Slide 1:
The German Tank Problem
Diane Evans

Rose-Hulman Institute of Technology
Terre Haute, Indiana

http://www.rose-hulman.edu/~evans/

Abstract

This webinar is based on the activity I found at http://www.lhs.logan.k12.ut.us/~jsmart/tank.htm and other on-line resources (see references). During World War II, the British and U.S. statisticians used estimation methods to deduce the productivity of Germany's armament factories using serial numbers found on captured equipment, such as tanks. The tanks were numbered in a manner similar to 1, 2, 3, …, *N*, and the goal of the allies was to estimate the population maximum *N* from their collected sample of serial numbers. The purpose of this activity is to introduce students to the concept of an unbiased estimator of a population parameter. Students will develop several estimators for the parameter *N* and compare them by running simulations in Minitab.

Slide 2:
Introduction to German Tank Problem

- During World War II, German tanks were sequentially numbered; assume 1, 2, 3, …, *N*
- Some of the numbers became known to Allied Forces when tanks were captured or records seized
- The Allied statisticians developed an estimation procedure to determine *N*
- At the end of WWII, the serial-number estimate for German tank production was very close to the actual figure

Today's German Tank Problem activity is based on this real-world problem

Slide 3:
Alternatives to German Tanks

- Number of buzzers at Panera Bread Company
- Number of taxis in New York City
- Number of iPhones purchased

In 2008 "A London investor called Tommo_UK started asking for people to post the serial numbers of the phone and the date they bought it so that he could decipher how many phones Apple is distributing."[12][13] from this information and using the above formula he was able to calculate that Apple Inc had sold 9.1 million iPhones to the end of September 2008,[14] which meant that they would probably sell more than 10 million in the year.[12] (Wikipedia)

Slide 4:
Learning Goals of the German Tank Problem Activity

- Bring up the topic of estimation before starting statistical inference

  Statistical Inference makes use of information from a sample to draw conclusions (inferences) about the population from which the sample was taken.

  Estimating an unknown parameter of a presumed data model is an intermediate, if not final step, in almost every inference problem.

- What is a parameter? What is an estimator, or a statistic?

  A *parameter* is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic.

  A function of a random sample whose objective is to approximate a parameter is called a *statistic*, or an *estimator*.

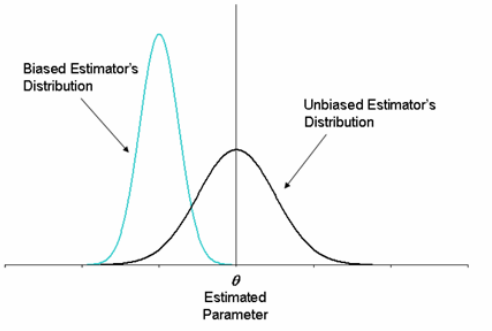
- What is a *good* estimator? What qualities does a good estimator have?
  - Biased versus unbiased estimators
  - Minimum variance estimators

If the mean value of an estimator equals the true value of the quantity it estimates, then the estimator is called an *unbiased* estimator.

If the mean value of an estimator is either less than or greater than the true value of the quantity it estimates, then the estimator is called a *biased*. For example, suppose you decide to choose the smallest observation in a sample to be the estimator of the population mean. Such an estimator would be biased because the average of the values of this estimator would always be less than the true population mean. In other words, the mean of the sampling distribution of this estimator would be less than the true value of the population mean it is trying to estimate. Consequently, the estimator is a biased estimator.

Suppose we are estimating the parameter theta associated with a distribution. Because estimators are random variables, they will take on different values from sample to sample. Typically, some samples will yield $\hat{\theta}$'s that underestimate θ while others will lead to $\hat{\theta}$'s that are numerically too large. Intuitively, we would like the underestimates to somehow "balance out" the overestimates—that is, $\hat{\theta}$ should not systematically err in any one particular direction. (Larsen & Marx, 2006)

Definition: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the discrete pdf $p_X(k, \theta)$, where θ is an unknown parameter. An estimator $\hat{\theta}$ is said to be *unbiased* (for θ) if $E(\hat{\theta}) = \theta$ for all θ.



