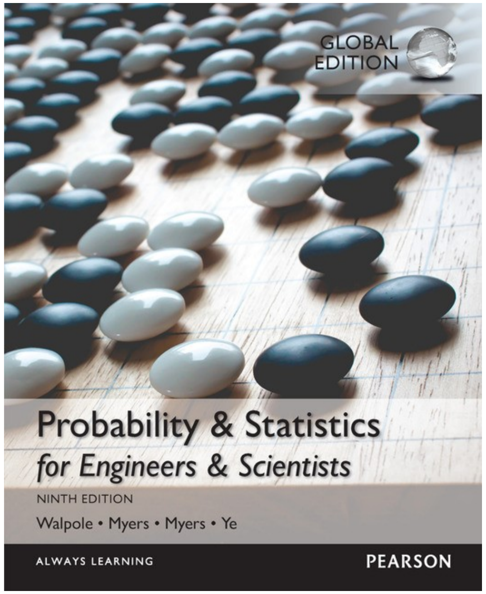
# Chapter 8

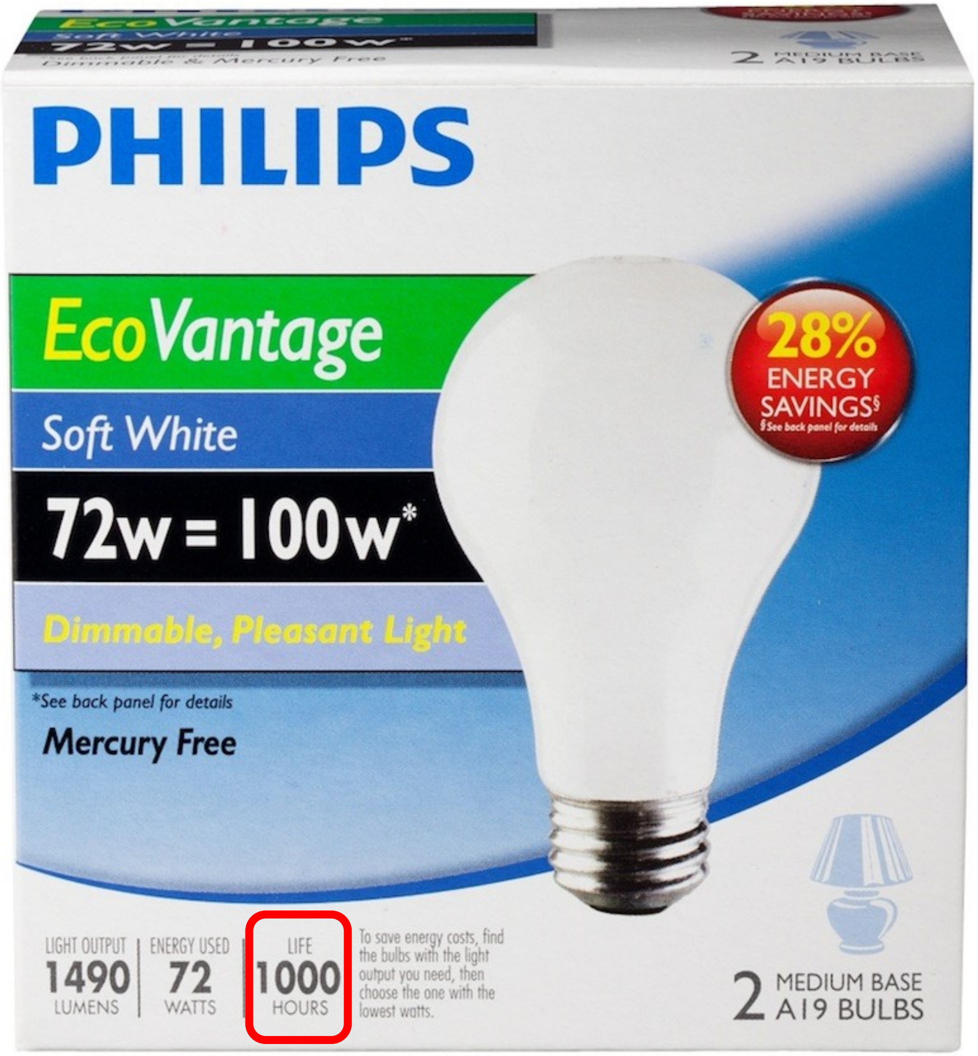
# Fundamental Sampling Distributions and Data Descriptions (1)



# Outline

- Introduction to
  random sampling and statistical inference
- Populations and samples
- Sampling distribution of means
  - Central Limit Theorem
- Other distributions
  - $S^2$
  - t-distribution
  - F-distribution

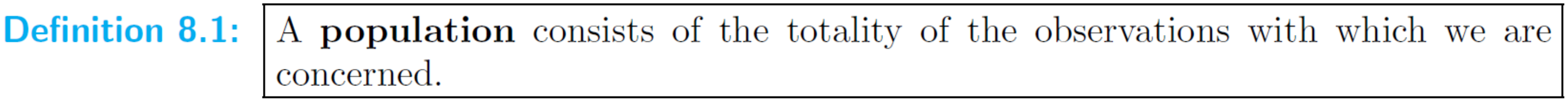
# How can we believe this ?



# Sampling

- Statistics **deal with data,**
  **but where does it come from?**

# Population



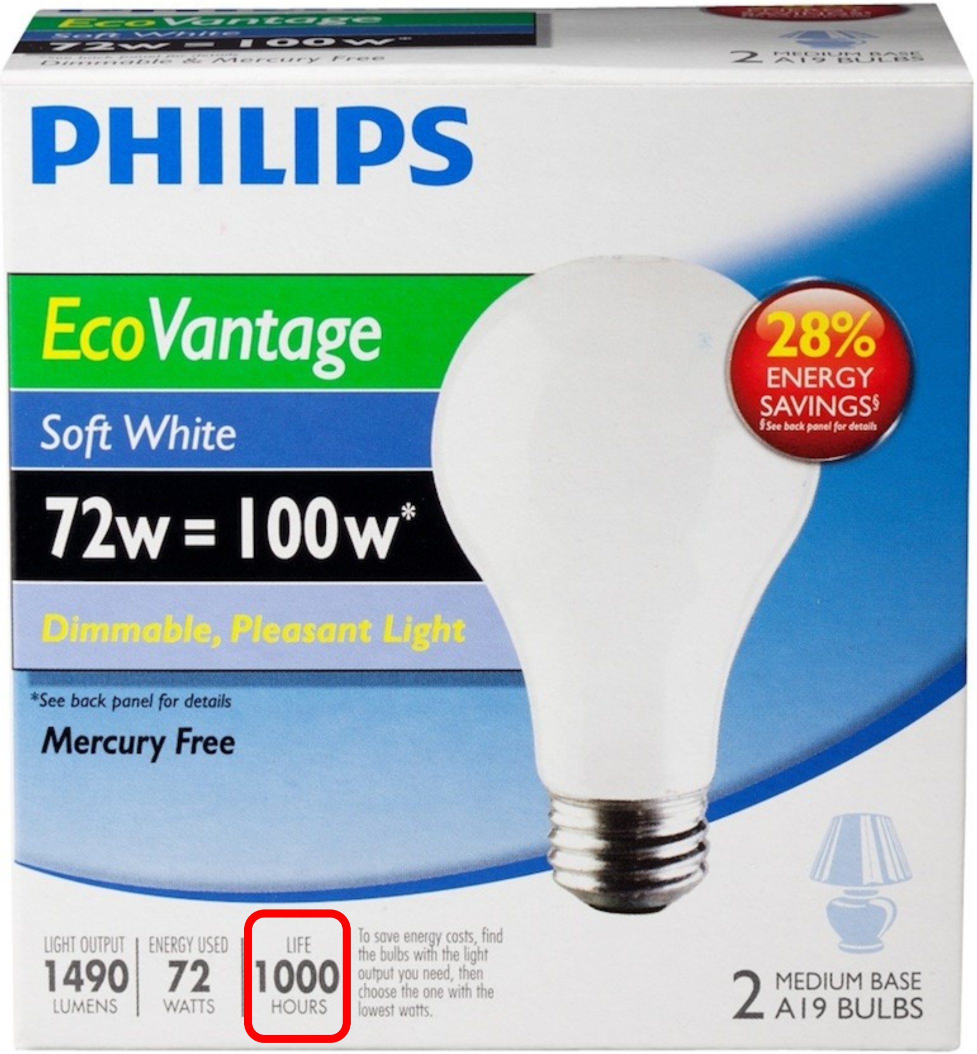
**Definition 8.1:** A **population** consists of the totality of the observations with which we are concerned.

