

3 Data checking, SAS Manual, basic tests

1. Phases of data checking: NHANES III example
2. Proc Insight: plots and histograms
3. Proc Univariate: checking data within SAS
4. Data step editing, arithmetic and comparisons
5. Proc TTest: two-sample t -test
6. Reading the SAS Manual
7. Making an indicator variable
8. Proc Freq: chi-square test

1

Phases of Data Checking

1. SAS recognizes several types for data (character, numeric, date) and each of these has various different formats. The first phase of checking asks:

Did SAS read the data as the correct type?

Did SAS read the correct number of observations and variables?

Proc Contents answers both these questions.

Usually the SAS Import wizard will handle things correctly. On the rare occasions when it doesn't, you can convert your data to CSV format and identify the type of each variable for SAS.

2

2. After you're satisfied that SAS has gotten the variable types right, the next questions are:

Are there mistakes or problems—outliers or incorrect values?

Try scatterplots, Proc Univariate to find extremes, Proc Freq for list of distinct values

What is the pattern of missing data?

Proc Means will count missing values for a variable

Should some variables be transformed?

For positive variables, when $\frac{\text{maximum value}}{\text{minimum value}} > 10$, take logs.

3

NHANES III.

The third National Health and Nutrition Examination Survey collected data from 33,994 people during 1988–1994, a representative sample of the whole US population.

We will consider the subset of survey participants aged 20 to 29, which is the data for HW 1.

One of the survey questions was: How many years of education have you had?

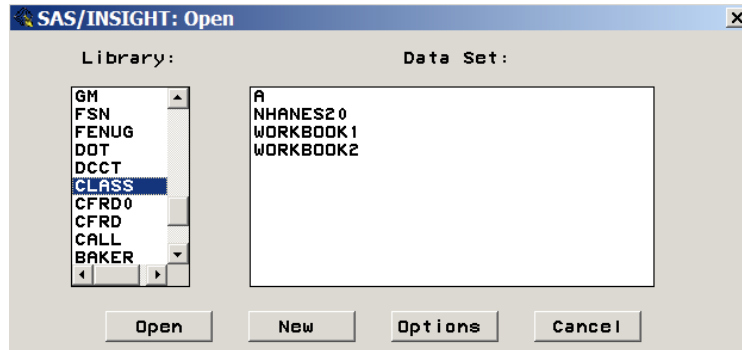
What do you expect a histogram of this data to look like?

4

Quick scatterplots and histograms: Proc Insight

From the menu: Solutions > Analysis > Interactive Data Analysis to start Insight.

Insight asks you to select a data set from a list of libraries and their SAS datasets.

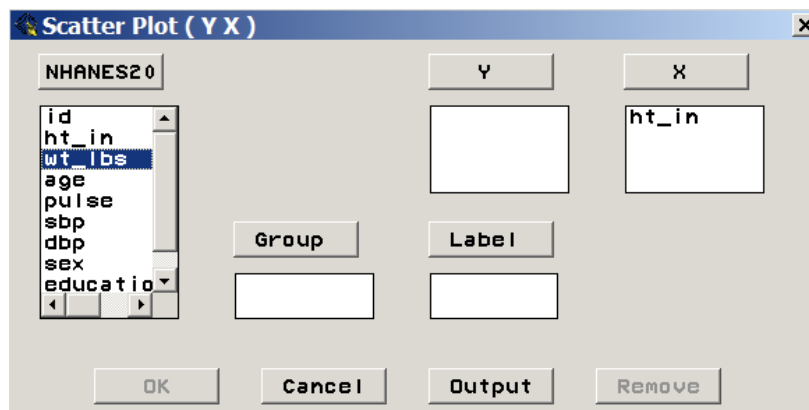


Select library CLASS, then dataset NHANES20.

(You must have previously imported NHANES20 .xls into this library.)

5

To make a scatterplot, select the menu: Analyze > Scatter Plot (Y X) which brings up a dialog box for selecting the variables to plot:

