Statistics from Simulations

A.1 Statistical Inference and Estimation of Population Parameters

When the goal is to estimate the parameter of some population (e.g., mean or variance), the usual procedure is to design a probability sample to provide a *sample statistic*. As an example, we might estimate the parameter 'true average cost' of an inventory system (population mean, μ) with the statistic 'average cost for 12 months' (sample mean, $\bar{x}$); or we might estimate the proportion of taxpayers whose income tax returns were audited by the IRS in a particular year (population proportion, π) by using the proportion of a random sample that were audited (sample proportion, p). When a single number based on a sample is used to estimate the population parameter ($\bar{x}$ for μ or p for π), it is termed a point estimate. While such estimates are indispensable for calibrating a model, they provide no insight into the extent of random sampling error. *Interval estimates*, on the other hand, allow us to specify the maximum expected error and associated probability that a point estimate would diverge from the population parameter.

By realizing that the sample statistic is only one of many possibilities, it is not difficult to conceptualize the existence of a distribution of possible sample statistics, termed the *sampling distribution of a statistic*. Now for the crux of sampling theory. If the shape or form of the sampling distribution can be specified, then it would be possible to base errors on probability judgments. For example, if p is used to estimate π, then the sampling distribution of p turns out to be binomial (which can be approximated by the normal curve if the sample size n is sufficiently large), assuming the conditions underlying the binomial process are met. If $\bar{x}$ is used to estimate μ, then the sampling distribution of $\bar{x}$ turns out to be approximately normally or t-distributed. Given a specified shape for the sampling distribution, interval estimates are made easily, as illustrated below.

Sampling and Experimental Design

For any study that requires data, there is a target set of objects about which information regarding some attribute is desired. This set of objects is termed the *universe* or *population*. Illustrative attributes and populations include the following: the probability distribution for "interarrival times" (attribute) for "all possible arrivals at the emergency room of a hospital" (population); the average "response velocity for a particular type of emergency vehicle" (attribute) among "all the possible responses in a given section of the city" (population); the expected "total cost of producing any given number of items over some planning horizon" (attribute) from among the "set of items of the same type which will be, could be, or would have been produced" (population).

If every item of the population is to be examined, a *census* is to be undertaken; if some subset of the population is to be examined, then a *sample* will be undertaken. *Inferential statistics* is the area of study which is concerned with the making of generalizations or inferences about some population characteristic based on the results of a sample. The motivation for sampling is strong. In many cases, a census either is not feasible (e.g., an infinite population) or is impractical (e.g., determination of the attribute requires the destruction of the object). In most cases, sampling is considerably less costly with little or no sacrifice in estimation error.

The method of selecting a sample is a field of study in itself requiring substantial expertise. Essentially, samples can be selected in one of two ways: judgment samples and probability samples. In the first approach, an individual (or set of individuals) selects items which are “known” to be typical with respect to the desired attribute(s); in the latter approach, the selection of items is based on a plan that requires a knowledge of the probability that any given item will be selected. Only for the latter can we estimate the degree of sampling error -- a measure of the variability in estimating a statistic such as the mean. Of the various methods for selecting probability samples (e.g., simple random, systematic, cluster, stratified random, and variants or combinations thereof), we will exclusively deal with simple random sampling where each element of the population has an equal probability of being selected.

Sampling Distribution of the Mean

Let *X* be a random variable with mean μ_X. If $x_1, x_2, \ldots, x_n$ represent independent and identically distributed random variates (i.e, a random sample of size *n*) drawn from the distribution of *X*, then the sample mean (point estimate for the population mean, μ_X) can be calculated as follows.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{A.1}$$

If *n* is sufficiently large (say, above 30) and the population standard deviation is known, then we have the result guaranteed by the *central limit theorem* (CLT):

> If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a population with $E[X] = \mu_X$ and $Var[X] = \sigma_X^2$, then the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approaches a *normal distribution* with
>
> $$E[\bar{X}] = \mu_X \tag{A.2}$$
>
> and
>
> $$Var[\bar{X}] = \sigma_{\bar{X}}^2 = \sigma_X^2 / n \tag{A.3}$$
>
> as *n* approaches infinity.

