Video: “L9HistogramComparison” (2:14)

Video: (00:00)
Our final tricky issue is the correct way to use histograms to compare data sets. We’re going to compare the data from 751 Daphne Island finches with only 43 Santa Cruz Island finches. We’re definitely going to use percentages rather than counts because there are so many more Daphne birds. The vertical scales on the histograms would be completely different.

(00:23)
When doing a comparison, we want the viewer to be able to use the height of the bars to compare the percentage of values. Now the heights of the bars are equivalent in terms of percentages, but we’re not done, because the two graphs have different vertical scales. We’ll need to make them the same for comparison. Let’s set the scales from 0 to 40.

(00:46)
Here’s the result of setting the vertical scales of the two histograms to be the same. We’re closer, but not done yet. Because the horizontal scales are not the same, we cannot directly use the heights of the bars to compare the data. The final step is to use a common set of bins to histogram the data.

(1:03)
Here’s the final result. We see that on the whole, Santa Cruz finches have larger beaks than Daphne finches. The beak sizes of Santa Cruz finches are centered around a value close to 11mm, while Daphne finch beak sizes are centered at a value under 10mm. There are no Santa Cruz beak sizes less than 8mm, while about 10% of the Daphne finches are less than 8. A greater percentage of Santa Cruz finches have beak sizes larger than 12mm. These are rough estimates and we can easily calculate the exact values, but they give us a quick starting point for comparison.

(1:40)
Getting a common set of bins is a bit technical, but here’s the way to do it:
First, you have to find the common range. You do this by finding the larger of the two data set maxima, finding the smaller of the two data set minima, and then subtracting the smallest value from the largest value.

(2:01)
Bin both data sets using the same bin positions in the common range. You’ll have to experiment with the number of bins to make both histograms look reasonable.