IX. Sampling Distributions & CLT

Instructor: **Yanlin Qi**

Institute of Transportation Studies
Department of Statistics
University of California, Davis

- Today:
  - Probability and Statistic
  - Random Sample
  - Statistics and Their distributions
  - The Distribution of the Sample Mean
  - Central Limit Theorem (CLT)

Probability & Statistics

Introduction - Probability

- **Probability**
  - A branch of mathematics that deals with the likelihood of occurrence of different events. It provides a framework for quantifying uncertainty.
  - **Focus:** Theoretical and model-based
  - **Key Concepts:**
    - Random Variables
    - Probability Distributions
    - Expected Value and Variance

Introduction - Statistics

- **Statistics**
  - A branch of mathematics that deals with collecting, analyzing, interpreting, and presenting empirical data. It uses probability theory to draw conclusions from data.
  - **Focus:** Empirical and data-based
  - **Key Concepts:**
    - Descriptive Statistics
    - Inferential Statistics
    - Sampling

Introduction - Probability vs Statistics

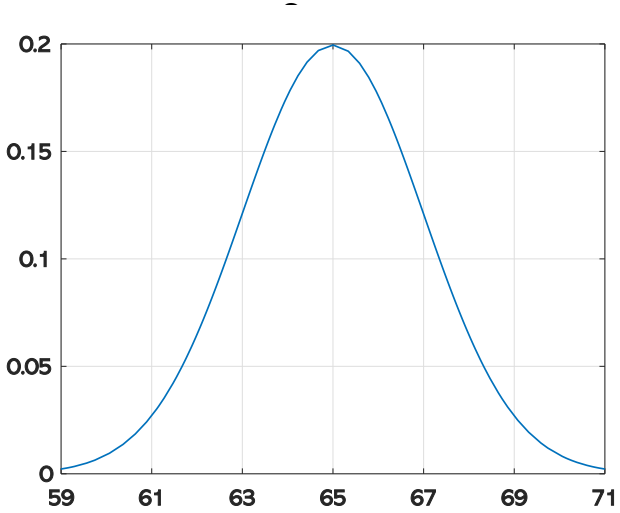
❖ Probability vs Statistics

Aspect	Probability	Statistics
Starting Point	Known model or distribution	Collected data
Primary Goal	Predict likelihood of events	Infer population characteristics from samples
Approach	Deductive	Inductive
Key Tools	Probability distributions, random variables, expected value	Descriptive statistics, inferential statistics, sampling techniques
Applications	Risk assessment, random processes, games of chance	Scientific research, business analytics, quality control

Random Sample

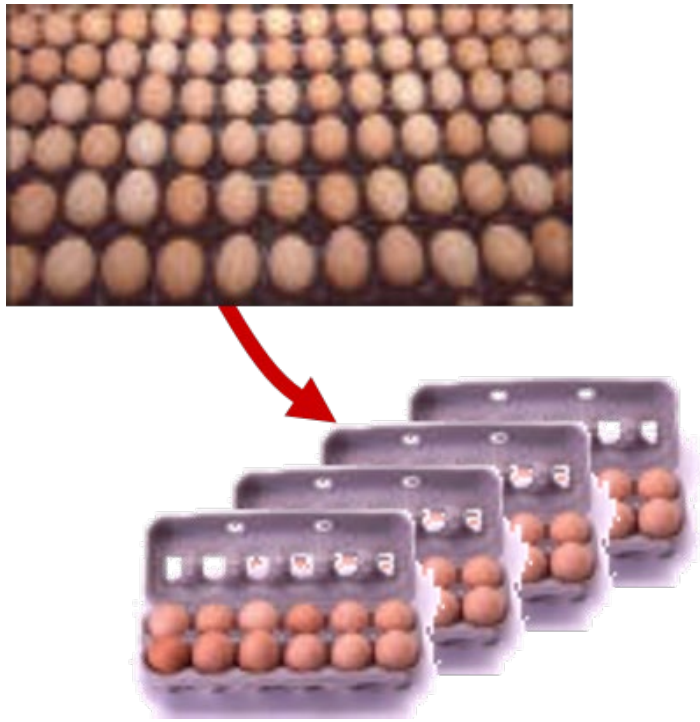
Motivating example

Suppose that the weights (in grams) of brown eggs produced at a local farm have a normal distribution:



Sampling distributions

1. Those eggs are divided into cartons of size 12, to be sold on the market.
2. You randomly select a carton and measure the weights of all the 12 eggs in it.
3. Let $\bar{X}$ be their **average weight**.
4. $\bar{X}$ clearly may vary from carton to carton, and thus is a (continuous) random variable.

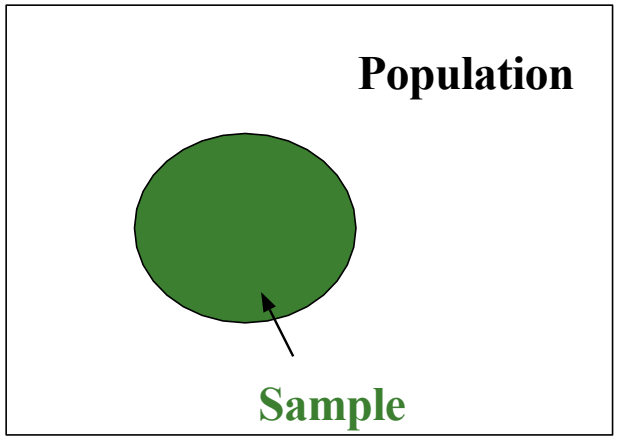


Question: What is the distribution of $\bar{X}$?

