Chapter 3 Class Notes – Intro to Probability

Concept: role a fair die, then: what is the probability of getting a '3'? Getting a '3' in one roll of a fair die is called an Event and denoted E. In general,

$$\Pr\{E\} = \frac{\text{Number of successful outcomes}}{\text{Total number of outcomes}}$$

This is the *frequency interpretation* of probability (and gives the answer 1/6 for the above); another interpretation is the *mathematical definition*.

Another example – throw a pair of dice and let E be the event that neither die shows a '6' (i.e., both are non-6's). By counting below, we see that Pr{E} = 25/36 = 69.44%.

		Second Die					
		1	2	3	4	5	6
First Die	1	*	*	*	*	*	
	2	*	*	*	*	*	
	3	*	*	*	*	*	
	4	*	*	*	*	*	
	5	*	*	*	*	*	
	6						

Here that the tosses are 'independent' and that we can also find the answer 25/36 by multiplication: Pr{E} = 5/6 × 5/6 = 0.6944.

In contrast, suppose that we have a box containing six pieces of paper (called 'tickets') – your name is on one of the tickets. Your friend randomly selects two tickets *without replacement*, and let E be the event that your name is not selected (in the two draws). Draws here are 'dependent'. Then, Pr{E} = 5/6 × 4/5 = 2/3 = 0.6667, and this differs from the above answer.

Illustration of Probability Trees (p.94 #3.2.5a)

An EPT (early pregnancy test) is such that if a woman is Pregnant, then the probability she tests Positive (+) is 98%. Further, if she is Not Pregnant, there is a 99% chance that she will test Negative (-). By p.92, we then say that the sensitivity of this EPT is 98% and the specificity of the test is 99%. Additionally, of every group of 100 women in a certain village, suppose 10 are indeed Pregnant. If one woman from this village is randomly selected, what is the probability that she tests Positive (+)?

		Product
	|-- 0.98 -- Tests +	0.098
|-- 0.10 -- Pregnant	|	
|	|-- 0.02 -- Tests -	0.002
|		
|	|-- 0.01 -- Tests +	0.009
|-- 0.90 -- Not Pregnant	|	
	|-- 0.99 -- Tests -	0.891

Answer = 0.098 + 0.009 = 0.107 = 10.7%.

[Aside: note that here the false negative rate is 0.2%, the false positive rate is 0.9%, & so this test is in error 1.1% of the time.]

Rules of Probability (Section 3.3)

1. **The Addition Rule** – for any two events E_1 and E_2,

$$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \text{ and } E_2\}.$$

   If the events are disjoint, then just add the probabilities.

2. **Conditional probability** – given that an event E_1 has occurred (i.e., it is certain), then the probability that E_2 will occur is

$$\Pr\{E_2 / E_1\} = \frac{\Pr\{E_1 \text{ and } E_2\}}{\Pr\{E_1\}}$$

   provided $\Pr\{E_1\} \neq 0$. For example, if a woman is Pregnant, then the probability that she tests + is

   $\Pr\{\text{Tests +} / \text{Pregnant}\} = 0.098 / 0.10 = 0.98$ (this is the given "sensitivity" of the EPT).

3. **The Multiplication Rule** – for any two events E_1 and E_2,

$$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2 / E_1\}$$

   We used the multiplication rule above for the example regarding drawing two tickets from the box. We also used the multiplication rule with the above pair of dice example since for *independent events* E_1 and E_2,

$$\Pr\{E_2 / E_1\} = \Pr\{E_2\} \text{ and } \Pr\{E_1 / E_2\} = \Pr\{E_1\}$$

4. Bayes Rule – relates *prior* and *posterior probabilities* by:

$$\Pr\{E_1 / E_2\} = \frac{\Pr\{E_2 / E_1\} \times \Pr\{E_1\}}{\Pr\{E_2\}}$$

For example, given that a woman tests positive on the EPT (E_2), what is the probability that she really is pregnant (E_1)?

$$\Pr\{E_1 / E_2\} = \frac{0.98 \times 0.10}{0.107} = 0.9159 = 91.59\%$$

Note how different this conditional probability is from the unconditional probability that a woman is pregnant (10%).

