

Sample Problems
Qualifying Exam for B01.1305
Statistics and Data Analysis

The sample problems listed here are similar in spirit and difficulty to those on the actual qualifying exam. Solutions appear on page 14. This set of sample problems is slightly longer than the actual qualifying exam.

Revision data MAY 2004

S1. You have just calculated the average and standard deviation of a list of 101 numbers, finding an average of 240.0 and a standard deviation of 25.88.

Unfortunately, a check of the list of numbers uncovers two errors.

A number originally listed as 230 should be 200.

A number originally listed as 250 should be 280.

Based on this information, you decide

- (a) It will be impossible now to give either the correct value of the mean or the correct value of the standard deviation.
- (b) The mean will remain at 240, and the standard deviation will increase.
- (c) The mean will remain at 240, and the standard deviation will decrease.
- (d) The mean will change, but the value of the standard deviation will remain unchanged.
- (e) The mean and standard deviation will remain at 240 and 25.88, respectively.

S2. Each of the following questions can be answered as *true* or *false*. Indicate your solutions on the blank spaces below. Your answers do not have to be explained.

- ____(a) The median of a set of numbers is always larger than the mean.
- ____(b) The upper quartile of a set of numbers is greater than or equal to the median of the set of numbers.
- ____(c) The interquartile range of a set of numbers must be at least as large as the median of the set of numbers.
- ____(d) Negative values of the standard deviation indicate that the set of values is even less dispersed than would be expected by chance alone.
- ____(e) The standard deviation is commonly used because of its ease of calculation using only paper and pencil.

- _____ (f) If the median salary of all male employees in a firm exceeds \$26,000, and if the median salary of all female employees in the firm exceeds \$25,000, then it is certain that the median salary of all employees exceeds \$25,000.
- _____ (g) It happens that the average pay of a floor supervisor exceeds the average pay of a line worker in factory Q, and the same statement also holds in factory R. If the data for the two factories are combined, then the average pay of all floor supervisors is guaranteed to exceed the average pay of all line workers.

S3. For each of the following situations, indicate whether list P or list Q has the smaller sample standard deviation. You need not justify your solution, nor do you need to actually compute any standard deviations.

- (a) List P has 100 values. The values are the integers 1, 2, 3, ..., 99, 100.
List Q has 100 values. The values are 2, 4, 6, 8, ..., 198, 200.
- (b) List P has 50 values.
The number 180 appears 10 times.
The number 200 appears 30 times.
The number 220 appears 10 times.
List Q has 50 values.
The number 180 appears 5 times.
The number 200 appears 40 times.
The number 220 appears 5 times.
- (c) List P contains 500 numbers.
The value 18 appears 100 times.
The value 19 appears 100 times.
The value 20 appears 100 times.
The value 21 appears 100 times.
The value 22 appears 100 times.
List Q contains 500 numbers.
The value 18 appears 150 times.
The value 20 appears 200 times.
The value 22 appears 150 times.
- (d) List P contains the 1,000 integers 1, 2, 3, ..., 1,000.
List Q contains 2,000 integers; these are 1, 1, 2, 2, 3, 3, 4, 4, ..., 999, 999, 1,000, 1,000.

S4. In each of the situations below, a set of data is described. Circle the value which, you believe, could be the standard deviation.

(a) Ms. Rivera's third grade class, containing 29 students, was asked about daily milk consumption. The standard deviation of this set of 29 values could be

16 oz 32 oz 64 oz 128 oz
480 cc 960 cc 1.9 ℓ 3.8 ℓ

Approximate metric equivalents are shown. You might remember that 32 oz = 1 quart.

(b) An auditor examined the numbers of cars handled by individual toll booth attendants at the Triborough Bridge during rush hours. The standard deviation of the cars-per-hour for these attendants could be

3.5 35 350 3,500

(c) The standard deviation of the heights in inches of the female members of the Stern MBA class of 1999 could be

0.3 inch 3 inches 9 inches 62 inches

(d) An accounting firm took a survey of its entry-level clerk-typists who have rental apartments in Manhattan. The standard deviation of the monthly rents could be

\$20 \$200 \$2,000 \$8,000

(e) For the purchase amounts by customers at the first-floor snack area of the K-MEC building, the standard deviation could be

12 cents 25 cents \$1.20 \$12.00

NOTE: The items sold here are coffee (50 cents), soft drinks (about \$1), sandwiches (about \$3), muffins and pastries (about \$1).

S5. Carolyn and Lou work in a laboratory which assesses the presence of environmental toxins in food. Thursday's task required that they measure the PCB content in the bluefish brought to a certain wholesale fish market. They were given a sample of 25 fish. These measurements were difficult, and they had processed only 18 of the 25 fish by 5:00 p.m. Carolyn had an appointment for dinner at 6:00, and Lou graciously agreed to finish the job. Carolyn, however, did take the 18 measurements home with her, and later in the evening she computed a 95% confidence interval for the population mean PCB content, based on her 18 values. Lou stayed late at the lab, and he completed the work, getting all 25 values. He also computed a 95% confidence interval for the population mean.

Please answer the following questions as *true* or *false*.

