Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. It only takes a minute to sign up.

Sign up to join this community

Anybody can ask a question
Anybody can answer
The best answers are voted up and rise to the top

SPONSORED BY aws machine learning

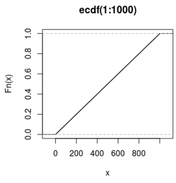
Why is the empirical cumulative distribution of 1:1000 a straight line?

Ask Question

Asked 4 years, 10 months ago Active 3 years, 9 months ago Viewed 3k times

6

Why does plot(ecdf(1:1000)) produce a straight line?



Since $Fn(x_n) = x_1/(\text{total sum}) + x_2/(\text{total sum}) + \ldots + x_n/\text{total sum} = (x_1 + x_2 + x_3 + \ldots + x_n)/\text{total sum}$. the fact that Fn(200) roughly being equal to 0.2 and sum(0:200) roughly being equal to 0.4 seems to indicate that sum(1:200) is roughly half of sum(1:400), which isn't true, the two expressions being 20,000 and 80,000 respectively.

What am i misunderstanding?

distributions proportion cumulative-distribution-function

Share Cite Improve this question Follow

edited Aug 7 '17 at 12:03 Firebug 13.8k 3 49 116

asked Jul 4 '16 at 18:56 Qwertford 275 3 9

3 Give a good look. It is a step function. – Khashaa Jul 4 '16 at 23:58

Add a comment

4 Answers

Active Oldest Votes

11

The cumulative distribution function of a random variable X has nothing to do with *summing* the random variable. It is

> the probability that X will take a value less than or equal to x.

And of course, the probability that a value randomly sampled from your vector $(1, \ldots, 1000)$ is less than or equal to 200 is exactly half the probability that it is less than or equal to 400.

Share Cite Improve this answer Follow

answered Jul 4 '16 at 19:00 Stephan Kolassa 82.5k 10 165 307

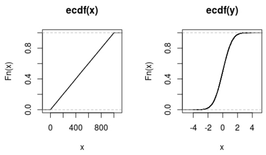
Add a comment

8

Empirical cumulative distribution function is a cumulative sum of frequencies of observed x_i's divided by total sample size. Your data is a vector of values from 1 to 1000, where each of the values appears exactly once. This means that your "variable" follows a discrete uniform distribution, that has a flat CDF.

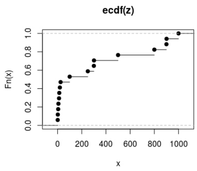
As you can see on the example below, it'd be different if you used other imput data.

```
set.seed(123)

x <- sample(0:1000, 1e5, replace = TRUE)
y <- rnorm(1e5)

def <- par(mfrow = c(1,2))
plot(ecdf(x))
plot(ecdf(y))
par(def)
```



or

```
z <- c(1,2,5,7,12,14,19,25,100,250,300,301,500,800,900,901,1000)
plot(ecdf(z))
```



Notice that in the second example distances between different values are different so no matter that each value appeared only once, the line is curved.

Share Cite Improve this answer Follow

edited Jun 11 '20 at 14:32 Community ♦ 1

answered Jul 4 '16 at 19:00 Tim ♦ 91.6k 17 193 332

Add a comment

3

You can think about it mechanically, too.

The ECDF $\hat{F}$ evaluated at x is the proportion of observations with value x or below. Since you have exactly 1,000 observations $\{y_i\}_{i=1}^{1000}$, the difference between $\hat{F}(y_i)$ and $\hat{F}(y_{i+1})$ is always 0.001 for any $1 \leq i < 1000$.

Moreover, your sample values are evenly spaced, so the difference between y_i and y_{i+1} is always 1. Therefore, for any $1 \leq i < 1000$, the slope between $(y_i, \hat{F}(y_i))$ and $(y_{i+1}, \hat{F}(y_{i+1}))$ is always $\frac{0.001}{1} = 1$. A curve with a constant slope is just a straight line.

As for what you're misunderstanding, the Fn you defined is definitely not the right formula. The denominator should be the number of observations, and the numerator should be the number of observations with value at or below x_n.

Share Cite Improve this answer Follow

edited Jul 5 '16 at 11:01 Nick Cox 46.1k 8 106 151

answered Jul 5 '16 at 5:50 shadowtalker 11k 3 46 104

Add a comment

2

The empirical distribution function of a sample $Y_1, \ldots, Y_n$ is defined as

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(Y_i \leq x)$$

In your data set, $Y_i = i$. So, $\widehat{F}(x) = x/n$, for $x = 1, 2, \ldots, 1000$. Plotted the way you did, this looks like a linear function of x.

Share Cite Improve this answer Follow

edited Jul 10 '16 at 20:36

answered Jul 4 '16 at 19:02 gammer 1,416 5 21

1 The final equation is incorrect. $\hat{F}$ is a *step function*, not a linear function. – whuber ♦ Jul 5 '16 at 14:06

@whuber, I never did thank you for that astute, and very substantively important, correction to my answer. Keep up the good work. – gammer Jan 9 '17 at 0:17

Add a comment

Your Answer

B I Link Image Code Bullets Numbers Indent Outdent Undo Redo

Sign up or log in

Sign up using Google
Sign up using Facebook
Sign up using Email and Password

Post as a guest

Name

Email
Required, but never shown

Post Your Answer

By clicking "Post Your Answer", you agree to our terms of service, privacy policy and cookie policy

Not the answer you're looking for? Browse other questions tagged distributions proportion cumulative-distribution-function or ask your own question.

Featured on Meta
Testing three-vote close and reopen on 13 network sites
We are switching to system fonts on May 10, 2021
Should we reduce the vote threshold for closing questions?

Linked
23 Empirical CDF vs CDF
1 How is this chart called? Percentile chart?

Related
1 Using monte carlo integration with bisection method to find a true value in R
2 Cádlág property of cumulative distribution functions
0 Empirical conditional cumulative distribution function
3 Accuracy of empirical cumulative distribution function
0 q-Gaussian Cumulative distribution function
1 Updating discrete probability distribution: Bayesian or something else?

Hot Network Questions
What's the opposite of the phrase "pay under the table"?
Why do we need air-filled tires?
What is this icon from shell32.dll
Structure of "affirming the consequent fallacy"
In GIMP, how can I identify and match the saturation of an image
Do certificates need to be stored as encrypted?
Can a person be operating an aircraft in a "Careless" and "Reckless" manner at the same time? Or are these two different standards?
What is Narcipasis?
What is the intuition behind Chebychev's Inequality in Measure Theory
Hypothetical conflict - incorrectly licensed code
Why are there more research papers on Android malware than iOS malware?
In which Chinese fairy tale does a painter disappear into his own painting?
Multiple cubes as just one object?
How do I fix Outlook only showing the first line of every message?
Why is the number 63 popular in electronics?
How does Planck's curve solve the ultraviolet catastrophe?
Guide my eyes in the right direction
Can Newton's gravity equation explain why black holes are so strong?
CGCC sings a song together
What is the purpose of identifier-first login screens?
Why can space telescopes see through a planetary nebula?
Is there an easy way to say you do not want to hear Lochon Hara?
Find Abecedarian Words
"Old" vs new Shimano 9-speed compatibility?
Question feed

site design / logo © 2021 Stack Exchange Inc; user contributions licensed under cc by-sa. rev 2021.5.12.39277