

TRUNCATION

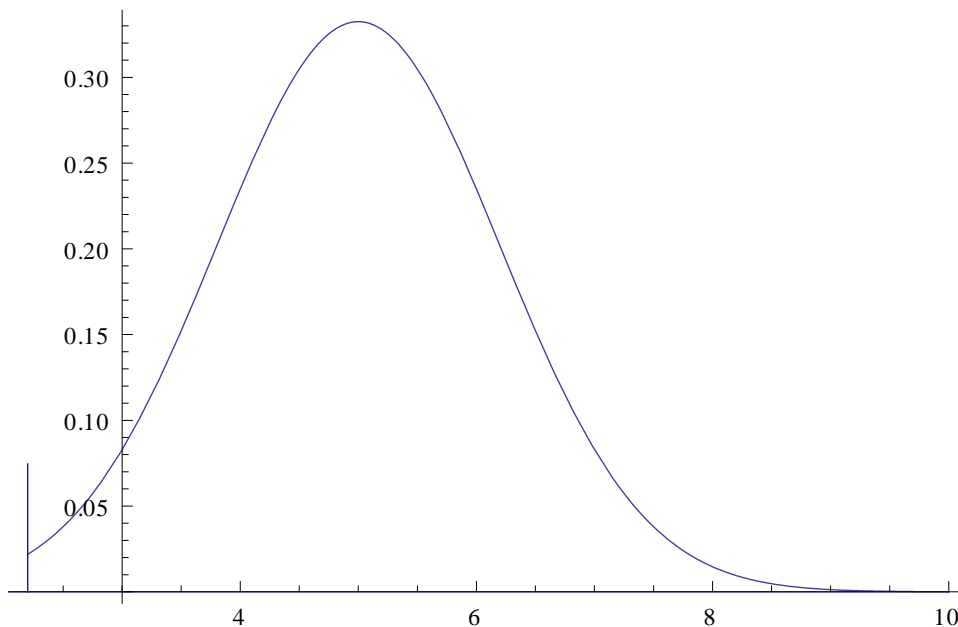
Weber, Kannan, Bordeman 11/18/10



In our discussion of random variables, we have talked about methods of narrowing our focus within the population of interest. One such method that we have discussed in class is censoring. In these notes, we address another method: truncation.

A QUICK REVIEW ON CENSORING

First, let's review what we know about censoring: To censor a population is to collapse the density over a range of the random variable onto a single value of the random variable. For example, if we were analyzing the distribution for annual income in a population, and we considered all families with annual income below the poverty level (say \$22,000) as a single group, we would collapse the density of the pdf for all values less than \$22,000 onto a single value of the random variable – the poverty level, or \$22,000. Therefore, a pile-up would occur at \$22,000 representing the collapsed density of all families with annual income at or below that amount. The whole population is still considered in the censored density function, but we consider all families earning less than the lower limit as one in the same. A graph of this left-censored probability density function might look like this, assuming a income is normally distributed with mean and variance of \$50,000 and \$12,000, respectively, where we left-censor at the poverty level of \$22,000:




Note that Mathematica drew the y-axis at $x=\$30,000$ instead of the origin and that the x-axis is scaled by \$10,000. The pile-up occurs at the poverty level of \$22,000, and the height of the pile-up is equal to the area of the un-censored distribution left of \$22,000.

In our poverty level example, we have a left-censored distribution. We can also have right-censored distributions (see the insurance example in the following section). However, censored distributions need not be only right- or left-censored. Any distribution for which we collapse the values of a range of a random variable onto a single value is a censored distribution. Consider the following example: Assume we have income data for a particular population and that data has some distribution. If we decided to round the income data to the nearest \$10,000, our distribution would be censored, although not necessarily right- or left-censored. Specifically, the density for all values between \$0 and \$4,999 would be collapsed at \$0, the density for all values between \$5,000 and \$14,999 would be collapsed at \$10,000, and so on.

Now that we've reviewed censoring, our task for the remainder of these notes is to discuss another method of refining our distribution called "truncation".

