Unit 9: Checking Assumptions of Normality

PREREQUISITES

Students need to be familiar with histograms (Unit 3) and boxplots (Unit 5). The background on normal distributions covered in Unit 7, Normal Distributions, and Unit 8, Normal Calculations, are essential to this unit.

ACTIVITY DESCRIPTION

Students learn how to use normal quantile plots by examining a variety of shapes of data and the corresponding normal quantile plots. Students should work in small groups on this activity.

MATERIALS

Access to technology to construct normal quantile plots.

In questions 1 – 4, students are presented with pairs of histograms and normal quantile plots. In their groups, students should discuss the shapes of the histograms and how a given shape affects the pattern of dots in the normal quantile plots. In questions 5 and 6 students are given real data and asked to assess whether or not it is reasonable to assume the data are normally distributed.

For the final question, students are asked to collect their own data. Then they construct normal quantile plots to assess whether the distributions of the data they have collected are normally distributed. They can collect data on themselves – height, forearm length, head circumference, foot length, etc. They can also collect data from online sources; for example, data on sports – salaries of baseball players, lengths of tennis matches, a favorite basketball team's scores for a season, etc. They could also collect data from their school – football scores for a season of games, exam scores from a class, and so forth.

In terms of technology to construct normal quantile plots (or normal probability plots), students can use statistical software, graphing calculators or spreadsheet software such as Excel.

Below are Excel instructions for constructing the simplified version of the normal quantile plot discussed in the Content Overview.

Step 1:
Label two columns: In cell A1, enter the variable name and in B1 enter Normal Quantile.

Step 2:
Enter your data in column A, starting in cell A2.

Step 3:
Use Sort to sort your data in column A from smallest to largest.

Step 4:
In cell B2 enter the following: =NORMINV((CELL(“row”,B2) -1)/(*n*+1),0,1). Replace *n* with the number of data values. Press Enter.

Step 5:
Click on cell B2. Then click on the lower right hand corner of cell B2 and drag down to create the quantiles.

Step 6:
Make a scatterplot of the data in column B versus the data in column A.

THE VIDEO SOLUTIONS

1. The curve is bell-shaped. It has one peak and is symmetric.

2. You expect a histogram that is roughly symmetric and mound-shaped.

3. You expect the boxplot to be roughly symmetric. The box should be concentrated in the middle of the display. The whiskers to the right and left of the box should be longer than the Q1 to median distance (or median to Q3 distance).

4. The pattern of the dots appears linear.

5. Skewed to the right.

UNIT ACTIVITY SOLUTIONS

1. a. The histogram is skewed to the right.

b. The pattern of the dots in the normal quantile is concave down, which you would expect when the histogram is skewed to the right.

c. No. The histogram is skewed to the right instead of being symmetric. The normal quantile plot is concave down instead of showing a straight-line pattern.

2. a. Sample answer: The histogram is mound-shaped and not quite symmetric. The first bar represents too many data values.

b. Sample answer: The normal quantile plot appears fairly linear. The dots all stay within the curved bands produced by Minitab. The three dots that are almost in a vertical line (corresponding to $v \approx 0$) are the result of the fact that too many data values fall in the first class interval.

c. Sample answer: Given that the pattern of dots is mostly linear, it seems reasonable to assume these data are from a normal distribution.

3. a. Sample answer: There appear to be three possible outliers. The major portion of the histogram is mound-shaped but not symmetric. However, the lack of symmetry may be due to choice of class interval size.

b. Sample answer: The overall pattern of the dots is fairly linear (there may be a slight concave up curvature). However, there are three dots on the plot that are separated from that pattern. These data values appear on the histogram to be outliers.

