⋮

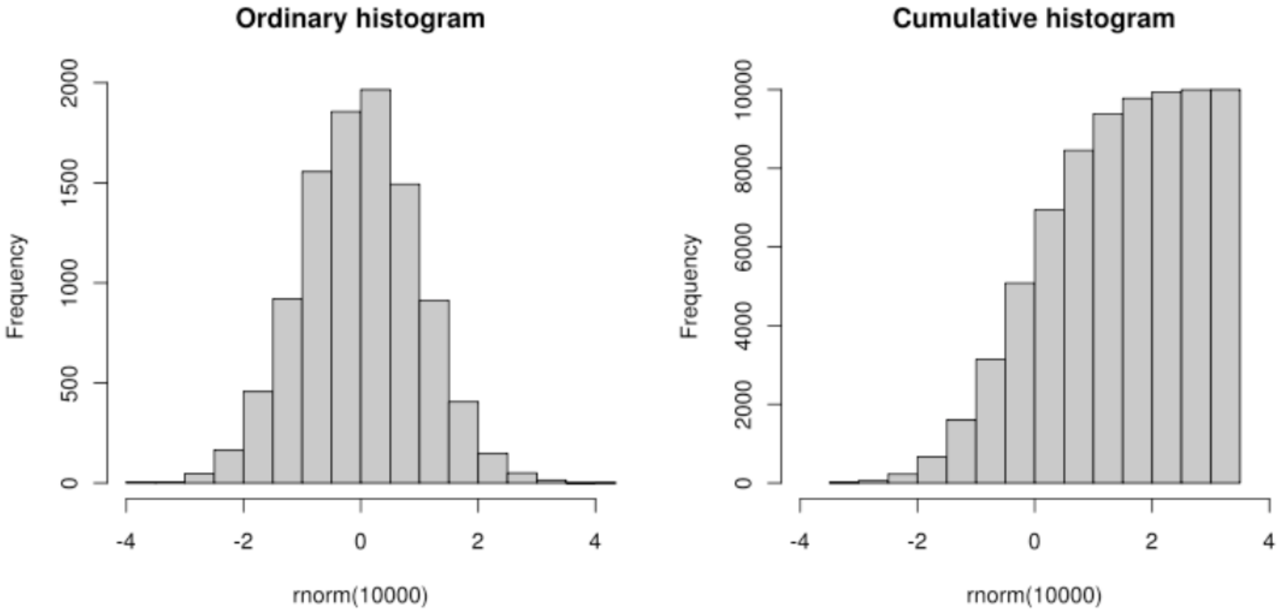
This is a graded discussion: 10 points possible

due -

34 38

Can you improve it even more?

Ok, there is a simple way to modify the normal histogram to display information about percentile values. It's called "**cumulative histogram**". The idea is very simple. The first (left-most) bar is the same in both cases, but from the second bar, you display the **total frequency up to this bin** rather than the frequency within this bin. It looks like the following:



From here, you can normalize the y-axis to [0, 1] if you want to get information about percentiles. If you want to know where the median lies, draw a horizontal line from 0.5, and see which bar meets this horizontal line. The bin of the bar is where the median is located. You can do the same thing with all other percentile values.

Ok that's great. But **can you improve this even further?** This is still not optimal because the resolution is limited by the size of each bin. Data is aggregated based on the fixed bins and if a bin contains lots of data points, you can't figure out what's going on within the bin. So my question is the following: **can you improve this cumulative histogram to have maximum resolution that the data can provide?** Again the goal is to show where each percentile point lies.

Search entries or author Unread ✓ Subscribed

Reply

○

https:/

Sneha Satish (https://iu.instructure.com/courses/2165942/users/6679606)

Sep 25, 2023

As you said, normalizing the y-axis to [0,1] would make it easier to read off percentile values. Additionally, you can calculate cumulative counts, interpolate the percentiles and display the percentiles. By using percentile interpolation, you can achieve higher resolution within each bin and get a more accurate representation of where specific percentiles lie within the data distribution. This way, we can gain more insights into the data distribution with greater precision, especially when dealing with large datasets or data that is not evenly distributed across bins.

Reply

○

https:/

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

Sep 25, 2023

To ensure the maximum resolution and to get as much information out of the cumulative histogram graph as possible you could put on top of the histogram chart a line graph depicting the individual cumulative growth. This will give you the most out of the data while still having an overarching view of what the histogram is showing. This could look messy with charts overlapping but could be a short term solution.

Reply

○

https:/

Dustin Cole (https://iu.instructure.com/courses/2165942/users/6701715)

Sep 26, 2023

I think you would have to do a density curve instead of a histogram. Or you could use more bins in areas where the percentage changes to ensure all value to the left on the x axis are within the stated percentile of the y axis.

