Sampling and Sampling Distributions

Random Sampling

A **sample** is a group of objects or readings taken from a population for counting or measurement. We shall distinguish between two kinds of populations — **finite populations** and **infinite populations.**

The sample must be random, meaning that all possible samples of a particular size must have equal probabilities of being chosen from the population. This will prevent bias in the sampling process.

Most modern sampling procedures involve the use of **probability sampling.** In probability sampling, a random devicesuch as tossing a coin, consulting a table of random numbers, or employing a random-number generatoris used to decide which members of the population will constitute the sample instead of leaving such decisions to human judgment.

Simple Random Sampling

A sample of size n from a finite population of size N is random if it is chosen in such a way that each of the $\binom{N}{n}$ possible samples has the same probability, $\frac{1}{\binom{N}{n}}$, of being selected.

EXAMPLE: As reported by the World Almanac, the top five state officials of Oklahoma are as shown in the table below. Consider these five officials a population of interest.

Governor (G)
Lieutenant Governor (L)
Secretary of State (S)
Attorney General (A)
Treasurer (T)

(a) List the possible samples (without replacement) of two officials from this population of five officials.

(b) Describe a method for obtaining a simple random sample of two officials from this population of five officials.

(c) For the sampling method described in part (b), what are the chances that any particular sample of two officials will be the one selected?

(d) Repeat parts (a)–(c) for samples of size 4.

EXAMPLE: As reported by the World Almanac, the top five state officials of Oklahoma are as shown in the table below. Consider these five officials a population of interest.

Governor (G)
Lieutenant Governor (L)
Secretary of State (S)
Attorney General (A)
Treasurer (T)

(a) List the possible samples (without replacement) of two officials from this population of five officials.

(b) Describe a method for obtaining a simple random sample of two officials from this population of five officials.

(c) For the sampling method described in part (b), what are the chances that any particular sample of two officials will be the one selected?

(d) Repeat parts (a)–(c) for samples of size 4.

Solution: For convenience, we represent the officials in the Table above by using the letters in parentheses.

(a) The Table below lists the 10 possible samples of two officials from this population of five officials.

G, L	G, S	G, A	G, T	L, S
L, A	L, T	S, A	S, T	A, T

(b) To obtain a simple random sample of size 2, we could write the letters that correspond to the five officials (G, L, S, A, and T) on separate pieces of paper. After placing these five slips of paper in a box and shaking it, we could, while blindfolded, pick two slips of paper.

(c) The procedure described in part (b) will provide a simple random sample. Consequently, each of the possible samples of two officials is equally likely to be the one selected. There are 10 possible samples, so the chances are $\frac{1}{10}$ (1 in 10) that any particular sample of two officials will be the one selected.

(d) The Table below lists the five possible samples of four officials from this population of five officials. A simple random sampling procedure, such as picking four slips of paper out of a box, gives each of these samples a 1 in 5 chance of being the one selected.

G, L, S, A	G, L, S, T
G, L, A, T	G, S, A, T
L, S, A, T	

Random-Number Tables

Obtaining a simple random sample by picking slips of paper out of a box is usually impractical, especially when the population is large. Fortunately, we can use several practical procedures to get simple random samples. One common method involves a **table of random numbers** — a table of randomly chosen digits

EXAMPLE: Student questionnaires, known as "teacher evaluations," gained widespread use in the late 1960s and early 1970s. Generally, professors hand out evaluation forms a week or so before the final.

That practice, however, poses several problems. On some days, less than 60% of students registered for a class may attend. Moreover, many of those who are present complete their evaluation forms in a hurry in order to prepare for other classes. A better method, therefore, might be to select a simple random sample of students from the class and interview them individually.

