Sampling Distribution for STAT112

Charles Fleming

May 14, 2018

1 Example of a Sampling Distribution

Define $\bar{x} = \frac{\sum_{i \in \mathcal{S}} x_i}{n}$ and $s^2 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2}{n-1}$. For each sample, $\bar{x}$ maps outcomes of the sample space to a number and s^2 maps outcomes of the sample space to a number according to a drawing of the sample. Both $\bar{x}$ and s^2 are, therefore, random variables. The schematic diagram shown in Figure 1 illustrates the mapping of $\bar{x}$ and s^2 from the set of outcomes which comprise the sample space Ω but to different numbers.

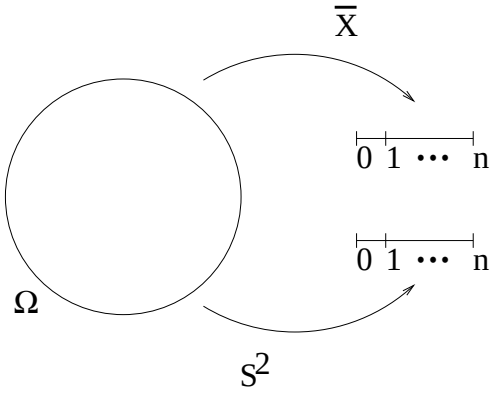

Figure 1:

Not only are $\bar{x}$ and s^2 random variables, but any mapping of a sample space to a number is a random variable like the sample median or the sample 1^{st} quartile or the sample range. Associated with a random variable is a probability distribution. There is one for $\bar{x}$ and a different one for s^2. The probability distribution which is associated with a sampling random variable is called a sampling distribution simply, in order to emphasize its association with a sample.

To illustrate the concept of a sampling distribution, consider the sample space of outcomes in which an outcome consists of a pair of numbers. Any place in the pair can be filled with either a 0, 2, 4, 6. As such, the sample space of all possible outcomes is shown in Figure 2.

$$\Omega = \begin{Bmatrix} (0,0) & (2,0) & (4,0) & (6,0) \\ (0,2) & (2,2) & (4,2) & (6,2) \\ (0,4) & (2,4) & (4,4) & (6,4) \\ (0,6) & (2,6) & (4,6) & (6,6) \end{Bmatrix}$$

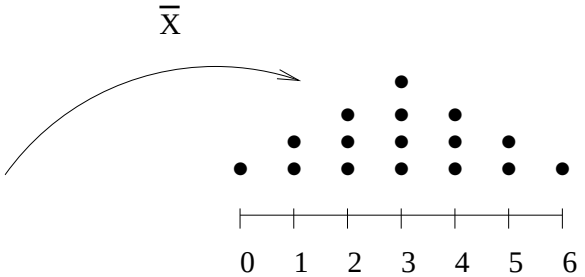

Figure 2:

Define $\bar{x} = \frac{a+b}{2}$ where a and b are the places in any pair (a,b). $\bar{x}$ maps an outcome to the average of its two members. Define $s^2 = \frac{(a-\bar{x})^2+(b-\bar{x})^2}{1} = \frac{(b-a)^2}{2}$, and the median as the median of a and b. Each random variable has a set of possible values. For $\bar{x}$, the possible values are: $\{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6\}$; for s^2, the possible values are: $\{0,\ 2,\ 8,\ 18\}$; for the median, the possible values are: $\{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6\}$. Co-incidentally, the distribution of $\bar{x}$ and of the median are the same in this example but not necessarily the same in general. Associated with each of these three random variables is a probability distribution; they are shown in Figure 3. None of the distributions is a common distribution which we know by a name, nevertheless, the diagram tells us everything we need to know about the distributions of $\bar{x}$, s^2, and the median. From the diagram, for instance, it is can be seen that $P(\bar{x} = 4) = \frac{4}{16}$. Similarly, $P(median = 3) = \frac{4}{16}$ and $P(s^2 = 8) = \frac{4}{16}$. It is clear that the sample mean, sample variance, and the sample median are random variables though they each have a different probability distribution.

