

The German Tank Problem



Diane Evans



Rose-Hulman Institute of Technology
Terre Haute, Indiana

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

Introduction to the German Tank Problem

- During World War II, German tanks were sequentially numbered; assume $1, 2, 3, \dots, N$
- 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

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 Londoner started asking for people to post the serial number of their phone and the date they bought it

From the posted information and using estimation formulas, he was able to calculate that Apple had sold 9.1 million iPhones by the end of September 2008

Learning Goals of the German Tank Problem Activity

- Bring up the topic of estimation before starting statistical inference
- What is a parameter? What is an estimator, or a statistic?
- What is a *good* estimator? What qualities does a good estimator have?
 - Biased versus unbiased estimators
 - Minimum variance estimators
- Simulation is a powerful tool for studying distributions and their properties

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
 - Brown lunch bags for each group for holding the cut out slips of paper 1 through N
 - Handouts available at the Cause webinar site

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.

2. Discuss in your group how you could 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:

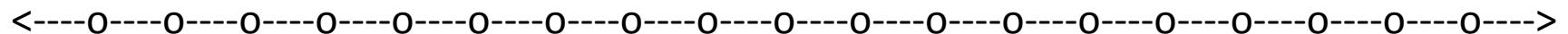
- (a). Use the largest of the five numbers in your sample.
- (b). Add the smallest and largest numbers of your sample.
- (c). Double the mean of the five numbers obtained in your sample.

3. Come up with an estimator for determining the total number of “tanks” (slips of paper) N in the bag. 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 :

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.



Estimates for N using each group's sample values

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 .

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

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.

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 or histogram centered about the value $N =$ _____ approximately? In other words, do any of the estimators (formulas) appear to be unbiased?

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

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

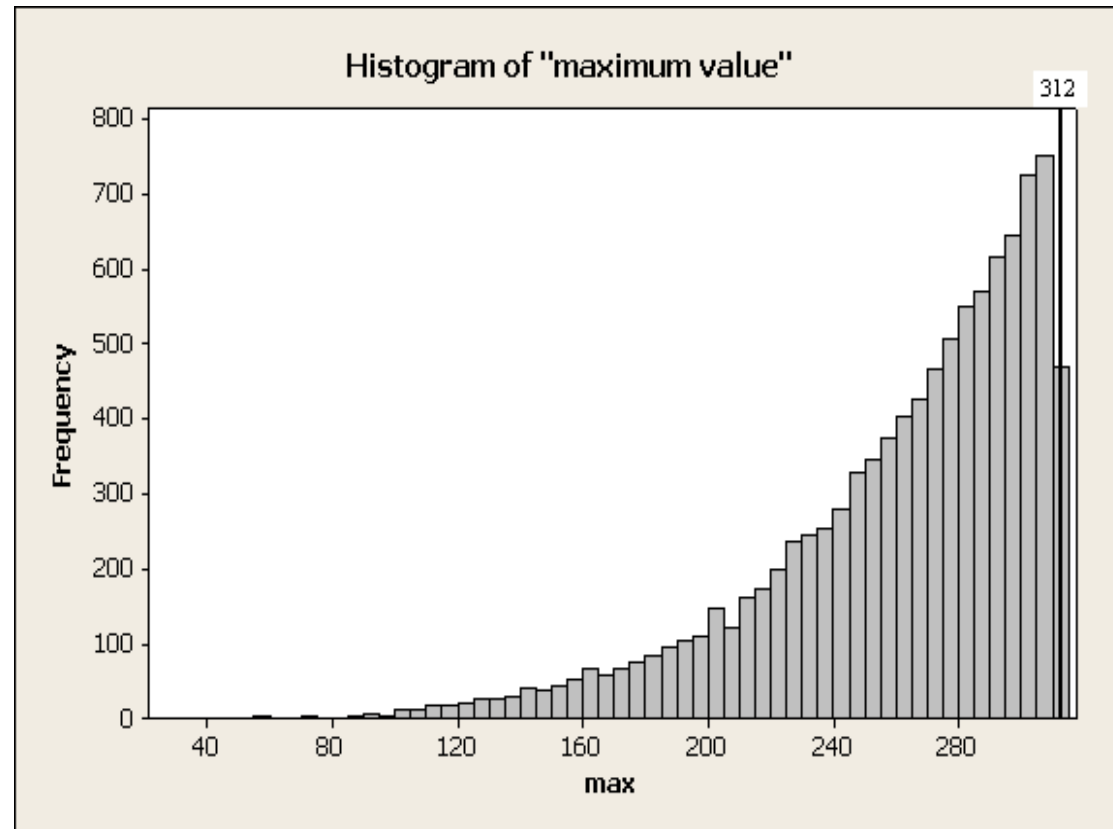
Number of rows of data to generate: 10000

