Suppose you reach into your pocket and pull out a coin; think of one side as "heads" and the other as "tails." You toss the coin three times. We could now define some random variables, at least informally, e.g.,

- the number of heads on the first toss (which is 0 or 1),
- the total number of heads (which is 0, 1, 2, or 3).

This experiment is a little hypothetical—hence the subjunctive voice. If we actually did it and got "head tail head," the *observed value* of our first random variable—the number of heads on the first toss—would be 1; the observed value of the other random variable—the total number of heads—would be 2.

Even after the experiment is done and the data are collected, we can still think about the things that could have happened but didn't. Random variables and probabilities are mainly *not* about actual outcomes. Rather, these concepts apply to an experiment and the various ways it could have turned out. In fact, most statistical calculations (e.g., confidence levels and significance levels) apply to the experiment and all of its possible outcomes, including the ones that did not materialize.

Reaching into pockets and pulling out coins are not mathematical ideas. The connection between real coins and mathematical models is a complicated topic. To continue the discussion, we need to start from another angle. A.N. Kolmogorov laid the foundations for modern probability theory with the "sample space representation" of probabilities. For the coin, this amounts to a list of the $2^3 = 8$ possible outcomes:

H H H
H H T
H T H
H T T
T H H
T H T
T T H
T T T

For instance, H T H corresponds to the sequence "heads tails heads."

Kolmogorov defined random variables as functions on the sample space. Here are some examples:

	X_1	X_2	X_3	S
H H H	1	1	1	3
H H T	1	1	0	2
H T H	1	0	1	2
H T T	1	0	0	1
T H H	0	1	1	2
T H T	0	1	0	1
T T H	0	0	1	1
T T T	0	0	0	0

- X_1 is the number of heads on the first toss,
- X_2 is the number of heads on the second toss,
- X_3 is the number of heads on the third toss,
- $S = X_1 + X_2 + X_3$ is the total number of heads.

Mathematically, a random variable is a function on the sample space.
Doing arithmetic on random variables gives you more random variables.

According to Kolmogorov, a probability assigns numbers to outcomes: these numbers are non-negative, and their sum has to be 1. (This is for the "discrete" case; other cases are discussed below.) If A is an "event," i.e., a subset of sample space, the probability of A is by definition the sum of the probabilities of the outcomes in A. Thus, $P\{A\} = \sum_{\omega \in A} P\{\omega\}$.

For example, it may be natural to assign probability 1/8 to each of the 8 outcomes. More generally, if we think the tosses are independent but the chance of heads is p, the probability assignment would be as follows:

	Probability
H H H	p^3
H H T	$p^2(1-p)$
H T H	$p^2(1-p)$
H T T	$p(1-p)^2$
T H H	$p^2(1-p)$
T H T	$p(1-p)^2$
T T H	$p(1-p)^2$
T T T	$(1-p)^3$

This is a "statistical model," in the sense that the probabilities depend on a parameter p that may be estimated from data.

For mathematicians, Kolmogorov brought great clarity to an area that was previously quite murky. For applied workers, the setup may be less satisfying. The connection between the physical coin and the mathematical coin takes some getting used to. The number of heads seems to be in one realm, the function on the sample space in quite another. Furthermore, "observed values"—which represent the data that statisticians have to work with—remain outside the formalism.

To what extent does the statistical model of a coin correspond to a real coin? For a few thousand tosses, the fit is excellent, in the sense that observed frequencies match expected frequencies, with margins of error that are in line with the theory. For instance, take John Kerrich's coin (Freedman-Pisani-Purves 2007, Chapter 16). With 10,000 tosses and $p = 1/2$, the expected number of heads is 5,000 and the standard error is 50: these are computed from the model (see below). Kerrich observed 5,067 heads on the real coin. Many similar tests can be done on his data—e.g., we could look at the frequency of pairs like "heads tails," or triples like "heads tails heads"—with similar results.