- (a) Carolyn's confidence interval will certainly be longer than Lou's.
- (b) Lou should use the t table with 25 degrees of freedom in constructing his confidence interval.
- (c) It is possible that one of their intervals covered the true population mean while the other did not.
- (d) The probability that both intervals will cover the true population mean is $0.95 \times 0.95 = 0.9025$.
- (e) Lou's confidence interval is more likely to cover the true population mean.

S6. For each of the following situation circle either the response **COULD HAPPEN** or the response **IMPOSSIBLE**. For example, " $\bar{x} = -19.4$ " gets the response **COULD HAPPEN** while " $x - 1 = x + 1$ " is **IMPOSSIBLE**. In dealing with these situations, you should assume that the arithmetic is always done correctly.

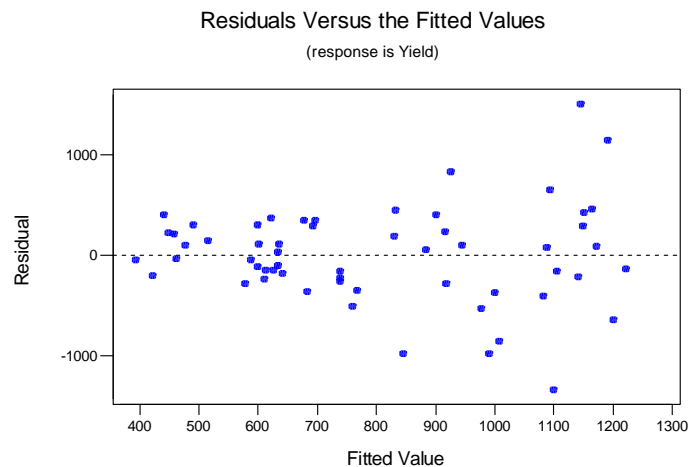
- (a) Sherry had a set of data of size $n = 6$ in which the mean and median were identical.
- (b) In comparing two samples of continuous data, Miles assumed that $\sigma_x = \sigma_y$ while Henrietta allowed $\sigma_x \neq \sigma_y$. Miles accepted the null hypothesis $H_0: \mu_x = \mu_y$ while Henrietta, using the same α , rejected the same null hypothesis.
- (c) Based on a sample of 38 yes-or-no values, the estimate \hat{p} was found to be 0.41.
- (d) In a single sample $\bar{x} = 141.2$ and $s = -32.1$.
- (e) In a regression of Y on X , we computed $s_Y = 21.04$ and $s_e = 21.52$. (Note: s_Y is the standard deviation of Y , while s_e is the standard error of regression, the root-mean-square residual.)
- (f) In a very strong multiple regression, the F statistic was found as $F = 12,810$.
- (g) Based on a sample of $n = 131$ yes-or-no values, the 95% confidence interval for the binomial parameter p was found to be 0.711 ± 0.083 . Also, the null hypothesis $H_0: p = 0.61$ (versus alternative $H_1: p \neq 0.61$) was rejected at the 5% level of significance.
- (h) The hypothesis $H_0: \mu = 1.8$ regarding the mean of a continuous population was tested against alternative $H_1: \mu \neq 1.8$ at the 0.05 level of significance, using a sample of size $n = 85$. Unknown to the statistical analyst, the true value of μ was 1.74 and yet H_0 was still accepted.

S7. Following a regression of annual salary (five-years after hiring) on academic qualifications, it is possible to make a prediction for a new hire, Michael Jamison, who has just obtained his MBA. This prediction results in the 95% prediction interval $\$80,000 \pm \$22,000$. Please answer the following as *true* or *false*.

- (a) Michael will certainly be making more than \$58,000 after five years.
- (b) The best prediction for Michael's annual salary after five years is \$80,000.
- (c) If the MBA is coded as 18 years of education, then we are 95% confident that the value of $\beta_0 + \beta_1 \times 18$ is in the interval $\$80,000 \pm \$22,000$. (In the relevant regression model, the equation of the line is $\text{SALARY} = \beta_0 + \beta_1 \text{YRS_EDUC}$.)
- (d) We can predict with 95% certainty that Michael's salary in five years will be between \$58,000 and \$102,000.

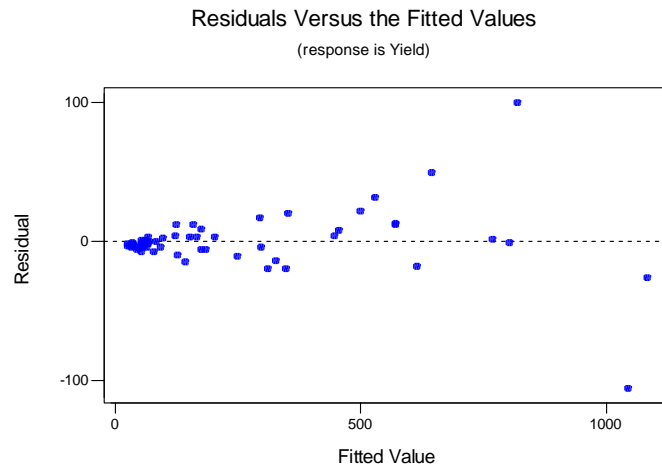
S8. Consider the regression of dependent variable *Yield* on independent variable *Temperature*. Consider the following possible situations for the residual versus fitted plot. Indicate the action that you would take.

