**Project ECSITE Activity:** Frequentist Statistics Made Easy, from First Principles

This activity lets students build the chi-squared test from scratch in Excel. In so doing, they will develop a more mechanistic understanding of frequentist statistics, while feeling empowered to tackle challenging scientific questions computationally with simulations.

**Content Areas:** data analysis & probability, reasoning & proof

**Associated Unit:** N/A

**Activity Dependencies:** N/A

**Grade Level:** 9 – 12 (best in a science or statistics class)

**Time Required:** 100 minutes (though this can be less by having students complete the worksheet of questions at home rather than in a class, possibly group-based, discussion).

**Group Size:** 2-5 students. Ideally no student will work alone, and groups of much more than 5 are too large for one computer to be used and seen by everyone.

**Expendable Cost per Group:** $0 (though each group of students will need a computer with Microsoft Excel installed).

**Key Academic Vocabulary:** chi-squared test, distribution, p-value, simulation, statistics

**Vocabulary Definitions:**

*Chi-squared test:* A statistical test for comparing observations to theoretical expectations.

*Distribution:* A function or method of assigning a probability of occurrence to all observable outcomes.

*p-value:* The probability of obtaining the collected/observed data, or more extreme data, assuming the null hypothesis is true.

*Simulation:* An imitation of some real set of occurrences.

*Statistics:* The practice of analyzing data so as to infer underlying processes.

**Pre-Requisite Knowledge:** Completion of this activity in a timely manner requires only a basic knowledge of Microsoft Excel for students. Teachers, however, must have a more in depth understanding of how functions in Excel work, and an intuition into frequentist statistics (which can be acquired by going through this activity before doing so with students).

**Learning Objectives:** This activity will help students improve in two important areas: 1) Students will understand how frequentist statistics work, and how they allow inference from raw data. 2) Students will gain an appreciation for, and ideally a limited ability to use, simulations to tackle meaningful scientific questions.

**Computational Thinking Connection:** This activity teaches students the power of computing simulations to explore data analysis in the context of probability and frequent statistics. This is a
skill with broader applications, as it emphasizes the power of modern computation in problem solving in general. It also helps students improve their computer skills, primarily in Excel.

**Materials List:**

*Per Group:* 1 computer with Excel (part of Microsoft Office) installed

*Per Class:* 1 computer with Excel (part of Microsoft Office) installed and a projector, so the students can follow along with the teacher

**Introduction/Motivation:** This lesson begins best by having one student roll a die 10 times, and another record the outcome on the board. Then, after the 10 rolls, ask the students if they think the die is rigged or not. Why do they think that? Can they ever know for certain? What does it mean to know things about the world based solely on data? Hopefully the students decide that the only way to know if the die is fair or not is to roll is many times (infinitely many to be actually certain). How many times is enough? Were those 10 rolls enough? 100? 1,000,000? Students quickly realize that it seems impossible to know with a reasonable number of rolls, but using a computer we can roll a truly fair “die” and see how often we see what we did on those 10 rolls, and the likelihood of the outcome we did see gives some sense about how likely it is the die is fair. This is the power of computational thinking! By the end of the lesson students will feel empowered to tackle questions about the processes that generate data using simulations.

**Procedure:**

*Before the Activity/Setup*

1. Teachers should only need to do the lesson themselves beforehand to get familiar with the functions in Excel, and the general intuition behind the activity. Above all else, students should see the value of thinking about data probabilistically and computationally, so knowing these aspects of the lesson are most important for teachers to teach it successfully.

*With the Students*

1. Pick the sample size (10 works well to compare to the motivating 10 die rolls the students did at the start of the lesson).
2. Generate truly random outcomes for rolling a die the sample size number of times using Excel’s function ‘randbetween(1,6)’.
3. Repeat step 2 as many times as the students want.
4. Count the number of times each possible outcome (1 through 6) occurred in each experiment (each time you do step 2 is a single experiment) using the ‘countif()’ function.
5. Subtract step 4 from the expected number of times you should see each possible outcome (sample size/# outcomes = sample size/6).
6. Sum all of the differences from step 5. This should result in having one number (usually not an integer) for each experiment.
7. Finally, create a histogram of all possible outcomes using ‘frequency’.
8. The outcome

