Basic Statistics 01

Describing Data

Describing Data

1. Numerical Measures
   - Measures of Location
   - Measures of Dispersion
   - Correlation Analysis
2. Frequency Distributions
   - (Relative) Frequency Distribution
   - Histogram

Measures of Location

- Arithmetic mean
- Weighted mean

$$\overline{X}_w = \frac{(w_1 X_1 + w_2 X_2 + \ldots + w_n X_n)}{(w_1 + w_2 + \ldots w_n)}$$

- Geometric mean

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\ldots(X_n)}$$

- Median
  - The midpoint of the values after they have been ordered from the smallest to the largest
- Mode
  - The value of the observation that appears most frequently

Central location

- *Population mean* $\mu = \frac{\sum X_i}{N}$

- *Sample mean* $\bar{X} = \frac{\sum X_i}{n}$

  - Properties of the arithmetic mean
    - Every data set has a unique mean.
    - All the values are included in computing the mean.
    - The mean is affected by outliers.
    - The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.

Measures of Dispersion

- *Range*
- *Mean deviation*

$$MD = \frac{\Sigma \left| X_i - \overline{X} \right|}{n}$$

- *Variance*
  - Population variance

$$\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N}$$

  - Sample variance

$$s^2 = \frac{\Sigma (X_i - \overline{X})^2}{n-1}$$

- *Standard deviation (s.d.)*
- *coefficient of variation*

$$CV = \frac{s}{\overline{X}} (100\%)$$

Properties of Variance & S.D.

- The major characteristics of variance are:
  - All values are used in the calculation.
  - Not influenced by extreme values.
  - The units are the square of the original units.
- The population (sample) ***standard deviation*** σ (s) is the square root of the population (sample) variance.
  - The units are the same as the original ones.

The Coefficient of Correlation

- The (sample) ***Coefficient of Correlation*** (r) is a measure of the strength of the ***linear*** relationship between two variables.

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}$$

  - It can range from -1 to 1.
  - Values of -1 or 1 indicate perfect & strong correlation.
  - Values close to 0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Frequency Distribution

- A ***frequency distribution*** is a grouping of data into mutually exclusive categories showing the number of observations in each class.
- A ***relative frequency*** distribution shows the percent of observations in each class.

Example 1

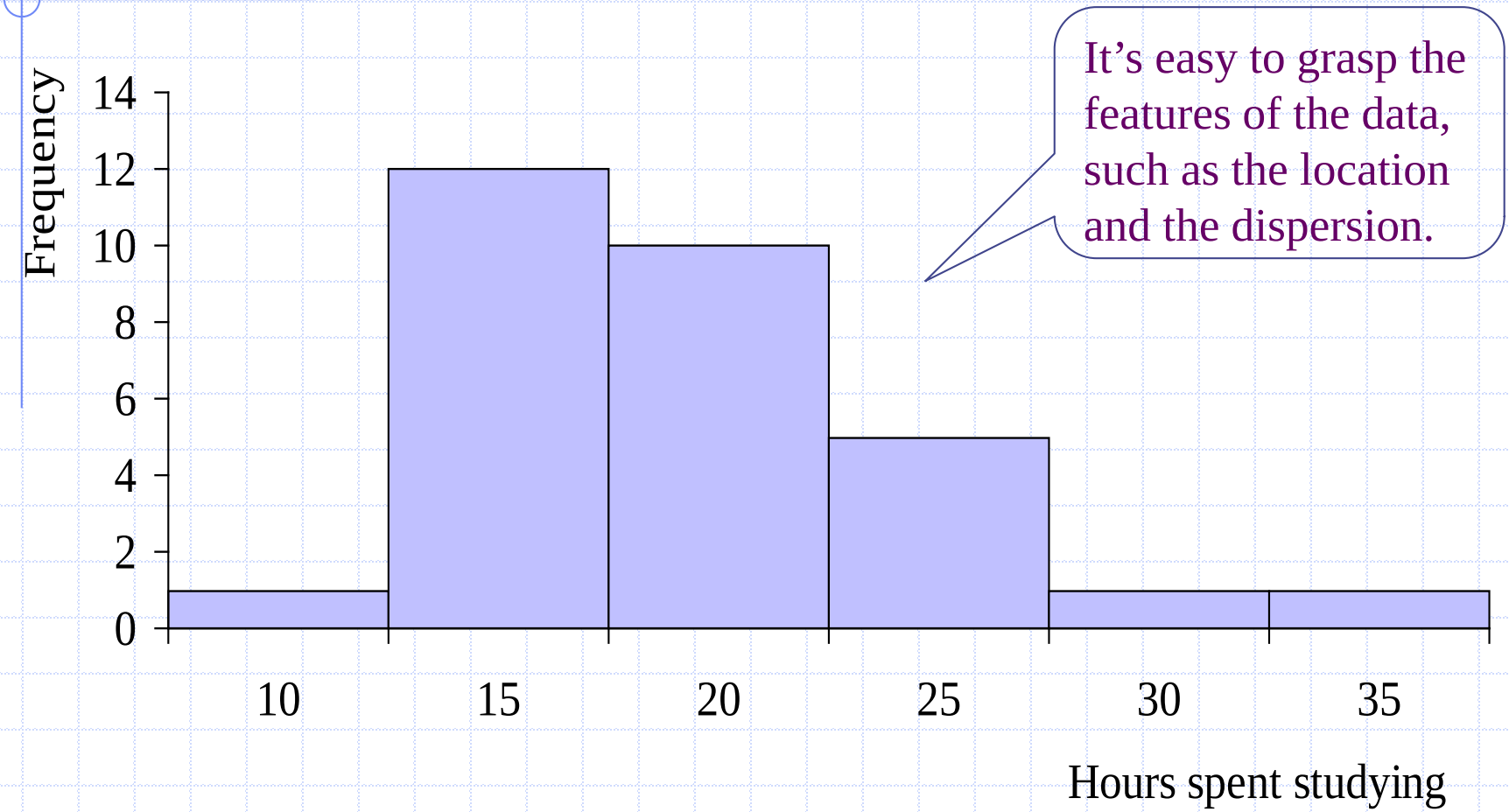
Dr. K is a professor of ABC University. He wishes to prepare a report showing the number of hours per week students spend studying. About his 30 students, he determines the number of hours each student studied last week.

