Epidemiology 640 – Fall 2017 NAME: Steph Mecham

SECTION: 4

Homework 3: Working with Datasets

This homework assignment continues our replication of **Bansil et al., 2011** using the 20013-2014 data from NHANES. You may discuss the assignment with others but you MUST write your own code and answers in your own words. Turn in a printed copy of your write-up and SAS commands (attached as a printout to the back of your write-up) at **9:30 AM on Friday, October 13th,** and **submit an electronic copy** of your write-up and code via Canvas Assignment tab. ONLY output that is relevant to the questions should be included.

Be sure that: 1) Your SAS code runs from start to finish

2) Your results make sense (check your sample size and look for unreasonable, unlikely, or impossible answers)

3) Your code is well commented. Commenting is done by either *type in any text here**;** or /* type in text here */. Be sure to include the homework number and your name at the top of your homework and code, identify each question in the code, and describe each new task. Format your code (indentation and carriage returns) to improve readability. **5% will be deducted** if these two tasks are not completed.

4) Make sure your homework has your name on EACH page and that your file is a .doc, .docx or .pdf in the following format: LASTNAME_FIRSTNAME_HW3.docx

---

Create a new folder on your computer or flash drive specifically for Homework 3. Make a permanent library called **nhanes** in SAS that points to the folder you just created. Download the included dataset from Canvas and place them in the folder on your computer/flash drive that corresponds to your nhanes library. **Include the code to create your library with the code submitted with the homework**.

DEMO_BP Dataset

The **demo_bp** dataset provided is a combination of demographics (DEMO_H), blood pressure measures (BPX_H), and self-reported medication use for hypertension (BPQ_H). It also includes created variables for the mean systolic (`sysbp_ave`) and diastolic (`diabp_ave`) blood pressure measures (there are up to three for each person in NHANES), an indicator for current blood pressure medication use (`bpmed`), and an indicator for hypertension (`htn`). You will create these variables yourself in future homework.

1. From the **demo_bp** dataset provided, create a new dataset in your **nhanes** library called **BPdata** that contains only those variables that we will need for our analysis (`SEQN, RIDAGEYR, RIAGENDR, RIDRETH1, INDFMPIR, DMDEDUC2, RIDEXPRG, sysbp_ave, diabp_ave, bpmed,` and `htn`).

a. Paste a copy of the log output that describes how many records and how many variables are in the **BPdata** dataset.

```
NOTE: There were 10175 observations read from the data set NHANES.DEMO_BP.
NOTE: The data set NHANES.BPDATA has 10175 observations and 11 variables.
```

b. From Homework 2, we saw that NHANES documents different groups of variables from the same person within different datasets. For example, the demographic variables and the blood pressure measure variables were stored in DEMO and BPX_H datasets, respectively. How can we identify the same groups of sampled participants across different NHANES datasets? What variable can you use to accomplish this?

We can look at the sequence number from each dataset to identify all the data for the same person across variables.

2. Run a PROC CONTENTS on the **BPdata** dataset and paste the lines from the Alphabetic List of Variables and Attributes in your homework.

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
6	DMDEDUC2	Num	8
5	INDFMPIR	Num	8
3	RIAGENDR	Num	8
2	RIDAGEYR	Num	8
7	RIDEXPRG	Num	8
4	RIDRETH1	Num	8
1	SEQN	Num	8

10	bpmed	Num	8
9	diabp_ave	Num	8
11	htn	Num	8
8	sysbp_ave	Num	8

3. Using the variable definition from the codebook that we created in Homework 2, now apply labels to `SEQN`, `RIDAGEYR`, `RIAGENDR`, `RIDRETH1`, `DMDEDUC2`, and `RIDEXPRG` in your **BPdata** dataset. To show that your labels have been properly applied, re-run the PROC CONTENTS procedure on the **BPdata** dataset and paste the Alphabetic List of Variables and Attributes into your homework.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
3	Age	Num	8
5	Education	Num	8
1	IDNumber	Num	8
7	INDFMPIR	Num	8
6	PregnancyStatus	Num	8
4	Race	Num	8
2	Sex	Num	8
10	bpmed	Num	8