With 100,000 tosses, the independence assumptions built into the model start to conflict with the data: there will be small serial correlations. Furthermore, the probability of heads may well differ from 50%, at least to a degree that is detectable. With similar caveats, the model also works

for dice, roulette wheels, and so forth. (Dice need to have counter-filled spots, roulette wheels have to be balanced. . . .) For card shuffling, the usual model of randomness (drawing at random without replacement, so that all permutations of the deck are equally likely) does not apply so well: it takes many, many shuffles to bring a deck of cards into anything like random order—a fact that professional card players may exploit to their advantage.

What is the expectation of a random variable?

A random variable X has *expectation* and *variance*, denoted $E(X)$ and $\mathrm{var}(X)$ respectively:

$$\mathrm{var}(X) = E\left\{[X - E(X)]^2\right\} = E(X^2) - [E(X)]^2.$$

The *standard error* of X is $\sqrt{\mathrm{var}(X)}$. The standard error is abbreviated as SE; "expected value" is a frequent synonym for expectation.

> A random error will be somewhere around its expected value. The SE gauges the likely size of the amount off.

Rule 1: the discrete case. Suppose $P\{X = x_i\} = p_i$ for $i = 1, 2, \ldots$ and $\sum_i p_i = 1$. Then $E(X) = \sum_i x_i p_i$, $E(X^2) = \sum_i x_i^2 p_i$, and so forth. More generally, $E\{g(X)\} = \sum_i g(x_i) p_i$.

Rule 2: the absolutely continuous case. If X has density f, i.e., $P\{X \le x\} = \int_{-\infty}^{x} f(u)\,du$, then $E(X) = \int_{-\infty}^{\infty} u f(u)\,du$, $E(X^2) = \int_{-\infty}^{\infty} u^2 f(u)\,du$, and so forth. More generally, $E\{g(X)\} = \int_{-\infty}^{\infty} g(u) f(u)\,du$.

Rule 3. If a is a real number, then $E(aX) = aE(X)$.

Rule 4. $E(X + Y) = E(X) + E(Y)$.

Rule 5. If a is a real number, then $\mathrm{var}(aX) = a^2\mathrm{var}(X)$. Standard errors are often more readily interpretable than variances. For instance, the standard error of aX is $|a|$ times the standard error of X.

Rule 6. $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X, Y)$, where

$$\mathrm{cov}(X, Y) = E\left\{[X - E(X)][Y - E(Y)]\right\} = E(XY) - E(X)E(Y).$$

Caveats

(i) In Rule 2, the function g has to be reasonable (e.g., Borel measurable—see below). Of course, f has to be reasonable too.

(ii) If a random variable has infinitely many values and the distribution has "long tails," the expectation may be infinite or ill-defined. This is a concern in probability courses, but is usually not an issue in applied statistics. See Nassim Taleb's book, *The Black Swan*, for examples (drawn from finance) where long tails really matter.

Conditional distributions and expectations

The formal definition of conditional probability, even for events, is somewhat opaque:

$$P\{A|B\} = \frac{P\{A \text{ and } B\}}{P\{B\}}$$

It takes a real effort to see that the definition matches the intuition—the probability of A, given that B occurs. Working through some examples is the best way to go. In essence, $P\{\bullet|B\}$ is a new probability assignment on the sample space. Probabilities outside B are reset to 0; inside B, probabilities are renormalized so the sum is 1.

The *conditional distribution* of Y given X is the distribution of Y, given the value of X. In the discrete case, this is just $P\{Y = y|X = x\}$. The *conditional expectation* of Y given X is

$$E(Y|X = x) = \sum_y y P\{Y = y|X = x\}$$

The sum will be finite or countable.

