STATISTICAL INFERENCE AND HYPOTHESIS TESTING PART IV

Nasir Mushtaq, PhD, MBBS
Associate Professor
Department of Biostatistics and Epidemiology
Hudson College of Public Health

Department of Family and Community Medicine
OU-TU School of Community Medicine

In the fourth part of this series entitled Statistical Inference and Hypothesis Testing, we will discuss types of hypothesis testing methods and statistical analysis approaches.

Objectives

- Identify data analysis methods commonly used in the biomedical literature

After viewing this module, you will be able to identify data analysis methods that are commonly used in the biomedical literature.

Statistical Hypothesis Testing and Inferential Techniques

I will present an overview of different statistical hypothesis testing and inferential methods.

This is a very brief overview. Additional details can be reviewed in the course reference textbooks.

Choosing the correct method of analysis

- Is the study objective one of superiority, non-inferiority, or equivalence?
- Is the outcome categorical, quantitative, or a survival distribution?
- If the outcome is quantitative, is it normally distributed?
- How many study groups do you have?
- Are the study groups independent or dependent (matched or repeated)?

The choice of an appropriate method of analysis depends on your answer to several questions.

Is the study objective one of superiority, non-inferiority, or equivalence?

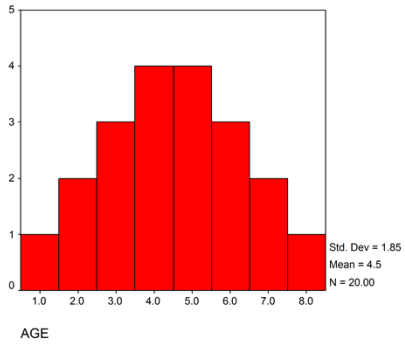
Is the outcome categorical, quantitative, or a survival distribution?

If the outcome is quantitative, is it normally distributed?

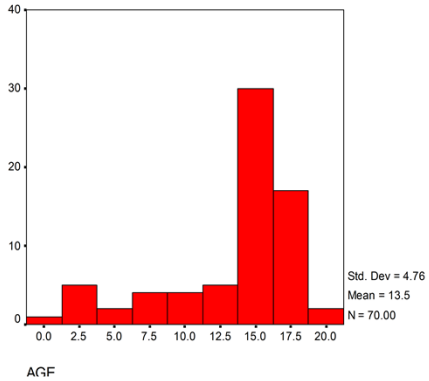
How many study groups do you have?

Are the study groups independent or dependent (matched or repeated)?

Assessing the distribution



Approximately normal distribution



Non-normal/skewed distribution

Many of the methods that we use are appropriate for normally-distributed variables. A normal distribution will have a bell-shaped curve.

When the distribution is not normal, for example, it may be positively or negatively skewed, we will use non-parametric analysis methods that do not require an assumption of normality.

The Normal Distribution

Characteristics of the normal distribution:

1. Symmetrical about its mean μ (mirror image)
2. Mean, median, and mode are all equal (bell shaped)
3. The area between -1 standard deviation and +1 standard deviation from the mean is approximately 68% of total area under the curve

   ± 2 standard deviation is about 95% of the total area under the curve

   ± 3 standard deviation is about 99.7% of the total area under the curve

There are several key characteristics of a normal distribution:

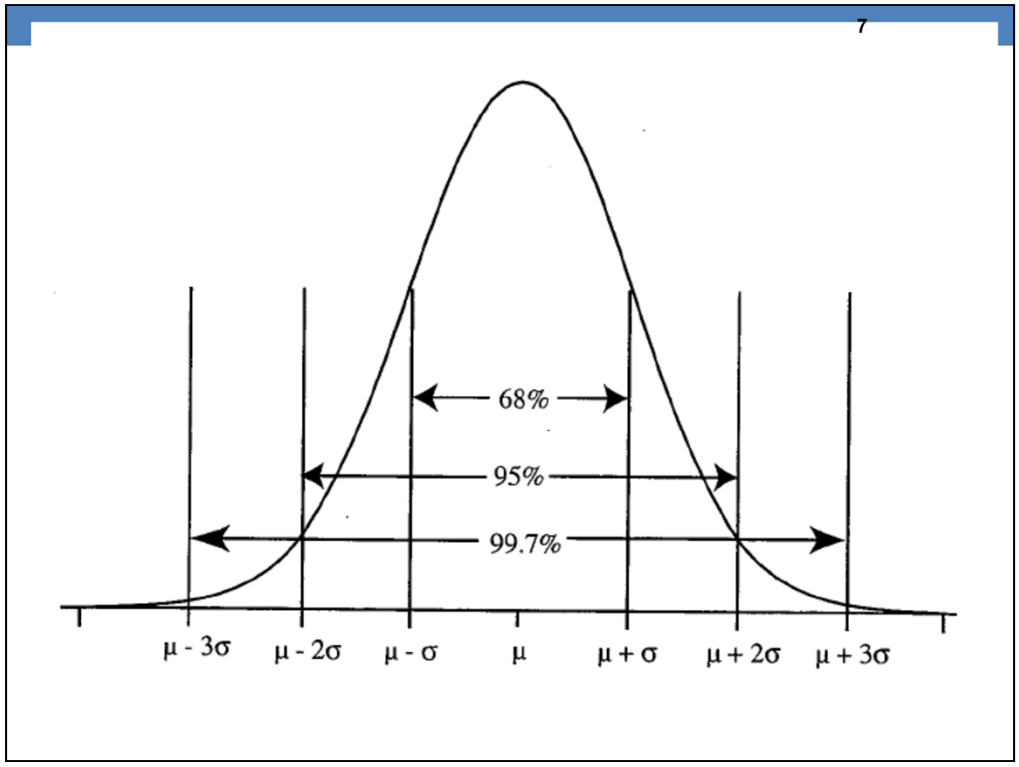
Symmetrical about its mean □ (mirror image)

Mean, median, and mode are all equal (bell shaped)

The area between -1 standard deviation and +1 standard deviation from the mean is approximately 68% of total area under the curve

± 2 standard deviation is about 95% of the total area under the curve

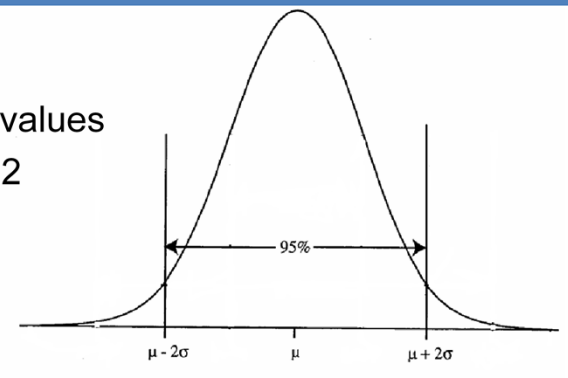
± 3 standard deviation is about 99.7% of the total area under the curve



This figure includes a summary of the area under the curve (or probability) corresponding to particular cut-points defined by +/- 1 standard deviation, +/- 2 standard deviations, and +/- 3 standard deviations. As we consider wider intervals, the area under the curve increases.

