


distribution types. When this occurs, one of your options is to use the Empirical Cumulative Distribution Function (ECDF) to use historical data to predict future data values. This historical data could come from an expert opinion, weighted existing data, or computer generated values in order to inform decisions going forward. It is assumed that the events are independent and the sum of the probabilities is 1.

The ECDF is calculated by ordering all the unique observations in the data sample and calculating the cumulative probability for each as the number of observations less than or equal to a given observation divided by the total number of observations.


In this try-it activity, you will use Python and Colab to analyze a sample dataset, specifically, the ECDF against the generated data. First let us start by creating a non-normally distributed data set. We can do this by creating a two-peak dataset by combining two normal distributions.

To complete this activity, please follow the instructions below:

1. Click the link: **Try-It Activity 3.1**  (https://mo-pcco.s3.us-east-1.amazonaws.com/MO-PCDS/module3/activity-name-1_starter.zip)
2. You will be redirected to a new tab where a corresponding .zip file will download.
3. Unzip the .zip file to extract the activity's files.
4. Finally, navigate to <https://colab.research.google.com> (<https://colab.research.google.com/>) and click the upload tab.
5. Select the file you downloaded and want to begin. You can now view the code.
For more information on this step, please see the **Google Colab instructions** (https://classroom.emeritus.org/courses/9054/pages/google-colab?module_item_id=1506888) from the Program Tools section.

Once you've gone through the Colab notebook, record your observations and share what you've learned about the ECDF. How could you use this tool in the future?

Be sure to read the statements posted by your peers. Engage with them by responding with thoughtful comments and questions to deepen the discussion.

For additional practice, we have included a bonus activity within **Colab**  (https://colab.research.google.com/drive/1iNwgdrqREKPtrHyZr7QtbkyKp_DE17VK?usp=sharing). Keep in mind, this portion of the activity will not be graded.

Suggested time: 20-30 minutes.

Rubric: Try-It Activity 3.1

Criteria	Exceeds expectations	Meets expectations	Below expectations
Thoughtful and complete response to the question(s)	4 pts Fully responds to the question(s), post is supported by connections to the reading and real-life examples, and post makes additional connections to the field of data engineering with novel ideas, critical thinking, or extensive application of how to use the topic in future work.	3 pts Fully responds to the question(s), and post is supported by connections to the content or real-life examples.	0 pts Partially responds to the question(s), or connections to the content are missing or vague.
Engagement with the learning community	2 pts Posts thoughtful questions or novel ideas to multiple peers that generate new ideas and group discussion.	1.5 pts Asks questions or posts thoughtful responses to generate a single peer's response.	0 pts No responses to peers or posts minimal or vague responses to peers that do not motivate a response (e.g., "I agree.").

Search entries or author

Unread

Subscribed

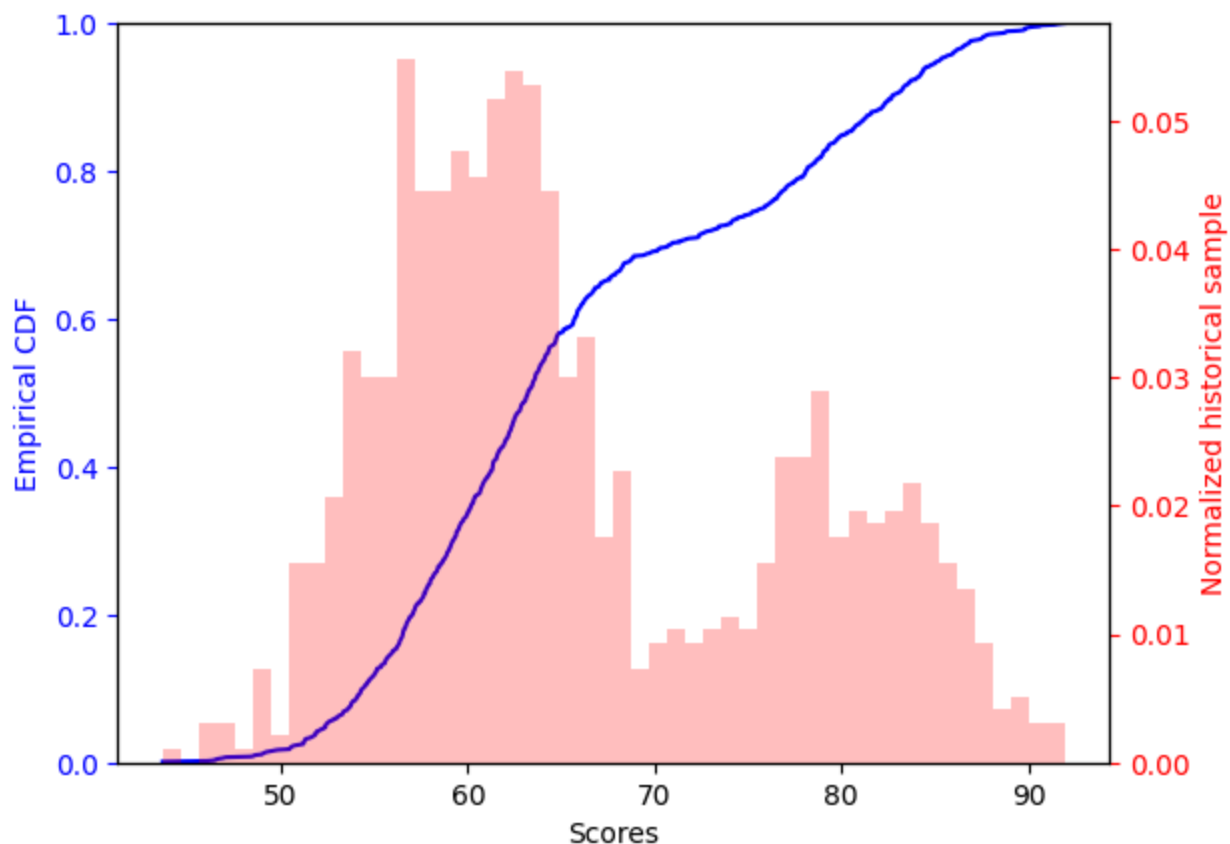
Reply

<https://classroom.emeritus.org/courses/9054/users/228518> **Diego Milanes (He/Him)** <https://classroom.emeritus.org/courses/9054/users/228518>

Apr 10, 2024

A cumulative distribution function (cdf) is a function created as the integral of another function (for instance, a probability density function, pdf) up to a given point. In previous exercises, we have used known pdfs and therefore the cdf will be known either by analytical or numerical integration of the given pdf. However, pdfs are not always well-known and well-behaved functions; instead, they can be sampled from historical data. After normalization (its integral sums up to 1), the empirical cdf can be computed numerically. In both cases, the meaning of the cdf is the same and represents the probability of obtaining the relevant parameter below a given point. For clarity, attached is a plot that shows the pdf and the cdf together for the try-it example.

Fun fact: the cdf never decreases, since you are always adding positive quantities. We are sure those quantities are positive since the probability is always positive.



← Reply 👍 (1 like)



Javier Di (<https://classroom.emeritus.org/courses/9054/users/226884>)

Apr 11, 2024

Great chart Diego. Curious how you put it together with the 2 graphs overlay? Thank you

← Reply 👍



Diego Milanes (He/Him) (<https://classroom.emeritus.org/courses/9054/users/228518>)

Apr 11, 2024

Hi Javier,

The code below does the job

cheers

```
import matplotlib.pyplot as plt
```

```
fig, ax1 = plt.subplots()
```

```
ax1.plot(ecdf.x, ecdf.y, 'b-')
```

```
ax1.set_xlabel('Scores')
```

```
ax1.set_ylabel('Empirical CDF', color='b')
```

```
ax1.set_ylim(0,1)
```

```
ax1.tick_params(axis='y', colors='b')
```

```
ax2 = ax1.twinx()
```

```
ax2.hist(sample, bins=50, color='r', density=True, alpha = 0.25)
```

```
ax2.set_ylabel('Normalized historical sample', color='r')
```

```
ax2.tick_params(axis='y', colors='r')
```

```
plt.show()
```

← Reply 👍 (3 likes)



Turki Alghusoon (<https://classroom.emeritus.org/courses/9054/users/229165>)

Apr 13, 2024

great idea! thank you for sharing the code!

← Reply 👍



Ricardo Anaya (<https://classroom.emeritus.org/courses/9054/users/228915>)

Apr 14, 2024

ditto, great idea and thanks for sharing

 Reply **Swati Sharma** (<https://classroom.emeritus.org/courses/9054/users/236938>)

Apr 15, 2024

thanks for sharing! I'm also diving into cumulative distribution functions (CDFs), and your explanation really helped clarify things for me. I liked how you broke down the process of constructing CDFs from both known PDFs and empirical data. The normalization step you mentioned was particularly useful for understanding how CDFs are interpreted. The plot you included was a nice touch for visualizing the concept. And I didn't know that fun fact about CDFs never decreasing – it adds an interesting perspective for me ! Keep up the great work!

