<u>*FUN*</u>damentals of Biostatistics (Lectures 1-4)

BIOSTAT 201A Fall 2025

Discussion 2 – October X, 2025

Cindy J. Pang

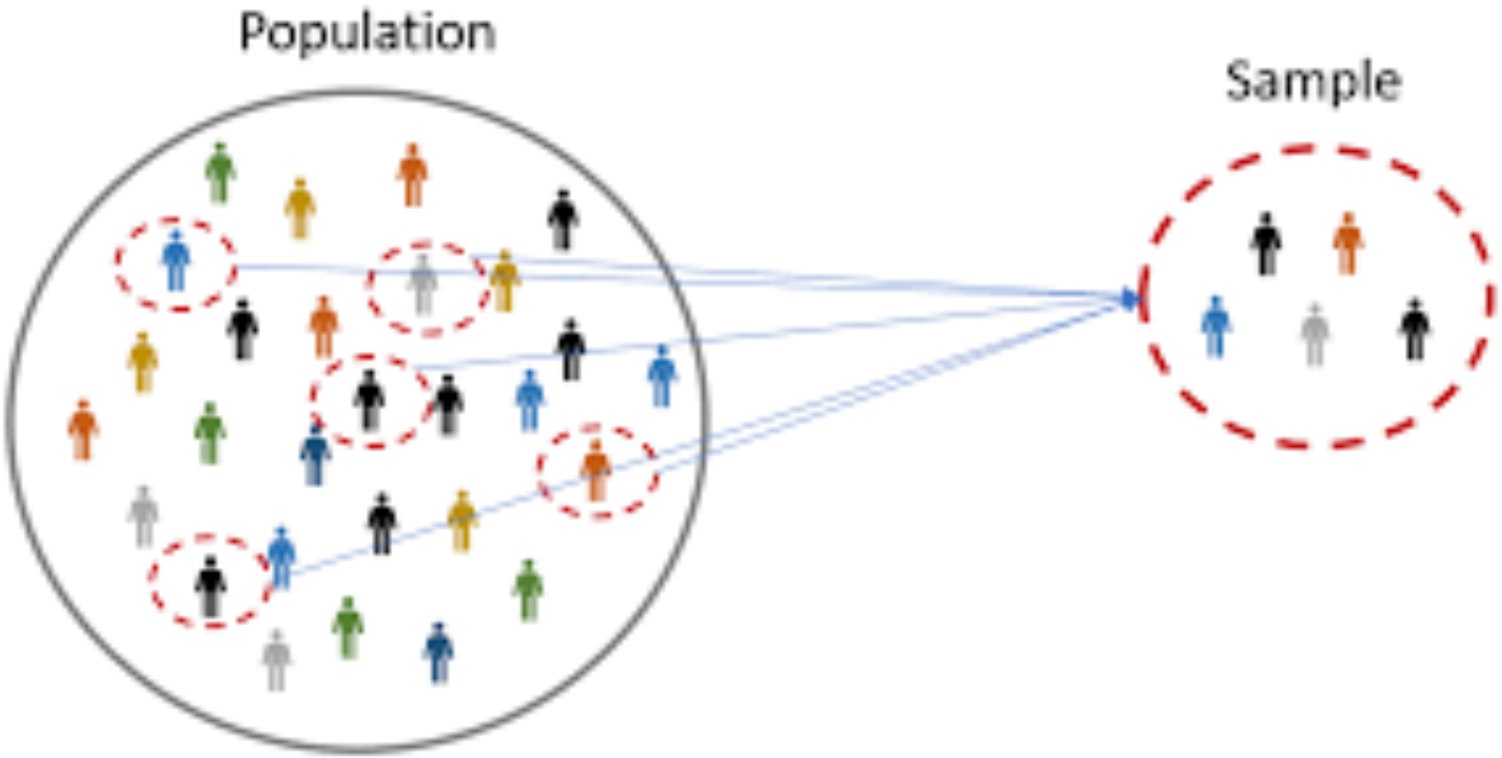
Outline

1. Lecture 1: Biostatistics, Inference, and Data
    1. Population vs. Sample, Parameter vs. Statistic
    2. Types of Data
2. Lecture 2: Parameters vs. Statistics, Summary Statistics, and Data Visualization (See Problem 1)
    1. Measures of Location
    2. Measures of Spread
    3. Describing a Distribution
    4. Data Visualization - Boxplots
3. Lecture 3: Probability, Conditioning, and Independence (see Problems 2 and 3)
4. Lecture 4: Random Variables, Sampling, and The Central Limit Theorem
    1. Discrete Random Variables
    2. Properties of the Normal Distribution
    3. Central Limit Theorem

Outline

1. Lecture 1: Biostatistics, Inference, and Data
    1. Population vs. Sample, Parameter vs. Statistic
    2. Types of Data
2. Lecture 2: Parameters vs. Statistics, Summary Statistics, and Data Visualization (See Problem 1)
    1. Measures of Location
    2. Measures of Spread
    3. Describing a Distribution
    4. Data Visualization - Boxplots
3. Lecture 3: Probability, Conditioning, and Independence (see Problems 2 and 3)
4. Lecture 4: Random Variables, Sampling, and The Central Limit Theorem
    1. Discrete Random Variables
    2. Properties of the Normal Distribution
    3. Central Limit Theorem

1.1 Population vs. Sample, Parameter vs. Statistic



1.2 Types of Data

