Introduction to SAS, Working with continuous variables

Steve Simon

Created 2021-05-30

Overview

- Using variable labels
- Simple descriptive statistics
- Printing row with smallest/largest value
- Missing value logic
- Simple transformations
- Histograms
- Correlations
- Scatterplots

author: Steve Simon

date: created 2021-05-30

purpose: to produce slides for module01 videos

Review definitions

- Categorical
  - Small number of possible values
  - Each value associated with a category
- Continuous
  - Large number of possible values
  - Potentially any value in an interval

Before we start, let's review a couple of definitions.

A **categorical variable** is a variable that can only take on a small number of values. Each value is usually associated with a particular category.

Examples of categorical variables are

sex (Male or Female),

race (White, Black, Native American, etc.),

cancer stage (I, II, III, or IV),

birth delivery type (Vaginal, C-section).

A **continuous variable** is a variable that can take on a large number of possible values, potentially any value in some interval.

Examples of continuous variables are

Birth weight in grams,

gestational age,

fasting LDL level.

There are some variables that are on the boundary between categorical and continuous, but it is not worth quibbling about here.

The point to remember is that the types of graphs that you use and the types of statistics that you compute are dependent on many things, but first and foremost on whether the variables are categorical, continuous, or a mixture.

Today, you will see examples involving mostly continuous variables.

Semicolons are important

- Ends every SAS statement
- Easy to forget
- Use this to your advantage
  - Several short lines
  - Indent continuations

Before I go too far, let me mention and important thing. Every SAS statement ends in a semicolon. This is important. You will forget a semicolon and it will lead to a cryptic error message. So here's a quick hint. If you get an error message on a certain line of code, look to see if you forgot a semicolon on the previous line. It happens to me all the time and I've been using SAS for decades.

Example of stretching statement across multiple lines.

One long line

```
statement option1 option2 option3 option4;
```

versus several short lines.

```
statement
  option1
  option2
  option3
  option4;
```

The use of semicolons is nice, in a way, because it allows you to stretch a complicated SAS statement across two or more rows of your program. This can often make your program more readable. It is hard to read a long line of code. Your eye has to scan left to right and you can sometimes lose track of which line you are on. Most newspapers place their articles in narrow columns because it makes them easier to read.

There is no official rule of thumb on this, but I do try to keep my lines below 50 characters. I also try to indent substatements with a data step or procedure. I use blank lines between data steps and procedures.

Don't obsess about this now, but you'll see a fairly consistent coding style that I use for my SAS code. You don't have to follow my format, of course, which might be a bit too extreme for your tastes. Just experiment with things a bit until you can settle on a layout that you are comfortable with.

Rules for variable names (1/2)

- Can use mix of
  - letters (A-Z, a-z),
  - numbers (0-9)
  - underscore (_)
  - no blanks, no symbols

There are important rules for variable names in SAS. You can use a mix of letters, numbers, and the underscore. You can't use blanks or any special symbols like the dollar sign ($) or the dash/minus sign.

I'm using the variable names provided but if you create your own names, use brief (but descriptive) name for EVERY variable in your data set. There's no precise rule, but names should be around 8 characters long. Longer variable names make your typing tedious and much shorter variable names makes your code terse and cryptic.

I'm a bit more terse with these variable names than I normally would be just to reduce the amount of typing you have to do.

You should avoid special symbols in your variable names especially symbols that are likely to cause confusion (the dash symbol, for example, which is also the symbol for subtraction). You should also avoid blanks. These rules are pretty much universal across most statistical software packages.

Rules for variable names (2/2)

- Can't start with a number
  - "a1" but not "1a"
- Capitalization not important
  - BMI, Bmi, bmi are same
- Up to 32 characters in length

You can't start with a number. So "a1" is okay, but "1a" is not.

Capitalization is not important in SAS. So you can call your variable "BMI" with all caps, or "Bmi" with mixed capitals or "bmi" with all lower case. SAS treats all of these the same.

Your variable name has to be 32 characters or less in length.

I'm using the variable names provided but if you create your own names, use brief (but descriptive) name for EVERY variable in your data set. There's no precise rule, but names should be around 8 characters long. Longer variable names make your typing tedious and much shorter variable names makes your code terse and cryptic.

I'm a bit more terse with these variable names than I normally would be just to reduce the amount of typing you have to do.

Recommendations for variable names (1/2)

- Avoid generic names (x1, var01, etc.)
- Keep it short
  - Use commonly known abbreviations…
  - …but nothing cryptic
- Use all lower case (age, not AGE or Age)

Your variable names should by descriptive. Avoid generic names like x1, var01, and so forth.

Keep things short. You can use commonly known abbreviations, such as “wt” for “weight”. But avoid any cryptic abbreviations.

I like to keep everything in lower case. It is more readable than all upper case, and easier to remember than a mixture of upper and lower case. Some people prefer an initial upper case, and there’s nothing wrong with that. It is important, however, to be consistent.

Recommendations for variable names (2/2)

- Separate words with underscores
  - fat_brozek, not fatbrozek
- Alternative: CamelCase
  - FatBrozek
- Caution: Writer's Exchange website
  - www.writersexchange.com

You can use two or three short words in a variable name, but be sure to separate them using underscores. So the variable for percentage body fat as measured by Brozek's equation is "fat" and "brozek" separated by an underscore. Some people prefer CamelCase, where each word starts with an initial capital letter: capital F fat, capital B brozek.