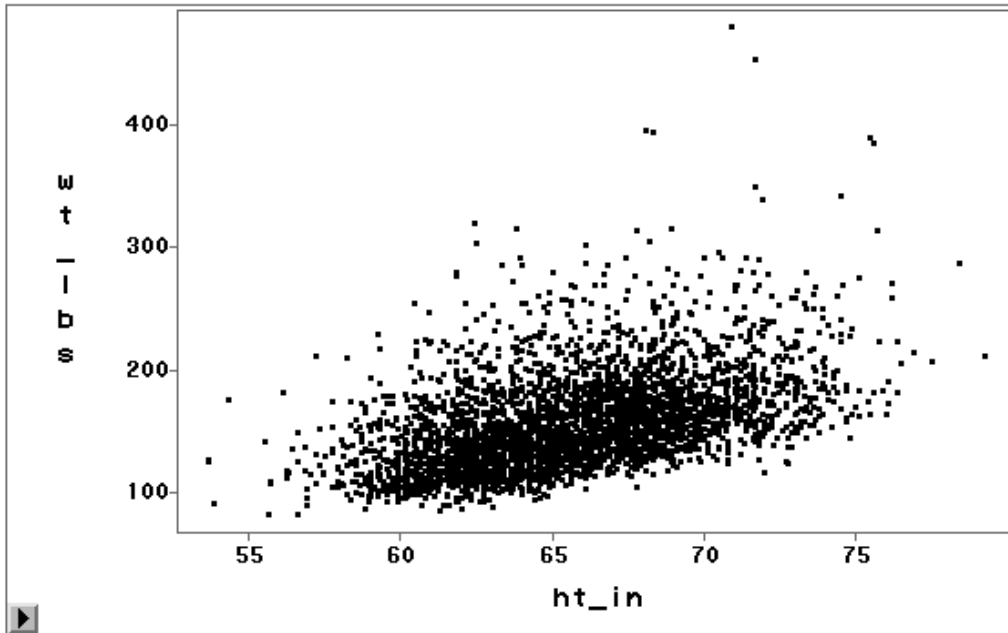


Select wt_lbs, then click Y. Select ht_in, then click X, then click OK.

Selecting sex as a group gives two plots: one for each gender.

Selecting more than one X or Y gives a matrix of scatter plots.

6



Click on a point to get observation number.

Button on lower left offers different size plotting characters.

Cannot get a plot with different symbols for males and females.

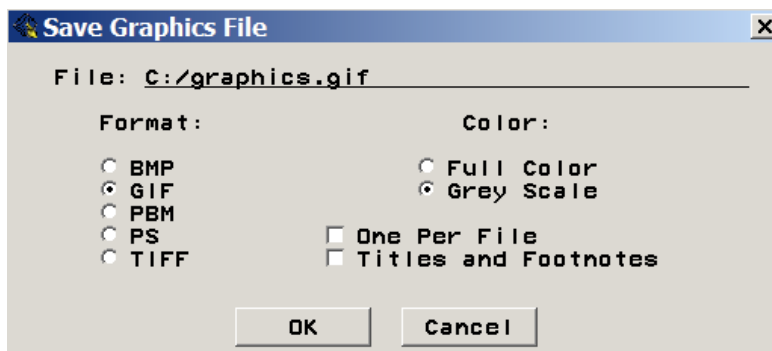
7

1. To save an Insight graphic on a PC, copy and paste it into MSWord.

2. On a Mac, use Grab to capture the picture (in TIFF format).

Use Preview to convert the file to GIF format, to insert into MSWord.

3. On a PC or a Mac, choose Edit > Save > Graphics File ... which brings up this dialog box:

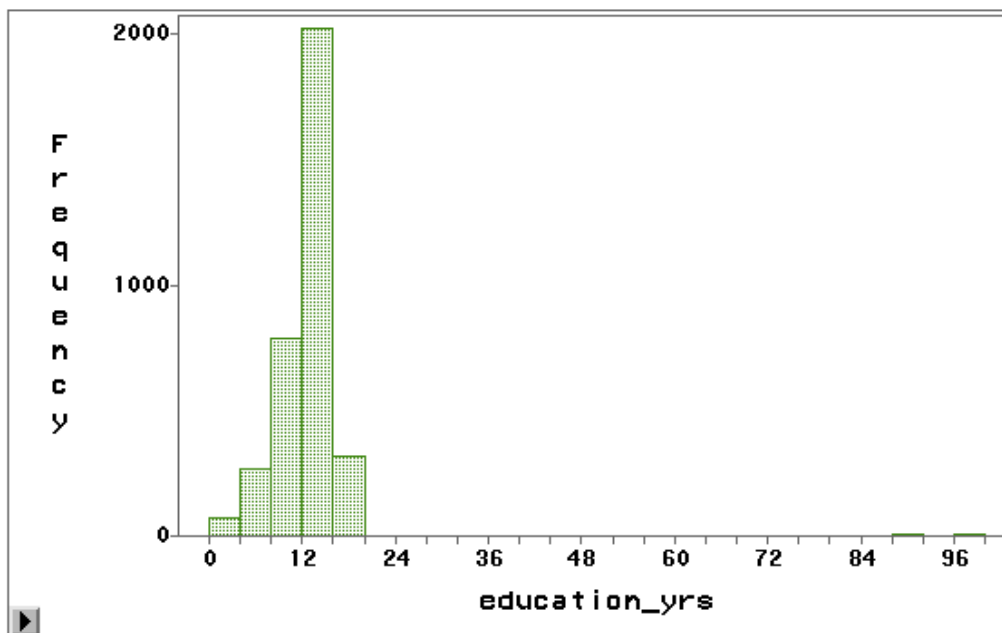


Type in a path to save to a particular location; default is graphics within the SAS folder in Program Files. Save in GIF format.

