STA 580 — Fall 2008 — 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. But if a different group of 24 hospital employees had been selected, the serum cholesterol reductions might have been $15, 3, \ldots, 28$. Or $17, 34, \ldots, 9$. Or $6, 24, \ldots, -5$. Indeed, before a sample is selected, we do not actually 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. 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 potentially 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

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

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

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

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

Moreover,

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

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

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

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 just 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 just 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.

Example (expected value and variance). For the "Genetics" example,

$$E[X] = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1$$

and

$$Var[X] = (0 - 1)^2 \times 1/4 + (1 - 1)^2 \times 1/2 + (2 - 1)^2 \times 1/4 = 1/2.$$

[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.

Binomial distributions

Definition. Suppose that we conduct n independent "trials", each of which can have only two possible outcomes: "success" and "failure". If the probability of success at each trial is p (a number between 0 and 1), 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 $p = 1/2$. Note that the word success has a different meaning in this statistical context than in ordinary English parlance.

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 20 trials and success probability 0.90. Suppose, moreover, that we want to find $P(X \geq 18)$. Let $Y := 20 - X$, so that Y is binomial with 20 trials and success probability $0.10 = 1 - 0.90$. Since $\{X \geq 18\}$ is the same as $\{Y \leq 2\}$, we obtain $P(X \geq 18)$ from Table 1 as $P(Y = 0) + P(Y = 1) + P(Y = 2) =$

$0.1216 + 0.2702 + 0.2852 = 0.6770.$

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 (at least in principle!) are 142.47, 141.5924, and 142.061358. The possible values cannot be enumerated.

The cumulative distribution function (Definition 5.2) is denoted $F(x)$ and defined as $P(X \leq x)$, just as for discrete random variables. If the cumulative distribution function is continuous, then we say that X is a continuous random variable.[4]

Probability density function and inequalities. The probability density function for a continuous random variable X is denoted $f(x)$ and satisfies the following relations for any numbers a and b with $b > a$ (Definition 5.1):

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

[4]Definition 4.3 — inability to enumerate the possible values — is not quite enough. In fact, there exist random variables that are neither discrete nor continuous. However, we will not be concerned with them in STA 580.

where $\int_a^b f(x)\, dx$ is mathematical shorthand for "the area under the curve of $f(x)$ from a to b".[5]

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

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

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.[6] The variance is, roughly speaking, an appropriately weighted average of the possible squared deviations from the mean.[7] 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 in the same manner as with discrete random variables.

Normal distributions

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

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

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

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

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

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

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).

Standard normal distribution. If $\mu = 0$ and $\sigma^2 = 1$, then we say (Definition 5.7) that X has a standard normal distribution. We have (Figure 5.9)

$$P(-1 < X \leq 1) \approx 0.68,$$

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

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

The cumulative distribution function is customarily denoted by $\Phi(x)$ rather than $F(x)$ (Definition 5.8, Figure 5.10), and it 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 $b \geq 0$, $P(X \leq b) = \Phi(b)$ is in row b, column A. For $b < 0$, $P(X \leq b) = \Phi(b) = 1 - \Phi(|b|)$ is in row $|b|$, column B.

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

$$P(-1.65 < X \leq 1.65) = \Phi(1.65) - \Phi(-1.65) = 0.9505 - 0.0495 = 0.9010$$

$$\text{and} \quad P(X > -2) = 1 - P(X \leq -2) = 1 - \Phi(-2) = 1 - 0.0228 = 0.9772.$$

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), so that (Equation 5.5)

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

In particular,

$$P(\mu - \sigma < X \leq \mu + \sigma) \approx 0.68,$$
$$P(\mu - 1.96\sigma < X \leq \mu + 1.96\sigma) \approx 0.95,$$
$$\text{and}\quad P(\mu - 2.58\sigma < X \leq \mu + 2.58\sigma) \approx 0.99.$$

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 $z_{0.95} = 1.645$ and $z_{0.975} = 1.96$.

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.[8] 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.

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 de-

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

pends both on n and on how "non-normal" X_1 through X_n are. Even so, a common rule of thumb is that the Central Limit Theorem may be comfortably invoked once $n \geq 30$. To summarize, for $n \geq 30$ we can treat $\bar{X}$ as approximately normally distributed with mean μ and variance σ^2/n.

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

$$\mu = E[X_1] = \cdots = E[X_n] = p$$

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

Hence, 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)

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

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

The above approximation works particularly well when $np(1-p) \geq 10$.

The addition and subtraction of 1/2 are called a continuity correction. To understand this, consider approximating $P(a \leq X \leq a) = P(X = a)$. Surely we would not want the approximation to equal 0, which is what would happen if the continuity correction were omitted.