**Module 6: Sampling and Sampling Distributions**

# The Population-Sample Paradigm

Treat the observed data as a sample from a population.

Use sample characteristics to make inferences about population characteristics.

# Some Managerial Examples Where Sampling is Useful

Operations:  
 Manufacturer estimates proportion of defectives in a shipment.

Marketing:   
 Retailer estimates share of executives shopping in outlets.

Economics:  
 Questionnaire measures influence of price on customer purchasing.

# Target Population Versus Sampled Population

|  |  |  |
| --- | --- | --- |
| Sample | Target Population | Sampled Population |
|  |  |  |
| 100 Incomes | U.S. Incomes |  |
|  |  |  |
| Political Poll | Actual Voters |  |
| 20 Volunteer Tasters | Potential Consumers |  |
|  |  |  |
| CNN Web Poll | U.S. Opinions |  |
|  |  |  |
| 10 Goats | All Goats |  |

*Sampling* *bias* is a mismatch between the target population and the sampled population.

Typical causes of sampling bias: self-selection, non-response, incentives to answer, interviewer characteristics, formulation of questions, and wording of questions.[[1]](#footnote-1)

# Hypothetical Populations

Suppose a genetic scientist at an agricultural company harvests 200 oranges, the first of a new variety. Can these be considered as a sample from a population of interest?

From which populations might the following be considered a sample?

The 2004-2007 Microsoft returns *Microsoft\_Subset.JMP* (used in Module 2)

A bag of M&M’s candies

# (Simple) Random Sampling

Random Sampling - every possible subset of a given size has an equal chance of being drawn.

Can be obtained as a sequence of individual random draws from the population.[[2]](#footnote-2)

Sampling without replacement - items can only be selected once

Sampling with replacement - items can be selected repeatedly

When will sampling with replacement be virtually the same as sampling without replacement?

Random sampling should be done with a device that provides random selection.

Careful! Haphazard ≠ Random.

Other sampling designs include systematic sampling, stratified sampling, cluster sampling, and multistage sampling.

# iid Sampling

We shall be especially interested in simple random samples obtained by sampling from a population of a conceptually infinite size.[[3]](#footnote-3)

Real populations have finite size, but it’s often reasonable to treat them as infinite when the size of the sample is small relative to the size of the population.

In this case, the data *x*1, *x*2, …, *xn* an be thought of as

*n* independent draws from the same population.

Such samples are called *iid samples*:[[4]](#footnote-4)

*iid* = independent and identically distributed.

Notation: We’ll use

*x*1, *x*2, …, *xn* iid ~ N(*μ*, σ2)

to denote data obtained as an *iid* sample from the normal distribution N(*μ*, σ2).

What aspects of the 2004-2007 Microsoft returns (p. 1-17) support an assumption that

*x*1, *x*2, …, *x*1006 iid ~ N(*μ*, σ2)?

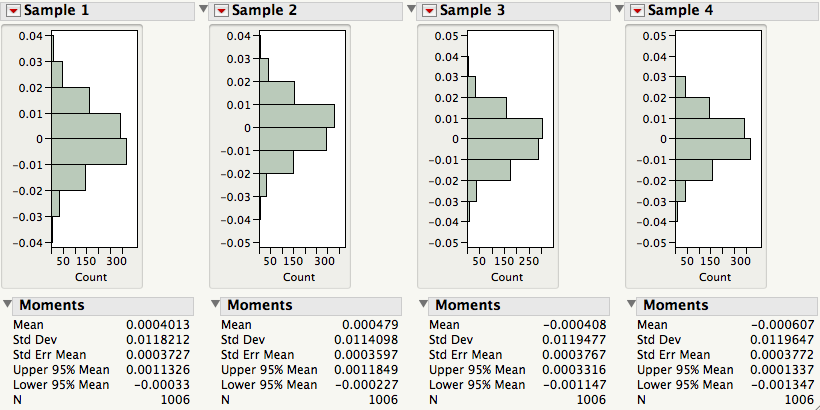
Suppose we had included the 2008 Microsoft returns. What then?

Key benefit of random sampling  
Statistical theory shows that characteristics of *iid* samples emulate population characteristics more and more closely as the sample size increases.

Example 1: Simulating *x*1,…,*x*50 *iid* selections with *Chipsim.JMP* from the “PICK A CHIP” distribution *p*(1) = .5, *p*(5) = .3, *p*(10) = .1, and *p*(20) = .1. Recall *μ* = 5 (p. 2-11) and σ = 5.744 (p. 2-14).



