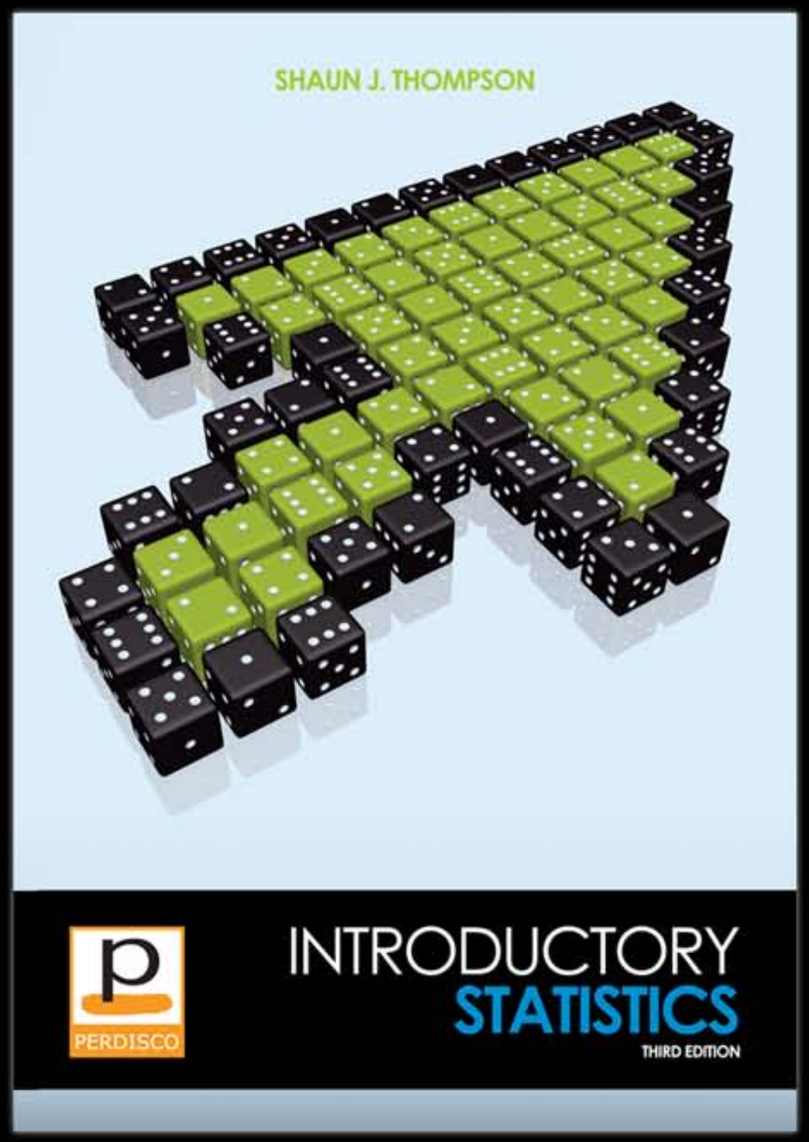


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

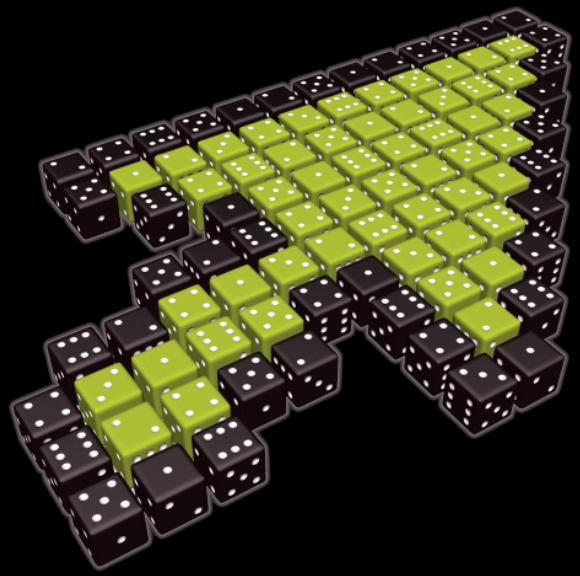
Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

