2.4: Frequency Distribution Tables

There are a variety of tables that can be used to summarize the frequency with which scores were observed for a quantitative variable. The most basic version of this is simply referred to as a frequency distribution.

Creating Frequency Distributions using Scores

A frequency distribution is a table used to summarize a quantitative variable by showing how frequently each score occurred. A frequency distribution has two columns. When the range of values is somewhat small (i.e. generally the difference between the highest score and the lowest score is 20 or less), the first column is titled *Score*. The score column shows the various scores which could have occurred starting from the highest which was observed (top row of data) to the lowest which was observed (bottom row of data). The second column is titled *Frequency*. To fill in frequency column, the statistician counts up how many times each score was observed in the data. Then, they put the count in the corresponding row. Take a look at Table 4 below. This is a frequency distribution for the variable Years of Experience built using the scores from Data Set 2.1.

Table 4 Frequency of Years of Experience (n = 18)

Score	Frequency
7	2
6	3
5	4
4	3
3	2
2	2
1	2

Interpretation

Let's review how to read the table before going over the rules used to create it. The title tells us the variable being summarized and the sample size. The variable is Years of Experience and the sample size is 18. The score column shows all the possible scores that could have occurred ranging from the highest which was observed down to the lowest which was observed. We can see that experience ranged from 1 to 7 years. The frequency column tells us how frequently each score occurred. In this table, the frequencies refer to how many cases in the data set were people who had 1, 2, 3, 4, 5, 6, or 7 years of experience. We can see that the most frequent number of years of experience was 5 years because 4 cases were of people with this many years of experience. The least common years of experience reported were 7, 3, 2, and 1 because each of these amounts of experience were reported by two people. We also know that no one in the dataset reported having 0 years of experience or more than 7 years of experience because if they had, we would have rows for those scores in the table. If you sum the values in the frequency column it will be equal to the sample size. This is because the frequency column shows how many cases in the sample had each score and, thus, was created by simply dividing up the sample.

Construction

To construct a frequency distribution a statistician starts by organizing the data to get a sense of the range. We did this with the data for Years of Experience in Table 3. The statistician must then identify the highest score, the lowest score, and the precision with which they were measured and/or rounded. Years of Experience was measured and shown to the year in intervals of 1 year. The statistician then uses this information to construct the score column. Though it varies for different fields and circumstances, it is common in many behavioral and social sciences to create the score column going from the lowest score at the bottom toward the highest score at the top. Therefore, the score column for Table 4 was constructed counting up from lower scores (starting at 1) to higher scores (ending at 7) in increments of 1. However, you may also see tables which are constructed from lowest on the top to highest at the bottom when this is the better fit for the field or data.

When a score column is constructed, no rows are needed for values outside the range within which data were observed. This makes it easy for anyone reading the table to quickly deduce the range and know that scores beyond it were not observed. If we want to know how many people had 0 years of experience or 10 years of experience in Table 4, for example, we can quickly deduce the

answer to both is 0 because these are outside the range shown and, thus, outside the range observed. However, a row for any score inside the observed range is retained in the table even it if has a frequency of 0. Suppose that no one had 3 years of experience but that all other data were the same for Table 4. In this case, we would simply put a 0 in the frequency column to the right of the score of 3 rather than deleting the row corresponding to this score.

Ideally, a frequency distribution should have between 10 and 20 rows, however, there are times when it is appropriate to have more than 20 rows or fewer than 10 rows. When there are more than 20 rows the table can be providing too much detail without summarizing enough. It also increases the cognitive load for the reader when there are many rows. When there are fewer than 10 rows, it sometimes means that the data are being over-simplified and too little detail is being shown. However, if the data have a small range, we sometimes must use fewer than 10 rows. Take a look again at Table 4. There are only 7 rows and those are sufficient to show all the detail needed and possible for the variable Years of Experience because the range was small. Therefore, it is appropriate to have a small table with only 7 rows for these data.