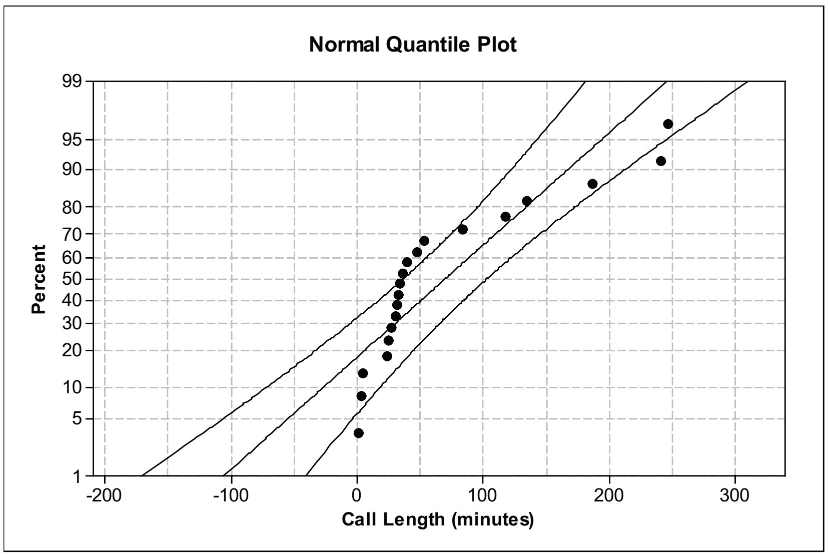
c. Sample answer: The three outliers probably indicate that these data are not normally distributed. (However, it is possible to observe outliers from normal distributions – but it would be unusual to observe three (in a data set of 25) that were this extreme.)

4. a. Sample answer: The histogram has two parts with a gap between them.

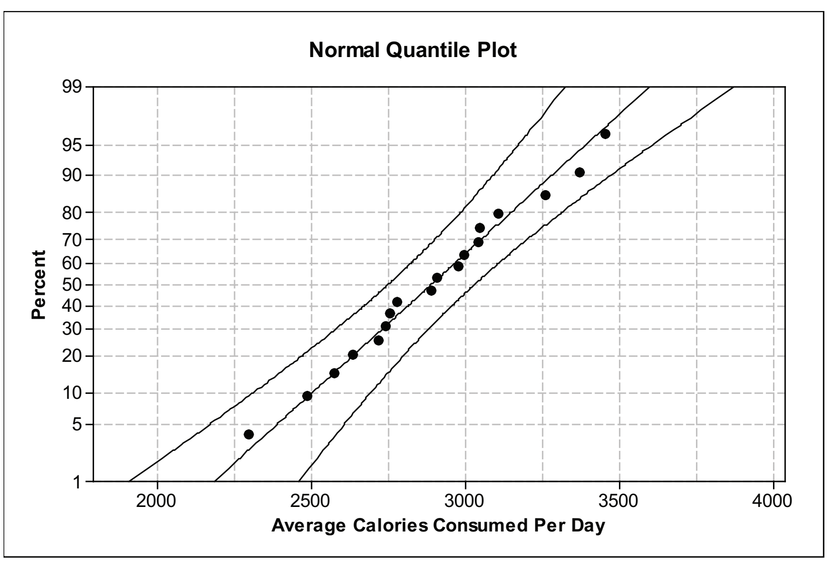
b. Sample answer: The plot appears to be in two pieces that are not aligned. So, the pattern does not appear mostly linear.

c. No. The normal quantile plot does not appear linear. The histogram is not unimodal and symmetric but instead has two peaks, one smaller than the other, with a gap in between.

5. The pattern of the dots in the normal quantile plot is severely curved concave down. Hence, the data are strongly skewed to the right. Hence, it is not reasonable to assume that call lengths are normally distributed.



