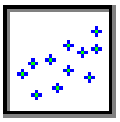
CHAPTER 14



Visualizing Bivariate Data Analysis

CONCEPTS

- Scatter Plot, Box Plot, Column Box Plot, Correlation Coefficient, Regression Line, Coefficient of Determination, t Statistic, F Statistic, Cross-Tabulation, Chi-Square Test for Independence, P-Value

OBJECTIVES

- Become familiar with alternative ways to display bivariate data and assess their strengths and weaknesses
- Understand alternative measures of association in bivariate data and recognize their underlying assumptions
- Be able to interpret bivariate regression statistics and assess their significance
- Know how to use and interpret a chi-square test for independence using cross-tabulated frequencies

Overview of Concepts

Bivariate data analysis refers to any technique that helps us analyze relationships between two variables X and Y with n observed data pairs (X_1,X_2), (X_2,X_2), . . . , (X_n,X_n). Simply displaying the data pairs on a **scatter plot** may reveal much of what we want to know. Descriptive statistics and **box plots** allow us to analyze X and Y separately (to appraise central tendency, dispersion, and skewness). **Column box plots** for subgroups of Y based on values of X may provide further insight. These initial analytical steps require few assumptions.

A further step is to examine the **correlation coefficient**, a statistical measure of association between X and Y. The correlation coefficient can range from –1 (perfect inverse relationship) to +1 (perfect direct relationship). Uncorrelated data will appear as a random collection of points with a correlation coefficient near zero. Correlation analysis does not require us to specify a "dependent" variable and an "independent" variable. Rather, the variables are considered to covary, without indication of causality. For example, students who tend to score well on history exams may also tend to score well on literature exams. However, high history scores do not "cause" high literature scores (both variables may instead reflect the students' general abilities and study habits).

Often (but not always) one variable is regarded as a cause and the other as an effect. For example, we might suppose that quarterly software revenue is affected by advertising expenditures (but not vice versa). From pairs of sample data, we may estimate the coefficients of a simple **regression line** $Y_i = \beta_0 + \beta_1 X_i$. Its slope (change in Y for a unit change in X) and intercept (value of Y when X = 0) can help us answer policy questions (e.g., how much extra revenue is generated by an extra dollar's advertising expenditure?). Although the regression line reveals average change in Y for a unit change in X, it does not prove causation.

Since the fit of a linear model to observed data is usually imperfect, there is a *residual* for each observation (difference between observed Y_i and estimated Y_i). The Ordinary Least Squares (OLS) method chooses the slope and intercept so the fitted regression yields the smallest possible sum of squared residuals. Residuals are the vertical distances from the fitted regression line to each point on the scatter plot. The significance of each estimated coefficient is assessed using its **t statistic** (ratio of the estimated slope or intercept to its standard error) and corresponding **p-value**.

To assess overall fit we examine R^2 (the **coefficient of determination**), the standard error of the regression, and the **F statistic** (and its p-value). R^2 ranges from 0 (poor fit) to 1 (perfect fit). A small standard error signifies a good fit. The larger the F statistic, the more likely it is that the observed association between Y and X is not due to chance.

Regression analysis assumes that X is nonstochastic and that the errors are independent, normally distributed stochastic disturbances with zero mean and constant variance. If we feel that these assumptions are unwarranted, we might merely construct a grid on the scatter plot and count the number of points within each grid cell (**cross-tabulation** of frequencies). This table of frequencies (also called a contingency table) permits a nonparametric **chi-square test for independence** to be used as a measure of association. The chi-square test compares the expected cell frequency in row j and column k with the corresponding observed cell frequency. The expected frequencies are calculated based on the assumption that X and Y are independent of one another. A chi-square statistic near zero indicates that X and Y are independent, while a large chi-square test statistic indicates that X and Y are not independent.

Illustration of Concepts

Renée is a statistics student who works part-time in a supermarket seafood department. As a project in her university statistics class, she decides to record the weight and cost of customers' fish purchases. For a sequential sample (every third customer) she records the type of fish, its weight, and the cost that is printed on the bar-code stick-on label when the fish is weighed. She records 51 purchases and makes a **scatter plot**. The **box plots** for X and Y (Figure 1) reveal that both variables are right-skewed, and a **column box plot** (Figure 2) reveals that higher weight groups have higher median cost and greater interquartile range.

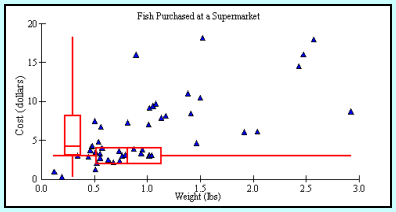


Figure 1: Individual Box Plots

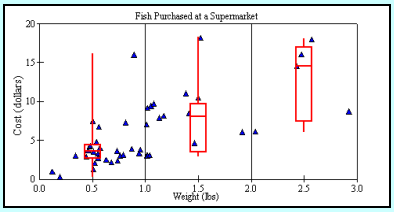


Figure 2: Column Box Plots

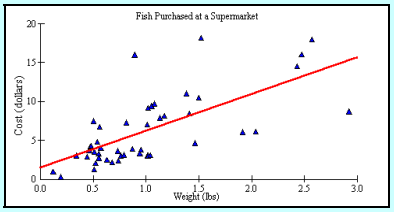


Figure 3: Regression Line

The scatter plot shows a positive association, as expected. Actually, the relationship between weight and cost would be linear (except for measurement error) for a given *type* of fish. But the sample contains 28 different kinds of fish, so the relationship isn't exact. The **correlation coefficient** ($R = 0.679$) indicates a strong association. Renée proposes a regression model: Cost = $\beta_0 + \beta_1$ Weight. Her hypotheses are based on her prior beliefs about the slope:

H_0: $\beta_1 \leq 0$ (cost does not increase as weight of fish increases)
H_1 $\beta_1 > 0$ (cost increases as weight of fish increases)

The estimated **regression line** (Figure 3) is Cost = 1.415 + 4.744 Weight. On average, cost rises $4.74 for each extra pound of weight. For d.f. = 49 and $\alpha = 0.05$, the one-tail critical value of Student's t is 1.677. The **t statistic** for the slope ($t = 6.48$) leads to a strong rejection of the null hypothesis. This decision is supported by its **p-value** ($p < 0.01$), which says that the slope estimate probably is not due to chance. In contrast, the intercept's t statistic ($t = 1.67$) suggests that the intercept does not differ significantly from zero. The **coefficient of determination** ($R^2 = 0.461$) says that fish weight "explains" about 46% of the variation in cost. The **F statistic** ($F = 41.97$) has a p-value below 0.01, so the overall regression is significant.

But is regression appropriate? As an alternative, Renée places a 2×2 grid on her scatter plot (Figure 4) and examines the **cross-tabulation** of frequencies in each grid cell (Figure 5). Expected and observed frequencies differ markedly, and the chi-square statistic (19.062) is significant ($p < 0.001$). But the small expected frequency in one cell ($E_{12} = 1.1$) violates Cochran's Rule (which requires $E_{ij} \geq 5$), so the **chi-square test for independence** is suspect in this case. A different grid might help.

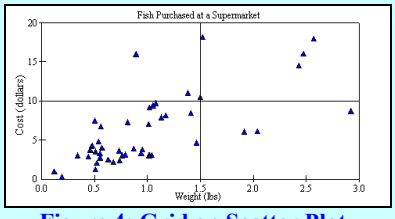


Figure 4: Grid on Scatter Plot

Cost (dollars)	Weight (lbs) Low	High	Total
High	2 / 5.9	5 / 1.1	7
Low	41 / 37.1	3 / 6.9	44
Total	43	8	51

Chi-square test statistic = 19.062 ($p < .001$)

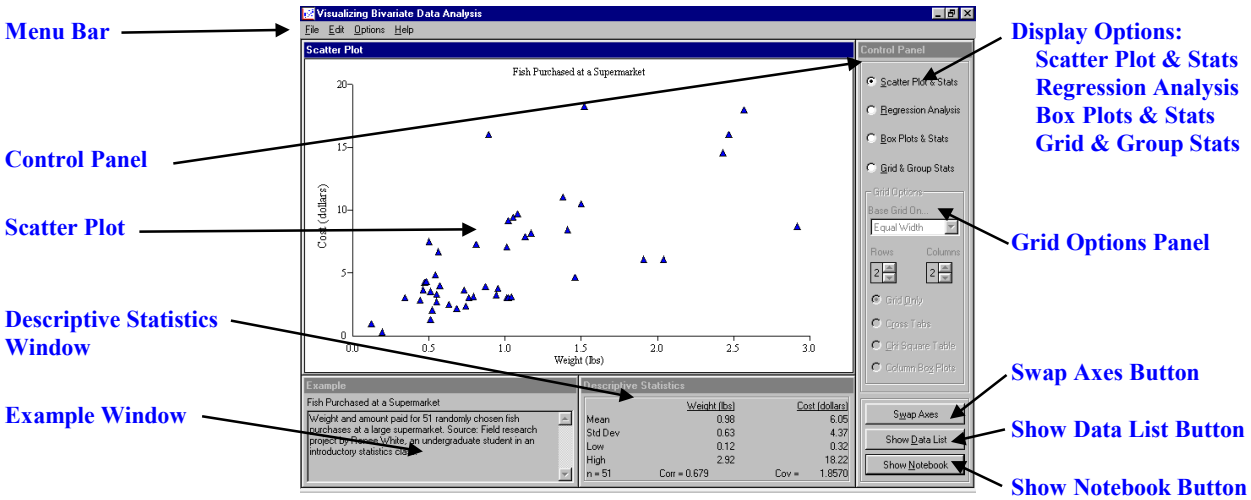
Figure 5: Cross-Tabulation

Orientation to Basic Features

This module helps you learn to recognize and display bivariate data, and to interpret common measures of association between two variables. It permits regression analysis and a chi-square test for independence of two variables. You can enter your own data, choose a scenario, use a Do-It-Yourself simulation control panel, use built-in real databases, or choose a real example.

1. **Select an Example**
   Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions and then click the Concepts tab to see the concepts that you will learn. Click the Examples tab. Click Consumer. Select an example, read it, and press OK. Read the Hint that appears. Press OK.

2. **Scatter Plot Display**
   The initial display contains a Scatter Plot, the Control Panel, an Example window, and a Descriptive Statistics window. The Grid Options panel is initially disabled. Click on any point to reveal its (X,Y) coordinates and its label (if any). Any data point can be dragged to a new location (this is not recommended at this time) to see what effect it has on the analysis.

