



This is a graded discussion: 10 points possible

due -

23 39

Histogram for showing percentiles

YY: "Ok, histograms are awesome. But, I'm thinking it'd be really cool if we can modify the histogram so that we can easily figure out where is the median, quartiles, and so on."

ZZ: "Well, why not just use a boxplot?"

YY: "Yeah that's true. Boxplot shows median and quartiles... But, how about other percentiles? Say I want to know where is the 10 percentile and 90 percentile points in the data range. You can't do that with the boxplot, can you?"

ZZ: "Hmm.. That's true."

Can you help YY & ZZ? Is there any way to modify a normal histogram so that it can not only show the distribution of data, but also let you know (roughly) where are the percentile points?



← Reply



Mukul Gharpure (<https://iu.instructure.com/courses/2165942/users/6678592>)

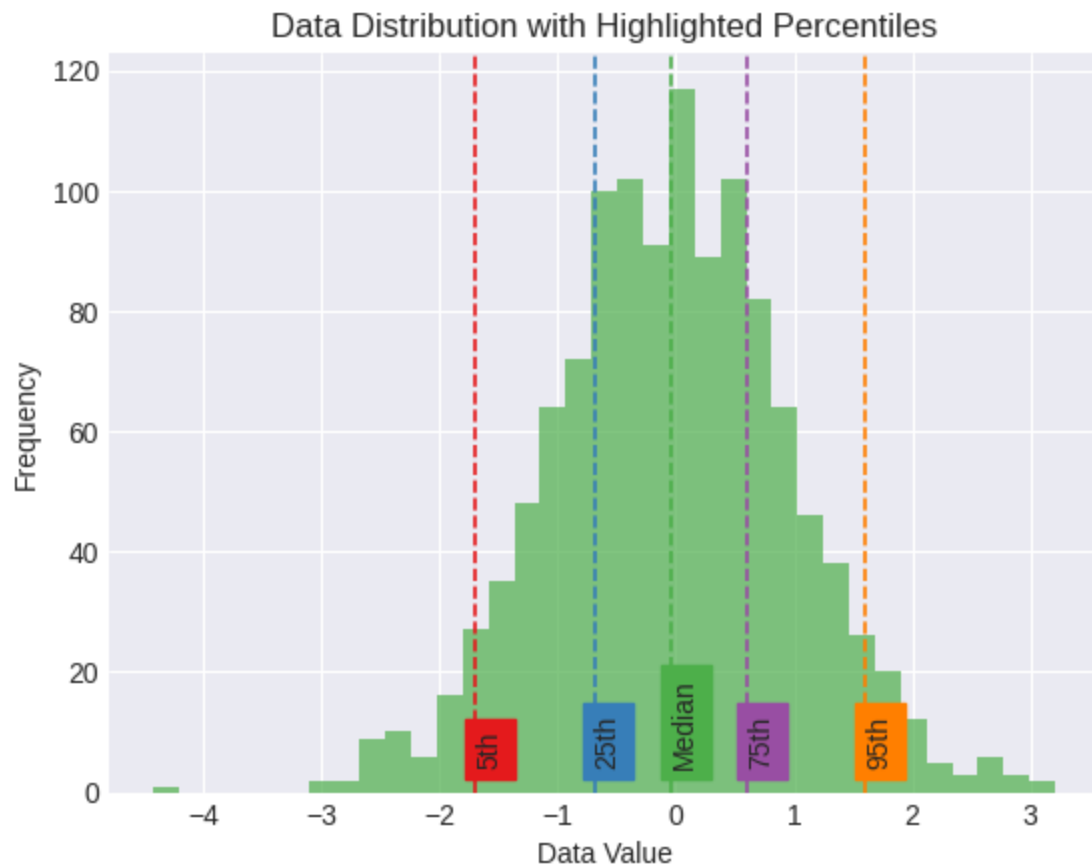
Saturday



Yes, A histogram can be modified to show percentiles by overlaying vertical lines representing the desired percentile values on top of the histogram. Here's how we achieve it:

1. Firstly, compute the desired percentiles of the dataset using a percentile computation function, like `np.percentile()`.
2. Then plot the histogram as you would normally.
3. After that, add vertical lines to the histogram at the values calculated.

Attached in the answer is representation for the same.



← [Reply](#)  (1 like)



<https://iu.instructure.com/courses/2165942/users/6679606>

Sep 25, 2023

YY and ZZ have a valid point. While histograms are excellent for visualizing the distribution of data, they do have limitations when it comes to pinpointing specific percentile points.

A way to modify a normal histogram would be to first create a histogram, as usual, and then calculate the cumulative frequency of the data by adding up the frequencies of all previous bins. The last cumulative frequency should be equal to the total number of data points. Then estimate the approximate location of specific percentiles by interpolating within the cumulative frequency curve.

This method provides a rough estimate of where specific percentile points fall within the data range, beyond median and quartiles.

← [Reply](#) 

<https://iu.instructure.com/courses/2165942/users/6701599>**Thomas Jablenski** (<https://iu.instructure.com/courses/2165942/users/6701599>)

Sep 25, 2023



A way to add to the normal histogram is to add vertical lines on top of your chart depicting every 10th percentile. This could get messy and look cluttered especially if a majority of the data is close together.

[← Reply](#) <https://iu.instructure.com/courses/2165942/users/6701715>**Dustin Cole** (<https://iu.instructure.com/courses/2165942/users/6701715>)

Sep 26, 2023



The best option would be creating a density curve, but if you want to do it with a histogram, you could use a large number of bins. Then you could do some math to plot the distribution percentages on the histogram so people can see the different percentiles on the x axis.

