BTS | Barcelona Technology School

Alex Kumenius – Business Intelligence Data Scientist
Date : October 2019

RANDOM VARIABLES

We call a variable or process with a **numerical outcome** a **Random Variable**, and we represent this **Random Variable** with a capital letter such as X, Y, or Z. The possible **outcomes** of X are labeled with a corresponding lower case letter x and subscripts.

The amount of money a single student will spend on her statistics books is a random variable, and we represent it by X.

When we computed the average outcome of X, we call this average the **expected value of** X, denoted by $E(X)$. The *expected value of a random variable* is computed by **adding** each outcome **weighted** by its probability.

Expected Value of a Discrete Random Variable If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$E(X) = x_1 * P(X = x_1) + \ldots + x_k * P(X = x_k)$$

The $Greek$ letter μ may be used in place of the notation $E(X)$.

SAMPLING WITHOUT REPLACEMENT

When a **sampling unit** is drawn from a **finite population** and is **NOT returned to that population**, after its characteristic(s) have been recorded, before the next unit is drawn, the **sampling** is said to be **"without replacement"**.

SAMPLING WITH REPLACEMENT

When a **sampling unit** is drawn from a **finite population** and is **returned to that population**, after its characteristic(s) have been recorded, before the next unit is drawn, the **sampling** is said to be **"with replacement"**.

SAMPLING FROM A SMALL POPULATION

If we sample from a small population **WITHOUT replacement**, we no longer have independence between our **observations**.

- the probability of **not being picked** for a *second draw* is condition on the event that it was **not picked** for the *first draw*.

However, if we sample from a small population **WITH replacement**: We repeatedly sample the entire **population** without regard to which **observation** we already selected.

> When the sample size is only a **small fraction** of the **population (under 10%)**, observations are **nearly independent** even when **sampling without replacement**

.

RANDOM SAMPLING METHODS

Statistical methods are based on the notion of implied **randomness**. If **observational data** are not collected in a random framework from a **population**, these statistical methods – the **estimates and errors** associated with the estimates – **are not reliable**.

Sampling takes on two forms in statistics:

- **Probability sampling** uses random sampling techniques to create a sample.
- **Non-probability sampling** techniques use non-random processes like researcher judgment or convenience sampling.

PROBABILITY SAMPLING

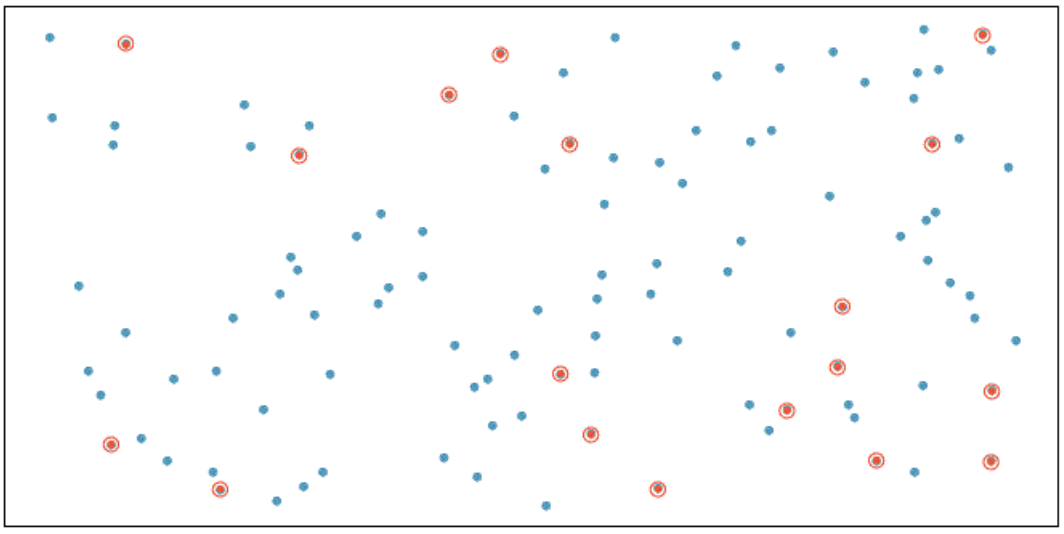
Probability sampling is based on the fact that every member of a population has a known and equal chance of being selected. **Probability sampling** gives us the best chance to create a sample that is truly representative of the population.

we consider four ***random sampling techniques*** :

1. Simple,
2. Stratified,
3. Cluster,
4. Multistage sampling.

Simple Random Sampling

> In general, a sample is referred to a Simple Random Sampling if each **case** in the population has an **equal chance** of being **included in the final sample** and knowing that a **case** is included in a sample **does not provide useful information** about which other **cases** are included.