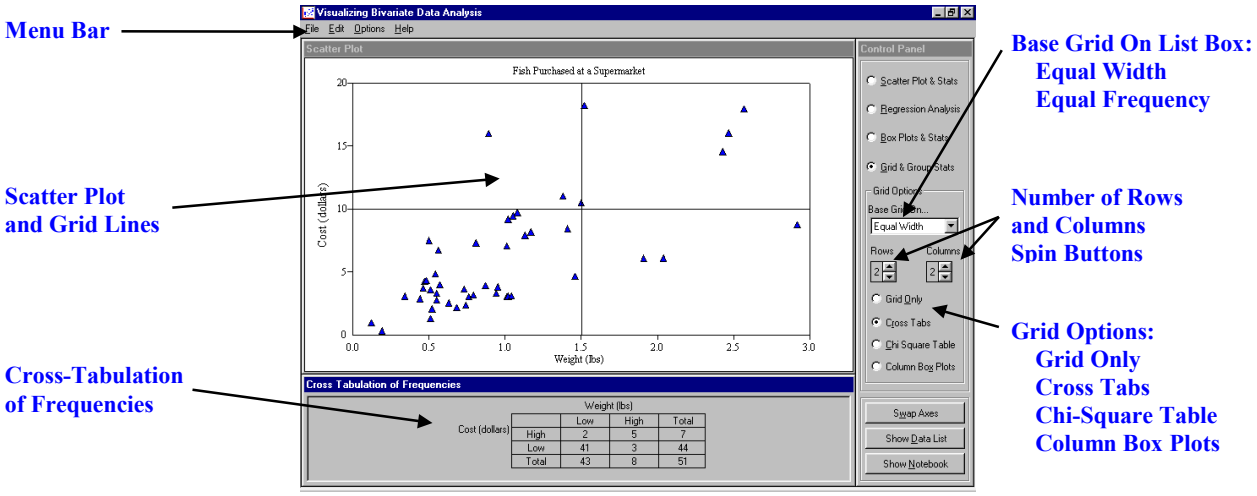


3. **Control Panel: Display Options**
   Click the Swap Axes button to interchange the X and Y variables. Both the data and the axis labels are switched. Click Swap Axes again to restore the initial display. Click the Show Data List button to see a list of the data pairs and their labels. At the top of the Control Panel, select Box Plots & Stats to display box plots (in red) for both variables along the axes of the scatter plot. Their descriptive statistics appear in the Descriptive Statistics window.

4. **Regression**
   Click Regression Analysis to display a fitted regression line (in red) on the scatter plot and to show the Regression Statistics window. Click Options on the menu bar and choose Regression. The Always Show Y Intercept option changes the scale of the scatter plot so that X = 0 is displayed. Otherwise, the scale is chosen to display the maximum graph detail. It has no effect if the scale already includes X = 0. If you choose Suppress Intercept, the regression is forced through the origin (a major change in model specification).

5. **Control Panel: Grid Options**

   Select Grid & Group Stats to superimpose a grid on the scatter plot. Use the Rows and Columns spin buttons to change the number of rows or columns. Experiment with the Base Grid On list box (Equal Width, Equal Frequency). Click Grid Only to display a simple grid. Click Cross Tabs to replace the two lower display windows with a single window showing a tabulation of frequencies. Click Chi-Square Table to see expected and actual frequencies along with a chi-square test statistic and its p-value. Click Column Box Plots to show column box plots.



6. **Scenarios**

   Press the Show Notebook button, choose the Scenarios tab, pick a category and click OK. The scenarios are hypothetical, but this type of data would be likely to exhibit the proposed correlation and population shape. Variable descriptions appear in the Data Sets window. Press Take Samples to see a sample. You can't change the correlation or population type, but you can modify the sample size (though this may distort the scenario's logic).

7. **Databases**

   Click Show Notebook and choose the Databases tab. Click on U.S. States or World Nations to see a list of variables. Categories labeled + can be clicked to display an expanded list of variables within the category (or click – to collapse the category). You must choose two variables. Click OK. Descriptions of each variable appear in the Data Sets window.

8. **Copying a Display**

   Click the display you wish to copy. Its window title will be highlighted. Select Copy from the Edit menu (on the menu bar at the top of the screen) or Ctrl-C to copy the display.

9. **Help**

   Click Help on the menu bar at the top of the screen. Search for Help lets you search a topic index, Contents shows a table of contents, Using Help gives instructions on how to use Help, and About gives licensing and copyright information about *Visual Statistics*.

