Session 11: Random Variable And Normal Distribution

Agenda

Sl.No.	Agenda Topics
1.	Random Variables
2.	Example
3.	Events
4.	Probability Of Events
5.	Probability
6.	Expected Value
7.	Example
8.	What Is A Large Enough Sample Size?
9.	Probability Distribution
10.	Standard Deviation And Variance
11.	Reason
12.	Types Of Distributions
13.	Discrete Probability Distributions- The Binomial Distribution
14.	Negative Binomial Distribution

Sl.No.	Agenda Topics
15.	Geometric Distribution
16.	Continuous Probability Distribution
17.	Working On Distributions Using R
18.	Normal Distribution
19.	What's So Important About The Normal Distribution?
20.	The Standard Normal Distribution
21.	Question
22.	How Skewness And Kurtosis Affect Your Distribution
23.	Skewness Interpretation
24.	Kurtosis
25.	Kurtosis: Interpretation
26.	Quantile-Quantile (q-q) Plots
27.	R Code For Summary Stats
28.	R Code for Normal Distribution
29.	Sampling Types

Random Variable

- A random variable describes the probability of an uncertain future numerical outcome of a random process
- It is random because there is some chance associated with each possible value
- All random variables have three aspects:
  - A distribution
  - A mean
  - A standard deviation

- X = a random variable designating the outcome of a single event

- Mean of X = μ; Standard deviation of X = σ

- $\bar{X}$ = a random variable designating the average outcome of n measurements of the event

- Mean of $\bar{X}$ = μ; Standard deviation of $\bar{X}$ = $\sigma/\sqrt{n}$

Example

Experiment- Two cards randomly selected

Let *X* be the number of diamonds selected

$$S = \left\{ \begin{array}{l} CC\ CD\ CH\ CS \\ DC\ DD\ DH\ DS \\ HC\ HD\ HH\ HS \\ SC\ SD\ SH\ SS \end{array} \right\}$$

Events

- Events can be described in terms of random variables

Example:

- X=1 is the event that exactly one diamond is selected
- X<=1 is the event that at most one diamond is selected

Probability Of Events

- Probabilities of events can be stated as probabilities of the corresponding values of *X*

$$P(X = 1) = P(F) = \frac{6}{16} = \frac{3}{8}$$

$$P(X \leq 1) = P\left(\begin{Bmatrix} CC & CD & CH & CS & DC \\ DH & DS & HC & HD & HH \\ HS & SC & SD & SH & SS \end{Bmatrix}\right) = \frac{15}{16}$$

Probability

In general,

P(X=x) is the probability that *X* takes on the value *x*

P(X<=x) is the probability that *X* takes on a value that is less than or equal to *x*

Suppose, *X* can only assume the values x_1, x_2, ... x_n. Then

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

Expected Value

- **The mean (or expected value)** of *X* gives the value that we would expect to observe on average in a large number of repetitions of the experiment

$$\mu_X = E(X) = \sum_{i=1}^{n} x_i * P(X = x_i)$$

Example

- Attendance at a basketball game averages 80000 with a standard deviation of 6000.
- X = Attendance at a gam
- $\mu = 80000$; $\sigma = 6000$
- X = Average attendance at n games
- Mean of $\bar{X} = 8000$

  $\bar{X} =$ $=$
- Standard deviation of $\bar{X} = 6000/\sqrt{n}$

What Is A Large Enough Sample Size?

- To determine whether or not X can be approximated by a normal distribution, typically n = 30 is used as a breakpoint

Probability Distribution

- A probability distribution is a rule that identifies possible outcomes of a random variable and assigns a probability to each
- A discrete distribution has a finite number of values, e.g. Price of coke, age of employees
- A continuous distribution has all possible values in some range, e.g. sales per week, weight of students

Standard Deviation And Variance

- The variance (σ2) of a data set is calculated by taking the arithmetic mean of the squared differences between each value and the mean value or the weighted average of the squared deviations from the mean

σ = population standard deviation

σ2 = population variance

s = estimate of population standard deviation based on sampled data

s^2 = estimate of population variance based on sampled data

The population variance is defined as:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

- **Standard deviation** is simply the square root of the **variance**

Note - *Because the differences are squared, the units of variance are not the same as the units of the data. Therefore, the standard deviation is reported as the square root of the variance and the units then correspond to those of the data set*.

Standard Deviation And Variance (Contd.)

