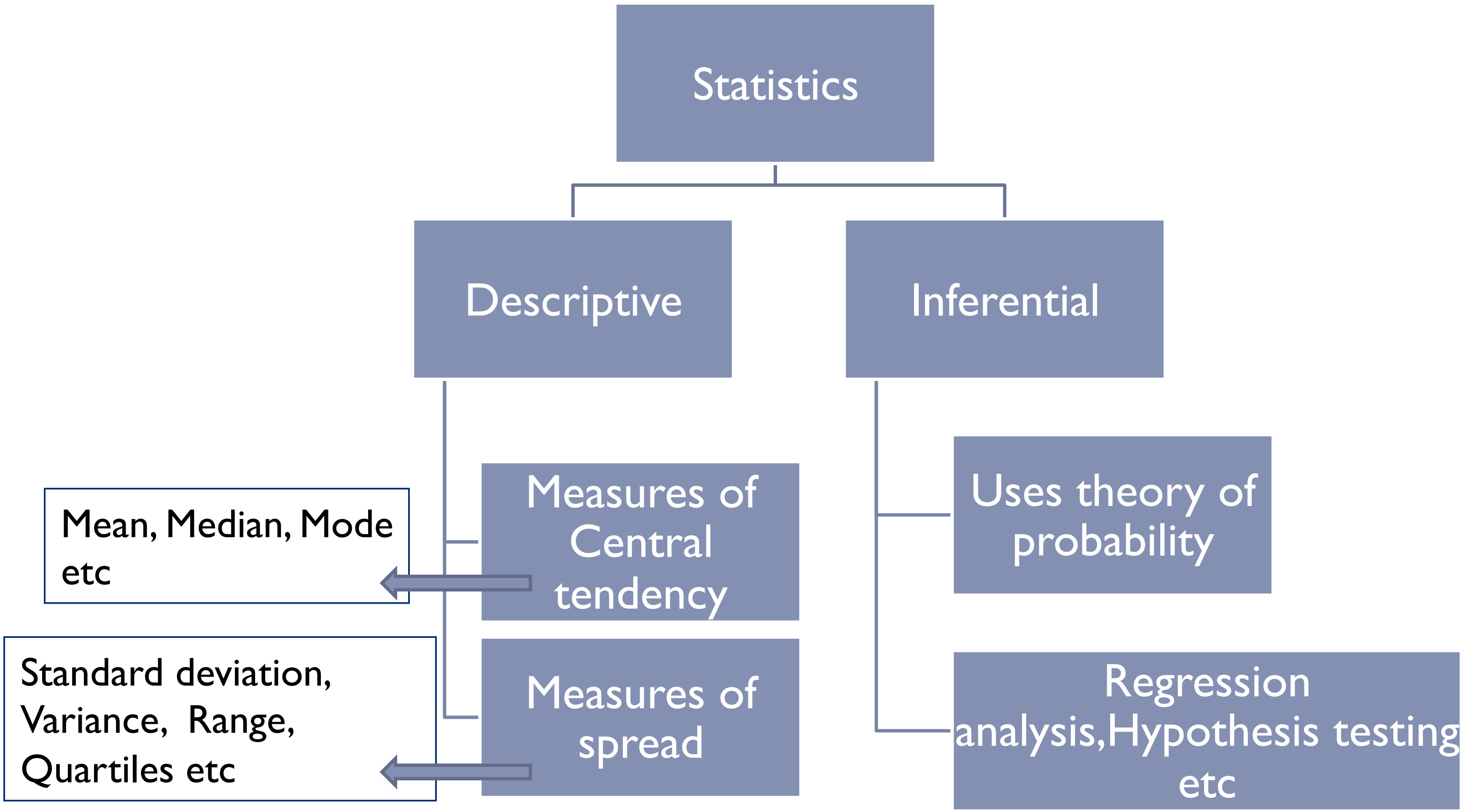
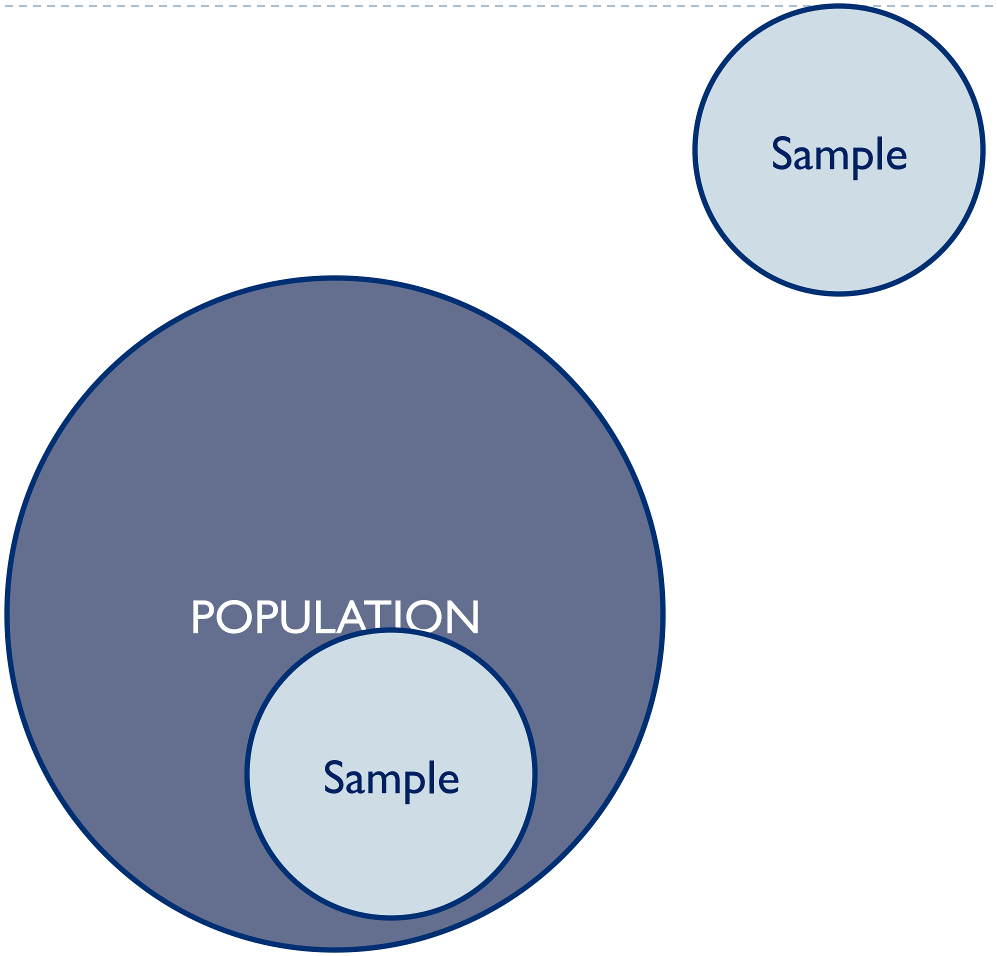
Introduction to statistics and probability

Prepared by Ranju Mohan, PhD student, CE dept, IITM
For CE302 class by Gitakrishnan Ramadurai, AP, CE dept, IITM

Statistics:


Statistics
Descriptive
Inferential
Measures of Central tendency
Measures of spread
Uses theory of probability
Regression analysis,Hypothesis testing etc
Mean, Median, Mode etc
Standard deviation, Variance, Range, Quartiles etc

Working with data…..



+ Descriptive statistics
(Measures of spread & Measures of dispersion)

Statistical estimate
+
Inferential statistics

Concepts of probability

INFERENCE about the population

Descriptive statistics

Speed data collected for 10 vehicles:

Speed (km/hr.)
27.8
29.5
32.4
32.4
32.4
36.9
43.5
49.8
52.3
58.9

$$\text{Mean} = \frac{32.4 \times 3 + 27.8 + 43.5 + 52.3 + 36.9 + 29.5 + 49.8 + 58.9}{10}$$

$$= 39.59$$

Arranging in increasing order,

$$\text{Median} = \frac{32.4 + 36.9}{2} = 34.65$$

Mode = 32.4

Measures of Central Tendency

Descriptive statistics – *Measures of spread*

Speed (km/hr.)
27.8
29.5
32.4
32.4
32.4
36.9
43.5
49.8
52.3
58.9

Range = Max value –Min value = 58.9- 27.8 = 31.1

Absolute deviation: |value - mean|

Ex: Absolute deviation = |52.3- 39.59|=12.71

Arranging in increasing order,

Quartiles	1st (25% of data)	2nd (25% of data)	3rd (25% of data)	4th (25% of data)
Value	[1(10)+2]/4th = 32.4 (**Q1**)	[2(10)+2]/4th =34.65 (**Q2**)	[3(10)+2]/4th =49.8 (**Q3**)	-

Interquartile range = Q3-Q1 = 17.4

Descriptive statistics - *Measures of spread*

$(32.4 - 39.59)^2$
$(27.8 - 39.59)^2$
$(43.5 - 39.59)^2$
$(52.3 - 39.59)^2$
$(36.9 - 39.59)^2$
$(29.5 - 39.59)^2$
$(32.4 - 39.59)^2$
$(49.8 - 39.59)^2$
$(58.9 - 39.59)^2$
$(32.4 - 39.59)^2$
∑ = 684.21

Variance :

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = 684.21/9 = 76.02$$

Standard deviation :

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} = \sqrt{684.21/9} = \sqrt{76.02} = 8.72$$

Data:

- Groups of information that represents the qualitative or quantitative attributes of a variable or a set of variables.

