

Chapter 6 Discrete Probability Distributions

While a frequency table shows the actual results of an experiment, a **probability distribution** shows all **possible outcomes** and their **probabilities**. This is exactly the same thing as a relative frequency table **except that it includes the probabilities for the whole population, not just a sample**.

A **random variable** associates a **single numerical value** with **each outcome** of an experiment. A random variable is treated *mathematically* the same as the class midpoint in a frequency table. For example, if a 4 question exam is given, the possible number of right answers on any randomly selected exam are 0 correct, 1 correct, 2 correct, 3 correct, or 4 correct. If we let **x** be a **random variable** defined as the number of correct answers on a 4 question test, then we can show **x** as follows

[Number of correct answers
on a four question test]

x
0
1
2
3
4

This is an example of a **discrete random variable**. That is, it has either a **finite number or a countable number** of values. “x” must be either 0, 1, 2, 3, or 4.

A **continuous random variable** is the same concept except that it has **infinitely many values of “x”** with no gaps or interruptions between the values.

A **probability distribution** gives the **probability** for **each value of the random variable**. It can be either discrete or continuous, but it must show each random variable and its associated probability. In our present case, we know that it is possible to get 0 correct, 1 correct, etc. up to 4 correct. The question is, what is the probability of getting a randomly selected exam that has 0 correct, or 1 correct, or 2 correct, etc.

[No. of correct answers on a four question test]	[probability of x number of correct answers on a test]
x	$P(x)$
0	
1	
2	
3	
4	

What we need to know is this: what is the sample space of random variables for the problem, and what is the probability of each random variable occurring given that **$P(x)$ = the percentage of the exams that would have 0 correct, 1 correct, etc.**

Example: Assume students are given a test consisting of 4 True/False questions and that each student is totally unfamiliar with the topic and must guess at the answers. Each student then has a 50/50 chance to get any question correct. The possible number of questions answered correctly on any exam are the discrete random variables – 0 correct, 1 correct, 2 correct, 3 correct, or 4 correct.

We must then determine every possible outcome. This will allow us to calculate the probability ratio for any number of correct answers.

f_0 f_1 f_2 f_3 f_4

Adding up the total possible ways these 4 questions could be answered gives $\Sigma f = \underline{\hspace{2cm}}$ possible ways.

Random variable (number correct)
0 1 2 3 4

And the probability of each random variable:

$P(0)$ $P(1)$ $P(2)$ $P(3)$ $P(4)$

Alternate solution using combinations: (Combinations are appropriate here because our interest is in number of correct answers - if an exam has two correct answers, it is not important which two are correct, only that SOME two are correct.)

Total possible outcomes (combinations) =

Probabilities for each possible outcome:

We can now fill in the probabilities of each random variable in the above probability distribution:

Random Variable	
<u>x</u>	<u>$P(x)$</u>
0	
1	
2	
3	
4	

This probability distribution can also be shown as a **probability histogram** as below, with the probabilities on the y-axis and the random variables on the x-axis.

Note that on the above histogram, each bar is 1 unit wide and its height is the same as its probability. Since area = width x height, we can calculate the area of each bar, and add them together to get the total area graphed [the same as had been previously done with relative frequency]:

Important concept: The area of a probability graph (histogram) is equal to one. Therefore, the area above any “x” on the graph must be its probability (it’s percentage of the total area of the graph).

Assume that we have administered this true/false exam to a class of 100 students with the following results;

# correct	# of students	% of 100 students	% of population
<u>x</u>	<u>f</u>	<u>rel. f</u>	<u>P(x)</u>
0	5	.05	.0625
1	23	.23	.2500
2	39	.39	.3750
3	26	.26	.2500
4	<u>7</u>	<u>.07</u>	<u>.0625</u>
	100	1.00	1.0000

Note that the relative frequency from a sample of 100 students approximates the probabilities we determined analytically. The more students that we test (larger the sample) the closer the relative frequency will come to being exactly the same as the probabilities (which represent percents of the total population). If we cannot derive the true probabilities (percentages), then we will have to use the results of our sample to approximate them. That is, we would like to INFER that sample statistics are close enough to the true population parameter to be able to use the sample results. We know the sample is not exactly correct in representing the population. It really only represents the sample, so there will be some error in our inference. Luckily, statistics gives us a way to specify how much error we will permit, as we will see in a later chapter.

