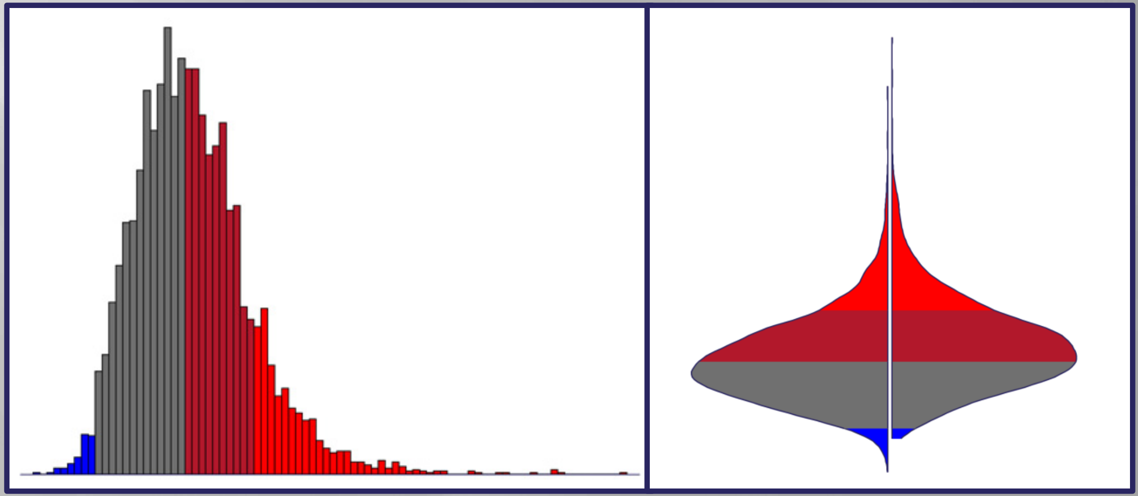
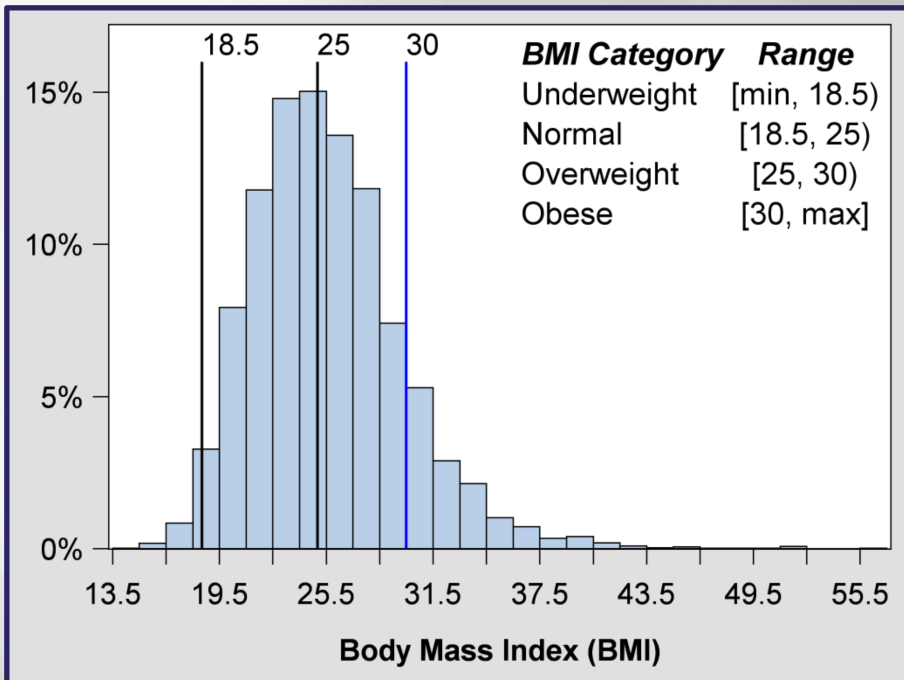
Increase Pattern Detection in SAS® GTL with New Categorical Histograms and Color Coded Asymmetric Violin Plots

