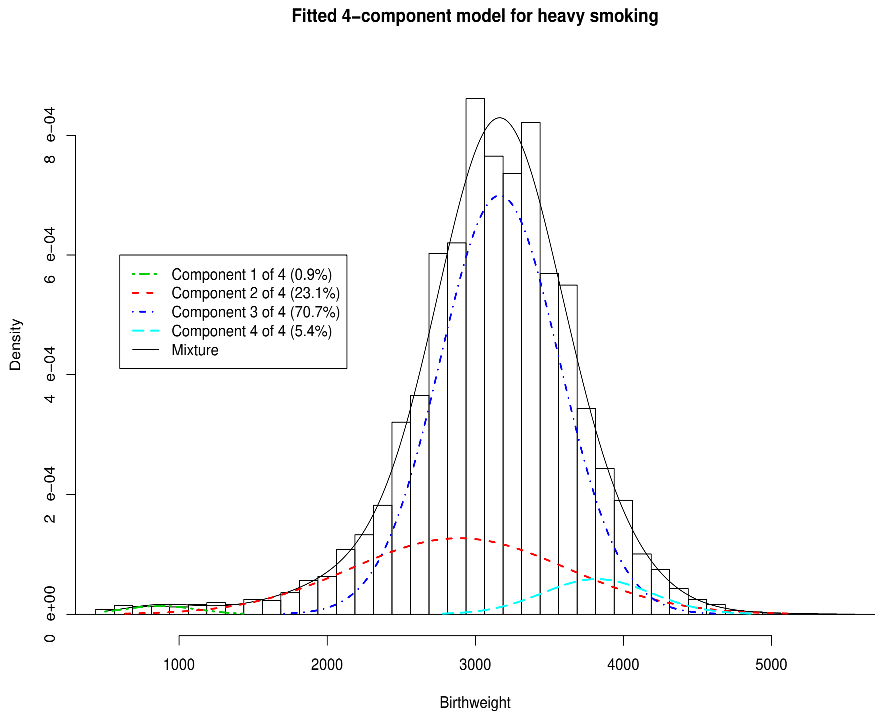
BST 675 – Fall 2010 – Dr. Charnigo

Unit III: Continuous Random Variables

a. Motivating Case Study #1: Describing a distribution of birthweights

One of my research areas in statistics is mixture modeling, which I have applied to the problem of describing a birthweight distribution and its connection to infant mortality. If interested, feel free to look at two papers that my colleagues and I have published, {http://www.biomedcentral.com/1471-2393/10/37} and {http://www.biomedcentral.com/1471-2393/10/44}. This motivating case study will touch on only a few basic ideas from the first paper.

Figure 1:



The figure above presents a histogram of birthweights for 50,000 singleton infants born to white mothers who smoked heavily during pregnancy. One way to interpret the histogram is as (the graph of) a probability mass function for the random variable of birthweight after rounding to the nearest 100 grams. Indeed, such a random variable is discrete because its support set is finite, consisting of multiples of 100 grams from (say) 500 to 5500 grams.

For instance, the probability that a randomly selected singleton infant (born to a white mother who smoked heavily during pregnancy) has a birthweight of 2800 grams after rounding to the nearest 100 grams is approximately 0.06. Visually, this corresponds to the area of the histogram bar centered at 2800 grams, whose width is 100 (the difference between 2850 grams and 2750 grams) and whose approximate height is $6 \times 10^{-4} = 0.0006$.

However, regarding birthweight as a discrete random variable in this manner seems rather artificial. Indeed, apart from some need for discretization in order to construct a histogram, there seems no compelling reason to treat 2849 grams and 2851 grams as different while treating 2849 grams and 2751 grams as the same. Instead, regarding birthweight as a continuous random variable seems far more natural.

Of course, a philosophical caveat applies here. Since the precision with which birthweight can be measured is limited, perhaps in practice birthweight truly is a discrete random variable, albeit one with a much larger support set than suggested previously, consisting of multiples of 1 gram rather than multiples of 100 grams. But what if we had "perfect" measuring instruments? Even that might not resolve the issue because of Heisenberg's Uncertainty Principle from quantum physics, which roughly stated says that an object's position and momentum are inherently (i.e., not because of our imperfect measuring instruments) defined only to a limited precision.

Be that as it may, we will adopt the perspective that a continuous random variable provides an acceptable and, indeed, tractable probabilistic model for birthweight. Once we have come to this agreement, the question becomes what specific probabilistic model should be employed.

Although we have yet to formally introduce a normal distribution in BST 675 (we will do so later in Unit III), you are already familiar with the basic idea. So, if you look at the figure and attempt to smooth out the choppiness of the histogram, as I have done by superimposing the black solid curve, your initial reaction may be that birthweight can be described by a normal distribution.

