A company is interested in studying job satisfaction within the firm. A simple random survey of $n=300$ full-time employees is drawn from a company list containing the names of all $N=5,000$ full-time employees in order to evaluate job satisfaction via a web-based self-administered survey.

**Homework, Question 1**
- Why do you think the company would want to use a sample?
  - Surveying everyone may be difficult and disruptive
  - It may be expensive
  - It may be impossible
- What are some of the issues involved in measuring job satisfaction?
  - Can a single question measure satisfaction?
  - Satisfaction with what?
    - Pay
    - Hours
    - The ability to effect change or make a difference
    - Their ability to grow and be promoted
    - Their relationship with managers or administrators

**Homework, Question 1**
- What potential problems might you anticipate with a survey of this kind?
  - Will employees trust the survey?
  - Will employees be honest with their answers or give back what they expect the company to hear?
  - Will they respond?

**Homework, Question 2**
- An Excel file can be found on the web site - DRINK.xls. The data are the amount of soft drink in a sample of 50 consecutive 2-liter bottles.
- Construct a histogram of the data using Excel. (you can choose the categories or let Excel do it).
Histogram of Distribution of Amount in 2 liter bottle of soft drink

Homework, Question 2
- How can you describe the distribution - is there any concentration of the amount of soft-drink around specific values?
- Much like the Stem and Leaf plot, the distribution seems to be symmetrical and centered around the 2 liter mark.
- Since the data are consecutive, add a new column of data which is a linear trend (make a new variable that goes from 1 to 50). Plot the amount of soft drink on the Y axis and the trend. What pattern, if any is present in the data?

Scatter Plot of Amount Versus Time

Homework, Question 2
- Say you were making a prediction about the level of the next bottle based on the histogram and then based on the scatter plot. What conclusion would you reach from each graph?
- Based on the histogram and the stem and leaf plot we would assume it is a random process and the level of the next bottle could be above or below the 2 liter mark.
- Based on the scatter plot, once we included time, it is clear that the next bottle will be under-filled. Our model is likely to predict the level very well.

Group Problem 1
- On the web site is a file called BULBS.xls. It contains data on the life of light bulbs (in hours) of two manufacturers. I want you to use Excel, PHStat to make a:
  - Stem and Leaf Plot for each company, and a
  - Box Plot comparing each company.
- The data are in a stacked formation, but you could move the data to have two columns of 40 observations. You can manipulate it any way you wish.
- What does the Stem and Leaf Plot show you about the light bulb life of the two manufacturers? Briefly present and describe the results in Word or Power Point.
- What does the Box Plot show you about the light bulb life of the two manufacturers? Briefly present and describe the results in Word or Power Point.
- Which graph do you think tells the best story?
Comparison of Light Bulb Life for the Two Manufacturers Using the Stem & Leaf Plot

- Manufacturer A has lower bulb life, with most of the values centered around 900 hours. The distribution is skewed to the left with a few low values, around 600 hours.
- In contrast, Manufacturer B has a more symmetrical distribution which is centered around 1,000 hours.

Problem 2

- The data are the amount of soft drink in a sample of 50 consecutive 2-liter bottles.
- Construct a Stem and Leaf Plot using Excel using PHStat. Briefly comment on the result.
- Excel does not do a great job with a Stem and Leaf Plot with this data. However, I want you to build your own stem and leaf plot, in Excel, simply by deciding on the relevant stems, entering the stems into a column, turning on the line in the right of the column, and then entering leafs. Make sure the 50 observations are sorted. With only 50 observations it should not take you too long to accomplish this.
- Briefly describe your new stem and leaf plot.

My first attempt with Excel

I did this in Excel, but by hand.

I used indents, bold, and a cell line to dress it up.

Statistics

- Sample Size: 50
- Mean: 2.0007
- Median: 2.0000
- Std. Deviation: 0.0446
- Minimum: 1.8940
- Maximum: 2.1090
Discussion

- The distribution shows there is variation in the amount of soda in a 2-liter bottle, but the variation is centered slightly above 2 liters, the target mark.
- The distribution is symmetrical and bell-shaped with a few outliers, but none that are extreme.
- This distribution shows no tendency to over or under fill the bottles.

Problem 3: Catalog Sales

- On the web is a data file called CATALOGS.xls. The data are for a random sample of 1,000 customers of a catalog sales company which sells electronic equipment. Each of the variable descriptions can be found as a comment in Excel. The prime dependent variable we will be interested in is SALES - the amount of catalog sales in the past year.
- We will create a XY Scatter Plot of Sales (Y) versus Salary (X). Briefly describe your graph.
- Next, create two new variables: LNSALES, the natural log of Sales and LNSALARY, The natural log of Salary.
- Plot these variables in a similar way, with LNSALES on the Y-axis.
- Describe and compare this graph with the first graph.

Sales versus Salary

- This graph shows a strong linear relationship between the amount of catalog sales and the salary of the customer – as the salary goes up, so do the level of sales.
- However, there is more spread of the data around the linear relationship as salary increases – wealthier customers have more variability in purchases than do less wealthy customers.
Scatter Plot of LNSALES versus LNSALARY

By taking the natural log of each variable we reduce the variability in the data while still preserving the order.

Once we take the log the relationship remains linear – the more salary the more sales.

However, now the variability around the linear relationships is constant. Taking the log of the variables reduced the variability and provided a better estimate of the relationship.