Example 2: Simulating *x*1,…, *x*1006 *iid* ~ N(0, .0122) with *Norm Sim.JMP*.



# Sample Estimates of Population Parameters

Simulated examples are artificial because we know population features such as *μ* and σ2. If we know the population, it’s easy to simulate samples. But what about inferences in the other direction? If we only observe a single sample *x*1, *x*2, …, *xn*, what can we *infer* about the population features such as *μ* and σ2?

In real problems, these population features **—** called *parameters* **—** are not known and must be *estimated* from data.

The sample statistics , *s*2 and *s* are typically used to estimate *μ*, σ2 and σ.

For Microsoft returns during 2004-2007,  = .00048 and *s* = .0120.

Based on our normal simulation results, does it seem plausible that *μ* = 0 and σ = .012 could be the true unknown values of the population for the 2004-2007 Microsoft returns?

# A Class Experiment

Organize into teams of 2 or 3 students.

Every team will receive a bag of M&M’s candies.

Is it reasonable to treat the contents of your bag as an *iid* sample from a population?

Which population?

Estimate the population proportion of blue M&M’s using only the information in your sample.

Will every team come up with the same estimate? Why not?

Note that a sample proportion is a special case of . Why?

# The Sampling Distribution of a Statistic

As previous examples show, sample estimates such as  or *s*, do not match *μ* and σ, and vary from sample to sample. Once we admit that we might have gotten a different value if we had gotten a different sample, we need to describe just how different the result might have been.

To quantify this *sample-to-sample variation*, we introduce two new populations:

The *population of samples* – the set of all possible samples (of a particular size *n*) that could be drawn from the original population.

The *population of values of the sample statistic* – the set of all possible values of the sample statistic –one for each sample.

Definition: the population of sample statistic values is called the *sampling distribution of the statistic*.

# Example: For the Class M&M’s Experiment[[5]](#footnote-5)

What is the population of samples? What is the population of sample statistic values?

How do the class samples and estimates relate to these populations?

# The Sampling Distribution of the Sample Mean

Astonishing Facts: For *x*1, *x*2, …, *xn*, *iid* from *any* population with mean *μ* and standard deviation *σ*, the sampling distribution of 

a) has mean *μ*

b) has standard deviation 

c) is approximately normal when *n* is large, and so is essentially determined by *μ* and *σ*.

Remarks:

These facts are known as the Central Limit Theorem (CLT). For practical purposes, normality can be assumed when *n* ≥ 15.[[6]](#footnote-6)

As we’ll see in Module 7, we can use these facts to estimate the sampling variation of  from the information in just *one sample*.

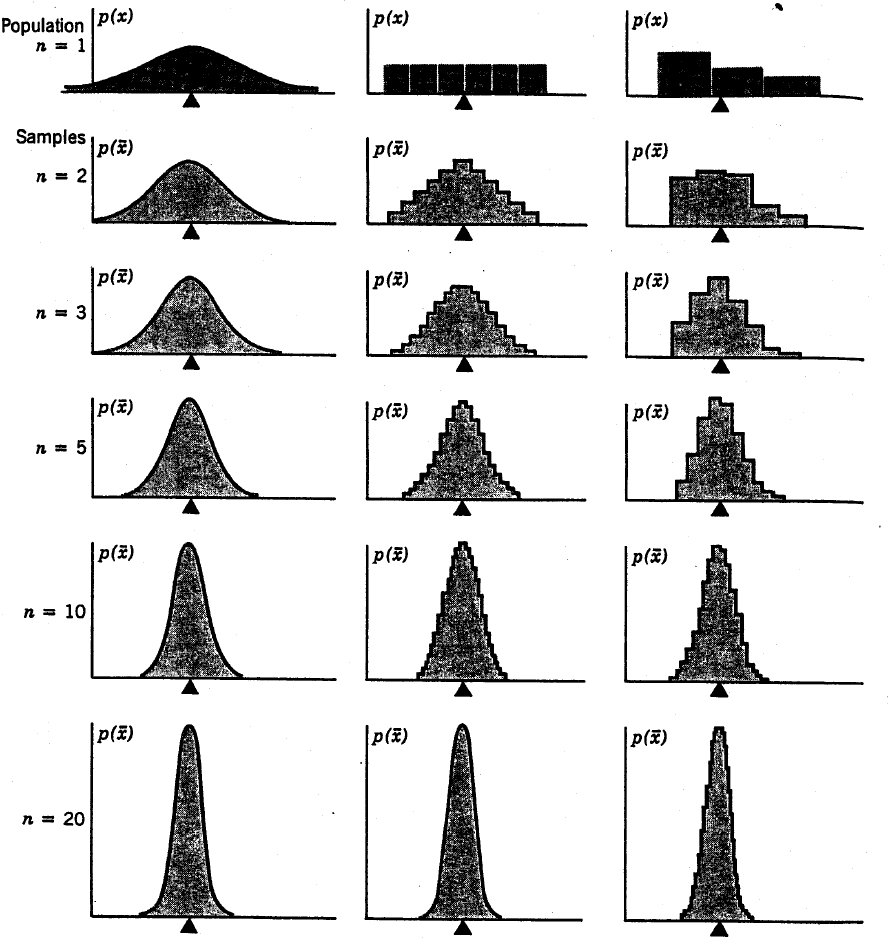
When the original population is exactly normal, then the sampling distribution of  will also be exactly normal.