(i)



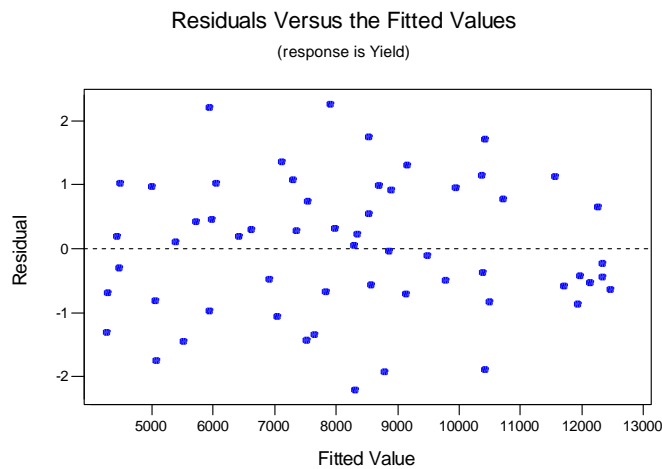
- (a) Replace *Temperature* by its logarithm.
- (b) Replace *Yield* by its logarithm.
- (c) Replace both *Temperature* and *Yield* by their logarithms.
- (d) Remove the highest and lowest points on the plot.
- (e) This problem is caused by a high leverage point, which should be identified and removed.
- (f) There is no problem with this plot.

(ii)



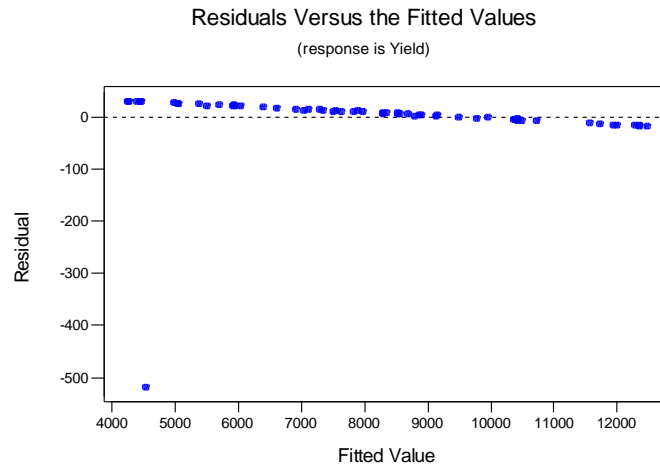
- (a) Replace *Temperature* by its logarithm.
- (b) Replace *Yield* by its logarithm.
- (c) Replace both *Temperature* and *Yield* by their logarithms.
- (d) Remove the highest and lowest points on the plot.
- (e) This problem is caused by a high leverage point, which should be identified and removed.
- (f) There is no problem with this plot.

(iii)



- (a) Replace *Temperature* by its logarithm.
- (b) Replace *Yield* by its logarithm.
- (c) Replace both *Temperature* and *Yield* by their logarithms.
- (d) Remove the two highest points on the plot.
- (e) This problem is caused by a high leverage point, which should be identified and removed.
- (f) There is no problem with this plot.

(iv)



- (a) Replace *Temperature* by its logarithm.
- (b) Replace *Yield* by its logarithm.
- (c) Replace both *Temperature* and *Yield* by their logarithms.
- (d) The relation between *Temperature* and *Yield* is clearly non-linear.
- (e) This problem is caused by an outlier point, which should be identified and removed.
- (f) There is no problem with this plot.

S9. Charlotte has been working as a “day trader” for her own account for several months now. She has been comparing the execution times for two on-line brokerage services. This execution time represents the time gap between her submitting the order and her receiving an electronic notice that the trade has been completed.

For the service Compu-Trade, she has timed 12 trades, resulting in an average of 82 seconds, with a standard deviation of 25 seconds.

For the service *E-Market*, she has timed 20 trades, resulting in an average of 104 seconds, with a standard deviation of 28 seconds.

She found that the pooled standard deviation is $s_p = 26.94$. The t statistic to compare these was found to be -2.2364 . This value was found to be significant at the 0.05 level but not at the 0.01 level.

For each statement below, select the most appropriate response.

- (i) The null hypothesis under test in this situation would be written as
- (a) $H_0: 82 = 104$
 - (b) $H_0: \mu_{CT} = 104$
 - (c) $H_0: \mu_{CT} = 82$
 - (d) $H_0: \mu_{CT} = \mu_{EM}$
 - (e) $H_0: \mu_{CT} \neq \mu_{EM}$
- (ii) Consider a confidence interval for the parameter difference $\mu_{CT} - \mu_{EM}$.
- (a) Both the 95% confidence interval and the 99% confidence interval would cover the value zero.
 - (b) The 95% confidence interval would cover the value zero, but the 99% confidence interval would not cover the value zero.
 - (c) The 95% confidence interval would not cover the value zero; however, the 99% confidence interval would cover the value zero.
 - (d) Neither the 95% confidence interval nor the 99% confidence interval would cover the value zero.
 - (e) Neither the 95% confidence interval nor the 99% confidence interval would cover the value $82 - 104 = -22$.
- (iii) Consider the p value for this hypothesis test.
- (a) The value of p is less than 0.01.
 - (b) The value of p is greater than 0.05.
 - (c) The value of p is less than 0.05, but it is greater than 0.01.
 - (d) The value of p is less than 0.01, but it is greater than 0.05.
 - (e) Nothing can be said about p , based on the facts presented.
- (iv) Charlotte showed her data to Hank, and Hank checked the arithmetic. Hank obtained the value $t = +2.2364$. The most plausible reason for the disagreement between Charlotte and Hank is
- (a) One of them has the correct answer, but the other made a numerical error in doing the arithmetic.
 - (b) Charlotte neglected to give Hank one of the samples; that is, Hank was working from only one of the two sets of values.
 - (c) Hank used the test in the form which allows for unequal standard deviations.
 - (d) Hank used $\bar{x}_{EM} - \bar{x}_{CT}$ in the numerator of his t statistic, while Charlotte used $\bar{x}_{CT} - \bar{x}_{EM}$.
 - (e) Hank replaced the values by their logarithms before doing the t test.

