Introduction

Welcome to Advanced Statistics course. This is a major course for the Doctor of Philosophy in Nursing program aimed. The course reviews important concepts in statistics and presents their applications in nursing situations. This course equips you with the necessary skills and knowledge on data processing, data analysis, and data presentation. The focus is on advanced statistics that doctoral students would need in preparation for their dissertation.

The course is divided into three units. Unit 1, Review of Basic Statistics, is about descriptive statistics – how to organize and present data, such that they can be summarized and interpreted more meaningfully and it includes basic principles in inferential statistics. Unit 2, Common Inferential Statistical Methods, showcase statistical applications common in nursing, which includes tests of differences, and tests of relationships. Unit 3, Multivariate Analysis, is about statistical techniques that are useful in modeling techniques especially when testing or validating theories or models, which are required in the PhD in nursing program.

Course outcomes

At the end of the course, students should be able to:

1. Explain basic concepts and principles in statistics;
2. Demonstrate application of common inferential statistics used in nursing;
3. Perform advanced statistical methods and analysis given a data set; and
4. Critique data analysis and interpretations of select statistical applications in nursing studies.

Content outline

I. Review of Basic Statistics
   1. Basic statistical concepts
   2. Frequency tables
   3. Measures of central tendency and variability
   4. Probability and sampling
   5. Estimation
   6. Testing of hypothesis

II. Common Inferential Statistical Methods
   1. Test of independence
   2. Analysis of variance
   3. Correlation
   4. Regression

III. Multivariate Analysis
   1. Multivariate analysis of variance
   2. Multiple regression
   3. Logistic regression
   4. Factor analysis
   5. Path analysis
   6. Structural equation modeling

UNIT 1. Review of Basic Statistics

Organizing Principle

Descriptive statistics are used to describe or characterize data by organizing and summarizing them into more understandable terms without losing or distorting much of the information. It is different from inferential statistics, which consist of a set of statistical techniques that provide predictions about population characteristics based on information in a sample from that population.

In this Unit, the basic statistical concepts in descriptive and inferential statistics are going to be discussed. Tools in descriptive statistics will also be reviewed, namely: frequency distribution, measures of central tendency and variability, as well as normal distribution. The concept of probability is also applied in sampling, estimation and testing of hypothesis.

Learning Objectives

After working on this module, you should be able to:

1. Differentiate descriptive statistics from inferential statistics;
2. Use frequency distribution in describing data;
3. Describe data using measures of central tendency and variability;
4. Discuss properties of normal distribution;
5. Discuss concept of probability and sampling; and
6. Apply concepts of estimation and hypothesis testing in nursing studies

Underlying Concepts

1. Basic Statistical Concepts
2. Frequency Distribution
3. Summary Measures: Central Tendency and Variability
4. Normal Distribution
5. Probability and Sampling
6. Estimation and Hypothesis Testing

Module 1. Basic Statistical Concepts

This first module provides an overview of descriptive statistics and introduces fundamental concepts in statistics as it applies to nursing. This is a basic foundation in statistics, which is just a brief review for doctoral students. Still, take time to read the concepts and perform the activities in this module to refresh our memory and skills in descriptive statistics

Learning Objectives

After working on this module, you should be able to:

1. Define statistics and its importance in nursing;
2. Differentiate between descriptive and inferential statistics;
3. Discuss the fundamental concepts in statistics; and
4. Demonstrate application of descriptive statistics in nursing.

1.1 Statistics

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing and interpreting numerical information. It is used in several different disciplines (both scientific and non-scientific) to make decisions and draw conclusions based on data (Singpurwalla, 2013).

Read "What are statistics" and "Importance of statistics" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Study Guide Question:

- How is statistics used in everyday life?
- How is statistics used in nursing and health care?
- In what ways can statistics be misused or misinterpreted? Cite examples from the above reading material or from your observations in nursing.

1.2 Descriptive vs. Inferential Statistics

Descriptive statistics is used to describe a data set using both numerical measures and graphical displays of data. On the other hand, inferential statistics is used to make estimates, decisions, predictions, or other generalizations about a larger set of data using a sample data derived from it.

Read more about descriptive and inferential statistics at Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Activity 1-1.

Fill in the table to distinguish between descriptive and inferential statistics

	Descriptive statistics	*Inferential statistics*
Purpose		
Examples		
Statistical tools used		

1.3 Fundamental Concepts in Statistics

There are several fundamental concepts in statistics that we should keep in mind.

1.3.1 Population and sample

Statistical methods are used to study, analyze and learn about a **population** (a well-defined collection of objects, people or events). But populations (or the entirety of the subject of interest) is usually very large, not completely accessible, and thus, too expensive to study all of its members. Therefore, a portion of the population referred to as a **sample** is typically randomly drawn in order to study it in its entirety to be able to make some sort of generalizations from the sample to the population.

