Sampling and Sampling distributions

Sampling and Sampling distributions

- Population and sample
- Random and simple random samples
- Other types of samplings
- Sampling from a given distribution
- Sample statistic and distribution
- Techniques for sample distributions

Population and sample

- **Population** is a set of all elements of interest for a particular study
- **Sample** is a subset of the population selected to represent the whole population
- Collecting all information from every possible item is called **census.** It is often, when number of elements in the population is very large, time consuming and cost inefficient. Sampling is an alternative of census that provides time and cost saving.
- For example: if a Television company wants to know popularity of some of the programs it would be expensive to ask everybody's opinion. Instead subset of viewers are interviewed.
- Sometimes trying out all items in the population would be impossible due to some reasons (e.g. trying out an item damages it). Census of continuous population is physically impossible.
- Surprisingly, sometimes studying carefully chosen random sample can give more accurate understanding of population than census. For census trained and experienced interviewers are needed. They may not be available. Then accurate information collected by well trained interviewers can give better result than inaccurate information about population as whole collected by not very well trained interviewers.

Example of population and sample

We want to measure average height, weight and correlation between them for men. We chose 15 individuals randomly and measure their parameters. Population is “all men”, sample is “15 individuals”. Their parameters are heights and weights.

	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164

Random and simple random samples

- Purpose of sampling: Given a sample from population, to infer from it some or all properties of population. To apply probability to this problem sampling should be random.
- **Random sample** is a sample that has been selected so that every possible sample has a calculable chance of selection. It means that any subset of population has some, not necessarily equal chance to be selected.
- **Simple random sample** of size n from population of size N is a sample selected so that each possible sample of the same size has the probability being selected and successive drawings are independent.
- For example if we want to infer about the next election in UK then we can choose subset of population of UK and ask their opinion. Care should be taken so that sample is representative. Sampling from conservative supporters would give completely wrong impression.
- Exercise: If the population size is N what is the number of possible samples of size n? What is the probability of each sample?
- **Simple random sample from infinite population** should have two properties: 1) each item chosen comes from the same population; 2) Each item is selected independently.

Other types of sampling

Stratified sampling

- Population is first divided into groups (strata). Each item of population belongs only to one strata. Strata is formed so that elements within each stratum is as much as possible alike. For example population of earth could be divided into ethnic groups. After strata are formed then simple random sampling from each stratum is carried out. If items within strata are alike then variance of estimation will be smaller and characteristics of estimation will be similar to that of simple random sampling. Stratified sampling is used to reduce sample size.

Cluster sampling

- Again population is divided into groups called clusters. Then clusters are taken in random and all elements of cluster are taking as sample. Best results can be achieved if elements of clusters are as little alike as possible. Clusters represent small scale version of entire population. For example clusters can be city blocks. Using cluster sampling can reduce the cost of sampling.

Other types of sampling

Systematic sampling

- If simple random sampling from population is very expensive then systematic sampling can be used. First population is enumerated from 1 onwards. If sample size of n from population size of N is required, every k=N/n-th item is selected. First a random number between 1 and k is selected and it is taken as the 1^{st} element. Then every k-th element is taken. It can be expected that systematic sampling will have similar properties to simple random sampling if enumeration of population was done randomly

Convenience sampling

- All above sampling are probability sampling. Convenience sampling is not probability sampling. For example for medical studies voluntaries are used. It can not be guarantied that voluntaries will represent whole population. This type of sampling is easy to perform but results should be interpreted carefully

Judgment sampling

- In this case persons most knowledgeable in the field select items for study. It is easy to perform but results again should be interpreted and used carefully.

Sampling from a given distribution

- Random numbers play an important role in random sampling. With advent of computers this task has become relatively easy. The simplest form of generating random numbers is using following equation

$$x_{i+1}=(a\, x_{i+1}+b)\ (mod(m))\ +\ 1$$

- This procedure will generate pseudo uniform random numbers between 1 and *m*. To generate pseudo random numbers in the interval (0,1] one can choose m large enough and then dived x_{i+1} by m at every stage. There are many various techniques for pseudo random number generation. For example see Knuth, The art of Programming, Volume 1.
- Often one needs to generate random number from a given distribution. Let us assume we want to generate random numbers with given probability distribution *F(x)*. Then we first generate uniform random numbers (say *u*) in the interval (0,1]. Then use following relation:

  $F(x)=u$ if distribution corresponds continuous random variable

  $F(x_{j-1}) < u \leq F(x_j)$ if distribution corresponds to a discrete random variable

Sampling from a given distribution

- To do that it is necessary to have inversion of $F(x)$ and solve the equation $x=F^{-1}(u)$.
- For example if probability distribution is exponential i.e. $F(x) = 1-e(-x)$ then random number from this distribution will be $x = -log(1-u)$.
- One simple way of generating random numbers from the distribution of $F(x)$ is to tabulate $F(x)$ to create look-up table and then use it to generate random numbers from this distribution.
- Generating random numbers from the given distribution plays important role in computer simulation, integration in multidimensional space, stochastic optimisation techniques such as Monte-Carlo, genetic, evolution algorithms.

