Math2 - Random samples, sampling distributions of estimators, Methods of Moments and

UNIT-2

Random samples, sampling distributions of estimators, Methods of Moments and Maximum Likelihood.

In statistics, a **simple random sample** is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals.[1] This process and technique is known as **simple random sampling**, and should not be confused with systematic random sampling. A simple random sample is an unbiased surveying technique.

Simple random sampling is a basic type of sampling, since it can be a component of other more complex sampling methods. The principle of simple random sampling is that every object has the same probability of being chosen. For example, suppose N college students want to get a ticket for a basketball game, but there are only X < N tickets for them, so they decide to have a fair way to see who gets to go. Then, everybody is given a number in the range from 0 to N-1, and random numbers are generated, either electronically or from a table of random numbers. Numbers outside the range from 0 to N-1 are ignored, as are any numbers previously selected. The first X numbers would identify the lucky ticket winners.

In small populations and often in large ones, such sampling is typically done "**without replacement**", i.e., one deliberately avoids choosing any member of the population more than once. Although simple random sampling can be conducted with replacement instead, this is less common and would normally be described more fully as simple random sampling **with replacement**. Sampling done without replacement is no longer independent, but still satisfies exchangeability, hence many results still hold. Further, for a small sample from a large population, sampling without replacement is approximately the same as sampling with replacement, since the odds of choosing the same individual twice is low.

An unbiased random selection of individuals is important so that if a large number of samples were drawn, the average sample would accurately represent the population. However, this does not guarantee that a particular sample is a perfect representation of the population. Simple random sampling merely allows one to draw externally valid conclusions about the entire population based on the sample.

Conceptually, simple random sampling is the simplest of the probability sampling techniques. It requires a complete sampling frame, which may not be available or feasible to construct for large populations. Even if a complete frame is available, more efficient approaches may be possible if other useful information is available about the units in the population.

Advantages are that it is free of classification error, and it requires minimum advance knowledge of the population other than the frame. Its simplicity also makes it relatively easy to interpret data collected in this manner. For these reasons, simple random sampling best suits situations where not much information is available about the population and data collection can be efficiently conducted on randomly distributed items, or where the cost of sampling is small enough to make efficiency less important than simplicity. If these conditions do not hold, stratified sampling or cluster sampling may be a better choice.

Algorithms[edit]

Several efficient algorithms for simple random sampling have been developed.[2][3] A naive algorithm is the draw-by-draw algorithm where at each step we remove the item at that step from the set with equal probability and put the item in the sample. We continue until we have sample of desired size k. The drawback of this method is that it requires random access in the set.

The selection-rejection algorithm developed by Fan et al. in 1962[4] requires single pass over data; however, it is a sequential algorithm and requires knowledge of total count of items n, which is not available in streaming scenarios.

A very simple random sort algorithm was proved by Sunter in 1977[5] which simply assigns a random number drawn from uniform distribution (0, 1) as key to each item, sorts all items using the key and selects the smallest k items.

J. Vitter in 1985[6] proposed reservoir sampling algorithm which is often widely used. This algorithm does not require advance knowledge of n and uses constant space.

Random sampling can also be accelerated by sampling from the distribution of gaps between samples,[7] and skipping over the gaps.

Random sampling is a technique used in selecting people or or items for research. There are many techniques that can be used; but, each technique makes sure that each person or item considered for the research has an equal opportunity to be chosen as part of the group.

Common Random Sampling Techniques

Random Number Table

Random number tables are created when every person or every item receives a number. The numbers are entered into a table with digits, starting with the number one and including a number for every person or item. The numbers are added onto the table in a random order.

To use this method for random sampling, each person in the population receives a unique number that is included on the table. Numbers are chosen at random from the table. The choice of one digit is unaffected by the choice of any other given digit.

An example of using a random number table is to assign a number to each of 100 people who have expressed interest in attending a special event. Individual numbers are chosen from the random number table. When a person's number is chosen, they are approved to attend the event.

- This is random sampling because any number can be chosen from the table. The numbers don't have to be chosen in numerical order.
- Each number has an equal opportunity to be chosen from the table.

Replacement Sampling

With Replacement

Sampling with replacement is the act of choosing an individual once and then replacing their number or name into the original group of potential people such that the same person has the ability to be chosen more than once.

Without Replacement

This procedure is the same as sampling with replacement, except that the name or number of the individual is not replaced into the original group. This results in only one opportunity to be chosen, rather than multiple.