What is the probability of having between 1 and 3 questions inclusive answered correctly? Use both the addition rule and the areas of the bars.

Generalizing from the above, the following are **requirements** that must be met for a distribution of random variables to be a **probability distribution**.

$$\sum P(x) = 1 \quad \text{where } x \text{ assumes all possible values}$$

$$0 \leq P(x) \leq 1 \quad \text{for every value of } x$$

Note that this is just a restatement of the requirements laid out in Chapter 5.

The probabilities of x , $P(x)$, can sometimes be expressed as a **function of x** . For example, $P(x) = 1/x^2$, for $x =$ stated values. In a case like this, merely substitute each possible value of x from the sample space and solve for $P(x)$. However, when all possible values of x have been solved, the two criteria above must be satisfied.

Example: Determine the probability distribution of the random variable x given that $P(x) = 1/x$ and the values that x can assume (random variables) are 2, 4, 5, 20.

<u>x</u>	<u>$P(x)$</u>
-----------------------	--------------------------

If all values for $P(x)$ are the same (i.e., the probability of an occurrence of all random variables is the same), then the probability distribution is said to be a **uniform probability distribution**.

Homework: p. 163-164, Exercises 3, 4

Mean and Standard Deviation of Probability Distributions

Note that a probability distribution is mathematically exactly the same as a frequency table where the classes have been replaced with the random variable and the frequencies replaced with the probabilities (i.e., relative frequencies). Therefore, the same tools used to calculate mean and standard deviation from a frequency table can be used for probability distributions, with only a change in notation used in the formulas.

Of course, a probability distribution usually represents the total possible outcomes of an item of interest divided by the total number of outcomes possible

over some sample space. When dealing with such a probability distribution, (rather than one derived from taking a sample), we are dealing with the true probabilities of events. If a large enough sample were taken from this population, the relative frequency distribution of the sample would approach the true probabilities as the sample size grew. Therefore, for probabilities we utilize the population notation instead of the sample notation. That is, **the mean of the probability distribution** is represented by μ and the **standard deviation** by σ .

Therefore,

$$\mu = \sum xP(x)$$

$$\sigma = \sqrt{\sum (x - \mu)^2 P(x)}$$

The shortcut formula for standard deviation is

$$\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$$

and the variance is, of course, the square of the standard deviation.

Here, μ is a weighted average with each random variable “x” weighted by its probability.

Note that the empirical rule and the range rule of thumb, therefore, apply here also.

Where do these formulas come from? They are really just a restatement of the original mean and standard deviation formulas developed in Chapter 3. Remember, the mean from a frequency table is calculated using the formula

$$\bar{X} = \sum fx / \sum f \text{ or } \bar{X} = \sum wx / \sum w$$

where x is the class midpoint of each class and f is the frequency of the class (all f have the assumed value of the class midpoint).

For a probability distribution, the random variable assumes the role of the class midpoint, while the probability assumes the role of the frequency. Substituting P(x) for each f and μ for the mean gives the formula:

$$\mu = \sum xP(x) / \sum P(x)$$

But, $\sum P(x) = 1$, so the resulting formula is just $\mu = \sum xP(x)$ as above.

The standard deviation is calculated using the formula

$$s = \sqrt{[n\sum fx^2 - (\sum fx)^2] / n(n-1)}$$

As above, f represents P(x) so that $\sum f = 1$. But remember, n and $\sum f$ are the same thing, so that n must also be 1. Note that in this formula, n is some quantifiable

number. For large n , the denominator approaches $n \times n$ or n^2 , and $\sqrt{n^2} = n = \Sigma f = 1$. Therefore, the denominator becomes just 1 for probability distributions.

$$\sigma = \sqrt{[\Sigma x^2 P(x) - (\Sigma x P(x))^2 / 1]}$$

But, $\Sigma x P(x) = \mu$, and substituting

$$\sigma = \sqrt{[\Sigma x^2 P(x)] - \mu^2}$$

Remember the 4-question true-false test above. The number of possible correct answers on a randomly selected exam were 0, 1, 2, 3, and 4 (random variable). The probability distribution was given by

