Name: Daljeet Maken (dmm6393)

## Lesson 1 – Homework 1 and practice homeworks

**Notes:**

* The following numbered homework assignments refer to the problems in the textbook*: An Introduction to Statistical Methods and Data Analysis*, e.g., 1.2 is problem 1.2 (found on p.14 in chapter 1 for the fifth edition and p.13 in chapter 1 for the sixth edition).
* The answers to the problems will be posted on the **Sunday after the due date**. To view the answers, click on the **Homework Solutions** link in the menu bar to the left.
* You may use minitab whenever appropriate and include the minitab output in your homework. **To place Minitab output into your homework, highlight the portion of the Minitab printout you want, copy that to the clipboard, and then paste it into the document**

**1.** Provide the following information:

* your name and preferred first name: Daljeet Maken
* your undergraduate major : Engineering
* your favorite book : The Emperor of All Maladies: A Biography of Cancer
* your previous knowledge about any statistical packages : Did 3 courses as a part of MBA

**2.** Problem 1.3 (same for both 5th and 6th edition)

1. Population: All the households in this particular city (city of interest to the researcher) who receive welfare support. Differently put it will be all the households on the city’s welfare rolls.
2. Sample: The subset of 400 households selected from the city welfare rolls.
3. Number of children in households that receive welfare support.
4. We are collecting data only for the 400 households and will use this information to understand/project the number of children in all the households in this particular city who receive welfare support. To go from the small sample to the entire population requires that the inferences are reliable. If the inference isn’t reliable we can’t be confident that the number for the entire population is well represented (will be close) by the number for the sample.

**3.** Problem 2.8 of the 6th edition. (which is problem 2.9 of 5th edition)

1. The surveys will not produce the same results since the framing of the question and the follow up in the second question will induce different kinds of biases.
2. The biases in the two different questions may include:
   1. The first question is somewhat leading in its composition. Instead of asking what type of peanut butter is used at the household, the interviewer asks the question in a formation that may lead to a yes answer
   2. The second question while may address the situation where there is an incorrect answer coming out of the first method, however it will have a different type of issue. Asking the person to show the peanut butter after they have answered may lead to embarrassment and thereby hostility towards the entire process.

**4.** In a study of the relationship for senior citizens between physical activity and frequency of colds, participants were asked to monitor their weekly time spent in exercise over a five-year period and the frequency of colds. The study demonstrated that a negative statistical relation exists between time spent in exercise and frequency of colds. The investigator concluded that increasing the time spent in exercise is an effective strategy for reducing the frequency of colds for senior citizens.

**a.** Were the data obtained in the study observational or experimental (scientific) data?

The data obtained in the study is observational data.

**b.** Comment on the validity of the conclusions reached by the investigator.

Given the limited information about the study, it can be said that the results are invalid because of the following reasons:

1. The natural differences in experimental units (senior citizens) have not been taken into consideration. We don’t know the frequency of colds of these subjects before the commencement of the study.
2. Variation in the measurement – it has not clearly specified that how the subjects are supposed to measure the frequency of colds. When is the condition sever enough to be counted and the mechanisms of how they record this information isn’t specified.
3. Variation in setting the treatment conditions. The subjects have simply been told to measure the frequency of exercise and colds. There are numerous factors that can impact if and when a person may develop the condition. The variations can be in the form of overall climate, house temperature, and diet differences amongst many others. Therefore there can be numerous confounding factors and to simply ignore them is invalid.

**c.** Identify two or three factors that might affect both the time spent in exercise and the frequency of colds for senior citizens simultaneously.

The factors may include diet differences, health supplements, type of exercise (cardio etc), location of exercise (outdoors, at home, in a gym), weather differences (cold vs hot climate), type of lodging, and ambient temperature in the house.

**d.** How might the study be changed so that a valid conclusion about “causal” relationship between the amount of exercise and the frequency of colds be reached?

