Analysis of Chance of Admission

AUTHOR
Michael Dang

PUBLISHED
November 20, 2024

The dataset that will be used is: https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

Where the data dictionary is found: https://github.com/micho0802/Bio-Stat/blob/main/Admission_Predict_data_dictionary.yaml

Load the library

```
library(broom)
library(tidyverse)
library(readr)
```

Load the dataset

```
dataset <- read_csv(
  file = "../data/Admission_Predict.csv", show_col_types = FALSE)
names(dataset) <- tolower(names(dataset))
glimpse(dataset)
```

```
Rows: 400
Columns: 9
$ `serial no.`        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,…
$ `gre score`         <dbl> 337, 324, 316, 322, 314, 330, 321, 308, 302, 323, …
$ `toefl score`       <dbl> 118, 107, 104, 110, 103, 115, 109, 101, 102, 108, …
$ `university rating` <dbl> 4, 4, 3, 3, 2, 5, 3, 2, 1, 3, 3, 4, 4, 3, 3, 3, 3,…
$ sop                 <dbl> 4.5, 4.0, 3.0, 3.5, 2.0, 4.5, 3.0, 3.0, 2.0, 3.5, …
$ lor                 <dbl> 4.5, 4.5, 3.5, 2.5, 3.0, 3.0, 4.0, 4.0, 1.5, 3.0, …
$ cgpa                <dbl> 9.65, 8.87, 8.00, 8.67, 8.21, 9.34, 8.20, 7.90, 8.…
$ research            <dbl> 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0,…
$ `chance of admit`   <dbl> 0.92, 0.76, 0.72, 0.80, 0.65, 0.90, 0.75, 0.68, 0.…
```

Rename the column

```
dataset |>
  rename(
    serial_no = "serial no.",
    gre_score = "gre score",
    toefl_score = "toefl score",
    university_rating = "university rating",
    chance_of_admit = "chance of admit"
  ) -> dataset
```

```
glimpse(dataset)
```

```
Rows: 400
Columns: 9
$ serial_no         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ gre_score         <dbl> 337, 324, 316, 322, 314, 330, 321, 308, 302, 323, 32…
$ toefl_score       <dbl> 118, 107, 104, 110, 103, 115, 109, 101, 102, 108, 10…
$ university_rating <dbl> 4, 4, 3, 3, 2, 5, 3, 2, 1, 3, 3, 4, 4, 3, 3, 3, 3, 3…
$ sop               <dbl> 4.5, 4.0, 3.0, 3.5, 2.0, 4.5, 3.0, 3.0, 2.0, 3.5, 3.…
$ lor               <dbl> 4.5, 4.5, 3.5, 2.5, 3.0, 3.0, 4.0, 4.0, 1.5, 3.0, 4.…
$ cgpa              <dbl> 9.65, 8.87, 8.00, 8.67, 8.21, 9.34, 8.20, 7.90, 8.00…
$ research          <dbl> 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1…
$ chance_of_admit   <dbl> 0.92, 0.76, 0.72, 0.80, 0.65, 0.90, 0.75, 0.68, 0.50…
```

