Chapter 10: Probability Theory and Sampling

Did you hear the one about the statistician?
Probably....

Learning Objectives

Upon completion of this chapter, students should know

- The basics of probability theory.
- How to compute probability of events using the addition and multiplication rules of probability.
- The importance of sampling methods.
- The distinct between random sampling and nonrandom sampling.
- How to define and explain the concepts of the sampling distribution of the mean and central limit theorem.
- How to define and compute the standard error of the mean.
- When to use and how to compute a *z* test.

Key Terms

Probability is a measure of how likely it is that a given even or behavior will happen.

Mutually exclusive is a term used to indicate that two events cannot occur simultaneously.

Addition rule of probability or addition theorem of probability is used to compute the probability of one event or another event occurring. *p* (A or B) = *p* (A) + *p* (B)

Multiplication rule of probability or multiplication theorem of probability is used to compute the probability of two mutually events occurring together. *p* (A or B) = $p(A) \cdot p(B)$

Representative sample is a sample in which all significant subgroups of the population are represented.

Random sampling is a sampling technique used to increase the chances of obtaining a representative sample. It assumes that by using a completely random selection process that everyone in the population is equally likely to be selected.

Random number table is a table of numbers that were randomly generated by a computer. It can be used as a means of generating a random sample of a population.

Distribution of sample means is a theoretical distribution made up of an unlimited number of sample means. In such a distribution, if all samples are the same size *n*, then the mean of the sample means will equal the population mean.

$$\mu_{\bar{X}} = \mu$$

Standard error of the mean is the standard deviation of the distribution of sample means. As sample sizes increase, the standard error of the mean decreases. When the standard error of the mean is small, it indicates that the distribution of sample means has less error estimating the true population mean.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Central limit theorem states as sample sizes increase, the distributions grow to more closely approximate the normal curve. When *n* is sufficiently large, then the distribution of sample means approaches a normal distribution. This is important for statistical inference, because when the sample is sufficiently large, it is assumed the distribution is normal.

***z* test** is a statistical test used to compare the mean of a particular sample to the mean of a population. It can be used to determine how likely it is that a sample drawn from the population.

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

Lecture and Demonstration Ideas

Sampling distributions are often the most difficult concept for students to grasp. Since sampling distributions and the central limit theorem provide the basis for making statistical inferences about a population from a sample, a review of the concept map (Transparency 10-1 and 10-2) may help students visualize the conceptual relationship. This review can lead into the M & M demonstration.

M & Ms and the Mean of the Means. This demonstration is designed to involve students and helps them understand the sampling distribution of the mean and standard error of the mean uses M&M candies. Bring a supply of M&M milk chocolate candy or M&M peanut candy to class. Usually two of the large economy-size bags of M&M's milk chocolate candies or four of the large economy-size bags of M&M's peanut candies are plenty for a class size of 30 to 35 students.

- Pass the bags around the class and have each student take a sample of 10 M&M's from the bag.
- Have students code the M & M's candy as follows: Brown = 1 point, Yellow = 2 points, Red = 3 points, Green = 4 points, Blue = 5 points, and Orange = 6 points.
- Ask students to compute the mean of the coded M&Ms

- Write each of the means on the board.
- Compute the mean of the means and the standard deviation of the means.

Using data from the M&M website, this population data are as follows:

M&M milk chocolate candies: $\mu = 2.8$ points and $\sigma = 1.67$ points

M&M milk chocolate with peanut candies: $\mu = 3.2$ points and $\sigma = 1.67$ points.

First, compare the mean of the means to the population mean. Then, use the population standard deviation to compute the standard error of the mean. Compare the standard error of the mean computed from the population value to the standard deviation of the mean computed from the class samples. Often, the mean of the means and the population mean are fairly close. Likewise, the standard error of the mean is often close to the standard deviation of the means from the class samples.

Random Sampling.