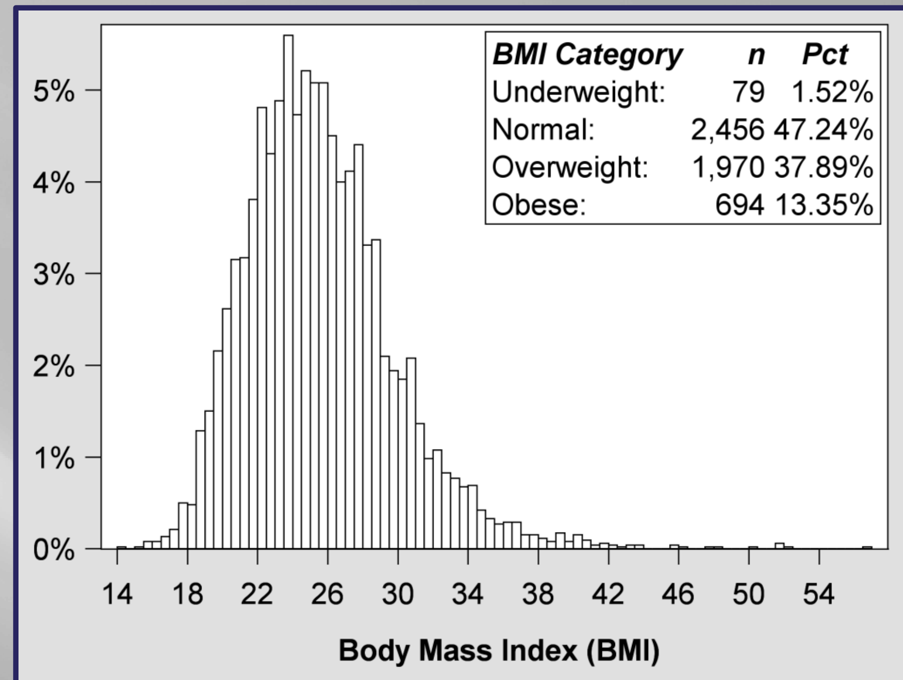
Perry Watts, Stakana Analytics, pwatts@stakana.com

Detecting patterns in graphics output is much easier when continuous data can be grouped categorically. Such is the case with the Body Mass Index and its four classifications: *underweight*, *normal weight*, *overweight* and *obese*. This presentation goes from conventional histogram to color-coded asymmetric violin plot with coverage of the categorical histogram along the way.



Problems with the Conventional Histogram





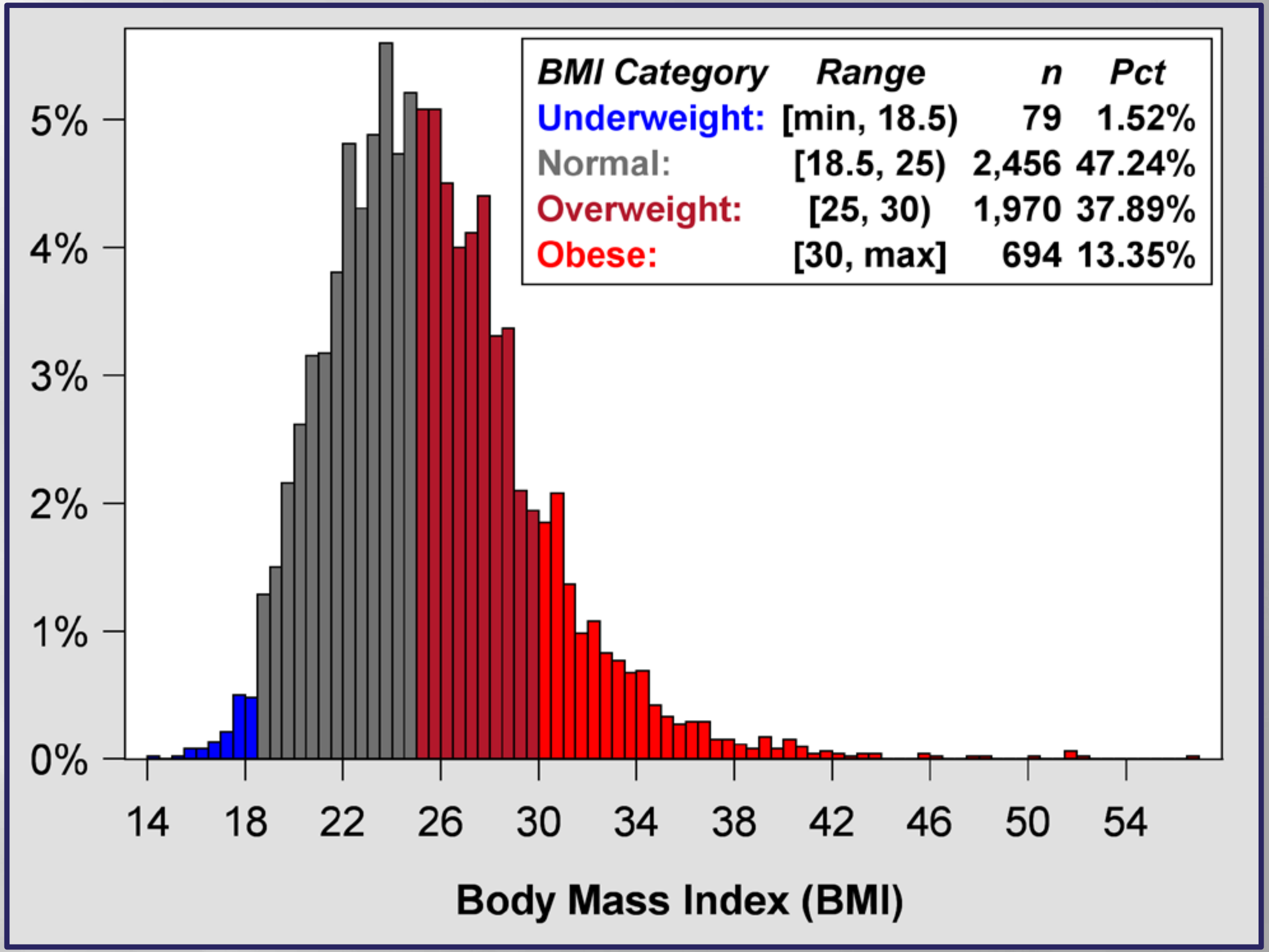
With default settings, bin boundaries and category endpoints are not aligned. You can't fix the problem by simply adding bin boundaries at **18.5** and **25**, because the HISTOGRAM statement in GTL requires equal bin widths.

