Outline of Module

- This module will introduce the following topics:
  - Data collection
    - Reliability and validity
    - Random vs. systematic error
  - Data classification
    - Categorical variables
    - Continuous variables
  - Descriptive statistics
    - Measures of central tendency (mean, median, mode)
    - Graphical distributions
    - Measures of variability (standard deviation and standard error)
  - Analytic statistics
    - Chi-square, Fisher’s exact test
    - T-test, Wilcoxon-Mann-Whitney, ANOVA
Objectives of Module

- Students who complete this module should be able to correctly define and apply the following terms as they pertain to evidence-based clinical practice:
  - Random vs. systematic error
  - Reliability, validity
  - Continuous vs. categorical variables
  - Mean, median, mode
  - Standard deviation, standard error

- Students who complete this module should be able to:
  - Distinguish between a continuous variable and a categorical variable.
  - Identify common methods by which continuous and categorical variables are described.
  - Identify common methods by which continuous and categorical variables are analyzed.
Role of Statistics in EBCP

- In previous modules, we have seen that data obtained from groups of patients or populations can provide clinicians with information about questions in the four basic *domains* of diagnosis, prognosis, risk/etiology, and treatment/prevention.

- We have also considered some basic study designs that are used to obtain this data and learned about the relative risk and odds ratios calculations of effect size, which are commonly seen in the literature.

- But how do we make sense of these statistics and translate them into clinical information?
Role of Statistics in EBCP

- Modules 4 and 5 will provide an overview of several basic principles of statistical analysis.
  - How data are collected, classified, described, and some basic approaches to statistical analysis (Module 4)
  - Basic principles of hypothesis testing (Module 5)

- The purpose of both of these modules is to encourage and equip students to become thoughtful users of medical research—we do not expect you to become statisticians or data analysts!
Key Terms / Vocabulary
Data Collection

- Good data requires good data collection.
  - The 2x2 tables presented in previous modules presented data (numbers) that had to have been collected.
  - The accuracy of any calculations we make using these tables depends on how well the numbers were measured.

- Data collection may be associated with random errors and systematic errors.
  - Random error is error that randomly varies from one measurement to another
  - Systematic error is error that does not randomly vary, but moves measurements systematically away from their true value

- Key concepts in data collection include reliability and validity.
  - These are described on the following slides.
Reliability and Validity

- **Reliability** in data collection is also known as precision.
  - It reflects the amount to which measurements are reproducible (if the measurement gives the same value every time).
  - It is impacted mainly by random error.

- **Validity** in data collection is also known as accuracy.
  - It reflects the extent to which measurements truly capture what is intended to be measured (if the measurements accurately reflect reality).
  - It is impacted mainly by systematic error.

These concepts are illustrated on the following slide using the analogy of a target.
Reliability and Validity

Reliable, not valid

Reliable and valid
Relating Errors and Reliability/Validity

1. Good reliability, poor validity.
2. Poor reliability, good validity (on average).
3. Good reliability, good validity.

Systematic Error
Random Error
Errors, Reliability, and Validity

The following graphic also conveys the relationship between error, reliability, and validity.

To orient yourself:
- The x-axis shows value of measurement
- The y-axis shows the number of times a measurement occurs
- M represents the true value that should be obtained

As shown in figure (a), random measurement error increases variability around the true measurement (and still obtains the true measurement).

On the other hand, systematic error as illustrated in figure (b) leads to systematic mis-estimation of the true measurement—also known as bias, a term you will come across in later modules.
Other Key Terms in Data Collection:

- **A variable** is any characteristic considered in a study.
  - A variable can take different values for different individuals within a group.
  - Example: A study interested in determining the relationship of smoking and lung cancer in a group of patients may use the following variables: lung cancer diagnosis (yes/no), pack-year history of smoking (#).

- **A distribution** is the set of values that a variable takes among a group of subjects.
  - We will return to this later. First, we will discuss the classification of variables.
Data Classification
Data Classification

Data come in different types. Two of the most common forms are:

- Categorical variables
- Continuous variables

The type of variable determines:

- The way it is described
- The way it is analyzed

We will study these first.

