Introduction to SAS, Working with a mix of categorical and continuous variables

Steve Simon

Created: 2019-07-19

Overview

- Review
  - proc format, proc freq, proc means
- proc corr
- proc sgplot
  - scatterplot
  - boxplot
- by statement

author: Steve Simon date: Created 2021-05-30 purpose: to produce slides for module05 videos license: public domain

Here is an overview of what I want to cover in module05.

SAS code: Part00. Documentation header

```
* 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
```

Notes00 This is the standard documentation header.

SAS code: Part01. Tell SAS where to find and store things.

```
options papersize=(6in 4in);
* This needed to have the output fit on
PowerPoint;

%let path=q:/introduction-to-sas;

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"
```

Notes01. You should already be familiar with this. The filename statement tells you where the raw data is stored. The libname statement tells you where SAS will store any permanent datsets. The ods statement tells you that SAS is going to store the results with a particular filename and in pdf format.

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

SAS code: Part02. Label your categorical variables

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

Notes02. There are several categorical variables in this data set with number codes, so you should define labels for those codes.

SAS code: Part03. Reading the data using a data step

```
data perm.fev;
  infile raw_data delimiter="," firstobs=2;
  input age fev ht sex smoke;
  label
    age=Age in years
    fev=Forced Expiratory Volume (liters)
    ht=Height in inches
    sex=Sex
    smoke=Smoking status
  ;
run
```

Notes03. The data file is comma delimited and the first row includes variable names.

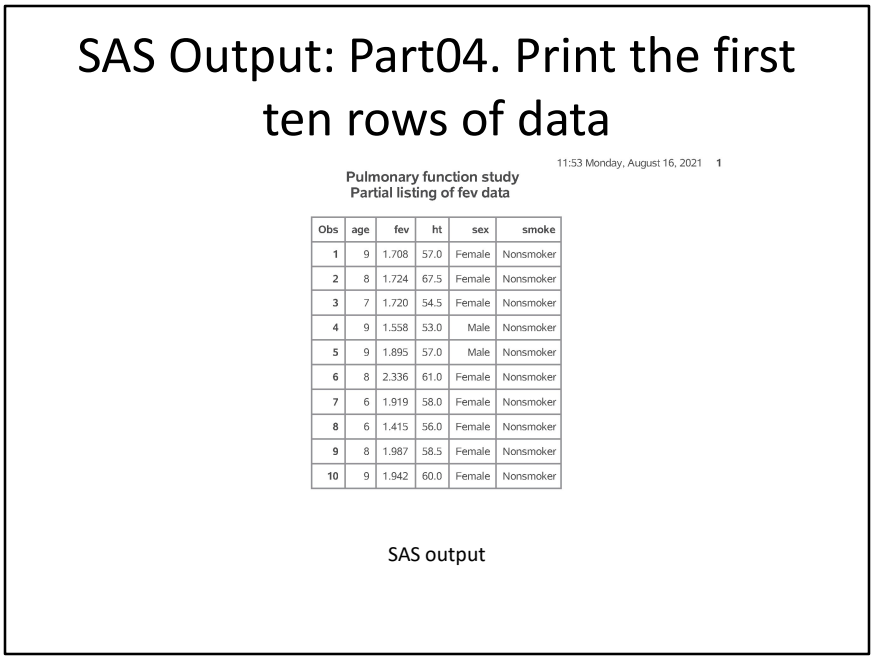
Normally, this means that you can save a bit of time by using proc import, but I chose to read in the data using a data step. The number of variables was so small that this didn't matter that much. It also allowed me to define variable labels in the initial data step rather than later.

SAS code: Part04. Print the first ten rows of data

```
proc print
    data=perm.fev(obs=10);
  format
    sex fsex.
    smoke fsmoke.
  ;
  title1 "Pulmonary function study";
  title2 "Listing of first ten rows of fev data";
run
```

Notes04. It's always a good idea to peek at the first few rows of data.