In short, this allows us to establish a probability range of μ_X. Carefully note that the CLT is operative if and only if (i) the variates are independently and identically distributed, (ii) *n* is sufficiently large, and (iii) σ_X is known. Condition (i) and the derivation of (A.3) require that sampling is from an infinite population or from a finite population with replacement; otherwise, a so-called finite population correction factor must be applied to (A.3):

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right). \tag{A.4}$$

If n is small or σ_X is unknown, then the sampling distribution of $\bar{X}$ will have a t-distribution, providing the population for X is normally distributed. Note that the normality of $\bar{X}$ as specified by the CLT makes absolutely no assumptions about the distribution of X. The application of the t-distribution, however, requires that X itself be normally distributed. If σ_X^2 is unknown (the usual case), it is estimated by the *unbiased sample variance*:

$$\hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{A.5}$$

If n is large or small, σ_X is unknown, and X is *not* normally distributed, then statistical theory does not tell us how to compute reliable estimates of σ_X. For practical purposes, however, a large n (above 50) under these conditions will give satisfactory results.

Confidence Intervals

Once $\bar{x}$ and the *standard error of the mean,* $\sigma_{\bar{X}}$, are determined, the confidence interval and maximum error for μ_X are given by

$$\mu_X = \bar{x} \pm z\sigma_{\bar{X}} \tag{A.6}$$

or

$$\mu_X = \bar{x} \pm t\hat{\sigma}_{\bar{X}} \tag{A.7}$$

as the case may be, where $\hat{\sigma}_{\bar{X}}$ is the estimated standard error when (A.5) is used in place of σ_X^2 in (A.3) or (A.4). If σ_X is known, then the maximum error ε for a given level of confidence can be found from

$$\varepsilon = z\sigma_{\bar{X}} \quad \text{or} \quad \varepsilon = z\sigma_X / \sqrt{n}$$

when (A.3) applies. It follows that

$$n = \left(\frac{z\sigma_X}{\varepsilon}\right)^2 \tag{A.8}$$

provides the required sample size which satisfies a given maximum error and confidence level.

Example (Response Velocity)

In computing typical police response times, a velocity of 20 mph is used for a major thoroughfare in Austin, Texas. Suppose this value represents a point estimate (i.e., $\bar{x}$ = 20) for the mean of the population of all possible response velocities based on a simple random sample of 61 observations (n = 61). The population variance (σ_X^2) is unknown, but has been estimated as $\hat{\sigma}_X^2$ = 4 using (A.5). Seasonal factors based on shift, day of week, and month of year have been controlled to ensure as much as possible that random variates are independent and identically distributed. The assumption of a random sample was further buttressed by a previous study which showed that response velocities do not vary significantly with individual drivers. Finally, a chi-square test on the null hypothesis is that the random sample of 61 observed velocities came from a normal distribution was accepted. This cleared the way for using the *t*-distribution as representative of the $\bar{X}$ distribution.

Given that the population is infinite, we substitute $\hat{\sigma}_X^2$ for σ_X^2 in (A.3), which gives $\hat{\sigma}_X^2 = {}^{4}\!/_{61}$, or $\hat{\sigma}_{\bar{x}}$ = 0.26 mph. For degrees of freedom, $df = n - 1 = 60$ and a 0.95 level of confidence (that is, 0.025 of the area lies in each tail), the value of t is 2.0; hence, according to (A.7),

$$\begin{aligned}\mu_X &= 20 \pm (2.0)(0.26)\\ &= 20 \pm 0.52.\end{aligned}$$

This means that there is a 95% is probability that the estimate of 20 mph is *within* 0.52 mph of the population (true) average velocity. Alternatively, the 95-percent confidence interval for μ_X is 19.48 to 20.52 mph. These confidence limits can be used as extreme values of this parameter for performing sensitivity analysis.

Follow-up exercises:

1. Determine the confidence limits if n = 21, all other things being equal. Conclusion?
2. Estimate maximum error for μ_X if $\bar{x}$ = 50, n = 30, and σ_X= 10. Use 95% and 99% confidence intervals.
3. Determine the required sample size if σ_X= 1.41 and an error of no more than 0.5 mph with 99% confidence is desired. What sample size must be taken if ε = 0.1? What conclusion can you draw as to behavior of n as ε varies?

A.2 Chi-Square Goodness-of-Fit Test

In most studies involving random phenomena, one of the first steps is to characterize the randomness by trying to associate a known probability distribution to it. In this section, we illustrate the use of the chi-square distribution to test the hypothesis that an empirical distribution based on some random variable represents a sample from some theoretical distribution.

