Webinar Transcript
Understanding Education Statistics

Slide No. 1:

Welcome to today's webinar: Understanding Education Statistics. This webinar is brought to you by District 180 in the Office of Continuous Improvement and Support at the Kentucky Department of Education.

Slide No. 2:

Here are our objectives for today:

By the end of this webinar, participants will be able to…

- recognize common statistical language and
- interpret the results of statistical tests commonly used in education research.

Slide No. 3:

The agenda for this webinar is on the screen. We will begin with an introduction to educational statistics and a discussion regarding this webinar's alignment to existing statutes, regulations, and guidance. Then, we will provide an overview of descriptive and inferential statistics, followed by a discussion on statistical significance and interpreting the magnitude of a finding.

Slide No. 4:

In 2015, the U.S. Congress passed the *Every Student Succeeds Act*, or ESSA. This law reauthorized the Elementary and Secondary Education Act and created new mandates for school improvement efforts across the country. One of those new mandates is that processes, practices, products, and strategies used for turnaround efforts be rooted in evidence.

ESSA defines four levels of evidence, each with the requirement that studies demonstrate statistically significant, or substantively important, improvement on a student outcome. In this webinar, we will present information related to common statistical tests to support educators as they seek to interpret and apply research findings in schools and classrooms.

Slide No. 5:

Effort has been made to ensure that the definitions and key concepts presented in this webinar align to multiple educational resources. Those references include the Code of Federal Regulations, the *Non-Regulatory Guidance: Using Evidence to Strengthen Education Investments*, the What Works Clearinghouse Version 4.0 Standards Handbook and the What Works Clearinghouse Version 4.0 Procedures Handbook. All of these resources have been hyperlinked

in the PowerPoint for your convenience. General definitions and interpretations for this webinar can be attributed to the book “Basic Statistics: Tales of Distributions” by Chris Spatz.

Slide No 6:

We will begin our discussion today with descriptive statistics.

Slide No. 7:

Descriptive statistics are mathematical tests that tell us general information about a body of data. They are generally well understood by most educators and include very common tests such as average, median, and standard deviation. Descriptive statistics usually provide a foundation that can be used for further analysis by more complex statistical tests and therefore will be included in most research studies.

Slide No. 8:

The most common set of descriptive statistics are a classification called “Measures of Central Tendency.” These measures produce a single number that describe a data set. There are three primary measures of central tendency: mean, median, and mode.

The mean represents the average of a set of scores, or the sum of the scores divided by the number of scores. The mean is a valuable measure that tells us generally how well a group of students performed on an assessment or how well a single student performed on a group of assessments. The mean is the formula used by most educators to calculate grades at the end of the term. Later in this webinar, we will discuss how the mean can be used to draw further conclusions about a body of data.

The median represents the middle number in a distribution of scores. To find the mean, the study author places the scores in ascending order, for lowest to highest, and then finds the middle. If there is no middle score, the average of the two center scores is calculated to determine the median. In other words, the median represents a hinge point, where half of the students scored higher than the median and half of the students scored lower.

Finally, the mode represents the score that occurs most often within a distribution of scores. This can be valuable with exploring clusters of students and discussing why they are or are not moving at the same rate as other students. A distribution of scores can have two modes, which is called bimodal, or three or more modes, which is often called a multimodal distribution.

Slide No. 9:

Another common way to describe a data set is through measures of variability. Variability provides a single number that describes how spread out the scores in a data set are. This is valuable information for educators who need to ensure that all students are moving. Generally, when variability is low, students are moving as a group, whereas a data set with high variability may suggest that some students are moving faster or slower than others. There are three main

measures of variability used in educational research: the range, interquartile range, and standard deviation.

Slide No. 10:

The range represents the distance between the highest score and the lowest score in a data set. While range represents the whole body of data, the interquartile range represents only the middle fifty percent of scores. Typically, these two calculations are represented in a box and whisker plot.