Density Curves (Section 3.4)

These curves correspond to the population (of a continuous RV) and are thus "theoretical"; the progression of histograms on p. 99 (as the sample size grows) is helpful to understand the connection between histograms and density curves. These curves help us connect Area on the one hand with {Percentages, Probabilities, or Proportions} on the other hand. Thus, the area under the density curve is one (i.e., 100%).

To illustrate, for a randomly selected individual from the population generating Figure 3.4.4 on p.100, the probability that his/her blood glucose is greater than 150 mg/dLi is Pr{Y > 150} = 8%. Also, referring to Figure 3.4.5 on p.101, the probability that a randomly selected Douglas fir tree has diameter of between 2" and 8" is Pr{2 < Y < 8} = 0.20 + 0.33 + 0.25 = 0.78 (78%); the

probability that one such tree's diameter is exactly 2" is Pr{Y = 2} = 0 (zero) since Y here is continuous. Thus, the area under the density curve between 'a' and 'b' equals the proportion of Y values between 'a' and 'b' and equals the probability that one such 'individual' has a Y between 'a' and 'b'.

Random Variables (Section 3.5)

A random variable (RV) Y is a variable whose value depends upon the outcome of a chance operation.

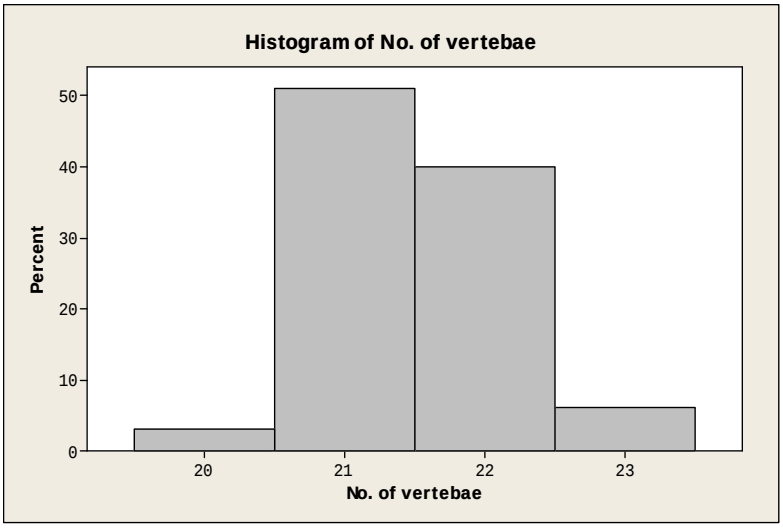
For example, randomly choose a card from a deck of cards, then Pr{face card} = 12/52 = 3/13 (face card is J or Q or K). Throw a pair of dice, Pr{sum exceeds 4} = 1 – 6/36 = 30/36 = 5/6 = 83.33%

SUM		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

In Ex. 3.5.5 on pp.103-4, we are given the distribution for Y = the number of tail vertebrae of the freshwater sculpin *Cottus rotheus*.

Y	Percent of Fish
20	3
21	51
22	40
23	6

Note that this RV is discrete, and note also that in the following frequency histogram that the bars extend from 19.5 to 20.5 for Y = 20, 20.5 to 21.5 for Y= 21, etc.



If we randomly select one such sculpin and let Y be the number of its tail vertebrae, then (a) Pr{Y = 21} = 0.51 and (b) Pr{Y ≤ 22} = 0.94. As pointed out above, the *probability* that a randomly selected 'individual' has a certain characteristic is equal to the *proportion* of the population with that characteristic.

The tree diameter illustration (p.101) is an example of a *continuous* RV; the vertebrae example (p.104) is a *discrete* RV.