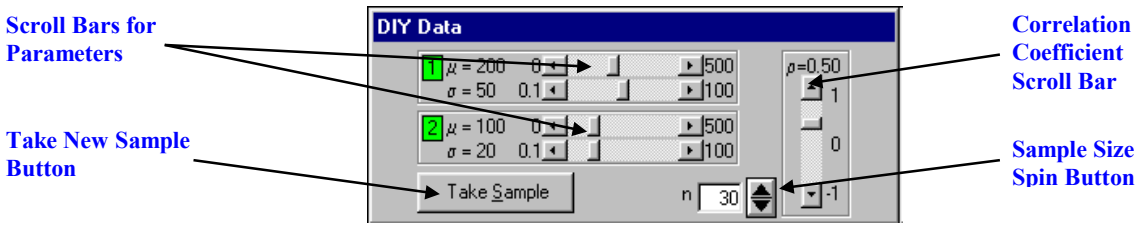
10. **Exit**

    Close the module by selecting Exit in the File menu (or click ☒ in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

1. **Do-It-Yourself Data**

   To create a simulated data set of your own, click the Scatter Plot & Stats option. Press the Show Notebook button. Choose the Do-It-Yourself tab, and click OK. On the DIY Data control panel, use the horizontal scroll bars to change μ and σ and the vertical scroll bar to set ρ (the desired correlation coefficient, where $-1 \le \rho \le +1$). Use n to set the sample size (up to 150). Click Take Sample to display a new bivariate sample from the populations you have specified.

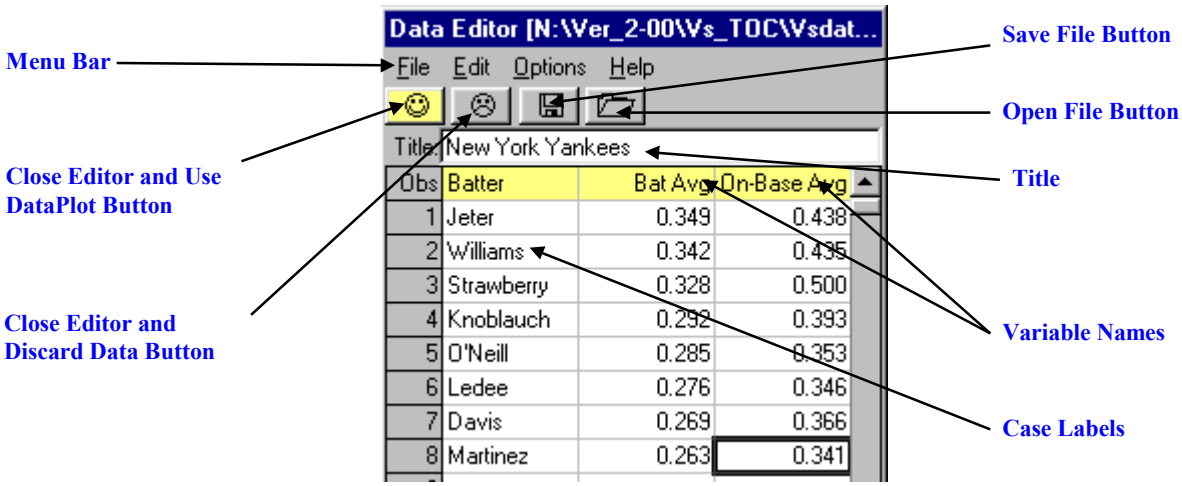


2. **Options: Population Being Sampled**

   Select Options on the menu bar and choose Distribution of Variables. You may specify the distribution of the populations to be sampled as Normal, Uniform, Skewed, or Very Skewed. The default is Normal. Both population distributions will be the same. Click Take Sample to see the effects of varying the distribution's shape.

3. **Data Editor**

   To enter your own data, click Show Notebook, choose the Data Editor tab, and click OK. Each row is an observation (a person, a team, etc.). The first column is an optional label for the observation (a person's name, a team's name, etc.). Use the menu bar File commands (New, Open, Save, Save As) to read existing files or save files in *Visual Statistics* format. Use the menu bar Edit commands (Cut, Copy, Paste, Insert Row, Delete Row) to edit your data or copy/paste data from another spreadsheet using the clipboard. Use the menu bar Options commands (Format, Sort, Typing Replaces Cell) to adjust the decimal places displayed or sort the data. When you are finished editing your data, click Close Editor and Use Data (or click the "smiling face" icon). If you don't want to use the data , click Close Editor and Discard Data (or click the "sad face" icon).



Basic Learning Exercises Name _____

Press the Show Notebook button and select the Databases tab. Choose U.S. States. For your first variable (upper window) select Demographics (click + to expand its categories) and highlight Pop 65 and Up (%). For your second variable (lower window) select Health (click + to expand its categories) and highlight Heart Deaths per 100,000. Then press OK.

Scatter Plots and Correlation

1. Define each variable. Are the variables corrected for the effects of population size? Why would this correction be important (i.e., why not use totals)?

2. Describe the general appearance of the scatter plot. If you see any unusual data points, identify them by clicking on them (but be careful *not* to drag them).

3. Click Show Data List. Scroll down the list. Identify the state with the highest and lowest values for each variable. Can you suggest reasons? **Hint:** You can also find out by clicking on individual data points.

4. What is the value of the correlation coefficient? What is its p-value? What do these statistics tell you about the association between X and Y?

5. Click a point in the middle of the scatter plot and drag and drop it near the extreme lower right corner. What happens to the correlation coefficient? The p-value? How influential was this single data point?

Box Plots

6. The data have been altered by dragging a data point. Restore the correct data by clicking Show Notebook and clicking OK (to select the same two database files). Make sure the scatter plot has its original appearance. Record the means for both variables, then click Box Plots & Stats. Record the medians. Explain what the means, box plots, and quartile statistics tell you about each variable.

