

Basic Concepts

A horizontal bar spanning the width of the slide, divided into three colored segments: blue, orange, and pink.

Statistics

Statistics is a branch of applied mathematics which deals with the collection, organization, presentation, analysis and interpretation of data to draw a valid conclusion.

It has various uses in sciences, education, psychology, sociology, management, economics and business that may include marketing, production and finance.

Types of Statistics

Descriptive Statistics deals with the collection and presentation of data and collection of summarizing values to describe its group characteristics.

Inferential Statistics deals with predictions and inferences based on analysis and interpretation of the results of the information gathered by statistician.

Variable

A numerical characteristic or attribute associated with the population being studied.

Types of Variables

Categorical or Qualitative Variable are classified according to some attributes or categories. Categories may be ordered, which may or may not be assigned specific numerical values.

Examples:

Gender, Blood Type, Year Level, Performance Rating (Poor, Fair, Good, Very Good, Excellent)

Types of Variables

Numerical – valued or Quantitative Variables are classified according to numerical characteristics. Numerical – valued are often grouped into class intervals.

Examples:

Age In Years - 5-9, 10-14

Height in cm – 100-149

Grade in Stat – 1.0, 1.25, 1.5

Numerical-Valued Variables Classifications

Discrete – is a variable whose values are obtained by counting.

Example: Number of children, Number of Male Students in Stat Class

Continuous – is a variable whose values are obtained by measuring.

Example: Temperature, Distance, Height, Weight, Age

Room Number	Gender	Age in Years	Job Description	Civil Status
101	M	25	IT Specialist	Single
102	F	32	Programmer	Married
103	F	38	Animator	Separated
104	M	46	Data Analyst	Married

1. How many variables?
2. How many are the categorical or qualitative variables?
3. How many are the numerical-valued or quantitative variables?

Scales of Measurement

- Nominal Scale is a measurement scale that classifies elements into two or more categories or classes, the numbers indicating that the elements are different but not according to order.
 - Example: Gender, Civil Status
- Ordinal Scale is a measurement scale that ranks individuals in terms of the degree to which they possess a characteristic of interest
 - Example: Class Standing (Excellent, Good, Poor)
- Interval Scale is a measurement scale, in addition to ordering scores from high to low, it also establishes a uniform unit in the scale so that any equal distance between two scores is of equal magnitude. No absolute zero in this scale.
 - Example: Aptitude Score(80-90), Aptitude Score(90-100)
- Ratio Scale is a measurement scale, in addition to being an interval scale, that also has an absolute zero in the scale.
 - Example: height, weight, speed, area

Definitions

Population is defined as groups of people, animals, places, things or ideas to which any conclusions based on characteristics of a sample will be applied.

Sample is a subgroup of the population.

Parameter is a numerical measure that describes a characteristics of a population.

Example: The population mean of the water bill of the residents in Valencia City is Php 500.00.

Statistic is a numerical measure that is used to describe a characteristic of a sample.

Example: The sample mean of the water bill of the 50 residents in Valencia City is Php 325.00.

Methods of Collecting Data



Collection of Data

a process of gathering data and processed it to become information to answer stated research questions, test hypotheses, and evaluate outcomes.

Methods of Collecting Data

1. Direct or Interview Method.

- Researcher personally interviews the respondent.
- One of the most effective methods of collecting original data.
- The method is appropriate to use if needed information is minimal and the number of respondents is few.
- Well-trained interviewers may do the interview to obtain accurate responses.

Advantages	Disadvantages
It can give complete information needed in the study	Very costly and time consuming if the number of respondents are very large and they are living very far apart
The interviewers help the respondent in explaining further the questions.	It can yield inaccurate information since the interviewer can influence the respondent's answer through his facial expression, tone of voice, or wording of the questions.
	The interview is subject to interviewer bias and respondent bias.

