

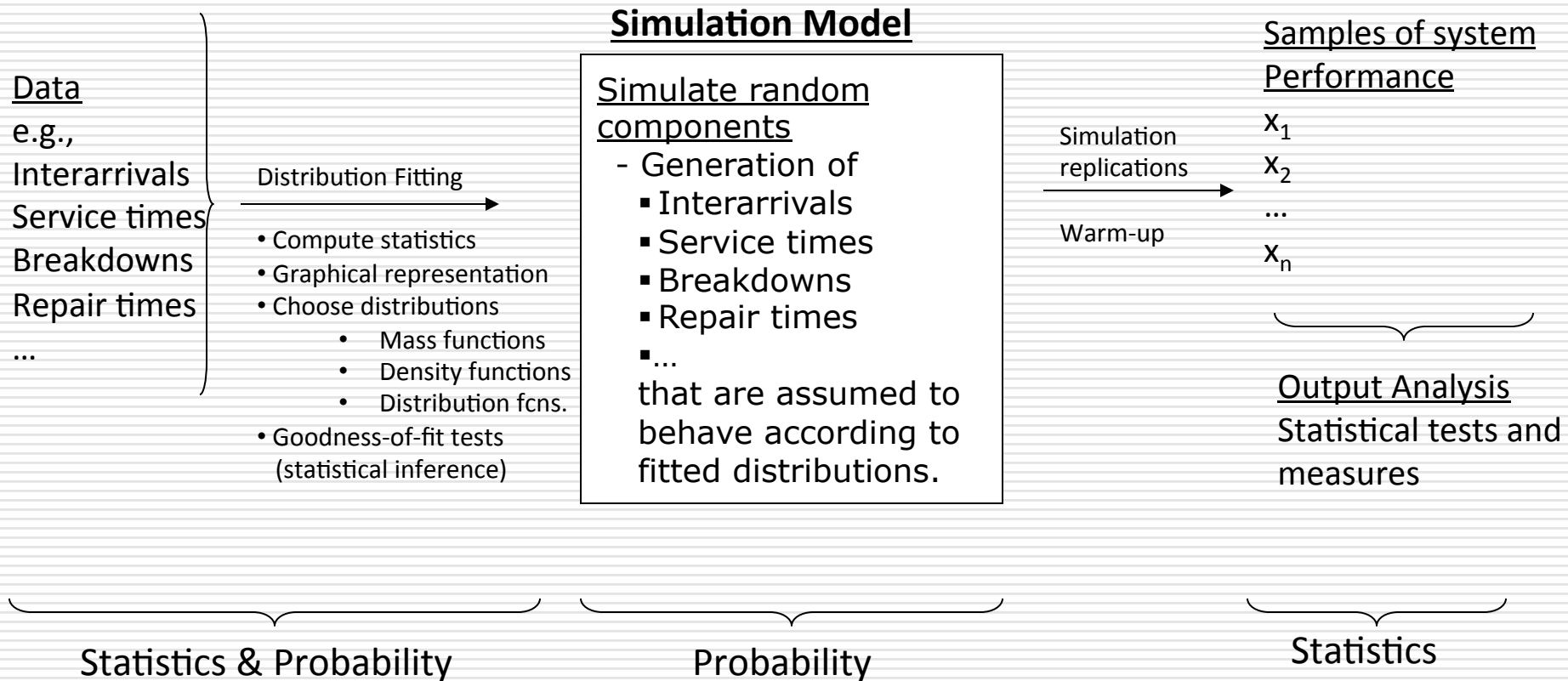
# Review of Basic Probability and Statistics Concepts

---

# Probability, Statistics, and Simulation

“Statistics” – processing/analyzing data

“Probability” – Manipulating/utilizing functions characterizing uncertainty



# Review Outline

---

- Random variables.
- Basic statistics and graphs computed from collected data.
- Probabilistic characterization of random variables – Mathematical model.
- Statistical inference.