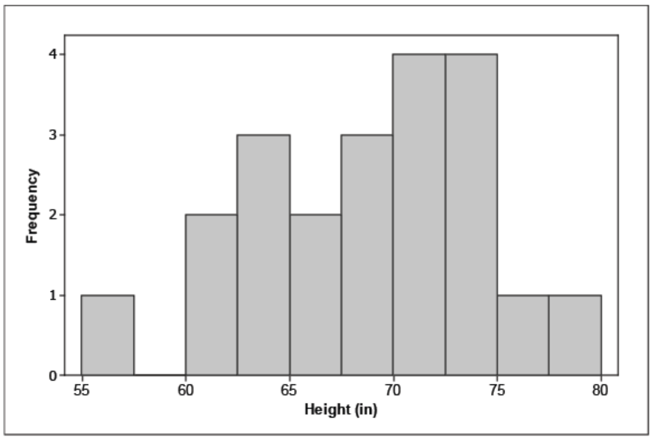
6. The normal quantile plot shows a fairly straight-line pattern. Therefore, it is reasonable to assume that these data are from a normal distribution.

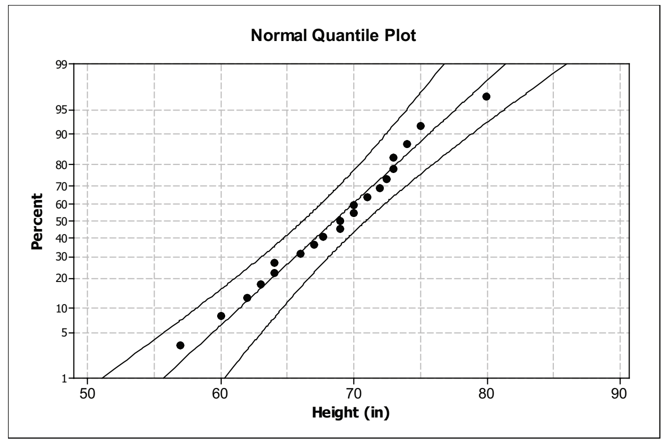


7. Sample answer:

Data on student heights collected from college students enrolled in an introductory statistics course:

62.0	60.0	64.0	69.0	66.0	63.0	64.0	57.0	67.0	72.5	70.0
73.0	69.0	74.0	70.0	71.0	80.0	73.0	75.0	72.0	67.7	

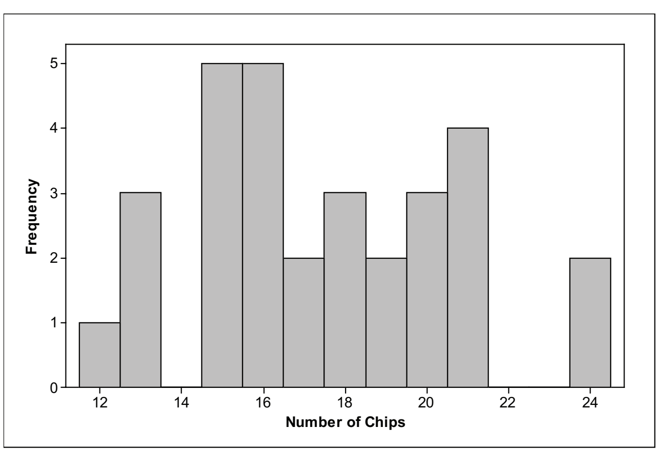


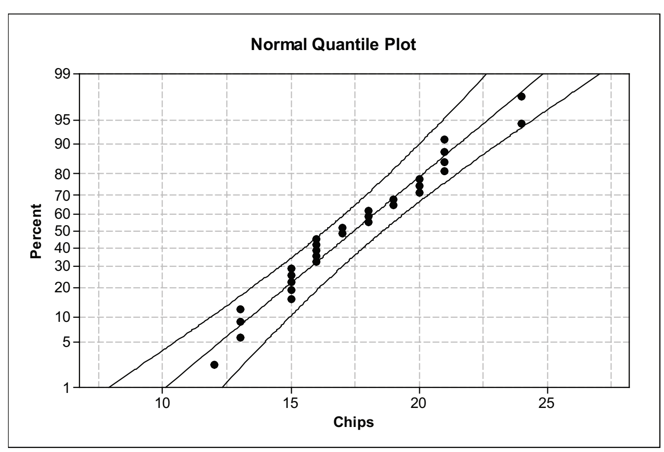


Based on the histogram, it appears that these data might not be normally distributed. However, the normal quantile plot looks pretty straight, so based on that plot, we conclude that it is reasonable to assume heights are normally distributed.