Given an empirical probability distribution with observed frequencies (o_i) and a theoretical probability distribution with expected frequencies (e_i), both with k categories as illustrated in Table A-1, the statistic

$$s = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{A.9}$$

approaches a chi-square (χ^2) distribution with $df = k - m$ degrees of freedom as the sample size n approaches infinity (m stands for the number of parameters or restrictions that apply in the calculations). In order to construct the empirical distribution, we divide the n observations into classes or intervals of roughly equal size. As a rule, $n \geq 50$ gives satisfactory results as long as all classes have more than five members; otherwise, additional refinements must be incorporated in the test. Note that small values for s imply "good" fits between the empirical and theoretical distributions; for example, if $o_i = e_i$ for all i, then the fit would be perfect (giving $s = 0$).

Table A-1. Setup for chi-square test

Range of random variable $a_i \leq X \leq b_i$ (class i)	Empirically observed frequency (o_i)	Theoretically expected frequency (e_i)
1	o_1	e_1
2	o_2	e_2
⋮	⋮	⋮
k	o_k	e_k

The idea then is to evaluate (A.9) utilizing the expected frequencies for some assumed theoretical distribution. The null hypothesis would state that the empirical distribution represents a sample of observations from the assumed theoretical distribution. If this were true, then differences between observed and expected frequencies would be due only to sampling error and values for s would be small; on the other hand, a large value for s would raise the suspicion that the empirical data are inconsistent with the assumed underlying distribution, which implies a rejection of the null hypothesis.

Example (Emergency Room Service)

Consider the quality of service in an emergency room. A simple random sample of 100 cases has been selected from among a large number of records, and the attribute “time to treat a patient” was recorded as shown in Table A-2. The data have been smoothed to ensure the elimination of identifiable “noise.” Furthermore, a previous statistical study showed no appreciable differences in the distribution of treatment times for various physicians.

Table A-2. Data and computations for exponential service distribution

Service time per patient (hours)		Observed frequency	For exponential pdf with $\lambda = 2.5$			Expected frequency	
			$P(X \le a_i)$	$P(X \le b_i)$	$P(a_i \le X \le b_i)$		
$(a_i \le X \le b_i)$		(o_i)	$F(a_i)$	$F(b_i)$	$F(b_i) - F(a_i)$	(e_i)	$\frac{(o_i - e_i)^2}{e_i}$
0.0	0.2	38	0.0000	0.3935	0.3935	39.35	0.05
0.2	0.4	25	0.3935	0.6321	0.2386	23.86	0.06
0.4	0.6	17	0.6321	0.7769	0.1448	14.48	0.44
0.6	0.8	9	0.7769	0.8647	0.0878	8.78	0.01
0.8	1.0	6	0.8647	0.9179	0.0532	5.32	0.09
1.0	1.2	5	0.9179	0.9507	0.0328	3.28	1.26
Over	1.2	0	0.9507	1.0000	0.0493	4.93	
		100					1.91

Using the definition of expected value, we calculate the sample mean as 0.4 hours. Suppose we formulate the null hypothesis that the observed distribution is a sample from an exponential distribution with $E[X] = 0.4$. Furthermore, suppose we specify a probability of no more than 0.01 for the wrong decision of rejecting a true hypothesis (this establishes a 0.01 level of significance ($\alpha = 0.01$) or right-tail area and a corresponding value for the critical value of χ^2 as in Fig. A-1).

For the exponential distribution, $E[X] = 1/\lambda$ so $\lambda = 1/0.40 = 2.5$. Thus the assumed CDF given by

$$F(x) = 1 - e^{-2.5x}.$$

We can now determine the probability of a theoretical observation within any specified category or class (a_i, b_i) in the table. For example, in the third class ($a_3 = 0.4$, $b_3 = 0.6$), we have

$$\begin{aligned}\Pr\{0.4 \le X \le 0.6\} &= (1 - e^{-2.5(0.6)}) - (1 - e^{-2.5(0.4)}) \\ &= 0.7769 - 0.6321 \\ &= 0.1448.\end{aligned}$$

Expected frequencies for each class are now determined as the product of the total number of observations (n = 100) and the probability of an observation in that class. For the third class, e_3 = 100(0.1448) = 14.48. Calculations for each class are shown in Table A-2. Note that the last two classes have been combined to avoid an expected frequency of less than five in any class. This changes the number of categories to k = 6.

The last column illustrates the calculation of s = 1.91 according to (A.9). For df = 6 − 2 = 4 (two parameters, n and λ, were needed so m = 2), the critical value of χ^2 that gives a 0.01 level of significance is 13.28 (See Fig. A-1). Note that critical values can be found in any statistical text. Since s < 13.28, we conclude that the exponential distribution with λ = 2.5 is a reasonably theoretical distribution to represent this process.