Descriptive statistics

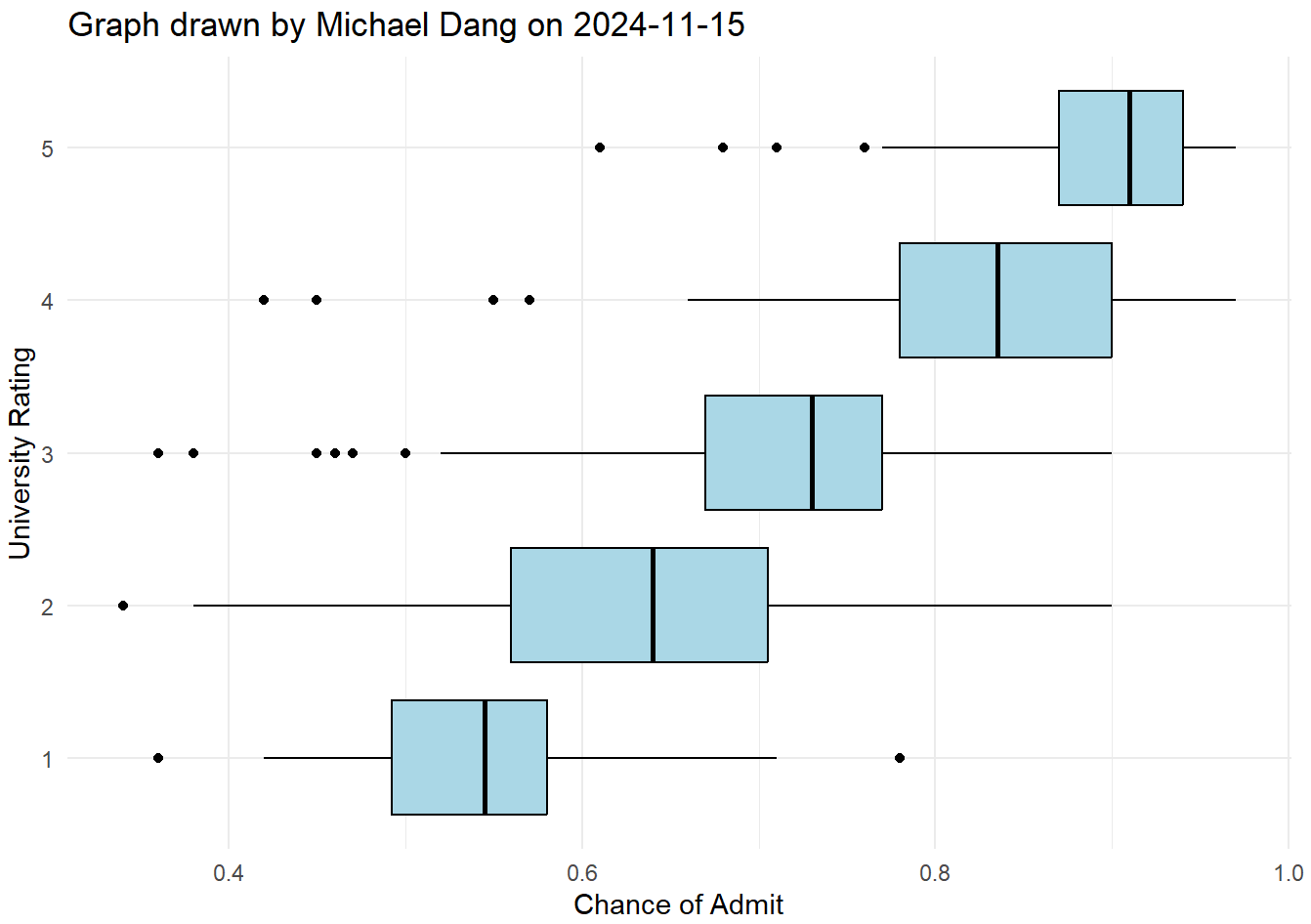
```
dataset |>
  group_by(university_rating) |>
  summarize(
    chance_of_admit_mn=mean(chance_of_admit),
    chance_of_admit_sd=sd(chance_of_admit),
    n=n())
```

```
# A tibble: 5 × 4
  university_rating chance_of_admit_mn chance_of_admit_sd     n
              <dbl>              <dbl>              <dbl> <int>
1                 1              0.548             0.0892    26
2                 2              0.626             0.112    107
3                 3              0.712             0.0958   133
4                 4              0.818             0.112     74
5                 5              0.888             0.0760    60
```

As the University Rating increase the Chance of Admit decrease. More than half of the applications were applied to 2nd and 3rd University Rating.

