HISTOGRAMS: THE BASICS
TEACHER NOTES

DESCRIPTION
Scientists use histograms to analyze data. In particular, particle physicists rely on histograms to find new particles and to measure particle characteristics. Sometimes the probability of a particular interaction occurring is small. In such cases, particle physicists collect large amounts of data in the hope of finding this interaction as a small bump in the histogram. This activity builds the histogram skills required in many of the other activities in the Data Activities Portfolio. Students will construct histograms and use them to make claims about the data.

STANDARDS ADDRESSED
Next Generation Science Standards

Science and Engineering Practices

4. Analyzing and interpreting data
5. Using mathematics and computational thinking

Crosscutting Concepts

3. Scale, proportion, and quantity

Common Core Literacy Standards

Reading

9-12.4 Determine the meaning of symbols, key terms . . .
9-12.7 Translate quantitative or technical information . . .

Common Core Mathematics Standards

MP2. Reason abstractly and quantitatively.

AP Physic 1: Algebra-Based and AP Physics 2: Algebra-Based Science Practices

Science Practice 4

The student can plan and implement data collection strategies in relation to a particular scientific question.

Science Practice 5

The student can perform data analysis and evaluation of evidence.

ENDURING UNDERSTANDINGS
Scientists can analyze data more effectively when they are properly organized; charts and histograms provide methods of finding patterns in large data sets.

LEARNING OBJECTIVES
Students will know and be able to:

- Generate a histogram given a list of events.
- Explain the meaning of the vertical axis on a histogram.
- Select appropriate bin size to best display the data.
- Identify how many discrete peaks appear in the histogram.

PRIOR KNOWLEDGE
Students should be able to:

- Collect and organize data.
- Explain the difference between representations of central value: mean, median and mode.

RESOURCES/MATERIALS
- Ruler

- Graph paper or graphing software
- Calculator
- The importance of histograms for CERN data analysis: https://home.cern/news/news/computing/big-data-takes-root
- Drawing a histogram by hand: https://www.youtube.com/watch?v=EqlHVMTaPiA
- Note that the presenter uses the term *classes* while particle physicists use the term ***bins***.
- The wiki page for histograms: https://en.m.wikipedia.org/wiki/Histogram

BACKGROUND MATERIAL

This activity is an introduction to making and interpreting histograms.

IMPLEMENTATION

Histogram Basics:

A histogram is a graph of how frequently a value lies in a particular range called a bin. Consider Data Set 1: {1, 2, 2, 3, 3, 3, 3, 4, 4, 5, 6}. Notice that there is a single value of 1, so the bin for one has a height of 1. There are two values of 2, so the bin for two has a height of 2, etc.

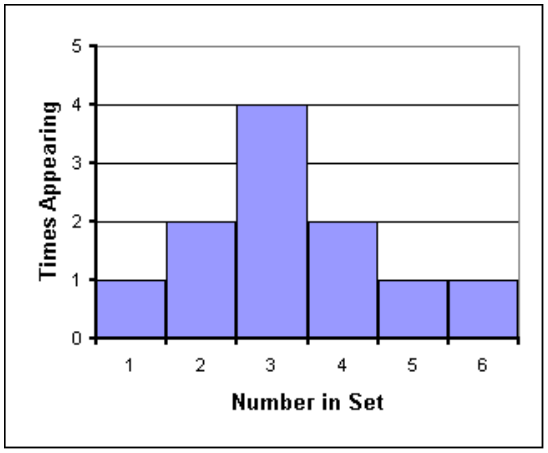


Figure 1: Histogram of Data Set 1.

This histogram in Figure 1 gives us useful data about the set. For example, the graph peaks at 3, which is also the median and the mode of the set. The mean of the set is 3.27—also not far from the peak. The shape of the graph gives us an idea of how the numbers in the set are distributed about the mean; the distribution of this graph is wide compared to size of the peak, indicating that values in the set are only loosely bunched round the mean.

