Random Variables, Sampling, Estimation and Inference

Parthasarathi Edupally[1]

[1]DataConscientious LLP, Mumbai

Elements of Econometrics,
Russell Square International College, Mumbai

Outline

Outline

Outline

Outline

Outline

Objectives

- Review basic concepts of Random Variables, Sampling, Estimation and Inference

Objectives

- Review basic concepts of Random Variables, Sampling, Estimation and Inference
- Intiutively understand these concepts with examples

Objectives

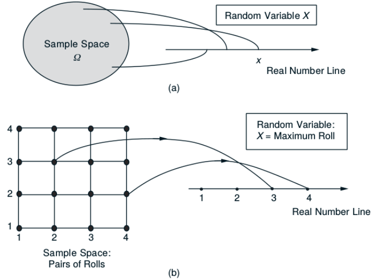
- Review basic concepts of Random Variables, Sampling, Estimation and Inference
- Intiutively understand these concepts with examples
- Be able to use them in a practical problem setting

Definition

- In a given probabilistic experiment, a random variable X is a function from sample space to real line

Definition

- In a given probabilistic experiment, a random variable X is a function from sample space to real line



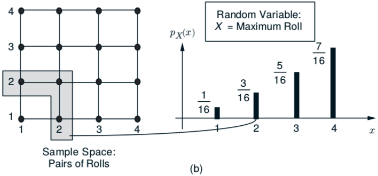
- Intuitively, we can think of rv as a bunch of values from real line
- It is often much easier to deal with summary variables than the actual outcomes

Probability Distribution of RV

- It is description of all possible values, a RV can take along with their frequencies of occurance (probabilities)

Probability Distribution of RV

- It is description of all possible values, a RV can take along with their frequencies of occurance (probabilities)

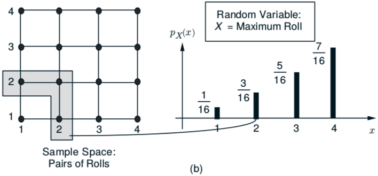


(b)

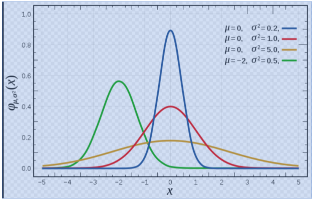
- Other way of describing a distribution of RV is through its cumulative distribution function (cdf)

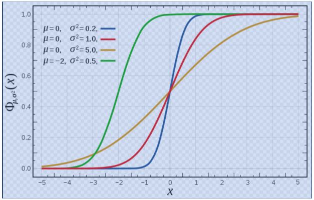
Probability Distribution of RV

- It is description of all possible values, a RV can take along with their frequencies of occurance (probabilities)



- Other way of describing a distribution of RV is through its cumulative distribution function (cdf)
- Example: Distribution of Normal RV





Expected Value of a RV

- It is merely average value of a RV, where average is taken after weighting values with their probabilities

Expected Value of a RV

- It is merely average value of a RV, where average is taken after weighting values with their probabilities
- Can think of it as a typical or expected value of an observation of the random variable

$$E(X) = \begin{cases} \sum_{x \in X} xP(X = x), & \text{if X is discrete} \\ \int_{-\infty}^{\infty} xf(x)dx, & \text{if X is continuous} \end{cases}$$

Expected Value of a RV

- It is merely average value of a RV, where average is taken after weighting values with their probabilities
- Can think of it as a typical or expected value of an observation of the random variable

$$E(X) = \begin{cases} \sum_{x \in X} xP(X = x), & \text{if X is discrete} \\ \int_{-\infty}^{\infty} xf(x)dx, & \text{if X is continuous} \end{cases}$$

- Expectation is a linear operation and hence following identities hold
  - $E(g(X)) = \sum_{x \in X} g(x)P(X = x)$
  - $E(aX + bY + c) = aE(X) + bE(Y) + c$

Properties of RVs

- Variance of a RV is measure of dispersion of its distribution and is given by

$$VarX = E[(X-\mu)^2], \quad \text{where } \mu \text{ is expected value of RV}$$

Properties of RVs

- Variance of a RV is measure of dispersion of its distribution and is given by

$$VarX = E[(X-\mu)^2], \quad \text{where } \mu \text{ is expected value of RV}$$

- If (X, Y) is a bivariate RV and have marginal pdfs $f(x)$ and $f(y)$, then they are independent if

$$f(x,y) = f(x)f(y)$$

Properties of RVs

- Variance of a RV is measure of dispersion of its distribution and is given by

$$VarX = E[(X-\mu)^2], \quad \text{where } \mu \text{ is expected value of RV}$$

- If (X, Y) is a bivariate RV and have marginal pdfs $f(x)$ and $f(y)$, then they are independent if

