

Introduction to Data Analysis

Suggested Answers

Uncertainty - Problem 1

Task 1. Answers will vary. In a typical class it is likely that some students will estimate length and width to the nearest ± 1 mm and that others will estimate them to the nearest ± 0.1 mm. The width of the rectangle's border is sufficient that students may have different results due to individual decisions about whether to measure from the outside edge, the inside edge, or from the border's middle. Taken together, there is an opportunity to engage students in a discussion of determinate and indeterminate errors.

Task 2. Answers will vary. In a typical class it is likely that some students will pay attention to significant figures and that others will not do so.

Task 3. Answers will vary. The goal here is for a student to begin considering how the precision of his or her measurements affects the precision which he or she can report the results of a calculation.

Uncertainty - Problem 2

Task 1. Students should recognize that this is a source of determinate error.

Task 2. For a length and width of 154 mm and 32 mm, students should report the area as $4.9 \times 10^3 \text{ mm}^2$. For a length and width of 154.3 mm and 31.8 mm, students should report the area as $4.91 \times 10^2 \text{ mm}^2$.

Uncertainty - Problem 3

Task 1. The mass of water is 0.1276 g. The percent water is 25.57%. Each answer has four significant figures.

Task 2. The exact answer for the percent water is 25.56601 (to five decimal places). The largest possible mass of water is

$$0.4992 \text{ g} - 0.3714 \text{ g} = 0.1278 \text{ g}.$$

The smallest possible mass of water is

$$0.4990 \text{ g} - 0.3716 \text{ g} = 0.1274 \text{ g}.$$

The largest possible percent water is

$$100 * (0.1278 \text{ g}/0.4990 \text{ g}) = 25.6112\%.$$

The smallest possible percent water is

$$100 * (0.1274 \text{ g}/0.4992 \text{ g}) = 25.5208\%.$$

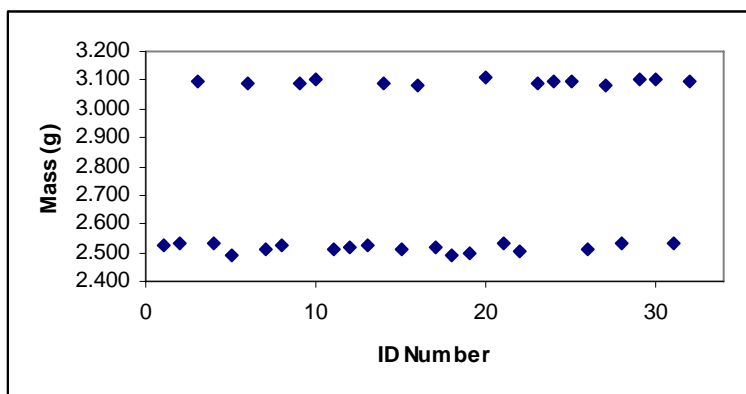
Comparing the two worst case results and the exact result, we see that rounding to the hundredth's place is the first instance where there is no agreement between the three results. The result of the exact calculation, therefore, is rounded to the hundredth's place, giving 25.57%, or four significant figures, which agrees with the results from Task 1.

The section on Further Study introduces an example where the worst case approach to uncertainty suggests one more significant figure than suggested by significant figures. If so desired, this provides an opportunity to discuss the concept of the propagation of uncertainty.

Preliminary Analysis of Data - Problem 1

Task 1. The mean and standard deviation are 2.769 g and 0.291 g, respectively. Observant students will note that none of the pennies has a mass near the mean, find this odd and suggest that the mean is not very representative. Less observant students will consider the mean to represent the best measure of the mass for a typical penny.

