Probability

Module 3g

Central Limit Theorem

Learning Objectives

- Reminder of the concepts of sample statistics & population parameters
- Understanding and applying the Central Limit Theorem

Parameters and Statistics

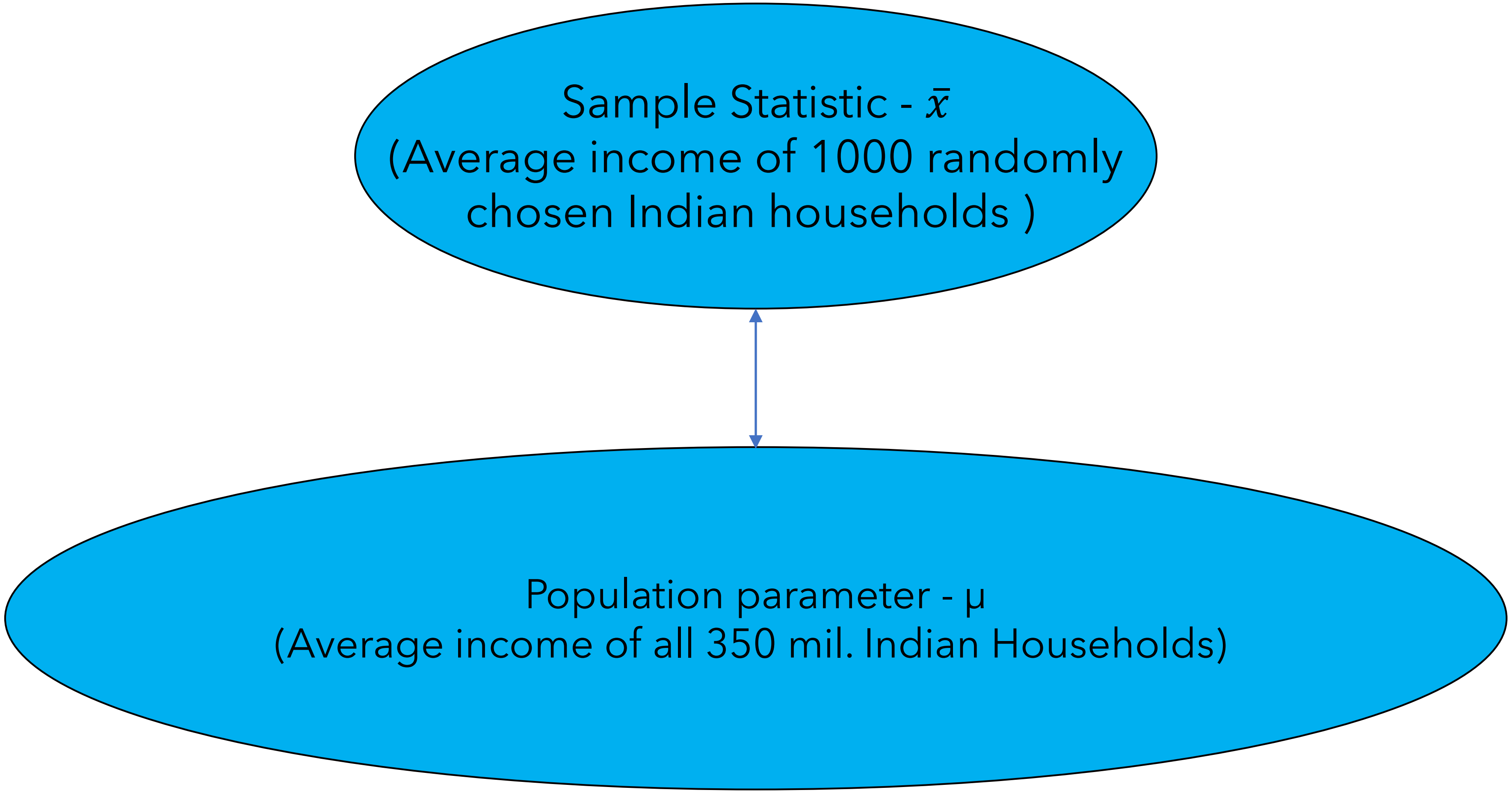
A **parameter** is a numerical descriptive measure of a population (e.g., μ, σ)

- Based on all the observations in the population
- Fixed value
- Almost always unknown
- **If known we wouldn’t bother to sample**

A **sample statistic** is a numerical descriptive measure of a sample (e.g. $\bar{x}$, s)

- Calculated from the observations in the sample
- Used as estimates of population parameter
- Varies from sample to sample–it’s a random variable!

