Lesson 09 - objectives

- appreciate the “bell” curve (normal distribution)
- what is the Empirical Rule? (68 , 95 , 99.7)
  - how to use it?
- understand the Central Limit Theorem
  - what is the implication?

till ~320pm

- ~ 20 mins
- Two data files on IVY:
  - temperatures (use now) and Boston Marathon (use later)
- Read the temperature csv file
- plot a histogram for the entire population. (using the daily high)
- sample a size of 100 → then plot a histogram for the sample

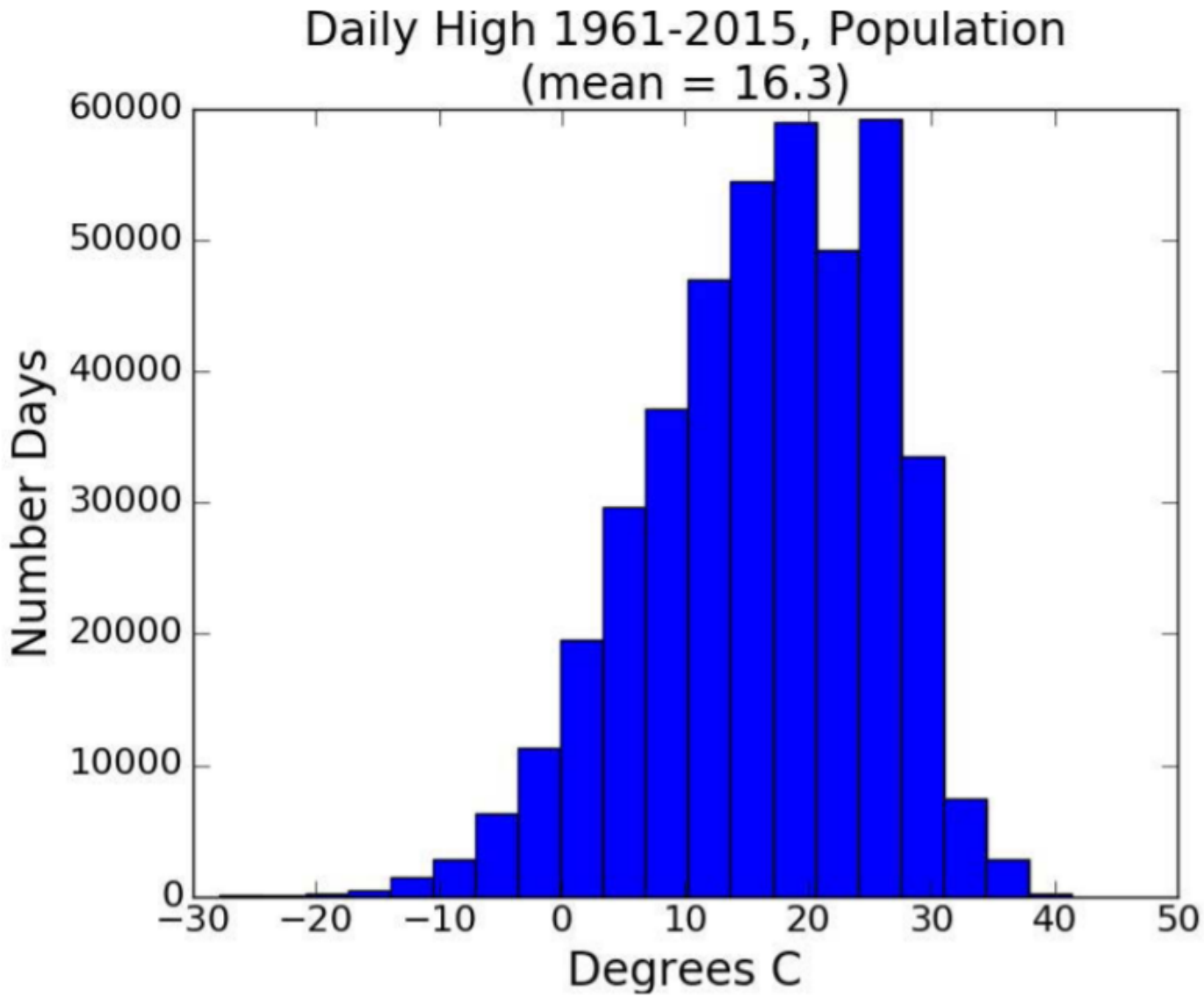
Data

- From U.S. National Centers for Environmental Information (NCEI)
- Daily high ~~and low~~ temperatures for
  - 21 different US cities
    - ALBUQUERQUE, BALTIMORE, BOSTON, CHARLOTTE, CHICAGO, DALLAS, DETROIT, LAS VEGAS, LOS ANGELES, MIAMI, NEW ORLEANS, NEW YORK, PHILADELPHIA, PHOENIX, PORTLAND, SAN DIEGO, SAN FRANCISCO, SAN JUAN, SEATTLE, ST LOUIS, TAMPA
  - 1961 – 2015
  - 421,848 data points (examples)
- Let's use some code to look at the data

