All the Probability you need for STA220

Introduction The theory of Statistics is based firmly in Probability. In a year-long sequence like STA257F, STA261S, the principles of probability are throughly developed, and the transition from Probability to Statistics is so smooth that students rarely notice when it happens.

In an elementary course like ours, the usual practice is to try to do the same thing at a more modest level, but it just does not work. All is well at first, with clear definitions and principles that build upon one another. Students learn to solve some moderately challenging problems; the professor is happy (assuming that he or she retains the capacity for happiness) and some of the students are happy too.

Unfortunately, many of these moderately challenging problems have to do with cards and dice and tossing coins. This is because historically, the area of Mathematics called "Probability" arose from attempts to model games of chance, and so that is what the simplest examples are about. Most students (except for those who are interested in gambling or just like puzzles of any kind) find the material quite dreary, and completely inapplicable to their chosen fields of study.

Building the bridge between elementary Probability and Statistics takes time, and many of the key concepts depend on Calculus (something like MAT132Y or MAT138Y). So our text, like most, makes a quick leap. They skip the hard part, and the material that follows makes use of only a few very simple probability principles.

The experience of many students is that they are going along studying cards and dice, and then suddenly they are studying something unrelated. The material of elementary probability sits there like an undigested lump, with no obvious relationship to either descriptive or inferential Statistics. In this course, our solution is not to swallow that coconut in the first place. This document will tell you all you really need to know, and you can skip Chapters Three and Four. This way, we can cover the most important material at a more moderate pace, and learn some things that usually come in STA221S.

Probability A statistical *experiment* is any procedure whose outcome is not known in advance with certainty. A *sample point* is the most basic outcome of an experiment, and the *sample space* is the set of sample points.

An example, and the *only* game of chance that will be considered here, is that you have a large jar filled with numbered marbles, and the marbles are thoroughly mixed up. You select a marble from the jar without looking.

The marbles correspond to units in a population; there are N marbles in the jar, where N is the population size. When you select marble number 1,347, you then locate population unit 1,347 and conduct some observations or do an interview or something. Each marble is a sample point, and the sample space is the entire set of marbles.

An *event* corresponds to a subset of the sample space – that is, a subset of the marbles in the jar. For example, the marbles could correspond to the set of students enrolled in our class, so that selecting a marble corresponds to selecting a student. Some possible events are:

- A female student is chosen
- A student with two or more jobs is chosen
- A student with resting pulse rate over 70 beats per minute is chosen
- A student weighing over 200 kg is chosen

You can see that the definition of an "event" corresponds to a way of splitting up the marbles (population units) into categories. By the way, these "categories" can correspond to quantitative measurements. For example, there could be one category for each dollar amount of annual earnings, and "A student making $18,297 is chosen" would be one event. Events may or may not overlap. If a female student holding three jobs is chosen, we would say that *both* the event "Female" and the event "Two or more jobs" occurred.

Now imagine conducting a statistical experiment a large number of times in exactly the same way. In our main example, you would draw a marble, look at it, put it back in the jar, mix up the marbles again and draw another, over and over. You could easily make a frequency distribution or histogram to describe the results of your repeated draws from the jar.

Again, we conduct a statistical experiment in the same way, a large number of times. The *probability* of an event is defined as its long-run relative frequency. This is not the most mathematically elegant definition of probability, but it is the clearest.

In our simple example with the marbles in the jar, the long-run relative frequency of an event is identical to the relative frequency of the event in the population. For example,

- If the relative frequency of females in our class is 0.55 (55%), and we randomly selected students over and over again with replacement, the relative frequency of the event "Female" will approach 0.55 as the number of draws increases, and we say that the probability of choosing a female is 0.55.
- If 22% of the class have resting pulse rates over 90 beats per minute, and we randomly selected students over and over again with replacement, the relative frequency of the event "Over 90 beats per minute" will approach 0.22, and we say that the probability of choosing a student with a resting pulse rate over 90 is 0.22.
- If 47.2% of students in the class usually take public transportation to school, and we randomly selected students over and over again with replacement, the relative frequency of the event "Usually takes public transportation to school" will approach 0.472, and we say that the probability of choosing a student who usually takes public transportation to school is 0.472.

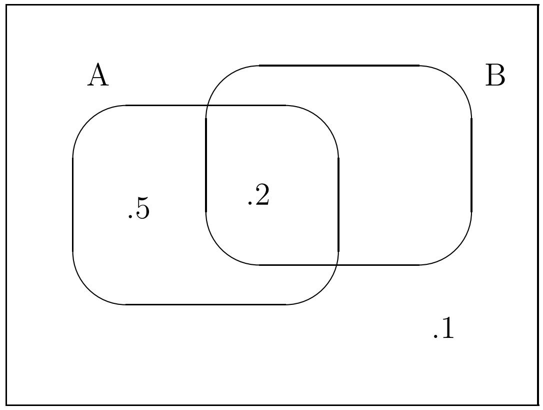
It should be clear that

1. Because they are relative frequencies, probabilities must be between zero and one inclusive, and
2. When a population is divided into a set of non-overlapping events that include every member of the population, the probabilities of those events must add up to one.

