# Section 1

## MANE 3332.05

Subsection 1

## Lecture 1, September 2

# Agenda

- Discuss Syllabus
- Me Talk
- Grit Lesson
- Lecture - Chapter 1
- Call roll

# Handouts

- Lecture 1 Slides-pdf
- Lecture 1 Slides - Powerpoint
- Lecture 1 Marked Slides

# Statistics and Statistical Thinking

“The field of **statistics** deals with the collection, presentation, analysis and use of data to make decisions, solve problems, and design products and processes.”

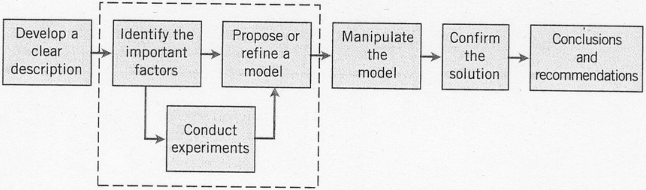

Develop a clear description
Identify the important factors
Propose or refine a model
Conduct experiments
Manipulate the model
Confirm the solution
Conclusions and recommendations

Figure 1-1 The engineering method.

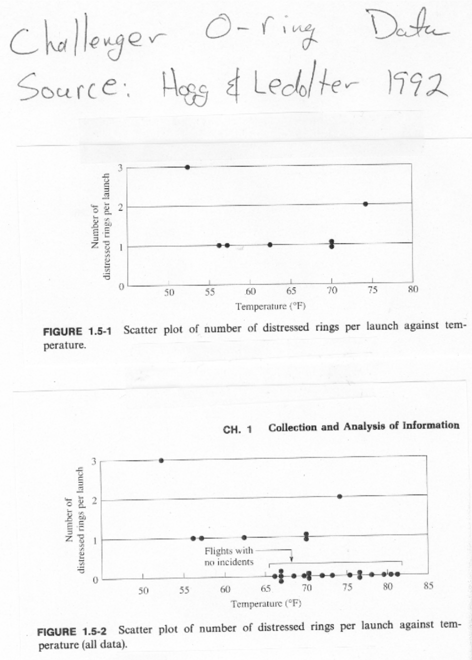
Figure 1: image

# Statistics

- Moore (Ostle, Turner, Hicks and McElrath 1996) defines statistics as "the science of gaining information in the face of uncertainty."
- Generally the field of statistics is divided into two major branches: **descriptive** and **inferential**

# Descriptive Statistics

- Devore and Farnum (1999) define descriptive statistics in the following manner.
  *an investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from* ***descriptive statistics.*** *Some of these methods are graphical in nature–the construction of histograms, boxplots, and scatter plots are primary examples. Other descriptive methods involve calculation of numerical summary measures, such as means, standard deviations and correlation coefficients.*
- Chapter 6 of the textbook emphasize descriptive statistics



Challenger O-ring Data

Source: Hogg & Ledolter 1992

**FIGURE 1.5-1** Scatter plot of number of distressed rings per launch against temperature.

**CH. 1 Collection and Analysis of Information**

**FIGURE 1.5-2** Scatter plot of number of distressed rings per launch against temperature (all data).

# Inferential Statistics

- Devore and Farnum (1999) gives the following description of inferential statistics
  *Having obtained a sample from a population, an investigator would frequently like to use sample information to draw some type of conclusion (make an inference of some sort) about the population. That is, the sample is a means to an end rather than an end in itself. Techniques for generalizing from a sample to a population are gathered within the branch of our discipline called **inferential statistics**.*
- The field of inferential statistics can be further subdivided into two general areas: **estimation** and **hypothesis testing**
- Chapters 4 - 14 of the textbook focus on inferential statistics

# Statistical Thinking

- The textbook points out that statistical methods are used to help us describe and understand variability
- **Variability** is the differences in successive observations of a system or phenomenon
- Vining (1998) gives the following definition of statistical thinking
  *Only by "thinking statistically" can engineers truly address the problems inherent in the variability in real data. When we think statistically, we come to know that all decisions based on real data involve risk and uncertainty. Good decisions require us to quantify this risk. As we become more mature in our thinking, we understand that there are sources or causes of variability. Discovering these sources and removing them are often the keys to engineering success.*