The example above is a little too simple. In most real data sets, almost all numbers will be unique. Consider Data Set 2: {3, 11, 12, 19, 22, 23, 24, 25, 27, 29, 35, 36, 37, 45, 49}. A graph which shows how many ones, how many twos, how many threes, etc. would be meaningless. Instead, we *bin* the data into convenient ranges. In this case, we can easily group the data as below with a bin width of 10. **Note:** When using bins, notice that the right value is repeated as the left value in the line below. Common practice is to place the value in the bin that has the value on the left. Therefore, a value of 20 is plotted in the third bin which has 20 on the left.

Data Set 2

Data Range	Frequency
0–10	1
10–20	3
20–30	6
30–40	4
40–50	2

Note that the median is 25 and that there is no mode; the mean is 26.5.

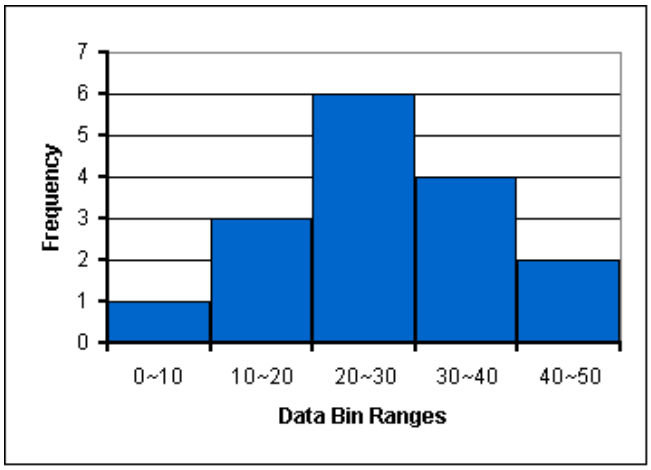


Figure 2: Histogram of Data Set 2.

Effect of Bin Size:

Consider the case of a teacher grading tests. Miss Chang's physics class has just taken a test. In order to come up with meaningful grades, Miss Chang will make a histogram to represent the distribution of grades and find a reasonable central value.

Data Set 3

Student	Grade
Bullwinkle	84
Rocky	91
Bugs	75
Daffy	68
Wylie	98
Mickey	78
Minnie	77
Lucy	86
Linus	94
Asterix	64
Obelix	59
Donald	54
Sam	89
Taz	76

Sorted Data Set 3

Student	Grade
Donald	54
Obelix	59
Asterix	64
Daffy	68
Bugs	75
Taz	76
Minnie	77
Mickey	78
Bullwinkle	84
Lucy	86
Sam	89
Rocky	91
Linus	94
Wylie	98

It is important to note that when you make a histogram by hand, the data must be sorted in order to determine how many values will be in each bin.

The table on the far left is the original data. The data on the right has been sorted. When there is an *odd number of values*, the ***median*** is the middle number.

Notice that there is an *even number of data values*. Based on this sort, the ***median*** is the average of the two middle numbers. For these data the median is 77.5.

The next question is that of bin size. Clearly, a bin size of 100 makes no sense, as it puts all the data in one bin, giving us no information. At the same time, a bin size of 1 or less makes no sense, as the bins would be so small as to look pretty much like a simple list of results. We already have that!

Let's try a few bin sizes:

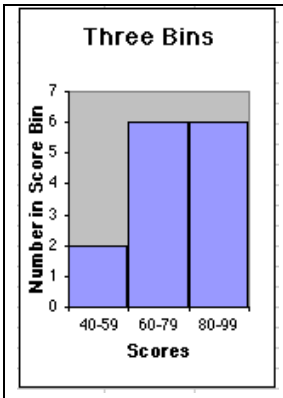


Figure 4: Data Set 3 with three bins.