Methods of Collecting Data

2. Questionnaire Method.

- One of the easiest methods of data gathering.
- A list of well planned questions written on the paper which can be either personally administered or mailed by the researcher to the respondents

Advantages	Disadvantages
Less expensive since questionnaires can be distributed personally or by mail	It takes time to prepare questionnaire that is precise, clear and self-explanatory.
Less time-consuming since it can be distributed over a wider geographical area in a shorter time.	It cannot be accomplished by illiterates.
It can give confidential responses since the respondents can answer the questionnaire privately	It has high proportion of nonresponse or nonreturn.
The answers obtained are free from any influence coming from the interviewer	It tend to give wrong information since answers cannot be corrected right away. It tends to give

Methods of Collecting Data

Forms of Questionnaire:

1. Guided-Response Type
2. Recall Type
3. Recognition Type (Through Figures)
4. Dichotomous Type
5. Multiple Choice Type
6. Multiple Response Type
7. Free Response Type
8. Rating Scale Type

Methods of Collecting Data

3. Empirical Observation Method

- This method is utilized to gather data regarding attitudes, behavior, values and cultural patterns of the samples under investigation.
- Obtaining data through seeing, hearing, testing, touching and smelling.
- Additional information maybe gathered which cannot be obtained using the other methods such as questionnaire.
- Commonly used in psychological and anthropological studies.
- The observer may participate in the activities of the group being studied (**participant observation**) or he may just be a bystander only (**non-participant observation**). This type of observation is called **controlled observation**.

Methods of Collecting Data

4. Test Method

- This method is widely used in psychological research and psychiatry. Standard test are used because of their validity, reliability and usability.
- Example: Aptitude tests, IQ tests and Achievement Test

5. Registration Method

- The examples of data gathered using this method are those that are obtained from the PSA, LTO, DEPED, CHED and many other government agencies.

6. Mechanical Devices

- Mechanical devices that can be used for social and educational research in data gathering are the camera, projector and etc. For chemical, biological, and medical researches are the x-ray machine, ultrasound, CT Scan and etc.

Methods of Collecting Data

7. Phone Interview

- Uses phone to interview the respondents
- The method is biased because people with no phone cannot have the chance to be included in the study.
- For some individuals, phone interview is not considered polite and proper.
- Some people feel offended when interviewed through the phone only.
- The researcher can never be sure if he or she is interviewing the right person since there is no personal contact and exchange of ideas.

Methods of Collecting Data

8. Experiments

- This method is applied to collect or gather data if the investigator wants to control the factors affecting the variable being studied.
- Example: When the researcher aims to determine the different factors affecting the academic performance of the students such as methods or approaches used in teaching

Topic 3: Sampling Techniques



Sampling Techniques

Random Sampling

- All the members of the population have equal chances of being included in the study.
- This is applicable if the target population is not classified into different clusters, sections, levels or classes.
- The method is easy to apply, but not when the population is very large, say a thousand or more.

Random Sampling

Lottery Method

➤ It is the most common and easiest method of random sampling.

Example: Names of the respondents are written on small pieces of paper then rolled and placed in a jar. The respondents who are included in the study are those whose names are written on the pieces of paper picked at random from the jar.

Sampling Techniques

Systematic Sampling

➤ This method involves selecting every n^{th} element of a series representing the population.

Example: The value of n may be obtained by dividing the total number of elements in the population by the desired sample size.

$$n = \frac{N}{n} = \frac{100}{10} = 10^{\text{th}}$$

where n is sample size, N is population size

The 10 sample units would be the persons holding the following numbers: 10, 20, 30,100 at 10 sample units. Variations may be added by choosing a random start. Take 10 pieces of paper and number it from 1 to 10. By lottery method, take one number as starting point. For example if you picked no. 7 then the 10 sample units should be 7, 17, 27, 37, 47, 57, 67, 77, 87, 97

Systematic Sampling

Stratified Random Sampling

- This method is applied when the population is divided into different strata or classes and each class must be represented in the study.

$$n = \frac{N}{1 + N(e)^2}$$

where **n** is sample size,
N is population size and
e is the margin of error

Systematic Sampling

Stratified Random Sampling

Example: Suppose a researcher wants to determine the average income of the families in a Barangay Poblacion having 3,000 families that are distributed in 5 Purok. Compute for the sample size n at 5% margin of error.

$$n = \frac{N}{1 + N(e)^2}$$

$$n = \frac{3000}{1 + 3000(.05)^2}$$

$$\mathbf{n = 353}$$

Systematic Sampling

Required Sample Size from each Purok

Purok	Population	Percentage	n_k
1	800	27%	$.27 \times 353 = 95$
2	400	13%	$.13 \times 353 = 46$
3	500	17%	$.17 \times 353 = 60$
4	600	20%	$.20 \times 353 = 71$
5	700	23%	$.23 \times 353 = 81$
N	3000	100%	353

