Disadvantages over Survey Method:

- things such as awareness, beliefs, feelings and preferences cannot be observed
- the observed behavior patterns can be rare or too unpredictable thus increasing the data collection costs and time requirements

3. Experimental method - a method designed for collecting data under controlled conditions. An experiment is an operation where there is actual human interference with the conditions that can affect the variable under study. This is an excellent method of collecting data for *causation studies*. If properly designed and executed, experiments will reveal with a good deal of accuracy, the effect of a change in one variable on another variable.
4. Use of existing studies - e.g., census, health statistics, and weather bureau reports

Two types:

- documentary sources - published or written reports, periodicals, unpublished documents, etc.
- field sources – researchers who have done studies on the area of interest are asked personally or directly for information needed

5. Registration method - e.g., car registration, student registration, and hospital admission

General Classification of Collecting Data

Definition. **Census** or **complete enumeration** is the process of gathering information from every unit in the population.

- not always possible to get timely, accurate and economical data
- costly, especially if the number of units in the population is too large

Definition. **Survey sampling** is the process of obtaining information from the units in the selected sample.

Advantages of Survey Sampling:

- reduced cost
- greater speed

Chapter 3 PRESENTATION OF DATA

Textual Presentation

- data incorporated to a paragraph of text

Example

> At last count, 38 airlines were operating Boeing 707's, 720's, and 727's over the world's airlines. The far-flung Boeing fleet has now logged an estimated 1,803,704,000 miles (22,855,948,000 kms.) and has massed approximately 4,096,000 revenue flight hours. Passenger totals stand at upwards of 71.6 million.

Advantages

- This presentation gives emphasis to significant figures and comparisons
- It is simplest and most appropriate approach when there are only a few numbers to be presented

Disadvantages

- When a large mass of quantitative data are included in a text or paragraph, the presentation becomes almost incomprehensible
- Paragraphs can be tiresome to read especially if the same words are repeated so many times

Tabular Presentation

- the systematic organization of data in rows and columns

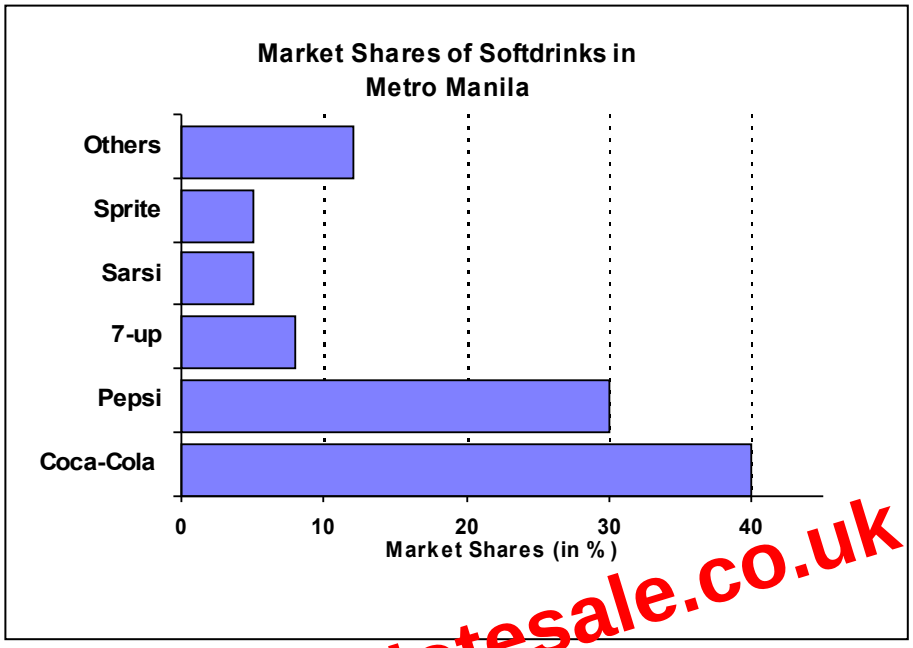
Advantages

- more concise than textual presentation
- easier to understand
- facilitates comparisons and analysis of relationship among different categories
- presents data in greater detail than a graph

