SAS Libraries, PROC FREQ, PROC UNIVARIATE

Shannon Pileggi

STAT 330

OUTLINE

Libraries: temporary vs permanent

SAS stores data sets in a *SAS Library*. Libraries can be:

Temporary

- Stored in the `WORK` folder
- SAS data sets are deleted when the SAS session closes

Permanent

- Some come with SAS, like `SASHELP`
- We can also create our own permanent libraries
- Allows us to create permanent SAS data sets that remain on your computer even after the SAS session closes

SAS data set names

- SAS data sets have two-level names: `SASHELP.BASEBALL`
  1. the library reference (`SASHELP`)
  2. the data set name (`BASEBALL`)
- The 2 levels are separated by a period
- Capitalization does not matter - these two level names work equivalently: `SASHELP.baseball`, `sashelp.BASEBALL`, `Sashelp.Baseball`
- This naming convention is used in both `DATA` steps and `PROCS`
- More generally, the naming convention is `LibRef.DataSetName`

SAS data set names, another example

SAS Code

```
PROC IMPORT OUT = WORK.babies
   DATAFILE = "X:/spileggi/Data Sets/babies.csv"
   DBMS = CSV REPLACE;
RUN;

PROC IMPORT OUT = babies
   DATAFILE = "X:/spileggi/Data Sets/babies.csv"
   DBMS = CSV REPLACE;
RUN;
```

SAS Code

These two code chunks are **equivalent**. If the library reference is missing/blank, then it defaults to `WORK`. For both,

- the library reference is `WORK`
- the data set name is `babies`

On your own:

For each of the following data set names, indicate if we are referring to a (1) temporary SAS data set or (2) a permanent SAS data set.

1. `baseball`
2. `mylib.baseball`
3. `work.baseball`
4. `x.baseball`
5. `temp.baseball`

Library reference

```
libname LibRef "Computer Address/Location";
```

- We can use our own *library reference* to access and save permanent SAS data. The `LibRef`
  - is limited to 8 characters
  - must begin with a character
  - can only contain characters/numbers/underscores
- You can think of this as a shortcut to a location on your computer
- You can see your SAS libraries in the *Explorer window of SAS*
- If you are navigating in your *computer's* explorer, you will **not** see the library reference name - just the data set name and extension (`.sas7bdat`)

Try it!

On your own:

1. Copy the data set `adni.sas7bdat` to your flash drive or desktop
2. Create a library reference called `flash` for the location of the data set on your flash drive

```
LIBNAME flash "Computer Address/Location";
```

Remember: You can explore to the data set in your computer and right click on it to identify the location.

3. View the contents of the data set

```
PROC CONTENTS DATA=flash.adni; RUN;
```


Your first DATA step

In a SAS `DATA` step we can *create* or *manipulate* data.

SAS Code

```
DATA work.adni_temp ;
RUN;

PROC CONTENTS DATA = work.adni_temp ;
RUN;

PROC PRINT DATA = work.adni_temp ;
RUN;
```

SAS Code

Here, we are creating a *brand new*, temporary data set called `adni_temp` in the work library.

On your own: How many observations and variables are in the `work.adni_temp` data set?

Using the SET statement, demo 1

The `SET` statement allows you to make a copy of an existing data set, as well as perform calculations/manipulations on the data.

SAS Code

```
DATA work.adni_temp ;
    SET flash.adni;
RUN;

PROC CONTENTS DATA = work.adni_temp ;
RUN;
```

SAS Code

Here, we are creating a *brand new*, temporary data set called `adni_temp` in the work library. This data set contains a copy of the permanent `adni` data set located in the `flash` library.

On your own: How many observations and variables are in the `work.adni_temp` data set?

Using the SET statement, demo 2

SAS Code

```
DATA flash.adni2 ;
    SET flash.adni;
RUN;

PROC CONTENTS DATA = flash.adni2 ;
RUN;
```

SAS Code

Here, we are creating a *brand new*, permanent SAS data set called `adni2` in the flash library. This data set contains a copy of the permanent `adni` data set located in the `flash` library.

On your own: How many observations and variables are in the `flash.adni2` data set? Examine your desktop / flash drive to verify that this data set was created.

Discussion

Suppose I want to create a permanent data set named `cookie` that is stored in the SAS library `monster`. Which `libname` statement is correct?

1. `libname cookie.monster "Computer Location";`
2. `libname cookie "Computer Location";`
3. `libname monster "Computer Location";`
4. `libname cookie monster "Computer Location";`
5. `libname monster.cookie "Computer Location";`

PROC FREQ

SAS Code

```
proc freq data=datasetname;
    table var1 var1*var2 / options;
run;
```

SAS Code

- Obtains counts of *numeric* and *character* variable values.
- For two way tables, `var1` goes on rows and `var2` goes on columns
- `table` options:
  - `list` - modifies output to list format
  - `missing` - includes number of missing in counts
  - `nopercent` - suppresses overall percentages
  - `nocol` - suppresses column percentages
  - `norow` - suppresses row percentages
  - `out=` - save frequencies/percents to a data set

Example code

SAS Code

```
*all variables in data set;
PROC FREQ DATA = flash.adni; RUN;

*one-way and two-way contingency table;
PROC FREQ DATA = flash.adni;
    TABLES dx dx*gender ;
RUN;

*two-way contingency table converted to list style;
PROC FREQ DATA = flash.adni;
    TABLES dx*gender / LIST MISSING NOPERCENT;
RUN;
```

SAS Code

On your own:

SAS Code

```
PROC FREQ DATA = flash.adni;
    TABLES dx*gender / options;
RUN;
```

SAS Code

Which options would you use to obtain the percent of males that have a normal diagnosis?