The one thing you do want to avoid is just running two or three words together and all lower case. There's a story about a group that started up in the era before the web called Writer's Exchange. As you can guess this was a resource for new authors. When the web came out, they decided to put their resources up on a website, www.writersexchange.com. That seemed logical enough, but then someone notices that you could read the website as "www dot writer sex change dot com". Not exactly the image they wanted.

SAS variable labels (1/2)

- Longer description of a variable
  - Can include blanks, special symbols
  - Internal documentation
  - Labels substituted on some (but not all) output
- Required in this class (see grading rubric)

SAS offers an opportunity for you to add documentation to your program about individual variables. These are called variable labels. They have almost no restrictions. You can use blanks, or special symbols like a dollar sign or a dash. The documentation that variable labels provide is mostly internal, but these labels are substituted in a few places like some graphs.

I strongly recommend use of variable labels and will require them for any homework you submit in this class. See the grading rubric for details.

SAS variable labels (2/2)

- Recommendations for variable labels
  - Judicious use of upper and lower case
  - Spell out abbreviations
  - Specify units of measurement
  - Any other important details

What makes a good variable label? First take advantage of a mixture of upper and lower case to make your labels more readable. Spell out any abbreviations, even obvious abbreviations. If your variable has a measurement unit, specify that unit in your variable label. If there are other important details, put these in the variable label as well.

Break #1

- What have you learned so far
  - Rules for variable names
  - Using variable labels
- What's next
  - SAS code with variable labels

SAS Code: Part01. Documentation header

```
* 5507-02-simon-continuous-variables.sas
* author: Steve Simon
* date: created 2021-05-30
* purpose: to work with continuous
variables
* license: public domain;

options papersize=(6in 4in);
```

Speaker notes: This is the standard documentation header.

SAS Code: Part02. Tell SAS where to find and store things.

```
filename fat
  "q:/introduction-to-sas/data/fat.txt";

libname intro
  "q:/introduction-to-sas/data";
```

SAS Code: Part02. Tell SAS where to find and store things.

```
ods pdf file=
  "q:/introduction-to-sas/results/5507-02-
simon-continuous-variables.pdf";
```

Speaker notes: 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 mostly continuous variables. The first dataset at body measurements.

The input statement is very long and does not fit on a single slide. Go to the Canvas site if you want to see the full code.

SAS Code: Part03. Read in your data

```
data intro.fat;
  infile fat;
  input
    case
    fat_brozek
    fat_siri
    dens
```

SAS Code: Part03. Read in your data

```
age
wt
ht
bmi
ffw
neck
chest
abdomen
```

SAS Code: Part03. Read in your data

```
hip
thigh
knee
ankle
biceps
forearm
wrist;
```

Speaker notes: This is the code to input all the variables in this data set. It is quite long and does not fit on a single Powerpoint slide.

SAS Code: Part04. Add variable labels

```
label
    case="Case number"
    fat_brozek="Fat (Brozek's equation)"
    fat_siri="Fat (Siri's equation)"
    dens="Density"
    age="Age (yrs)"
    wt="Weight (lbs)"
```

SAS Code: Part04. Add variable labels

```
ht="Height (inches)"
bmi="Body mass index (kg/m^2)"
ffw="Fat Free Weight (lbs)"
neck="Neck circumference (cm)"
chest="Chest circumference (cm)"
abdomen="Abdomen circumference (cm)"
hip="Hip circumference (cm)"
```

SAS Code: Part04. Add variable labels

```
    thigh="Thigh circumference (cm)"
    knee="Knee circumference (cm)"
    ankle="Ankle circumference (cm)"
    biceps="Biceps circumference (cm)"
    forearm="Forearm circumference (cm)"
    wrist="Wrist circumference (cm)";
run;
```

Speaker notes: SAS offers an opportunity for you to add documentation to your program about individual variables. These are called variable labels. They have almost no restrictions. You can use blanks, or special symbols like a dollar sign or a dash. The documentation that variable labels provide is mostly internal, but these labels are substituted in a few places like some graphs.

I strongly recommend use of variable labels and will require them for any homework you submit in this class. See the grading rubric for details.

What makes a good variable label? First take advantage of a mixture of upper and lower case to make your labels more readable. Spell out any abbreviations, even obvious abbreviations. If your variable has a measurement unit, specify that unit in your variable label. If there are other important details, put these in the variable label as well.

Every variable in a SAS program should have a label. This label will make some (but not all) of the SAS output more readable. it is also part of the internal documentation of your program. Note that some of these labels do not fit well in this Powerpoint slide, but that's okay.

SAS Code: Part04. Add variable labels

```
* Some additional details about this data:

Brozek's equation is 457/Density - 414.2

Siri's equation is 495/Density - 450
```

SAS Code: Part04. Add variable labels

```
Abdomen circumference is measured at the
umbilicus and level with the iliac crest

Wrist circumference is distal to the
styloid processes;
```

Speaker notes: I am including some additional details that would not fit easily into the variable labels. How much documentation you include is a judgment call. I am including this extra documentation just to remind you that such documentation is possible.

SAS Code: Part05. Print a small piece of the data

```
proc print
    data=intro.fat(obs=10);
  var case fat_brozek fat_siri dens age;
  title1 "Ten rows and five columns";
  title2 "of the fat data set";
run;
```

Speaker notes: It's always a good idea to print out a small piece of your data to make sure everything is okay.

The data option tells SAS what data set you want to print. If you omit this, SAS will print the most recently created data set.

The obs=10 option limits the number of rows printed to the first 10. For large data sets, you should always take advantage of this option.

The var statement limits the variables that you print to those that you specify. Again, this is important for large data sets.

Please do not ever print more than ten rows or more than five variables, if you can help it. Excessively lengthy outputs will lose you a few points (see the grading rubric).

