

6.2 German Tank Problem

Purpose: Examine the sampling distribution of different statistics used as point estimators of a population parameter in terms bias and variance.

Reading Assignment: Read through section 6.2.

Problem Description: During World War II, the American and British Intelligence tried to estimate German tank strength but their results were inaccurate and contradictory. Statisticians were consulted to improve the estimates. It was discovered that the serial numbers on the Mark V tanks were consecutive. In other words, the tanks were numbered in a manner similar to 1, 2, 3, ..., N. The goal of the allies was to estimate the total number of tanks N, from a sample of serial numbers on captured tanks. (This lab is loosely based on the lab entitled "How Many Tanks" from the book "Activity-Based Statistics" by Scheaffer, Gnanadesikan, Watkins, and Witmer, published by Springer, 1996.)

Step 1: Suppose the Allies have captured 5 German tanks with the following serial numbers:

399 125 219 216 175

We can regard this as a random sample of size 5 from a discrete uniform distribution on the integers 1 to N. Your goal is to use these 5 numbers to obtain a point estimator of N. Recall the definition of Point Estimator of a population parameter is "a rule or formula that tells us how to use the sample data to calculate a single number that can be used as an *estimate* of the population parameter," (Definition 6.4 on page 269 of your text). The estimator in this case is N - "the total number of tanks manufactured".

STOP AND THINK: With your lab partner, come up with a point estimator for N. That is, develop a rule or formula to plug the 5 serial numbers into for estimating N. Write down the rule or formula for your point estimator. Plug in the 5 numbers above to get an estimate of N using your formula. Your instructor should allow 5 to 10 minutes for the groups to determine a point estimator. If you are unclear as to what is expected of you, please ask your instructor.

Step 2: (Optional) Your instructor will ask each group to describe the point estimator they came up with in step 1. As a class, have a discussion on all the point estimators suggested.

STOP, THINK, AND DISCUSS: Do you think your point estimator is unbiased? Or do you think your estimator systematically under or over estimates the true value of N?

Step 3: We will also compare your point estimator with the following point estimator:

MAX = maximum tank serial number

For the 5 serial numbers given in Step 1, the maximum value is 399. Thus the estimator defined by the maximum serial number gives a point estimate for N of:

MAX = 399 .

STOP AND THINK: The estimator MAX given in step 3 will be biased for N. Why?

Step 4: Start Minitab. In this step we will compare different point estimators of N. To do this, we will have Minitab simulate the **sampling distributions** (see Definition 6.3 on page 263) of the point estimator you determined in step 1 and the point estimator defined in step 3. We will also consider a new point estimator, called MAX2, defined as follows (you don't need to type this formula into minitab at the moment):

$$\text{MAX2} = (6/5) * (\text{maximum tank serial number}) + 1$$

In order to simulate the sampling distribution for each of the point estimators, we will assume a value for N. Let us consider a value $N = 600$. We will now have Minitab simulate a random sample of size 5 from a uniform distribution on the integers 1 to $N = 600$. To do this, we will have to type a small program in Minitab. In your session window, type the following commands:

```
MTB > name c2 'max' c3 'max2' c4 'MyEst'
```

Here is the idea: we will have minitab generate a random sample of 5 tank serial numbers and then compute the point estimates (MAX, MAX2, and MyEst, the estimator you decided upon). These estimates will be put into columns 2, 3 and 4 respectively. Then we will have Minitab repeat this process many, many times to see what happens in repeated sampling.

To do this, select **File** from the menu at the top of the screen followed by **Other Files** and then **Start Storing Macro**. Fill out the dialog box that appears as follows (Note - your dialog box may look slightly different than the one shown below):



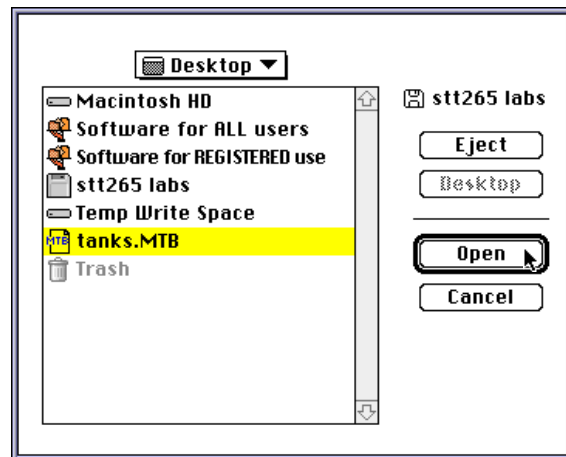
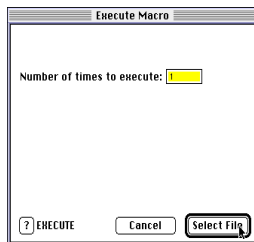
In your session window, **CAREFULLY** type the following commands:

```
STOR> noecho
STOR> random 5 c1;
STOR> integers 1 to 600.
STOR> let k2=max(c1)
STOR> let c2(k1)=k2
STOR> let c3(k1)=(6/5)*k2+1
STOR> let c4(k1)= TYPE YOUR FORMULA HERE!!!!
STOR> let k1=k1+1
STOR> end
```

```
MTB > let k1=1
```

Now we will run the program - we will create 1000 samples. For each sample of 5 tank serial numbers we will compute the MAX, MAX2, and the value of your estimator.

First we run the program just one time to make sure it runs correctly: from the **File** menu, chose **Other files** followed by **Execute Macro**. Fill out the dialog boxes as shown below. In the second dialog box that appears, highlight **tanks.MTB** and click open. This will run the program. Take a look on your data window and see what the macro just did. If your program is running correctly, then next run the program 999 times: follow the same steps above, but replace the number of times to execute by 999. Now sit back for a few minutes because it will take awhile for Minitab to perform the simulations.



Look in your data window. In columns 2, 3, and 4 are the MAX, MAX2, and your estimator for each of the 1000 simulated samples.

Step 5: Make dotplots for columns c2, c3, and c4 putting each on the same scale. Type the following commands in your session window:

```
MTB> dotPlot c2-c4;
SUBC> Same.
```

On each dotplot, mark on the horizontal axis the value $N = 600$.

Also, compute descriptive statistics for each estimator. Choose the menu **Stats**, followed by **Basic Statistics**, followed by **Descriptive Statistics** and select columns c2, c3, and c4.

STOP AND THINK: The dotplots estimate the sampling distribution of each of the point estimators we have considered. Are any of the dotplots centered over the value $N = 600$ approximately? In other words, do any of the statistics appear to be unbiased? Look at the mean of the 1000 estimates given by the descriptive statistics. Which estimator has a mean closest to $N = 600$? Compare the "spread" of each estimator's sampling distribution by looking at the dotplots and the standard deviation given in the descriptive statistics.

Lab Report: Describe the sampling distributions for each of the estimators considered in step 5 by looking at their dotplots and the basic statistics. Which estimator would you choose to estimate N ? Give your reasons. How did your estimator perform in estimating N compared to the other 2?

Lab 6.2, 6/2002