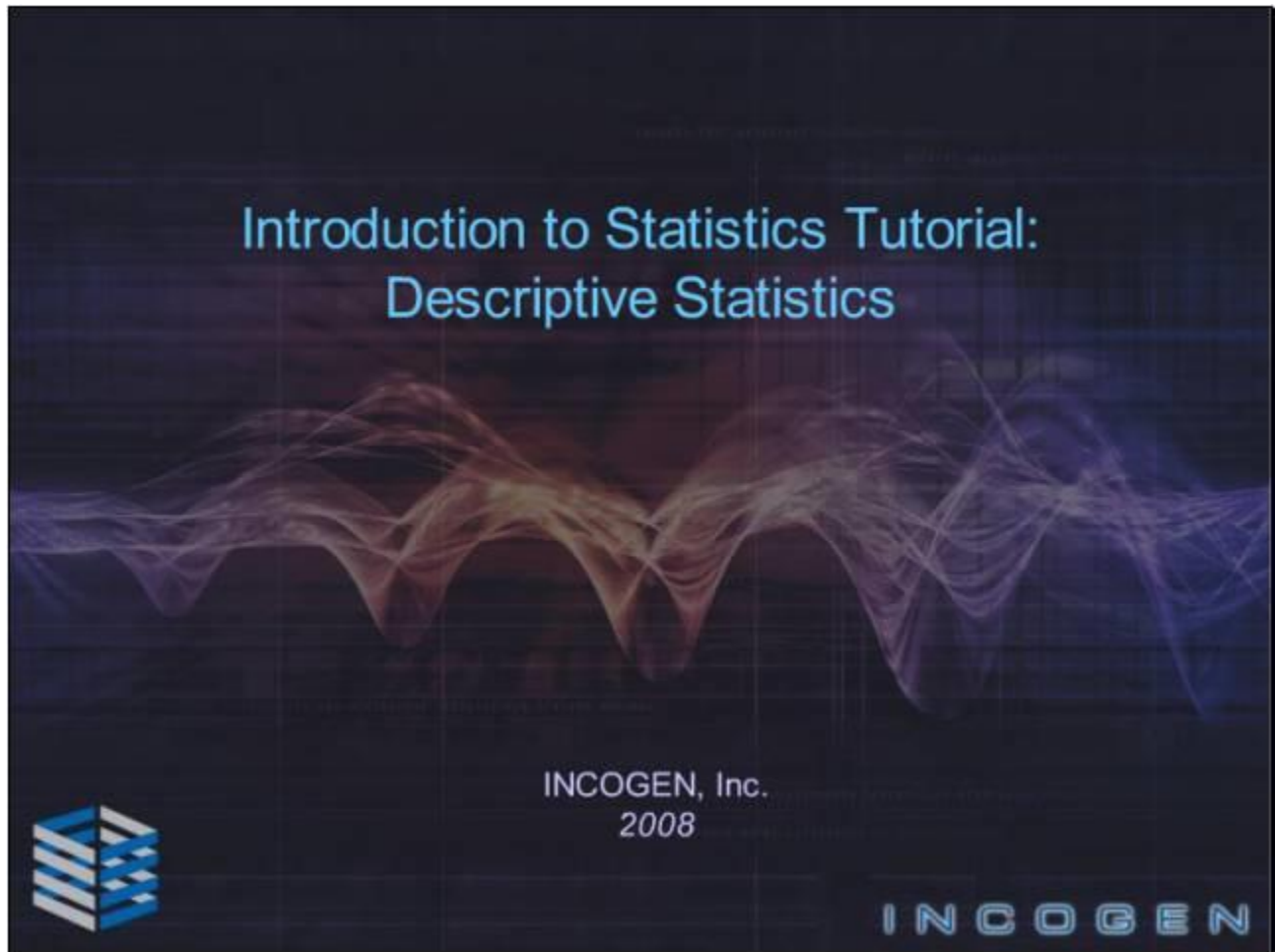


Slide 1 - Introduction to Statistics Tutorial: Descriptive Statistics



Slide notes

Introduction to Statistics Tutorial: Descriptive Statistics. This tutorial is part of a series of several tutorials that introduce probability and statistics. Here we will concentrate on descriptive statistics.

Outline

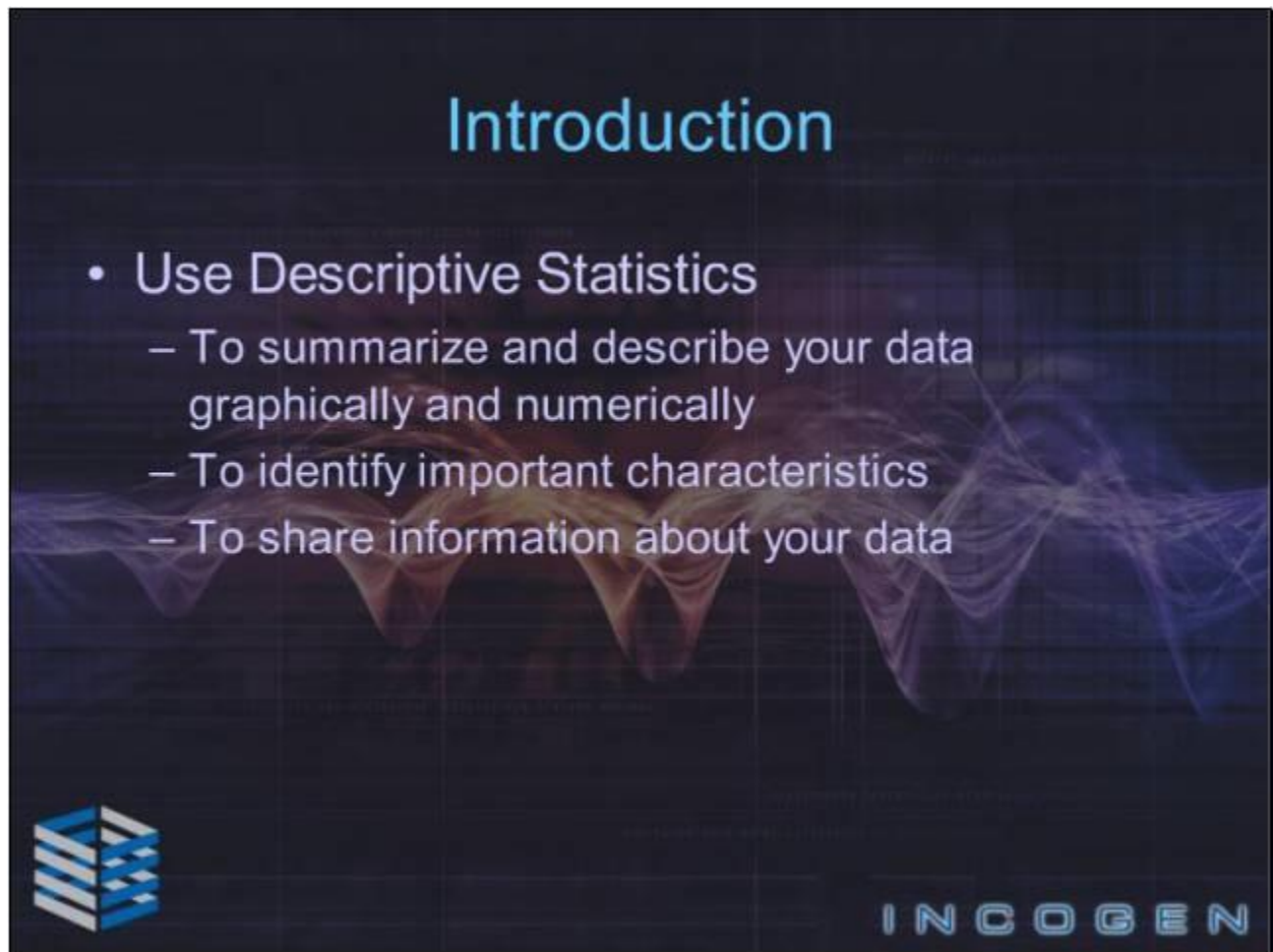
- Graphical Techniques
 - Bar graphs
 - Line graphs
 - Scatterplots
 - Relative frequency histograms
 - Pie charts
- Numerical Techniques
 - Measures for central tendency
 - Variability
 - Relative standing
 - Boxplots (graphical and numerical)

INCOGEN

Slide notes



We will present graphical and numerical techniques to summarize and describe the important characteristics of a set of data. Bar graphs, pie charts, line graphs, scatterplots, relative frequency histograms, and box plots are some of the most common graphical techniques that will be introduced. How to interpret these graphs along with the possible interpretation flaws will be stressed. Along with the graphical techniques, numerical descriptions for central tendency (mean, median and mode), variability (range, variance, and standard deviation) and relative standing (z-score, percentile, and quartiles) will be explained. We will conclude with discussion of box plots which are graphical and numerical.

Slide 3 - Introduction



Introduction

- Use Descriptive Statistics
 - To summarize and describe your data graphically and numerically
 - To identify important characteristics
 - To share information about your data



Slide notes

When statisticians have lots of data (a bunch of numbers), they need a way to look at in an organized way that will be useful. They use Descriptive Statistics:

To summarize and describe data graphically and numerically

To identify important characteristics

To share information about the data

They want to see if there are any patterns or basic shapes that the data follow. They want to summarize the data to be able to say something about them, other than "Here is a list of numbers" which is how we will begin.

Example 1

- **Bioinformatics Trend:**
Since 1997, the number of universities that offer bioinformatics training programs has drastically increased. The numeric data is to the right.

