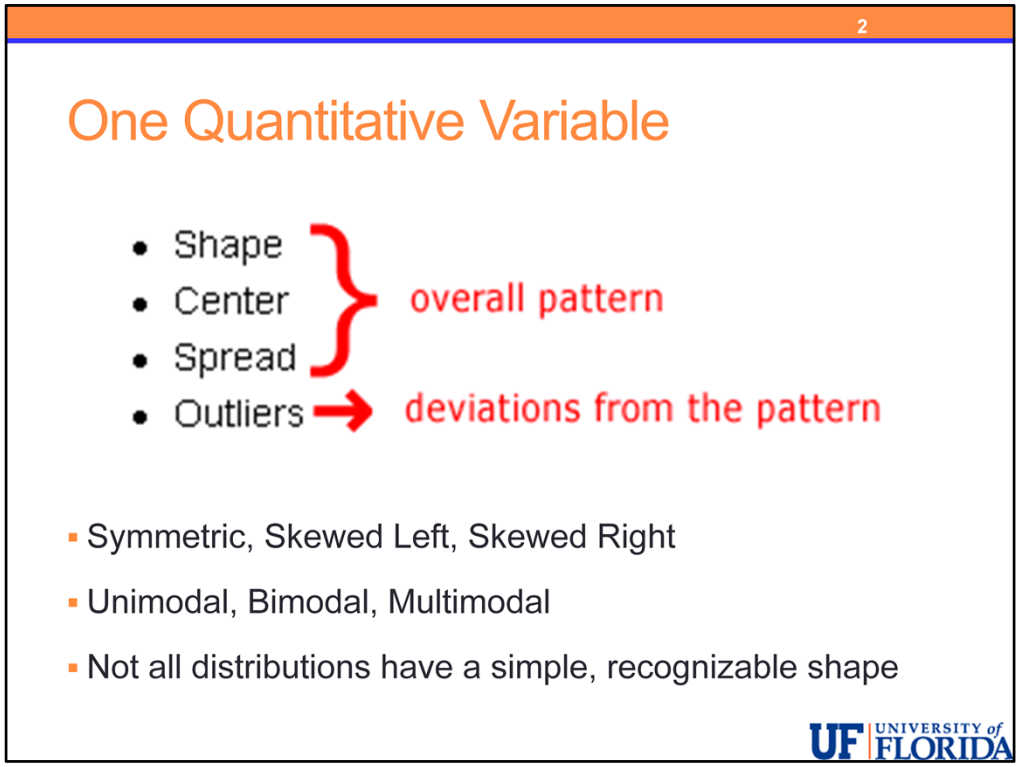
DESCRIBING DISTRIBUTIONS

One Quantitative Variable

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

Before continuing into greater detail, we need to discuss the basic features of distributions which we want to describe.



For one quantitative variable, we are interested in four main components

Shape – this includes whether this distribution is symmetric or skewed and if there is a pattern to its peakedness (modality)

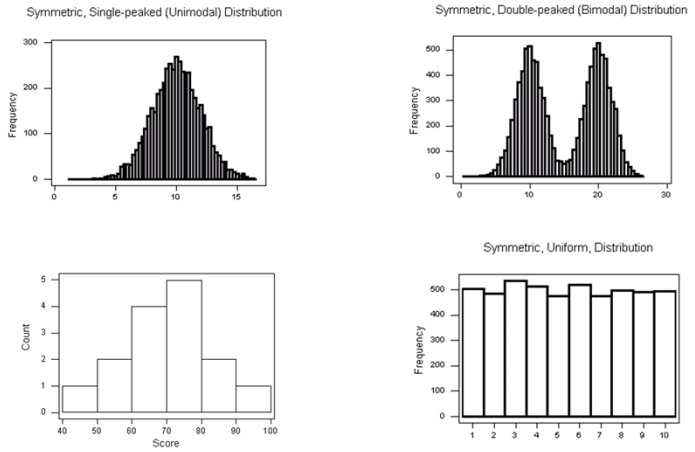
Center – gives a location of the approximate center of the distribution on the scale for the variable

Spread – measures how much variation exists in the distributions, how spread out the values are on the scale for the variable

And we also look for outliers – values which seem to fall outside the norm for the variable

It is important to note that not all distributions have a simple recognizable shape.

Symmetric Distributions



UF UNIVERSITY of FLORIDA

Here we have four symmetric (or reasonably symmetric) distributions. Remember – we don't expect perfection! Look at these types of questions with slightly blurry vision – what is the general trend – and don't allow yourself to be stuck on any minor details.