# Population vs. Sample

The concept of populations, samples, parameters and statistics is very important. The following definitions are taken from Ostle, Turner, Hicks and McElrath (1996)

- **Population:** the totality of all possible values (measurements, counts, and so on) of a particular characteristic for a specific group of objects
- **Population parameters:** a numerical descriptive measure of a population characteristics
- **Sample:** a portion of the population that is selected according to some rule or plan
- **Sample statistics:** a numerical descriptive measure of a particular characteristic based upon the sample values

# Enumerative vs. Analytical Studies

- Deming introduced the concept of enumerative versus analytical studies
- **Enumerative study** is one in which a sample is used to make inference on the current population. This is the safest use of statistical estimation.
- **Analytical study** is one in which inference is applied to future populations. There is nothing wrong with this approach. However, you must be aware that there is an inherent assumption of stability

# Data

- Most statistical methods are data-driven
- Data are almost always a sample from a population or populations
- Engineering data are usually collected in 3 ways:
    - A **retrospective study** based on historical data,
    - An **observational study**,
    - A **designed experiment**

# “Happenstance” Data

- Box, Hunter and Hunter (1978) group retrospective studies and observational studies as “happenstance” data
- They point out the following dangers:
  1. Inconsistent data
  2. Range of variables limited by control
  3. Semiconfounding of effects
  4. Nonsense correlation – beware the lurking variable
  5. Serially correlated errors
  6. Dynamic relationships
  7. Feedback
- So why use happenstance data?

# Designed Experiments

- “In a designed experiment, the engineer makes deliberate or purposeful changed in controllable variables (called **factors**) of the system, observes the resulting system output, and then makes a decision or an inference about which variables are responsible for the changed that he or she observes in the output performance.”
- “An important distinction between a designed experiment and either an observational or retrospective study is that the different combinations of the factors of interest are applied randomly to a set of experimental units.”
- Box, Hunter and Hunter (1978) present a table (shown below) that clarifies how designed experiments avoid the problems that occur in the analysis of happenstance data

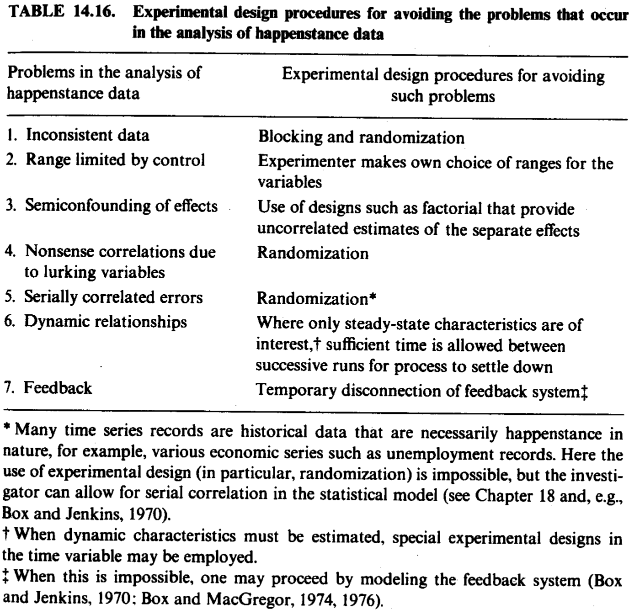


**TABLE 14.16. Experimental design procedures for avoiding the problems that occur in the analysis of happenstance data**

| Problems in the analysis of happenstance data | Experimental design procedures for avoiding such problems |
|---|---|
| 1. Inconsistent data | Blocking and randomization |
| 2. Range limited by control | Experimenter makes own choice of ranges for the variables |
| 3. Semiconfounding of effects | Use of designs such as factorial that provide uncorrelated estimates of the separate effects |
| 4. Nonsense correlations due to lurking variables | Randomization |
| 5. Serially correlated errors | Randomization* |
| 6. Dynamic relationships | Where only steady-state characteristics are of interest,† sufficient time is allowed between successive runs for process to settle down |
| 7. Feedback | Temporary disconnection of feedback system‡ |

* Many time series records are historical data that are necessarily happenstance in nature, for example, various economic series such as unemployment records. Here the use of experimental design (in particular, randomization) is impossible, but the investigator can allow for serial correlation in the statistical model (see Chapter 18 and, e.g., Box and Jenkins, 1970).

