**CLASSROOM EXERCISES**

***I. Data Structures and Histograms***

1. Consider the data set below, which consists of observations on five production facilities (identified by their group ID).
   1. What is an elementary unit for this data set?
   2. What kind of data set is this, univariate, bivariate, or multivariate?
   3. Identify the qualitative variables, if any.
   4. Is there an ordinal variable here? If so, please identify it.
   5. Is this a time-series, or are these cross-sectional data?

|  |  |  |  |
| --- | --- | --- | --- |
| **Group Id** | **Part** | **Quality** | **Employ** |
| A-235-86 | Brakes | good | 53 |
| W-186-74 | Fuel line | better | 37 |
| X-937-85 | Radio | fair | 26 |
| C-447-91 | Chassis | excellent | 85 |
| F-258-89 | Wire | good | 16 |

1. The Dow Jones company calculates a number of stock market index numbers that are used as indicators of the performance of the New York Stock Exchange. The best known of these is the Dow Jones Industrial Average (DJIA), which is calculated based on the performance of 30 stocks from companies categorized as general industry. Another index is the Dow Jones Transportation Average (DJTA), which is based on the perofrmance of 20 stocks from companies that specialize in transportation services (railroads, airlines, truck lines, shipping lines, and so on). Observations for each of the 20 transportation companies in the DJTA are shown in Table below.
   1. What is the elementary unit for this data set?
   2. What kind of data set is this, univariate, bivariate, or multivariate?
   3. Which of these variables are quantitative? Which are qualitative?
   4. If there are any qualitative variables in this data set, are they nominal or ordinal?
   5. Is this a cross-sectional or time-series data?

|  |  |  |
| --- | --- | --- |
| **Company Name** | **Closing Price on 02/01/12** | **Percent Change from 01/02/12** |
| **Close** | **Change** |
| 3M | 113.27 | -3.30% |
| Alcoa | 35.15 | -1.49% |
| American Express | 35.10 | -2.17% |
| AT&T | 17.33 | -7.33% |
| Boeing | 41.46 | 8.82% |
| Caterpillar | 50.52 | -2.19% |
| Citigroup | 46.49 | -8.86% |
| Coca-Cola | 44.68 | -5.88% |
| DuPont | 43.58 | 1.54% |
| Eastman Kodak | 28.25 | -3.62% |
| Exxon Mobil | 39.00 | -1.52% |
| General Electric | 36.85 | -10.01% |
| General Motors | 51.11 | 5.08% |
| Hewlett-Packard | 22.00 | 1.62% |
| Home Depot | 49.40 | -1.96% |
| Honeywell | 33.68 | 0.84% |
| IBM | 108.00 | -11.11% |
| Intel Corp. | 34.67 | 5.06% |
| International Paper | 41.77 | 4.03% |
| J.P. Morgan Chase & Co. | 32.16 | -11.53% |
| Johnson & Johnson | 57.60 | -1.87% |
| McDonald's | 26.63 | 0.53% |
| Merck | 59.43 | -0.55% |
| Microsoft Corp. | 62.66 | -6.53% |
| Philip Morris | 49.73 | 6.63% |
| Procter & Gamble | 82.57 | 3.21% |
| SBC Communications | 39.96 | 0.15% |
| United Technologies | 70.05 | 8.87% |
| Wal-Mart Stores | 59.26 | 2.08% |
| Walt Disney | 22.45 | 4.66% |

1. Consider CREF, the College Retirement Equities Fund, which manages retirement accounts for employees of non-profit educational and research organizations. CREF manages a large and diversified portfolio in its stock account, somewhere around $67 billion. Investment in furniture and home-furnishing stores represents 0.18% of this portfolio. Data on the market value of these investments is shown in the Table below:
2. Construct a histogram of this data set
3. Based on this histogram, describe the distribution of CREF’s investment in furniture and home-furnishing stores.
4. Describe the shape of the distribution. In particular, is it skewed or symmetric?
5. Find the logarithm of each data value
6. Construct a histogram of these logarithms.
7. Describe the distribution shape of the logarithms. In particular, is it skewed or symmetric?

