

Stat 321 - Day 36
Introduction to Point Estimation

Recall that a *random sample* is a series of independent and identically distributed (i.i.d.) random variables, typically denoted by X_1, X_2, \dots, X_n . When a sample is drawn from a finite *population* of values in such a manner that all possible samples are equally likely, then the observations are not strictly independent, but we will treat them as approximately so provided that the population is at least twenty times larger than the population. We typically denote a population size by N and a sample size by n , so this requires that $N \geq 20n$ in order for the sample to be considered a random sample in the technical sense.

Recall that a *statistic* is a function of the random variables in a random sample, and the probability distribution of a statistic is known as its *sampling distribution*. You have studied how to find sampling distributions exactly for very small samples, approximately through simulation, and by the Central Limit Theorem when the statistic of interest is a sample mean.

Those random variables X_1, X_2, \dots, X_n have some probability distribution, and associated with that probability distribution are parameters. One important goal is to use a sample statistic to estimate a population parameter. A *point estimator* is a statistic (i.e., a function of random variables) used to estimate a parameter. A *point estimate* is a calculated value of a point estimator for given sample data.

Example: Uniform Guessing

Consider a random sample of five random variables X_1, X_2, X_3, X_4, X_5 from a uniform distribution on the interval $(\theta, 2\theta)$, where θ is the unknown parameter to be estimated. In particular, suppose that the sample values turn out to be: 37.4, 32.5, 44.9, 23.9, 31.6.

- (a) Use these sample values to make a guess for the unknown value of θ .
- (b) Calculate the sample mean of these five values. Does this produce a reasonable estimate of θ ? Explain.

(c) Let $\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$ denote the sample mean of these five random variables.

Determine the expected value of \bar{X} .

- (d) Propose an estimator (call it T_1) based on a slight adjustment to \bar{X} , with the property that $E(T_1) = \theta$.

- (e) What estimate does this new estimator produce for the sample of five values above? Does this seem like a more reasonable estimate of θ ? Explain.

In evaluating the performance of competing point estimators, two criteria are often used:

1. A point estimator is said to be *unbiased* if its expected value equals the parameter it seeks to estimate. In other words, an unbiased estimator averages out to the desired parameter value if the sampling is repeated indefinitely.
2. Other things (such as unbiasedness) being equal, a point estimator with small variance is preferred to one with large variance.

The ideal estimator is unbiased and has *minimum variance*. Unbiasedness is but a minor virtue if individual estimates are widely varied. On the other hand, small variance is not helpful if the long-term average value of the estimator misses its target. These properties of estimators can be investigated theoretically using properties of expected value and variance. They can also be investigated empirically through simulation.

- (f) Determine the variance of the estimator T_1 .

- (g) Use simulation to investigate the performance of the estimator T_1 when the actual parameter value is $\theta=23$. Generate 1000 simulated samples of size 5, and calculate the estimate for each:

```
MTB> random 1000 c1-c5;
SUBC> uniform 23 46.
MTB> rmean c1-c5 c6
MTB> let c7=2*c6/3
MTB> name c6 'x-bar' c7 't1'
MTB> hist c7
MTB> describe c7
```

- (h) Does the simulation reveal the unbiasedness of the estimator T_1 ? Explain how.

- (i) Is the standard deviation of the simulated estimates close to the theoretical value? Verify.

Now consider a new estimator of θ : $T_2 = \frac{\max\{X_i\} + \min\{X_i\}}{3}$. Determining the expected value

and variance of this estimator theoretically is beyond our scope, but we can use simulation to assess its performance empirically.

(j) Calculate estimates based on this estimator for your 1000 simulated samples:

```
MTB> rmax c1-c5 c8
MTB> rmin c1-c5 c9
MTB> let c10=(c8+c9)/3
MTB> name c8 'max' c9 'min' c10 't2'
MTB> hist c10
MTB> describe c10
```

(k) Based on your simulation analysis, does T_2 appear to be an unbiased estimator? Explain.

