Q-core Assessment Report

Course: Stat 216Q

Semesters: Fall 2014 and Spring 2015

Instructor(s) and/or supervisor: Course Supervisor: Jim Robison-Cox

Student Success Coordinator: Jade Schmidt

Assessment done by : *Jim Robison-Cox and Jade Schmidt*

Number of students in course: *Fall 2014: 820, Spring 2015: 620*

Number of students assessed (at least 6): *Approximately 1140*

Description of assignment, problems, and/or questions used for assessment*:*

We participated in a national assessment led by Dr. Nathan Tintle of Dordt College in which students were given pre and post test questions about common statistical topics. Not all students participated – it was a voluntary response sample - but of 820 enrolled in Fall 2014, 620 took both pre and post test, and in Spring 2015, of 620 enrolled, we had 519 take both. The authors of the assessment have shared section-by-section summaries of the responses for each question. For this report we have extracted several questions from the assessment which apply to the specific learning outcomes of interest.

The study also provides a comparison group which was again a voluntary sample at two levels. First, statistics teachers were invited to use the assessment, and secondly, students in the participating classrooms were given the opportunity to answer the questions. No students were graded on the correctness of their response, but many, like ours, were given a small bonus for participation. The students included are taking introductory statistics from teachers who have an interest in participating in a statistics education study, and this is important information when considering the reference group we are using to compare results.

Learning Outcome 1: *Interpret and draw inferences from mathematical or statistical models represented as formulas, graphs, or tables*.

We interpret this outcome, in the context of STAT216Q, as the ability to properly interpret visual representations of data, for instance, to understand that we have stronger evidence that two means differ when a) the two sample means are further apart and b) the distributions have less spread. Assessment questions 23, 49 provide information on these outcomes.

- Total number assessed: Fall 2014: 620; Spring 2015: 519

- Proportion correct: Because we averaged over two assessment questions, only proportions (not raw counts) are reported. In Fall 2014 77.6% were correct and in Spring, 81.7% were correct (on average). These are similar to the 80.8% correct which Tintle et al. report for their group of US students.

- Is this over the specified threshold of 2/3? Yes

Learning Outcome 2: *Represent mathematical or statistical information numerically and visually.*

Students were asked to identify skewed distributions (Questions 33-34) and to compare variances of samples from plots. (Questions 42-43)

- Total number assessed: Fall 2014: 620; Spring 2015: 519

- Proportion correct: In Fall 2014, 51.1% of answers were correct, on average, over the 4 assessment items. In Spring 2015 the number was similar: 49.6%. The Tintle et al. comparison group did a bit better at 55.9%

- Is this over the specified threshold of 2/3? No.

Learning Outcome 3: *Employ quantitative methods such as arithmetic, algebra, geometry, or statistical inference to solve problems.*

Many of the questions involve interpretation of p-values in hypothesis testing and confidence intervals which are the main tools of statistical inference. We will average proportions correct over questions 29, 30, 31, and 33 to assess understanding of p-values, and over questions 20, 21, 22, and 48a-b for interpretation of interval estimates.

- Total number assessed: Fall 2014: 620; Spring 2015: 519

- Proportions correct:

Interpretation of:	MSU Fall 2014	MSU Spring 2015	US Comparison
P-value	76.5%	75.8%	75.5%
Confidence Intervals	56.1%	56.0%	56.5%
Combined	65.2%	64.8%	64.9%

- Is this over the specified threshold of 2/3?

For interpretation of p-values we seem to have succeeded, but we have work to do on confidence interval interpretation. Both sets of results are strongly similar to the US comparison group.

Reflections on improvements:

We need a better way to evaluate the second set of objectives. Our students are using graphs in class every day to separate out extreme observations in a null distribution, and they transfer such plots to their activity work books, so we think the assessment did not ask them the right questions. We also need to think about what it means to "represent statistical information numerically". Computing appropriate summary statistics could fit into this category, and certainly computing a p-value or a confidence interval could as well.

