What does frequency mean?

What does distribution mean?

Objectives

- Construct a frequency distribution
- Know the characteristics of a frequency distribution

Aug 28-2:04 PM

Frequency Distributions

Statistics exists because of variation. The statistician's job is to comprehend variation by looking for structure. Frequency distributions are one method of examining a data set's structure. To examine structural characteristics, ask questions such as,

Where are most of the observations located?

Do the data cluster around one central point or are there several points that data seem to cluster around?

Do the data seem to be uniformly spread out over some interval or bunched in some range?

These questions all relate to the concept of "distribution."

Sep 8-9:59 AM

Definition

A **frequency distribution** summarizes data into classes and provides in tabular form a list of the classes along with the number of observations in each class.

Sep 8-10:07 AM

The process of refining information is interesting.

The analyst begins with raw data, then organizes that data by counting the number of observations in each classification.

In Table 3.1, the raw data consist of population counts in each state for the years 2000 and 2008.

By comparing the populations in 2008 with the populations in 2000, a population growth percentage can be computed for each state over the 8-year period.

Table 3.1 – Population Counts of Individual States in 2000 and 2008 (Rounded to the nearest 1000)

State	2000 Population	2008 Population	Growth	Percent Growth
Alabama	4447	4662	215	4.84
Alaska	627	686	59	9.41
Arizona	5131	6500	1369	26.69
Arkansas	2673	2855	182	6.81
California	33872	36757	2886	8.52
Colorado	4302	4939	637	14.81
Connecticut	3498	3574	85	2.43
Delaware	784	873	89	11.35
D.C.	572	592	20	3.50
Florida	15982	18328	2346	14.67

Sep 8-8:59 PM

The process of refining information is interesting. The analyst begins with raw data, then organizes that data by counting the number of observations in each classification. In Table 3.1, the raw data consist of population counts in each state for the years 2000 and 2008. By comparing the populations in 2008 with the populations in 2000, a population growth percentage can be computed for each state over the 8-year period.

Table 3.1 – Population Counts of Individual States in 2000 and 2008 (Rounded to the nearest 1000)

State	2000 Population	2008 Population	Growth	Percent Growth
Alabama	4447	4662	215	4.83
Alaska	627	686	59	9.41
Arizona	5131	6500	1369	26.68
Arkansas	2673	2855	182	6.81
California	33872	36757	2885	8.52
Colorado	4302	4939	637	14.81
Connecticut	3406	3501	95	2.79
Delaware	784	873	89	11.35
D.C.	572	592	20	3.50
Florida	15983	18328	2345	14.67

With the frequency distribution, we are able to see the broader structure of the data. It is now easy to see that most state population

There are only two steps in the construction of a frequency distribution.

Step 1: Choose the classifications.

Step 2: Count the number in each class.

For simple data, such as the results from tossing a coin, the choice of classifications is easy. Heads is one category and tails the other. However, for continuous data, such as weights, heights, and volumes, the choice of classification scheme becomes less obvious, since there are an enormous number of possibilities. There are two requirements that should be met when setting up the categories for classification: the categories must be both **mutually exclusive** and **exhaustive**. Essentially, this means categories should not overlap and should cover all possible values.

Since choosing the classification depends on whether the data are qualitative (nominal or ordinal) or quantitative (interval or ratio), the discussion of frequency distributions will be presented on the basis of these data types.

Sep 8-9:01 PM

Constructing Frequency Distributions for Qualitative Data

To construct a frequency distribution for qualitative data, **choose the categories** to classify the data. In many instances, the problem at hand will suggest the classification scheme. For instance, in the coin-tossing example, there are only two classes: heads and tails. If we are classifying students by gender, again there are only two classes: male and female. If we are classifying animals found on land, we could have six classes: mammal, reptile, avian, insect, arachnid, and amphibian. For qualitative data, it would be unusual if a reasonable set of categories is not relatively obvious.

After the categories have been chosen, **count the items belonging to each class** in order to construct the frequency distribution.

Sep 8-9:18 PM

Example 3.1: Constructing a Frequency Distribution

Apple Inc. introduced its third generation iPad in March of 2012. After months of anticipation, owners were thrilled with their new iPads. Two of the main features that separated the 3rd generation iPad from its predecessors were an improved camera and a higher resolution display. In spite of the excitement of the launch, there were still things that owners did not like about the new device. The following table contains the responses from a survey of 30 new iPad owners when asked what they dislike about the iPad.

