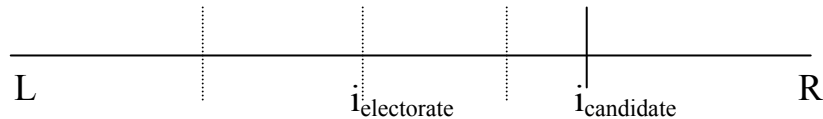


POL 602
Prof. Matthew Lebo
Week 5 – September 27th and 29th, 2005

Continuous Random Variables and Probability Distributions

As you have probably already realized, not all random variables of interest to political scientists (or others) are discrete. Examples could include the length of time that a coalition government lasts, or the “distance” between a candidate’s ideal point and the electorate’s median voter’s ideal point in a simple, 1-dimensional spatial model:



For discrete random variables, we created probability distributions by assigning (by table, graph or formula) a positive probability to each possible value that the variable could assume.

What if we tried to do this for a continuous variable, such as the line segment in the spatial model example? As we try to assign a probability to each point, we discover that there is an infinite number of points.

If there is an infinite number of points, the probability of any one point must be zero.

Therefore it is impossible to assign probabilities to each point without violating the second axiom of probability: that the sum of all probabilities in a space must equal 1.

We must develop a different method to describe the probability distribution of a continuous random variable.

The first idea we need to introduce actually applies to both discrete and continuous random variables: the *cumulative distribution function* (cdf).

The definition of a cdf is as follows:

Let Y denote any random variable. The cdf of Y , written $F(Y)$ is given by:

$$F(y) = P(Y \leq y) \text{ for } -\infty < y < \infty$$

In words, a cdf (often called just a *distribution function* in the text) is a function that tells us, for each value $Y=y$, what is the probability that Y takes on a value less than y .

Example 4.1, p. 151: take a discrete random variable which is distributed binomial, $n = 2$, $p = 0.5$. Find $F(y)$:

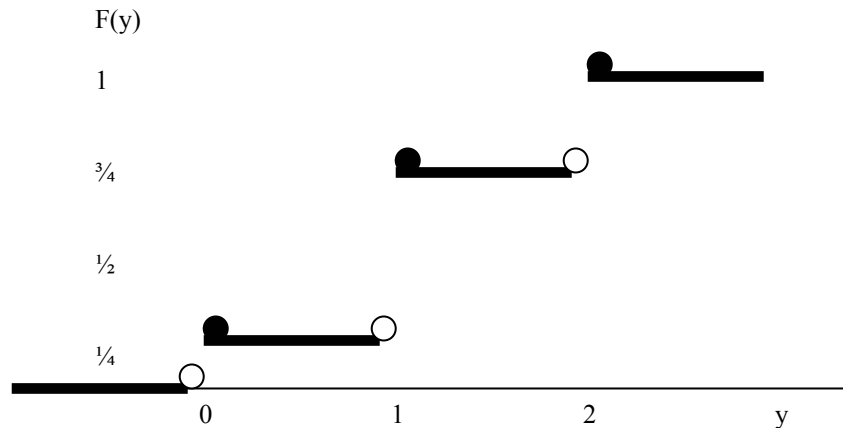
$$p(y) = \binom{2}{y} \cdot .5^y \cdot .5^{2-y}, y = 0, 1, 2$$

Solve the three $p(0,1,2)$ points.

So, $p(0) = 0.25, p(1) = 0.5, p(2) = 0.25$. What, then, is $F(y)$?

$$F(y) = P(Y \leq y) = \begin{cases} 0, & y < 0 \\ 0.25, & 0 \leq y < 1 \\ 0.75, & 1 \leq y < 2 \\ 1, & y \geq 2 \end{cases}$$

Graphically, the cdf of a variable distributed binomially is always an increasing step function, as the cdf can only increase at a countable number of points (it jumps up in discrete “steps”):



There are three general properties of cdf's that are of importance to us:

$$F(-\infty) = \lim_{y \rightarrow -\infty} F(y) = 0$$

$$F(\infty) = \lim_{y \rightarrow \infty} F(y) = 1$$

$F(y)$ is a right continuous, non-decreasing function of y

$$(y_1 < y_2 \Rightarrow F(y_1) \leq F(y_2))$$

Notice each of these features on the graph above.

We can illustrate right continuity by drawing open circles on the right ends of the segments and closed circles on the left sides of the next ones.

What about the *cdf* for a continuous random variable?

Let's take the spatial model example again, and assume that the distance between the candidate and the median voter cannot be less than 0 or greater than d . If we define y_1 and y_2 such that:

$$0 \leq y_1 \leq y_2 \leq 1$$

then it is true that the interval (y_1, y_2) has a positive probability of including Y no matter how small this interval is.

