Visible Probability: User's Manual

Contents

1 Probability Distribution Calculators

Recall that a *probability distribution* defines how likely a random variable is to take on each value in its domain. For example, the amount of rain that will fall in Los Angeles in any particular year is a random variable, and there is a probability distribution defining how likely it is that there will be 1 inch, how likely it is that there will be 2 inches, etc.

Visible Probability includes tools to help you visualize and perform calculations on a variety of popular families of probability distributions. Supported families include *Bernoulli*, *normal*, *binomial*, *negative binomial*, *exponential*, *Poisson*, *logistic*, *geometric*, *hypergeometric*, *Zipf*, *gamma*, *Cauchy*, χ*-square*, *Student's t* distribution, *F* distribution, *Weibull* and *beta*. You can configure the parameters of your distribution to suite your needs or accept the defaults.

1.1 Getting Started

To get started, just open up Visible Probability in your web browser. If you are working with a discrete random variable (i.e. only countably many possible values), you should start at *http://www.covariable.com/discrete.html*. Otherwise you should start at *http://www.covariable.com/continuous.html*.

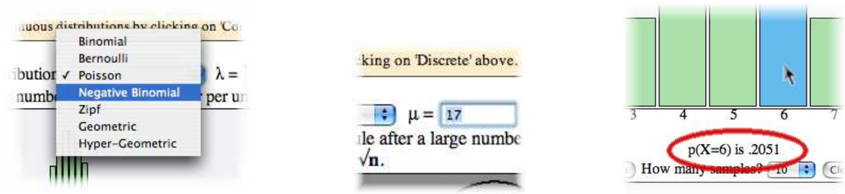

Figure 1: (a) Picking a distribution family, (b) Setting the μ parameter, and (c) Viewing a probability

Next, pick the distribution family you want to study from the popup menu (see Figure 1(a)). The picture below it should change to show the shape of the distribution. If you know specific values for the parameters of the distribution, enter them in the boxes (e.g. if the mean of your normal distribution should be 17, then enter "17" into the box labeled "$\mu =$", as in Figure 1(b)).

1.2 Using the Discrete Distribution Tools

The discrete probability calculator shows a histogram with a bar for each possible value that the random variable X being described can take. The height of the bar shows you the probability that X will take on that value. To see the probability numerically, move the mouse over the bar. The data represented by the bar should appear below the graph (Figure 1(c)).

You can also easily determine the probability that X is within a given range. Simply drag across the bars of interest. If you aren't sure which bar corresponds to a given number, remember that you can see the value represented below the graph whenever you move the mouse over a bar. For example, to see the probability that X is somewhere between 3 and 8 inclusive, just drag the mouse from the 3 bar to the 8 bar. All of the bars in between should hilight blue, indicating that the calculator has selected that range. The information below the graph now indicates the probability that X is between 3 and 8 (or whatever range you selected). To clear the range, click anywhere in the graph.

If you click on the "Sample from this distribution" button, Visible Probability will pick a number of random values distributed according to the distribution you have configured. The values appear in an area at the bottom of the page, and (if you ask for a sufficient number of samples) stack up to help you see which values come up more often (Figure 2). With a sufficiently large sample size, you should be able to confirm that the shape of the

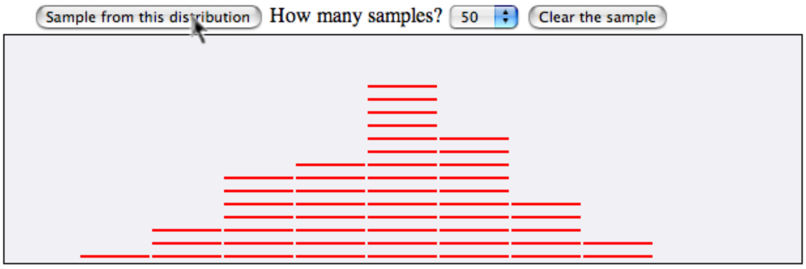

Figure 2: Sampling from a discrete distribution

theoretical histogram on top matches the shape of the sampling histogram on the bottom.

1.3 Using the Continuous Distribution Tools

The continuous probability calculator is very similar to the discrete probability calculator. A continuous random variable is described by a *probability density* curve, which shows which values are more likely than others. Instead of a discrete histogram showing the exact probability of each possible outcome, it plots the probability density curve, showing the relative probability of each possible outcome with respect to other outcomes. For example, with the normal distribution (which is selected by default when you open the continuous probability tools), values near the mean (i.e. 0) have higher probabilities than values far from the mean (Figure 3).

