Data Acquisition, Sampling, and the IID Assumption

GIGO - garbage in, garbage out - is a core principle in statistics. No amount of sophisticated data analysis can compensate for botched data acquisition. The ability to accurately model a population or process using sample data is critically dependent on how the sample data are acquired.

IID Assumption. ALL analysis procedures in this course require that the data satisfy the *IID ASSUMPTION* in some form. Suppose the sample values are numeric, IQ scores for example. Then the observations comprising a sample of size n - before it is acquired - consist of n random variables: $\{X_1, X_2, \ldots, X_n\}$. The sample is IID if the random variables have the following two properties:

Independent: The random variables $X_1, X_2, \ldots, X_n$ are independent.

Identically Distributed: The random variables $X_1, X_2, \ldots, X_n$ are from the same population and thus have the same distribution or CDF F:

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

Independent Random Variables. Two RV's X and Y are independent if all events involving X and independent of all events involving Y. More precisely, X and Y are independent RV's if, for all numbers a, b, c, and d

$$P(a \le X \le b \cap c \le Y \le d) \quad = \quad P(a \le X \le b)P(c \le Y \le d).$$

This definition generalizes to any number of RV's.

Example 1. Suppose I draw cards from a well-shuffled 52-card deck one by one without replacement. Let $X_1 = 1$ if the first card is black, 0 if it is red. Let X_2, X_3, …, X_{52} be similarly defined. Are X_1,X_2, …, X_{52} independent RV's? Why or why not?

Example 2. How can I modify my procedure in example 1 so that X_1, X_2, …, X_{52} are independent RV's?

Example 3. How can I modify my procedure in example 1 so that X_1, X_2, …, X_{52} are IID RV's?

Why the IID Data Requirement? Why do all our procedures require that the data be IID? Because IID samples have very desirable properties:

1. By consisting of independent observations the sample provides maximum information about the population or process. Although it may not be intuitively obvious, dependent observations contain less information than independent ones.

2. By consisting of independent and identically distributed observations the sample is mathematically tractable. Recall that independent random variables are much easier to analyze than dependent ones. For instance, consider our card examples above. In example

1, what's the probability the first 10 cards are red, i.e., $P(X_1 = 0 \cap X_2 = 0 \cap ... \cap X_10 = 0)$? Now, in example 2 where the $\{X_i\}$ are independent and identically distributed, what's this probability?

3. IID samples have the important property that the larger the sample becomes, the greater the probability the sample will closely resemble the population.

Sampling a Population: Simple Random Sampling

There are two basic sampling scenarios: sampling a population and sampling a process. The usual method for sampling a population is **simple random sampling.**

> A **simple random sample of size n** is any sample acquired in such a way that each subset of size n from the population has the same probability of being the sample.

Note that simple random sampling is *sampling without replacement* and thus the observations comprising the sample are **not** independent. However, if the sample size n is small compared to the population size, then the observations are approximately independent and so a simple random sample is approximately IID.

Example: It is highly likely that you have done simple random sampling at some point in your life. Can you think of a situation in which you have randomly sampled a collection of objects?

Note: It is important to be able to distinguish simple random sampling from other sampling procedures. A common mistake is to confuse simple random sampling with procedures in which *each individual has an equal chance of being in the sample.* While it is true that if you use simple random sampling then each individual will have an equal chance of being in the sample, the converse is not true. In other words, the fact that a sampling procedure ensures that each individual has an equal chance of being in the sample does NOT imply that the procedure is a simple random sampling procedure. For example, suppose we select a five card hand from a deck of 52 cards as follows. We split the deck into half with respect to color: the 26 black cards go into one stack, the 26 red into a second. We toss a nickel. If its heads, we select the top five cards from the black stack; if it's tails we select the top five cards from the red stack. Assuming both stacks are thoroughly shuffled and the nickel toss is fair, each card in the deck has an equal chance of being in the resulting hand. However, this procedure is not simple random sampling.

Exercise 1: Why is the above procedure not simple random sampling?

Summary: The simplest way to get an approximately IID sample from a population is to

1. use simple random sampling, and
2. use a sample size n which is small compared to the population size, i.e., $n < 5\%$ of the population size.

Exercise 2: IID Sampling Simulation. In this exercise you will use Minitab to simulate getting IID samples from a standard normal population. By increasing the sample size and inspecting the resulting normality plot and histogram, you will verify that the larger the IID sample, the more similar the sample distribution resembles the population distribution, i.e., the normal distribution. Do the following:

i. Open Minitab by clicking on the Minitab icon or else executing Minitab from the *start* -> *all programs* menu.

ii. Generate an IID sample from the standard normal population as follows:

1. Select the `Calc` menu off the main menu bar at the top of the Minitab window.
2. Select `Random Data -> Normal....`
3. In the `Normal Distribution` window, enter 10 for `Number of data rows to generate`.
4. Enter "c1" in the `Store in column(s)` text box then click OK. You should see 10 numbers appear in column c1 in the worksheet.

iii. Create a normality plot of the IID sample in column c1 using Minitab's normality test routine, `Stat -> Basic Statistics -> Normality Test`.

iv. Create a histogram of the IID sample in column c1 as follows:

1. Select the `Graph` menu of the main menu bar then select `Histogram`.
2. In the `Histograms` window, select the `With Fit` window then click OK.
3. In the `Histogram - With Fit` window, select column c1 by double-clicking on c1 in the left textbox then click OK. You should now see a histogram of the data with a normal density curve superimposed on it. Do they look similar?
4. If you want delete the histogram by clicking on the "x" in the red box in the upper right corner and then clicking `no`.

v. Repeat steps ii and iii several times with a sample of size 10.

vi. Repeat steps ii and iii several times with a sample of size 100.

vii. Repeat steps ii and iii several times with a sample of size 1000.

viii. Repeat steps ii and iii several times with a sample of size 10000.

viii. As the sample size is increased does it become more clear that the sample is from a normally distributed population?

Sampling a Process: Systematic Random Sampling

When sampling a process, it is typically neither desirable nor necessary to acquire data using simple random sampling. Instead **systematic random sampling** is often used. Systematic random sampling is a sampling procedure which

1. selects every mth item, and
2. selects the first item randomly from among the first m items.

Example 4. Why is systematic random sampling preferable to simple random sampling when sampling a process? Consider getting a simple random sample of size 100 from the output of a production line for a given day in order to estimate the percentage of defectives. Suppose the line produces 10,000 units/day. In order to get a simple random sample of size 100 we use a random number generator to select which 100 of the 10,000 will be selected. What can go wrong with this procedure? (Hint: recall that under simple random sampling any size 100 subset of the 10,000 units can be the sample.)

Population or Process?

One final word about IID sampling. Whenever you acquire data you should always record the order in which you acquire the measurements. For example, suppose you get a simple random sample of water samples from a lake. If you analyze the samples sequentially using an electronic instrument, you now have process data since the instrument might drift over time.