Thus, a *cdf* of a continuous function is smooth and non-decreasing and defined over some interval of real numbers.

Actually, if we reverse this chain of reasoning, we find that the a continuous random variable is actually defined in terms of its distribution function:

Let Y be a random variable with cdf $F(y)$. Y is said to be *continuous* if the distribution function, $F(y)$ is continuous for $-\infty < y < \infty$.

It must also be true that, for any continuous r.v.: $P(Y = y) = 0$

This seems unusual given our intuitive understanding of probability, but it makes sense (what is the probability that you are exactly my height? what does exactly mean? if we can measure only to a certain discrete value, such as the nearest inch, then our measure is not continuous). However, we need another way to talk about the probability of a given interval, since we can not meaningful discuss $P(y = y)$.

This leads us to *probability density functions*, or pdf's.

If $F(y)$ is the cdf for a continuous random variable Y , we can define $f(y)$:

$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

This new function, the derivative of the cdf with respect to y , is the probability density function of Y .

We are trying to find the probability of an interval, $P(y_1 \leq y \leq y_2)$:

$$P(y_1 \leq y \leq y_2) = P(Y \leq y_2) - P(Y \leq y_1) = F(y_2) - F(y_1) = \int_{y_1}^{y_2} f(y) dy$$

Similarly, to find the probability of up to and including an event a , as in the discrete example above, we evaluated the cdf at a ; we took $F(a)$. Alternatively:

$$P(Y \leq a) = F(a) = \int_{-\infty}^a f(y)dy$$

The best way to understand this is:

The probability of a random variable taking on a value in a given interval, then, is equivalent to the area under the probability density function between the points of that interval. This, in turn, is equivalent to the cumulative distribution function evaluated at one point of the interval minus the cdf evaluated at the other end.

Example 4.3 p. 156: Let Y be a continuous random variable with pdf given by

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

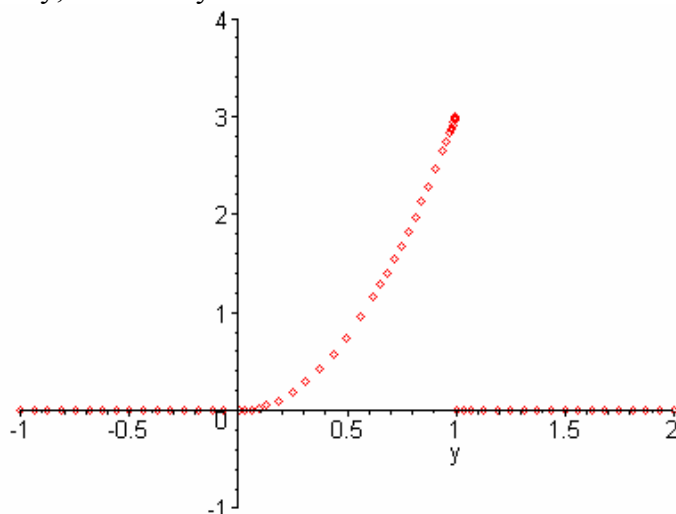
what is F(y)?

First, we can restate the result above in terms of a new variable, t, which will avoid

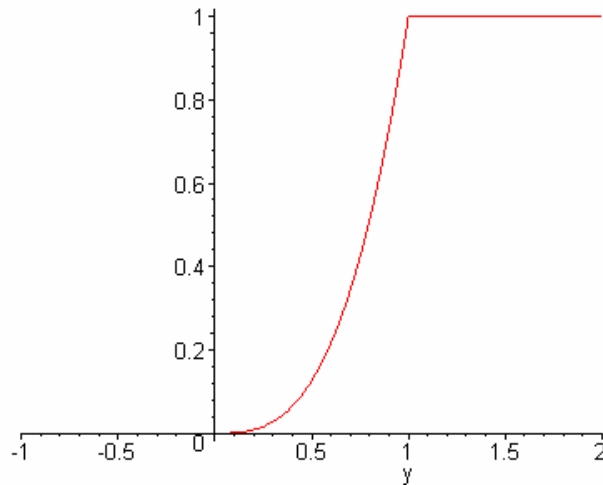
confusion: $F(y) = \int_{-\infty}^y f(t)dt$. Then,

$$F(y) = \begin{cases} \int_{-\infty}^y 0dt = 0, & \text{for } y < 0 \\ \int_{-\infty}^0 0dt + \int_0^y 3t^2 dt = 0 + t^3 \Big|_0^y = y^3, & \text{for } 0 \leq y \leq 1 \\ \int_{-\infty}^0 0dt + \int_0^1 3t^2 dt + \int_1^y 0dt = 0 + t^3 + 0 \Big|_0^1 = 1, & \text{for } y > 1 \end{cases}$$

Graphically, the density function is:



And the distribution function is:



Given this function, what is the probability that Y lies between 1/4 and 1/2?