<u>x</u>	<u>P(x)</u>	<u>xP(x)</u>	<u>x²P(x)</u>
0	.0625		
1	.2500		
2	.3750		
3	.2500		
4	<u>.0625</u>		
	1.000		

In this case, “ n ” is the four questions on the exam (4). Solving for the mean and standard deviation using the above equations gives:

$$\mu = \Sigma x P(x) =$$

Does it make sense that 2 would be the mean number of correct questions on a 4-question true-false exam where everyone guesses at all the answers?

$$\sigma = \sqrt{[\Sigma x^2 P(x)] - \mu^2} =$$

Does it make sense that 1 would be the average deviation of the number of correct answers from the mean of 2 correct answers?

Test with range rule of thumb: $\sigma \approx \text{range}/4 = (4-0)/4 =$

Approximately 2/3 of the scores would be between what range of correct answers?

$$\mu + \sigma =$$

$$\mu - \sigma =$$

According to the empirical rule, ___ of the exams would have between ___ and ___ correct answers. That is, _____ correct answers.

What percentage have 1, 2 or 3 correct answers? Use the Addition Rule.

Why does the empirical rule give inaccurate results in this case?

Example: Arthur Andersen, LLP provides audits for major public companies. It is aware that its auditors (mostly new college graduates), do accurate audits 70% of the time. There are 6 areas of each company that are subject to audit. A senior partner wants to know the likelihood that a company randomly selected by the SEC for audit review will have errors in one or more of the 6 areas. Statisticians have provided the following distribution to the senior partner. (x is the random variable and represents the number of auditable areas that have errors out of six auditable areas).

(No. of areas with errors)	<u>x</u>	<u>$P(x)$</u>	<u>$xP(x)$</u>	<u>$x^2P(x)$</u>
	0	.118		
	1	.303		
	2	.324		
	3	.184		
	4	.060		
	5	.010		
	6	<u>.001</u>		
		1.000		

What is the probability of an error free review by the SEC?

What is the probability that 1 or more of the 6 areas will have errors?

What is the probability that the SEC will find errors in 3 or more of the six areas on any randomly selected review?

What is the mean number of incorrectly audited areas that the SEC will find?

$$\mu = \sum xP(x)$$

What is the standard deviation of the areas found in error on the companies audited?

$$\sigma = \sqrt{[\sum x^2P(x)] - \mu^2}$$

If you were the senior partner, would you be anxious to have the SEC review the audits?

Note the major drawback of calculating the mean and standard deviation of discrete probability distribution is that it is necessary to know the value of each random variable x and its probability.

Homework: p. 163-164, Exercises 1, 2, 7

The Binomial Distribution

A binomial distribution is a probability distribution that meets special requirements. That is, it can be treated just like any other discrete probability distribution, but since it is a special case we have some special tools to deal with it.

The binomial is special because it meets certain requirements that simplify how we can treat it statistically. The name "binomial" means "two-named" and a binomial distribution is one that exhibits "twoness", such as pass-fail, good-bad, or defective-non-defective. Note that these are always "complementary" events (D and $\sim D$).

To use the tools of dealing with a binomial experiment, all of the following conditions must be met:

1. There are only two possible outcomes, one identified as a **Success (the event of interest)** and the other as a **Failure**.
2. The probability " π " of a success on any one draw is known and constant. This " π " is also known as the proportion, or percentage, of the population that exhibits the characteristics of the "success". Thus, in the statement "90% of the parts are not defective", $\pi = 0.90$. Hence, 10% of the parts ARE defective. The 10% that are defective are designated by " $(1 - \pi)$ ", since together they comprise 100% of the population. In this example, $(1 - \pi) = 0.10$.
3. The events "success" and "failure" are independent of each other. In other words, the percentage of the population that is a success " π " or a failure " $(1 - \pi)$ " is assumed not to change as a result of our experiment. Therefore, on any random draw from the population, the probability of getting a "success" will not change (sampling with replacement or from very large populations).
4. We will always deal with samples of the same size "n". If samples of another size are taken, then the probability distribution will change and results are invalid. This is true because the sample size determines the random variables. For example, a sample of size $n = 3$ has four possible outcomes - 0 successes, 1 success, 2 successes, and 3

successes and an associated probability with each. If we take a sample of size $n = 5$, there will be 6 possible successes (0 successes through 5 successes), so the probability of getting 0 successes, or 1 success, etc. will change.