The title statement tells SAS to provide a descriptive title at the top of the page of output.

The run statement says you're done with the procedure.

SAS output: Part05. Print a small piece of the data

13:41 Thursday, June 16, 2022 1

Ten rows and five columns
of the fat data set

Obs	case	fat_brozek	fat_siri	dens	age
1	1	12.6	12.3	1.0708	23
2	2	6.9	6.1	1.0853	22
3	3	24.6	25.3	1.0414	22
4	4	10.9	10.4	1.0751	26
5	5	27.8	28.7	1.0340	24
6	6	20.6	20.9	1.0502	24
7	7	19.0	19.2	1.0549	26
8	8	12.8	12.4	1.0704	25
9	9	5.1	4.1	1.0900	25
10	10	12.0	11.7	1.0722	23

Figure 1. proc print.

Speaker notes: There are no obvious problems with this dataset.

SAS Code: Part05. Print a small piece of the data

```
proc contents
    data=intro.fat;
  title1 "Internal description of fat
dataset";
run;
```

Speaker notes. The contents procedure produces information about any dataset produced by SAS, including both temporary datasets (one part names) and permanent datasets (two part names).

For a dataset that you just created and one that is not all that complicated, using proc contents is overkill. I am showing it so you will know how to use proc contents for very complex datasets, especially ones that were created by someone other than yourself.

SAS output: Part05. Print a small piece of the data

13:41 Thursday, June 16, 2022 2

Internal description of fat dataset

The CONTENTS Procedure

Data Set Name	INTRO.FAT	Observations	252
Member Type	DATA	Variables	19
Engine	V9	Indexes	0
Created	06/16/2022 13:50:26	Observation Length	152
Last Modified	06/16/2022 13:50:26	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Figure 2. proc contents (1 of 4).

Speaker notes: SAS produces a lot of information and much of it is only relevant for advanced applications. You have to wade through the details to get the important information. The important information on this page is

date created,

date modified,

observations, and

variables.

SAS output: Part05. Print a small piece of the data

Internal description of fat dataset

The CONTENTS Procedure

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	430
Obs in First Data Page	252
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	q:\introduction-to-sas\data\fat.sas7bdat
Release Created	9.0401M3
Host Created	X64_8PRO

Figure 3. proc contents (2 of 4).

Speaker notes. The only important things on this page are

filename, and

release created (which tells the precise version of SAS that was used to create this dataset.

SAS output: Part05. Print a small piece of the data

13:41 Thursday, June 16, 2022 4

Internal description of fat dataset

The CONTENTS Procedure

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
12	abdomen	Num	8	Abdomen circumference (cm)
5	age	Num	8	Age (yrs)
16	ankle	Num	8	Ankle circumference (cm)
17	biceps	Num	8	Biceps circumference (cm)
8	bmi	Num	8	Body mass index (kg/m^2)
1	case	Num	8	Case number
11	chest	Num	8	Chest circumference (cm)
4	dens	Num	8	Density
2	fat_brozek	Num	8	Fat (Brozek's equation)
3	fat_siri	Num	8	Fat (Siri's equation)

Figure 4. proc contents (3 of 4).

Speaker notes: This page and the following page lists all the variables in the dataset, their type (all numeric in this dataset), and the variable label.

SAS output: Part05. Print a small piece of the data

13:41 Thursday, June 16, 2022 5

Internal description of fat dataset

The CONTENTS Procedure

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
9	ffw	Num	8	Fat Free Weight (lbs)
18	forearm	Num	8	Forearm circumference (cm)
13	hip	Num	8	Hip circumference (cm)
7	ht	Num	8	Height (inches)
15	knee	Num	8	Knee circumference (cm)
10	neck	Num	8	Neck circumference (cm)
14	thigh	Num	8	Thigh circumference (cm)
19	wrist	Num	8	Wrist circumference (cm)
6	wt	Num	8	Weight (lbs)

Figure 5. proc contents (4 of 4).

Break #2

- What have you learned so far
  - SAS code with variable labels
- What's next
  - Simple statistics
  - Printing row with smallest/largest value

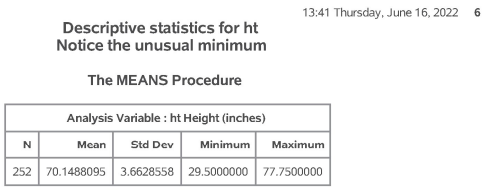
SAS Code: Part06. Calculate simple statistics for ht

```
proc means
    n mean std min max
    data=intro.fat;
  var ht;
  title1 "Descriptive statistics for ht";
  title2 "Notice the unusual minimum";
run;
```

Speaker notes. The means procedure will produce descriptive statistics for your data. By default, it will produce the count of non-missing values, the mean, the standard deviation, and the minimum and maximum values, but I am listing them explicitly here, just for show.

The data option tells SAS which data set you want descriptive statistics on, and the var statement tells SAS which variable(s) you want descriptive statistics on.

SAS output: Part06. Calculate simple statistics for ht



13:41 Thursday, June 16, 2022 6

Descriptive statistics for ht
Notice the unusual minimum

The MEANS Procedure

Analysis Variable : ht Height (inches)				
N	Mean	Std Dev	Minimum	Maximum
252	70.1488095	3.6628558	29.5000000	77.7500000

Figure 6. proc means.

Speaker notes: This is what your output looks like.

Notice the unusual minimum value. While this is not totally outside the realm of possibility, you should always ask when you see something unusual like this.

