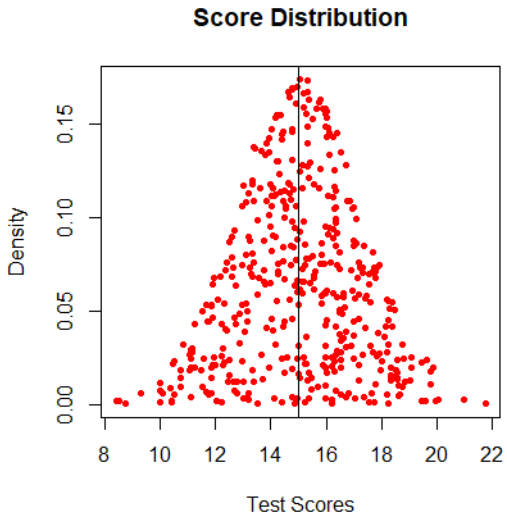
Sample Question to Demonstrate the Nature of Normal Distribution

Question 1: Suppose a form of human knowledge (e.g., use of English article *'the'*) can be measured by 15 questions that are each worth up to 2 points (i.e., partial credit is allowed). What could be a likely distribution of overall test scores (i.e., sum of 15 questions' points) for 5000 randomly selected test takers?

Find a visual answer in R (see interactive figure HERE):

```
source("https://raw.githubusercontent.com/rnorouzian/m/master/qs.r")
add.norm(n.test.taker = 5000, n.question = 15, pt.worth = 2)
```



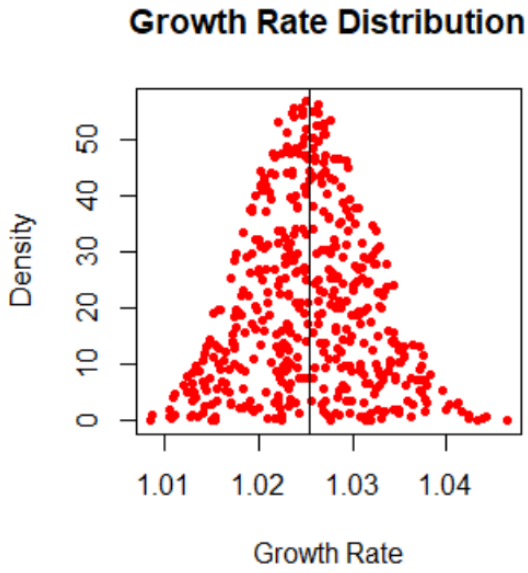
Explanation: Think of each test taker as being able to obtain any equally possible points (from 0 to 2 including any possible partial credit e.g., 0.25 etc.) on each question which when added together leads to an overall test score for that test taker. Additive phenomena (e.g., sum of 15 questions' point worth) in nature tend to cluster more heavily around their average when we study them in a population (the scientific reason is not exactly clear, see Breiman, 1968). That is, although extremely high or low realizations (e.g., high or low test scores) of that phenomenon is possible, mid-level realizations (e.g., mid-level test scores) often occur more frequently. Here because we think of each test taker as being able to obtain any points on each question on an equally possible basis, the average overall score among all test scores is simply the midpoint of the lowest (i.e., 0) and the highest (i.e., 30) possible overall test score (i.e. average overall test score = 15). Most likely, the distribution of overall test scores across all test takers is going to be a normal one centered at a mean of 15 (see figure above).

Reflection: In the absence of any other evidence, if a continuous phenomenon in psychological and educational research may consist of addition of sum measurable subcomponents, it is likely for that phenomenon to have a bell-shaped, normal distribution when studied in a population.

Question 2: Suppose the growth rate of a human skill (e.g., driving) is determined by 5 other interacting subskills (i.e., these subskills' effects multiply), each of which can contribute to the growth rate of driving by a small percentage (e.g., 1% max or by a factor of 1.01). What can be a likely distribution for the driving skill's growth rate among 5000 randomly selected subjects?

Find a visual answer in R:

```
mult.norm(n.subject = 5000, n.subskill = 5, max.small.growth = .01)
```



Explanation: Think of each interacting subskill (e.g., a, b, c, d, e) as being able to stimulate the driving skill growth rate by anything between 0% and 1% on an equally likely basis. However, because we have 5 '*interacting*' subskills, these subskills multiplicatively (i.e., a × b × c × d × e) stimulate the growth rate of the driving skill. In this case, multiplication of small contributions of the subskills to the driving growth rate (e.g., at max: [1.01×1.01×1.01×1.01×1.01]) can be approximately re-expressed as the addition of their percentage contributions (e.g., 1 + [.01+.01+.01+.01+.01]). Thus, although multiplicative, the resulting effects of the interacting subskills can be well approximated by the same principle of addition discussed in **Question 1**. As such, we again expect a normal distribution to arise. The mean of the driving skill growth rate among all drivers is, then, simply the midpoint of the lowest growth rate (i.e., rate of 1 or no change) and the highest (i.e., ~ 1.05 or 5% growth) growth rate (i.e. average driving skill growth rate = 1.025). Most likely, the distribution of driving skill growth rate for all subjects is going to be approximately a normal one centered at a mean of ~1.025 (see figure above).

