Unit 9: Checking Assumptions of Normality

SUMMARY OF VIDEO

Examples of normal distributions can be found in all sorts of settings. Remember normal curves are symmetric and bell-shaped. But in the real world, data can be a lot messier than the idealized examples you might find in a textbook. At times it can be difficult to eyeball whether or not data are normally distributed. Down the road, it will be important to feel confident in our assumption that a particular data distribution is, in fact, normal because that is a prerequisite for some more advanced statistical techniques.

We begin our study into methods of sorting out normal data from data that is not normal at Pete & Gerry's organic egg farm. Plenty of data get collected at the egg farm: for example, how much water and feed the birds are eating, how they are growing, how many and what sizes of eggs they are laying, and how production is running on the packaging line. Are any of the data distributions they see normal?

Let's start by taking a look at the weights of 7-week-old hens in one flock. Pete & Gerry's Jesse Laflamme explains that they get them as day-old chicks and they are essentially the same size. But as they grow, Pete & Gerry's test the young hens by taking a random sample of around 100 pullets. They weigh the hens and then graph the weights from the sample. They expect the graph to have the shape of a normal curve with a target weight of where they expect the middle of that curve to be. Figure 9.1 shows a histogram of the weights from one sample.

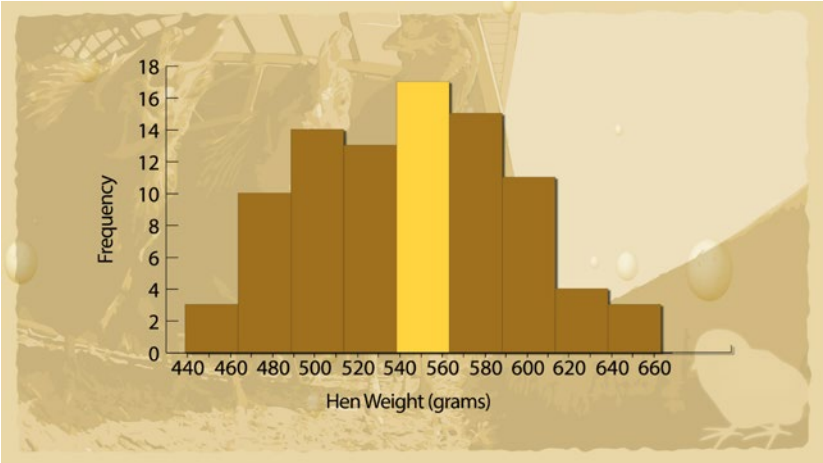


Figure 9.1. Weights of 7-week-old hens.

The histogram for this flock does appear to be normal with the one peak in the middle at around 550 grams. The normal curve drawn over the histogram in Figure 9.2 seems to be a pretty good fit.

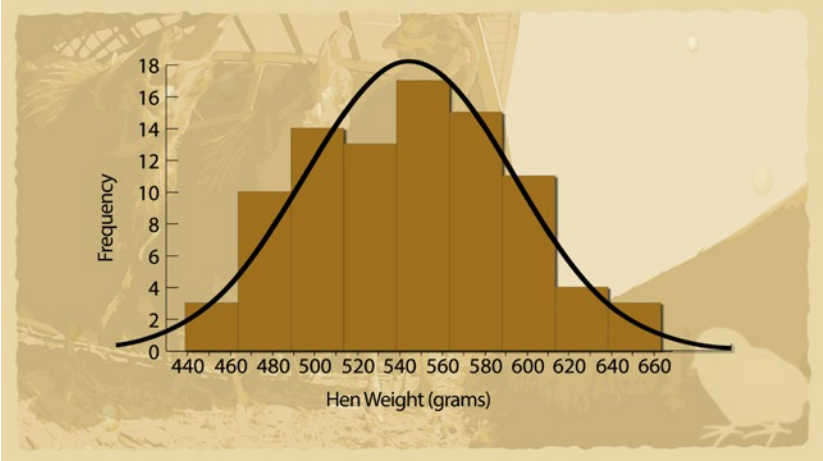


Figure 9.2. Overlaying a normal curve.

It's important to consider the class size when we are eyeballing a histogram to see if the data are normal. Sometimes, changing the class size can really change the way the histogram looks and what once appeared perfectly bell-shaped now looks quite different. The histogram of hen weights in Figure 9.3 looks less like normal data than the histogram of the same data in Figure 9.2.

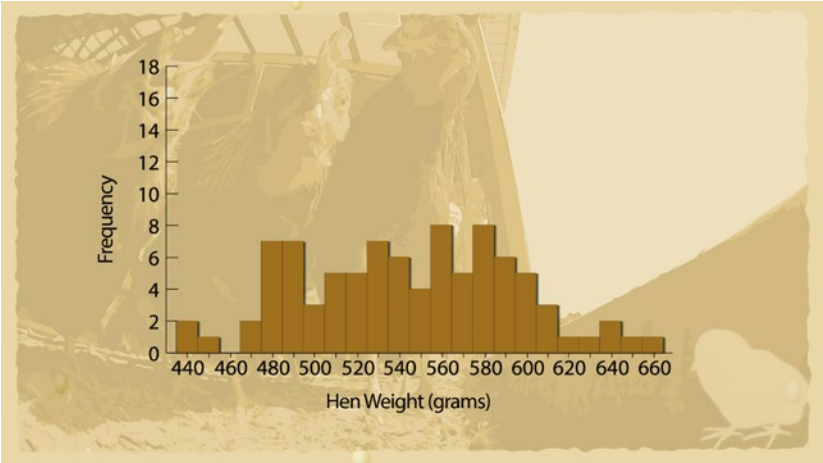


Figure 9.3. Changing the class size.

A second option as we assess normality is to use the same hen weights to construct a boxplot. Boxplots can act as another graphical display test to see if our data are normally distributed. If a distribution is normal, we would expect to see the box containing the middle 50% of the data to be pretty tightly grouped in the center of the distribution, with longer whiskers indicating the increased spread of the upper and lower quarters of the data. In Figure 9.4, take a look at how a truly normal distribution translates into a boxplot and compare it with our hens. The weight distribution does appear to be approximately normal, with the whiskers each longer than the Q1 to median distance and the median to Q3 distance.

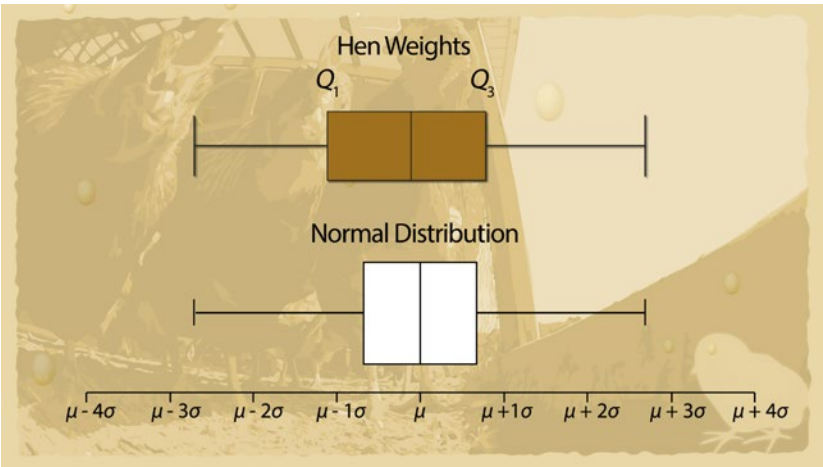


Figure 9.4. Comparing histograms of hen weight and normal data.

Going beyond eyeballing these kinds of graphic displays, there is another more precise way to check whether a distribution is approximately normal. Statisticians use software to construct what is known as a normal quantile plot. The basic idea is to compare the ordered data values you have with values you would expect from a standard normal distribution. If your data are normally distributed, the normal-quantile-plot points will fall close to a straight line.