15.0, 23.7, 19.7, 15.4, 18.3, 23.0, 14.2, 20.8, 13.5, 20.7, 17.4, 18.6, 12.9, 20.3, 13.7, 21.4, 18.3, 29.8, 17.1, 18.9, 10.3, 26.1, 15.7, 14.0, 17.8, 33.8, 23.2, 12.9, 27.1, 16.6.

Example 1 *continued*

Hours	Frequency f	Relative Frequency
7.5 up to 12.5	1	1/30=.0333
12.5 up to 17.5	12	12/30=.400
17.5 up to 22.5	10	10/30=.333
22.5 up to 27.5	5	5/30=.1667
27.5 up to 32.5	1	1/30=.0333
32.5 up to 37.5	1	1/30=.0333
TOTAL	30	30/30=1

Histogram for Studying-Hours



It's easy to grasp the features of the data, such as the location and the dispersion.

Basic Statistics 02

Probability Distributions

Probability Distributions

1. Probability Distributions
   - Random variable and probability distribution.
   - The mean, variance, and standard deviation of a (discrete) probability distribution.
2. Normal Distribution
   - Normal & Standard Normal Distribution
   - Calculating *z* value
   - Determining the probability

Random Variables & Probability Dist.

- A ***random variable*** is a numerical value determined by the outcome of an experiment.
- A ***probability distribution*** is the listing of all possible outcomes of an experiment and the corresponding probability.
  - The sum of the probabilities of the various outcomes is 1.
  - The probability of a particular outcome is between 0 and 1.

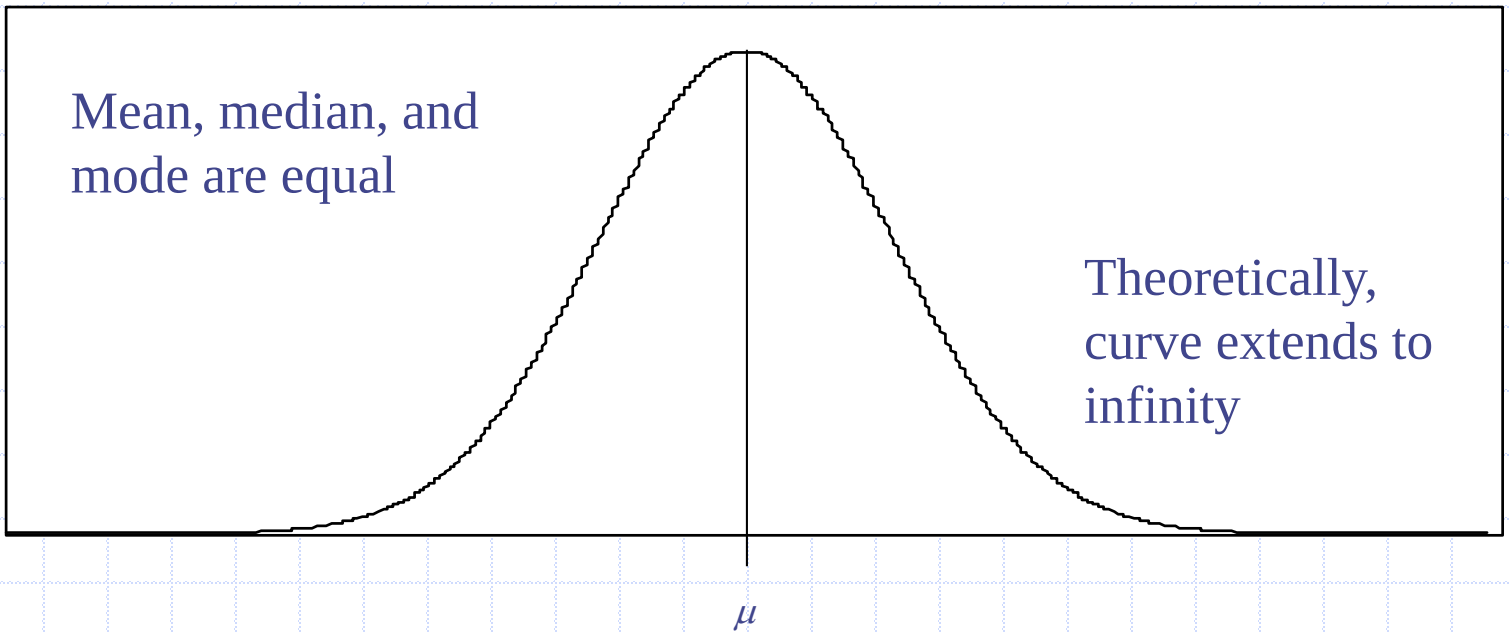
Mean & Variance of a Probability Dist.