Reply

○

(https:/

Yumeng Liang (https://iu.instructure.com/courses/2165942/users/6587577)

Wednesday

Increase the number of bins in the histogram to match the granularity of the data. I believe more bins will provide higher resolution and better reveal the data's distribution?

Reply

○

(https:/

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

Thursday

I would recommend leveraging an empirical cumulative distribution function, which has cumulative percentile along the y axis, and can be easily drawn without leveraging bins.

Reply

○

(https:/

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

Thursday

You can use a smoothed out line to visualize both the ordinary and cumulative histograms. After that you can have labeled dotted horizontal lines for each percentile over the chart. Having them labeled with their value makes it easy to see which percentile point is what value.

Reply

○

(https:/

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

Friday

We can increase the number of bins to have maximum resolution.

Reply

○

(https:/

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

Friday

My solution is the way I answered the question first. The method is to draw a cumulative line on the original histogram without the bins fixed. By doing this, it seems possible to check the percentile without distorting the original data form.

Reply

○

https:/

Shantanu Dixit (https://iu.instructure.com/courses/2165942/users/6684610)

Friday

To show percentiles on a histogram, I'd use the Empirical Cumulative Distribution Function. I would sort the data, plot each data point with its position as the y-value, and then I can easily read percentiles directly from the y-axis.

Reply

○

https:/

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

Friday

I think I can use a heat map on the cumulative histogram to show results better.

Also, another of my idea is drawing vertical lines of each 10th of quartiles on the cumulative histogram.

Reply

○

https:/

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

Saturday

you could do fit a curve on the cumulative histogram to indicate cumulative growth , also you could just mark Q1,Q2,Q3 inorder to get an idea about the other quantiles aswell

Edited by Shreedeep Sadasivan Nair (https://iu.instructure.com/courses/2165942/users/6813278) on Sep 30 at 12:33am

Reply

○

https:/

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

Saturday

The cumulative histogram you mentioned is a good way to get an idea of percentiles in a dataset, but it still relies on fixed bins, which can limit the detail you can see within each bin. To improve this and get the maximum resolution, we can use a technique called kernel density estimation (KDE). Instead of using fixed bins, KDE provides a smooth estimate of how the data is distributed. It creates a continuous curve that approximates the shape of the data. With KDE, we create a smoother representation of data distribution, allowing us to bypass fixed bins. This approach yields a cumulative density function (CDF), where we can easily locate specific percentiles like the median. By referencing the CDF, we can pinpoint any desired percentile's precise position, enhancing our data analysis. For example, to find the median, we locate where the CDF crosses the 0.5 mark on the y-axis. The x-coordinate at that point is the median.

Reply

○

https:/

Mothi Gowtham Ashok Kumar (*he/him/his*) (https://iu.instructure.com/courses/2165942/users/6683278)

Saturday

Yes, it is possible to improve the resolution of a cumulative histogram to have maximum resolution that the data can provide.

One way to do this is to use a **step cumulative histogram**. A step cumulative histogram is similar to a regular cumulative histogram, but the bars are drawn as steps instead of solids. This allows the histogram to show more detail about the data distribution.

To create a step 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.
3. Plot the cumulative frequencies on a graph, using a step line instead of a solid line.

Another way to improve the resolution of a cumulative histogram is to use a **kernel density estimate (KDE)**. A KDE is a non-parametric method for estimating the probability density function of a random variable.

To create a KDE plot, you can use the following steps:

1. Choose a kernel function. A common kernel function is the Gaussian kernel.
2. Calculate the KDE value for each value in the data set.
3. Plot the KDE values on a graph.

Reply

○

https:/

Vedant Tapadia (https://iu.instructure.com/courses/2165942/users/6678810)

Saturday

To improve the resolution we an simply plot the CDF as a points and see the exact position instead of bars. This will not give us a histogram exactly but will make sure that we can see all the points clearly.

Reply

○

https:/

Andrea Chung (https://iu.instructure.com/courses/2165942/users/6443321)

Saturday

To improve cumulative histogram, we can add corresponding box plots to show have maximum resolution that data can provide.

