

## II. Descriptive Statistics - Organization & Display

### A. Summarizing Qualitative Data

1. Frequency Distribution - tabular summary showing the absolute or relative number of elements in a data set that are in each of several mutually exclusive (non-overlapping) categories (or classes).

---

---

---

---

---

---

---

---

#### Types of Frequency Distributions

(Absolute) Frequency Distribution - summarizes the number of elements from the data set in each category

Relative Frequency Distribution - summarizes the proportion of elements from the data set in each category

Percent Frequency Distribution - summarizes the percentage of elements from the data set in each category

---

---

---

---

---

---

---

---

#### An Example of Three Types of Frequency Distributions

##### MLB HALL OF FAME MEMBERSHIP BY POSITION

Position	Frequency		
	Absolute	Relative	Percentage
C	11	0.089	8.9%
1	19	0.153	15.3%
2	14	0.113	11.3%
3	8	0.065	6.5%
S	17	0.137	13.7%
O	55	0.444	44.4%
D	0	0.000	0.0%
<b>Total</b>	<b>124</b>	<b>1.000</b>	<b>100.0%</b>

---

---

---

---

---

---

---

---

So how will a *Relative Frequency* or *Percentage Frequency* chart help me better interpret the data in this *absolute frequency distribution*?

**HOF ELIGIBLE PLAYERS BY POSITION**

Position	IN MLB HALL	NOT IN MLB
	OF FAME	HALL OF FAME
C	11	243
1	18	120
2	13	134
3	9	137
S	18	137
O	55	437
D	0	8
<b>Total</b>	<b>124</b>	<b>1216</b>

---

---

---

---

---

---

---

---

---

---

Making comparisons between the two groups of players who are eligible for the MLB HOF (those who are in and those who are not) is much easier using a *relative frequency* (why?)

**HOF ELIGIBLE PLAYERS BY POSITION - IN HOF vs. NOT IN HOF**

Position	IN MLB HALL OF FAME		NOT IN MLB HALL OF FAME	
	Absolute	Relative	Absolute	Relative
C	11	0.089	243	0.200
1	18	0.145	120	0.099
2	13	0.105	134	0.110
3	9	0.073	137	0.113
S	18	0.145	137	0.113
O	55	0.444	437	0.359
D	0	0.000	8	0.007
<b>Total</b>	<b>124</b>	<b>1.000</b>	<b>1216</b>	<b>1.000</b>

---

---

---

---

---

---

---

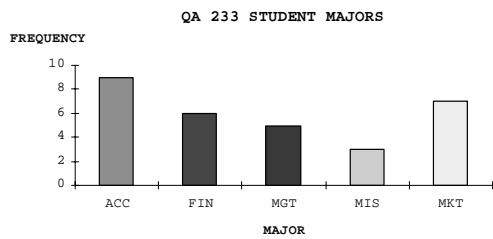
---

---

---

## 2. Graphical Displays of Frequency Distributions

a. Bar Chart - graphical depiction of the information in a frequency distribution




---

---

---

---

---

---

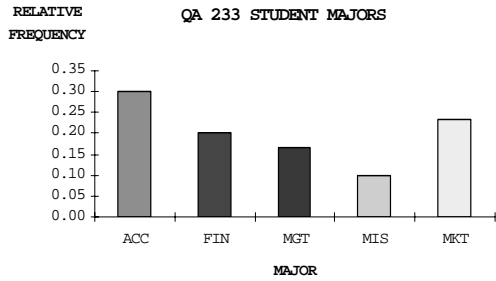
---

---

---

---

A Bar Chart for a relative frequency:



---

---

---

---

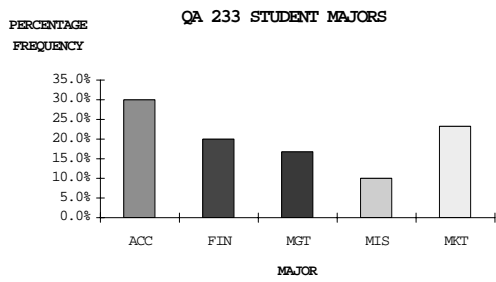
---

---

---

---

A Bar Chart for a percentage frequency:



---

---

---

---

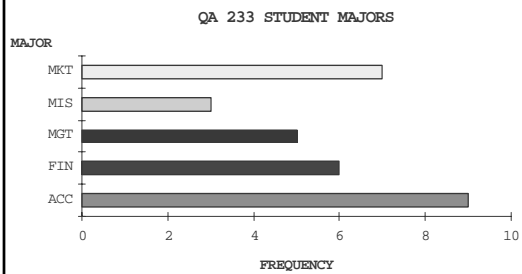
---

---

---

---

The previous displays are often referred to as 'Horizontal Bar Charts.' Bar charts can also be displayed vertically.



