**Describing the Shapes of Frequency Distributions**

 This document is intended to help beginning students in statistics learn a little bit about the shapes of frequency distributions and how to describe them. For each of several shapes, I provide a stem and leaf plot for a sample of 100 scores randomly drawn from a population of known shape. I also provide a few summary statistics and some commentary.

**Skewness and Kurtosis**

 The **skewness** statistic tells you whether the sample has a negative or a positive skewness. Keep in mind that even when drawn from a perfectly symmetric distribution a sample is likely to have a little bit of skew. If you get an absolute value of skewness below .5, best not to infer that the population is skewed (unless you have a very large sample).

 The kurtosis statistic can be very helpful, but as with skewness, expect there to be some sampling error. In a normal distribution, both skewness and kurtosis are zero. In a uniform distribution the skewness is zero and the kurtosis is -1.2. A binomial distribution with *n* =1, *p* = .5 has the lowest possible kurtosis, -2. Such a distribution will have only two values with nonzero frequencies. Kurtosis will be quite low in such distributions, often approaching -2. Kurtosis is of little value when determining whether or not a distribution is bimodal. Best to rely on plots for that.

**Normal Distribution.**

 The population was a normal distribution with a mean of 35 and a standard deviation of 1.5. I created this distribution with the following SAS code:

DO K=**1** TO N; Y=round(**35**+**1.5**\*NORMAL(**0**)); OUTPUT; END;

 Here is what a sample of 100 scores looked like:

 Stem Leaf # Boxplot

 38 000 3 |

 37 |

 37 00000000000000 14 |

 36 |

 36 000000000000000000000000 24 +-----+

 35 | |

 35 0000000000000000000000 22 \*--+--\*

 34 | |

 34 000000000000000000000000 24 +-----+

 33 |

 33 000000000 9 |

 32 |

 32 0000 4 |

 ----+----+----+----+----

You might be tempted to describe this distribution as “bimodal,” but the frequencies of the scores of 34, 35, and 36 differ by only trivial amounts, so it would be better to describe it as unimodal. **Keep in mind that you are attempting to describe the shape of the population distribution from which the sample data were randomly obtained. The sample above came from a population that was perfectly normal.** The skewness was -.10 and the kurtosis -.57, both values deviating from 0 due to sampling error. I should add that the sample does lean towards the uniform (kurtosis -1.2), but is closer to normal (kurtosis 0) than to normal.

 Deciding whether the sample is more like a uniform distribution or a normal distribution can be difficult. In both cases the skewness should not be very far away from 0. Although I am not a big fan of statistical tests of normality, I shall mention here how to test for skewness and for kurtosis. For skewness, compute  If the resulting absolute value of z is 2 or more, skewness differs significantly from 0. For our sample,  For kurtosis, compute  For our sample,  I should point out that the test statistics here do not become approximately normal until sample size is up to 150 for *g1* and 1000 for *g2*, so take them with a grain of salt.

 Here is an SPSS-produced histogram from a different sample drawn from a normal population. The skewness of this sample was -.10 and the kurtosis -.27.



**Uniform Distribution**

 The population was a uniform distribution ranging from 30 to 40. I created this distribution with the following SAS code:

DO K=**1** TO **100**; Y=round(**30**+**10**\*UNIFORM(**0**)); OUTPUT; end;

 Here is what the sample of 100 scores looked like:

 Stem Leaf # Boxplot

 40 00000000 8 |

 39 000000000 9 |

 38 0000000000 10 +-----+

 37 00000 5 | |

 36 000000000 9 | |

 35 000000000 9 | |

 34 000000000 \*--+--\*

 33 000000000 9 | |

 32 0000000000 10 +-----+

 31 0000000000000 13 |

 30 000000000 9 |

 ----+----+----+----+

 Not all of the scores are equally frequent, due to sampling error, but the plot clearly indicates a uniform distribution. One could describe the score of 31 as being the mode, since it is a bit more frequent than the other scores, but it would be, IMHO, be better to say that the sample really has no mode. The skewness is 0.15 (essentially symmetric), and the mean (34.7) and the median (34.5) are not separated by much. The kurtosis of a uniform population is -1.2. The kurtosis in in this sample is -1.2665781 (about what one should expect for a sample from a uniform distribution).

