Your Name: ____________________

Section: ____________________

36-201 INTRODUCTION TO STATISTICAL REASONING
Computer Lab #15
Analysis of Infant Mortality in 19th Century Sweden

Objectives:

1. To use your statistical and data analytic expertise to answer substantive research questions about a data set of historical interest. Specifically, you will use exploratory methods (e.g., boxplots, scatterplots, and numerical summaries) and more formal procedures (e.g., confidence intervals and p-values) to investigate relationships in a data set.

2. To apply the general data-analysis approach, which involves the following three main steps: (1) identifying the question, (2) analyzing the data, and (3) drawing conclusions from the data.

Getting Started: For today's lab you will need to copy one data file using the fileserver. The data file is a special Minitab formatted file called *InfMort.MTW*. Get the file now.

Statistical Review

Recall that in any data analysis problem you should first identify which variable is the *response* variable and which is the *explanatory* variable. Then you need to identify whether each variable is quantitative or categorical. The following table summarizes the methods you could use in each case. You should refer to it as needed as you do the lab.

Variables		Methods		
Response (Y)	Explanatory (X)	Exploratory	Formal	Null Hypothesis
Quantitative	Categorical (2 categories)	Boxplots; numerical summary	CI for $\mu_1 - \mu_2$; **Two Sample t-test**	$H_0 : \mu_1 = \mu_2$
Categorical	Categorical	Two-Way Table; Conditional prob	Std. Residuals; **Chi-Square Test**	H_0: No Association between Y and X
Quantitative	Quantitative	Scatterplot; Correlation	**Regression Analysis**	H_0: No Association between Y and X or H_0: Correlation = 0

General Data-Analysis Approach — A Review

Earlier in the semester we discussed a general approach to attacking data analysis problems in Statistics. This approach has three main steps:

1. **Identify the Question**
2. **Analyze the Data**
3. **Draw Conclusions from the Data**

These main steps are all important parts of a thorough data analysis, and each one may require the completion of a series of sub-steps. For example, the first main step **Identify the Question** involves the following sub-steps: describe the problem, state the question in your own words, check the data format, and reflect on the study design. Keep in mind that each sub-step may, in turn, require several "sub-sub-steps" to be completed.

For a given data-analysis problem, the way a step is performed may be different, but the general sequence of steps is the same. This sequence can be viewed as a general template for how to solve data analysis problems. The outline below shows the main steps and their sub-steps. You should always consider these when solving a data-analysis problem.

1. **Identify the Question**
   - A. **Describe the problem**
   - B. **State the question(s)**
   - C. **Check the data format**
   - D. **Reflect on the study design**
2. **Analyze the Data**
   - A. **Identify the relevant variables**
   - B. **Determine the appropriate analysis**
   - C. **Conduct the analysis**
   - D. **Interpret the results**
   - E. **Consider whether additional analyses are necessary**
3. **Draw Conclusions from the Data**
   - A. **Re-state the question(s)**
   - B. **Answer the question(s) based on analyses.**
   - C. **Evaluate the strengths and weaknesses**

This week's lab is structured according to the above outline template. In fact, the above steps and sub-steps appear in bold as headings in this lab handout to structure the material. You will notice, however, that the details under each heading are specific to the data set you are analyzing. This organization is designed to help you see what is common across different statistics problems as well as how the details can vary.

1. Identify the Question

1A. Describe the problem

A common measure of health and quality-of-life of a society is the *infant mortality rate*. The infant mortality rate is the number of infants who die within the first year of life divided by the total number of live births. It is generally believed that the lower the infant mortality rate the better the health and quality-of-life of a society. For example, the infant mortality rate in the United States in 1988 was 1% (10 per 1000 live births).

