Statistics 252 (Winter 2007) January 8, 2007
Prof. Michael Kozdron

Review of Stat 251

The purpose of these notes is for me to outline what I believe are the basic and most fundamental concepts about random variables that every Stat 252 student ***MUST*** know. While the exposition in these notes varies from that in [1], the ideas are nonetheless the same. My suggestion to you is that you read these notes, and do the exercises *without* looking back at your Stat 251 material. Try and solve them using *only* what is given.

1 Review of Random Variables

Suppose that Ω is the sample space of outcomes of an experiment.

Example 1.1. Flip a coin once: $\Omega = \{H, T\}$.

Example 1.2. Toss a die once: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 1.3. Toss a die twice: $\Omega = \{(i, j) : 1 \le i \le 6,\ 1 \le j \le 6\}$.

Note that in each of the previous three examples, Ω is a *finite* set. In the next example, however, Ω is an *uncountably infinite* set.

Example 1.4. Consider a needle attached to a spinning wheel centred at the origin. When the wheel is spun, the angle ω made with the tip of the needle and the positive x-axis is measured. The possible values of ω are $\Omega = [0, 2\pi)$.

Definition 1.5. A *random variable* X is a function from the sample space Ω to the real numbers $\mathbb{R} = (-\infty, \infty)$. Symbolically, $X : \Omega \to \mathbb{R}$ via

$$\omega \in \Omega \mapsto X(\omega) \in \mathbb{R}.$$

Example 1.1 (continued). Let X denote the number of heads on a single flip of a coin. Then, $X(H) = 1$ and $X(T) = 0$.

Example 1.2 (continued). Let X denote the upmost face when a die is tossed. Then, $X(i) = i$, $i = 1, \ldots, 6$.

Example 1.3 (continued). Let X denote the sum of the upmost faces when two dice are tossed. Then, $X((i, j)) = i + j$ for $i = 1, \ldots, 6$, $j = 1, \ldots, 6$. Note that the elements of Ω are ordered pairs, so that the function $X(\cdot)$ acts on (i, j) giving $X((i, j))$. We will often omit the inner parentheses and simply write $X(i, j)$.

Example 1.4 (continued). Let X denote the cosine of the angle made by the needle on the spinning wheel and the positive x-axis. Then $X(\omega) = \cos(\omega)$ so that $X(\omega) \in [-1, 1]$.

Remark. The use of the notation X and $X(\omega)$ is EXACTLY analogous to elementary calculus. There, the function f is described by its action on elements of its domain. For example, $f(x) = x^2$, $f(t) = t^2$, and $f(\omega) = \omega^2$ all describe EXACTLY the same function, namely, the function which takes a number and squares it.

Remark. For historical reasons, the term *random variable* (written X) is used in place of *function* (written f) and generic elements of the domain are denoted by ω instead of by x.

Remark. If X is a random variable, then we call $X(\omega)$ a *realization* of the random variable. The physical interpretation is that if X denotes the UNKNOWN outcome (*a priori*) of the experiment before it happens, then $X(\omega)$ represents the realization or observed outcome (*a posterior*) of the experiment after it happens.

Remark. It was A.N. Kolmogorov in the 1930's who formalized probability and realized the need to treat random variables as *measurable* functions. See Math 810: Analysis I or Stat 851: Probability.

Suppose that X is a random variable. The *distribution function* of X is the function $F : \mathbb{R} \to \mathbb{R}$ given by

$$F(t) := P(X \le t) := P(\{\omega : X(\omega) \le t\}).$$

Sometimes F is called the *probability distribution function* of X.

It is a fact that the distribution function exists for every random variable. Furthermore, the following theorem characterizes distribution functions.