[← Reply](#) <https://iu.instructure.com/courses/2165942/users/6587577>**Yumeng Liang** (<https://iu.instructure.com/courses/2165942/users/6587577>)

Wednesday



I think YY and ZZ can modify a normal histogram to show percentile points by adding a cumulative distribution curve. Calculate the cumulative distribution function for the data, representing the probability that a data point is below a certain value. Plot it alongside the histogram, and mark specific percentiles by drawing lines at corresponding to the percentage values.

[← Reply](#) <https://iu.instructure.com/courses/2165942/users/6684610>**Shantanu Dixit** (<https://iu.instructure.com/courses/2165942/users/6684610>)

Thursday



After plotting the histogram, calculate the data values for the percentiles. Then, draw vertical lines at these values on the histogram. These lines show where the percentiles are in the data.

Edited by **Shantanu Dixit** (<https://iu.instructure.com/courses/2165942/users/6684610>) on Sep 28 at 7:09pm

← Reply 👍



Erik Gonzalez (<https://iu.instructure.com/courses/2165942/users/6352173>)

Thursday

I would recommend plotting a normal histogram, but overlaying reference lines on the graph to call out the relevant percentiles by using the np.percentile function

← Reply 👍



Erik Gonzalez (<https://iu.instructure.com/courses/2165942/users/6352173>)

Thursday

Reading through some of the above suggestions made me think, a good alternative may be leveraging 10 bins (assuming your goal is to show every 10th percentile) and variable bin widths, so each bin end point acts as a reference line for the percentiles.

← Reply 👍



Carmen Galgano (<https://iu.instructure.com/courses/2165942/users/6762945>)

Thursday

You can modify a histogram to show percentile points as well. All you need to do is overlay percentile lines on the histogram and label so people can tell which quartile is there/where the median is. In matplotlib, you can add dotted percentile lines.

← Reply 👍



Andi Mai (<https://iu.instructure.com/courses/2165942/users/6705680>)

Thursday

When plot the histogram, we can include approximate locations of percentile points by adding vertical lines or markers at those points.

← Reply 👍



<https://iu.instructure.com/courses/2165942/users/6813278> **Shreedeeep Sadasivan Nair (he/him/his)** (<https://iu.instructure.com/courses/2165942/users/6813278>)

Friday



We can find out the cdf for a distribution plot the cdf over the standard histogram and the mark the points for the desired quantiles

[Reply](#)



<https://iu.instructure.com/courses/2165942/users/6703376> **Sangzun Park** (<https://iu.instructure.com/courses/2165942/users/6703376>)

Friday



The first thing that came to mind was to scale the number of bins to 100 or a ratio equal to 100. But a better way is to insert a PMF line on top of the histogram. I think this is a good way to compensate for the shortcomings of the histogram.

[Reply](#)



<https://iu.instructure.com/courses/2165942/users/6758180> **Onur Tekiner** (<https://iu.instructure.com/courses/2165942/users/6758180>)

Friday



I think I can visualize the data with a histogram with a boxplot on the corner of the graph simultaneously.

Also, if the dataset is a normal distribution, each standard deviation from the mean point is considered 33 percent of the data. Creating colorful bins of each standard deviation from the mean could make it easier to interpret quartiles.

[Reply](#)



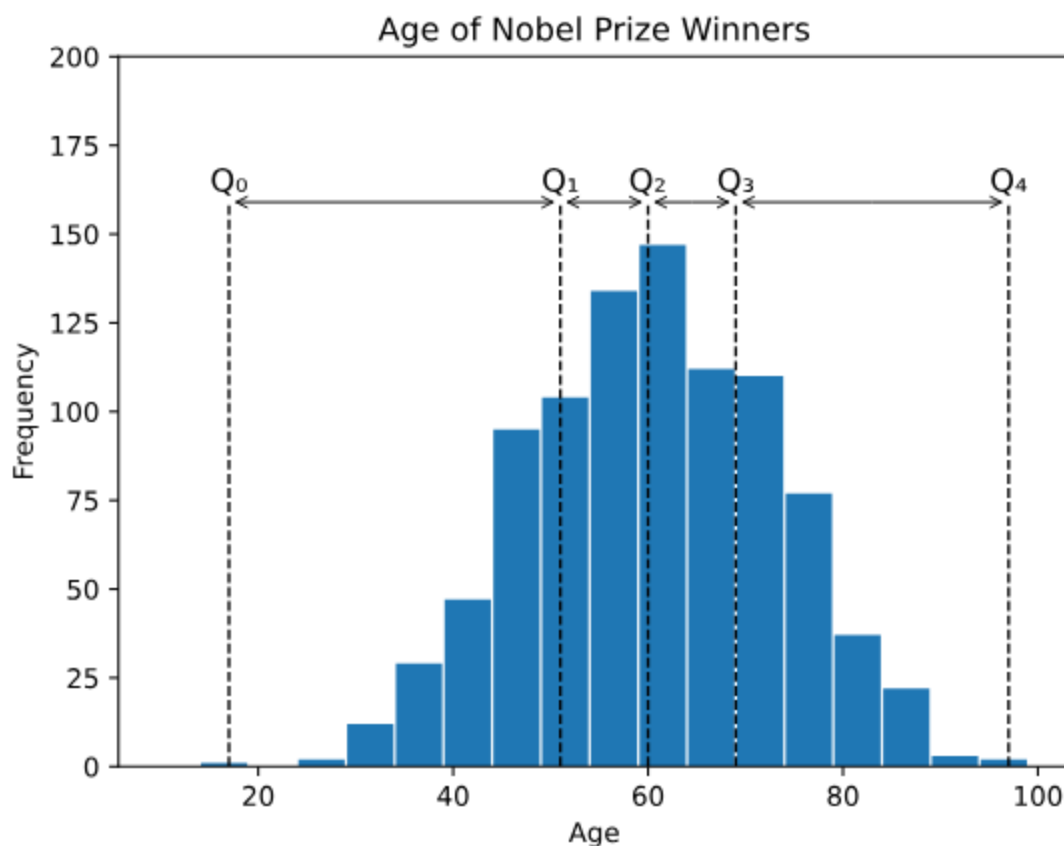
<https://iu.instructure.com/courses/2165942/users/6684842> **Prem Amal** (<https://iu.instructure.com/courses/2165942/users/6684842>)