|  |  |
| --- | --- |
| **Company** | **Market Value (Thousands)** |
| AAR Corp. | 2,035 |
| Alliant Techsystems, Inc. | 5,133 |
| Armor Holdings, Inc. | 1,758 |
| BAE Systems PLC | 31,984 |
| Boeing Co. | 3,64,299 |
| Echostar Communications Corp. | 14,464 |
| Empresa Brasileira de Aeronautica S.A. | 317 |
| General Dynamics Corp. | 1,50,671 |
| General Motors Corp. | 1,83,967 |
| Heico Corp. | 740 |
| Hexcel Corp. | 1,162 |
| Kaman Corp. | 2,141 |
| Lockheed Martin Corp. | 81,234 |
| Moog, Inc. | 745 |
| Motient Corp. | 784 |
| Northrop Grumman Corp. | 29,878 |
| Orbital Sciences Corp. | 770 |
| Panamsat Corp. | 4,861 |
| Pegasus Communications Corp. | 4,640 |
| Perkinelmer, Inc. | 28,371 |
| Precision Cast Parts Corp. | 9,822 |
| Raytheon Co. A | 31,952 |
| Raytheon Co. B | 25,787 |
| Remec, Inc. | 2,147 |
| Rolls-Royce PLC | 40,110 |
| Smith Group PLC | 9,263 |
| Teledyne Technologies, Inc. | 4,009 |
| Thales (Ex Thomson CFS) | 45,169 |
| Triumph Group, Inc. | 2,875 |
| Zodiac S.A. | 13,429 |

1. Consider the salaries (in thousands of Rupees) for a group of business executives:

177, 54, 98, 57, 209, 56, 45, 98, 58, 90, 116, 42, 142, 152, 85, 53, 52, 85, 72, 45, 168, 47, 93, 49, 79, 145, 149, 60, 58

* 1. Construct a histogram of this data set
  2. Describe the distribution shape
  3. Based on the histogram, what values appear to have been typical for this group of salaries

***II. Landmark Summaries, Variability and Normal Distribution: Dealing with Diversity***

1. Consider the quality of cars, as measured by the number of cars requiring extra work after assembly, in each day’s production for 15 days:

30, 34, 9, 14, 28, 9, 23, 0, 5, 23, 25, 7, 0, 3, 24

1. Find the average number of defects per day.
2. Find the median number of defects per day.
3. Draw the histogram of the data.
4. Find the mode number of defects per day for your histogram in part C.
5. Find the quartiles.
6. Find the extremes (the smallest and largest).
7. Draw a box plot of the data
8. Draw a cumulative distribution function of the data
9. Find the 90th percentile for this data set
10. Find the percentile ranking for the next day’s value of 29 defects
11. Many countries (but not the United States) have a “value-added tax” that is paid by businesses based upon how much value they add to a product (e.g., the difference between sales revenues and the cost of materials). This is different from a sales taxt because the consumer does not see it added on at the cash register. Consider the VAT (value-added tax) percentages for various countries as shown in the Table below:
    1. Draw a histogram of this data set and briefly describe the shape of the distribution
    2. Find the VAT tax level of the average country
    3. Find the median VAT tax level
    4. Compare the average and median. Is this what you expect for a distribution with this shape?
    5. Draw the cumulative distribution function.
    6. What VAT tax level is at the 20th percentile? The 80th percentile?
    7. What percentile is a VAT tax of 10%?

|  |  |
| --- | --- |
| **Country** | **VAT** |
| Belgium | 19.5% |
| Canada | 7.0% |
| Denmark | 15.0% |
| France | 18.6% |
| Germany | 15.0% |
| Greece | 18.0% |
| Italy | 19.0% |
| Japan | 3.0% |
| Luxembourg | 15.0% |
| Netherlands | 18.5% |
| New Zealand | 12.5% |
| Norway | 22.0% |
| Portugal | 16.0% |
| Spain | 15.0% |
| Switzerland | 6.5% |
| Turkey | 12.0% |
| United Kingdom | 17.5% |