9	diabp_ave	Num	8
11	htn	Num	8
8	sysbp_ave	Num	8

4. Next, create and assign formats to the variables `RIAGENDR`, `RIDRETH1`, `DMDEDUC2 and RIDEXPRG` in your **BPdata** dataset again using the codebook that you created in Homework 2.

    a. Copy the code that you used to create and assign these formats into your homework.

```
libname nhanes 'C:\Users\smecham\Desktop\nhanes';
options fmtsearch= (nhanes);

proc format library=nhanes;
        value sex
        1 = 'Male'
        2 = 'Female'
        . = 'Missing';
        value race
        1 = 'Mexican American'
        2 = 'Other Hispanic'
        3 = 'Non-Hispanic White'
        4 = 'Non-Hispanic Black'
        5 = 'Other Race - Including Multi-Racial'
        . = 'Missing';

        value education
        1 = 'Less than 9th grade'
        2 = '9-11th grade (includes 12th grade with no diploma)'
        3 = 'High school graduate/GED or equivalent'
        4 = 'Some college or AA degree'
        5 = 'College graduate or above'
        7 = 'Refused'
        9 = 'Don’t know'
        . = 'Missing';

        value pregnancystatus
        1 = 'Yes, positive lab pregnancy test or self-reported pregnant at exam'
        2 = 'The participant was not pregnant at exam'
        3 = 'Cannot ascertain if the participant is pregnant at exam'
        . = 'Missing';

        run;


data nhanes.BPdata;
        set nhanes.BPdata;
        format sex sex. race race. education education. pregnancystatus pregnancystatus.;
        run;
```

b. If you saved these permanently, what SAS code would you use to direct SAS to your permanent format library in future SAS sessions?

```
libname nhanes 'C:\Users\smecham\Desktop\nhanes';
options fmtsearch= (nhanes);
```

c. What is the purpose of formatting a variable like RIDRETH1?

Because it can be confusing when there are a lot of coded options (RIDRETH1 had six possible options, for example), so it is easier to format the variable into the options that they actually represent.

d. Run a PROC FREQ on RIDRETH1 in the **BPdata** dataset. Does SAS display the raw or formatted values in the results viewer?

Displayed the formatted values in the results viewer.

e. Does formatting a variable alter how the data is stored in SAS? For example, if you wanted to select a certain race in your code, would you ask SAS to subset based on the original values or the formats assigned to those values?

Formatting variables just changes how SAS shows us, the viewers, the information. It does not change how the data is stored within SAS, so we would ask SAS to subset based on the original values.

5. Create a new permanent dataset saved in your **nhanes** directory called **sample** from your labeled and formatted **BPdata** dataset that contains those individuals who are aged **18 years or more**. Exclude individuals who have been classified as **pregnant or pregnancy status**

could not be determined, keeping those who have missing information on pregnancy status. Also exclude those with **missing** data on hypertension status. This subset will represent your analytic sample.

a. How many individuals and variables are in this file?

5708 individuals and 11 variables.

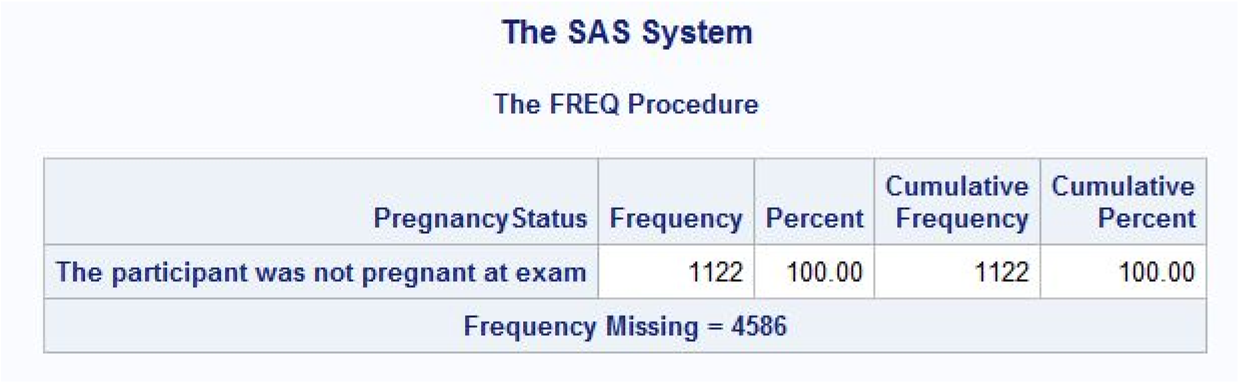
b. What would have been the gender distribution of the file had we excluded all of those participants with missing information on pregnancy status?