(1) **Quantitative (“Quantifiable”)** – discrete or continuous response of measurable magnitude
  - Continuous – can take on an uncountable set of values
  - Discrete – can take on a limited, and often fixed number of possible values

(2) **Qualitative –** no corresponding set of numbers

Can you produce some examples of each type of data?

1.2 Types of Data

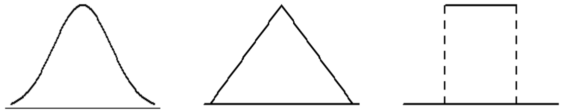
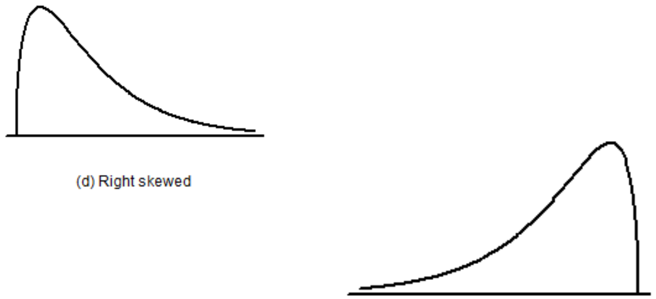
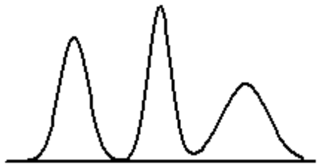
- Stevens' Scale of Data Classification

	Categorical	Order	Equal Spacing	True Zero	Examples
Nominal – categorical data without any sensible ordering of values	X				
Ordinal – categorical or discrete data with a natural ranking/ordering relationship	X	X			
Interval – has a notion of distance, defined by measurement units of equal size	X	X	X		
Ratio – has a notion of magnitudes or "x times greater"	X	X	X	X	

Outline

1. Lecture 1: Biostatistics, Inference, and Data
    1. Population vs. Sample, Parameter vs. Statistic
    2. Types of Data
2. Lecture 2: Parameters vs. Statistics, Summary Statistics, and Data Visualization (See Problem 1)
    1. Measures of Location
    2. Measures of Spread
    3. Describing a Distribution
    4. Data Visualization - Boxplots
3. Lecture 3: Probability, Conditioning, and Independence (see Problems 2 and 3)
4. Lecture 4: Random Variables, Sampling, and The Central Limit Theorem
    1. Discrete Random Variables
    2. Properties of the Normal Distribution
    3. Central Limit Theorem