Sampling distributions

The above question is about the **sampling distribution of a statistic**.

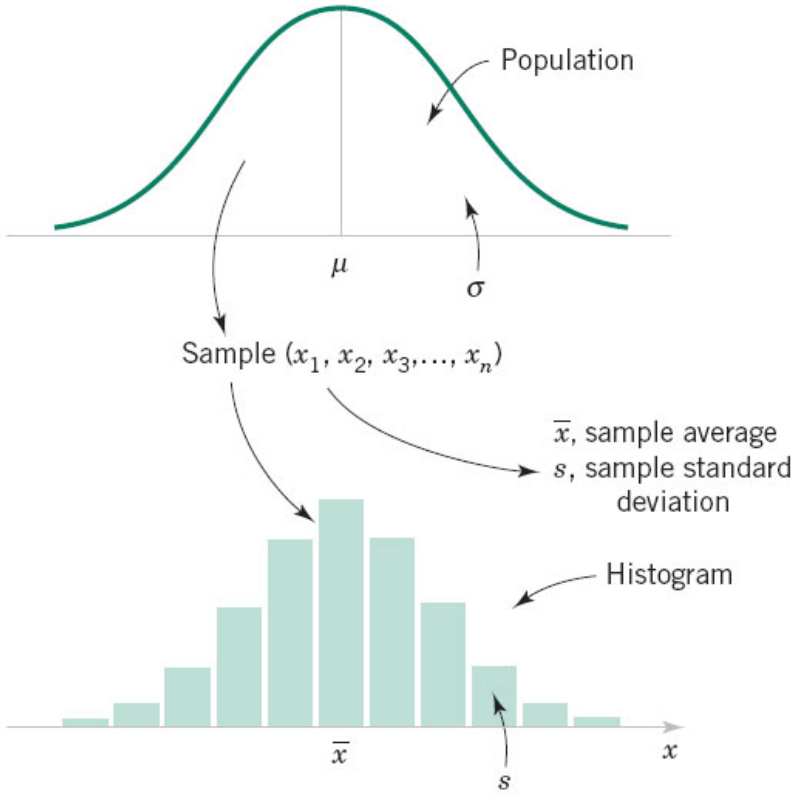
- **Population**: all brown eggs produced at the farm
- **Sample**: a carton of 12 eggs
- **Statistic**: $\bar{X}$ (average weight of the 12 eggs in the sample)



Understanding Sampling Distributions

Population vs. Sample:

- **Population:** The entire group you're interested in studying.
- **Sample:** A subset of the population selected for analysis.



Random Sample

To study the distribution of $\bar{X}$, we denote individual weights of the 12 to-be-selected eggs as $X_1, \ldots, X_{12}$.

We then have

$$\bar{X} = \frac{X_1 + \cdots + X_{12}}{12}.$$

What we know about $X_1, \ldots, X_{12}$:

They are *identically and independently distributed (iid)*:

$$X_1, \ldots, X_{12} \overset{iid}{\sim} N(65, 2^2)$$

and are called a **random sample** (of size 12) from the distribution $N(65, 2^2)$.

Sampling distributions - Random sample

Random sample

Def 0.1. More generally, a collection of n random variables $X_1, \ldots, X_n$ is called a random sample, if they are

(1) identically distributed according to some pmf/pdf $f(x)$, and

(2) independent.

In short, we write $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$.

Random sample

Example 0.1. Coin Tossing

- **Scenario:** Suppose you toss a coin repeatedly and independently for a total of n times. The probability of getting heads is p.
- **Numerical Outcomes:** let $X_1, \ldots, X_n$ denote the numerical outcomes of individual trials:
  - 1 represents heads.
  - 0 represents tails.
- **Random Sample:** This setup constitutes a random sample from the $Bernoulli(p)$ distribution because each trial (toss) is
  - **Independent:** The outcome of one trial does not affect another.
  - **Identically distributed:** Each trial has the same probability p of heads (1) and 1−p of tails (0).
- **Mathematical Representation:**

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p).$$

Random sample

Example 0.2.

- Let $X_1, \ldots, X_n$ represent n repeated and independent measurements of an object's length.
- These measurements can be thought of as a random sample from a normal distribution

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

where

- μ: true length (assuming the measurement process is unbiased)
- σ^2: variance of the measurement error.

Specific realizations of a random sample

Example 0.3.

- Suppose you ***actually*** buy a carton of $n = 12$ eggs from the farm and measure their weights individually. Then you may obtain a data set like the following (called a specific sample):

$x_1 = 65.4,\ x_2 = 65.0,\ x_3 = 64.8,\ x_4 = 65.1,\ x_5 = 64.8,\ x_6 = 64.4,$

$x_7 = 65.0,\ x_8 = 65.1,\ x_9 = 65.5,\ x_{10} = 64.8,\ x_{11} = 64.8,\ x_{12} = 65.2$

