Here are the 10 selected questions, complete with their answers and full explanations from the document, presented in the recommended order for your lecture.

**Part 1: Describing Clinical Data (The Fundamentals)**

**Question 1 of 10 (Original Question #4)**

A physician has been collecting data on adolescent patients’ body mass indices (BMIs). The physician is now trying to determine whether BMI is associated with glycated hemoglobin (HbA1c) levels. It is hypothesized that there should be an association because higher BMIs are associated with type 2 diabetes mellitus and because people with uncontrolled diabetes mellitus have elevated HbA1c levels. The physician has collected data on 300 patients over the past 6 months.Of the following, the variable type that BEST describes BMI and HbA1c is:

A. continuous

B. discrete

C. nominal

D. ordinal

Correct Answer: A. continuous

**PREP Pearl(s):**

* Numerical scales include 1) Continuous scales, where the variables can be measured and have an infinite number of values, and 2) Discrete scales, which can be counted and have a finite number of values.
* Nominal scales are used when the data can be placed into categories that do not have a hierarchical relationship to each other, such as hair color or blood type.
* Ordinal scales are used when the categories are placed in an order that has meaning or hierarchy, such as Likert scale.

**Critique:**

Measurement scales are divided into 3 main groups: nominal, ordinal, and numerical. Nominal scales are used when the data fit into categories. An example of a variable that uses a nominal scale is blood type; ie, there are 4 possible blood type categories: A, B, AB, and O. The categories have names, but there is no hierarchical relationship between them. If there are only 2 categories, then the variable is said to be dichotomous or binary; eg, responses to a true–false statement.When a hierarchical relationship exists, the variable uses an ordinal scale. The categories can be placed in an order that has meaning. A Likert scale uses ordinal variables. A typical Likert item requires the study participant to rate their agreement with a statement. Possible responses are rated as follows: (1) strongly agree, (2) agree, (3) neither agree nor disagree, (4) disagree, and (5) strongly disagree. There is a directionality or order to these responses. An important characteristic of an ordinal scale is that the difference between categories is not the same across the scale. Using the Likert scale example, the difference between an “agree” and a “strongly agree” response is not inherently the same as the difference between a “neither agree nor disagree” and an “agree.”When the difference between values of a variable is consistent or equal, the variable uses a numerical scale. Numerical scales are further divided into continuous and discrete. Continuous scales have an infinite number of values, while discrete scales have a finite number of values. For example, if one is measuring body mass indices (BMIs) and glycated hemoglobin (HbA1c) levels in a population, theoretically one can carry the measure out to an infinite number of decimal places depending on the precision of the height and weight scale and the HbA1c instrument. In reality, BMI is usually only carried out to the 10th or 100th decimal place and the same with HbA1c, but these variables can be measured on a continuous scale. On the other hand, if one wanted to determine the number of cigarettes per day that individuals smoke, one would use a discrete scale of integers. If the range of cigarettes smoked by the population is between 0 and 60 per day, then one has a finite number of values.Knowing what type of scale the variable uses is important in order to make the right choice about the statistical test to use. As an example, if one wanted to test for a significant relationship or association between frequencies of 2 dichotomous variables, an appropriate statistical test would be the Chi-square test. The Chi-square can also be used for nominal variables with more than 2 groups; however, caution is needed because the test will not be valid when there are too many groups and an inadequately sized sample. For the example in the vignette, the Pearson Correlation would be appropriate to test for an association between the 2 continuous variables of BMI and HgbA1c.

**Question 2 of 10 (Original Question #5)**

A researcher is designing a clinical trial and developing the data collection forms. The researcher wants to collect information regarding household income, and, instead of leaving a blank area for the participant to report their income, the researcher provides the following choices: less than $50,000, $50,000 to $75,000, $75,001 to $100,000, and greater than $100,000.Of the following, the BEST way to describe this income variable is:

A. continuous variable

B. dichotomous variable

C. nominal variable

D. ordinal variable

**Correct Answer: D. ordinal variable**

PREP Pearl(s):

* Variables are either continuous or categorical.
* Categorical variables can further be described as dichotomous (2 distinct choices), nominal (>2 categories, no order/hierarchy), or ordinal (>2 categories with a specific order/hierarchy).