(l) Based on your simulation analysis, does T_2 appear to have smaller variance than T_1 ? Which estimator would you prefer? Explain.

Example: “German Tank” Problem

During World War II, Allied Intelligence (the spies) gave reports on the production of tanks and other war materials that varied widely and were somewhat contradictory. In 1943 Allied statisticians (the geeks) trying to improve on these estimates developed a method that used information in the serial numbers stamped on captured equipment. One particularly successful venture was the estimation of the number of Mark V tanks, whose serial numbers were conveniently highly correlated to the order of its manufacture. It wasn't long before statisticians figured this out and exploited it to come up with such an estimate. Capturing tanks was like randomly drawing an integer from this sequence. Let the parameter N = total number of tanks produced based on the serial numbers. Consider the following potential estimators of N :

- | | | |
|--------------------------|--------------------------|-----------------------|
| 1. sum of all the values | 2. mean + median | 3. mean*2 |
| 4. mean*3 | 5. median*2 | 6. median*3 |
| 7. mean + 2*(std dev) | 8. mean + 3*(std dev) | 9. maximum value |
| 10. maximum + minimum | 11. maximum + minimum -1 | 12. maximum + range/2 |
| 13. maximum + mean | 14. maximum + median | 15. maximum + std dev |
| 16. maximum *(n+1)/n | 17. 2*range | 18. 3*range |

Your task is to evaluate *three* estimators based on the criteria of (1) unbiasedness and (2) minimum variance. You are to choose at least two of your estimators from the list above: one must involve the maximum value and one must involve the mean or median. The third estimator can be one of your own devising or another one from the list.

Your analysis will use simulation to approximate the sampling distributions for your estimators to determine which perform “better.” To do this, you will have to specify what value N has in order to see which estimators come “closest.” You are to analyze at least two different values of N and at least two different values of n (in order to investigate whether your results are dependent on those values) for each of your three estimators. Thus, you will do a total of twelve ($3 \times 2 \times 2$) analyses. Remember to keep $N \geq 20n$.

Minitab hints:

- To create a population of, say, $N=100$ tanks:
MTB> set c1 DATA> 1:100 DATA> end
- To sample, say, $n=5$ tanks from that population:
MTB> sample 5 c1 c2
- To calculate the value of your point estimate (say, for estimator #11 from the list above) for the sample in C2:
MTB> let c3=max(c2) + min(c2) - 1
- Set up a macro to repeat these commands, e.g., create a file in Notepad with the following commands:
sample 5 c1 c2
let c3(k1)=max(c2)+min(c2)-1 **substitute your estimator**
let k1=k1+1
- Save this file with an .mtb extension (e.g., “tanks.mtb”), putting quotes around the filename.
- To execute the macro:
 1. Initialize your counter: MTB> let k1=1
 2. Choose File > Other Files > Run an Exec from the menu indicate that you want to execute the macro say 1000 times, this will control the number of samples.
 3. Change folders to find the file on your disk (e.g., search for file name: *.mtb).
- You might consider storing the results for your three estimators in three different columns, e.g., C3, C4, and C5. You can put all three into a single macro. Don’t forget to reinitialize k1 and erase previous results (MTB> erase c3-c5) whenever you restart the macro.

Creating your estimators (below are some potentially helpful Minitab reminders):

- To add numbers in a column: sum(c2)
- To find the max or min of the column: max(c2), min(c2)
- To find the mean and :mean(c2) median(c2)
- To find the standard deviation: stan(c2) or std(c2)
- To raise all values in a column to a power (say 2): let c4=c2**2

Examining your empirical sampling distribution:

You will want to examine the mean, standard deviation, and a graphical summary (e.g., histogram) of your sample results. An unbiased estimator should have the mean of the sampling distribution close to N . Compare the standard deviations to determine which has the smallest variance.

Note: Investigation 17 will ask you to write a report summarizing your analysis of these estimators. Be sure to save your Minitab work (File> Save Project As...) often.