All participants would be female because all the males have 'missing' for pregnancy status.

c. Use PROC MEANS to check your age restriction. Report the minimum, mean, and maximum age for all participants in the **sample** file.

Minimum age is 18 years, mean age is 47.977 years, and the maximum age is 80 years.

d. Now run a PROC FREQ to check your pregnancy restriction among women. Paste in your results that confirm that there are only those women who are not pregnant (i.e., no positive pregnancy tests or unknown pregnancy status) in the dataset.



The SAS System

The FREQ Procedure

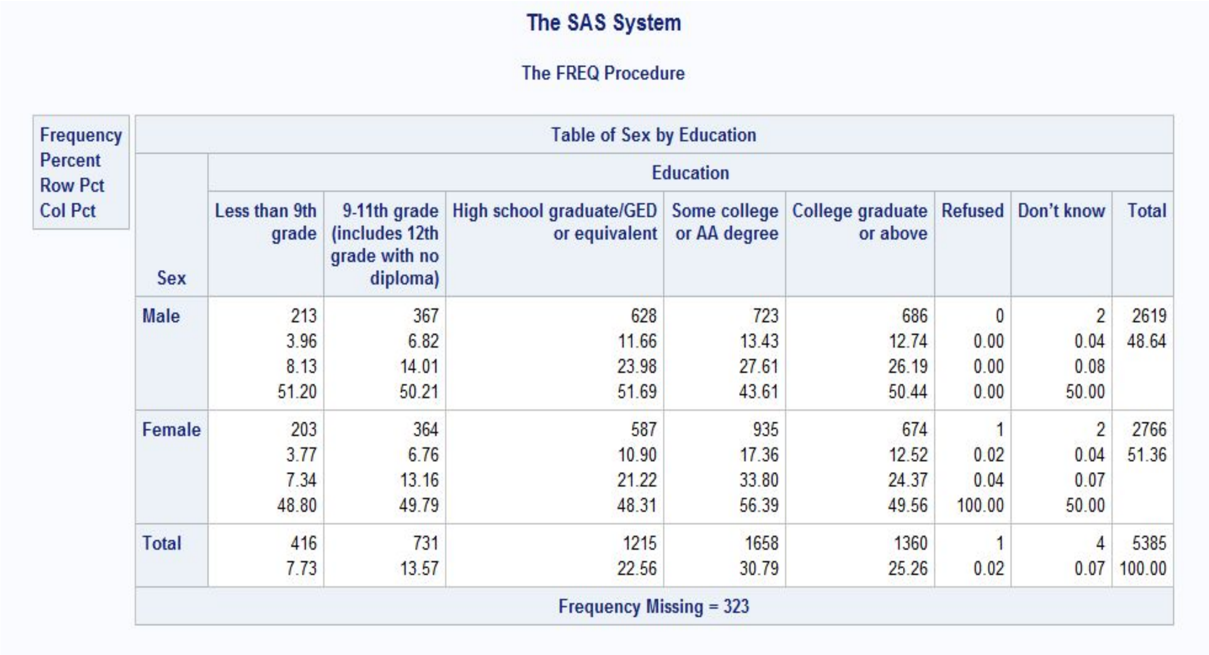
Pregnancy Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
The participant was not pregnant at exam	1122	100.00	1122	100.00
Frequency Missing = 4586				

e. If you still wanted only those individuals over age 18 and non-pregnant but did not want to include the RIDEXPRG variable in your **sample** dataset, where in your code would you drop the pregnancy variable (i.e., in the SET or DATA statement)? Why?

You would exclude the pregnancy variable in the data statement, because this will allow SAS to still sift out individuals by pregnancy status (just not showing us individuals who are pregnant). If we excluded the pregnancy variable in the set statement, SAS would not consider pregnancy as a factor at all.