- Population
  - the entire group of things that you're trying to measure, study, or analyze
  - "a group of individual persons, objects, or items from which *samples* are taken for statistical measurement"

# Population

- Each observation in a population is a value of a random variable $X$ having some probability distribution $f(x)$.
  - Example: In the blood-type experiment, the random variable $X$ represents the type of blood and is assumed to take on values from 1 to 8 (AB, A, B, or O, each with a plus or minus sign): **multinomial population**
  - Example: The lives of the storage batteries are values assumed by a continuous random variable having perhaps a normal distribution: **normal population**
- Hence, the mean and variance of a random variable or probability distribution are also referred to as **mean and variance of the population**

# How can we believe this ?

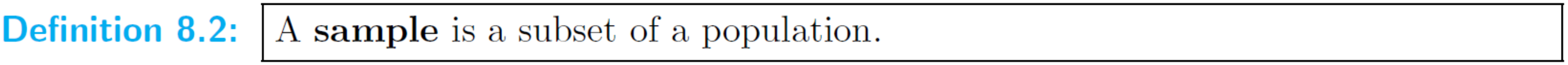


- Population
  - test all bulbs' lifetime
    - Impossible: no bulbs left to sell
    - Too much cost
- Sample
  - Test a **subset** of the observations from the entire population

# Examples

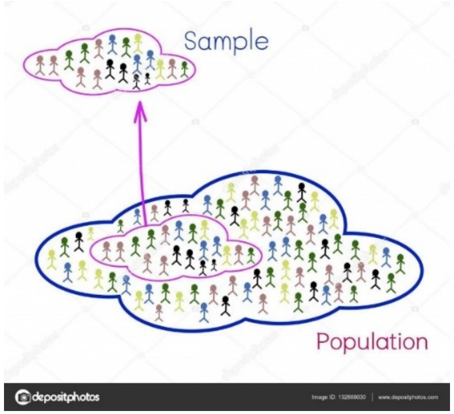
| Population | Sample |
|---|---|
| Students pursuing undergraduate engineering degrees | 1000 engineering students selected at random from all engineering programs in the US |
| Cars capable of speeds in excess of 160 mph. | 50 cars selected at random from among those certified as having achieved 160 mph or more during 2003 |

# Populations and Samples



**Definition 8.2:** A **sample** is a subset of a population.

- Sample
  - a selection of items taken from a population
  - “a finite part of a statistical population whose properties are studied to gain information about the whole”



# Question

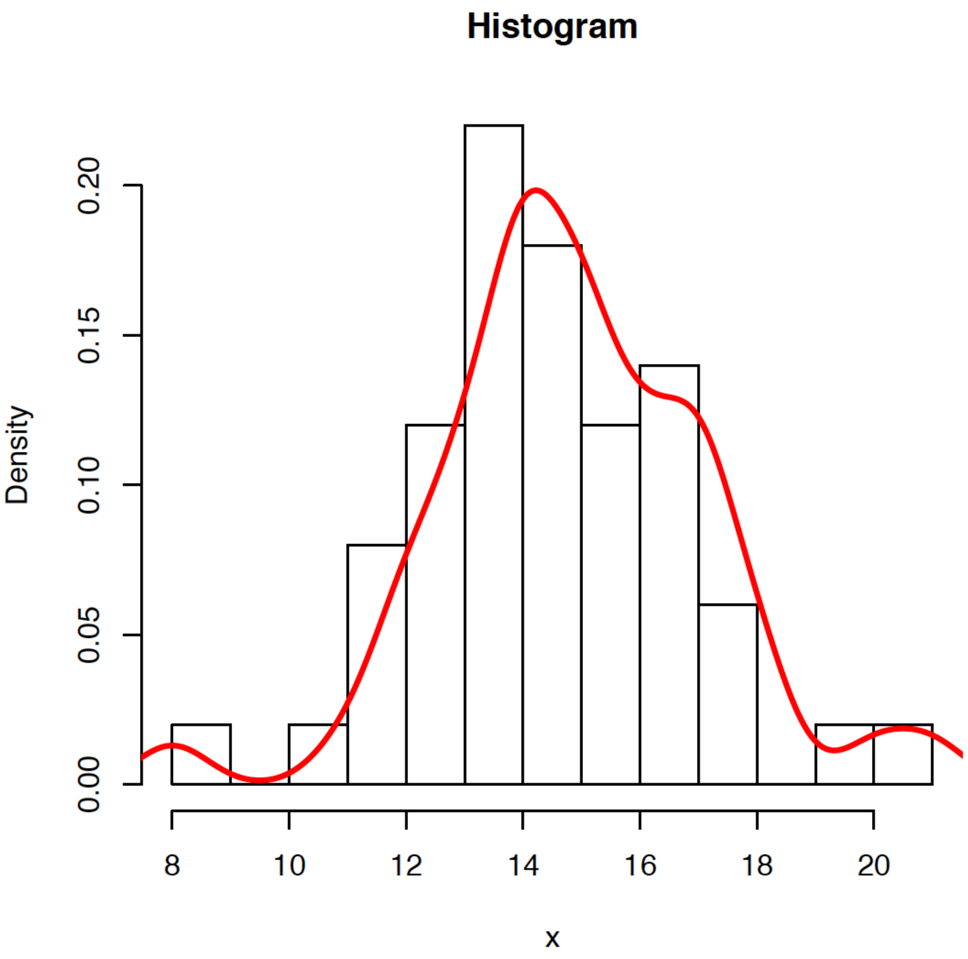
- Traveling a university in a city via shuttle bus takes, on average, 15 minutes with a standard deviation of 3 minutes.
- In a given week, a bus transported passengers 50 times. What is the probability that the average transport time was more than 18 minutes?

# Sample Distribution

- Sample distribution
  - the distribution resulting from the collection of actual data
  - example:
    - 15 14 15 18 15 20 15 16 17 14 17 13 11 14 18 12 17 12 21 8 14 17 14 12 13 15 15 16 17 14 16 13 14 15 18 16 16 17 14 15 16 15 17 12 14 14 13 13 13 14

→ These numbers constitute a sample distribution

# Sample Distribution

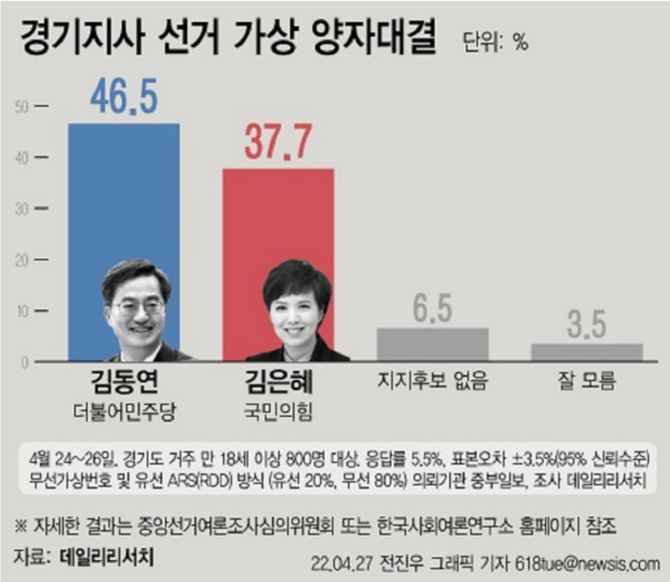


- Characteristic of a sample
  - it contains a finite (countable) number (frequency) of scores, the number of scores represented by the letter n.

In addition to the frequency distribution, the sample distribution can be described with other numbers. (next page)

- In addition to the frequency distribution, the sample distribution can be described with numbers, called statistics.

- Examples of statistics (We will study soon)
  - the mean,
  - median,
  - mode,
  - standard deviation,
  - range, and
  - correlation coefficient, among others.


경기지사 선거 가상 양자대결

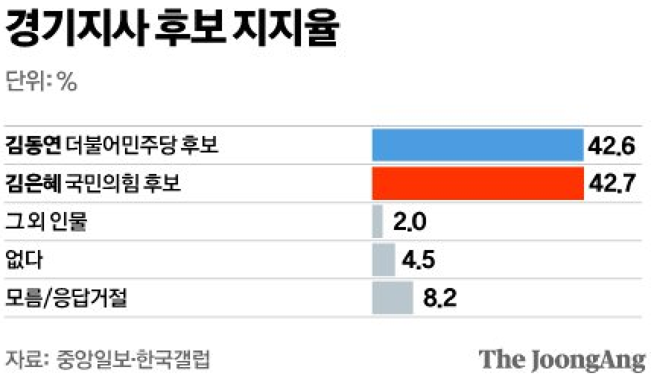
4월 24~26일. 경기도 거주 만 18세 이상 800명 대상. 응답률 5.5%, 표본오차 ±3.5%(95% 신뢰수준) 무선가상번호 및 유선 ARS(RDD) 방식 (유선 20%, 무선 80%) 의뢰기관 중부일보, 조사 데일리리서치

※ 자세한 결과는 중앙선거여론조사심의위원회 또는 한국사회여론연구소 홈페이지 참조

자료: 데일리리서치

22.04.27 전진우 그래픽 기자 618tue@newsis.com

http://www.wolyo.co.kr/news/articleView.html?idxno=204410


경기지사 후보 지지율

자료: 중앙일보·한국갤럽

https://www.joongang.co.kr/article/25068017#home

- Results are from random sampling ?