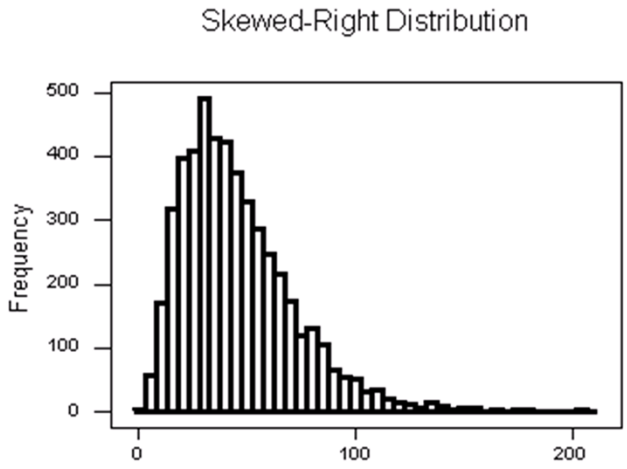
The first graph is symmetric and unimodal - there is only one cluster or high point – called a mode.

The second graph is symmetric and bimodal – there are two equal clusters or high points but the symmetry is maintained

The third graph is again symmetric and unimodal – these were our 15 exam scores from our discussion of histograms.

The final graph is symmetric and uniform – there is no mode – all values are approximately equally likely – which is what we mean by a uniform distribution

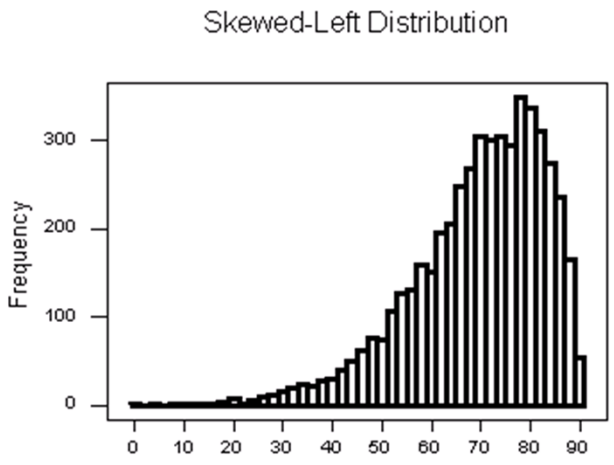
Skewed Distributions



UF UNIVERSITY of FLORIDA

Here we have a unimodal distribution which is skewed right – the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

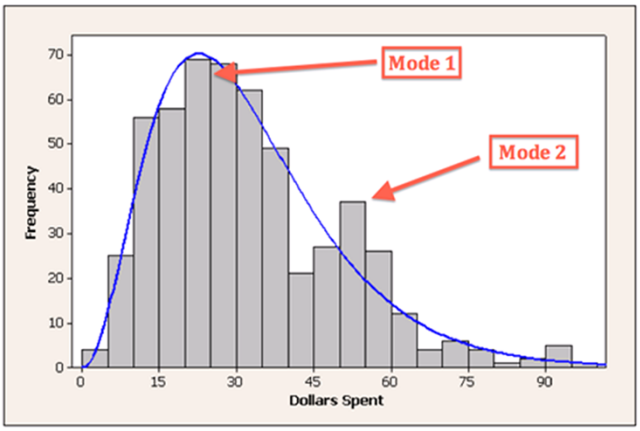
Skewed Distributions



UF UNIVERSITY of FLORIDA

Here we have a unimodal distribution which is skewed left – the left tail of the distribution is longer than the right. High values are more common in a skewed left distribution.

