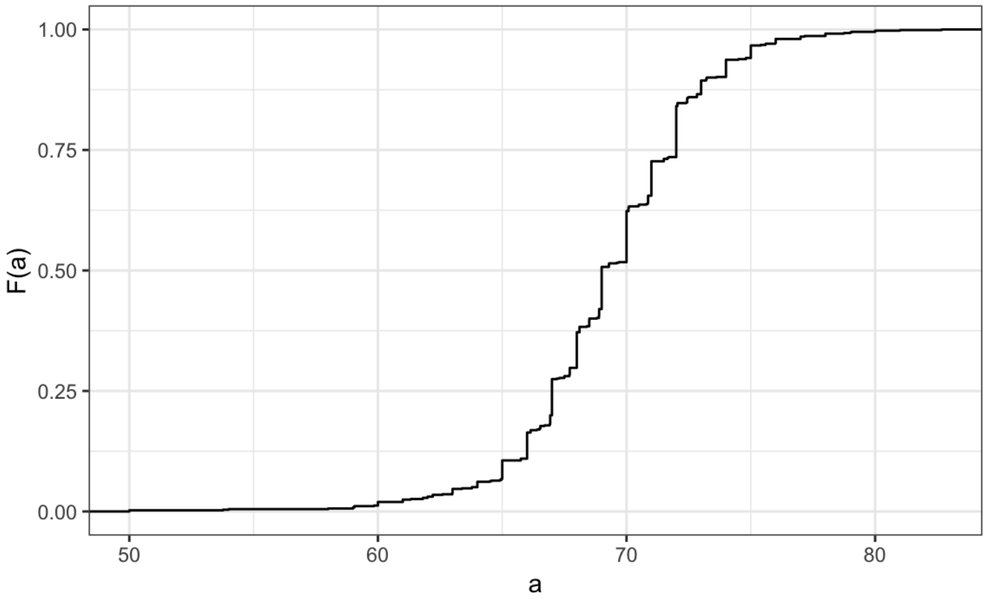
Course › Section... › 1.2 Intr... › Cumul...

Cumulative Distribution Function

Every continuous distribution has a *cumulative distribution function (CDF)*. The CDF defines the proportion of the data below a given value a for all values of a:

$$F(a) = \Pr(x \leq a)$$

Any continuous dataset has a CDF, not only normal distributions. For example, the male heights data we used in the previous section has this CDF:



As defined above, this plot of the CDF for male heights has height values a on the x-axis and the proportion of students with heights of that value or lower ($F(a)$) on the y-axis.

The CDF is essential for calculating probabilities related to continuous data. In a continuous dataset, the probability of a specific exact value is not informative because most entries are unique. For example, in the student heights data, only one individual reported a height of 68.8976377952726 inches, but many students rounded similar heights to 69 inches. If we computed exact value probabilities, we would find that being exactly 69 inches is much more likely than being a non-integer exact height, which does not match our understanding that height is continuous. We can instead use the CDF to obtain a useful summary, such as the probability that a student is between 68.5 and 69.5 inches.

For datasets that are not normal, the CDF can be calculated manually by defining a function to compute the probability above. This function can then be applied to a range of values across the range of the dataset to calculate a CDF. Given a dataset `my_data`, the CDF can be calculated and plotted like this:

```
a <- seq(min(my_data), max(my_data), length = 100)    # define range of
cdf_function <- function(x) {    # computes prob. for a single value
    mean(my_data <= x)
}
cdf_values <- sapply(a, cdf_function)
plot(a, cdf_values)
```

The CDF defines that proportion of data below a cutoff a. To define the proportion of values above a, we compute:

$$1 - F(a)$$

To define the proportion of values between a and b, we compute:

$$F(b) - F(a)$$

Note that the CDF can help compute probabilities. The probability of observing a randomly chosen value between a and b is equal to the proportion of values between a and b, which we compute with the CDF.

Learn About Verified Certificates