Solution: $\int_{1/4}^{1/2} 3t^2 dt = t^3 \Big|_{1/4}^{1/2} = \left(\frac{1}{2}\right)^3 - \left(\frac{1}{4}\right)^3 = \frac{7}{64}$

Note how the second step is equivalent to evaluating the cdf at $F(1/2) - F(1/4)$.

Note that, from the axiomatic definition of probability, there are two properties of density functions:

$$f(y) \geq 0, \text{ for any value } y$$

$$\int_{-\infty}^{\infty} f(y) = 1$$

Example (4.4):

Given $f(y) = \begin{cases} cy^2, & 0 \leq y \leq 2 \\ 0, & \text{elsewhere} \end{cases}$, find the value of c that makes f(y) a valid density function:

Solution:

$$F(\infty) = 1 = \int_0^2 cy^2 dy = \left. \frac{cy^3}{3} \right|_0^2 = \frac{8c}{3}$$

$$\frac{8c}{3} = 1$$

$$c = \frac{3}{8}$$

Just as in the case of discrete random variables, we can take the expected value of continuous random variables and of functions of continuous random variables to calculate useful quantities such as the mean and variance.

Recall that the expected value of a discrete random variable is given by:

$$E(Y) = \sum_y yp(y)$$

The analogous formula for a continuous random variable is:

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy$$

Looking at the formula, this should make intuitive sense. The integral is the continuous analog of the summation (recall integral theory—the sum of many small rectangles).

Note, however, that the expectation operation is not possible for all functions, (i.e.

$\int_{-\infty}^{\infty} |y|f(y)dy < \infty$ must hold for $E(Y)$ to exist) and that often, even if existence is proven, analytical evaluation of the integral is impossible and it must be evaluated numerically.

Similarly, we can find the expected value of a function of continuous random variable Y :

$$E[g(y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

And, as is the case for discrete r.v.'s, the following results hold:

$$E(c) = c$$

$$E[cg(Y)] = cE[g(Y)]$$

$$E[g_1(Y) + g_2(Y) + \dots + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + \dots + E[g_k(Y)]$$

$$V(Y) = E(Y^2) - \mu^2$$

As an example, let us return to the probability density function we computed in the last example, $f(y) = \frac{3}{8}y^2, 0 \leq y \leq 2$ and compute its mean and variance:

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy = \int_0^2 y \frac{3}{8} y^2 dy = \left. \frac{3}{8} \frac{1}{4} y^4 \right|_0^2 = 1.5$$

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y)dy = \int_0^2 y^2 \frac{3}{8} y^2 dy = \left. \frac{3}{8} \frac{1}{5} y^5 \right|_0^2 = 2.4$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = 2.4 - (1.5)^2 = 0.15$$

Some Useful Distributions:

The Uniform Distribution (again)

As in the case of discrete random variables, the first continuous random variable that we will discuss is the Uniform probability distribution (sometimes called the Uniform on an Interval to distinguish it from the Uniform on Integers which is discrete).

Remember our example from the discussion of the Poisson distribution: the number of bills reported out of the House Armed Services committee.

Let us assume further that, in a given 10 hour work day in the session, the probability that a bill is reported out of committee in any subinterval of the 10 hours is equally likely.

That is, the bill is as likely to be reported between 9:30 and 10:30 as it is between 11:30 and 12:30.

Let Y denote a random variable representing the length of time we must wait during the day for the bill to be reported out. This is a (somewhat contrived) example of a variable with a continuous uniform probability distribution.

More formally, given an interval (θ_1, θ_2) such that $\theta_1 < \theta_2$, a random variable Y is said to have a uniform distribution on that interval if and only if the density function of Y is given by:

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \leq y \leq \theta_2 \\ 0, & \text{elsewhere} \end{cases}$$

(Note that θ_1 and θ_2 are *parameters* of this distribution, just as n and p are the parameters of the Bernoulli distribution. Parameters are constants that determine the specific form of the distribution function). Graphically:



Given this equation, what is the probability that the bill is reported out in the last hour of the 10-hour working period?

$$P(9 \leq Y \leq 10) = \int_9^{10} \frac{1}{10-0} dy = \left. \frac{y}{10} \right|_9^{10} = \frac{10-9}{10} = \frac{1}{10}$$

What about the mean and variance of a uniform random variable?