Definitions: The **mean** (or **expected value**) of a **discrete** random variable is

$$E(Y) = \mu_Y = \Sigma\, y_k \times Pr\{Y = y_k\}$$

The **variance** of such a RV is

$$VAR(Y) = \sigma_Y^2 = \Sigma\, (y_k - \mu_Y)^2 \times Pr\{Y = y_k\}$$

Notes:

- For **continuous** RVs, substitute $\int$ (integral) signs above instead of Σ signs
- $E(aY + bZ) = aE(Y) + bE(Z)$
- $VAR(cX + d) = c^2 VAR(X)$
- The variance short-cut formula is

$$\sigma_Y^2 = \Sigma\,[y_k^2 \times Pr\{Y = y_k\}] - \mu_Y^2$$

For the sculpins, $\mu = 20 \times 0.03 + 21 \times 0.51 + 22 \times 0.40 + 23 \times 0.06$
$= 21.49$ (vertebrae)

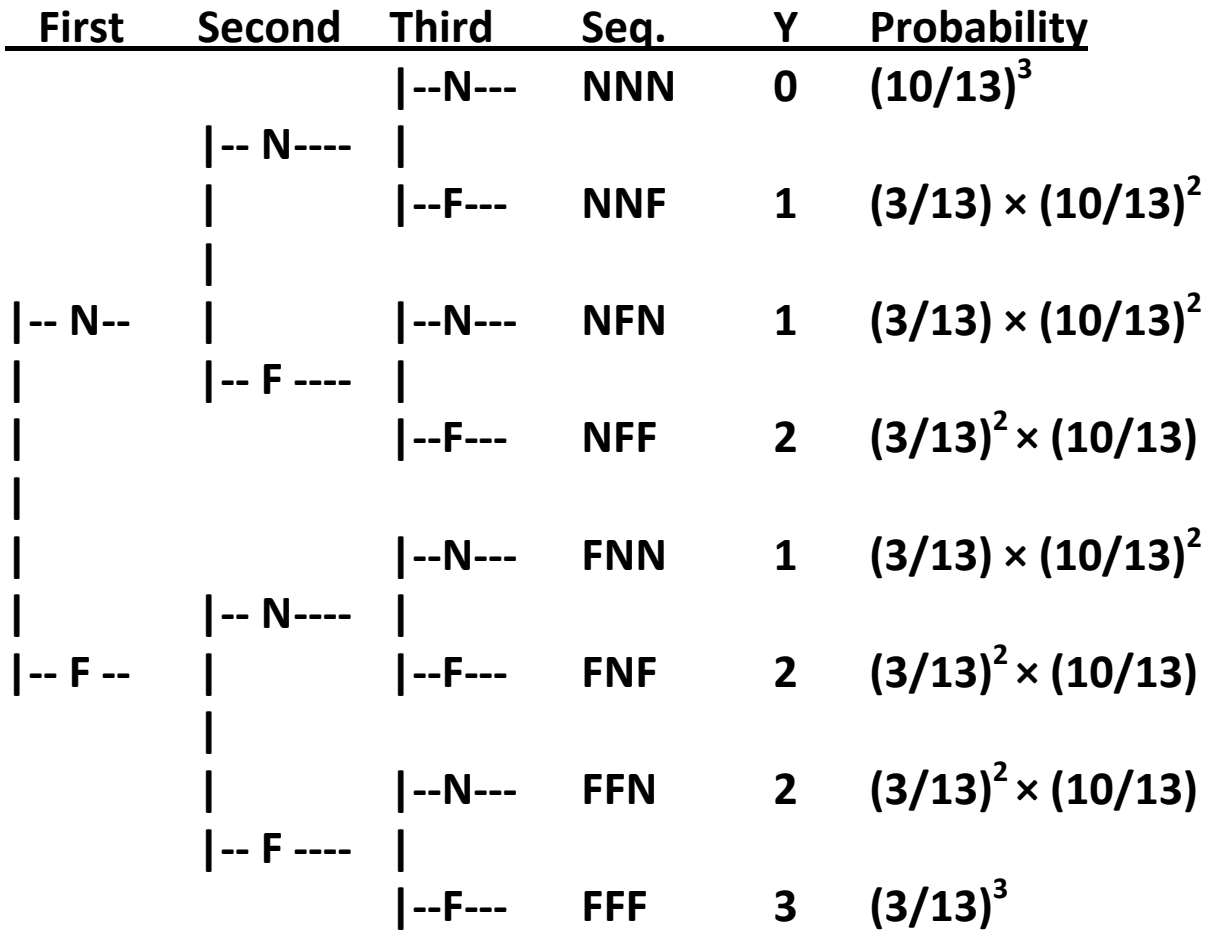
$\sigma^2 = (20 - 21.49)^2 \times 0.03 + (21 - 21.49)^2 \times 0.51 + (22 - 21.49)^2 \times 0.40$
$+ (23 - 21.49)^2 \times 0.06 = 0.4299$

