simon-5507-05-slides

Topics to be covered

- What you will learn
  - Labels and formats
  - Descriptive statistics
  - Correlations and scatterplots
  - Boxplots and group statistics
  - Investiage anomaly

Overview

Today, you will analyze some data sets that have a mix of categorical and continuous variables. The first data set looks at pulmonary function in a group of children.

You can find a description of this data set at

http://jse.amstat.org/datasets/fev.txt

;

Labels and formats

- Document your data
  - Formats for number codes
  - Labels for all variables

Speaker notes

Documenting your code is important. If your categorical data uses number codes, make sure to define what those codes represent using proc format.

Also use labels for all variables, with the possible exception of simple variables that are self-explanatory with the variable name alone.

Documentation header; (1)

```
* 5507-05-simon-working-with-a-mix-of-variables.sas
  author: Steve Simon
  date created: 2018-11-27
  purpose: to illustrate how to work with
    data that has a mix of categorical and
    continuous variables.
  license: public domain;
```

Speaker notes

- Comments on the code: Documentation header;

File locations; (2)

```
%let path=q:/5507-2025b/05;

ods pdf
  file="&path/results/5507-05-simon-working-with-mix-of-variables.pdf";

filename raw_data
  "&path/data/fev.txt";

libname perm
  "&path/data";
```

Speaker notes

- Comments on the code: File locations;

Reading the data using a data step (3)

```
data perm.fev;
  infile raw_data delimiter=",";
  input
    age 2-3
    fev 5-11
    ht 12-16
    sex 19
    smoke 25;
  label
    age=Age in years
    fev=Forced Expiratory Volume (liters)
    ht=Height in inches
    sex=Sex
    smoke=Smoking status
```

Speaker notes

- Comments on the code: Reading the data using a data step

The data file is comma delimited and the first row includes variable names.

Normally, this means that you can save a bit of time by using proc import, but I chose to read in the data using a data step. The number of variables was so small that this didn’t matter that much. It also allowed me to define variable labels in the initial data step rather than later.;

Label your categorical variables; (4)

```
proc format;
  value fsex
    0 = "Female"
    1 = "Male"
  ;
  value fsmoke
    0 = "Nonsmoker"
    1 = "Smoker"
  ;
run;
```

Speaker notes

- Comments on the code: Label your categorical variables;

Print the first ten rows of data (5)

```
proc print
    data=perm.fev(obs=10);
  format
      sex fsex.
      smoke fsmoke.;
  title1 "Pulmonary function study";
  title2 "There are no obvious problems with this dataset";
run;
```

Speaker notes

- Comments on the code: Print the first ten rows of data

It’s always a good idea to peek at the first few rows of data.

Break #1

- What you have learned
  - Labels and formats
- What’s coming next
  - Descriptive statistics

Descriptive statistics

- Use proc freq to look for
  - Inconsistencies
  - Rare/missing category levels
- Use proc means to look for
  - Unusual minimum/maximum values
  - Zero standard deviation

Speaker notes

There is a mix of categorical and continuous variables in this data set. Recall that you use proc freq for categorical variables and proc means for continuous variables.

Always get in the habit of checking for missing values.

Look for problems. This could mean a lot more categories than you expected, a particular category level that is unexpectedly small, or multiple categories caused by misspelling or inconsistent capitalization. There are no problems here.

Look for minimum or maximum values that are unusual. Also make sure that you don’t have a continuous variable that is constant (zero variation).

Get statistics for categorical variables; (6)

```
proc freq
    data=perm.fev;
  tables sex smoke / missing;
  format
      sex fsex.
      smoke fsmoke.;
  title2 "Frequency counts";
run;
```

Speaker notes

- Comments on the code: Get statistics for categorical variables;

Get statistics for continuous variables; (7)

```
proc means
    n nmiss mean std min max
    data=perm.fev;
  var age fev ht;
  title2 "Descriptive statistics";
run;
```

Speaker notes

- Comments on the code: Get statistics for continuous variables;

Break #2

- What you have learned
  - Descriptive statistics
- What’s coming next
  - Correlations and scatterplots

Look at relationships

- Between two continuous variables
  - Correlations, Scatterplots
- Between two categorical variables
  - Crosstabulations
- Between a categorical and a continuous variables
  - Boxplots

Speaker notes

After looking at descriptive statistics for individual variables, you should look for relationships between pairs of variables.

Correlations, covered earlier, are a simple way to examine relationships between two continuous variables.

