International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 6, No. 2 (Nov-2019)

http://dx.doi.org/10.12785/ijcts/060204

Summarizing Test Grades Using Descriptive Statistical Tools

Mohammad Fraiwan Al-Saleh[1]

[1]*Department of Statistics, Yarmouk University, Irbid, Jordan*

Received January 28, 2019, Revised May 29, 2019, Accepted June 21, 2019, Published November 1, 2019

Abstract: In this paper, several Educational Statistical Tools for summarizing a test marks are discussed. The mean, variance, 5-number summary, Difficulty index and Discrimination index are discussed in details. The tools are simple, so that they can be understood by almost all instructors regardless of their backgrounds in statistics. The contents of the paper can be very useful for users of statistics at different areas and in particular, teachers.

Keywords: Descriptive statistics, 5-number summary, Difficulty index, Discrimination index.

1. Introduction and Basic Setup

Summarizing the results of a test is very important for students, **teachers and** decision makers. Students can see their relative standing with respect to their classmates. Teachers can see the overall achievement of their students and the appropriateness of tests as learning assessment. Decision makers (for examples Deans of colleges or heads of Departments) can check for problems in assessments methods used by instructors and can look for any abnormalities in the results of the tests.

Assume that a class consists of *n* students. A test, which consists of a number of questions, is to be given to the students. The questions can be of any kind (True- False, Multiple choices, matching problems, Open-ended, etc.). It is assumed that each problem has a specific number of points out of a total number of points which could be (20, 50, 100, etc.). The allocated number of points for each problem should be known to the students before answering the test. The test should be graded uniformly against a standard (key) solution. Grades should be recorded for each problem and then the total number of points is obtained. It is proper to start grading one problem at a time for all students; this will guarantee a high level of uniformity in the assignments of grades. Once the grade of each student is determined, some summary statistics can be obtained to describe the test results (Data). Some of the measures are to describe the *center* of the data and others to describe the *spread* of the data. Some Indices are used to measure the *appropriateness* of the questions such as *difficulty* and *discrimination* indices. For more details about these measures see Conover (1980), Siegel & Morgan (1996) and Merrens, & Lehmann (1991).

In the next section, several summary measures are discussed; these measures will be illustrated using a real data. In section 3, difficulty and discrimination indices for each question are discussed and illustrated using a real data. Summarizing results of more than one test is addressed in section 4. Suggested future work is mentioned in section 5.

2. Summary Measures of Grades

2.1 Five Number Summary

“Five Number Summary”, as the name implies consists of five numbers: the minimum, the first quartile, the median, the third quartile and the maximum. It is denoted by

$$n(Min, Q_1, M, Q_3, Max)$$

E-mail address: m-saleh@yu.edu.jo

Min: the minimum grade; Q_1 : the minimum grade that exceeds 25% of the grades (Lower Quartile); M : the minimum grade that exceeds 50% of the grades (Median); Q_3 : the minimum grade that exceeds 75% of the grades (Upper Quartile); **Max**: the maximum grade; n : The total number of students; it is not part of the Five-Number Summary:

$$(Min \underset{25\%}{\longleftrightarrow} Q_1 \underset{25\%}{\longleftrightarrow} M \underset{25\%}{\longleftrightarrow} Q_3 \underset{25\%}{\longleftrightarrow} Max\ )$$

The above display shows that the Five-Number Summary divides the students into four groups: Lower, Second, Third and Upper Quarter. In addition, it can be thought of as dividing the students into lower, middle and upper level group.

To find the Five-Number Summary of the data, we follow the following steps:

1. Arrange the grades from smallest to largest;
2. Identify the Min. and Max. grade;
3. Identify the middle grade. Note that if n is odd, then we have a unique middle grade; if n is even then we have two middles; the median is the average of the two middles.
4. As in the previous step, identify the median of the grades that are at or below the median; this is denoted by Q_1;
5. Identify the median of the grades that are at or above the median; this is denoted by Q_3;

<u>Example 1</u>**: The following example is to illustrate the use of the techniques discussed in the paper. The conclusions are only for this data. The analysis of this data can help users to analyze their own data and deduce their own conclusions.**

The following data are the grades of students (in ascending order) (out of **20** points) in the first test of a course in Mathematical Statistics ;(a course that I was teaching at Qatar University in 2006). The total number of students is $n = 23$:

08.0 08.0 10.5 10.5 11.5 12.0 12.5 13.0 13.5 14.5
14.5 15.0 15.0 15.5 15.5 15.5 15.5 15.5 16.0 16.0
17.5 17.5 18.0.