Chapter 6

Sampling distributions



Using samples to talk about populations

- Often we will want to know something about a population, meaning we want to know a parameter
- Examples
  - In a nationwide test, what is the average score?
  - What proportion of people choose chocolate as their favorite ice cream flavor?
- We answer such questions by studying a sample
  - e.g. get sample of 100 test scores, calculate sample mean
- We use sample measurements (sample mean) to infer population measurements (population mean)

Why trust samples?

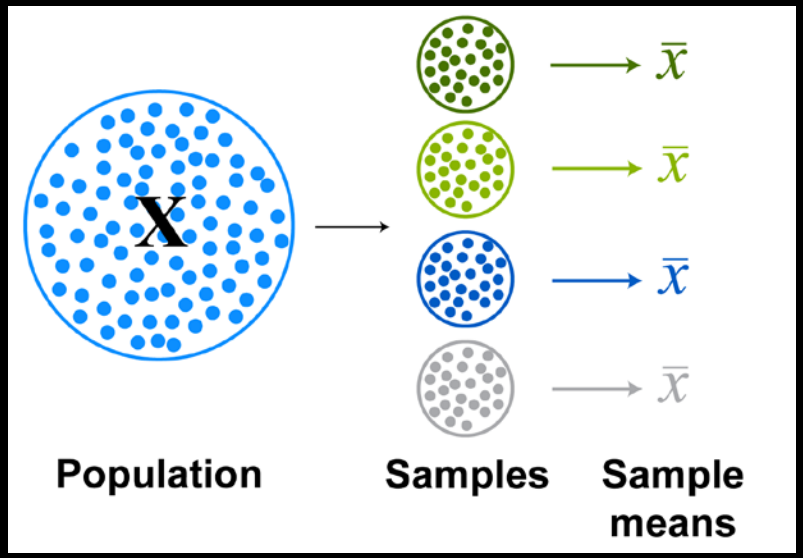
- But why can we do this?
- For example, if a different 100 test scores happened to be collected, we'd get a different sample mean
- Samples, and sample statistics, vary
- So why should we trust the sample mean to tell us anything about the (fixed) population mean?
- The reason we trust the sample mean is that samples behave 'nicely'

Behavior of samples

- So what does nicely mean?
- There's two important facts about sample means
- While sample means $\bar{x}$ do vary from sample to sample:
  1. they vary about the population mean, μ
  2. they don't vary 'much'
- This second property is a bit vague
- We need to be more precise than this!

Population of samples

- We must acknowledge that, for a given population, there are entire populations of sample statistics
- For example, suppose X is a numerical variable and we collect a sample and calculate $\bar{x}$
- There are lots of different possible values for $\bar{x}$
- In fact, there is an entire population of values!



