HISTOGRAM LESSON: Questions

FOCUS QUESTION: How can I understand and compare the distributions of two data sets?

Contents

- EXAMPLE 1: Load the Daphne Island and Santa Cruz Island beak size data
- EXAMPLE 2: Display a histogram of the Daphne Island beak size data
- EXAMPLE 3: Use different choices of number of bins for Daphne Island histograms
- EXAMPLE 4: Compare beak distributions of Daphne and Santa Cruz Islands
- EXAMPLE 5: Calculate explicit histogram bin positions
- EXAMPLE 6: Compare percentages using scaling and explicit bin positions
- EXAMPLE 7: Calculate and display a histogram using a bar chart, line graph and stair plot
- EXAMPLE 8: Generate "random" numbers from three common probability distributions
- EXAMPLE 9: Display the histograms of the generated distributions

EXAMPLE 1: Load the Daphne Island and Santa Cruz Island beak size data

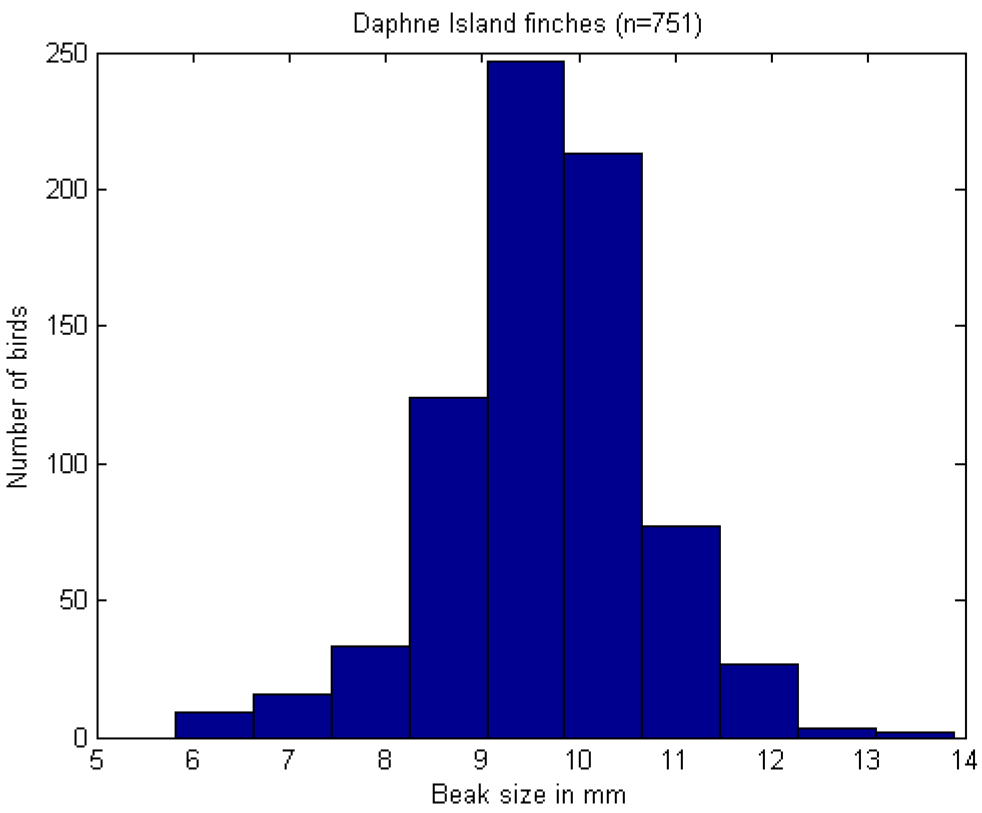
```
Daphne = load('DaphneBeaks.txt');

SantaCruz = load('SantaCruzBeaks.txt');
```

Questions	Answers
Why was the data for the two islands put in separate files?	The two data sets were not the same size, so you could not simply put them into two columns. Although you could use a missing designator to artificially make the two data sets the same size, the measurements are independent. Thus, arranging the two data sets as side-by-side columns in the same file might be misleading.

