2016 Statistics in Science Olympiad Disease Detectives Events – Division C

Descriptive and analytic statistical methods are important in the life and health sciences and it is reasonable to expect competitors at the high school level to demonstrate a basic competency in these methods in events such as Disease Detectives. However, statistics is not a major component of the event and should not define the event. Many competitors will not have had statistics and event supervisors should use it like a habanero, sparingly if at all but with purpose. One or two short, well-placed questions are adequate.

Key principles

- Competitors should be able to use statistics – not do statistics. Competitors should not be expected to do lengthy calculations such as paired T-tests, life tables or analysis of variance using hand calculators. In real life, these are done with computers and statistical software packages such as SAS, SPSS or EpiInfo.
- A more realistic approach would be to explain the study design and data collection methods and have competitors interpret printouts from one of the above programs, to use intermediate statistics such as standard deviations and population means to calculate test statistics or tell how changes in the value of intermediate statistics would affect final results.
- Students should be provided formulas for all calculations and tables. It is reasonable to expect them to know which test to use, assumptions and limitations of each but not to memorize the formulas.
- Interpretation and translation of statistics into plain language is a valid competency that fits easily into a multiple choice format. For example, "Which of the below statements best describes these results?"
- Asking students to select a test or what test they would use to test a particular hypothesis is also appropriate. This question may be best suited for multiple choice unless graders have sufficient expertise to know which tests are clearly inappropriate and which are marginally appropriate and recognize the different names for identical tests.
- Questions concerning statistics should be part of a larger scenario so that competitors can see how they fit into the larger analysis.
- Free-standing questions on statistics just for the sake of including statistics into the competition are contrary to the intent of the event. Event supervisors should not feel compelled to include questions about statistics in the events if they do not fit into the larger picture.

Statistical Concepts for Disease Detectives
Division C

Descriptive Epidemiology - deals with the frequency and the distribution of risk factors in populations and enables to assess the extent of a disease - the epidemiologist collects information to characterize and summarize the health event or problem

- Used as a first step to look at health-related outcomes
- Examine numbers of cases to identify an increase
- Examine patterns of cases to see who gets sick and where and when they get sick

Mean – average – sum of all the given elements divided by the number of elements. Means are typically reported when data are normally distributed.

$$\bar{x} = \frac{\sum x}{n}$$

where sum of x scores and n = number of units in a sample

$\sum x =$

Median - The median of a series of numbers is the number that appears in the middle of the list when arranged from smallest to largest. Medians are typically reported when data are skewed.

For a list with an odd number of members, the way to find the middle number is to take the number of members and add one. Then divide that value by two. In our case, there are 9 numbers in the series. 9+1 = 10 and half of 10 is 5. The fifth number in the series is the median or 14.

If the number of members of the series was even, the average of the two middle numbers would be the median.

Mode - The mode is the number in the series that appears the most often. If there is no single number that appears more than any other number in the series, there is no value for the mode.

Variance - measures how far the results are from the expected results

$$S^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Standard deviation - expresses how close the results are to each other - ***it is the square root of the variance***

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Standard error of the mean - standard deviation (sample standard deviation) divided by the square root of the sample size

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Confidence intervals of means - specifies a range of **v**alues within which the mean may lie.

$$\overline{X} \pm Z\frac{s}{\sqrt{n}}$$

Z Scores for Commonly Used Confidence Intervals

Desired Confidence Interval	Z Score
90%	1.645
95%	1.96
99%	2.576

Analytic Epidemiology - the epidemiologist relies on comparisons between groups to determine the role of various risk factors in causing the problem

Z-test - compares sample and population means to determine if there is a significant difference. It requires a simple random sample from a population with a Normal distribution and where the mean is known. The Z-test is preferable when the sample number ***n*** is greater than 30. The Z value indicates the number of standard deviation units of the sample from the population mean.

The Z measure is calculated as: **z = (x - m) / SE**

where **x** is the ***mean sample*** to be standardized, **m (mu)** is the ***population mean*** and **SE** is the ***standard error of the mean***. where s is the population standard deviation and n is the sample size.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The z value is then looked up in a z-table. A negative z value means it is below the population mean (the sign is ignored in the lookup table).

T-test - An independent one-sample t-test is used to test whether the average of a sample differ significantly from a population mean, a specified value μ0

Paired T-test -The dependent t-test for paired samples is used when the samples are paired. This implies that each individual observation of one sample has a unique corresponding member in the other sample

Chi-square - Any statistical test that uses the chi square distribution used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value - Chi square test for independence of two attributes

McNemar test for paired data - McNemar's test is basically a paired version of Chi-square test – a form of chi-square (X^2) test for matched paired data which is used to compare paired proportions. It can be used to analyze retrospective case-control studies, where each case is matched to a particular control. Or it can be used to analyze experimental studies, where the two treatments are given to matched subjects.

Fischers exact test - a statistical test used to determine if there are nonrandom associations between two categorical variables

Cochran Mantel-Haenszel summary odds ratio (often called the Mantel-Haenszel test) is a hypothesis test for association between two variables while controlling for one or more nuisance or control variables. Mantel and Haenszel proposed stratification techniques to account for confounding.