Theorem 1.6. *A function $F : \mathbb{R} \to \mathbb{R}$ is a distribution function for some random variable X if and only if the following hold:*

- $\lim_{t \to -\infty} F(t) = 0,$
- $\lim_{t \to \infty} F(t) = 1,$ *and*
- *F is non-decreasing and right-continuous.*

2 Discrete and Continuous Random Variables

There are two extremely important classes of random variables, namely the so-called discrete and continuous. In a sense, these two classes are the same since the random variable is described in terms of a density function. However, there are slight differences in the handling of sums and integrals so these two classes are taught separately in undergraduate courses.

Important Observation. Recall from elementary calculus that the Riemann integral

$$\int_a^b f(x)\, dx$$

is *defined* as an appropriate limit of Riemann sums $\sum_{i=1}^N f(x_i^*)\ \Delta x_i$. Thus, you are ALREADY FAMILIAR with the fact that SOME RELATIONSHIP exists between integrals and sums.

Definition 2.1. A random variable X is said to be *discrete* if it can assume only a finite or countably infinite number of distinct values.

Definition 2.2. Suppose that X is a discrete random variable. Suppose that there exists a function $p : \mathbb{Z} \to \mathbb{R}$ with the properties that $p(k) \geq 0$ for all k, $\sum_{k=-\infty}^{\infty} p(k) = 1$, and

$$P(\{\omega \in \Omega : X(\omega) \leq N\}) =: P(X \leq N) = \sum_{k=-\infty}^{N} p(k).$$

We call p the *(probability mass function or) density* of X. Note that $p(k) = P(X = k)$.

Example 1.3 (continued). If X is defined to be the sum of the upmost faces when two dice are tossed, then the density of X, written $p(k) = P(X = k)$, is given by

$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$	$p(11)$	$p(12)$
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and $p(k) = 0$ for any other $k \in \mathbb{Z}$.

Remark. Note that the distribution function of a discrete random variable is piecewise constant. If, however, the distribution function is continuous, we call the corresponding random variable continuous.

Definition 2.3. We say that a random variable X is *continuous* if the distribution function of X is a continuous function. (Note that continuous random variables are sometimes called *absolutely continuous*.)

Definition 2.4. Suppose that X is a continuous random variable. Suppose that there exists a function $f : \mathbb{R} \to \mathbb{R}$ with the properties that $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f(x)\, dx = 1$, and

$$P(\{\omega \in \Omega : X(\omega) \leq t\}) =: P(X \leq t) = \int_{-\infty}^{t} f(x)\, dx.$$

We call f the *(probability) density (function)* of X.

Fact. By the Fundamental Theorem of Calculus, $F'(x) = f(x)$.

Exercise 2.5. Prove the fact that $F'(x) = f(x)$, being sure to carefully state the necessary assumptions on f. Convince me that you understand the use of the dummy variables x and t in your argument.

Remark. There exist continuous random variables which *do not* have densities. For our purposes, though, we will always assume that our continuous random variables are ones with a density.

Example 2.6. A random variable X is said to be *normally distributed with parameters μ and σ^2*, if the density of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < \mu < \infty,\ 0 < \sigma < \infty.$$

This is sometimes written $X \sim \mathcal{N}(\mu, \sigma^2)$. In Exercises 3.4 and 4.9, you will show that the mean of X is μ and the variance of X is σ^2, respectively.

Remark. There do exist random variables which are neither discrete nor continuous; however, such random variables will not concern us.

Remark. If we know the distribution of a random variable, then we know all of the information about that random variable. For example, if we know that X is a normal random variable with parameters 0 and 1, then we know everything possible about X without actually realizing it.

3 Law of the Unconscious Statistician

Suppose that $X : \Omega \to \mathbb{R}$ is a random variable (either discrete or continuous), and that $g : \mathbb{R} \to \mathbb{R}$ is a (piecewise) continuous function. Then $Y := g \circ X : \Omega \to \mathbb{R}$ defined by $Y(\omega) = g(X(\omega))$ is also a random variable.

We now define the expectation of the random variable Y, distinguishing the discrete and continuous cases.

