

EEB 600A, Lecture 3

Continuous Distributions

For continuous random variables, an event is no longer a single point (i.e. $x = 5$), as any single point for a truly continuous distribution has probability zero. Rather, events are defined by the probability of the random variable falling in some interval, for example $\Pr(x \leq 10)$, $\Pr(1 \leq x \leq 1.5)$, or $\Pr(x \geq 2.5)$. The probability for any given interval is computed using the **probability density function**, or **pdf**, denoted $p(z)$, for the random variable of interest. The pdf satisfies the integral

$$P(z_1 \leq z \leq z_2) = \int_{z_1}^{z_2} p(z) dz$$

If z_{min} and z_{max} are the upper and lower bounds to z , then $p(z) = 0$ outside of this range, and over the entire range $\int_{z_{min}}^{z_{max}} p(z) dz = 1$. Both of these properties are in accord with common sense — a probability is never negative, and the total probability of all possible outcomes is one.

The **cumulative distribution function**, or **cdf** is defined by

$$cdf(x) = \Pr(z \leq x) = \int_{-\infty}^x p(z) dz$$

Example 1. Suppose that z is continuously distributed in the range of 0 to ∞ with probability density function

$$p(z) = \lambda e^{-z\lambda}$$

This is the **negative exponential distribution** in which the density has a maximum at $z = 0$ and declines to zero as $z \rightarrow \infty$. Since the integral of $p(z)$ is $-e^{-z\lambda}$,

$$\int_0^{\infty} p(z) dz = -e^{-z\lambda} \Big|_0^{\infty} = 0 - (-1) = 1$$

showing that $p(z)$ fulfills the properties of a probability density, as it integrates to one and $p(z) \geq 0$.

What is the probability that a randomly drawn individual will have z in the range of $1/4$ to $1/2$?

$$\Pr(1/4 \leq z \leq 1/2) = \int_{1/4}^{1/2} p(z) dz = -e^{-z\lambda} \Big|_{1/4}^{1/2} = e^{-\lambda/4} - e^{-\lambda/2}$$

The numerical answer depends on the parameter λ . If, for example, $\lambda = 2$, then $\Pr(1/4 \leq z \leq 1/2) = 0.239$.

Finally, the cumulative distribution function is given by

$$\text{cdf}(x) = \int_0^x p(z) dz = -e^{-z\lambda} \Big|_0^x = 1 - e^{-x\lambda}$$

Hence,

$$\Pr(z \leq x) = 1 - e^{-x\lambda}$$

and

$$\Pr(z \geq x) = e^{-x\lambda}$$

Common Continuous Distributions:

1. Uniform. Parameters: The range a, b of the distribution

$$p(x) = \begin{cases} 1/(b-a) & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Mean $\mu = (a + b)/2$, variance $\sigma^2 = (b - a)^2/12$. A uniform random variable has equal probability at any point within the interval (a, b)

2. Exponential. Parameter $\lambda > 0$, the probability of a success per unit time. The exponential distribution arises as the waiting until a success, and is the continuous analog of the geometric distribution. The mean waiting time is $1/\lambda$.

$$p(x) = \begin{cases} \lambda e^{-x\lambda} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Example 1 shows for an exponential that

$$\Pr(z \leq x) = 1 - e^{-x\lambda}$$

and

$$\Pr(z \geq x) = e^{-x\lambda}$$

An exponential random variable has a **memory-less property**: Independent of how much time has passed since a failure, the waiting time until the next failure has the same distribution. Consider two light bulbs with the same exponential failure time distribution, where one light bulb has been on for 2 hours and the other for 10,000 hours. The distribution for additional time each light bulb burns is independent of how long they have burned in the past.

3. Gamma. Parameters $\lambda, r > 0$. The waiting time until the r th success given a success rate of λ per unit time. This is the continuous analogue of the negative binomial.

$$p(x) = \begin{cases} \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Here $\Gamma(r) = \int_0^\infty \lambda^r x^{r-1} e^{-\lambda x} dx$ is the gamma function. Since the gamma has two "shape" parameters, it can take on a variety of forms. Mean $\mu = r/\lambda$, variance $\sigma^2 = r/\lambda^2$.

