Section 9.5

24. In testing $H_o : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$ with $\nu_1 = 4$ and $\nu_2 = 6$ degrees of freedom, if the test statistic value f = 4.53, then P-value = __________.

ANSWER: .05

75. The sample standard deviation of sodium concentration in whole blood (mEq/L) for m = 20 marine eels was found to be $s_1 = 40.5$, whereas the sample standard deviation of concentration for n = 20 freshwater eels was $s_2 = 32.5$. Assuming normality of the two concentration distributions, test at level .10 to see whether the data suggests any difference between concentration variances for the two types of eels.

ANSWER:
$H_o : \sigma_1 = \sigma_2$ will be rejected in favor of $H_a : \sigma_1 \neq \sigma_2$ if either $f \geq F_{.05,19,19} \approx 2.18$ or if $f \leq \frac{1}{2.18} = .459$. Since $f = \frac{(40.5)^2}{(32.1)^2} = 1.59$, which is neither ≥ 2.18 nor $\leq .459$, H_o is not rejected. The data does not suggest a difference in the two variances.

76. In a study of copper deficiency in cattle, the copper values (*ug* Cu/100mL blood) were determined both for cattle grazing in an area known to have well-defined molybdenum anomalies (metal values in excess of the normal range of regional variation) and for cattle grazing in a nonanomalous area, resulting in $s_1 = 21.5$ (m = 48) for the anomalous condition and $s2 = 19.45$ (n = 45) for the nonanomalous condition. Test for the equality versus inequality of population variances at significance level .10 by using the P-value approach.

ANSWER:
$H_o : \sigma_1 = \sigma_2$ will be rejected in favor of $H_a : \sigma_1 \neq \sigma_2$ if either $f \leq F_{.975,47,44} \approx .56$ or if $f \geq F_{.025,47,44} \approx 1.8$. Because f = 1.22, H_o is not rejected. The data does not suggest a difference in the two variances.

Section 10.1

3. An experiment is conducted to study the effectiveness of three teaching methods on student performance. In this experiment, the factor of interest is __________, and there are __________ different levels of the factor.

ANSWER: teaching method, three

5. Single-factor ANOVA focuses on a comparison of more than two population or treatment __________.

ANSWER: means

6. In a one-way ANOVA problem involving four populations or treatments, the null hypothesis of interest is H_o : __________.

ANSWER: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

7. In single-factor ANOVA, the __________ is a measure of between samples variation, and is denoted by __________.

ANSWER: mean square between groups, MSA

8. In single-factor ANOVA, the __________ is a measure of within-samples variation, and is denoted by __________.

ANSWER: mean square for error, MSE or MSW

9. In one-factor ANOVA, both mean square for treatments (MSA) and mean square for error (MSE) are unbiased estimators for estimating the common population variance σ^2 when __________, but MSA tends to overestimate σ^2 when __________.

ANSWER: H_o is true, H_o is false

10. In single-factor ANOVA, SST – SSA= __________.

ANSWER: SSE

11. In one-factor ANOVA, __________ denoted by __________ is the part of total variation that is unexplained by the truth or falsity of H_o.

ANSWER: sum square for error, SSE

12. In one-factor ANOVA, __________ denoted by __________ is the part of total variation that can be explained by possible differences in the population means.

ANSWER: sum square for treatments, SSA

13. Let *F* =MSTr/MSE be the test statistic in a single-factor ANOVA problem involving four populations or treatments with a random sample of six observations from each one. When H_o is true and the four population or treatment distributions are all normal with the same variance σ^2, then *F* has an *F* distribution with degrees of freedom $\nu_1 =$ __________ and $\nu_2 =$ __________. With *f* denoting the computed value of *F*, the rejection region for level .05 test is __________.

ANSWER: 3, 20, $f \geq 3.10$

23. In a single-factor ANOVA problem involving five populations or treatments, which of the following statements are true about the alternative hypothesis?

A. All five population means are equal.
B. All five population means are different.
C. At least two of the population mean are different.
D. At least three of the population mean are different.
E. At most, two of the population means are equal.

ANSWER: C

24. Which of the following statements are true?

A. In some experiments, different samples contain different numbers of observations. However, the concepts and methods of single-factor ANOVA are most easily developed for the case of equal sample sizes.
B. The population or treatment distributions in single-factor ANOVA are all assumed to be normally distributed with the same variance σ^2.
C. In one-way ANOVA, if either the normality assumption or the assumption of equal variances is judged implausible, a method of analysis other than the usual *F* test must be employed.
D. The test statistic for single-factor ANOVA is *F* = MSA/MSE, where MSA is the mean square for treatments, and MSE is the mean square for error.
E. All of the above statements are true.

ANSWER: E

25. In single-factor ANOVA, MSA is the mean square for treatments, and MSE is the mean square for error. Which of the following statements are not true?

A. MSE is a measure of between-samples variation.
B. MSE is a measure of within-samples variation.
C. MSA is a measure of between-samples variation.
D. The value of MSA is affected by the status of H_o (true or false).
E. All of the above statements are true

ANSWER: A

27. In a single-factor ANOVA problem involving four populations or treatments, the four sample standard deviations are 25.6, 30.4, 28.7, and 32.50. Then, the mean square for error is

A. 29.3
B. 117.2
C. 864.865
D. 29.409
E. None of the above answers are correct.

ANSWER: C

29. In one-way ANOVA, which of the following statements are true?

A. SST is a measure of the total variation in the data.
B. SSE measures variation that would be present within treatments even if H_o were true, and is thus the part of total variation that is *unexplained* by the truth or falsity of H_o.
C. SSA is the amount of variation between treatments that can be *explained* by possible differences in the population or treatments' means.
D. If explained variation is large relative to unexplained variation, then H_o is rejected in favor of H_a.
E. All of the above statements are true.

ANSWER: E

42. In an experiment to compare the tensile strengths of *I* = 5 different types of copper wire, *J* = 4 samples of each type were used. The between-samples and within-samples estimates of σ^2 were computed as MSA = 2578.9 and MSE = 1394, respectively. Use the *F* test at level .05 to test $H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ versus H_a : at least two μ_i's are unequal.

ANSWER:
H_o will be rejected if $f \geq F_{.05,4,15} = 3.06$ (since I – 1 = 4, and I (J – 1) = (5)(3) = 15). The computed value of *F* is $f = \frac{2573.3}{1394.2} = 1.85$. Since 1.85 is not $\geq 3.06, H_o$ is not rejected. The data does not indicate a difference in the mean tensile strengths of the different types of copper wires.