Sep 8-9:20 PM

Table 3.3 – Survey Responses		
Cost of Device	Size/Weight	Excessive Heat from Device
Battery Life Too Short	Size/Weight	Cost of Device
Cost of Device	Battery Life Too Short	Excessive Heat from Device
Cost of Wireless Data Plan	Amount of Flash Memory Storage	Size/Weight
Cost of Device	Battery Life Too Short	Cost of Device
Amount of Flash Memory Storage	Cost of Device	Integration with Other Devices
Amount of Flash Memory Storage	Battery Life Too Short	Integration with Other Devices
Cost of Wireless Data Plan	Cost of Device	Excessive Heat from Device
Amount of Flash Memory Storage	Cost of Wireless Data Plan	Cost of Device
Cost of Device	Battery Life Too Short	Cost of Wireless Data Plan

Sep 8-9:20 PM

Table 3.3 – Survey Responses		
Cost of Device	Size/Weight	Excessive Heat from Device
Battery Life Too Short	Size/Weight	Cost of Device
Cost of Device	Battery Life Too Short	Excessive Heat from Device
Cost of Wireless Data Plan	Amount of Flash Memory Storage	Size/Weight
Cost of Device	Battery Life Too Short	Cost of Device
Amount of Flash Memory Storage	Cost of Device	Integration with Other Devices
Amount of Flash Memory Storage	Battery Life Too Short	Integration with Other Devices
Cost of Wireless Data Plan	Cost of Device	Excessive Heat from Device
Amount of Flash Memory Storage	Cost of Wireless Data Plan	Cost of Device
Cost of Device	Battery Life Too Short	Cost of Wireless Data Plan

Examining the table of responses, one can see that there are seven responses (or categories) that were given by the 30 new iPad owners. Summarizing the responses in a table (by counting the number of times that each reason was given), we can create a frequency distribution

Table 3.4 – Frequency Distribution of Survey Responses	
Response	Frequency
Cost of Device	9
Battery Life Too Short	5
Cost of Wireless Data Plan	4
Amount of Flash Memory Storage	4
Size/Weight	3
Excessive Heat from Device	3
Integration with Other Devices	2

Sep 8-9:20 PM

Table 3.4 – Frequency Distribution of Survey Responses	
Response	Frequency
Cost of Device	9
Battery Life Too Short	5
Cost of Wireless Data Plan	4
Amount of Flash Memory Storage	4
Size/Weight	3
Excessive Heat from Device	3
Integration with Other Devices	2

The frequency distribution makes it easy to see the greatest concern for iPad owners. From the table, it is evident that in spite of the anticipation and the rush for many consumers to purchase the new iPad, the cost of the device was the number one reason that new owners disliked it. We can also see that iPad integration with other devices is not much of a concern for new customers. These conclusions would have been much more difficult to make if we considered the raw data alone without consolidating the responses into a frequency distribution

Sep 8-9:22 PM

Example 3.2: A Model to Determine Virtual Security Practices

Table 3.5 – Frequency Distribution of Responses

"I am aware that there are measures that I can take to help protect my personal information on my personal computer."	
Strongly Agree	419
Slightly Agree	327
Neutral	250
Slightly Disagree	124
Strongly Disagree	85

Table 3.7 – Relative Frequency Distribution of Responses

"I am aware that there are measures that I can take to help protect my personal information on my personal computer."	
Strongly Agree	35%
Slightly Agree	27%
Neutral	21%
Slightly Disagree	10%
Strongly Disagree	7%

Sep 8-9:24 PM

Constructing Frequency Distributions for Quantitative Data

Example 3.3

Table 3.9 – Revenues of the Top 100 Companies on the Fortune 500 (Millions of Dollars)				
110	486	31	26	25
39	77	99	118	76
72	66	45	36	71
37	32	40	33	100
26	109	45	31	39
38	23	38	44	66
45	115	27	50	27
41	43	87	80	66
140	107	63	65	33
47	24	108	45	44
25	164	62	157	41
26	70	1	36	109
112	26	63	35	58
100	32	255	30	31
69	30	33	27	65
36	29	32	35	31
37	32	26	27	25
29	32	25	27	25
45	69	28	66	29
25	150	40	33	34

Of course, looking at the raw numbers is not very revealing. Given the number of observations, it is difficult to get a good idea of the measurements. Instead of trying to examine or analyze all 100 values, a frequency distribution is given in Table 3.10. Examining the frequency distribution, an analyst can look at a table divided into ten categories and ten frequencies, thus reducing its complexity.