Saturday



We can achieve this by first creating a standard histogram that provides an overview of data frequencies within specified bins. Then, we can calculate the percentiles of interest, such as


the median, quartiles, or any other specific percentiles, using mathematical tools or libraries. The magic happens when we overlay vertical lines onto the histogram at the positions corresponding to these percentile values. These lines act as visual indicators, clearly showing where these key data points are situated along the x-axis. To make the visualization even more informative, we can label each percentile line with its respective value and include a legend for clarity. This approach transforms a standard histogram into a powerful tool for not only grasping the overall data distribution but also pinpointing the exact locations of specific percentiles



Reference:

[https://www.google.com/url?](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.w3schools.com%2Fstatistics%2Fstatistics_quartiles_and_percentiles.php&psig=AOvVaw2RtJ3BXmkpAwL1ILT6z2OI&ust=1696172192411000&source=images&cd=vfe&opi=89978449&ved=0CBAQjRxqFwoTCKi-lfrL0oEDFQAAAAAdAAAAABAE)

[sa=i&url=https%3A%2F%2Fwww.w3schools.com%2Fstatistics%2Fstatistics_quartiles_and_percentiles.php&psig=AOvVaw2RtJ3BXmkpAwL1ILT6z2OI&ust=1696172192411000&source=images&cd=vfe&opi=89978449&ved=0CBAQjRxqFwoTCKi-lfrL0oEDFQAAAAAdAAAAABAE](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.w3schools.com%2Fstatistics%2Fstatistics_quartiles_and_percentiles.php&psig=AOvVaw2RtJ3BXmkpAwL1ILT6z2OI&ust=1696172192411000&source=images&cd=vfe&opi=89978449&ved=0CBAQjRxqFwoTCKi-lfrL0oEDFQAAAAAdAAAAABAE)

[← Reply](#) **Mothi Gowtham Ashok Kumar** ([he/him/his](https://iu.instructure.com/courses/2165942/users/6683278)) (<https://iu.instructure.com/courses/2165942/users/6683278>)

Saturday


There are a few ways to modify a normal histogram so that it can also show percentiles.

One way is to add vertical lines to the histogram at the desired percentile points. For example, to add lines for the median, quartiles, and 10th and 90th percentiles, you could do the following:

1. Calculate the percentiles of the data.
2. Add vertical lines to the histogram at the percentile values. You can use a different color or line style for each percentile.
3. Label the lines with the corresponding percentiles.

Another way to modify a histogram to show percentiles is to use a cumulative histogram. A cumulative histogram shows the number of data points that are less than or equal to each value in the data set. To create a cumulative histogram, you can use the following steps:

1. Sort the data in ascending order.
2. Calculate the cumulative frequency for each value in the data set. The cumulative frequency is the number of data points that are less than or equal to the current value.
3. Create a histogram of the cumulative frequencies.

[← Reply](#) **Vedant Tapadia** (<https://iu.instructure.com/courses/2165942/users/6678810>)

Saturday

If we set the number of bins to 10 we can always see that 10th percentile is the 1st bin 20th percentile is the 2nd bin and so on until the 10th bin.


We can also make vertical lines in the histogram that shows every 10th percentile.

[← Reply](#) 