 Reply **Manjari Vellanki** (<https://classroom.emeritus.org/courses/9054/users/231480>)

Apr 10, 2024

Takeaways from Module3.1:

The distribution refers to how the data is spread out or clustered around certain values or ranges. By examining the distribution, we can gain insights into the characteristics and patterns of the data, which can be useful in making informed decisions and predictions.

Sometimes the observations in a collected data sample do not fit any known probability distribution and cannot be easily forced into an existing distribution by data transforms or parameterization of the distribution function. Instead, an empirical probability distribution must be used.

There are two main types of probability distribution functions we may need to sample; they are:

- Probability Density Function (PDF): returns the expected probability for observing a value.
- Cumulative Distribution Function (CDF): CDF returns the expected probability for observing a value less than or equal to a given value.

The Empirical Cumulative Distribution Function is calculated by ordering all of the unique observations in the data sample and calculating the cumulative probability for each as the

number of observations less than or equal to a given observation divided by the total number of observations.

```
ECDF(x) = number of observations <= x / n
```

To demonstrate this, in Try-it activity, a sample data is generated by combining samples from two different normal distributions, 300 examples with a mean of 80 and a standard deviation of five (the larger peak), and 700 examples with a mean of 60 and a standard deviation of five (the smaller peak). The distribution is fit by calling `ECDF()` and passing in the raw data sample (`ecdf = ECDF(sample)`)

The function can be called to calculate the cumulative probability for a given observation:

```
ecdf(94) to calculate  $x < 94$ 
```

```
ecdf(44) to calculate  $x < 44$ 
```

```
ecdf(np.mean(sample)) to calculate  $x < \text{mean}$ 
```

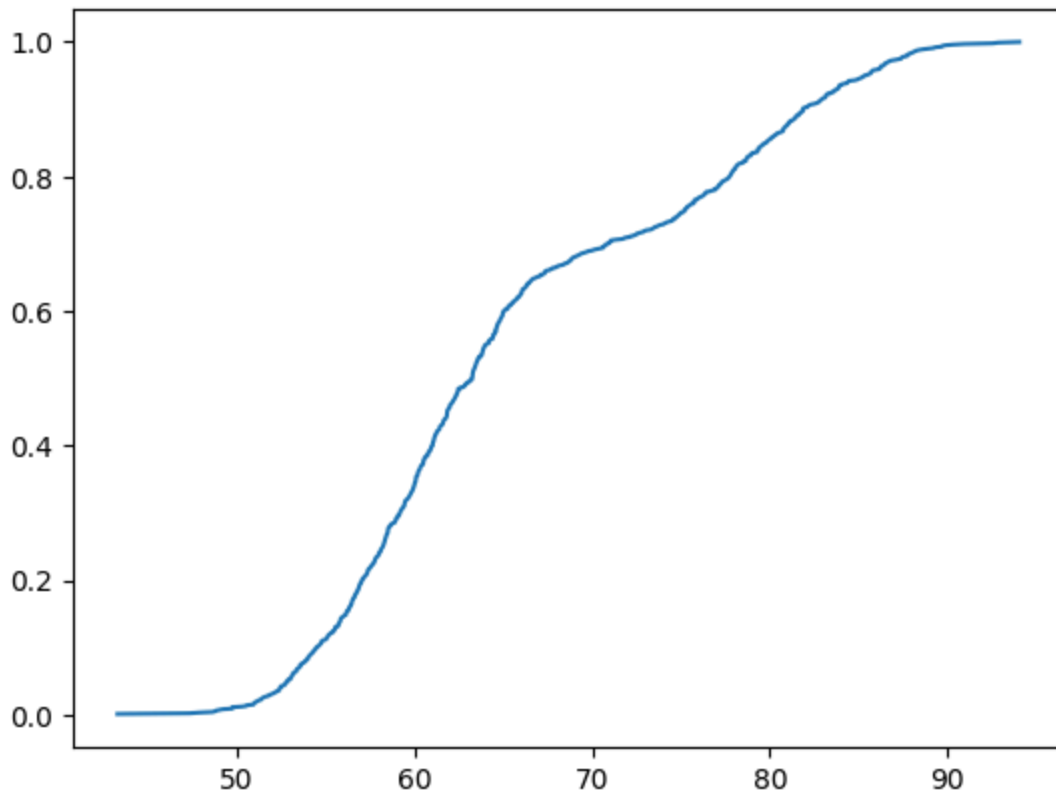
The function can be called to calculate the cumulative probability for observing a value less than or equal to a given value.

```
1-ecdf(94) to calculate  $x > 94$ 
```

```
1-ecdf(75) to calculate  $x > 75$ 
```

The class also provides an ordered list of unique observations in the data (the `.x` attribute) and their associated probabilities (`.y` attribute).

```
pyplot.plot(ecdf.x, ecdf.y)
```



From the plot, we can observe the peaks at means for 2 samples and also to identify the probability of given observation in a combined sample.

In clinical firm, we can use this tool for observational studies where data can't fit in any other distribution and where decisions have to be made based on historic data and predict future outcomes, to compare the outcomes between Treatment vs Placebo , to estimate patient retention etc by collecting related variables.

Edited by [Manjari Vellanki \(https://classroom.emeritus.org/courses/9054/users/231480\)](https://classroom.emeritus.org/courses/9054/users/231480) on Apr 10 at 7:45pm

← Reply 



Javier Di (<https://classroom.emeritus.org/courses/9054/users/226884>)

Apr 11, 2024

Manjari, Very interesting example on the clinical firm and curious how it would work there. Do you need a large sample to map out the Empirical Cumulative Distribution Function and then apply it to predict future data? How would it work in a real life example? Thank you

← Reply 

**Manjari Vellanki** (<https://classroom.emeritus.org/courses/9054/users/231480>)

Apr 11, 2024

Hi Javier-

Yes, definitely we need good collection of data. There is one example of collecting patient data for Clinical Outcome assessments vs Percentage of patients involved in trial by comparing between two models RMM model and eCDF estimates which results the distribution of the COA score changes affected the degree of concordance between RMM and eCDF estimates. The COA score changes from simulated normally distributed data led to greater concordance between the two approaches than did COA score changes from the actual clinical data. The confidence intervals of MWPC estimate based on eCDF methods were much wider than that by RMM methods, and the point estimate of eCDF methods varied noticeably across visits.

[← Reply](#) **Ricardo Anaya** (<https://classroom.emeritus.org/courses/9054/users/228915>)

Apr 14, 2024

great idea on pointing this:

- Probability Density Function (PDF): returns the expected probability for observing a value.
- Cumulative Distribution Function (CDF): CDF returns the expected probability for observing a value less than or equal to a given value

it would be great to have the same example plotted in both ways to compare also

[← Reply](#) **Manjari Vellanki** (<https://classroom.emeritus.org/courses/9054/users/231480>)

Apr 15, 2024

Thanks Ricardo :)

[← Reply](#)



Javier Di (<https://classroom.emeritus.org/courses/9054/users/226884>)

