◊◊

Here are problems used on the final exams in previous versions of B01.1305. These are identified as A, B, C, and D. E2 is an additional problem. Problems in the F set are from the actual 2006 final exam. Some sequence numbers are missing; these correspond to problems on linear programming, which was not part of the course material after 1997, or to problems on topics no longer covered.

Solutions follow, beginning on page 23.

A1. Marisol took a sample of 18 values from a population which was assumed normal. With an average of $\bar{x}$ = 510.6 and a standard deviation of s = 64.2, she gave a 95% confidence interval as $510.6 \pm 2.1098 \frac{64.2}{\sqrt{18}}$ or 510.6 ± 31.9. She reported this interval as (478.7, 542.5).

Unknown to Marisol, the true population mean was μ = 549.2. Did Marisol make a mistake? If so, what was it?

A2. The information that follows comes from a computer run in which the objective was to find a useful regression. This shows the results of these steps:

Descriptive statistics
Pearson correlations
Linear regression using four independent variables
Linear regression using two independent variables
Best subset regression
Stepwise regression

Questions follow.

Descriptive Statistics

Variable	N	Mean	Median	StDev
ALUM	49	41.280	40.900	5.251
BORON	49	34.143	34.400	4.340
COBALT	49	48.31	48.90	9.85
DIAMOND	49	47.724	48.100	—a—
SALES	49	2176.6	2136.0	242.9

Variable	Min	Max	Q1	Q3
ALUM	31.100	51.600	37.900	44.700
BORON	23.700	44.600	30.800	37.450
COBALT	27.60	70.10	41.85	55.40
DIAMOND	31.700	62.800	44.600	51.550
SALES	1741.0	2720.0	2020.5	2332.0

Correlations (Pearson)

	ALUM	BORON	COBALT	DIAMOND
BORON	0.533			
COBALT	-0.424	-0.513		
DIAMOND	-0.504	-0.336	0.397	
SALES	0.064	-0.299	0.698	0.053

Regression Analysis

The regression equation is
SALES = 765 + 20.5 ALUM - 7.29 BORON + 21.3 COBALT - 4.47 DIAMOND

Predictor	Coef	SECoef	T	P	VIF
Constant	764.8	404.3	1.89	0.065	
ALUM	20.519	5.270	3.89	0.000	1.7
BORON	-7.287	6.243	-1.17	0.249	1.6
COBALT	21.261	2.632	8.08	0.000	1.5
DIAMOND	-4.473	4.251	-1.05	0.298	1.4

S = 147.2 R-Sq = 66.3% R-Sq(adj) = 63.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1877299	469325	21.65	0.000
Error	44	953959	—d—		
Total	48	2831258			

Unusual Observations

Obs	ALUM	SALES	Fit	StDev Fit	Residual	St Resid
3	51.6	1903.0	2193.0	54.0	-290.0	-2.12R

R denotes an observation with a large standardized residual

Regression Analysis

The regression equation is
SALES = 288 + 20.3 ALUM + 21.8 COBALT

Predictor	Coef	SECoef	T	P	VIF
Constant	287.6	257.6	1.12	0.270	
ALUM	20.273	4.490	4.51	0.000	1.2
COBALT	21.778	—e—	9.10	0.000	1.2

S = 148.0 R-Sq = 64.4% R-Sq(adj) = 62.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1824009	912005	41.65	0.000
Error	46	1007249	21897		
Total	48	2831258			

Unusual Observations

Obs	ALUM	SALES	Fit	StDev Fit	Residual	St Resid
3	51.6	1903.0	2233.2	47.1	-330.2	-2.35R

R denotes an observation with a large standardized residual


```
Best Subsets Regression                               D
Response is SALES                                   C I
                                                  B O A
                                                A O B M
                                                L R A O
              R-Sq                              U O L N
Vars   R-Sq   (adj)   C-p          S            M N T D
   1   48.7   47.6   22.0    175.86                 X
   2   64.4   62.9    3.5    147.98             X   X
   3   65.5   63.2    4.1    147.42             X X X
   4   66.3   63.2    5.0    147.24             X X X X
```

```
Stepwise Regression
 F-to-Enter:      4.00    F-to-Remove:      4.00
 Response is  SALES   on  4 predictors, with N =   49

    Step        1        2
Constant   1345.7    287.6

COBALT       17.2     21.8
T-Value      6.67     9.10

ALUM                  20.3
T-Value               4.51

S             176      148
R-Sq        48.66    64.42
```

Please answer the following questions.

(a) Find the position marked —a—. Guess the value that appeared there.
(b) Which variable was used as the dependent variable in the regression work?
(c) Which pairs of variables have negative correlations?
(d) Find the position marked —d—. What value should appear there?
(e) Find the position marked —e—. What value should appear there?
(f) Following the initial regression run, there is a second regression of SALES on only two of the independent variables, ALUM and COBALT. What is the fitted model in this second regression run?
(g) Why, in your opinion, was this second regression done?
(h) The BEST SUBSET section shows four models. Which of these four models would provide an adequate fit to the data? (This is not asking you to choose the best model.)
(i) The BEST SUBSET section seems to rank the independent variables in terms of their usefulness, from best to worst. What is this ordering?

    DISCLAIMER: Not all BEST SUBSET results suggest an ordering. Moreover, the claimed ordering does not reflect any scientific reality.

(j) The STEPWISE section also ranks the independent variables in terms of usefulness. What is the ordering?

    DISCLAIMER: Same disclaimer as in (i).

A4. Maple ice cream is made at Chauncey's Superlative Ice Cream factory by squirting high-quality maple syrup into an ice cream base. The resulting mixture is then stirred and frozen. Because the maple syrup is sticky, the dispenser cannot always squirt out the same quantity. The management is concerned about the concentration of maple syrup in the ice cream. A sample of 24 half-gallon containers was taken, and the maple syrup content of each was measured. The average quantity in this sample was 41.6 ounces, with a standard deviation of 3.5 ounces.

(a) Give a 95% confidence interval for the mean amount of maple syrup in a half-gallon container.
(b) Provide an interval with the property that about 95% of all containers will contain maple syrup amounts within the interval.

A7. The baloney used at the Super Submarine Drive-In is prepared by a local meat-packing plant. The baloney was ordered to the specification that each slice have a weight of 25 grams and contain no more than 11 grams of fat. The requirement that the weight be 25 grams is easily checked, and there has never been a problem here. There is, however, some concern that the baloney slices might contain more than 11 grams of fat. You decide to do a quality control inspection, and you arrange for the random sampling of 80 baloney slices. These 80 baloney slices had an average fat content of 14 grams, with a sample standard deviation of 1.2 grams. You compute the *t*-statistic as

$$t = \sqrt{80}\,\frac{14-11}{1.2} \approx 22.36.$$

Please respond T (true) or F (false) to each of the following.

(a) This problem can be regarded as a test of the null hypothesis $H_0: \mu = 11$.
(b) The parameter μ represents the mean weight of the 80 baloney slices in the sample.
(c) The *t*-statistic has 79 degrees of freedom.
(d) Most published *t* tables have no line for 79 degrees of freedom.
(e) At the 5% level of significance, one would accept the hypothesis in (a).
(f) Though the sampled baloney did have an average fat content above 11 grams, the amount of extra fat seems to be small.
(g) The standard deviation of the amount of fat seems to be close to 1.2 grams.
(h) Type II error here consists of claiming that the baloney is acceptable when it is really too fatty.
(i) Given the actual data, it appears that the probability of a type II error is quite small.
(j) An approximate 95% prediction interval for the amount of fat in a baloney slice is about 14 grams ± 2.4 grams.

A9. A fruit dealer receives regular shipments of pears. He gets to make a yes/no purchase decision on each shipment by taking a sample of pears. The rule that he has devised is the following:

Select ten pears.
If the total weight equals or exceeds 2.5 pounds, then the shipment is accepted.

If the true distribution of pear weights in the first shipment in October is approximately normal with mean $\mu = 0.28$ pound and standard deviation $\sigma = 0.06$ pound, what is the probability that this shipment will be accepted?

A10. A regression of RETURN on SIZE and PARK was performed for a set of 109 randomly-selected toy stores. RETURN was measured in thousands of dollars, SIZE was floor space in thousands of square feet, and PARK was obtained by counting available numbers of parking spaces. (Stores located in indoor shopping malls were not used in the data base.) The fitted regression was

$$\widehat{\text{RETURN}} = \underset{(2.02)}{14.66} + \underset{(4.66)}{31.20}\ \text{SIZE} + \underset{(1.88)}{1.12}\ \text{PARK}$$

The figures in ( ) are standard errors.

(a) State the regression model on which this is based.
(b) What profit would you predict for a store with 10,000 square feet of floor space and parking for 100 cars?
(c) Which of the estimated coefficients is/are significantly different from zero?
(d) It was noticed that the correlation coefficient between SIZE and PARK was 0.67. Will this present a problem?

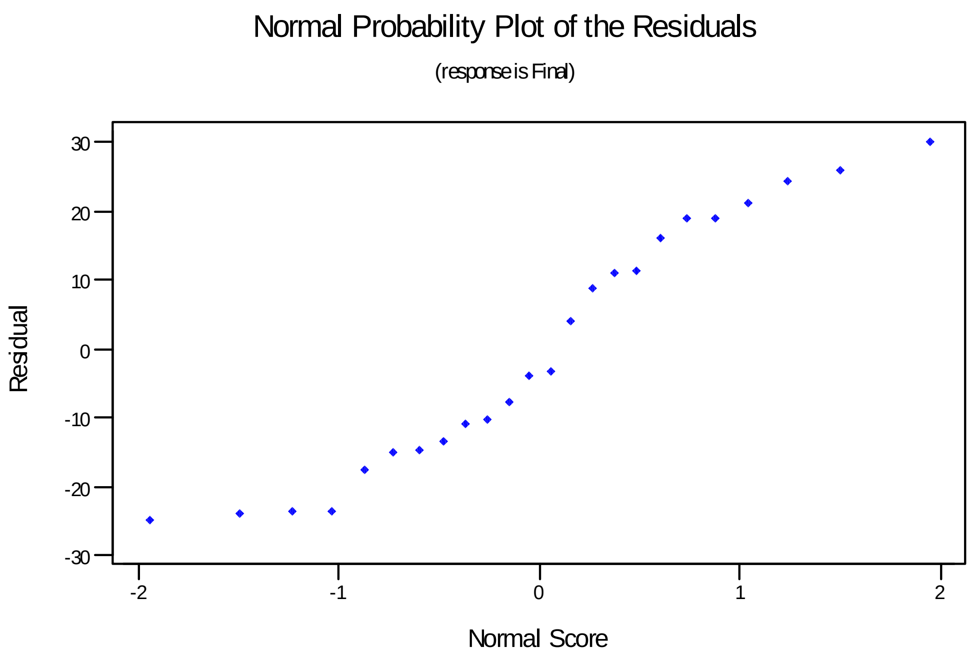
B1. In a certain course, there were two tests and a final exam. The data are listed here (in alphabetic order, by last name) for the 24 people who took the course:

Test1	Test2	Final	Test1	Test2	Final
110	80	105	47	65	84
110	102	103	135	80	120
75	81	85	97	95	132
118	108	128	100	95	130
30	33	83	100	90	103
118	64	90	32	51	100
59	56	115	42	12	45
52	78	79	105	99	121
57	69	75	60	101	150
105	75	128	75	77	96
45	42	106	125	102	150
140	110	155	92	91	100

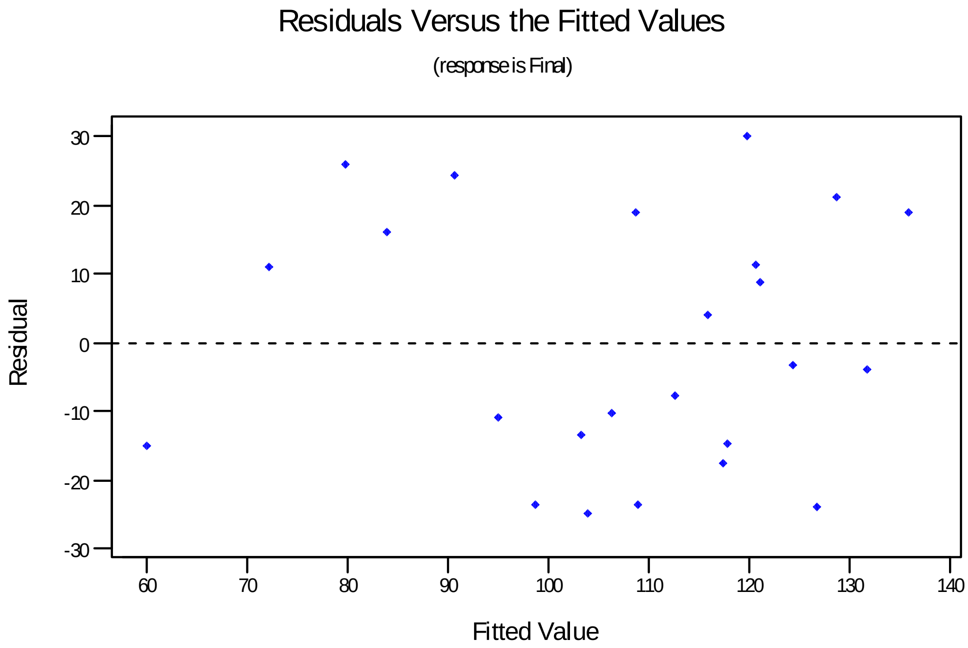
There was some curiousity as to whether the final exam could reasonably be predicted from the two tests. Accordingly, the linear regression of Final on (Test1, Test2) was done, with these results:


```
Regression Analysis

The regression equation is
Final = 46.8 + 0.128 Test1 + 0.646 Test2

Predictor        Coef       SECoef         T        P
Constant        46.85        12.94      3.62    0.002
Test1          0.1277       0.1702      0.75    0.461
Test2          0.6463       0.2299      2.81    0.010

S = 18.95       R-Sq = 53.6%     R-Sq(adj) = 49.2%

Analysis of Variance
Source        DF          SS          MS         F        P
Regression     2      8704.0      4352.0     12.12    0.000
Error         21      7539.6       359.0
Total         23     16243.6

Source        DF      Seq SS
Test1          1      5867.2
Test2          1      2836.7

Unusual Observations
Obs     Test1      Final      Fit  StDev Fit   Residual St Resid
 19        42      45.00    59.97      11.64     -14.97    -1.00 X
X denotes an observation whose X value gives it large influence.
```

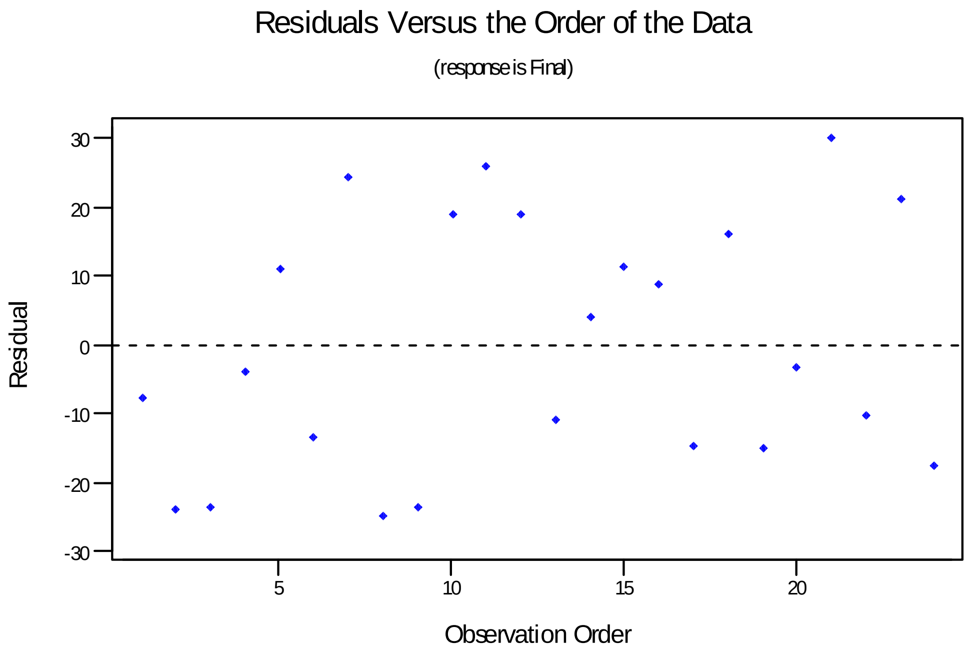
Additionally, some graphs were provided. Here first is the normal probability plot of the residuals.



Here is the residual-versus-fitted plot:



Here is a plot of the residuals in sequence order:



Please provide answers to the following questions:

(a) What fraction of the variability in the Final is explained by the two quizzes?
(b) Give the fitted regression equation.
(c) It appears that one of the quizzes was more useful than the other in terms of predicting the final. Which quiz? How do you know?
(d) The normal plot of the residuals looks reasonably straight, except for the four oddly-placed points at the bottom left. Locate these four points on the residual-versus-fitted plot. (If you verbally locate these points, that's fine. If not, just circle them on the residual-versus-fitted plot and indicate with letter "d.")
(e) Is there any useful information on the plot of the residuals versus the order of the data? If yes, tell what that information is. If not, please indicate why not.
(f) One of the students got a 45 on the final exam. Locate that student's point on the residual-versus-fitted plot. (If you verbally locate this point, that's fine. If not, just circle it on the residual-versus-fitted plot and indicate with letter "f.")
(g) One data point is listed as an unusual observation. Give the three test scores for this data point. Indicate also what you think is unusual about this point.
(h) Based on this output, it's possible to find the correlation coefficient between Test1 and Final. What is this correlation?
(*i*) The *F* value is given as 12.12. Exactly what hypothesis is being tested with this *F* value? What conclusion do you reach about this hypothesis?
(j) Find the standard deviation of Final. Using this value make a judgment as to whether the regression was useful.

B5. 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) In a single sample $\bar{x}$ = 141.2 and s = -32.1.
(b) In a very strong linear relationship, the correlation was found as r = 1.18.
(c) In a single sample $\bar{x}$ = 15.2 and s = 33.5.
(d) In a regression of Y on X, we computed s_Y = 21.04 and s_ε = 21.52.
(e) In a very strong multiple regression, the F statistic was found as F = 12,810.
(f) A hypothesis test was done at the 5% level of significance, and H_0 was rejected. With the same data, but using the 1% level of significance, H_0 was again rejected.
(g) With a single sample of n = 22 values, the 95% confidence interval for μ was found as 85.0 ± 6.2 and the null hypothesis $H_0: \mu = 92$ (tested against alternative $H_1: \mu \neq 92$) was accepted at the 5% level of significance.
(h) 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.

(i) 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.

(j) Sherry had a set of data of size *n* = 6 in which the mean and median were identical.

(k) 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.

(ℓ) In comparing two samples of continuous data, the standard deviations were found to be $s_x = 42.4$ and $s_y = 47.2$; also, the pooled standard deviation was found to be $s_p = 48.5$.

(m) Martha had a sample of data $x_1, x_2, \ldots, x_n$ in inches and tested the hypothesis $H_0 : \mu = 10$ inches against the alternative $H_1 : \mu \neq 10$ inches. Edna converted the data to centimeters by multiplying by $2.54 \frac{\text{cm}}{\text{in}}$ and tested $H_0 : \mu = 25.4$ cm against $H_1 : \mu \neq 25.4$ cm. Martha accepted the null hypothesis while Edna (using the same α) rejected it.

(n) Based on a sample of 38 yes-or-no values, the estimate $\hat{p}$ was found to be 0.41.

(*o*) In a regression involving dependent variable *Y* and possible independent variables *G*, *H*, *K*, *L*, and *M*, the regression of *Y* on (*G*, *H*, *L*) had R^2 = 53.2% and the regression of *Y* on (*G*, *K*, *M*) had R^2 = 56.4%.

(p) (a problem about linear programming)

(q) In a regression involving dependent variable *D* and possible independent variables *R*, *S*, and *V*, the *F* statistic was significant at the 5% level while all three *t* statistics were non-significant at the 5% level.

(r) For a regression with two independent variables, it was found that SS_{regr} = 58,400 and SS_{error} = 245,800.

(s) In a linear regression problem, AGE was used as one of the independent variables, even though AGE = 6 for every one of the data points (which were first-grade children). The estimated coefficient was found to be $b_{AGE} = 0.31$.

B6. The fill amounts which go into bottles of Nutri-Kleen (a home-use rubbing alcohol) are supposed to be normally distributed with mean 8.00 oz. and standard deviation 0.05 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely on a control chart. There are bars on the control chart at 7.88 oz. and 8.12 oz. If the actual amount goes below 7.88 oz. or above 8.12 oz., then the bottle will be declared out of control.

(a) If the process is in control, meaning μ = 8.00 oz. and σ = 0.05 oz., find the probability that a bottle will be declared out of control.

(b) In the situation of (a), find the probability that the number of bottles found out of control in an eight-hour day (16 inspections) will be zero.

(c) In the situation of (a), find the probability that the number of bottles found out of control in an eight-hour day (16 inspections) will be exactly one.

(d) If the process shifts so that $\mu = 7.98$ oz and $\sigma = 0.06$ oz, find the probability that a bottle will be declared out of control. (μ and σ both change)

If you can't solve (a), use the symbol γ as the solution to (a); then express the solutions to (b) and (c) in terms of γ.

B8. Two groups of economists were surveyed regarding opinions as to the price of gold on December 31, 1998. These economists were either academic (major employment by college or university) or professional (major employment by bank or securities firm). The findings were the following:

Group	Number	Average	Standard deviation
Academic	24	250	40
Professional	30	210	30

Give a 95% confidence interval for the population mean difference between these groups. Be sure to state any assumptions that you make.