- Visual/Graphical Representation:
  - Frequency distributions
  - Graphs
  - Box plot
  - Scatter plot
  - Stem and leaf

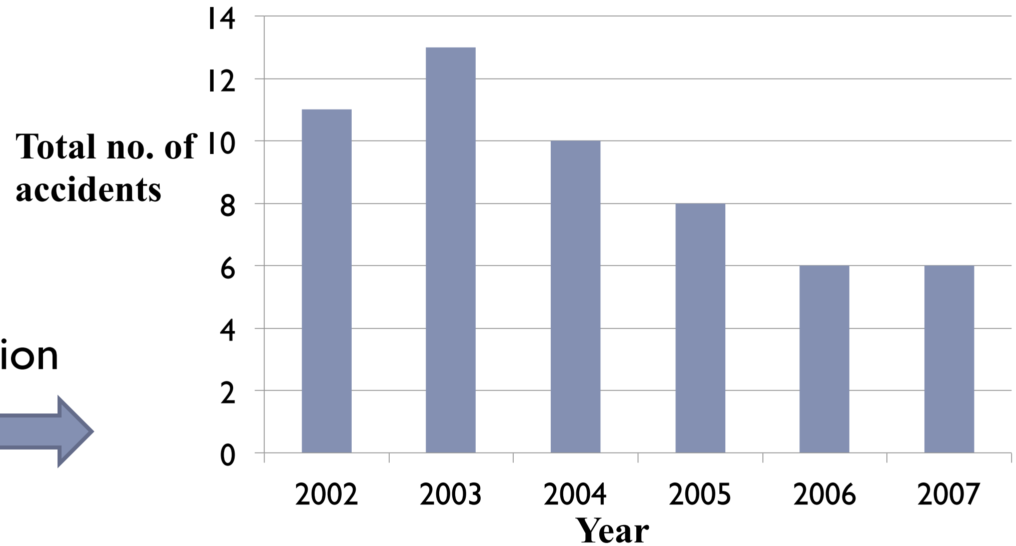
Data representation: *Examples*

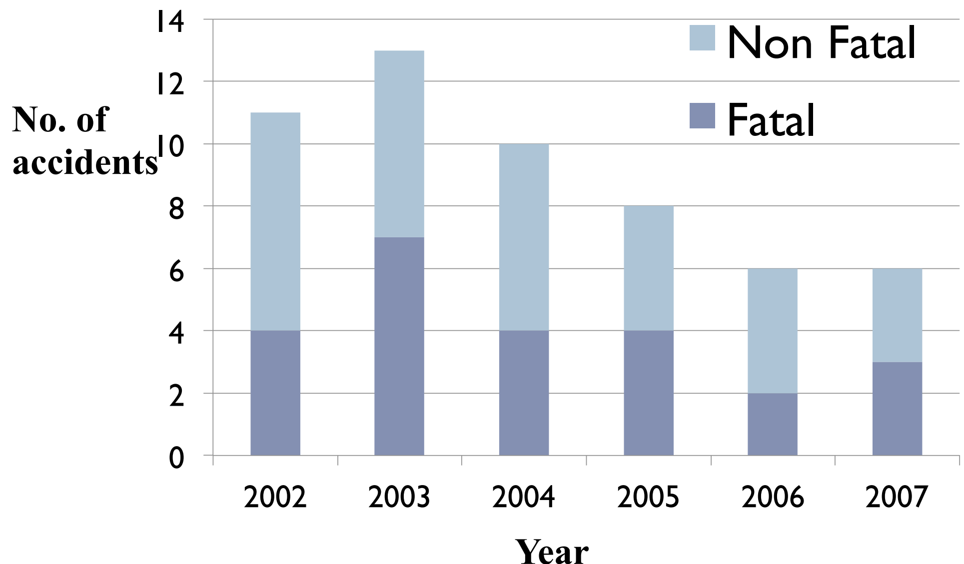
Year	No. of Accidents		
	Fatal	Non-fatal	Total
2002	4	7	11
2003	7	6	13
2004	4	6	10
2005	4	4	8
2006	2	4	6
2007	3	3	6

Frequency distribution table

Graphical representation

Bar chart

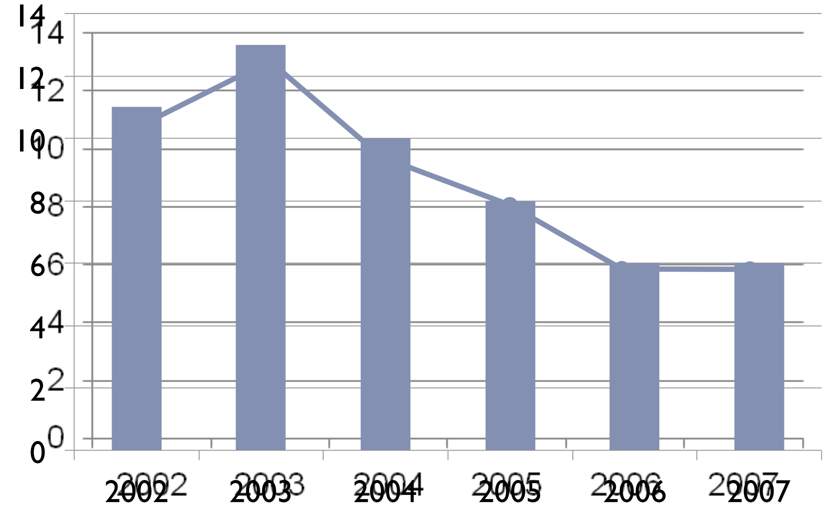




Data representation: *Examples*

Previous data ,

Year	No. of Accidents
2002	11
2003	13
2004	10
2005	8
2006	6
2007	6



Frequency polygon

Data representation: *Examples*

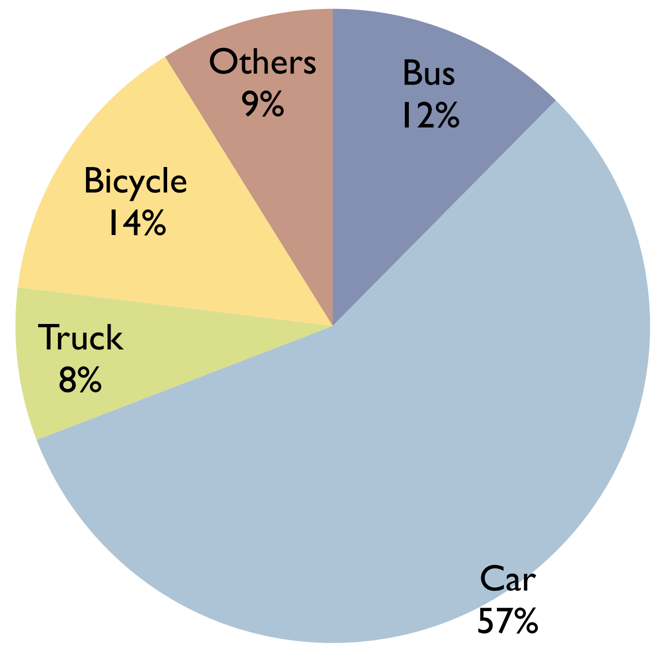
Daily volume of vehicles observed :

Vehicle type	frequency
Bus	35
Car/Jeep	160
Truck	22
Bicycle	40
Others	25
TOTAL	282

Vehicle composition

Pie Diagram

No. of vehicles (%)



Data representation: *Examples*

Given speed data:

63.2	49.9	36.9	44.2	54.8	49	42.9	32.4
54.3	37.5	45.1	51.7	47.5	43.8	55.9	48.8
41.1	47.5	52.3	39.2	57.3	36.3	42.8	58.7
52.9	42.5	46.4	53.3	46.5	43.2	56.9	47.7
47.8	35.6	50.3	44.7	46.2	38.4	62.4	39.4
56.4	55.1	64.8	52.8				

3	2.4	5.6	6.3	6.9	7.5	8.4	9.2	9.4											
4	1.1	2.5	2.8	2.9	3.2	3.8	4.2	4.7	5.1	6.2	6.4	6.5	7.5	7.5	7.7	7.8	8.8	9	9.9
5	0.3	1.7	2.3	2.8	2.9	3.3	4.3	4.8	5.1	5.9	6.4	6.9	7.3	8.7					
6	2.4	3.2	4.8																

57.3

Stem and Leaf plot

Data representation: *Examples*

Given speed data (km/hr.),

63.2	49.9	36.9	44.2	54.8	49	42.9	32.4
54.3	37.5	45.1	51.7	47.5	43.8	55.9	48.8
41.1	47.5	52.3	39.2	57.3	36.3	42.8	58.7
52.9	42.5	46.4	53.3	46.5	43.2	56.9	47.7
47.8	35.6	50.3	44.7	46.2	38.4	62.4	49.4
56.4	55.1	64.8	52.8				

Group into different speed class

Speed class	No. of vehicles
30-35	1
35-40	6
40-45	8
45-50	12
50-55	8
55-60	6
60-65	3