1997	2
1998	6
1999	11
2000	20
2001	30
2002	41
2003	49
2004	78

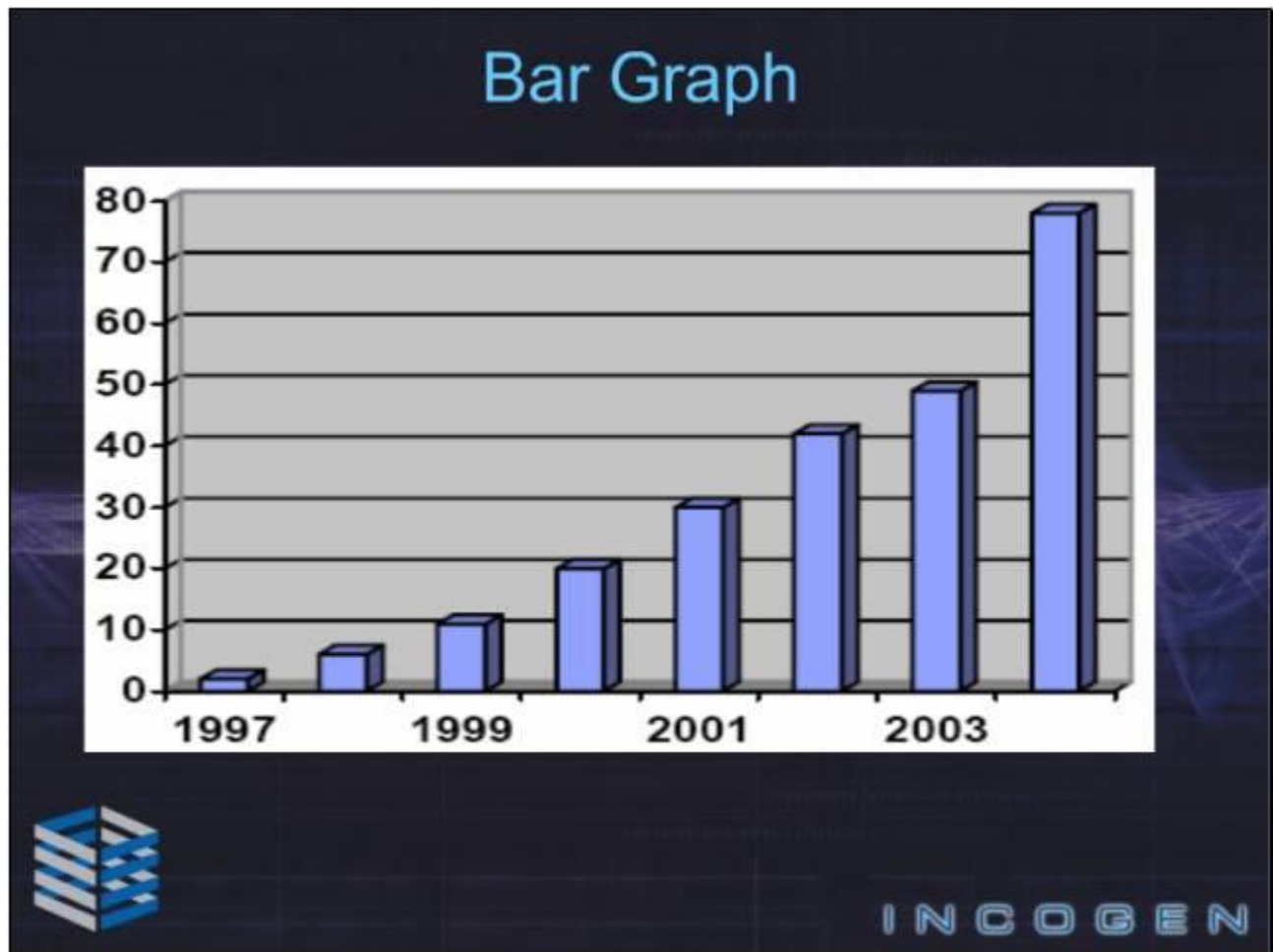


INCOGEN

Slide notes

Our first example shows a small data set. For each year from 1997 to 2004, the data lists the number of universities offering bioinformatics programs. Since 1997, the number of universities that offer bioinformatics training programs has drastically increased. The numeric data is shown in the table on the right.

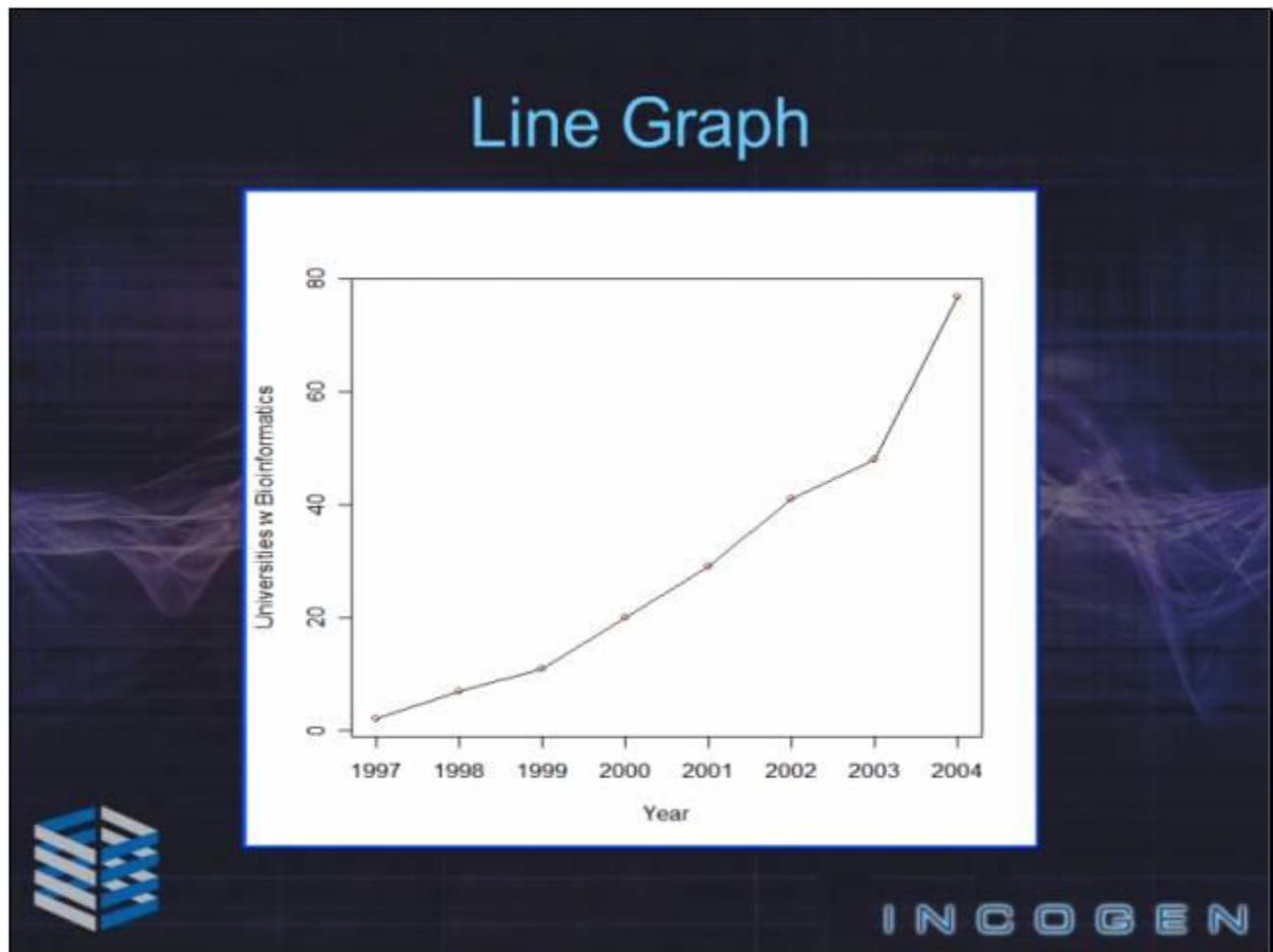
Slide 5 - Bar Graph



Slide notes

Here is a bar graph of the same data. A bar graph can be used to show amounts or frequencies in categories. In our example, the year is on the x-axis and the number of universities offering bioinformatics programs is on the y-axis. For each year a bar rises to the number of schools offering the program. Instead of just viewing the numbers, we can instantly observe that there is an increasing trend in the number of schools that offer bioinformatics programs.

Slide 6 - Line Graph





Slide notes

Here is a line graph of the same data. A line graph is a good way to display the change in your data over time. The time intervals or categories are on one axis and the data to be plotted are on the other. In our example, the year is on the x-axis and the number of universities offering bioinformatics programs is on the y-axis. For each year a point is plotted at the corresponding height. Then the points are connected with lines. Again, we can instantly observe that there is an increasing trend in the number of schools that offer bioinformatics programs.

Example 2

- Classroom Data: The test scores of 25 biology students were recorded. Out of 150 points the students' scores were as follows:

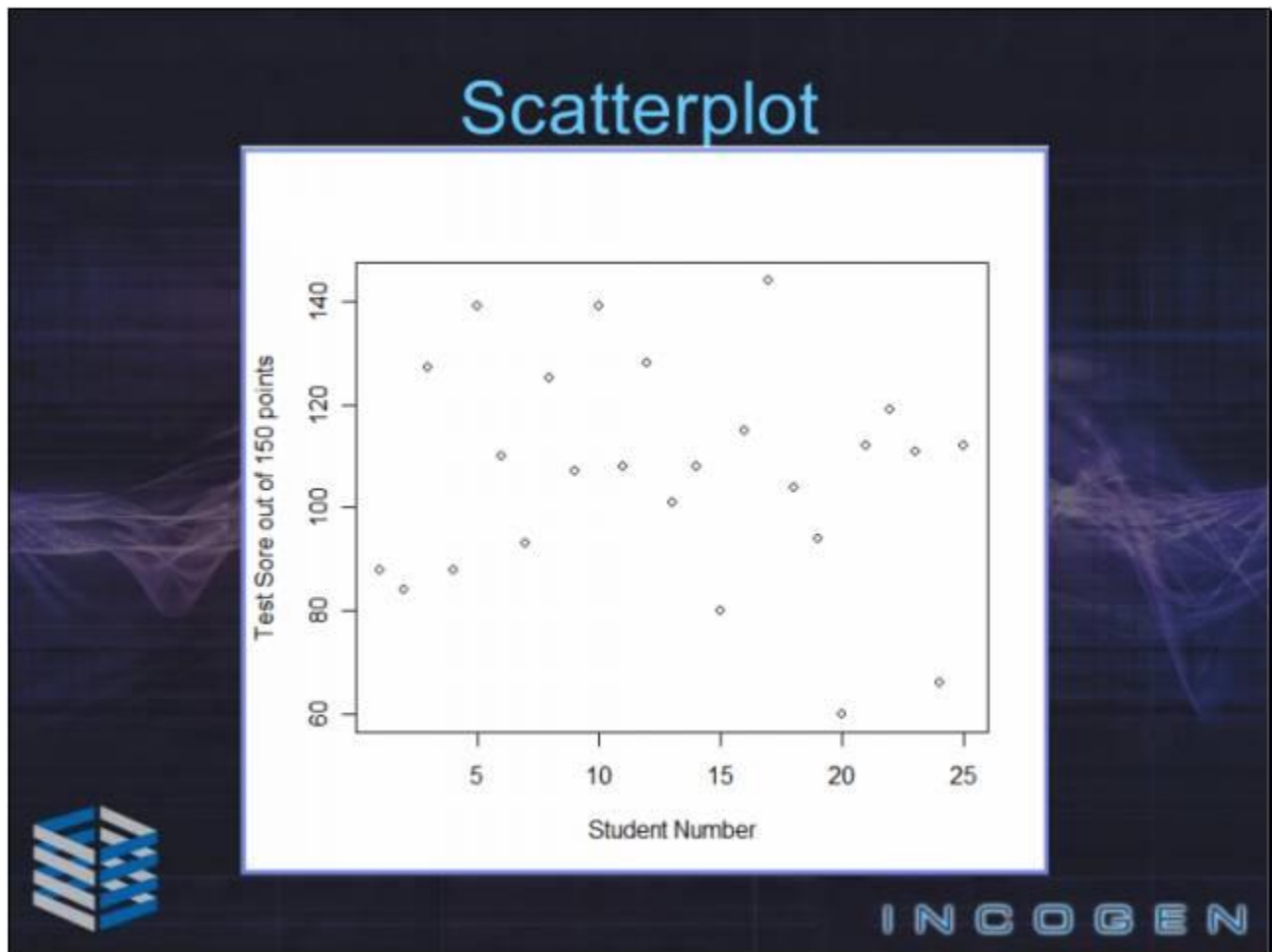
88 84 127 88 139 110 93 125 107 139 108 128 101
108 80 115 144 104 94 60 112 119 111 66 112



Slide notes

Our second example lists the test scores of 25 biology students. Out of 150 points the students' scores were recorded. We will look at the data in a scatterplot, a relative frequency histogram, and a pie chart. The numeric data is shown.

