

Your Name: \_\_\_\_\_

Section: \_\_\_\_\_

**36-201 INTRODUCTION TO STATISTICAL REASONING**  
**Computer Lab Exercise #5**  
**Analysis of Time of Death Data for Soldiers in Vietnam**


**Objectives:**

1. To use exploratory data analysis to investigate the relationship between two categorical variables. Specifically, to study how the conditional distribution of a categorical response variable changes for different categories of an explanatory variable.
2. To review methods for studying the relationship between two quantitative variables, and for the relationship between a quantitative response variable and a categorical explanatory variable.
3. To apply these methods to a data set of historical interest.

**Getting Started**

For today's lab you will need to copy a file using the file server.

**COPYING FILES**

1. Open the **Students** folder. To do this, double click on the "Cluster" icon. Then double click the "students" folder. You will eventually move the data into this folder.
2. In the following order: From the  menu in the upper left hand corner of the titlebar, choose **Chooser**. In the dialog box that appears *click AppleShare*. In the bottom left box called "AppleTalk Zones" *select BH*. In the top right box where it says "Select a file server" select **HSSHE-LIOS**. *Click Ok*.
3. In the dialog box that appears login using "36201" as your Name and "36201" as your Password. Click **Ok**. In the dialog box that appears put a check next to "Class." Click **Ok**.
4. A "Class" icon will appear to the right. Double click on the icon. Then double click on 36201. You will see many files in a folder.
5. Copy all of the files that you will need to the **Students** folder. Do this by dragging the files into the Students folder. For today, the data file you need is *vietnam.dat*.
6. Close all windows.
7. **IMPORTANT:** Drag the *class* icon into the trash. This is necessary to let others in the lab get access to the data.

To start Minitab:

1. *Double-click* the "Cluster" icon.
2. *Double-click* the Applications icon.
3. *Double-click* the Minitab 10.5 folder.

#### 4. Double-click Minitab 10.5.

The data file is not a Minitab formatted file but rather a text file called *vietnam.dat*. Opening the file is a little different than in previous labs.

- From the **File** menu, choose **Open Worksheet**.
- From “List Files of Type” select “Text”.
- Click the **Desktop** button on the right side of the dialog box. Select “Cluster.” Click **Open**. Select “Student Folder.” Click **Open**.
- Select the file called “vietnam.dat” and then *click Open*.

### Part I. Background

#### The Data

Professor John Modell from the CMU history department is interested in investigating what factors, if any, are related to how soldiers learn to survive during wartime. Specifically, he wants to know (1) what factors are related to the rate at which soldiers adapt to the perils of war, and (2) do soldiers learn to avoid death? We will use data from the Vietnam war to study these questions.

The United States official involvement in Vietnam roughly spanned the years 1962 to 1975; the first group of official “advisors” were sent by President Kennedy in November, 1961, and Saigon fell in April, 1975. Approximately 2.7 million troops were deployed over this period, and roughly 58,000 (2.1%) died. The U.S. National Archives maintains a data set which contains various demographic, service history, and date of death data for soldiers *who died* in Vietnam. This is the data set provided to us by Professor Modell that we will use for this analysis. We will use a random sample of 200 male soldiers from this data set.

Professor Modell’s working hypothesis is that if soldiers with certain characteristics tend to die shortly after arriving (i.e., being deployed) in Vietnam, then perhaps these individuals learn more slowly about how to survive in a war zone than others.

**Question #1** There is a key limitation in these data that may prevent us from being able to evaluate Professor Modell’s working hypothesis. What do you think it is? (Don’t spend a lot of time on this.)

The variables in the data set are:

<u>Variable Name</u>	<u>Description</u>
DTIME	Weeks until death following deployment
DATE	Date of deployment (in years past 1900)
MARITAL	Marital Status: 0=single 1=married
RELIGION	Religion: 0=Catholic 1=Protestant
REGION	Region of Origin 1=Northeast 2=South 3=Midwest 4=West
AGE	Approximate age at time of entry into military service
DCAUSE	Cause of Death: 1=Hostile–killed immediately 2=All other

Notice that the different levels of the categorical variables in this data set are given numerical values to denote the different categories. When the levels of a categorical variable are given numerical values, this is called **coding** the variable.

**Question #2** From the data worksheet, fill in the blanks below for case #2:

How many weeks did he survive until he died? \_\_\_\_\_  
What year was he deployed? \_\_\_\_\_  
Was he married? \_\_\_\_\_  
What was his religion? \_\_\_\_\_  
What region of the country was he from? \_\_\_\_\_  
How old was he when he entered military service? \_\_\_\_\_  
Did he die immediately from hostile causes? \_\_\_\_\_

### **Statistical Tools**

The variable of primary interest will be time-to-death following deployment, DTIME. The usual amount of time a soldier served in Vietnam was not longer than 52 weeks. The objective of this exercise is to use exploratory statistical methods to examine relationships among variables in order to investigate Professor Modell's working hypothesis. We will use scatterplots and correlation coefficients to look at relationships between quantitative variables.