EXAMPLE 2: Display a histogram of the Daphne Island beak size data

```
nDaphne = length(Daphne);

titleDaphne = ['Daphne Island finches (n=' num2str(nDaphne) ')'];

figure('Name', titleDaphne);

hist(Daphne)

xlabel('Beak size in mm');
```

```
    ylabel('Number of birds');

    title(titleDaphne);
```

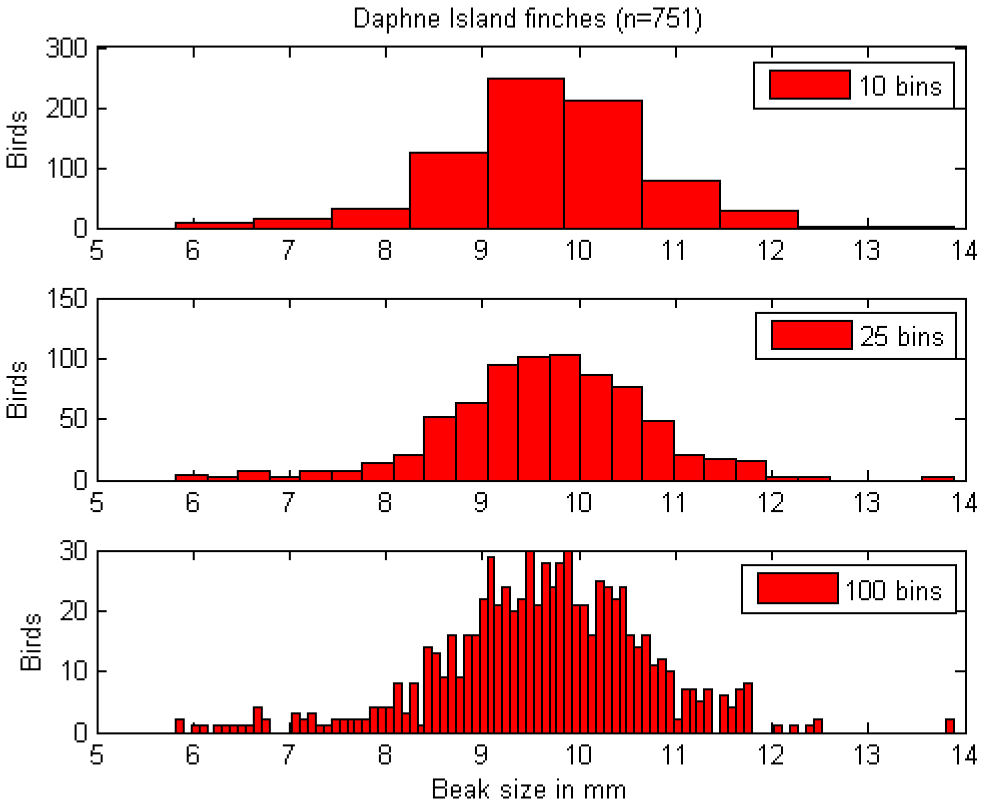


Questions	Answers
What is a histogram?	A histogram refers to a frequency table, that is a table listing how many times each value (or range of values) appears in a data set or to the display of that data, often using a bar chart.
What is a frequency table?	A frequency table records how many times each data value occurs in the data set. If the data set only has a small number of values, we keep a count for each possible value. For data sets that contain real numbers or have a large number of possible discrete values, we use a binned frequency table.
What does the `hist` function do?	The `hist` function computes a binned frequency table or histogram for the data. Since the example did not have output arguments, the resulting frequency table is plotted as a bar chart rather than returned as an array.
What is a binned frequency table?	A binned frequency table divides the possible data values into subranges called bins and counts how many values fall into each bin.
How many bins does `hist` use?	By default, the `hist` function uses 10 equal-sized bins that span the range of the data. (You may also explicitly specify the bins as in later examples.)

Questions	Answers
Does a histogram always have to be displayed as a bar chart?	No. The bar chart is a common visual representation of a histogram but not the only useful one.
Does the `hist` function always display a figure?	No. If you use the output arguments, as shown in the next example, the `hist` function does not produce a figure.

