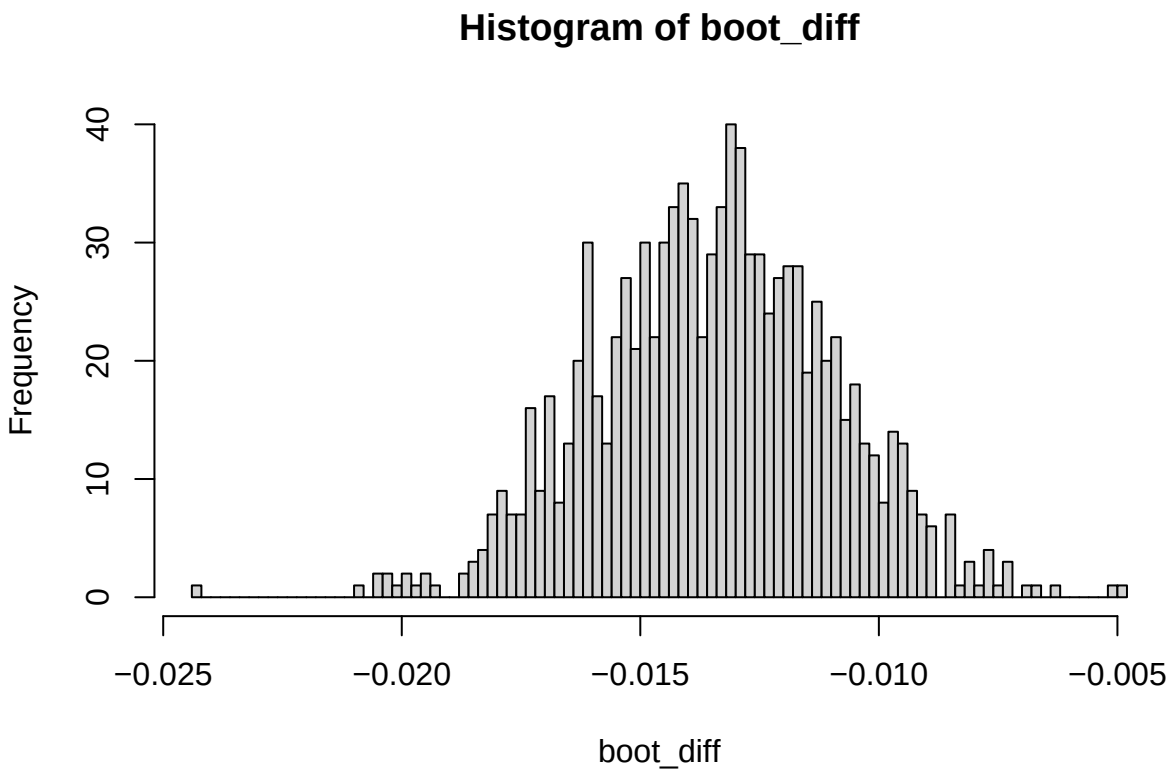
Under H_0, the proportion of males and females who died are the same. The pooled proportion π^{hat} is the same for both groups. The bootstrap procedure is as follow: - Under the null, generate new datasets by randomly sampling deaths for both males and females based on the pooled proportion π^{hat}, 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  : -0.009"
```

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.