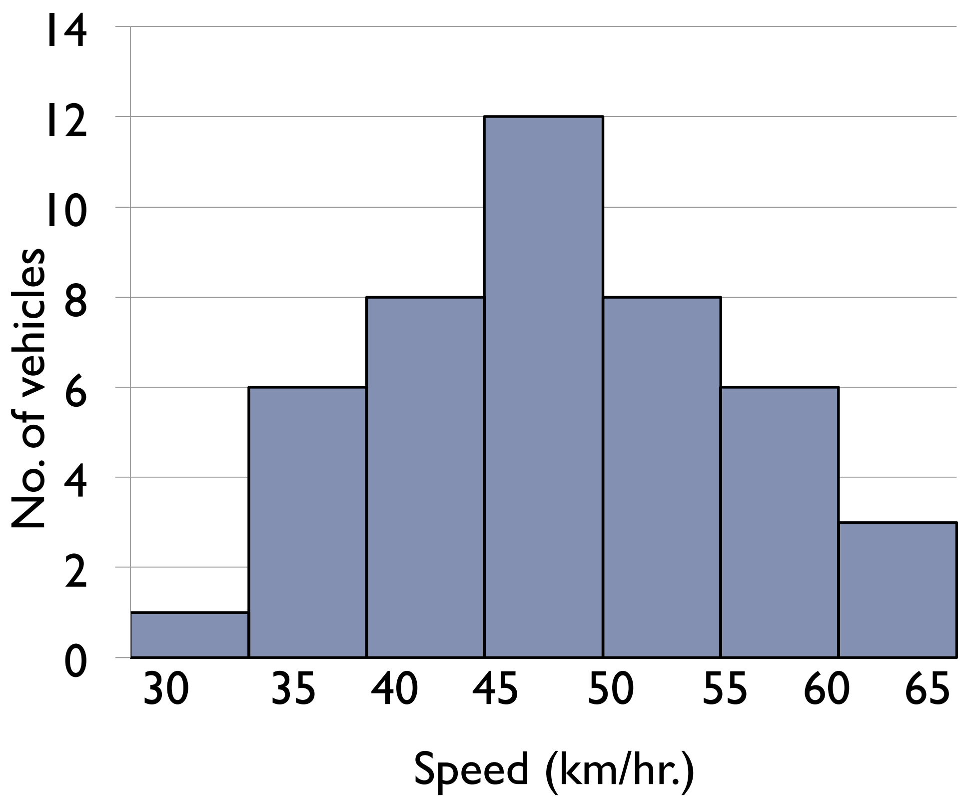
$$Class\ Interval = \frac{\max value - \min value}{1 + 3.22 \log\left(No.\ of\ veh\right)}$$

$$= \frac{64.8 - 32.4}{1 + 3.22 \log(48)}$$

$$= 5.05,\ say\ 5$$

Data representation: *Examples*

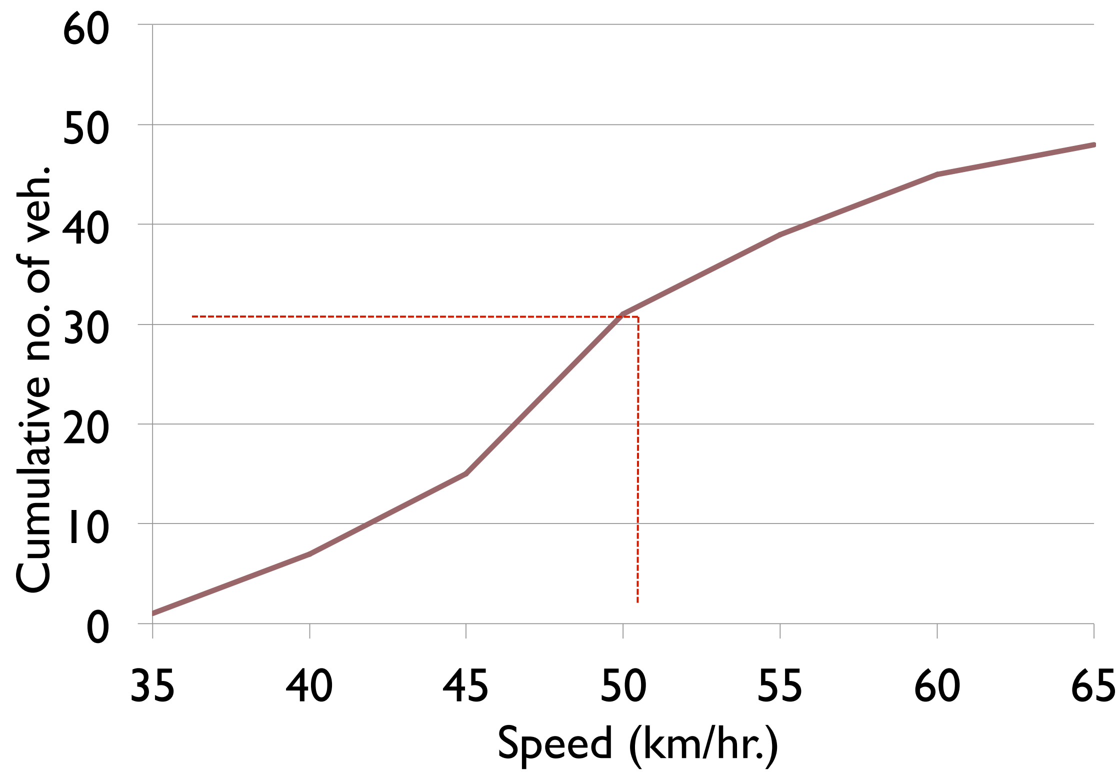
Speed class	No. of vehicles
30-35	1
35-40	6
40-45	8
45-50	12
50-55	8
55-60	6
60-65	3



Histogram

Data representation: *Examples*

Speed class	No. of vehicles	Cumulative no. of veh.
30-35	1	1
35-40	6	7
40-45	8	15
45-50	12	27
50-55	8	35
55-60	6	41
60-65	3	44



Ogive

Q: Number of vehicles with speed less than 50 km/hr. ?

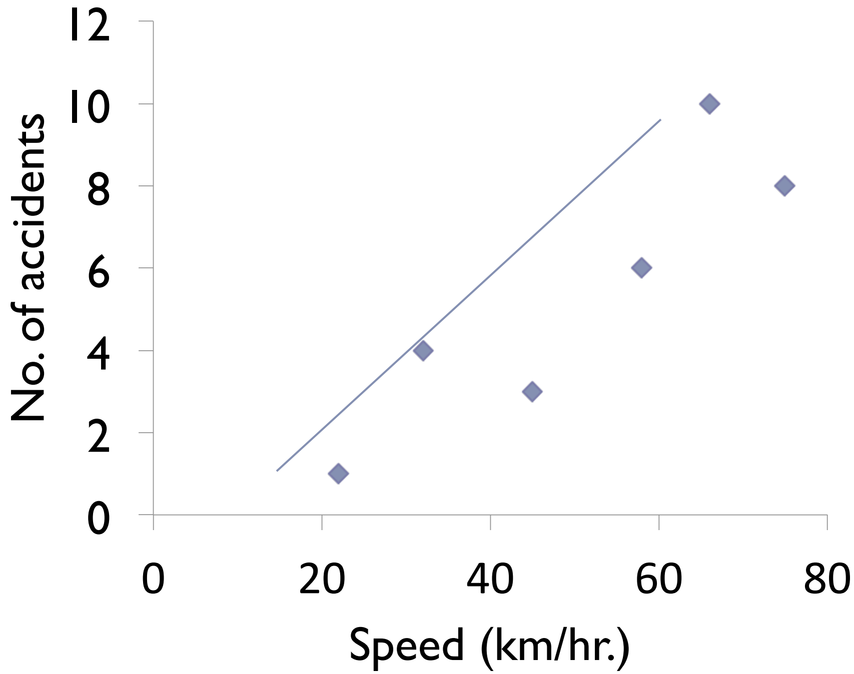
Ans: 31

Data representation: *Examples*

Paired data set:
(Number of accidents ,Vehicle speed)

Speed (km/ hr.)	No. of accidents
22	1
45	3
32	4
75	8
66	10
58	6

How to relate?



Scatter plot

Data representation: *Examples*

When having several simultaneous comparison: *11 observations in 3 diff. days. (n=11)*

→ ***Box Plot***

	Speed (km/hr)		
	Mon	Wed	Fri
L	25	22	24
	28	29	29
Q1	33	30	31
	38	33	33
	39	40	36
Q2	42	42	38
	46	55	40
	55	60	41
Q3	59	64	45
	60	70	50
H	65	72	58

Points to be plotted	**Speed (km/hr.)**		
	Mon	Wed	Fri
Lowest (L)	25	28	28
Q1 = [1(n)+2]/4th value = 3rd value	33	30	31
Q2 = median	42	42	38
Q3= [3(n)+2]/4th value =9th value	59	64	45
Highest (H)	65	72	58