Since a computer will do the work for you, it is less important to understand the steps taken to construct the normal quantile plot than it is to know how to interpret it. Figure 9.5 shows the normal quantile plot for the hen weights. Our observed weights are on the *x*-axis and the expected values on the *y*-axis. The pattern of dots in the plot lies close to a straight line. So we can conclude that our data come from an approximate normal distribution.

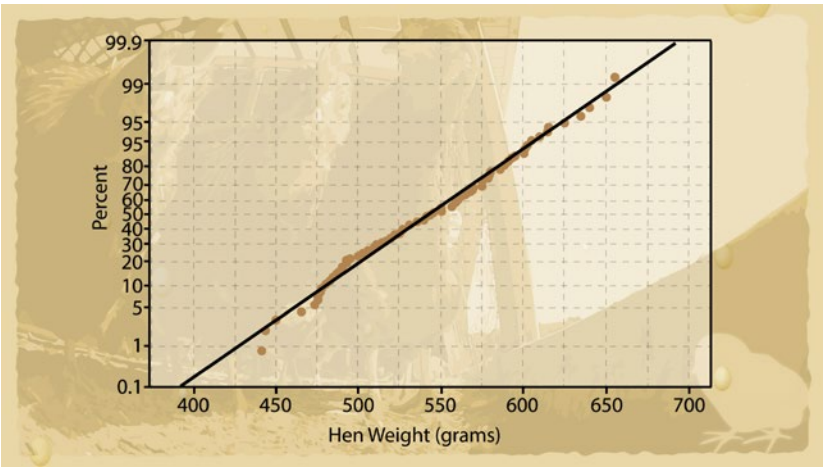


Figure 9.5. Normal quantile plot for hen weights.

Next, we try out these tests with another size range – egg weights. In the wild, we would expect the size of one bird species' eggs to be normally distributed. Figure 9.6 shows a histogram of the egg weights for a day's worth of eggs from Pete & Gerry's.

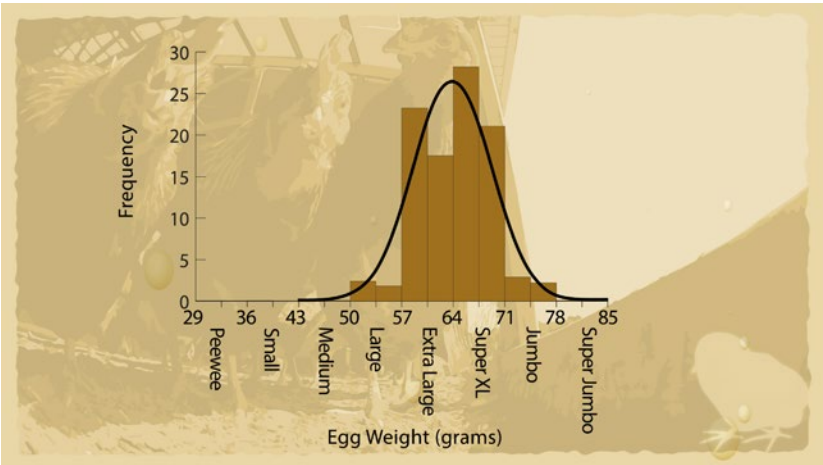


Figure 9.6. Egg weights from a week's worth of data from Pete & Gerry's.

It is a little challenging to decide if Pete & Gerry's egg size distribution looks normal. Could this be an example of messy real-world numbers? Or has the farm's careful control over hatching and breeding had an influence on the size range? To find out, we investigate using the normal quantile plot for these data, which is shown in Figure 9.7.

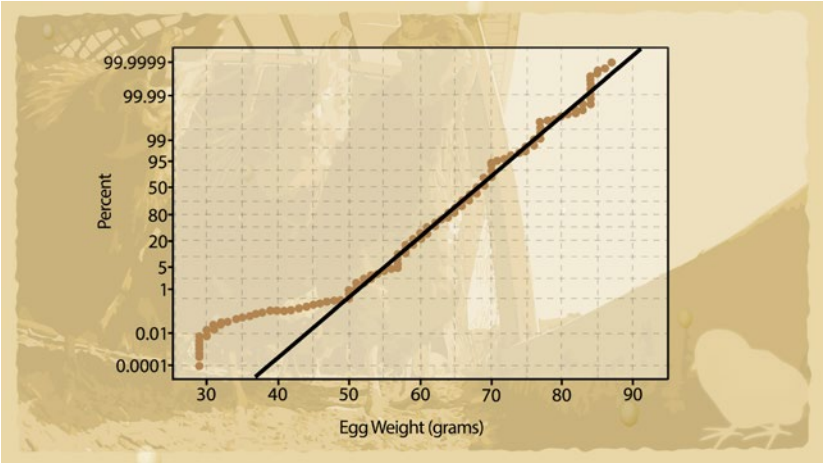


Figure 9.7. Normal quantile plot of egg weight.

The middle of the plot looks like what you would expect from normal data (roughly a straight-line pattern). But the lower tail of the distribution does not look at all linear. That lower tail shows us that we have more eggs at the lower weight range than we would expect to see if the distribution were actually normal. We can conclude that the egg size distribution at Pete & Gerry's is not normal, at least on the day that these data were collected. That is logical if you understand that the size of the eggs a chicken lays increases over her lifetime. The egg business has seasonal cycles.

For instance, sales increase for the year-end holidays when people are stocking up for their baking needs. Pete & Gerry's tries to prepare for those cycles by knowing the age of the hens they will need laying to meet that heightened demand for the most desirable egg sizes, large and extra large. On the day these data were recorded, months before the peak season when the demand for large eggs is highest, there were more younger flocks laying smaller eggs.

As you get more familiar with normal quantile plots, you might start to recognize predictable patterns for various non-normal distributions. For instance, Figure 9.8 shows a histogram that is skewed to the right, indicating that there were a majority of very young birds laying smaller eggs. The normal quantile plot in this case, shown in Figure 9.9, is curved with a concave down pattern.

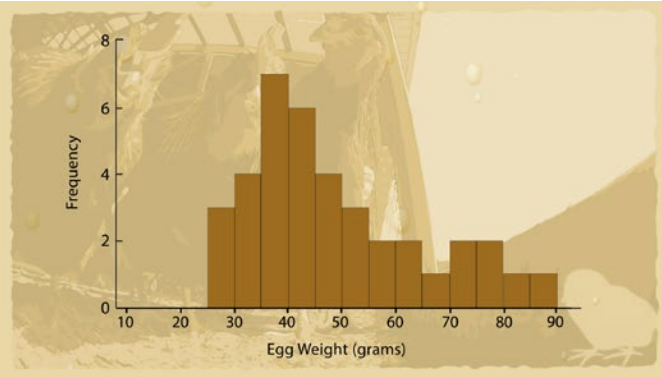


Figure 9.8. Histogram skewed to right.

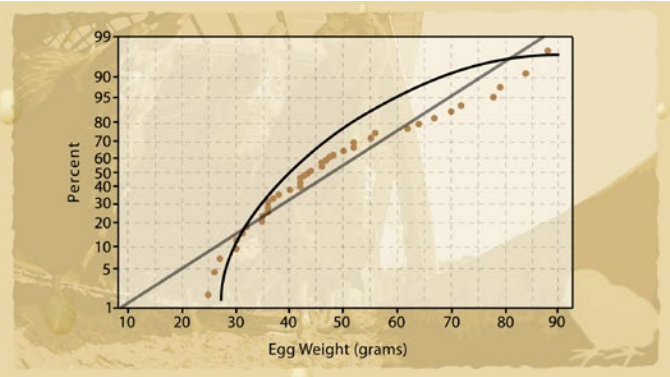
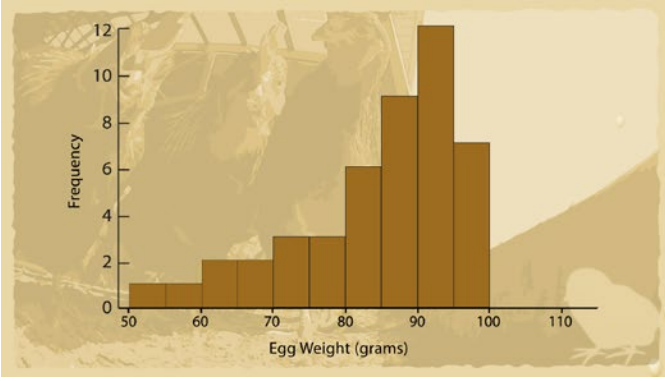


Figure 9.9. Normal quantile plot concave down.

