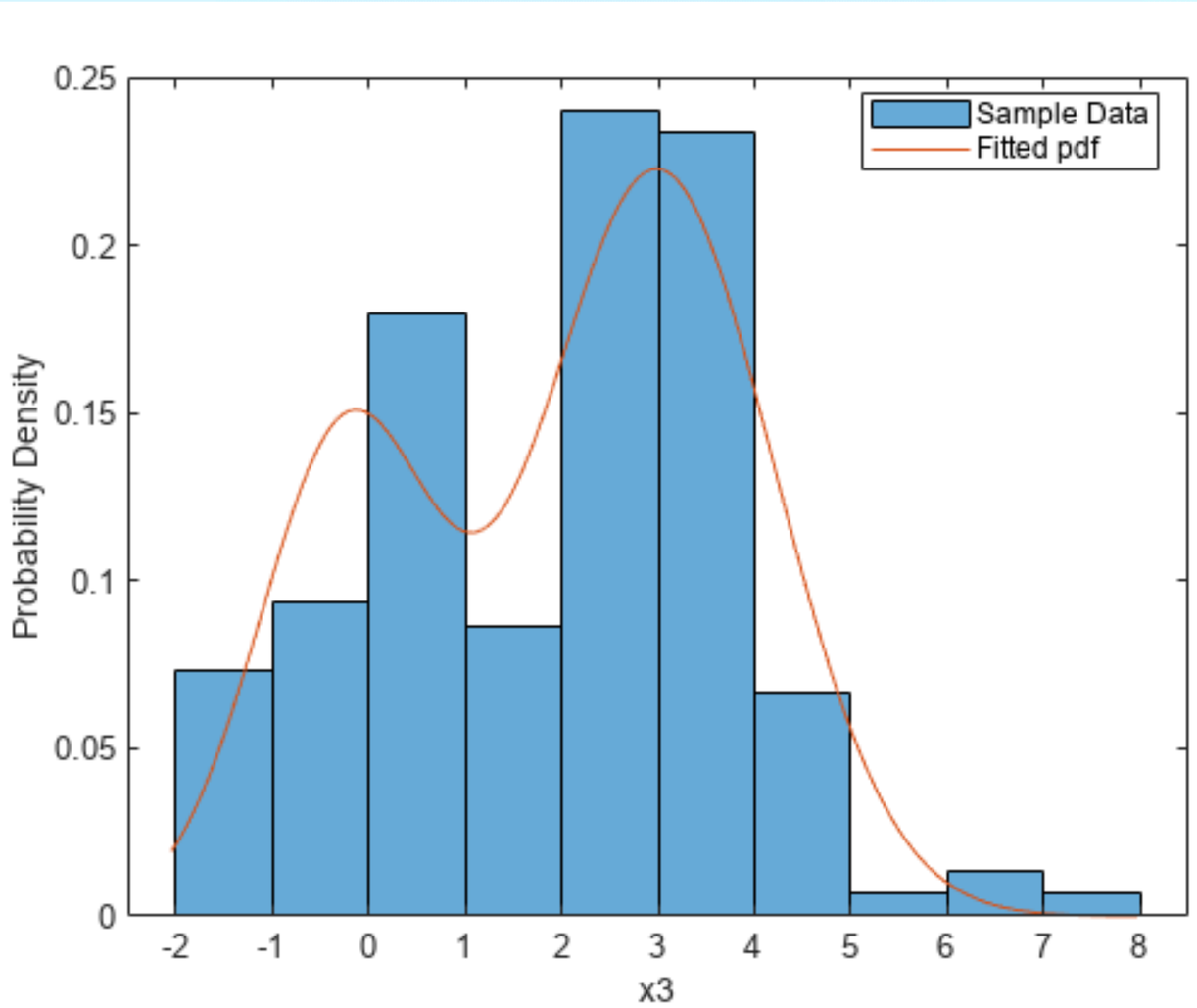
Resampling

PHYS 2601

No Generative Model

- We have discussed how we can use a probability density function to create ‘theoretical’ samples of data which we can then test our analysis on.
- But what if we don’t have a probability density function for our data?
- This might occur because our data doesn’t look like any probability density function we can think of, or because we don’t have enough data or a theory to favor a particular function.