Task 2. A typical plot is shown here



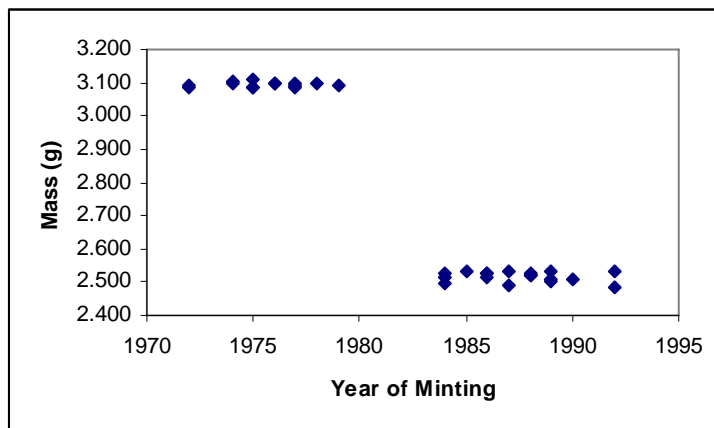
Although both the raw data and the plot show the same information, students are more likely to notice that no single penny has a mass similar to the calculated mean. Most students will begin to consider that the pennies come from two different populations.

Task 3. With a more direct question, all students should reach the conclusion that the data set consists of samples drawn from two populations. Suggestions for the difference between the populations will vary.

Task 4. Students should recognize that the balance's limit is ± 0.001 g and that this much smaller than the standard deviation of 0.291 g.

Preliminary Analysis of Data - Problem 2

Task 1. A typical plot is shown here:



Students now can readily see that the two populations are segregated by the year of minting. Explanations may vary, but some students will probably suggest a change in the penny's composition.

Task 2. The mean and standard deviation for the pennies minted before 1980 are 3.094 g and 0.007 g, respectively, and the mean and standard deviation for pennies minted after 1980 are 2.517 g and 0.014 g, respectively. Students should recognize that dividing the pennies into two populations explains most, but not all of the uncertainty in the results. Possible explanations for uncertainties larger than that for the balance used to measure the mass include uncertainty due to manufacturing and "wear and tear."

Task 3. Students should note that information is missing for the years 1980 – 1983. They might also suggest collecting more pennies from each year and plotting their respective averages in place of the individual masses.

Preliminary Analysis of Data - Problem 3

Task 1. Students should note that the mean mass of pennies from 1977 – 1981 are similar, that the mean mass of pennies from 1983 – 1987 are similar, and that both are different than the mean mass of pennies from 1982. Explanations will vary.

Task 2. The unusually large standard deviation for the 1982 pennies is similar to that seen earlier when the pennies were treated as a single population. Most students will suggest that the change in the penny's composition must have occurred in 1982 and that pennies of both types were produced.

Task 3. The data in this spreadsheet should convince all students that the penny underwent a change in composition in 1982.

Comparing Data Sets - Problem 1

Task 1. The three data points are well separated and students can easily see that the t-test predicts that the data are from two populations.

Task 2. In this case the two samples usually are found to be different at the 90% CL, but not at the 95% or 99% CL. Adding one additional point to each data set is usually sufficient to show a difference between the two samples at the 99% CL.

Task 3. In this case the two samples are not found to be different at the 90% CL. Adding 4 – 5 points to each data set usually leads to a significant difference at the 95% CL. It may take as many as 10+ total points in each data set to achieve means that are significantly different at the 99% CL.

Task 4. The interplay of the difference between the means, the relative standard deviations and the number of samples on the discriminating power of the t-test is evident.

Comparing Data Sets - Problem 2

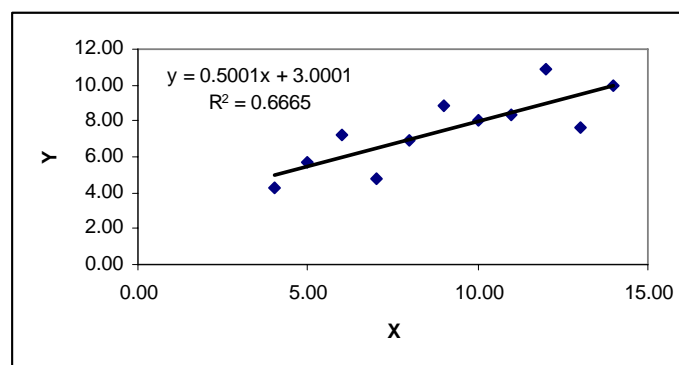
Task 1. The default data are different at the 90% CL, but not at the 95% or 99% CL. Adding the additional data changes the results so that the difference is now significant at the 95% CL.

