SMDM Week 2 FAQs

Q1. My understating is that a box plot can be used to find if a distribution is normal or not based on the median line in the box plot. Is this correct?

Ans. *Boxplots can tell you whether a data is skewed or not depending on the median line and the length of the tails/whiskers, (one longer than the other is). While it can be used to gauge if a distribution is symmetric or not, it is not a concrete measure. Secondly, it will not tell you what distribution it is (in case it's not a normal distribution). Hence, it is a great tool for identifying outliers and skewness but not distribution.*

Q2. I have a few questions related to quantile-quantile plot:
1. What is a quantile-quantile plot? It is said that it shows the comparison between two types of data. But I am unable to interpret the meaning correctly.
2. How is it different from Probability Plot?

Ans. *The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, you would see the points forming a line that's roughly straight. For more info: https://towardsdatascience.com/what-in-the-world-are-qq-plots-20d0e41dece1. The probability plot is a graphical technique for assessing whether or not a data set follows a given distribution such as the normal distribution. The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. For more info: https://towardsdatascience.com/explaining-probability-plots-9e5c5d304703*

Q3. p=0.6 n=15 k=15 binomial = stats.binom.pmf(k,n,p)
in the above example simple binomial formula is used and in the below example cumulative frequency formula is used
binomial = stats.binom.cdf(k,n,p)
binomial not Passing=1-binomial
The question I have here is how to identify when I have to calculate cumulative frequency and when I have to calculate binomial. Can you tell me any specific word / line that I can see in the question and identify which formula to use?

Ans. *Generally, the CDF (cumulative distribution function) function is used when you want to find the cumulative probability. So, if a question has some keywords like, "less than", "at-most", "up to", etc which will try to ask cumulative probability up to some value, in that case, you have to use the CDF function. Similarly, there is a function called SF (Survival Function) which can be used in cases where the question might have keywords like, "at least", "more than/higher than", "greater than", etc. A probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.*

Q4. When we are writing answers for probability, it will be less than 1. For example, 0.46 cannot be represented as 46% because then it will become % not probability. Kindly guide.

Ans. *That's right. Probability is represented between 0 to 1 whereas percentage is represented between 0 to 100. So please read the question properly as to what has been asked. If it says probability, the answer will be expected as 0.46, but if the question says percentage, then the answer will be expected as 46 %.*

greatlearning
Power Ahead

Q5. A consumer research survey sampled 200 men and 200 women to find out whether they prefer to drink plan water or soft drink when they are really thirsty. 280 reported they prefer to drink plain water. Of the group preferring a soft drink, 80 were men and 40 were women. What is the probability that a randomly chosen man will prefer a soft drink? Ans I got was 80/120 please explain the correct Ans
Ans. *You have taken 120 which includes 40 women too. This (80/120) will give the proportion of men out of people who prefer soft drinks but not the proportion of men among total men who prefer soft drinks. The question asks "probability that a randomly chosen man will prefer a soft drink". Now, a man will be randomly chosen from the total men i.e. 200. Then, the number of men who prefer a soft drink is 80.* *Probability (man will prefer a soft drink)* *= 80/200* *= 0.4* *= 0.4*100* *= 40%*
Q6. F-test and Levene Test both are used for comparing equality of variance of 2 samples. Please explain the difference.
Ans. That is correct. Both F-Test & Levene Test are used to compare the equality of variance of samples. F-Test strictly demands and assumes that the distributions are normal. In case, sample distributions are not normal, F-Test is not used and Levene Test takes the priority there. The rationale for choosing amongst them is based on their performance if the original data are not truly normal or normal. Levene Test is chosen as the preferred method if distributions are not normal in nature
Q7. We studied in the lecture that trials in Poisson distribution is infinite. Please explain what this means
Ans *When you perform an instance of the event, for example, a coin toss, it is one instance. You toss a coin one time, it is an experiment. (Sometimes you may have read the term Bernoulli trial, it refers to this only) When you repeat the experiment n number of times, that n is called trial. Therefore, if tossing a coin 3 times were an experiment, then a trial would mean running the experiment 10 times, or 100 times, or 1000 times, even 100000000000 times. There is no upper bound to the number of trials you can do.*
Q8. What is Quantile and Quantile Plot. It will be great if you can provide me with an example of the same.
Ans. *The kth percentile of a set of values divides them so that k % of the values lie below and (100 − k) % of the values lie above.* *• The 25th percentile is known as the lower quartile.* *• The 50th percentile is known as the median.* *• The 75th percentile is known as the upper quartile.* *Also, referred to as Quantiles in statistics. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, you would see the points forming a line that is roughly straight. For more info: https://towardsdatascience.com/what-in-the-world-are-qq-plots-20d0e41dece1*