

# **Chapter One**

## **What is Statistics?**

# What is Statistics?

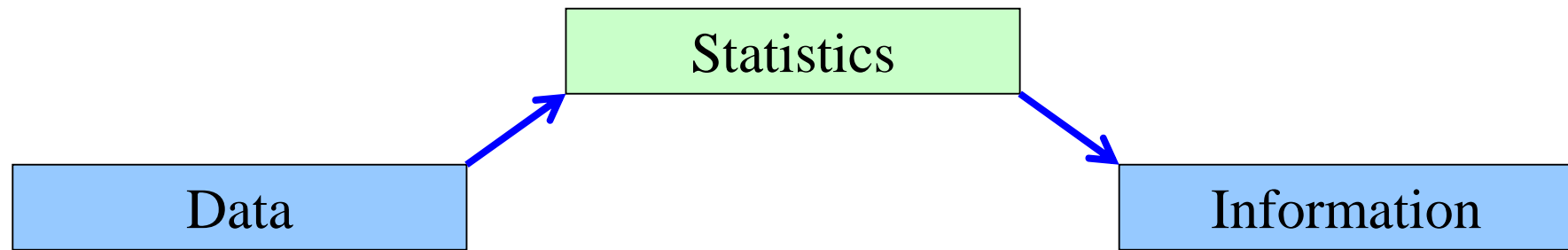
---

“Statistics is a way to get information from data.”

# What is Statistics?

---

“Statistics is a way to get information from data”



Statistics is a *tool* for creating *new understanding* from a set of numbers.

Q: If that's all there is to statistics, why is this book so large?

## Example 2.6 Stats Anxiety

- A student enrolled in a business program is attending the first class of the required statistics course.
- The student is somewhat apprehensive because he believes the myth that the course is difficult.
- To alleviate his anxiety the student asks the professor about last year's marks.
- The professor obliges and provides a list of the final marks, which is composed of term work plus the final exam.
- What information can the student obtain from the list?

## Example 2.6 Stats Anxiety

---

65	81	72	59
71	53	85	66
66	70	72	71
79	76	77	68
65	73	64	72
82	73	77	75
80	85	89	74
86	83	87	77
67	80	78	69
64	67	79	60
62	78	59	92
74	68	63	69
67	67	84	69
72	62	74	73
68	83	74	65

Q: Can you find any information from the above raw data without any statistical work?

## Example 2.6 Stats Anxiety

“Typical mark”

- Mean (average mark)
- Median (mark such that 50% above and 50% below)

Mean = 72.67

Median = 72

Is this enough information?

## Example 2.6 Stats Anxiety

Are most of the marks clustered around the mean or are they more spread out?