† When dynamic characteristics must be estimated, special experimental designs in the time variable may be employed.

‡ When this is impossible, one may proceed by modeling the feedback system (Box and Jenkins, 1970; Box and MacGregor, 1974, 1976).

# Data Collected Over Time

- Often data is collected over time (either retrospective, observational or designed experiments)
- Most elementary statistical techniques assume that the observations are independent (not always a good assumption)
- The correct term for data collected over time is a **time series**
- Time series analysis does not assume that the observations are independent over time

# Models

- “Models play an important part in engineering analysis”
- From the statistical point of view, we will divide models into two categories: mechanistic and empirical
- All models have these characteristics:
    - One or more observed outcomes that we wish to understand or predict
    - These outcomes are referred to as the **response** or **dependent** variable(s)
    - The set of variables (factors) that influence response variables is called the **independent** or **regressor** variables
    - A functional relationship between the dependent and independent variables

# Mechanistic Models

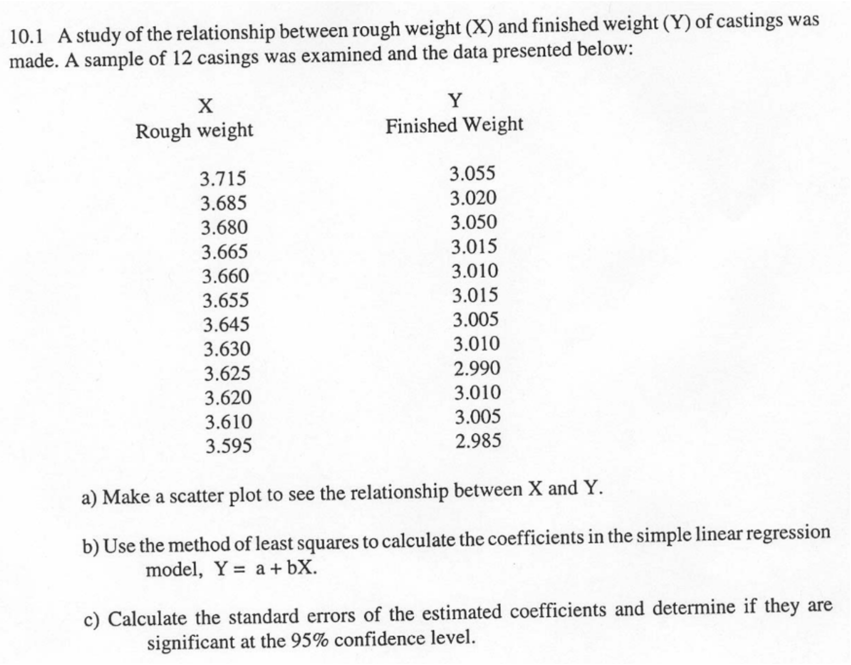
- Mechanistic models are based upon our understanding of the physical systems affecting the response variable
- An engineer uses his knowledge, experience and training in mathematics, physics, chemistry and engineering to develop a mathematical expression that defines the response variable as a function of the regressor variable
- Textbook uses Ohm's law; consider a wind generator
- Statistical techniques can augment mechanistic models

# Empirical Models

- There are many situations in which engineers and scientist do not have a clear understanding of the physical systems of a phenomenon. However, there is some knowledge that a set of regressor variables influences a response variable(s)
- An **empirical model** is not build upon explicit knowledge of the physical phenomenon
- Empirical variables do not prove or disprove that a variable has an affect on the response variable(s)