Data on number of chips in Chips Ahoy chocolate chip cookies collected by an introductory statistics class:

17	18	13	15	15	16	16	16	20	17	21	24	18	19	20
18	21	15	15	16	13	15	13	12	20	21	21	24	19	16





Based on the histogram, the data on number of chips do not look at all normal. However, the normal quantile plot tells a somewhat different story. Certainly these data are not perfectly normal. For one thing, the data are counts and as such take on only integer values. Data from a normal distribution can take on values in an interval.

However, the dots in the normal quantile plot stayed within the guide bands produced by Minitab. So, while the pattern in the normal quantile plot is not perfectly linear, they are close enough to being linear that we can conclude these data come from an approximately normal distribution.

EXERCISE SOLUTIONS

1. a. The 5th percentile is -1.645.

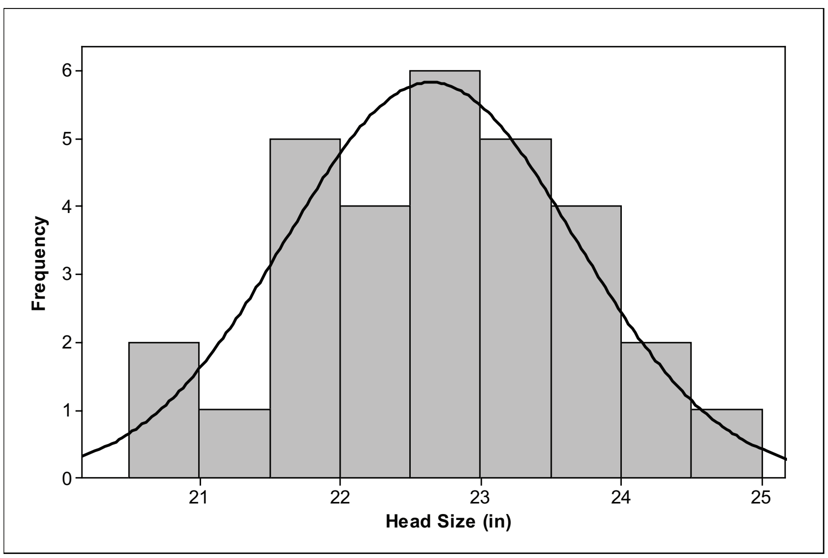
b. The 10th percentile is -1.282.

c. The 90th percentile is 1.282.

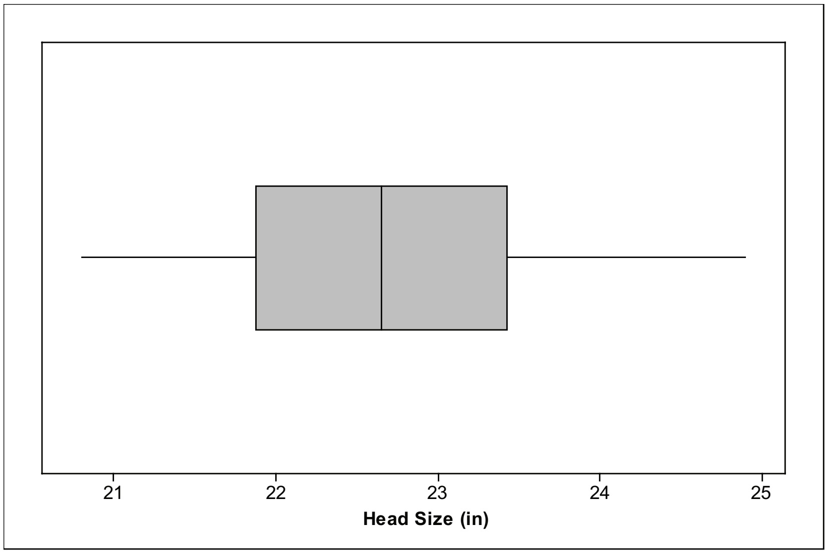
d. The 95th percentile is 1.645.

e. The 5th percentile is the opposite (or negative) of the 95th percentile. The 10th percentile is the opposite of the 90th percentile.

