Chapter 1

Basic Concepts in Statistics

Peng HUA 华鹏

huapeng@hit.edu.cn

办公室 Office: G620

Outline

1.1 Populations and Samples

1.2 Selecting a Random Sample

1.3 Ratio, Interval, Ordinal, and Nominative Scales of Measurement (Optional)

1.4 An Introduction to Survey Sampling (Optional)

1.5 More About Data Acquisition and Survey Sampling (Optional)

1.1 Populations and Samples

Data

- **Data**: facts and figures from which conclusions can be drawn
- **Data set**: the data that are collected for a particular study
- **Elements**: a single member or unit of a population or sample: people, objects, events, or other entries
- **Variable**: a characteristic or attribute of an element that can take on different values
- **Measurement**: A way to assign a value of a variable to the element
- Example: A bank might measure the time it takes for a credit card-holder's bill to be paid to the nearest day

1.1 Populations and Samples

- **Population**: All students in a school.
  - **Element**: Each student.
  - **Variables**: Height, weight, age, gender, etc.

- **Population**: All cars in a parking lot.
  - **Element**: Each car.
  - **Variables**: Color, make, model, year, etc.

1.1 Data

- **Quantitative** variable : Measurements that represent quantities are numbers (for example, “how much” or “how many”). For example, **annual starting salary** is quantitative, **age and number of children** is also quantitative
- **Qualitative** or **categorical** variable: Measurements that represent quantities fall into several categories. For example, **a person’s gender**, **the make of an automobile** and **whether a person who purchases a product is satisfied with the product** are qualitative.

1.1 Two types of qualitative variables:

- Nominative
  - Unranked categorization
    - Example: gender, car color
- Ordinal
  - Rank-order categories
  - Ranks are relative to each other
    - Example: Low (1), moderate (2), or high (3) risk