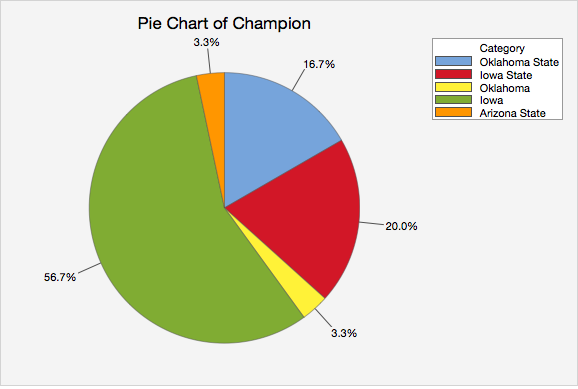
The study has to be changed to incorporate the complexity of the multitude of the factors involved. Since this is an observational study, we can look into dividing the senior citizens (experimental units) into multiple blocks. The blocks are based on demographics and the past health criteria.

**5.** Use minitab to provide **two** different graphs for the following data set:

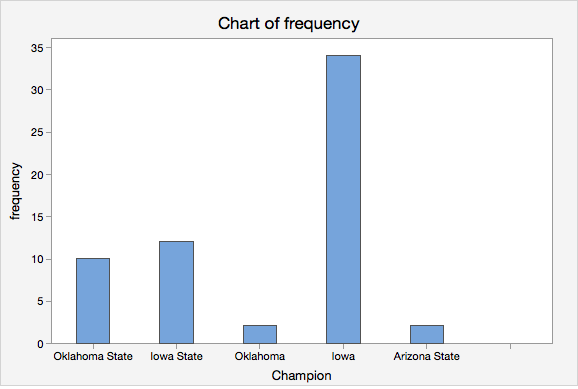
The winners of the NCAA wrestling championships for the years 1968-1997:

|  |  |
| --- | --- |
| Champion | frequency |
| Oklahoma State | 5 |
| Iowa State | 6 |
| Oklahoma | 1 |
| Iowa | 17 |
| Arizona State | 1 |

Pie Chart



Bar Chart



**6**. Get data from <http://lib.stat.cmu.edu/DASL/Datafiles/Singers.html> and use minitab to:

a. draw frequency histograms for Soprano and Alto separately.

|  |  |
| --- | --- |
| Soprano | Alto |
|  |  |

b. draw relative frequency histograms for Soprano and Alto separately.

|  |  |
| --- | --- |
| Soprano | Alto |
|  |  |

c. draw stem-and-leaf diagrams for each type of singer (Soprano, Alto, Tenor, Bass). Use the stem-and-leaf diagrams to find the median height of each group.

|  |  |  |  |
| --- | --- | --- | --- |
| Soprano | Alto | Tenor | Bass |
| Stem-and-leaf of Soprano  N = 36  Leaf Unit = 0.10  1 60 0  3 61 00  9 62 000000  12 63 000  13 64 0  (15) 65 000000000000000  8 66 000000  2 67 0  1 68 0 | Stem-and-leaf of Alto  N = 35  Leaf Unit = 0.10  1 60 0  5 61 0000  8 62 000  12 63 0000  15 64 000  (4) 65 0000  16 66 0000000  9 67 0000  5 68 00  3 69 0  2 70 0  1 71  1 72 0 | Stem-and-leaf of Tenor  N = 20  Leaf Unit = 0.10  1 64 0  2 65 0  5 66 000  7 67 00  10 68 000  10 69 0  9 70 00  7 71 00  5 72 00  3 73 0  2 74 0  1 75  1 76 0 | Stem-and-leaf of Bass  N = 39  Leaf Unit = 0.10  2 66 00  2 67  8 68 000000  11 69 000  19 70 00000000  (6) 71 000000  14 72 000000  8 73 000  5 74 0  4 75 0000 |
| Median: 65 | Median: 65 | Median: 68.5 | Median: 71 |

