

# **STATISTICAL TABLES**

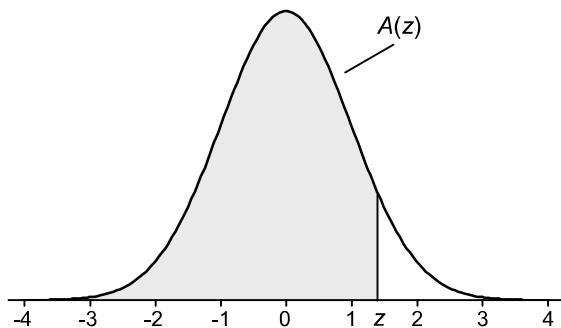
**Cumulative normal distribution**

**Critical values of the  $t$  distribution**

**Critical values of the  $F$  distribution**

**Critical values of the chi-squared distribution**

**TABLE A.1**  
**Cumulative Standardized Normal Distribution**



$A(z)$  is the integral of the standardized normal distribution from  $-\infty$  to  $z$  (in other words, the area under the curve to the left of  $z$ ). It gives the probability of a normal random variable not being more than  $z$  standard deviations above its mean. Values of  $z$  of particular importance:

$z$	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9998	0.9999						

**TABLE A.2**  
***t* Distribution: Critical Values of *t***

<i>Degrees of freedom</i>	<i>Two-tailed test:</i> <i>One-tailed test:</i>	<i>Significance level</i>					
		10%	5%	2%	1%	0.2%	0.1%
5%	2.5%	1%	0.5%	0.1%	0.05%		
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
$\infty$		1.645	1.960	2.326	2.576	3.090	3.291

TABLE A.3

**F Distribution: Critical Values of F (5% significance level)**

$v_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
$v_2$															
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
600	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
750	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58

**TABLE A.3 (continued)*****F* Distribution: Critical Values of *F* (5% significance level)**

<i>v</i> <sub>1</sub>	25	30	35	40	50	60	75	100	150	200
<i>v</i> <sub>2</sub>										
<b>1</b>	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68
<b>2</b>	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49
<b>3</b>	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54
<b>4</b>	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65
<b>5</b>	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39
<b>6</b>	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69
<b>7</b>	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25
<b>8</b>	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95
<b>9</b>	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73
<b>10</b>	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56
<b>11</b>	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43
<b>12</b>	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32
<b>13</b>	2.41	2.38	2.36	2.34	2.31	2.30	2.28	2.26	2.24	2.23
<b>14</b>	2.34	2.31	2.28	2.27	2.24	2.22	2.21	2.19	2.17	2.16
<b>15</b>	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.10	2.10
<b>16</b>	2.23	2.19	2.17	2.15	2.12	2.11	2.09	2.07	2.05	2.04
<b>17</b>	2.18	2.15	2.12	2.10	2.08	2.06	2.04	2.02	2.00	1.99
<b>18</b>	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.98	1.96	1.95
<b>19</b>	2.11	2.07	2.05	2.03	2.00	1.98	1.96	1.94	1.92	1.91
<b>20</b>	2.07	2.04	2.01	1.99	1.97	1.95	1.93	1.91	1.89	1.88
<b>21</b>	2.05	2.01	1.98	1.96	1.94	1.92	1.90	1.88	1.86	1.84
<b>22</b>	2.02	1.98	1.96	1.94	1.91	1.89	1.87	1.85	1.83	1.82
<b>23</b>	2.00	1.96	1.93	1.91	1.88	1.86	1.84	1.82	1.80	1.79
<b>24</b>	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.80	1.78	1.77
<b>25</b>	1.96	1.92	1.89	1.87	1.84	1.82	1.80	1.78	1.76	1.75
<b>26</b>	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76	1.74	1.73
<b>27</b>	1.92	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72	1.71
<b>28</b>	1.91	1.87	1.84	1.82	1.79	1.77	1.75	1.73	1.70	1.69
<b>29</b>	1.89	1.85	1.83	1.81	1.77	1.75	1.73	1.71	1.69	1.67
<b>30</b>	1.88	1.84	1.81	1.79	1.76	1.74	1.72	1.70	1.67	1.66
<b>35</b>	1.82	1.79	1.76	1.74	1.70	1.68	1.66	1.63	1.61	1.60
<b>40</b>	1.78	1.74	1.72	1.69	1.66	1.64	1.61	1.59	1.56	1.55
<b>50</b>	1.73	1.69	1.66	1.63	1.60	1.58	1.55	1.52	1.50	1.48
<b>60</b>	1.69	1.65	1.62	1.59	1.56	1.53	1.51	1.48	1.45	1.44
<b>70</b>	1.66	1.62	1.59	1.57	1.53	1.50	1.48	1.45	1.42	1.40
<b>80</b>	1.64	1.60	1.57	1.54	1.51	1.48	1.45	1.43	1.39	1.38
<b>90</b>	1.63	1.59	1.55	1.53	1.49	1.46	1.44	1.41	1.38	1.36
<b>100</b>	1.62	1.57	1.54	1.52	1.48	1.45	1.42	1.39	1.36	1.34
<b>120</b>	1.60	1.55	1.52	1.50	1.46	1.43	1.40	1.37	1.33	1.32
<b>150</b>	1.58	1.54	1.50	1.48	1.44	1.41	1.38	1.34	1.31	1.29
<b>200</b>	1.56	1.52	1.48	1.46	1.41	1.39	1.35	1.32	1.28	1.26
<b>250</b>	1.55	1.50	1.47	1.44	1.40	1.37	1.34	1.31	1.27	1.25
<b>300</b>	1.54	1.50	1.46	1.43	1.39	1.36	1.33	1.30	1.26	1.23
<b>400</b>	1.53	1.49	1.45	1.42	1.38	1.35	1.32	1.28	1.24	1.22
<b>500</b>	1.53	1.48	1.45	1.42	1.38	1.35	1.31	1.28	1.23	1.21
<b>600</b>	1.52	1.48	1.44	1.41	1.37	1.34	1.31	1.27	1.23	1.20
<b>750</b>	1.52	1.47	1.44	1.41	1.37	1.34	1.30	1.26	1.22	1.20
<b>1000</b>	1.52	1.47	1.43	1.41	1.36	1.33	1.30	1.26	1.22	1.19

TABLE A.3 (continued)

F Distribution: Critical Values of F (1% significance level)

$v_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
$v_2$															
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6142.67	6170.10	6191.53	6208.73
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.75	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.08	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.77	9.68	9.61	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.64	2.56	2.50	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.35	2.27	2.20	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.29	2.21	2.14	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.23	2.15	2.09	2.03
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97
250	6.74	4.69	3.86	3.40	3.09	2.87	2.71	2.58	2.48	2.39	2.26	2.15	2.07	2.01	1.95
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92
600	6.68	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.11	2.03	1.96	1.91
750	6.67	4.63	3.81	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.21	2.11	2.02	1.96	1.90
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90

**TABLE A.3 (continued)*****F* Distribution: Critical Values of *F* (1% significance level)**

<i>v</i> <sub>1</sub>	25	30	35	40	50	60	75	100	150	200
<i>v</i> <sub>2</sub>										
<b>1</b>	6239.83	6260.65	6275.57	6286.78	6302.52	6313.03	6323.56	6334.11	6344.68	6349.97
<b>2</b>	99.46	99.47	99.47	99.47	99.48	99.48	99.49	99.49	99.49	99.49
<b>3</b>	26.58	26.50	26.45	26.41	26.35	26.32	26.28	26.24	26.20	26.18
<b>4</b>	13.91	13.84	13.79	13.75	13.69	13.65	13.61	13.58	13.54	13.52
<b>5</b>	9.45	9.38	9.33	9.29	9.24	9.20	9.17	9.13	9.09	9.08
<b>6</b>	7.30	7.23	7.18	7.14	7.09	7.06	7.02	6.99	6.95	6.93
<b>7</b>	6.06	5.99	5.94	5.91	5.86	5.82	5.79	5.75	5.72	5.70
<b>8</b>	5.26	5.20	5.15	5.12	5.07	5.03	5.00	4.96	4.93	4.91
<b>9</b>	4.71	4.65	4.60	4.57	4.52	4.48	4.45	4.41	4.38	4.36
<b>10</b>	4.31	4.25	4.20	4.17	4.12	4.08	4.05	4.01	3.98	3.96
<b>11</b>	4.01	3.94	3.89	3.86	3.81	3.78	3.74	3.71	3.67	3.66
<b>12</b>	3.76	3.70	3.65	3.62	3.57	3.54	3.50	3.47	3.43	3.41
<b>13</b>	3.57	3.51	3.46	3.43	3.38	3.34	3.31	3.27	3.24	3.22
<b>14</b>	3.41	3.35	3.30	3.27	3.22	3.18	3.15	3.11	3.08	3.06
<b>15</b>	3.28	3.21	3.17	3.13	3.08	3.05	3.01	2.98	2.94	2.92
<b>16</b>	3.16	3.10	3.05	3.02	2.97	2.93	2.90	2.86	2.83	2.81
<b>17</b>	3.07	3.00	2.96	2.92	2.87	2.83	2.80	2.76	2.73	2.71
<b>18</b>	2.98	2.92	2.87	2.84	2.78	2.75	2.71	2.68	2.64	2.62
<b>19</b>	2.91	2.84	2.80	2.76	2.71	2.67	2.64	2.60	2.57	2.55
<b>20</b>	2.84	2.78	2.73	2.69	2.64	2.61	2.57	2.54	2.50	2.48
<b>21</b>	2.79	2.72	2.67	2.64	2.58	2.55	2.51	2.48	2.44	2.42
<b>22</b>	2.73	2.67	2.62	2.58	2.53	2.50	2.46	2.42	2.38	2.36
<b>23</b>	2.69	2.62	2.57	2.54	2.48	2.45	2.41	2.37	2.34	2.32
<b>24</b>	2.64	2.58	2.53	2.49	2.44	2.40	2.37	2.33	2.29	2.27
<b>25</b>	2.60	2.54	2.49	2.45	2.40	2.36	2.33	2.29	2.25	2.23
<b>26</b>	2.57	2.50	2.45	2.42	2.36	2.33	2.29	2.25	2.21	2.19
<b>27</b>	2.54	2.47	2.42	2.38	2.33	2.29	2.26	2.22	2.18	2.16
<b>28</b>	2.51	2.44	2.39	2.35	2.30	2.26	2.23	2.19	2.15	2.13
<b>29</b>	2.48	2.41	2.36	2.33	2.27	2.23	2.20	2.16	2.12	2.10
<b>30</b>	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.13	2.09	2.07
<b>35</b>	2.35	2.28	2.23	2.19	2.14	2.10	2.06	2.02	1.98	1.96
<b>40</b>	2.27	2.20	2.15	2.11	2.06	2.02	1.98	1.94	1.90	1.87
<b>50</b>	2.17	2.10	2.05	2.01	1.95	1.91	1.87	1.82	1.78	1.76
<b>60</b>	2.10	2.03	1.98	1.94	1.88	1.84	1.79	1.75	1.70	1.68
<b>70</b>	2.05	1.98	1.93	1.89	1.83	1.78	1.74	1.70	1.65	1.62
<b>80</b>	2.01	1.94	1.89	1.85	1.79	1.75	1.70	1.65	1.61	1.58
<b>90</b>	1.99	1.92	1.86	1.82	1.76	1.72	1.67	1.62	1.57	1.55
<b>100</b>	1.97	1.89	1.84	1.80	1.74	1.69	1.65	1.60	1.55	1.52
<b>120</b>	1.93	1.86	1.81	1.76	1.70	1.66	1.61	1.56	1.51	1.48
<b>150</b>	1.90	1.83	1.77	1.73	1.66	1.62	1.57	1.52	1.46	1.43
<b>200</b>	1.87	1.79	1.74	1.69	1.63	1.58	1.53	1.48	1.42	1.39
<b>250</b>	1.85	1.77	1.72	1.67	1.61	1.56	1.51	1.46	1.40	1.36
<b>300</b>	1.84	1.76	1.70	1.66	1.59	1.55	1.50	1.44	1.38	1.35
<b>400</b>	1.82	1.75	1.69	1.64	1.58	1.53	1.48	1.42	1.36	1.32
<b>500</b>	1.81	1.74	1.68	1.63	1.57	1.52	1.47	1.41	1.34	1.31
<b>600</b>	1.80	1.73	1.67	1.63	1.56	1.51	1.46	1.40	1.34	1.30
<b>750</b>	1.80	1.72	1.66	1.62	1.55	1.50	1.45	1.39	1.33	1.29
<b>1000</b>	1.79	1.72	1.66	1.61	1.54	1.50	1.44	1.38	1.32	1.28

TABLE A.3 (continued)

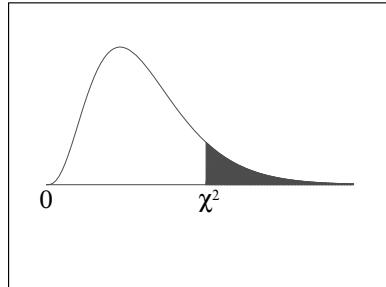
**F Distribution: Critical Values of F (0.1% significance level)**

$v_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
$v_2$															
<b>1</b>	4.05e05	5.00e05	5.40e05	5.62e05	5.76e05	5.86e05	5.93e05	5.98e05	6.02e05	6.06e05	6.11e05	6.14e05	6.17e05	6.19e05	6.21e05
<b>2</b>	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.42	999.43	999.44	999.44	999.45
<b>3</b>	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	128.32	127.64	127.14	126.74	126.42
<b>4</b>	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.95	46.60	46.32	46.10
<b>5</b>	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	26.42	26.06	25.78	25.57	25.39
<b>6</b>	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.99	17.68	17.45	17.27	17.12
<b>7</b>	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.43	13.23	13.06	12.93
<b>8</b>	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	11.19	10.94	10.75	10.60	10.48
<b>9</b>	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.33	9.15	9.01	8.90
<b>10</b>	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.45	8.22	8.05	7.91	7.80
<b>11</b>	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.41	7.24	7.11	7.01
<b>12</b>	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.79	6.63	6.51	6.40
<b>13</b>	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.31	6.16	6.03	5.93
<b>14</b>	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.93	5.78	5.66	5.56
<b>15</b>	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.62	5.46	5.35	5.25
<b>16</b>	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.55	5.35	5.20	5.09	4.99
<b>17</b>	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.13	4.99	4.87	4.78
<b>18</b>	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.94	4.80	4.68	4.59
<b>19</b>	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.78	4.64	4.52	4.43
<b>20</b>	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.64	4.49	4.38	4.29
<b>21</b>	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.51	4.37	4.26	4.17
<b>22</b>	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.40	4.26	4.15	4.06
<b>23</b>	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.30	4.16	4.05	3.96
<b>24</b>	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.21	4.07	3.96	3.87
<b>25</b>	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.13	3.99	3.88	3.79
<b>26</b>	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	4.06	3.92	3.81	3.72
<b>27</b>	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.99	3.86	3.75	3.66
<b>28</b>	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.93	3.80	3.69	3.60
<b>29</b>	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.88	3.74	3.63	3.54
<b>30</b>	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.82	3.69	3.58	3.49
<b>35</b>	12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18	4.03	3.79	3.62	3.48	3.38	3.29
<b>40</b>	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.47	3.34	3.23	3.14
<b>50</b>	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.27	3.41	3.04	2.95
<b>60</b>	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.15	3.02	2.91	2.83
<b>70</b>	11.80	7.64	6.06	5.20	4.66	4.28	3.99	3.77	3.60	3.45	3.23	3.06	2.93	2.83	2.74
<b>80</b>	11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.16	3.00	2.87	2.76	2.68
<b>90</b>	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34	3.11	2.95	2.82	2.71	2.63
<b>100</b>	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.91	2.78	2.68	2.59
<b>120</b>	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.85	2.72	2.62	2.53
<b>150</b>	11.27	7.24	5.71	4.88	4.35	3.98	3.71	3.49	3.32	3.18	2.96	2.80	2.67	2.56	2.48
<b>200</b>	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12	2.90	2.74	2.61	2.51	2.42
<b>250</b>	11.09	7.10	5.59	4.77	4.25	3.88	3.61	3.40	3.23	3.09	2.87	2.71	2.58	2.48	2.39
<b>300</b>	11.04	7.07	5.56	4.75	4.22	3.86	3.59	3.38	3.21	3.07	2.85	2.69	2.56	2.46	2.37
<b>400</b>	10.99	7.03	5.53	4.71	4.19	3.83	3.56	3.35	3.18	3.04	2.82	2.66	2.53	2.43	2.34
<b>500</b>	10.96	7.00	5.51	4.69	4.18	3.81	3.54	3.33	3.16	3.02	2.81	2.64	2.52	2.41	2.33
<b>600</b>	10.94	6.99	5.49	4.68	4.16	3.80	3.53	3.32	3.15	3.01	2.80	2.63	2.51	2.40	2.32
<b>750</b>	10.91	6.97	5.48	4.67	4.15	3.79	3.52	3.31	3.14	3.00	2.78	2.62	2.49	2.39	2.31
<b>1000</b>	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99	2.77	2.61	2.48	2.38	2.30

**TABLE A.3 (continued)****F Distribution: Critical Values of F (0.1% significance level)**

<i>v<sub>1</sub></i>	25	30	35	40	50	60	75	100	150	200
<i>v<sub>2</sub></i>										
<b>1</b>	6.24e05	6.26e05	6.28e05	6.29e05	6.30e05	6.31e05	6.32e05	6.33e05	6.35e05	6.35e05
<b>2</b>	999.46	999.47	999.47	999.47	999.48	999.48	999.49	999.49	999.49	999.49
<b>3</b>	125.84	125.45	125.17	124.96	124.66	124.47	124.27	124.07	123.87	123.77
<b>4</b>	45.70	45.43	45.23	45.09	44.88	44.75	44.61	44.47	44.33	44.26
<b>5</b>	25.08	24.87	24.72	24.60	24.44	24.33	24.22	24.12	24.01	23.95
<b>6</b>	16.85	16.67	16.54	16.44	16.31	16.21	16.12	16.03	15.93	15.89
<b>7</b>	12.69	12.53	12.41	12.33	12.20	12.12	12.04	11.95	11.87	11.82
<b>8</b>	10.26	10.11	10.00	9.92	9.80	9.73	9.65	9.57	9.49	9.45
<b>9</b>	8.69	8.55	8.46	8.37	8.26	8.19	8.11	8.04	7.96	7.93
<b>10</b>	7.60	7.47	7.37	7.30	7.19	7.12	7.05	6.98	6.91	6.87
<b>11</b>	6.81	6.68	6.59	6.52	6.42	6.35	6.28	6.21	6.14	6.10
<b>12</b>	6.22	6.09	6.00	5.93	5.83	5.76	5.70	5.63	5.56	5.52
<b>13</b>	5.75	5.63	5.54	5.47	5.37	5.30	5.24	5.17	5.10	5.07
<b>14</b>	5.38	5.25	5.17	5.10	5.00	4.94	4.87	4.81	4.74	4.71
<b>15</b>	5.07	4.95	4.86	4.80	4.70	4.64	4.57	4.51	4.44	4.41
<b>16</b>	4.82	4.70	4.61	4.54	4.45	4.39	4.32	4.26	4.19	4.16
<b>17</b>	4.60	4.48	4.40	4.33	4.24	4.18	4.11	4.05	3.98	3.95
<b>18</b>	4.42	4.30	4.22	4.15	4.06	4.00	3.93	3.87	3.80	3.77
<b>19</b>	4.26	4.14	4.06	3.99	3.90	3.84	3.78	3.71	3.65	3.61
<b>20</b>	4.12	4.00	3.92	3.86	3.77	3.70	3.64	3.58	3.51	3.48
<b>21</b>	4.00	3.88	3.80	3.74	3.64	3.58	3.52	3.46	3.39	3.36
<b>22</b>	3.89	3.78	3.70	3.63	3.54	3.48	3.41	3.35	3.28	3.25
<b>23</b>	3.79	3.68	3.60	3.53	3.44	3.38	3.32	3.25	3.19	3.16
<b>24</b>	3.71	3.59	3.51	3.45	3.36	3.29	3.23	3.17	3.10	3.07
<b>25</b>	3.63	3.52	3.43	3.37	3.28	3.22	3.15	3.09	3.03	2.99
<b>26</b>	3.56	3.44	3.36	3.30	3.21	3.15	3.08	3.02	2.95	2.92
<b>27</b>	3.49	3.38	3.30	3.23	3.14	3.08	3.02	2.96	2.89	2.86
<b>28</b>	3.43	3.32	3.24	3.18	3.09	3.02	2.96	2.90	2.83	2.80
<b>29</b>	3.38	3.27	3.18	3.12	3.03	2.97	2.91	2.84	2.78	2.74
<b>30</b>	3.33	3.22	3.13	3.07	2.98	2.92	2.86	2.79	2.73	2.69
<b>35</b>	3.13	3.02	2.93	2.87	2.78	2.72	2.66	2.59	2.52	2.49
<b>40</b>	2.98	2.87	2.79	2.73	2.64	2.57	2.51	2.44	2.38	2.34
<b>50</b>	2.79	2.68	2.60	2.53	2.44	2.38	2.31	2.25	2.18	2.14
<b>60</b>	2.67	2.55	2.47	2.41	2.32	2.25	2.19	2.12	2.05	2.01
<b>70</b>	2.58	2.47	2.39	2.32	2.23	2.16	2.10	2.03	1.95	1.92
<b>80</b>	2.52	2.41	2.32	2.26	2.16	2.10	2.03	1.96	1.89	1.85
<b>90</b>	2.47	2.36	2.27	2.21	2.11	2.05	1.98	1.91	1.83	1.79
<b>100</b>	2.43	2.32	2.24	2.17	2.08	2.01	1.94	1.87	1.79	1.75
<b>120</b>	2.37	2.26	2.18	2.11	2.02	1.95	1.88	1.81	1.73	1.68
<b>150</b>	2.32	2.21	2.12	2.06	1.96	1.89	1.82	1.74	1.66	1.62
<b>200</b>	2.26	2.15	2.07	2.00	1.90	1.83	1.76	1.68	1.60	1.55
<b>250</b>	2.23	2.12	2.03	1.97	1.87	1.80	1.72	1.65	1.56	1.51
<b>300</b>	2.21	2.10	2.01	1.94	1.85	1.78	1.70	1.62	1.53	1.48
<b>400</b>	2.18	2.07	1.98	1.92	1.82	1.75	1.67	1.59	1.50	1.45
<b>500</b>	2.17	2.05	1.97	1.90	1.80	1.73	1.65	1.57	1.48	1.43
<b>600</b>	2.16	2.04	1.96	1.89	1.79	1.72	1.64	1.56	1.46	1.41
<b>750</b>	2.15	2.03	1.95	1.88	1.78	1.71	1.63	1.55	1.45	1.40
<b>1000</b>	2.14	2.02	1.94	1.87	1.77	1.69	1.62	1.53	1.44	1.38

## Chi-Square Distribution Table



The shaded area is equal to  $\alpha$  for  $\chi^2 = \chi_{\alpha}^2$ .

$df$	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

## Critical values of Z

Critical values( $Z_\alpha$ )	Level of significance( $\alpha$ )			
	1%	2%	5%	10%
Two-tailed test	$ z_\alpha  = 2.58$	$ z_\alpha  = 2.33$	$ z_\alpha  = 1.96$	$ z_\alpha  = 1.645$
Right-tailed	$z_\alpha = 2.33$	$z_\alpha = 2.054$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left-tailed	$z_\alpha = -2.33$	$z_\alpha = -2.054$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

# **MAT2001**

# **Statistics for Engineers**

## **Module 1**

## **Introduction to Statistics**

### **Syllabus**

#### **Introduction to Statistics:**

Introduction to statistics and data analysis-Measures of central tendency - Measures of variability- [Moments-Skewness-Kurtosis (Concepts only)].

# **Statistics**

- **Introduction**
- **Data Science**
- **Data Analysis**

## **Frequency Distribution**

1. Discrete Frequency Distribution
2. Continuous Frequency Distribution

# **Measures of Central Tendency**

## **List of Measures of Central Tendency:**

1. Arithmetic Mean or Average
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

## 1. Arithmetic Mean or Average ( $M$ )

**Arithmetic Mean.** Arithmetic mean of a set of observations is their sum divided by the number of observations, e.g., the arithmetic mean  $\bar{x}$  of  $n$  observations  $x_1, x_2, \dots, x_n$  is given by

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

In case of frequency distribution  $x_i | f_i$ ,  $i = 1, 2, \dots, n$ , where  $f_i$  is the frequency of the variable  $x_i$ ;

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{1}{N} \sum_{i=1}^n f_i x_i, \quad \left[ \sum_{i=1}^n f_i = N \right]$$

In case of grouped or continuous frequency distribution,  $x$  is taken as the mid-value of the corresponding class.

## Arithmetic Mean or Average

$x_i :$	$x_1$	$x_2$	$\dots$	$x_n$
$f_i :$	$f_1$	$f_2$	$\dots$	$f_n$

$$\text{Mean} = M = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

where  $N = \sum_{i=1}^n f_i$ .

## Example:

Find the arithmetic mean of the following frequency distribution:

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6

## Solution:

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	73	299

$$\bar{x} = \frac{1}{N} \sum f_i x_i = \frac{299}{73} = 4.09$$

## Example:

Calculate the arithmetic mean of the marks from the following table :

Marks	: 0-10	10-20	20-30	30-40	40-50	50-60
No. of students	: 12	18	27	20	17	6

## Solution:

Marks	No. of students (f)	Mid - point (x)	fx
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	100		2,800

$$\text{Arithmetic mean or } \bar{x} = \frac{1}{N} \sum f x = \frac{1}{100} \times 2,800 = 28$$

## Arithmetic Mean or Average

It may be noted that if the values of  $x$  or (and)  $f$  are large, the calculation of mean by formula is quite time-consuming and tedious. The arithmetic is reduced to a great extent, by taking the deviations of the given values from any arbitrary point 'A', as explained below.

Let  $d_i = x_i - A$ , then  $f_i d_i = f_i (x_i - A) = f_i x_i - Af_i$

Summing both sides over  $i$  from 1 to  $n$ , we get

$$\sum_{i=1}^n f_i d_i = \sum_{i=1}^n f_i x_i - A \sum_{i=1}^n f_i = \sum_{i=1}^n f_i x_i - A \cdot N.$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^n f_i d_i = \frac{1}{N} \sum_{i=1}^n f_i x_i - A = \bar{x} - A$$

where  $\bar{x}$  is the arithmetic mean of the distribution.

$$\therefore \bar{x} = A + \frac{1}{N} \sum_{i=1}^n f_i d_i$$

## Arithmetic Mean or Average

In case of grouped or continuous frequency distribution, the arithmetic is reduced to a still greater extent by taking

$$d_i = \frac{x_i - A}{h},$$

where  $A$  is an arbitrary point and  $h$  is the common magnitude of class interval. In this case, we have

$$h d_i = x_i - A,$$

and proceeding exactly similarly as above, we get

$$\bar{x} = A + \frac{h}{N} \sum_{i=1}^n f_i d_i$$

## Example:

Calculate the mean for the following frequency distribution.

Class-interval :	0-8	8-16	16-24	24-32	32-40	40-48
Frequency :	8	7	16	24	15	7

## Solution:

Class-interval	mid-value (x)	Frequency (f)	$d = (x - A) / h$	$fd$
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16
24-32	28	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
		77		-25

Here we take  $A = 28$  and  $h = 8$ .

$$\therefore \bar{x} = A + \frac{h \sum f d}{N} = 28 + \frac{8 \times (-25)}{77} = 28 - \frac{200}{77} = 25.404$$

## 2. Median ( $M_d$ )

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, i.e., it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a *positional average*.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

### Example:

For example, the median of the values 25, 20, 15, 35, 18, i.e., 15, 18, 20, 25, 35 is 20 and the median of 8, 20, 50, 25, 15, 30, i.e., of 8, 15, 20, 25, 30, 50 is  $\frac{1}{2}(20 + 25) = 22.5$ .

## Median for Discrete Frequency Distribution

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

- (i) Find  $N/2$ , where  $N = \sum_i f_i$ .
- (ii) See the (less than) cumulative frequency (c.f.) just greater than  $N/2$ .
- (iii) The corresponding value of  $x$  is median.

### Example:

Obtain the median for the following frequency distribution:

$x$ :	1	2	3	4	5	6	7	8	9
$f$ :	8	10	11	16	20	25	15	9	6

### Solution:

$x$	$f$	c.f.
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	120	

$$\text{Hence } N = 120 \Rightarrow N/2 = 60$$

Cumulative frequency (c.f.) just greater than  $N/2$ , is 65 and the value of  $x$  corresponding to 65 is 5. Therefore, median is 5.

## Median for Continuous Frequency Distribution

In the case of continuous frequency distribution, the class corresponding to the c.f. just greater than  $N/2$  is called the *median class* and the value of median is obtained by the following formula :

$$\text{Median} = l + \frac{h}{f} \left( \frac{N}{2} - c \right)$$

where  $l$  is the lower limit of the median class,

$f$  is the frequency of the median class,

$h$  is the magnitude of the median class,

' $c$ ' is the c.f. of the class preceding the median class,

and  $N = \Sigma f$ .

