STATISTICS TOPIC NOTES

Uncertainty Distributions and Random Variables

Introduction to Statistics

The Big Picture:
Statistics is about managing uncertainty.

If you have ever thought about statistics before, you probably have not thought about it in terms of managing uncertainty. But that is the way that business people actually use statistics, whether they are simply summarizing some data or forecasting next year's sales. I will not attempt to convince you now that statistics really is about managing uncertainty. I will let the reality emerge as the course develops. But I will tell you now that I will point out repeatedly throughout the course how we are managing uncertainty. And you will see repeatedly that we are.

Uncertainty is the most fundamental concept in statistics. In this course:

- You will learn how we **model** uncertainty.
- You will learn how we **measure** uncertainty.
- You will learn how we **find** hidden knowledge in uncertainty.

But before we begin, I want to explain how I will develop and present the concepts in this course. I want you to understand the basic concepts and principles clearly and intuitively. It is much more important to understand statistics intuitively than to memorize formulas. If you know the intuition, the formulas will make sense. If you don't know the intuition, the formulas will not make sense. Besides, with the steady increase in computer power, we have put most of the formulas out of sight inside computer software. So you don't *have* to memorize formulas.

Experience shows that concepts and principles are most readily mastered in a two-step process:

1. First, begin with one or more very simple examples that have been stripped of complicated real-world context so that the concept or principle can be seen in pure form. These simple examples may not be very practical. But once the concept/principle has been grasped, then it will be much easier to apply to the real-world problems that we are actually interested in.
2. The second step is to present one or more real-world examples with more complicated contexts in which the basic concept/principle can now be recognized. Developing competence at statistics requires developing the ability to strip away the interesting but distracting real-world aspects of a problem to expose the underlying statistical principles. If you try to learn the concepts in complicated contexts, you can get confused by the interesting real-world detail and fail to recognize the underlying principles.

Let us begin.

The Concepts of Uncertainty Distribution and Random Variable

Definition. A **distribution** is a set of numbers that represent all of the *possible outcomes* of an uncertain phenomenon together with the *probabilities* with which those outcomes occur.

There are several important aspects to this definition of distribution that are worth further emphasis:

1. There is an uncertain phenomenon that can produce different outcomes.
2. The outcomes are numbers.
3. The uncertainty of each outcome is measured by the probability of that outcome.
4. The distribution includes ALL possible outcomes.
5. The concept of distribution includes BOTH the set of all outcomes and the corresponding probabilities.

Example 1. Toss a fair coin. Record a 1 if a head occurs. Record a 0 if a tail occurs.

Tossing a fair coin is the simplest possible uncertain phenomenon. It is uncertain because we do not know whether a head will occur, or a tail, until we have tossed the coin. Although we see either a head or a tail, we record a number (0 or 1) to represent the outcome. All possible outcomes are the set of numbers {0, 1}. The corresponding probabilities for a fair coin are {½, ½}. So the distribution for a fair coin toss can be represented by a table:

Outcome	Probability
0	½
1	½

Note that in the preceding five sentences, I have just run through the list of five points above that I wanted to emphasize further:

1. The coin toss is uncertain because we do not know which of head or tail will occur.
2. The outcomes are the numbers 0 (tail) and 1 (head).
3. The probability of the outcome 0 is ½, and the probability of the outcome 1 is ½.
4. The distribution of the coin toss includes both the 0 outcome and the 1 outcome.
5. In addition, the distribution includes the probability of ½ for each outcome.

Note also that probabilities are nonnegative and add up to 1 over all of the possible outcomes.

All you need to do to specify a distribution is to list all possible outcomes together with the corresponding probabilities.

Very important idea: There is uncertainty about the outcome of the coin toss. This uncertainty is represented by a distribution. It is critical to understand and to remember that ANYTIME there is uncertainty about the outcome of a numerical phenomenon – whether it be something as simple as tossing a coin or more complicated, like forecasting sales for next year – the uncertainty can be represented by a distribution consisting of outcomes and probabilities. We will see *much* more of this throughout the course.