6. Run a cross-tabulation (frequency) of sex by education within the **sample** dataset, requesting the Chi-Squared statistic.

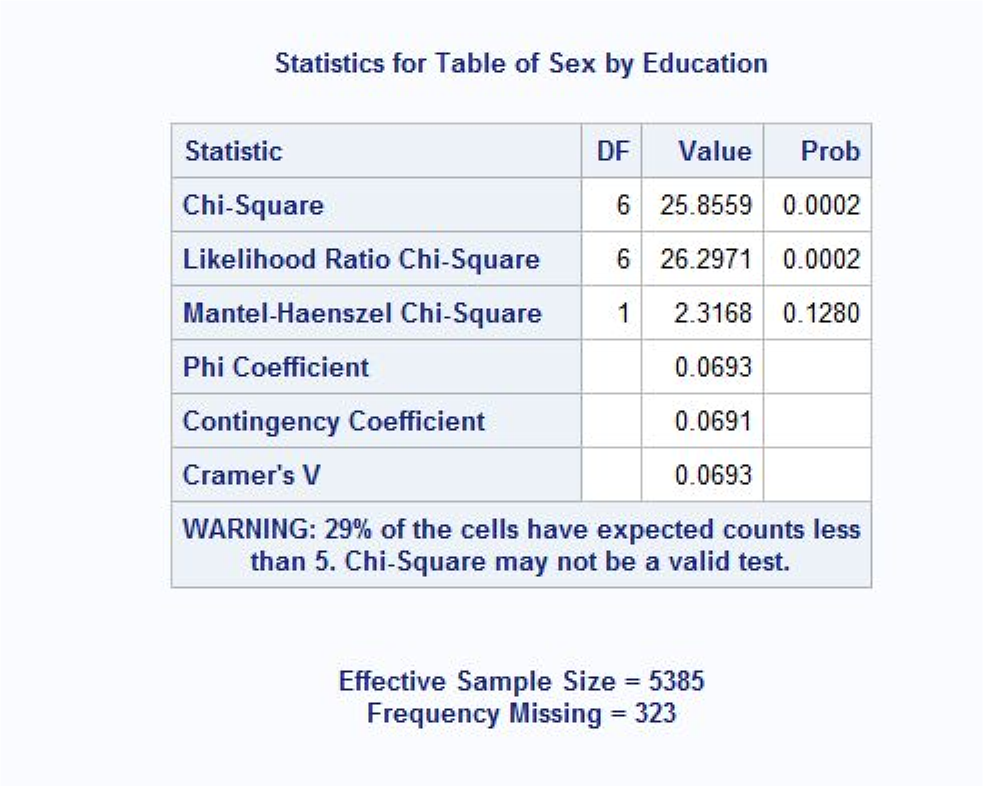
   a. Copy the cross-tabulation table into your homework:



The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Sex by Education							
	Education							
Sex	Less than 9th grade	9-11th grade (includes 12th grade with no diploma)	High school graduate/GED or equivalent	Some college or AA degree	College graduate or above	Refused	Don't know	Total
Male	213 3.96 8.13 51.20	367 6.82 14.01 50.21	628 11.66 23.98 51.69	723 13.43 27.61 43.61	686 12.74 26.19 50.44	0 0.00 0.00 0.00	2 0.04 0.08 50.00	2619 48.64
Female	203 3.77 7.34 48.80	364 6.76 13.16 49.79	587 10.90 21.22 48.31	935 17.36 33.80 56.39	674 12.52 24.37 49.56	1 0.02 0.04 100.00	2 0.04 0.07 50.00	2766 51.36
Total	416 7.73	731 13.57	1215 22.56	1658 30.79	1360 25.26	1 0.02	4 0.07	5385 100.00
Frequency Missing = 323								



Statistics for Table of Sex by Education

Statistic	DF	Value	Prob
Chi-Square	6	25.8559	0.0002
Likelihood Ratio Chi-Square	6	26.2971	0.0002
Mantel-Haenszel Chi-Square	1	2.3168	0.1280
Phi Coefficient		0.0693	
Contingency Coefficient		0.0691	
Cramer's V		0.0693	
WARNING: 29% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Effective Sample Size = 5385
Frequency Missing = 323

b. Report the fraction of the individuals in the **sample** dataset whose highest educational attainment was a high school degree or equivalent.