Keep in mind that for a continuous random variable X, any *particular* value has infinitessimally small probability, that is $p(X = 3.7) = 0$ for example, because there are infinitely many possible values. However, the probability that X will be within some *range* is not always 0. Visible Probability lets you select a particular range of your random variable X's domain, and it tells you the probability that X will be within that range by displaying an equation like $p(-2.0 \leq X \leq 2.0) = .95450$ below the graph. The currently selected range is depicted by the dark gray section on the density curve, and is bounded by the dotted vertical lines. You can change the range by dragging the dotted lines left or right, or by typing in exact values in the "Integrate from" and "to" boxes below the curve. You can also change the scale of the graph by entering different "Min" and "Max" values.

A continuous distribution also has associated to it a *cummulative distribution function* or CDF, which shows the probability that X will be less than any particular value. The CDF will only be useful in unusual circum-

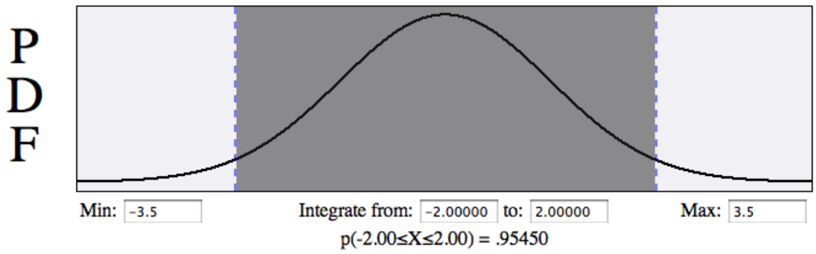

Figure 3: The PDF for a normal distribution, showing an integration.

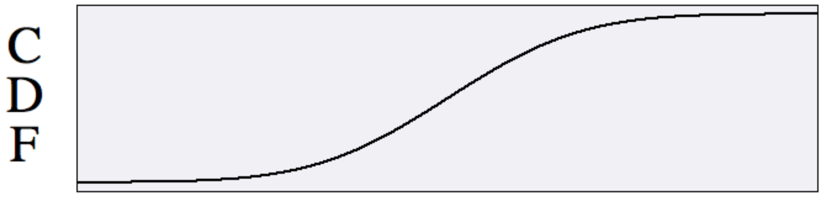

Figure 4: The CDF for a normal distribution.

stances, but Visible Probability includes a plot of the CDF below the main plot (Figure 4).

Like the discrete probability calculator, the continuous calculator also lets you perform random sampling from your distribution. Just click the "Sample from this distribution" button below the CDF, and a set of red lines will appear on the CDF plot showing where the samples fell. Usually you will see that more samples fell in horizontal positions where the PDF is taller, especially if you increase the number of samples. Visible Probability also finds the average of the red sample lines, and plots that as a blue line. The average of the samples is called the *sample mean*, and when the true

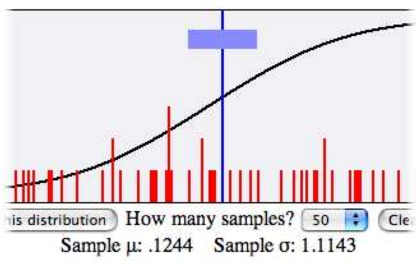

Figure 5: A random sample taken from a normal distribution.

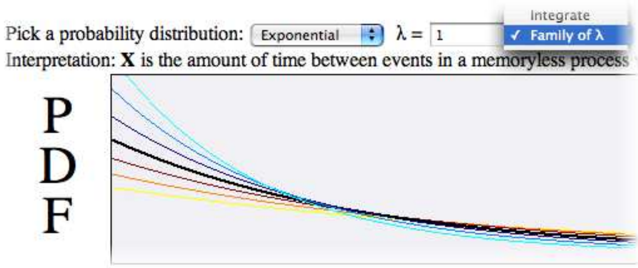

Figure 6: Plotting a family of exponential distributions

mean of the distribution is unknown, the sample mean is often used as an estimate (Figure 5). The blue bar at the top of the sample mean line is called a *95% confidence interval for the mean*, because in 95% of all samples of a given size, the resulting confidence interval will contain the true mean of the distribution. The sample mean and sample standard deviation are displayed numerically below the sample.