8

To make a histogram, choose Analyze > Histogram/Bar Chart (Y) which only makes frequency histograms.

Select education_yrs, click Y, then click OK.



9

Proc Univariate

Univariate produces univariate statistics and lists extreme values. The `plot` option gives stem-leaf diagrams, histograms, Q-Q-plots.

```
Proc Univariate plot data=class.nhanes20;
var education_yrs;
```

The UNIVARIATE Procedure
Variable: education_yrs

Moments

N	3507	Sum Weights	3507
Mean	12.1103507	Sum Observations	42471
Std Deviation	7.31065017	Variance	53.4456059
Skewness	9.14954351	Kurtosis	102.081185
Uncorrected SS	701719	Corrected SS	187380.294
Coeff Variation	60.3669566	Std Error Mean	0.12344915

Basic Statistical Measures

Location		Variability	
Mean	12.11035	Std Deviation	7.31065
Median	12.00000	Variance	53.44561
Mode	12.00000	Range	99.00000
		Interquartile Range	3.00000

Quantile	Estimate
100% Max	99
99%	17
95%	16
90%	15
75% Q3	13
50% Median	12
25% Q1	10
10%	8
5%	6
1%	1
0% Min	0

11

List of 5 smallest and 5 largest observations can identify outliers or errors

The UNIVARIATE Procedure
Variable: education_yrs

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0	3382	99	2309
0	3185	99	2526
0	3141	99	2534
0	2885	99	2771
0	2829	99	3058

What does 99 years of education mean?

12

What are these extremely high values of education_yrs?

Use Proc Freq to list all distinct values.

```
Proc FREQ data=class.nhanes20;  
  tables education_yrs;
```

education_yrs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	29	0.83	29	0.83
1	8	0.23	37	1.06
2	17	0.48	54	1.54
.
11	246	7.01	1139	32.48
12	1268	36.16	2407	68.63
13	297	8.47	2704	77.10
14	290	8.27	2994	85.37
15	169	4.82	3163	90.19
16	219	6.24	3382	96.44
17	102	2.91	3484	99.34
88	11	0.31	3495	99.66
99	12	0.34	3507	100.00

23 observations have nonsense values for years of education.

15

The variable education_yrs is really HFA8R in the NHANES III data.

From the documentation for NHANES III adult data:

```
HFA8R    What is the highest grade or year of  
         regular school -- has completed?
```

00 Never attended or kindergarten only

01-17

88 Blank but applicable

99 Don't know

88 and 99 are codes for missing data.

```
So exclude values larger than 17 years.
```

16

Editing data in a DATA step

We want to exclude values of `education_yrs` larger than 17 years.

Don't edit the Excel spreadsheet. Make the change in the SAS code.

We work with a new dataset that begins as a copy of the original data. We could delete these observations in a DATA step, using `IF-THEN` (LSB§3.4)

Data one;

```
SET class.nhanes20;      SET means "copy as"
IF (education_yrs > 17) THEN delete;  when IF condition is true, delete the obs
```

Better approach—replace 88 and 99 with “missing;” a period indicates a missing number:

Data one;

```
SET class.nhanes20;
IF (education_yrs > 17) THEN education_yrs = . ;
```

17

Still better—keep the original variable and make a new corrected variable:

Data one;

```
SET class.nhanes20;
education_yrs_corrected = education_yrs;
IF (education_yrs_corrected > 17) THEN education_yrs_corrected = .
```

This doesn't work:

Data one;

```
SET class.nhanes20;
IF (education_yrs > 17) THEN education_yrs_corrected = . ;
```

`education_yrs_corrected` will be created but set to missing for all observations. Make the variable first, then edit it.

Data step arithmetic and comparisons

	*	multiplication
	/	division
	+	addition
	-	subtraction
	**	exponentiation
Comparison ⁴		
	= or EQ	equal to
	^=, ^=, ~=, or NE ¹	not equal to
	> or GT	greater than
	< or LT	less than
	>= or GE	greater than or equal to
	<= or LE	less than or equal to
	IN	equal to one of a list
Logical (Boolean)		
	& or AND	logical and
	or OR ²	logical or ¹
	~, ^, ~, or NOT ¹	logical not

19

Don't use the exponentiation symbol ** for any exponents except 2 or 3.