<https://iu.instructure.com/courses/2165942/users/6443321>

Saturday

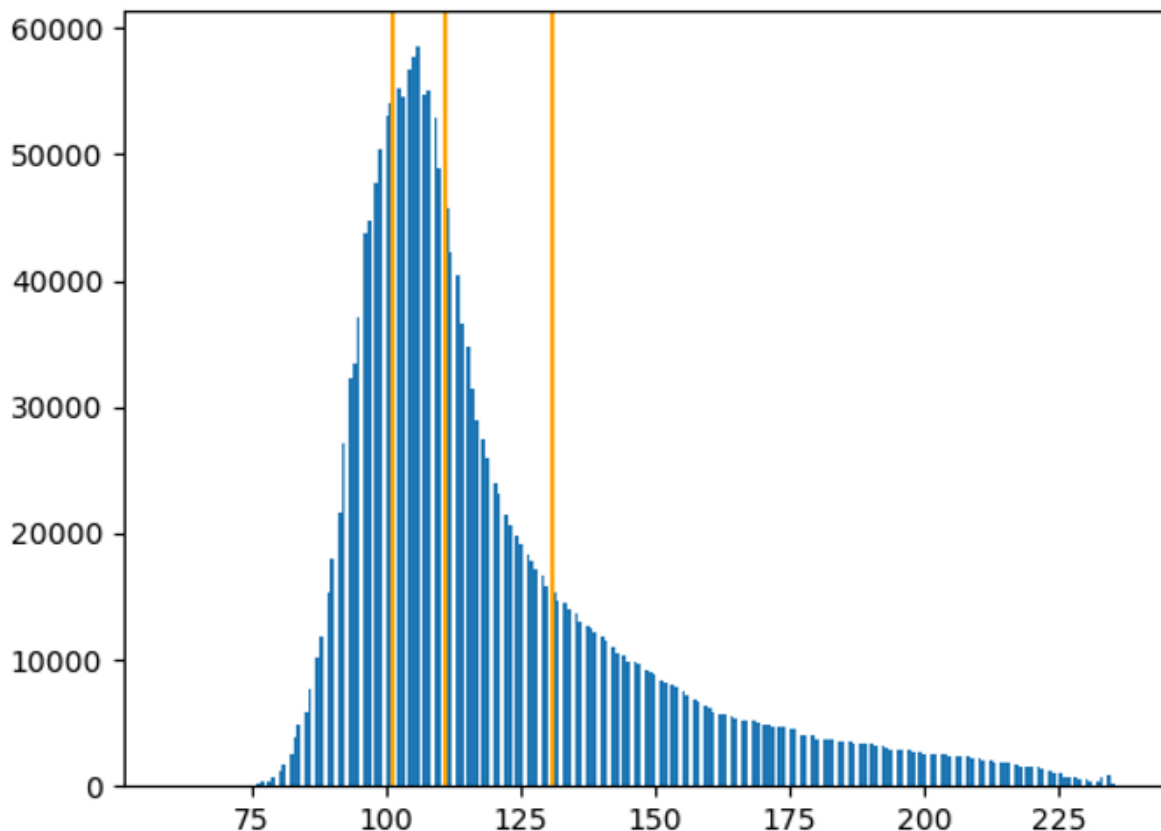
If histogram can show the distribution and box plots can show median and quartiles, I think that box plots can be added to the normal histogram.

[← Reply](#) 


<https://iu.instructure.com/courses/2165942/users/6701521>



Saturday

One method to display percentiles on a histogram is to first specify the percentiles you want to call out (for example, the quartiles) in an array. Then plot that array on the histogram with the distribution you're visualizing, using a color that stands apart. The result is something that looks a little like this:



You can view this source image and read about the method here:

<https://www.aivia-software.com/post/python-quick-tip-2-plotting-image-histograms> 
(<https://www.aivia-software.com/post/python-quick-tip-2-plotting-image-histograms>)

 [Reply](#) 





Hymavathi Gummudala (<https://iu.instructure.com/courses/2165942/users/6679250>)



Saturday

To roughly estimate percentiles in a histogram, we can add vertical lines and labels for specific percentiles (e.g., 10th and 90th percentiles) to the plot.

 [Reply](#) 



Jeevan Deep Mankar (<https://iu.instructure.com/courses/2165942/users/6644229>)



Sunday

Yes, there are a few ways to modify a normal histogram so that it can show not only the distribution of data, but also show the percentile points.

One way is to add vertical lines to the histogram at the desired percentiles. For example, to add lines at the 10th and 90th percentiles, we would first need to calculate those percentiles. Once we have the percentiles, we can add vertical lines to the histogram at those values.

Another way to modify a histogram to show percentiles is to use a cumulative histogram. A cumulative histogram shows the percentage of data points that are less than or equal to a given value. To create a cumulative histogram, we would first need to sort the data in ascending order. Then, we would need to calculate the cumulative percentage for each data point. The cumulative percentage for a data point is the number of data points that are less than or equal to that data point divided by the total number of data points. Once we have the cumulative percentages, we can plot a cumulative histogram.

 [Reply](#) 

[https://](https://iu.instructure.com/courses/2165942/users/6760559)
Madhuri Patibandla (she/her/hers) (<https://iu.instructure.com/courses/2165942/users/6760559>)

Sunday

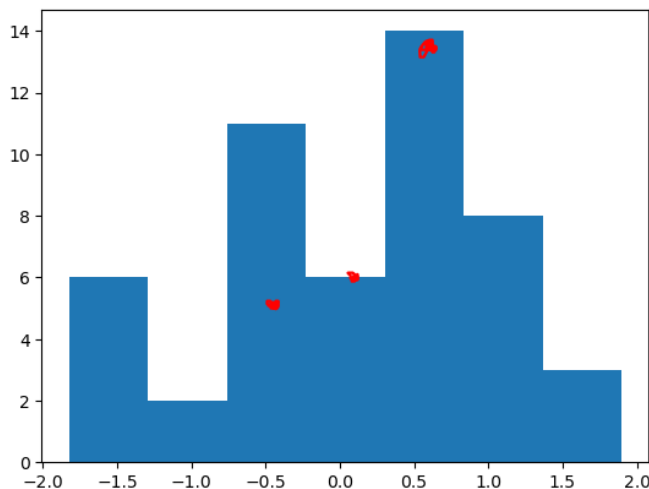
As YY said. In histogram we can figure out the median, 25th and 75th percentiles.