- (v) The t statistics has this number of degrees of freedom:
- 12
 - 20
 - 30
 - 32
 - ∞
- (vi) What is a proper conclusion to draw from these data?
- E -Market is significantly slower than Compu-Trade.
 - We must reserve judgment.
 - We would need larger sample sizes in order to assert a significant difference.
 - Compu-Trade is faster, but not significantly so.
 - No conclusion should be made.
- (vii) Suppose that you consider a *new* transaction with E -Market. Which of the following would be an approximate 95% prediction interval for the execution time?
- 104 ± 28
 - 104 ± 56
 - $104 \pm \frac{28}{\sqrt{20}}$
 - $104 \pm 2 \times \frac{28}{\sqrt{20}}$
 - $104 \pm t_{0.025;19} \frac{28}{\sqrt{20}}$
- (viii) Charlotte used the pooled standard deviation s_p in doing her arithmetic. This means that
- she apparently assumed that $\sigma_{CT} = \sigma_{EM}$.
 - she was willing to allow $\sigma_{CT} \neq \sigma_{EM}$ but found the pooled standard deviation version to be computationally easier.
 - she found that σ_{CT} and σ_{EM} were significantly different by Bartlett's test.
 - she wanted to publish her results in *The Journal of Finance*.
 - she made an error in judgment, as clearly $25 = s_{CT} \neq s_{EM} = 28$.
- (ix) In comparing the execution times for these two services, which of the following devices might have been useful?
- A scatterplot, with Compu-Trade on the horizontal axis and E -Market on the vertical axis.
 - A time plot, showing the results in the sequence order that they were collected, using difference colors for Compu-Trade and E -Market.
 - Side-by-side boxplots.
 - A histogram showing the full set of $12 + 20 = 32$ trades.
 - Separate normal probability plots for the two sets of values.

- (x) Suppose that the sample sizes were doubled, to 24 and 40, while keep the sample averages and standard deviations the same. Then
- the value of s_p would change only slightly, and t would grow by a factor about $\sqrt{2}$.
 - the values of s_p and t would remain unchanged.
 - the values of s_p and t would double.
 - the value of s_p would double, but t would be unchanged.
 - the value of s_p would change only slightly, but t would double.

S10. The following is a stem-and-leaf display for a small set of data:

```

Stem-and-leaf of Elastic    N = 30
Leaf Unit = 1.0

 1   3 4
 3   3 79
 9   4 112334
11   4 59
15   5 1134
15   5 68
13   6 03
11   6 566
 8   7 133
 5   7 58
 3   8 4
 2   8 59

```

- How many data points are in this set of values?
- What is the largest data value?
- What is the median value?

S11. The section of information that follows contains information about five variables: Y, CHIPS, RAISINS, MARSHM, REPACK, FLAKE. Regression information follows. Some positions are marked with letters and will be involved in the questions which follow.

| DESCRIPTIVE STATISTICS | | | | | | |
|------------------------|-----------|--------|---------|---------|---------|---------|
| | Y | CHIPS | RAISINS | MARSHM | REPACK | FLAKE |
| N | 59 | 59 | 59 | 59 | 59 | 59 |
| MEAN | 9478.8 | 54.330 | 24.141 | 21.764 | 498.85 | 24.748 |
| SD | 432.77 | 6.6679 | 6.2581 | 1.1350 | 42.321 | 10.151 |
| MINIMUM | 8288.6 | 41.161 | 10.414 | 18.364 | 413.08 | -1.9414 |
| MEDIAN | 9516.4 | 54.122 | 24.399 | 21.861 | 500.82 | 25.781 |
| MAXIMUM | 1.050E+04 | 73.326 | 41.801 | 23.795 | 591.60 | 48.762 |
| SKEW | -0.2631 | 0.4241 | 0.1347 | -0.4273 | -0.1337 | -0.3175 |

CORRELATIONS (PEARSON)

| | Y | CHIPS | RAISINS | MARSHM | REPACK |
|---------|--------|---------|---------|---------|---------|
| CHIPS | 0.3705 | | | | |
| RAISINS | 0.1799 | 0.0993 | | | |
| MARSHM | 0.0897 | 0.1129 | 0.3522 | | |
| REPACK | 0.4511 | 0.7377 | 0.2067 | 0.0622 | |
| FLAKE | 0.4198 | -0.4498 | -0.1667 | -0.0417 | -0.5137 |