Since a *distribution* represents all possible outcomes and corresponding probabilities for an uncertain phenomenon, I can call a distribution a **distribution of uncertainty** or an **uncertainty**

distribution. I will use the term **uncertainty distribution** to mean the same thing as distribution. It represents how the uncertainty is distributed over the possible outcomes.

How to interpret probabilities: There are two equivalent ways to think about the probabilities of outcomes. In the coin toss example, they are:

1. *One-time relative frequency*. There are probably millions of slightly different ways that we can toss a coin. We can vary the speed of the spin and height of the toss – even the wind speed and temperature. About half of those ways would turn out to be heads (1) and half would turn out to be tails (0). Then whether we get a head or a tail depends upon which way – which combination of spin, height of toss, wind speed, etc. – we happen to pick. Moreover, the "head" ways are thoroughly mixed up with the "tail" ways, so that it is very hard to reliably target a group of "head" ways versus a group of "tail" ways. So we pick blindly. If by chance we pick one of the "head" ways, we will get a 1; if by chance we pick one of the "tail" ways, we will get a 0. Since the number of head ways is about the same as the number of tail ways, the odds are about even for getting a head or a tail.
2. *Long-run relative frequency*. If we repeatedly toss the coin a large number of times, then the proportion of heads will get closer and closer to the probability of heads, namely ½. Likewise the proportion of tails will get closer and closer to the probability of tails, namely ½. This is because about half of the ways we blindly pick for our tosses will be "head" ways and about half will be "tail" ways. Repeated tossing provides a means to verify empirically what the probability of heads really is.

The concept of **random variable** is very closely related to the concept of distribution. For many purposes it is not important to distinguish them. So we may think of *random variable* as a rough synonym for *distribution*. Like a distribution, a random variable also has outcomes and probabilities. The concept of random variable describes what happens when we make a random draw from the outcome set of a distribution, where the probability of drawing an outcome is governed by the probabilities of the distribution. In Example 1, when we toss a coin, we make a random draw from the outcome set and get a 0 or a 1. So a random variable has an associated distribution. Thus, Example 1 above, which describes the distribution for tossing a fair coin, also describes the random variable for tossing a fair coin. For our purposes, the most important thing about a random variable will be its distribution. So we can think of random variables as we think of distributions.

So why have two concepts and distinguish random variable from distribution …?

The two concepts are actually subtly different. But the distinction is useful. In essence, "distribution" is a generic concept and a particular "random variable" is a specific instance of a distribution. "Distribution" is like the concept "dog", whereas my Bowser and your Spot are like "random variables" in being particular instances of "dog". As a consequence, two different random variables can have the same distribution, in the same way that two different animals (Bowser and Spot) can both be dogs. For example, we can toss the coin twice. We can draw more than once from the same distribution, and each toss or draw describes a different random variable with the same distribution. Different draws from the same distribution share important statistical features by being replicates of the same distribution, in the same way that different

dogs share important anatomical and behavioral features by virtue of being dogs. A simple example should clarify this point:

Example 2.

- Toss a fair coin one time. What is the distribution for the first toss? Clearly the distribution is given by Example 1, since the possible outcomes are {0, 1}, and the probabilities are {½, ½}.
- Toss the same fair coin a second time. What is the distribution for the second toss? Once again, it is clear that the distribution is given by Example 1, for the possible outcomes are again {0, 1}, and the probabilities are again {½, ½}. So Example 1 is "dog", and the first toss is "Spot", and the second toss is "Bowser".
- So the distribution for toss #1 and the distribution for toss #2 are the same because the set of possible outcomes and the corresponding probabilities are the same. But the actual outcomes of toss #1 and toss #2 need not be the same. For example, #1 could turn out to be a head (1) and #2 could turn out to be a tail (0). This point is important: The *potential* outcomes are the same, but the *actual* outcomes may be different.

It is common to denote random variables by capital letters. So in Example 2, we could let X_1 denote the outcome for toss #1 and X_2 denote the outcome for toss #2. Then X_1 and X_2 are two separate random variables that have the same distribution, but it is not necessarily true that $X_1 = X_2$.

Example 3.

