Kathryn Bolton, Lara Haase, Nicholas Maier Homework 3

Problem #1:

a) Is every uncertain quantity in the world well described as a normal random variable with a bell shaped distribution? Answer in one word. (Hint: This question is very easy. The answer is "No".)

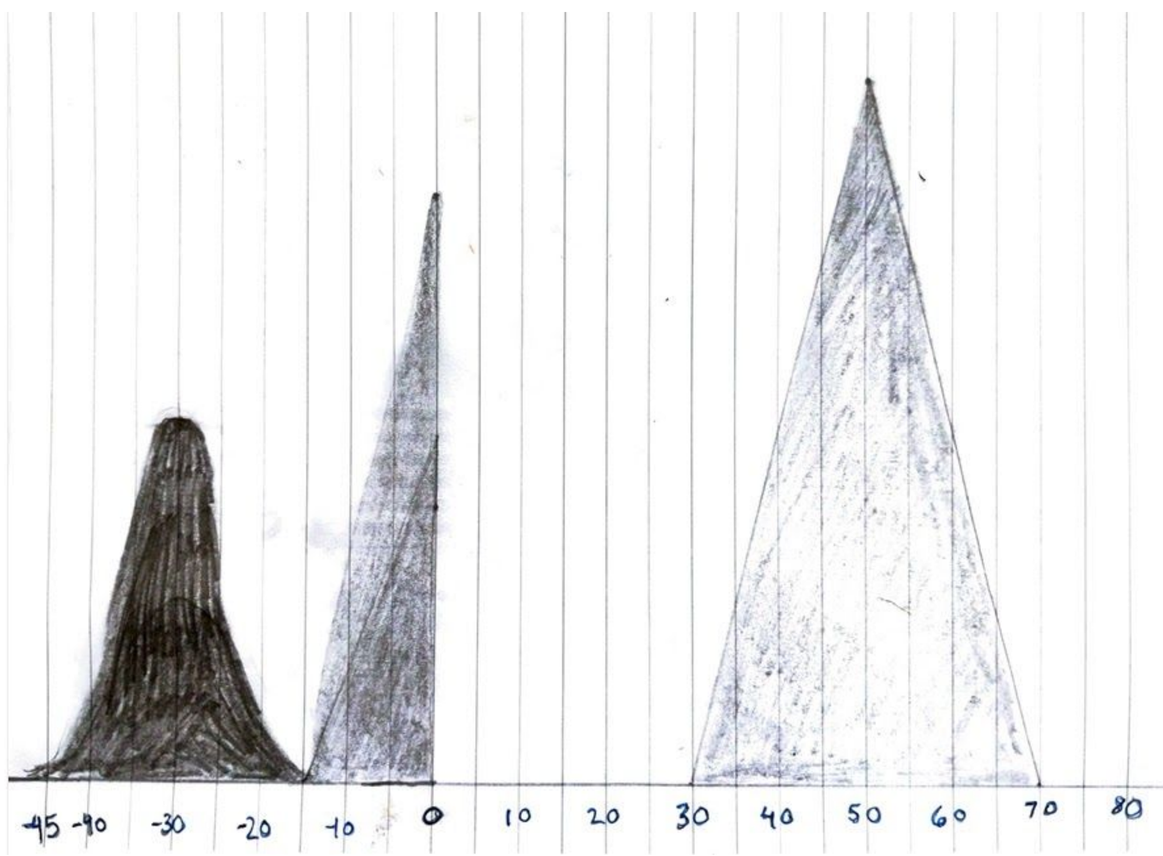
No.

b) In two to three sentences, give an example of an uncertain quantity that is not well-described by a Normal random variable, and explain why its distribution is not bell shaped.