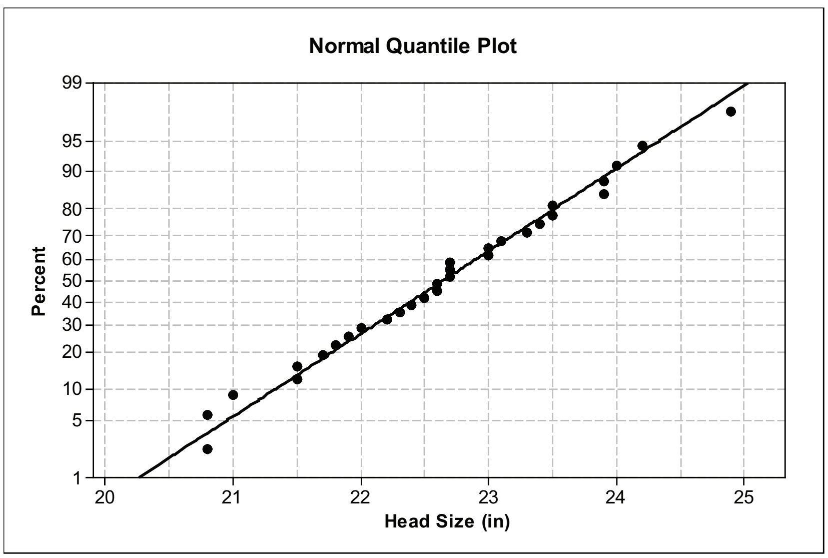
2. a. Sample answer: The histogram below is fairly mound-shaped and symmetric. However, the left side appears closer to the shape of the normal curve than the right side. Still, the fit of the normal curve appears to fit the shape of the histogram reasonably well.



b. The boxplot appears to reasonably represent normal data. Although the right whisker is slightly longer than the left whisker, the plot is fairly symmetric. The whiskers are each a bit longer than the half-width of the box. So, based on this boxplot it seems reasonable to conclude that these data are normally distributed.



c. We created the plot below using Minitab. The dots in the plot hug the line very closely. Therefore, based on this plot, it is reasonable to assume that soldiers' head sizes are normally distributed.



3. a. This histogram has a single peak and is roughly symmetric. The matching normal quantile plot should be fairly linear. So, it matches with Normal Quantile Plot #3 in Figure 9.29.

b. This histogram is skewed to the right. The matching normal quantile plot should show a concave down pattern. So, it matches with Normal Quantile Plot #1 in Figure 9.27.

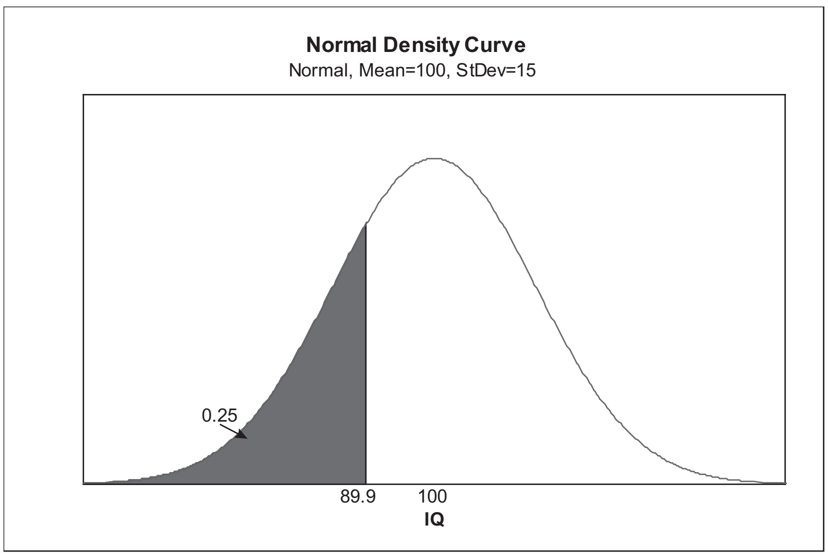
c. This histogram has a lot of data in the tails. It doesn't taper off in the tails the way a normal distribution should. So, its match is Normal Quantile Plot #2 in Figure 9.28.