43. The lumen output was determined for each of *I* = 3 different brands of 60-watt soft-white light bulbs, with *J* = 8 bulbs of each brand tested. The sums of squares were computed as *SSE* = 4725 and *SSA* = 590. State the hypotheses of interest (including word definitions of parameters), and use the *F* test of ANOVA $(\alpha = .05)$ to decide whether there are any differences in true average lumen outputs among the three brands for this type of bulb by obtaining as much information as possible about the *P*-value.

ANSWER: (note: MSTr = MSA)
With μ_i = true average lumen output for brand *i* bulbs, we wish to test

$H_{:\mu_1} = \mu_2 = \mu_3$ versus H_a : at least two μ_i's are unequal.

$$MSTr = \hat{\sigma}_B^2 = \frac{590}{2} = 295, \quad MSE = \hat{\sigma}_W^2 = \frac{4725}{21} = 225, \text{so}$$

$f = \frac{MSTr}{MSE} = \frac{295}{225} = 1.31$ For finding the p-value, we need degrees of freedom (I – 1) = 2 and I $(J-1) = 21$. Since $f = 1.31 < F_{.10,2,21} = 2.57$, then the p-value $> .10$. Since .10 is not $< .05$, we cannot reject H_o. There are no differences in the average lumen outputs among the three brands of bulbs.

44. In an experiment to investigate the performance of four different brands of spark plugs intended for use on a 125-cc two-stroke motorcycle, five plugs of each brand were tested and the number of miles (at a constant speed) until failure was observed. The partial ANOVA table for the data is given below. Fill in the missing entries, state the relevant hypotheses, and carry out a test by obtaining as much information as you can about the *P*-value.

Source	**df**	**SS**	**MS**	***f***
Brand				
Error			14,700	
Total		310,200		

ANSWER:

Source	**df**	**SS**	**MS**	***f***
Treatments	3	75,000	25,000	1.70
Error	16	235,200	14,700	
Total	19	310,200		

The hypotheses are $H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. $H_a :$ at least two μ_i's are unequal.
$1.70 < F_{.10,3,16} = 2.46$, so p-value > .10, and we fail to reject H_o.

45. Six samples of each of four types of cereal grain grown in a certain region were analyzed to determine thiamin content, resulting in the following data (*ug*/*g*):

Wheat	5.5	4.8	6.3	6.4	7.1	6.1
Barley	6.8	8.3	6.4	7.8	6.2	5.9
Maize	6.1	5.0	6.7	5.2	6.3	5.5
Oats	8.6	6.4	8.1	7.3	5.8	7.5

Does this data suggest that at least two of the grains differ with respect to true average thiamin content? Use a level $\alpha = .05$ based on the *P*-value.

ANSWER:

Source	**df**	**SS**	**MS**	***f***
Treatments	3	8.983	2.996	3.958
Error	20	15.137	.757	
Total	23	24.12		

Since $3.10 = F_{.05,3,20} < 3.958 < 4.94 = F_{.01,3,20}, .01 < p-value < .05$ and H_o is rejected at level .05.

Section 10.2

36. Consider calculating a 95% confidence interval for a population mean μ based on a sample from a population, and then a 95% confidence interval for a population proportion *p* based on another sample selected independently from the same population. Which of the following statements are true?

A. Prior to obtaining data, the probability that the first interval will include μ is .95.
B. Prior to obtaining data, the probability that the second interval will include *p* is .95.
C. The probability that *both* intervals will include the values of the respective parameters is about .90.
D. All of the above statements are correct.

ANSWER: D

37. If three 90% confidence intervals for a population proportion *p* are calculated based on three independent samples selected randomly from the population, then the simultaneous confidence level will be

A. about 73%
B. exactly 90%
C. exactly 81%
D. exactly 270%
E. None of the above answers are correct.

ANSWER: A

38. The assumptions of single-factor ANOVA can be described succinctly by means of the "model equation" $X_{ij} = \mu_i + \varepsilon_{ij}$ where ε_{ij} represents a random deviation from the population or true treatment mean μ_i. Which of the following statements are true?

A. The ε_{ij}'s are assumed to be independent.
B. The ε_{ij}'s are normally distributed random variables.
C. $E(\varepsilon_{ij}) = 0$ for every i and *j*.
D. $V(\varepsilon_{ij}) = \sigma^2$ for every i and *j*.
E. All of the above statements are true.

ANSWER: E

39. Which of the following statements are not true?

A. ANOVA can be used to test $H_o : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$.
B. ANOVA cannot be used to test $H_o : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$.
C. ANOVA can be used to test $H_o : \mu_1 = \mu_2 = \mu_3$ versus H_a : at least two of the μ_i's are different.
D. The two-sample *t* test can be used to test $H_o : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$.
E. All of the above statements are true.

ANSWER: B

40. Which of the following statements are not true?

A. The two-sample *t* test is more flexible than the *F* test when the number of treatments or populations is 2.
B. The two-sample *t* test is valid without the assumption that the two population variances are equal.
C. The two-sample *t* test can be used to test $H_a : \mu_1 > \mu_2$ or $H_a : \mu_1 < \mu_2$ as well as $H_a : \mu_1 \neq \mu_2$.
D. The *F* test can be used to test $H_a : \mu_1 > \mu_2$ or $H_a : \mu_1 < \mu_2$ as well as $H_a : \mu_1 \neq \mu_2$.
E. When the number of treatments or populations is at least 3, there is no general test procedure known to have good properties without assuming equal populations variances.

ANSWER: D

41. In a single-factor ANOVA problem involving 4 populations, the sample sizes are 7,5,6, and 6. If *SST* = 65.27 and *SSA* = 23.49, then the test statistic value *f* is

A. 3.75
B. 2.09
C. 7.83
D. 0.56
E. 6.67

ANSWER: A

50. The following data refers to yield of tomatoes (kg/plot) for four different levels of salinity: salinity level here refers to electrical conductivity (EC), where the chosen levels were EC = 1.6, 3.8, 6.0, and 10.2 nmhos/cm:

1.6	59.8	53.6	57.1	63.4	59.0
3.8	55.5	59.4	53.2	54.8	
6.0	52.0	49.2	54.1	49.3	
10.2	44.9	48.8	41.3	48.0	46.4

Use the *F* test at level $\alpha = .05$ to test for any differences in true average yield due to the different salinity levels.