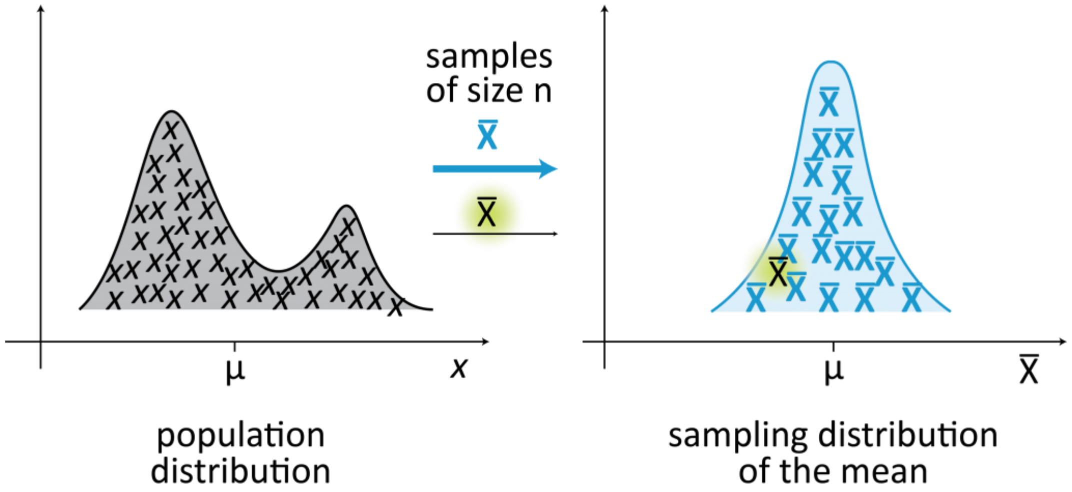
Inferential Statistics: Drawing Conclusions about a Population from a Sample



Sample Statistic - $\bar{x}$
(Average income of 1000 randomly chosen Indian households)

Population parameter - μ
(Average income of all 350 mil. Indian Households)

Sampling Distributions



What is a sampling distribution?

- Sample statistics are random variables (RV)
- Sample statistics like all RV have probability distributions
- A "sampling distribution" is the probability distribution of a sample statistic

Any statistic you can calculate from a sample has a sampling distribution

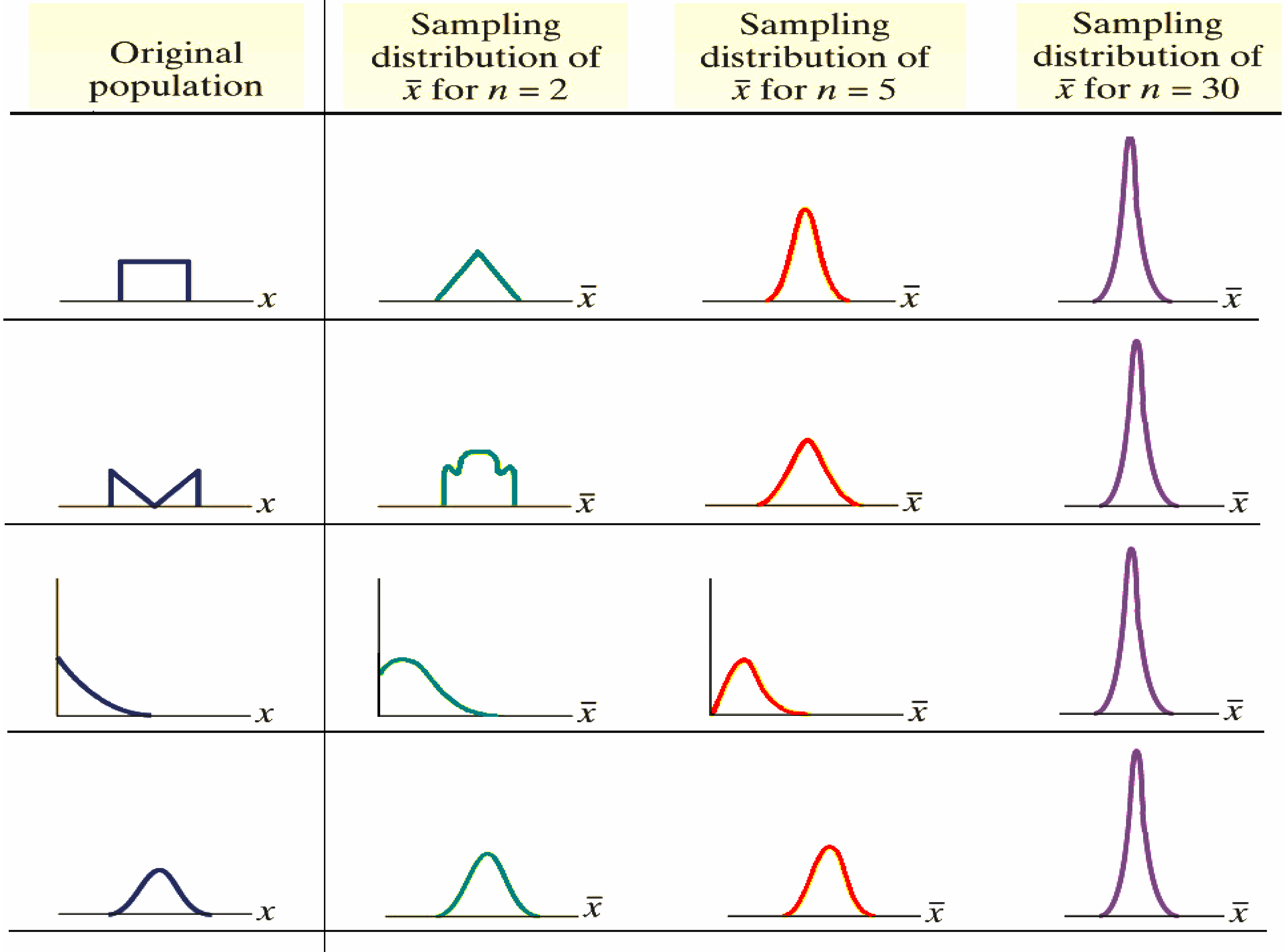
Sampling distribution specifies the probability of all possible sample values for a fixed sample size, n.