8

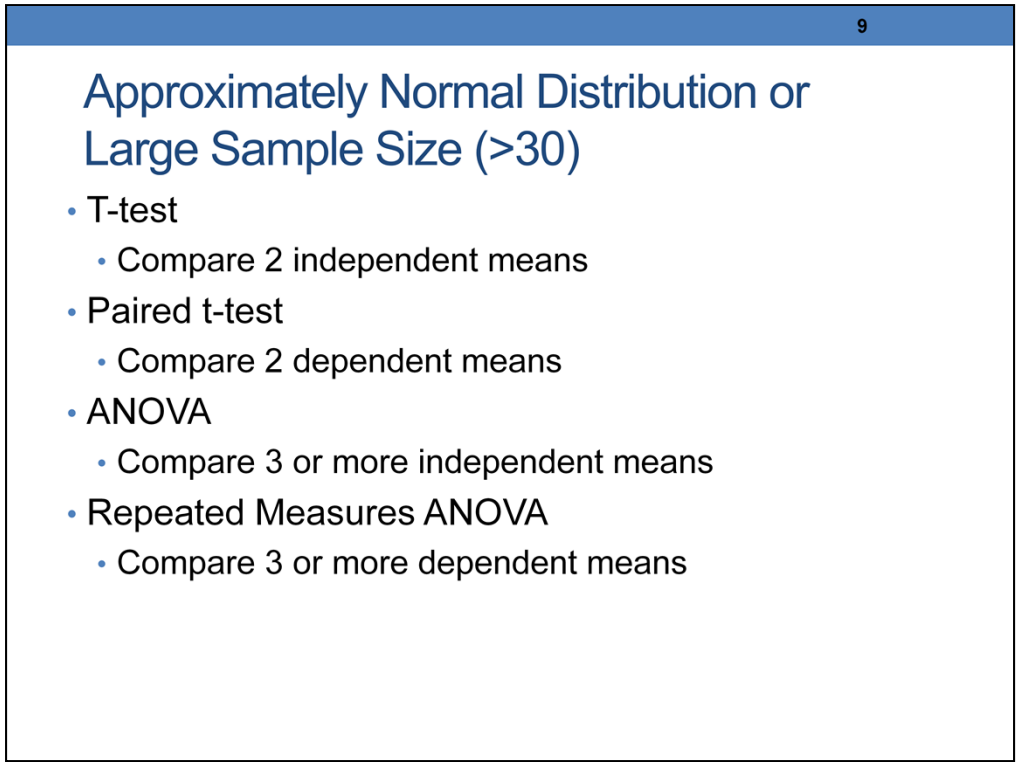
So, 95% of our data values are between -2 and 2 σ's from the mean



- Therefore, it's common practice that if we want to evaluate whether a value is unusually high or low, we usually see if it is within ± 2 standard deviations from the mean

When given a normal distribution, we expect that 95% of the observations will fall within +/- 2 standard deviations of the mean.

Therefore, it's common practice that if we want to evaluate whether a value is unusually high or low, we usually see if it is within ± 2 standard deviations from the mean (95% of the values are expected to be within this interval and 5% are expected to be outside this interval).



When we are analyzing a continuous outcome measure, such as cholesterol or blood pressure, and the data distribution is approximately normally distributed or the sample size is larger than 30, we will use the following tests:

T-test to compare means from two independent groups, such as a treatment and control group where group membership was randomly assigned.

A paired t-test to compare means from two paired or dependent groups, such as measures on the same subject before and after an intervention, twin studies, matched individuals in each group in the beginning of the study based on gender or age.

Analysis of variance (ANOVA) is used to compare the means among 3 or more independent groups.

Repeated measures ANOVA is used to compare the means among 3 or more groups that are paired or dependent such as repeated, longitudinal measures on a patient over time.

Approximately normal distribution

- Linear Regression
  - Describes the relationship between an explanatory variable (independent) and a continuous outcome variable (dependent)
  - Independent variable: categorical or continuous
  - How well does variable X predict variable Y?
  - Multiple linear regression – used to include multiple independent variables
- Correlation
  - Describes the strength of a linear relationship between two continuous variables
  - Pearson correlation coefficient (r)
    - $-1 \leq r \leq 1$

Linear regression can be used to describe the linear relation between an explanatory variable (independent) and a continuous outcome variable (dependent). The independent variable can be categorical or continuous while the dependent variable must be continuous.

We can use linear regression to determine how well does variable X predict variable Y?

Multiple linear regression can be used to include multiple independent variables when predicting an outcome Y.