These principles apply to any statistical experiment, not just our primary example of selecting a marble from a jar. They are all you need to solve some problems that look fairly challenging on the surface. For example,

> Consider an arbitrary sample space and two events A and B. The probability of A is 0.5, the probability of both A and B is 0.20 and the probability of neither A nor B is 0.10. What is the probability of B?

To solve such problems, it helps to draw a *Venn Diagram*, in which the sample space is represented by a rectangle, and the events by overlapping circles or ovals. Fill in what you know, and solve for the missing information by making the separate pieces add up.



Now it's easy to see that B-not-A has probability 0.2 (so the separate pieces add up to one), and then the probability of B is $0.2 + 0.2 = 0.4$. By the way, it is common to write the probability of an event as $P(\text{Event})$, so the answer to the question could be written $P(B) = 0.4$.

Getting back to our prime example of sampling a marble from a jar, let each of the N marbles in the jar represent a distinct event. Because the physical nature of the selection procedure does not give any marble an advantage over the others, we know that the probability of selecting each marble must be the same. All the probabilities must add to one, so the probability of selecting each marble must equal $\frac{1}{N}$.

Now one can see that the probability of any event is just the number of marbles in the event divided by N. And for any probability problem, we can get the probability of an event by adding the probabilities of the sample points that make it up.

Random Variables Recall how the jar of marbles can be subdivided into events, according to the value taken on by some variable like age or income. The statistical experiment now consists of selecting a marble at random as usual, and then ascertaining the value of the variable for the population unit (person or whatever) corresponding to the marble. The resulting value of the variable is not known in advance, and is subject to the laws of chance because of the random selection of the marble. For this reason, it is called a *random variable*. Obviously, you can make probability statements about the observed value of a random variable. For example,

> A jar has 10 numbered balls. There are 4 ones, 3 twos, 2 threes and 1 four. What is the probability of selecting a ball with a number between 1.5 and 3.5?

The answer is (4+1)/10 = 1/2.

All you need to answer probability questions about a random variable is a relative frequency histogram. Just add up relative frequencies (heights of the bars) to get the probability of any event.

Equivalently, you can do it with symbols. We will follow the book's convention and use x to represent the numerical value of a random variable. The notation $p(x)$ represents the probability that the random variable will equal the value x. It is often helpful to arrange the values of x and $p(x)$ in a tabular form to represent the *probability distribution* of the random variable. The distribution from the problem above would look like this:

x	1	2	3	4
$p(x)$	0.4	0.3	0.2	0.1

Because $p(x)$ values are probabilities, they must satisfy two rules:

$$0 \leq p(x) \leq 1 \text{ and } \sum p(x) = 1,$$

where we are summing over all value of x. To see whether a table of $p(x)$ could really be a probability distribution, these are the properties you need to check.

Continuous Random Variables Our text (like almost all others) makes a distinction between "discrete" and "continuous" random variables. Discrete random variables are random variables that can assume only a countable number of values. All the random variables we have been considering are discrete, because all populations that actually exist are finite. The number of values a random variable can assume cannot exceed the number of population units, and so they are always countable. The $p(x)$ notation refers specifically to discrete random variables, not continuous ones.

But imagine a variable measured on a scale, like length, that is theoretically continuous. No matter how accurately you measure length, finer accuracy is always possible in principle. Now imagine measuring such a continuous variable with greater and greater accuracy,for an *extremely* large population. Make a sequence of relative frequency histograms, with finer and finer subdivisions along the x axis. After a while you will no longer be able to see the blockiness of the individual bars of the histogram. It will look like a smooth curve. If all the relative frequency histograms are drawn so that the area of the bars (with time height) adds to one, then the area between the smooth curve and the x-axis will also equal one.

So, with an infinite population (just pretend, please) and finer and finer measurement, the random variable approaches a theoretical abstraction called a *continuous* random variable. The smooth curve is called its *probability density function*, denoted by $f(x)$, and probabilities correspond to *areas* between the smooth curve and regions of the x-axis.

Remember how the mean is the physical balance point of a discrete distribution? Well, make a thin sheet of metal shaped like the limiting relative frequency histogram; the top is the smooth curve and the bottom is the x axis. The physical balance point of the sheet (a point on the x axis, the "first moment" of Physics) is called the "Expected Value" of the continuous random variable. It is symbolized by μ, and is a model of the population mean. Continuous random variables can have a variance too; σ^2 is a quantity proportional to the amount of energy required to spin the sheet to a given velocity around its balance point (the "second moment" of Physics).

Now you are ready for Chapter 5. First, read the discussion of discrete and continuous random variables on pages 174-176 in Chapter 4. Then read Chapter 5's Introduction and Section 5.1. Skip Section 5.2, but pay a lot of attention to Section 5.3. Skip the rest.