Slide No. 11:

On this plot, the box represents the interquartile range while the whiskers represent the full range of scores. A box and whisker plot will also typically indicate the median scores for comparison.

Let's pretend that the example on the screen represents student test scores. You can see that there are roughly seventy points between the highest and lowest score plotted, while the median score is around fifty. A plot such as this can help us see that fifty percent of our students (those represented by the interquartile range) are clustered within the forty to sixty percent range, and we have at least one student who is dramatically outscoring the rest of the students in this group.

Now, if this plot represented a pre-test, we could place a posttest plot next to it to see how far our students have moved, as a group, after instruction or intervention.

Slide No. 12:

This plot gives us new information about how the students performed on our fictional assessment. You can see that the class, as a whole, performed much higher on the post-test than the pre-test. The whiskers on the post-test plot, representing the range, are much shorter than on the pre-test plot, telling us that the student scores are much closer together. We can also see that our median score has risen by about twenty-five percent and the outer edges of the box, representing our interquartile range, are closer to the median. This suggests that fifty percent of our students are scoring closer to the median score than before. Graphic representations such as this are very common in the reporting of educational research.

Slide No. 13:

Standard deviation is another common measures of variability. Like range, it helps us understand how spread out the data is, but it goes a step farther by exploring how the data is distributed around the mean. When a standard deviation measure is low, it means that the scores are clustered around the mean, while a higher standard deviation means that are scores are more spread out.

Slide No. 14:

Standard deviation is represented graphically with a bell curve chart like the one on the screen. The tall vertical line in the middle, labeled zero on the chart, represents the mean. You can see that in this example, 34.1 percent of scores fall within one standard deviation of the mean in either direction. This means that roughly 68 percent of scores are clustered together near the mean. As you move farther away from the mean, we see that 13.6 percent of scores are one standard deviation away from the mean, 2.1 percent of scores are located two standard deviations from the mean, and so-on.

Standard deviation is an important measure for educators because it helps us not only see how spread out student scores are, but also how far away from the mean different groups of students may lie. For example, a teacher may assign students who scored two standard deviations above the mean to enrichment activities, while students who scored two standard deviations below the mean may be identified for intensive intervention.

This type of grouping is one way that education research may assign students to group. This is an assignment method called stratification.

Slide No. 15:

Altogether, the descriptive statistics discussed in this webinar help educators and education researchers to gain a general understanding of a body of data. These easily calculated statistical measures can be used to explore student performance and make instructional decisions. Next, we will see how descriptive statistics can be used to further make inferences from the data.

Slide No. 16:

Let's stop here to check for understanding. On the slides that follow, you will be asked to interpret the findings of three statistical tests.

Slide No. 17:

Question One. In the chart below, you will find information reported from a set of student test scores. What do these measures tell you about the students who took the assessment? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 18:

Now that you have had a chance to look at the data, let's talk about what it tells us. First, we can look at the spread of the scores. Here, the scores are pretty widely spread, from 18 percent to 100 percent. This tells us that some students demonstrated mastery while others still need help. The median, or middles core, and the mean, or average, are pretty close together. That tell us that the majority of our students fell near the middle of the distribution. We also know that half of our students scored between 25 percent and 79.75 percent. Assuming a mastery score is 80

percent or above, this tells us that the majority of our students missed the mark and some re-teaching is required.

Slide No. 19:

Question Two. The histogram below shows the distribution of the average years of teaching experience in Kentucky public schools. What information can you learn from the distribution? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 20:

A histogram shows the distribution of a set of scores. Since this distribution looks similar to the bell curve we discussed earlier, we know that this data follows a normal distribution pattern. In other words, the data falls the way we would expect it to. We also know that the majority of schools in Kentucky have a teaching staff with an average experience of 10-15 years because the highest bars fall within that range. You may have noticed that the average years of teaching experience in a school drops sharply after 15 years. A next step may be to look at which schools have a very high average teaching experience and see what other characteristics in those schools contribute to the longevity of their teachers.