TRUNCATION VS. CENSORING

The literal meaning of truncation is to 'shorten' or 'cut-off' something. Extending this definition to our world of statistics, we can define the truncation of a distribution as a process which results in certain values being 'cut-off,' thereby resulting in a 'shortened' distribution. As with censoring, we can have left- or right-truncated distributions, and we can also have distributions that are truncated in the middle, such as a distribution where we can observe values of the random variable between one and ten, but we truncate to exclude values between two and three (see example 2 under Some Examples of Truncated Distributions). The density of the pdf at all values that are 'cut-off' are omitted from the truncated distribution, and the remaining distribution is shifted upward so that the area beneath it is still one. Note the difference from censoring where we "pile-up" the CDF of the distribution outside of our limit.

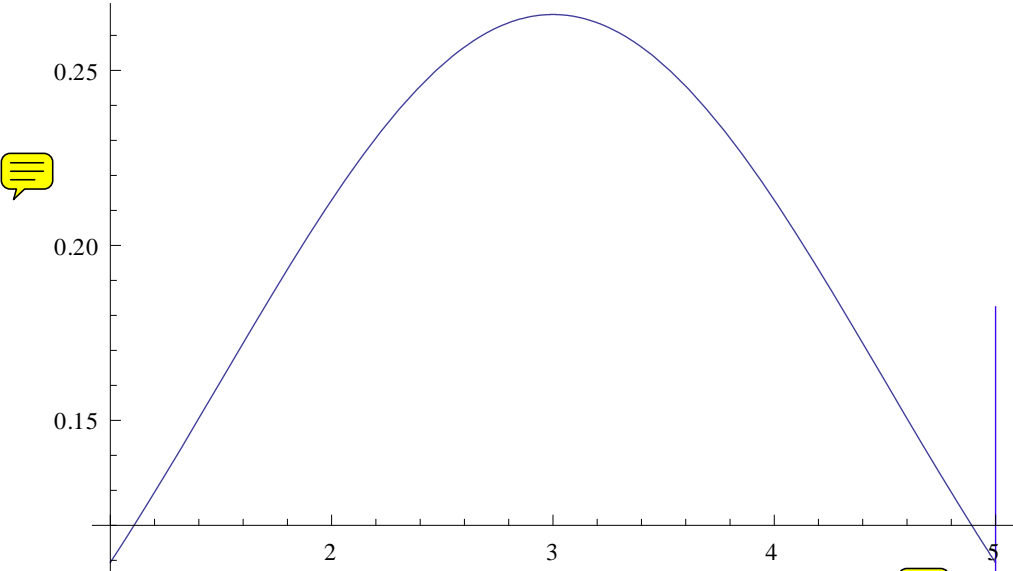
We use  a truncated distribution when certain values within the distribution cannot be observed. Unlike censoring, truncation allows us to shorten our scope of analysis of the distribution and disregard the data outside of the limit.

To see the difference between truncation and censoring, consider the following example in the insurance industry.

Insurance companies levy a policy limit on insurance policy holders. Suppose this limit is \$5,000. Any damages submitted by the policyholder above \$5,000 will be reimbursed by the insurance company as \$5,000 (because of the policy limit). So, the distribution for the random variable of interest (damages paid) is obtained by censoring damage amounts. This is a case of '*right censoring*' of data as the insurer only pays up to the policy limit, even if actual damages exceed that limit. As a result, there is a pile-up at the limit, since all losses exceeding \$5,000 will be reimbursed for \$5,000.

Now suppose the policyholder is subject to a deductible of \$1,000. Any loss incurred below the deductible (\$1,000) will not be reported by the policyholder. Although losses will occur below the deductible limit (\$1,000), the insurance company is unaware of them since the policyholders will not report them. This is a case of 'left truncation' of data, as the insurance company does not observe damages below the limit of \$1,000.

Let's graph a probability density function assuming that damages are normally distributed with mean of \$3,000 and variance of \$1,500. We see that the graph is truncated on the left at \$1,000, since these values are never observed by the insurance company, and we see a pile-up at \$5,000, since the sum of the probabilities of damages exceeding \$5,000 are piled-up at this limit.



Note that Mathematica drew the x-axis at $y=0.12$ and the y-axis at $x=1$ rather than at the origin.

SOME EQUATIONS FOR TRUNCATION

1. If a continuous random variable x , has pdf $f(x)$, then

$$f(x | x > a) = f(x) / \text{Prob}(x > a)$$

Note that the truncated distribution is a conditional distribution.

