Permutation Tests for ANOVA

for Total Charge Estimation

TABLE OF CONTENTS

Introduction

Goal:

Assessing the Effectiveness of a Permutation test in the analysis of variance (ANOVA) model for Estimating Total Charges of Patients in a Hospital

OUR PROCESS

01
Data Collection

02
EDA

03
Traditional ANOVA

04
Permutation ANOVA

05
Calculate 95% C.I.

Data Description

Data Source

The data we are using to show these techniques focuses on the hospital inpatient discharges from 2017 within the state of New York. They use a statewide planning and research cooperative system (SPARCS) to contain patient, characteristics, diagnoses, treatments, services and charges.

The data is provided from the New York State Department of Health.

Website https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-z7e7

Data Cleaning

The original data set included **34** different variables with a total of **2.34 million** rows of data. Before we could do anything we had to cut down this, or it would have taken ages and way too much memory to work in R with.

We narrowed it down to just a handful of important variables (**total charges, age group, and hospital county**) to allow us to show the methods from our presentation. Most importantly, we are focusing only on hospital data from **Manhattan**.

Our Variables

Total Charges : Total charges for the discharge

Age Group : Age in years at time of discharge. Grouped into the following age groups: 0 to 17, 18 to 29, 30 to 49, 50 to 69, and 70 or Older.

Hospital County : A description of the county in which the hospital is located.

```r
manhattan_data <- readr::read_csv("hospital.csv") %>%
  janitor::clean_names() %>%
  filter(hospital_county == "Manhattan") %>%
  select(age_group, total_charges)
```

Variable Explanations

Age Group ~ The age of the patient at the time of discharge

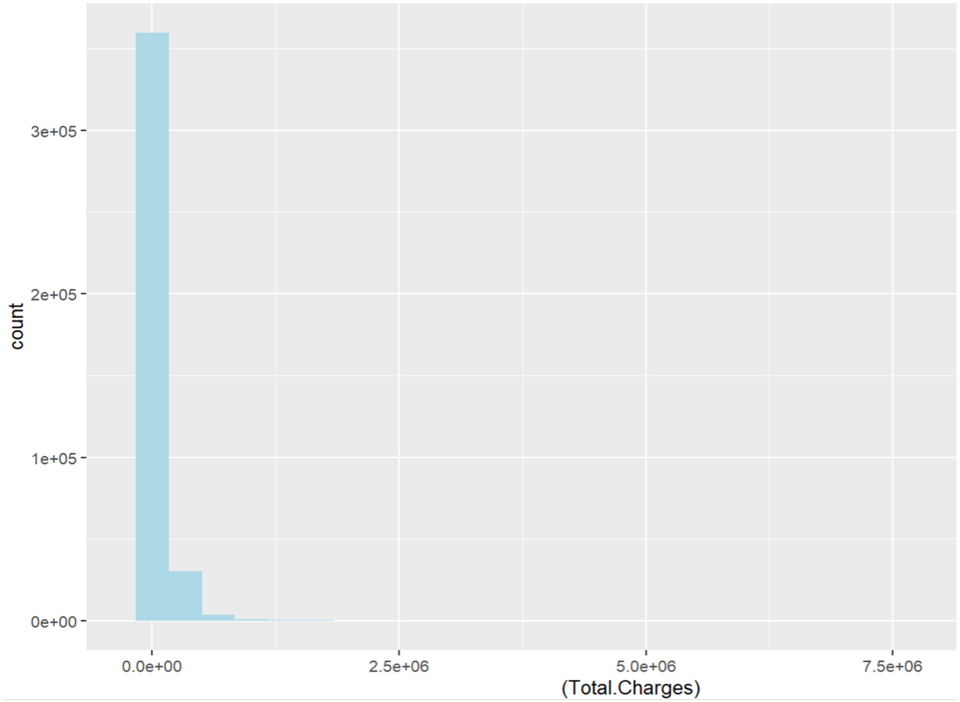
- Age Groups are broken up into
  0 - 17 ,   18 - 29 ,   30 - 49 , 50 - 69 , 70 <
- The groups are roughly equal in size, with the smallest group, 18 to 29, making up 8.9% of the data while the largest group, 50 to 69, makes up 29% of the data.

Total Charges ~ The total amount charged for the hospital discharge

- This is a continuous variable with mean $76,006.14, standard deviation $143,871.4, and a range from 0.25 to $9,696,645.
- This data is incredibly right-skewed.

EDA

Distribution of Charges



Without transformation

Transformations?

