*Math. Model. Nat. Phenom.*
**Vol. 6, No. 6, 2011, pp. 76-95**
DOI: 10.1051/mmnp/20116605

# The $p$ and the Peas: An Intuitive Modeling Approach to Hypothesis Testing

**C. Neuhauser**[1] * **and E. Stanley** [2]

[1] Biomedical Informatics and Computational Biology
University of Minnesota Rochester, Rochester MN†
[2] BioQUEST, Biology Department, Beloit College, Beloit WI†

**Abstract.** We propose a novel approach to introducing hypothesis testing into the biology curriculum. Instead of telling students the hypothesis and what kind of data to collect followed by a rigid recipe of testing the hypothesis with a given test statistic, we ask students to develop a hypothesis and a mathematical model that describes the null hypothesis. Simulation of the model under the null hypothesis allows students to compare their experimental data to what they would expect under the null hypothesis, thus leading to a much more intuitive understanding of hypothesis testing. This approach has been tested both in the classroom and in faculty workshops, and we provide some suggestions for implementations based on our experiences.

**Key words:** data analysis, hypothesis testing, mathematical and biological reasoning, model building, model revision, normal distribution, peas, probability, sampling, seed development
**AMS subject classification:** 62-01, 62F03, 92-01, 97U50

## 1. Introduction

Much time in science labs is spent on developing working hypotheses both to explain why something is happening and to predict what is going to happen. While students have opportunities to learn how to devise experiments to test hypotheses, the analysis of the data resulting from experiments is often given short shrift. Students are provided with a recipe for data analysis and proceed

*Corresponding author. E-mail: neuha001@umn.edu
†Partially supported by Howard Hughes Medical Institute through an HHMI Professor grant to C. Neuhauser.

without considering the nature of the data. National reports, such as Bio2010 or the AAMC-HHMI report *Scientific Foundations for Future Physicians*, have expressed the need for students to gain competency in "[a]pply[ing] quantitative reasoning and appropriate mathematics to describe or explain phenomena in the natural world" [1]. Furthermore, the report explicitly states that students need to be able to "[m]ake statistical inferences from data sets [1].

One of the fundamental tools of making inferences from data is hypothesis testing. Students encounter this tool in their freshman biology lab where a lab manual gives detailed instructions on what type of data should be collected and how to analyze the data. For instance, to test the goodness of fit students are given the formula for calculating the chi-square statistic, determine the degrees of freedom in the experiment, and consult a table that lists the probability that the value or a greater value of the calculated statistic would occur by chance alone. Armed with this probability, students are then ready to draw a conclusion about whether or not to reject the null hypothesis. Without any prior knowledge of hypothesis testing, however, this approach does little to convey to the student the principles behind reaching the conclusion. Instead, hypothesis testing remains a rote experience. Unfortunately, more thorough mathematical introductions to hypothesis testing are relegated to advanced statistics courses and are therefore rarely seen by life science undergraduate students.

The basic components of a statistical hypothesis testing procedure are (1) identification of a null hypothesis and appropriate statistical test, (2) calculation of a test statistic, and (3) calculation of the acceptance and rejection region or calculation of the $p$-value to conclude whether or not to reject the null hypothesis. This systematic approach works well for students who have a solid understanding of the concepts of hypothesis testing, and allows students to "solve the problem in an organized fashion" [5]. However, to students who have not yet mastered the concepts, the process is both passive and not informative. Students are guided to a specific null hypothesis, told what kind of data to collect, given the formula for computing the test statistic, and led through the steps of finding the rejection region or the $p$-value. The problems with this approach are manifold: (1) the appropriate statistical test and test statistic are not motivated; (2) the model assumptions for calculating the distribution of the test statistic under the null hypothesis are not made explicit; and (3) the meaning of the rejection region or the $p$-value remains obscure.

We propose a more intuitive simulation approach that emphasizes the mathematical model that underlies the null hypothesis. Students must consider biological data and how to build a model that satisfies the assumptions expressed in the null hypothesis. Further, students explore their model by running simulations to sample the distribution under the null hypothesis. If the sample results do not fit with their observations (expected results), students must reject the null hypothesis. Students then build a new model requiring further biological and mathematical reasoning.

Undergraduate and high school students with a wide range of biological and mathematical preparedness can engage in this simulation approach since it requires very little prior knowledge beyond some introductory biology, an elementary familiarity with spreadsheets, and some basic

probability. Students with some experience using statistical tests may find this approach meaningful even if traditional assumptions, such as normality, do not hold, and they are asked to come up with their own test statistic. We do assume that students are sufficiently familiar with the scientific process to understand that scientific knowledge is advanced by disproving null hypotheses.

## 2. The Null Hypothesis

To illustrate the process from formulating a working hypothesis to statistical hypothesis testing we chose a topic in plant biology lab where students learn about the development of seeds. Depending on the laboratory experience, students may be asked to come up with a hypothesis or are given the hypothesis. In the following, we will state the hypothesis and use it as a starting point for our discussion of how to move through the process of hypothesis testing. Seed development is a complex process in which much can go wrong. For instance, a seed may fail to develop because the flower was not pollinated. Even if pollination is successful, the location of the ovule in the ovary may play a role in successful fertilization as the distance a pollen tube must travel to reach an ovule varies in some plants. If fertilization is successful, the embryo within the ovule may not receive enough nutrients from the maternal plant to complete development resulting in seed abortion. The delivery of nutrients also depends on the relative distance an ovule is from maternal resources. (See Figure 1.)