*Lesson Plan*

*Day 1*
Minutes 0 – 5: Go over two main objectives:
- Begin to think computationally
- Understand where p-values come from, and how frequentist statistics work

Minutes 6 – 15: Talk conceptually about p-values
- Ask “Why do we use statistics?” and “What is a p-value?”
  - Get away, at first, from anything the students know and emphasize a probabilistic intuition (which will turn out to be correct!)
  - Don’t do much correcting at this point, but rather just make a list of ideas the students have about statistics
- (Optional) Very briefly show Pearson’s original derivation
  - Use as motivation for “there must be a simpler way to understand this stuff!”
  - Natural history of any system matters
    - There was a time before these tests existed that people asked the very same questions!

Minutes 16 – 25: Coin experiment
- Have one volunteer flip a coin ten times while another volunteer records the results on the board
- Pose to the class the question “How would you determine if this coin is unusual?”
  - Talk about their ideas
  - Emphasize the value of computational thinking and simulations

Minutes 26 – 45: Generate chi-squared distribution as a class and determine what the p-value is for our coin-flipping experiment
- This should be done in groups, following the teacher’s lead using a projector

Minutes 46 – 50: Wrap up as a class
- Did everyone get the same p-value?
- Does the way the chi-squared test works make sense to everyone?

It will likely take more than 20 minutes for to generate the chi-squared distribution. If so, this can spill over into Day 2, but it is important to regroup and ensure everyone understands.

Day 2
Minutes 0 – 20: Finish generating the chi-squared distribution (if needed)

Minutes 21 – 50: Work through the worksheet in groups, and then cycle through having the groups present to the class on the questions one at a time.

Worksheet
1. Describe the general steps you took to determine if the die was likely fair or not, and why you did each one. (Don’t spend much time on this - just demonstrate that you understand the general approach.)

2. What is a p-value?

3. Why have scientists historically relied on p-values? (This is not meant to be a historical question, but just your sense of why they help us acquire knowledge.) Do statistical distributions really model real-world scenarios?

4. How might you accidentally determine there was a significant effect when in fact there was not? What is the proper term for this type of mistake? How could you use what you did in Excel to determine how often this type of mistake will occur? Go ahead and do this several times (just press F9 until you get a significantly abnormal series of rolls), and record how many tries it took to make this type of mistake. Are you surprised by how often it occurs?

5. What would change if the experiment had a different number of possible outcomes? How does this relate to the concept of degrees of freedom? How does this relate to the actual equation of the distribution for the test you mimicked? (Need a hint? Look at this figure from Wikipedia of the Chi-square distribution for different numbers of possible outcomes)

6. How would you suspect p-values and sample size to be related? How about p-values and “effect size” (the actual magnitude of the effect you observed)? Consider how likely it is to get 100% heads when flipping a coin 2 times compared to 200 times.

7. What does the distribution you generated look like near zero? Is this surprising?

8. Try to explain what the figure (next page) from a recent Nature article is saying about the role p-values are capable of playing in science. Does this change your answer to #2? (Need a hint? Consider this quote from an accompanying article: “…rather than being convenient shorthand for significance, the P value is a specific measure developed to test whether results touted as evidence for an effect are likely to be observed if the effect is not real. It says nothing about the likelihood of the effect in the first place.”)

9. To you, what does it mean to think about statistics “computationally”? 
The observed result can be attributed to chance. But it cannot answer a question: are the odds that a hypothesis is correct? Those odds depend on how plausibly the hypothesis is in the first place.

**The Long Shot**
19-to-1 odds against

- 95% chance of no real effect

- P = 0.05

- 11% chance of real effect

- 89% chance of no real effect

**The Toss-Up**
1-to-1 odds

- 50% chance of no real effect

- 50% chance of real effect

- P = 0.05

- P = 0.01

**The Good Bet**
9-to-1 odds in favour

- 90% chance of real effect

- 10% chance of no real effect

- P = 0.05

- P = 0.01

- 99% chance of no real effect

- 1% chance of real effect
Related Lessons/Activities: N/A
Attachments: N/A
Links: N/A
Other Notes: N/A
Contributors: Ryan Langendorf, Dr. Paul Strode
Supporting Program: National Science Foundation, eCSite