Apr 11, 2024



Takeaways from Module 3.1: ECDF

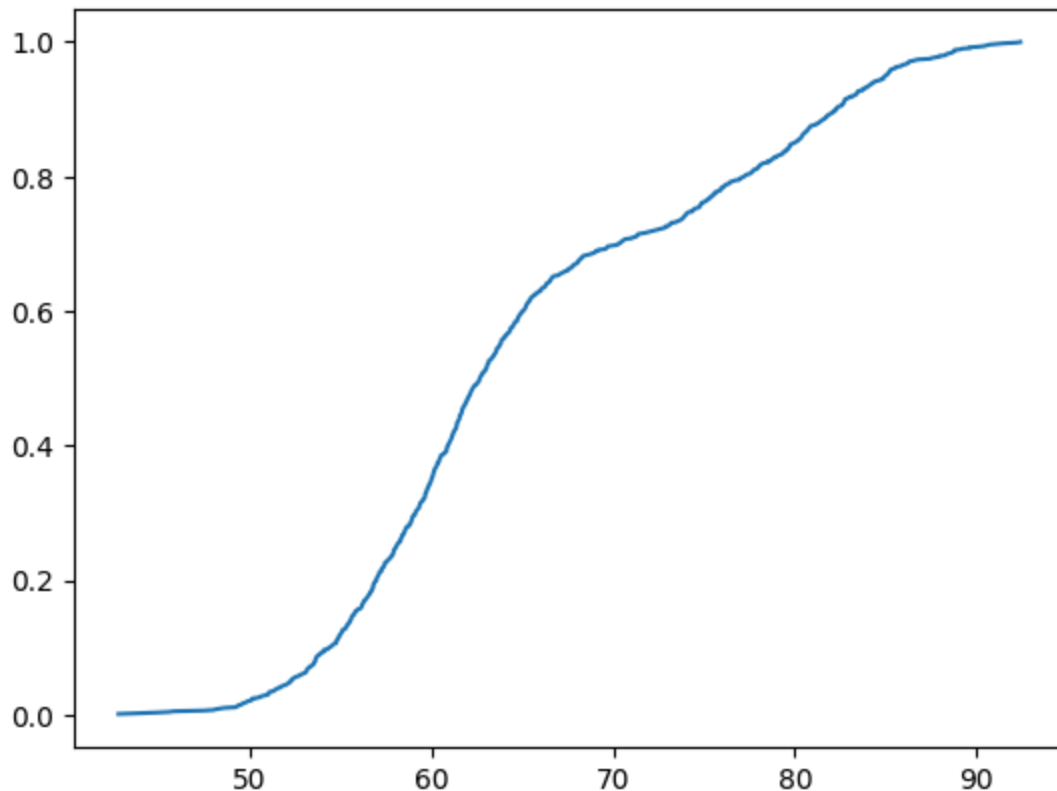
The ECDF is a very powerful tool when having enough data about a particular distribution, to be able to use it to predict future values.

From the Colab example, random data is generated then yields a particular distribution and histogram that doesn't fit to a Normal or Binomial Distribution. The ECDF allows to model the distribution which is not normally distributed for exams scored.

The Empirical Cumulative Distribution Function is calculated by ordering all of the unique observations in the data sample and calculating the cumulative probability for each as the number of observations less than or equal to a given observation divided by the total number of observations

The distribution has a high concentration of values in the 50-70 range with a mean of 65.8 and a 62.6% probability of following below the mean.

The Empirical CDF below shows peaking around the mean in the 60s and then the slope reduces as it goes into higher scores indicating lower probability of reaching those scores.



One application in my experience of this would be in the investing/financial world where the normal distribution assumptions are used frequently to calculate returns. A Q-Q plot could be run to test the normal distribution hypothesis. For example, a Normal Distribution wouldn't fit as

well for Tech volatile stock returns like Apple. This is because Apple will have fat tails and years where the stock is up +100% or more and years in which is down more than 50% and this will not fit a normal distribution and an Empirical CDF is needed.

A normal distribution would do a good job for a broad low volatility index such as the S&P 500 but not for individual volatile stocks.

An Empirical CDF is needed to predict future returns for single name volatile stocks

Edited by [Javier Di \(https://classroom.emeritus.org/courses/9054/users/226884\)](https://classroom.emeritus.org/courses/9054/users/226884) on Apr 11 at 1:22am

← [Reply](#) 



[Diego Milanes \(He/Him\) \(https://classroom.emeritus.org/courses/9054/users/228518\)](https://classroom.emeritus.org/courses/9054/users/228518)

Apr 11, 2024

Hi Javier

Great example. I wonder if there is no historical data to build an empirical CDF, are there analytical models to predict high volatile stock returns?

Thanks

Diego

← [Reply](#) 



[Yossr Hammad \(https://classroom.emeritus.org/courses/9054/users/229118\)](https://classroom.emeritus.org/courses/9054/users/229118)

Apr 12, 2024

Javier,

Thank you for your explanation. it helped me for a better understanding of the ECDF. interesting example of Apple that clearly explains the difference.

greatly appreciate it :)

← [Reply](#) 



[Lee Lanzafame \(https://classroom.emeritus.org/courses/9054/users/231975\)](https://classroom.emeritus.org/courses/9054/users/231975)

Apr 15, 2024

I love practical examples like this. Thanks for brining it to our attention, it wasn't immediately obvious to me that stocks like apple have fat tails and extreme fluctuations and normal distributions wouldn't suffice, good to see that EDCF can still offer a solution.

← Reply 👍

○

[https://](https://classroom.emeritus.org/courses/9054/users/233864)**Haitham Farag** (<https://classroom.emeritus.org/courses/9054/users/233864>)

⋮

Apr 11, 2024

Definition

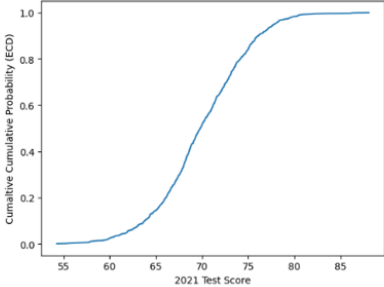
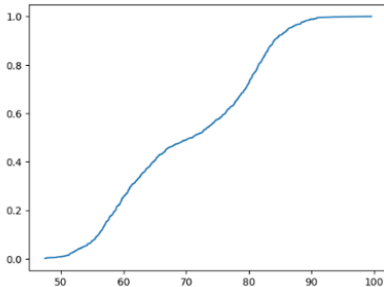
Empirical Cumulative Distribution Function-ECDF provides the cumulative probability for an outcome (observation) and is calculated by dividing *the number of observations equal or less than a given observation* by the total number of *observations*.

Potential Use:

1. Apply the ECDF as a statistical metric, an alternative (e.g. to a histogram) to ensure that data behaves in a way that doesn't completely contradict the assumption (e.g. certain distribution).
2. Use ECDF to construct a more precise predictive model aligned with the summary statistics.

Example

In the 2 scenarios below, I would have made the wrong assumption about the second scenario distribution thinking it is similar to the first., Then erroneously attempted to apply *norm.cdf* where ECDF would be more accurate.

	Scenario 1 (e.g. 2021 test results)	Scenario 2 (e.g. 2022 test results)
Sample creation	1000 random scores (mean 70)	500 random scores (mean 60) & 500 random scores (mean 80)
Summary Stats	<pre> ----- Mean: 69.9732 Max Score 88.0665 Min Score 54.2686 Range: 33.7979 ----- </pre>	<pre> ----- Mean: 70.0032 Max Score 93.9283 Min Score 46.5524 Range: 47.3759 ----- </pre>
Alternative to Assess the Distribution model		

Edited by [Haitham Farag \(https://classroom.emeritus.org/courses/9054/users/233864\)](https://classroom.emeritus.org/courses/9054/users/233864) on Apr 11 at 10:38am

← [Reply](#) 👍 (2 likes)



[Todd Engle \(https://classroom.emeritus.org/courses/9054/users/228910\)](https://classroom.emeritus.org/courses/9054/users/228910)

Apr 13, 2024

Thank you, Haitham for that example. It helped me understand the different interpretations.

← [Reply](#) 👍 (1 like)



[Haitham Farag \(https://classroom.emeritus.org/courses/9054/users/233864\)](https://classroom.emeritus.org/courses/9054/users/233864)

Apr 16, 2024

Thanks Todd for reviewing my response and the kind feedback.

← [Reply](#) 👍



[Turki Alghusoon \(https://classroom.emeritus.org/courses/9054/users/229165\)](https://classroom.emeritus.org/courses/9054/users/229165)

Apr 13, 2024

Hi Haitham,

Great example. It is interesting to see the 2 plots side by side. In fact, the ECDF plot does a better job of incorporating the 2 peaks. I noticed that while the normal distribution

has one single "s" figure, the ECDF looks like 2 connected "s" figures. This could give a quick indication of the number of peaks on the dataset in addition to allowing for statistical inference on non-normal distributions.

← Reply 👍 (1 like)



Haitham Farag (<https://classroom.emeritus.org/courses/9054/users/233864>)

Apr 16, 2024

Thanks Turki for reviewing my response and the feedback. The 2 connected S was also my observation.

Edited by **Haitham Farag** (<https://classroom.emeritus.org/courses/9054/users/233864>) on May 14 at 2:37pm

← Reply 👍



Gustavo Santana (<https://classroom.emeritus.org/courses/9054/users/120927>)

Apr 20, 2024

Thanks for the example Haitham, this really shows how only looking into the mean can be deceiving. Showing the ECDF help to understand te whole scenario.

← Reply 👍 (1 like)



Swati Sharma (<https://classroom.emeritus.org/courses/9054/users/236938>)

Apr 11, 2024

Based on the activity ,here are my observations.

