

Probability and Simulations in the AP Curriculum

In the AP Statistics syllabus the concepts of probability that are stressed are those that serve the understanding of statistics, and to some extent deviate from the traditional topics presented in high school algebra texts. We would like to briefly discuss some of the concepts of probability that are useful in statistics, and point out differences and similarities in their use by mathematicians and statisticians.

Probability

The usual presentation of probability is a blend of what is usually termed the “Classical” understanding of probability and a set of mathematical axioms defining the behavior of probability. The classical understanding of probability, formalized by the mathematician Pierre Simon de Laplace in the early 19th century, grew as a method for analyzing games of chance. In most gambling scenarios, probabilities are associated with the outcomes when dice are rolled or coins are flipped, or cards are drawn. In this arena it is reasonable to consider probability of an event, E , as a ratio:

$$P(E) = \frac{\text{number of favorable outcomes}}{\text{number of outcomes in the sample space}}$$

There are some limitations to this conception of probability -- not all events in life are equally likely. For example, if you take your car out for a spin, you may or may not have a flat tire; thankfully, these events are not equally likely!

A strictly mathematical formulation of probability was provided by the Russian mathematician, A. N. Kolmogorov, in the 1930's. His axiomatic formulation for probability may be successfully applied in all chance situations, not just the equally likely outcomes of the classical probabilists. His axioms include the following familiar statements:

1. For all events, E , $0 \leq P(E) \leq 1.0$.
2. For the sample space, S , $P(S) = 1.0$.
3. If two events, E and F , are disjoint, then

$$P(E \text{ or } F) = P(E) + P(F).$$

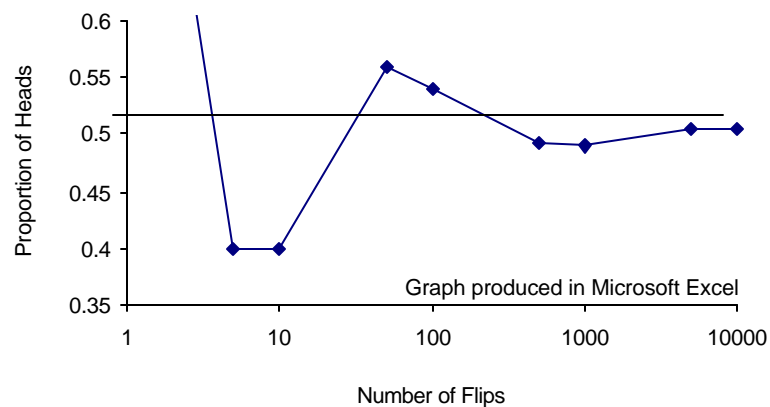
Neither the classical nor the axiomatic approach to probability completely meets the needs of the statistician. The understanding of probability used by statisticians is known as the frequentist approach to probability. From this perspective, probability is considered to be a constant long-run relative frequency. Borrowing some terminology from calculus, the frequentist might express the probability of a (possibly biased) coin landing heads to be the long run relative frequency of heads in an “infinite” number of coin tosses, and think of the probability as a limiting value of the proportion of heads in the sequence of long run relative frequencies.

For example, if a fair coin is repeatedly flipped and the relative frequency of heads is noted after each flip, observed relative frequencies will tend to differ less from what the frequentist would regard as the “true” probability. This tendency is formulated as the *Law of Large Numbers*. This long-run stability in a statistic is what makes statistics a credible mathematical science.

The table at right shows the results of a simulated series of coin flips and the proportion of heads at indicated positions in the series. The accompanying graph illustrates the long-term trend. The last two columns in the table illustrate that even though the deviation from $\frac{1}{2}n$ in the number of heads increases with the number of flips, the deviation in proportion from $\frac{1}{2}$ tends to decrease.