On the other extreme, Figure 9.10 shows how things might look for a weight distribution of eggs laid by much older chickens. The histogram would be skewed to the left. The normal quantile plot is again curved but this time it is concave up.

Figure 9.10. Histogram skewed left.



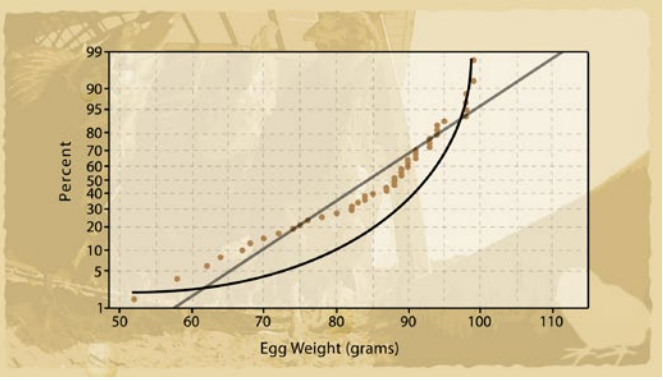


Figure 9.11. Normal quantile plot concave up.

So, don't count your chickens before they hatch. Or, in this case, don't count your eggs before they are laid, at least when it comes to assuming data are normal!

STUDENT LEARNING OBJECTIVES

A. Be familiar with the characteristic shape of histograms and boxplots of normal data.

B. Know how to calculate percentiles from a normal distribution.

C. Understand how normal *n*-quantiles are used to divide the area under a normal curve into *n*-equal areas.

D. Understand the basic construction of a normal quantile plot.

E. Use a normal quantile plot to assess whether data are from a normal distribution.

CONTENT OVERVIEW

Normal distributions are a very important class of statistical distributions and certainly the most common distributions you will encounter in this course. Data produced by many natural processes can be described as normally distributed. Some examples include human bone lengths, people's IQ scores, bird weights, and flower lengths. Later in this course, you will encounter statistical procedures that are based on an assumption of normality. So, how do we decide whether it is reasonable to assume data come from a normal distribution?

Not all mound-shaped, symmetric data turn out to be normally distributed. Furthermore, histograms and boxplots of normal data don't always fit the pattern we expect. For example, take a look at the histogram and boxplot in Figures 9.12 and 9.13.

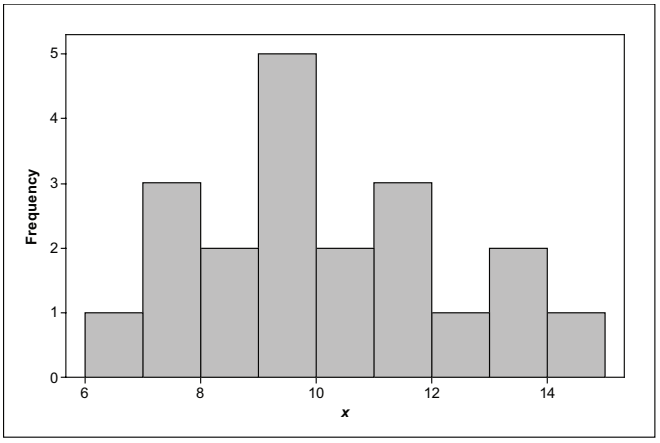


Figure 9.12. Histogram of normal data.

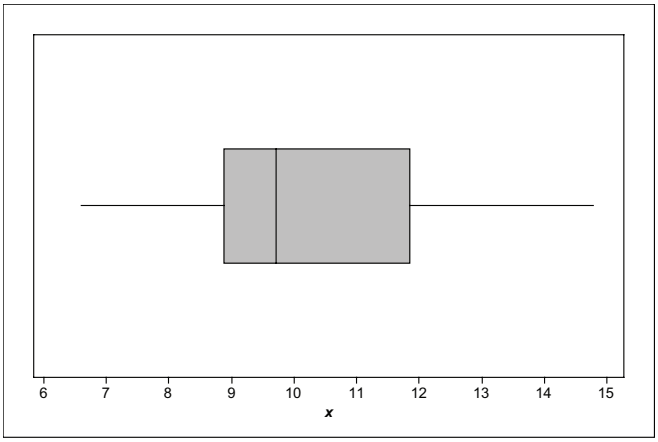


Figure 9.13. Boxplot of normal data.

The histogram looks more ragged than the characteristic mound-shaped, symmetric histogram we might expect from normal data. From the boxplot, the inner 50% of the data do not appear symmetric about the median, again a characteristic we would expect from normal data. So, we might conclude that these data do not come from a normal distribution – but we would be wrong! Particularly when the sample size is small, here n = 20, histograms and boxplots may not provide sufficient information to decide whether or not data are normally distributed. So, we need another tool to differentiate normal data from non-normal data – a **normal quantile plot**. This plot compares the ordered data with what we would expect from perfectly normal data.

Generally, we use technology to create normal quantile plots, which are difficult and tedious to construct by hand. However, we will discuss a simplified version of normal quantile plots. But first, we need some background on percentiles and quantiles.

Finding Percentiles

In Unit 8, Normal Calculations, we discussed how to use a *z*-table to find the proportion of standard normal data that would fall below a specific value of *z*. That was done by finding the area under the standard normal curve that lies to the left of that specific *z*-value. For example, if we use the partial *z*-table that appears in Figure 8.5 of Unit 8, we learn that the area to the left of $z = 1.25$ is 0.8944. So, we know that around 89.44% of standard normal data falls below 1.25. (See Figure 9.14.)

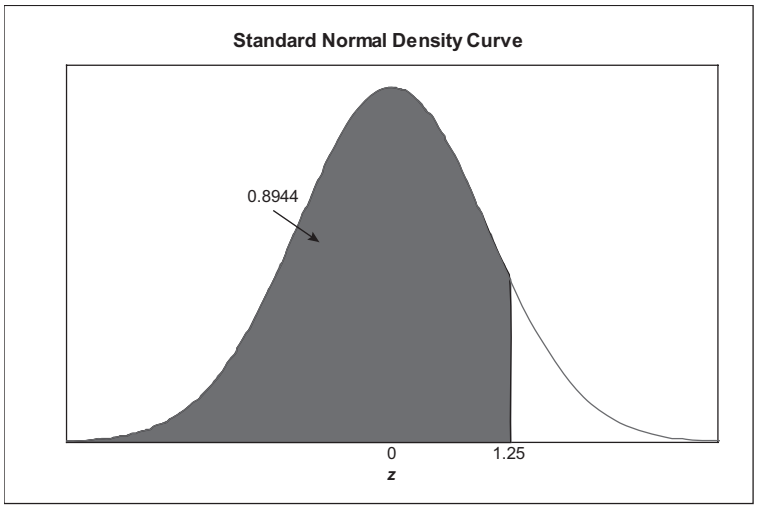


Figure 9.14. Finding the area that lies to the left of 1.25.

For percentiles, we work this process in reverse – we start with the area or percentage, and then determine the corresponding value of *z*, which we call a **percentile**. For example, finding the 50th percentile of a standard normal distribution is easy – that's the value of z for which the area under the density curve to its left is 0.50: $z = 0$. To find the 70th percentile of a standard normal distribution, we need to find the *z*-value such that the area under the curve to the left of that *z*-value is 0.70. Table 9.1 shows a portion of the *z*-table. Start in the body of the table (which contains the areas) and find numbers as close to 0.70 as possible: 0.6985 and 0.7019. Then read off the corresponding *z*-values: 0.52 and 0.53. So, we know the 70th percentile is between $z = 0.52$ and $z = 0.53$.

z	0	0.01	0.02	0.03	0.04
0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389

Table. 9.1. Partial z-table.