Box plot

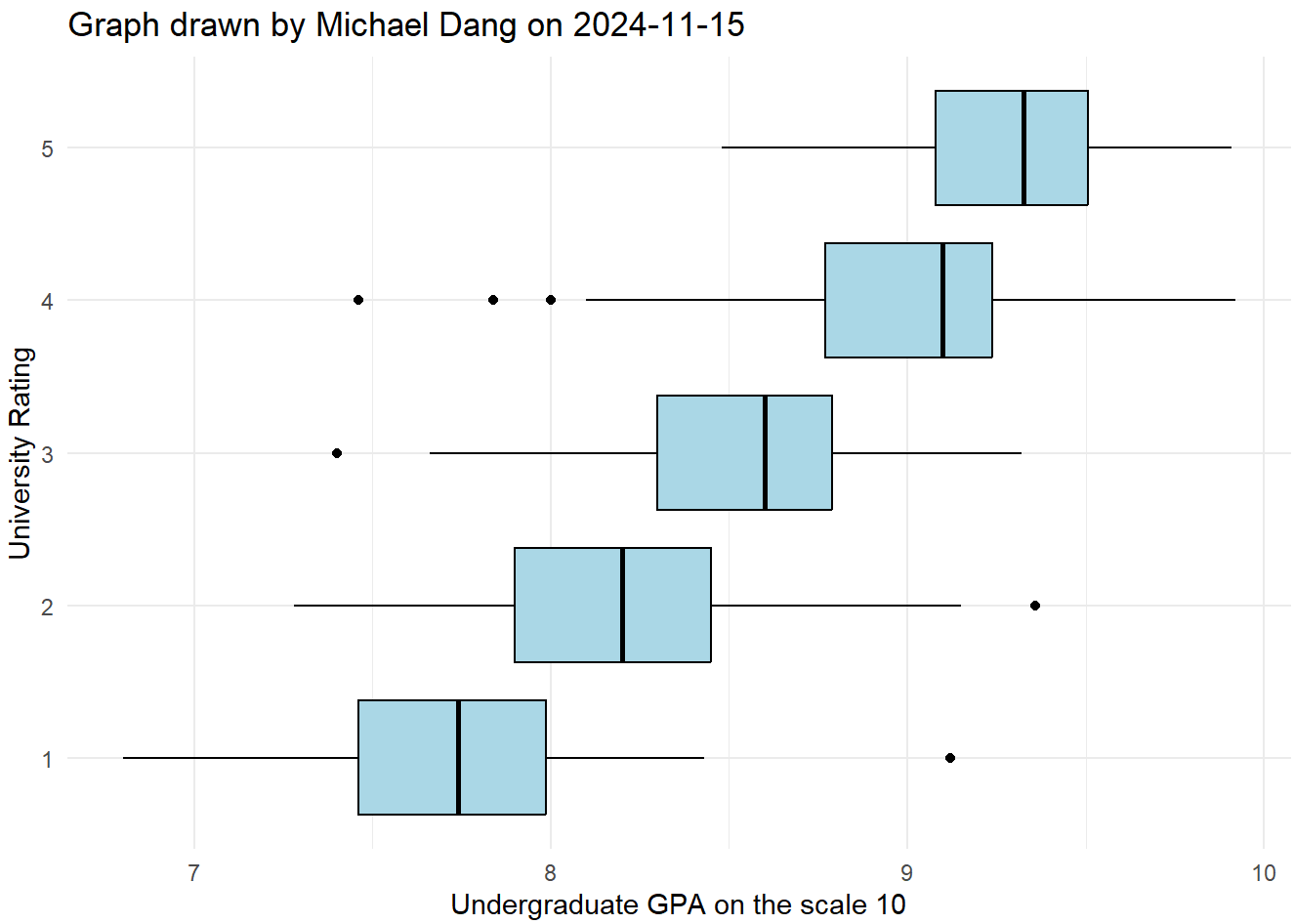
```
dataset |>
  ggplot(aes(x = as.factor(university_rating), y = chance_of_admit)) +
    geom_boxplot(fill = "lightblue", color = "black") +
    ggtitle("Graph drawn by Michael Dang on 2024-11-15") +
    xlab("University Rating") +
    ylab("Chance of Admit") +
    coord_flip() +
    theme_minimal()
```



- There is evidence of non-normality in the distribution of 'Chance of Admit' across university ratings. The graph exhibits right skewness, especially in lower ratings, and includes outliers across all ratings, with a concentration in the 3rd and 4th ratings.
- Hence non-normality assumption maybe violated and we may assume for heterogeneity.