Slide No. 21:

Question 3. The box and whisker plot to the right represents the distribution of proficiency ratings in Kentucky's accountability system in the Fall of 2018. What insights can you glean from the boxplot? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 22:

Remember that a box and whisker plot gives you descriptive information about a data set. The bold line in the middle represents the median. Here, we can see that the median proficiency rate for schools in Kentucky is around 70 percent. The box represents the middle half of your scores. In this case, we can see that half of Kentucky's schools fall within 60 and 80 percent, a range of 20 points. Finally, dots represent outliers. We can see that a handful of schools scored far above the majority of schools. If we were looking to improve our school's scores, it may be beneficial to identify who those outlier schools are and examine the characteristics of those schools.

Slide No. 25:

Great work! We have explored the world of descriptive statistics and how they can be useful in analyzing a set of data. If descriptive statistics give us descriptive information from a data set, then inferential statistics are those that allow us to make an inference, or draw a conclusion, from a set of data. In part two we will explore a set of inferential statistics that are commonly found in education research.

Slide No. 24:

Correlation is arguably the most common inferential statistic used in educational research. That is because it can be on its own to make an inference about a set of data, or combined with other tests to draw a more concrete conclusion. Correlation describes the degree of relationship between to variables. It is used to answer questions like: “Is there a relationship between academic achievement in math and student gender?” or “Does household income impact student attendance rates?”

In addition to these simple comparison questions, correlation is also used to determine the reliability (or consistency) of academic tests. In this model, a sample test is given to a group of students, and then given to them again, without instruction in between. If the test is reliable the two sets of scores will have a high level of agreement – or a high correlation – if the test is not-reliable, the two sets of scores will have a low correlation. This calculation is often seen in studies that use teacher-created assessments rather than established standardized assessments in their collection protocol.

When calculated, the correlation is reported as a “Pearson product-moment correlation coefficient” and is represented by a lower-cased *r*.

Slide No. 25:

Correlation is graphically represented by a scatter plot. You can see three scatter plots on the screen. Scatterplot one represents a strong linear correlation. You can see that all of the scores are clustered together and slope positively to the right. Plot two represents a weak correlation. In this plot, the scores slope positively to the right, but they are more spread out than in the first plot. Finally, scatter plot three represents a data set with no correlation. The data is not clustered, nor does it have a slope.

Slide No. 26:

While correlation can help us gain information about the relationship between two items, it cannot tell us if one variable directly influences another. For example, if we find that low household income has a positive correlation with low student attendance, then we may want to direct some of our resources towards influencing the attendance of lower-income households, but we cannot assume that having a low household income causes low student attendance.

It is important that we interpret the results of correlational tests correctly when using them for education decision making. This test alone does not give us the whole picture and the best practice would be to seek other data points to supplement this finding before committing time and resources towards a potential solution.

Slide No. 27:

When a researcher is trying to determine a causal relationship between two variables, the t-test is commonly deployed. The t-test is used in hypothesis testing – or tests that seek to prove that one variable does, or does not, impact another variable. For example, a researcher may use the

t-test to answer the question, “Does this intervention improve student scores on this assessment?”

The t-test uses the data to examine a hypothesis and either reject, or not reject, the hypothesis. This is called the “null hypothesis.” Essentially, the null hypothesis is a hypothetical statement that the data may or may not support. It is common for researchers to write something like “this result rejects the null hypothesis” to help the put the results of a statistical test in context.

The t-test uses a variety of calculations already discussed in this webinar. The t-test is, at is core, a comparison of mean scores, and some formulas also include *r* or the correlation coefficient.