CASES INCLUDED 59 MISSING CASES 0

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF Y

PREDICTOR

| VARIABLES | COEFFICIENT | STD ERROR | STUDENT'S T | P | VIF |
|-----------|-------------|-----------|-------------|--------|-----|
| CONSTANT | 3754.33 | 608.983 | 6.16 | 0.0000 | |
| CHIPS | 15.9204 | 5.60154 | 2.84 | 0.0063 | 2.3 |
| RAISINS | 11.2591 | -- (f) -- | 2.59 | 0.0124 | 1.2 |
| MARSHM | -0.39297 | 23.4672 | -0.02 | 0.9867 | 1.2 |
| REPACK | 7.26345 | 0.92305 | 7.87 | 0.0000 | 2.5 |
| FLAKE | 39.3137 | 2.86676 | 13.71 | 0.0000 | 1.4 |

R-SQUARED 0.8276 RESID. MEAN SQUARE (MSE) 35330.8
 ADJUSTED R-SQUARED 0.8114 STANDARD ERROR OF ESTIMATE 187.965

| SOURCE | DF | SS | MS | F | P |
|------------|----|-----------|-----------|-------|--------|
| REGRESSION | 5 | 8.990E+06 | 1.798E+06 | 50.89 | 0.0000 |
| RESIDUAL | 53 | -- (h) -- | 35330.8 | | |
| TOTAL | 58 | 1.086E+07 | | | |

CASES INCLUDED 59 MISSING CASES 0

- What is the correlation coefficient between CHIPS and REPACK?
- In the regression, which variable was designated as the dependent variable?
- What is the maximum value of Y in the data base?
- How many data points are used in this problem?
- What is the value of s_e , the standard error of regression?
- What value belongs in the position marked --(f)-- ?
- For which of the independent variables would you accept the null hypothesis that the true coefficient value is zero?
- What value belongs in the position marked --(h)-- ?
- The value of the F statistic was given as 50.89. What null hypothesis is tested by this statistic? Do you accept or reject the null hypothesis?
- Suppose that you wanted to run the regression of REPACK on FLAKE. That is, you want to treat REPACK as the dependent variable and FLAKE as the independent variable. What is the estimated slope in this regression?

S12. A BEST SUBSETS regression was done in a problem involving four independent variables. The starting model was

$$Y_i = \beta_0 + \beta_P P_i + \beta_Q Q_i + \beta_R R_i + \beta_S S_i + \varepsilon_i$$

The computer listing (partially edited) that resulted from this was the following:

| Vars | R-Sq | C-p | P | Q | R | S |
|------|--------|-----|---|---|---|---|
| 1 | 0.5288 | 5.3 | | X | | |
| 2 | 0.5621 | 1.4 | | X | X | |
| 3 | 0.5631 | 3.0 | X | X | X | |
| 4 | 0.5632 | 5.0 | X | X | X | X |

Based on this listing, which simplified model would you recommend? Why?

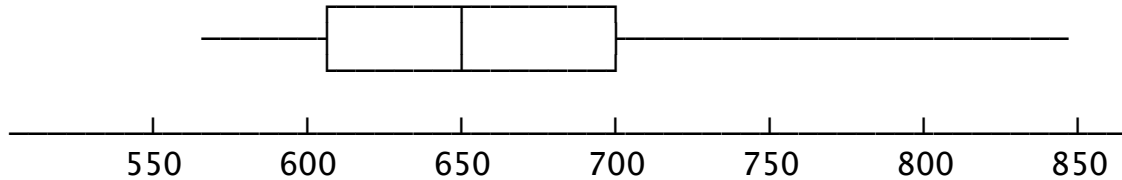
S13. The computer listing shown below is the result of a multiple regression. Questions follow.

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF
WRHSCOST Warehousing costs

| PREDICTOR VARIABLES | COEFFICIENT | STD ERROR | STUDENT'S T | P | VIF |
|------------------------|-------------|----------------------------|-------------|---------|--------|
| CONSTANT | 4387.94 | 280.510 | 15.64 | 0.0000 | |
| OLDSTOCK | 1.14280 | 0.19971 | 5.72 | 0.0000 | 2.3 |
| SERVCHRG | 0.25193 | 0.66771 | 0.38 | 0.7071 | 1.2 |
| COOLING | 0.18350 | 1.79847 | 0.10 | 0.9190 | 2.3 |
| INSURNCE | 0.69458 | 0.30378 | 2.29 | 0.0253 | 1.0 |
| R-SQUARED | 0.5632 | RESIDUAL MEAN SQUARE (MSE) | | 49632.4 | |
| ADJUSTED R-SQUARED | 0.5375 | STANDARD ERROR OF ESTIMATE | | 222.783 | |
| SOURCE | DF | SS | MS | F | P |
| REGRESSION | 4 | 4.352E+06 | 1.088E+06 | 21.92 | 0.0000 |
| RESIDUAL | 68 | 3.375E+06 | 49632.4 | | |
| TOTAL | 72 | 7.727E+06 | | | |
| CASES INCLUDED 73 | | MISSING CASES 0 | | | |

- How many data points were used in this analysis?
- What is the name of the dependent variable?
- How many independent variables were used for this regression?
- Give the fitted regression equation.
- The usual regression model contains noise terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$. These are assumed to have standard deviation σ . What is the estimate of σ ?
- What fraction of the variability in the dependent variable is explained by the regression?
- What is the standard deviation of the dependent variable?
- Would you describe this regression as useful?