To the right of the parameter specifications for your distribution is a popup menu. By default, it is set to the *integrate* mode, which lets you perform the sort of calculations described above for finding the probability that X falls in a particular range. There is also a *family* display mode for each parameter, which shows you how the shape of the distribution depends on the value of that parameter. For example, if you configured an exponential distribution with $\lambda = 1.0$ and then selected "Family of λ" from the popup menu, you would see how the shape of the exponential distribution depends on the λ parameter (Figure 6). Larger λ values are plotted in blue shades, while smaller λ values are plotted in red shades. The thick black line represents the original $\lambda = 1.0$ distribution. To leave the family plotting mode, just select "Integrate" again from the popup menu.

2 Custom Distributions

Often the random variable that you are studying doesn't follow one of the common distribution families built in to Visible Probability. In many cases, however, it is possible to express your random variable in terms of a number of simpler random variables whose distributions *are* built in to Visible Probability. In such situations, the *Bayesian Computation* feature (accessible by clicking *Custom* in the Visible Probability screen, or going directly to *http://www.covariable.com/bayes.html*) makes it easy to visualize the probability distribution of your complex random variable.

2.1 Operation

The Bayesian Computation screen lets you define a number of *independent variables* whose probability distributions are built in to Visible Probability. Then, you can define one or more *dependent variables* whose values are determined as an algebraic formula of the other variables. On the left is the list of variables; click on one to display its definition and probability distribution on the right. If you selected an independent variable, you can edit the parameters of its built-in probability distribution family to suit your needs. If you selected a dependent variable, you can edit its formula. Your formula can be any expression using the four basic arithmetic operations (+, -, /, *), parentheses, the independent variables names (e.g. X and Y), and numbers like 3 or -14.237. For example, your formula for variable Z might be $X/(1 - Y/2)$.

Below the variable's definition, Visible Probability displays the probability distribution curve associated to the selected random variable. For dependent variables, the distribution cannot generally be determined exactly, so Visible Probability computes and displays an approximate histogram. The mean, standard deviation, and mode of the dependent variable are also estimated and displayed.

In addition to showing the names of each variable you have defined, the variable list displays a quick summary of the variable's definition; for dependent variables, it shows the formula, and for independent variables it shows the probability distribution family's name. Initially, Visible Probability sets up two independent and one dependent random variable for you. If you need more of either type of variable, you can create them by clicking the buttons below the variable list.

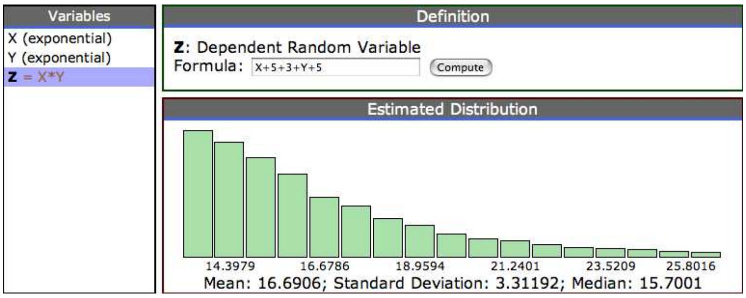

Figure 7: Viewing a custom probability distribution

2.2 Example Custom Distribution

Suppose you live in a busy city and you need to get the post office and back by taxi. You want to predict how long your round trip will take. Taxis pass in front of your house and the amount of time you have to wait follows an exponential distribution with $\lambda = 0.3$ taxis per minute. At the post office it will take 3 minutes to check your PO Box, and each direction driving will take 5 minutes. Taxis pass by the post office more frequently, at a rate of $\lambda = 2.5$ taxis per minute.

To model this scenario in Visible Probability's *Bayesian Computation* mode, you would set up an exponential variable X with $\lambda = 0.3$ representing the time you wait for a taxi from home to the post office, and another exponential variable Y with $\lambda = 2.5$ representing the time you wait for a taxi from the post office back to home. Then you would set up a dependent variable Z and enter $X+5+3+Y+5$ for its formula, since the total time your trip will take is X minutes waiting for a taxi, plus 5 minutes driving, plus 3 minutes at the post office, plus Y minutes waiting, plus 5 more minutes driving. Just push the "compute" button, and Visible Probability will make a histogram showing the probability distribution of the length of your trip! In our example, the expected (mean) time that the round trip would take is 16.69 minutes (about 16 minutes and 41 seconds).