Probability theory

The class is about statistics. This session is about probability. They are not the same, but they are related.

Probability theory is a branch of mathematics that precisely describes random processes. It initially centered around gambling, but by the early 1900's, it took hold in science, and it was growing in response to questions in genetics, physics, and financial economics. Here's a classic probability problem. *Suppose a stock price changes by 0.1% every hour, going up with 51% probability and down with 49% probability. What is the chance it will exceed a 3% increase some time in the first 30 days?*

Mathematical statistics is a complement to probability, working backwards rather than forwards. *Suppose a stock price changes by a factor of θ every hour, going up with probability p and down with probability $1-p$. Given records of stock prices, what are the ranges of plausible values for p and θ?* These questions can be answered mathematically for a given set of model assumptions, without needing to look at a particular dataset.

Applied statistics, now often called **data science**, faces more than just mathematical questions. *Is the idea of regularly updating a stock by small percentages workable, or are there irregular or violent shocks that must be considered? What additional data would give us a deeper understanding of this stock price variation?* These are empirical questions, meaning their answers cannot be deduced from first principles. The answers depend on observation.

This course will consider mathematical questions and empirical questions, as well as questions whose answers require both modes of thought. But, we need some probability theory as a jumping-off point. Today, as a bridge into mathematical statistics, we'll discuss some basic probability concepts using one of the simplest and most common statistical models: the "Binomial distribution."

Learning objectives

- Describe binomial random variables.
- Define probability mass functions. Write them out for binomial random variables.
- Define expected value $E[X]$ and variance $Var[X]$ of a random variable X.
- Define independence for random variables. Describe pairs of random variables that are independent or not.
- Calculate expected value and variance of a binomial random variable in two different ways.
- Describe the difference between a parameter and an estimate of that parameter.

Background

This might seem like too much information at once. Try it anyway, and be patient. The exercises below will give you a chance to practice at a slower pace.

- A “random variable” is exactly what it sounds like: a number that results from a random process, such as a coin flip, a card deck shuffle, a decaying radioisotope, a meiotic recombination, a stock market fluctuation, or a thermal motion at nanometer scale.
- A “binomial” random variable is what you get by flipping a coin N times (each with probability p for heads and $1-p$ for tails) and counting the number of heads.
- A “probability mass function” takes possible values of a random variable as input and yields their probabilities as output. If my random variable is the number of heads after two coin flips, the inputs to the PMF can be 0, 1, or 2. The outputs would be 25%, 50%, and 25% respectively. In symbols, $f_X(0) = f_X(2) = 0.25$ and $f_X(1) = 0.5$, where f is the probability mass function and X is the random variable.
- The “expected value” or “the mean” is the average over all possible results, each weighted by its probability. In the example, it is $0.25 \times 0 + 0.5 \times 1 + 0.25 \times 2 = 1$. For a random variable X, the expected value is written $E[X]$.
- The “variance” is the expected value of the square of the distance to the mean. (Squaring it prevents it from being negative.) In the example, it is $0.25 \times (0-1)^2 + 0.5 \times (1-1)^2 + 0.25 \times (2-1)^2 = 0.5$. For a random variable X, the variance is written $Var[X]$.

Coin 1	Coin 2	Total	Prob	Contribution to mean	Contribution to variance
H	H	2	0.25	0.5	0.25
H	T	1	0.25	0.25	0
T	H	1	0.25	0.25	0
T	T	0	0.25	0	0.25

- Two random variables are “independent” if each contains no information about the other. Mathematically, this happens when $P(E_1 \text{ and } E_2) = P(E_1)P(E_2)$ for any pair of events E_1 and E_2 where each event involves only the respective variable. In all the calculations above, we have assumed the coins are independent.
- For any two random variables, $E[X+Y] = E[X] + E[Y]$, even if they are not independent. If they are also independent, then $Var[X+Y] = Var[X] + Var[Y]$. This can be proven mathematically – take a proper stat class if you’d like to see how.

Next, we will take some time to experience the probability rules and definitions that were laid out above, by implementing them in software.

Exercises

Type your answers in a Word document and your code in an R script and email them to Eric as `LAST_FIRST_binomial.docx` and `LAST_FIRST_binomial.R`. If you want to write math symbols more easily, you could also try writing **markdown** using an editor like this one.

1. If the coin is weighted to have a 1/3 chance of landing heads, and you flip it twice, then what is the PMF of the total number of heads? The expectation? The variance?
2. Suppose the coins we flip follow a "herd mentality". The second coin, if it sees the first land heads, will also land heads. If it sees the first show tails, it will show tails as well. How does the table above change? What is the expected value? The variance?
3. Write a function in R to compute the expected value and variance of a binomial random variable given the number of trials and the success probability. Use the definitions, even if you are aware of a more convenient method. You may use the built-in function for the PMF, which is called `dbinom`. Write out problem 1 as a test case to ensure your code is correct.
4. Draw 10,000 independent binomial random variables, each with 25 trials and success probability 0.12345. You can use the built-in function `rbinom` to generate random samples. Estimate the mean and variance of the results from your samples. You can use the built-in functions `mean` and `var`. Also compute the mean and variance from your code for the previous question. Are the estimates close to what your code would predict?
5. You can represent a binomial random variable as a sum of independent random variables, each of which is mathematically simple on its own. This leads to simple, efficient formulas for the mean and variance that work for any p and n. Figure out how to do this. Report your thought process and your results in terms of p and n. (For instance, this is wrong but the format is correct: $E[X] = n^2/p^2$.) Use the properties $E[X + Y] = E[X] + E[Y]$ and $Var[X + Y] = Var[X] + Var[Y]$. This could be disorienting. If you can't get started, we can provide a scaffold to structure your calculations.
6. Above, you used the built-in `R` function `dbinom`. It would be nice to know what that does. What is the PMF of a binomial random variable with success probability p over n independent trials? Look it up on Wikipedia or a reference of your choice. Explain why each component of the formula is present and what would happen if you left each part out.