Number of Flips	Number of Heads	Proportion of Heads	Deviation in Number from 50%	Deviation in Proportion from 50%
1	1	1.000	0.5	0.500
5	2	0.400	0.5	0.100
10	4	0.400	1	0.100
50	28	0.560	3	0.040
100	54	0.540	4	0.040
500	246	0.492	4	0.008
1,000	490	0.490	10	0.010
5,000	2,525	0.505	25	0.005
10,000	5,040	0.504	40	0.004

Simulated Coin Flipping



The different traditions of understandings of probability can and do lead to different uses of some of the terminology related to probability, and this can be disconcerting to teachers who, in their quest to better understand and teach the “probability chapters,” refer to probability books, as distinguished from statistics and mathematical statistics books. We will attempt to sort out some of these usages below.

Random or chance experiment

In the field of probability the term “random experiment” is formally abstractly defined consistently with Kolmogorov’s axioms. A random experiment is a mathematical entity characterized by:

1. A set, S , the sample space,
2. a collection, E , of subsets of S , the event space, and
3. a real valued function, Pr , the probability measure, defined on E , that has the Kolmogorov properties.

The term “chance experiment” is sometimes used by statisticians to mean any activity or situation in which there is uncertainty about which of two or more possible outcomes will result.

Some statisticians prefer to reserve the word “experiment” for randomized comparative experiments and refer to the probabilists’ random experiment as a random “circumstance.” In the context of AP Statistics an experiment is understood to be a type of study wherein treatments are randomly assigned to subjects. This is understood to be distinguished from an observational study where the investigator does not have the ability to control the experimental environment to the extent of assigning treatments.

Random variable

Another concept shared by mathematicians and statisticians is the idea of a random variable. A random variable is defined as a real valued function X , defined on a sample space. For statisticians, a random variable is a function whose numerical value depends on the outcome of a chance experiment. The particular chance experiment statisticians have most in mind is taking a random sample. The numerical value that depends on the outcome of the sample is calculated from observations taken on the elements that make up the random sample, and to the statisticians, this numerical value is known as a statistic. For statisticians, random variables are particularly important because a sample statistics *is* a random variable.

Probability Distributions

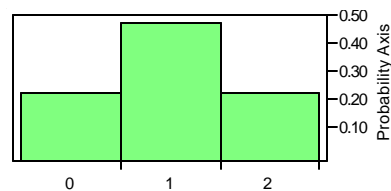
For both mathematicians and statisticians the probability distribution for a random variable, X , represents an assignment of probabilities to the values of a discrete variable or to intervals of a continuous variable random variable, X . For the statistician, these probabilities are the limiting relative frequencies of outcomes when a chance experiment is performed.

For instance, the random variable X might represent the number of heads observed in two flips of a fair coin. X can have values of 0, 1, and 2. The question is then how often each of those values occurs. There are two ways of determining this. First, one might determine the values theoretically from a priori statements about the “fairness” of the coins. Second, one can estimate the values by performing the chance experiment of flipping

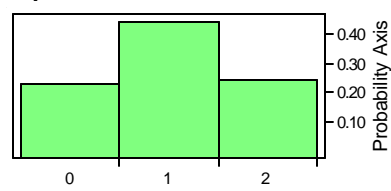
Theoretical	
X	$Pr(X)$
0	0.250
1	0.500
2	0.250

Experimental (100 trials)	
X	$Pr(X)$
0	0.260
1	0.470
2	0.270

Theoretical



Experimental



Graphs produced in JMP-INTRO by SAS Institute, Inc.

two coins many times. The table and accompanying graphs show both the theoretical probability distribution and the results of a simulation of flipping two coins. For the statisticians, these estimates in the long run are by definition the probabilities by virtue of the Law of Large Numbers.

Returning to the favorite chance experiments of statisticians – sampling from a population and calculating a statistic – the concept of a probability distribution is enshrined as one of the fundamental concepts of inferential statistics: the sampling distribution. The sampling distribution of a statistic is a probability distribution of a random variable.

References:

Bean, M. A. *Probability: The Science of Uncertainty with Applications to Investments, Insurance, and Engineering*. Brooks/Cole, Pacific Grove, CA. 2001.