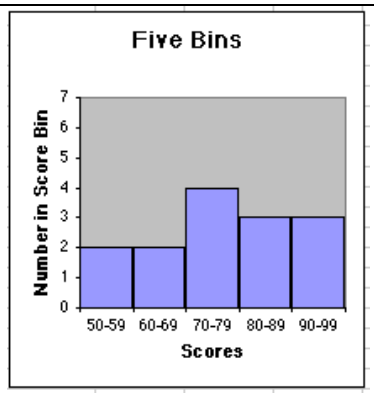


Figure 5: Data Set 3 with five bins.

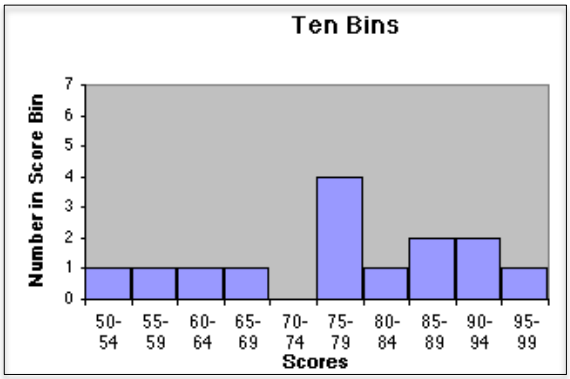


Figure 6: Data Set 3 with ten bins.

When Data Set 3 is graphed with three bins, the bin size is too large. The claim that can be made based on this histogram is that two students failed the test. The assumption that the cutoff for a grade for D is 60.

When Data Set 3 is graphed with five bins, you can get a better picture of what happened. This bin size tells you that the median is between 70–79. The assumption is that A = 90–99, B = 80–89, C = 70–79, and D = 60–69. Another claim is that Miss Chang can tell at a glance the number of students who scored A, B, C, D or F.

When Data Set 3 is graphed with ten bins, more information is available. The median is clearly in the range of 75–79. Four students are in trouble and may need extra assistance in learning the material.

Of course, part of the power of histograms is that they allow us to analyze extremely large datasets by reducing them to a single graph that can show primary, secondary and tertiary peaks in data as well as give a visual representation of the statistical significance of those peaks. To get an idea, look at these histograms for three different data sets: The median is in the range of 35–40; without a data list, the mean cannot be found.

Identifying Peaks:

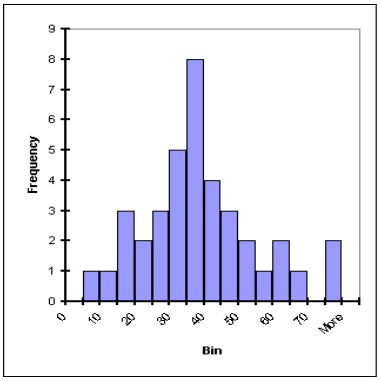


Figure 7: Histogram with a well-defined peak.

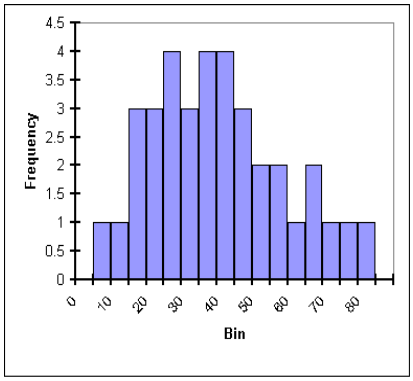


Figure 8: Histogram with less-defined peak.

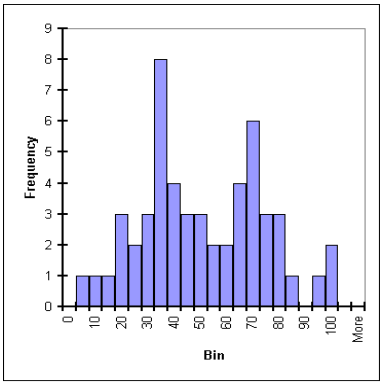


Figure 9: Histogram with two peaks.