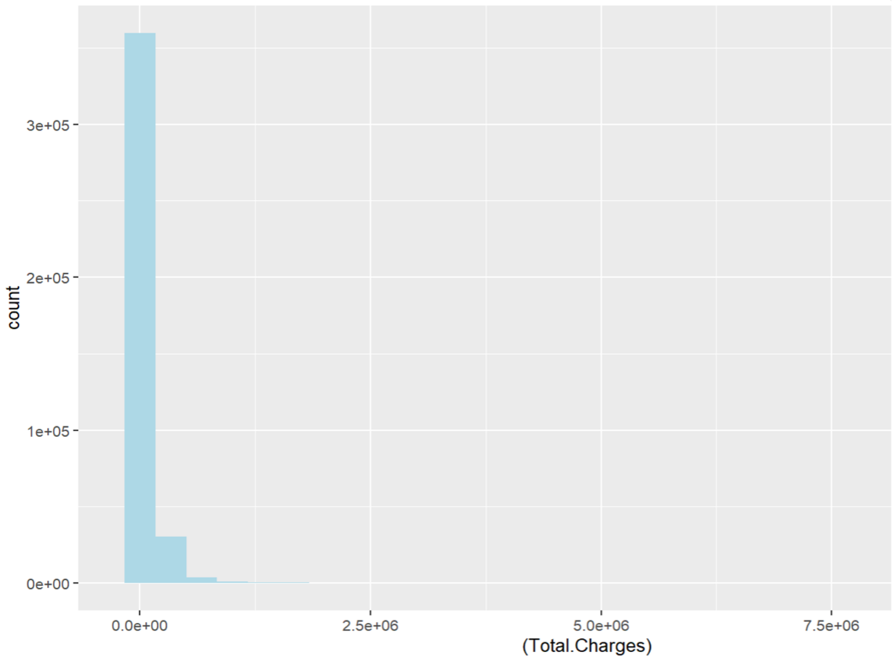
From the box-cox, we got **λ=-0.07** which indicates that we may need a power transformation

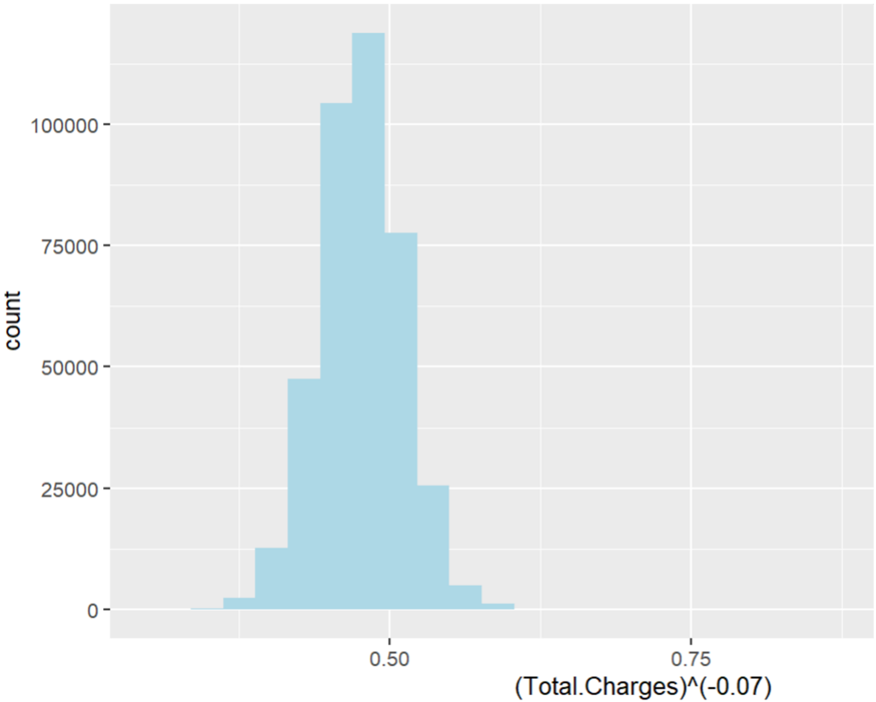
```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0725       -0.07      -0.0745      -0.0704
```