For Outcome 3, we do intend to focus more effort on confidence interval interpretation, and will build questions to probe this understanding into future final exams. We plan to reassess this effort in Spring 2016 with results from the final exam this semester.

Reflections on this assessment:

- In general, our students have succeeded in the areas which the US group showed success, and had difficulty with concepts over which others have also had difficulty. Because the background group is being taught by instructors with a strong interest in statistics education, comparable numbers indicate that we are doing as well as other institutions which take a lead in statistics education.

- Alignment with core objectives:
  The Q core is designed to teach students quantitative reasoning, and introductory statistics should be a primary course in which students learn to "make sense of data".

- Assessment Process:
  With the large numbers of students enrolled in Stat 216 each semester, we feel uncomfortable with assessing just a few assignments from a few students who were taught by a subset of our instructors. We instead are using a nationally developed multiple choice web-based survey tool which had a high student participation rate. The NSF funded assessment will likely not be available in future years. As a substitute, we have adapted some similar questions (derived from the GOALS exam[1]) and use them as part of our final exam, so these could be used to assess many of these learning outcomes. We need to make some changes or additions to get information on learning outcomes 2 and 3 above.

Appendix

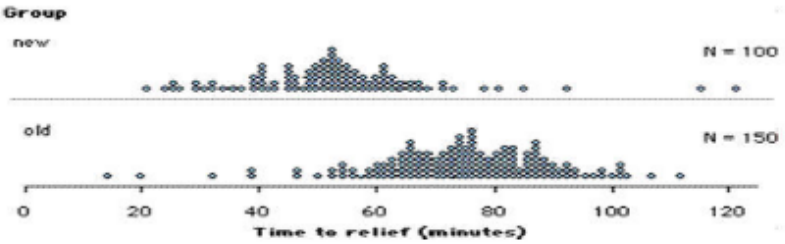
Questions from the Tintle assessment.

---

1 Zieffler, Andrew, Joan Garfield, and Robert Delmas. "Development of an instrument to assess statistical thinking." (2010). ICOTS8 Invited Paper

Outcome 1 Question 23 and 49

Question 23. Two hundred fifty people who frequently suffer from headaches agreed to participate in a study. One hundred of these people were randomly assigned to receive a new headache medication when they had a headache, and the other 150 people received the old headache medication. The time until the patient reported that they no longer had a headache was recorded. The results are shown below:



Which of the following is the most valid conclusion for these data?

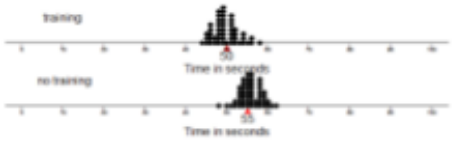
Select one:

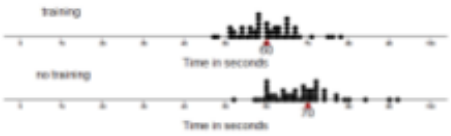
A. The new medication may be preferable. People taking the new medication tended to feel relief about 20 minutes sooner, on average, than those taking the old medication.

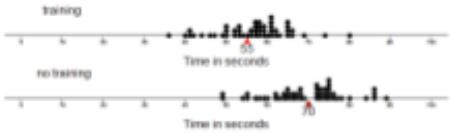
B. Neither medication is preferable. The number of patients in the two groups is not the same so there is no fair way to compare the two medications.

C. The old medication works better. Two people who took the old medication felt relief in less than 20 minutes, compared to none who took the new medication. Also, the worst result near 120 minutes was with the new medication.

Question 49. Which pair of dotplots provides the strongest statistical evidence that the Training group ran faster (smaller times), on average, than the No Training group?





