PH614: 03. Univariate Analysis

Professional Skills for Public Health

2022

Learning objectives

- Understand numeric summaries
- Understand categorical summaries
- Carry out a univariate analysis
- Understaning how to set text length
- Understanding how to create labels for data
- Understanding formats

Quick review of what we should know

- Data is generally classed in two:
  - Numeric (Quantitative)
  - Categorical (Non-numeric or Qualitative)
- Categorical data may be:
  - Binary (with only two levels)
  - Have more than two levels

```
/*
Also remember
Numeric
  Continuous
  Count
Categorical
  Binary
  Nominal
  Ordinal
*/
```

Univariate analysis

- This simply refers to analysing a single variable
- Broadly we could consider variables as **numeric** or **categorical**
- Univariate analysis will involve three parts generally being:

  - **Variable summaries**
  - **Visualization of variable distibution**
  - **Hypothesis testing**

- Our focus today is on variable summaries
- Every data analysis begins with understanding the data; the number of records (observations) and the number of variables and what type each is

First look

```
libname PH614 '/path/to/folder';

/*
Let's explore the contents of a dataset in our library
This will inform us of the number of records and variables
It will also make a guess of each variable type
*/

PROC CONTENTS DATA=PH614.NHANES_mini;
RUN;
```

Numeric summaries

- What summaries we need from a numeric variable
  - Number of expected records (**PROC CONTENTS** lets us know this)
  - Number of missing records
  - Data distribution (**histogram**)
    - Normal or not Normal
  - Minimum value
  - Maximum value
  - First Quartile (Q1)
  - Median (Q2)
  - Third Quartile (Q3)
  - Mean
  - Standard Deviation
  - Range (Maximum - Minimum)
  - Inter Quartile range (Q3 - Q1)

Measures of center and spread

- A histogram will tell us if our data is normally distributed or not
- A bell shaped curve suggests normality, similar curves may be considered approximately normal (this is

a judgement call)

- When a numeric variable is normally distributed then:
  - It's center is represented by the mean (arithmetic average)
  - It's spread is represented by the standard deviation ($\sigma = 68\%$, $2\sigma = 95\%$, $3\sigma = 99.7\%$)

- When a variable is not normally distributed then:
  - It's center is best represented by the median (Q2) as this is closer to the true center of the data and more resilient than the mean which tends to be dragged in the direction of outliers

  - It's spread is consequently best represented by the interquartile range (IQR) which represents the central 50%

The univariarte analysis with SAS

- There are two ways of getting this done with SAS using either the **UNIVARIATE** proceedure or the **MEANS** procedure

- First lets look at the univariate procedure AKA **PROC UNIVARIATE**

```
PROC UNIVARIATE DATA=PH614.NHANES_mini;
VAR Age; *This line tells SAS to carry out an analysis of the numeric variable Age;
HISTOGRAM Age; *This line tells SAS to draw a histogram of the Age variable;
RUN;
```

- Next we consider PROC MEANS which will allow us specify what we want to see as summaries

```
/*
Here we list what summaries we want and in no particular order
MIN: Minimum
MAX: Maximum
Q1: First Quartile
Q3: Third Quartile
MEDIAN: Median
MEAN: Mean
RANGE: MAX - MIN
QRANGE: Q3 - Q1
MEAN: Mean
STD: Standard Deviation
*/
PROC MEANS DATA=PH614.NHANES_mini MIN Q1 MEDIAN Q3 MAX RANGE QRANGE MEAN STD;
VAR Age;
RUN;
```

Categorical summaries

- For categorical variables we are interested in
  - Number of expected records

  - Number of records missing the variable

  - Number of levels/categories of the variable

  - Names of the levels/categories
  - Number of records in each category (frequency)
  - Proportion of records in each category
  - Percentage of records in each category

Using SAS for categorical univariate analysis

- The name UNIVARIATE was already used for numeric summaries so SAS uses something different for categorical variables
- It looks for frequencies hence the name **FREQ**

```
PROC FREQ DATA=PH614.NHANES_mini;
*This simply says create a frequency table for the education variable;
TABLE Education;
RUN;
```

Determining length of variables

```
DATA new_data;
SET existing_data;
*Usually you dont need this for numeric variables;
LENGTH new_categorical $10.;
RUN;
```

Descriptive labels for variables

- Often variable names are made so they can make some sort of sense without being completely intelligible
- This can be a problem when sharing analysis!
- To fix this problem temporary labels can be added to variables in datasets

```
DATA new_data;
SET existing_data;
LABEL cat_var='Descriptive name';
LABEL num_var='A numeric variable';
RUN;

PROC FREQ DATA=new_data;
TABLE cat_var;
RUN;

PROC UNIVARIATE DATA=new_data;
TABLE num_var;
RUN;
```

Creating formats

- Sometimes data is encoded and you need to have a way of decoding the data without having to edit every single record

- This can be done using the SAS procedure to create formats (FORMAT)

- Formats can be used by multiple datasets so does not require setting a dataset when defining them

```
PROC FORMAT;
*Numeric formats require just the format name;
VALUE yesno
    0 = 'no'
    1 = 'yes';

*Categorical variables require a dollar sign;
VALUE $sex
    'm' = 'Male'
    'f' = 'Female';
RUN;
```

Using the format

- Formats can be used in **DATA** or **PROC** steps

```
*You can now use the format in a DATA step;
DATA new_var;
SET existing_var;
FORMAT Diabetes yesno. Sex $sex.;
RUN;

*You can also use the format in a PROC step;
PROC FREQ DATA=new_data;
TABLE cat_var;
FORMAT cat_var $sex.;
RUN;
```

Review of learning objectives

- Understand numeric summaries

- Understand categorical summaries

- Carry out a univariate analysis

- Understaning how to set text length
- Understanding how to create labels for data

- Understanding formats

Q&A

Next…

- Bivarite analysis

- Probability and two way tables

- Interpreting correlation (Pearson's correlation coefficient)
- Looking for correlation between two numeric variables (**PROC COR**)
- Looking for association between two categorical variables with (**PROC FREQ**)
- Looking for association accross a numeric and categorical variable (**PROC MEANS**)