### **Example:**

*Find the median wage of the following distribution :*

<i>Wages (in Rs.)</i> :	20—30	30—40	40—50	50—60	60—70
<i>No. of labourers</i> :	3	5	20	10	5

### **Solution:**

<i>Wages (in Rs.)</i>	<i>No. of labourers</i>	<i>c.f.</i>
20—30	3	3
30—40	5	8
40—50	20	28
50—60	10	38
60—70	5	43

Here  $N/2 = 43/2 = 21.5$ . Cumulative frequency just greater than 21.5 is 28 and the corresponding class is 40—50. Thus median class is 40—50.

$$\text{Median} = 40 + \frac{10}{20}(21.5 - 8) = 40 + 6.75 = 46.75$$

Thus median wage is Rs. 46.75.

### **3. Mode ( $M_o$ )**

**Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series.**

## Mode for Discrete Frequency Distribution

In case of discrete frequency distribution mode is the value of  $x$  corresponding to maximum frequency.

### Example:

For the following discrete frequency distribution,

$x$	:	1	2	3	4	5	6	7	8
$f$	:	4	9	16	25	22	15	7	3

the value of  $x$  corresponding to the maximum frequency, viz., 25 is 4. Hence mode is 4.

## Mode for Discrete Frequency Distribution

But in any one (or more) of the following cases :

- (i) if the maximum frequency is repeated,
  - (ii) if the maximum frequency occurs in the very beginning or at the end of the distribution, and
  - (iii) if there are irregularities in the distribution,
- the value of mode is determined by the *method of grouping*, which is illustrated below by an example.

## Mode for Discrete Frequency Distribution

### Method of Grouping

#### Example:

*Find the mode of the following frequency distribution :*

Size (x) :	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f) :	3	8	15	23	35	40	32	28	20	45	14	6

#### Solution:

Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below :

The frequencies in column (i) are the original frequencies. Column (ii) is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two we get column (iii). Combining the frequencies two by two after leaving the first two frequencies results in a repetition of column (ii). Hence, we proceed to combine the frequencies three by three, thus getting column (iv). The combination of frequencies three by three after leaving the first frequency results in column (v) and after leaving the first two frequencies results in column (vi).

## Solution (Continued):

Size (x)	Frequency					
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
1	3					
2	8	11				
3	15	23	26			
4	23	38	58			
5	35	58	98			
6	40	75	72	107		
7	32	72				
8	28	60	80			
9	20	48	80	93		
10	45	65	59	65		
11	14	59				
12	6	20				

## Solution (Continued):

The maximum frequency in each column is given in black type. To find mode we form the following table :

ANALYSIS TABLE

<i>Column Number (1)</i>	<i>Maximum Frequency (2)</i>	<i>Value or combination of values of <math>x</math> giving max. frequency in (2) (3)</i>
(i)	45	10
(ii)	75	5, 6
(iii)	72 .....	6, 7
(iv)	98	4, 5, 6,
(v)	107	5, 6, 7
(vi)	100	6, 7, 8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

## Mode for Continuous Frequency Distribution

In case of continuous frequency distribution, mode is given by the formula :

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

where  $l$  is the lower limit,  $h$  the magnitude and  $f_1$  the frequency of the modal class,  $f_0$  and  $f_2$  are the frequencies of the classes preceding and succeeding the modal class respectively.

### **Example:**

*Find the mode for the following distribution :*

<i>Class - interval</i> :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
<i>Frequency</i> :	5	8	7	12	28	20	10	10

### **Solution:**

Here maximum frequency is 28. Thus the class 40-50 is the modal class.  
the value of mode is given by

$$\text{Mode} = 40 + \frac{10(28 - 12)}{(2 \times 28 - 12 - 20)} = 40 + 6.666 = 46.67 \text{ (approx.)}$$

## 4. Geometric Mean ( $G$ )

Geometric mean of a set of  $n$  observations is the  $n$ th root of their product. Thus the geometric mean  $G$ , of  $n$  observations  $x_i, i = 1, 2, \dots, n$  is

$$G = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$$

The computation is facilitated by the use of logarithms. Taking logarithm of both sides, we get

$$\log G = \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n) = \frac{1}{n} \sum_{i=1}^n \log x_i$$

$$\therefore G = \text{Antilog} \left[ \frac{1}{n} \sum_{i=1}^n \log x_i \right]$$

## 4. Geometric Mean (Continued)

In case of frequency distribution  $x_i | f_i$ , ( $i = 1, 2, \dots, n$ ) geometric mean,  $G$  is given by

$$G = [x_1^{f_1} \cdot x_2^{f_2} \cdots \cdot x_n^{f_n}]^{\frac{1}{N}}, \text{ where } N = \sum_{i=1}^n f_i$$

Taking logarithms of both sides, we get

$$\begin{aligned}\log G &= \frac{1}{N} (f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n) \\ &= \frac{1}{N} \sum_{i=1}^n f_i \log x_i\end{aligned}$$

Thus we see that logarithm of  $G$  is the arithmetic mean of the logarithms of the given values.

$$G = \text{Antilog} \left( \frac{1}{N} \sum_{i=1}^n f_i \log x_i \right)$$

In the case of grouped or continuous frequency distribution,  $x$  is taken to be the value corresponding to the mid-point of the class-intervals.

## 5. Harmonic Mean ( $H$ )

Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus, harmonic mean  $H$ , of  $n$  observations  $x_i$ ,  $i = 1, 2, \dots, n$  is

$$H = \frac{1}{\frac{1}{n} \sum_{i=1}^n (1/x_i)}$$

In case of frequency distribution  $x_i | f_i$ , ( $i = 1, 2, \dots, n$ ),

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^n (f_i/x_i)}, \quad \left[ N = \sum_{i=1}^n f_i \right]$$

# **Measures of Variability or Measures of Dispersion**

Dispersion - Variations or Scatteredness

## **List of Measures of Dispersion:**

1. Range
2. Quartile Deviation or Semi-interquartile Range
3. Mean Deviation and
4. Standard Deviation

## **1. Range**

**Range = Maximum Value - Minimum Value**

The range is the difference between two extreme observations of the distribution. If  $A$  and  $B$  are the greatest and smallest observations respectively in a distribution, then its range is  $A - B$ .

Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

## 2. Quartile Deviation or Semi-interquartile Range

### Partition Values

These are the values which divide the series into a number of equal parts.

### Quartiles

The three points which divide the series into four equal parts are called *quartiles*. The first, second and third points are known as the first, second and third quartiles respectively. The first quartile,  $Q_1$ , is the value which exceed 25% of the observations and is exceeded by 75% of the observations. The second quartile,  $Q_2$ , coincides with median. The third quartile,  $Q_3$ , is the point which has 75% observations before it and 25% observations after it.

## Quartile Deviation or Semi-interquartile Range (Q)

$Q$  is given by

$$Q = \frac{1}{2} (Q_3 - Q_1),$$

where  $Q_1$  and  $Q_3$  are the first and third quartiles of the distribution respectively.

Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

Quartile Deviation (Q) =  $(1/2)(Q_3 - Q_1)$ .

$$Q = \frac{Q_3 - Q_1}{2}$$

## Example:

Eight coins were tossed together and the number of heads resulting was noted. The operation was repeated 256 times and the frequencies ( $f$ ) that were obtained for different values of  $x$ , the number of heads, are shown in the following table. Calculate median, quartiles.

$x :$	0	1	2	3	4	5	6	7	8
$f :$	1	9	26	59	72	52	29	7	1

## Solution:

$x :$	0	1	2	3	4	5	6	7	8
$f :$	1	9	26	59	72	52	29	7	1
$c.f. :$	1	10	36	95	167	219	248	255	256

Median : Here  $N/2 = 256/2 = 128$ . Cumulative frequency ( $c.f.$ ) just greater than 128 is 167. Thus, median = 4.

$Q_1$  : Here  $\underline{N/4} = 64$ .  $c.f.$  just greater than 64 is 95. Hence,  $Q_1 = 3$ .

$Q_3$  : Here  $\underline{3N/4} = 192$  and  $c.f.$  just greater than 192 is 219. Thus  $Q_3 = 5$ .

$$\text{Quartile Deviation (Q)} = (1/2)(Q_3 - Q_1).$$

### 3. Mean Deviation

If  $x_i | f_i, i = 1, 2, \dots, n$  is the frequency distribution, then mean deviation from the average A, (usually mean, median or mode), is given by

$$\text{Mean deviation} = \frac{1}{N} \sum_i f_i |x_i - A|, \quad \sum f_i = N$$

where  $|x_i - A|$  represents the modulus or the absolute value of the deviation  $(x_i - A)$ , when the -ive sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations  $(x_i - A)$  creates artificiality and renders it useless for further mathematical treatment.

It may be pointed out here that mean deviation is least when taken from median.

$$MD = \frac{1}{N} \sum_{i=1}^n f_i |x_i - A|, \quad N = \sum_{i=1}^n f_i$$

## 4. Standard Deviation ( $\sigma$ )

Standard

deviation, usually denoted by the Greek letter small sigma ( $\sigma$ ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution  $x_i | f_i, i = 1, 2, \dots, n$ ,

$$\sigma = \sqrt{\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2}$$

where  $\bar{x}$  is the arithmetic mean of the distribution and  $\sum_i f_i = N$ .

Mean

## Variance

The square of standard deviation is called the *variance* and is given by

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

## Root Mean Square Deviation

Root mean square deviation, denoted by 's' is given by

$$s = \sqrt{\frac{1}{N} \sum_i f_i (x_i - A)^2}$$

where  $A$  is any arbitrary number.  $s^2$  is called mean square deviation.

## Difference Formulae for Calculating Variance

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i x_i^2 - \left( \frac{1}{N} \sum_i f_i x_i \right)^2$$

We know that if  $d_i = x_i - A$  then  $\bar{x} = A + \frac{1}{N} \sum_i f_i d_i$

$$A - \bar{x} = -\frac{1}{N} \sum_i f_i d_i$$

Hence

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i d_i^2 + \left( -\frac{1}{N} \sum_i f_i d_i \right)^2 + 2 \left( -\frac{1}{N} \sum_i f_i d_i \right) \left( \frac{1}{N} \sum_i f_i d_i \right)$$

$$= \frac{1}{N} \sum_i f_i d_i^2 - \left( \frac{1}{N} \sum_i f_i d_i \right)^2$$

$\Rightarrow$

$$\sigma_x^2 = \sigma_d^2$$

Hence variance and consequently standard deviation is independent of change of origin.

## Difference Formulae for Calculating Variance

If we take  $d_i = (x_i - A)/h$  so that  $(x_i - A) = hd_i$ , then

$$\begin{aligned}\sigma_x^2 &= \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i - A + A - \bar{x})^2 \\&= \frac{1}{N} \sum_i f_i (hd_i + A - \bar{x})^2 \\&= h^2 \frac{1}{N} \sum_i f_i d_i^2 + (A - \bar{x})^2 + 2(A - \bar{x}) \cdot h \cdot \frac{1}{N} \sum_i f_i d_i\end{aligned}$$

Using  $\bar{x} = A + h \frac{\sum f_i d_i}{N}$ , we get

$$\sigma_x^2 = h^2 \left[ \frac{1}{N} \sum_i f_i d_i^2 - \left( \frac{1}{N} \sum_i f_i d_i \right)^2 \right] = h^2 \sigma_d^2,$$

$$\boxed{\sigma_x^2 = h^2 \sigma_d^2}$$

which shows that variance is not independent of change of scale.  
Hence variance is independent of change of origin but not of scale.

## Co-efficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages or which are measured in different units, we do not merely calculate the measures of dispersion but we calculate the co-efficients of dispersion which are pure numbers independent of the units of measurement. The co-efficients of dispersion (C.D.) based on different measures of dispersion are as follows :

1. C.D. based upon range =  $\frac{A - B}{A + B}$ , where  $A$  and  $B$  are the greatest and the smallest items in the series.

2. Based upon quartile deviation :

$$C.D. = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Based upon mean deviation :

$$C.D. = \frac{\text{Mean deviation}}{\text{Average from which it is calculated}}$$

4. Based upon standard deviation :

$$C.D. = \frac{S.D.}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

## Co-efficient of Variation

100 times the co-efficient of dispersion based upon standard deviation is called co-efficient of variation (C.V.),

$$C.V. = 100 \times \frac{\sigma}{\bar{x}}$$

According to Professor Karl Pearson who suggested this measure, C.V. is *the percentage variation in the mean, standard deviation being considered as the total variation in the mean.*

For comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater C.V. is said to be more variable than the other and the series having lesser C.V. is said to be more consistent (or homogenous) than the other.

### Example:

Calculate the mean and standard deviation for the following table giving the age distribution of 542 members.

Age in years : 20—30 30—40 40—50 50—60 60—70 70—80 80—90

No. of members : 3 61 132 153 140 51 2

**Solution:** ALSO find CD & CV.

$$\text{Here we take } d = \frac{x - A}{h} = \frac{x - 55}{10}$$

Age group	Mid-value (x)	Frequency (f)	$d = \frac{x - 55}{10}$	$fd$	$fd^2$
20 — 30	25	3	-3	-9	27
30 — 40	35	61	-2	-122	244
40 — 50	45	132	-1	-132	132
50 — 60	55	153	0	0	0
60 — 70	65	140	1	140	140
70 — 80	75	51	2	102	204
80 — 90	85	2	3	6	18
		$N = \sum f = 542$		$\sum fd = -15$	$\sum fd^2 = 765$

$$\bar{x} = A + h \frac{\sum fd}{N} = 55 + \frac{10 \times (-15)}{542} = 55 - 0.28 = 54.72 \text{ years.}$$

$$\sigma^2 = h^2 \left[ \frac{1}{N} \sum fd^2 - \left( \frac{1}{N} \sum fd \right)^2 \right] = 100 \left[ \frac{765}{542} - (0.28)^2 \right]$$

$$CD = \frac{\sigma}{\bar{x}}$$

$$= 100 \times 1.333 = 133.3$$

$$\sigma (\text{standard deviation}) = 11.55 \text{ years}$$

$$CV = 100 \times \frac{\sigma}{\bar{x}}$$

## Moments

The  $r$ th moment of a variable  $x$  about any point  $x = A$ , usually denoted by  $\mu'_r$  is given by

$$\mu'_r = \frac{1}{N} \sum_i f_i (x_i - A)^r, \quad \sum_i f_i = N$$

$$= \frac{1}{N} \sum_i f_i d_i^r,$$

$$\mu'_r = \frac{1}{N} \sum_{i=1}^n f_i (x_i - A)^r$$

where  $d_i = x_i - A$ .

The  $r$ th moment of a variable about the mean  $\bar{x}$ , usually denoted by  $\mu_r$  is given by

$$\mu_r = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^r = \frac{1}{N} \sum_i f_i z_i^r$$

where  $z_i = x_i - \bar{x}$ .

$$\mu_r = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^r$$

$A = \bar{x}$

## Particular Cases

In particular

$$\mu_0 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^0 = \frac{1}{N} \sum_i f_i = 1$$

and  $\mu_1 = \frac{1}{N} \sum_i f_i (x_i - \bar{x}) = 0$ , being the algebraic sum of deviations from the mean. Also

$$\mu_2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \sigma^2$$

These results, viz.,  $\mu_0 = 1$ ,  $\mu_1 = 0$ , and  $\mu_2 = \sigma^2$ , are of fundamental importance and should be committed to memory.

## Pearson's Co-efficients

Karl Pearson defined the following four coefficients, based upon the first four moments about mean :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2} \quad , \quad \gamma_1 = + \sqrt{\beta_1} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$


## Skewness

Literally, skewness means '*lack of symmetry*'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

- (i) Mean, median and mode fall at different points,  
*i.e.*,  $\text{Mean} \neq \text{Median} \neq \text{Mode}$ ,
- (ii) Quartiles are not equidistant from median, and
- (iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

## Measures of Skewness

**Measures of Skewness.** Various measures of skewness are

$$(1) S_k = M - M_d \quad (2) S_k = M' - M_0,$$

where  $M$  is the mean,  $M_d$ , the median and  $M_0$ , the mode of the distribution.

$$(3) S_k = (Q_3 - M_d) - (M_d - Q_1).$$

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the *co-efficients of skewness* which are pure numbers independent of units of measurement.

## Co-efficients of Skewness

I. Prof. Karl Pearson's Coefficient of Skewness.

$$S_k = \frac{(M - M_0)}{\sigma}$$

II. Prof. Bowley's Coefficient of Skewness. Based on quartiles,

$$S_K = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

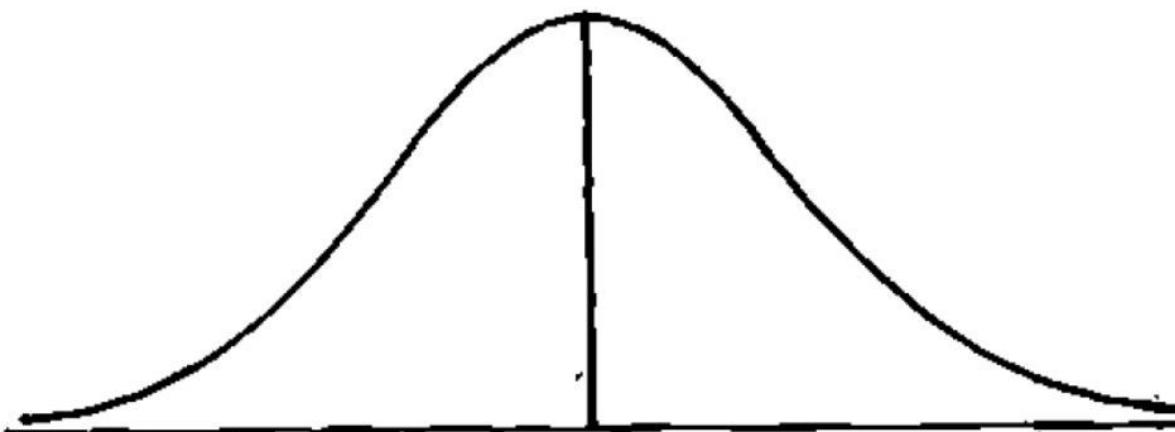
III. Based upon moments, co-efficient of skewness is

$$S_k = \frac{\sqrt{\beta_1} (\beta_2 + 3)}{2 (5\beta_2 - 6\beta_1 - 9)}$$

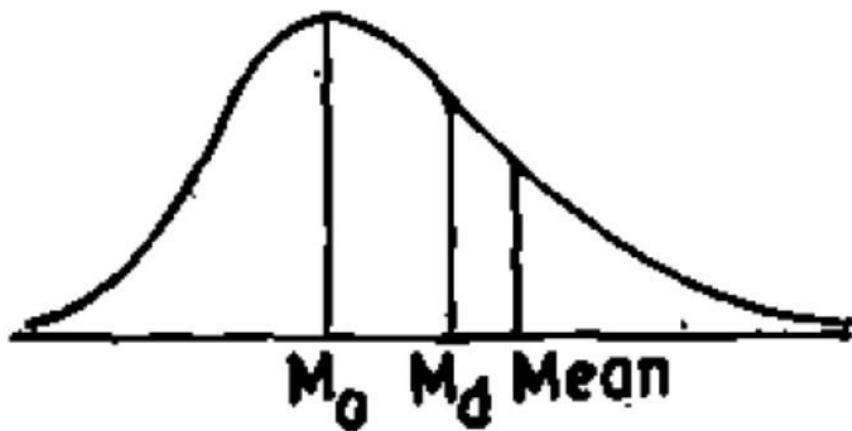
where symbols have their usual meaning. Thus  $S_k = 0$  if either  $\beta_1 = 0$  or  $\beta_2 = -3$ . But since  $\beta_2 = \mu_4/\mu_2^2$ , cannot be negative,  $S_k = 0$  if and only if  $\beta_1 = 0$ . Thus for a symmetrical distribution  $\beta_1 = 0$ . In this respect  $\beta_1$  is taken to be a measure of skewness.

The skewness is positive if the larger tail of the distribution lies towards the higher values of the variate (the right), i.e., if the curve drawn with the help of the given data is stretched more to the right than to the left and is negative

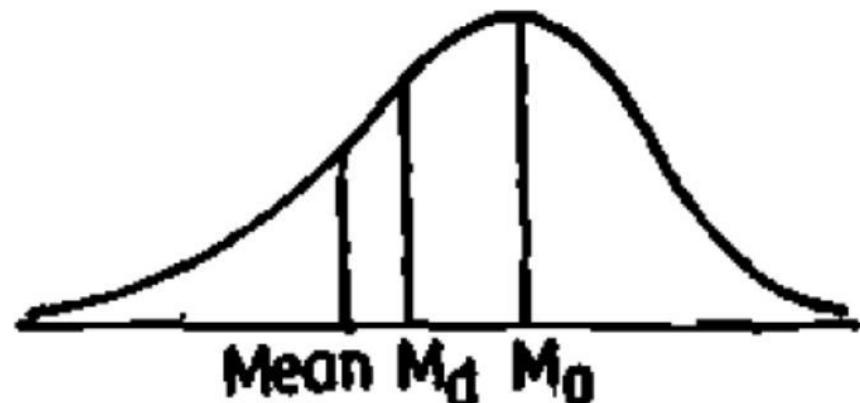
## Graphical Representation of Skewness



$\bar{x}$  (Mean) =  $M_0$  =  $M_d$   
(Symmetrical Distribution)



(Positively Skewed Distribution)



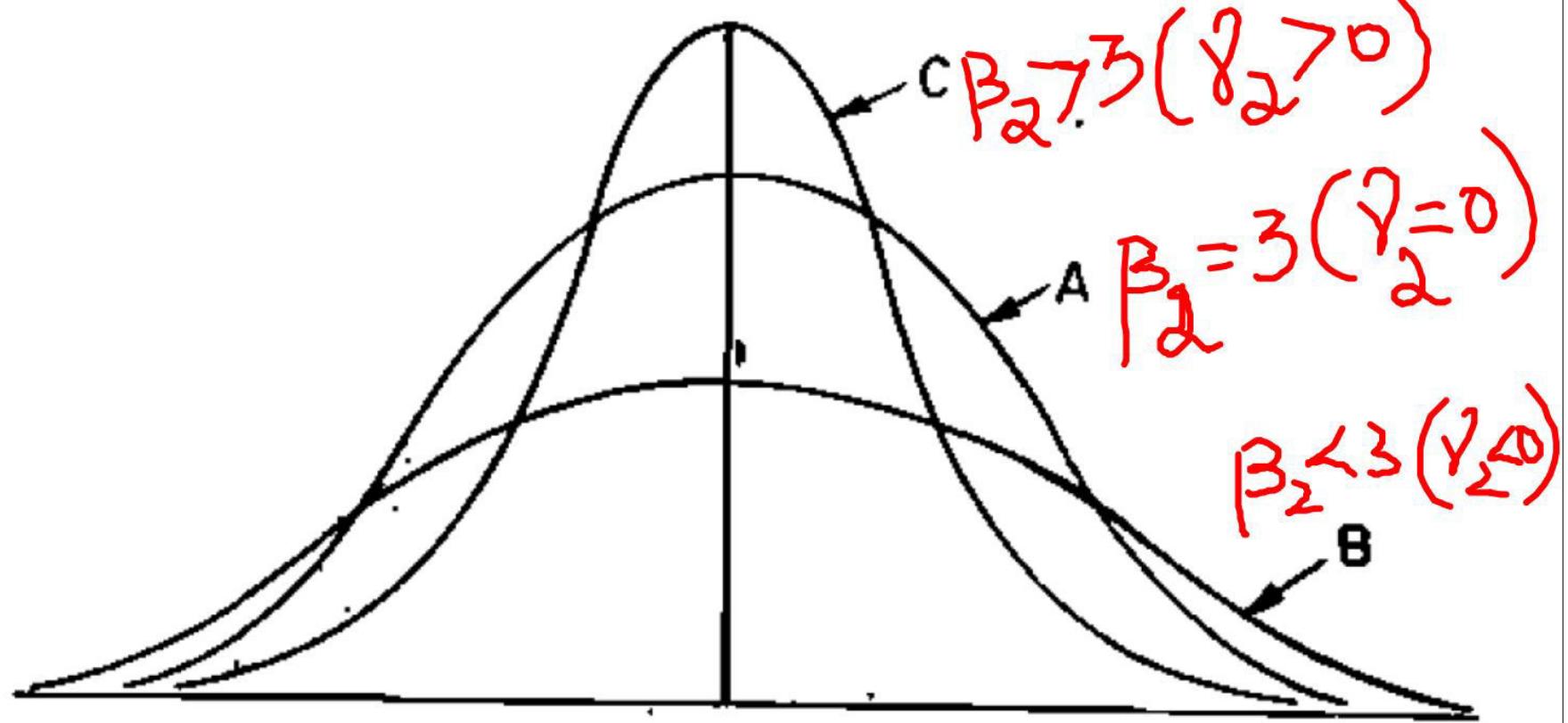
(Negatively Skewed Distribution)

## Kurtosis

If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the following figure in which all the three curves A, B and C are symmetrical about the mean ' $m$ ' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the 'Convexity of curve' or Kurtosis. Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the co-efficient  $\beta_2$  or its derivation  $r_2$  given by:

$$\beta_2 = \mu_4/\mu_2^2, \gamma_2 = \beta_2 - 3$$



Curve of the type 'A' which is neither flat nor peaked is called the *normal curve or mesokurtic curve* and for such a curve  $\beta_2 = 3$ , i.e.,  $\gamma_2 = 0$ . Curve of the type 'B' which is flatter than the normal curve is known as *platykurtic* and for such a curve  $\beta_2 < 3$ , i.e.,  $\gamma_2 < 0$ . Curve of the type 'C' which is more peaked than the normal curve is called *leptokurtic* and for such a curve  $\beta_2 > 3$ , i.e.,  $\gamma_2 > 0$ .



# **MAT2001**

## **Statistics for Engineers**

### **Module 2**

### **Random Variables**

## **Syllabus**

### **Random variables:**

Introduction -random variables-Probability mass Function, distribution and density functions - joint Probability distribution and joint density functions- Marginal, conditional distribution and density functions- Mathematical expectation, and its properties Covariance , moment generating function – characteristic function.

# Probability Theory

Probability theory had its origin in the analysis of certain games of chance that were popular in the seventeenth century. It has since found applications in many branches of Science and Engineering and this extensive application makes it an important branch of study. Probability theory, as a matter of fact, is a study of random or unpredictable experiments and is helpful in investigating the important features of these random experiments.

- Random Experiment
- Sample Space
- Events

## Random Experiment, Sample Space and Event

Problem (Non-deterministic)

Random Experiment:

$S \rightarrow$  Sample Space (All possible outcomes)

$\Omega = \mathcal{P}(S) =$  Power Set of  $S$

$$= \{A : A \subseteq S\}$$

Event:  $\rightarrow$  a subset of a sample space  $S$

$A$  is called an event, if  $A \subseteq S$

$$(i.e.) A \in \Omega = \mathcal{P}(S)$$

# Mathematical Definition of Probability

Let E - Experiment with Sample Space S.

and let A - an event.

Here S is finite

$n(S)$  is finite

$$\overline{P(A) = \frac{n(A)}{n(S)} = \frac{|A|}{|S|}}$$

$n(A)$  = no. of elements in A  
 $|A|$

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{Number of cases favourable to } A}{\text{Exhaustive number of cases in } S}$$

**Example:**

①.  $S = \{H, T\} \Rightarrow |S| = 2$

$$A = \{H\} \Rightarrow |A| = 1$$

$$P(A) = \frac{1}{2}$$

②.  $S = \{1, 2, \dots, 5\} \Rightarrow |S| = 5$

$$A = \{1, 3, 5\} \Rightarrow |A| = 3$$

$$P(A) = \frac{1}{2}$$

## Statistical Definition of Probability

n - Trials

A = Expected Event

$n_A$  = No. of occurrences of A

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

## Axiomatic Definition of Probability

Probability | Probability function |

Probability Measure

A function

$$P : \mathcal{P}(S) \longrightarrow [0, 1]$$

Let  $E$ -Exp's  
with S.S.S

Let  $A \subseteq S$

is called probability, if  $P$  satisfies

(i).  $P(S) = 1$  [or  $P(\emptyset) = 0$ ]

(ii). Addition theorem of Probability.

## Mutually Exclusive Events

Let  $A \times B$  are 2 events

$A$  and  $B$  are said to be mutually exclusive events, if  $A \cap B = \emptyset$

$$\text{(Or)} \quad P(A \cap B) = P(\emptyset) = 0$$

## Addition Theorem of Probability

Let  $A \times B$  be an events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Suppose  $A \times B$  are mutually exclusive

then

$$P(A \cup B) = P(A) + P(B)$$

In general

Suppose  $A_1, A_2, \dots, A_n, \dots$  are pair-wise  
mutually exclusive events,  
then

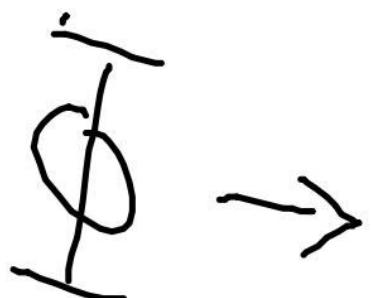
$$P(A_1 \cup A_2 \cup \dots \cup A_n \cup \dots) = P(A_1) + P(A_2) + \dots + \dots + P(A_n) + \dots$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Sure Event

Sample Event

$$P(S) = 1$$



Impossible Event

**Theorem 1**

The probability of the impossible event is zero, i.e., if  $\phi$  is the subset (event) containing no sample point,  $P(\phi) = 0$ .

**Theorem 2**

If  $\bar{A}$  is the complementary event of  $A$ ,  $P(\bar{A}) = 1 - P(A) \leq 1$ .

**Theorem 3**

If  $A$  and  $B$  are any 2 events,  $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$

**Theorem 4**

If  $B \subset A$ ,  $P(B) \leq P(A)$ .