Creating Frequency Distributions using Intervals

However, there are times when the range of values for a variable is so large that all scores could not be listed in about 20 or fewer rows. When this occurs, statisticians create a frequency distribution using an interval column in place of a score column. Intervals break the range up into useful segments which summarize and organize the data. To enhance clarity and consistency, there are rules that should be followed when creating intervals. First, the intervals should all be the same size. Second, the intervals should be intuitive. By intuitive here we mean that the intervals should be in amounts that are easy to count in or understand such as ones (0, 1, 2, 3, 4, 5, and so on), twos (0, 2, 4, 6, 8, 10 and so on), fives (0, 5, 10, 15, 20, 25 and so on), tens (0, 10, 20, 30, 40, 50 and so on) or hundreds (0, 100, 200, 300, 400, 500 and so on). Third, the intervals must be mutually exclusive. This means the intervals cannot overlap so that each score fits into only one interval. Finally, it is best to have approximately 10 to 20 intervals as appropriate for the data. This last rule must be a bit flexible to allow the other three rules to be followed.

A variable such as Annual Salary from Data Set 2.1 necessitates the use of intervals. This is because values were measured to the dollar and ranged from 34,000 to 190,000. If each dollar amount from 34,000 to 190,000 got its own row, the table would have 156,001 rows. This would be overwhelming both to create and to read and wouldn't meet the primary goal of making the data easier to understand through a summary. No one wants that. Instead, intervals can be created to organize the data more efficiently.

Interval Construction

Creating intervals takes a bit of practice and thought but a good place to start is by finding the range for the variable. The number of dollar amounts starting from the highest salary down to and including the lowest salary, also known as the inclusive range, is 156,001. If we want 10-20 rows, we need to figure out how to divide this range intuitively and evenly. A good strategy is to divide the range by 10 to see the approximate interval size it would take to have 10 rows, then divide the range by 20 to see the approximate interval size it would take to have 20 rows. If we do this we will find the approximate interval size to get 10 rows is 15,600 and for 20 rows is 7,800. Neither of these are intuitive but we can choose a value somewhere between them that is intuitive such as 10,000. Intervals of 10,000 are easy to count in and understand and would cause our table to have about 17 rows to accommodate the full range of data. That is between 10 and 20 rows so it meets two of our criteria. Last, we need to create the intervals of 10,000 which are mutually exclusive. An easy and advisable place to start creating intervals for ratio level data such as salary is 0. Therefore, we can start by creating our lowest interval going from 0-9,999. This is an interval with 10,000 dollar amounts in it because 0 counts as the first value (making 1 the second value, 2 the third value and so on until we get to 9,999 as the 10,000th value).

Let's look at the construction process for creating a frequency distribution for Annual Salary. We started with 0-9,999. We continue to the interval 190,000-199,999 because this top interval will include the highest salary observed in the data. This causes us to have three intervals at the bottom that we do not need because the lowest income observed was 34,000. Therefore, we can remove those three before we create our frequency column. Drafting those bottom intervals and removing them can help folks know where to start their intervals so they won't be tempted to start the interval at the lowest value of 34,000 which could cause the bottom interval to go from 34,000-43,999 which is less intuitive and, thus, creates more work for the reader. However, some people can skip the steps of drafting and removing the bottom rows if they can remember to start their intervals at intuitive numbers (i.e. start at 30,000 rather than 34,000 for this table). Drafting and removing the bottom rows or simply starting at 30,000 are equally appropriate ways to create the interval column so it is best to use the method that makes the most sense to you.

Table 5 Interval Colum for Annual Salary

Interval

Interval
190,000-199,999
180,000-189,999
170,000-179,999
160,000-169,999
150,000-159,999
140,000-159,999
130,000-139,999
120,000-129,999
110,000-119,999
100,000-109,999
90,000-99,999
80,000-89,999
70,000-79,999
60,000-69,999
50,000-59,999
40,000-49,999
30,000-39,999
~~20,000-29,999~~
~~10,000-19,999~~
~~0-9,999~~

Next, we count how many salaries fell into each interval. This is easiest to do by first putting the data for the variable in order and then counting the occurrences. Therefore, we can look at Table 2 and begin counting. One raw score for salary was between 30,000 and 39,000 so the frequency for this interval was 1. One raw score was between 40,000 and 49,999 so the frequency for this interval was also 1. However, two raw scores were between 50,000 and 59,999 so the frequency for this interval is 2. We continue counting in this way until we have identified and filled in the frequencies for all intervals as shown in Table 6. Once all the frequencies are filled in it is good practice to sum them to make sure they are equal to the sample size. If we add the frequencies in Table 6 we get 18 which is equal to the sample size so no errors are readily apparent.