Yet, if you look more closely, you will see that a normal distribution is a rather crude approximation. The mean of the normal distribution would be about 3150 grams. Recalling that a normal distribution is symmetric about its mean, one would then anticipate a comparable number of births in the vicinity of 5150 grams as in the vicinity of 1150 grams. This is clearly not the case, though, as there are more births near 1150 grams than near 5150 grams. The spread of birthweights, although vaguely bell shaped, is slightly left skewed.

How I modeled birthweight, then, and why I modeled it the way I did, will be addressed at the end of Unit III.

b. Motivating Case Study #2: Modeling a distribution of times to smoking recidivism

Quitting smoking is notoriously difficult. Yet, because of the very serious health issues faced by smokers as well as the bystanders whom they subject to secondhand smoke, a priority for public health practitioners and behavioral science researchers is to develop programs (e.g., counseling) and/or technologies (e.g., nicotine replacement) that may help smokers quit.

If we wish to analyze smoking recidivism data, one approach is simply to identify a fixed time point (such as six months) and then define a dichotomous dependent variable to indicate whether a person remained abstinent from cigarettes at least until that time point. This approach is limited, though, because it does not distinguish someone who "almost" made it until that time point (say, someone who remained abstinent for five and a half months) from someone who did not come close (say, someone who resumed smoking within one week). [Question: Since both of these individuals ultimately failed to quit smoking, why should we wish to make a distinction?]

Thus, we may instead prefer to analyze smoking recidivism data by defining a time-to-event dependent variable that indicates the length of time for which a person remained abstinent. Of course, we then face the possibility of right censoring, but that can be addressed in the context of the proportional hazards regression model frequently employed in conjunction with a time-to-event dependent variable. If you have not already encountered the proportional hazards regression model, you will see it in BST 760 and/or BST 761. My aim with this motivating case study is comparatively modest. I want to provide some intuition for an upcoming definition of a hazard function and, when the time comes to resolve the motivating case study at the end of Unit III, some considerations relevant to the selection of a probabilistic model for a continuous random variable.

So, let the continuous random variable T denote the length of time until someone resumes smoking. [One may object that some people will never resume smoking, in which case T appears to equal $+\infty$ with positive probability and hence does not meet the definition of a continuous random variable. This is a valid objection, but we can circumvent it by re-defining T as, say, the length of time that a person lives smoke free. Then, since everyone dies eventually, T is always finite.] Consider the conditional probability

$$P(T \leq a + b \mid T > a),$$

where a and b are nonnegative real numbers. In words, this is the probability that a person resumes smoking in the next b time units given that he/she had remained abstinent for more than a time units.

Thinking about how this conditional probability depends on a and b is informative. First, for a fixed b, we anticipate that this conditional probability will be a decreasing function of a. If you have been abstinent for six months, you are less likely to resume smoking in the near future than if you have been abstinent for only one month. Second, for a fixed a, this conditional probability is necessarily an increasing function of b. To see that, write it as

$$\frac{P(a < T \leq a + b)}{P(T > a)} = \frac{F_T(a + b) - F_T(a)}{1 - F_T(a)},$$

where $F_T(\cdot)$ denotes the cumulative distribution function of T. Since a cumulative distribution function is necessarily nondecreasing, $F_T(a + b)$ and hence the entire quotient must become larger as b increases. Intuitively, you are more likely to resume smoking in the next two months than you are in the next one month.

I will resume this motivating case study at the end of Unit III.

c. Probability density functions and hazard functions (Cf. pp. 161-173 of Larsen and Marx)

In Unit II we learned that X was called a continuous random variable if its cumulative distribution function $F_X(x)$ was continuous. In practice, most continuous cumulative distribution functions are also differentiable. That is,

$$F'_X(x) := \lim_{\delta \to 0} [F_X(x+\delta) - F_X(x)]/\delta$$

exists for (almost) all x. Moreover, in practice, such a derivative is itself typically a continuous function at (almost) all x. If this is indeed the case, then the Fundamental Theorem of Calculus provides

$$F_X(b) - F_X(a) = \int_a^b F'_X(x)\, dx \tag{1}$$

for $-\infty \le a < b \le \infty$. Here $F_X(\infty) := \lim_{x\to\infty} F_X(x) = 1$ and $F_X(-\infty) := \lim_{x\to -\infty} F_X(x) = 0$.

To understand the implications of (1), note that the left side is $P(a < X \le b)$. Moreover, $P(X = b) = 0$ because otherwise $F_X(x)$ would have a step of nonzero size $P(X = b)$ at $x = b$, contradicting the supposition that $F_X(x)$ is continuous. Likewise, $P(X = a) = 0$. Hence, the left side of (1) is also equal to any of $P(a \le X \le b)$, $P(a < X < b)$, and $P(a \le X < b)$. Thus, the probability that a continuous random variable X falls between a and b (inclusive or exclusive does not matter) equals the area under the curve of $F'_X(x)$ from a to b.