B9. A local bakery uses its stock of sugar according to demand for its products. Indeed, the weekly sugar use follows approximately a normal distribution with mean 2,400 lb. and with standard deviation 400 lb. The starting supply of sugar is 4,000 lb. and there is a scheduled delivery of 2,500 lb at the end of each week. Find the probability that, after 12 weeks (including the delivery that comes at the end of the twelfth week), the supply of sugar will exceed 5,000 lb. (Ignore the very small probability that the sugar runs out completely during this time period.)

As a suggestion, let $X_0, X_1, X_2, \ldots, X_{12}$ be the supply at the ends of the twelve weeks. Certainly $X_0 = 4{,}000$ lb and X_1 is the sum of X_0, the 2,500 lb delivery at the end of week 1, less the random usage during week 1. Similarly, X_2 is the sum of X_1, the 2,500 lb delivery at the end of week 2, less the random usage during week 2.

C2. Thumbs Up Marketing Research is a firm that employs consumer panels to explore preferences for new products. The current inquiry involves a taste test comparison between a standard JELL-O gelatin dessert and a new reduced sugar preparation. Thumbs Up recruits panels of five consumers at a time and determines how many of the five prefer the new product. Their objective is to obtain a panel of five in which all five consumers prefer the new product. The actual probability that an individual will prefer the new product is 0.40. Give an expression for the probability that Thumbs Up will go through 60 panels (each of five people) without finding a single panel of five which will unanimously prefer the new product. (You need not evaluate the expression.)

C5. It has been determined that 40% of all people who walk into a *Dasher* automobile dealership will eventually purchase a *Dasher*.

(a) Among four people walking into a *Dasher* dealership (and presumably a random sample), what is the probability that all four will eventually purchase a *Dasher*?

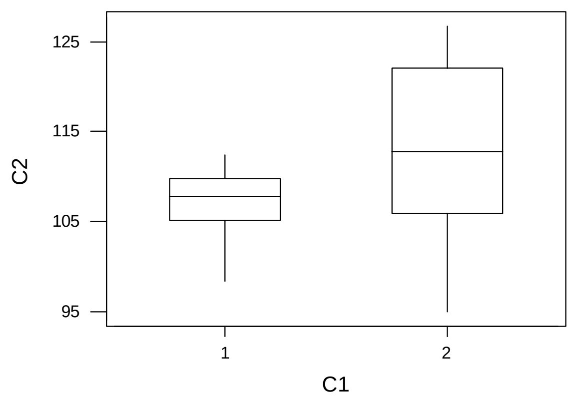
(b) Among 12 people walking into a *Dasher* dealership (and presumably a random sample), is it more likely that four will eventually purchase a *Dasher* or that five will eventually purchase a *Dasher*? (You need a calculation to support the solution.)

(c) Find the smallest sample size *n* for which

P(all *n* eventually purchase a *Dasher*) ≤ 0.10.

(d) Among five randomly-selected people walking into a *Dasher* dealership (and presumably a random sample), what is the probability that *at most one* will *not* eventually purchase a *Dasher*?

C6. This problem considers the comparison of two groups. These boxplots compare the groups



Minitab output related to the comparison is given next.

ITEM 1

Two Sample T-Test and Confidence Interval

```
Two sample T for C2
C1           N      Mean     StDev   SE Mean
1            9    107.06      4.07       1.4
2           17    112.87      9.45       2.3

95% CI for mu (1) - mu (2): ( -11.3,  -0.3)
T-Test mu (1) = mu (2) (vs not =): T = -2.18  P = 0.040  DF = 23
```

ITEM 2

Two Sample T-Test and Confidence Interval

```
Two sample T for C2
C1           N      Mean     StDev   SE Mean
1            9    107.06      4.07       1.4
2           17    112.87      9.45       2.3

95% CI for mu (1) - mu (2): ( -12.7,  1.1)
T-Test mu (1) = mu (2) (vs not =): T = -1.75  P = 0.093  DF = 24
Both use Pooled StDev = 8.07
```

ITEM 3

Homogeneity of Variance

```
ConfLvl     95.0000

Bonferroni confidence intervals for standard deviations
Lower     Sigma     Upper   N  Factor Levels

 2.61139   4.07478    8.6868   9  1
 6.76488   9.45422   15.3695  17  2

F-Test (normal distribution)
Test Statistic: 5.383
P-Value       : 0.021
```

Please provide answers to the following.

(a) State the null hypothesis that is being tested. Be sure to identify any parameters which appear.

(b) State the alternative hypothesis.

(c) What assumptions about the population standard deviations are at work in ITEM 1?

(d) What assumptions about the population standard deviations are at work in ITEM 2?

(e) What conclusion, using $\alpha = 0.05$, about the hypotheses do you make in ITEM 1?

(f) What conclusion, using $\alpha = 0.05$, about the hypotheses do you make in ITEM 2?

(g) Show the calculation that gives the value -2.18 in ITEM 1. That is, indicate a numeric expression which, when simplified, will produce the value -2.18.

(h) ITEM 3 may be somewhat unfamiliar, but see if you can use it to reach a decision as to whether you prefer the assumptions in (c) or the assumptions in (d).

(i) Are there any other assumptions that should be stated in this problem?

(j) Why are the degrees of freedom different in ITEM 1 and ITEM 2 even though the sample sizes are unchanged? You need not show any elaborate formulas.

C9. The following Minitab output represents a "best subsets" run in the regression of YNOT on AL, BE, GA, and DE.

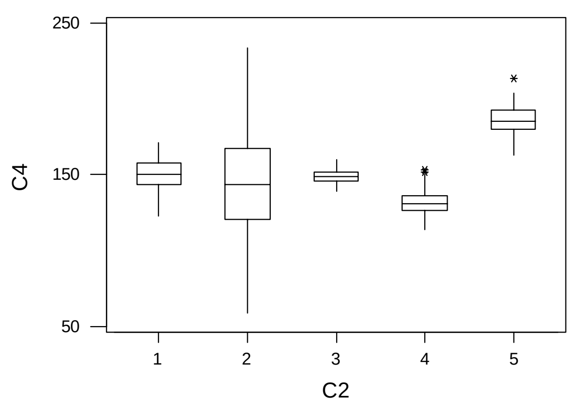
Best Subsets Regression

Response is YNOT

Vars	R-Sq	Adj. R-Sq	C-p	s	A L	B E	G A	D E
1	68.8	68.5	132.1	393.96		X		
1	45.8	45.1	293.8	519.85	X			
2	87.8	87.5	1.5	248.03		X		X
2	68.9	68.2	133.9	396.00		X	X	
3	87.9	87.4	3.1	248.87		X	X	X
3	87.8	87.4	3.4	249.32	X	X		X
4	87.9	87.3	5.0	250.25	X	X	X	X

Based on this information, which of the independent variables are useful in predicting YNOT? On what do you base your opinion?

C10. Here is a display showing five boxplots:



The five variables pictured here have the following selected numerical summaries:

	N	Mean	Median	StDev
ARTICHOKE	123	149.03	148.95	3.83
BERRY	117	150.42	150.32	10.02
COLLAR	124	144.06	143.39	33.30
DIANTHUS	119	185.91	185.44	8.78
EGGPLANT	117	131.93	130.83	8.31

Match the boxplots to the variables.

D2. The voters in a large state are confronted with a decision about voting for Proposition 18. The Attimetrix polling service took a random sample of 720 voters, and found that 450 of them planned to vote "Yes" on Proposition 18.

(a) Based on the results of Attimetrix, give a 95% confidence interval for the population proportion that plan to vote "Yes" on this proposition.

(b) Suppose that Attimetrix tested the null hypothesis H_0: $p = 0.58$ versus H_1: $p \neq 0.58$ at significance level 0.05, based on its sample of 720. Would H_0 be accepted or rejected?

Please note that for this problem, there is not an exact correspondence between the confidence interval and the hypothesis test.

D3. A company's human resources department offers financial planning services, ranging from all the way from pension planning to debt management. This human resources department has kept track of meeting times with employees over the last calendar year. The data below give information on this meeting time for a random sample of 15 secretarial positions and 12 management positions. The individual values represent hours per year using these human resources services.

Group	Number of people	Average time (hours)	Standard deviation of time (hours)
Secretarial	15	6.6	2.1
Management	12	8.0	2.9

You would like to test, at the 0.05 level of significance, whether μ_S (secretarial mean) is equal to μ_M (management mean) or not.

(a) State null and alternative hypotheses appropriate for this situation.

(b) It seems reasonable to assume that the population standard deviations σ_S and σ_M are equal. Find the value of s_p , the pooled standard deviation.

(c) Give the value of the t statistic which tests the hypothesis given in (a). NOTE: If you were not able to get a value for s_p in part (b), use $s_p = 2.5$.

(d) Give the comparison point to which the t statistic should be compared. Indicate whether this causes you to accept or to reject the null hypothesis.

D5. A marketing group would like to locate regular drinkers of the liqueur Campari. The target population is here defined as adults in the city of Philadelphia. This is a situation of detecting events of low probability, as this is not a very popular drink.

(a) Suppose that 0.004 of the adults in Philadelphia are regular drinkers of Campari. What sample size should the marketing group use if it wants a probability of at least 80% of finding at least one Campari drinker in the sample?

(b) Suppose that 0.02 of the adults in Philadelphia are regular drinkers of Campari. What sample size should the marketing group use if it wants a probability of at least 80% of finding at least one Campari drinker in the sample?

D6. Dark corn syrup is particularly difficult to manage, in that the equipment frequently becomes sticky. The management is concerned about the specific gravity of this syrup. A sample of 25 quart containers was taken, and the syrup specific gravity in each was measured. The average quantity in this sample was 1.062, with a standard deviation of 0.015. Assume that the sample comes from a normal population.

(a) Give a 95% confidence interval for the mean specific gravity of this syrup in a quart container.
(b) Provide an interval with the property that about 95% of all quart containers will have specific gravity values within the interval.

D7. A small bank would like to estimate the average monthly dollar value of checks written by its customers. A random sample of 160 of its customers had a mean monthly transaction amount of $1,060, with a standard deviation of $445.
(a) Give a 95% confidence interval for μ.
(b) Identify the meaning of μ in this problem.
(c) State the assumptions behind this confidence interval.
(d) Suppose that the null hypothesis H_0: μ = $1,120 were to be tested at the level 0.05. Based on the confidence interval, would H_0 be accepted or rejected?

D8. The Market Maker Newsletter has just invented the Recreation Index, intended to track the performance of companies involved in sports, leisure, and recreation products. Using market records, they were able to construct a history as to how this index would have performed. It is believed that the weekly changes in the market index have mean 0.06 and standard deviation 0.24. If the current value of the index is 276, find a value c, so that P[Index value after 52 weeks > c] = 0.90.