EXAMPLE 3: Use different choices of number of bins for Daphne Island histograms

```
figure

colormap autumn

subplot(3, 1, 1)

hist(Daphne, 10)

title(titleDaphne)

legend('10 bins')

ylabel('Birds')

subplot(3, 1, 2)

hist(Daphne, 25)

legend('25 bins')

ylabel('Birds')

subplot(3, 1, 3)

hist(Daphne, 100)

legend('100 bins')

ylabel('Birds')

xlabel('Beak size in mm')
```



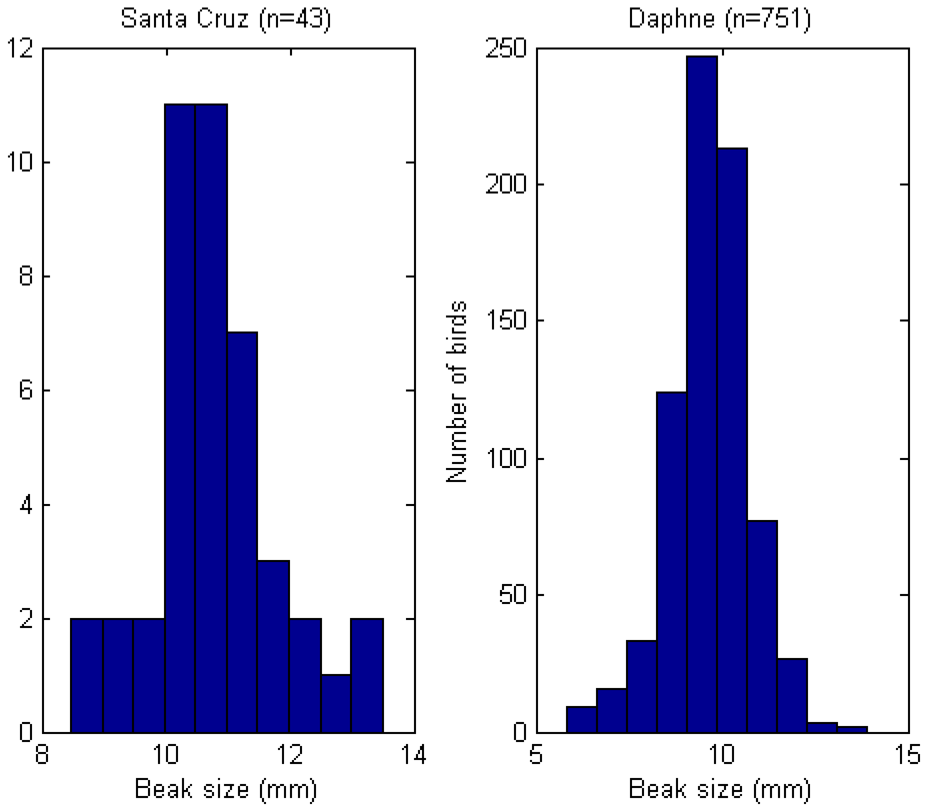
Questions	Answers
What does the `10` represent in the first call to `hist`?	The `10` specifies the number of bins to use in the frequency table. The default number of bins is 10. So the first call to `hist` behaves the same `hist(Daphne)`. The second call to `hist` uses 25 bins. Notice that the bars on the corresponding graph are thinner because more of them must fit in the same area.
Should I always use a large number of bins for a histogram?	Choosing the right bin size is sometimes a tricky trade-off. If you choose too few bins, the poor resolution may hide interesting features. If you choose too many bins, some bins will be sparsely occupied and the histogram may take on a jagged appearance. You may also miss essential features. It is usually good to experiment with the bin size to see what the trade-offs are.
How does MATLAB determine the positions of the bins?	MATLAB divides the range of data values into the specified number of bins. Your data can't contain `+inf` or `-inf`.
What happens if I move the `xlabel` statement after the first `hist`?	The `xlabel` adds an x-axis label to the current axis. The top histogram's x-axis will be labeled. Currently, the `xlabel` appears after the third `hist`, so only the third axis is labeled.
What happens if I move the `xlabel` statement directly after the first `subplot`?	The `xlabel` adds an x-axis label to the current axis, which was created by the `subplot`. However, the `hist` function creates a new axis, so the label is lost.