S14. Boxplots mark five critical values in a set of numbers: the minimum, the 25th percentile, the median, the 75th percentile, and the maximum. The boxplot below shows the weekly wages for a sample of skilled electrical workers in the greater Kansas City area. The horizontal axis shows prices in dollars. For example, 600 means \$600.



Please answer each of the following as either “true” or “false” or “no way to tell.”

- (a) The median weekly wage for this sample is very close to \$650.
- (b) The lowest-paid electrician in this sample gets about \$560 per week.
- (c) Nearly all the electricians in this sample earn between \$600 and \$700 per week.
- (d) Electricians earn more in St. Louis.
- (e) Only 10% of the electricians in this sample earn more than \$700 per week.
- (f) None of the skilled electricians in the Kansas City area earn less than (about) \$560 per week.
- (g) The average weekly wage for the skilled electricians in the Kansas City area is \$650.
- (h) The sample size used to create this boxplot was at least 20.
- (i) The range of the data is (about) $\$850 - \$560 = \$290$.
- (j) The standard deviation of this set of data is less than \$290.

SOLUTIONS TO SAMPLE QUALIFYING EXAM

S1. (b). The correction of this error removed two values with a total of $230 + 250 = 480$ and replaced them with two values with a total of $200 + 280 = 480$. Certainly the average does not change. Moreover, the corrected values are farther away from $\bar{x} = 240$, so the standard deviation must increase.

S2.

- (a) *False*
- (b) *True*
- (c) *False*. The interquartile range is a measure of dispersion, and the median is a measure of location. These bear no particular relationship to each other.
- (d) *False*. Negative standard deviations indicate computational errors.
- (e) *False*. Standard deviations are not easy to compute by hand.
- (f) *True*. Half the women have salaries over \$25,000, and *at least* half the men also have salaries over \$25,000. When the two groups are combined, it is guaranteed that at least half have salaries over \$25,000.
- (g) *False*. Here is an example:

| Factory | Average salary Line worker | Average salary Floor supervisor |
|----------|-------------------------------|------------------------------------|
| Q | \$20,000 ($n = 10$) | \$25,000 ($n = 2$) |
| R | \$32,000 ($n = 25$) | \$34,000 ($n = 1$). |
| Combined | \$28,571 ($n = 35$) | \$28,000 ($n = 3$) |

There are 10 line workers getting a total salary of $10 \times \$20,000 = \$200,000$ in factory Q. The 25 line workers in factory R get a total salary of $25 \times \$32,000 = \$800,000$. The average salary over 35 workers is $\frac{\$200,000 + \$800,000}{35} \approx \$28,571$.

The average salary for the three floor supervisors is $\frac{\$84,000}{3} = \$28,000$.

S3.

- (a) List P has the smaller standard deviation.
(b) List Q has the smaller standard deviation.
(c) List P has the smaller standard deviation. This is tricky. The calculation of the standard deviation involves finding the sum $\sum_{i=1}^{500} (x_i - \bar{x})^2$. For both lists \bar{x} is 20.

List P involves the sum

$$100 \times (-2)^2 + 100 \times (-1)^2 + 100 \times 0^2 + 100 \times 1^2 + 100 \times 2^2 = 1,000$$

List Q involves the sum

$$150 \times (-2)^2 + 200 \times 0^2 + 150 \times 2^2 = 1,200$$

This will lead to a smaller standard deviation for list P.

- (d) List X will have the smaller standard deviation. The standard deviation is calculated as $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$. Let the value of the sum $\sum_{i=1}^n (x_i - \bar{x})^2$ be M for list P. Certainly the value of the sum $\sum_{i=1}^n (x_i - \bar{x})^2$ for list Q must be $2M$. The standard deviation for list P is then $\sqrt{\frac{1}{999} M}$, and the standard deviation for list Q is $\sqrt{\frac{1}{1,999} 2M}$. As $\frac{1}{999} > \frac{2}{1,999}$, it must follow that list Q has the smaller standard deviation. This is a very tricky question.

S4.

- (a) 16 oz
(b) 35
(c) 3 inches
(d) \$200
(e) \$1.20

As a common statistical fact, about $\frac{2}{3}$ of the observations should be within one standard deviation of the average, and therefore about $\frac{1}{3}$ of the observations should be more than one standard deviation away from the average. Thus, for the milk problem in (a), the values larger than 16 oz are simply not believable.

S5.

(a) *False.* Carolyn used the interval $\bar{x}_{18} \pm t_{0.025;17} \frac{s_{18}}{\sqrt{18}}$, where the subscript “18” indicates that she used only her 18 values. Lou used the interval $\bar{x}_{25} \pm t_{0.025;24} \frac{s_{25}}{\sqrt{25}}$, using all the data. The lengths of these intervals depend on:

(i) $t_{0.025;17}$ versus $t_{0.025;24}$, which are nearly identical

(ii) $\frac{1}{\sqrt{18}}$ versus $\frac{1}{\sqrt{25}}$, which certainly favors Lou’s having a shorter interval

(iii) s_{18} versus s_{25} , which can be somewhat different

Thus, it’s likely that Lou has a shorter interval, but the fact that s_{25} can be larger than s_{18} makes this fall short of a guarantee.