Remember the cut-offs. A correlation between +0.7 and 1.0 implies a strong positive association. A correlation between +0.3 and +0.7 implies a weak positive association. A correlation between -0.3 and +0.3 implies little or no association. A correlation between -0.3 and -0.7 implies a weak negative association. A correlation between -0.7 and -1.0 implies a strong negative association.;

Compute a correlation matrix; (8)

```
proc corr
    data=perm.fev
    noprint
    outp=correlations;
  var age fev ht;
run;
```

Speaker notes

- Comments on the code: Compute a correlation matrix;

Round the correlations; (9)

```
data correlations;
  set correlations;
  if _type_ NE "CORR" then delete;
  drop _type_;
  age=round(age, 0.01);
  fev=round(fev, 0.01);
  ht=round(ht, 0.01);
run;
```

Speaker notes

- Comments on the code: Round the correlations;

Print the correlations (10)

```
proc print
    data=correlations;
  title2 "All variables show a positive correlations";
run;
```

Speaker notes

- Comments on the code: Print the correlations

With a small number of variables, there is no need to sort the correlations when there are just a few of them.;

Draw scatterplots; (11)

```
proc sgplot
    data=perm.fev;
  scatter x=age y=fev /
      markerattrs=(size=10 symbol=circle);
  pbspline x=abdomen y=fat_brozek /
      lineattrs=(pattern=dash color=red)
      nomarkers;
  title2 "There is a positive association, close to linear";
run;
```

Speaker notes

- Comments on the code: Draw scatterplots;

Break #3

- What you have learned
  - Correlations and scatterplots
- What’s coming next
  - Boxplots and group statistics

05-04

Speaker notes

When you want to look at a relationship between a categorical variable and a continuous variable, you should use a boxplot. Notice that you use proc sgplot for both a scatterplot and a boxplot. This is a big improvement over previous methods in SAS to produce plots because it is easier to learn one procedure and minor variations in that procedure rather than having to learn multiple procedures.

The bottom and top of the boxplot represents the 25th and 75th percentiles, respectively. A thin line, or whisker, is drawn down to the minimum value and up to the maximum value. Extreme values are shown as individual data points. Notice the discrepancy in fev. Smokers seem to have a much higher FEV than non-smokers. This is quite surprising.;

- Notes10. Also look at how the means and standard deviations of your continuous variable change for each level of your categorical variable.

Output, page 8. Notice again the discrepancy in fev by smoking status. This is quite surprising.;

Draw a boxplot; (12)

```
ods graphics / height=1.5 in width=6 in;

proc sgplot
    data=perm.fev;
  hbox fev / category=smoke;
  format smoke fsmoke.;
  title2 "Smokers tend to have higher fev values";
  title3 "This is a surprising and counter-intutive finding";
run;
```

Speaker notes

- Comments on the code: Draw a boxplot;

Compute statistics with a by statement; (13)

```
proc sort
    data=perm.fev;
  by smoke;
run;

proc means
    data=perm.fev;
  var fev;
  by smoke;
  title2 "Descriptive statistics by group";
run;
```

Speaker notes

- Comments on the code: Compute statistics with a by statement;

Break #4

- What you have learned
  - Boxplots and group statistics
- What’s coming next
  - Investiage anomaly

05-05

Speaker notes

This is very odd. You can get a hint as to why smokers might have higher fev values than non-smokers by looking at how height and smoking status are related.

Smokers are taller than non-smokers, and by quite a bit.

These statistics show the same trend. It is obvious that smoking is confined to mostly older children. And since the older children are bigger, that may explain the odd relationship we saw earlier. You should also examine the relationship between sex and fev. Do this on your own, but there is no need to turn anything in. ;

Investigate unusual trend with boxplots; (14)

```
proc sgplot
    data=perm.fev;
  hbox age / category=smoke;
  format smoke fsmoke.;
  title2 "Boxplots";
run;
```

Speaker notes

- Comments on the code: Investigate unusual trend with boxplots;

Investigate unusual trend with descriptive statistics; (15)

```
proc sort
    data=perm.fev;
  by smoke;
run;

proc means
    data=perm.fev;
  var age;
  by smoke;
  format smoke fsmoke.;
  title2 "Descriptive statistics by group";
run;

ods pdf close;
```

Speaker notes

- Comments on the code: Investigate unusual trend with descriptive statistics;

Summary

- What you have learned
  - Labels and formats
  - Descriptive statistics
  - Correlations and scatterplots
  - Boxplots and group statistics
  - Investiage anomaly