Handout 1 SAS: Introduction to data

Getting Started

Load the data set of 20,000 observations

In [1]:
```
filename cdc url 'http://www.openintro.org/stat/data/cdc.csv';

proc import datafile=cdc
            out=work.cdc
            dbms=csv
            replace;
            getnames=yes;
run;
```

```
SAS Connection established. Subprocess id is 18327
```

Out[1]:

```
34   ods listing close;ods html5 (id=saspy_internal) file=stdout options(bitmap_mode='inline') device=svg; ods graphics on /
34 ! outputfmt=png;
NOTE: Writing HTML5(SASPY_INTERNAL) Body file: STDOUT
35
36   filename cdc url 'http://www.openintro.org/stat/data/cdc.csv';
37
38   proc import datafile=cdc
39               out=work.cdc
40               dbms=csv
41               replace;
42               getnames=yes;
43   run;
NOTE: Unable to open SASUSER.PROFILE. WORK.PROFILE will be opened instead.
NOTE: All profile changes will be lost at the end of the session.
44    /**********************************
**********************************
45    *   PRODUCT:   SAS
46    *   VERSION:   9.4
47    *   CREATOR:   External File Interface
48    *   DATE:      07JUL18
49    *   DESC:      Generated SAS Datastep Code
50    *   TEMPLATE SOURCE:  (None Specifie
```

```
d.)
51    ************************************
**********************************/
52       data WORK.CDC    ;
53       %let _EFIERR_ = 0; /* set the ERRO
R detection macro variable */
54       infile CDC delimiter = ',' MISSOVE
R DSD lrecl=32767 firstobs=2 ;
55          informat genhlth $9. ;
56          informat exerany best32. ;
57          informat hlthplan best32. ;
58          informat smoke100 best32. ;
59          informat height best32. ;
60          informat weight best32. ;
61          informat wtdesire best32. ;
62          informat age best32. ;
63          informat gender $1. ;
64          format genhlth $9. ;
65          format exerany best12. ;
66          format hlthplan best12. ;
67          format smoke100 best12. ;
68          format height best12. ;
69          format weight best12. ;
70          format wtdesire best12. ;
71          format age best12. ;
72          format gender $1. ;
73       input
74                   genhlth  $
75                   exerany
76                   hlthplan
77                   smoke100
78                   height
79                   weight
80                   wtdesire
81                   age
82                   gender  $
83       ;
84       if _ERROR_ then call symputx('_EFI
ERR_',1);  /* set ERROR detection macro var
iable */
85       run;
NOTE: The infile CDC is:
      Filename=http://www.openintro.org/sta
t/data/cdc.csv,
      Local Host Name=localhost.localdomai
n,
      Local Host IP addr=::1,
      Service Hostname Name=www.openintro.o
rg,
      Service IP addr=192.185.65.127,
      Service Name=httpd,Service Portno=80,
      Lrecl=32767,Recfm=Variable

NOTE: 20000 records were read from the infi
le CDC.
      The minimum record length was 24.
      The maximum record length was 31.
```

```
NOTE: The data set WORK.CDC has 20000 obser
vations and 9 variables.
NOTE: DATA statement used (Total process ti
me):
      real time           1.53 seconds
      cpu time            0.09 seconds

20000 rows created in WORK.CDC from CDC.



NOTE: WORK.CDC data set was successfully cr
eated.
NOTE: The data set WORK.CDC has 20000 obser
vations and 9 variables.
NOTE: PROCEDURE IMPORT used (Total process
 time):
      real time           2.86 seconds
      cpu time            0.20 seconds

86
87   ods html5 (id=saspy_internal) close;od
s listing;

88
```

View the names of the variables

In [2]:
```
proc contents data=work.cdc short;
run;
```

Out[2]:

The SAS System

The CONTENTS Procedure

Alphabetic List of Variables for WORK.CDC
age exerany gender genhlth height hlthplan smoke100 weight wtdesire

Exercise 1

How many cases are there in this data set? How many variables? For each variable, identify its data type (for example, categorical, numeric).

There are 9 variables.

Data Source: US CDC website.

- genhlth: A categorical vector indicating general health, with categories excellent, very good, good, fair, and poor.
- exerany: A categorical vector, 1 if the respondent exercised in the past month and 0 otherwise.
- hlthplan: A categorical vector, 1 if the respondent has some form of health coverage and 0 otherwise.
- smoke100: A categorical vector, 1 if the respondent has smoked at least 100 cigarettes in their entire life and 0 otherwise.
- height: A numerical vector, respondent's height in inches.
- weight: A numerical vector, respondent's weight in pounds.
- wtdesire: A numerical vector, respondent's desired weight in pounds.
- age: A numerical vector, respondent's age in years.

Look at the first 10 rows of our data

In [3]:
```
proc print data=work.cdc (obs=10);
run;
```

Out[3]:

The SAS System

Obs	genhlth	exerany	hlthplan	smoke100
1	good	0	1	0
2	good	0	1	1
3	good	1	1	1
4	good	1	1	0
5	very good	0	1	0
6	very good	1	1	0
7	very good	1	1	0
8	very good	0	1	0
9	good	0	1	1
10	good	1	1	0

Summaries and Tables

Look at numerical summary, such as mean, variance, standard deviation, minimum, maximum, and extreme observations

Summary statistic for weight

In [4]:
```
proc univariate data=work.cdc;
   var weight;
run;
```

Out[4]:

The SAS System

The UNIVARIATE Procedure
Variable: weight

Moments			
N	20000	Sum Weights	200(
Mean	169.68295	Sum Observations	3393
Std Deviation	40.08097	Variance	160(
Skewness	0.95572799	Kurtosis	1.99
Uncorrected SS	607974147	Corrected SS	3212
Coeff Variation	23.6210945	Std Error Mean	0.28

Basic Statistical Measures			
Location		Variability	
Mean	169.6830	Std Deviation	40.08097
Median	165.0000	Variance	1606
Mode	160.0000	Range	432.00000

Basic Statistical Measures			
Location		**Variability**	
		Interquartile Range	50.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	598.7079	**Pr > |t|**	<.0001
Sign	**M**	10000	**Pr >= |M|**	<.0001
Signed Rank	**S**	1.0001E8	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	500
99%	290
95%	240
90%	220
75% Q3	190
50% Median	165
25% Q1	140
10%	124
5%	115
1%	100
0% Min	68

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
68	18743	400	2944
70	16531	400	19319
78	18065	405	15720
78	11299	495	4445
79	7614	500	1995

The sample frequency distribution for smoke100

In [5]:
```
proc freq data=work.cdc;
    tables Smoke100;
run;
```

Out[5]:

The SAS System

The FREQ Procedure

smoke100	Frequency	Percent	Cumulative Frequency
0	10559	52.80	10559
1	9441	47.21	20000

Graph the sample frequency distribution for smoke100

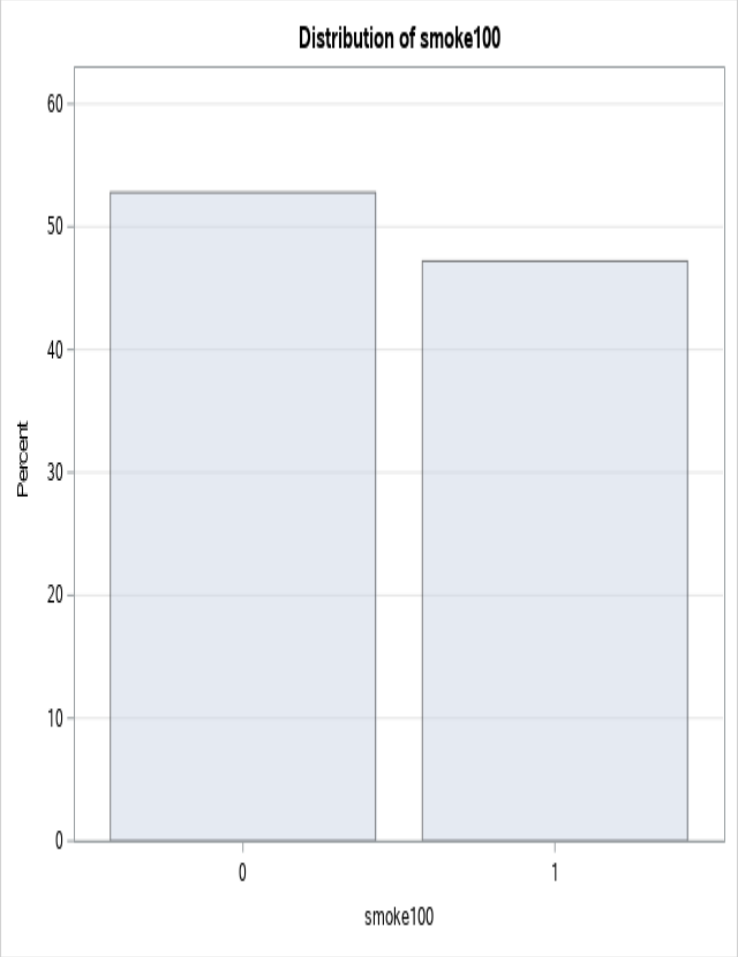
In [6]:
```
proc freq data=work.cdc;
    tables Smoke100 / plots=freqplot(scale=pe
rcent);
run;
```

Out[6]:

The SAS System

The FREQ Procedure

smoke100	Frequency	Percent	Cumulative Frequency
0	10559	52.80	10559
1	9441	47.21	20000



Exercise 2

Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

There are 9569 males in the sample. A total of 4657 out out of 20000 reported being in excellent health. This is 23.29% of the sample.