Distribution of Charges

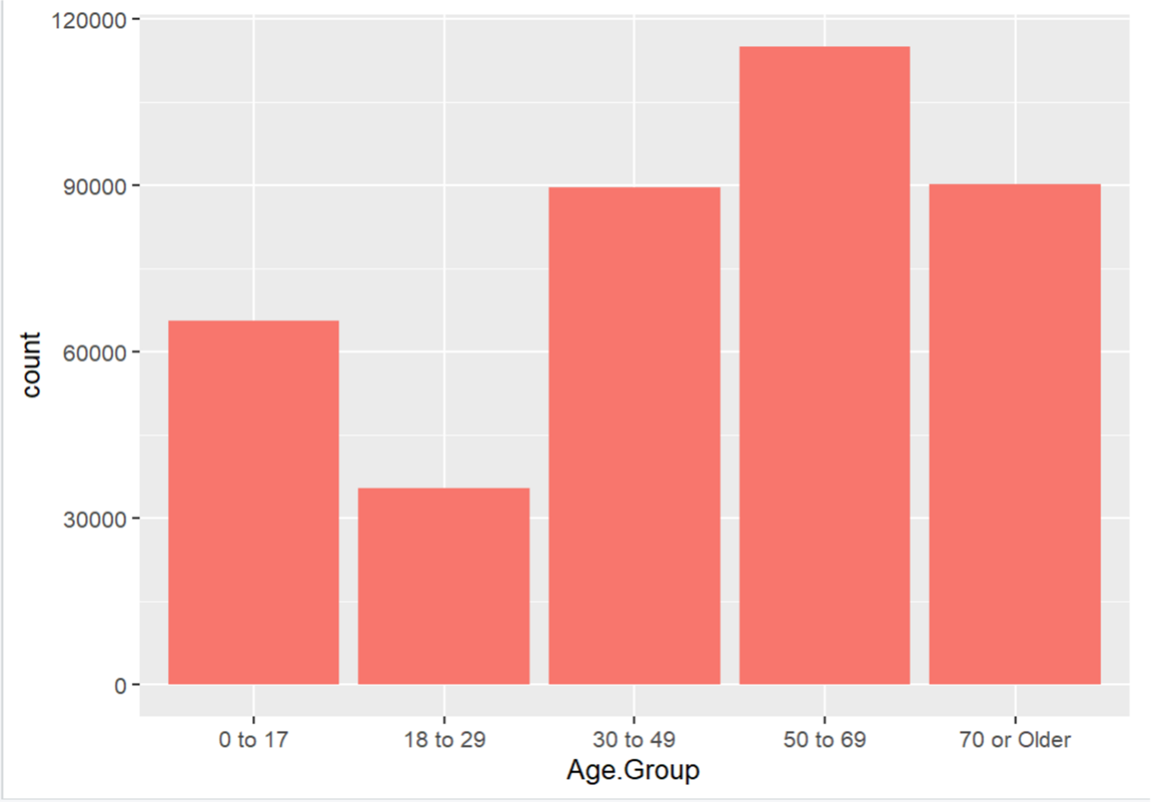
Without transformation

With transformation

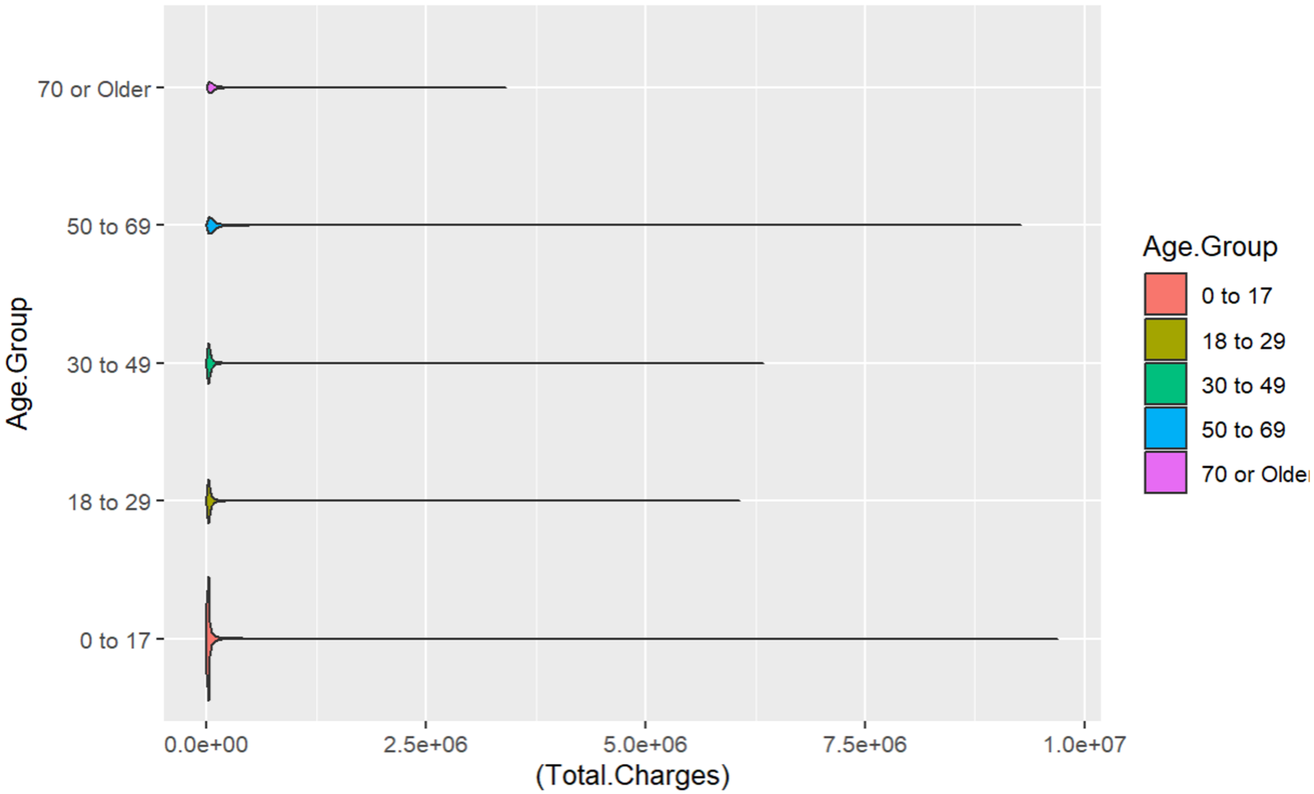




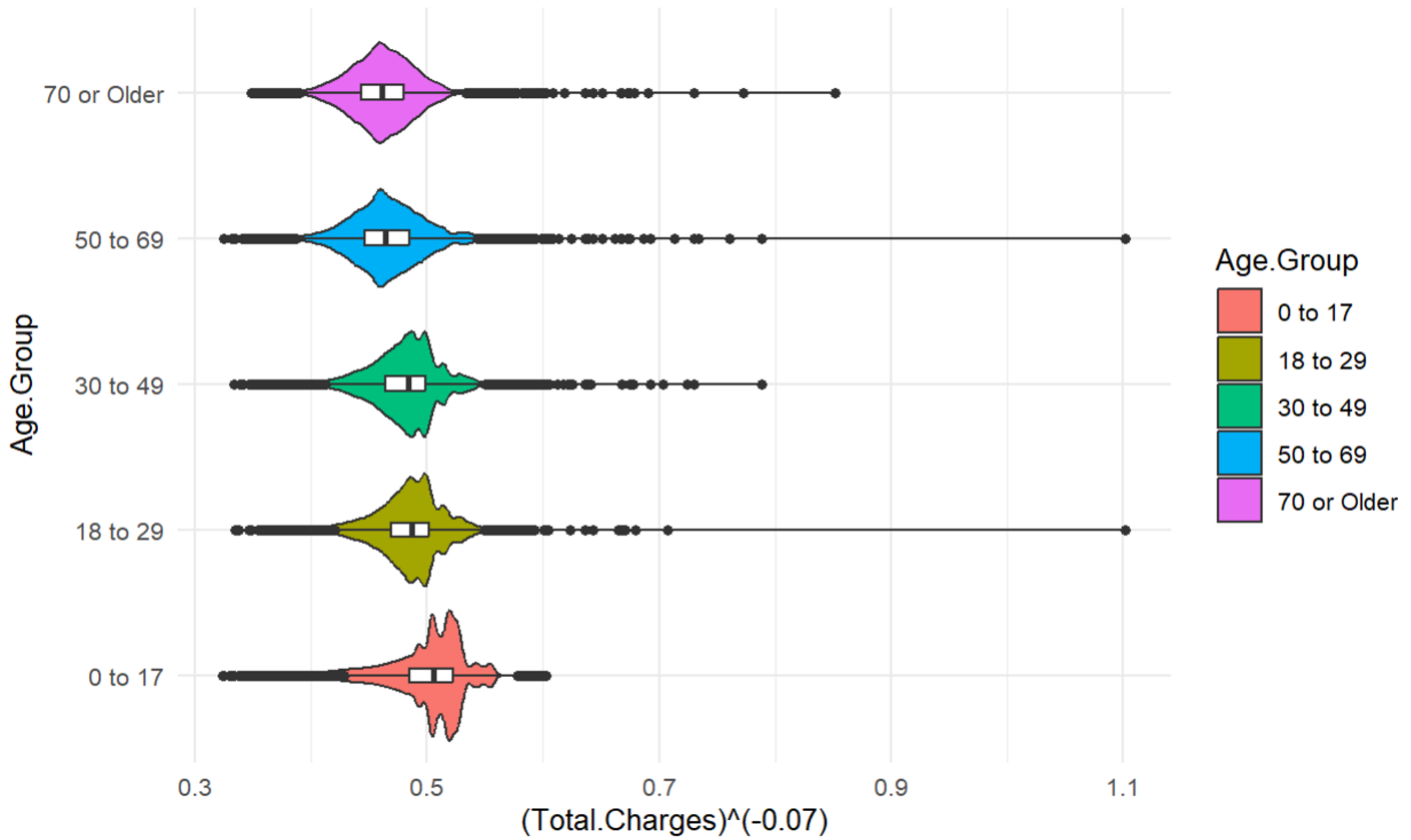
Barplot of age group



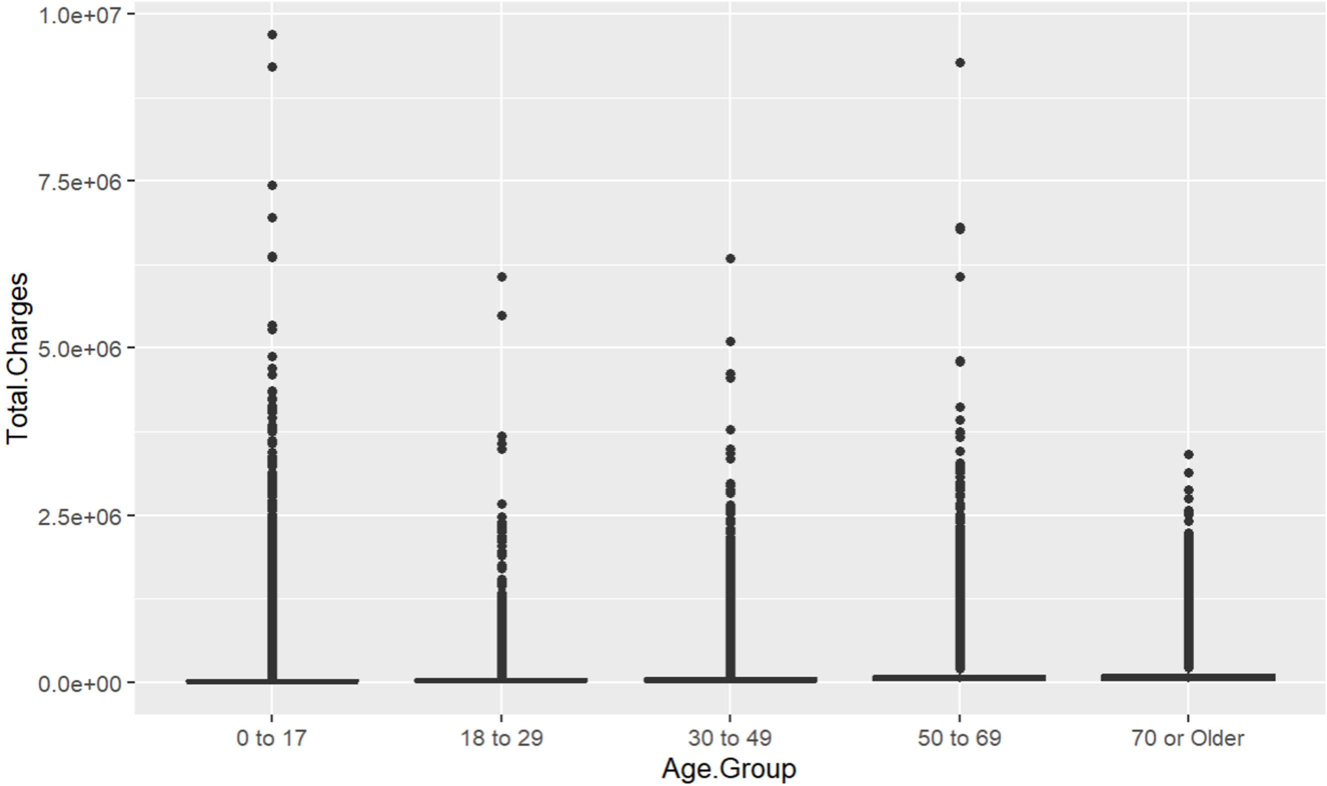
Violin plot without transformation



