Chapter 4: Correlation and Experimental Design

Created	@April 16, 2024 6:41 PM
Class	Introdiction to Statistics in Python

Correlation and Experimental Design

Relationships Between Numeric Variables

- Visualized using scatter plots
- Example: Relationship between total sleep and REM sleep in mammals
  - x-axis: Explanatory/Independent Variable
  - y-axis: Response/Dependent Variable

Correlation Coefficient

- Quantifies the linear relationship between two variables
- A number between -1 and 1
- Magnitude represents strength of relationship
- Sign (+ or -) represents direction of relationship

Strength of Correlation

- 0.99: Near-perfect/very strong relationship
- 0.75: Strong relationship
- 0.56: Moderate relationship
- ~0.2: Weak relationship
- ~0: No relationship (random scatter)

Direction of Correlation

- Positive: As x increases, y increases
- Negative: As x increases, y decreases

Visualizing with Scatter Plots

- Import seaborn: `import seaborn as sns`
- Create scatter plot:

```
import seaborn as sns
sns.scatterplot(x="col_x", y="col_y", data=df)
```

- Add trendline:

```
sns.lmplot(x="col_x", y="col_y", data=df, ci=None)
```

Calculating Correlation

- Use `.corr()` method on pandas Series
  - `df["col_x"].corr(df["col_y"])`
  - Order doesn't matter: `corr(x,y) == corr(y,x)`
- Pearson correlation (most common):
  - Represented by r
  - Formula:

$$\sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x . \sigma_y}$$

- Other methods: Kendall's tau, Spearman's rho (beyond scope)

Caveats of Correlation

Non-Linear Relationships

- Correlation only measures linear relationships
- For non-linear relationships like quadratic, correlation can be misleading
  - Example: x and y had a clear relationship, but low correlation of 0.18 due to quadratic pattern
- Always visualize data in addition to calculating correlation

Skewed Data

- Highly skewed data can obscure relationships
  - Example: Weak correlation (0.3) between body weight and awake time in mammals
- Can apply transformations to make relationship more linear:
  - Log transformation: `msleep['log_bodywt'] = np.log(msleep['bodywt'])`
  - Other options: Square root, reciprocal, combinations
- After log transforming body weight, correlation with awake time increased to 0.57

Importance of Transformations

- Some statistical methods require linear relationships
  - Correlation coefficients
  - Linear regression

Correlation Does Not Imply Causation

- If x and y are correlated, x does not necessarily cause y
- Example: High correlation (0.99) between margarine consumption and divorce rate
  - But margarine doesn't cause divorces (spurious correlation)

Confounding Variables

- A third, hidden variable can cause a spurious correlation
  - Example: Coffee and lung cancer are correlated
    - But smoking is a confounder that causes both
- These confounders are also called "lurking variables"
- Example: Holidays and retail sales
  - Promotions/deals are a potential confounder

Practice Responsibly

- Don't blindly use correlation
- Visualize data and account for potential confounders

Design of Experiments

Introduction

- Data is often created as a result of studies aimed at answering specific questions
- Data analysis and interpretation depends on how the data was generated and the study design

Experiments vs Observational Studies

Experiments

- Aim to answer: "What is the effect of the treatment on the response?"
- Treatment = Explanatory/Independent Variable
- Response = Dependent Variable
- Example: Effect of advertisement on products purchased
  - Treatment = Advertisement
  - Response = Number of products purchased

Controlled Experiments

- Participants randomly assigned to treatment or control group

```
import random
participants = [1, 2, 3, ...]# List of participant ids
treatment_group = random.sample(participants, k=50)
```

```
control_group = [p for p in participants if p not in treatmen
t_group]
```

- Treatment group receives treatment, control does not
- Groups should be comparable to isolate treatment effect
- Example: A/B testing

Eliminating Bias

- Use randomized controlled trials
  - Random assignment to treatment/control
- Use placebos
  - Resembles treatment but has no effect
  - Participants don't know if receiving actual treatment
  - Used in clinical trials where patients are grouped into treatment and controlled group.
- Double-blind experiments
  - Administrators also don't know actual treatment
  - Used in clinical trials where the doctor also doesn’t know whether they are assigned to treatment or controlled group

Observational Studies

- Participants not randomly assigned, self-selected
- Cannot establish causation, only association
- Used when cannot force treatment
  - `smoking_data = data[data['smoker'] == True]`
  - `purchase_data = data.groupby('past_purchases')...`
- Groups may not be comparable
- Can control for confounders to strengthen conclusions

Longitudinal vs Cross-Sectional Studies

- Longitudinal: Same participants followed over time
- Cross-sectional: Single snapshot in time
  - Can be confounded by external factors (generation, lifestyle, etc.)
- Longitudinal more expensive but controls for confounders

Key Principles

Study Type	Causation	Bias Control	Confounding
Experiment	Yes	Random assignment, placebos, blinding	Low
Observational	No	Control for known confounders	Potential confounding
Longitudinal	Yes	Follows same participants	Minimizes confounding
Cross-Sectional	Limited	-	High potential for confounding