Table 3.10 – Frequency Distribution	
Revenue (Millions of Dollars)	Number of Companies
0 to 41	50
41 to 81	29
82 to 122	14
123 to 163	3
164 to 204	1
205 to 245	0
246 to 286	1
287 to 327	0
328 to 368	0
369 to 409	1

Sep 8-9:26 PM

There are only two steps in the construction of a frequency distribution

Step 1: Choose the number of classifications.

Step 2: Count the number in each class.

Selecting the Number of Classes

Choosing the number of classes is arbitrary and should depend on the amount of data available. In general, the more observations one has in a data set, the more classes or intervals that can be used in the frequency table or histogram. Consequently, only very general guidelines exist. Generally, fewer than four classes would be too much compression of the data and greater than 20 classes provides too little summary information. To start, a good rule of thumb for the number of classes to create is to round $\sqrt{n}$ or $\sqrt[3]{2n}$ to the nearest whole number.

Sep 8-9:32 PM

Determining Class Width

However, a good starting point for determining class width is to divide the difference between the largest and the smallest observations by the number of classes.

$$\text{Class Width} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Number of Classes}}$$

Suppose we wanted to create a frequency distribution from the revenue data in Example 3.3. If there are to be 10 classes, determine a class width.

$$\text{Class Width} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Number of Classes}} = \frac{408 - 1}{10} = \frac{407}{10} = 40.7$$

Sep 8-9:35 PM

Finding Class Limits

After determining the class width, the class limits for all the classes in the frequency distribution must be specified. The lower class limit is the smallest number that can belong to a particular class, and the upper class limit is the largest number that can belong to a class.

Definition

The **class width** is the difference between the lower limits or upper limits of two consecutive classes of a frequency distribution.

The **lower class limit** is the smallest number that can belong to a particular class.

The **upper class limit** is the largest number that can belong to a particular class.

Sep 8-9:37 PM

Example 3.4

Create a frequency distribution using five classes for the following list of 3-D TV prices

Table 3.11 – 3-D TV Prices (in Dollars)				
1999	1599	1999	1899	1899
1699	1699	1899	1799	1685
1888	1787	1984	1699	1799
1699	1885	1999	1595	1757

$$\frac{1999-1595}{5}=80.8\approx 81$$

Table 3.12 – 3-D TV Prices	
Class	Frequency
\$1500 – \$1599	
\$1600 – \$1699	
\$1700 – \$1799	
\$1800 – \$1899	
\$1900 – \$1999	

Sep 8-9:39 PM

Calculating Class Boundaries

There are other characteristics of a frequency distribution that can be calculated once the basic frequency table has been constructed. Let's look at them.

The first calculation we will consider is that of class boundaries, which are similar to the class limits. The class boundaries split the difference in the gap between the upper limit of one class and the lower limit of the next class

$$\frac{10+11}{2}=10.5$$

Definition

A **class boundary** is the value that lies halfway between the upper limit of one class and the lower limit of the next class. After finding one class boundary, add (or subtract) the class width to find the next class boundary. The boundaries of a class are typically given in interval form: lower boundary–upper boundary.

Sep 8-9:40 PM

Example 3.5

Calculate the class boundaries for each class in the frequency distribution from Example 3.4.

Solution

Look at the first and second classes. The upper limit of class one is 1599. The lower limit of class two is 1600. Thus, the class boundary between the first two classes is calculated as follows.

$$\frac{1599+1600}{2}=1599.5$$

Recall that the class width is 100. Adding 100 to 1599.5 gives the next class boundary. You can repeat this step to find the remaining class boundaries.