$$Min. = 08.0\ ;\ Q_1 = \frac{12.5+13.0}{2} = 12.75\ ;\ M = 15.0\ ;\ Q_3 = \frac{15.5+15.5}{2} = 15.5\ ;\ Max. = 18.0.$$

Thus, the **<u>Five-Number Summary</u>** is

$$23(8, 12.8, 15, 15.5, 18.0)$$

$$(8.0 \underset{25\%}{\longleftrightarrow} 12.8 \underset{25\%}{\longleftrightarrow} 15.0 \underset{25\%}{\longleftrightarrow} 15.5 \underset{25\%}{\longleftrightarrow} 18.0)$$

Based on this summary we may conclude the following:

1. About 50% of the students got 15 or higher; about 75% of the students got about 13 or higher; about 25% of the students got about 15.5 or higher; about 50% of the students got between 13 or 15.5;

2. If the test is a true measure of the mastery of course objectives, then we may say that about 50% of students fulfilled at least 75% of the course objectives**;** about 50% of students mastered the objectives of the course with about 75% proficiency**.**

3. Since the median is closer to the third quartile than to the first quartile, we may say that the distribution of grades is skewed to the left. This can be shown using the Boxplot given below(Figure 1):

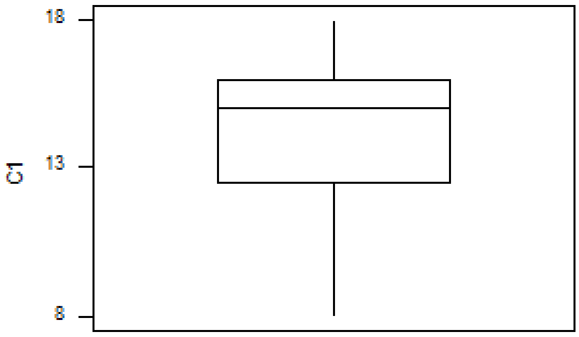

Figure 1. Boxplot of the Data

2.2 Mean μ and Standard Deviation σ

The mean (or average) of the grades is their total divided by the total number of students: If $a_1, a_2, ..., a_n$ represent the grades of n students in a test, then the mean is

$$\bar{a} = \mu = \sum_{i=1}^{n} \frac{a_i}{n}$$

The value of μ is between the **Min**. and **Max.** of the grades; it is regarded as a measure of central tendency (center) of the grades. Reporting the value of the mean alone does not give a clear picture of the distribution of grades. For example, if the mean is μ =**15** then at one extreme, it could be that all grades are equal to 15 and hence their average is 15; on the other extreme it could mean that half of students get **0** grade and half of them get **20**. Thus, there is a need for another measure, which gives us an idea about the spread of the grades. One popular measure of dispersion (variation) is the *standard deviation* **(** σ **)**. It is the square root of average of the squared deviations of the grades from their mean value. In symbols,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (a_i - \mu)^2}{n}} .$$

In this formula, $\mu = \bar{a}$, is actually the value of t that minimizes the quantity

$$\sum_{i=1}^{n} (a_i - t)^2 ,$$

i.e.

$$\min_t \left\{ \sqrt{\frac{\sum_{i=1}^{n} (a_i - t)^2}{n}} \right\} = \sqrt{\frac{\sum_{i=1}^{n} (a_i - \bar{a})^2}{n}} = \sigma .$$

The smallest value of σ is zero, occurs when all grades are equal (to their mean value), i.e. the students form one stratum. The largest value of σ is $(b-a)/2$, occurs when 50% of the grades are equal to **a** & 50% of the grades are equal to **b**, where **a** is the minimum possible grade and **b** is the maximum possible grade; the grades classify the students into two separate strata: the extremely higher level and the extremely lower level. For example if a=0 and b=20 then largest value of σ is 10. (See Al-Saleh and Yusef (2009)). If the shape of the distribution is mound shape (Bell or normal shape) then σ is about $(b-a)/5$ which is equal to **4** when a=0 & b=20.