- Toss fair coin #1 and let *X* denote the outcome.
- Toss fair coin #2 and let *Y* denote the outcome.
- Is it clear to you that Example 3 is essentially the same as Example 2? The only difference is that we toss the same coin twice in Example 2, whereas we toss two different coins in Example 3. But all of the coins have the same properties. *X* and *Y* are random variables that have the same distribution, but it is not necessarily true that *X* = *Y*. The coins are different, but the outcomes and probabilities are the same. So the random variables have the same distributions – although their actual values need not be the same.

Example 4.

- Toss fair coin #1 one time and let *X* denote the outcome.
  The distribution for *X* is

Outcome	Probability
0	½
1	½

- Toss <u>*un*</u>fair coin #2 one time and let *Y* denote the outcome. Suppose that coin #2 is biased in favor of heads according the following distribution:

Outcome	Probability
0	¼
1	¾

Although the outcome set for coin #2 is the same as the outcome set for coin #1 (both may be 0 or 1), the distribution for coin #1 and the distribution for coin #2 are not the same. For example,

the probability for outcome 0 is not the same for *X* as for *Y*, and the probabilities for outcome 1 are different, as well. In this example, *X* and *Y* are different random variables and they have different distributions. To have the same distribution, both the outcome set and the corresponding probabilities must be the same. If *X* had probability ¼ for outcome 0 and probability ¾ for outcome 1, then *X* and *Y* would have the same distribution.

How to interpret the unequal probabilities for unfair coin #2? Again, there are two equivalent ways to think about the probabilities:

1. *One-time relative frequency*. There are probably millions of slightly different ways that we can toss coin #2. About 3/4 of those ways would turn out to be heads (1) and 1/4 would turn out to be tails (0). Then whether we get a head or a tail depends upon which way – which combination of spin, height of toss, point of catch, etc. – we happen to pick. If by chance we pick one of the "head" ways, we will get a 1; if by chance we pick one of the tail-ways, we will get a 0. Since the number of "head" ways outnumbers the number of "tail" ways by 3 to 1, the odds favor heads by 3 to 1. Physically, we can make a biased coin by making the tail side from a heavy metal, like lead or uranium, so that the bottom-heavy tail side is more likely to end down, and the lighter head side more likely to end up.
2. *Long-run relative frequency*. If we repeatedly toss coin #2 a large number of times, then the proportion of heads will get closer and closer to the probability of heads, namely 3/4. Likewise the proportion of tails will get closer and closer to the probability of tails, namely 1/4.

In summary, a random variable can be thought of as making a random draw from the outcome set of a distribution, where the probability of selecting any particular outcome is given by the probabilities of the distribution. The distinction between random variable and distribution is most useful when making repeated draws from the same population (same outcome set). Each draw represents a different random variable based on the same distribution. Repeatedly drawing from the same population is the foundation of sampling, which is very important in statistics because it helps us learn by induction about uncertain phenomena.

The probability function for a random variable *X* gives the probability for each specific outcome *x*. The probability function may be written $pr(X = x)$, in which the lower case *x* refers to a specific outcome.

So the distribution for a toss of a fair coin (Example 1) may be written as

x	$pr(X = x)$
0	½
1	½

The "*X*" refers to the random variable. The "*x*" refers to either the 0 or the 1 that are the potential outcomes.

Example 5. Throw a fair die once and let X denote the number of spots on the side facing up. The outcome set is {1, 2, 3, 4, 5, 6}. Since the die is putatively fair, each side has the same probability of 1/6 of occurring. So the distribution of X may be represented in table form as:

X	1	2	3	4	5	6
$pr(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6

The table is rotated sideways to save space. It should be clear from the preceding discussion that the roll of a fair die represents a distribution/random variable.

Now I will start to increase the complexity of the examples. But with mastery of the preceding elementary examples, you should be able to discern the concepts of uncertainty, distribution, and random variable in the following more complex examples. Don't worry, I will guide you.

Example 6. Throw a fair die twice and let X_1 be the number of spots on the up side on throw 1 and let X_2 be the number of spots on the up side on throw 2. Then X_1 and X_2 each have the distribution shown in Example 5. But now let $Y = X_1 + X_2$. So Y is the total number of spots in two throws. Is Y a random variable? If so, what is its distribution?