REVIEW QUESTIONS SOLUTIONS

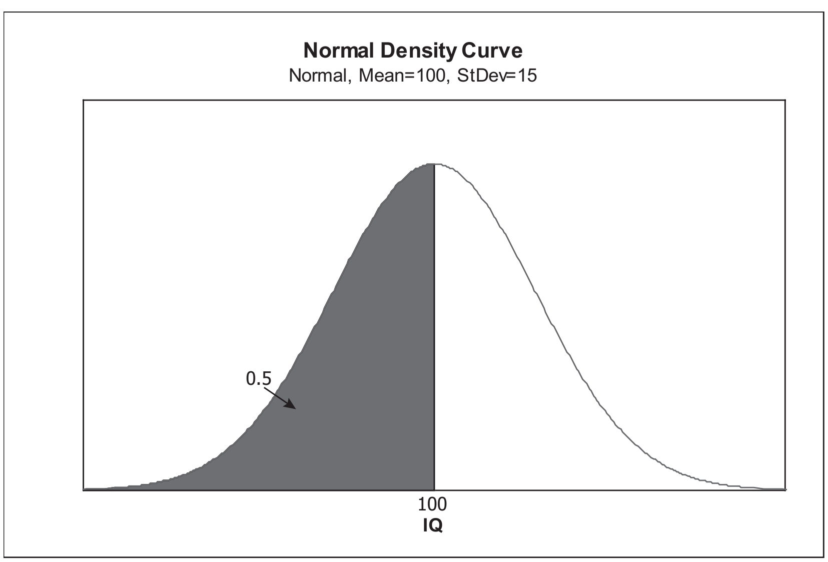
1. a. The quintiles divide the horizontal axis into five intervals. The area under the normal density curve over each of these intervals is 0.20. In other words, 20% of standard normal data will fall between two consecutive quintiles.

b. -0.8416, -0.2533, 0.2533, 0.8416

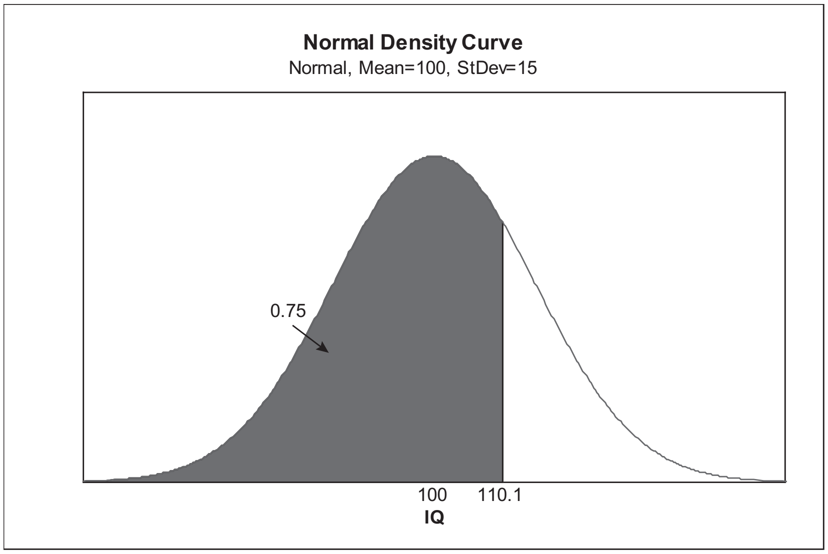
2. a. The 25th percentile for IQ scores is 89.9; 25% of IQ scores fall below this value.



b. Since the normal curve is symmetric about its mean, the 50^{th} percentile is 100.