1. Based on demand analysis forecasts, a factory plans to produce 80,000 video game cartridges this quarter, on average, with an estimated uncertainty of 25,000 cartridges as the standard deviation. The fixed costs for this equipment are Rs. 72,000 per quarter, and the variable cost is Rs. 1.43 per cartridge produced.
   1. What is the forecast expected total cost of the cartridges produced?
   2. What is the uncertainty involved in this forecast of total cost, expressed as a standard deviation?
   3. Find the coefficient of variation for the number of cartridges produced and for the total cost. Interpret and compare these coefficients of variation.
   4. After the quarter is over, you find that the factory actually produced 100,000 cartridges. How many standard deviations above or below the average is this figure?
   5. Suppose the firm actually produces 200,000 cartridges. How many standard deviations above or below the average is this figure? Would this be a surprise in light of the earlier forecast? Why or why not?
2. The upper management at Simplified Technologies Ltd., have finally noticed that sales forecasts are usually wrong. Last quarter’s sales were forecast at Rs. 18 Lakhs but came in at Rs. 21.3 Lakhs. Sales for the next quarter are forecast at Rs. 20 Lakhs, with a standard deviation (based on previous experience) of Rs. 3 Lakhs. Assuming a normal distribution centered at the forecast value, find
   1. the probability of a “really bad quarter,” which is defined as sales lower than Rs. 15 Lakhs.
   2. the probability of a “really good quarter” with sales in excess of Rs 24 Lakhs.
   3. The probability of a “typical quarter” with sales between Rs. 16 Lakhs and Rs. 23 Lakhs.
3. Assume that the stock market closed at 13,246 points today. Tomorrow you expect the market to rise a mean of 4 points, with a standard deviation of 115 points. Assume a normal distribution.
   1. Find the probability that the stock market goes down tomorrow.
   2. Find the probability that the stock market goes up more than 50 points tomorrow.
   3. Find the probability that the stock market goes up more than 100 points tomorrow.
   4. Find the probability that the stock market goes down more than 150 points tomorrow.
   5. Find the probability that the market changes by more than 200 points in either direction.

***III. Random Sampling, Confidence Interval and Hypothesis Testing***

1. A study used the discharge records of 30 patients from a university hospital in Delhi to find that 35% of patients received antibiotics during their hospital stay. Describe the *population* and *sample* for this study.
2. A study of eighteen 35- to 44-year old male diabetics found that mean body weight was 13% above ideal. Describe the *population* and *sample* for this study.
3. Select a random sample of three without replacement from the following (very small) population of firms: IBM, GM, Ford, Shell, HP, Boeing, and ITT. Use the following sequence of random digits: 5887053671352339.
4. Your agricultural firm is considering the purchase of some farmland, and an indication of the quality of the land will be helpful. A random sample of 62 selected locations planted with corn indicates an average yield of 103.6 bushels per acre, with a standard deviation of 9.4 bushels per acre. Find the two-sided 95% confidence interval for the mean yield for the entire area under consideration.
5. A survey of 15 engineers in fabricated metal products industries, who have been in their current positions from one to three years and were members of the American Society for Quality Control, revealed $37.496 as the average salary. Assume a random sample with a standard deviation of $9000.
   1. Find the 95% confidence interval for the population mean salary.
   2. Complete the following sentence: We are 95% sure that the population mean salary is at least \_\_\_\_\_\_\_\_\_.
6. You are concerned about waste in the newspaper publishing process. Previously, no measurements have been taken, although it is clear that frequent mistakes often require many pounds of newsprint to be thrown away. To judge the severity of the problem and to help you decide if action is warranted, you have begun collecting data. You will take action only if the amount if waste is large enough. So far, on 27 selected mornings, the weight of the wastepaper has been recorded. The average is 273.1 pounds per day, with a standard deviation of 64.2 pounds.
   1. Is it appropriate for you to conduct a one-sided confidence interval for this situation? Why or why not?
   2. Find the most useful one-sided 99% confidence interval. Why did you choose the side you did?
   3. Express your confidence interval in terms of pounds per year, assuming operations continue 365 days per year.
7. The amount of caffeine (milligrams) in randomly sampled cups of coffee was as follows:

112, 86, 16, 110, 100, 93, 102, 116, 93, 117, 106, 82

Find the one-sided 99% confidence interval for the population mean caffeine content of a cup of coffee that claims “at least …. “

1. Some of your advertisements seem to get no reaction, as though they are being ignored by the public. You have arranged for a study to measure the public’s awareness of your brand before and after viewing a TV show that includes the advertisement in question. You wish to see if the ad has a statistically significant effect as compared with zero, representing no effect. Your brand awareness, measured on a scale from 1 to 5, was found to have increased an average of 0.22 point when 200 people were shown an advertisement and questioned before and after. The standard deviation of the increase was 1.39 points.
   1. Identify the null and research hypotheses for a two-sided test, using both words and mathematical symbols.
   2. Perform a two-sided test at the 5% significance level and describe the result.
   3. Perform a two-sided test at the 1% significance level and describe the result.
   4. State the p-value as either *p > 0.05*, *p < 0.05*, *p < 0.01*, or *p < 0.001.*
2. In a random sample of 725 selected for interview from your database of 13,916 customers, 113 said they are dissatisfied with your company’s service.
   1. Find the best estimate of the percentage of all customers in your entire database who are dissatisfied.
   2. Find the standard error of your estimate of the percentage of all customers who are dissatisfied.
   3. Find the best estimate of the overall number of dissatisfied customers within your database.
   4. Find the 95% confidence interval for the percentage of dissatisfied customers.
   5. The company’s goal has been to keep the percentage of dissatisfied customers at or below 10%. Could this reasonably still be the case, or do you have convincing evidence that the percentage is larger than 10%? Justify your answer.
3. The current advertising campaign for a major soft drink brand would be changed if less than 30 percent of the consumers like it.
   1. Formulate a null and alternative hypothesis.
   2. Discuss the type I and type II errors that could occur in hypothesis testing.
   3. Which statistical test would you use? Why?
   4. A random sample of 300 consumers was surveyed, and 84 respondents indicated that they liked the campaign. Should the campaign be changed? Why?
4. To understand your competitive position, you have examined the reliability of your product as well as the reliability of your closest competitor’s product. You have subjected each product to abuse that represents about a year’s worth of wear-and-tear per day. Table below shows the data indicating how long each item lasted.

|  |  |
| --- | --- |
| **Days Until Failure** | |
| You | Competitor |
| 1.0 | 0.2 |
| 8.9 | 2.8 |
| 1.2 | 1.7 |
| 10.3 | 7.2 |
| 4.9 | 2.2 |
| 1.8 | 2.5 |
| 3.1 | 2.6 |
| 3.6 | 2.0 |
| 2.1 | 0.5 |
| 2.9 | 2.3 |
| 8.6 | 1.9 |
| 5.3 | 1.2 |
|  | 6.6 |
|  | 0.5 |
|  | 1.2 |

* 1. Find the average time to failure for your and your competitor’s products. Find the average difference (yours minus your competitor’s).
  2. Find the appropriate standard average for this average difference. In particular, is this a paired or an unpaired situation? Why?
  3. Find the two-sided 99% confidence interval for the mean difference in reliability.
  4. Test at the 1% level if there is a significant difference in reliability between your products and your competitor’s.
  5. Find the p-value for the difference in reliability (as either *p>0.05, p<0.05, p<0.01, p<0.001*).