 For skewness in this sample,  For kurtosis, -- the kurtosis is significantly less than zero, as is expected in a uniform distribution.

 Here is an SPSS-produced histogram of a different sample drawn from a uniform distribution. The skewness in this sample was .042 and the kurtosis -1.28.



**A Two-Point Binomial Distribution, *p* = .5**

I threw this in the mix just to show the distribution that has the minimum possible value of kurtosis, -2. The SAS code used to generate the data is:

DO K=**1** TO N;IF UNIFORM(**0**) <**.5** THEN Y=**32**;

ELSE Y=**38**; OUTPUT; END;

Here is a sample of 100 generated from this code:

 Histogram # Boxplot

 38.25+\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* 52 +-----+

 . | |

 . | |

 . | |

 . | |

 . | |

 35.25+ | + |

 . | |

 . | |

 . | |

 . | |

 . | |

 32.25+\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* 48 +-----+

 ----+----+----+----+----+-

 \* may represent up to 2 counts

The estimated kurtosis was -2.03. Yikes, how is that possible when the minimum possible value of kurtosis is -2? Well, SAS reports an estimate of kurtosis that can fall below -2, and, when the kurtosis in the population is -2, samples from that population will frequently have values of estimated kurtosis that fall a bit below -2. There were only two values in the data, 32 and 38. This distribution is properly described as U-shaped, bimodal, and two-point binomial.

 Here is an SPSS-produced histogram of a different sample drawn from a two-point binomial population. The skewness in this sample was -.20 and the kurtosis -1.999.



**U-Shaped Distribution**

 I created an approximately U-shaped distribution with the following code:

DO K=**1** TO **100**; IF Select <**.25** THEN Y=**25**;

Else if **.25** < Select < **.5** then Y=round(**25**+ABS(**2**\*NORMAL(**0**)));

Else if Select < **.75** then Y=**45**;

ELSE Y=round(**45**-ABS(**2**\*NORMAL(**0**))); OUTPUT; END;

 Here is what the sample looked like:



 This distribution is, I opine, properly described as approximately U-shaped and as bimodal -- the peak with a frequency of 31 is of about the same height as the peak with a frequency of 27. Although the mean (35.7) is separated from the median (42), the distribution is nearly symmetric, as indicated by the skewness of -0.15. The kurtosis of -1.95 is about what one would expect for a sample drawn from a U-shaped distribution – that is, just a bit less than would be expected to a two-point binomial distribution.

 Here is an SPSS-produced histogram from a different sample drawn from a U-shaped distribution population. The skewness in this sample was 0.24 and the kurtosis -1.92.



**A Bimodal Distribution, Not U-Shaped**

 Here I combined, in approximately equal proportions, scores from normal distributions with means of 30 and 40, using the code:

N=**100**;D=UNIFORM(**0**);

BIMOD: Shape = 'Bimodal'; DO K=**1** TO N;

IF UNIFORM(**0**) <**.5** THEN Y=round(**40**+**1.75**\*NORMAL(**0**));

ELSE Y=round(**30**+**1.75**\*NORMAL(**0**)); OUTPUT; END;

 I would describe this sample as having two modes (30 and 40), even though they do not have identical frequencies. The mean (34.9) and median (33) are quite close here, and the skewness (.06) nearly zero. The kurtosis (-1.55) reflects the displacement of scores from the shoulders toward the tails of the distribution. Values of kurtosis less than that expected for a uniform distribution (-1.2) may suggest that the distribution is bimodal (Darlington, 1970), but bimodal distributions can have high kurtosis if the modes are distant from the shoulders. See the stem and leaf plot and the histogram below.