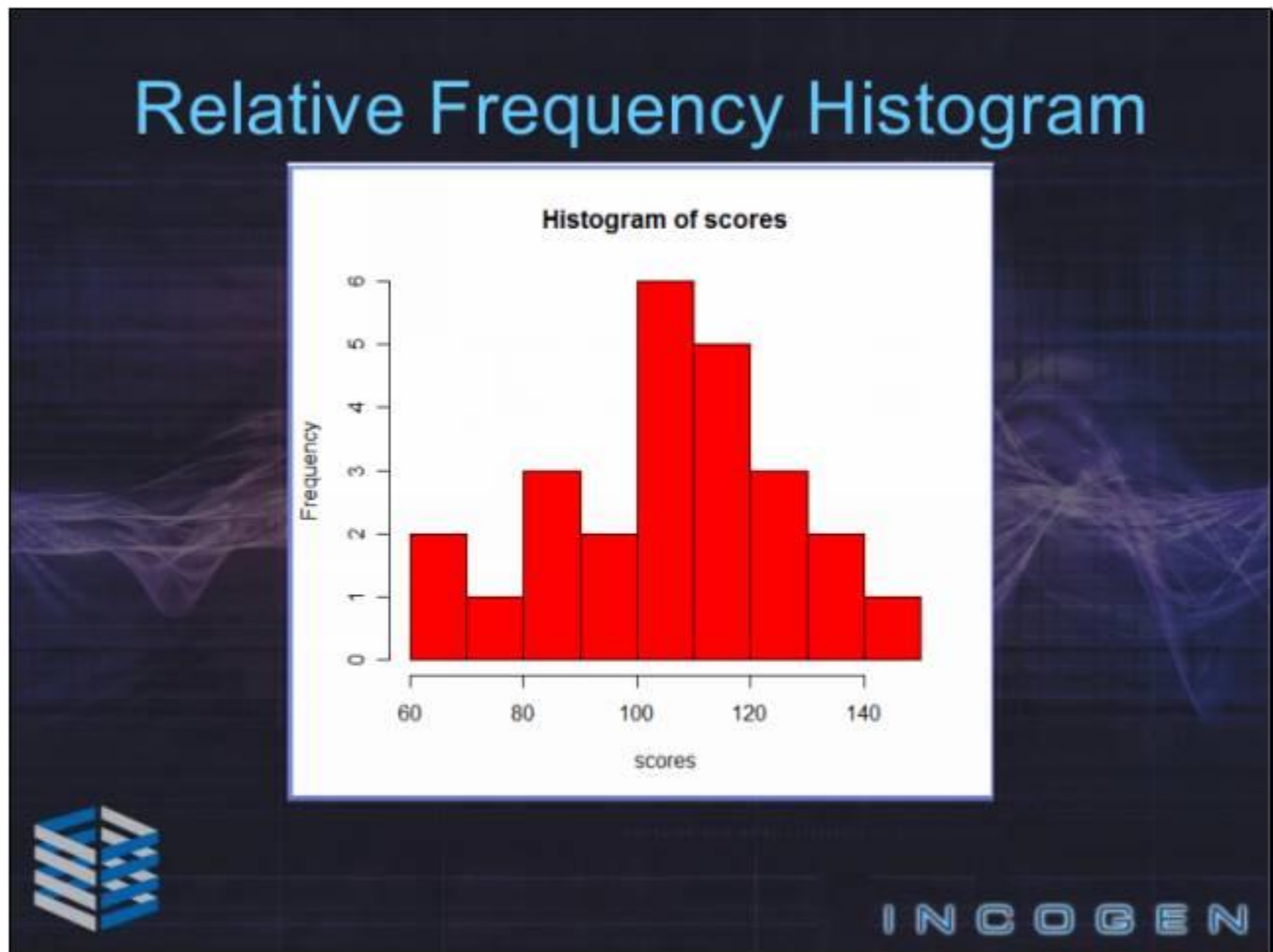
Slide 8 - Scatterplot



Slide notes

Here is a scatterplot. It is simply a plot of all the test scores. For each student (Numbered 1-25), place a point above the student number at his/her test score. This plot shows that the scores are distributed all over and that there isn't a lump of bad or good scores, but other than that this plot is uninformative. So let's look at something better.

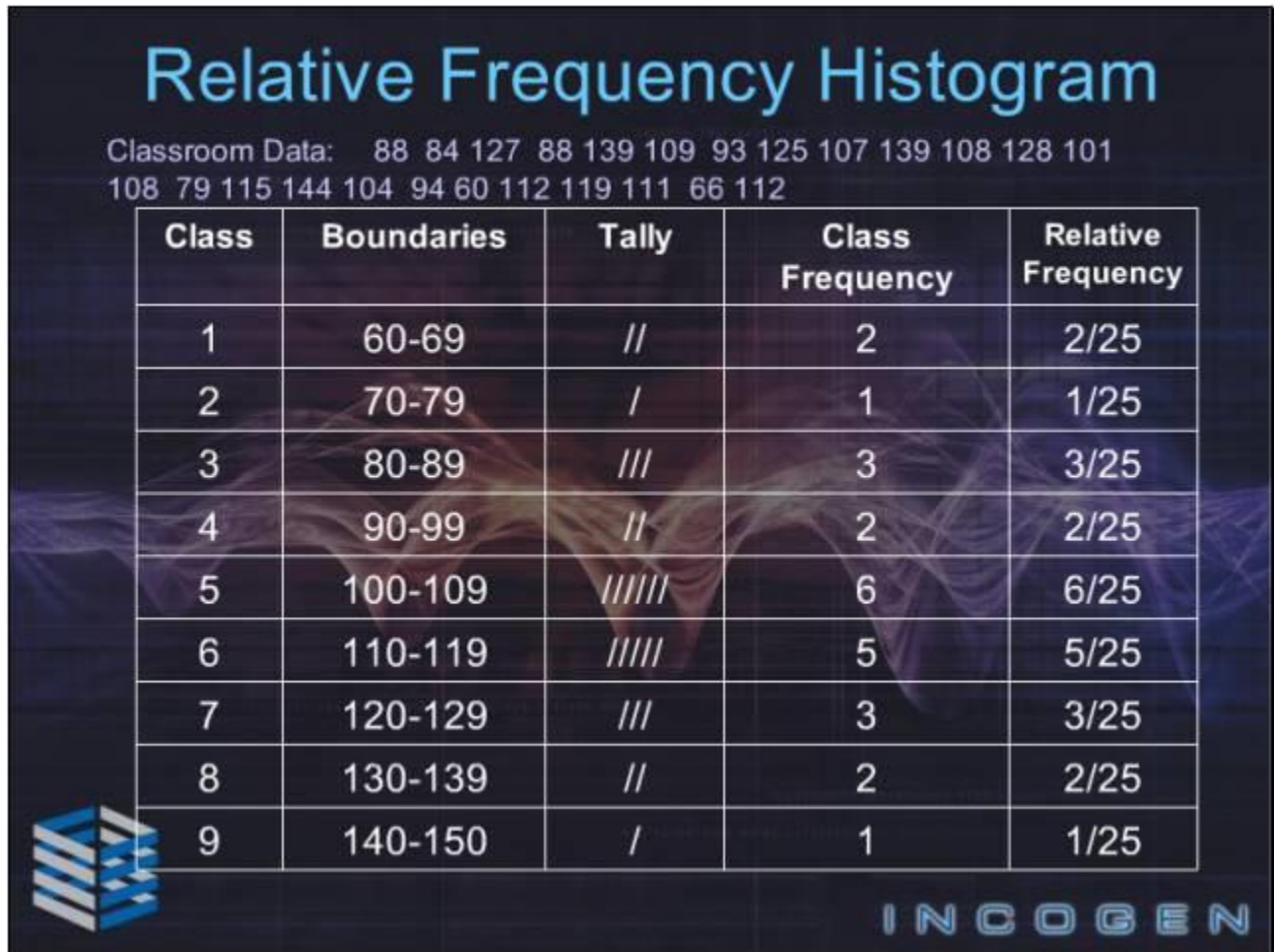
Slide 9 - Relative Frequency Histogram - 2



Slide notes

This is a relative frequency histogram. It shows you the shape of the distribution which is very important. When administering an exam, we hope to see this bell shaped diagram or curve. This is the "curve" that people are referring when they say, he grades on a "curve". Most of the scores are in the middle, or average and then some are above and some are below. We will now spend some time on understanding this plot and how to make it.