$$f(x,y) = f(x)f(y)$$

Properties of RVs

- Variance of a RV is measure of dispersion of its distribution and is given by

$$VarX = E[(X-\mu)^2], \quad \text{where } \mu \text{ is expected value of RV}$$

- If (X,Y) is a bivariate RV and have marginal pdfs $f(x)$ and $f(y)$, then they are independent if

$$f(x,y) = f(x)f(y)$$

- To quantify extent of linear relation between two RVs X and Y, we define
    - Covariance of X and Y as

    $$Cov(X,Y) = E((X-E(X))(Y-E(Y)))$$

    - Correlation of X and Y as

    $$\rho_{xy} = Cov(X,Y)/\sqrt{Var(X)}\sqrt{Var(Y)}$$

Population, Sample and Realization

- Data collected in an experiment consists of several observations of variables of interest
- For example, to understand the effect of education on future employment of individuals in a particular state
    - data consists of age, education, current employment and other factors of n individuals from the population of interest

Population, Sample and Realization

- Data collected in an experiment consists of several observations of variables of interest
- For example, to understand the effect of education on future employment of individuals in a particular state
  - data consists of age, education, current employment and other factors of n individuals from the population of interest
- From statistical point of view, we have two sets of RVs in this experiment

Population, Sample and Realization

- Data collected in an experiment consists of several observations of variables of interest
- For example, to understand the effect of education on future employment of individuals in a particular state
  - data consists of age, education, current employment and other factors of n individuals from the population of interest
- From statistical point of view, we have two sets of RVs in this experiment
  - one corresponding to the whole population called population RVs
  - other corresponds to the sample, sample RVs, because we could have chosen any set of n individuals from population (so sample is itself random)
- After the observation is made, it is called 'realization'of the corresponding RV

Random Sample

- Consider observing age of individuals in a population:
  - these sample observations $X_1, X_2 \ldots X_n$ of age are from the same population RV X
  - also each observation is independent of the other

Random Sample

- Consider observing age of individuals in a population:
  - these sample observations $X_1, X_2 ... X_n$ of age are from the same population RV X
  - also each observation is independent of the other
- This is an example of Random sample
- Note: $X_1, X_2 ... X_n$ are RVs (sample RVs) because we could have easily realized different set of n individuals if experiment was conducted again!
- Note: They are all having same distribution as X (population RV)

Random Sample

- Consider observing age of individuals in a population:
  - these sample observations $X_1, X_2 ... X_n$ of age are from the same population RV X
  - also each observation is independent of the other
- This is an example of Random sample
- Note: $X_1, X_2 ... X_n$ are RVs (sample RVs) because we could have easily realized different set of n individuals if experiment was conducted again!
- Note: They are all having same distribution as X (population RV)
- Alternatively, $X_1, X_2 ... X_n$ are called independent and identically distributed RVs

Point Estimation

- Why do we need to estimate population parameter?

Point Estimation

- Why do we need to estimate population parameter?
  - The population of interest is completely defined by the probability density function $f(x|\theta)$
  - Thus parameter θ yields knowledge of the entire population
  - Also there might be important physical interpretation for θ (for example population mean)

Point Estimation

- Why do we need to estimate population parameter?
  - The population of interest is completely defined by the probability density function $f(x|\theta)$
  - Thus parameter θ yields knowledge of the entire population
  - Also there might be important physical interpretation for θ (for example population mean)
- Estimator vs Estimate
  - Estimator of a population parameter is some function of the sample,

$$\theta = W(X_1, X_2 ... X_n), \quad \text{where } X_1, X_2 ... X_n \text{ are sample RVs}$$

Point Estimation

- Why do we need to estimate population parameter?
  - The population of interest is completely defined by the probability density function $f(x|\theta)$
  - Thus parameter θ yields knowledge of the entire population
  - Also there might be important physical interpretation for θ (for example population mean)
- Estimator vs Estimate
  - Estimator of a population parameter is some function of the sample,

$$\theta = W(X_1, X_2...X_n), \quad \text{where } X_1, X_2...X_n \text{ are sample RVs}$$

  - So θ is a Random Variable in itself

Point Estimation

- Why do we need to estimate population parameter?
  - The population of interest is completely defined by the probability density function $f(x|\theta)$
  - Thus parameter θ yields knowledge of the entire population
  - Also there might be important physical interpretation for θ (for example population mean)
- Estimator vs Estimate
  - Estimator of a population parameter is some function of the sample,

$$\theta = W(X_1, X_2 ... X_n), \quad \text{where } X_1, X_2 ... X_n \text{ are sample RVs}$$

  - So θ is a Random Variable in itself
  - Estimate is the realized value of an estimator,

$$\theta = W(x_1, x_2 ... x_n), \quad \text{where } x_1, x_2 ... x_n \text{ are realized values}$$

