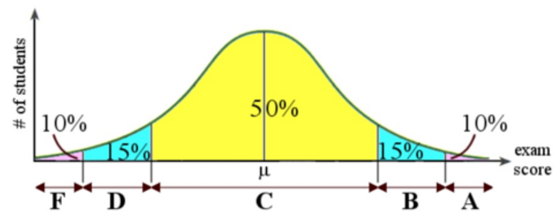
2.1 Probability Distributions

Saturday, 08 October 2022 18:39

Summary	• 3 ways to fit distributions to data 1. Trace-driven simulation 2. Theoretical distribution 3. Empirical distribution
	Probability distributions What are distributions? How to identify correct distribution of the data?
	Probability distributions • A statistical model that shows possible outcomes of a particular event or course of action as well as the statistical likelihood of each event. • What do we mean by "Grades in a course follows a normal distribution"? • What do we mean by "Sales for the next month may be uniformly distributed"?  • How do you fit distributions to data?
• In how many ways can we create models once we have the data? ○ Name them. ○ Explain.	**How to go about this?** How do we use the collected business data (sales volume, loan defaulters, Salary hikes in an organization, etc.)? 1. The data values themselves are used directly in the simulation. This is called trace-driven simulation. 2. "Fit" a theoretical distribution to the data (and check whether that "fit" is good!). 3. The data values could be used to define an empirical distribution function in some way. • In point 1, we don't need to fit any distribution. We directly use the data in our analysis. • Theoretical distributions are the distributions we've already studies, eg, normal, uniform distribution, etc. • In point 2, we first fit a distribution in the data. Check how good is the fit. • Point 3: If the data doesn't fit any theoretical distribution, then instead of trying to fit already available distributions, we create our own distributions. Those distributions are called **empirical distributions.** And then we use this distribution in our future analysis.
• What are the essential building blocks of empirical distribution?	**What are these empirical distributions?** • Using the data, we build our own distributions. • How does one build a distribution? • Essential building blocks: Define the density/distribution functions. Estimate the parameters (mean, standard deviation, etc.) • When the professor says distribution function, he means CDF.
	Empirical distributions For ungrouped data: Let $X_{(i)}$ denote the ith smallest of the X_j's so that: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

	$$F(x)=\begin{cases}0 & \text{if} \quad x<X_{(1)} \\ \dfrac{i-1}{n-1}+\dfrac{x-X_{(i)}}{(n-1)\left(X_{(i+1)}-X_{(i)}\right)} & \text{if} \quad X_{(i)} \le x < X_{(i+1)} \quad \text{for} \quad i=1,2,\cdots,n-1 \\ 1 & \text{if} \quad X_{(n)} \le x\end{cases}$$ • Given a CDF, we already know how to find the density function.
	Empirical distributions For grouped data: • Suppose that n X_j's are grouped in k adjacent intervals $[a_0,a_1)$, $[a_1,a_2)$,…$[a_{k-1},a_k)$ so that jth interval contains n_j observations. $n_1+n_2+\ldots n_k=n$. • Let a piecewise linear function G be such that $G(a_0)=0$, $G(a_j)=(n_1+n_2+\ldots n_j)/n$, then: $$G(x)=\begin{cases}0 & \text{if } x<a_0 \\ G(a_{j-1})+\dfrac{x-a_{j-1}}{a_j-a_{j-1}}\left[G(a_j)-G(a_{j-1})\right] & \text{if } a_{j-1}\le x<a_j, j=1,2,\cdots k \\ 1 & \text{if } a_k \le x.\end{cases}$$ • We have k intervals, and in each interval we have $n_1, n_2, \ldots, n_k$ values. • $G(a_j)$ is proportional to the observations up to that point/interval.
• When do we use trace driven simulation? ○ What are its drawbacks? ☐ Approach 1 is not very clear to me	**The three approaches…** • Approach 1 is used to validate simulation model when comparing model output for an existing system with the corresponding output for the system itself. • Two drawbacks of approach 1: simulation can only reproduce only what happened historically; and there is seldom enough data to make all simulation runs. • Approaches 2 and 3 avoid these shortcomings so that any value between minimum and maximum can be generated. So approaches 2 and 3 are preferred over approach 1. • If theoretical distributions can be found that fits the observed data (approach 2), then it is preferred over approach 3. • Approach 1: You have the output, and you want to validate if the output is correct or not. You push the already available data into your model and your model generates an output and you compare that output with the reality (the existing system, what happens in future) and check whether it matches. • So trace-driven simulation is used to validate a model that you already may have built using some approach. • Problem with approach 1 is you're going to test the model only with the data you already have. This may not be enough. • What if the data was collected in a certain circumstance, and as the circumstances change will not give you a fair values.
• What are the drawbacks of empirical approach?	**Approach 3 v/s Approach 2** • Empirical distribution may have some irregularities if small number of data points are available. Approach 2 smoothens out the data and may provide information on the overall underlying distribution. • In approach 3, it is usually not possible to generate values outside the range of observed data in the simulation. • If one wants to test the performance of the simulated system under extreme conditions, that can not be done using approach 3. • There may be compelling (physical) reasons in some situations for using a particular theoretical distribution. In that case too, it is better to get empirical support for that distribution from the observed data.