You have exploited the above fact many times in your introductory methods course to calculate probabilities involving, for example, a standard normal random variable. Indeed, now you see that the familiar bell curve is nothing but the derivative of a cumulative distribution function.

Typically we give the derivative of the cumulative distribution function its own name of probability density function and symbolize it as $f_X(x)$. That is, $f_X(x) := F'_X(x)$. If there exists an x at which $F'_X(x)$ is undefined, then usually we assign $f_X(x)$ to be either $\lim_{\delta \searrow 0} f_X(x+\delta)$ or $\lim_{\delta \nearrow 0} f_X(x+\delta)$.

Since the cumulative distribution function $F_X(x)$ is nondecreasing, the probability density function $f_X(x)$ is nonnegative. Moreover, by putting $a := -\infty$ and $b := \infty$ in (1), we see from the second axiom of probability that the integral of the probability density function over the real line equals 1.

I also want to mention here that the support set of a continuous random variable X is defined to consist of all x for which $f_X(x) > 0$. This looks like the definition of the support set of a discrete random variable X. However, the interpretation is subtly different. With a discrete random variable X, the support set identifies those real numbers assumed by X with positive probability. With a continuous random variable X, the support set identifies those real numbers over which the cumulative distribution function is strictly increasing; as noted earlier, a continuous random variable X does not assume any specific real number with positive probability.

Now let us consider an example. Let λ be a positive real number and put $f_X(x) := 1_{\{x>0\}} C(\lambda) \exp[-\lambda x]$, where $C(\lambda)$ is a function of λ but not x. How can $C(\lambda)$ be chosen so that $f_X(x)$ is a probability density function? What is $P(X > 1)$? What is the cumulative distribution function of X?

When X is a continuous nonnegative random variable (that is, a continuous random variable whose support set is contained in the set of nonnegative real numbers), we often speak not only of its probability density function but also of its survival function and its hazard function.

The survival function of X is defined as

$$S_X(x) := 1 - F_X(x) = P(X > x).$$

If X represents the lifetime of a person or object, then $S_X(x)$ is the probability that the person or object is still alive at time x.

However, just like "success" on a Bernoulli trial does not necessarily correspond to "success" in ordinary English parlance, "survival" for a continuous nonnegative random variable does not necessarily correspond to "survival" in ordinary English parlance. For instance, X may be taken to represent the length of time that a person lives smoke free. In this case, X is bounded above by but not identified with the person's survival. Only if the person never smokes again will X literally represent the person's survival.

The hazard function of X is defined as

$$h_X(x) := \lim_{\delta \searrow 0} P(X \le x + \delta \mid X > x)/\delta \tag{2}$$

for all x with $S_X(x) > 0$. (Why do we require $S_X(x) > 0$?) The numerator of the quotient in (2) is the probability that a person "dies" within the next δ time units given that he/she had "survived" to time x. Because X is a continuous random variable, this numerator must tend to 0 as $\delta \searrow 0$. Therefore, to avoid a vacuous definition, we include the denominator δ in (2). The limit can be and typically is a positive real number, which may depend on x.

To understand the interplay among the probability density function, survival function, and hazard function, note that (2) can be expressed as

$$\begin{aligned}
h_X(x) &= \lim_{\delta \searrow 0} \frac{P(x < X \le x + \delta)}{P(X > x)\delta} \\
&= \lim_{\delta \searrow 0} \frac{F_X(x+\delta) - F_X(x)}{S_X(x)\delta} \\
&= \frac{1}{S_X(x)} \lim_{\delta \searrow 0} [F_X(x+\delta) - F_X(x)]/\delta \\
&= \frac{1}{S_X(x)} f_X(x).
\end{aligned}$$

Since $f_X(x) = -S'_X(x)$, we then obtain

$$h_X(x) = -\frac{S'_X(x)}{S_X(x)} = -\frac{d}{dx} \log[S_X(x)]. \tag{3}$$

Applying the Fundamental Theorem of Calculus with the "initial condition" $S_X(0) = 1$, we see that

$$\int_0^x h_X(t)\ dt = -\log[S_X(x)]$$

and, hence,

$$S_X(x) = \exp\left[-\int_0^x h_X(t)\ dt\right]. \tag{4}$$

Result (3) provides an alternative interpretation for the hazard function, namely that the hazard function reflects how quickly the survival function is decaying on the logarithmic scale.

