# Business Analytics


## Sampling Distributions

Faton Berisha

Chapter 6

# Sampling Distributions

# Sampling Distributions

# Sampling Distribution of the Sample Mean

A *sampling distribution of the sample mean* is the probability distribution of the population of the sample means obtainable from all possible samples of size *n* from a population of size N

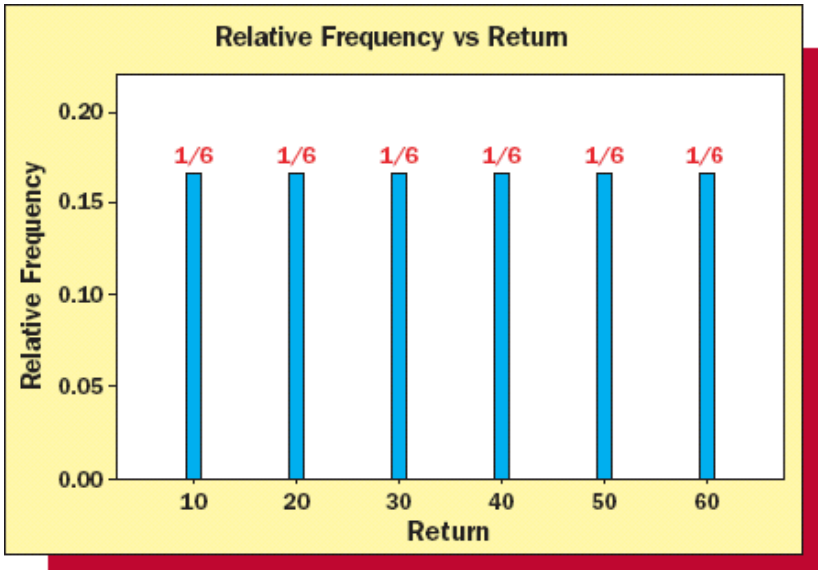
# Example: Sampling Annual % Return
## on 6 Stocks #1

- Population of the percent returns from six stocks
  - In order, the values of % return are:

    10%, 20%, 30%, 40%, 50% and 60%
    - Label each stock A, B, C, …, F in order of increasing % return
    - The mean rate of return is 35% with a standard deviation of 17.078%
  - Any one stock of these stocks is as likely to be picked as any other of the six
    - Uniform distribution with $N = 6$
    - Each stock has a probability of being picked of 1/6 = 0.1667

# Example: Sampling Annual % Return
## on 6 Stocks #2

| *Stock* | *% Return* | *Frequency* | *Relative Frequency* |
|---|---|---|---|
| Stock A | 10 | 1 | 1/6 |
| Stock B | 20 | 1 | 1/6 |
| Stock C | 30 | 1 | 1/6 |
| Stock D | 40 | 1 | 1/6 |
| Stock E | 50 | 1 | 1/6 |
| Stock F | 60 | 1 | 1/6 |
| **Total** | | **6** | **1** |



# Example: Sampling Annual % Return
## on 6 Stocks #3

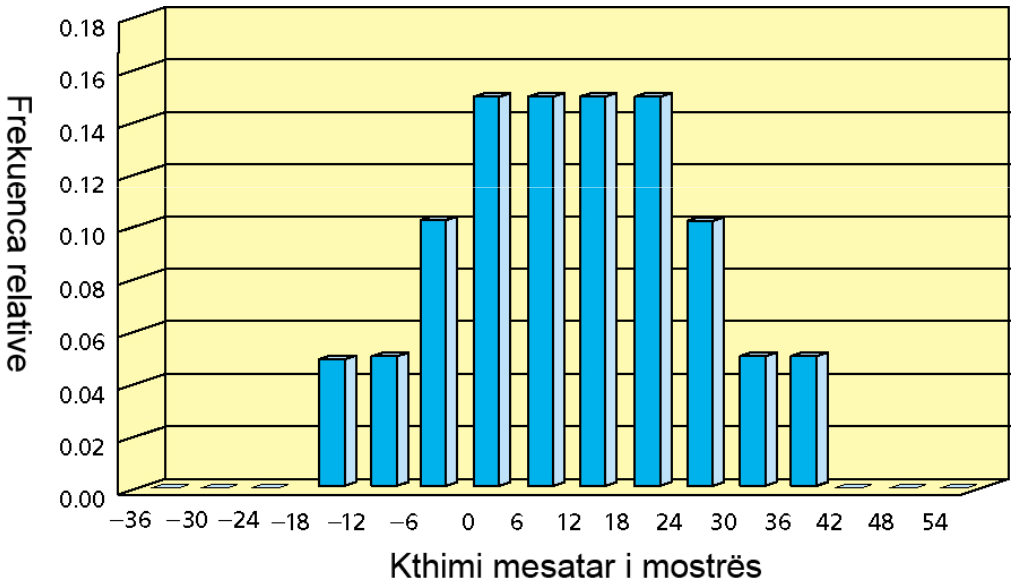
- Now, select all possible samples of size $n = 3$ from this population of stocks of size $N = 6$
  - Now select all possible pairs of stocks
- How to select?
  - Sample randomly
  - Sample without replacement
  - Sample without regard to order