# Random Variable

---

- Engineering definition - A quantity of interest whose exact value is unpredictable.
- Notation

# Characterizing Random Variables – Describing Variation

---

1. Start with data – no assumptions made
  - Numerical and graphical descriptions
2. Describe the “behavior” of the random variable with mathematical models (functions)
  - What numerical quantities are used to characterize unpredictability?
  - The function used is a model
    - Justify/verify that it is a good model with data

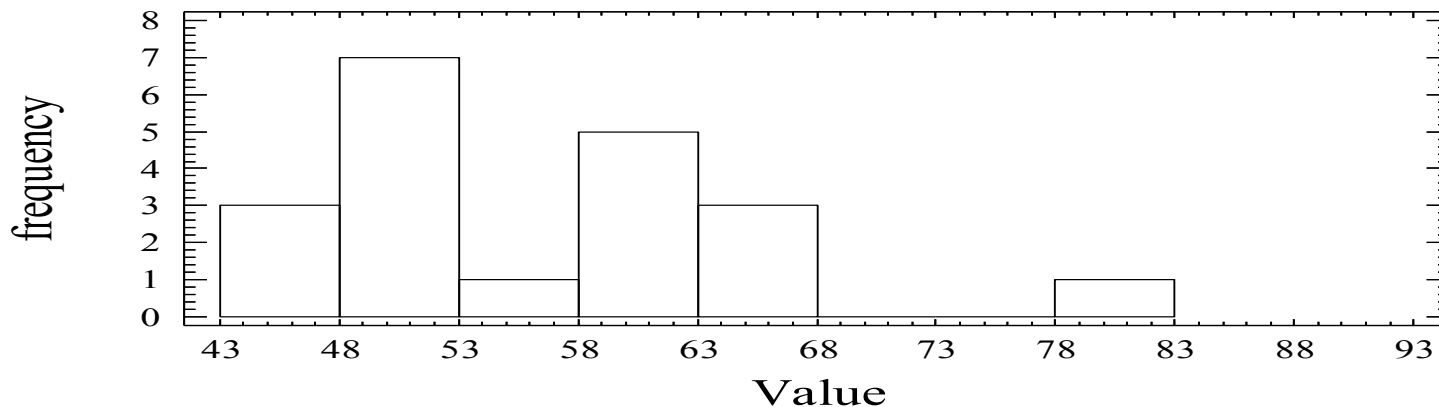
# Statistics and Graphical Representations of Data

---

- ☐ No assumptions made – data collected and processed
  - Graphical methods
    - ☐ Histogram
    - ☐ Box Plot
  - Statistics
    - ☐ Central tendency
    - ☐ Variability
    - ☐ Other numerical characterizations

# Histogram

- Graph of observed frequencies vs. value



- Visual display of
  - Shape
  - Location or central tendency
  - Scatter or spread