1.3.2 Concept and variable

Studying a population involves collecting information or data about specific characteristics (or observations) of interest in the study based on a **concept** or construct. A characteristic that cannot be changed is a constant; while characteristic that can change across conditions is a **variable**. After a method of measurement is applied to the concept or construct, it is then referred to as a variable (measured characteristics that takes on different values).

Variables can be categorized into several types. Variables that can assume an infinite number of values in between are called **continuous variables**, such as age or blood pressure; while those with a finite or limited set of values (or no intermediate values possible) are called **discrete variables**, like the number of patients or nurses.

1.3.3 Quantitative and qualitative measurements

Quantitative measurements use a naturally occurring numerical scale to describe the size of a particular variable**. Qualitative** measurements involve classification of observation into categories.

1.3.4 Levels of measurement of variable

There are four types of measurement scales for variables: nominal, ordinal, interval and ratio. Before analyzing data, it is best to assign first the type of measurement scale for each of the variables as this will help in deciding the organization and analysis of data.

Read "Levels of Measurement" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Activity 1-2

Given the following variables, list ways how you can measure them and identify the level of measurement

Variable	*How variable is measured*	*Level of measurement*
1. Age	• Age in years • Age group (pediatrics, adults, older adults)	• Interval • Nominal
2. Sex		
3. Educational attainment		
4. Body temperature		
5. Blood pressure		
6. Pain		
7. Knowledge of statistics		
8. Attitude towards statistics		

1.3.5 Univariate, bivariate and multivariate analysis

Data is obtained from making observations on a single variable (univariate) or simultaneously on two (bivariate) or more variables (multivariate). In the same manner, they can be analyzed as single variable, two variables or more variables. Because of this, classification of statistical

analyses is sometimes referred to as univariate analysis, bivariate analysis and multivariate analysis.

As a summary of fundamental concepts in statistics, you may want to watch this video developed by Quinnipiac University Health Professions Biostatistics
https://www.youtube.com/watch?v=moFyqGw0_0g

This video explains the role of statistics within the general field of scientific inquiry, and introduces some of the vocabulary and notations that are necessary for the statistical methods that follow.

1.4 Application of descriptive statistics in nursing

Read the abstract of a study on "Nursing students' personal qualities: a descriptive study from http://www.ncbi.nlm.nih.gov/pubmed/24907895

Activity 1-3

Questions	*Answers*
1. What is the purpose of descriptive statistics in this article?	
2. How is descriptive statistics used? Cite these.	
3. Describe the population and sample	
4. Identify what are the concepts and what are the variables in the study.	
5. In the identified variables, determine what are their level of measurements	

Module 2. Frequency distribution

Data collected in their original form are raw data. They can be organized for examination in order to make sense and draw meaning from them. The first thing that we do when we have collected data is to try to organize them. Creating a frequency distribution is the best way to organize and summarize data and to get simple interpretations. Use of frequency distribution is useful not only for organizing data, but also for determining the components of frequency distribution, describing the shape of distribution and facilitating creation of graphical or tabular presentation of data.

Learning Objectives

After working on this module, you should be able to:

1. Organize a given set of data using the ungrouped or grouped frequency distributions
2. Examine ungrouped and grouped frequency distributions for data interpretation
3. Determine the components of frequency distribution
4. Describe shapes of distribution
5. Create graphical or tabular presentation of data from frequency distribution

2.1 Organizing discrete and continuous variables

Frequency distribution organizes data into a table using categories for the data in one column and the frequencies (the count of the number of occurrences of a value) for each category in the second column.

Categories of data (variable)	*Count (frequency)*
• Single	
• Married	
• Widowed	

Frequency distribution can be classified according to type of data. For qualitative data, the **categorical frequency distribution** lists all the possible categories or attributes for the data in the first column and the count of the number of data in each category in the second column.

For quantitative data with discrete variables such as sex, marital status, ethnicity, religion, etc. where the number of categories can be enumerated or listed and the range of values in the dataset is small and the sample size (n) is large, **ungrouped frequency distribution** is used displaying all numerical values obtained for a particular variable (Burns and Grove, 2013).

In examining continuous variables, **grouped frequency distribution** is used where the range of values in the dataset is large. In which case, frequencies are displayed for ranges of data rather than for individual values (e.g. age, weight, blood pressure, scores in the final exams, etc.).

Read more about "Distribution" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and clarify the types of frequency distribution that can be produced from discrete and continuous variables.

Activity 2.1

Given the following variables related to distance education students, identify the type of data and classify whether you would use grouped or ungrouped frequency distribution in organizing them.