Definition 3.1. Suppose that X is a discrete random variable with probability mass function p. If g is as above and

$$\sum_k |g(k)|\, p(k) < \infty,$$

then we say that X has *finite expectation* and we define the *expectation* of $g \circ X$ to be the number

$$\mathbb{E}(g \circ X) := \sum_k g(k)\, p(k).$$

On the other hand, if

$$\sum_k |g(k)|\, p(k) = \infty$$

then we say that X has *infinite expectation* and we define $\mathbb{E}(g \circ X) := \infty$.

Definition 3.2. Suppose that X is a continuous random variable with probability density function f. If g is as above and

$$\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx < \infty,$$

then we say that X has *finite expectation* and we define the *expectation* of $g \circ X$ to be the number

$$\mathbb{E}(g \circ X) := \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

On the other hand, if

$$\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx = \infty,$$

then we say that X has *infinite expectation* and we define $\mathbb{E}(g \circ X) := \infty$.

It is worth noting that $g \circ X$ has finite expectation if and only $\mathbb{E}(\,|g \circ X|\,) < \infty$.

Remark. Exercise 4.2 below provides an example of a continuous random variable with infinite expectation. It also illustrates the importance of the hypothesis that the sum or integral defining the expectation be absolutely convergent.

Notice that if $g(x) = x$ for all x and $\mathbb{E}(\,|g \circ X|\,) < \infty$, then the expectation of X itself is

- $\mathbb{E}(X) := \sum_k k\, p(k)$, if X is discrete, and
- $\mathbb{E}(X) := \int_{-\infty}^{\infty} x\, f(x)\, dx$ if X is continuous.

Exercise 3.3. Suppose that X is a Bernoulli(p) random variable. That is, $P(X = 1) = p$ and $P(X = 0) = 1 - p$ for some $p \in [0, 1]$. Carefully verify that

- $\mathbb{E}(X) = p$,
- $\mathbb{E}(X^2) = p$, and
- $\mathbb{E}(e^{\theta X}) = 1 - p(1 - e^{\theta})$, for $0 \le \theta < \infty$.

Exercise 3.4. The purpose of this exercise is to make sure you can compute some straightforward (but messy) integrals. Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$; that is, X is a normally distributed random variable with parameters μ, σ^2. (See Example 2.6 for the density of X.) Show directly (without using any unstated properties of expectations or distributions) that

- $\mathbb{E}(X) = \mu$,
- $\mathbb{E}(X^2) = \sigma^2 + \mu^2$, and
- $\mathbb{E}(e^{\theta X}) = \exp\left\{\theta\mu + \dfrac{\sigma^2\theta^2}{2}\right\}$, for $0 \le \theta < \infty$.

Together with Exercise 4.9, this is the reason that if $X \sim \mathcal{N}(\mu, \sigma^2)$, we say that X is normally distributed with mean μ and variance σ^2.

4 Summarizing Random Variables

Definition 4.1. If X is a random variable and $\mathbb{E}(\,|X|\,) < \infty$ (that is, X has finite expectation), then the *mean* of X is the number $\mu := \mathbb{E}(X)$, and we say that X has a *finite mean*, or that X is an *integrable random variable*, written $X \in L^1$. If $\mathbb{E}(\,|X|\,) = \infty$ (that is, X has infinite expectation), then we say that X *does not have finite mean* and write $X \notin L^1$.

Exercise 4.2. Suppose that X is a Cauchy-distributed random variable. That is, X is a continuous random variable with density function

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{x^2+1}.$$

Carefully show that $X \notin L^1$; that is, show that $E(\,|X|\,) = \infty$.

Definition 4.3. If X is a random variable with $\mathbb{E}(X^2) < \infty$, then we say that X has a *finite second moment* and write $X \in L^2$. If $X \in L^2$, then we define the *variance* of X to be the number $\sigma^2 := \mathbb{E}\left((X-\mu)^2\right)$. The *standard deviation* of X is the number $\sigma := \sqrt{\sigma^2}$. (As usual, this is the positive square root.)