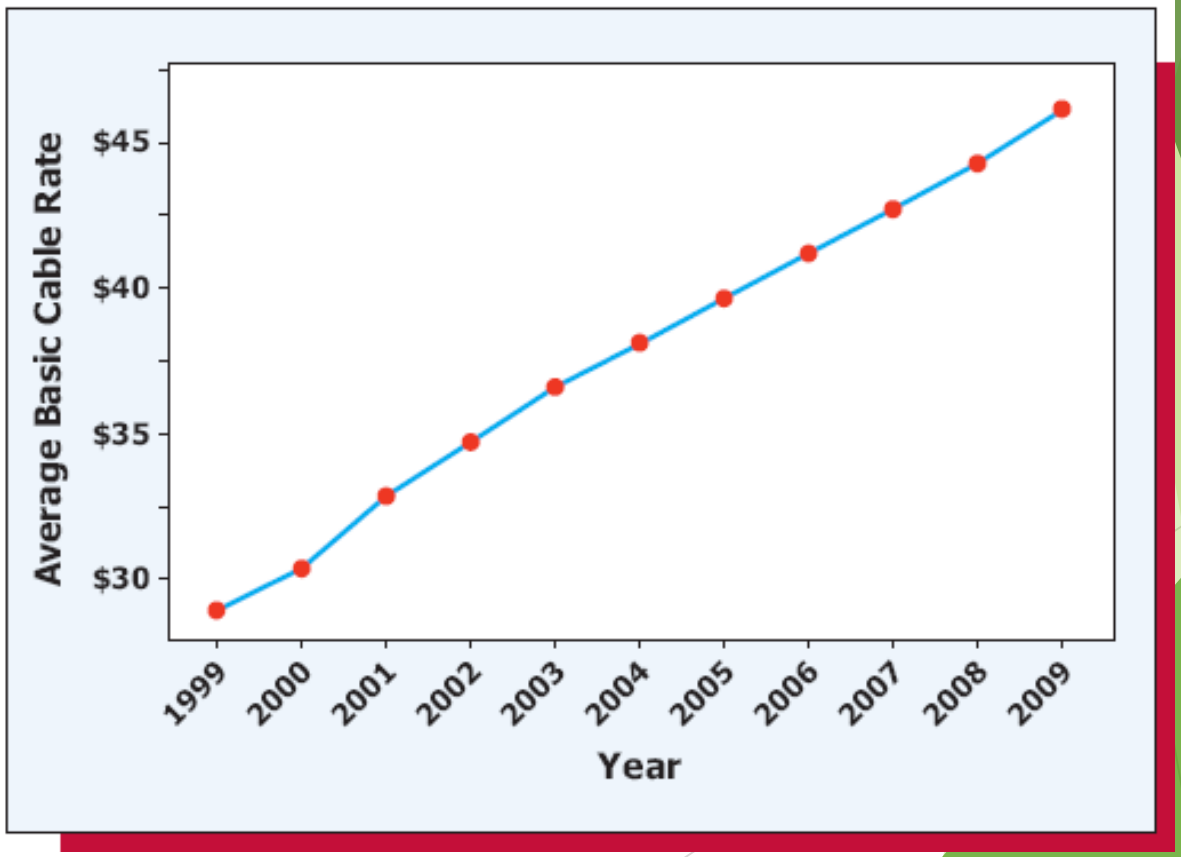
1.1 Cross-Sectional Data

1. **Cross-sectional data**: Data collected at the same or approximately the same point in time
2. **Time series data**: data collected over different time periods

Time Series Data

Year	Average Basic Cable Rate
1999	$ 28.92
2000	30.37
2001	32.87
2002	34.71
2003	36.59
2004	38.14
2005	39.63
2006	41.17
2007	42.72
2008	44.28
2009	46.13

Source: U.S. Energy Information Administration, http://www.eia.gov/



Populations and Samples

Population

the complete set of all individuals, objects, or elements (people, objects or events) that share common characteristics and are the subject of a statistical analysis.

- Complete Set: A population includes every member or element that fits the criteria being studied. E.g., when studying the heights of all students in a particular school, the population would be all the students in the school.
- Finite vs. Infinite: Populations can be finite (having a specific, countable number of elements) or infinite (theoretically unlimited). E.g., the number of grains of sand on a beach can be considered an infinite population for practical purposes.
- Parameters: Characteristics of a population are called parameters. Usually denoted by Greek letters (e.g., μ for the population mean, σ for the population standard deviation). Parameters are fixed values that describe the entire population.

Populations and Samples

Population	the complete set of all individuals, objects, or elements (people, objects or events) that share common characteristics and are the subject of a statistical analysis.
Sample	A smaller subset of the population. Make inferences about the sampled population

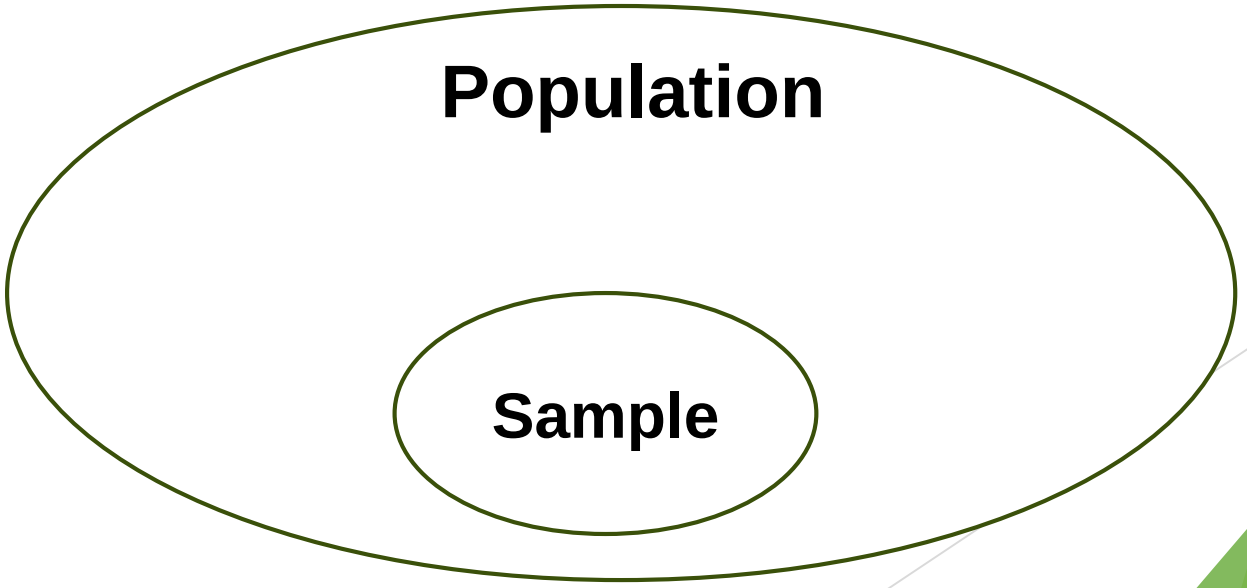
- Sampling: Because populations can be very large or even infinite, it is often impractical or impossible to study every member. Instead, statisticians use samples, which are subsets of the population, to make inferences about the population as a whole.
- Population vs. Sample: The population is the entire group of interest, while a sample is a smaller subset selected from the population. For example, if you want to know the average height of all adult males in a country (population), you might measure the heights of a few thousand adult males (sample) and use that data to estimate the population parameter.