Another boxplot

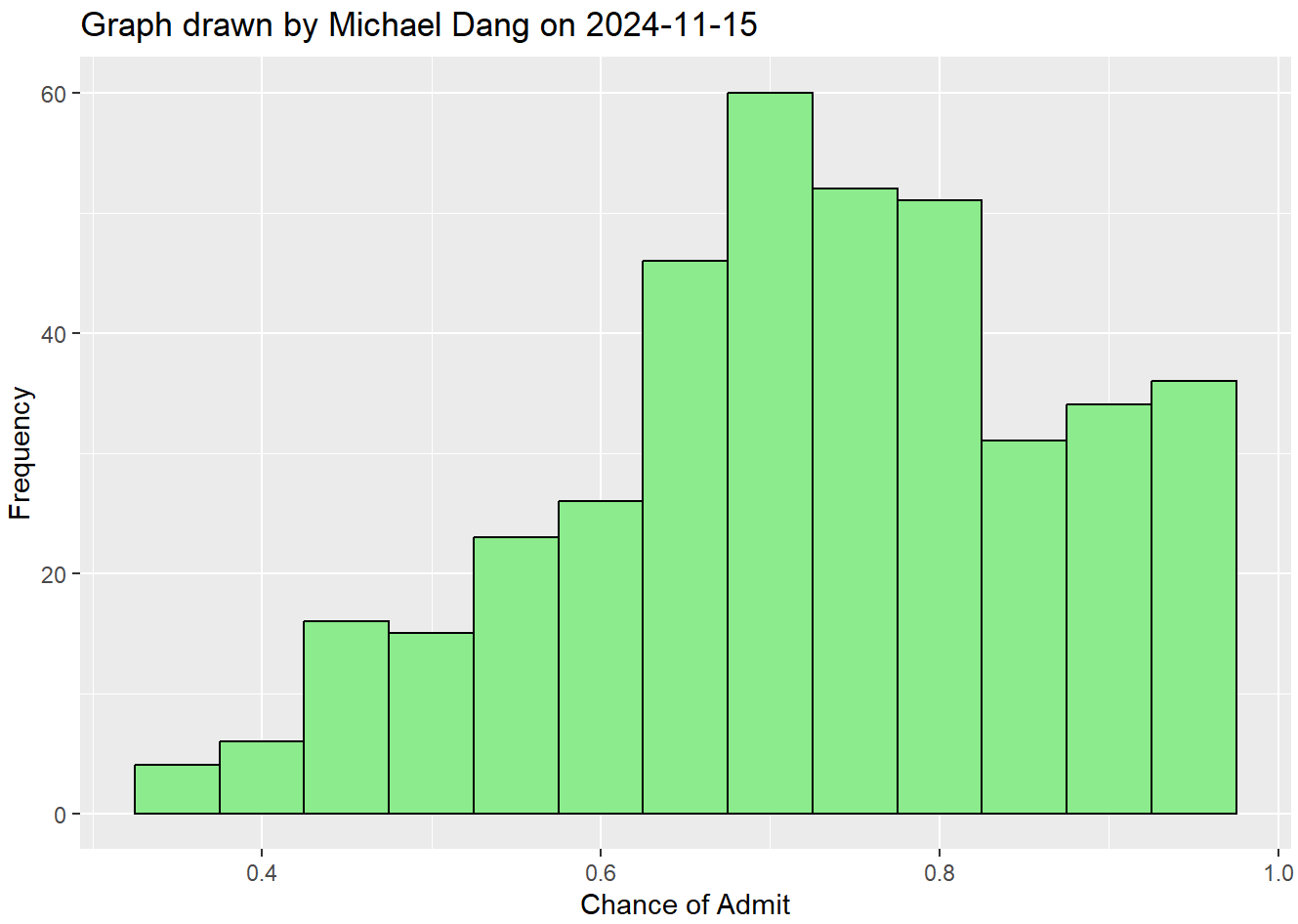
```
dataset |>
  ggplot(aes(x = as.factor(university_rating), y = cgpa)) +
    geom_boxplot(fill = "lightblue", color = "black") +
    ggtitle("Graph drawn by Michael Dang on 2024-11-15") +
    xlab("University Rating") +
    ylab("Undergraduate GPA on the scale 10") +
    coord_flip() +
    theme_minimal()
```