Store in Columns: C1 – C5

Minimum Value: 1; Maximum Value: N

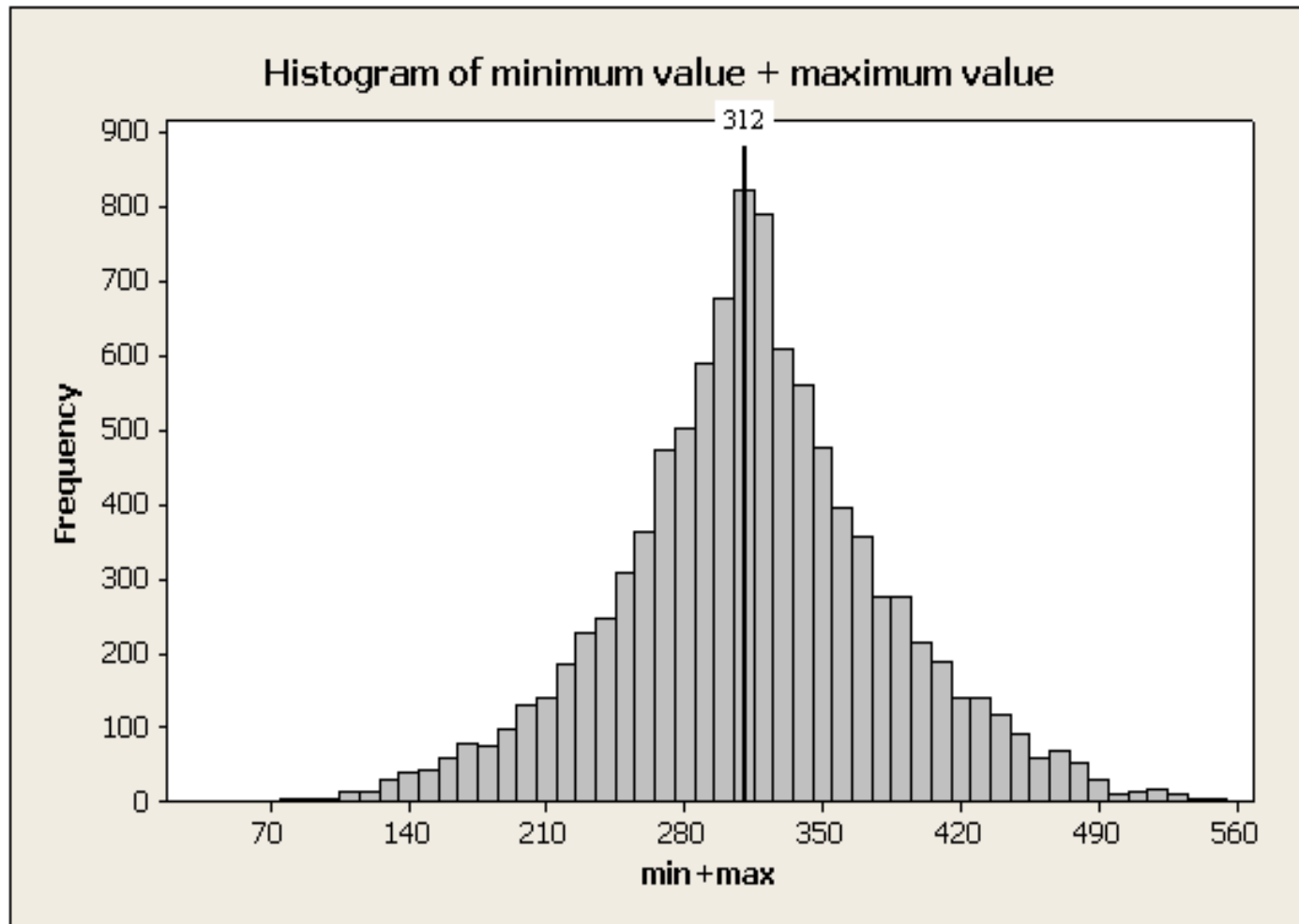
Then use Calc > Row Statistics to enter your specific formula.