## Conditional Probability

The conditional probability of an event  $B$ , assuming that the event  $A$  has happened, is denoted by  $P(B/A)$  and defined as

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ provided } P(A) \neq 0$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A) \times P(B/A)$$

### Conditional Probability

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \neq 0$$

Example

$$S = \{1, 2, \dots, 6\}$$

$$E = \{2, 4, 6\} \Rightarrow P(E) = \frac{1}{2}$$

$$O = \{1, 3, 5\} \Rightarrow P(O) = \frac{1}{2}$$

$$A = \{1\} \quad A \cap O = \{1\} \Rightarrow P(A \cap O) = \frac{1}{6}$$

$$P(A) = \frac{1}{6}$$

$$P(A/O) = \frac{P(A \cap O)}{P(O)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(A/E) = \frac{P(A \cap E)}{P(E)} = \frac{P(\emptyset)}{\frac{1}{2}} = 0$$

*Product theorem of probability*

$$P(A \cap B) = P(A) \times P(B/A)$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

The product theorem can be extended to 3 events  $A$ ,  $B$  and  $C$  as follows:

$$P(A \cap B \cap C) = P(A) \times P(B/A) \times P(C/A \text{ and } B)$$

### **Independent Events**

$$P(A \cap B) = P(A) \times P(B/A)$$

If  $A$  and  $B$  are independent to each other,

$$P(B/A) = P(B)$$

$$P(A \cap B) = P(A) \times P(B)$$

## **Independent Events**

If  $P(A \cap B) = P(A) \times P(B)$ ,

the events  $A$  and  $B$  are said to be independent (pairwise independent).

*In General,*

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n)$$

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i)$$

**Theorem 1**

If the events  $A$  and  $B$  are independent, the events  $\bar{A}$  and  $B$  (and similarly  $A$  and  $\bar{B}$ ) are also independent.

$$\bar{A} = S - A = S/A$$

**Theorem 2**

If the events  $A$  and  $B$  are independent, then so are  $\bar{A}$  and  $\bar{B}$ .

**Random Variables ( $X$ )**

$$X: S \rightarrow \mathbb{R}$$

$$\begin{aligned} X(s) &= 1 \\ R_X &= \{1\} \end{aligned}$$

$$\{X \leq x\} = \{\underline{s \in S : X(s) \leq x}\}$$

$$P(\{X \geq x\}) = P(\{s \in S : X(s) = x\})$$

Example:

$$S = \{1, 2, 3, \dots, 6\}$$

$$A = \{1, 2, 3\}, B = \{4, 5, 6\}$$

$$P(A) \text{ vs } P(B) = ?$$

$$X: S \rightarrow \mathbb{R} \implies R_X = \{-3, 3\}$$

$$X(s) = \begin{cases} -3; & \text{if } s \in \{1, 2, 3\} \\ +3; & \text{if } s \in \{4, 5, 6\} \end{cases}$$

$$\begin{aligned} \{X = -3\} &= \{s \in S : X(s) = -3\} \\ &= \{1, 2, 3\} \end{aligned}$$

$$P(A) = P(X = -3)$$

$$P(B) = P(X = 3)$$

# Random Variables

**Definition:** A random variable (abbreviatively RV) is a function that assigns a real number  $X(s)$  to every element  $s \in S$ , where  $S$  is the sample space corresponding to a random experiment  $E$ .)

Hereafter,  $R_x$  will be referred to as **Range space**.

Similarly  $\{X \leq x\}$  represents the subset  $\{s: X(s) \leq x\}$  and hence an event associated with the experiment.

## Discrete Random Variable

## Continuous Random Variable

$X = \Sigma x_i$ , where  $x_i \in \{x_1, x_2, \dots, x_n\}$

$X = x$ , where  $x \in [a, b] \quad \forall a, b \in \mathbb{R}$

$\Rightarrow R_x$

### Discrete Random Variable

$X - DRV$

$$X = x_1, x_2, \dots, x_n$$

$$x_i \in \mathbb{R}$$

Probability Mass function

$x_i$	$x_1$	...	$x_n$
$p_i$	$p_1$	...	$p_n$

(i)  $p_i$ 's  $\geq 0$

(ii).  $\sum_{i=1}^n p_i = 1$

$p_i$ 's are p.m.f

$\{x_i, p_i\}_{i=1}^n$  is called

prob. distribution of.

a DRV  $X$

### Continuous Random Variable

$X - CRV$

$$X = x \in [a, b] = R_x$$

$$[a, b] \subseteq \mathbb{R}$$

Prob. Density function

(p.d.f)

$f(x)$  is said p.d.f

if (i).  $f(x) \geq 0$ ,

(ii). (fre. ness)  $\forall x \in R_x$

$$\int_{R_x} f(x) dx = 1$$

$\{x, f(x)\}$

$x \in R_x$  is

called  $P.D.f.$  of a  
CRV  $X$

## **Discrete Random Variable**

If  $X$  is a random variable (RV) which can take a finite number or countably infinite number of values,  $X$  is called a discrete RV.

## **Probability Function**

If  $X$  is a discrete RV which can take the values  $x_1, x_2, x_3, \dots$  such that  $P(X = x_i) = p_i$ , then  $p_i$  is called the *probability function* or *probability mass function* or *point probability function*, provided  $p_i$  ( $i = 1, 2, 3, \dots$ ) satisfy the following conditions:

(i)  $p_i \geq 0$ , for all  $i$ , and

(ii)  $\sum_i p_i = 1$

The collection of pairs  $\{x_i, p_i\}$ ,  $i = 1, 2, 3, \dots$ , is called the *probability distribution of the RV X*, which is sometimes displayed in the form of a table as given below:

$X = x_i$	$P(X = x_i)$
$x_1$	$p_1$
$x_2$	$p_2$
$\vdots$	$\vdots$
$x_r$	$p_r$
$\vdots$	$\vdots$

## Continuous Random Variable

If  $X$  is an RV which can take all values (i.e., *infinite number* of values) in an interval, then  $X$  is called a *continuous* RV.

### Probability Density Function

If  $X$  is a continuous RV such that

$$P\left\{x - \frac{1}{2} dx \leq X \leq x + \frac{1}{2} dx\right\} = f(x)dx$$

then  $f(x)$  is called the *probability density function* (shortly denoted as pdf) of  $X$ , provided  $f(x)$  satisfies the following conditions:

(i)  $f(x) \geq 0$ , for all  $x \in R_x$ , and

$$(ii) \int_{R_X} f(x)dx = 1$$

Moreover,  $P(a \leq X \leq b)$  or  $P(a < X < b)$  is defined as

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

---

When  $X$  is a continuous RV

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0$$

## Cumulative Distribution Function (cdf)

If  $X$  is an R V, discrete or continuous, then  $P(X \leq x)$  is called the *cumulative distribution function* of  $X$  or *distribution function* of  $X$  and denoted as  $F(x)$ .

If  $X$  is discrete,

$$F(x) = \sum_j p_j$$
$$x_j \leq x$$

If  $X$  is continuous,

$$F(x) = P(-\infty < X \leq x) = \int_{-\infty}^x f(x) dx$$

### Properties of the cdf $F(x)$

1.  $F(x)$  is a non-decreasing function of  $x$ , i.e., if  $x_1 < x_2$ , then  $F(x_1) \leq F(x_2)$ .
2.  $F(-\infty) = 0$  and  $F(\infty) = 1$ .
3. If  $X$  is a discrete R V taking values  $x_1, x_2, \dots$ , where  $x_1 < x_2 < x_3 < \dots < x_{i-1} < x_i < \dots$ , then  $P(X = x_i) = F(x_i) - F(x_{i-1})$ .
4. If  $X$  is a continuous R V, then  $\frac{d}{dx} F(x) = f(x)$ , at all points where  $F(x)$  is differentiable.

## **Special Distributions**

### **Discrete Distributions**

- 1. Binomial Distribution*
- 2. Poisson Distribution*

### **Continuous Distributions**

- 1. Normal Distribution*
- 2. Exponential Distribution*
- 3. Gamma Distribution*
- 4. Weibull Distribution*

$X - RV$ 

$$\underset{\text{Mean}}{\frac{E(X)}{\sum p_i}} = \begin{cases} \sum_{i=1}^n x_i p_i & \text{if } X \text{ is DRV} \\ \int x f(x) dx & \text{if } X \text{ is CRV} \end{cases}$$

$$E(X^2) = \sum_{i=1}^n x_i^2 p_i ; \text{ if } X \text{ is DRV}$$

$$\int x^2 f(x) dx ; \text{ if } X \text{ is CRV}$$

Let  $g(x)$  be a function on a RV  $X$ .

$$E(g(x)) = \begin{cases} \sum_{i=1}^n g(x_i) p_i & \text{if } X \text{ is DRV} \\ \int g(x) f(x) dx & \text{if } X \text{ is CRV} \end{cases}$$

Suppose  $g(x) = x$  (Identity function)

$$E(g(x)) = E(X)$$

Suppose  $g(x) = x^2$

$$\text{then } E(g(x)) = E(X^2)$$

Properties:

$$E(aX) = a E(X)$$

$$E(ax+by) = a E(X) + b E(Y)$$

$$E(X+Y) = E(X) + E(Y)$$

$$E(1) = 1$$

$$E(a) = a$$

$$\begin{aligned} \text{Variance} &= \frac{1}{N} \sum_{i=1}^N p_i \cdot \left( \sum_{j=1}^N x_j p_j \right)^2 - E(X)^2 \\ &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 p_i = \sum_{i=1}^N (x_i - E(X))^2 p_i \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (E(X))^2 = E(X - E(X))^2 \\ \sigma_X^2 &= E[(X - E(X))^2] \end{aligned}$$

$$SD(X) = \sigma_X = \sqrt{\text{Var}(X)}$$

$$E(g(x)) = \begin{cases} \sum_{i=1}^n g(x_i) \cdot p_i & ; \text{if } X \text{- DRV} \\ \int_{-\infty}^{\infty} g(x) \cdot f(x) dx & ; \text{if } X \text{- CRV} \end{cases}$$

choose  $g(x) = x$

Mean of X

$$E(X) = \begin{cases} \sum_{i=1}^n x_i \cdot p_i & ; \text{if } X \text{- DRV} \\ \int_{-\infty}^{\infty} x \cdot f(x) dx & ; \text{if } X \text{- CRV} \end{cases}$$

choose:  $g(x) = x \cdot x = x^2$

$$E(X^2) = \begin{cases} \sum_{i=1}^n x_i^2 \cdot p_i & ; \text{if } X \text{- DRV} \\ \int_{-\infty}^{\infty} x^2 \cdot f(x) dx & ; \text{if } X \text{- CRV} \end{cases}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (E(X))^2 \\ &= E[(X - E(X))^2] \end{aligned}$$

$$SD = +\sqrt{\text{Var}(X)}$$

## ***Mathematical Expectation of One Dimensional Random Variable***

**Definitions:** If  $X$  is a discrete RV, then *the expected value* or the mean value of  $g(X)$  is defined as

$$E\{g(X)\} = \sum_i g(x_i)p_i,$$

where  $p_i = P(X = x_i)$  is the probability mass function of  $X$ .

If  $X$  is a continuous RV with pdf  $f(x)$ , then

$$E\{g(X)\} = \int_{R_X} g(x)f(x)dx$$

## **Mean, Variance and Standard Deviation of a Random Variable**

Two expected values which are most commonly used for characterising a RV  $X$  are its **mean**  $\mu_X$  and **variance**  $\sigma_X^2$ , which are defined as follows:

$$\mu_X = E(X)$$

$$= \sum_i x_i p_i, \text{ if } X \text{ is discrete}$$

$$= \int_{R_X} x f(x) dx, \text{ if } X \text{ is continuous}$$

$$\text{Var}(X) = \sigma_X^2 = E\{(X - \mu_X)^2\}$$

$$= \sum_i (x_i - \mu_X)^2 p_i, \text{ if } X \text{ is discrete}$$

$$\int_{R_X} (x - \mu_X)^2 f(x) dx, \text{ if } X \text{ is continuous}$$

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2$$

The square root of variance is called **the standard deviation**.

## Properties of Expected Value :

$$1). E(aX) = a E(X)$$

$$2). \text{Var}(aX) = a^2 \text{Var}(X)$$

$$3). E(X+Y) = E(X) + E(Y)$$

4). If  $X$  and  $Y$  are independent RVs, then  $E(XY) = E(X) \cdot E(Y)$

## Example:

If the random variable  $X$  takes the values 1, 2, 3 and 4 such that  $2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$ , find the probability distribution and cumulative distribution function of  $X$ .

## Solution:

Let  $P(X = 3) = 30K$ . Since  $2P(X = 1) = 30K$ ,  $P(X = 1) = 15K$ .

Similarly  $P(X = 2) = 10K$  and  $P(X = 4) = 6K$ .

Since  $\sum p_i = 1$ ,  $15K + 10K + 30K + 6K = 1$ .

$$\therefore K = \frac{1}{61}$$

The probability distribution of  $X$  is given in the following table:

$X = i$	1	2	3	4
$p_i$	$\frac{15}{61}$	$\frac{10}{61}$	$\frac{30}{61}$	$\frac{6}{61}$

The cdf  $F(x)$  is defined as  $F(x) = P(X \leq x)$ . Accordingly the cdf for the above distribution is found out as follows:

When  $x < 1$ ,  $F(x) = 0$

When  $1 \leq x < 2$ ,  $F(x) = P(X = 1) = \frac{15}{61}$

When  $2 \leq x < 3$ ,  $F(x) = P(X = 1) + P(X = 2) = \frac{25}{61}$

When  $3 \leq x < 4$ ,  $F(x) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{55}{61}$

When  $x \geq 4$ ,  $F(x) = P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) = 1$ .

## Example:

A random variable  $X$  has the following probability distribution.

$x:$	-2	-1	0	1	2	3
$p(x):$	0.1	$K$	0.2	$2K$	0.3	$3K$

- (a) Find  $K$ , (b) Evaluate  $P(X < 2)$  and  $P(-2 < X < 2)$ , (c) find the cdf of  $X$  and  
(d) evaluate the mean of  $X$ .

## Solution:

(a) Since  $\sum P(x) = 1$ ,  $6K + 0.6 = 1$

$$\therefore K = \frac{1}{15}$$

$\therefore$  the probability distribution becomes

$x$	-2	-1	0	1	2	3
$p(x)$	$1/10$	$1/15$	$1/5$	$2/15$	$3/10$	$1/5$

$$(b) P(X < 2) = P(X = -2, -1, 0 \text{ or } 1)$$

$$= P(X = -2) + P(X = -1) + P(X = 0) + P(X = 1)$$

[since the events  $(X = -2)$ ,  $(X = -1)$  etc. are mutually exclusive]

$$= \frac{1}{10} + \frac{1}{15} + \frac{1}{5} + \frac{2}{15} = \frac{1}{2}$$

$$P(-2 < X < 2) = P(X = -1, 0 \text{ or } 1)$$

$$= P(X = -1) + P(X = 0) + P(X = 1)$$

$$= \frac{1}{15} + \frac{1}{5} + \frac{2}{15} = \frac{2}{5}$$

## Solution (Continued):

(c)  $F(x) = 0$ , when  $x < -2$

$$= \frac{1}{10}, \text{ when } -2 \leq x < -1$$

$$= \frac{1}{6}, \text{ when } -1 \leq x < 0$$

$$= \frac{11}{30}, \text{ when } 0 \leq x < 1$$

$$= \frac{1}{2}, \text{ when } 1 \leq x < 2$$

$$= \frac{4}{5}, \text{ when } 2 \leq x < 3$$

$$= 1, \text{ when } 3 \leq x$$

(d) The mean of  $X$  is defined as  $E(X) = \Sigma xp(x)$

$$\therefore \text{Mean of } X = \left(-2 \times \frac{1}{10}\right) + \left(-1 \times \frac{1}{15}\right) + \left(0 \times \frac{1}{5}\right)$$

$$+ \left(1 \times \frac{2}{15}\right) + \left(2 \times \frac{3}{10}\right) + \left(3 \times \frac{1}{5}\right)$$

$$= -\frac{1}{5} - \frac{1}{15} + \frac{2}{15} + \frac{3}{5} + \frac{3}{5} = \frac{16}{15}$$

## Exercise:

Find the value of  $a$ ,  $P(X < 3)$ , cumulative distribution function, mean, variance and standard deviation of the discrete random variable ( $X$ ) with the following probability distribution.

$X = x:$	0	1	2	3	4	5	6	7	8
$p = P(X = x):$	$a$	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

## Exercise:

A random variable  $X$  has the following probability distribution.

$x :$	0	1	2	3	4	5	6	7
$p(x) :$	0	$K$	$2K$	$2K$	$3K$	$K^2$	$2K^2$	$7K^2 + K$

Find (i) the value of  $K$ , (ii)  $P(1.5 < X < 4.5 | X > 2)$

## Example:

$$\text{If } p(x) = \begin{cases} x e^{-x^2/2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- show that  $p(x)$  is a pdf (of a continuous RV  $X$ .)
- find its distribution function  $P(x)$ .

**Example:**

$$\text{If } p(x) = \begin{cases} x e^{-x^2/2}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- (a) show that  $p(x)$  is a pdf (of a continuous RV  $X$ ).  
 (b) find its distribution function  $F(x)$ .

a)  
 i)  $I = \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^{\infty} f(x) dx$   
 $I = \int_{-\infty}^0 x \cdot e^{-x^2/2} dx$

Let  $t = x^2 \Rightarrow x = \sqrt{t}$      $\left| \begin{array}{l} x=0 \Rightarrow t=0 \\ x=\infty \Rightarrow t=\infty \end{array} \right.$   
 $dt = 2x dx$

$$I = \frac{1}{2} \int_0^{\infty} e^{-t/2} dt$$

$$I = \frac{1}{2} \left[ \frac{e^{-t/2}}{-\frac{1}{2}} \right]_0^{\infty} = - \left[ 0 - 1 \right] = 1$$

b)  $F(x) = \int_{-\infty}^x f(x) dx$

$$-\infty < x < 0, F(x) = 0$$

$$0 \leq x < \infty, F(x) = \int_0^x f(x) dx$$

$$= \int_0^x x \cdot e^{-x^2/2} dx$$

$$F(x) = 1 - e^{-x^2/2} \quad [t=x^2]$$

$$F(x) = \begin{cases} 0; & -\infty < x < 0 \\ 1 - e^{-x^2/2}; & 0 \leq x < \infty \end{cases}$$

## Solution:

(a) If  $p(x)$  is to be a pdf,  $p(x) \geq 0$  and

$$\int_{R_X} p(x) dx = 1$$

Obviously,  $p(x) = xe^{-x^2/2} \geq 0$ , when  $x \geq 0$

$$\begin{aligned} \text{Now } \int_0^\infty p(x)dx &= \int_0^\infty xe^{-x^2/2} dx = \int_0^\infty e^{-t} dt \text{ (putting } t = x^2/2) \\ &= 1 \end{aligned}$$

$\therefore p(x)$  is a legitimate pdf of a RV  $X$ .

$$F(x) = P(X \leq x) = \int_0^x f(x)dx$$

$\therefore F(x) = 0$ , when  $x < 0$

and  $F(x) = \int_0^x xe^{-t^2/2} dt = 1 - e^{-x^2/2}$ , when  $x \geq 0$ .

## Example:

If the density function of a continuous RV  $X$  is given by

$$\begin{aligned}f(x) &= ax, & 0 \leq x \leq 1 \\&= a, & 1 \leq x \leq 2 \\&= 3a - ax, & 2 \leq x \leq 3 \\&= 0, & \text{elsewhere}\end{aligned}$$

- (i) find the value of  $a$
- (ii) find the cdf of  $X$

## Solution:

(i) Since  $f(x)$  is a pdf,  $\int_{R_x} f(x)dx = 1$ .

i.e.,  $\int_0^3 f(x)dx = 1$

i.e.,  $\int_0^1 axdx + \int_1^2 adx + \int_2^3 (3a - ax)dx = 1$

i.e.,  $2a = 1$

$\therefore a = \frac{1}{2}$

(ii)  $F(x) = P(X \leq x) = 0$ , when  $x < 0$

$$F(x) = \int_0^x \frac{x}{2} dx = \frac{x^4}{4}, \text{ when } 0 \leq x \leq 1$$

$$= \int_0^1 \frac{x}{2} dx + \int_1^x \frac{1}{2} dx = \frac{x}{2} - \frac{1}{4} \text{ when } 1 \leq x \leq 2$$

$$\begin{aligned} &= \int_0^1 \frac{x}{2} dx + \int_1^2 \frac{1}{2} dx + \int_2^x \left(\frac{3}{2} - \frac{x}{2}\right) dx = \frac{3}{2}x - \frac{x^2}{4} - \frac{5}{4}, \text{ when } 2 \leq x \leq 3 \\ &= 1, \text{ when } x > 3 \end{aligned}$$

## Example:

A line of length  $a$  units is divided into two parts. If the first part is of length  $X$ , find  $E(X)$ ,  $\text{var}(X)$  and  $E\{X(a - X)\}$ .

## Solution:

Since the positions of the point of division are equally likely,  $X$  is uniformly distributed in  $(0, a)$ .

$$\therefore f(x) = \frac{1}{a}$$

$$E(X) = \int_0^a xf(x) dx = \frac{1}{a} \int_0^a x dx = \frac{a}{2}$$

$$E(X^2) = \int_0^a x^2 f(x) dx = \frac{a^2}{3}$$

$$\therefore \text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}$$

$$E\{X(a - X)\} = a E(X) - E(X^2) = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}$$

### Exercise:

A continuous RV  $X$  that can assume any value between  $x = 2$  and  $x = 5$  has a density function given by  $f(x) = k(1 + x)$ . Find  $P(X < 4)$ .

### Exercise:

A continuous RV  $X$  has a pdf  $f(x) = kx^2e^{-x}$ ;  $x >$ . Find  $k$ , mean and variance.

### Exercise:

A continuous RV has a pdf  $f(x) = 3x^2$ ;  $0 \leq x \leq 1$ . Find  $a$  and  $b$  such that

- (i)  $P(X \leq a) = P(X > a)$  and
- (ii)  $P(X > b) = 0.05$

## Exercise:

Suppose  $X$  is a continuous random variable with the probability density function given

$$\text{by } f(x) = \begin{cases} kx, & 0 \leq x \leq 2; \\ 2k, & 2 \leq x \leq 4, \\ 6k - kx, & 4 \leq x \leq 6. \end{cases}$$

Find the value of  $k$ , cumulative distribution function, mean, variance and standard deviation of  $X$ .

## Two-Dimensional Random Variables

**Definitions:** Let  $S$  be the sample space associated with a random experiment  $E$ . Let  $X = X(s)$  and  $Y = Y(s)$  be two functions each assigning a real number to each outcomes  $s \in S$ . Then  $(X, Y)$  is called a *two-dimensional random variable*.

If the possible values of  $(X, Y)$  are finite or countably infinite,  $(X, Y)$  is called a *two-dimensional discrete RV*. When  $(X, Y)$  is a two-dimensional discrete RV the possible values of  $(X, Y)$  may be represented as  $(x_i, y_j)$ ,  $i = 1, 2, \dots, m, \dots; j = 1, 2, \dots, n, \dots$ .

If  $(X, Y)$  can assume all values in a specified region  $R$  in the  $xy$ -plane,  $(X, Y)$  is called a *two-dimensional continuous RV*.

## Probability Function of $(X, Y)$

If  $(X, Y)$  is a two-dimensional discrete RV such that  $P(x = x_i, y = y_j) = p_{ij}$ , then  $p_{ij}$  is called the *probability mass function* or simply the *probability function* of  $(X, Y)$  provided the following conditions are satisfied.

(i)  $p_{ij} \geq 0$ , for all  $i$  and  $j$

(ii)  $\sum_j \sum_i p_{ij} = 1$

The set of triples  $\{x_i, y_j, p_{ij}\}$ ,  $i = 1, 2, \dots, m, \dots, j = 1, 2, \dots, n, \dots$ , is called *the joint probability distribution of  $(X, Y)$* .

## Joint Probability Density Function

If  $(X, Y)$  is a two-dimensional continuous RV such that,

$$P\left\{x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2} \text{ and } y - \frac{dy}{2} \leq Y \leq y + \frac{dy}{2}\right\} = f(x, y) dx dy, \text{ then } f(x, y) \text{ is}$$

called *the joint pdf* of  $(X, Y)$ , provided  $f(x, y)$  satisfies the following conditions.

(i)  $f(x, y) \geq 0$ , for all  $(x, y) \in R$ , where  $R$  is the range space.

(ii)  $\iint_R f(x, y) dx dy \equiv 1$ .

Moreover if  $D$  is a subspace of the range space  $R$ ,  $P\{(X, Y) \in D\}$  is defined as

$$P\{(X, Y) \in D\} = \iint_D f(x, y) dx dy. \text{ In particular}$$

$$P\{a \leq X \leq b, c \leq Y \leq d\} = \int_c^d \int_a^b f(x, y) dx dy$$

## Cumulative Distribution Function

If  $(X, Y)$  is a two-dimensional RV (discrete or continuous), then  $F(x, y) = P\{X \leq x \text{ and } Y \leq y\}$  is called *the cdf of  $(X, Y)$* .

In the discrete case,

$$F(x, y) = \sum_j \sum_i p_{ij} \quad y_j \leq y, x_i \leq x$$

In the continuous case,

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy$$

## Properties of $F(x, y)$

- (i)  $F(-\infty, y) = 0 = F(x, -\infty)$  and  $F(\infty, \infty) = 1$
- (ii)  $P\{a < X < b, Y \leq y\} = F(b, y) - F(a, y)$
- (iii)  $P\{X \leq x, c < Y < d\} = F(x, d) - F(x, c)$
- (iv)  $P\{a < X < b, c < Y < d\} = F(b, d) - F(a, d) - F(b, c) + F(a, c)$
- (v) At points of continuity of  $f(x, y)$

$$\frac{\partial^2 F}{\partial x \partial y} = f(x, y)$$

## Marginal Probability Distribution

$$P(X = x_i) = P\{(X = x_i \text{ and } Y = y_1) \text{ or } (X = x_i \text{ and } Y = y_2) \text{ or etc.}\}$$

$$= p_{i1} + p_{i2} + \dots = \sum_j p_{ij}$$

$P(X = x_i) = \sum_j p_{ij}$  is called the *marginal probability function of X*. It is defined

for  $X = x_1, x_2, \dots$  and denoted as  $P_{i*}$ . The collection of pairs  $\{x_i, p_{i*}\}, i = 1, 2, 3, \dots$  is called the *marginal probability distribution of X*.

Similarly the collection of pairs  $\{y_j, p_{*j}\}, j = 1, 2, 3, \dots$  is called the *marginal probability distribution of Y*, where  $p_{*j} = \sum_i P_{ij} = P(Y = y_j)$ .

In the continuous case,

$$P\left\{x - \frac{1}{2}dx \leq X \leq x + \frac{1}{2}dx, -\infty < Y < \infty\right\}$$

$$= \int_{-\infty}^{\infty} \int_{x - \frac{1}{2}dx}^{x + \frac{1}{2}dx} f(x, y) dx dy$$

$$\begin{aligned} &= \left[ \int_{-\infty}^{\infty} f(x, y) dy \right] dx \quad [\text{since } f(x, y) \text{ may be treated a constant in } \\ &\quad (x - 1/2dx, x + 1/2dx)] \\ &= f_X(x)dx, \text{ say} \end{aligned}$$

$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$  is called the *marginal density of X*.

Similarly,  $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$  is called the *marginal density of Y*.

**Note**

$$P(a \leq X \leq b) = P(a \leq X \leq b, -\infty < Y < \infty)$$

$$= \int_{-\infty}^{\infty} \int_a^b f(x, y) dx dy$$

$$= \int_a^b \left[ \int_{-\infty}^{\infty} f(x, y) dy \right] dx = \int_a^b f_X(x) dx$$

Similarly,  $P(c \leq Y \leq d) = \int_c^d f_Y(y) dy$

## Conditional Probability Distribution

$P\{X = x_i / Y = y_j\} = \frac{P\{X = x_i, Y = y_j\}}{P\{Y = y_j\}} = \frac{p_{ij}}{p_{*j}}$  is called *the conditional probability function of X, given that Y = y<sub>j</sub>*.

The collection of pairs,  $\left\{x_i, \frac{p_{ij}}{p_{*j}}\right\}, i = 1, 2, 3, \dots,$

is called *the conditional probability distribution of X, given Y = y<sub>j</sub>*.

Similarly, the collection of pairs,  $\left\{y_j, \frac{p_{ij}}{p_{*i}}\right\}, j = 1, 2, 3, \dots$ , is called the *conditional probability distribution of Y given X = x<sub>i</sub>*. In the continuous case,

$$\begin{aligned} & P\left\{x - \frac{1}{2} dx \leq X \leq x + \frac{1}{2} dx / Y = y\right\} \\ &= P\left\{x - \frac{1}{2} dx \leq X \leq x + \frac{1}{2} dx / y - \frac{1}{2} dy \leq Y \leq y + \frac{1}{2} dy\right\} \\ &= \frac{f(x, y) dx dy}{f_Y(y) dy} = \left\{\frac{f(x, y)}{f_Y(y)}\right\} dx. \end{aligned}$$

$\frac{f(x, y)}{f_Y(y)}$  is called *the conditional density of X, given Y*, and is denoted by  $f(x/y)$ .

Similarly,  $\frac{f(x, y)}{f_X(x)}$  is called *the conditional density of Y, given X*, and is denoted by  $f(y/x)$ .

