# The IID Assumption

All formal inference procedures in this course assume or require that the data are, in some manner, **IID**, where IID stands for **Independent and Identically Distributed**.

Recall that any random sample - such as a simple random sample of size $n$ - is a collection of random variables which we will denote by $\{X_1, X_2, \ldots, X_n\}$. Also recall that the sample is IID if

**I:** $X_1, X_2, \ldots, X_n$ are **Independent** random variables. Roughly stated this means that the value assumed by any given variable does not depend on nor influence the value assumed by any of the other variables.

**ID:** $X_1, X_2, \ldots, X_n$ are **Identically Distributed**, i.e.,

$$F_{X_1} = F_{X_2} = \cdots = F_{X_n} = F_p$$

where $F_p$ is the distribution of the population or process being sampled.

# Got IID Data? Assessing the IID Assumption

We'll consider two scenarios:

**I.** Simple random sampling of a population

**II.** Sequential Sampling of a process

**I. Simple random sampling.** If the size of the sample is small relative to the size of the population, i.e., $n \leq 5\%$ of the population size, then a simple random sample is approximately an IID sample since

1. $X_1, X_2, \ldots, X_n$ are *approximately* independent since although we are sampling without replacement, the sample is small relative to the population
2. $X_1, X_2, \ldots, X_n$ are approximately*identically distributed*, i.e., have nearly the same distribution because even though we sampling without replacement the population is essentially unchanged by sampling since the sample is small.

Since all the statistical procedures in this course assume IID data of some form, this is a crucial approximation.

**II. Sequential sampling of a process.** By sequential sampling of a process we mean acquiring data **ordered** with respect to time and/or space. The usual engineering application in which this occurs is process control. A typical science scenario is acquiring data using instruments which drift over time. To check that process data are IID we first check that the data are ID (identically distributed) by checking that the distribution of the data is not changing over time or space, i.e., that the data generating process is *stationary*. We do this by plotting the observations in the order collected. Such a graph is called a **time series plot** or **run chart**. You can create a time series plot using the Minitab macro `tsplot`. If the data are ID then the time series plot should consist of points which

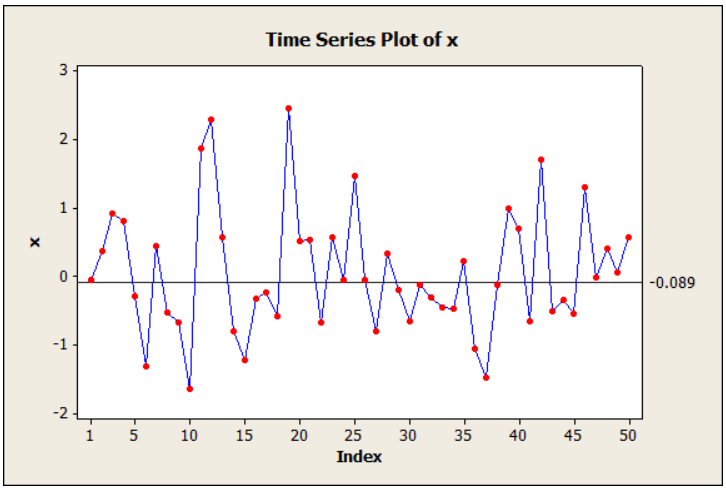
1. appear randomly scattered vertically about the center (sample median) of the data (depicted by a horizontal line), that is, the points do not exhibit an increasing, decreasing, or otherwise systematic trend about the center over time, and
2. appear to have a constant degree of vertical dispersion or scatter, that is, the amount of variability does not exhibit an increasing, decreasing, or otherwise systematic trend over time.

If the data appear appear to be ID, then we check that the data are I (independent) using a **lag plot**. You can create a lag plot using the Minitab macro `lagplot`. Usually lack of independence in process data is due to positive autocorrelation in which adjacent observations are similar in value. We check for autocorrelation by constructing a lag plot of the data. A lag plot is a scatter plot of the points $(x_{i-1}, x_i)$ for $i = 2, 3, \ldots, n$. If these points tend to fall on a line or curve, then adjacent observations tend to obey a relationship and are hence not independent. If the observations are positively autocorrelated, then the points tend to fall on a line with positive slope. If they are negatively autocorrelated, then the points tend to fall on a line with a negative slope. If the points are independent and therefore uncorrelated, the points should be randomly distributed in the x-y plane.

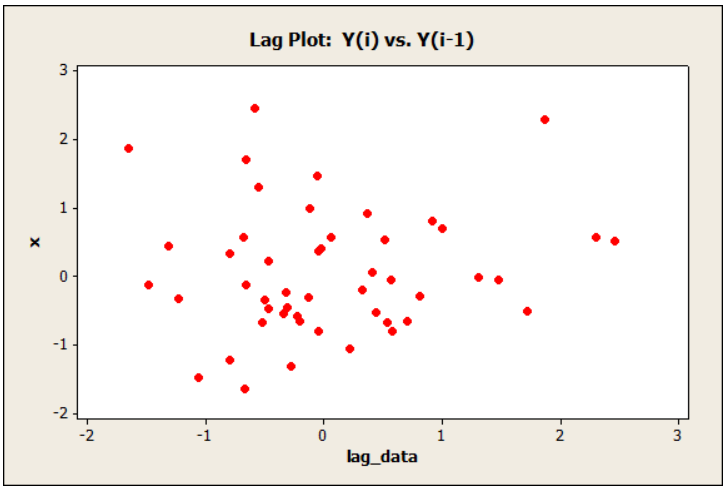
**NOTE: You cannot create time series plot and lag plots of unordered data. In other words, if you do not know or are not informed that the data are in order you cannot create these plots and doing so is incorrect.**

# Examples

The graph below is a time series plot of data from a stationary process. Note that the points satisfy both properties indicating the data are ID (identically distributed).




Since the data appear to be ID, we construct a lag plot to check for lack of independence due to autocorrelation.




Since the points are randomly distributed about the x-y plane we conclude that autocorrelation is not present and the data are independent. Thus we conclude the sample is IID.

**Exercise 1: Time Series Plot/Run Chart Practice.** Open the following data set

`www.rose-hulman.edu/~inlow/examples1.MTW`

then do the following:

**i.** Create a time series plot of the data in column C5 using the macro command `%tsplot c5` and then assess the stationarity of the data by

1. checking if there are systematic trends in location and
2. checking if there are systematic trends in variability.

What do you conclude about the stationarity of the data and why?

**ii.** Create a time series plot of the data in column C7 and assess the stationarity of the data as in part i.

**iii.** Create a time series plot of the data in column C8 and assess the stationarity of the data as in part i.

**Exercise 2: Assessing the IID-ness of Sequential Data.** Determine if the data in column C14 is IID by

1. determining if the data are stationary by creating and interpreting a run chart and
2. determining if the data are independent by creating a lag plot using the macro command `%lagplot c14` and interpreting it.

What do you conclude?

**Exercise 3: Assessing the IID-ness of Process Data.** Yashchin (*Technometrics*, 37, 311-323.) discusses a process for the chemical etching of silicon wafers used in integrated circuits. This process etches the layer of silicon dioxide until the layer of metal beneath is reached. The company monitors the thickness of the silicon dioxide layers because thicker layers require longer etching times. The data - provided in column C16 - consists of the average thickness of 40 consecutive groups of four wafers. Do the following:

1. Assess the stationarity of the data. What do you conclude and why?
2. Assess the independence of the data. What you conclude and why?