Example: Finished product coming off an assembly line in batches of 20 has a defect rate of 6%. $n =$ fixed sample size = 20. There are two categories of outcomes (defective and not defective), and the probability of any one randomly selected part being defective is always constant at 6% because that is the percentage of parts that are defective (assuming statistical stability of the assembly process). Also, the probability of any randomly selected assembly being defective is unaffected by having previously found a defective product (independence).

Of interest in this example is the probability that a given batch of 20 (being shipped to a customer, for example), will have 0 defects out of the 20, or 1 defect out of the 20, or 2 defects out of the 20, etc., up to 20 defects out of the 20. This is what will define our **probability distribution - the random variable being the number of defects out of a batch of size 20 and the associated probabilities of each number of defects. If we denote**

$P(\text{Success}) = \pi$ (note : a success is defined as the event of interest in the experiment. If we are looking for failures (defects), then actually finding a defect is a success!)

$P(\text{Failure}) = (1 - \pi)$ note: π and $(1 - \pi)$ are complementary events.

$n =$ fixed sample size on which the experiment is conducted.

$x =$ the number of successes we are interested in out of the trial of size n . That is, x is a random variable that can have any value between 0 and " n ".

$P(x) =$ probability of getting exactly x successes out of the n trials.

Then the probability distribution is a listing of each possible x and its probability $P(x)$.

For a binomial problem as described, the probability of any specific x , $P(x)$ is given by the following formula.

$$P(x) = (n! / [(n-x)!x!]) \pi^x (1 - \pi)^{n-x}$$

Which can also be written

$$P(x) = {}_n C_x \pi^x (1 - \pi)^{n-x}$$

where ${}_n C_x$ is the number of possible combinations with exactly x successes among n trials, and $\pi^x(1-\pi)^{n-x}$ is the probability of x successes among n trials for any one particular combination.

That is, the probability of x successes in a random sample of "n" items = the number of possible ways (combinations) that "x" success can happen (${}_n C_x$) multiplied by the constant probability of getting any one of the combinations on a random draw: $\pi^x(1-\pi)^{n-x}$.

Example: An experiment is run that has 6 fixed trials repeated, and the percentage of successes in the population is known to be 95%. What is the probability of getting 4 successes out of a trial, or $P(4)$? (Note: to find the entire probability distribution, the binomial formula must be solved 7 times, for $x=0$, for $x=1$, for $x=2$, etc., up to $x=6$. Fortunately, for "n" up to 15 and certain values of " π ", a table (Appendix A) is provided to give us the complete distribution).

$n =$

$x =$

$\pi =$

$(1-\pi) =$

x

$P(x)$

Now, fill in the entire distribution using the table. How close was your calculation to the value for $P(4)$ given in the table? Why might they be different?

x

$P(x)$

Example: The space shuttle has three flight control systems, one principal system and two backups in the event of a failure of the principal system. Tests have shown that a flight control system operates properly 90% of the time. Find the probability distribution for successful operation of the space shuttle flight control system. That is, find the probability that at least one of the three flight control systems works. Find using the binomial formula (4 times) and compare your results to the table.

(Note: the random variable "x" is the number of possible working systems on a flight, so x = 0 means that the principal and both back-up systems have failed.

n = $\pi =$ (1- π) = x =

(no. of working systems)

x P(x)

What is the probability that shuttle will crash due to having no working flight control system?

Mean and Standard Deviation of Binomial Distributions

One of the great things about binomial distributions is that we can determine their mean and standard deviation without having to know each x and its associated P(x). Since the binomial distribution is a probability distribution, the mean and standard deviation formulas for probability distributions (presented above) can be used for binomial distributions also. However, some very useful shortcut formulas have been developed. They are

$$\mu = n\pi$$

$$\sigma = \sqrt{n\pi(1-\pi)}$$

Again, the variance is just the square of the standard deviation.