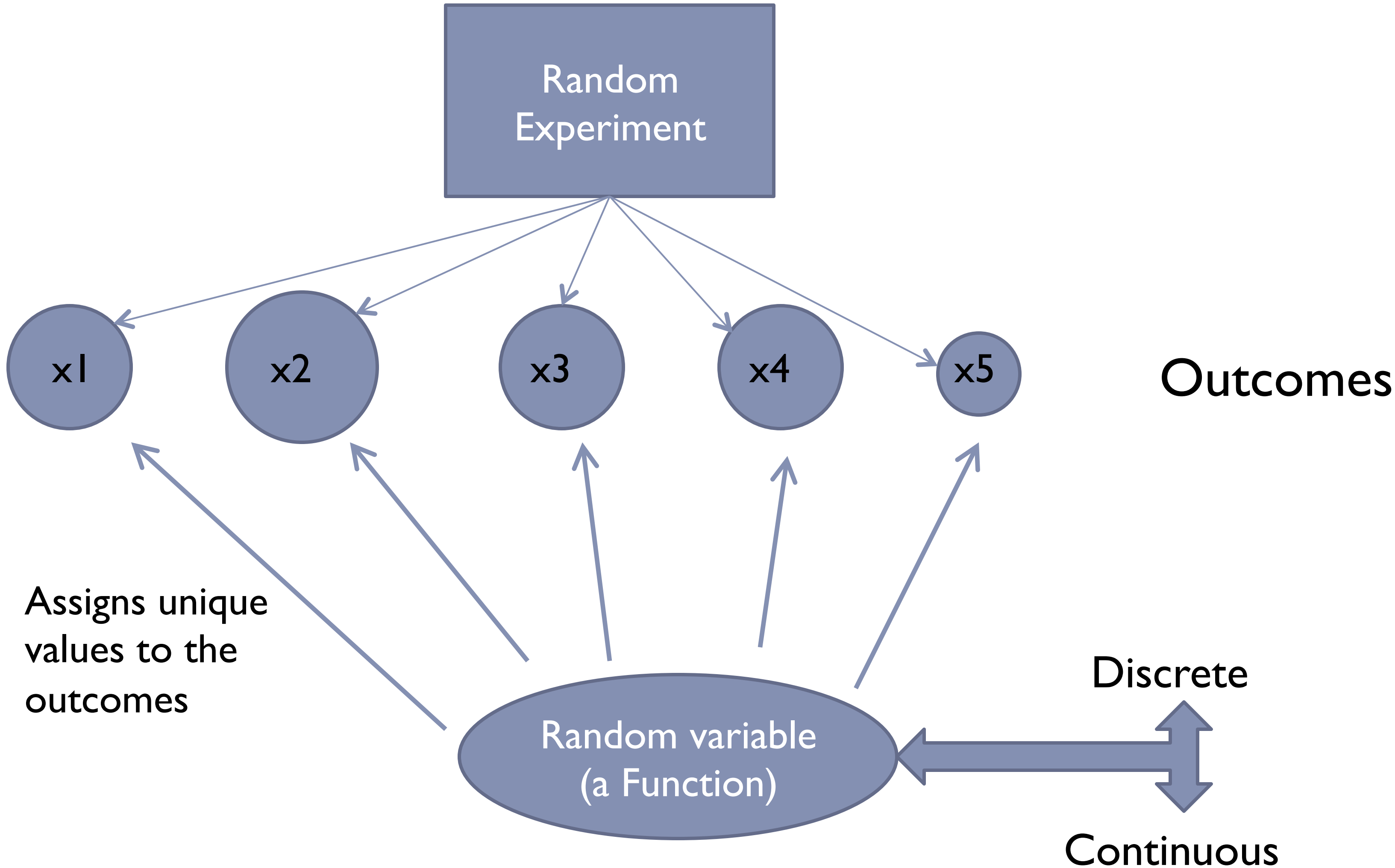
Inferential statistics- basic concepts

- Use sample statistics and probability concepts to make inferences about the population

- Probability (P): The likelihood of something happening or being true.

- Based on the assumption that sampling is random

Inferential statistics : Probability concepts-Random variables



Probability concepts – Random variables

Discrete Random variable	Continuous Random variable
Probability mass function, pmf = p(x)	Probability density function, pdf = f(x)
$p(x) = P(X{=}x)$ $0 \leq p(x) \leq 1$ $\sum p(x) = 1$	$\int_a^b f(x)dx = P(a < x < b)$ $f(x) > 1$ $\int f(x) = 1$

Cumulative distribution function $F(X \leq x)$ Examples:

x	P(X=x)
1	0.13
2	0.27
3	0.25
4	0.15
5	0.20

F(3) =0.13+0.27+0.25

= 0.65

$$\int f(x)dx = \begin{cases} 2x; x \geq 0 \\ 0; x < 0 \end{cases}$$

$$F(3) = \int_{-\infty}^{0} 0dx + \int_{0}^{3} 2xdx = 9$$

Probability concepts – Random variables

Discrete Random variable	Continuous Random variable
Given x_i' s and $p(x_i)$' s	Given x_i' s and $f(x_i)$' s
Expectation of a random variable, E(X):	Weighted average of the possible values
$E(X) = \sum_i x_i p(x_i)$ $E(X^2) = \sum_i x_i^2 p(x_i)$	$E(X) = \int x f(x)$ $E(X^2) = \int x^2 f(x)$
Mean = E(X) = First moment about origin	
Variance =V(X) Second moment about mean	$V(X) = E(x-\mu)^2$ or $V(X) = E(x^2) - E(x)^2$

Some common probability distributions used in traffic engineering

Discrete data	Continuous data
Bernoulli distribution Binomial distribution Multinomial distribution Poisson distribution	Exponential distribution Normal distribution & distribution arising from normal Chi-square distribution t- distribution F – distribution

Special Random variables and probability distributions

Discrete Random variables

Bernoulli :

Two possible outcomes for one trial:
'success' (X=1) or 'failure' (X=0)

$$\text{pmf} = \begin{cases} P(X=0) = 1-p \\ P(X=1) = p \end{cases} \qquad 0 \le p \le 1$$

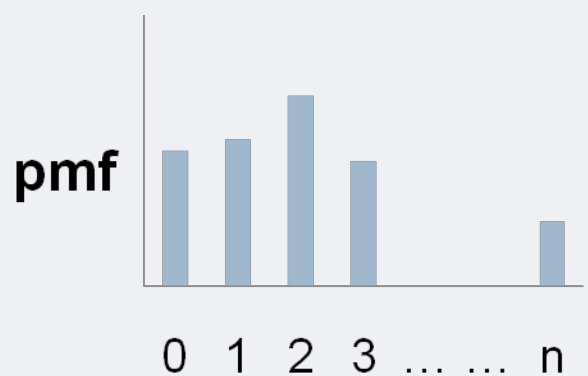
Mean = p; Variance = p(1-p)

Binomial :

'n' independent trials, each having two outcomes

$$\text{pmf} = p(X=x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \; ; x = 0, 1, \dots, n$$

Mean = np; Variance = np(1-p)



Special Random variables and probability distributions

Discrete Random variables

Poisson:

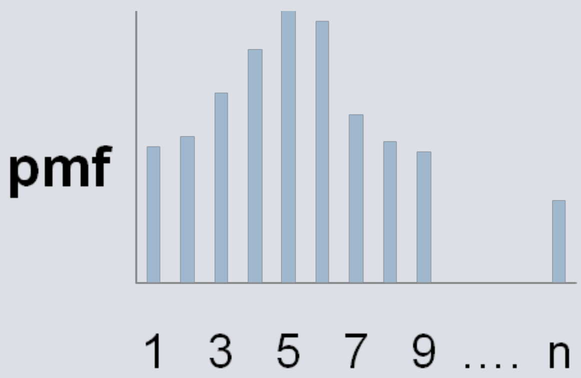
When 'n' is large and p is small

$$\text{pmf} = P(X = x) = \frac{e^{-\nu}\nu^{x}}{x!}\ ; \ i = 0, 1, \ldots.$$

ν = mean number of successes = np

x = actual number of successes

Mean = ν; Variance = ν

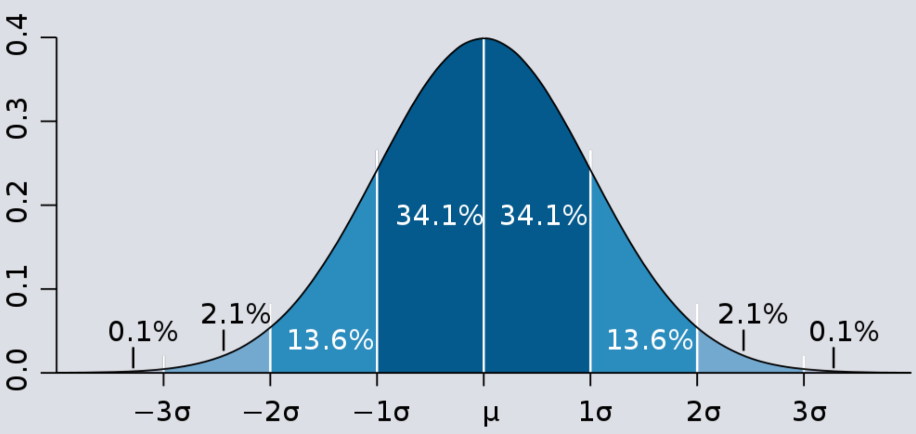


Continuous Random variables

Normal:

$$\text{pdf} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)2}{2\sigma^2}}\ ; -\infty < x < \infty$$

Mean = μ ; Variance = σ^2



Special Random variables and probability distributions

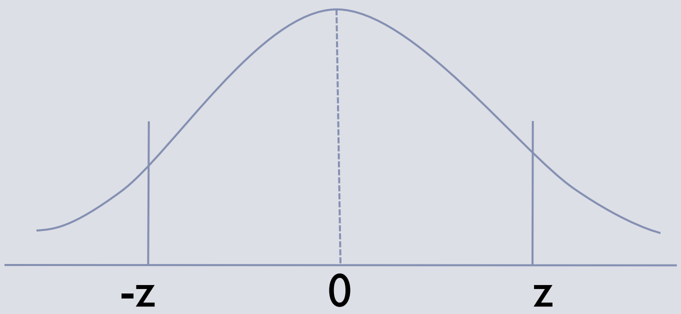
Continuous Random variables

Normal random variable, z

$$z = \frac{x - \mu}{\sigma}$$

When μ=0 and σ =1;