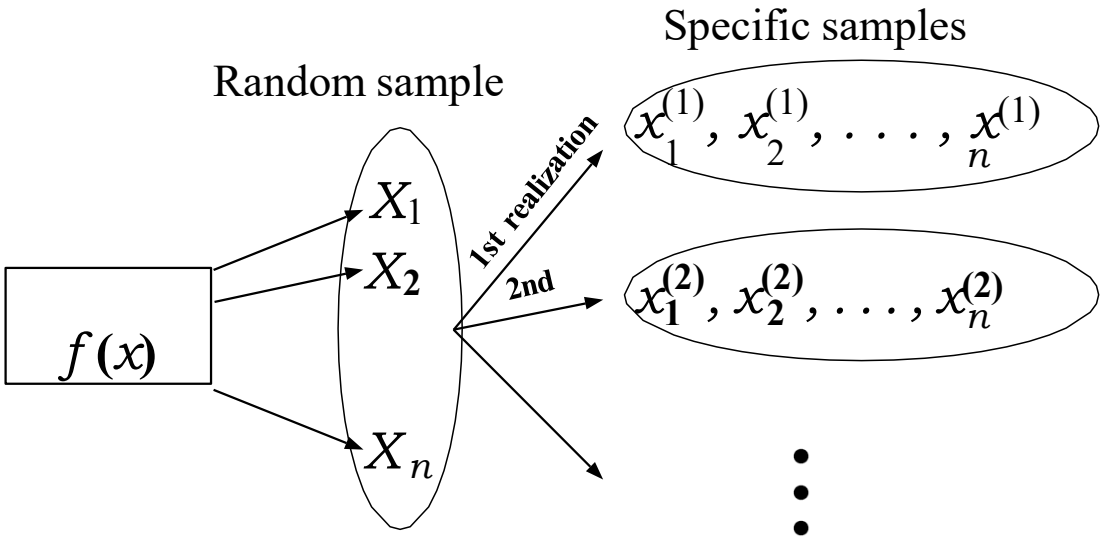
Notation:

We use lowercase letters such as $x_1, x_2, \ldots$ to represent specific values of the random variables ($X_1, X_2, \ldots$) in a random sample.

Random Sample

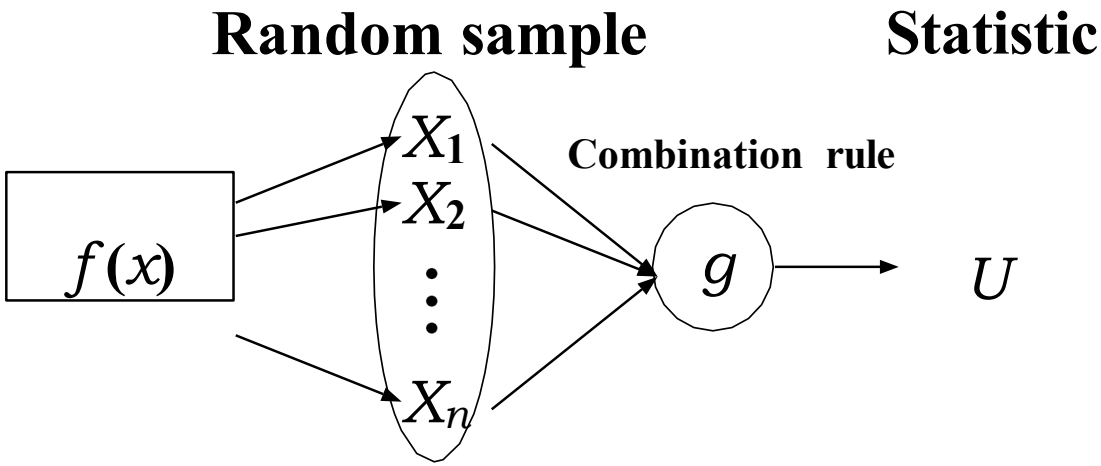
Remark. If we realize the sampling process again, then we may obtain a different set of weights. For example,

$$x_1 = 65.6,\ x_2 = 64.3,\ x_3 = 64.2,\ x_4 = 65.4,\ x_5 = 64.9,\ x_6 = 64.4,$$
$$x_7 = 65.2,\ x_8 = 65.2,\ x_9 = 65.0,\ x_{10} = 64.7,\ x_{11} = 64.5,\ x_{12} = 65.1$$



Statistic

- **Def 0.2.** Mathematically, a statistic is just a **summary** of a random sample by certain combination rule g:
  - $U = g(X_1, X_2, \ldots, X_n)$



Statistic

***Remark*.** Depending on purpose, different statistics may be defined on the same random sample. Two common ones are

- **Sample mean**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ ← **a measure of center, or location**

- **Sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ ← **a measure of variability**

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

Statistic - Sample Mean

- We can characterize the central tendency in the data by the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, x_2, \ldots, x_n$ are n observations in a sample selected from some larger population

- The sample mean $\bar{x}$ is a reasonable estimate of the population mean μ

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Statistic - Sample Variance

- We can characterize the variability or scatter in the data by the sample variance or sample standard deviation

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

where $x_1, x_2, \ldots, x_n$ are n observations in a sample selected from some larger population

- The sample variance s^2 is a reasonable estimate of the population variance σ^2

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \sigma = \sqrt{\sigma^2}$$

Other Statistics

Other examples of statistics include

- **sample median** (also a measure of center)
- **sample minimum** or **maximum**
- **sample range** (i.e., sample maximum - sample minimum)

See Chapter 1 for details.