SAS Code: Part07. Look at smallest value

```
proc sort
    data=intro.fat;
  by ht;
run;
```

SAS Code: Part07. Look at smallest value

```
proc print
    data=intro.fat(obs=1);
  title1 "The row with the smallest ht";
  title2 "Note the inconsistency with wt";
run;
```

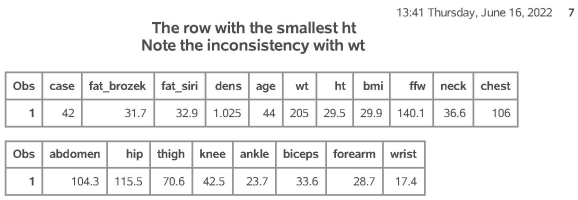
Speaker notes. First, let's look at this value in the context of the other values in this row of data.

You do this by sorting the data so that the shortest subject becomes the first row of the data and the tallest subject becomes the last. Then print just the very first row of your data.

Warning! Be careful about sorting your data if you can't get the data easily back to the original order. It might be okay, but there are times when you'd like your data all the way back and that means data in the original order. This data set has a case variable that you can resort by in order to get back ot the original order.

If you don't have a case variable, store the sorted data in a separate location: something along the lines of proc sort data=x out=y.

SAS output: Part07. Look at smallest value



13:41 Thursday, June 16, 2022 7

The row with the smallest ht
Note the inconsistency with wt

Obs	case	fat_brozek	fat_siri	dens	age	wt	ht	bmi	ffw	neck	chest
1	42	31.7	32.9	1.025	44	205	29.5	29.9	140.1	36.6	106

Obs	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
1	104.3	115.5	70.6	42.5	23.7	33.6	28.7	17.4

Figure 7. proc print

Speaker notes: This is what your output looks like.

There is no possible way that a height of 29.5 inches could be paired with a weight of 205 pounds.

With this outlier on the low end, you might consider doing nothing other than noting the unusual value.

Alternately, you could delete the entire row associated with this value. Finally, you might consider converting the small ht value to a missing value code.

There is no one method that is preferred in this setting. If you encounter an unusual value, you should discuss it with your research team, investigate the original data sources, if possible, and review any procedures for handling unusual data values that might be specified in your research protocol.

Your data set may arrive with missing values in it already. Data might be designated as missing for a variety of reasons (lab result lost, value below the limit of detection, patient refused to answer this question) and how you handle missing values is way beyond the scope of this class. Just remember to tread cautiously around missing values as they are a

minefield.

Notice that I store the revised data sets with the row removed and with the 29.5 replaced by a missing value in different data frames. This is good programming practice. If you ever have to make a destructive change to your data set (a change that wipes out one or more values or a change that is difficult to undo), it is good form to store the new results in a fresh spot. That way, if you get cold feet, you can easily backtrack.

We'll use the data set with the 29.5 changed to a missing value for all of the remaining analyses of this data set.

Logic statements involving missing value codes are tricky. SAS stores missing value codes as the most extreme legal negative number. So if you want, for example, to exclude negative values, make sure that you account for missing values as well.

```
(ht < 0) & (ht ~= .)
```

SAS Code: Part08. Look at the largest value

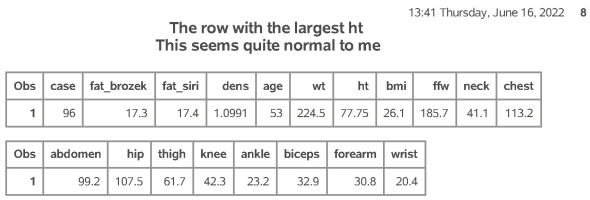
```
proc sort
    data=intro.fat;
  by descending ht;
run;
```

SAS Code: Part08. Look at the largest value

```
proc print
    data=intro.fat(obs=1);
  title1 "The row with the largest ht";
  title2 "This seems quite normal to me";
run;
```

Speaker notes: Just for the sake of completeness, let's look at the row of data with the largest height value. Add the keyword desc to sort the data in reverse order.

SAS output: Part08. Look at the largest value



13:41 Thursday, June 16, 2022 8

The row with the largest ht
This seems quite normal to me

Obs	case	fat_brozek	fat_siri	dens	age	wt	ht	bmi	ffw	neck	chest
1	96	17.3	17.4	1.0991	53	224.5	77.75	26.1	185.7	41.1	113.2

Obs	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
1	99.2	107.5	61.7	42.3	23.2	32.9	30.8	20.4

Figure 8. proc print.

Speaker notes: This is what your output looks like. These values seem reasonable to me.

SAS Code: Part09. Removing the entire row

```
data intro.fat1;
  set intro.fat;
  if ht > 29.5;
run;
```

Speaker notes: This code removes the entire row of data. Notice that I store the modified data under a new name. That way, if I regret tossing the entire row out, I can easily revert to the original data.

SAS Code: Part09. Removing the entire row

```
proc contents
    data=intro.fat;
  title1 "Internal description of fat
dataset";
run;
```

Speaker notes: It is reasonable to check the contents when you create a new file. In this case, the change is so small that it is definitely overkill. I just want to encourage you to think about using proc contents as a way of reviewing your work in more complex settings.

SAS output: Part09. Removing the entire row

Internal description of fat dataset

The CONTENTS Procedure

Data Set Name	INTRO.FAT	**Observations**	252
Member Type	DATA	**Variables**	19
Engine	V9	**Indexes**	0
Created	06/16/2022 13:50:27	**Observation Length**	152
Last Modified	06/16/2022 13:50:27	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	YES
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Figure 9. proc contents (1 of 5).