Variable	*Type of data*	*Grouped or ungrouped frequency distribution*
1. Age		
2. Sex		
3. Marital status		
4. Residence (address)		
5. Final grade in statistics		

There are advantages and disadvantages in using ungrouped frequency distribution and grouped frequency distribution. Ungrouped frequency distribution gives complete information about the data. However, if there are too many data points or wide range of values, this would be difficult to understand and meaningless. On the other hand, grouped frequency distribution may give organized and more meaningful tables, but some information are lost due to the process of condensing or summarizing data into groups or intervals. Therefore, the knowledge of when to use and for which type of data should ungrouped and grouped frequency distribution be used is important.

2.2 Examining grouped and ungrouped frequency distributions

The frequency distribution provides several components that allow interpretation of data. These components are: (1) frequencies (the count or number of occurrences), (2) percentage associated with each score, (3) cumulative frequencies (the number of observations at or below a particular value in a data set, or simply the "running total" of frequencies), and (4) cumulative percentage (the percentage of cases scoring at or below each score). A percentage distribution indicates the percentage of subjects in a sample whose scores fall into a specific group and the number of scores in that group. This is useful for comparing data among different studies that have different sample sizes. A cumulative distribution is a type of percentage distribution in which the frequencies and percentages are summed as one moves from the top of the table to the bottom. This is useful when ascertaining the position of a particular data point or score in the dataset.

Watch this video developed by Quinnipiac University Health Professions Biostatistics https://www.youtube.com/watch?v=a5-H4tl0U_8. Follow how ungrouped and grouped frequency distributions are created. This also examines components of frequency distribution and discusses percentile and percentile rank.

Study Questions

1. What is the advantage of grouped frequency distribution over ungrouped frequency distribution (and vice versa)?
2. What is the basis for the categories and class intervals when creating grouped frequency distribution?
3. What is percentile and when is it used?
4. What is percentile rank and when is it used?

2.3 Shapes of frequency distribution

Frequency distribution may reveal different shapes of distribution. Refer back to "Distribution" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and discover the different shapes of frequency distribution.

Study Questions

1. What is a positively skewed distribution?
2. What is a negatively skewed distribution?
3. What is a bimodal distribution?
4. What is a leptokurtic distribution?
5. What is a platykurtic distribution?

2.4 Presenting frequency data in tables and graphs

Going back to the video developed by Quinnipiac University Health Professions Biostatistics https://www.youtube.com/watch?v=a5-H4tl0U_8, examine the different graphs that can be derived from frequency distribution.

Study Guide Questions

- What is a bar graph? What component of frequency distribution do you use for the bar graph?
- What is a pie chart? What component of frequency distribution do you use for the pie chart?
- What is a histogram? How do you create a histogram?
- What is a frequency polygon? How do you create frequency polygon?
-

Activity 2.2

Create frequency tables and graphs using the data below

Students' scores in a 100-item long exam in statistics

81	94	90	80	87	80	85	95
83	92	87	70	96	76	87	89
86	79	75	83	84	75	81	81
81	84	70	78	96	94	88	78
80	77	93	87	77	78	79	72

Module 3. Measures of central tendency and variability

Central tendency is a concept that has to do with the location of the center of a distribution. The measures of central tendency give the most concise statement of the nature of the data in a study (Burns and Grove, 2013). The three most common measures of central tendency used in statistical analyses are mean, median, and mode. This module will present the characteristics, computational procedures and application of the different measures of central tendency.

Measures of variability refer to how much the numbers or values in the distribution differ from each other. While measures of central tendency describes the central location of data, measures of variability show the spread of the data or how close they are to each other. This complement and gives a better picture of the data. This module describes the different ways that we can measure and variability and how they should be reported together with the measures of central tendency.

Learning Objectives

After working on this module, you should be able to:

1. Define central tendency
2. Define the three basic measures of central tendency (mean, median, and mode)
3. Calculate the mean, median, and mode of both ungrouped and grouped data.
4. Determine which is the best estimator (mean, median, and mode) given a set of data
5. Describe measures of variability
6. Compute for measures of variability (range, interquartile range, variance, standard deviation)
7. Interpret results from measures of central tendency and variability

3.1 Definition of central tendency

Central tendency of a group of scores is an important concept in statistics to describe and summarize data. Knowing the central tendency of a set of scores allow us to imagine how that relates to another set of scores. Central tendency has been defined as: (1) a balance scale, (2) smallest absolute deviation, and (3) smallest squared deviation.

Read more about these formal definitions of central tendency from "What is Central Tendency" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Study Questions

1. What is a useful application of measures of central tendency?
2. Which of the three formal definitions of central tendency is easier to understand?
3. Which of the three formal definitions of central tendency is more difficult to understand?

