Study Guide for Research Methods in Human Relations

William O. Ray

The University of Oklahoma

I. Descriptive Statistics: Percentiles

Example. International students applying for admission to the University of Oklahoma must take the Test of English as a Foreign Language (TOEFL) to ensure that their English language skills are adequate to benefit from instruction. Applicants to OU with less than a 550 TOEFL must complete an intensive English Language course prior to taking classes.

The TOEFL is administered by ETS. ETS does not report the raw TOEFL score but instead reports a score which is between 300 and 700 and always ends in a zero, a three or a seven. Thus, for example, a student who scores a 553 on the TOEFL will have gotten 84 items correct, a student who scores 550 will have gotten 83 items correct and a student who scores 547 will have gotten 82 items correct. (Why do you suppose TOEFL scores are reported in this manner?)

Table I.

Score	Frequency	Score	Frequency	Score	Frequency
513	2	517	1	520	1
523	1	527	1	530	1
533	5	537	4	540	7
543	8	547	10	550	49
553	46	557	50	560	33
563	39	567	38	570	36
573	37	577	25	580	22
583	35	587	24	590	21
593	23	597	19	600	25
603	21	607	23	610	15
613	18	617	14	620	17
623	19	627	18	630	18
633	11	637	13	640	6
643	12	647	5	650	7
653	4	657	7	660	4
663	4	667	2	670	1

Listed above are the TOEFL scores for the 802 international students enrolled on the Norman Campus of the University of Oklahoma. Our goal will be to summarize these scores in a more understandable manner: in a *histogram*. ▮

Solution. **Step 1.** The first step is to list the scores in order from lowest to highest; Table I already does this.

Step 2. The next step is to decide how many bars we wish to display; we will shoot for sixteen bars.

Step 3. Next we need to find the range of the observations (highest minus lowest):

$$\text{range} = 670 - 513 = 157$$

Step 4. Next divide the range by the number of bars to obtain the *cell ranges* needed for each bar:

$$\text{cell width} = \frac{157}{16} = 9.8$$

Since 9.8 is inconvenient, we will use 10 for our cell ranges.

Step 5. Since our lowest score is 513, we will select 511 as the bottom of the first cell (so that the top of the first cell, 520, is divisible by 10).

Step 6. Next construct a *frequency table*. Our cells look like:

Table II.

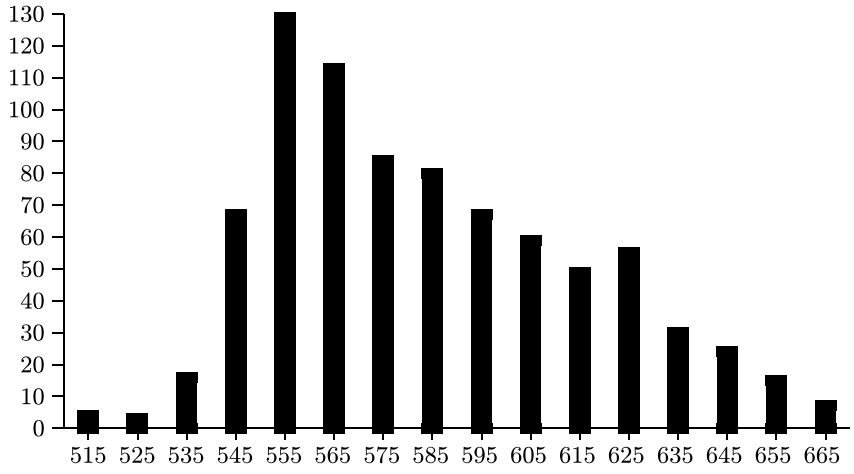
Range	Frequency
661-670	
651-660	
641-650	
631-640	
621-630	
611-620	
601-610	
591-600	
581-590	
571-580	
561-570	
551-560	
541-550	
531-540	
521-530	
511-520	

Counting how many scores occur in each range gives a frequency table of:

Table III.

Range	Frequency
661-670	7
651-660	15
641-650	24
631-640	30
621-630	55
611-620	49
601-610	59
591-600	67
581-590	80
571-580	84
561-570	113
551-560	129
541-550	67
531-540	16
521-530	3
511-520	4

Step 7. The resulting histogram looks like;



Question. Why do you suppose that the TOEFL scores for students admitted to OU do not look like a "bell curve?" Do you suppose that the scores for *applicants* (as opposed to the more select admitted group) would have a different appearance?

Solution Template

Step 1. Make a list of all of the observations in order from lowest to highest.