New in Code

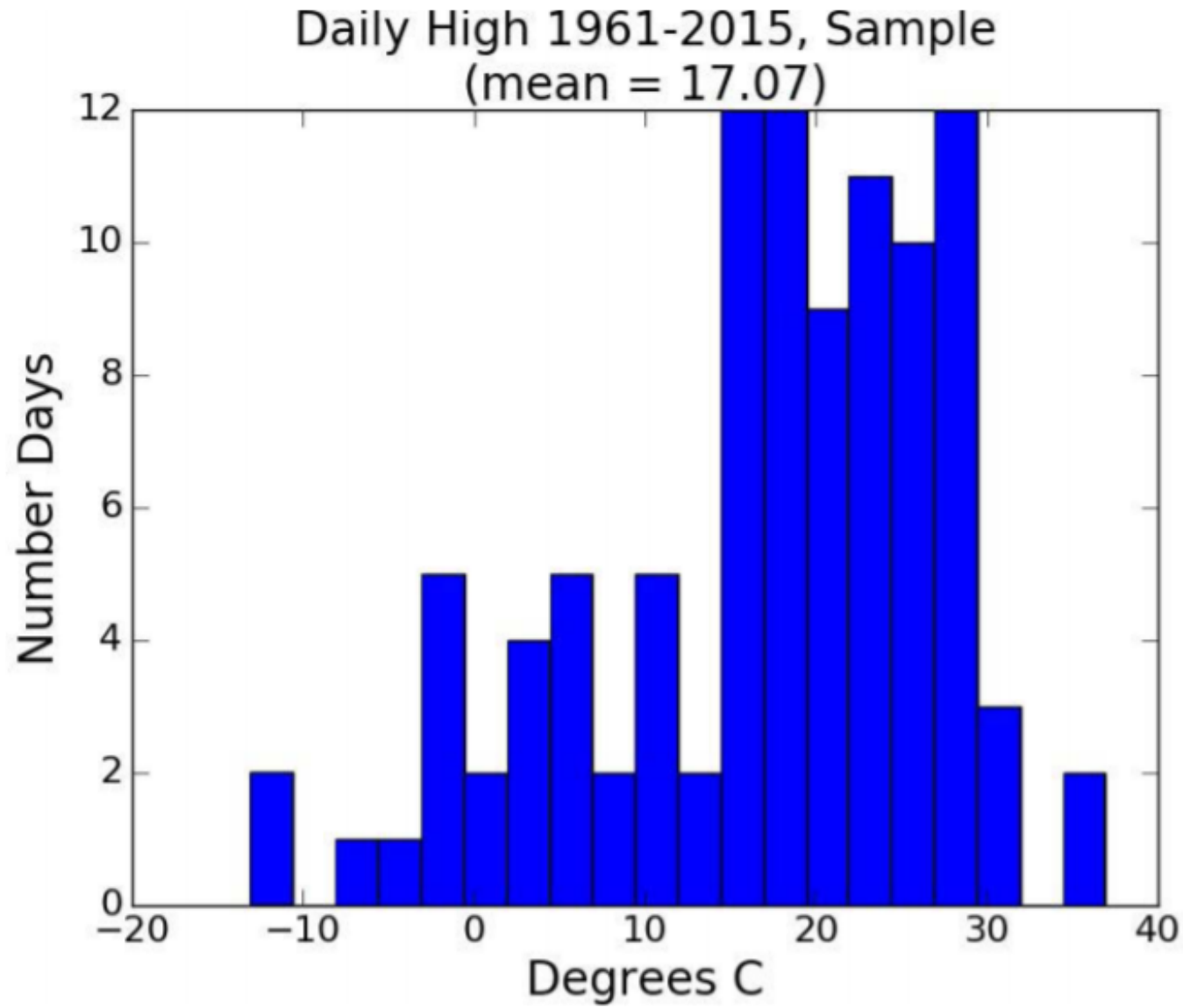
- `numpy.std` is function in the numpy module that returns the standard deviation
- `random.sample(population, sampleSize)` returns a list containing `sampleSize` randomly chosen distinct elements of `population`
  - Sampling without replacement

Histogram of Entire Population



σ = ~9.4

Histogram of Random Sample of Size 100



σ = ~10.4

Means and Standard Deviations

- Population mean = 16.3
- Sample mean = 17.1
- Standard deviation of population = 9.44
- Standard deviation of sample = 10.4
- A happy accident, or something we should expect?
- Let's try it 1000 times and plot the results

To do.

- Empirical Rule: https://colab.research.google.com/drive/1Hj8AV7eBYH6m6Jp3U6zAsaVEP0FBV2V3?usp=sharing
- Central Limit Theorem: https://colab.research.google.com/drive/1lfqdqJ_BnzOvTJ7GZ_eYS8rA_fCnT9zL?usp=sharing
- Assignment 07 is out on IVY.
- Reminder for Take-Home Quiz 03 (due on 11 April 2022)

Recall Central Limit Theorem

- Given a sufficiently large sample:
  - 1) The means of the samples in a set of samples (the sample means) will be approximately normally distributed,
  - 2) This normal distribution will have a mean close to the mean the population, and
  - 3) The variance of the sample means will be close to the variance of the population divided by the sample size.
- Time to use the 3rd feature
- Compute *standard error of the mean* (SEM or SE)