STA 580 — Spring 2009 — Dr. Charnigo

Lecture 3

Introduction. In Lecture 2 we discussed notions of probability and conditional probability along with some useful rules for their computation. Today we enlarge our probabilistic framework by defining random variables and describing how to compute probabilities involving them.

One motivation for defining random variables is that events can often be expressed succinctly in terms of a variable X assuming numerical values. This variable X is random in the sense that a numerical value is not assumed until we know which events have occurred. Consider again the "Genetics" example on page 69, and let X denote the number of affected siblings. Possible values for X are 0, 1, and 2. But we do not know whether $X = 0$, $X = 1$, or $X = 2$ until we find out whether each sibling has been affected. Problem 3.32 can now be succinctly rephrased as, "What is $P(X = 1)$?"

A second motivation is that sample values $x_1, \ldots, x_n$ (as in Lecture 1) can be viewed as realizations of random variables. This is because what we have for sample values depends on which n members of the population have been included in the sample. For instance, Table 2.13 showed serum cholesterol reductions of $49, -10, \ldots, 12$ for 24 specific hospital employees. Yet, if a different group of 24 hospital employees had been selected, the serum cholesterol reductions might have been $15, 3, \ldots, 28$. Indeed, before a sample is selected, we do not have fixed numerical values $x_1, \ldots, x_n$. However, we can use (capital) $X_1, \ldots, X_n$ to symbolize the random variables for which fixed numerical values will be realized after the sample is selected. The inferential methods in Lectures 4 through 14 will invoke the following important principle: probabilistic statements involving $X_1, \ldots, X_n$ can help us to determine whether certain assertions about the population are compatible with the sample values that we observe.

Discrete random variables

Definition. A random variable X is discrete (Definition 4.2) if its possible values can be enumerated. Examples include the number of siblings affected by the disease in "Genetics" (possible values are 0, 1, 2) and the number of people in the world who will contract influenza next year (possible values are 0, 1, 2, ..., up through about six billion).

Probability mass function and cumulative distribution function. The probability mass function (Definition 4.4) of a discrete random variable X is denoted $f(x)$ and defined as $P(X = x)$ for each numerical value x that can be taken on by X.

The cumulative distribution function (Definition 4.7) is denoted $F(x)$ and defined as $P(X \leq x)$. Let a and b be any two numbers with $b > a$. A useful formula for computing probabilities is $P(a < X \leq b) = F(b) - F(a)$.[1]

Example (probability mass function and cumulative distribution function). Consider the "Genetics" example, with X denoting the number of affected siblings. Recalling our solutions to exercises 3.31, 3.32, 3.33 in Lecture 2, we note that

$$f(0) = P(X = 0) = P(\text{neither sibling affected}) = 1/4,$$

$$f(1) = P(X = 1) = P(\text{exactly one sibling affected}) = 1/2,$$

$$\text{and } f(2) = P(X = 2) = P(\text{both siblings affected}) = 1/4.$$

[1]Since $\{X \leq a\}$ and $\{a < X \leq b\}$ are mutually exclusive with union $\{X \leq b\}$ (draw a picture!), we have $F(a) + P(a < X \leq b) = P(X \leq a) + P(a < X \leq b) = P(X \leq b) = F(b)$.

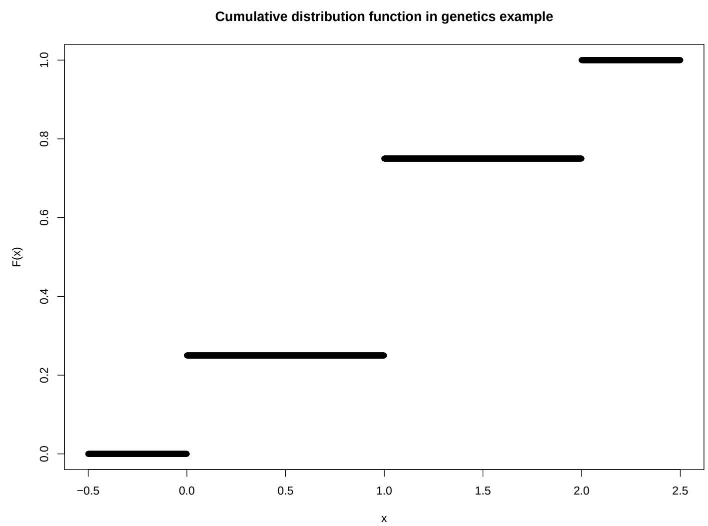
Moreover,

$$F(0) = P(X \leq 0) = P(X = 0) = 1/4,$$

$$F(1) = P(X \leq 1) = P(X = 0 \cup X = 1) = P(X = 0) + P(X = 1) = 3/4,$$

and $F(2) = P(X \leq 2) = P(X \leq 1 \cup X = 2) = P(X \leq 1) + P(X = 2) = 1.$

Figure 1:



Note that we can define $F(x)$ for any number x, not just a numerical value that can be taken on by X. So, for instance, $F(1.5) = P(X \leq 1.5) = 3/4$. We can visualize the cumulative distribution function of a discrete random variable as a step function with discontinuities at the numerical values that can be taken on by X. Figure 1 above displays $F(x)$ for the "Genetics" example. The size of the discontinuity at $x = 0$ is precisely $f(0) = 1/4$, the size of the discontinuity at $x = 1$ is precisely $f(1) = 1/2$, and the size of the discontinuity at $x = 2$ is precisely $f(2) = 1/4$.

Expected value and variance. The expected value of X (Definition 4.5) is denoted $E[X]$ or μ and is defined as $\sum\{x \times f(x)\}$, where the summation is taken over all possible values for X.[2] Note that $E[X]$ is a weighted average of the possible values for X, the weights being the probabilities attached to the possible values. Thus, we also refer to $E[X]$ as the mean of X.

The variance of X (Definition 4.6, Equation 4.1) is denoted $Var[X]$ or σ^2 and is defined as $\sum\{(x-\mu)^2 \times f(x)\}$.[3] So, the variance is a weighted average of the possible squared deviations from the mean. The standard deviation of X is the square root of the variance and is denoted $SD[X]$ or σ.

Binomial distributions

Definition. Suppose that we conduct n "trials", each of which has only two possible outcomes: "success" and "failure". In this statistical context, "success" does not neccessarily have the same meaning that we might assign it in ordinary English conversation. Suppose also that the trials are independent, in that the outcome on one trial does not influence the outcome on any other trial,[4] and that the probability of success — call it p, a number between 0 and 1 — is the same for each trial.

To build intuition about independent trials with identical success probability, let us consider a situation in which trials are <u>not</u> independent and do <u>not</u> have identical success probability: basketball free throw shooting over a college player's career. The trials are not independent because the player may get on a "hot streak" and hit one free throw after another. The trials do not have identical success probability because, over the months and years of the player's career, some improvement is anticipated.

[2] Those who have taken calculus will recall that not all infinite sums converge. Hence, there are discrete random variables for which the expected value is undefined. Fortunately, we will not encounter them in STA 580.

[3] Again, we are assuming that the sum converges.

[4] A more formal characterization of independence involves generalizing Definition 3.7 to accommodate more than two constitutent events. While not terribly difficult to pursue, doing so would take us too far afield.

If we conduct n trials that are independent with identical success probability, then the total number of successes in the n trials — call it X — is a binomial random variable (or "has a binomial distribution"). The probability mass function is (Equation 4.5)

$$f(x) = P(X = x) = \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x}$$

for $x = 0, 1, \ldots, n$. In fact, X from the "Genetics" example is a binomial random variable with $n = 2$ trials and probability of success $p = 1/2$.

Table 1 of Rosner. We can use a desk calculator and Equation 4.5 to evaluate probabilities involving binomial random variables. Or we can employ SAS. A third option, if $n \leq 20$ and p is a multiple of 0.05, is to use Table 1 of Rosner. Table 1 lists $f(x)$ for all x between 0 and n. If $p > 0.50$ (but is still a multiple of 0.05), you can define $Y := n - X$ and recast your statement about X into a statement about Y. This is because (Equation 4.6) Y is binomial with n trials and success probability $p_Y = 1 - p$.

Example (Table 1 of Rosner). Suppose that X is binomial with 5 trials and success probability 0.25. We have $P(X = 0) = 0.2373$, $P(X = 1) = 0.3955$, and so forth. Thus, $P(X \leq 1) = 0.2373 + 0.3955 = 0.6328$ and $P(X \geq 2) = P(X > 1) = 1 - P(X \leq 1) = 0.3672$.