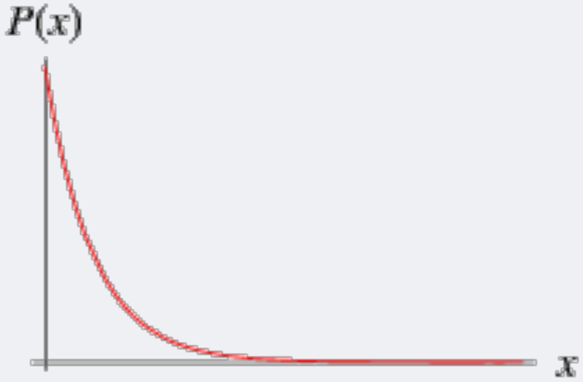
$$\text{pdf} = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad ; -\infty < x < \infty$$



Exponential:

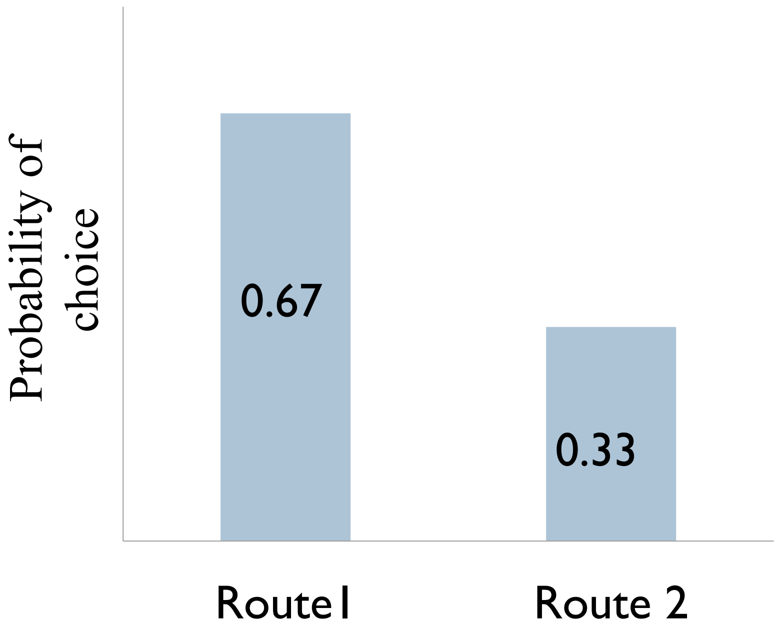
$$P(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Mean = $1/\lambda$;Variance = $1/\lambda^2$



Examples

- *Que: On a particular junction, out of two routes to a particular destination, probability of choosing 1st route is twice as that of the 2nd route. How many number of vehicles will turn to Route 1 when a total of 5 vehicles reach at the junction at a specified time?*



Bernoulli distribution

$$p(\text{route 1}) = p^x (1-p)^{1-x}$$

$x = 0$ with probability 0.33

$x = 1$ with probability 0.67

$$\left.\begin{array}{l} p(\text{route 1}) = 0.33^0 (1-0.33)^1 \\ \text{or} \\ p(\text{route 1}) = 0.67^1 (1-0.67)^0 \end{array}\right\} = 0.67$$

Ans: Number of vehicles choosing Route 1 = 0.67×5 = 3.3, say 3

Examples

- *Que: Probability of choosing a particular route is 1/5. Find out the probabilities that out of 5 vehicles reaching that location, exactly 0, 1, 2, 3, 4, 5 vehicles will choose that particular route.*

With Binomial distribution,

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$n = 5$; $p = 1/5$

Ans:

x	P(x)
0	0.33
1	0.41
2	0.20
3	.05
4	.006
5	.0003

Examples

- *Ques: For 3 different routes at a particular location, probability of choice is given by 0.35, 0.40, 0.25 respectively. What is the probability that out of 5 vehicles reaching at the location, one, three and two vehicles will choose the route 1, 2 and 3 respectively.*

By multinomial distribution,

$$p(x_1, x_2, \ldots . x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots . p_k^{x_k}$$

$$p(1, 3, 2) = \frac{5!}{1!\ 3!\ 2!} 0.35^1 \times 0.4^3 \times 0.25^2$$

Ans: 0.014

Examples

- *On a motorway, the number of vehicles arriving from one direction in successive 10 sec intervals was counted and is given below. Find out the probabilities P(0), P(> 3), P(3< X< 6) etc.*

By Poisson distribution,

$$p(x) = \frac{e^{-\nu}\nu^{x}}{x!}$$

ν = (200/1000)*10 = 2

Ans:

P(0) = 0.135

P(X>3) = 0.144

P(3< X< 6) =0.127

No. of veh. in 10 sec (i)	Frequency (ii)
0	11
1	28
2	30
3	18
4	8
5	4
6	1
7	0

Total no. of veh. (i*ii)	Total time (ii*10)
0	110
28	280
60	300
54	180
32	80
20	40
6	10
0	0
∑ = 200	∑ =1000

Example:

- *Ques: If an average of 3 trucks arrive per hour to be unloaded at a warehouse, what are the probability that the time between the arrivals of successive trucks will be (i) less than 5 min., (ii) at least 45 min.*

Using exponential distribution,

$$P(x) = \lambda e^{-\lambda x}$$

$$P(X < x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$$

$\lambda = 3/60 = 0.05$ veh/min.

Ans:

$P(X<5) = 1 - e^{-0.05(5)} = 0.2212$

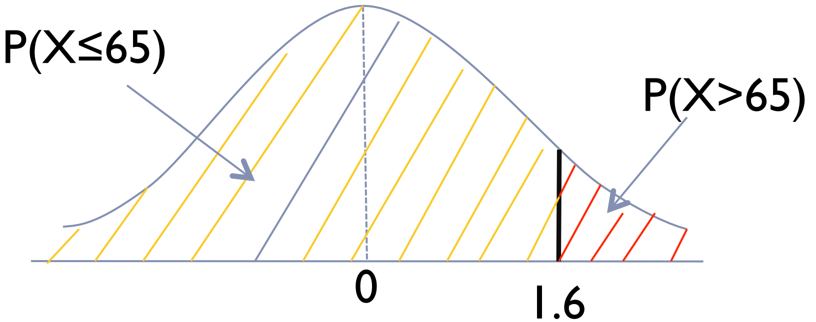
$P(X \geq 45) = e^{-0.05(45)} = 0.105$

Example:

- *The spot speed at a particular location are normally distributed with a mean of 51.7 km/hr. and std. deviation of 8.3 km/hr. what is (ii)the probability that speed exceeds 65 km/hr. (ii) the 85th percentile speed.*

$$z = \frac{x - \mu}{\sigma}$$

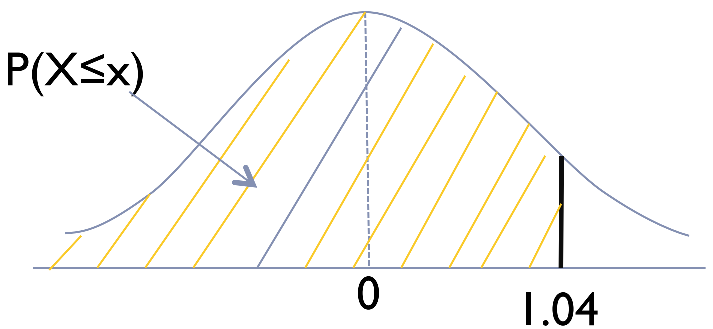
(i) z = (65-51.7)/8.3 =1.6



From the standard normal distribution table,
F(1.6)= P(X≤65) = 0.9452

Ans: P(X>65) = 0.0548

(ii) P(X≤x) = 0.85 = F(z)= 0.85



From the standard normal distribution table,
z= 1.04
x = 1.04(8.3)+51.7 = 60.33

Ans: x =60.33 km/hr.

Standard Normal distribution table

- Shows the cumulative probability associated with a particular z- score

z	0.00	0.01	…	…..	0.07	0.08	0.09
-3.0	0.0013	0.0013	…	…..	0.0012	0.0010	0.0010
….	…	…	…	…	…	…	….
-1.3	0.0968	0.0951	….	…	0.0853	0.0838	0.0823
….	….	…..	…..	….	…	…	….
3.0	0.9987	0.9987	…..	….	0.9989	0.9990	0.9990

Example: $P(z<-1.31) = 0.0951$

Inferential statistics: Sampling distributions

- ## Sampling Theory:

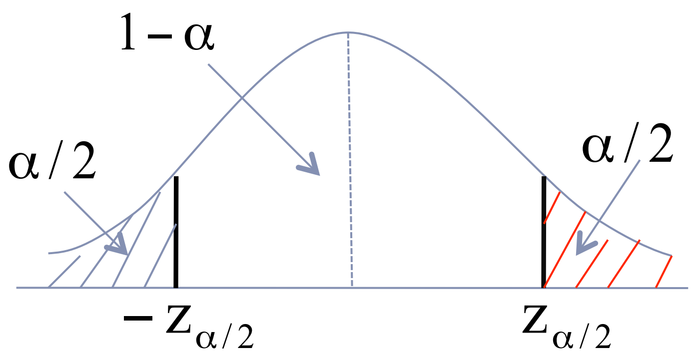
If a random sample of size n is taken from a population of mean μ and variance σ^2, *then* the sample mean $\overline{X}$ follows a normal distribution with mean μ and variance σ^2/n.

The standard error of mean is given by $\dfrac{\sigma}{\sqrt{\text{n}}}$.