Bin boundaries and category endpoints are now aligned. However with 86 bins it is impossible to define category boundaries.

Data: SASHELP.HEART from the Framingham Heart Study with

$$BMI = 703 \times \frac{Weight}{Height^2}$$

The (Ordinal) Categorical Histogram



With legend and bin colors it is possible to identify BMI categories and associated ranges in the histogram. Later, you will see an example where cholesterol is ranked as "desirable", "borderline" and "high". Can you think of additional applications from your own data?

Building a Categorical Histogram

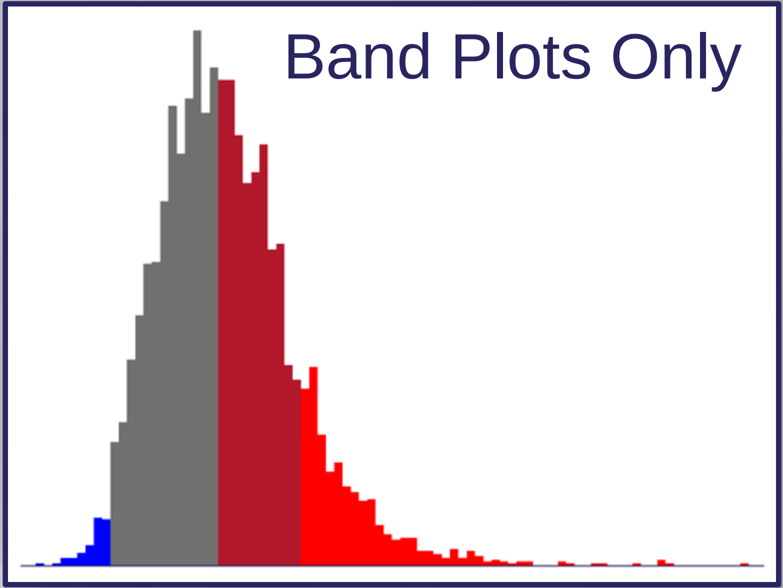
Requires a Two-Pass Solution

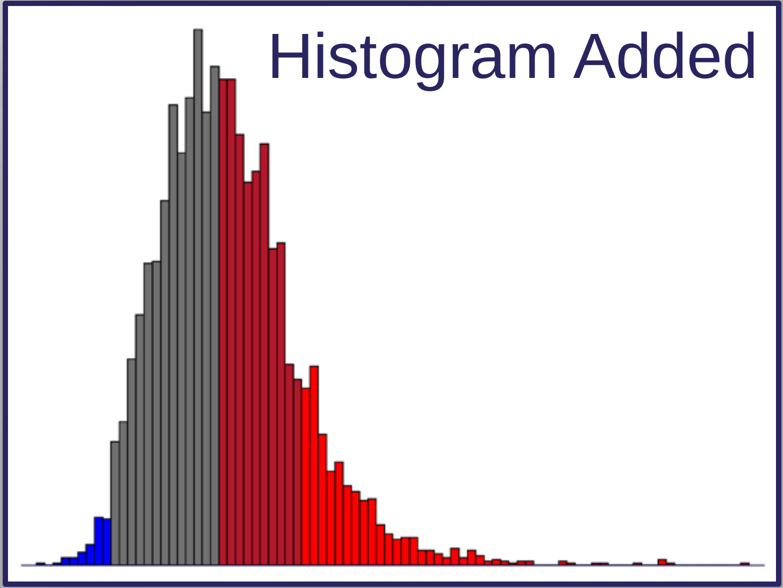
In the **first pass**, an object data set, **`PlotObjDS`**, is created to capture histogram coordinates. In the **second pass**, **`PlotObjDS`** is used to create four band plots; one for each BMI category. Below is Partial Code for the red obese band that uses variables from **`PlotObjDS`**:

```
BANDPLOT X=xxObese
LIMITUPPER=yyObese
LIMITLOWER=0 /
    TYPE=STEP JUSTIFY=LEFT
    DISPLAY=(FILL)
    FILLATTRS=(COLOR=cxFF0000);
```

As a final step, a hollow histogram is plotted over the band plots:

```
HISTOGRAM BMIScore /
BINSTART=14
    BINWIDTH=0.5
    DISPLAY=(OUTLINE) ...;
```