Using technology to find percentiles is easier and more accurate than using the *z*-table. Just use the inverse function of what you used to find the proportions (areas). (For example, in Excel use Norm.Dist to find the area to the left of $z = 1.25$ and use Norm.Inv to find the 70th percentile, the *z*-value associated with area 0.70.) Figure 9.15 shows the result using Minitab's Probability Distribution Plot, which gives $z = 0.5244$ as the 70th percentile of the standard normal distribution.

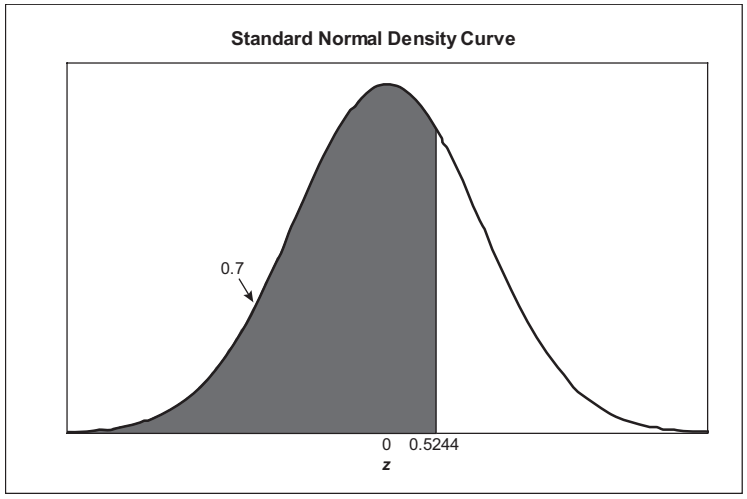


Figure 9.15. Specify area 0.7 and read off percentile z = 0.5244.

Finding Normal Quantiles

Quantiles are points of a distribution taken at regular percentile intervals. Suppose we want to find the 10-quantiles, or deciles, of the standard normal distribution. That means we need to determine 9 *z*-values that divide the *x*-axis into 10 intervals such that the areas over consecutive intervals are equal. In this case, the area is 0.10. So, to find the deciles, we need to find the 10^{th}, 20^{th}, 30^{th}, . . . 90^{th} percentiles, which are:

-1.282 -0.842 -0.524 -0.253 0.000 0.253 0.524 0.842 1.282

Figure 9.16 shows the 10-quartiles dividing the area under the standard normal curve into equal areas.

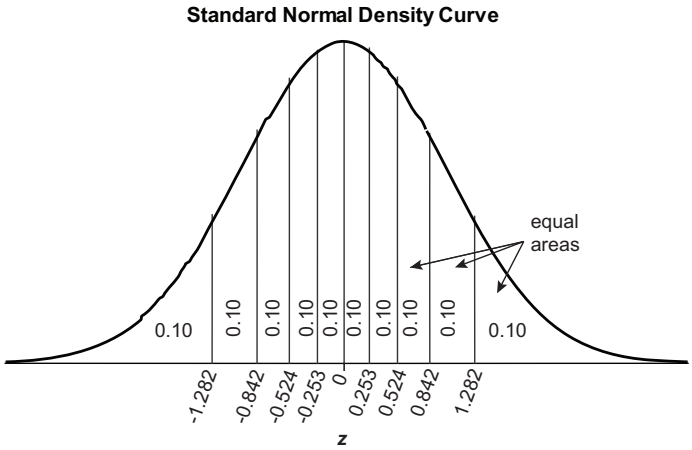


Figure 9.16. Deciles of standard normal distribution.

Normal Quantile Plot

A normal quantile plot is a graphical method for assessing whether data come from a normal distribution. The plot compares the ordered data with what would be expected of perfectly normal data, which in a simplified version of the plot will be the normal quantiles. If the data are normally distributed, then the dots on a normal quantile plot will fall close to a straight line. We'll test this out with the following data:

4.45 5.76 5.81 5.34 4.68 5.15 6.25 4.88 4.19

Next, we construct a normal quantile plot in order to decide whether these data come from a normal distribution. Here are the instructions for making a simplified version of a normal quantile plot.

Constructing a Simplified Normal Quantile Plot

Step 1: Given a sample of n data values, order the data from smallest to largest.

Step 2. Find the (n+1)-quantiles. This gives n z-values that divide the area under the normal density curve into n + 1 equal areas of size 1/(n+1).

Step 3. Plot the quantiles versus the ordered data values.

We begin construction of the normal quantile plot by ordering the data from smallest to largest:

4.19 4.45 4.68 4.88 5.15 5.34 5.76 5.81 6.25

Since there are nine data values, we find the (9 + 1)-quantiles of a standard normal distribution, which we have already done (See Figure 9.16.). The final step is to plot the normal quantiles versus the ordered data. In other words, plot (4.19, -1.282), . . . , (6.25, 1.282), which gives us the plot in Figure 9.17.

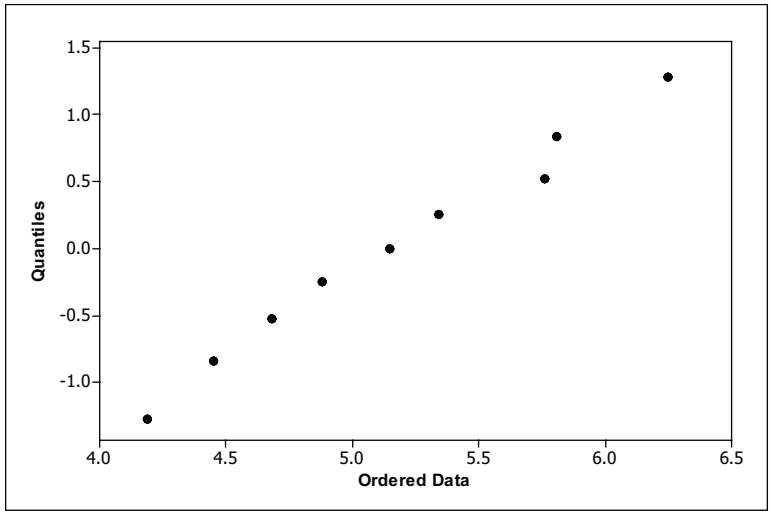


Figure 9.17. Simplified normal quantile plot.

Now, compare Figure 9.17 with the normal quantile plot in Figure 9.18, which was constructed using the statistical software Minitab. The patterns of the dots in the two plots are similar even though the scaling on the vertical axes differs. Minitab adds a line to help us see whether the pattern is linear. In addition, Minitab adds some curved bands that help us decide when the points are straying too far from the line compared to what is expected of normal data.

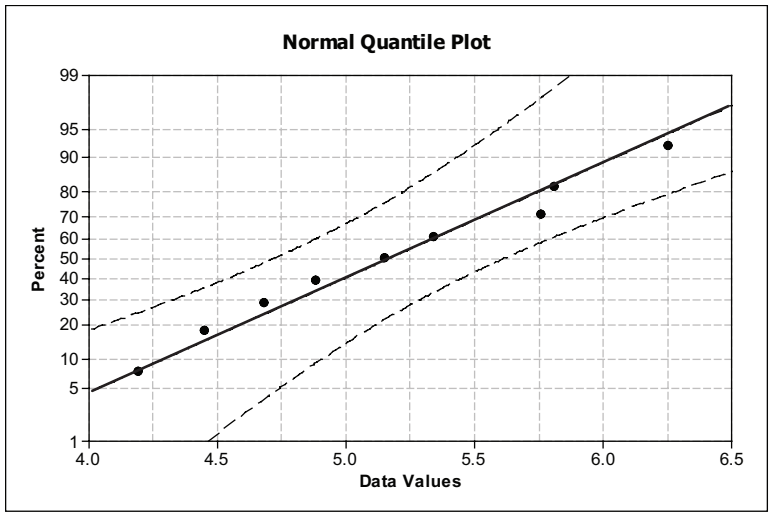


Figure 9.18. Normal quantile plot using Minitab.

