Sampling and Empirical Distributions

References:

Chapter 10 - Sampling and Empirical Distributions

https://www.inferentialthinking.com/chapters/intro

Overview

- Deterministic sample vs Probability sample
- Empirical Distribution
- Sample of Convenience
- Systematic Sample
- Law of Averages

Sampling

- Deterministic sample:
  - Sampling scheme doesn’t involve chance

- Probability sample:
  - A *population* is the set of all elements from whom a sample will be drawn.
  - A *probability sample* is one for which it is possible to calculate, before the sample is drawn, the chance with which any subset of elements will enter the sample.
  - Not all individuals have to have equal chance of being selected.

A Random Sampling Scheme

- For example, suppose you choose two people from a population that consists of three people A, B, and C, according to the following scheme:
    - Person A is chosen with probability 1.
    - One of Persons B or C is chosen according to the toss of a coin: if the coin lands heads, you choose B, and if it lands tails you choose C.

- This is a probability sample of size 2. Here are the chances of entry for all non-empty subsets:

A : 1
B : ½
C : ½
AB : ½
AC : ½
BC : 0
ABC : 0

A Systematic Sample

- Imagine all the elements of the population listed in a sequence. One method of sampling starts by choosing a random position early in the list, and then evenly spaced positions after that. The sample consists of the elements in those positions. Such a sample is called a systematic sample.

- Here we will choose a systematic sample of the rows of top. We will start by picking one of the first 10 rows at random, and then we will pick every 10th row after that.

Random Samples Drawn With or Without Replacement

- The first is random sampling with replacement (default behavior of *np.random.choice* when it samples from an array).

- Other is "simple random sample", is a sample drawn at random *without* replacement. Sampled items are not replaced in the population before the next item is drawn

Sample of Convenience

- Example: Sample consists of first ten people who walk by a street corner.
- Just because you think you’re sampling “at random”, doesn’t mean you are.
- If you can’t figure out ahead of time
    - what’s the population
    - what’s the chance of selection, for each group in the population

  then you don’t have a random sample.

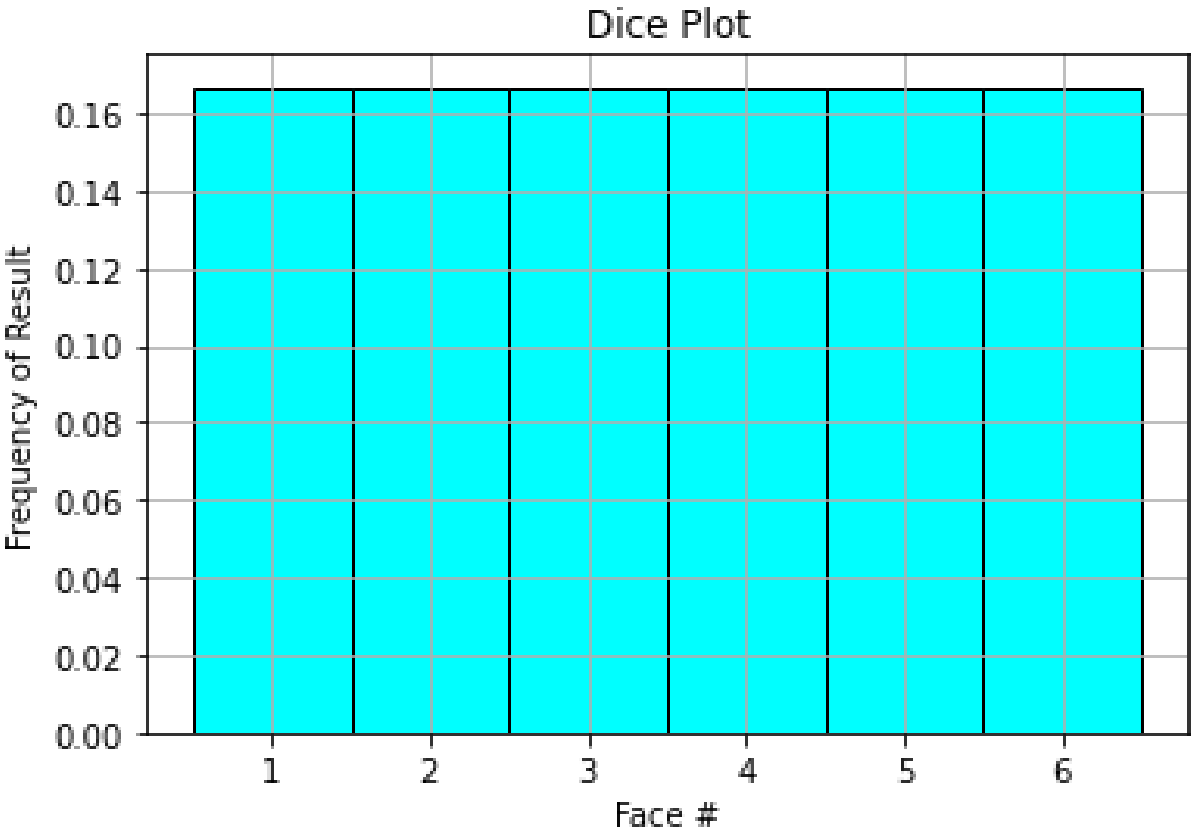
The example discussed above is a ‘sample of convenience’.

