15. Statistics - AB testing

3ikakke

Monday 18th April 2022

Outline

- Objectives
- Data types
- Univariate Analysis
- Bivariate Analysis
- Null and Alternate Hypothesis
- Comparing Categorical vs Categorical
- A single Categorical variable (Goodness of fit)
- Comparing Binary vs Numeric
- Comparing Categorical (more than 2 levels) to Numeric
- Comparing Numeric to Numeric
- Review Objectives
- Q&A
- Gist of the day

Objectives

- Understand hypothesis testing
- Knowing what tests to use when testing a hypothesis

Data types

- Numeric (Quantitative)
  - Continuous
  - Count (Discrete)
- Categorical (Qualitative)

  - Binary
  - Nominal
  - Ordinal

Univariate Analysis

- Numeric
  - 5 number summary
  - Normally distributed - Mean and Standard Deviation
  - Range and IQR
- Categorical
  - Frequencies
  - Proportions
  - Percentages
- Visualizations
  - Numeric - Histograms, Boxplots
  - Categorical - Barplots

Bivariate Analysis

- Numeric vs Numeric
  - Correlation
- Categorical vs Categorical
  - 2 way tables
  - Frequencies
  - Propotions
  - Percentages
- Categorical vs Numeric
  - Numeric summaries at each level of the categorical variable
- Visualizations
  - Numeric vs Numeric - Scatterplots
  - Categorical vs Categorical - Grouped barplots and Stacked bar plots
  - Numeric vs Categorical - Grouped boxplots

Null and Alternate Hypothesis

- H_0 or the null hypothesis assumes there is no difference or things are as expected
- H_A or the alternative hypothesis is what we believe and are testing for
- p-value is the probability that we observe what we see in our data under the assumption that the H_0 is true
- The cut point used for significance is usually 0.05 (5%) and is referred to as the alpha
- Anything above 0.05 means there is no significance and we cannot reject the H_0
- Anything below 0.05 means there is significance and that we can reject the H_0
- AB Testing

Comparing Categorical vs Categorical

- Chisquare tests

```
import pandas as pd
from scipy.stats import chis2_contingency

#import dataset
nhanes = pd.read_csv("nhanes.csv")

Gender_Race = pd.crosstab(nhanes['Gender'], nhanes['Race1'])
print(Gender_Race)

chi, pvalue, dof, expected  = chis2_contingency(Gender_Race)
print(chi)
print(pvalue)
print(dof)
print(expected)
```

- Interpreting a result

A single Categorical variable (Goodness of fit)

- Chisquare test

```
import pandas as pd
from scipy.stats import chisquare

Race = nhanes['Race1'].value_counts()

chi, pvalue  = chisquare(Gender_Race)
print(chi)
print(pvalue)
```

- ?Interpretation

Comparing Binary vs Numeric

- T-Test

```
from scipy.stats import ttest_ind

male_age = nhanes.loc[nhanes['Gender'] == 'male', 'Age']
female_age = nhanes.loc[nhanes['Gender'] == 'female', 'Age']

print(male_age.mean(), female_age.mean())

tstat, pvalue = ttest_ind(male_age, female_age)
```

Comparing Categorical (more than 2 levels) to Numeric

- ANOVA (Analysis of variance)

```
from scipy.stats import f_oneway as anova

white_age = nhanes.loc[nhanes['Race1'] == 'White', 'Age']
black_age = nhanes.loc[nhanes['Race1'] == 'Black', 'Age']
hispanic_age = nhanes.loc[nhanes['Race1'] == 'Hispanic', 'Age']

fstat, pvalue = anova(white_age, black_age, hispanic_age)
```

Comparing Numeric to Numeric

- A regression is simply a line
- All straight lines can be defined by an equation

- Simple Linear Regression
- $y = mx + c$
  - where y = what we want to predict or the dependednt variable

  - x = the predictor variable or independent variable

  - c = intercept (the value of y when x = 0)
  - m = the slope also known as $\frac{rise}{run}$
- $y = c + mx$

- $y = \beta_0 + \beta_1 x$
  - where y = what we want to predict or the dependednt variable

  - x = the predictor variable or independent variable

  - β_0 = intercept (the value of y when x = 0)
  - β_1 = the slope also known as $\frac{rise}{run}$

Review Objectives

- Understand hypothesis testing

- Knowing what tests to use when testing a hypothesis

Q&A

Gist of the day

- Get the gist
- Get the pdf

- The Jupyter Notebook will be uploaded

We continue on Friday