- ## Central limit theorem:

If $\bar{x}$ is the mean of a sample of size n taken from a population of mean μ and variance σ^2, then the variate $z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$ *approaches a normal distribution as* $n \longrightarrow \infty$

Central limit theorem-Error estimate



$$-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

$$\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}$$

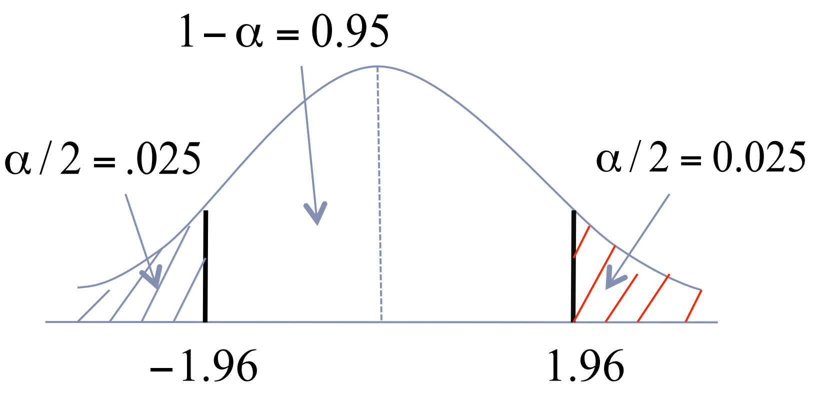
$$E = \overline{X} - \mu = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$, where 'α' is th level of significance

- *Level of significance: the probability that the computed estimate will lie outside the indicated range .Here the range is the confidence level,* $1-\alpha$

Example

- *While determining the mean speed of veh. on a section of a road, engineer wants to be able to assert with 95% confidence that the mean speed is off by 2.5 km/hr. If std. deviation is 8.2 km/hr., how large the sample is?*

$$E = \overline{X} - \mu = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$



$1-\alpha = 0.95; \quad \alpha = 0.05; \quad \alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

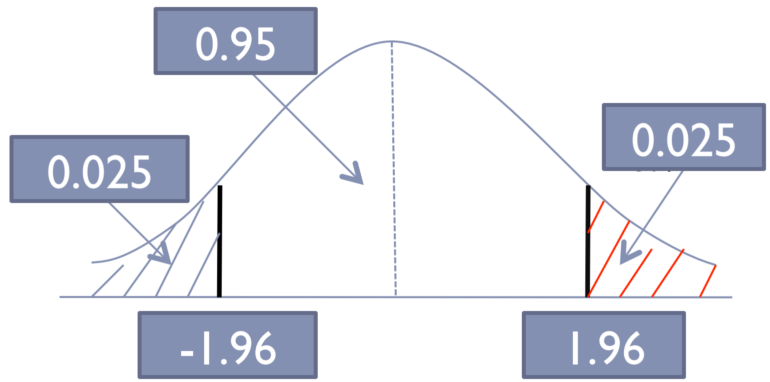
$$2.5 = \frac{1.96 \times 8.2}{\sqrt{n}}$$

Sample size, n = 41

Central Limit theorem

- Confidence interval (C.I.) for the population mean μ

$$\text{C.I.} = \left(\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right)$$



- *Example:*

A random sample of size 100 is taken from a population with std. deviation 5.1, given that the sample mean is 21.6, construct a 95% confidence interval.

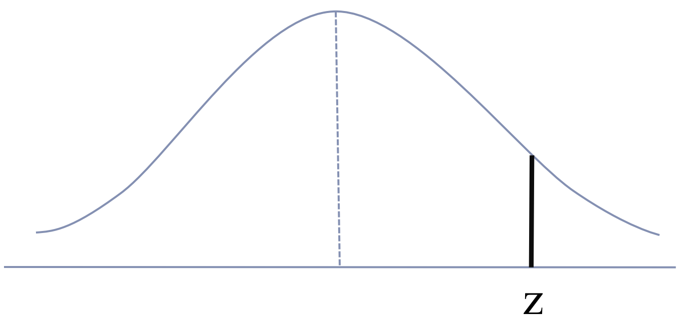
$$\text{C.I.} = \left(21.6 - \frac{1.96 \times 5.1}{\sqrt{100}},\ 21.6 + \frac{1.96 \times 5.1}{\sqrt{100}} \right)$$

Ans: (21.5, 22.6)

Distributions from Normal distribution

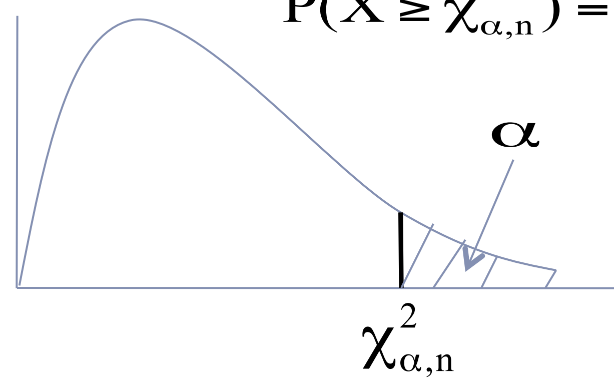
$z_1, z_2, \ldots\ldots, z_n$ → Independent standard normal random variables

$$\chi_n^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$



Standard normal distribution

$$P(X \geq \chi^2_{\alpha,n}) = \alpha$$

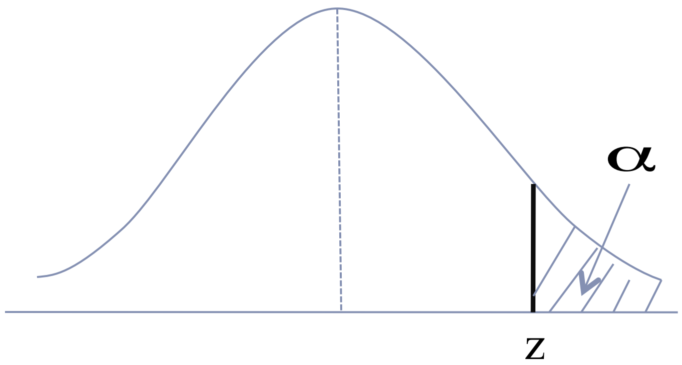


Chi-square distribution with 'n' degrees of freedom

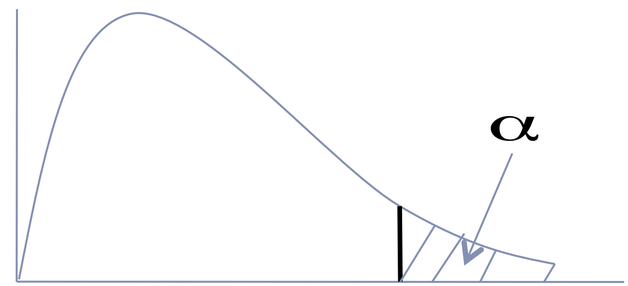
$$\text{pdf} = \frac{\frac{1}{2} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{\left(\frac{n}{2}-1\right)!}, x > 0$$

Distributions from Normal distribution

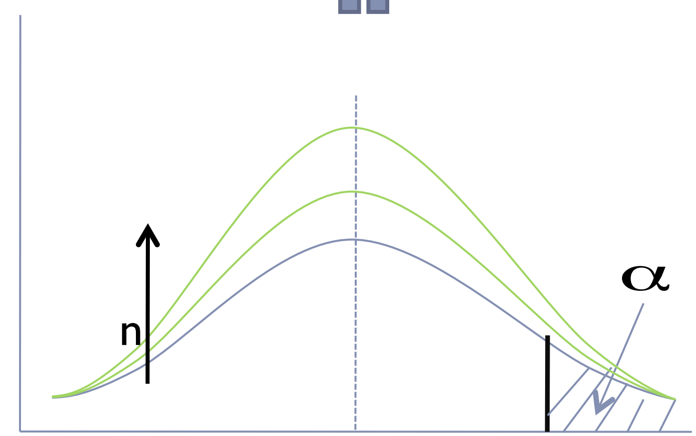
z – Random variable with standard normal distribution



$\chi^2_{\alpha,n}$ – Random variable with Chi-square distribution



$\chi^2_{\alpha,n}$



$t_{\alpha,n}$

t- distribution with ‘n’ degrees of freedom

$$P(t_n \geq t_{\alpha,n}) = \alpha$$

$$t_n = \frac{z}{\sqrt{\chi^2_n / n}}$$

As n becomes large, $\chi^2_n = 1 \rightarrow t_n \approx z$

Distributions from Normal distribution

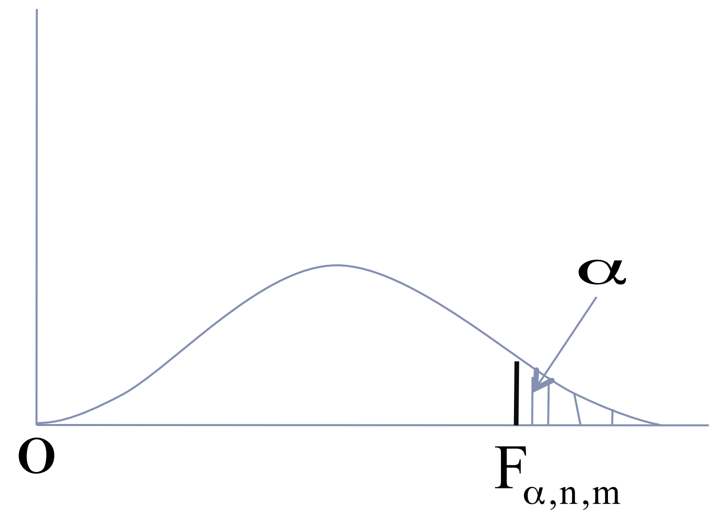
For independent chi-square random variables χ_n^2 and χ_m^2