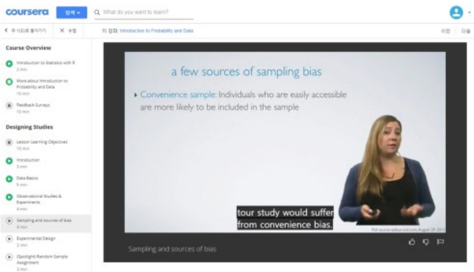
- If a different sample was taken, different scores would result

- However, there would also be some consistency in that while the statistics would not be exactly the same, they would be similar.

- To achieve order in this chaos, statisticians have developed probability models for sampling.

# 8.1 Random Sampling

# Random Sampling


coursera
탐색
What do you want to learn?
이 강좌: Introduction to Probability and Data
이전
다음
Course Overview
Introduction to Statistics with R
2 min
More about Introduction to Probability and Data
10 min
Feedback Surveys
10 min
Designing Studies
Lesson Learning Objectives
10 min
Introduction
3 min
Data Basics
5 min
Observational Studies & Experiments
4 min
Sampling and sources of bias
8 min
Experimental Design
2 min
(Spotlight) Random Sample Assignment
3 min

a few sources of sampling bias

- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample

tour study would suffer from convenience bias.

Sampling and sources of bias

- To eliminate bias in the sampling procedure, we select a random sample in the sense that the observations are made independently and at random.
  - every set of *n* individuals has an equal chance to be the sample actually selected.

- It consists of *n* observations selected independently and randomly from the population.

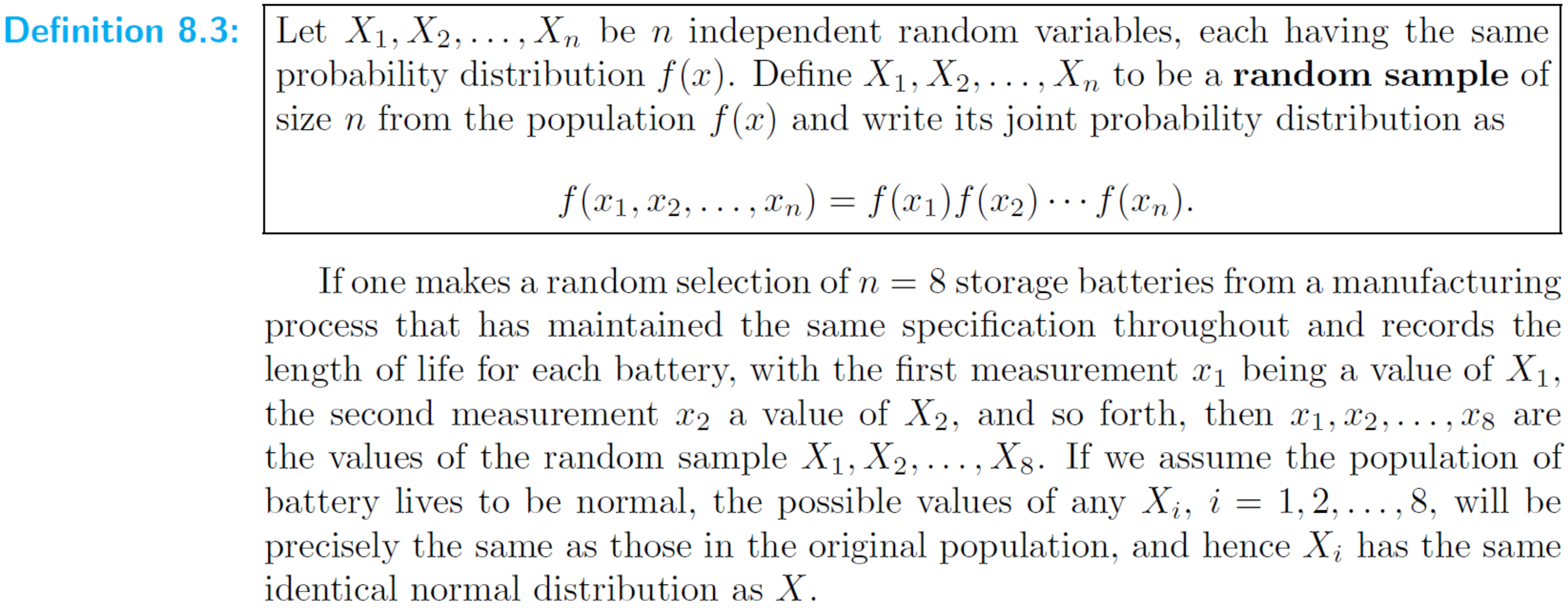
# Random Sampling

- Each observation in a population is a value of a random variable ***X*** having some probability distribution ***f*(*x*)**.

- **Def. 8.3: Random Sampling**
  - Let $X_1$, $X_2$, …, $X_n$ be *n* **independent** random variables, each having the same probability distribution ***f*(*x*)**

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

# Definition 8.3



**Definition 8.3:** Let $X_1, X_2, \ldots, X_n$ be $n$ independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size $n$ from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

If one makes a random selection of $n = 8$ storage batteries from a manufacturing process that has maintained the same specification throughout and records the length of life for each battery, with the first measurement $x_1$ being a value of $X_1$, the second measurement $x_2$ a value of $X_2$, and so forth, then $x_1, x_2, \ldots, x_8$ are the values of the random sample $X_1, X_2, \ldots, X_8$. If we assume the population of battery lives to be normal, the possible values of any $X_i$, $i = 1, 2, \ldots, 8$, will be precisely the same as those in the original population, and hence $X_i$ has the same identical normal distribution as $X$.

# 8.2 Some Important Statistics

# Statistic

- Question: What is the proportion of coffee-drinkers in the United States who prefer a certain brand of coffee?
  - It would be impossible to question every coffee drinking American in order to compute the value of the parameter $\boldsymbol{p}$ representing the **population** proportion.
  - Instead, a large random **sample** is selected and the proportion $\hat{\boldsymbol{p}}$ of people in this sample favoring the brand of coffee in question is calculated. The value $\hat{\boldsymbol{p}}$ is now used to make an ***inference*** concerning the true proportion $\boldsymbol{p}$.
- Now, $\hat{\boldsymbol{p}}$ is a function of the observed values in the random sample; since many random samples are possible from the same population, we would expect $\hat{\boldsymbol{p}}$ to vary somewhat from sample to sample.
- That is, $\hat{\boldsymbol{p}}$ is a value of a **random variable** that we represent by $\boldsymbol{P}$. Such a random variable is called a ***statistic***.

# Statistic

- ***Def. 8.4:*** statistic
  - Any function of the random sample $X_1$, $X_2$, …, $X_n$ is called a statistic.

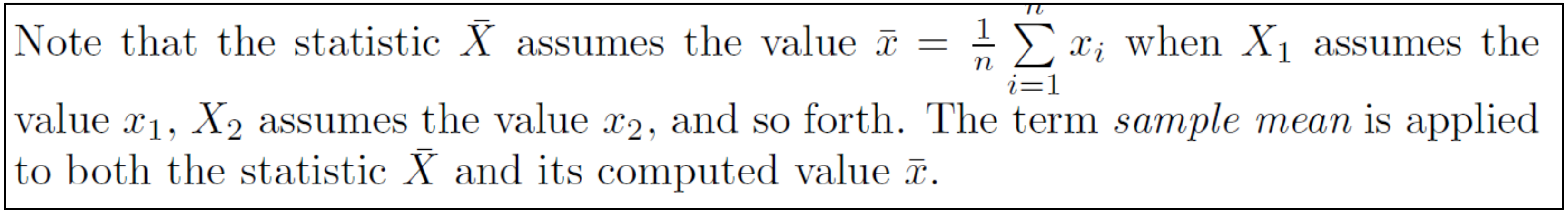
- Statistic examples
  - sample mean
  - sample variance
  - median
  - mode
  - sample standard deviation
  - sample range
  - …

# Statistic : Sample Mean

- Sample Mean
  - If $X_1$, $X_2$, …, $X_n$ represents a random sample of size *n* (*n* random variables), then the sample mean is defined to be the statistic:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{(unit)}$$

$\overline{X}$ is a statistic because it is a function of the random sample $X_1$, $X_2$, …, $X_n$ !

