

15

Distributions and Graphs

Creating an Ungrouped Frequency Distribution
Creating a Grouped Frequency Distribution
Visualizing the Distribution: the Histogram
Visualizing the Distribution: the Frequency Polygon
Common Distribution Shapes
Distribution-Free Data

The end of the research part of a study comes after the data has been collected through tests, attitude scales, questionnaires, or other instruments. **Raw data presents us an incomprehensible mass of numbers.** The first step in statistical analysis is to reduce this incomprehensible mass of numbers into meaningful forms. This is done by using frequency distributions and associated graphs. In this chapter we'll look at several ways to organize data so that you see its meaning. We will look at both Ungrouped and Grouped Frequency Distributions.



Creating An Ungrouped Frequency Distribution

Let's say that you have given a Bible knowledge test to 38 high school seniors. The maximum score is 120. Here are the scores:

90	59	75	81	66	95	75	71	100	78	51
84	70	105	109	104	47	89	62	83	95	58
99	98	59	93	82	69	72	84	97	44	74
80	75	68	91	97						

As you can see, this collection of numbers makes little sense as it is. But we can organize and summarize the data in such a way to make it meaningful. Let's start by *rank ordering the numbers from high (109) to low (44).*

↓	109	97	90	81	74	66	47
↓	105	97	89	80	72	62	44
↓	104	95	84	78	71	59	
↓	100	95	84	75	70	59	
↓	99	93	83	75	69	58	
↓	98	91	82	75	68	51	

This ranking helps us to see where any given score fell along the whole range of scores. But the list is still rather long and difficult to manage. Let's now go through the list and *count the number of times each score occurs*. This is the score's **frequency**, represented by the letter "f."

Score	f	Score	f	Score	f
109	1	89	1	70	1
105	1	84	1	69	1
104	1	83	1	68	1
100	1	82	1	66	1
99	1	81	1	62	1
98	1	80	1	59	2
97	2	78	1	58	1
95	2	75	3	51	1
93	1	74	1	47	1
91	1	72	1	44	1
90	1	71	1		

The *ungrouped frequency distribution* above removes the redundancy of repeating scores. But the large number of single scores ($f=1$) still confuses the picture. If we were to *group ranges of scores together in classes*, we would get a *better picture* of the data. Grouping scores into classes produces a *grouped frequency distribution*.

Creating a Grouped Frequency Distribution

The steps in constructing a grouped frequency distribution are as follows: calculate the range of scores, compute the class width (i), determine the lowest class limit, determine the limits of each class, and finally group the scores into the classes.

Calculate the Range

The range of scores is found by subtracting the lowest score from the highest and adding one. Or, in statistical shorthand,

$$\text{Range} = X_{\max} - X_{\min} + 1$$

The "X" represents a score. The term " X_{\max} " refers to the highest (maximum) score and " X_{\min} " to the lowest (minimum) score. Putting the above formula into English, we read,

The range of a group of scores is equal to the difference between the maximum and minimum scores in the group, plus 1.

In our case, the **range** equals $(109 - 44 + 1) = 66$.

Compute the Class Width

We approximate the size of each category of scores, called the **class width** (i), by *dividing the range by the number of intervals we wish to have*. Conventional practice suggests we use 5 to 15 classes. We'll use **10 classes** here.

The tentative class width (i) is equal to the range of **66**, computed above, **divided** by the number of intervals desired, **10**.

$$66 / 10 = 6.6$$

We need to round up or down to a whole number. **Odd class widths are better than even ones** because the midpoint of an odd-width class is a whole number. So let's round up to "7." (In this context, we would even round a number like '6.1' up to 7). The distribution will have a **class interval (i)** of 7.

Determine the Lowest Class Limit

Each class of scores should begin with a *multiple of the class width*. The lowest class limit should be a **multiple of i** (in our case, $i=7$) AND **include the lowest score**. Our lowest score is 44. *The value "42" includes the score of 44 and is a multiple of 7*. So our first class begins with 42 and includes 7 scores. As a result, all scores with a value of 42, 43, 44, 45, 46, 47, or 48 will be counted in this class. The lowest class is 42-48.

Determine the Limits of Each Class

The next higher class will begin with $(42+7=)$ 49, the next with $(49+7=)$ 55, and so on, until we reach the last class, 105-111. All classes are listed below.

Group the Scores in Classes

Move through the data and count how many scores fall into each class. The result looks like this:

Class	Counts	f
105-111	//	2
98-104	////	4
91-97	////////	6
84-90	////	4
77-83	////	5
70-76	////////	7
63-69	///	3
56-62	////	4
49-55	/	1
42-48	//	2

$n = \Sigma f = 38$ scores

This grouped frequency distribution reveals much more about the Bible knowledge of high school seniors than we could discern in previous listings. On the down side, by grouping our scores into classes, we actually lost some detail. But "losing detail" is necessary when the aim is to derive meaning from the numbers. We can combine our scores even more by increasing the class width i . Let's look at a frequency distribution of the same data with $i = 14$.