1. We created a data set from two normal datasets. After plotting the data set in an histogram presentation, the distribution does does not look like a normal distribution. (Normal distribution typically has bell shaped curve which is symmetric around the mean and tapers off smoothly towards both ends.)
2. To observe it further more we calculated the Mean, Max value,Min Value, and the range(which is max value - Min value) to understand the data better.
3. We used different labels to find the probabilities associated with certain values or ranges of values in the dataset ($P(x < 94)$: 1 $P(x < 44)$: 0.001 $P(x < \text{Mean})$: 0.617). by using these labels or ranges we understood that there are 61.7% that are under the mean %. that means that there are studend that score above the Mean range.

4. We again used different sets of labels/ranges $P(x > 90)$: 0.007 $P(x > 61)$: 0.258 $P(x < 50)$: 0.02 and observed that only .7% student score more than than maximum value(90) and and majority scored between mean(61) and max value(91)

ECDF: i would use this tool when the underlying dataset is complex or whenever i need to visualize the cumulative probabilities of observed data points like we learnt in our activity today.

← Reply 👍



Ahmad Abu Baker (<https://classroom.emeritus.org/courses/9054/users/234460>)

Apr 12, 2024

Hello Swati,

Thanks for sharing your observations! It's neat to see how your analysis of the ECDF reflects a thorough examination of our dataset. I agree, the histogram's departure from the bell curve typically associated with a normal distribution clearly indicated the bimodal nature of our data. It's these kinds of insights that can lead to a deeper understanding of our subject matter.

Your point about the ECDF being a powerful tool in the face of complex datasets is well-taken. It really does shine when the data doesn't fit into a neat, predefined box. The probabilities you've calculated also tell an interesting story about our data - with only 0.7% scoring above 90, it suggests a high level of difficulty or perhaps that the content wasn't well absorbed.

Considering your findings, it seems like the majority of students are clustered around the mean score, which could indicate a tight central grouping in the data. This can be particularly useful when looking to target interventions or support for students that fall outside this central group.

In the future, I'm thinking about using the ECDF as a means to inform decisions. For example, in a business context, understanding the cumulative percentage of sales below a certain threshold might help in setting benchmarks for sales teams.

It's interesting to think about how such a simple graph can provide such a wealth of information. Are there any specific scenarios you can envision using the ECDF for in your future endeavors?

Cheers!

← Reply 👍

[https://](https://classroom.emeritus.org/courses/9054/users/225803)**Roman Jazmin** (<https://classroom.emeritus.org/courses/9054/users/225803>)

Apr 12, 2024



The Empirical Cumulative Distribution Function (ECDF), is a function that uses observations to create a cumulative distribution.


ECDF is a stepped function that displays the cumulative distribution observed in the given sample. The step function increases by a percentage equal to $1/N$ for each observation in your dataset of N observations. The ECDF's step function shows how a set of observed feature's value approaches 1 on the vertical Y-axis.

An ECDF plot displays data points in a sample from lowest to highest against their vertical percentiles on the Y-axis. The Y-axis represents a percentile scale and the X-axis represents the data values. It's empirical because it represents your observed values and the corresponding data percentiles.

ECDFs allow you to plot a feature of your data from least to greatest, which lets you see the whole feature as if it is distributed across a given sample.

How can I use ECDF in the future? Well given a list of observed factors, graphed in the X-axis, I can tell or at least estimate the probability of its effect to a given observation as whole and how it compares to other observed factors.

An more realistic and practical example is when evaluating the performance, in terms of class grades, for a class population of college students after being given a test. We organized each students ascending order with the student with the lowest score to the far left and the student with the highest grade in the class placed to the far right. Once we finish plotting the grades, we can see and better understand the class performance as a whole.

← Reply  (1 like)

[http](http://classroom.emeritus.org/courses/9054/users/229552)**Roy Nunez** (<https://classroom.emeritus.org/courses/9054/users/229552>)

Apr 14, 2024



Hi Roman,

Thanks for sharing.

When I initially saw your post, I thought why would one need to use ECDF when grades for a class are normally used in examples for a normal distributions especially in large

classes.

As I did some more digging and thinking I found that even if the distribution is a normal one the ECDF can be valuable. It can be used to identify outliers and observe the percentile ranks as you alluded to.

There is also the possibility that the grades are not normally distributed as one would expect and one can go into the analysis with out the assumption of abnormality, improving the intuition on how the data is actually spread.

I would have not thought about this and have probably made normality assumptions if you would have not shared your example. Thanks again!

← Reply 👍 (1 like)

○



Ahmad Abu Baker (<https://classroom.emeritus.org/courses/9054/users/234460>)

Apr 12, 2024

⋮

Hey everyone,

I've been diving into the Empirical Cumulative Distribution Function (ECDF) and it's been quite enlightening. By generating a bimodal dataset from two normal distributions, we get a more complex, real-world scenario where the data isn't simply following a single, neat distribution. The histogram visualization was particularly helpful in seeing how our data is distributed with two clear peaks.

Now, when it comes to the ECDF, it's a great tool for understanding the proportion of scores that fall below a certain threshold. For example, we see that no students scored above 94, which indicates the upper limit of our dataset. On the flip side, the probability of a score less than 44 is practically zero, which could signal that our students have a decent understanding of the material, or perhaps our test doesn't effectively differentiate lower abilities.

A cool thing I learned was how the ECDF can be more informative than basic statistics like the mean or median. It provides a full picture of the distribution, which can be more insightful for understanding the complete range of data. For instance, knowing that around 61.5% of students scored below the mean could lead us to examine our teaching methods or the difficulty of the exam.

Looking forward, I can see the ECDF being a super useful tool in various situations. Like if I'm working on project outcomes or examining user engagement metrics, the ECDF could tell me

exactly what percentage of data points are below a certain level. It's a clear and straightforward method to understand the data without making any assumptions about its underlying distribution.

I'm curious to hear how you all might apply the ECDF in your fields. And for those who've also run through this analysis, did you interpret the results in the same way, or did you notice something different?

Looking forward to your insights!

← Reply 👍 (1 like)



Mariana Flores (<https://classroom.emeritus.org/courses/9054/users/237198>)

Apr 13, 2024

Hi Ahmad, so nice to meet you. Great post, I agree with you - ECDF is a great tool for understanding the proportion of student scores at certain thresholds and for understanding trends without a normal distribution. It is interesting that although the test scores can be from 0 to 100, the range for our sample falls between 92.55 and 44.67.

Examining trends of non-normally distributed data through ECDF is fascinating – thank you for sharing.

← Reply 👍 (1 like)



Yossr Hammad (<https://classroom.emeritus.org/courses/9054/users/229118>)

Apr 12, 2024

Empirical Cumulative distribution Function helps to analyze and visualize the distribution of data.

It provides visual presentation of the dataset and how it is distributed. it helps find the probability of certain dataset that are below or within certain values and that helps understanding the spread of the data. When created 2 data set and plotted them in a histogram didnt fit the normal distribution bell curve shape. For further analysis we calculated the following :

Mean: 65.758

Max Score 93.4866

Min Score 45.7435

Range: 47.7431

further analysis we wanted to see the probability that a student get score below the average 65%,

$P(x < 94)$: 1

$P(x < 44)$: 0

$P(x < 65)$: 0.609

we found there is a high probability of 61% that scores falls below the average of 65%

then we checked the probability that a student's score fall within a certain range, we found the following

$P(x > 90)$: 0.007

$P(x > 75)$: 0.249

$P(x < 50)$: 0.018

$P(x > 90)$: 0.007

$P(x > 60)$: 0.643

$P(x < 60)$: 0.357

which we can conclude the probability that most of the scores fall between 90% and 60% is about 64%.

In the future the model can be used to get insights regarding the distribution of more complex data and identify patterns. It can be used to apply comparison between 2 groups of dataset also to get an accurate range that the data fall within.

Edited by **Yossr Hammad** (<https://classroom.emeritus.org/courses/9054/users/229118>) on Apr 12 at 8:50pm

← **Reply** (1 like)



Mariana Flores (<https://classroom.emeritus.org/courses/9054/users/237198>)

Apr 13, 2024

Hi Yossr, so nice to meet you. Great post, I agree with you, the Empirical Cumulative Distribution Function is a great tool to understand trends of data that do not fit a normal distribution. It is interesting that the majority that although the range of the exam is between 0 and 100, about 64%, of student's scores fall between 60 and 90.

Examining trends of non-normally distributed data through ECDF is fascinating – thank you for sharing.

 [Reply](#) **Roman Jazmin** (<https://classroom.emeritus.org/courses/9054/users/225803>)

Apr 17, 2024

Evening all. I was thinking about another use for ECDF and I believe you can check on the performance of all your stocks you invested in your retirement portfolio. That should be an interesting analysis to determine how soon you can retire and start enjoying your golden year or make changes to your portfolio.