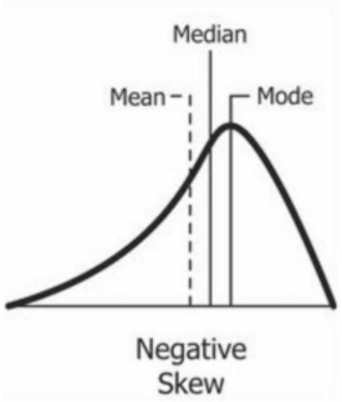
2.1 Measures of Location

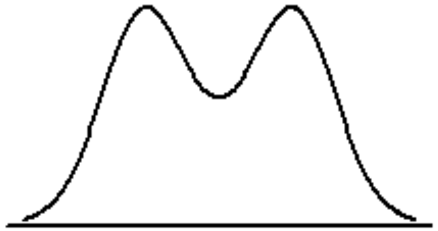
	Formula/How to Calculate it	What the Data Looks Like
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	 (a) Bell shaped (b) Triangular (c) Uniform (retangular)
Median	Let observations $x_1, \ldots, x_n$ be ordered such that $x_1 \leq x_2 \leq \cdots \leq x_n$ where x_1 is the minimum and x_n is the maximum If n is even, the median is the average of $x_{\frac{n}{2}}$ and $x_{\frac{n+1}{2}}$ If n is odd, the median is $x_{\frac{n+1}{2}}$	 (d) Right skewed (e) Left skewed
Mode	Most common value	 (i) Multi-modal

2.2 Measures of Spread

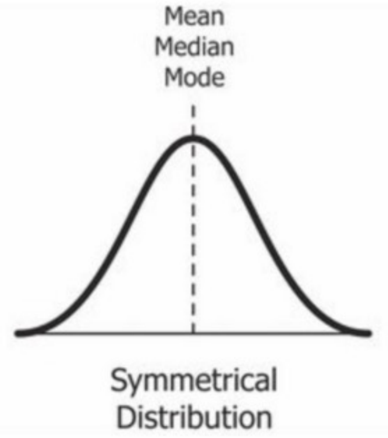
	Formula/How to Calculate it
Range	Range = $x_{max} - x_{min}$
Interquartile Range (IQR)	IQR = $x_{75th\ percentile} - x_{25th\ percentile} = Q3 - Q1$
Sample Variance/Standard deviation ($\sqrt{s^2}$)	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