ANSWER:

Source	**df**	**SS**	**MS**	***f***
Treatments	3	456.505	152.168	17.11
Error	14	124.498	8.893	
Total	17	581.003		

Since $17.11 \ge F_{.05,3,14} = 3.34$, $H_0 : \mu_1 = ... = \mu_4$ is rejected at level .05. There is a difference in yield of tomatoes for the four different levels of salinity.

51. The following partial ANOVA table is taken from a study in which the abilities of three different groups to identify a perceptual incongruity were assessed and compared. All individuals in the experiment had been hospitalized to undergo psychiatric treatment. There were 21 individuals in the depressive group, 32 individuals in the functional "other" group, and 21 individuals in the brain-damaged group. Complete the ANOVA table and carry out the *F* test at level $\alpha = .01$.

Source	**df**	**SS**	**MS**	***f***
Groups			75	
Error				
Total		1144		

ANSWER:

Source	**df**	**SS**	**MS**	***f***
Groups	2	150	75	5.36
Error	71	994	14	
Total	73	1144		

Since $5.36 \ge F_{.01,2,71} \approx 4.94$, reject $H_o : \mu_1 = \mu_2 = \mu_3$ at level .01.

Section 11.1

7. The parameters for the fixed effects model with interaction are $\alpha_i = \mu_{i.} - \mu, \beta_j = \mu_{.j} - \mu,$ and $\gamma_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$. Thus the model is $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. The α_i's are called the __________ for factor *A*, whereas the β_j's are the __________ for factor *B*. The γ_{ij}'s are referred to as the __________ parameters.

ANSWER: main effects, main effects, interaction

24. Which of the following statements are not true?

A. The model specified by $X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$ and $\mu_{ij} = \alpha_i + \beta_j (i = 1,\ldots,I$ and $j = 1,\ldots,J)$ is called an additive model.

B. The model $X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ where $\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0,$ and the ε_{ij}'s are assumed independent, normally distributed with mean 0 and common variance σ^2 is an additive model in which the parameters are uniquely determined.

C. In two-way ANOVA, when the model is additive, additivity means that the difference in mean responses for two levels of one of the factors is the same for all levels of the other factor.

D. All of the above statements are true.

E. None of the above statements are true.

ANSWER: D

26. In a two-factor experiment where factor *A* consists of 4 levels, factor *B* consists of 3 levels, and there is only one observation on each of the 12 treatments, which of the following statements are not true?

A. *SST* has 12 degrees of freedom

B. *SSA* has 3 degrees of freedom

C. *SSB* has 2 degrees of freedom

D. *SSE* has 6 degrees of freedom

E. None of the above statements are correct.

ANSWER: A

30. In the fixed effects model with interaction, assume that there are 5 levels of factor *A*, 4 levels of factor *B*, and 3 observations (replications) for each of the 20 combinations of levels of the two factors. Then the number of degrees of freedom of the interaction sum of squares (*SSAB*) is

A. 60

B. 20

C. 15

D. 12

E. 59

ANSWER: D

35. The following equation *SST* = *SSA* + *SSB* +*SSAB* +*SSE* applies to which ANOVA model?

A. One-factor ANOVA

B. Two-factor ANOVA with interaction

C. Three-factor ANOVA

D. Randomized block design

E. All of the above

ANSWER: B

47. The number of miles useful tread wear (in 1000's) was determined for tires of five different makes of subcompact car (factor A, with $I = 5$) in combination with each of four different brands of radial tires (factor B, with $J = 4$), resulting in $IJ = 20$ observations. The values SSA = 30, SSB = 45, and SSE = 60 were then computed. Assume that an additive model is appropriate.

a. Test $H_o : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ (no differences in true average tire lifetime due to makes of cars) versus H_a : at least one $\alpha_i \neq 0$ using a level .05 test.

b. $H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (no differences in true average tire lifetime due to brands of tires) versus H_a : at least one $\beta_\mathrm{j} \neq 0$ using a level .05 test.

ANSWER:

a. $MSA = \frac{30}{4} = 7.50$, $MSE = \frac{60}{12} = 5.0$, $f_A = \frac{7.50}{4.93} = 1.50$. Since 1.50 is not $\geq F_{.05,4,12} = 3.26$, don't reject H_{oB}. There is no difference in true average tire lifetime due to different makes of cars.

b. $MSB = \frac{45}{3} = 15.0$, $f_B = \frac{15.0}{5.0} = 3.0$. Since 3.0 is not $\geq F_{.05,3,12} = 3.49$, don't reject H_{oB}. There is no difference in true average tire lifetime due to different brands of tires.

48. In an experiment to see whether the amount of coverage of light-blue interior paint depends either on the brand of paint or on the brand of roller used, 1 gallon of each of four brands of paint was applied using each of three brands of roller, resulting in the following data (number of square feet covered).

		Roller Brand		
		1	**2**	**3**
Paint Brand	**1**	404	396	401
	2	396	394	397
	3	389	392	394
	4	394	387	393

a. Construct the ANOVA table.
b. State and test hypotheses appropriate for deciding whether paint has any effect on coverage. Use $\alpha - .05$.
c. Repeat part (b) for brand of roller.
d. Use Tukey's method to identify significant differences among brands. Is there one brand that seems clearly preferable to the others?

ANSWER:

a

Source	**df**	**SS**	**MS**	***f***	$f_{.05}$
A	3	159.58	53.19	7.85	4.76
B	2	38.00	19.00	2.80	5.14
Error	6	40.67	6.78		
Total	11	238.25			

b. Since $7.85 \geq 4.76$, reject $H_{0A} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$: The amount of coverage depends on the paint brand.

c. Since 2.80 is not ≥ 5.14, do not reject $H_{oA} : \beta_1 = \beta_2 = \beta_3 = 0$. The amount of coverage does not depend on the roller brand.

d. Because H_{oB} was not rejected. Tukey's method is used only to identify differences in levels . of factor A (brands of paint). $Q_{.05,4,6} = 4.90, w = 7.37$.

i:	4	3	2	1
$\bar{x}_{ig}$:	231.75	325.25	441.00	613.25

50. The strength of concrete used in commercial construction tends to vary from one batch to another. Consequently, small test cylinders of concrete sampled from a batch are “cured” for periods up to about 28 days in temperature- and moisture-controlled environments before strength measurements are made. Concrete is then “bought and sold on the basis of strength test cylinders”. The accompanying data resulted from an experiment carried out to compare three different curing methods with respect to compressive strength (MPa). Analyze this data.