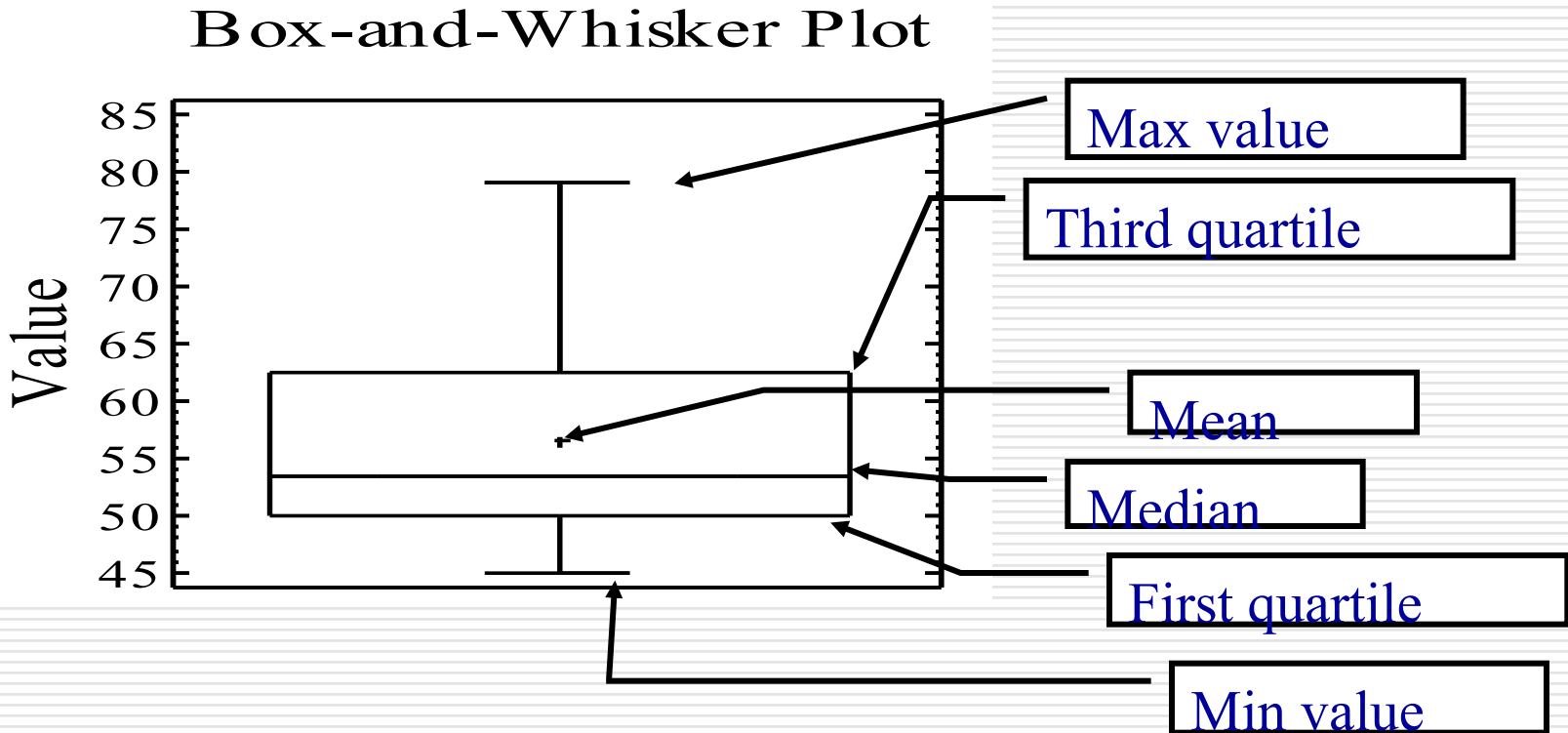
# Histogram

---

- ❑ Decisions to make when constructing a histogram.
- ❑ # of intervals (  $\approx \sqrt{n}$  ).
- ❑ Minimum, Maximum ( $\pm\infty$  OK).
- ❑ Can affect the visual impression.
- ❑ We'll manually construct these later.



# Box (and Whisker) Plot



- Visual display of
  - Central tendency, Variability, Departure from symmetry, Outliers

# Numerical Characterization: Values Computed from Data = Statistics

---

## ☐ Measures of central tendency:

- ☐ Sample Average (Sample mean)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ☐ Median (50th percentile)
- ☐ Mode (Most frequent)

# Numerical Characterization: Values Computed from Data = Statistics

---

- Measures of variability reflected in the data
  - Sample variance

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


- Sample standard deviation:

$$s = \sqrt{s^2}$$

- Sample Range:

$$R = x_{\text{largest value}} - x_{\text{smallest value}}$$

# Example



# Things to Notice

---

- The sample average and sample variance are random variables
  - They are “quantities of interest whose exact value is unpredictable”
- The sample variance is not affected by the location/magnitude of data, only by scatter about the sample average.

# Coefficient of Variation

---

- Measure of relative variability
  - Coefficient of variation (for X continuous) –  $CV(X)$ 
    - $CV(X) = \text{Std Dev}(X)/\text{Mean}(X)$  – Typically estimated from data
  - $X \sim \text{Exponential}$  –  $CV = 1$ .
  - $X \sim U(a, b)$  –