Systematic Sampling

Cluster Sampling

- The cluster sampling technique is appropriate when the geographical area where the study is too big and the target population is too large.
- The selection of sample units is not by individuals but by groups called clusters.
- The area is divided into clusters then select a desired number of clusters at random

Example: A doctor wants to make a nationwide study on the correlation between smoking and death rate. He decides to take the **13 regions** of the country which can be considered as **clusters**. If the 3 of the 13 clusters or regions are the desired sample units, then the 3 clusters are selected using lottery method. **All residents of the selected 3 clusters** are included in the study.

Sampling Techniques

Purposive Sampling

- The respondents of the study are chosen based on their knowledge of the information required by the researcher.

Example: Suppose a researcher wants to make a historical study about Town A. The target population is the senior citizens of the town since they are the most reliable persons who know the history of the town. If there are 2,000 senior citizens and a margin of error of 3% is allowed, the sample size is 714.

Sampling Techniques

Quota Sampling

➤ This technique is commonly used in opinion polls.

Example: Suppose a salesman is required to gather information as to the most common hair shampoo used by the female Filipino clients. If he wants 2,000 sample units and should do the survey in a short time, he can station himself in a public place such as the park or in malls then ask the females he meets regarding the hair shampoo they usually use. After meeting the required number of sample points, the researcher is done with his collection of data.

Sampling Techniques

Convenience Sampling

➤ This technique is resorted to by researchers who need the information the fastest way possible.

Example: The phone can be used to interview the respondents about their opinions on a certain issue. This method may be fast but biased because those who have no phones do not have the chance to be included in the study.

Another example is the case of a teacher who makes a research which requires the inclusion of students as respondents. Conveniently, the teacher may use his own students as respondents.

Topic 4: Statistical Notations and Operations



Notations

- Statistics involves summation of several numbers, which necessitates the use of symbols to express numerical ideas, minimizing the excessive use of words. The symbols used are referred to as **standard notation**.
- A shorthand notation for such a sum involves **sigma notation**. The term comes from the use of the upper case Greek letter sigma Σ .

The expression $\sum_{i=1}^n x$ reads “the summation of x; i is from 1 to n”

Notations

The expression $\sum_{i=1}^n x$ reads “the summation of x ; i is from 1 to n ”

- This means, **taking the sum of n number of observations** or values of the **variable represented by x** . The subscript i represents the order of an observation, whether it is the first, second, third, or the last. The notation **$i=1$ under the summation sign Σ , denotes the lower limit** and indicates the start of counting. The number above, **n , is the upper limit** and tells the total number of observations to be added.

Example

Find the following sums:

1. $\sum_{i=1}^6 i$

2. $\sum_{i=1}^5 \frac{1}{i}$

Answer

Find the following sums:

1. $\sum_{i=1}^6 i = 1+2+3+4+5+6 = 21$

2. $\sum_{i=1}^5 \frac{1}{i} = 1+1/2+1/3+1/4+1/5 = 137/60$

Properties of Sums

A. Sum of the Product of a Constant and a Variable

The sum of the product of a constant and a variable is equal to the constant multiplied by the sum of the variables.

Notation: $\sum_{i=3}^7 cx_i$ and $c = 5$

i	x_i	c	cx_i
3	3	5	15
4	4	5	20
5	5	5	25
6	6	5	30
7	7	5	35
		Sum=	125

Properties of Sums

B. Sum of a Constant

The sum of a constant taken from 1 to n equals n times the constant symbols

Notation : $\sum_{i=1}^n C_i$ or

$$\sum_{i=1}^n C_i = nC, \text{ c is constant}$$

The sum of this set of values: 6, 6, 6, 6, 6 is 30. The five values are constant and equal to 6. The sum is equal to the product of 5 and the constant is 6.