## **Independent RVs**

If  $(X, Y)$  is a two-dimensional discrete RV such that  $P\{X = x_i | Y = y_j\} = P(X = x_i)$  i.e.,  $\frac{P_{ij}}{P_{*j}} = p_{i*}$ , i.e.,  $P_{ij} = p_{i*} \times p_{*j}$  for all  $i, j$  then  $X$  and  $Y$  are said to be independent RVs.

Similarly if  $(X, Y)$  is a two-dimensional continuous RV such that  $f(x, y) = f_X(x) \times f_Y(y)$ , then  $X$  and  $Y$  are said to be independent RVs.

## **Random Vectors**

**Definitions:** A vector  $X: [X_1, X_2, \dots, X_n]$  whose components  $X_i$  are RVs is called a *random vector*.  $(X_1, X_2, \dots, X_n)$  can assume all values in some region  $R_n$  of the  $n$ -dimensional space.  $R_n$  is called the *range space*.

# Mathematical Expectation of Two Dimensional Random Variable

## Expected Values of a Two-Dimensional RV

If  $(X, Y)$  is a two-dimensional discrete RV with joint probability mass function  $P_{ij}$ , then  $E\{g(X, Y)\} = \sum_j \sum_i g(x_i, y_i)P_{ij}$ .

If  $(X, Y)$  is a two-dimensional continuous RV with joint pdf  $f(x, y)$ , then

$$E\{g(X, Y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y) dx dy$$

## Example:

Three balls are drawn at random without replacement from a box containing 2 white, 3 red and 4 black balls. If  $X$  denotes the number of white balls drawn and  $Y$  denotes the number of red balls drawn, find the joint probability distribution of  $(X, Y)$ .

### Solution:

As there are only 2 white balls in the box,  $X$  can take the values 0, 1 and 2 and  $Y$  can take the values 0, 1, 2 and 3.

$$\begin{aligned} P(X=0, Y=0) &= P(\text{drawing 3 balls none of which is white or red}) \\ &= P(\text{all the 3 balls drawn are black}) \\ &= 4C_3/9C_3 = \frac{1}{21} \end{aligned}$$

$$\begin{aligned} P(X=0, Y=1) &= P(\text{drawing 1 red and 2 black balls}) \\ &= \frac{3C_1 \times 4C_2}{9C_3} = \frac{3}{14} \end{aligned}$$

$$\text{Similarly, } P(X=0, Y=2) = \frac{3C_2 \times 4C_1}{9C_3} = \frac{1}{7}; P(X=0, Y=3) = \frac{1}{84}$$

$$P(X=1, Y=0) = \frac{1}{7}; P(X=1, Y=1) = \frac{2}{7}; P(X=1, Y=2) = \frac{1}{14};$$

$$P(X=1, Y=3) = 0 \text{ (since only 3 balls are drawn)}$$

$$P(X=2, Y=0) = \frac{1}{21}; P(X=2, Y=1) = \frac{1}{28}; P(X=2, Y=2) = 0;$$

$$P(X=2, Y=3) = 0$$

The joint probability distribution of  $(X, Y)$  may be represented in the form of a table as given below:

$X$	$Y$			
	0	1	2	3
0	$\frac{1}{21}$	$\frac{3}{14}$	$\frac{1}{7}$	$\frac{1}{84}$
1	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{1}{14}$	0
2	$\frac{1}{21}$	$\frac{1}{28}$	0	0

## Example:

For the bivariate probability distribution of  $(X, Y)$  given below, find  $P(X \leq 1)$ ,  $P(Y \leq 3)$ ,  $P(X \leq 1, Y \leq 3)$ ,  $P(X \leq 1/Y \leq 3)$ ,  $P(Y \leq 3/X \leq 1)$  and  $P(X + Y \leq 4)$ .

$X \backslash Y$	1	2	3	4	5	6
0	0	0	$1/32$	$2/32$	$2/32$	$3/32$
1	$1/16$	$1/16$	$1/8$	$1/8$	$1/8$	$1/8$
2	$1/32$	$1/32$	$1/64$	$1/64$	0	$2/64$

**Solution:**

$$\begin{aligned}P(X \leq 1) &= P(X = 0) + P(X = 1) \\&= \sum_{j=1}^6 P(X = 0, Y=j) + \sum_{j=1}^6 P(X = 1, Y=j) \\&= \left(0 + 0 + \frac{1}{32} + \frac{2}{32} + \frac{2}{32} + \frac{3}{32}\right) + \left(\frac{1}{16} + \frac{1}{16} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\&= \frac{1}{4} + \frac{5}{8} = \frac{7}{8}\end{aligned}$$

$$\begin{aligned}P(Y \leq 3) &= P(Y = 1) + P(Y = 2) + P(Y = 3) \\&= \sum_{i=0}^2 P(X = i, Y = 1) + \sum_{i=0}^2 P(X = i, Y = 2) \\&\quad + \sum_{i=0}^2 P(X = i, Y = 3) \\&= \left(0 + \frac{1}{16} + \frac{1}{32}\right) + \left(0 + \frac{1}{16} + \frac{1}{32}\right) + \left(\frac{1}{32} + \frac{1}{8} + \frac{1}{64}\right) \\&= \frac{3}{32} + \frac{3}{32} + \frac{11}{64} = \frac{23}{64}\end{aligned}$$

$$P(Y \leq 3/X \leq 1) = \frac{P(X \leq 1, Y \leq 3)}{P(Y \leq 3)} = \frac{9/32}{23/64} = \frac{9}{23}$$

$$P(Y \leq 3/X \leq 1) = \frac{P(X \leq 1, Y \leq 3)}{P(X \leq 1)} = \frac{9/32}{7/8} = \frac{9}{28}$$

$$\begin{aligned}P(X + Y \leq 4) &= \sum_{j=1}^4 P(X = 0, Y=j) + \sum_{j=1}^3 P(X = 1, Y=j) + \sum_{j=1}^2 P(X = 2, Y=j) \\&= \frac{3}{32} + \frac{1}{4} + \frac{1}{16} = \frac{13}{32}\end{aligned}$$

## Example:

The joint probability mass function of  $(X, Y)$  is given by  $p(x, y) = k(2x + 3y)$ ,  $x = 0, 1, 2$ ;  $y = 1, 2, 3$ . Find all the marginal and conditional probability distributions. Also find the probability distribution of  $(X + Y)$ .

## Solution:

The joint probability distribution of  $(X, Y)$  is given below. The relevant probabilities have been computed by using the given law.

X	Y		
	1	2	3
0	$3k$	$6k$	$9k$
1	$5k$	$8k$	$11k$
2	$7k$	$10k$	$13k$

$$\sum_{j=1}^3 \sum_{i=0}^2 p(x_i, y_j) = 1$$

i.e., the sum of all the probabilities in the table is equal to 1.

$$\text{i.e., } 72k = 1.$$

$$\therefore k = \frac{1}{72}$$

### Solution (Continued):

#### Marginal Probability Distribution of $X$ : $\{i, p_{i*}\}$

$X = i$	$p_{i*} = \sum_{j=1}^3 p_{ij}$
0	$p_{01} + p_{02} + p_{03} = \frac{18}{72}$
1	$p_{11} + p_{12} + p_{13} = \frac{24}{72}$
2	$p_{21} + p_{22} + p_{23} = \frac{30}{72}$
Total = 1	

#### Marginal Probability Distribution of $Y$ : $\{j, p_{*j}\}$

$Y = j$	$p_{*j} = \sum_{i=0}^2 p_{ij}$
1	15/27
2	24/72
3	33/72
Total = 1	

## Solution (Continued):

Conditional distribution of  $X$ , given  $Y = 1$ , is given by  $\{i, P(X = i|Y = 1)\} = \{i, P(X = i, Y = 1)/P(Y = 1)\}$   
 $= \{i, p_{i1}/p_{*1}\}, i = 0, 1, 2.$

The tabular representation is given below:

$X = i$	$p_{i1}/p_{*1}$
0	$3k/15k = \frac{1}{5}$
1	$5k/15k = \frac{1}{3}$
2	$7k/15k = \frac{7}{15}$
Total = 1	

The other conditional distributions are given below:

C.P.D. of $X$ , given $Y = 2$	
$X = i$	$p_{i2}/p_{*2}$
0	$\frac{6k}{24k} = \frac{1}{4}$
1	$\frac{8k}{24k} = \frac{1}{3}$
2	$\frac{10k}{24k} = \frac{5}{12}$
	Total = 1

C.P.D. of $X$ , given $Y = 3$	
$X = i$	$p_{i3}/p_{*3}$
0	$\frac{9k}{33k} = \frac{3}{11}$
1	$\frac{11k}{33k} = \frac{1}{3}$
2	$\frac{13k}{33k} = \frac{13}{33}$
	Total = 1

## Solution (Continued):

C.P.D. of  $Y$ , given  $X = 0$

$Y = j$	$p_{0j}/p_{0*}$
1	$\frac{3k}{18k} = \frac{1}{6}$
2	$\frac{6k}{18k} = \frac{1}{3}$
3	$\frac{9k}{18k} = \frac{1}{2}$
	Total = 1

C.P.D. of  $Y$ , given  $X = 1$

$Y = j$	$p_{1j}/p_{1*}$
1	$\frac{5k}{24k} = \frac{5}{24}$
2	$\frac{8k}{24k} = \frac{1}{3}$
3	$\frac{11k}{24k} = \frac{11}{24}$
	Total = 1

C.P.D. of  $Y$ , given  $X = 2$

$Y = j$	$p_{2j}/p_{2*}$
1	$\frac{7k}{30k} = \frac{7}{30}$
2	$\frac{10k}{30k} = \frac{1}{3}$
3	$\frac{13k}{30k} = \frac{13}{30}$
	Total = 1

## Solution (Continued):

Probability distribution of $(X + Y)$	
$(X + Y)$	$P$
1	$p_{01} = \frac{3}{72}$
2	$p_{02} + p_{11} = \frac{11}{72}$
3	$p_{03} + p_{12} + p_{21} = \frac{24}{72}$
4	$p_{13} + p_{22} = \frac{21}{72}$
5	$p_{23} = \frac{13}{72}$
	Total = 1

## Exercise:

If  $X$  denotes the number of aces and  $Y$  denotes the number of queens obtained when 2 cards are randomly drawn without replacement from a standard deck of cards, then obtain the joint probability distribution of  $(X, Y)$ .

## Exercise:

The following table represents the joint probability distribution of the two dimensional discrete random variable  $(X, Y)$ .

Y	X		
	1	2	3
1	1/12	1/6	0
2	0	1/9	1/5
3	1/18	1/4	2/15

Find all the marginal and conditional probability distributions.

## Example:

The joint pdf of a two-dimensional RV  $(X, Y)$  is given by  $f(x, y) = xy^2 + \frac{x^2}{8}$ ,  
 $0 \leq x \leq 2$ ,  $0 \leq y \leq 1$ .

Compute  $P(X > 1)$ ,  $P(Y < \frac{1}{2})$ ,  $P(X > 1 | Y < 1/2)$

$P(Y < \frac{1}{2} | X > 1)$ ,  $P(X < Y)$  and  $P(X + Y \leq 1)$ .

### Solution:

Here the rectangle defined by  $0 \leq x \leq 2$ ,  $0 \leq y \leq 1$  is the range space  $R$ .  $R_1, R_2, \dots$  are event spaces.

$$(i) P(X > 1) = \int \int f(x, y) dx dy$$

$$= \int_0^1 \int_0^2 \left( xy^2 + \frac{x^2}{8} \right) dx dy = \frac{19}{24}$$

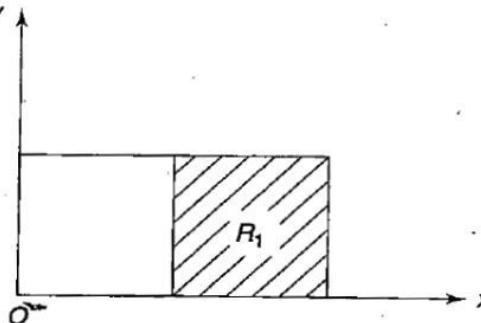


Fig.

$$(ii) P(Y < 1/2) = \int_{R_2} \left( xy^2 + \frac{x^2}{8} \right) dx dy$$

$$= \int_0^{1/2} \int_0^2 \left( xy^2 + \frac{x^2}{8} \right) dx dy$$

$$= \frac{1}{4}$$

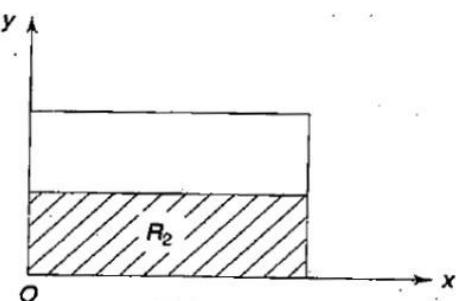
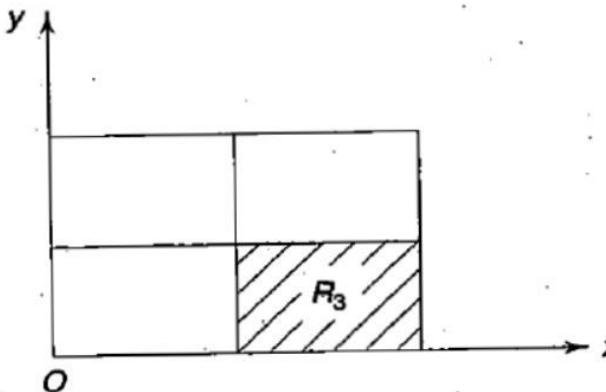


Fig.

## Solution (Continued):

$$\begin{aligned}
 \text{(iii)} \quad P(X > 1, Y < 1/2) &= \int_{R_3} \left( xy^2 + \frac{x^2}{8} \right) dx dy \\
 &\quad \left( x > 1 \& y < \frac{1}{2} \right) \\
 &= \int_0^{1/2} \int_1^2 \left( xy^2 + \frac{x^2}{8} \right) dx dy \\
 &= \frac{5}{24}
 \end{aligned}$$



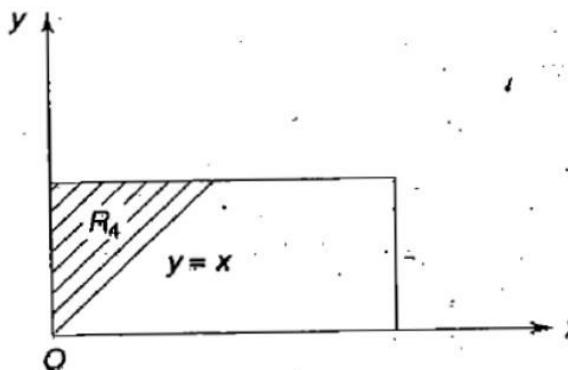
**Fig.**

$$\text{(iv)} \quad P\left(X > 1 / Y < \frac{1}{2}\right) = \frac{P\left(X > 1, Y < \frac{1}{2}\right)}{P\left(Y < \frac{1}{2}\right)} = \frac{5/24}{1/4} = \frac{5}{6}$$

$$\text{(v)} \quad P\left(Y < \frac{1}{2} / X > 1\right) = \frac{P\left(X > 1, Y < \frac{1}{2}\right)}{P(X > 1)} = \frac{5/24}{19/24} = \frac{5}{19}$$

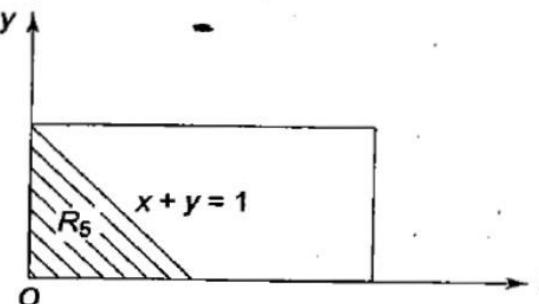
### Solution (Continued):

$$\begin{aligned}
 \text{(vi)} \quad P(X < Y) &= \int_{R_4} \int \left( xy^2 + \frac{x^2}{8} \right) dx dy \\
 &\quad (x < y) \\
 &= \int_0^1 \int_0^y \left( xy^2 + \frac{x^2}{8} \right) dx dy = \frac{53}{480}
 \end{aligned}$$



**Fig.**

$$\begin{aligned}
 \text{(vii)} \quad P(X + Y \leq 1) &= \int_{R_5} \int \left( xy^2 + \frac{x^2}{8} \right) dx dy \\
 &\quad (x + y \leq 1) \\
 &= \int_0^1 \int_0^{1-y} \left( xy^2 + \frac{x^2}{8} \right) dx dy = \frac{13}{480}
 \end{aligned}$$



**Fig.**

## Example:

The joint pdf of the RV  $(X, Y)$  is given by  $f(x, y) = k x y e^{-(x^2 + y^2)}$   $x > 0, y > 0$ .  
Find the value of  $k$  and prove also that  $X$  and  $Y$  are independent.

## Solution:

Here the range space is the entire first quadrant of the  $xy$ -plane.  
By the property of the joint pdf

$$\iint_{x>0, y>0} kxy e^{-(x^2+y^2)} dx dy = 1$$

i.e.,  $k \int_0^\infty ye^{-y^2} dy \int_0^\infty xe^{-x^2} dx = 1$

i.e.,  $\frac{k}{4} = 1$

$\therefore k = 4$

Now  $f_X(x) = \int_0^\infty 4x e^{-x^2} \times ye^{-y^2} dy = 2x e^{-x^2}$ ,  $x > 0$

Similarly,  $f_Y(y) = 2ye^{-y^2}$ ,  $y > 0$ .

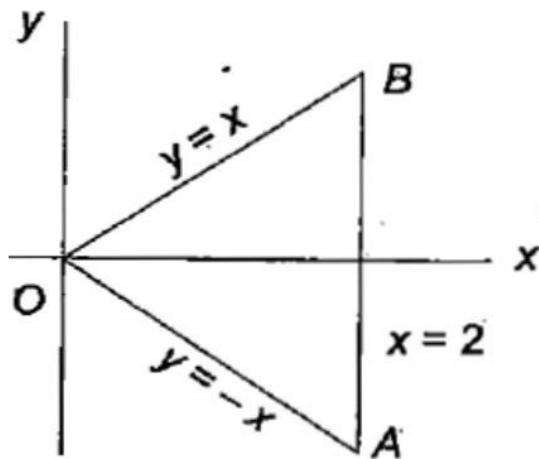
Now  $f_X(x) \times f_Y(y) = 4xy e^{-(x^2+y^2)} = f(x, y)$

$\therefore$  The RVs  $x$  and  $y$  are independent.

## Example:

Given  $f_{XY}(x, y) = cx(x - y)$ ,  $0 < x < 2$ ,  $-x < y < x$ , and 0 elsewhere, (a) evaluate  $c$ , (b) find  $f_X(x)$ , (c)  $f_{Y|X}(y/x)$  and (d)  $f_Y(y)$ .

## Solution:



Here the range space is the area within the triangle  $OAB$  (shown in the figure), defined by  $0 < x < 2$  and  $-x < y < x$ .

(a) By the property of jpdf

$$\int \int_{\Delta OAB} cx(x-y) dx dy = 1$$

$\Delta OAB$

$$\int_0^2 \int_{-x}^x cx(x-y) dy dx = 1$$

$$\text{i.e.,} \quad 8c = 1$$

$$c = \frac{1}{8}$$

## Solution (Continued):

$$(b) f_X(x) = \int_{-x}^x \frac{1}{8} x (x - y) dy \\ = \frac{x^3}{4}, \text{ in } 0 < x < 2$$

$$(c) f(y/x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{2x^2} (x - y), -x < y < x$$

$$(d) f_Y(y) = \int_{-y}^2 \frac{1}{8} x (x - y) dx, \text{ in } -2 \leq y \leq 0 \\ = \int_y^2 \frac{1}{8} x (x - y) dx, \text{ in } 0 \leq y \leq 2$$

$$\text{i.e., } f_Y(y) = \begin{cases} \frac{1}{3} - \frac{y}{4} + \frac{5}{48} y^3, & \text{in } -2 \leq y \leq 0 \\ \frac{1}{3} - \frac{y}{4} + \frac{1}{48} y^3, & \text{in } 0 \leq y \leq 2 \end{cases}$$

## Exercise:

If the joint probability density function of a two-dimensional random variable  $(X, Y)$  is given by

$$f(x, y) = \begin{cases} k(6 - x - y), & 0 < x < 2, 2 < y < 4; \\ 0, & \text{otherwise,} \end{cases}$$

then find (i). the value of  $k$ , (ii).  $P(Y < 3)$ , (iii).  $P(X < 1, Y < 3)$ , (iv).  $P(X + Y < 3)$ , and (v).  $P(X < 1 | Y < 3)$ .

## Exercise:

If the joint density for the random variables  $(X, Y)$ , where  $X$  is the unit temperature change and  $Y$  is the proportion of spectrum shift that a certain atomic particle produces, is given by

$$f(x, y) = \begin{cases} cxy^2, & 0 < x < y < 1; \\ 0, & \text{otherwise,} \end{cases}$$

then find (i). the value of  $c$ , (ii).  $E(XY)$ , (iii).  $f_X(x)$ , (iv).  $f_Y(y)$  and (v).  $f_{Y/X}(y/x)$ .

## Covariance

As the variance  $E\{X - E(X)\}^2$  measures the variations of the R.V.  $X$  from its mean value  $E(X)$ , the quantity  $E\{[X - E(X)][Y - E(Y)]\}$  measures the simultaneous variation of two R.V.'s  $X$  and  $Y$  from their respective means and hence it is called *the covariance of  $X$ ,  $Y$*  and denoted as  $\text{Cov}(X, Y)$ .

$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$  is also called the *product moment* of  $X$  and  $Y$  and is also denoted as  $p(X, Y)$ .

$$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\} = E(XY) - E(X) \cdot E(Y)$$

$$\text{Cov}(X, Y) = \frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i$$

## Example:

Consider the joint probability distribution

		Y		
		-1	0	1
		0	0	1/3
X	0	1/3	0	1/3
	1	1/3	0	1/3

Find the co-variance between  $X$  and  $Y$ .

## Solution:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.$$

## Example:

Let  $X$  and  $Y$  be continuous random variables with joint pdf

$$f_{X,Y}(x, y) = 3x, \quad 0 \leq y \leq x \leq 1,$$

and zero otherwise. Find the co-variance between  $X$  and  $Y$ .

## Solution:

The marginal pdfs, expectations and variances of  $X$  and  $Y$  are

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^x 3xdy = 3x^2, \quad 0 \leq x \leq 1,$$

$$\Rightarrow E_{f_X}[X] = \int_{-\infty}^{\infty} xf_X(x)dx = \int_0^1 x \times 3x^2 dx = \left[ \frac{3}{4}x^4 \right]_0^1 = \frac{3}{4},$$

$$E_{f_X}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 x^2 \times 3x^2 dx = \left[ \frac{3}{5}x^5 \right]_0^1 = \frac{3}{5},$$

$$\Rightarrow Var_{f_X}[X] = E_{f_X}[X^2] - \{E_{f_X}[X]\}^2 = \frac{3}{5} - \left\{ \frac{3}{4} \right\}^2 = \frac{3}{80}.$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_y^1 3xdx = \left[ \frac{3}{2}x^2 \right]_y^1 = \frac{3}{2}(1-y^2), \quad 0 \leq y \leq 1,$$

$$\Rightarrow E_{f_Y}[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_0^1 y \times \frac{3}{2}(1-y^2)dy = \left[ \frac{3}{2} \left( \frac{y^2}{2} - \frac{y^4}{4} \right) \right]_0^1 = \frac{3}{2} \left( \frac{1}{2} - \frac{1}{4} \right) = \frac{3}{8},$$

$$E_{f_Y}[Y^2] = \int_{-\infty}^{\infty} y^2 f_Y(y)dy = \int_0^1 y^2 \times \frac{3}{2}(1-y^2)dy = \left[ \frac{3}{2} \left( \frac{y^3}{3} - \frac{y^5}{5} \right) \right]_0^1 = \frac{3}{2} \left( \frac{1}{3} - \frac{1}{5} \right) = \frac{1}{5},$$

$$\Rightarrow Var_{f_Y}[Y] = E_{f_Y}[Y^2] - \{E_{f_Y}[Y]\}^2 = \frac{1}{5} - \left\{ \frac{3}{8} \right\}^2 = \frac{19}{320},$$

and to compute the covariance we also need to compute  $E_{f_{X,Y}}[XY]$

## Solution (Continued):

$$\begin{aligned} E_{f_{X,Y}}[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dy dx = \int_0^1 \int_0^x xy \times 3x dy dx \\ &= \int_0^1 \left\{ \int_0^x y dy \right\} 3x^2 dx = \int_0^1 \left[ \frac{y^2}{2} \right]_0^x 3x^2 dx = \int_0^1 \frac{x^2}{2} \times 3x^2 dx \\ &= \frac{3}{2} \left[ \frac{x^5}{5} \right]_0^1 = \frac{3}{10}, \end{aligned}$$

$$\Rightarrow Cov_{f_{X,Y}}[X, Y] = E_{f_{X,Y}}[XY] - E_{f_X}[X] E_{f_Y}[Y] = \frac{3}{10} - \frac{3}{4} \times \frac{3}{8} = \frac{3}{160}$$

## Exercise:

Let  $X$  and  $Y$  be discrete random variables with joint mass function defined by

$$f_{X,Y}(x,y) = \frac{1}{4}, \quad (x,y) \in \{(0,0), (1,1), (1,-1), (2,0)\},$$

and zero otherwise. Find the co-variance between  $X$  and  $Y$ .

## Exercise:

Two random variables  $X$  and  $Y$  have the following joint probability density function :

$$f(x, y) = \begin{cases} 2 - x - y & ; 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & , \text{ otherwise} \end{cases}$$

Find

- (i) Marginal probability density functions of  $X$  and  $Y$ .
- (ii) Conditional density functions.
- (iii)  $\text{Var}(X)$  and  $\text{Var}(Y)$ .
- (iv) Co-variance between  $X$  and  $Y$ .

# Moment Generating Function (MGF)

Moment Generating Function (MGF) of a RV  $X$  (discrete or continuous) is defined as  $E(e^{tX})$ , where  $t$  is a real variable and denoted as  $M(t)$ .

If  $X$  is discrete, then  $M(t) = \sum_r e^{tx_r} p_r$ ,

where  $X$  takes the values  $x_1, x_2, x_3, \dots$ , with probabilities  $p_1, p_2, p_3, \dots$

If  $X$  is a continuous RV with density function  $f(x)$ , then

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

## Properties of MGF

(Proofs of the properties are omitted, as the proofs of the corresponding properties of characteristic function will be given later.)

$$1. M(t) = \sum_{n=0}^{\infty} t^n E(X^n)/n!$$

i.e.,  $E(X^n) = \mu'_n$  is the co-efficient of  $\frac{t^n}{n!}$  in the expansion of  $M(t)$  in series of powers of  $t$ .

$$2. \mu'_n = E(X^n) = \left[ \frac{d^n}{dt^n} M(t) \right]_{t=0}$$

3. If the MGF of  $X$  is  $M_X(t)$  and if  $Y = aX + b$ , then  $M_Y(t) = e^{bt}M_X(at)$ .

4. If  $X$  and  $Y$  are independent RVs and  $Z = X + Y$ , then  $M_Z(t) = M_X(t)M_Y(t)$ .

## Example:

If  $X$  represents the outcome, when a fair die is tossed, find the MGF of  $X$  and hence find  $E(X)$  and  $\text{Var}(X)$ .

## Solution:

The probability distribution of  $X$  is given by

$$p_i = p(X = i) = \frac{1}{6}, i = 1, 2, \dots, 6$$

$$\begin{aligned}M(t) &= \sum_i e^{tx_i} p_i = \sum_{i=1}^6 e^{ti} p_i \\&= \frac{1}{6} (e^t + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t})\end{aligned}$$

$$E(X) = [M'(t)]_{t=0} = \frac{7}{2}$$

$$E(X^2) = [M''(t)]_{t=0}$$

$$= \frac{1}{6} [e^t + 4e^{2t} + 9e^{3t} + 16e^{4t} + 25e^{5t} + 36e^{6t}]_{t=0} = \frac{91}{6}$$

$$\text{Var}(x) = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

## Example:

If a RV  $X$  has the MGF  $M(t) = \frac{3}{3-t}$ , obtain the standard deviation of  $X$ .

**Solution:**

$$M(t) = \frac{3}{3\left(1 - \frac{t}{3}\right)} = 1 + t/3 + t^2/9 + \dots + \infty$$

$$E(X) = \text{coefficient of } \frac{t}{[1]} \text{ in (1)} = \frac{1}{3}$$

$$E(X^2) = \text{coefficient of } \frac{t^2}{[2]} \text{ in (1)} = \frac{2}{9}$$

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{1}{9}$$

$$\sigma_X = \frac{1}{3}$$

## Example:

If  $x = 1, 2, 3, \dots$  has the geometric distribution  $f(x) = pq^{x-1}$ , where  $q = 1 - p$ , show that the moment generating function is

$$M(t) = \frac{pe^t}{1 - qe^t}.$$

Find  $E(x)$ .

## Solution:

The moment generating function of  $x$  is

$$\begin{aligned} M(t) &= \sum_{x=1}^{\infty} e^{xt} pq^{x-1} = \frac{p}{q} \sum_{x=1}^{\infty} (qe^t)^x \\ &= pe^t \sum_{x=0}^{\infty} (qe^t)^x = \frac{pe^t}{1 - qe^t}. \end{aligned}$$

To find  $E(x)$ , we may use the quotient rule to differentiate the expression  $M(t)$  with respect to  $t$ . This gives

$$\frac{dM(t)}{dt} = \frac{(1 - qe^t)pe^t - pe^t(-qe^t)}{(1 - qe^t)^2}.$$

Setting  $t = 0$  gives  $E(x) = 1/p$ .