Range = Maximum – minimum =  $92 - 53 = 39$

Variance

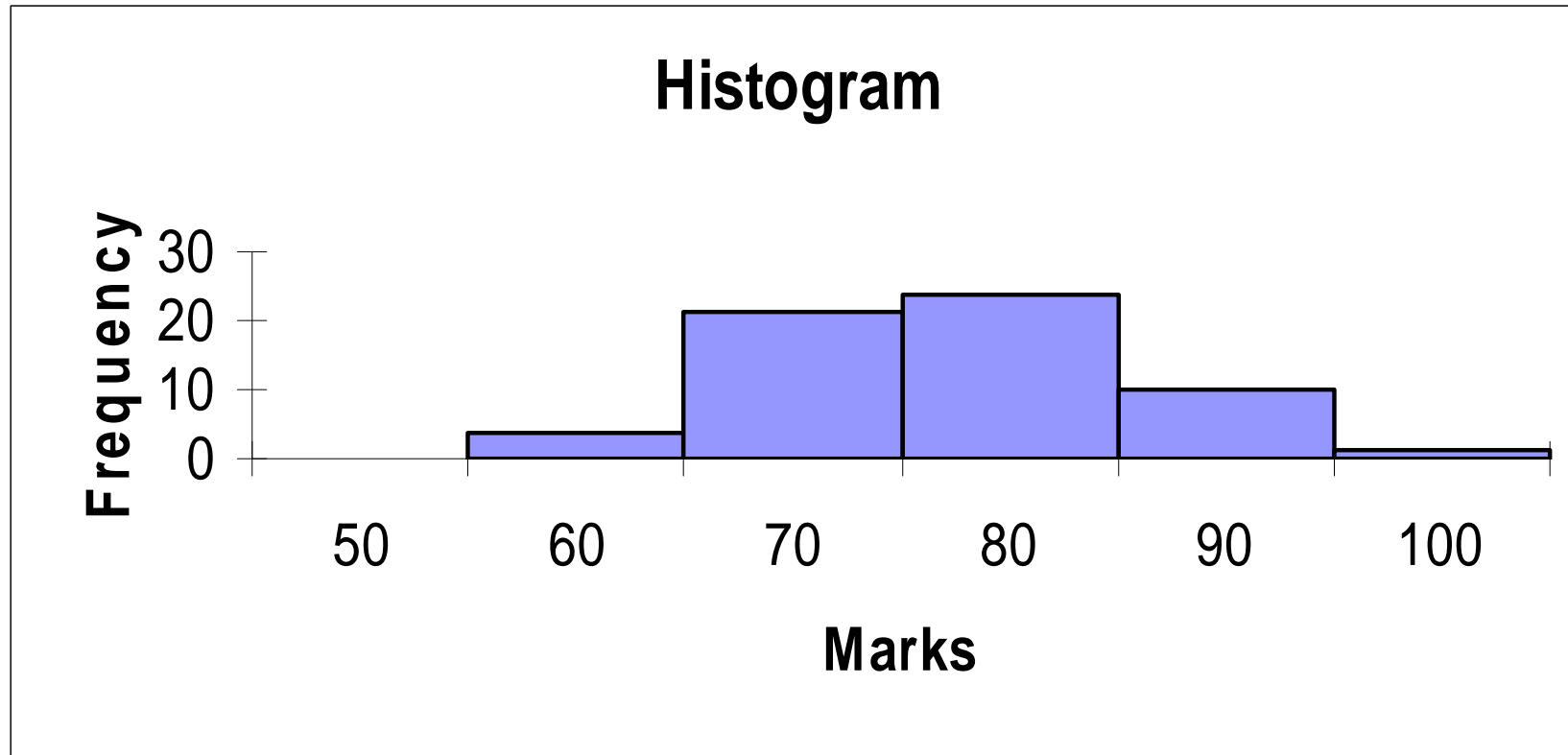
Standard deviation

## Example 2.6 Stats Anxiety

- Q1: Are there many marks below 60 or above 80?
- Q2: What proportions are A, B, C, D grades?

A graphical technique –histogram can provide us with this and other information

## Example 2.6 Stats Anxiety



Q: Will the student's anxiety be released with the information from the histogram?

# Descriptive Statistics

---

- Descriptive statistics deals with methods of **organizing**, **summarizing**, and **presenting** data in a convenient and informative way.
- One form of descriptive statistics uses **graphical techniques**, which allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information.
- Chapter 2 introduces several graphical methods.

# Descriptive Statistics

---

- Another form of descriptive statistics uses numerical techniques to summarize data.
- The mean and median are popular numerical techniques to describe the location of the data.
- The range, variance, and standard deviation measure the variability of the data
- Chapter 4 introduces several numerical statistical measures that describe different features of the data.

# Case 12.1 Pepsi's Exclusivity Agreement

- A large university with a total enrollment of about 50,000 students has offered Pepsi-Cola an exclusivity agreement
  - that would give Pepsi exclusive rights to sell its products at all university facilities for the next year with an option for future years.
- In return, the university would receive
  - 35% of the on-campus revenues and
  - an additional lump sum (總金額) of \$200,000 per year.
- Pepsi has been given 2 weeks to respond.

# Case 12.1 Pepsi's Exclusivity Agreement

- The market for soft drinks is measured in terms of 12-ounce cans.
- Pepsi currently sells an average of 22,000 cans per week (over the 40 weeks of the year that the university operates).
- The cans sell for an average of 75 cents each. The costs including labor amount to 20 cents per can.
- Pepsi is unsure of its market share but suspects it is considerably less than 50%.

