# 13. Statistics - Basics

3ikakke

Monday 4th April 2022

## Outline

- Objectives
- Data types
- Normal Distributions
- Center and spread
- Probability
- Exploring single numeric data
- Exploring single categorical data
- Exploring numeric and categorical data
- Exploring 2 categorical variables
- Exploring 2 Numeric:
- Review Objectives
- Q&A
- Gist of the day

## Objectives

- Understand univariate analysis
- Understand bivariate analysis
- Understand probabilities
- Understand numeric distributions
- Understand correlations

## Data types

- Numeric

  - Discrete (Count)
  - Continuous
- Categorical
  - Nominal
  - Ordinal
  - Binary
- Missing
- Recall Python mapping

## Normal Distributions

- Bell curve
- Mean usually at the peak
- How variable is this data?
- How much does it deviate from the center?
- Standard deviation

## Center and spread

- Normal
  - Mean
  - Standard Deviation
- Not normal
  - Median
  - IQR

## Probability

- Based on proportions: how often would we find one from a category?
- Mathematical notation:
  - $\cup$ Union
  - $\cap$ Intersection
  - $|$ given
- Rules of probability
  1. Probabilities range between 0 and 1.
  2. Sum of all probabilities = 1
  3. For only 2 parts $1 - P(A) = P(B)$ This is called the compliment rule where P(B) is called the complement of P(A)

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \cap B) = P(A) * P(B|A)$ or $P(A \cap B) = P(A|B) * P(B)$ and if A and B are independent then $P(A \cap B) = P(A) * P(B)$

6. $P(A|B) = \frac{P(A)*P(B \cap A)}{P(B)}$ This is Bayes rule.

## Exploring single numeric data

- Distribution

- Normal vs Not Normal

- Non Normal
  - Five number summary
    - Minimum

    - First Quartile (Q1)

    - Median (Q2) - Reported for non normal distributions

    - Third Quartile (Q3)

    - Maximim
  - Additional summaries
    - Range (Maximum - Minimum)

    - Inter Quartile Range (Q3 - Q1)- reported for non normal distributions
    - mean - reported for normal distributions

    - standard deviation - reported for normal distributions

## Exploring single numeric data with python

```
import pandas as pd
dataset = pd.read_csv("some_dataset.csv")

var_min = dataset['var'].min()
var_max = dataset['var'].max()
var_Q1 = dataset['var'].quantile(0.25)
var_Q3 = dataset['var'].quantile(0.75)
var_median = dataset['var'].median()
var_mean = dataset['var'].mean()
var_sd = dataset['var'].std()

var_range = vat_max - var_min
var_iqr = var_Q3 - var_Q1
```

## Exploring single categorical data

- Frequencies (How many in each category?)

- Proportions (Frequencies expressed as fractions)

- Percentages (Proportions multiplied by 100 to convert to percentages)

## Exploring single categorical data (Pandas)

```
dataset['var'].value_counts() #Frequencies
dataset['var'].value_counts()/dataset['var'].value_counts().sum()
```

## Exploring numeric and categorical data

- Look at the numeric summaries at each level of the categorical variable
- Example - Sex vs Age
  - Report 5 number summary for males
  - Report 5 number summary for females
  - Compare the 2 summaries

## Exploring numeric and categorical data

```
dataset[['quant', 'cat']].groupby('cat').min()
dataset[['quant', 'cat']].groupby('cat').max()
dataset[['quant', 'cat']].groupby('cat').quantile(0.25)
dataset[['quant', 'cat']].groupby('cat').quantile(0.75)
dataset[['quant', 'cat']].groupby('cat').mean()
dataset[['quant', 'cat']].groupby('cat').median()
dataset[['quant', 'cat']].groupby('cat').std()
```

## Exploring 2 categorical variables

- Frequencies and Proportions
- Use a two way table
- This reinforces the concept of $\cup$ Unions, $\cap$ Intersections and $|$ conditional probability (given)
- Example comparing Sex to Employment Status

| | Employed | Unemployed |
|---|---|---|
| Female | 25 | 27 |
| Male | 20 | 28 |

- A two-way table tells so many stories all at once

## Exploring 2 categorical variables

```
pd.crosstab(dataset['cat1'], dataset['cat2'])
```

## Exploring 2 Numeric:

- Correlation:
- Pearson's correlation coefficient (R)

- $R^2$
- Ranging from -1 through 0 to +1
  - Negative correlation means as one increases the other reduces
  - Positive correlation means they both increase or decrease together
  - 0 means there is no correlation
  - The closer to -1 ot +1 the stronger the correlation
- Does one change with the other?

### Exploring 2 Numeric:

```
dataset[['quant1', 'quant2']].corr()
```

### Review Objectives

- Understand univariate analysis
- Understand bivariate analysis
- Understand probabilities
- Understand numeric distributions
- Understand correlations

### Q&A

### Gist of the day

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

### We continue on Friday