Sampling from a Population

References:

Chapter 10 - Sampling and Empirical Distributions

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

Overview

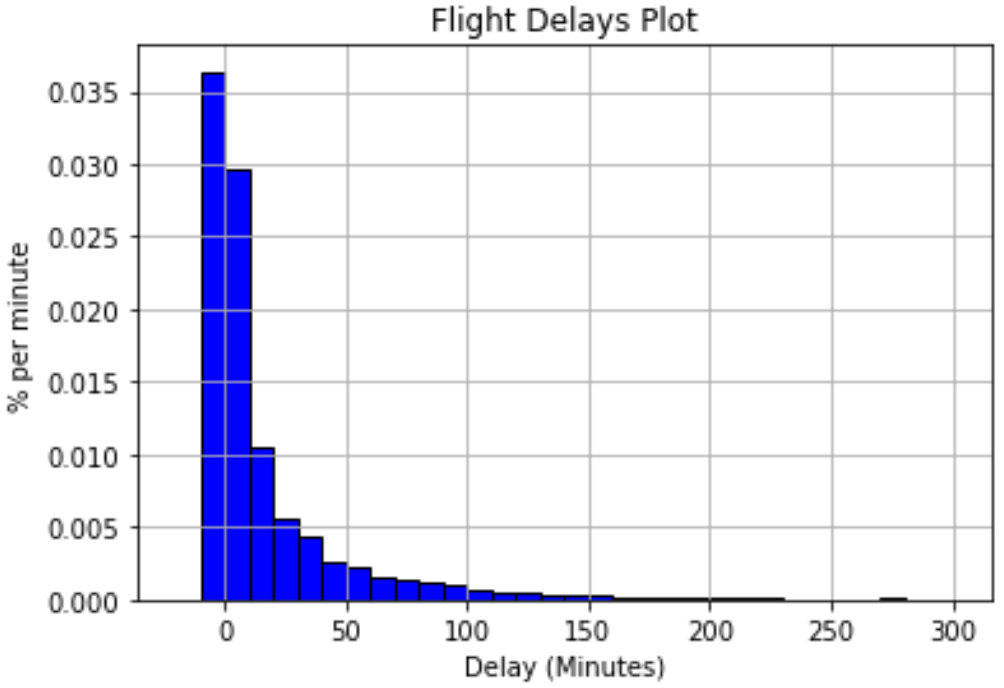
- **Sampling from a Population**
- **Convergence of the Empirical Histogram of the Sample**
- **Inference**
- **Empirical Distribution of a Statistic**
- **Simulating a Statistic**

Sampling from a Population

- If the sample size is large, then the empirical distribution of a random sample resembles the distribution of the population, with high probability.

- There are 13,825 rows, each corresponding to a flight. The columns are the date of the flight, the flight number, the destination airport code, and the departure delay time in minutes. Some delay times are negative; those flights left early

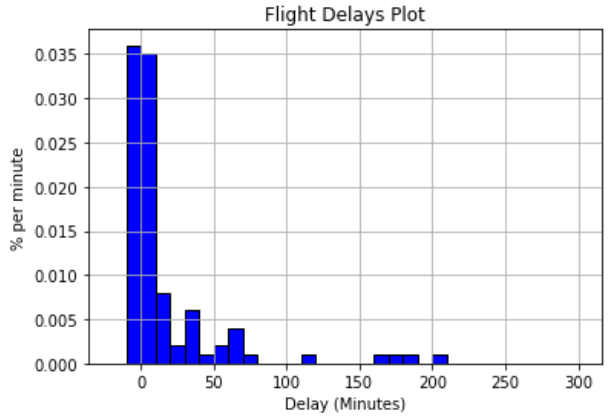
- The height of the [0, 10) bar is just under 3% per minute, which means that just under 30% of the flights had delays between 0 and 10 minutes. That is confirmed by counting rows: matching this criteria



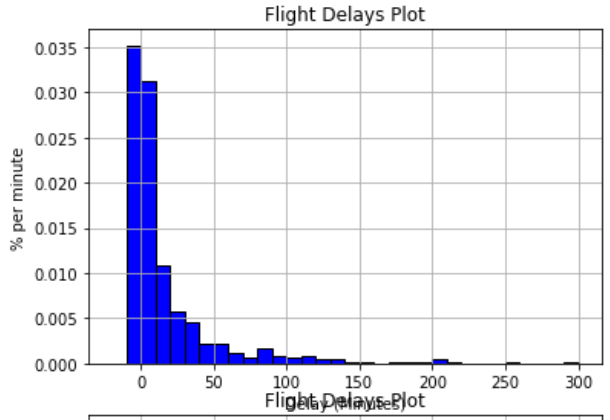
Convergence of the Empirical Histogram of the Sample

- For a large random sample, the empirical histogram of the sample resembles the histogram of the population, with high probability.
- This justifies the use of large random samples in statistical inference.
- The idea is that since a large random sample is likely to resemble the population from which it is drawn, quantities computed from the values in the sample are likely to be close to the corresponding quantities in the population.