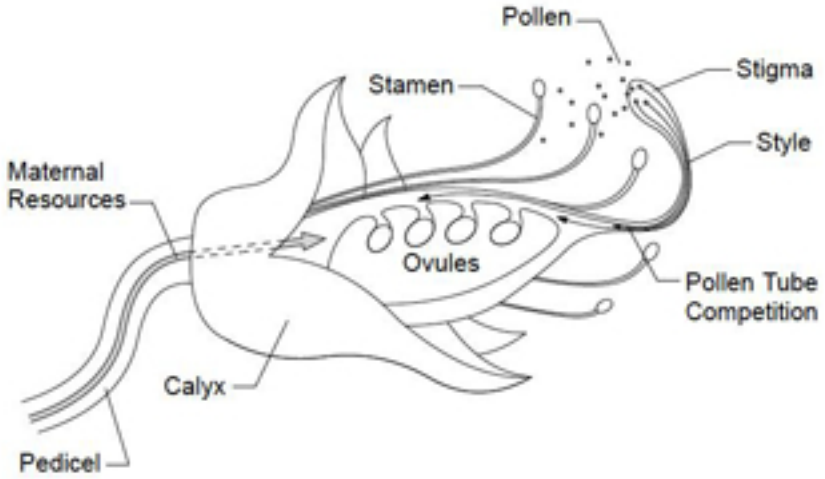


Figure 1: Diagram of a pea flower without petals showing spatial relationship of individual ovules with respect to maternal resources and pollen tubes. (Courtesy of Micah Stanley, 2010)

While there are other factors besides pollen tube and resource competition that can affect individual seed development, such as sibling rivalry and kin selection [2], we focused on the following

working hypothesis:

Working Hypothesis: Because of pollen tube and resource competition, the likelihood of seed development will differ from ovule to ovule depending on the location of the ovule in the ovary of the flower.

Based on this working hypothesis, students may predict that depending on whether pollen tube competition is more/less important than resource competition, ovules that are closer to the calyx have a smaller/larger probability of developing than ovules that are further away from the calyx. On the other hand, if pea development is independent of pollen tube or resource competition or if the two sources of competition cancel each other out, we would expect that there is no difference in the likelihood of developing from ovule to seed based on location of the ovule in the flower. This reasoning leads to the null hypothesis H0. The null hypothesis reflects the "no effect" scenario, and the development of the null hypothesis is an important step. Depending on the level of experience of the students, students could be asked to come up with the null hypothesis independently or the instructor could facilitate a class discussion with the goal of jointly developing the null hypothesis. Whether the student comes up with the null hypothesis independently or in a class discussion, the student needs to be able to articulate that the null hypothesis reflects the "no effect" scenario.

$H_0$: Each ovule, independently of all other ovules in a flower, has the same likelihood of developing into a pea.

The process of testing the null hypothesis involves the following conceptual steps. To

1. describe the development into a pea under the "no effect" scenario to formulate a mathematical model;
2. identify a quantity, the test statistic, which appropriately summarizes the data; and
3. determine the probability distribution of the statistic under the null hypothesis.

In a traditional lab experiment, students are told what data to collect and are then given the "recipe" for analyzing the data to test the null hypothesis, using a standard goodness-of-fit test: They would be given a sample of pea pods and asked to tabulate the number of pods with 0,1,2,... peas. They would then be asked to calculate the expected frequencies for each of the categories according to a formula, and finally to calculate the value of the chi-square statistic together with the degrees of freedom and find the value in a table. While this is an efficient way to complete the lab, the students' role is passive. They have no opportunity to decide what data to collect for building a mathematical model that would become the basis for the hypothesis testing.

# 3. Observation, Data Collection, and Model Building

We continue with the development of a mathematical model that reflects the null hypothesis. It is a pedagogical issue whether to start with observations or with developing the mathematical model.

We believe that it is more intuitive for students to make observations first and then develop the model. This "observation first" approach caters to the observational skills of many biology students. Students first develop a mental picture of pea pods during their observation of the peas in their pods. Then they are better prepared to develop a mathematical model that allows them to more easily abstract from the actual pea pods to simulated pea pods. This strategy engages students in biological thinking as well. After making observation on a sample of pea pods, questions like "how do the peas get into the pods?" are likely to lead to an exploration of reproduction and seed development in this type of plant. This approach would mimic more closely the discovery-based approach in contemporary biology where knowledge discovery follows data collection without first formulating a hypothesis.

For the data collection, we bought snow peas, *Pisum sativum* var. macrocarpon, a commercial variety from a local supermarket. Snow peas are particularly suitable for this study since the peas in a pod can be observed by holding the pod against a light without cutting the pod open. Students observe the peas in the pods and begin to develop an approach to test the null hypothesis. Students frequently find it difficult that there is not just one correct approach. For instance, students could keep track of the number of peas per pod or try to determine which ovules developed into peas by recording the spatial locations of peas in the pods. A discussion on potential strategies can focus on the type of data that would need to be collected, the feasibility of the data collection, and whether the type of data could be used to reject the null hypothesis. Different student groups may pursue different strategies, but whatever the strategy, students will need to simulate the peas in the pods under the null hypothesis to mimic their observations with the actual pea pods.

By observing the peas in the pods, students will need to realize that in order to simulate the peas in the pods under the null hypothesis, they would need to know the maximum number of peas that could develop in a pod and the likelihood of development. Students may need help with this step, and a whole-class discussion can facilitate arriving at this realization. Under the null hypothesis knowing these two numbers would be enough to simulate peas in pods since they could then decide for each potential pea, independently of all other peas, whether or not to turn it into an actual pea. Before building a simulation model, students would collect data to obtain estimates for these two numbers.

The data come from a NUMB3R5 COUNT! workshop for college instructors at NIMBioS at the University of Tennessee in 2009 where participants counted the number of peas per pod in 253 pods [8]. The result is listed in Table 1:

The table lists the data in eleven categories, from 0 peas per pod to 10 peas per pod. The estimate of the two parameters requires calculations where students may need some guided assistance. To calculate the likelihood of a pea to develop, we need to know the ratio of the total number of *actual* peas to the total number of *potential* peas in the sample. The total number of potential peas in the sample is the product of the number of pods and the maximum number of peas per pod. To

| Number of Peas per Pod | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 0 | 0 | 1 | 3 | 7 | 21 | 61 | 104 | 47 | 7 | 2 |

Table 1: Data from observations: The first row lists the eleven categories from the minimum of 0 peas per pod to the maximum of 10 peas per pod. The second row lists the frequencies in each of the eleven categories. For instance, we observed 21 pea pods with exactly 5 peas.

find the total number of pods in the sample, we add up the frequencies in the eleven categories:

$$\text{Total number of pods} = 0 + 0 + 1 + 3 + 7 + 21 + 61 + 104 + 47 + 7 + 2 = 253$$

The total number of pods is 253. Pods can contain a variable number of ovules. This sample had a maximum of ten peas per pod, and we assume that this is the number of potential peas per pod for all the pods in the sample. With 10 being the maximum number of peas per pod, the potential number of peas in the sample is thus (253)(10)=2530.