Correlation can be used to describe the strength of a linear relationship between two continuous variables. The correlation coefficient (r) ranges from -1 to 1.

If r is between 0 and +1 then you have a positive slope, as x increases so does y

If r is between -1 and 0 then you have a negative slope, as x increases y decreases

If r = 1 or -1 then you have a perfect linear relationship

If r = 0 then you have no linear relationship

Nonparametric tests

- Nonparametric tests should typically be used if:
  - the variable (or a transformed version) does not have an approximately normal distribution
  - the distribution is unknown and cannot rely on large sample (>30) theory

Nonparmetric tests are used in situations where the variable (or a transformed version) does not have an approximately normal distribution and the sample size is small, for example, when considering a pain measures on 10 patients using an ordinal scale with values ranging from 1 to 5, the distribution would not be normally distributed.

Nonparametric tests are also used when the distribution is unknown and cannot rely on large sample (>30) theory.

Nonparametric tests

Parametric	Nonparametric
One-sample t-test	Sign Test (ordinal data)
Paired t-test	Signed-Rank Test
t-test: 2 independent samples	Mann-Whitney Test Wilcoxon Rank Sum Test
Pearson Correlation	Spearman Correlation
ANOVA	Kruskal-Wallis 1-way ANOVA

This table provides a listing of the non-parametric alternative to each parametric test that we discussed previously.

Chi-square test

- Used to compare proportions between two or more populations
  - If the groups are independent – a general chi-square is appropriate
  - If the groups are dependent – a McNemar chi-square is appropriate
  - If expected cell counts are too small – a Fisher's Exact test is appropriate

A Chi-square test is used to compare proportions between two or more populations or between two variables with two or more categories each. If the groups are independent – a general chi-square is appropriate. If the groups are dependent or paired – a McNemar chi-square is appropriate. If expected cross-tabulation counts are too small or are 0 – a Fisher's Exact test is appropriate.

Logistic Regression

- Used to predict dichotomous outcome from an explanatory (independent) variable
- Independent variable: categorical or continuous
- Modeling concept similar to linear regression
- Interpretation deals with log odds
- Multiple logistic regression used to include multiple independent variables

Similar to linear regression where we use independent variables to predict continuous outcome measures, we can use logistic regression to predict a dichotomous outcome from an explanatory (independent variable).

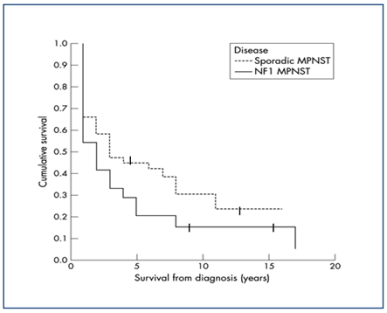
The independent variable can be either categorical or continuous while the outcome must be categorical.

Modeling concepts are similar to linear regression. The model estimates are interpreted in terms of the log odds of an event or the odds ratio of the event associated with a particular between-group comparison.

Multiple logistic regression is used to include multiple independent variables when modeling the outcome.

Survival Analysis: estimation

- Time to event data, censored data
- Kaplan-Meier curves
  - Graphs that illustrate the survivorship function for different groups

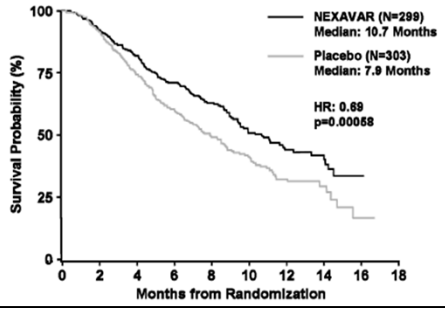


When analyzing time to event data, such as the time to death or time to treatment response, our method of analysis needs to account for the censoring, meaning, not all patients experience the endpoint by the end of the study and therefore, the time to event is censored for those individuals. Also, the time to event distributions are often positively skewed.

