Chapter 2

Descriptive Statistics and Simple Inference

Descriptive Statistics

- Start with samples obtained from some population
- Descriptive statistics tell us about features of the data
- Data visualizations can allow us to see features in the data
- Allow us to infer characteristics of the population

Some Uses of Descriptive Analysis

- Preliminary exploration
- Checks of assumptions
- Foundations for test statistics
- Foundations for more complicated models

Water Example

The data:

- 61 data points from towns in England
- Mortality rate per 100,000 males (averaged over 1958-1964)
- Calcium concentration (higher = harder water) in ppm in the town's drinking water
- Indicator for northern town

Plot of the Data

- Start with a scatter plot of mortality vs. hardness
- Do we notice anything interesting?

Univariate Statistics

- Obtain univariate statistics for mortality and hardness (ignoring location)
- What do they tell us about the mortality and calcium concentrations?
- What univariate visualizations might be useful?
- Let's add **ods** statements to simplify the results

Distributional Check

- T test assumes an underlying normal population
- Qualitatively check by looking at the univariate plots
- Quantitatively check by using distributional goodness of fit tests (EDF tests)

Goodness of Fit Tests

- Null hypothesis: population is of the assumed type of distribution
- Alternative: the null is not true
- Test statistic typically based on some measure of distance between the theoretical and empirical CDFs
- See **Details>Goodness-of-Fit Tests** under **proc univariate**

Normality vs. EDF

- **normal** option to the **histogram** statement specifies distribution to compare against
- **normal** option to **proc univariate** specifies that tests of normality should be performed
- Normality tests included the EDF tests
- The **ods** tables have different names

Testing for Correlation

- Pearson correlation
- Spearman rank correlation

Exercise: Analysis by Location

- Previous analyses were for all the data
- Could perform same analyses for sub-populations (north and south)
- Perform visualizations, univariate analyses, & correlation checks by **location**
- What conclusions can we draw for north?
- For south?

Exercise: Non-zero Null Value

Proc univariate assumes μ_0=0.

- Instead, take μ_0=1500 for **mortal** and μ_0=45 for **hardness**.
- Test the alternatives that the northern population values are significantly different from μ_0.
- Do the same for the south.

One Sample T-Tests with **proc ttest**

Proc ttest is specifically for t tests

- Null value is 0 by default
- Can set non-default null value
- Can choose which tail of distribution for one-sided alternatives
- Can still test **by** a categorical variable (like geographic location)

Tests of Population Differences

- **Proc ttest** to test difference of means assuming normality
- Should check assumption of equal variance via an F-test (generated by **ttest** by default)
- **Proc npar1way** with **wilcoxon** option if normality is not reasonable

class & **by** Statements

- Classifying according to a variable, use **class** statement
- Doing separate analyses, use **by** statement
- Testing north vs. south, **location** will be a classification variable

Exercise: Differences between North & South

- Use **proc ttest** or **proc npar1way** as appropriate
- Test for a significant difference between the mortality rates in the north and south
- Do the same for the water hardness values
- What are our conclusions?

Example: Data Transformation

- Could transform variable toward normality
- Try log transform of hardness variable
- Add variable **lhardnes** = log(**hardness**)
- Check normality assumption on **lhardnes**
- Perform t-test if normality is not too unreasonable