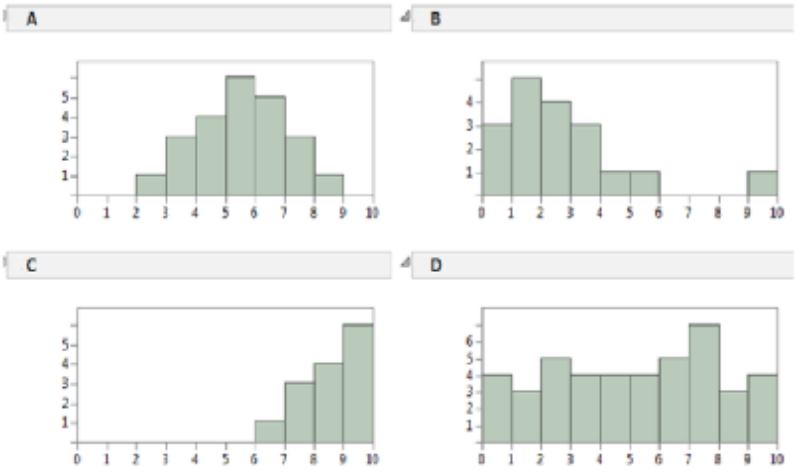


Outcome 2 Questions 33, 34, 42, 43

Question 33. A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article gives a p-value of 0.04 in the analysis section.

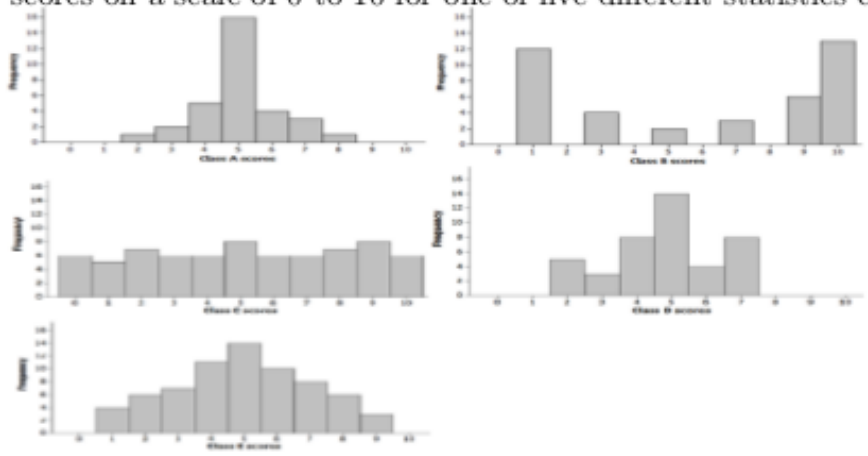
Indicate whether the following interpretation is a valid or invalid interpretation of this p-value:
We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.
Select one:
A. Valid
B. Invalid

Question 34. Four histograms are displayed below. Match the description to the appropriate histogram.



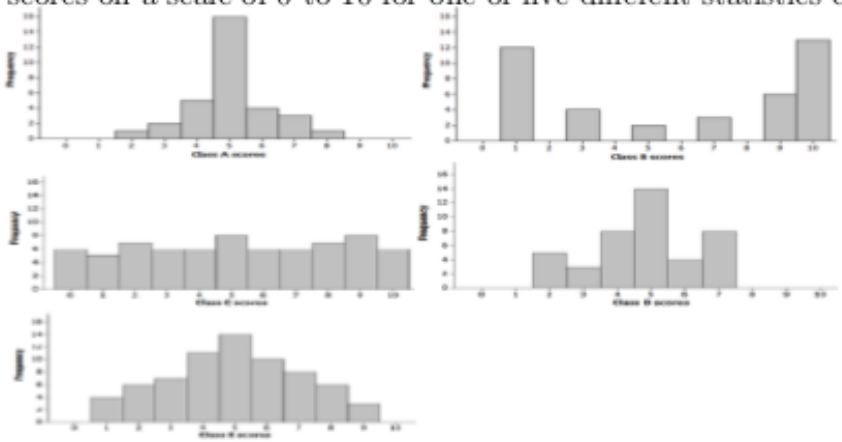
A distribution for the second to last digit of phone numbers sampled from students in a class (i.e., for the phone number 9689667, the second to last digit is 6) is best represented by:

Question 42. Five histograms are presented below. Each histogram displays test scores on a scale of 0 to 10 for one of five different statistics classes.