2. If X is a random variable with density $f_x(\cdot)$ and cumulative distribution $F_x(\cdot)$, then the density of X truncated on the left at a and on the right at b is given by:

$$\frac{f(x) I_{(a,b)}(x)}{F_x(b) - F_x(a)}$$

Note: $I_{(a,b)}(x)$ is an indicator function where

- $I_{(a,b)}(x) = 1$, if $a \leq x \leq b$
- $I_{(a,b)}(x) = 0$, otherwise

MOMENTS OF TRUNCATED DISTRIBUTIONS

We are often interested in the mean and variance of a truncated distribution as they are ways to characterize the distribution. The moments can be obtained using the following formulae:

$$\text{Mean} : E[x | x > a] = \int_a^{\infty} x f(x | x > a) dx$$

$$\text{Variance} : = \int_a^{\infty} [x - E(x | x > a)]^2 f(x | x > a) dx$$

$$3^{\text{rd}} \text{ Moment} : = \int_a^{\infty} [x - E(x | x > a)]^3 f(x | x > a) dx$$

$$4^{\text{th}} \text{ Moment} = \int_a^{\infty} [x - E(x | x > a)]^4 f(x | x > a) dx$$

SOME EXAMPLES OF TRUNCATED DISTRIBUTIONS

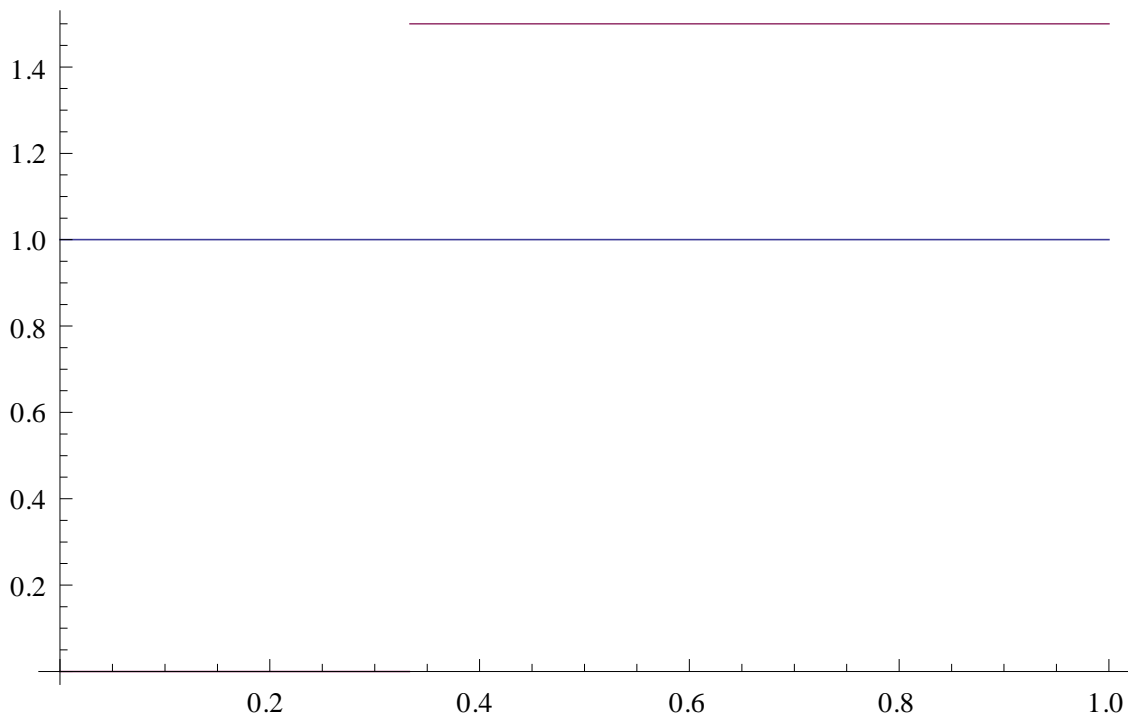
1. Truncated uniform distribution

Suppose x has a standard uniform distribution, $U(0,1)$, then

$$f(x) = 1, 0 \leq x \leq 1.$$

Let the distribution be left truncated at $x=1/3$. The truncated distribution is also uniform.