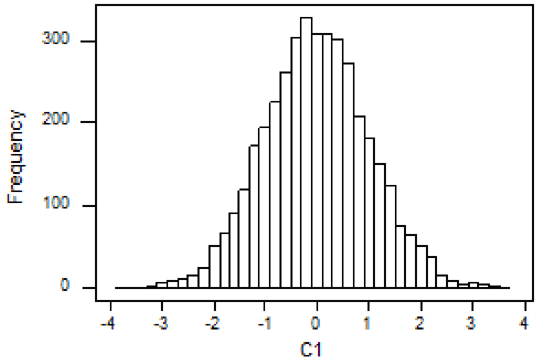

Figure 2. Mound Shape Distribution

Mean and **Standard Deviation** are usually suitable to use, when the distribution of the data is nearly bell-shaped. In this case, we can say that about 68% of the grades are in the interval $(\mu-\sigma,\mu+\sigma)$ and about 95% are in the interval $(\mu-2\sigma,\mu+2\sigma)$. If the shape of the distribution is <u>not a mound</u> shape then one can use **<u>Chebyshev's Inequality</u>**:

For any $k>1$, ***<u>at least</u>*** $\left(1-\frac{1}{k^2}\right)100\%$ *of the grades are in the interval* $(\mu-k\sigma,\mu+k\sigma)$. This percentage is about 56%, 75% and 89% for k=1.5, 2 and 3, respectively.

<u>Example 2</u>:

Using the previous data set of the grades, we have:

$$\mu=\bar{a}=\sum_{i=1}^{23}\frac{a_i}{n}=\frac{319}{23}=13.9\;;\sigma=\sqrt{\frac{\sum_{i=1}^{n}(a_i-\mu)^2}{23}}=2.8.$$

The maximum possible average is 20 and the maximum possible standard deviation is 10. Therefore, the grades of students are suitable with relatively small variability.

2.3 A Measure of Relative Standing

Using μ & σ, one can construct a measure of a relative standing that compares a student grade to the other students in the class or even in another section. Relative standing of a student with grade "a" is defined as follows:

$$RS=\frac{a-\mu}{\sigma}.$$

RS is the distance between the grade of a student and the mean in terms of the standard deviation as a unit of measurement. Based on **Chebyshev's Inequality**, the RS of almost all students is between -5 to 5.

For example, the relative standing for the student with grade 15.5 in the previous example is

$$RS=\frac{15.5-13.9}{2.8}=0.57.$$

So, she is **0.57** standard deviation **<u>above</u>** the average of the class. The relative standing for the student with grade 18 is

$$RS=\frac{18-13.9}{2.8}=1.46.$$

So, she is **1.46,** standard dev. **<u>above</u>** the average of the class. The relative standing for the student with grade 10.5 is

$$RS=\frac{10.5-13.9}{2.8}=-1.21.$$

So, she is **<u>1.21</u>** standard deviation **<u>below</u>** the average of the class.

3. Item Analysis

After giving the test and grade it, item analysis is a suitable thing to do to make sure that the items are appropriate: not too difficult & not too easy, effectively differentiate between students who do well on the overall test and those who do not. Two important measures of item analysis can be used: **<u>Difficulty Index</u>** (DFI) and **<u>Discrimination Index</u>** (DSI).

3.1 Difficulty Index of a Question

The **difficulty index** of a question is a measure of the question's suitability. Neither a very difficult question nor a trivial question is suitable. The difficulty index of a test problems is one important factor to take into account, if the teacher is planning to include this question in future tests. It is time consuming if teachers have to prepare new questions whenever they want to prepare a test; overtime, they should have built a "Test Bank" of good questions to be reused in the future. Difficulty index is one characteristic among others that should be taken into account when preparing such bank.

The difficulty index of a question is obtained based on the proportion of the total points earned by all students for this question from the maximum points that could have been earned for the question:

$$DFI\,(i\,) = \frac{Sum\ of\ \ poi\,\mathrm{nt}s\ earned\ by\ students\ on\ question\ \ \ i}{\max imum\ \ possible\ \ po\,\mathrm{int}s}$$

- The values of DFI ranges from **zero** (extremely difficult) to **one** (extremely easy).
- A rough "rule-of-thumb" (for multiple-choice items) is that if the item DFI is less than 0.25, it is a difficult item; if it is larger than 0.75, it is an easy item. Suitable items are those that has difficulty index between 0.25 and 0.75. This rule can still be used in the case of other types of problems. See: **http://fcit.usf.edu/assessment/selected/responsec.html**

Example 3

Continue with the previous data set, the total number of points is 20 and the total number of students is n=21; table 1 contains the results of the 4 questions. Note that the students are arranged with the top overall scorers at the top of the table:

Table 1. Students Results on the Questions of the Exam & the DFI of Each Question

