AA501 Review

Institute for Advanced Analytics
MSA Class of 2024

Boxplots illustrate the relationship between what types of variables?

A. Categorical versus Categorical
B. Quantitative versus Quantitative
C. Categorical versus Quantitative
D. None of these

Boxplots illustrate the relationship between what types of variables?

A. Categorical versus Categorical
B. Quantitative versus Quantitative
C. Categorical versus Quantitative
D. None of these

When we say we estimate parameters, we …

A. Are trying to estimate the theoretical parameter from the entire population by using a sample
B. Are using the entire population.
C. Are rounding the parameters to some decimal place.
D. Are seeing if the null hypothesis is true.

When we say we estimate parameters, we …

A. Are trying to estimate the theoretical parameter from the entire population by using a sample
B. Are using the entire population.
C. Are rounding the parameters to some decimal place.
D. Are seeing if the null hypothesis is true.

Which of the following is *not* used to describe a variable's distribution?

A. Spread (Variance)
B. Center/Location
C. Kurtosis
D. Heteroscedasticity
E. Skewness
F. Outliers

Which of the following is *not* used to describe a variable's distribution?

A. Spread (Variance)
B. Center/Location
C. Kurtosis
D. Heteroscedasticity
E. Skewness
F. Outliers

Which of the following is *not* a characteristic of the normal distribution?

A. Symmetric
B. Skewed
C. Bell shaped
D. Mean = Median
E. Unimodal

Which of the following is *not* a characteristic of the normal distribution?

A. Symmetric
B. Skewed
C. Bell shaped
D. Mean = Median
E. Unimodal

Which plots can help you visualize if your residuals are normally distributed?

A. Histograms
B. Box Plots
C. QQ-Plots
D. Residual scatter plots
E. A and C
F. All of the above
G. Nothing helps me with this task.

Which plots can help you visualize if your residuals are normally distributed?

A. Histograms
B. Box Plots
C. QQ-Plots
D. Residual scatter plots
E. A and C
F. All of the above
G. Nothing helps me with this task.

A 95% confidence interval for the mean represents an interval …

A. …in which the probability of the true mean being inside the interval is 95%.

B. …in which of 100 intervals created using the same technique, we would expect approximately 95 of them to contain the true mean.

C. Both A and B

D. Neither A nor B

E. of time in which ego reigned

A 95% confidence interval for the mean represents an interval …

A. …in which the probability of the true mean being inside the interval is 95%.

B. …in which of 100 intervals created using the same technique, we would expect approximately 95 of them to contain the true mean.

C. Both A and B

D. Neither A nor B

E. of time in which ego reigned

To satisfy the assumption that the sample means are normally distributed…

A. Verify that the population distribution is approximately normal
B. Apply the central theorem by verifying that the sample size is ”large enough” (n=50 sufficient for α=0.05)
C. The sample means are always normally distributed
D. EITHER A or B
E. BOTH A and B
F. check the ‘almost totally normal’ button

To satisfy the assumption that the sample means are normally distributed…

A. Verify that the population distribution is approximately normal
B. Apply the central theorem by verifying that the sample size is ”large enough” (n=50 sufficient for α=0.05)
C. The sample means are always normally distributed
D. EITHER A or B
E. BOTH A and B
F. check the ‘almost totally normal’ button

α = P(Type I Error). A Type I error occurs when…

A. You reject a null hypothesis that is true.
B. You accept a null hypothesis that is false.
C. You fail to reject a null hypothesis that is false.
D. You calculate the test statistic using the wrong theoretical degrees of freedom
E. You say the F-test.

α = P(Type I Error). A Type I error occurs when…

A. You reject a null hypothesis that is true.
B. You accept a null hypothesis that is false.
C. You fail to reject a null hypothesis that is false.
D. You calculate the test statistic using the wrong theoretical degrees of freedom
E. You say the F-test.