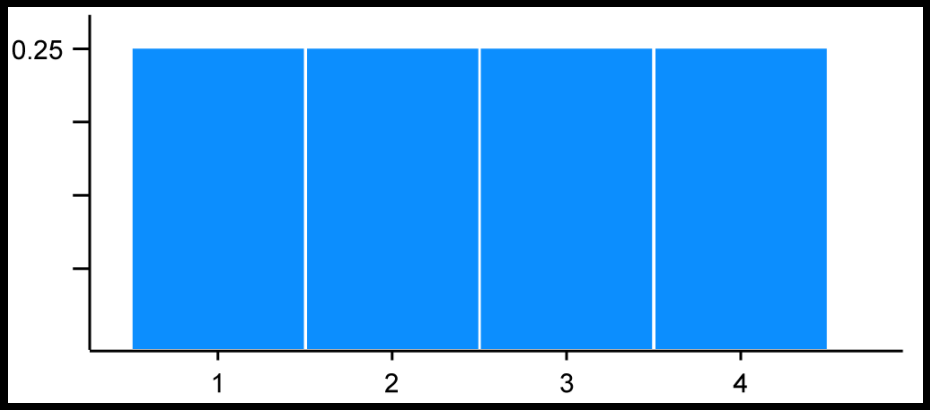
Sampling distribution

- The population of all possible sample means is denoted by $\overline{X}$
- The **sampling distribution of the mean** is the probability distribution for this new variable
- It describes how the sample means 'behave'
  - What sort of sample means are possible?
  - Are certain sample means more likely than others?
  - How much do the sample means vary from one another, and from the population mean?

Example – a small population

- Let X be the population of four values {1, 2, 3, 4}
- Assume every value is equally likely:

x	p(x)
1	0.25
2	0.25
3	0.25
4	0.25

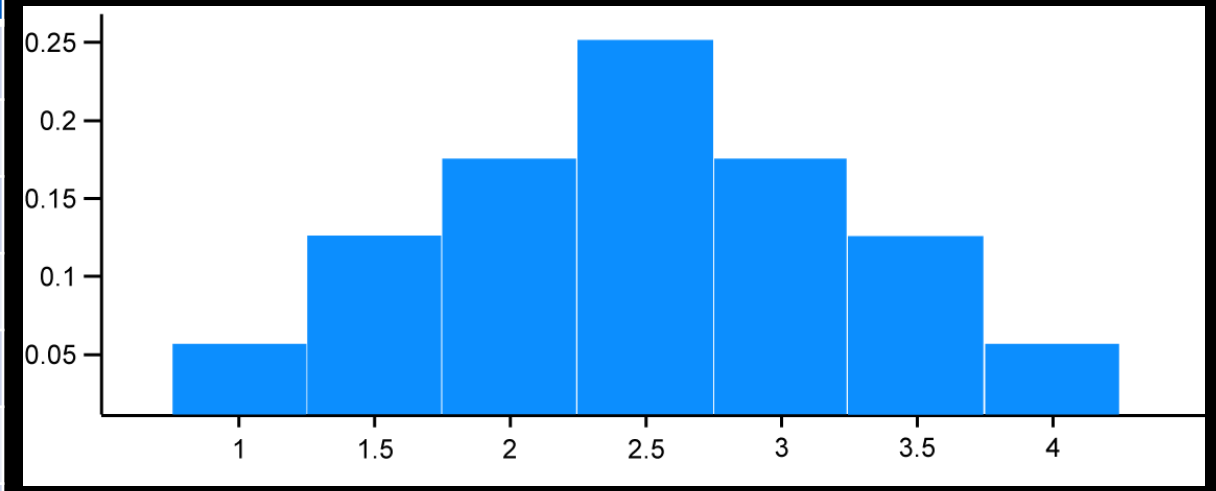


- X has mean μ = 2.5 and standard deviation σ = 1.12
- Now suppose we draw samples of size 2
- What sort of sample means could we get?

Example – a small population (continued)

- For samples of size 2, it can be shown that $\overline{X}$ has the following probability distribution:

$\overline{x}$	$p(\overline{x})$
1	0.0625
1.5	0.125
2	0.1875
2.5	0.25
3	0.1875
3.5	0.125
4	0.0625



- $\overline{X}$ has a mean of 2.5 and standard deviation of 0.79

Example – a small population (continued)

- Notice the mean of $\bar{X}$ is the same as the mean of X
- So sample means center on the population mean

- There is less variation in $\bar{X}$ than in X
- The sample means don't just center on the population mean – they are concentrated there

- Sample means tend to stay 'close' to μ

Properties of the sampling distribution

- To be precise, suppose we have any numerical variable X with mean μ and standard deviation σ
- Denote by $\overline{X}$ the set of sample means for samples of size n
- Then the mean of $\overline{X}$ is μ and the standard deviation is equal to $\sigma/\sqrt{n}$
- So the variation in sample means decreases if:
  - the variation in X itself is small
  - the sample size is large