Properties of Sums

C. Summation of a Sum

The summation of the sum of several variables is equal to the sum of terms taken separately. In symbols:

Notation : $\sum_{i=1}^n (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \dots + (x_n + y_n)$

$$\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$$

Properties of Sums

Given the table of values below, evaluate the summations that follow:

i	x_i	y_i
1	2	3
2	1	4
3	4	5
4	3	2

1. $\sum_{i=1}^2 (x_i + y_i)$ 2. $\sum_{i=1}^4 (x_i + y_i)$

Answer

$$1. \sum_{i=1}^2 (xi + y_i) = 5 + 5 = 10$$

$$2. \sum_{i=1}^4 (xi + y_i) = 5 + 5 + 9 + 5 = 24$$

Properties of Sums

D. Summation of a Variable and a Constant

The summation of a variable and a constant is equal to the sum of the variables added to the product of n and the constant. Notation : $\sum_{i=1}^n x_i + c = \sum_{i=1}^n x_i + nc$

i	x_i	c	$x_i + c$
1	3	6	9
2	6	6	12
	$\sum_{i=1}^2 x_i = 9$	$\sum_{i=1}^2 c = 2(6)=12$	$\sum_{i=1}^2 x_i + c = 9+12=21$

Properties of Sums

Given the table of values below, evaluate the summations that follow:

i	x_i	y_i
1	2	3
2	1	4
3	4	5
4	3	2

1. $\sum_{i=1}^2 x_i + 3$ 2. $\sum_{i=1}^4 y_i + 5$

Answer

$$1. \sum_{i=1}^2 x_i + 3 = 3 + 2(3) = 9$$

$$2. \sum_{i=1}^4 y_i + 5 = 14 + 4(5) = 34$$

Properties of Sums

E. Sum of the Squares of Variables

The sum of the square of n number of observation

Notation: $\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \dots + x_n^2$

Given the set of numbers 3,4,and 5. The sum of squares is:

$$\sum_{i=1}^3 x_i^2 = 3^2 + 4^2 + 5^2 = 9 + 16 + 25 = 50$$

Properties of Sums

F. Square of the Sum of Variables

The square of the sum of n number of observations

Notation : $[\sum_{i=1}^n x_i]^2 = [x_1 + x_2 + \dots + x_n]^2$

Given the set of numbers 3,4,and 5. The sum of squares is:

$$[\sum_{i=1}^3 x_i]^2 = [3+4+5]^2 = 144$$

Properties of Sums

G. Sum of a Product

The sum of the product of n pairs of variables

Notation : $\sum_{i=1}^n (x_i) (y_i) = (x_1)(y_1) + (x_2)(y_2) + \dots + (x_n)(y_n)$

i	x_i	y_i	$x_i y_i$
1	1	3	3
2	2	4	8
3	3	5	15
4	4	6	24
5	5	7	35
$\sum_{i=1}^5 (x_i) (y_i)$		Sum=	85

Topic 5: Organization and Presentation of Data



3 Different Ways of Presenting Data

1. Textual Form – the presentation is in narrative or paragraph form. The data are within the text of the paragraph. However, it can present a more comprehensive picture of the data because of further written explanation of its nature.

3 Different Ways of Presenting Data

2. Tabular Method – makes use of rows and columns like a frequency table or frequency distribution. The data are presented in a systematic and orderly manner which may facilitate the comprehension and analysis of the data presented.

3 Different Ways of Presenting Data

3. Graphical Form – a graph is a pictorial or geometrical representation of given data. They may be in the form of a frequency polygon, bar graph, stem and leaf display, pie graph or pictograph.

Frequency Distribution Table

- It is a device for organizing and representing grouped data. There is no need to construct the frequency distribution table if the number of observations is **less than 30**.
- Data that are not presented in a frequency distribution table are called **ungrouped data**.

Steps in Constructing Frequency Distribution Table

1. Find the range **R**

$$\mathbf{R = Highest\ Score - Lowest\ Score}$$

2. Compute the number of intervals **n**

$$\mathbf{n = 1 + 3.3 \log N}$$

Where

n is the number of class intervals

N is the population or total number of observations

Steps in Constructing Frequency Distribution Table

3. Compute the class size i

$$i = R/n$$

4. Using the lowest score as lower limit, add $i-1$ to it to obtain the higher limit of the desired class interval.

5. The lower limit of the second interval is obtained by adding the class size to the lower limit of the first interval then add $i-1$ to the result to obtain the higher limit of the second interval.

