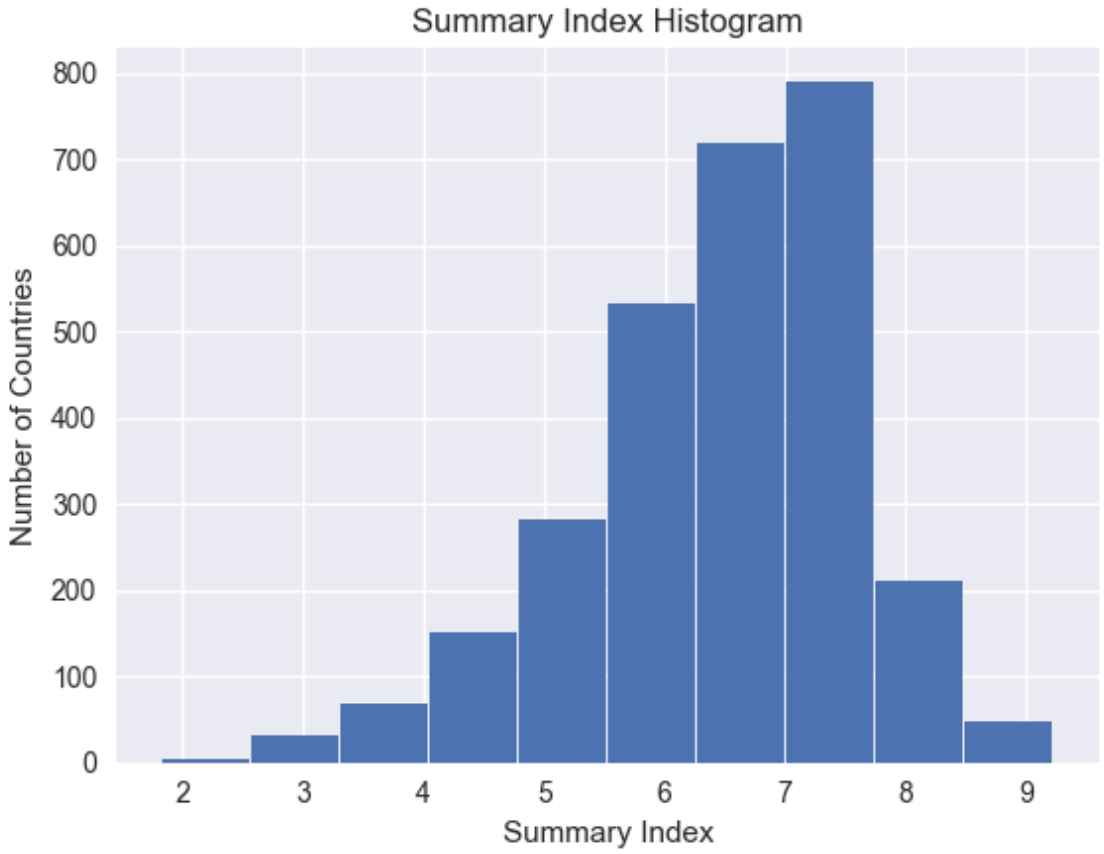
What, Why, and How to Read Empirical CDF

John DeJesus Mar 4, 2019 · 6 min read ★

Exploratory Data Analysis (EDA) is encouraged to get digestible glimpses of your data. This includes histograms for summary statistics:



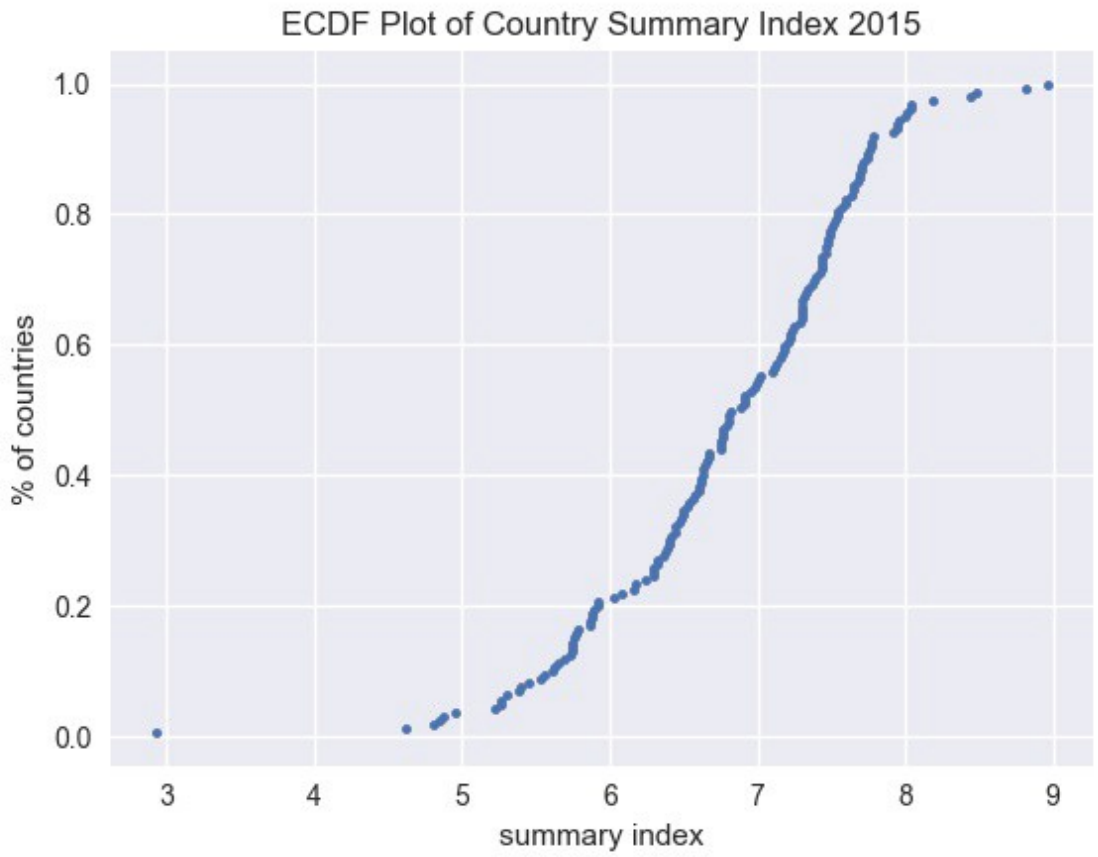
Scatter plot comparing the Economic Summary Index of Countries with their Size of Government.

But there is one summary statistic visualization that I did not learn about until I explored a statistical thinking course from Datacamp. It is known as the Empirical Cumulative Distribution Function (try saying that 10 times fast…we will call it ECDF for short).

In this post, we will explore what an ECDF is, why to use it and the insights we can read from it using our Economic Freedom of the World dataset provided by the folks at #MakeoverMonday. That data was also used to make the plots above. The code used for these charts and a copy of the excel file are in this Github Repo. The code isn't included in this post so you can focus on the visuals and interpret them.

So what is an Empirical Cumul….What's an ECDF?

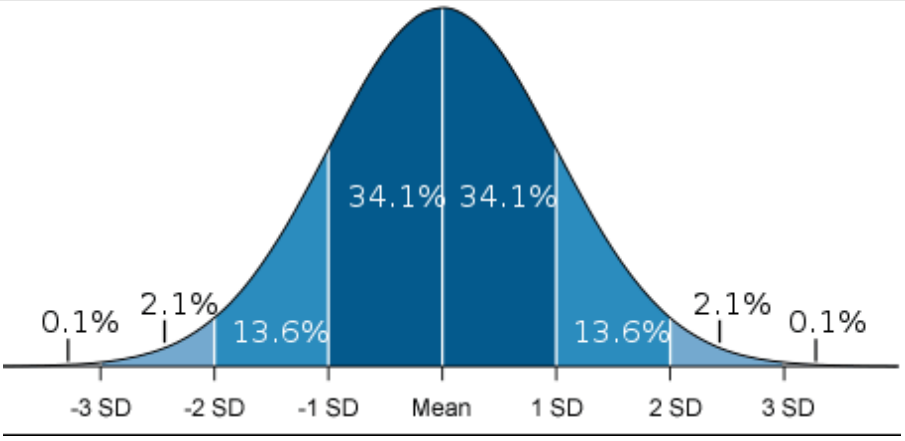
An ECDF is an estimator of the Cumulative Distribution Function. The ECDF essentially allows you to plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set. Let's take a look at the ECDF chart above in the post.



Here we can see the variance in Economic Freedom Summary Indexes of countries (for 2015). This chart can provide us with some summary statistics about how the economies of these countries are performing. One quick insight is that the range of the indexes goes from a little under 3 to 9 (range of about 6). So no country has complete economic freedom.

Sounds Great! But how do I read the ECDF to get those Summary Statistics?

I had trouble reading this at first too but it is easier than you think. The explanation I received wasn't enough for me. Then I thought there must be something similar I can use that was similar to the ECDF that could improve my understanding. Let me provide you with that scaffolding through the Gaussian (normal) distribution which you may be more familiar with.