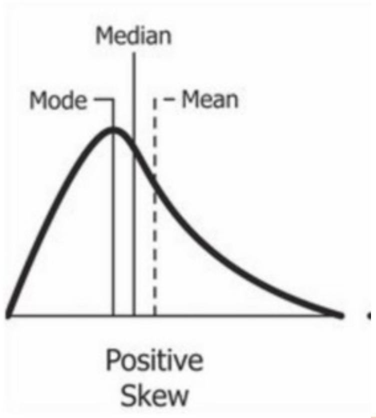
2.3 Describing a Distribution



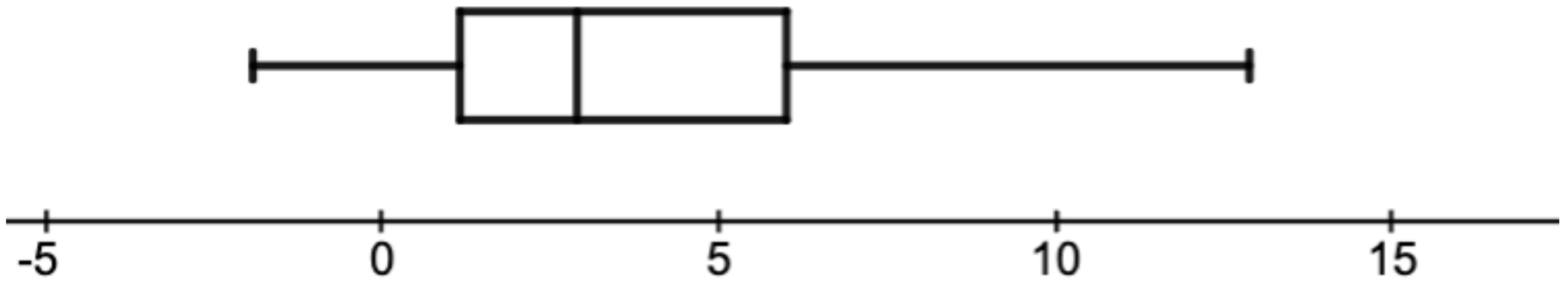


(h) Bimodal





2.4 Data Visualization - Boxplots



2.4 Data Visualization – Stem and Leaf

15, 16, 21, 23, 23, 31, 35, 39, 40, 41, 46, 55, 56

Outline

1. Lecture 1: Biostatistics, Inference, and Data
    1. Population vs. Sample, Parameter vs. Statistic
    2. Types of Data
2. Lecture 2: Parameters vs. Statistics, Summary Statistics, and Data Visualization (See Problem 1)
    1. Measures of Location
    2. Measures of Spread
    3. Describing a Distribution
    4. Data Visualization - Boxplots
3. Lecture 3: Probability, Conditioning, and Independence (see Problems 2 and 3)
4. Lecture 4: Random Variables, Sampling, and The Central Limit Theorem
    1. Discrete Random Variables
    2. Properties of the Normal Distribution
    3. Central Limit Theorem

Some general probability rules

- Complements: $\mathrm{P}(\overline{\mathrm{E}}) = 1 - \mathrm{P}(\mathrm{E})$
- Independence of Events:
  (1) Event A is independent of Event B if the joint probability of A and B can be expressed as:
  $$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B})$$
  (2) Event A is independent of Event B is the following conditional holds:
  $$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \mathrm{P}(\mathrm{A}),\ \mathrm{P}(\mathrm{B}) > 0$$
- Conditional Probability: $\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{B})}{\mathrm{P}(\mathrm{B})} = \frac{\mathrm{P}(\mathrm{A}|B)P(B)}{\mathrm{P}(\mathrm{B})}$
- Rule of Total Probability: $\mathrm{P}(\mathrm{A}) = \mathrm{P}(\mathrm{A}|\mathrm{B})\mathrm{P}(\mathrm{B}) + \mathrm{P}(\mathrm{A}|\overline{\mathrm{B}})\mathrm{P}(\mathrm{B})$

Outline

1. Lecture 1: Biostatistics, Inference, and Data
    1. Population vs. Sample, Parameter vs. Statistic
    2. Types of Data
2. Lecture 2: Parameters vs. Statistics, Summary Statistics, and Data Visualization (See Problem 1)
    1. Measures of Location
    2. Measures of Spread
    3. Describing a Distribution
    4. Data Visualization - Boxplots
3. Lecture 3: Probability, Conditioning, and Independence (see Problems 2 and 3)
4. Lecture 4: Random Variables, Sampling, and The Central Limit Theorem
    1. Discrete Random Variables
    2. Properties of the Normal Distribution
    3. Central Limit Theorem

4.1 Discrete Random Variables - Example

Probability Mass Function (PMF) – probability distribution for a discrete variable

Value	**1**	**2**	**4**	**6**
Probability	0.2	0.1	0.4	0.3

Exercise. Compute

(a) Expected Value of X:

(b) Variance of X:

(c) Construct the cumulative distribution function (cdf)

Expected Value of a Discrete RV:

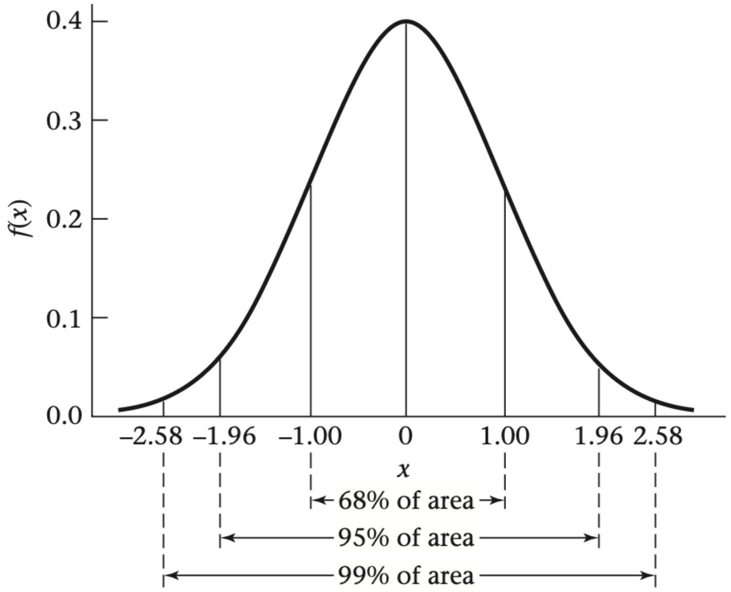
$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \Pr(X = x_i)$$

Variance of a Discrete RV:

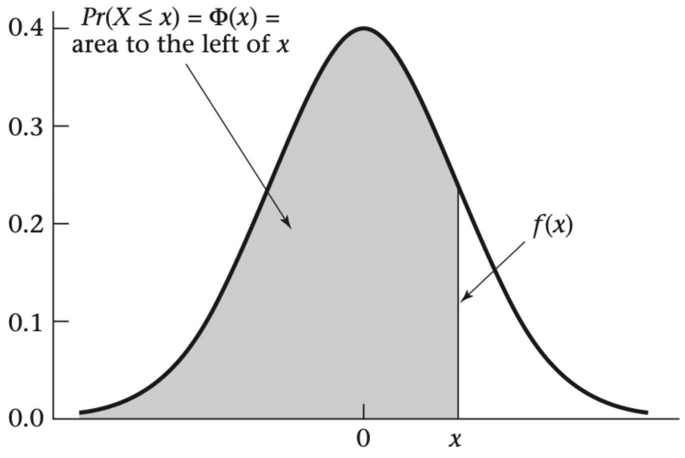
$$Var(X) = \mathbb{E}(X - \mu)^2 = \sum_{i=1}^{n} x_i^2 \Pr(X = x_i) - \mu^2$$

4.2 Properties of the Normal Distribution – The Empirical Rule and CDF

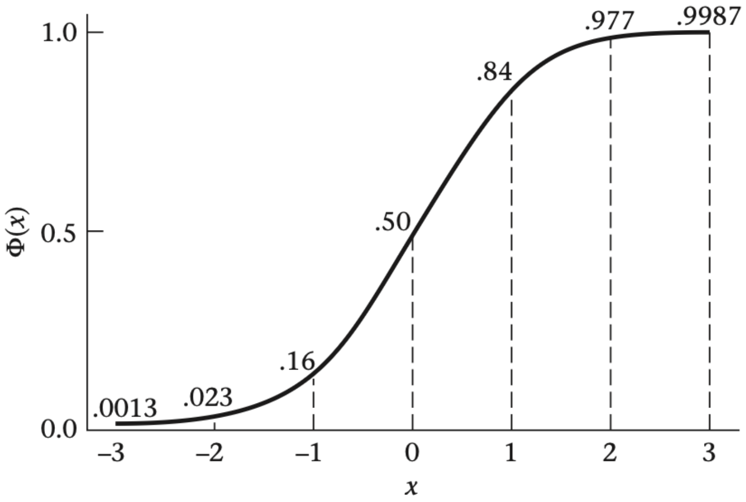
Empirical properties of the standard normal distribution