Table 6 Frequency Distribution for Annual Salary (n = 18)

Interval	Frequency
190,000-199,999	1
180,000-189,999	2
170,000-179,999	0
160,000-169,999	1
150,000-159,999	0
140,000-159,999	0
130,000-139,999	1
120,000-129,999	2
110,000-119,999	1
100,000-109,999	1
90,000-99,999	0
80,000-89,999	1
70,000-79,999	2

Interval	Frequency
60,000-69,999	2
50,000-59,999	2
40,000-49,999	1
30,000-39,999	1

Interpretation

Now that the frequency distribution has been created, let's take a moment to read it, paying close attention to a few thing. First, we can quickly deduce that no interval was especially common or uncommon relative to other intervals as there were between 0 and 2 cases for all 17 intervals shown. Thus, the incomes were spread out somewhat evenly with half the sample (9 cases) earning 100,000 or more and half (the other 9 cases) earning 89,999 or less. It is also quick and easy for us to identify how many salaries in the sample were in any given interval. For example, if you want to quickly know how many people earned between 50,000 and 59,999, you can look down to that row and easily see the count was 2. Compare this to trying to quickly identify how many salaries were between 50,000 and 59,999 in the raw data show in Table 1. It is going to take more time and effort to answer the question from the raw data.

Retaining Empty Rows

Some may wonder why we don't remove the intervals from the table with a frequency of 0 to make the table look even simpler. The irony is that removing those intervals makes the table shorter yet increases the cognitive load. Let's take a look to see why. Try to find the frequency with which salaries between 90,000 and 99,000 were observed in the improperly constructed version shown in the right side Table 7. Once you have your answer try it again using the properly constructed version shown in the left side Table 7.

Table 7 Proper and Improper Interval Columns for Annual Salary

Proper Construction		Improper Construction	
Interval	**Frequency**	**Interval**	**Frequency**
190,000-199,999	1	190,000-199,999	1
180,000-189,999	2	180,000-189,999	2
170,000-179,999	0	160,000-169,999	1
160,000-169,999	1	130,000-139,999	1
150,000-159,999	0	120,000-129,999	2
140,000-159,999	0	110,000-119,999	1
130,000-139,999	1	100,000-109,999	1
120,000-129,999	2	80,000-89,999	1
110,000-119,999	1	70,000-79,999	2
100,000-109,999	1	60,000-69,999	2
90,000-99,999	0	50,000-59,999	2
80,000-89,999	1	40,000-49,999	1
70,000-79,999	2	30,000-39,999	1
60,000-69,999	2		
50,000-59,999	2		
40,000-49,999	1		

Proper Construction		Improper Construction	
30,000-39,999	1		

It is easier to confidently deduce that the frequency of salaries between 90,000 and 99,999 in the sample was 0 using the properly constructed table than in the improperly constructed table. Next, take a look down the frequency column in each table. In the properly constructed table, the pattern is clear that some intervals occurred once or twice while others did not occur at all. We don't need to look at the interval column to know if any intervals had a frequency of 0; instead we only need to look at the frequency column to see which intervals had a frequency of 0. However, if we look down the frequency column of the improperly constructed version of the table, we cannot easily tell which, if any, intervals had a frequency of 0. Instead, we have to look at the interval column and try to figure out which intervals are missing. If an interval is missing we would assume that it was because the frequency was 0 but may not feel as confident. We might wonder if the person who created the table made a mistake and left out an interval. To check this we could add all the frequencies to see it if is equal to the sample size. If so, we can be fairly confident that the missing interval isn't a mistake but rather is meant to indicate the frequency for that interval was 0. We *could* do all this extra work but why? This is a lot more work than simply looking at the row for the interval in the properly constructed table to see if the frequency was 0. Thus, the reason intervals within the range with a frequency of 0 are retained is to meet the goal of making the data easier to understand and summarize.

Reading Review 2.1

Two of the tables below have construction errors and one is properly constructed. Identify which two have errors and specify the nature of the errors.

