Lecture 2

Introduction to Statistics and Data Analysis II

Lecture Information

Ceng272 *Statistical Computations* at February 22, 2010

Dr. Cem Özdoğan
Computer Engineering Department
Çankaya University

Contents

Role of Probability I

- Concepts in probability allows us

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.
- It is natural to study probability prior to studying statistical inference.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.
- It is natural to study probability prior to studying statistical inference.
- **Example 1.1**. In a manufacturing process, 100 items are sampled and 10 are found to be defective.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.
- It is natural to study probability prior to studying statistical inference.
- **Example 1.1**. In a manufacturing process, 100 items are sampled and 10 are found to be defective.
- However, in the long run, the company can only tolerate 5% defective in the process.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.
- It is natural to study probability prior to studying statistical inference.
- **Example 1.1**. In a manufacturing process, 100 items are sampled and 10 are found to be defective.
- However, in the long run, the company can only tolerate 5% defective in the process.
- Suppose we learn that; if it does produce items 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or <u>confidence</u> in our conclusion.
- It is natural to study probability prior to studying statistical inference.
- **Example 1.1**. In a manufacturing process, 100 items are sampled and 10 are found to be defective.
- However, in the long run, the company can only tolerate 5% defective in the process.
- Suppose we learn that; if it does produce items 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process.
- The small probability suggests that the process indeed have a long-run defective exceeding 5%.

Role of Probability I

- Concepts in probability allows us
  - to have a better understanding of statistical inference,
  - to quantify the strength or confidence in our conclusion.
- It is natural to study probability prior to studying statistical inference.
- **Example 1.1**. In a manufacturing process, 100 items are sampled and 10 are found to be defective.
- However, in the long run, the company can only tolerate 5% defective in the process.
- Suppose we learn that; if it does produce items 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process.
- The small probability suggests that the process indeed have a long-run defective exceeding 5%.
- Probability aids in translation of sample information into conclusions.

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.
  - All other environmental conditions are held constant.

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.
  - All other environmental conditions are held constant.
- The stem weights in grams were recorded after the end of 140 days.

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.
  - All other environmental conditions are held constant.
- The stem weights in grams were recorded after the end of 140 days.
- Would the data set indicate that nitrogen is effective? We observed:

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.
  - All other environmental conditions are held constant.
- The stem weights in grams were recorded after the end of 140 days.
- Would the data set indicate that nitrogen is effective? We observed:
  - Four nitrogen observations are larger than any of the no-nitrogen observations (see underlined elements in Table 1).

Role of Probability II

- **Example 1.2**. We want to determine if the use of nitrogen influences the growth of the roots?
- Experimental Design:
  - Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
  - Here, we have two samples from two populations.
  - All other environmental conditions are held constant.
- The stem weights in grams were recorded after the end of 140 days.
- Would the data set indicate that nitrogen is effective? We observed:
  - Four nitrogen observations are larger than any of the no-nitrogen observations (see underlined elements in Table 1).
  - Most of the no-nitrogen observations appear to be below the center of the data (see underlined element in Table 1).

Role of Probability III

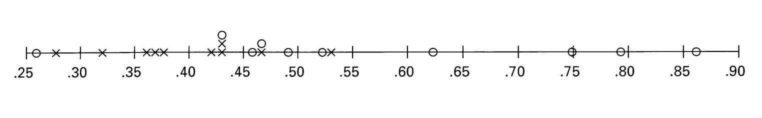


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

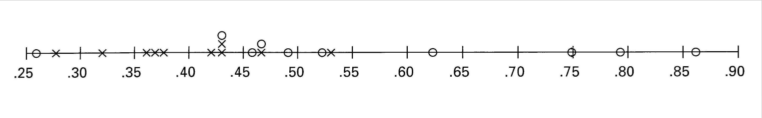


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

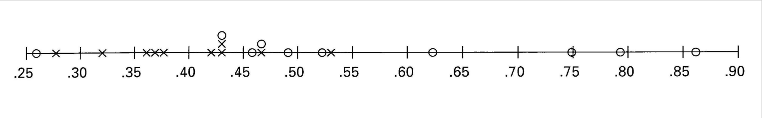


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

- How this can be quantified or summarized in some sense?

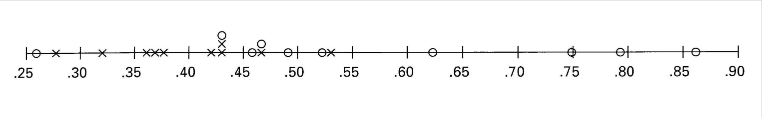


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

- How this can be quantified or summarized in some sense?
- The conclusions may be summarized in a probability statement:

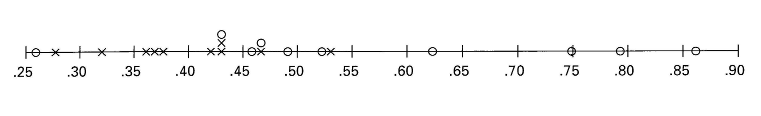


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

- How this can be quantified or summarized in some sense?
- The conclusions may be summarized in a probability statement:

Summary

The probability that data like these could be observed given that nitrogen has no effect is small, say 0.03.

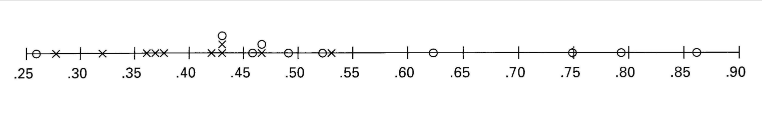


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability III

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

- How this can be quantified or summarized in some sense?
- The conclusions may be summarized in a probability statement:

Summary

The probability that data like these could be observed given that nitrogen has no effect is small, say 0.03.

- That would be strong evidence that the use of nitrogen does have influence.

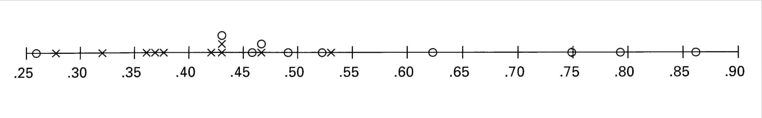


Figure: Stem weight data. (o: the with nitrogen data. x: the without nitrogen data.