From Categorical Histogram to Categorical KDE Plot

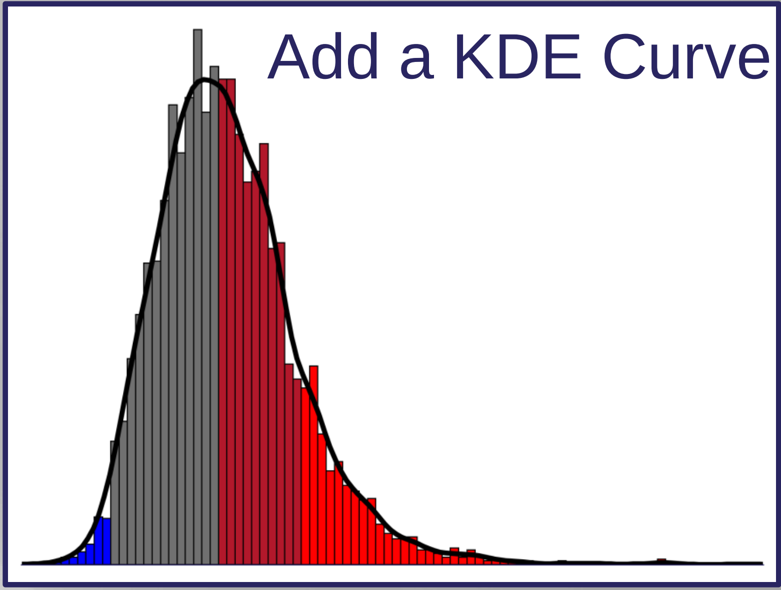
Add a KDE curve to the categorical histogram to see how they are related:

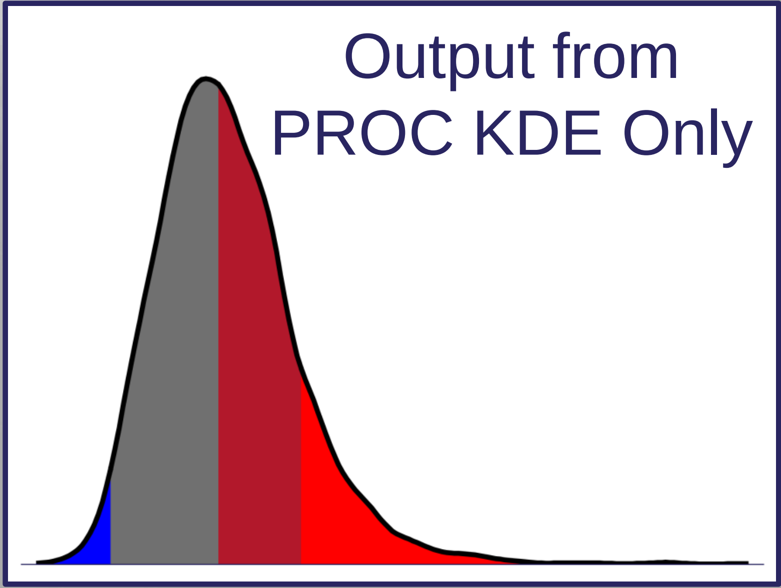
```
HISTOGRAM   BMIScore / ...;
DENSITYPLOT BMIScore / kernel()
  LINEATTRS=(COLOR=black THICKNESS=2);
```

Replace the HISTOGRAM statement with a SERIESPLOT statement that uses output from PROC KDE:

```
PROC KDE DATA = heartBMIdat;
   UNIVAR bmiscore (GRIDL=14 GRIDU=57) /
     NGRID=173 PLOTS=none ...;
     OUT=KDEOUT(...); RUN;
```

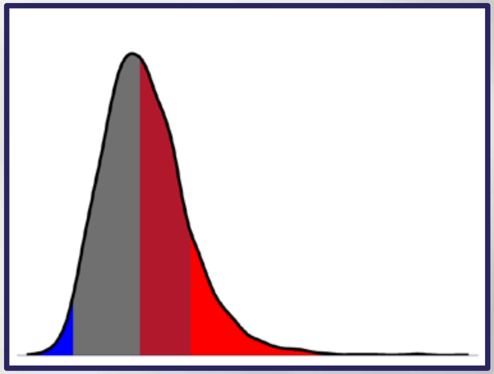
GRIDL, **GRIDU** and **NGRID** are set up so that there will be an **X** coordinate at each category boundary in the output data set.