Gamma distributions are often used in molecular evolution as models of the distribution of mutation rates or selection coefficients.

Fun fact : The sum of r exponentials is a Gamma (a gamma with $r = 1$ is an exponential).

4. Beta. Parameters $p, q > 0$. The random variable falls in the interval 0 to 1.

$$p(x) = \begin{cases} \frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1} & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Here $B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} dx$ is the beta function.

Beta distributions arise in population genetics when describing the distribution of allele frequencies (which are, of course, in the 0-1 interval) under genetic drift and mutation.

Parameters vs. Estimates

It is very important to distinguish between true **parameters** of distributions and **estimates** of those parameters obtained by sampling. True parameter values can only be obtained if every member of a population is measured with absolute accuracy. We must therefore almost always settle for approximations, the accuracy of which depends on the experimental setting, the measurement apparatus, and the sample size. **Statisticians often denote parameters of a population with Greek symbols and to sample estimates with Roman symbols.** We will adhere to this protocol as much as possible, although there will be some instances where traditional quantitative-genetic notation prevents us from doing so.

Momements and Expectations.

The **Expectation** of a random variable is the average value of that quantity over the distribution. The simplest case is the **expected value** or **expectation** of the random variable X , where

$$E[X] = \begin{cases} \int_{-\infty}^{+\infty} z p_X(z) dz & \text{if } X \text{ is continuous} \\ \sum_{k=-\infty}^{\infty} x_k \Pr(X = x_k) & \text{if } X \text{ is discrete} \end{cases}$$

This is often denoted by μ or μ_X (to indicate that it refers to the random variable X), and the simple expectation is also referred to as the **arithmetic mean** or **first moment about the origin**. The sample estimate of the mean is generally denoted by \bar{z} , and estimated as the average of the n measures,

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$$

Example 2. What is the mean of the distribution discussed in Example 1? Since the integral of $(z\lambda) e^{-z\lambda}$ is $-(z + 1/\lambda) e^{-z\lambda}$,

$$\mu = \int_0^{\infty} z p(z) dz = -(z + 1/\lambda) e^{-z/\lambda} \Big|_0^{\infty} = 1/\lambda$$

Thus, the parameter $1/\lambda$ is the mean of the distribution defined by the density function $p(x) = \lambda e^{-z\lambda}$.

More generally, the expectation for any function $f(X)$ of the random variable is given by

$$E[f(X)] = \begin{cases} \int_{-\infty}^{+\infty} f(z) p_X(z) dz & \text{if } X \text{ is continuous} \\ \sum_{k=-\infty}^{\infty} f(x_k) \Pr(X = x_k) & \text{if } X \text{ is discrete} \end{cases}$$

Two useful properties of expectations are

$$\begin{aligned} E(x + y) &= E(x) + E(y) \\ E(cx) &= c E(x) \quad \text{where } c \text{ is a constant} \end{aligned}$$

The Variance

Higher-order moments provide measures of the dispersion of a frequency distribution. The most useful measure is the population **variance**, or **second moment about the mean**. The variance is the expected squared deviation of an observation from its mean,

$$\sigma^2 = \int_{-\infty}^{+\infty} (z - \mu)^2 p(z) dz = E [(z - \mu)^2]$$

Because $\mu = E(z)$, this quantity can be expressed more simply by expanding $(z - \mu)^2$ to obtain

$$\sigma^2 = E(z^2 - 2z\mu + \mu^2) = E(z^2) - 2\mu E(z) + \mu^2 = E(z^2) - \mu^2$$