(b) *False.* Lou uses 24 degrees of freedom.

(c) *True.* As long as they have different intervals, this is a possibility.

(d) *False.* This would be true if their intervals were completely independent. However, they had 18 data values in common.

(e) *False.* This is a bit of a shock. As they both used 95% confidence intervals, the coverage probability is 95% for each of them.

S6.

(a) *Could happen.* Here is such a set: 12 13 14 15 16 17.

(b) *Could happen.* As long as they make different assumptions and different calculations, they could get inconsistent results.

(c) *Impossible.* The sample fraction \hat{p} must be computed as $\frac{x}{38}$ for some integer x .

However, $\frac{15}{38} \approx 0.3947$ would be rounded to 0.39 and $\frac{16}{38} \approx 0.4211$ would be rounded to 0.42. Thus, it is impossible to report $\hat{p} = 0.41$ with $n = 38$.

(d) *Impossible.* The sample standard deviation s cannot be negative.

(e) *Could happen.* Usually s_Y is larger, but $s_Y < s_e$ can happen.

(f) *Could happen.*

(g) *Could happen.* The value 0.61 is outside the confidence interval, so rejection of the null hypothesis $H_0: p = 0.61$ is the usual state of affairs.

(h) *Could happen.* This is just a case of Type II error, which happens all the time.

S7.

- (a) *False.*
- (b) *True.*
- (c) *False.* The interval $\$80,000 \pm \$22,000$ was given as the *prediction* for one new individual. The interval for the parameter expression $\beta_0 + \beta_1 \times 18$, which represents the mean salary for *all* MBAs, would be much narrower.
- (d) *True.* This is the literal meaning of the prediction interval.

S8.

- (i-b). The right-expanding residuals suggest that the noise standard deviation is proportional to the value of *Yield*.
- (ii-c). Not only are the residuals right-expanding, but there is also extreme clumping at the low values. This can be cured by taking the logarithms of both variables.
- (iii-f). This is an ideal residual versus fitted plot.
- (iv-e). The point at the lower left is clearly a high leverage point, and it has certainly caused some damage. This point is masking the regression slope, and it must be removed.

S9.

- (i-d) Data values cannot appear in the hypotheses, ruling out (a), (b), and (c). The = version, which is (d), becomes the null hypothesis. The statement given in (e) will be H_1 .
- (ii-c) As the data are significant at the 0.05 level, the 95% confidence interval for $\mu_{CT} - \mu_{EM}$ would exclude zero. However, the data are not significant at the 0.01 level, so the 99% interval would indeed cover zero.
- (iii-c) If H_0 is rejected at the 0.05 level, then $p \leq 0.05$. Similarly, if H_0 is accepted at the 0.01 level, then $p > 0.01$.
- (iv-d) Charlotte and Hank had results which were exact negatives of each other. Here (d) is the most reasonable explanation by far.
- (v-c) The degrees of freedom number is $n_{CT} + n_{EM} - 2$. This value is 30.
- (vi-a) This is the direct conclusion from this experiment. Any other conclusion would be either incorrect or seriously indecisive.

- (vii-b) This is (estimate) $\pm 2 \times (\text{sd})$. Version (a) is (estimate) $\pm (\text{sd})$, which would be a 68% prediction interval. The other intervals all involve $\frac{s_{EM}}{\sqrt{n_{EM}}}$; this is the standard error of the mean, and it would come into play if making a confidence interval for μ_{ET} .
- (viii-a) Story (c) is unbelievable as s_{CT} and s_{EM} are very close. As for part (e), it is simply unrealistic to expect two estimates of the same quantity to be exactly equal.
- (ix-c) Side-by-side boxplots will show plainly the differences between the two samples. The scatterplot in (a) could not even be constructed, as there are 12 trades from one company and 20 from the other. Item (b), the color-coded time plot, will be disastrously confusing. Item (d) talks about combining the two samples and must necessarily be useless for distinguishing them. Item (e) wants to explore whether the data are (separately) normally distributed. This can't help you.
- (x-a) Note that s_p^2 is a weighted average of s_{CT}^2 and s_{EM}^2 with weights 11 and 19. After the sample sizes are doubled, the weights will be 23 and 39. Thus, it happens that s_p will change very slightly. The calculation of the t statistic is

$$t = \frac{\sqrt{\frac{n_{CT} n_{EM}}{n_{CT} + n_{EM}}} (\bar{x}_{CT} - \bar{x}_{EM})}{s_p}$$

When the sample sizes are doubled, the fraction $\frac{\bar{x}_{CT} - \bar{x}_{EM}}{s_p}$ will change very little, but the expression $\sqrt{\frac{n_{CT} n_{EM}}{n_{CT} + n_{EM}}}$ will grow to $\sqrt{\frac{2n_{CT} 2n_{EM}}{2n_{CT} + 2n_{EM}}} = \sqrt{2} \sqrt{\frac{n_{CT} n_{EM}}{n_{CT} + n_{EM}}}$. Thus t will grow by the factor $\sqrt{2}$.

S10.

- As indicate by the line $N = 30$ there are 30 points.
- The largest value is 89.
- The median occurs between the 15th and 16th values. These are 54 and 56, so the median would conventionally be reported as 55.