Lottery

All of the names or assigned numbers of individuals are entered into a given group and then chosen at random.

Using a Computer

Software and other programs are available for the purpose of making random samples.

Real-World Examples of Random Sampling

- At a birthday party, teams are chosen by putting everyone's name into a jar, then choosing names for each team.
- To test the quality of a river's water, samples of the water would be taken from multiple places in the river, on different dates and at different times of the day. All of the samples would be numbered with the numbers entered into a table. As numbers are chosen from the table, quality tests are run on the chosen sample.
- For a study, psychologists assign numbers to all volunteers for the trial. A number table is used to assign volunteers to the control group and to the four test groups.
- To determine what class to put students into at a school, names are entered into a software program, which then randomly places students in each class.
- To determine the quality of items produced on a factory line, a group of produced items would be removed from the line, each receiving a unique number and then the numbers would be entered into a table and selected one at a time until the quality control team has the number of items required for testing.
- At a fundraising event, every attendees name is entered for the chance to win a prize. Names are selected at random, then returned to the jar, giving individuals multiple opportunities to win prizes.
- Students are given an opportunity to win one of five prizes if they have displayed good behavior. Each child who has displayed good behavior throughout the marking period gets one ticket to put into a bowl. Then, five separate names are pulled from the bowl, without replacing them.
- In a warehouse with an assembly line operation, each employee is assigned a random number using computer software. The same software is used periodically to choose a number of one of the employees to be observed to ensure he is employing best practices.
- A company assigns random numbers using a number table to employees. Every week, a number is chosen and that employee is expected to go for blood tests to ensure sobriety.
- A restaurant leaves a fishbowl on the counter for diners to drop their business cards. Once a month, a business card is pulled out to award one lucky diner a discounted dinner.
- An organization uses a computer program to assign numbers to all of their clients. Once a week, the organization uses the computer program to choose a client for the firm to provide lunch for as a reward for their business.

Remember, there are many ways to implement random sampling. What they all have in common is that each person or item has an equal chance to be chosen.

Sampling distributions of estimators

• Since our estimators are statistics (particular functions of random variables), their distribution can be derived from the joint distribution of X1 . . . Xn. It is called the sampling distribution because it is based on the joint distribution of the random sample.

• Given a sampling distribution, we can

– calculate the probability that an estimator will not differ from the parameter θ by more than a specified amount

– obtain interval estimates rather than point estimates after we have a sample- an interval estimate is a random interval such that the true parameter lies within this interval with a given probability (say 95%).

– choose between to estimators- we can, for instance, calculate the mean-squared error of the estimator, $E_\theta[(\hat{\theta} - \theta)^2]$ using the distribution of $\hat{\theta}$.

Sampling distributions of estimators depend on sample size, and we want to know exactly how the distribution changes as we change this size so that we can make the right trade-offs between cost and accuracy.

Sampling distributions: sample size and precision

Examples:

1.What if $X_i \sim N(\theta, 4)$, and we want $E(\bar{X}_n - \theta)^2 \leq .1$? This is simply the variance of $\bar{X}_n$, and we know $\bar{X}_n \sim N(\theta, 4/n)$.

$$4/n \leq .1 \text{ if } n \geq 40$$

2. Consider a random sample of size n from a Uniform distribution on [0, θ], and the statistic U = max{X1, . . . ,Xn}. The CDF of U is given by:

$$F(X) = \begin{cases} 0 & \text{if } u \leq 0 \\ \left(\frac{u}{\theta}\right)^n & \text{if } 0 < u < \theta \\ 1 & \text{if } u \geq \theta \end{cases}$$

We can now use this to see how large our sample must be if we want a certain level of precision in our estimate for θ. Suppose we want the probability that our estimate lies within .1θ for any level of θ to be bigger than 0.95:

$$\Pr(|U - \theta| \leq .1\theta) = \Pr(\theta - U \leq .1\theta) = \Pr(U \geq .9\theta) = 1 - F(.9\theta) = 1 - 0.9^n$$

We want this to be bigger than 0.95, or $0.9^n \leq 0.05$. With the LHS decreasing in n, we choose $n \geq \frac{\log(.05)}{\log(.9)} = 28.43$. Our minimum sample size is therefore 29.

The Method of Moments

- One of the oldest methods; very simple procedure
- What is Moment?
- Based on the assumption that sample moments should provide GOOD ESTIMATES of the corresponding population moments.

How it works?

THE METHOD OF MOMENTS PROCEDURE

Suppose there are l parameters to be estimated, say $\theta = (\theta_1, \ldots, \theta_l)$.