SAS Output: Part04. Print the first ten rows of data



11:53 Monday, August 16, 2021 1

Pulmonary function study
Partial listing of fev data

Obs	age	fev	ht	sex	smoke
1	9	1.708	57.0	Female	Nonsmoker
2	8	1.724	67.5	Female	Nonsmoker
3	7	1.720	54.5	Female	Nonsmoker
4	9	1.558	53.0	Male	Nonsmoker
5	9	1.895	57.0	Male	Nonsmoker
6	8	2.336	61.0	Female	Nonsmoker
7	6	1.919	58.0	Female	Nonsmoker
8	6	1.415	56.0	Female	Nonsmoker
9	8	1.987	58.5	Female	Nonsmoker
10	9	1.942	60.0	Female	Nonsmoker

SAS output

Output, page 1. There are no obvious problems with this dataset.

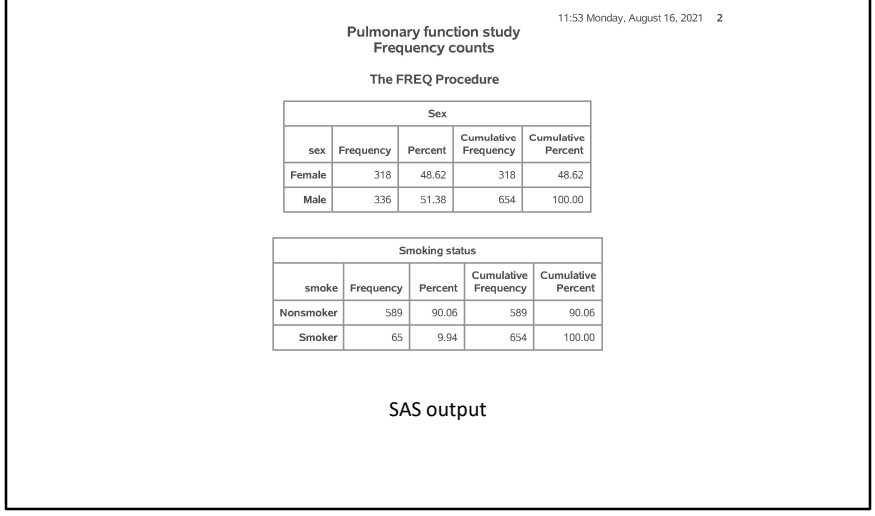
SAS code: Part05. Proc freq and proc means

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

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

Notes05. 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.

SAS Output: Part05. Proc freq and proc means



11:53 Monday, August 16, 2021 2

Pulmonary function study
Frequency counts

The FREQ Procedure

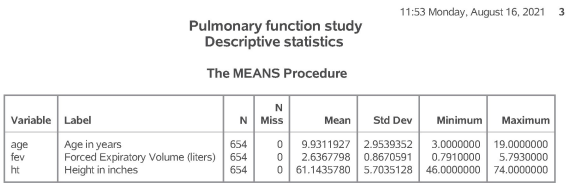
Sex				
sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	318	48.62	318	48.62
Male	336	51.38	654	100.00

Smoking status				
smoke	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Nonsmoker	589	90.06	589	90.06
Smoker	65	9.94	654	100.00

SAS output

Output, page 2. 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.

SAS Output: Part05. Proc freq and proc means



11:53 Monday, August 16, 2021 3

Pulmonary function study
Descriptive statistics

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum
age	Age in years	654	0	9.9311927	2.9539352	3.0000000	19.0000000
fev	Forced Expiratory Volume (liters)	654	0	2.6367798	0.8670591	0.7910000	5.7930000
ht	Height in inches	654	0	61.1435780	5.7035128	46.0000000	74.0000000

SAS output

Output, page 3. 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).