Census and Sample

Census An official count or survey of a population, typically recording various details about individuals: age, job, income...

Note: Census usually too expensive, too time consuming, and too much effort for a large population.

Sample A selected subset of the units of a population.



Example

A university graduated 8,742 students

a. This is too large for a census.

b. So, we select a sample of these graduates (e.g., 100 graduates) and learn their annual starting salaries.

Sample of measurements

- Measured values of the variable of interest for the sample units.
- For example, the actual annual starting salaries of the sampled graduates.

Population of Measurements

- Measurement of the variable of interest for each and every population unit

  For example, annual starting salaries of all graduates from last year's MBA program

- Sometimes called *observations*
- If population too large, will analyze a subset

Descriptive statistics

The science of describing the main features of a dataset using numerical and graphical techniques. provide a concise summary of the data's central tendency, variability, and distribution shape, giving you an initial understanding without making inferences beyond the data at hand.

- For a dataset of exam scores:
- Data: 55, 65, 70, 75, 75, 80, 85, 90, 95, 100
- Mean: 79; Median: 77.5; Mode: 75; Range: 45; Standard Deviation: ~13.4
- If the population is small enough, could take a census and not have to sample and make any statistical inference
- But if the population is too large, then use sampling and inference

Statistical Inference

The science of using a sample of measurements to make generalizations about the important aspects of a population of measurements.

- For example, use a sample of starting salaries to estimate the important aspects of the population of starting salaries

There is a criteria on how to choose a sample:

the information contained in a sample can accurately reflect the population under study.

1.2 Selecting a Random Sample

Random sample

A random sample is a sample selected from a population so that:

- Each population unit has the same chance of being selected as every other unit
  - Each possible sample (of the same size) has the same chance of being selected
- For example, randomly pick two different people from a group of 101:
  - Number the people from 1 to 101; and write their numbers on 101 different slips of paper
  - Thoroughly mix the papers and randomly pick two of them (random draw; draw lots)
  - The numbers on the slips identifies the people for the sample

Sample with replacement

Replace each sampled unit before picking next unit

- The unit is placed back into the population for possible reselection
- However, the same unit in the sample does not contribute new information

Sample without replacement

A sampled unit is withheld from possibly being selected again in the same sample

- Guarantees a sample of different units
  - Each sampled unit contributes different information
  - Sampling without replacement is the usual and customary sampling method

Approximately Random Samples

Sometimes it is not possible to list and number all the units in a population. In such a situation we often select a systematic sample, which approximates a random sample.

A Systematic Sample

Randomly enter the population and systematically sample every *k*th unit.

Three Case Studies that Illustrate Sampling and Statistical Inference

1. The Cell Phone Case: Estimating Cell Phone Costs
2. The Marketing Research Case: Rating a New Bottle Design
3. The Car Mileage Case: Estimating Mileage

Example 1.1: The Cell Phone Case: Estimating Cell Phone Costs

- A bank considering using a service to manage their cellular resources
- Many overages and underage
- Random sample of 100 employees on 500-minute plan

- The bank has over 10,000 employees on many different types of calling plans.
- A cellular management service suggests that by studying the calling patterns of cellular users on 500-minute-per-month plans, the bank can accurately assess whether its cell phone costs can be substantially reduced. The bank has 2,136 employees on a variety of 500-minute-per-month plans with different basic monthly rates, different overage charges, and different additional charges for long distance and roaming.
- The bank will estimate its cellular costs for the 500-minute plans by analyzing last month's cell phone bills for a random sample of 100 employees on these plans.

