# Initial Posts

|  |
| --- |
| **Probability Mass Function (PMF)**  A Probability Mass Function (PMF) is a way to represent a distribution and is very similar to a frequency distribution where the total for each value is calculated. Instead of frequency however, the values are normalized by calculating the probability of each value within the sample. “Pmf and Hist objects are similar in many ways; The biggest difference is that a Hist maps from values to integer counters; a Pmf maps from values to floating-point probabilities.” PMF’s are good at representing discrete variables but are sensitive to the number of distinct values within the variable. The larger the number of distinct values within the variable, the harder it is to interpret the distribution using PMF.  Reference:  Downey, Allen B.. Think Stats: Exploratory Data Analysis (pp. 43-44). O'Reilly Media. Kindle Edition. |
| **Class Size Paradox**  The class size paradox is a paradox where the experienced size of a group is overestimated when group members are asked about it’s group size. In the example of the education system, students will report that the average class size is larger than the actual average. This is because there are more students in larger classes which means that more students will experience larger classes than those students who experience smaller ones.  The class size paradox is not restricted only to the education system but to anything that has different sized groups within a parent group. For example, in the 1991 paper by Scott Feld “Why Your Friends Have More Friends Than You Do”, Feld explains how this paradox relates to a person’s experience regarding how many friends their friends have and how the experienced number of friends is overestimated. Finally! Scientific proof that I’m not anti-social, I’m just overestimating how many friends I need to have! |
| **Percentile vs. Percentile Rank**  Well…looks like I’ve been mixing up the definition of Percentile vs. Percentile Rank pretty much since forever. Nobody every told me otherwise so I blame the world for not correcting me. Nevertheless, after this week, I’ve now learned the difference which is below:   * Percentile Rank: 2 different definitions with no universal/definitive definition.   + % of values that are less than a specified value   + % of values that are less than **or equal to** a specified value * Percentile: 2 different definitions with no universal/definitive definition.   + Given a specified percentile rank, the minimum value that has a percentile rank that is greater than the percentile rank you are evaluating.   + Given a specified percentile rank, the minimum value that has a percentile rank that is greater than **or equal to** the percentile rank you are evaluating. |

|  |
| --- |
| **Option 6**  Q: Imagine you work for an organization that is interested in getting more value out of the data. What steps would you suggest to help them start finding additional insight on their data?  A: I could be misinterpreting the point of the question but this task seems rather vague. It would be the equivalent of going to a psychiatrist and saying, "I'm interested in being a happier and more positive person". The psychiatrist would probably start by saying, "Well...obviously you think you're not as happy or positive as you should be. Why is that?". That would probably be the first step I would take in the task of getting more value out of data in the system. Obviously they think that there is room for improvement and I would ask why they think that. |
| **Cumulative Distribution Function (CDF)**  "The CDF is the function that maps from a value to its percentile rank." (Downey 2015). It is a great way to interpret both the central tendency of the data as well as the spread. For me personally, the CDF plot was somewhat hard to interpret at first but once I got used to it I actually love them and find that they show a lot of information in a simple yet effective way.   Downey, Allen B.. Think Stats: Exploratory Data Analysis (p. 61). O'Reilly Media. Kindle Edition. |

# Replies

|  |
| --- |
| Good post Jahedur. I like how you simplified the different ways to interpret PMF. When I first saw this week’s assignments, I thought PMF was going to be a crazy topic but after learning more about it it's really just looking at the percentage a value within a dimension makes up. |
| I like how the link provided gives you the formula to calculate the overestimation of the students observations. Thanks Saima. |
| Good link Sameer. The article does a great job explaining how histograms can be interpreted very differently based on the number of bins. I also liked how it explained how to build the other graphs in Python and the pros and cons of each. Intuitively, I find it easier to interpret PMF and KDE plots. CDF plots at first were a bit harder to interpret but once it "clicks" I found that they are very intuitive and provide a lot of information. |
| I could be wrong (likely) but I thought indexing was the generation of identifiers for rows/columns, not necessarily selecting rows/columns? |
| Good post Madeleine. I've actually really enjoyed this topic this week and it has made me start to wonder how much other stuff I've been overestimating in my life. |