To calculate the total number of actual peas in the sample, we first calculate the number of peas in each category by multiplying the number of peas per pod by the frequency of the category. Adding up these numbers then gives the total number of actual peas:

$$\begin{aligned}&\text{Total number of acual peas} = \\ &= (0)(0) + (1)(0) + (2)(1) + (3)(3) + (4)(7) + (5)(21) + (6)(61) \\ &+ (7)(104) + (8)(47) + (9)(7) + (10)(2) \\ &= 1697\end{aligned}$$

The total number of actual peas is 1697. The likelihood of a pea to develop is therefore

$$\frac{\text{Total number of actual peas}}{\text{Total number of potenial peas}} = \frac{1907}{2530} = 0.67.$$

Figure 2 shows screen shot of the calculations in an Excel spreadsheet:

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | # of Peas per Pod | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | TOTAL |
| 2 | Frequency | 0 | 0 | 1 | 3 | 7 | 21 | 61 | 104 | 47 | 7 | 2 | 253 |
| 3 | # of Peas per Category | 0 | 0 | 2 | 9 | 28 | 105 | 366 | 728 | 376 | 63 | 20 | 1697 |

Figure 2: Screenshot of an Excel spreadsheet for calculating the parameters of the model. The first two rows [# of Peas per Pod (Row 1) and Frequency (Row 2)] are a repetition of Table 1. In Row 3, we multiply the two numbers in the two cells above to obtain the # of Peas per Category.

To obtain the total number of pods, we enter: "=SUM(B2:L2)" in Cell M2. To obtain the total number of peas, we first calculate the number of peas in each category. These numbers are listed

in Cells B3 to L3 for each category. For instance, to obtain the total number of peas in pods that have exactly 5 peas, we enter "=G1∗G2" in Cell G3. The total sum is then found in Cell M3 where we enter "=SUM(B3:L3)." To find the ratio, we enter into a cell "=1697/2530" or "=M3/(L1∗M2)" (not shown).

To answer the question "What to expect under the null hypothesis?" we build a mathematical model where each ovule, independently of all other ovules in a flower, has the same probability of developing into a pea. Model building is a skill that requires the ability to identify the most pertinent aspects of reality, making simplifying assumptions, and combining the assumptions to come up with a caricature of reality. The caricature captures enough of reality to allow us to address the question we are interested in, which suggests that the aspects of reality that are important may depend on the question we ask. The goal is to build a model that is simple, yet captures enough of reality to allow us to model pea pods under the null hypothesis. Depending on how experienced students are with model building, students may need a fair amount of assistance, and so this step may be better done in a whole-class discussion instead of each student or each student group trying to come up with a model.

Recall the assumptions of the null hypothesis: we assume that all pods have the same properties, or are identical, each has the same maximum number of potential peas, and each ovule has the same probability, independently of all other ovules, to develop into an actual pea. We formulate our first assumption for the model based on the observation from our pea pod sample:

(A1) Each pod has a maximum of 10 potential peas.

The second assumption states the independence of peas in a pod:

(A2) Each ovule, independently of all other ovules, has the same likelihood of developing into a pea.

We use the fraction of peas per pod as an estimate for the probability that an ovule develops into a pea. The data give the estimate 0.67, which is our third assumption:

(A3) The probability of a pea to develop is 0.67.

We can now build a model: Given that an ovule has a 67% likelihood of developing into a pea, we can imagine that each ovule, independently of all other ovules, tosses a coin with probability 0.67 of Heads and turns into a pea if Heads come up. We can group ten ovules together and call it a pod. This thought experiment can be implemented into a spreadsheet.

# 4. Simulation

Coin tossing in an Excel spreadsheet can be implemented using the RAND() function. This function generates a uniformly distributed random variable over the interval between 0 and 1. We can use this to simulate pea development. If the random variable is less than 0.67, we call the outcome of the coin toss Heads, if it is greater than 0.67, we call the outcome Tails. If Heads result, the pea develops; if Tails result, the pea does not develop. With 67% of the coin tosses resulting in Heads, 67% of the ovules develop into peas. To implement this in an Excel spreadsheet, we type the function command "=IF(RAND()<0.67,1,0)" into a cell. This command will produce a "1" 67% of the time and a "0" the remaining time. A "1" thus represents the presence of a pea; a "0" the absence of a pea. Figure 3 shows an Excel screen shot for simulating peas in their pods.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ovule | | | | | | | | | | Total # of Peas |
| 1 | Pea ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | per Pod |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 9 |
| 3 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 7 |
| 4 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 9 |
| 5 | 4 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 6 |
| 6 | 5 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 7 |

Figure 3: Screenshot of an Excel spreadsheet for simulating peas in their pods. Each of the columns B-K represents a different ovule. A "1" indicates that the ovule developed into a pea; a "0" indicates that the ovule did not develop. In Column L, we sum up the number of 1s. This sum represents the number of peas in the pod represented in the respective row.

Rows 2-6 represent pea pods. The ID number of the pea pod is given in Column A. Columns B-K represent the ten ovules of each of the pods, numbered 1-10 in the first row. The entry "1" in Cell B2 was generated by the command "=IF(RAND()<0.67,1,0)" and the "1" in that cell tells us that the first ovule of the first pea pod developed into a pea. Repeating the command "=IF(RAND()<0.67,1,0)" in each of the cells in the array B2:K6 then results in the array of 0s and 1s. In Column L, the total number of peas in each pea pod is listed, for instance, the first pea pod has 9 peas. To obtain this number, we enter the command "=SUM(B2:K2)" in Cell L2. The other sum totals are obtained similarly by replacing B2:K2 with the appropriate range of cells. Instead of just five pea pods as in the table, we can do this for 253 pea pods, which then represent a computer-generated replicate of the experiment.