As your sample size increases and all else remains the same…

A. Your p-value increases
B. Your p-value decreases
C. Your p-value is unaffected
D. Your p-value becomes a q-value

As your sample size increases and all else remains the same…

A. Your p-value increases
B. Your p-value decreases
C. Your p-value is unaffected
D. Your p-value becomes a q-value

When testing for a difference between two sample means from *normally distributed* populations, we find that the variances of the two populations are different…

A. We cannot use the t-test and instead have to use ANOVA
B. We can use the unequal variances t-test procedure
C. We cannot test for a difference in this situation because a difference would not have any meaning
D. I was told there would be snacks at this review.

When testing for a difference between two sample means from *normally distributed* populations, we find that the variances of the two populations are different…

A. We cannot use the t-test and instead have to use ANOVA
B. We can use the unequal variances t-test procedure
C. We cannot test for a difference in this situation because a difference would not have any meaning
D. I was told there would be snacks at this review.

I have a categorical variable with 3 levels. If I built a One-Way ANOVA model with this variable, how many unique predictions would I have?

A. 2
B. 3
C. 0
D. 1
E. Each prediction is unique in its own way

I have a categorical variable with 3 levels. If I built a One-Way ANOVA model with this variable, how many unique predictions would I have?

A. 2
B. 3
C. 0
D. 1
E. Each prediction is unique in its own way

The overall F-test in ANOVA tests…

A. If the model is specified correctly
B. If all of the levels of the categorical variables are different from each other
C. If any level of a categorical variable is different from the others
D. If F is for failure.

The overall F-test in ANOVA tests…

A. If the model is specified correctly
B. If all of the levels of the categorical variables are different from each other
C. If any level of a categorical variable is different from the others
D. If F is for failure.

The table below represents what type of coding for a categorical variable with 3 levels?

A. reference coding
B. effects coding
C. effective coding
D. -1 and 1 coding
E. python

	X_A	X_B
A	1	0
B	0	1
C	-1	-1

The table below represents what type of coding for a categorical variable with 3 levels?

A. reference coding
B. effects coding
C. effective coding
D. -1 and 1 coding
E. python

	X_A	X_B
A	1	0
B	0	1
C	-1	-1

Why might we adjust our p-values or α levels when we perform hypothesis tests?

A. If we don't like the result of a hypothesis test, we can always lower the confidence level, α
B. Because the assumption of normality was violated so a different p-value chart must be used.
C. To control the experimentwise error rate when performing many hypothesis tests at once
D. #statistics

Why might we adjust our p-values or α levels when we perform hypothesis tests?

A. If we don't like the result of a hypothesis test, we can always lower the confidence level, α
B. Because the assumption of normality was violated so a different p-value chart must be used.
C. To control the experimentwise error rate when performing many hypothesis tests at once
D. #statistics