Break #1

- What have you learned
  - Reviewing descriptive statistics
- What's next
  - Correlations and scatterplots

SAS code: Part06. Pearson correlation, proc corr

```
title2 "Correlations";
proc corr
    nosimple noprob
    data=perm.fev;
  var age fev ht;
run
```

Notes06. The Pearson correlation coefficient gives you a numeric measure of the strength of association between two continuous variables.

SAS Output: Part06. Pearson correlation, proc corr

Pulmonary function study
Correlations

The CORR Procedure

3 Variables:	age	fev	ht

Pearson Correlation Coefficients, N = 654

	age	fev	ht
age Age in years	1.00000	0.75646	0.79194
fev Forced Expiratory Volume (liters)	0.75646	1.00000	0.86814
ht Height in inches	0.79194	0.86814	1.00000

SAS output

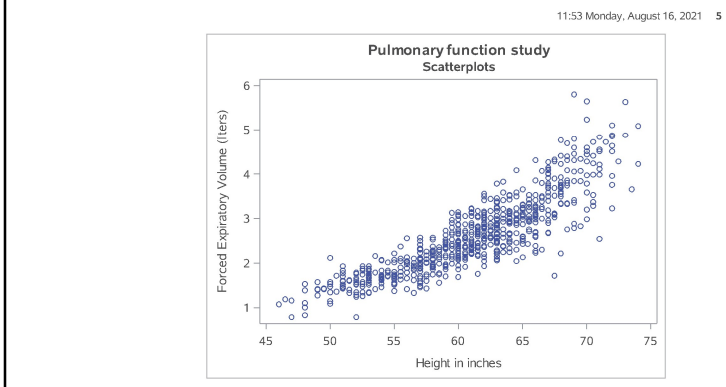
Output, page 4. Remember the cut-offs. A correlation between +0.7 and 1.0 implies a strong positive association. A correlation between +0.3 and +0.7 implies a weak positive association. A correlation between -0.3 and +0.3 implies little or no association. A correlation between -0.3 and -0.7 implies a weak negative association. A correlation between -0.7 and -1.0 implies a strong negative association.

SAS code: Part07. Scatterplot, proc sgplot

```
title2 "Scatterplots";
proc sgplot
    data=perm.fev;
  scatter x=ht y=fev;
run
```

Notes07. You should also examine the association between continuous variables using a scatterplot.

SAS Output: Part07. Scatterplot, proc sgplot



SAS output

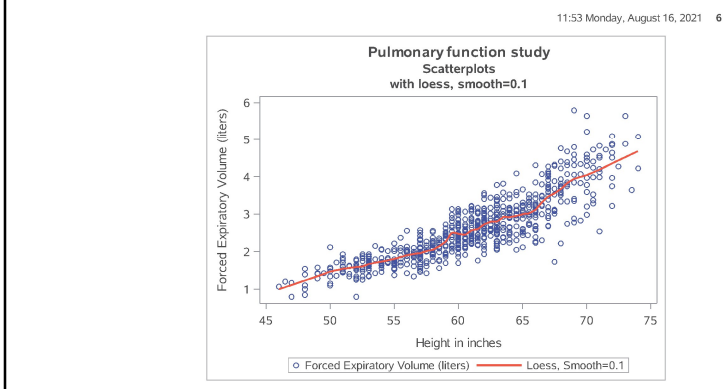
Output, page 5. I am only showing the plot of ht versus fev, but you should also examine the plot of age versus fev.

SAS code: Part08. Scatterplot, smoothing curve

```
title3 "with loess, smooth=0.1";
proc sgplot
    data=perm.fev;
  scatter x=ht y=fev;
  loess x=ht y=fev /
    nomarkers
    smooth=0.1
    lineattrs=(color=Red);
run
```

Notes08. Sometimes a trend line can help. You should consider a smoothing method like loess or pbspline, as this will help you visualize any potential nonlinear relationships.

SAS Output: Part08. Scatterplot, smoothing curve



SAS output

Output, page 6. The relationship looks reasonably close to linear.