$$CV(X) = \frac{b - a}{\sqrt{3}(b + a)}$$

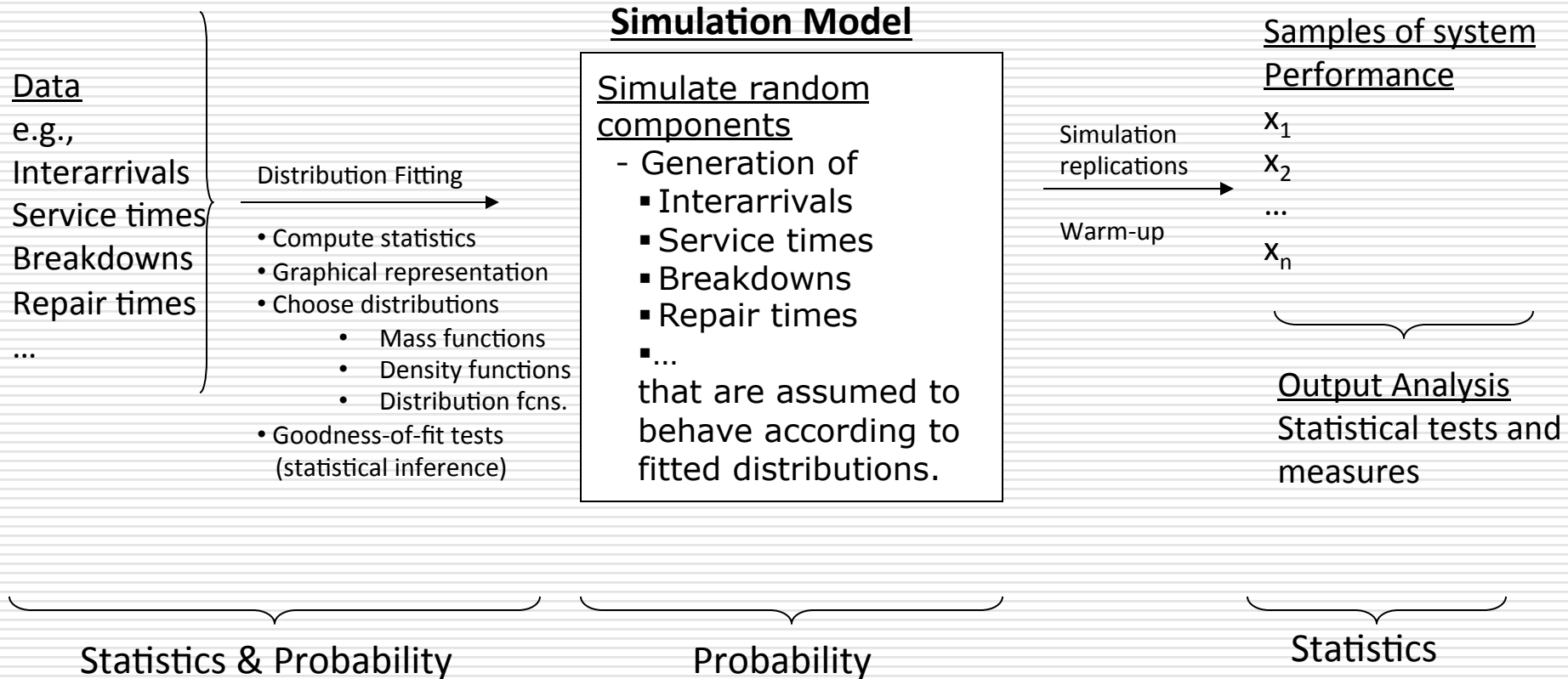
- $X \sim \text{Lognormal}$  –  $0 < CV < \infty$ .

# Numerical Summary of Data

---

- Many other numerical measures.
  - E.g., skewness, kurtosis, ...

# Connection to Simulation





# Connection to Simulation

---

- Histograms are used as a basis for suggesting a probability model for random behavior.
- Statistics computed from data are used to estimate parameters of the probability models.

# Probabilistic Descriptions/Characterization of Random Variables

---

- Mathematical models – functions.

# Mathematical Models

---

- You have seen a variety of mathematical models – exact outputs for given inputs
  - Thermodynamics, statics/dynamics, engineering economy

# Mathematical Models of Random Variables

---

- Mathematical models of unpredictable phenomena are different
  - If the model is accurate, it is accurate over many observations of the phenomena
  - Examples

# Mathematical Models of Random Variables

---

- The uncertain behavior of a random variable is expressed in a mathematical function (this function is the model) - Functional characterization of random variables
  - There are different types of functions utilized

# Mathematical Models of Random Variables

---

- How do we describe uncertain behavior?
  - i.e., What can be calculated from the mathematical functions describing uncertain behavior?

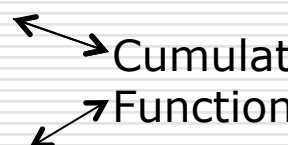
# Mathematical Models of Random Variables

---

- Two general types of random variables
  - Discrete –
  - Continuous –
  
- Functions - terminology
  - Probability/cumulative distribution functions
  - Density functions
  - Probability mass functions