Numerical summary for height and age

In [7]:
```
proc univariate data=work.cdc;
   var height;
   var age;
run;
```

Out[7]:

The SAS System

The UNIVARIATE Procedure
Variable: height

Moments			
N	20000	Sum Weights	2000
Mean	67.1829	Sum Observations	1343
Std Deviation	4.12595429	Variance	17.0
Skewness	0.1036124	Kurtosis	-0.37
Uncorrected SS	90611294	Corrected SS	3404
Coeff Variation	6.14137569	Std Error Mean	0.02

Basic Statistical Measures			
Location		Variability	
Mean	67.18290	Std Deviation	4.12595
Median	67.00000	Variance	17.02350
Mode	66.00000	Range	45.00000
		Interquartile Range	6.00000

Tests for Location: Mu0=0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2302.763	Pr > |t|	<.0001
Sign	M	10000	Pr >= |M|	<.0001
Signed Rank	S	1.0001E8	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	93
99%	76
95%	74
90%	73
75% Q3	70
50% Median	67
25% Q1	64
10%	62
5%	61
1%	59
0% Min	48

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
48	15465	82	10160
48	5412	82	10322
49	8871	83	3691
50	3905	84	18817
51	11948	93	17534

The SAS System

The UNIVARIATE Procedure
Variable: age

Moments			
N	20000	Sum Weights	200(
Mean	45.06825	Sum Observations	9013
Std Deviation	17.1926895	Variance	295.
Skewness	0.45170032	Kurtosis	-0.6
Uncorrected SS	46534419	Corrected SS	5911
Coeff Variation	38.1481186	Std Error Mean	0.12

Basic Statistical Measures	
Location	Variability

Basic Statistical Measures			
Location		**Variability**	
Mean	45.06825	**Std Deviation**	17.19269
Median	43.00000	**Variance**	295.58857
Mode	40.00000	**Range**	81.00000
		Interquartile Range	26.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	370.7165	**Pr > |t|**	<.0001
Sign	**M**	10000	**Pr >= |M|**	<.0001
Signed Rank	**S**	1.0001E8	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	99
99%	84
95%	76
90%	71
75% Q3	57
50% Median	43
25% Q1	31

Quantiles (Definition 5)	
Level	**Quantile**
10%	24
5%	21
1%	18
0% Min	18

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
18	19885	95	16084
18	19860	96	17051
18	19832	97	10350
18	19706	99	900
18	19622	99	6710

Compute the interquartile range for height and age

Interquartile range for height: between 48 and 93 with a range of 45 Interquartile range for age: between 18 and 99 with a range of 81

Relative frequency distribution for gender

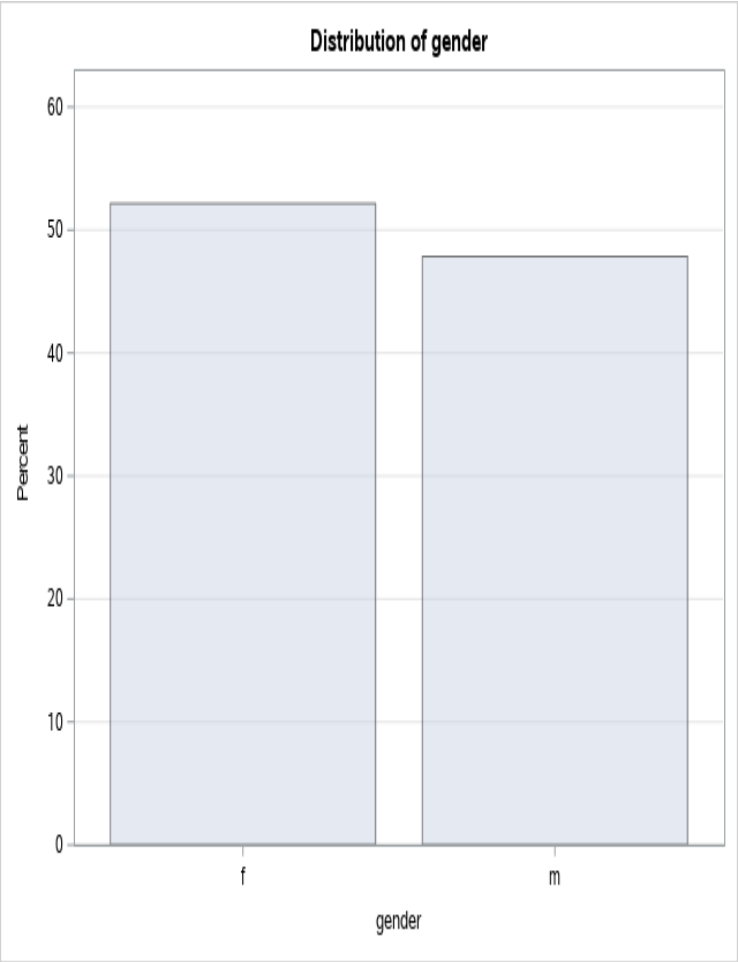
In [8]:
```
proc freq data=work.cdc;
    tables gender / plots=freqplot(scale=perc
ent);
run;
```

Out[8]:

The SAS System

The FREQ Procedure

gender	Frequency	Percent	Cumulative Frequency	Cu Pe
f	10431	52.16	10431	52
m	9569	47.85	20000	10



Relative frequency distribution for exerany

In [9]:
```
proc freq data=work.cdc;
    tables exerany / plots=freqplot(scale=per
cent);
run;
```

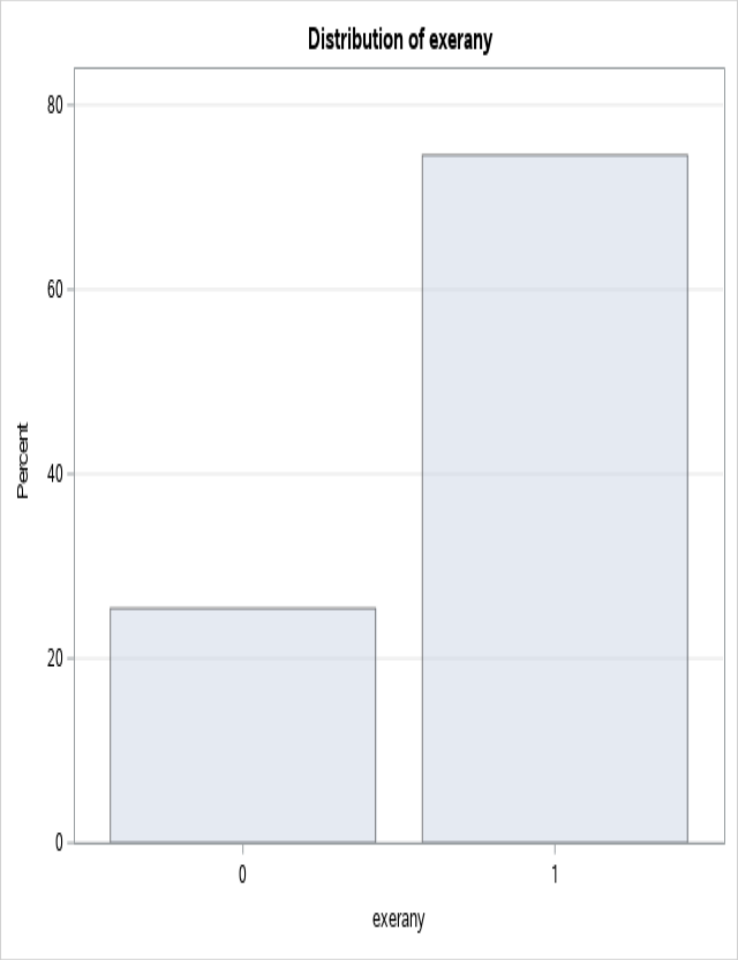
Out[9]:

The SAS System

The FREQ Procedure

exerany	Frequency	Percent	Cumulative Frequency	C P

exerany	Frequency	Percent	Cumulative Frequency	C P
0	5086	25.43	5086	2
1	14914	74.57	20000	1



Relative frequency distribution for smoke100

In [10]:
```
proc freq data=work.cdc;
    tables smoke100 / plots=freqplot(scale=pe
rcent);
run;
```

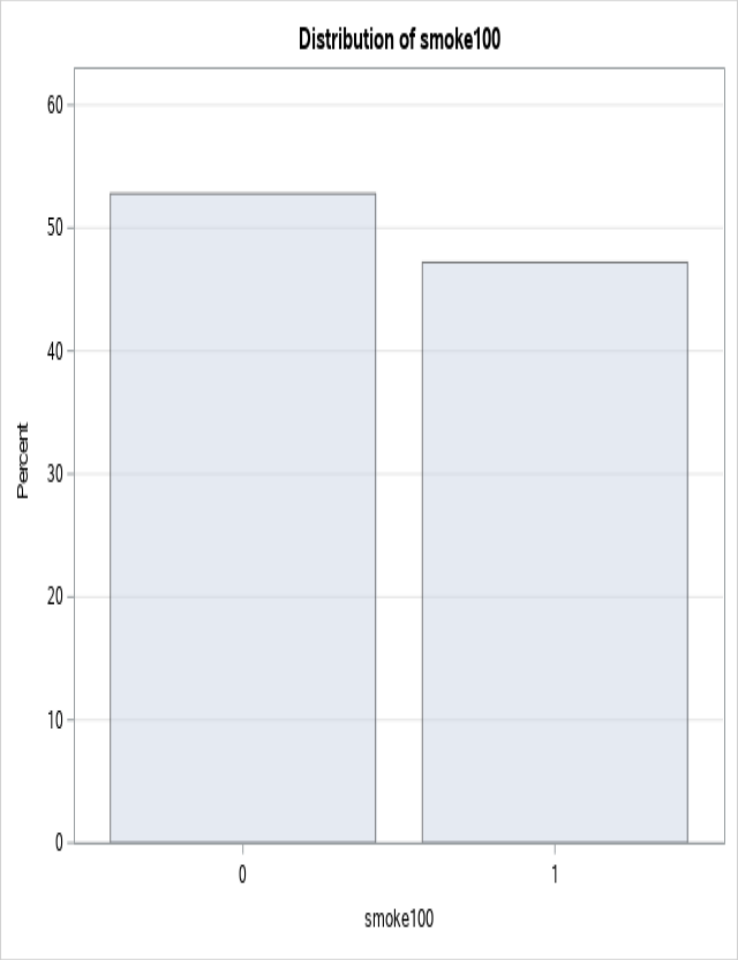
Out[10]:

The SAS System

The FREQ Procedure

smoke100	Frequency	Percent	Cumulative Frequency

smoke100	Frequency	Percent	Cumulative Frequency
0	10559	52.80	10559
1	9441	47.21	20000



Relative frequency distribution for genhlth

In [11]:
```
proc freq data=work.cdc;
    tables genhlth / plots=freqplot(scale=per
cent);
run;
```