# Case 12.1 Pepsi's Exclusivity Agreement

- A quick analysis reveals that if its current market share were 25%, then, with an exclusivity agreement,
  - Pepsi would sell 88,000 (22,000 is 25% of 88,000) cans per week or 3,520,000 cans per year.
- The profit or loss can be calculated.
- The only problem is that we do not know how many soft drinks are sold weekly at the university.

# Case 12.1 Pepsi's Exclusivity Agreement

- Pepsi assigned a recent university graduate to survey the university's students to supply the missing information.
- Accordingly, she organizes a survey that asks 500 students to keep track of the number of soft drinks they purchase in the next 7 days.
- The responses are stored in a file on the disk that accompanies this book. [Case 12.1](#)

# Inferential statistics

---

- The information we would like to acquire in Case 12.1 is
  - an estimate of annual profits from the exclusivity agreement.
- The data are the numbers of cans of soft drinks consumed in 7 days by the 500 students in the sample.
- We want to know the mean number of soft drinks consumed by all 50,000 students on campus.
- To accomplish this goal we need another branch of statistics- inferential statistics.

# Inferential statistics

---

- Inferential statistics is a body of methods used to
  - draw conclusions or inferences about characteristics of populations based on sample data.
- The population in question in this case is
  - the soft drink consumption of the university's 50,000 students.
- The cost of interviewing each student would be prohibitive and extremely time consuming.

# Inferential statistics

---

- Statistical techniques make such endeavors unnecessary.
- Instead, we can sample a much smaller number of students (the sample size is 500) and infer from the data the number of soft drinks consumed by all 50,000 students.
- We can then estimate annual profits for Pepsi.

## Example 12.5

---

When an election for political office takes place, the television networks cancel regular programming and instead provide election coverage.

When the ballots are counted the results are reported.

However, for important offices such as president or senator in large states, the networks actively compete to see which will be the first to predict a winner.

## Example 12.5

---

- This is done through **exit polls**, wherein a random sample of voters who exit the polling booth is asked for whom they voted.
- From the data the sample proportion of voters supporting the candidates is computed.
- A statistical technique is applied to determine
  - whether there is enough evidence to infer that the leading candidate will garner enough votes to win.

## Example 12.5

---

- The exit poll results from the state of Florida during the 2000 year elections were recorded (only the votes of the Republican candidate George W. Bush and the Democrat Albert Gore).
- Suppose that the results (765 people who voted for either Bush or Gore) were stored on a file on the disk. (1 = Gore and 2 = Bush) [Xm12-05](#)
- The network analysts would like to know whether they can conclude that
  - George W. Bush will win the state of Florida.

# Example 12.5

---

- Example 12.5 describes a very common application of statistical inference.
- The population the television networks wanted to make inferences about is the approximately 5 million Floridians who voted for Bush or Gore for president.
- The sample consisted of the 765 people randomly selected by the polling company who voted for either of the two main candidates.

# Example 12.5

---

- The characteristic of the population that we would like to know is
  - the proportion of the total electorate that voted for Bush.
- Specifically, we would like to know
  - whether more than 50% of the electorate voted for Bush (counting only those who voted for either the Republican or Democratic candidate).

## Example 12.5

---

Because we will not ask every one of the 5 million actual voters for whom they voted, we cannot predict the outcome with 100% certainty.

A sample that is only a small fraction of the size of the population can lead to correct inferences only a certain percentage of the time.

You will find that statistics practitioners can control that fraction and usually set it between 90% and 99%.

# Key Statistical Concepts

---

## Population

— a *population* is the group of all items of interest to a statistics practitioner.

— frequently very large; sometimes infinite.

E.g. All 5 million Florida voters, per Example 12.5

## Sample

— A *sample* is a set of data drawn from the population.

— Potentially very large, but less than the population.

E.g. a sample of 765 voters exit polled on election day.

# Key Statistical Concepts

---

## Parameter

— A descriptive measure of a *population*.