Batch	Method A	Method B	Method C
1	30.2	33.2	30.0
2	28.6	30.1	32.1
3	29.5	31.7	30.0
4	31.4	34.1	33.0
5	30.0	32.5	31.9
6	26.4	28.8	27.3
7	27.7	27.9	30.2
8	31.9	31.9	33.1
9	26.1	29.0	28.7
10	28.1	28.9	32.7

ANSWER:

Source	df	SS	MS	*f*
Method	2	23.23	11.61	8.69
Batch	9	86.79	9.64	7.22
Error	18	24.04	1.34	
Total	29	134.07		

$F_{.01,2,18} = 6.01 < 8.69 < F_{.001,2,18} = 10.39$, so $.001 < p - value < .01$, which is significant. At least two of the curing methods produce differing average compressive strengths. (With *p*-value < .001, there are differences between batches as well.)

$Q_{.05,3,18} = 3.61;\ w = (3.61)\sqrt{\frac{1.34}{10}} = 1.32$

Method A	Method B	Method C
28.99	30.81	30.90

Methods B and C produce strengths that are not significantly different, but Method A produces strengths that are different (less) than those of both B and C.

Section 14.1

12. In a two-way contingency table, if the second row total is 125, the third column total is 60, and the total number of observations is 375, then the estimated expected count in cell (2, 3) is __________.

ANSWER: 20

13. A two-way contingency table has 3 rows and 5 columns. Then, the number of degrees of freedom associated with the chi-squared test for homogeneity is __________.

ANSWER: 8

15. A two-way contingency table has r rows and c columns. Then, the number of degrees of freedom associated with the chi-squared test for independence is __________.

ANSWER: (r-1)(c-1)

25. The number of degrees of freedom for a two-way contingency table with *I* rows and *J* columns is

A. $I \cdot J$
B. $(I-1) \cdot J$
C. $I \cdot (J-1)$
D. $(I-1) \cdot (J-1)$
E. $I + J - 1$

ANSWER: D

26. In a two-way contingency table with 3 rows and 5 columns, assume that the second row total is 120 and the fourth column total is 50, and the total number of observations is 600. Then, the estimated expected count in cell (2, 4) is

A. 50
B. 40
C. 30
D. 20
E. 10

ANSWER: E

29. The number of degrees of freedom in testing for independence when using a contingency table with 6 rows and 4 columns is:

A. 24
B. 10
C. 15
D. 20
E. 12

ANSWER: C

32. A statistics department at a state university maintains a tutoring service for students in its introductory service courses. The service has been staffed with the expectation that 40% of its students would be from the business statistics course, 30% from engineering statistics, 20% from the statistics course for social science students, and the other 10% from the course for agriculture students. A random sample of n=120 students revealed 50, 40, 18, and 12 from the four courses. Does this data suggest that the percentages on which staffing was based are not correct? State and test the relevant hypotheses using $\alpha = .05$.

ANSWER:

Using the number 1 for business, 2 for engineering, 3 for social science, and 4 for agriculture, let $p_i =$ the true proportion of all clients from discipline i. If the Statistics department's expectations are correct, the relevant null hypothesis is $H_0 : p_1 = .35, p_2 = .30, p_3 = .20, p_4 = .15,$ versus H_a : The Statistics department's expectations are not correct. With d.f = k-1=4-1=3, we reject H_0 if $\chi^2 \geq \chi^2_{.05,3} = 7.815$.

Cell	1	2	3	4
Observed (n_i)	50	40	18	12
Expected $(n_i p_i)$	42	36	24	18
X^2 term	1.524	.444	1.5	2.0

Since all the expected counts are at least 5, the chi-squared test can be used. The value of the test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{all cells} \frac{(observed - \exp ected)^2}{\exp ected} = 1.524 + .444 + 1.5 + 2.0 = 5.468$$

Since $X^2 = 5.468$ is not ≥ 7.815, we fail to reject H_0. (alternatively, p-value = $P(\chi^2 \geq 1.57)$ which is >.10, and since the p-value is not < .05, we reject H_0). Thus we have no evidence to suggest that the statistics department's expectations are incorrect.

41. A study reports on research into the effect of different injection treatments on the frequencies of audiogenic seizures.

Treatment	No response	Wild running	Clonic seizure	Tonic seizure
Thienylalanine	22	8	25	45
Solvent	14	14	20	52
Sham	22	10	23	45
Unhandled	47	13	28	32

Does the data suggest that the true percentages in the different response categories depend on the nature of the injection treatment? State and test the appropriate hypotheses using $\alpha = .005$.

ANSWER:
With p_{ij} denoting the probability of a type *j* response when treatment is applied, $H_o : p_{1j} = p_{2j} = p_{3j} = p_{4j}$ for *j* = 1, 2, 3, 4 will be rejected at level .005 if $\chi^2 \geq \chi^2_{.005,9} = 23,587$.

$\hat{E}_{ij}$	1	2	3	4
1	25.0	10.71	22.86	41.43
2	25.0	10.71	22.86	41.43
3	25.0	10.71	22.86	41.43
4	30.0	12.86	27.43	49.71

The computed value of the test statistic χ^2 is

χ^2 = 0.36 + 0.69 + 0.20 + 0.31 + 4.84 + 1.01 + 0.36 + 2.70 + 0.36 + 0.05 + 0.00 + 0.31 +
9.63 + 0.00 + 0.01 + 6.31 = 27.14

Since $\chi^2 = 27.14 \geq 23.587$, so reject H_o at level .005

42. Each individual in a random sample of high school and college students was cross-classified with respect to both political views and marijuana usage, resulting in the data displayed in the accompanying two-way table. Does the data support the hypothesis that political views and marijuana usage level are independent within the population? Test the appropriate hypotheses using level of significance .01.

	Usage Level		
Political	Never	Rarely	Frequently

Views			
Liberal	480	180	120
Conservative	215	50	15
Other	170	45	85

ANSWER:

H_0 : political views and marijuana usage level are independent within the population.

H_a : political views and marijuana usage level are not independent within the population.

