Lecture03 - A mix of categorical and continuous variables

Steve Simon

7/19/2019

Topics in this lecture

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

Here are the topics we will cover. It is a lot of material, but you will pick it up very easily.

1. Introduction

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

You can find a description of this data set at

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

fev data set

```
"V1","V2","V3","V4","V5"
9,1.708,57,0,0
8,1.724,67.5,0,0
7,1.72,54.5,0,0
9,1.558,53,1,0
9,1.895,57,1,0
8,2.336,61,0,0
6,1.919,58,0,0
6,1.415,56,0,0
8,1.987,58.5,0,0
```

Here are the first ten rows of the fev data set.

2. Data and output locations

```
filename raw_data
  "../data/fev.txt";

libname intro
  "../data";

ods pdf
  file="lecture03.pdf";
```

Here are the standard commands to tell SAS where to find the data, where to place its own data files and where to store the output.

3. Labels for categorical data

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

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

4. Reading the data using a data step

```
data intro.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;
```

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.

5. Print the first ten rows of data

```
title1 "Pulmonary function study";
title2 "Partial listing of fev data";
proc print
    data=intro.fev(obs=10);
  format
    sex fsex.
    smoke fsmoke.
  ;
run;
```

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

5. Print the first ten rows of data

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

Output, proc print

Here are the first ten rows of the data .

6. Proc freq and proc means

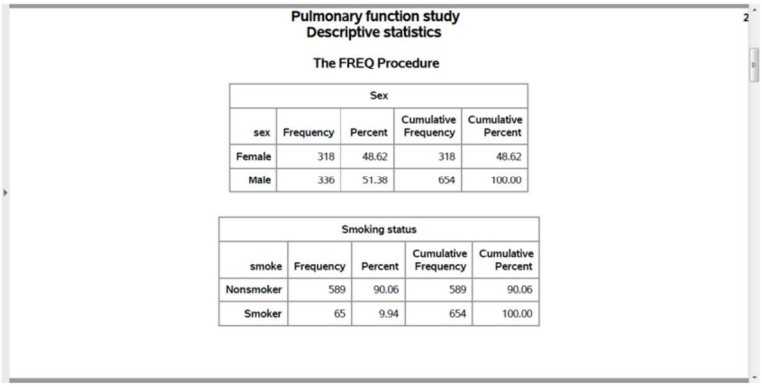
```
title2 "Frequency counts";
proc freq
    data=intro.fev;
  tables sex smoke / missing;
  format
    sex fsex.
    smoke fsmoke.
run;

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

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.

6. Proc freq and proc means



Pulmonary function study
Descriptive statistics

2

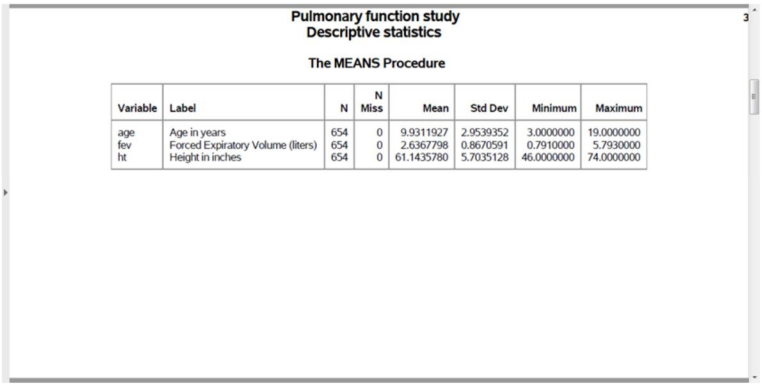
The FREQ Procedure

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

Output, proc freq

6. Proc freq and proc means



Pulmonary function study
Descriptive statistics

3

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

Output, proc means

7. Pearson correlation, proc corr

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

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

7. Pearson correlation, proc corr

Pulmonary function study
Correlations

4

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

Output, proc corr

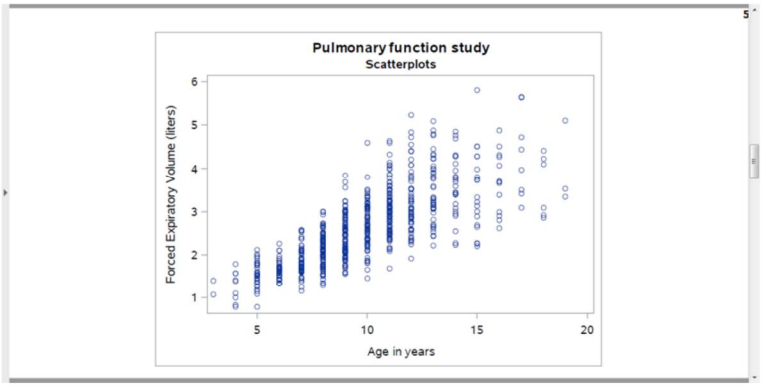
8. Scatterplot, proc sgplot