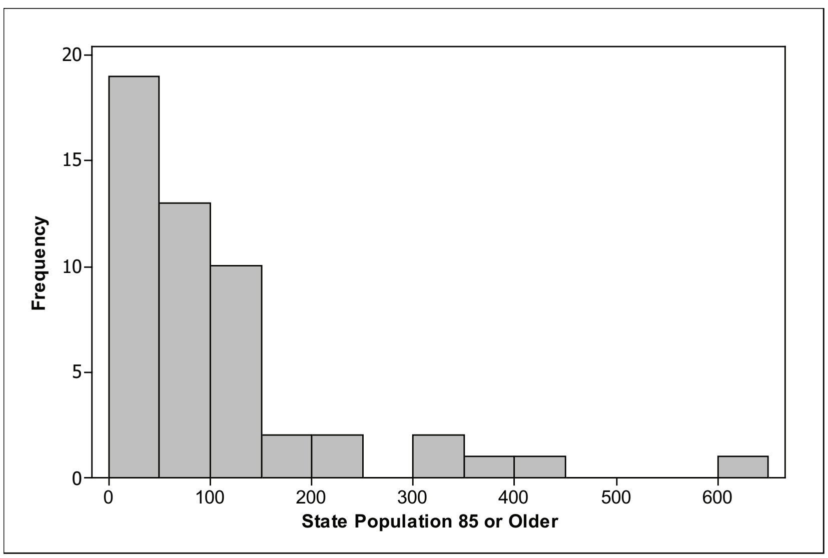


c. The 75th percentile for IQ scores is 110.1; 75% of IQ scores fall below this value.



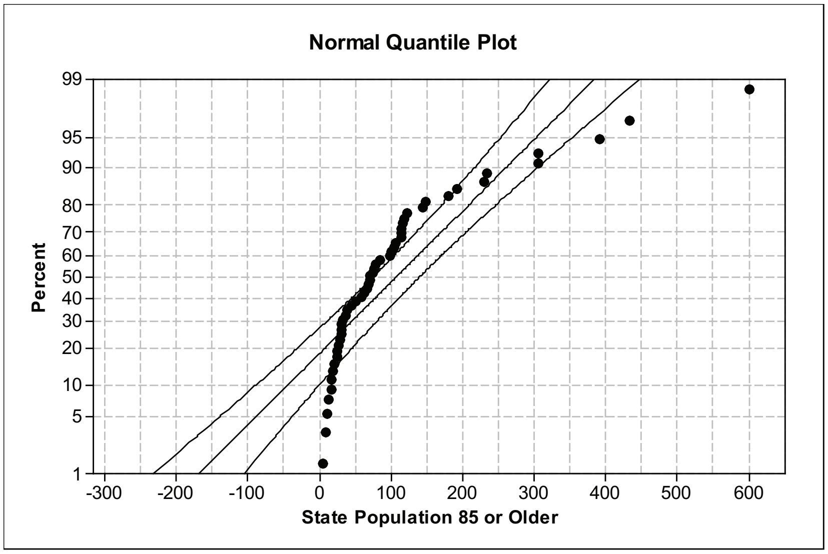
d. They are both approximately 10.1 units away from the mean of 100.

3. a. The histogram of the state populations of residents 85 and older is strongly skewed to the right. It is not reasonable to assume that these data are normally distributed.

