*Statistics and their Sampling Distributions* Very often we would like to make an inference or just say something interesting about a population. Since data may be difficult or expensive to come by we often find ourselves working with a sample from a population and trying to use sample data to estimate a characteristic of a population.

Two simple examples.

1. We would like to know the weight of the heaviest elephant in Cameroon. For obvious reasons we can't weigh all the elephants, so we weigh a sample. On the basis of this sample data how would you estimate the weight of the heaviest elephant?
2. You would like to know how well students from SUNYIT who wish to attend graduate school do on the Graduate Record Exam. For privacy and other reasons you won't have access to everyone's scores. How would you estimate the average score?

Since we are often led to compute quantities from a sample, define a **statistic** simply as a value computed from a sample ("something you compute from the data"). Notice that a statistic is itself a random variable and so will have a probability distribution, called a sampling distribution.

As an example, suppose a population is distributed normally with unknown mean $\mu$ and unknown variance $\sigma$. If we collect $n = 10$ individuals at random (more on this later) and denote the quantity of interest for each individual as $X_1, X_2, \dots X_{10}$, we might want to use the sample mean, denoted $\bar{X}$, as an estimate of $\mu$. Here

$$\bar{X} \equiv \frac{\sum_{i=1}^{n} X_i}{n}$$

is a statistic. If we wish to estimate $\sigma$ the following statistic is useful.

$$\chi^2 \equiv \frac{(n-1)S^2}{\sigma^2}$$

where

$$S^2 \equiv \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}$$

Other examples we will look at include

$$t \equiv \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$z \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$F \equiv \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

We will spend some time seeing how these quantities are distributed. To begin, first define a simple random sample as a set of random variables $X_1, X_2, \ldots, X_n$ where the $X_i$'s are independent, identically distributed. For example, when sampling from a normal population we would write $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

*Example* Using moment generating functions we have shown that the sum of independent normally distributed random variables is again normally distributed. Let $X_1, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find the sampling distribution of $\bar{X}$. Also find the sampling distribution of $Z$ as defined above.

*Estimating the mean of a population, $\mu$.* We often wish to know the mean of a population. To do this, a natural procedure would seem to be to take a simple random sample from the target population, calculate the sample mean $\bar{x}$ and use this as an estimate of $\mu$. Some questions immediately come to mind:

- Is $\bar{X}$ a good way to estimate $\mu$?
- Are there better ways?
- Do we, upon repeated sampling, have $\bar{x}$'s which cluster around $\mu$?
- How spread out, upon repeated sampling, are the $\bar{x}$'s from one another?

We need a few definitions. To begin, we will call a statistic $\widehat{\theta}$ a *point estimator* of a population parameter $\theta$ if it is a sensible value to use in the place of $\theta$. In our example $\bar{X} = \widehat{\theta}$ and $\mu = \theta$.

A point estimator is said to be unbiased if $E[\widehat{\theta}] = \theta$. Call the bias of an estimator $E[\widehat{\theta}] - \theta$.

Show that $\bar{X}$ is an unbiased estimator of $\mu$.

Show that $S^2$ is an unbiased estimator of $\sigma^2$.

Show that $\hat{p} = \frac{X}{n}$ is an unbiased estimator of a (infinite) population proportion.

If we wish to estimate a population parameter $\theta$ in an unbiased way (and not everyone does) then how do we select among competing estimators? For example, suppose you have a population uniformly distributed from $-c$ to $c$. That is,