Here is an SPSS-produced histogram from a sample drawn from a bimodal distribution. This sample had skewness -.10 and kurtosis -1.72.



**A Skewed Distribution**

 I created a skewed distribution by taking about 80% of the scores from a normal distribution with a mean of 35 and about 20% from a normal distribution with a mean of 43 (for half of the samples) or 27 (the other half), using this code:

IF UNIFORM(**0**)<**.5** THEN S=-**1**; ELSE S=**1**;

DO K=**1** TO **100**; Y=round(**35**+**1.0**\*NORMAL(**0**));

IF UNIFORM(**0**) <**.2** THEN Y=round(Y+S\***8**);OUTPUT; end;

 Here is an example of one of the negatively skewed distributions created with this code:

 The mean (33.7) is less than the median (35.0), and the skewness is -1.35. I would not call this distribution bimodal, but some might. It was created by combining two normal distributions with different means. Would it be meaningful to say that there is one mode at 36 and another at 27? If you say yes, are you uncomfortable with the fact that the score values of 35 and 34 are considerably more frequent than is the 27?

 For skewness in this sample, .

 See the plots below.

Stem Leaf # Boxplot

 38 0 1 |

 37 00000 5 |

 36 000000000000000000000000000000 30 +-----+

 35 0000000000000000000000000 25 \*-----\*

 34 00000000000000000 17 +-----+

 33 00 2 |

 32 0 1 |

 31

 30

 29

 28 000000 6 0

 27 00000000 8 \*

 26 00000 5 \*

 ----+----+----+----+----+----+

IMHO, a distribution that has two peaks but with one much lower than the other is not properly described as “bimodal.” Consider this distribution:

 This distribution clearly has two peaks, but I would not call the score of 43 a second mode, given that its frequency (9) is much less than that of several other scores – 35 (frequency = 34), 34 (frequency = 17) and 36 (frequency = 17).

 Here is an SPSS-produced histogram from a different sample drawn from a skewed distribution. The skewness in this sample was 1.66 and the kurtosis 1.51.



**A Warning About SPSS Stem and Leaf Plots**

 Look at the following SPSS stem and leaf plot from a negatively skewed distribution (g1 = −1.58).

Circum Stem-and-Leaf Plot

 Frequency Stem & Leaf

 15.00 Extremes (=<28)

 1.00 32 . 0

 .00 32 .

 6.00 33 . 000000

 .00 33 .

 19.00 34 . 0000000000000000000

 .00 34 .

 33.00 35 . 000000000000000000000000000000000

 .00 35 .

 21.00 36 . 000000000000000000000

 .00 36 .

 5.00 37 . 00000

 Stem width: 1

 Each leaf: 1 case(s)

 Notice that there are 15 extreme scores on the low end of the distribution but none on the high end. This plot makes the distribution look less skewed than it is, especially if you don’t notice SPSS’ warning about “15 Extremes”), which I have highlighted here.

 Now look at the SAS stem and leaf plot for the same data, which much better displays the shape of the distribution.



**What Does “Bimodal” Mean?**

 The term “bimodal” is used early in almost all introductory statistics texts, but it is not really well defined. A “mode” is defined as the most frequent score, so one could conclude that a bimodal distribution is one which has two scores tied for most frequent -- but if those scores were very close to one another, I would not describe the distribution as bimodal. On the other hand, if there were two peaks of about equal height (but not exactly equal) and not very close to one another, I would describe the distribution as bimodal. More troublesome is the case where there are two peaks, but the one peak is very much smaller than the other. While it might be appropriate to describe such a distribution as bimodal, I find it discomforting that the score at the lower peak is likely much less frequent than many other scores that are not considered to be modes. I have a really great class of graduate students right now (Autumn, 2001), and they asked me about this (I had never really given it much thought before). I posted a query to the EDSTAT list (a wonderful resource for those interested in teaching statistics). I include here my query and responses thereto.