The cdf [Φ(x)] for a standard normal distribution



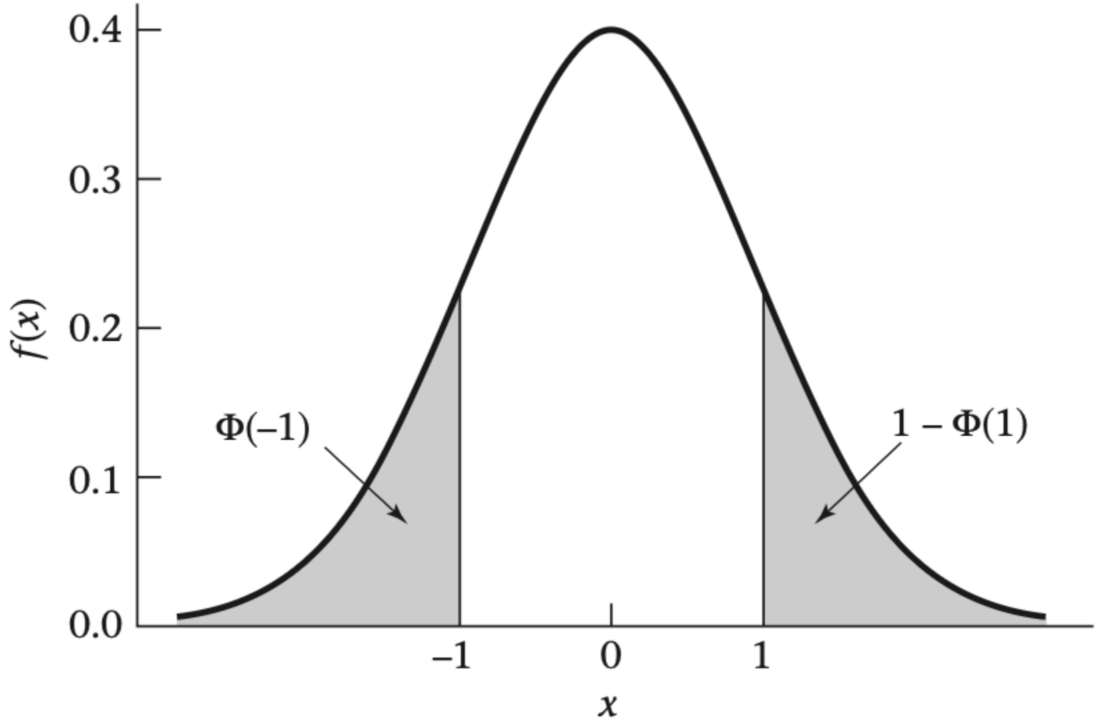
The cdf for a standard normal distribution [Φ(x)]



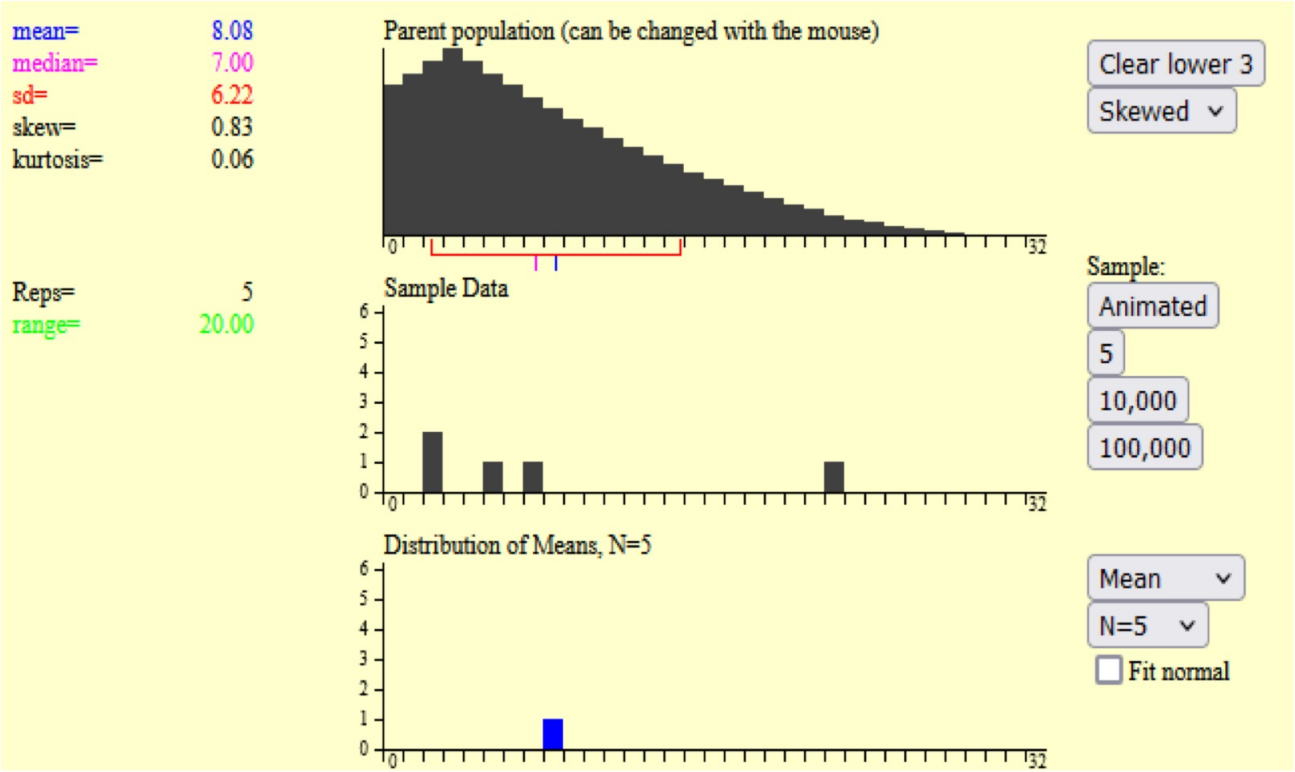
4.2 Properties of the Normal Distribution - Symmetry

