STATISTICS

Unit 2

1. Probability
2. Probability Distribution

2.1 Discrete probability distribution

2.1.1 Binomial distribution

2.1.2 Poisson distribution

2.2 Continuous probability distribution

2.2.1 Normal distribution

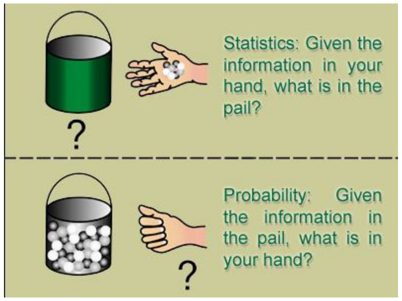
2.2.2 Student's t distribution

2.2.3 Chi-Square distribution

1 Probability

Definitions

- An **experiment** is a situation involving chance that leads to results called outcomes.
- An **outcome** is the result of a single trial of an experiment



- An **event** is one or more outcomes of an experiment.
- **Probability** is the measure of how likely an event is.

Formula for finding the probability

P(A) = Number of ways an event can occur / Total number of possible outcomes

Example: A balanced coin is tossed. The probability of Head coming up is given as P(1) = 1 / 2 = 0.5

1 Probability – continued

Definition

Conditional Probability:

- If E_1 and E_2 are two events, then the probability that E_2 occurs, if E_1 occurs is $P(E_2|E_1)$.
- If $P(E_2|E_1) = P(E_2)$, then E_1 and E_2 are independent events; otherwise they are dependent,
- For independent events, $P(E_1E_2) = P(E_1)P(E_2)$
- For dependent events, $P(E_1E_2) = P(E_1)\,P(E_2|E_1)$
- For three dependent events,

  $$\mathbf{P(E_1E_2E_2) = P(E_1)P(E_2|E_1)\,P(E_3|E_1E_2)}$$

Example: *A math teacher gave her class two tests. 25% of the class passed both tests and 42% of the class passed the first test. What percent of those who passed the first test also passed the second test?*

Solution: *P(Second| First) = 0.25 / 0.42 = 0.60 = 60%*

2 Probability Distribution

Types of Probability distributions

A Discrete Probability Distribution:

- If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.
- Suppose you flip a coin two times. There are four outcomes: HH,HT,TH,TT The random variable X can take on the values HH,HT,TH,TT, so it is a discrete random variable.

Some discrete probability distributions:

- Binomial probability distribution
- Poisson probability distribution

2 Probability Distribution – continued

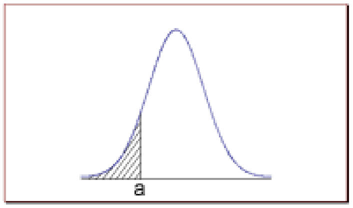
Types of Probability distributions – continued

B Continuous Probability Distribution:

- If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.
- Consider the probability density function shown below and assume we want to know the probability that the random variable X is less than or equal to a. This is a **cumulative** or **continuous** probability.

Some discrete probability distributions:

- Normal probability distribution
- Student's t probability distribution
- Chi-square probability distribution



2 Probability Distribution – continued

Types of Probability distributions – continued

B Continuous Probability Distribution – continued

- The probability of a value lying between a and b is the area under the curve between the coordinates a and b, expressed as a ratio to the total area under the curve.
- The probability of an exact value a is zero.
- Example: Even though a fast-food restaurant advertises a hamburger as weighing 200 grams, you observe that it may not be exactly 200 grams, some weighing less (say- 180 grams) or more (220 grams) than 200 grams.
- Suppose X denote the weight of a randomly selected 200-gram hamburger, ranging from 180 grams to 220 grams, it is a continuous random variable.

2 Probability Distribution – continued

2.1 Discrete probability distribution

2.1.1 Binomial distribution

- The binomial distribution arises naturally when events depend upon a fixed probability of occurrence p, when the number of trials is limited.
- If the probability of an event occurring at any trial is p, and n trials are made, the probability of exactly x successes in the n trials is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$ *binomial distribution and denoted as*

$X \sim b(n,p)$

The binomial approaches normal distribution, when n becomes large.