 [Reply](#) **Roy Nunez** (<https://classroom.emeritus.org/courses/9054/users/229552>)

Apr 13, 2024

I learned that ECDF is a great tool to use for a bimodal distribution, a non normal or non binomial distribution. In this example we randomly generated and combined 2 normal distributions combined with two means, 80 and 60, sizes 300 and 700, respectively. The outcome was a bimodal distribution with a combined average score of 65%.

As stated in the example it was notable is that more than 61.5% of scores are below the mean. This is the majority of students. In contrast in a normal distribution, it would normally be expected that half of the students would be scoring below the mean and half above.

The ECDF was also used to analyze other segments of the bimodal distribution and allowed to calculate probabilities for scores greater than 90, greater than 75 and less than 50.

I can use this tool in the future for transactions data and group transactions of all categories together, for trend analysis. I can plot the transaction amounts for each category with the ECDF since they will not be normally distributed. This is a great tool to combined datasets like the normal distribution. One can analyze and test hypothesis with this tool without relying on the assumption that the distribution is a normal distribution. It can be used to compute for probabilities in other segments of the distribution. I also learned that I can be used to detect outliers, by observing steep rises or sudden jumps at the tails, which was not observed in this case.

 [Reply](#)  (1 like)

**Timothy Andrew Ramkissoon** (<https://classroom.emeritus.org/courses/9054/users/226697>)

Apr 16, 2024

Your idea of apply the ECDF to transaction data is great. Transactions across different categories often exhibit non-normal behavior. One thing to remember is that real world data rarely conforms perfectly to theoretical models, and ECDF provides a valuable bridge between theory and practice.

[← Reply](#) **Roy Nunez** (<https://classroom.emeritus.org/courses/9054/users/229552>)

Apr 16, 2024

Thanks Timothy! Yes, the ECDF is a valuable tool that I have never heard of before this course and will be very practical and useful when applying to many scenarios in addition to transactions data moving forward.

[← Reply](#) **Mariana Flores** (<https://classroom.emeritus.org/courses/9054/users/237198>)

Apr 13, 2024

When data is not normally distributed, the Empirical Cumulative Distribution Function (ECDF) is helpful to predict future values based on historical. ECDF holds the assumption that events are independent. By calculating the cumulative probability for each observation as the number less than or equal to a given observation divided by the total number the sum of the probabilities is equal to 1.

For an entire state's benchmark score on their state exam, the scores can range between 0 and 100. In our specific sample, the maximum score was 92.55 and minimum score 44.67. The average score was 66.15. However, there is an above average probability, 60.9%, a random student's score is below the average score.

A few additional interesting insights around probability of students receiving certain test scores:

- There's a 0% probability student's score is below 40
- .1% probability student's score is below 45
- 1.2% probability student's score is below 50
- 11.4% probability student's score is below 55
- 33.4% probability student's score is below 60

- 57.6% probability student's score is below 65
- 68.9% probability student's score is below 70
- 74.4% probability student's score is below 75
- 84.9% probability student's score is below 80
- 95.4% probability student's score is below 85
- 99.6% probability student's score is below 90
- 100% probability student's score is below 95

ECDF could be used in the future to find the probability of student's receiving certain scores or identify where students' scores are likely to fall and have a better idea of how students will score. To compare performance across various years or student cohorts, or identify the percentage of students with a certain score for the state's benchmark score on their state exam.

← Reply 👍



Priscilla Annor-Gyamfi (<https://classroom.emeritus.org/courses/9054/users/226376>)

Apr 16, 2024

Great post Mariana and I love the example you cited with how ECDF could be helpful with analyzing students performance with their scores.

← Reply 👍



Isabella Tockman (<https://classroom.emeritus.org/courses/9054/users/207395>)

Apr 24, 2024

Nice breakdown of ECDF's application in analyzing test scores! Your explanation really nails how it predicts future scores and checks out score distributions across percentiles.

← Reply 👍

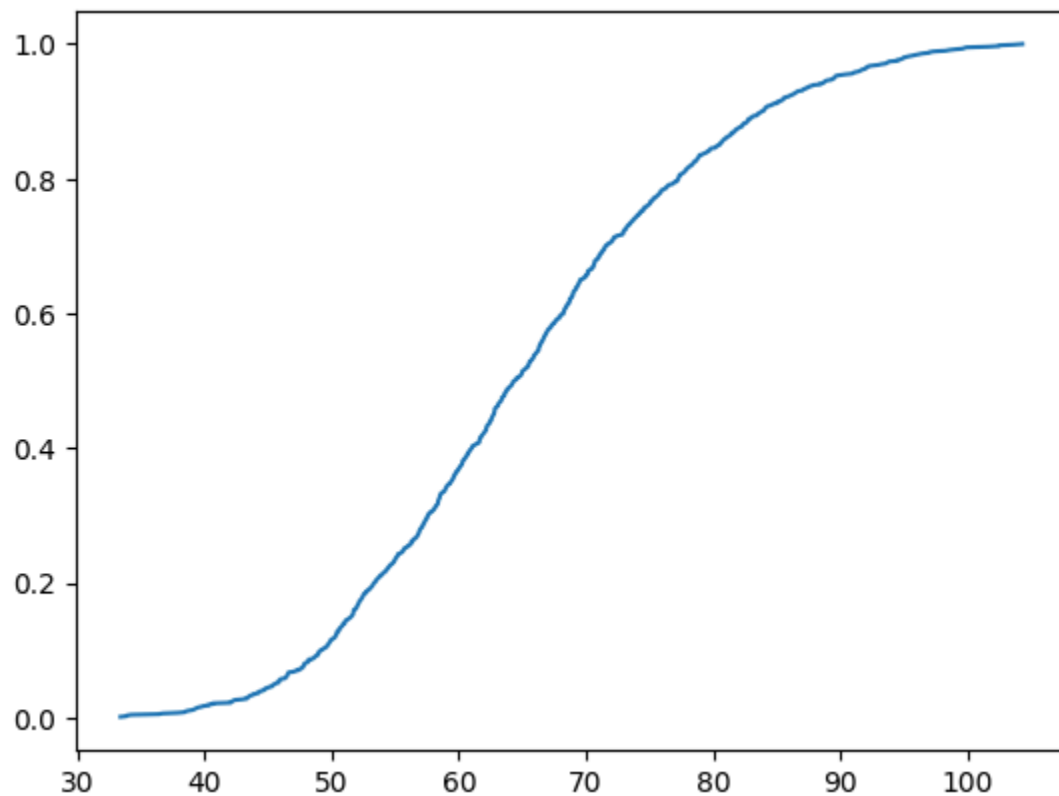
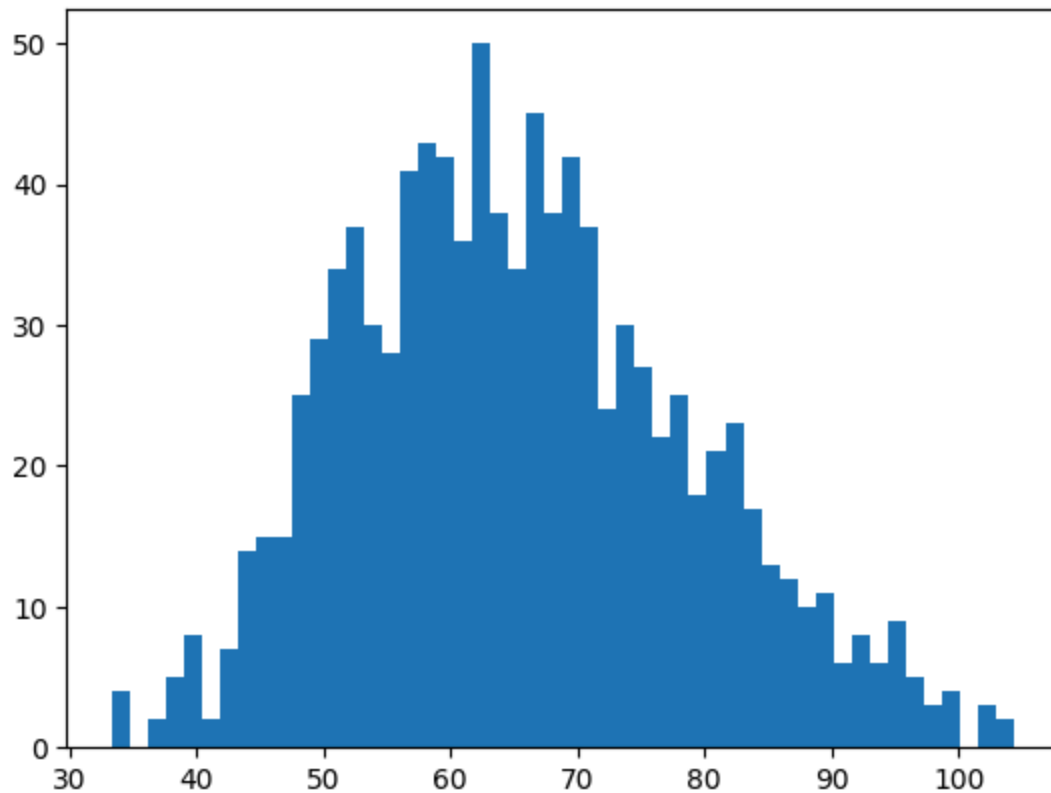


Todd Engle (<https://classroom.emeritus.org/courses/9054/users/228910>)

Apr 13, 2024

Putting the two normal distributions was interesting. I played with the deviation of each (scale=#) and to my expectations as I increased the number the peaks became taller as the deviation shortened, and wider as the deviation increased.