Peck, R., Olsen, C., and Devore, J. *Introduction to Statistics and Data Analysis*. Duxbury Press, Pacific Grove, CA. 2001.

The Role of Probability in the AP Syllabus

Where do the ideas of probability and random variables arise in statistics? Why are the “Probability Chapters” included in the syllabus?

A) Probability in the context of surveys.

It may seem odd to connect the real world of surveys with the abstract world of probability, and it is true that calculations of probability do not seem to appear in reports of surveys. But it is also true that probability underpins the conduct and interpretation of surveys in significant ways. In the paragraphs to follow, notice how many of the ideas and concepts of the “chapter on probability” appear in the discussion of sample surveys.

Consider the choice of a simple random sample from the population. In a simple random sample, each collection of n elements from the population has an equal chance of being selected for the study. The chance of getting any particular sample, that is any particular subset of n elements from a population of size N , is:

$$\begin{aligned} P(\text{Any particular set of } n \text{ elements chosen}) &= \frac{1}{{}_N C_n} \text{ or} \\ &= \frac{1}{\binom{N}{n}}, \text{ depending on your notation.} \end{aligned}$$

$\binom{N}{n}$ and ${}_N C_n$ commonly denote the number of different groups of size n that can be drawn (without replacement) from a population of size N .

Suppose now that we are gathering information about whether the individuals in a sample are favorably disposed to a recent movie; each of the ${}_N C_n$ possible simple random samples has associated with it a sample proportion of people favorable to the movie. Possible values for this sample proportion are $\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}$. Each of these possible values, in turn, has a probability of occurring. In the language of probability, the assignment of a numeric value (the sample proportion) to the outcome of the sampling process creates a random variable. The probability distribution of the sample proportion is the “sampling distribution” of the sample proportion, and it is determined by the probabilities associated with each of the possible outcomes of the study.

B) Probability in the context of comparative studies.

To continue our example above, suppose that it is of interest to compare the proportions of men and women who are favorably disposed toward a particular movie. All of the probability considerations discussed above would apply to sampling from the population of males and the population of females. In addition, there is an added concern – the interpretation of a difference in proportions.

The sampling distributions of the individual sample proportions, together with the properties of random variables, allow us to determine the sampling distribution (probability distribution!) of the statistic that we can use in this new context. We know, for example, that under the appropriate conditions the sampling distribution of the sample proportion, \hat{p} , has a mean, or expected value, equal to the population mean, p . We also know that the variance of this sampling distribution is equal to $\frac{p(1-p)}{n}$.

Fast forward now to the problem of comparing two proportions. A natural way to investigate the difference between the two proportions, p_M and p_F in our example, is to consider the difference, $\hat{p}_M - \hat{p}_F$. Textbooks suggest that for large samples the statistic $\hat{p}_M - \hat{p}_F$ is approximately normally distributed with mean of $p_M - p_F$ and variance equal to $\frac{p_M(1-p_M)}{n_M} + \frac{p_F(1-p_F)}{n_F}$. While this is certainly true, it must also be said that these are not just more pretty formulae! They are derived using the properties of random variables:

$$\begin{aligned} \mathbf{m}_{\hat{p}_M - \hat{p}_F} &= \mathbf{m}_{\hat{p}_M} - \mathbf{m}_{\hat{p}_F} \\ &= p_M - p_F \end{aligned}$$

and with the additional assumption that the samples are independent,

$$\begin{aligned} \mathbf{s}_{\hat{p}_M - \hat{p}_F}^2 &= \mathbf{s}_{\hat{p}_M}^2 + \mathbf{s}_{\hat{p}_F}^2 \\ &= \frac{p_M(1-p_M)}{n_M} + \frac{p_F(1-p_F)}{n_F} \end{aligned}$$

While we may not explicitly “derive” these results for our students, it might be worth mentioning that the chapters on probability and random variables actually are relevant to the study of statistical inference.

Generic Simulations

In the AP Statistics course we take on the task of helping students design and execute simulations, and analyze the results of such simulations. By simulation we mean an artificial chance experiment designed to mimic a real world problem in all relevant aspects. It is hoped that the distribution of outcomes of such a simulation will lead to an understanding of a corresponding real world problem, and quite possibly a “solution” to a posed problem.