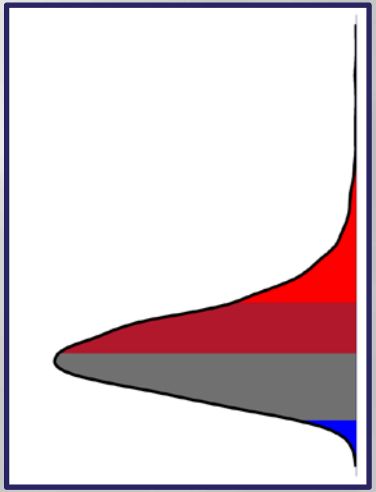


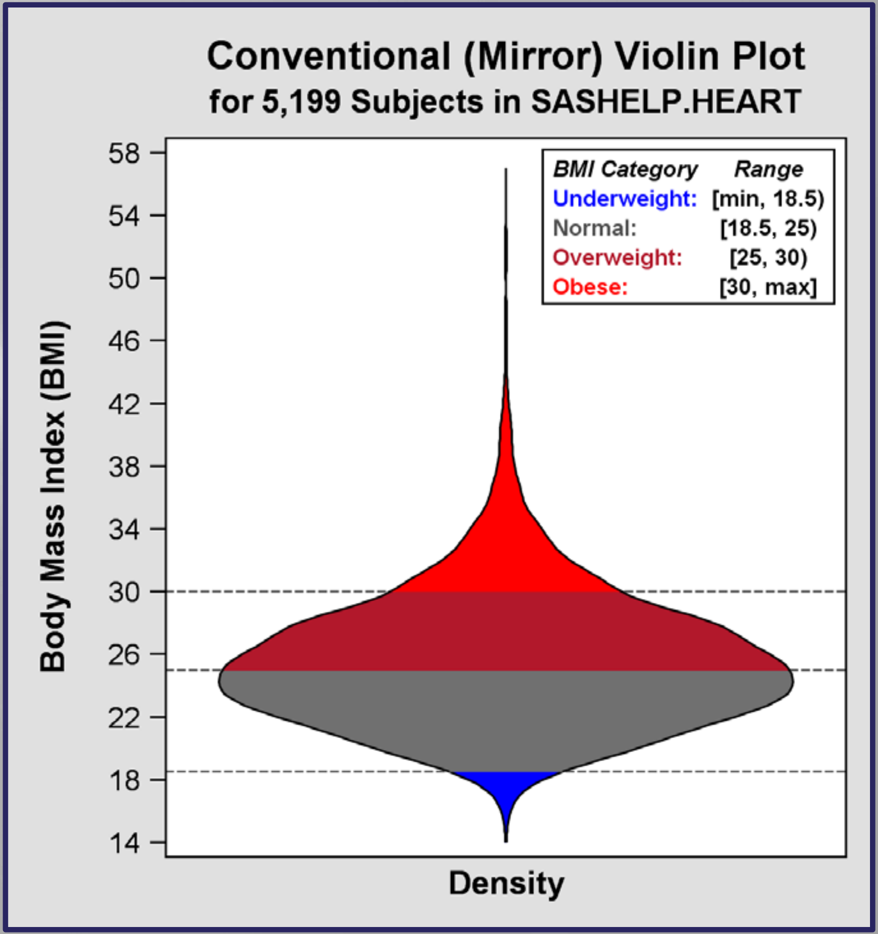


From KDE Plot to Symmetric Violin Plot

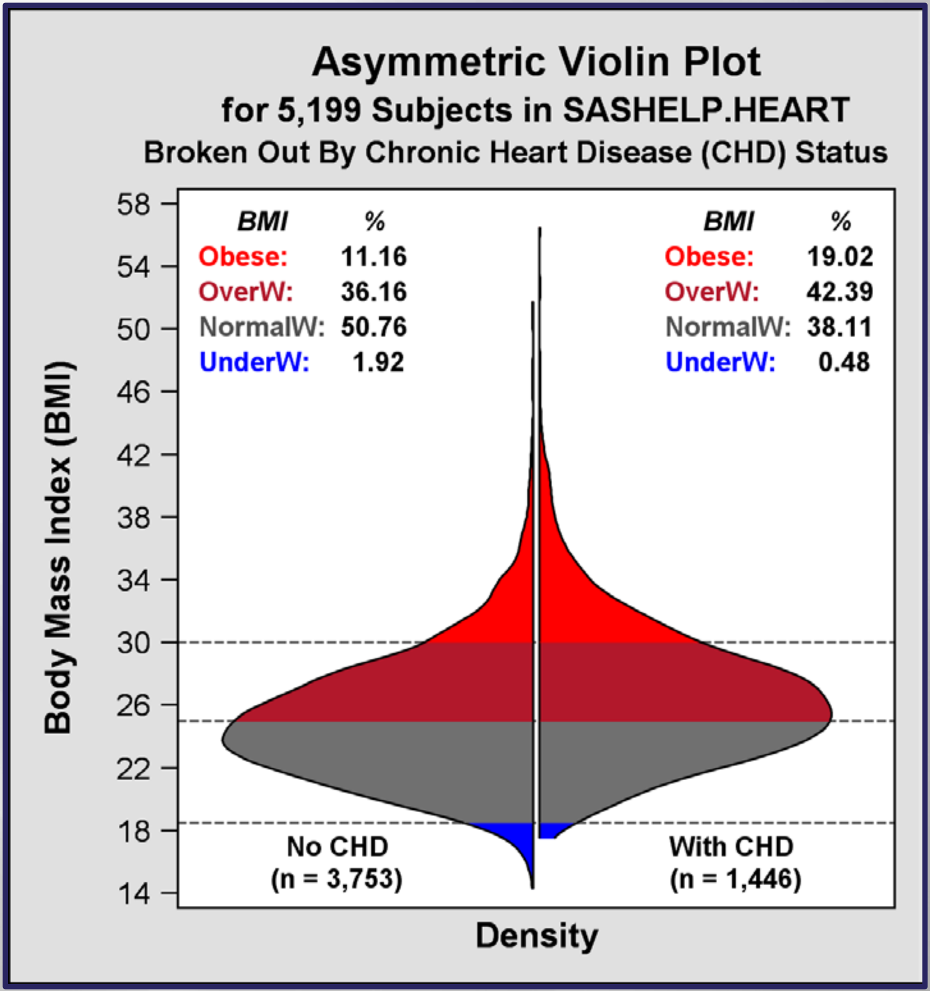
Rotate the KDE Plot 90° and use it to create identical halves of a Violin Plot







