**Chapter 2 Notes**

We now have a clear strategy for exploring data from a single quantitative variable.

1. Plot the data: make a graph
2. Look for overall pattern (shape, center, spread and unusual values)
3. Calculate a numerical summary to describe center and spread.
4. The new step: Use a smooth curve to help describe overall patterns.

The curve is a MATHMATICAL MODEL for the distribution. A model gives a compact picture of the overall pattern of the data but ignores minor irregularities as well as any outliers.

A DENSITY CURVE is a curve that

* Is always on or above the horizontal axis, and
* Has area exactly 1 underneath it.

Density curves come in different shapes. We will focus mainly on then normal curve for now.

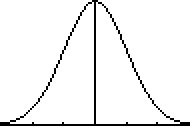
When distributions are unimodal and symmetric, they often can be modeled by a normal curve.

*ex: draw a dotplot and then overlay a normal curve*

Technically, the normal curve is a mathematical function that extends from negative infinity to positive infinity. Thus, a normal curve will never model a data set perfectly and data sets should be described as “approximately normal”, never “normal.”

When we use a mathematical function to describe a distribution of real data, we call the function a MODEL. The normal model is centered at the mean  with a standard deviation of : N(,)

When graphing a normal distribution, the inflection points are located at 1 standard deviation above and below the mean. An inflection point is where a graph goes from concave up (happy face) to concave down (frowny face).



We use Greek letters to denote that these values are not calculated from the data in a sample. They are either calculated from the entire population or they are assumed values to help solve problems. Numbers like these are called PARAMETERS (for example,  and ). In contrast, summaries of data from samples are called STATISTICS and are represented with non-Greek letters (such as and s).

The 68-95-99.7 Rule

When a distribution is approximately normal, then approximately 68% of the observations will be within 1 standard deviation of the mean, 95% of the observations will be within 2 standard deviations of the mean, and 99.7% of the observations will be within 3 standard deviations of the mean.

Note: How to draw a normal curve: Start with the axis and label the mean and 3SD in each direction. Place a dot above the mean for the top of the curve. Place dots at 60% of this height above 1 SD. Connect these with a concave down curve. Place dots at 1/8 of this height for 2 SD and very close to the axis for 3 SD. Connect these with a concave up curve.

Suppose that the birth weights of newborn babies are approximately normal with a mean of 7.6 lb and a standard deviation of 1.3 pounds.

1. sketch the normal curve with the correct scale
2. What percentage of babies will weigh less than 8.9 lbs?
3. What percentage will weigh more than 8.9 lbs?
4. What percentage will weigh more than 5 lbs?
5. Between which two weights will the middle 95% of babies be?
6. What proportion of babies weigh less than 8 lb?

The z-score for a weight of 8 lb is: * = .31.*

This z-score is between 0 and 1 so we know at least 50% weigh less than 8 lb. We also know that at most 84% weigh less than 8 lb. Thus, between 50%-84% of babies weigh less than 8 lb.

\*\*recall the information already taught on Z-score. Converting scores like this from original values to standard deviation units is know as ***standardizing.***

When we standardize data that is approximately normal in shape, we calculate:



If the distribution of x was approximately normal, then the distribution of z is also approximately normal with  = 0 and  = 1. That is z ~ N(0,1). This distribution is called the STANDARD NORMAL DISTRIBUTION.

Note: z-scores only follow a standard normal model when the original data was approximately normal!

To be more precise, we can use the standard normal table, which gives normal percentiles for z-scores between -3.50 and 3.50.

*introduce table*

The circumferences of oranges from a certain tree have an approximately normal distribution with a mean of 20 cm with a standard deviation of 4 cm. *Note: Always define the variable, state its distribution, and draw a picture!*

1. What proportion of oranges have circumferences less than 18 cm?
2. What proportion of oranges have circumferences less than 18 cm?
3. What proportion have circumferences greater than 25 cm?
4. What proportion have circumferences between 17 cm and 19 cm?
5. What is the 90th percentile for this distribution?
6. What is the first quartile of this distribution?
7. What is the interquartile range for this distribution?

Suppose that the number of M&M’s in a 1-pound bag has an approximately normal distribution with a mean of 550 and a standard deviation of 12.

1. Draw and label the normal model

Using the TI-83: Dist: Normalpdf(x, , ) gives the probability density function f(x) for a normal distribution with mean  and standard deviation . In other words, it gives the height of the graph for a particular value of x.

Note: 

To graph the M&M distribution, put Normalpdf(x,550,12) in Y1 and set an appropriate window.

1. What proportion of bags will have fewer than 525 M&M’s?

*Draw picture and use P(y < 525) = Normalcdf for all problems!*

Using the TI-83: Dist: Normalcdf(lower boundary, upper boundary, , ) gives you the area under the pdf bounded by the lower boundary and upper boundary values. If you do not enter values for  and , it will assume you are using the standard normal curve.

To find P(x < 525), enter Normalcdf(,525,500,12)

*Note: to get full credit on an exam, you must define the variable, state the distribution, sketch the picture, state the question, how you got the answer, and the answer.*

1. What proportion will have fewer than 600 M&M’s?
2. What proportion will have more than 565 M&M’s?
3. What proportion will have between 530 and 570 M&M’s?
4. What is the 30th percentile for this distribution?

Using the TI-83: Dist: Invnorm(area to left, , ) gives the value of the random variable x that has the given area to its left.

To find the 30th percentile on the TI-83, enter Invnorm(.3,550,12)

1. What is the interquartile range for this distribution?
2. If the company wants to guarantee that no more than 5% of bags have fewer than 540 M&M’s, what mean should they try to achieve (assuming the SD is 12)?
3. If the mean must stay at 550, but they still want to have no more than 5% of bags have fewer than 540 M&M’s, what standard deviation should they try to achieve?

**Assessing Normality**

In many problems we assume that a distribution is normal. In real life, however, a set of data will never be exactly normal. Still, the normal model can still be a useful approximation if the data are close to normally distributed.

To check if a data set is normally distributed, we could look at a dotplot, histogram or stemplot since these show both symmetry/skewness and modes. A boxplot can show symmetry/skewness but not where the modes are.

However, even dotplots, histograms, and stemplots are not as useful as a Normal Probability Plot (also called a Normal Quantile Plot).

In a Normal Probability Plot, the actual data values are plotted on one axis and each value’s theoretical z-score is plotted on the other axis. For example, in a data set with 100 values from a normal distribution, the smallest value should have a z-score of -2.58.

If a data set closely follows a normal distribution, then the NPP should look linear. However, if the NPP is not linear, this gives evidence that a Normal Model may not be appropriate.

*Examples with Fathom:*





Using the TI-83: In the Stat Plot menu (2nd Y=), the 6th graph choice is a Normal Probability Plot.