Role of Probability IV

- For a statistical problem,
  the sample along with inferential statistics allow us to draw conclusions about the population, with inferential statistics making clear use of elements of probability. (inductive in nature)

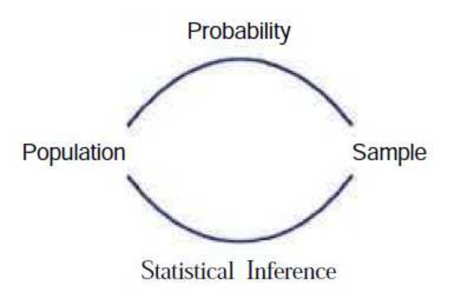


Figure: Fundamental relationship between probability and inferential

Role of Probability IV

- For a statistical problem,
  the sample along with inferential statistics allow us to draw conclusions about the population, with inferential statistics making clear use of elements of probability. (inductive in nature)
- For a probability problem, we can draw conclusions about characteristics of hypothetical data taken from the population based on known features of the population. (deductive in nature)

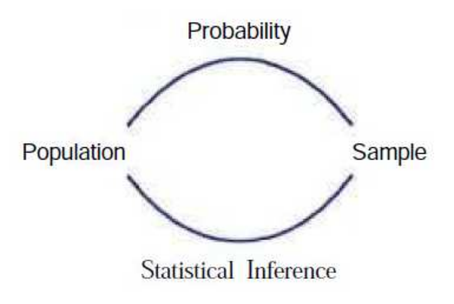


Figure: Fundamental relationship between probability and inferential

Role of Probability V

- The procedure we recognize is that we may want to know about the parameter not from the entire population but from the sample.

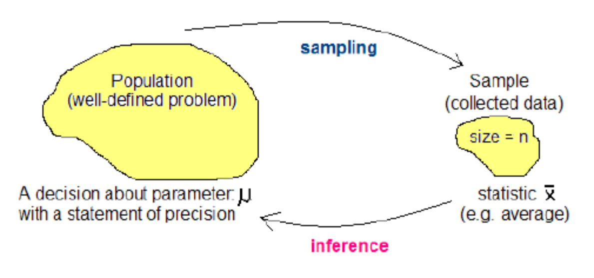


Figure: The Cycle of Statistical Procedure.

Role of Probability V

- The procedure we recognize is that we may want to know about the parameter not from the entire population but from the sample.
- The procedure involves two main different jobs. Those are

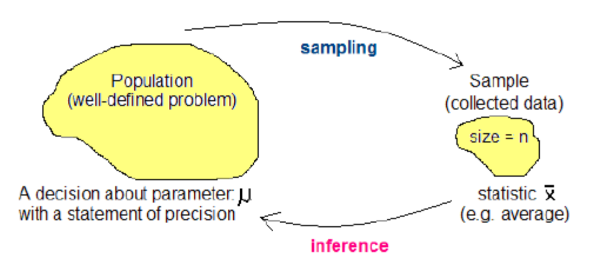


Figure: The Cycle of Statistical Procedure.

Role of Probability V

- The procedure we recognize is that we may want to know about the parameter not from the entire population but from the sample.
- The procedure involves two main different jobs. Those are
  1. estimate a parameter of the population through sample,

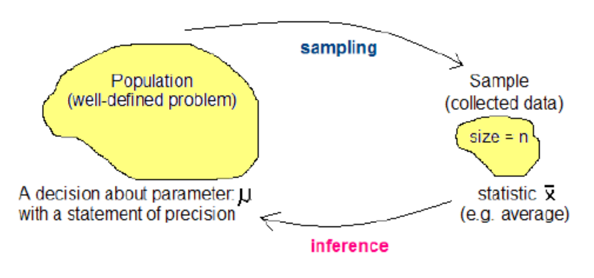


Figure: The Cycle of Statistical Procedure.

Role of Probability V

- The procedure we recognize is that we may want to know about the parameter not from the entire population but from the sample.
- The procedure involves two main different jobs. Those are
  1. estimate a parameter of the population through sample,
  2. testing hypotheses (or conjectures/claims) about the parameter.

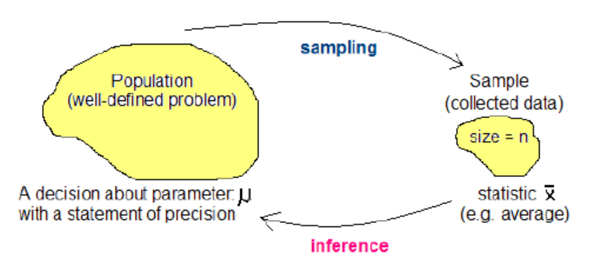


Figure: The Cycle of Statistical Procedure.

Role of Probability V

- The procedure we recognize is that we may want to know about the parameter not from the entire population but from the sample.
- The procedure involves two main different jobs. Those are
  1. estimate a parameter of the population through sample,
  2. testing hypotheses (or conjectures/claims) about the parameter.
- Usually the above two procedures are called collectively statistical inference

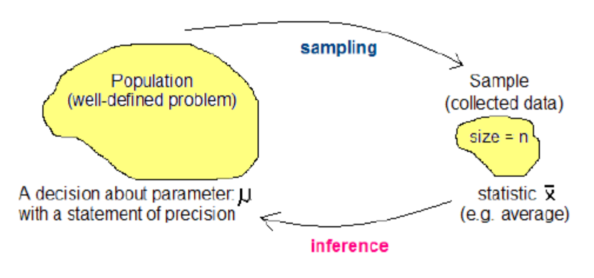


Figure: The Cycle of Statistical Procedure.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
  - Sample size means the number of items in the sample.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
  - Sample size means the number of items in the sample.
- **Biased sample:**

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
  - Sample size means the number of items in the sample.
- **Biased sample:**
  - Example: A sample is chosen to answer certain questions regarding political preferences in a certain state.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
  - Sample size means the number of items in the sample.
- **Biased sample:**
  - Example: A sample is chosen to answer certain questions regarding political preferences in a certain state.
  - Now, suppose that all or nearly all of the 1000 sampling families chosen live in urban (vs. rural) areas.