Clearly, Y is an uncertain numerical phenomenon. Reflection shows that Y could be any integer from 2 to 12, inclusively. Then the possible outcome set is {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. There are no other possibilities. But what are the probabilities of the outcomes? Reflection shows that there are 36 (= 6 × 6) equally likely combinations of the 6 outcomes on throw 1 with the 6 outcomes on throw 2. Furthermore, only 1 of these 36 combinations produces $Y = 2$ – namely, $X_1 = 1$ and $X_2 = 1$. So $pr(Y = 2) = 1/36$. Furthermore, only 2 of the 36 combinations produce $Y = 3$ – namely, $X_1 = 1$ and $X_2 = 2$, or $X_1 = 2$ and $X_2 = 1$. So $pr(Y = 3) = 2/36$. Continuing in this manner, you can derive the following distribution for Y:

y	2	3	4	5	6	7	8	9	10	11	12
$pr(Y=y)$	1/36	2/36	3/36/	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Example 7. Ten students were randomly selected from a fulltime UT MBA class. Each student was asked to provide anonymously, to the nearest $1000, the annual salary that he/she anticipates to make in his/her first job after graduation (including bonus). The ten students reported the following anticipated salaries (in $1,000s): 110, 160, 120, 95, 175, 80, 120, 130, 120, 110. Suppose one student is picked at random from the ten. Let X denote his/her anticipated salary. Is X a random variable? If so, what is the distribution of X?

Clearly, there is uncertainty about the value of X. To get the distribution for X, we must specify the possible values for X and their probabilities. Since each student occurs once and each student is equally likely to be chosen, the probability of picking any one *student* is 1/10. However, some students anticipate to make the same amount – so some of the possible *values* of X occur more than once. $X = 110$ occurs two times, and $X = 120$ occurs three times. Although each *student* is equally likely, each *value* is not. Although there are ten distinct students, there are only seven distinct values in the distribution of X. So the distribution of X is:

X	80	95	110	120	130	160	175
$pr(X=x)$	0.1	0.1	0.2	0.3	0.1	0.1	0.1

Example 8. In response to a class survey, each of the 65 students in a fulltime UT MBA class writes down the annual salary that he/she anticipates to make in his/her first job after graduation (including bonus). One of the 65 students is then selected at random and the number that he/she wrote down is revealed. Let X denote the revealed number. Is X a random variable? If so, what is the distribution of X?

Clearly, there is uncertainty about the value of X. Before the number is revealed, we do not know what it will be. So X is a random variable. So X must have an outcome set and corresponding probabilities.

What is the outcome set? The revealed number could be any of the 65 that were written down. We do not know what any of the 65 is. We do not even know how many distinct values there are. Certainly, there are no more than 65 distinct values. But some students may have unknowingly selected the same value.

OK, so what about their probabilities? We do know that the probabilities are 1/65 for each *student's* response. But some *values* might have probabilities that are multiples of 1/65, depending upon how many students chose the same value.

So what do these ruminations mean about the distribution for X? Can we still say that X has a distribution and is a random variable? Yes! In principle, an umpire could collect the 65 values, write down the distinct values and tally the number of times that each value occurs. That would provide the distribution, even though we do not know what it is!

An important follow-up: In Example 8, you encountered your first example of a distribution that exists, although you do not know what it is. You do not know its possible values, nor do you know its probabilities. In real-world applications of statistics, this is usually the case. Most of the time in real-world statistics, we do not have complete knowledge of the distribution of interest. For example, we may be trying to estimate sales for next year. Our uncertainty about sales is represented *in principle* by a distribution – a list of possible values for sales and their probabilities. But we may not know anything about the possible values for sales other than that they will be positive numbers. And we may have no idea what the probabilities are for those values. *The whole objective of our statistical endeavor may be to infer some things about that uncertainty distribution* – such as, what is the most likely value for sales? And, how much is the probability that sales will be between \$19 million and \$20 million? It is astonishing that we can actually do this without knowing much about the uncertainty distribution of sales.[1] But we need to be able to posit the existence *in principle* of a well-defined distribution. This means that *if we were all-knowing*, we could make a list of all possible outcomes and specify their uncertainties in a probability function, like the hypothetical umpire in Example 8. We do not actually need to be able to make the list in the real world.