Simulations of Some Possible Methods



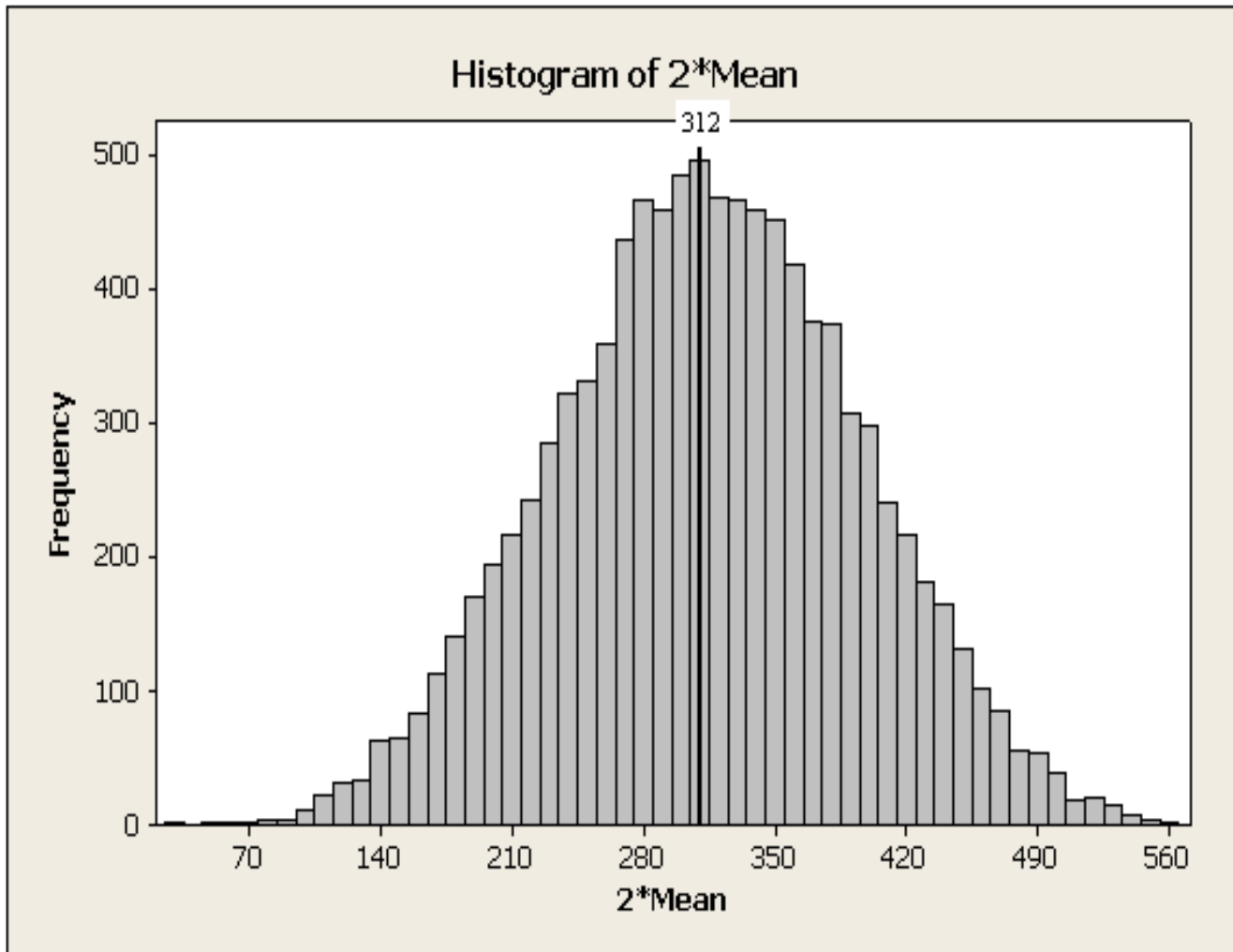
Descriptive Statistics: maximum value