so $\sigma = 0.6557$.

For the sum of 2 dice, verify that $\mu = 7$ and $\sigma^2 = 1974/36 - 7^2 = 5.8333$ so $\sigma = 2.4152$.

Now draw n = 3 cards from a fair deck of cards *with replacement* and note whether the card is a face card (F) or not (N); remember, the probability of F is 3/13 and the probability of N is 10/13.

We can list the possibilities using a tree (Y = # of face cards):



The Binomial Distribution (Section 3.6)

In the above example, Y = the number of picture cards in n draws with replacement is an example of a Binomial random variable. For such a variable, we need all of these conditions:

- **Each trial results in a 'Success' or a 'Failure' and we are interested in Y = # of successes [B]**
- **The outcomes of the trials are independent of each other [I]**
- **The number of trials (n) is fixed in advanced [n]**
- **The probability of a success (p) stays the same from trial to trial [S]**

These are called the 'BInS' requirements (see p.110) – all four must be met in order for a RV to be a Binomial RV. Note that not all 'S/F' random variables are Binomial – two counterexamples are (1) count the number of cards until we see the 5th Face card, and (2) if the n = 3 cards above are drawn *without* replacement.

For Y a Binomial RV with parameters 'n' and 'p' – we write Y ~ Bin(n,p) – then the probability of *exactly k successes* can be computed using the binomial probability formula:

$$Pr(Y = k) = {}_nC_k p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots n$$

Here, ${}_nC_k = \frac{n!}{k!(n-k)!}$ **and (n factorial) n! = n(n-1)…(2)(1) and 0! = 1. Most calculators can find $_nC_k$ or we can use Table 2 on p.615.**

For the Binomial distribution -- and only for this distribution -- the following short-cut formulas hold for the mean and variance:

$\mu_Y = np$, **and** $\sigma_Y^2 = np(1-p) = npq$ **(where q = 1-p)**

For the card example above, a success = a face card, n = 3, p = 3/13 (recall Y is the number of face cards!), so

$$\Pr\{Y = 0\} = \frac{3!}{0!3!}(3/13)^0(10/13)^3 = 0.455166$$
$$\Pr\{Y = 1\} = \frac{3!}{1!2!}(3/13)^1(10/13)^2 = 0.409650$$
$$\Pr\{Y = 2\} = \frac{3!}{2!1!}(3/13)^2(10/13)^1 = 0.122895$$
$$\Pr\{Y = 3\} = \frac{3!}{3!0!}(3/13)^3(10/13)^0 = 0.012289$$

Of course, the probabilities sum to one. In n = 3 draws (with replacement), the expected number of face cards is μ_Y = np = 3(0.2308) = 0.6923, give or take $\sigma_Y = \sqrt{3(0.2308)(0.7692)} = 0.7298$.