Qu1(5 pts)	Qu2(6 pts)	Qu3(5 pts)	Qu4(4 pts)	Total
5	6	4	3	18
4.5	6	5	2	17.5
4.5	6	4	3	17.5
4	6	4	2	16
4.5	4.5	5	2	16
4.5	6	4	1	15.5
4	5.5	4	2	15.5
5	5.5	4	1	15.5
4.5	4	5	2	15.5
5	4	5	1	15
4	4	4.5	2	14.5
5	3	4.5	2	14.5
5	4	3.5	2	14.5
4	4	4	2	14
4.5	4	4	1	13.5
4	5	3	1	13
3.5	4	3.5	2	13
4.5	4.5	1.5	2	12.5
2.5	5	3.5	1	12
4.5	3	1	2	10.5
3	4	0	1	8
Total 90	**98**	**77**	**37**	**302**
Max. Possible pts 105	**126**	**105**	**84**	**420**
DFI 0.86	**0.78**	**0.73**	**0.44**	**0.72**

The difficulty index is obtained by dividing the total points obtained by the maximum possible points. It can be seen that **question 1** was the easiest with 86% DFI followed by question 2 and then question 3; the most difficult question was question 4 with 44% DFI. The overall level of difficulty for the test is: **(302/420)=0.72**.

3.2 Discrimination Index of a Question

Another measure, the *Discrimination Index*, refers to how well a question discriminate between high and low level students. The two levels of students are constructed based on their total scores. If the number of students is large enough (about 60 or more) then the upper level group may consist of the highest 27% in the total scores and the lower level consists of the lower 27%. In case the total number of students who took the test is small, then the upper level group consists of the highest 50% in the total scores and the lower level group consists of the lower 50%. The discrimination index of a specific question (DSI) is defined as:

$$DSI = \frac{U - L}{T}$$

where, **U**: Total points earned on the question by the students in the upper level group; **L**: Total earned points on the question by students in the lower level group; **T**: The maximum of total possible points that could have been earned by any of the two groups. A proper question should have a *positive discrimination index* (between 0 & 1), indicating that students who received a high total score did better in the specific item than the students who had a lower overall score. If however, it is found that low-performing students did better in the item than the upper-performing students, then the item has a *negative discrimination index* (between **-1** & **0**).

Discrimination Index can be found using the following steps:

1. Arrange the students with the highest overall scores at the top.
2. For each student, write the score earned for each question.
3. For each question, obtain the total scores earned for all students in the upper level group (top half), U;
4. Also obtain the total scores earned for all students in the lower level group (bottom half), L.
5. Determine the Discrimination Index by subtracting the total of the lower level group from total of the upper level group, U-L and then, dividing by the maximum of total possible points that could have been earned by any of the two groups, T.

A question with negative discrimination index should be carefully reviewed. Ambiguity could be the cause; the question may be wrong, or it has some problem that was not recognized by good students or even by the teacher.

Example 4: Consider again the data of example 3 (Table 2):

Table 2. Students Results on the Questions of the Exam & the DSI of Each Question

Qu1(5pts)	**Qu2(6pts)**	**Qu3(5pts)**	**Qu4(4pts)**	**Total(20)**
5	6	4	3	18
4.5	6	5	2	17.5
4.5	6	4	3	17.5
4	6	4	2	16
4.5	4.5	5	2	16
4.5	6	4	1	15.5
4	5.5	4	2	15.5
5	5.5	4	1	15.5
4.5	4	5	2	15.5
5	4	5	1	15
4	4	4.5	2	14.5
Total 49.5	**57.5**	**48.5**	**21**	**176.5**
4	4	4.5	2	14.5
5	3	4.5	2	14.5
5	4	3.5	2	14.5
4	4	4	2	14
4.5	4	4	1	13.5
4	5	3	1	13
3.5	4	3.5	2	13
4.5	4.5	1.5	2	12.5
2.5	5	3.5	1	12
4.5	3	1	2	10.5

3	4	0	1	8
Total 44.5	**44.5**	**33**	**18**	**140**
DSI 0.09	**0.20**	**0.28**	**0.07**	**0.17**

For example, for Question 1,

$$DSI = \frac{49.5 - 44.5}{11 \times 5} = 0.09$$

Usually easy questions have small DSI (close to zero) while difficult questions have large DSI; Question 4 may has some ambiguity because it is little bit difficult but has low level of discrimination. (This question was a multiple-choice problem)

Table 3. DFI and DSI of the Test Questions

Question	DFI	DSI
1	0.86	0.09
2	0.78	0.20
3	0.73	0.28
4	0.44	0.07
Overall	**0.72**	**0.16**

4. Pearson Correlation Coefficient

Pearson Correlation coefficient (ρ) is a measure of the linear relation between two variables. It can be used to study the relation between two tests for the same students at two different times. Its value is between -1 and 1. The closer the value to 1, the stronger is the relation between the two variables in the positive direction, i.e. students who get high grades in the first test get also high grades in the second test and vice versa. Negative value means that the two variables are negatively associated. A value close to zero indicates that the two tests are unrelated. If the two tests are valid measures of students' ability, then the correlation tends to be high between the two tests.