***IV. Hypothesis Testing for Ordinal Data***

1. Consider the profits of the aerospace firms in the Fortune 500, shown in Table below:

|  |  |
| --- | --- |
| **Aerospace Firm Profits** | |
| Firm | Aerospace Profit |
| United Technologies | 4 |
| Boeing | 4 |
| McDonnell Douglas | 2 |
| Rockwell International | 7 |
| Allied-Signal | 4 |
| Lockheed | 6 |
| General Dynamics | 4 |
| Textron | 3 |
| Northrop | 2 |
| Martin Marietta | 6 |
| Grumman | 2 |
| Gencorp | 3 |
| Sequa | 4 |
| Colt Industries | 5 |
| Sundstrand | -5 |
| Rohr Industries | 4 |
| Kaman | 3 |

* 1. Draw a histogram of these profit percentages. Describe the distribution.
  2. Find the average and the median. Explain why they are either similar or different.
  3. Use the t-test to see if the mean profit (for the idealized population of similar firms operating under similar circumstances) is significantly different from zero.
  4. Use the sign test to see whether the median profit of this idealized population is significantly different from zero.

1. Your firm is being sued for gender discrimination, and you are evaluating the documents filed by the other side. Their data set is shown in the Table below.

|  |  |
| --- | --- |
| **Women** | **Men** |
| 21,100 | 38,700 |
| 29,700 | 30,300 |
| 26,200 | 32,800 |
| 23,000 | 34,100 |
| 25,800 | 30,700 |
| 23,100 | 33,300 |
| 21,900 | 34,000 |
| 20,700 | 38,600 |
| 26,900 | 36,900 |
| 20,900 | 35,700 |
| 24,700 | 26,200 |
| 22,800 | 27,300 |
| 28,100 | 32,100 |
| 25,000 | 35,800 |
| 27,100 | 26,100 |
|  | 38,100 |
|  | 25,500 |
|  | 34,000 |
|  | 37,400 |
|  | 35,700 |
|  | 35,700 |
|  | 29,100 |

* 1. Draw box plots for this data set on the same scale and comment on their appearance.
  2. Use a nonparametric method to test whether these salary distributions are significantly different.
  3. Briefly summarize your conclusions based on the result of this test.

***V. Hypothesis Testing for Nominal Data***

1. Your firm is considering expansion to a nearby city. A survey of employees in that city, asked to respond to the question “Will business conditions in this area get better, stay the same, or get worse?” produced the data-set shown in Table below:

|  |  |  |
| --- | --- | --- |
|  | Managers | Employees |
| Better | 23 | 185 |
| Same | 37 | 336 |
| Worse | 11 | 161 |
| Not sure | 15 | 87 |

* 1. Fill in the “Total” row and column.
  2. Find the table of overall percentages. Interpret these as estimates of probabilities in the population. In particular, what probabilities do they represent?
  3. Find the table of percentages by type of employee. Interpret these as estimates of probabilities in the population. In particular, what probabilities do they represent?
  4. Find the table of percentages by response. Interpret these as estimates of probabilities in the population. In particular, what probabilities do they represent?
  5. Does the response appear to be independent of the employee’s classification? Why or why not?

1. A commercial bank is reviewing the status of recent real estate mortgage applications. Some applications have been accepted, some rejected, and some are pending while waiting for further information. The data is shown in Table below:

|  |  |  |
| --- | --- | --- |
|  | **Residential** | **Commercial** |
| Accepted | 78 | 57 |
| Information requested | 30 | 6 |
| Rejected | 44 | 13 |

* 1. Compare the status of residential to commercial loan applications.
  2. Are the differences between residential and commercial customers significant? How do you know?

1. One group of households was asked how satisfied they were with the car, while the other group was asked how dissatisfied they were. Results are shown in Table below:

|  |  |  |
| --- | --- | --- |
|  | **Satisfied** | **Dissatisfied** |
| Very satisfied | 139 | 128 |
| Somewhat satisfied | 82 | 69 |
| Somewhat dissatisfied | 12 | 20 |
| Very dissatisfied | 10 | 23 |

* 1. Which group was more likely to report that they were satisfied?
  2. Which group was more likely to report that they were dissatisfied?
  3. Are the differences significant? Justify your answer by reporting the chi-squared statistic and its degrees of freedom.