**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 effect of price on customer purchasing.

HR: personnel manager estimates variability of hourly wages across an industry.

# Target Population Versus Sampling Frame

|  |  |  |
| --- | --- | --- |
| **Sample** | **Target Pop’n** | **Sampling Frame** |
|  |  |  |
| 100 Incomes | U.S. Incomes |  |
|  |  |  |
| Political Poll | Actual Voters |  |
|  |  |  |
| CNN Poll of Callers | U.S. Opinions |  |
|  |  |  |
| Sample 10 Goats | All Goats |  |

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

Typical causes of sampling bias: self selection, non-response, incentives to answer, interviewer characteristics, formulation of questions, and sensitivity of questions (see BBS, p. 108).[[1]](#footnote-2)

# 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 1,006 daily returns in *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-3)

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 - 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-4)

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 x1,…,x*n* can be thought of as

*n* independent draws from the same population.

Such samples are called *iid samples*:

*iid* = independent and identically distributed.

Notation: We’ll use

x1,…,x*n* iid ~ N(μ,σ2)

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

What aspects of the Microsoft daily returns (pp. 2-15 and BBS pp. 27-28)

support an assumption that

x1,…,x1,006 iid ~ N(μ,σ2)?

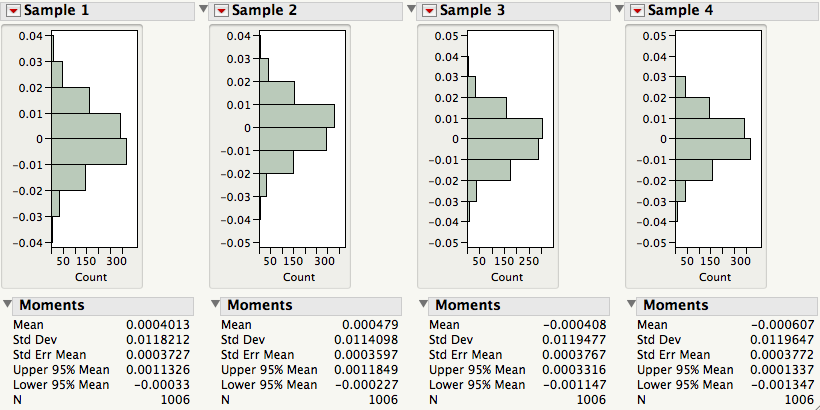
Key benefit: statistical theory shows that characteristics of iid samples tend to emulate population characteristics.

Example 1: Simulating x1,…,x50 iid selections with *Chipsim.JMP* from the chip distribution p(1) = .5, p(5) = .3, p(10) = .1, and p(20) = .1. Recall μ = 5 (pp. 3-11) and

σ = 5.744 (pp. 3-14).



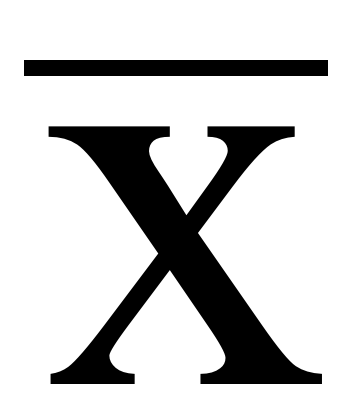
Example 2: Simulating x1,…,x507 iid ~ N(0, .0122) with *Norm Sim.JMP*.

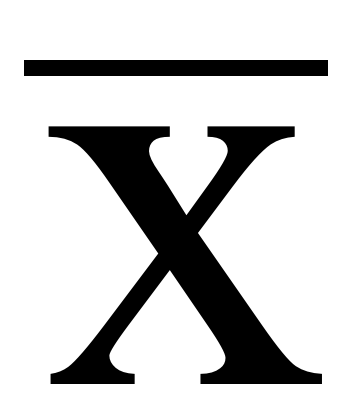


# Sample Estimates of Population Parameters

Simulation 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 the other direction? If we only observe a single sample x1,…,x*n*, 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 , s2 and s are typically used to estimate μ, σ2 and σ.

For the returns on Microsoft, we computed  = .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 for the population of 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*.