Simple Random Sampling is probably the most intuitive form of random sampling.

In [ ]:
```
import random
```

In [ ]:
```
population=[2,5,7,8,9]
k=3
random.sample(population, k)
```

- Members of the list need not be hashable or unique, and it allows duplicate entries.
- **k** must be less than the size of the list.
- The population can be any sequence or set from which you want to select a **k** length numbers.
  The sequence can be string, List.

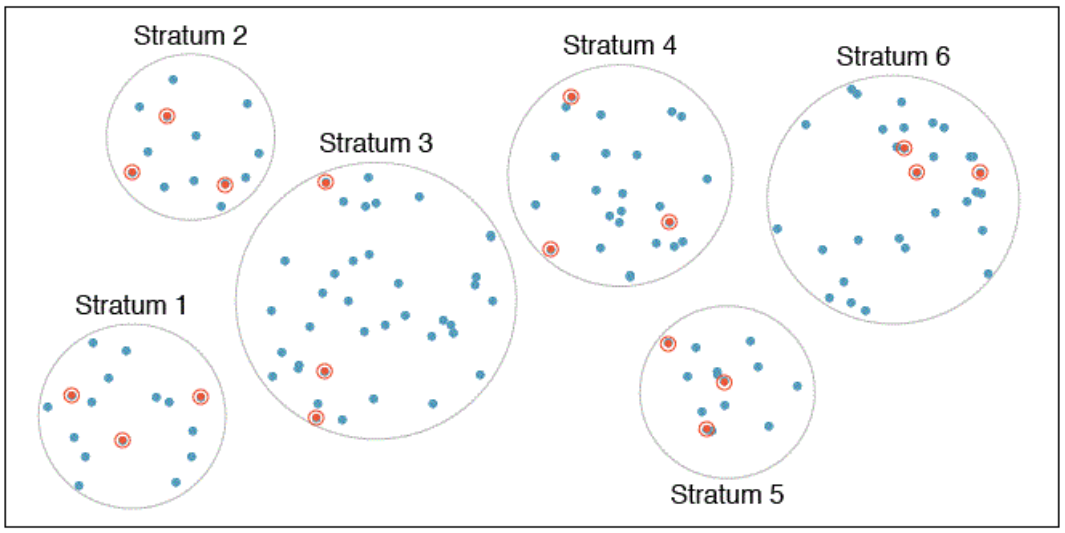
If the list contains repeats, then each occurrence is a possible selection in the sample.

In [ ]:
```
population=[2,5,2,8,2]
k=3
random.sample(population, k)
```

In [ ]:
```
population=[2,5,2,8,2]
k=3
np.random.choice(population, k, replace=True)
```

Stratified Random Sampling

Stratified Random Sampling is a divide-and-conquer sampling strategy. The population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each **stratum**.



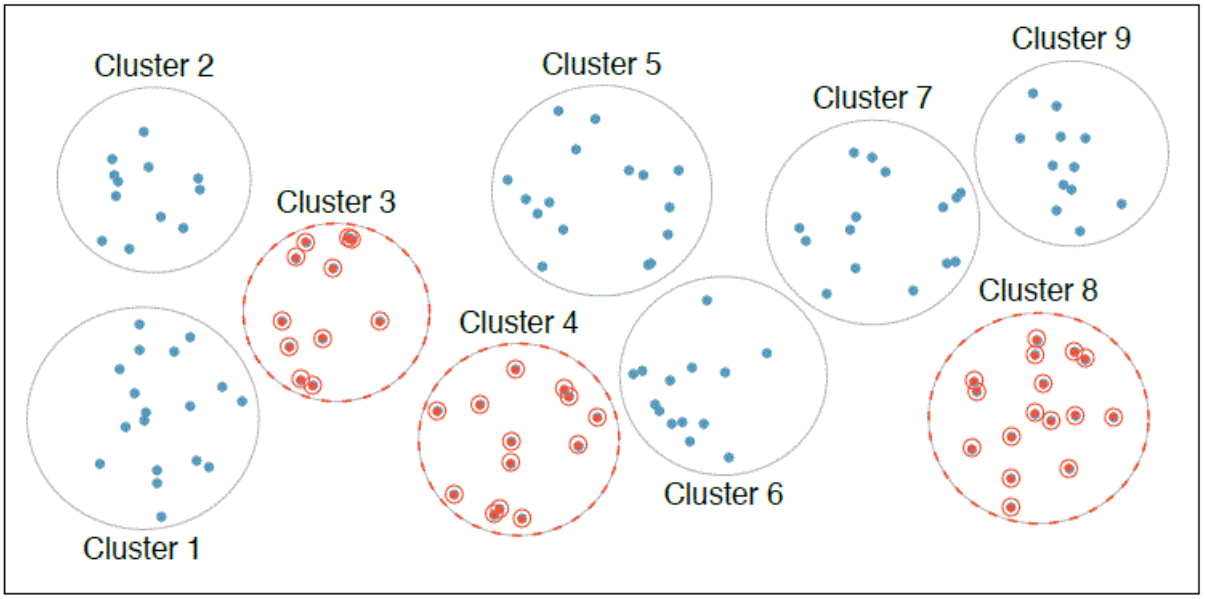
Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. Analyzing data from a Stratified Sample is a more complex task than analyzing data from a Simple Random Sample.