- The mean (*expected value*)
  - The long-run average value of the random variable
  - $\mu = \Sigma[xP(x)] = E[x]$
    - where μ is the mean and $P(x)$ is the probability of the various outcomes x.
- The variance
  - The amount of spread (variation) of a distribution
  - $\sigma^2 = \Sigma[(x-\mu)^2 P(x)]$

    $= E[(x-\mu)^2] = Var[x]$

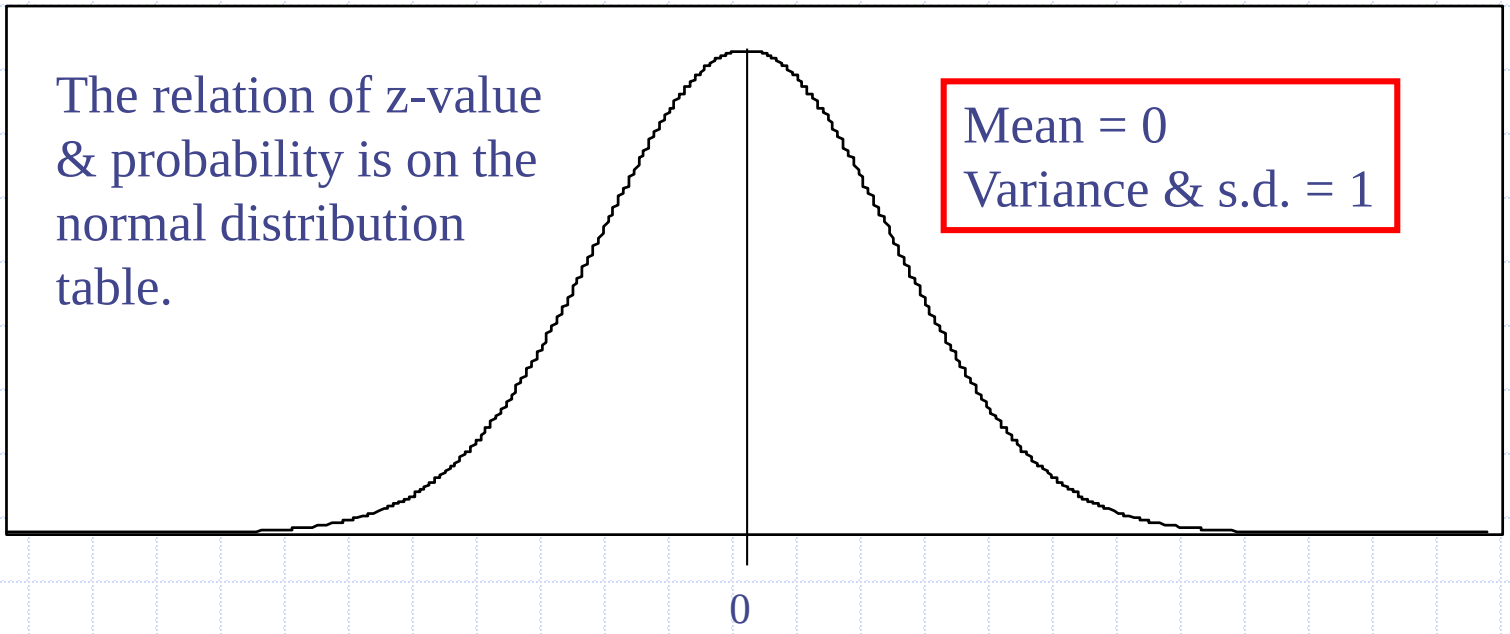
Characteristics of a Normal Dist.

- Bell-shaped
- Symmetrical
- Asymptotic



The Standard Normal Distribution

- Standard normal distribution
  - The normal dist. with mean 0 & variance 1
  - *z-value*: standardized index, $z = \frac{X - \mu}{\sigma}$



Example 1

The monthly starting salaries of recent MBA graduates follows the normal distribution with a mean of $2,000 and a standard deviation of $200. What is the ***z-value*** for a salary of $2,200?

$$z = \frac{X - \mu}{\sigma} = \frac{\$2{,}200 - \$2{,}000}{\$200} = 1.00$$

What is the z-value corresponding to $1,600?

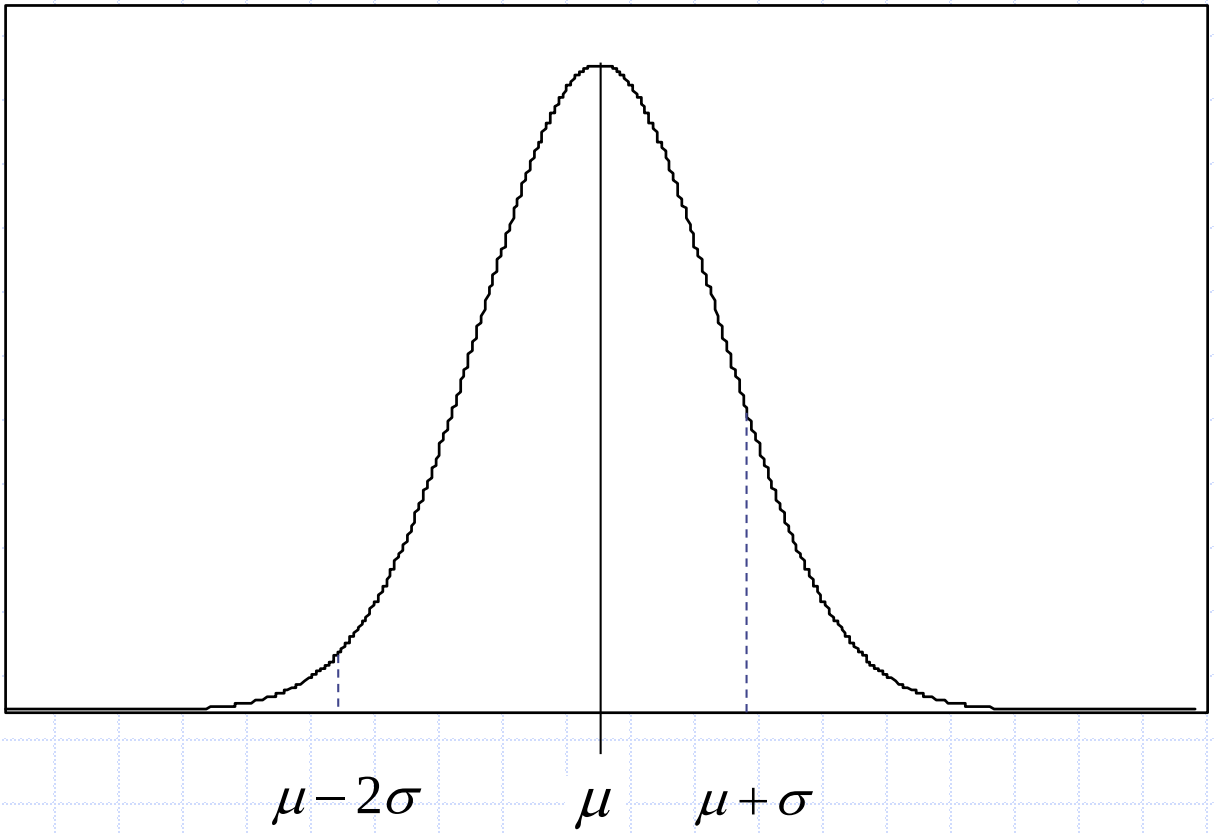
$$z = \frac{X - \mu}{\sigma} = \frac{\$1{,}600 - \$2{,}000}{\$200} = -2.0$$