   X = Percent of population age 65 and over
   Mean _____ Median _____

   Y = Death rate from heart disease
   Mean _____ Median _____

7. Click Grid & Group Stats to display a 2 x 2 grid. Click Column Box Plots. Note the sample sizes (column statistics). In the Base Grid On list box, choose Equal Frequency instead of Equal Width and again note the sample sizes. Why is there a difference? Under what circumstances would Equal Width and Equal Frequency yield the same statistics?

8. Using Equal Width, increase the number of columns to 3, and then to 4. What happens to the column box plots, and why?

9. Repeat exercise 8 using the Equal Frequencies option. Discuss the advantages and disadvantages of these two options.

10. Using Equal Frequencies and 2 columns, compare the column box plots and their medians. What does this comparison suggest about the relationship between X and Y? Increase the number of columns to 3, and then to 4. Does the same conclusion hold?

Intermediate Learning Exercises Name _____

Press the Show Notebook button, select the Databases tab, and choose U.S. States. For your first variable select Demographics and highlight Pop 65 and Up (%). For your second variable select Health and highlight Heart Deaths per 100,000. Then press OK.

Regression Analysis

11. Select Regression Analysis. Examine the regression line. Does it represent the data well?

12. Examine the regression statistics. a) Write the equation for the estimated regression line and interpret it. b) Is the sign of the slope believable? c) In this particular example, is the intercept meaningful? d) Would you have any reservations if a state were to use this estimated regression equation to make a health care policy decision? Explain.

13. What does the R^2 tell you? How much variation in the dependent variable is unexplained?

14. Click any point in the middle of the scatter plot and drag and drop it near the extreme lower right corner. a) What happens to the R^2? b) What happens to the estimated slope? c) How great was the effect of this single data point on the fitted regression line?

Cross Tabulations

15. The data have been altered by dragging a data point. Restore the correct data by clicking Show Notebook and clicking OK (to select the same two database files). Click Grid & Group Stats. Verify that the Equal Frequency option is selected. Choose 2 rows and 2 columns. Choose the Cross Tabs option. Examine the cross-tabulation and compare its frequencies to the grid (i.e., pick any grid cell and try to count its data points to see if your count agrees with the cross-tabulation). If you encounter any difficulties checking the frequencies, explain why.

16. a) Are the row totals equal? Are the column totals equal? b) Would this be the case in any 2×2 table? Explain. c) How do you imagine the categories (Low, High) were formed? d) Are the cell frequencies similar in all four cells of the contingency table?

Chi-Square Table

17. Click Chi-Square Table. a) How different are the observed and expected cell frequencies? b) Does the chi-square test indicate that the two variables are independent? c) Is Cochran's Rule violated? **Hint:** If you're unfamiliar with chi-square tests, click Help or consult the Glossary.

18. In this particular bivariate data analysis, why might the chi-square test offer advantages (compared with regression) in assessing the possibility of association between X and Y?

19. Use the Rows and Columns spin buttons to increase the table size to 3 x 3. a) What happens to the chi-square test statistic? b) Does the test indicate independence? c) Is Cochran's Rule violated? **Hint:** You may have to scroll down the chi-square window to see everything.

20. Should the size of the table be increased further (say to 4 x 4)? Explain.

Advanced Learning Exercises Name _____

Press the Show Notebook button, select the Do-It-Yourself tab, and click OK. Use the default means and standard deviations ($\mu_1 = 200$, $\sigma_1 = 50$, $\mu_2 = 100$, $\sigma_2 = 20$) and keep the default sample size ($n = 30$). Set the desired correlation at $\rho = 0.25$.

Correlation and Regression: Effects of Sample Size

21. Click Scatter Plot & Stats. Press Take Sample 10 times and record each estimated correlation coefficient and its p-value. Find the average correlation coefficient and average p-value. How close did the average correlation coefficient come to the desired correlation? How many of the estimated correlations were negative? How much variation was there? How many p-values were below 0.10? How much *visual* evidence of a positive correlation did you see in a typical scatter plot (take a few more samples if you wish).

Sample	1	2	3	4	5	6	7	8	9	10	Average
r	___	___	___	___	___	___	___	___	___	___	0
p-value	___	___	___	___	___	___	___	___	___	___	0

22. Increase the sample size to $n = 100$ and repeat exercise 20. What difference does sample size have on the estimates and their p-values?

Sample	1	2	3	4	5	6	7	8	9	10	Average
r	___	___	___	___	___	___	___	___	___	___	0
p-value	___	___	___	___	___	___	___	___	___	___	0

23. Click Regression Analysis. Press Take Sample 10 times. How often was the estimated slope positive (just by looking at the red regression line on the scatter plot)? Does having the regression line help you see evidence of a positive correlation?

Correlation and Regression: Varying the Correlation

24. Choose n = 30 and set the desired correlation to ρ = 0.50. a) Click Take Sample until you obtain an estimated correlation coefficient within the range 0.48 to 0.52. How hard was it? b) Now click the point farthest from the regression line and drag and drop it on the line. What happened to the correlation coefficient? c) Continue this exercise, watching the estimated correlation coefficient, until you have increased the estimated correlation to 0.99. How nearly linear is the data when you are finished? d) What is the t statistic for the slope? The F statistic? e) Do you expect such samples to occur very often?