The Cell Phone Case: The Data

Table 1.2

A Sample of Cellular Usages (in minutes) for 100 Randomly Selected Employees

75	485	37	547	753	93	897	694	797	477
654	578	504	670	490	225	509	247	597	173
496	553	0	198	507	157	672	296	774	479
0	822	705	814	20	513	546	801	721	273
879	433	420	521	648	41	528	359	367	948
511	704	535	585	341	530	216	512	491	0
542	562	49	505	461	496	241	624	885	259
571	338	503	529	737	444	372	555	290	830
719	120	468	730	853	18	479	144	24	513
482	683	212	418	399	376	323	173	669	611

- There is substantial overage and underage—many employees used far more than 500 minutes, while many others failed to use all of the 500 minutes allowed by their plan.
- In Chapter 3 we will use these 100 usage figures to estimate the bank's cellular costs and decide whether the bank should hire a cellular management service.

Example 1.2: The Marketing Research Case: Rating a New Bottle Design

- A brand group wishes to research consumer reaction to a new bottle design for a popular soft drink.
- Using “mall intercept method” to select a sample of consumers
- Interviewed a sample of 60 shoppers at a mall on a particular Saturday (Sample size = 60)

FIGURE 1.4 The Bottle Design Survey Instrument

Please circle the response that most accurately describes whether you agree or disagree with each statement about the bottle you have examined.

Statement	Strongly Disagree						Strongly Agree
The size of this bottle is convenient.	1	2	3	4	5	6	7
The contoured shape of this bottle is easy to handle.	1	2	3	4	5	6	7
The label on this bottle is easy to read.	1	2	3	4	5	6	7
This bottle is easy to open.	1	2	3	4	5	6	7
Based on its overall appeal, I like this bottle design.	1	2	3	4	5	6	7

The Marketing Research Case: The Form and the Data

A Sample of Bottle Design Ratings (Composite Scores for a Sample of 60 Shoppers)

DS Design

34	33	33	29	26	33	28	25	32	33
32	25	27	33	22	27	32	33	32	29
24	30	20	34	31	32	30	35	33	31
32	28	30	31	31	33	29	27	34	31
31	28	33	31	32	28	26	29	32	34
32	30	34	32	30	30	32	31	29	33

- 57 of the 60 composite scores are at least 25.
- A proportion of 57/60 =0.95 (that is, 95 percent) of all consumers would give the bottle design a composite score of at least 25.

Terms

- **Process**: a sequence of operations that takes inputs and turns them into outputs
- **Finite population**: a population of limited size
- **Infinite population**: a population of unlimited size

Sampling a Process

Process

A sequence of operations that takes *inputs* (labor, raw materials, methods, machines, and so on) and turns them into *outputs* (products, services, and the like)

Processes produce output over time

- The “population” from a process is all output produced in the past, present, and the yet-to-occur future.
- For example, coffees of a particular coffee shop, like Starbucks
  - Coffees will continue to be made over time

Example

The Coffee Temperature Case: Monitoring Coffee Temperatures

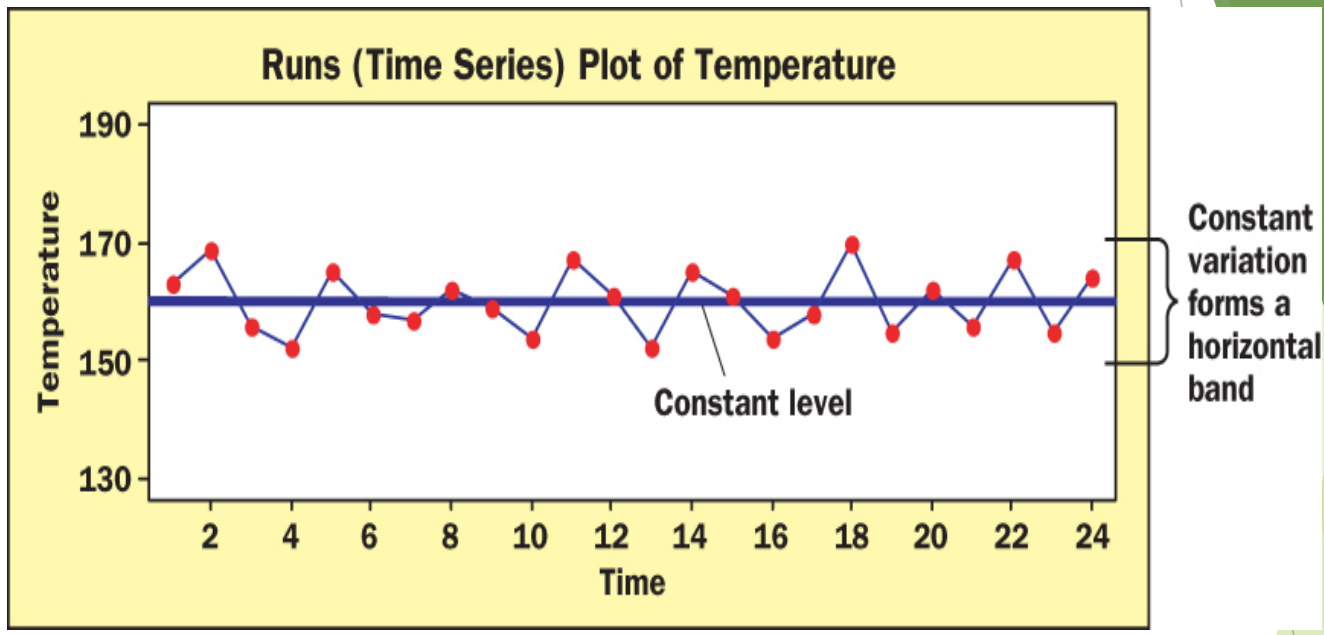
Coffee temperatures at a fast-food restaurant. The restaurant personnel measures the temperature of the coffee being dispensed (in degrees Fahrenheit (°F)) at half-hour intervals from 10 A.M. to 9:30 P.M. on a given day.