Table 3.14 – 3-D TV Prices with Class Boundaries

Class	Frequency	Class Boundaries
\$1500–\$1599	2	1499.5–1599.5
\$1600–\$1699	5	1599.5–1699.5
\$1700–\$1799	4	1699.5–1799.5
\$1800–\$1899	5	1799.5–1899.5
\$1900–\$1999	4	1899.5–1999.5

Sep 8-9:42 PM

Calculating Class Midpoints

The **midpoint**, or class mark, of a class is the sum of the lower and upper limits of the class divided by 2. The midpoints are often used for estimating the average value in each class.

Formula: Class Midpoint

$$\text{Class Midpoint} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

Sep 8-9:43 PM

Example 3.6

Calculate the midpoint of each class in the frequency distribution from Example 3.4.

Solution

The midpoint is the sum of the class limits divided by two. For the first class, the midpoint is calculated as follows.

$$\frac{1500+1599}{2}=1549.5$$

We can use this same calculation to find the midpoints of the remaining classes. Another method is to add 100 (the class width) to the first midpoint, as we did with class boundaries.

Table 3.15 – 3-D TV Prices with Class Midpoints

Class	Frequency	Midpoint
\$1500–\$1599	2	1549.5
\$1600–\$1699	5	1649.5
\$1700–\$1799	4	1749.5
\$1800–\$1899	5	1849.5
\$1900–\$1999	4	1949.5

Sep 8-9:44 PM

Relative Frequency Distribution

Relative frequency represents the proportion of the total observations in a given class. The **relative frequency distribution** enables the reader to view the number in each category in relation to the total number of observations. Relative frequency is a standardizing technique. Converting the frequency in each class to a proportion in each class enables us to compare data sets with different numbers of observations.

Formula: Relative Frequency

The **relative frequency** of any class is the number of observations in the class divided by the total number of observations.

$$\text{Relative Frequency} = \frac{\text{Number in Class}}{\text{Total Number of Observations}}$$

Sep 8-9:44 PM

A relative frequency distribution of the revenue data from Example 3.5 is given in Table 3.16. Notice that the relative frequencies are obtained by dividing the frequencies in Table 3.10 by the total number of observations, which is 100.

Table 3.16 – Relative Frequency Distribution of Revenue Data

Revenue (Millions of Dollars)	Relative Frequency
0 to 40	$\frac{50}{100}=0.50$
41 to 81	$\frac{30}{100}=0.30$
82 to 122	$\frac{14}{100}=0.14$
123 to 163	$\frac{3}{100}=0.03$
164 to 204	$\frac{1}{100}=0.01$
205 to 245	$\frac{0}{100}=0.00$
246 to 286	$\frac{1}{100}=0.01$
287 to 327	$\frac{0}{100}=0.00$
328 to 368	$\frac{0}{100}=0.00$
369 to 409	$\frac{1}{100}=0.01$

Cumulative Frequency Distribution

The cumulative frequency distribution gives the reader an opportunity to look at any category and determine immediately the number of observations that belong to a particular category and all categories below it.

Definition

The **cumulative frequency** is the sum of the frequency of a particular class and all preceding classes.

Table 3.17 – Cumulative Frequency Distribution of Revenue Data

Revenue (Millions of Dollars)	Frequency	Cumulative Frequency
0 to 40	50	50
41 to 81	30	80
82 to 122	14	94
123 to 163	3	97
164 to 204	1	98
205 to 245	0	98
246 to 286	1	99
287 to 327	0	99
328 to 368	0	99
369 to 409	1	100

In this example, the reader can easily see in Table 3.17 that 97 out of 100 revenues are less than or equal to $163 million.

Sep 8-9:45 PM

Cumulative Relative Frequency

To obtain the cumulative relative frequency, add the relative frequencies of all preceding classes to the relative frequency of the current class.