Based on the normal quantile plot (either Figure 9.17 or Figure 9.18), it seems reasonable to assume these data are normally distributed.

For data that are not normal, the normal quantile plot can help us determine how the data deviate from normality. For example, normal quantile plots can tell us that data are skewed. In Figures 9.8 – 9.11, we see that if data are skewed to the right, then normal quantile plots are concave down, and if data are skewed to the left, then normal quantile plots are concave up. Instead, suppose data come from the uniform distribution pictured in Figure 9.19, which compares a uniform density curve to a normal density curve with the same mean.

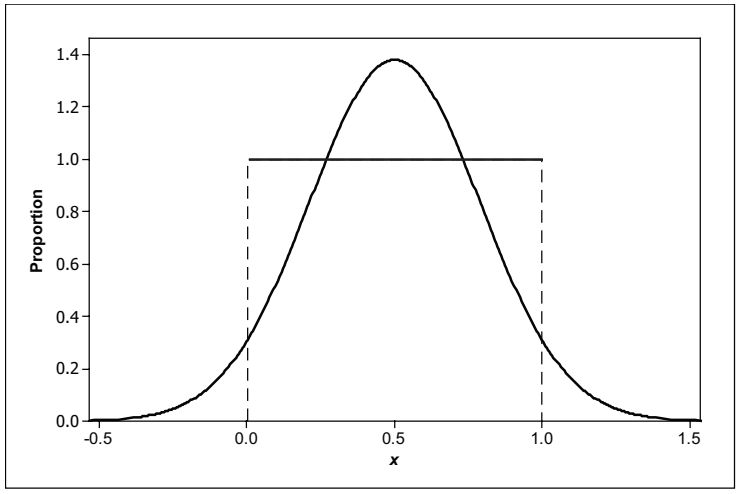


Figure 9.19. Comparing a normal density curve to a uniform density curve.

Data from this uniform distribution will be fairly evenly spread out between 0 and 1. There are no values out in the tails (data values less than 0 or greater than 1) as you would expect from normal data. So, how does the lack of data in the tails show up in a normal quantile plot? Take a look at Figure 9.20, which is a normal quantile plot for 20 data values from this uniform distribution.

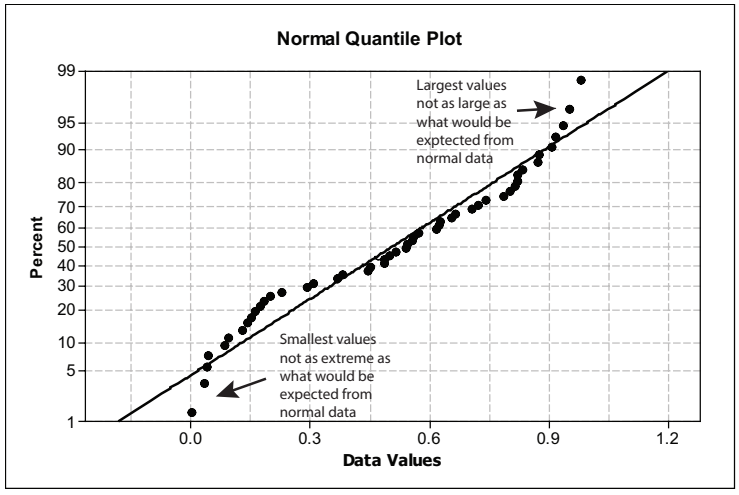


Figure 9.20. Normal quantile plot of data from a uniform distribution.

In the middle of the plot, the dots stay close to the line and hence are similar to what we would expect from normal data. However, the data at the extremes, the smallest and largest data values, are associated with curved patterns in the normal quantile plot. Consider the largest data value, marked with the double circle. The *x*-value associated with this point is 0.98. In order for this point to lie on the line, that *x*-value would need to be about 1.2. So, the largest uniform data values are too small compared to what we would expect from normal data values, which explains the pattern of an upward bend at the right end of the graph. On the other extreme, the smallest uniform data values are too large compared to what we would expect from normal data values, which explains the downward bend at the left end of the graph.

KEY TERMS

A **percentile** of a distribution is a value such that a certain percentage of observations from the distribution fall at or below that value. For example, a 16th percentile of a standard normal distribution is a value such that 16% of the area under the standard normal curve falls at or below that value. According to the 68-95-99.7% Rule, that value should be around $z = -1$, as shown in Figure 9.21.

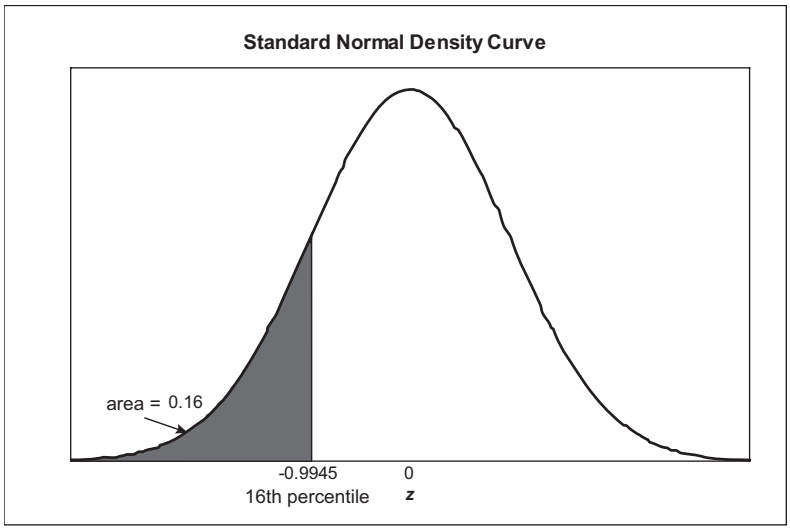


Figure 9.21. The 16th percentile of a standard normal distribution.

Quantiles are points of a distribution taken at regular percentile intervals. For example, the 4-quantiles, or quartiles, are the 25th, 50th, and 75th percentiles. The quartiles for a standard normal distribution are shown in Figure 9.22.

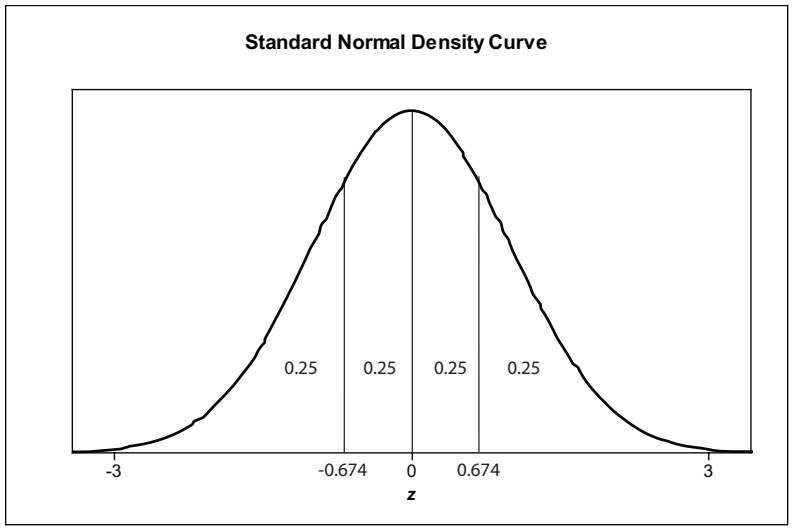


Figure 9.22. Quartiles of a standard normal distribution.

A **normal quantile plot** (also known as a **normal probability plot**) is a graphical method for assessing whether data come from a normal distribution. The plot compares the ordered data with what would be expected of perfectly normal data. A fairly linear pattern in a normal quantile plot suggests that it is reasonable to assume that the data come from a normal distribution.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is the shape of a normal distribution curve?