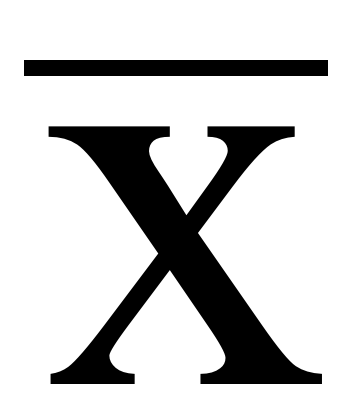
# Example: For the Class M&M’s Experiment

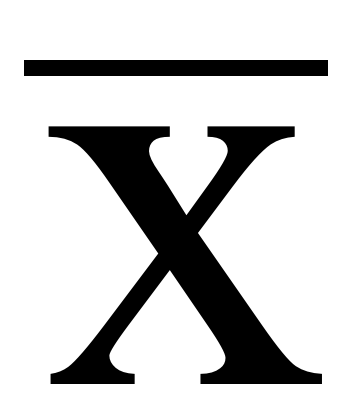
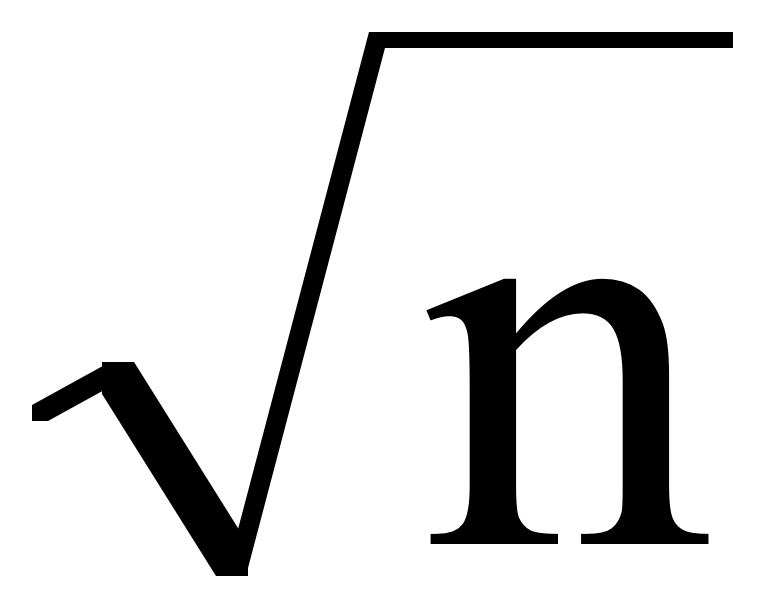
What is the population of samples? What is the population of sample statistic values?

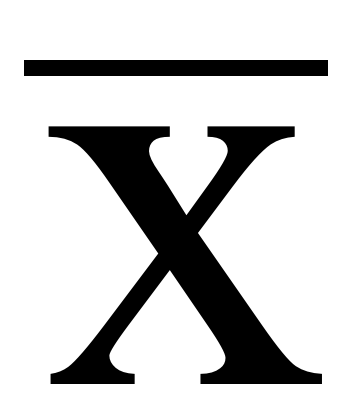
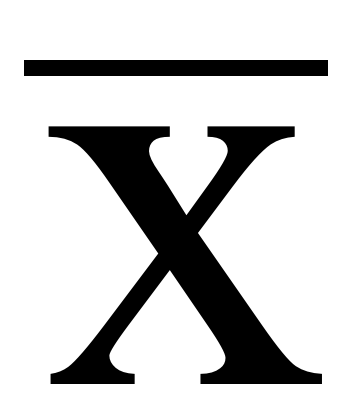
How do the class samples and estimates relate to these populations?

# The Sampling Distribution of

Astonishing Fact: For x1,…, x*n* iid from *any* population with mean μ and standard deviation σ, the sampling distribution of 

a) has mean μ  = μ.

b) has standard deviation σ  = (σ / ).[[4]](#footnote-5)

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

We can use this fact to estimate the sampling variation of  from the information in just *one sample*.

Remarks:

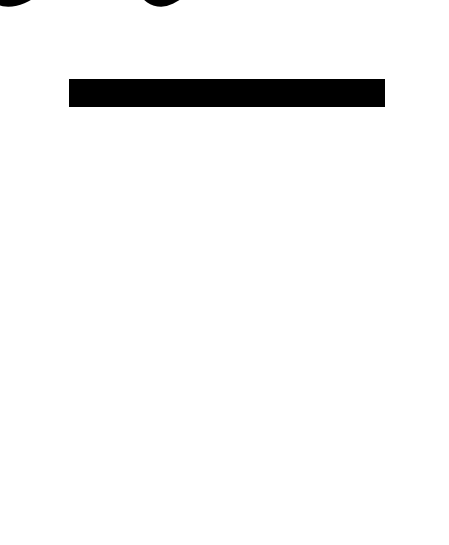
This is known as the Central Limit Theorem (CLT). For practical purposes, normality can be assumed when *n* ≥ 15.