Sampling Procedures I

- The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- **Simple random sampling:**
  - Any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
  - Sample size means the number of items in the sample.
- **Biased sample:**
  - Example: A sample is chosen to answer certain questions regarding political preferences in a certain state.
  - Now, suppose that all or nearly all of the 1000 sampling families chosen live in urban (vs. rural) areas.
  - Biased sample confined the population and thus the inferences need to be confined to the limited population.

Sampling Procedures II

- **Stratified random sampling:**

Sampling Procedures II

- **Stratified random sampling:**
  - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.

Sampling Procedures II

- **Stratified random sampling:**
  - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.
  - Separate random samples are chosen from each stratum with sample sizes proportional to the size of the stratum.

Sampling Procedures II

- **Stratified random sampling:**
  - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.
  - Separate random samples are chosen from each stratum with sample sizes proportional to the size of the stratum.
  - The purpose is to be sure that each of the strata is neither over- or under-represented. For example,

Sampling Procedures II

- **Stratified random sampling:**
    - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.
    - Separate random samples are chosen from each stratum with sample sizes proportional to the size of the stratum.
    - The purpose is to be sure that each of the strata is neither over- or under-represented. For example,
        - A sample survey is conducted to gather some political opinions in a city,

Sampling Procedures II

- **Stratified random sampling:**
  - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.
  - Separate random samples are chosen from each stratum with sample sizes proportional to the size of the stratum.
  - The purpose is to be sure that each of the strata is neither over- or under-represented. For example,
    - A sample survey is conducted to gather some political opinions in a city,
    - The city is subdivided into several ethnical group,

Sampling Procedures II

- **Stratified random sampling:**
  - The sampling units are not homogeneous and divide themselves into non-overlapping groups, called *strata*.
  - Separate random samples are chosen from each stratum with sample sizes proportional to the size of the stratum.
  - The purpose is to be sure that each of the strata is neither over- or under-represented. For example,
    - A sample survey is conducted to gather some political opinions in a city,
    - The city is subdivided into several ethnical group,
    - Separate random samples of families could be chosen from each group.

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.
- Excessive variability among experimental units will wash out any detectable difference among populations.

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.
- Excessive variability among experimental units will wash out any detectable difference among populations.
- A standard approach is to assign the experimental units randomly to different treatments. For example,

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.
- Excessive variability among experimental units will wash out any detectable difference among populations.
- A standard approach is to assign the experimental units randomly to different treatments. For example,
  - In a drug study, we use a total of 200 available patients.

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.
- Excessive variability among experimental units will wash out any detectable difference among populations.
- A standard approach is to assign the experimental units randomly to different treatments. For example,
    - In a drug study, we use a total of 200 available patients.
    - Age, gender, weight, and other characteristics of the patients may produce variability in the results.

Random Assignment I

- In an experiment, we apply treatments to experimental units and proceed to observe the effect.
- Excessive variability among experimental units will wash out any detectable difference among populations.
- A standard approach is to assign the experimental units randomly to different treatments. For example,
  - In a drug study, we use a total of 200 available patients.
  - Age, gender, weight, and other characteristics of the patients may produce variability in the results.
  - In a completely randomized design, 100 patients are assigned randomly to placebo and 100 to the active drug.

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)
- Four treatment combinations:

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)
- Four treatment combinations:
  - two levels of coating: no coating and chemical coating

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)
- Four treatment combinations:
  - two levels of coating: no coating and chemical coating
  - two relative humidity level: 20% and 80%

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)
- Four treatment combinations:
  - two levels of coating: no coating and chemical coating
  - two relative humidity level: 20% and 80%
- Eight experiment units are used, with two assigned randomly to each of four treatment combinations.

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment II

- **Example 1.3**. A corrosion study to determine if coating of an aluminium reduces the amount of corrosion.
- A corrosion measurement can be expressed in thousands of cycles to failure. (more cycles means less corrosion)
- Four treatment combinations:
  - two levels of coating: no coating and chemical coating
  - two relative humidity level: 20% and 80%
- Eight experiment units are used, with two assigned randomly to each of four treatment combinations.
  - The corrosion data are averages of 2 specimens.

Table: Data for Example 1.3

Coating	Humidity	Thousands of Cycles to Failure
Uncoated	20%	975
Uncoated	80%	350
Chemical Coated	20%	1750
Chemical Coated	80%	1550

Random Assignment III

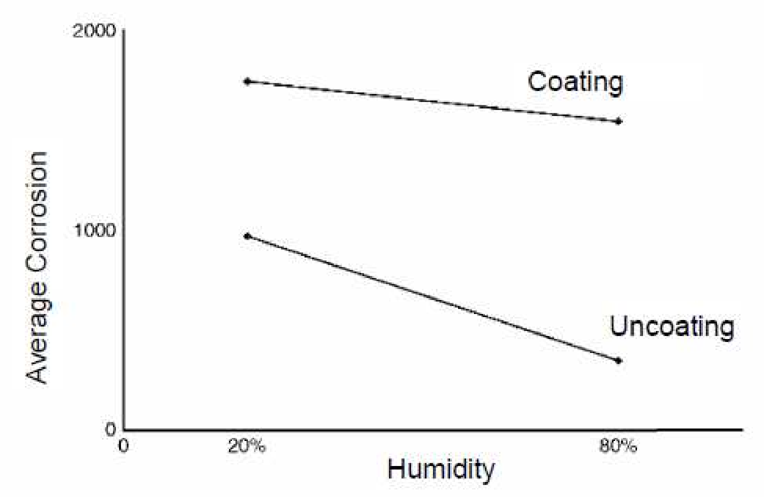


Figure: Corrosion results for Example 1.3.

Random Assignment IV

- Consider the variability around the average:

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.
  - If each corrosion value is an average of two values that are widely dispersed, then this variability wash away any information we obtain.

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.
  - If each corrosion value is an average of two values that are widely dispersed, then this variability wash away any information we obtain.
- Three concepts are illustrated:

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.
  - If each corrosion value is an average of two values that are widely dispersed, then this variability wash away any information we obtain.
- Three concepts are illustrated:
  - Random assignment of treatment combination to experiment units.

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.
  - If each corrosion value is an average of two values that are widely dispersed, then this variability wash away any information we obtain.
- Three concepts are illustrated:
  - Random assignment of treatment combination to experiment units.
  - The use of sample average in summarizing sample information.

