doi:10.1111/j.1467-9639.2010.00446.x

What is a *P*-Value?

KEYWORDS:
Teaching;
Hypothesis tests;
Simulation;
Power.

Patricia Humphrey
Georgia Southern University, Statesboro, Georgia, USA.
e-mail: phumphrey@georgiasouthern.edu

Summary
An in-class activity is described that can be used not only to motivate hypothesis testing, but also to understand and compute the *p*-value and power in a statistical test.

◆ INTRODUCTION ◆

Very often students have a difficult time understanding the *p*-value of a statistical test of hypotheses. They (erroneously) believe it is the probability that the null hypothesis is true, or the probability it is false, along with several other mutilations of the truth. Even the definition confuses them. “The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed” is the definition given in Moore and McCabe’s (2003) *Introduction to the Practice of Statistics*. The students understand we have a sample with a particular mean or proportion. Why the focus on the test statistic, and why the “as or more extreme” part?

Years ago, when introducing the subject of hypothesis testing and trying to focus on the idea of how much information (evidence) is needed to reject the null, I brought a deck of cards into class and offered a “free homework” pass to anyone who pulled a black card from the deck. As can be expected, the students immediately assumed the deck had been rigged – even to the extent of believing there were no black cards in the deck! I pondered the problem and decided to pose the question somewhat differently. I later modified the activity to include a simulation of the expected distribution of card colours under the null hypothesis so that students get a feeling for the randomness that can occur even when the deck is fair. If time permits (or at a later class period), we return to the problem and simulate the power of the test for our particular “stacked” deck.

◆ THE DEMONSTRATION ◆

Alter, or have someone else alter, the composition of a deck of cards by mixing cards from two identical decks. I have had my daughter do this several times in the past; this worked well when she was a student at our university so at least a few in each class knew her – she was a believable scapegoat; lately, I have asked a colleague to alter the deck. The altered deck’s actual composition is marked with a hidden label. Present the deck to the class, giving the background information that someone has (possibly) altered the composition of red and black cards so that it may or may not be fair – we want to find out the truth. The deck could have been altered either in favour of red cards or black cards, so this leads naturally into a test of the hypotheses

$$H_0 : p = 0.5 \text{ versus } H_A : p \neq 0.5$$

where we will consider p to be the proportion of red cards.

Since hypothesis tests are always conducted under the assumption that the null is true, we determine the sampling distribution of the sample proportion,

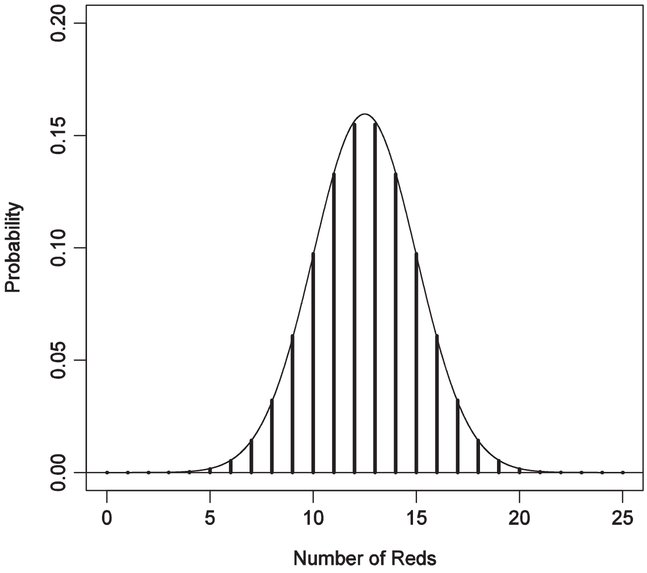


Fig. 1. The sampling distribution of the number of reds under the null hypothesis.

$\hat{p}$ under the null. This depends on the number of students in the class. For a class with 25 students with each student drawing a single card, the sampling distribution of the number of successes will be Bin($n = 25$, $p = 0.5$) or approximately $\mathrm{N}\left(\mu = 12.5, \sigma = \sqrt{25 \cdot 0.5 \cdot 0.5} = 2.5\right)$, and can be depicted as in figure 1. For the same class of 25 students, the distribution of $\hat{p}$ will be approximately Normal with mean 0.5 and standard deviation $\sigma_{\hat{p}} = \sqrt{0.5 \cdot 0.5/25} = 0.1$.

In estimating the proportion of red cards in the deck, we know, if the deck is fair, that approximately 95% of all samples should give results between 30% and 70% red cards (based on the 68-95-99.7 rule). In terms of the actual number of red cards obtained in our sample, this means that for the class of 25 students, with each student selecting a card, obtaining between 7.5 and 17.5 (or really between 8 and 17) red cards will agree with the null hypothesis; fewer than 8 or more than 17 will suggest the deck has been altered.

Now, get a student to keep track of the frequency of each colour drawn on the blackboard. Have each student successively draw a card, replacing the card and mixing a little between students. At this point, I have even heard the class begin to cheer for their favourite colour!

◆ FINDING THE *P*-VALUE ◆

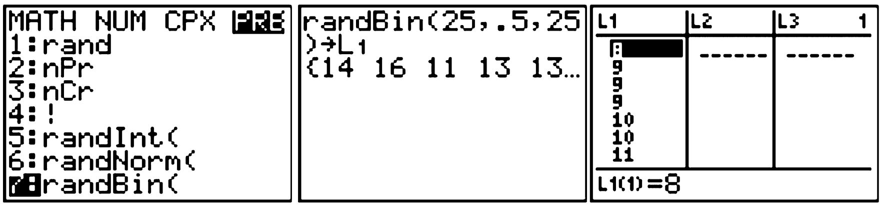
In one class last semester, there were 19 red cards drawn in a class of 25 students. We know from the above that the null hypothesis will be rejected, and conclude the deck has been stacked in favour of red cards. But how likely is our observed result? From the graph in figure 1, we know that any of the possible numbers of red cards is fairly unlikely. How likely are we to see either 19 (or more) or 6 (or fewer – an equal distance from the mean of 12.5) reds if the deck were fair? This is the *p*-value of the test, and it tends to be a probability students can understand.