- The variance of a sampled subset of observations is calculated in a similar manner
- However, while the sample mean is an unbiased estimator of the population mean, the same is not true for the sample variance, if it is calculated in the same manner as the population variance
- This corrected sample variance is defined as:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Reason

- If one takes all possible samples of **n** members and calculate the sample variance of each combination using **n** in the denominator and get an average of the results, the value would not be equal to the true value of the population variance; that is, it would be biased. This bias can be corrected by using **(n - 1)** in the denominator instead of just **n**, in which case the sample variance becomes an unbiased estimator of the population variance.

Types Of Distributions

- Discrete probability distributions
- Continuous probability distributions

Discrete Probability Distributions- The Binomial Distribution

- A binomial random variable is the number of successes in a series of trials, for example, the number of 'tails' occurring when a coin is tossed 200 times.
- A discrete random variable, X is said to follow a Binomial distribution with parameters n and p,if it has probability distribution

P(X=x) =(nCx)*p^x* (1-p)* (n-x)

where

x = 0, 1, … , n

n = 1, 2, …

p = success probability; 0 < p < 1

Binomial Distribution (Contd..)

The trials must have the following characteristics:

- the total number of trials is fixed in advance
- there are just two possible outcomes of each trial: success or failure
- the outcomes of all the trials are statistically independent
- all the trials have the same probability of success

- Mean (μ) of the Binomial Distribution:

$\mu = np$

- Variance ($\sigma 2$) of the Binomial Distribution:

$\sigma 2 = np(1-p)$

Negative Binomial Distribution

- Consider a statistical experiment where a success occurs with probability p. If the experiment is repeated indefinitely and the trials are independent of each other, then the random variable X, whose value is the number of the trial on which the r^{th} success occurs, has a negative binomial distribution with parameters r and p.
- Mean (μ) of the Negative Binomial Distribution:

  $\mu = k/p$

- Variance (σ2) of the Negative Binomial Distribution:

  $\sigma 2 = k(1-p)/p2$

Geometric Distribution

- A Geometric random variable is the number of trials required to obtain the first success.
- A discrete random variable X is said to follow a Geometric distribution with parameter p, if it has probability distribution:

$$P(X=x) = p(1-p)x-1$$

where

$x = 1, 2, 3, \ldots$

p = success probability; $0 < p < 1$

Geometric Distribution (Contd..)

- The trials must meet the following requirements:
  - the total number of trials is potentially infinite
  - there are just two outcomes of each trial; success and failure
  - the outcomes of all the trials are statistically independent
  - all the trials have the same probability of success

- **Mean (μ) of the Geometric Distribution:**

$$\mu = 1/p$$

- **Variance (σ^2) of the Geometric Distribution:**

$$\sigma^2 = (1-p)/p^2$$

Continuous Probability Distribution

- Normal distribution
- Exponential distribution (covered later)
- Gamma Distribution (out of scope)
- Beta Distribution (out of scope)

Working On Distributions Using R

- pnorm is the R function that calculates the c. d. f. (the cumulative distribution function (c. d. f.))
- $F(x) = P(X <= x)$, where X is normal

```
>pnorm(32.6, mean=45, sd=15)
>pnorm(32.6, 45, 15)
```

- qnorm is the R function that calculates the inverse c. d. f. , F^{-1} of the normal distribution. The c. d. f. and the inverse c. d. f. are related by:

$$p = F(x)$$

$x = F^{-1}(p)$ So given a number p between zero and one, qnorm looks up the p-th quantile of the normal distribution. As with pnorm, optional arguments specify the mean and standard deviation of the distribution.

Working On Distributions Using R (Contd.)

- For every distribution there are four commands. The commands for each distribution are prepended with a letter to indicate the functionality:
  - "d" returns the height of the probability density function
  - "p" returns the cumulative density function
  - "q" returns the inverse cumulative density function (quantiles)
  - "r" returns randomly generated numbers

```
> help(Distributions)
> help(Normal)
> help(TDist)
> help(Binomial)
> help(Chisquare)
```

Working On Distributions Using R (Contd.)