Statistics as Random Variables

Statistics are random variables

Clearly, for different realizations of the sampling process, the values of the statistic may vary. For the egg weight example (and the statistic $\bar{X}$),

(1) One realization ($\bar{x} = 64.992$):

$$x_1 = 65.4,\ x_2 = 65.0,\ x_3 = 64.8,\ x_4 = 65.1,\ x_5 = 64.8,\ x_6 = 64.4,$$
$$x_7 = 65.0,\ x_8 = 65.1,\ x_9 = 65.5,\ x_{10} = 64.8,\ x_{11} = 64.8,\ x_{12} = 65.2$$

(2) Another realization ($\bar{x} = 64.875$) :

$$x_1 = 65.6,\ x_2 = 64.3,\ x_3 = 64.2,\ x_4 = 65.4,\ x_5 = 64.9,\ x_6 = 64.4,$$
$$x_7 = 65.2,\ x_8 = 65.2,\ x_9 = 65.0,\ x_{10} = 64.7,\ x_{11} = 64.5,\ x_{12} = 65.1$$

Sampling Distributions

Sampling distributions

- **Definition:**
  - The distribution of a given statistic (like the mean) calculated from multiple samples of the same size drawn from the population.
  - It concern the randomness associated to a **statistic** based on a **random sample** from a population.

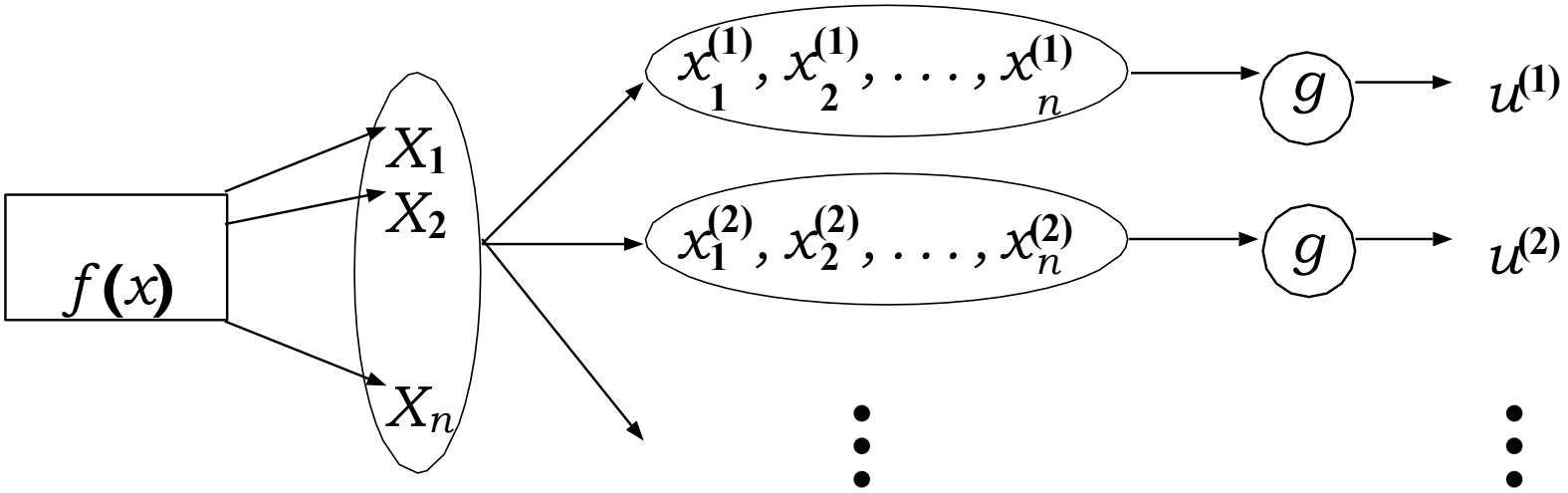
- It serves as the bridge between probability and statistics.

Sampling distribution of a statistic

Def 0.3. The probabilistic distribution of a statistic (as a random variable)

$$U = g(X_1, X_2, \ldots, X_n)$$

is called the *sampling distribution* of the statistic.



