INTRODUCTION TO STATISTICAL REASONING

36-201 **Lab #5 -Partial Solutions**

Question #1: The working hypothesis is that soldiers with certain characteristics tend to die earlier after deployment and therefore perhaps learn how to survive more slowly than others. The limitation is that this data set consists only of the soldiers who died. The successes, that is, the soldiers who "learned" how to survive and did survive, are not in this data set. What we can hope to learn about in this data set is the answer to the question among soldiers who died in Vietnam (i.e., conditional on having died) are there certain characteristics that differentiate those who died early from those who died later.

Questions #6-8:

	MARITAL		
DCAUSE	Single (0)	Married (1)	Total
Hostile-Immediately (1)	105 (71%)	31 (58 %)	136
Other (2)	42 (29%)	22 (42%)	64
Total	147 (100%)	53 (100%)	200

About 71% of the single soldiers died immediately from hostile causes, whereas among married soldiers about 58% died immediately from hostile cuases. Thus, single soliders were $\frac{.71}{.58} = 1.2$ times more likely to die from hostile causes. We do not know whether this is a meaningful or important difference. A possible explanation for this difference might be that single soldiers take more risks and are less cautious than married soldiers, or that perhaps single soldiers are given more dangerous assignments than married soldiers.

Questions #10-11: $\bar{x}$=20.3; median=18; $Q_1 = 10$; $Q_3 = 28$. The median time to death is 18 weeks, although 25% of the soldiers who are killed die within the first 10 weeks (Q_1) of deployment. The distribution is positively skewed and has two peaks, one around 10 weeks and the other around 25 weeks. Interestingly, a number of soldiers died within the first week of deployment.

Questions #12-13: Response: DTIME; Explanatory: AGE. AGE is not a variable that helps explain DTIME because the scatterplot does not show any paticular pattern or relationship between AGE and DTIME. The correlation coefficient between these two variables is -0.023, indicating no linear association. In other words knowing a soldiers age doesn't tell us anything about his death time.

Question #14 There doesn't seem to be a substantial change in age as the war progressed, though there appears to be slightly greater variability in the distribution of ages as the war went on. There are two outliers — both are 30-year-olds, one who entered the war in 1964 and the other in 1969.

Question #15 The distribution of ages doesn't seem to change much. The correlation between age and time of entry is 0.072, indicating that there is virtually no relationship between these two variables.

Questions #16-19: In each of these questions we are looking for a relationship between a quantitative response variable (DTIME) and a categorical explanatory (marital status, religion, and cause of death). The method we use is to compare the distribution of the response variable for each level or category of the categorical variable using boxplots. In other words, we look at the distribution of the response variable *conditioned on* or *given* the different values of the explanatory variable. This is called the conditional distribution of the response variable given the explanatory variable.

We do not see any relationship between the response variable and any of the explanatory variables considered. In each case the boxplots for each category of the explanatory variable overlapped almost completely with one another. No differences.