D9. In multiple regression it is useful to examine the standard deviation ratio $\frac{s_\varepsilon}{s_y}$, where s_y is the standard deviation of the dependent variable and s_ε is the standard error of regression. There is a corrected, or adjusted, form of the R^2 statistic, called R^2_{adj} which uses $\frac{s_\varepsilon}{s_y}$. This adjusted R^2 is defined as $R^2_{\text{adj}} = 1 - \frac{s^2_\varepsilon}{s^2_y}$.

Now consider the following regression analysis of variance table:

Source	Degrees of freedom	Sum Squares	Mean Squares	F
Regression	5	150,000	30,000	30.00
Residual	70	70,000	1,000	
Total	75	220,000		

(a) How many independent variables were used in this regression?
(b) What is the sample size n ?
(c) Find the value of s_ε .
(d) Find the value of s_y .
(e) Give the value of R^2.
(f) Give the value of R^2_{adj} .
(g) Would these regression results be considered significant at the 0.05 level?
(h) True or false…. There is at least one significant t statistic.
(i) True or false…. If Y is measured in units of days, then the units associated with the value 70,000 will be days^2.
(j) True or false…. If the value M is added to all the Y-values, then all the entries of the analysis of variance table will remain unchanged.

D10. If you do a regression of Y on x, with n data points, and if you make the usual assumptions, then the 1 - α prediction interval for a new Y-value based on a new x-value x_{new} is

$$b_0 + b_1\, x_{new} \ \pm\ t_{\alpha/2;n-2}\ s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_{new} - \bar{x}\right)^2}{S_{xx}}}$$

You have sales data for your company's battery-powered Pokémon dolls collected over 11 weeks. You ran the regression of sales (Y) on the week numbers (x) over these 11 weeks. Specifically,

Y_i = sales in week i $\qquad\qquad$ $x_i = i$ = week number

This resulted in the fitted regression $\hat{Y}$ = 240 + 5 x, with an associated standard error of regression s_ε = 7.0.

(a) Note that the x's are the integers 1, 2, …, 11. Please give $\bar{x}$ and then show that S_{xx} = 110.
(b) Please give the point prediction for the 12th week.
(c) Give the 95% prediction interval for the sales for the 12th week.
(d) Please give the point prediction for the 20th week. (This is to be based only on the data that you've seen in weeks 1 through 11. This is a long-range prediction.)
(e) Give the 95% prediction interval for the sales for the 20th week.
(f) The intervals in (c) and (e) will not be equally long. Please indicate which is longer and why it is reasonable that it be longer.

E2. A lognormal random walk model has been assumed for the stock price of Consolidated Food Products, one of the country's heaviest users of artificial food coloring. The model declares that $\log\left(\frac{P_t}{P_{t-1}}\right)$ follows a normal distribution with mean 0 and with standard deviation 0.016. Here P_t is the price at the end of the t^{th} trading day. The model further assumes the statistical independence of each day from all the other days. If the current price is P_0 = \$52.18, find the probability that the price will exceed \$55 after 40 trading days.

F1. During the first three months of 1987, the Downtown Stock Letter made 12 stock purchase recommendations. From the date of recommendation to the end of 1987, these 12 stocks had an average rate of return of -6.5%, with a sample standard deviation of 10.8%. You might remember that 1987 was not a good year. A competing letter, The Wallstreeter, made 10 purchase recommendations during the same period. For these, the average rate of return was -11.5%, with a sample standard deviation of 12.4%. (It can be shown that the pooled standard deviation is 11.55%.) Use the t test (with the 5% level of significance) to compare these two stock letters. Be sure to

(a) state the null and alternative hypotheses, identifying any parameters you create,
(b) give the value of the t statistic,
(c) state the value from the statistical table to which t is to be compared, and
(d) state a conclusion about the two stock letters.

F2. Crates of grapes which arrive at the Eden's Paradise fruit juice factory are supposed to contain 200 pounds. Indeed, the payment schedule is based on this assumption. You are suspicious of fraud, so you select 25 crates at random and weigh them. You find a sample average of 184 pounds, with a standard deviation of 10 pounds. What conclusion do you reach about the grapes?

F3. The "rework" rate at the McKeesport factory of Dynatek, a computer motor manufacturing company, is defined as the proportion of finished items that must be sent to a special room for repairs before being packaged for shipment. Traditionally, this rework rate has been 22%. In November, the McKeesport factory began a program of changes in procurement procedures. The first 1,000 items produced after these changes had a rework rate of 0.194; that is, 194 of 1,000 items required reworking.

At the 5% level of significance, does this represent a significant change from the traditional value of 22%?

Formulate this problem as a statistical hypothesis test, and be sure to state the null and alternative hypotheses, defining any symbols that you create.

F4. For each of the following situations, complete the statement by circling the most appropriate word or phrase.

(a) A 90% confidence interval will be
- shorter than a 95% confidence interval.
- longer than a 95% confidence interval.
- the same length as a 95% confidence interval.

(b) The probability distribution of the number of consumers, out of 40 on a taste-test panel, who prefer Red Dog beer to Miller beer would be modeled as
- Poisson
- binomial
- normal

(c) The probability distribution of the number of defects found on a new automobile, just off the production line would be modeled as
- Poisson
- binomial
- normal

(d) We get to use the normal distribution for sample averages based on 30 or more observations, as justified by
- the Law of Averages
- the Bureau of Standards 1958 report
- the Central Limit theorem

(e) In performing a hypothesis test, the *p*-value has been reported as 0.0241. This means that
- the null hypothesis would be rejected at the 0.05 level of significance
- the alternative hypothesis would be rejected at the 0.05 level of significance
- the probability that the null hypothesis is true is 0.0241

(f) The *t* table stops somewhere between 30 and 40 degrees of freedom because
- the printed page is not large enough to go beyond this many degrees of freedom
- Gosset’s original work on the *t* distribution only developed the theory up through 40 degrees of freedom
- after 40 degrees of freedom, the tabled values pretty much stabilize anyhow

(g) The test, based on a single sample $X_1, X_2, \ldots, X_n$ of measured data, of the hypothesis H_0: $\mu = 410$ versus H_1: $\mu \neq 410$ at significance level 0.05 would

be accepted if the t statistic takes a value between 409 and 411
be accepted if the 95% confidence interval for μ covers the value 410
be accepted if the sample average is more than 2 standard errors away from 410

(h) If, in testing H_0: $\mu = 45$ versus H_1: $\mu \neq 45$, we decided to reject H_0 when $\mu = 45.2$, we would be

making a Type II error
making a Type I error
making a correct decision

(i) If, in testing H_0: $\mu = 2{,}250$ versus H_1: $\mu \neq 2{,}250$, we described the results as "statistically significant" that means that

we accepted the null hypothesis
we rejected the null hypothesis
we got a value of $\overline{X}$ which was likely to prove interesting

(j) If, in testing H_0: $p = 0.40$ versus H_1: $p \neq 0.40$ with regard to a binomial experiment, we obtained a sample fraction $\hat{p}$ = 0.52, then

we would actually need to perform a statistical test to determine if this is significant
we can assert that the difference between 0.40 and p is so large that it is unlikely to be due to chance
we should claim that the results are inconclusive without more data

F5. The following linear regression output has been covered up with a number of blanks. Please fill in these blanks.

Predictor	Coef	SECoef	T	P	VIF
Constant	3571.58	158.567	(a)_________	0.0000	
ACID	0.13099	(b)_______	1.35	0.1801	1.1
BUFFER	-0.56833	(c)_______	-12.45	0.0000	(d)____

S = (e)________ R-Sq = (f)________ R-Sq(adj) = 63.5%

◊◊

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	(g)___	2452000	1226000	(h)_________	0.0000
Residual Error	(i)___	(j)________	13028.0		
TOTAL	107	3820000			

F6. The computer listing shown below is the result of multiple regression. The dependent variable was CRRYCOST, and the independent variables were INVENTRY, REFRIG, LIABLTY, and GUARDS. Questions follow.

Predictor	Coef	SECoef	T	P	VIF
CONSTANT	4387.94	280.510	15.64	0.0000	
INVENTRY	1.14280	0.19971	5.72	0.0000	2.3
REFRIG	0.25193	0.66771	0.38	0.7071	1.2
LIABLTY	0.18350	1.79847	0.10	0.9190	2.3
GUARDS	0.69458	0.30378	2.29	0.0253	1.0

S = 222.783 R-Sq = 0.5632 R-Sq(adj) = 0.5375

Analysis of Variance

SOURCE	DF	SS	MS	F	P
REGRESSION	4	4352000	1088000	21.92	0.0000
RESIDUAL	68	3370000	49632.4		
TOTAL	72	7727000			

CASES INCLUDED 73 MISSING CASES 0

(a) How many data points were used in this analysis?

(b) What is the name of the dependent variable?

(c) How many independent variables were used for this regression?

(d) How many data points were eliminated because of missing data?

(e) Give the fitted regression equation.

(f) The usual regression model contains noise terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. These are assumed to have standard deviation σ. What is the numeric estimate of σ?

(g) What fraction of the variability in the dependent variable is explained by the regression?

(h) What is the standard deviation of the dependent variable?

(i) Would you say that this regression is statistically significant? What statistical fact tells you this?

(j) Suppose that you were to regress INVENTRY on the independent variables (REFRIG, LIABLTY, GUARDS). What would be the value of R^2 for this regression?

F7. The regression that appears below was done as a continuation of the work on the previous problem. The dependent variable was again CRRYCOST. Questions follow.

```
Predictor       Coef          SECoef          T           P       VIF
CONSTANT        4467.62       191.490       23.33       0.0000
INVENTRY        1.17532       0.12884        9.12       0.0000      1.0
GUARDS          0.67601       0.29296        2.31       0.0240      1.0

S = 219.843      R-Sq = 0.5621     R-Sq(adj) = 0.5496

Analysis of Variance

SOURCE        DF       SS           MS         F         P
REGRESSION     2   4344000      2172000    44.94   0.0000
RESIDUAL      70   3383000      48330.8
TOTAL         72   7727000

CASES INCLUDED 73   MISSING CASES 0
```

(a) Given that we already have the results of the previous problem, why was this regression attempted?

(b) Based on this regression, how would you test at the 5% level the null hypothesis H_0: $\beta_{GUARDS} = 0$?

(c) Based on this regression, how would you test at the 5% level the null hypothesis H_0: $\beta_{INVENTRY} = 1$? Please note that the comparison value is 1.

(d) The SS value in the TOTAL line is the same in both regressions. Is there a reason that this has happened, or is this just a lucky result?

(e) Based on this regression, how would we predict the carrying costs for an operation with INVENTRY = 42,000 and GUARDS = 3,500?

F8. A scientist concerned about global warming and polar ice melt has formed a model regarding Y_0 , Y_1 , Y_2 , … where

Y_0 = 24.4 cm is the annual mean high tide elevation for 2006
Y_1 is the annual mean high tide elevation for 2007
Y_2 is the annual mean high tide elevation for 2008
. . . and so on

The measurements are made on a ruler embedded in a sea wall in the Maldive Islands. The model is of course very controversial. This scientist has proposed that the annual changes $X_t = Y_t - Y_{t-1}$ are independent and normally distributed with mean 0.8 cm and with standard deviation 0.6 cm. If this model is correct (and that's a big *if*), find the probability that Y_{44} will exceed 65.0 cm.