Variable	Mean	StDev
max	261.25	43.53



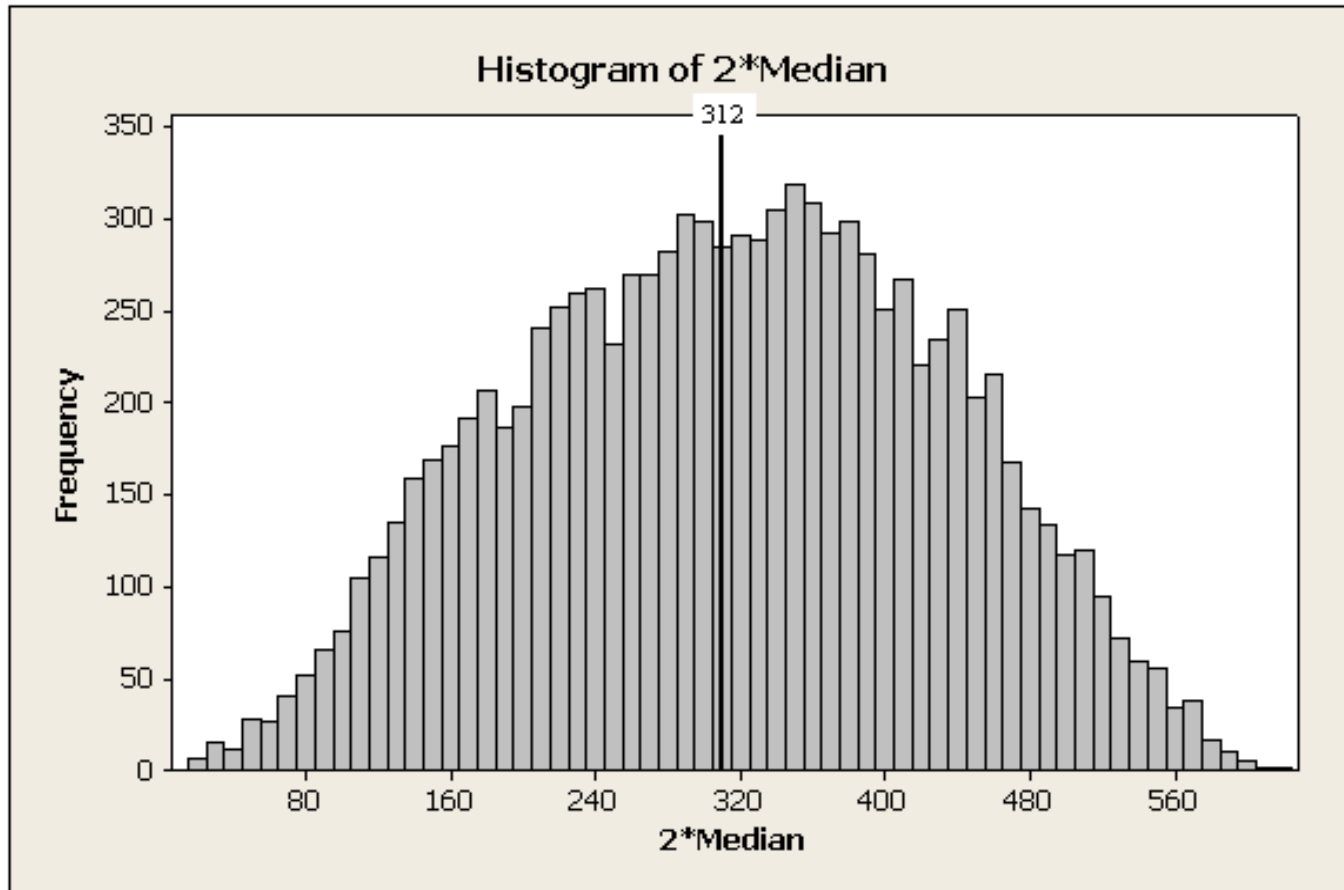
Descriptive Statistics: min value + max value