The methods of Tukey and Dunnett covered in class were used to…

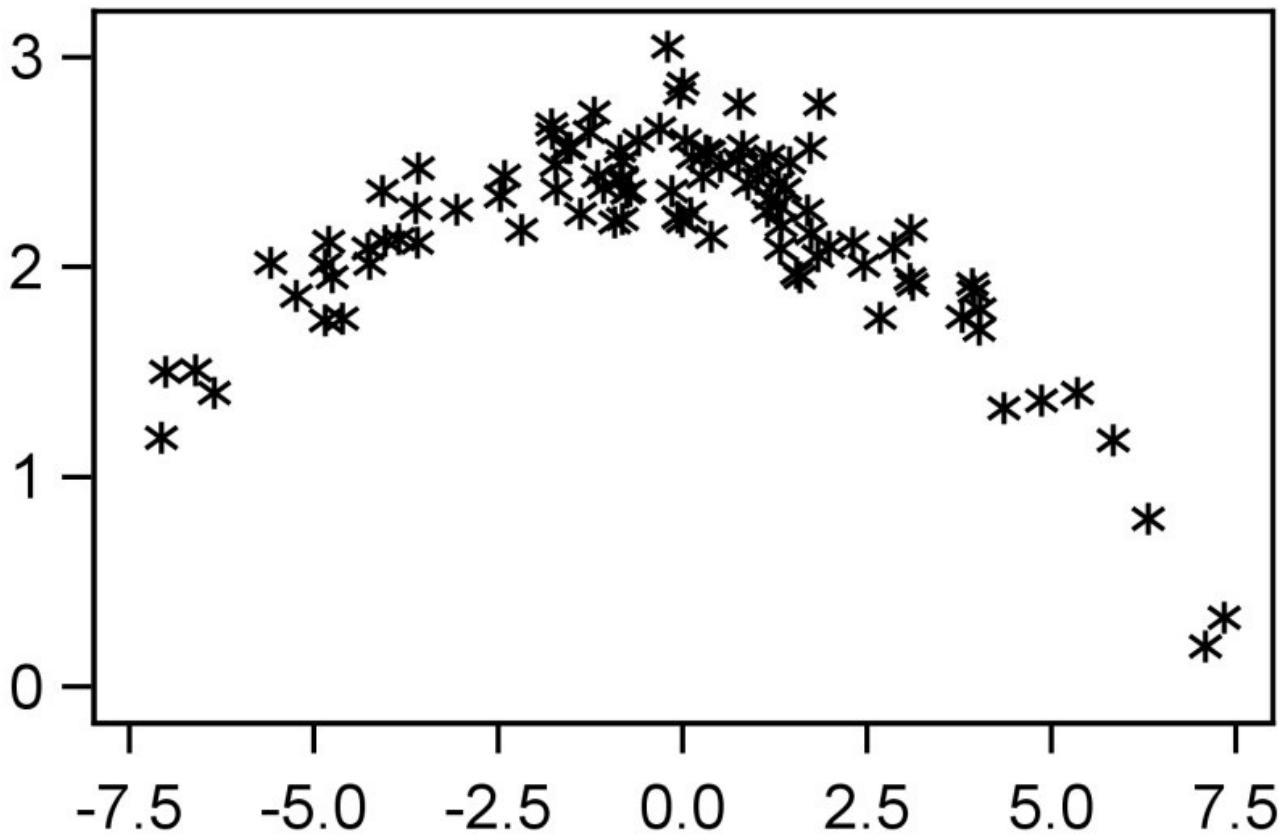
A. Run an ANOVA with unbalanced designs
B. Test a hypothesis about means with unequal variances
C. To control the experimentwise error rate for multiple hypothesis tests
D. Test for equality of variances
E. Torture me.

The methods of Tukey and Dunnett covered in class were used to…

A. Run an ANOVA with unbalanced designs
B. Test a hypothesis about means with unequal variances
C. To control the experimentwise error rate for multiple hypothesis tests
D. Test for equality of variances
E. Torture me.

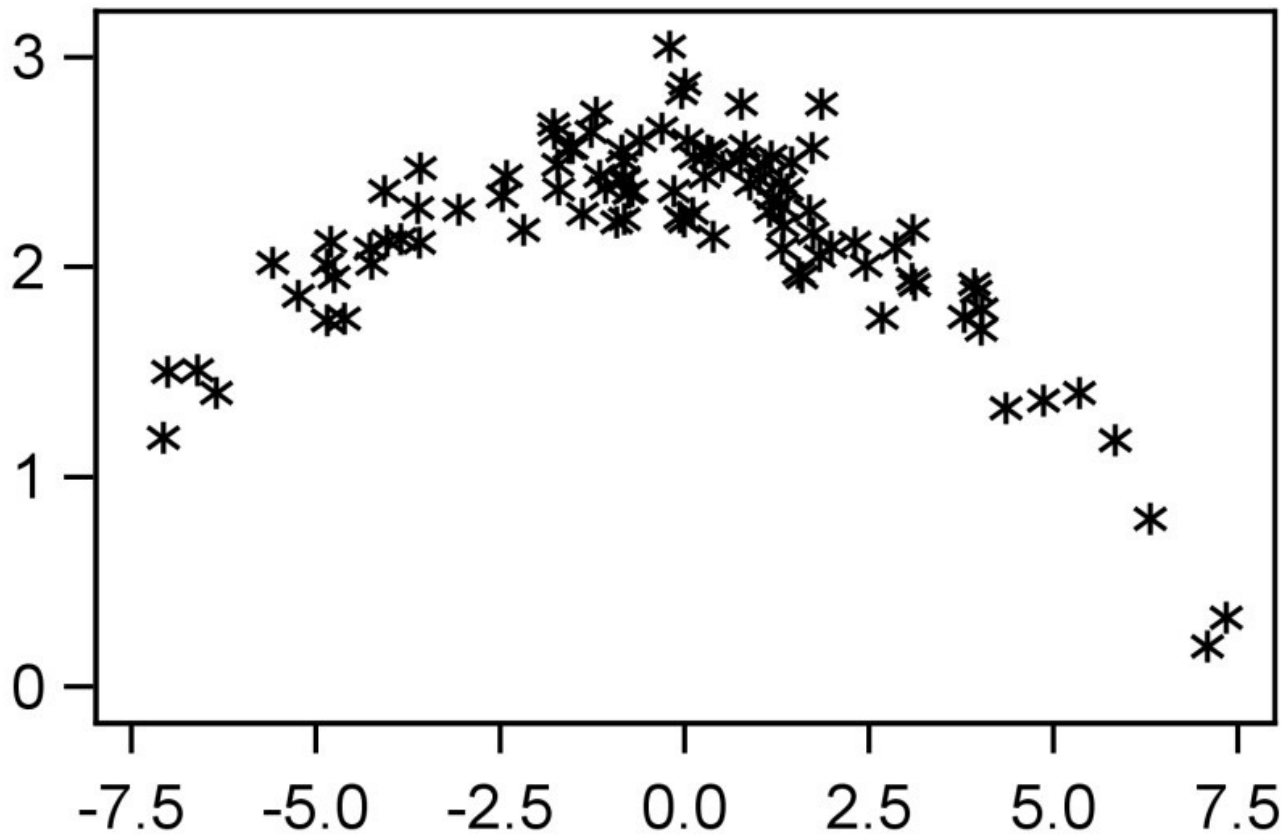
For the scatter plot below, what is a reasonable value for Pearson's correlation?