## Example:

Find the MGF of the binomial distribution and hence find its mean and variance.

Binomial distribution is given by

$$p_r = p(X = r) = nC_r p^r q^{n-r}, \quad r = 0, 1, 2, \dots, n$$

$$p + q = 1$$

## Solution:

$$M(t) = \sum_{r=0}^n e^{tr} p_r$$

$$= \sum_{r=0}^n e^{tr} n C_r p^r q^{n-r}$$

$$= \sum_{r=0}^n n C_r (p e^t)^r q^{n-r}$$

$$= (p e^t + q)^n$$

$$M'(t) = n(p e^t + q)^{n-1} \times p e^t$$

$$M''(t) = np[(p e^t + q)^{n-1} \times e^t + (n-1)(p e^t + q)^{n-2} p e^{2t}]$$

$$E(X) = M'(0) = np \quad (\text{since } p + q = 1)$$

$$E(X^2) = M''(0) = np [1 + (n-1)p]$$

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2$$

$$= np - np^2$$

$$= npq$$

## **Exercise:**

Find the MGF of a RV which is uniformly distributed over  $(-1, 2)$  and hence find its mean and variance.

# Characteristic Function (CF)

**Characteristic function** of a RV  $X$  (discrete or continuous) is defined as  $E(e^{i\omega X})$  and denoted as  $\phi(\omega)$ .

If  $X$  is a discrete RV that can take the values  $x_1, x_2, \dots$ , such that  $P(X = x_r) = p_r$ , then

$$\phi(\omega) = \sum_r e^{i\omega x_r} p_r$$

If  $X$  is a continuous RV with density function  $f(x)$ , then

$$\phi(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x) dx$$

## Properties of Characteristic Function

1.  $\mu'_n = E(X^n) =$  the coefficient of  $\frac{i^n \omega^n}{n!}$  in the expansion of  $\phi(\omega)$  in series of ascending powers of  $i\omega$ .

2. 
$$\mu'_n = \frac{1}{i^n} \left[ \frac{d^n}{d\omega^n} \phi(\omega) \right]_{\omega=0}$$

3. If the characteristic function of a RV  $X$  is  $f_X(\omega)$  and if  $Y = aX + b$ , then

$$\phi_Y(\omega) = e^{ib\omega} \phi_X(a\omega)$$

4. If  $X$  and  $Y$  are independent RVs, then

$$\phi_{X+Y}(\omega) = \phi_X(\omega) \times \phi_Y(\omega)$$

## Example:

Find the characteristic function of the geometric distribution given by  $P(X = r) = q^r p$ ,  $r = 0, 1, 2, \dots, \infty$ ,  $p + q = 1$ .

Hence find the mean and variance.

**Solution:**

$$\phi(\omega) = \sum_{r=0}^{\infty} e^{i\omega r} pq^r$$

$$= p \sum_{r=0}^{\infty} (qe^{i\omega})^r = p(1 - qe^{i\omega})^{-1}$$

$$\phi^{(1)}(\omega) = p(1 - qe^{i\omega})^{-2} i q e^{i\omega}$$

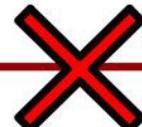
$$\phi^{(2)}(\omega) = i^2 pq [2(1 - q e^{i\omega})^{-3} q e^{i2\omega} + (1 - q e^{i\omega})^{-2} e^{i\omega}]$$

$$E(X) = \frac{1}{i} \phi^{(1)}(0) = \frac{q}{p} \text{ and } E(X^2) = \frac{1}{i^2} \phi^{(2)}(0) = \frac{q}{p^2} (1 + q)$$

$$\mu_X = \frac{q}{p} \text{ and } \sigma_X^2 = \frac{q}{p^2}$$

## Exercise:

Find the CF of  $X$  whose pdf is given by  $f(x) = \frac{1}{2} e^{-|x|}$ ,  $-\infty < x < \infty$ .



## **Module-3**

### **Correlation and Regression**

In this Module, we study the relationship between the variables. Also, the interest lies in establishing the actual relationship between two or more variables. This problem is dealt with regression. On the other hand, we are often not interested to know the actual relationship but are only interested in knowing the degree of relationship between two or more variables. This problem is dealt with correlation analysis.

Linear relationship between two variables is represented by a straight line which is known as regression line. In the study of linear relationship between two variables  $X$  and  $Y$ , suppose the variable  $Y$  is such that it depends on  $X$ , then we call it as the regression line of  $Y$  on  $X$ . If  $X$  depends on  $Y$ , then it is called as the regression line of  $X$  on  $Y$ .

To find out the regression line, the observations  $(x_i, y_i)$  on the variable  $X$  and  $Y$  are necessarily taken in pairs. For example, a chemical engineer may run a chemical process several times in order to study the relationship between the concentration of a certain catalyst and the yield of the process. Each time the process is run, the concentration  $X$  and the yield  $Y$  are recorded. Generally, the studies are based on samples of size ' $n$ ' and hence ' $n$ ' pairs of sample observations can be written as  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

### **Correlation**

In a bivariate distribution, we are interested to find out whether there is any relationship between two variables. The correlation is a statistical technique which studies the relationship between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between the two variables. When two variables are related in such a way that a change in the value of one is accompanied either by a direct change or by an inverse change in the values of the other, the two variables are said to be correlated. In the correlated variables an increase in one variable is accompanied by an increase or decrease in the other

variable. For instance, relationship exists between the price and demand of a commodity because keeping other things equal, an increase in the price of a commodity shall cause a decrease in the demand for that commodity. Relationship might exist between the heights and weights of the students and between amount of rainfall in a city and the sales of raincoats in that city.

### **Utility of Correlation**

The study of correlation is very useful in practical life as revealed by these points.

1. With the help of correlation analysis, we can measure in one figure, the degree of relationship existing between variables like price, demand, supply, income, expenditure etc. Once we know that two variables are correlated then we can easily estimate the value of one variable, given the value of other.
2. Correlation analysis is of great use to economists and businessmen; it reveals to the economists the disturbing factors and suggests to him the stabilizing forces. In business, it enables the executive to estimate costs, sales etc. and plan accordingly.
3. Correlation analysis is helpful to scientists. Nature has been found to be a multiplicity of interrelated forces.

### **Types of Correlation**

Correlation can be categorized as one of the following:

- (i) Positive and Negative,
- (ii) Simple and Multiple.
- (iii) Partial and Total.
- (iv) Linear and Non-Linear (Curvilinear)

**(i) Positive and Negative Correlation :** Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is

accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.

**(ii) Simple and Multiple Correlation :** Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlation, the relationship is studied among three or more variables. For example, the relationship of yield of wheat may be studied with both chemical fertilizers and the pesticides.

**(iii) Partial and Total Correlation :** There are two categories of multiple correlation analysis. Under partial correlation, the relationship of two or more variables is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant. For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. Total correlation is based upon all the variables.

**(iv) Linear and Non-Linear Correlation:** When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

## **Methods of Studying Correlation**

There are different methods which helps us to find out whether the variables are related or not.

1. Scatter Diagram Method.
2. Graphic Method.
3. Karl Pearson's Coefficient of correlation.
4. Rank Method.

### **Karl Pearson's Co-efficient of Correlation.**

Karl Pearson's method, popularly known as Pearsonian co-efficient of correlation, is most widely applied in practice to measure correlation. The Pearsonian co-efficient of correlation is represented by the symbol  $r$ . Degree of correlation varies between + 1 and -1; the result will be + 1 in case of perfect positive correlation and -1 in case of perfect negative correlation. Computation of correlation coefficient can be simplified by dividing the given data by a common factor. In such a case, the final result is not multiplied by the common factor because coefficient of correlation is independent of change of scale and origin.

$$r(X, Y) = \rho(X, Y) = \frac{Cov(x, Y)}{\sigma_X \cdot \sigma_Y}$$
$$Cov(X, Y) = \frac{1}{n} \sum XY - \bar{X}\bar{Y}$$
$$\sigma_X = \sqrt{\frac{1}{n} \sum X^2 - \bar{X}^2}, \quad \sigma_Y = \sqrt{\frac{1}{n} \sum Y^2 - \bar{Y}^2}$$

n - number of items in the given data

## **Standard Error**

The standard error is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample represents a population.

In statistics, a sample means deviates from the actual mean of a population; this deviation is the standard error.

$$S.E(r) = \frac{1 - r^2}{\sqrt{n}}$$

Probable Error=  $P.E(r) = 0.675 \times S.E(r)$

Range:

$$r - S.E(r) \leq Population \leq r + S.E(r)$$

**Note:** Two independent variables are uncorrelated when  $\text{Cov}(X,Y) = 0$

### Problems:

1. Find the correlation coefficient between annual advertising expenditures and annual sales revenue for the following data:

Year ( $i$ )	1	2	3	4	5	6	7	8	9	10
Annual advertising expenditure ( $X_i$ )	10	12	14	16	18	20	22	24	26	28
Annual sales ( $Y_i$ )	20	30	37	50	56	78	89	100	120	110

Solution: Now,  $\bar{X} = \frac{\sum X}{n} = \frac{190}{10} = 19$ ,  $\bar{Y} = \frac{\sum Y}{n} = \frac{690}{10} = 69$

$i$	$X_i$	$Y_i$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	10	20	-9	-49	81	2401	441
2	12	30	-7	-39	49	1521	273
3	14	37	-5	-32	25	1024	160
4	16	50	-3	-19	9	364	57
5	18	56	-1	-13	1	169	13
6	20	78	1	9	1	81	9
7	22	89	3	20	9	400	60
8	24	100	5	31	25	961	155
9	26	120	7	51	49	2601	357
10	28	110	9	41	81	1681	369
	<b>190</b>	<b>690</b>	<b>0</b>	<b>0</b>	<b>330</b>	<b>11200</b>	<b>1894</b>

Correlation coefficient is  $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{1894}{\sqrt{330} \sqrt{11200}} = 0.985$

The correlation coefficient between annual expenditure and annual sales revenue is 0.985.

2. Let X, Y and Z be uncorrelated random variables with zero means and standard deviations 5, 12 and 9 respectively. If U = X + Y and V = Y + Z, find the correlation coefficient between U and V.

**Solution:**

Given that all the three random variables have zero mean.

Hence,  $E(X) = E(Y) = E(Z) = 0$ .

Now,  $\text{Var}(X) = E(X^2) - [E(X)]^2$

$$\Rightarrow E(X^2) = \text{Var}(X) \quad \{ \text{since, } E(X) = 0 \}$$

$$= 5^2 = 25$$

Similarly,  $E(Y^2) = 12^2 = 144$  and  $E(Z^2) = 9^2 = 81$

Since X and Y are uncorrelated we have  $\text{Cov}(X, Y) = 0$

$$\Rightarrow E(XY) = E(X).E(Y) = 0$$

Similarly,  $E(YZ) = 0$  and  $E(ZX) = 0$ .

To find  $\rho(U, V)$ :

$$\text{Now, } \rho(U, V) = \frac{E(UV) - E(U).E(V)}{\sigma_U \cdot \sigma_V}$$

$$E(U) = E[X + Y] = E[X] + E[Y] = 0$$

$$E(V) = E[Y + Z] = E[Y] + E[Z] = 0$$

$$\begin{aligned} E(U^2) &= E[(X + Y)^2] = E[X^2] + E[Y^2] + 2E[XY] \\ &= 25 + 144 + 0 \\ &= 169 \end{aligned}$$

Similarly,  $E(V^2) = 225$

$$\text{Now, } \text{Var}(U) = E(U^2) - [E(U)]^2 = 169$$

$$\Rightarrow \sigma_U = \sqrt{169} = 13$$

$$\text{Similarly, } \text{Var}(V) = E(V^2) - [E(V)]^2 = 225$$

$$\Rightarrow \sigma_V = \sqrt{225} = 15$$

$$\begin{aligned}
E(UV) &= E[(X+Y)(Y+Z)] \\
&= E(XY) + E(Y^2) + E(XZ) + E(YZ) \\
&= 144
\end{aligned}$$

$$\text{Therefore, } \rho(U, V) = \frac{E(UV) - E(U)E(V)}{\sigma_U \cdot \sigma_V} = \frac{144}{195} = \frac{48}{65}$$

3. If the joint pdf of  $(X, Y)$  is given by  $f(x, y) = x + y$ ,  $0 \leq x, y \leq 1$ . Find  $\rho_{XY}$ .

Solution:

$$\text{We know that, } \rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \cdot \sigma_Y}$$

$$\begin{aligned}
\text{Now, } E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy \\
&= \int_0^1 \int_0^1 xy(x + y) dx dy \\
&= \int_0^1 \left[ \frac{x^3 y}{3} + \frac{x^2 y^2}{2} \right]_0^1 dy \\
&= \int_0^1 \left[ \frac{y}{3} + \frac{y^2}{2} \right] dy \\
&= \left[ \frac{y^2}{6} + \frac{y^3}{6} \right]_0^1 \\
&= \frac{1}{3}
\end{aligned}$$

The pdf of  $X$  and  $Y$  is given by

$$\begin{aligned}
f(x) &= \int_0^1 f(x, y) dy = \int_0^1 (x + y) dy = \left[ xy + \frac{y^2}{2} \right]_0^1 = x + \frac{1}{2} \\
f(y) &= \int_0^1 f(x, y) dx = \int_0^1 (x + y) dx = \left[ \frac{x^2}{2} + xy \right]_0^1 = y + \frac{1}{2} \\
E(X) &= \int_0^1 x f(x) dx = \int_0^1 x \left( x + \frac{1}{2} \right) dx = \left[ \frac{x^3}{3} + \frac{x^2}{4} \right]_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \\
E(Y) &= \int_0^1 y f(y) dy = \int_0^1 y \left( y + \frac{1}{2} \right) dy = \left[ \frac{y^3}{3} + \frac{y^2}{4} \right]_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \\
E(X^2) &= \int_0^1 x^2 f(x) dx = \int_0^1 x^2 \left( x + \frac{1}{2} \right) dx = \left[ \frac{x^4}{4} + \frac{x^3}{6} \right]_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}
\end{aligned}$$

$$E(Y^2) = \int_0^1 y^2 f(y) dy = \int_0^1 y^2 \left( y + \frac{1}{2} \right) dy = \left[ \frac{y^4}{4} + \frac{y^3}{6} \right]_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{5}{12} + \left( \frac{7}{12} \right)^2 = \frac{11}{144}$$

$$\Rightarrow \sigma_X = \frac{\sqrt{11}}{12}$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{5}{12} + \left( \frac{7}{12} \right)^2 = \frac{11}{144}$$

$$\Rightarrow \sigma_Y = \frac{\sqrt{11}}{12}$$

$$\text{Therefore, } \rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \cdot \sigma_Y} = \frac{\frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12}}{\frac{\sqrt{11}}{12} \cdot \frac{\sqrt{11}}{12}} = \frac{-1}{11}$$

4. The independent random variables X and Y have the pdf given by  $f_X(x) = \begin{cases} 4ax & , 0 \leq x \leq 1 \\ 0 & , \text{otherwise} \end{cases}$ ,  $f_Y(y) = \begin{cases} 4by & , 0 \leq y \leq 1 \\ 0 & , \text{otherwise} \end{cases}$
- Find the correlation coefficient.

**Solution:**

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx = \int_0^1 x 4ax dx = 4a \int_0^1 x^2 dx = 4a \left[ \frac{x^3}{3} \right]_0^1 = \frac{4a}{3}$$

$$E(Y) = \int_{-\infty}^{\infty} yf(y) dy = \int_0^1 y 4by dy = 4b \int_0^1 y^2 dy = 4b \left[ \frac{y^3}{3} \right]_0^1 = \frac{4b}{3}$$

Since X and Y are independent, the joint pdf of X and Y is given by  $f(x, y) = f(x) \cdot f(y)$

$$= (4ax)(4by)$$

$$= 16abxy, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1$$

$$\text{Now, } E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dxdy \\ = \int_0^1 \int_0^1 xy(16abxy)dxdy = \frac{16ab}{9}$$

$$\text{Therefore we get, } \text{Cov}(X,Y) = E(XY) - E(X)E(Y) \\ = \frac{16ab}{9} - \left(\frac{4a}{3}\right)\left(\frac{4b}{3}\right) = 0$$

Which implies that the  $\text{cor}(X,Y)=0$

That is, the variables X and Y are independent and there is no relationship between them.

### **SPEARMAN'S RANK CORRELATION COEFFICIENT**

Rank correlation coefficient is useful for finding correlation between any two qualitative characteristics such as Beauty, Honesty, and Intelligence etc., which cannot be measured quantitatively but can be arranged serially in order of merit or proficiency possessing the two characteristics.

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient  $r$  is given by

$$\rho = 1 - \left[ \frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where,

$\sum d^2$  = Sum of squares of differences of ranks between paired items in two series

$n$  = Number of paired items

### **Remarks**

Spearman's rank correlation coefficient can be used to find the correlation between two quantitative characteristics or variables. In this case, we associate the ranks to the observations based on their magnitudes for X and Y series separately. Let  $R_X$  and  $R_Y$  be the ranks of observations on two variables X and

Y respectively for a pair. Then the Spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[ \frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where,  $\sum d^2 = (R_x - R_y)^2$  = sum of squares of differences between the ranks of variables X and Y

n = number of pairs of observations

### **SPEARMAN'S RANK CORRELATION COFFICIENT FOR A DATA WITH TIED OBSERVATIONS**

In any series, if two or more observations are having same values then the observations are said to be tied observations. If tie occurs for two or more observations in a series, then common ranks have to be given to the tied observations in that series; these common ranks are the average of the ranks, which these observations would have assumed if they were slightly different from each other and the next observation will get the rank next to the rank already assumed.

In the case of data with tied observations, the Spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[ \frac{6(Adj \sum d^2)}{n(n^2 - 1)} \right]$$

Where,

$$Adj \sum d^2 = \sum d^2 + \left[ \frac{S_1^3 - S_1}{12} \right] + \left[ \frac{S_2^3 - S_2}{12} \right] + \left[ \frac{S_3^3 - S_3}{12} \right] + \dots$$

Here,

$S_1$  is the number of times first tied observation is repeated

$S_2$  is the number of times second tied observation is repeated

$S_3$  is the number of times third observation is repeated etc.

**Problem:** In a quantitative aptitude test, two judges rank the ten competitors in the following order.

Competitor	1	2	3	4	5	6	7	8	9	10
Ranking of judge I	4	5	2	7	8	1	6	9	3	10
Ranking of judge II	8	3	9	10	6	7	2	5	1	4

Is there any concordance between the two judges ?

**Solution:** Let Rx: Ranking by Judge I and Ry: Ranking by Judge II The Spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[ \frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where,  $\sum d^2 = (R_x - R_y)^2$  and n = number of competitors.

R <sub>x</sub>	R <sub>y</sub>	d = R <sub>x</sub> - R <sub>y</sub>	d <sup>2</sup>
4	8	-4	16
5	3	2	4
2	9	-7	49
7	10	-3	9
8	6	2	4
1	7	-6	36
6	2	4	16
9	5	4	16
3	1	2	4
10	4	6	36
		TOT	190

$$\rho = 1 - \left[ \frac{6(190)}{10(100-1)} \right]$$

$$= 1 - 1.1515$$

$$= -0.1515$$

We say that there is low degree of negative rank correlation between the two judges.

**Problem :** Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below:

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test Score	48	55	45	60	43	80	58	50	77	46	47	65

calculate rank correlation coefficient and comment on your result

**Solution:** Let selection test score be a variable X and proficiency test score be a variable Y. We associate the ranks to the scores based on their magnitudes. The spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[ \frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where,  $\sum d^2 = (R_x - R_y)^2$  = sum of squares of differences between the ranks of observations X and Y

n = number of recruits.

Given,

X	Y	R <sub>x</sub>	R <sub>y</sub>	d = R <sub>x</sub> - R <sub>y</sub>	d <sup>2</sup>
44	48	12	8	4	16
49	55	10	6	4	16
52	45	8	11	-3	9
54	60	7	4	3	9

47	43	11	12	-1	1
76	80	1	1	0	0
65	58	3	5	-2	4
60	50	5	7	-2	4
63	77	4	2	2	4
58	46	6	10	-4	16
50	47	9	9	0	0
67	65	2	3	-1	1

From the table, we have,

$$\sum d^2 = 80, n = 12$$

$$\begin{aligned}\rho &= 1 - \left[ \frac{6(80)}{12(144 - 1)} \right] \\ &= 1 - 0.2797 \\ &= 0.7203\end{aligned}$$

We say that there is high degree of positive rank correlation between the scores of selection and proficiency tests.

**Example:**

Following is the data on heights and weights of ten students in a class:

Heights (in cm)	140	142	140	160	150	155	160	157	140	170
Weights (in cm)	43	45	42	50	45	52	57	48	49	53

Calculate rank correlation coefficient between heights and weights of students.

**Solution:**

Let height be a variable X and weight be a variable Y. Since, the data contains tied observations, we associate average ranks to the tied observations. The spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[ \frac{6(Adj \sum d^2)}{n(n^2 - 1)} \right]$$

Where,

$$Adj \sum d^2 = \sum d^2 + \left[ \frac{S_1^3 - S_1}{12} \right] + \left[ \frac{S_2^3 - S_2}{12} \right] + \left[ \frac{S_3^3 - S_3}{12} \right] + \dots$$

N= No. of students

X	Y	R <sub>x</sub>	R <sub>y</sub>	d= R <sub>x</sub> - R <sub>y</sub>	d <sup>2</sup>
140	43	9	9	0	0
142	45	7	7.5	-0.5	0.25
140	42	9	10	-1	1
160	50	2.5	4	-1.5	2.25
150	45	6	7.5	-1.5	2.25
155	52	5	3	2	4
160	57	2.5	1	1.5	2.25
157	48	4	6	-2	4
140	49	9	5	4	16
170	53	1	2	-1	1
				TOT	33

From the table, we have,

$$n = 10, \sum d^2 = 33, S_1 = 3, S_2 = 2, S_3 = 33$$

Thus,

$$Adj \sum d^2 = \sum d^2 + \left[ \frac{S_1^3 - S_1}{12} \right] + \left[ \frac{S_2^3 - S_2}{12} \right] + \left[ \frac{S_3^3 - S_3}{12} \right] + \dots$$

$$Adj \sum d^2 = 33 + \left[ \frac{3^3 - 3}{12} \right] + \left[ \frac{2^3 - 2}{12} \right] + \left[ \frac{2^3 - 2}{12} \right] + \dots$$

$$= 33 + 2 + 0.5 + 0.5$$

$$= 36$$

$$\rho = 1 - \left[ \frac{6(36)}{10(100-1)} \right]$$

$$\rho = 1 - 0.2182$$

$$= 0.7818$$

We say that there is high degree of positive rank correlation between heights and weights of students.

### **Partial and Multiple Correlation**

Let us consider the example of yield of rice in a firm. It may be affected by the type of soil, temperature, amount of rainfall, usage of fertilizers etc. It will be useful to determine how yield of rice is influenced by one factor or how yield of rice is affected by several other factors. This is done with the help of partial and multiple correlation analysis.

The basic distinction between multiple and partial correlation analysis is that in the former, the degree of relationship between the variable  $Y$  and all the other variables  $X_1, X_2, \dots, X_n$  taken together is measured, whereas, in the later, the degree of relationship between  $Y$  and one of the variables  $X_1, X_2, \dots, X_n$  is measured by removing the effect of all the other variables.

### **Partial correlation**

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variable, with the effect of the rest of the variables eliminated. If there are three variables  $X_1, X_2$  and  $X_3$ , there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by  $r_{12.3}$ , that is, the coefficient of partial correlation  $X_1$  and  $X_2$  keeping  $X_3$  constant, it is calculated as

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}, \quad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}},$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

1. In a trivariate distribution, it is found that  $r_{12} = 0.7$ ,  $r_{13} = 0.61$  and  $r_{23} = 0.4$ . Find the partial correlation coefficients.

**Solution:**

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.7 - (0.61 \times 0.4)}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}} = 0.628$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \frac{0.61 - (0.7 \times 0.4)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.4)^2}} = 0.504$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} = \frac{0.4 - (0.7 \times 0.61)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.61)^2}} = -0.048$$

### Multiple Correlation

In multiple correlation, we are trying to make estimates of the value of one of the variable based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables.

The coefficient of multiple correlation with three variables  $X_1, X_2$  and  $X_3$  are  $R_{1.23}$ ,  $R_{2.13}$  and  $R_{3.21}$ .  $R_{1.23}$  is the coefficient of multiple correlation related to  $X_1$  as a dependent variable and  $X_2, X_3$  as two independent variables and it can be expressed in terms of  $r_{12}, r_{23}$  and  $r_{13}$  as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}},$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}},$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}$$

## **PROPERTIES OF MULTIPLE CORRELATION COEFFICIENT**

The following are some of the properties of multiple correlation coefficients:

1. Multiple correlation coefficient is the degree of association between observed value of the dependent variable and its estimate obtained by multiple regression,
2. Multiple Correlation coefficient lies between 0 and 1.
3. If multiple correlation coefficient is 1, then association is perfect and multiple regression equation may said to be perfect prediction formula.
4. If multiple correlation coefficient is 0, dependent variable is uncorrelated with other independent variables. From this, it can be concluded that multiple regression equation fails to predict the value of dependent variable when values of independent variables are known.
5. Multiple correlation coefficient is always greater or equal than any total correlation coefficient. If  $R_{1.23}$  is the multiple correlation coefficient than  $R_{1.23} \geq r_{12}$  or  $r_{13}$  or  $r_{23}$  and
6. Multiple correlation coefficient obtained by method of least squares would always be greater than the multiple correlation coefficient obtained by any other method.

**Example:**

1. The following zero-order correlation coefficients are given:  
 $r_{12} = 0.98$ ,  $r_{13} = 0.44$  and  $r_{23} = 0.54$ . Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

**Solution:**

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(0.54)(0.44)}{1 - (0.54)^2}} = 0.986$$

2. From the following data, obtain  $R_{1.23}$ ,  $R_{2.13}$  and  $R_{3.12}$

X <sub>1</sub>	2	5	7	11
X <sub>2</sub>	3	6	10	12
X <sub>3</sub>	1	3	6	10

**Solution:**

We need  $r_{12}$ ,  $r_{13}$  and  $r_{23}$  which are obtained from the following table:

S. No	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	(X <sub>1</sub> ) <sup>2</sup>	(X <sub>2</sub> ) <sup>2</sup>	(X <sub>3</sub> ) <sup>2</sup>	X <sub>1</sub> X <sub>2</sub>	X <sub>1</sub> X <sub>3</sub>	X <sub>2</sub> X <sub>3</sub>
1	2	3	1	4	9	1	6	2	3
2	5	6	3	25	36	9	30	15	18
3	7	10	6	49	100	36	70	42	60
4	11	12	10	121	144	100	132	110	120
TOT	25	31	20	199	289	146	238	169	201

Now we get the total correlation coefficient  $r_{12}$ ,  $r_{13}$  and  $r_{23}$

$$r_{12} = \frac{N(\sum X_1 X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}} \sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}}} \\ r_{12} = 0.97$$

$$r_{13} = \frac{N(\sum X_1 X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}} \sqrt{\{N(\sum X_3^2) - (\sum X_3)^2\}}} \\ r_{13} = 0.99$$

$$r_{23} = \frac{N(\sum X_2 X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}} \sqrt{\{N(\sum X_3^2) - (\sum X_3)^2\}}} \\ r_{23} = 0.97$$

Now, we calculate  $R_{1.23}$

We have,  $r_{12} = 0.97$ ,  $r_{13} = 0.99$  and  $r_{23} = 0.97$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$R_{1.23} = 0.99$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}}$$

$$R_{2.13} = 0.97$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{13}r_{23}r_{12}}{1 - r_{12}^2}}$$

$$R_{3.12} = 0.99$$

## **Regression:**

Regression is a mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

### **➤ Lines of regression:**

1. The line of regression of  $Y$  on  $X$  is given by  $y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X} (x - \bar{x})$
2. The line of regression of  $X$  on  $Y$  is given by  $x - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y} (y - \bar{y})$

### **➤ Regression Coefficients:**

1. Regression coefficient of  $Y$  on  $X$ :  $r \cdot \frac{\sigma_Y}{\sigma_X} = b_{YX}$

$$\text{Where } b_{YX} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

2. Regression coefficient of  $X$  on  $Y$ :  $r \cdot \frac{\sigma_X}{\sigma_Y} = b_{XY}$

$$\text{Where } b_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}$$

3. Correlation coefficient  $r = \pm \sqrt{b_{XY} \times b_{YX}}$

### **Remarks:**

1. The lines of regression of  $Y$  on  $X$  and  $X$  on  $Y$  passes through the mean value of  $x$  and  $y$ . In other words, the mean value of  $x$  and  $y$  can be obtained as the point of intersection of the two regression lines.
2. In case of perfect correlation. ( $r = \pm 1$ ), both the lines of regression coincide. Therefore, in general, we always have two lines pf regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.
3. The sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term. Thus, if the regression coefficients are positive, ' $r$ ' is positive and if the

regression coefficients are negative ' $r$ ' is negative.

4. If one of the regression coefficients is greater than unity, the other must be less than unity.
5. If the two variables are uncorrelated, the lines of regression become perpendicular to each other.