3.2 Characteristics of mean, median and mode

The most common measures of central tendency are the mean, median and mode. Read about these from "Measures of Central Tendency" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Study Questions

1. What is the most commonly used measure of central tendency? Why?
2. What is otherwise known as 50^{th} percentile? Why?
3. How is mode derived from continuous data?

3.3 Computing for mean, median and mode

The computations of the measures of central tendency: mean, median and mode are also shown in "Measures of Central Tendency" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

When working with nominal variables, the only measure of central tendency that can be used is the mode. The mode of continuous data is normally computed from a grouped frequency distribution.

Activity 3.2

Using the formula given in the above article, compute for the mean, median and mode the previous dataset on students' scores in a statistics long exam. After doing the manual computation, you may verify your answers by using statistical tools and software.

Students' scores in a 100-item long exam in statistics

81	94	90	80	87	80	85	95
83	92	87	70	96	76	87	89
86	79	75	83	84	75	81	81
81	84	70	78	96	94	88	78
80	77	93	87	77	78	79	72

Mean ➔

Median ➔

Mode ➔

3.4 Choosing the best estimator of central tendency

For symmetric distributions, the mean, median and mode are equal. Differences among the measures occur with skewed distributions. And mode is not equal to mean and median in bimodal distributions.

Read "Comparing Measures of Central Tendency" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Study Guide Questions
1. What is the best estimator of central tendency when there are extreme values in the data set?
2. What happens to the mean when there is a positive skew?
3. What measures of central tendency should be reported when there is a large skew in the dataset?

Optional resource

Watch this video if you want another overview of "Measures of Central Tendency" from Qunnipiac University: Health Professions Biostatistics. Retrieved from https://www.youtube.com/watch?v=ciGfHZVgNtg

3.5 Variability

Read "Variability" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ This resource illustrates that two sets of scores may have the same measure of central tendency but they could have varied distribution of scores between them. This is the reason why reporting measures of central tendency should always be accompanied by a measure of variability.

Activity 3.3

Look for a journal article (on nursing studies) where measures of central tendency are reported with measures of variability. Given this study, describe the data using the two measures together. Continue reading "Variability" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and discover the four frequently used measures of variability: range, interquartile range, variance, and standard deviation.

Study Questions
1. What is the simplest measure of variability to calculate?
2. What is the usefulness of knowing the middle 50% scores in a distribution?
3. What is the relationship between variance and standard deviation?

3.6 Computing for variability

Study the computations of the different measures of variability: range, interquartile range, variance and standard deviation in the "Variability" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Activity 3.4
Given the formula for measures of variability in the above article, compute for the
(1) range, (2) interquartile range, (3) variance and (4) standard deviation of the following dataset.

Students' scores in a 100-item long exam in statistics

81	94	90	80	87	80	85	95
83	92	87	70	96	76	87	89
86	79	75	83	84	75	81	81
81	84	70	78	96	94	88	78
80	77	93	87	77	78	79	72

Measures of variability	*Computations*
Range	
Interquartile range	
Variance	
Standard deviation	

3.7 Interpreting results

Study the results of the computations of the different measures of variability: range, interquartile range, variance and standard deviation in the "Variability" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Study questions

1. What is the meaning of range? How is a low range score compared to high range score in terms of variability?
2. Which is more variable: IQR of 2 or IQR of 5?
3. What is the meaning of higher variance?
4. Why is the standard deviation useful when the distribution of the data is normal?

Optional resource

Watch this video from Quinnipiac University: Health Professions Biostatistics. (2013, July 5). Module 4 - Describing Data: Variability (Video file). Retrieved from https://www.youtube.com/watch?v=nguOqpR73Eg to learn techniques that describe the differences within a distribution of scores in order to study diversity and consistency within that distribution.

Module 4. Normal Distribution

The normal distribution, sometimes referred to as a normal curve, is a concept considered to be of great significance in statistics. Many statistical tests are based on the statistical assumptions in normal distribution. This module presents the properties and uses of the normal distribution and its role in data analysis.

Learning Objectives

After working on this module, you should be able to:

1. Identify the properties of normal distribution
2. Find areas under normal distributions
3. Describe standard normal distribution
4. Discuss applications of normal distribution in inferential statistics

4.1 Properties of normal distribution

Read "Introduction to Normal Distributions" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and answer the following study questions.

Study Questions

1. What determines the height of (or how tall is) the normal curve?
2. What determines the width of (or how flat is) the normal curve?
3. Why is the concept of normal distribution important?
4. What are examples of application of normal distribution in life?

To speak specifically of any normal distribution, two quantities have to be specified: the mean μ (pronounced "mu"), where the peak of the density occurs, and the standard deviation σ (pronounced sigma), which indicates the spread or girth of the bell curve. Different values of μ and σ yield different normal density curves and hence different normal distributions.

