

<Answers Version 1 – White>

1) Use the Excel sheet labeled “fish_ver1_white” to answer the following questions. This dataset presents a random sample of fish caught in Hosu Lake. In the sheet, values for “SPECIES”, “LENGTH” and “WEIGHT” are recorded for 157 fish of 7 species. Read the data description in the Excel sheet carefully. (23 pts. plus Extra credit 5 pts.)

A) When you describe the central tendency of “SPECIES”, (MODE) is the most desirable measure because “SPECIES” is a(n) (NOMINAL) variable. The value which the measure takes is (PERCH). (5 pts)

Since “SPECIES” is a **nominal** variable, the **MODE** is the only measure of center.

The Excel command (=Mode(Data Array)) produces the value of 7.

Since 7 represents “Perch”, the mode is “Perch”.

2 points for mentioning “Mode” in the first parenthesis

2 points for mentioning “Nominal” variable in the second parenthesis

Note: Although “qualitative (or categorical)” is not a correct answer, give a partial credit (1 point). Even though students write down “nominal” in the third parenthesis, give 2 points. The same rule applies to “qualitative (or categorical)”, i.e. 1 point.

1 point if a student presents “Perch” as her/his conclusion. Treat “7” as correct.

B) Determine the shape of the distribution for the variable “WEIGHT”. Provide numerical justification. (5 pts.)

Numerical value(s):

“The mean of “WEIGHT” (401.23) is greater than the median (273)” or

“Pearson’s 2nd Skewness is 1.07 ($=3*(\text{mean}-\text{median})/\text{SD}$) which is positive.”

Conclusion:

right-skewed (positively skewed)

2 points, if a student uses “comparison of mean and median” or “Pearson’s 2nd Skewness”.

2 points, if the above intermediate value(s) are correct.

1 point, if the conclusion is correct (right-skewness).

Note: If a student shows some efforts to answer this question without mentioning skewness, give one point in total.

C) Between “WEIGHT” and “LENGTH”, which variable is relatively more homogeneous? Provide numerical justification. (5 pts.)

(CV) of “WEIGHT” = $100 \times 358.8162 / 401.2287 = 89.43$ (%)

(CV) of “LENGTH” = $100 \times 11.7678 / 31.1783 = 37.74$ (%)

Conclusion: (LENGTH) is relatively more homogeneous.

Since the CV of LENGTH is less than that of WEIGHT, we can conclude that **LENGTH is relatively more homogeneous.** (Or **WEIGHT is relatively more dispersed.**)

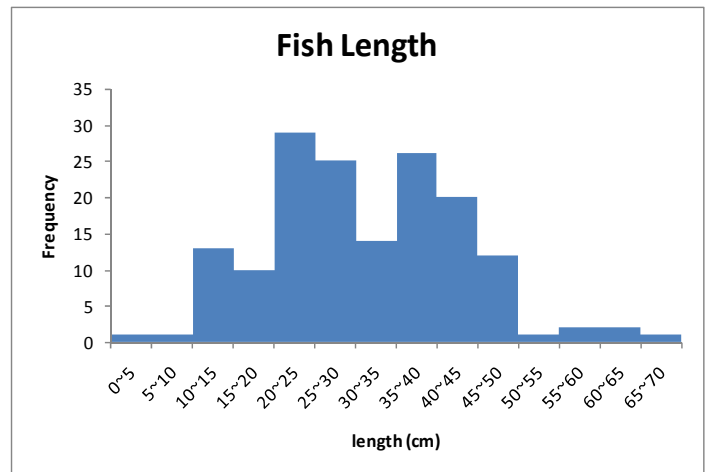
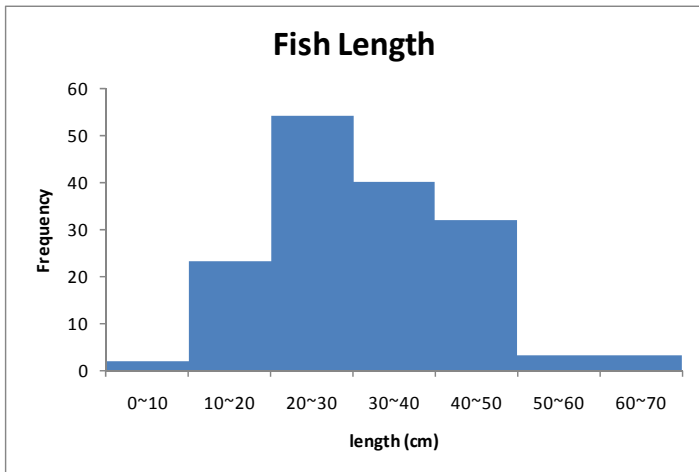
1 point for using CV (coefficient of variation)