$$\begin{aligned} \text{Now, } f(x | x > 1/3) &= f(x) / \text{Prob}(x > 1/3) \\ &= (1)/(2/3) \\ &= 3/2, \quad 1/3 \leq x \leq 1 \end{aligned}$$



In the above figure we have plotted the un-truncated uniform distribution (blue line) and the truncated uniform distribution (maroon line).

$$\text{Calculating mean: } E[x | x > 1/3] = \int_{1/3}^1 x(3/2) dx = 2/3$$

The variance for a variable that is distributed uniformly between any 2 value a and b is $(b-a)^2/12$.

$$\text{Therefore variance} = [1-(1/3)]^2/12 = 1/27$$

This example illustrates two results:¹

- A. When the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable will be less than the mean of the original one.
- B. Truncation reduces the variance when compared with the variance of the non-truncated distribution.

¹ Greene, William H. *Econometric analysis*(Pg 759). Fifth Edition. Pearson Education.

2. Another truncated uniform distribution (A SPECIAL CASE)

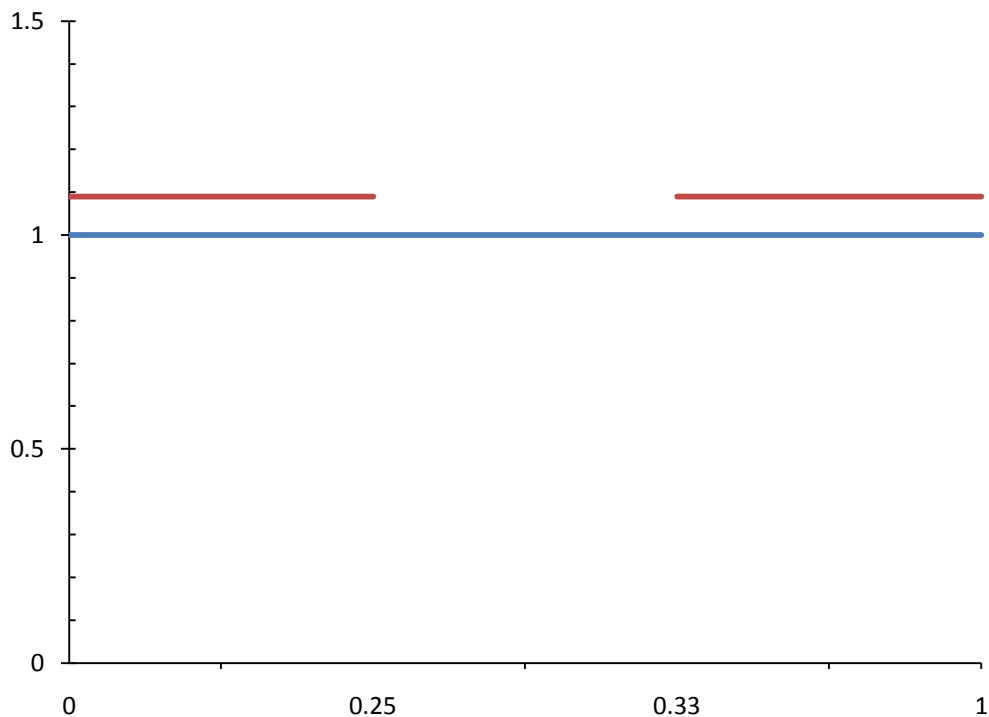
Suppose x has a standard uniform distribution, $U(0,1)$, then $f(x) = 1, 0 \leq x \leq 1$.

Let the distribution now be truncated between $x=1/4$ and $x=1/3$ (The distribution has a probability of zero in that range). The truncated distribution is also uniform. To truncate our uniform distribution in the middle, as opposed to truncating at a bound, we need to divide the PDF by the sum of the CDF from 0 to our lower truncation point and the CDF of our upper truncation point to 1.

$$\text{Now, } f(x) = \begin{cases} \frac{1}{F_x(0 < x < .25) + F_x(.33 < x < 1)} & 0 < x < .25 \text{ \& } .33 < x < 1 \\ 0 & .25 \leq x \leq .33 \end{cases}$$

In this example, $F_x(0 < x < .25) = .25$ and $F_x(.33 < x < 1) = .66$

So the height of our new pdf is: $\frac{1}{.25 + .66}$ in the range $0 < x < .25$ and $.33 < x < 1$



In the above figure we have plotted the un-truncated uniform distribution (blue line) and the truncated uniform distribution (maroon line).