# Mathematical Models of Random Variables

---

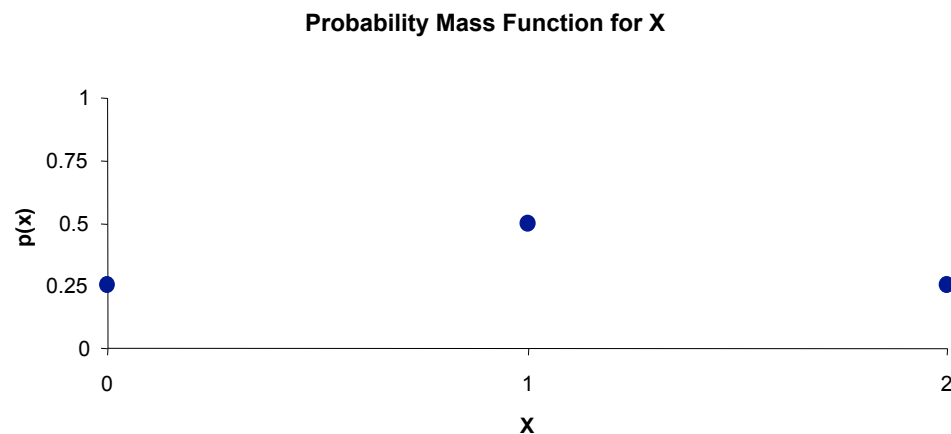
<u>Type of Random Variable</u>	<u>Functions from which probabilities can be found</u>
• Discrete      -	Probability Mass Function 
• Continuous      -	Probability Density Function



# Discrete Random Variable

- Discrete random variable – the possible values of a r.v. are “countable” .
  - E.g. Tossing two coins where the r.v. ***X*** is the number of heads

$$p(x) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 1 \\ 0.25 & x = 2 \\ 0 & \textit{otherwise} \end{cases}$$



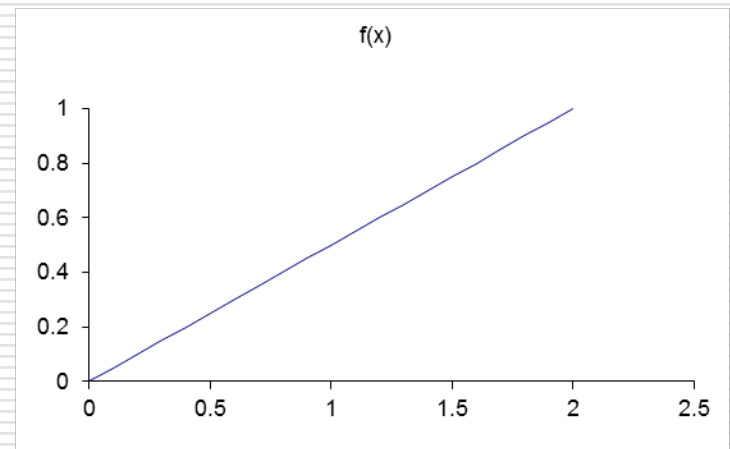
# Continuous Random Variable

- Continuous random variable can take on values in a continuous interval.
  - E.g. let r.v.  $X$  take values in interval  $[0, 2]$

$$f(x) = 0.5x \quad 0 \leq x \leq 2$$

$$\begin{aligned} P[1 \leq x \leq 1.5] &= \int_1^{1.5} 0.5x \, dx = 0.25x^2 \Big|_1^{1.5} \\ &= 0.56 - 0.25 = 0.31 \end{aligned}$$

Density Function



# Probability/Cumulative Distribution Functions

---

- Probability/cumulative distribution function = Probability a random variable takes on a value less than or equal to some value  $x$
- Denoted

$$F(x) = P(X \leq x)$$

# Probability/Cumulative Distribution Functions

---

- For discrete random variables
  - $F(x)$  is a summation of  $p(x)$
- For continuous random variables
  - $F(x)$  is an integration of  $f(x)$

# Probability/Cumulative Distribution Functions


---

## □ Examples


$$p(x) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 1 \\ 0.25 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = 0.5x \quad 0 \leq x \leq 2$$

# Example



# Example



# Connection to Simulation

---

- Generating observations from a probability model of some random phenomena (probability model = probability distribution) is based on the:
  - Probability/cumulative distribution functions
  - Density functions
  - Probability mass functions



# Commonly Used Random Variables

---

- Some random variables are often used as models of unpredictable phenomena – They have names and known functional forms for their mass/density functions and/or their cumulative distribution function.

# Commonly Used Discrete Distributions

---

- ☐ Hypergeometric
- ☐ Binomial
- ☐ Poisson
- ☐ Pascal/Geometric

# Commonly Used Continuous Distributions

---

- ☐ Normal
- ☐ Lognormal
- ☐ Exponential
- ☐ Chi-Square
- ☐ t-distribution
- ☐ F-distribution
- ☐ ...