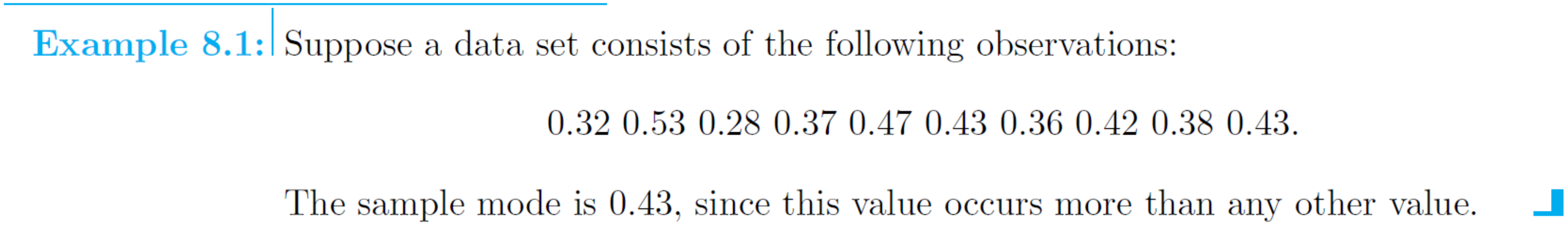


Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ when $X_1$ assumes the value $x_1$, $X_2$ assumes the value $x_2$, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.

# Median and Mode

- **Median** (중간값)
  - The median is the middle value in your list. When the totals of the list are odd, the median is the middle entry in the list after sorting the list into increasing order.
- Mode
  - The mode in a list of numbers refers to the list of numbers that occur most frequently.

# Example 8.1



**Example 8.1:** Suppose a data set consists of the following observations:

$$0.32\ 0.53\ 0.28\ 0.37\ 0.47\ 0.43\ 0.36\ 0.42\ 0.38\ 0.43.$$

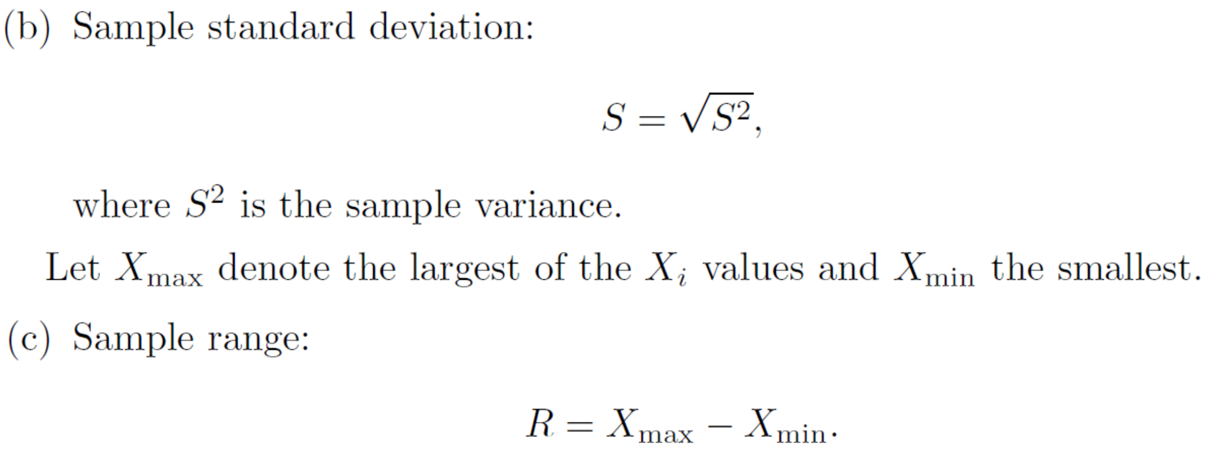
The sample mode is 0.43, since this value occurs more than any other value. ▟

- Sample Variance **$S^2$**
  - If $X_1, X_2, \ldots, X_n$ represents a random sample of size $n$, then the sample variance is defined to be the statistic:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \qquad (\text{unit})^2$$

*S*$^2$ is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$ !

- NOTE: *S*$^2$ measures the variability in the sample.



(b) Sample standard deviation:

$$S = \sqrt{S^2},$$

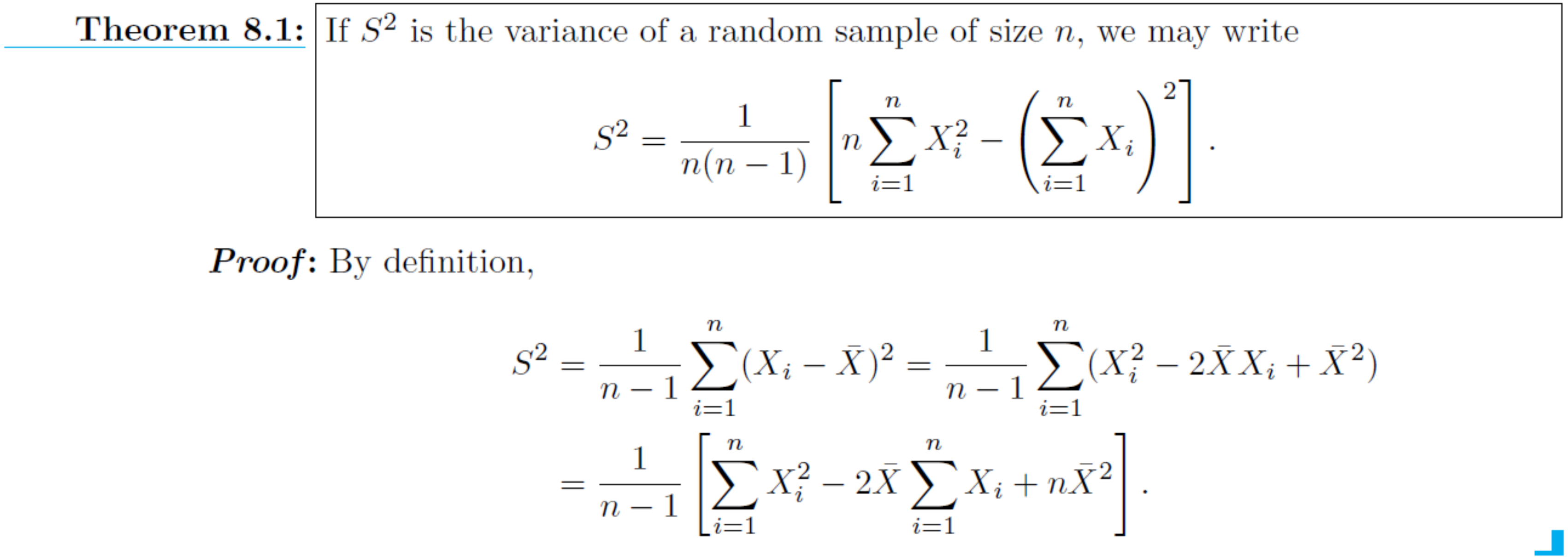
where $S^2$ is the sample variance.

Let $X_{\max}$ denote the largest of the $X_i$ values and $X_{\min}$ the smallest.

(c) Sample range:

$$R = X_{\max} - X_{\min}.$$

# Theorem 8.1



**Theorem 8.1:** If $S^2$ is the variance of a random sample of size $n$, we may write

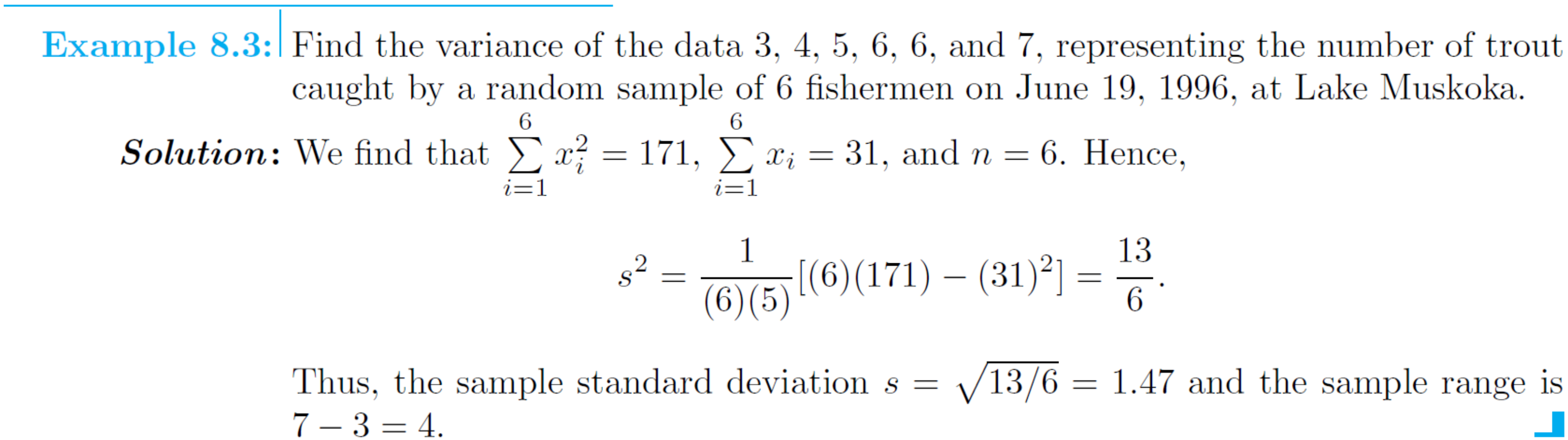
$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

***Proof*:** By definition,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right].$$

# Example 8.3

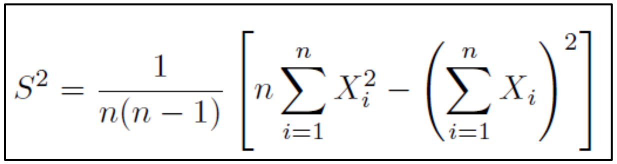


**Example 8.3:** Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen on June 19, 1996, at Lake Muskoka.

***Solution*:** We find that $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$, and $n = 6$. Hence,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Thus, the sample standard deviation $s = \sqrt{13/6} = 1.47$ and the sample range is $7 - 3 = 4$.