Steps in Constructing Frequency Distribution Table

6. Repeat step 5 to obtain the third class interval and so on and so forth.
7. When the n number of class intervals is completed, determine the frequency for each class interval by counting the elements.

Try this!

Construct the frequency distribution table of the data below:

Ages of Patients in Hospital ABC as of February 2020

25	28	27	30	32	25	31	26	29	6
31	20	21	32	18	50	53	60	50	54
45	40	37	25	20	27	32	24	29	30
25	24	10	12	15	28				

Answer

25	28	27	30	32	25	31	26	29	6
31	20	21	32	18	50	53	60	50	54
45	40	37	25	20	27	32	24	29	30
25	24	10	12	15	28				

1. $R=60-6$ $R=54$
2. $n=1+3.3 \log 36$ $n=6.1357$ $n=6$
3. $i=54/6$ $i=9$

Answer

1. $R=60-6$ $R=54$
2. $n=1+3.3 \log 36$ $n=6.1357$ $n=6$
3. $i=54/6$ $i=9$
4. $N=36$

Ages of Patients in Hospital ABC as of February 2020

Ages in Years	Tally Marks	Frequency
60 – 68		1
51 – 59		2
42 – 50		3
33 – 41		2
24 – 32	- - -	20
15 – 23		5

Class Mark and Class Boundary

1. Class Mark is the midpoint of a class interval. To obtain this point, **add the lower limit and the upper limit then divide by 2.**
2. Class Boundary is also known as the exact limit, and it can be obtained by **subtracting 0.5 from the lower limit** of each interval and **adding 0.5 to the upper limit.**

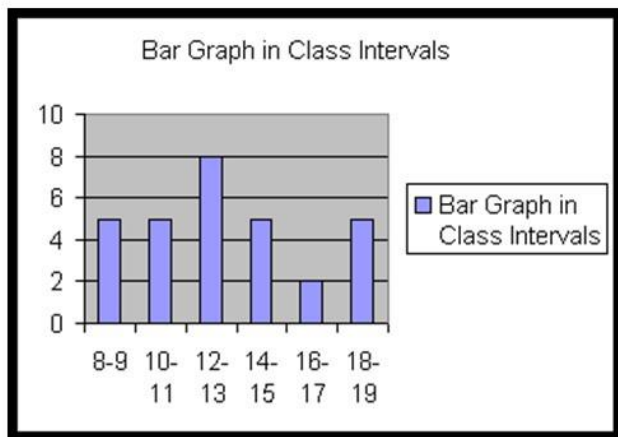
Class Mark and Class Boundary

Ages of Patients in Hospital ABC as of February 2020

Ages in Years	Class Mark	Class Boundary
60 – 68	$(60+68)/2=64$	55.5 – 68.5
51 – 59	55	50.5 – 59.5
42 – 50	46	41.5 – 50.5
33 – 41	37	32.5 – 41.5
24 – 32	28	23.5 – 32.5
15 – 23	19	14.5 – 23.5
6 - 14	10	5.5 – 14.5

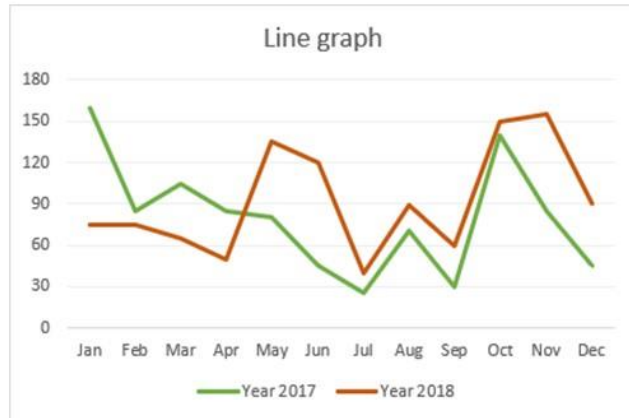
Graphs

1. Bar Graph are usually applied to compare data and to determine which class or interval is more common or frequently appears in text.



Graphs

2. Line Graph show trends and increase in sales, improvement of score, rise or fall of temperature of patients, enrollment of students in certain courses and comparison of population per year.

