171:161: Introduction to Biostatistics
Breheny

Lab #3

The focus of this lab will be on using SAS and R to provide you with summary statistics of different variables with a data set. We will look at both categorical and continuous variables.

Viewing Data in SAS

SAS has two types commands that it will run, PROC steps and DATA steps. Today will will look at couple of PROC steps that can be used to provide information or perform certain calculations on a data set. From now on our focus in SAS will be using and interpretting the results of the different PROC steps that are built in to SAS.

To do this lab we will need to have our data sets ready in SAS. Today will use a categorical data set 'titanic' and a continuous data set 'tips'. Note that in the code below you will have to change the file path to match where you have the data set saved.

```
proc import datafile = "H:\BIOS-161-TA\titanic.txt"
    out = titanic
    dbms = tab
    replace;
    getnames = yes;
run;
```

The two simpliest procedures in SAS are PROC PRINT and PROC CONTENTS. PROC PRINT will simply print a table of what the data set looks like into the output window. PROC CONTENTS will provide a more detailed report of what is in the data set, including variable names and information about the variables.

```
PROC PRINT data = titanic;
RUN;

PROC CONTENTS data = titanic;
RUN;
```

Note that SAS is not case sensitive. However it is useful to get in the habit of capitalizing keywords to enhance the readability of your code.

Tables in SAS

PROC FREQ is a procedure that can be used to obtain tables for categorical data. Within PROC FREQ you can create 1-way, 2-way, 3-way, etc. tables using a TABLES statement. The code below demonstrates the syntax for creating a 1-way, 2-way and 3-way table (each are created with their own TABLES statement within the procedure).

```
PROC FREQ data = titanic;
   TABLES class;
   TABLES sex*survived;
   TABLES age*sex*survived;
RUN;
```

A lot information is presented in these tables. Be sure that you are comfortable reading all of the output that SAS provides you. For each cell in the table make sure that you know what is being counted and what is meant by 'Percent', 'Row Percent' and 'Col Percent'.

How would you interpret the two tables provided when by the statement “TABLES age*sex*survived”

Graphs in SAS

The same information contained in the tables created by PROC FREQ can also be displayed using graphs. This can be done using the GPLOT procedure and with the statement VBAR to create a vertical bar chart. If you would like to save this bar chart and use it in something like a word document you can right click on it and select save as to save a copy of it.

```
PROC SGPLOT DATA=titanic;
     VBAR class;
RUN;
```

This graph however doesn't offer any explanation as to how different variables are linked to survival. IE if survival differs based upon factors like age, gender, etc. To look into these relationships we have to options: The first is conditioning, which is subsetting the bar chart by another variable. The second is grouping, which is grouping which includes more than one relationship in a single plot using colors to differentiate.

The code below is an example grouping using “survived” as the grouping variable with respect to the variable “class”

```
PROC SGPLOT DATA=titanic;
     VBAR class / GROUP=survived;
RUN;
```

How does this representation compare to using “class” as the grouping variable with respect to “survived”? Which do you find to be more useful?

Now we will look at conditioning. The code below will use “sex” to condition our prior graph of “class” grouped by “survived”. Note that we now need to use the procedure PROC SGPANEL instead of PROC SGPLOT.

```
PROC SGPANEL DATA=titanic;
     PANELBY sex;
     VBAR class / GROUP=survived;
RUN;
```

SAS can also condition on multiple variables at the same time:

```
PROC SGPANEL DATA=titanic;
PANELBY age sex;
      VBAR class / GROUP=survived;
RUN;
```

Sometimes its useful to have the same scale across all of your graphs, in other cases you might want to scale to change to make it easier to identify what is happening. The option UNISCALE as shown below will allow you to manipulate this.

```
PROC SGPANEL DATA=titanic;
     PANELBY age sex / UNISCALE=COLUMN;
     VBAR class / GROUP=survived;
RUN;
```

Now lets compare the results of our bar chart to the table we obtained using PROC FREQ:

```
PROC FREQ data = titanic;
    TABLES age*sex*survived;
RUN;
```

Are you able to identify the same relationships using both types of output (graphs vs tables)? Which is easier?