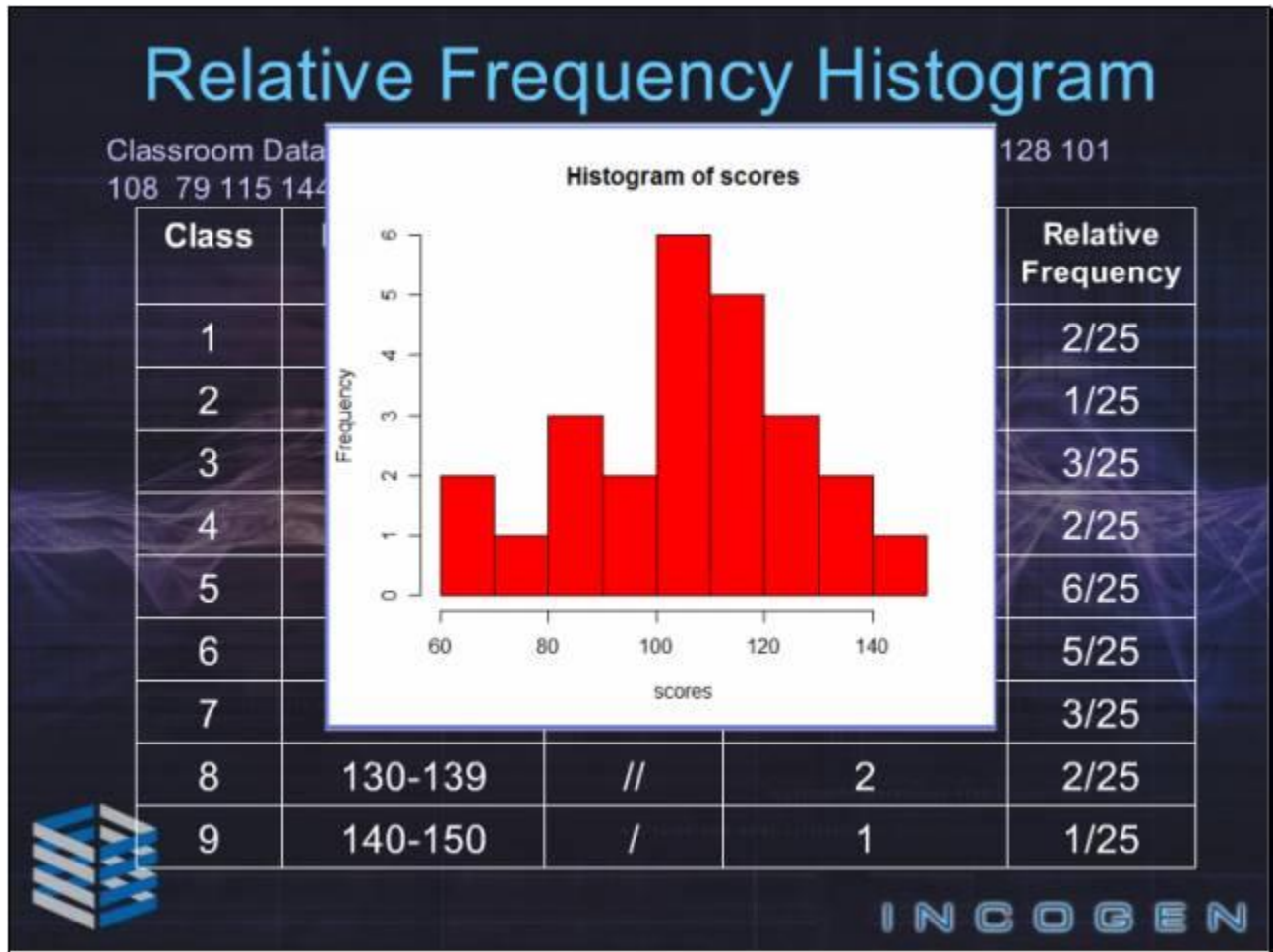
Slide 10 - Relative Frequency Histogram - 3



Slide notes

We created this table to help generate the relative frequency histogram on the last slide. To begin, create the number of desired classes. In this example, we have 9 classes. Then we need to create even bins or boundaries for our classes. In this example our scores can all fit within 60 to 150. So I divided this range into 9 groups of 10 points each; ie., $150-60 = 90 / 9 = 10$. So the first class contains scores from 60 to 69, the second from 70 to 79, and so forth. Then we count (or tally) the number of test scores that fall into each class. For example, there are 3 test scores that fall into class 3: 88, 84, 88. This is called the class frequency. When you divide this number by the total number of test scores, you get the relative frequency.

Slide 11 - Relative Frequency Histogram - 4



Slide notes



When you create the histogram, each bar is as wide as the class boundaries, and extends as high as the relative frequency (or in our case, simply the frequency).

Slide 12 - Relative Frequency Histogram - 5

Relative Frequency Histogram

Classroom Data: 88 84 127 88 139 109 93 125 107 139 108 128 101 108
79 115 144 104 94 60 112 119 111 66 112

Class	Boundaries	Tally	Class Frequency	Relative Frequency
1:F				
2:D				
3:C				
4:B				
5:A				



Slide notes



Let's create a new relative frequency histogram using the same data, but with different classes. This time we will have five classes: A,B,C,D, and F representing the letter grades. We, again, will start by making a table. We start by dividing the range by the 5 groups. Take a moment to create the bins. Hit the pause button below to work on the exercise and the play button when you are ready to resume.

Slide 13 - Relative Frequency Histogram -6

Relative Frequency Histogram

Classroom Data: 88 84 127 88 139 109 93 125 107 139 108 128 101 108
79 115 144 104 94 60 112 119 111 66 112

Class	Boundaries	Tally	Class Frequency	Relative Frequency
1:F	60-77			
2:D	78-95			
3:C	96-113			
4:B	114-131			
5:A	132-150			



Slide notes



The range is $150 - 60 = 90$; $90 / 5 = 18$. So each bin should be 18 points long; that is 60-77 should be the first bin, 78-95 the second and so on. Continue on your own to tally up the number of scores in each bin. Hit the pause button below to work on the exercise and the play button when you are ready to continue.

Slide 14 - Relative Frequency Histogram -7

Relative Frequency Histogram

Classroom Data: 88 84 127 88 139 109 93 125 107 139 108 128 101 108
79 115 144 104 94 60 112 119 111 66 112

Class	Boundaries	Tally	Class Frequency	Relative Frequency
1:F	60-77	//		
2:D	78-95	////		
3:C	96-113	////////		
4:B	114-131	////		
5:A	132-150	///		



Slide notes



Add up the tallied scores for the class frequency and divide by 25 (the total number of data points) for the relative frequency. Hit the pause button below to work on the exercise and the play button when you are ready to continue.

Slide 15 - Relative Frequency Histogram - 8

Relative Frequency Histogram

Classroom Data: 88 84 127 88 139 109 93 125 107 139 108 128 101 108
79 115 144 104 94 60 112 119 111 66 112

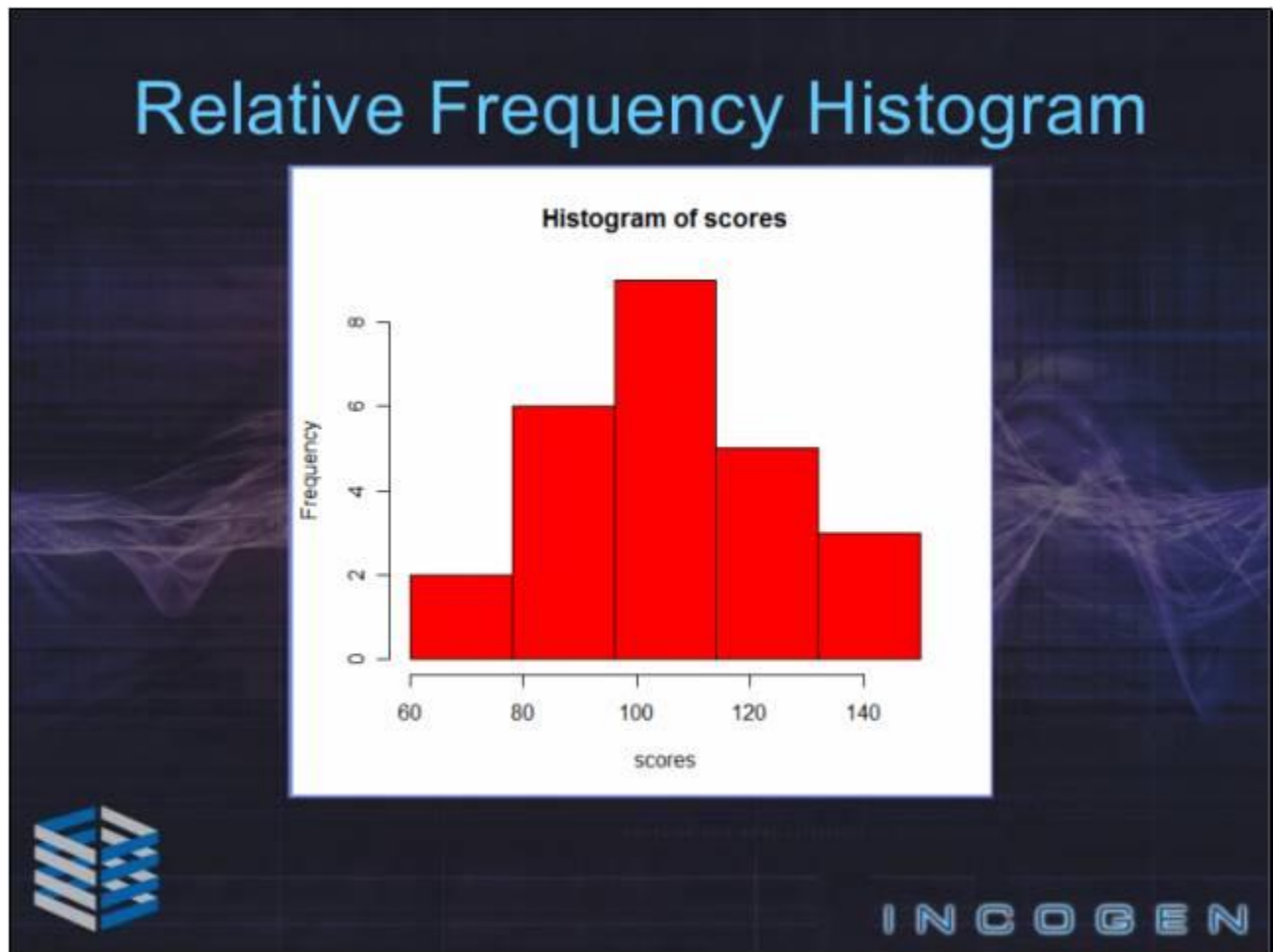
Class	Boundaries	Tally	Class Frequency	Relative Frequency
1:F	60-77	//	2	2/25
2:D	78-95	////	6	6/25
3:C	96-113	////////	9	9/25
4:B	114-131	////	5	5/25
5:A	132-150	///	3	3/25



Slide notes

With the frequencies and the bins, we are ready to make the histogram. Pause the tutorial to make sure you completed your table correctly and try to create your own frequency histogram.