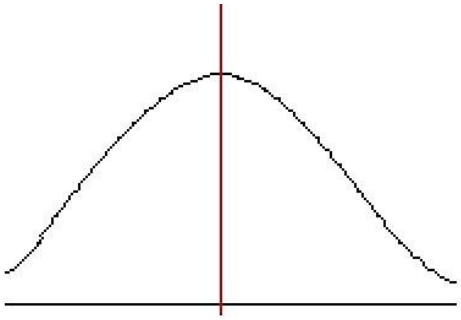
- dnorm is the R function that calculates the p. d. f. f of the normal distribution. As with pnorm and qnorm, optional arguments specify the mean and standard deviation of the distribution
- rnorm is the R function that simulates random variates having a specified normal distribution. As with pnorm, qnorm, and dnorm, optional arguments specify the mean and standard deviation of the distribution

Working On Distributions Using R (Contd.)

- dbinom is the R function that calculates the p. f. of the binomial distribution
- pbinom is the R function that calculates the c. d. f. of the binomial distribution
- qbinom is the R function that calculates the "inverse c. d. f." of the binomial distribution.
- The quantile is defined as the smallest value *x* such that *F(x) >= p*, where *F* is the distribution function

The Normal Distribution

- A continuous random variable X follows a normal distribution if it has the following probability density function (p.d.f.)
- The parameters of the distribution are m and s2, where m is the mean (expectation) of the distribution and s2 is the variance. We write X ~ N(m, s2) to mean that the random variable X has a normal distribution with parameters m and s2.
- The normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
- symmetrical about its mean:



What's So Important About The Normal Distribution?

- One reason why the normal distribution is important is because many psychological and educational variables are distributed approximately, normally
- The second reason why the normal distribution is so important is because it is easy for mathematical statisticians to work with
- If a random variable X follows the normal distribution, then we write:

$$X \sim N(\mu, \sigma^2)$$

The Standard Normal Distribution

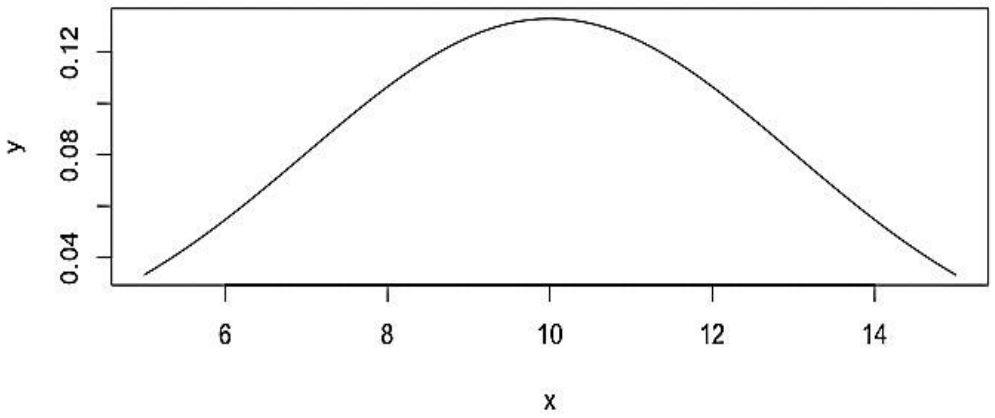
- If $Z \sim N(0, 1)$, then Z is said to follow a standard normal distribution

- $P(Z < z)$ is known as the cumulative distribution function of the random variable Z

Question

- Using R, plot a normal distribution for mean =10 and standard deviation =3

```
>x <- seq(5,15,length=1000)
>y <- dnorm(x,mean=10, sd=3)
>plot(x,y, type="l", lwd=1)
```



Question

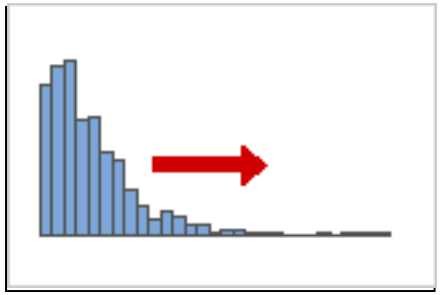
- Question - Assume that the scores of a college entrance test has normal distribution. Also, the mean score is 64, and the standard deviation is 10.5. What is the percentage of students scoring 88 or more in the exam?
- Answer - Using R to solve this question, we apply the function pnorm of the normal distribution with mean 64 and standard deviation 10.5. Since we are looking for the percentage of students scoring higher than 88, we are interested in the upper tail of the normal distribution.
- Thus 11.1% students scored above 88 marks

```
> pnorm(88, mean=64, sd=10.5, lower.tail=FALSE)
[1] 0.01113549
```