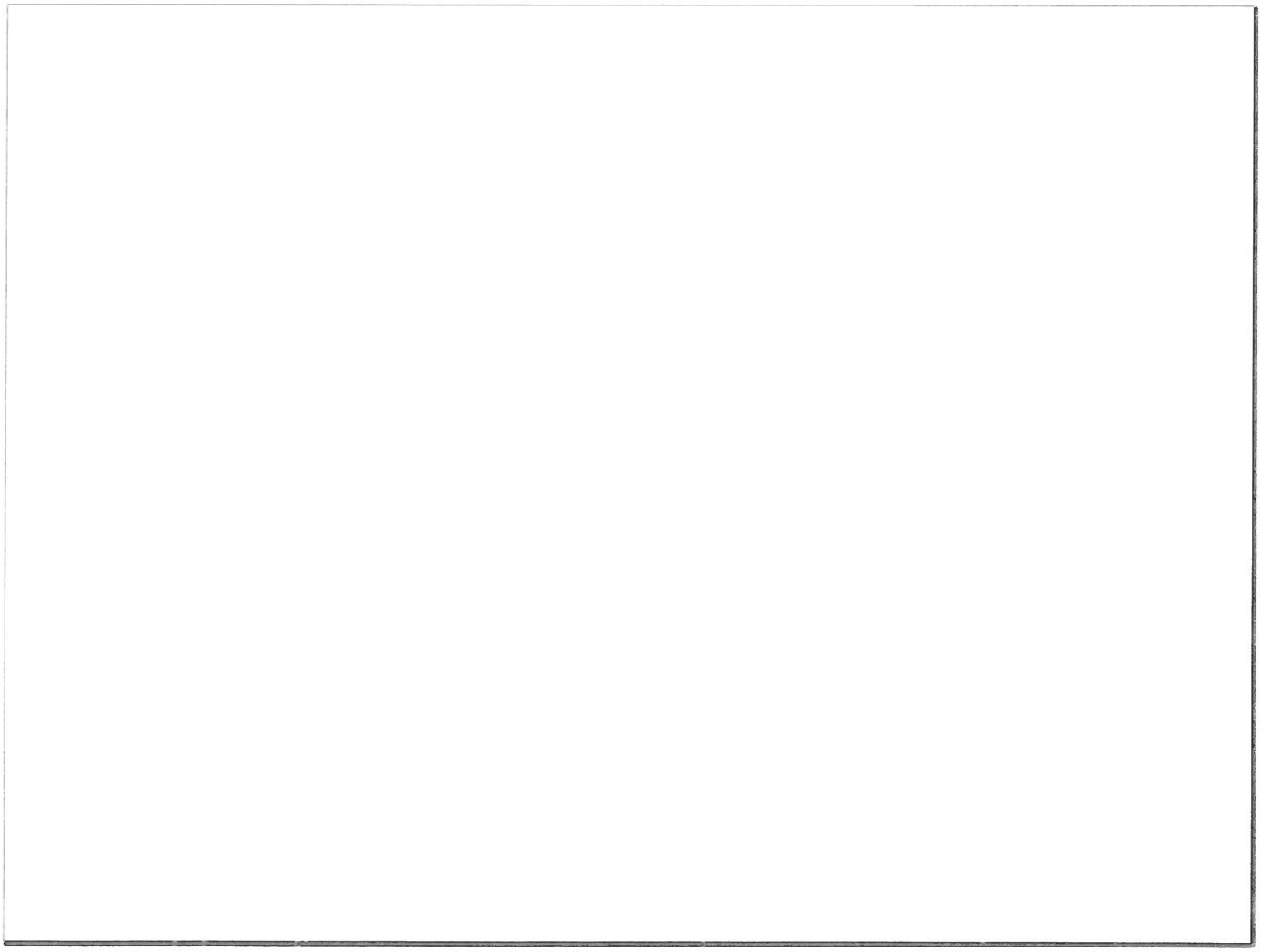
Definition

The **cumulative relative frequency** is the proportion of observations in a particular class and all preceding classes.

Table 3.18 – Cumulative Relative Frequency Distribution of Revenue Data

Revenue (Millions of Dollars)	Frequency	Relative Frequency	Cumulative Relative Frequency
0 to 40	50	0.50	0.50
41 to 81	30	0.30	0.80
82 to 122	14	0.14	0.94
123 to 163	3	0.03	0.97
164 to 204	1	0.01	0.98
205 to 245	0	0.00	0.98
246 to 286	1	0.01	0.99
287 to 327	0	0.00	0.99
328 to 368	0	0.00	0.99
369 to 409	1	0.01	1.00

Sep 8-9:45 PM



Sep 8-9:46 PM