We will return to these later in the module.
Types of Variables - Categorical

- Categorical variables place individuals into one of several groups or categories. Examples include:
  - Sex (male or female)
  - Currently cigarette smoker (yes or no)
  - Cancer stage (1,2,3,4)
  - Infection (HIV+ or HIV-)
  - Risk for developing hypertension (low, moderate, high)

- Categorical variables may be further classified as:
  - **Ordinal variables**: Have a natural order (low, moderate, high)
  - **Nominal variables**: Have names and no order (Pittsburg, London, Melbourne)
  - **Dichotomous variables**: Have exactly two categories (yes/no or HIV+/HIV-)
Types of Variables - Continuous

- Continuous variables have numerical values for which arithmetic operations such as adding and averaging make sense. Examples include:
  - Height (in cm)
  - Weight (in kg)
  - Systolic blood pressure (in mmHg)
  - Hemoglobin level (in g/ml)

- Note, however, that just because a variable has numerical values does not mean that it is continuous. For example, a zip code is a categorical variable. (It would not make sense to add numbers from zip codes for any statistical purpose).
Test your knowledge

- Which terms in the table below describe categorical variables?

- Which terms describe continuous variables?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of Oral Contraceptive use (years)</td>
<td></td>
</tr>
<tr>
<td>Birth weight (extremely low, very low, low, normal)</td>
<td></td>
</tr>
<tr>
<td>Asthma Diagnosis (yes, no)</td>
<td></td>
</tr>
<tr>
<td>Duration between successive menstrual periods (days)</td>
<td></td>
</tr>
<tr>
<td>Smoking Status (former, current, never)</td>
<td></td>
</tr>
<tr>
<td>Pack Years of Smoking (avg. packs per day times years smoked)</td>
<td></td>
</tr>
</tbody>
</table>
Test your knowledge

- Which terms in the table below describe categorical variables?

- Which terms describe continuous variables?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of Oral Contraceptive use (years)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Birth weight (extremely low, very low, low, normal)</td>
<td>Categorical</td>
</tr>
<tr>
<td>Asthma Diagnosis (yes, no)</td>
<td>Categorical</td>
</tr>
<tr>
<td>Duration between successive menstrual periods (days)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Smoking Status (former, current, never)</td>
<td>Categorical</td>
</tr>
<tr>
<td>Pack Years of Smoking (avg. packs per day times years smoked)</td>
<td>Continuous</td>
</tr>
</tbody>
</table>
Clinical Relevance of Data Types

- The classification of data in a clinical study may affect your interpretation of the results. Three issues to consider are:
  - Is the variable clinically important? Are the researchers studying a variable that is relevant in terms of the overall disease process?
  - If a continuous variable was used to measure the effect of a treatment, what constitutes a clinically important change in that variable? For instance, does it matter if a drug changes the systolic blood pressure by only a few points, or is a larger difference needed to make a clinically significant difference?
  - If a categorical variable was used, do the categories make clinical sense? For instance, if a disease is classified as mild, moderate, or severe for a clinical study, are these categories clearly defined? Do these categories correspond with clinical knowledge about the disease?

- Classification of data is also important to understanding description and analysis of data.
Descriptive Statistics
How Data Are Described

- **Categorical** data are often described using proportions or percentages.
  - For example:
    - Percent of individuals with each blood type
    - Proportion of individuals with no exposure

- **Continuous** data are often described using measures of central tendency and measures of variability.
  - *Measures of central tendency* include mean, median, and mode
  - *Measures of variability* include standard deviation and standard error
Measures of Central Tendency

- **Mean**: Sum of observations divided by the total number of observations
  - e.g., average of 10, 13, 10, and 11 is \( \frac{10 + 13 + 10 + 11}{4} = 11 \)

- **Median**: Represents the 50th percentile
  - The point at which half the observations are above, and half are below
  - e.g., of 9 total observations, observation #5 is the median, when observations are ordered from smallest to largest
  - e.g., of 10 total observations, the average of #5 and #6 is the median when observations are ordered from smallest to largest.

- **Mode**: Most frequent observation
  - e.g., mode of 10, 13, 10, and 11 is 10.
  - A data set may have more than one mode.

Continuous variables can also be displayed graphically. Let’s look at some common examples.
Graphical Distributions

- We can plot the values (e.g., the readings for blood pressure or hemoglobin) for the variable of interest (x-axis) by the number of times each value is observed in a group of patients or a population (y-axis).
- The resulting graph gives the distribution of values for that variable. Distributions may take different forms or shapes.