Remark. It is an important fact that if $X \in L^2$, then it must be the case that $X \in L^1$. This follows from the so-called Cauchy-Schwarz Inequality. (See Exercises 4.18 and 4.19 below.)

Definition 4.4. If X and Y are both random variables in L^2, then the *covariance* of X and Y, written $\mathrm{Cov}(X,Y)$ is defined to be

$$\mathrm{Cov}(X,Y) := \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right)$$

where $\mu_X := \mathbb{E}(X)$, $\mu_Y := \mathbb{E}(Y)$. Whenever the covariance of X and Y exists, we define the *correlation* of X and Y to be

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{$\dagger$}$$

where σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Remark. By *fiat*, $0/0 := 0$ in ($\dagger$). Although this is sinful in calculus, we advanced mathematicians understand that such a decree is permitted as long as we recognize that it is only a *convenience* which allows us to simplify the formula. We need not bother with the extra conditions about dividing by zero. (See Exercise 4.20.)

Definition 4.5. We say that X and Y are *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$ (or, equivalently, if $\mathrm{Corr}(X,Y) = 0$).

Theorem 4.6 (Linearity of Expectation). *Suppose that $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are (discrete or continuous) random variables with $X \in L^1$ and $Y \in L^1$. Suppose also that $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are both (piecewise) continuous and such that $g \circ X \in L^1$ and $h \circ Y \in L^1$. Then, $g \circ X + h \circ Y \in L^1$ and, furthermore,*

$$\mathbb{E}(g \circ X + h \circ Y) = \mathbb{E}(g \circ X) + \mathbb{E}(h \circ Y).$$

Exercise 4.7. Prove the above theorem separately for both the discrete case and the continuous case. Be sure to state any assumptions or theorems from elementary calculus that you use.

Fact. If $X \in L^2$ and $Y \in L^2$, then the following computational formulæ hold:

- $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$;
- $\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \sigma^2$;
- $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Exercise 4.8. Verify the three computational formulæ above.

Exercise 4.9. Using the third computational formula, and the results of Exercise 3.4, quickly show that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathrm{Var}(X) = \sigma^2$. Together with Exercise 3.4, this is the reason that if $X \sim \mathcal{N}(\mu, \sigma^2)$, we say that X is normally distributed with mean μ and variance σ^2.

Definition 4.10. The random variables X and Y are said to be *independent* if $f(x, y)$, the joint density of (X, Y), can be expressed as

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

where f_X is the density of X and f_Y is the density of Y.

Remark. Notice that we have combined the cases of a discrete and a continuous random variable into one definition. You can substitute the phrases *probability mass function* or *probability density function* as appropriate.

The following is an extremely deep, and important, result.

Theorem 4.11. *If X and Y are* independent *random variables with $X \in L^1$ and $Y \in L^1$, then*

- *the product XY is a random variable with $XY \in L^1$, and*
- $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$.

Exercise 4.12. Using this theorem, quickly prove that if X and Y are independent random variables, then they are necessarily uncorrelated. (As the next exercise shows, the converse, however, is not true: there do exist uncorrelated, dependent random variables.)

Exercise 4.13. Consider the random variable X defined by $P(X = -1) = 1/4$, $P(X = 0) = 1/2$, $P(X = 1) = 1/4$. Let the random variable Y be defined as $Y := X^2$. Hence, $P(Y = 0|X = 0) = 1$, $P(Y = 1|X = -1) = 1$, $P(Y = 1|X = 1) = 1$.

- Show that the density of Y is $P(Y = 0) = 1/2$, $P(Y = 1) = 1/2$.
- Find the joint density of (X, Y), and show that X and Y are not independent.

- Find the density of XY, compute $\mathbb{E}(XY)$, and show that X and Y are uncorrelated.

Exercise 4.14. Prove Theorem 4.11 in the case when both X and Y are continuous random variables.