Random Assignment IV

- Consider the variability around the average:
  - The use of the chemical corrosion coating procedures appears to reduce corrosion if two corrosion values at each treatment combination are close together.
  - If each corrosion value is an average of two values that are widely dispersed, then this variability wash away any information we obtain.
- Three concepts are illustrated:
  - Random assignment of treatment combination to experiment units.
  - The use of sample average in summarizing sample information.
  - The need for consideration of measures variability in the analysis of sample sets.

Measures of Location: Sample Mean and Median I

- Location measures in a data set provide the analyst some quantitative measure of where the data center is in a sample.

Measures of Location: Sample Mean and Median I

- Location measures in a data set provide the analyst some quantitative measure of where the data center is in a sample.
- One obvious measure is the **sample mean** (Mathematical Average), which is simply a numerical average.

Measures of Location: Sample Mean and Median I

- Location measures in a data set provide the analyst some quantitative measure of where the data center is in a sample.
- One obvious measure is the **sample mean** (Mathematical Average), which is simply a numerical average.
- The **sample median** is to reflect the central tendency of the sample and is uninfluenced by extreme values.

Measures of Location: Sample Mean and Median I

- Location measures in a data set provide the analyst some quantitative measure of where the data center is in a sample.
- One obvious measure is the **sample mean** (Mathematical Average), which is simply a numerical average.
- The **sample median** is to reflect the central tendency of the sample and is uninfluenced by extreme values.
- Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean, denoted by $\bar{x}$ is

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

  Sensitive to outliers (or extreme values).

Measures of Location: Sample Mean and Median I

- Location measures in a data set provide the analyst some quantitative measure of where the data center is in a sample.
- One obvious measure is the **sample mean** (Mathematical Average), which is simply a numerical average.
- The **sample median** is to reflect the central tendency of the sample and is uninfluenced by extreme values.
- Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean, denoted by $\bar{x}$ is

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

  Sensitive to outliers (or extreme values).
- **Sample Median -** middle value in the observations of ordered data set. It divides a data set into two equal parts, denoted by $\tilde{x}$;

$$\tilde{x} = \left\{ \begin{array}{cc} x_{(n+1)/2} & \text{if } n \text{ is odd}, \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even, average of two middle observations} \end{array} \right\}$$

Measures of Location: Sample Mean and Median II

- For example, if the data set is the following: 1.7, 2.2, 3.11, 3.9, and 14.7. The sample mean is 5.12 and the sample median is 3.9.

Measures of Location: Sample Mean and Median II

- For example, if the data set is the following: 1.7, 2.2, 3.11, 3.9, and 14.7. The sample mean is 5.12 and the sample median is 3.9.
- The centroid of the data

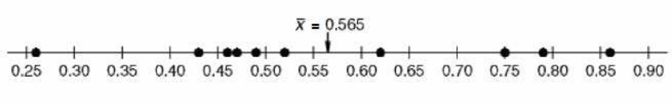


Figure: Sample mean as a centroid of the "with nitrogen" stem weight.

Measures of Location: Sample Mean and Median III

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

- A **trimmed mean** is computed by "trimming away" a certain percent of both the largest and smallest set of values.

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

- A **trimmed mean** is computed by "trimming away" a certain percent of both the largest and smallest set of values.
- Example:

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

- A **trimmed mean** is computed by "trimming away" a certain percent of both the largest and smallest set of values.
- Example:
  - The 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

- A **trimmed mean** is computed by "trimming away" a certain percent of both the largest and smallest set of values.
- Example:
  - The 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.
  - So, for the with nitrogen group the 10% trimmed mean is

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Location: Sample Mean and Median III

- A **trimmed mean** is computed by "trimming away" a certain percent of both the largest and smallest set of values.
- Example:
  - The 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.
  - So, for the with nitrogen group the 10% trimmed mean is

Table: Observation of nitrogen influences.

No nitrogen	Nitrogen
0.32	**0.26**
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	**0.86**
0.38	0.62
0.43	0.46

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625$$

Measures of Variability I

- Location measures do not provide a proper summary of the nature of a data set.

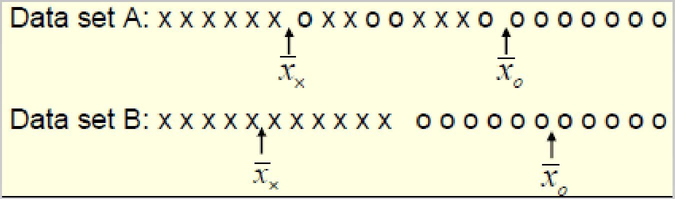


Figure: Different data sets. Difference in the means is roughly the same.

Measures of Variability I

- Location measures do not provide a proper summary of the <u>nature of a data set</u>.
- We can not make meaningful conclusion without considering sample variability. Example:

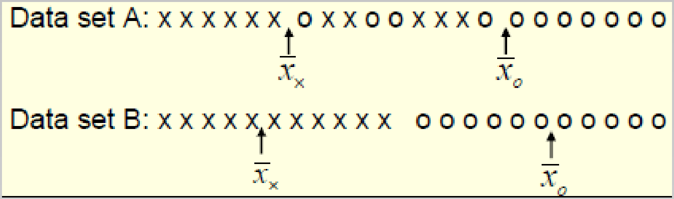


Figure: Different data sets. Difference in the means is roughly the same.

Measures of Variability I

- Location measures do not provide a proper summary of the nature of a data set.
- We can not make meaningful conclusion without considering sample variability. Example:
    - Each data set contains two samples and the difference in the means is roughly the same.

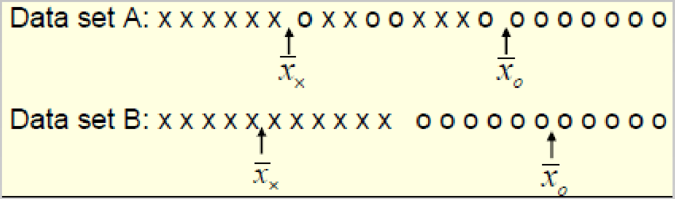


Figure: Different data sets. Difference in the means is roughly the same.

Measures of Variability I

- Location measures do not provide a proper summary of the nature of a data set.
- We can not make meaningful conclusion without considering sample variability. Example:
  - Each data set contains two samples and the difference in the means is roughly the same.
  - Data set B provides sharper distinction between two populations.