Consider the random variable, $\bar{x}$. It has an expected value and a variance, that is:

$$E[\bar{x}] = 0(\frac{1}{16}) + 1(\frac{2}{16}) + \cdots + 5(\frac{2}{16}) + 6(\frac{1}{16}) = 3$$

and

$$var(\bar{x}) = (0-3)^2\frac{1}{16} + \cdots + (6-3)^2(\frac{1}{16}) = \frac{5}{2}$$

We know by Theorem 1 [1] which is discussed in the *Probability* lecture notes that $E[\bar{x}] = E[x_i] = \mu$. In this problem, $x_i = \{0,\ 2,\ 4,\ 6\}$ and that the probabilities of selecting the

[1]

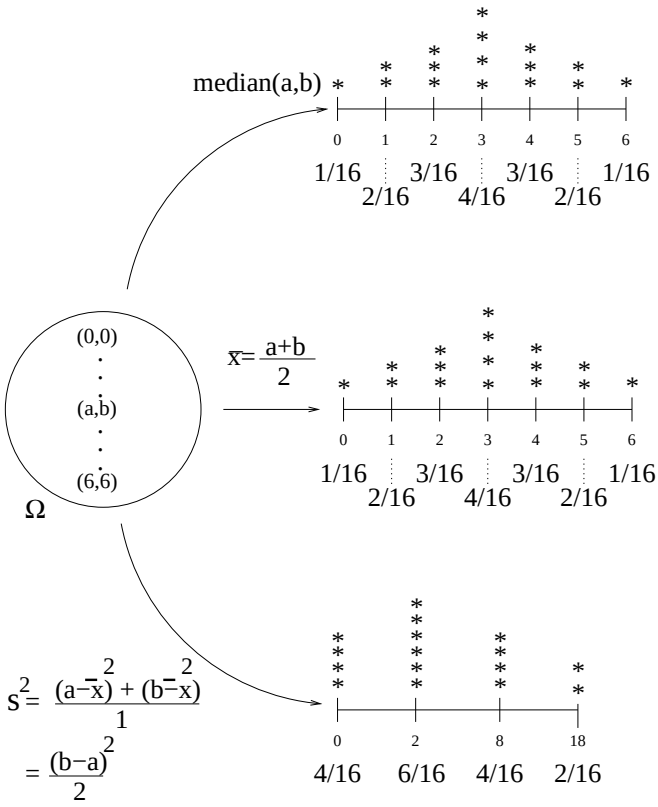

Figure 3:

elements of the sample space are equally likely; therefore, $E[x_i] = (0+2+4+6)\frac{1}{4} = 3 = \mu$. The short way of determining $E[\bar{x}]$ is to invoke Theorem 1 which tells us that $E[\bar{x}] = 3$. Likewise, by Theorem 1, since the x_i's are i.i.d., $var(\bar{x}) = \frac{\sigma^2}{n}$ where $\sigma^2 = var(x_i)$ and where n is the number of random variables which constitute $\bar{x}$. By direct computation,

$$var(x_i) = \frac{1}{4}(0-3)^2 + \frac{1}{4}(2-3)^2 + \frac{1}{4}(4-3)^2 + \frac{1}{4}(6-3)^2 = 5$$

Theorem 1. *If X_1, X_2, …, X_n are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$, then*

$$E[\bar{x}] = \mu \text{ and } var(\bar{x}) = \frac{\sigma^2}{n}$$

Note: for a finite population $\widehat{var(\bar{x})} = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$. See Theory of Survey Sampling for STAT112

$var(\bar{x}) = \frac{5}{2}$. We note that Theorem 1 does not apply to the median nor to s^2. For those cases, we need to resort to the definitions of median and variance as tedious as that might be.

The results of our computations are compiled in Table 1.

$\bar{x}$	median	s^2
$E[\bar{x}] = 3$	$E[median] = 3$	$E[s^2] = \frac{15}{4}$
$median(\bar{x}) = 3$	$median(median) = 3$	$median(s^2) = 2$
$var(\bar{x}) = \frac{5}{2}$	$var(median) = \frac{5}{2}$	$var(s^2) = \frac{553}{16}$

2 Empirical Sampling Distribution

In this section, a probability distribution will be compared to an empirical probability distribution which is obtained from a process of drawing pairs of numbers from: {0, 2, 4, 6} such that each drawing is equally likely. According to the theory of simulations, the resulting histogram should bear a resemblance to the theoretical probability distribution.

Simulations are performed when it is impossible or highly impractical to derive a theoretical probability distribution, because the mathematical problem is too complex to solve. The method of obtaining an empirical probability distribution by a process of simulations is called a Monte Carlo simulation technique. It was invented in the 1930's by physicists to evaluate complex molecular interactions and atomic physics which defied the derivation of exact mathematical solutions. Statisticians have used these Monte Carlo techniques to evaluate complex statistical problems.