Shape of the sampling distribution

- What about the shape of $\overline{X}$?
- If X is normal, then so is $\overline{X}$
- But even if X isn't normal, the sample means $\overline{x}$ still approximately follow the normal distribution, provided the sample size is large enough!
- This fact is a powerful result in statistics, known as the **Central Limit Theorem**

Central Limit Theorem

- This is one of the most useful results in statistics:

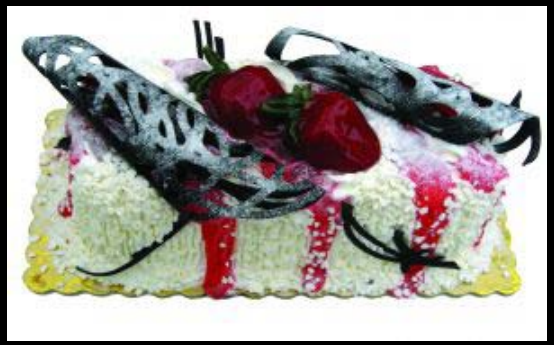
Suppose X is any numerical random variable with mean μ and standard deviation σ. Then for sufficiently large sample sizes n, the sampling distribution of the mean, $\overline{X}$, is approximately normal with mean μ, standard deviation $\sigma/\sqrt{n}$

- Typically, sample sizes of 30 are 'sufficiently' large

Proportions in categorical variables

- So far we have been looking at numerical variables
- What if X is categorical?
- Then the values of X are categories, and we can ask what proportion of values fall in each category
- Example: Favorite ice cream flavor out of chocolate, strawberry and vanilla
- These 3 flavors are

  the 3 categories



- What proportion of people prefer chocolate?

Population proportion

- Suppose we happen to know the proportion is 40%
- This is known as a **population proportion**, and is typically denoted π
- Just as with the population mean, we can estimate the population proportion by drawing samples from X and calculating a sample proportion, p
- Example: You might run a survey of 100 people and find that p = 39% of this survey prefers chocolate

Distribution of sample proportions

- Just as with the mean, the sample proportion we get will change, depending on the sample we choose!
  - e.g. someone else might run a survey and get p = 42%
- But, like with the mean, we'd expect the sample proportions to stick close to π
- And this is true!
- The variable of all sample proportions is denoted P
- The **sampling distribution of the proportion** is the probability distribution of P

Sampling distribution of the proportion

- Our scenario: X is a categorical variable, and we are interested in the proportion, π , of values that falls in one of its categories
- P is the variable of possible sample proportions, p, for samples of size n
- Then P approximately follows the normal distribution with:
  - mean equal to the population proportion, $\mu = \pi$
  - standard deviation given by $\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}$

Sampling distributions and description

- So far we've been able to describe the behavior of samples relative to the population parameters
- That is, given a numerical variable X with a population mean μ, we have been able to describe the behavior of sample means $\overline{x}$
- And given a categorical variable P with a population proportion π, we have been able to describe the behavior of sample proportions p
- But we really would like to go the other way around!

Sampling distributions and inference

What sampling distributions tell us already

μ or π known

Using a sampling distribution

We can say how samples behave

What we'd like sampling distributions to tell us

We've collected a sample and calculated the sample mean or sample proportion

Using a sampling distribution

What might μ or π look like?

