1.017/1.010 Class 13
Populations and Samples

Sampling and Repeated Trials:

Statistics relies on sets of observations of random variables, commonly called **samples**.

A sample can be viewed as a collection of outcomes from **repeated trials**. Random trial i yields an outcome (a particular number) corresponding to an underlying random variable $\boldsymbol{x}_i$. The collection of random variables associated with a sample of N observations is $[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_N]$.

Samples are used to estimate distributional properties of the underlying random variables or to make inferences about relationships among different variables.

Samples are sometimes supposed to be drawn from a larger **population** of (which may actually exist or may be hypothesized).

Example:

Define the random variable $\boldsymbol{x}$ to be the height of an MIT student drawn at random from the **population** of all MIT students. Suppose we record the heights of 10 MIT students selected at random. This can be viewed as a composite experiment composed of 10 repeated trials with 10 corresponding random heights $[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{10}]$. The outcomes obtained from this set of trials constitute a **sample**.

The 10 student sample can be used to estimate the unknown mean and variance (or other properties) of the height probability distribution or to make inferences about this distribution (e.g. to compare it to a similar distribution from Stanford).

The following is a sample of 10 student heights (in inches).

[56.25, 59.90, 54.64, 65.73, 56.45, 57.45, 61.26, 57.23, 56.61, 53.67]

What are reasonable estimates of the mean and variance of the student height probability distribution?

Random samples:

$[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_N]$ is a **random sample** if:

1. The $\boldsymbol{x}_i$'s are **independent** $F_{\boldsymbol{x}1,\boldsymbol{x}2,\ldots\boldsymbol{x}N} = F_{\boldsymbol{x}1}(x_1)F_{\boldsymbol{x}2}(x_1)\ldots F_{\boldsymbol{x}N}(x_N)$
2. The $\boldsymbol{x}_i$'s all have the **same distribution** $F_{\boldsymbol{x}i}(x_i) = F_{\boldsymbol{x}}(x)$

Functions of many random variables

A function $\boldsymbol{y} = g(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N)$ of a sample of N random variables is a random variable with CDF $F_{\boldsymbol{y}}(y)$.

If $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N]$ is a random sample $F_{\boldsymbol{y}}(y)$ can be derived from $g(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N)$ and $F_{\boldsymbol{x}}(x)$. In general, this derived distribution problem can be solved with stochastic simulation.

Mean and variance of a linear function (prove)

Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are random variables (not necessarily independent) and a_1, a_2, b are constants (not random):

$$E[a_1\boldsymbol{x}_1 + a_1\boldsymbol{x}_2] = a_1E[\boldsymbol{x}_1] + a_2E[\boldsymbol{x}_2] + b$$

$$Var[a_1\boldsymbol{x}_1 + a_1\boldsymbol{x}_2 + b] = a_1^2Var[\boldsymbol{x}_1] + a_2^2Var[\boldsymbol{x}_2] + 2a_1a_2Cov[\boldsymbol{x}_1, \boldsymbol{x}_2]$$

If $\boldsymbol{x}_1, \boldsymbol{x}_2$ are uncorrelated

$$Var[a_1\boldsymbol{x}_1 + a_1\boldsymbol{x}_2 + b] = a_1^2Var[\boldsymbol{x}_1] + a_2^2Var[\boldsymbol{x}_2]$$

Example:

A reasonable **estimate of the mean** of an unknown probability distribution is the arithmetic average (sample mean) $\boldsymbol{m}_{\boldsymbol{x}}$ of N observations drawn at random from this distribution.

If the observations are outcomes associated with a random sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N$ the sample mean is:

$$\boldsymbol{m}_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

This is a linear function of the random sample. The mean and variance of $\boldsymbol{m}_{\boldsymbol{x}}$ are (prove):

$$E[\boldsymbol{m}_{\boldsymbol{x}}] = E[\boldsymbol{x}] = \overline{\boldsymbol{x}}$$

$$Var[\boldsymbol{m}_{\boldsymbol{x}}] = \frac{1}{N}Var[\boldsymbol{x}]$$

A reasonable **estimate of the standard deviation** of the unknown distribution is the mean sum of squares (sample variance) $\boldsymbol{s_x}^2$ computed from the random sample:

$$\boldsymbol{s}_{\boldsymbol{x}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}]^2$$

This is a linear function of the set of variables $[\boldsymbol{x}_i - \boldsymbol{m}_x]^2$., $i = 1 \ldots N$. The mean and variance of $\boldsymbol{s_x}^2$ are:

$$E[\boldsymbol{s}_{\boldsymbol{x}}^2] = \sigma_x^2$$

$$Var[\boldsymbol{s}_{\boldsymbol{x}}^2] \propto \frac{1}{N}$$