Sampling and Confidence Intervals
Distributions with scipy

Chapter 19

Scipy

- SciPy is a collection of mathematical algorithms and convenience functions built on the Numpy extension of Python.
- With SciPy an interactive Python session becomes a data-processing and system-prototyping environment rivaling systems such as MATLAB, IDL, Octave, R-Lab, and SciLab.
- The additional benefit of basing SciPy on Python is that this also makes a powerful programming language available for use in developing sophisticated programs and specialized applications.

Scipy subpackages

https://docs.scipy.org/doc/scipy/reference/

- SciPy is organized into sub-packages for different scientific computing domains:
  - `cluster` – clustering algorithms
  - `constants` – physical and mathematical constants
  - `io` – input and output
  - `linalg` – linear algebra
  - `spatial` – spatial data structures and algorithms
  - `stats` – statistical distributions and functions
- Each sub-package needs to be imported separately as follows:
  - `from scipy import linalg, stats`

Sampling and Confidence Intervals

Chapter 19

Population

- All possible examples
- Measurable characteristics are **parameters**
- Population mean is $\boldsymbol{\mu}$

Sample

- A subset of the population that is (hopefully) representative of the population
    - Multiple samples can come from a population
    - Sample sizes can vary
- Measurable characteristic are **statistics**
- Sample mean = $\bar{\mathbf{x}}$

Sampling

- **Sampling** is the method by which samples are selected from the population
- **Probability Sampling** each member of the population has some non-zero chance of being selected
  - **Simple random sampling** each member of the population has an equal chance of being selected
  - **Stratified sampling**
    - The population is *stratified* along one or more characteristics.
    - Samples are taken to represent each subgroup.
    - Sample has a better chance to represent the population

Let’s look at our friendly Boston Marathon data

How Big is Big Enough?

- The Law of Large Numbers says as sample size grows the more it should represent the population (e.g. distribution, mean, sd)
- The more **variance** in the population, the larger the sample required

- Compare two normal distributions with different standard deviations

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central Limit Theorem

- Explains why we can use sampling to estimate a population
  - Given a sufficiently large set of samples form a population the means of those samples will approximate a normal distribution
  - The mean of the distribution (mean of means) will be close to the population mean
  - The variance of the sample means will be close to the variance of the population divided by sample size
- Allows us to compute confidence levels and intervals even with the population is *not* normally distributed

Standard Error of the Mean

- **Standard Error of the Mean** (SE or SEM) is the standard deviation of an infinite number of samples of size *n* taken from the same population
  - So it should only take infinity to compute, right?
  - $SE = \sigma_m = \frac{\sigma}{\sqrt{n}}$