As was mentioned in the *Probability* lecture notes, probability begins with a sampling space; it is abstract in that it only exits in man's imagination. Whereas statistics begins with a population of real objects which can be touched and examined. As such, even though there is no direct connection between the world of probability and the world of statistics, we can use the tools which are developed in the science of probability to make inferences about the population from a set of data which is obtained from the study of a real phenomenon. The validity of those inferences are expressed in part by of confidence intervals and equivalently by testing hypotheses.

The present example of creating an empirical probability distribution by means of a Monte Carlo simulation will illustrate that as the number of drawings increase the histogram will appear to converge to the probability density function. How close the histogram is to the probability density function will be measured by the goodness-of-fit test statistic. If the discrepancy between the histogram and the probability distribution is too large, then we reject the hypothesis that the histogram adequately fits the probability density function, otherwise if the discrepancy is small, a good appears to exit between the two.

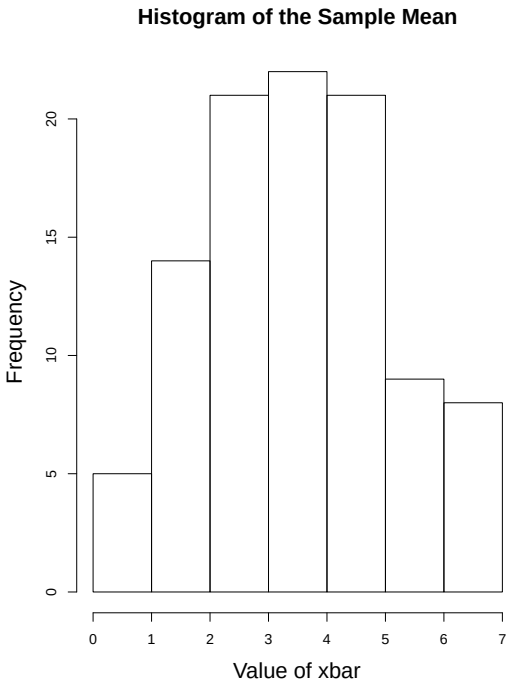

Figure 4:

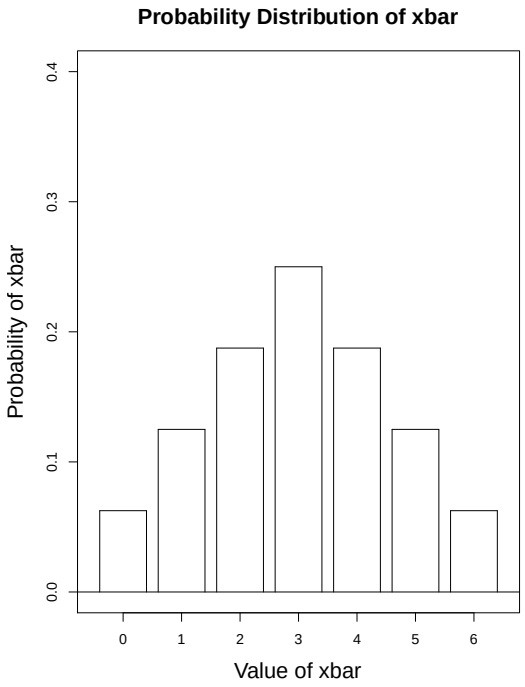

Figure 5:

The following simple computer program written in R will produce a histogram, X^2 test statistic, the X^2 quantile at $\alpha = .05$, and the p-value for the goodness-of-fit test in the case of creating an empirical sampling distribution like the actual one shown in Figure 3.

```
x<-c(0,2,4,6)
dist<-c()
for(i in 1:100){
y<-sample(x,size=2,replace=TRUE)    ## size=2 -> drawing a pair.
dist<-rbind(dist,c(y,mean(y),median(y),var(y)))
}
breaks<-0:7
r<-hist(dist[,3]+.0001,breaks=breaks, freq=T,
        main="Histogram of the Sample Mean", xlab="Value of xbar")
counts<-r$counts
prob<-c(1/16,2/16,3/16,4/16,3/16,2/16,1/16)
obs<-counts
exp<-sum(counts)*prob
chi2<-sum((obs-exp)^2/exp)
```

```
n<-length(prob)
print(c("Chi-square Test Statistic=",chi2))
print(c("Chi-square Quantile=",qchisq(.95,n-1)))
print(c("p-value=",1-pchisq(chi2,n-1)))
```

A different histogram and set of statistics is produced every time this program is executed. One such histogram appears in Figure 4 and a picture of the probability density function is shown in Figures 4 and 5.