A Bimodal Skewed Right Distribution

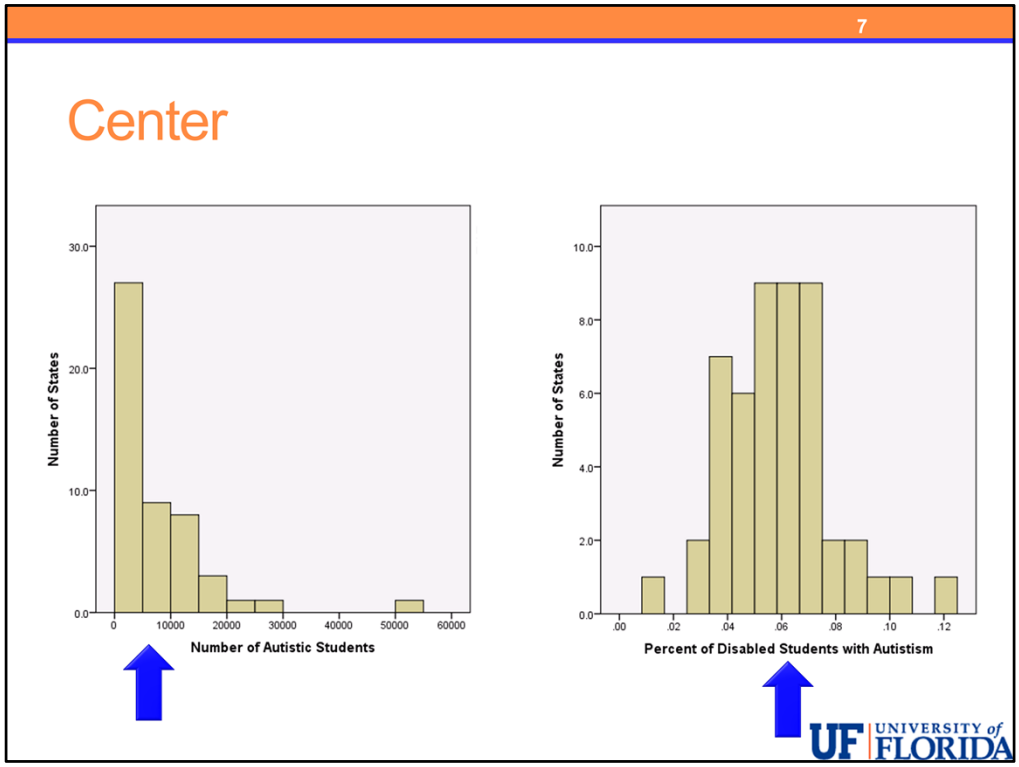


UF UNIVERSITY of FLORIDA

Bimodal histograms can be skewed right as seen in this example where the second mode is less pronounced than the first.

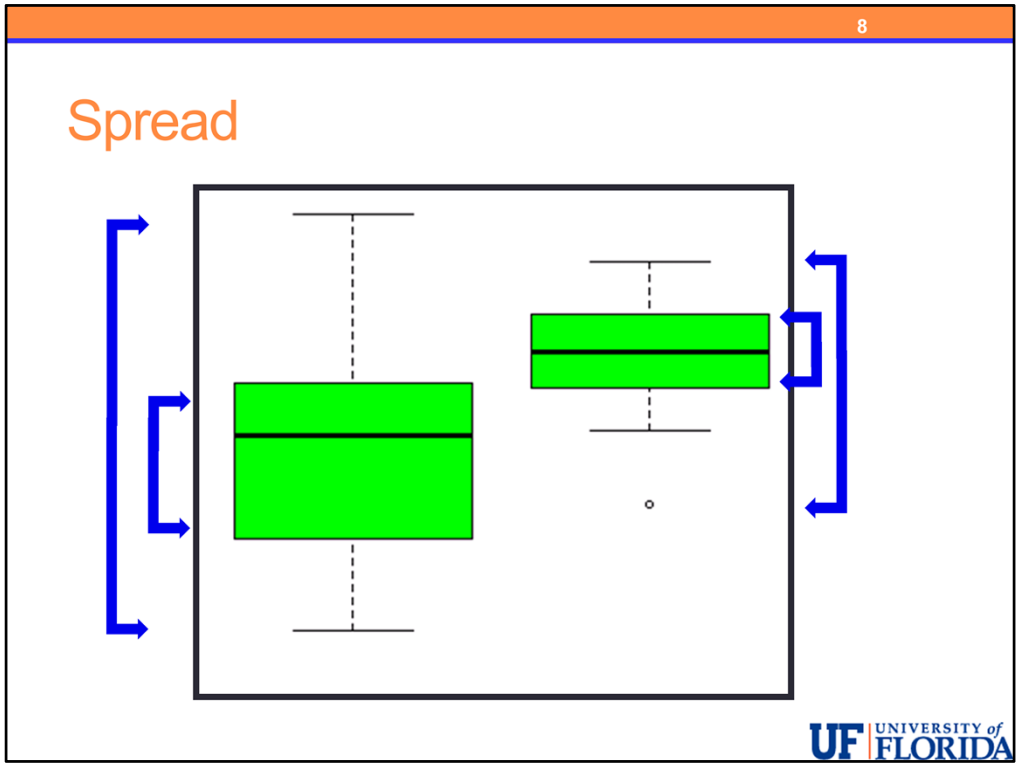
This second mode may or may not have any practical meaning but could be worth investigating to be certain. Sometimes a mode such as this is simply a strange "random" chance and does not reflect any bimodality in the population from which the sample was taken.

Distributions having more than two modes are called multi-modal.