When there is no ambiguity as to the variable being considered, σ^2 suffices. More generally, the variance of z is denoted by σ_z^2 or $\sigma^2(z)$. One useful identity for variances is that

$$\sigma^2(cz) = c^2 \sigma^2(z) \quad \text{where } c \text{ is a constant}$$

A slight complication arises when one wishes to estimate the parameter σ^2 from a random sample of the population. As noted above, the true parameters μ and $E(z^2)$ cannot be known with certainty unless the entire population is sampled. Because the estimated mean (\bar{z}) is a function of the data, individual measures tend to be closer to the observed mean than to the true mean, and as a consequence, observed values of $\bar{z}^2 - \bar{z}^2$ tend to be slightly less than the parametric value $[E(z^2) - \mu^2]$. Thus, the estimator $(\bar{z}^2 - \bar{z}^2)$ is biased in the sense that it tends to underestimate the parameter $\sigma^2(z)$ to a degree that decreases with increasing sample size (n). A major goal of applied statistics is to obtain unbiased estimators that account for these kinds of small sample size limitations. In the case of the variance, the solution is

$$\text{Var}(z) = \frac{n(\bar{z}^2 - \bar{z}^2)}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2$$

provides an unbiased estimate of $\sigma^2(z)$. This equation should be used whenever the true population variance, $\sigma^2(z)$, is being estimated from actual sample data.

The variance is measured in units that are the square of those of the mean, but it is often desirable to describe the dispersion of a frequency distribution on the same scale as the mean. The square root of the variance of z is called the **standard deviation** of z . The parametric value is denoted by $\sigma(z)$, σ_z , or just σ , and the statistic by $\text{SD}(z) = \sqrt{\text{Var}(z)}$. The ratio of the standard deviation to the mean, the **coefficient of variation**, is frequently used as a relative measure of dispersion. It is known that the statistic $\text{CV}(z) = \text{SD}(z)/\bar{z}$ is a downwardly biased estimator of the parametric index (σ/μ) , but the bias is expected to be negligible in most cases.

The Normal, or Guassian, distribution

When large data sets are displayed in the form of frequency histograms, they often approximate a bell-shaped distribution. Three famous mathematicians, DeMoivre (1738), LaPlace (1778), and Gauss (1809), worked out the properties of a very useful description of this form — the **normal distribution**, also referred to as the **Gaussian distribution**. If z is a normally distributed variable, its density function is given by

$$p(z) = (2\pi\sigma^2)^{-1/2} \exp \left[-\frac{(z - \mu)^2}{2\sigma^2} \right]$$

The normal distribution is a function of only two parameters, the population mean (μ) and variance (σ^2). The normal density attains a maximum when $z = \mu$ and declines continuously and symmetrically in both directions as z deviates from μ . A normally distributed variable with mean μ and variance σ^2 is often denoted by $z \sim N(\mu, \sigma^2)$, where \sim means “is distributed as.” The **unit normal** is a normal random variable with mean zero and variance one, which is often called a **standard normal**. If z is normally distributed with mean μ and variance σ^2 , then $z' = (z - \mu)/\sigma$ follows a unit normal distribution.

The normal distribution plays a central role in statistical theory for two reasons. First, the normal probability density function has many simple mathematical features that allow the derivation of practical statistical tests. Second, even when actual distributions of phenotypes are inconsistent with the normal density function, after an appropriate scale transformation (such as $\ln(x)$), many can be rendered approximately normal. A general reason why many traits are distributed normally or nearly so is provided by the **central limit theorem**, which states that the sum of a number of independent random variables approaches normality as the number of variables increases.

There are, of course, limitations of the normal density function and of distribution functions in general. For instance, the normal distribution gives small positive values, rather than zero, for negative z , an unrealistic situation for traits such as body size or bone length, which cannot take on negative values. Nevertheless, if the mean of a distribution is sufficiently greater than zero, the theoretical incidence of negative values is minuscule and not problematical. It should also be emphasized that the normal distribution is a continuous function, giving positive values for noninteger values of z . It is, therefore, not strictly applicable to meristic traits such as egg number or spine count, although it provides a close approximation when the number of classes is large.

Skewness and Kurtosis, the 3rd and 4th moments.

While the normal distribution is the standard for most statistical tests, the true distribution of some test statistic can depart from a normal. The third and fourth moments provide a measure of the amount of departure from normality.