Which of the classes has the least variability of scores?

Question 43. Five histograms are presented below. Each histogram displays test scores on a scale of 0 to 10 for one of five different statistics classes.



Which of the classes has the greatest variability in scores?

Outcome 3 Questions 29, 30, 31, and 33 on p-values

Question 29. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of pvalue would she want to obtain?

Select one:
A. The magnitude of a p-value has no impact on statistical significance
B. A large p-value
C. A small p-value

Question 30. A research article reports the results of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a pvalue of 0.04 in the analysis section.

Indicate whether the following interpretation is a valid or invalid interpretation of this pvalue:
We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.
Select one:
A. Valid
B. Invalid

Question 31. A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article gives a pvalue of 0.04 in the analysis section.

Indicate whether the following interpretation is a valid or invalid interpretation of this pvalue:
We conclude that the new drug is effective because results like they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective.
Select one:
A. Valid
B. Invalid

Question 31. A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article gives a pvalue of 0.04 in the analysis section.

Indicate whether the following interpretation is a valid or invalid interpretation of this pvalue:
We conclude that the new drug is effective because results like they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective.
Select one:
A. Valid
B. Invalid

Question 33. A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article gives a p-value of 0.04 in the analysis section.

Indicate whether the following interpretation is a valid or invalid interpretation of this p-value:
We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.
Select one:
A. Valid
B. Invalid

Question 20. A high school statistics class wants to estimate the average cookie weight of a generic brand of chocolate chip cookies. They collect a random sample of 50 cookies from the manufacturing process and obtain the weight (in grams) for each cookie. Based on their data, the 95% confidence interval for the average weight per cookie is 25.65 to 26.35 grams.

Indicate whether the interpretation of the interval provided is valid or invalid:
We can infer with 95% confidence that a randomly selected cookie manufactured for this generic brand will weigh between 25.65 to 26.35 grams.

Select one:
A. Valid
B. Invalid

Question 21. A high school statistics class wants to estimate the average cookie weight of a generic brand of chocolate chip cookies. They collect a random sample of 50 cookies from the manufacturing process and obtain the weight (in grams) for each cookie. Based on their data, the 95% confidence interval for the average weight per cookie is 25.65 to 26.35 grams.

Indicate whether the interpretation of the interval provided is valid or invalid:
We can infer with 95% confidence that mean weight of all cookies manufactured for this generic brand is between 25.65 and 26.35 grams.

Select one:
A. Valid
B. Invalid

Question 22. A high school statistics class wants to estimate the average cookie weight of a generic brand of chocolate chip cookies. They collect a random sample of 50 cookies from the manufacturing process and obtain the weight (in grams) for each cookie. Based on their data, the 95% confidence interval for the average weight per cookie is 25.65 to 26.35 grams.

Indicate whether the interpretation of the interval provided is valid or invalid:
We can infer with 95% confidence that the average weight for 50 cookies randomly selected from those manufactured for this generic brand will be between 25.65 and 26.35 grams.

Select one:
A. Valid
B. Invalid

Question 48a. Suppose your teacher believes the confidence interval found in question 47 is too wide. She wants to know what could have been done to produce a narrower confidence interval and therefore a more precise estimate of the mean foot length for students at this university.
If you increase the sample size to 150, would this change produce a narrower confidence interval?
A. True B. False

Question 48b. Suppose your teacher believes the confidence interval found in question 47 is too wide. She wants to know what could have been done to produce a narrower confidence interval and therefore a more precise estimate of the mean foot length for students at this university.
If you increase the confidence level to 99%, would this change produce a narrower confidence interval?
A. True B. False