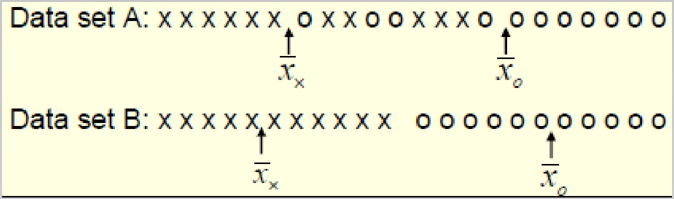


Figure: Different data sets. Difference in the means is roughly the same.

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.
- The question "How far is each x from the mean?"

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.
- The question "How far is each x from the mean?"
- The one that is used most often is the **sample standard deviation**.

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.
- The question “How far is each x from the mean?”
- The one that is used most often is the **sample standard deviation**.
- The **sample variance**, denoted by σ^2

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

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.
- The question "How far is each x from the mean?"
- The one that is used most often is the **sample standard deviation**.
- The **sample variance**, denoted by σ^2

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

- The sample standard deviation, denoted by σ, is the positive square root of σ^2, that is

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

Measures of Variability II

- The simplest measures of sample variability is the **sample range** $X_{max} - X_{min}$.
- The range can be very useful in statistical quality control.
- The question “How far is each x from the mean?”
- The one that is used most often is the **sample standard deviation**.
- The **sample variance**, denoted by σ^2

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

- The sample standard deviation, denoted by σ, is the positive square root of σ^2, that is

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

- The sample variance is measured in squared units. The sample standard deviation is in linear units.

Measures of Variability III

- For a bell-shaped distribution,

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;
  - within two standard deviations of the mean there will be approximately 95% of the data;

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;
  - within two standard deviations of the mean there will be approximately 95% of the data;
  - within three standard deviations of the mean there will be approximately 99.7% of the data.

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;
  - within two standard deviations of the mean there will be approximately 95% of the data;
  - within three standard deviations of the mean there will be approximately 99.7% of the data.
- That is,

$$x \pm \sigma \approx 68\%, \;\; x \pm 2\sigma \approx 95\%, \;\; x \pm 3\sigma \approx 99.7\%.$$

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;
  - within two standard deviations of the mean there will be approximately 95% of the data;
  - within three standard deviations of the mean there will be approximately 99.7% of the data.
- That is,

$$x \pm \sigma \approx 68\%, \;\; x \pm 2\sigma \approx 95\%, \;\; x \pm 3\sigma \approx 99.7\%.$$

- This is a rule of thumb. Since the range $R \approx 6\sigma$, the rule is also called the 6σ-rule.

Measures of Variability III

- For a bell-shaped distribution,
  - within one standard deviation of the mean there will be approximately (empirically) 68% of the data;
  - within two standard deviations of the mean there will be approximately 95% of the data;
  - within three standard deviations of the mean there will be approximately 99.7% of the data.
- That is,

$$x \pm \sigma \approx 68\%, \;\; x \pm 2\sigma \approx 95\%, \;\; x \pm 3\sigma \approx 99.7\%.$$

- This is a rule of thumb. Since the range $R \approx 6\sigma$, the rule is also called the 6σ-rule.
- An observation beyond $(x 2\sigma, x + 2\sigma)$ can be declared as an outlier.

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n - 1$ terms can vary freely.

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n - 1$ terms can vary freely.
- In general,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n - 1$ terms can vary freely.
- In general,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- The computation of a sample variance does not involve n independent squared deviations from the mean. For example,

Measures of Variability IV

- The quantity $n-1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n-1$ terms can vary freely.
- In general,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- The computation of a sample variance does not involve n independent squared deviations from the mean. For example,
  - for the data set (5, 17, 6, 4), the sample mean is 8.

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n - 1$ terms can vary freely.
- In general,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- The computation of a sample variance does not involve n independent squared deviations from the mean. For example,
  - for the data set (5, 17, 6, 4), the sample mean is 8.
  - The variance is

$$(5 - 8)^2 + (17 - 8)^2 + (6 - 8)^2 + (4 - 8)^2$$
$$= (-3)^2 + (9)^2 + (-2)^2 + (-4)^2$$

Measures of Variability IV

- The quantity $n - 1$ is called the **degrees of freedom** associated with the variance estimate.
- It depicts the number of independent pieces of information available for computing variability. Only $n - 1$ terms can vary freely.
- In general,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- The computation of a sample variance does not involve n independent squared deviations from the mean. For example,
  - for the data set (5, 17, 6, 4), the sample mean is 8.
  - The variance is

$$(5-8)^2 + (17-8)^2 + (6-8)^2 + (4-8)^2$$
$$= (-3)^2 + (9)^2 + (-2)^2 + (-4)^2$$

  - The quantities inside parentheses sum to zero.

Measures of Variability V

- **Example 1.4**. An engineer is interested in testing the “bias” in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken with results given by

  7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

$$\bar{x} = 7.0205$$
$$\sigma^2 = 0.001939$$
$$\sigma = \sqrt{0.001939} = 0.0440$$

  with 9 degrees of freedom.

Measures of Variability VI

- In statistical inference, we like to draw conclusions about characteristics of populations, called population parameters.

Measures of Variability VI

- In statistical inference, we like to draw conclusions about characteristics of populations, called population parameters.
- Population mean and population variance are two important parameters.

Measures of Variability VI

- In statistical inference, we like to draw conclusions about characteristics of populations, called population parameters.
- Population mean and population variance are two important parameters.
- The sample variance is used to draw inferences about the population variance.

Measures of Variability VI

- In statistical inference, we like to draw conclusions about characteristics of populations, called population parameters.
- Population mean and population variance are two important parameters.
- The sample variance is used to draw inferences about the population variance.
- The sample standard deviation and the sample mean are used to draw inferences about the population mean.

Measures of Variability VI

- In statistical inference, we like to draw conclusions about characteristics of populations, called population parameters.
- Population mean and population variance are two important parameters.
- The sample variance is used to draw inferences about the population variance.
- The sample standard deviation and the sample mean are used to draw inferences about the population mean.
- In general, the variance is considered more in inferential theory, while the standard deviation is used more in applications.

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or <u>continuous</u>.