1. `list missing`
2. `missing nopercent`
3. `norow nocol`
4. `nocol nopercent`
5. `norow nopercent`

Chi-square test

SAS Code

```
PROC FREQ DATA = flash.adni;
    TABLES dx*gender / CHISQ;
RUN;
```

SAS Code

Statistics for Table of dx by GENDER

Statistic	DF	Value	Prob
Chi-Square	2	1.2443	0.5368
Likelihood Ratio Chi-Square	2	1.2434	0.5370
Mantel-Haenszel Chi-Square	1	0.4941	0.4821
Phi Coefficient		0.0671	
Contingency Coefficient		0.0670	
Cramer's V		0.0671	

H_0: there is no association between gender and diagnosis
H_a: there is an association between gender and diagnosis

On your own: What is the conclusion?

PROC UNIVARIATE

- `PROC UNIVARIATE` also produces descriptive and some inferential statistics
- much more detailed output than `PROC MEANS`
- default is to produce results for all numeric variables
- can also produce graphs

SAS Code

```
PROC UNIVARIATE DATA = flash.adni ;
RUN;
```

SAS Code

One sample t-test

SAS Code

```
PROC UNIVARIATE
    DATA = flash.adni
    CIBASIC LOCATION = 70 ;
    VAR age ;
RUN;
```

SAS Code

- `CIBASIC` computes a confidence interval for the population mean age
- `LOCATION` specifies the null hypothesis value

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence	Limits
Mean	73.58261	72.75374	74.41148
Std Deviation	6.99484	6.45591	7.63270
Variance	48.92777	41.67880	58.25810

Tests for Location: Mu0=70				
Test	Statistic		p Value	
Student's t	t	8.508955	Pr > |t|	<.0001
Sign	M	61	Pr >= |M|	<.0001
Signed Rank	S	10146.5	Pr >= |S|	<.0001

H_0: $\mu = 70$ vs H_a: $\mu \neq 70$

On your own: What is the interpretation of the result?

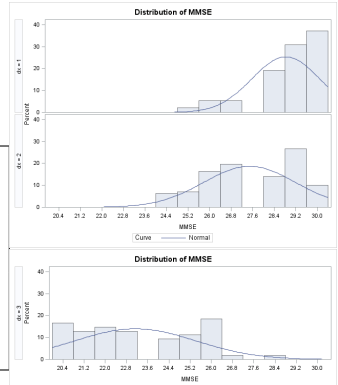
PROC UNIVARIATE

- `PROC UNIVARIATE` has many more options
- Like `PROC UNIVARIATE`, can use `CLASS` or `BY` statements
- Can also produce graphs

SAS Code

```
PROC UNIVARIATE DATA = flash.adni ;
   CLASS dx;
   VAR MMSE;
   HISTOGRAM MMSE / NORMAL;
RUN;
```

SAS Code



On your own:

`APOE4`	Type of APOE4 variant (genetics) 0 - No copies of the ApoE4 allele 1 - One copy of the ApoE4 allele 2 - Two copies of the ApoE4 allele

Which `PROC` would you use to *summarize* `APOE4`?

1. `PROC PRINT`
2. `PROC CONTENTS`
3. `PROC MEANS`
4. `PROC UNIVARIATE`
5. `PROC FREQ`

On your own:

`ADAS`	Alzheimer's Disease Assessment Scale (larger scores indicate greater dysfuction)

Which `PROC` would you use to see if there are any unusual/outlier/erroneous values of `ADAS`?

1. `PROC PRINT`
2. `PROC CONTENTS`
3. `PROC MEANS`
4. `PROC UNIVARIATE`
5. `PROC FREQ`

The Data

- Alzheimer's Disease (AD) is a serious mental illness that affects an estimated 5.3 million Americans; it is the most common cause of dementia among the elderly.
- Characterized by a progressive cognitive decline, AD has been notoriously difficult to diagnose due to symptom-overlap with other mental disorders; until recently, AD could only be confirmed posthumously.
- The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal study that began in 2005, and is designed to track AD biomarkers, identify at-risk patients, and evaluate the efficacy of novel treatments.
- The study consists of healthy individuals (the control group) as well as adults with early Alzheimer's Disease (AD). http://adni.loni.usc.edu/about/.

The Variables

`DX`	Alzheimer's disease diagnosis 1 - Normal cognitive function 2 - Mild cognitive impairment 3 - Alzheimer's disease
`AGE`	Age (years)
`APOE4`	Type of APOE4 variant (genetics) 0 - No copies of the ApoE4 allele 1 - One copy of the ApoE4 allele 2 - Two copies of the ApoE4 allele
`GENDER`	Patient gender
`MMSE`	Mini Mental State Exam (score out of 30, lower scores indicate more cognitive impairment)
`ADAS`	Alzheimer's Disease Assessment Scale (larger scores indicate greater dysfuction)
`WholeBrain`	Brain volume (mm^3)

More Details

- The mini mental state exam (MMSE) is a 30 question assessment commonly used to assess cognitive impairment.
- The Alzheimer's Disease Assessment Scale (ADAS) is a more comprehensive measure of cognitive impairment.
- The apolipoprotein E (APOE) gene, on chromosome 19, has variants associated with high risk of AD.