### Problems:

1. From the following data find (i) two regression equations (ii) the coefficient of correlation (iii) Find  $Y$  when  $X = 30$

$X$	25	28	35	32	31	36	29	38	34	32
$Y$	43	46	49	41	36	32	31	30	33	39

### Solution:

$X$	$Y$	$X - \bar{X}$ $= X - 32$	$Y - \bar{Y}$ $= Y - 38$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
<b>320</b>	<b>380</b>	<b>0</b>	<b>0</b>	<b>140</b>	<b>398</b>	<b>-93</b>

$$\text{Here, } \bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{380}{10} = 38$$

The line of regression of  $X$  on  $Y$  is given by  $x - \bar{x} = b_{XY}(y - \bar{y})$

$$\begin{aligned} b_{XY} &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{-93}{398} = -0.2337 \\ \Rightarrow (x - 32) &= -0.2337(y - 38) \\ &= -0.2337y + 0.2337 \times 38 \\ \Rightarrow x &= -0.2337y + 40.8806 \end{aligned}$$

The line of regression of  $Y$  on  $X$  is given by  $y - \bar{y} = b_{YX}(x - \bar{x})$

$$\begin{aligned} b_{YX} &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{-93}{140} = -0.6643 \\ \Rightarrow (y - 38) &= -0.6643(x - 32) \\ &= -0.6643x + 0.6643 \times 32 \\ \Rightarrow y &= -0.6643x + 59.2576 \end{aligned}$$

$$\begin{aligned} \text{Coefficient of correlation } r^2 &= b_{YX} \times b_{XY} \\ &= (-0.6643)(-0.2337) = 0.1552 \\ r &= \pm \sqrt{0.1552} = \pm 0.394 \end{aligned}$$

$$\begin{aligned} \text{When } X = 30, Y &= (-0.6643)(30) + 59.2576 \\ &= 39.3286 \end{aligned}$$

2. The two lines of regression are  $8x - 10y + 66 = 0$ ,  $40x - 18y - 214 = 0$ . The variance of  $X$  is 9. Find the mean values of  $X$  and  $Y$ .

### Solution:

Since both the lines of regression passes through the mean values  $\bar{x}$  and  $\bar{y}$ , the point  $(\bar{x}, \bar{y})$  must satisfy the two given regression lines.

$$8\bar{x} - 10\bar{y} = -66 \quad \dots \dots \dots (1)$$

$$40\bar{x} - 18\bar{y} = 214 \quad \dots \dots \dots (2)$$

Solving these (1) and (2) we get,  $\bar{x} = 13$ ,  $\bar{y} = 17$

3. Estimate the regression line from the given information:

*Solution:*  $\sum_{i=1}^{33} x_i = 1104$ ,  $\sum_{i=1}^{33} y_i = 1124$ ,  $\sum_{i=1}^{33} x_i y_i = 41,355$ ,  $\sum_{i=1}^{33} x_i^2 = 41,086$

Therefore,

$$b_1 = \frac{(33)(41,355) - (1104)(1124)}{(33)(41,086) - (1104)^2} = 0.903643 \text{ and}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

4. The two regression lines are given as  $x+2y-5=0$  and  $2x+3y-8=0$ . Which one is the regression line of  $x$  on  $y$ ?

*Suppose  $x + 2y - 5 = 0$  is the equn. of the reg. line of  $x$  on  $y$  &  $2x + 3y - 8 = 0$  is the equn. of the reg. line of  $y$  on  $x$ ,*

*then the 2 equns can be written as  $x = -2y + 5$*

$$\& y = -\frac{2}{3}x + \frac{8}{3} \quad \text{Hence } b_{yx} = -\frac{2}{3} \& b_{xy} = -2$$

$$\text{Now } r^2 = \frac{4}{3} > 1$$

**This is impossible. Hence our assumption is wrong**

*$\therefore 2x + 3y - 8 = 0$  is the equn. Of reg. line of  $x$  on  $y$*

5. The Two Lines of Regressions Are  $X + 2y - 5 = 0$  and  $2x + 3y - 8 = 0$  and the Variance of X is 12. Find the Variance of Y and the Coefficient of Correlation.

Let  $y = -\frac{1}{2}x + \frac{5}{2}$  be the regression line of y on x

and  $x = -\frac{3}{2}y + \frac{8}{2}$  be the regression line of x on y

Now,  $b_{yx} = -\frac{1}{2}$     $b_{xy} = -\frac{3}{2}$

$$\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{-1}{2} \cdot \frac{-3}{2}}$$

$$= \sqrt{\frac{3}{4}} = \frac{-\sqrt{3}}{2} < 1$$

$$r = \frac{-\sqrt{3}}{2}$$

$$\text{Now, } \sigma_x = \sqrt{12} = 2\sqrt{3}$$

We have:  $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

$$-\frac{1}{2} = -\frac{\sqrt{3}}{2} \cdot \frac{\sigma_y}{2\sqrt{3}}$$

$$\Rightarrow \sigma_y = 2$$

$\therefore$  Variance of y = 4

coefficient of correlation =  $\frac{-\sqrt{3}}{2}$  ... (same sign as  $b_{yx}$  and  $b_{xy}$ )

### Advanced types of linear regression (not in Syllabus)

Linear models are the oldest type of regression. It was designed so that statisticians can do the calculations by hand. However, OLS ( Ordinary Least squares) has several weaknesses, including a sensitivity to both outliers and multicollinearity, and it is prone to overfitting. To address these problems, statisticians have developed several advanced variants:

- **Lasso regression** (least absolute shrinkage and selection operator) performs variable selection that aims to increase prediction accuracy by identifying a simpler model. It is similar to Ridge regression but with variable selection.
- **Ridge regression** allows you to analyse data even when severe multicollinearity is present and helps prevent overfitting. This type of model reduces the large, problematic variance that multicollinearity causes

by introducing a slight bias in the estimates. The procedure trades away much of the variance in exchange for a little bias, which produces more useful coefficient estimates when multicollinearity is present.

- **Partial least squares (PLS) regression** is useful when you have very few observations compared to the number of independent variables or when your independent variables are highly correlated. PLS decreases the independent variables down to a smaller number of uncorrelated components, similar to Principal Components Analysis. Then, the procedure performs linear regression on these components rather than the original data. PLS emphasizes developing predictive models and is not used for screening variables. Unlike OLS, you can include multiple continuous *dependent* variables. PLS uses the correlation structure to identify smaller effects and model multivariate patterns in the dependent variables.

### **Practice Problem:**

1. Find the regression equations for the following data:

X	1	3	5	7	9
Y	15	18	21	23	22

**Solution:**  $x = 0.887y - 12.562$ ,  $y = 0.95x + 15.05$

### **Multiple Regression**

If the number of independent variables in a regression model is more than one, then the model is called as multiple regression. In fact, many of the real-world applications demand the use of multiple regression models.

#### **Assumptions of multiple linear regression**

**Homogeneity of variance (homoscedasticity):** the size of the error in our prediction doesn't change significantly across the values of the independent variable.

**Independence of observations:** the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated ( $r^2 > \sim 0.6$ ), then only one of them should be used in the regression model.

**Normality:** The data follows a normal distribution.

**Linearity:** the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

### Multiple linear Regression formula

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

- $y$  = the predicted value of the dependent variable
- $b_0$  = the  $y$ -intercept (value of  $y$  when all other parameters are set to 0)
- $b_1 X_1$  = the regression coefficient ( $B_1$ ) of the first independent variable ( $X_1$ ) (a.k.a. the effect that increasing the value of the independent variable has on the predicted  $y$  value)
- ... = do the same for however many independent variables you are testing
- $b_n X_n$  = the regression coefficient of the last independent variable

### Application:

where  $Y$  represents the economic growth rate of a country,  $X_1$  represents the time period,  $X_2$  represents the size of the populations of the country,  $X_3$  represents the level of employment in percentage,  $X_4$  represents the percentage of literacy,  $b_0$  is the intercept and  $b_1, b_2, b_3$  and  $b_4$  are the slopes of the variables  $X_1, X_2, X_3$  and  $X_4$  respectively. In this regression model,  $X_1, X_2, X_3$  and  $X_4$  are the independent variables and  $Y$  is the dependent variable.

### Regression model with two independent variables using normal equations:

If the regression equation with two independent variables is

$$Y = b_o + b_1 X_1 + b_2 X_2$$

Then, the normal equations are

$$\Sigma Y = nb_o + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

$$\Sigma YX_1 = b_o \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2$$

$$\Sigma YX_2 = b_o \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2$$

### Problems:

1. The annual sales revenue (in crores of rupees) of a product as a function of sales force (number of salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table.

Annual sales revenue $Y$	20	23	25	27	21	29	22	24	27	35
Sales force $X_1$	8	13	8	18	23	16	10	12	14	20
Annual advertising expenditures $X_2$	28	23	38	16	20	28	23	30	26	32

Let the regression model be

$$Y = b_o + b_1 X_1 + b_2 X_2$$

$Y$	$X_1$	$X_2$	$X_1^2$	$X_2^2$	$X_1 X_2$	$YX_1$	$YX_2$
20	8	28	64	784	224	160	560
23	13	23	169	529	299	299	529
25	8	38	64	1444	304	200	950
27	18	16	324	256	288	486	432
21	23	20	529	400	460	483	420
29	16	28	256	784	448	464	812
22	10	23	100	529	230	220	506

24	12	30	144	900	360	288	720	
27	14	26	196	676	364	378	702	
35	20	32	400	1024	640	700	1120	
253	142	264	2246	7326	3617	3678	6751	<b>Total</b>

Substituting the required values in the normal equations, we get the following simultaneous equations

$$253 = 10b_o + 142b_1 + 264b_2$$

$$3678 = 142b_o + 2246b_1 + 3617b_2$$

$$6751 = 264b_o + 3617b_1 + 7326b_2$$

The solution to the above set of simultaneous equation is

$$b_o = 5.1483, \quad b_1 = 0.6190 \quad \text{and} \quad b_2 = 0.4304$$

Therefore, the regression model is  $Y = 5.1483 + 0.6190X_1 + 0.4304X_2$

If mean, standard deviation and partial correlation of the trivariate distribution are known, then the multiple regression of  $X_1$  on  $X_2$  and  $X_3$  is given by

$$(X_1 - \bar{X}_1) \frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2) \frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3) \frac{\omega_{13}}{\sigma_3} = 0$$

where  $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

$$\omega_{12} = - \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13} r_{23} - r_{21}$$

$$\omega_{13} = r_{23} r_{12} - r_{13}$$

Example :

Find the regression equation of  $X_1$  on  $X_2$  and  $X_3$  given the following results :—

Trait	Mean	Standard deviation	$r_{12}$	$r_{23}$	$r_{31}$
$X_1$	28.02	4.42	+ 0.80	—	—
$X_2$	4.91	1.10	—	-0.56	—
$X_3$	594	85	—	—	-0.40

where  $X_1$  = Seed per acre;  $X_2$  = Rainfall in inches

$X_3$  = Accumulated temperature above 42°F.

Solution:

Regression equation of  $X_1$  on  $X_2$  and  $X_3$  is given by

$$(X_1 - \bar{X}_1) \frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2) \frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3) \frac{\omega_{13}}{\sigma_3} = 0$$

$$\text{where } \omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0.56)^2 = 0.686$$

$$\omega_{12} = - \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13} r_{23} - r_{21} = -0.576$$

$$\omega_{13} = r_{23} r_{12} - r_{13} = (-0.56) (0.80) - (-0.40) = -0.048$$

$\therefore$  Required equation of plane of regression of  $X_1$  on  $X_2$  and  $X_3$  is given by

$$\frac{0.686}{4.42} (X_1 - 28.02) + \frac{(-0.576)}{1.10} (X_2 - 4.91) + \frac{(-0.048)}{85.00} (X_3 - 594) = 0$$

### Practice Problems :

1.

Data Collected From Random Sample of 5 General Motors Salespeople

Independent Variable 1 (X1)	Independent Variable 2 (X2)	Dependent Variable (Y)
Highest Year of School Completed	Motivation as Measured by Higgins Motivation Scale	Annual Sales in Dollars
12	32	\$350,000
14	35	\$399,765
15	45	\$429,000
16	50	\$435,000
18	65	\$433,000

Solution :  $r = 0.9360$  .

# **MAT2001**

## **Statistics for Engineers**

### **Module 4**

### **Probability Distributions**

## **Syllabus**

.

#### **Probability Distributions:**

Binomial and Poisson distributions – Normal distribution – Gamma distribution – Exponential distribution – Weibull distribution.

# **Special Probability Distributions**

## **Discrete Probability Distributions**

- 1. Binomial Distribution*
- 2. Poisson Distribution*

## **Continuous Probability Distributions**

- 1. Normal Distribution*
- 2. Exponential Distribution*
- 3. Gamma Distribution*
- 4. Weibull Distribution*

# Binomial Distribution

**Definition:** Let  $A$  be some event associated with a random experiment  $E$ , such that  $P(A) = p$  and  $P(\bar{A}) = 1 - p = q$ . Assuming that  $p$  remains the same for all repetitions, if we consider  $n$  independent repetitions (or trials) of  $E$  and if the random variable (RV)  $X$  denotes the number of times the event  $A$  has occurred, then  $X$  is called a *binomial random variable* with parameters  $n$  and  $p$  or we say that  $X$  follows a *binomial distribution* with parameters  $n$  and  $p$ , or symbolically  $B(n, p)$ . Obviously the possible values that  $X$  can take, are  $0, 1, 2, \dots, n$ .

## Probability Mass Function of the Binomial Distribution

$$P(X = r) = nC_r p^r q^{n-r}; \quad r = 0, 1, 2, \dots, n \text{ where } p + q = 1$$

---

(i) Binomial distribution is a legitimate probability distribution since

$$\begin{aligned} \sum_{r=0}^n P(X = r) &= \sum_{r=0}^n nC_r q^{n-r} p^r \\ &= (q + p)^n = 1 \end{aligned}$$

## Binomial Distribution $B(n, p)$

A DRV ' $X$ ' is said to follows B.D if the pmf of  $X$  is defined as

$$P_r = P(X=x_r) = P(X=r) = nC_r \cdot p^r \cdot q^{n-r}$$

$X$  represents the trials of the experiment.

$$x = 0, 1, 2, 3, \dots, n$$

$$r \in \{0, 1, 2, \dots, n\}$$

$$nC_r = \frac{n!}{(n-r)! r!}$$

$p$  = probability of success

$q$  = probability of failure

$$\Rightarrow p+q=1 \quad ((1-p)q=1-p)$$

For pmf

(i). tve near

$$p_r \geq 0, \forall r = 0, 1, 2, \dots, n$$

$$(ii). \sum_{r=0}^n p_r = \sum_{r=0}^n nC_r \cdot p^r \cdot q^{n-r}$$

$$= (p+q)^n = 1$$

$$(x+q)^n = \sum_{r=0}^n nC_r x^r \cdot q^{n-r}$$

## Mean and Variance of the Binomial Distribution

$$\begin{aligned} E(X) &= \sum_r x_r p_r \\ &= \sum_{r=0}^n r \cdot nC_r p^r q^{n-r} \\ &= \sum_{r=0}^n r \cdot \frac{n!}{r!(n-r)!} p^r q^{n-r} \end{aligned} \tag{1}$$

$$\begin{aligned} &= np \cdot \sum_{r=1}^{n-1} \frac{(n-1)!}{(r-1)!\{(n-1)-(r-1)\}!} p^{r-1} q^{(n-1)-(r-1)} \\ &= np \sum_{r=1}^{n-1} (n-1) C_{r-1} \cdot p^{r-1} \cdot q^{(n-1)-(r-1)} \\ &= np (q+p)^{n-1} \\ &= np \end{aligned} \tag{2}$$

$$\begin{aligned}
E(X^2) &= \sum_r x_r^2 p_r = \sum_0^n r^2 p_r \\
&= \sum_{r=0}^n \{r(r-1) + r\} \frac{n!}{r!(n-r)!} p^r q^{n-r} \\
&= n(n-1)p^2 \sum_{r=2}^n (n-2)C_{r-2} p^{r-2} q^{n-r} + np, \\
&\quad \text{[by (1) and (2)]} \\
&= n(n-1)p^2 (q+p)^{n-2} + np \\
&= n(n-1)p^2 + np
\end{aligned}$$

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - \{E(X)\}^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= np(1-p) \\
&= npq
\end{aligned}$$

# Poisson Distribution

**Definition:** If  $X$  is a discrete RV that can assume the values  $0, 1, 2, \dots$ , such that its probability mass function is given by

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}; \quad r = 0, 1, 2, \dots; \quad \lambda > 0$$

then  $X$  is said to follow a *Poisson distribution* with parameter  $\lambda$  or symbolically  $X$  is said to follow  $P(\lambda)$ .

**(Note:** Poisson distribution is a legitimate probability distribution, since

$$\begin{aligned} \sum_{r=0}^{\infty} P(x = r) &= \sum_{r=0}^{\infty} \frac{e^{-\lambda} \lambda^r}{r!} \\ &= e^{-\lambda} e^{\lambda} = 1 \end{aligned}$$

## Mean and Variance of the Poisson Distribution

$$\begin{aligned} E(X) &= \sum_r x_r p_r \\ &= \sum_{r=0}^{\infty} r \frac{e^{-\lambda} \cdot \lambda^r}{r!} \end{aligned} \tag{1}$$

$$\begin{aligned} &= \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned} \tag{2}$$

$$\begin{aligned} E(X^2) &= \sum_r x_r^2 p_r \\ &= \sum_{r=0}^{\infty} \{r(r-1) + r\} e^{-\lambda} \frac{\lambda^r}{r!} \\ &= \lambda^2 e^{-\lambda} \sum_{r=2}^{\infty} \frac{\lambda^{r-2}}{(r-2)!} + \lambda \quad [\text{by (1) and (2)}] \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - \{E(X)\}^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

## Poisson Distribution as Limiting Form of Binomial Distribution

Poisson distribution is a limiting case of binomial distribution under the following conditions:

- (i)  $n$ , the number of trials is indefinitely large, i.e.,  $n \rightarrow \infty$ .
- (ii)  $p$ , the constant probability of success in each trial is very small, i.e.,  $p \rightarrow 0$ .
- (iii)  $np$  ( $= \lambda$ ) is finite or  $p = \frac{\lambda}{n}$  and  $q = 1 - \frac{\lambda}{n}$ , where  $\lambda$  is a positive real number.

**Note:**

if  $X$  is the Poisson RV

$$\boxed{\lim_{n \rightarrow \infty} B(n, p) = P(X)}$$

$$E(X) = \lim_{\substack{n \rightarrow \infty \\ np = \lambda}} (np) = \lambda$$

$$\text{and } \text{Var}(X) = \lim_{\substack{p \rightarrow 0 \\ np = \lambda}} (npq) = \lim_{p \rightarrow 0} [\lambda(1-p)] = \lambda.$$

## **Example:**

Out of 800 families with 4 children each, how many families would be expected to have (i) 2 boys and 2 girls, (ii) at least 1 boy, (iii) at most 2 girls and (iv) children of both sexes. Assume equal probabilities for boys and girls.

Example:

Out of  $\textcircled{800}$  families with  $\textcircled{4}$  children each, how many families would be expected to have (i) 2 boys and 2 girls, (ii) at least 1 boy, (iii) at most 2 girls and (iv) children of both sexes. Assume equal probabilities for boys and girls.

Soln.:

Let  $X$  represents the no. of boys

$X = 0, 1, 2, 3 \times \textcircled{4}$ . (girls)

$\Rightarrow X \rightarrow \text{DRV}$

$n = \textcircled{4}$  &  $p = \text{probability of success}$

$$p = \frac{1}{2} \Rightarrow q = \frac{1}{2}$$

(i)  $P(2 \text{ Boys} \& 2 \text{ Girls})$

$$P(X=2) = 4C_2 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^{4-2}$$

**Solution:**

Considering each child as a trial,  $n = 4$ . Assuming that birth of a boy is a success,  $p = \frac{1}{2}$  and  $q = \frac{1}{2}$ . Let  $X$  denote the number of successes (boys).

(i)  $P(2 \text{ boys and } 2 \text{ girls}) = P(X = 2)$

$$\begin{aligned} &= 4C_2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^{4-2} \\ &= 6 \times \left(\frac{1}{2}\right)^4 = \frac{3}{8} \end{aligned}$$

$\therefore$  No. of families having 2 boys and 2 girls

$$\begin{aligned} &= N \cdot (P(X = 2)) \text{ (where } N \text{ is the total no. of families considered)} \\ &= 800 \times \frac{3}{8} \\ &= 300. \end{aligned}$$

(ii)  $P(\text{at least } 1 \text{ boy}) = P(X \geq 1)$

$$\begin{aligned} &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= 1 - P(X = 0) \\ &= 1 - 4C_0 \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^4 \\ &= 1 - \frac{1}{16} = \frac{15}{16} \end{aligned}$$

$\therefore$  No. of families having at least 1 boy

$$= 800 \times \frac{15}{16} = 750.$$

(iii)  $P(\text{at most } 2 \text{ girls}) = P(\text{exactly } 0 \text{ girl, } 1 \text{ girl or } 2 \text{ girls})$

$$\begin{aligned} &= P(X = 4, X = 3 \text{ or } X = 2) \\ &= 1 - \{P(X = 0) + P(X = 1)\} \\ &= 1 - \left\{ 4C_0 \cdot \left(\frac{1}{2}\right)^4 + 4C_1 \cdot \left(\frac{1}{2}\right)^4 \right\} \\ &= \frac{11}{16} \end{aligned}$$

$\therefore$  No. of families having at most 2 girls

$$= 800 \times \frac{11}{16} = 550.$$

(iv)  $P(\text{children of both sexes})$

$$\begin{aligned} &= 1 - P(\text{children of the same sex}) \\ &= 1 - \{P(\text{all are boys}) + P(\text{all are girls})\} \\ &= 1 - \{P(X = 4) + P(X = 0)\} \\ &= 1 - \left\{ 4C_4 \cdot \left(\frac{1}{2}\right)^4 + 4C_0 \cdot \left(\frac{1}{2}\right)^4 \right\} \\ &= 1 - \frac{1}{8} = \frac{7}{8} \end{aligned}$$

$\therefore$  No. of families having children of both sexes

$$= 800 \times \frac{7}{8} = 700.$$

**Example:**

An irregular 6-faced die is such that the probability that it gives 3 even numbers in 5 throws is twice the probability that it gives 2 even numbers in 5 throws. How many sets of exactly 5 trials can be expected to give no even number out of 2500 sets?

Soln.

Let  $X$  represents the no. of even faces of a die.

$$X = 0, 1, 2, 3, 4, 5 \rightarrow B(n, p)$$
$$n = 5, r \in \{0, 1, \dots, 5\}$$
$$p = ?$$

Given,

$$P(X=3) = 2 \times P(X=2) \checkmark$$

$${}^5C_3 p^3 \cdot (1-p)^2 = 2 \times {}^5C_2 p^2 \cdot (1-p)^3$$

$$\Rightarrow p=? \quad \& \quad q=?$$

$$P(X=0) = {}^5C_0 \cdot p^0 \cdot q^5$$
$$= ?$$

$$N \times P(X=0) = 2500 \times ?$$

## Solution:

Let the probability of getting an even number with the unfair die be  $p$ .

Let  $X$  denote the number of even numbers obtained in 5 trials (throws).

**Given:**  $P(X = 3) = 2 \times P(X = 2)$

i.e.,  $5C_3 p^3 q^2 = 2 \times 5C_2 p^2 q^3$

i.e.,  $p = 2q = 2(1 - p)$

$\therefore 3p = 2$  or  $p = \frac{2}{3}$  and  $q = \frac{1}{3}$

Now  $P(\text{getting no even number})$

$$= P(X = 0)$$

$$= 5C_0 \cdot p^0 \cdot q^5 = \left(\frac{1}{3}\right)^5 = \frac{1}{243}$$

$\therefore$  Number of sets having no success (even number) out of  $N$  sets  $= N \times P(X = 0)$

$$\therefore \text{Required number of sets} = 2500 \times \frac{1}{243}$$

$$= 10, \text{ nearly}$$

Example:

Two dice are thrown 120 times. Find the average number of times in which the number on the first dice exceeds the number on the second dice.

mean

Soln.       $B(n, p)$

$$n = 120 \checkmark$$

$$p = ?$$

$$S = \{(1,1), (1,2), \dots, (1,6), \\ (2,1), (2,2), \dots, (2,5), \\ \dots \\ (6,1), (6,2), \dots, (6,5)\}$$
$$|S| = 36$$

$$\text{Expected } E(X) = \{(2,1), (3,1), (3,2), \\ \dots, (6,5)\}$$

$$|E| = 15$$

$$p = \frac{|E|}{|S|} = \frac{15}{36}$$

$$\text{Average} = E(X) = n \times p = 120 \times \frac{15}{36} \\ = \underline{\underline{50}}$$

## Solution:

The number on the first dice exceeds that on the second die, in the following combinations:

(2, 1); (3, 1), (3, 2); (4, 1), (4, 2), (4, 3); (5, 1), (5, 2), (5, 3); (5, 4); (6, 1),  
(6, 2), (6, 3), (6, 4), (6, 5),

where the numbers in the parentheses represent the numbers in the first and second dice respectively.

$$\therefore P(\text{success}) = P(\text{no. in the first die exceeds the no. in the second die})$$

$$= \frac{15}{36} = \frac{5}{12}$$

This probability remains the same in all the throws that are independent.

If  $X$  is the no. of successes, then  $X$  follows a binomial distribution with parameters  $n (= 120)$  and  $p\left(=\frac{5}{12}\right)$ .

$$\therefore E(X) = np = 120 \times \frac{5}{12} = 50$$

## Example:

Fit a binomial distribution for the following data:

$x:$	0	1	2	3	4	5	6	Total
$f:$	5	18	28	12	7	6	4	80

frequency

### Solution:

Fitting a binomial distribution means assuming that the given distribution is approximately binomial and hence finding the probability mass function and then finding the theoretical frequencies.

To find the binomial frequency distribution  $N(q + p)^n$ , which fits the given data, we require  $N$ ,  $n$  and  $p$ . We assume  $N = \text{total frequency} = 80$  and  $n = \text{no. of trials} = 6$  from the given data.

To find  $p$ , we compute the mean of the given frequency distribution and equate it to  $np$  (mean of the binomial distribution).

$x$ :	0	1	2	3	4	5	6	Total
$f$ :	5	18	28	12	7	6	4	80
$fx$ :	0	18	56	36	28	30	24	192

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{192}{80} = 2.4$$

i.e.,

$$np = 2.4 \text{ or } 6p = 2.4$$

∴

$$p = 0.4 \text{ and } q = 0.6$$

If the given distribution is nearly binomial, the theoretical frequencies are given by the successive terms in the expansion of  $80(0.6 + 0.4)^6$ . Thus we get,

$x$ :	0	1	2	3	4	5	6
Theoretical $f$ :	3.73	14.93	24.88	22.12	11.06	2.95	0.33

Converting these values into whole numbers consistent with the condition that the total frequency is 80, the corresponding binomial frequency distribution is as follows:

$x$ :	0	1	2	3	4	5	6	Total
$f$ :	4	15	25	22	11	3	0	80

### Example:

The number of monthly breakdowns of a computer is a RV having a Poisson distribution with mean equal to 1.8. Find the probability that this computer will function for a month

- (a) without a breakdown,
- (b) with only one breakdown and
- (c) with atleast one breakdown.

Soln.:

Let  $X = 0, 1, 2, 3, \dots$

$X \sim PD(\lambda)$  with the mean  $= \lambda = 1.8$

$$\lambda = 1.8$$

$$P_r = P(X=r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!}, r=0, 1, 2, \dots$$

$$P(X=r) = \frac{e^{-1.8} \cdot (1.8)^r}{r!}$$

## Solution:

Let  $X$  denote the number of breakdowns of the computer in a month.  $X$  follows a Poisson distribution with mean (parameter)  $\lambda = 1.8$ .

$$\therefore P\{X = r\} = \frac{e^{-\lambda} \cdot \lambda^r}{r!} = \frac{e^{-1.8} \cdot (1.8)^r}{r!}$$

- (a)  $P(X = 0) = e^{-1.8} = 0.1653$
- (b)  $P(X = 1) = e^{-1.8} (1.8) = 0.2975$
- (c)  $P(X \geq 1) = 1 - P(X = 0) = 0.8347$

### Example:

Fit a Poisson distribution for the following distribution:

$x:$	0	1	2	3	4	5	Total
$f:$	142	156	69	27	5	1	400

**Solution:**

Fitting a Poisson distribution for a given distribution means assuming that the given distribution is approximately Poisson and hence finding the probability mass function and then finding the theoretical frequencies.

To find the probability mass function

$$P\{X = r\} = \frac{e^{-\lambda} \cdot \lambda^r}{r!} \quad r = 0, 1, 2, \dots, \infty$$

of the approximate Poisson distribution, we require  $\lambda$ , which is the mean of the Poisson distribution.

We find the mean of the given distribution and assume it as  $\lambda$ .

$x$ :	0	1	2	3	4	5	Total
$f$ :	142	156	69	27	5	1	400
$fx$ :	0	156	138	81	20	5	400

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{400}{400} = 1 = \lambda$$

The theoretical frequencies are given by

$$\begin{aligned} & \frac{N e^{-\lambda} \cdot \lambda^r}{r!} \text{ where } N=400, \text{ obtained from the given distribution.} \\ & = \frac{400 e^{-1}}{r!}, \quad r = 0, 1, 2, \dots, \infty \end{aligned}$$

Thus, we get

$x$ :	0	1	2	3	4	5
Theoretical $f$ :	147.15	147.15	73.58	24.53	6.13	1.23

The theoretical frequencies for  $x = 6, 7, 8, \dots$  are very small and hence neglected.