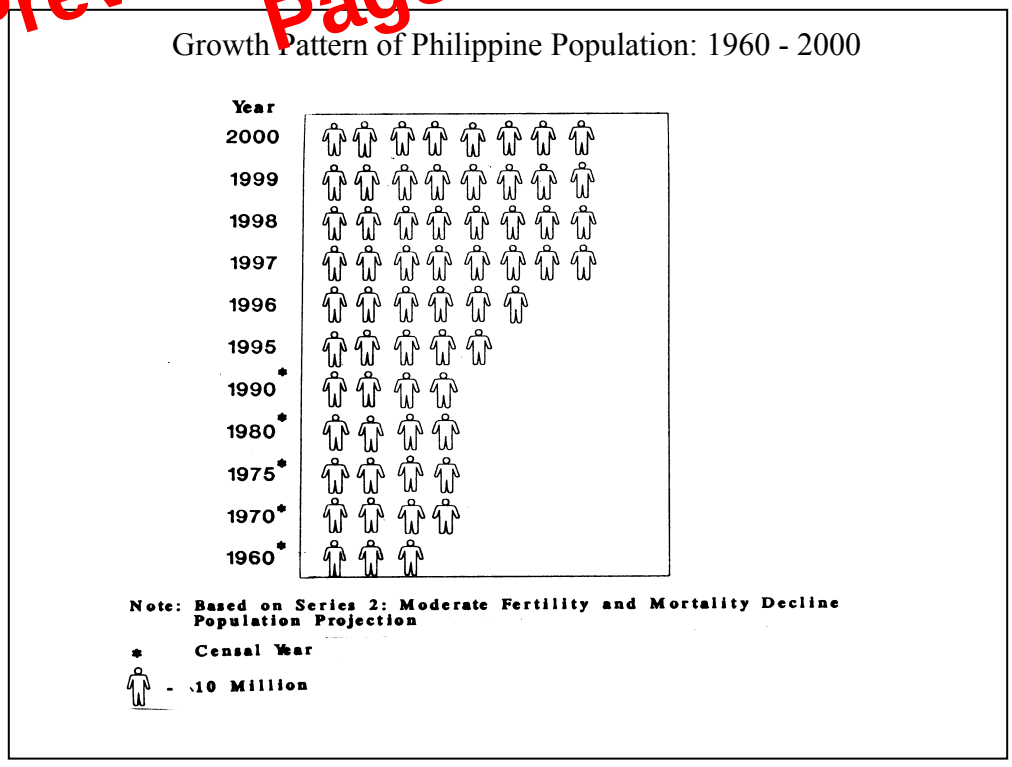
Parts of a Formal Statistical Table

1. Heading - consists of a *table number, title*, and *headnote*. The title is a brief statement of the nature, classification and time reference of the information presented and the area to which the statistics refer. The headnote is a statement enclosed in

3. Bar Chart - consists of a series of rectangular bars where the length of the bar represents the quantity or frequency for each category if the bars are arranged horizontally. If the bars are arranged vertically, the height of the bar represents the quantity.



4. Pictorial unit chart - a pictorial chart in which each symbol represents a definite and uniform value



Pearson's First and Second Coefficients of Skewness

1. $Sk = \dfrac{\overline{X} - Mo}{s}$

where $\overline{X}$ = mean
Mo = mode
s = standard deviation

2. $Sk = \dfrac{3(\overline{X} - Md)}{s}$

where $\overline{X}$ = mean
Md = median
s = standard deviation

Remarks:

1. Since the mode is frequently only an approximation, formula 2 is preferred.

2. Interpretation of the measure of skewness:

Sk > 0: positively skewed since $\overline{X}$ > Md > Mo
Sk < 0: negatively skewed since $\overline{X}$ < Md < Mo
Sk = 0: symmetric since $\overline{X}$ = Md = Mo

Example: Refer to the final grade of 110 Soc Sci 101 students.