*Note: Frequency indicates the number of times a value is observed.

- Left (Negative) Skewed
- Symmetric (not skewed)
- Right (Positive) Skewed
Normal Distributions

- A normal distribution is symmetric (“bell-shaped”). In this case, the mean, median, and mode are equal.

- The mean is a good measure of central tendency for normal distributions.
Skewed Distributions

- Distributions that are non-symmetric may be skewed to the left or right.
  - Right skew: Mean > Median > Mode
  - Left skew: Mean < Median < Mode
  - The median is a good measure of central tendency for skewed distributions.

Left (Negative) Skewed

Symmetric (not skewed)

Right (Positive) Skewed
Bimodal Distributions

- Similarly, distributions may have more than one peak.

- A distribution with two peaks is called a bimodal distribution.
  - The *modes* are good measures of central tendency for a distribution with two or more peaks.
Standard Deviation

- In addition to describing the central tendency of a distribution (with mean, median or mode), we should also describe variability within a distribution.

- **Standard deviation** (also SD or \( \sigma \)) is an indicator of the average distance that a single data value (e.g., a single patient’s value) lies from the average value for that group.

- For a *normal distribution* (with SD on x-axis)
  - Mean + 1 SD contains 68% of group
  - Mean + 2 SD contains 95% of group
  - Mean + 3 SD contains 99% of group

- If we know that our data are normally distributed (as in the graph), we can use the SD to determine how different an individual is from the mean.
  - For example: If the mean of the distribution is 5, the SD is 2, and our patient’s value is 12, we can determine that the patient’s value is greater than 3SD from the mean (5+3SD = 5+3*2 = 11). Therefore, we can say that this patient is farther from the mean than 99% of the group.

- We will return to this important concept in the next module.

*Note: Students are not expected to calculate standard deviations here. However, they should know percentages of individuals that fall within the standard deviations of a normal distribution.*
Standard Error

- A related term to standard deviation is standard error.

- Suppose that we collect data from a group of individuals for a clinical study.

- Values collected for each variable in the study will vary.
  - The indicator of the average variation of these values from the mean is the SD.

- We expect that the estimate of the average value of each variable will become more precise as more and more people are included in the study.
  - A larger number of participants will increase the reliability of our results.
    - Note, however, that it will not necessarily increase the validity of our results. It will not correct systematic errors.

- We can account for the reliability of our measurements by calculating standard error (s.e.).
  - Also known as the “standard error of the mean” or SEM
  - $\text{s.e.} = \frac{\text{SD}}{\sqrt{n}}$ where $n$ reflects the sample size (the number of people in the study)
Standard Deviation vs. Standard Error

- These two similarly named concepts are often confused. A few notes to help clarify why they are different (and why this is important):
  - Standard deviation is expected (and desirable) variation from the mean. We recognize, for example, that there is a natural variation in the height of medical students—we could quantify this expected variation as S.D.
  - Standard error, however, is as it sounds—an error. It affects estimation of a sample characteristic (e.g., the mean or median). If we wanted to estimate the mean height of medical students, we would want to reduce standard error to obtain the most accurate estimate of this value.

- To illustrate this difference, suppose that we have two studies:
  - Study 1: n=100 patients, mean=50, SD=15
    - s.e.=SD/sqrt(n)=15/10=1.5
  - Study 2: n=400 patients, mean=50, SD=15
    - s.e.=SD/sqrt(n)=15/20=0.75

- As this example shows, a larger number of patients in the second study may increase the reliability of the estimate (and decrease the standard error), even though the variability (the standard deviation) in both studies is the same.
Standard Deviation vs. Standard Error

- This figure illustrates how sample size affects standard error.
We’ve reviewed a number of concepts so far. Before we move on to statistical analysis, let’s review some important points:

- Variables may be classified as categorical or continuous.
  - Categorical variables may be described using proportions or percentages.
  - Continuous variables may be described using measures of central tendency (mean, median, or mode) and measures of variability (standard deviation, standard error).
- Graphical distributions of data may indicate the relationship of mean, median, and mode.
- A standard error may be calculated using the standard deviation along with the number of participants in the study. It reflects the reliability of a measurement.
Analytic Statistics
Data analysis is a complex process. There are many methods used in scientific literature and many are quite nuanced.