# Statistical Inference

---

- Suppose we have a data– a sample,  $x_1, x_2, \dots, x_n$ , with values that are assumed independent, and are assumed to be realizations from the same distribution.
  - E.g., results from simulation experiments.
- What conclusions can we make from this data?

# Statistical Inference

---

- In statistical inference, both numerical characterizations of data and probability models (distributions) are being utilized to answer the last question.
  - Numerical characterization - statistics computed from data.
  - Probability model - sampling distributions.

# Sample Statistics

---

- Sample statistic – computed from data
  - e.g.  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  Is the sample mean
  - Point estimator – A statistic that is a single numerical estimate of a distribution parameter.
    - e.g. The mean is a parameter of a normal distribution.
- Do you expect the sample mean (or other sample statistics) to be the same the next time data is collected?

# Sampling Distribution

---

- Sampling Distribution – Probability distribution of a statistic.
  - A statistic (e.g., sample mean) is an observation of a random variable.

# Sampling Distribution

---

- If the distribution (probability model) of the random variable from which the sample was taken is known (an assumption), it may be possible to determine the distribution of various statistics computed from the sample.



# The Average of Observations from a Normal Distribution

---

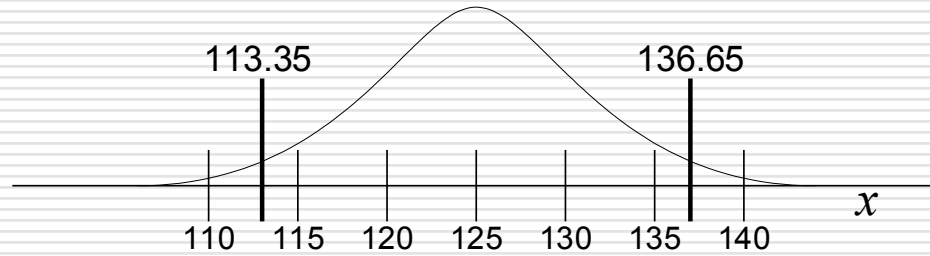
- ❑ Take a random sample,  $x_1, x_2, \dots, x_n$ , from a normal distributed random variable with mean  $\mu$  and standard deviation  $\sigma$ , i.e.,  $N(\mu, \sigma)$
- ❑ Compute the sample average  $\bar{x}$
- ❑ Then  $\bar{x}$  is an observation from a normal distribution with mean  $\mu$  and std dev  $\frac{\sigma}{\sqrt{n}}$
- ❑ That is  $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ . The sample average is an observation of a different random variable.

# Example

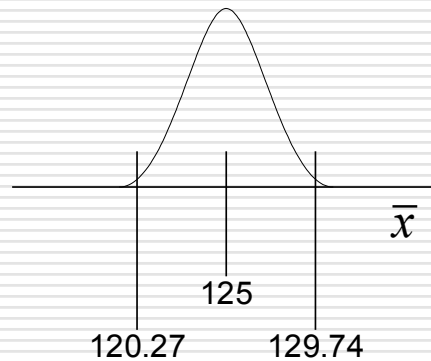
---

- When a process is operating properly, the density of a liquid is normally distributed with a mean of 125 and standard deviation of 5. Ten observations are taken and the average density is 150.
- What is the distribution of the sample average?
- What is the probability a sample average will be greater than or equal to 150?

# Examine the Sampling Distribution



$$r.v. \ x \sim N(\mu = 125, \sigma = 5)$$



$$r.v. \ \bar{x} \sim N(\mu = 125, \sigma_{\bar{x}} = 5/\sqrt{10} = 1.58)$$

# Question

---

- Why do various statistical tests that you have encountered use “z values”, “t values”, “Chi-square values”, “F values”?

# Answer

---

- The “z values”, “t values”, “Chi-square values”, “F values” are statistics computed from data.
- Assuming the data are observations from normal distributions, the sampling distributions of these statistics is known.

# Other Sampling Distributions

---

- If  $x_1, \dots, x_n$  is a random sample from a  $N(\mu, \sigma)$  distribution and  $s^2$  is the sample variance, then the sampling distribution of  $y = \frac{(n-1)s^2}{\sigma^2}$  is  $\chi_{n-1}^2$

# Other Sampling Distributions

- If  $x_1, \dots, x_n$  is a random sample from a  $N(\mu, \sigma)$  distribution and  $\bar{x}$  and  $s^2$  are the sample mean and variance, then

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

is an observation from a t-distribution with  $n-1$  degrees of freedom.