Slide 5:
Requirements of the Activity

- Level of students: Introductory statistics, probability, or mathematical statistics students
- Classroom size: Works well with 25-30 students; students work in small groups of sizes 3 or 4
- Time to do activity in class: 60 minutes
- Preferable software requirement: Students have access to statistical software, such as Minitab
- Teaching materials
  - Paper sheets with numbers 1 through *N* printed on them
  - Plastic or Styrofoam cups for holding slips of paper 1 through *N*
  - Handouts available at the Cause webinar site

Slide 6:
Instructions for Students

0. Form Allied Statistician Units of size 3 or 4

1. Your unit will obtain (through non-violent military action) a bag filled with the serial numbers of the entire fleet of tanks. Please do not look at the numbers in the bag.

Randomly draw five slips of paper out of the bag without replacement. DO NOT LOOK IN THE BAG. Record your sample:

Sample:
_____,_____,_____,_____,_____

Have someone from your unit write your sample results on the board for your military unit.

Slide 7:

2. Discuss in your group how you could you use the data above (and only this data) to estimate the total number of "tanks" (slips of paper) in the bag. Allow yourself to think "outside the box."

Here are some ideas (not necessarily correct or incorrect) to get you started:

(a). Use the largest of the five numbers in your sample.
(b). Add the smallest and largest numbers of your sample.
(c). Double the mean of the five numbers obtained in your sample.

Slide 8:

3. Come up with an estimator for determining the total number of "tanks" (slips of paper) *N* in the cup. That is, develop a rule or formula to plug the 5 serial numbers into for estimating *N*.

Write down your military unit's formula for estimating *N*:

Slide 9:

4. Plug in your sample of 5 serial numbers from #1 to get an estimate of *N* using the formula your unit constructed.

5. Apply your rule to each of the samples drawn by the other groups (on the board) to come up with estimates for *N*. Construct a dot plot of these estimates below.

<----o----o----o----o----o----o----o----o----o----o----o----o----o----o----o----o----o----o----o---->

Estimates for *N* using each group's sample values

Slide 10:

6. Do you think your point estimator is ***unbiased***? Or do you think your estimator systematically under or over estimates the true value of *N*, which would mean it is ***biased***?

For example, the formula or rule "choose the max of the sample" is biased – why?

7. Calculate the mean of the estimates you obtained for *N* (using each unit's data) from #5.

Sample mean =

Calculate the variance of the estimates you obtained for *N*.

Sample variance = $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

Have a person from your unit record the mean and variance in the front of the room on the white board in the designated area.

Slide 11:

8. In your group, decide on what you think the true value of *N* is. Record it.

9. I will give you the correct value of *N* after the majority of the units are done. It is:

N =

Did you make a "good" estimate in #8? Why or why not? Did you have a good estimation formula?

Is any unit's dotplot centered about the value *N* = ______ approximately? In other words, do any of the estimators (formulas) appear to be unbiased?

Slide 12:

10. The records of the Speer Ministry, which was in charge of Germany's war production, were recovered after the war. The table below gives the actual tank production for three different months, the estimate by statisticians from serial number analysis, and the number obtained by traditional American/British "intelligence" gathering.

Month	Actual # of Tanks Produced	Allied Statisticians Estimate	Estimate by Intelligence Agencies
June 1940	122	169	1000
June 1941	271	244	1550
Sept 1942	342	327	1550

Slide 13:

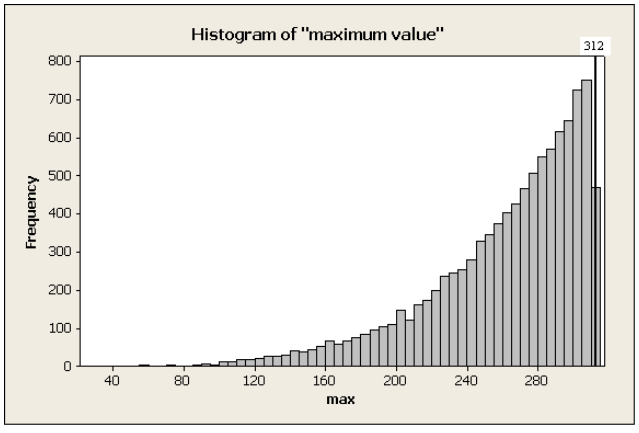
11. In Minitab, simulate this experiment of drawing 5 numbers and using your formula to estimate the number of tanks. Plot the values (in histograms or dotplots) you obtain for *N* using 10,000 simulations (of drawing 5 numbers and then computing *N*).