Out of curiosity, I set both at '10' and the ECDF looked similar to a standard deviation chart as both distributions bled into one another. This also allowed the data to go beyond 100. What I don't understand is how we chose what deviation number to use, why was five chosen as a value?



As I was playing around with this, my niece came in and started talking about taking her test at the DMV, and how she needed to schedule an appointment. You could use this model to predict how many people will have to reschedule another DMV appointment based on the probability they would fail their first or second try. This could give the DMV data points to predict scheduling demands around driver testing.

← Reply  (2 likes)



Haitham Farag (<https://classroom.emeritus.org/courses/9054/users/233864>)

May 14, 2024

Good day Todd

I found your observation on increasing the scale to be interesting. I used the code kindly shared by Diego Milanes to plot (attached). Again only looking at the histogram, the mean and the median I would have made the wrong assumption that data followed a normal distribution. While plotting ECDF reveals a bimodel.

Thanks for the insightful question.

Regards

[Dimodel.png \(https://classroom.emeritus.org/files/2598397/download?download_frd=1&verifier=eA2KsumaYtLpf1ccwugO8NnaYyUZ3vx5Z2tvCwBT\)](https://classroom.emeritus.org/files/2598397/download?download_frd=1&verifier=eA2KsumaYtLpf1ccwugO8NnaYyUZ3vx5Z2tvCwBT)

← Reply 



Turki Alghusoon (<https://classroom.emeritus.org/courses/9054/users/229165>)

Apr 13, 2024

Empirical Cumulative Distribution Function (ECDF) is a great option when there is a need to for exploratory data analysis on data that is not normally distributed, such as when data is highly skewed (e.g. wealth distribution), data distribution is bimodal (e.g. average height of humans across genders) or when data distribution is multimodal (e.g. average of age of kids in middle school).

ECDF is also great for getting insights from data without needing to fully understand the data, which could be valuable when time and resources required to fully analyze data are limited, or in cases where additional information around the data is not available. For example, in the

dataset in the 3.2 try it activity, the student data only included the exam scores which had a 2-peak distribution. That meant the data could not be modeled using a normal distribution. Despite that, decision makers needed to get some insights from the data and if they were not aware of the ECDF, they could have relied on general intuition and concluded that half of the students would score below the average, where in fact over 60% percent of student did.

I am curious to see what additional variables could have helped explain the data. It could be that the 2 peaks represented the distribution of scores for groups of students who studied 2 different sets of material or students from 2 different counties. If that is the case, then it might be worthwhile to split the data accordingly and continue to dive deeper into the data. However, it is hard to make such a determination without the additional variables.

I could see myself using ECDF to analyze elapsed time for data sets containing different types of projects and determine the percentage of projects that would finish in less than 90 days, regardless of the project type.

← Reply 



Dawn Prewett (<https://classroom.emeritus.org/courses/9054/users/233112>)

Apr 14, 2024

I am also very curious to see what additional variables could have helped explain the data. Splitting the data into separate datasets is an interesting idea - though I wonder that would look like. If you split the dataset into two based on ECDF, then analyzed them, how would we reintegrate the data back together to all us to see the full picture? While I found ECDF really interesting and insightful, I feel like there is definitely more information needed to make ready use of this tool.

← Reply 



Turki Alghusoon (<https://classroom.emeritus.org/courses/9054/users/229165>)

Apr 14, 2024

Hi Dawn,

I think once you have gathered the additional variables, you won't need to reconsolidate the data. Instead, you would calcsilicate the findings and recommendations. In the example of the student test score, you would determine that those 2 samples are representatives of 2 distinct groups with 2 different catachrestic, maybe one county has stronger educational faculty than the other, or maybe one set of educational material is more effective than the other, and so on.

Another example would be shoe manufacturing, once you determine the shoe sizes for men and women are fundamentally different, you would not need to reconsolidate the data. Instead, you would proceed to make shoes differently for each segment, and so on.

This is how it made sense to me. However, I am interested to hear your thoughts on it.

← Reply 👍



Dawn Prewett (<https://classroom.emeritus.org/courses/9054/users/233112>)

Apr 14, 2024

I had thought about how we could just keep them separate, but in the case of the data I am considering from work, the inputs aren't fundamentally different, but created by a flawed system - human perception. Keeping them separate would mean I have a separate data set for each person who inputs the data. This might actually be a sign that we need to create a better system, but that isn't always feasible. I'm sure there are some new tools that will be added to our toolbox to help us answer this very question.

← Reply 👍 (1 like)



Haitham Farag (<https://classroom.emeritus.org/courses/9054/users/233864>)

Apr 16, 2024

The examples you kindly shared of nonnormally distributed datasets grounded ECDF into real life scenarios, and enhanced my knowledge on the subject.

Thanks, Turki

Edited by **Haitham Farag** (<https://classroom.emeritus.org/courses/9054/users/233864>) on Apr 16 at 11:03am

← Reply 👍



Dawn Prewett (<https://classroom.emeritus.org/courses/9054/users/233112>)

Apr 14, 2024

Working through the ECDF try-it provided some interesting insights. Since we merged data from two different normal distributions, two distinct bell curves were revealed when the data was graphed using a histogram. Initially, I didn't realize the impact these separate distributions would have, but quickly realized the impact when I identified that the likelihood of scoring below the mean of 65% this was actually higher than 50%. This shows how the size and nature of each dataset can skew the results.

To better understand how this skew came to be, it is important to understand that each bell curve effectively has its own mean and the actual mean falls somewhere between the two means of the separate datasets. If one dataset has more data points, a higher average, or both then the mean will be skewed in its direction. Thus, the mean doesn't always truly indicate the middle of the dataset.

Understanding this through the ECDF isn't just about knowing what the average is but seeing how the data stacks up around it. The ECDF graph I generated laid this all out, showing the cumulative probability of each score, which really helps visualize where the bulk of my data points fall.

Looking forward, I can see how this understanding would be useful when working with some of the datasets I analyze at work. These datasets represent information collected by a number of unique individuals and are based on their perceptions of the situation. So, while the inputs are the same, they don't have a 1-to-1 correlation. If I were to graph these, I suspect there would be skewing, especially since some inputs are more prolific from certain individuals than others. Correcting for this and understanding it will be crucial to making the most out of this data.

Looking forward, I can see how this will be useful when working with some of the datasets I analyze at work. These datasets represent information collected by a number of unique individuals and are based on their perceptions of a given situation. So, while the inputs are the same, they do not have a 1-to-1 correlation. If I were to graph these, I suspect there would be skewing, especially since some inputs are more prolific from certain individuals and/or inquiries than others. Correcting for this and understanding it will be crucial to making the most out of this data.

In short, the ECDF doesn't just identify what is typical, but also provides a more comprehensive picture of our data's distribution. It's a tool that can help spot trends, understand ranges, and make informed decisions based on the actual layout of our data.

← Reply 👍 (1 like)



Ricardo Anaya (<https://classroom.emeritus.org/courses/9054/users/228915>)

Apr 14, 2024



thanks for this comment:

In short, the ECDF doesn't just identify what is typical, but also provides a more comprehensive picture of our data's distribution. It's a tool that can help spot trends, understand ranges, and make informed decisions based on the actual layout of our data

to be honest I had a really hard time understing practical uses of this, this gave me clartiy

← Reply 



Dawn Prewett (<https://classroom.emeritus.org/courses/9054/users/233112>)

Apr 14, 2024

Yes, you are correct about the more comprehensive picture. I actually thought I had included it in there!

← Reply 



Lee Lanzafame (<https://classroom.emeritus.org/courses/9054/users/231975>)

Apr 15, 2024

well done, love the way you summarised this, you have a deep understanding of ECDF.