$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]$$

# 8.3 Sampling Distributions

# Question

- Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes.
- In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes?

# Statistical Inference

- Statistical inference
  - From samples, make various statements concerning values of the population parameters (generalization and prediction)

- Statistical inference: Example
  - Based on the opinions of several people interviewed on the street, that in a forthcoming election 60% of the eligible voters in the city of Detroit favor a certain candidate.
  - In this case, we are dealing with a random sample of opinions from a very large finite population

# Statistical inference: Example

The company official made the decision that the soft-drink machine dispenses drinks with an average content of 240ml, even though the sample mean was 236ml, because he knows from sampling theory that, if μ = 240ml, such a sample value could easily occur.

If a different sample was taken, different scores would result.

In fact, if he ran similar tests, say every hour, he would expect the values of the statistic $\bar{x}$ to fluctuate above and below μ = 240 milliliters.

Only when the value of $\bar{x}$ is substantially different from 240 millimeters will the company official imitate action to adjust the machine.

# Sampling distribution

- If we conduct the same experiment several times with the same sample size, the probability distribution of the resulting statistic (e.g., **mean**, **variance**) is called a *sampling distribution.*
  - In the previous example, the values of the statistic $\bar{X}$ to fluctuate above and below μ = 240 milliliters.

# Sampling distribution

- ***Def. 8.5:*** The probability distribution of a statistic is called a sampling distribution.

- The probability distribution of $\bar{X}$ is called the sampling distribution of the mean.
  - Example:
    - if $n$ observations are taken from a normal population with mean $\mu$ and variance $\sigma^2$ and the experiment is conducted over and over (always with sample size **n**),
    - the many values of $\bar{X}$ result → the distribution of $\bar{X}$

# 8.4 Sampling Distribution of Means and the Central Limit Theorem

# Sampling Distributions of Means

- If $X_1, X_2, \ldots, X_n$ is a random sample of size $n$ taken from a normal population with mean $\mu$ and variance $\sigma^2$, i.e. N($\mu$,$\sigma$)
- Each observation $\boldsymbol{X_i}, i = 1,2, \ldots, n$, of the random sample will then have the same normal distribution as the population being sampled
- Then the sample mean $\bar{X}$ has a normal distribution

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

**Theorem 7.11**

- The sum of independent normal random variables is also normal random variable with

  mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$

# Sampling Distributions of Means

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- The sample mean $\bar{X}$ has a normal distribution with mean and variance:

$$\mu_{\overline{x}} = \frac{\mu + \mu + \mu + \ldots + \mu}{n} = \mu$$

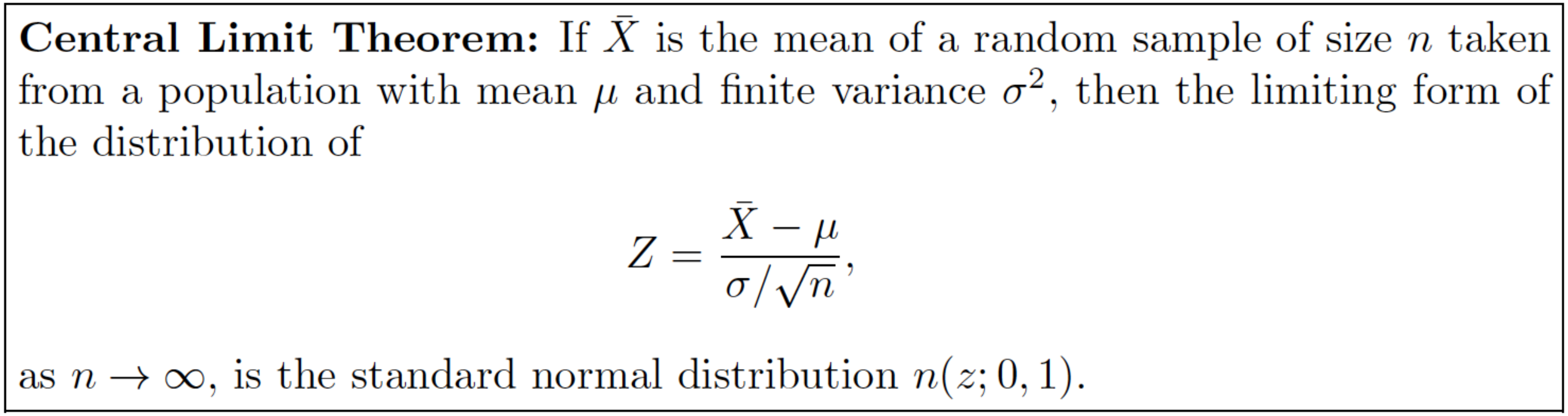
$$\mathrm{E}(\overline{\mathrm{X}}) = \mu_{\overline{\mathrm{X}}} = \mu$$

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2 + \sigma^2 + \sigma^2 + \ldots \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Proof

# Central Limit Theorem

- If we are sampling from a population with unknown distribution, either finite or infinite, the sampling distribution of $\bar{X}$ will still have an approximately **normal** distribution with mean $\mu$ and variance $\sigma^2$, provided that the sample size ***n*** is large ($n \geq 30$).
- **Theorem 8.2 :**



**Central Limit Theorem:** If $\bar{X}$ is the mean of a random sample of size $n$ taken from a population with mean $\mu$ and finite variance $\sigma^2$, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

# Central Limit Theorem

- The normal approximation for $\bar{X}$ will generally be good if n ≥ 30, provided the population distribution is not terribly skewed.
- If n < 30, the approximation is good only if the population is not too different from a normal distribution (or approximately normal)
- If the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a normal distribution exactly, no matter how small the size of the samples.
  - Sampling distribution of means

$$\mu_{\overline{x}} = \frac{\mu + \mu + \mu + \ldots + \mu}{n} = \mu$$

$$\sigma_{\overline{x}}^{2} = \frac{\sigma^{2} + \sigma^{2} + \sigma^{2} + \ldots \sigma^{2}}{n^{2}} = \frac{\sigma^{2}}{n}$$

- We consider *n* large when *n*≥30.

$$\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

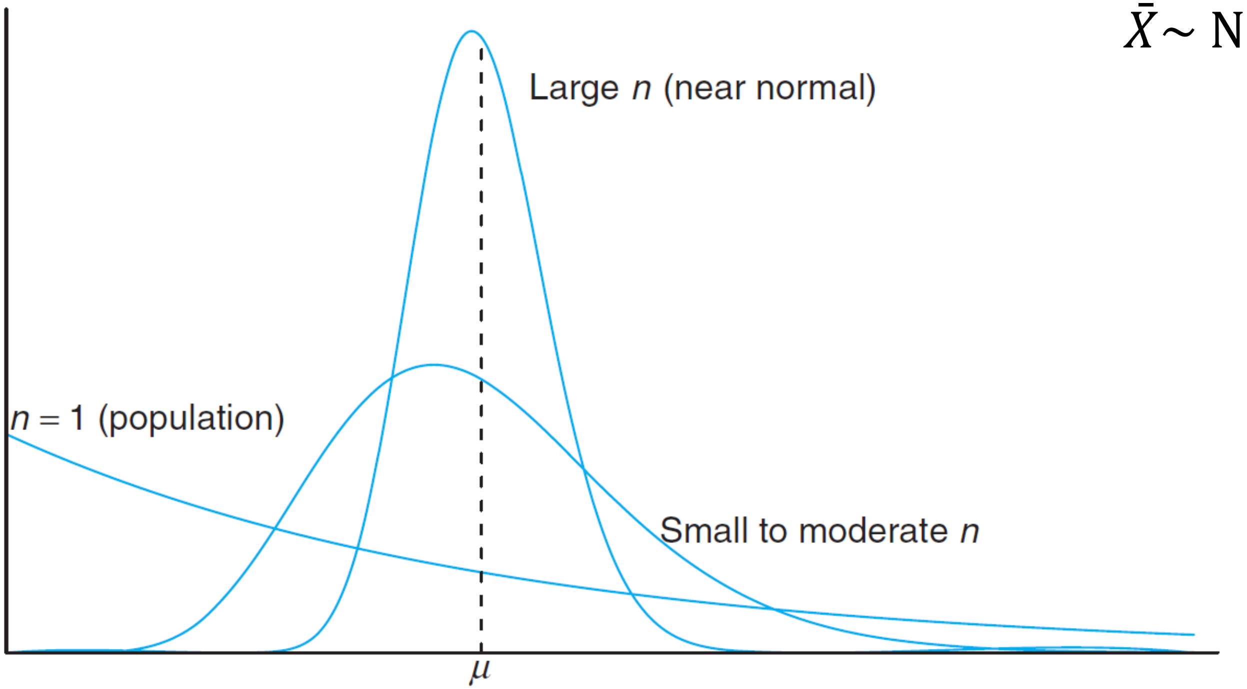


Figure 8.1: Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for $n = 1$, moderate $n$, and large $n$).