How to do this in Minitab?
Calc > Random Data > Integer

Number of rows of data to generate: 10000
Store in Columns: C1 – C5
Minimum Value: 1; Maximum Value: *N*

Then use Calc > Row Statistics to enter your specific formula; or Calc > Calculator if your specific formula is not available in Row Statistics.

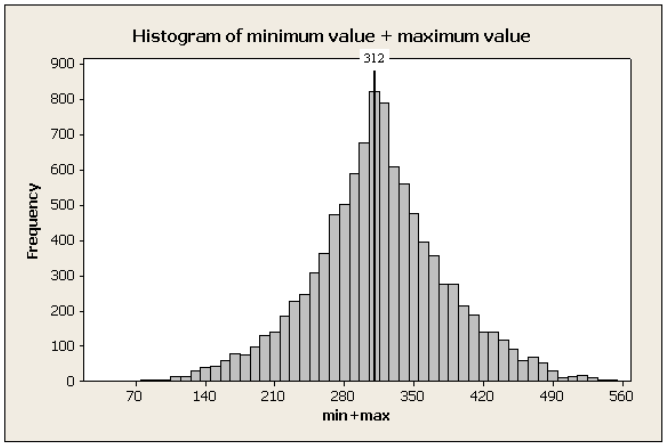
Slide 14:
Simulations of Some Possible Methods



Descriptive Statistics: max value

```
Variable    Mean  StDev
max       261.25  43.53
```

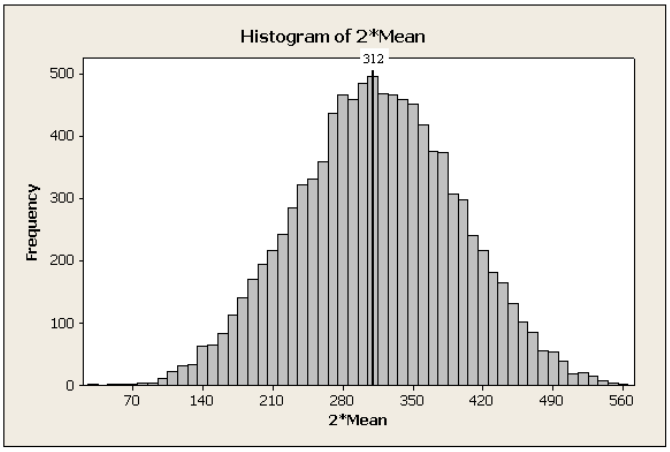
Slide 15:



Descriptive Statistics: min value + max value

Variable	Mean	StDev
min+max	314.00	68.12

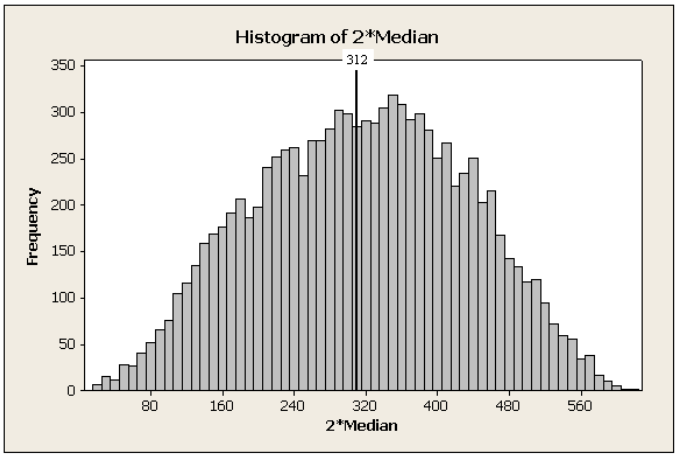
Slide 16:



Descriptive Statistics: 2*Mean

Variable	Mean	StDev
2*Mean	314.17	80.72

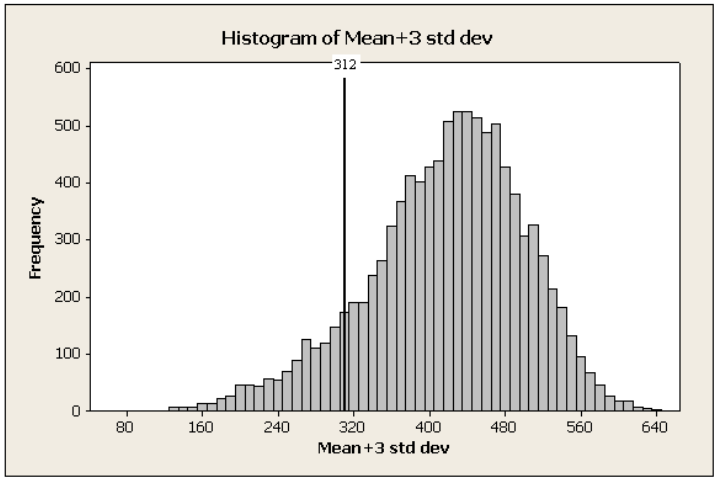
Slide 17:



Descriptive Statistics: 2*Median

Variable	Mean	StDev
2*Median	315.06	117.35

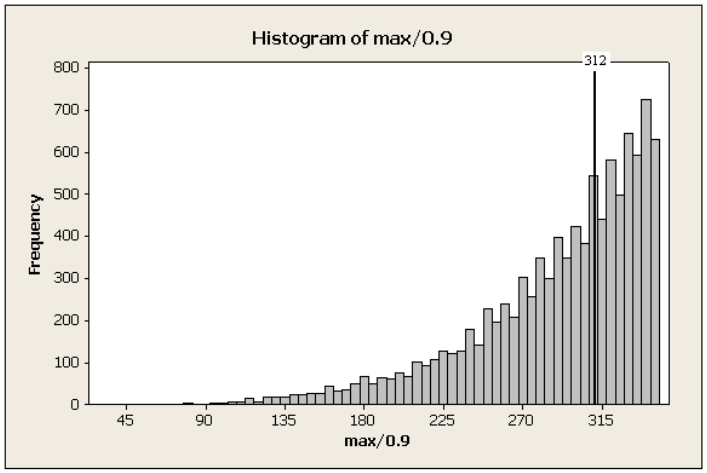
Slide 18:



Descriptive Statistics: Mean+3 std dev

Variable	Mean	StDev
Mean+3 std dev	417.78	82.96

Slide 19:



Descriptive Statistics: max/0.9

Variable	Mean	StDev
max/0.9	290.28	48.37

Slide 20:

Show that $\hat{\theta} = 2\bar{Y}$ is an unbiased estimator of θ, where θ is the finite upper bound of the continuous uniform distribution $f_Y(y, \theta) = \frac{1}{\theta}, 0 \leq y \leq \theta$.

$$E(\hat{\theta}) = E\left(\frac{2}{n}\sum_{i=1}^{n} Y_i\right) = \frac{2}{n}\sum_{i=1}^{n} E(Y_i) = \frac{2}{n}\sum_{i=1}^{n} \frac{\theta}{2} = \frac{2}{n} \cdot \frac{n\theta}{2} = \theta.$$

Slide 21:

Show that $\hat{\theta} = Y_{max}$ is a biased estimator of θ, where θ is the finite upper bound of the continuous uniform distribution $f_Y(y,\theta) = \frac{1}{\theta}, 0 \leq y \leq \theta$. Determine what must be done to $\hat{\theta}$ so that it is an unbiased estimator.

Recall from order statistics: $f_{\hat{\theta}}(x) = f_{Y_{max}}(x) = n \cdot \frac{1}{\theta} \cdot \left(\frac{x}{\theta}\right)^{n-1}, 0 \leq x \leq \theta.$

Thus,

$$E(\hat{\theta}) = \int_0^\theta x \cdot \frac{n}{\theta} \cdot \left(\frac{x}{\theta}\right)^{n-1} dx = \frac{n}{\theta^n} \cdot \frac{x^{n+1}}{n+1}\Big|_0^\theta = \frac{n}{n+1}\theta.$$

Slide 22:

References:

http://mtsu32.mtsu.edu:11281/classes/math2050_new/coursepack/final/12_germantank_bcL7.doc

http://web.mac.com/statsmonkey/APStats_at_LSHS/Teacher_Activities_files/GermanTanksTeacher.pdf

http://web.monroecc.edu/manila/webfiles/beyond/2003S022S071Bullard.pdf

http://www.lhs.logan.k12.ut.us/~jsmart/tank.htm

http://www.math.wright.edu/Statistics/lab/stt264/lab6_2.pdf

http://www.weibull.com/DOEWeb/unbiased_and_biased_estimators.htm

Larsen, R J. and M. L. Marx (2006). *An Introduction to Mathematical Statistics and Its Applications*, 4th Edition, Prentice Hall, Upper Saddle River, NJ.