Empirical Distribution

- Based on observations
- Observations are typically from repetitions of an experiment
- ‘Empirical Distribution’
  - All observed values
  - The proportion of repetitions that produced each value
- Consider a simple experiment: rolling a die multiple times and keeping track of which face appears.

A Probability Distribution

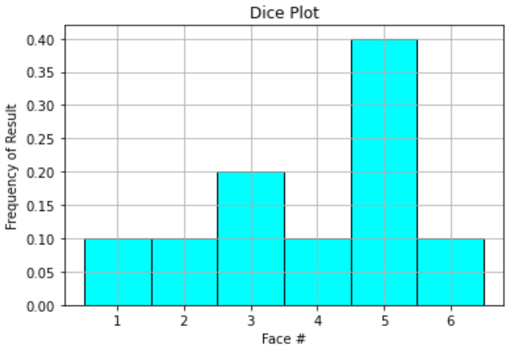
- The histogram below helps us visualize the fact that every face appears with probability 1/6.
- We say that the histogram shows the distribution of probabilities over all the possible faces.
- Since all the bars represent the same percent chance, the distribution is called uniform on the integers 1 through 6.



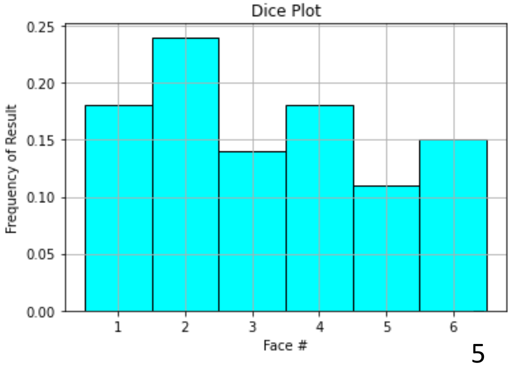
Empirical Histograms

- Here is an *empirical histogram* of 10 rolls. It doesn't look very much like the probability histogram discussed earlier.
- When the sample size increases, the empirical histogram begins to look more like the histogram of theoretical probabilities.
- As we increase the number of rolls in the simulation, the area of each bar gets closer to 16.67%, which is the area of each bar in the probability histogram.

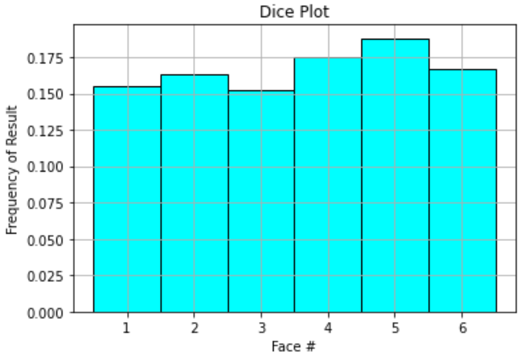
#Rolls = 10



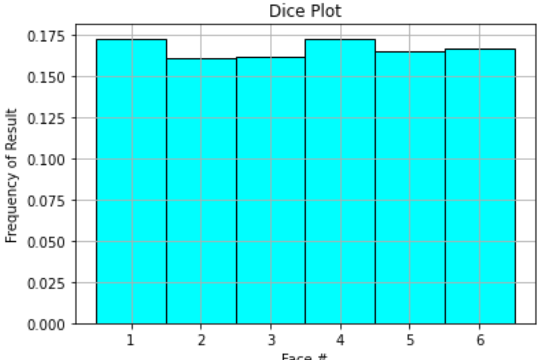
#Rolls = 100



#Rolls = 1000



#Rolls = 10000



Law of Averages

- If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event.

- As you increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6.

Summary

- A *probability sample* is one for which it is possible to calculate, before the sample is drawn, the chance with which any subset of elements will enter the sample.

- An empirical distribution is based on observations where observations are typically from repetitions of an experiment

- Law of Averages: If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event.