Another example – variation on Ex. 3.6.5 on p.112 – Drosophila, where 30% are black (B) and 70% are gray (G). Randomly select n = 10 such flies. (Even though sampling here is without replacement, the population of flies is infinite, so we can act as though sampling here is with replacement.) We get:

a) The probability exactly 2 flies are B.

$$\Pr\{Y = 2\} = \frac{10!}{2!8!}(0.3)^2(0.7)^8 = 0.233474$$

b) The probability that exactly 20% of the flies is B.

c) The probability that all of the flies are G.

$$\Pr\{Y = 0\} = \frac{10!}{0!10!}(0.3)^0(0.7)^{10} = 0.028248$$

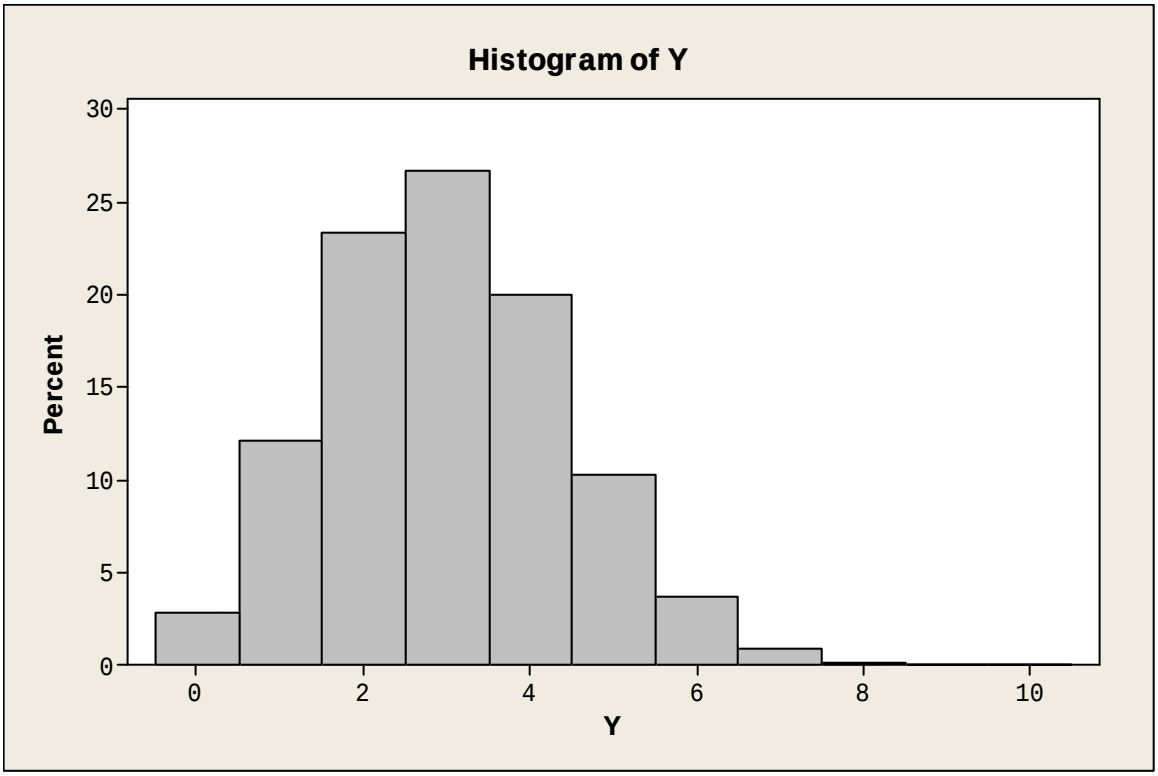
d) The probability that at least one of the flies is B.

$$\Pr\{Y \geq 1\} = \Pr\{Y > 0\} = 1 - \Pr\{Y = 0\} = 0.971752$$

e) The expected number of B flies and the give-or-take number.

μ = 10(0.3) = 3 B flies give or take $\sigma = \sqrt{10(0.3)(0.7)} = 1.45$ flies.

The probability histogram for this example is the following:



[Skip Section 3.7 for now.]