,

$$F_{n,m} = \frac{\chi_n^2 / n}{\chi_m^2 / m}$$

$$P(F_{n,m} > F_{\alpha,n,m}) = \alpha$$

$$\frac{1}{F_{\alpha,n,m}} = F_{1-\alpha,m,n}$$



F- distribution with degrees of freedom ‘n’ and ‘m’

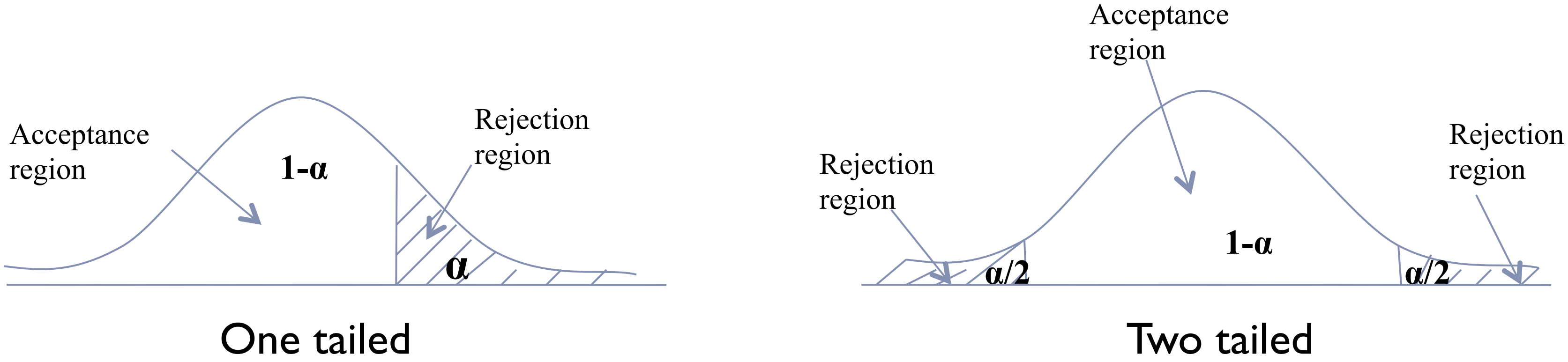
How to use these sampling distributions to draw conclusion?

- Hypothesis testing
  - Concerned with two distinct choices:
    - Null Hypothesis (H_0)
    - Alternate hypothesis (H_1)
  - Test whether to accept or reject H_0 using various test statistics.
  - Two types of errors:

Two possibilities	Decision	
	Accept H_0	Reject H_0
H_0 True	Correct !	Type I error
H_1 True	Type II error	Correct !

Testing the hypothesis

- One tail or two tail?



- Confidence level: 1-α : probability that the computed estimate will lie in the acceptance region
- Level of significance: α :probability that the computed estimate will lie in the rejection region

Distribution statistics in hypothesis testing

- *Que.No.1: The spot speed at a particular location in an expressway are known to be normally distributed with a mean of 80km/hr. and std. dev. of 15km/hr. A new radar speed meter was bought by traffic dept. and a set of 100 observations were taken. The mean speed observed was 77.3km/hr. Is there any evidence to prove that :*

  (i) the new speed meter might have been faulty

  (ii) the new speed meter is showing lesser speed than actual. Assume 5% level of significance.

Solution

Solution to Ques.No.1(i)

Here we have to test:

H_0: The speedometer is not faulty ($\bar{x}$ μ=80km/hr.)

against

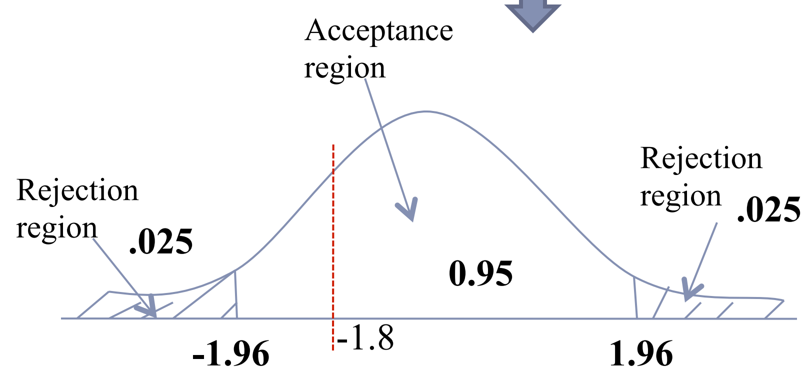
H_1: The speedometer is faulty ($\neq$ $\bar{x}$ 80km/hr. i.e either >80 or <80)

Two -tailed

Given α = 5%

n=100, large sample

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{77.3-80}{15/\sqrt{100}} = -1.8$$



Accept H_0

Inference: The speedometer is not faulty

Solution to Ques.No.1(ii)

Here we have to test:

H_0: $\mu = 80$km/hr.

against

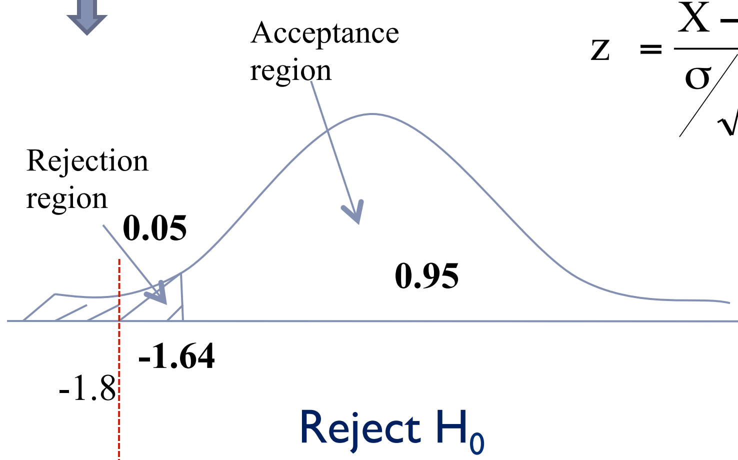
H_1: $\mu < 80$

One tailed

Given $\alpha = 5\%$

n=100, large sample

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{77.3 - 80}{15 / \sqrt{100}} = -1.8$$



Reject H_0

Inference: The new speedometer is showing lesser speed than actual

BACK

Distribution statistics in hypothesis testing

- *Que. No. 2: The mean spot speed of 15 vehicles observed on a Sunday at a particular roadway was 81.2km/hr. The mean speeds of all vehicles at this location as per previous records was 75.5 km/hr. and std. dev. 10.2km/hr. Is there sufficient evidence to show that the speeds of vehicles on that Sunday was higher than the average speed? Take level of significance as 5%*

Solution

Solution to Ques.No.2

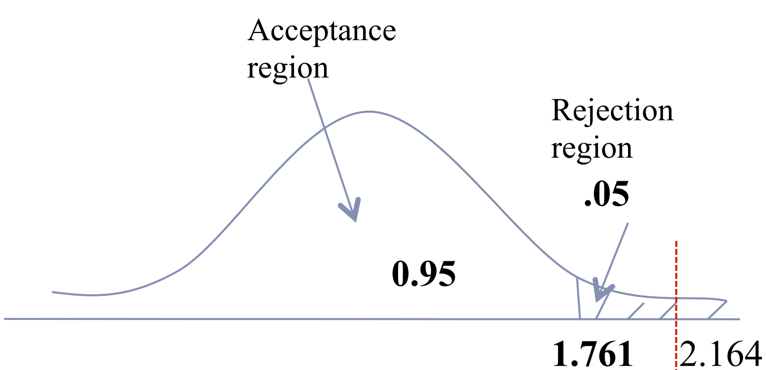
Here we have to test:

H_0: $\mu = 75.5$km/hr.

against

H_1: $\mu > 75.5$km/hr.

One tailed



Reject H_0

Given $\alpha = 5\%$

n=15, small sample

Also sample std. dev. is given, hence use t-statistics

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{75.5 - 81.2}{10.2/\sqrt{15}} = 2.164$$

Inference: The speeds of vehicles on that Sunday is higher than the average speed

BACK

Distribution statistics in hypothesis testing

- *Ques. No.3: Two samples of speed data are collected are as follows:*

  For sample 1, mean speed is 74.3km/hr. and std. dev. is 7km/hr. (n_1=120)

  For sample 2, mean speed is 72.5km/hr. and std. dev. is 8km/hr. (n_2=120)

  Is there any evidence to prove that the mean speed reduced by more than 0.5km/hr. when using these samples? Assume level of significance as 10%.

Solution

Solution to Ques.No.3

Two samples and hence concerned with two means μ_1 andμ_2

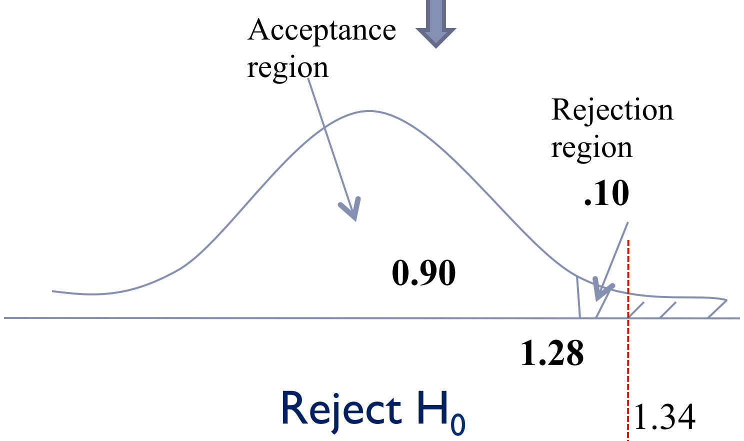
Have to test:

H_0: $\mu_1 - \mu_2 = 0.5$km/hr.

against

H_1: $\mu_1 - \mu_2 > 0.5$km/hr.