H_0 will be rejected if $X^2 \geq X^2_{.01,4} = 13.277$

$\hat{E}_{ij}$	1	2	3
1	496.10	157.72	126.18
2	178.09	56.62	45.29
3	190.81	60.66	48.53

The computed value of the test statistics χ^2 is

χ^2 = 0.52 + 3.15 + 0.30 + 7.65 + 0.77 + 20.26 + 2.27 + 4.04 + 27.41 = 66.37

Since $X^2 = 66.37 \geq 13.277$, the independence hypothesis is rejected in favor of the conclusion that political views and level of marijuana usage are dependent (related).

Section 12.5

35. The __________ is a measure of how strongly related two variables x and y are in a sample.

ANSWER: sample correlation coefficient

36. Given n pairs of observations $(x_1 y_1),(x_2 y_2),.........,(x_n,y_n)$, if large x's are paired with large y's and small x's are paired with small y's, then a __________ relationship between the variables is implied. Similarly, it is natural to speak of x and y having a __________ relationship if large x's are paired with small y's and small x's are paired with large y's.

ANSWER: positive, negative

37. The value of the sample correlation coefficient r is always between __________ and __________.

ANSWER: -1, +1

38. The sample correlation coefficient r equals 1 if and only if all (x_i,y_i) pairs lie on a straight line with __________ slope.

ANSWER: positive

39. The sample correlation coefficient r equals -1 if and only if all (x_i,y_i) pairs lie on a straight line with __________ slope.

ANSWER: negative

40. If the sample correlation coefficient r equals -.80, then the value of the coefficient of determinations is __________.

ANSWER: 64

43. When $H_0 : \rho = 0$ is true, the test statistic $T = R\sqrt{n-2}/\sqrt{1-R^2}$ has a t distribution with __________ degrees of Freedom, where n is the sample size.

ANSWER: n-2

84. A data set consists of 15 pairs of observations $(x_1,y_1),(x_2,y_2),........(x_{15},y_{15})$. If each x_i is replaced by $3x_i$ and if each y_1 is replaced by $4y_i$, then the sample correlation coefficient r

A. increases by 3/15
B. increases by 4/15
C. remains unchanged
D. decreases by 3/15
E. decreases by 4/15

ANSWER: C

85. A data set consists of 20 pairs of observations $(x_1,y_1),(x_2,y_2),.........(x_{20},y_{20})$. If each x_i is replaced by $x_i - 1$ and if each y_i is replaced by $y_i - 2$, then the sample correlation coefficient r

A. decreases by .05
B. decreases by .10
C. increases by.05
D. increases by .10
E. remains unchanged

ANSWER: E

103. The Turbine Oil Oxidation Test (TOST) and the Rotating Bomb Oxidation Test (RBOT) are two different procedures for evaluating the oxidation stability of steam turbine oils. The accompanying observations on x = TOST time (hr) and y = RBOT time (min) for 12 oil specimens have been reported:

TOST	4200	3600	3750	3675	4050	2770
RBOT	370	340	375	310	350	200

TOST	4870	4500	3450	2700	3750	3300
RBOT	400	375	285	225	345	285

a. Calculate and interpret the value of the sample correlation coefficient .
b. How would the value of r be affected if we had let x = RBOT time and y = TOST time?
c. How would the value of r be affected if RBOT time were expressed in hours?
d. Normal probability plots indicate that Both TOST and ROBT time appear to have come from normally distributed populations. Carry out a test of hypotheses to decide whether RBOT time and TOST time are linearly related.

ANSWER:

a. Summary values: $\sum x = 44,615, \sum x^2 = 170,355,425, \sum y = 3,860,$ $\sum y^2 = 1,284,450, \sum xy = 14,755,500, n = 12.$ Using these values we calculate $S_{xx} = 4,480,572.92, S_{yy} = 42,816.67,$ and $S_{xy} = 404,391.67.$ So

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = .9233.$$

b. The value of r does not depend on which of the two variables is labeled as the x variable. Thus, had we let x = RBOT time and y = TOST time, the value of r would have remained the same.

c. The value of r does no depend on the unit of measure for either variable. Thus, had we expressed RBOT time in hours instead of minutes, the value of r would have remained the same.

d. $H_o : \rho_1 = 0$ vs $H_a : \rho \neq 0.$ $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; Reject H_0should be rejected. The model is useful.

105. Hydrogen content is conjectured to be an important factor in porosity of aluminum alloy castings. The accompanying data on x = content and y = gas porosity for one particular measurement technique have been reported:

x	.18	.20	.21	.21	.21	.22	.23
y	.46	.70	.41	.45	.55	.44	.24

x	.23	.24	.24	.25	.28	.30	.37
y	.47	.22	.80	.88	.70	.72	.75

MINITAB gives the following output in response to a CORRELATION command:

Correlation of Hydrogen and Porosity = 0.449

a. Test at level .05 to see whether the population correlation coefficient differs from 0.
b. If a simple linear regression analysis had been carried out, what percentage of observed variation in porosity could be attributed to the model relationship?

ANSWER:

a. $H_o : \rho_1 = 0$ vs $H_a : \rho \neq 0,$ Reject H_o if; Reject at level .05if either

$t \geq t_{.025,12} = 2.179$ or $t \leq -2.179.$ $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(.449)\sqrt{12}}{1-(.449)^2} = 1.74.$ Fail to reject H_o / the data does not suggest that the population correlation coefficient differs from 0.

b. $(.449)^2 = .20$ so 20 percent of the observed variation in gas porosity can be accounted or by variation in hydrogen content.

Section 13.1

19. The regression coefficient β_2 in the multiple regression model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \cdots + \beta_k x^k + \varepsilon$ is interpreted as the expected change in ___________ associated with a 1-unit increase in ___________,while___________ are held fixed.

ANSWER: Y, x_2, $(x_1, x_3, x_4, \ldots \ldots, x_k)$

20. A dichotomous variable, one with just two possible categories, can be incorporated into a regression model via a ___________ or __________ variable x whose possible values 0 and 1 indicate which category is relevant for any particular observations.

ANSWER: dummy, indicator

23. If a data set on at least five predictors is available, regressions involving all possible subsets of the predictors involve at least __________different models

ANSWER: 32

25. If SSE_k is the error sum of squares computed from a model with k predictors and n observations, then the mean squared error for the model is MSE_k = __________/__________.

ANSWER: SSE_k n-k-1

26. When the numbers of predictors is too large to allow for an explicit or implicit examination of all possible subsets, several alternative selection procedures generally will identify good models. The simplest such procedure is the __________, known as BE method.