Class	Tally	f
98-111	//// /	6
84-97	//// //	10
70-83	//// // //	12
56-69	//// //	7
42-55	///	3

$n = 38$

This last graph gives a smoother picture of the data set, though we notice the loss of more detail because we reduced the number of classes. Frequency distributions certainly simplify data sets, but we can present the data even more clearly by graphing the frequency distributions.

Graphing Grouped Frequency Distributions

Graphs display frequencies in a visual form. We can see a bit of this visual form in the “counts” columns above. The length of the counts (\\) gives a rough visual image of the data distribution. But we can do better with a graph. A frequency distribution graph consists of two axes which frame the frequency of each score interval.

X-and Y-axes

A graph is composed of a vertical line, called the *ordinate* or the *Y-axis*, and a horizontal line, called the *abscissa* or the *X-axis*. These two lines intersect to form a right angle. By convention, the Y-axis should be three-fourths the length of the X-axis. *Axis* is pronounced AX-is. *Axes* is pronounced AX-ees.

Scaled Axes

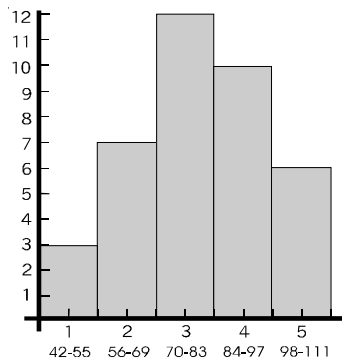
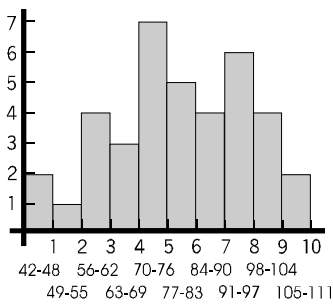
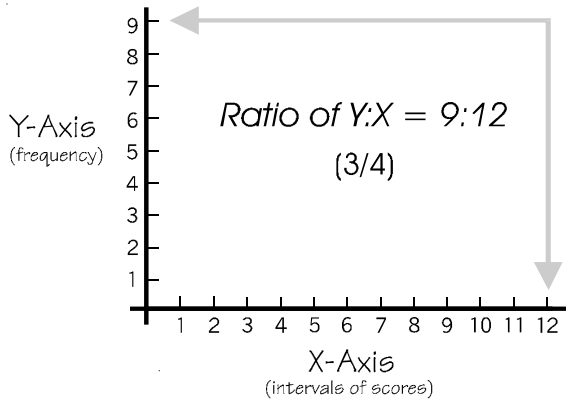
Numbers are placed on the X- and Y-axes at equal intervals to represent the scale values of the variable being graphed. In a graph of a grouped frequency distribution, the X-axis is scaled by the range and class intervals, the Y-axis is scaled by frequency. There are two major graph types used to display information from a grouped frequency distribution. The first is the histogram and the other is the frequency polygon.

Histogram

A *histogram* (HISS-ta-gram) is a special type of bar graph. The width of the bars equals the class interval and the heights of the bars equal class frequencies. Let's use the example data to build a histogram with a range of 44-111 and class width (i) of 7. The frequencies for this graph are located in the middle of page 15-3. Look at the graph at left.

Class limits are listed along the X-axis. The widths of all classes equal 7. The height of each bar equals the frequency of scores contained in each category. The shape of the graph provides us a clear and meaningful picture of the entire data set.

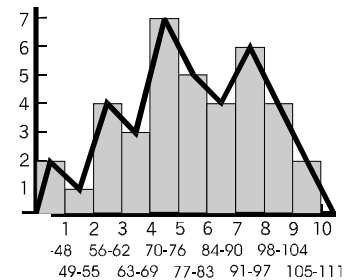
Then we reduced the number of categories from ten to five (increased *i* from 7 to 14). The graph at



left shows the effect of reducing the number of classes. Irregularities have been smoothed out, but some of the more specific (irregular) data has been glossed over. Choosing class width and the number of classes is a trial and error process. Our goal is to reflect the shape of the data as clearly as possible while attaining as much precision as possible.

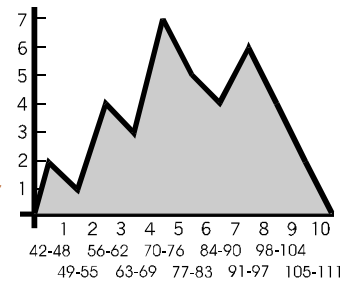
Frequency Polygon

By connecting the **midpoints of the bars with lines**, we produce a frequency polygon. The frequency polygon displays the same information as the histogram, but in a different form. The frequency polygon at right is based on the ten-class histogram or the previous page. If we remove the bars of the histogram, we obtain a frequency polygon graph, below right.



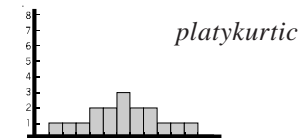
Distribution Shapes

The graphic image of a histogram or frequency polygon tells us at a glance the group profile of the data. **The incomprehensibility of a set of numbers is transformed into a meaningful visual portrait.** This visual portrait displays two special characteristics: *kurtosis* and *skewness*.