To produce a similar table as in Table 1, we need to count the number of times a 0, 1, 2, etc. appears in the sum total of the 253 pea pods (Column L). Excel has a command that does just this: "=COUNTIF(*range*,"= *value*")" where *range* is the range of cells where the sum totals are listed and *value* is the number of peas per pod that we want to count. Table 2 lists the outcome of one such simulation with 253 pea pods.

| # peas/pod | Simulation |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | 3 |
| 4 | 18 |
| 5 | 30 |
| 6 | 51 |
| 7 | 73 |
| 8 | 49 |
| 9 | 24 |
| 10 | 5 |
| | 253 |

Table 2: Outcome of one run of the pea pod simulation. We simulated 253 pea pods. The first column lists the categories from the minimum of 0 peas per pod to the maximum of 10 peas per pod. The second column lists the frequency of each category. For instance, in 18 of the 253 simulated pea pods, we saw pea pods with exactly 4 peas.

Students can repeat the simulation by hitting the F9 key, which recalculates the cells in the spreadsheet, thus resulting in a new run. Students will observe that the outcomes differ from run to run, which is important for their understanding of a "sample." Each run represents a simulated sample. Students can then make the connection that the data they collected on the actual pea pods are just another sample and that they would get a different result if they repeated the experiment with a different batch of pea pods.

# 5. Data Analysis

After students have gained some familiarity with the variation observed in the simulation, they can begin to ask the question of how they can quantify what to expect under the null hypothesis. This question combines the second and third step of the process of testing the null hypothesis, and is perhaps the most difficult part for students. Designing a test statistic involves ambiguity and may result in more than one strategy. This ambiguity, however, can lead to a rich discussion and deepen students' understanding of the process of hypothesis testing. Students should also not be prevented from false starts: They may, for instance, initially attempt to come up with a measure of the difference between the observed and simulated data in each category only to realize that this measure would not allow them to conclude whether the observed data are consistent with the null hypothesis.

Students should be encouraged to pursue different strategies. We explain two such strategies.

The first strategy uses a measure of dispersion, namely the sample variance, to compare the sample variance of the observed data to that of the simulated data. The second strategy explores ways to measure the difference between the simulated data and the theoretical expectation in each category of pea counts. The latter strategy will eventually lead to the chi-square statistic. In each case, we will simulate repeatedly the test statistic under the null hypothesis to obtain the distribution of the test statistic. This will then allow us to determine whether or not the value of the test statistic based on the actual data is consistent with the null hypothesis.

*Strategy 1*

For each simulation run, the sample variance can be computed according to the following formula:

$$S^2 = \frac{1}{N-1}\left(\sum_{j=1}^{k} x_j^2 f_j - \frac{1}{N}\left(\sum_{j=1}^{k} x_j f_j\right)^2\right) \tag{5.1}$$

where $N$ is the number of pods, $k$ is the number of categories, $x_j, j = 1, 2, \ldots, k$ are the distinct values in each category, and $f_j, j = 1, 2, \ldots, k$ are the corresponding frequencies. Table 3 lists an example of a simulation together with required calculations.

| $x_j$ | $f_j$ | $x_j f_j$ | $x_j^2 f_j$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 1 | 3 | 9 |
| 4 | 12 | 48 | 192 |
| 5 | 31 | 155 | 775 |
| 6 | 49 | 294 | 1764 |
| 7 | 77 | 539 | 3773 |
| 8 | 48 | 384 | 3072 |
| 9 | 30 | 270 | 2430 |
| 10 | 5 | 50 | 500 |
| Sums: | 253 | 1,743 | 12,515 |

Table 3: Calculating the sample variance of simulated peas: This is a simulation of 253 pea pods. The first column lists the eleven categories $x_j, j = 1, 2, \ldots, 11$, from the minimum of 0 peas per pod (category $x_1$) to the maximum of 10 peas per pod (category $x_{11}$). The second column lists the frequency of each category. In the third and fourth column, we calculate terms for (5.1). In the last row, we calculate the sums in each of columns 2-4.

We find

$$S^2 = \frac{1}{N-1}\left(\sum_{j=1}^{k} x_j^2 f_j - \frac{1}{N}\left(\sum_{j=1}^{k} x_j f_j\right)^2\right) = \frac{1}{252}\left(12,515 - \frac{(1,743)^2}{253}\right) = 2.012$$

To compare the sample variance of the observed data to that of the simulated data, we need to calculate the sample variance of the observed data first.

| $x_j$ | $f_j$ | $x_j f_j$ | $x_j^2 f_j$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 2 | 1 | 2 | 4 |
| 3 | 3 | 9 | 27 |
| 4 | 7 | 28 | 112 |
| 5 | 21 | 105 | 525 |
| 6 | 61 | 366 | 2196 |
| 7 | 104 | 728 | 5096 |
| 8 | 47 | 376 | 3008 |
| 9 | 7 | 63 | 567 |
| 10 | 2 | 20 | 200 |
| Sums: | 253 | 1,697 | 11,735 |

Table 4: Calculating the sample variance of observed pea pods. This is the data from the experiment of counting peas in 253 pea pods. The first column lists the categories $x_j, j = 1, 2, \ldots, 11$, from the minimum of 0 peas per pod (category $x_1$) to the maximum of 10 peas per pod (category $x_{11}$). The second column lists the frequency of each category. In the third and fourth column, we calculate terms for (5.1). In the last row, we calculate the sums in each of columns 2-4.

We find

$$S^2 = \frac{1}{N-1}\left(\sum_{j=1}^{k} x_j^2 f_j - \frac{1}{N}\left(\sum_{j=1}^{k} x_j f_j\right)^2\right) = \frac{1}{252}\left(11,735 - \frac{(1,697)^2}{253}\right) = 1.398$$

This value is quite a bit smaller than the one we found when we simulated the sample variance under the null hypothesis. Of course, one simulation is not enough, and realizing this is at the heart of this approach: by repeating the simulation many times, we get a sense for the distribution of the sample variance and can decide whether the observed value is "typical" or "unusual." When we simulated the sample variance under the null hypothesis 1000 times in Excel, we found that the sample variance ranged from 1.611 to 2.825. (These repeated runs are most easily done using a macro in Excel. See below for a comment on using macros in Excel.) Students will notice that

the observed value is much smaller than the simulated ones. Furthermore, they can see from the simulations that values as small or smaller than the observed sample variance seem to occur in less than 1 in 1000 simulations. These two observations are key insights for students to conclude ultimately that the observed value is far from typical. Time needs to be spent in class on discussing what these insights tell us. We found it particularly useful to have a whole-class discussion at this point.