Step 2. Decide how many bars (cells) you wish to display. In general you should display at least five bars but no more than twenty.

Step 3. Find the *range* of the observations:

$$\text{range} = \text{highest} - \text{lowest}$$

Step 4. Divide the range by the number of bars to obtain the size of the individual cells. Round to a convenient whole number. This way all of the cells will be the same length; only bad histograms have cells of varying length.

Step 5. Select a convenient score for the bottom of the first cell; this should be no higher than the smallest observation. Write down the cells; make sure that the cells are collectively exhaustive (include all the scores) and mutually exclusive (don't overlap).

Step 6. Construct a frequency table, counting the number of scores in each range.

Step 7. Draw the histogram.

End of Solution Template

Percentiles

Another approach to summarize the data would be to compute percentiles. Given a score $\boldsymbol{x}$, the *percentile* corresponding to $\boldsymbol{x}$ is

$$\text{percentile} = \frac{\text{scores} \leq \boldsymbol{x}}{\text{total number of scores}} \times 100\%$$

Suppose, for example, we wished to know the percentile corresponding to a TOEFL score of 540. Based on the frequency table above, there are

$$4 + 3 + 16 = 23$$

scores less than or equal to 540. Since there are 802 scores altogether, this gives

$$\text{percentile} = \frac{23}{802} \times 100\% = 2.87\%.$$

We can in fact find the percentile corresponding to the top of each cell range. To do this, add a "cumulative frequency" column to Table III:

Table IV.

TOEFL Range	Frequency	Cumulative Frequency
661-670	7	802
651-660	15	795
641-650	24	780
631-640	30	756
621-630	55	726
611-620	49	671
601-610	59	622
591-600	67	563
581-590	80	496
571-580	84	416
561-570	113	332
551-560	129	219
541-550	67	90
531-540	16	23
521-530	3	7
511-520	4	4

The cumulative frequency for a range is the sum of all of the scores in that range and below that range.

Next, convert the cumulative frequencies to percentiles. The percentile for a score is the percent of scores equal to or less than the given score. Thus, for example, the percentile corresponding to 580 is $\frac{416}{802} \times 100\% = 51.87\%$ since 416 of the 802 scores are equal to or less than 580. Since the cumulative frequency column is the cumulative frequency for the entire range, we will have found the percentile corresponding to 580; the table does not provide sufficient detail to directly deduce the percentile corresponding to 577 or any other intermediate value.

In this manner we can then fill in Table IV by adding a cumulative percentile column:

Table V.

A TOEFL Range	*B* Frequency	*C* Cumulative Frequency	*D* Cumulative Percentile
661-670	7	802	100.00%
651-660	15	795	99.13%
641-650	24	780	97.26%
631-640	30	756	94.26%
621-630	55	726	90.52%
611-620	49	671	83.67%
601-610	59	622	77.56%
591-600	67	563	70.20%
581-590	80	496	61.85%
571-580	84	416	51.87%
561-570	113	332	41.40%
551-560	129	219	27.31%
541-550	67	90	11.22%
531-540	16	23	2.87%
521-530	3	7	0.87%
511-520	4	4	0.50%

To fill in column "*D*" use the following formula:

$$\boldsymbol{D} = \frac{\boldsymbol{C}}{\textit{total \# of scores}} \times 100\%.$$

The cumulative percentile computed is for the *upper* score in the range in column "*A*."

Problems

1. An insurance company wishes to estimate the speed of vehicles travelling on the Broken Arrow Expressway in Tulsa. The following data were gathered on 83 randomly selected

vehicles:

36	54	61	68
40	54	61	69
40	55	63	69
40	55	63	69
41	55	63	69
43	55	65	70
45	55	65	70
46	55	65	71
47	55	65	71
47	56	66	72
47	56	66	72
47	57	66	72
48	58	66	72
48	58	67	72
49	58	67	73
50	59	67	72
50	60	67	72
52	60	68	72
52	61	68	72
52	61	68	86
54	61	68	

Attempt to construct a histogram with approximately 11 bars. Why do you wind up with fewer than eleven bars? Is there "bad data" in this sample which should be eliminated?

2. The 83 vehicles in the above sample fell into the following categories:

automobiles	28
pickup trucks	40
semis	14
ambulance	1

Construct a pie chart to present this data. The ambulance was the vehicle travelling at 86 miles per hour and had its emergency lights flashing. Does this change your answer to question one about "bad data?"

3. A researcher records the age (in months) and the weight