e.g., For the sample statistic $\bar{x}$, graph of all possible pairs $\left(\bar{x}, P(\bar{x})\right)$

Central Limit Theorem (CLT)

- If we select a random sample of n observations from any population with mean **μ** and standard deviation **σ**
- Regardless of the distribution of x, as n gets large the sampling distribution of $\bar{x}$ becomes **approximately** normal
- The mean and standard deviation of this (normal) sampling distribution of $\bar{x}$ is:

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



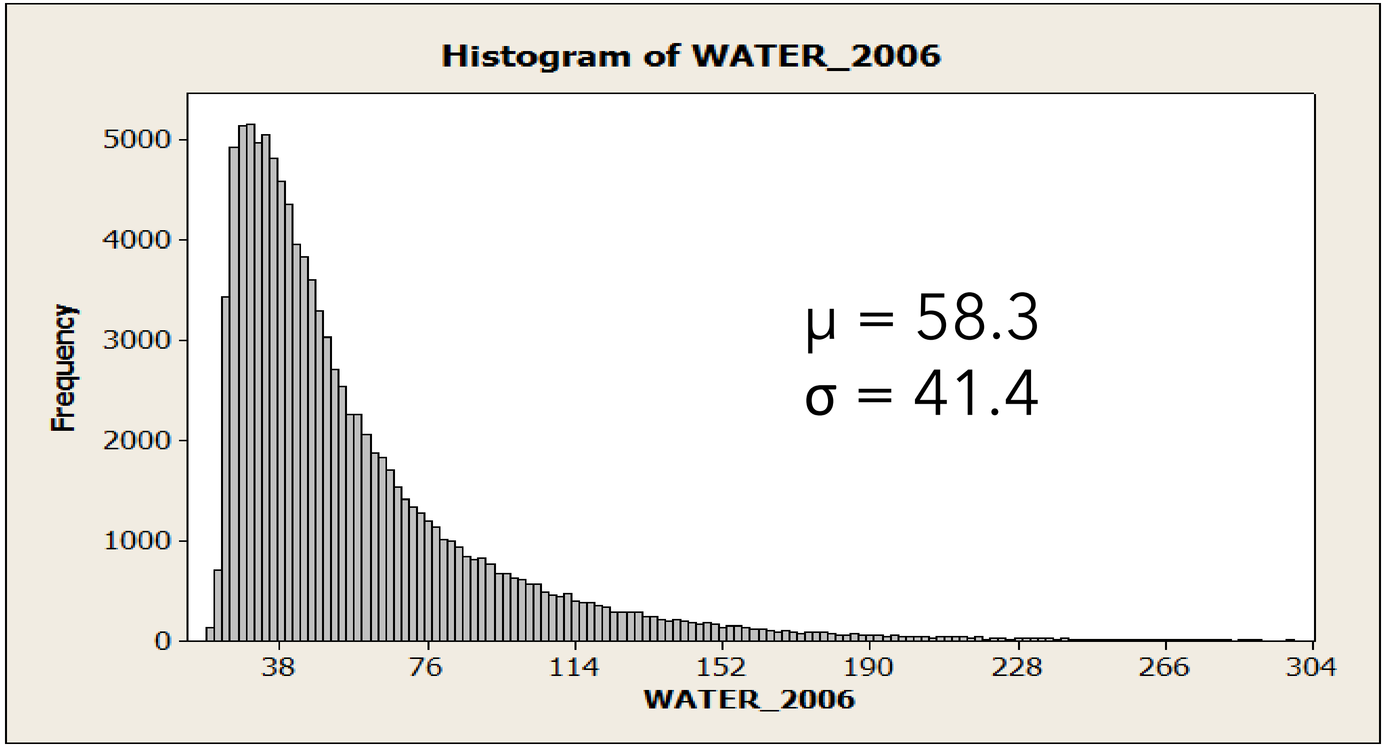
Original population
Sampling distribution of $\bar{x}$ for $n = 2$
Sampling distribution of $\bar{x}$ for $n = 5$
Sampling distribution of $\bar{x}$ for $n = 30$
x
$\bar{x}$
$\bar{x}$
$\bar{x}$
x
$\bar{x}$
$\bar{x}$
$\bar{x}$
x
$\bar{x}$
$\bar{x}$
$\bar{x}$
x
$\bar{x}$
$\bar{x}$
$\bar{x}$