Slide 16 - Relative Frequency Histogram -9



Slide notes

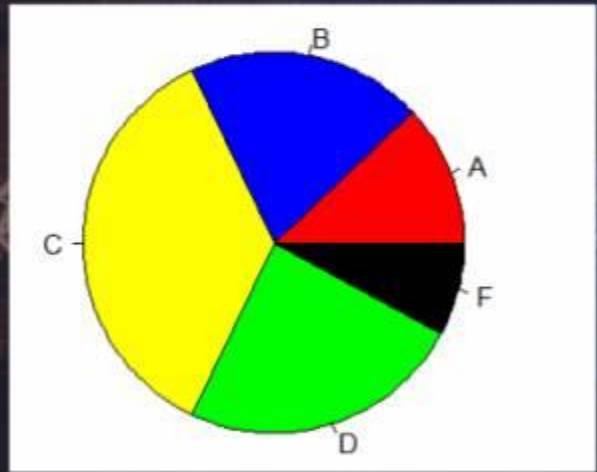
We have 2 test scores in the 60-77 bin so the first bar goes from 60 to 77 on the horizontal axis and it goes up 2 (or 2/25 for a relative frequency histogram) on the vertical axis. Each bar is created similarly.

Slide 17 - Pie Chart

Pie Chart

Classroom Data: 88 84 127 88 139 110 93 125 107 139 108 128 101 108
80 115 144 104 94 60 112 119 111 66 112

Class	Relative frequency	Percent
1:F	$2/25$	0.12
2:D	$6/25$	0.20
3:C	$9/25$	0.36
4:B	$5/25$	0.24
5:A	$3/25$	0.08





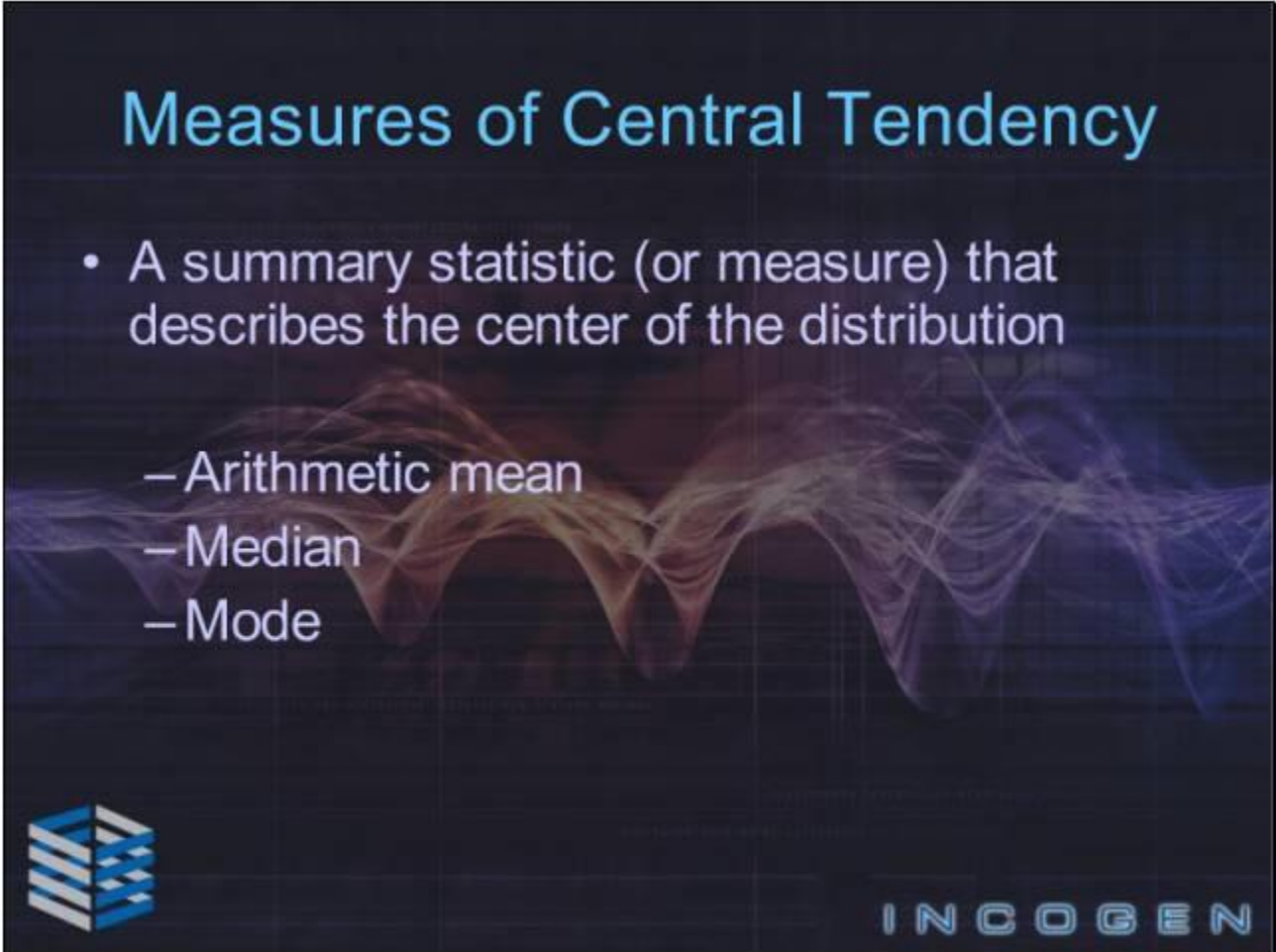
INCOGEN

Slide notes

From the table we created for our histogram, we can also make a pie chart (or pie graph) of the same data. A pie chart is a circular graph that shows how a total quantity can be distributed among categories. In our example, the categories are the letter grades: A, B, C, D, and F. I added a new column to our table (Percent) to show the percent of the circle shaded for each class. The area of each piece of the pie corresponds to the percent of test scores in each class or category, in this case for each letter grade.

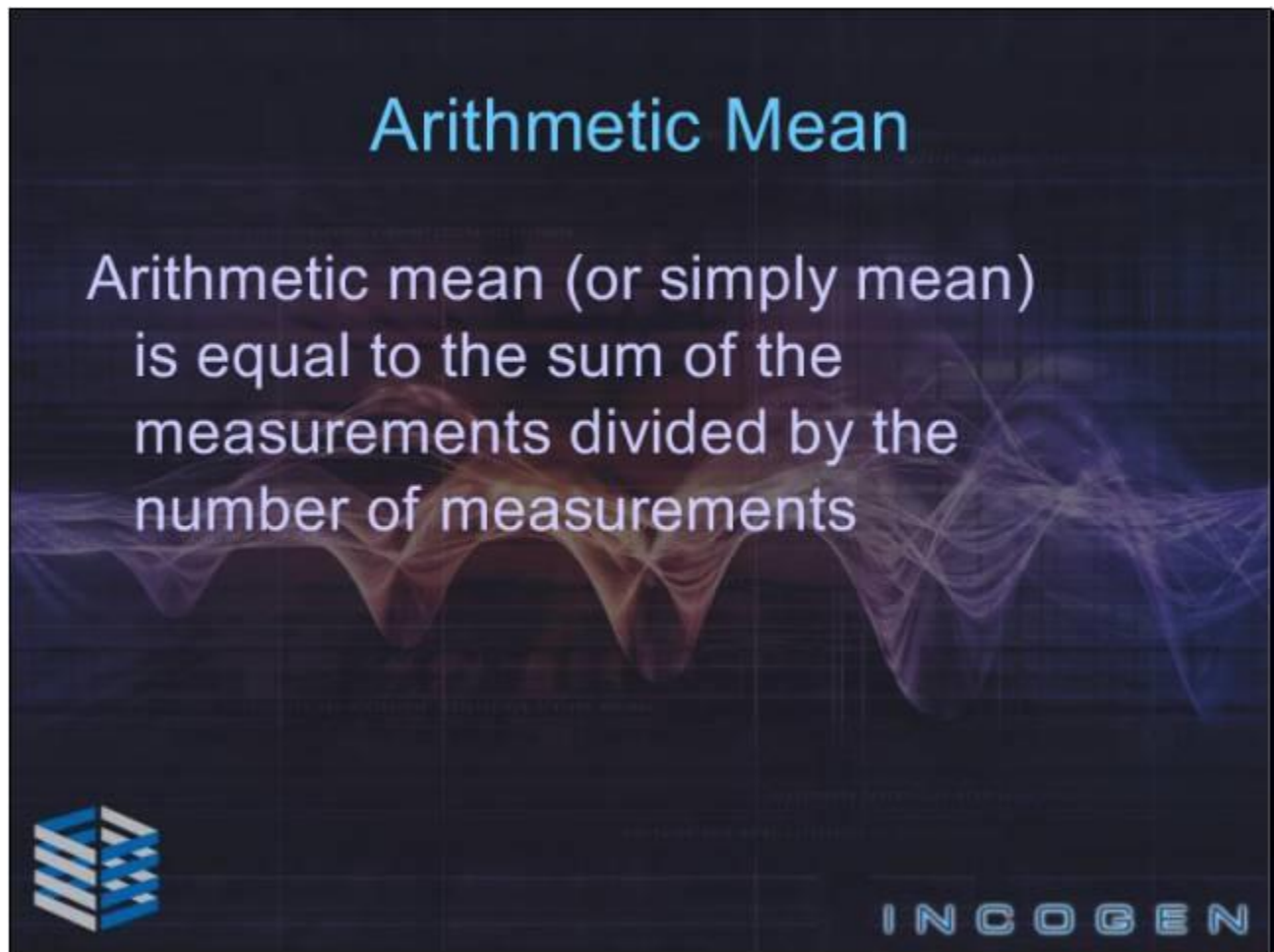
Measures of Central Tendency

- A summary statistic (or measure) that describes the center of the distribution
 - Arithmetic mean
 - Median
 - Mode




Slide notes

Now, let's look at some descriptive statistics that are numerical, instead of graphical. Measures of Central Tendency are summary statistics that describes the center of the distribution. Three examples of central tendency are mean, median, and mode.

The slide features a dark blue background with a faint grid and abstract, colorful wave-like patterns in shades of purple, blue, and orange. The title 'Arithmetic Mean' is centered at the top in a light blue, sans-serif font. Below it, the definition is written in a white, sans-serif font. In the bottom left corner, there is a logo consisting of a 3D cube-like structure made of blue and white lines. In the bottom right corner, the word 'INCOGEN' is written in a light blue, spaced-out, sans-serif font.