← Reply 



Priscilla Annor-Gyamfi (<https://classroom.emeritus.org/courses/9054/users/226376>)

Apr 16, 2024

Great post Dawn. I like the observations shared from the standpoint of the skewness of the data and its impact on the mean. I love how it is a great tool to spot trends, understand ranges and even the quartiles of a data to make a quick but well informed data-driven decision.

← Reply 

 [https://](https://classroom.emeritus.org/courses/9054/users/228915)[Ricardo Anaya \(https://classroom.emeritus.org/courses/9054/users/228915\)](https://classroom.emeritus.org/courses/9054/users/228915)

Apr 14, 2024

The Empirical Cumulative Distribution Function (eCDF) is a way to describe how data points are distributed in a sample.

In this set of data points, test scores from a class. The eCDF shows how many data points are less than or equal to a specific value.

Sorting our data points in ascending order.

For each data point, we calculate the proportion of data points that are less than or equal to it.

Plotting these proportions as steps on a graph. Each step represents a data point, and the height of the step will tell us the proportion of data points below that value.

The ECDF helps us understand how data is spread out in the whole range of values, and present us a visual representation of the cumulative distribution of our sample.

I found that it can be used to Compare Distributions, to determine if it fits in theoretical data distributions, and Quantile estimations

This is a little research on the uses of the ECDF practical uses:

“Comparing Distributions:

You can overlay multiple eCDFs to compare distributions from different samples or populations.

Example: Suppose you want to compare the heights of two different plant species. Plotting their eCDFs allows you to see which species tends to be taller. “

Source:

https://lazymodellingcrew.com/post/post_14_reading_like_a_fourthgrader_ta/ 
(https://lazymodellingcrew.com/post/post_14_reading_like_a_fourthgrader_ta/)


“Goodness-of-Fit Testing:

The eCDF can be used to assess how well your sample data fits a theoretical distribution (e.g., normal, exponential, etc.).

Example: If you suspect your data follows a normal distribution, compare the eCDF of your sample to the theoretical normal cumulative distribution function. Deviations may indicate a

poor fit”

Source:

<https://www.library.virginia.edu/data/articles/understanding-empirical-cumulative-distribution-functions>  (<https://www.library.virginia.edu/data/articles/understanding-empirical-cumulative-distribution-functions>)

“Quantile Estimation:

The eCDF provides estimates for percentiles or quantiles.

Example: To find the 75th percentile (Q3) of a dataset, locate the value on the x-axis where the eCDF reaches 0.751.”

Source:

https://en.wikipedia.org/wiki/Empirical_distribution_function 
(https://en.wikipedia.org/wiki/Empirical_distribution_function)

 Reply  (2 likes)



Haitham Farag (<https://classroom.emeritus.org/courses/9054/users/233864>)

Apr 16, 2024

Good day Ricardo

Thank you for the delineated response. One of the sources you kindly shared helped me grasp one of the probability examples from Medel 2 office hour (i.e. probability of 6 heads in an x number of grouped tosses)

Thanks again

 Reply 



STEPHEN HUTSON (<https://classroom.emeritus.org/courses/9054/users/233645>)

Apr 17, 2024

Thanks for sharing these examples Ricardo! I think by being able to look at realistic examples like the goodness of fit testing and the heights between different species was a great way to get a practical sense of how the eCDF can be used in future analyses. This

response was helpful in me deepening my understanding of the concept and reminds me of the lessons in prior modules around how it's important to use a variety of tools to understand data and not just rely on the basic statistics like mean.

← Reply 



Shahrod Hemassi (He/Him) (<https://classroom.emeritus.org/courses/9054/users/224267>)

Apr 20, 2024

Hi Ricardo. I like the examples that you have shared. It helped to gain a wider understanding of the ECDF and it's applicability. It also helped to understand what to look out for and when the ECDF could be a good option to use. Thanks for sharing!

← Reply 



Koffi Henri Charles Koffi (<https://classroom.emeritus.org/courses/9054/users/208039>)

Apr 23, 2024

hi Ricardo , thank you for share the resource and the tips here , it really helpful

← Reply 



Lee Lanzafame (<https://classroom.emeritus.org/courses/9054/users/231975>)

Apr 15, 2024

1. ECDF allows us to use historical data to predict future data values
2. ECDF, doesn't assume a specific distribution, it can be used on Non-normally distributed data set
3. There is an above average probability a random students score is below the average
4. Calculating using statistics can be very different to calculating a probability instead
5. This tool can be used to predict/forecast future values for any type of distribution.

← Reply 



STEPHEN HUTSON (<https://classroom.emeritus.org/courses/9054/users/233645>)

Apr 15, 2024

The ECDF is a powerful tool that we were able to leverage for this example to examine a case where data is not normally distributed. One interesting observation that came out of this example is that rather than seeing student test scores on a normal bell curve, we saw there were really 2 clusters of data that indicated that a larger population of students tended to score lower in one cluster, and a smaller subset of students tended to score higher around another cluster. When we looked at the more simplistic statistics like the mean, by using the ECDF we were able to determine that although the mean was ~65%, the majority of students scored lower than this mean for the entire dataset which would be counterintuitive if the data were to fall on a normal curve. Given that the ECDF is generated based on real data, it is a good strategy that can be utilized when conducting exploratory data analyses in the future to get a better sense of looking at datasets.

← Reply 👍



Timothy Andrew Ramkissoo (<https://classroom.emeritus.org/courses/9054/users/226697>)

Apr 16, 2024

Empirical Cumulative Distribution Function (ECDF) models empirical (observed) data. It helps us understand how our data behaves, as you collect more data, the ECDF gets closer to the true CDF. As an Asset Integrity Manager, I would like to utilize ECDF for risk assessment and decision making - utilizing historical data to assess the probability of certain events or failure occurring. This information helps make informed decisions regarding asset maintenance, inspection schedules and risk mitigation strategies.

One of the projects that I'm currently working on is the development of a reliability dashboard to provide insights into the performance of assets over time. This utilizes historical data from global projects with similar product designs and compares with data from current project history. ECDF can be used to estimate the probability of failure at different levels and aids in optimizing maintenance resources and prioritizing critical assets.

← Reply 👍 (1 like)



Jignesh Dalal (<https://classroom.emeritus.org/courses/9054/users/229173>)

Apr 16, 2024

Hi Timothy, Thank you for your clear explanation on using the Empirical Cumulative Distribution Function for risk assessment and decision-making in assess integrity

management.

Your application of ECDF in developing a reliability dashboard is intriguing. Please can you help how the dashboard data put together will answer to challenges in data standardization across different projects?

Also, I am curious to know potential of incorporating real time data to dynamically update ECDF's.

Much Appreciated.

← Reply 👍

○



Priscilla Annor-Gyamfi (<https://classroom.emeritus.org/courses/9054/users/226376>)

Apr 16, 2024

The Empirical Cumulative Distribution Function (ECDF) is derived by arranging all unique observations in a dataset and computing the cumulative probability for each observation. It converges with probability 1 to that underlying distribution, according to the **Glivenko–Cantelli theorem** (https://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theorem). This is done by dividing the number of observations less than or equal to a particular value by the total number of observations. While most distributions like normal and binomial may not always accurately represent the data, the ECDF serves as a valuable alternative. It, along with historical dataset information, becomes pivotal in predicting future data values.

In this analysis, we utilized a dataset representing benchmark scores from a state exam. Graphical representation revealed non-normal distribution patterns within the datasets. Calculations were performed for mean, maximum, minimum, and range scores, providing historical context for future predictions alongside the ECDF. Notably, the ECDF was constructed using probabilities derived from mean, maximum, and minimum scores. Interestingly, it was found that there is an above-average likelihood (65%) of a random student's score falling below the mean. This underscores the significant distinction between statistically calculated predictions for a random variable and those derived solely from probability assessments.

I have learned that:

1. The ECDF offers a clear graphical depiction of your data being distributed across the data set from least to greatest.
2. You can easily compare differences and similarities between the distribution of different datasets or variables within the dataset.

3. The ECDF allows for easy calculation of percentiles within a dataset such as the median(50th percentile), 25th and 75th percentiles.
4. It is easy to assess the range of your data as well as identifying outliers.

This can be used in the future by:

1. offering effective predictive analysis on data by generating great insights into future events or occurrences making use of historical data distribution.
2. helping stakeholders to make well data-driven/ informed decisions.
3. analyzing and helping to understand the distribution of new datasets, identifying patterns, outliers, and trends.

