

3.8: A Heuristic View of the Cumulative Distribution Function

We can use these ideas to create a plot that approximates the cumulative probability distribution function given any set of N measurements of a random variable u . To do so, we put the u_i values found in our N measurements in order from smallest to largest. We label the ordered values $u_1, u_2, \dots, u_i, \dots, u_N$, where u_1 is the smallest. By the argument that we develop in the previous section, the probability of observing a value less than u_1 is about $1/(N+1)$. If we were to make a large number of additional measurements, a fraction of about $1/(N+1)$ of this large number of additional measurements would be less than u_1 . This fraction is just $f(u_1)$, so we reason that $f(u_1) \approx 1/(N+1)$. The probability of observing a value between u_1 and u_2 is also about $1/(N+1)$; so the probability of observing a value less than u_2 is about $2/(N+1)$, and we expect $f(u_2) \approx 2/(N+1)$. In general, the probability of observing a value between u_{i-1} and u_i is also about $1/(N+1)$, and the probability of observing a value less than u_i is about $i/(N+1)$. In other words, we expect the cumulative probability distribution function for u_i to be such that the i^{th} smallest observation corresponds to $f(u_i) \approx i/(N+1)$. The quantity $i/(N+1)$ is often called the **rank probability** of the i^{th} data point.

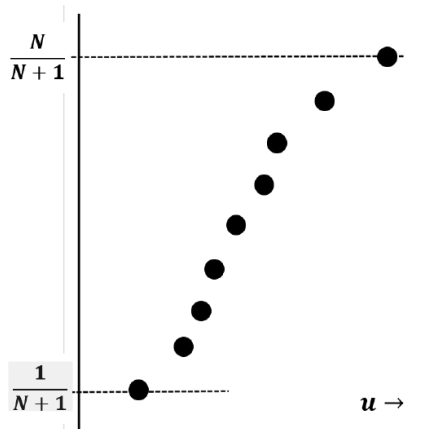


Figure 11. An approximate cumulative probability distribution function.

Figure 11 is a sketch of the sigmoid shape that we usually expect to find when we plot $i/(N+1)$ versus the i^{th} value of u . This plot approximates the cumulative probability distribution function, $f(u)$. We expect the sigmoid shape because we expect the observed values of u to bunch up around their average value. (If, within some domain of u values, all possible values of u were equally likely, we would expect the difference between successive observed values of u to be roughly constant, which would make the plot look approximately linear.) At any value of u , the slope of the curve is just the probability-density function, $df(u)/du$.

These ideas mean that we can test whether the experimental data are described by any particular mathematical model, say $F(u)$. To do so, we use the mathematical model to predict each of the N rank probability values: $1/(N+1), 2/(N+1), \dots, i/(N+1), \dots, N/(N+1)$. That is to say, we calculate $F(u_1), F(u_2), \dots, F(u_i), \dots, F(u_N)$; if $F(u)$ describes the data well, we will find, for all i , $F(u_i) \approx i/(N+1)$. Graphically, we can test the validity of the relationship by plotting $i/(N+1)$ versus $F(u_i)$. If $F(u)$ describes the data well, this plot will be approximately linear, with a slope of one.

In Section 3.12, we introduce the **normal distribution**, which is a mathematical model that describes a great many sources of experimental observations. The normal distribution is a distribution function that involves two parameters, the mean, μ , and the standard deviation, σ . The ideas we have discussed can be used to develop a particular graph paper—usually called normal probability paper. If the data are normally distributed, plotting them on this paper produces an approximately straight line.

We can do essentially the same test without benefit of special graph paper, by calculating the average, $\bar{u} \approx \mu$, and the estimated standard deviation, $s \approx \sigma$, from the experimental data. (Calculating \bar{u} and s is discussed below.) Using \bar{u} and s as estimates of μ and σ , we can find the model-predicted probability of observing a value of the random variable that is less than u_i . This value is $f(u_i)$ for a normal distribution whose mean is \bar{u} and whose standard deviation is s . We can find $f(u_i)$ by using standard tables (usually called the **normal curve of error** in mathematical compilations), by numerically integrating the normal distribution's probability density function, or by using a function embedded in a spreadsheet program, like Excel®. If the data are described by the normal distribution function, this value must be approximately equal to the rank probability; that is, we expect $f(u_i) \approx i/(N+1)$. A plot of $i/(N+1)$ versus $f(u_i)$ will be approximately linear with a slope of about one.

This page titled [3.8: A Heuristic View of the Cumulative Distribution Function](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Paul Ellgen](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.