Reply

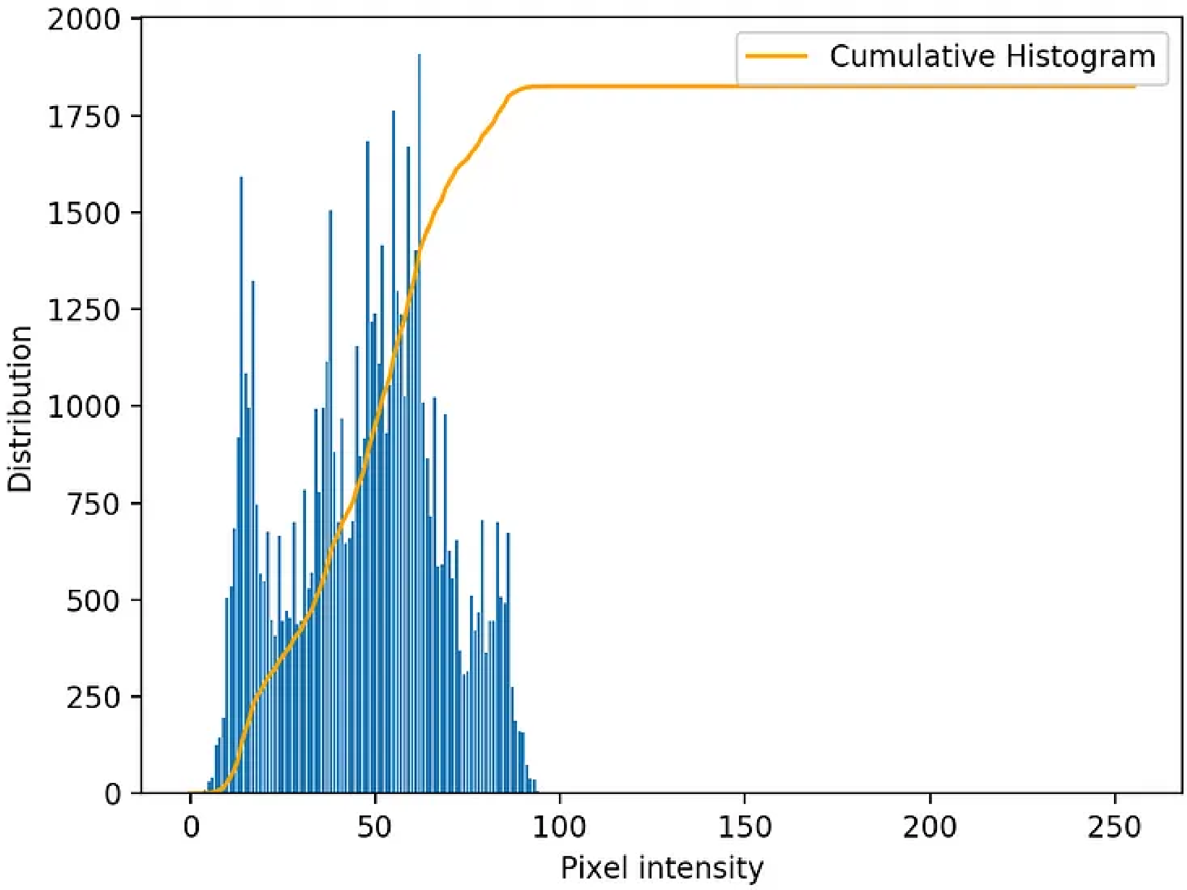
○

https:/

Robert Perez (*he/him/his*) (https://iu.instructure.com/courses/2165942/users/6701521)

Saturday

How about we combine the two? If we plot the cumulative histogram as a line in a color that stands out against the ordinary histogram, won't that give us the best of both worlds? Here's an example of what I am describing:



(Source: **https://levelup.gitconnected.com/introduction-to-histogram-equalization-for-digital-image-enhancement-420696db9e43** (https://levelup.gitconnected.com/introduction-to-histogram-equalization-for-digital-image-enhancement-420696db9e43))

I would further refine the above example to show the quartile percentages on the right y-axis and faint gridlines so it would be easier to pick out the percentile you're interested in seeing.

Reply

○

(https:/

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

Saturday

To improve upon the cumulative histogram and provide maximum resolution, we can make use of the Empirical Cumulative Distribution Function (ECDF). Unlike the cumulative histogram which bins data, the ECDF represents each unique data point, offering the highest possible resolution.

Steps for using ECDF:

1. Sort the Data: Arrange all data points in ascending order.
2. Plot the ECDF: For each data point, plot its value (x-coordinate) against the proportion of data points less than or equal to it (y-coordinate).

Benefits of ECDF:

1. No Binning: Avoids issues related to bin width or origin.
2. High Resolution: Each data point is represented.
3. Easy Percentile Visualization: For instance, to find the median, draw a horizontal line at y=0.5 and note the x-value.

Using the ECDF, we can effortlessly visualize where each percentile lies in your data.

Edited by **Mukul Gharpure** **(https://iu.instructure.com/courses/2165942/users/6678592)** on Sep 30 at 6:49pm

Reply

○

(https:/

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

Sunday

We can use a kernel density estimate (KDE) with a configurable bandwidth to enhance the cumulative histogram and give it the highest resolution the data allow.