$$\Phi(-x) = 1 - \Pr(X \leq x) = 1 - \Phi(\mathrm{x})$$

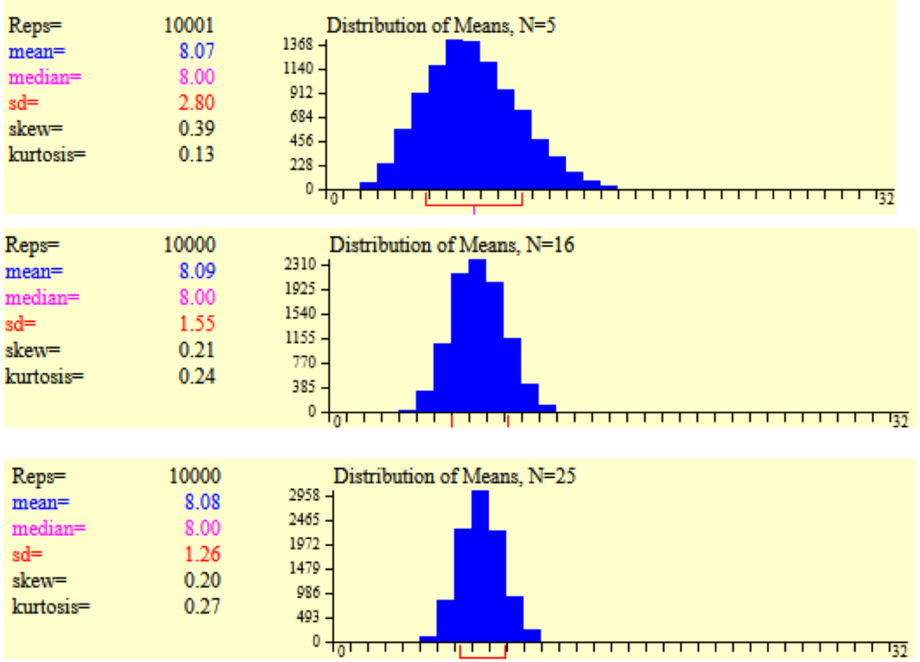
Illustration of the symmetry properties of the normal distribution



4.3 Central Limit Theorem - Illustration



10,000 iterations later…



What do you notice?

- About the Shape?
- What happens as we increase the sample size?

Simulation Link: https://onlinestatbook.com/stat_sim/sampling_dist/