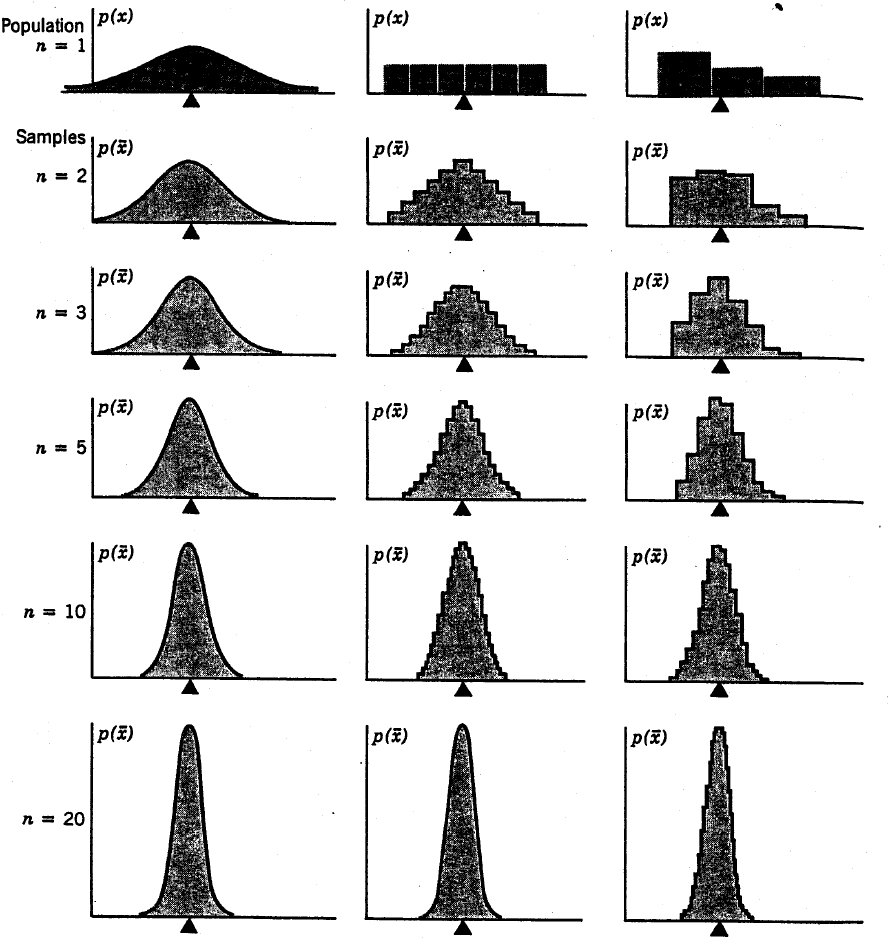
When the original population is exactly normal, then the sampling distribution of  will also be exactly normal.

The Astonishing Fact says that the sampling distribution of the sample mean  is always approximately N( , ).

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.



Dice

# Motor Shaft Data (BBS, p. 68)

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.



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





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

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

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

Summaries of the individual, daily, and weekly means (BBS, p. 69) 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 the number averaged, gets larger.

# Take-Away Review

The information in iid samples allows us to

Estimate population parameters, such as μ and σ, by their corresponding sample statistics,  and *s*.

as well as to

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.

The estimated standard deviation of  is *s*/ which can be calculated from the same sample used to calculate . It estimates σ/, the standard deviation of the sampling distribution of .

# Next Module

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 Basic Business Statistics (BBS) casebook. In some cases, the casebook uses the same data that we consider in class and in others, the casebook considers a similar data set. [↑](#footnote-ref-2)
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-3)
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-4)
4. This formula comes from the same methods used for portfolios. The average is a sum divided by n, and we know that the variance of a sum of independent items is the sum of the variances. Factor out the divisor and take a square root to get this expression. [↑](#footnote-ref-5)