# *DS100: Probability Samples*

*Prof. Deborah Nolan*

Scribe: Simon Mo

## *Probability Samples*

Suppose we have a population of 6 individuals. They are uniquely identified by letters $A - F$.

## *Simple Random Sample (SRS) sample*

To take a SRS, we can think of filling an urn with 6 marbles, each identical in size and shape and marked with the unique identifier. We mix the marbles in the urn well. We draw (without looking) 2 marbles from the urn. [1]

What are the possible samples?

AB
AC
AD
AE
AF

BC
BD
BE
BF

CD
CE
CF

DE
DF

DF

There are 15 possible samples of size 2 from our population of 6. Another way to count the number of samples is:

$$\binom{6}{2} = \frac{6!}{2!4!} = 15$$

Here notation $\binom{6}{2}$ means 6 choose 2

Each of these 15 samples are *equally likely* to be chosen. So

$$\mathbb{P}(AB) = \mathbb{P}(CD) = \ldots = \mathbb{P}(DF) = \frac{1}{15}$$

Notation $\mathbb{P}$ means chance.

This notion of each possible sample being equally likely defines the SRS. We can also use this chance mechanism to answer other questions about the composition of the sample. For example:

$$\mathbb{P}(A \text{ in sample}) = \frac{5}{15} = \frac{1}{3}$$

$\frac{5}{15}$ comes from that there are 5 of the 15 possible samples include $A$.

By symmetry, we can say:

$$\mathbb{P}(A \text{ in sample}) = \mathbb{P}(F \text{ in sample}) = \frac{1}{3}$$

Another approach is:

$$\mathbb{P}(A \text{ in sample}) = \frac{2}{6} = \frac{1}{3}$$

$\frac{2}{6}$ comes from: 2 marbles chosen out of 6.

There are many other probability samples used for selecting subjects from a population. We briefly describe the cluster sample and the stratified sample.

[1] We can either draw both marble at once, or draw 1, set it aside, draw the next.

## *Cluster Sample*

We have the same 6 individuals, but they are organized us buddies as follows [2]:

$$(A,B) \quad (C,D) \quad (E,F)$$

Suppose we chose 1 cluster at random and observe all units in the cluster. Now what is the change $A$ is in the sample?

$$\mathbb{P}(A \text{ in sample}) = \mathbb{P}(\text{cluster } (A,B) \text{ chosen}) = \frac{1}{3}$$
$$\mathbb{P}(B \text{ in sample}) = \frac{1}{3}$$
$$\mathbb{P}(C \text{ in sample}) = \mathbb{P}(D \text{ in sample}) = \frac{1}{3}$$

Is this a SRS? Are all subsets of 2 individuals equally likely to be chosen? NO! Because:

$$\mathbb{P}(A,C \text{ in sample}) = 0 \qquad \text{because they are in different cluster}$$
$$\mathbb{P}(A,B \text{ in sample}) = 1$$

Is it a probability sample? YES! We can still compute the chance of various units being chosen. [3]

## *Stratified Sample*

We have the same 6 individuals, but they are divided into 2 strata as follows:

$$\begin{array}{ll} \text{Strata 1} & \{A,B,C,D\} \\ \text{Strata 2} & \{E,F\} \end{array}$$

We take a SRS of 1 unit from each strada. Again, what is the chance $A$ is in our sample?

Now our possible samples are:

$$(A,E) \quad (A,F) \quad (B,E) \quad (B,F) \quad (C,E) \quad (C,F) \quad (D,E) \quad (D,F)$$

Each is equally likely, so $\mathbb{P}(A \text{ in sample}) = \frac{2}{8} = \frac{1}{4}$.

Since we know $A$ is in strata # 1 and 1 unit is taken from this stratum. We have $\frac{1}{4}$

But,

$$\mathbb{P}(A,B \text{ in sample}) = 0 \qquad \text{Why?}$$
$$\mathbb{P}(A,E \text{ in sample}) = \frac{1}{8} \qquad \text{Why?}$$

Strata do no need to be the same size. We can take a different number of units from each stratum. [4] [5]

[2] That is, we have 3 clusters of 2.

[3] Clusters do not need to be the same size in practice. We can choose more than one cluster for our sample. Cluster sampling is a SRS of clusters, rather than a SRS of individual units.

[4] Strata: plural;
Stratum: singular.

[5] Stratified sampling takes a SRS from each stratum.

## *What's so good about simple random sampling?*

1. Sample is often representative of population.
2. We can compute chance of units being in sample.
3. We can measure the variability/accuracy in the sample.

## *Why would we do any other kind of sampling?*

*Convenience* cluster sample tend to be more convient but units in a cluster tend to be similar so accuracy suffers.

*Increased Accuracy* if stratification is done well.