To use these formulas, we need only know the sample size and the fixed probability of success. To illustrate, take another look at the Arthur Andersen audit example given previously. This is a binomial problem because it meets the four criteria. Here, n = 6, p = 0.3 (we're looking for errors, so finding an error is a success), and q = 0.7. Therefore,

$$\mu = n\pi =$$

$$\sigma = \sqrt{n\pi(1-\pi)} =$$

.How does this compare with the mean and standard deviation calculated earlier using non-binomial techniques? (see p. 8)

What is the mean and standard deviation for number of correct answers on the four-question true-false test (also a binomial problem)?

$$n =$$

$$\pi =$$

$$(1-\pi) =$$

$$\mu = n\pi =$$

$$\sigma = \sqrt{n\pi(1-\pi)} =$$

How does this compare to the conventional calculation? (see p. 7)

In the two examples above, verify $P(x)$ for each x using the binomial formula.

Homework: pp. 170-172, Exercises 9, 11, 14, 15, 18

p. 173, Exercises 20, 22, 24

pp. 178-180, Exercises 32, 33, 35, 37, 39, 42, 45, 50,

Comprehensive Example – random variables, probability distributions, binomial distributions.

Based on a campus-wide survey, it is determined that 40% of MTC students smoke. That is, if we selected a student at random, the probability that the student smokes would be $\pi = 0.40$. Therefore, the probability that the student doesn't smoke is $(1 - \pi) = 0.60$.

Suppose that we are interested in determining the probabilities of the number of smokers in a class of 4 students ($n = 4$). For a class of 4 students, there could be 0 smokers, 1 smoker, 2 smokers, 3 smokers, or 4 smokers. Note that there are five possible outcomes for the "number of smokers in a class of 4 students". These five outcomes, 0, 1, 2, 3, 4, are the **random variables (x)**, and there is always one more random variable than there is n.

We now need to assess the probability of getting each of these outcomes (0 smokers, 1 smoker, etc.).

The probability of getting 0 smokers, $P(0)$, is determined by the probability of a random individual being a smoker and the total possible number of outcomes. This is true of all probabilities: $P(\text{event } x) = \text{no. of event } x\text{'s} / \text{total possible number of events}$.

For an outcome of no smokers, there is only one possible way to have this outcome – no one smokes. Therefore, since there is a probability of 0.60 that a random individual does not smoke, we use the multiplication rule to determine $P(0)$. That is, for an outcome of 0 smokers, student 1 doesn't smoke AND student 2 doesn't smoke AND student 3 doesn't smoke AND student 4 doesn't smoke. [$S_i = \text{student } i=1, 2, 3, 4, \text{ smoking}$; $\sim S_i = \text{student } i=1, 2, 3, 4 \text{ not smoking}$].

Are the $P(S)$ independent?

Only one way to get 0 smokers:

$$P(0) = P(\sim S_1) \times P(\sim S_2) \times P(\sim S_3) \times P(\sim S_4) = 0.6 \times 0.6 \times 0.6 \times 0.6 = \underline{\underline{0.1296}}$$

Or 0.1296×1

How many combinations for 4 students taken 0 smokers at a time?

Using the above, write the equation for $P(0)$.

Ways to get 1 smoker – one of the student smokes AND none of the other three students smoke:

$$\mathbf{P(1)} = P(S_1) \times P(\sim S_2) \times P(\sim S_3) \times P(\sim S_4) = 0.4 \times 0.6 \times 0.6 \times 0.6 = \mathbf{0.0864}$$

$$\mathbf{OR} = P(\sim S_1) \times P(S_2) \times P(\sim S_3) \times P(\sim S_4) = 0.6 \times 0.4 \times 0.6 \times 0.6 = \mathbf{0.0864}$$

$$\mathbf{OR} = P(\sim S_1) \times P(\sim S_2) \times P(S_3) \times P(\sim S_4) = 0.6 \times 0.6 \times 0.4 \times 0.6 = \mathbf{0.0864}$$

$$\mathbf{OR} = P(\sim S_1) \times P(\sim S_2) \times P(\sim S_3) \times P(S_4) = 0.6 \times 0.6 \times 0.6 \times 0.4 = \mathbf{0.0864}$$

$$\text{Using ADDITION RULE for "OR", we add them up} = \mathbf{0.3456}$$

Adding them up is the same as multiplying 0.0864×4 (number of possible outcomes). Are the $P(1)$'s mutually exclusive?

