Integrating Statistics into the Secondary Curriculum

Joe H Ward Jr and Paul A Foerster - San Antonio, Texas, USA

1. Introduction

Our experiences with high school students have indicated that statistical concepts can be introduced meaningfully into the secondary curriculum by integrating them not only into the existing mathematics curriculum, but also into other non-mathematics subjects. Curricula must be changed to reduce the amount of material covered and to focus on fewer basic concepts that can be applied to a wide variety of problem situations in the real world. Statistical and other scientific approaches to problem-solving should play an important role in the education of *all* students.

Foerster has already taken steps to integrate some concepts of probability and statistics into the secondary mathematics curriculum through his book *Precalculus with Trigonometry : Functions and Applications* (Foerster, 1987). Real-world phenomena are modelled by finding equations for linear, exponential, or variation functions which fit several data points *exactly*, then which fit many scattered data points *approximately*. The idea of least squares which arises in these regression problems sets the stage for standard deviation, and then to predictive statistics later in the text.

In order to simultaneously *reduce* the sheer number of statistical procedures to be learned and *increase* the students' power to solve a wide variety of problems across different disciplines, a different approach to teaching statistical problem-solving must be used. One such approach that has been used in real-world research settings and has been introduced successfully to high school students can be termed a *Prediction Modelling Approach*. The approach greatly reduces the amount of material to be learned. It unifies many seemingly different statistical procedures under one model (e.g. t-tests for comparison of means, F-tests for analysis of variance, simple regression). And, of most importance, this instructional approach, when combined with the power of computers, gives the student exceptional problem-solving capability. The basic ideas have been described by Bottenberg and Ward (1963) and Ward and Jennings (1973).

2. Instructional goals and strategies

(i) The student should give examples of questions that might occur in a wide variety of real-world settings. Get the student involved in identifying an interesting problem.
(ii) The student should be able to describe the problem in terms of prediction. This can be introduced by using "averages" (means or medians) from predictor information to make predictions of numerical outcomes. At some point, the student should begin to view the prediction problem as a model of the form:

Outcomes (Dependent Variable) = Function of Independent Variables + Errors

This form of expression should come easily for students who have been introduced to functions *without* errors.
(iii) The student should be able to describe hypothesis tests as "comparing errors of prediction from two models". The first prediction model, called an Assumed Model, contains more predictor information than the second model, called a Restricted Model. The Restricted Model is obtained by imposing restrictions on the Assumed Model that are implied by the hypothesis of interest. The beginning student should work with problems that require Assumed and Restricted Models that are easy to create and that require computations that can be accomplished without a high-speed computer. These models should involve problems that require only the use of averages for mutually exclusive categories of predictor information. A problem of this type is shown below.
(iv) The student should be able to explain the difference between *rejecting* and *failing to reject* a hypothesis, and that the cut-off probability at which the decision is made is arbitrary. Uncertainty about decisions is the nature of problems in the real world.
(v) After the fundamental ideas are mastered, the student can create and manipulate more complex models that require the use of computer software for solutions. The student should be able to prepare the data for input to a computer program that will perform the required analysis.
(vi) The student must be able to interpret the computer output to answer the questions of interest. If the student has created the appropriate model and specified the computing requirement precisely then the meaning of the output should be clear.

3. An example of a Prediction Model Approach

The following example illustrates a problem that uses prediction models that require only computation of arithmetic averages of data associated with mutually exclusive groups. Assume that we are interested in comparing daily profits of the San Antonio, Texas, Sea World and the Orlando, Florida, Sea World. A random sample of ten days' profits was selected from San Antonio and a random sample of ten days' profits was selected from Orlando.

Question 1: Is there a difference in expected profits between the two Sea World locations? This question disregards any information (e.g. rainfall conditions) that might "contaminate" the conclusions.

Question 2: Is there a difference in expected profits between the two locations, after "removing" any differences that might be attributable to uncontrollable environ-

mental factors such as rainfall, temperature, etc.? We would like to compare the profits of the two locations under the "same" environmental conditions.

Assume that we have the following hypothetical daily profit data:

Location	Rain?	Daily Profit (Thousands of Dollars)
(T) San Antonio, TX	Rain (R)	0, 1, 2
(T) San Antonio, TX	*NO* Rain (N)	8, 9, 10, 11, 12, 8, 12
(F) Orlando, FL	Rain (R)	2, 4, 4, 1, 3, 5, 6, 7
(F) Orlando, FL	*NO* Rain (N)	14, 16

(i) *A standard approach to Question 1:* To compare the average (mean) profit of the two locations, without considering the Rain information, compute the average profit for Texas (M_T) and for Florida (M_F):

$$M_T = \frac{(0+1+2+8+9+10+11+12+8+12)}{10} = \frac{73}{10} = 7.3$$

$$M_F = \frac{(2+4+4+1+3+5+6+7+14+16)}{10} = \frac{62}{10} = 6.2$$

We observe that M_T (7.3) is greater than M_F (6.2) by 1.1 thousand dollars, and ask "Is the difference of 1.1 of practical importance"? If we answer "yes", then we can ask "Could the difference have happened by 'chance' when there is no real difference"? We can explore the "chance" difference in one or more of the following ways: (a) with simple numerical and graphical summaries similar to the box plots shown in Figure 1; (b) with a randomisation test using computer simulations (Barbella, Denby and Landwehr, 1990); or (c) using a t (or F) statistic to test the hypothesis. In this first question we wish to determine how often a difference of 1.1 or greater could have occurred by "chance", when there is no difference.

(ii) *A Prediction Model Approach to Question 1:* This approach provides a general method for investigating not only Questions 1 and 2 above, but also a wide variety of unforeseen and novel questions.