Clinicians will benefit neither from merely memorizing these details nor from ignoring them. There are three important things to consider when reviewing a statistical analysis in a clinical study.

- Statistical methods must match the type of data being analyzed (e.g., continuous vs. categorical).

- Many statistical methods are based on assumptions. You should consider these assumptions when making decisions about using the scientific literature.

- All statistics have context. That context should be clearly described in the methods section of a research paper.
Analysis Example

- Suppose that we have two groups in a clinical study.  
  - A treatment group that received a neuro-protective drug  
  - A control group that received a placebo

- Now, suppose we want to know whether the patients in the treatment group were at lower risk for stroke than those in the placebo group.
  - Is stroke a categorical or continuous variable in this situation?
  - How would we describe the data for stroke in this study?
  - How would we analyze the data to answer our question about whether the drug worked?
Analysis Example

- In this case, each patient will be classified as either having a stroke or not having a stroke.
  - This is a categorical variable. Since there are only two values possible in this case, we could also call this a dichotomous variable.

- To describe data for a categorical variable, we would use proportions or percentages.
  - We would present the percentage of patients with stroke in the treatment group and the percentage of patients with stroke in the placebo group.

- In order to determine whether there is a significant difference between these groups, we will need to use statistical analysis.
Analyzing Categorical Variables

Relative risks and odds ratios can be used to analyze certain types of categorical data (Module 3).
- RR may be used when incidence can be calculated.
- OR provides a measure of association that does not require incidence.

Another way to analyze categorical variables is to assess statistical differences in the proportions or percentages (e.g., percentage with stroke) across the groups we’re comparing. We can use:

- Chi-square test (or $\chi^2$ test)
- Fisher’s exact test
  - Similar to the chi-square test
  - Used when the number of people in an analysis is small
Analyzing Continuous Variables

Now, suppose that, in our study of the neuro-protective drug, we are also interested in whether systolic blood pressure was reduced in patients taking the drug.

Systolic blood pressure would represent a continuous variable in this case.

Thus, we could describe the blood pressure data using measures of central tendency and measures of variability.

How would we analyze the data to see if the drug actually reduced the blood pressure among the patients in the treatment group?
Analyzing Continuous Variables

- To analyze continuous variables, we will look for statistical differences in the averages (e.g., mean or median blood pressure) across the groups we’re comparing. We can use:

  - **T-tests**
    - If we are comparing **two groups** (as in this study)
    - If these groups have a **normal distribution** (compares their **means**)

  - **Wilcoxon-Mann-Whitney tests**
    - If we are comparing **two groups** (as in this study)
    - If these groups **do not have normal distributions** (compares their **medians**)

  - **ANOVA (Analysis of Variance)**
    - If we are comparing **three or more groups** (e.g., if we added a third group to the study so that we had groups for placebo, neuro-protective drug 1, and neuro-protective drug 2)
    - **Does not require normal distributions**
A Note on Multivariate Regression

- The tests mentioned so far are generally used to provide **bivariate** statistics—that is, to compare groups on a single variable.
  - e.g. Does the placebo or intervention group have a higher systolic blood pressure?

- In addition to these methods, you will often see **multivariate** statistics used in the medical literature. These methods can be used to compare groups on multiple variables simultaneously.
  - e.g. Does the placebo or intervention group have a higher systolic blood pressure, after accounting for differences in age and gender in these two groups?
  - As both increasing age and male gender are associated with differences in systolic blood pressure, it is important to account for them in the analysis.
    - This is also known as **controlling for** gender and age in the analysis.

- In future modules, you will learn more about how and why multivariate methods are used in the medical literature.
Categorical variables, including nominal, ordinal, and dichotomous variables, involve distinct groups, into which participants in a study may be classified.
- These may be described using proportions or percentages.
- These may be analyzed using chi-square tests or Fisher’s exact test, and relative risks or odds ratios, when appropriate.

Continuous variables involve numerical values which can be measured for participants in a study.
- These may be described using measures of central tendency (mean, median, and mode) and measures of variability (standard deviation and standard error)
- These may be analyzed using t-tests, Wilcoxon-Mann-Whitney tests, and ANOVA.
Please complete the online quiz.

Thank you!

*Module Design*: Timothy Bahr, Martha Carvour, Matthew Rysavy
*Revisions & updates* by T. Hegmann and E. Kiscaden