20200210【Data Analyst Nanodegree】P04M01L03

> Part 04 : Practical Statistics
> Learn how to apply inferential statistics and probability to important, real-world scenarios, such as analyzing A/B tests and building supervised learning models.

- 20200210【Data Analyst Nanodegree】P04M01L03
  - Module 01: Practical Stats
    - Lesson 03: Admissions Case Study
      - 01. Admissions Case Study Introduction
      - 02. Admissions 1
      - 03. Admissions 2
      - 04. Admissions 3
      - 05. Admissions 4
      - 06. Gender Bias
      - 07. Aggregation
      - 08. Aggregation 2
      - 09. Aggregation 3
      - 10. Gender Bias Revisited
      - 11. Dangers of Statistics
      - 12. Text: Recap + Next Steps
      - 13. Case Study in Python
      - 14. Conclusion
      - 15. Appendix: Glossary

Module 01: Practical Stats

Lesson 03: Admissions Case Study

> Learn to ask the right questions, as you learn about Simpson's Paradox.

01. Admissions Case Study Introduction

In this case study, you're going to witness an instance of Simpson's paradox. A phenomenon that shows how powerful and dangerous statistics can be. Sometimes just grouping your data differently for your analysis can make your conclusions disappear or even be reversed.

02. Admissions 1

The problem I'd like to tell you about is motivated by an actual study the University of California Berkeley, which many years back wanted to know whether it's admissions procedure is gender biased. The paradox is indeed the same and is often called, "**Simpson's Paradox(辛普森悖论)**".

Data:

Male	Applied	admitted	rate
major A	900	450	50%

03. Admissions 2

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%

04. Admissions 3

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%

Female	Applied	admitted	rate
major A	100	80	80%
major B	900	180	20%

05. Admissions 4

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%

Female	Applied	admitted	rate
major A	100	80	80%
major B	900	180	20%

06. Gender Bias

- Is there a gender bias?

And I would say yes, in part because the acceptance rate is so different for the different student populations, even though the numbers are relatively large. So, it doesn't seem just like random deviations.

But the thing that will blow your mind away is a different question.

07. Aggregation

Who is being favored--the male students or the female students?

And looking at the data alone, it makes sense to say the **female students** are favored because for both majors, they have a better admission rate than the corresponding male students.

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%
both	1000	460	

Female	Applied	admitted	rate
major A	100	80	80%
major B	900	180	20%

08. Aggregation 2

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%
both	1000	460	46%

09. Aggregation 3

Data:

Female	Applied	admitted	rate
major A	100	80	80%
major B	900	180	20%
both	1000	260	26%

10. Gender Bias Revisited

And surprisingly, when you look at both majors together, you find that males have a much higher admissions rate than females.

Data:

Male	Applied	admitted	rate
major A	900	450	50%
major B	100	10	10%
both	1000	460	46%

Female	Applied	admitted	rate
major A	100	80	80%
major B	900	180	20%
both	1000	260	26%

So how come, when you do this, what looks like an admissions bias in favor of females switches into admissions bias in favor of males?

11. Dangers of Statistics

As you've seen in this example, on Simpson's paradox, the way you choose to look at your data can lead to completely different results.

And often, you can majorly impact what people believe to be true with how you choose to communicate your findings. You can guess how people intentionally or unintentionally come to false conclusions with these choices.

12. Text: Recap + Next Steps

Simpson's Paradox

In this example lesson, you learned about **Simpson's Paradox**, and you had the opportunity to apply it to a small example with Sebastian, as well as work through similar example in Python.

In the lessons ahead, you will be learning a lot by following along with Sebastian, but it is really important to put these ideas to practice using data and computing, because that is how you will apply these skills in a day to day environment as a **Data Analyst** or **Data Scientist**.

It is so easy to get caught up in looking at full aggregates of your data. Hopefully, the examples here serve as a reminder to look at your data multiple ways.

Upcoming Lessons

In the upcoming lessons, you will learn the fundamentals of probability by working through some examples. After finishing the lessons on probability with Sebastian, you will put what you learned to practice using Python!

13. Case Study in Python

Use the Jupyter notebook to analyze `admission_data.csv` to find the following values and for the quizzes below. Indexing, query, and groupby may come in handy!