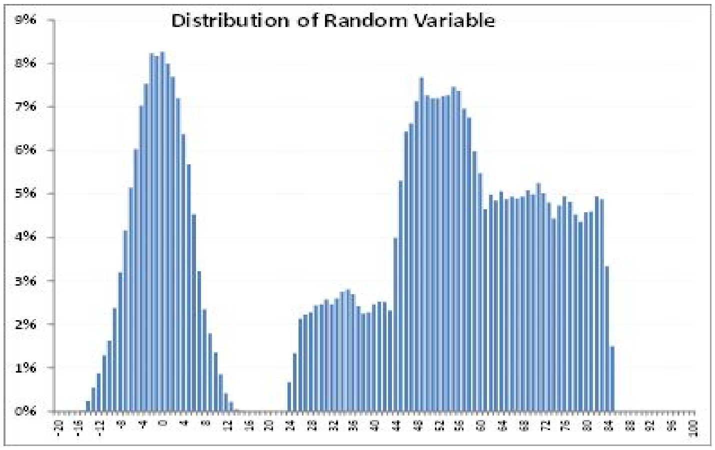
Density of mosquitos in the Florida Everglades over time. Sometimes in the Everglades, there are many mosquitos, and other times there are many many many mosquitos. This density is hard to quantify accurately, and it is not well described by a Normal random variable, because the true quantity is dependent on a number of other factors, primarily the temperature and human interventions (such as pesticide use).

Problem #2:

a) Convert the following verbal description of a distribution into a sketch. (Note: I expect most of you will submit a (neatly) hand drawn sketch, which you can scan/ embed into your pdf submission. This can also be graphed in excel if you prefer.) "About 20% of the probability mass is in a bell-shaped distribution centered around -30 and with most of its mass between -45 and -15. About 20% of the probability mass is in a roughly asymmetric triangle between -15 and 0, with values at the upper end of that range being the most likely and probability declining roughly linearly to near 0 around -15. The greatest bulk of the probability, about 60%, is in a symmetric triangle shape between 30 and 70."



b) Convert this picture of a distribution into a verbal description that would enable someone else to sketch a fair approximation to this picture without actually seeing the picture.

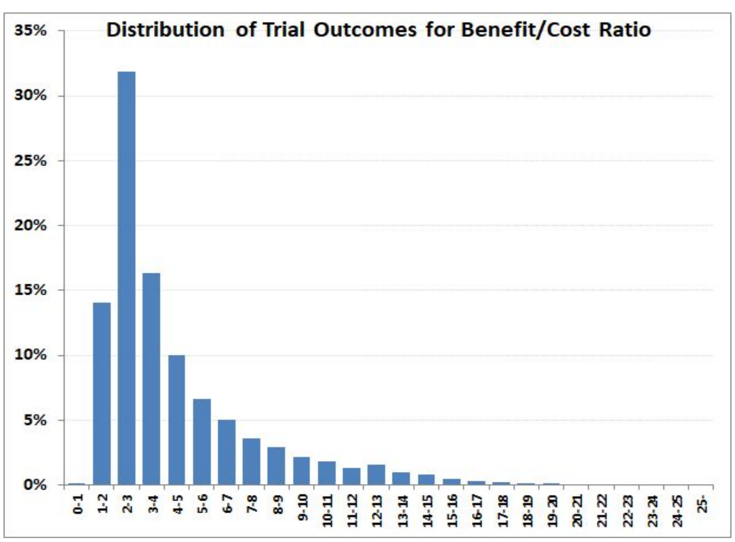


About 25% of the probability mass is in a bell-shaped distribution centered around 0 and with most of its mass between -12 and 12. About 15% of the probability mass is in a roughly square between 24 and 44. Another 30% is in a rectangle with three times the height of the previous square, spanning between 44 and 60. The last 30% comes from a rectangle roughly 2/3rds the height of the previous rectangle spanning between 60 and 84.

Problem #3:

Suppose you are responsible for computing the benefit-cost ratio of a government program whose benefits and costs are both uncertain. Experts best guess estimates are that the program's benefits will be $16 (in millions) and its costs will be $5M. Experts also agree that there is uncertainty in the world, and the benefits should really be modeled as being normally distributed with a standard deviation of $2M, and that the program costs are equally likely to take on any value between $1M and $9M. Your job is to forecast the Benefit-Cost ratio that will be reported. Simulate the resulting BC ratio 10,000 times.

a) Provide a histogram of the resulting trial outcomes. Is it bell-shaped? **No, it is not bell shaped.**



b) Give a quick intuitive explanation for what is driving the shape of the distribution.