Data Set A, Daily Inches of Rainfall: 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 7 7

Table A Frequency of Daily Inches of Rainfall (n = 31)

Score	Frequency
7	2
6	0
5	0
4	3
3	5
2	4
1	8
0	9

Data Set B, Age in Years: 2 4 6 6 7 9 10 11 11 14 15 18 18 18 19 20 22 23 23 24 26 26 28 33

Table B Age in Years (n = 24)

Interval	Frequency
20-39	9
0-19	15

Data Set C, Length in Inches: 2 2 3 4 5 7 8 9 9 11 12 12

Table C Length in Inches (n = 12)

Score	Frequency
12	2
11	1

Score	Frequency
9	2
8	1
7	1
5	1
4	1
3	1
2	2

Cumulative Frequency Distributions

A third column can be added to a frequency distribution to create a table known as a cumulative frequency distribution. **Cumulative frequency distributions** are used to summarize the frequencies and cumulative frequencies of ordered scores for quantitative variables. The third column is added to the right of the frequency column and is given the title "Cumulative Frequency." This cumulative column counts up from the bottom row to tell how many cases occur at or below each row.

Let's take a look at an example. Table 8 is a cumulative frequency distribution for the variable Annual Salary. The first two columns are the same as those used in the frequency distribution for these data but a third column has been added. Starting from the bottom and counting up, the cumulative frequency column summarizes how many people were at or below each salary interval. In this example, starting from the bottom we can see that 1 case for salary was at or

below 39,999, 2 cases were at or below 49,999, 4 cases were at or below 59,999 and so on until we reach the highest interval. The total sample size appears as the top cumulative frequency because all scores are at or below the highest score or interval observed.

Table 8 Cumulative Frequency Distribution for Annual Salary (n = 18)

Interval	Frequency	Cumulative Frequency
190,000-199,999	1	18
180,000-189,999	2	17
170,000-179,999	0	15
160,000-169,999	1	15
150,000-159,999	0	14
140,000-159,999	0	14
130,000-139,999	1	14
120,000-129,999	2	13
110,000-119,999	1	11
100,000-109,999	1	10
90,000-99,999	0	9
80,000-89,999	1	9
70,000-79,999	2	8
60,000-69,999	2	6
50,000-59,999	2	4
40,000-49,999	1	2
30,000-39,999	1	1

Let’s review it one more time to better understand what the third column shows us. Notice that the bottom right shows a cumulative frequency of 1 because there is only 1 case at the interval of 30,000-39,999 or below (since there are no data below). Notice also that the second row from the bottom has a cumulative frequency of 2 because it adds 1 (the number of cases in the second interval from the bottom) to 1 (the total number of cases that occur below the second interval from the bottom). Cases are added as we move up the cumulative column so that we know how many cases are at or below each interval. This allows us identify specific thresholds easily.

Relative Frequency Distributions

There are other options for creating a third column, each of which makes a different kind of frequency distribution. One option is to add a third column to make a relative frequency distribution. **Relative frequency distributions** (also known as proportional frequency distributions) are used to summarize the relative frequencies of scores or intervals using percentages. The third column is given the title “Relative Frequency.” This column what percentage of the raw scores for the variable are represented in each row.

Let’s take a look at an example. Table 9 is a relative frequency distribution for the variable Annual Salary. The first two columns are the same as those used in the frequency distribution for these data but a third column has been added. Each row of the relative frequency column is computed by dividing the frequency for each row by the total sample size and then multiplying by 100 to calculate the relative percentage. For example, the frequency for the top interval of 190,000-199,999 is 1. The sample size is 18. Thus, the percentage is computed as follows:

$(1 \div 18)100 = (0.555\ldots)100 = 5.56\%$

The sum of all values in the relative frequency column must be 100% (give or take any rounding error introduced when reporting percentages for each interval).