What is new in this lab is that we will also look at relationships between two categorical variables. Relationships between categorical variables are described in a display called a **contingency table** and by calculating appropriate percentages from the counts given.

When we describe the distribution of a single categorical variable, we look at the percent of responses in each category, which as we sum these percentages across all categories must total 100%. For example, we might be interested in the distribution of DCAUSE, or specifically, the percentage of soldiers who died immediately from hostile causes. To use Minitab to calculate this value, go to the **Stat** menu, and select the **Tables** sub-menu. Choose **Cross Tabulation**. Type DCAUSE in the box under "Classification Variables". *Click* the check boxes next to "Counts" and "Column Percents". *Click OK*.

♣ **Question #3** How many died immediately from hostile causes? What is the percentage who died immediately from hostile causes? How many total observations in the data set? (Look for the results in the "Session" window.)

**Question #4** Make a **Cross Tabulation** table for the variable REGION. i) What region of the country (i.e., NE, S, MW, W) supplied the most soldiers? What percent of the sample was this?

ii) What region of the country (i.e., NE, S, MW, W) supplied the fewest soldiers? What percent of the sample was this?

### Conditional Distributions

To compare two categorical variables we will look at the distribution of the response variable separately for each category (or level) of the explanatory variable. This is called the *conditional distribution* of the response variable given or conditioned on the values of the explanatory variable. For example, to see if there is any relationship between cause of death, DCAUSE (response variable), and marital status, MARITAL (explanatory variable), we will look at the distribution of DCAUSE for soldiers who were single and compare that to the distribution of DCAUSE for soldiers who were married. This is called the conditional distribution of DCAUSE given MARITAL status.

From the **Stat** menu, select the **Tables** sub-menu and choose **Cross Tabulation**. Type MARITAL and DCAUSE in the box under “Classification Variables”. Make sure that you include a space between the two variable names. Select the “Counts” and “Row Percent” check boxes. *Click OK.*

**Question #5** Copy the table (including counts and row percents) in the space below and label the rows and columns (not with numerical labels but with the appropriate words). Make the table large enough so that you can put additional numbers in the cells. *Be sure you know which variable is on the rows and which is on the columns and what each of the codes for the variables stand for.*

♣ **Question #6 i)** What percent of single men died from hostile causes?

ii) What percent of married men died from hostile causes?

iii) Based on these row percents, is there a relationship between cause of death and marital status? What is it and offer a possible explanation?

## **Part II: Further Analysis of the Vietnam Data**

**Question #7** Is the variable DTIME quantitative or categorical?

**Question #8** Make a histogram of the distribution of death times, DTIME. Describe the distribution of death times.

**Question #9** Find the mean, median, and the quartiles of the distribution of DTIME (time-to-death following deployment).

Next lets look at the relationship between DTIME and AGE. To do this, make a scatterplot of DTIME versus AGE, i.e., age of soldier at time of entry into military service.

**Question #10**

(i) Which variable is the response variable and which is the explanatory variable?

(ii) Describe the relationship between DTIME and AGE.

**Question #11** Find the correlation coefficient between AGE and DTIME. Based on the scatterplot and the correlation coefficient, is AGE a variable that helps explains who will die early after deployment? Why or why not?

Did the distribution of age of soldiers change over the course of the Vietnam War? To address this

question we will look at the relationship between AGE (Y-variable) of the soldiers and the DATE (X-variable) of deployment.

**Question #12** Make a scatterplot of AGE versus DATE. Describe the relationship between AGE and DATE. This will be easier if you use the **Lowess** command. After selecting **Plot**, proceed as follows: Under **Data Display** highlight row 2 in the “display” column. From the pop-up (arrow) menu next to **Display**, choose **Lowess**. Then highlight row 2 in the “For each” column. From the pop-up (arrow) menu next to **For each**, choose **Graph**. Click **Ok**.

**Question #13** Based on your analysis in question #12 does it look like the age distribution of soldiers changed over the course of the Vietnam War? Why or why not?

For the following set of analyses, DTIME (death time) will be the response variable, Y. We want to explore the relationship between DTIME and several categorical explanatory variables: MARITAL, RELIGION, and DCAUSE. For each analysis requested below determine whether there is a relationship between DTIME (Y) and the specified explanatory (X) variable.

**Question #14** What graphical display do we use to investigate the relationship between a quantitative response variable and a categorical explanatory variable?

♣ **Question #15.** Using the display you suggested in question #14, do you think there a relationship between death time, DTIME, and marital status, MARITAL? Without giving too much detail briefly explain why or why not.

**Question #16.** Is there a relationship between death time, DTIME, and religion, RELIGION? Without giving too much detail briefly explain why or why not.

**Question #17** Is there a relationship between death time, DTIME, and cause of death, DCAUSE? Without giving too much detail briefly explain why or why not.

**Question #18** If you have time, go back and rethink your answer to question #1.

*To quit Minitab from the **File** menu, choose **Quit**. Do not save any files. Remember to **delete** files and folders that you might have created.*