EXAMPLE 4: Compare beak distributions of Daphne and Santa Cruz Islands

```
    nSantaCruz = length(SantaCruz);
```

```
figure

subplot(1, 2, 1)

hist(SantaCruz)

title(['Santa Cruz (n=' num2str(nSantaCruz) ')'])

xlabel('Beak size (mm)')

subplot(1, 2, 2)

hist(Daphne)

title(['Daphne (n=' num2str(nDaphne) ')'])

xlabel('Beak size (mm)')

ylabel('Number of birds')
```

Questions	**Answers**
Why are the vertical scales of the two histograms different?	The counts depend on how many values each data set has. These data sets are of different size.
Why are the horizontal scales of the two histograms different?	The horizontal scales depend on the maximum and minimum values in the data set.
Can I still compare the distributions?	These histograms do not allow very effective comparison of the data. A more effective comparison would use the same bins and scale the data to be fractions of the data set rather than actual counts.



EXAMPLE 5: Calculate explicit histogram bin positions

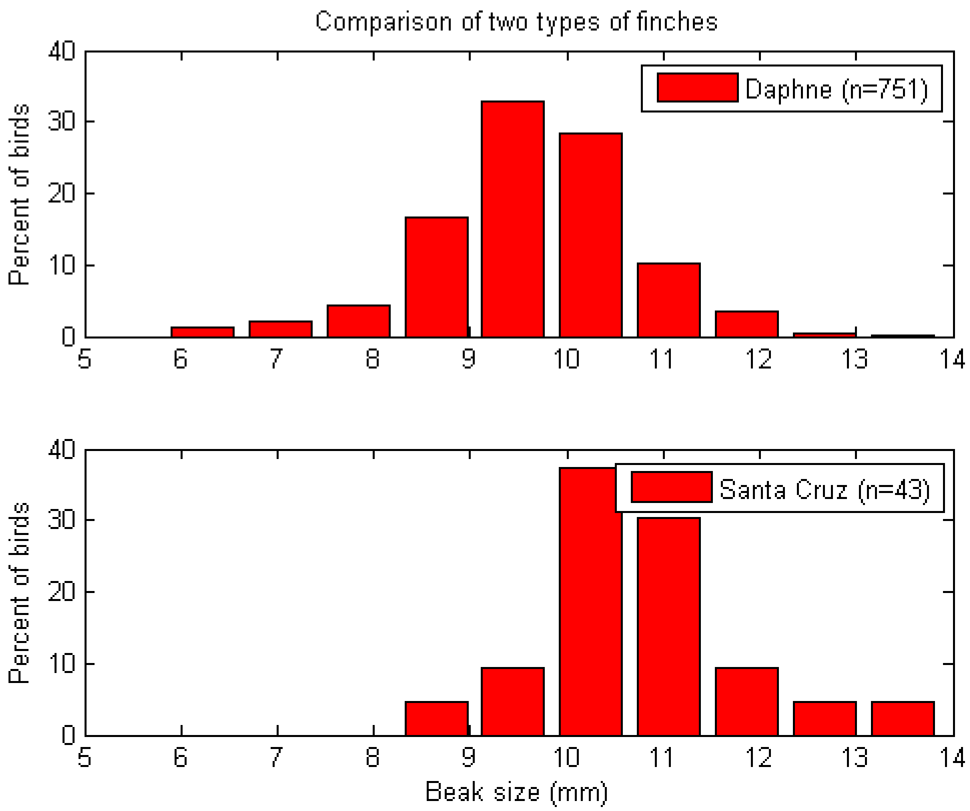
```
    minBeak = min([min(Daphne), min(SantaCruz)]);

    maxBeak = max([max(Daphne), max(SantaCruz)]);

    xEdges = linspace(minBeak, maxBeak, 11);

    xCenters = 0.5*(xEdges(2:end) + xEdges(1:end-1));

    nD = hist(Daphne, xCenters);

    nS = hist(SantaCruz, xCenters);
```

Questions	Answers
Why was the `min` of the `min` needed to find the minimum beak size?	The inner pair of `min` functions finds the minimum values of the Daphne and Santa Cruz data individually. The square brackets combine these values into a two-element vector. We need to apply another `min` to find the overall minimum.
What is the first element in	The first element is the value of `minBeak`.

Questions	Answers
linspace(minBeak, maxBeak, 10)?	