By the way, go back to Example 7 for a moment. What if I had not told you that the ten students had picked 110, 160, 120, 95, 175, 80, 120, 130, 120, 110? What if I had concealed this information from you? Would my concealment have changed the distribution of X? The ten numbers are still there. Selecting one of the students would still select one of these ten numbers. Would the outcomes for X be any different just because you do not know what they are? Would the probabilities be any different? Do you really need to know what the outcomes and their probabilities are for the distribution to *exist*?

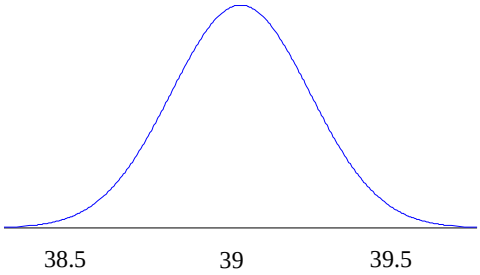
[1] Stay tuned for the rest of the course to find out how we do this.

In all of the examples that I have shown you so far, the distribution and the related random variables are called **discrete** because the outcome set consists of discrete, or isolated, numbers. However, most of the distributions and random variables that we will deal with in this course will be of a different type – at least approximately. We will deal mostly with **continuous** distributions and continuous random variables – especially the special type called **normal** distributions. The outcome set for a continuous distribution is a continuous range of numbers. Correspondingly, the probability function for a continuous random variable is a continuous function. You get probabilities for continuous distributions and continuous random variables indirectly – from the area under the graph of the probability function. (I will explain later how to do this.) The subsequent examples in this Topic Note will be examples of continuous distributions and continuous random variables.

Example 9. The weight of a can of Folger's coffee.

Suppose we pick a can of Folger's coffee at random from the section of the grocery store shelf where 39-ounce cans of Folger's coffee are displayed. The can says "Net weight 39 ounces." But is the net weight really 39 ounces? Reflection suggests that the process of filling the can is imperfect and is unlikely to put exactly 39 ounces of coffee into the can. So we are uncertain about the true weight of the coffee in the can. Therefore, there is a distribution that describes our uncertainty. The actual weight of coffee in the can is a random variable X that has that distribution.

Now weight is a continuous variable with an infinite number of possible values. So we cannot list all possible values and their probabilities. However, we can display the distribution graphically, in principle. One possible distribution is shown below.[2] The possible outcomes are represented by the horizontal axis and the probabilities by the curve.



But what does the picture mean? It means that there are a lot of cans that weigh around 39 ounces. There are not many cans that weigh less than 38.5 ounces. There are also not many cans that weigh more than 39.5 ounces. We can think of making the curve by stacking up the cans at their weights. There are a lot of cans at or around 39 ounces, so the stack is high there. There are few cans around 38.5 ounces or 39.5 ounces, so the stacks are short there.

Furthermore, we will scale the curve so that it represents the percentage or proportion of cans at each weight, instead of the number of cans at each weight. In that way, the curve can apply equally when the totality of cans number in the thousands as when the cans number in the millions. Probabilities concern percentages or proportions, rather than counts.

Because most of the cans weigh around 39 ounces, when we select a can at random, we are likely to get a can that weighs around 39 ounces. We are unlikely to get a can that weighs

[2] This is not the only possible distribution. In the real world, we will probably not know enough about the actual distribution to draw a specific curve like this one. But an actual distribution will exist to an all-knowing umpire!

less than 38.5 ounces or more than 39.5 ounces. The *cans* may be selected with equal probabilities, but the *weights* are not. This is because more cans weigh close to 39 ounces than to 38.5 or to 39.5 ounces. This discussion parallels Examples 7 and 8, in which students were selected with equal probabilities, but their anticipated incomes were not – because some incomes are popular with students – similarly, some weights are popular with cans.