# Regression Models

- The most popular method of developing empirical models is the use of linear regression models
- It is assumed that the response variable can be modelled by a low-order polynomial equation of the regressor variables
- Very common approach
- Consider the example from Lawson and Erjavec (2001)



10.1 A study of the relationship between rough weight (X) and finished weight (Y) of castings was made. A sample of 12 casings was examined and the data presented below:

| X<br>Rough weight | Y<br>Finished Weight |
|---|---|
| 3.715 | 3.055 |
| 3.685 | 3.020 |
| 3.680 | 3.050 |
| 3.665 | 3.015 |
| 3.660 | 3.010 |
| 3.655 | 3.015 |
| 3.645 | 3.005 |
| 3.630 | 3.010 |
| 3.625 | 2.990 |
| 3.620 | 3.010 |
| 3.610 | 3.005 |
| 3.595 | 2.985 |

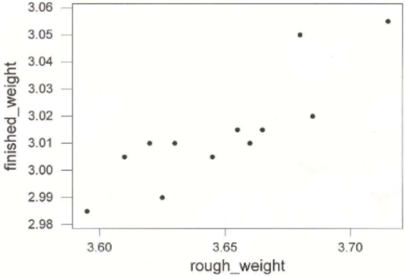
a) Make a scatter plot to see the relationship between X and Y.

b) Use the method of least squares to calculate the coefficients in the simple linear regression model, Y = a + bX.

c) Calculate the standard errors of the estimated coefficients and determine if they are significant at the 95% confidence level.

Scatter plot produced in Minitab

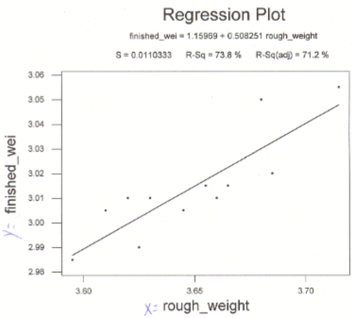
The regressor variable is the rough weight of a casting and the response variable is the finished weight.



This graph suggests a linear relationship between rough and finished weights.

# Fitted Line Plot from Minitab

This Command Calculates the linear regression model and plots the fitted line with the data



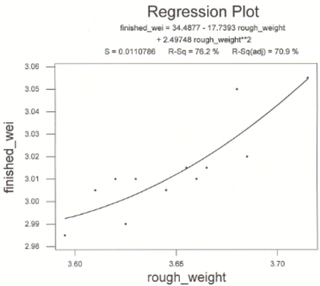
To predict finished weight (y).

$$\hat{y} = 1.15969 + 0.508251x.$$

This model explains 73.8% of the variability present in the data ($R^2$ statistic).

# Fitted Line Plot

This time a quadratic term is added
to the model (notice the slight curvature).



This model is not as good as the first model. Notice
that the adjusted $R^2$-statistic is lower. The curvature
is not very strong and can be removed from the Model.
Also the presence of the quadratic term causes the
linear term to have a negative correlation with
finished Weight (this is counter-intuitive). Finally notice
that the error standard deviation, S is slightly larger
for the Quadratic model (indicates worse fit).

Schubert, Kerber, Schmidt and Jones (1992). "The Catapult Problem: Enhanced Engineering Modeling using Experimental Design." Quality Engineering, 463–473.

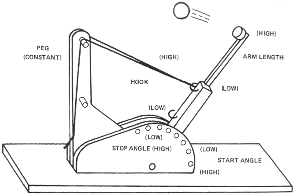


*Figure 1.* The catapult.