Result (4), on the other hand, shows how to construct a continuous nonnegative random variable with a desired hazard function. For instance, suppose that we want to construct a continuous nonnegative random variable with $h_X(x) = \lambda > 0$ for $x > 0$. Then $S_X(x)$ must equal which determines both $F_X(x)$ and $f_X(x)$,

d. Expected values, means, and variances (Cf. pp. 173-203 of Larsen and Marx)

Let X be a continuous random variable with probability density function $f_X(x)$ and support set $\mathcal{X}$. We define the expected value of $g(X)$ as

$$E[g(X)] := \int_{\mathcal{X}} g(x) f_X(x)\ dx,$$

provided that the integral is absolutely convergent. If the integral is not absolutely convergent, then we say that $E[g(X)]$ does not exist as a finite number. If the integral is not absolutely convergent and $g(x) \geq 0$ for all $x \in \mathcal{X}$, then we may also say that $E[g(X)] = \infty$. Two special cases of interest are $g(X) := X$, which yields the mean of X, and $g(X) := (X - E[X])^2$, which yields the variance of X.

To illustrate the computation of an expected value, let α be a positive real number. The gamma function is defined by

$$\Gamma[\alpha] := \int_0^\infty x^{\alpha-1} \exp[-x]\, dx.$$

Since $n! = \Gamma[n+1]$ for any positive integer n and $\Gamma[\alpha+1] = \alpha\Gamma[\alpha]$ for any positive real number α, the gamma function may be viewed as an extension of the factorial function from the positive integers to the positive real numbers. Now let X have probability density function

$$f_X(x) := \frac{\lambda^\alpha}{\Gamma[\alpha]} x^{\alpha-1} \exp[-\lambda x] 1_{\{x>0\}}$$

for some positive real numbers α and λ. Putting $g(x) := x^\beta$ for $x \in \mathcal{X}$ (and defining $g(x)$ to be, say, 0 for $x \in \mathcal{X}^c$), where β is a positive real number, we have

$$E[X^\beta] = \int_0^\infty \frac{\lambda^\alpha}{\Gamma[\alpha]} x^{\alpha+\beta-1} \exp[-\lambda x]\, dx =$$

Let us do another example. Say that X has probability density function $f_X(x) := (\alpha-1)x^{-\alpha} 1_{\{x\geq 1\}}$ for some real constant $\alpha > 1$. Let β be a positive real constant, and take $g(x) := x^\beta$ for $x \in [1, \infty)$. We have

$$E[X^\beta] = \int_1^\infty (\alpha-1)x^{\beta-\alpha}\, dx = \lim_{M\to\infty} \int_1^M (\alpha-1)x^{\beta-\alpha}\, dx.$$

If $\beta = \alpha - 1$, then we have

$$E[X^\beta] =$$

If $\beta > \alpha - 1$, then we have

$$E[X^\beta] = \lim_{M\to\infty} \frac{(\alpha-1)}{(\beta-\alpha+1)}(M^{\beta-\alpha+1} - 1) = \infty.$$

If $\beta < \alpha - 1$, then we have

$$E[X^\beta] = \lim_{M\to\infty} \frac{(\alpha-1)}{(\beta-\alpha+1)}(M^{\beta-\alpha+1} - 1) = \frac{(1-\alpha)}{(\beta-\alpha+1)}.$$

Now we provide a useful technique for calculating the expected value of a continuous nonnegative random variable. Let X have probability density function $f_X(x)$ and survival function $S_X(x)$ with $\mathcal{X} = [0, \infty)$. Then integrating by parts with $u := x$, $dv := f_X(x)\ dx$, $v := -S_X(x)$, and $du := dx$ yields

$$\begin{aligned} E[X] &= \int_0^\infty x f_X(x)\ dx = \lim_{M\to\infty} \int_0^M x f_X(x)\ dx \\ &= \lim_{M\to\infty} \left\{ -M S_X(M) + \int_0^M S_X(x)\ dx \right\}. \end{aligned}$$

If $\lim_{M\to\infty} M S_X(M) = 0$, then we obtain

$$E[X] = \lim_{M\to\infty} \int_0^M S_X(x)\ dx = \int_0^\infty S_X(x)\ dx.$$

To illustrate use of this technique, let X have probability density function

$$f_X(x) := \lambda \exp[-\lambda x]$$

for $x \geq 0$, where λ is a positive real constant. Then, for $x \geq 0$, we have

$$S_X(x) = P(X > x) = \int_x^\infty f_X(t)\ dt = \int_x^\infty \lambda \exp[-\lambda t]\ dt = \exp[-\lambda x].$$

Since

$$\lim_{M\to\infty} M S_X(M) = \lim_{M\to\infty} M \exp[-\lambda M] = 0$$

(apply L'Hopital's rule), we may conclude that

$$E[X] = \int_0^\infty S_X(x)\ dx = \int_0^\infty \exp[-\lambda x]\ dx = \lambda^{-1} \int_0^\infty f_X(x)\ dx = \lambda^{-1}.$$

The preceding calculation is clearly easier than the more direct approach,

$$\int_0^\infty x f_X(x)\ dx = \int_0^\infty x \lambda \exp[-\lambda x]\ dx,$$

which is left to you as an exercise.