Using this approach for Question 1, we predict the observed profit values using the *two* different averages. We use $M_T = 7.3$ as our prediction when we know an observation comes from Texas and use $M_F = 6.2$ as our prediction when we know an observation comes from Florida. This first prediction model will be called the *Assumed Model.*

Then we will predict the same value for both locations using a single average, M_A, (the average of *all 20* profit values) as our prediction for each one of the 20 profit values. This model will be called the *Restricted Model.* The average of all 20 profit values can be computed as:

$$M_A = \frac{135}{20} = 6.75.$$

Then we can compare the prediction-error sum of squares for the Assumed Model (SSEA) using the *two* averages with the prediction-error sum of squares for the Restricted Model (SSER) using *only one* average. The numerical values are: SSEA = 413.70 and SSER = 419.75. Obviously, if the two sample means were identical then SSER - SSEA = 0, and our decision would be that there is no difference. However, if the two means are far apart then the difference SSER - SSEA becomes large. The evaluation of the difference (SSER - SSEA = 419.75 - 413.70 = 6.05) between the errors of prediction of the two prediction systems forms the basis for deciding if the two Sea Worlds "differ by chance". (The probability associated with the F statistic for these data indicates that we would most likely "fail to reject" the hypothesis that the means are equal.)

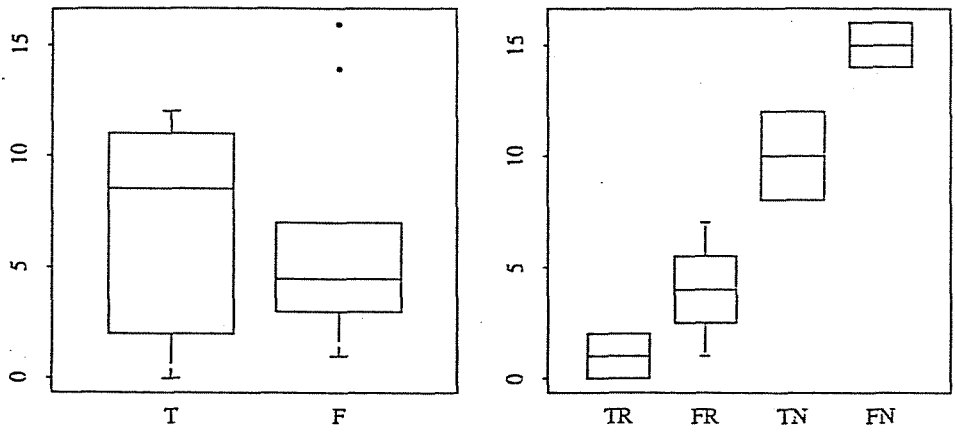

FIGURE 1
Box plots for Sea World data

(iii) *A Prediction Model Approach to Question 2:* To compare averages (means) of the *two* Sea World locations within each of the *two* rainfall conditions, compute the average profit for Texas with Rain (M_{TR}), for Texas with No Rain (M_{TN}), for Florida with Rain (M_{FR}), and for Florida with No Rain (M_{FN}).

$$M_{TR} = \frac{(0+1+2)}{3} = \frac{3}{3} = 1.0$$

$$M_{TN} = \frac{(8+9+10+11+12+8+12)}{7} = \frac{70}{7} = 10.0$$

$$M_{FR} = \frac{(2+4+4+1+3+5+6+7)}{8} = \frac{32}{8} = 4.0$$

$$M_{FN} = \frac{(14+16)}{2} = \frac{30}{2} = 15.0$$

Now we observe from the above (and from the box plots in Figure 1) that the situation is different when we compare the averages within the same rainfall conditions. We have "removed differences due to rainfall conditions".

$$M_{FR}\ (4.0) \text{ is greater than } M_{TR}\ (1.0)$$

and

$$M_{FN}\ (15.0) \text{ is greater than } M_{TN}\ (10.0).$$

We would conclude that the Florida Sea World profit is greater than the Texas Sea World profit when operating under the same rainfall conditions. Similar to the analysis procedure of Question 1, *if the observed differences are of practical importance*, we can compare the prediction errors using the Assumed Model with *four* averages (SSEA = 50) with the prediction errors using the Restricted Model with *two* averages (SSER = 108.525). The evaluation of the difference (SSER - SSEA = 108.525 - 50 = 58.525) forms the basis for comparing the two locations' average daily profits, when "controlling for differences due to rainfall conditions". (The probability associated with the F statistic for these data indicates that we would most likely "reject the hypothesis" that the means are equal under the same rainfall conditions.)

This problem illustrates the possibility that, when we compare the profits of the two Sea Worlds *without controlling for* differences due to the presence or absence of rain, we would come to a different conclusion than if we *control for* differences due to rainfall conditions.

Other interesting examples can be included involving two or more environmental variables, but still *using only models that require the comparisons of averages* of the dependent variable among mutually exclusive categories created from predictor information.

References

Barbella, Peter, Denby, Lorraine and Landwehr, James M (1990) Beyond exploratory data analysis : the randomisation test. *Mathematics Teacher* **83**, 144-149.

Bottenberg, Robert A and Ward, Joe H Jr (1963) *Applied Multiple Linear Regression.* PRL-TDR-63-6, AD-413 128, Personnel Research Laboratory, Lackland AFB, TX.

Foerster, Paul A (1987) *Precalculus with Trigonometry : Functions and Applications.* Addison-Wesley.

Ward, Joe H Jr and Jennings, Earl (1973) *Introduction to Linear Models.* Prentice-Hall, Englewood Cliffs, NJ.