1. Find l population moments, $\mu'_k, k = 1, 2, \ldots, l$. μ'_k will contain one or more parameters $\theta_1, \ldots, \theta_l$.
2. Find the corresponding l sample moments, $m'_k, k = 1, 2, \ldots, l$. The number of sample moments should equal the number of parameters to be estimated.
3. From the system of equations, $\mu'_k = m'_k, k = 1, 2, \ldots, l$, solve for the parameter $\theta = (\theta_1, \ldots, \theta_l)$; this will be a moment estimator of $\hat{\theta}$.

$$\mu'_k = E[X^k]$$

$$m'_k = (1/n)\sum_{i=1}^{n} X_i^k \qquad m_1' = \bar{X};\ m_2' = (1/n)\sum_{i=1}^{n} X_i^2$$

$$\mu'_k = m'_k$$

Example: normal distribution

$$X_1, X_2, \ldots X_n \ \text{iid}\sim N(\tau, \sigma^2).$$

step 1, $\mu_1' = E(X) = \tau;\ \mu_2' = E(X^2) = \tau^2 + \sigma^2.$

step 2, $m_1' = \bar{X};\ m_2' = (1/n)\sum_{i=1}^{n} X_i^2.$

step 3, Set $\mu_1' = m_1', \mu_2' = m_2'$, therefore,

$$\tau = \bar{X},$$

$$\tau^2 + \sigma^2 = (1/n)\sum_{i=1}^{n} X_i^2$$

Solving the two equations, we get $\hat{\tau} = \bar{X}, \hat{\sigma}^2 = (1/n)\sum_{i=1}^{n} X_i^2 - \bar{X}^2$

Example: Bernoulli Distribution

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with parameter p.

(a) Find the moment estimator for p.

Solution

(a) *For the Bernoulli random variable, $\mu_k' = E[X] = p$, so we can use m_1' to estimate p. Thus,*

$$m_1' = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

X follows a Bernoulli distribution, if $P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$

Example: Poisson distribution

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\lambda > 0$. Show that both $(1/n)\sum_{i=1}^{n} X_i$ and $(1/n)\sum_{i=1}^{n} X_i^2 - \left((1/n)\sum_{i=1}^{n} X_i\right)^2$ are moment estimators of λ.

Solution

We know that $E(X) = \lambda$, from which we have a moment estimator of λ as $(1/n)\sum_{i=1}^{n} X_i$. Also, because we have $Var(X) = \lambda$, equating the second moments, we can see that

$$\lambda = E(X^2) - (EX)^2,$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

so that

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Both are moment estimators of λ. Thus, the moment estimators may not be unique. We generally choose $\overline{X}$ as an estimator of λ, for its simplicity.

Pros of Method of Moments

Easy to compute and always work:

- The method often provides estimators when other methods fail to do so or when estimators are hard to obtain (as in the case of gamma distribution).
- MME is consistent.

Cons of Method of Moments

- They are usually not the “best estimators” available. By best, we mean most efficient, i.e., achieving minimum MSE.
- Sometimes it may be meaningless.

The Method of Maximum Likelihood

- Proposed by geneticist/statistician:

  Sir Ronald A. Fisher in 1922
- Idea: We attempt to find the values of the parameters which would have most likely produced the data that we in fact observed.

What is likelihood?

➢ Definition 5.3.1 *Let* $f(x_1, \ldots, x_n; \theta), \theta \in \Theta \subseteq \mathbb{R}^k$, *be the joint probability (or density) function of n random variables* $X_1, \ldots, X_n$ *with sample values* $x_1, \ldots, x_n$. *The* likelihood function *of the sample is given by*

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta), [= L(\theta), \textit{in a briefer notation}].$$

We emphasize that L is a function of θ *for fixed sample values.*

➢ E.g., Likelihood of θ=1 is the chance of observing $X_1, X_2, \ldots, X_n$ when θ=1.

How to compute Likelihood?

- If $X_1, \ldots, X_n$ are discrete iid random variables with probability function $p(x, \theta)$, then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i), \quad \text{(by multiplication rule for independent random variables)} \\ &= \prod_{i=1}^{n} p(x_i, \theta) \end{aligned}$$

- and in the continuous case, if the density is $f(x, \theta)$, then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Example of computing likelihood (discrete case)

Suppose $X_1, \ldots, X_n$ are a random sample from a geometric distribution with parameter $p, 0 \leq p \leq 1$.