Exercise 4.15. Suppose that $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are independent, integrable, continuous random variables with densities f_X, f_Y, respectively. Let $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be continuous functions such that $g \circ X \in L^1$ and $h \circ Y \in L^1$. Prove that $\mathbb{E}\left((g \circ X) \cdot (h \circ Y)\right) = \mathbb{E}(g \circ X)\,\mathbb{E}(h \circ Y)$.

As a consequence of the previous exercise, we have the following very important result.

Theorem 4.16 (Linearity of Variance when Independent). *Suppose that $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are (discrete or continuous) random variables with $X \in L^2$ and $Y \in L^2$. If X and Y are* independent, *then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

It turns out that Theorem 4.11 is not quite true when X and Y are not independent. However, the following is a probabilistic form of the ubiquitous Cauchy-Schwarz inequality, and usually turns out to be good enough.

Theorem 4.17 (Cauchy-Schwarz Inequality). *Suppose that X and Y are both random variables with finite second moments. That is, $X \in L^2$, and $Y \in L^2$. It then follows that*

- *the product XY is a random variable with $XY \in L^1$, and*
- $(\mathbb{E}(XY))^2 \le \mathbb{E}(X^2)\,\mathbb{E}(Y^2)$, *and*
- $(\mathrm{Cov}(X, Y))^2 \le \mathrm{Var}(X)\ \mathrm{Var}(Y)$.

Exercise 4.18. Using the first part of the Cauchy-Schwarz inequality, show that if $X \in L^2$, then $X \in L^1$.

Exercise 4.19. Using the second part of the Cauchy-Schwarz inequality, show that if $X \in L^2$, then $X \in L^1$.

Exercise 4.20. Using the third part of the Cauchy-Schwarz inequality, you can now make sense of the Remark following Definition 4.4. Show that if X and Y are random variables with $\mathrm{Var}(X) = \mathrm{Var}(Y) = 0$, then $\mathrm{Cov}(X, Y) = 0$.

The following facts are also worth mentioning.

Theorem 4.21. *If $a \in \mathbb{R}$ and $X \in L^2$, then $aX \in L^2$ and $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$. In particular, $\mathrm{Var}(-X) = \mathrm{Var}(X)$.*

Theorem 4.22. *If $X_1, X_2, \dots, X_n$ are L^2 random variables, then*

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n \operatorname{Cov}(X_i, X_j) = \sum_{k=1}^n \operatorname{Var}(X_k) + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

In particular, if $X_1, X_2, \dots, X_n$ are uncorrelated *L^2 random variables, then*

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var}(X_i).$$

Remark. The previous theorem is a generalization of Theorem 4.16. In the particular case of two random variables we have the following. If X and Y are L^2 random variables, then

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$$

and

$$\operatorname{Var}(X-Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) - 2\operatorname{Cov}(X,Y).$$

Theorem 4.23. *If $X_1, \dots, X_n, Y_1, \dots, Y_m$ are L^2 random variables, and $a_1, \dots, a_n, b_1, \dots, b_m$ are constants, then*

$$\operatorname{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \operatorname{Cov}(X_i, Y_j).$$

Remark. Although variance is *not* a linear operator (as shown by Theorems 4.21 and 4.22), this last result shows that covariance *is* a linear operator. The statement of Theorem 4.23 might be a little confusing to understand, but with a bit of practice, it turns out to be rather easy to apply. The next exercise illustrates this.

Exercise 4.24. Suppose that X_1, X_2, X_3 are normal random variables each with a $\mathcal{N}(0,1)$ distribution. Suppose further that $\operatorname{Cov}(X_1, X_2) = -2$, $\operatorname{Cov}(X_1, X_3) = -2$, and $\operatorname{Cov}(X_2, X_3) = 2$. Let $Z_1 = 2X_1 + X_2$, and let $Z_2 = X_2 - 2X_3$. Compute $\operatorname{Cov}(Z_1, Z_2)$.