# Other Sampling Distributions

- If  $x_{11}, \dots, x_{1n_1}$  is a random sample from a  $N(\mu_1, \sigma_1^2)$  distribution and  $x_{21}, \dots, x_{2n_2}$  is a random sample from a  $N(\mu_2, \sigma_2^2)$  distribution and  $s_1^2$  and  $s_2^2$  are their sample variances, then the sampling distribution of  $\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$  is  $F_{n_1-1, n_2-1}$



# Other Sampling Distributions

---

- Based on the central limit theorem, if the number of samples  $n$ , is sufficiently large, then  $\bar{x}$  will be approximately normally distributed (regardless of the distribution of  $x_1, x_2, \dots, x_n$ ).

# Statistical Inference

---

- ❑ Statistics are computed from sample data.
- ❑ The sampling distribution is known (based on assumptions).
- ❑ The sampling distribution can then be used to compute information about the data source (random variable ).

# Statistical Inference Procedures

---

- Hypothesis testing & confidence intervals
  - In their construction, they use the same information.
  - They are consistent with each other.

# Hypothesis Tests

---

- A ***Hypothesis*** is a statement about the parameters of a distribution that is evaluated with collected data.



# Hypothesis Tests

---

## ☐ General procedure

- State the null hypothesis  $H_0$  and appropriate alternative hypothesis  $H_1$ .
- Choose a significance level  $\alpha$  (probability of type I error – covered later).
- Select the appropriate test statistic and compute the value of the test statistic from the sample data.
  - ☐ Based on the hypothesis and assumptions/knowledge about the distribution of the random variable.
- Determine the chance of obtaining the test statistic.
  - ☐ Based on the distribution of the test statistic (sampling distribution).

# Statistical Inference

More Information Needed	Underlying Dist. Of Random Variable	Variance of Random Variable	Test Statistic Sampling Distribution
Mean	Normal or unknown with large enough sample size	Known	Z (std. normal)
	Normal	Unknown	t-distribution
Variance	Normal	N/A	Chi-Square
Population Proportion	Bernoulli/Binomial	Based on proportion tested	Z (std. normal)
Difference in mean	Normal or other/unknown with large enough sample size	Known	Z (std. normal)
Difference in mean	Normal	Unknown (Equal/Non-equal)	t-distribution
Equality of variance	Normal	Unknown	F-distribution
Equality of Population Proportion	Bernoulli/Binomial	Unknown	Z (std. normal)

## EXAMPLE 4.3 Rubberized Asphalt

Rubber can be added to asphalt to reduce road noise when the material is used as pavement. Table 4.1 shows the stabilized viscosity (cP) of 15 specimens of asphalt paving material. To be suitable for the intended pavement application, the mean stabilized viscosity should be equal to 3200. Test this hypothesis using  $\alpha = 0.05$ . Based on experience we are willing to initially assume that stabilized viscosity is normally distributed.

■ **TABLE 4.1**  
**Stabilized Viscosity of Rubberized Asphalt**

Specimen	Stabilized Viscosity
1	3193
2	3124
3	3153
4	3145
5	3093
6	3466
7	3355
8	2979
9	3182
10	3227
11	3256
12	3332
13	3204
14	3282
15	3170

# Example

To be suitable for the intended pavement application, the mean stabilized viscosity should be equal to 3200. Suppose that we wish to test this hypothesis using  $\alpha = 0.05$ . Based on experience we are willing to initially assume that stabilized viscosity is normally distributed. The appropriate hypotheses are

$$H_0: \mu = 3200$$

$$H_1: \mu \neq 3200$$


The sample mean and sample standard deviation are

$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{48,161}{15} = 3210.73$$

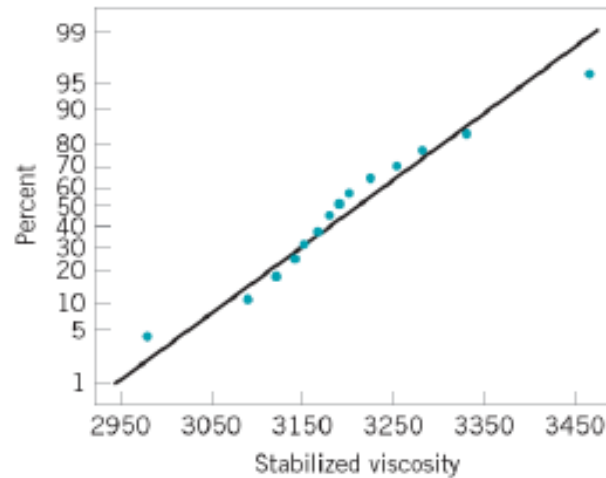
$$s = \sqrt{\frac{\sum_{i=1}^{15} x_i^2 - \frac{\left(\sum_{i=1}^{15} x_i\right)^2}{15}}{15 - 1}} = \sqrt{\frac{154,825,783 - \frac{(48,161)^2}{15}}{14}} = 117.61$$