Experience – and looking in a lot of reference books! – teaches that the variety of situations that can be effectively simulated overwhelms any attempt at constructing a perfect “ n -step method.” For purposes of our discussion, we have constructed a 5-step method; we are not under the illusion that this will work perfectly for all simulations, and certainly we do not want to impose upon teachers a “right” way to think about simulations. We do hope to offer a consistent terminology as we talk about simulations, so we can more effectively communicate our thoughts and ideas.

Let us examine this 5-step process and try to suggest – one step at a time -- what we mean at each juncture. Since simulations are generally performed when scientists or statisticians are confronted with “real” problems, we will not shirk our responsibility to use a real life situation.

An arguably decent 5-step method for constructing simulations

1. Understand the problem, especially the random aspects.
2. Identify the statistical elements of the problem, represent the components as statistical elements, and define outcomes.
3. Construct mechanisms and algorithms for modeling the statistical elements of the process.
4. Execute the algorithm.
5. Analyze the results statistically – i.e. answer the question.

The scenario we will use is from a paper by Bakker & Milinski (1991), which discusses the three-spined stickleback, *Gasterosteus aculeatus*. The three-spined stickleback is a fish, typically about 3 inches long as an adult, found in Europe, northern Asia, and North America. In early spring (mating season) the male stickleback stakes out a territory and prepares a nest for the next generation. His belly turns bright red, and this attracts females. Females deposit eggs in the nest and leave; the male stays with the nest, protecting newly hatched sticklebacks from predators. The female chooses a male by evaluating the color brightness of male candidates sequentially, and having chosen

one she deposits her eggs and leaves. (Actually it is a little more complicated than that, but in the interest of decorum we will avoid excessive detail.)



Our question of interest in this example centers on the strategy used by the female to select the male. Biologists have proposed different possible strategies, one of which is the “fixed threshold” rule. A female using this strategy would mate with the first male she found that was brighter than some fixed level of brightness. Formally, our question is this: if the female uses a fixed threshold rule, how many males will be passed by before the female selects a male?

Simulation Step 1: Understand the problem, especially the random aspects.

“Understanding” the problem is a vague, but necessary intuitive first step. By the components of a problem, we mean those small “chunks” that together define the characteristics of the problem. The various components of a problem to be simulated must be understood, as well as the relationships between and among the components. In the female stickleback choice problem, for example, we need to understand components such as the following: the problem involves a sequential choice, the brightness of the males is a variable rather than a constant, and there is a “stopping” point where the process under consideration would end. It probably does not matter that the males are colored red (as opposed to blue), or that these creatures are not found in, say, South America, or that biologists rather than statisticians have formulated a mate selection strategy. This first simulation step, understanding the problem, must surely involve repeated and careful readings of the problem, some reflection about what aspects of the information presented are relevant, and especially the outcome measure of concern for the question(s) of interest.

Simulation Step 2: Identify the statistical elements of the problem, represent the components as statistical elements, and define outcomes.

The “components” of the problem are those features that are the parts of the process. In a simulation, generally these components will consist of some number of variables, possibly some number of constants, and some idea of the relationships between and among these components. After identification, we must translate the “real” components into mathematical representations. In the present circumstance, for example, it would

seem that the brightness of the male stickleback is a necessary component. We observe that the choice behavior of the female is related to the brightness of the male stickleback in a pretty well-defined manner, but the actual brightness of the individual fish is random. It may be that all females have the same threshold, but more likely the threshold varies from female to female. Even if it is determined that the thresholds vary, we may choose to make the simplifying assumption, for the purposes of simulation, that they do not. (Possibly different simulations would be undertaken for different values of the threshold.) The process terminates with the female's final choice of a mate. The outcome of interest is often identified after scanning the sentences in the problem scenario and considering those with question marks at the end. Perhaps something like: "if the female uses a fixed threshold rule, how many males will be passed by before the female selects a male?"