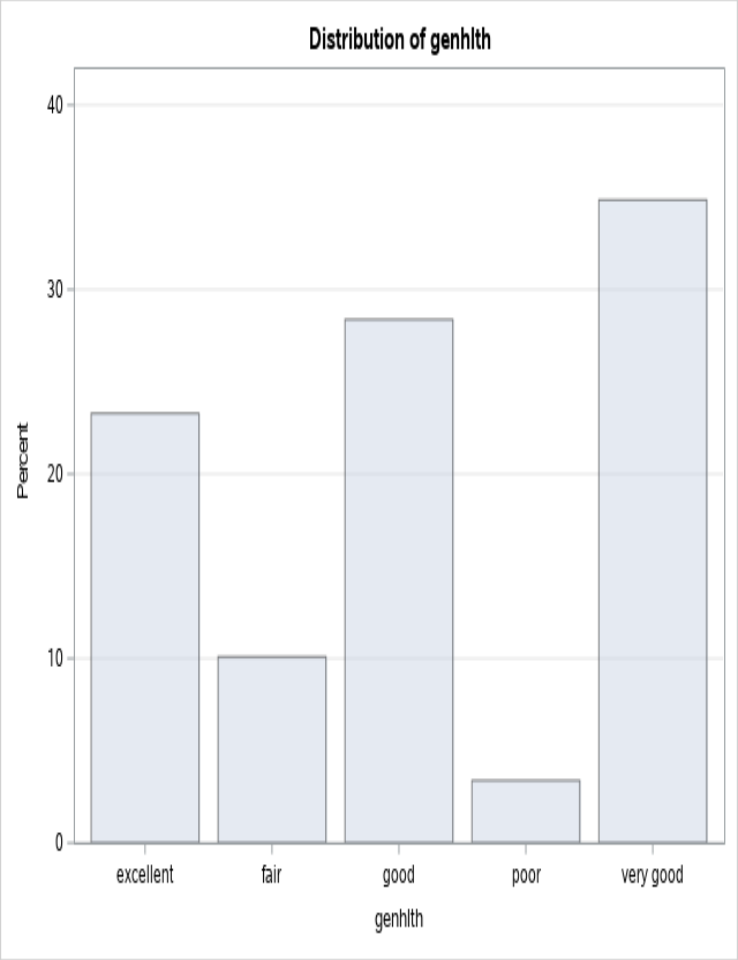
Out[11]:

The SAS System

The FREQ Procedure

genhlth	Frequency	Percent	Cumulative Frequency

genhlth	Frequency	Percent	Cumulative Frequency	C
excellent	4657	23.29	4657	2
fair	2019	10.10	6676	1
good	5675	28.38	12351	6
poor	677	3.39	13028	6
very good	6972	34.86	20000	



Create multi-way frequency tables: gender and smoke100

In [12]:
```
proc freq data=work.cdc;
   tables gender*smoke100;
run;
```

Out[12]: **The SAS System**

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of gender by smoke100		
	smoke100		
gender	0	1	Total
f	6012 30.06 57.64 56.94	4419 22.10 42.36 46.81	10431 52.16
m	4547 22.74 47.52 43.06	5022 25.11 52.48 53.19	9569 47.85
Total	10559 52.80	9441 47.21	20000 100.00

Create a mosaic plot

In [13]:
```
proc freq data=work.cdc;
    tables Gender*Smoke100 / plots=mosaicplo
t;
run;
```

Out[13]:

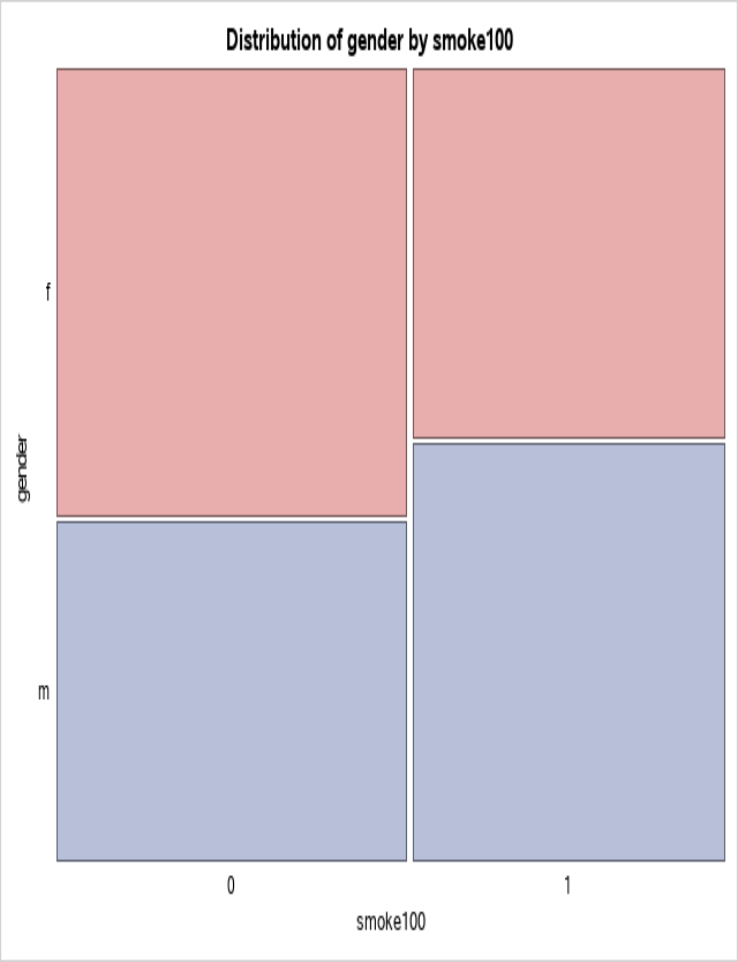
The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of gender by smoke100		
	smoke100		
gender	0	1	Total
f	6012 30.06 57.64 56.94	4419 22.10 42.36 46.81	10431 52.16

Table of gender by smoke100

gender	smoke100 0	smoke100 1	Total
m	4547 22.74 47.52 43.06	5022 25.11 52.48 53.19	9569 47.85
Total	10559 52.80	9441 47.21	20000 100.00



Exercise 3

What does the mosaic plot reveal about smoking habits and gender?

The percentage of males that smoked at least 100 cigarettes in their entire life is larger than percentage of females

Interlude: How SAS Processes Data

See the first 10 values in the data portion

In [14]:
```
proc print data=work.cdc (obs=10);
run;
```

Out[14]: The SAS System

Obs	genhlth	exerany	hlthplan	smoke100
1	good	0	1	0
2	good	0	1	1
3	good	1	1	1
4	good	1	1	0
5	very good	0	1	0
6	very good	1	1	0
7	very good	1	1	0
8	very good	0	1	0
9	good	0	1	1
10	good	1	1	0

See the descriptor portion such as the names, types, and lengths of the variables

In [15]:
```
proc contents data=work.cdc;
run;
```

Out[15]: The SAS System

The CONTENTS Procedure

Data Set Name	WORK.CDC	Observati
Member Type	DATA	Variables
Engine	V9	Indexes
Created	07/07/2018 06:06:24	Observati Length
Last Modified	07/07/2018 06:06:24	Deleted Observati
Protection		Compress
Data Set Type		Sorted
Label		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64	
Encoding	utf-8 Unicode (UTF-8)	

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	23
First Data Page	1
Max Obs per Page	908

Engine/Host Dependent Information	
Obs in First Data Page	867
Number of Data Set Repairs	0
Filename	/tmp/SAS_work89DD00004797_loca
Release Created	9.0401M5
Host Created	Linux
Inode Number	409728
Access Permission	rw-r--r--
Owner Name	sasdemo
File Size	2MB
File Size (bytes)	1572864

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
8	age	Num	8	BEST12.	BEST32.
2	exerany	Num	8	BEST12.	BEST32.
9	gender	Char	1	$1.	$1.
1	genhlth	Char	9	$9.	$9.
5	height	Num	8	BEST12.	BEST32.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
3	hlthplan	Num	8	BEST12.	BEST32.
4	smoke100	Num	8	BEST12.	BEST32.
6	weight	Num	8	BEST12.	BEST32.
7	wtdesire	Num	8	BEST12.	BEST32.

Create subset of observations of people who are men or anyone over the age of 30. See the first 10 values of work.newcdc

In [16]:
```
data work.newcdc;
   set work.cdc;
   if gender="m" and age>30;
run;

proc print data=work.newcdc (obs=10);
run;
```

Out[16]:

The SAS System

Obs	genhlth	exerany	hlthplan	smoke100
1	good	0	1	0
2	very good	1	1	0
3	very good	0	1	0
4	good	1	1	0
5	excellent	1	1	1
6	fair	1	1	1
7	excellent	1	1	1
8	good	1	1	1

Obs	genhlth	exerany	hlthplan	smoke100
9	good	0	0	1
10	fair	0	1	1

Create subset of observations of people who are men or female and less the age of 30. See the first 10 values of work.newcdc

In [17]:
```
data work.newcdc;
   set work.cdc;
   if gender="m" or (gender="f" and age>30);
run;

proc print data=work.newcdc (obs=10);
run;
```

Out[17]:

The SAS System

Obs	genhlth	exerany	hlthplan	smoke100
1	good	0	1	0
2	good	0	1	1
3	good	1	1	1
4	good	1	1	0
5	very good	0	1	0
6	very good	1	1	0
7	very good	1	1	0
8	very good	0	1	0
9	good	1	1	0
10	excellent	1	1	1

Exercise 4

Create a new data set named under23smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the programming statements you used to create the new data set as the answer to this exercise.