For ex: for 50 consecutive data points, if we define the bins correctly, we can figure the out median is 25.5 round as 26, 25th quartile is 13.25 and 75th quartile is 37.75.

for same 50 consecutive data points developed a histogram and highlighted median, 25th and 75th quartile.

```
In [6]: import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import altair as alt
import pandas as pd
%matplotlib inline
norm_dist = np.random.randn(50)
plt.hist(norm_dist, bins=7)
```

```
Out[6]: (array([ 6.,  2., 11.,  6., 14.,  8.,  3.]),
array([-1.8218155, -1.29076742, -0.75971935, -0.22867128,  0.3023768 ,
        0.83342487,  1.36447294,  1.89552102]),
<BarContainer object of 7 artists>)
```



In Box plot it is easy to identify median, 25th and 75th, 95th and 5th percentile, but not 90th and 10th percentile. .

← [Reply](#)

[https://](https://iu.instructure.com/courses/2165942/users/6684840)
Jash Shah (<https://iu.instructure.com/courses/2165942/users/6684840>)

Sunday

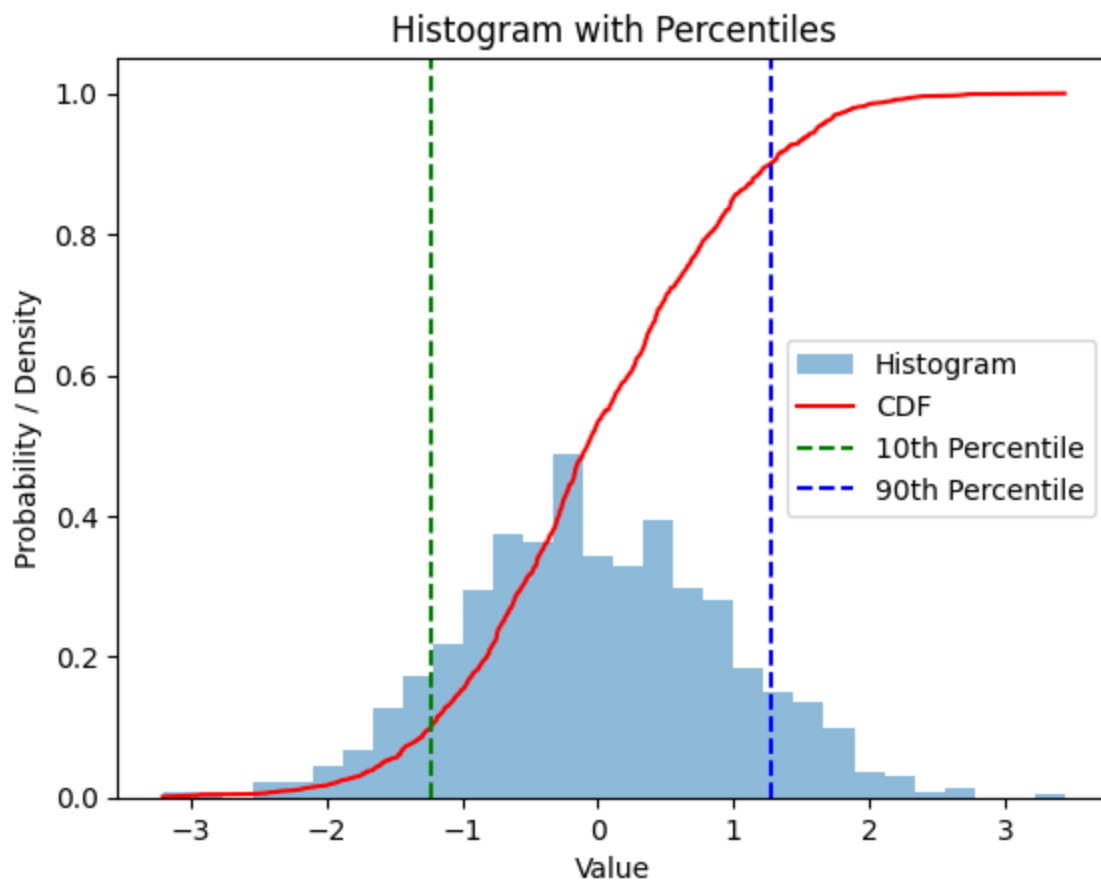
Creating a Histogram with Percentile Lines

To enhance a standard histogram's capability to display key percentile values, such as the median, quartiles, or any other desired percentiles, we can follow these steps:

1. Begin by constructing a standard histogram, which offers an initial view of data frequencies grouped within predefined bins.
2. Calculate the desired percentiles (e.g., median, quartiles) using mathematical methods or dedicated libraries.
3. Overlay vertical lines onto the histogram plot at positions corresponding to the calculated percentile values. These lines serve as visual cues, precisely indicating where these significant data points are located along the x-axis.
4. To enhance clarity, label each percentile line with its respective percentile value. This labeling provides an informative reference for interpreting the histogram.
5. Including a legend within the visualization ensures that the purpose and meaning of the percentile lines are readily understood, making the histogram a powerful tool for grasping both the overall data distribution and the specific locations of key percentiles.