1. Proportion and admission rate for each gender
2. Proportion and admission rate for physics majors of each gender
3. Proportion and admission rate for chemistry majors of each gender
4. Admission rate for each major

```python
# admission_analysis

import pandas as pd

df = pd.read_csv('admission_data.csv')

# Proportion of students that are female
df.query('gender=="female"').gender.count()/df.gender.count()


    0.514


# Proportion of students that are male
df.query('gender=="male"').gender.count()/df.gender.count()


    0.486


# Admission rate for females
df.query('gender=="female" and admitted==True').admitted.count()/df.query('gender=="female"').admitted.count()


    0.28793774319066145


# Admission rate for males
df.query('gender=="male" and admitted==True').admitted.count()/df.query('gender=="male"').admitted.count()


    0.48559670781893005


# Proportion of females with physics majors
df.query('gender=="female" and major=="Physics"').gender.count()/df.query('major=="Physics"').gender.count()


    0.12109375


# Proportion of males with physics majors
df.query('gender=="male" and major=="Physics"').gender.count()/df.query('major=="Physics"').gender.count()


    0.87890625


# Admission rate for female physics majors
df.query('gender=="female" and major=="Physics" and admitted==True').admitted.count()/df.query('gender=="female" and major=="Physics"').admitted.count()


    0.7419354838709677


# Admission rate for male physics majors
df.query('gender=="male" and major=="Physics" and admitted==True').admitted.count()/df.query('gender=="male" and major=="Physics"').admitted.count()


    0.5155555555555555


# Difference Physics majors and chemistry majors for female
df.query('gender=="female" and major=="Physics"').major.count()-df.query('gender=="female" and major=="Chemistry"').major.count()


    -195


# Difference Physics majors and chemistry majors for female
df.query('gender=="male" and major=="Physics"').major.count()-df.query('gender=="male" and major=="Chemistry"').major.count()


    207


# Proportion of males with chemistry majors
df.query('gender=="female" and major=="Chemistry"').gender.count()/df.query('major=="Chemistry"').gender.count()


    0.9262295081967213


# Proportion of males with chemistry majors
df.query('gender=="male" and major=="Chemistry"').gender.count()/df.query('major=="Chemistry"').gender.count()


    0.07377049180327869


# Admission rate for female physics majors
df.query('gender=="female" and major=="Chemistry" and admitted==True').admitted.count()/df.query('gender=="female" and major=="Chemistry"').admitted.count()


    0.22566371681415928


# Admission rate for male physics majors
df.query('gender=="male" and major=="Chemistry" and admitted==True').admitted.count()/df.query('gender=="male" and major=="Chemistry"').admitted.count()


    0.1111111111111111


df.query('admitted==True and major=="Physics"').admitted.count()/df.query('major=="Physics"').admitted.count()


    0.54296875


df.query('admitted==True and major=="Chemistry"').admitted.count()/df.query('major=="Chemistry"').admitted.count()


    0.21721311475409835
```

Quiz

1. Match the correct values

Feature	Value
Proportion of students that are female	0.514
Proportion of students that are male	0.486
Admission rate for females	0.287938
Admission rate for males	0.485597

2. By only looking at gender and admission rates, who appears to be favored in the admissions process?

- [ ] Females
- [x] Males

3. Match the correct values

Feature	Value
Proportion of females with physics majors	0.121
Proportion of males with physics majors	0.879
Admission rate for female physics majors	0.742
Admission rate for male physics majors	0.516

4. Of the students applying as physics majors, who appears to be favored in the admissions process?

- [x] Females
- [ ] Males

5. Who tends to have more physics majors than chemistry majors?

- [ ] Females
- [x] Males

6. Match the correct values

Feature	Value
Proportion of females with chemistry majors	0.926
Proportion of males with chemistry majors	0.074
Admission rate for female chemistry majors	0.226
Admission rate for male chemistry majors	0.111

7. Of the students applying as chemistry majors, who appears to be favored in the admissions process?

- [x] Females
- [ ] Males

8. Who tends to have more chemistry majors than physics majors?

- [x] Females
- [ ] Males

9. Which major has a lower admission rate?

- [ ] Physics
- [x] Chemistry

10. Take a moment to organize and explain what just happened.

There are many Simpson's Paradox happened. Can you think of other situations where Simpson's Paradox could occur?

14. Conclusion

I hope this example made you think and learn to be skeptical, of your own results and the results from others. Moving forward even when you feel very confident about the statistics you use for your analysis, take a moment to reconsider other ways of looking at your data and whether you chose wisely. Stay tuned as we dive into the basics of statistics. We'll begin with probability theory.

15. Appendix: Glossary

- Simpson's Paradox(辛普森悖论)