In the absolutely continuous case, the pair (X, Y) has a density f, i.e.,

$$P\{X \le x \text{ and } Y \le y\} = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\,du\,dv$$

Then X has density g and Y has density h:

$$g(x) = \int_{-\infty}^{\infty} f(x, v)\,dv, \quad h(y) = \int_{-\infty}^{\infty} f(u, y)\,du$$

(These are sometimes called "marginal" densities, by contrast with the "joint" density f.) Furthermore, Y has a conditional density given that $X = x$, viz., $h(y|x) = f(x, y)/g(x)$. Said another way, the conditional distribution of Y given $X = x$ has the density $h(y|x)$. For instance,

$$P\{Y \le w|X = x\} = \int_{-\infty}^{w} h(y|x)\,dy, \quad E(Y|X = x) = \int_{-\infty}^{\infty} yh(y|x)\,dy,$$

$$E(Y^2|X = x) = \int_{-\infty}^{\infty} y^2 h(y|x)\,dy, \quad E\big(g(Y)|X = x\big) = \int_{-\infty}^{\infty} g(y)h(y|x)\,dy$$

What is independence?

Suppose we make two draws at random from the box $\boxed{1}\ \boxed{2}\ \boxed{2}\ \boxed{5}$. Let X be the first draw, and Y the second.

(i) Suppose the draws are made with replacement.

If $X = 1$, the chance that $Y = 5$ is 1/4.
If $X = 2$, the chance that $Y = 5$ is 1/4.
If $X = 5$, the chance that $Y = 5$ is 1/4.

This is *independence*: $P\{Y = y|X = x\}$ is the same for all x. (Equality has to hold for each y, not just one special y; this definition is good only in the discrete case.) Here is an equivalent condition.

Factorization. Discrete random variables X and Y are independent provided

$$P\{X = x \text{ and } Y = y\} = P\{X = x\}P\{Y = y\} \text{ for all } x \text{ and } y.$$

(ii) In the box example, if the draws are made without replacement, the two random variables are dependent: $P\{Y = y|X = x\}$ may be different for different x's.

If $X = 1$, the chance that $Y = 5$ is 1/3.
If $X = 2$, the chance that $Y = 5$ is 1/3.
If $X = 5$, the chance that $Y = 5$ is 0.

The sample space representation for the draws is shown below.

	X	Y	Probability with replacement	Probability without replacement
1 1	1	1	1/16	0/12
1 2	1	2	2/16	2/12
1 5	1	5	1/16	1/12
2 1	2	1	2/16	2/12
2 2	2	2	4/16	2/12
2 5	2	5	2/16	2/12
5 1	5	1	1/16	1/12
5 2	5	2	2/16	2/12
5 5	5	5	1/16	0/12

It may be irritating to compute the probabilities. One way to do it is to label the four tickets as a, b, c, d; then display the possible outcomes, as below. (Notice that b and c label the two tickets marked "2.") All possible pairs are equally likely. Without replacement, the pairs $a\,a$, $b\,b$, $c\,c$, $d\,d$ cannot occur. That is what the blanks mean. The body of each table shows the values of X and Y. For example, with replacement, the chance that $X = 1$ and $Y = 2$ is 2/16, because this pattern occurs in two out of the 16 entries. Without replacement, the chance is 2/12.

With replacement

	a	b	c	d
a	1 1	1 2	1 2	1 5
b	2 1	2 2	2 2	2 5
c	2 1	2 2	2 2	2 5
d	5 1	5 2	5 2	5 5

Without replacement

	a	b	c	d
a		1 2	1 2	1 5
b	2 1		2 2	2 5
c	2 1	2 2		2 5
d	5 1	5 2	5 2	

In the absolutely continuous case, we have to start over, because $P\{X = x\} = 0$ for all x and 0/0 is ill-defined ($0 \bullet x = 0$ for all x). Suppose the pair (X, Y) has a joint density f. The independence condition here is that $h(y|x)$ is the same for all x, where h is the conditional density of Y given $X = x$, as discussed above. An equivalent condition is factorization.

Absolutely continuous random variables X, Y are independent provided the joint density f factors: $f(x, y) = g(x)h(y)$ for all x, y. (Minor technical difficulties are elided.)