Professor Katherine Lynch from the CMU History Department is an historical demographer who is interested in studying changes in the infant mortality rate in Sweden, during the 19th century (1800-1899). During this period the infant mortality rate in Sweden dropped from 21% to 10%. What factors are related to this change? Sweden was moving from an agricultural society to a more industrialized society during this period. Could this have been a factor? Did families on the average change their reproductive patterns during this period (e.g., have more/fewer babies, have their babies earlier/later in life, allow more/less time to pass between births)? What factors are related to the infant mortality rate?

All of these questions are of interest to Professor Lynch. The data come from excellent records of births and deaths in Sweden kept by parish priests for all individuals living in their parish. (A parish is a geographical unit like a county.) Our data set is a random sample of size $n = 150$ of families from a much larger data set. Here we are interested in the infant mortality rate of an individual family, which we will call FIMR. For example, if a family has a total of 5 live births and 2 of these children died during their first year, the family infant mortality rate is 2/5=40%.

1B. State the question(s)

Today you will analyze data to answer the following questions: (I) Does our measure of infant mortality show a change during the 1800s? and (II) Is there a change in family occupations during the 1800s?

1C. Check the data format

Use **Open Worksheet** to read the data from the special formatted Minitab file *InfMort.MTW*.

The following is a list of the variables in our data set.

Variable Name	Description
FIMR	Estimate of the family infant mortality rate
MOM.AGE	Mother's age at birth of *first* child
NUM.KIDS	Total number of live births in the family
AVG.BINTV	The average number of months between live births
YR.B1	The year in which the *first* birth occurred
DECADE.B1	The decade in which the first birth occurred. (1 = 1810-1819, 2 = 1820-1829, etc.)
PRE.1850	First birth occurred before 1850 (1800-1849) or after (1850-1899) (1 = 1800-1849, 2 = 1850-1899)
WORKER	Family's working class: Farmer or Other (1 = Farmer, 2 = Other)

Question #1: From the data worksheet fill in the blanks below for the first case (family).

What was the infant mortality rate for this family? ____________________
How old was the mother at the birth of her first child? ____________________
How many children were born to this family? ____________________
What was the average number of months between births? ____________________
What year was the first child born in? ____________________
What was the family's working class? ____________________

1D. Reflect on the study design

Question #2: Is this an experiment or an observational study? Explain why.

2. Analyze the Data

I. Does our measure of infant mortality change during the 1800s?

2A. Identify the relevant variables

To assess whether infant mortality changed during the 1800s, we will examine the relationship between YR.B1 and FIMR.

2B. Determine the appropriate analysis

Question #3:

- Is YR.B1 an explanatory variable or a response? Is it qualitative or quantitative?
- Is FIMR an explanatory variable or a response? Is it qualitative or quantitative?

Question #4: Given your answers above, what type of plot or table is most appropriate (choose: scatter plot, side-by-side boxplots, contingency table, etc.)?

2C. Conduct the analysis

Produce the appropriate plot.

2D. Interpret the results

♣ **Question #5:** Describe the relationship between FIMR and the year in which the mother had her first child, YR.B1.

Question #6: To get a better idea of the relationship described above, use the *lowess* function, i.e., exploratory regression analysis, to describe the time trend for the FIMR. (Click *Display* and choose *Lowess*.) What does this show?

2E. Consider whether additional analyses are necessary

Besides being interested in the above analyses, Professor Lynch believes that there was a change in FIMR from pre-1850 to post-1850. That is, she hypothesizes $\mu_{preFIMR} \neq \mu_{postFIMR}$ or, alternatively, $\mu_{preFIMR} - \mu_{postFIMR} \neq 0$.
Below, we will perform additional analyses to address this specifically.

Question #7: Write down Professor Lynch's null hypothesis and her alternative hypothesis.

Addressing this hypothesis involves examining the relationship between PRE.1850 and FIMR.

Question #8:

- Is PRE.1850 an explanatory variable or a response? Is it qualitative or quantitative?
- Is FIMR an explanatory variable or a response? Is it qualitative or quantitative?