1215/5385

c. Report the percent of women who were college graduates or above.

24.37%

d. If the p-value of the chi-squared test is <0.05, then you can conclude that there is a relationship between sex and education that is more than expected due to chance alone (i.e., there is a statistical association). Do you find evidence of a statistically significant relationship between education and sex?

Yes, because the p-value of the chi-square is 0.0002, which is less than the significance value of <0.05 standard for concluding a relationship due to more than chance alone.

7. Run a PROC UNIVARIATE on the **sample** dataset to look at the distribution of `sysbp_ave` and `diabp_ave` in the cohort. Request a histogram with test-statistics for normality and a QQ plot on each variable.

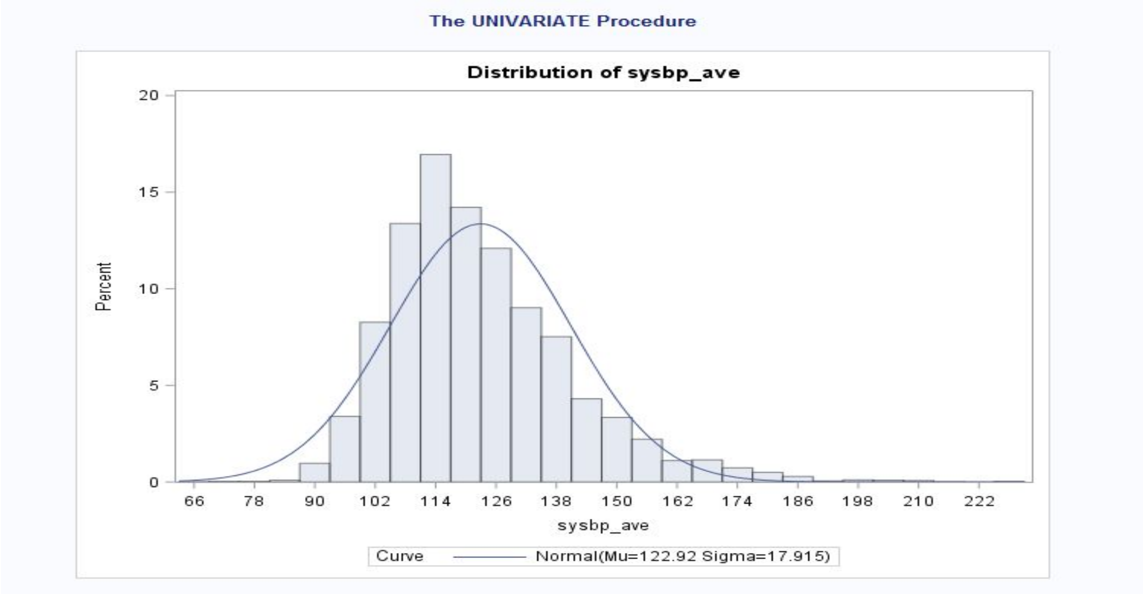
a. What was the minimum of systolic and diastolic pressure in this cohort?

64.6667 for systolic and 31.3333 for diastolic

b. What was the maximum of systolic and diastolic pressure in this cohort?