1 point if the CV of WEIGHT is correct.

1 point if the CV of LENGTH is correct.

2 points for the correct conclusion

D) The histogram below is plotted using the variable “LENGTH”. Now plot another histogram for the same variable “LENGTH”. Use the same lower limit of the first class, which is zero, however, make the class interval 5. (1) Report the frequency values for 5th, 6th, 7th, and 8th classes. (2) Describe the major difference in the two plots, paying particular attention to revealed characteristics of the distribution. (8 pts.) Explain why the data would have the distribution you see in the graph which you plotted. (Extra credit 5 pts.)



(1) Frequency values (4 pts.)

5th: 29, 6th: 25, 7th: 14, 8th: 26

1 point for each correct response

(2) Major difference (4 pts.)

First of all, choosing the number of bins and class limits requires our judgment. The judgment may affect our interpretation of a distribution. 7-bin histogram shows some concentration between 20 and 50, but it may be a little coarse for giving a precise perception of distribution. This point will be clearer when we look at 14-bin histogram. There may be many possible (and legitimate) interpretations for this result. Among them, the characteristic which only the second histogram conceives is that the data appears to have a multi-modal distribution. To put it differently, the second histogram shows the possible existence of heterogeneity of the data.

A bimodal or multimodal distribution occurs when dissimilar populations are combined into one sample. In other words, multimodal distribution most commonly arises as a mixture of two or more different uni-modal distributions. Since 7 different species of fishes are combined into one sample, we can expect that heterogeneity may exist.

1 point in total for any try not belonging to the below categories (if a student mentions only 'skewness', she/he will get 1 point. We cannot say that there is a discernable distinction related skewness from the revealed distribution.)

4 points in total for simply providing the above correct points (multi-modal (or bimodal) distribution, locally multiple peaks, etc)

<For extra credits>

1 point in total for any try not belonging to the below categories

2 points in total for a response showing that "more bins imply more (frequency) information" in this case (although this statement is not always true.)

5 points in total for a response showing that observations are sampled from populations with different categories (in this case, species) or that heterogeneity exists in a sample due to different species of fish.

Name _____ Team Number _____

2) Use the Excel sheet labeled “software_ver1_white” to answer the following questions. This data presents a random sample of software projects completed by DaumAgora Inc . In the sheet, values for “TEAM”, “WORKHOUR”, “FP” and “VC” are recorded for 104 projects. Read the data description carefully. **(17 points)**

A) Among “TEAM”, “WORKHOUR” and “FP”, which has the strongest linear relationship with “VC”? Provide Numerical justification. **(6 pts.)**

Numerical values:

correlation between TEAM and VC (0.79)

correlation between WORKHOUR and VC (0.93)

correlation between FP and VC (0.75)

Conclusion: (WORKHOUR) has the strongest linear relationship with “VC”

2 point for using correlation

1 point for the correct correlation between TEAM and VC (0.79)

1 point for the correct correlation between WORKHOUR and VC (0.93)

1 point for the correct correlation between FP and VC (0.75)

1 point for the correct conclusion (WORKHOUR)

Note: Please ignore minor differences resulted from using different decimal points.