Figure 7 represents data with a well-defined peak that is close in value to the median. While there are "outliers," they are of relatively low frequency. The data is also symmetric about the median. Thus, it can be said that deviations in this data group from the mean are small. If this were a mass plot in particle physics, we'd say the mass is understood with good precision.	In Figure 8, the peak is still fairly close to the median, but it is much less defined. It is harder to tell from the plot the exact location of the peak. There are almost as many values close to the peak as at the peak itself, and outliers are frequent. As a particle physics mass plot, this gives an imprecise and uncertain mass of a particle.	Where is the median for Figure 9? It is hard to tell; it also may not be relevant. There are two peaks in this plot: a taller primary peak as well as a shorter secondary peak. This could indicate either very poor definition of one signal in the data or, more likely, two signals. In particle physics, this could show two separate particles or, as is often the case, a large signal with "background" particles and a smaller signal (sometimes very small), called a "bump," which shows the actual particle under study.

ASSESSMENT

For a summative assessment, divide the class into teams and tell each team to master their portion of the activity. Then direct the students to make a presentation using whiteboards or electronic methods to explain their part to the rest of the class.

You can print off the pages with the histograms below as a summative assessment.

Student Instructions:

For each of the histograms below, answer the following questions:

1. How many peaks are present?
2. What aspect of each run results in histograms that look “blocky” and histograms that look smooth?
3. What is the mass, in GeV/c^2, associated with each peak?
4. Describe the range of values for the mass based on the width of the peak.
5. Histograms 3 and 4 may represent the same particle. Which histogram indicates a more precise estimate of the particle mass? Explain your reasoning.
6. How many particles are represented in Histogram 6? Defend your claim.

Answer Key:

1. What quantity is plotted in each graph?
    - *Histogram 1: invariant mass in GeV*
    - *Histogram 2: energy in GeV*
    - *Histogram 3: invariant mass in GeV*
    - *Histogram 4: invariant mass in GeV*
    - *Histogram 5: transverse momentum in GeV*
    - *Histogram 6: invariant mass in GeV*
2. How many peaks are present?
    - *Histogram 1: One peak*

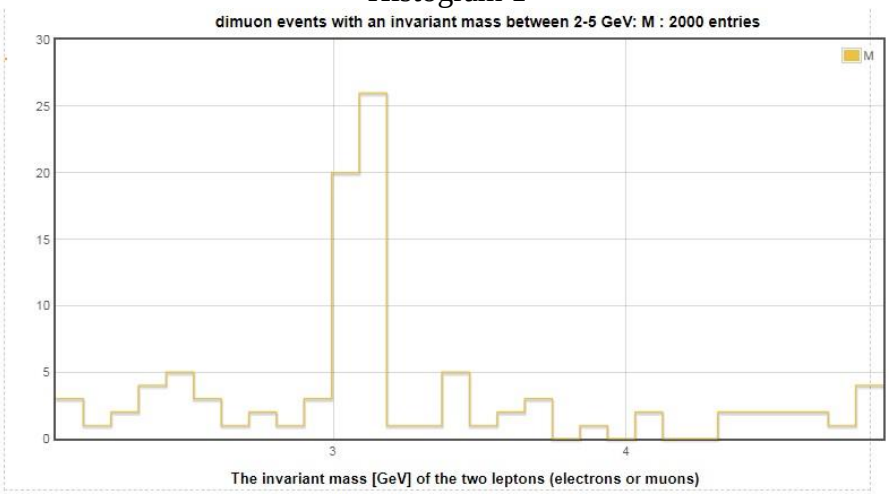
- *Histograms 2 and 5: One clear peak with a possible peak at very low energy. These histograms provide an opportunity for students to explain about low-energy background if the concept was discussed during the activity.*
- *Histograms 3 and 4: One peak and possibly representing the same particle*
- *Histogram 6: Two peaks and one possible peak at very low mass*