ANSWER: backward elimination

27. In many multiple regression data sets, the predictors $x_1, x_3, x_4, \ldots \ldots, x_k$ are highly interdependent. When the sample x_i values can be predicted very well from the other predictor values, for at least one predictor, the data is said to exhibit __________.

ANSWER: multicollinearity

28. Which of the following statements are true?

A. One way to study the fit of a model is to superimpose a graph of the best-fit function on the scatter plot of the data.
B. An effective approach to assessment of model adequacy is to compute the fitted or predicted values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$, then plot various functions of these computed quantities, and examine the plots either to confirm our choice of model or for indications that the model is not appropriate.
C. Multiple regression analysis involves building models for relating the dependent variable y to two or more independent variables.
D. All of the above statements are true.
E. None of the above statements are true.

ANSWER: D

32. A multiple regression model has

A. One independent variable.
B. Two dependent variables
C. Two or more dependent variables.
D. Two or more independent variables.
E. One independent variable and one independent variable.

ANSWER: D

33. In multiple regression models, the error term ε is assumed to have:

F. a mean of 1.
G. a standard deviation of 1.
H. a variance of 0.
I. negative values.
J. normal distribution.

ANSWER: E

38. The coefficient of multiple determination R is

A. *SSE*/*SST*
B. *SST*/*SSE*
C. 1-*SSE*/*SST*
D. 1-*SST*/*SSE*
E. (*SSE* + *SST*) / 2

ANSWER: C

52. A first-order no-interaction model has the form $\hat{Y} = 5 + 3x_1 + 2x_2$. As x_1 increases by 1-unit, while holding x_2 fixed, then *y* will be expected to

A. increase by 10
B. increase by 5
C. increase by 3
D. decrease by 3
E. decrease by 6

ANSWER: C

Section 12.1

10. The vertical deviations $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \mathrm{K\ K}, y_n - \hat{y}_n$ from the estimated regression line are referred to as the __________.

ANSWER: residuals

11. When the estimated regression line is obtained via the principle of least squares, the sum of the residuals $y_i - \hat{y}_i$ (i = 1, 3, …….., n) should in theory be __________.

ANSWER: zero

13. In simple linear regression analysis, the __________, denoted by __________, can be interpreted as a measure of how much variability in y left unexplained by the model - that is, how much cannot be attributed to a linear relationship.

ANSWER: error sum of squares, *SSE*

14. In simple linear regression analysis, a quantitative measure of the total amount of variation in observed y values is given by the __________, denoted by __________.

ANSWER: total sum of squares, *SST*

15. If *SSE* = 36 and *SST* = 500, then the proportion of total variation that can be explained by the simple linear regression model is_ _________.

ANSWER: .928

16. In simple linear regression analysis, *SST* is the total sum of squares, *SSE* is the error sum of squares, and *SSR* is the regression sum of squares. The coefficient of determination r^2 is given by $r^2 =$ _____/ *SST* or $r^2 = 1 - ($_____/ *SST*).

ANSWER: *SSR*, *SSE*

21. A 100(1 - α) % confidence interval for the slope β_1 of the true regression line is $\hat{\beta}_1$ ± __________ $\cdot$ $s_{\hat{\beta}_1}$.

ANSWER: $t_{\alpha/2,n-2}$

22. Given that $\hat{\beta}_1 = 1.5, s_{\hat{\beta}_1} = .12$, and n = 15, the 95% confidence interval for the slope β_1 of the true regression line (__________,__________).

ANSWER: -1.7592, -1.2408

28. The null hypothesis $H_0 : \beta_1 = 0$ can be tested against $H_a : \beta_1 \neq 0$ by constructing an ANOVA table, and rejecting H_0 at α level of significance if the test statistic value f ≥ __________, where n is the sample size.

ANSWER: $F_{\alpha,1,n-2}$

52. In simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$, which of the following statements are not required assumptions about the random error term ε ?

A. The expected value of ε is zero.
B. The variance of ε is the same for all values of the independent variable x.
C. The error term is normally distributed.
D. The values of the error term are independent of one another.
E. All of the above are required assumptions about ε .

ANSWER: E

53. A procedure used to estimate the regression parameters β_1 and $\beta_{2,}$ and to find the least squares line which provides the best approximation for the relationship between the explanatory variable x and the response variable Y is known as the

A. least squares method
B. best squares method
C. regression analysis method
D. coefficient of determination method
E. prediction analysis method

ANSWER: A

54. The principle of least squares results in values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the sum of squared deviations between

F. the observed values of the explanatory variable x and the estimated values $\hat{x}$
G. the observed values of the response variable y and the estimated values $\hat{y}$
H. the observed values of the explanatory variable x and the response variable y
I. the observed values of the explanatory variable x and the response values $\hat{y}$
J. the estimated values of the explanatory variable x and the observed values of the response variable y

ANSWER: B

63. If the error sum of squares is 12 and the total sum of squares is 400, then the proportion of observed y variation explained by the simple linear regression model is

A. 0.030
B. 0.173
C. 0.970
D. 0.985
E. None of the above answers are correct.

ANSWER: C

64. Which of the following statements are not correct?

A. The coefficient of determination, denoted by r^2, is interpreted as the proportion of observed y variation that cannot be explained by the simple linear regression model.
B. The higher the value of the coefficient of determination, the more successful is the simple linear regression model in explaining y variation.
C. If the coefficient of determination is small, an analyst will usually want to search for an alternative model (either a nonlinear model or a multiple regression model that involves more than a single independent variable).
D. The coefficient of determination can be calculated as the ratio of the regression sum of squares (*SSR*) to the total sum of squares.
E. All of the above statements are correct.