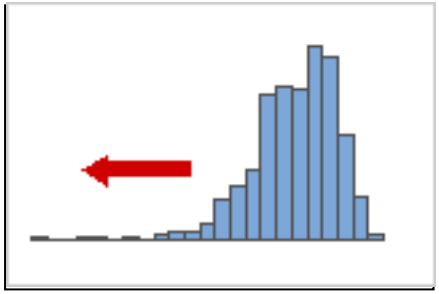
How Skewness And Kurtosis Affect Your Distribution

- Skewness is the extent to which the data is non-symmetrical
- Whether the skewness value is 0, positive, or negative reveals information about the shape of the data. As data becomes more symmetrical, its skewness value approaches zero

Positive or right skewed distributions
Positive skewed or right skewed data is so named because the "tail" of the distribution points to the right and its skewness value will be greater than 0 (or positive).



Negative or left skewed distributions
Left skewed or negative skewed data is so named because the "tail" of the distribution points to the left, and it produces a negative skewness value.



Skewness Interpretation

- Skewness > 0: Right skewed distribution - most values are concentrated on the left of the mean, with extreme values to the right

- Skewness < 0: Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left

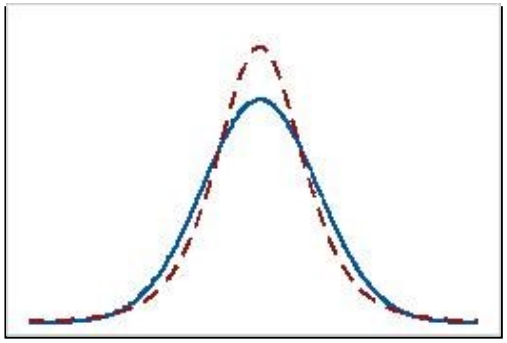
- Skewness = 0: mean = median, the distribution is symmetrical around the mean.

Kurtosis

- Kurtosis indicates how the peak and tails of a distribution differ from the normal distribution. Use kurtosis to help you initially understand general characteristics about the distribution of your data.

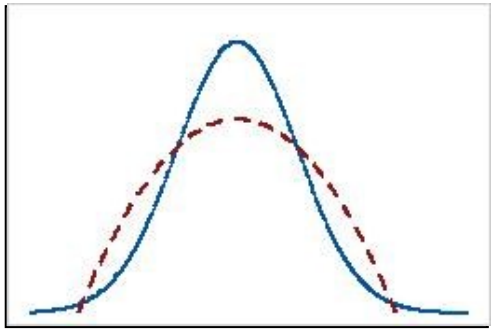
Positive kurtosis

A distribution with a positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.



Negative kurtosis

A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.



Kurtosis: Interpretation

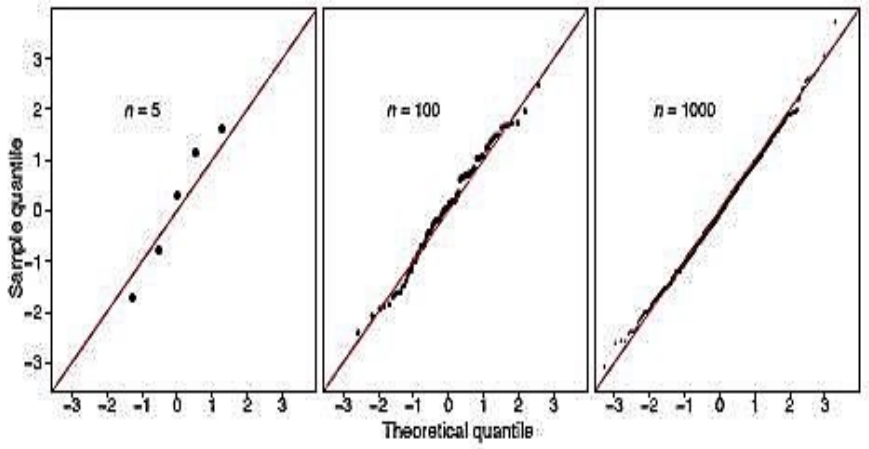
- Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values

- Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme value is less than for a normal distribution, and the values are wider spread around the mean

- Kurtosis = 3 - Mesokurtic distribution has normal distribution

Quantile-Quantile (q-q) Plots

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- The advantages of the q-q plot are:
  - The sample size need not equal
  - Many distributional aspects can be simultaneously tested
- 3rd Graph shows QQ plot of a normal distribution