In [18]:
```
data work.under23smoke;
   set work.cdc;
   if smoke100=1 and age<23;
run;

proc print data=work.under23smoke (obs=10);
run;
```

Out[18]:

The SAS System

Obs	genhlth	exerany	hlthplan	smoke100
1	excellent	1	0	1
2	very good	1	0	1
3	excellent	1	1	1
4	good	1	1	1
5	very good	1	1	1
6	very good	1	0	1
7	fair	0	1	1
8	fair	1	1	1
9	excellent	1	0	1
10	fair	1	1	1

Quantitative Data

Create box-and-whisker plot and a histogram for weight

In [19]:
```
ods graphics;
proc univariate data=work.cdc plots;
    var weight;
run;
```

Out[19]:

The SAS System

The UNIVARIATE Procedure
Variable: weight

Moments			
N	20000	Sum Weights	2000
Mean	169.68295	Sum Observations	3393
Std Deviation	40.08097	Variance	1606
Skewness	0.95572799	Kurtosis	1.99
Uncorrected SS	607974147	Corrected SS	3212
Coeff Variation	23.6210945	Std Error Mean	0.28

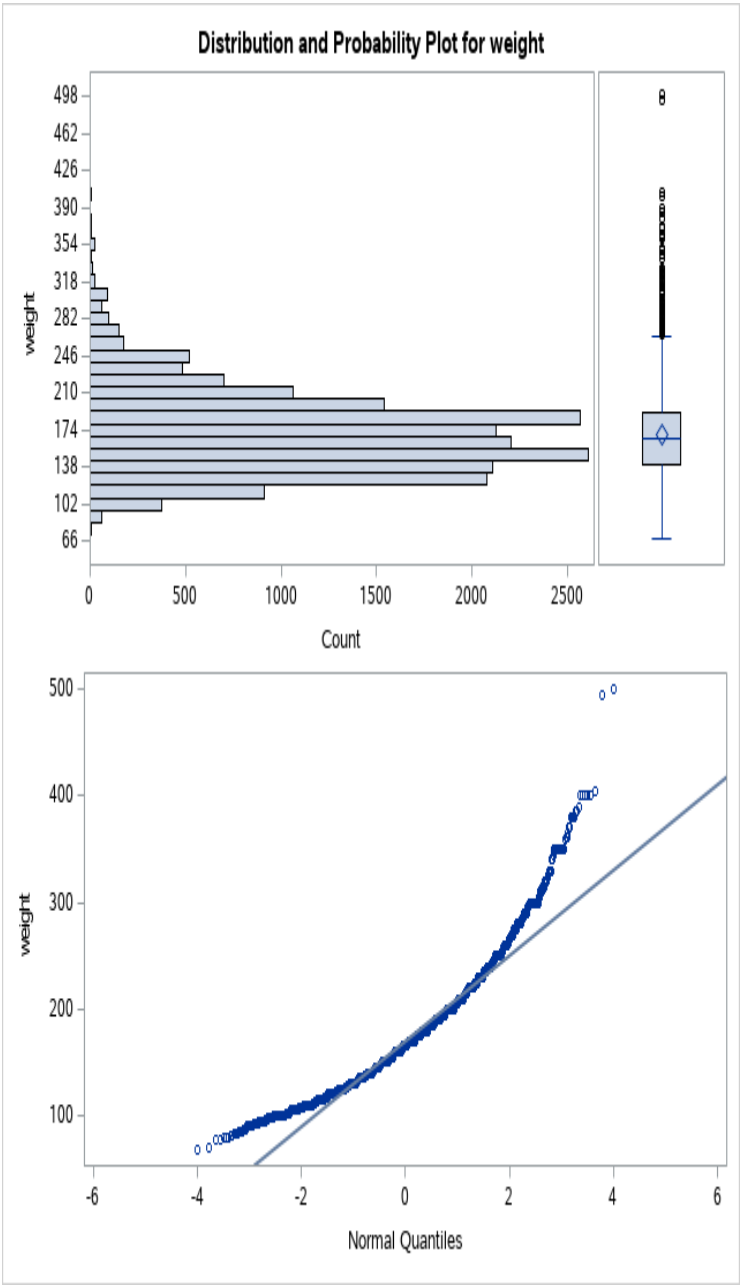
Basic Statistical Measures			
Location		Variability	
Mean	169.6830	Std Deviation	40.08097
Median	165.0000	Variance	1606
Mode	160.0000	Range	432.00000
		Interquartile Range	50.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	598.7079	Pr > |t|	<.0001
Sign	M	10000	Pr >= |M|	<.0001
Signed Rank	S	1.0001E8	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	500
99%	290
95%	240
90%	220
75% Q3	190
50% Median	165
25% Q1	140
10%	124
5%	115
1%	100
0% Min	68

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
68	18743	400	2944
70	16531	400	19319
78	18065	405	15720
78	11299	495	4445
79	7614	500	1995



Create box-and-whisker plot and a histogram for weight subset by gender using "class" statement

In [20]:
```
ods graphics;
proc univariate data=work.cdc plots;
    class gender;
    var weight;
run;
```

Out[20]:

The SAS System

The UNIVARIATE Procedure
Variable: weight
gender = f

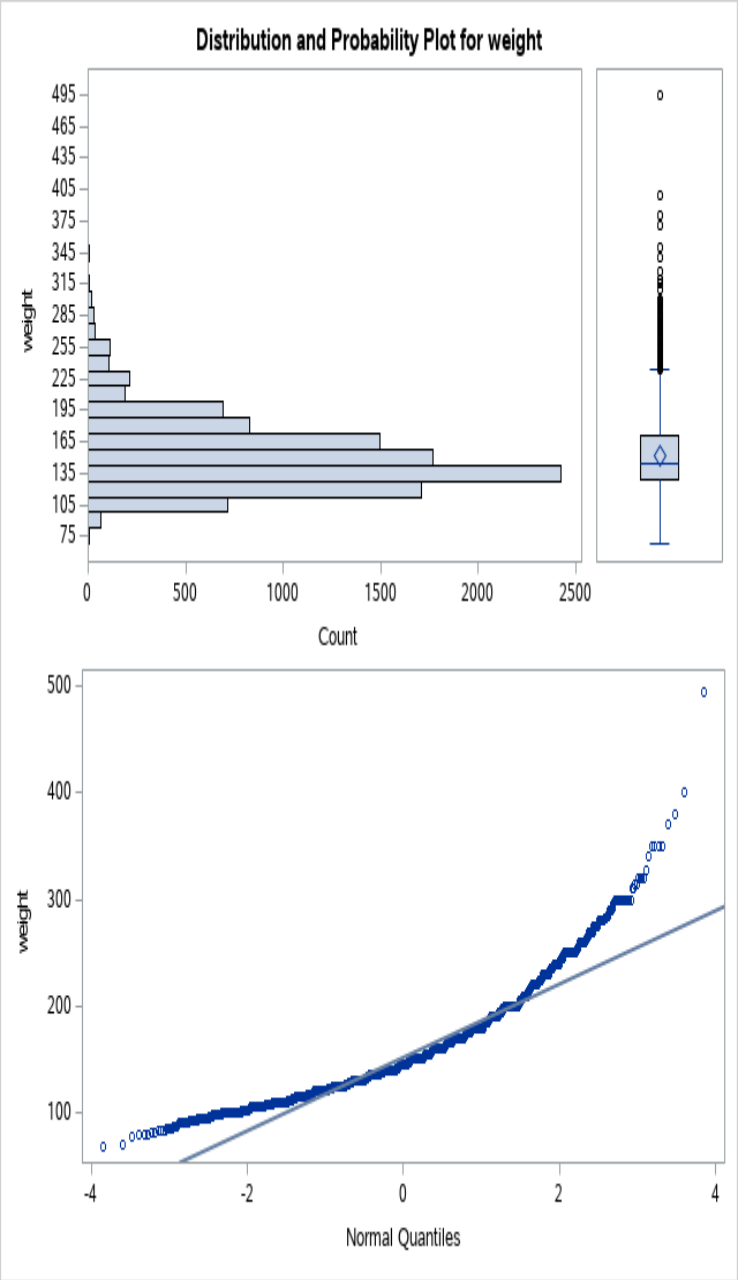
Moments			
N	10431	Sum Weights	1043
Mean	151.666187	Sum Observations	1582
Std Deviation	34.2975191	Variance	1176
Skewness	1.38530136	Kurtosis	3.77
Uncorrected SS	252209474	Corrected SS	1226
Coeff Variation	22.6138203	Std Error Mean	0.33

Basic Statistical Measures			
Location		Variability	
Mean	151.6662	Std Deviation	34.29752
Median	145.0000	Variance	1176
Mode	140.0000	Range	427.00000
		Interquartile Range	42.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	451.6365	Pr > |t|	<.0001
Sign	M	5215.5	Pr >= |M|	<.0001
Signed Rank	S	27204048	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	495
99%	260
95%	220
90%	198
75% Q3	170
50% Median	145
25% Q1	128
10%	115
5%	110
1%	100
0% Min	68

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
68	18743	350	13607
70	16531	371	4612
78	11299	380	7160
79	7614	400	19319
80	15673	495	4445



The SAS System

The UNIVARIATE Procedure
Variable: weight
gender = m

Moments			
N	9569	**Sum Weights**	9569
Mean	189.322709	**Sum Observations**	1811

Moments			
Std Deviation	36.5503551	Variance	1335
Skewness	1.16960127	Kurtosis	3.04
Uncorrected SS	355764673	Corrected SS	1278
Coeff Variation	19.3058484	Std Error Mean	0.37

Basic Statistical Measures			
Location		Variability	
Mean	189.3227	Std Deviation	36.55036
Median	185.0000	Variance	1336
Mode	180.0000	Range	422.00000
		Interquartile Range	45.00000