Demonstration of CLT Cobb County Water Authority Data

- Data on water consumption of the Cobb County Water Authority’s customer base from 2006-2009
- Customer base was the target of an experiment to reduce water consumption (to be discussed in later lectures)
- Because the data set includes (nearly) the entire population, we can actually measure population parameters!

Cobb County Water Authority Data

Distribution of *x*, the 2006 water usage (000 gallons) for a household.



$\mu = 58.3$
$\sigma = 41.4$

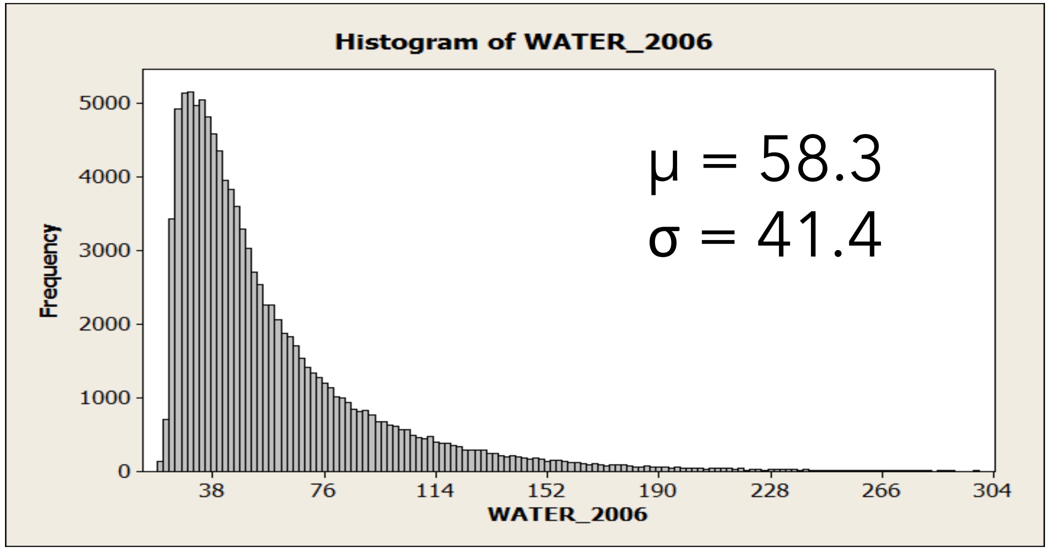
Central Limit Theorem Demo

Some Important Takeaway Points

- Sample estimates of μ vary from sample to sample –They are random variables!
- Per the Central Limit Theorem the distribution of sample averages for fixed sample size, n, follows the normal distribution
- No estimate is exactly equal to μ – some are pretty close but others are not
- In the real world we don’t know which estimates are close and which are far away
- If we did, we wouldn’t bother to sample because we’d already know μ