---

---

---

---

---

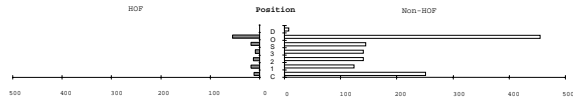
---

---

---

Why do we need *Relative Frequency* or *Percentage Frequency* charts?

ABSOLUTE FREQUENCY BY POSITION



Our population sizes are:

$$N_{\text{HOF}} = 124 \quad N_{\text{not HOF}} = 1263$$

---

---

---

---

---

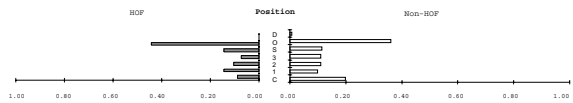
---

---

---

So how will a *Relative Frequency* or *Percentage Frequency* chart help me here?

RELATIVE FREQUENCY BY POSITION




---

---

---

---

---

---

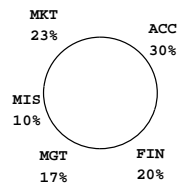
---

---

b. Pie Chart - circular graphical depiction of the information in a relative frequency distribution. The angle for each segment is calculated as

$$\text{Degree of Angle for the Segment} = 360 \cdot \left( \frac{\text{Proportion of Elements that Belong to the Corresponding Class}}{\text{Total}} \right)$$

SAMPLE OF QA 233 STUDENTS




---

---

---

---

---

---

---

---

## B. Summarizing Quantitative Data

1. Data Array - listing of the observed values which belong to a data set in either ascending or descending order.

Example - suppose that the sample of  $n = 20$  observations have been collected for some random variable  $x$ :

10 21 16 14 28 32 19  
17 26 31 24 18 12 14  
16 16 31 11 21 36

---

---

---

---

---

---

---

---

the (ascending) data array would be:

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

Suppose that the values 10, 11, 19, 24, 26, and 32 were taken from males, while all other observations were taken from females. How could we adapt this display to illustrate both variables?

Why not use color to differentiate gender for the observations?

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

where values corresponding to males are light blue and values corresponding to females are pink.

---

---

---

---

---

---

---

---

2. Frequency Distribution Table - tabular summary showing the relative number of elements in a data set that are in each of several mutually exclusive (non-overlapping) categories (or class). The steps for building a frequency distribution for quantitative data are:

- Find the range ( $r$ ) of the data set

$$r = \frac{\text{Largest Value}}{\text{in Raw Data}} - \frac{\text{Smallest Value}}{\text{in Raw Data}}$$

- Decide on the number of mutually exclusive and collectively exhaustive classes  $K$  (usually between 5 and 15). Can use *Sturges' Rule*

$$K = 1 + 3.322(\log_{10} n)$$

the *Sample Size Square Root Rule*

$$K = \sqrt{n}$$

or a combination of common sense and trial and error.

---

---

---

---

---

---

---

---

- Establish the approximate class width (c.w.). This is calculated as

$$c.w. \approx \frac{\text{Largest Value in Raw Data} - \text{Smallest Value in Raw Data}}{\text{Number of Classes Desired}} = \frac{r}{K}$$

- The lower class limits (LCLs) are then equal to i) the smallest value in the raw data for the first class and ii) the smallest value in the raw data + (k - 1)(approximate class width) for the k<sup>th</sup> class (k = 1,...,K). The upper class limit (UCL) for the k<sup>th</sup> class is the lower limit for the class + the approximate class width.
- Record the frequency/number of sample elements whose values fall into each of the classes.

---

---

---

---

---

---

---

---

---

---

Example - for the previous sample of n = 20 observations that have been collected for some random variable x:

- Find the range of the data set.

$$r = \frac{\text{Largest Value in Raw Data} - \text{Smallest Value in Raw Data}}{\text{in Raw Data}} = 36 - 10 = 26$$

- Decide on the number classes. Since there are only 20 observations and

$$K = \sqrt{n} = \sqrt{20} = 4.47$$

we will use K = 5 classes.