A KDE with variable bandwidth is one in which the bandwidth is altered according to the local density of the data for each data point. In locations with dense data, the bandwidth will be reduced, whereas in areas with sparse data, it will be larger. As a result, a KDE is created that is better able to precisely reflect the specifics of the data distribution.

Reply

○

(https:/

Madhuri Patibandla (*she/her/hers*) **(https://iu.instructure.com/courses/2165942/users/6760559)**

Sunday

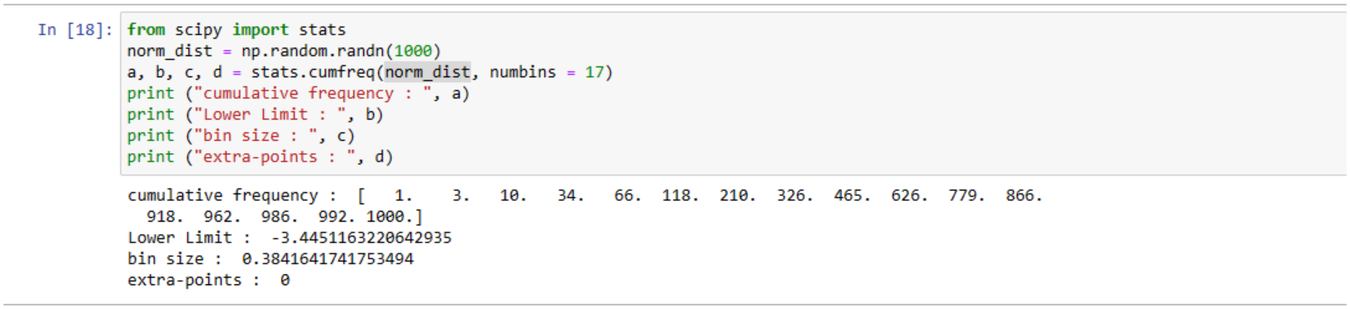
A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified Bin. We can improve the maximum resolution for the data we can provide and verify the percentiles else we can add the cumulative line on the histogram.

Cumulative Frequency values.