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} yf(y)dy = \int_{\theta_1}^{\theta_2} y \frac{1}{\theta_2 - \theta_1} dy \\ &= \left. \frac{y^2}{2} \frac{1}{\theta_2 - \theta_1} \right|_{\theta_1}^{\theta_2} = \frac{\theta_2^2 - \theta_1^2}{2(\theta_2 - \theta_1)} = \frac{\theta_2 + \theta_1}{2} \end{aligned}$$

The variance is similarly easy to calculate (and is left as an exercise):

$$\sigma^2 = V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}$$

Example: What, then, is the mean amount of time we should expect to wait before a bill is reported out of the committee in our example? What about the variance?

$$\begin{aligned} \mu = E(Y) &= \frac{10+0}{2} = 5 \text{ hours} \\ V(Y) &= \frac{(10-0)^2}{12} = \frac{100}{12} = 8.333 \text{ hrs} \end{aligned}$$

This variance makes sense, as the “spread” around the mean of a uniform distribution is quite large given the shape of the distribution function.

The Normal Distribution

We now have enough of a background to discuss more formally a distribution that we have been referring to frequently along the way: the Gaussian or Normal distribution. This is a very important distribution for many reasons, among them the fact that many naturally and socially occurring data appears to be distributed approximately normal, or in the familiar “bell curve.”

A random variable Y is said to have a normal probability distribution if for parameters $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

The mean and variance are simply:

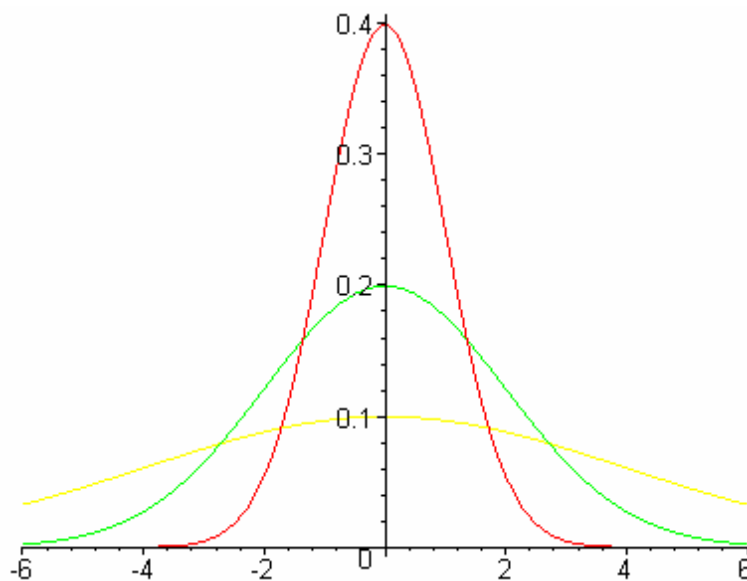
$$E(Y) = \mu$$

$$V(Y) = \sigma^2$$

Note: these are proven in 4.9, which we may cover in late October.

These (μ and σ) are the only two parameters needed to define the shape of this distribution:

Three normal distributions with mean 0 and s.d. 1, 2 and 4.



If we wish to find the probability of a given interval $[a,b]$ we should expect to integrate this pdf over the interval.

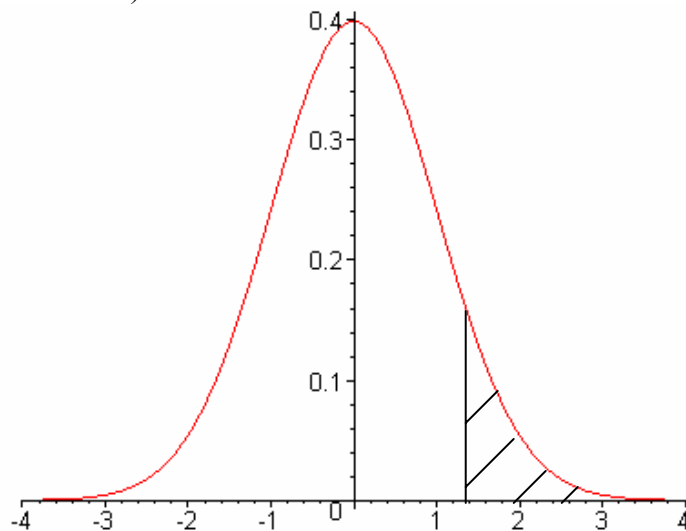
However, there is no closed-form, analytic solution, so the integration must be done numerically. In practice, until computing speed caught up with this problem recently, these computations were given in tables, an example of which is in the back of the text (Table 4, Appendix III).

Since it is impractical to provide many tables, we usually must convert our normal distribution of interest into the so-called “standard normal” which has a mean of 0 and a standard deviation of 1.