**Critique:**

In the vignette, there are 4 distinct categories of income, with the continuous variable category indicating the lowest income level and the ordinal variable category indicating the highest income level. This classifies this variable as an ordinal value.A continuous variable is a quantitative, numerical data point with an infinite number of possibilities along a continuum, such as height, weight, blood pressure, IQ, and yearly income. One family’s yearly income may be $80,125, another may be $80,126.50, and yet another may be $81,100.16.This is in contrast to a categorical variable, which is a qualitative, discrete variable with a limited number of possibilities. There are 3 main types of categorical variables: dichotomous, nominal, and ordinal.

* **Dichotomous variables** are variables or descriptive characteristics that can only be divided into 2 categories, for example, has disease/doesn’t have disease, yes/no, or male/female.
* Variables with more than 2 categories are either **nominal** or **ordinal** variables. Examples are eye color (green, blue, brown), race (per US Census Bureau: white, Black/African American, American Indian/Alaska Native, Asian, Native Hawaiian/other Pacific Islander), and Likert pain scales (no hurt, hurts little bit, hurts little more, hurts even more, hurts whole lot, hurts worst). When the categorical variable has an order or hierarchy to it (hurts worst > no hurt), this is considered an ordinal variable. When there is no order/hierarchy (eye color, race), then the variable is a nominal variable.

Sometimes variables that could be continuous variables (eg, income level) are collected or reported as categorical variables (as described in the vignette: <$50,000, $50,000-$75,000, $75,001-$100,000, >$100,000) instead of being reported as a distinct/exact number (ie, $80,101). Another common example of this is with blood pressure (ordinal variables: normotensive, prehypertensive, stage I hypertension, stage II hypertension) instead of being reported as a continuous variable (135/88 mm Hg). This is occasionally done for ease of reporting or describing study populations or because of statistical considerations (calculating an odds ratio, investigating for interaction, etc).

**Question 3 of 10 (Original Question #1)**

A hospitalist is working on a quality improvement project related to bronchiolitis. A team member has proposed an intervention package that they believe will reduce length of stay by 25%. The hospitalist is gathering length-of-stay data from the past 3 bronchiolitis seasons to set the baseline metric for the project. The histogram and descriptive statistics are shown in Figure 1.Of the following, the MOST accurate assessment of the baseline data is that:

A. the large number of outliers precludes accurate estimation of length of stay

B. the mean is the most appropriate measure of length of stay

C. the median is the most appropriate measure of length of stay

D. there is a 95% certainty that the true mean is between 30 and 88 hours

**Correct Answer: C. the median is the most appropriate measure of length of stay**

PREP Pearl(s):

* Preferred measures of central tendency include the mean for normally distributed data and the median for skewed data.
* The mean is sensitive to change from outliers, while the median is resistant to change from outliers.
* Measures of dispersion include variance and standard deviation when the data are normally distributed and the interquartile range when data are skewed.

**Critique:**

When data are normally distributed, they can be described and characterized using any measure of central tendency, mean or median, because they are equal to each other. However, when the data are skewed, the mean and the median may differ significantly. The data in the vignette are skewed. Specifically, they are right skewed; the majority of points are clustered at lower values while higher values are more widely distributed. Because the baseline data are skewed, the median is the more appropriate measure of the length of stay.Core knowledge in basic statistics is essential for all pediatricians. When data are summarized, it is critical to apply the appropriate measures of central tendency and variation. There are 3 main measures of central tendency: mean, median, and mode:

* The **mode** is the result that occurs most frequently.
* The **mean** is the arithmetic average of all the data points.
* The **median** is the middle point of the data set; half of the data points are greater than the median, while half of the data points are below the median.

When data are normally distributed, the mean, median, and mode are all equal; therefore, it does not matter which method of central tendency is used to describe the data set. However, when the mode < median < mean, the distribution is right or positively skewed. Conversely, the distribution is left skewed when the mean < median < mode. When skewed data are encountered, the mean is not the preferred measure of central tendency because the value of the mean is greatly influenced by outliers. In this situation, the median is a better measure of central tendency because it is resistant to the presence of outliers. This is shown in Figure 1, when the mean and median are compared. There are several very large outliers (several children who have greater than 500-hour-long hospital stays) in this sample. These large outliers inflate the value of the mean. However, because the median is affected by the total number of observations and not by outlying observations, it is a more accurate representation of the typical length of stay in a skewed sample.