A whole-class discussion should also include talking about one-sided versus two-sided tests without necessarily defining these technical terms. If there is an *a priori* reason to expect a smaller variance, a one-sided test would be appropriate, and the null hypothesis should be rejected if the observed sample variance is too small. How small is too small naturally leads to the definition of the *rejection region* for a given *Type 1 error* without mentioning either technical term in the discussion: Students may decide that if the value of the test statistic of the observed data falls in the bottom 5%, they would reject the null hypothesis since they would observe this or a smaller value in only 5% of the samples that come from the null hypothesis. This number in our simulation turned out to be 1.927. That is, if the observed sample variance is less than 1.927, we would reject the null hypothesis at the 5% significance level.

Students discuss whether they should choose 5% or 1% or some other value for the significance level. They realize that this percentage value is the likelihood of rejecting the null hypothesis even though the null hypothesis is true. The possibility of making this type of error can be somewhat unsettling to students but can also result in a rich discussion that can extend into other areas, such as the medical field, where false positive results are of concern.

*Strategy 2*

The goal of the second strategy is to find a quantity that can compare the expected and the observed distribution of pea counts. The expected distribution of the number of peas per pod can either be determined empirically or theoretically. An empirical determination of the expectation involves simulating pea counts multiple times and keeping track of the frequency of each count category. Averaging the simulated frequencies for each category then results in an empirical distribution that is close to the expected distribution if the number of simulations is sufficiently large. To obtain the expected distribution theoretically requires a higher level of mathematical knowledge. Namely, the theoretical distribution of the number of peas per pod under the null hypothesis follows a binomial distribution with $n$ trials, where $n$ is the number of potential peas per pod, and success probability $p$. Based on our observed pea pods, we estimated that $n$ is equal to 10 and $p$ is equal to 0.67.

The expected number of pods with $k$ peas,$k = 0, 1, \ldots, 10$, can be calculated using the follow-

ing formula:

$$\begin{aligned}\text{E(\# Pods with } k \text{ peas)} &= (\text{\# Pods})P(\text{Pod has } k \text{ peas}) \\ &= 253 \times \binom{10}{k} (0.67)^k (1-0.67)^{10-k}.\end{aligned}$$

Excel has a function that calculates the probability of the event that there are $k$ successes in $n$ trials: "=BINOMDIST(*number of successes, number of trials, success probability,* FALSE)." The "FALSE" at the end gives the probability of the event. If it is replaced by "TRUE," the cumulative probability would be calculated. For instance, in Cell E2 in Figure 4, we list the probability of 3 successes in 10 trials where the success probability is 0.67. To calculate this probability, we would enter

"=BINOMDIST(3,10,0.67,FALSE)"

into Cell E2 in the spreadsheet. Similar calculations were done in all cells in the second row in Figure 4. Multiplying this number by 253 results in the expected number of pods with exactly three peas, namely 3.8916, which is entered in Cell E3.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | # Peas/Pod | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | Probabilty | 2E-05 | 0.0003 | 0.0028 | 0.0154 | 0.0547 | 0.1332 | 0.2253 | 0.2614 | 0.199 | 0.0898 | 0.0182 |
| 3 | Expected # Peas | 0.0039 | 0.0787 | 0.7188 | 3.8916 | 13.827 | 33.687 | 56.996 | 66.125 | 50.345 | 22.715 | 4.6118 |

Figure 4: Screenshot of an Excel spreadsheet for the expected distribution of the number of peas in a pod under the null hypothesis. Row 1 lists the eleven categories for the number of peas per pod from the minimum of 0 peas per pod to the maximum of 10 peas per pod. In the second row, we take advantage of a built-in function to calculate the probability of having $k$ successes in $n$ trials (see text for details). Finally, by multiplying the number of peas per pod by the probability of this event, we arrive at the numbers in the third row, namely the expected number of peas per pod.

Students can plot histograms of the simulated data and the expected values in a single plot and compare the result (Figure 5).

Repeating the simulation by pressing the F9 key shows the variation of the simulated data from run to run and will motivate how to measure the difference between the simulated data and the theoretical expectation. Students may measure the difference by adding up the absolute values of the differences between the simulated value and the expected value of each bin, that is,

$$\sum_{k=0}^{10} |Sim_k - Exp_k| = \sum_{k=0}^{10} |\text{\# (simulated pods with } k \text{ peas)} - 253 \times P(k \text{ peas in a pod})|. \quad (5.2)$$

Note that the summation index $k$ runs from 0 to 10 instead of 1 to 11 as previously. This suggested statistic is a good starting point for motivating the chi-square statistic, which is quite similar in

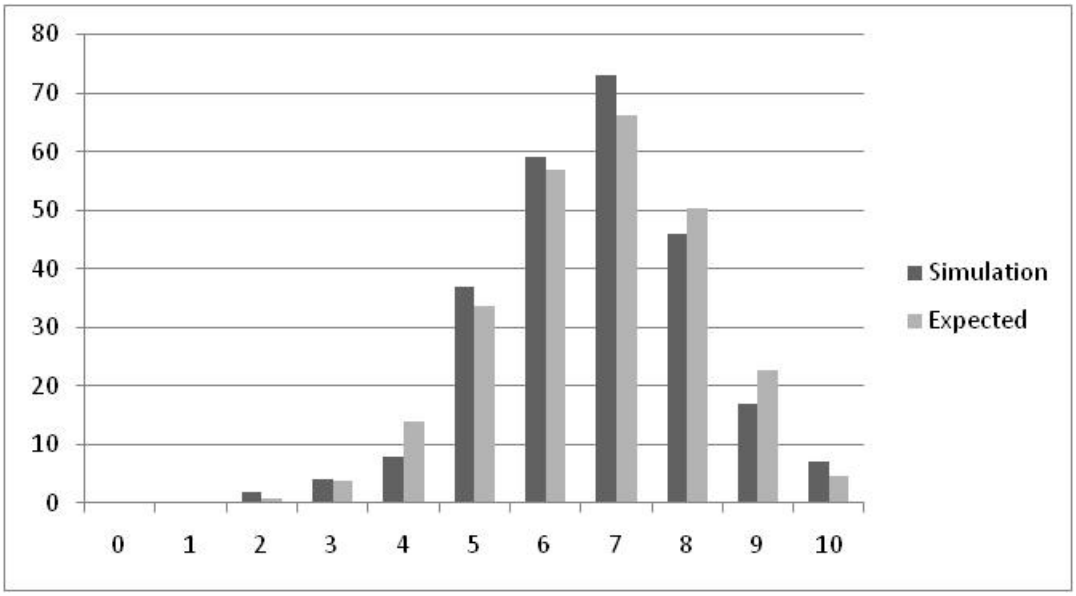


Figure 5: Comparing simulations and expectations for pea counts: The simulation values change from run to run. They are compared against the expected values listed in Row 3 in the screenshot in Figure 4.