Variable	Mean	StDev
min+max	314.00	68.12



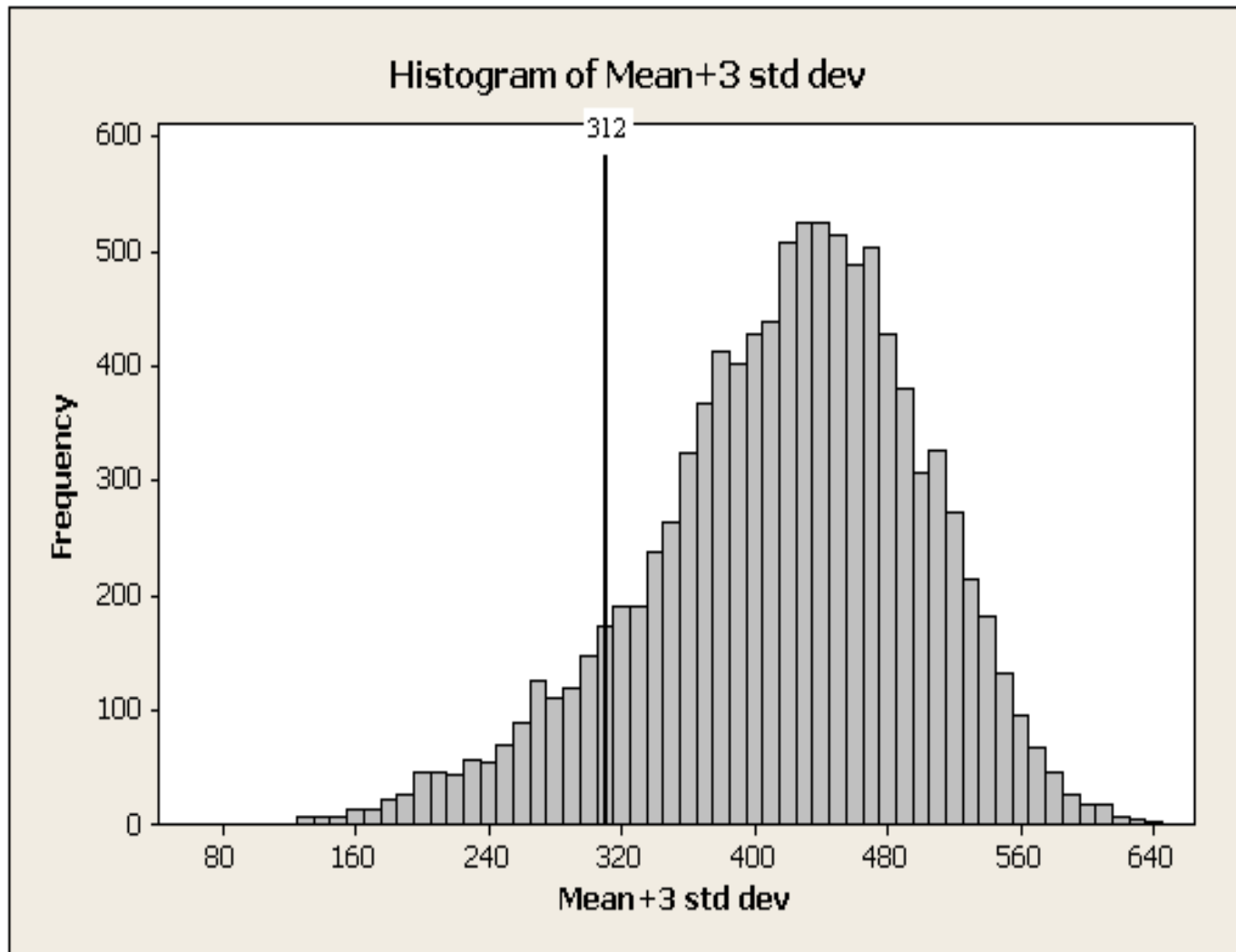
Descriptive Statistics: 2*Mean

Variable	Mean	StDev
2*Mean	314.17	80.72



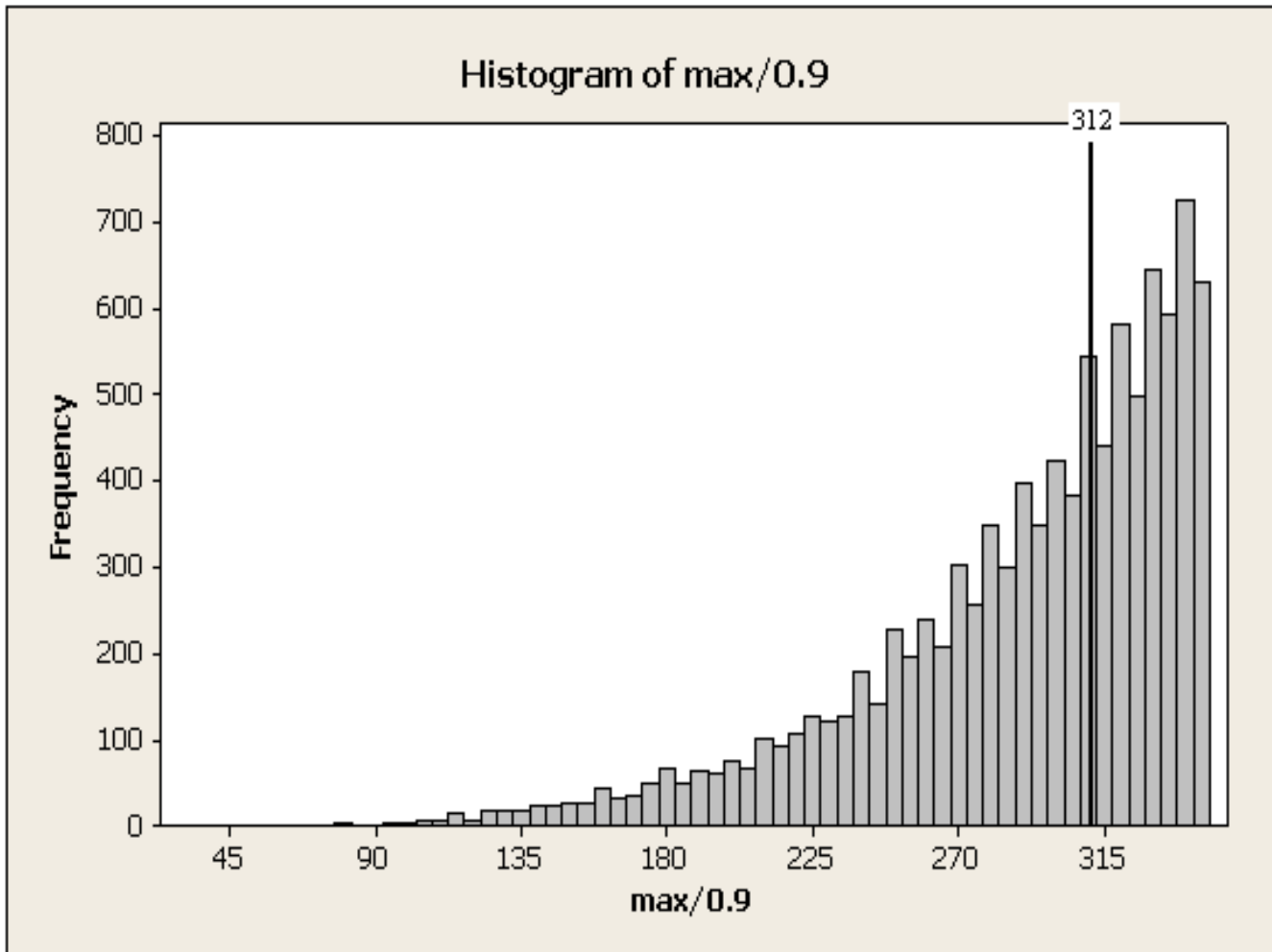
Descriptive Statistics: 2*Median

Variable	Mean	StDev
2*Median	315.06	117.35



Descriptive Statistics: Mean+3 std dev

Variable	Mean	StDev
Mean+3 std dev	417.78	82.96



Descriptive Statistics: max/0.9

Variable	Mean	StDev
max/0.9	290.28	48.37

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 $f_Y(y, \theta) = \frac{1}{\theta}, 0 \leq y \leq \theta$.

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

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,

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

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.