Correlation and Regression: Intercept Options

25. Click Take Sample a few times and watch the left end of the X-axis scale. Note that the Y-intercept is not revealed in most of these samples, because X = 0 is not visible. Choose Options and then Regression and select Always Show Y Intercept. Click Take Sample a few times. Discuss advantages and disadvantages of showing the intercept.

26. Choose Options and then Regression and select Suppress Intercept. Examine the scatter plot and regression statistics as you press Take Sample a few times. How is suppressing the intercept different from merely turning off the Always Show Y Intercept option?

27. Choose Options and Regression. Deselect both Suppress Intercept and Always Show Y Intercept (so that neither option is checked). Click Scatter Plot & Stats to eliminate the regression line. Click Take Sample. Click Options again and choose Distribution of Variables. Instead of Normal select Uniform. Press Take Sample a few times. Repeat for Right Skewed and Left-Skewed. Discuss the appearance of the scatter plots. What effects would you expect the non-normal distributions to have on a fitted regression line?

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs in your report.

1. Use the Data Editor to create a bivariate data set of your own. Choose X, Y data from your own experience (e.g., phone call cost versus length of call from your phone bill) or from a secondary source (e.g., an almanac or the *Statistical Abstract of the United States*) or, if nothing else is available, from a textbook. Explore the data set thoroughly using the tools of bivariate data analysis. Which tools are most helpful, and why? Include in your report a copy of graphs and/or tables you feel are relevant for each different experimental setup.
2. Use column box plots and chi-square analysis to analyze a pair of variables from the U.S. States database, or from the World Nations database that you believe *a priori* will be related. Vary the number of columns in the box plots, and vary the number of rows and columns in the chi-square analysis, to see whether a consistent impression is given of the possible association between the two variables (i.e., does varying the size of the contingency table affect the outcome of the significance test). Tell whether a regression or correlation analysis would be appropriate, and why (or why not). Repeat using a pair of variables from the World Nations database. Include in your report a copy of graphs and/or tables you feel are relevant for each different experimental setup.
3. Select the Scenarios tab in the Notebook. Choose a scenario representing each degree of correlation (ranging from Large Positive to Large Negative) and examine a scenario from each. Set each sample size to 30 so the scatter plots will be comparable. Compare the appearance of the scatter plots and comment on the differences that are due to population shape and note the effects of the data's characteristics on the appropriate methods of analysis. What conclusions can you draw from this experiment? Include in your report a copy of graphs and/or tables you feel are relevant for each different experimental setup.

Team Learning Projects

Select one of the three projects listed below. In each case, produce a team project that is suitable for an oral presentation. Use presentation software or large poster boards to display your results. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Ask your instructor if a written report is also expected.

1. This is a project for a team of two. One team member should investigate the U.S. States database and the other should investigate the World Nations database. Using logic and trial-and-error, try to find examples of variables that exhibit strong negative correlation, near-zero correlation, and positive correlation. For each example that you finally decide to present, discuss the characteristics of each variable (central tendency, dispersion, skewness) as revealed by box plots, and comment on the possible policy implications and degree of believability of a regression fitted to the data. Discuss which tools are best for each data set, and explain why. Include in your report a copy of all graphs and statistics that you evaluated.
2. This is a project for a team of two. Investigate the effects of population non-normality and sample size on the accuracy of estimates for the regression slope. Use the Do-It-Yourself controls with a true correlation of 0.60. Use the Options to select each of the four population types. Using five samples, find the average and the range of estimates for the slope. One team member should use a sample of size 10 and the other should use a sample of size 50. Explain your conclusions clearly. Include in your report a copy of all graphs and statistics that you evaluated.
3. This is a project for a team of three or more. Investigate the effects of sample size on accuracy of estimation of the correlation coefficient. Use the Do-It-Yourself controls, and set the true correlation to $\rho = 0.00$. Each team member should select a different sample size between 10 and 120 so that the range is covered. Take several samples and average the estimated correlation coefficients. How often would you mistakenly conclude there is a non-zero correlation? Repeat for a true correlation of $\rho = 0.80$. Repeat the experiment using a skewed population and then a uniform population. Display the team's findings in a simple visual summary (table or chart). Include in your report a copy of all graphs and statistics that you evaluated.

Self-Evaluation Quiz

1. A scatter plot of X and Y
   a. gives little information about the actual correlation.
   b. requires that a linear regression be calculated and displayed.
   c. indicates causal direction since X is the independent variable.
   d. makes too many restrictive assumptions to be used in Exploratory Data Analysis.
   e. has none of the above characteristics.

2. Which is *not* a common characteristic of real bivariate data sets?
   a. The populations are not normally distributed.
   b. The data are skewed to the right.
   c. The data may contain extreme values.
   d. The data are inversely related.
   e. The data are not a random sample.

3. The p-value for the correlation coefficient shows
   a. the probability that the true correlation is non-zero.
   b. the probability of obtaining the sample correlation if the true correlation is zero.
   c. the probability of Type I error if the hypothesis of zero correlation is rejected.
   d. more than one of the above.
   e. none of the above.

4. Which is indicative of an *inverse* relationship between X and Y?
   a. A scatter plot whose points are shaped like a circle.
   b. A scatter plot with points mostly in the lower left and upper right quadrants.
   c. A negative correlation coefficient.
   d. A negative p-value for the correlation coefficient.
   e. None of the above.