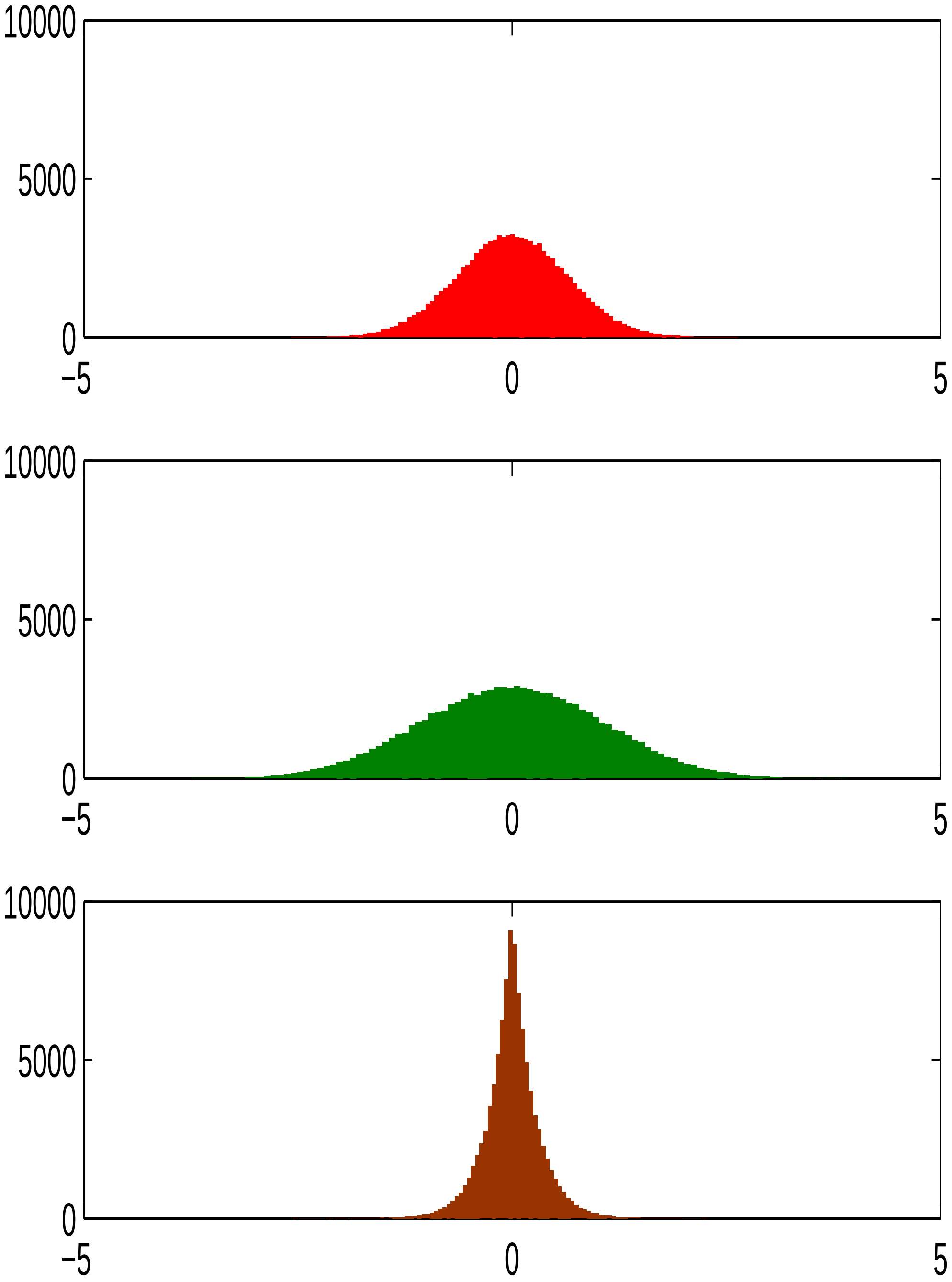
$$f(x) = \begin{cases} \frac{1}{2c} & if \quad -c < x < c \\ 0 & \qquad else \end{cases}$$

We could estimate the population mean of 0 with the following estimators. (There are, of course, many others).

- The sample mean, $\bar{X}$.
- The sample median, $\tilde{X}$.
- The average of the lowest and highest value, $\bar{X}_e$.

We may use $MATLAB$ to explore this situation to see which estimator performs the best (and to validate the fact that all three are unbiased).

```
sampleSize = 20;

popMax =  5;

popMin = -5;

numberSamples  =  100000;

sampleMeans    =  zeros(numberSamples,1);

sampleMedians  =  zeros(numberSamples,1);

sampleExAve    =  zeros(numberSamples,1);

for  i = 1:numberSamples

    sampleData =popMin + (popMax-popMin)*rand(sampleSize,1);

    sampleMeans(i)   = mean(sampleData);

    sampleMedians(i) = median(sampleData);

    sampleExAve(i)   = (max(sampleData)+min(sampleData))/2;

end


subplot(3,1,1); hist(sampleMeans,100); axis([-5 5 0 10000]);
subplot(3,1,2); hist(sampleMedians,100); axis([-5 5 0 10000]);

subplot(3,1,3); hist(sampleExAve,100); axis([-5 5 0 10000]);
```



We define the ***Standard Error***, $\sigma_{\hat{\theta}}$ of an estimator as the **standard deviation of its sampling distribution**.

We often want to estimate a mean with more than a single number in order to provide a measure of the quality of the estimate. This leads us to define a confidence interval.