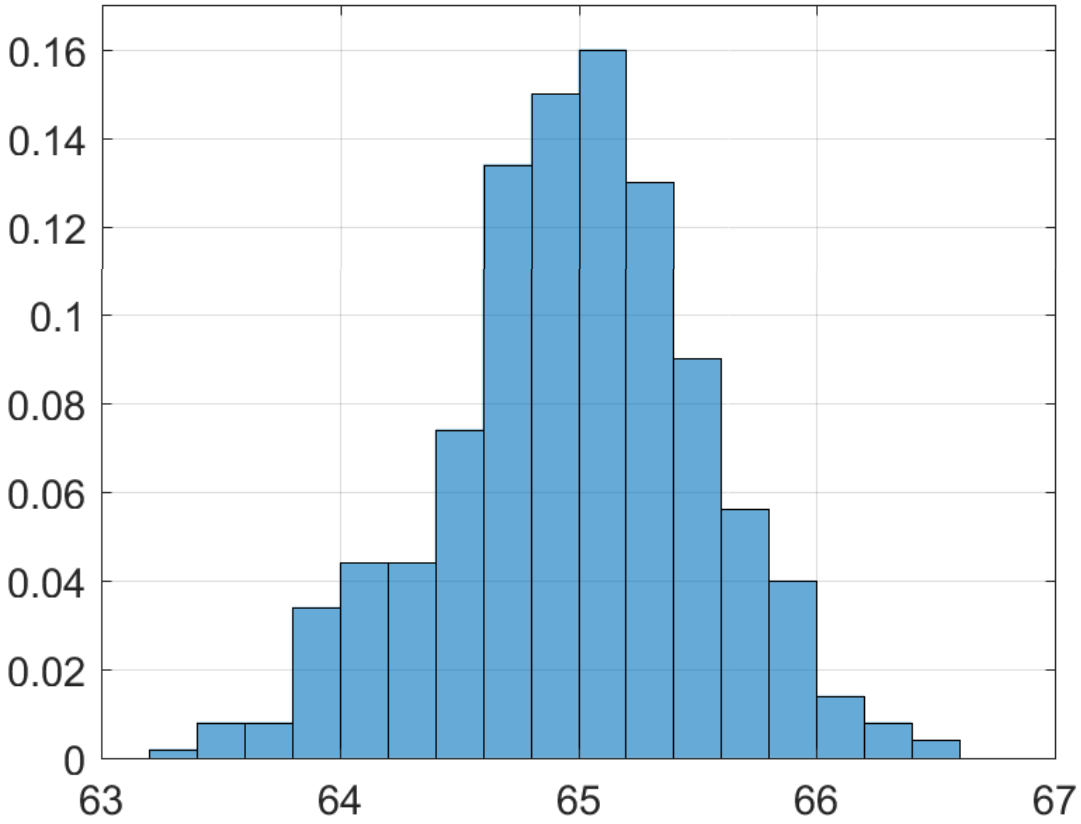
Sampling distributions

Simulation

We "selected" 500 cartons of eggs randomly from the farm (through computer simulation) and computed their average weights. Below shows 50 observations of $\bar{X}$:

65.0506 64.7592 65.0571 64.9674 65.4973 64.7503 65.0393 64.6714
65.3764 65.2525 65.2012 64.4910 65.6002 65.1868 65.0916 63.8280
65.2636 64.9638 65.2998 65.5587 63.9801 65.3903 64.9052 65.7352
64.6329 64.5109 65.7044 64.3291 65.1044 64.8036 66.0407 65.3560
65.3534 65.4668 64.7394 65.1690 64.5668 64.8478 64.0334 65.7562
64.8553 64.9939 65.6044 64.5237 64.2092 64.5860 65.2096 65.5114
64.6195 65.0312

We can display all 500 values of $\bar{X}$ through a histogram shown below



Sampling distributions - sample mean

We focus on the sample mean statistic

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x).$$

and

$$\mathbf{E}(X_i) = \mu, \qquad \mathbf{Var}(X_i) = \sigma^2, \textbf{ for all } i.$$

Sampling distributions

We present three different results for **the statistic $\bar{X}$** :

1. **Expectation and variance of $\bar{X}$** (for any distribution $f(x)$)
2. **Exact distribution of $\bar{X}$** when $f(x)$ is a normal distribution
3. **Approximate distribution of $\bar{X}$** for nonnomral distributions in the setting of a large sample

Sampling distributions

General distributions:

Expectation and variance of $\bar{X}$

Theorem 0.1. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$, with $\mathrm{E}(X_i) = \mu$ (population mean) and $\mathrm{Var}(X_i) = \sigma^2$ (population variance). Then

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Remark. This result does **NOT** concern the specific distribution of $\bar{X}$!

Sampling distributions

Proof. By linearity and independence,

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\left(\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)\right) = \frac{1}{n}(\mu + \cdots + \mu) = \mu$$

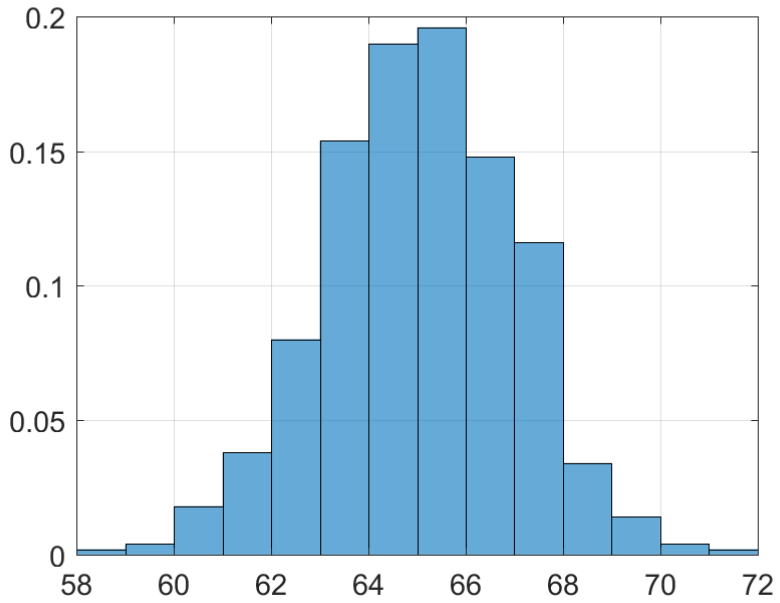
$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