2 Probability Distribution – continued

2.1.1 Binomial distribution – continued

- The binomial distribution arises naturally when events depend upon a fixed probability of occurrence p, when the number of trials is limited.
- The binomial distribution provides a suitable model for very many statistical distributions which occur in nature, economics, business, psychological and educational testing etc.
- Main properties of the binomial distribution:

Mean	$\mu = np$
Variance	$\sigma^2 = npq$
Standard Deviation	$\sigma = \sqrt{npq}$
Moment coefficient of skewness	$\alpha_3 = (q - p) / \sigma$
Moment coefficient of kurtosis	$\alpha_4 = 3 + (1 - 6pq) / npq$

When $p = q$, *this distribution is symmetrical and* $\alpha_3 = 0$

2 Probability Distribution – continued

2.1.1 Binomial distribution – continued

Example:

Three dice are rolled. What is the probability of 0, 1, 2 and 3 sixes?

x	$x!$	$n!$	$(n-x)!$	$n!/x!(n-x)!$	$p^x q^{n-x}$	$f(x)$
0	1	6	6	1	$125/6^3$	$125/6^3$
1	1	6	2	3	$25/6^3$	$75/6^3$
2	2	6	1	3	$5/6^3$	$15/6^3$
3	6	6	1	1	$1/6^3$	$1/6^3$
						$216/6^3 = 1$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The probabilities of 0, 1,2, and 3 sixes are 125/ 216, 75/216, 15/216, and 1/216 respectively.

2 Probability Distribution – continued

2.1.2 Poisson distribution

- The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time or space.
- Let X be the number of events in a given interval.
- If the mean number of the events per interval is λ.
- The probability of observing x events in a given interval is given as

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, 3, 4, .$$

Here e is approximately equal to 2.718282

- $X \sim Po(\lambda)$

2 Probability Distribution – continued

2.1.2 Poisson distribution – continued

Properties

Mean	$\mu = \lambda$
Standard Deviation	$\sigma = \sqrt{\lambda}$
Moment coefficient of skewness	$\alpha_3 = 1/\sqrt{\lambda}$
Moment coefficient of kurtosis	$\alpha_4 = 3 + (1/\lambda)$

- *The Poisson distribution is the limit of binomial distribution as $n \Rrightarrow \propto$ and $p \Rrightarrow 0$*

Applications

- Car accidents
- Number of deaths by horse kicking in the Prussian army (first application)
- Birth defects and genetic mutations

2 Probability Distribution – continued

2.1.2 Poisson distribution – continued

Example

A life insurance salesman sells on the average 3 life insurance policies in a week. Use Poisson law to calculate the probability that in a given week he will sell some policies.

Here, $\mu = 3$

$P(X > 0) = 1 - P(x_0)$

$P(x_0) = e^{-3}\, 3^0 / 0! = 4.9787 \times 10^{-2} = 0.04979$

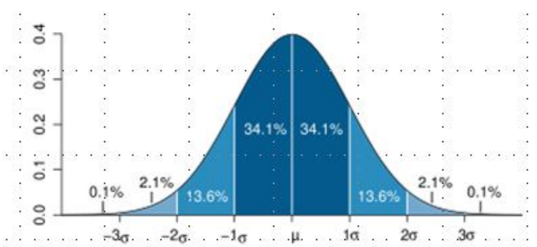
$P(X > 0) = 1 - 0.04979 =$ ***0.95021***

2 Probability Distribution – continued

2.2 Continuous Probability Distribution

2.2.1 Normal distribution

- The normal distribution or bell curve is most familiar distribution and has unique characteristics:
- It is always symmetrical with equal areas on both sides of the curve.
- The highest point on the curve corresponds to the mean score, which equals the median and the mode of the distribution.
- *The area between given standard deviation units includes a determined percentage area.*
- *Total area under the curve $\pm\ 1\sigma$ (68.26%); $\pm\ 2\sigma$ (95.44%); $\pm\ 3\sigma$ (99.74%);*



2 Probability Distribution – continued

2.2.2 Normal distribution – continued

Properties

Mean $= \mu$

Standard Deviation $= \sigma$

Moment coefficient of skewness, $\alpha_3 = 0$

Moment coefficient of kurtosis, $\alpha_4 = 3$

- *The Normal distribution is the limit of binomial distribution when n is large.*

Applications

- *Biological measurements of all sorts*
- *Number of events accumulated in time*
  - *a. Amount of rainfall per interval*
  - *b. Number of stock orders per (longer) interval*

2 Probability Distribution – continued

2.2.1 Normal distribution – continued

Example

The amount of cash filled in a ATM machine by a bank each day is normally distributed with mean Rs. 10 lakhs and standard deviation Rs. 3.5 lakhs. If they keep Rs. 15 lakhs on hand, what is the probability that the ATM will run out of money for the customers.