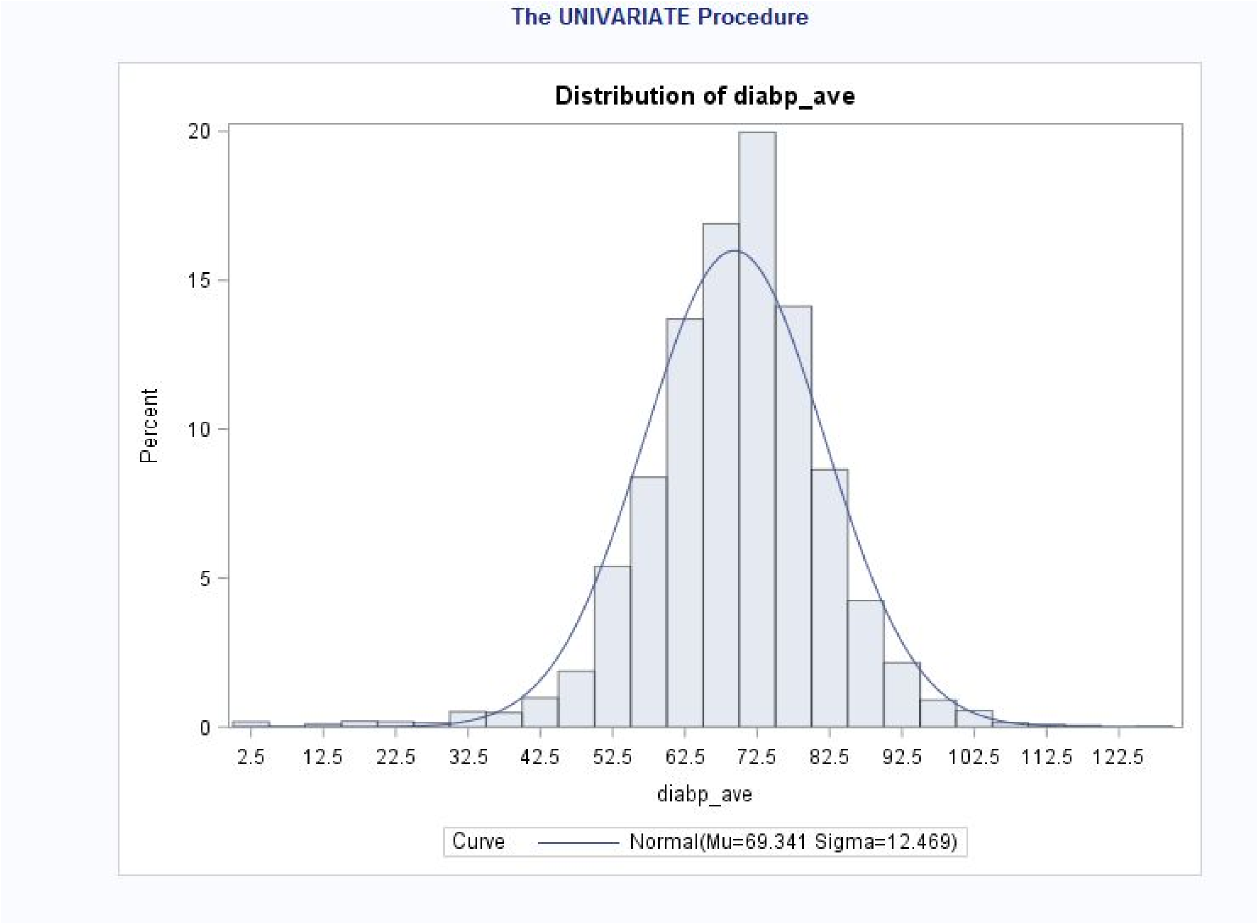
228.667 for systolic and 128 for diastolic

c. Paste your histograms of the distributions of systolic and diastolic pressure into your homework. Briefly describe the distributions in terms of their normality.



The UNIVARIATE Procedure

The distribution for systolic blood pressure among this cohort is relatively normal, but slightly positively skewed (left-modal).



The UNIVARIATE Procedure

The distribution for diastolic blood pressure in this cohort is relatively normal, but slightly negatively-skewed (right modal).

d. Given that the null hypothesis is that the distribution is normally distributed, a p-value <0.05 would indicate that the distribution is not consistent with a normal distribution. Report the p-value for the test statistic for the Kolmogorov-Smirnov test of normality for each distribution.

The Kolmogorov-Smirnov p-value for systolic is <0.010 and the Kolmogorov-Smirnov p-value for diastolic is also <0.0100, which signifies distribution not consistent with a normal distribution.

e. Based on the histograms, QQ plots, Kolmogorov-Smirnov test results, and any other information from PROC UNIVARIATE, conclude if the distributions are reasonably normal.

The QQ plots for systolic and diastolic blood pressure also signify skewness of the data. Along with the asymmetrical tails of the histograms and the <0.05 Kolmogorov-Smirnov p-values, we can conclude the distributions are skewed away from a normal distribution.