Arithmetic Mean

Arithmetic mean (or simply mean)
is equal to the sum of the
measurements divided by the
number of measurements

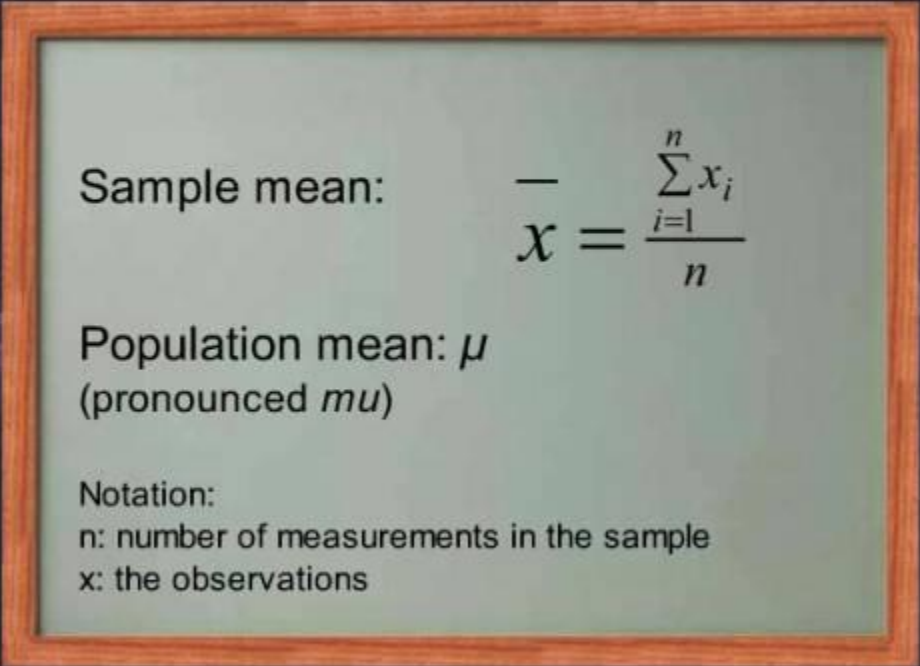


INCOGEN

Slide notes

The arithmetic mean (or simply mean) is equal to the sum of the measurements divided by the number of measurements. The value is commonly called the average, in layman's terms.



Arithmetic Mean Formula



Sample mean: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Population mean: μ
(pronounced *mu*)

Notation:
n: number of measurements in the sample
x: the observations





Slide notes

Mathematically, the formula for the sample mean looks like this. You sum the observations in your sample, then divide by the number of observations in your sample. The sample mean estimates the population mean, which is represented by the Greek letter mu.

Median

The median is the value of the measurements that falls in the middle position when the measurements are ordered from smallest to largest.



Slide notes

The median is the value of the measurements that falls in the middle position when the measurements are ordered from smallest to largest.

Median

Method for calculating the Median:

Order the n measurements from smallest to largest. Then,

1. If n is odd, the median m is the value of the middle measurement, or the one "ranked" $(n+1)/2$
2. If n is even, then the median m is the value halfway between the two middle measurements---it is the average of the two middle values



INCOGEN

Slide notes

To calculate the Median: Order the data (n measurements) from smallest to largest. Then, If n is odd, the median m is the value of the middle measurement...you simply figure out which one is in the middle of the data set. If n is even, then the median m is the value halfway between the two middle measurements---it is the average of the two middle values.

Mode

The mode is the category/data that occurs most often. On a histogram, the mode is the class with the highest frequency. (If there are more than one value in the class, then the midpoint of the class is the mode.)



INCOGEN



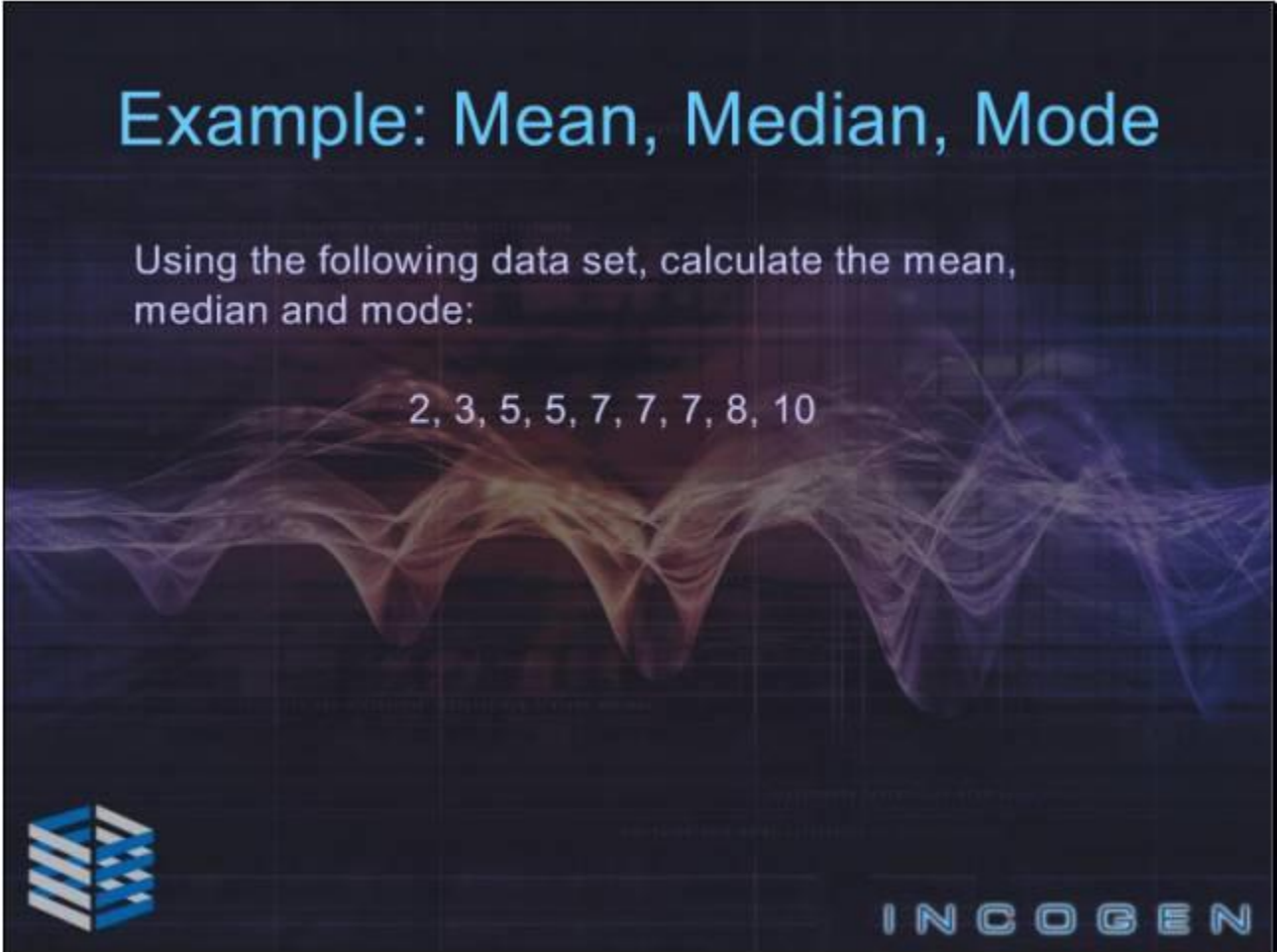
Slide notes

The mode is the category or data that occurs most often. On a histogram, the mode is the class with the highest frequency or bar.

Example: Mean, Median, Mode

Using the following data set, calculate the mean, median and mode:

2, 3, 5, 5, 7, 7, 7, 8, 10



Slide notes

Using the following data set, calculate the mean, median and mode: 2, 3, 5, 5, 7, 7, 7, 8, 10. Note that these data are already ordered for you from smallest to largest. Pause the tutorial until you have completed this exercise.

Example: Mean, Median, Mode

Data set: 2, 3, 5, 5, 7, 7, 7, 8, 10

mode
↑

↑
median

n = 9 observations

$$\text{Mean} = (2+3+5+5+7+7+7+8+10)/9 = 54/9 = 6$$

Median = 7

Mode = 7





INCOGEN

Slide notes

To find the mean, we simply add up all the data (2+3+5+5+7+7+7+8+10), then divide by 9. This gives us 6. The median is 7 because the middle value is 7---see the arrow below the data set. The mode is 7 because there are 3 values of 7 (see the arrow above the data set), only 2 values of 5, and only 1 of the rest of the data. Therefore, the data point that occurs most often is 7.

Measures of Variability


- A summary statistic (or measure) that describes how the data varies or how it is dispersed
 - Range
 - Variance
 - Standard deviation



INCOGEN

Slide notes

Measures of Variability are summary statistics that describes how the data varies or how it is dispersed. Three examples of variability are range, variance, and standard deviation.



Range

The Range of a set of measurements is the difference between the largest and smallest measurements.