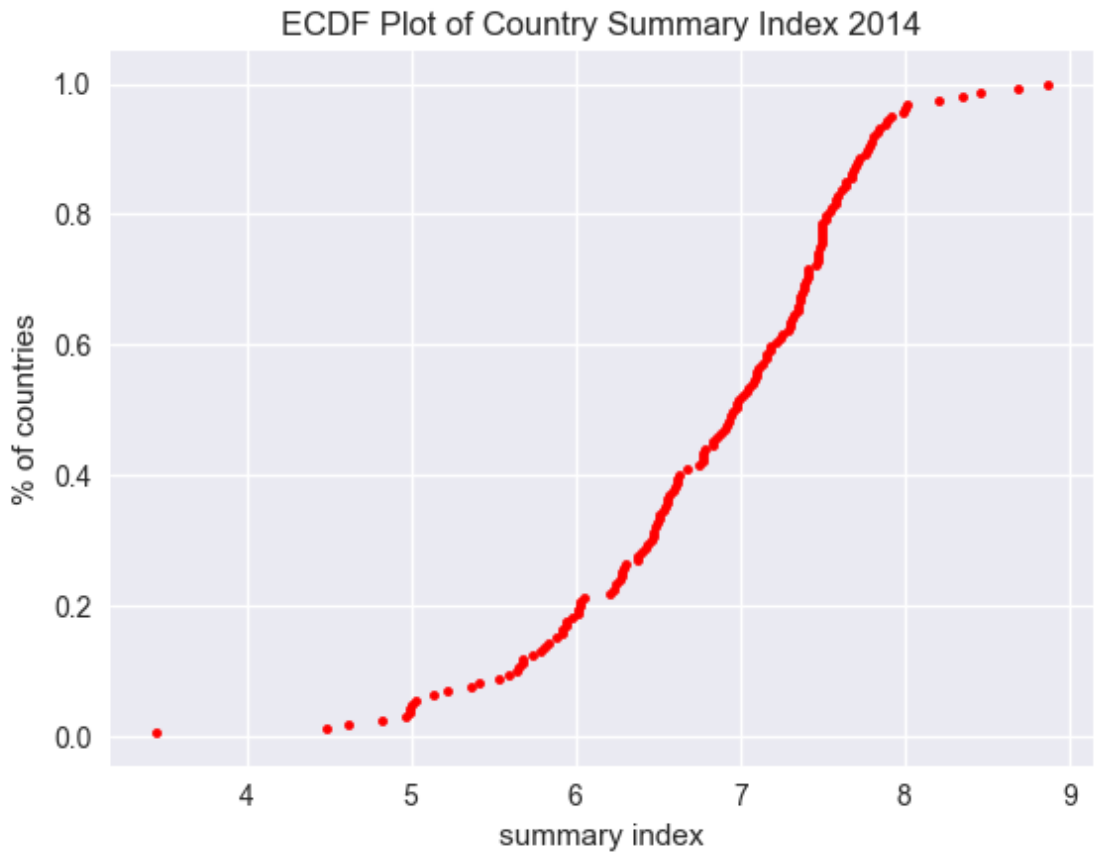


Gaussian Distribution. Image from etfhq.com

If you are familiar with the Gaussian Distribution, let me ask you this question. What percentage of data points are less than the mean? If you answered 50%, then you are right!

The ECDF can be read in a similar way! For example, what percentage of countries have a summary index of less than 6? Look on the x-axis to 6, then move vertically until you hit the curve. You get to about 20%. Therefore, about 20% of the countries have a summary index of less than 6. So they have some economic freedom.

Here is another ECDF plot with the same data but from 2014.



ECDF of the Countries' Economic Summary Index in 2014.

Try answering the following questions using the chart above. I will leave the answers underneath the questions.