The results of a t-test are reported using a lower-cased *t.*

Slide No. 28:

This is also a good time to introduce the concept of *degrees of freedom*. Similar to standard deviation, the degrees of freedom measurement tell us how much variability there may be within the results of a statistical test. This measurement, typically reported as a lower-cased *df*, should be included in all t-test results, as well as the results for other common inferential statistical tests. This number is important for determining statistical significance, which we will discuss at the end of the webinar.

Slide No. 29:

The t-test is a useful tool for researchers who want to compare two variables, but when more than two variables are necessary they turn to the Analysis of Variance test, commonly called ANOVA. Like the t-test, ANOVA tests are calculated using the means of the data sets. In education, an ANVOA test may be used to answer a question like “Do students from urban, suburban, and rural districts differ in their attitudes towards the importance of school?”.

The ANOVA test is reported as the *F* distribution, written as a capitalized *F.*

You may see other variations of this test, such as the Multiple Analysis of Variance (MANOVA) and the Analysis of Covariance (ANCOVA). We will not explore the differences between these tests in this webinar. For our purposes, it is sufficient enough to know that these tests help researchers create more accurate findings under certain conditions.

The ANOVA tests can only tell you whether or not relationships exist between variables, but it cannot tell you where those relationships lie. For that, researchers must deploy one of a variety of post-hoc, or after-the-fact, tests to identify the specific relationships between variables.

Slide No. 30:

The final inferential test we will discuss in this webinar is the chi-square. The chi-square is used when we want to compare a data set with our expectations of a data set.

Consider this hypothetical example. We asked a group of high-school freshman whether they wanted pizza or ice cream as a reward for on-time homework assignments. The results of that survey are on the screen. We might expect that the results would be split evenly between male and female students, as in Table One. If we found, however, that female students seem to want pizza and male students seem to want ice cream, as you can see in Table Two, the chi-square test can help us to determine if there is any statistical significance between gender and reward preference. In other words, the chi-square test tell us how confident we can be in making the statement that there is a difference between our expected results and our observed results.

While most education leaders would not conduct a chi-square to determine the reward patterns for on-time-assignments, the chi-square method can be used by education researchers to answer much more serious questions that can impact policy decisions. Any time the data can be broken down into a chart like the one on the screen, a chi-square can be used to draw a conclusion. For example, we may want to compare the opinions of parents who send their kids to public school versus those who send their kids to private school on a given variable, or compare the proficiency rate of students in schools with a population over 500 with schools with a smaller population.

Slide No. 31:

Inferential statistics help researchers draw conclusions about a set of data. They typically build off of the differential statistics discussed in the previous section of this webinar and provide greater insight into a population. So far, we have simply discussed inputs and outputs. Now, let's talk about how we know what those numbers mean.

Slide No. 32:

The result of a statistical test is simply a number, until we give it some context. That is where measures of statistical significance come into play.

The significance of a statistical test is reported by the p-value, reported as a lower-cased *p.* The p-value comes into play for all hypothesis tests, and tell us the probability of the statistical test value when the null hypothesis is true.

For example, if you make a null hypothesis that says "this intervention will improve outcomes for this student population", then the p-value tells you the probability of the data observed.

It is generally accepted that a statistical test is significant when it is less than 0.05, written as $p<0.05$. If we apply this number to the null hypothesis stated a moment ago, this would mean that the sample results you got would occur fewer than five times in one hundred attempts.

The p-value is considered highly significant if it is less than 0.01 or 0.001.

Slide No. 33:

To determine the p-value, researchers compare the scores of their statistical test and the degrees of freedom to a distribution chart like the one on the screen. This chart represents the first seven rows of a *t* distribution table. These standardized tables are widely available and regularly referred to by even the most senior statisticians. In order for a *t* score to be significant, it must be equal to or greater than the value shown in the column that aligns with the degree of freedom.

For example, a score of t=5.64 with a degree of freedom measurement of four would be considered statistically significant with a *p* value of 0.01.