- The plot shows evidence of variability in GPA across university ratings. Higher university ratings are associated with higher cumulative GPAs, but the distributions are non-uniform. Outliers are present in all university ratings, with a concentrations in the 1st and 4th ratings. The data suggests some skewness, particularly in the lower ratings, indicating a diverse range of student profiles across these ratings.
- Hence non-normality assumption maybe violated and we may assume for heterogeneity.

Histogram plot

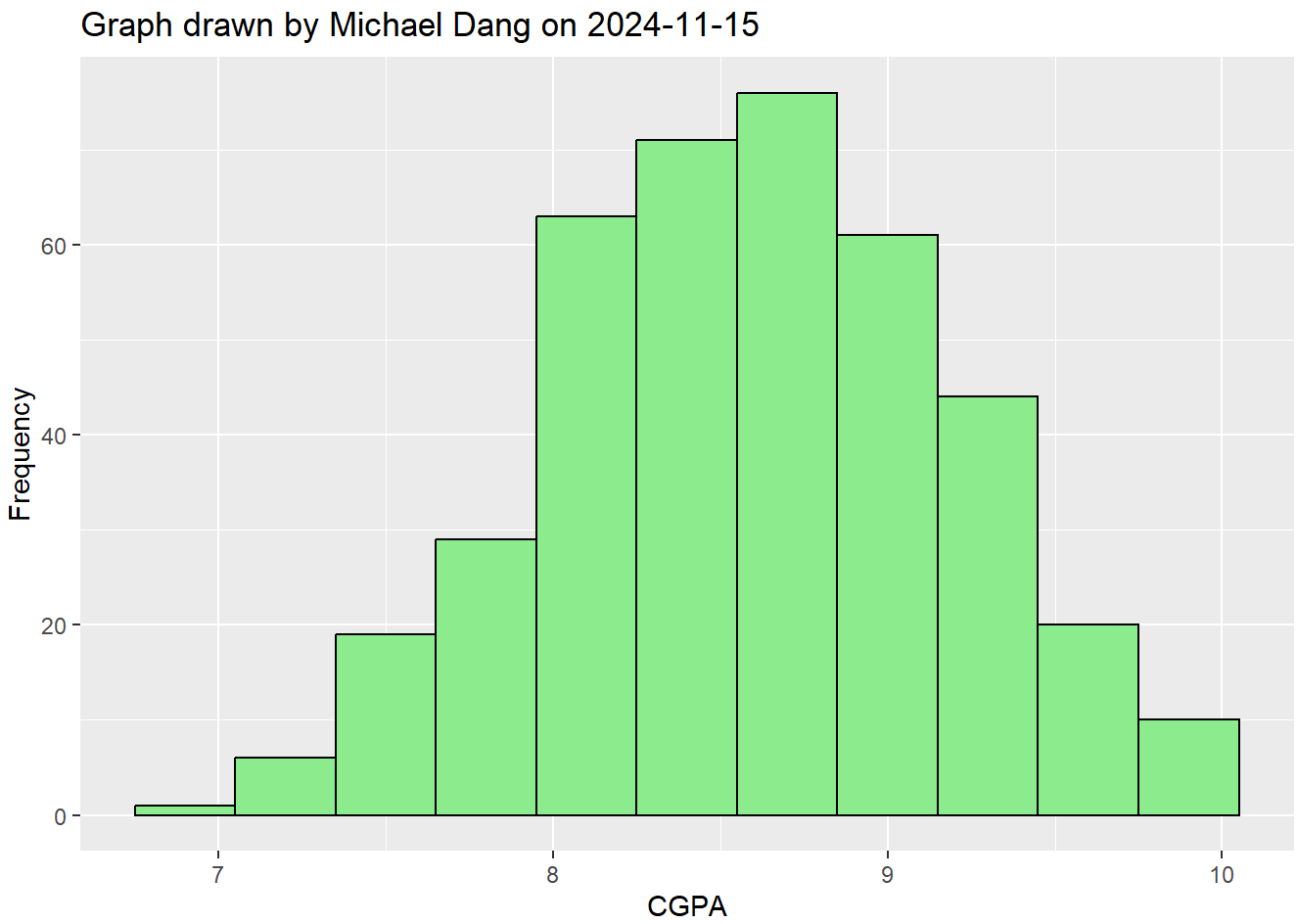
```
dataset |>
  ggplot(aes(x = chance_of_admit)) +
    geom_histogram(binwidth = 0.05, fill = "lightgreen", color = "black") +
    ggtitle("Graph drawn by Michael Dang on 2024-11-15") +
    xlab("Chance of Admit") +
    ylab("Frequency")
```