2. What characteristics do you expect to see in a histogram of normal data?

3. What characteristics do you expect to see in a boxplot of normal data?

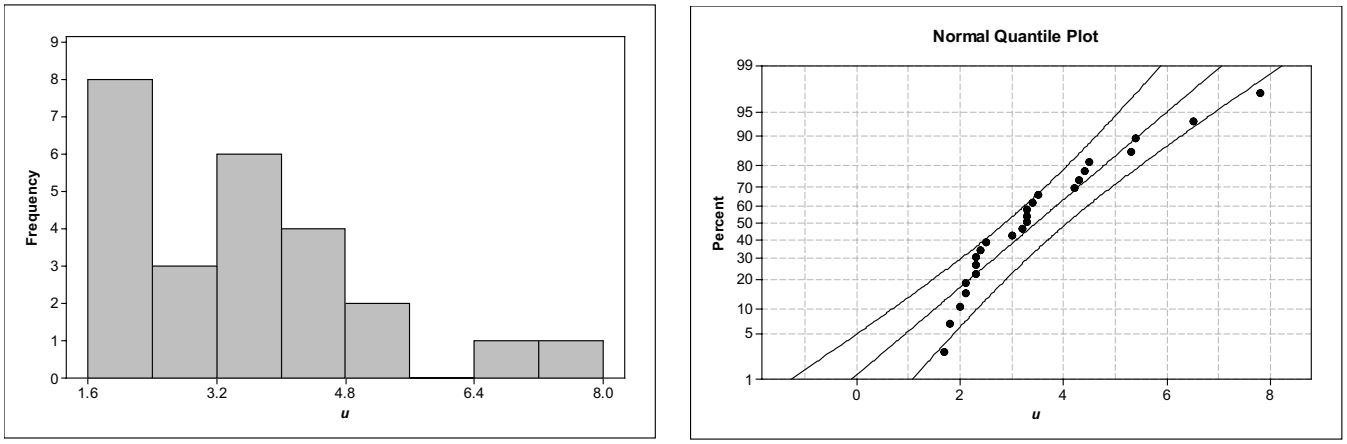
4. What pattern in a normal quantile plot tells you that data come from a normal distribution?

5. Suppose a normal quantile plot has a curved, concave down pattern. Would you expect a histogram of the data to be symmetric, skewed to the right, or skewed to the left?

UNIT ACTIVITY:
WORKING WITH NORMAL QUANTILE PLOTS

In questions 1 – 4 you are given a histogram and normal quantile plot for each of four datasets. Compare the histograms to the normal quantile plots for each dataset. These activity questions should give you a better idea of how to interpret patterns in normal quantile plots. For the rest of the activity, you will construct normal quantile plots to assess whether data distributions are normal.

1. Figure 9.23 shows a histogram and normal quantile plot for 20 data values collected on a variable *u*.



(a) *(b)*

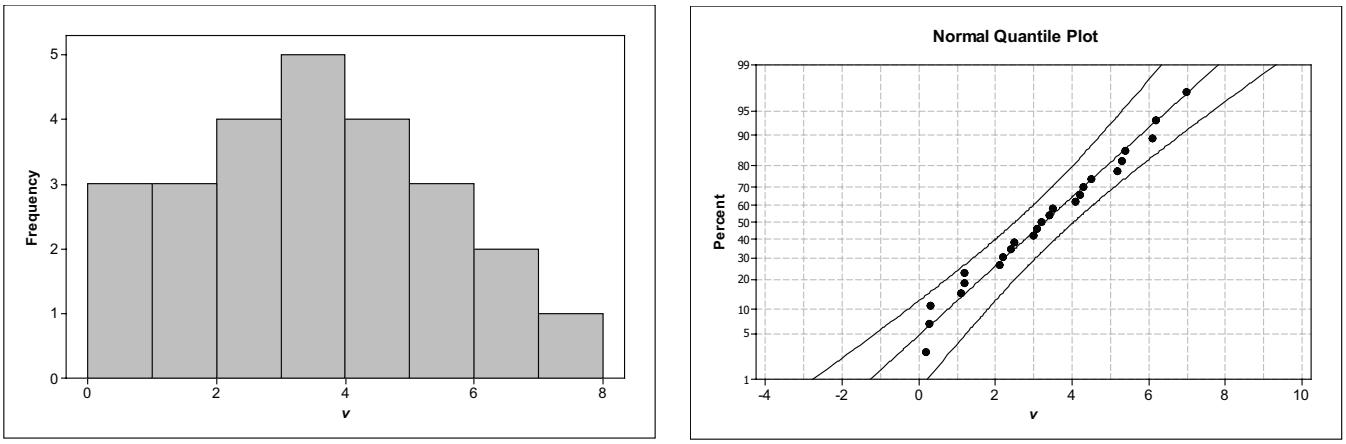
Figure 9.23. Histogram (a) and normal quantile plot (b) for 20 data values.

a. Describe the shape of the histogram.

b. Does the pattern of the dots in the normal quantile plot look fairly straight? If not, describe the general overall pattern of the dots, or if the overall pattern is straight, describe departures from the straight-line pattern.

c. Do you think it is reasonable to assume these data come from a normal distribution? Explain.

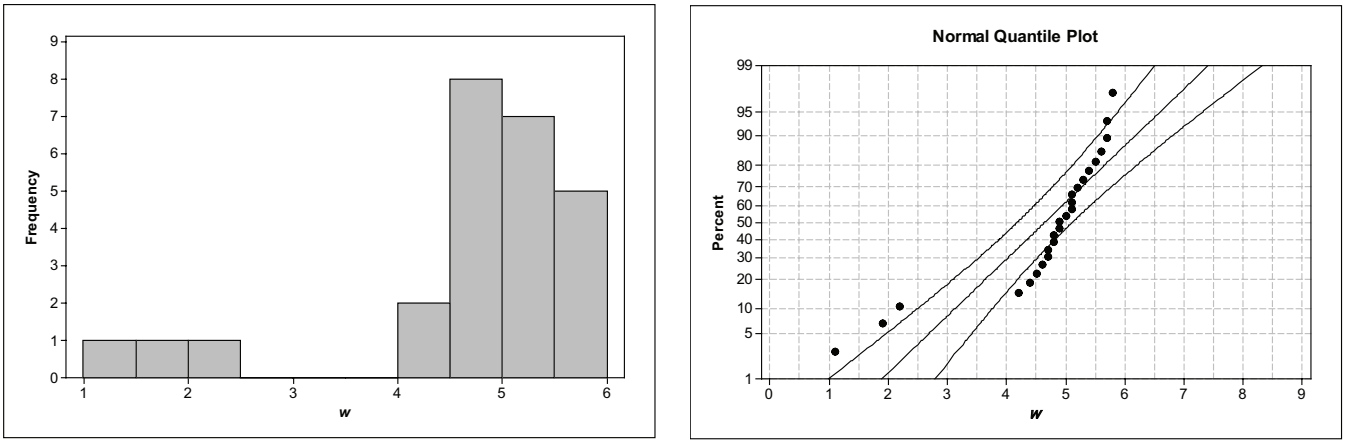
2. Figure 9.24 shows a histogram and normal quantile plot for 20 data values collected on a variable *v*. Repeat question 1 for variable *v*.



(a) *(b)*

Figure 9.24. Histogram (a) and normal quantile plot (b) for 20 data values.

3. Figure 9.25 shows a histogram and normal quantile plot for 20 data values collected on a variable *w*. Repeat question 1 for variable *w*.



(a) *(b)*

Figure 9.25. Histogram (a) and normal quantile plot (b) for 20 data values.

4. Figure 9.26 shows a histogram and normal quantile plot for 20 data values collected on a variable *x*. Repeat question 1 for variable *x*.