5. Column box plots on a scatter plot *cannot* reveal
   a. significance of the correlation coefficient.
   b. skewness of the data within each column group.
   c. dispersion of the data within each column group.
   d. central tendency of the data within each column group.
   e. whether the relationship between X and Y is direct or inverse.

6. A box plot on the X-axis of a scatter plot is *least* likely to reveal which of these?
   a. Number of modes in X.
   b. Skewness in X.
   c. Non-normality in X.
   d. The range of X.
   e. Central tendency of X.

7. The regression line that is superimposed on the scatter plot
   a. is computed using the Ordinary Least Squares method.
   b. guarantees the largest possible sample variance.
   c. guarantees that the slope and intercept are minimized.
   d. guarantees all of the above.
   e. guarantees none of the above.

8. Which is *not* correct regarding the estimated slope of the regression line?
   a. It is divided by its standard error to obtain its t statistic.
   b. It shows the change in Y for a unit change in X.
   c. It is chosen so as to minimize the sum of squared errors.
   d. It may effectively be regarded as zero if its p-value is below 0.01.
   e. Its magnitude is an unreliable indicator of significance.

9. An important attractive feature of the chi-square test for independence is
   a. its one-to-one correspondence to the magnitude of the slope.
   b. its lack of reliance on sophisticated assumptions.
   c. its insensitivity to small expected frequencies.
   d. its ability to reveal directionality of the relationship (direct or inverse).
   e. its high power relative to regression or correlation analysis.

10. Which is *least* likely to be adversely affected by outliers?
    a. the chi-square test statistic.
    b. the regression slope.
    c. the sample correlation coefficient.
    d. the R^2 statistic.
    e. the p-value for the slope.

11. Which is a characteristic of the chi-square test for independence?
    a. It requires a large observed frequency in each cell.
    b. It is a parametric test.
    c. It requires at least 30 observations for a 4×4 table.
    d. Its results do not depend on the X–Y grid boundaries.
    e. None of the above is a correct characteristic.

12. In a simple regression, which would suggest a relationship between X and Y?
    a. Large p-value for the estimated slope.
    b. Large t statistic for the slope.
    c. Large p-value for the F statistic.
    d. All of the above suggest a relationship.
    e. None of the above suggests a relationship.

Glossary of Terms

ANOVA table Summary of decomposition of variance. For a bivariate regression, the ANOVA table is shown below. See also **Sums of squares**, **Mean square**, and **Degrees of freedom**.

Source	Sum of Squares	d.f.	Mean Square
Regression	SSR	1	MSR = SSR/1
Error (residual)	SSE	n – 2	MSE = SSE/(n – 2)
Total	SST	n – 1	

Bivariate data Sample of n observations on two random variables X and Y. Each data pair is denoted (X_i,Y_i) where i = 1, 2, . . . , n.

Box plot Graphic representation of X_{Min}, Q_1, Q_2, Q_3, and X_{Max} . A simple box plot is sometimes called a "five-number summary" or a "box and whisker plot". The box encloses the quartiles and the span of the "whiskers" indicates the range.

Chi-square test for independence Test for association between two cross-tabulated variables X and Y in a contingency table with R rows and C columns. If X and Y are truly independent, the chi-square test statistic should be near zero. The test statistic is

$$\sum_{j=1}^{R}\sum_{k=1}^{C}\frac{\left(O_{jk} - E_{jk}\right)^2}{E_{jk}}$$

where O_{jk} and E_{jk} are the respective observed and expected cell frequencies in row j and column k. For a valid test, E_{jk} should not be too small. See **Cochran's rule** and **Degrees of freedom**.

Cochran's rule A rule of thumb that suggests that the minimum expected frequency (E_{jk}) in each cell of a contingency table should be at least 5 for a valid chi-square test.

Coefficient of determination In a bivariate regression, the coefficient of determination is a measure of overall fit. R^2 near 0 signifies a lack of fit while R^2 near 1 signifies a near-linear fit. It is calculated from sums of squares using the equation R^2 = SSR / SST. See **ANOVA table**.

Column box plot For a bivariate data set, a box plot for Y values corresponding to X values within a certain range.

Conditional mean In a regression, the expected value of the dependent variable *given* the observed value(s) of the dependent variable(s). In contrast, the unconditional mean is just the expected value of the dependent variable (i.e., the mean).

Contingency table A cross-tabulation of sample observations (X_i, Y_i) on two random variables into categories (the number of categories need not be the same for X and Y). A contingency table with R rows and C columns is called an R × C contingency table. See **Bivariate data**, **Observed frequency**, **Expected frequency**, and **Chi-square test for independence**.

Correlation coefficient Measure of fit in a bivariate regression, equal to the sample covariance divided by the product of the sample standard deviations of X and Y. A correlation of –1 indicates a perfect inverse relationship, 0 indicates no relationship, and +1 indicates a perfect direct relationship. The formula for the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \qquad \text{or} \qquad r = \frac{\text{Cov}(X,Y)}{S_X S_Y}$$

Cross-tabulation See **Contingency table**.

Degrees of freedom Number of independent pieces of information in a sample. In a chi-square test, degrees of freedom will be (R – 1)(C – 1), where R is the number of rows and C is the number of columns in the contingency table. In a regression ANOVA table, total degrees of freedom will be n – 1, error degrees of freedom will be n – k – 1, and the regression degrees of freedom will be k, where n is the sample size and k is the number of independent predictors in the model (k = 1 for a bivariate model).