Activity 4.1

List some properties of normal distribution that you learned from the above article

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

4.2 Areas under normal distribution

The total area under the normal curve is equal to 1.0 or 100%. This can be interpreted as a probability value. The baseline of the normal curve is measured off in standard deviation units. These units are indicated by small letter z. A score that is one standard deviation above the mean (to its right) is +1 z and a score one standard deviation below the mean (to its left) is -1 z.

Read "Areas Under Normal Distributions" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and answer the following study questions.

Study Questions

1. What areas under the curve is approximately two standard deviations away from the mean?
2. In Figure 1 of the above article, the shaded values between 40 to 60 contains what % of the distribution?
3. In Figure 2 of the above article, what are the values that are two standard deviations away from the mean?

All normal density curves satisfy the following property which is often referred to as the Empirical Rule.

- 68%of the observations fall within 1 standard deviation of the mean, the probability expression for this is $P=(\mu - \sigma \leq x \leq \mu + \sigma) = 0.6826$
- 95%of the observations fall within 2 standard deviations of the mean, the probability expression for this is $P=(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.9544$
- 99.7 %of the observations fall within 3 standard deviations of the mean, the probability expression for this is $P=(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.9974$

4.3 Standard normal distribution

A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution. A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula: $Z = (X - \mu)/\sigma$

where, Z = the value on the standard normal distribution,
X = the value on the original distribution,
μ = the mean of the original distribution, and
σ = the standard deviation of the original distribution

Read "Standard Normal Distributions" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/ and answer the following study questions.

Study Questions

1. What is the meaning of the shaded area in Figure 1?
2. In Figure 1, what is the computed area below -2.5 on the standard normal distribution?
3. In Figure 2, scores are normally distributed with a mean of 50 and standard deviation of 10, what percent of scores is below 26? Above 26?

4.4 Application of normal distribution

Normal distribution is at the core of most of inferential statistics. Many statistical tests are based on the assumptions of normal distribution.

See how the Empirical Rule is applied when estimating probabilities such as: probability of a particular measurement and probability that a particular measurement is in a range.

Watch these videos from Khan Academy and work on Activity 1-3.

https://www.khanacademy.org/math/probability/statistics-inferential/normal_distribution/v/ck12-org-normal-distribution-problems-empirical-rule

https://www.khanacademy.org/math/probability/statistics-inferential/normal_distribution/v/ck12-org-normal-distribution-problems-z-score

https://www.khanacademy.org/math/probability/statistics-inferential/normal_distribution/v/ck12-org-exercise-standard-normal-distribution-and-the-empirical-rule

https://www.khanacademy.org/math/probability/statistics-inferential/normal_distribution/v/ck12-org-more-empirical-rule-and-z-score-practice

Activity 4-2

Based on what you have learned from previous module on normal distribution and the videos above, read the scenario below and answer the questions that follow.

Scenario

A study is recruiting women in a weight loss program. Weights of adult women in the city are normally distributed with a population mean of $\mu = 63.5$ kilograms and a population standard deviation of $\sigma = 2.5$

1. *What is the standard value or z-score for Alma whose weight is 67 kilograms? Interpret the result.*

2. *Alma's weight is higher than what percentage of the population? Interpret the result.*

3. *What fraction of the population had a weight between 60 and 65? Show computations step by step. Interpret the result.*

Module 5. Probability and Sampling Designs

Inferential statistics enables one to determine if the results derived from the sample can be generalized to the population from which it was drawn. It is then important for the sample to accurately represent the population. This module discusses the important principles in samples and sampling designs including factors that affect sample size determination.

Learning Objectives

After working on this module, you should be able to:

1. Discuss the principles behind choosing a sample;
2. Describe the different sampling techniques; and
3. Discuss the factors in sample size determination

5.1 Principles in choosing a sample

Inferential statistics are based on the assumptions that sampling is random. We trust a random sample to represent different segments of population close to the appropriate proportions.

Two important principles in choosing a sample are: (1) It should represent the population and (2) It should have adequate size. If the sample were not representative of the population or large enough, any inferences made would not be valid or reliable.

Read more about the principles from this link:
http://www.skillsyouneed.com/learn/sampling-sample-design.html

Study Guide Questions

1. How would you know if a sample is representative of a population?
2. How would you know if sample size is adequate?
3. What is the effect on the sample if it is not representative of the population?
4. What is the effect on the sample if the sample size is too small?

5.2 Different sampling techniques

There are two general types of sample techniques: (1) probability sampling and (2) non-probability sampling.