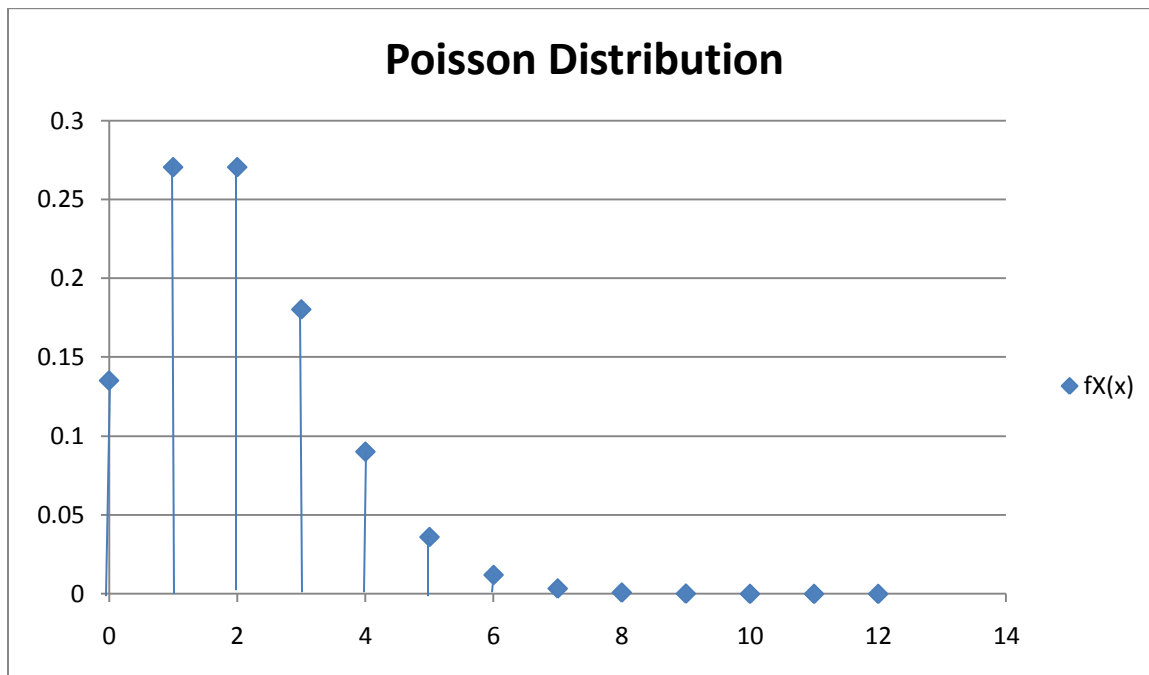
3. Truncated discrete distribution

Let's assume it's an election year and the local government is trying to stimulate economic expansion so it has proposed the following tax break. Families with between four and eight children will receive a tax credit, lowering their taxable income for the year. The credit will be variable based on the number of children in the family; the credit per child will be lower as the number of children in the family increases. The government wants to do an analysis of exactly how much the proposed tax credit will cost, so they want a breakdown of the families with between four and eight children; of these families, what percent have 4 kids, 5 kids, etc.

Since last year was a census year, the government has a pretty accurate count of the total number of local families and has determined that family size can be fairly accurately represented with a Poisson distribution with $\lambda=2$ (Note: In the Poisson distribution, λ represents the mean). Obviously this example will utilize a discrete distribution since one can only have a "whole number" of children. So our Poisson distribution of all of the families looks like:



$$f_x(x) = \frac{(e^{-\lambda}\lambda^x)}{x!} \quad \Longrightarrow \quad f_x(x) = \frac{(e^{-2}2^x)}{x!}$$



Here is our Poisson distribution of all of the families that would be in the jurisdiction of the new tax law. From what we know from discrete distributions, the height of each spike corresponds to the probability of that particular realization of X. For example, we can see that over 50% of the families have either 1 or 2 kids.

Now let's truncate this previous Poisson distribution to only look at families with between (and including) four and eight children. In order to determine exactly what size to make the tax

break at each level, the government is really interested in only the families in the particular range (4-8).