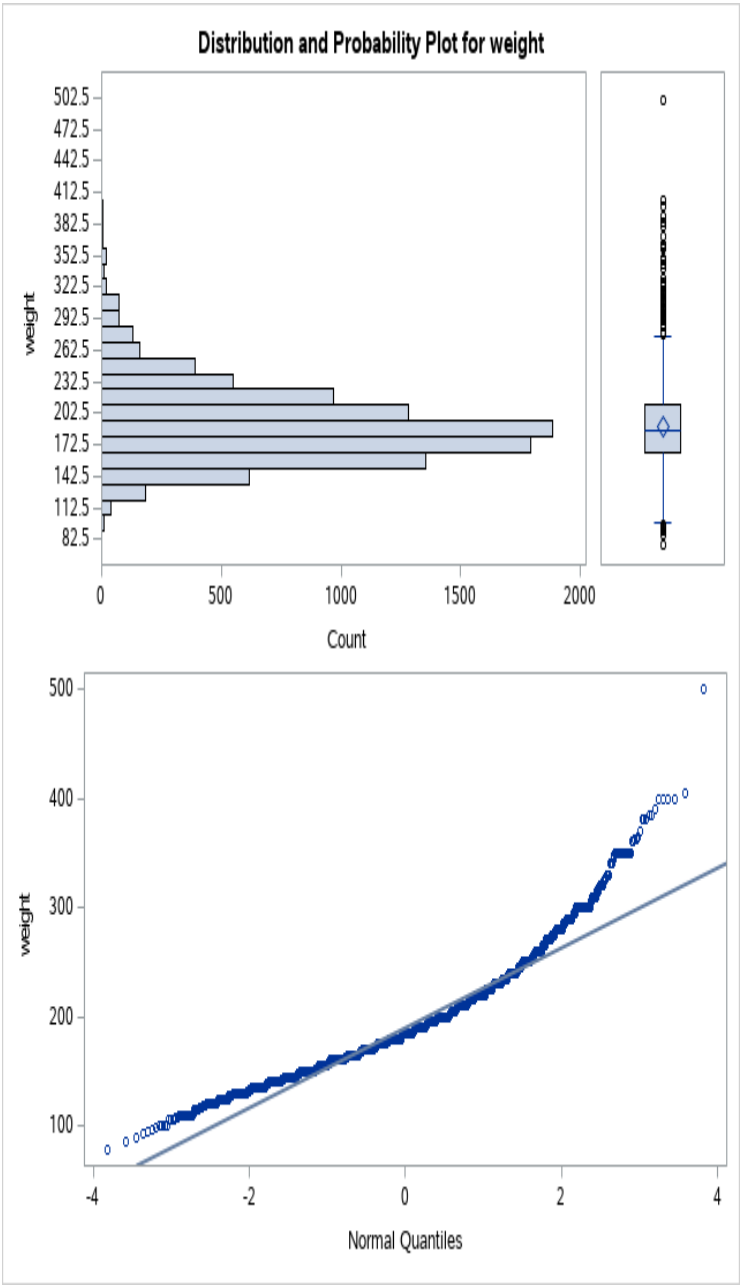
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	506.6924	Pr > |t|	<.0001
Sign	M	4784.5	Pr >= |M|	<.0001
Signed Rank	S	22893833	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantiles (Definition 5)	Quantile

Level	Quantile
100% Max	500
99%	300
95%	256
90%	235
75% Q3	210
50% Median	185
25% Q1	165
10%	150
5%	140
1%	125
0% Min	78

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
78	18065	400	1279
86	15967	400	2659
90	303	400	2944
93	9558	405	15720
94	11333	500	1995



Create box-and-whisker plot and a histogram for weight subset by gender using "by" statement

In [21]:
```
proc sort data=work.cdc;
    by gender;
run;

ods graphics;
proc univariate data=work.cdc plots;
    by gender;
    var Weight;
run;
```

Out[21]:

The SAS System

The UNIVARIATE Procedure

Variable: weight

gender=f

Moments			
N	10431	Sum Weights	1043
Mean	151.666187	Sum Observations	1582
Std Deviation	34.2975191	Variance	1176
Skewness	1.38530136	Kurtosis	3.77
Uncorrected SS	252209474	Corrected SS	1226
Coeff Variation	22.6138203	Std Error Mean	0.33

Basic Statistical Measures			
Location		Variability	
Mean	151.6662	Std Deviation	34.29752
Median	145.0000	Variance	1176
Mode	140.0000	Range	427.00000
		Interquartile Range	42.00000

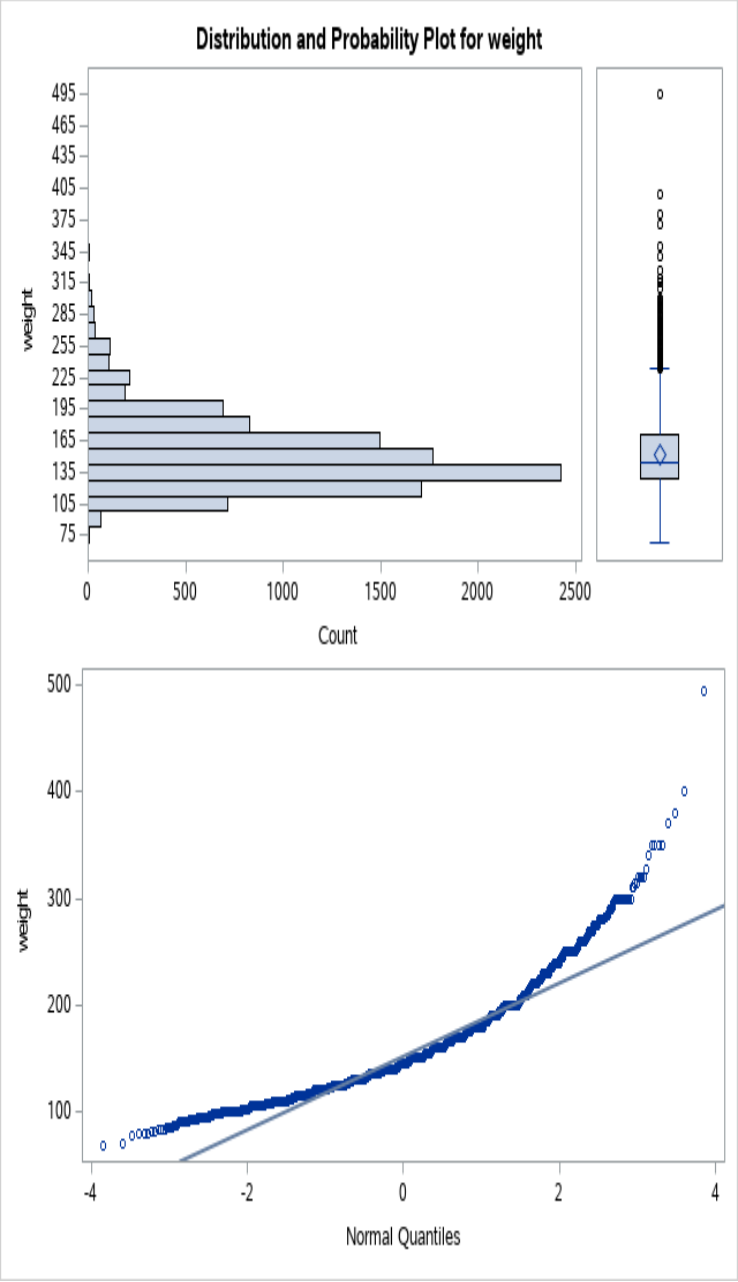
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	451.6365	Pr > |t|	<.0001

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Sign	M	5215.5	Pr >= |M|	<.0001
Signed Rank	S	27204048	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	495
99%	260
95%	220
90%	198
75% Q3	170
50% Median	145
25% Q1	128
10%	115
5%	110
1%	100
0% Min	68

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
68	9739	350	7051

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
70	8592	371	2354
78	5860	380	3663
79	3905	400	10060
80	8154	495	2284



The SAS System

The UNIVARIATE Procedure
Variable: weight

gender=m

Moments			
N	9569	Sum Weights	9569
Mean	189.322709	Sum Observations	1811
Std Deviation	36.5503551	Variance	1335
Skewness	1.16960127	Kurtosis	3.04
Uncorrected SS	355764673	Corrected SS	1278
Coeff Variation	19.3058484	Std Error Mean	0.37

Basic Statistical Measures			
Location		Variability	
Mean	189.3227	Std Deviation	36.55036
Median	185.0000	Variance	1336
Mode	180.0000	Range	422.00000
		Interquartile Range	45.00000

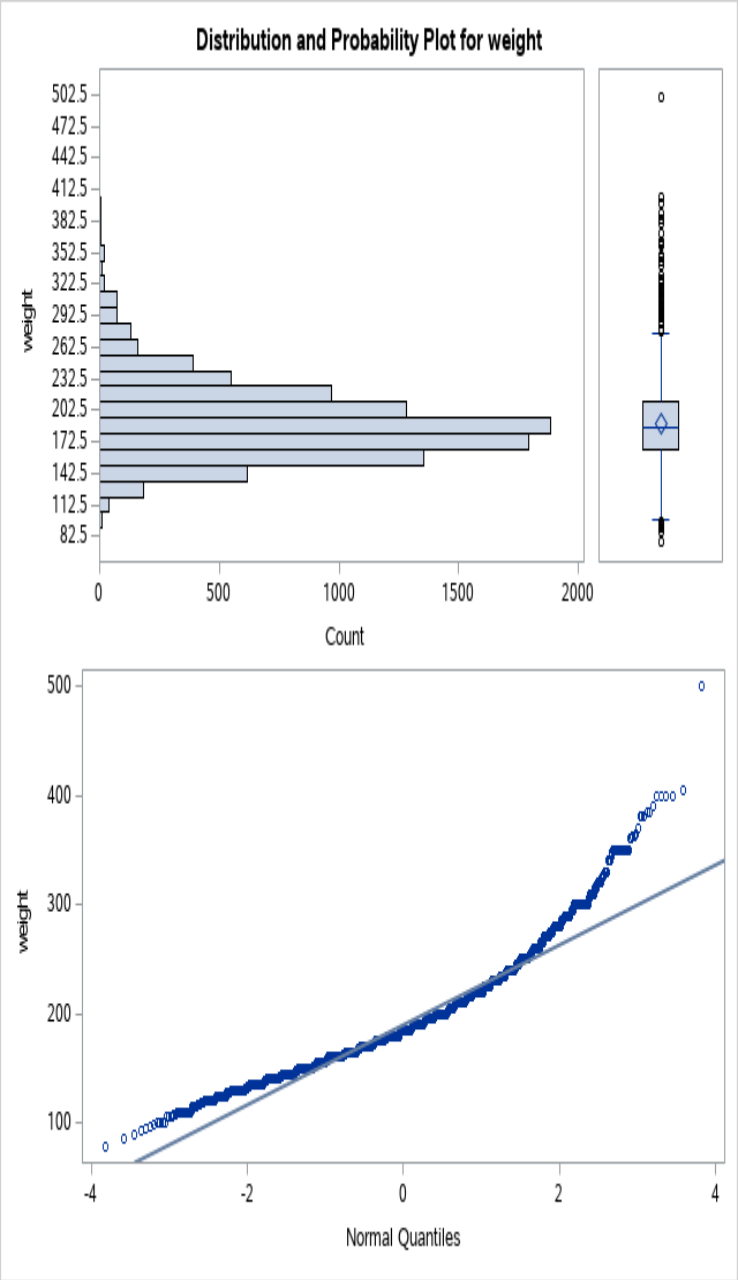
Tests for Location: Mu0=0		
Test	Statistic	p Value