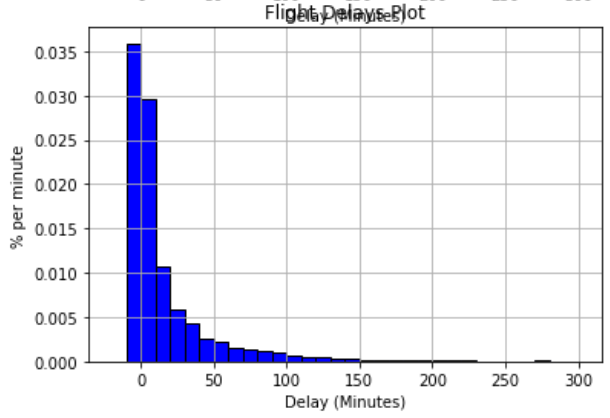
#Flights = 100



#Flights = 1000



#Flights = 10000



Inference

- **Statistical Inference:** Making conclusions based on data in random samples

- **Example**: Use the data to guess the value of an unknown number

Create an **estimate** of the unknown quantity

Empirical Distribution of a Statistic

- **Parameter:** A number associated with the population
  - In a population of voters, what percent will vote for Candidate A?
  - In a population of Facebook users, what is the largest number of Facebook friends that the users have?
  - In a population of United flights, what is the median departure delay?
    - Among all the flights in united, the median delay was 2 minutes. That is, about 50% of flights in the population had delays of 2 or fewer minutes:

Empirical Distribution of a Statistic

- **Statistic**
  - A number calculated from the sample
  - A *statistic* is any number computed using the data in a sample.
  - Therefore, the sample median in this example, is a *statistic*.

In this flight delay example, often the median for a large sample is equal to 2, the same value as the population parameter.

But sometimes it is different.

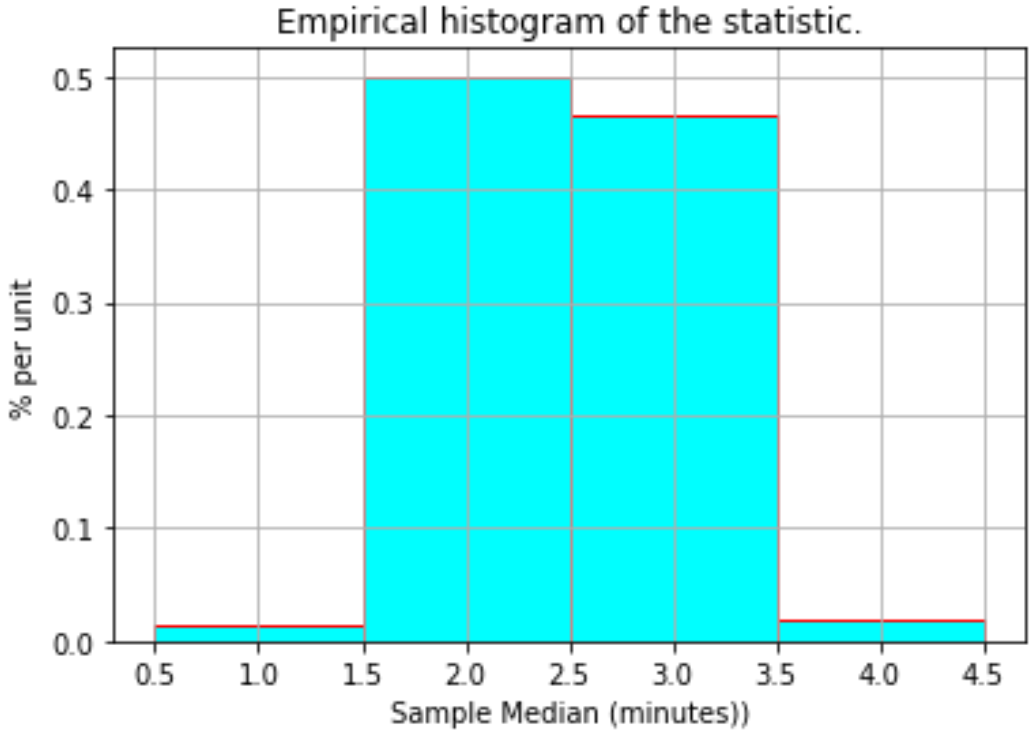
Simulating a Statistic

Step 1: Decide which statistic to simulate.

Step 2: Write the code to generate one value of the statistic.

Step 3: Decide how many simulated values to generate

Step 4: Write the code to generate an array of simulated values



Probability Distribution of a Statistic

- Values of a statistic vary because random samples vary
- "Sampling distribution" or "probability distribution" of the statistic:
  - All possible values of the statistic, and all the corresponding probabilities
- Can be hard to calculate
  - Either have to do the math
  - Or have to generate all possible samples and calculate the statistic based on each sample

Empirical Distribution of a Statistic

- Based on simulated values of the statistic
- Consists of all the observed values of the statistic and the proportion of times each value appeared
- Good approximation to the probability distribution of the statistic if the number of repetitions in the simulation is large

Summary

- If the sample size is large, then the empirical distribution of a random sample resembles the distribution of the population, with high probability.

- **Statistical Inference** involves **m**aking conclusions based on data in random samples where the *parameter* associated with a *population* is estimated by determining the *statistic* associated with large random sample drawn from the *population*.

- We know that by the Law of Averages, the empirical histogram of the statistic is likely to resemble the probability histogram of the statistic, if the sample size is large and if you repeat the random sampling process numerous times.

- Simulating random processes repeatedly is a way of approximating probability distributions *without figuring out the probabilities mathematically or generating all possible random samples*.