Now suppose that X is binomial with 6 trials and success probability 0.90. Let $Y := 6 - X$, so that Y is binomial with 6 trials and success probability $0.10 = 1 - 0.90$. We can use Table 1 to find $P(X \geq 4)$ by noting that $P(X \geq 4) = P(Y \leq 2) = 0.5314 + 0.3543 + 0.0984 = 0.9841$.

Expected value and variance. The mean of a binomial random variable (Equation 4.7) is known to be np, while the variance is known to be $np(1-p)$. Hence, when you have a binomial random variable, there is no need for explicit calculation of $\sum\{x \times f(x)\}$ or $\sum\{(x-\mu)^2 \times f(x)\}$.

Continuous random variables

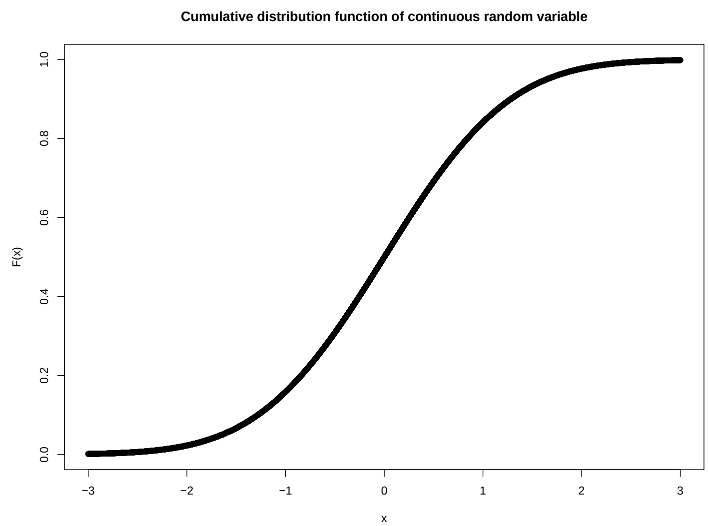
Definition and cumulative distribution function. Now suppose that we have a random variable X whose possible values cannot be enumerated. For instance, let X be the systolic blood pressure measurement for a randomly selected person walking along Rose Street. A possible value for X is 142. Other possible values are 142.47, 141.5924, 142.061358, and 141.9370421865. Whether we have a device that can measure systolic blood pressure with arbitrary precision is another matter. Yet, in principle, the possible values cannot be enumerated because they include a continuum of numbers.

The cumulative distribution function (Definition 5.2) is denoted $F(x)$ and defined as $P(X \leq x)$, just as before with discrete random variables. If the cumulative distribution function is continuous, as in Figure 2 below, then we say that X is a continuous random variable.[5]

Probability density function and inequalities. The probability density function for a continuous random variable X is denoted $f(x)$ and satisfies the

[5] Thus, some random variables are neither discrete nor continuous. An example would be the amount of time spent on Facebook today for a randomly selected person walking along Rose Street. The cumulative distribution function would have a discontinuity at $x = 0$ corresponding to a positive probability that the randomly selected person did not use Facebook at all. The bottom line is that Definition 4.3 in the textbook is not quite correct; the inability to enumerate the possible values does not by itself ensure that we have a continuous random variable.

Figure 2:



following relations for any numbers a and b with $b > a$ (Definition 5.1):

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)\, dx,$$

where $\int_a^b f(x)\, dx$ is mathematical shorthand for "the area under the curve of $f(x)$ from a to b". See Figure 3 below for an illustration.[6]

For continuous random variables — but, beware, not for discrete random variables — we are permitted to interchange strict inequalities ("<") with weak inequalities ("≤"):