Reference

https://www.w3schools.com/statistics/statistics_quartiles_and_percentiles.php 
(https://www.w3schools.com/statistics/statistics_quartiles_and_percentiles.php)



Edited by **Jash Shah** (<https://iu.instructure.com/courses/2165942/users/6684840>) on Oct 1 at 12:59am

← Reply

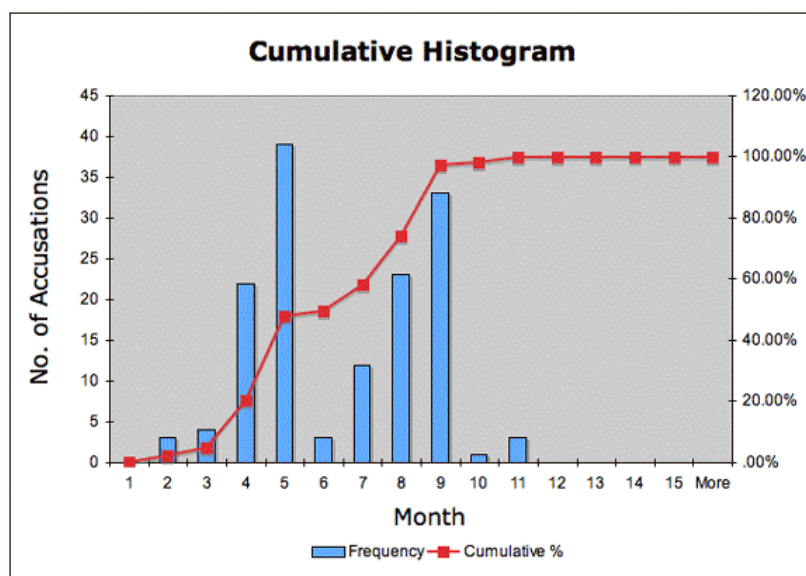


([https://](https://iu.instructure.com/courses/2165942/users/6818242)

Simon Driver (<https://iu.instructure.com/courses/2165942/users/6818242>)

Sunday

Yes, what we could do is use a cumulative histogram; in addition to showing the frequency of each observation, it also adds in a line that shows what % of the total observations are at what point. Hence, you could easily read off the 10th or 90th percentile, or the median, from that additional line. An example is shown below, which is charting the frequency of accusations during the Salem witch trials by month:



As one can see, the red line show the cumulative frequency, so I can see that the 50% (or median) mark is at around mark 6, and so forth.

Source: <https://www2.tulane.edu/~salem/Cumulative%20Histogram.html>

← Reply




([https://](https://iu.instructure.com/courses/2165942/users/5667580)

Sarah Biggs (<https://iu.instructure.com/courses/2165942/users/5667580>)

Sunday

I had to do some research for this, but I think I found a method at this link: <https://www.aivia-software.com/post/python-quick-tip-2-plotting-image-histograms> → (<https://www.aivia-software.com/post/python-quick-tip-2-plotting-image-histograms>)

Functionally, you're going through your data points that align with a set of named percentiles (that you're most interested in), and plotting those as vertical lines on your histogram. You can do this utilizing numpy and matplotlib.

← [Reply](#) 




Sydney Dicks (<https://iu.instructure.com/courses/2165942/users/6819877>)

Sunday

Though it may difficult to interpret the histogram, if you wanted to see the percentile points, you could make the number of bins equal to 10 with variable bin size so that each 10 percentage points would have its own bin. With the axis labeled, you could then see where the percentile points were located.

You could also add vertical lines to the histogram to indicate where the specific percentile points are located.

← [Reply](#) 



Maria Klein (<https://iu.instructure.com/courses/2165942/users/5444499>)

Sunday

I think so...you could draw vertical lines at the standard deviations away from the mean corresponding to whatever percentile points you were interested in, or at percentiles that are of frequent interest, like including 1 line each at ± 1.96 standard deviations away from the mean, leaving 0.025 and 0.975 of the distribution to the right of each line.

Edited by **Maria Klein** (<https://iu.instructure.com/courses/2165942/users/5444499>) on Oct 1 at 7:14pm

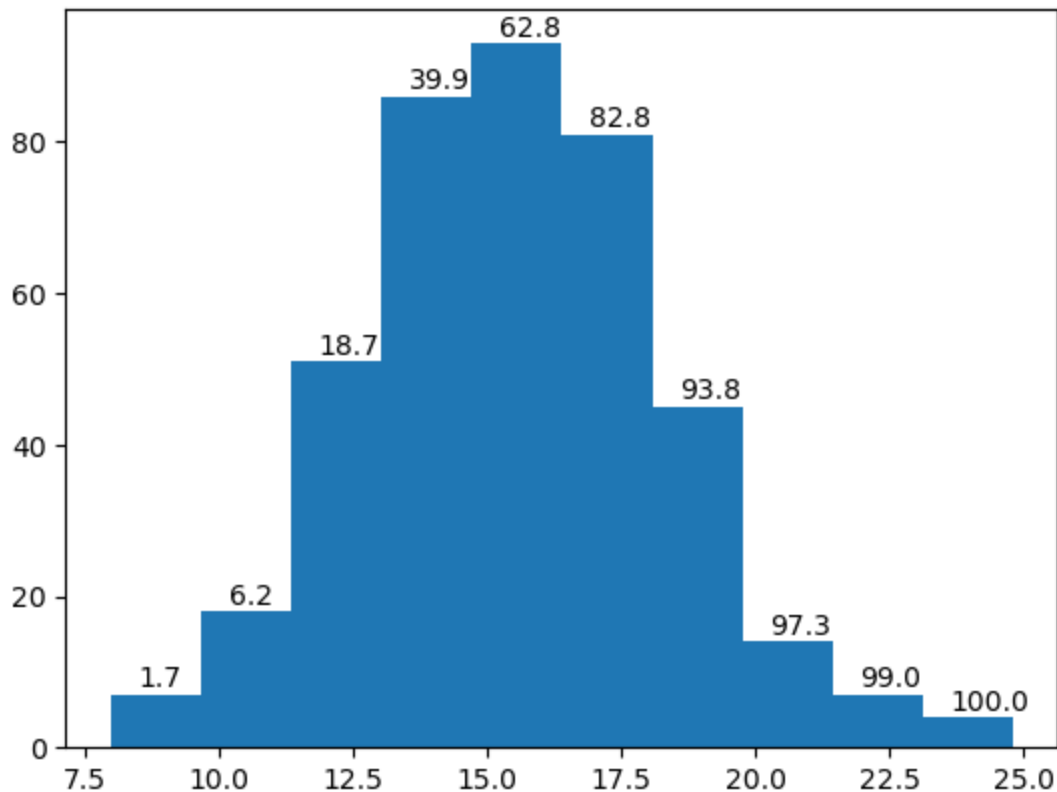
← [Reply](#) 