Using sampling distributions

- When we use our knowledge of a sample to conclude something about the population, we call this **inference**
- Sampling distributions can help us with this
- Basic idea:
  - You collect a sample and measure the sample statistic (e.g. mean or proportion)
  - You then judge the likelihood of this sample relative to different values of the population parameter, by considering the sampling distribution

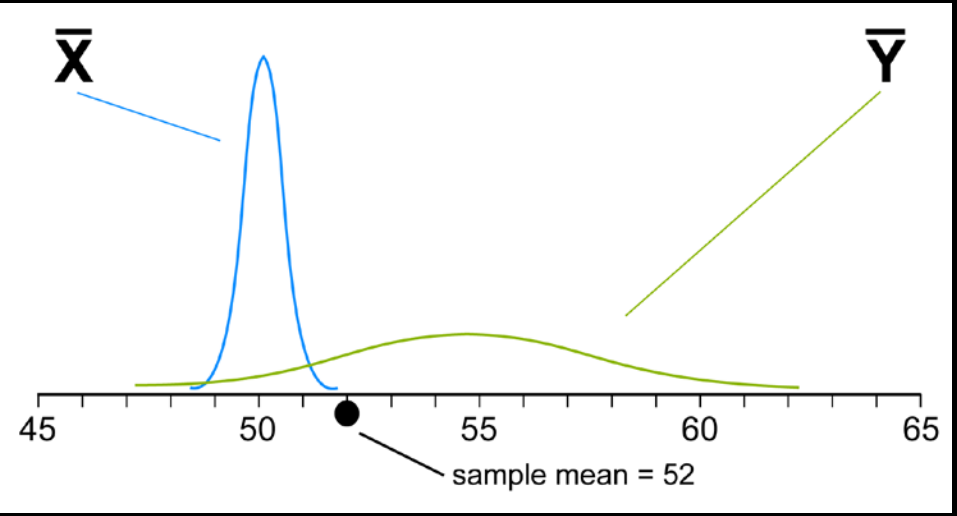
Example

- Suppose your friend has two normal distributions
  - X has a mean of 50 and a standard deviation of 5
  - Y has a mean of 55 and a standard deviation of 30
- You are given a sample of 100 values from one of these populations, but you don't know which one
- The sample mean is 52

- Which population do you think it came from?

Example continued

- Notice that the sample mean of 52 is a value in one of the two sampling distributions, $\overline{X}$ or $\overline{Y}$
- So let's look at those two sampling distributions:
  - $\overline{X}$ is normal, mean 50 and standard deviation $5/\sqrt{100} = 0.5$
  - $\overline{Y}$ is normal, mean 55 and standard deviation $30/\sqrt{100} = 3$



Example continued

- As a value in the sampling distribution $\overline{X}$, 52 seems extreme, it is in a very high part of the tail of $\overline{X}$
  - So if the sample came from X, it was an unlikely one
- In comparison, 52 does not seem like an extreme value in sampling distribution $\overline{Y}$
  - So if the sample came from Y, it was not an unlikely one
- So if we had to guess, we would guess the sample came from population Y
- That is a basic example of inference!

Fields of inference

- There are two major fields in statistical inference:

  Estimation

  - a sample is collected and a sample statistic calculated
  - a range of values for the population parameter is then given, based on the value of this sample statistic

  Testing

  - a value for the population parameter is hypothesized
  - a sample is collected and sample statistic calculated
  - the hypothesized value is then tested by comparing it to this calculated sample statistic

Nature of inference

- Both methods of inference are big fields in statistics
- We will study each in detail in later chapters
- But the philosophy behind both methods is in sampling distributions
- Both methods involve studying a sample in relation to the possible and unknown sampling distributions it could have come from