spirit (we use $Sim_k$ in the formula instead of the more commonly used $Obs_k$ to indicate that the observed values come from the simulation):

$$\chi^2 = \sum_{k=0}^{10} \frac{(Sim_k - Exp_k)^2}{Exp_k}. \tag{5.3}$$

In essence, the larger this value, the more the observations must have differed from the expectation. This suggests that we reject the null hypothesis if the chi-square value is too large. But how large is "too large?"

We can again use simulations to answer this question. Calculating the chi-square value for a large number of simulations, gives the range of values that may occur. We simulated chi-square values 1000 times in Excel using a macro and recorded their values. Sorting the values in ascending order shows the range of values and allows for easy identification of percentiles. Selected percentiles are listed in Table 5, indicating, for instance, that 95% of the simulated chi-square values are less than or equal to 19.19, or, equivalently, 5% of the simulated values are above 19.19.

It is again natural to define a *rejection region* for a given *Type 1 error* without mentioning either technical term in the discussion: Students may conclude that if the chi-square value of the experimental data exceeds 19.19, they would reject the null hypothesis since they would observe this or a larger value in only 5% of the samples that come from the null hypothesis. Or, if students wanted to be more conservative, they might choose 24.49 as the threshold value, which would

| Percentile | Value |
|---|---|
| 0.0% | 1.13 |
| 25.0% | 5.26 |
| 50.0% | 7.82 |
| 75.0% | 11.40 |
| 95.0% | 19.19 |
| 99.0% | 24.49 |
| 100.0% | 53.26 |

Table 5: Simulated percentiles of the chi-square distribution.

mean that they would only reject the null hypothesis in 1% of samples that come from the null hypothesis. Another approach would be to calculate the chi-square value from the observed data and determine what fraction of samples would have this or a larger value if the samples came from the null hypothesis.

Calculating the chi-square statistic for the experimental data, we find $\chi^2 = 43.1$. Since 43.1 is greater than 19.9, we conclude that we reject the null hypothesis at the 5% level. In fact, we only observed the value 43.1 or larger once in 1000 simulations, which is an indication that it is quite rare to see a value that exceeds 43.1. While the number of runs of the simulation is not sufficient to bracket a $p$-value, students can discuss the concept of a $p$-value. Since the discussion is based on values that were obtained from simulating the null hypothesis, it becomes clearer in a student's mind that the $p$-value is the probability of rejecting the null hypothesis if, in fact, it is true. A student can then learn about Type 1 error and has a way to obtain the Type 1 error trough simulations. This approach emphasizes that the Type 1 error depends on the null hypothesis.

# 6. Using Excel for Simulations

We have had years of experience in using Excel in the classroom and found that students quickly learn how to navigate this tool. Students like the almost instantaneous response time of Excel. However, Excel is not without problems; in particular, it is known that the random generator that was implemented in Excel prior to 2003 did not pass standard tests, which is an issue when a user calls a large number of random numbers. The random number generator that is implemented in Excel 2003 or later versions is of considerably better quality. In order to test the quality of the Excel simulations that resulted in Table 5, we performed simulations in Matlab and also compared both simulations to the theoretical distribution of the chi-square statistic.

For the Matlab simulations, we obtained percentiles (75%, 80%, 90%, 95%, and 99%) based on 1000 runs, similarly to the Excel simulations. We repeated the runs 50 times and averaged the percentile scores and calculated the standard errors. The averaged scores for each of the percentiles

and the corresponding standard errors are displayed in Figure 6 as "Mean (Matlab)." We compare the Matlab simulations to both the Excel simulation (shown in Figure 6 as Excel) and to tabulated chi-square values with ten degrees of freedom (shown in Figure 6 as "chi-square").