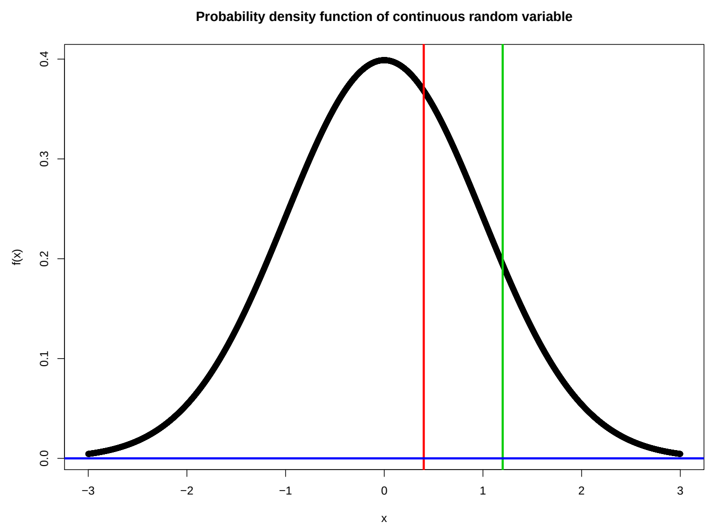
$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Expected value and variance. The expected value of a continuous random variable X is, roughly speaking, an appropriately weighted average of the possible values for X.[7] The variance is, roughly speaking, an appropriately

[6] Those who have taken calculus will recognize the definite integral and deduce that $f(x)$ is the derivative of the cumulative distribution function $F(x)$ at all points where the derivative exists.

[7] The expected value is $\int_{-\infty}^{+\infty} \{x \times f(x)\}\, dx$, provided that the integral converges.

Figure 3:



The area bounded above by the probability density function $f(x)$, bounded below by the blue horizontal line at 0, bounded on the left by the red vertical line at $a = 0.4$, and bounded on the right by the green vertical line at $b = 1.2$, equals $P(0.4 < X \leq 1.2)$.

weighted average of the possible squared deviations from the mean.[8] The standard deviation is the square root of the variance. Notations such as $E[X]$, μ, $Var[X]$, σ^2, $SD[X]$, and σ are used as before with discrete random variables.

Normal distributions

Definition and description. A continuous random variable X having the probability density function (Definition 5.5)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

[8]The variance is $\int_{-\infty}^{+\infty} \{(x-\mu)^2 \times f(x)\}\, dx$, provided that the integral converges.

is a normal random variable (or "has a normal distribution") with mean μ and variance σ^2.

The probability density function $f(x)$ is bell-shaped (Figure 5.5). The mean μ determines the center of the distribution (Figure 5.6), while the variance σ^2 determines the spread of the distribution (Figure 5.7). Figure 3 above shows the special case when $\mu = 0$ and $\sigma^2 = 1$.

Generally speaking, we have the following results for a normal random variable X:

$$P(\mu - \sigma < X \leq \mu + \sigma) \approx 0.68,$$

$$P(\mu - 1.96\sigma < X \leq \mu + 1.96\sigma) \approx 0.95,$$

$$\text{and} \quad P(\mu - 2.58\sigma < X \leq \mu + 2.58\sigma) \approx 0.99.$$

Standard normal distribution. In the special case when $\mu = 0$ and $\sigma^2 = 1$, we say (Definition 5.7) that X has a standard normal distribution. In this case, the cumulative distribution function is customarily denoted by $\Phi(x)$ rather than $F(x)$ (Definition 5.8, Figure 5.10). Moreover, the cumulative distribution function satisfies the symmetry relation $P(X \leq -x) = \Phi(-x) = 1 - \Phi(x) = P(X > x)$ (Equation 5.3, Figure 5.12).

Table 3 of Rosner. We may use Table 3 of Rosner to calculate probabilities involving a standard normal random variable X. For selected nonnegative numbers c, $P(X \leq c) = \Phi(c)$ is in row c column A and $P(X \leq -c) = 1 - \Phi(c)$ is in row c column B.

Example (Table 3 of Rosner). Suppose that X is a standard normal random variable. We have

$$P(X \leq 1.2) = \Phi(1.2) = 0.8849 \text{ (area left of green line in Figure 3)},$$

$$P(X \leq 0.4) = \Phi(0.4) = 0.6554 \text{ (area left of red line)},$$

$$P(0.4 < X \leq 1.2) = \Phi(1.2) - \Phi(0.4) = 0.2295 \text{ (area between red and green lines)},$$

$$P(X > 0.4) = 1 - \Phi(0.4) = 0.3446 \text{ (area right of red line)},$$

$$P(X \leq -0.4) = 1 - \Phi(0.4) = 0.3446 \text{ (equivalent to area right of red line)}.$$

Standardization. Let X be a normal random variable with mean μ and variance σ^2. Define

$$Z := \frac{X - \mu}{\sigma}.$$

Then Z is a standard normal random variable (Equation 5.4), which yields (Equation 5.5)

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

The above is an "all-purpose" equation for computing probabilities involving a normal random variable, provided we adopt the conventions that $\Phi(+\infty) = 1$ and $\Phi(-\infty) = 0$.