median : 465

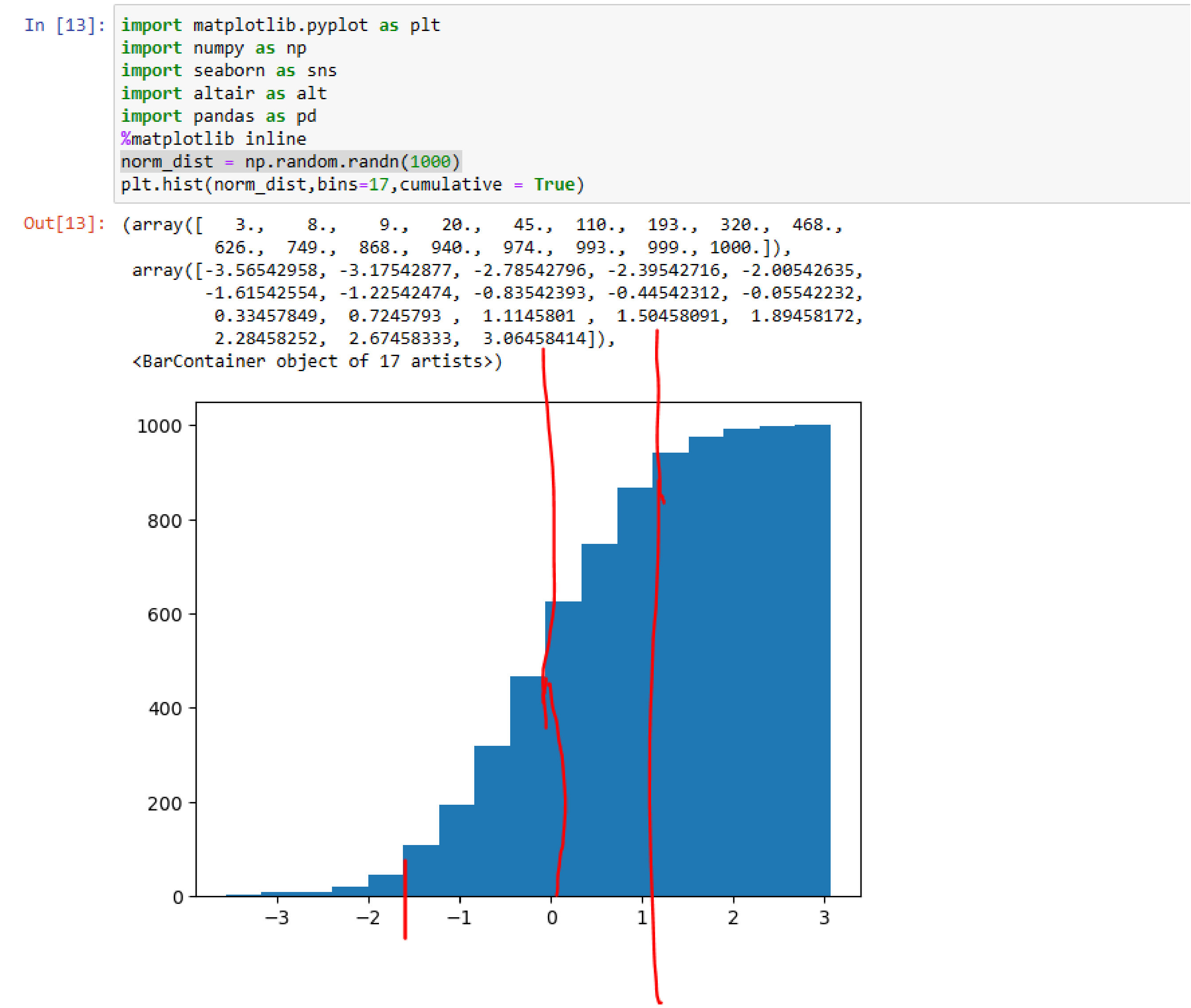
25th Quartile is : 66

3td quartile : 918.



```
In [18]: from scipy import stats
         norm_dist = np.random.randn(1000)
         a, b, c, d = stats.cumfreq(norm_dist, numbins = 17)
         print ("cumulative frequency : ", a)
         print ("Lower Limit : ", b)
         print ("bin size : ", c)
         print ("extra-points : ", d)

         cumulative frequency :  [   1.    3.   10.   34.   66.  118.  210.  326.  465.  626.  779.  866.
           918.  962.  986.  992. 1000.]
         Lower Limit :  -3.4451163220642935
         bin size :  0.3841641741753494
         extra-points :  0
```



```
In [13]: 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(1000)
         plt.hist(norm_dist,bins=17,cumulative = True)

Out[13]: (array([   3.,    8.,    9.,   20.,   45.,  110.,  193.,  320.,  468.,
                  626.,  749.,  868.,  940.,  974.,  993.,  999., 1000.]),
          array([-3.56542958, -3.17542877, -2.78542796, -2.39542716, -2.00542635,
                 -1.61542554, -1.22542474, -0.83542393, -0.44542312, -0.05542232,
                  0.33457849,  0.7245793 ,  1.1145801 ,  1.50458091,  1.89458172,
                  2.28458252,  2.67458333,  3.06458414]),
          <BarContainer object of 17 artists>)
```

Edited by **Madhuri Patibandla (https://iu.instructure.com/courses/2165942/users/6760559)** on Oct 1 at 1:45am

Reply

https:/

Jash Shah (https://iu.instructure.com/courses/2165942/users/6684840)

Sunday

Although the cumulative histogram described is a useful tool for understanding percentiles in a dataset, it still uses fixed bins, which can constrict the amount of detail you can see within each bin. Using a method known as kernel density estimation (KDE), we may enhance this and obtain the highest resolution. KDE offers a smooth estimate of how the data is distributed rather than utilizing preset bins. It produces a continuous curve that closely resembles the data's form. We can avoid fixed bins by creating a smoother depiction of the data distribution with KDE. With the use of the cumulative density function (CDF) produced by this method, we may quickly find particular percentiles like the median.

Reply

https:/

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

Sunday

I suppose at this point you directly draw each data point and which percentile it falls into onto the graph? That way you would end up with a smooth line eventually (with enough data points) or could simply connect the line between each dot for the data point. By plotting each data point directly, you would then end up with a curve ranging from 0 - 1 (eg which percentile it is) rather than the choppiness of the histogram bins.

Reply

https:/

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

Sunday

Kernal Density Estimation Plot or
np.percentile()

Edited by Hymavathi Gummudala (https://iu.instructure.com/courses/2165942/users/6679250) on Oct 1 at 12:46pm

Reply

https:/

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

Sunday

I would add vertical lines to indicate where the percentile points lie. This would directly indicate to the viewer where these percentile points are located instead of relying on the perception of differences in the heights of the bars.

Reply

○

https:/

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

Sunday

Maybe you could also insert a vertical line that intersects the horizontal percentile line at the histogram point, and then color the section of the histogram to the left of that vertical line a different color? At the next point percentile point you would use a new color that ended when you got to the previous percentile's color.

Reply

○

https:/

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

Sunday

I am trying to think about using points for every value instead of using bins so that the size of each bin won't affect the result in the figure. Or we could set the size of bin smaller enough to ensure they cannot contain lots of data point.