A. -0.8
B. -1
C. 1
D. -0.1
E. ruh rho



For the scatter plot below, what is a reasonable value for Pearson's correlation?

A. -0.8
B. -1
C. 1
D. -0.1
E. ruh rho



Multiple linear regression can be used for the following tasks:

A. To predict values of a response variable when data is collected on the input variables
B. To explain how the input variables are related to the target variable
C. To explain which input variables cause changes in the target variable
D. All of the above
E. A and B

Multiple linear regression can be used for the following tasks:

A. To predict values of a response variable when data is collected on the input variables
B. To explain how the input variables are related to the target variable
C. To explain which input variables cause changes in the target variable
D. All of the above
E. A and B

Why is adjusted-R^2 adjusted?

A. To account for collinearity of the predictor variables
B. To account for the number of variables in the model
C. To account for the number of observations in the training data
D. To more accurately reflect the variance of the response variable
E. It was too cold.

Why is adjusted-R^2 adjusted?

A. To account for collinearity of the predictor variables
B. To account for the number of variables in the model
C. To account for the number of observations in the training data
D. To more accurately reflect the variance of the response variable
E. It was too cold.

For the information criterion metrics (AIC, BIC) a more desirable model has…

A. Smaller values on these metrics
B. Larger values on these metrics
C. As many variables as possible

For the information criterion metrics (AIC, BIC) a more desirable model has…

A. Smaller values on these metrics
B. Larger values on these metrics
C. As many variables as possible

Which of the following violations of linear regression assumptions will affect the parameter estimates?

A. Violation of normality
B. Violation of constant variance
C. Violation of linearity
D. None of the above
E. All of the above
F. Linear regression has no assumptions.

Which of the following violations of linear regression assumptions will affect the parameter estimates?

A. Violation of normality
B. Violation of constant variance
C. Violation of linearity
D. None of the above
E. All of the above
F. Linear regression has no assumptions.

Which of the following violations of linear regression assumptions can you NOT use residual *scatter plots* to diagnose?

A. Violation of multicollinearity
B. Violation of constant variance
C. Violation of linearity
D. None of the above
E. All of the above
F. Linear regression has no assumptions.

Which of the following violations of linear regression assumptions can you NOT use residual *scatter plots* to diagnose?

A. Violation of multicollinearity
B. Violation of constant variance
C. Violation of linearity
D. None of the above
E. All of the above
F. Linear regression has no assumptions.