Percentiles. The $(100u)^{th}$ percentile (or "u quantile") of a standard normal distribution (Definition 5.10) is the number z_u satisfying $P(Z \leq z_u) = u$, where Z has a standard normal distribution.

Example (percentiles). We have $z_{0.95} = 1.645$ because $P(Z \leq 1.645) = 0.95$. We have $z_{0.975} = 1.96$ because $P(Z \leq 1.96) = 0.975$.

Central Limit Theorem

Motivation. Consider random variables $X_1, \ldots, X_n$ that will assume numerical values $x_1, \ldots, x_n$ after a sample is drawn. Let $\bar{X} := (X_1 + \cdots + X_n)/n$. Then $\bar{X}$ is a random variable that corresponds to the sample mean $\bar{x}$ in the same way that $X_1, \ldots, X_n$ correspond to $x_1, \ldots, x_n$. We will see in Lecture 4 that probabilistic statements about $\bar{X}$ can help us to make inferences about the population from which the sample is drawn. Hence, we want to know as much as we can about $\bar{X}$.

Expected value and variance. Suppose that $X_1, \ldots, X_n$ have common expected value $\mu = E[X_1] = \cdots = E[X_n]$ and common variance $\sigma^2 = Var[X_1] = \cdots = Var[X_n]$. Suppose also that $X_1, \ldots, X_n$ are independent in that, for example, knowing the value assumed by X_1 does not provide any hint about what value will be assumed by X_2.[9] These assumptions of a common expected value, a common variance, and independence are generally accepted when we speak of a simple random sample, in which any group of n individuals (or objects) has the same probability of being selected as any other group of n individuals (or objects).

With the above assumptions, we have (Equation 5.8) $E[\bar{X}] = \mu$ and (Equation 5.9) $Var[\bar{X}] = \sigma^2/n$. Thus, $\bar{X}$ has the same expected value as $X_1, \ldots, X_n$ but (if n is large) a much smaller variance.

[9] A formal mathematical definition of independence entails more than this, but an intuitive understanding is sufficient for the present purpose.

Central Limit Theorem. If, in addition to what we assumed earlier, X_1 through X_n are normally distributed, then $\bar{X}$ is also normally distributed. But what if X_1 through X_n are not normally distributed? What if, for example, X_1 through X_n are discrete random variables whose only possible values are 0 and 1?

The Central Limit Theorem says that, if n is large enough, then $\bar{X}$ is approximately normally distributed even if X_1 through X_n are not normally distributed (Equation 6.3)! The quality of the normal approximation depends both on n and on how "non-normal" X_1 through X_n are.

Example (Central Limit Theorem). Let Y be a binomial random variable based on n independent trials with common success probability p. We can write $Y = X_1 + \cdots + X_n$, where $X_i = 1$ if the i^{th} trial is successful and $X_i = 0$ if it is not. We can easily show that

$$\mu = E[X_1] = \cdots = E[X_n] = p \quad \text{and} \quad \sigma^2 = Var[X_1] = \cdots = Var[X_n] = p(1-p).$$

Since $Y/n = (X_1 + \cdots + X_n)/n = \bar{X}$, we may conclude that Y/n is approximately normally distributed with mean p and variance $p(1-p)/n$. For any nonnegative integers a and b with $b \geq a$ we have (Equation 5.14)

$$P(a \leq Y \leq b) = P(a - 1/2 \leq Y \leq b + 1/2) =$$

$$P([a-1/2]/n \leq Y/n \leq [b+1/2]/n) \approx \Phi\left(\frac{[b+1/2]/n - p}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{[a-1/2]/n - p}{\sqrt{p(1-p)/n}}\right).$$

The above approximation works well when $np(1-p) \geq 10$. The addition and subtraction of 1/2 are called a "continuity correction".[10]

[10] Consider approximating $P(a \leq Y \leq a) = P(Y = a)$. Without a continuity correction, our approximation would necessarily equal 0, which seems unreasonable since Y is a discrete random variable.