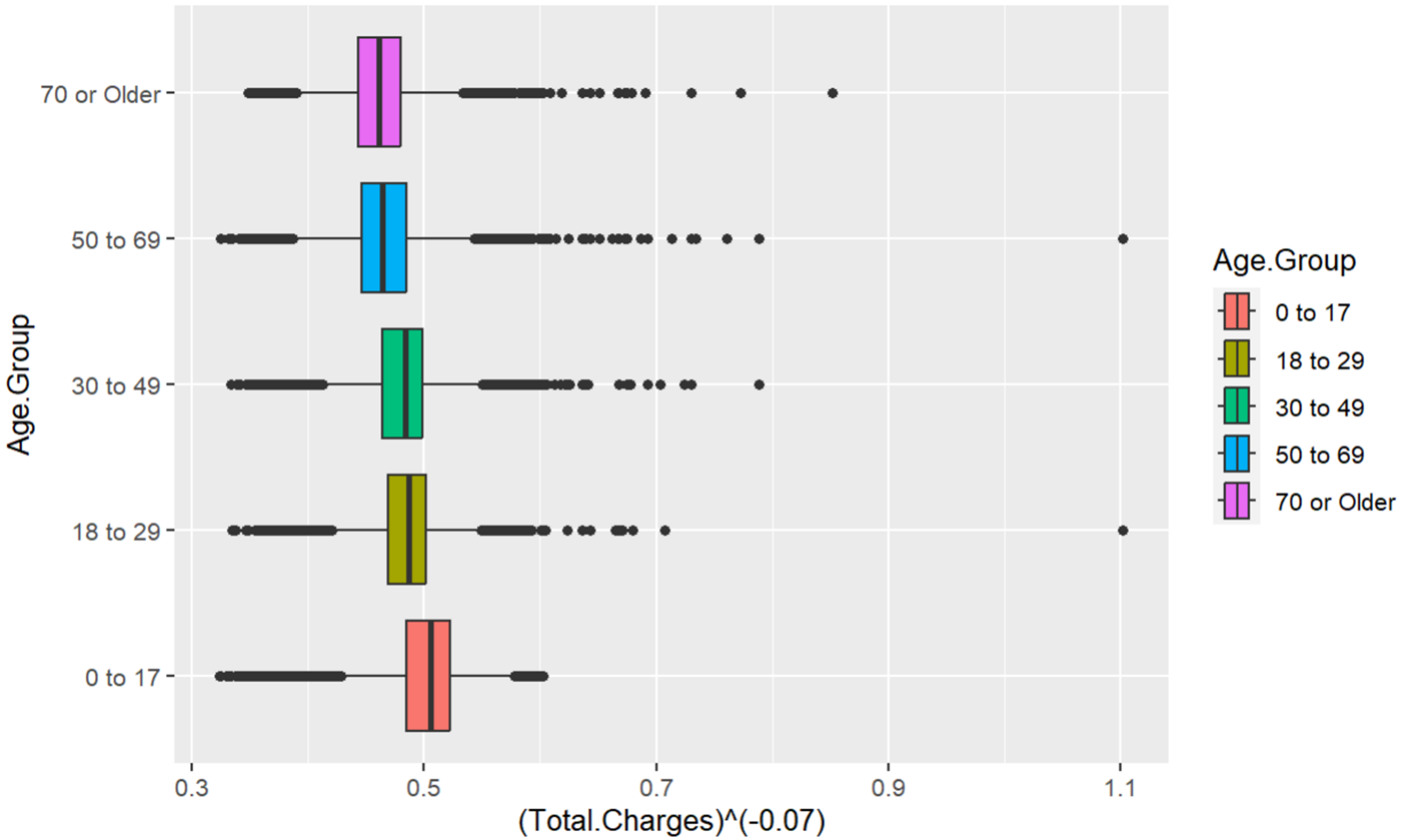
VIOLIN PLOT


70 or Older
50 to 69
30 to 49
18 to 29
0 to 17
Age.Group
0.3
0.5
0.7
0.9
1.1
(Total.Charges)^(-0.07)
Age.Group
0 to 17
18 to 29
30 to 49
50 to 69
70 or Older

Box-plot without transformation



BOX PLOT



70 or Older
50 to 69
30 to 49
18 to 29
0 to 17
Age.Group
0.3
0.5
0.7
0.9
1.1
(Total.Charges)^(-0.07)
Age.Group
0 to 17
18 to 29
30 to 49
50 to 69
70 or Older

Summary Statistics:

Age Group	Mean	SD	(using transformed response) Mean	(using transformed response) SD
0 to 17	55932.	200566.	0.499	0.0354
18 to 29	52993.	114447.	0.486	0.0310
30 to 49	57794.	111648.	0.482	0.0309
50 to 69	93276.	148779.	0.466	0.0322
70 or Older	95673.	119245.	0.462	0.0278

Methods

I. Traditional ANOVA

Traditional analysis of variance (ANOVA) is a statistical method used to compare means of different groups.