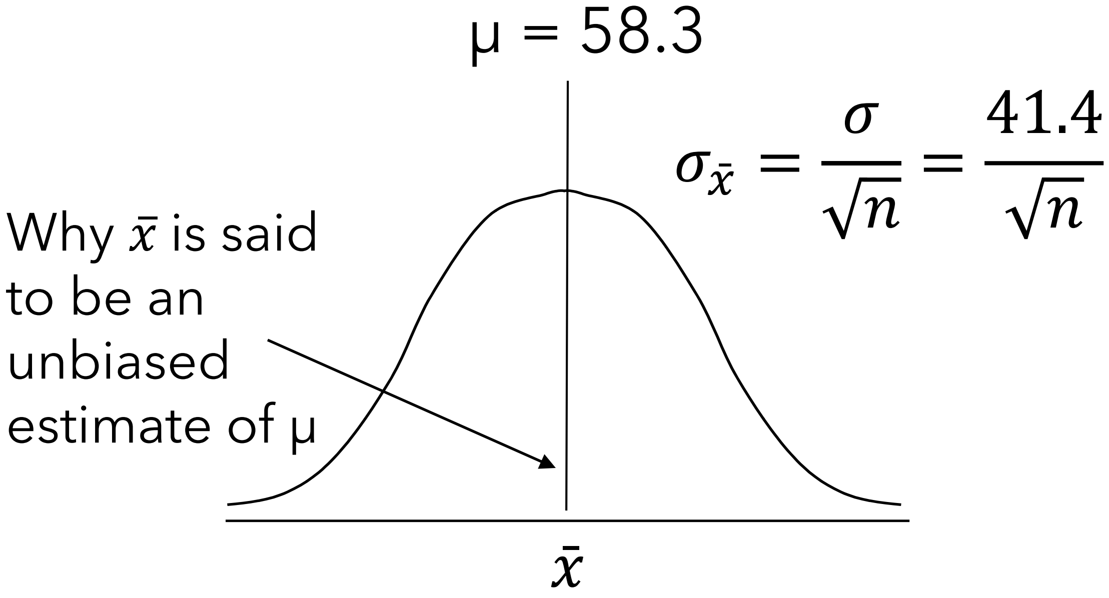
CLT for Water 2006

Distribution of *x*, the 2006 water bill for a household.



μ = 58.3
σ = 41.4

Sampling distribution of $\bar{x}$, the average 2006 water bill for n households.



μ = 58.3

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{41.4}{\sqrt{n}}$$

Why $\bar{x}$ is said to be an unbiased estimate of μ

Why is the Central Limit Theorem Important?

- Sampling distribution of $\bar{x}$ is normally distributed no matter how x is distributed.
- Allows us to make probabilistic statements about $\bar{x}$'s relationship to μ without knowing how x is actually distributed in the population.

How Large Does the Sample Size, *n*, Have to Be?

No clear cut answer but…

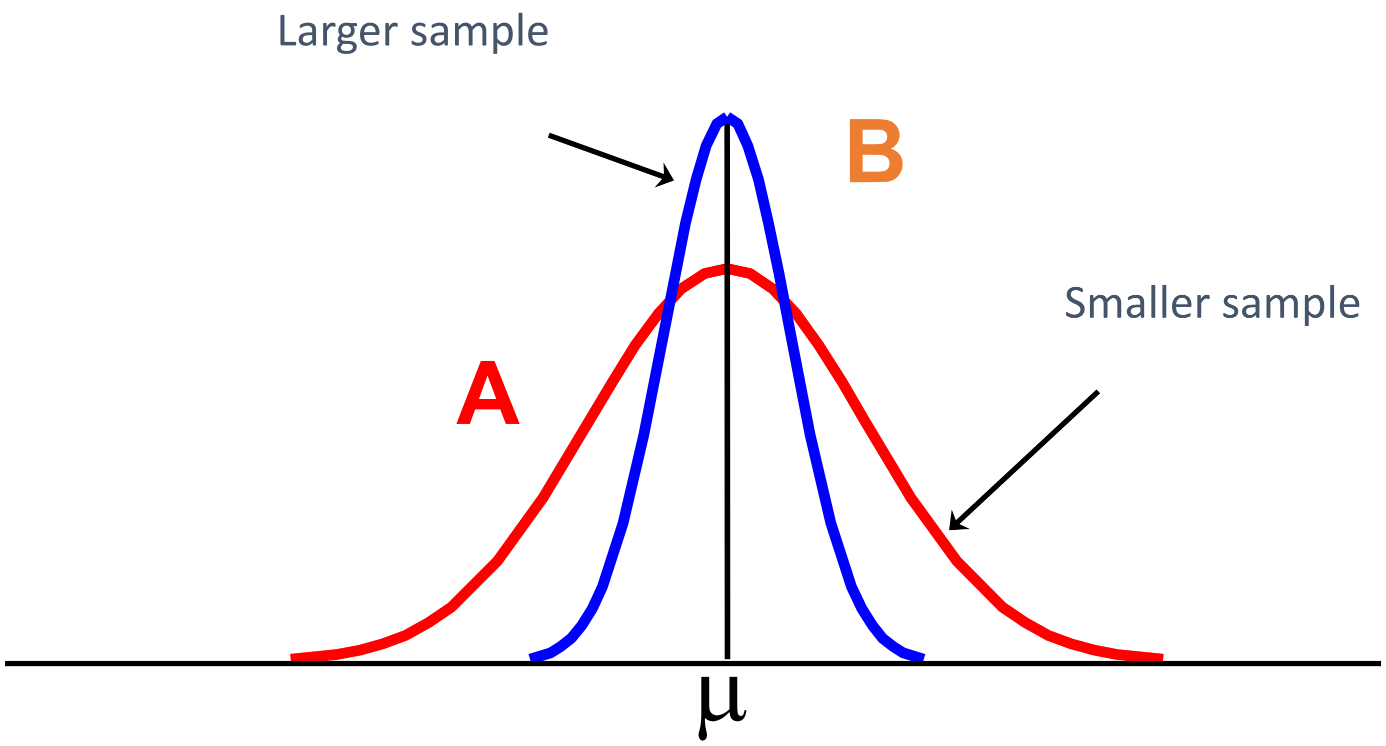
- If x is normally distributed, CLT holds regardless of sample size. This means – if you **know** that x is normally distributed, the sampling distribution will be normally distributed regardless of the sample size.
- If x is not normal but not terribly skewed CLT is likely to hold for samples sizes of 30 or more and can sometimes holds for samples of about 5 or 10.