The third moment about the mean (μ_3) is a useful measure of the asymmetry of a distribution. Also known as the **skewness**, μ_3 is the expected cubic deviation from the mean. As in the case of the variance, it can be expressed in terms of the moments about the origin,

$$\begin{aligned}\mu_3 &= \int_{-\infty}^{+\infty} (z - \mu)^3 p(z) dz = E[(z - \mu)^3] \\ &= E(z^3) - 3\mu E(z^2) + 3\mu[E(z)]^2 - [E(z)]^3 \\ &= E(z^3) - 3\mu E(z^2) + 2\mu^3\end{aligned}$$

An unbiased sample estimator for μ_3 is

$$\text{Skw}(z) = \frac{n^2 (\bar{z}^3 - 3 \bar{z}^2 \bar{z} + 2 \bar{z}^3)}{(n-1)(n-2)}$$

where \bar{z}^3 denotes the observed mean cubed value of z . The degree of asymmetry can also be described with a dimensionless index, the **coefficient of skewness**, which is estimated by the ratio

$$k_3 = \frac{\text{Skw}(z)}{\text{Var}(z)^{3/2}}$$

k_3 is positive when the longer tail of a distribution is to the right, negative when the tail is to the left, and zero for a perfectly symmetrical distribution.

For a symmetric distribution (such as a normal), the third moment (μ_3) is equal to zero. For the normal, the fourth moment has an expected value equal to $3\sigma^4$. Thus, if we let $\text{Kur}(z)$ be the sample estimate of μ_4 , where Kur denotes **kurtosis**, the index

$$k_4 = \frac{\text{Kur}(z) - 3 [\text{Var}(z)]^2}{[\text{Var}(z)]^2}$$

where

$$\text{Kur}(z) = \frac{n^2(n+1)(\bar{z}^4 - 4 \bar{z}^3 \bar{z} + 6 \bar{z}^2 \bar{z}^2 - 3 \bar{z}^4)}{(n-1)(n-2)(n-3)}$$

provides a measure of the peakedness of a distribution. For a truly normal distribution, $k_4 = 0$. A distribution with a high narrow peak relative to the normal ($k_4 > 0$) is said to be **leptokurtic**. A broader peak than normal ($k_4 < 0$) is referred to as **platykurtic**.

In general, the r th moment (or central moment, or moment about the mean) is given by

$$\mu_r = E[(z - \mu)^r] = \int_{-\infty}^{+\infty} (z - \mu)^r p(z) dz$$

is a general expression for the r th moment about the mean. It also follows that μ_r can always be expressed in terms of moments about the origin [$E(z)$, $E(z^2)$, \dots , $E(z^r)$]. As was shown for the variance and the skewness, these terms are obtainable from the binomial expansion of $(z - \mu)^r$.

CONFIDENCE INTERVALS

Estimates such as \bar{z} and $\text{Var}(z)$ vary from one sample to the next because of sampling error, so it is useful to know how far an observed statistic is likely to deviate from the true parameter that is being estimated. Although the true values

are unknown, if something is known about the sampling error of the estimate, it is possible to evaluate the probability that the observed value lies within a specific range of the true value. Generally, we do not estimate the sampling error of statistics by sampling populations over and over again, but by using known algebraic expressions that themselves depend on sample statistics.

The square root of the variance of the estimator (or other test statistic) is usually called its **standard error**. Ideally, the full distribution of the estimator (the distribution of the possible true values of the unknown parameter being estimated) would be examined. This is what is done under **Bayesian statistics**. Often, instead, one simply obtains a probability interval (or **confidence interval** or **confidence limits**) such that there is a 90 or 95 percent chance that this interval enclosed the true parameter value given the value of the estimator.

It is often the case that an estimator or test statistic is approximately normal (or at least assumed to be approximately normal), in which case a 95% interval is given by (test statistic) $\pm 1.96 \cdot SE$ (Se = standard error of the estimator). The phrase **large sample variance** often appears, and this refers to an approximation for the sampling variance of a test statistic (or estimator) for large sample values. It is often the case that for large samples the distribution of the statistic is also approximately normal.