Simulation Step 3: Construct mechanisms and algorithms for modeling the statistical elements of the process.

After the individual elements have been identified, they must be modeled with mathematical representations. If a particular element is random, modeling the element will usually mean assigning some sort of structure using a probability model. For example, we might think of the random male stickleback as possessing a "probability of success" equal to some fixed value. We might also think of the population of the brightness values of male sticklebacks. Upon randomly selecting a male, it can be determined whether its brightness exceeds the female's threshold. The choosing by the female might be modeled as a sequence of mate/not mate decisions that terminates with the first decision to mate (because she cannot, having laid her eggs, mate again.)

A mathematical representation of the process might be geometric, algebraic, or verbal. It might even utilize some sort of algorithmic representation or flow chart, such as is common in the computer programming field. The "best" representation will depend mostly on the individual attacking the problem and the tools he or she brings to bear! A reasonable place to begin the problem representation would be to define the concept of a "run" for the simulation. By a "run" we mean the steps necessary to determine a single agreed-upon outcome for further analysis. We will perform the runs a very large number of times, keeping track of outcomes; these outcomes will constitute the "simulation." In general, consideration of the data from the runs (i.e. the distribution of outcomes) will be our analysis of the simulation results. Generally the simulation of a process should result in a large number of outcomes, which will be subsequently analyzed statistically.

Simulation Step 4: Execute the algorithm.

The execution of the algorithm might consist of the presentation of various mathematical representations, geometric, algebraic, or algorithmic, and unfolds as a sequence of well-defined operations. If the algorithm is executed for someone else, such as in the classroom or on the AP Statistics test, it is very important to represent the algorithm, including the random numbers generated during the process, as well as any decision points in a run. If a random number table is used, it should be "marked up" with appropriate notes for the reader to follow. If a calculator or computer generates random

numbers, the actual random numbers should be presented in a list and subsequently “marked up.” In some cases – e.g. not in a classroom or taking the AP Statistics test – this communication step might seem to be unimportant, but skipping this step would be an error! It is the responsibility of the simulator to check and verify the algorithm. One excellent way to check it is to write out the random numbers and keep track of the process as it initially unfolds – long enough to be sure it is working correctly. This is especially of concern if a program or script has been written for a computer or calculator to do the actual calculations. It is oh, so dangerous to assume the simulation works just by looking at the program steps.

Simulation Step 5: Analyze the results statistically – i.e. answer the question.

A common error when performing a simulation is to lose sight of the fact that there actually was a question that is being asked. The simulation is generally not the goal; it is the means toward the goal. The output of the simulation provides the raw data for further examination, usually statistical in nature. Which statistical analyses are appropriate will, of course, depend on the goal of the simulation and the nature of the output of the trials.

A possible simulation of this problem.

1. Understand the problem, especially the random aspects.

The problem is to analyze a particular mate-choosing strategy by a species of fish. The female encounters males in sequence, and chooses either to mate with that male or to continue on to the next. For each male, the decision is made based on the brightness of the male’s belly. Once a mating occurs the female stops her choice behavior.

2. Identify the statistical elements of the problem, represent the components as statistical elements, and define outcomes.

There appear to be two components to this problem: (1) the presentation of the male and subsequent decision by the female, and (2) the sequential nature of the presentation of males. The brightness of the male varies, and will need to be modeled with probabilities. The simplifying assumption that all females have the same threshold is made. The sequential nature of the choice behavior could be modeled using a sequence of random outputs from a probability model. The run ends with the female’s selection of a sufficiently bright-bellied male. The stopping point of a run will be the yes decision by the female. The outcome of each run will be the number of males presented, excluding the one chosen. Some number of repetitions of those runs will generate the distribution of outcomes.

3. Construct mechanisms and algorithms for modeling the statistical elements of the process.

The distribution of brightness is not specified in the problem, but it seems reasonable to model the brightness as a normally distributed random variable. Let the female be somewhat picky, requiring the male to have brightness greater than one standard deviation above the mean before she will mate. The algorithm for the simulation can be constructed as follows:

Step 1: Represent the male brightness by generating a random number, z , distributed as a standard normal random variable.