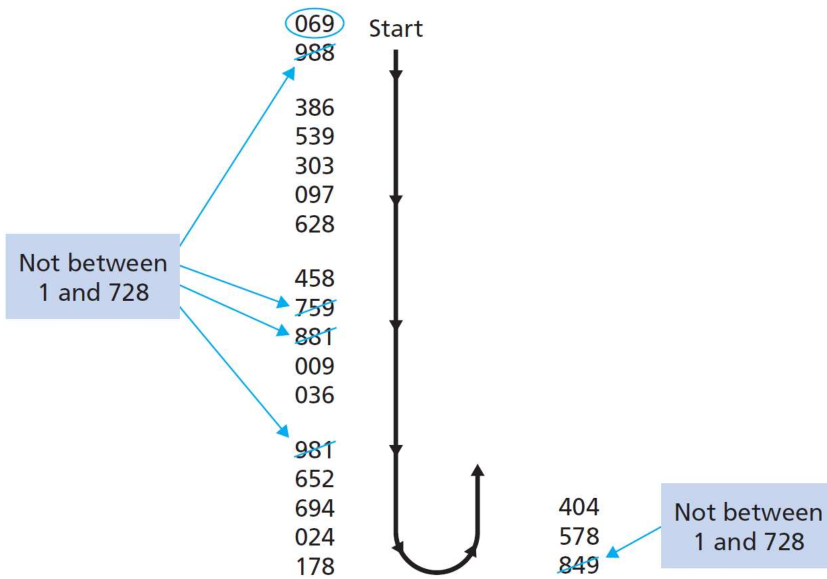
During one semester, Professor Hassett wanted to sample the attitudes of the students taking college algebra at his school. He decided to interview 15 of the 728 students enrolled in the course. Using a registration list on which the 728 students were numbered 1-728, he obtained a simple random sample of 15 students by randomly selecting 15 numbers between 1 and 728. To do so, he used the following random-number table:

Line number	Column number									
	00–09		*10–19*		*20–29*		*30–39*		*40–49*	
00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210
02	47435	53308	40718	29050	74858	64517	93573	51058	68501	42723
03	91312	75137	86274	59834	69844	19853	06917	17413	44474	86530
04	12775	08768	80791	16298	22934	09630	98862	39746	64623	32768
05	31466	43761	94872	92230	52367	13205	38634	55882	77518	36252
06	09300	43847	40881	51243	97810	18903	53914	31688	06220	40422
07	73582	13810	57784	72454	68997	72229	30340	08844	53924	89630
08	11092	81392	58189	22697	41063	09451	09789	00637	06450	85990
09	93322	98567	00116	35605	66790	52965	62877	21740	56476	49296
10	80134	12484	67089	08674	70753	90959	45842	59844	45214	36505
11	97888	31797	95037	84400	76041	96668	75920	68482	56855	97417
12	92612	27082	59459	69380	98654	20407	88151	56263	27126	63797
13	72744	45586	43279	44218	83638	05422	00995	70217	78925	39097
14	96256	70653	45285	26293	78305	80252	03625	40159	68760	84716
15	07851	47452	66742	83331	54701	06573	98169	37499	67756	68301
16	25594	41552	96475	56151	02089	33748	65289	89956	89559	33687
17	65358	15155	59374	80940	03411	94656	69440	47156	77115	99463
18	09402	31008	53424	21928	02198	61201	02457	87214	59750	51330
19	97424	90765	01634	37328	41243	33564	17884	94747	93650	77668

To select 15 random numbers between 1 and 728, we first pick a random starting point, say, by closing our eyes and placing a finger on the Table above. Then, beginning with the three digits under the finger, we go down the table and record the numbers as we go. Because we want numbers between 1 and 728 only, we discard the number 000 and numbers between 729 and 999. To avoid repetition, we also eliminate duplicate numbers. If we have not found enough numbers by the time we reach the bottom of the table, we move over to the next column of three-digit numbers and go up.

Using this procedure, Professor Hassett began with 069, circled in the Table above. Reading down from 069 to the bottom of the Table above and then up the next column of three-digit numbers, he found the 15 random numbers displayed in the Figure below and in the Table on the right. Professor Hassett then interviewed the 15 students whose registration numbers are shown in the Table on the right.

69	303	458	652	178
386	97	9	694	578
539	628	36	24	404



Random-Number Generators

Nowadays, statisticians prefer statistical software packages or graphing calculators, rather than random-number tables, to obtain simple random samples. The built-in programs for doing so are called **random-number generators**. When using randomnumber generators, be aware of whether they provide samples with replacement or samples without replacement.

Other Sampling Designs

Simple random sampling is the most natural and easily understood method of probability sampling — it corresponds to our intuitive notion of random selection by lot. However, simple random sampling does have drawbacks. For instance, it may fail to provide sufficient coverage when information about subpopulations is required and may be impractical when the members of the population are widely scattered geographically. Below we examine some commonly used sampling procedures that are often more appropriate than simple random sampling.

Systematic Random Sampling

One method that takes less effort to implement than simple random sampling is **systematic random sampling**.