3. What aspect of each run results in histograms that look "blocky" and histograms that look smooth?
    - *The "blocky" appearance seems to result for histograms with fewer events. When there are a large number of events, histograms seem to smooth out.*
4. What is the value, in GeV, associated with each peak?
    - *Histogram 1: ~3.1 GeV*
    - *Histograms 2 and 5: ~40 GeV*
    - *Histograms 3 and 4: ~90 GeV*
    - *Histogram 6: ~15 GeV and ~90 GeV*
5. Describe the range of values for the the width of the peak.
    - *Histogram 1: ~3.0 GeV to 3.1 GeV*
    - *Histogram 2: ~40 GeV to ~55 GeV*
    - *Histogram 3: ~85 GeV to ~95 GeV*
    - *Histogram 4: ~90 GeV to ~95 GeV*
    - *Histogram 5: ~30 GeV to ~49 GeV*
    - *Histogram 6: ~9 GeV to ~20 GeV and ~88 GeV to ~95 GeV*
6. Histograms 3 and 4 may represent the same particle. Which histogram indicates a more precise estimate of the particle mass? Explain your reasoning.
    - *Histogram 3 had far fewer events; the maximum scale value was a frequency of 60 events. Histogram 4 had many more events with a maximum scale of 350 events. The plot with the most events has the narrower peak. Therefore, Histogram 4 has the most precise estimate.*
7. How many types of particles are represented in Histogram 6? Defend your claim.
    - *There are three peaks, but the low-mass peak may be background. Therefore, there are only two peaks that represent types of particles.*

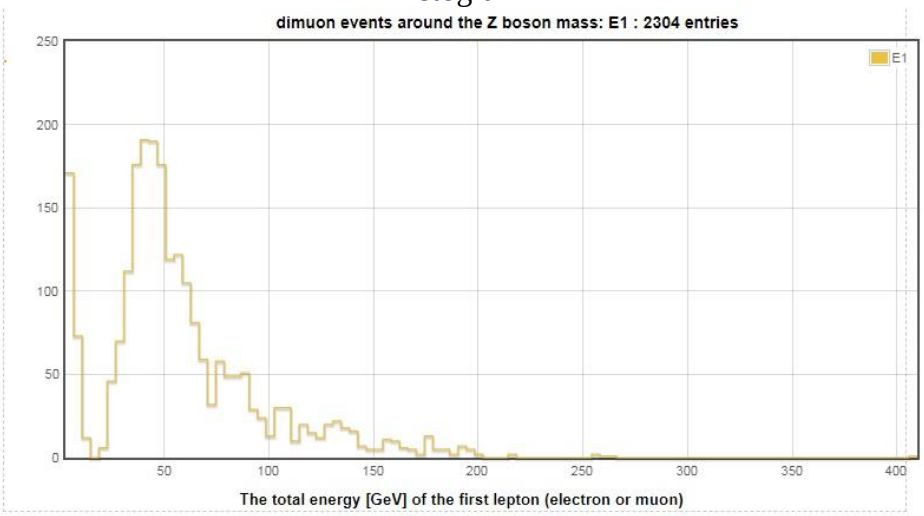
Histograms for Assessment

Note: In these plots, mass (GeV/c^2), momentum (GeV/c) and energy (GeV) are all reported in GeV in accordance with common usage by particle physicists in which c = 1.