$\overline{X}$ = 74.1 Md = 75 Mo = 84 s = 11.25

Using the first formula,

$$Sk = \frac{74.1 - 84}{11.25} = -0.88$$

Using the second formula,

$$Sk = \frac{3(74.1 - 75)}{11.25} = -0.24$$

THE PROBABILITY CONCEPT AND SOME PROPERTIES

Probability analysis is based on the following simple postulates.

Postulate 1. $0 \le P(O_i) \le 1$ for any simple event O_i

Postulate 2. The probability for any event E is the sum of the probability of the simple events that constitute E.

Postulate 3. $P(S) = 1$, where S is the sample space, and $P(\phi) = 0$, where ϕ is the null space.

Approaches to Assigning Probabilities

1. **A Priori or Classical Probability** – probability is determined even before the experiment is performed using the following rule:

   If an experiment can result in any one of N different equally likely outcomes, and if exactly n of these outcomes correspond to event A, then the probability of event A is

$$P(A) = \frac{\text{no. of sample points in A}}{\text{no. of sample points in S}} = \frac{n}{N}$$

2. **A Posteriori or Relative Frequency or Empirical Probability** - probability is determined by repeating the experiment a large number of times using the following rule:

$$P(A) = \frac{\text{no. of times event A occurred}}{\text{no. of times experiment was repeated}}$$

3. **Subjective Probability** – probability is determined by the use of intuition, personal beliefs, and other indirect information.

Examples:

1. Find the errors in each of the following statements:

   a. The probability that it will rain tomorrow is 0.40 and the probability that it will not rain tomorrow is 0.52.
   b. The probabilities that a printer will make 0, 1, 2, 3, or 4 or more mistakes in printing a document are, respectively, 0.19, 0.34, -0.25, 0.43, and 0.29.
   c. The probabilities that an automobile salesperson will sell 0, 1, 2, or 3 cars on any given day in February are, respectively, 0.19, 0.38, 0.29, and 0.15.
   d. On a single draw from a deck of playing cards the probability of selecting a heart is 1/4, the probability of selecting a black card is 1/2, and the probability of selecting both a heart and a black card is 1/8.

2. Answer the following:
   a. In tossing a fair coin, what is the probability of getting a head? Of either a head or tail? Of neither a head nor tail?
   b. In tossing a fair die, what is the probability of getting a 3? Of getting an even number? Of getting a number greater than 6?

3. A coin is biased so that a head is twice as likely to occur as a tail. If the coin is tossed once, what is the probability of getting a head?

Rules of Counting *(Optional)*

Theorem. If an operation can be performed in n_1 ways, and for each of these a second operation can be performed in n_2 ways, then the two operations can be performed in $n_1 n_2$ ways.

Example:

1. How many sample points are there in the sample space when a pair of balanced dice is thrown once?
2. If a travel agency offers special weekend trips to 12 different cities, by air, rail, or bus, in how many ways can such a trip be arranged?

Theorem. **(Multiplication Rule)** If an operation can be performed in n_1 ways, if for each of these a second operation can be performed in n_2 ways, if for each of the first two a third operation can be performed in n_3 ways, and so on, then the sequence of k operations can be performed in $n_1 n_2 \ldots n_k$ ways.

Examples:

1. How many even three-digit numbers can be formed from the digits 3, 5, 6, 7, 9, 2 if
   a. all digits are distinct?
   b. numbers must be divisible by 5 and repetition is not allowed?
   c. Numbers must be even and repetition is not allowed?

2. The probability density function can not be represented in tabular form.

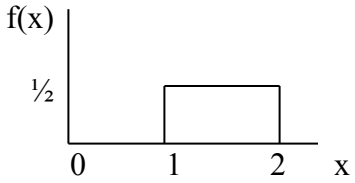
Example:

A continuous random variable X that can assume values between 0 and 2 has a density function given by

$$f(x) = \begin{cases} 0.5 & \text{for } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the following probabilities:

a. P(1 < X < 2).



area of shaded region = (length)(width) = P(1 < X < 2) = (1/2)(1) = 1/2

b. P(X > 1.5)
c. P(X < 0.75)
d. P(X = 0.75)
e. P(X ≤ 0.75).