```
title2 "Scatterplots";
proc sgplot
    data=intro.fev;
  scatter x=age y=fev;
run;
```

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

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

8. Scatterplot, proc sgplot



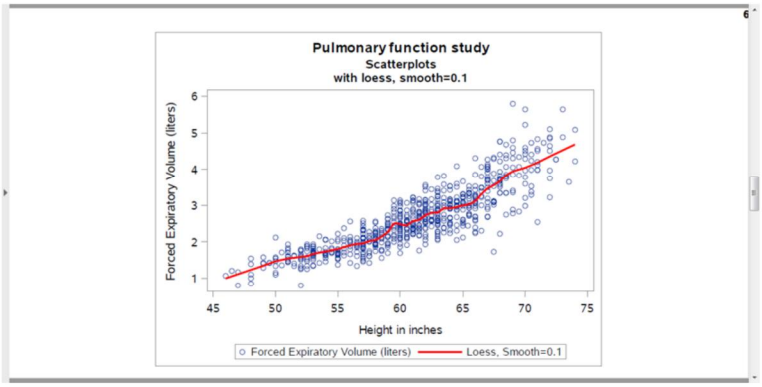
Output, proc sgplot

9. Scatterplot, smoothing curve

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

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.

9. Scatterplot, smoothing curve



Output, proc sgplot

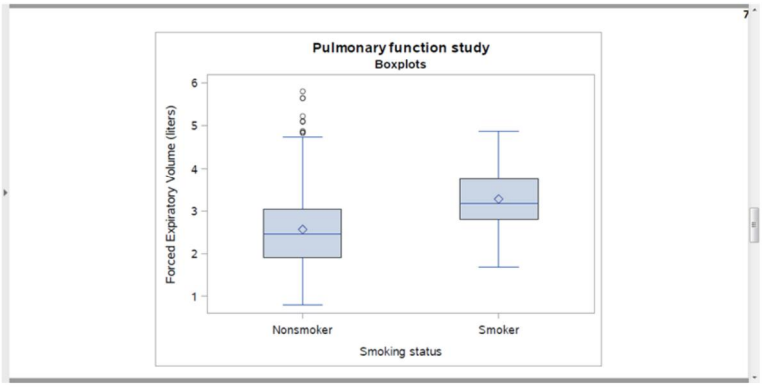
10. Boxplot, proc sgplot

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

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.

10. Boxplot, proc sgplot



Output, proc sgplot

11. Descriptive statistics, by statement

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

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

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

11. 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

Output, proc means with by statement

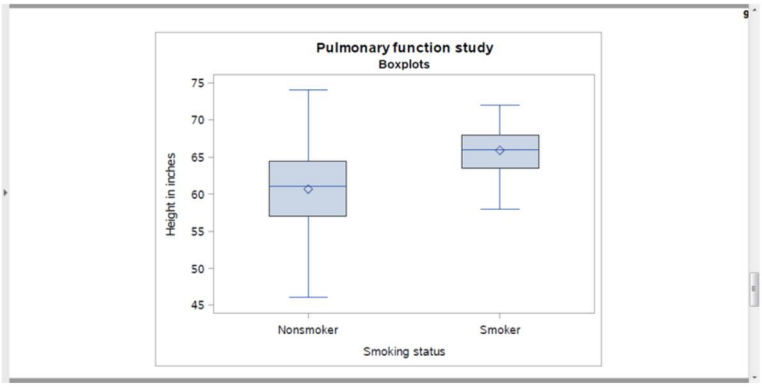
12. Investigate unusual trend, proc sgplot and means

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

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

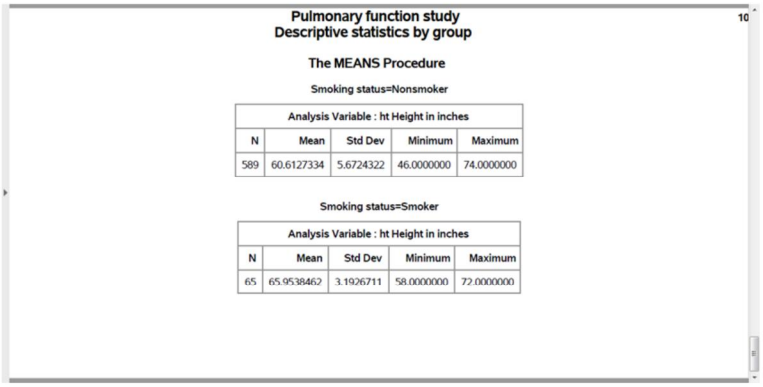
This is very odd. You can get a hint as to why smokers might have higher fev values than non-smokers by looking at how age and smoking status are related.

12. Investigate unusual trend, proc sgplot and means



Output, proc sgplot

12. Investigate unusual trend, proc sgplot and means



Pulmonary function study
Descriptive statistics by group

10

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

Output, proc means with by statement

13. Further investigation on your own

You should also examine the relationship between sex and fev. Do this on your own, but there is no need to turn anything in.

Homework assignment (1/5)

Your homework assignment will use a data set of housing prices and factors that influence the price.

Details for this data set can be found on the DASL web site.

Homework assignment (2/5)

- Read in the data housing.txt.
- Convert the asterisks in the AGE and TAX variables to missing. How many missing values are there for AGE and for TAX?
- Create factors for NE, CUST, and COR. Draw bar charts for each of these factors.

Homework assignment (3/5)

- Find the largest house (biggest SQFT) in the data set. Is the largest house also the most expensive house?
- Calculate frequency counts for FEATS. Are there any houses with the no features? Are there any houses with every possible feature?

Homework assignment (4/5)

- Evaluate the relationship between PRICE and SQFT using a scatterplot. Include a smooth curve. Do larger houses tend to cost more?
- Evaluate the relationship between CUST and PRICE using a boxplot. Calculate the mean price and standard deviation by CUST.

Homework assignment (5/5)

- What is the difference in average prices between northeast houses and other houses? What is the difference in average taxes?
- Are custombuilt houses more likely to appear on corner lots? Calculate the percentages and compute a relative risk.