Question #9: Given your answers above, what type of plot or table is most appropriate (choose: scatter plot, side-by-side boxplots, contingency table, etc.)?

Question #10: The variable PRE.1850 is a categorical variable indicating whether the first birth was pre or post 1850. Using boxplots, compare the distribution of FIMR before 1850 to the distribution after 1850.

Question #11: Now, we would like to find a 95% confidence interval for the difference between the average FIMR pre-1850 versus post-1850, i.e., $\mu_{preFIMR} - \mu_{postFIMR} \neq 0$. From the **Stat** menu, select the **Basic Statistics** sub-menu, and *choose* **2-sample t**. Write down the interval for the difference in the FIMR rates.

Question #12: Interpret the interval in the above question. Does it mean that the infant mortality rates before and after 1850 were the same or different? Explain.

♣ **Question #13:** To test Professor Lynch's hypothesis, find the p-value for this test. Is there strong evidence against the null hypothesis?

2F. Summarize analyses produced thus far

Question #14: Summarize the results of your analyses regarding question I: Does our measure of infant mortality show a change during the 1800s?

II. Is there a change in family occupations during the 1800s?

2A. Identify the relevant variables

It is believed that during the latter part of the 1800's Sweden moved from an agricultural to an industrialized society. As such, a smaller proportion of families should be classified as farmers after 1850 than before 1850. We would like to test this.

Question #15: Which two variables in this data set are relevant to this question?

2B. Determine the appropriate analysis

Question #16:

- Is PRE.1850 an explanatory variable or a response? Is it qualitative or quantitative?
- Is WORKER an explanatory variable or a response? Is it qualitative or quantitative?

Question #17: Given your answers above, what type of plot or table is most appropriate (choose: scatter plot, side-by-side boxplots, contingency table, etc.)?

2C. Conduct the analysis

From the **Stat** menu, select the **Tables** sub-menu, and choose **Cross Tabulation**. The explanatory variable is entered first and will be the row variable. Enter the response variable next and it will appear as the column variable in your table. *Select* "Row percents" and "Counts". Also check "Chisquare analysis."

2D. Interpret the results

♣ **Question #18:** Based on the row percents does it appear as if the proportion of farmers has decreased after 1850? Explain.

2E. Consider whether additional analyses are necessary

In addition to the exploratory analysis above, we can perform the relevant formal procedure to test the relationship between WORKER and PRE.1850. Because both of these variables are categorical, a chi-square test is appropriate.

Question #19: In words, state the null and alternative hypothesis of interest.

Question #20: Write down the value of "Chi-square" from your table. We need to find the p-value that corresponds to your Chi-square value. Recall that if $\chi^2 > 3.84$ then the $p - value < .05$. Otherwise the $p - value \geq .05$. What is the p-value in this example? Is there strong evidence against the null hypothesis? Explain.

2F. Summarize analyses produced thus far

Question #21: Summarize the results of your analyses regarding question II: Is there a change in family occupations during the 1800s?

3. Draw Conclusions from the Data

3A. Re-state the Question(s)

(I) Does our measure of infant mortality show a change during the 1800s? and (II) Is there a change in family occupations during the 1800s?

3B. Answer the Question(s) Based on Analyses

Question #22: Given all the analyses above, how would you answer each question:

(I) Does our measure of infant mortality show a change during the 1800s?

(II) Is there a change in family occupations during the 1800s?

♣ **Question #23:** From your interpretation of the above results, would you conclude that it was the industrialization of Sweden after 1850 that caused the decrease in infant mortality? Why or why not?

3C. Evaluate Strengths and Weaknesses

Question #24: Name one alternative hypothesis for why infant mortality decreased in Sweden during the 1800s. Choose a hypothesis that you could test using the current data set.

If you have time, take a look at the relationship between FIMR and AVG.BINTV. You also might like to explore the relationship between FIMR and some of the other variables in the data set.

Remember to **delete** *files and folders that you might have created.*