The histogram of the Chance of Admit showed the distribution is right-skewed but it is roughly center at the middle.

Another histogram plot

```
dataset |>
  ggplot(aes(x = cgpa)) +
    geom_histogram(binwidth = 0.3, fill = "lightgreen", color = "black") +
    ggtitle("Graph drawn by Michael Dang on 2024-11-15") +
    xlab("CGPA") +
    ylab("Frequency")
```



The histogram of CGPA shows roughly bell-shaped distribution, which suggests it may approximate a normal distribution.

Hypothesis test

- Null hypothesis (H_0): The mean "Chance of Admit" is the same across all university ratings.
- Alternative hypothesis (H_a): The mean "Chance of Admit" differs for at least one university rating group.
- $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$

One-way ANOVA

```
m1 <- aov(chance_of_admit ~ as.factor(university_rating), data = dataset)
tidy(m1)
```

```
# A tibble: 2 × 6
  term                          df sumsq meansq statistic   p.value
  <chr>                      <dbl> <dbl>  <dbl>     <dbl>     <dbl>
```

```
1 as.factor(university_rating)     4  4.12 1.03        102.  1.31e-59
2 Residuals                      395  3.99 0.0101       NA  NA
```

The F-ratio is large and the p-value is small. Conclude there is a differs for at least one university rating group.

Since ANOVA indicates significant differences, perform post-hoc tests (e.g., Tukey's HSD) to identify which specific groups (university ratings) differ from each other.

```
TukeyHSD(m1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = chance_of_admit ~ as.factor(university_rating), data = dataset)

$`as.factor(university_rating)`
          diff        lwr       upr     p adj
2-1 0.07790439 0.01768247 0.1381263 0.0039993
3-1 0.16380278 0.10474279 0.2228628 0.0000000
4-1 0.27003119 0.20723912 0.3328232 0.0000000
5-1 0.34008974 0.27542103 0.4047585 0.0000000
3-2 0.08589839 0.05013035 0.1216664 0.0000000
4-2 0.19212680 0.15048411 0.2337695 0.0000000
5-2 0.26218536 0.21776337 0.3066073 0.0000000
4-3 0.10622841 0.06628448 0.1461723 0.0000000
5-3 0.17628697 0.13345338 0.2191206 0.0000000
5-4 0.07005856 0.02221007 0.1179070 0.0006834
```

- Higher Ratings Have Higher Chance: Each higher university rating has a significantly higher "Chance of Admit" compared to lower ratings
- Largest Difference: The largest difference is between ratings 5 and 1, diff = 0.3401, showing that students with a university_rating = 5 have a substantially higher average chance of admit compared to rating = 1.
- Smallest Significant Difference: The smallest difference is between 5 and 4, diff = 0.0701, suggesting a smaller improvement in "Chance of Admit" between these higher-rated groups.
- Statistical Significance: All pairwise comparisons have p-values less than 0.05, indicating that the differences in means are statistically significant.