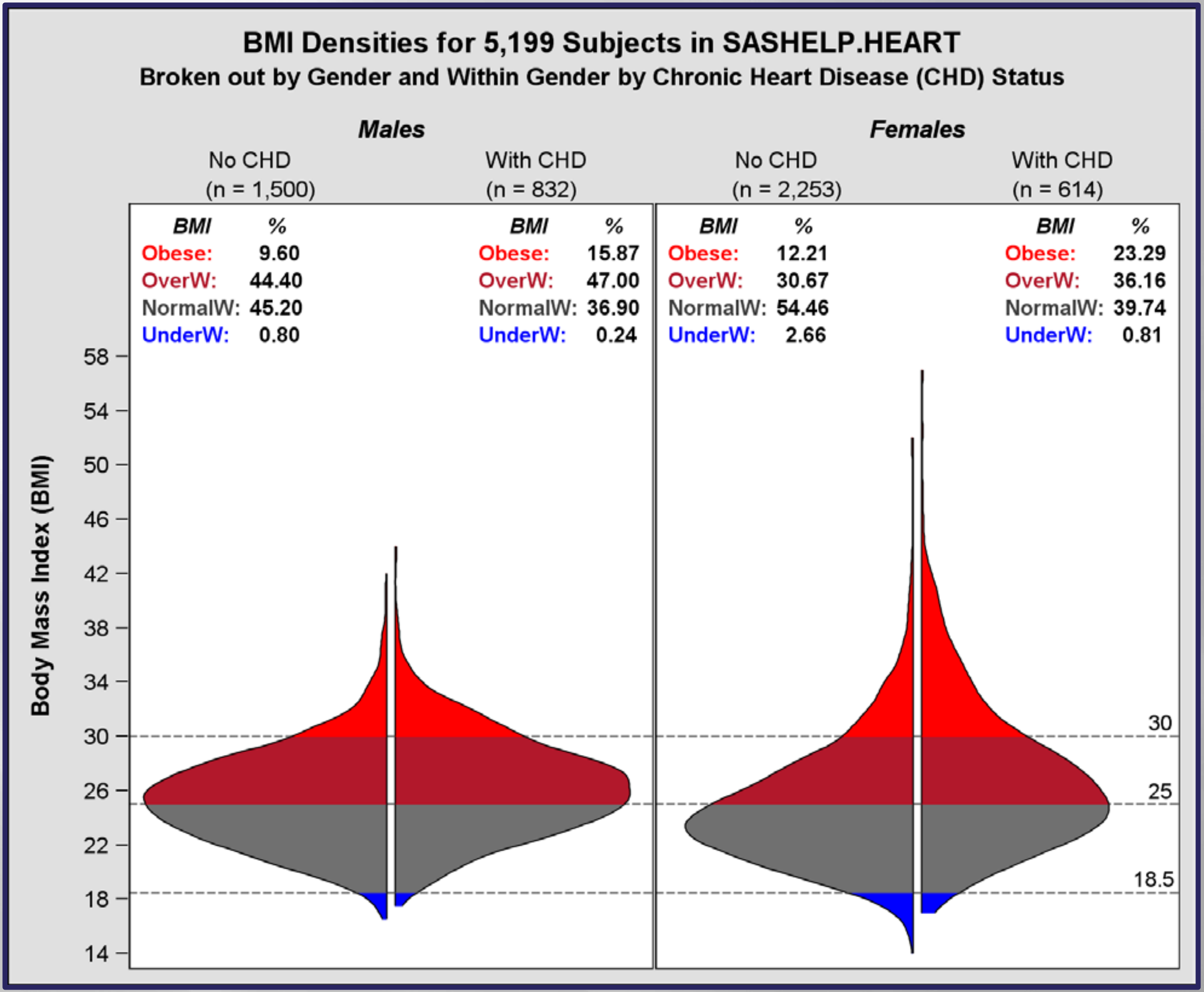
From Symmetric to Asymmetric Violin Plot



