Statistics

Definition: Sample

- A *sample* is a subset of randomly selected members from a population.

- A *physical sample* is an actual random selection of members from a population where each member's measurement is denoted by x_i.

- A *random sample* is a theoretical collection of random variables X_i such that the X_i are:

    1. independent
    2. identically distributed according to some pmf $f(x)$

    denoted by:

    $$X_i \overset{\text{iid}}{\sim} f(x)$$

Samples are used when it is not feasible to address an entire population, either because the population is too large or only exists theoretically.

Example

Consider a farm that produces brown eggs. All of the eggs produced by the farm is a population. A carton of a dozen randomly selected eggs is a sample. Since the weight of each egg is iid to a normal distribution:

$$X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$$

Example

Consider tossing a coin with the probability of heads p independently for N times, where X_i denotes the numerical value of the ith toss:

$$X_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$$

Definition: Statistic

A *statistic* is any value that can be calculated from a sample of a population.

Note that since the measurement of each value in a sample has uncertainly and is thus a random variable described by a distribution, a statistic also exhibits uncertainty and is therefore also a random variable with a distribution.

Definition: Mean

The *sample mean* of a sample X of size N, denoted $\bar{X}$, is given by:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Definition: Median

The *sample median* of a sample X of size N ordered by increasing value, denoted $\tilde{X}$, is given by:

$$\tilde{X} = x_{\left(\frac{N+1}{2}\right)}$$

for N odd, and:

$$\tilde{X} = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2}$$

for N even.

Note that both the mean and median are measures of center; however, the mean is more sensitive to outliers. They are equal for symmetric distributions, but the median will move towards the side of the mean with the most values.

Definition: Variance

The *sample variance* of a sample X of size N, denoted s^2, is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Also, the *sample standard deviation* is given by:

$$S = \sqrt{S^2}$$

Definition: Sampling Distribution

The distribution of a statistic random variable is called the *sampling distribution* of the statistic.