**indexReading:** *An Introduction to Statistical Methods and Data Analysis*, chapters 1, 2, 3.1, and 3.3.

**Practice problems and its solution:**

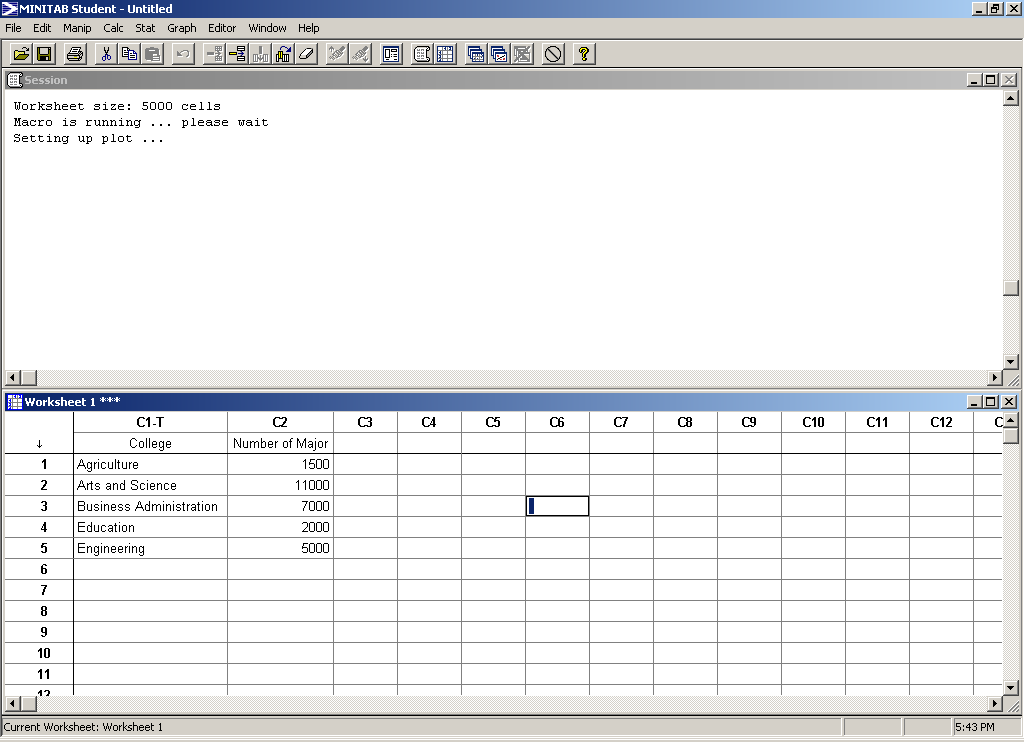
**Practice problem 1: Draw pie chart and bar chart the following problem:**

University officials periodically review the distribution of undergraduate majors within the colleges of the university to help determine a fair allocation of resources to departments within the colleges. At one such review, the following data were obtained:

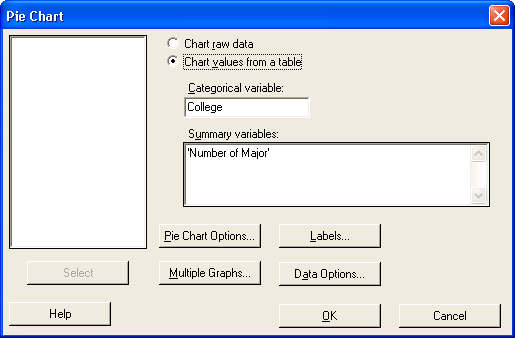
|  |  |
| --- | --- |
| **College** | **Number of majors** |
| Agriculture | 1,500 |
| Arts and Sciences | 11,000 |
| Business Administration | 7,000 |
| Education | 2,000 |
| Engineering | 5,000 |

**Solution:**

Enter the data into minitab as:

:

Here the data has been summarized by a table, thus you choose “chart values from a table” and click in **College** as the “Categorical variable” and **“Number of major”** as the “summary variables”:

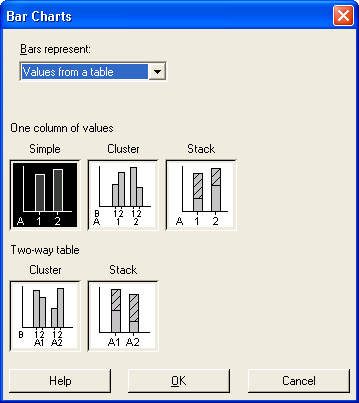


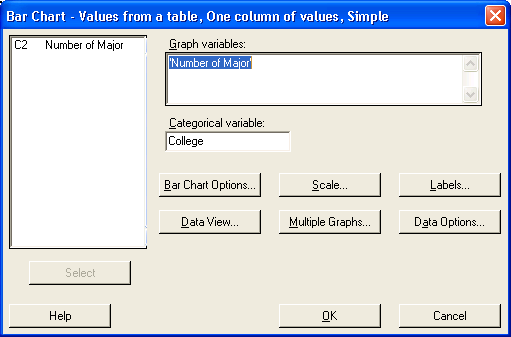
**The pie chart will be generated as given below.** (you can right click on the graph generated by minitab and then choose “Copy graph” to copy that to the buffer of your computer and then paste that to your homework document.)



To draw the bar chart for the problem using minitab, Minitab>Graph>Bar chart,

Then one chooses Bar represent “Values from a table” since the data for this problem is given in a table. You then choose Simple and then click ok.





**The bar chart will be given as:**

.

**Practice problem 2** **: draw frequency histogram and relative frequency histogram for the 1994 annual per capita city given below:**

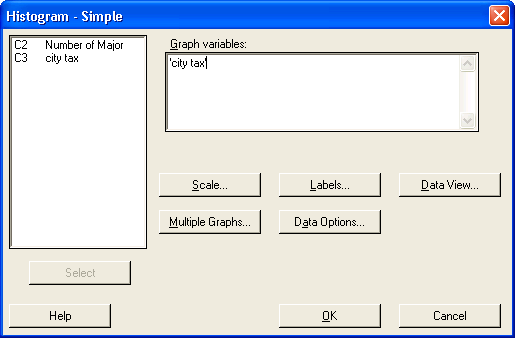
|  |
| --- |
| 2470, 520, 561, 488, 986, 359, 1305, 512, 467, 270, 360, 451, 4904, 572, 498, 382, 271, 634, 1682, 784, 298, 643, 947, 686 |

*Source: Statistical Abstract of the United States, 1997.*

For the data on city tax, one can construct the histogram using minitab by entering the city tax in a column. Minitab will automatically decide the number of intervals and draw a nice histogram.

|  |
| --- |
|  |

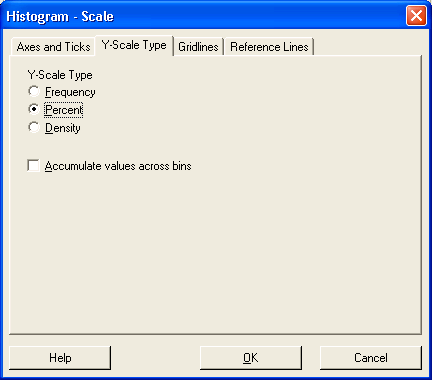
Select “Simple” and then choose ‘city tax’ into graph variables.



The histogram will be given as the output:



If you want the relative frequency histogram, then instead of using the default Y-scale, one has to get into “Scale” and set Y-Scale Type as “Percent”:



The output will then be a **relative frequency histogram** instead of a frequency histogram.



**Practice problem 4: draw the stem and leaf diagram for city tax data given before.**

Minitab>Graph>Stem and leaf, choose “city tax” , then the stem-and-leaf diagram will show up in the session window. One can highlight the display in the session window and copy that to the buffer and then paste to “Word document”.

**Stem-and-Leaf Display: city tax**

Stem-and-leaf of city tax N = 24

Leaf Unit = 100

10 0 2223334444

(10) 0 5555666799

4 1 3

3 1 6

2 2 4

1 2

1 3

1 3

1 4

1 4 9

**Practice problem 3: Problem 1.1 of the textbook**

Answer: a. population of measurements of interest to the researcher are weight of shrimp maintained on a specific diet for a period of 6 months

b. weight of the 100 shrimps randomly sampled

c. mean weight of shrimp maintained on a specific diet for a period of 6 months is the characteristic of interest to the researcher

d. A measure of the reliability of the inferences is important since the sample is only a small proportion of the population, it is necessary to evaluate what the mean weight may be for any other randomly sampled 100 shrimps.