# Mechanistic Modeling

We develop an elementary mathematical model by making certain simplifying assumptions about the launch conditions and neglecting all frictional effects. Most of the physical parameters are determined from simple experiments [see Kerber (6) for the details]. Under these conditions, a mechanical analysis of the catapult leads to the equation

$$\frac{I_0 g}{4r_B^2}\frac{(R + R_B\cos\theta_1)^2}{(\sin\theta_1(R\cos\theta_1 + r_B)} = r_F\int_{\theta_0}^{\theta_1} F(\theta)\sin(\phi+\theta)\,d\theta$$
$$+ a\int_{\theta_0}^{\theta_1} F(\theta)\cos(\phi+\theta)\,d\theta - (Mgr_G + mgr_B)(\sin\theta_1 - \sin\theta_0) \qquad (1)$$

where $I_0$ is the moment of inertia of the moment arm/ball combination relative to the pivot point 0. $\theta_0$ and $\theta_1$ are the initial and launching angles, $M$ and $m$ are the masses of the moment arm and ball, respectively, and $r_G$ is the distance from 0 to the mass center $G$ of the moment arm. The remaining dimensions are shown in Figure 2.

In Eq. (1), $F(\theta)$ is the force exerted by the rubber band on the moment arm. In general, this force is a function of the current arm position, $\theta$. However, it also

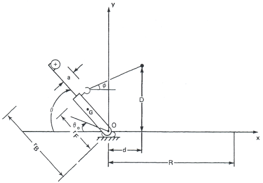


*Figure 2.* Schematic of the catapult with dimensions.

# Mechanistic Modeling -2

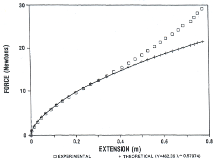


*Figure 3.* A typical force/extension curve for rubber band used in the experiment.

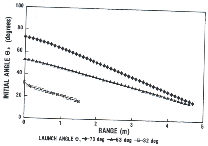
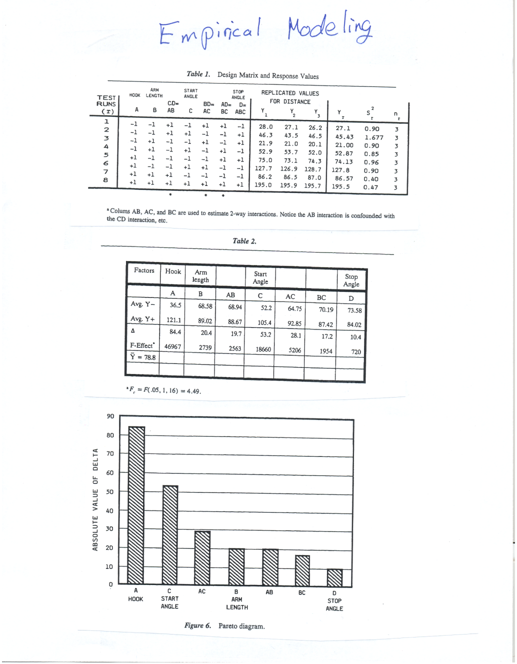


*Figure 4.* A typical output from Eq. (1). Shown are initial and launch angles for prescribed range values.



# Empirical Modeling

*Table 1.* Design Matrix and Response Values

| TEST RUNS | HOOK | ARM LENGTH | | START ANGLE | | | STOP ANGLE | REPLICATED VALUES FOR DISTANCE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (r) | A | B | CD= AB | C | BD= AC | AD= BC | D= ABC | $Y_1$ | $Y_2$ | $Y_3$ | $Y_r$ | $S_r^2$ | $n_r$ |
| 1 | −1 | −1 | +1 | −1 | +1 | +1 | −1 | 28.0 | 27.1 | 26.2 | 27.1 | 0.90 | 3 |
| 2 | −1 | −1 | +1 | +1 | −1 | −1 | +1 | 46.3 | 43.5 | 46.5 | 45.43 | 1.677 | 3 |
| 3 | −1 | +1 | −1 | −1 | +1 | −1 | +1 | 21.9 | 21.0 | 20.1 | 21.00 | 0.90 | 3 |
| 4 | −1 | +1 | −1 | +1 | −1 | +1 | −1 | 52.9 | 53.7 | 52.0 | 52.87 | 0.85 | 3 |
| 5 | +1 | −1 | −1 | −1 | −1 | +1 | +1 | 75.0 | 73.1 | 74.3 | 74.13 | 0.96 | 3 |
| 6 | +1 | −1 | −1 | +1 | +1 | −1 | −1 | 127.7 | 126.9 | 128.7 | 127.8 | 0.90 | 3 |
| 7 | +1 | +1 | +1 | −1 | −1 | −1 | −1 | 86.2 | 86.5 | 87.0 | 86.57 | 0.40 | 3 |
| 8 | +1 | +1 | +1 | +1 | +1 | +1 | +1 | 195.0 | 195.9 | 195.7 | 195.5 | 0.47 | 3 |
| | | | * | | * | * | | | | | | | |