From: "Wuensch, Karl L." <WUENSCHK@MAIL.ECU.EDU>

To: "edstat (E-mail)" <edstat@jse.stat.ncsu.edu>

Subject: Bimodal distributions

Date: Thursday, August 30, 2001 12:54 PM

 Does a bimodal distribution necessarily have two modes? This might seem like a silly question, but in my experience many folks apply the term "bimodal" whenever the PDF has two peaks that are not very close to one another, even if the one peak is much lower than the other. For example, David Howell (*Statistical Methods for Psychology*, 5th, p. 29) presents Bradley's (1963) reaction time data as an example of a bimodal distribution. The frequency distribution shows a peak at about 10 hundredths of a second (freq about 520), no observations between about 18 and 33 hundredths, and then a second (much lower) peak at about 50 hundredths (freq about 25).

From: "David C. Howell" <David.Howell@uvm.edu>

Date: Thursday, August 30, 2001 2:00 PM

Karl Wuensch asks an interesting question, though I would phrase it somewhat more generally. "At what point does a bimodal distribution become just a distribution with two peaks?" Except for a few quite extreme situations, dealing with mixtures of distributions and the like, it will rarely ever be the case that the two peaks of a distribution are EXACTLY the same height. But if they are extremely similar, no one would ever quibble. The case that I use from Bradley (1963) has two peaks that are quite clearly different in height--in fact, one might argue that the second peak is so diffuse as not to deserve to be called a peak. And yet it seems to me that calling the distribution bimodal is saying something useful about the distribution. Perhaps someone can suggest a better term.

Dave Howell

From: "Paul R. Swank" <Paul.R.Swank@uth.tmc.edu>

Date: Thursday, August 30, 2001 2:19 PM

A bimodal distribution is often thought to be a mixture of two other distributions with different modes. If the distributions have different sizes, then it is possible to have two or more "humps". I once read somewhere (and now can't remember where) that this may be referred to as bimodal (or multimodal). In the bimodal case, some refer to the higher "hump" as the major mode and the other as the minor mode.

Paul R. Swank, Ph.D.

From: "Dennis Roberts" <dmr@psu.edu>

Date: Thursday, August 30, 2001 3:54 PM

this is an interesting point but, one we have to be careful about ... in the minitab pulse data set ... c6 is heights of 92 college students ... a mixture of males and females ...

 : :

 : . : : : .

 : : : : : : : : .

 : : : : : : : : : : : : .

 . .: .: : : . : : : : . : : . : : : : :

 -----+---------+---------+---------+---------+---------+-Height

 62.5 65.0 67.5 70.0 72.5 75.0

now, if we were to 'roughly' see the 'peaks' ... around 68/69 ... and 72/73 one might say that THIS is because of the gender differences (ie, where the modes or averages BY sex were)... but look at the separate dotplots

Dotplot: Height by Sex

 . : .

 Sex : . : : : .

 1 : : : : : : : : : .

 : : : : . : : . : : : : :

 -----+---------+---------+---------+---------+---------+-Height

 :

 Sex : : : : . : .

 2 . .: .: : : . : : : : .

 -----+---------+---------+---------+---------+---------+-Height

 62.5 65.0 67.5 70.0 72.5 75.0

but look at the desc. stats ...

Descriptive Statistics: Height by Sex

Variable Sex N Mean Median TrMean StDev

Height 1 57 70.754 71.000 70.784 2.583

 2 35 65.400 65.500 65.395 2.563

using the modes ... as approximations for the averages ... means or medians ... might not be a good idea ...

in this case ... we get a 'peak' around 68/69 not because of ONE gender concentrating there ... but, OVERLAPPING between the sexes ... at this approximate location of heights modes are tricky.

[Return to Wuensch’s Stats Lessons Page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm)

[Karl L. Wuensch](https://core.ecu.edu/wuenschk/klw.htm), February, 2023