Avril Kenney
AP Statistics
Project #1
11/22/06

A Survey Regarding Selection of Numbers from 1 to 10

The purpose of this survey, aside from the general purpose of practicing writing and giving surveys, was to look for patterns in random numbers selected by people. The survey was given by email (posted on CPSNet), so the sample was entirely self-selected from the part of the CPS population who read Items of Interest during the few days that the survey was posted (of 141 people who opened the email, 40 responded). Since this particular survey was, or at least was intended to be, based on processes of the brain which should be similar for everyone, there are no obvious lurking variables, so non-random selection of the sample should not have a significant effect on the data.

The survey consisted of the following questions, preceded by instructions telling the readers to answer each question before reading the next question and not to change answers once written. There were several lines of empty space between the questions in order to prevent people from accidentally looking ahead.

1. Choose at random a number from 1 to 10. What is it?
2. Of the numbers from 1 to 10, which is your favorite?
3. In questions 1 and 2, did you assume it meant inclusive?
4. In questions 1 and 2, did you assume it meant integers?
5. In question 1, is there any conscious reason that you might have chosen that number? If so, what was the reason?

Question 1 by itself was used to evaluate people's ability to choose numbers "at random." If the selections were truly random, then they would be uniformly distributed from 1 to 10. Question 2 by itself is interesting because here people are actually asked to have a preference, so there was no mathematical prediction for the distribution. Question 2 could also be compared with question 1 to determine whether a person's favoritism had any effect on their choosing of a "random" number, and question 5 also contributed to explaining non-random selections in question 1. Questions 3 and 4 were to find out how people had interpreted questions 1 and 2; question 3 was about the interpretation of the English phrase "from ... to," and question 4 was a distinction between mathematical versus common usage of the word "number".

One rather flawed aspect of this survey is that there was no particular question it was trying to answer. It would probably have been better to have an intended research question and then to create survey questions that could lead to an answer. As it is, this survey was created with the fairly vague purpose of getting some numbers and then looking for patterns in them.

This type of survey is very straightforward, a style which has two clear advantages: people can answer it quickly and easily, and the data are completely numerical with no room for people to give answers different from the type intended (with the exception to the second part of question 5, which was in fact just out of curiosity and is not analyzed anywhere). The disadvantage is that there are not many types of research questions for which such a survey can be used.

The results of the survey, while not very different from what was expected (an aspect of the experiment which was also quite vague, with no specific hypotheses defined before the administration of the survey), were interesting.

Values for people's answers to question 1 are termed "R" for "random."
Values for people's answers to question 2 are termed "F" for "favorite."

Distribution of R:

R	Frequency
1	1
2	4
3	5
4	2
5	4
6	7
7	8
8	5
9	3
10	1

Distribution of F:

F	Frequency
1	1
Φ	1
2	4
3	7
4	2
5	5
6	0
7	12
8	5
9	2
10	1

$R_{avg} = 5.675$
$s_R \approx 2.336$
Q1 = 3.5
Median = 6
Q3 = 7

$F_{avg} \approx 5.415$
$s_F \approx 2.463$
Q1 = 3
Median = 6
Q3 = 7

Both R and F have very non-uniform distributions, though their means are both very close to what the mean would be if the numbers were selected completely randomly (μ=5.5). It is also interesting that they both have the same basic shape: increase from 1 to 3, drop at 4, increase from 5 to 7, decrease from 7 to 10. In the distribution of F, however, there is a very noticeable extra low point at 6, which was the only integer value never selected. This exception is especially curious because in the distribution of R, 6 has a frequency between those of 5 and 7, and also because 7 has an extremely high frequency in the distribution of F.

In addition to their individual distributions, R and F were also compared to each other, based on the belief that a person's choice of a "random" number might be influenced by his favorite number. The regression equation of R on F was $R \approx .137F + 4.933$, with an r^2 value of about .021, indicating that there is no significant correlation in general. However, when R and F values were compared for individual people, it was found that 12 (30%) of people gave the same values for R and F (if it were due to random chance, only 10% would be expected to have equal R and F values). This correlation was confirmed to be causation for all 12 (100%) of these people in their answers to question 5; in contrast, only 12 (~43%) of the 28 people whose R and F were different answered yes to question 5.

As for frequencies of specific numbers, those of the F values do have a wider range, indicating that favorites are more biased than "random" numbers. The number 7 has the highest frequency for both R and F; this could be because 7 is traditionally considered a lucky number. Similarly, 4 has low frequencies for both R and F, and it is often thought to be unlucky. The most interesting number is 6, because it has the second-highest frequency for R but a frequency of 0 for F. A possible explanation is that the R category is considered more mathematical while the F category is more related to real life; the average number between 1 and 10, 5.5, would be rounded to 6 mathematically, but 6 is also the digit that makes up 666, a number that supposedly represents the devil. However, further investigation would need to be done to check the validity of this explanation.

Questions 3 and 4 were for the most part not related individually to the other questions. Of the 40 people who took the survey, 37 (92.5%) answered yes to question 3 (regarding the assumption of inclusiveness), and 39 (97.5%) answered yes to question 4 (regarding the assumption of limitation to integers). If it can be assumed that the people were correct in the reporting of their assumptions (i.e., for question 3, they actually assumed it to be inclusive at the time rather than not considering it until afterward), then it seems that people have a tendency to avoid numbers toward the edges of the range, since for both R and F, 1 and 10 only occurred once, whereas in a uniform distribution they would each have occurred 4 times. The results of question 4 are not surprising, for the word "number" as used in common speech generally denotes a counting number.

The results of this survey are quite intriguing, but if anything they raise more questions than they answer. As a matter of fact, this was probably not a very good way to learn about surveys, because this topic was so contrived that it is much simpler than actual useful surveys would be, and the process by which it was administered required virtually no effort on the part of the writer. The topic was chosen so that analysis would be simple but results could still be interesting, and this purpose was achieved, but perhaps at the expense of the greater learning experience.