F9. You are interviewing applicants, in sequence, for a clerical position. You have three openings that need to be filled, and you will hire each applicant immediately if he or she has a high school diploma. In the population of applications, 90% of the people have high school diplomas. What is the probability that you will need to interview *at most four* people to fill your three positions?

SOLUTIONS:

A1. Marisol's interval did not cover μ. From a procedural point of view, Marisol did not make a mistake. Her "mistake" is completely honest. Marisol was simply using a procedure which works 95% of the time. This time it didn't work.

A2.
(a) Find the position marked —a—. Guess the value that appeared there.
SOLUTION: The variable DIAMOND lies between 31.7 and 62.8, with a mean of 47.724. With a sample size of 49, one would guess that the range is about 4 standard deviations, which would be about 7.5. (The actual value was 5.955.)

(b) Which variable was used as the dependent variable in the regression work?
SOLUTION: SALES

(c) Which pairs of variables have negative correlations?
SOLUTION: These pairs...

COBALT, ALUM
COBALT, BORON
DIAMOND, ALUM
DIAMOND, BORON
SALES, BORON

(d) Find the position marked —d—. What value should appear there?
SOLUTION: This asks for the mean square error. It can be recovered in several ways, but the easiest is probably 953,959 ÷ 44 ≈ 21681.

(e) Find the position marked —e—. What value should appear there?
SOLUTION: In the line for variable COBALT, we have $t = \frac{\text{Coefficient}}{\text{SE}} =$

$\frac{21.778}{\text{- e -}}$ = 9.10. From this we can recover —e— = $\frac{21.778}{9.10} \approx 2.393$.

(f) Following the initial regression run, there is a second regression of SALES on only two of the independent variables, ALUM and COBALT. What is the fitted model in this second regression run?
SOLUTION: $\widehat{\text{SALES}}$ = 288 + 20.3 ALUM + 21.8 COBALT

(g) Why, in your opinion, was this second regression done?
SOLUTION: The first regression showed significant *t* values for ALUM and COBLAT, but non-significant values for the other two variables. The second regression seems to represent the decision to remove those variables with non-significant *t* values.

(h) The BEST SUBSET section shows four models. Which of these four models would provide an adequate fit to the data? (This is not asking you to choose the best model.)
SOLUTION: The model listed first, using only COBALT, would not be acceptable. It has an unfavorable (high) value of C_p, along with values of s_ε and R^2 that can easily be improved. The other three models would all be acceptable, depending on your decisions about cutting off C_p, s_ε, and R^2. Probably most persons would stick with the COBALT, ALUM regression.

(i) The BEST SUBSET section seems to rank the independent variables in terms of their usefulness, from best to worst. What is this ordering?

DISCLAIMER: Not all BEST SUBSET results suggest an ordering. Moreover, the claimed ordering does not reflect any scientific reality.

SOLUTION: The BEST SUBSET work, in this problem at least, gets the independent variables in order. This ordering is COBALT, ALUM, BORON, DIAMOND. But read the disclaimer again.

(j) The STEPWISE section also ranks the independent variables in terms of usefulness. What is the ordering?

DISCLAIMER: Same disclaimer as in (i).

SOLUTION: The STEPWISE section gets the variables in the order COBALT, ALUM and then stops.

A4.
(a) The sample size n = 24 is not large, so we should begin by stating the assumption that the amount follow, at least approximately, a normal distribution. The interval is $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$, which is here $41.6 \pm 2.069 \frac{3.5}{\sqrt{24}}$ or 41.6 ± 1.48. Thus, we're 95% confident that the mean amount is between 40.12 ounces and 43.08 ounces.

(b) About 95% of all the values will be within two standard deviations of the mean. Thus, an interval for this purpose would be $\bar{x} \pm 2s$, or 41.6 ± 7.0, meaning 34.6 ounces to 48.6 ounces. This interval is rather wide. Chauncey should be concerned.

A7.
(a) This problem can be regarded as a test of the null hypothesis $H_0 : \mu = 11$. *True*.

(b) The parameter μ represents the mean weight of the 80 baloney slices in the sample. *False*. The parameter is the true-but-unknown mean fat content in the whole population of these baloney slices.

(c) The t-statistic has 79 degrees of freedom. *True*.

(d) Most published *t* tables have no line for 79 degrees of freedom. *True*. However you can still use such tables, because you can make a crude interpolation using the lines for 60 and 120 degrees of freedom. The infinity line will do fine, also.

(e) At the 5% level of significance, one would accept the hypothesis in (a). *False*. The value of *t* is outrageous.

(f) Though the sampled baloney did have an average fat content above 11 grams, the amount of extra fat seems to be small. *False*. Don't be shy or evasive here. The baloney appears to be 2.5 standard deviations away from the requirement—and 14 grams is not close to 11 grams.

(g) The standard deviation of the amount of fat seems to be close to 1.2 grams. *True*.

(h) Type II error here consists of claiming that the baloney is acceptable when it is really too fatty. *True*.

(i) Given the actual data, it appears that the probability of a type II error is quite small. *True*. The data suggest that the baloney is so fatty that you could not very likely mistake them for adequate.

(j) An approximate 95% prediction interval for the amount of fat in a baloney slice is about 14 grams ± 2.4 grams. *True*. Since the population standard deviation seems to be about 1.2 grams, the interval mean ± 2 SD should capture about 95% of the values. This question is not the same as finding a confidence interval for the population mean μ, which can be trapped in a much shorter interval based on the sample size $n = 80$.

A9. The requirement is apparently $\overline{X} \geq 0.25$ pound. That is, the shipment is accepted if the 10 selected pears have an average weight 0.25 pound or over. This condition is identical to having a total weight over 2.5 pounds. For the sampling distribution of $\overline{X}$, we have $E(\overline{X}) = 0.28$ pound and $SD(\overline{X}) = \frac{0.06}{\sqrt{10}} \approx 0.0190$. Then

$$P[\ \overline{X} \geq 0.25\ ] = P\left[\frac{\overline{X} - 0.28}{0.0190} \geq \frac{0.25 - 0.28}{0.0190}\right] \approx P[\ Z \geq -1.58\ ]$$

$$= 0.50 + P[\ 0 \leq Z \leq 1.58\ ] = 0.50 + 0.4429 = 0.9429 \approx 94\%.$$

It would also be worth the trouble to state the assumption that the ten pears constitute a random sample. It's not that easy to take a genuine random sample of objects that come to you in large piles.

A10.

(a) State the regression model on which this is based.

SOLUTION: This is $\text{RETURN}_i = \beta_0 + \beta_{\text{SIZE}}\ \text{SIZE}_i + \beta_{\text{PARK}}\ \text{PARK}_i + \varepsilon_i$, where the noise terms $\varepsilon_1, \varepsilon_2, \ldots\ \varepsilon_{109}$ are independent with means 0 and equal standard deviations σ_ε. (Other notational schemes would be permitted here, of course.)

(b) What profit would you predict for a store with 10,000 square feet of floor space and parking for 100 cars?

SOLUTION: Find this as

$$14.66 + 31.20\ (10) + 1.12\ (100) = 438.66 ,$$

meaning $438,660.

(c) Which of the estimated coefficients is/are significantly different from zero?

SOLUTION: The constant is $14.66/2.02 \approx 7.26$ standard errors away from zero, so it is certainly significant. (Most people would not even bother about asking for the significance of the constant.)

The coefficient of SIZE is $31.20/4.66 \approx 6.70$ standard errors away from zero, so it is certainly significant.

The coefficient of PARK is $1.12/1.88 \approx 0.60$ standard errors away from zero; it would be judged not statistically significant.

(d) It was noticed that the correlation coefficient between SIZE and PARK was 0.67. Will this present a problem?

SOLUTION: No.

B1.

(a) What fraction of the variability in the Final is explained by the two quizzes?

SOLUTION: This is asking for R^2, which is 53.6%.

(b) Give the fitted regression equation.

SOLUTION: This is

$$\widehat{\text{Final}} = 46.8 + 0.128\ \text{Test1} + 0.646\ \text{Test2}$$

(c) It appears that one of the quizzes was more useful than the other in terms of predicting the final. Which quiz? How do you know?

SOLUTION: Certainly Test2 was more useful. It had a very strong *t* statistic. You should not dismiss Test1 as being useless; it happens that Test1 and Test2 are somewhat collinear, and the conditional role of Test1, given Test2, turns out to be small. Indeed, in part (g), you'll show $r_{\text{Final, Test1}} = 0.60$.

(d) The normal plot of the residuals looks reasonably straight, except for the four oddly-placed points at the bottom left. Locate these four points on the residual-versus-fitted plot. (If you verbally locate these points, that's fine. If not, just circle them on the residual-versus-fitted plot and indicate with letter "d.")
SOLUTION: The four odd points have the lowest residuals. Thus they must be the four lowest points on the residual-versus-fitted plot. (These are at the bottom right.) These four residuals are approximately equal, which explain the flat section on the normal plot.

(e) Is there any useful information on the plot of the residuals versus the order of the data? If yes, tell what that information is. If not, please indicate why not.
SOLUTION: The data points were listed alphabetically by last name. There is no reason at all to even look at this plot.

(f) One of the students got a 45 on the final exam. Locate that student's point on the residual-versus-fitted plot. (If you verbally locate this point, that's fine. If not, just circle it on the residual-versus-fitted plot and indicate with letter "f.")
SOLUTION: This must be the leftmost point on the residual-versus-fitted plot. This has a fitted value of about 60 and a residual of about -15. Thus, 45 = 60 - 15.

(g) One data point is listed as an unusual observation. Give the three test scores for this data point. Indicate also what you think is unusual about this point.
SOLUTION: We see directly from the unusual observation note that this is point 19, with Test1 = 42 and Final = 45. It does not take much detective work to identify Test2 = 12. This is a possibly influential point because of its very unusual value on Test2.

(h) Based on this output, it's possible to find the correlation coefficient between Test1 and Final. What is this correlation?
SOLUTION: This is tricky. The output reveals SS_{total} = 16,243.6 and we then use the Seq SS (sequential sums of squares) section to find that the regression of Final on Test1 (only) would have a regression sum of squares of 5,867.2. Thus, the R^2 for the regression of Final on Test1 (only) would be $\frac{5{,}867.2}{16{,}243.6} \approx 0.3612$. The simple correlation here is the square root of this, namely $\sqrt{0.3612} \approx 0.60$. Of course, it's conceptually possible that this correlation could be the negative square root, namely -0.60, but that's quite unlikely in this context.

(*i*) The *F* value is given as 12.12. Exactly what hypothesis is being tested with this *F* value? What conclusion do you reach about this hypothesis?
SOLUTION: It's testing $H_0 : \beta_{Test1} = 0, \beta_{Test2} = 0$ against the alternative that at least one of these β's is different from zero. With a *p*-value printed as 0.000, we certainly reject H_0.