Point Estimation

- Sample mean can be used as an esimator of population mean,

$$\overline{\mathrm{X}} = (X_1 + X_2 + X_3 ... + X_n)/n$$

- Sample variance is a good estimator of population variance,

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{\mathrm{X}})^2/(n-1)$$

Point Estimation

- Sample mean can be used as an esimator of population mean,

$$\overline{\mathrm{X}} = (X_1 + X_2 + X_3 ... + X_n)/n$$

- Sample variance is a good estimator of population variance,

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{\mathrm{X}})^2/(n-1)$$

- When we are using a qualifier 'good ', we have to be more precise,
  - this leads us to three properties of estimators - unbiasedness, efficiency and consistency

Unbiasedness and Efficiency

- Mean Square Error (MSE) to evaluate an estimator:
  - MSE of an estimator W of a parameter θ is a function of θ defined as $E_\theta(W-\theta)^2$,

$$E_\theta(W-\theta)^2 = \underbrace{Var_\theta W}_{\text{Variance}} + \underbrace{(E_\theta W-\theta)^2}_{\text{Bias}}$$

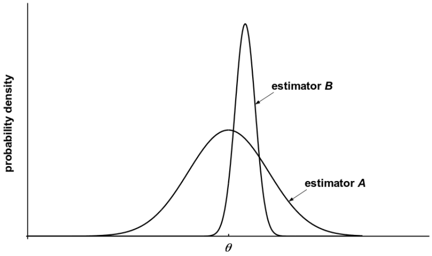
  - MSE incorporates to components, one measuring the variability of estimator (precision) and the other its bias (accuracy)
  - An estimator is '**Unbiased**'if its bias term is zero
  - For example, population mean estimator $\overline{X}$ is unbiased since $E\overline{X} = \mu_x$
- An estimator is said to be '**Efficient** 'if its variance component is least among all the estimators with same bias

Unbiasedness and Efficiency

- Evaluating an estimator based on MSE is a trade-off between Variance and bias components
- An estimator which is biased could be a better estimator due to less variance

Unbiasedness and Efficiency

- Evaluating an estimator based on MSE is a trade-off between Variance and bias components
- An estimator which is biased could be a better estimator due to less variance



- Unbiasedness and efficiency are finite sample properties of estimators

Consistency of an Estimator

- Asymptotic properties like consistency are used to evaluate an estimator as the sample size tends to infinity

Consistency of an Estimator

- Asymptotic properties like consistency are used to evaluate an estimator as the sample size tends to infinity
- Probability limits (Convergence in probability)
  - A sequence of random variables, $X_1, X_2...$, converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

  - Intuitively, limit implies that for large n ‘bulk ’of the distribution of X_n is concentrated near X

Consistency of an Estimator

- Asymptotic properties like consistency are used to evaluate an estimator as the sample size tends to infinity
- Probability limits (Convergence in probability)
  - A sequence of random variables, $X_1, X_2 ...$, converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0$$

  - Intuitively, limit implies that for large n 'bulk 'of the distribution of X_n is concentrated near X
- An estimator is said to be consistent if it satisfies these two properties:
  - the estimator possesses probability limit
  - and the limit is the true value of the population parameter

Consistency of an Estimator

- For example, Sample mean is a consistent estimator of the population mean

$$\lim_{n \to \infty} P(|\overline{X} - \mu_x| \geq \epsilon) = 0$$

Consistency of an Estimator

- For example, Sample mean is a consistent estimator of the population mean

$$\lim_{n\to\infty} P(|\overline{X} - \mu_x| \geq \epsilon) = 0$$

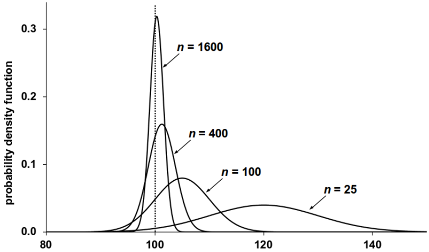
- Unbiasedness is sufficient but not necessary condition for consistency
- An estimator may become unbiased as the sample size increases:

Consistency of an Estimator

- For example, Sample mean is a consistent estimator of the population mean

$$\lim_{n \to \infty} P(|\overline{X} - \mu_x| \geq \epsilon) = 0$$

- Unbiasedness is sufficient but not necessary condition for consistency
- An estimator may become unbiased as the sample size increases:



Hypothesis Testing

- It is yet another method of inference like point estimation
- A hypothesis is a statement about the population parameter
- Goal of a hypothesis test is to decide, based on sample from the population, which of two complementary hypothesis (Null and Alternative) is true
- Hypothesis testing procedure is a rule that species for which values of sample do we accept one hypothesis over the other
- Some of the popular Hypothesis testing procedures include LRT, Bayesian tests etc

Hypothesis Testing - Example

Dougherty Slides from Study Guide

Evaluating Hypothesis tests

Dougherty Slides from Study Guide

Hypothesis tests - Confidence Interval approach

Dougherty Slides from Study Guide

"In God we trust, all others bring data."

William Edwards Deming (1900 - 1993).