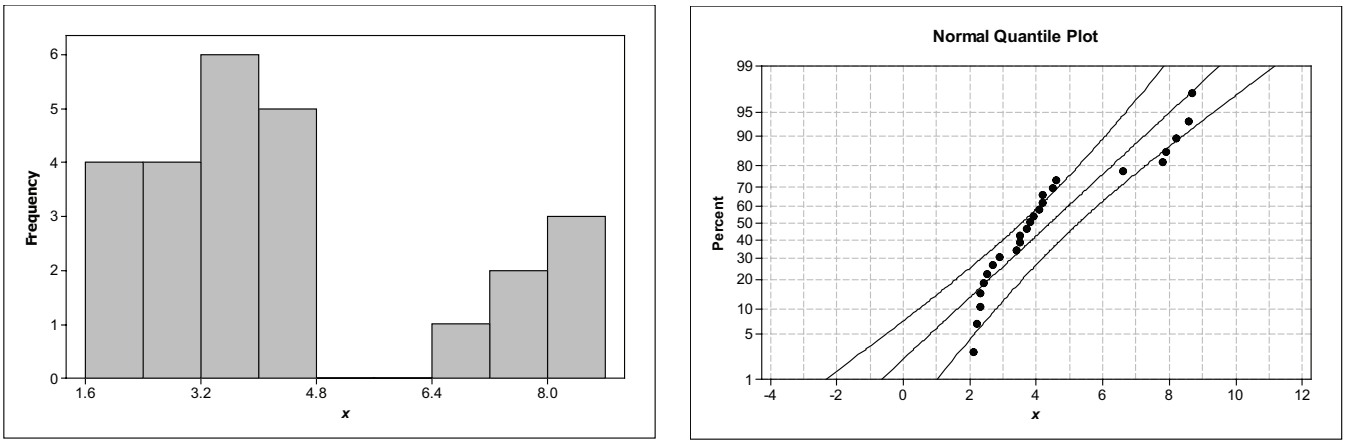


Figure 9.26. Histogram (a) and normal quantile plot (b) for 20 data values.

Next, you will decide whether it is reasonable to assume that data collected from two studies are normally distributed.

5. The length in minutes of 20 calls made to a call center are given below:

25	27	135	36	4	83	39	34	233	
33	241	186	246	53	30	47	32	118	1

Construct a normal quantile plot for these data. Based on your plot, is it reasonable to assume that call lengths are normally distributed? Explain.

6. The average daily calories consumed by a sample of men are as follows:

2716	2754	2484	2995	2635	2296	2741	3262	2572
3371	2778	3041	3045	2888	2908	3457	3109	2977

Construct a normal quantile plot for these data. Based on your plot, is it reasonable to assume that the average daily calories consumed by men are normally distributed? Explain.

7. Collect some data on two or more quantitative variables. The data can be from your class, from an online source (such as sports statistics), or from another source. Make histograms and normal quantile plots for each dataset. Decide which of your data could be described as approximately normally distributed.

EXERCISES

1. Find the following percentiles for a standard normal distribution.

a. 5th percentile.

b. 10th percentile.

c. 90th percentile.

d. 95th percentile.

e. What is the relationship between the 5th and 95th percentile? What is the relationship between the 10th and 90th percentile?

2. For data from a normal distribution, $\bar{x}$ and s are appropriate measures of center and spread, respectively. However, for non-normal data, it is better to use a five-number summary to describe center and spread. In Unit 6, you were asked to compute $\bar{x}$ and s for data on soldiers' head size.

23.0	22.2	21.7	22.0	22.3	22.6
22.7	21.5	22.7	24.9	20.8	23.3
24.2	23.5	23.9	23.4	20.8	21.5
23.0	24.0	22.7	22.6	23.9	21.8
23.1	21.9	21.0	22.4	23.5	22.5

a. Draw a histogram of these data. Then sketch a normal curve over the histogram. Based on your histogram, does it appear reasonable that these data come from a normal distribution? Explain.

b. Represent these data with a boxplot. Describe what features of the boxplot would lead you to believe that these data are normally distributed.

c. Not all data that is mound-shaped and roughly symmetric is normal. Make a normal quantile plot of these data. Based on your plot, is it reasonable to assume these data come from a normal distribution? Explain.

3. Three normal quantile plots are displayed in Figures 9.27 – 9.29. In parts (a) – (c) match the histogram of 50 data values to the normal quantile plot.

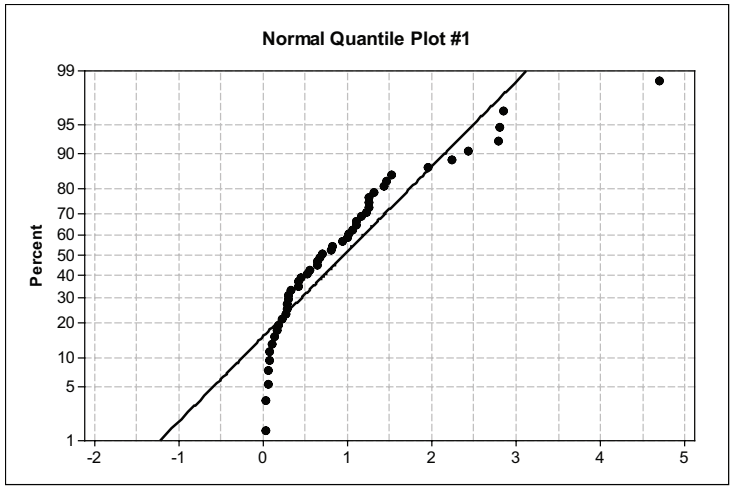


Figure 9.27. Normal Quantile Plot #1.

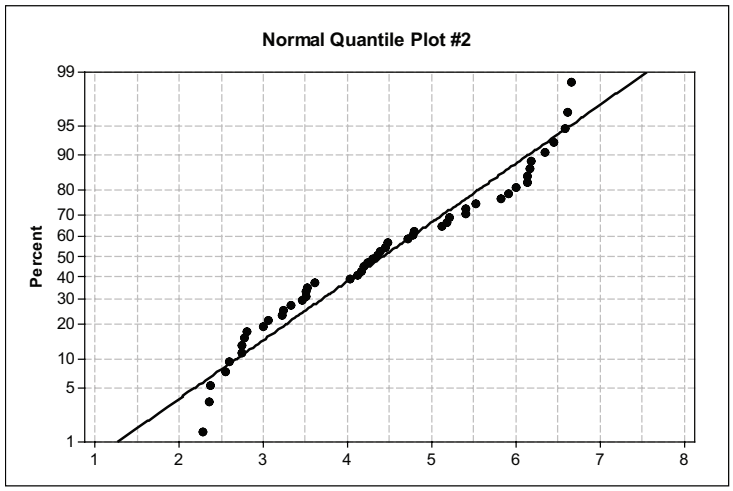


Figure 9.28. Normal Quantile Plot #2.

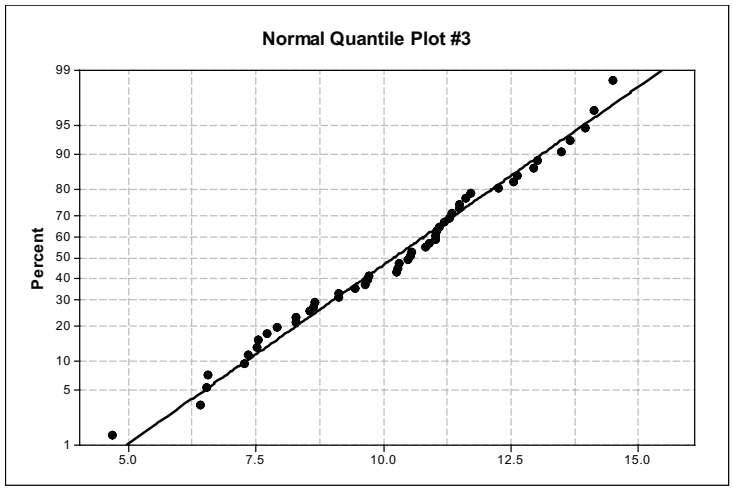


Figure 9.29. Normal Quantile Plot #3.

a. Match the histogram of data on variable *v* (Figure 9.30) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.