Reply

○

https:/

Adam Hume (https://iu.instructure.com/courses/2165942/users/6056428)

Yesterday

I think we can modify this diagram by using a ECDF plot. This will give a smooth representation of the cumulative distribution and allows for a better determination of percentiles values without being constrained by fix sized bins. This type of graph takes into account the total number of data points and allows the viewer to easily identify where specific percentiles lie by looking at the y axis values that correspond to the percentile.

Reply

○

(https:/

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

Yesterday

One way to improve this cumulative histogram to have maximum resolution is to normalize the data points. This should take care of the outliers and allows for ease calculation of the cumulative histogram.

Reply

○

(https:/

Shubham Agarwal (https://iu.instructure.com/courses/2165942/users/6682743)

Yesterday

We can do that by defining custom bins so that each interval on the x-axis has similar frequency

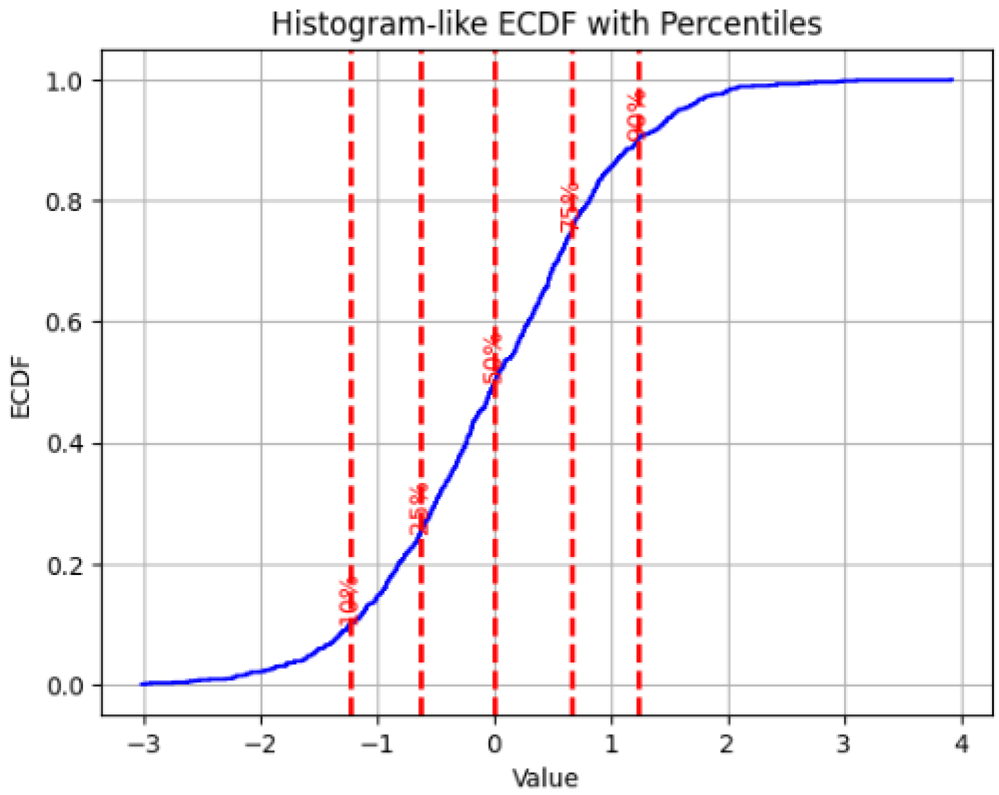
Reply

○

(https:/

Vaibhav Piyushkumar Lodhiya (https://iu.instructure.com/courses/2165942/users/6694681)

Yesterday



To improve the resolution of the cumulative histogram and show the exact location of percentile points, you can use an "empirical cumulative distribution function (ECDF)" plot. An ECDF provides a step function that increases by 1/N at each data point, where N is the total number of data points.

Reply

○

(https:/

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

Yesterday

Perhaps the simplest way is to find the cumulative points that equal your percentile of interest and add those to the graph. I'd almost want to either add red vertical lines to the appropriate places or perhaps shade over the graph with a series of colors, which would change depending on what you're trying to convey.