EXAMPLE 6: Compare percentages using scaling and explicit bin positions

```
figure

colormap autumn

subplot(2, 1, 1)

bar(xCenters, 100*nD/nDaphne)

legend(['Daphne (n=' num2str(nDaphne) ')'])

title('Comparison of two types of finches')

ylabel('Percent of birds')

subplot(2, 1, 2)

bar(xCenters, 100*nS/nSantaCruz)

legend(['Santa Cruz (n=' num2str(nSantaCruz) ')'])

ylabel('Percent of birds')

xlabel('Beak size (mm)')
```

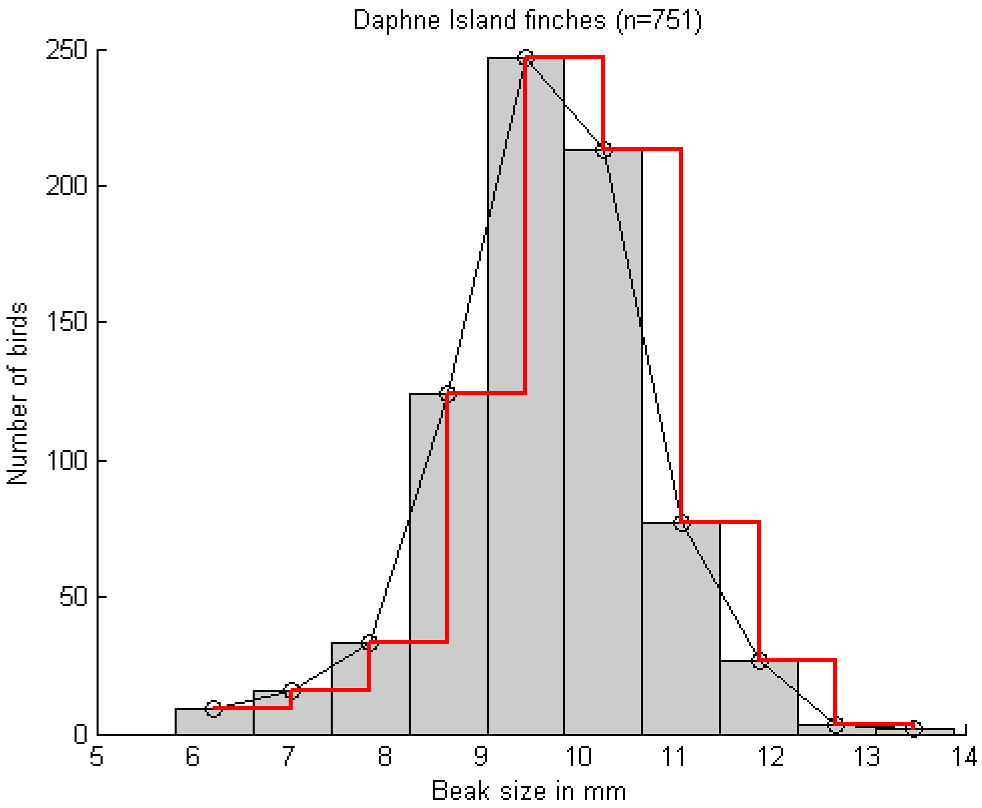


Questions	**Answers**
Why was the `nDaphne` divided by `sum(nDapne)`?	Since the data sets did not have the same number of elements, a comparison of the counts is not meaningful. Dividing by the total number of elements plots the fractions, which are comparable.

EXAMPLE 7: Calculate and display a histogram using a bar chart, line graph and stair plot

```
[n, xout] = hist(Daphne);

figure

hold on

bar(xout, n, 1.0, 'FaceColor', [0.8, 0.8, 0.8]);

plot(xout, n, '-ok')

stairs(xout, n, 'r', 'LineWidth', 2)

hold off
```

```
    xlabel('Beak size in mm');

    ylabel('Number of birds');

    title(titleDaphne);

    datacursormode on
```



Questions	Answers
How does `[n, xout] = hist(Daphne)` differ from the `hist(Daphne)` of EXAMPLE 2?	When you use output arguments with the `hist` function, MATLAB does not draw a figure. Rather the `hist` function returns the frequency counts and the centers of the bins.
Did I need to assign the result of `hist` to variables?	Yes, if you want to get the values in the frequency table rather then to just see a plot. Use the form with output arguments when you want to do your own display or if you want to compute something else from the frequency table.
When would I need the bin positions and counts from a histogram?	This example illustrates using these values to display the histogram in three different ways.
When would I need the bin positions	This example illustrates using these values to display the histogram in three different ways.