SAS output: Part09. Removing the entire row

13:41 Thursday, June 16, 2022 10

Internal description of fat dataset

The CONTENTS Procedure

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	430
Obs in First Data Page	252
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	q:\introduction-to-sas\data\fat.sas7bdat
Release Created	9.0401M3
Host Created	X64_8PRO

Figure 10. proc contents (2 of 5).

Speaker notes. The only important things on this page are

filename, and

release created (which tells the precise version of SAS that was used to create this dataset.

SAS output: Part09. Removing the entire row

13:41 Thursday, June 16, 2022 11

Internal description of fat dataset

The CONTENTS Procedure

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
12	abdomen	Num	8	Abdomen circumference (cm)
5	age	Num	8	Age (yrs)
16	ankle	Num	8	Ankle circumference (cm)
17	biceps	Num	8	Biceps circumference (cm)
8	bmi	Num	8	Body mass index (kg/m^2)
1	case	Num	8	Case number
11	chest	Num	8	Chest circumference (cm)
4	dens	Num	8	Density
2	fat_brozek	Num	8	Fat (Brozek's equation)
3	fat_siri	Num	8	Fat (Siri's equation)

Figure 11. proc contents (3 of 5).

Speaker notes: This page and the following page lists all the variables in the dataset, their type (all numeric in this dataset), and the variable label.

SAS output: Part09. Removing the entire row

13:41 Thursday, June 16, 2022 12

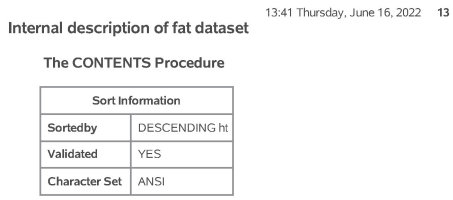
Internal description of fat dataset

The CONTENTS Procedure

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
9	ffw	Num	8	Fat Free Weight (lbs)
18	forearm	Num	8	Forearm circumference (cm)
13	hip	Num	8	Hip circumference (cm)
7	ht	Num	8	Height (inches)
15	knee	Num	8	Knee circumference (cm)
10	neck	Num	8	Neck circumference (cm)
14	thigh	Num	8	Thigh circumference (cm)
19	wrist	Num	8	Wrist circumference (cm)
6	wt	Num	8	Weight (lbs)

Figure 12. proc contents (4 of 5).

SAS output: Part09. Removing the entire row



13:41 Thursday, June 16, 2022 13

Internal description of fat dataset

The CONTENTS Procedure

Sort Information	
Sortedby	DESCENDING ht
Validated	YES
Character Set	ANSI

Figure 13. proc contents (5 of 5).

Speaker notes: You get an extra page for this dataset because it notes that your data is sorted by descending height.

”

SAS Code: Part10. Converting the outlier to a missing value

```
data intro.fat2;
  set intro.fat;
  if ht=29.5 then ht=.;
run;
```

Speaker notes: This code converts the height to a missing value, but keeps the original data.

I won't use proc contents a third time.

There is no one method that is preferred in this setting. If you encounter an unusual value, you should discuss it with your research team, investigate the original data sources, if possible, and review any procedures for handling unusual data values that might be specified in your research protocol.

Your data set may arrive with missing values in it already. Data might be designated as missing for a variety of reasons (lab result lost, value below the limit of detection, patient refused to answer this question) and how you handle missing values is way beyond the scope of this class. Just remember to tread cautiously around missing values as they are a minefield.

Notice that I store the revised data sets with the row removed and with the 29.5 replaced by a missing value in different data frames. This is good programming practice. If you ever have to make a destructive change to your data set (a change that wipes out one or more values or a change that is difficult to undo), it is good form to store the new results in a fresh spot. That way, if you get cold feet, you can easily backtrack.

We'll use the data set with the 29.5 changed to a missing value for all of the remaining analyses of this data set.

SAS Code: Part11. Faulty approach for filtering out negative values

```
proc print
    data=intro.fat2;
  where ht < 0;
  title1 "ht < 0 will include ht = .";
run;
```

Speaker notes: Here's an important thing to remember about missing values. SAS stores missing value codes as the most extreme legal negative number. This can sometimes lead to surprising and misleading results.

Every procedure in SAS has its own default approach to missing values and often provides you with one or more alternatives. You have to review this carefully for each and every statistical procedure that you run. If you do data manipulations involving missing values, you have to make sure that the result correctly reflects what you want.

SAS output: Part11. Faulty approach for filtering out negative values

ht < 0 will include ht = .

Obs	case	fat_brozek	fat_siri	dens	age	wt	ht	bmi	ffw	neck	chest
252	42	31.7	32.9	1.025	44	205	.	29.9	140.1	36.6	106

Obs	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
252	104.3	115.5	70.6	42.5	23.7	33.6	28.7	17.4

Figure 14. proc print.

Speaker notes: This is what your output looks like.

In order to prevent this from happening, you need to check for missingness before applying any other logic statement.

The proper way to search for negative ht values

```
where (ht < 0) & (ht ~= .)
```

You may hate having to do this and wish that SAS would have handled things differently. Different packages, like R, have a three valued logic system where every logic statement (well, almost every logic statement) involving missing values codes to MISSING rather than to TRUE or FALSE. This sometimes works better, but sometimes the SAS approach works better.

SAS Code: Part12. Counting missing values

```
proc means
    n nmiss mean std min max
    data=intro.fat2;
  var ht;
  title "Using the nmiss statistic";
run;
```

Speaker notes: If you are concerned at all about missing values (and you should be), ask for the number of missing values in proc means using nmiss.