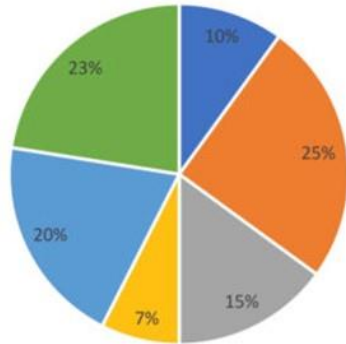


Graphs

3. Pie Graph is helpful when the categories represent parts of the whole.

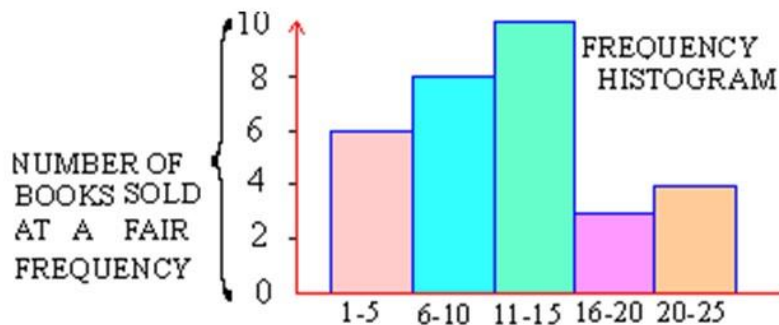
Grocery Store Sales (Jan 2010)

■ Pharmaceuticals ■ Produce ■ Dry Foods
■ Canned Foods ■ Dairy Products ■ Frozen Food



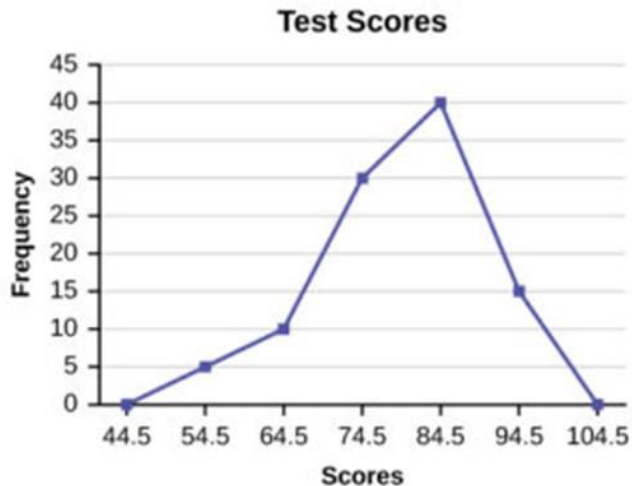
Graphs

4. Frequency Histogram – frequency is represented by points in the vertical axis and the class intervals in the horizontal axis. The ordered pair of points in the vertical and horizontal axes is plotted by placing bars in the graph area. Each bar represents the class interval with its corresponding frequency.



Graphs

5. Frequency Polygon graph is just a line connecting the points representing the important data in the xy-plane.



Graphs






6. The Stem and Leaf – is another visual illustration of the distribution of data. This form is feasible only for small number of observations with at least two digit numbers

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

Graphs

7. Pictograph – picture symbols are used to illustrate or represent the data under consideration. Usually, in depicting population data, the figures of persons are used or the data on car sales.

Small Towns	Number of illiterate children
Melrose	
Marengo	
Midway	
Parral	
Rushville	

Graphs

8. Cumulative Frequency Ogive – commonly used in statistical reports and text.

Less Than Cumulative Frequency
"<CF" – starting with the interval with Lowest Score to Highest Score

Greater Than Cumulative Frequency
">CF" – starting with the interval with Highest Score to Lowest Score

Age in Years	Frequency	<CF	>CF
60-68	1	36	1
51-59	2	35	3
42-50	3	33	6
33-41	2	30	8
24-32	20	28	28
15-23	5	8	33
6-14	3	3	36

Graphs

8. Cumulative Frequency Ogive – commonly used in statistical reports and text.

Less Than Cumulative Frequency " $<CF$ " – starting with the interval with Lowest Score to Highest Score

Greater Than Cumulative Frequency " $>CF$ " – starting with the interval with Highest Score to Lowest Score