What % of the population below these salaries?

Areas under the Normal Curve

- A rule of thumb
  - About 68 percent of the area under the normal curve is within one standard deviation of the mean. ($\mu \pm \sigma$)
  - About 95 percent is within two standard deviations of the mean. ($\mu \pm 2\sigma$)
  - Practically all is within three standard deviations of the mean. ($\mu \pm 3\sigma$)

Areas under the Normal Curve



Between:

$\pm 1\sigma$ - 68.26%

$\pm 2\sigma$ - 95.44%

$\pm 3\sigma$ - 99.74%

Example 2

The daily water usage per person in a city follows a normal distribution with a mean of 20 gallons and a standard deviation of 5 gallons. About 68 percent of those living in a city will use how many gallons of water?

About 68% of the daily water usage will lie between 15 and 25 gallons since 20 ± 5.

Example 3

What is the probability that a person from this city selected at random will use between 20 and 25 gallons per day?

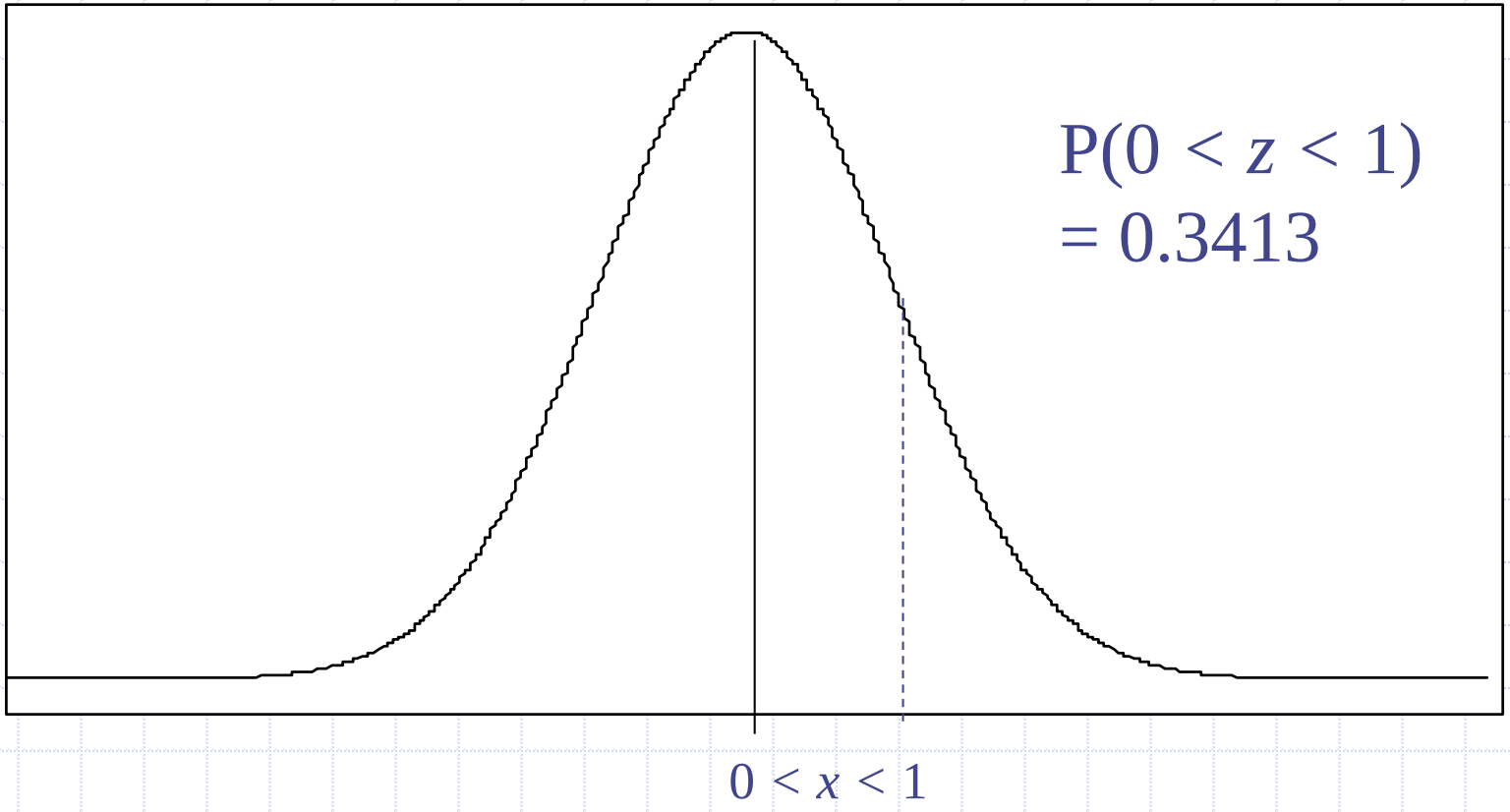
$$z = \frac{X - \mu}{\sigma} = \frac{20 - 20}{5} = 0.00$$

$$z = \frac{X - \mu}{\sigma} = \frac{25 - 20}{5} = 1$$

Example 3 *continued*

- The area under a normal curve between a *z*-value of 0 and a *z*-value of 1 is 0.3413 (=0.8413-0.5)
- We conclude that 34.13% of the residents use between 20 and 25 gallons of water per day.
- See the following diagram.

