14. Statistics Visualizations

3ikakke

Friday 8th of April 2022

Outline

- Objectives
- Review of probabilities
- Introducing two-way tables
- Joint probabilities
- Conditional probabilities
- Review of Distributions
- Univariate numeric analysis
- Univariate categorical analysis
- Bivariate: numeric vs numeric
- Bivariate: categorical vs categorical
- Bivariate: numeric vs categorical

Objectives

- Reinforce data exploration
- Understand the value of 2 way tables
- Learn data vizualization rules based on univariate and bivariate analysis

Review of probabilities

- Remember the rules?
  - Probabilities exist between 0 and 1
  - All probabilities in a system add up to 1
  - The compliment rule: if there are only 2 probabilities then 1 - known probability = unknown probability

  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$: remember the double counting problem

- Joint probabilities: $P(A \cap B)$ means both conditions are true
- Conditional probabilities: $P(A|B)$ means probability of A being true only in the group of B (read as Probability of A given B)

Introducing two-way tables

Distribution of Pass and Fail among Private and Public Schools

	Private	Public
Passed	23	17
Failed	12	44

- Let's use keys to identify Private as **R**, Public as **U**, Pass as **A** and Fail as **F**
- Total = 23 + 12 + 17 + 44 = **96**
- Probability of attending a private school or $P(R)$ = (23 + 12)/96 = 0.3646 = **36.46%**
- Probability of attending a public school or $P(U)$ = (17 + 44)/96 = 0.6354 = **63.54%**
- Probability of passing or $P(A)$ = (23+17)/96 = 0.4167 = **41.67%**
- Probability of failing or $P(F)$ = (12+44)/96 = 0.5833 = **58.33%**

Two-way tables: Joint probabilities

	Private	Public
Passed	23	17
Failed	12	44

- Probability of attending a private school and passing (joint probability) is $P(R \cap A)$ = 23/96 = 0.2396 = **23.96%**
- Probability of attending a private school and failing (joint prbability) is $P(R \cap F)$ = 12/96 = 0.1250 = **12.50%**
- Probability of attending a public school and passing (joint probability) is $P(U \cap A)$ = 17/96 = 0.1771 = **17.71%**
- Probability of attending a pubic school and failing (joint probability) is $P(U \cap F)$ = 44/ = 0.4583 = **45.83%**
- All probabilities are between 0 and 1
- Add all joint probabilities and they come to 1! **0.2396 + 0.1771 + 0.1250 + 0.4583 = 1**

Two-way tables: Conditional probabilities

	Private	Public
Passed	23	17
Failed	12	44

- Probability of a person passing if they attend a private school (conditional probability) is $P(A|R) = 23/(23+12) =$ **0.66**
- Probability of a person failing if they attend a private school (conditional probability) is $P(F|R) = 12/(23+12) =$ **0.34**
- Probability of a person passing if they attend a public school (conditional probability) is $P(A|U) = 17/(17+44) =$ **0.28**
- Probability of a person failing if they attend a public school (conditional probability) is $P(F|U) = 44/(44+17) =$ **0.72**
- Probability that a person attended a private school if they passed (conditional probability) is $P(R|P) = 23/(23+17) =$ **0.58**
- Probability that a person attended a public school if they passed (conditional probability) is $P(U|P) = 17/(23+17) =$ **0.43**
- Probability that a person attended a private school if they failed (conditional probability) is $P(R|F) = 12/(12+44) =$ **0.21**
- Probability that a person attended a public school if they failed (conditional probability) is $P(U|F) = 44/(12+44) =$ **0.79**
- **How would these help you make a decision?**

Review of Distributions

- The histogram
  - A bell curve is a normal distribution
  - The mean approximates the median in a normal distribution
  - Report median and IQR for non-normal distributions
  - Report mean and standard deviation for normal distributions
- The five number summary:
  - Minimum
  - First Quartile (Q1)
  - Median
  - Third Quartile (Q3)
  - Maximum
- The skeleton of a boxplot

Univariate numeric analysis

- We need matplotlib and seaborn

```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns

#lets read in our data
dataset = pd.read_csv("nhanes.csv")
```

- Histograms
- Boxplots

```
sns.histplot(x='num_var', data=dataset)
plt.show()

sns.boxplot(x='num_var', data=dataset)
plt.show()
```

Univariate categorical analyis

- Frequencies
- Proportions
- Percentages
- Barplots (count plots)
- Pie charts are a bad idea!

```
sns.countplot(x='cat_var', data=dataset)
```

Bivariate: numeric vs numeric

- Correlation (Pearson correlation coefficient)
  - -1 to 0 to +1
  - -1 indicates negative coorelation (the numbers trend in opposite directions)
  - 0 means no correlation
  - +1 means the numbers trend in the same direction
- Scatterplots

```
sns.scatterplot(x='num_var1', y='num_var2', data=dataset)
```

Bivariate: categorical vs categorical

- Two way tables

- Grouped barplots

```
sns.countplot(x='cat_var1', hue='cat_var2', data=dataset)
```

Bivariate: numeric vs categorical

- Rememeber - report the numeric summaries at each level of the categorical variable

```
sns.boxplot(x='num_var', y='cat_var', data=dataset)
```

Q&A

Review of objectives

- Reinforce data exploration
- Understand the value of 2 way tables
- Learn data vizualization rules based on univariate and bivariate analysis

Gist of the day

- Get the pdf from here
- Get the gist
- The Jupyter Notebook will be uploaded

Thanks for being awesome students!