Read more about the sampling methods from this link:
http://www.skillsyouneed.com/learn/sampling-sample-design.html

Study Guide Questions

1. What is the advantage of probability sampling over non-probability sampling?
2. When is the use of non-probability sampling justified?
3. What are the considerations when choosing sampling methods?

Activity 5-1.

Differentiate between probability and non-probability sampling by filling in the table below.

	Probability sampling	*Non-probability sampling*
Purpose		
Examples		
Advantages		
Disadvantages		

Activity 5-2.

Choose which sampling designs are appropriate for the following research problems

Research objective	*Chosen sampling design*	*Justification*
1. Estimate the number of obese children in a community		
2. Describe the nutritional status of children under-5 years old in the Philippines		
3. List the top ten grocery items purchased by mothers in a supermarket		
4. Determine childrearing practices in indigenous populations		
5. Compare two treatments for breast cancer in a hospital		
6. Compare income of males and females in management positions		
7. Field trial of dengue vaccine		
8. Describe risk behaviors of patients who are HIV positive		
9. List top ten morbidity causes in a community after a natural disaster		
10. Survey of people living in condominium units about their lifestyle (diet and physical activity)		

5.3 Sample size

Sample size calculation is important in ensuring the study is valid, especially when making inferences about a population. There are several things to consider when determining sample size; among which are: (1) research design, sampling procedure, formula used for estimating optimum sample size, degree of precision required, heterogeneity of the attributes under investigation, relative frequency that the phenomenon of interest occurs in the population, and projected cost of using a particular sampling strategy.

Read the following article, which describes the principles and methods used to calculate sample size. After reading, answer the study guide questions below to check what you have learned from the article.

Kadam, P. & Bhalerao, S. (2010). Sample size calculation. *International Journal of Ayurveda Research*, Jan-Mar 1(1): 55-57. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2876926/

Study Guide Questions

1. List the factors that affect sample size.
2. What is the meaning of level of significance? What is the common level of significance in nursing studies?
3. What is the meaning of power of a study? What is the most common power in nursing studies?
4. What happens to the sample size if the effect size is small?
5. What happens to the sample size if the event rate in the population is low?
6. What happens to the sample size if the variable being measured in the population is more homogenous or has smaller variance?

Sample size calculation depends on the research design. Research designs such as surveys or cross-sectional studies require larger sample size compared to experimental studies.

For different study designs, read more about sample size calculation from:

Dean, A.G., Sullivan, K.M., and Soe, M.M. (2015). OpenEpi: Open source epidemiologic statistics for public health. Available at http://www.openepi.com/Menu/OE_Menu.htm

To use the above resource, click “sample size” from the left menu bar, and choose which sample size calculator you need depending on your research design. These research designs include:

- proportion or descriptive study
- unmatched case control study
- cross-sectional, cohort, and randomized clinical trials
- comparing two means

Activity 5-3.

Compute for the sample size for the following study conditions using Open Epi.

Study	*Sample size*
1. A simple random sampling is going to be conducted to study nurses practice of universal precautions in an Intensive Care Unit as part of patient safety. The population size of nurses in the hospital is 1,500. The expected result is unknown. The design effect is for random sample. The confidence limit is 5%.	
2. The same study above is to be conducted in all tertiary hospitals across the country using cluster sampling. The population size of nurses in all these hospitals is 25,000. The expected result based on previous study is 86% observe universal precautions. The design effect based on previous studies is 1.5. The confidence limit is 5%.	
3. A case-control study was done to see whether the nurse presence (measured in terms of number of visits to patient) increase rate of infection post-operatively. How many cases and controls are needed to detect an odds ratio of 2.0 or greater given 95% confidence, 80% power and 50 percent of controls exposed.	
4. In the same study above, a group of patients were followed up post-operatively. About 20% of patients 'exposed' to nurses and 10% of others develop post-operative infections. Using 95% level of significance, 80% power, 1 ratio of unexposed to exposed in the sample, compute for the number of exposed and unexposed in the population	
5. A study is to be conducted on the effect of “moringa tea” on weight loss among obese women. Given the weight of women taking “moringa tea” has a mean of 72 kgs and standard deviation of 5.5; and the control group has a mean of 77 kgs and standard deviation of 8.0, what would be the minimal sample size in each group to detect a difference with a power of 80% at 95% confidence level?	

Module 6. Estimation and Hypothesis Testing

Estimating population parameters from sample statistics is one of the processes in inferential statistics. It allows scientific guess on estimates of characteristics or properties of population based on the sample. This module describes the types of estimates, characteristics of good estimators and shows computations of confidence interval.

Hypothesis testing is an important process in inferential statistics. It is this process that allows us to make scientific guess about a research question. The hypothesis translates the research question into a prediction of expected outcomes. This module describes the concepts behind hypothesis testing, the steps in doing it, and how to interpret its results.