Finally, I note that the linearity and monotonicity properties reported for expected values of discrete random variables in Unit II also hold for expected values of continuous random variables.

e. Moment generating functions (Cf. pp. 257-269 of Larsen and Marx)

Moments and moment generating functions were defined for discrete random variables in Unit II. The same definitions apply for continuous random variables except that summations are replaced by integrals. In particular, the moment generating function of a continuous random variable X with probability density function $f_X(x)$ and support set $\mathcal{X}$ is

$$M_X(t) = E[\exp(tX)] = \int_{\mathcal{X}} \exp(tx) f_X(x)\ dx.$$

Also, the three results for moment generating functions described in Unit II apply here as well. The first of these results, recall, was that

$$\frac{d^n}{dt^n} M_X(t)|_{t=0} = E[X^n]$$

for any positive integer n, provided there exists $h > 0$ such that $M_X(t) < \infty$ for $t \in [-h, h]$.

To illustrate moment calculations for a continuous random variable, suppose that X has probability density function

$$f_X(x) := (2\pi)^{-1/2}\sigma^{-1} \exp[-(x-\mu)^2/(2\sigma^2)],$$

where $\mu \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$. Because directly calculating moments of X would be difficult, we will employ a "trick". We will standardize X by defining $Z := (X - \mu)/\sigma$. Then we will calculate the moments of Z, from which the moments of X can be recovered by linearity of expectation.

For any real number z, we have

$$P(Z \le z) = P(X \le \sigma z + \mu) = \int_{-\infty}^{\sigma z+\mu} (2\pi)^{-1/2}\sigma^{-1} \exp[-(x-\mu)^2/(2\sigma^2)]\ dx.$$

Making the substitutions $y := (x - \mu)/\sigma$ and $dy := dx/\sigma$, we can express the integral as

$$\int_{-\infty}^{z} (2\pi)^{-1/2} \exp[-y^2/2]\ dy.$$

Recalling the Fundamental Theorem of Calculus, we differentiate with respect to z to find that the probability density function of Z is

$$f_Z(z) := (2\pi)^{-1/2} \exp[-z^2/2].$$

Next, we note that, for any positive integer p,

$$E[Z^{2p}] = (2\pi)^{-1/2} \int_{\mathbb{R}} z^{2p} \exp[-z^2/2]\ dz = (2\pi)^{-1/2} \int_{\mathbb{R}} z^{2p-1} \exp[-z^2/2] z\ dz.$$

Integrating by parts with $u := z^{2p-1}$, $du := (2p-1)z^{2p-2}\ dz$, $dv := \exp[-z^2/2]z\ dz$, and $v := -\exp[-z^2/2]$, we obtain

$$E[Z^{2p}] = (2\pi)^{-1/2}(2p-1) \int_{\mathbb{R}} z^{2p-2} \exp[-z^2/2]\ dz = (2p-1)E[Z^{2p-2}].$$

Thus, by mathematical induction,

$$E[Z^{2p}] = (2p-1) \times (2p-3) \times \cdots \times 3 \times 1,$$

which some mathematicians denote $(2p-1)!!$, read "double factorial".

Note that $E[Z^{2p-1}]$ must exist as a finite number (because $E[Z^{2p}]$ does) and must equal zero because $E[Z^{2p-1}]$ is the integral of an odd function over $\mathbb{R}$.

Now, finally, we are in a good position to find some moments of $X = \sigma Z + \mu$. We have

$E[X] =$

$E[X^2] =$

$Var[X] =$

$$E[X^3] = E[(\sigma Z + \mu)^3] = \sigma^3 E[Z^3] + 3\sigma^2 \mu E[Z^2] + 3\sigma\mu^2 E[Z] + \mu^3 = 3\sigma^2\mu + \mu^3,$$

$$\text{and } E[X^4] = E[(\sigma Z + \mu)^4] = \sigma^4 E[Z^4] + 6\sigma^2\mu^2 E[Z^2] + \mu^4 = 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4.$$

A similar standardization "trick" can be employed to derive the moment generating function of X, $M_X(t)$. A sketch of the derivation is as follows; BST 675 Written Assignment 3, Fall 2010, requests details. First, show that the moment generating function of Z, $M_Z(t)$, equals $\exp[t^2/2]$. Second, show that $M_X(t) = \exp(t\mu)M_Z(t\sigma)$. Third, simplify to find that $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$.

f. Normal family (Cf. pp. 292-317 of Larsen and Marx)