The concepts of symmetry, skewness, and modality represent the overall shape of the distribution.

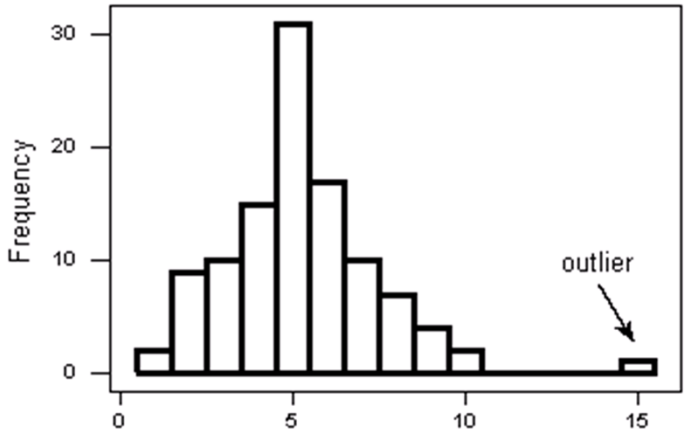
The next feature of the distribution to consider is the "center" of the distribution. We will discuss specific measures of center, but for now we again point out that we desire a value that locates the "bulk" of the data. This will be easier to do for a symmetric distribution than for a skewed one.



We will define three ways to measure the spread of the distribution two of which, the range and Inter-quartile range, we will see are easily visible in the boxplot. The third measure of spread we will cover is the standard deviation.

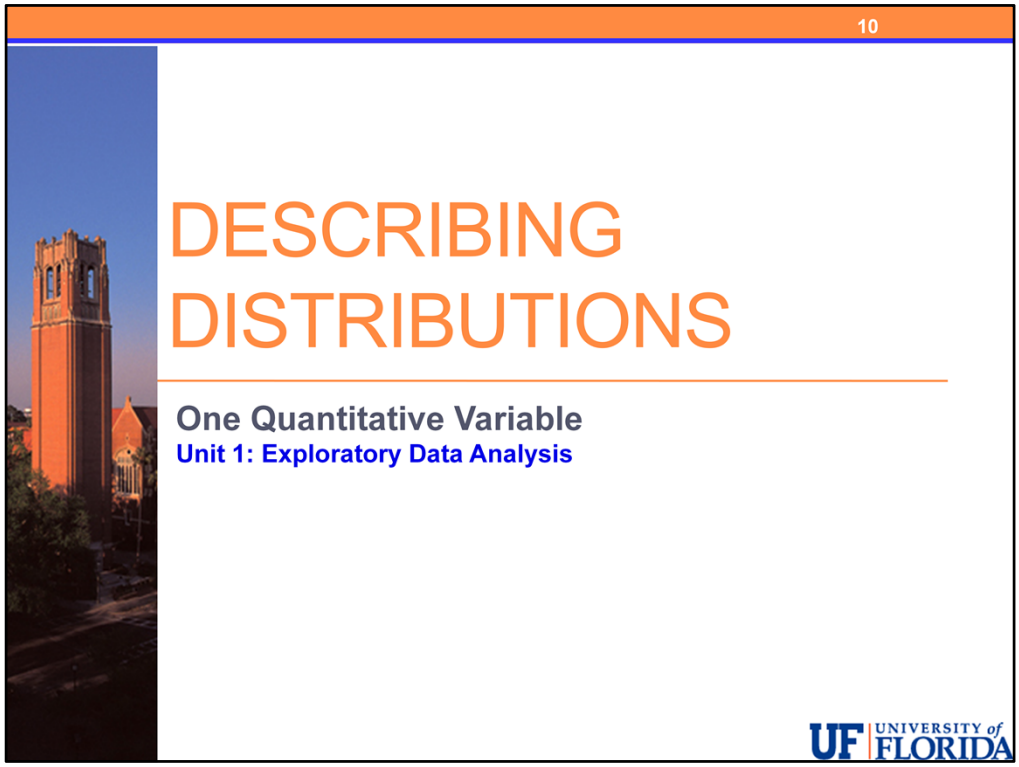
When presented on the same scale, we can easily compare the variation of two distributions visually. We will need to formalize methods for measuring this variability or spread.

Outliers



UF UNIVERSITY of FLORIDA

Unusually large or small values can occur for many reasons. Identifying these values, called outliers, is the first step.



The remainder of the material regarding one quantitative variable will add to our knowledge and ability to describe the distribution of these types of variables.

Remember as we learn the specifics that our primary goal is to describe and summarize the distribution of this variable. In the case of one quantitative variable we do this by investigating the shape, center, spread, and outliers.