(j) Find the standard deviation of Final. Using this value make a judgment as to whether the regression was useful.

SOLUTION: This standard deviation is available as $\sqrt{\frac{SS_{total}}{n-1}} = \sqrt{\frac{16{,}243.6}{23}} \approx \sqrt{706.2435} \approx 26.58$. The standard error of regression is s_ε = 18.95. Thus, the noise remaining after the regression was pleasantly less than 26.58, so we'd say that the regression was fairly useful. This is a situation of course where we do not want the regression to be *too useful*. After all, if the final exam was easily predictable, why give it?

B5.

(a) In a single sample $\bar{x}$ = 141.2 and s = -32.1. IMPOSSIBLE. Standard deviations cannot be negative.

(b) In a very strong linear relationship, the correlation was found as r = 1.18. IMPOSSIBLE. Correlations cannot exceed 1.

(c) In a single sample $\bar{x}$ = 15.2 and s = 33.5. COULD HAPPEN.

(d) In a regression of Y on X, we computed s_Y = 21.04 and s_ε = 21.52. COULD HAPPEN. It's far more common to have $s_Y > s_\varepsilon$, but the situation given here is possible.

(e) In a very strong multiple regression, the F statistic was found as F = 12,810. COULD HAPPEN.

(f) A hypothesis test was done at the 5% level of significance, and H_0 was rejected. With the same data, but using the 1% level of significance, H_0 was again rejected. COULD HAPPEN. Indeed rejection at the 1% level implies rejection at the 5% level.

(g) With a single sample of n = 22 values, the 95% confidence interval for μ was found as 85.0 ± 6.2 and the null hypothesis $H_0 : \mu = 92$ (tested against alternative $H_1 : \mu \neq 92$) was accepted at the 5% level of significance. IMPOSSIBLE. There is an exact correspondence for the t test; H_0 is accepted if and only if the comparison value is inside the confidence interval.

(h) 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. COULD HAPPEN. For the binomial situation, there is not necessarily exact agreement between the test and the confidence interval. Here however, the comparison value 0.61 is not even in the interval, so it's not surprising that H_0 was rejected.

(i) 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. COULD HAPPEN. This is just a real-life case of Type II error.

(j) Sherry had a set of data of size n = 6 in which the mean and median were identical. COULD HAPPEN. Here is such a data set: 3 4 5 5 6 7.

(k) 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. COULD HAPPEN. The procedures are slightly different, and it's certainly possible that one could accept and one could reject.

(ℓ) In comparing two samples of continuous data, the standard deviations were found to be $s_x = 42.4$ and $s_y = 47.2$; also, the pooled standard deviation was found to be $s_p = 48.5$. IMPOSSIBLE. The pooled standard deviation must be between s_x and s_y.

(m) Martha had a sample of data $x_1, x_2, \ldots, x_n$ in inches and tested the hypothesis $H_0 : \mu = 10$ inches against the alternative $H_1 : \mu \neq 10$ inches. Edna converted the data to centimeters by multiplying by $2.54 \frac{\text{cm}}{\text{in}}$ and tested $H_0 : \mu = 25.4$ cm against $H_1 : \mu \neq 25.4$ cm. Martha accepted the null hypothesis while Edna (using the same α) rejected it. IMPOSSIBLE. The t test carries consistently across changes of units.

(n) Based on a sample of 38 yes-or-no values, the estimate $\hat{p}$ was found to be 0.41. IMPOSSIBLE: There is no integer x for which $\frac{x}{38}$ would round to 0.41. You can check that $\frac{15}{38} \approx 0.3947 \approx 0.39$ and $\frac{16}{38} \approx 0.4211 \approx 0.42$.

(*o*) In a regression involving dependent variable *Y* and possible independent variables *G*, *H*, *K*, *L*, and *M*, the regression of *Y* on (*G*, *H*, *L*) had R^2 = 53.2% and the regression of *Y* on (*G*, *K*, *M*) had R^2 = 56.4%. COULD HAPPEN. The sets of predictors are not nested, so there is no required ordering on the R^2 values.

(p) (a linear programming problem)

(q) In a regression involving dependent variable *D* and possible independent variables *R*, *S*, and *V*, the *F* statistic was significant at the 5% level while all three *t* statistics were non-significant at the 5% level. COULD HAPPEN. In situations with strong collinearity, this is easily possible.

(r) For a regression with two independent variables, it was found that SS_{regr} = 58,400 and SS_{error} = 245,800. COULD HAPPEN. There are no required interrelationships involving these two quantities.

(s) In a linear regression problem, AGE was used as one of the independent variables, even though AGE = 6 for every one of the data points (which were first-grade children). The estimated coefficient was found to be b_{AGE} = 0.31. IMPOSSIBLE. If all values of AGE are identical, then the regression cannot be performed.

B6. For part (a), we let *X* be the random amount in one bottle. We can ask for the probability that a bottle will be *in* control as

$$P[\, 7.88 \leq X \leq 8.12 \,] \;=\; P\left[\frac{7.88 - 8.00}{0.05} \leq \frac{X - 8.00}{0.05} \leq \frac{8.12 - 8.00}{0.05} \right]$$

$$= \; P[\, -2.40 \leq Z \leq 2.40 \,] \;=\; 2 \times P[\, 0 \leq Z \leq 2.40 \,] \;=\; 2 \times 0.4918 \;=\; 0.9836$$

The probability that the bottle will be out of control is 1 - 0.9836 = 0.0164. This solves (a), and we can call this answer γ.

For (b), the probability of no bottles out of control in 16 inspections is $(1 - \gamma)^{16} = 0.9836^{16} \approx 0.7675$.

For (c), the probability of exactly one bottle out of control in 16 inspections is
$\binom{16}{1}(1-\gamma)^{15}\,\gamma^{1} \;=\; \binom{16}{1} 0.9836^{15}\; 0.0164^{1} \;\approx\; 0.2048.$

For part (d), we can ask for the probability that a bottle will be *in* control as

$$P[\ 7.88 \le X \le 8.12\ ] \ = \ P\left[\frac{7.88-7.98}{0.06} \le \frac{X-7.98}{0.06} \le \frac{8.12-7.98}{0.06}\right]$$

$$\approx \ P[\ -1.67 \le Z \le 2.33\ ] \ = \ P[\ 0 \le Z \le 1.67\ ] \ + P[\ 0 \le Z \le 2.33\ ]$$

$$= \ 0.4525 \ + \ 0.4901 \ = \ 0.9426$$

The probability that the bottle will be declared out of control is then 1 - 0.9426 = 0.0574.

B8. The assumption should be that these are samples from populations with equal standard deviations. The calculation depends on the pooled standard deviation, found as follows:

$$s_p^2 = \frac{(24-1)\times 40^2 \ + \ (30-1)\times 30^2}{24+30-2} \ \approx 1{,}209.6154$$

$$s_p \ = \ \sqrt{1{,}209.6154} \ \approx 34.78$$

It is no surprise that s_p is about halfway between 40 and 30. The professional-minus-academic confidence interval is then

$$210 - 250 \ \pm \ 2.0067 \times \ 34.78\sqrt{\frac{24+30}{24\times 30}}$$

The value 2.0067 is $t_{0.025;52}$. If your t table does not a line for 52 degrees of freedom, you might reasonably use $t_{0.025;\infty}$ = 1.96. Numerically the interval is

$$-40 \ \pm 19.11$$

This is (-59.11, -20.89) or about (-59, -21).

If you had used 1.96, the interval would have been -40 ± 18.67. This works out to (-58.67, -21.33), or about (-59, -21).

B9. The question asks for P[X_{12} > 5,000]. The easiest way to do this is to realize that X_{12} is the sum 4,000 + 12 × 2,500 = 34,000, less the total of 12 weeks of sugar usage. This accounts for X_{12} as the initial 4,000 lb of sugar plus the total of 12 deliveries of 2,500 lb each, less 12 weeks of sugar usage. Following this logic, we write X_{12} = 34,000 - Y, where Y is the total of 12 weeks of sugar usage. Note then that E Y = 12 × 2,400 lb = 28,800 lb and SD(Y) = 400 lb $\sqrt{12}$ ≈ 1,386 lb. Then

$$P[\, X_{12} > 5{,}000 \,] \;=\; P[\, 34{,}000 - Y > 5{,}000 \,] \;=\; P[\, Y < 29{,}000 \,]$$

$$= P\left[\frac{Y - 28{,}800}{1{,}386} < \frac{29{,}000 - 28{,}800}{1{,}386} \right] \approx P[\, Z < 0.14 \,] = 0.5 + P[\, 0 \le Z < 0.14 \,]$$

$$= 0.5 + 0.0557 = 0.5557 \approx 56\%$$

It would be very tricky, by the way, to find the probability that the sugar does not run out at some time during this 12-week period. That probability is certainly small, however.

C2. The probability that a group of five will be unanimous in favor of the new product is $0.40^5 = 0.01024$. Thus, the probability that it will not be unanimous in favor of the new product is $1 - 0.40^5 = 0.98976$. The probability that 60 such panels will be non-unanimous is then $(1 - 0.40^5)^{60} \approx 0.5393$. Thus, there is about a 54% chance that, even after 60 tries, Thumbs Up will be unable to find a unanimous panel.

C5. For part (a), the probability is $0.4^4 = 0.0256 \approx 2.5\%$.

For part (b), you have to compare

$$P[\text{ 4 purchases }] = \binom{12}{4} \times 0.4^4 \times 0.6^8 = 495 \times 0.4^4 \times 0.6^8$$

with

$$P[\text{ 5 purchases }] = \binom{12}{5} \times 0.4^5 \times 0.6^7 = 792 \times 0.4^5 \times 0.6^7$$

Thus we ask $495 \times 0.4^4 \times 0.6^8 \overset{\text{how related}}{?} 792 \times 0.4^5 \times 0.6^7$. By canceling factors 0.4^4 and 0.6^7, we rephrase this as

$$495 \times 0.6 \overset{\text{how related}}{?} 792 \times 0.4$$

The values being compared are 297 (left) and 316.8 (right). The number on the right is bigger, so that five purchases are more likely than four.

In part (c), the probability of n out of n purchases is 0.4^n. We ask for the smallest n for which $0.4^n \le 0.10$. This is easily done by trial-and-error:

$0.4^1 = 0.40$	This exceeds 0.10.
$0.4^2 = 0.16$	This exceeds 0.10.
$0.4^3 = 0.064$	This is below 0.10.

Thus $n = 3$ is the smallest.

In part (d), we have

P[at most one does not purchase] = P[n-1 or n make purchase]

$$= \binom{n}{n-1} \times 0.4^{n-1} \times 0.6 \ + \ 0.4^n$$

$$= \ n \times 0.4^{n-1} \times 0.6 \ + \ 0.4^n \ = \ 0.4^{n-1} (0.6n + 0.4)$$

For $n = 5$, this is $0.4^4 \times 3.4 = 0.08704 \approx 9\%$.

C6.

(a, b) We are testing H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means of the two populations being compared.