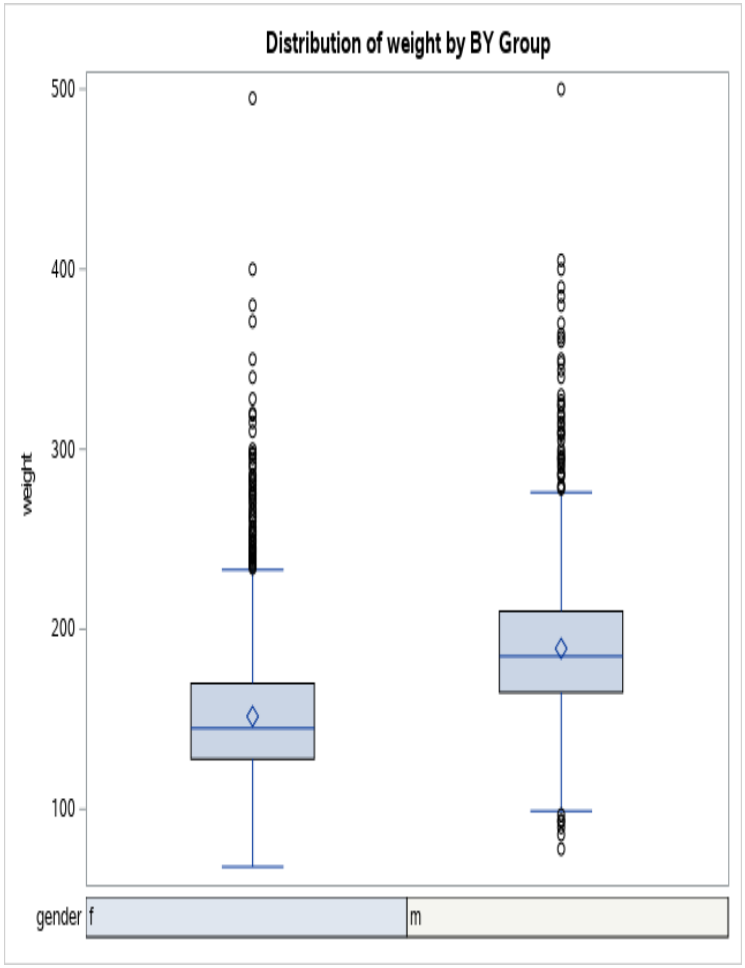
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	506.6924	Pr > |t|	<.0001
Sign	M	4784.5	Pr >= |M|	<.0001
Signed Rank	S	22893833	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	500
99%	300
95%	256
90%	235
75% Q3	210
50% Median	185
25% Q1	165
10%	150
5%	140
1%	125
0% Min	78

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
78	19127	400	11036
86	18099	400	11725
90	10578	400	11867
93	15074	405	17968
94	15887	500	11392


Distribution and Probability Plot for weight
weight
502.5
472.5
442.5
412.5
382.5
352.5
322.5
292.5
262.5
232.5
202.5
172.5
142.5
112.5
82.5
0
500
1000
1500
2000
Count
weight
500
400
300
200
100
-4
-2
0
2
4
Normal Quantiles



Data set with a new variable named bmi

In [22]:
```
data work.cdcbmi;
    set work.cdc;
    bmi = (weight / height**2) * 703;
run;
```

Out[22]:

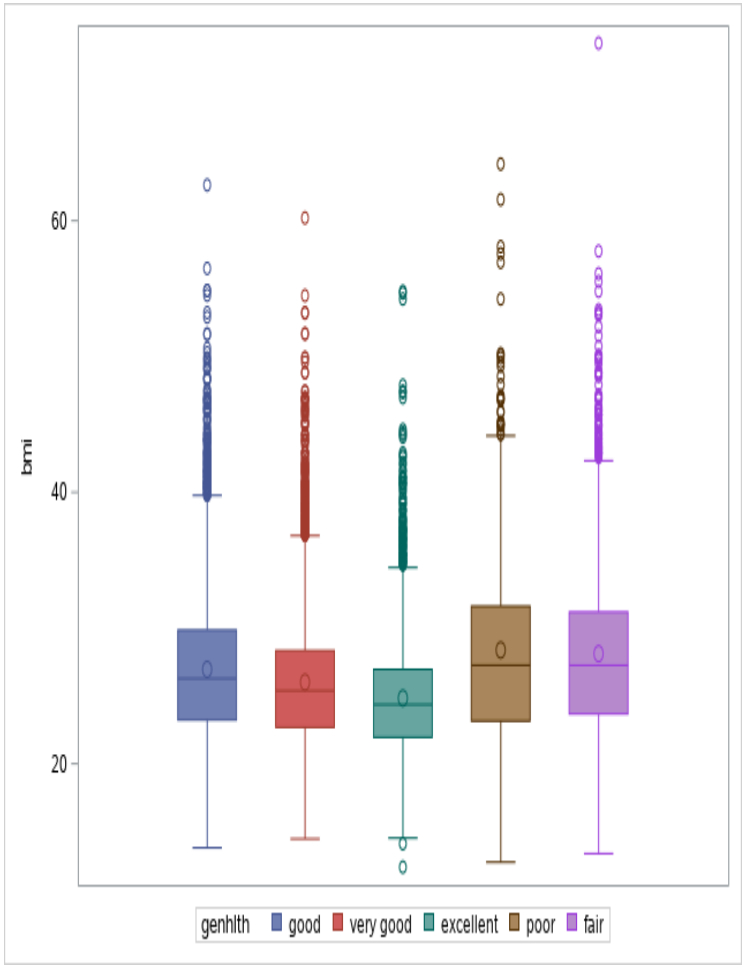
```
288  ods listing close;ods html5 (id=saspy_
internal) file=stdout options(bitmap_mode
='inline') device=svg; ods graphics on /
288! outputfmt=png;
NOTE: Writing HTML5(SASPY_INTERNAL) Body fi
le: STDOUT
289
290  data work.cdcbmi;
291      set work.cdc;
292      bmi = (weight / height**2) * 703;
293  run;
NOTE: There were 20000 observations read fr
om the data set WORK.CDC.
NOTE: The data set WORK.CDCBMI has 20000 ob
servations and 10 variables.
NOTE: DATA statement used (Total process ti
me):
      real time           0.00 seconds
```

```
      cpu time               0.00 seconds

294
295  ods html5 (id=saspy_internal) close;od
s listing;

296
```

Alternative method to producing box-and-whisker plots using group statement

In [23]:
```
proc sgplot data=work.cdcbmi;
    vbox bmi / group=genhlth;
run;
```

Out[23]:



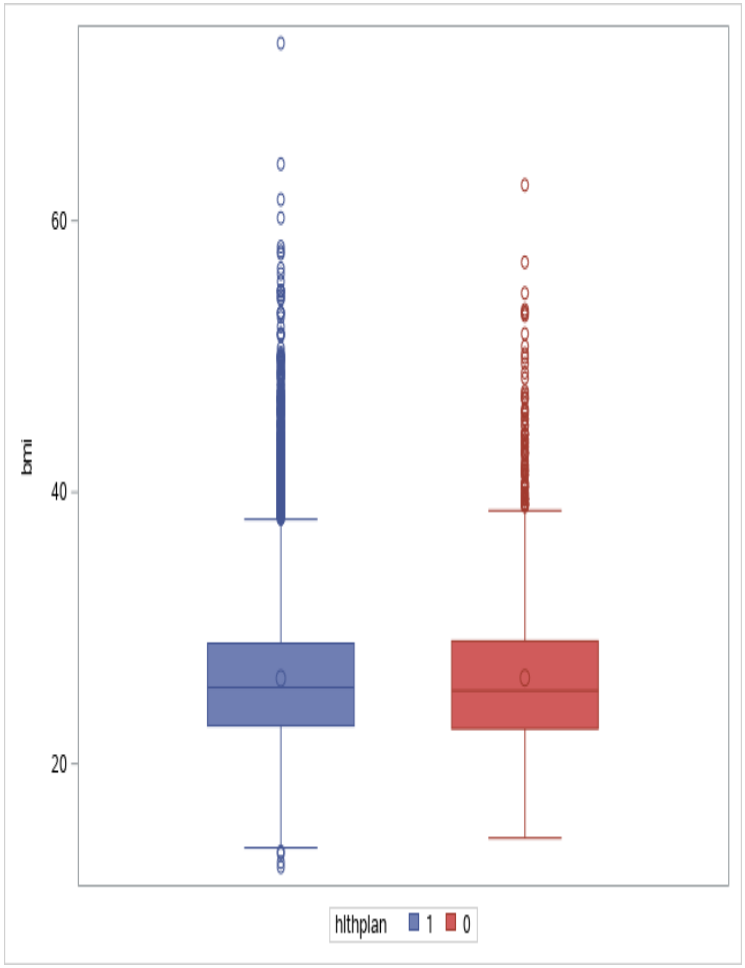
Exercise 5

What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

In [24]:
```
proc sgplot data=work.cdcbmi;
```

```
    vbox bmi / group=hlthplan;
run;
```

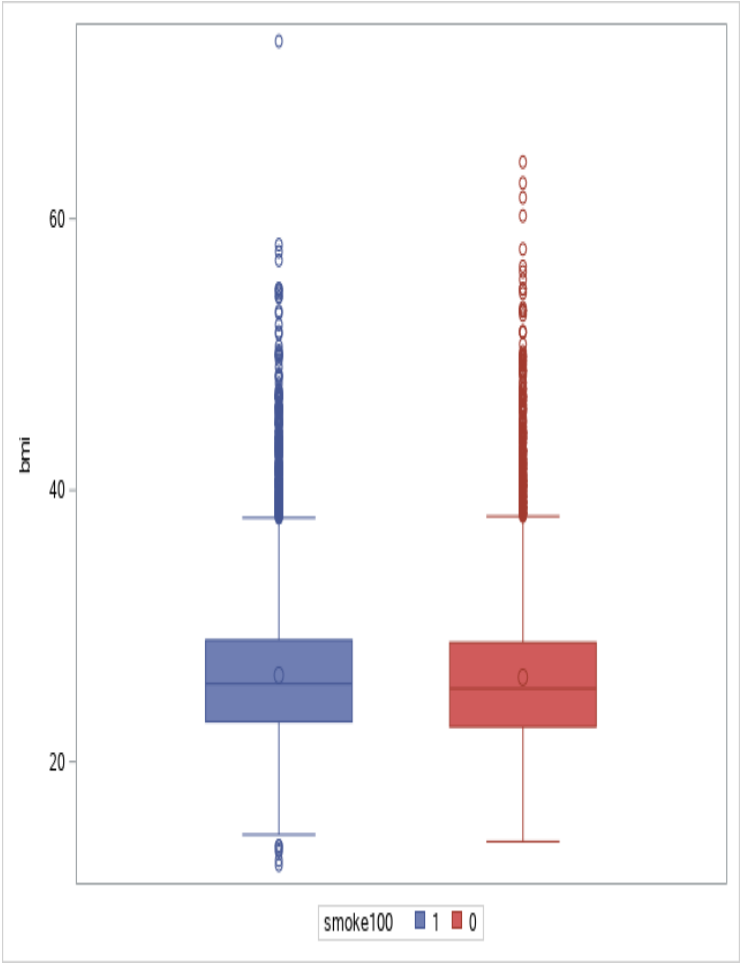
Out[24]:



In [25]:

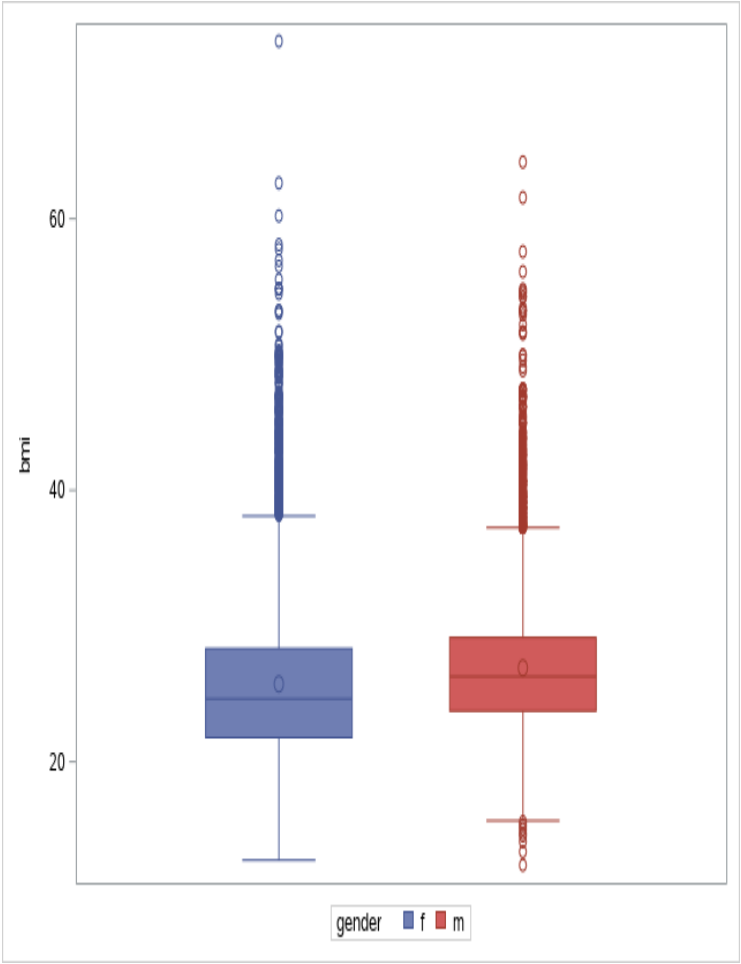
```
proc sgplot data=work.cdcbmi;
    vbox bmi / group=smoke100;
run;
```

Out[25]:



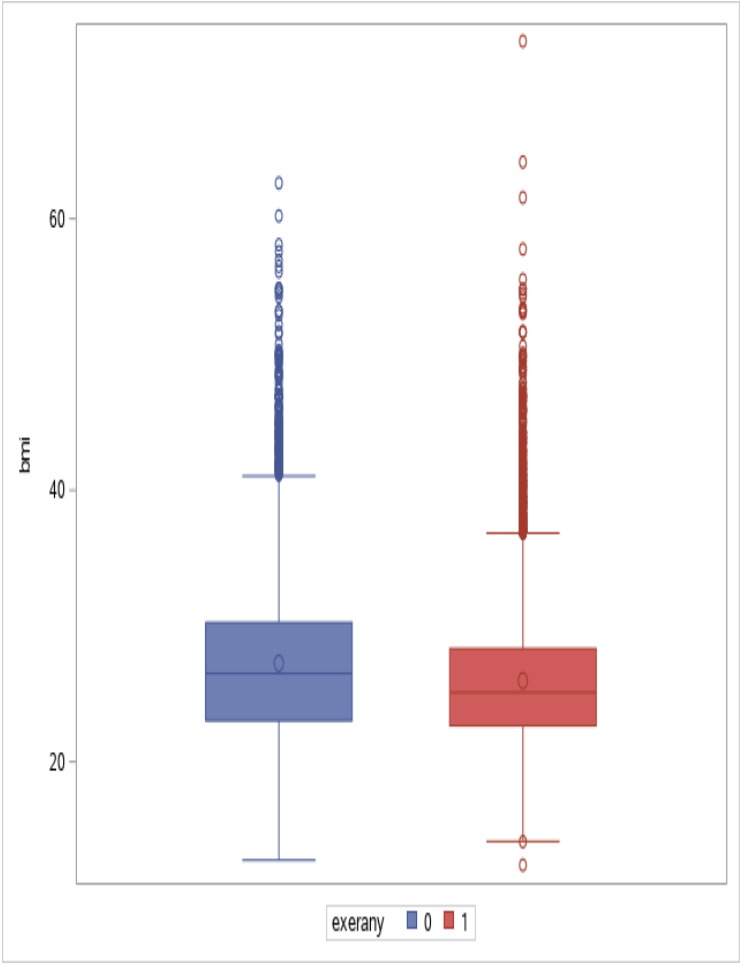
In [26]:
```
proc sgplot data=work.cdcbmi;
    vbox bmi / group=gender;
run;
```

Out[26]:



In [27]:
```
proc sgplot data=work.cdcbmi;
    vbox bmi / group=exerany;
run;
```

Out[27]:



Histogram for the bmi of the respondents

In [28]:
```
ods graphics;
proc univariate data=work.cdcbmi;
    var bmi;
    histogram bmi;
run;
```

Out[28]:

The SAS System

The UNIVARIATE Procedure
Variable: bmi

Moments			
N	20000	**Sum Weights**	2000
Mean	26.3069252	**Sum Observations**	5261
Std Deviation	5.21810488	**Variance**	27.2

Moments			
Skewness	1.27589014	Kurtosis	3.32
Uncorrected SS	14385631.4	Corrected SS	5445
Coeff Variation	19.83548	Std Error Mean	0.03

Basic Statistical Measures			
Location		Variability	
Mean	26.30693	Std Deviation	5.21810
Median	25.60354	Variance	27.22862
Mode	27.12191	Range	60.69029
		Interquartile Range	6.17852

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	712.9717	Pr > |t|	<.0001
Sign	M	10000	Pr >= |M|	<.0001
Signed Rank	S	1.0001E8	Pr >= |S|	<.0001

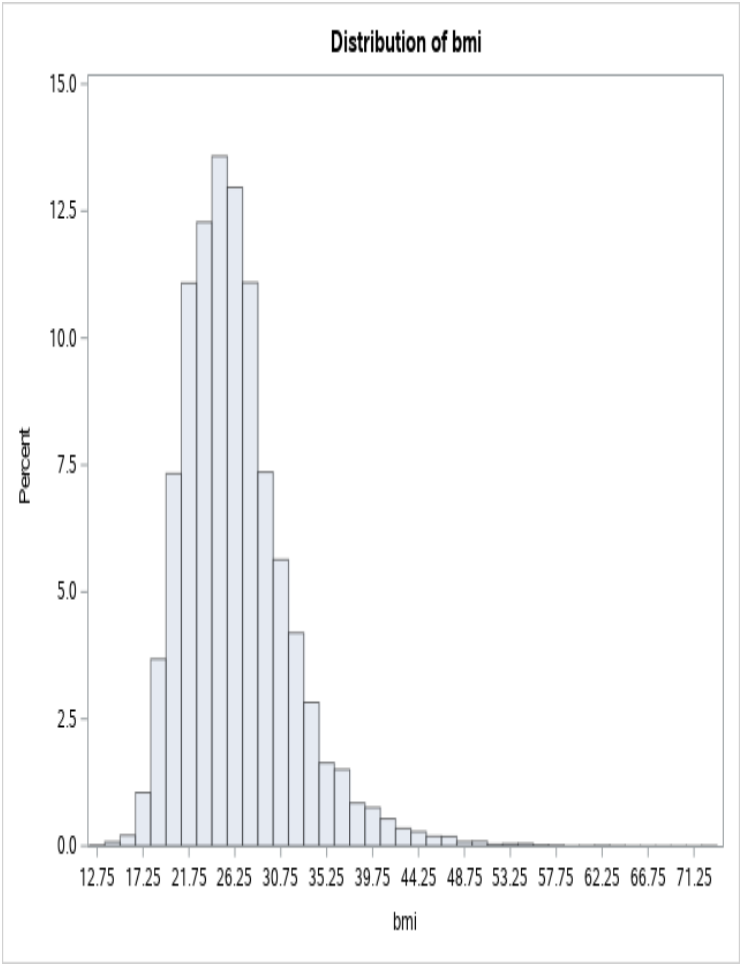
Quantiles (Definition 5)	
Level	Quantile
100% Max	73.0907

Quantiles (Definition 5)	
Level	**Quantile**
99%	43.5553
95%	35.9509
90%	32.8802
75% Q3	28.8862
50% Median	25.6035
25% Q1	22.7077
10%	20.5957
5%	19.4835
1%	17.7123
0% Min	12.4005

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
12.4005	15887	60.2217	443
12.7495	3905	61.5733	17968
13.3872	19127	62.6420	10060
13.5565	8698	64.1892	11392
13.8104	2493	73.0907	2284

The SAS System

The UNIVARIATE Procedure



Histogram for the bmi of the respondents grouped by gender

In [29]:
```
ods graphics;
proc univariate data=work.cdcbmi;
    class gender;
    var bmi;
    histogram bmi;
run;
```

Out[29]:

The SAS System

The UNIVARIATE Procedure
Variable: bmi
gender = f

Moments			
N	10431	**Sum Weights**	1043
Mean	25.7411474	**Sum Observations**	2685