Danger ahead. Independence is special. Many statistical calculations ride on the assumption of independence. This assumption is often left implicit. In applications, independence is often questionable.

Notation

$X \perp\!\!\!\perp Y$ means that X and Y are independent.

If $E(Y|X = x) = \varphi(x)$, we often write $E(Y|X) = \varphi(X)$.

Sums of independent variables

Proposition. If $X \perp\!\!\!\perp Y$ then

(i) $E(XY) = E(X)E(Y)$,

(ii) $\text{cov}(X, Y) = 0$,

(iii) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$.

Let's restate (iii). If the standard error (SE) of X is σ and the SE of Y is τ, and the two variables are independent, then the SE of $X + Y$ is $\sqrt{\sigma^2 + \tau^2}$: this is the *square root law*. The *correlation* between X and Y is the covariance divided by the product of the standard errors. If the covariance is 0, the variables are "uncorrelated:" independence implies a correlation of 0. The converse holds for jointly normal variables, but not in general.

Corollary. Suppose $X_1, X_2, \ldots$ are independent and identically distributed (IID). Let $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then

(i) $E(S_n) = n\mu$,

(ii) $\text{var}(S_n) = n\sigma^2$.

In translation: (i) The sum of IID random variables has expected value equal to n times the common expected value of the summands (n being the number of summands). (ii) The standard error of the sum is $\sqrt{n}$ times the common standard error of the summands. Thus, the uncertainty in the sum increases rather slowly relative to the expected value—the square root law in action.

Let $\overline{X} = S_n/n$ be the average of $X_1, X_2, \ldots, X_n$: this is often called "the sample average" or "the sample mean." Then $E\big(\overline{X}\big) = n\mu/n = \mu$ and $\text{var}\big(\overline{X}\big) = n\sigma^2/n^2 = \sigma^2/n$. A more helpful statement, perhaps, is this: $\overline{X}$ has expectation μ and standard error $\sigma/\sqrt{n}$. When n is large, $\overline{X}$ is going to be close to its expected value. This is the *law of large numbers*. (Remember, the X_i are IID by assumption.)

Terminology

$E(X)$ is often referred to as the "mean" of X. A statistician with a tin ear may therefore say, "the mean of the sample mean is the population mean." Mathematically, this translates to the equation $E\big(\overline{X}\big) = \mu$. In other words, if you take the expected value of the average of independent, identically distributed random variables, you get their common expected value.

Some examples

Example 1. A coin lands heads with probability p. It is tossed once. Let X be the number of heads (0 or 1). Find $E(X)$ and $\text{var}(X)$. To begin with, $E(X) = p$ by Rule 1 for computing expected values. But $X^2 = X$, so $E(X^2) = p$ as well. Now $\text{var}(X) = p - p^2 = p(1 - p)$.

Example 2. A coin lands heads with probability p. It is tossed n times. (The implicit assumption: tosses are independent, and the probability of heads stays the same from one toss to another.) Let X be the number of heads. How big is X?

The distribution of X is Bin(n, p), i.e., binomial with parameters n and p:

$$P\{X = j\} = \binom{n}{j} p^j (1-p)^{n-j}$$

Moreover, $E(X) = np$ and $\text{var}(X) = np(1-p)$: see the previous example, and the Corollary before that. So X is around np, but is off by something like $\sqrt{np(1-p)}$.

The *central limit theorem* makes this more precise: as n gets large,

$$P\left\{np - z\sqrt{np(1-p)} < X < np + z\sqrt{np(1-p)}\right\}$$

converges to

$$\frac{1}{\sqrt{2\pi}} \int_{-z}^{+z} e^{-x^2/2}\, dx, \qquad (*)$$

the area under the normal curve between $-z$ and $+z$.