EXAMPLE 3 *continued*



EXAMPLE 3 *continued*

- What percent of the population use between 15 and 30 gallons per day?

$$z = \frac{X - \mu}{\sigma} = \frac{15 - 20}{5} = -1$$

$$z = \frac{X - \mu}{\sigma} = \frac{30 - 20}{5} = 2$$

Example 3 *continued*

- The area below the z-value of –1 is 0.9773.
- The area below the z-value of 2 is 0.1587.
- Since 0.9773 – 0.1587, the result is 0.8186.
- We conclude that 82% of the residents use between 15 and 30 gallons of water per day.

Basic Statistics 03

Sampling & Central Limit Theorem

Sampling & C.L.T.

1. The distribution of the sample mean
2. The Central Limit Theorem
3. The application of CLT

1. Distribution of the Sample Means

The ***sampling distribution of the sample mean*** is a probability distribution consisting of all possible sample means of a given sample size selected from a population.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{n\mu}{n} = \mu$$

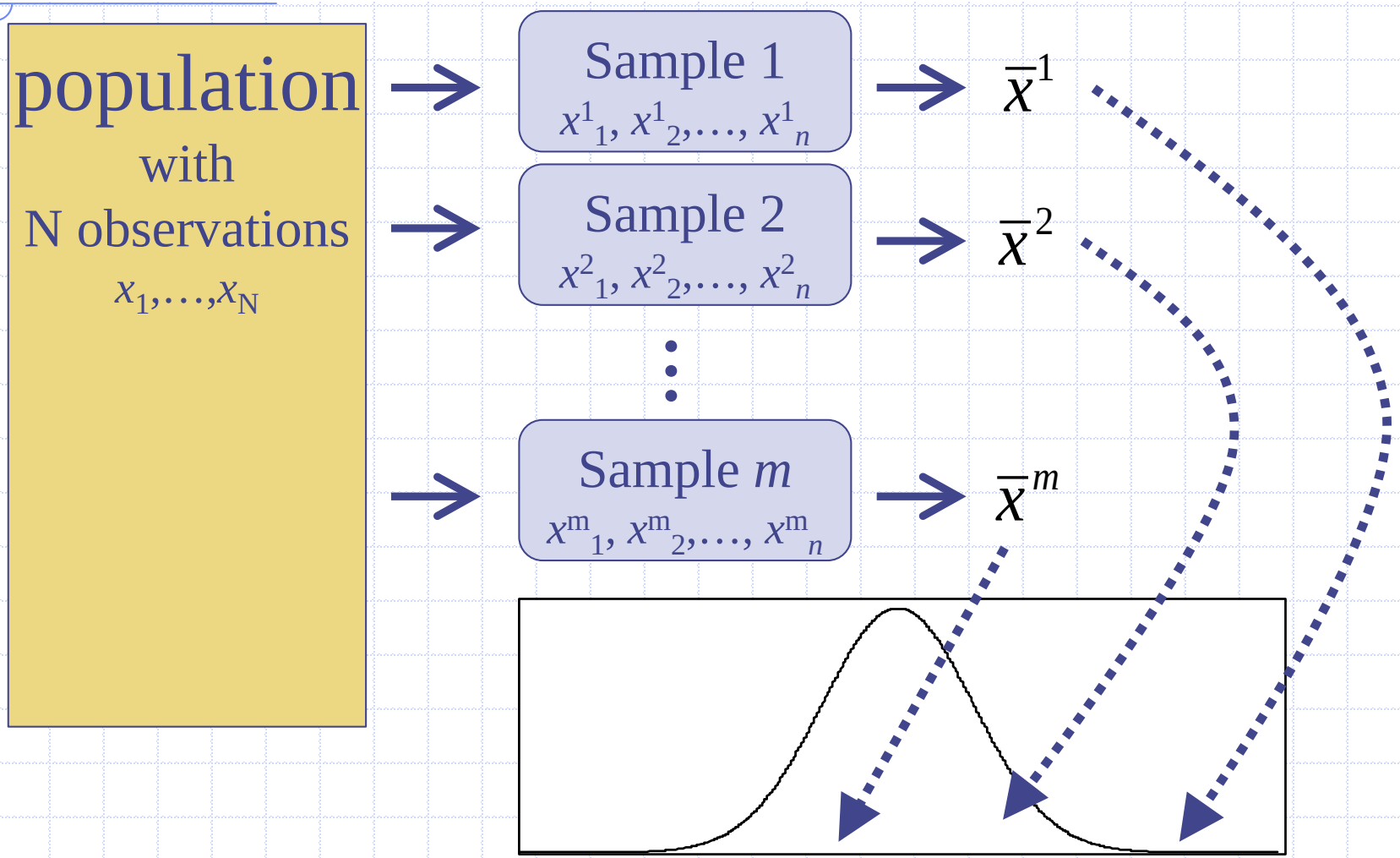
Population mean

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

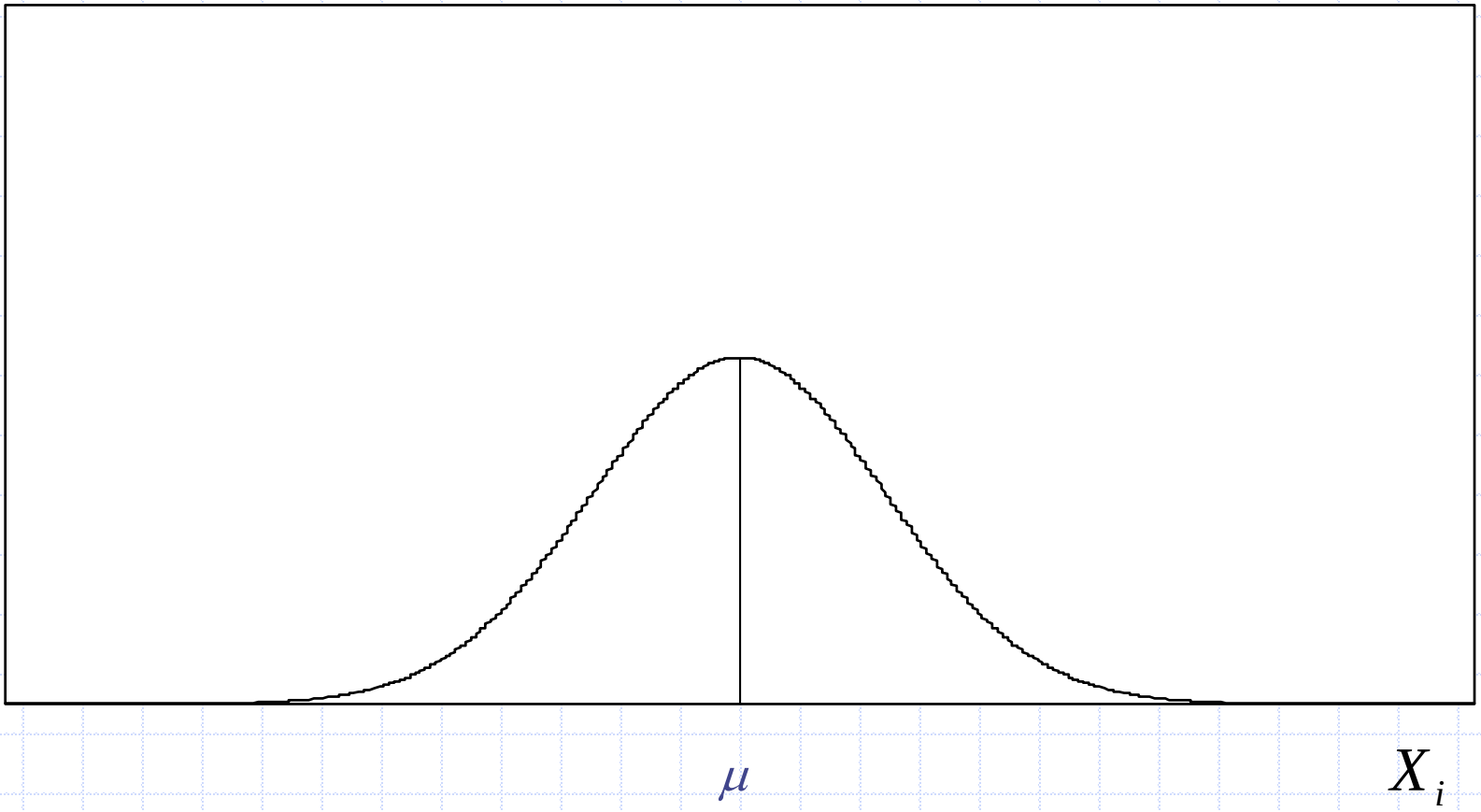
Population variance

*in the case of finite population, $Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