A random variable X has the normal distribution with mean $\mu \in (-\infty, \infty)$ and standard deviation $\sigma \in (0, \infty)$ if its probability density function is

$$f(x) = (2\pi)^{-1/2}\sigma^{-1}\exp[-(x-\mu)^2/(2\sigma^2)].$$

Suppose that X has the normal distribution with mean μ and standard deviation σ. Let $a \in (-\infty, 0) \cup (0, \infty)$ and $b \in (-\infty, \infty)$. Then we can readily verify that $Y := aX + b$ has the normal distribution with mean $a\mu + b$ and standard deviation $|a|\sigma$. Below I prove this assertion for the case in which $a \in (0, \infty)$. The case in which $a \in (-\infty, 0)$ is left to you.

The cumulative distribution function of Y is, for any real number y,

$$\begin{aligned} P(Y \le y) &= P(aX + b \le y) \\ &= P(X \le (y-b)/a) \\ &= \int_{-\infty}^{(y-b)/a} (2\pi)^{-1/2}\sigma^{-1}\exp[-(x-\mu)^2/(2\sigma^2)]\ dx \\ &= \int_{-\infty}^{y} (2\pi)^{-1/2}(a\sigma)^{-1}\exp[-(t-a\mu-b)^2/(2\sigma^2a^2)]\ dt, \end{aligned}$$

where the last step follows from the substitution $t := ax + b$, $dt := a\ dx$. Then differentiation with respect to y yields the probability density function of Y, $(2\pi)^{-1/2}(a\sigma)^{-1}\exp[-(y-a\mu-b)^2/(2\sigma^2a^2)]$, which is the probability density function of the normal distribution with mean $a\mu+b$ and standard deviation $a\sigma$.

An alternative proof is given through moment generating functions. We have

$$\begin{aligned} M_Y(t) &= E[\exp(tY)] \\ &= E[\exp(t\{aX+b\})] \\ &= E[\exp(taX)\exp(tb)] \\ &= E[\exp(taX)]\exp(tb) \\ &= M_X(ta)\exp(tb) \\ &= \exp(\mu[at] + \sigma^2[at]^2/2)\exp(tb) \\ &= \exp(t[a\mu+b] + [a\sigma]^2t^2/2). \end{aligned}$$

Since Y has the same moment generating function as a normal random variable with mean $a\mu+b$ and standard deviation $a\sigma$, Y itself must be a normal random variable with mean $a\mu + b$ and standard deviation $a\sigma$ by the second result for moment generating functions.

A special case of interest is $a := \sigma^{-1}$ and $b := -\mu\sigma^{-1}$, which yields $Y = (X-\mu)/\sigma$. Then Y has the normal distribution with mean 0 and standard deviation 1, which is called the standard normal distribution.

The practical importance of the above result is that any probability involving a normal random variable can be expressed as a probability involving a standard normal random variable, and the cumulative distribution function of a standard normal random variable has been tabulated. An abbreviated table is shown below, where Z denotes a standard normal random variable. The process of defining $Z := (X - \mu)/\sigma$ and calculating a probability involving X in terms of Z is called standardization.

Table 1:

z	$P(Z \leq z)$	z	$P(Z \leq z)$
-3	0.0013	1	0.8413
-2	0.0228	2	0.9772
-1	0.1587	3	0.9987
0	0.5000		

To illustrate, suppose that X has the normal distribution with mean 100 and standard deviation 10. What is $P(X \geq 70)$? What is $P(90 \leq X \leq 120)$?

Normal distributions are of special interest for two reasons. First, many physical, biological, or social phenomena can reasonably be modeled using a normal distribution. Second, if a random variable X can be expressed as a sum of independent random variables $X_1, \ldots, X_n$, then under fairly general conditions the cumulative distribution function of X can be approximated by the cumulative distribution function of a normal random variable with mean $E[X]$ and standard deviation $\sqrt{Var[X]}$. This is a consequence of the Central Limit Theorem, with which you will become familiar in BST 676.

For instance, we know that a binomial random variable X with parameters p and n can be expressed as $X_1 + \cdots + X_n$, where $X_i := 1_{\{\text{success on trial i}\}}$ for $i \in \{1, \ldots, n\}$. Since X has mean np and standard deviation $\sqrt{np(1-p)}$, the Central Limit Theorem tells us that the cumulative distribution function of X can be approximated by the cumulative distribution function of a normal random variable with mean np and standard deviation $\sqrt{np(1-p)}$. Or, put differently, $(X - np)/\sqrt{np(1-p)}$ "looks" like a standard normal random variable. The quality of the approximation gets better as $np(1-p)$ gets larger.