More generally, suppose X_i are IID, $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$, and $S_n = X_1 + \cdots + X_n$. As n gets large,

$$P\left\{n\mu - z\sigma\sqrt{n} < S_n < n\mu + z\sigma\sqrt{n}\right\}$$

converges to $(*)$. Independence is a key assumption here: the central limit theorem need not hold if the X_i are merely uncorrelated.

Example 3. A coin is tossed 10 times. (The implicit assumptions: the tosses are independent, and the chance of heads is 1/2.) Let S be the number of heads among the first 6 tosses. Let T be the number of heads among the last 6 tosses. Are S and T independent? Clearly not: they have in common the number of heads on tosses 5 and 6. However, S and T are functionally independent: you cannot compute one from the other. Statistical independence is a much stronger condition than functional independence.

Example 4. As before, a coin is tossed 10 times. Let X be $+1$ if the first toss is heads; else, $X = -1$. So $E(X) = 0$. Let V be the number of heads among the last 9 tosses. So $E(V) = 4.5$.

(a) Are X and V independent? Yes: X comes from the first toss, V from the last 9 tosses, and the tosses are independent.

(b) Let $U = XV$. Are U and V independent? Clearly not. Among other things, $|U| = V$. So U tells you everything there is to know about V. On the other hand, U and V are uncorrelated. Indeed, $E(UV) = 0$ because $E(UV) = E(XV^2) = E(X)E(V^2) = 0 \bullet E(V^2) = 0$. Furthermore, $E(U) = 0$ so $E(U)E(V) = 0$. Consequently, $\text{cov}(U, V) = 0$. The take-home message: uncorrelated variables can be dependent. Independence is a strong condition. Zero correlation is a weaker condition.

Example 5. A coin is tossed 3 times; the probability of heads is p. Let S be the number of heads. Given that $S = 2$, what is the conditional probability of H H T? There are only three sequences with $S = 2$, viz., H H T, H T H, T H H. Unconditionally, they are all equally likely: each has chance $p^2(1-p)$. Hence, each has conditional probability 1/3. Unconditionally, the tosses are independent. Conditionally, they are dependent.

Example 6. Let U, V, W be independent random variables, each having positive variance.

(a) Are $U + W$ and $V + W$ independent? Clearly not: they have W in common.

(b) Are $U + W$ and $V + W$ conditionally independent, given W? Yes: conditioning on W converts W to a constant, but does not affect the distribution of U and V. For any value w of W, the variables $U + w$ and $V + w$ are independent, which completes the argument for conditional independence. (By strict mathematical standards, the arguments in these notes may seem a little informal—but these arguments can made entirely rigorous, if that is desired.)

Example 7. A coin is tossed 10 times. (Implicit assumptions: the tosses are independent, and the chance of heads on each toss is 1/2.) Let X be the number of heads on the first 5 tosses, and Y the total number of heads. What is $E(Y|X = x)$? $\text{var}(Y|X = x)$? Conditionally, if $X = x$, then Y is distributed like $x + Z$, where Z is the number of heads in 5 tosses of a coin. So $E(Y|X = x) = x + 2.5$ and $\text{var}(Y|X = x) = 5/4$. Here, the possible values of x are $0, 1, \ldots, 5$.

Example 8. Suppose (X, Y) is a point chosen at random in the unit square.

(a) Are X and Y independent? Yes.

(b) Find $E(X)$ and $\text{var}(X)$. Well, $E(X) = \int_0^1 x\,dx = 1/2$ and $E(X^2) = \int_0^1 x^2\,dx = 1/3$, so $\text{var}(X) = 1/3 - (1/2)^2 = 1/12$. But the SE is $\sqrt{1/12} \doteq 0.29$. Compared to $\text{var}(X)$, the SE gives a better idea of how far X is likely to stray from the expected value.

(c) Given that $Y < X$, what is the conditional distribution of (X, Y)? Let $\mathcal{T}$ be the triangular region in the unit square below the main diagonal (see diagram). Conditionally, (X, Y) is uniform in $\mathcal{T}$.