The graph indicates that the most likely value for the net weight is 39 ounces, since the curve reaches its highest value at 39. The curve drops rapidly as the possible value for net weight moves away from 39 in either direction. The curve suggests that there is only a tiny probability that the net weight could be less than 38.5 ounces or greater than 39.5 ounces. Thus, if this is the actual distribution for net weight, we may be confident that the actual net weight of the can that we pick will be very close to the 39 ounces marked on the can. How to compute probabilities for continuous distributions will be explained in the section on the Normal Distribution later in these Topic Notes.

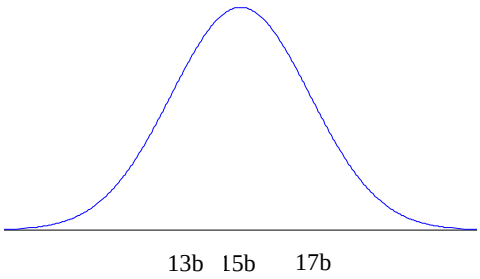
The important take away for Example 9 is that there is uncertainty about the net weight of a can of coffee and that this uncertainty is represented by a distribution of outcomes and probabilities.

Example 10. Sales next quarter for Dell Computer Company.

Let X denote sales for Dell Computer Company for next quarter. Is X a random variable? Are we uncertain what the value of sales will be for Dell next quarter? Of course, we are! So everything that I have taught you so far suggests that there must be a distribution that represents this uncertainty.

The outcomes for that distribution will be all possible values for sales next quarter. Some of those outcomes are more likely than others, so the probabilities for the distribution will reflect the greater likelihood of some values and lesser likelihood of others. A vast number of forces affect sales, including the general health of the economy, Dell's advertising, interest rates, materials availability, consumer demand, business capital spending plans, etc. All of these forces make some outcomes for sales more likely than other outcomes.

Do we know what the distribution is? Do we know the possible outcomes and their probabilities? No. But the graph below shows one possible distribution. The possible values (outcomes) for sales are shown on the horizontal axis and the curve represents corresponding probabilities. We will learn how to compute probabilities for continuous distributions in the section on the Normal Distribution in a later Topic Note.[3]



[3] Is the distribution of sales really continuous, as shown in the graph? You can argue that the possible values for sales are in fact discrete, since sales must be a whole number multiple of the smallest unit of a currency (like the penny). However, a penny is really small compared to the $ billions that sales could be. It is often a useful simplification to approximate a discrete distribution with a close continuous distribution, as I have done in this example. No harm is done thereby.

The graph suggests that the most likely outcome for sales is $15 billion. But there is considerable probability that sales could be less than $13 billion or greater than $17 billion. If this graph shows the actual distribution of sales next quarter, then there is considerable uncertainty over the actual value of sales next quarter.

As will be discussed later in the term, this distribution could be used to make a prediction of Dell's sales next quarter. It will probably not surprise you to learn that the best prediction is the value where the probability curve is highest. Just as importantly, the distribution can give a meaningful measure of how accurate the prediction is likely to be. Intuitively, the more the probability is concentrated around the prediction, the more certain and more accurate the prediction will be. Conversely, the more the probability is spread out far from the prediction, the less certain and less accurate the prediction will be. In the Topic Note (Mean and Standard Deviation), we will learn how to measure how spread out a distribution is.

As with the preceding examples, the main take away for Example 10 is that there is uncertainty about the value of Dell's sales for next quarter and that this uncertainty is represented by a distribution.

SUMMARY

- **Uncertainty** is the most fundamental concept in statistics.
- Whenever there is some numerical phenomenon about which we are uncertain, there is a **distribution** of uncertainty.
- An uncertainty distribution is determined by its set of **outcomes** and their corresponding **probabilities**.
- A **random variable** is a random draw from a particular uncertainty distribution. Like the uncertainty distribution itself, a random variable has a set of outcomes and associated probabilities.
- If the uncertainty distribution/random variable has a finite number of outcomes, then it is **discrete**. Its distribution can be written down as a table, with a list of all of the outcomes and a list of all of the corresponding probabilities.
- If the uncertainty distribution/random variable has a continuous range of values for its outcomes, then it is **continuous**. Its distribution can be displayed as a graph, with the outcomes on a horizontal line and the probabilities displayed as a curve above the line.

A number of examples illustrated these concepts, from very simple to increasingly complex.