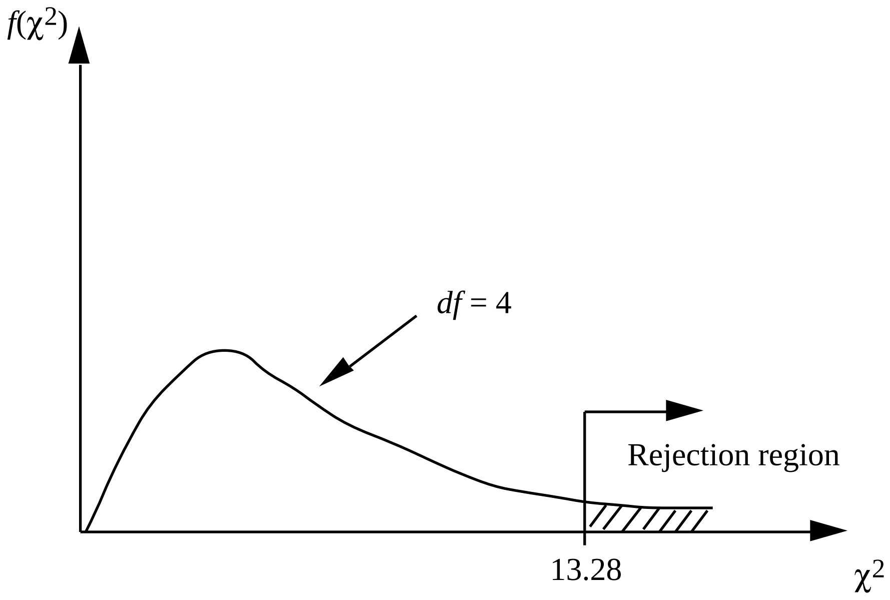

Figure A-1. Chi-square distribution for hypothesis test

Follow-up exercises:

1. Plot the empirical and theoretical CDF's for this example and compare.
2. Test the null hypothesis that the empirical distribution in Table A-2 is a sample from an underlying normal distribution with $E[X]$ = 0.4 and $Var[X]$ = 0.16. Use 1% and 5% levels of significance.

3. Test the null hypothesis that the empirical distribution in Table A-3 for n = 100 is a sample from an underlying Poisson distribution. Use a 5% level of significance. For k = 7, calculate

$$E[X] = \sum_{i=1}^{k} x_i f(x_i)$$

to determine λ based on $E[X] = \lambda t$, where X is a Poisson r.v. with pdf

$$f(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x = 0, 1, 2, \ldots$$

Table A-3. Distribution of emergency room arrivals

Number of arrivals per hour, x	Number of one-hour time periods	pdf	CDF,
		$f(x)$	$F(x)$
0	10	0.10	0.10
1	28	0.28	0.38
2	29	0.29	0.67
3	16	0.26	0.83
4	10	0.10	0.93
5	6	0.06	0.99
6	1	0.01	1.00
	n = 100	1.0	

A.3 Kolmogorov-Smirnov Goodness-of-Fit Test

This test compares the continuous CDF, $F(x)$, of a uniformly distributed random variable X over the interval [0,1] to the empirical CDF, $S_n(x)$, of a sample of size n. By definition,

$$F(x) = x, \quad 0 \le x \le 1.$$

If the sample from a random number generator is $u_1, u_2, \ldots, u_n$, the empirical CDF, $S_n(x)$ is defined as follows

$$S_n(x) = \frac{\text{Number of } u_1,\ldots,u_n \text{ that are} \le x}{n}$$

As n becomes larger, $S_n(x)$ should become a better approximation to $F(x)$, provided that the null hypothesis is true.

The Kolmogorov-Smirnov test is based on the largest absolute difference between $S_n(x)$ and $F(x)$ over the range of the random variable X. If we let $D = \max |S_n(x) - F(x)|$, the sampling distribution of D is known and is tabulated as a function of N below. Table A-4 provides one-tailed values of $d_\alpha(n)$ such that $\Pr\{\max[\, |S_n(x) - F(x)|\,] > d_\alpha(n)\} = \alpha$, where α is desired level of significance.

Table A-4. Critical values for one-tailed K-S test

Sample size (n)	Level of significance (α)				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
> 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.35}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$