8. Write SAS code to identify the 98th percentile of systolic pressure. To do this, write a PROC UNIVARIATE step examining your systolic pressure variable that includes the following statement (as its own line of code within the procedure):

   ```
   output out=sys_pct pctlpre=P_ pctlpts= 98;
   ```

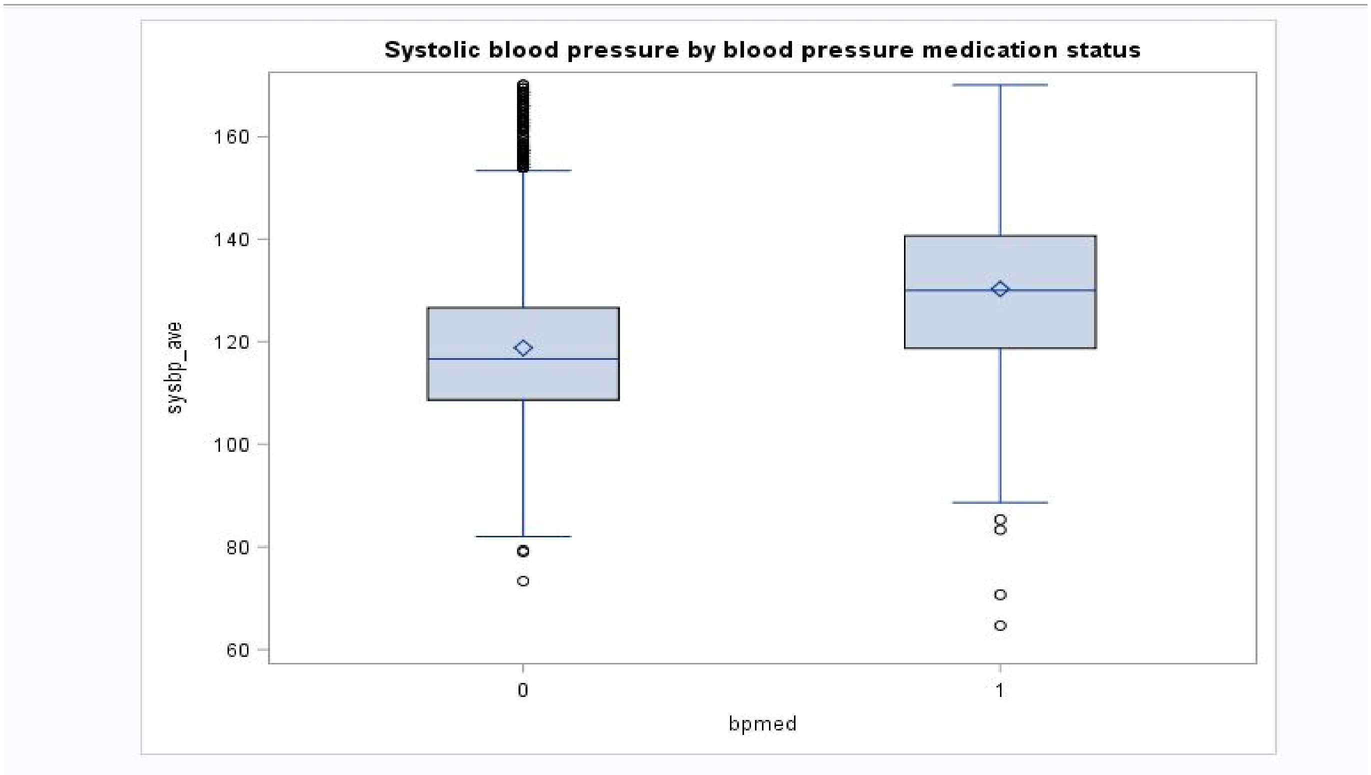
   This piece of code will create a new dataset called sys_pct with 1 observation that identifies the value of the 98th percentile of the systolic pressure variable. Open the dataset or use PROC PRINT to view the observation. What is the 98th percentile of systolic pressure?

   170 mmHg

9. Create vertical box plots to explore the distribution of systolic pressure by whether a respondent is currently using a blood pressure medication (`bpmed`). Due to large outliers, first subset the dataset to only contain individuals with systolic values less than or equal to the 98th percentile of systolic pressure. Present the plot and write a sentence or two about your findings.