Assume that the grades of the first and second test are (a_i, b_i), $i = 1, \ldots, n$, then the correlation coefficient is given by

$$\rho = \frac{\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2 \sum_{i=1}^{n} (b_i - \bar{b})^2}}$$

<u>Example 5</u>: The grades of two tests in a stat. course were as follows (Table 4), their summaries are given in Table 5.

Table 4. The Grades of Test 1 & Test 2

Test 1	Test 2
16.0	**16.0**
15.0	**11.0**
16.0	**15.5**
15.5	**16.5**
16.0	**12.0**
15.5	**17.5**
12.0	**6.0**
17.5	**16.0**
10.5	**12.5**
13.5	**14.5**
15.5	**6.5**
08.0	**7.0**
13.0	**8.5**
14.5	**12.5**

14.5	**9.5**
12.5	**14.0**
15.5	**14.5**
18.0	**12.5**
15.5	**14.5**
17.5	**14.5**
10.5	**7.5**
15.0	**19.5**
11.5	**6.5**

Table 5. Summaries of Test 1 & Test 2

Variable	N	Mean	Median	St. De.	Min.	Max.	Q1	Q3
Test 1	**23**	**14.3**	**15.0**	**2.5**	**8.0**	**18.0**	**12.0**	**16.0**
Test 2	**23**	**12.4**	**12.5**	**3.9**	**6.0**	**19.5**	**8.5**	**15.5**

- Five number summary for test 1 is :

$$23\left(8.0,12.0,15.0,16.0,18.0\right)$$

- Five number summary for test 2 is

$$23\left(6.0,8.5,12.5,15.5,19.5\right)$$

$$\sum_{i=1}^{23}(a_i-\bar{a})(b_i-\bar{b})=126.2;\sum_{i=1}^{23}(a_i-\bar{a})^2=138.4;\sum_{i=1}^{23}(b_i-\bar{b})^2=336.6$$

- $\rho=\dfrac{126.2}{\sqrt{(138.4)(336.6)}}=0.585$.
- The correlation between the two tests is positive and the value is more than 0.5, indicating that those who got high grades in one test tend to do well in the second test and vice versa.
- Thus, there is a suitable level of validity of the tests in measuring students' actual achievement, suggested by this moderate value of ρ .

Note: The data discussed above is a population and 0.585 is the actual value of ρ . If the data is a random sample from a population, then the value r of the sample correlation need to be tested for significant.

It is not easy to judge whether the distribution of final grades of a course in one semester is normal or abnormal, as each course, and students of a course may have different properties. One possible method to judge on this issue is to relay on all previous information to establish what is called Reference curve, or reference distribution. This is something similar to reference ranges of Blood Pressure, BMI (Body mass index), etc.

http://www.lotrel.com/info/answers/high_blood_pressure.jsp

http://en.wikipedia.org/wiki/Body_mass_index

It is hoped for something similar to that for each course, so that the teacher, the chairperson and anyone interested in such thing can with some simple calculations see if there is any abnormality of a course grades distribution. See Al-Saleh et al. (2010).

ACKNOWLEDGMENT

The author is very thankful to the referee for the comments and the editor of the journal for his efforts.

REFERENCES

[1] M. Fraiwan Al-Saleh, Dareen Ali & Laila Dahshal (2010). Toward a Reference Curve for the Grades of Each Course. International Journal of Mathematical Education in Science and Technology 41, 547-555.

[2] M. Fraiwan Al-Saleh and Adel Yusef (2009). Properties of the Standard Deviation that are Rarely Mentioned in Classrooms. Austrian Journal of Statistics 38, 193--202.

[3] W. J. Conover (1980). Practical Nonparametric Statistics, 2nd edition, John Wiley and Sons, New York.

[4] F. Siegel and J. Morgan (1996). Statistics and Data Analysis, 2nd edition, John Wiley and Sons, New York.

[5] A. Merrens, and J. Lehmann, (1991). Measurement and Evaluation in Education and Psychology, 2nd edition, Halt, Rinhart, Winston.

[6] http://www.lotrel.com/info/answers/high_blood_pressure.jsp?

[7] http://en.wikipedia.org/wiki/Body_mass_index

[8] http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html