Standard Error of the Mean— $\sigma_{\bar{x}}$

- Term for describing standard deviation of the normally distributed sampling distribution of $\bar{x}$, which measures dispersion in possible values of sample means.
- Formula:

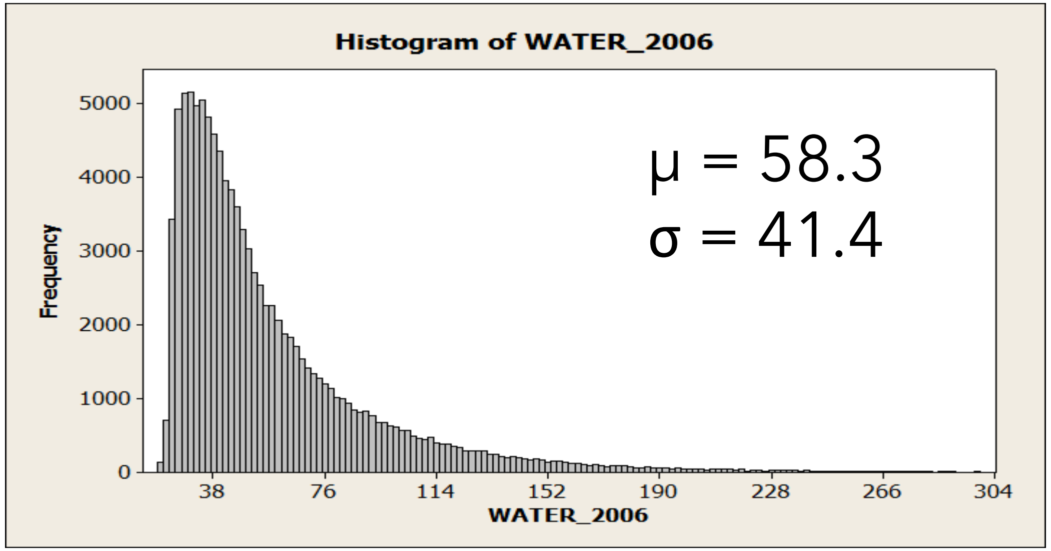
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

More Data: Better Estimate



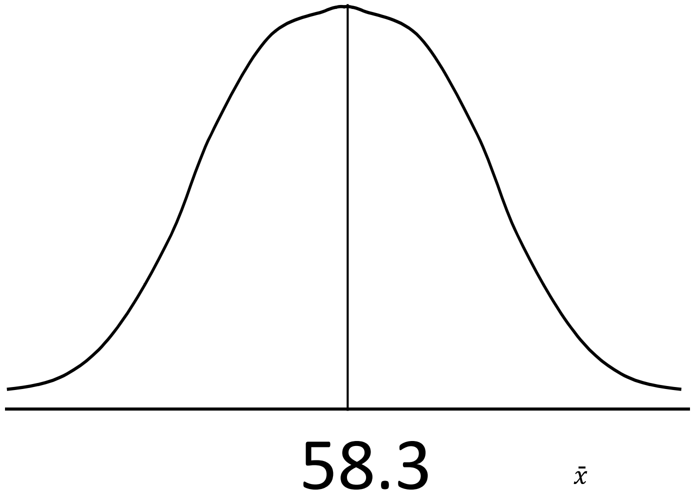
Example

For the water_2006 variable, in a random sample of 144 customers what is the probability that the sample mean will fall between 52.6 and 64.0?