> ### Systematic Random Sampling
>
> **Step 1** **Divide the population size by the sample size and round the result down to the nearest whole number, m.**
>
> **Step 2** **Use a random-number table or a similar device to obtain a number, k, between 1 and m.**
>
> **Step 3** **Select for the sample those members of the population that are numbered $k, k+m, k+2m, \ldots$.**

EXAMPLE: Recall the previous Example, in which Professor Hassett wanted a sample of 15 of the 728 students enrolled in college algebra at his school. Use systematic random sampling to obtain the sample.

Solution:

Step 1: Divide the population size by the sample size and round the result down to the nearest whole number, m.

The population size is the number of students in the class, which is 728, and the sample size is 15. Dividing the population size by the sample size and rounding down to the nearest whole number, we get $728/15 = 48$ (rounded down). Thus, $m = 48$.

Step 2: Use a random-number table or a similar device to obtain a number, k, between 1 and m.

Referring to Step 1, we see that we need to randomly select a number between 1 and 48. Using a random-number table, we obtained the number 22 (but we could have conceivably gotten any number between 1 and 48, inclusive). Thus, $k = 22$.

Step 3: Select for the sample those members of the population that are numbered $k, k+m, k+2m, \ldots$

From Steps 1 and 2, we see that $k = 22$ and $m = 48$. Hence, we need to list every 48th number, starting at 22, until we have 15 numbers. Doing so, we get the 15 numbers displayed in the Table on the right.

22	166	310	454	598
70	214	358	502	646
118	262	406	550	694

Interpretation: If Professor Hassatt had used systematic random sampling and had begun with the number 22, he would have interviewed the 15 students whose registration numbers are shown in the Table above.

Cluster Sampling

Another sampling method is cluster sampling, which is particularly useful when the members of the population are widely scattered geographically.

> **Cluster Sampling**
>
> **Step 1** **Divide the population into groups (clusters).**
>
> **Step 2** **Obtain a simple random sample of the clusters.**
>
> **Step 3** **Use all the members of the clusters obtained in Step 2 as the sample.**

Many years ago, citizens' groups pressured the city council of Tempe, Arizona, to install bike paths in the city. The council members wanted to be sure that they were supported by a majority of the taxpayers, so they decided to poll the city's homeowners.

Their first survey of public opinion was a questionnaire mailed out with the city's 18,000 homeowner water bills. Unfortunately, this method did not work very well. Only 19.4% of the questionnaires were returned, and a large number of those had written comments that indicated they came from avid bicyclists or from people who strongly resented bicyclists. The city council realized that the questionnaire generally had not been returned by the average homeowner.

An employee in the city's planning department had sample survey experience, so the council asked her to do a survey. She was given two assistants to help her interview 300 homeowners and 10 days to complete the project.

The planner first considered taking a simple random sample of 300 homes: 100 interviews for herself and for each of her two assistants. However, the city was so spread out that an interviewer of 100 randomly scattered homeowners would have to drive an average of 18 minutes from one interview to the next. Doing so would require approximately 30 hours of driving time for each interviewer and could delay completion of the report. The planner needed a different sampling design.

EXAMPLE: To save time, the planner decided to use cluster sampling. The residential portion of the city was divided into 947 blocks, each containing 20 homes, as shown in the Figure below. Explain how the planner used cluster sampling to obtain a sample of 300 homes.

EXAMPLE: To save time, the planner decided to use cluster sampling. The residential portion of the city was divided into 947 blocks, each containing 20 homes, as shown in the Figure below. Explain how the planner used cluster sampling to obtain a sample of 300 homes.

Solution:

Step 1: Divide the population into groups (clusters).

The planner used the 947 blocks as the clusters, thus dividing the population (residential portion of the city) into 947 groups.

Step 2: Obtain a simple random sample of the clusters.

The planner numbered the blocks (clusters) from 1 to 947 and then used a table of random numbers to obtain a simple random sample of 15 of the 947 blocks.

Step 3: Use all the members of the clusters obtained in Step 2 as the sample.