The Kaplan-Meier method is used to illustrate the survivorship function for groups of patients. The curves represent the estimated survival probability at a given time. Curves located in the lower left corer correspond to worse outcomes, meaning, events occur more rapidly over time.

Survival Analysis: comparisons

- Log-rank test
  - Nonparametric method for statistically comparing survival distributions
- Cox proportional hazards model: regression model for time-to-event outcome data; continuous or categorical independent factors



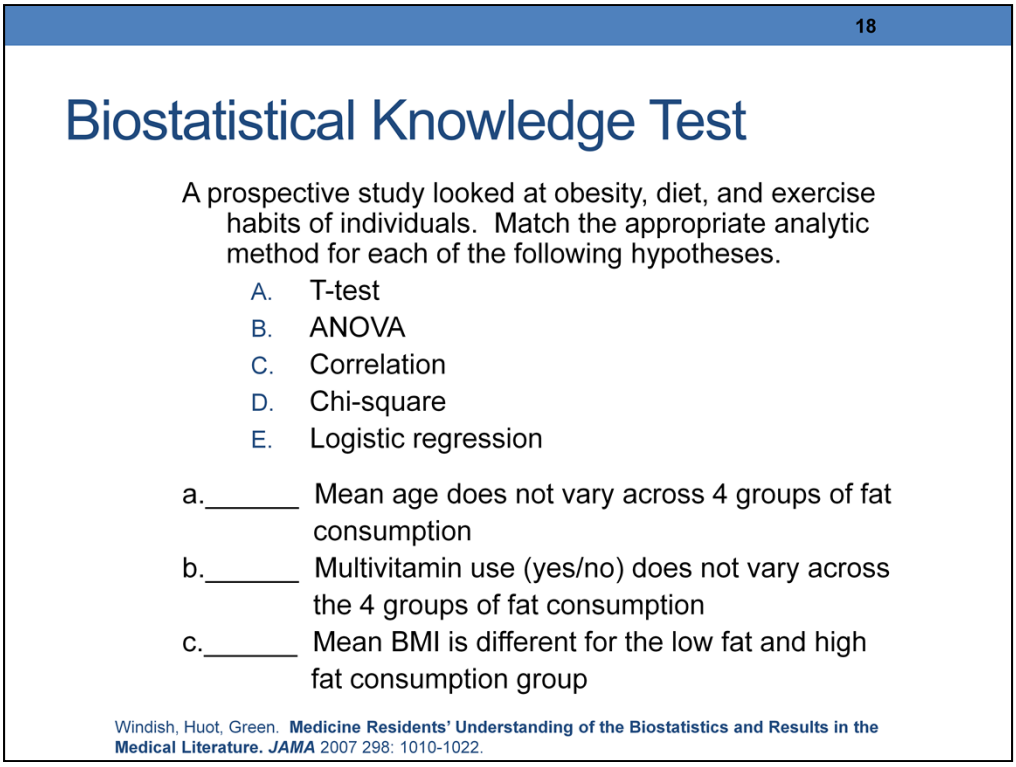
The distribution of survival times are compared between groups using a log-rank test.

Cox proportional hazards regression models can be used to quantify the association between multiple independent factors and the hazard or risk of an event.

Basic Analysis Techniques Summary

	Explanatory Variable	
Outcome Variable	Categorical	Continuous
Categorical	Chi-square, Logistic Regression	Logistic Regression
Continuous	ANOVA, t-test, Linear Regression	Linear Regression/ Correlation

This table provides a summary of analytic methods depending on the type of explanatory (independent) variable and the type of outcome variable. Analysis methods for time to event data are not shown in this table and methods for small sample sizes or non-normal distributions also are not shown.



Now, lets consider a series of questions to test your understanding.

A prospective study looked at obesity, diet, and exercise habits of individuals. Match the appropriate analytic method for each of the following hypotheses.