TABLE 1.7 24 Coffee Temperatures Observed in Time Order (°F) Coffee

Time		Coffee Temperature	Time		Coffee Temperature	Time		Coffee Temperature
(10:00 A.M.)	1	163°F	(2:00 P.M.)	9	159°F	(6:00 P.M.)	17	158°F
	2	169		10	154		18	170
	3	156		11	167		19	155
	4	152		12	161		20	162
(12:00 noon)	5	165	(4:00 P.M.)	13	152	(8:00 P.M.)	21	156
	6	158		14	165		22	167
	7	157		15	161		23	155
	8	162		16	154		24	164

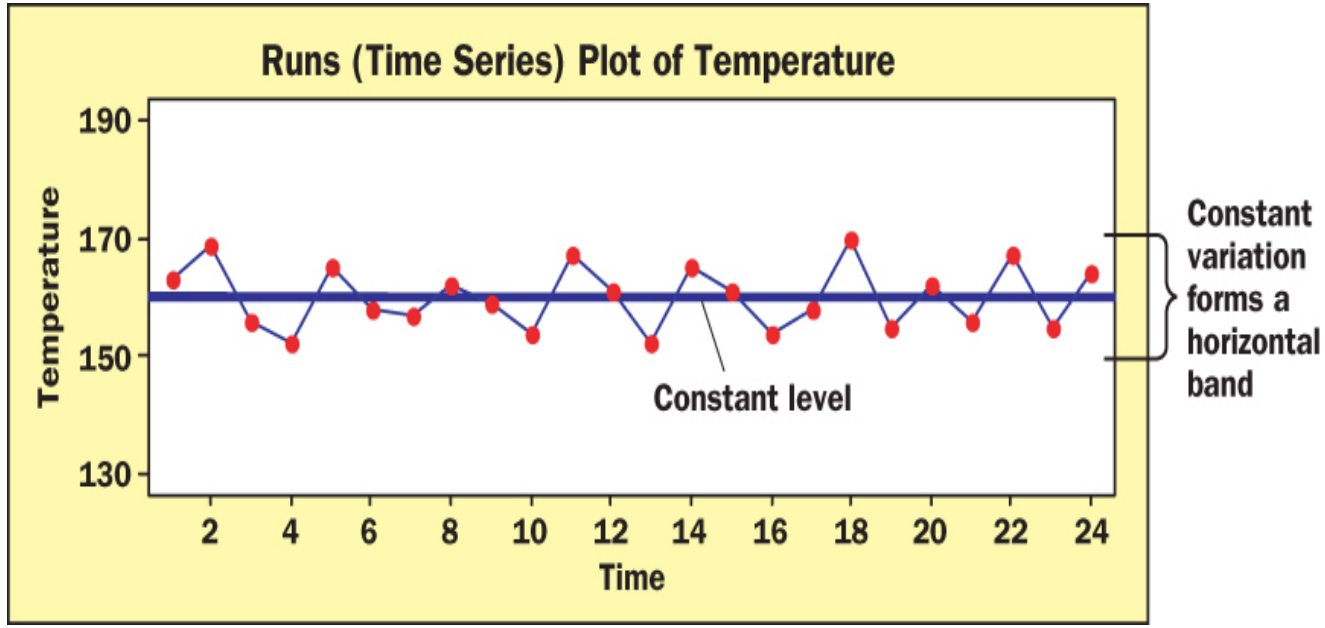
185 Fahrenheit (°F)= 85 °C
158 °F= 70 °C