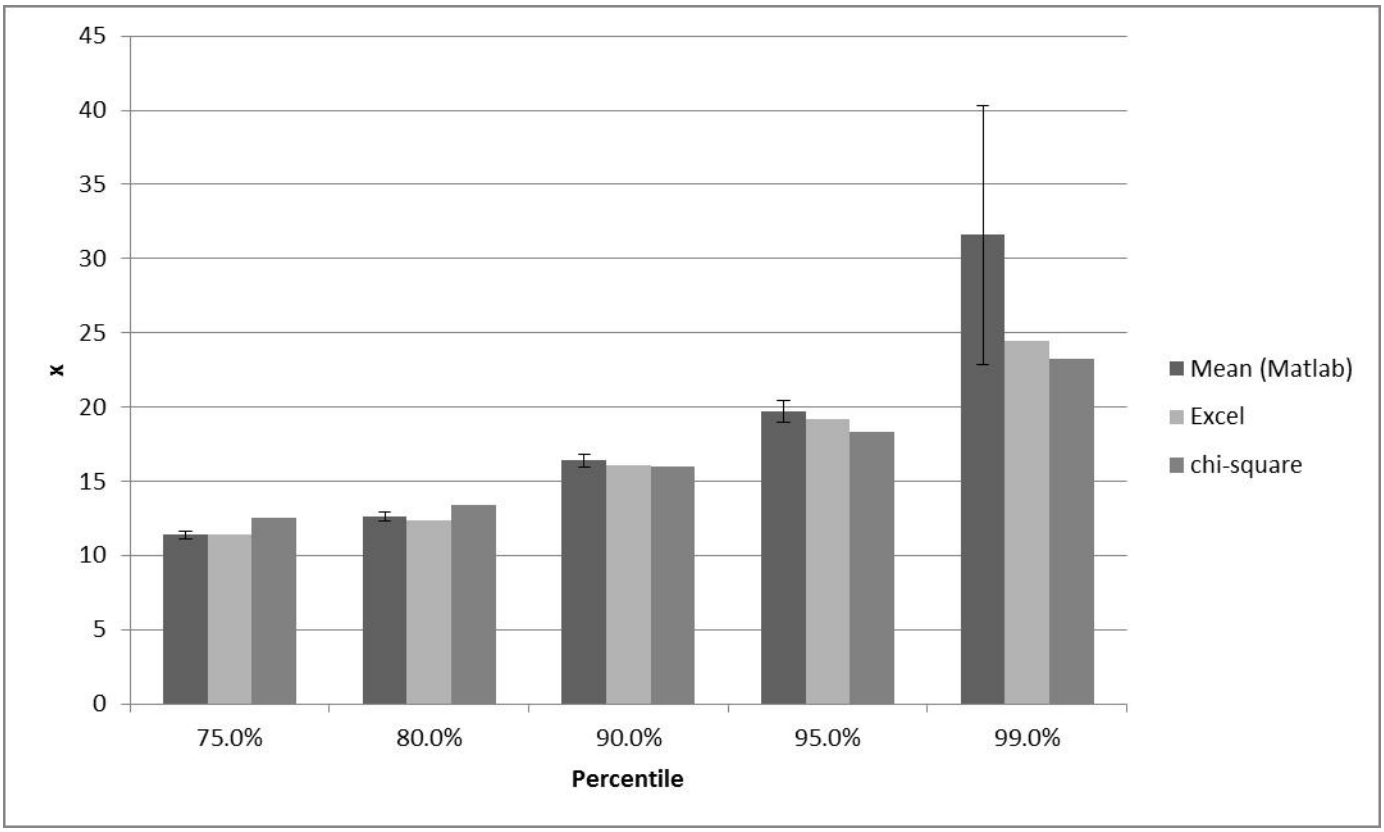


Figure 6: Comparing Matlab simulations to the Excel simulation and the theoretical chi-square percentiles.

First, we observe that the Matlab and Excel simulations are in quite good agreement. Second, likely because the expected frequencies in some of the categories are well below the recommended size, the chi-square approximation appears not to be as close, which is reflected in the difference in values. The 99th percentile score for the Matlab simulation exhibits significant variation due to the relatively small number of runs (1000). We conclude that the Excel simulation is quite good when using 1000 runs and a significance value of 5%. In fact, it is closer to the Matlab simulation than the theoretically calculated cut-off based on a chi-square distribution with 10 degrees of freedom.

The simulations in Excel are not very fast, which limits the number of runs that can be done in a class period. We found that 1000 runs are doable when students use a macro. Based on the Matlab simulations and chi-square distribution comparisons, we conclude that the quality of the Excel simulation is sufficiently good for classroom purposes.

# 7. Using Macros in the Classroom

Macros in Excel are a way to automate steps that need to be repeated many times. During the recording of a macro, Excel keeps tracks of all the key strokes. After recording the macro it can be called up and the key strokes are repeated in the order they were recorded. One way to introduce students to macros is to have them repeat the same steps multiple times until they realize the repetitiveness of the steps. At this point, they are ready to start using macros. In our example where we need to run a large number of simulations, we have students perform each run manually and record the outcome in a separate column on the same sheet. To use a macro, it is important that the simulation is set up so that students repeat the exact same steps in each run, which includes entering of simulation outcome into the same cell in the spreadsheet. Excel allows for moving cells down by selecting the cell one wishes to move down and then clicking on "Insert Cells," which can be found on the "Home" tab in the "Cells" group by clicking the arrow next to "Insert." Moving the cells down opens the same cell for entering the next outcome.

When students work in groups of 3–4 students, it is possible to avoid macros all together. We have had students enter the outcomes of simulation runs manually for up to about 100 runs. This is reasonably fast and not too tedious. If this method is used repeatedly throughout a course, it is worthwhile, however, to teach the use of macros. We have also found that student become interested in looking at the virtual basic code that is written during the recording and is accessible on the "Developer" tab in the "Code" group by clicking on "Macros" and then on "Edit" after highlighting the macro that one wishes to look at.

We often teach macros to the entire class by asking students to follow along and repeat the steps on their computers as we present the macro. Since these macros are short, we repeat the steps until most of the students successfully implement the macros. We leave some time to work individually with the few remaining students who experienced problems during the recording.

# 8. Future Directions

The lab reached a clear conclusion, namely, the null hypothesis can be rejected at the 5% or 1% level, and, in fact, the $p$-value appears to be much smaller than 1%. Students should now be sufficiently confident that at least one of the three assumptions (A1)-(A3) in the null hypothesis is violated, but also realize that the conclusion may be wrong. This opens new questions that students can explore. We suggest two areas of exploration: (1) the role of the spatial position of the ovule, and (2) more advanced statistical concepts, such as Type 2 error and properties of statistics.

Positional data and spatial model

During the process of counting peas, students may have observed that peas closer to the stigma were more often absent than peas closer to the calyx. The non-randomness of the distribution of

peas within the pod can lead to a discussion in which alternative explanations such as pollen tube competition [4] and maternal resource competition [3] could be tested. The predicted outcome of this type of competition is illustrated in Figure 7 where we assume that each pod has four peas and make predictions according to the four groups as indicated in the figure. When pollen competition and resource competition are absent, all four ovules would develop into peas after fertilization. If pollen competition is more important than resource competition, the seeds closer to the stigma have a higher chance to develop than the seeds closer to the calyx. The reverse would be the case if resource competition is more important than pollen competition. If both types of competition are important, no peas may develop.