B) If the Chief Technology Officer is interested in predicting **variable costs (“VC”)** of a certain software development project given **function points (“FP”)**, what do you suggest he/she do? (1) Report the numerical values of your estimation. (2) If a specific project has 10,000 function points, what is your prediction of the variable cost? **(11 pts.)**

(1) Estimation result (6 pts.):

Expected VC = $0.0664 * FP + 892.27$

3 points for trying to derive a least squares line (trend line, line of best fit, etc)

3 points for the correct slope and intercept (the slope of “0.07” is also correct in this case)

Note: As long as values for slope and intercept are correct, give points regardless of the use of general terms for variables (i.e. X and Y instead of FP and VC)

(2) Prediction (5 pts.):

If a specific project has 10,000 function points, the variable is predicted (estimated, expected or on average) to be 1556.27 hundred dollars ($=0.0664 * 10000 + 892.27$) or 155,627 dollars.

2 points in total for trying to derive a predicted value using a least squares line (regardless of the correctness of the least squares line)

5 points for the correct predicted value of VC given 10000 function points [1 point off for the wrong use of unit of measurements]

Name _____ Team Number _____

3) A student is taking a multiple-choice quiz in which each question has five choices. Suppose that she has no clue of the correct answers to any of the questions, and thus decides on a strategy in which she will place five balls (marked A, B, C, D and E) into a box. She randomly selects one ball for each question and puts the ball back in the box. The marking on the ball selected will determine her answer to the question. **(20 pts)**

A) If there are 10 questions of five options in the exam, what is the probability that she will get:

For the following four subquestions, 2 pts for the right command and 1pt for the right final numerical solution.

(1) five questions correct? **(3 pts.)**

Excel Formula: [BINOM.DIST\(5,10,0.2,0\)](#)

Answer: [0.0264 \(or 0.03\)](#)

(2) at least four questions correct? **(3 pts.)**

Excel Formula: [1-BINOM.DIST\(3,10,0.2,1\)](#)

Answer: [0.1209 \(or 0.12\)](#)

(3) no more than two questions correct? **(3 pts.)**

Excel Formula: [BINOM.DIST\(2,10,0.2,1\)](#)

Answer: [0.6778 \(or 0.68\)](#)

(4) more than four but less than eight questions correct? **(3 pts.)**

Excel Formula: [BINOM.DIST\(7,10,0.2,1\)-BINOM.DIST\(4,10,0.2,1\)](#)

Answer: [0.0327 \(or 0.03\)](#)

B) Describe briefly two assumptions necessary to use the above distribution? **(2 pts.)**

That each outcome is independent of any other outcome, and that all balls are identical with the exception of the letter marked on the ball.

1 pt for the independence; 1 pt for the identical point.

Name _____ Team Number _____

C) What are the mean and the standard deviation of the number of questions that she will get correct? **(6 pts.)**

(1) Mean (3 pts.): $=n \cdot \pi = 10 \cdot 0.2 = 2$

3pts for the right final number; 1 pt for the right formula but wrong number

(2) Standard deviation (3 pts.): $=\sqrt{n \cdot \pi \cdot (1-\pi)} = \sqrt{10 \cdot 0.2 \cdot 0.8} = 1.2649$ or (1.26)

3pts for the right final number; 1 pt for the right formula but wrong number

4) You have sample data for four stocks on the Excel sheet labeled "**stock_ver1_white.**" Answer the following questions. **(15 pts.)**

A) Estimate the expected returns and variances for the following simple portfolios. **(10 pts.)**

Portfolio A: Stock 1 - 40%, Stock 2 - 60%,

Portfolio B: Stock 3 - 30%, Stock 4 - 70%.

	Portfolio A	Portfolio B
Expected Return	-0.0031	0.0392
Variance	0.0181	0.0279

Note: Include at least 4 decimal places when you write down your final answers.

B) If the objective of the investors is to maximize their expected return (profitability), then which portfolio would you recommend? **(2 pts.)**

"B" since it has higher mean.

2 pts. for the right choice

C) If the objective of the investors is to minimize their expected risks (variance), then which portfolio would you recommend? **(3 pts.)**

"A" since it has smaller variance.

3pts. for the right choice