By visual inspection, the histogram bears a resemblance to the probability density function. To provide an analytical way to measure the goodness-of-fit between the two, we will calculate the X^2 test statistic as follows:

$$X^2 = \sum_{i=1}^{n} \frac{(observed_i - expected_i)^2}{expected_i} \tag{1}$$

where the observed values of $\bar{x}$ came from the computer simulation of the sampling distribution. The results of one execution of the program is shown in Table 1 . The observed frequencies for every value of $\bar{x}$ which were obtained from this particular execution of the program are: $\{5,\ 14,\ 21,\ 22,\ 21,\ 9,\ 8\}$. The expected number of such occurrences is $100 \times prob$ where $prob = \{\frac{1}{16},\ \frac{2}{16},\ \frac{3}{16},\ \frac{4}{16},\ \frac{3}{16},\ \frac{2}{16},\ \frac{1}{16}\}$, that is $expected = \{6.25,\ 12.50,\ 18.75,\ 25.00,\ 18.75,\ 12.50,\ 6.25\}$. A tabulation of these calculations is given in Table 1.

Table 1

Value of $\bar{x}$	0	1	2	3	4	5	6
Observed	5	14	21	22	21	9	8
Expected	$100(\frac{1}{16})$	$100(\frac{2}{16})$	$100(\frac{3}{16})$	$100(\frac{4}{16})$	$100(\frac{3}{36})$	$100(\frac{2}{16})$	$100(\frac{1}{16})$
Deviation	$-20(\frac{1}{16})$	$24(\frac{1}{16})$	$36(\frac{1}{16})$	$-48(\frac{1}{16})$	$36(\frac{1}{16})$	$-56(\frac{1}{16})$	$28(\frac{1}{16})$
Squared Deviation	$\frac{25}{16}$	$\frac{36}{16}$	$\frac{81}{16}$	$\frac{144}{16}$	$\frac{81}{16}$	$\frac{196}{16}$	$\frac{49}{16}$
X^2 Terms	$\frac{1}{4}$	$\frac{9}{50}$	$\frac{27}{100}$	$\frac{9}{25}$	$\frac{27}{100}$	$\frac{49}{50}$	$\frac{49}{100}$

Therefore, after using equation (1) whose terms appear in the fifth row of Table 1, $X^2 = \frac{14}{5} = 2.8$; the $X^2_{5,.05}$ quantile is: 12.59159. Because $X^2 = 2.8 \ngtr 12.59159 = X^2_{5,.05}$, we cannot reject the null hypothesis that the probability density function adequately explains the histogram. The p-value happens to be .8335. It appears that the simulated distribution

makes an adequate fit and that the simulation whereby 100 sample means were generated by drawing pairs of elements at random agrees with the theoretical probability distribution.

Similarly, the same exercise may be performed for finding an empirical distributions of the median and for the variance. The simulation of the distribution of the median is shown in Figure 6 and its theoretical probability distribution is shown beside it in Figure 7. The goodness-of-fit test statistic is $X^2 = 2.8$ with a p-value of .8335. The simulation of the distribution of the variance is shown in Figure 8 and its theoretical probability distribution is shown in Figure 9. The goodness-of-fit test statistic is $X^2 = 2.306$ with a p-value of .5112. Note that n=4 because there are only four terms in computing X^2.

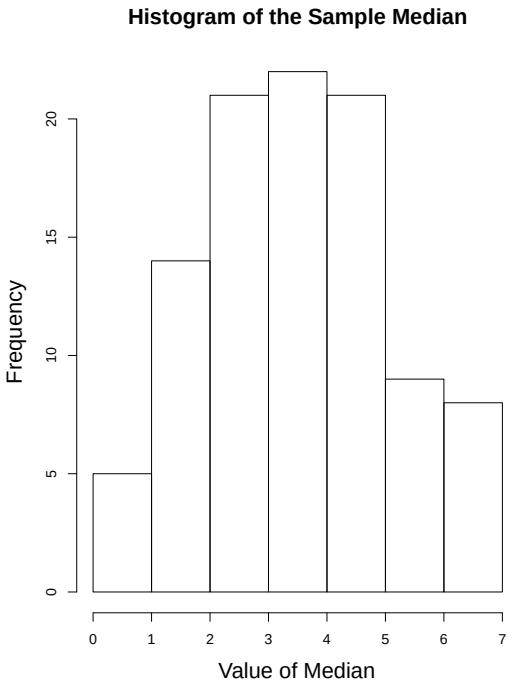

Figure 6:

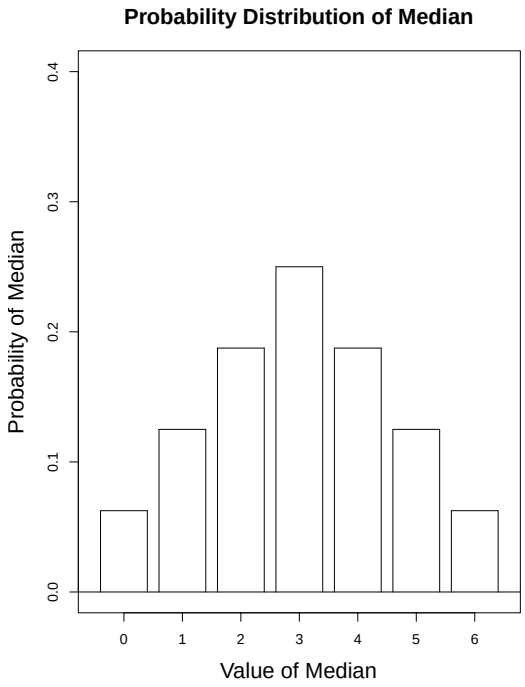

Figure 7:

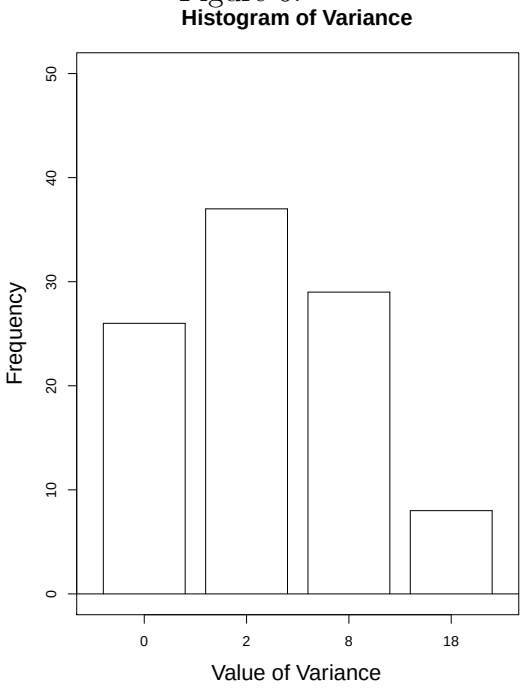

Figure 8:

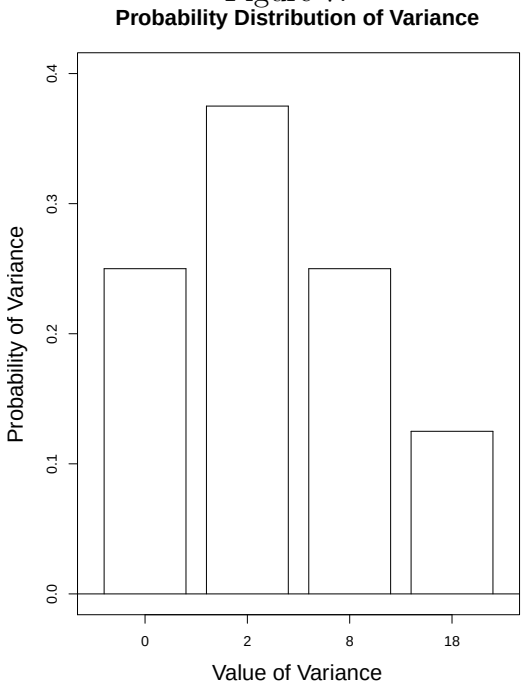

Figure 9:

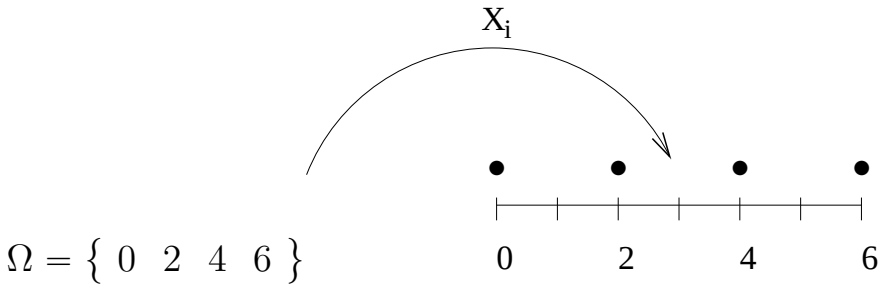

Figure 10:

3 Modeling a Survey Sample

When creating the example which was described in Section 1, a random variable was defined on a sample space by mapping the name of an outcome to its corresponding numerical value as depicted in Figure 10. The random variable, X_i follows a discrete Uniform distribution with a probability of $P(X_i = k) = \frac{1}{4} \forall k$. This sample space and random variable could describe the four possible scores of a quiz which student, i, can receive. The Uniform distribution suggests that the student merely guessed at the four possible answers as if the quiz consisted of three multiple choice problems where each problem is worth two points.