The Astonishing Fact says that the sampling distribution of the sample mean  is approximately N(*μ*,) when the data are an iid sample.[[7]](#footnote-7)

Pictorially:

# The Astonishing Fact in Action!

The following slide shows the exact sampling distribution of  for three different populations and various values of *n.* Regardless of the shape of the population, the sampling distribution of the average gets closer and closer to a normal distribution.[[8]](#footnote-8)



# Motor Shaft Data

A quality control application provides another chance for us to see the sample-to-sample variation of the average. The file *ShaftXtr.JMP* contains 400 observations of diameters (in thousands of an inch) of motor shafts produced at a manufacturing plant. Five observations were taken per day for 16 weeks.[[9]](#footnote-9)



Look what happens when we plot daily means and weekly means instead.





The individual diameters can be considered as the means for 400 samples of size *n* = 1.

The daily averages can be considered as the means for 80 samples of size *n* = 5.

The weekly averages can be considered as the means for 16 samples of size *n* = 25.

Summaries of the individual, daily, and weekly means show the effects of larger and larger samples on the variation of the average.



Averages concentrate closer and closer around the overall mean as *n*, the number of shafts that are averaged, gets larger.

# Take-Away Review

When the data is a sample from a population, we can use the characteristics of the sample to make inference about characteristics of the sampled population. However, one must be careful to ensure that the sampled population is same as the target population.

When the sample is *iid*, we can estimate population parameters, such as *μ* and σ, by their corresponding sample statistics,  and *s*.

When the sample is *iid*, the sampling distribution of  has mean *μ*, standard deviation  and is approximately normal.

# Next Module

When we estimate a population parameter, we can also estimate how close the sample statistics are likely to come to the corresponding population parameters. The key ingredient is the *standard error* of the statistic.

When we combine standard error with the implications of the Central Limit Theorem, we can make profound statements about features of the population with *confidence intervals*.

1. Many of the examples that we use in these notes appear with further discussion in the course textbook. In some cases, the text uses the same data that we consider in class and in others, the text considers a similar data set. FS, section 13.4, p. 317, discusses these and other causes of sampling bias in surveys. [↑](#footnote-ref-1)
2. A sequence of independent draws from specified probability distributions can be obtained with JMP. From the list of “random” functions, consider choices such as random uniform or random normal. [↑](#footnote-ref-2)
3. An alternative, artificial way to avoid the complications of sampling from a finite population is to assume that we sample with replacement. If we put them back, the result of one case does not influence other cases. [↑](#footnote-ref-3)
4. Section 10.5 of SF has more of the background on the use of acronym iid. [↑](#footnote-ref-4)
5. SF discusses this experiment further in a case study that begins on page 296. [↑](#footnote-ref-5)
6. Convergence to normality depends on how weird the distribution of the data is. SF describes an explicit condition derived from the data on page 328. [↑](#footnote-ref-6)
7. If the data are not an iid sample, then almost anything can happen. For example, if the sample is a large share of the population, then the observations in the sample are dependent and the variance of the sample average is less than . [↑](#footnote-ref-7)
8. The case study “Modeling Sampling Variation” (SF, pages 296-302) describes the CLT in the context of counting M&Ms. The case also includes more figures to explain why counts becomes normally distributed as the number of summands increases. [↑](#footnote-ref-8)
9. Chapter 14 of SF gives further examples of the use of statistics in quality control. [↑](#footnote-ref-9)