# Example



# Example




■ **FIGURE 4.5** Normal probability plot of the stabilized viscosity data.

# Example - Comparing Means From Two Samples

---

20 rats were exposed to a predetermined level of ozone for 30 days after which lung volume was measured. The sample average was 9.28 ml with a sample standard deviation of 0.37. A control group of 17 rats had a sample average of 7.97 ml with a sample standard deviation of 0.41. Is there a significant ( $\alpha = 0.01$ ) difference average lung volume?

# Comparing Means From Two Samples

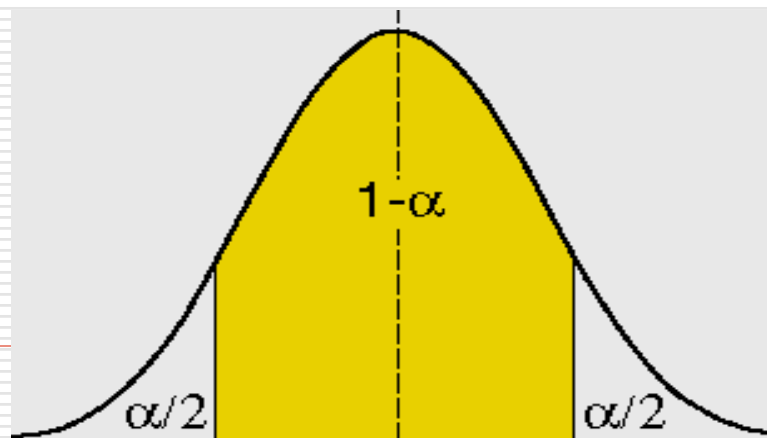


# Statistical Inference

---

## □ Confidence Intervals

- A **confidence interval** for an unknown parameter is an interval that contains a set of plausible values of the parameter. It is associated with a confidence level  **$1 - \alpha$** , which is the percentage of confidence intervals (constructed from random data samples) that contain the unknown parameter.



# Statistical Inference

---

- Confidence intervals for a parameter can be constructed from the same information used for hypothesis testing (for the same parameter and confidence level).

# Example - Confidence Interval

To be suitable for the intended pavement application, the mean stabilized viscosity should be equal to 3200. Suppose that we wish to test this hypothesis using  $\alpha = 0.05$ . Based on experience we are willing to initially assume that stabilized viscosity is normally distributed. The appropriate hypotheses are

$$H_0: \mu = 3200$$

$$H_1: \mu \neq 3200$$

The sample mean and sample standard deviation are

$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{48,161}{15} = 3210.73$$

$$s = \sqrt{\frac{\sum_{i=1}^{15} x_i^2 - \frac{\left(\sum_{i=1}^{15} x_i\right)^2}{15}}{15 - 1}} = \sqrt{\frac{154,825,783 - \frac{(48,161)^2}{15}}{14}} = 117.61$$

# Example - Confidence Interval





# Statistical Testing Errors

---

- Typically presented in the context of hypothesis testing.
  - Equivalent information in confidence intervals.

# Statistical Testing Errors

---

- ■ **Type I Error Level** ( $\alpha$ ) – the probability of rejecting a null hypothesis when it is true.
  - $1 - \alpha$  refers to a confidence level
  
- **Type II Errors** ( $\beta$ ) – the probability of accepting a null hypothesis when it is false.
  - $1 - \beta$  refers to “power”
  
- Can use a confidence interval to test a hypothesis.
  - E.g., Does the interval contain zero?

# Statistical Testing Errors

---

- ***Type I Error Level*** – the percentage of confidence intervals that will not contain the true value of the parameter of interest.

# Connection to Simulation

---

- Simulation results represent outcomes of random variables = results of experiments
  - Statistical inference procedures are utilized to make conclusions from experimental results