(c) ITEM 1 is not making an assumption about σ_1 and σ_2. It is permitting $\sigma_1 \neq \sigma_2$.

(d) ITEM 2 is using the assumption $\sigma_1 = \sigma_2$, where the σ's are the standard deviations of the two populations.

(e) Reject H_0 as the p-value is 0.04.

(f) Accept H_0 as the p-value is 0.093, which exceeds 0.05.

(g) The expression is $\dfrac{107.06 - 112.87}{\sqrt{\dfrac{4.07^2}{9} + \dfrac{9.45^2}{17}}}$.

(h) ITEM 3 is suggesting that you should *not* invoke the assumption $\sigma_1 = \sigma_2$. There is a test provided, and the p-value is 0.021, which is significant. The confidence intervals overlap a little bit, but this does not constitute a test.

(i) As the sample sizes are small, we should invoke the assumption that the samples come from normal populations.

(j) ITEM 2, which uses the pooled standard deviation, gets the degrees of freedom as 9 + 17 - 2 = 24. ITEM 1 uses an approximation formula for an approximate t version of the test statistic.

C9. It appears that you should use the two-predictor model (BE, DE). This is the first model that produces a reasonably small C_p value. Also, the model has an R^2 value which is virtually identical with that for the full model.

C10. In order from 1 to 5, the boxplots are BERRY, COLLAR, ARTICHOKE, EGGPLANT, DIANTHUS.

D2.

(a) The sample proportion for Attimetrix is $\hat{p} = \frac{450}{720} = 0.625$. Then the 95% continuity-corrected confidence interval is

$$0.625 \pm \left[1.96 \sqrt{\frac{0.625 \times 0.375}{720}} + \frac{1}{1,440} \right]$$

which is 0.625 ± 0.0361, approximately. This is (0.5889, 0.6611), or about 59% to 66%.

Most people would now use the Agresti-Coull interval. This starts from $\tilde{p} = \frac{452}{724} \approx 0.6243$, and the interval is given as $\tilde{p} \pm 1.96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$. This works out to 0.6243 ± 0.0353. This is (0.5890, 0.6596).

(b) It would seem that, for the Attimetrix data, that H_0: p = 0.58 would be rejected, as it is outside the confidence interval. However, this is an approximate correspondence, so we ought to check. The test statistic is

$$Z = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} = \sqrt{720} \frac{0.625 - 0.58}{\sqrt{0.58 \times 0.42}} \approx 2.4465$$

This is easily outside the interval (-1.96, 1.96), so we reject the null hypothesis.

D3. Certainly for part (a) we use H_0: $\mu_S = \mu_M$ versus H_1: $\mu_S \neq \mu_M$. There is nothing in this problem which should lead you to a one-sided H_1.

Part (b) asks for s_p. This will be found as

$$s_p = \sqrt{\frac{(15-1) \times 2.1^2 + (12-1) \times 2.9^2}{15+12-2}} = \sqrt{\frac{14 \times 2.1^2 + 11 \times 2.9^2}{25}} \approx 2.4839$$

(c) The t statistic is $t = \sqrt{\frac{15 \times 12}{15+12}} \times \frac{6.6 - 8.0}{2.4839} \approx -1.4553$

(d) The t statistic should be compared to $t_{0.025;25}$ = 2.0595. As | -1.4553 | is not greater than 2.0595, we accept H_0. We do not find a significant difference here between the management people and the secretarial people.

D5. Let p be the proportion of fraudulent claims. The probability that a sample of n turns up no bad cases is $(1 - p)^n$. Thus, the probability that the sample turns up at least one bad (fraudulent) case is $1 - (1 - p)^n$. We want this to be at least 80%.

(a) Using $p = 0.004$, we want $1 - (1 - 0.004)^n = 1 - 0.996^n \geq 0.80$. This statement is $0.996^n \leq 0.20$. This can be solved either by trial and error or by logarithms. In logarithm form, this is

$$n \log 0.996 \leq \log 0.20$$

or

$$n \geq \frac{\log(0.20)}{\log(0.996)}$$

The inequality direction is reversed, as this results from the division by log(0.996), which is negative. Any base can be used for these logarithms. The result comes to $n \geq 401.55$. The marketers will need a sample of size 402 in order to have at least an 80% probability of seeing at least one Campari drinker.

(b) Using $p = 0.02$, we want $1 - (1 - 0.02)^n = 1 - 0.98^n \geq 0.80$. This statement is $0.98^n \leq 0.20$. Now we solve

$$n \geq \frac{\log(0.20)}{\log(0.98)} \approx 79.66$$

The required sample size is at least 80.

D6.

(a) The interval is $1.062 \pm t_{0.025;24} \frac{s}{\sqrt{n}}$, or $1.062 \pm 2.0639 \frac{0.015}{\sqrt{25}}$. This comes down to 1.062 ± 0.0062.

(b) About 95% of all the values will be within two standard deviations of the mean. Thus, an interval for this purpose would be $1.062 \pm (2 \times 0.015)$, or 1.062 ± 0.03. We can state that about 95% of all the containers will have specific gravities in the interval (1.032, 1.092). This interval is distressingly wide. Management should be concerned.

D7.

The confidence interval is certainly $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$, which here is

$$\$1{,}060 \ \pm\ 1.96\ \frac{\$445}{\sqrt{160}}$$

This computes to \$1,060 ± \$68.95. This interval runs from \$991.05 to \$1,128.95.

In this context, μ is the true-but-unknown mean of the population of the monthly check amounts for the customers of this bank.

In setting up this interval, we assume simply that we have a random sample and that this sample represents only a small proportion of the population. As n is quite large, we have no need to assume a normal population.

The value \$1,120 is inside this confidence interval, so we would certainly accept the null hypothesis H_0: μ = \$1,120.

D8. Let the index sequence be T_0 = current value = 276, T_1, T_2, …, T_{52} . Our problem is to find c so that P[$T_{52} > c$] = 0.90.

However, T_{52} is T_0 plus the sum of 52 weekly changes. The sum of 52 random variables with mean 0.06 and standard deviation 0.24 will itself have mean 52 × 0.06 = 3.12 and will have standard deviation $\sqrt{52} \times 0.24 \approx 1.7307$. Thus, T_{52} will have mean 276 + 3.12 = 279.12 and standard deviation 1.7307.

$$P[\ T_{52} > c\ ] \ =\ P\left[\frac{T_{52} - 279.12}{1.7307} > \frac{c - 279.12}{1.7307}\right] \approx P\left[Z > \frac{c - 279.12}{1.7307}\right]$$

Here Z represents a standard normal random variable, and the ≈ (approximate equal) sign indicates that we've used the Central Limit theorem in approximating the distribution of T_{52}.

The normal table tells us that P[$Z > -1.28$] = 0.90. Thus we solve

$$-1.28 \ =\ \frac{c - 279.12}{1.7307}$$

The solution is $c = 279.12 - 1.28 \times 1.7307 \approx 276.90$.

D9.

(a) There are five predictors, or independent variables, used in this regression.

(b) The sample size n is 76.

(c) s_ε is, as always, the square root of the residual mean square. Here that value is $\sqrt{1{,}000} \approx 31.62$.

(d) We can calculate s_y from information in the Total line. This is $s_y = \sqrt{\dfrac{220{,}000}{75}} \approx 54.16$.

(e) $R^2 = \dfrac{SS_{\text{regression}}}{SS_{\text{total}}} = \dfrac{150{,}000}{220{,}000} \approx 0.6818 = 68.18\%$.

(f) $R^2_{\text{adj}} = 1 - \dfrac{s_\varepsilon^2}{s_y^2} = 1 - \left(\dfrac{31.62}{54.16}\right)^2 \approx 0.6591 = 65.91\%$.

(g) With an F statistic of 30, these data would certainly be regarded as significant. Specifically this means that H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ would be rejected.

(h) *False*. Even though the F is large, it can happen in highly collinear cases that none of the individual t statistics will be significant.

(i) *True*. The sums of squares and the mean squares in the analysis of variance table are all in the units of Y^2.

(j) *True*. Adding a number to all the Y-values changes nothing but b_0, the estimate of β_0.

SIDE NOTE: The formula usually given for the adjusted R^2 is

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-1-K} \times \left(1 - R^2\right)$$

This is algebraically identical to $1 - \dfrac{s_\varepsilon^2}{s_y^2}$.

D10.

(a) If the x's are the integers 1, 2, …, 11, then clearly $\bar{x} = 6$. To find S_{xx}, simply do this:

$$S_{xx} = \sum_{i=1}^{11} \left(x_i - \bar{x}\right)^2$$

$$= \sum_{i=1}^{11} \left(x_i - 6\right)^2 = (1 - 6)^2 + (2 - 6)^2 + \ldots + (6 - 6)^2 + \ldots + (11 - 6)^2$$

$$= (-5)^2 + (-4)^2 + \ldots + 0^2 + 1^2 + \ldots + 5^2 = 2\,[\,1^2 + \ldots + 4^2 + 5^2\,] = 110$$

(b) $\hat{Y}_{12} = 240 + 5(12) = 240 + 60 = 300.$

(c) The 95% prediction interval is

$$300 \pm t_{\alpha/2;n\text{-}2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_{new} - \bar{x}\right)^2}{S_{xx}}} \quad \text{or} \quad 300 \pm t_{0.025;9}\, s_\varepsilon \sqrt{1 + \frac{1}{11} + \frac{\left(12-6\right)^2}{110}}$$

This is 300 ± 2.2622 × 7.0 × 1.4182, which is about 300 ± 26.84.

(d) $\hat{Y}_{20} = 240 + 5(20) = 240 + 100 = 340.$

(e) The 95% prediction interval is then

$$340 \pm t_{0.025;9}\, s_\varepsilon \sqrt{1 + \frac{1}{11} + \frac{\left(20-6\right)^2}{110}} \quad \text{or about } 340 \pm 2.2622 \times 7.0 \times 1.6949$$

This becomes about 340 ± 26.84.

(f) Certainly the prediction interval in (c), which is done for week 12, is shorter than that in (d), which is done for week 20. This makes sense. It should be easier to give a prediction for the very next week, rather than for nine weeks into the future.

E2. If we let $X_t = \log\left(\frac{P_t}{P_{t-1}}\right)$, then it's routine to show that $\log\left(\frac{P_{40}}{P_0}\right)$ is the sum of 40 independent normal random variables, each with mean 0 and standard deviation 0.016. Thus, $Y = \log\left(\frac{P_{40}}{P_0}\right)$ will be normally distributed with mean 0 and with standard deviation $0.016\sqrt{40} \approx 0.1012$. Then we associate the event $P_{40} > \$55$ with $Y > \log\left(\frac{\$55}{\$52.18}\right) \approx 0.0526$. We get then P[$P_{40} > \$55$] = P[$Y > 0.0526$] = $P\left[\frac{Y}{0.1012} > \frac{0.0526}{0.1012}\right] \approx$ P[$Z > 0.52$] = 0.3015.

F1.