Our truncated Poisson distribution would therefore look like:

$$\text{Prob}(x|4 \leq x \leq 8) = \frac{\frac{(e^{-2}2^x)}{x!}}{\text{Prob}[4 \leq x \leq 8]}$$

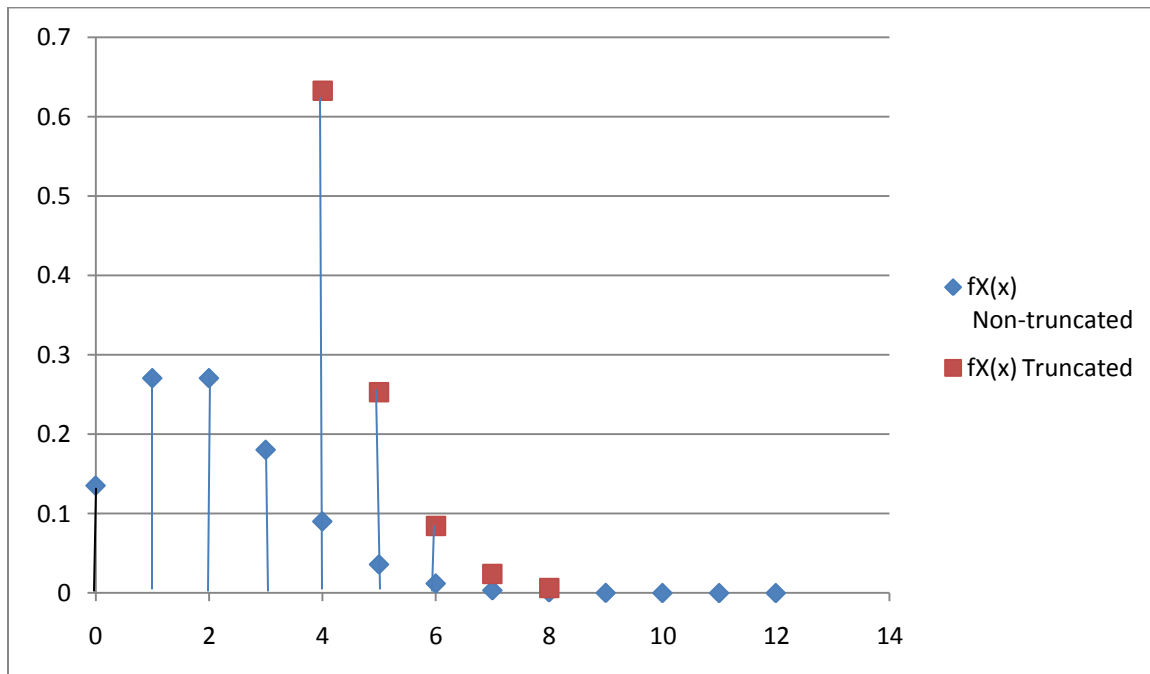
Another method (other than mentioned earlier) to find the Probability that $4 \leq x \leq 8$ in a discrete distribution is to simply take $1 - (\text{sum of the probabilities of } x < 4 \text{ and } x > 8)$ We will round to 4 decimal places for this example...

So from our Poisson, we can find $\sum_{i=0}^3 f_X(x) = .8571$ and $\sum_{i=9}^{+\infty} f_X(x) = .0002$

Therefore, $1 - \text{Pr}[x < 4, x > 8] = .1427$ and our truncated distribution will be calculated as:

$$\text{Prob}(x|4 \leq x \leq 8) = \frac{\frac{(e^{-2}2^x)}{x!}}{.1427}$$

Let's now graph both the original function and our truncation



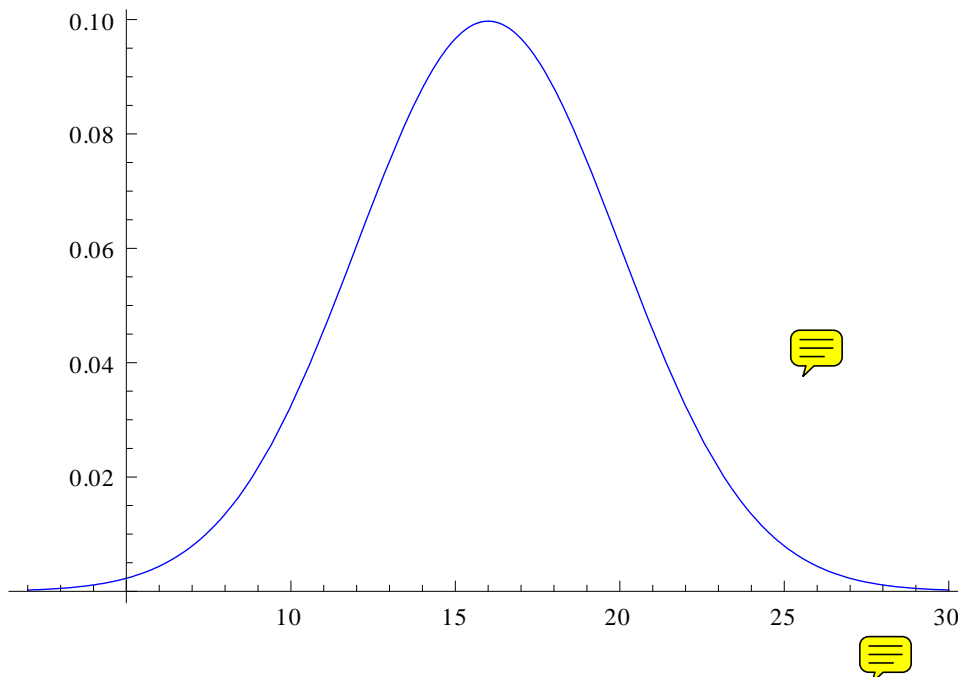
# of Children	Original Distribution	Truncated Distribution
0	0.1353	0
1	0.2707	0
2	0.2707	0
3	0.1804	0
4	0.0902	0.6323
5	0.0361	0.2529
6	0.0120	0.0843
7	0.0034	0.0241
8	0.0009	0.0060
9	0.0002	0
10	0.00004	0
11	0.00001	0
12	0.000001	0

Now we can see the effects of the truncation on the distribution. When we only want to look at families with between four and eight children (inclusive), we can see that families with exactly 4 children make up over 60% of the new population, whereas four child families were less than 10% of the non-truncated distribution. The government can now use the original distribution to get a sense of the population as a whole and the truncated distribution to see the breakdown of the families that will be impacted by the proposed legislation. This could be a starting point in determining the costs/benefits of the tax breaks.

Note that in both discrete distributions the height of each spike in the probability density function is the probability of that particular X value (as mentioned earlier) and the CDF (or the sum of the heights at each x value) totals to 1 (rounded in this example).

4. Truncated continuous distribution

Let's assume an electronics store sells televisions ranging from \$200 to \$3,000, and the prices of all televisions sold has some distribution that resembles the normal, but only has density for values between \$200 and \$3,000. Assume the mean price of televisions sold is \$1,600 and the variance is \$400. The distribution of prices of televisions sold would look like this (note the x-axis is scaled down by \$100):