Remark. The theorem indicates that

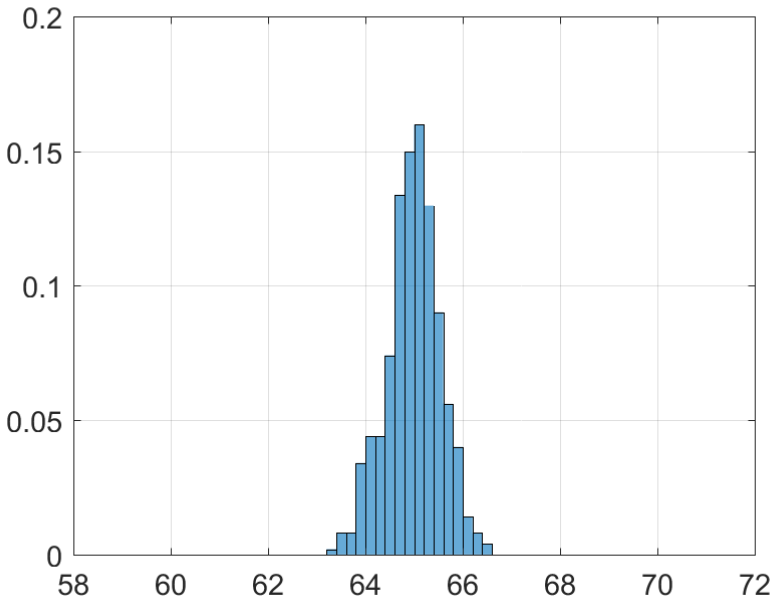
- expectation of $\bar{X}$ is μ (population mean), and
- variance of $\bar{X}$ is only $\frac{1}{n}$ of the population variance (for single X_i)

Sampling distributions

Example 0.4. Weights of 500 single eggs (left) and average weights of 500 cartons (right), all selected at random.



X_i



$\bar{X}$

Sampling distributions

Normal populations: Exact distribution of $\overline{X}$

Assume a random sample

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2).$$

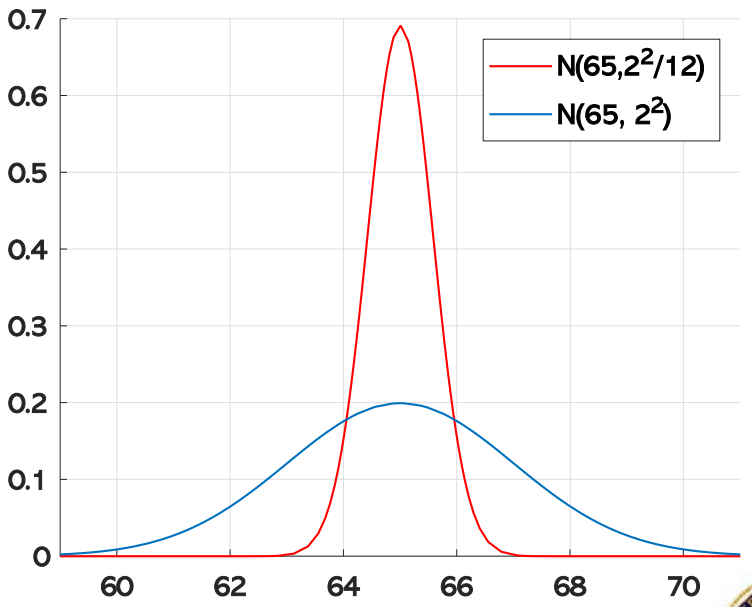
We assume Normal distribution of $X_1, X_2, \ldots, X_n$ here

***Theorem* 0.2.** We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

This also implies that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$



Sampling distributions - Example

- Example 0.5. Brown Egg Example
- Suppose population distribution is $N(65, 2^2)$.
- 1. Probability for Sample Mean:
  - For a random sample of size 12, what is the probability that the sample mean $\bar{X}$ is within 65 ± 1?
- 2. Probability for an Individual Egg:
  - What is the probability that an individual egg is within 65±1?

Example - Sample Mean Probability

- Explanation:
  - Population mean (μ) = 65
  - Population standard deviation (σ) = 2
  - Sample size (n) = 12

Sampling distribution of the sample mean:

- The variance of the sample mean (μ) = 65
- The variance of the sample mean $\sigma_{\bar{X}}{}^2 = \frac{\sigma^2}{n} = \frac{4}{12} = \frac{1}{3}$
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = \left(65, \frac{1}{3}\right)$
- $\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}{}^2} = \sqrt{\frac{1}{3}} \approx 0.577$

Example - Sample Mean $(\bar{X})$ Probability

- For Q1. We need to find the probability that **the sample mean $\bar{X}$** is within 64 and 66 (65 ± 1).
- This is done by finding the Z-scores for 64 and 66:

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

For 64: $$z = \frac{\mathbf{64 - 65}}{\mathbf{0.577}} \approx -1.73$$

For 66: $$z = \frac{\mathbf{66 - 65}}{\mathbf{0.577}} \approx 1.73$$

Using the standard normal distribution table:

$$P(64 \leq \bar{X} \leq 66) = \Phi(Z = 1.73) - \Phi(Z = -1.73) \approx 0.9164$$

Example - Individual Egg (X) Probability

- Population mean (μ) = 65
- Population standard deviation (σ) = 2

- $\bar{X} \sim N(\mu, \sigma^2) = (65,4)$
- We need to find the probability that **an individual egg's weight** is within 64 and 66.