Learning Objectives

After working on this module, you should be able to:

1. Define point estimate and interval estimate;
2. Discuss characteristics of good estimators; and
3. Compute for confidence intervals.
4. Describe the logic of hypothesis testing
5. Define null hypothesis and alternative hypothesis
6. Distinguish between one-tailed and two-tailed tests
7. Describe the steps in hypothesis testing
8. Discuss Type I and Type II errors
9. Define statistically significant
10. Determine confidence interval whether a test is significant

6.1 Point and Interval Estimates

An important aspect of statistical inference is using estimates to approximate the value of an unknown population parameter based on information obtained from a sample. An estimator is a statistical parameter that provides an estimation of a population parameter.

An estimate of a population parameter may be expressed in two ways:

- A point estimate is a single numerical value or estimate of a population parameter.
- An interval estimate places the unknown population parameter between 2 limits, expressed as $a < \mu < b$. It assumes or considers the errors associated with the sampling procedure.

Read "Introduction to Estimation" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/estimation/introM.html

After reading/ watching the above resource, answer the questions that follow.
Study Guide Questions

1. What is the point estimate for the people's support to the proposition to build a new sports stadium?
2. If the margin of error in the polls is 10%, what is the interval estimate for the support to the proposal?
3. What is better, a point estimate or an interval estimate?

6.2 Characteristics of Good Estimators

Two important characteristics of estimates are bias and sampling variability. Good estimates are not biased and their sampling variability must be low.

Read "Characteristics of Estimators" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/estimation/characteristicsM.html

Study Guide Questions

1. What is an unbiased estimate?
2. What is the reason for the difference in the denominator of a population variance and sample variance?
3. What is standard error?
4. What is the effect of sample size on the standard error of the mean or sampling variability?

6.3 Computing for Confidence Intervals

Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time.

Read "Introduction to Confidence Intervals" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/estimation/confidenceM.html

Study Guide Questions

1. Why is a confidence interval better than a point estimate?
2. What is the meaning of 95% confidence interval?
3. Why is confidence interval not interpreted as the probability the interval contains the parameter?

Confidence interval can be computed for various parameters, such as mean, proportion, and Pearson's r. For this module, we will only look only into computing for the confidence interval for the mean since the other topics will be discussed later.

Read "Confidence Interval for the Mean" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/estimation/meanM.html

Study Guide Questions

1. What type of distribution should be used when the variance is known? when it is not known?
2. What is the formula for confidence interval when σ is known?
3. What is the formula for confidence interval when σ is estimated?

4. Where did 1.96 in the formula for 95% confidence interval using normal distribution come from?
5. How do you interpret the results of 95% confidence interval?
6. Why is a 99% confidence interval wider than a 95% confidence interval?

Activity 6-1.

Compute for confidence intervals of the following cases

1. IQ in a certain population is known to be normally distributed with a standard deviation of 5.

    a. Compute the 95% confidence interval on the mean based on the following sample of ten: 108, 109, 110, 113, 114, 116, 117,120, 121, 130
    b. Compute the 99% confidence interval using the same data.
    c. Interpret the results.

2. A sample of 30 from a population of test scores yields a sample mean of 70.

    a. If the standard deviation of the population is 10, what is the 99% confidence interval on the population mean?
    b. Assuming the population standard deviation is not known, but the standard deviation in your sample is 10, what is the 99% confidence interval on the mean now?
    c. Interpret the results.

6.4 Logic of hypothesis testing

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. "Hypothesis is a formal statement that presents the expected relationship between an independent and dependent variable" (Creswell, 1994). Hypothesis is usually structured by describing what will happen to the dependent variable if changes are made to the independent variable.

Why is it not possible to directly prove a hypothesis whether it is true or not?

Read "Introduction to Hypothesis Testing" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/introM.html

Study Guide Questions

1. Refer to the James Bond example in the article.
    a. What is the hypothesis in the James Bond example?
    b. What was the result of the hypothesis testing in the James Bond example?
    c. How did the researcher come to the conclusion that there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred?

2. Refer to the Physicians' Reactions in the article
    a. What is the hypothesis in the Physicians' Reactions in the article?
    b. What was the result of the hypothesis testing in the Physicians' Reactions example?

c. How did the researcher come to the conclusion that there is confidence that the difference in times is due to the patient's weight and is not due to chance?

3. Try to illustrate the logic of hypothesis testing with a diagram or drawing.

6.5 Null and alternative hypotheses

There are two types of statistical hypotheses: Null hypothesis (H_0) and Alternative hypothesis (H_1 or H_A).

The hypothesis that an apparent effect is due to chance is called the null hypothesis. The alternative hypothesis is the hypothesis that is accepted when the null hypothesis is rejected. It is also known as the research hypothesis. The research hypothesis could be about finding an estimate, looking for difference or relationships between variables.