Tables in R

We can create many of the same tables and graphs in R as we just saw in SAS. First we must read in the titanic dataset:

```
titanic <- read.delim("http://myweb.uiowa.edu/pbreheny/161/data/titanic.txt")
attach(titanic)
```

In R the table function will create a table for a specified variable.

```
Table1 <- table(Class)
Table1
```

To see proportions or percentages in R we need to change our code a little bit:

```
Props1 <- prop.table(Class)
Props1
Perc1 <- 100 * prop.table(Class)
Perc1
```

Two and three way tables can be created using multiple variables in the table function:

```
Table2 <- table(Sex,Survived)
Table2
Table3 <- table(Sex,Survived,Age)
Table3
```

You can still use prop.table to see proportions for these two/three way tables. You can also use prop.table to see percentages across rows or across columns.

```
> prop.table(Table2)    ## Overall using Table2 created previously
> prop.table(Table2,1)  ## Across row in Table2
> prop.table(Table2,2)  ## Across col in Table2
```

For three way tables things get a little trickier. The code below will calculate proportions for each combination of Sex and Age using our Table3 from before. (Note 1 and 3 are the positions of the variables Sex and Age in the construction of Table3).

```
prop.table(Table3,c(1,3))
```

Graphs in R

To replicate many of the graphs we saw in SAS will need to load a package into R. The code to load the "Lattice" package is shown below:

```
require(lattice)
```

You should see a message in red saying the package was loaded.

Now we can create barcharts using the function barchart. Several annotated examples are shown below:

```
barchart(Table1)
barchart(table(Survived))
## A simple barchart showing the counts of survivors and non-survivors,
## just like what we saw first in SAS

barchart(table(Survived),horizontal=FALSE)
## Changing horizontal from its default to =FALSE will create a vertical barchart

barchart(table(Class,Survived))
## Using "Survived" as a grouping variable for "Class"

barchart(table(Class,Survived),auto.key=TRUE)
## Includes a key indentifying the groups by color

barchart(table(Class,Sex,Survived),auto.key=TRUE)
barchart(table(Class,Sex,Age,Survived),auto.key=TRUE)

barchart(table(Class,Sex,Age,Survived),auto.key=TRUE,scales="free")
#Allows the scale to change across graphs, similar to UNISCALE in SAS
```

Weighted Average Example

We are still using the Titanic data set. We've included helpful information below.

Class	Survived Count	N	Rate
1st	203	325	62.46154
2nd	118	285	41.40351
3rd	178	706	25.21246
Crew	212	885	23.95480

Class	Sex	Survived Count	N	Rate
1st	F	141	145	97.24138
	M	62	180	34.44444
2nd	F	93	106	87.73585
	M	25	179	13.96648
3rd	F	90	196	45.91837
	M	88	510	17.25490
Crew	F	20	23	86.95652
	M	192	832	22.27378

Based on the information above, what percentage of the passengers were female? What percentage were male?

$$PROP_{female} = \frac{470}{2201}$$
$$PROP_{male} = \frac{1731}{2201}$$

Note: These are our weights.

Using these rates as well as the tables above, calculate the weighted average of the percentage of 1^{st}, 2^{nd}, 3^{rd}, and Crew members who survived, controlling for the effect of sex.

First Class Survival Rate Adjusting for Sex

$$\begin{aligned}
RATE_{1^{st}} &= \text{Weighted Average for Females} + \text{Weighted Average for Males}\\
&= \left(\frac{\text{\# Female Passnegers}}{\text{Total \# Passengers}}\right)\left(\frac{\text{\# Female Survived in } 1^{st} \text{ Class}}{\text{\# Female Passengers in } 1^{st} \text{ Class}}\right) + \left(\frac{\text{\# Male Passnegers}}{\text{Total \# Passengers}}\right)\left(\frac{\text{\# Male Survived in } 1^{st} \text{ Class}}{\text{\# Male Passengers in } 1^{st} \text{ Class}}\right)\\
&= \left(\frac{470}{2201}\right)\left(\frac{141}{145}\right) + \left(\frac{1731}{2201}\right)\left(\frac{62}{180}\right)\\
&= \boxed{0.47854}
\end{aligned}$$

Second Class Survival Rate Adjusting for Sex

$$\begin{aligned}
RATE_{2^{nd}} &= \text{Weighted Average for Females} + \text{Weighted Average for Males}\\
&= \left(\frac{\text{\# Female Passnegers}}{\text{Total \# Passengers}}\right)\left(\frac{\text{\# Female Survived in } 2^{nd} \text{ Class}}{\text{\# Female Passengers in } 2^{nd} \text{ Class}}\right) + \left(\frac{\text{\# Male Passnegers}}{\text{Total \# Passengers}}\right)\left(\frac{\text{\# Male Survived in } 2^{nd} \text{ Class}}{\text{\# Male Passengers in } 2^{nd} \text{ Class}}\right)\\
&= \left(\frac{470}{2201}\right)\left(\frac{93}{106}\right) + \left(\frac{1731}{2201}\right)\left(\frac{25}{179}\right)\\
&= \boxed{0.29719}
\end{aligned}$$

Weighted Average Class Survival Rates Controlling for Sex

Class	Rate
1st	47.85
2nd	29.72
3rd	23.38
Crew	36.09