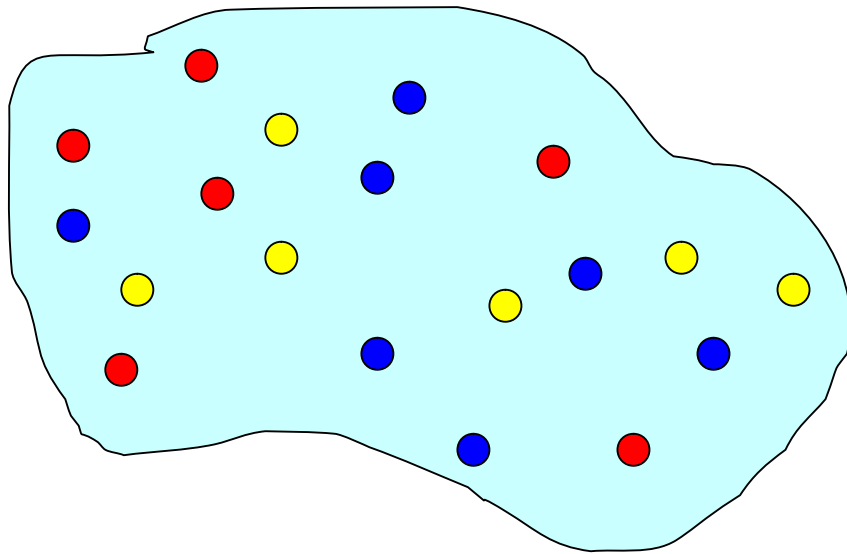
## Statistic

— A descriptive measure of a *sample*.

# Key Statistical Concepts

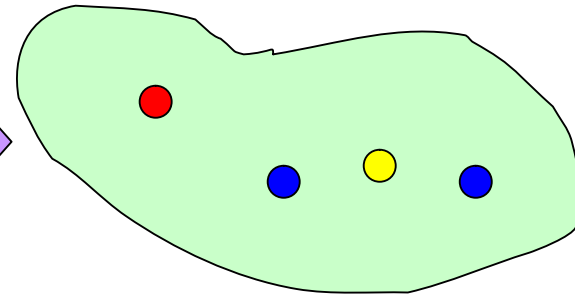
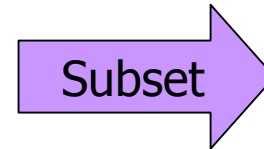
---

Population



Parameter

Sample



Statistic

Populations have Parameters,

Samples have Statistics.

# Descriptive Statistics

---

...are *methods* of organizing, summarizing, and presenting data in a convenient and informative way. These methods include:

Graphical Techniques (Chapter 2), and  
Numerical Techniques (Chapter 4).

The actual method used depends on what *information* we would like to extract. Are we interested in...

- measure(s) of central location? and/or
- measure(s) of variability (dispersion)?

Descriptive Statistics helps to answer these questions...