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1) \Leftrightarrow \bar{X} \sim \mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$$ for large n

- The sampling distribution of $\overline{X}$ is used for inferences about the population mean $\mu$.

- ***Example 8.13:***
  - An electric firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours.
  - Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

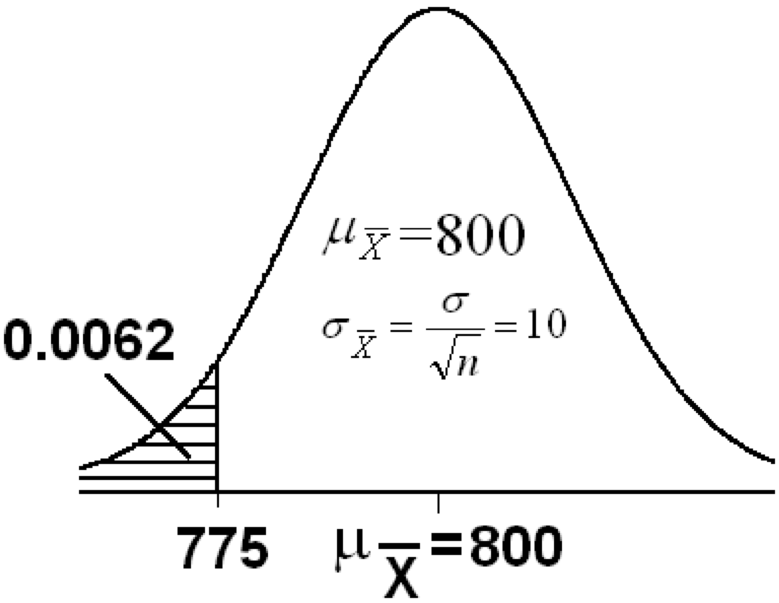
## *Solution:*

- X= the length of life
- $\mu$=800 , $\sigma$=40
- X~N(800, 40)
- *n*=16

$$\mu_{\overline{X}} = \mu = 800$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$$

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(800, 10)$$



- ***Cont.***

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(800,10)$$

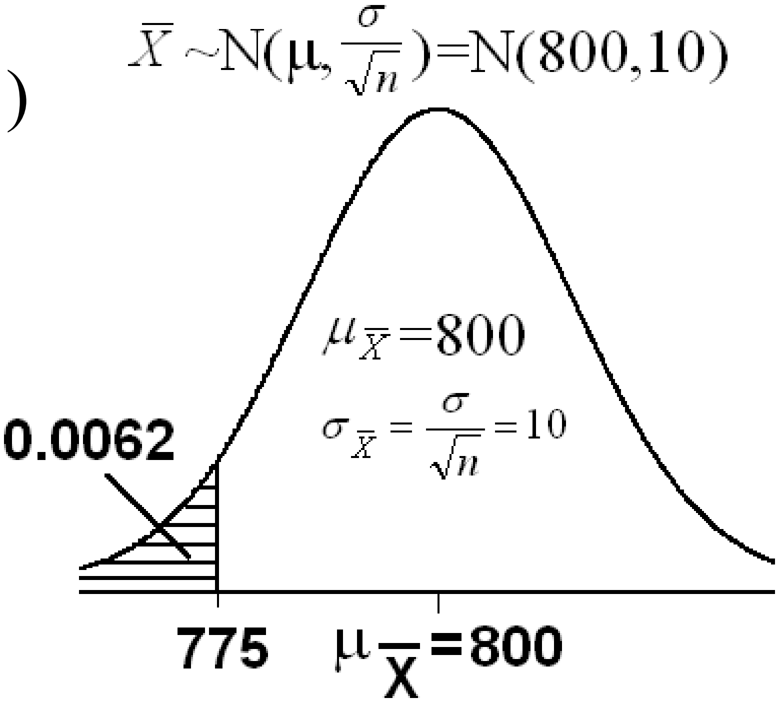
$$\Leftrightarrow Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = Z = \frac{\overline{X} - 800}{10} \sim N(0,1)$$

$$= P\left(\frac{\overline{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

$$= P\left(Z < \frac{775 - 800}{10}\right)$$

$$= P(Z < -2.50)$$

$$= 0.0062$$

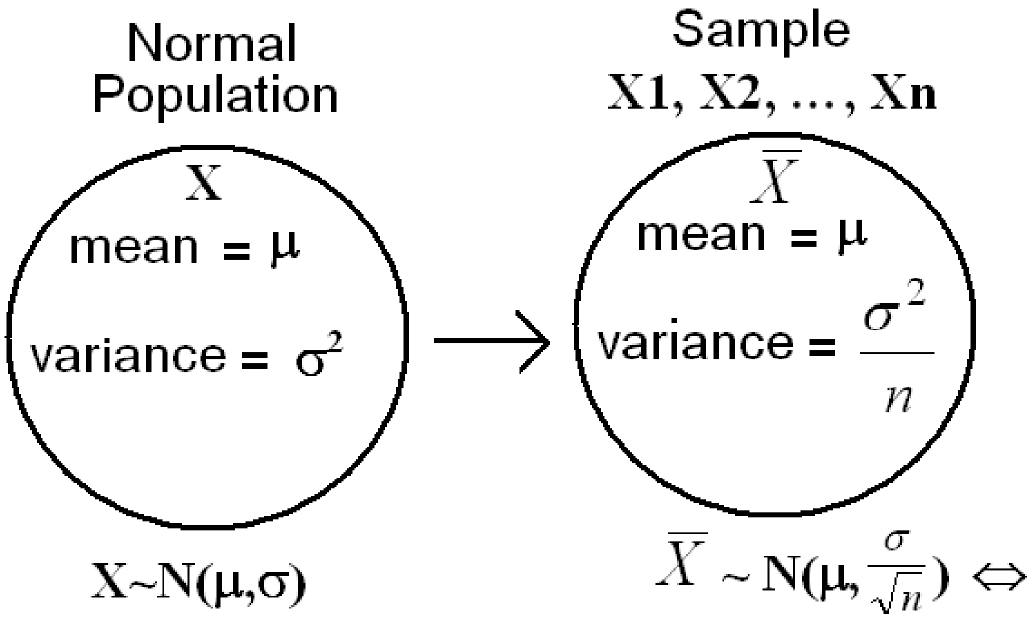


- ***Recall***

·If $X_1$, $X_2$, …, $X_n$ is a random sample of size *n* from N(μ,σ), then

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

~~for large n~~
regardless of size n



$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow$$

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

- Example 8.5
  - Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes.
  - In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

***Solution:*** In this case, $\mu = 28$ and $\sigma = 3$. We need to calculate the probability $P(\bar{X} > 30)$ with $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{x}$ greater than 30 is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An illustrative graph is shown in Figure 8.4.

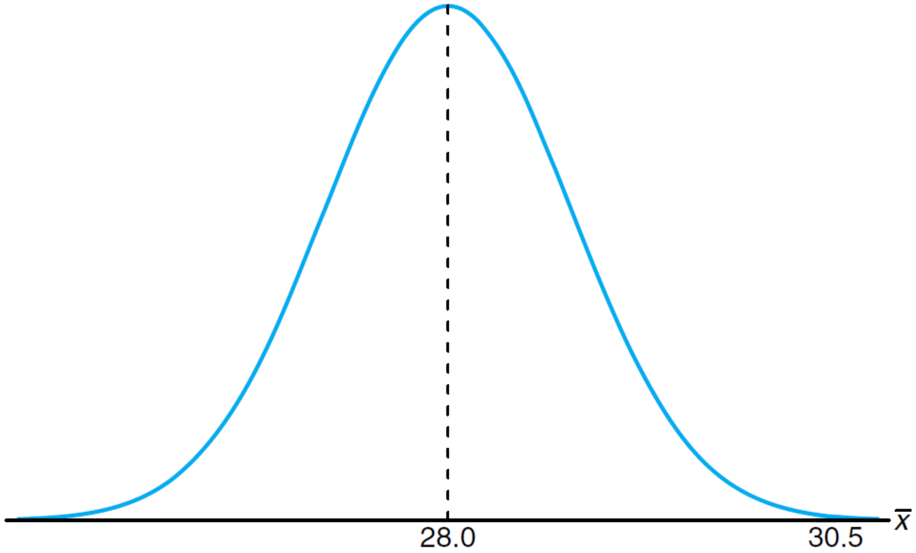


Figure 8.4: Area for Example 8.5.