Reflection: In the absence of any other evidence, if a continuous phenomenon in psychological and educational research may consist of multiplication of subcomponents of small effects, it is likely for that phenomenon to have a bell-shaped, normal distribution when studied in a population.

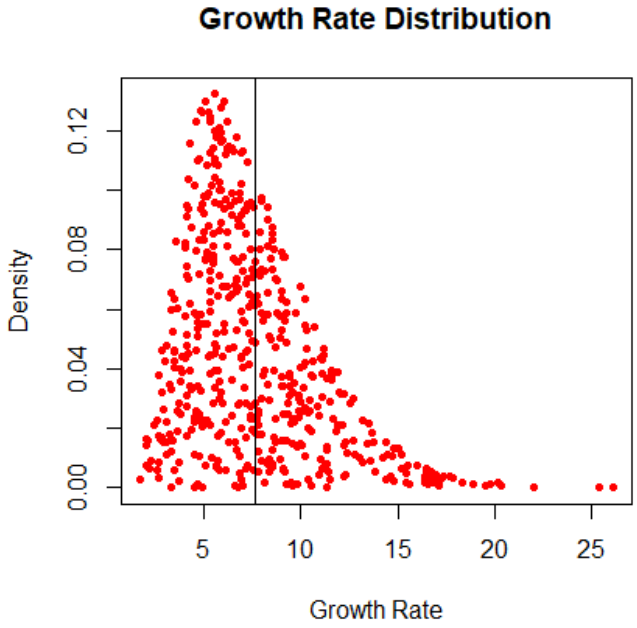
Question 3: Go back to Question 2. Now, suppose that each interacting subskill can stimulate the growth rate of driving skill by a much higher percentage (e.g., 100% max or by a factor of 2):

Part (A)

Is the distribution of driving skill's growth rate among 5000 randomly selected subjects still the same as in Question 2?

Find a visual answer for (A) in R:

```
mult.norm.free(n.subject = 5000, n.subskill = 5, max.big.growth = 1, log = FALSE)
```



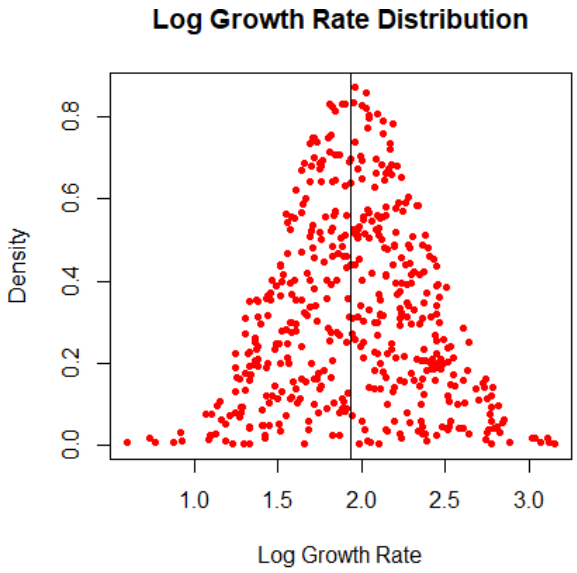
Explanation: When multiplicative phenomena consist of contributing subcomponents with large power to influence the growth of a skill, the principle of multiplication ≈ addition as in **Question 2** no longer holds, so we cannot expect a normal distribution to arise as shown in the figure above. In fact, we could have a right-skewed distribution given that most driver's growth rates are concentrated below the average. Therefore, normality or any other symmetric distribution for that matter cannot be reasonably assumed. [As we will later learn, the most likely distribution, in this case, is a gamma distribution].

Part (B)

Suppose we are open to re-scaling the growth rate contributions of our 5 interacting subskills on a logarithmic scale (`log = TRUE`) as a way to reduce the size of the growth rates (e.g., 10, 20, 25) that we saw in part (A), now what is the likely distribution of the log-scaled growth rates?

Find a visual answer for (B) in R:

```
mult.norm.free(n.subject = 5000, n.subskill = 5, max.big.growth = 1, log = TRUE)
```



Explanation: If we are willing to re-express our 5 subskills growth rate contributions to driving on a logarithmic scale, then the distribution of the log of driving growth rate will again be normal, why? Simply, because taking the log of the multiplicative (i.e., interacting) subskills growth rate contributions is equivalent to summing (i.e., addition) the log of them. For example, if the growth factor contributions for the 5 interacting subskills are: a = 1 (no contribution), b = 2, c = 3, d = 4, and e = 5, then in R you can do `identical(log(prod(1:5)), sum(log(1:5)))` to see this equivalency. Therefore, because the log transformation is equivalent to an addition phenomenon as demonstrated in **Question 1**, we expect a normal distribution to arise. Now back to part (B) question, because of the symmetric shape of the normal distribution, we would expect the mean of the normal distribution to be the midpoint of the lowest possible contributions [i.e., `sum(log(c(1,1,1,1,1)))`] from all interacting subskills resulting in 0 growth in the driving skill and the highest possible contributions [i.e., `sum(log(c(2,2,2,2,2)))`] from all interacting subskills resulting in ~3.465 of growth in the driving skill. Thus, the mean will be ~1.732 (see figure above).