Gary Croke (<https://iu.instructure.com/courses/2165942/users/6706306>)

Sunday

It's fairly straight forward to calculate percentiles for any set of histogram bins given their divisions and counts. Here the percentiles are calculated after bins are determined. It would also be possible to set bin divisions for a desired set of percentiles.



Edited by [Gary Croke \(https://iu.instructure.com/courses/2165942/users/6706306\)](https://iu.instructure.com/courses/2165942/users/6706306) on Oct 1 at 7:54pm
[m08_histogram_percentiles.html \(https://iu.instructure.com/files/163006921/download?download_frd=1&verifier=MBAteDyWwJA3mk3wb3hDEfB3yseqhOpmnhTreyud\)](https://iu.instructure.com/files/163006921/download?download_frd=1&verifier=MBAteDyWwJA3mk3wb3hDEfB3yseqhOpmnhTreyud)

← [Reply](#)



[Ao Zhang \(https://iu.instructure.com/courses/2165942/users/6703098\)](https://iu.instructure.com/courses/2165942/users/6703098)

Sunday

It is likely to show the frequency of different bins. Since we know the total sample size, we can calculate to get the rough results of percentile points. Another way is using different colors to show the different groups in different percentiles. For example, we could use blue to stand for the groups in 10%, yellow stands for the groups from 10% to 50%, and grey stands for the rest of groups.


← [Reply](#)



[Olufisola Oladipo \(https://iu.instructure.com/courses/2165942/users/6469527\)](https://iu.instructure.com/courses/2165942/users/6469527)

Yesterday

Yes, there is a way to calculate percentile points in a normal histogram. One way is to determine the value for each of the bins in the histogram, then adding the values together to give a total. Then to determine either 10 or 90 percentile, one has to take each bin value and divide by the total bin value and multiply by 100 to get the percentile for each bin. Therefore, 10 or 90 percentile will fall within one of the bins calculated percentile.

← [Reply](#) 

○



<https://iu.instructure.com/courses/2165942/users/6056428>

Yesterday

⋮

Yes there is a way this can be done. You can first create the histogram, then calculate the percentile values, then add vertical lines to the plot at the positions based on their calculated percentile values.

← [Reply](#) 

○




<https://iu.instructure.com/courses/2165942/users/6682743>

Yesterday

⋮

We can add vertical lines at several points in the graph showing the percentile at that point

← [Reply](#) 

○



<https://iu.instructure.com/courses/2165942/users/6694681>

Yesterday

⋮

Here's how we can do it:

1. Calculate the percentiles you want to display (e.g., 10th, 25th, 50th - median, 75th, and 90th percentiles).
2. Create a histogram in python or R
3. After creating the histogram, add vertical lines or markers at the calculated percentile positions.

← [Reply](#) 

○

**Harsh Patel** ([he/him/his](https://iu.instructure.com/courses/2165942/users/6825193))

Yesterday

⋮

To modify a normal histogram so that it can not only show the distribution of data we can be add percentile lines to the histogram.

```
np.percentile(data, 10)
```

```
np.percentile(data, 90)
```

We can use this to calculate the percentiles. Then we can add the percentile lines. By adding percentile lines to your histogram, you can get a rough idea of where specific percentiles are located within your data distribution, in addition to visualizing the overall data distribution.

[← Reply](#)

○

**Ram Kiran Devireddy** (<https://iu.instructure.com/courses/2165942/users/6677399>)

Yesterday

⋮

Yes, we can modify a histogram to show not only the distribution of data but also show approximate percentile points. One way to do that is by adding vertical lines or markers to the histogram plot to show specific percentiles. We can use numpy to calculate these percentiles and then add vertical lines at these percentile points.

[← Reply](#)

○

**Rohan Isaac** (<https://iu.instructure.com/courses/2165942/users/6694525>)

1:21pm

⋮

Most probably by providing x-ticks that show percentile information would help know where the percentile points are.

[← Reply](#)

○


**Anudeep Devulapally** ([he/him/his](https://iu.instructure.com/courses/2165942/users/6696028))

⋮

2:01pm

You can draw vertical lines to your histogram at the corresponding data values to get an idea of where percentile points like the median, quartiles, 10th percentile, and 90th percentile are.

By adding these vertical lines to your histogram, you can visually estimate where these percentile points are located within the data range.

[← Reply](#) 

○


<https://iu.instructure.com/courses/2165942/users/6692441>

4:11pm

⋮

They can modify a normal histogram to include percentile markers by calculating the desired percentiles for the dataset and adding markers at those positions on the histogram plot. This approach will allow them to not only visualize the data's distribution but also identify specific percentile points, providing a better understanding of the data's distribution beyond the median and quartiles.