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or <u>continuous</u>.
- Both binary data and count data are discrete data.

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or <u>continuous</u>.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or <u>continuous</u>.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.
- Binary data and sample proportion:

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or continuous.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.
- Binary data and sample proportion:
    - Two categories are involved.

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or continuous.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.
- Binary data and sample proportion:
  - Two categories are involved.
  - If there are n units involved in the data and x units is in category 1 and $n - x$ units are in category 2

Discrete and continuous data

- Depending on the area of application, the data gathered may be **discrete** or continuous.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.
- Binary data and sample proportion:
  - Two categories are involved.
  - If there are n units involved in the data and x units is in category 1 and $n - x$ units are in category 2
  - The sample proportion in category 1 is x/n

Discrete and continuous data

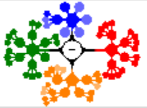
- Depending on the area of application, the data gathered may be **discrete** or <u>continuous</u>.
- Both binary data and count data are discrete data.
- Great distinctions are made between discrete and continuous data in the probability theory.
- Binary data and sample proportion:
    - Two categories are involved.
    - If there are n units involved in the data and x units is in category 1 and $n - x$ units are in category 2
    - The sample proportion in category 1 is x/n
    - The sample proportion in category 2 is $1 - x/n$

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.



Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst.
  - Example 1.2 scientists draw some distinction between “nitrogen” and “no-nitrogen” populations through the sample information.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst.
  - Example 1.2 scientists draw some distinction between “nitrogen” and “no-nitrogen” populations through the sample information.
  - The analysis may require a certain model for the data, e.g., normal (Gaussian) distributions.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst.
  - Example 1.2 scientists draw some distinction between “nitrogen” and “no-nitrogen” populations through the sample information.
  - The analysis may require a certain model for the data, e.g., normal (Gaussian) distributions.
- Some simple graphics (plots) can suggest the clear distinction between the samples, e.g., means and variability.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics I

- The result of a statistical analysis is the estimation of parameters of a postulated model.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst.
  - Example 1.2 scientists draw some distinction between “nitrogen” and “no-nitrogen” populations through the sample information.
  - The analysis may require a certain model for the data, e.g., normal (Gaussian) distributions.
- Some simple graphics (plots) can suggest the clear distinction between the samples, e.g., means and variability.
- Often, plots can illustrate information that sometimes are not retrieved from the formal analysis.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength.** A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.

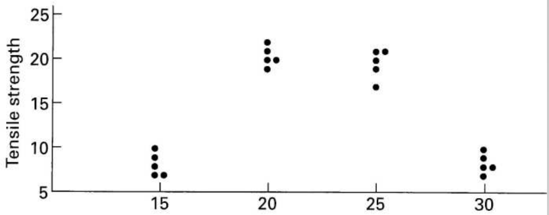


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength**. A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.

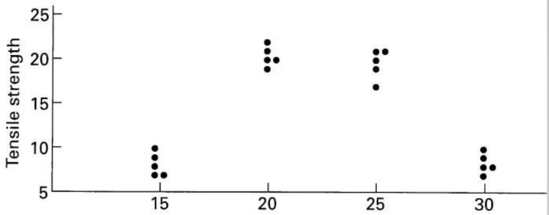


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength.** A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.

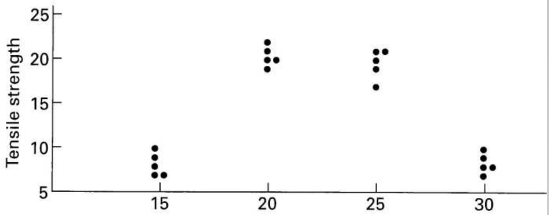


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength.** A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.

- Five cloth specimens are tested for each of the four cotton percentages.

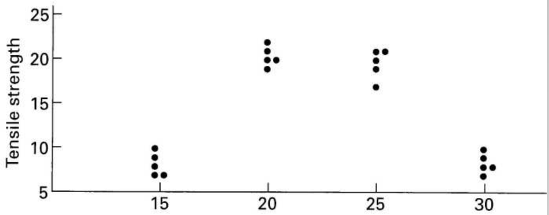


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength.** A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.
- Five cloth specimens are tested for each of the four cotton percentages.
- A reasonable model is that each sample comes from a normal distribution.

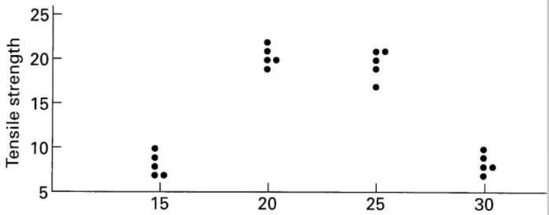


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics II

- **Example of tensile strength.** A textile manufacturer design an experiment to determine the relationship between the tensile strength and the cotton percentage of the cloth specimens.

- Five cloth specimens are tested for each of the four cotton percentages.
- A reasonable model is that each sample comes from a normal distribution.

Table: Observation of nitrogen influences.

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

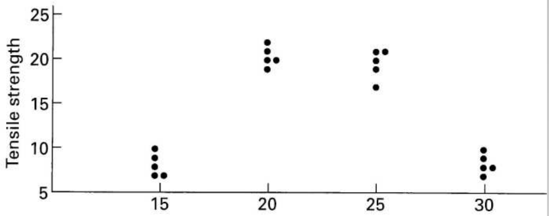


Figure: Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics III

- It is likely that the scientist anticipates the existence of a maximum population mean of tensile strength.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics III

- It is likely that the scientist anticipates the existence of a maximum population mean of tensile strength.
- Here the analysis of the data may revolve around a different type of model, whose structure relating the population mean tensile strength to the cotton concentration.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics III

- It is likely that the scientist anticipates the existence of a maximum population mean of tensile strength.
- Here the analysis of the data may revolve around a different type of model, whose structure relating the population mean tensile strength to the cotton concentration.
  - E.g., a **regression model**; $\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2$ where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics III

- It is likely that the scientist anticipates the existence of a maximum population mean of tensile strength.
- Here the analysis of the data may revolve around a different type of model, whose structure relating the population mean tensile strength to the cotton concentration.
    - E.g., a **regression model**; $\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2$ where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C.
    - The use of an empirical model is accompanied by estimation theory, where $\beta_0, \beta_1, \beta_2$ are estimated by the data.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics IV

- The type of model used to describe the data often depends on the goal of the experiment.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics IV

- The type of model used to describe the data often depends on the goal of the experiment.
- The structure of the model should take advantage of nonstatistical scientific input.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics IV

- The type of model used to describe the data often depends on the goal of the experiment.
- The structure of the model should take advantage of nonstatistical scientific input.
- A selection of a model represents a fundamental assumption upon which the resulting statistical inference is based.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics IV

- The type of model used to describe the data often depends on the goal of the experiment.
- The structure of the model should take advantage of nonstatistical scientific input.
- A selection of a model represents a fundamental assumption upon which the resulting statistical inference is based.
- Often, plots (graphics)can illustrate information that allows the results of the formal statistical inference to be better communicated to the scientist or engineer, and teach the analyst something not retrieved from the formal analysis.