# Inferential Statistics

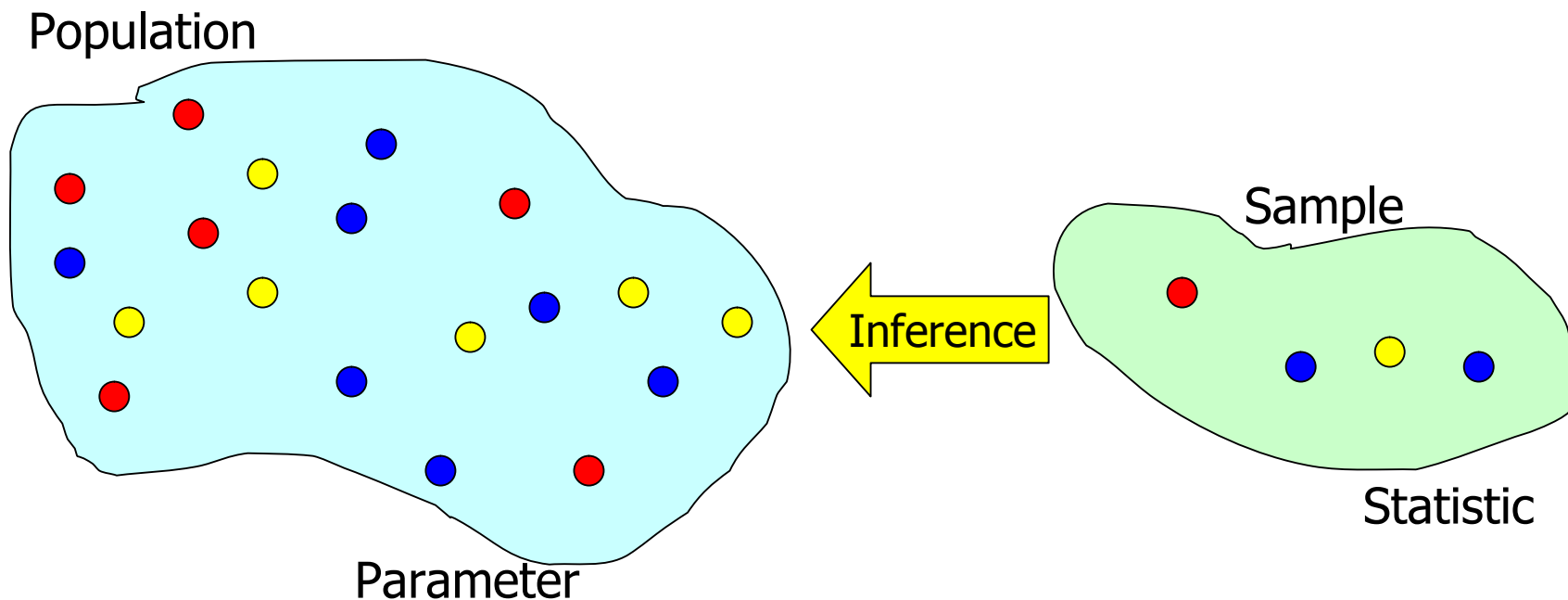
---

Descriptive Statistics describe the data set that's being analyzed, but doesn't allow us to draw any conclusions or make any interferences about the data. Hence we need another branch of statistics: *inferential statistics*.

Inferential statistics is also a set of methods, but it is used to draw conclusions or inferences about characteristics of *populations* based on data from a *sample*.

# Statistical Inference

**Statistical inference** is the *process* of making an estimate, prediction, or decision about a population based on a sample.



What can we *infer* about a Population's Parameters based on a Sample's Statistics?

# Statistical Inference

---

We use statistics to make inferences about parameters.

Therefore, we can make an estimate, prediction, or decision about a **population** based on **sample** data.

Thus, we can apply what we know about a sample to the larger population from which it was drawn!

# Statistical Inference

---

## Rationale:

- Large populations make investigating each member impractical and expensive.
- Easier and cheaper to take a sample and make estimates about the population from the sample.

## However:

Such conclusions and estimates are not always going to be correct.

For this reason, we build into the statistical inference “measures of reliability,” namely **confidence level** and **significance level**.

# Confidence & Significance Levels

The *confidence level* is the proportion of times that an estimating procedure will be correct.

E.g. a confidence level of 95% means that, estimates based on this form of statistical inference will be correct 95% of the time.

When the purpose of the statistical inference is to draw a conclusion about a population, the *significance level* measures how frequently the conclusion will be wrong in the long run.

E.g. a 5% significance level means that, in the long run, this type of conclusion will be wrong 5% of the time.

# Confidence & Significance Levels

If we use  $\alpha$  (Greek letter “alpha”) to represent significance, then our confidence level is  $1 - \alpha$ .

This relationship can also be stated as:

$$\begin{array}{r} \text{Confidence Level} \\ + \text{Significance Level} \\ \hline = 1 \end{array}$$

# Confidence & Significance Levels

Consider a statement from polling data you may hear about in the news:

`"This poll is considered accurate within 3.4 percentage points, 19 times out of 20."`

In this case, our confidence level is 95% ( $19/20 = 0.95$ ), while our significance level is 5%.

# Statistical Applications in Business

Statistical analysis plays an important role in virtually *all* aspects of business and economics.

Throughout this course, we will see applications of statistics in accounting, economics, finance, human resources management, marketing, and operations management.

Exercise for Chapter 1 (p8-9): 1.3, 1.4, 1.6, 1.7