One tailed



Reject H_0

Given α = 5%

$n_1 = n_2 = 50$, large sample

For test concerning two means, z-statistics is given by,

$$z = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{(74.3 - 72.5) - (0.5)}{\sqrt{\frac{7^2}{120} + \frac{8^2}{120}}} = 1.34$$

Inference: the mean speed reduced by more than 0.5km/hr.

BACK

Distribution statistics in hypothesis testing

- *Que.No.4:For a given vehicle speed data sample of size 20, the standard deviation observed was 12.5km/hr. The data can be used only if the standard deviation is near to approximately equal to10km/hr. Check whether the data can be accepted at 5% level of significance.*

Solution

Solution to Ques.No.4

Problem is related to the sampling distribution of variance

Have to test:

H_0: $\sigma = 10$km/hr.

against

H_1: $\sigma > 10$km/hr.

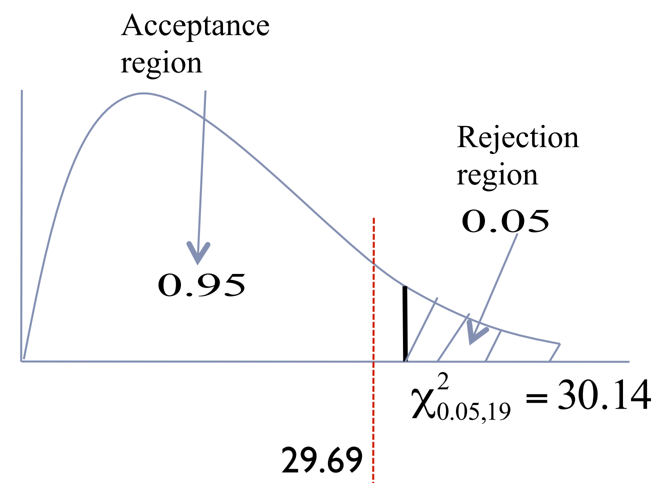
Given α = 5%

Degrees of freedom = sample size-1 =19

χ^2 statistics for variance is :

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$= \frac{(20-1)12.5^2}{10^2} = 29.69$$



Accept H_0

Inference: The given speed data can be accepted

BACK

Distribution statistics in hypothesis testing

- *Que.No.5:It is desired to determine whether there is less variability in the speed data collected for day 1 than for day2. If independent random samples are taken for these two days as below:*

  For day 1: std. dev.=12km/hr. ;sample size=12

  For day 2:std. dev.=10km/hr. ;sample size=14,

  test the given hypothesis with a level of significance 5%.

Solution

Solution to Ques.No.5

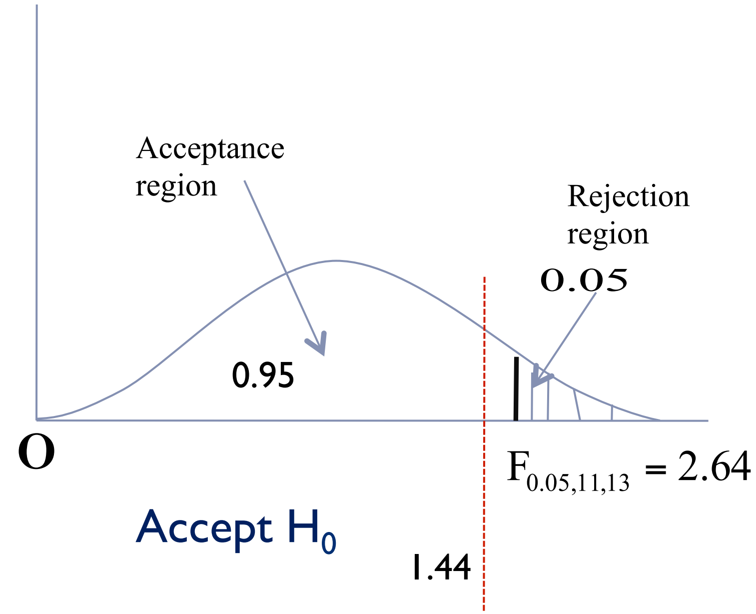
Here the question concerned with the comparison of variance.

Have to test:

H_0: $\sigma_1^2 = \sigma_2^2$

against

H_1: $\sigma_1^2 < \sigma_2^2$



Given α = 5%

Statistics that can be used: F-statistics

Degrees of freedom = sample size-1.

Here, 11 and 13

For comparing sample variance, $F = (s_1^2/n_1) / (s_2^2/n_2)$

Where s_1 and s_2 are the sample standard deviations.

$$F = (12^2/12)/(10^2/14) = 1.68$$

Inference: There is no variability in the speed data measured for day 1 and 2

BACK

Distribution statistics – Hypothesis testing

- *Que.No.6: Every minute vehicle count data was collected for a period of 65 minutes. Determine at 95% confidence level , whether the data follows a poisson distribution.*

No. of arrival	Observed frequency
0	2
1	6
2	7
3	12
4	13
5	9
6	9
7	4
8	2
9	1

To test the fit of data to a particular distribution,

'GOODNESS OF FIT' test

Solution

Solution to Que.No.6

H_0: Data follows poisson distribution
H_1: Data not follows poisson distribution

O_i: Observed frequency
E_i: Expected frequency

Poisson probability:

$$p(x) = \frac{e^{-\nu}\nu^{x}}{x!}$$

ν = mean number of arrival = 260/65 =4

$$p(x) = \frac{e^{-4}4^{x}}{x!}$$

Arrival (x_i)	Obsv. freq (min)	Total no. of veh.	Prob. $p(x_i)$	E_i (prob.*65)
0	2	0	0.018	1.17
1	6	6	.0733	4.76
2	7	14	0.1465	9.52
3	12	36	0.1954	12.7
4	13	52	0.1954	12.7
5	9	45	0.1563	10.16
6	9	54	0.1042	6.77
7	4	28	0.0595	3.87
8	2	16	0.0298	1.94
9	1	9	0.0132	0.858
	∑ =65	∑ =260		

Goodness of fit – solution to Que.No.6

At least 5 groups and at least 5 nos. in each group

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$= 2.31$

Degrees of freedom
$= N-1-g = 5$

g - *no. of statistics used to calculate E_i ; here only ν*

$$\chi^2_{0.05,5} = 11.07 > 2.31$$

Accept H_0

No. of arrival	Observed frequency (mintute), O_i			Expected frequency (E_i)		$(O_i-E_i)^2/E_i$
0	2	8	1	1.17	5.93	0.7189
1	6			4.76		
2	7		2	9.52		0.6671
3	12		3	12.7		0.0386
4	13		4	12.7		0.007
5	9		5	10.16		0.132
6	9		6	6.77		0.7345
7	4			3.87	6.67	
8	2	7	7	1.94		0.0165
9	1			0.858		

N=7

$\sum = 2.31$

Inference: The given data follows poisson distribution

BACK

Summary of test statistics for Hypothesis testing

TEST STATISTICS **Hint: μ_0 = population mean** **σ_0 = population std. dev.**	**H_1**	**Reject H_0 if**
Large sample – concerning mean $H_0 : \mu = \mu_0$		
$z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$	$\mu < \mu_0$	$z < -z_\alpha$
	$\mu > \mu_0$	$z > z_\alpha$
	$\mu \neq \mu_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
Small sample – concerning mean $H_0 : \mu = \mu_0$		
$t = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$	$\mu < \mu_0$	$t < -t_\alpha$
	$\mu > \mu_0$	$t > t_\alpha$
	$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Summary of test statistics for Hypothesis testing

TEST STATISTICS **Hint: μ_0 = population mean** **σ_0 = population std. dev.**	**H_1**	**Reject H_0 if**
Comparison of sample mean $H_0 : \mu_1 - \mu_2 = \delta$		
$z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^{\,2}}{n_1} + \dfrac{\sigma_2^{\,2}}{n_2}}}$	$\mu_1 - \mu_2 < \delta$	$z < -z_\alpha$
	$\mu_1 - \mu_2 > \delta$	$z > z_\alpha$
	$\mu_1 - \mu_2 \neq \delta$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
One variance $H_0 : \sigma^2 = \sigma_0^{\,2}$		
$\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$	$\sigma^2 > \sigma_0^{\,2}$	$\chi^2 > \chi^2_\alpha$
	$\sigma^2 < \sigma_0^{\,2}$	$\chi^2 > \chi^2_{1-\alpha}$
	$\sigma^2 \neq \sigma_0^{\,2}$	$\chi^2 < \chi^2_{1-\alpha/2}$ or $\chi^2 > \chi^2_{\alpha/2}$

Summary of test statistics for Hypothesis testing

TEST STATISTICS **Hint: μ_0 = population mean** **σ_0 = population std. dev.**		**H_1**	**Reject H_0 if**
Two variance $H_0 : \sigma_1^2 = \sigma_2^2$			
F	$(s_1^2 / n_1)/(s_2^2 / n_2)$	$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha, n_1 - 1, n_2 - 1}$
	$(s_2^2 / n_2)/(s_1^2 / n_2)$	$\sigma_1^2 < \sigma_2^2$	$F > F_{\alpha, n_2 - 1, n_1 - 1}$
	$(s_{large}^2 / n_L)/(s_{small}^2 / n_S)$	$\sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha, n_{large} - 1, n_{small} - 1}$
Underlying distribution H_0 : Data follows given distribution			
$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$		Data not follows given distribution	$\chi^2 > \chi_\alpha^2$

Thank You