- Establish the approximate class width as

$$c.w. \approx \frac{r}{K} = \frac{26}{5} = 5.2$$

- Then the five sets of class limits are  
10.0 - 15.2, 15.2 - 20.4, 20.4 - 25.6, 25.6 - 30.8, 30.8 - 36.0

---

---

---

---

---

---

---

---

---

---

- The frequencies/number of sample elements whose values fall into each of the classes are:

Class	Frequency
10.0 - 15.2	5
15.2 - 20.4	6
20.4 - 25.6	3
25.6 - 30.8	2
30.8 - 36.0	4

Of course, we can also produce relative and percent frequency distributions for quantitative data as well:

Class	Frequency	Relative Frequency	Percentage
10.0 - 15.2	5	5/20=.25	25.0%
15.2 - 20.4	6	6/20=.30	30.0%
20.4 - 25.6	3	3/20=.15	15.0%
25.6 - 30.8	2	2/20=.10	10.0%
30.8 - 36.0	4	4/20=.20	20.0%

---

---

---

---

---

---

---

---

---

---

What if we decided to round the class width up from 5.2 to 6.0?

Our classes will now cover 30 units instead of 26 units (why?) or four more units than necessary - we must adjust for that by starting two units below the minimum when we define our class intervals (again, why?)

<u>Class</u>	<u>Frequency</u>
8.0 - 14.0	5
14.0 - 20.0	8
20.0 - 26.0	4
26.0 - 32.0	5
32.0 - 38.0	2
Total	24 $\neq$ n!

What is wrong?

---

---

---

---

---

---

---

---

Suppose we decide to also subtract a small amount (say, 0.1) from each upper class limit to correct for the overlap between the classes. The frequency distribution would look like this

<u>Class</u>	<u>Frequency</u>
8.0 - 13.9	3
14.0 - 19.9	8
20.0 - 25.9	3
26.0 - 31.9	4
32.0 - 37.9	2
Total	20 = n!

Why did this work?

---

---

---

---

---

---

---

---

We could have added a small amount (say, 0.1) to each lower class limit to correct for the overlap between the classes. The frequency distribution would look like this

<u>Class</u>	<u>Frequency</u>
8.1 - 14.0	5
14.1 - 20.0	6
20.1 - 26.0	4
26.1 - 32.0	4
32.1 - 38.0	1
Total	20 = n!

Why did this work?

---

---

---

---

---

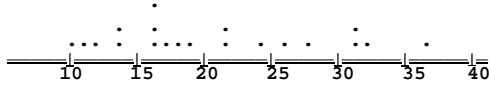
---

---

---

3. Dot Plot - tabular summary in which the relative number of elements in a data set that take on each potential value of the variable are represented by dots.

Example - for the previous sample of  $n = 20$  observations have been collected for some random variable  $x$ :




---

---

---

---

---

---

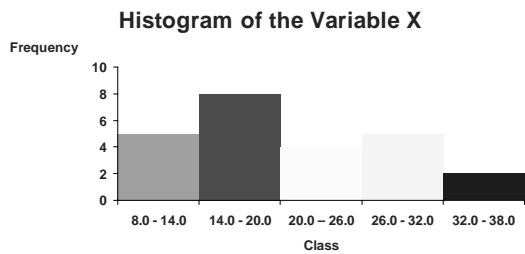
---

---

---

---

4. Histogram - bar chart for the frequency distribution of a quantitative variable



This is a histogram of our original frequency distribution of random variable  $X$ .

---

---

---

---

---

---

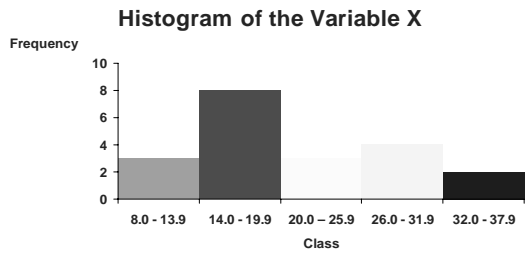
---

---

---

---

If we rounded the class width up from 5.2 to 6.0 and subtract a small amount (say, 0.1) from each upper class limit to correct for the overlap between the classes:



this is the resulting histogram - how does it differ from the previous histogram?