Read "Introduction to Hypothesis Testing" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

Activity 6-1.

Write the null and alternative hypothesis for the following examples.

	Null hypothesis	*Alternative hypothesis*
1. Determining whether James Bond is able to tell whether his martini is shaken or stirred		
2. Comparing the time spent by physicians with patients who are obese compared with patients with normal weight		
3. Finding relationship between number of hours of studying and final grade in statistics		

6.6 One-tailed and two-tailed tests

Statistical tests that compute one-tailed probabilities are called **one-tailed tests**; those that compute two-tailed probabilities are called **two-tailed tests**. Two-tailed tests are much more common than one-tailed tests in scientific research because a result that signifies something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to

distinguish between no effect and an effect in the unexpected direction. The areas of rejection or **critical region** are also set by one-tailed or two-tailed tests. These critical regions are both tails of the distribution in two-tailed test, and either left or right tail in one-tailed test.

Read "One- and Two-tailed Tests" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/tailsM.html

Activity 6-2.

Write the hypothesis for the one-tailed and two-tailed tests of the following examples.

	One-tailed	*Two-tailed*
1. Determining whether James Bond is able to tell whether his martini is shaken or stirred		
2. Comparing the time spent by physicians with patients who are obese compared with patients with normal weight		
3. Finding relationship between number of hours of studying and final grade in statistics		

6.7 Steps in hypothesis testing

Hypothesis testing is the formal process to determine whether to reject the null hypothesis or accept it. There is a series of logical steps when doing hypothesis testing.

1. State the null and alternative hypotheses
2. Determine the level of significance
3. Identify the test statistic to be used
4. Determine the critical region
5. Compute the test statistic and probability value
6. Decide whether the result is significant or not
7. Make conclusions about the research hypothesis

Read "Steps in Hypothesis Testing" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/stepsM.html

Study Guide Questions

1. Why is the null hypothesis the important first step in hypothesis testing?
2. What is alpha or level of significance?
3. What is the probability value?
4. When is the result statistically significant?
5. Does failure to reject the null hypothesis mean the null hypothesis is true?

Activity 6-3.

Given the following scenario, do the hypothesis testing using the steps below.

> *A study compares the blood level of a particular vitamin among vegetarians and non-vegetarians. This vitamin is shown to be normally distributed in the population. The study showed that vegetarians do have more of the vitamin, but the difference is not significant. The probability value is 0.13*

Steps in Hypothesis Testing	*Application to scenario*
1. State the null and alternative hypotheses	
2. Determine the level of significance	
3. Identify the test statistic to be used	
4. Determine the critical region	
5. Compute the test statistic and the probability value	
6. Decide whether the result is significant or not	
7. Make conclusions about the research hypothesis	

6.8 Type I and Type II errors

There are two types of decision errors that can occur in hypothesis testing: Type I error and Type II error.

Type I error occurs when a significance test results in the rejection of a true null hypothesis. Type II error occurs when the null hypothesis is false and the significance test fail to reject it.

Read "Type I and Type II Errors" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/errorsM.html

Study Guide Questions

1. What is the relationship between alpha level (level of significance) and Type I error rate?
2. When is it impossible to make a Type I error?
3. Why is it erroneous to conclude that the null hypothesis is true when result of a statistical test is not significant?
4. What is beta?
5. What is power?

6.9 Statistically significant

Making decisions whether a result is statistically significant or not is based on the computed probability value or p-value. This value is compared against the level of significance. When a probability value is below the α level, the effect is statistically significant and the null hypothesis is rejected. If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted.

Read "Interpreting Significant Results" and "Interpreting Non-Significant Results" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the videos of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/significantM.html
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/nonsignificantM.html

Study Guide Questions

1. When is a result called statistically significant?
2. What is level of significance and probability value?
3. What is statistical significance and practical significance?
4. What is the difference between the two approaches to conducting significance tests: Fisher versus Neyman and Pearson?
5. Why is it not correct to accept the null hypothesis when you do not reject it?
6. What statistical analysis can demonstrate that an effect is most likely small?

6.10 Computing confidence intervals in significant testing

There is a close relationship between confidence interval and significance testing.

Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative).

If the 95% confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then the effect will not be significant at the 0.05 level.

Read "Significance Testing and Confidence Intervals" from Online Statistics Education: A Multimedia Course of Study from http://onlinestatbook.com/

You can also watch the video of this resource:
http://www.onlinestatbook.com/2/logic_of_hypothesis_testing/sign_confM.html

Study Guide Questions

1. How do you determine from a confidence interval whether a test is significant?
2. How does the confidence interval explain why non-significant results should not accept the null hypothesis