13.3: Simple Random Samples and Statistics

Simple Random Samples and Statistics

We formulate the notion of a (simple) *random sample*, which is basic to much of classical statistics. Once formulated, we may apply probability theory to exhibit several basic ideas of statistical analysis.

We begin with the notion of a *population distribution*. A population may be most any collection of individuals or entities. Associated with each member is a quantity or a feature that can be assigned a number. The quantity varies throughout the population. The population distribution is the distribution of that quantity among the members of the population.

If each member could be observed, the population distribution could be determined completely. However, that is not always feasible. In order to obtain information about the population distribution, we select "at random" a subset of the population and observe how the quantity varies over the sample. Hopefully, the sample distribution will give a useful approximation to the population distribution.

The sampling process

We take a sample of size n, which means we select *n* members of the population and observe the quantity associated with each. The selection is done in such a manner that on any trial each member is equally likely to be selected. Also, the sampling is done in such a way that the result of any one selection does not affect, and is not affected by, the others. It appears that we are describing a composite trial. We model the *sampling process* as follows:

> Let X_i, $1 \le i \le n$ be the random variable for the *i*th component trial. Then the class $\{X_i : 1 \le i \le n\}$ is iid, with each member having the population distribution.

This provides a model for sampling either from a very large population (often referred to as an infinite population) or sampling with replacement from a small population.

The goal is to determine as much as possible about the character of the population. Two important parameters are the mean and the variance. We want the population mean and the population variance. If the sample is representative of the population, then the sample mean and the sample variance should approximate the population quantities.

- The *sampling process* is the iid class $\{X_i : 1 \le i \le n\}$.
- A *random sample* is an observation, or realization, $(t_1, t_2, \cdots, t_n)$ of the sampling process.

The sample average and the population mean

Consider the numerical average of the values in the sample $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} t_i$. This is an observation of the *sample average*

$$A_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} S_n$$

The sample sum S_n and the sample average A_n are random variables. If another observation were made (another sample taken), the observed value of these quantities would probably be different. Now S_n and A_n are functions of the random variables $\{X_i : 1 \le i \le n\}$ in the sampling process. As such, they have distributions related to the *population distribution* (the common distribution of the X_i). According to the central limit theorem, for any reasonable sized sample they should be approximately normally distributed. As the examples demonstrating the central limit theorem show, the sample size need not be large in many cases. Now if the *population* mean $E[X]$ is μ and the *population variance* $\text{Var}[X]$ is σ^2, then

$$E[S_n] = \sum_{i=1}^{n} E[X_i] = nE[X] = n\mu \text{ and } \text{Var}[S_n] = \sum_{i=1}^{n} \text{Var}[X_i] = n\text{Var}[X] = n\sigma^2$$

so that

$$E[A_n] = \frac{1}{n}E[S_n] = \mu \text{ and } \text{Var}[A_n] = \frac{1}{n^2}\text{Var}[S_n] = \sigma^2/n$$

Herein lies the key to the usefulness of a large sample. The mean of the sample average A_n is the same as the population mean, but the variance of the sample average is $1/n$ times the population variance. Thus, *for large enough sample, the probability is high that the observed value of the sample average will be close to the population mean*. The population standard deviation, as a measure of the variation is reduced by a factor $1/\sqrt{n}$.

Example 13.3.1 Sample size

Suppose a population has mean μ and variance σ^2. A sample of size n is to be taken. There are complementary questions:

1. If n is given, what is the probability the sample average lies within distance a from the population mean?
2. What value of n is required to ensure a probability of at least p that the sample average lies within distance a from the population mean?

Solution

Suppose the sample variance is known or can be approximated reasonably. If the sample size n is reasonably large, depending on the population distribution (as seen in the previous demonstrations), then A_n is approximately $N(\mu, \sigma^2/n)$.

1. Sample size given, probability to be determined.

$$p = P(|A_n - \mu| \le a) = P(|\frac{A_n - \mu}{\sigma/\sqrt{n}}| \le \frac{a\sqrt{n}}{\sigma} = 2\phi(a\sqrt{n}/\sigma) - 1$$

2. Sample size to be determined, probability specified.

$$2\phi(a\sqrt{n}/\sigma) - 1 \ge p \text{ iff } \phi(a\sqrt{n}/\sigma) \ge \frac{p+1}{2}$$

Find from a table or by use of the inverse normal function the value of $x = a\sqrt{n}/\sigma$ required to make $\phi(x)$ at least $(p+1)/2$. Then

$$n \ge \sigma^2(x/a)^2 = (\frac{\sigma}{a})^2 x^2$$

We may use the MATLAB function *norminv* to calculate values of x for various p.

```
p = [0.8 0.9 0.95 0.98 0.99];
x = norminv(0,1,(1+p)/2);
disp([p;x;x.^2]')
    0.8000    1.2816    1.6424
    0.9000    1.6449    2.7055
    0.9500    1.9600    3.8415
    0.9800    2.3263    5.4119
    0.9900    2.5758    6.6349
```

For $p = 0.95$, $\sigma = 2$, $a = 0.2$, $n \ge (2/0.2)^2 3.8415 = 384.15$ Use at least 385 or perhaps 400 because of uncertainty about the actual σ^2

The idea of a statistic

As a function of the random variables in the sampling process, the sample average is an example of a statistic.

Definition: statistic

A *statistic* is a function of the class $\{X_i : 1 \le i \le n\}$ which uses explicitly no unknown parameters of the population.