Example 9. As in the previous example, let $\mathcal{T}$ be the triangular region in the unit square below the main diagonal. Let (X, Y) be a point chosen at random in $\mathcal{T}$.

(a) Is X uniformly distributed? If so, on what interval? No, X isn't uniform: the density of X at x is $2x$. (You are more likely to get a point at the right end of the triangle than the left end—the right end is bigger.)

(b) Given $X = x$, is Y uniformly distributed? If so, on what interval? Yes, Y is conditionally uniform given $X = x$, on the interval $[0, x]$.

(c) Are X and Y independent? No, see part (b).

(d) Find $E(Y|X = x)$. The answer is $x/2$.

What about general probabilities and random variables?

Discrete probability models are enough for many purposes. For other purposes, the mathematics gets very complicated very fast, especially for people who don't intend to be professional mathematicians. Few of the complexities are relevant to applications, but an example may be interesting nonetheless. Suppose we want to pick a number at random between 0 and 1. The sample space is the closed unit interval [0, 1]. If $0 \le a \le b \le 1$, the chance of picking a number in $[a, b]$ should be $b - a$. However, it rarely makes sense to define probabilities for *all* subsets of the unit interval. Instead, attention may be restricted to the "Borel σ-field" $\mathcal{B}$, i.e., the smallest collection of subsets of the unit interval that contains all the intervals $[a, b]$ with $0 \le a \le b \le 1$, and

(i) $\mathcal{B}$ contains the empty set $\emptyset$,

(ii) $\mathcal{B}$ is closed under complementation, i.e., if A is in $\mathcal{B}$ then the complement of A is also in $\mathcal{B}$,

(iii) $\mathcal{B}$ is closed under the formation of countable unions, i.e., if A_i is in $\mathcal{B}$ for $i = 1, 2, \ldots$, then the union $\bigcup_i A_i$ is also in $\mathcal{B}$.

Conditions (i)-(ii)-(iii) are what make $\mathcal{B}$ a σ-field. "Closed under complementation" has nothing to do with being a closed set, although closed sets are indeed in $\mathcal{B}$, as are open sets, and pretty much any other sets that might be of interest.

Henri Lebesgue showed there was a unique countably additive function λ defined on $\mathcal{B}$ with $\lambda\{[a, b]\} = b - a$ for all a, b with $0 \le a \le b \le 1$. His λ is now called "Lebesgue measure." "Countable additivity" is discussed again, below.

Further technical detail

Let $\mathcal{F}$ be a σ-field of subsets of a set Ω. In other words, $\Omega \in \mathcal{F}$; furthermore, $\mathcal{F}$ is closed under complementation and the formation of countable unions. A sequence $A_1, A_2, \ldots$ of sets in $\mathcal{F}$ is *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for all $i \ne j$: in other words, A_i and A_j cannot occur simultaneously. A real-valued function P on $\mathcal{F}$ is a probability provided (i) $P \ge 0$, (ii) P assigns measure 1 to the whole space Ω, and (iii) P is *countably additive*, i.e., $P\{\bigcup_i A_i\} = \sum_i P\{A_i\}$ for pairwise disjoint sets $A_1, A_2, \ldots$ in $\mathcal{F}$. This is how Kolmogorov defined probability. To restate (iii), the probability of a countable union of pairwise disjoint sets must be the sum of the individual probabilities. The ur-probability is Lebesgue measure on the unit interval.

In Kolmogorov's setup, a random variable X is a measurable function on Ω. That is to say, for each real number y,

$$\{X \le y\} = \{\omega : \omega \in \Omega \text{ and } X(\omega) \le y\} \in \mathcal{F}$$

In other words, you fix y. Then you think about the set of ω in Ω for which $X(\omega) \le y$. This set has to be in $\mathcal{F}$. Now the distribution of X makes sense: at least in principle, $P\{X \le y\}$ is computable. Without measurability, $P\{X \le y\}$ would be undefined, because $\{X \le y\}$ wouldn't be an element of $\mathcal{F}$.