To compute a^b , don't use `a**b`. Instead, use the more numerically stable `log` and `exp` functions:

```
exp(b * log(a))
```

`log` is the natural log function; `log10` is common log or base 10 log.

Inverse of `log` is `exp`. Inverse of `b = log10(c)` is `exp(b*log(10.0))`.

See *LSB* §3.3 for a short list of SAS functions. For a complete list, see the SAS Documentation:

SAS Products > Base SAS > SAS Language Dictionary >
Dictionary of Language Elements > Functions and CALL Routines

20

TTEST: two-sample t-test

Compare years of education between men and women in their 20s (NHANES III) with a two-sample t -test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{with df} = (n_1 + n_2 - 2).$$

```
Proc TTEST  ci=none ;      omit CI for standard deviation
  class    group-variable ;
  var     response-variable ;
```

```
Proc TTEST  ci=none  data = one ;
  class    male ;
  var     education_yrs ;    corrected variable
```

21

Variable	male	N	Statistics				Std Dev	Std Err
			Lower CL	Mean	Upper CL	Mean		
education_yrs		0	1859	11.636	11.769	11.902	2.9235	0.0678
education_yrs		1	1625	11.191	11.345	11.5	3.1762	0.0788
education_yrs	Diff (1-2)			0.2213	0.424	0.6267	3.044	0.1034

Variable	Method	T-Tests		DF	t Value	Pr > t
		Variances	DF			
education_yrs	Pooled	Equal		3482	4.10	<.0001
education_yrs	Satterthwaite	Unequal		3326	4.08	<.0001

Variable	Method	Equality of Variances			Pr > F
		Num DF	Den DF	F Value	
education_yrs	Folded F	1624	1858	1.18	0.0005

Pooled t -test uses pooled standard deviation, which is the Root Mean Square Error in one-factor ANOVA:

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Satterthwaite t -test adjusts the test statistic and its degrees of freedom for differences in the group SDs.

22

When are the variances unequal?

The “Folded F -test” needs normally distributed data, so it is usually not a helpful test.

Instead, when the ratio

$$\frac{\text{larger standard deviation}}{\text{smaller standard deviation}} > 3$$

then the SDs are probably different. Satterthwaite t -test degrees of freedom will be much smaller than pooled df.

Report the larger p -value, or take logs, or try a nonparametric test. More on these later.

SAS Help Manual

Each procedure has a chapter in the Help File. Under the Help menu, select

SAS Help and Documentation > SAS Products > SAS/STAT >

SAS/STAT User’s Guide > The TTEST Procedure



The TTEST Procedure

The TTEST Procedure

- [Overview](#)
- [Getting Started](#)
- [Syntax](#)
- [Details](#)
- [Examples](#)
- [References](#)

[Previous](#) | [Next](#) | [Top of Page](#)

Getting Started a simple example with data, code and output

Syntax possible statements (commands), and definitions of options

The TTEST Procedure

Syntax

The following statements are available in PROC TTEST.

```
PROC TTEST < options > ;  
  CLASS variable ;  
  PAIRED variables ;  
  BY variables ;  
  VAR variables ;  
  FREQ variable ;  
  WEIGHT variable ;
```

No statement can be used more than once. There is no restriction on the order of the statements after the PROC statement.

Details Mathematical description of the computations, with references.

Examples Examples with data, code and output. Code can be copied and pasted

into the Editor window.

Making an indicator variable (0/1)

Compare high-school graduation rates between males and females in their 20s using the NHANES III data.

First we need to calculate an indicator variable for high-school graduation.

```
data two;
  set class.nhanes20;
  high_school_grad = 0;
  if (11 < education_yrs < 88) then high_school_grad = 1;
```

or equivalently

```
high_school_grad = (11 < education_yrs < 88);
= 1 when condition is true, = 0 when false
```

What happens when education_yrs is missing?

27

Proc FREQ: chi-square test

```
Proc FREQ;
  tables row-variable * column-variable / chisq relrisk ;

proc freq data=two;
  tables male * high_school_grad / chisq ;
```

Manual chapters for some procedures, such as Freq, Means, Univariate, Sort, are in

SAS Help and Documentation > SAS Products > Base SAS > SAS Procedures

male high_school_grad

Frequency			Total
Percent			
Row Pct			
Col Pct	0	1	
0	548	1311	1859
	15.73	37.63	53.36
	29.48	70.52	
	48.11	55.91	
1	591	1034	1625
	16.96	29.68	46.64
	36.37	63.63	
	51.89	44.09	
Total	1139	2345	3484
	32.69	67.31	100.00

Statistic	DF	Value	Prob
Chi-Square	1	18.7116	<.0001
Likelihood Ratio Chi-Square	1	18.6923	<.0001
Continuity Adj. Chi-Square	1	18.3997	<.0001