Epidemiology 640 – Fall 2017 NAME: Steph Mecham

SECTION: 4



People currently taking prescription medication for high blood pressure have, on average, higher levels of systolic blood pressure than people within normal blood pressure range. There is also a little bit of a greater range in values among people taking blood pressure medications than people not taking blood pressure medication (which makes sense because some people might be responding to the medications better than others and therefore there would be a pretty big variation in blood pressure for that group).

CODE

```
*Homework 3: Working with Data Sets - Stephanie Mecham | EPID 640 Section 4;


*Creating a permanent library;

LIBNAME nhanes "C:\Users\smecham\Desktop\nhanes";

RUN;


*Question 1: Create a new dataset with select variables;

data nhanes.BPdata;
	set nhanes.demo_bp;
	keep SEQN RIDAGEYR RIAGENDR RIDRETH1 INDFMPIR DMDEDUC2 RIDEXPRG
sysbp_ave diabp_ave bpmed htn;
run;


*Question 2: Running a PROC CONTENTS step;

proc contents data=nhanes.BPdata;
run;

*Question 3: Renaming variables;

DATA nhanes.BPdata (RENAME=(SEQN = IDNumber
RIDAGEYR = Age
RIAGENDR = Sex
RIDRETH1 = Race
DMDEDUC2 = Education
RIDEXPRG = PregnancyStatus));
SET nhanes.BPdata;
RUN;

proc contents data=nhanes.BPdata;
run;
```

```
*Question 4: Applying Formats;

libname nhanes 'C:\Users\smecham\Desktop\nhanes';
options fmtsearch= (nhanes);

proc format library=nhanes;
        value sex
        1 = 'Male'
        2 = 'Female'
        . = 'Missing';
        value race
        1 = 'Mexican American'
        2 = 'Other Hispanic'
        3 = 'Non-Hispanic White'
        4 = 'Non-Hispanic Black'
        5 = 'Other Race - Including Multi-Racial'
        . = 'Missing';

        value education
        1 = 'Less than 9th grade'
        2 = '9-11th grade (includes 12th grade with no diploma)'
        3 = 'High school graduate/GED or equivalent'
        4 = 'Some college or AA degree'
        5 = 'College graduate or above'
        7 = 'Refused'
        9 = 'Don’t know'
        . = 'Missing';

        value pregnancystatus
        1 = 'Yes, positive lab pregnancy test or self-reported pregnant at exam'
        2 = 'The participant was not pregnant at exam'
        3 = 'Cannot ascertain if the participant is pregnant at exam'
        . = 'Missing';

        run;


data nhanes.BPdata;
        set nhanes.BPdata;
        format sex sex. race race. education education. pregnancystatus pregnancystatus.;
        run;
```

```
*PROC FREQ step;

proc freq data=nhanes.BPdata;
        tables race;
        run;

*Question 5: Creating a new sample dataset;

data nhanes.sample;
        set nhanes.BPdata;
        IF age < 18 THEN DELETE;
        IF pregnancystatus = 1 THEN DELETE;
        IF pregnancystatus = 3 THEN DELETE;
        IF htn = . THEN DELETE;
        run;


*PROC MEANS step;

proc means data=nhanes.sample;
run;

*PROC FREQ step;

proc freq data=nhanes.sample;
        tables pregnancystatus;
        run;

*Question 6: Multi-level cross-tabulation and Chi Square;

proc freq data=nhanes.sample;
table sex*education / chisq;
run;

*Question 7: PROC UNIVARIATE step;

proc univariate data=nhanes.sample normaltest;
var sysbp_ave diabp_ave;
histogram sysbp_ave diabp_ave / normal;
qqplot;
run;
```

```
*Question 8: Identifying percentiles using PROC UNIVARIATE;

proc univariate data=nhanes.sample;
var sysbp_ave;
output out=sys_pct pctlpre=P_ pctlpts= 98;
run;

*Question 9: Vertical box plots;

proc sgplot data=nhanes.sample;
vbox sysbp_ave / category=bpmed;
title 'Systolic blood pressure by blood pressure medication status';
where sysbp_ave<=170;
run;
```