STAT40830 Adv Data Programming with R - HW1

Nishanth Chennagiri Keerthi (24212022)

Introduction

The `UCBAdmissions` dataset is a real-world dataset included in base R. It contains aggregated data on the admissions of men and women to six departments at the University of California, Berkeley, for the Fall 1973 admissions cycle. The dataset records the number of applicants by gender and admission outcome (Admitted or Rejected) across departments.

This dataset is historically significant because it was used in a widely cited example of **Simpson's Paradox**, where aggregated data seemed to show gender discrimination, but detailed department-level analysis revealed a more detailed story. It serves as an important example of how grouped categorical data must be interpreted with care.

Data Summary

We first convert the table into a data frame using base R. We then calculate the total number of admitted and rejected applicants grouped by gender using base R's `aggregate()` function.

```
# Load and convert the dataset
data(UCBAdmissions)
ucb_df <- as.data.frame(UCBAdmissions)

#total admitted vs rejected by gender
gender_summary <- aggregate(Freq ~ Gender + Admit, data = ucb_df, sum)
gender_summary
```

```
  Gender    Admit Freq
1   Male Admitted 1198
2 Female Admitted  557
3   Male Rejected 1493
4 Female Rejected 1278
```

The resulting table shows the total number of male and female applicants who were either admitted or rejected:

- Among men, 1198 were admitted and 1493 were rejected.
- Among women, 557 were admitted and 1278 were rejected.

This results in an admission rate of approximately 44.7% for men and 30.4% for women. At first glance, this appears to indicate a gender bias in admissions.

However, this observation is deceptive when not considering which departments applicants applied to. Women were more likely to apply to competitive departments with lower acceptance rates, while men applied more frequently to departments with higher acceptance rates. This phenomenon is known as Simpson's Paradox — where trends in aggregated data disappear or reverse when examined in subgroups.

> **i** Did you know?
>
> This dataset sparked a landmark 1975 study in Science by Bickel et al., which concluded that women were not discriminated against by departments — instead, they tended to apply more to departments with lower overall acceptance rates. This real-world demonstration of Simpson's Paradox changed how analysts and policymakers think about bias, fairness, and statistical storytelling. It's now one of the most frequently cited examples in the fields of statistics and data ethics.

Faceted Plot: Admission Proportions by Department and Gender

To clearly demonstrate **Simpson's Paradox**, we now visualize the admission proportions for men and women within each department using a **faceted bar chart**. This shows what proportion of applicants in each group were admitted or rejected — department by department — and makes it easier to see subgroup-specific trends that are lost in the aggregated totals.

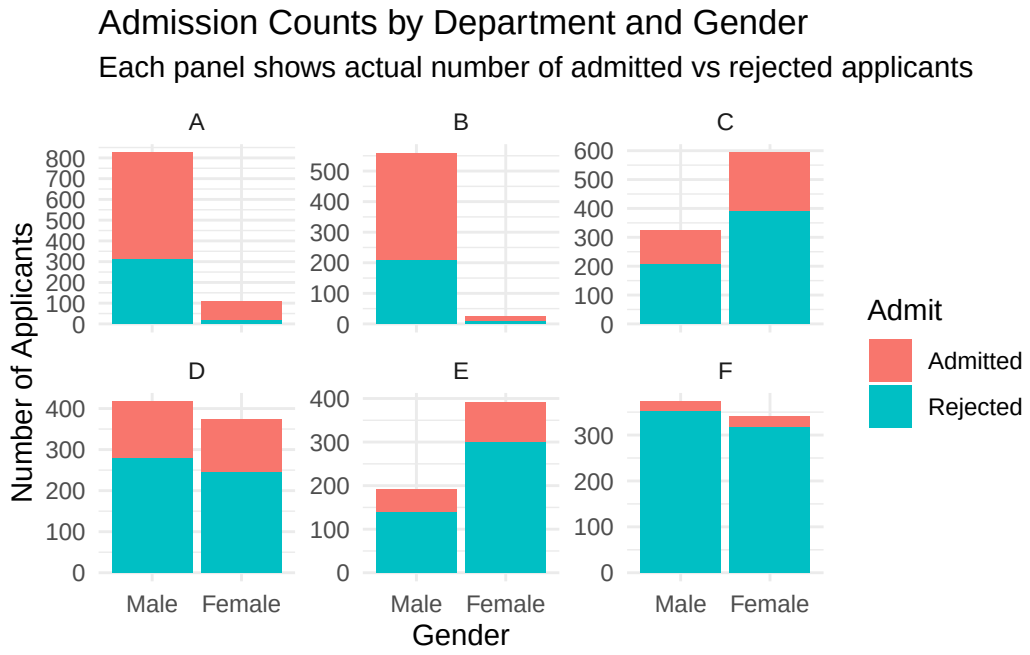

Figure 1: Bar Chart of Admission Counts by Department and Gender

Key Observations:

- Departments A and B:
  - Far more men applied than women.
  - In Department A, nearly 800 men applied versus just over 100 women.
  - Despite the higher male admission rate (visible in the larger red portion), the low number of female applicants distorts the overall picture in the aggregate.
  - This means women had a better chance of being admitted in A and B but applied less frequently.
- Departments C and D:
  - Application numbers are more balanced between men and women.
  - The number of rejections outweighs admissions for both genders, especially in Department C.
  - Gender-wise, there is no major discrepancy in treatment visible here.
- Departments E and F:
  - More women applied than men, reversing the trend seen in A and B.
  - These departments have low admission rates for both genders.
  - The stacked bars are tall and largely composed of rejections, especially for women.
  - This reveals a key insight: Women disproportionately applied to the most competitive departments (E and F), which skews the overall aggregated admission rate against them.

This makes it clear that the overall lower admission rate for women is not due to departmental bias, but rather due to differences in the departments they applied to. Women disproportionately applied to more selective departments, which had lower acceptance rates for everyone — regardless of gender.

As a result, the aggregate data misleadingly suggests gender bias, even though the department-level data tells a different, more accurate story.

This is the essence of Simpson's Paradox: a statistical illusion where trends within groups disappear or reverse when the data is combined.