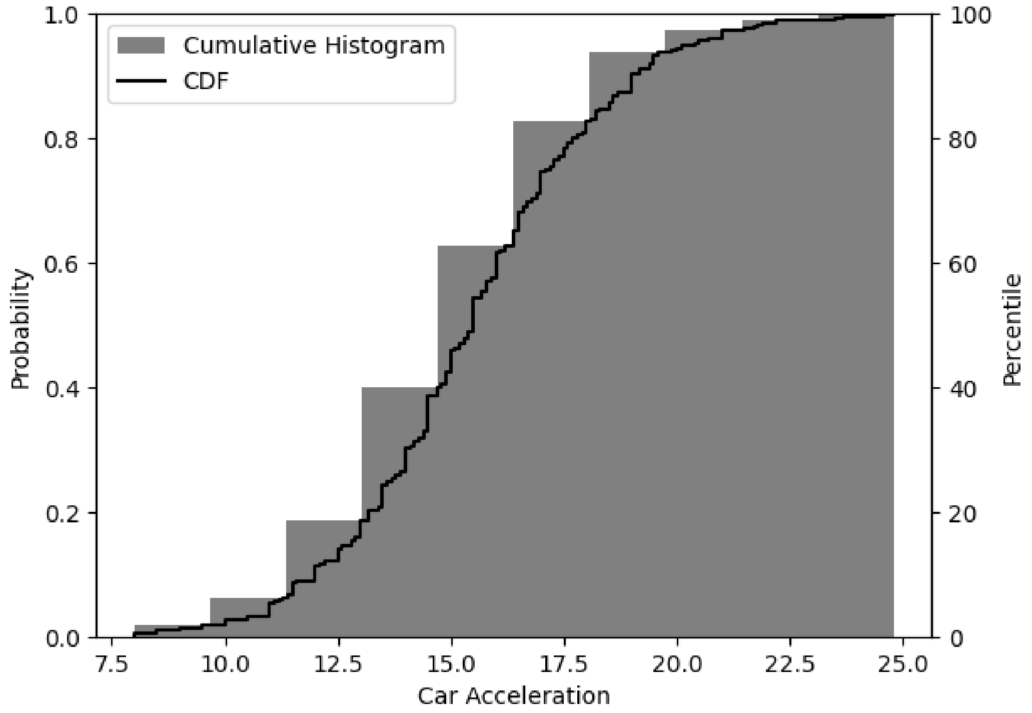
Reply

○

(https:// [Gary Croke (https://iu.instructure.com/courses/2165942/users/6706306)](https://iu.instructure.com/courses/2165942/users/6706306)

Yesterday

One further improvement might be to overlay the CDF onto the cumulative histogram.



[m08_histogram_percentiles-1.html (https://iu.instructure.com/files/163088363/download?download_frd=1&verifier=UOWflfhi7BjkhBcOg7TjAMOAljR5K1E69hMDMLgz)](https://iu.instructure.com/files/163088363/download?download_frd=1&verifier=UOWflfhi7BjkhBcOg7TjAMOAljR5K1E69hMDMLgz)

Reply

○

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

Yesterday

We can use **Empirical Cumulative Distribution Function (**ECDF) plot to represent data values and the cumulative proportion. Also by adding percentile lines we will get accurate percentiles within the data set.

Reply

○

(https:// [Ram Kiran Devireddy (https://iu.instructure.com/courses/2165942/users/6677399)](https://iu.instructure.com/courses/2165942/users/6677399)

12:04am

That's, interesting. The idea of cumulative histogram really simplifies locating the specific percentiles. When reaching about this, I came across the Empirical Cumulative Distribution Function (ECDF) plot, which can probably improve this! The ECDF plot offers maximum resolution and allows us to precisely identify the location of each percentile point within the dataset.

Reply 👍

○

https:/

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

1:26pm

Since it is a cumulative histogram we can try to split the size of the bins only for the cumulative histogram.

Reply 👍

○

https:/

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

3:01pm

Yes, there are a few methods that can improve a cumulative histogram's resolution to its highest level. Kernel density estimation (KDE) is one approach. KDE is a non-parametric technique for calculating a random variable's probability density function. To estimate the probability density function of the data, it uses a kernel function rather than using bins at all. A extremely smooth and high-resolution cumulative histogram may result from this.

To create a cumulative histogram using KDE, we can use the following steps:

1. Estimate the probability density function of the data using KDE.
2. Calculate the cumulative probability function of the estimated probability density function.
3. Plot the cumulative probability function against the values of the data.

Reply 👍

○

https:/

Yashada Nikam (*she/her/hers*) (https://iu.instructure.com/courses/2165942/users/6692441)

4:14pm

The cumulative histogram can be improved to have maximum resolution by using the Empirical Cumulative Distribution Function (ECDF) plot. This plot displays each data point's cumulative percentile position, allowing for a detailed view of the data distribution. This can precisely determine the location of specific percentiles, such as the median or quartiles, with high accuracy.

Reply

○

https:/

David Rosenthal (https://iu.instructure.com/courses/2165942/users/6762824)

4:57pm

I believe that if you code in the percentile this would help, otherwise you may be able to break up the x axis to be you percentiles instead of the numbers. There are better charts out there that would show you this percentile and give a more accurate picture. Calculating the ECDF would also be a helpful way to accomplish this. You would overlay the ECDF and label the percentiles as desired.