# Example: Sampling Annual % Return
## on 6 Stocks #4

- Result: There are 20 possible samples of size $n = 3$
- Calculate the sample mean of each and every sample

# Example: Sampling Annual % Return
## on 6 Stocks #5

| | *n = 3 Returns* | | | *Sample* |
|---|---|---|---|---|
| *Sample* | *in the sample* | | | *mean* |
| 1 | -36 | -15 | 3 | -16.00 |
| 2 | -36 | -15 | 15 | -12.00 |
| 3 | -36 | -15 | 33 | -6.00 |
| 4 | -36 | -15 | 54 | 1.00 |
| 5 | -36 | 3 | 15 | -6.00 |
| 6 | -36 | 3 | 33 | 0.00 |
| 7 | -36 | 3 | 54 | 7.00 |
| 8 | -36 | 15 | 33 | 4.00 |
| 9 | -36 | 15 | 54 | 11.00 |
| 10 | -36 | 33 | 54 | 17.00 |
| 11 | -15 | 3 | 15 | 1.00 |
| 12 | -15 | 3 | 33 | 7.00 |
| 13 | -15 | 3 | 54 | 14.00 |
| 14 | -15 | 15 | 33 | 11.00 |
| 15 | -15 | 15 | 54 | 18.00 |
| 16 | -15 | 33 | 54 | 24.00 |
| 17 | 3 | 15 | 33 | 17.00 |
| 18 | 3 | 15 | 54 | 24.00 |
| 19 | 3 | 33 | 54 | 30.00 |
| 20 | 15 | 33 | 54 | 34.00 |



# Observations

- Although the population of $N = 6$ stock returns has a uniform distribution, …
- … the histogram of *20* sample mean returns:
  1. Seems to be centered over the sample mean return of 35%, and
  2. Appears to be bell-shaped and less spread out than the histogram of individual returns

# General Conclusions

- If the population of individual items is normal, then the population of all sample means is also normal
- Even if the population of individual items is not normal, there are circumstances when the population of all sample means is normal (Central Limit Theorem)

# General Conclusions Continued

- The mean of all possible sample means equals the population mean
  - That is, $\mu = \mu_{\bar{x}}$
- The standard deviation $\sigma_{\bar{x}}$ of all sample means is less than the standard deviation of the population
  - That is, $\sigma_{\bar{x}} < \sigma$
    - Each sample mean averages out the high and the low measurements, and so are closer to $\mu$ than many of the individual population measurements

# And the Empirical Rule

- The empirical rule holds for the sampling distribution of the sample mean
  - 68.26% of all possible sample means are within (plus or minus) one standard deviation $\sigma_{\bar{x}}$ of $\mu$
  - 95.44% of all possible observed values of *x* are within (plus or minus) two $\sigma_{\bar{x}}$ of $\mu$
  - 99.73% of all possible observed values of *x* are within (plus or minus) three $\sigma_{\bar{x}}$ of $\mu$

# Properties of the Sampling
## Distribution of the Sample Mean #1

- If the population being sampled is normal, then so is the sampling distribution of the sample mean, $\bar{x}$
- The mean $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu$$

- That is, the mean of all possible sample means is the same as the population mean