This is done by transferring our random variable Y into a new standard variable Z with the following formula:

$$Z = \frac{Y - \mu}{\sigma}$$

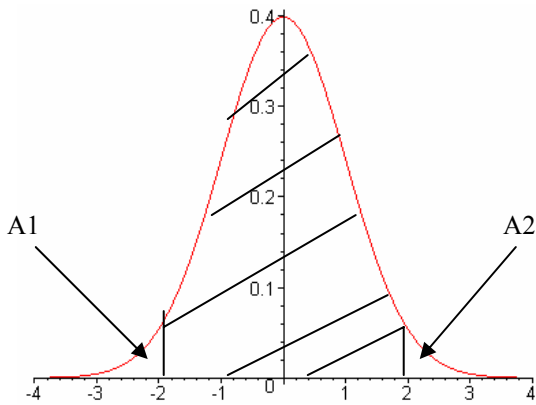
Note also that, since the normal distribution is symmetric with respect to the mean, most (including our) tables only provide areas on one side of the distribution. Our table, specifically, gives areas to the right of point z (the distance from the mean in std. deviations):



Example (4.8): If Z is a standard normal random variable $(0,1)$ find:

a) $P(Z > 2)$. Go to page 792. This value is 2 standard deviations from the mean, so go down the first column until you get to 2 → value for 2.00 is .0228

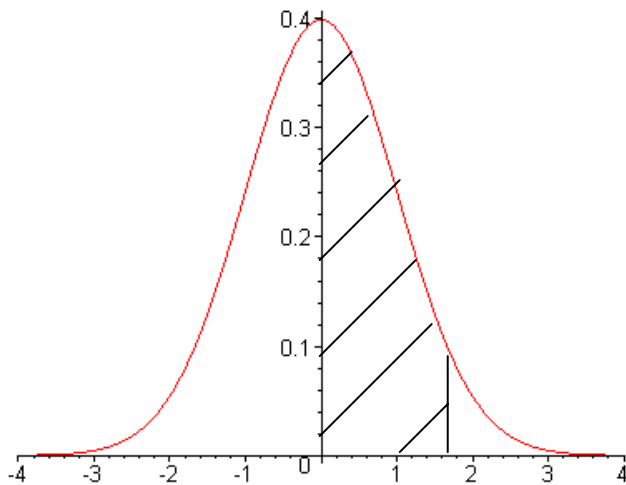
b) $P(-2 \leq Z \leq 2)$:



We are looking for the area $1 - A_1 - A_2$. We know from the previous problem that $A_2 = .0228$ and that, by symmetry, $A_1 = A_2$. Therefore:

$$P = 1 - A_1 - A_2 = 1 - 2A_2 = 1 - 2(.0228) = .9544$$

c) $P(0 \leq Z \leq 1.73)$:



First, we can find $A(1.73)$ from the table as before. It is .0418. But we are looking for the area between 0 and 1.73. Since the total area equals 1.0, half the area must equal 0.5.

Thus, our answer is:

$$P = 0.5 - 0.0418 = 0.4582$$

A second example illustrates what to do when the random variable is not given in standard form and must be transformed:

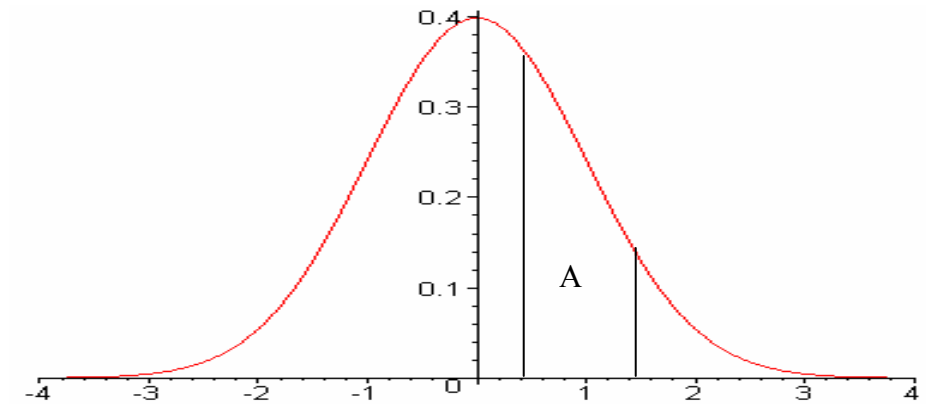
Example: (4.9) Achievement scores for a college entrance exam are assumed to be distributed normal with mean 75 and s.d. 10. What fraction of the scores lies between 80 and 90?