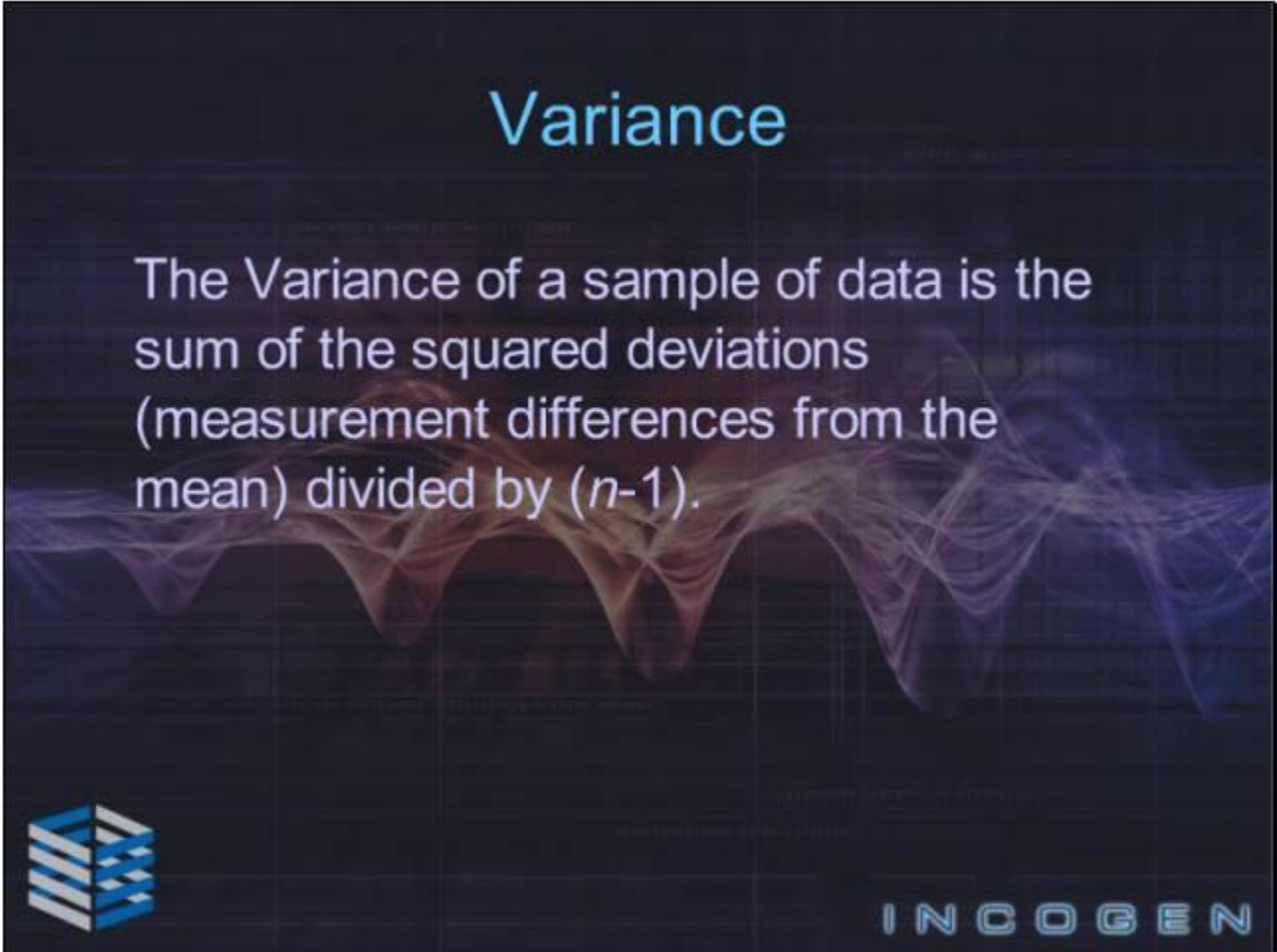


Slide notes

The range of a set of measurements is the difference between the largest and smallest measurements.

Variance

The Variance of a sample of data is the sum of the squared deviations (measurement differences from the mean) divided by $(n-1)$.



Slide notes

The variance of sample data is the sum of the squared deviations (that is, the differences from the measurements to the mean) divided by $(n-1)$.

Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad \text{Sample Variance}$$
$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{Population Variance}$$

INCOGEN

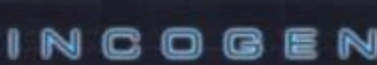

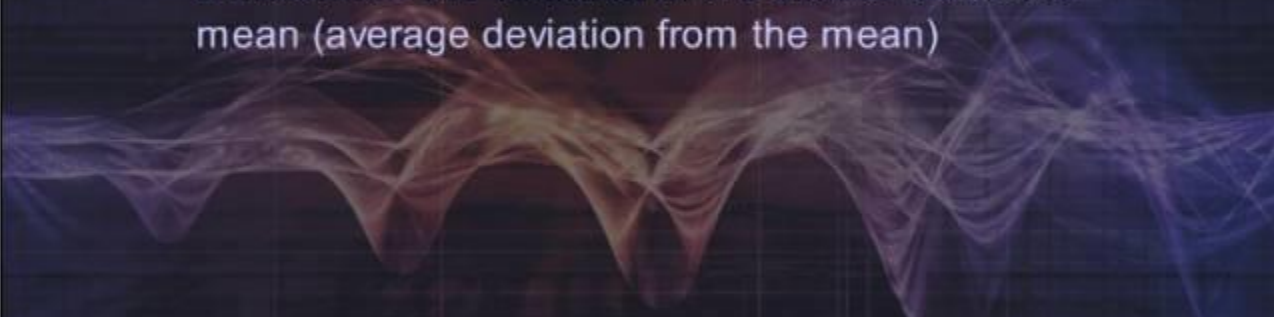
Slide notes

Mathematically, the formula for the sample variance looks like this. You subtract the mean from each observation, then square those values, then sum the squared values. Finally, divide by n-1, where n is the number observations in your sample. The sample variance estimates the population variance, which is represented by the Greek letter sigma-squared.

Standard Deviation

The sample standard deviation is the positive square root of the variance.

It measures the variation of the scores about the mean (average deviation from the mean)



Slide notes

The sample standard deviation is the positive square root of the variance. It measures the variation of the observations about the mean (average deviation from the mean)

Sample Standard Deviation Formula

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

The sample standard deviation is the positive square root of the variance.



INCOGEN

Slide notes

The formula for the sample standard deviation is the positive square root of the variance. Some calculators are able to calculate this for you.

Sample Standard Deviation

$$s = \sqrt{\frac{n (\sum x^2) - (\sum x)^2}{n (n - 1)}}$$

The Shortcut Formula



INCOGEN



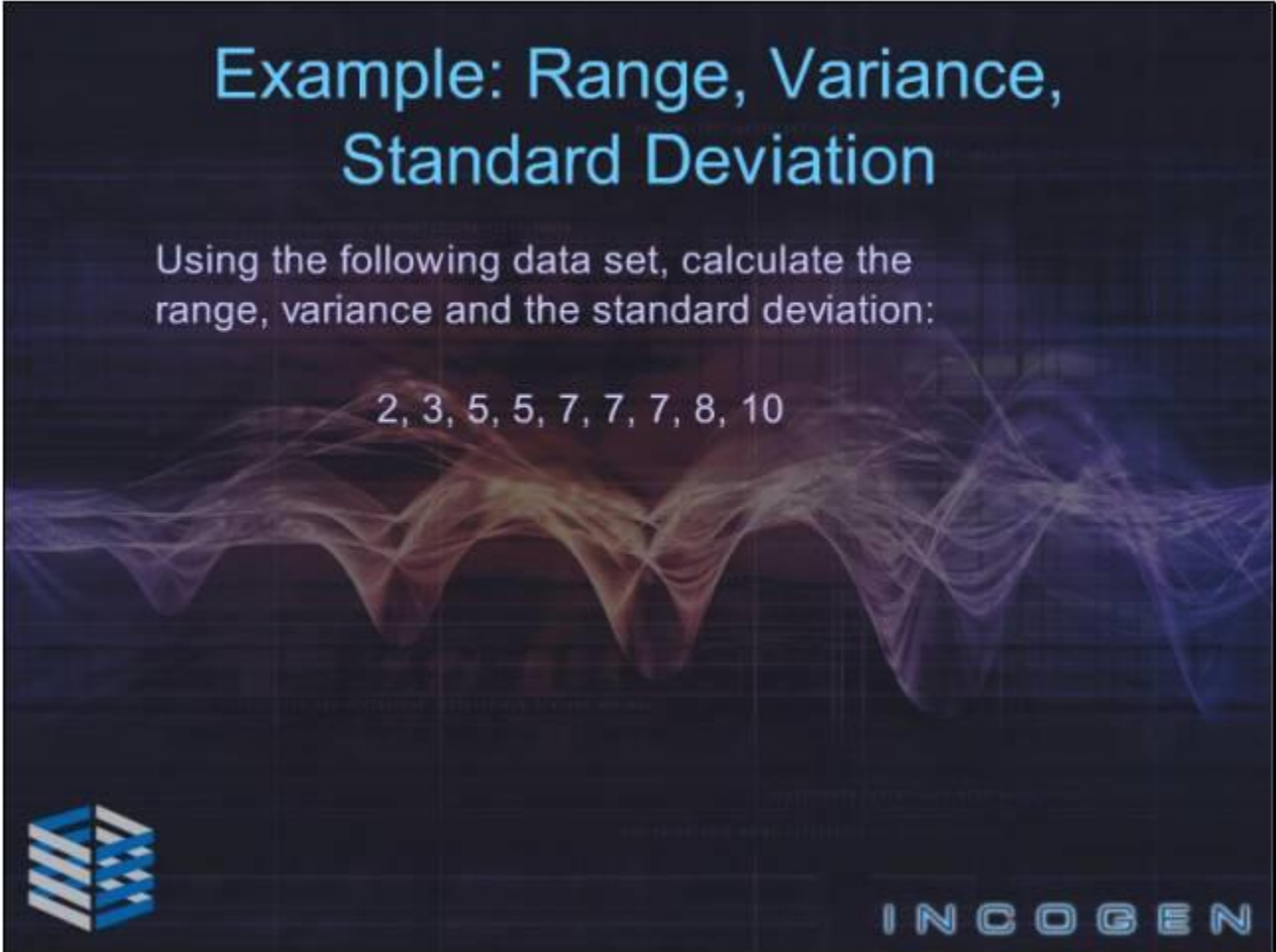
Slide notes

This formula is easier to use if computing the standard deviation 'by hand' as it does not rely on the use of the mean. Only three values are needed: n, sum of x, and sum of x².

Example: Range, Variance, Standard Deviation

Using the following data set, calculate the range, variance and the standard deviation:

2, 3, 5, 5, 7, 7, 7, 8, 10



Slide notes

Using the data set listed, calculate the range, variance and the standard deviation: Take your time by pausing the tutorial. When you are finished, click play to see the answers on the following page and to continue.

Example: Range, Variance, Standard Deviation

Data set: 2, 3, 5, 5, 7, 7, 7, 8, 10

Range = $10 - 2 = 8$

Variance =

$$= \frac{(2-6)^2 + (3-6)^2 + (5-6)^2 + (5-6)^2 + (7-6)^2 + (7-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2}{9-1}$$
$$= \frac{(-4)^2 + (-3)^2 + (-1)^2 + (-1)^2 + (1)^2 + (1)^2 + (1)^2 + (2)^2 + (4)^2}{8}$$
$$= \frac{16 + 9 + 1 + 1 + 1 + 1 + 1 + 4 + 16}{8}$$
$$= \frac{50}{8} = 6.25$$

Standard Deviation = $\sqrt{6.25} = 2.5$

Slide notes

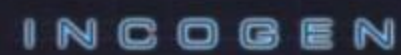

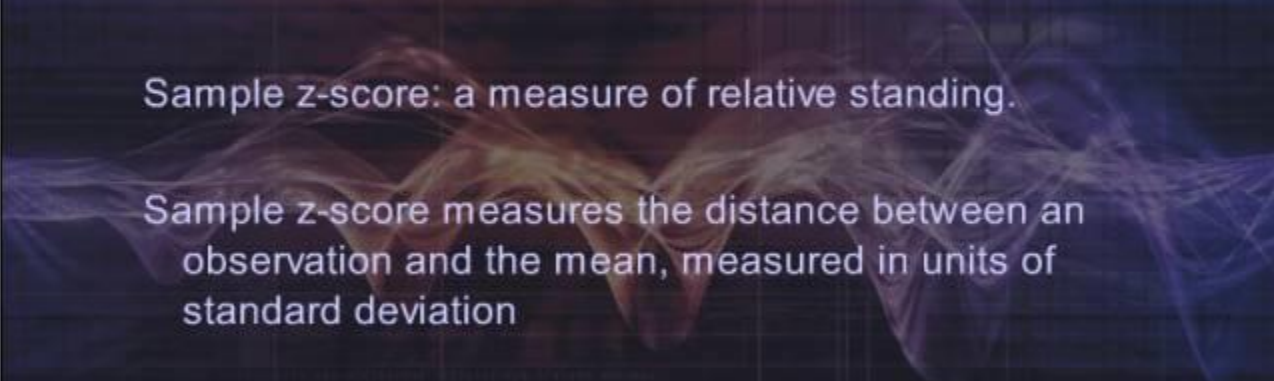
The range is found by subtracting the smallest measurement from the largest. So $10 - 2 = 8$. Using the formula for the variance, we take each data point and subtract off the mean, then square each difference, add them up, and divide by $n-1$, or 8. So the variance is 6.25. Then to find the standard deviation, we simply take the square root of 6.25 to get 2.5.