Extend this demonstration to discuss representativeness of a sample. Since, the most predominant color of M&Ms milk chocolate candies is brown and the least prominent are orange, green, and blue, ask students how the results might differ if only 3 M&Ms were sampled? Would they be more or less likely to select blue, green, and orange M&Ms? Would the standard deviation be larger or smaller? How does this relate to central limit theorem? Finally, after the class demonstration and discussion everyone in the class, including professors gets to eat the M&Ms.

Instructional Video. *Against All Odds: Inside Statistics.* Program Fifteen, "What is Probability?" has many examples of probability, including dice, car accidents, assessing risk, etc. The final 10 minutes uses a stimulation of New York traffic to demonstration the addition rule of probability. You may want to show the entire 30-minute video. These videos are produced by the Consortium for Mathematics and Its Applications and Chad-Angier (1989) and available through Annenberg/CPB.

Active-Learning Activities

Sleep. According to statistics from the National Sleep Foundation (2001), Americans are sleeping less and working more hours. It seems Americans tend to add more responsibilities and activities to their daily lives by reducing sleep time. Use the table of sleep statistics, extracted from the results of the NSF study, (Transparency 10-1) to demonstrate and discuss probability. Note that the responses to various reasons people reported that they would "try" to sleep longer are not mutually exclusive.

Random Sampling . The National Sleep Foundation commissioned WB & A Research to conduct the *2001 Sleep in America* poll. To illustrate random sampling, you may want to mention the technique used by WB & A. The report indicated that a random list of telephone numbers was used. In addition, quotas were established by region based on the 2000 U.S. Census household data. Interviewers telephoned an equal number of men and women (over 18 years) between 5:00 p.m. and 9:00 p.m. on weekdays and 10:00 a.m. to 4:00 p.m. on weekends. Since, the text discusses representativeness of the sample, this is a good time to mention that the use of phone numbers excluded individuals without phones. Thus, the random sampling method used may not be as representative as it could be if individuals with and without phones were given an opportunity to respond. Most likely, cost factors were involved in this sampling decision.

***z* test.** Collect data from the class on the number of hours students sleep workdays and weekends. Separate data by gender and use the *2001 Sleep in America* data (listed below) as population sleep statistics to determine if the sample data from the class came from the population.

Sleep in America data:

Adult women: Mean hours on workdays = 7.1 hours, standard deviation = 1.4 hours; mean hours on weekends = 7.7 hours, standard deviation = 1.5 hours. Adult men: Mean hours on workdays = 6.9 hours, standard deviation 1.5; mean hours on weekends = 7.8 hours, standard deviation = 1.3.

