Topic 11: Probability: Part 2

Recap of material so far:

- **An experiment is a task, a process that moves from an initial condition to a final condition but where we cannot predict exactly what that final condition will be.**
- **A trial is one instance of performing an experiment.**
- **A sample space is the collection of all possible outcomes of a performing a trial for an experiment.**

Let us assume that there is a non-negative number assigned to each outcome in a spacific sample space and that the sum of all of these numbers is 1. Then we call that number the probability of getting that outcome. Here is a small sample space and an assigned probability for each outcome.

pet	cat	dog	rabbit	canary	hamster	parrot	boa
probability as a fraction	23/85	31/85	7/85	11/85	6/85	4/85	3/85
probability as a decimal	0.270588	0.364706	0.082353	0.129412	0.070588	0.047059	0.035294

This is a sample space so these are the only pets under consideration. The probabilities can be given as fractions or as their decimal equivalents. The decimal versions are approximations since we cannot show infinitely many digits. The fractional probabilities add to 1. The decimal probabilities add to 1 but, as displayed, with rounding errors the displayed values may add to be really close to 1.

An event is achieving one or more of the outcomes. Thus, an event, X, might be to select "dog" from the sample space. The probability of this is 31/85. We write this in symbols as

P(X=dog) = 31/85 or, in shortened form, P(dog) = 31/85

These should be read as "The probability of an event X being equal to "dog" is 31/85."

Because the sum of the probabilities of all outcomes is 1 we know that the probability of not selecting "dog" must be 1 minus the probability of selecting "dog". That is

$$P(X \neq dog) = 1 - P(X = dog) = 1 - 31/85 = 85/85 - 31/85 = (85-31)/85 = 54/85$$

An event need not be so simple. For example, we could look at P(X is a mammal) and that would be P(X =cat or X=dog or X=rabbit or X=hamster) = 23/85 + 31/85 +7/85 + 6/85 = 67/85.

See web page on Probabilities for discussion of tree diagrams and the issue of sampling with or without replacement.

See web page for law of large numbers and getting approximations for probabilities.

Here is a contingency table for the frequency of items that have both a color and a design on them. We see that there are 122 Yellow Dot items, a total of 221 Green items, a total of 529 triangle items, and a grand total of 2,941 items. We can use this to answer some probability questions.

	Red	Blue	Yellow	Green	Purple	column total
Star	143	246	321	92	211	1013
Dot	120	124	122	11	68	445
Crescent	143	279	266	76	190	954
Triangle	59	163	158	42	107	529
row total	465	812	867	221	576	2941

If we randomly choose an item from our population, what is the probability that the item is:

1) a blue triangle? ______________

2) a purple star? ______________

3) a green item? ______________

4) a dot item? ______________

5) not a purple item? ____________

6) not a crescent item? ___________

7) either yellow or green? _________

8) either blue or dot? _____________

9) is neither green nor crescent? ___________

If we randomly choose an item from our population and I tell you that

10) the item is green, then what is the probability the item is a dot? ___________

11) the item is a triangle, then what is the probability the item is blue? ____________

Questions 7 and 8 illustrate the use of "or" in our probability. We have a general formula for expanding an "or" condition whose values can be read from the table. If we have condition A and condition B then

$$P(\text{ A or B }) = P(\text{ A }) + P(\text{ B }) - P(\text{ A and B })$$

Questions 10 and 11 illustrate the use of a conditional probability. We could restate problem 10 as **P(dot | green)** which is read as "the probability of getting dot given that we know the item is green. The vertical bar, |, is read as "given that." Problem 11 would be stated as **P(blue | triangle)**. We have a general formula for expanding a conditional probability to probabilities that are not conditional. If we have events A and B then

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

Questions 5 and 6 also have a general form. When we say we want something that is "not" a value then we are asking for the complement of a value. Recall that **P(X=A)** is often written in a shorter form as **P(A)**. We have two ways to write a shorter form for **P(X ≠A)**. They are **P(A')** and **P(A^c)**. Then we can write our rule for calculating a complement as

$$P(A') = P(A^c) = 1 - P(\text{ A })$$

Finally, we will look at sampling with and without replacement. From the table above we know that the probability of choosing an item that is a Yellow Triangle is 158/2941. If we choose an item it is either a Yellow Triangle or it is not. If we take the item that we chose and put it back with the other items and then draw a new random item then the probability that this new item is a Yellow Triangle remains at 158/2941. That is sampling with replacement. In that case the probability does not change.

However, we might not put the first selected item back into the population. That would be sampling without replacement. In that case, what is the probability that the second item randomly selected is a Yellow Triangle? That depends of what the first item was. If the first item was a Yellow triangle then there are only 157 Yellow Triangles left in the population of 2940 remaining items. Therefore, the probabily will be 157/2940≈0.05340136, different from the 158/2941≈0.05372333. And, if the first item chosen was not a Yellow Triangle then we still have 158 of those in the population but the population is now down to 2940 so the probability for the second item is 158/2940≈0.0537415, yet another value.

Dealing with the changing probabilities of sampling without replacement is a royal pain. It would be nice if we could ignore those changes, and, in fact the three values that we saw are really not very different. Our rule will be that as long as we sample less than 5% of the population we can ignore the changes in the probabilities due to sampling without replacement.