PH614: 04. Bivariate Analysis

Professional Skills for Public Health

2022

Learning objectives

- Understand two way tables
- Understand marginal probabilities
- Understand joint probabilities
- Understaning conditional probabilities
- Difference between probabilites and odds
- Understanding Pearsons correlation coefficient
- Understanding how to assess association between:
  - Two numeric variables
  - Two categorical variables
  - A numeric and a categorical variable

Bivarite analysis

- This refers to comparing two variables to see how they are related - if there is some association between them or not
- Since we have established that broadly speaking there are two types of variables: numeric and categorical we can consider each possible pair combination
  - Numeric vs Numeric
  - Categorical vs Categorical
  - Numeric vs Categorical

Probability and 2-way tables

- When considering a single categorical variable we look at frequencies and proportions
- When comparing two categorical variables similarly we look at two way tables
  Two way table

	Cancer	No Cancer
Smoking	8	12
Non Smoking	15	42

- **Marginal probabilities** $P(A)$ - Probability of being a smoker, probability of being a non smoker,

probability of having cancer and probability of not having cancer

- **Joint probabilities** $P(A \cap B)$ - Probability of being a smoker and having cancer, probability of being a smoker and not having cancer, probability of not being a smoker and having cancer, probability of not being a smoker and not having cancer. Read as probability of A intersection B.
- **Conditional probabilities** $P(A|B)$ - probability of being a smoker among people with cancer, probability of being a smoker among people without cancer, probability of being a non smoker among people with cancer and probability of being a non smoker among people without cancer. Read as probability of A given B

- Row frequencies and proportions versus column frequencies and percentages

- Probability vs odds

Categorical vs Categorical

- Looking for association between two categorical variables with (**PROC FREQ**)

```
*Association between Sex and Diabetes;
PROC FREQ DATA=PH614.NHANES_MINI;
TABLE Sex * Diabetes;
RUN;

*Multiple two way tables;
PROC FREQ DATA=PH614.NHANES_MINI;
TABLE Sex * Diabetes;
TABLE Race * Education;
RUN;

*This is different from association as it does not include the asterisk;
PROC FREQ DATA=PH614.NHANES_MINI;
TABLE Sex Diabetes;
RUN;
```

Correlation

- The Pearson's correlation coefficient is a measure of how much a numeric variable changes in comparision to another

- The range is from -1 to +1

- -1 indicates a strong negative correlation meaning the as one numeric variable increases the other decreases and vice versa example in HIV as HIV viral load increases CD4 decreases and vice versa

- +1 indicates a strong positive correlation meaning the numeric variables increase together and decrease together eg in the first 18 years of life as age increases height increases

- 0 means there is no association

Numeric vs Numeric

- Looking for correlation between two numeric variables (**PROC COR**)

```
PROC CORR DATA=PH614.NHANES_MINI;
VAR Weight Height;
RUN;

*More variables may be included to create a correlation matrix;
PROC CORR DATA=PH614.NHANES_MINI;
VAR Weight Height Age;
RUN;
```

Numeric vs Categorical

- Looking for association accross a numeric and categorical variable (**PROC MEANS**)
- Here we consider the numeric summaries as we did in univariate analysis but at every level of the categorical variable so see how those summaries change accross the categories

```
PROC MEANS DATA=PH614.NHANES_MINI MEAN STD;
CLASS Sex;
VAR Height;
RUN;
```

Review of learning objectives

- Understand two way tables
- Understand marginal probabilities
- Understand joint probabilities
- Understaning conditional probabilities
- Difference between probabilites and odds
- Understanding Pearsons correlation coefficient
- Understanding how to assess association between:
  - Two numeric variables
  - Two categorical variables
  - A numeric and a categorical variable

Q&A

Next...

- Hypothesis testing (Biostats - Random variables, significance and hypothesis testing)
  - PROC TTEST
  - PROC ANOVA
  - PROC FREQ - with CHISQ
  - PROC REG