First, transform 80 and 90 to standard units or “z-scores” as some of you may be familiar with:

$$z_1 = \frac{y - \mu}{\sigma} = \frac{80 - 75}{10} = .5$$

$$z_2 = \frac{y - \mu}{\sigma} = \frac{90 - 75}{10} = 1.5$$

So we want to find the area between .5 and 1.5 standard normal units:



Calculate this from the table:

$$A(.5) - A(1.5) = .3085 - .0668 = .2417$$

As final note, the p.d.f. of the normal distribution is often given the symbol ϕ (small phi).

The c.d.f. of the normal (the equation for which was not presented) is capital phi: Φ .

[Note: $\Phi = \int_{-\infty}^y \phi(u) du$]

Example: In many states a motorist is legally drunk or driving under the influence (DUI), if his or her blood alcohol concentration, Y , is 0.10% or higher. When a suspected DUI offender is pulled over, police often request a sobriety test. Although the breath analyzers used are remarkably precise, the machines do exhibit a certain amount of measurement error. Because of that variability, the possibility exists that a driver’s *true* blood alcohol concentration may be under 0.10% even though the analyzer gives a reading over 0.10%.

Experience has shown that repeated breath analyzer measurements taken on the same person produce a distribution of responses that can be described by a normal pdf with μ equal to the person’s true blood alcohol concentration and σ equal to 0.004%. Suppose a driver is stopped and has a *true* blood alcohol concentration of 0.095%, just barely

under the legal limit. If he takes a breath analyzer test, what are the chances that he will be incorrectly booked on a DUI charge?

Since a DUI arrest occurs when $Y \geq 0.10\%$, we need to find $P(Y \geq 0.10)$ when $\mu = 0.095$ and $\sigma = 0.004$ (the percentage is irrelevant to any probability calculation and can be ignored). An application of the Z transformation shows that the driver has almost an 11% chance of being falsely accused:

$$\begin{aligned}
 P(Y \geq 0.10) &= P\left(\frac{Y - 0.095}{0.004} \geq \frac{0.10 - 0.095}{0.004}\right) & P(Y \geq 0.10) &= P\left(\frac{Y - 0.095}{0.004} \geq \frac{0.10 - 0.095}{0.004}\right) \\
 &= P(Z \geq 1.25) = 1 - P(Z < 1.25) & &= P(Z \geq 1.25) = 1 - P(Z < 1.25) \\
 &= 1 - 0.8944 & &= 1 - 0.8944 \\
 &= 0.1056 & &= 0.1056
 \end{aligned}$$

The Gamma Distribution

This distribution is useful for modeling random variables that are always non-negative and are skewed to the right. Examples might include duration of events in continuous time (such as armed conflicts, stability of parliamentary governments, etc).

A random variable Y is said to have a gamma distribution with parameters α, β (both >0) if the density function is given by:

$$f(y) = \begin{cases} \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}, & 0 \leq y < \infty \\ 0, & \text{elsewhere} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

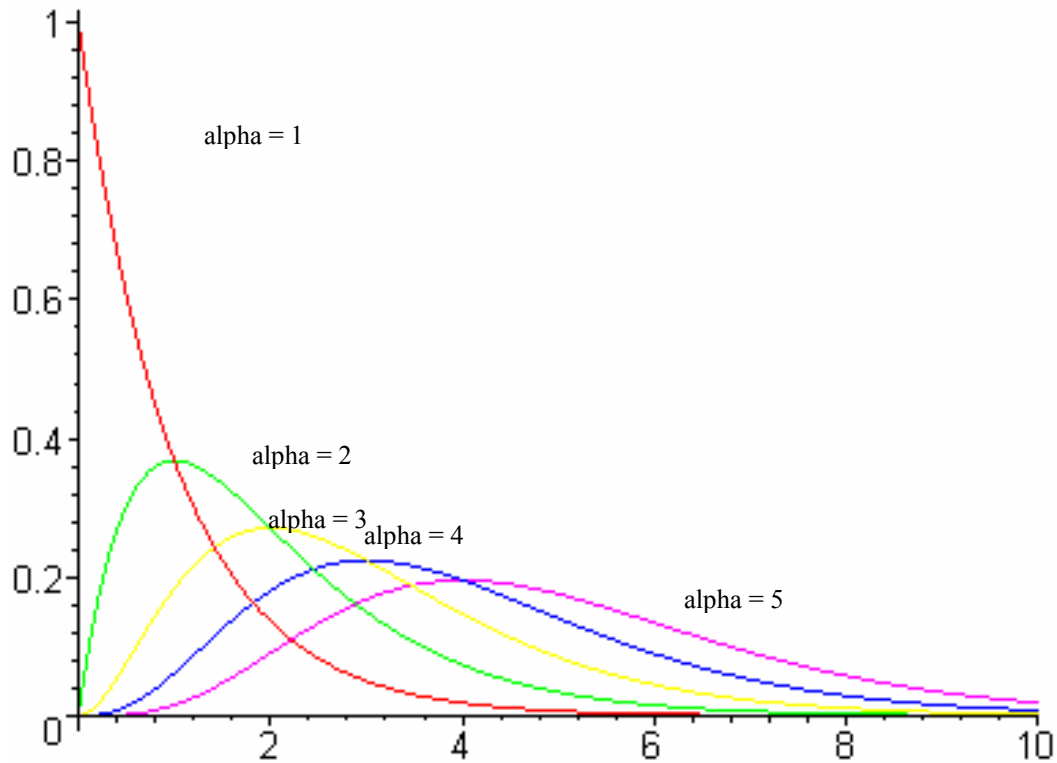
α, β are known as the shape and scale parameters, respectively. The gamma function defined above can be solved for positive integers via a computational shortcut:

$$\Gamma(1) = 1$$

$$\Gamma(n) = (n-1)! \quad \forall \text{ positive integers } n$$

(you can verify this by integrating the function).

As an example, here are gamma distributions with different shape parameters, betas held constant at 1.0:



The mean and variance of gamma-type distributions are given by:

$$\mu = E(Y) = \alpha\beta$$

$$\sigma^2 = V(Y) = \alpha\beta^2$$

(Proofs are in the text).

A special case of the Gamma distribution is the Chi-Square distribution (χ^2) which plays a central role in much of the statistical theory we will study later.

A Chi-Square is a Gamma with parameters: $\alpha = \nu/2, \beta = 2$. The only parameter, then, is ν which is often referred to as “degrees of freedom.”

The mean and variance of an RV distributed Chi-Square are given by:

$$\mu = E(Y) = \nu$$

$$\sigma^2 = V(Y) = 2\nu$$

Another special case of the gamma is when alpha is equal to 1. This is called the *exponential distribution function* with form:

$$f(y) = \begin{cases} \frac{1}{\beta} e^{-\frac{y}{\beta}}, & y \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

This is a useful distribution for many physical phenomena and engineering applications. In political science, an application would be modeling the duration of something for which the probability of “surviving” for an additional time period is not conditional on how long it has survived to date (a so-called *memoryless* process).

The mean and variance are given by:

$$\begin{aligned} \mu &= E(Y) = \beta \\ \sigma^2 &= V(Y) = \beta^2 \end{aligned}$$

As an example of an exponential distribution: if we assume that the mean number of individuals failing in a large group under a certain situation is 5, what is the probability that more than 20 will fail at the same time?

$$P(Y > 20) = \int_{20}^{\infty} \frac{1}{5} e^{-\frac{y}{5}} dy = -e^{-\frac{1}{5}y} \Big|_{20}^{\infty} = -e^{-\frac{1}{5}\infty} - (-e^{-\frac{20}{5}}) = 0 + e^{-4} \approx .0183$$

Here’s the math trick to do that:

To get the antiderivative for a simple exponential function use this rule:

$$\int A e^{kx} dx = \frac{A}{k} e^{kx}$$

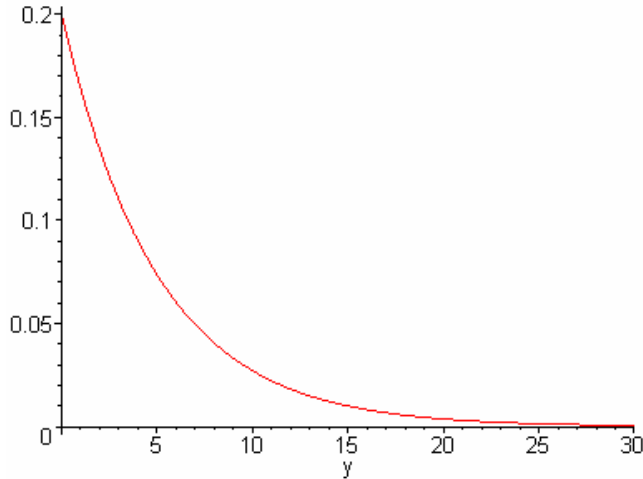
Here, $k = -\frac{1}{5}$ and $A = \frac{1}{5}$ in $\int \frac{1}{5} e^{-\frac{1}{5}y} dy$

$$= \frac{\frac{1}{5}}{-\frac{1}{5}} e^{-\frac{1}{5}y} = -e^{-\frac{1}{5}y}$$

And:

$$-e^{-\frac{1}{5}\infty} = -e^{-\frac{\infty}{5}} = -e^{-\infty} = -\frac{e}{\infty} = -0 = 0$$

[Note that this is actually a discrete problem, but we are using a continuous distribution as an approximation]. This function looks like:



Beta Distribution

The final continuous probability distribution we will cover is the beta. This distribution is defined only over $0 \leq y \leq 1$, and is thus an excellent way to model proportions, such as the proportion of a legislature that belongs to one party, or the proportion of a budget spent on servicing the national debt.