SAS output: Part12. Counting missing values

13:41 Thursday, June 16, 2022 15

Using the nmiss statistic

The MEANS Procedure

Analysis Variable : ht Height (inches)					
N	N Miss	Mean	Std Dev	Minimum	Maximum
251	1	70.3107570	2.6142960	64.0000000	77.7500000

Figure 15. proc means.

Speaker notes: This is what your output looks like. Note that your data set has 251 observations and 1 missing value.

Break #3

- What have you learned
  - Handling outliers
  - Missing values
- What's next
  - Simple transformations
  - Histograms

SAS Code: Part13. Simple transformations

```
data converted_units;
  set intro.fat2;
  ht_cm = round(ht * 2.54, 0.01);
  wt_kg = round(wt / 2.2, 0.01);
run;
```

SAS Code: Part13. Simple transformations

```
proc print
    data=converted_units(obs=10);
  var ht ht_cm wt wt_kg;
  title1 "Original and converted units";
run;
```

Speaker notes: You can do simple transformations like unit conversions in SAS. Create a new dataset with the data statement. Use the set command to tell SAS that you plan to use and modify an existing dataset.

The conversions done here will turn height and weight into centimeters and kilograms, respectively.

SAS output: Part13. Simple transformations

13:41 Thursday, June 16, 2022 16

Original and converted units

Obs	ht	ht_cm	wt	wt_kg
1	77.75	197.49	224.50	102.05
2	77.50	196.85	188.15	85.52
3	76.00	193.04	216.00	98.18
4	76.00	193.04	244.25	111.02
5	75.50	191.77	194.00	88.18
6	75.25	191.14	171.50	77.95
7	75.00	190.50	212.75	96.70
8	74.75	189.87	210.25	95.57
9	74.75	189.87	224.75	102.16
10	74.50	189.23	186.25	84.66

Figure 16. proc print.

Speaker notes: This is your output with measurements both in the original units and metric. Notice that I did not print any more than 10 rows of data.

SAS Code: Part14. Display a histogram

```
proc sgplot
    data=intro.fat2;
  histogram ht;
  title1 "Histogram with default bins";
run;
```

Speaker notes: Here is the code to create a histogram with the default option. Generally, it is wise to modify the defaults for any graphic image.

SAS output: Part14. Display a histogram

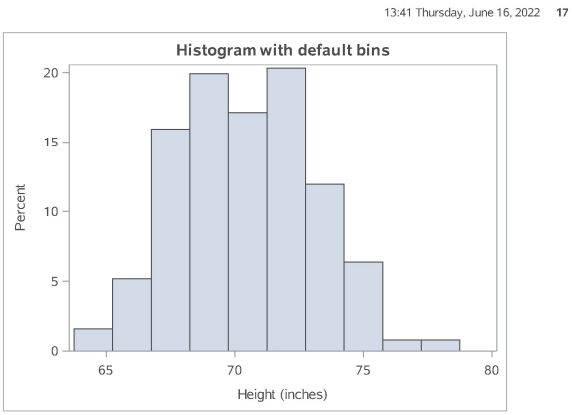


Figure 17. proc sgplot.

Speaker notes: This is the default histogram.

SAS Code: Part15. Revised histogram with narrow bins

```
proc sgplot
    data=intro.fat2;
  histogram ht / binstart=60 binwidth=1;
  title "Histogram with narrow bins";
run;
```

Speaker notes: Here's the code to create a histogram with many bars. The first bar is centered at 60, and each bin has a width of 1 inch (plus or minus 0.5 inches)

SAS output: Part15. Revised histogram with narrow bins

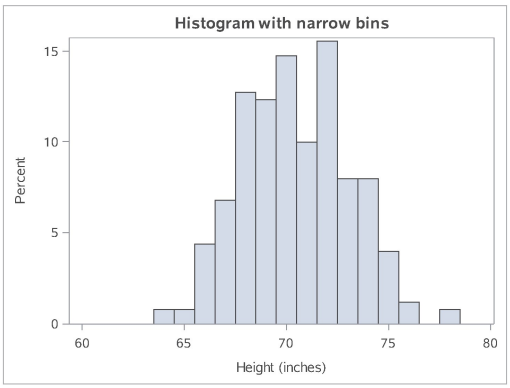


Figure 18. proc sgplot.

Speaker notes: This is what you get. You can also go in the opposite direction.

SAS Code: Part16. Revised histogram with wide bins

```
proc sgplot
    data=intro.fat2;
  histogram ht / binstart=60 binwidth=5;
  title "Histogram with wide bins";
run;
```

Speaker notes: Here's the code to create a histogram with few bars. The first bar is again centered at 60, but now each bin has a width of 5 inches (plus or minus 2.5 inches).

SAS output: Part16. Revised histogram with wide bins

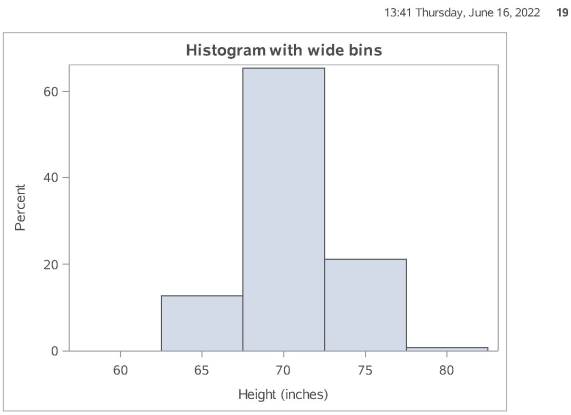


Figure 19. proc sgplot.