ACTIVITY 3.1.docx ([https://classroom.emeritus.org/files/2473778/download?](https://classroom.emeritus.org/files/2473778/download?download_frd=1&verifier=R2gOAr4tBiKI3dXYgoD0JdlRFSz4HrQWF9Jq1FCa)

[download_frd=1&verifier=R2gOAr4tBiKI3dXYgoD0JdlRFSz4HrQWF9Jq1FCa](https://classroom.emeritus.org/files/2473778/download?download_frd=1&verifier=R2gOAr4tBiKI3dXYgoD0JdlRFSz4HrQWF9Jq1FCa))

← Reply 👍



Chris Cosmas (He/Him) (<https://classroom.emeritus.org/courses/9054/users/226607>)

Apr 16, 2024

After reviewing the functionality of the ECDF it has shown how simple statistical metrics are not sufficient because they do not fully capture the characteristics of the parameter, even though the average of students' grades was computed at 66.1339 this proved to be misleading as students are more likely to score below the average which was shown through the ECD function. This is due to the non-normal distribution of the data as it presents two different peaks which pulls the mean away from the modes, presenting confusing conclusions. The exercise has also shown the importance of visualizing data before picking a statistical treatment, as the data will not always follow preset distribution it allows us to use other tools which might be a better fit for the data at hand. It also highlights the importance of choosing an empirical vs a theoretical distribution.

← Reply 👍



Jignesh Dalal (<https://classroom.emeritus.org/courses/9054/users/229173>)

Apr 16, 2024

Practical Applications of the Empirical Cumulative Distribution Function (ECDF) in Non-Normally Distributed Data

In our latest activity, we explored the use of the Empirical Cumulative Distribution Function (ECDF) to analyze non-normally distributed datasets. While data often conforms to normal or binomial distributions, real-world datasets sometimes do not.

When data does not fit known distribution types, ECDF is an alternative. It uses historical data, which may be based on expert opinion, weighted data, or computer-generated values, to predict future data points. Key assumptions for ECDF include the independence of events and that the sum of the probabilities is one.

We created a two-peak dataset by merging two normal distributions to simulate complex data scenarios. We then used ECDF to analyze how data is distributed across different intervals.

The ECDF is calculated by sorting all unique data points and calculating the cumulative probability for each. This method allows us to visually inspect the percentile ranking of data points, offering insights into the distribution's shape and spread.

Observations from the Activity

The ECDF plot showed how data clusters and where potential outliers might be. Unlike parametric methods that assume a specific distribution, ECDF is non-parametric and adapts well for predictive analytics in fields with irregular data.

Future Applications

ECDF can be useful in several areas:

Risk Management: Understanding distributions of financial losses or claims to predict risk exposure.

Quality Control: Assessing if product batches meet specifications by comparing quality metrics against benchmarks.

Environmental Studies: Monitoring environmental data changes to evaluate intervention impacts or climate shifts.

Engaging with Peer Insights

I encourage everyone to share their insights on using ECDF in their areas. How do you see this tool in your work? What implementation challenges do you anticipate?

Conclusion

Exploring ECDF has expanded our tools for managing complex datasets. I look forward to your experiences and insights on using ECDF in various contexts and engaging in discussions to deepen our understanding.

[← Reply](#) 



Shahrod Hemassi (He/Him) (<https://classroom.emeritus.org/courses/9054/users/224267>)

Apr 20, 2024

In this exercise, I was able to see how ECDF can use historical data to predict future values. We had 2 peaks in our historical data set. We had a large peak at 60% and a smaller peak at 80%. We generated our historical data with a standard deviation of 5 from these 2 target values.

We could do some basic analysis of the historical data and find that the average (mean point) was 65% and this might have been useful if we had a linear distribution, but our data had 2 peaks which required us to use the ECDF to produce a prediction of future values.

We used ECDF to predict future values. When we analyzed the future values and used the mean-point of the historical data set, we found that there was an above-average probability that a student would score lower than the mean point.

Additionally, we found that our S-curve plot of future values was not perfectly smooth, which makes sense as we were basing our analysis on historical data that had 2 peaks associated with it. In fact, we notice a steeper climb at around these 2 values which makes sense.

In the future, I could envision this being used in sports prediction of future performance of players based on historical performance. In baseball for example, when we look at hitting average, we often see 2 or more peaks in the data. Using ECDF, we could evaluate the historical data appropriately to predict the future performance of a random baseball player.

Edited by **Shahrod Hemassi** (<https://classroom.emeritus.org/courses/9054/users/224267>) on Apr 20 at 11:15am

← **Reply** 👍



Gustavo Santana (<https://classroom.emeritus.org/courses/9054/users/120927>)

Apr 20, 2024

The ECDF will be important when we want to calculate the probabilities of an event using historical data and its distribution isn't normal.

One occasion I faced this problem was dealing with Support Tickets, we had an exponential distribution skewed right, where many tickets were closed after 1 day but some stayed open for months. Using ECDF would make it possible to understand the probabilities of closing the tickets much better than the average completion time.

 [Reply](#) **Mhelissa Yayalar** (<https://classroom.emeritus.org/courses/9054/users/233590>)

Apr 24, 2024

Hi Gustavo,

Your real-life scenario example, "Support Tickets," I think is great example of using ECFD. Limiting the data analysis to just average tickets closed per time-period, such per day, does not provide actionable insights. For instance, your company wanted to understand what's the lowest number of support tickets being closed at specific time period during the day. Calculating ECFD can be provide your company insights as to further understand what's driving the lowest transaction during the day or whether they need to evaluate resourcing during that time-period. Regardless, using a simple aggregate data set to drive business outcomes is not a competitive advantage. Using cumulative functions, like ECFD or CFD, can be instrumental for businesses.

 [Reply](#) **Koffi Henri Charles Koffi** (<https://classroom.emeritus.org/courses/9054/users/208039>)

Apr 23, 2024

the Empirical Distribution (ECDF) is a powerful function that allow us to easy combine different distribution . and I think it can help to analyse more complex dataset or analyse dataset by combining different source of data .

this tool is really a powerful one .

 [Reply](#) **Isabella Tockman** (<https://classroom.emeritus.org/courses/9054/users/207395>)

Apr 24, 2024

I've learned that we can create and study distributions with two sets of data, each having different averages. When dealing with large datasets, it's hard to make sense of the information. That's where the Empirical Cumulative Distribution Function (ECDF) comes in

handy. It helps us visualize how the data is spread out, and from the plot, we can tell that it doesn't follow a normal pattern.

Recognizing this deviation from normality is crucial because assuming a normal distribution can lead us to wrong conclusions and make our models ineffective.

ECDF is also great for handling large datasets. It lets us spot even small changes in patterns, thanks to the curves we see in the plot. Plus, it's useful in many areas. In finance, it helps us analyze how stocks are doing, and in social services, it lets us understand things like employment rates. It's a versatile tool that gives us deeper insights and helps us make better decisions.

← Reply 👍



Mhelissa Yayalar (<https://classroom.emeritus.org/courses/9054/users/233590>)

Apr 24, 2024

In my observation, Empirical Cumulative Distribution Function (ECDF) provides a more broader understanding of how the data distribution is of real data set in comparison to simple aggregated measures like mean and median. For instance, at face-value, if we just calculate the mean of the students score, it shows the average score is 65%, which is just pointing out just that, an average score from all the array of scores from our real-time observations. However, if we wanted to further understand what the chances of students is attaining lower test scores than 65%, that's when we use the ECDF. In the colab exercise, using ECDF, we discovered that although the average score was a 65%, there is an above average probability a random student's score is below the average. Which describes that there many students who did not do well on the exam. Instead of just focusing on the average, using ECDF you can determine the holistic picture of the entire data set to determine the nuances or patterns. In contrast, if we wanted to predict what's the mean of new students taking the test next year, then rather than using ECDF, we would use CDF since CDF handles hypothetical data in order to predict possible outcomes.

← Reply 👍



Haitham Farag (<https://classroom.emeritus.org/courses/9054/users/233864>)

May 15, 2024

In the above Act 3.1.

Was the "additional practice, we have included a bonus activity within [Colab](https://colab.research.google.com/drive/1iNwgdrqREKPtrHyZr7QtbkyKp_DE17VK?usp=sharing) (https://colab.research.google.com/drive/1iNwgdrqREKPtrHyZr7QtbkyKp_DE17VK?usp=sharing)" tested?

getting an error running the second patch of code. (attached)

link

https://colab.research.google.com/drive/1iNwgdrqREKPtrHyZr7QtbkyKp_DE17VK?usp=sharing#scrollTo=3F0vrUX5Bqux

[3.1.try it .png \(https://classroom.emeritus.org/files/2603270/download?download_frd=1&verifier=iYV4YBXLOp9uZ6DQb4Nui992VisyAMCuJ83Ga8b0\)](https://classroom.emeritus.org/files/2603270/download?download_frd=1&verifier=iYV4YBXLOp9uZ6DQb4Nui992VisyAMCuJ83Ga8b0)

← Reply 