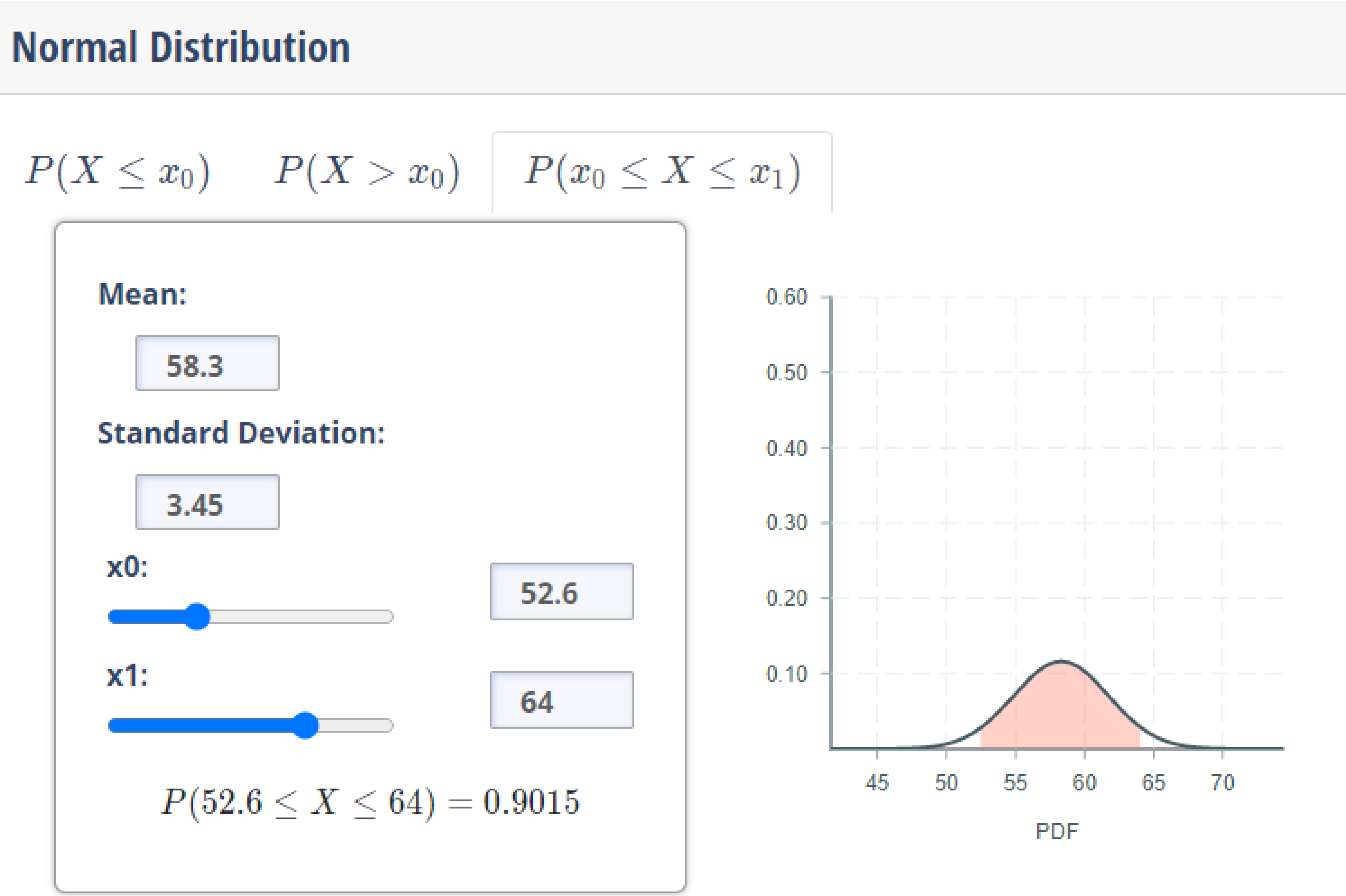


μ = 58.3
σ = 41.4

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{41.4}{\sqrt{144}} = 3.45$$



Solution



Normal Distribution

Mean: 58.3

Standard Deviation: 3.45

x0: 52.6

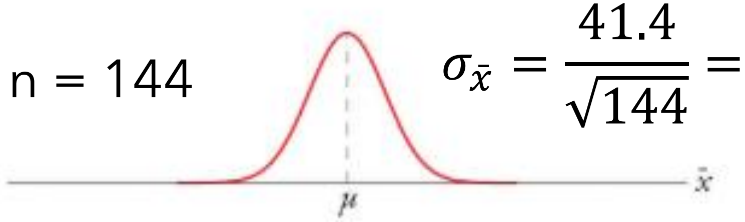
x1: 64

$P(52.6 \leq X \leq 64) = 0.9015$

Solution Continued

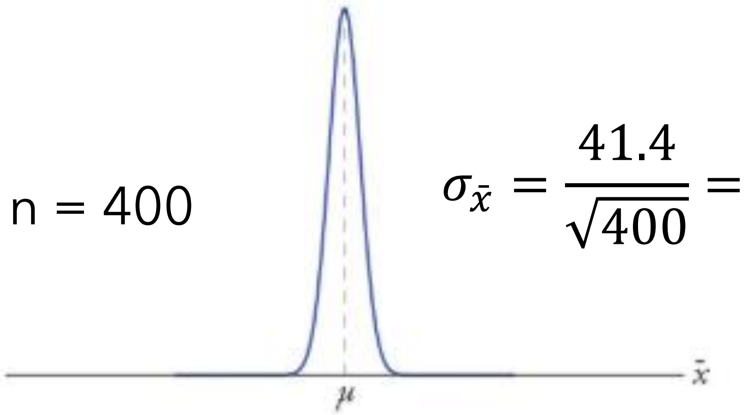
What happens with the probability if we increase sample size to 400?