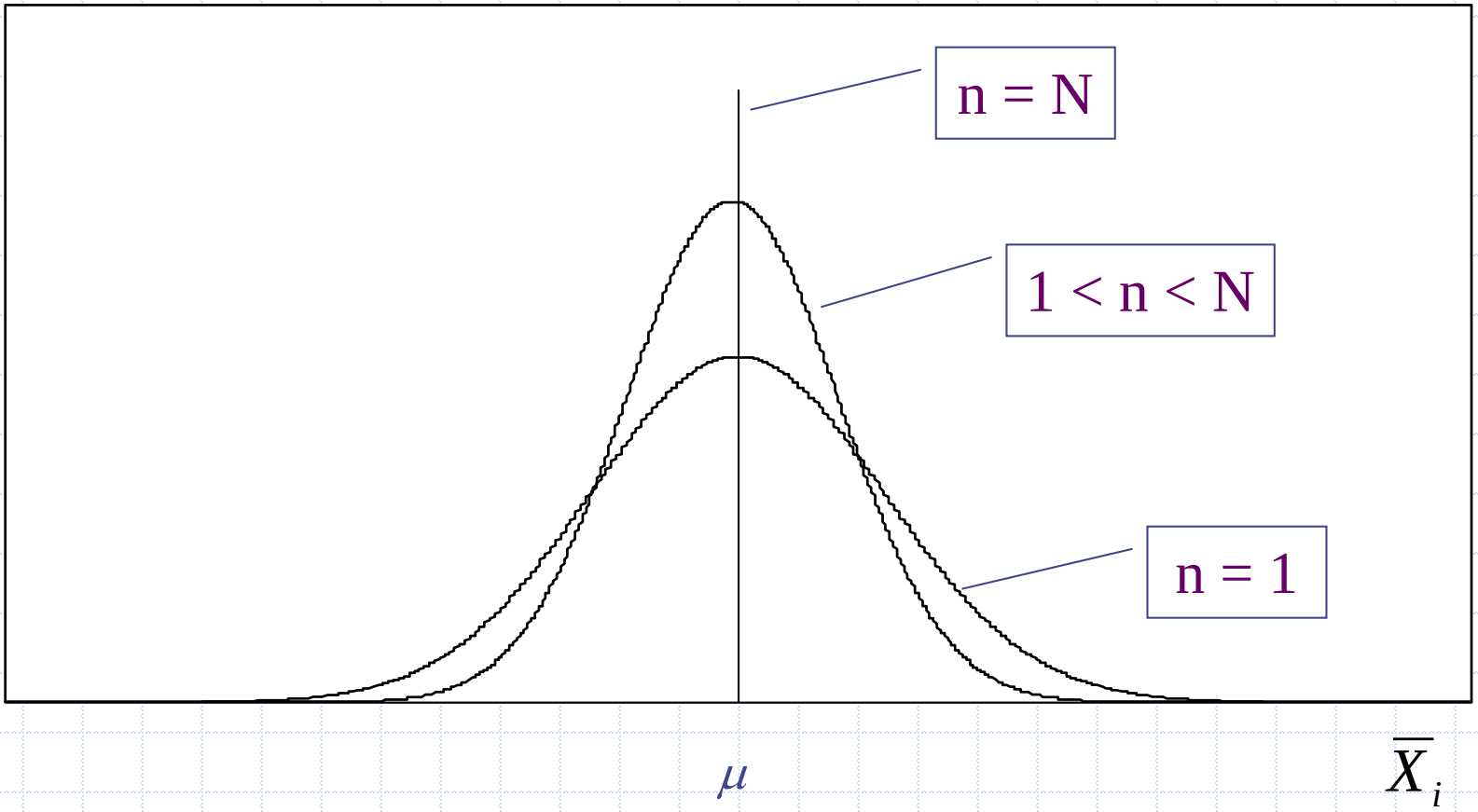
The distribution of sample mean



Population Distribution



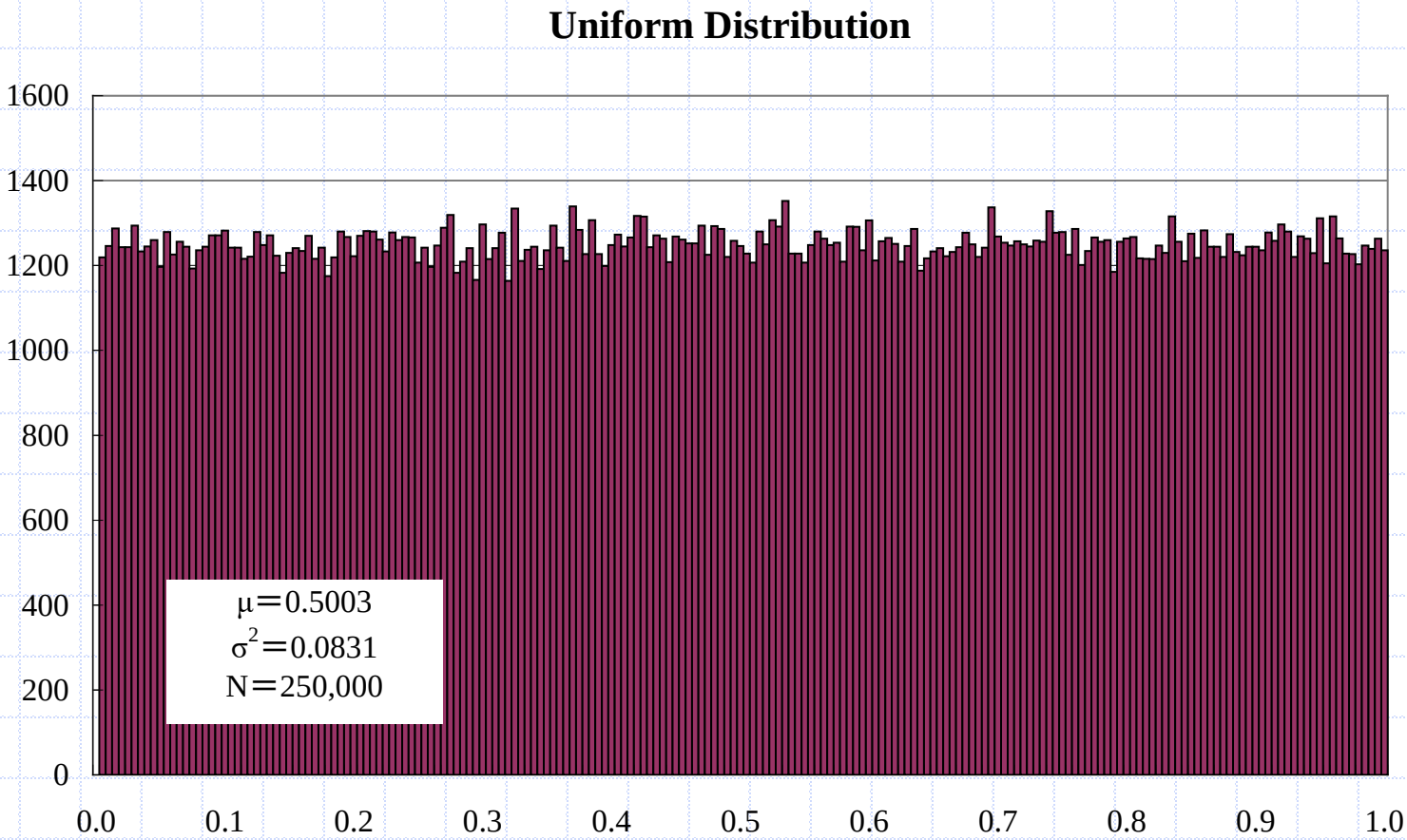
Sampling Dist. of the Sample Means



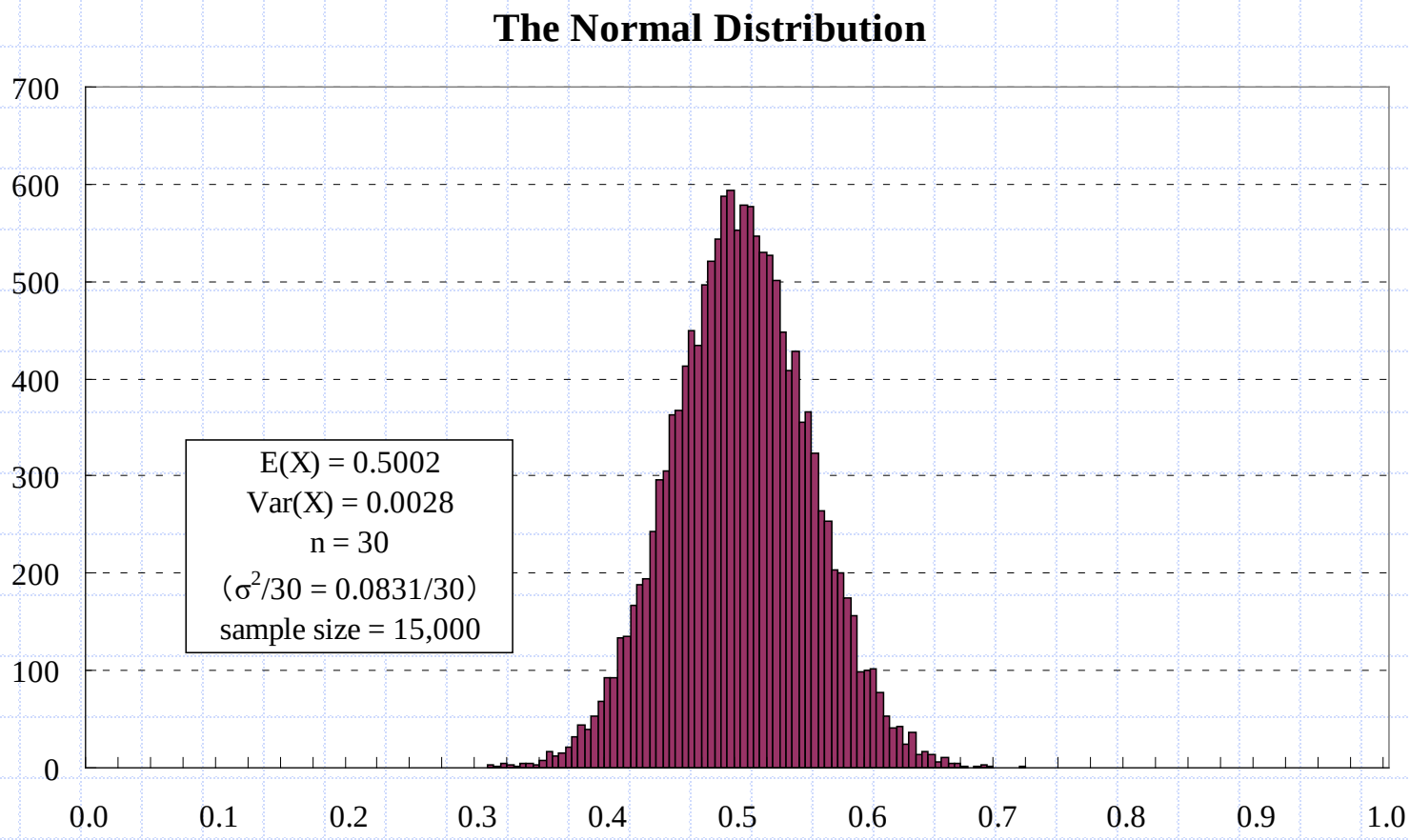
2. Central Limit Theorem

- For ***any*** population with a mean μ and a variance σ^2, the ***sampling distribution of the means*** of all possible samples of size n will be approximately normally distributed, with larger sample size n.
- The mean of the sampling distribution equal to μ and the variance equal to σ^2/n.

CLT *The Population Distribution*



CLT *The Sample Distribution of Sample mean*



CLT 0 *the case of $X \sim N(\mu, \sigma^2)$*

- If ***a population follows the normal distribution***, the sampling distribution of the sample mean will also follow the normal distribution for ***any sample size***.
- To determine the probability a sample mean falls within a particular region, use:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

CLT 1 *the case of X~ non-normal dist.*(μ, σ^2)

- If the population ***isn't normally distributed*** (with known σ^2) and ***sample size is large***, the sample means will follow the normal distribution. (See the above figure.)

$$\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n}) \qquad \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0,1)$$

CLT 2 *the case of $X \sim N(\mu$, unknown $\sigma^2)$*

- If ***the population follows the normal distribution but σ^2 is unknown***, the sample means will follow the ***t*** distribution.