1. What percentage of countries have a summary index less than 6?

2. What is an approximate percentage of the countries that have a summary index less than 8?

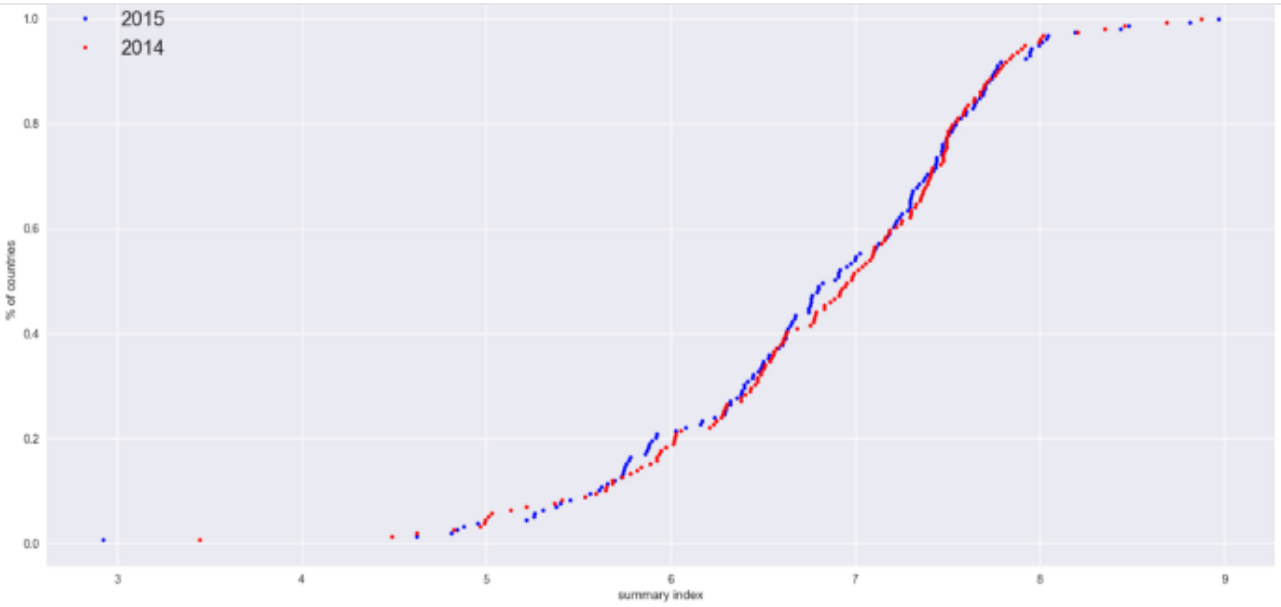
Answers:

1. About 20%.

2. About 97–98%.

Awesome! How about Plotting Multiple ECDFs Simultaneously?

We can plot multiple ECDF plots! For example, let's plot the two ECDFs onto a single chart for 2014 and 2015.



Now you can determine summary statistics for two years and compare them. You can see that the summary indexes for both years are almost identical within certain summary index ranges such as from 7 to 8. But in the range of 6 to 7, 2015 had a slightly larger percentage of countries that had a summary index less than 7.

Now you may be wondering, how many ECDFs can you plot on a single chart? The most I saw was about 3. This data has 22 years' worth of summary indexes. Let's just see how good plotting 22 ECDFs is (spoiler alert: it won't be good…)

ECDF plot with all 22 years worth of Country Summary Indexes. Nulls were dropped.

of ECDFs plotted simultaneously. Even with the legend, it is difficult to distinguish one year from the others. This is a case where showing less will reveal more. Let's instead plot 4 years' worth. Specifically, we will plot from 1970 to 2015, going to every 15th year in that interval.

ECDF plot of summary indexes for 1970, 1985, 2000, and 2015

Much clearer isn't it? Plus this allows you to see distinctions in the variation of the summary indexes over 15-year jumps. What observations can you make from this plot? Try to think about this before reading about them in the next paragraph.

The immediate observation you could make is that the summary indexes were overall lower in 1970 and 1985 versus the later years. Therefore economic freedom increased overall from 2000 and at least in 2015 also. There are a couple of other interesting observations with 1970 and 1985. One is that 1970 appears to have the least amount of data points, which may indicate that very few countries had economic freedom or some quantities weren't recorded.

The second observation is that in 1985 even though the number of countries with some economic freedom increased (more data points), the variability in summary index values is smaller due to how its ECDF curves. That is, in 1985 there was a greater percentage of countries with a summary index less than 6 than the same percentage in 1970.

Thanks for reading this post! If you would like to get some more info on ECDF, check out the following:

1. Conceptual Foundations of ECDF-Great blog post showing the math behind the ECDF.

2. ECDF and CDF Wikipedia Pages for additional reading.

3. Datacamp Statistical Thinking ECDF Video-Intro video by Justin Bois where I learned about the existence of ECDFs. Good for visual learners and to hear the advantages of it versus a bee swarm plot. He also shows how to create an ECDF plot using Python.

4. Below is a video version of this article with a dataset on avocado sales. I go through the same explanation as above. So if you are more of a visual learner this will be helpful for you.

Hope you found ECDFs as interesting as I did (last time I type ECDF I promise… including this one…). I plan on adding this to my EDA toolkit and using it whenever possible. Will you do the same? What observations in the data did you make that I missed? Let me know in the comments below.

Linear regression is a classic technique to determine the correlation between two or more continuous features of a data

Data Visualization Ecdf Statistics Data Science Data Scientist

towardsdatascience.com

Until next time,

John DeJesus