Fixed groups have assigned observations, they are not random

Residuals are i.i.d and have a normal distribution with a shared variance

With independence being the most important assumption

I. Traditional ANOVA

Since we are trying to see if there is a difference between the medical costs for different age groups in Manhattan, we are going to be testing this hypothesis,

H_0: There is no difference in the true mean total medical costs between the different age groups in Manhattan.

$H_0 : \mu_1 = \mu_2 = \mu_2 = \ldots = \mu_n$

Vs.

H_a: There is some difference in the true mean total medical costs between the different age groups in Manhattan.

$H_a : \mu_i \neq \mu_j$ for some $i \neq j$

1. Traditional ANOVA

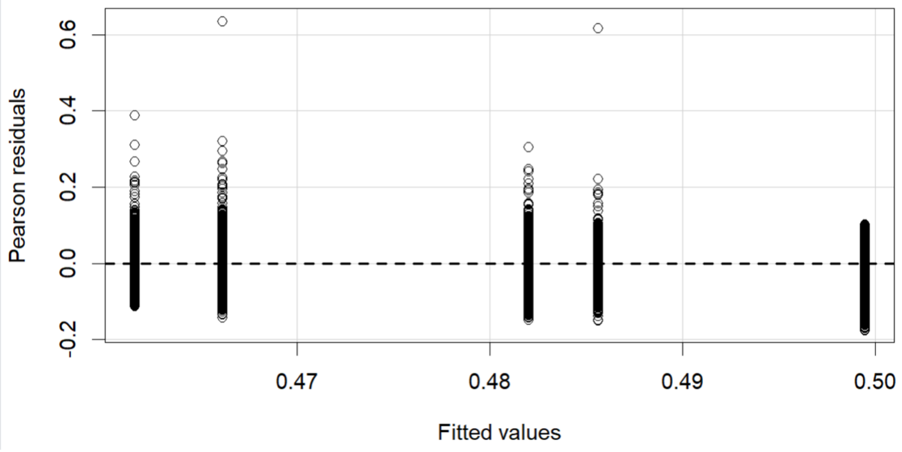
Traditional ANOVA

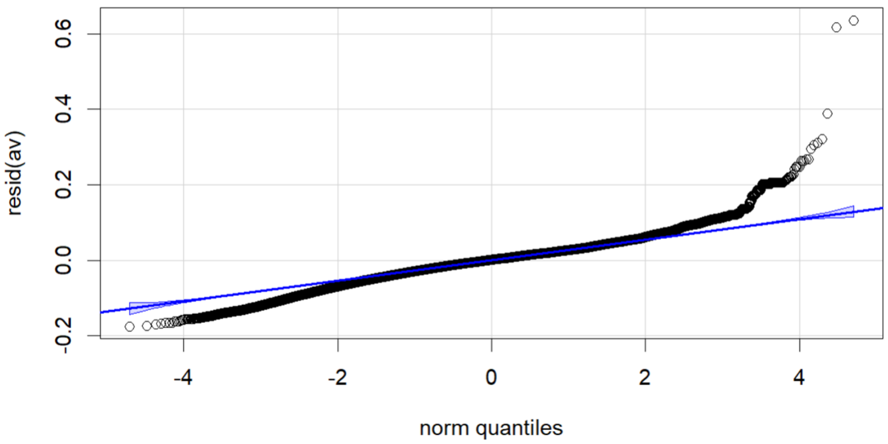
```
manhattan_data2 <- manhattan_data %>%
  mutate(transformed_charges = total_charges^-0.07)

aov(
  transformed_charges ~ age_group,
  data = manhattan_data2
) %>%
  anova()
```

```
Analysis of Variance Table

Response: transformed_charges
              Df Sum Sq Mean Sq F value    Pr(>F)
age_group      4  72.63  18.157   18377 < 2.2e-16 ***
Residuals 395712 390.97   0.001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value is significant, so we **reject** the null hypothesis.

1. Checking Assumptions





Advantages Vs Disadvantages of permutation test of Anova

Non-parametric

Normality and constant variance of residuals not required

(Chihara & Hesterberg 64, 428)

Computationally intensive

2. Permutation ANOVA

1. Computing the ANOVA model and F- statistic from the original data

1. Generating a random permutation of the response while keeping the predictors in the same order, this creates the permuted data

   We compute the ANOVA model and F-statistic for the permuted data

   Save the permuted F- statistics

1. Repeat Step 2 N-1 more times

1. Compute the P-value as the proportion of permuted F-statistics greater than our original from Step 1

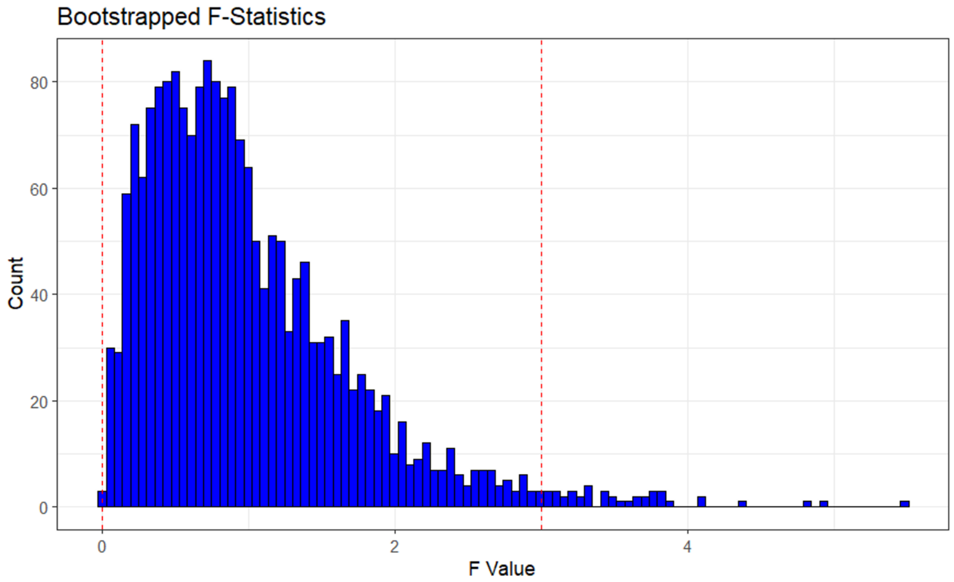