Sample statistic

- In random sampling every item is considered as a random variable therefore they have distribution.
- If each draw is independent from others then sample distribution will be product of the distributions of the individual items.
- Let us assume that we have sample of size n and each of them come from the population with density of distribution (continuous case) f(x) then probability distribution of sample $(x_1,x_2,,,x_n)$ is

$$f(x_1,x_2,,,x_n) = \Pi f(x_i)$$

- Often we are not interested on individual outcomes but some general properties. Usually general properties are calculated as a function of the elements of the sample – $t(x_1,x_2,,,,x_n)$. Function of the sample elements is called a statistic.
- Since every element of sample is random variable any function of them is also random variable. Therefore sample statistic is random variable and it has a probability distribution.
- To analyse variance, reliability of the sample statistics we need their distribution. If we could find this distribution all probabilistic properties can be derived.

Sample statistic

Examples of sample statistics:

Sample mean. Assume that we have sample of size n.

Sample mean is defined as: $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Example of sample and sample statistic

Population of size N=6:

Unit	1	2	3	4	5	6
X	1	1	2	2	2	3

Probability of elements

x	1	2	3
p(x)	1/3	1/2	1/6

Probability distribution and sample
Statistics of sample size 2

(x_1,x_2)	p(x1,x2)	m	S^2
(1,1)			
(2,2)			
(3,3)			
(1,2)			
(1,3)			
(2,3)			
(2,1)			
(3,1)			
(3,2)			

Example of sample and sample statistic

Population of size N=6:

Unit	1	2	3	4	5	6
X	1	1	2	2	2	3

Probability of elements

x	1	2	3
p(x)	1/3	1/2	1/6

Probability distribution and sample Statistics of sample size 2

(x_1,x_2)	p(x1,x2)	m	S^2
(1,1)	1/9		
(2,2)	1/4		
(3,3)	1/36		
(1,2)	1/6		
(1,3)	1/18		
(2,3)	1/12		
(2,1)	1/6		
(3,1)	1/18		
(3,2)	1/12		

Example of sample and sample statistic

Population of size N=6:

Unit	1	2	3	4	5	6
X	1	1	2	2	2	3

Probability of elements

x	1	2	3
p(x)	1/3	1/2	1/6

Probability distribution and sample Statistics of sample size 2

(x_1,x_2)	p(x1,x2)	m	S^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,3)	1/12	2.5	0.5
(2,1)	1/6	1.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Example of sample and sample statistic: Distribution

Distribution of sample statistics:

Derive sample distribution by counting and summing all possible values and using their probabilities.

(x_1,x_2)	p(x1,x2)	m	S^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,3)	1/12	2.5	0.5
(2,1)	1/6	1.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Mean:

m	p(m)
1	
1.5	
2	
2.5	
3	

Variance:

S^2	$p(S^2)$
0	
0.5	
2	

Example of sample and sample statistic: Distribution

Distribution of sample statistics:

Derive sample distribution by counting and summing all possible values and using their probabilities.

(x_1,x_2)	p(x1,x2)	m	S^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,3)	1/12	2.5	0.5
(2,1)	1/6	1.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Mean:

m	p(m)
1	1/9
1.5	1/3
2	13/36
2.5	1/6
3	1/36

Variance:

S^2	$p(S^2)$
0	14/36
0.5	1/2
2	1/9

In reality it is not easy when number of possible sample of size n is very large.

Exercise: Calculate expected value for mean and variance.

Techniques to derive sample distributions

- If the distribution of random variables is known then it may allow to derive exact sampling distribution. Usual techniques used for that are:
- Characteristic function. It is known that characteristic function and distribution define each other. The characteristic function of sample statistics can be derived and then its inversion will give the probability distribution.
- Example. Let us assume that sample is from population with standardized normal distribution. Then each element of the sample has normal distribution with 0 mean and unit variance.
- Let us consider mean value of the sample. If random variables are independent then characteristic function for the sum of the random variables is a product of the characteristic functions for individual random variables

Techniques to derive sample distributions

Recall that:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And characteristic function for normal variable:

$$c(t) = e(-t^2)$$

Then characteristic function for the sum of n random variable will have:

$$c_n(t) = (c(t))^n = e(-nt^2)$$

And it is easy to see that characteristic function for mean is exactly same as characteristic function for normal variable with variance=1/n and 0 mean=0. It can be extended for normal distribution with a given variance and mean

Techniques to derive sample distributions: Contd.