Moments			
Std Deviation	5.62057729	Variance	31.5
Skewness	1.36746824	Kurtosis	3.13
Uncorrected SS	7241143.13	Corrected SS	329
Coeff Variation	21.8349913	Std Error Mean	0.05

Basic Statistical Measures			
Location		Variability	
Mean	25.74115	Std Deviation	5.62058
Median	24.63832	Variance	31.59089
Mode	27.46094	Range	60.34120
		Interquartile Range	6.55244

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	467.7459	Pr > |t|	<.0001
Sign	M	5215.5	Pr >= |M|	<.0001
Signed Rank	S	27204048	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	73.0907
99%	44.3980
95%	36.5765
90%	32.9453
75% Q3	28.3396
50% Median	24.6383
25% Q1	21.7872
10%	19.8529
5%	18.9853
1%	17.2684
0% Min	12.7495

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
12.7495	3905	57.7725	3663
13.5565	8698	58.1005	2354
13.8104	2493	60.2217	443
13.8156	5860	62.6420	10060
14.1367	8592	73.0907	2284

The SAS System

The UNIVARIATE Procedure

Variable: bmi
gender = m

Moments			
N	9569	Sum Weights	9569
Mean	26.9236698	Sum Observations	2576
Std Deviation	4.6633462	Variance	21.7
Skewness	1.30603659	Kurtosis	3.92
Uncorrected SS	7144488.31	Corrected SS	2080
Coeff Variation	17.3206188	Std Error Mean	0.04

Basic Statistical Measures			
Location		Variability	
Mean	26.92367	Std Deviation	4.66335
Median	26.28313	Variance	21.74680
Mode	27.12191	Range	51.78874
		Interquartile Range	5.41028

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	564.7677	Pr > |t|	<.0001
Sign	M	4784.5	Pr >= |M|	<.0001

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Signed Rank	**S**	22893833	**Pr >= |S|**	<.0001

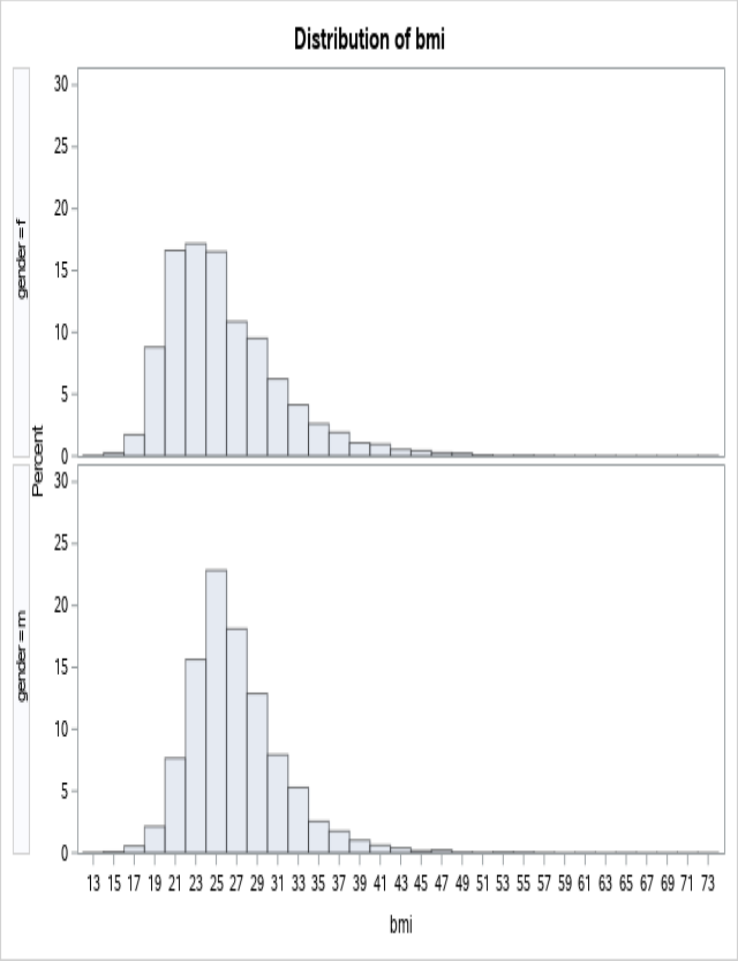
Quantiles (Definition 5)	
Level	**Quantile**
100% Max	64.1892
99%	42.0390
95%	35.4379
90%	32.5463
75% Q3	29.1561
50% Median	26.2831
25% Q1	23.7458
10%	21.9247
5%	20.8031
1%	18.6510
0% Min	12.4005

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
12.4005	15887	54.8118	17510
13.3872	19127	56.1101	10796
14.1216	11534	57.5866	16778

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14.5493	18866	61.5733	17968
14.7603	18099	64.1892	11392

The SAS System

The UNIVARIATE Procedure



Histogram for the age of the respondents with 50 bins

In [30]:
```
proc univariate data=work.cdcbmi;
   var bmi;
   histogram bmi / nmidpoints=50;
run;
```

Out[30]:

The SAS System

The UNIVARIATE Procedure
Variable: bmi

Moments			
N	20000	Sum Weights	200(
Mean	26.3069252	Sum Observations	526(
Std Deviation	5.21810488	Variance	27.2
Skewness	1.27589014	Kurtosis	3.32
Uncorrected SS	14385631.4	Corrected SS	544(
Coeff Variation	19.83548	Std Error Mean	0.03

Basic Statistical Measures			
Location		Variability	
Mean	26.30693	Std Deviation	5.21810
Median	25.60354	Variance	27.22862
Mode	27.12191	Range	60.69029
		Interquartile Range	6.17852

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	712.9717	Pr > |t|	<.0001

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Sign	**M**	10000	**Pr >= |M|**	<.0001
Signed Rank	**S**	1.0001E8	**Pr >= |S|**	<.0001

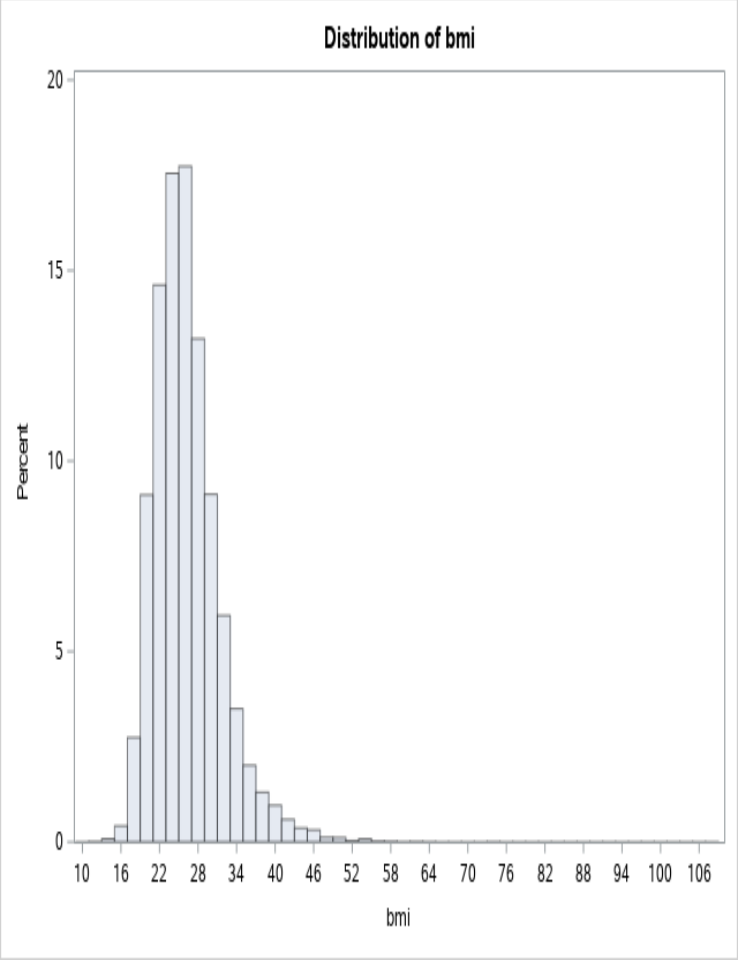
Quantiles (Definition 5)	
Level	**Quantile**
100% Max	73.0907
99%	43.5553
95%	35.9509
90%	32.8802
75% Q3	28.8862
50% Median	25.6035
25% Q1	22.7077
10%	20.5957
5%	19.4835
1%	17.7123
0% Min	12.4005

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
12.4005	15887	60.2217	443

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12.7495	3905	61.5733	17968
13.3872	19127	62.6420	10060
13.5565	8698	64.1892	11392
13.8104	2493	73.0907	2284

The SAS System

The UNIVARIATE Procedure



Histogram for the age of the respondents with 50 bins include the mean, standard deviation, and median

In [31]:
```
ods graphics;
proc univariate data=work.cdcbmi;
    var bmi;
    histogram bmi / nmidpoints=50;
    inset mean std median / position=NE;
run;
```

Out[31]:

The SAS System

The UNIVARIATE Procedure
Variable: bmi

Moments			
N	20000	Sum Weights	2000
Mean	26.3069252	Sum Observations	5261
Std Deviation	5.21810488	Variance	27.2
Skewness	1.27589014	Kurtosis	3.32
Uncorrected SS	14385631.4	Corrected SS	5445
Coeff Variation	19.83548	Std Error Mean	0.03

Basic Statistical Measures			
Location		Variability	
Mean	26.30693	Std Deviation	5.21810
Median	25.60354	Variance	27.22862
Mode	27.12191	Range	60.69029
		Interquartile Range	6.17852

Tests for Location: Mu0=0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	712.9717	Pr > |t|	<.0001
Sign	M	10000	Pr >= |M|	<.0001
Signed Rank	S	1.0001E8	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	73.0907
99%	43.5553
95%	35.9509
90%	32.8802
75% Q3	28.8862
50% Median	25.6035
25% Q1	22.7077
10%	20.5957
5%	19.4835
1%	17.7123
0% Min	12.4005

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12.4005	15887	60.2217	443
12.7495	3905	61.5733	17968
13.3872	19127	62.6420	10060
13.5565	8698	64.1892	11392
13.8104	2493	73.0907	2284

The SAS System

The UNIVARIATE Procedure