ANSWER: A

65. The quantity ε in the simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$ is a random variable, assumed to be normally distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. The estimated standard deviation $\hat{\sigma}$ is given by

A. $SSE / (n-2)$
B. $\sqrt{SSE/(n-2)}$
C. $[SSE/(n-2)]^2$

D. $SSE/\sqrt{n-2}$

E. $\sqrt{SSE}/(n-2)$

ANSWER: B

66. In simple linear regression analysis, if the residual sum of squares is zero, then the coefficient of determination r^2 must be

A. -1
B. 0
C. between -1 and zero
D. 1
E. between -1 and 1

ANSWER: D

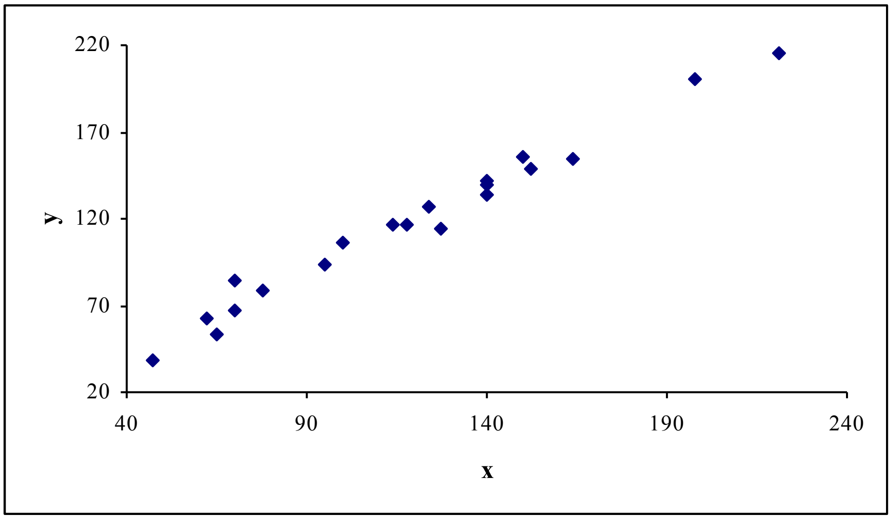
87. The accompanying observations on x = hydrogen concentration (ppm) using a gas chromatography method and y = concentration using a new sensor method were obtained in a recent study

x	47	62	65	70	70	78	95	100	114	118
y	38	62	53	67	84	79	93	106	117	116

x	124	127	140	140	140	150	152	164	198	221
y	127	114	134	139	142	156	149	154	200	215

Construct a scatter plot. Does there appear to be a very strong relationship between the two types of concentration measurements? Do the two methods appear to be measuring roughly the same quantity? Explain your reasoning.

ANSWER:
A scatter plot of the data appears below. The points fall very close to a straight line with an intercept of approximately 0 and a slope of about 1. This suggests that the two methods are producing substantially the same concentration measurements.



90. Suppose that in a certain chemical process the reaction time y (hour) is related to the temperature $(^{\circ}\mathrm{F})$ in the chamber in which the reaction takes place according to the simple linear regression model with equation y = 5.00 - .01x and σ = .075.

a. What is the expected change in reaction time for a 1°F increase in temperature? For a 10°F increase in temperature?

b. What is the expected reaction time when temperature is 200°F? When temperature is 250°F?

c. Suppose five observations are made independently on reaction time, each one for a temperature of 250°F. What is the probability that all five times are between 2.4 and 2.6 hours?

d. What is the probability that two independently observed reaction times for temperatures 1° apart are such that the time at the higher temperature exceeds the time at the lower temperature?

ANSWER:

a. β_1 = expected change for a one degree increase = -.01, and $10\beta_1 = -.1$ is the expected change for a 10 degree increase.

b. $\mu_{Y\cdot 200} = 5.00 - .01(200) = 3$, and $\mu_{Y\cdot 250} = 2.5$.

c. The probability that the first observation is between 2.4 and 2.6 is

$$P(2.4 \le Y \le 2.6) = P\left(\frac{2.4-2.5}{.075} \le Z \le \frac{2.6-2.5}{.075}\right)$$

$= P(-1.33 \le Z \le 1.33) = .8164$. The probability that any particular one of the other four observations is between 2.4 and 2.6 is also .8164, so the probability that all five are between 2.4 and 2.6 is $(.8164)^5 = .3627$.

d. Let Y_1 and Y_2 denote the times at the higher and lower temperatures, respectively. Then $Y_1 - Y_2$ has expected value $5.00 - .01(x+1) - (5.00 - .01x) = -.01$. The standard deviation of $Y_1 - Y_2$ is $\sqrt{(.075)^2 + (.075)^2} = .10607$. Thus

$$P(Y_1 - Y_2 > 0) = P\left(z > \frac{-(-.01)}{.10607}\right) = P(Z > .09) = .4641.$$

91. The accompanying data on x = current density (mA/cm^2) and y = rate of deposition $(\mu\mathrm{m/min})$ appeared in a recent study. Do you agree with the claim by the article's author that "a linear relationship was obtained from the tin-lead rate of deposition as a function of current density"? Explain your reasoning.

x	20	40	60	80
y	.24	1.20	1.71	2.22

ANSWER:

For this data, $n = 4, \sum x_i = 200, \sum y_i = 5.37, \sum x_i^2 = 12.000, \sum y_i^2 = 9.3501,$

$\sum x_i y_i = 333.$ $S_{xx} = 12{,}000 - \frac{(200)^2}{4} = 2000, S_{yy} = 9.3501 - \frac{(5.37)^2}{4} = 2.140875,$

and $S_{xy} = 333 - \frac{(200)(5.37)}{4} = 64.5.$ $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{64.5}{2000} = .03225$ and

$$\hat{\beta}_0 = \frac{5.37}{4} - (.03225)\frac{200}{4} = -.27000.$$

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy} = 2.14085 - (.03225)(64.5) = .060750.$$

$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{.060750}{2.14085} = .972.$ This is a very high value of r^2, which confirms the authors' claim that there is a strong linear relationship between the two variables.

92. A scatter plot, along with the least squares line, of x = rainfall volume (m^3) and y = runoff volume (m^3) for a particular location were given. The accompanying values were read from the plot.

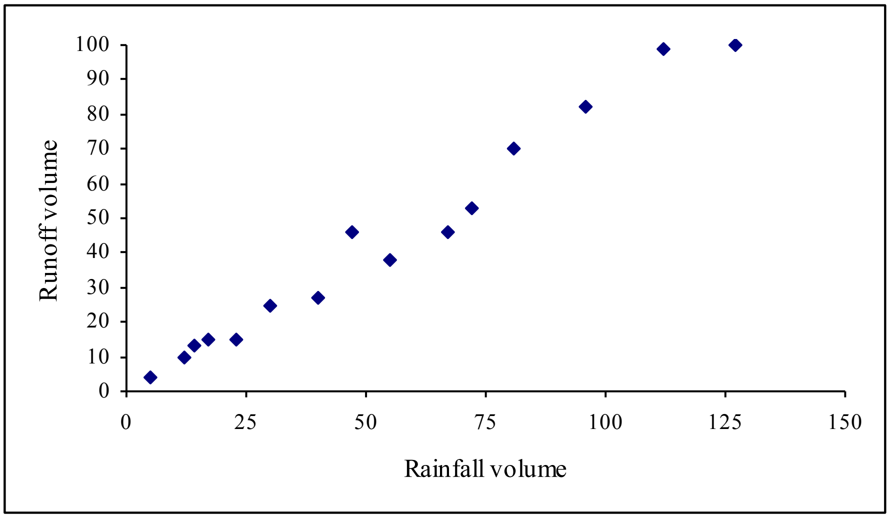
x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46

x	55	67	72	81	96	112	127
y	38	46	53	70	82	99	100

a. Does a scatter plot of the data support the use of the simple linear regression model?
b. Calculate point estimates of the slope and intercept of the population regression line.
c. Calculate a point estimate of the true average runoff volume when rainfall volume is 50.
d. Calculate a point estimate of the standard deviation σ.
e. What proportion of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

ANSWER:

a.