If X is a random variable, its expected value is the "Lebesgue-Stieltjes" integral with respect to P:

$$E(X) = \int_\Omega X \, dP$$

Constructing the Lebesgue-Stieltjes integral is quite a project, even for a mathematician today; just the construction of Lebesgue measure on the unit interval might take some concentrated effort.

An infinite set is "countable" if it can be put into a 1–1 correspondence with the positive integers $\{1, 2, 3, \ldots\}$. Georg Cantor demonstrated that the unit interval is uncountable: $[0, 1]$ has the "cardinality of the continuum," which is larger than the cardinality of the positive integers. It takes some time to accept the idea that there are infinities of different sizes.

Why "absolutely continuous" not just "continuous"? As it turns out, there are random variables X such that (i) $P\{X = x\} = 0$ for all x, but (ii) X does not have a density. Condition (i) is *continuity*. This is a weaker condition than absolute continuity.

Conditional distributions and expectations can be defined in general, but that would take us too far afield. These few pages are no substitute for good undergraduate courses in real analysis and probability theory—but they may help you find your way through some of the thickets.

If you want to read more. . . .

R. M. Dudley (2002). *Real Analysis and Probability*. Cambridge, University Press. A textbook for advanced students.

W. Feller (1968, 1971). *An Introduction to Probability Theory and Its Applications*. Vol. 1, 3rd ed. Vol. 2, 2nd. ed. Wiley, New York. Although some details are out of focus and some examples need to be taken with a grain or two of salt, volume 1 is still *the* great undergraduate text. Volume 2 is a graduate text, and a very good one.

D. A. Freedman (1971). *Markov Chains*. Reprinted by Springer in 1983. Also see *Brownian Motion and Diffusion*. The Appendix has a detailed discussion of many topics mentioned here, especially, conditioning. Again, a graduate-level text.

D. A. Freedman, R. Pisani, and R. Purves (1998). *Statistics*. 3rd ed. Norton. Chapters 13-14-15 discuss probabilities. Chapters 16-17-18 are about random variables. An informal introduction to the ideas. Might be the place to start.

D. A. Freedman and P. B. Stark. "What is the probability of an earthquake?" In *Earthquake Science and Seismic Risk Reduction*. NATO Science Series IV: Earth and Environmental Sciences, vol. 32, Kluwer, Dordrecht, The Netherlands (2003) pp. 201–213. F. Mulargia and R.J. Geller, editors. On the interpretation of statistical models.

P. Humphreys (2005). "Probability theory and its models." Corcoran Philosophy Department, University of Virginia. On the interpretation of statistical models.

A. N. Kolmogorov and S.V. Fomin (1970). *Introductory Real Analysis*. Translated from the Russian by R.A. Silverman. Dover, New York. This is a lovely book, but very mathematical. Chapters 1, 7, and 8 are the most relevant.

J. C. Oxtoby (1980). *Measure and Category*. Springer, New York. Beautiful, very very mathematical. Chapters 1–5 are the most relevant.

W. Rudin (1976). *Principles of Mathematical Analysis*. 3rd ed. McGraw Hill, Undergraduate level, there are generally words to explain the equations, quite friendly as such books go. Chapters 1, 2, and 11 are the most relevant.

A. N. Shiryaev (2000). *Kolmogorov in Perspective*. Translated from the Russian by H. H. McFaden. *History of Mathematics*, vol. 20, American Mathematical Society, Providence, R.I. Review of Kolmogorov's work. For mathematicians.

More on observed values

The object here is to sketch the connection between data on the one hand and random variables in formal statistical models on the other. The concept of "observed value" is often used for this purpose: see, e.g., Lehmann (1983: 4) or Freedman (2005: 25, 42). However, the project is a little more difficult than it sounds, because the concept of an observed value is informal, i.e., extra-mathematical. In the end, we would suggest, such extra-mathematical elements are needed. That is what differentiates applied work from theory.