Slide No. 34:

If the p-value tells you whether or not a statistical finding is significant, the effect size provides context as to the magnitude of a finding. The effect size compares the distance between the means of two sets of data. In a pre-test, post-test model, the effect size would be calculated by subtracting the mean of the pre-test from the mean of the post-test, and dividing by the standard deviation. This would tell you how far apart the two sets of scores are.

Effect size can be calculated using either the formula for Cohen's *d* or Hedge's *g,* and is reported as a lower-cased *d* or *g* respectively.

There is much debate in the scholarly world related to the interpretation of effect size. While there are many arguments for or against various interpretations, the Kentucky Department of Education aligns its standards with the What Works Clearinghouses, which states that an effect size is considered substantively important when it is equal to, or greater than, 0.25.

Effect size is an important measurement for educational decision making, but should not be the only measurement considered. It is important for education leaders to consider as much data as possible when making the decision to adopt or discontinue a practice, program, or strategy.

Slide No. 35:

It is time for another skill check. Just as before, the following slides will ask you to interpret the results of some statistical tests.

Slide No. 36:

Question four. The chart on the screen presents the correlation of various teacher characteristics in Kentucky public schools. Which traits are strongly correlated? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 37:

There are many different findings outlined in this chart. Let's take a look at three of them. First, we see that there is a positive correlation between the percentage of new teachers and percentage of teacher turnover in Kentucky schools. This makes sense because if you have a

school that has a high level of teacher turnover, they have to hire new teachers. We also see that there is a positive correlation between the average years of teaching experience and the percentage of teachers in Kentucky with a Rank I. Again, that makes sense, because the longer you teach the more likely you are to earn a Rank I.

While the correlation is lower, there is a slight, yet positive correlation between the percentage of teachers in Kentucky with a national board certification and the percentage of teachers in Kentucky with a doctorate degree. This may suggest that teachers who pursue national board certification are more like to pursue other forms of higher education or that teachers who pursue terminal degrees are more like to pursue other forms of professional learning.

Remember, correlation does not equal causation, so we cannot say any of this for certain, but it does allow us to make some informed guesses and create a platform for our decision making moving forward.

Slide No. 38:

Question Five. Below are the results of an analysis of variance test that compares the rate of emergency certified teachers in schools identified as TSI, CSI, and Other. Can you describe the outcome? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 39:

The reporting you saw on the previous screen shows how the results of an Analyses of Variance test are often reported in scholarly journals. Remember that an Analysis of Variance, or ANOVA, test looks for relationships between the means of two or more variables. You can tell this is an ANOVA test because it is reported with a capital “F”. If you look at the p value, you can see that p = 0.133. In order for it to be significant, *p* must be less than 0.05. So these results tell us that there is no significant relationship between the rate of emergency certified teachers in schools identified as TSI, CSI, or Other.

Slide No. 40:

Question six. Last one! You are designing an intervention class for third graders who are falling behind in reading. You need to select a teaching strategy to deploy in the intervention class. Below are the effect sizes of several teaching strategies. Which strategies are you going to deploy in your class? Pause the webinar now and replay when you are ready to hear the discussion.

Slide No. 41:

Okay – so this one was kind of a trick question. As we mentioned earlier, effect size is an excellent way to estimate how much change you can expect an intervention to have, but it really should not be the sole descriptor used to inform your decision. When reflecting on the effect sizes of the strategies listed, all of them are over the substantively important threshold of 0.25.

In theory, any of these strategies, or a combination of them, should help your struggling learners.

Slide No. 42:

Thank you for taking time to view this webinar. During this webinar, we provided an overview of common statistical tests for educational research. We discussed a variety of descriptive statistics, along with their use and calculation. We reviewed some basic information about inferential statistics and how they help education researchers draw conclusions. Finally, we touched on the importance of statistical significance and interpreting the magnitude of a finding.

Slide No. 43:

If you have questions regarding evidence-based practices or educational statistics, please contact the District 180 branch in the Office of Continuous Improvement and Support at (502) 564-2116.