This curve depicts the likelihood of possible outcomes from dividing a normal distribution with a mean of 16 by a uniform random distribution with a mean of 5. Dividing the means of each distribution would give a "center" value of just over 3, and this is about where most of the possible outcomes are concentrated, since most of the possible outcomes in a normal distribution are centered around the center of the bell curve. Since the cost variable can't be negative and it is extremely unlikely for the benefits component to be negative, the distribution drops off rapidly as it approaches 0 from the mode. Moving to the right, the probability trails off, representing increasingly high and unlikely values of the normally distributed benefits divided by the unlikely low values of the costs.

c) What is the expected value of the B/C ratio and what is a balanced 90% confidence interval for the BC ratio (balanced meaning 5% chance the GAO gives a number below this range, and a 5% chance it gives a number outside this range on the high side)?

E[f(X)] = $4.40 million. 90% confidence interval: ($1.74M, $11.46M)

d) What ratio would a naïve person report if they thought it was sufficient to just plug the expected revenues and expected costs into the formula for a BC ratio?

3.2

e) What is a 90% confidence interval for the expected value of the ratio? (report both the numerical value and show the equations and arithmetic used to produce this value.)

Confidence interval = 4.35 - 4.45 $ million

90% C.I. for E[X] = sample average +/- 1.645* sample standard deviation / SQRT(n),
Where n = 10,000 trials in this simulation

Conf. interval = 4.40 +/- (1.645 * 3.092)/sqrt(10,000) = (4.35, 4.45) $ million

Problem #4:
Suppose an agency or organization wants to find the solution to a problem fast, e.g., develop a COVID vaccine, and has two options. Option #1 is to fund one well-resourced effort for which the time to completion would be well-modeled by a Normal random variable with a mean of 60 days and a standard deviation of 20 days. (Don't worry for this problem about the very slim possibility that the time could be negative. We'll not be working with the extreme tails of the distributions.)

The second option is to split the funding across ten completely separate efforts, none of which would be adequately resourced, so much so that their individual performance would be twice as bad. That is, each of the ten little efforts would be expected to take twice as long and with twice as much variability in completion time independent of each other (i.e., well modeled by independent Normal distribution with a mean of 120 and standard deviation of 40). But the advantage of the second option is that the problem would be solved as soon as any one of the ten efforts solved it. That can be modeled using Excel's =Min() function.

Use Monte Carlo simulation with 10,000 trials to assess whether option #1 or option #2 is better in terms of expected time, standard deviation, and the 90% confidence interval for the outcome. You should also use the 95% CI for the mean in discussing which option offers the better expected time.

Present your answer as a table containing the requested values for option #1 and option #2. Furthermore, include 2-3 sentences interpreting your findings from the table.

Estimated value	**Option #1**	**Option #2**
Average	59.9	58.5
Std Dev	20.1	23.4
90% CI for Time to Vaccine: Lower Bound	27.2	17.3
90% CI for Time to Vaccine: Upper Bound	92.6	94.2
95% CI for Mean: Lower Bound	59.5	58.0
95% CI for Mean: Upper Bound	60.3	59.0

Each option possesses advantages, though option #2 (10 uncertain projects) is likely to develop a vaccine more rapidly than option #1. The expected time to a vaccine under option #2 is 58.5 days, with a 95% confidence interval for this average time estimate of 58.0 to 59.0 days. This is slightly faster than option #1, which has a mean of 59.9 days with a 95% CI for the mean of 59.5 to 60.3 days. Option #2 does have a greater standard deviation (meaning more uncertainty for planners/decisionmakers), but the main result of this greater spread is that the minimum time to a vaccine is much lower (10 days shorter at the 5th percentile), while the longest possible time to a vaccine is only slightly longer (1.6 days), since only the *fastest* time to completion of all the 10 projects matters.