Suppose that there are two students in the class who both guess at the questions. According to the discussion of Section 1, the sample space of quiz scores for the class is shown in Figure 2 and the probability distribution of the class mean is shown in Figure 3 where $\bar{x} = \frac{x_1+x_2}{2}$.

Rather than guessing, suppose the two students answered the questions after having carefully studied the subject of the quiz and having had attended the lectures. Assuming that the two students are of equal caliber such that their probability distributions of getting the correct answers are the same. In other words, the x_i's are i.i.d. Let us assume that the probability distribution is the one shown in Table 2 and shown in Figure 11.

Table 2

0	2	4	6
$\frac{1}{10}$	$\frac{1}{6}$	$\frac{2}{5}$	$\frac{1}{3}$

Whereas, in the case of x_i following a Uniform distribution as shown in Figure 10, the

probability distribution of $\frac{x_1+x_2}{2}$ which is shown in Figures 2 and 5 was easy to deduce, but it is a challenge to deduce the probability distribution of $\frac{x_1+x_2}{2}$ when x_i's follow another probability distribution. The mathematical technique which is used to find the probability distribution of $\frac{x_1+x_2}{2}$ is called the method of convolutions. To find the probability distribution of $\frac{x_1+x_2}{2}$, we note that because $P(\bar{x}=k) = P(\frac{x_1+x_2}{2}=k) = P(x_1+x_2=2k)$, it is sufficient to find the convolution of x_1+x_2. The process of finding the convolution begins by defining the probability generating function, $p(s) = p_0s^0 + p_1s^1 + p_2s^2 + \ldots + p_ns^n$. For x_i, the probability generating function is, $p(s) = p_0+p_1s^1+p_2s^2+p_3s^3 = \frac{1}{10}+\frac{1}{6}s+\frac{2}{5}s^2+\frac{1}{3}s^3$. According to the theory of convolutions, the probability generation function for x_1+x_2 is $p(s)^2$, because x_i's are i.i.d. Continuing in this way, the probability generation function for $x_1+x_2+x_3$ will be $p(s)^3$. Or in general, the probability generation function for $x_1+x_2+x_3+\ldots+x_n$ is $p(s)^n$ when the x_i's are i.i.d.

The coefficients of the probability generating function correspond to the probabilities. For example,

$$\begin{aligned} p(s)^2 &= (\frac{1}{10}+\frac{1}{6}s+\frac{2}{5}s^2+\frac{1}{3}s^3)^2 \\ &= \frac{1}{100}+\frac{1}{10}s^1+\frac{97}{900}s^2+\frac{1}{5}s^3+\frac{61}{225}s^4+\frac{4}{15}s^5+\frac{1}{9}s^6 \end{aligned}$$

Therefore, the probability distribution of x_1+x_2 is:

Table 3

Value of $\bar{x}$	0	1	2	3	4	5	6
Probability	$\frac{1}{100}$	$\frac{1}{10}$	$\frac{97}{900}$	$\frac{1}{5}$	$\frac{61}{225}$	$\frac{4}{15}$	$\frac{1}{9}$

and it is shown in Figure 12.

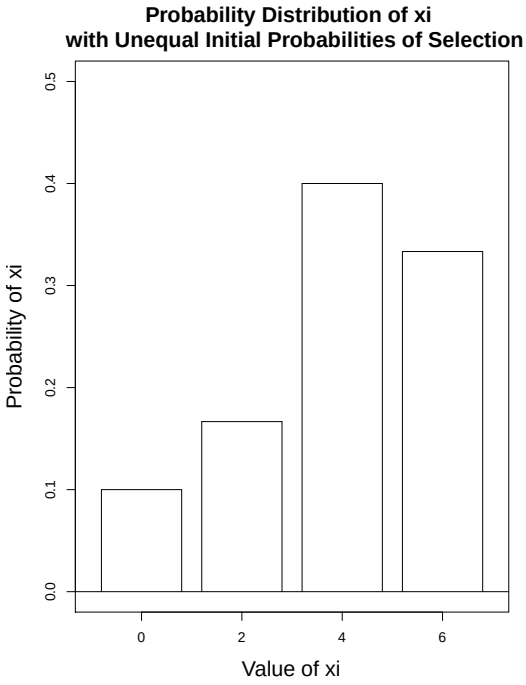

Figure 11:

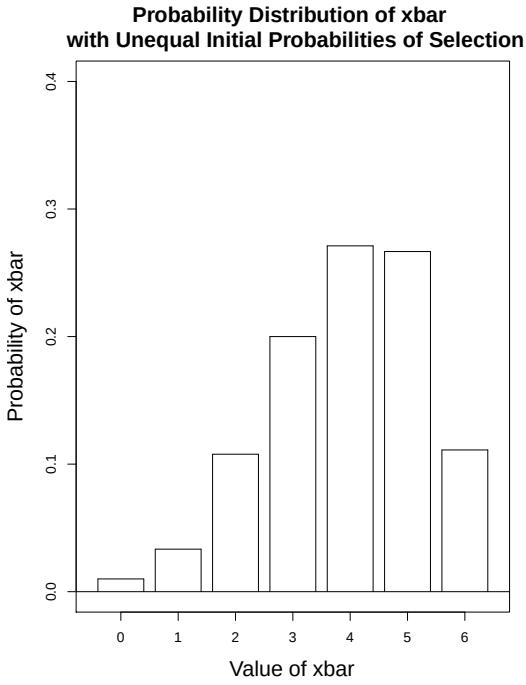

Figure 12:

For a class of two students, the number of elements in the sample space is $4^2 = 16$. For three students, the size of the sampling space will be 4^3; for a class of 55 students, the size of Ω will be 4^{55}. The corresponding probability distribution of $\bar{x} = \frac{x_1+\ldots+x_{55}}{55}$ will have $4 \times 55 = 210$ terms and its corresponding probability generating function will be $p(x)^{55}$. With so many terms, the practicality of finding the expected value of $\bar{x}$ and of s^2 from the definitions is formidable. Fortunately, Theorem 1 provides easy answers. By means of Theorem 1, $E[\bar{x}] = E[x_i]$ and $var(\bar{x}) = \frac{\sigma^2}{n}$. By referring to Table 3, $E[x_i] = 0(\frac{1}{100}) + \ldots + 6(\frac{1}{9}) = \frac{59}{15}$ and $\sigma^2 = var(x_i) = (0 - \frac{59}{15})^2\frac{1}{100} + \ldots (6 - \frac{59}{15})^2\frac{1}{9} = \frac{809}{450}$, so that $var(\bar{x}) = \frac{\frac{809}{450}}{55} = \frac{809}{24750} = .03268$.

Unfortunately, there is no equivalent Theorem 1 for finding the class median nor the class variance, s^2. Instead, we resort to a Monte Carlo technique like the one which we used in Section 2. A computer program for finding the mean of $\bar{x}$ using a simulation technique of drawing a sample of 55 elements 100 times is given below:

```
x<-c(0,2,4,6)
dist<-c()
p0<-c(1/10,1/6,2/5,1/3)
for(i in 1:100){
```

```
y<-sample(x,size=55,replace=TRUE,prob=p0)
dist<-rbind(dist,c(y,mean(y),median(y),var(y)))
}
breaks<-0:7
r<-hist(dist[,3]+.0001,breaks=breaks, freq=T, main="Histogram of the Sample Mean",
counts<-r$counts
p<-counts/(sum(counts))
```

The empirical probability density function of $\bar{x}$ is shown in Figure 13. Superimposed on the histogram is a plot of a Normal distribution. We see that the Normal distribution fits the histogram quite well in accordance with the Central Limit Theorem.