In essence, Kolmogorov (1933) formalized the ideas of probability theory in terms of a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, $\mathcal{F}$ is a σ-field of subsets of Ω, and P is a non-negative, countably additive function on $\mathcal{F}$ with $P(\Omega) = 1$. Crudely speaking, P distributes mass 1 over the various ω in Ω. In this setup, a random variable X is an $\mathcal{F}$-measurable function on Ω; and P assigns mass to any set $A \in \mathcal{F}$, including sets of the form $\{\omega : \omega \in \Omega \text{ and } X(\omega) \in B\}$.

What is missing here is the possibility of choosing a particularized ω at random from Ω, according to the mass distributed by P. If ω could be chosen, that would determine the observed value of $X(\omega)$ of X. Choosing ω, however, remains an informal activity; formalizing such choices does not seem to be on the cards. The issues may be more obvious when P is continuous, but remain even when P is discrete.

A *statistical model* replaces the probability P on $(\Omega, \mathcal{F})$ by an indexed family P_θ of probabilities, where the parameter θ runs through a parameter space Θ. The model also specifies a sequence of special ("observable") random variables $Y_1, Y_2, \ldots$. Typically, it is Θ and the joint distribution of $Y_1, Y_2, \ldots$ under P_θ that matter, rather than particular features of Ω, $\mathcal{F}$, P_θ, $Y_1, Y_2, \ldots$ themselves. If the model includes other random variables, as it often does, these may be irrelevant for present purposes; or, they may serve only to define the joint distribution that has just been mentioned.

From the present perspective, random variables, probabilities, expected values and the like are confined to statistical models. When we choose a model to apply, we assume that the data are like observed values of the random variables $Y_1, Y_2, \ldots$ specified in the model. This step is necessarily informal: it leads us away from pure mathematics, and into the empirical world. Moreover—although it is rarely articulated—this step is critical in applied work, because it brings the pretty assumptions behind the model (e.g., linearity, independence) into collision with ugly reality.

From the point of view outlined here, there is an important circuit-breaker between statistical theory and its applications. One application after another could prove to be a fiasco, without undermining the theory; of course, a reasonable bystander might conclude that the subject-matter domain to which the theory applies is narrower than had previously been thought. Conversely, the mathematical rigor of the theory cannot validate any particular application, the issue being the extent to which the assumptions behind the models fit the circumstances of interest.

For examples and more discussion, see Freedman (2005). "Observed value" was already used early in the twentieth century, referring to something in the data; see, e.g., Pearson (1916: 248) or Pearson (1925: 205), where the observed value is distinguished from what we would now call the random variable. See also Wald (1940: 286), Wald and Wolfowitz (1950: 82). "Realized value" was also in use by mid-century, as a synonym; see, e.g., Welch (1958: 784).

References

K. Pearson (1916). On the Application of "Goodness of Fit" Tables to Test Regression Curves and Theoretical Curves Used to Describe Observational or Experimental Data. *Biometrika* 11: 239–61.

K. Pearson (1925). James Bernoulli's Theorem. *Biometrika* 17: 201–10.

A.N. Kolmogorov (1933). *Foundations of the Theory of Probability*. Original in German. English translation reprinted by Chelsea, New York (1956).

A. Wald (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. *The Annals of Mathematical Statistics* 11: 284–300.

A. Wald and J. Wolfowitz (1950). Bayes Solutions of Sequential Decision Problems. *The Annals of Mathematical Statistics* 21: 82–99.

B.L. Welch (1958). "Student" and Small Sample Theory. *Journal of the American Statistical Association* 53: 777–88.

E.L. Lehmann (1983). *Theory of Point Estimation*. Wiley. Reissued by Wadsworth & Brooks/Cole (1991). 2nd ed. with G. Casella, Springer (1998).

D.A. Freedman (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Note. The page on observed values summarizes joint work with Paul Humphreys. Russ Lyons and Don Ylvisaker made helpful comments.