Additional Assignments

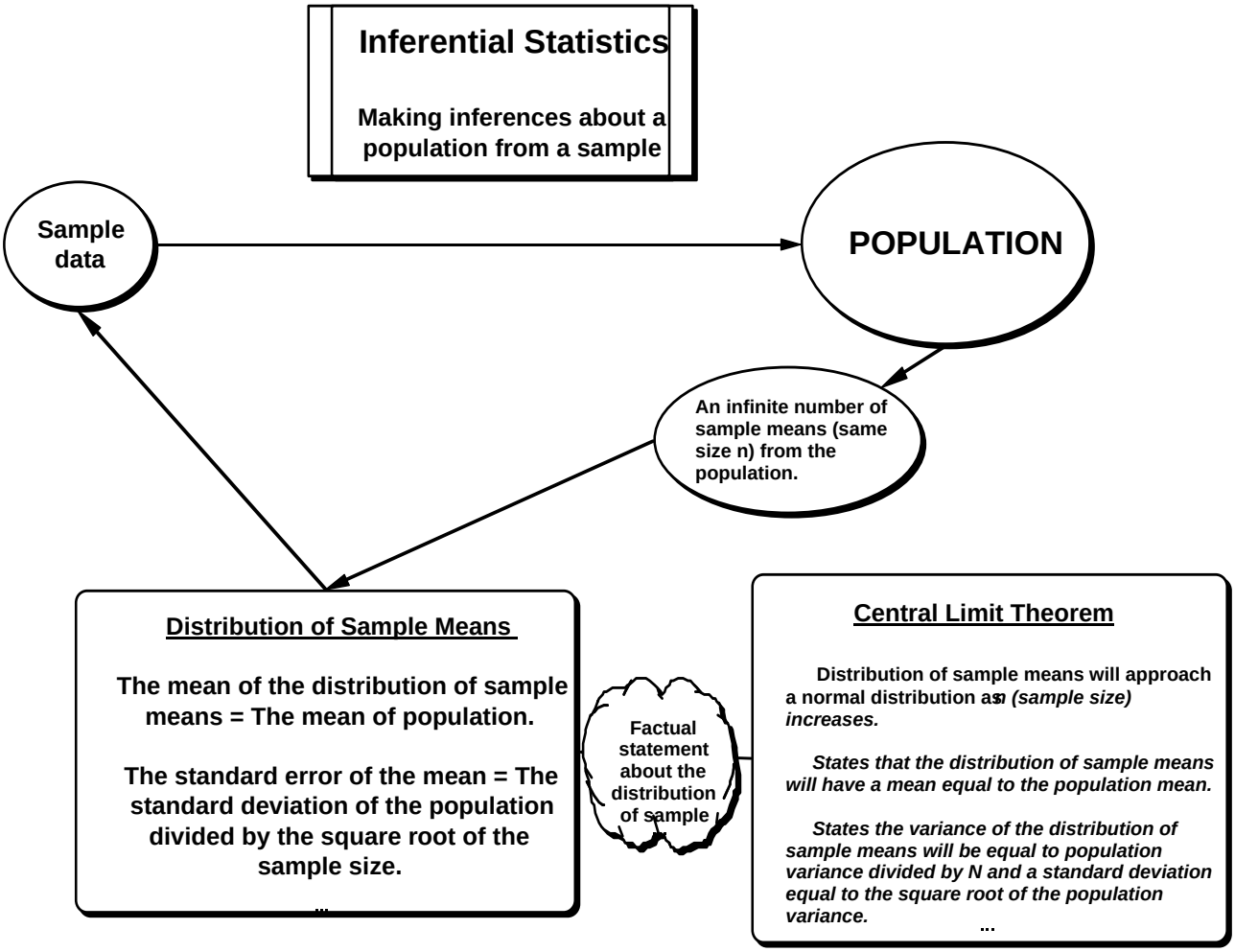
Assignment 1. Straight Dope discusses the topic of career field and suicide rate. The myth that the profession of dentistry has the highest number of suicides is explored from a probability standpoint. Source statistics, sample size, etc. are discussed. Have students visit this website, read the statistical argument against this myth and write a reaction paper about their experience.
http://www.straightdope.com/columns/010420.html

References:

M&Ms Home Page (2001). *Frequently asked questions*. http://www.m-ms.com/us/index.jsp

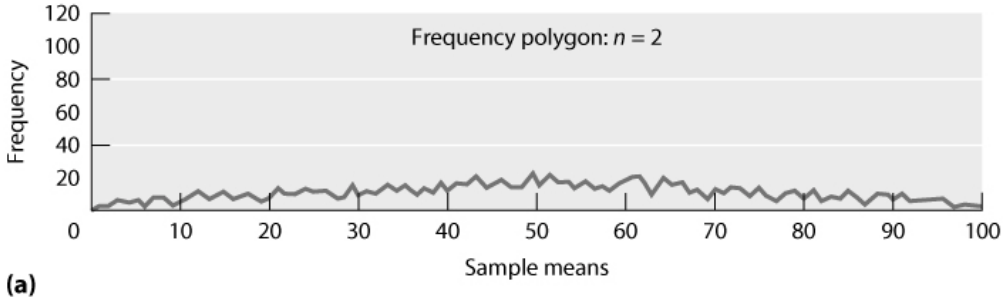
National Sleep Foundation (2001). *2001 sleep in American poll*. http:// www.sleepfoundation.org

Transparency 10-1.