- Another widely used technique is change of variables.
- Let us assume we want to derive distribution of sample statistic $t(x_1,x_2,,,x_n)$. Then we define new variables so that one of them is t. Then we find joint distribution for new variables and integrate all variables but t.
- Example: Assume that sample of size is from normal distribution with 0 mean and unit variance. We want to find distribution of sample variance:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- Joint distribution of n independent random variables is product of the distributions of the individual components. I.e.

$$f(x_1,,,x_n)=\frac{1}{(2\pi)^{n/2}}e(-1/2(\sum_{i=1}^{n}x_i^2))$$

Techniques to derive sample distributions: Contd.

- We want first the distribution of $z=\Sigma x_i^2$. Now if we use polar transformations:

$$x_1 = z^{1/2} \cos\theta_1 \ldots \cos\theta_{n-1}$$

$$x_j = z^{1/2} \cos\theta_1 \ldots \cos\theta_{n-j} \sin\theta_{n-j+1}$$

$$\ldots\ldots\ldots$$

$$x_n = z^{1/2} \sin\theta_1$$

- Jacobian of this transformation is

$$1/2 z^{\frac{n}{2}-1} \cos^{n-2}\theta_1 \ldots \cos\theta_{n-2}$$

- To derive distribution for z we want integrate all θs out. If we multiply distribution $f(x_1,,,x_n)$ with the Jacobian integrate θs out we will get χ^2-distribution with n degrees of freedom. Having derived distribution for z we can derive distribution for z/n also.

Techniques to derive sample distributions: Cont.

One popular technique is to use a limiting distribution to approximate the sampling distribution of interest. Usually two type of limiting concepts are used: 1) Convergence in probability 2) convergence in distribution:

Defintion 1: A sequence of random variables X_m converges in probability to a randam variable c if for every positive number ε and δ, there exists positive integer $m_0 = m_0(\varepsilon,\delta)$ so that

$$P(|X_m - c| > \varepsilon) < \delta, \; m \geq m_0$$

Example: Consider sequence X_m with probabilities $P(x_m=0) = 1 - 1/m$, $P(x_m=m) = 1/m$. It converges in probability to c=0.

Techniques to derive sample distributions: Cont.

Defintion 2. A sequence of random variables X_m converges in distribution to a random variable X if for every ε there exists an integer $m_0=m_0(\varepsilon)$ such that every point where F(x) continuous

$$|F_m(x)-F(x)|< \varepsilon, m>m_0$$

Where $F_m(x)$ is distribution of X_m and $F(x)$ is distribution of X.

Example: Sequence of random variables $P(x_m=1) = ½+1/m$ and $P(x_m=2) = ½-1/m$. It converges to random variable with probability distribution $P(x_m=1) = P(x_m=2) =1/2$.

These definitions are used often to derive limiting probability distributions for sample statistics

Example of limiting distribution: Poisson as limiting distribution of binomial

- For any distribution with finite mean (μ) and finite variance (σ^2) sample mean distribution converges to normal distribution

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \xrightarrow{D} N(0,1)$$

- It is true regardless of the parent distributions. Where D means convergence in distribution. For some distributions convergence is faster than for other distributions. If distribution is extremely skewed then number of sample size should be very large in order normal approximation to be accurate.

Example of limiting distribution: Poisson as limiting distribution of binomial

- If we are making draw from “success” (with probability p) and “failure” scheme and we are interested in fraction of successes then for large number of trials its distribution will converge to normal distribution – *N(p, p (1-p)/n)*. It is good approximation when *np>5* and *n(1- p)>5*.
- I.e. neither success nor failure have very small probability. When probability of “success” or that of “failure” becomes very small then Poisson distribution becomes better approximation:

$$\binom{n}{j} p^j (1-p)^{n-j} \xrightarrow{D} \frac{(np)^n}{j!} e(-np)$$

- That is why it is called probability of rare events. That is reason why when probability of success (or failure) is very small for moderate size of samples Poisson rather than normal approximation is used.

References

Berthold, M. and Hand, D.J. (eds) Intelligent Data Analysis, 1998, Springer

Anderson, D.R., Sweeney, D.J. and Williams T.A. Introduction to statistics; Concepts and Applications, 1991, West Publishing Company

Box, G.E.p., Hunter, W.G. and Hunter, J.S. Statistics for Experimenters, 1978, John Wiley & Sons

Stuart, A. and Ord, JK (1994) Kendall's Advanced theory of statistics: Volume 1. Distribution theory