EXPECTED VALUES

Definition. Let X be a discrete random variable with probability distribution

x	x_1	x_2	. . .	x_n
P(X=x)	$f(x_1)$	$f(x_2)$	. . .	$f(x_n)$

The **mean** or **expected value** of X is

$$\mu = E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Examples:

1. Find the mean of the random variables X and Y of Experiment No. 1.

X	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

E(X) = (0)(1/8) + (1)(3/8) + (2)(3/8) + (3)(1/8) = 12/8 or 1.5

Y	-3	-1	1	3

P(Y=y)	1/8	3/8	3/8	1/8

$E(Y) = (-3)(1/8) + (-1)(3/8) + (1)(3/8) + (3)(1/8) = 0$

2. Find the expected number of correct matches in Experiment No. 2.
3. In a gambling game a man is paid P50 if he gets all heads or all tails when 3 coins are tossed, and he pays out P30 if either 1 or 2 heads show. What is his expected gain?

Theorem. Let X be a discrete random variable with probability distribution

x	x_1	x_2	. . .	x_n
P(X=x)	$f(x_1)$	$f(x_2)$	. . .	$f(x_n)$

The **mean** or **expected value** of the random variable g(X) is

$$E(g(X)) = \sum_{i=1}^{n} g(x_i) f(x_i)$$

Example: A used car dealer finds that in any day, the probability of selling no car is 0.4, one car is 0.2, two cars is 0.15, 3 cars is 0.10, 4 cars is 0.08, five cars is 0.06 and six cars is 0.01. Let g(X) = 500 + 1500X represent the salesman's daily earnings, where X is the number of cars sold. Find the salesman's expected daily earnings.

Definition. **Let X be a random variable with mean μ, then the** variance **of X is**

$$\sigma^2 = Var(X) = E(X - \mu)^2$$

Definition. **Let X be a discrete random variable with probability distribution**

X	x_1	x_2	. . .	x_n
P(X=x)	$f(x_1)$	$f(x_2)$	. . .	$f(x_n)$

The **variance** of X is

$$\sigma^2 = Var(X) = E(X - \mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$$

Theorem. Computational Formula for σ^2

$$Var(X) = E(X^2) - [E(X)]^2$$

Example : In Experiment No. 1, find the variance of X..

Using the definition of Var(X),

$E(X) = 1.5$

$$Var(X) = \sum_{i=1}^{4} (x_i - 1.5)^2 f(x_i)$$

$$= (0-1.5)^2(1/8) + (1-1.5)^2(3/8) + (2-1.5)^2(3/8) + (3-1.5)^2(1/8) = 0.75$$

Using the computational formula of the Var(X),

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 3 - (1.5)^2 = 0.75$$

Properties of the Mean and Variance

Let X and Y be random variables (discrete or continuous) and let **a** and **b** be constants.

1. $E(aX + b) = a\,E(X) + b$

   Special Cases:
   a. if $b = 0$, then $E(aX) = a\,E(X)$.
   b. if $a = 0$, then $E(b) = b$.

2. $E(X+Y) = E(X) + E(Y)$
   $E(X-Y) = E(X) - E(Y)$

3. $E(XY) = E(X)E(Y)$ if X and Y are independent.

4. $E[\,X - E(X)\,] = 0$.

5. $\text{Var}(aX + b) = a^2\text{Var}(X)$.

   Special Cases:
   a. if $b = 0$, then $\text{Var}(aX) = a^2\text{Var}(X)$.
   b. if $a = 0$, then $\text{Var}(b) = 0$.