S11.

- The value is 0.7377, read directly from the output.
- Y was used as the dependent variable.
- It's 1.050E+04, meaning 10,500.

- (d) There are 59 data points.
- (e) It's 187.965.
- (f) Since $\text{STUDENT'S T} = \text{COEFFICIENT} \div \text{STD ERROR}$, we must have

$$\text{STD ERROR} = \frac{\text{COEFFICIENT}}{\text{STUDENT'S T}} = \frac{11.2591}{2.59} \approx 4.34714.$$

The value actually covered up was 4.34916.

- (g) Only for MARSHM, for which the value of STUDENT'S T is only -0.02. That is, we accept $H_0 : \beta_{\text{MARSHM}} = 0$.
- (h) Since $\text{SS}(\text{REGRESSION}) + \text{SS}(\text{RESIDUAL}) = \text{SS}(\text{TOTAL})$, we have

$$8.990\text{E}+06 + \text{SS}(\text{RESIDUAL}) = 1.086\text{E}+07$$

or

$$8,990,000 + \text{SS}(\text{RESIDUAL}) = 10,860,000$$

which leads to $\text{SS}(\text{RESIDUAL}) = 10,860,000 - 8,990,000 = 1,870,000$. The value was actually printed as 1.873E+06.

- (i) This is a test of

$$H_0 : \beta_{\text{CHIPS}} = 0, \beta_{\text{RAISINS}} = 0, \beta_{\text{MARSHM}} = 0, \beta_{\text{REPACK}} = 0, \beta_{\text{FLAKE}} = 0$$

which says that all five coefficients are equal to zero. With a calculated value of 50.89, the null hypothesis is soundly rejected. The official 5% cutoff point, using (5, 53) degrees of freedom, is 2.39; the 1% cutoff point is 3.38.

- (j) This estimated slope will have units which are $\frac{\text{units of REPACK}}{\text{units of FLAKE}}$. The actual formula is $b_{\text{REPACK on FLAKE}} = r_{\text{REPACK, FLAKE}} \frac{\text{SD}(\text{REPACK})}{\text{SD}(\text{FLAKE})}$ and the numeric value is $-0.5137 \frac{42.321}{10.151} \approx -2.142$

S12. You like to pick the simplest model which is adequate. The R^2 criterion would lead you to the one-variable or perhaps the two-variable model. Certainly you would not go to the more complicated models.

The C_p statistic should be at or below 1+the number of variables. A desirable C_p statistic for the one-variable line should be 2.0 or below, and a desirable value for the two-variable line should be 3.0 or below. Thus, the C_p statistic leads us to the two-variable model

$$Y_i = \beta_0 + \beta_Q Q_i + \beta_S S_i + \varepsilon_i$$

S13.

- (a) There were $n = 73$ points in this computer run.
- (b) The dependent variable is WRHSCOST. With this software, it was also identified as Warehousing costs.
- (c) Four independent variables were used.
- (d) The fitted equation is

$$\begin{aligned} \widehat{\text{WRHSCOST}} &= 4387.94 \\ &+ 1.14280 \text{ OLDSTOCK} \\ &+ 0.25193 \text{ SERVCHRG} \\ &+ 0.18350 \text{ COOLING} \\ &+ 0.69458 \text{ INSURNCE} \end{aligned}$$
- (e) The value of s_ε is listed as STANDARD ERROR OF ESTIMATE. The number is 222.783.
- (f) This asks for R^2 , and the value is $0.5632 = 56.32\%$.
- (g) The standard deviation of the dependent variable takes a little detective work. The TOTAL line of the analysis of variance table gives $n - 1$ as 72 and gives S_{yy} as $7.727\text{E}+06 = 7,727,000$. Then $s_y = \sqrt{\frac{7,727,000}{72}} \approx \sqrt{107,319.44} \approx 327.60$.
- (h) This would be deemed a useful regression with $R^2 = 56.32\%$ and with s_ε neatly less than s_y . Of course, it may not be quite good enough for the context in which it was done, but that's a separate judgment.

S14.

- (a) *True.* This is a correct reading of the boxplot.
- (b) *True.* This is a correct reading of the boxplot.
- (c) *False.* The values \$600 and \$700 are (about) the 25th and 75th percentiles. It would be correct to say that about half the electricians are in this pay range.
- (d) *No way to tell.* The statement may be true, but this judgment goes beyond the scope of the data.
- (e) *False.* Since \$700 is (about) the 75th percentile, it would be correct to say that about 25% of the electricians in this sample earn more than \$700 per week.
- (f) *No way to tell.* This is the best answer. The statement would be correct if and only if the lowest-paid electrician ended up in the sample. If the sample involved 4% of the skilled electricians in the Kansas City area, then the probability that the lowest-paid one ends up in the sample is 0.04.
- (g) *False.* The boxplot is designed to show medians, not averages. Moreover, the upper whisker is much longer than the lower whisker, and this makes it almost certain that the mean is larger than the median.
- (h) *No way to tell.* The boxplot is not designed to reveal sample sizes.
- (i) *True.* The range is the maximum value *minus* the minimum value.
- (j) *True.* One is tempted to say *no way to tell*, since the boxplot is not designed to reveal standard deviations. However, the standard deviation is less than the range (as long as the data set has at least two different values).