# Properties of the Sampling
## Distribution of the Sample Mean #2

- The variance $\sigma^2_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

  - That is, the variance of the sampling distribution of $\bar{x}$ is
    - directly proportional to the variance of the population, and
    - inversely proportional to the sample size

# Properties of the Sampling
## Distribution of the Sample Mean #3

- The standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

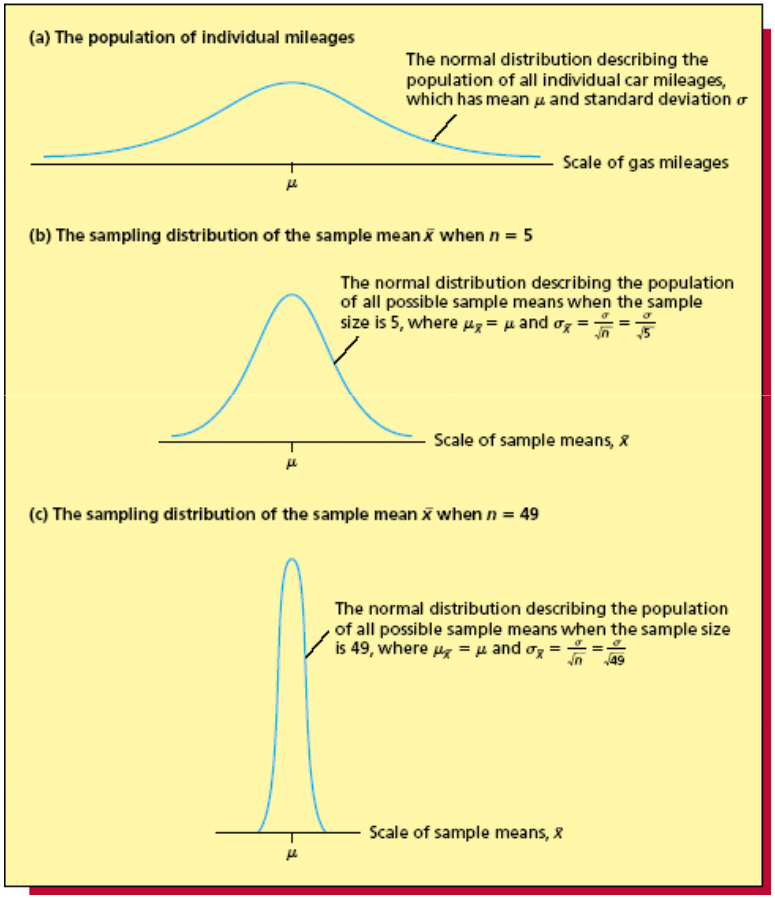
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - That is, the standard deviation of the sampling distribution of $\bar{x}$ is
    - directly proportional to the standard deviation of the population, and
    - inversely proportional to the square root of the sample size

# Notes

- The formulas for $\sigma^2_{\bar{x}}$ and $\sigma_{\bar{x}}$ hold if the sampled population is infinite
- The formulas hold approximately if the sampled population is finite but if *N* is much larger (at least 20 times larger) than the *n* ($N/n \geq 20$)
  - $\bar{x}$ is the point estimate of $\mu$, and the larger the sample size *n*, the more accurate the estimate
  - Because as *n* increases, $\sigma_{\bar{x}}$ decreases as $1/\sqrt{n}$
    - Additionally, as *n* increases, the more representative is the sample of the population
      - So, to reduce $\sigma_{\bar{x}}$, take bigger samples!

# Effect of sample size




$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{5}} = \frac{\sigma}{2.2361}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{49}} = \frac{\sigma}{7}$$

So, the different possible sample means will be more closely clustered around μ if $n$ = 49 than if $n$ =5

# Reasoning from the Sampling Distribution

- Recall from Chapter 2 mileage example, $\bar{x}$ = 31.5531 mpg for a sample of size n=49
  - With s = 0.7992
- Does this give statistical evidence that the population mean $\mu$ is greater than 31 mpg?
  - That is, does the sample mean give evidence that $\mu$ is at least 31 mpg?
- Calculate the probability of observing a sample mean that is greater than or equal to 31.5531 mpg if $\mu$ = 31 mpg
  - Want $P(\bar{x} \geq 31.5531 \text{ if } \mu = 31)$

