Statistics test

निरंजन*
Department of Linguistics
University of Mumbai

PhD coursework examination
Paper II: Quantitative and qualitative methods in linguistics
Term-paper submission 2
February 1, 2024

Question 1

The idea of bell curve relies on the distance from average. The tip of the bell represents the average and the slope represents the dispersion. Usually when the data is collected in natural settings, e.g., rolling dice, flipping coin etc., a range of dispersion is seen. Usually the dispersion is equal on both the positive as well as the negative side of the average. Suppose the average marks scored in a group of students is 50 and one student scores 30 which is lowest. Now, in order to make the representation skewed, another student will have to score at least 90 marks. Because, till 70, all the data-points will fall in the area of bell curve. Also, strictly only a few students will have to score such big score, because if more students score 90 or so, then the average itself will shift and again we will start seeing a bell curve. This is the reason why in most of the natural settings the representation is of a bell.

Question 2

Median:

1. First we order the data-points.
2. Since there are 20 data-points, we divide the sum of 10^{th} and 11^{th} data-points by 2.

3. The median is 80.

Mean:

1. First we take the sum of all the data-points, which is 1745.
2. Since there are 20 data-points, we divide the sum by 20.
3. The mean is 87.25.

Variance:

1. First we subtract the mean from each of the data-points to get the following numbers.

   -50.25, -47.25, -47.25, -44.25, -29.25, -28.25, -24.25, -20.25, -13.25, -10.25, -4.25, -3.25, -1.25, 3.75, 4.75, 9.75, 15.75, 17.75, 38.75, 232.75

2. We square these numbers to get the following set.

   2525.06, 2232.56, 2232.56, 1958.06, 855.56, 798.06, 588.06, 410.06, 175.56, 105.06, 18.06, 10.56, 1.56, 14.06, 22.56, 95.06, 248.06, 315.06, 1501.56, 54172.56

3. We add the set given above, divide it by 19 (since this is a variance of the sample, we use the number of data-points minus 1 ($N - 1$) in the denominator, i.e., 19).
4. The variance is 3593.67.
5. We take the square root of the variance to get the standard deviation.
6. The standard deviation is 59.94.
7. For population, the standard deviation is calculated by dividing by the number of data-points itself (i.e., 20 in this case).
8. The standard deviation for the population is 58.42.

Question 3

1. Firstly, we arrange the marks obtained.

   20, 25, 28, 29, 33, 38, 42, 43.

2. Then create a frequency distribution table for them.

Marks	Number of students	Cumulative frequency
20	𝍸 |	6
25	𝍸 𝍸 𝍸 𝍸	26
28	𝍸 𝍸 𝍸 𝍸 ||||	50
29	𝍸 𝍸 𝍸 𝍸 𝍸 |||	78
33	𝍸 𝍸 𝍸	93
38	||||	97
42	||	99
43	|	100

3. Since the midpoint of the number of students is between 50—51, $\frac{28+29}{2}$ will give us the median.

4. The median is 28.5.

Question 4

1. 53.28% lie between the range 45—60.

2. 93.32% population is below 65.

3. 99.38% population is above 25.

Question 5

While investigating the correlation between two variables, the standard methodology of statistics assumes that there is no correlation between the two variables. This is called the null hypothesis. It is important to notice that this assumption is made without actual investigation and the null hypothesis undergoes verification after actually handling the data. Suppose the data suggests that there is a good correlation between the two variables, the null hypothesis is proven wrong and thus researcher is supposed to reject it.

The diagram presented shows that in total there are 4 possibilities when it comes to verifying the null hypothesis. When the data suggests no correlation, the null hypothesis is true. If the researcher accepts the null hypothesis, they have successfully inferred. If they reject the null hypothesis despite of seeing no correlation in the variables, they commit what is called Type 1 error. When the data shows correlation, the null hypothesis is automatically proven wrong. In this case, if the researcher rejects it, they successfully inferred, but if they still hold on to the null hypothesis, they commit what is called Type 2 error.

Both types of errors can be avoided in different circumstances. I believe, both the errors are results of researcher's bias or negligence. They should be careful while inferring from the data. The researcher may have a pre-investigation bias against the two variables both ways. They may believe that these two variables

are unrelated while they actually aren't or vice versa. If there is negligence in handling the data and inferring from it, it is likely to result in a mistake.

In linguistics too, researchers should be concerned about both the errors. I believe none of these two is even slightly tolerable. Suppose a researcher is investigating the correlation between the use of nominalisers and the speed of speech, none of these two errors would be less serious. One risk, though, is that linguists may commit type-I error more often. The reason is they hypothesise their claims at a very abstract level. Only when they collect and analyse data, they are acquainted with the variance, but till then they may have developed a strong bias towards the correlation because of their work. Maybe they should be more cautious when rejecting the null hypothesis.

Question 6

1. The new student's score *decreases* the average.
2. 73.6 is the new average.
3. The new student's score *increases* the standard deviation.

Question 7

The M in the diagram represents the mean of the data. The bell shape represents the dispersion. The measurement of this dispersion is standard deviation. 64% of the data-points are dispersed in the region from the average to the point till 1 standard deviation on both the sides. If we go ahead, 2 standard deviations incorporate 95% of the data-points. 3 standard deviations incorporate 99.7% of the data-points which means only 0.3% of the population is outside the dispersion. The SE-diff in the given diagram precisely marks that region which has the 0.3% data-points.

Question 8

The diagram shown represents the dispersion of the population. Researchers can only analyse samples. Then they try to predict things for the population. This diagram proposes that there are many such potential samples and many such dispersions in the entire population. Thus the dispersion of our sample lies on a plot of the population mean which has several such potential dispersions. The P-mean is a value which is taken as a reference point and the distance till our mean is measured in terms of standard error. The lesser the standard error is, we can be more confident about our predictions. This is the relevance of the hypothetical P-mean.