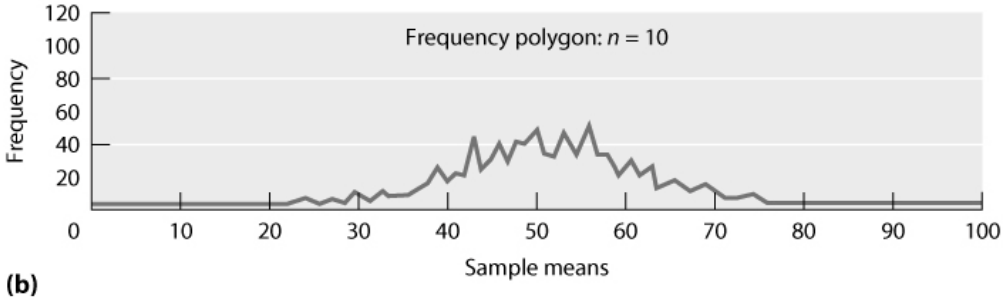


Transparency 10-2.

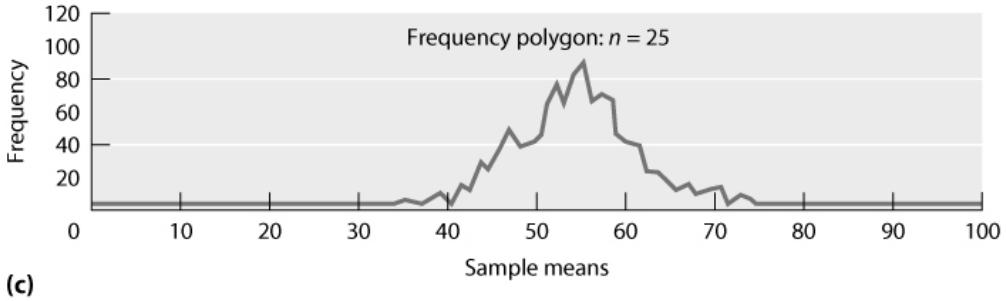
Central Limit Theorem



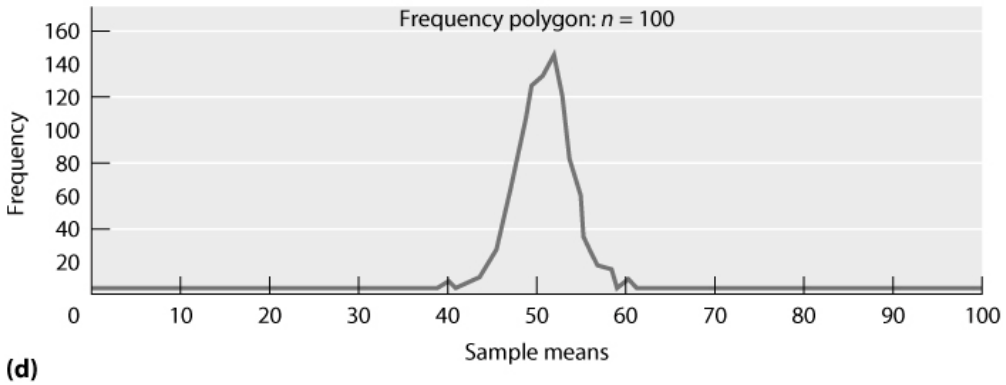
(a)



(b)



(c)



(d)

Transparency 10-3.

National Sleep Foundation – Sleep in America Poll (2001) Random Sample of 1004 (Quotas by Region, Random List of Telephone Numbers)				
Gender:			**Driven While Drowsy:**	
Men	500		Yes	532
Women	504		No	472
Hours Slept on Weekday:			**Dozed Off While Driving:**	
8 hours or more	372		Yes	192
8 hours or less	632		No	812
Hours Slept on Weekend:			**Time Spent Sleeping Compare to 5 years ago:**	
8 hours or more	607		More	186
8 hours or less	397		Same	428
			Less	390
Would Try To Sleep More Hours If....			**Days Per Week Need Alarm Clock to Wake Up**	
They could look better	713		0 days	361
It would improve sex drive	703		1 to 3 days	160
They could lose weight	643		4 to 5 days	261
Could slow the aging process	783		6 to 7 days	222