Graphical Methods and Data Description I

- Characterizing or summarizing the nature of collections of data is important.

Graphical Methods and Data Description I

- Characterizing or summarizing the nature of collections of data is important.
- A summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken.

Graphical Methods and Data Description I

- Characterizing or summarizing the nature of collections of data is important.
- A summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken.
- A **Stem-and-leaf** plot, a combined tabular and graphic display, can be used to study the behavior of the mass statistical data.

Graphical Methods and Data Description I

- Characterizing or summarizing the nature of collections of data is important.
- A summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken.
- A **Stem-and-leaf** plot, a combined tabular and graphic display, can be used to study the behavior of the mass statistical data.
- Example: the following table show the life of 40 car batteries

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Figure: Table of Car Battery Life (in years).

Graphical Methods and Data Description II

Table: Stem (integer part)-and-Leaf (decimal part) Plot of Battery Life.

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

Table: Double-Stem-and-Leaf Plot of Battery Life.

Stem	Leaf	Frequency
1.	69	2
2*	2	1
2.	5669	4
3*	001111222333444	15
3.	5567778899	10
4*	11234	5
4.	577	3

Graphical Methods and Data Description III

Table: Relative Frequency Distribution of Battery Life.

Class Interval	Class Midpoint	Frequency *f*	Relative Frequency
1.5-1.9	1.7	2	0.05
2.0-2.4	2.2	1	0.025
2.5-2.9	2.7	4	0.100
3.0-3.4	3.2	15	0.375
3.5-3.9	3.7	10	0.250
4.0-4.4	4.2	5	0.125
4.5-4.9	4.7	3	0.075

Graphical Methods and Data Description IV

Graphical Methods and Data Description IV

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

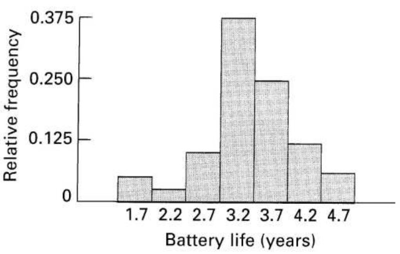


Figure: Relative frequency histogram.

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

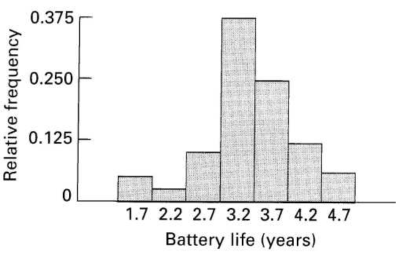


Figure: Relative frequency histogram.

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

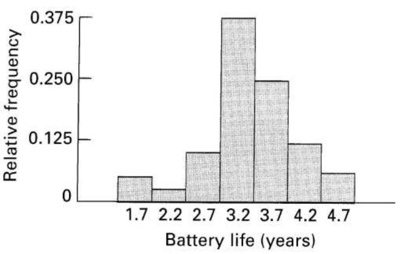


Figure: Relative frequency histogram.

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

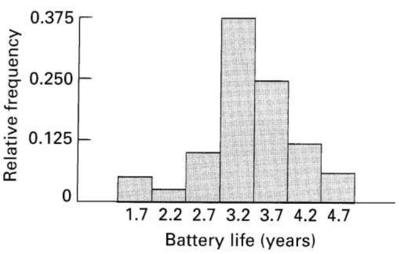


Figure: Relative frequency histogram.

- As the sample size becomes larger, the frequency histogram would approach a bell-shaped continuous **probability distribution**.

Graphical Methods and Data Description IV

- By rotating a stem-and-leaf plot counter-clockwise through an angle of 90, the resulting columns of leaves form a picture that is similar to a histogram.

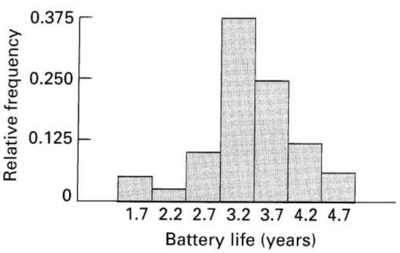


Figure: Relative frequency histogram.

- As the sample size becomes larger, the frequency histogram would approach a bell-shaped continuous **probability distribution**.

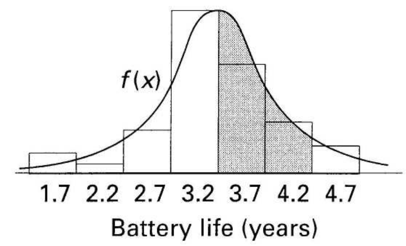


Figure: Estimating frequency distribution.

Graphical Methods and Data Description V

- **Skewness of data.** A distribution is symmetric if it can be folded along a vertical axis so that the two sides coincide, otherwise skewed.

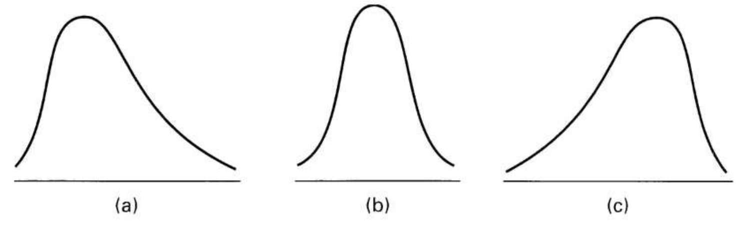


Figure: Skewness of data.