**Question 4 of 10 (Original Question #2)**

Data obtained on elements of the complete blood count in 2- to 24-month-old infants and children with viral infections are being compared with data in those with confirmed bacterial infection.Of the following, and assuming continuous variables and a normal distribution, the measure of central tendency that yields the BEST descriptive information for evaluating the components of the complete blood count in this study is:

A. mean

B. median

C. mode

D. range

Correct Answer: A. mean

**PREP Pearl(s):**

* Mean, median, and mode are collectively called measures of central tendency and are numerical representations of descriptive statistics.
* Mean is the calculated average of all reported numerical data and is most useful in reporting a set of normally distributed continuous data points (interval and ratio observations).
* Median is the midpoint of a reported set of data (average of the 2 middle data points if there is an even number of observations) and is best used to report ordinal data.
* Mode is the most commonly reported numerical observation in a set of data and is best used to describe nominal data.

**Critique:**

Mean, median, and mode are numerical representations of descriptive statistics and are collectively called measures of central tendency. They are used to summarize and characterize a collection of study observations and are often presented in tables, graphs, or histograms. In this example, the mean of the data set is the most important description of central tendency. Although median and mode are also measures of central tendency and could be reported for this data set, they are better measures of central tendency for other types of outcome data. Range is a measure of the dispersion of data.Mean (or average) is the central value of a discrete set of numbers and is derived by summing all the numerical observations and dividing by the number of individual observations. Mean is most useful in describing central tendency for interval and ratio data, which are also called continuous data.

**Question 5 of 10 (Original Question #3)**

An 11-year-old boy is experiencing unexplained weight loss. He has vague symptoms of decreased appetite, and his parents have seen a gradual decrease in oral intake. When he was seen a year ago for a health supervision visit, he was growing along the 30th percentile for weight and 50th percentile for height. At his visit this year, he has lost 3.63 kg (8.00 lb). His current height and weight are shown in Figure 1.Of the following, assuming weights are distributed normally, this boy’s weight falls:

A. 1 to 2 standard deviations from the mean

B. 2 to 3 standard deviations from the mean

C. greater than 3 standard deviations from the mean

D. less than 1 standard deviation from the mean

Correct Answer: A. 1 to 2 standard deviations from the mean

**PREP Pearl(s):**

* Ninety-five percent of normally distributed data points fall within 2 standard deviations of the mean, which is equivalent to the 3rd to 97th percentiles.
* Z scores measure how far away a plot is from the population mean.

**Critique:**

In this vignette, the boy’s weight-for-age has fallen to about the 5th percentile, which is between 1 standard deviation (SD) (2.5th percentile) below the average weight for boys his age. SD is a measure of how far the data are spread out from the mean, or average, of the group.When data are normally distributed and follow a bell-shaped curve, empiric rules exist about the distribution of data around the mean.

* Sixty-eight percent of the data points fall within **1 SD** of the mean. On a growth curve, this would represent the 16th to 84th percentiles.
* Ninety-five percent of all data points are distributed within **2 SD** of the mean, which are about the 3rd and 97th percentiles.
* Three SD represents **99.7%** of all data points distributed around the mean.

Z scores for height, weight, and body mass index also use SD to compare an individual’s growth to that of the general population. For the child in the vignette, the 5th percentile is equivalent to 1.65 SD below the mean (−1.65). This SD “score” is the Z score, which is a numeric representation of how far above or below the mean a data point lies. A Z score is derived by subtracting the individual’s value from the population mean and dividing by the SD of the entire reference population. An individual at the 50th percentile has a Z score of zero. Z scores of ±2.0 are essentially equivalent to 2 SD above and below the mean, or again, the 3rd and 97th percentiles. Therefore, Z scores greater than 2.0 or less than −2.0 are outside the “normal” cutoff values.

**Part 2: Foundational Epidemiology & Diagnostic Testing**

**Question 6 of 10 (Original Question #8)**

A screening program conducted at a high school is screening student athletes with electrocardiography to identify undiagnosed heart disease. Of 1,000 students, 2 students are found to have a prolonged QT interval.Of the following, the parameter of long QT syndrome that can BEST be calculated by these data is the:

A. incidence

B. odds ratio

C. prevalence

D. relative risk

Correct Answer: C. prevalence

**PREP Pearl(s):**

* Prevalence is the number of affected cases in a specific population at a given point in time.
* Incidence is the number of new cases that develop over a specific period.
* Pretest probability is the probability of disease before a diagnostic test is ordered; posttest probability is the likelihood of disease after the result of a diagnostic test is known.