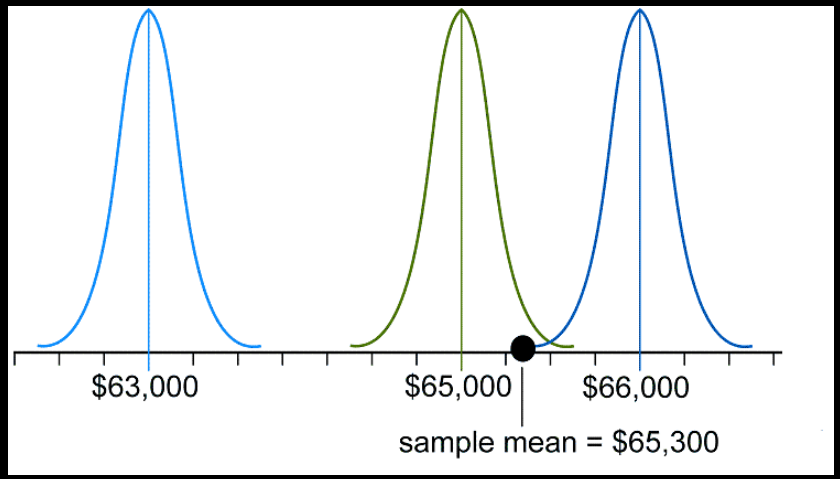
Estimation – example

- A survey of 625 annual incomes produces a sample mean income of $\bar{x}$ = $65,300
- A statistician then makes the following claim:

  "I am 95% confident that the true average income is somewhere between $64,900 and $65,700."
- This is an estimate
- Notice that it is:
  - imprecise (a whole range of values is provided)
  - uncertain (the statistician is only 95% confident)

Estimation – example continued

- But where does the estimate come from?
- Whatever the unknown true population mean μ is, the sample mean of $65,300 came from a sampling distribution related to it
- The range in the estimate then includes only those values μ for which the sample mean is not 'extreme'



- The details of this are treated in the next chapter!

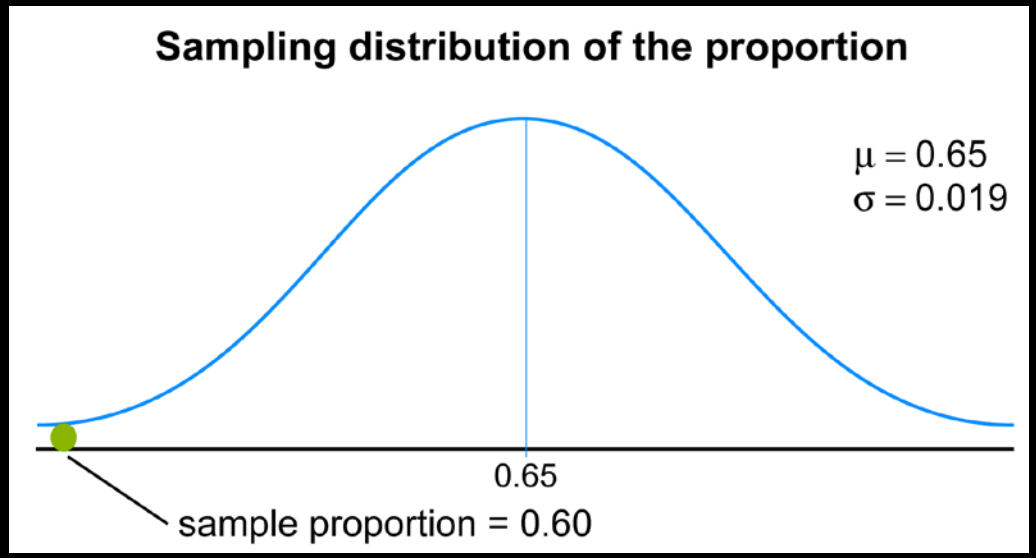
Testing – example

- A sociologist claims that the current approval rating of the government is 65%
- This is known as the **null hypothesis**
- A survey of n = 625 is conducted and the approval rating in the sample is only 60%
- This sample will do one of the following:
  - constitute enough evidence to reject the null hypothesis
  - not constitute enough evidence, in which case the sociologist's claim is not rejected by this test
- Note that 'not reject' is not the same as 'accept'!

Testing – example continued

- As with estimation, we do this by going to the sampling distribution
- If the sociologist is right, then the sampling distribution of the proportion is approximately normal with:
  - mean $\mu = 0.65$
  - standard deviation

    $\sigma \approx 0.019$



Testing – example continued

- A sample proportion of only 60% seems unlikely in this sampling distribution

- So we conclude that this isn't the sampling distribution

- That is, we conclude that the sociologist was wrong!