# Example

- Insurance company
  - An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million

http://www.thepitstopcarcare.com/insurance.html

# Example

- Solution
  - X=total yearly claim
  - Xi=the yearly claim of policy holder i
  - By the CLT, $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution for n=25,000 with mean $320 \times 25{,}000 = 8 \times 10^6$ and SD $540\sqrt{25{,}000} = 8.5381 \times 10^4$

$$
\begin{aligned}
P\{X > 8.3 \times 10^6\} &= P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.5381 \times 10^4}\right\} \\
&= P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{.3 \times 10^6}{8.5381 \times 10^4}\right\} \\
&\approx P\{Z > 3.51\} \qquad \text{where } Z \text{ is a standard normal} \\
&\approx .00023
\end{aligned}
$$

2.3 chances out of 10,000 that the yearly claim will exceed 8.3 million!

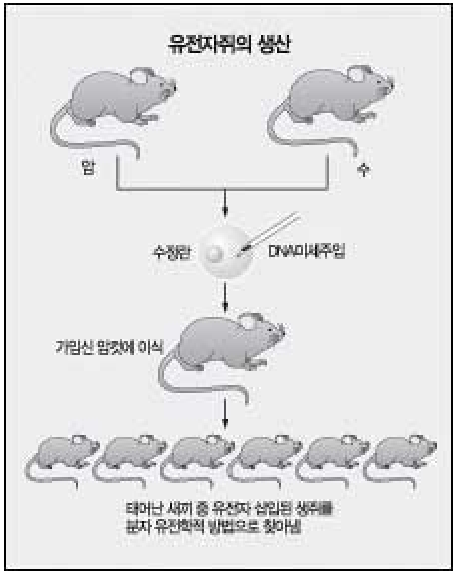
# Question

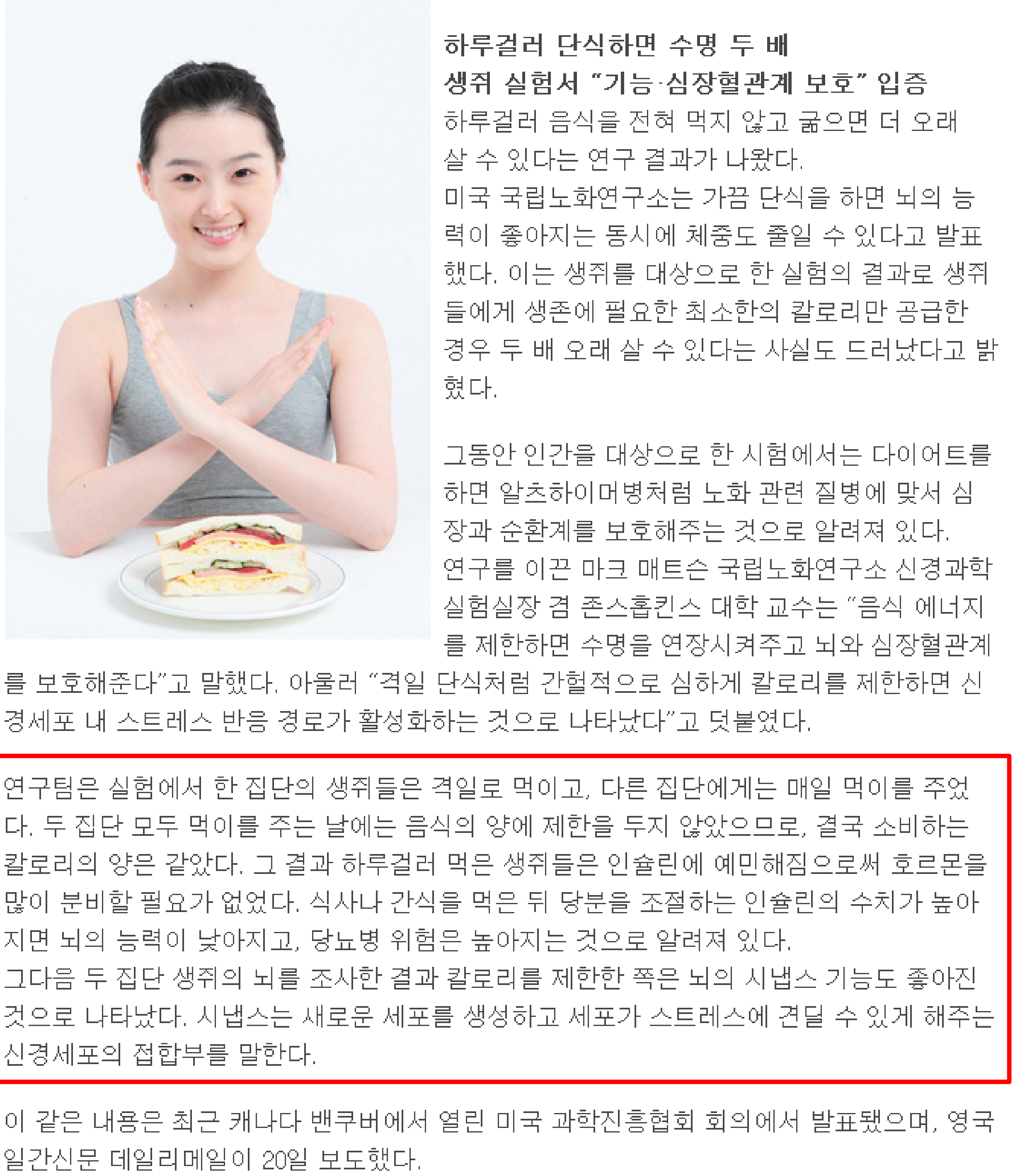
- The mean score for freshmen on an aptitude test at a certain college is 540, with a standard deviation of 50. What is the probability that two groups of students selected at random, consisting of 64 and 100 students, respectively, will differ in their mean scores by more than 10 points?

# *Sampling Distribution of the Difference between Two Means*

- Statistical analyses are very often concerned with the difference between means.

- Typical example:
  - A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group.


유전자쥐의 생산
암
수
수정란
DNA미세주입
가임신 암컷에 이식
태어난 새끼 중 유전자 삽입된 생쥐를
분자 유전학적 방법으로 찾아냄



**하루걸러 단식하면 수명 두 배**

**생쥐 실험서 "기능·심장혈관계 보호" 입증**

하루걸러 음식을 전혀 먹지 않고 굶으면 더 오래 살 수 있다는 연구 결과가 나왔다.

미국 국립노화연구소는 가끔 단식을 하면 뇌의 능력이 좋아지는 동시에 체중도 줄일 수 있다고 발표했다. 이는 생쥐를 대상으로 한 실험의 결과로 생쥐들에게 생존에 필요한 최소한의 칼로리만 공급한 경우 두 배 오래 살 수 있다는 사실도 드러났다고 밝혔다.

그동안 인간을 대상으로 한 시험에서는 다이어트를 하면 알츠하이머병처럼 노화 관련 질병에 맞서 심장과 순환계를 보호해주는 것으로 알려져 있다.

연구를 이끈 마크 매트슨 국립노화연구소 신경과학 실험실장 겸 존스홉킨스 대학 교수는 "음식 에너지를 제한하면 수명을 연장시켜주고 뇌와 심장혈관계를 보호해준다"고 말했다. 아울러 "격일 단식처럼 간헐적으로 심하게 칼로리를 제한하면 신경세포 내 스트레스 반응 경로가 활성화하는 것으로 나타났다"고 덧붙였다.

연구팀은 실험에서 한 집단의 생쥐들은 격일로 먹이고, 다른 집단에게는 매일 먹이를 주었다. 두 집단 모두 먹이를 주는 날에는 음식의 양에 제한을 두지 않았으므로, 결국 소비하는 칼로리의 양은 같았다. 그 결과 하루걸러 먹은 생쥐들은 인슐린에 예민해짐으로써 호르몬을 많이 분비할 필요가 없었다. 식사나 간식을 먹은 뒤 당분을 조절하는 인슐린의 수치가 높아지면 뇌의 능력이 낮아지고, 당뇨병 위험은 높아지는 것으로 알려져 있다.

그다음 두 집단 생쥐의 뇌를 조사한 결과 칼로리를 제한한 쪽은 뇌의 시냅스 기능도 좋아진 것으로 나타났다. 시냅스는 새로운 세포를 생성하고 세포가 스트레스에 견딜 수 있게 해주는 신경세포의 접합부를 말한다.

이 같은 내용은 최근 캐나다 밴쿠버에서 열린 미국 과학진흥협회 회의에서 발표됐으며, 영국 일간신문 데일리메일이 20일 보도했다.

# Really???

- **Key Question**
  - Difference between the mean of two samples really does mean the difference between the mean of the two populations ??

  - How can you ensure that it works for the real population ???