Relative Standing

Often it is important to know how one data point relates to others, or its Relative Standing

Sample z-score: a measure of relative standing.

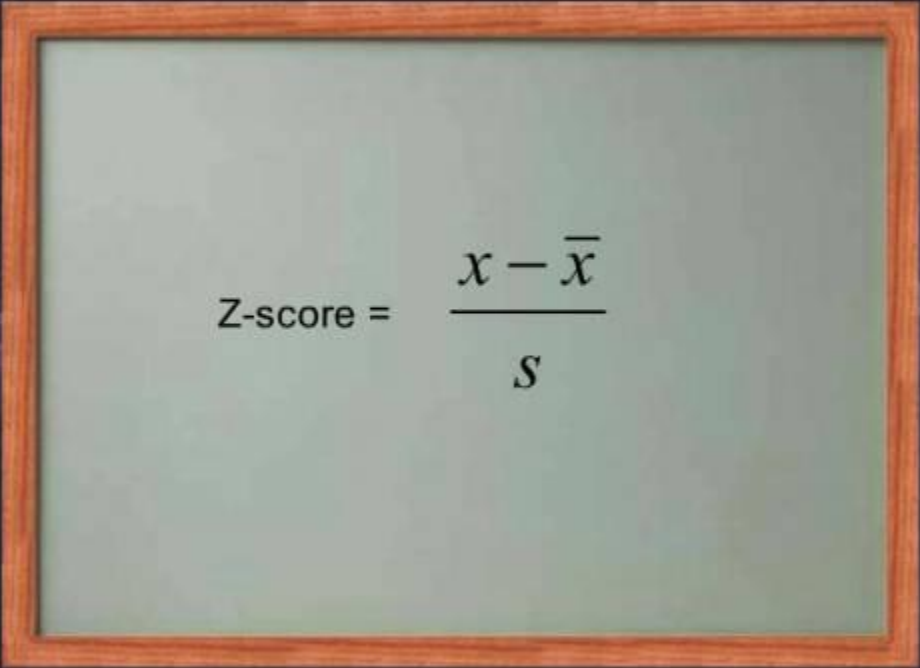


Sample z-score measures the distance between an observation and the mean, measured in units of standard deviation



Slide notes

Often it is important to know how one data point relates to others, or its Relative Standing. The sample z-score is a measure of relative standing. It measures the distance between an observation and the mean, measured in units of standard deviation.

Sample z-score




$$\text{Z-score} = \frac{x - \bar{x}}{s}$$


Slide notes

The formula for the z-score of an observation is to subtract the mean from the observation and divide by the standard deviation.

Relative Standing: Percentiles

Let x_1, x_2, \dots, x_n be a set of n measurements arranged on order of magnitude. The p th percentile is the value of x that exceeds $p\%$ of the measurements and is less than the remaining $(100-p)\%$.



Slide notes



Percentile is another type of relative standing. Let x_1, x_2, \dots, x_n be a set of n measurements arranged on order of magnitude. The p th percentile is the value of x that exceeds $p\%$ of the measurements and is less than the remaining $(100-p)\%$.

Relative Standing: Percentiles/Quartiles

Lower quartile (first quartile) = 25th Percentile

Second quartile (also called the median) = 50th Percentile

Upper quartile (third quartile) = 75th Percentile



Slide notes



We can define some useful percentiles: The lower quartile (or first quartile) = 25th Percentile; the second quartile (also called the median) = 50th Percentile; and the upper quartile (or third quartile) = 75th Percentile. These values will be needed in a very popular and useful plot, the box plot.

Box plots

Exploratory Data Analysis (EDA)

Exploratory data analysis is the process of using statistical tools (such as graphs, measures of central tendency and variation) to investigate the data sets in order to understand their important characteristics – good for seeing outlying observations.

Outliers - extreme values located very far away from most of the other values. They can have a dramatic effect on the mean, standard deviation, and on the scale of the histogram so that the true nature of the distribution is totally obscured.

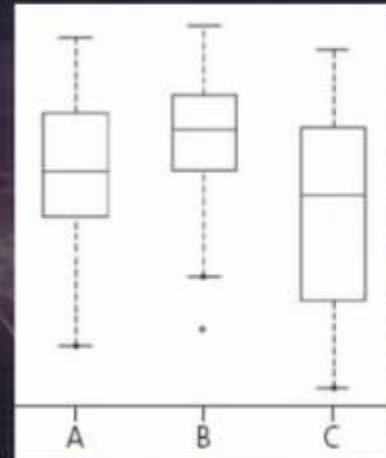


Slide notes

Box plots are for Exploratory Data Analysis (or EDA). Exploratory data analysis is the process of using statistical tools (such as graphs, measures of central tendency and variation) to investigate the data sets in order to understand their important characteristics – they are good for seeing outlying observations, or outliers. Outliers are extreme values located very far away from most of the other values. They can have a dramatic effect on the mean, standard deviation, and on the scale of the histogram so that the true nature of the distribution is totally obscured.

Box plots

Reveal the:
center of the data
spread of the data
distribution of the data
presence of outliers
Excellent for comparing two or
more data sets



INCOGEN

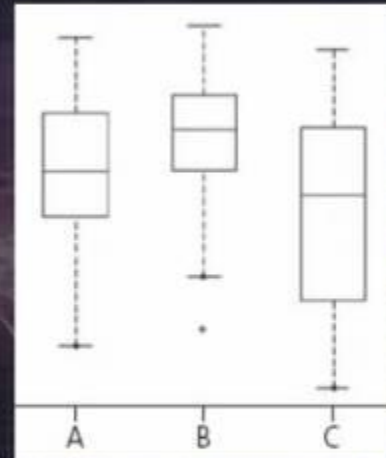
Slide notes

Box plots reveal the center of the data, the spread of the data, the distribution of the data, and the presence of outliers. They are excellent for comparing two or more data sets as shown here in the figure of 3 box plots displaying datasets A, B and C.

Box plots

To construct a Box Plot – need 5 summary statistics:

- Minimum
- First quartile Q1
- Median (Q2)
- Third quartile Q3
- Maximum



INCOGEN

Slide notes

To construct a Box Plot, we need 5 summary statistics: the Minimum value, the First quartile, Q1, the Median (Q2), the Third quartile Q3, and the Maximum value.

Box plots: Construction

To construct a Box Plot:

- Calculate first quartile $Q1$, median M , and third quartile $Q3$.
- Calculate the InterQuartile Range (IQR) by subtracting the first quartile from the third quartile ($IQR = Q3 - Q1$)
- Construct a box above the number line bounded on the left by the first quartile $Q1$ and on the right by the third quartile $Q3$
- Indicate where the median lies inside of the box with a line dividing the box at the median value.
- The mean value of the data can also be labeled with a point



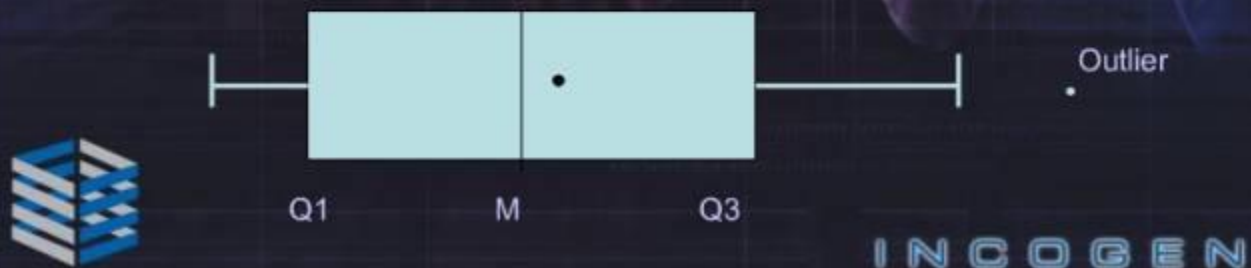
INCOGEN

Slide notes

After we find the first quartile $Q1$, median M , and third quartile $Q3$, we calculate the InterQuartile Range (IQR) by subtracting the first quartile from the third quartile ($IQR = Q3 - Q1$). Start by constructing a box above the number line bounded on the left by the first quartile $Q1$ and on the right by the third quartile $Q3$. Then indicate where the median lies inside of the box with a line dividing the box at the median value. The mean value of the data can also be labeled with a point, if desired.

Box plots: Construction

- Any data observation which lies more than 1.5 IQR lower than the first quartile or 1.5 IQR higher than the third quartile is considered an outlier. Indicate where the smallest value that is not an outlier is by connecting it to the box with a horizontal line. Also, mark the position of this value more clearly using a small vertical line. Likewise, connect the largest value that is not an outlier to the box with a horizontal line (and mark it with another small vertical line)
- Indicate outliers by dots



Slide notes

Let's continue the construction of our box plot. Any data observation which lies more than 1.5 IQR lower than the first quartile or 1.5 IQR higher than the third quartile is considered an outlier. Indicate where the smallest value that is not an outlier is by connecting it to the box with a horizontal line. Also, mark the position of this value more clearly using a small vertical line. Likewise, connect the largest value that is not an outlier to the box with a horizontal line (and mark it with another small vertical line). Indicate outliers by dots. As a side note, box plots can be either horizontal (as shown here) or vertical and are created the same way in either case.



References

This tutorial is comprised of materials from the following sources:

Introduction to Probability and Statistics by Mendenhall and Beaver. ITP/Duxbury.

The Cartoon Guide to Statistics by Gonick and Smith. HarperCollins.

Elementary Statistics by Triola. Addison-Wesley-Longman



Slide notes

This concludes our tutorial on descriptive statistics. This material and more details can be found in the listed resources.