(Chihara & Hesterberg 428-29)

2. Permutation ANOVA

```r
set.seed(0)

f_stat <- function(data, idx) {
  anova(
    tibble(
      age_group = data$age_group,
      total_charges = data$total_charges[idx]
    ) |>
      aov(total_charges ~ age_group, data = _)
  )$F[1]
}

f_boot <- boot(
  statistic = f_stat,
  data = manhattan_data,
  R = 2000,
  sim = "permutation",
  parallel = "multicore",
  ncpus = 6
)
```

2. Permutation ANOVA

```r
ggplot(mapping = aes(f_boot$t)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "blue") +
  labs(
    title = "Bootstrapped F-Statistics",
    x = "F Value",
    y = "Count") +
  geom_vline(aes(xintercept = 3),
             color = "red", linetype = 2) +
  geom_vline(aes(xintercept = 0),
             color = "red", linetype = 2) +
  theme_bw()
```



2. Permutation ANOVA

The permutation method has no assumptions, making it more robust when the residuals are not normally distributed

From this, we got a F - statistic of **1771.019**.
This indicates that we have a p-value of **1/2001 ≈ 0.0005**.

With this p-value and a significance level of 0.05, we are able to **reject** our null, so there is proof that there is some difference of total mean costs between the different age groups.

Percentile Confidence Interval

At 95% confidence we get a percentile CI of **(0,3)**.

Since our observed F value is much larger than 3, we can verify that our observed value exceeds our values from the permutation test, and that there is a significant difference in mean costs between age groups.

Conclusion

Summary of Analysis

- From traditional and permutation based ANOVA analyses, there is evidence of a difference in mean total medical costs between age groups in Manhattan.

- **Assumptions**

  Traditional ANOVA: Assumptions

  Permutation ANOVA: No assumptions necessary

- Permutation ANOVA requires fitting a large number of ANOVA models

Reference:

Chihara, Laura M., and Tim C. Hesterberg. *Mathematical Statistics with Resampling and R*. Second ed., Wiley, 2018, pp. 64, 428-29.


Questions


Thank You