Task 2. The pennies from 1978 and 1980 are not different at the 90% CL. The pennies from 1978 and 1984, however, clearly are different at the 99% CL.

Linear Regression - Problem 1

Task 1. The mean and standard deviation are 9.00 and 7.50, respectively.

Task 2. A typical plot is shown here;



Most students will accept that the data are explained by a linear model, although many will find the value for R^2 suspicious. Here it helps to remind students that not all data sets have R^2 values of >0.95 .

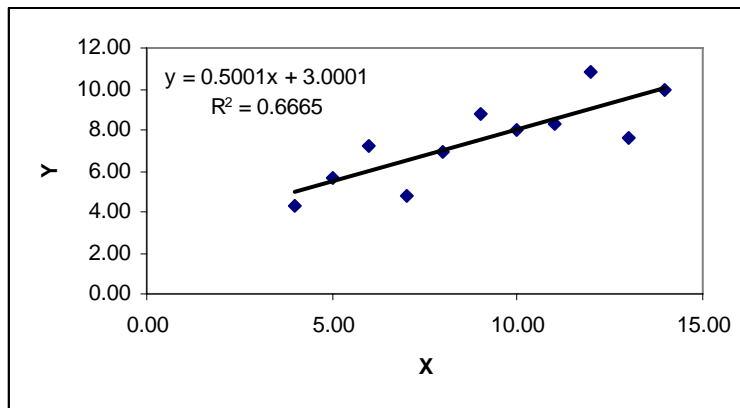
Linear Regression - Problem 2

Task 1. Because the X values are the same for all three data sets, the students need only examine the Y values. Most students will conclude that there is a lot of variation in the value of Y for any value of X, but it is unlikely that they will observe a discernable pattern.

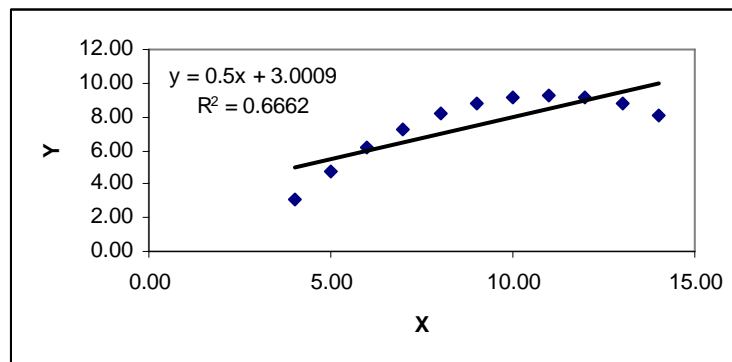
Task 2. Knowing this additional information, most students will conclude that the data are explained equally well by the same model.

Linear Regression - Problem 3

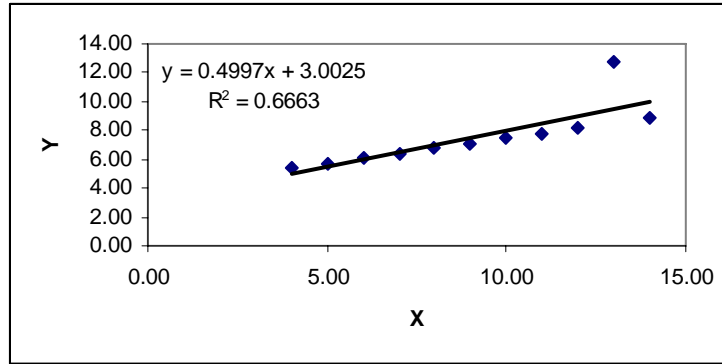
Task 1. Typical plots are shown here for Data Set 1