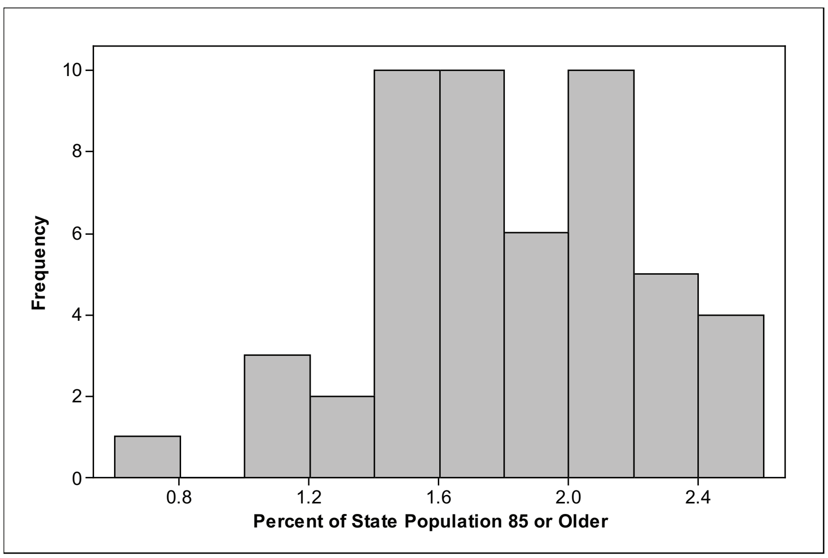


b. Because the shape of the histogram is skewed to the right, the normal quantile plot should be concave down.

c. Since the normal quantile plot below is concave down and not linear, it is not reasonable to assume that population sizes of residents 85 or older are normally distributed.

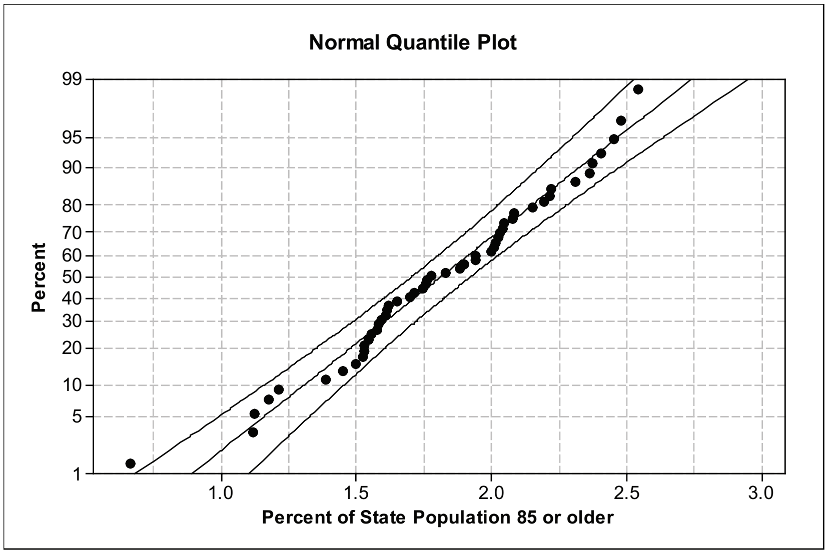


4. a. Sample answer: The histogram is roughly mound-shaped and symmetric. However, there is one valley with two peaks on either side, which may just be the messy nature of real world data. Whether or not it is reasonable to assume these data are approximately normally distributed is somewhat difficult to say. However, these data are closer to having a normal distribution than the population sizes dealt with in question 3.



b. Sample answer: Mostly linear. The data is mound-shaped (in a ragged sort of way) and very roughly symmetric. This could be an example of messy real-world data that is approximately normally distributed.

c. The normal quantile plot is mostly linear. All but one of the dots lie inside the curved bands of the normal probability plot. So, based on this plot, it is reasonable to assume that the percentages are approximately normally distributed.


Normal Quantile Plot
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
1.0
1.5
2.0
2.5
3.0
Percent of State Population 85 or older