Resampling

- If we don’t have a generative function that describes our data all we can do is use the data itself to describe our data.
- Under the assumption (often not correct) that we have enough data that our sample approximates the true generating function we can simply use the sample itself as a means of generating *”new”* samples.
- This procedure goes under the general name of ‘resampling’ meaning one draws a new sample from the data. There are a number of specific techniques used in resampling including bootstrap, jackknife and cross-validation.

with and without Replacement

- One of the most important concepts in resampling is the idea of replacement.
- This simply means if one data value is chosen from my data set do I now continue on with that data point gone (without replacement) or do I still include that data point in my data set (with replacement).
- For example, if we are sampling a 6 sided die and we role a 6 then on our next role we could still get a 6. The 6 is replaced in our choices.
- On the other hand if we have a bag with 6 different colored M&Ms in it and we pull out the green one, then the next person randomly pulling out an M&M could only get the colors not green. There is no replacement, unless we put the green M&M back in the bag.

Bootstrap

- The bootstrap method is an incredibly powerful way to resample data and derive errors on your statistics.
- It almost seems magical which is how it got its name. Pulling yourself up by your bootstraps, which should be impossible to do.

Bootstrap

- For this method, one simply resamples your data randomly **with replacement**. One draws the same number of data points as your sample has.
- So for example if your data was [1, 2, 3, 4, 5 ,6], then you would randomly draw from these values six times. Your resamples could look like [1,1,4,5,6,6], [2,3,2,5,6,3] and [3,3,3,4,5,6].

Bootstrap

- If you did this 100 times then one could calculate the mean and standard deviation on each of your 100 samples and determine the range of 90% or the means and standard deviations.
- Thus one can determine error bars from just the data, without any assumptions about the probability density function or generating function.
- Better yet, one can use the bootstrap method and a generating function and test if they give comparable results. This way one has confidence that your generating function is correct.

Bootstrap

- The one weakness of the bootstrap method is you have to believe you have a fair sample of the underlying distribution.
- If your data set is large this is a reasonable assumption, if it is small almost certainly not.
- Even if large you are likely to have a few outlier data points that will skew your bootstrap resampling.

Jackknife

- Another way to resample your data is called the jackknife method.
- In this case one simply removes one data point and performs your analysis again. This is done until you have gone through all the data points.
- Note this is technically a delete-one jackknife, one could also delete a block of data instead.

Bootstrap vrs. Jackknife

- Generally the bootstrap method is preferred to the jackknife because it is simpler to implement. As we have seen choosing random numbers is vary easy with a computer.
- The jackknife has a few advantages that one should consider when deciding between the two.

    1. The bootstrap is random, while the jackknife always gives the same answer. If having multiple people checking an answer or trying to give a definitive result, one might want to consider the jackknife.

    2. The jackknife can be very good and identifying outliers. If your results suddenly shift when one data point is removed, this could make you suspicious of that data point.

    3. This is when doing jackknife on blocks of data become interesting. For example if one has collected data on different days, one could jackknife removing data from one day. Then one could find one day where the data was very off and decide not to use it. This is when the jackknife is much better than randomly selecting the data.

Cross - Validation

- Ideally if we analyzed any data and came to some conclusion, the best test of our conclusion, from the scientific method, is that we would predict the results of a new experiment and then test our model. If our predictions are statistically in agreement with the new data we would say that new experiment validated our model.

- If we can’t do this, an alternative based on resampling is to validate our data with our own data. To do this we can randomly remove a fraction of our data (like 20%-25%) perform our analysis on the remaining data and then validate our model on the data that was removed.

Implementation of Bootstrapping

- Bootstrapping is very easy to implement in python. Let us say we have a data set of size N containing x and y values. Then to randomly resample with replacement all we have to do is the following:

```
rng = np.random.default_rng()
ids = rng.integers(low=0,high=N,size=N)
new_x = x[ids]
new_y = y[ids]
```

- Note that if using a data frame one can simple sample it keeping all rows together

```
df_new = df.sample(n=N)
```

Randomization

- Another application of resampling is to shuffle a variable so that it looses its correlation.
- For example, image we have x and y data that we think make a line, but we worry maybe it is just because our x or y values are funny.
- We can test this by shuffling our y data randomly. That is instead of our original (x,y) pairs, we randomly reorder y so to make new (x,y) pairs. Then we can perform our analysis on the new data pairs, which shouldn’t retain any correlation between x and y.
- If we still find evidence of a line, or other statistic then we know it is just he y values by themselves and not their correlation with the x values that is giving rise to this result.

Resampling unbalanced data sets

- What if we think our data set is very unbalanced? This happens for example in random polling of people.
- One can try to rebalance the dataset by random sampling according to some population statistics.
- This is always done for example with telephone polls, the pollsters try to correct their sample to the population average by weighting by race, education level, etc.
- This can be dangerous if ones estimate of the ‘correct’ population has any errors, but for many situations it is possible to get a good estimate of an entire population (e.g. a census) and hard to get other data (people to answer phones).
- One should consider balancing the data if one thinks the sample one has is highly biased or unbalanced, otherwise ones conclusions will also be biased.