(a) Test H_0: $\mu_D = \mu_W$ versus H_1: $\mu_D \neq \mu_W$ where μ_D and μ_W are the true means for "Downtown" and "Wallstreeter." Somewhere along the line it would be nice to state the assumption that the data represent random samples from the population of possible predictions. If you use the pooled standard deviation, you are also making the assumption that $\sigma_D = \sigma_W$, where these are the standard deviations of the two populations. The assumption seems harmless here.

(b) You need the t statistic and hence the pooled standard deviation. Here

$$s_p = \sqrt{\frac{(12-1)\times 10.8^2 + (10-1)\times 12.4^2}{12+10-2}} \approx 11.55$$

and this value was given above. The t statistic is found then as

$$t = \sqrt{\frac{12\times 10}{12+10}} \times \frac{(-6.5)-(-11.5)}{11.55} \approx 1.01$$

(c) The two-sided point corresponding to 20 (= 12+10-2) degrees of freedom for a 5% test is $t_{0.025;\,20}$ = 2.086.

(d) The value of t is clearly not significant. We accept the null hypothesis H_0 and say that we cannot find a significant difference between the two newsletters.

F2. Formally, this is a test of H_0: $\mu = 200$ versus H_1: $\mu \neq 200$, where μ is the true-but-unknown mean weight of the crates which are being sent to you. (It would be OK to formulate the alternative as H_1: $\mu < 200$.) The t statistic is

$$t = \sqrt{25}\,\frac{184-200}{10} = -8.00$$

If you do a 5% level test, the comparison value is $t_{0.025;\,24}$ = 2.064. But no matter what standard is used, this value is outrageous. The crates are significantly underweight, and you have every right to be upset. Also, the average shortfall is 16 pounds out of 200, which is 8%.

If you used H_1: $\mu < 200$, the comparison value would be $-t_{0.05;\,24}$ = -1.711. The rule would consist of rejecting H_0 when $t \leq -1.711$.

You could reasonably make a case for giving a confidence interval for μ. Let's suppose that you decide on 95% confidence. This interval would be $\overline{x} \pm t_{\alpha/2;\,n-1}\,\frac{s}{\sqrt{n}}$, and numerically this is $184 \pm 2.064\,\frac{10}{\sqrt{25}}$, giving 184 ± 4.128 or (179.872, 188.128). This comes nowhere close to covering the comparison value 200, so you'll reach the same conclusion about the grapes.

F3. This is a test of the hypothesis H_0: $p = 0.22$ against the alternative hypothesis H_1: $p \neq 0.22$. The symbol p represents the true-but-unknown rework rate under the new changes.

The observed sample fraction is $\hat{p} = \frac{194}{1{,}000} = 0.194$. The test statistic is $Z =$

$\sqrt{n}\,\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} = \sqrt{1{,}000} \times \frac{0.194 - 0.22}{\sqrt{0.22 \times 0.78}} \approx -1.9848$. The null hypothesis is to be rejected at the 5% level of significance if $|Z| \geq z_{0.025} = 1.96$.

In this case, the null hypothesis is rejected, but just barely. We conclude that the changes have made a significant impact on the rework rate; apparently the rework rate has gone down.

It's hard to say if the estimated reduction of 2.6% (from 22% to 19.4%) has any material consequence.

F4.

(a) Choice 1:
A 90% confidence interval will be shorter than a 95% confidence interval.

(b) Choice 2:
The probability distribution of the number of consumers, out of 40 on a taste-test panel, who prefer Red Dog beer to Miller beer would be modeled as binomial

(c) Choice 1:
The probability distribution of the number of defects found on a new automobile, just off the production line, would be modeled as Poisson.

(d) Choice 3:
We get to use the normal distribution for sample averages based on 30 or more observations, as justified by the Central Limit theorem.

(e) Choice 1:
In performing a hypothesis test, the p-value has been reported as 0.0241. This means that the null hypothesis would be rejected at the 0.05 level of significance.

(f) Choice 3:
The t table stops somewhere between 30 and 40 degrees of freedom because after 40 degrees of freedom, the tabled values pretty much stabilize anyhow

(g) Choice 2:

The test, based on a single sample X_1, X_2, …, X_n of measured data, of the hypothesis H_0: μ = 410 versus H_1: μ ≠ 410 at significance level 0.05 would be accepted if the 95% confidence interval for μ covers the value 410.

(h) Choice 3:

If, in testing H_0: μ = 45 versus H_1: μ ≠ 45, we decided to reject H_0 when μ = 45.2, we would be making a correct decision.

(i) Choice 2:

If, in testing H_0: μ = 2,250 versus H_1: μ ≠ 2,250, we described the results as “statistically significant” that means that we rejected the null hypothesis.

(j) Choice 1:

If, in testing H_0: p = 0.40 versus H_1: p ≠ 0.40 with regard to a binomial experiment, we obtained a sample fraction $\hat{p}$ = 0.52, then we would actually need to perform a statistical test to determine if this is significant.

F5.

Predictor	Coef	SECoef	T	P	VIF
Constant	3571.58	158.567	(a) **22.54**	0.0000	
ACID	0.13099	(b) **0.0970**	1.35	0.1801	1.1
BUFFER	-0.56833	(c) **0.0456**	-12.45	0.0000	(d) **1.1**

S = (e) **114.14**   R-Sq = (f) **0.6419**   R-Sq(adj) = 63.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	(g) **2**	2452000	1226000	(h) **94.11**	0.0000
Residual Error	(i) **105**	(j) **1368000**	13028.0		
TOTAL	107	3820000			

The genuinely mysterious item here is (d), the VIF for buffer. Use these facts:

TOLERANCE(ACID) = 1 – R^2(regression of ACID on BUFFER)
TOLERANCE(BUFFER) = 1 – R^2(regression of BUFFER on ACID)
The two R^2 values are the same, since each is the square of $r_{\text{ACID, BUFFER}}$.
TOLERANCE(ACID) = TOLERANCE(BUFFER)

$$\text{VIF} = \frac{1}{\text{TOLERANCE}}$$

In a problem with two independent variables, the VIF values must be the same.

F6.

(a) 73

(b) CRRYCOST

(c) 4

(d) none

(e) The fitted equation can be read from the COEF column. It is

$$\widehat{\text{CRRYCOST}} = 4{,}387.94 + 1.14280 \text{ INVENTRY} + 0.25193 \text{ REFRIG}$$
$$+ 0.18350 \text{ LIABLTY} + 0.69458 \text{ GUARDS}$$

(f) This is s_ε (and noted as S by Minitab); the value is 222.783.

(g) This is R^2 = 0.5632 = 56.32%.

(h) This can be obtained from the TOTAL line of the analysis of variance table. The value is $\sqrt{\frac{7{,}727{,}000}{72}} \approx 327.60$.

(i) The regression is statistically significant, as evidenced by the F statistic of 21.92, with associated p-value printed as 0.0000.

(j) This fact can be pulled from the VIF for the variable INVENTRY. Recall that $\text{VIF} = \frac{1}{\text{TOLERANCE}}$. Thus, TOLERANCE(INVENTRY) = $\frac{1}{2.3} \approx 0.4348$.
Since TOLERANCE(INVENTRY) =
1 - R^2(regression of INVENTRY on (REFRIG, LIABLTY, GUARDS))
we decide that the R^2 must be 0.5652.

F7.

(a) The regression from the previous problem had non-significant coefficients for REFRIG and LIABLTY, and it would seem that this regression was done to examine the effects of eliminating those two variables.

(b) Just use the *t* statistic given for GUARDS. The value is $t = 2.31$, and the *p*-value is listed as 0.0240. Certainly H_0 would be rejected; we believe the true value is *not* zero.

(c) The test statistic would be $\frac{\text{Estimated statistic - Comparison value}}{\text{SE}\left(\text{Estimated statistic}\right)}$, and the value is $\frac{1.17532 - 1}{0.12884} = \frac{0.17532}{0.12884} \approx 1.3608$. This would follow the *t* distribution with 70 degrees of freedom. (The degrees of freedom number comes from the RESIDUAL line of the analysis of variance table.) The comparison point would $t_{0.025;\,70}$. We don't have this value exactly, but it's somewhere around 2. Here H_0 would be accepted, and we would say that the estimated coefficient (1.17532) is not significantly different from 1.

(d) Since $SS_{TOTAL} = S_{yy} = \sum_{i=1}^{73}\left(Y_i - \overline{Y}\right)^2$, its computation does not use the independent variables at all. The value of SS_{TOTAL} will be the same in every regression using CRRYCOST as the dependent variable.

(e) The fitted equation is

$$\text{CRRY}\hat{\text{C}}\text{OST} = 4{,}467.62 + 1.17532 \text{ INVENTRY} + 0.67601 \text{ GUARDS}$$

The prediction at this new point would be

$$\text{CRRY}\hat{\text{C}}\text{OST}_{\text{new}} = 4{,}467.62 + 1.17532 \times 42{,}000 + 0.67601 \times 3{,}500$$

$$= 4{,}467.62 + 49{,}363.44 + 2{,}366.035 = 56{,}197.095$$

F8. This is a random walk in which $Y_{44} = Y_0 + (X_1 + X_2 + \ldots + X_{44})$. Let T be the total $X_1 + X_2 + \ldots + X_{44}$. Then $Y_{44} = Y_0 + T$. In this model, T is normally distributed with mean $44 \times 0.8 = 35.2$ and with standard deviation $0.6\sqrt{44} \approx 3.9900$. Then find

$$P[\ Y_{44} > 65.0\ ] = P[\ Y_0 + T > 65.0\ ] = P[\ 24.4 + T > 65.0\ ] = P[\ T > 40.6\ ]$$

$$= P\left[\frac{T - 35.2}{3.99} > \frac{40.6 - 35.2}{3.99}\right] \approx P[\ Z > 1.35\ ] = 0.50 - 0.4115 = 0.0885 \approx 9\%$$

F9. There are several ways to conceptualize this problem. The simplest is to create random variable X which is the number of qualified people among the first four. Then X is a binomial random variable with $n = 4$ and $p = 0.90$. Your condition is satisfied if $X \geq 3$. Find the probability as

$$P[\ X \geq 3\ ] = P[\ X = 3\ ] + P[\ X = 4\ ]$$

$$= \binom{4}{3} 0.90^3 \times 0.10^1 + \binom{4}{4} 0.90^4 = 4 \times 0.729 \times 0.10 + 0.6561$$

$$= 0.9477$$

Of course, you might be successful on the first three interviews and not need the fourth person. The random variable X exists conceptually whether you interview the fourth person or not!

You can also, with more difficulty, conceptualize the needed event as

$$\{ \text{ first three good } \} \cup \{ \text{ two of first three good} \cap \text{fourth good } \}$$

The probability for this is $0.90^3 + \left\{ \binom{3}{2} 0.90^2 \times 0.10^1 \times 0.90 \right\}$

$= 0.90^3 + 3 \times 0.90^3 \times 0.10 = 0.90^3 [\ 1 + 0.3\ ] = 0.9477$