Table 9 Relative Frequency Distribution for Annual Salary (n = 18)

Interval	Frequency	Relative Frequency
190,000-199,999	1	5.56%
180,000-189,999	2	11.11%
170,000-179,999	0	0.00%
160,000-169,999	1	5.56%
150,000-159,999	0	0.00%
140,000-159,999	0	0.00%
130,000-139,999	1	5.56%
120,000-129,999	2	11.11%
110,000-119,999	1	5.56%
100,000-109,999	1	5.56%
90,000-99,999	0	0.00%
80,000-89,999	1	5.56%
70,000-79,999	2	11.11%
60,000-69,999	2	11.11%
50,000-59,999	2	11.11%
40,000-49,999	1	5.56%
30,000-39,999	1	5.56%

Cumulative Percentage Distributions

A cumulative frequency distribution can also be made using percentages by either, creating a third column that reports the cumulative relative proportions or by adding a cumulative column to a relative frequency distribution. Thus, **cumulative percentage distributions** are often created as relative frequency distributions with a fourth column showing the cumulative percentages as shown in Table 10. Just like a cumulative frequency distribution table, the cumulative column is made by summing computations starting from the bottom row and working up. The sum of all values in the relative frequency column must be 100% (give or take any rounding error introduced when reporting percentages for each interval). To help avoid any such rounding error, it is best practice to find the cumulative frequency for each row, to divide that by the sample size, and then multiply by 100 to get the cumulative percentage rather than adding the relative frequencies which may have rounding error. For example, the cumulative frequency for the interval of 50,000-59,999 was 4. When 4 is divided by the sample size of 18 the result is 0.222… When this is multiplied by 100 to translate from a decimal to a percentage and then rounded to the hundredths place the result is 22.22%. The same procedure was used for each row to create the Cumulative Percentage column of Table 10.

Table 10 Cumulative Percentage Distribution for Annual Salary (n = 18)

Interval	Frequency	Relative Frequency	Cumulative Percentage
190,000-199,999	1	5.56%	100.00%
180,000-189,999	2	11.11%	94.44%
170,000-179,999	0	0.00%	83.33%
160,000-169,999	1	5.56%	83.33%
150,000-159,999	0	0.00%	77.78%
140,000-159,999	0	0.00%	77.78%
130,000-139,999	1	5.56%	77.78%
120,000-129,999	2	11.11%	72.22%
110,000-119,999	1	5.56%	61.11%
100,000-109,999	1	5.56%	55.56%
90,000-99,999	0	0.00%	50.00%
80,000-89,999	1	5.56%	50.00%
70,000-79,999	2	11.11%	44.44%
60,000-69,999	2	11.11%	33.33%
50,000-59,999	2	11.11%	22.22%
40,000-49,999	1	5.56%	11.11%
30,000-39,999	1	5.56%	5.56%

Percentile Rank Distributions

You may also see or need a table that reports the percentile rank of each score or interval. A **percentile rank distributions** can be created as relative frequency distributions with a fourth column showing the percentile rank (i.e. the percentage of scores below) for each score or interval as shown in Table 11. A percentile ranks refer to the percentage of raw scores below a given score of interval. For example, 0 raw scores in the data set are below the interval of 30,000-39,999 in Data Set 2.1 Thus, the percentile rank for that row is 0.00%. However, there are two raw scores below the interval of 50,000-59,999 and, thus, the percentile rank of this interval is 11.11%. Notice that the values in the percentile rank column for Table 11 are the same as those in Table 10 just shifted up one by one row. This is because Table 10 focused only on scores below each row.

Table 11 Percentile Rank Distribution for Annual Salary (n = 18)

Interval	Frequency	Relative Frequency	Percentile Rank

Interval	Frequency	Relative Frequency	Percentile Rank
190,000-199,999	1	5.56%	94.44%
180,000-189,999	2	11.11%	83.33%
170,000-179,999	0	0.00%	83.33%
160,000-169,999	1	5.56%	77.78%
150,000-159,999	0	0.00%	77.78%
140,000-159,999	0	0.00%	77.78%
130,000-139,999	1	5.56%	72.22%
120,000-129,999	2	11.11%	61.11%
110,000-119,999	1	5.56%	55.56%
100,000-109,999	1	5.56%	50.00%
90,000-99,999	0	0.00%	50.00%
80,000-89,999	1	5.56%	44.44%
70,000-79,999	2	11.11%	33.33%
60,000-69,999	2	11.11%	22.22%
50,000-59,999	2	11.11%	11.11%
40,000-49,999	1	5.56%	5.56%
30,000-39,999	1	5.56%	0.00%