For 64: $$z = \frac{\mathbf{64 - 65}}{\mathbf{2}} = -0.5$$

For 66: $$z = \frac{\mathbf{66 - 65}}{\mathbf{2}} \approx 0.5$$

Using the standard normal distribution table:

$$P(64 \leq \mathrm{X} \leq 66) = \Phi(\mathrm{Z} = 0.5) - \Phi(\mathrm{Z} = -0.5) \approx \mathbf{0.3829}$$

Sampling distributions

- **Example 0.6. Library Elevator Weight Limit**
  - In the library elevator of a large university, there is a sign indicating a 16-person limit as well as a weight limit of 2500 lbs. When the elevator is full, we can think of the 16 people in the elevator as a random sample of people on campus.
  - Suppose that the weight of students, faculty, and staff is normally distributed with:
    - Mean weight (μ) = 150 lbs
    - Standard deviation (σ) = 27 lbs
- **Question:** What is the probability that the total weight of a random sample of 16 people in the elevator will exceed the weight limit of 2500 lbs?

Sampling distributions

- **Given Data:**
  - Sample size (n) = 16 people
  - Mean weight (μ) = 150 lbs
  - Standard deviation (σ) = 27 lbs
  - Weight limit = 2500 lbs
- $\bar{X} = \frac{1}{16}(X_1 + X_2 + \cdots + X_{16})$
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \bar{X} \sim N\left(150, \frac{27^2}{16}\right)$
- The total weight of 16 people exceeding 2500 lbs means the sample mean $\bar{X}$ exceeds:

$$\bar{X} > \frac{2500}{16} = 156.25 \text{ lbs}$$

Sampling distributions

- **Given Data:**
  - Sample size (n) = 16 people
  - Mean weight (μ) = 150 lbs
  - Standard deviation (σ) = 27 lbs
  - Weight limit = 2500 lbs
- $\bar{X} = \frac{1}{16}(X_1 + X_2 + \cdots + X_{16})$
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \bar{X} \sim N\left(150, \frac{27^2}{16}\right)$

Sampling distributions

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \bar{X} \sim N\left(150, \frac{27^2}{16}\right)$
- The total weight of 16 people exceeding 2500 lbs means $\bar{X}$ exceeds:

$$\bar{X} > \frac{2500}{16} = 156.25 \text{ lbs}$$

- Convert $\bar{X}$ to a standard normal variable Z:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{156.25 - 150}{6.75} = \frac{6.25}{6.75} \approx 0.9259$$

- Using the Z-table:

$$P(Z > 0.9259) = 1 - P(Z \leq 0.9259) = 1 - 0.8222 = 0.1778$$

Central Limit Theorem (CLT)

Central Limit Theorem (CLT)

- **What is the Central Limit Theorem?**
  - It states that the distribution of the sample mean (or sum) of a large number of *i.i.d.* random variables approaches a normal distribution.
  - This holds true regardless of the original population distribution, provided the sample size is sufficiently large.
- **Key Components:**
  - i.i.d. RVs
  - Large Sample Size (typically n≥30).
  - Normal Distribution of Sample Mean ($\bar{X}$)

Central Limit Theorem (CLT)

- **Mathematical Statement of the CLT**
- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with mean μ and standard deviation σ.

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$$

$f(x)$ could be any distribution

- The sample mean $\bar{X}$ is given by:

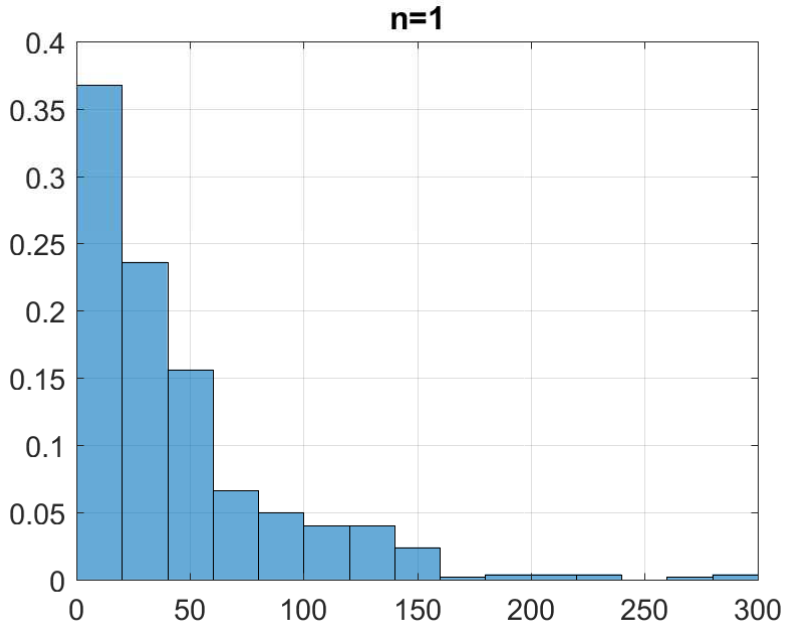
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

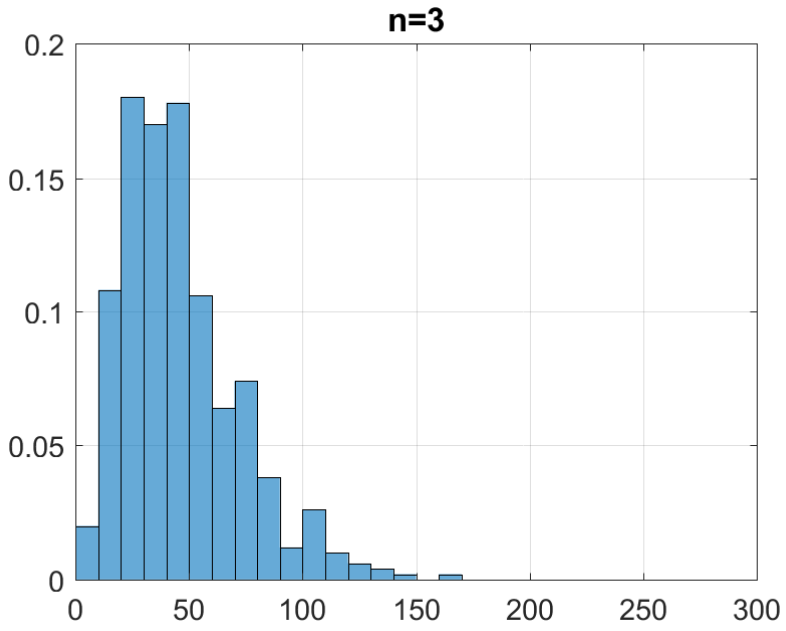
- As n approaches infinity:

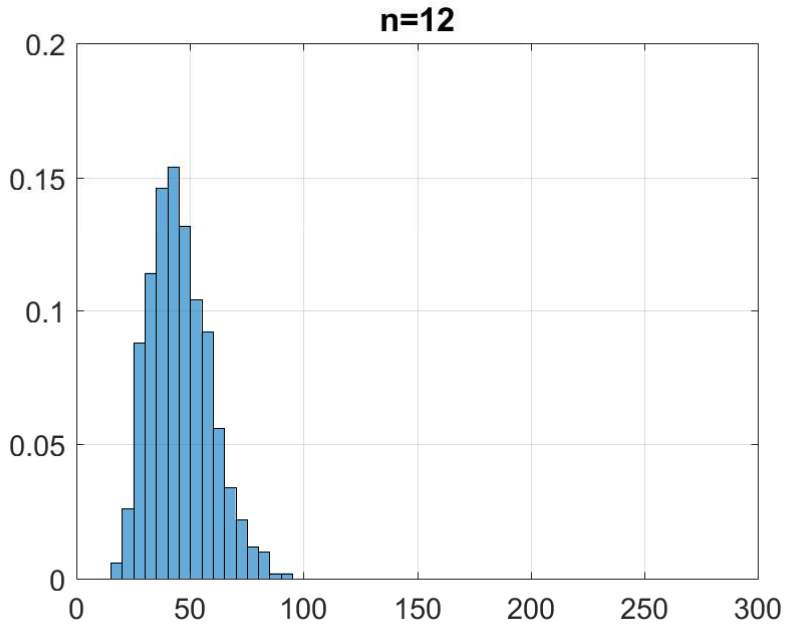
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

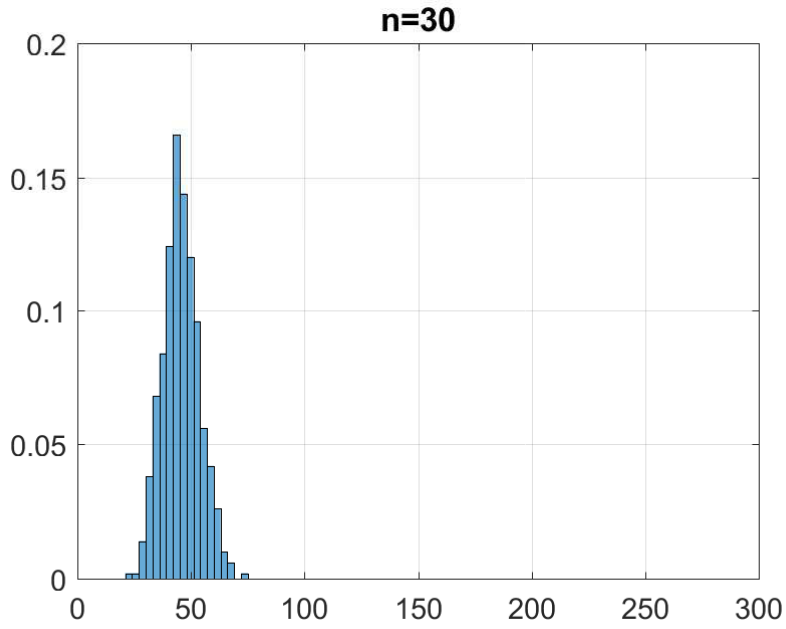
Sampling distributions

- **Example 0.7.**
- Suppose salaries of all UCD employees follow an exponential distribution with average salary = 45K (which means that $\lambda = \frac{1}{45}$)
- We display the histograms of the simulated values of $\bar{X}$ through 500 repetitions for each of $n = 1, 3, 12, 30$.


n=1
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300


n=3
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300


n=12
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300


n=30
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300

Summary

Basic concepts

- **Population:** set of all individuals (whose certain characteristic is of interest)
- **Sample:** a subset of the population (to be measured)
- **Random sample:** a collection of random variables $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x)$, where $f(x)$ represents the PMF/PDF of the population
- **Statistic:** a numerical summary of the sample, such as $\bar{X}, S^2$

Summary

- **Sampling distribution of a statistic:** probabilistic distribution of the statistic as a random variable
- The sample mean statistic: For any random sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x)$, define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- If the population distribution $f(x)$ has mean μ and variance σ^2 , then

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Summary

- Sampling distributions of $\bar{X}$
  - If the population is normal ($N(\mu, \sigma^2)$), then the sample mean has the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  - **Central limit theorem (CLT):** For non-normal populations, if the sample size is large (i.e., n ≥ 30), then

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$