Graphical Methods and Data Description V

- **Skewness of data.** A distribution is symmetric if it can be folded along a vertical axis so that the two sides coincide, otherwise skewed.
- Relationship between the Mean, Median, and Mode;

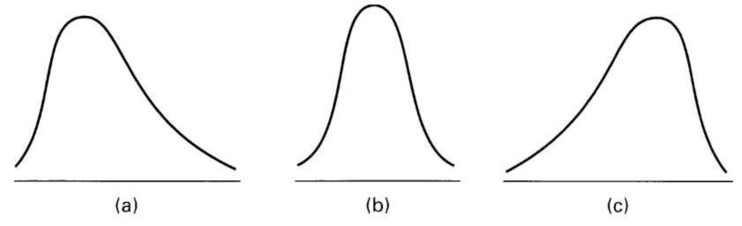


Figure: Skewness of data.

Graphical Methods and Data Description V

- **Skewness of data.** A distribution is symmetric if it can be folded along a vertical axis so that the two sides coincide, otherwise skewed.
- Relationship between the Mean, Median, and Mode;
  - For a symmetric histogram or frequency curve; mode = median = mean,

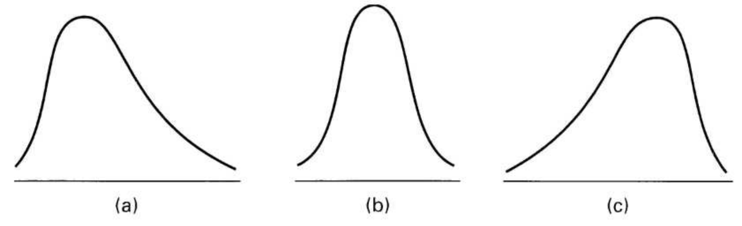


Figure: Skewness of data.

Graphical Methods and Data Description V

- **Skewness of data.** A distribution is symmetric if it can be folded along a vertical axis so that the two sides coincide, otherwise skewed.
- Relationship between the Mean, Median, and Mode;
  - For a symmetric histogram or frequency curve; mode = median = mean,
  - Skewed to the right; mode < median < mean,

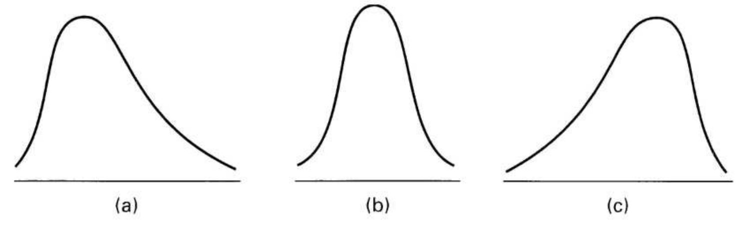


Figure: Skewness of data.

Graphical Methods and Data Description V

- **Skewness of data.** A distribution is symmetric if it can be folded along a vertical axis so that the two sides coincide, otherwise skewed.
- Relationship between the Mean, Median, and Mode;
  - For a symmetric histogram or frequency curve; mode = median = mean,
  - Skewed to the right; mode < median < mean,
  - Skewed to the left; mean < median < mode.

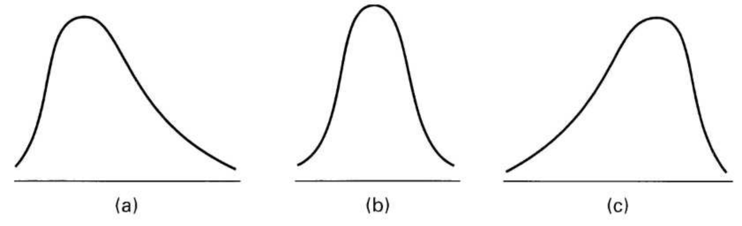


Figure: Skewness of data.

Graphical Methods and Data Description VI

- **Other distinguishing features of a sample.** The distribution can be divided by computing percentiles of the distribution.

Graphical Methods and Data Description VI

- **Other distinguishing features of a sample.** The distribution can be divided by computing percentiles of the distribution.
- These quantities give the analyst a sense of the tails of the distribution.

Graphical Methods and Data Description VI

- **Other distinguishing features of a sample.** The distribution can be divided by computing percentiles of the distribution.
- These quantities give the analyst a sense of the tails of the distribution.
- Tails are the relatively extreme values, either small or large. For example,

Graphical Methods and Data Description VI

- **Other distinguishing features of a sample.** The distribution can be divided by computing percentiles of the distribution.
- These quantities give the analyst a sense of the tails of the distribution.
- Tails are the relatively extreme values, either small or large. For example,
    - the 95th percentile separates the highest 5% from the bottom 95%.

Graphical Methods and Data Description VI

- **Other distinguishing features of a sample.** The distribution can be divided by computing percentiles of the distribution.
- These quantities give the analyst a sense of the tails of the distribution.
- Tails are the relatively extreme values, either small or large. For example,
    - the 95th percentile separates the highest 5% from the bottom 95%.
    - the 1st percentile separate the bottom 1% from the rest of the distribution.

General Types of Statistical Studies

- **Designed experiment.**

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.

General Types of Statistical Studies

- **Designed experiment.**
    - The analyst chooses and controls range of factors.
    - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
    - Factors of interest can not be controlled.
    - It is at the mercy of nature.
- **Retrospective study.**
    - Historical data are used.
    - Advantages:

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.
  - Disadvantages:

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.
  - Disadvantages:
    - validity of data is often in doubt.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.
  - Disadvantages:
    - validity of data is often in doubt.
    - there may be data missing.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.
  - Disadvantages:
    - validity of data is often in doubt.
    - there may be data missing.
    - there may be unknown errors in data.

General Types of Statistical Studies

- **Designed experiment.**
  - The analyst chooses and controls range of factors.
  - Nuisance factors would be equalized via the randomized process.
- **Observational study.**
  - Factors of interest can not be controlled.
  - It is at the mercy of nature.
- **Retrospective study.**
  - Historical data are used.
  - Advantages:
    - no cost in collecting the data.
  - Disadvantages:
    - validity of data is often in doubt.
    - there may be data missing.
    - there may be unknown errors in data.
    - there is no control on the range of the measured variables.