Solution

For the geometric distribution, the pmf is given $p(1-p)^{x-1}, \quad 0 \leq p \leq 1, \quad x = 1, 2, 3, \ldots.$

Hence, the likelihood function is

$$L(p) = \prod_{i=1}^{n} \left[p(1-p)^{x-1} \right] = p^n (1-p)^{-n + \sum_{i=1}^{n} x_i}.$$

Example of computing likelihood (continuous case)

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. Let $x_1, \ldots, x_n$ be the sample values. Find the likelihood function.

Solution

The density function for the normal variable is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. *Hence, the likelihood function is*

$$L\left(\mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

Definition of MLE

➢Definition 5.3.2 *The* maximum likelihood estimators (MLEs) *are those values of the parameters that maximize the likelihood function with respect to the parameter* θ. *That is,*

$$L\left(\hat{\theta}; x_1, \ldots, x_n\right) = \max_{\theta \in \Theta} L\left(\theta; x_1, \ldots, x_n\right)$$

where Θ *is the set of possible values of the parameter* θ.

- In general, the method of ML results in the problem of maximizing a function of single or several parameters. One way to do the maximization is to take derivative.

Procedure to find MLE

1. Define the likelihood function, $L(\theta)$.
2. Often it is easier to take the natural logarithm (ln) of $L(\theta)$.
3. When applicable, differentiate $\ln L(\theta)$ with respect to θ, and then equate the derivative to zero.
4. Solve for the parameter θ, and we will obtain $\hat{\theta}$.
5. Check whether it is a maximizer or global maximizer.

Example: Poisson Distribution

Suppose $X_1, \ldots, X_n$ are random samples from a Poisson distribution with parameter λ. Find MLE $\hat{\lambda}$.

Solution

We have the probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0.$$

Hence, the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.$$

Then, taking the natural logarithm, we have

$$\ln L(\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln (x_i!)$$

Example cont'd

and differentiating with respect to λ results in

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

and

$$\frac{d \ln L(\lambda)}{d\lambda} = 0, \text{ implies } \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

That is,

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

Hence, the MLE of λ is

$$\hat{\lambda} = \overline{X}.$$

Example: Uniform Distribution

Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta), \theta > 0$. Find the MLE of θ.

Solution

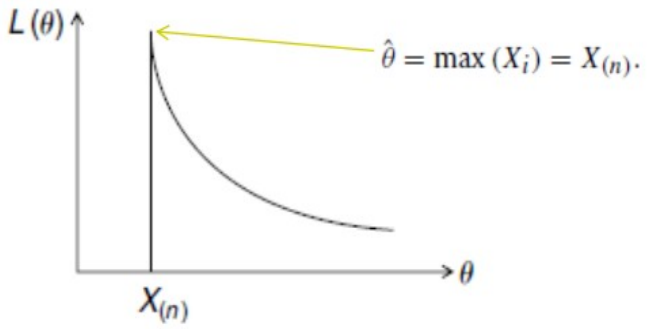
Note that the pdf of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the likelihood function is given by

$$L(\theta, x_1, x_2, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & 0 \le x_1, x_2, \ldots, x_n \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

Example cont'd



■ **FIGURE 5.1** Likelihood function for uniform probability distribution.

More than one parameter

As mentioned earlier, if the unknown parameter θ represents a vector of parameters, say $\theta = (\theta_1, \ldots, \theta_l)$, then the MLEs can be obtained from solutions of the system of equations

$$\frac{\partial}{\partial \theta} \ln L(\theta_1, \ldots, \theta_n) = 0, \quad \text{for} \quad i = 1, \ldots, l.$$

These are called the *maximum likelihood equations* and the solutions are denoted by $(\hat{\theta}_1, \ldots, \hat{\theta}_l)$.

Pros of Method of ML

- When sample size n is large (n>30), MLE is unbiased, consistent, normally distributed, and efficient ("regularity conditions")
- "Efficient" means it produces the minimum MSE than other methods including Method of Moments
- More useful in statistical inference.

Cons of Method of ML

- MLE can be highly biased for small samples.
- Sometimes, MLE has no closed-form solution.

- MLE can be sensitive to starting values, which might not give a global optimum. Common when (THETA SIGN) is of high dimension

How to maximize Likelihood

1. Take derivative and solve analytically (as aforementioned)

2. Apply maximization techniques including Newton's method, quasi-Newton method (Broyden 1970), direct search method (Nelder and Mead 1965), etc.

- These methods can be implemented by R function optimize(), optim()