T-test

ANOVA

Correlation

Chi-square

Logistic regression

Mean age does not vary across 4 groups of fat consumption.

Biostatistical Knowledge Test

A prospective study looked at obesity, diet, and exercise habits of individuals. Match the appropriate analytic method for each of the following hypotheses.

A. T-test
B. ANOVA
C. Correlation
D. Chi-square
E. Logistic regression

a._B_____ Mean age does not vary across 4 groups of fat consumption

b._____ Multivitamin use (yes/no) does not vary across the 4 groups of fat consumption

c._____ Mean BMI is different for the low fat and high fat consumption group

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

Give that there are 4 groups among which we are comparing means, we will use B. ANOVA.

Biostatistical Knowledge Test

A prospective study looked at obesity, diet, and exercise habits of individuals. Match the appropriate analytic method for each of the following hypotheses.

A. T-test
B. ANOVA
C. Correlation
D. Chi-square
E. Logistic regression

a._B_____ Mean age does not vary across 4 groups of fat consumption

b._D_____ Multivitamin use (yes/no) does not vary across the 4 groups of fat consumption

c._____ Mean BMI is different for the low fat and high fat consumption group

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

Question of interest: Multivitamin use (yes/no) does not vary across the 4 groups of fat consumption

In this example, the outcome is categorical and the independent factor is also categorical, so we will use a Chi-square test. We could have also used logistic regression to address this question.

Biostatistical Knowledge Test

A prospective study looked at obesity, diet, and exercise habits of individuals. Match the appropriate analytic method for each of the following hypotheses.

A. T-test
B. ANOVA
C. Correlation
D. Chi-square
E. Logistic regression

a._B_____ Mean age does not vary across 4 groups of fat consumption

b._D_____ Multivitamin use (yes/no) does not vary across the 4 groups of fat consumption

c._A_____ Mean BMI is different for the low fat and high fat consumption group

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

Question of interest: Mean BMI is different for the low fat and high fat consumption group

In this example, we are comparing the mean between two independent groups and therefore, the appropriate method is a T-test.

Biostatistical Knowledge Test

In an aspirin vs. drug D study, the researchers wished to assess if there were any difference between groups with respect to the primary endpoint of time to restenosis while controlling for other potential risk factors. What analytic method would be most appropriate in assessing their questions?

a. Log-rank test
b. Logistic regression
c. Linear regression
d. Cox proportional hazard regression
e. Chi-square test

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

Let's consider another example.

In an aspirin vs. drug D study, the researchers wished to assess if there were any difference between groups with respect to the primary endpoint of time to restenosis while controlling for other potential risk factors. What analytic method would be most appropriate in assessing their questions?

Biostatistical Knowledge Test

In an aspirin vs. drug D study, the researchers wished to assess if there were any difference between groups with respect to the primary endpoint of time to restenosis while controlling for other potential risk factors. What analytic method would be most appropriate in assessing their questions?

a. Log-rank test
b. Logistic regression
c. Linear regression
d. Cox proportional hazard regression
e. Chi-square test

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

The appropriate method to understand the association between multiple risk factors and a time to event outcome is the Cox proportional hazards regression method.

Biostatistical Knowledge Test

In a research study, the mean age of the participants was 26 years ± 5 years (standard deviation), where age followed a normal distribution. Which of the following statements is the most correct?

a. It is 95% certain that the true mean lies within the interval of 16-36 years.
b. Most of the patients were aged 26 years; the remainder were aged between 21 and 31 years.
c. We would expect that approximately 95% of the patients were aged between 16 years and 36 years.
d. No patients were younger than 16 or older than age 36.

Windish, Huot, Green. **Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature.** ***JAMA*** 2007 298: 1010-1022.