Runs plot and statistical control



- A *runs plot* is a graph of individual process measurements over time. It help us see the trend of the variable of interest.
- A process is in statistical control if it does not exhibit any unusual process variations.
- To determine if a process is in statistical control or not, sample the process often enough to detect unusual variations

Runs plot and statistical control



- Over time, temperatures have a fairly constant amount of variation around a fairly constant level
  - The temperature is expected to be at the constant level shown by the horizontal blue line
    - Sometimes the temperature is higher and sometimes lower than the constant level
  - About the same amount of spread of the values (data points) around the constant level
    - The points are as far above the line as below it
    - The data points appear to form a horizontal band
- So, the process is in statistical control
  - Coffee-making process is operating “consistently”

Example 1.3: The Car Mileage Case: Estimating Mileage

- Personal budgets, national energy security, and the global environment are all affected by our gasoline consumption. Hybrid and electric cars are a vital part of a long-term strategy to reduce our nation's gasoline consumption.
- Government has decided to offer a tax credit to any automaker selling a midsize model with an automatic transmission that achieves an EPA combined city and highway mileage estimate of at least 31 mpg (miles per gallon).
- Automaker has introduced a new model and wishes to demonstrate that it qualifies for the tax credit.
- Sample of 50 cars

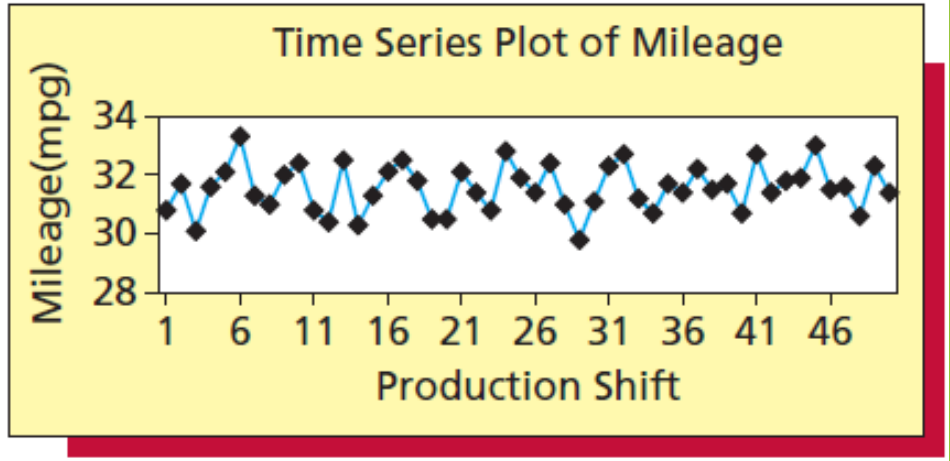
The Care Mileage Case: The Data

A Sample of 50 Mileages

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

Note: Time order is given by reading down the columns from left to right.

A Time Series Plot of the 50 Mileages



- the mileages vary over time, but do not vary in any unusual way
- the mileages do not tend to either decrease or increase
- the maximum and minimum are not too far from the constant level (31.56 mpg)
- The data points appear to form a horizontal band with even spread of the values (data points)
- the midsize car manufacturing process is producing consistent car mileages over time