**Critique:**

The data that are given include the number of affected cases in a specific population at a given point in time—in other words, the **prevalence**.

**Incidence** is the number of new cases that develop over a specific period.

The **relative risk** is an estimate of the extent of the association between an exposure to a risk factor and a disease. It can be thought of as the likelihood of developing a certain disease after exposure to a certain risk factor, or as the incidence of disease in the exposed group divided by the incidence of disease in the unexposed group.

The **odds ratio**, on the other hand, is the odds of exposure in those with the disease compared with the odds of exposure in those without the disease.

**Question 7 of 10 (Original Question #9)**

A recent article described the epidemiology of celiac disease in an inner city community. The study evaluated a city with 50,000 children without celiac disease at the start of the study. At the end of the 5-year period, the following number of children were diagnosed with celiac disease.

* Year 1: 5 New Celiac Cases
* Year 2: 5 New Celiac Cases
* Year 3: 5 New Celiac Cases
* Year 4: 10 New Celiac Cases
* Year 5: 10 New Celiac Cases

Of the following, the annual INCIDENCE rate of celiac disease per 100,000 patients in this study is:

A. 5

B. 7

C. 14

D. 35

Correct Answer: C. 14

**PREP Pearl(s):**

* Incidence measures new occurrences of a disease in a population over a specified period and indicates the risk of contracting a disease. It is generally seen in cohort studies.
* Prevalence represents the magnitude of a disease, is widespread, and is seen in cross-sectional studies.

Critique:

Disease frequency is measured using prevalence and incidence. **Incidence** is the number of new cases in a given period divided by the size of the population initially at risk for the disease. Incidence represents the risk probability of diseases in a population within a set period. In the vignette, if a population initially contains 50,000 at-risk individuals with 35 people who develop celiac disease over a 5-year period while they were under the risk, the incidence is 7 in 50,000 or 14 per 100,000.

* Total new cases = 5 + 5 + 5 + 10 + 10 = 35 cases over 5 years.
* Average annual cases = 35 / 5 = 7 new cases per year.
* Annual incidence rate = (7 new cases per year / 50,000 population) = 0.00014
* Rate per 100,000 = 0.00014 \* 100,000 = 14 per 100,000.

**Prevalence** is a proportion representing the number of existing cases of a disease, both new and old cases, in a population over a specified time. Prevalence measures how widespread a disease is in a population.

**Question 8 of 10 (Original Question #6)**

An infant is admitted to the intensive care unit with septic shock. There is concern regarding the development of acute kidney injury and uncertainty about the ability of a new blood test to identify acute kidney injury compared to the current methods. A recent study presented the following information: among 200 children admitted with septic shock, 100 developed acute kidney injury. Of those who developed acute kidney injury, the test was positive in 70 children. However, the test was also positive in 10 children who did not develop acute kidney injury.Of the following, based on these results, the SENSITIVITY of the new blood test for detecting acute kidney injury is:

A. 10%

B. 30%

C. 70%

D. 90%

**Correct Answer: C. 70%**

**PREP Pearl(s):**

* The sensitivity of a test is the number of positive results among those with the condition. A highly sensitive test is good for ruling out disease when negative.
* The specificity of a test is the number of negative results among those without the condition. A highly specific test is good for ruling in disease when positive.

**Critique:**

Sensitivity and specificity are common measures of a test’s accuracy. **Sensitivity** describes the ability of a test to identify patients with the condition of interest (ie, true positives), and is expressed as the percentage of patients with the condition who have a positive test result. **Specificity**, on the other hand, describes the ability to identify patients without the condition of interest (ie, true negatives), expressed as the percentages of patients without the condition who have a negative test result.To determine the sensitivity and specificity of a test, it is often helpful to create a 2x2 table.