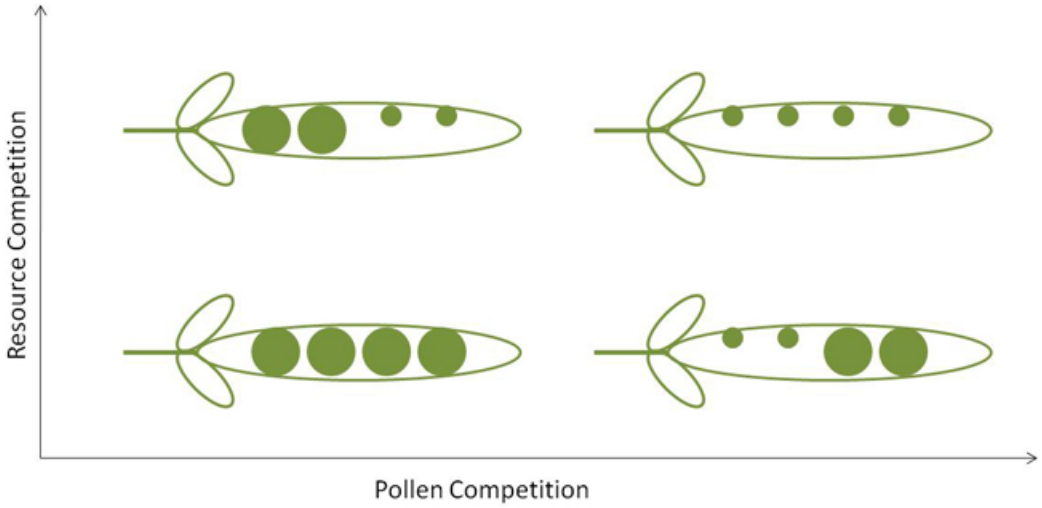


Figure 7: Position of peas within the pod under varying competition conditions: The calyx is on the left side of each of the pea pods. The small circles indicate undeveloped ovules; the large circles indicate ovules that developed into peas.

The same pea pods can be used to collect positional data where students keep track of which of the ovules developed into peas. The data set resembles the simulated data of 0s and 1s (see Figure 3). A probability of presence can be estimated for each of the ten positions separately, and students can develop a more complex model of seed development. Students, for instance, may group the ovule positions into two or three groups, where ovules within each group, independent of all other ovules, have the same probability of developing into a seed. The probabilities could vary among the groups. This model can be implemented in Excel and parameterized with estimates from the positional data. Students can simulate different scenarios and get a sense of what to expect under alternative hypotheses.

Investigating properties of statistic beyond Type I error

We suggest two projects. The first project could investigate the properties of the statistic sug-

gested in (5.1)

$$S^2 = \frac{1}{N-1}\left(\sum_{j=1}^{k} x_j^2 f_j - \frac{1}{N}\left(\sum_{j=1}^{k} x_j f_j\right)^2\right)$$

and in (5.2):

$$\sum_{k=0}^{10} |Sim_k - Exp_k| = \sum_{k=0}^{10} |\#\text{ (simulated pods with } k \text{ peas)} - 253 \times P(k\text{peas in a pod})|.$$

For either statistic, simulating the statistic under the null hypothesis would establish percentiles and students could compare this statistic with the chi-square statistic to see whether the rejection regions are similar. This could lead into further explorations on the assumptions of the chi-square distribution, and furthermore on what properties of a statistic are desirable.

The second project would utilize the positional model discussed in the first project. A model that takes the spatial position into account can lead to a discussion about what difference in probabilities of spatial location can be detected. The positional model would serve as the alternative hypothesis, and students could start exploring the *power of the test* and the *Type 2 error*.

# 9. Conclusion

Data analysis is a critical skill that students in the life sciences need to acquire, and statistical hypothesis testing will remain a mainstay of life scientists. Because hypothesis testing lends itself to an algorithmic approach where students follow a flow chart to select the most appropriate test and then follow a sequence of systematic steps to reach the conclusion of whether or not to reject the null hypothesis, most students lose the opportunity to gain a conceptual understanding of statistical hypothesis testing. The conceptual understanding is required when students encounter non-traditional data whose analyses do not fall into the standard categories that are taught in most introductory statistics courses. Non-traditional data are playing an increasingly important role in all areas of life and health sciences, in particular with the advent of "omics" data.

Every statistical hypothesis test relies on a mathematical model. Yet, model assumptions are often implicit in data analysis and as a result, may not be fully appreciated. The suggested modeling approach to hypothesis testing emphasizes the mathematical model that underlies the null hypothesis. Students begin to appreciate the value of simplified mathematical models as a tool to elucidate biological phenomena. Furthermore, during model building, students must make all assumptions explicit. Explicit assumptions facilitate further study when the null hypothesis is rejected since students can address each assumption separately with new experiments. This leads to open-ended discussions and encourages students to think deeply and creatively about how to design biological experiments and test their hypotheses.

With the suggested simulation approach to hypothesis testing, students will gain greater facility with mathematical models and the process of using models to analyze complex biological situations. We presented this approach in a specific context, namely developing and testing a hypothesis on pea development. This method can be adapted to other applications and other statistical tests. With the repeated use of this method in different applications, mathematical modeling will become an integral part of designing biological experiments to test hypotheses and bring us a step closer to realizing the goal of students being able to "[m]ake statistical inferences from data sets" [1].

# References

[1] AAMC-HHMI. *Scientific foundations for future physicians*. (2009). http://www.hhmi.org/grants/pdf/08-209_AAMC-HHMI_report.pdf [accessed 11 September 2010]

[2] M. Banuelos, J. Obeso. *Maternal provisioning, sibling rivalry and seed mass variability in the dioecious shrub Rhamnus alpinis*. Evolutionary Ecology, (2003), No. 17, 19–31.

[3] K. Bawa, C. Webb. *Flower, fruit and seed abortion in tropical rainforest trees: Implications for the evolution of paternal and maternal reproductive patterns*. American Journal of Botany, (1984), No. 71, 736–751.

[4] D. Buckley, M. Cohen. *Developmental selection: Pollen tube competition and seed abortion*. In J. Jungck and V. Vaughan, (Eds.). The BioQUEST Library VI. Academic Press, San Diego, 2001.

[5] L. Kamin. *Using a five-step procedure for inferential statistical analyses*. The American Biology Teacher, 72 (2010) No. 3, 186–188.

[6] Microsoft Knowledgebase Article ID 828795. *Description of the RAND function in Excel. Revision 6.0*. (2010). http://support.microsoft.com/kb/828795 [accessed 6 September 2010]

[7] National Research Council of the National Academies. BIO2010: Transforming undergraduate education for future research biologists. National Academies of Science, Washington, DC., 2003.

[8] NUMB3R5 COUNT! Workshop (2009). http://www.bioquest.org/NumbersCount/utk_2009/resources.php [accessed 24 September 2010]