- The asymmetric violin plot facilitates **distribution comparison**. The two halves of the violin plot continue to share a common Y-axis. Since areas are still difficult to compare, the graph is annotated.
- The single-panel display is created with **LAYOUT OVERLAY**. **`Density`** = **`-Density`** for the ***No CHD*** half-plot.***** In addition, X coordinates have been moved by a "fuzz" amount to accommodate the small vertical break at **`Density=0.`**

*****Adapted from the ***Graphically Speaking*** SAS BLOG:
http://blogs.sas.com/content/graphicallyspeaking/2012/10/30/violin-plots/

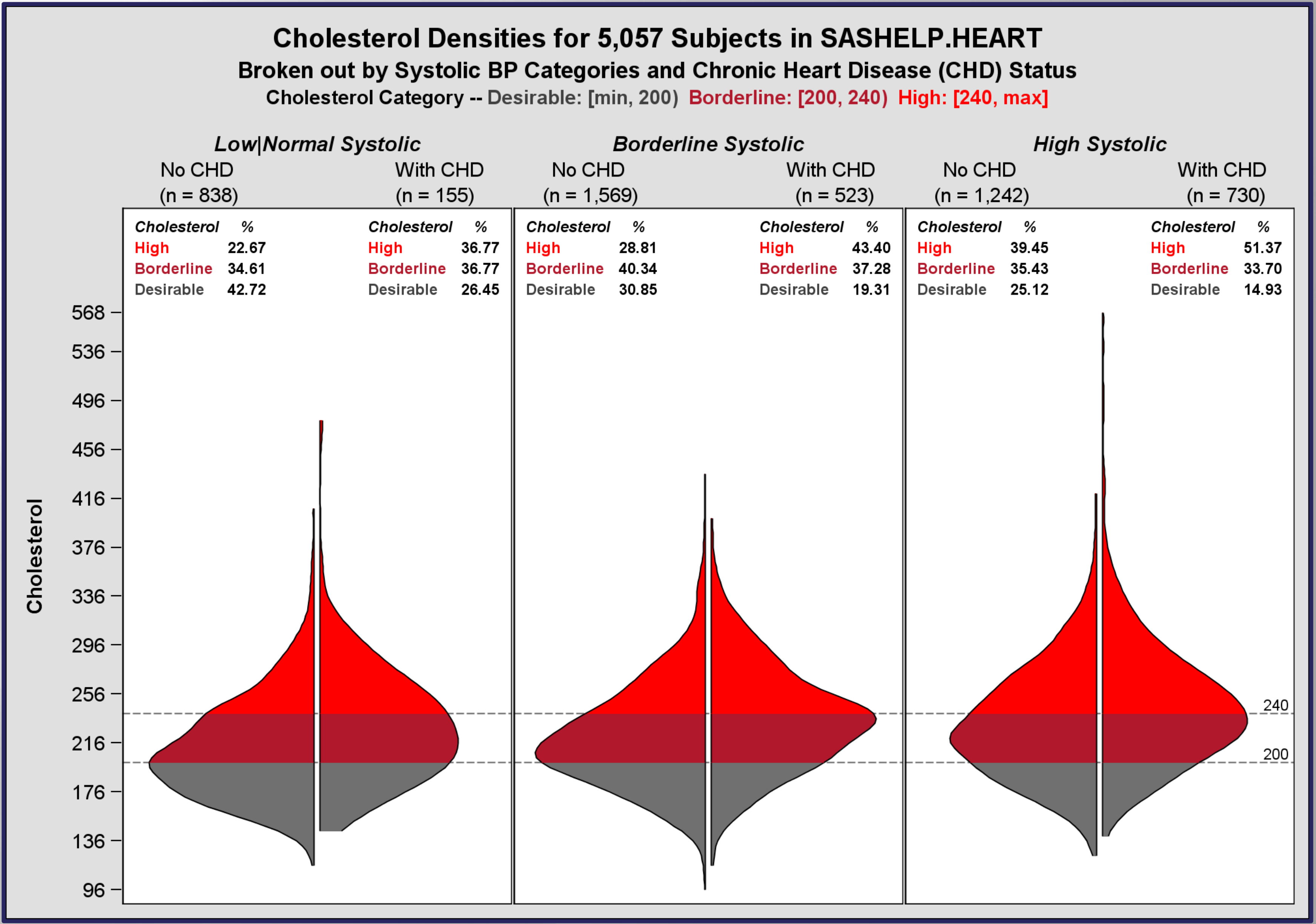
A Group of Asymmetric Violin Plots



- LAYOUT LATTICE is used in this graph to get enhanced cell headers for males and females.
- BMI category percents are displayed with embedded LAYOUT GRIDDED statements.
- The gridlines at 18.5, 25 and 30 are created with DROPLINE statements. They are labeled with DRAWTEXT statements.