Reply

○

https:/

Sarthak Vivek Chawathe (*he/him/his*) (https://iu.instructure.com/courses/2165942/users/6688770)

4:59pm

We can further improve the representation of percentile values using a cumulative histogram with maximum resolution. One way to achieve this is by using a "kernel density estimate (KDE) cumulative histogram" or "cumulative distribution function (CDF) estimate." This method provides a smooth, continuous estimate of the cumulative distribution, allowing for a more accurate determination of percentile points.

In Python

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

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

# Create a KDE cumulative histogram
sns.histplot(data, stat="density", cumulative=True, common_norm=False, kde_kws={"cumulativ
e": True})
```

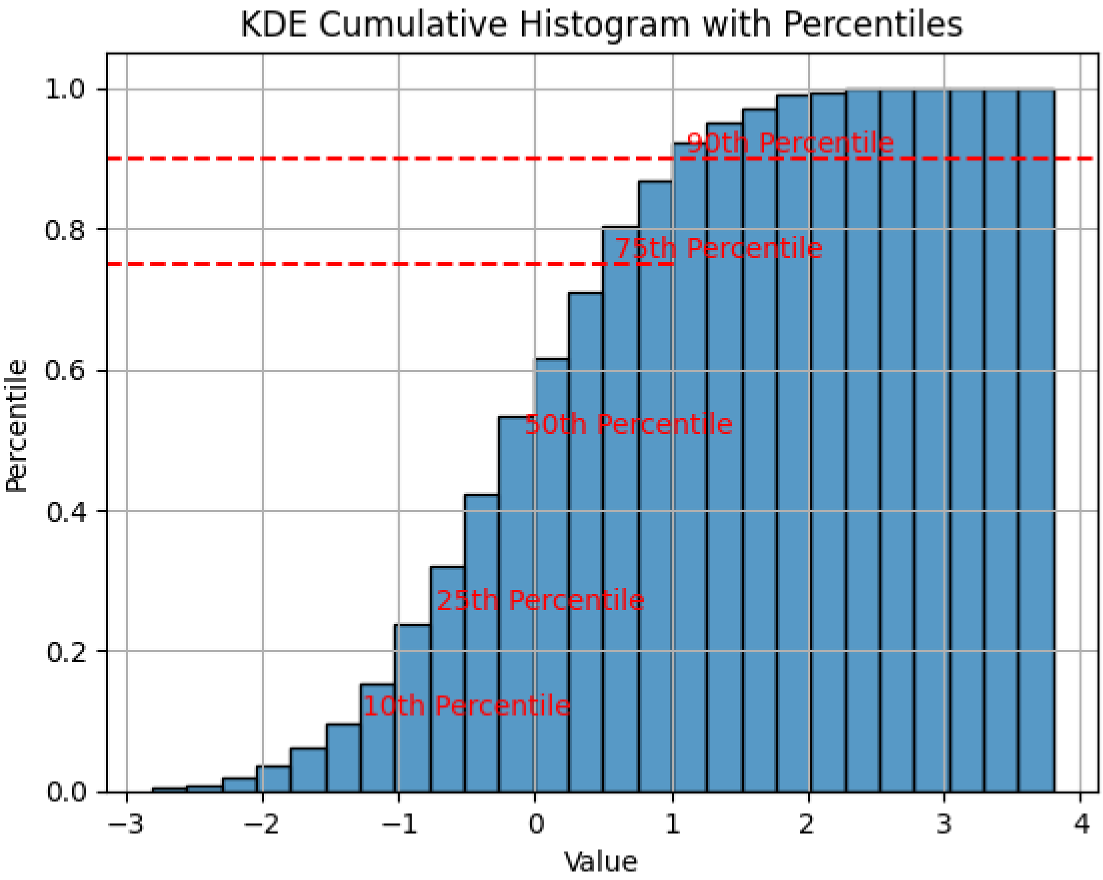
```
# Calculate and plot horizontal lines for percentile values
percentiles = [10, 25, 50, 75, 90]
percentile_values = np.percentile(data, percentiles)

for i, percentile in enumerate(percentiles):
 plt.axhline(y=percentile / 100, color='red', linestyle='--', xmax=percentile_values[i])

# Label the lines
for i, percentile in enumerate(percentiles):
 plt.text(percentile_values[i], percentile / 100, f'{percentile}th Percentile', color='re
d', va='bottom')

# Set x-axis and y-axis labels
plt.xlabel('Value')
plt.ylabel('Percentile')
plt.title('KDE Cumulative Histogram with Percentiles')

# Show the plot
plt.grid(True)
plt.show()
```



Reply