Yes, the scatterplot shows a strong linear relationship between rainfall volume and runoff volume, thus it supports the use of the simple linear regression model.

b. $\bar{x} = 53.200, \bar{y} = 42.867, S_{xx} = 63040 - \frac{(798)^2}{15} = 20,586.4,$

$S_{yy} = 41,999 - \frac{(643)^2}{15} = 14,435.7,$ and $S_{xy} = 51,232 - \frac{(798)(643)}{15} = 17,024.4.$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{17,024.4}{20,586.4} = .82697$ and $\hat{\beta}_0 = 42.867 - (.82697)53.2 = -1.1278.$

c. $\mu_{y \cdot 50} = -1.1278 + .82697(50) = 40.2207.$

d. $SSE = S_{yy} - \hat{\beta}_1 S_{xy} = 14,435.7 - (.82697)(17,324.4) = 357.07.$

$$s = \hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{357.07}{13}} = 5.24.$$

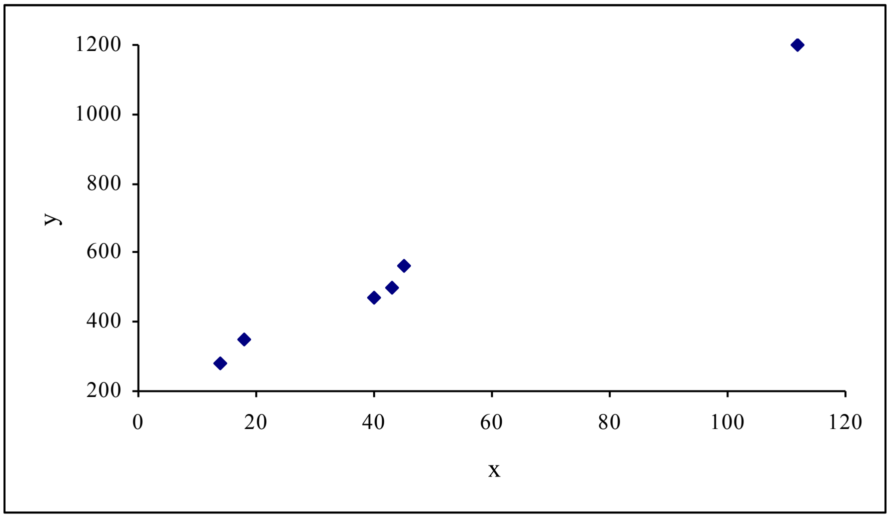
e. $r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{357.07}{14,435.7} = .9753$. So 97.53% of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall.

93. The accompanying data was read from a graph that appeared in a recent study. The independent variable is SO_2 deposition rate $(\text{mg/m}^2/\text{day})$ and the dependent variable is steel weight loss (g/m^2).

x	14	18	40	43	45	112
y	280	350	470	500	560	1200

a. Construct a scatter plot. Does the simple linear regression model appear to be reasonable in this situation?
b. Calculate the equation of the estimated regression line.
c. What percentage of observed variation in steel weight loss can be attributed to the model relationship in combination with variation in deposition rate?
d. Because the largest x value in the sample greatly exceeds the others, this observation may have been very influential in determining the equation of the estimated line. Delete this observation and recalculate the equation. Does the new equation appear to differ substantially from the original one (you might consider predicted values)?

ANSWER:
a.



According to the scatter plot of the data, a simple linear regression model does appear to be plausible.

b. The regression equation is $y = 138 + 9.31x$
c. The desired value is the coefficient of determination, $r^2 = 99.0\%$.
d. The new equation is $y^* = 190 + 7.55x^*$. This new equation appears to differ significantly. If we were to predict a value of y for $x = 50$, the value would be 567.9, where using the original data, the predicted value for $x = 50$ would be 603.5.

98. An investigation of the relationship between traffic flow x (1000's of cars per 24 hours) and lead content y bark on trees near the highway ($\mu g / g$ dry wt) yielded the data in the accompanying table.

x	8.3	8.3	12.1	12.1	17.0	17.0	17.0	24.3	24.3	24.3	33.6
y	227	312	362	521	640	539	728	945	738	759	1263

The summary statistics are:

$n = 11,\ \sum x_i = 198.3,\ \sum y_i = 7034,\ \sum x_i^2 = 4198.03,\ \sum y_i^2 = 5{,}390{,}382,\ \sum x_i y_i = 149{,}354.4$

In addition, the least squares estimates are given by: $\hat{\beta}_0 = -12.84159$, and $\hat{\beta}_1 = 36.18385$

Carry out the model utility test using the ANOVA approach for the traffic flow/lead-content data of Example 12.6. Verify that it gives a result equivalent to that of the *t* test.

ANSWER:

SSE = 5,390,382 – (-12.84159)(7034)-(36.18385)(149,354.4)=76,492.54, and

SST = 5,390,382 - $(7034)^2/11$ = 892,458.73

Source	**df**	**SS**	**MS**	**F**
Regression	1	815,966.19	815,966.19	96.0
Error	9	76,492.54	8499.17	
Total	10	892,458.73		

Since no α is specified, let's use $\alpha = .01$. Then $F_{.01,1,19} = 10.56 < 96.0$, so $H_o : \beta_1 = 0$ is rejected and the model is judged useful.

Now, $s = \sqrt{8499.17} = 92.19,\ \sum (x_i - \bar{x})^2 = 4198.03 - (198.3)^2/11 = 623.2218,$

Then, $t = \dfrac{36.184}{92.19/\sqrt{623.2218}} = 9.80,$ and $t^2 = (9.80)^2 = 96.0 = f.$