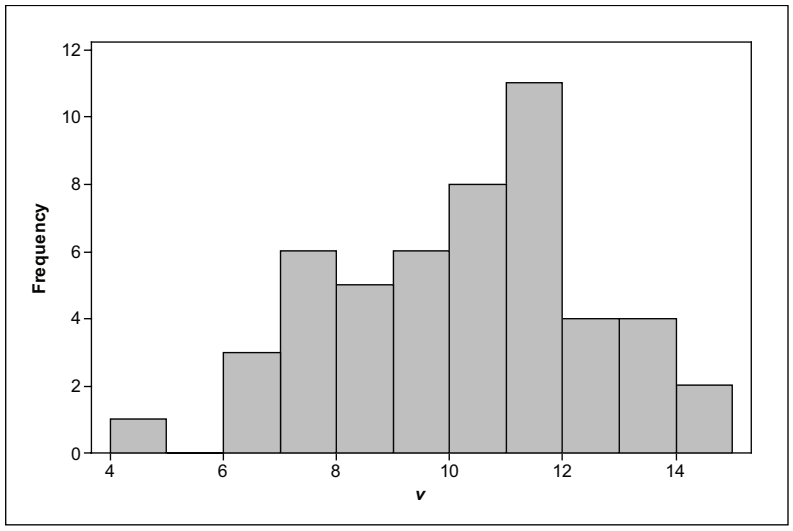


Figure 9.30. Histogram of data values for variable v.

b. Match the histogram of data on variable *w* (Figure 9.31) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.

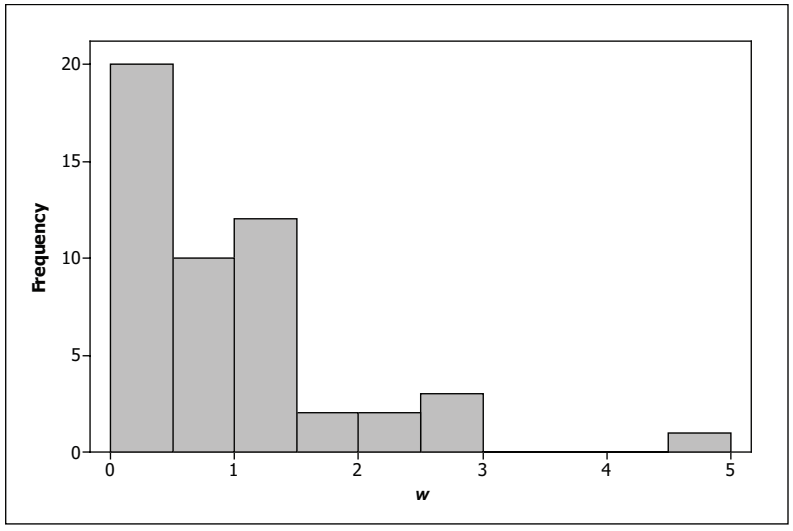


Figure 9.31. Histogram of data values for w.

c. Match the histogram of data on variable *x* (Figure 9.32) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.

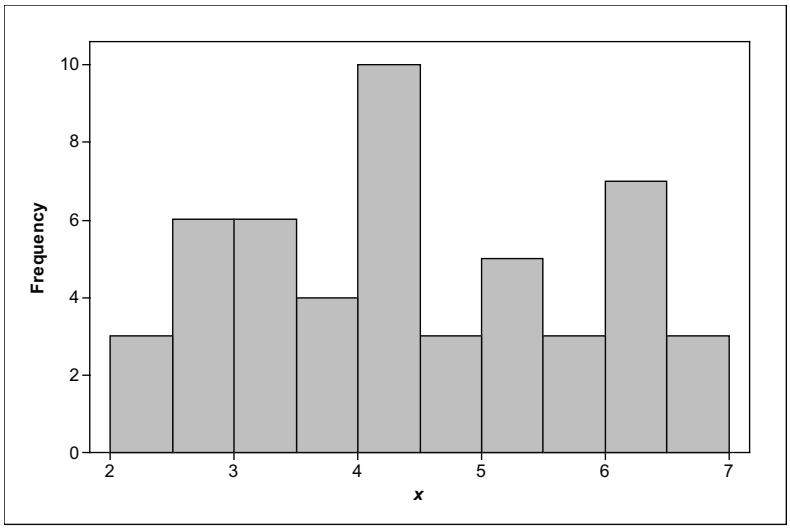


Figure 9.32. Histogram of data on variable x.

REVIEW QUESTIONS

1. The 5-quartiles are called quintiles.

a. What proportion of standard normal data falls between two consecutive quintiles?

b. What are the quintiles for a standard normal distribution?

2. IQ scores follow a normal distribution with mean and standard deviation. For each of the following questions, find the percentile. Then draw a graph of the normal distribution of IQ scores and shade the area associated with the percentile.

a. What is the 25^{th} percentile of IQ scores?

b. What is the 50^{th} percentile of IQ scores?

c. What is the 75^{th} percentile of IQ scores?

d. How are the 25^{th} and 75^{th} percentiles related?

The aging population is of interest to states because of costs connected with caring for older citizens. Table 9.2 gives the population in thousands of residents 85 or older in each state and the District of Columbia. But it isn't just the population size of older residents that is problematic. More important is the percentage of a state's residents who are older. Questions 3 and 4 are based on the data in Table 9.2, which contains information on the population size and percentage of state residents who are 85 years old or older.

State	Population 85 and Over (thousands)	Percent	State	Population 85 and Over (thousands)	Percent
Alabama	76	1.58%	Montana	20	2.02%
Alaska	5	0.66%	Nebraska	39	2.15%
Arizona	103	1.62%	Nevada	30	1.12%
Arkansas	51	1.76%	New Hampshire	25	1.88%
California	601	1.61%	New Jersey	180	2.04%
Colorado	70	1.38%	New Mexico	32	1.55%
Connecticut	85	2.38%	New York	391	2.02%
Delaware	16	1.75%	North Carolina	147	1.55%
District of Columbia	10	1.71%	North Dakota	17	2.48%
Florida	434	2.31%	Ohio	230	2.00%
Georgia	114	1.17%	Oklahoma	62	1.65%
Hawaii	30	2.22%	Oregon	78	2.03%
Idaho	25	1.61%	Pennsylvania	306	2.41%
Illinois	235	1.83%	Rhode Island	27	2.54%
Indiana	115	1.78%	South Carolina	71	1.53%
Iowa	75	2.45%	South Dakota	19	2.36%

Continued...

Kansas	59	2.08%	Tennessee	100	1.57%
Kentucky	69	1.59%	Texas	305	1.21%
Louisiana	66	1.45%	Utah	31	1.12%
Maine	29	2.19%	Vermont	13	2.05%
Maryland	98	1.70%	Virginia	122	1.53%
Massachusetts	145	2.22%	Washington	117	1.74%
Michigan	192	1.94%	West Virginia	36	1.94%
Minnesota	107	2.01%	Wisconsin	119	2.08%
Mississippi	44	1.49%	Wyoming	9	1.53%
Missouri	114	1.90%			

Table 9.2. Population and percentage of state residents 85 and over.

3. a. Make a histogram of the state population sizes of residents 85 or older. (The data are in Table 9.2). Describe the shape of the histogram. Is it reasonable to assume that the state populations of residents 85 or older are normally distributed?

b. Based on your histogram in (a), do you think a normal quantile plot for these data would be concave up, concave down, or mostly linear? Explain.

c. Make a normal quantile plot for these data. Based on your normal quantile plot, is it reasonable to assume that the state populations of residents 85 or older are approximately normally distributed? Explain.

4. Return to the data in Table 9.2. The focus here is on the percentage of state residents who are 85 or older.

a. Make a histogram of the percent data. Describe the shape of the histogram. Is it reasonable to assume that the state percentages of residents 85 years old or over are approximately normally distributed?

b. Based on your histogram in (a), do you think a normal quantile plot for these data would be concave up, concave down, or mostly linear? Explain.

c. Make a normal quantile plot for the percentage data. Based on your normal quantile plot, is it reasonable to assume that the state percentages of residents 85 years old or older are normally distributed? Explain.