# Reasoning from the Sampling Distribution Continued

- Use $s$ as the point estimate for $\sigma$ so that

- Then $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.7992}{\sqrt{49}} = 0.1143$

$$P(\bar{x} \geq 31.5531 \text{ if } \mu = 31) = P\left( z \geq \frac{31.5531 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \right)$$

$$= P\left( z \geq \frac{31.5531 - 31}{0.1143} \right)$$

$$= P(z \geq 4.84)$$

- But $z = 4.84$ is off the standard normal table
- The largest $z$ value in the table is 3.09, which has a right hand tail area of 0.001

# Reasoning from the Sampling Distribution #3

- $z = 4.84 > 3.09$, so $P(z \geq 4.84) < 0.001$
- That is, if $\mu = 31$ mpg, then fewer than 1 in 1,000 of all possible samples have a mean at least as large as observed
- Have either of the following explanations:
  - If $\mu$ is actually 31 mpg, then very unlucky in picking this sample

    OR
  - Not unlucky, but $\mu$ is not 31 mpg, but is really larger
- Difficult to believe such a small chance would occur, so conclude that there is strong evidence that $\mu$ does not equal 31 mpg
  - Also, $\mu$ is, in fact, actually larger than 31 mpg

# Central Limit Theorem

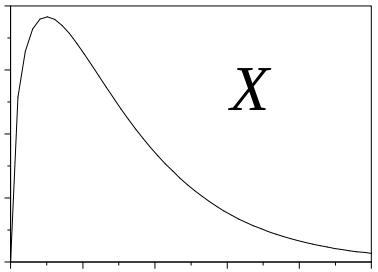
- Now consider sampling a non-normal population
- Still have: $\mu_{\bar{x}}=\mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
  - Exactly correct if infinite population
  - Approximately correct if population size *N* finite but much larger than sample size *n*
    - Especially if $N \geq 20 \times n$
- But if population is non-normal, what is the shape of the sampling distribution of the sample mean?
  - Is it normal, like it is if the population is normal?
  - Yes, the sampling distribution is approximately normal if the sample is large enough, even if the population is non-normal
    - By the “Central Limit Theorem”

# The Central Limit Theorem #2

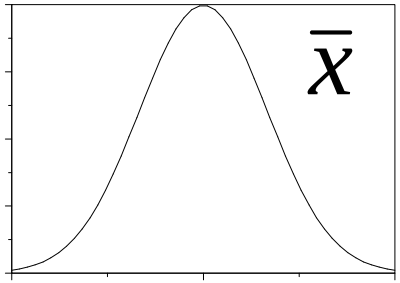
- No matter what is the probability distribution that describes the population, if the sample size *n* is large enough, then the population of all possible sample means is ***approximately*** normal with mean $\mu_{\bar{x}}=\mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- Further, the larger the sample size *n*, the closer the sampling distribution of the sample mean is to being normal
  - In other words, the larger *n*, the better the approximation