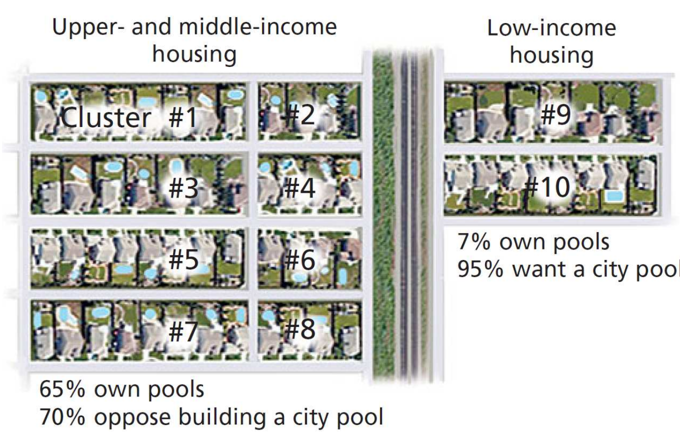
The sample consisted of the 300 homes comprising the 15 sampled blocks:

$$15 \text{ blocks} \times 20 \text{ homes per block} = 300 \text{ homes.}$$

Interpretation: The planner used cluster sampling to obtain a sample of 300 homes: 15 blocks of 20 homes per block. Each of the three interviewers was then assigned 5 of these 15 blocks. This method gave each interviewer 100 homes to visit (5 blocks of 20 homes per block) but saved much travel time because an interviewer could complete the interviews on an entire block before driving to another neighborhood. The report was finished on time.

Although cluster sampling can save time and money, it does have disadvantages. Ideally, each cluster should mirror the entire population. In practice, however, members of a cluster may be more homogeneous than the members of the entire population, which can cause problems.

For instance, consider a simplified small town, as depicted in the Figure below. The town council wants to build a town swimming pool. A town planner needs to sample homeowner opinion about using public funds to build the pool. Many upper-income and middle-income homeowners may say "No" if they own or can access pools. Many low-income homeowners may say "Yes" if they do not have access to pools.



If the planner uses cluster sampling and interviews the homeowners of, say, three randomly selected clusters, there is a good chance that no low-income homeowners will be interviewed. And if no low-income homeowners are interviewed, the results of the survey will be misleading. If, for instance, the planner surveyed clusters #3, #5, and #8, then his survey would show that only about 30% of the homeowners want a pool. However, that is not true, because more than 40% of the homeowners actually want a pool. The clusters most strongly in favor of the pool would not have been included in the survey.

In this hypothetical example, the town is so small that common sense indicates that a cluster sample may not be representative. However, in situations with hundreds of clusters, such problems may be difficult to detect.

Stratified Sampling

Another sampling method, known as **stratified sampling**, is often more reliable than cluster sampling. In stratified sampling the population is first divided into subpopulations, called strata, and then sampling is done from each stratum. Ideally, the members of each stratum should be homogeneous relative to the characteristic under consideration.

In stratified sampling, the strata are often sampled in proportion to their size, which is called proportional allocation.

Stratified Random Sampling with Proportional Allocation

Step 1 **Divide the population into subpopulations (strata).**

Step 2 **From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.**

Step 3 **Use all the members obtained in Step 2 as the sample.**

EXAMPLE: Consider again the town swimming pool situation. The town has 250 homeowners of which 25, 175, and 50 are upper income, middle income, and low income, respectively. Explain how we can obtain a sample of 20 homeowners, using stratified sampling with proportional allocation, stratifying by income group.

EXAMPLE: Consider again the town swimming pool situation. The town has 250 homeowners of which 25, 175, and 50 are upper income, middle income, and low income, respectively. Explain how we can obtain a sample of 20 homeowners, using stratified sampling with proportional allocation, stratifying by income group.

Solution:

Step 1: Divide the population into subpopulations (strata).

We divide the homeowners in the town into three strata according to income group: upper income, middle income, and low income.

Step 2: From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.

Of the 250 homeowners, 25 are upper income, 175 are middle income, and 50 are lower income. The sample size for the upper-income homeowners is, therefore,

$$\text{Total sample size} \times \frac{\text{Number of high-income homeowners}}{\text{Total number of homeowners}} = 20 \cdot \frac{25}{250} = 2$$

Similarly, we find that the sample sizes for the middle-income and lower-income homeowners are 14 and 4, respectively. Thus, we take a simple random sample of size 2 from the 25 upper-income homeowners, of size 14 from the 175 middle-income homeowners, and of size 4 from the 50 lower-income homeowners.

Step 3: Use all the members obtained in Step 2 as the sample.

The sample consists of the 20 homeowners selected in Step 2, namely, the 2 upper-income, 14 middle-income, and 4 lower-income homeowners.

Interpretation: This stratified sampling procedure ensures that no income group is missed. It also improves the precision of the statistical estimates (because the homeowners within each income group tend to be homogeneous) and makes it possible to estimate the separate opinions of each of the three strata (income groups).