[a] Colums AB, AC, and BC are used to estimate 2-way interactions. Notice the AB interaction is confounded with the CD interaction, etc.

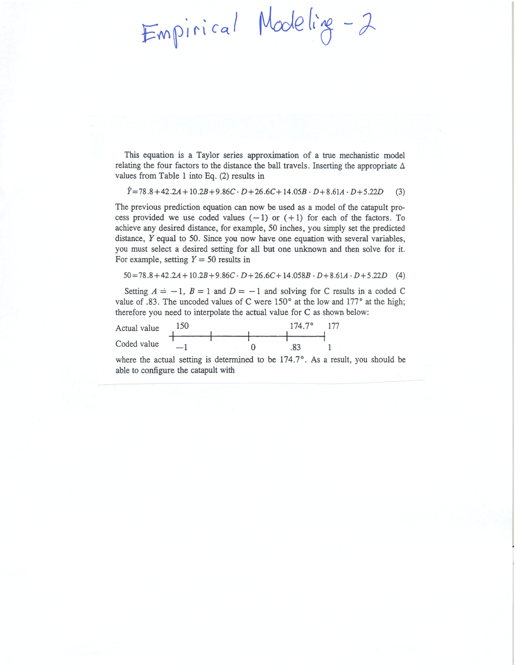
*Table 2.*

| Factors | Hook | Arm length | | Start Angle | | | Stop Angle |
|---|---|---|---|---|---|---|---|
| | A | B | AB | C | AC | BC | D |
| Avg. Y− | 36.5 | 68.58 | 68.94 | 52.2 | 64.75 | 70.19 | 73.58 |
| Avg. Y+ | 121.1 | 89.02 | 88.67 | 105.4 | 92.85 | 87.42 | 84.02 |
| Δ | 84.4 | 20.4 | 19.7 | 53.2 | 28.1 | 17.2 | 10.4 |
| F-Effect[a] | 46967 | 2739 | 2563 | 18660 | 5206 | 1954 | 720 |
| $\bar{Y}$ = 78.8 | | | | | | | |
| | | | | | | | |

[a] $F_c = F(.05, 1, 16) = 4.49$.

*Figure 6.* Pareto diagram.

# Empirical Modeling - 2



This equation is a Taylor series approximation of a true mechanistic model relating the four factors to the distance the ball travels. Inserting the appropriate Δ values from Table 1 into Eq. (2) results in

$$\hat{Y}=78.8+42.2A+10.2B+9.86C\cdot D+26.6C+14.05B\cdot D+8.61A\cdot D+5.22D \qquad (3)$$

The previous prediction equation can now be used as a model of the catapult process provided we use coded values $(-1)$ or $(+1)$ for each of the factors. To achieve any desired distance, for example, 50 inches, you simply set the predicted distance, $Y$ equal to 50. Since you now have one equation with several variables, you must select a desired setting for all but one unknown and then solve for it. For example, setting $Y=50$ results in

$$50=78.8+42.2A+10.2B+9.86C\cdot D+26.6C+14.058B\cdot D+8.61A\cdot D+5.22D \qquad (4)$$

Setting $A \doteq -1$, $B=1$ and $D=-1$ and solving for C results in a coded C value of .83. The uncoded values of C were 150° at the low and 177° at the high; therefore you need to interpolate the actual value for C as shown below:

| | | | | |
|---|---|---|---|---|
| Actual value | 150 | | 174.7° | 177 |
| Coded value | —1 | 0 | .83 | 1 |

where the actual setting is determined to be 174.7°. As a result, you should be able to configure the catapult with