---

---

---

---

---

---

---

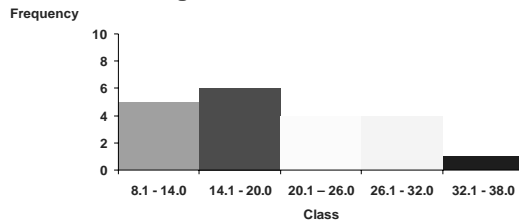
---

---

---

If we rounded the class width up from 5.2 to 6.0 and add a small amount (say, 0.1) to each lower class limit to correct for the overlap between the classes:

**Histogram of the Variable X**



this is the resulting histogram - how does it differ from the previous histograms?

---

---

---

---

---

---

---

---

---

---

**5. Line Graph - connected points used to identify trends in the consecutive values of the data**

Example - for the 1999 NFL Season, the following table summarizes the points scored and allowed for each game by the Cincinnati Bengals:

Game Number	Points Scored	Points Allowed	Home/Away
1	35	36	a
2	7	34	h
3	3	27	a
4	10	38	h
5	18	17	a
6	3	17	h
7	10	31	a
8	10	41	h
9	20	37	a
10	14	24	h
11	31	34	h
12	27	20	a
13	44	30	h
14	44	28	h
15	0	22	a
16	0	22	a
17	7	24	a

---

---

---

---

---

---

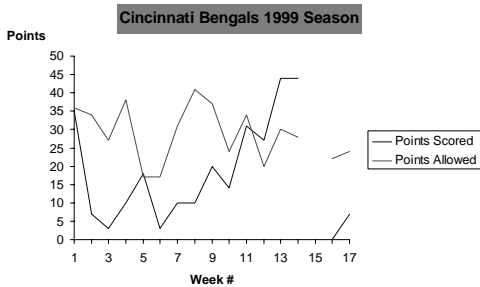
---

---

---

---

A line graph summarizing the the points scored and allowed for each game by the Cincinnati Bengals could look like this:



Note that a line chart for which the x-axis represents time is often called a *Runs Chart*.

---

---

---

---

---

---

---

---

---

---

6. Ogive - line graph of a cumulative frequency distribution of a quantitative variable

Example - for the previous sample of  $n = 20$  observations have been collected for some random variable  $x$ , we have the following various cumulative frequency distributions:

Limits	Frequency	Frequency	Frequency	Frequency
15.2	5	5	0.25	25.0%
20.4	6	11	0.55	55.0%
25.6	3	14	0.70	70.0%
30.8	2	16	0.80	80.0%
36.0	4	20	1.00	100.0%

$n = 20$

---

---

---

---

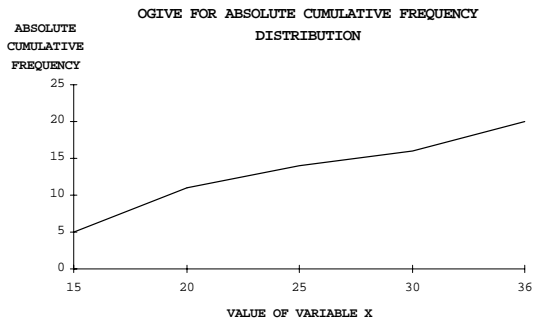
---

---

---

---

The ogive for the absolute cumulative frequency distribution would look like:




---

---

---

---

---

---

---

---

7. Density Curves - line graph for which the y-axis represents relative (proportional) or percentage frequency

Note that because proportions must add to 1.00 (if all outcomes in our data are included on the graph), the area under the density curve *must equal 1*. This allows us to interpret areas under a density curve as probabilities!

8. Pareto Charts - a bar graph or histogram overlaid with a cumulative line graph

Note that these charts are often used to isolate problems in quality control.

---

---

---

---

---

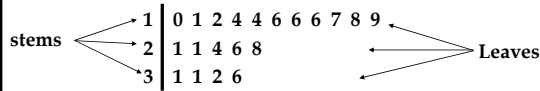
---

---

---

9. Stem (& Leaf) Plot - display, by class, of the actual values in a data set in a histogram-like manner. Steps in constructing a stem & leaf plot are:

- record the leading digit(s) that occur in the data set on the left side of a vertical line. These are referred to as 'stems.'



- record the trailing digit that occurs in the data set on the right side of a vertical line. These are referred to as 'leaves.'

---

---

---

---

---

---

---

---

---

---

---

---

10. Simultaneous Display of Multiple Variables

a. Crosstabulation - table displaying the frequency of occurrence for classes of values for two (or more) variables simultaneously.