6. If X and Y are independent then

   $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
   $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$

Example :

If X and Y are independent random variables with $E(X) = 3$, $E(Y) = 2$, $\text{Var}(X) = 2$ and $\text{Var}(Y)=1$, find

a. $E(3X + 5)$
b. $\text{Var}(3X +5)$
c. $E(XY)$
c. $\text{Var}(3X - 2Y)$

- Notation: t_α is the t-value leaving an area of α in the right-tail of the t-distribution. That is, if $T \sim t_{(v)}$ then t_α is such that $P(T > t_\alpha) = \alpha$.

- Since the t-distribution is symmetric about zero, $t_{1-\alpha} = -t_\alpha$

Examples:

1. Find the following values on the t -table:

   (a) $t_{0.025}$ when v = 14.
   (b) $t_{0.99}$ when v=10.

2. Find k such that $P(k < T < 2.807) = 0.945$ when $T \sim t_{(23)}$

3. A manufacturing firm claims that the batteries used in their electronic games will last an average of 30 hours. To maintain this average, 16 batteries are tested each month. If the computed t-value falls between $-t_{0.025}$ and $t_{0.025}$, the firm is satisfied with its claim. What conclusion should the firm draw from a sample that has mean $\overline{X}$ = 27.5 hours and standard deviation S = 5 hours? Assume the distribution of battery lives to be approximately normal.

Chapter 12 Estimation of Parameters

Definition. **Statistical inference** refers to methods by which one uses sample information to make inferences or generalizations about a population.

Two Areas of Statistical Inference

1. Estimation
   - point estimation
   - interval estimation

2. Hypothesis Testing

BASIC CONCEPTS IN ESTIMATION

Point Estimation

Definition. An **estimator** is any statistic whose value is used to estimate an unknown parameter. A realized value of an estimator is called an **estimate**.

For example, the sample mean $\overline{X}$ is an estimator of the population mean μ.

Remarks:

1. An estimator is said to be **unbiased** if the average of the estimates it produces under repeated sampling is equal to the true value of the parameter being estimated.

   Examples: Under random sampling, the sample mean is an unbiased estimator of the population mean, that is, $E(\overline{X}) = \mu$.

   Under random sampling with replacement, S^2 is an unbiased estimator of σ^2, but S on the other hand is a biased estimator of σ with the bias becoming insignificant for large samples.

2. A parameter can have more than one unbiased estimator. We would naturally choose the unbiased estimator with the smallest variance.

Interval Estimation

Definition. An **interval estimator** of a population parameter is a rule that tells us how to calculate two numbers based on sample data, forming an interval within which the parameter is expected to lie. This pair of numbers, *(a,b),* is called an **interval estimate** or **confidence interval**.

Examples:

1. An electrical firm manufactures light bulbs that have a length of life that is normally distributed, with a standard deviation of 40 hours. If a random sample of 25 bulbs has a mean life of 780 hours, find a 95% confidence interval for the population mean of all bulbs produced by this firm.

2. Regular consumption of presweetened cereals contribute to tooth decay, heart disease, and other degenerative diseases, according to a study by Dr. M. Albreight of the National Institute of Health and Dr. D. Solomon, Professor of Nutrition and Dietetics at the University of London. In a random sample of 20 similar servings of Alpha-Bits, the mean sugar content was 11.3 grams with a standard deviation of 2.45 grams. Assuming that the sugar content is normally distributed, construct a 95% confidence interval for the mean sugar content for single servings of Alpha-Bits.

3. A random sample of 100 automobile owners shows that an automobile is driven on the average 23,500 kilometers per year, in the state of Virginia, with a standard deviation of 3900 kilometers. Construct a 99% confidence interval for the average number of kilometers an automobile is driven annually in Virginia.

ESTIMATING THE DIFFERENCE BETWEEN TWO POPULATION MEANS

If we have two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively, a point estimator of the difference between μ_1 and μ_2 is the statistic $\overline{X}_1 - \overline{X}_2$.