Break #2

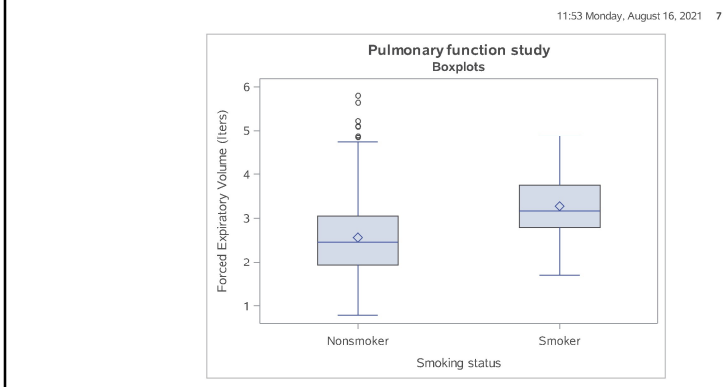
- What have you learned
  - Correlations and scatterplots
- What's next
  - Boxplots

SAS code: Part09. Boxplot, proc sgplot

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

Notes09. 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.

SAS Output: Part09. Boxplot, proc sgplot



SAS output

Output, page 7. 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.

Break #3

- What have you learned
  - Boxplots
- What's next
  - Means by group

SAS code: Part10. Descriptive statistics, by statement

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

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

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

SAS Output: Part10. Descriptive statistics, by statement

Pulmonary function study
Descriptive statistics by group

The MEANS Procedure

Smoking status=Nonsmoker

Analysis Variable : fev Forced Expiratory Volume (liters)				
N	Mean	Std Dev	Minimum	Maximum
589	2.5661426	0.8505215	0.7910000	5.7930000

Smoking status=Smoker

Analysis Variable : fev Forced Expiratory Volume (liters)				
N	Mean	Std Dev	Minimum	Maximum
65	3.2768615	0.7499863	1.6940000	4.8720000

SAS output

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

Break #4

- What have you learned
  - Means by group
- What's next
  - Investigating an odd association

SAS code: Part11. Investigate unusual trend, proc sgplot and means

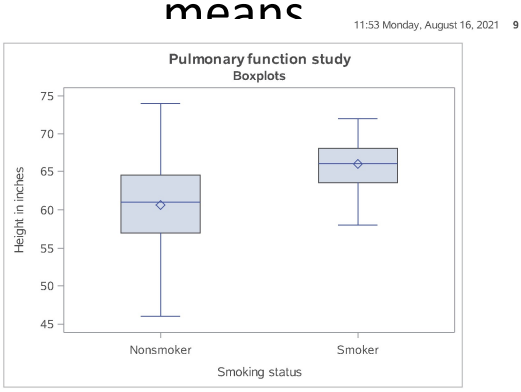
```
proc sgplot
    data=perm.fev;
  vbox ht / category=smoke;
  format smoke fsmoke.;
  title2 "Boxplots";
run;

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

proc means
    data=perm.fev;
  var ht;
```

Notes11. 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.

SAS Output: Part11. Investigate unusual trend, proc sgplot and means



SAS output

Output, page 9. Smokers are taller than non-smokers, and by quite a bit.

SAS Output: Part11. Investigate unusual trend, proc sgplot and means

Pulmonary function study
Descriptive statistics by group

The MEANS Procedure

Smoking status=Nonsmoker

Analysis Variable : ht Height in inches				
N	Mean	Std Dev	Minimum	Maximum
589	60.6127334	5.6724322	46.0000000	74.0000000

Smoking status=Smoker

Analysis Variable : ht Height in inches				
N	Mean	Std Dev	Minimum	Maximum
65	65.9538462	3.1926711	58.0000000	72.0000000

SAS output

Output, page 10. 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.

Summary

- Reviewing descriptive statistcs
- Correlations and scatterplots
- Boxplots
- Means by group
- Investigating an odd association