Why would it be good for cases within each stratum to be very similar?

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population.

Cluster Random Sampling

In a **Cluster Random Sample**, we break up the population into many groups, called clusters. Then we **sample** a fixed number of clusters and include all **observations** from each of those clusters in the **sample**.



cluster or multistage sampling can be more economical than the alternative sampling techniques. Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another.

For example, if neighborhoods represented clusters, then cluster or multistage sampling work best when the neighborhoods are very diverse.

A downside of these methods is that more advanced analysis techniques are typically required.

EXAMPLE

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria.

What sampling method should be employed?

SOLUTION

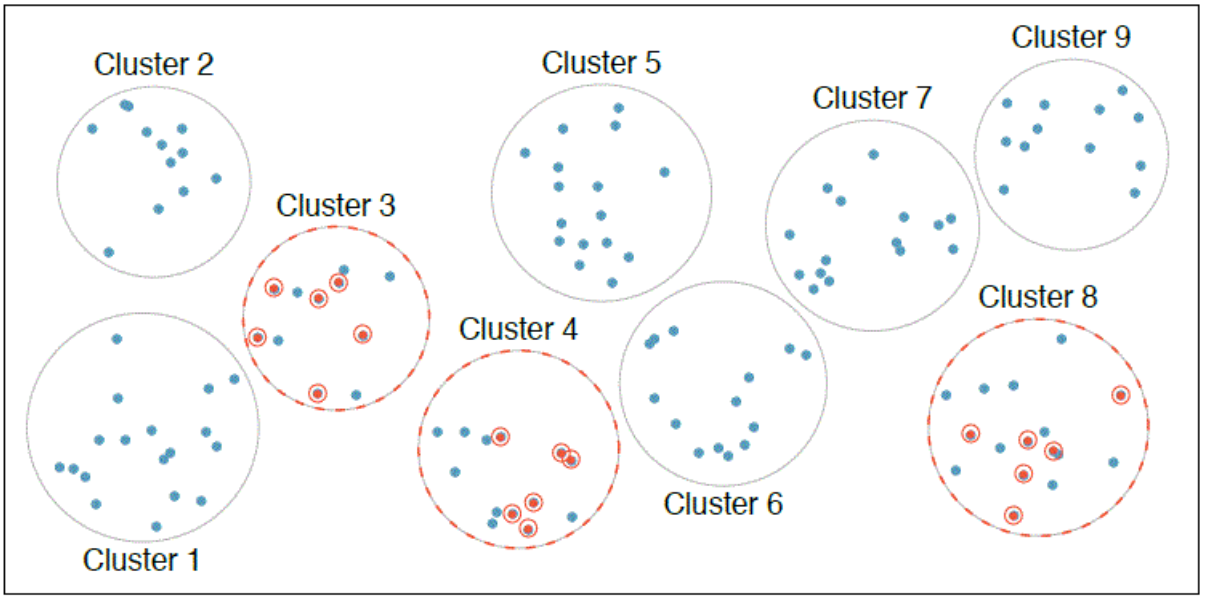
1. simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive.

2. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals.

3. However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and this approach would still give us reliable information.

Multistage Random Sampling

A **Multistage Random Sample** is like a cluster sample, but rather than keeping all **observations** in each cluster, we collect a random sample **within each** selected cluster.</DIV>



Transforming Data to Log

When data are very strongly skewed, we sometimes **transform** them so they are easier to model.

Most of the data are collected into **one bin** in the histogram and the data are so strongly skewed that many details in the data are **obscured**.

There are some **standard transformations** that are often applied when much of the data cluster near zero (*relative to the larger values in the dataset*) and all observations are positive. A **transformation** is a **rescaling** of the data using a function.

Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

In [ ]:

In [ ]:

Transformations can also be applied to one or both variables in a scatterplot.

We can see a positive association between the variables and that many observations are clustered near zero. However, we'll find that the data in their current state cannot be modeled very well.

Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

In [ ]:

The scatterplot where both the **fed_spending** and **poverty** variables have been transformed using a log (base e) transformation.

While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.