This one looks very confusing, but it is very useful. The density function is given by:

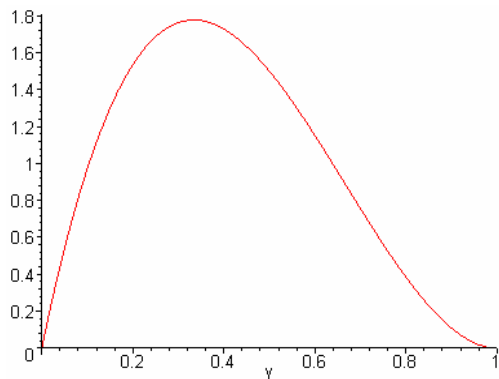
$$f(y) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, & \alpha, \beta > 0; 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where

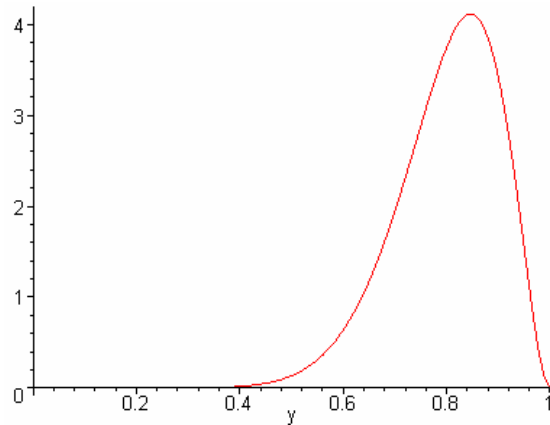
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

[The bottom function is the Beta function. Note that, when alpha and beta are positive integers, the cdf (called the incomplete beta function) reduces to the sum of binomial probabilities].

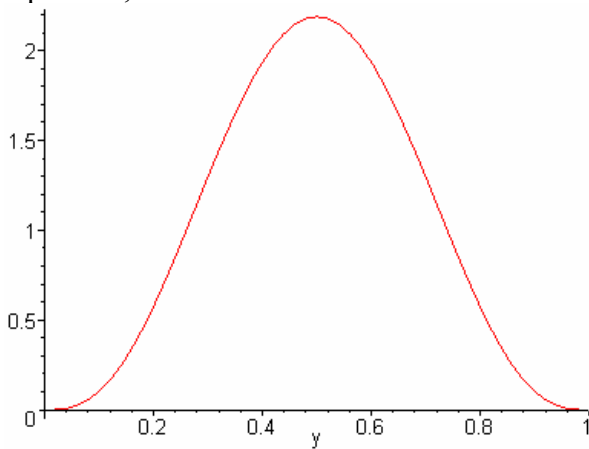
Graphically, beta distributions look like:



alpha=2, beta=3



alpha=12, beta=3



alpha=4, beta=4

The mean and variance of beta distributions are easy to calculate (and not too difficult to prove):

$$\mu = E(Y) = \frac{\alpha}{\alpha + \beta}$$

$$\sigma^2 = V(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Example (4.11 in book):

A gasoline distributor has tanks filled every Monday. The proportion of the supply sold every week is modeled by a Beta distribution, with $\alpha = 4$ and $\beta = 2$. What is the probability that the owner will sell at least 90% in a given week?

$$\begin{aligned}
P(Y > .9) &= \int_{.9}^1 \frac{y^3(1-y)dy}{B(4,2)} \\
P(Y > .9) &= \int_{.9}^1 \frac{\Gamma(4+2)}{\Gamma(4)\Gamma(2)} y^3(1-y)dy \\
&= \int_{.9}^1 \frac{5!}{3!1!} y^3(1-y)dy \\
&= \int_{.9}^1 20(y^3 - y^4)dy = 20 \left[\frac{y^4}{4} - \frac{y^5}{5} \right]_{.9}^1 = 20(.004) = .08
\end{aligned}$$

Recommended problems: 4.3, 4.4, 4.10, 4.14, 4.19, 4.31, 4.36, 4.46, 4.54, 4.60, 4.65*, 4.69, 4.90, 4.91, 4.94, 4.100

Additional problems: 4.27, 4.37, 4.38, 4.50, 4.58, 4.72, 4.73, 4.74, 4.86, 4.92.