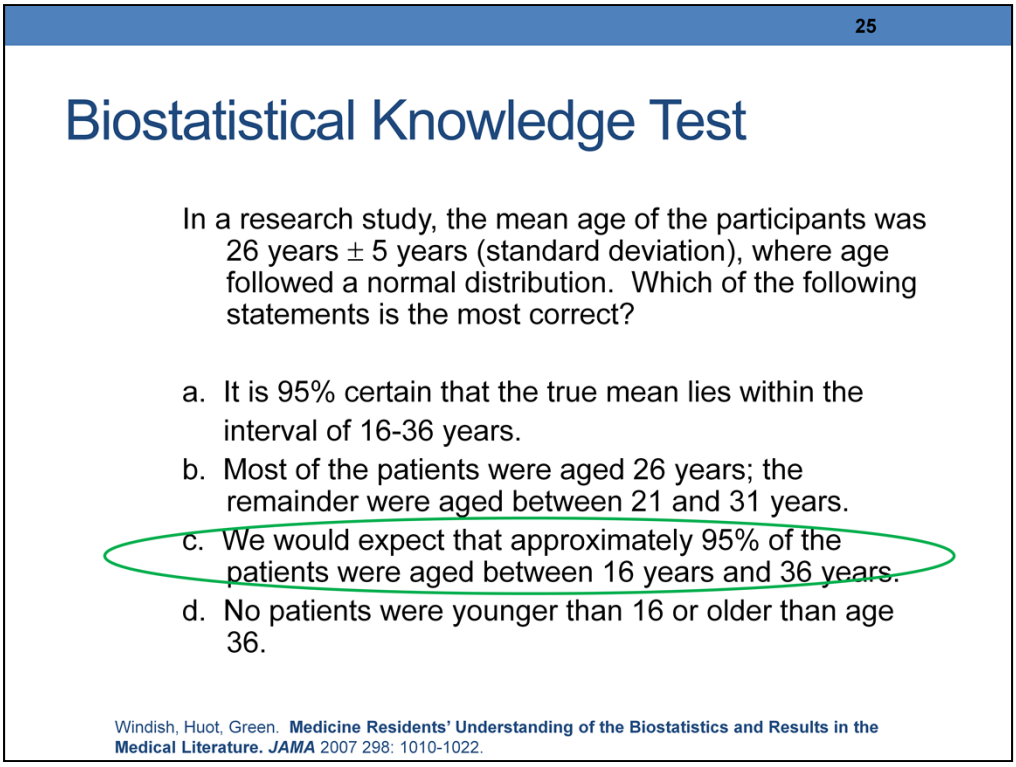
In another example, the mean age of the participants was 26 years □ 5 years (standard deviation), where age followed a normal distribution. Which of the following statements is the most correct?

a. It is 95% certain that the true mean lies within the interval of 16-36 years.

b. Most of the patients were aged 26 years; the remainder were aged between 21 and 31 years.

c. We would expect that approximately 95% of the patients were aged between 16 years and 36 years.

d. No patients were younger than 16 or older than age 36.



Based on a normal distribution, we expect 95% of the observations to fall within +/- 2 standard deviations of the mean for a normally distributed random variable.

In this case, the interval will be 26 +/- 2*5 = 16 to 36.

Based on a normal distribution, we would expect that approximately 95% of the patients were aged between 16 years and 36 years.

Summary

- Selecting the analytic method:
  - Is the study objective one of superiority, non-inferiority, or equivalence?
  - Is the outcome categorical, quantitative, or a survival distribution?
  - If the outcome is quantitative, is it normally distributed?
  - How many study groups do you have?
  - Are the study groups independent or dependent (matched or repeated)?

In summary, we have briefly reviewed common statistical methods of analysis. Our decision to choose one analysis method over another was drive by answers to the following questions:

Is the study objective one of superiority, non-inferiority, or equivalence?

Is the outcome categorical, quantitative, or a survival distribution?

If the outcome is quantitative, is it normally distributed?

How many study groups do you have?

Are the study groups independent or dependent (matched or repeated)?

This concludes the series focused on hypothesis testing and statistical inference.