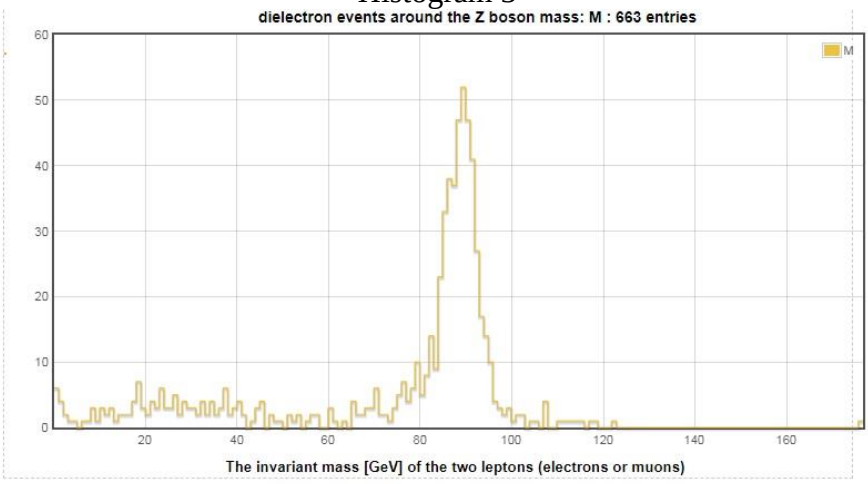
Histogram 1



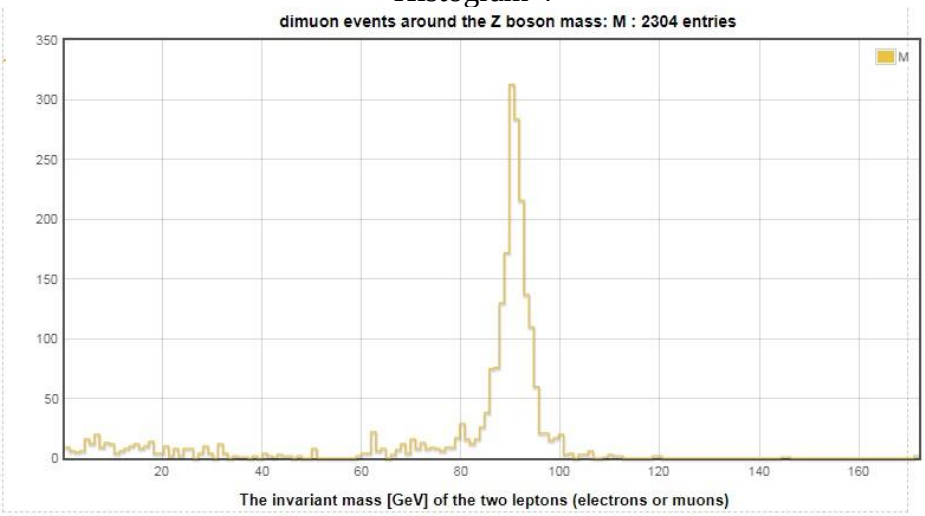
Histogram 2



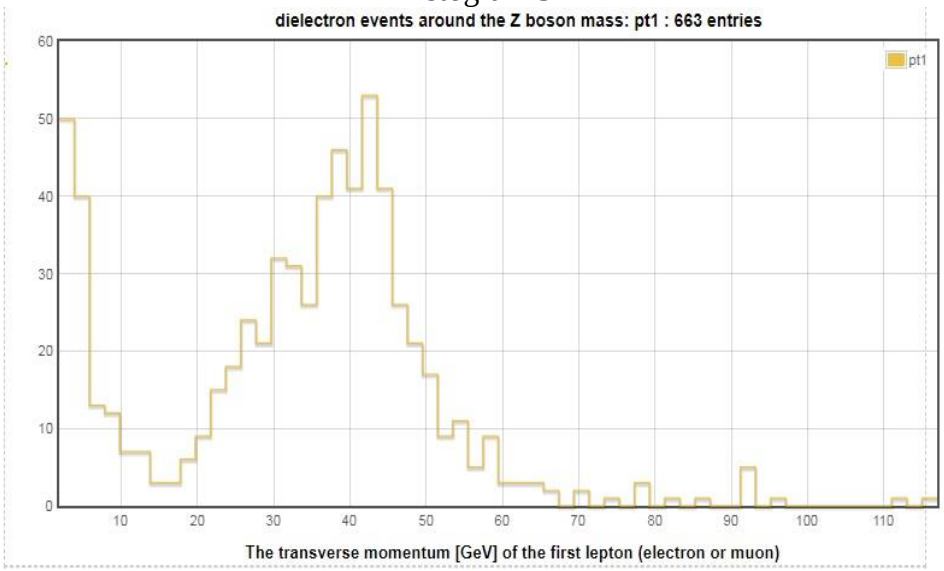
Histogram 3



Histogram 4



Histogram 5



Histogram 6