How many combinations for 4 students taken 1 smoker at a time?

Using the above, write the equation for $P(1)$.

Ways to get 2 smokers – 1 student smokes AND another student smokes AND another student doesn't smoke AND the last student doesn't smoke.

$$\mathbf{P(2)} = P(S_1) \times P(S_2) \times P(\sim S_3) \times P(\sim S_4) = 0.4 \times 0.4 \times 0.6 \times 0.6 = \mathbf{0.0576}$$

$$\mathbf{OR} = P(\overline{\sim S_1}) \times P(S_2) \times P(S_3) \times P(\overline{\sim S_4}) = 0.6 \times 0.4 \times 0.4 \times 0.6 = \mathbf{0.0576}$$

$$\mathbf{OR} = P(\sim S_1) \times P(\sim S_2) \times P(S_3) \times P(S_4) = 0.6 \times 0.6 \times 0.4 \times 0.4 = \mathbf{0.0576}$$

$$\mathbf{OR} = P(S_1) \times P(\sim S_2) \times P(\sim S_3) \times P(S_4) = 0.4 \times 0.6 \times 0.6 \times 0.4 = \mathbf{0.0576}$$

$$\mathbf{OR} = P(\sim S_1) \times P(S_2) \times P(\sim S_3) \times P(S_4) = 0.6 \times 0.4 \times 0.6 \times 0.4 = \mathbf{0.0576}$$

$$\mathbf{OR} = P(S_1) \times P(\sim S_2) \times P(S_3) \times P(\sim S_4) = 0.4 \times 0.6 \times 0.4 \times 0.6 = \mathbf{0.0576}$$

$$\text{Again, using the ADDITION RULE to add up the "OR's"} = \mathbf{0.3456}$$

$$\text{Or } 0.0576 \times 6 = 0.3456.$$

How many combinations for 4 students taken 2 smokers at a time?

Using the above, write the equation for $P(2)$.

Ways to get 3 smokers - 1 student doesn't smoke AND all of the other 3 students smoke.

$$\mathbf{P(3)} = P(\sim S_1) \times P(S_2) \times P(S_3) \times P(S_4) = 0.6 \times 0.4 \times 0.4 \times 0.4 = \mathbf{0.0384}$$

$$\mathbf{OR} = P(S_1) \times P(\sim S_2) \times P(S_3) \times P(S_4) = 0.4 \times 0.6 \times 0.4 \times 0.4 = \mathbf{0.0384}$$

$$\mathbf{OR} = P(S_1) \times P(S_2) \times P(\sim S_3) \times P(S_4) = 0.4 \times 0.4 \times 0.6 \times 0.4 = \mathbf{0.0384}$$

$$\mathbf{OR} = P(S_1) \times P(S_2) \times P(S_3) \times P(\sim S_4) = 0.4 \times 0.4 \times 0.4 \times 0.6 = \mathbf{0.0384}$$

Using the ADDITION RULE to add the "OR's" = **0.1536**

Or $.0384 \times 4$.

How many combinations for 4 students taken 3 smokers at a time?

Using the above, write the equation for P(3).

Only one way to get 4 smokers – they all smoke.

$$\mathbf{P(4)} = P(S_1) \times P(S_2) \times P(S_3) \times P(S_4) = 0.4 \times 0.4 \times 0.4 \times 0.4 = \mathbf{0.0256}$$

How many combinations for 4 students taken 4 smokers at a time?

Using the above, write the equation for P(4).

Write the generalized equation for P(x).

The probability distribution for this problem is:

<u>x</u>	<u>P(x)</u>
----------	-------------

0

1

2

3

4

Draw the probability histogram for this distribution.

Calculate the area of each bar of the histogram. What is the total area under the graph?

What is the mean of the distribution? Use both the formula for a probability distribution and the one for binomial distributions.

What is the standard deviation of the distribution? Use both the formula for a probability distribution and the one for binomial distributions.

What is the variance of the distribution?