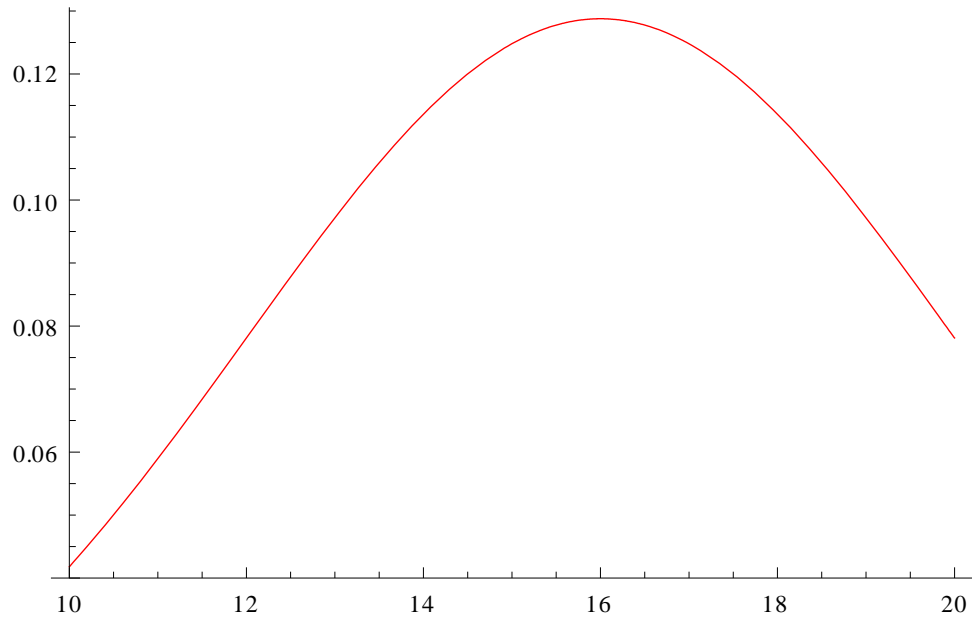


Now let's say the store is offering a holiday special where customers purchasing a television with a selling price over \$1,000 receive a \$100 mail-in rebate and customers purchasing a television with a selling price over \$2,000 receive a \$250 mail-in rebate. We want to analyze the television sales that result in a \$100 rebate being issued, so we must truncate the distribution of prices of televisions sold to include only those that sold for prices over \$1,000 but under \$2,000.

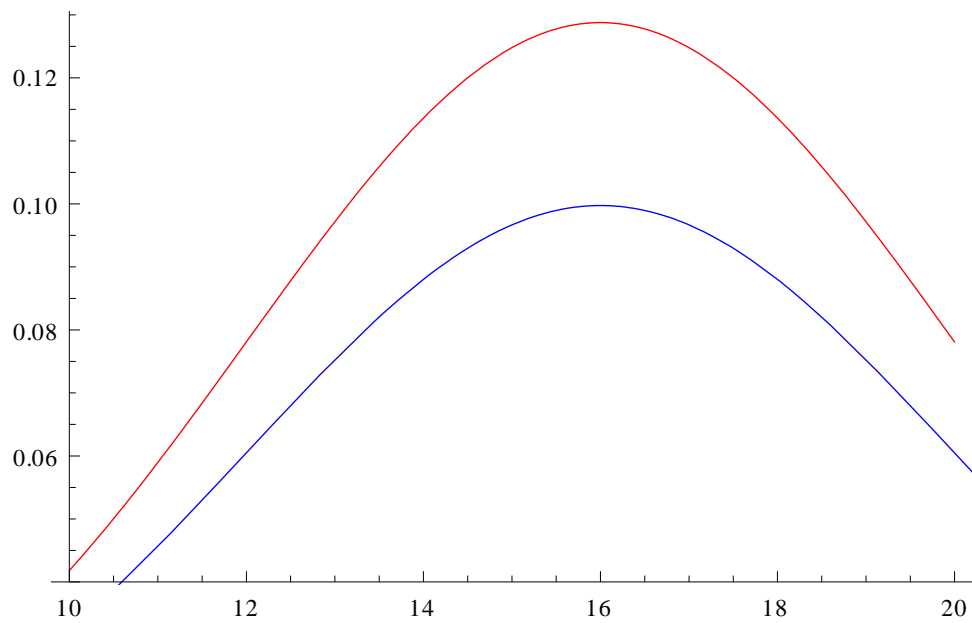
Our distribution that is truncated at \$1,000 on the left and \$2,000 on the right has the following density function:

$$f(x) = f(x;\mu,\sigma) = \frac{\phi_{\mu,\sigma^2}(x) \cdot I_{(1000,2000)}(x)}{\Phi_{\mu,\sigma^2}(2000) - \Phi_{\mu,\sigma^2}(1000)}$$

Where ϕ is defined as the un-truncated probability density function and Φ is defined as the cumulative distribution function. The truncated distribution for the prices of televisions sold between \$1,000 and \$2,000 would look like this:



When we graph both the un-truncated and the truncated distributions on the same set of axes over the range of \$1,000 to \$2,000, we can see that the truncated (red) distribution is shifted up so that the density will still be one after the density of the values less than \$1,000 and greater than \$2,000 has been truncated.



References

1. Greene, William H. *Econometric analysis*. Fifth Edition. Pearson Education.
2. Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. New York: McGraw-Hill, 1973.

SOME RELATED PAPERS OF INTEREST

1. <http://www.math.siu.edu/olive/ch4.pdf>
2. Thomas, Timothy S. *Appendix on Censored Data Analysis*. World Bank, Washington.