Step 2: If the z generated in step 1 is greater than 1.0, the male is chosen and the number of males considered so far – excluding the one chosen – is reported as an outcome, thus completing one run. If the z generated in step 1 is less than or equal to 1.0, the male is not chosen and Step 1 is executed again.

Step 3: The outcomes from a large number of runs of Steps 1 and 2 will be analyzed statistically.

4. Execute the algorithm.

Using the RandNorm command on the TI-83, the following z 's were generated. Notes tracing the execution of the simulation are included for communication purposes. [This is the “show-your-work” that is so essential when performing a simulation on a test.] The outcomes are in parentheses.

Generated z	Execution Trace	Generated z	Execution Trace
.0839	1	-0.867	1
.5650	2	-0.091	2
-0.904	3	-1.441	3
-2.090	4	-0.317	4
0.201	5	0.430	5
0.243	6	0.114	6
0.742	7	-0.860	7
-0.326	8	-1.115	8
-0.299	9	0.652	9
0.800	10	-0.248	10
0.572	11	0.591	11
0.796	12	2.286	Brightness exceeds $z=1$; (11)
1.312	Brightness exceeds $z=1$; (12)	-1.409	1
0.355	1	-0.978	2
-0.344	2	0.658	3
1.092	Brightness exceeds $z=1$; (2)	-0.600	4
-0.430	1	0.685	5
2.240	Brightness exceeds $z=1$; (1)	1.279	Brightness exceeds $z=1$; (5)

5. Analyze the results statistically – i.e. answer the question.

Of course, in this example we don't have anything like the number of trials we would need to analyze our outputs, but it is anticipated that with many trials the analysis might focus on the center, variability, and shape of the distribution of the number of “rejected” males.

Comment:

The “fixed threshold rule” is relatively simple to set up in a simulation, but is not the only or even the most interesting female choice strategy. Other strategies mentioned in the Bakker and Malinski (1991) article are:

1. The threshold-criterion rule. The female compares males sequentially until the most recently encountered male is of less brightness than the previous male encountered and she then mates with the previous male.

2. The sequential comparison rule. The female compares males sequentially. She mates when the brightness of the male encountered is greater than the male brightness expected from continued search. (Such a female would not only have spent a long time schooling, but studied statistics as well!)

3. Best-of- n rule. The female samples as many males as possible and picks the brightest one. In this scenario, the possible number sampled is limited by the threat of predation, her memory capacity for remembering male brightness, and the ticking of her biological clock; she cannot continue to reject males without cost. (Obviously, for this scenario the question of interest must change, perhaps relating to expected brightness of the chosen male.

In closing

Our attempt at formulating and illustrating a reasonably generic approach and language for simulation purposes in AP Statistics is now complete. Thus, a few “afterwords” might be in order. First, we restate that the terminology we have used is not intended to be representative of standard or professionally accepted terminology. We do not believe there is such a standard at this time. The terminology herein is offered as reasonable and allows us to maintain a certain consistency in our examples.

Secondly, the stickleback example was chosen to provide a fairly realistic setting wherein a practicing scientist might choose to use simulation. In the simulation above, once the brightness criterion and distribution of brightness in the population were specified, the distribution of numbers of rejected males was doomed to fall in the pattern of the well-known geometric distribution. In “real life,” this problem would in all

likelihood be solved analytically. The three alternative strategies mentioned immediately above, however, are less likely to admit an analytic solution, at least at the high school level. In that sense, the alternative choice strategies are better examples of simulations that are “reality based” as distinguished from pedagogical.

References:

Bakker, T., Milinski, M. Sequential female choice and the previous male effect in sticklebacks. Behavioral Ecology and Sociobiology, 1991, 29:205-210.

Stickleback picture downloaded from:

<http://www.sarep.cornell.edu/Sarep/fish/Gasterosteidae/stickleback.html>