Solution

Let X denote the demand.

$P(X > 15L) =$

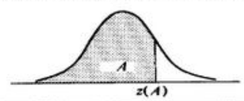
$P[(X - 10L) / 3.5L > (15L - 10L) / 3.5L]$

$= P[Z > 1.43]$

$= 1 - P[Z <= 1.43] = 1 - 0.9236$

$= 0.0764$

Entry is area A under the standard normal curve from $-\infty$ to $z(A)$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	$\boxed{.9236}$	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

2 Probability Distribution – continued

2.2.2 Student's t distribution

- This is covered in the later chapters.

2.2.3 Chi-square distribution

- The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- It is one of the most widely used probability distributions in inferential statistics, e.g., in hypothesis testing or in construction of confidence intervals.

2 Probability Distribution – continued

2.2.3 Chi-square distribution – continued

- The chi-squared distribution is used in

  1. the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one,
  2. the independence of two criteria of classification of qualitative data, and in
  3. the confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.

The probability density function (pdf) of the chi-squared distribution is

$$f(x;\,k) = \begin{cases} \frac{x^{(k/2-1)} e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

where $\Gamma(k/2)$ denotes the Gamma function, which has closed-form values for integer k.

2 Probability Distribution – continued

2.2.3 Chi-square distribution – continued

Properties

Mean of the distribution = the degrees of freedom, $\mu = v$

Variance = twice the degrees of freedom $\sigma^2 = v^2$

When the degrees of freedom ≥ 2*, the maximum value for the function occurs when* $\chi^2 = v - 2$

As the degrees of freedom increases, the chi-square curve approaches the normal distribution.

Applications

1. *Chi-square test – goodness of fit – A number of marketing problems involve decision situations in which it is important for a marketing manager to know whether the pattern of frequencies that are observed fit well with the expected ones.*

2 Probability Distribution – continued

2.2.3 Chi-square distribution – continued

Applications

2. *Chi-square test of independence – If there are two categorical varaibles (example: brand preference and income level), and our interest is to examine whether these two variables are associated with each other.*
3. *ANOVA problems via its role in the F distribution, which is the distribution of the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom.*
4. *Tests for equality of population proportions (three or more)*

2 Probability Distribution – continued

2.2.3 Chi-square distribution – continued

Example

In consumer marketing, it is always a challenge to select the appropriate design for packages. Assume that a marketing manager wishes to compare four different designs for packages and identify the most preferred design. A random sample of 500 consumers reveals the following:

Package design	*Preferred by consumers*
D1	*75*
D2	*144*
D3	*156*
D4	*125*

Do consumer preferences for package design show any significant difference?

2 Probability Distribution – continued

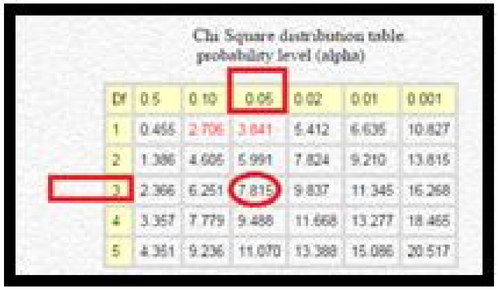
2.2.3 Chi-square distribution – continued

Solution

Null hypothesis: H_0 *All designs are equally preferred.*
Alternative hypothesis: H_1 *All designs are not equally preferred.*

$$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$$

Package Design	Observed frequencies (O)	Expected frequencies (E)	$(O-E)^2$	$(O-E)^2 / E$
D1	75	125	2500	20
D2	144	125	361	2.88
D3	156	125	961	7.69
D4	125	125	0	0
			Σ	30.57


Chi Square distribution table
probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

We reject the null hypothesis since
χ^2 calculated value is 30.57 > χ^2 tabulated value of 7.815