# Sampling distribution of the difference between means

- Two means from two different populations
  - (1) sample $n_1$ scores from Population 1 and $n_2$ scores from Population 2,
  - (2) compute the means of the two samples ($\mu_1$ and $\mu_2$), and
  - (3) compute the difference between means, $\mu_1 - \mu_2$.

- This sampling distribution of the difference between means.
  - is used to make inferences about $\mu_1 - \mu_2$ of populations

- Question:
  - what is the mean and its error of
    the sampling distribution of the difference between means ?

- Example:
  - The mean test score of all 12-year-olds in a population is **34**
  - and the mean of 10-year-olds is 25.
  - If numerous samples were taken from each age group and the mean difference computed each time,
    what would be the mean of these numerous differences between sample means?

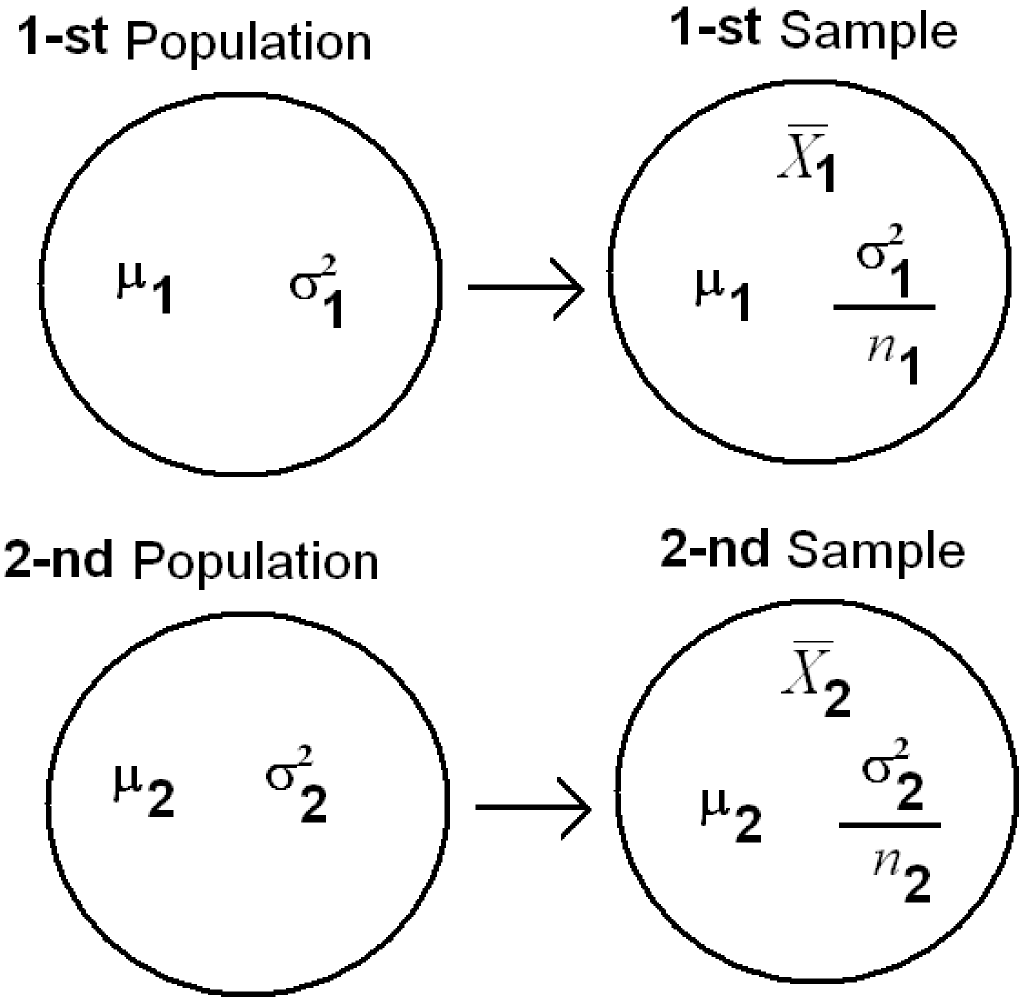
$$E(\overline{X}_1 - \overline{X}_2) = \mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

Answer of this question : 34 - 25 = 9

Error of the mean ? → described by variance

# Again, Sampling Distribution of the Difference between Two Means

- Suppose that we have two populations:
  - 1-st population with mean $\mu_1$ and variance $\sigma_1^2$
  - 2-nd population with mean $\mu_2$ and variance $\sigma_2^2$
- We are interested in comparing $\mu_1$ and $\mu_2$, or equivalently, making inferences about $\mu_1-\mu_2$.
- We independently select a random sample of size $n_1$ from the 1-st population and another random sample of size $n_2$ from the 2-nd population:
  - Let $\bar{X}_1$ be the sample mean of the 1-st sample.
  - Let $\bar{X}_2$ be the sample mean of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1-\mu_2$.



**1-st** Population

$\mu_1$ $\sigma_1^2$

**1-st** Sample

$\overline{X}_1$

$\mu_1$ $\dfrac{\sigma_1^2}{n_1}$

**2-nd** Population

$\mu_2$ $\sigma_2^2$

**2-nd** Sample

$\overline{X}_2$

$\mu_2$ $\dfrac{\sigma_2^2}{n_2}$

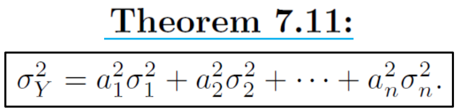
## *Theorem 8.3:*

If $n_1$ and $n_2$ are large, then the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean

$$E(\overline{X}_1 - \overline{X}_2) = \mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

and variance

$$Var(\overline{X}_1 - \overline{X}_2) = \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$



**Theorem 7.11:**

$$\boxed{\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.}$$

$a_1 = 1$ and $a_2 = -1$,

that is:

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

# Example 1

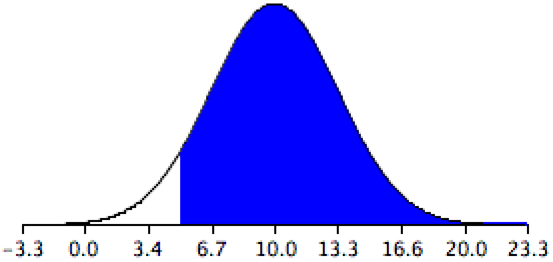
- Assume there are two species of green beings on Mars.
  - The mean height of Species 1 is 32
    while the mean height of Species 2 is 22.
  - The variances of the two species are 60 and 70, respectively and the heights of both species are normally distributed.
  - You randomly sample 10 members of Species 1 and 14 members of Species 2.
  - What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more?

What do you think ? Is the probability high or low ?

- Solution

$$\mu_{M_1-M_2} = 32 - 22 = 10$$

$$\sigma_{M_1-M_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$



$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

## *Example 8.16:*

The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 years and standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6 years and standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a random sample of 49 tubes from manufacturer *B*?

### *Solution:*

Population A

$\mu_1=6.5$

$\sigma_1=0.9$

$n_1=36$

Population B

$\mu_2=6.0$

$\sigma_2=0.8$

$n_2=49$

- We need to find the probability that the mean lifetime of manufacturer *A* is at least 1 year more than the mean lifetime of manufacturer *B* which is P( $\bar{X}_1 \geq \bar{X}_2 + 1$ ).
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$\mathrm{E}(\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2) = \mu_{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2} = \mu_1 - \mu_2 = 6.5 - 6.0 = 0.5$$

$$\mathrm{Var}(\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2) = \sigma^2_{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(0.9)^2}{36} + \frac{(0.8)^2}{49} = 0.03556$$

$$\sigma_{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{0.03556} = 0.189$$

$$\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 \sim \mathrm{N}(0.5, 0.189)$$

# Recall

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 1)$$

$$= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{1 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

$$= P\left(Z \geq \frac{1 - 0.5}{0.189}\right)$$

$$= \mathrm{P}(Z \geq 2.65) = 1 - \mathrm{P}(Z < 2.65)$$
$$= 1 - 0.9960 \;\; = 0.0040$$

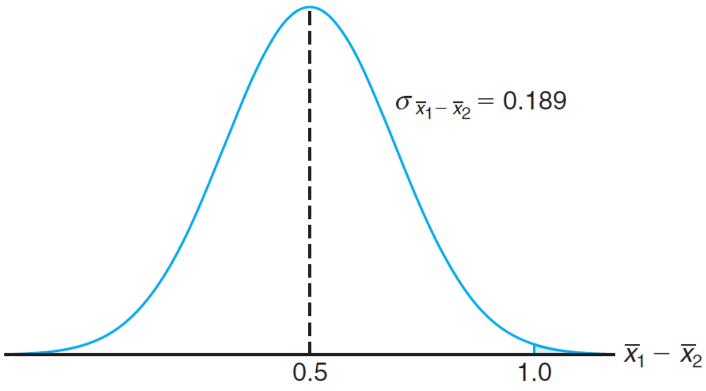


Figure 8.6: Area for Example 8.6.

Q
&
A