for Data Set 2

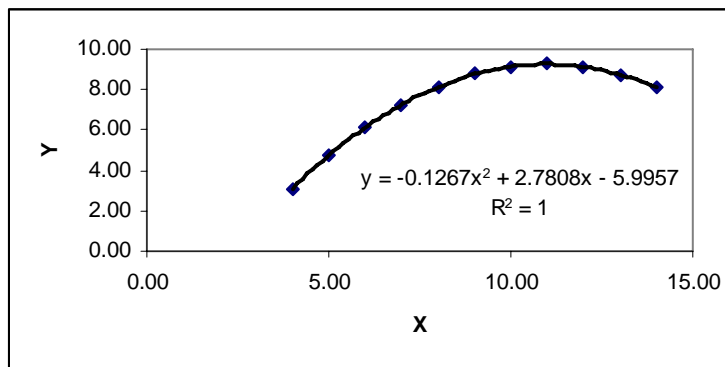


and for Data Set 3

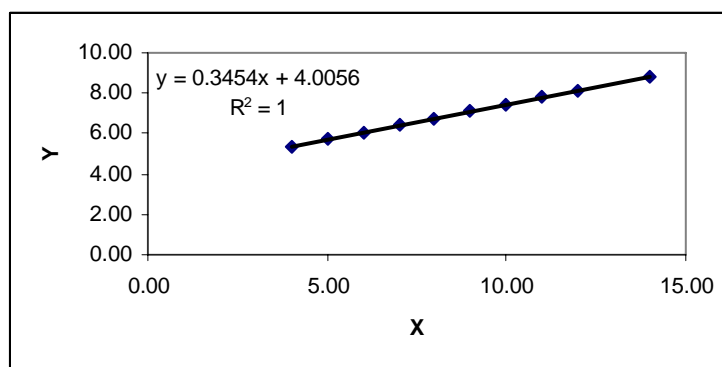


Students will easily conclude that a linear model is inappropriate for the last two data sets and recognize the importance of visually examining data.

Task 2. Students should easily find that a quadratic equation fits the data, as shown here:



and that removing the data point X = 13.00 and Y = 12.74 produces the following plot:



Outliers - Problem 1

Task 1. In the original configuration there is insufficient information to reject the potential outlier at any of the provided confidence levels. Adding several additional points generally leads to the conclusion that the suspect data point is an outlier. Students

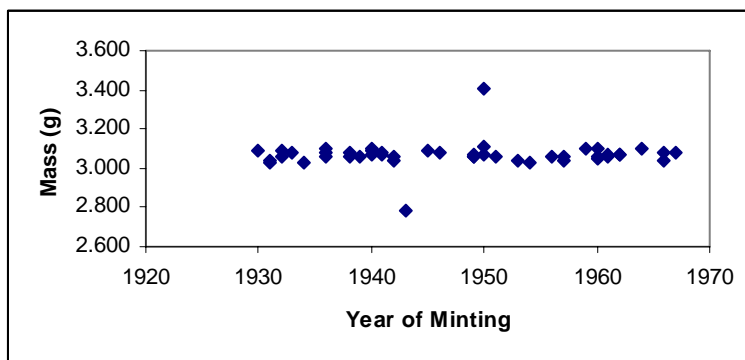
should appreciate that it always is a good idea to run additional experiments when a value is in question.

Task 2. In the original configuration there usually is sufficient information to reject the suspect data point at the 90% CL (or greater). Adding an additional point at the opposite end makes it impossible to find that the suspect point is an outlier. Students should appreciate that poor precision makes it harder to evaluate suspect data points.

Task 3. Students can easily see the relative importance of the gap and the range in determining the success of a Q-test.

Outliers - Problem 2

Task 1. A typical plot is shown here:



Students should identify the 1943 penny and one of the pennies from 1950 as possible outliers.

Task 2. The web site provides a discussion of the 1943 steel penny, which students easily recognize as a clear outlier. With no clear explanation for the strange 1950 penny, students might suggest examining it closely to see if there is an obvious reason for its larger than expected mass.

Outliers - Problem 3

Task 1. The gap is $9.9 - 6.2$, or 3.7 . The range is $9.9 - 4.2$, or 5.7 . The value of Q , therefore, is 0.649 , which is smaller than a Q_{crit} of 0.680 but larger than a Q_{crit} of 0.568 ; thus, the result can be treated as an outlier at the 95% CL, but not at the 99% CL.