Which of the following metrics is NOT used to diagnose multicollinearity

A. Correlation
B. Variance Inflation Factor
C. Cramer's V

Which of the following metrics is NOT used to diagnose multicollinearity

A. Correlation
B. Variance Inflation Factor
C. Cramer's V

Which Chi-square test is specifically designed to test the association between two ordinal variables?

A. Pearson Chi-square
B. Likelihood Ratio Chi-square
C. Mantel-Haenzel Chi-square
D. My Chi-square
E. Insert historical name here Chi-square.

Which Chi-square test is specifically designed to test the association between two ordinal variables?

A. Pearson Chi-square
B. Likelihood Ratio Chi-square
C. Mantel-Haenzel Chi-square
D. My Chi-square
E. Insert historical name here Chi-square.

Which of the following can tell you the *strength* of association between two ordinal variables?

A. Pearson Chi-square
B. Spearman Correlation
C. Mantel-Haenzel Chi-square
D. Pearson's correlation
E. Strengths' finder

Which of the following can tell you the *strength* of association between two ordinal variables?

A. Pearson Chi-square
B. Spearman Correlation
C. Mantel-Haenzel Chi-square
D. Pearson's correlation
E. Strengths' finder

If group A has 2 times the odds of group B to have an event, then …

A. Group B has 50% the odds of group A to have the event.
B. Group A has 50% the odds of group B to NOT have the event.
C. Group A has 2 times the odds of NOT having the event compared to group B.
D. Both A and B
E. Stop it. Just stop it.

If group A has 2 times the odds of group B to have an event, then …

A. Group B has 50% the odds of group A to have the event.
B. Group A has 50% the odds of group B to NOT have the event.
C. Group A has 2 times the odds of NOT having the event compared to group B.
D. Both A and B
E. Stop it. Just stop it.

If a logistic model predicts the probability that you fail this course, which of the following is an example of a concordant pair?

A. A passing student with predicted value 0.9 and a failing student with predicted value 0.5
B. A passing student with predicted value 0.1 and a failing student with predicted value 0.5
C. A passing student with predicted value 0.9 and a failing student with predicted value 0.9
D. A passing student with predicted value 0.9 and a
   - passing student with predicted value 0.5
E. Logarithm.

If a logistic model predicts the probability that a student fails a test, which of the following is an example of a concordant pair?

A. A passing student with predicted value 0.9 and a failing student with predicted value 0.5
B. A passing student with predicted value 0.1 and a failing student with predicted value 0.5
C. A passing student with predicted value 0.9 and a failing student with predicted value 0.9
D. A passing student with predicted value 0.9 and a passing student with predicted value 0.5
E. Logarithm.

What is the default coding of categorical variables in R?

A. GLM in R doesn't allow for categorical variables
B. Effects Coding
C. Reference Coding
D. Python coding

What is the default coding of categorical variables in R?

A. GLM in R doesn't allow for categorical variables
B. Effects Coding
C. Reference Coding
D. Python coding

Which of the following is true about the logits $(\log\left(\frac{p}{1-p}\right))$?

A. They are unbounded
B. They take on values between -1 and 1
C. They take on values between 0 and 1
D. They are equal to the probability that an event happens.
E. They are logitimate.

Which of the following is true about the logits $(\log\left(\frac{p}{1-p}\right))$?

A. They are unbounded
B. They take on values between -1 and 1
C. They take on values between 0 and 1
D. They are equal to the probability that an event happens.
E. They are logitimate.

Which of the following statements is true?

A. It is impossible to consider an odds ratio for continuous predictor variables
B. The coefficients in logistic regression have no interpretation in the context of the problem.
C. Logistic regression cannot be used for explanation, only for prediction of probabilities.
D. Logistic regression can predict response variables that are binary, ordinal, or nominal.
E. This statement is false.

Which of the following statements is true?

A. It is impossible to consider an odds ratio for continuous predictor variables
B. The coefficients in logistic regression have no interpretation in the context of the problem.
C. Logistic regression cannot be used for explanation, only for prediction of probabilities.
D. Logistic regression can predict response variables that are binary, ordinal, or nominal.
E. This statement is false.