To illustrate, suppose that X has the binomial distribution with parameters $p = 0.5$ and $n = 100$. Then $np = 50$ and $\sqrt{np(1-p)} = \sqrt{25} = 5$. Letting Z denote a standard normal random variable, we have

$$P(45 \leq X \leq 55) = P(-1 \leq (X - 50)/5 \leq 1) \approx P(-1 \leq Z \leq 1) =$$

Actually, since $P(45 \leq X \leq 55) = P(45-\delta < X < 55+\delta)$ for any $\delta \in (0, 1]$, we can validly approximate this probability by $P(-1-\delta/5 < Z < 1+\delta/5)$. Such a δ is referred to as a continuity correction. The best choice of δ is arguably 0.5, on the grounds that $5Z+50$ is meant to "look" like X, so $X = 45$ should translate to $44.5 < 5Z + 50 < 45.5$ rather than to (say) $5Z + 50 = 45$ or $44 < 5Z + 50 < 46$. In fact, with $\delta = 0.5$, we obtain $P(-1.1 < Z < 1.1) = 0.7287$, which is in agreement with the actual value of $P(45 \leq X \leq 55)$ to four decimal places.

g. Exponential and gamma families (Cf. pp. 327-333 of Larsen and Marx)

A random variable X has the gamma distribution with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$ if its probability density function is

$$f(x) = \frac{1}{\Gamma[\alpha]\beta^{\alpha}} x^{\alpha-1} \exp[-x/\beta] 1_{\{x>0\}}.$$

We have $E[X] = \alpha\beta$ and $Var[X] = \alpha\beta^2$. We refer to α as a shape parameter and to β as a scale parameter.

An alternative parametrization replaces β with $1/\lambda$, where $\lambda \in (0, \infty)$. Then $E[X] = \alpha/\lambda$ and $Var[X] = \alpha/\lambda^2$. When necessary to distinguish between the two parametrizations, we call the one with β a "mean" parametrization (because $E[X]$ is proportional to β) and the one with λ a "rate" parametrization (because, if α is a positive integer, X can be interpreted as the time required for α Poisson events to occur when the rate of Poisson events per unit time is λ).

Worth noting is that the shape of $f(x)$ is highly sensitive to α, which is why α is called a shape parameter. When $\alpha \in (0, 1]$, $f(x)$ is strictly decreasing on $(0, \infty)$. When α exceeds 1, $f(x)$ has a mode — i.e., a point at which $f(x)$ is maximized — interior to $(0, \infty)$. As α continues increasing, the mode of $f(x)$ moves rightward and $f(x)$ takes on a bell-shaped appearance. In fact, for very large α, a gamma distribution is well approximated by a normal distribution. Thus, for purposes of modeling physical, biological, or social phenomena not well described by a normal distribution, a gamma distribution with a small or modest α is a more viable choice than a gamma distribution with a large α.

In the special case that $\alpha = 1$, we say that X has the exponential distribution with scale parameter β. In the special case that $\alpha = p/2$ and $\beta = 2$, where p is a positive integer, we say that X has the chi-square distribution on p degrees of freedom. (In fact, other than for convenience in producing tables for the backs of methods textbooks, there is no real reason that a chi-square distribution must have integer degrees of freedom. So, if we like, we can just let p be a positive real number.) What are the mean and standard deviation of a chi-square random variable on p degrees of freedom?

h. Resolution of motivating case studies

Our first motivating case study considered the problem of modeling a birthweight distribution. Since the distribution of birthweights is left skewed, a normal probability model may not be adequate if our research objectives entail accurately representing the frequency of very low birthweights (as may be the case if, for instance, we want to quantify the association between birthweight and infant mortality).

What else can we try? A gamma probability model is not viable because a gamma distribution is either right skewed (if the shape parameter is small) or approximately symmetric (if the shape parameter is large). Of course, there are other commonly encountered families of continuous distributions that I have not presented in Unit III, and perhaps one of them may work. However, some researchers in perinatal epidemiology have adopted a different approach, that of mixture modeling.

The basic idea is that a population is assumed to have a certain number of nonoverlapping subpopulations or components, say k of them, each of which is governed by a different normal distribution. The distribution governing the full population is then called a mixture of the normal distributions governing the components.

More formally, let p_k denote the proportion of the full population contained in component k, μ_k the mean in component k, and σ_k the standard deviation in component k. Then the probability density function governing the full population is

$$\sum_{j=1}^{k} p_j (2\pi)^{-1/2} \sigma_j^{-1} \exp[-(x-\mu_j)^2/(2\sigma_j^2)]. \tag{5}$$

To see this, let X denote an observation on a randomly selected member of the full population. Let A_j denote the event that this observation arises from component j, for $j \in \{1, 2, \ldots, k\}$, and x a real number. The cumulative distribution function of X is

$$\begin{aligned} P(X \le x) &= \sum_{j=1}^{k} P(X \le x \cap A_j) \\ &= \sum_{j=1}^{k} P(A_j) P(X \le x \mid A_j) \\ &= \sum_{j=1}^{k} p_j \int_{-\infty}^{x} (2\pi)^{-1/2} \sigma_j^{-1} \exp[-(t-\mu_j)^2/(2\sigma_j^2)]\, dt. \end{aligned}$$

Differentiation with respect to x yields (5) as the probability density function of X.