Types of Sampling:

- selecting two independent samples
- paired sampling

Paired sampling is used to overcome the difficulty imposed by extraneous differences between two groups when testing the difference between 2 means. This is achieved by "matching" or studying 2 related samples. Matching may be achieved by:

- using the same subject in the 2 samples
- pairing of subjects with respect to any extraneous variable which might affect or influence the outcome.

Example:

In a random sample of 200 students who enrolled in Math 17, 138 passed on their first take. Construct a 95% confidence interval for the population proportion of students who passed Math 17 on their first take.

ESTIMATING THE DIFFERENCE OF TWO PROPORTIONS

Given 2 independent random samples of size n_1 and n_2 , a point estimator of the difference between the two proportions p_1 and p_2 is given by $\hat{p}_1 - \hat{p}_2 = \frac{X}{n_1} - \frac{Y}{n_2}$, where X is the number of successes in n_1 trials (first sample) and Y is the number of successes in n_2 trials (second sample).

An approximate (1-α)100% confidence interval for $p_1 - p_2$ when n_1 and n_2 are large is

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \right)$$

Example:

In a random sample of 200 students, 78 of the 120 females and 60 of the 80 males passed Math 17 on their first take. Construct a 95% confidence interval for $p_1 - p_2$, where p_1 and p_2 are the true proportions of females and males, respectively, who passed Math 17 on their first take.

SAMPLE SIZE DETERMINATION

Sample Size for Estimating μ

In random sampling, if $\overline{X}$ will be used to estimate μ, we can be (1-α)100% confident that that the error will not exceed a specified amount, e, when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e} \right)^2$$

- Based on 2 independent samples

Ho	Test Statistic	Ha	Critical region
a. σ_1^2 and σ_2^2 known			
$\mu_1 - \mu_2 = d_o$	$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_o}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$	$\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$ $\mu_1 - \mu_2 \neq d_o$	$z < -z_\alpha$ $z > z_\alpha$ $\lvert z \rvert > z_{\alpha/2}$
b. $\sigma_1^2 = \sigma_2^2$ but unknown			
$\mu_1 - \mu_2 = d_o$	$t = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_o}{S_p\sqrt{(1/n_1) + (1/n_2)}}$ $\upsilon = n_1 + n_2 - 2$ $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$ $\mu_1 - \mu_2 \neq d_o$	$t < -t_\alpha$ $t > t_\alpha$ $\lvert t \rvert > t_{\alpha/2}$
c. $\sigma_1^2 \neq \sigma_2^2$ and unknown			
$\mu_1 - \mu_2 = d_o$	$t = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_o}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}$ $\upsilon = \dfrac{(S_1^2/n_1 + S_2^2/n_2)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$	$\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$ $\mu_1 - \mu_2 \neq d_o$	$t < -t_\alpha$ $t > t_\alpha$ $\lvert t \rvert > t_{\alpha/2}$

- Based on 2 related samples

Ho	Test Statistic	Ha	Critical region
$\mu_D = d_o$	$t = \dfrac{\bar{d} - d_o}{S_d/\sqrt{n}}$ $\upsilon = n - 1$	$\mu_D < d_o$ $\mu_D > d_o$ $\mu_D \neq d_o$	$t < -t_\alpha$ $t > t_\alpha$ $\lvert t \rvert > t_{\alpha/2}$

Remark: The remarks made in Chapter 8.3 relative to the use of a given statistic apply to the tests described here.

Examples:

1. A statistics test was given to 50 girls and 75 boys. The girls made an average of 80 with a standard deviation of 4 and the boys had an average of 86 with a standard deviation of 6. Is there sufficient evidence at 0.05 level of significance that the average grades of girls and boys differ?

2. A study was made to determine if the subject matter in a physics course is better understood when a lab constitutes part of the course. Students were allowed to choose between a 3-unit course without lab and a 4-unit course with lab. In the section with lab, a sample of 11 students had an average grade of 85 with a standard deviation of 4.7, and in the section without lab, a sample of 17 students had an average grade of 79 with a standard deviation of