Converting the theoretical frequencies into whole numbers consistent with the condition that the total frequency = 400, we get the following Poisson frequency distribution which fits the given distribution:

$x$ :	0	1	2	3	4	5
Theoretical $f$ :	147	147	74	25	6	1

### **Exercise:**

It is known that the probability of an item produced by a certain machine will be defective is 0.05. If the produced items are sent to the market in packets of 20, find the number of packets containing at least, exactly and at most 2 defective items in a consignment of 1000 packets using (i) binomial distribution and (ii) Poisson approximation to binomial distribution.

### **Exercise:**

If a fair coin is tossed at random 5 independent times, find the conditional probability of 5 heads relative to the hypothesis that there are at least 4 heads.

### **Exercise:**

A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day follows a Poisson distribution with mean 1.5. Calculate the proportion of days on which (i) neither car is used and (ii) some demand is not fulfilled.

# Exponential Distribution $ED(\lambda)$

**Definitions:** A continuous RV  $X$  is said to follow an *exponential distribution* or *negative exponential distribution* with parameter  $\underline{\lambda > 0}$ , if its probability density function is given by

$$\text{Pdt} = f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We note that  $\int_0^{\infty} f(x) dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1$  and hence  $f(x)$  is a legitimate density function.

For pdt:

- ①.  $f(x) \geq 0$  (for non)
- ②.

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^0 0 dx + \int_0^{\infty} t e^{-\lambda x} dt = 1$$

## Mean and Variance of the Exponential Distribution

$E(X)$  = Mean of the exponential distribution

$$= \mu_1' = \underline{\frac{1}{\lambda}}$$

$$E(X^2) = \mu_2' = \underline{\frac{2}{\lambda^2}}$$

$$\therefore \text{Var}(X) = E(X^2) - \{E(X)\}^2$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \underline{\frac{1}{\lambda^2}}$$

## Gamma Function

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx, \quad 1 \leq n \leq 2.$$

Some properties of the gamma function:

$$\Gamma(n+1) = n\Gamma(n), \quad n > 0,$$

and when  $n = \text{integer} > 0$ , we have  $\Gamma(n) = (n - 1)!$

## Beta Function

.

The gamma function is related to the beta function,  $B(m,n)$ , as follows:

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx$$

$$B(m, n) = B(n, m) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

## Table for Gamma Function

$n$	$\Gamma(n)$	$n$	$\Gamma(n)$	$n$	$\Gamma(n)$	$n$	$\Gamma(n)$
1.00	1.00000	1.25	0.90640	1.50	0.88623	1.75	0.91906
1.01	0.99433	1.26	0.90440	1.51	0.88659	1.76	0.92137
1.02	0.98884	1.27	0.90250	1.52	0.88704	1.77	0.92376
1.03	0.98355	1.28	0.90072	1.53	0.88757	1.78	0.92623
1.04	0.97844	1.29	0.89904	1.54	0.88818	1.79	0.92877
1.05	0.97350	1.30	0.89747	1.55	0.88887	1.80	0.93138
1.06	0.96874	1.31	0.89600	1.56	0.88964	1.81	0.93408
1.07	0.96415	1.32	0.89464	1.57	0.89049	1.82	0.93685
1.08	0.95973	1.33	0.89338	1.58	0.89142	1.83	0.93969
1.09	0.95546	1.34	0.89222	1.59	0.89243	1.84	0.94261
1.10	0.95135	1.35	0.89115	1.60	0.89352	1.85	0.94561
1.11	0.94739	1.36	0.89018	1.61	0.89468	1.86	0.94869
1.12	0.94359	1.37	0.88931	1.62	0.89592	1.87	0.95184
1.13	0.93993	1.38	0.88854	1.63	0.89724	1.88	0.95507
1.14	0.93642	1.39	0.88785	1.64	0.89864	1.89	0.95838
1.15	0.93304	1.40	0.88726	1.65	0.90012	1.90	0.96177
1.16	0.92980	1.41	0.88676	1.66	0.90167	1.91	0.96523
1.17	0.92670	1.42	0.88636	1.67	0.90330	1.92	0.96878
1.18	0.92373	1.43	0.88604	1.68	0.90500	1.93	0.97240
1.19	0.92088	1.44	0.88580	1.69	0.90678	1.94	0.97610
1.20	0.91817	1.45	0.88565	1.70	0.90864	1.95	0.97988
1.21	0.91558	1.46	0.88560	1.71	0.91057	1.96	0.98374
1.22	0.91311	1.47	0.88563	1.72	0.91258	1.97	0.98768
1.23	0.91075	1.48	0.88575	1.73	0.91466	1.98	0.99171
1.24	0.90852	1.49	0.88595	1.74	0.91683	1.99	0.99581
						2.00	1.00000

## General Gamma or Erlang Distribution

**Definition:** A continuous RV  $X$  is said to follow an *Erlang distribution* or *General Gamma distribution* with parameters  $\lambda > 0$  and  $k > 0$ , if its probability density function is given by

$$E_Y(k, \lambda)$$

Pdf =  $f(x) = \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, & \text{for } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$

We note that  $\int_0^\infty f(x) dx = \frac{\lambda^k}{\Gamma(k)} \int_0^\infty x^{k-1} e^{-\lambda x} dx$

$$= \frac{1}{\Gamma(k)} \int_0^\infty t^{k-1} e^{-t} dt, [\text{on putting } \lambda x = t]$$
$$= 1$$

Hence  $f(x)$  is a legitimate density function.

## Mean and Variance of the General Gamma or Erlang Distribution

$$\text{Mean} = E(X) = \frac{1}{\lambda} \cdot \frac{\Gamma(k+1)}{\Gamma(k)} = \frac{k}{\lambda}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \quad \underline{\underline{=}}$$

$$= \frac{1}{\lambda^2} \cdot \frac{\Gamma(k+2)}{\Gamma(k)} - \left( \frac{k}{\lambda} \right)^2$$

$$= \frac{1}{\lambda^2} \{k(k+1) - k^2\} = \frac{k}{\lambda^2}$$

# Gamma Distribution

GD(k)

## Note

- When  $\lambda = 1$ , the Erlang distribution is called Gamma distribution or simple Gamma distribution with parameter  $k$  whose density function is  $f(x) = \frac{1}{\Gamma(k)} x^{k-1} e^{-x}; x \geq 0; k > 0$ .
- When  $k = 1$ , the Erlang distribution reduces to the exponential distribution with parameter  $\lambda > 0$ .
- Sometimes, the Erlang distribution itself is called Gamma distribution.

## Mean and Variance of the Gamma Distribution

$$\text{Mean} = E(X) = k$$

$$\text{Var}(X) = k$$

# Weibull Distribution $WD(\alpha, \beta)$

**Definition:** A continuous RV  $X$  is said to follow a *Weibull distribution* with parameters  $\alpha, \beta > 0$ , if the RV  $Y = \alpha X^\beta$  follows the exponential distribution with density function  $f_Y(y) = e^{-y}$ ,  $y > 0$ .

## Density Function of the Weibull Distribution

Since  $Y = \alpha \cdot X^\beta$ , we have  $y = \alpha \cdot x^\beta$ .

By the transformation rule, derived in chapter 3, we have  $f_X(x) = f_Y(y) \left| \frac{dy}{dx} \right|$ ,

where  $f_X(x)$  and  $f_Y(y)$  are the density functions of  $X$  and  $Y$  respectively.

$$\therefore f_X(x) = e^{-y} \alpha \beta x^{\beta-1}$$
$$= \alpha \beta x^{\beta-1} e^{-\alpha x^\beta}; x > 0 \quad [\because y > 0]$$

pdf:

### Note

When  $\beta = 1$ , Weibull distribution reduces to the exponential distribution with parameter  $\alpha$ .

## Mean and Variance of the Weibull Distribution

$$\text{Mean} = E(X) = \mu_1' = \alpha^{-\frac{1}{\beta}} \left[ \left( \frac{1}{\beta} + 1 \right) \right]$$

$$\begin{aligned}\text{Var}(X) &= E(X^2) - \{E(X)\}^2 \\ &= \alpha^{-2/\beta} \left[ \left( \frac{2}{\beta} + 1 \right) - \left\{ \left( \frac{1}{\beta} + 1 \right) \right\}^2 \right]\end{aligned}$$

# Normal (Gaussian) Distribution $N(\mu, \sigma)$

**Definition:** A continuous RV  $X$  is said to follow a *normal distribution* or *Gaussian distribution* with parameters  $\mu$  and  $\sigma$ , if its probability density function is given by

Pdf = 
$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}; \quad -\infty < x < \infty$$
  
$$-\infty < \mu < \infty \quad \underline{\sigma > 0} \quad (1)$$

Symbolically 'X follows  $N(\mu, \sigma)$ '. Sometimes it is also given as  $N(\mu, \sigma^2)$ .

## Note:

$f(x)$  is a legitimate density function, as

$$\begin{aligned}\int_{-\infty}^{\infty} f(x) dx &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx \\&= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} \sigma \sqrt{2} dt, \quad \left( \text{on putting } t = \frac{x-\mu}{\sigma \sqrt{2}} \right) \\&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt \\&= \frac{1}{\sqrt{\pi}} 2 \int_0^{\infty} e^{-t^2} dt = \frac{1}{\sqrt{\pi}} \cdot \left[ \left( \frac{1}{2} \right) \right] = \frac{1}{\sqrt{\pi}} \cdot \sqrt{\pi} = 1\end{aligned}$$

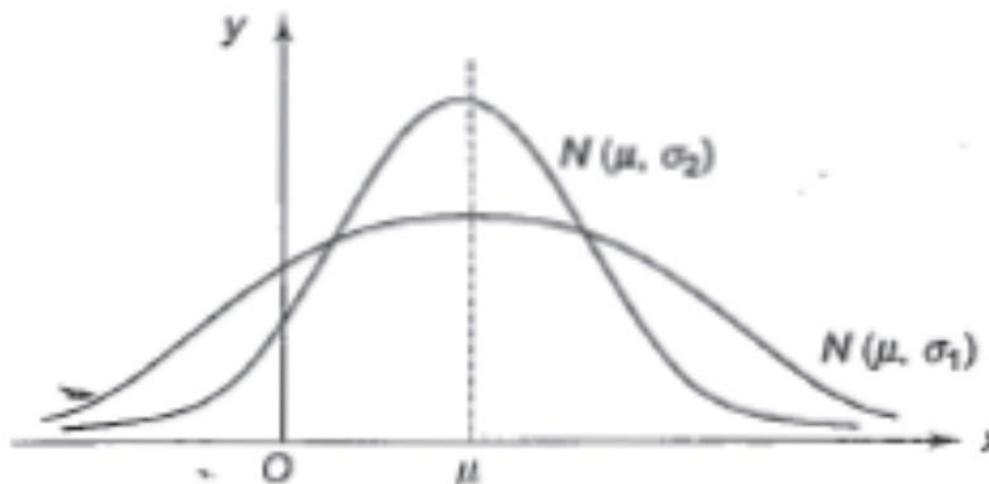
## Standard Normal Distribution

The normal distribution  $N(0, 1)$  is called the standardised or simply the standard normal distribution, whose density function is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$$

This is obtained by putting  $\mu = 0$  and  $\sigma = 1$  and by changing  $x$  and  $f$  respectively into  $z$  and  $\phi$ . If  $X$  has distribution  $N(\mu, \sigma)$  and if  $Z = \frac{X - \mu}{\sigma}$ , then we can prove that  $Z$  has distribution  $N(0, 1)$ .

## Normal Probability Curve



## AREAS UNDER NORMAL CURVE

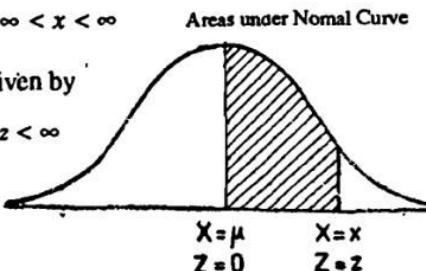
**Normal probability curve is given by**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right\} \quad -\infty < x < \infty$$

and standard normal probability curve is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), -\infty < z < \infty$$

where  $Z = \frac{X - E(X)}{\sigma_x} \sim N(0, 1)$



The following table gives the shaded area in the diagram viz.,  $P(0 < Z < z)$  for different values of  $z$ .

---

**TABLE OF AREAS**

---

## Mean and Variance of the Normal Distribution

If  $X$  follows  $N(\mu, \sigma^2)$ , then  $E(X) = \mu$  and  $\text{Var}(X) = \sigma^2$  and  $SD = \sigma$

## Mean and Variance of the Standard Normal Distribution

$$\text{Mean}(Z) = \mu = 0$$

$$\text{Var}(Z) = \sigma^2 = \underline{1}$$

**Example:**

The mileage which car owners get with a certain kind of radial tire is a RV having an exponential distribution with mean 40,000 km. Find the probabilities that one of these tires will last (i) at least 20,000 km and (ii) at most 30,000 km.

Soln/

$X \sim ED(\lambda)$  with Mean =  $\frac{1}{\lambda} = 40,000$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & ; x \geq 0 \\ 0 & ; \text{otherwise} \end{cases} \Rightarrow \lambda = \frac{1}{40,000}$$

(i).  $P(X \geq 20,000) = \int_{20,000}^{\infty} \frac{1}{40,000} e^{-\frac{1}{40,000}(x)} dx$

(ii)  $P(X \leq 30,000) = \int_{-\infty}^{30,000} f(x) dx$

$$= \int_{-\infty}^{0} (0) dx + \int_{0}^{30,000} f(x) dx$$

**Solution:**

Let  $X$  denote the mileage obtained with the tire

$$f(x) = \frac{1}{40,000} e^{-x/40,000} \quad x > 0$$

$$\begin{aligned}\text{(i)} \quad P(X \geq 20,000) &= \int_{20,000}^{\infty} \frac{1}{40,000} e^{-x/40,000} dx \\&= \left[ -e^{-x/40,000} \right]_{20,000}^{\infty} \\&= e^{-0.5} = 0.6065\end{aligned}$$

$$\begin{aligned}\text{(ii)} \quad P(X \leq 30,000) &= \int_0^{30,000} \frac{1}{40,000} e^{-x/40,000} dx \\&= \left[ -e^{-x/40,000} \right]_0^{30,000} \\&= 1 - e^{-0.75} = 0.5270\end{aligned}$$

### Exercise:

The time (in hours) required to repair a machine is exponentially distributed with parameter  $\lambda = 1/2$ .

- What is the probability that the repair time exceeds 2 h?
- What is the conditional probability that a repair takes at least 10 h given that its duration exceeds 9 h?

HINT:

$$X \sim \text{ED}(\lambda), \text{ with } \lambda = \frac{1}{2}$$

$$(a) P(X > 2)$$

$$(b)$$

$$P(X \geq 10 | X > 9) = \frac{P(X \geq 10, X > 9)}{P(X > 9)}$$

## Example:

In a certain city, the daily consumption of electric power in millions of kilowatt-hours can be treated as a RV having an Erlang distribution with parameters  $\lambda = \frac{1}{2}$  and  $k = 3$ . If the power plant of this city has a daily capacity of 12 millions kilowatt-hours, what is the probability that this power supply will be inadequate on any given day.

Soln.

$$X \sim E_r(\lambda, k)$$

$$P(X > 12)$$

## Solution:

Let  $X$  represent the daily consumption of electric power (in millions of kilo-watt-hours). Then the density function of  $X$  is given as

$$f(x) = \frac{\left(\frac{1}{2}\right)^3}{\Gamma(3)} x^2 e^{-x/2}, x > 0$$

$P(\text{the power supply is inadequate})$

$$= P(X > 12) = \int_{12}^{\infty} f(x) dx \quad [\because \text{The daily capacity is only 12}]$$

$$= \int_{12}^{\infty} \frac{1}{\Gamma(3)} \cdot \frac{1}{2^3} x^2 e^{-x/2} dx$$

$$= \frac{1}{16} \left[ x^2 \left( \frac{e^{-x/2}}{-\frac{1}{2}} \right) - 2x \left( \frac{e^{-x/2}}{\frac{1}{4}} \right) + 2 \left( \frac{e^{-x/2}}{-\frac{1}{8}} \right) \right]_{12}^{\infty}$$

$$= \frac{1}{16} e^{-6} (288 + 96 + 16)$$

$$= 25 e^{-6} = 0.0625$$

## Example:

If the life  $X$  (in years) of a certain type of car has a Weibull distribution with the parameter  $\beta = 2$ , find the value of the parameter  $\alpha$ , given that probability that the life of the car exceeds 5 years is  $e^{-0.25}$ . For these values of  $\alpha$  and  $\beta$ , find the mean and variance of  $X$ .

Soln.

$$X \sim WD(\alpha, \beta),$$

$$\beta = 2$$

$$\alpha = ?$$

$$P(X > 5) = e^{-0.25}$$

## Solution:

The density function of  $X$  is given by

$$f(x) = 2\alpha x e^{-\alpha x^2}, x > 0 \quad [\because \beta = 2]$$

$$\text{Now } P(X > 5) = \int_5^{\infty} 2\alpha x e^{-\alpha x^2} dx$$

$$= \left( -e^{-\alpha x^2} \right)_5^{\infty}$$
$$= e^{-25\alpha}$$

$$\text{Given that } P(X > 5) = e^{-0.25}$$

$$\therefore e^{-25\alpha} = e^{-0.25}$$

$$\therefore \alpha = \frac{1}{100}$$

$$\text{For the Weibull distribution with parameters } \alpha \text{ and } \beta, E(X) = \alpha^{-1/\beta} \sqrt{\left(\frac{1}{\beta} + 1\right)}$$

$$\therefore \text{Required mean} = \left(\frac{1}{100}\right)^{-\frac{1}{2}} \cdot \sqrt{\left(\frac{3}{2}\right)}$$

$$= 10 \times \frac{1}{2} \sqrt{\left(\frac{1}{2}\right)}$$
$$= 5 \sqrt{\pi}.$$

$$\text{Var}(X) = \alpha^{-\frac{2}{\beta}} \left[ \left( \frac{2}{\beta} + 1 \right) - \left\{ \left( \frac{1}{\beta} + 1 \right)^2 \right\} \right]$$

$$= \left(\frac{1}{100}\right)^{-1} \left[ \left( \frac{2}{2} + 1 \right) - \left\{ \left( \frac{3}{2} \right)^2 \right\} \right]$$

$$= 100 \left[ 1 - \left( \frac{1}{2} \sqrt{\pi} \right)^2 \right]$$

$$= 100 \left( 1 - \frac{\pi}{4} \right)$$

## **Exercise:**

Each of the 6 tubes of a radio set has a life length (in years) which may be considered as a RV that follows a Weibull distribution with parameters  $\alpha = 25$  and  $\beta = 2$ . If these tubes function independently of one another, what is the probability that no tube will have to be replaced during the first 2 months of service?

### **Example:**

The marks obtained by a number of students in a certain subject are approximately normally distributed with mean 65 and standard deviation 5. If 3 students are selected at random from this group, what is the probability that at least 1 of them would have scored above 75?

## Solution:

If  $X$  represents the marks obtained by the students,  $X$  follows the distribution  $N(65, 5)$ .

$P(\text{a student scores above } 75)$

$$= P(X > 75) = P\left(\frac{75 - 65}{5} < \frac{X - 65}{5} < \infty\right)$$

$= P(2 < Z < \infty)$ , (where  $Z$  is the standard normal variate)

$$= 0.5 - P(0 < Z < 2)$$

$$= 0.5 - 0.4772, \text{ (from the table of areas)}$$

$$= 0.0228$$

Let  $p = P(\text{a student scores above } 75) = 0.0228$  then  $q = 0.9772$  and  $n = 3$ .

Since  $p$  is the same for all the students, the number  $Y$ , of (successes) students scoring above 75, follows a binomial distribution.

$P(\text{at least 1 student scores above } 75)$

$$= P(\text{at least 1 success})$$

$$= P(Y \geq 1) = 1 - P(Y = 0)$$

$$= 1 - nC_0 \times p^0 q^n$$

$$= 1 - 3C_0 (0.9772)^3$$

$$= 1 - 0.9333$$

$$= 0.0667$$

## **Example:**

In an engineering examination, a student is considered to have failed, secured second class, first class and distinction, according as he scores less than 45%, between 45% and 60%, between 60% and 75% and above 75% respectively. In a particular year 10% of the students failed in the examination and 5% of the students got distinction. Find the percentages of students who have got first class and second class. (Assume normal distribution of marks).

## Solution:

Let  $X$  represent the percentage of marks scored by the students in the examination.

Let  $X$  follow the distribution  $N(\mu, \sigma)$ .

Given:  $P(X < 45) = 0.10$  and  $P(X > 75) = 0.05$

i.e.,  $P\left(-\infty < \frac{X-\mu}{\sigma} < \frac{45-\mu}{\sigma}\right) = 0.10$  and

$$P\left(\frac{75-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \infty\right) = 0.05$$

i.e.,  $P\left(-\infty < Z < \frac{45-\mu}{\sigma}\right) = 0.10$  and

$$P\left(\frac{75-\mu}{\sigma} < Z < \infty\right) = 0.05$$

$$\therefore P\left(0 < Z < \frac{\mu - 45}{\sigma}\right) = 0.40 \text{ and}$$

$$P\left(0 < Z < \frac{75-\mu}{\sigma}\right) = 0.45$$

From the table of areas, we get

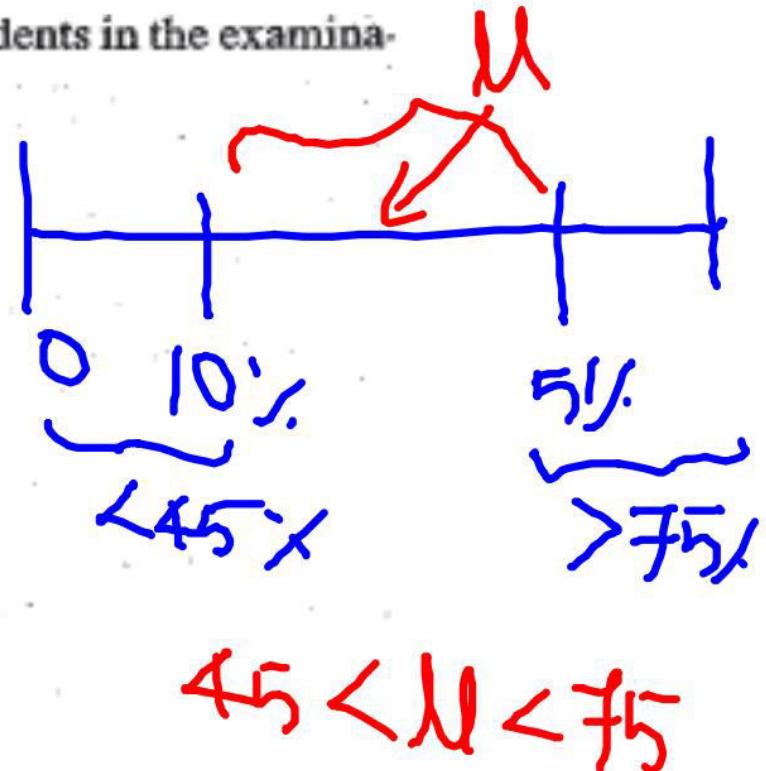
$$\frac{\mu - 45}{\sigma} = 1.28 \text{ and } \frac{75-\mu}{\sigma} = 1.64$$

i.e.,  $\mu - 1.28 \sigma = 45 \quad (1)$

and  $\mu + 1.64 \sigma = 75 \quad (2)$

Solving equations (1) and (2), we get

$$\mu = 58.15 \text{ and } \sigma = 10.28$$



## Solution (Continued):

Now  $P$  (a student gets first class)

$$= P(60 < X < 75)$$

$$= P \left\{ \frac{60 - 58.15}{10.28} < Z < \frac{75 - 58.15}{10.28} \right\}$$

$$= P\{0.18 < Z < 1.64\}$$

$$= P\{0 < Z < 1.64\} - P\{0 < Z < 0.18\}$$

$$= 0.4495 - 0.0714 = 0.3781$$

∴ Percentage of students getting first class = 38 (approximately)

Now percentage of students getting second class

= 100 – (sum of the percentages of students who have failed,  
got first class and got distinction)

$$= 100 - (10 + 38 + 5), \text{ approximately.}$$

$$= 47 \text{ (approximately)}$$

### **Exercise:**

If the actual amount of instant coffee which a filling machine puts into '6-ounce' jars is a RV having a normal distribution with  $SD = 0.05$  ounce and if only 3% of the jars are to contain less than 6 ounces of coffee, what must be the mean fill of these jars?

### **Exercise:**

The local corporation authorities in a certain city install 10,000 electric lamps in the streets of the city with the assumption that the life of lamps is normally distributed. If these lamps have an average life of 1,000 burning hours with a standard deviation of 200 hours, then how many lamps might be expected to fail in the first 800 burning hours and also how many lamps might be expected to fail between 800 and 1,200 burning hours.

### **Exercise:**

The marks obtained by a number of students in a certain subject are assumed to be approximately normally distributed with mean 55 and a SD of 5. If 5 students are taken at random from this set, then what is the probability that 3 of them would have scored marks above 60?



# **MAT2001**

## **Statistics for Engineers**

### **Module 5**

### **Hypothesis Testing I**

## **Syllabus**

### **Hypothesis Testing I:**

Testing of hypothesis - Introduction-Types of errors, critical region, procedure of testing hypothesis- Large sample tests- Z test for Single Proportion, Difference of Proportion, mean and difference of means.

# Introduction

## Population and Sample

Every statistical investigation aims at collecting information about some aggregate or collection of individuals or of their attributes, rather than the individuals themselves. In statistical language, such a collection is called *a population or universe*. For example, we have the population of products turned out by a machine, of lives of electric bulbs manufactured by a company etc. A population is finite or infinite, according as the number of elements is finite or infinite. In most situations, the population may be considered infinitely large. A finite subset of a population is called *a sample* and the process of selection of such samples is called *sampling*. The basic objective of the theory of sampling is to draw inference about the population using the information of the sample.

# Parameters and Statistics

Generally in statistical investigations, our ultimate interest will lie in one or more characteristics possessed by the members of the population. If there is only one characteristic of importance, it can be assumed to be a variable  $x$ . If  $x_i$  be the value of  $x$  for the  $i$ th member of the sample, then  $(x_1, x_2, \dots, x_n)$  are referred to as sample observations. Our primary interest will be to know the values of different statistical measures such as mean and variance of the population distribution of  $x$ . Statistical measures, calculated on the basis of population values of  $x$  are called parameters. Corresponding measures computed on the basis of sample observations are called statistics.

$\theta_0 \rightarrow$  Parameter

$\theta \rightarrow$  Statistic

# Sampling Distribution

If the number of samples is large, the values of the statistic may be classified in the form of a frequency table. The probability distribution of the statistic that would be obtained if the number of samples, each of same size were infinitely large is called the sampling distribution of the statistic. If we adopt random sampling technique that is the most popular and frequently used method of sampling [the discussion of which is beyond the scope of this book], the nature of the sampling distribution of a statistic can be obtained theoretically, using the theory of probability, provided the nature of the population distribution is known.

## Standard Error

Like any other distribution, a sampling distribution will have its mean, standard deviation and moments of higher order. The standard deviation of the sampling distribution of a statistic is of particular importance in tests of Hypotheses and is called the standard error of the statistic.

# Statistical Hypotheses

When we attempt to make decisions about the population on the basis of sample information, we have to make assumptions or guesses about the nature of the population involved or about the value of some parameter of the population. Such assumptions, which may or may not be true, are called statistical hypotheses.

## Null and Alternative Hypotheses

Very often, we set up a hypothesis which assumes that there is no significant difference between the sample statistic and the corresponding population parameter or between two sample statistics. Such a hypothesis of no difference is called a null hypothesis and is denoted by  $H_0$ . A hypothesis that is different from (or complementary to) the null hypothesis is called an alternative hypothesis and is denoted by  $H_1$ . A procedure for deciding whether to accept or to reject a null hypothesis (and hence to reject or to accept the alternative hypothesis respectively) is called the test of hypothesis.

## Significant Difference and Test of Significance

If  $\theta_0$  is a parameter of the population and  $\theta$  is the corresponding sample statistic, usually there will be some difference between  $\theta_0$  and  $\theta$  since  $\theta$  is based on sample observations and is different for different samples. Such a difference which is caused due to sampling fluctuations is called insignificant difference. The difference that arises due to the reason that either the sampling procedure is not purely random or that the sample has not been drawn from the given population is known as significant difference. This procedure of testing whether the difference between  $\theta_0$  and  $\theta$  is significant or not is called the test of significance.

# Critical Region

If we are prepared to reject a null hypothesis when it is true or if we are prepared to accept that the difference between a sample statistic and the corresponding parameter is significant, when the sample statistic lies in a certain region or interval, then that region is called the critical region or region of rejection. The region complementary to the critical region is called the region of acceptance.