Figure 1 on page 1 of Unit III shows the probability density function (5) for $k = 4$, $p_1 = 0.009$, $p_2 = 0.231$, $p_3 = 0.707$, $p_4 = 0.054$, $\mu_1 = 872$, $\mu_2 = 2890$, $\mu_3 = 3165$, $\mu_4 = 3821$, $\sigma_1 = 247$, $\sigma_2 = 726$, $\sigma_3 = 403$, $\sigma_4 = 365$.

With appropriate choices of k, p_j, μ_j, and σ_j, one can approximate virtually any configuration of data. So, why are not mixture models used by everyone? One reason is that, despite expressing a distribution of data in terms of finitely many parameters, mixture models are not easy to work with when performing hypothesis tests.

For instance, consider testing the null hypothesis that $k = 1$ against the alternative hypothesis that $k = 2$. At first, you may think that this is as simple as testing whether $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$ in a two-component mixture model, since $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$ effectively reduces a two-component mixture model to a one-component mixture model (i.e., an ordinary normal distribution). However, putting $p_2 = 0$ accompishes the same, so the answer is not as simple as testing whether $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$.

I have done some theoretical work on performing hypothesis testing in mixture models. However, another approach for data analysts is to use model selection criteria like the AIC and BIC to choose between competing mixture models. My paper {http://www.biomedcentral.com/1471-2393/10/37} takes this approach, which provides the basis for choosing $k = 4$ in modeling a birthweight distribution.

Another challenge with mixture models is that component membership is not observable, even if we believe that the components are biologically meaningful and not merely convenient mathematical tools. Indeed, all we can do is make a statement about the probability that a certain member of the population belongs to a particular component.

For instance, the probability that a population member with observation x belongs to component 1 is (Cf. Written Assignment 4, BST 675, Fall 2010)

$$\frac{p_1\sigma_1^{-1}\exp[-(x-\mu_1)^2/(2\sigma_1^2)]}{\sum_{j=1}^k p_j\sigma_j^{-1}\exp[-(x-\mu_j)^2/(2\sigma_j^2)]}.$$

Our second motivating case study questioned how one might model the time to smoking recidivism (or, more precisely, the time of smoke-free living). Letting T denote this time, we reasoned that the conditional probability

$$P(T \le a+b \mid T > a), \tag{6}$$

where a and b were nonnegative real numbers, should be a decreasing function of a.

We later saw that the hazard function of T was closely related to the conditional probability (6),

$$h(a) = \lim_{b\searrow 0} P(T \le a+b \mid T > a)/b \tag{7}$$

in the present notation. From expression (7), we conclude that the hazard function should also be decreasing in its argument.

A simple choice for the hazard function to make it decreasing in its argument is

$$h(t) := Ct^{D-1}$$

for $t > 0$, where $C \in (0,\infty)$ and $D \in (0,1)$. What is the corresponding probability density function?

To find out, first we obtain the survival function

$$S(t) = \exp[-\int_0^t h(x)\ dx] = \exp[-Ct^D/D]$$

and the cumulative distribution function

$$F(t) = 1 - S(t) = 1 - \exp[-Ct^D/D].$$

Note that the above expressions (and all expressions below) are valid only for $t > 0$.

Next, we differentiate the cumulative distribution function to obtain the probability density function

$$f(t) = \frac{d}{dt}F(t) = Ct^{D-1}\exp[-Ct^D/D]. \tag{8}$$

The probability density function (8) does not belong to either the normal or gamma family. Rather, (8) belongs to the so-called Weibull family.

Two remarks are in order at this juncture. First, all of the above computations would have gone through with $D \in [1, \infty)$. The choice $D = 1$ would have returned an exponential distribution with rate parameter C, demonstrating that the family of exponential distributions is in the intersection of the gamma and Weibull families. The choice $D > 1$ would have produced an increasing hazard function. Although inappropriate for our smoking recidivism example, there are some phenomena for which an increasing hazard function would be realistic. Can you name some?

Second, if we put $Y := T^D$, then for any $y \in (0, \infty)$ we have

$$P(Y \leq y) = P(T^D \leq y) = P(T \leq y^{1/D}) = 1 - \exp[-Cy/D].$$

Since $Y = T^D$ is equivalent to $T = Y^{1/D}$, this shows that any Weibull random variable is a power transformation of an exponential random variable.