Problem #5: This problem uses simulation to estimate the benefits of social distancing.. The motivation for this (stylized) model is thinking about the airline boarding process. If I have a seat at the front of the plane, especially an aisle seat, I want to board last to minimize the number of people who "brush" past me on the way to their seat further back in the plane. That got me to thinking how much benefit there would be if airlines reverted to the practice from the past of boarding the back of the plane first, rather than boarding primarily by groups defined by ticket prices.

I didn't take the time to build out a detailed model, but just imagine a situation in which there are twenty rows each with one seat. Row #1 is at the front, so if that person boards first, then 19 people will brush past. If that person boards 6^{th}, then 14 will brush past.

In the first version of the model, people arrive and board in random order. In the second version, the people going to Seats #11 - #20 all arrive before those going to seats #1 - #10, but arrivals are random within those bins.

The basic questions are: (1) How many "brushes" are there initially and (2) How much are they reduced by boarding in two waves?

I have provided you with a model of this situation. Connect it to a Monte Carlo simulation with 10,000 trials and report the resulting mean, standard deviation, and 90% confidence interval for the number of brushes for both boarding options.

For boarding option (1):

Mean = 94.8

Standard deviation = 15.3

90% confidence interval for number of brushes = 69 - 120 brushes

For boarding option (2):

Mean = 45

Standard deviation = 7.9

90% confidence interval for number of brushes = 32 - 58 brushes

Comment: If one has taken some probability classes, it is not hard to compute the expected value of these distributions exactly. The answers are 95 and 45, respectively. But it is not so easy to compute the standard deviation or spread of the distributions. When I tried to do it informally, I was off considerably. I erred on the side of being too low, because I under-estimated the extent to which the existence of a brush between one pair of customers is positively associated with the chance that another pair brush. So even for me, there was value in building and running the simulation for this stylized model, and a richer model that paid attention to there being aisle, middle and window seats, people boarding in groups, etc. would definitely require a simulation.

Problem #6: This problem is similar to the last, but I'll ask you to build the model. Use MC simulation with 100,000 (not just 10,000) trials to estimate the probability that two people will be in a store at the same time if they choose their arrival times independently and each stay for 15 minutes, if the store is open for 10 hours and:

Hint: model arrival times for two customers and determine if the arrival times differ by less than 15 minutes.

a) The arrival times are distributed as an asymmetric triangle RV that runs from 0 to 10 with 8 being most likely and

6.7% chance they will be in the store at the same time

b) Arrival times are uniformly distributed between 0 and 10.

5.0% chance

c) Convert the reduction between your answer for a and b into a percentage (e.g., if the probabilities were 12% and 6% you'd describe that as a 50% reduction). This is an indication of the reduction in transmission risk that could be achieved by spreading people's visits out more uniformly over the day.

25.4% reduction

d) Extend the model in part a to allow for four customer arrivals, all independent and following that same triangle distribution, and count the number of pairs of people who are in the store within 15 minutes of each other. E.g., for each pair of arrival times, see whether the absolute value of the difference in arrival times, measured in hours, is less than 0.25. If it is, count that as 1 pair who are there at the same time. Add up those 0-1 indicator variables across all pairs of people, and make that the outcome you simulate 100,000 times. Report its expected value.

0.4028

e) In part a you modeled the number of interactions when two people go to the store, and in part d you did the same with four people. Report by what multiple does doubling the number of people going to the store increase the expected number of pairs who are in the store at the same time? (Note: The expected number of pairs in part a is just equal to the probability that the one pair are there at the same time, because the number of pairs in part a is either 0 or 1. That is, the answer to part a is like a Bernoulli random variable, and the expected value of a Bernoulli RV is just p, the probability of a "success".)

5.967

Hint: The exact answer for part b is 1 – (1 – 0.025)^2 = 4.9375%.