Questions	Answers
and counts from a histogram?	
What does `'-ok'` mean in `plot`?	The `'-ok'` is shorthand for black (`k`) circular markers (`o`) that are connected with a solid lines (`-`).
What is the 1.0 argument of `bar` do?	This argument specifies the relative width of the bars. By default, this value is 0.8, meaning that the bars only take up a fraction 0.8 (80%) of the available space, leaving a gap of 20%. The value 1.0 specifies that the bars should take up 100% of the available space with no gap between. Histograms typically use a bar chart without the gap.
What is the difference between `plot` and `stairs`?	The `plot` function connects each consecutive (x, y) pair with a straight line. The `stairs` function connects each consecutive (x, y) pair with a staircase. MATLAB draws a horizontal line between the x values at the level of the first y value. At the second x value, MATLAB draws a vertical line between the two y values to form a stair.

EXAMPLE 8: Generate "random" numbers from three common probability distributions

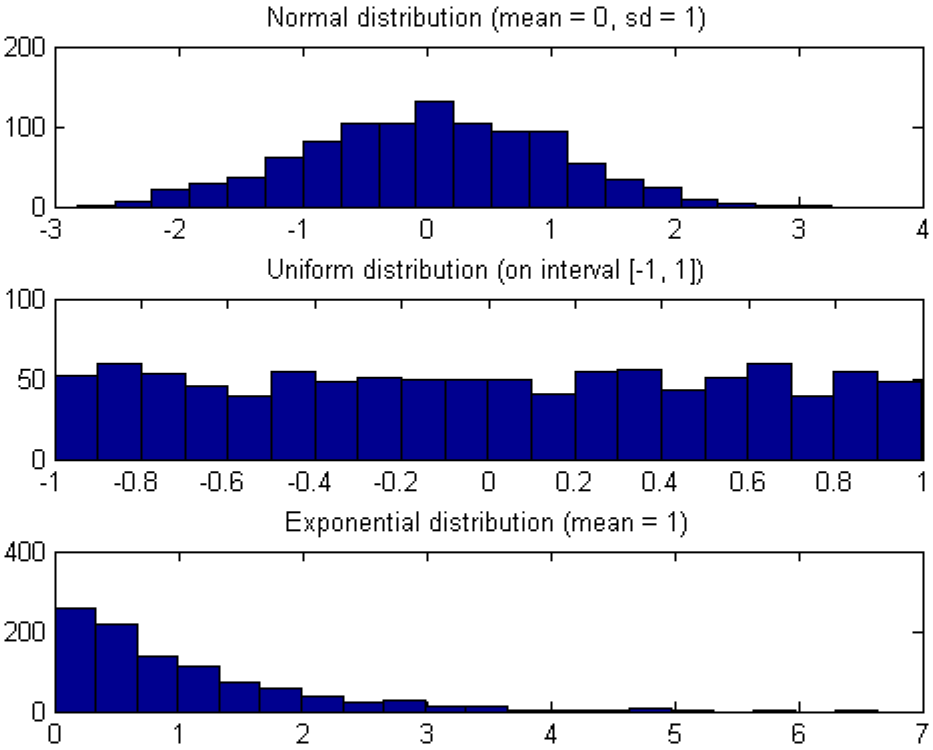
```
    yNormal = random('norm', 0, 1, [1000, 1]);

    yUniform = random('unif', -1, 1, [1000,1]);

    yExp = random('exp', 1, [1000, 1]);
```

Questions	Answers
Why are the values returned by `random` called pseudo-random?	The sequence of values produced by `random` is generated by a formula and completely predictable from the implementation of `random`.
Won't I always generate the same values each time I call `random`?	Although the sequence of values is predictable, you can pick different places (the seed) to start, giving the appearance of unpredictability.

EXAMPLE 9: Display the histograms of the generated distributions

```
    figure

    subplot(3,1,1)

    hist(yNormal, 20)

    title('Normal distribution (mean = 0, sd = 1)')



    subplot(3,1,2)
```

```
hist(yUniform, 20)

title('Uniform distribution (on interval [-1, 1])')



subplot(3,1,3)

hist(yExp, 20)

title('Exponential distribution (mean = 1)')
```



This lesson was written by Kay A. Robbins of the University of Texas at San Antonio and last modified by Dawn Roberson on 21-Jan-2014. Please contact kay.robbins@utsa.edu with comments or suggestions.

Published with MATLAB® 7.14