Dependent variable In a regression, the variable (denoted Y) that is placed on the left-hand side of the equation and may be assumed to be affected by the independent variable (denoted X). On the scatter plot, the dependent variable is customarily shown on the vertical axis.

Expected frequency In a contingency table, the number of observations in a cell that would be expected if X and Y were independent. For row j and column k, the observed frequency may be denoted E_{jk} and is $n_j n_k/n$, where n_j and n_k are the respective row and column totals for the contingency table. See **Observed frequency** and **Chi-square test for independence**.

F statistic In a regression ANOVA table, the ratio of the regression mean square to the error mean square (F = MSR / MSE). The larger the F statistic, the less likely it is that the association between Y and X is due to chance. An F statistic close to zero would suggest the regression line does not give a good fit to the sample data.

Frequency See **Expected frequency** and **Observed frequency**.

Grid Grouping of X and Y into categories on a scatter plot.

Independence See **Chi-square test for independence**.

Independent variable In a regression, the variable (denoted X) that appears on the right-hand side of the equation and is thought to cause variation in the dependent variable (denoted Y). On a scatter plot, the independent variable is usually shown on the horizontal axis.

Intercept Value of the dependent variable when X = 0 in the regression model $Y = \beta_0 + \beta_1 X$. On a graph, the intercept β_0 is the point where the regression line intersects the Y-axis. See **Ordinary Least Squares**.

Mean square In a regression ANOVA table, the sums of squares due to regression or error, divided by their respective degrees of freedom. In a bivariate regression the regression mean square is MSR = SSR/1 and the error mean square is MSE = SSE/(n – 2). See **ANOVA table**.

Model Proposed regression equation whose coefficients (i.e., slope and intercept) are to be estimated from sample data. The model is generally based on a theory.

Observed frequency In a contingency table, the number of observations in each cell. For row i and column j, the observed frequency is often denoted O_{ij}. See **Expected frequency**.

Ordinary Least Squares Abbreviated OLS, this is a calculus-based method of choosing the regression coefficients so that the fitted regression model yields the smallest possible sum of squared errors (SSE). OLS estimates are widely used in spreadsheets and statistical packages.

Sum of Squared Errors:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Estimated Slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

Estimated Intercept:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

P-value In a regression, the probability of Type I error if we reject a particular hypothesis about a parameter (such as an estimated slope). For example, if we hypothesize that the true slope of a regression line is zero, a small p-value (say, p = 0.01) would suggest rejection of the hypothesis because in so doing we are unlikely to commit Type I error.

Regression line Numerical estimates of the slope and intercept of a bivariate regression model.

Residual Difference between the actual and estimated value of the dependent variable.

R-squared See **Coefficient of determination**.

Scatter plot Visual display in which each of n observed (X_i, Y_i) pairs is plotted as a symbol (usually a dot) at the correct coordinate on the graph.

Slope The change in Y for a unit change in X in the bivariate model $Y = \beta_0 + \beta_1 X$. On a graph, the slope β_1 is the rise divided by the run. See **Ordinary Least Squares**.

Standard error Estimated standard deviation of an unknown parameter. Formulas for some of the standard errors used in this chapter are:

Regression:

$$S_{Y|X} = \sqrt{\frac{SSE}{n - 2}}$$

Intercept:

$$S_{\hat{\beta}_0} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - X\right)^2}}$$

Slope:

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{\sqrt{\sum_{i=1}^{n} \left(X_i - X\right)^2}}$$

Sum of squares In regression, total variation in the dependent variable around its mean (SST) is partitioned into variation explained by the regression (SSR) and variation that is unexplained by the regression (SSE). They are used in ANOVA to find the mean squares and F statistic. Their formulas are shown below.

Total Variation:

$$SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

Explained (Regression):

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

Unexplained (Error):

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

t statistic Generally, the ratio of an estimated coefficient in a regression model to its standard error, which can be used to test whether the estimated coefficient is equal to zero. If the errors in a regression are normal, this ratio is distributed as Student's t and its magnitude may be used to judge the null hypothesis. For example, a large t statistic for the slope would suggest that the true slope is not zero.

Test for Independence See **Chi-square test for independence**.

Y-intercept The value of Y when X is zero. See **Intercept**.

Solutions to Self-Evaluation Quiz

1. e Do Exercises 2–5. Read the Overview of Concepts. Consult the Glossary.
2. d Do Exercises 1–3. Read the Overview of Concepts. Consult the Glossary.
3. d Do Exercises 4–5. Read the Overview of Concepts.
4. c Do Exercises 1–5. Do Individual Learning Project 3.
5. a Do Exercises 6–10. Read the Illustration of Concepts. Consult the Glossary.
6. a Do Exercise 6. Consult the Glossary.
7. a Read the Overview of Concepts. Consult the Glossary.
8. d Do Exercises 11–14. Read the Illustration of Concepts. Consult the Glossary.
9. b Read the Overview of Concepts. Consult the Glossary.
10. a Do Exercises 5, 14, 18. Read the Overview of Concepts. Consult the Glossary.
11. e Do Exercises 15–20. Consult the Glossary. Read Overview of Concepts.
12. b Do Exercises 21–24. Read the Overview of Concepts and Illustration of Concepts.