# The Central Limit Theorem #3

Random Sample $(x_1, x_2, \ldots, x_n)$




as $n \rightarrow$ large




Population Distribution

$(\mu, \sigma)$

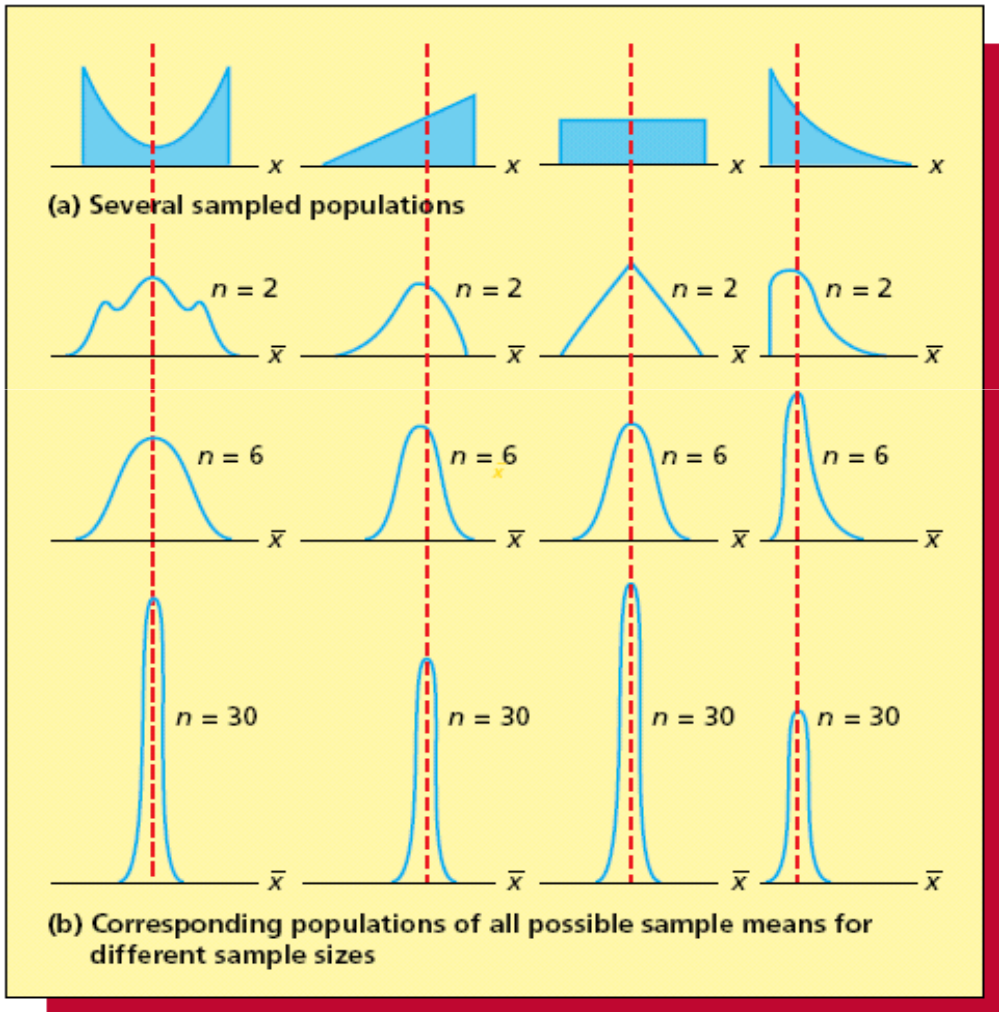
(right-skewed)

Sampling Distribution of Sample Mean

$\left(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \sigma/\sqrt{n}\right)$

(nearly normal)

# Example: Effect of Sample Size




The larger the sample size, the more nearly normally distributed is the population of all possible sample means

Also, as the sample size increases, the spread of the sampling distribution decreases

# How Large?

- How large is “large enough?”
- If the sample size is at least 30, then for most sampled populations, the sampling distribution of sample means is approximately normal
  - Here, if *n* is at least 30, it will be assumed that the sampling distribution of $\bar{x}$ is approximately normal
    - If the population is normal, then the sampling distribution of $\bar{x}$ is normal regardless of the sample size

# Sampling Distribution of the Sample Proportion

The probability distribution of all possible sample proportions is the *sampling distribution of the sample proportion* $\hat{p}$

If a random sample of size $n$ is taken from a population, then the sampling distribution of $\hat{p}$ is

- approximately normal, if $n$ is large, i.e., if $np \geq 5$ and $n(1-p) \geq 5$.
- has mean $\mu_{\hat{p}} = p$
- has standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

where $p$ is the population proportion and $\hat{p}$ is a sampled proportion

# Sampling Distribution of the
## Sample Proportion (Cont.)

- Example 6.4. The cheese spread case
  - Estimate if the proprtion is $p < 0.10$.
  - Assume $p = 0.10$.
  - Out of 1,000 interviewed 63 against

$$\hat{p} = \frac{63}{1000} = 0.063$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.10(1-0.10)}{1000}} = 0.0094868$$

$$P(\hat{p} \le 0.063 \mid p = 0.10) = P\left( z \le \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} \right)$$

$$= P\left( z \le \frac{0.063 - 0.10}{0.0094868} \right) < 0.001$$