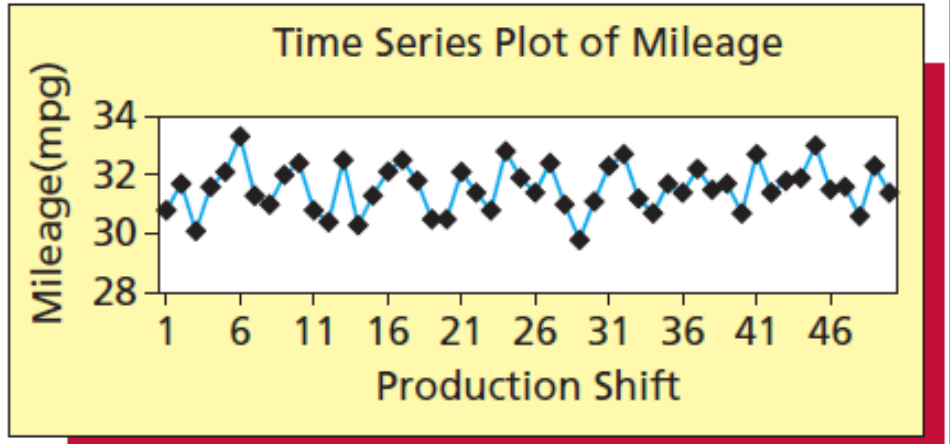
The Care Mileage Case: The Data

A Sample of 50 Mileages

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

Note: Time order is given by reading down the columns from left to right.

A Time Series Plot of the 50 Mileages



the mileages vary over time, but do not vary in any unusual way

- A minimum of 29.8 mpg to a maximum of 33.3 mpg
- 38 out of the 50 mileages—or 76 % mileages—are greater than or equal to the tax credit standard of 31 mpg
- the “typical car” produced by the process will meet or exceed the tax credit standard.

Summary

- We began this chapter by discussing data. We learned that the data that are collected for a particular study are referred to as a data set, and we learned that elements are the entities described by a data set. Quantitative variables are variables that use numbers to measure quantities (that is, “how much” or “how many”) and qualitative, or categorical, variables simply record into which of several categories an element falls.
- We next discussed the difference between cross-sectional data and time series data. Cross-sectional data are data collected at the same or approximately the same point in time. Time series data are data collected over different time periods.
- We often collect data to study a population, which is the set of all elements about which we wish to draw conclusions. We saw that, because many populations are too large to examine in their entirety, we frequently study a population by selecting a sample, which is a subset of the population elements. Next we learned that, if the information contained in a sample is to accurately represent the population, then the sample should be randomly selected from the population.

Thank you!

1.3 Ratio, Interval, Ordinal, and Nominative Scales of Measurement (Optional)

- Nominative
- Ordinal
- Interval
- Ratio

Qualitative Variables

- **Nominative:** A qualitative variable for which there is no meaningful ordering, or ranking, of the categories
  - Example: gender, car color
- **Ordinal:** A qualitative variable for which there is a meaningful ordering, or ranking, of the categories
  - Example: teaching effectiveness

Interval Variable

- All of the characteristics of ordinal
- Measurements are on a numerical scale with an arbitrary zero point
  - The “zero” is assigned: it is nonphysical and not meaningful
  - Zero does not mean the absence of the quantity that we are trying to measure

Interval Variable

- Can only meaningfully compare values by the interval between them
  - Cannot compare values by taking their ratios
  - "Interval" is the arithmetic difference between the values
- Example: temperature
  - 0 ℃ means "cold," not "no heat"
  - 60 ℃ is ***not*** twice as warm as 30 ℃

Ratio Variable

- Measurements are on a numerical scale with a meaningful zero point
  - Zero means “none” or “nothing”
- Values can be compared in terms of their interval and ratio
  - $30 is $20 more than $10
  - $0 means no money

Ratio Variable

- In business and finance, most quantitative variables are ratio variables, such as anything to do with money
  - Examples: Earnings, profit, loss, age, distance, height, weight

1.4 An Introduction to Survey Sampling (Optional)

- Methods for obtaining a sample are called sampling designs. The sample we take is sometimes called a sample survey.
- Stratified random sampling, cluster sampling, and systematic sampling
- In order to select a stratified random sample, we divide the population into nonoverlapping groups of similar units (people, objects, etc.). These groups are called strata. Then a random sample is selected from each stratum, and these samples are combined to form the full sample.

- Multi-stage cluster sampling
- Systematic sampling

1.5 More about Data Acquisition and Survey Sampling (Optional)

- Existing sources: data already gathered by public or private sources
  - Internet
  - Library
  - Private data sources
- Experimental and observational studies: data we collect ourselves for a specific purpose
  - Response variable: variable of interest
  - Factors: other variables related to response variable

  Example: A designed experiment

  A survey

  An observational study

Initiating a Study

- First, define the variable of interest, called a **response variable**
- Next, define other variables that may be related to the variable of interest and will be measured, called **independent variables**
- If we manipulate the independent variables, we have an **experimental study**
- If unable to control independent variables, the study is **observational**