* Total patients = 200
* Have acute kidney injury (AKI) = 100
* Do not have AKI = 100
* Test positive in those with AKI (True Positives) = 70
* Test positive in those without AKI (False Positives) = 10

The sensitivity of the test would be the number of true positives divided by the total number of children with the disease: 70 / 100 (ie, 70%).

The specificity would be the number of true negatives divided by the total number of children without the disease. The number of true negatives is 100 (without AKI) - 10 (false positives) = 90. So, specificity is 90 / 100 (ie, 90%).

**Question 9 of 10 (Original Question #7)**

A research article describes a potential new serum screening test for eosinophilic esophagitis (EoE). One hundred children are recruited in the study. Of these 100 children, esophageal biopsy findings were consistent with EoE in 25, and 75 children did not have histologic evidence of EoE. The screening test results in all subjects demonstrated the following:

* Of the 25 with EoE, the screening test was positive in 20 and negative in 5.
* Of the 75 without EoE, the screening test was negative in 65 and positive in 10.

Of the following, the sensitivity and specificity of the screening test are:

A. Sensitivity 20.0%, Specificity 75.0%

B. Sensitivity 80.0%, Specificity 86.7%

C. Sensitivity 80.0%, Specificity 25.0%

D. Sensitivity 20.0%, Specificity 25.0%

(Note: The original document has a typo in the provided options/answer for specificity. The calculation based on the text is 65/75 = 86.7%, not 75%. This is a good teaching point about critically reviewing source material.)

Correct Answer: B. Sensitivity 80.0%, Specificity 86.7%

**PREP Pearl(s):**

* Sensitivity is the ability of a test to correctly detect those patients with the disease, while specificity is the ability of a test to correctly identify those without the disease.
* Sensitivity and specificity values do not vary based on the prevalence of disease.

Critique:

When calculating sensitivity and specificity, a table may be helpful:

|  | **Presence of Disease (+)** | **Absence of Disease (-)** |
| --- | --- | --- |
| Test Positive (+) | 20 (True Positive) | 10 (False Positive) |
| Test Negative (-) | 5 (False Negative) | 65 (True Negative) |

The equation used to calculate **sensitivity** is: True positive / (True positive + False negative)

Therefore, the sensitivity of the blood test for EoE would be: 20 / (20 + 5) = **80%Specificity** can be calculated with the following equation: True negative / (True negative + False positive)

Therefore, the specificity of the blood test for EoE would be: 65 / (65 + 10) = **86.7%**

**Part 3: A Critical Capstone Concept**

**Question 10 of 10 (Original Question #26)**

The health editor of a parenting magazine inquires about a study that was recently published in a peer-reviewed journal. The study examined the relationship between toddler thumb-sucking and later food allergy.Of the following, based on the provided scatterplot (Figure 1), the MOST appropriate conclusion to draw from this study is that thumb-sucking:

A. is associated with a lower incidence of food allergy

B. is a result of food allergy

C. is unrelated to the incidence of food allergy

D. lowers the incidence of food allergy

Correct Answer: A. is associated with a lower incidence of food allergy

**PREP Pearl(s):**

* The terms “association” and “dependence” describe a relationship between 2 variables.
* Two variables may be associated with or without correlation. Correlation means a positive or negative trend between 2 variables.
* Correlation does not automatically imply causation.

Critique:

In particular, the concepts of association, correlation, and causation are often used interchangeably, when in fact they have very different meanings. Two values are dependent, or **associated**, if one value affects the other. However, association does not automatically imply **causality**, because there may be other confounding factors that affect the relationship between the 2 variables.Association and **correlation**, however, have different meanings. A scatterplot is useful for determining whether association and correlation are present. Association means that one value is affected by another or provides information about another. Correlation means that there is a positive or negative trend between the 2 variables.

* In a **positive correlation**, as one variable (x) increases, the other variable (y) also increases.
* In a **negative correlation**, as x increases, y decreases.

The scatter plot from the study described in the vignette shows an association with negative correlation. As the number of months of thumb-sucking increases, the number of food allergies decreases. Thus, the 2 variables are related, so it is inaccurate to conclude that thumb-sucking is not related to the incidence of food allergy. However, the demonstrated correlation does not imply causation, therefore "thumb-sucking lowering the incidence of food allergy" and "being the result of food allergy" are also incorrect. One might hypothesize about a mechanism by which the relationship occurs, but this study is not adequate to prove causation.