Example 13.3.2 Statistics as functions of the sampling progress

The random variable

$$W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \text{ , where } \mu = E[X]$$

is *not* a statistic, since it uses the unknown parameter μ. However, the following *is* a statistic.

$$V_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - A_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - A_n^2$$

It would appear that V_n^* might be a reasonable estimate of the population variance. However, the following result shows that a slight modification is desirable.

Example 13.3.3 An estimator for the population variance

The statistic

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - A_n)^2$$

is an estimator for the population variance.

VERIFICATION

Consider the statistic

$$V_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - A_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - A_n^2$$

Noting that $E[X^2] = \sigma^2 + \mu^2$, we use the last expression to show

$$E[V_n^*] = \frac{1}{n}n(\sigma^2+\mu^2) - (\frac{\sigma^2}{n} + \mu^2) = \frac{n-1}{n}\sigma^2$$

The quantity has a *bias* in the average. If we consider

$$V_n = \frac{n}{n-1}V_n^* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - A_n)^2 \text{ , then } E[V_n] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

The quantity V_n with $1/(n-1)$ rather than $1/n$ is often called the *sample variance* to distinguish it from the population variance. If the set of numbers

$$(t_1, t_2, \cdots, t_N)$$

represent the complete set of values in a population of N members, the variance for the population would be given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} t_i^2 - (\frac{1}{N}\sum_{i=1}^{N} t_i)^2$$

Here we use $1/N$ rather than $1/(N-1)$.

Since the statistic V_n has mean value σ^2, it seems a reasonable candidate for an estimator of the population variance. If we ask how good is it, we need to consider its variance. As a random variable, it has a variance. An evaluation similar to that for the mean, but more complicated in detail, shows that

$$\text{Var}[V_n] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4) \text{ where } \mu_4 = E[(X-\mu)^4]$$

For large n, $\text{Var}[V_n]$ is small, so that V_n is a good large-sample estimator for σ^2.

Example 13.3.4 A sampling demonstration of the CLT

Consider a population random variable X ~ uniform [-1, 1]. Then $E[X] = 0$ and $\text{Var}[X] = 1/3$. We take 100 samples of size 100, and determine the sample sums. This gives a sample of size 100 of the sample sum random variable S_{100}, which has mean zero and variance 100/3. For each observed value of the sample sum random variable, we plot the fraction of observed sums less than or equal to that value. This yields an experimental distribution function for S_{100}, which is compared with the distribution function for a random variable $Y \sim N(0, 100/3)$.

```
rand('seed',0)    % Seeds random number generator for later comparison
tappr                                        % Approximation setup
Enter matrix [a b] of x-range endpoints  [-1 1]
Enter number of x approximation points  100
Enter density as a function of t  0.5*(t<=1)
Use row matrices X and PX as in the simple case
```

```
qsample                                  % Creates sample
Enter row matrix of VALUES  X
Enter row matrix of PROBABILITIES  PX
Sample size n =  10000                   % Master sample size 10,000
Sample average ex = 0.003746
Approximate population mean E(X) = 1.561e-17
Sample variance vx = 0.3344
Approximate population variance V(X) = 0.3333
m = 100;
a = reshape(T,m,m);                      % Forms 100 samples of size 100
A = sum(a);                              % Matrix A of sample sums
[t,f] = csort(A,ones(1,m));              % Sorts A and determines cumulative
p = cumsum(f)/m;                         % fraction of elements <= each value
pg = gaussian(0,100/3,t);                % Gaussian dbn for sample sum values
plot(t,p,'k-',t,pg,'k-.')                % Comparative plot
% Plotting details                       (see Figure 13.3.1)
```

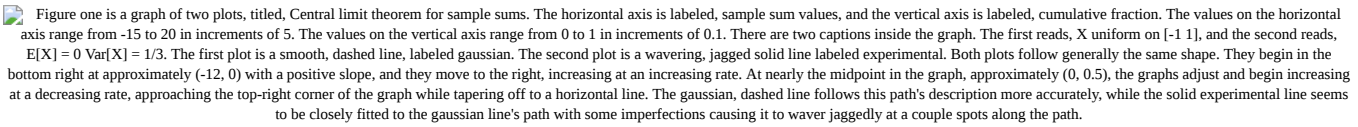
Figure one is a graph of two plots, titled, Central limit theorem for sample sums. The horizontal axis is labeled, sample sum values, and the vertical axis is labeled, cumulative fraction. The values on the horizontal axis range from -15 to 20 in increments of 5. The values on the vertical axis range from 0 to 1 in increments of 0.1. There are two captions inside the graph. The first reads, X uniform on [-1 1], and the second reads, E[X] = 0 Var[X] = 1/3. The first plot is a smooth, dashed line, labeled gaussian. The second plot is a wavering, jagged solid line labeled experimental. Both plots follow generally the same shape. They begin in the bottom right at approximately (-12, 0) with a positive slope, and they move to the right, increasing at an increasing rate. At nearly the midpoint in the graph, approximately (0, 0.5), the graphs adjust and begin increasing at a decreasing rate, approaching the top-right corner of the graph while tapering off to a horizontal line. The gaussian, dashed line follows this path's description more accurately, while the solid experimental line seems to be closely fitted to the gaussian line's path with some imperfections causing it to waver jaggedly at a couple spots along the path.

Figure 13.3.1. The central limit theorem for sample sums.