R Code For Summary Stats

Summary stats

- Mean: mean (variable)
- Variance: var (variable)
- Standard deviation: sd (variable)
- Skewness: skewness (variable), needs package ‘moments’
- Kurtosis: kurtosis (variable), needs package ‘moments’

R Code for Normal Distribution

- dnorm(data value, mu, sigma): gives the density i.e. the function returns the height of the normal distribution, at some value along the x-axis
- pnorm(data value, mu, sigma): gives the distribution function
- qnorm(quantile, mu, sigma): gives the Quantile function for calculating critical values
- rnorm(n,mu,sigma): generates ‘n’ samples from a Normal distribution with mean=mu and standard deviation=sigma

R Code for Normal Distribution (Contd.)

- qqnorm (variable): Without ‘extRemes’ package, it will create a plot but without bands and with ‘extRemes’ package, it will create a plot with bands.
- qqPlot (variable, distribution=”norm”):With ‘car’ package, creates a normal probability plot with bands
- qqline(variable,col="red") with or without extRemes package, it will draw a red line passing through the qqplot
- To generate a new variable as a linear combination of two normal variables plot a qqplot, use the following commands:

  X1<-rnorm(500,15,3)
  X2<-rnorm(500,25,5)

- Y<-3*X1+4*X2 library(car) qqPlot(Y)
- To draw a Histogram with a Normal curve superposed on it: hist(variable,prob=TRUE)
- curve(dnorm(x,mean(variable),sd(variable)),col="red",add=TRUE)

Sampling Types

Simple Random Sample (SRS)	A sample of size n such that every pair of unit in the population has the same chance of appearing in the sample. This implies that every possible sample of size n has the same chance of being the actual sample. This also implies that every individual unit has the same chance of appearing in the sample, but some other kinds of random samples also have this property
Systematic Random Sample	A random sample of size n drawn from a simple sampling frame, such that each of the first N/n (i.e., the inverse of the sampling fraction) units on the list has the same chance of being selected and every $(N/n)^{th}$ subsequent unit on the list is also selected. This implies that every unit, but not every subset of n units, in the population has the same chance of being in the sample
Multi-Stage Random Sample	A sample selected by random mechanisms in several stages, most likely because it is impossible or impractical to acquire a list of all units in the population, because no simple sampling frame is available.

Sampling Types (Contd..)

- Most population parameters and sample statistics that we consider are percentages. For example:
  - the percent of the population or sample who approve of the way the government is functioning
  - the percent of the population or sample who intend to vote a particular party in next election
- A sample statistic is unbiased if its expected value is equal to the corresponding population parameter.
  - This means that, as we take more and more samples from the same population, the average of all the sample statistics "converges" (comes closer and closer to) with the true population parameter.
- The variation in sample statistics from sample to sample is called sampling error
- (Random) Sampling Error is the magnitude of the inherent variability of sample statistics (from sample to sample)

Sampling Types (Contd.)

- most random sample statistics are
  - (approximately) normally distributed
    - with an average value equal to the corresponding population parameter, and
    - a variability (sampling error) that
      - is mainly a function of sample size n (as well as variability within the population sampled), and
      - can be calculated on the basis of the laws of probability.

- Sample mean varies from one sample to another
- Sample mean can be (and most likely) is different from the population mean
- Sample mean is a random variable

Next Class
Central Limit Theorem & Confidence Interval

Sl No.	Agenda Topics
1.	Central Limit Theorem
2.	R Code For Understanding Central Limit Theorem
3.	How does Central Limit Theorem work
4.	Mechanism
5.	Plotting Now
6.	CLT Facts
7.	Practical Application 1 Of CLT
8.	Practical Application 2 of CLT
9.	Confidence Interval & Probability
10.	(Mis)interpreting The Confidence Interval

Sl No.	Agenda Topics
11.	How To Interpret Confidence Interval
12.	P-value , Z-score
13.	Confidence Intervals For Unknown Mean And Known Standard Deviation
14.	t-Distribution
15.	t-Distribution Using R
16.	Poison Distribution
17.	Simulation Of Poisson Random Variables
18.	Question And Answer on Poison Distribution
19.	Exercises

Contact Info:

- **LinkedIn https://www.linkedin.com/in/gautham111/**