- Example

		AGE				Total
		0-20	21-40	41-60	61-80	
I	\$0,000 -					
N	\$40,000	25	33	61	19	138
C	\$40,000 -					
O	\$80,000	44	19	61	23	147
M	\$80,000 +	81	17	71	44	213
E						
	Total	150	69	193	86	498

---

---

---

---

---

---

---

---

---

---

---

---

b. Scatter Diagram - graphical simultaneous presentation of the values of two variables on a Cartesian coordinate system

- Example: suppose we have collected the following sample of 6 observations on age and income:

AGE	INCOME (in \$1,000's)
25	21
47	57
35	44
62	65
41	64
33	23

---

---

---

---

---

---

---

---

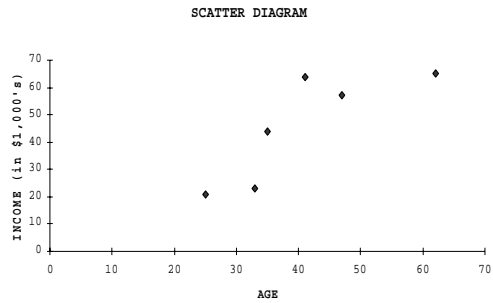
---

---

---

---

The resulting scatter diagram would look like this:




---

---

---

---

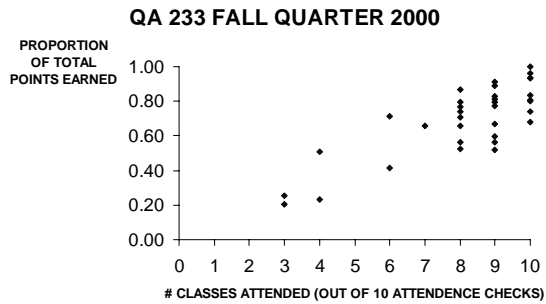
---

---

---

---

Here's a very interesting and important scatter diagram:



How would you interpret this scatter diagram?

---

---

---

---

---

---

---

---

c. Star Glyphs - graphical simultaneous presentation of the values of more than two variables on a coordinate system

- Example: suppose we have collected the following sample of 6 observations on age, income, years of education, and years married:

AGE	INCOME (in \$1,000's)	YEARS OF EDUCATION	YEARS MARRIED
25	21	14	3
47	57	17	20
35	44	12	7
62	65	10	33
41	64	18	15
33	23	16	17

---

---

---

---

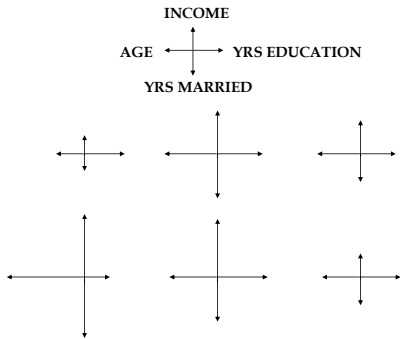
---

---

---

---

The six star glyphs could look like this:




---

---

---

---

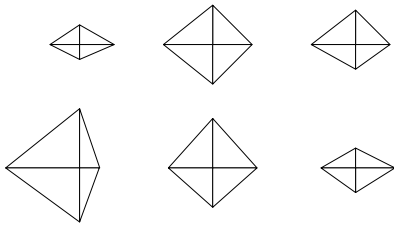
---

---

---

---

Sometimes they are filled and displayed in this manner:




---

---

---

---

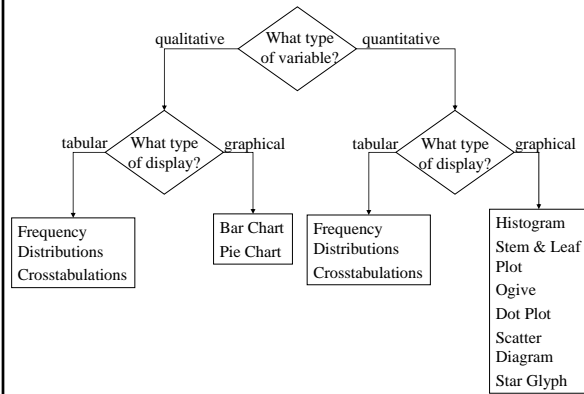
---

---

---

---

Summary of tabular and graphical displays:




---

---

---

---

---

---

---

---