  - But ***with larger sample size*** (*at least n >30*), the sample means will follow the normal distribution.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(d.f.) \qquad t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \xrightarrow{d} N(0,1)$$

CLT 3 *the case of X~ non-N (μ, unknown σ^2)*

- If the population ***isn't normally distributed with unknown σ^2 and sample size is large***, the sample means will follow the ***t*** distribution. (with larger sample size, the normal distribution)

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Example 1

- The mean selling price of a gallon of gasoline in the United States is $1.30.
- The distribution is positively *skewed*, with a standard deviation of $0.28.
- What is the probability of selecting a sample of 35 gasoline stations and finding the sample mean within $0.08?

Example 1 *continued*

- The first step is to find the *z*-values corresponding to \$1.22 and \$1.38 (= 1.30 ± 0.08). These are the two points within \$0.08 of the population mean.

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\$1.38 - \$1.30}{\$0.28 / \sqrt{35}} = 1.69$$

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\$1.22 - \$1.30}{\$0.28 / \sqrt{35}} = -1.69$$

Example 1 *continued*

- Next we determine the probability of a *z*-value between -1.69 and 1.69. It is:

$$P(-1.69 \le z \le 1.69) = 2(.4545) = .9090$$

We would expect about 91 percent of the sample means to be within \$0.08 of the population mean.