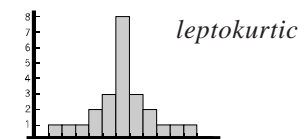


The *kurtosis* of a curve describes how flat or peaked it is. The three basic profiles of kurtosis are *platykurtic* (flat), *leptokurtic* (peaked), and *mesokurtic* (balanced).

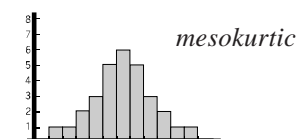
A flat curve is called *platykurtic*. Think of the flatness of a plate and you'll remember "platey-kurtic." Notice that there are low frequencies for all the categories.



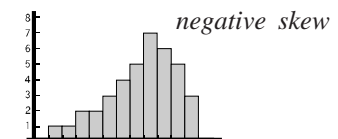
A peaked curve is called *leptokurtic*. Think of the central frequencies "leaping" away from the others and you'll remember "leap-tokurtic." Notice that outer categories have lower frequencies while the central categories have high frequencies.



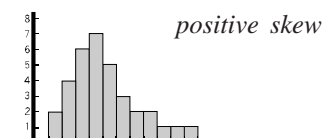
A curve that falls between platykurtic and leptokurtic is called *mesokurtic*. Think of medium (meso-) and you'll remember meso-kurtic. The familiar bell shaped curve is mesokurtic.



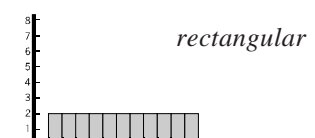
The *skewness* of a curve describes how horizontally distorted a curve is from the familiar bell-shaped curve. A curve with *negative skew* has its left tail pulled outward to the left, to the negative end of the scale.



A curve with *positive skew* has its right tail pulled outward to the right, to the positive end of the scale. A common mistake is to focus on the "mound of scores" rather than the distorted tail. **Remember: the direction the tail is pulled is the direction of the skew.**



A distribution where all categories of scores have equal



frequency is called a rectangular distribution.

Distribution-Free Measures

Our discussion on distributions applies to **ratio or interval data only**, called *parametric* data. Two other types of statistics deal with the *non-parametric* measures: either ordinal (ranks) or nominal (counts) data. Non-parametric data is often called "distribution-free." We will spend the next few chapters dealing with parametric statistics, and then deal with non-parametric types in Chapters 22, 23, and 24.

Summary

This chapter carried you through the first step in data analysis: reducing a series of chaotic numbers to orderly distributions and graphs. Before engaging in more sophisticated statistical procedures, you should initially analyze your data with these data reduction techniques. All good introductory statistics texts have chapters on data reduction techniques.

Vocabulary

Absissa	number along the horizontal (x-) axis of a graph
Class width (i)	distance between upper and lower limits in a given class
Class	a subset of scores defined by upper and lower limits in a frequency distribution
Exponential curve	line on a graph produced by the equation $y = x^2$
Frequency (f)	the number of scores in a given class
Frequency polygon	graph that depicts class frequencies: uses class midpoints
Histogram	graph that depicts class frequencies: uses class limits
Kurtosis	amount of flatness (or peakedness) in a distribution of scores
Leptokurtic	highly peaked distribution ("leaps up" in the middle)
Mesokurtic	moderately peaked distribution (normal curve)
Midpoint	halfway point between class limits in a given class: x'
Negative skew	negative end of skewed distribution: tail pulled left in a negative direction
Non-parametric measures	ranks or counts; ordinal or nominal; distribution-free
Ordinate	number along the vertical (y-) axis
Parametric measures	scales or tests; interval or ratio; normal distribution
Platykurtic	flat distribution ("like a plate")
Positive skew	positive end of skewed distribution: tail pulled right in a positive direction
Rectangular distribution	all classes have same frequency
Skew	the degree a tail in a frequency distribution is pulled away from the mean
X-axis	the horizontal axis in a graph
Y-axis	the vertical axis in a graph

Study Question

Using the following data and the guidelines provided in this chapter...

89, 92, 83, 98, 98, 80, 89, 97, 83, 87, 86, 84, 97, 97,
99, 90, 95, 90, 91, 96, 95, 91, 91, 92, 94, 93, 94, 100

- a) ...to construct a grouped frequency distribution with $i=3$.
- b) ...to construct a histogram of this distribution.
- c) ...to construct a frequency polygon of this distribution.
- d) How would you describe this distribution? (What type?)

Sample Test Questions

1. Frequency distributions and graphs perform what statistical function?
 - A. reduce massive data sets to meaningful forms
 - B. infer characteristics of populations from samples
 - C. predict future trends or behaviors of subjects
 - D. depict significant differences between groups

2. A distribution has a range of 55 points. The best value for “i” is
 - A. 55
 - B. 11
 - C. 7
 - D. 2

3. In a positively skewed distribution,
 - A. the scores are “piled up” on the right
 - B. the right tail curves away from the x-axis
 - C. the long tail points to the right
 - D. the curve is narrow and pointed

4. Which of the following best describes a negatively skewed distribution?
 - A. The test was too easy for the sample of subjects
 - B. The test was too difficult for the sample of subjects
 - C. Scores on the test were evenly distributed among subjects
 - D. Few subjects scored high on the test.