Speaker notes: This is the revised histogram. There is no “correct” version of the histogram. Try several widths and see which one gives the clearest picture of your data.

Break #4

- What have you learned
  - Simple transformations
  - Histograms
- What's next
  - Correlations
  - Scatterplots

Correlations

- Informal interpretation
  - between +0.7 and +1.0: strong positive association
  - between +0.3 and +0.7: weak positive association
  - between -0.3 and +0.3: little or no association
  - between -0.3 and -0.7: weak positive association
  - between -0.7 and -1.0: strong negative association

The correlation coefficient is a single number between -1 and +1 that quantifies the strength and direction of a relationship between two continuous variables. As a rough rule of thumb, a correlation larger than +0.7 indicates a strong positive association and a correlation smaller than -0.7 indicates a strong negative association. A correlation between +0.3 and +0.7 (-0.3 and -0.7) indicates a weak positive (negative) association. A correlation between -0.3 and +0.3 indicates little or no association.

Don't take these rules too literally. You're not trying to make definitive statements about your data set. You are just trying to get comfortable with some general patterns that occur in your data set. A complex and definitive statistical analysis will almost certainly not agree with at least some of the preliminary correlations noted here.

The corr procedure produces, by default, a square correlation matrix of all the numeric variables. The noprob and nosimple options cut down on the amount of information printed. The with statement produces a rectangular correlation matrix.

SAS Code: Part17. Calculate correlations

```
proc corr
    data=intro.fat2
    noprob nosimple;
  var fat_brozek fat_siri;
  with neck -- wrist;
  title "Correlation matrix";
run;
```

Speaker notes: Here's the code to compute correlations.

SAS output: Part17. Calculate correlations

The CORR Procedure

10 With Variables:	neck chest abdomen hip thigh knee ankle biceps forearm wrist
2 Variables:	fat_brozek fat_siri

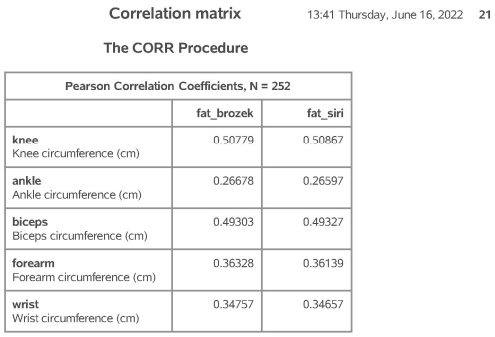
Pearson Correlation Coefficients, N = 252

	fat_brozek	fat_siri
neck Neck circumference (cm)	0.49149	0.49059
chest Chest circumference (cm)	0.70289	0.70262
abdomen Abdomen circumference (cm)	0.81371	0.81343
hip Hip circumference (cm)	0.62570	0.62520
thigh Thigh circumference (cm)	0.56128	0.55961

Figure 20. proc corr (1 of 2).

Speaker notes: The output here extends to a fresh page.

SAS output: Part17. Calculate correlations



Correlation matrix 13:41 Thursday, June 16, 2022 21

The CORR Procedure

Pearson Correlation Coefficients, N = 252		
	fat_brozek	fat_siri
knee Knee circumference (cm)	0.50779	0.50867
ankle Ankle circumference (cm)	0.26678	0.26597
biceps Biceps circumference (cm)	0.49303	0.49327
forearm Forearm circumference (cm)	0.36328	0.36139
wrist Wrist circumference (cm)	0.34757	0.34657

Figure 21. proc corr (2 of 2).

Speaker notes: The output here really annoys me. I want to show something a bit advanced here.

SAS Code: Part18. Save the correlations in a separate data file.

```
proc corr
    data=intro.fat2
    noprint
    outp=correlations;
  var fat_brozek fat_siri;
  with neck -- wrist;
run;
```

Speaker notes: You can save the correlations in a separate data file.

SAS output: Part18. Save the correlations in a separate data file.

13:41 Thursday, June 16, 2022 22

Correlation matrix output to a data set

Obs	_TYPE_	_NAME_	fat_brozek	fat_siri
1	MEAN		18.938	19.151
2	STD		7.751	8.369
3	N		252.000	252.000
4	CORR	neck	0.491	0.491
5	CORR	chest	0.703	0.703
6	CORR	abdomen	0.814	0.813
7	CORR	hip	0.626	0.625
8	CORR	thigh	0.561	0.560
9	CORR	knee	0.508	0.509
10	CORR	ankle	0.267	0.266
11	CORR	biceps	0.493	0.493

Figure 22. proc print (1 of 2).

Speaker notes: Continues on the next slide.

SAS output: Part18. Save the correlations in a separate data file.

Correlation matrix output to a data set

Obs	_TYPE_	_NAME_	fat_brozek	fat_siri
12	CORR	forearm	0.363	0.361
13	CORR	wrist	0.348	0.347

Figure 23. proc print (2 of 2).

Speaker notes: The output is a bit unusual because SAS wants to include means and standard deviations in your output. You can and should remove this. It would be easy enough to do (use the where statement), but I wanted to show you the full data set.

SAS Code: Part19. Modify these correlations.

```
data correlations;
  set correlations;
  if _type_="CORR";
  drop _type_;
  fat_brozek=round(100*fat_brozek);
  fat_siri=round(100*fat_siri);
run;
```

Speaker notes: Saving as a data file allows you to manipulate the individual correlations. Here we multiply the correlations by 100, round them, and sort them. This can often simplify the interpretation of large correlation matrices.

This code does the reordering and printing

SAS Code: Part20. Print the modified correlations.

```
proc print
    data=correlations;
  title "Rounded and re-ordered
correlation matrix";
run;
```

Speaker notes: Just to help visualize things, let's print the file before we modify it.