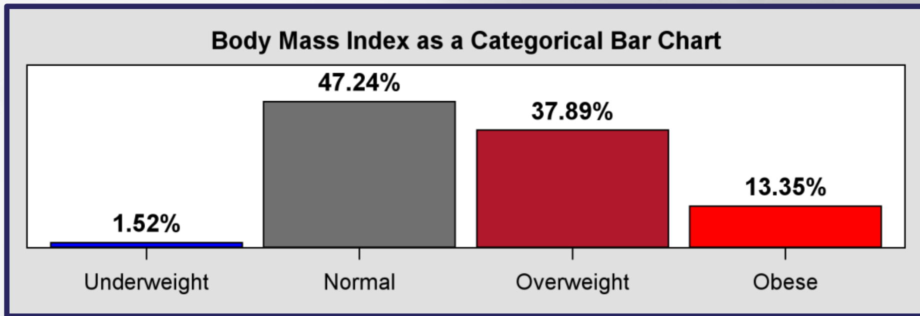
Question: What patterns can you detect by looking at the graph?

Another Group of Asymmetric Violin Plots

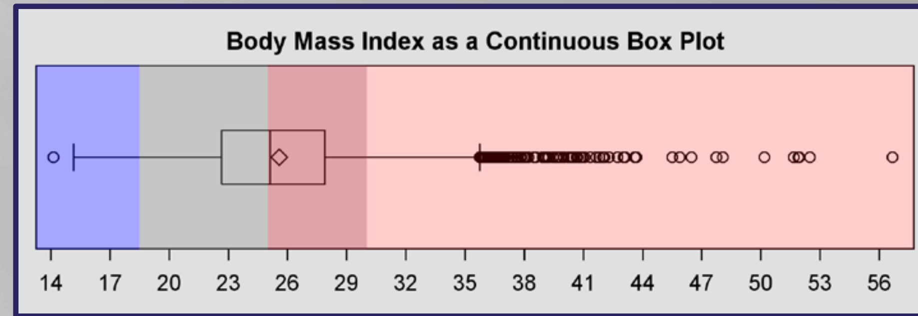


A Violin Plot Combines Categorical and Continuous Data

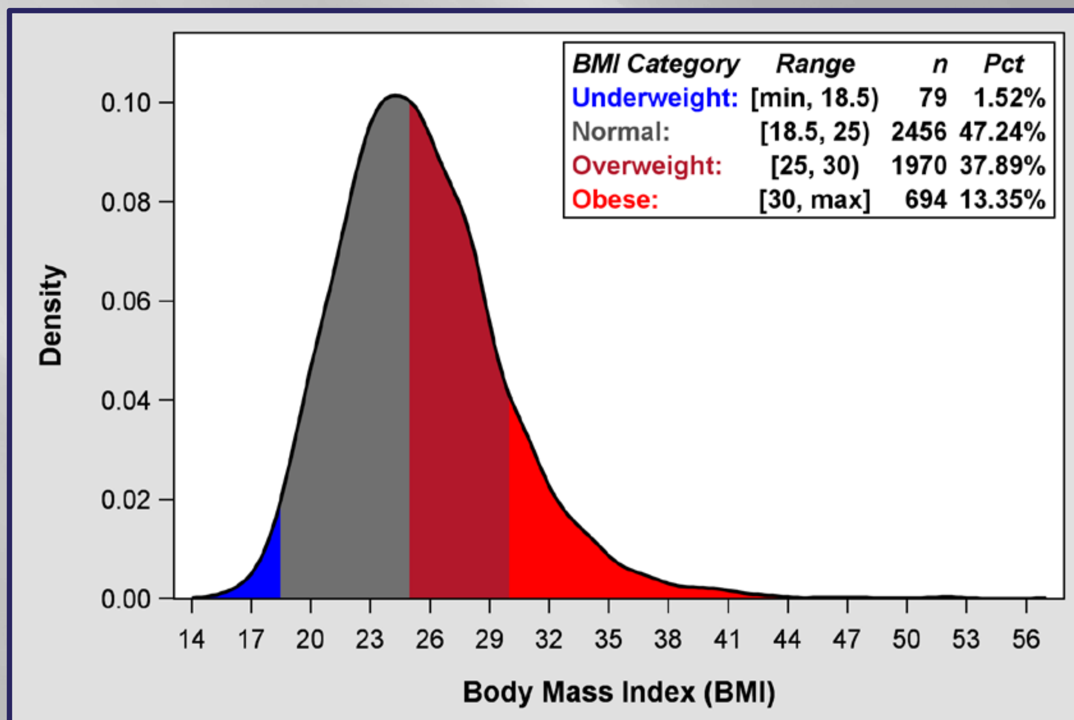
Categorical



Continuous



Are Expressed in a Violin Plot



Conclusion

Pattern detection is facilitated with the new **categorical histogram** and **asymmetric violin plot**. From the **histogram** we can see that subjects who participated in the Framingham Heart Study are considerably overweight. When the **asymmetric violin plot** is used to separate subjects by Chronic Heart Disease, an upward shift in the BMI is observed in those subjects with CHD.

The breakout by gender shown in Slide #8 is revealing. While males are more ***overweight*** than females, females are more ***obese*** than males. However, when "No CHD" is defined as baseline, the increase in both ***overweight*** and ***obesity*** that occurs during the transition to "With CHD" is greater for females than it is for males.

STAKANA