n = 144



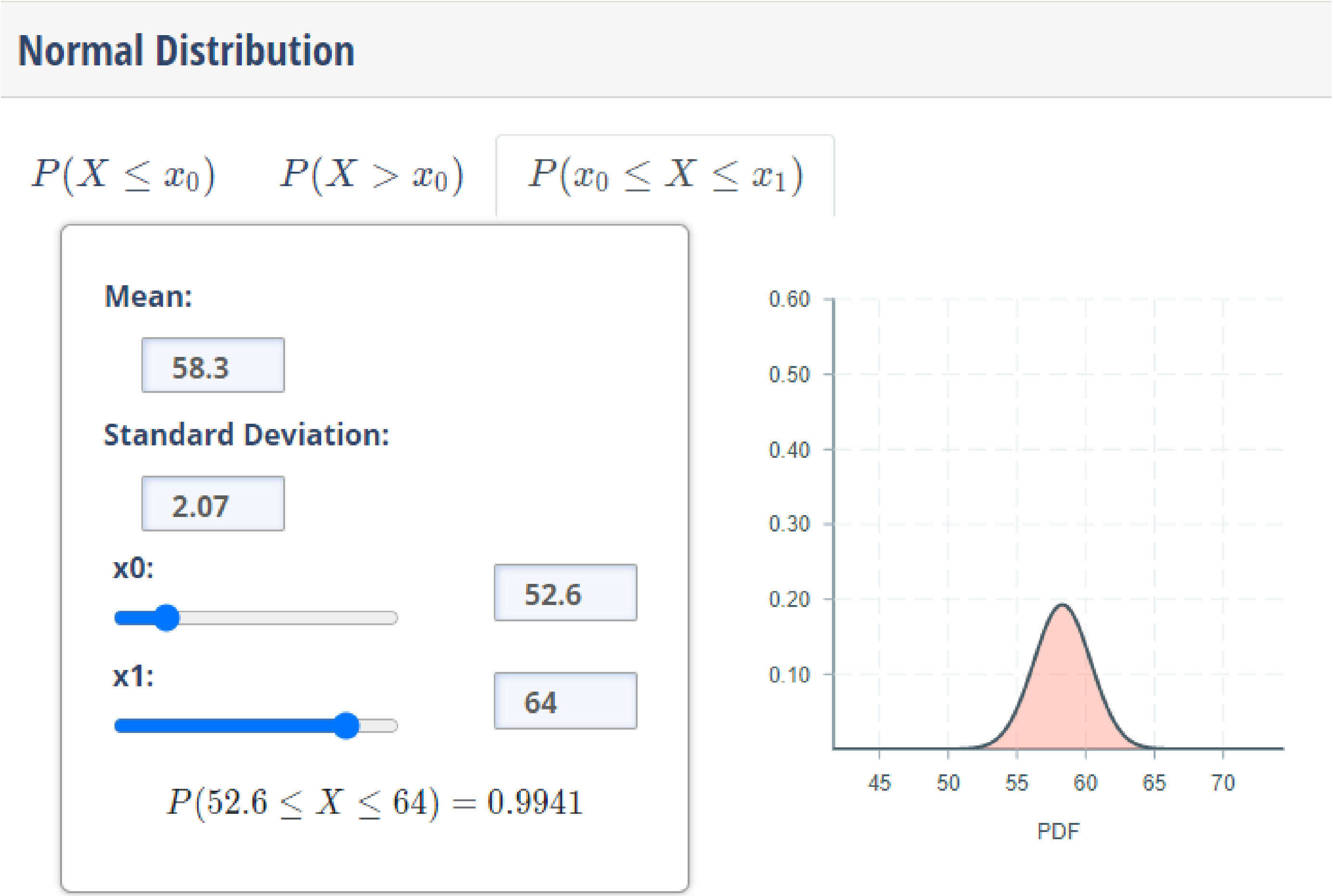
$$\sigma_{\bar{x}} = \frac{41.4}{\sqrt{144}} = 3.45$$

n = 400



$$\sigma_{\bar{x}} = \frac{41.4}{\sqrt{400}} = 2.07$$

μ=58.3



Normal Distribution

$P(X \leq x_0)$ $P(X > x_0)$ $P(x_0 \leq X \leq x_1)$

Mean: 58.3

Standard Deviation: 2.07

x0: 52.6

x1: 64

$P(52.6 \leq X \leq 64) = 0.9941$

Tests of Understanding

1) The taxi and takeoff times of commercial jets is a random variable with a mean of 8.5 minutes and standard deviation of 2.5 minutes. What is the probability that for 36 jets, the total taxi and take off time will be less than 320 minutes?