SAS output: Part20. Print the modified correlations.

13:41 Thursday, June 16, 2022 24

Rounded and re-ordered correlation matrix

Obs	_NAME_	fat_brozek	fat_siri
1	abdomen	81	81
2	chest	70	70
3	hip	63	63
4	thigh	56	56
5	knee	51	51
6	neck	49	49
7	biceps	49	49
8	forearm	36	36
9	wrist	35	35
10	ankle	27	27

Figure 24. proc print.

Speaker notes: This is the output. You can see that measurements at the extremities are poor predictors of body fat. Apparently, we grow fat from the middle outward.

SAS Code: Part21. Draw a scatterplot.

```
proc sgplot
    data=intro.fat2;
  scatter x=abdomen y=fat_brozek;
  title1 "Simple scatterplot";
run;
```

Speaker notes: A scatterplot is also useful for examining the relationship among variables. You can produce scatterplots several different ways, but the scatterplots produced by the sgplot procedure have the most flexibility.

SAS output: Part21. Draw a scatterplot.

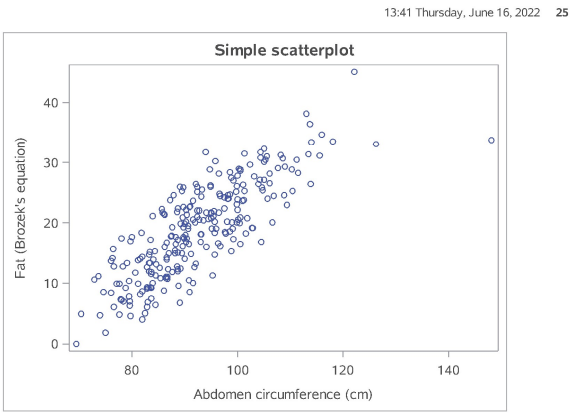


Figure 25. proc sgplot.

Speaker notes: This plot shows a general upward trend.

SAS Code: Part22. Adding linear trend line.

```
proc sgplot
    data=intro.fat2;
  scatter x=abdomen y=fat_brozek;
  reg x=abdomen y=fat_brozek;
  title2 "with linear trend";
run;
```

Speaker notes: The trend line is very useful for large and noisy data sets. It also allows you to more quickly visualize extreme values.

Notice that there is no title1. When you leave this out, SAS will pull the title1 used in the previous procedure, if it is available. This allows you to repeat the top line title across broad sections of your program.

SAS output: Part22. Adding linear trend line.

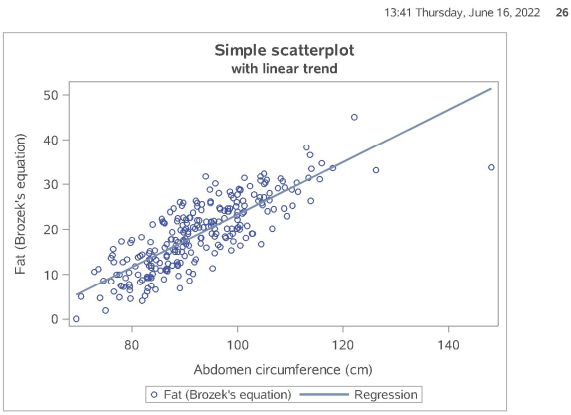


Figure 26. proc sgplot.

Speaker notes: Notice, for example, that the person with the largest abdomen measure (the biggest gut, if I can be informal) is quite out of line with what you might expect the relationship to be.

SAS Code: Part23. Adding a smooth curve.

```
proc sgplot
    data=intro.fat2;
  scatter x=abdomen y=fat_brozek;
  pbspline x=abdomen y=fat_brozek;
  title2 "with a smooth curve";
run;

ods pdf close;
```

Speaker notes: Here's the code to compute a smoothing spline. It helps you visualize whether the trend is linear or not.

SAS output: Part23. Adding a smooth curve.

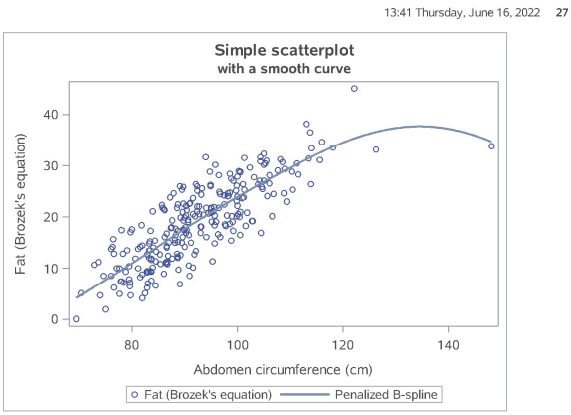


Figure 27. proc sgplot.

Speaker notes: The smoothing spline provides some evidence that the relationship is roughly linear a low and medium abdomen measurements, but tends to level off a bit at higher levels. Interpret this with caution, of course, because you have very little data at extrmemy high adbomen measures.

Don’t forget!

```
ods pdf close;
```

I always seem to forget this last statement and then I get upset with SAS for not providing the PDF output. But SAS can’t produce the PDF output until you tell it you are done. So don’t yell at your computer when it’s your own darn fault (just like Jimmy Buffet in the Margaritaville song).

Summary

– What have you learned
  - Using variable labels
  - Printing a small piece of data
  - Simple descriptive statistics
  - Printing row with smallest/largest value
  - Missing value logic
  - Simple transformations
  - Histograms
  - Correlations
  - Scatterplots