# Initial Posts

|  |
| --- |
| **Descriptive Statistics**  Unlike Inferential Statistics which is used to make inferences, or guess, what future data will be based on existing data, Descriptive Statistics is the branch of statistics that summarizes existing data. From the branch of Descriptive Statistics there are 2 measures commonly used (see below).   * Measures of Central Tendancy   + Mean   + Median   + Mode * Measures of Dispersion (or Variability)   + Variance   + Standard Deviation   + Range   + Quartiles   + Skewness   + Kurtosis |
| **What is Pandas?**  Pandas is a module that is used primarily for data analysis and allows you to do all sorts of things to data in the form of a data frame. A data frame can be thought of as a table in an Excel file which a lot of people are familiar with. Once data has been turned into a data frame in Pandas, you can manipulate the data, analyze the data, join multiple data frames, and much more. Pandas also allows you to connect easily to multiple different data sources right from Python for example, you can connect to Excel files, text files, databases, etc. |
| **Python Dictionaries & Pandas Methods**  Python dictionaries are used to store data in key:value pairs. The also take on the below attributes.  They are mutable meaning that they can be modified after creation.  Can contain multiple different data types both for the keys and the values  Are ordered meaning that the key:value elements within the dictionary are indexed.  Do not allow duplicate keys but do allow duplicate values for the values. If a duplicate key:value is added to a dictionary where the key already exists, the key will stay the same but the value will be overwritten with the new value.  One popular usage for dictionaries would be frequency counts (see below).  Python  nl **=** '\n'  t **=** '\t'  e **=** '='  list1 **=** **[**'a'**,**'a'**,**'a'**,**'a'**,**'b'**,**'b'**,**'b'**,**'c'**]**  freq\_dict **=** **{}**  **for** elem **in** list1**:**  **if** elem **in** freq\_dict**.**keys**():**  freq\_dict**[**elem**]** **+=** 1  **else:**  freq\_dict**[**elem**]** **=** 1    **print(**f"Key{t}Value{nl}{e**\***13}"**)**  **for** key**,** value **in** freq\_dict**.**items**():**  **print(**f"{key} {t} {value}"**)**  One of the more popular methods used in pandas is the data frame method. It takes data from a source and converts it into a data frame table where it can be further manipulated and analyzed in Python. In the example below I'll create a data frame from a Python dictionary.  Python  **import** pandas **as** pd  df\_dict **=** **{**  'ID'**:[**1**,**2**,**3**,**4**]**  **,** 'Name'**:[**'Patrick'**,**'Leigha'**,**'Blake'**,**'Gizmo'**]**  **,** 'Sex'**:[**'M'**,**'F'**,**'M'**,**'F'**]**  **}**  pd\_df **=** pd**.**DataFrame**(**df\_dict**)**  **print(**pd\_df**)** |
| **Method Resolution Order (MRO)** Is the order in which Python will use an attribute or method of an object. Regarding building a class in Python and subclasses within subclasses which inherit attributes and methods of preceding classes. This can become rather complex and hard to interpret...well it was for me anyway . I ran across a great video on YouTube which explains the process and clearly explains the order of operations that Python takes to associate an attribute or method with an object of a class with multiple inherited classes.  [[Language skills Python] MRO Method Resolution Order - Advanced [Tutorial] - YouTube](https://www.youtube.com/watch?v=7XXR3s_fYU8) |
| **Codebook – Think Stats**  I found that the codebook link for the ThinkStats2 repo was not working but was able to find it on the CDC website (link below).  [2003 National Survey of Family Growth, Female Pregnancy File (cdc.gov)](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NSFG/Cycle6Codebook-Pregnancy.pdf) |

# Replies

|  |
| --- |
| *Regarding EDA in the Workplace*  Madeleine Sharp, sure no problem! I'm actually in the middle of a project at work now that is going through an issue like this. I can't really go into the specifics but basically there are medications that now require a certain question to be answered when it is ordered which was currently not in the ordering process. Each specific order (there are thousands..) had to be updated so that when a specific value was selected during order entry, a question would pop up on the user's screen which was made as a hard stop so they were required to answer it to continue ordering the medication. There are of course certain scenarios where this can be circumvented if the provider fails to enter correct information on the order but a report was created to show orders where, in theory, the questions should be generated but were not so that we can keep an eye on it. After the order questions were added to our ordering process, the training team will then send out a mass communication advising of the new workflow and offer training if needed. |
| *Regarding Cohen’s d and effect sizes*  Hi Arti, I like the technical explanation and formulas provided! I also like how you mentioned that although something can be statistically significant that based on Cohen's d the effect size may be small, medium, or large. This meaning that although an effect does exist and rejects the NULL hypothesis that no effect is present, that the effect can be classified as small. One important thing to note is that although an effect size may be small, it does not necessarily mean that it is unimportant. The importance of an effect size of course depends on what you're measuring, expected effect size, confidence intervals, sample size, etc. Take for example Olympic swimmers. The average swim speed may be only fractions of a second different from swimmer to swimmer so an effect size considered small < 0.2 may actually be large regarding the avg. swim speed of Olympic swimmers. |
| *Regarding Outliers*  Good post Madeleine! Outliers seem to be a heavily debated topic in the Data Science community. The best answer on how to handle them seem to be  it depends! Shocker...  Some things that could affect how you handle outliers might be:   * How is the data distributed? * What's the sample size? * What predictive model your using? Is it robust to outliers? * Is there some workflow causing the outliers that should be addressed? * Does it make sense for the outlier to be there? * Are they being caused by an error in the application storing the data?   Again, this is a topic that is heavily debated and I personally am torn myself on how to address them. I guess my first response would probably be to investigate how the outlier got there in the first place and then get insight from others within my organization on how the outliers should be addressed. |
| *Regarding EDA in the Workplace*  Madeleine, yes, a doozy indeed! In this scenario we did not backfill values onto the historical orders. There are millions of order records in our system and due to the sensitivity of healthcare data, it would probably be unethical to backfill the data and 'assume' what value a provider "would have" entered on the historical orders. This means that historical orders will not have values for these questions but future ones will. |
| Hi Jeff, I like your explanation of a histogram vs. a bar chart. Yes, they are pretty similar. The main different being that histograms are used to measure frequency for a continuous variable whereas a bar chart is used for categorical variables and is not limited to just measuring frequency of instances. |
| *Regarding Think Stats Codebook*  No problem Amelia! Also, if you're curious about the page the PDF link was found on, I'll post the page below:  [NSFG - Cycle 6 (cdc.gov)](https://www.cdc.gov/nchs/nsfg/nsfg_cycle6.htm) |