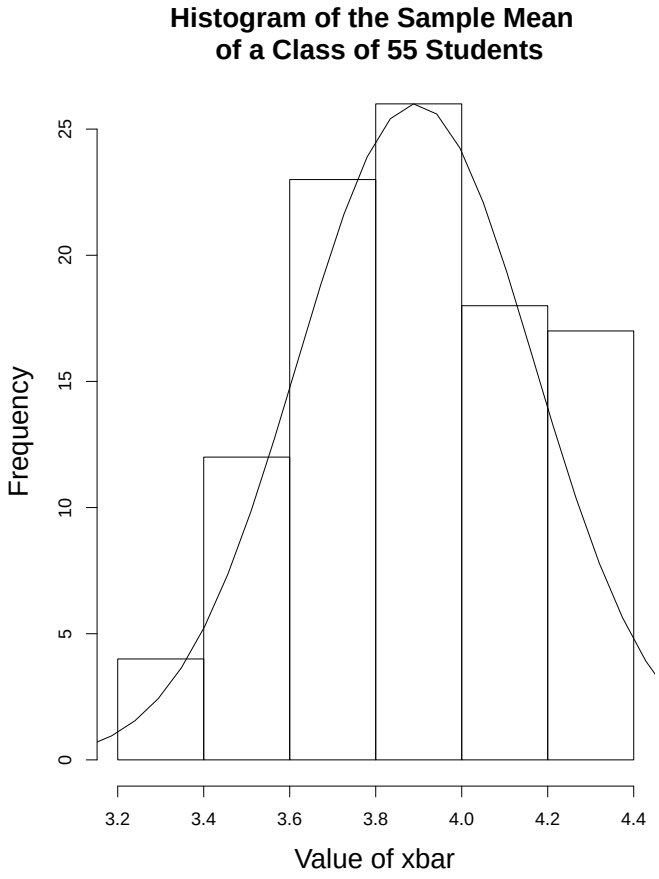

Figure 13:

By mean of the same Monte Carlo technique, the probability density functions of the median and of the sample variance for a class of 55 students are shown in Figure 14 and

15. Superimposed on the graph of the graph of the sample variance, there appears a plot of a X^2 distribution.

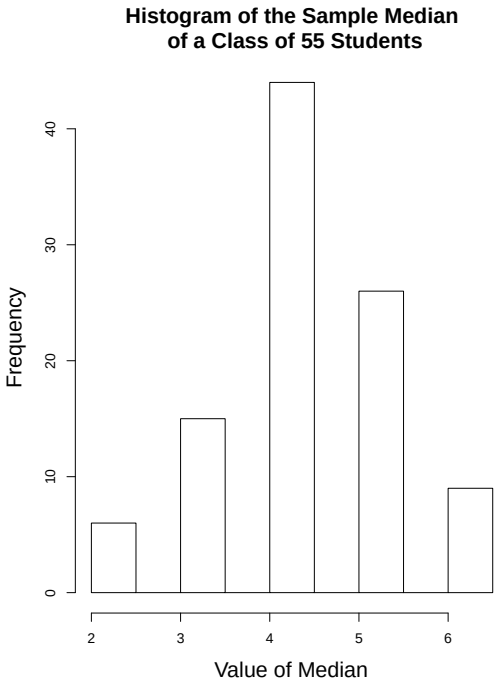

Figure 14:

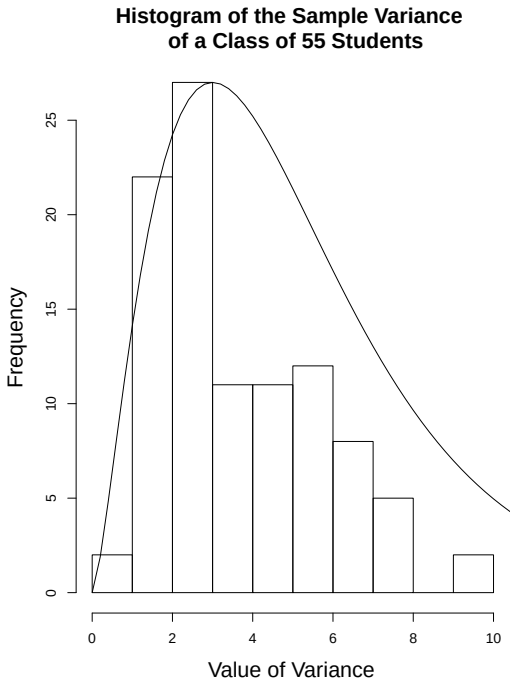

Figure 15:

The discussion on finding the class mean and its variance presupposes an infinite population of students from which the class of 55 students was drawn. Suppose that there are four sections of students at a certain college with a total size of 203 students, then the population of students is finite. Consequently, the estimate of the variance of the sample mean must be modified by the finite population correction factor, $\frac{N-n}{N}$, so that $\widehat{var(\bar{x})} = (\frac{203-55}{203})var(\bar{x}) = (\frac{148}{203})\frac{809}{14025} = .04205$. For a discussion of the finite population correction factor see the lecture notes, *Theory of Survey Sampling for STAT112*.