Edited by [Yashada Nikam \(https://iu.instructure.com/courses/2165942/users/6692441\)](https://iu.instructure.com/courses/2165942/users/6692441) on Oct 3 at 4:11pm

[← Reply](#) 

○

<https://iu.instructure.com/courses/2165942/users/6762824>

4:49pm

⋮

Yes there is, you would need to code this in but all you would have to do is hard code the percentiles you would like to have in the histogram. If the bin falls into that percentile you can call it out by color or even with hash marks.

[← Reply](#) 

○

<https://iu.instructure.com/courses/2165942/users/6688770>

4:51pm

⋮

Yes, there is a way to modify a normal histogram to approximate the location of percentile points, including the median, quartiles, and other percentiles. You can achieve this by adding

vertical lines or markers to the histogram plot at the positions corresponding to these percentiles. Here's how we can do it:

1. **Calculate Percentiles:** First, calculate the values of the percentiles you want to display (e.g., median, quartiles, 10th percentile, 90th percentile) from your dataset.
2. **Create the Histogram:** Create the histogram as you normally would, using your data and choosing an appropriate number of bins. This will show the data distribution.
3. **Add Vertical Lines or Markers:** Add vertical lines or markers at the positions corresponding to the calculated percentile values. You can use Matplotlib or a similar library to do this.

eg) In python

```
import numpy as np
import matplotlib.pyplot as plt

# Sample data
data = np.random.normal(0, 1, 1000)

# Calculate percentiles
percentiles = np.percentile(data, [10, 25, 50, 75, 90])

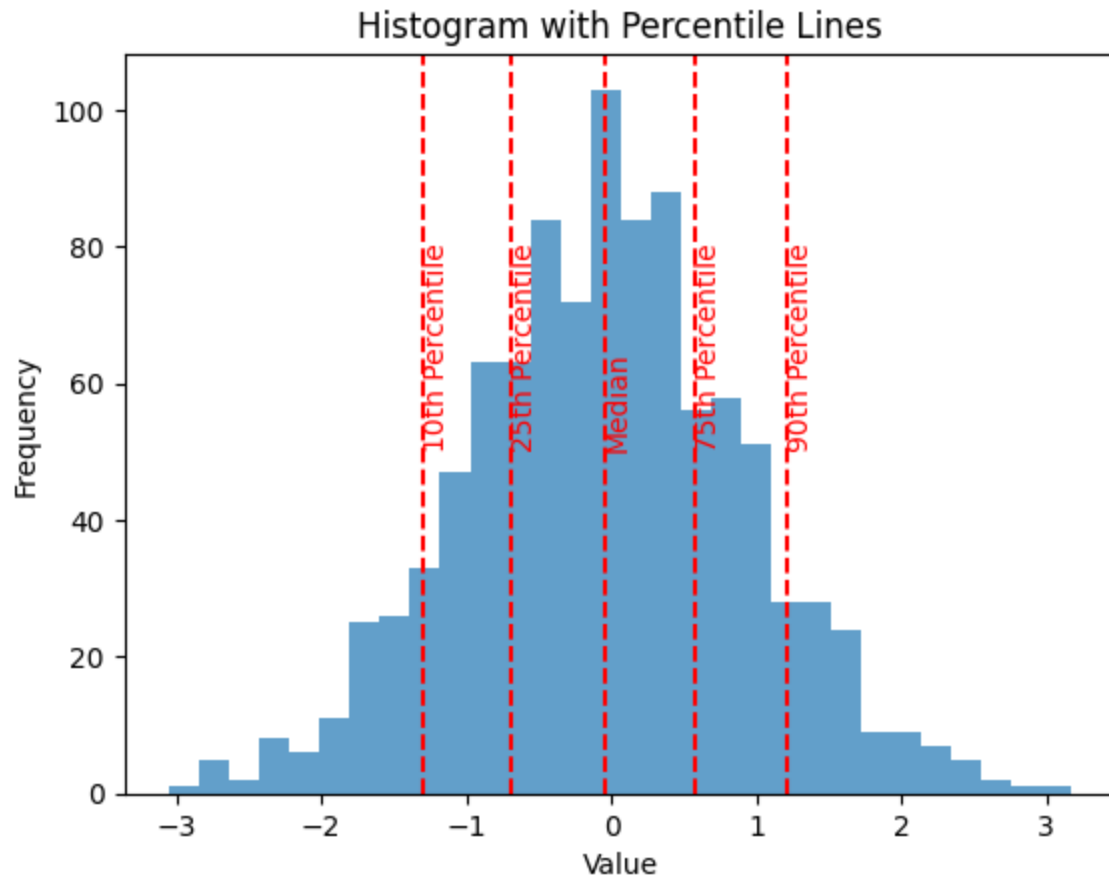
# Create the histogram
plt.hist(data, bins=30, alpha=0.7)

# Add vertical lines for percentiles
for percentile in percentiles:
    plt.axvline(x=percentile, color='red', linestyle='--')

# Label the lines
plt.text(percentiles[0], 50, '10th Percentile', rotation=90, va='bottom', color='red')
plt.text(percentiles[1], 50, '25th Percentile', rotation=90, va='bottom', color='red')
plt.text(percentiles[2], 50, 'Median', rotation=90, va='bottom', color='red')
plt.text(percentiles[3], 50, '75th Percentile', rotation=90, va='bottom', color='red')
plt.text(percentiles[4], 50, '90th Percentile', rotation=90, va='bottom', color='red')

# Show the plot
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.title('Histogram with Percentile Lines')
plt.show()
```

This will generate picture like:



← Reply 👍