The first step (which may be omitted) is to repeat the process of having students draw cards as before from a known, *fair* deck. In the same class, this yielded 12 red cards, which is much more in line with what is expected, but it is only one repetition. We know probabilities are long-term relative frequencies. How can we approach estimating the *p*-value? Simulation!

In our classes, we use the TI-83/84 as the technology of choice. It can simulate binomial experiments such as this one. On the MATH, PRB menu choice 7 is *randBin*. Selecting that option will transfer the command to the Home screen. There, one enters the parameters for the simulation: *n* (the number of trials, or students), *p* (the probability of a red card under the null hypothesis, 0.5) and the number of repetitions of the simulation (I usually have each student do 25 for time's sake). We also store the results into a list, L1 in my example below (see figure 2). To make finding 6 or fewer or 19 or more in the list easier, it is convenient to sort the list using either the *SortA* or *SortD* options from the STAT menu. In this particular simulation, none of the repetitions had less than 7 successes (or more than 17).

From the students, accumulate the number of times each had 6 or less and 19 or more (for this example). This also provides an opportunity to highlight the variability of the results. Divide this by the number of students times the number of repetitions per student, and we have an approximation of the *p*-value for the test. In the class being discussed in this example, there were 12 repetitions out of a total of 625 which fit our criterion, so the approximate *p*-value is $12/625 = 0.0192$. The exact *p*-value (obtained from the appropriate binomial calculation; *binomialcdf(25,0.5,6)* + 1-*binomialcdf (25,0.5,18)* on the calculator) was 0.0146. One can also discuss why the two (in this case) disagree – here, we have a relatively small number of repetitions of the simulation.

If your classroom is equipped with computers (ours are not), the simulation could be done easily with



```
MATH NUM CPX PRB
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(
```

```
randBin(25,.5,25
)→L1
{14 16 11 13 13…
```

```
L1      L2      L3     1
8       ------  ------
9
9
9
10
10
11
L1(1)=8
```

Fig. 2. The first screen shows the location of the *randBin* command under the MATH, PRB menu; the second shows the actual command entered to perform 25 repetitions of the simulation; and the third shows a portion of the sorted list of results.

Minitab or other software and yield much better results, as several thousand repetitions can be done almost instantaneously. In Minitab, from the Calc menu, select Random data, Binomial. Enter the number of repetitions, n and p (0.5) and a destination column. Then from the Stat menu, select Tables, and Tally to tabulate the results easily.

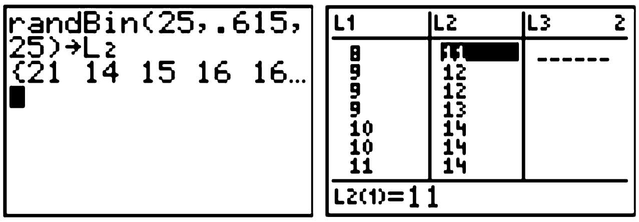
At this point of the demonstration (or somewhere earlier), students will want to know the "answer." Reveal the actual composition of the deck. For this class, my daughter had created a deck with 32 red cards, giving $P(\text{red}) = 0.615$.

◆ SIMULATING POWER ◆

The demonstration can also be used to discuss the power of a test, the probability of correctly rejecting the null hypothesis, or $P(\text{reject } H_0 \mid H_0 \text{ false})$. Since that terminology also sometimes confuses students, we translate this into the question "How likely were we to correctly detect the deck had been tampered with?" This idea came to me in a pinch one time when the class drew exactly half red cards (it *had* been altered).

Remind the students that from the discussion of the sampling distribution under the null, we decided that any result of more than 17 or fewer than 8 red cards (for the class of 25) would be inconsistent. Knowing the actual composition of the test deck, how likely is this to happen?

We could try several draws with the altered deck to begin the approximation, counting the number of times we found either more than 17 or fewer than 8 reds, but this can be time consuming. Since we found that the actual composition was $P(\text{red}) = 0.615$, we repeat the simulation above, using the actual proportion. Again, accumulate the results as was done in the simulation for the p-value, obtaining an estimate of the power of the test (see figure 3). In this particular class, we had 124 of 625 simulations result in at least 18 "reds" (with none less than 8), so the estimate of the power is 19.8%.



```
randBin(25,.615,
25)→L2
{21 14 15 16 16…
```

```
L1      L2      L3     2
8       11      ------
9       12
9       12
9       13
10      14
10      14
11      14
L2(1)=11
```

Fig. 3. Results of one simulation to estimate power when p is 32/52 = 0.615 "reds" (with none less than 8), so the estimate of the power is 19.8%.

One can repeat this if desired, having your students compute and/or simulate the p-value and power under different configurations of sample size (this makes a good demonstration of increasing power with increasing sample size) and deck composition (another good demonstration of power increasing with increasing distance from the null hypothesis).

◆ CONCLUSIONS ◆

This activity has been used for several semesters. I have found that students really have a good concept of what the p-value means after having been through the demonstration. They also develop a feeling for power. It can usually be accomplished in about half of a normal class period, so it is not too time consuming; the time spent is a good investment. If you try this, I hope your students enjoy it as much as mine have!

Reference

Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics* (4th edn). New York: W. H. Freeman.