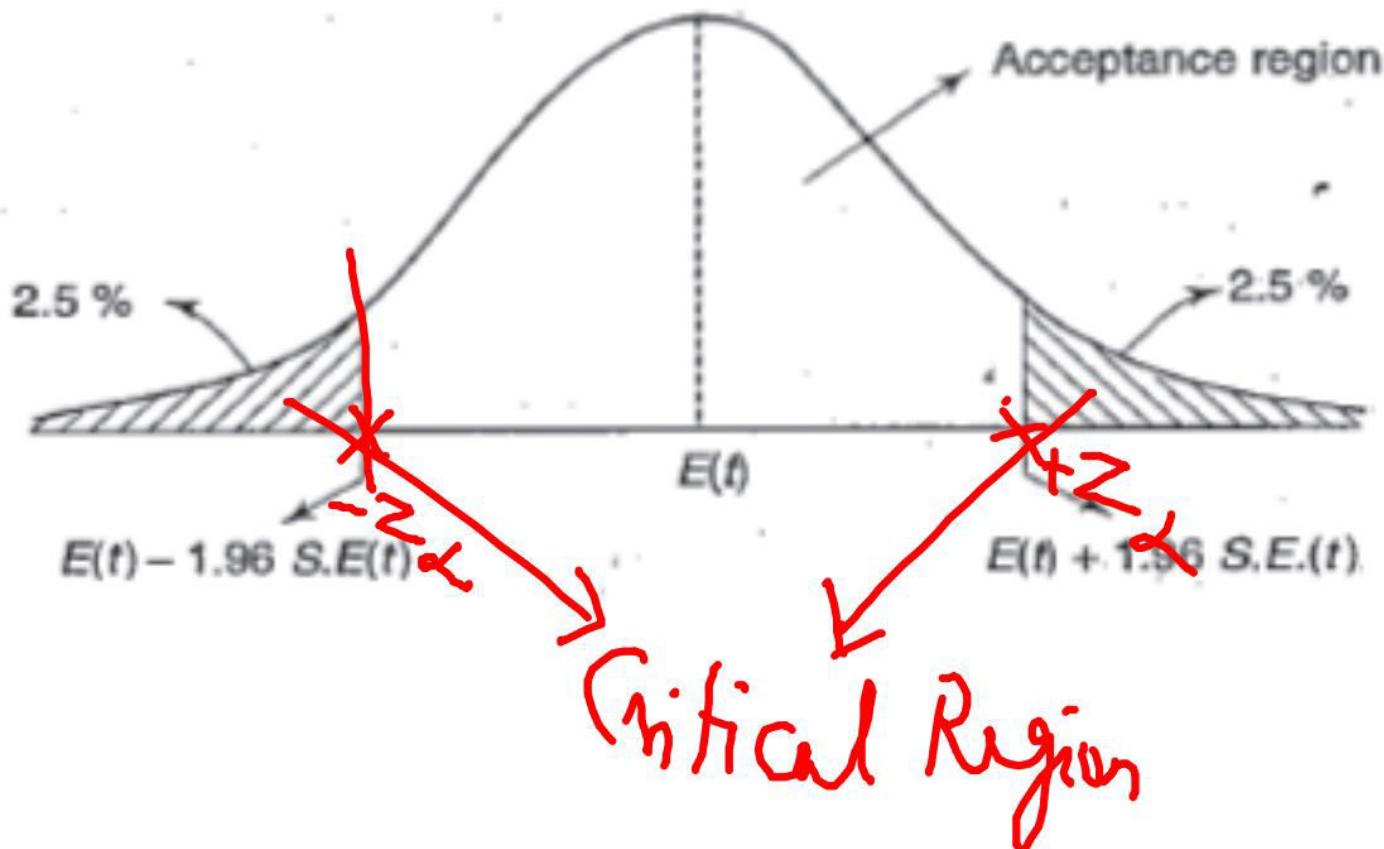
In the case of large samples, the sampling distributions of many statistics tend to become normal distributions. If 't' is a statistic in large samples, then  $t$  follows a normal distribution with mean  $E(t)$ , which is the corresponding population

parameter, and S.D. equal to S.E. ( $t$ ). Hence  $Z = \frac{t - E(t)}{\text{S.E.}(t)}$  is a standard normal variate i.e.,  $Z$  (called *the test statistic*) follows a normal distribution with mean zero and S.D. unity.

$$Z = \frac{X - \mu}{\sigma}$$

# Critical Region

It is known from the study of normal distribution, that the area under the standard normal curve between  $t = -1.96$  and  $t = +1.96$  is 0.95. Equivalently the area under the general normal curve of ' $t$ ' between  $[E(t) - 1.96 \text{ S.E.}(t)]$  and  $[E(t) + 1.96 \text{ S.E.}(t)]$  is 0.95. In other words, 95 per cent of the values of  $t$  will lie between  $[E(t) \mp 1.96 \text{ S.E.}(t)]$  or only 5 per cent of values of  $t$  will lie outside this interval.



# Level of Significance

The probability ' $\alpha$ ' that a random value of the statistic lies in the critical region is called *the level of significance* and is usually expressed as a percentage.

## Note

*The level of significance can also be defined as the maximum probability with which we are prepared to reject  $H_0$  when it is true. In other words, the total area of the region of rejection expressed as a percentage is called the level of significance.*

# Critical Values or Significant Values

The value of the test statistic  $z$  for which the critical region and acceptance region are separated is called *the critical value or the significant value* of  $z$  and denoted by  $z_\alpha$ , when  $\alpha$  is the level of significance. It is clear that the value of  $z_\alpha$  depends not only on  $\alpha$  but also on the nature of alternative hypothesis. -

# Types of Errors

The level of significance is fixed by the investigator and as such it may be fixed at a higher level by his wrong judgement. Due to this, the region of rejection becomes larger and the probability of rejecting a null hypothesis, when it is true, becomes greater. The error committed in rejecting  $H_0$ , when it is really true, is called Type I error. This is similar to a good product being rejected by the consumer and hence Type I error is also known as producer's risk. The error committed in accepting  $H_0$ , when it is false, is called Type II error. As this error is similar to that of accepting a product of inferior quality, it is also known as consumer's risk.

The probabilities of committing Type I and II errors are denoted by  $\alpha$  and  $\beta$  respectively. It is to be noted that the probability  $\alpha$  of committing Type I error is the level of significance.

# One-Tailed and Two-Tailed Tests

If  $\theta_0$  is a population parameter and  $\theta$  is the corresponding sample statistic and if we set up the null hypothesis  $H_0 : \theta = \theta_0$ , then the alternative hypothesis which is complementary to  $H_0$  can be any one of the following:

- (i)  $H_1 : \theta \neq \theta_0$ , i.e.  $\theta > \theta_0$  or  $\theta < \theta_0$
- (ii)  $H_1 : \theta > \theta_0$
- (iii)  $H_1 : \theta < \theta_0$ .

$H_1$  given in (i) is called a two tailed alternative hypothesis, whereas  $H_1$  given in (ii) is called a right-tailed alternative hypothesis and  $H_1$  given in (iii) is called a left-tailed alternative hypothesis.

When  $H_0$  is tested while  $H_1$  is a one-tailed alternative (right or left), the test of hypothesis is called a one-tailed test.

When  $H_0$  is tested while  $H_1$  is two-tailed alternative, the test of hypothesis is called a two-tailed test.

## Table for Z-Test

The critical values for some standard LOS's are given in the following table both for two-tailed and one-tailed tests

Nature of test	LOS $\alpha\%$	1% (.01)	2% (.02)	5% (.05)	10% (.1)
Two-tailed		$ z_\alpha  = 2.58$	$ z_\alpha  = 2.33$	$ z_\alpha  = 1.96$	$ z_\alpha  = 1.645$
Right-tailed		$z_\alpha = 2.33$	$z_\alpha = 2.055$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left-tailed		$z_\alpha = -2.33$	$z_\alpha = -2.055$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

# Procedure for Testing Hypothesis

1. Null hypothesis  $H_0$  is defined.
2. Alternative hypothesis  $H_1$  is also defined after a careful study of the problem and also the nature of the test (whether one-tailed or two tailed) is decided.
3. LOS ' $\alpha$ ' is fixed or taken from the problem if specified and  $z_\alpha$  is noted.
4. The test-statistic  $z = \frac{t - E(t)}{S.E.(t)}$  is computed.
5. Comparison is made between  $|z|$  and  $z_\alpha$ . If  $|z| < z_\alpha$ ,  $H_0$  is accepted or  $H_1$  is rejected, i.e. it is concluded that the difference between  $t$  and  $E(t)$  is not significant at  $\alpha$  % L.O.S.

On the other hand, if  $|z| > z_\alpha$ ,  $H_0$  is rejected or  $H_1$  is accepted, i.e. it is concluded that the difference between  $t$  and  $E(t)$  is significant at  $\alpha$  % L.O.S.

### Procedure for Testing Hypothesis

#### ①. Null Hypothesis:

$$H_0 : \theta = \theta_0$$

$$H_0 : \theta_1 = \theta_2$$

#### ②. Alternative Hypothesis:

$$* H_1 : \theta \neq \theta_0 (\theta_1 \neq \theta_2)$$

$$* H_1 : \theta > \theta_0 \rightarrow \text{Two-Tailed Test}$$

$$* H_1 : \theta < \theta_0 \rightarrow \text{Right One-Tailed Test}$$

$$* H_1 : \theta < \theta_0 \rightarrow \text{Left One-Tailed Test}$$

$$\text{LOS} = \alpha \% = 1\% \text{ (or) } 5\%$$

$$Z_{\alpha \%} = Z_{\text{Tab}} = 0.01 \text{ (or) } 0.05$$

$$④. \text{Test Statistic: } Z_{\text{Cal}} = \frac{t - E(t)}{S.E(t)}$$

#### ⑤. Comparison and Conclusion:

$$(i). |Z_{\text{Cal}}| < |Z_{\text{Tab}}|$$

Then,  $H_0$  is accepted (or)  $H_1$  is rejected.

$$(ii). |Z_{\text{Cal}}| > |Z_{\text{Tab}}|$$

Then,  $H_0$  is rejected (or)  $H_1$  is accepted.

# **Types of Sampling Theory**

## **1. Large Sample:**

**Size of the Sample ( $n$ ) is greater than or equal to 30.**

## **2. Small Sample:**

**Size of the Sample ( $n$ ) is less than 30.**

# **Large Sample Tests (Z-Tests)**

## **Tests of Significance for Large Samples**

It is generally agreed that, if the size of the sample exceeds 30, it should be regarded as a large sample. The tests of significance used for large samples are different from the ones used for small samples for the reason that the following assumptions made for large samples do not hold for small samples

1. The sampling distribution of a statistic is approximately normal, irrespective of whether the distribution of the population is normal or not.
2. Sample statistics are sufficiently close to the corresponding population parameters and hence may be used to calculate the standard error of the sampling distribution.

## Z-Test for Single Proportion

### Test I

*Test of significance of the difference between sample proportion and population proportion.*

The test statistic  $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

**Note** 1. If  $P$  is not known, we assume that  $p$  is nearly equal to  $P$  and hence S.E. ( $p$ )

is taken as  $\sqrt{\frac{pq}{n}}$ . Thus  $z = \frac{p - P}{\sqrt{\frac{pq}{n}}}$ .

2. 95 per cent confidence limits for  $P$  are then given by  $\left| \frac{P - p}{\sqrt{\frac{pq}{n}}} \right| \leq 1.96$ , i.e. they are

$$\left( p - 1.96 \sqrt{\frac{pq}{n}}, p + 1.96 \sqrt{\frac{pq}{n}} \right).$$

## Z-Test for Difference of Proportions

### Test 2

*Test of significance of the difference between two sample proportions.*

Let  $p_1$  and  $p_2$  be the proportions of successes in two large samples of size  $n_1$  and  $n_2$  respectively drawn from the same population or from two population with the same proportion  $P$ .

The test statistic  $z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ .

### Note:

If  $P$  is not known, an unbiased estimate of  $P$  based on both samples, given by

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}, \text{ is used in the place of } P.$$

## Z-Test for Single Mean

### Test 3

*Test of significance of the difference between sample mean and population mean.*

The test statistic  $z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ .

#### Note

1. If  $\sigma$  is not known, the sample S.D. 's' can be used in its place, as  $s$  is nearly equal to  $\sigma$  when  $n$  is large.

2. 95% confidence limits for  $\mu$  are given by  $\frac{|\mu - \bar{X}|}{\sigma / \sqrt{n}} \leq 1.96$ , i.e.

$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ , if  $\sigma$  is known. If  $\sigma$  is not known, then the 95% confidence interval is

$\left( \bar{X} - \frac{1.96 s}{\sqrt{n}}, \bar{X} + \frac{1.96 s}{\sqrt{n}} \right)$

## Z-Test for Difference of Means

### Test 4

*Test of significance of the difference between the means of two samples.*

Let  $\bar{X}_1$  and  $\bar{X}_2$  be the means of two large samples of sizes  $n_1$  and  $n_2$  drawn from two populations (normal or non-normal) with the same mean  $\mu$  and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively.

$$\text{The test statistic } z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (1)$$

## Z-Test for Difference of Means

**Note**

1. If the samples are drawn from the same population, i.e. if  $\sigma_1 = \sigma_2 = \sigma$  then

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

2. If  $\sigma_1$  and  $\sigma_2$  are not known and  $\sigma_1 \neq \sigma_2$ ,  $\sigma_1$  and  $\sigma_2$  can be approximated by the sample S.D.'s  $s_1$  and  $s_2$ . Hence, in such a situation,

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (3) \text{ [from (1)]}$$

3. If  $\sigma_1$  and  $\sigma_2$  are equal and not known, then  $\sigma_1 = \sigma_2 = \sigma$  is approximated by

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}. \text{ Hence, in such a situation,}$$

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left( \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \text{ from (2)}$$

i.e.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}} \quad (4)$$

4. The difference in the denominators of the values of  $z$  given in (3) and (4) may be noted.

### **Example:**

The fatality rate of typhoid patients is believed to be 17.26 per cent. In a certain year 640 patients suffering from typhoid were treated in a metropolitan hospital and only 63 patients died. Can you consider the hospital efficient?

## Solution:

$H_0 : p = P$ , i.e. the hospital is not efficient.  $H_1 : p < P$ .

One-tailed (left-tailed) test is to be used

Let us assume that LOS = 1%.  $\therefore z_\alpha = -2.33$

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}, \text{ where } p = \frac{63}{640} = 0.0984 \text{ and}$$

$$P = 0.1726 \quad \text{and hence} \quad Q = 0.8274.$$

$$z = \frac{0.0984 - 0.1726}{\sqrt{\frac{0.1726 \times 0.8274}{640}}} = -4.96$$

$$\therefore |z| > |z_\alpha|$$

$\therefore$  The difference between  $p$  and  $P$  is significant. i.e.,  $H_0$  is rejected and  $H_1$  is accepted.

i.e. The hospital is efficient in bringing down the fatality rate of typhoid patients.

### **Exercise:**

A salesman in a departmental store claims that at most 60 percent of the shoppers entering the store leaves without making a purchase. A random sample of 50 shoppers showed that 35 of them left without making a purchase. Are these sample results consistent with the claim of the salesman? Use a level of significance of 0.05.

## **Example:**

In a large city  $A$ , 20 per cent of a random sample of 900 school boys had a slight physical defect. In another large city  $B$ , 18.5 percent of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant?

**Solution:**

$$p_1 = 0.2, \quad p_2 = 0.185, \quad n_1 = 900 \quad \text{and} \quad n_2 = 1600$$

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Two tailed test is to be used.

Let L.O.S. be 5%  $\therefore z_\alpha = 1.96$

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

Since  $P$ , the population proportion, is not given, we estimate it as  $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{180 + 296}{900 + 1600} = 0.1904$ .

Using in (1), we have

$$z = \frac{0.2 - 0.185}{\sqrt{0.1904 \times 0.8096 \times \left(\frac{1}{900} + \frac{1}{1600}\right)}} = 0.92$$

$|z| \leq z_\alpha$ . Therefore The difference between  $p_1$  and  $p_2$  is not significant at 5 per cent level.

### **Exercise:**

Before an increase in excise duty on tea, 800 people out of a sample of 1000 were consumers of tea. After the increase in duty, 800 people were consumers of tea in a sample of 1200 persons. Find whether there is significant decrease in the consumption of tea after the increase in duty.

### **Example:**

A sample of 100 students is taken from a large population. The mean height of the students in this sample is 160 cm. Can it be reasonably regarded that, in the population, the mean height is 165 cm, and the S.D. is 10 cm?

## Solution:

A sample of 100 students is taken from a large population. The mean height of the students in this sample is 160 cm. Can it be reasonably regarded that, in the population, the mean height is 165 cm, and the S.D. is 10 cm?

$$\bar{x} = 160, \quad n = 100, \quad \mu = 165 \quad \text{and} \quad \sigma = 10.$$

$H_0$ :  $\bar{x} = \mu$  (i.e. the difference between  $\bar{x}$  and  $\mu$  is not significant)

$H_1$ :  $\bar{x} \neq \mu$ .

Two-tailed test is to be used.

Let LOS be 1%  $\therefore z_\alpha = 2.58$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{160 - 165}{10 / \sqrt{100}} = -5$$

Now  $|z| > z_\alpha$ .

$\therefore$  The difference between  $\bar{x}$  and  $\mu$  is significant at 1% level.

i.e.  $H_0$  is rejected.

i.e. it is not statistically correct to assume that  $\mu = 165$ .

### **Exercise:**

The mean breaking strength of the cables supplied by a manufacturer is 1800 with a S.D. of 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. In order to test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850. Can we support the claim at 1 per cent level of significance?

## **Example:**

A simple sample of heights of 6400 English men has a mean of 170 cm and a S.D. of 6.4 cm, while a simple sample of heights of 1600 Americans has a mean of 172 cm and a S.D. of 6.3 cm. Do the data indicate that Americans are, on the average, taller than the Englishmen?

**Solution:**

$$n_1 = 6400, \bar{x}_1 = 170 \quad \text{and} \quad s_1 = 6.4$$

$$n_2 = 1600, \bar{x}_2 = 172 \quad \text{and} \quad s_2 = 6.3$$

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad \bar{x}_1 = \bar{x}_2,$$

i.e. the samples have been drawn from two different populations with the same mean.

$$H_1 : \bar{x}_1 < \bar{x}_2 \quad \text{or} \quad \mu_1 < \mu_2.$$

Left-tailed test is to be used.

Let LOS be 1%.  $\therefore z_\alpha = -2.33$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

[ $\because \sigma_1 \approx s_1$  and  $\sigma_2 \approx s_2$ . Refer to Note 2 under Test 4]

$$= \frac{170 - 172}{\sqrt{\frac{(6.4)^2}{6400} + \frac{(6.3)^2}{1600}}} = -11.32$$

Now

$$|z| > |z_\alpha|$$

$\therefore$  The difference between  $\bar{x}_1$  and  $\bar{x}_2$  (or  $\mu_1$  and  $\mu_2$ ) is significant at 1% level.

i.e.  $H_0$  is rejected and  $H_1$  is accepted.

i.e. The Americans are, on the average, taller than the Englishmen.

### **Exercise:**

In a random sample of size 500, the mean is found to be 20. In another independent sample of size 400, the mean is 15. Could the samples have been drawn from the same population with S.D. 4?



# **MAT2001**

## **Statistics for Engineers**

### **Module 6**

### **Hypothesis Testing II**

## **Syllabus**

### **Hypothesis Testing II:**

Small sample tests- Student's t-test, F-test- chi-square test- goodness of fit - independence of attributes- Design of Experiments - Analysis of variance - one and two way classifications - CRD- RBD- LSD.

# **Types of Sampling Theory**

## **1. Large Sample:**

**Size of the Sample ( $n$ ) is greater than or equal to 30.**

## **2. Small Sample:**

**Size of the Sample ( $n$ ) is less than 30.**

# **Small Sample Tests**

## **TESTS OF SIGNIFICANCE FOR SMALL SAMPLES**

The tests of significance discussed in the previous section hold good only for large samples, i.e. only when the size of the sample  $n \geq 30$ . When the sample is small, i.e.  $n < 30$ , the sampling distributions of many statistics are not normal, even though the parent populations may be normal. Moreover the assumption of near equality of population parameters and the corresponding sample statistics will not be justified for small samples. Consequently we have to develop entirely different tests of significance that are applicable to small samples.

# Student's $t$ -Distribution $t(v)$

A random variable  $T$  is said to follow student's  $t$ -distribution or simply  $t$ -distribution, if its probability density function is given by

$$f(t) = \frac{1}{\sqrt{v} \beta\left(\frac{v}{2}, \frac{1}{2}\right)} \cdot \left(1 + \frac{t^2}{v}\right)^{-\frac{(v+1)}{2}}, -\infty < t < \infty.$$

$v$  is called the number of degrees of freedom of the  $t$ -distribution.

(Note:  $t$ -distribution was defined by the mathematician W.S.D. Gosset whose pen name is Student.)

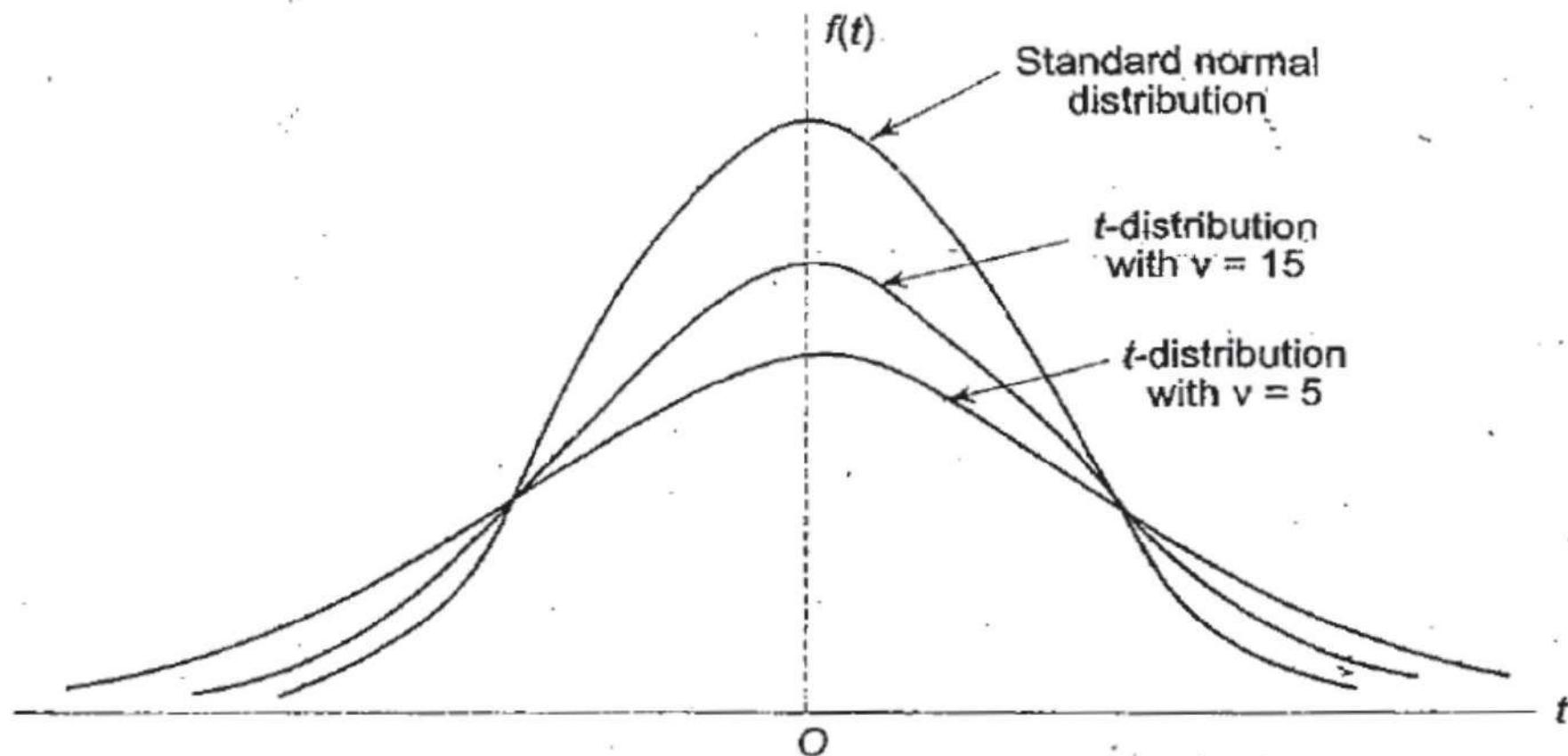
## Properties of *t*-Distribution

1. The probability curve of the *t*-distribution is similar to the standard normal curve and is symmetric about  $t = 0$ , bell-shaped and asymptotic to the *t*-axis.
2. For sufficiently large value of  $n$ , the *t*-distribution tends to the standard normal distribution.
3. The mean of the *t*-distribution is zero.
4. The variance of the *t*-distribution is  $\frac{v}{v - 2}$ , if  $v > 2$  and is greater than 1, but it tends to 1 as  $v \rightarrow \infty$ .

# Uses of $t$ -Distribution

The  $t$ -distribution is used to test the significance of the difference between

1. The mean of a small sample and the mean of the population.
2. The means of two small samples



**Table for *t*-Test*****t* Distribution: Critical Values of *t***

Degrees of freedom	Two-tailed test: One-tailed test: 5%	Significance level					
		10%	5%	2%	1%	0.2%	0.1%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
$\infty$		1.645	1.960	2.326	2.576	3.090	3.291

## *t*-Test for Single Mean

### Test 1

*Test of significance of the difference between sample mean and population mean.*

the test-statistic

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$$

degrees of freedom  $v = n - 1$

$\bar{x}$  - Mean ( $S$ )

$\mu$  - Mean ( $P$ )

$s$  - S.D ( $S$ )

$n$  - size ( $S$ )

## **t-Test for Difference of Means**

### **Test 2**

*Test of significance of the difference between means of two small samples drawn from the same normal population.*

the test-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

degrees of freedom  $v = (n_1 + n_2 - 2)$

**Note:**

If  $\sigma$  is not known, we may assume that  $\sigma \approx \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

with  $(n_1 + n_2 - 2)$  degrees of freedom, the test statistic becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

## **t-Test for Difference of Means**

**Note** 1. If  $n_1 = n_2 = n$  and if the samples are independent i.e., the observations in the two samples are not at all related, then the test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n-1}}} \quad \text{with } v = 2n - 2 \quad (2)$$

## **Paired t-Test**

2. If  $n_1 = n_2 = n$  and if the pairs of values of  $x_1$  and  $x_2$  are associated in some way (or correlated), the formula (2) for  $t$  in Note (1) should not be used. In this case, we shall assume that  $H_0 : \bar{d} (= \bar{x} - \bar{y}) = 0$  and test the significance of the difference between  $\bar{d}$  and 0,

using the test statistic  $t = \frac{\bar{d}}{s / \sqrt{n-1}}$  with  $v = n - 1$ , where  $d_i = x_i - y_i$  ( $i = 1, 2, \dots, n$ ),  $\bar{d} = \bar{x} - \bar{y}$ ; and  $s = S.D. \text{ of } d's = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_i - \bar{d})^2}$ .

### **Example:**

Tests made on the breaking strength of 10 pieces of a metal wire gave the results: 578, 572, 570, 568, 572, 570, 570, ~~572~~, 596 and 584 kg. Test if the mean breaking strength of the wire can be assumed as 577 kg.

**Solution:**

Let us first compute sample mean  $\bar{x}$  and sample S.D.'s and then test if  $\bar{x}$  differs significantly from the population mean  $\mu = 577$ .

$$\text{We take the assumed mean } A = \frac{568 + 596}{2} = 582$$

$$\begin{aligned} d_i &= x_i - A \\ \therefore x_i &= d_i + A \\ \therefore \bar{x} &= \frac{1}{n} \sum x_i = \frac{1}{n} \sum d_i + A \\ &= \frac{1}{10} \times (-68) + 582 = 575.2 \text{ (see Table 9.7 given below)} \end{aligned}$$

**Table 9.7**

$x_i$	$d_i = x_i - A$	$d_i^2$
578	-4	16
572	-10	100
570	-12	144
568	-14	196
572	-10	100
570	-12	144
570	-12	144
572	-10	100
596	14	196
584	2	4
Total	-68	1144

$$\begin{aligned} s^2 &= \frac{1}{n} \sum d_i^2 - \left( \frac{1}{n} \sum d_i \right)^2 \\ &= \frac{1}{10} \times 1144 - \left( \frac{1}{10} \times -68 \right)^2 = 68.16 \end{aligned}$$

$$s = 8.26$$

## Solution (Continued):

Now

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}} = \frac{575.2 - 577}{8.26 / \sqrt{9}}$$
$$= -0.65$$

and

$$v = n - 1 = 9.$$

$$H_0 : \bar{x} = \mu \quad \text{and} \quad H_1 : \bar{x} \neq \mu.$$

Let LOS be 5%. Two tailed test is to be used.

From the  $t$ -table, for  $v = 9$ ,  $t_{5\%} = 2.26$ . Since  $|t| < t_{5\%}$ , the difference between  $\bar{x}$  and  $\mu$  is not significant or  $H_0$  is accepted.  $\therefore$  The mean breaking strength of the wire can be assumed as 577 kg at 5% LOS

### **Exercise:**

The mean lifetime of a sample of 25 bulbs is found as 1550 hours with a S.D. of 120 hours. The company manufacturing the bulbs claims that the average life of their bulbs is 1600 hours. Is the claim acceptable at 5% level of significance?

### **Example:**

Two independent samples of sizes 8 and 7 contained the following values:

Sample I: 19, 17, 15, 21, 16, 18, 16, 14

Sample II: 15, 14, 15, 19, 15, 18, 16

Is the difference between the sample means significant?

**Solution:**

Sample I			Sample II		
$x_1$	$d_1 = x_1 - 18$	$d_1^2$	$x_2$	$d_2 = x_2 - 16$	$d_2^2$
19	1	1	15	-1	1
17	-1	1	14	-2	4
15	-3	9	15	-1	1
21	3	9	19	3	9
16	-2	4	15	-1	1
18	0	0	18	2	4
16	-2	4	16	0	0
14	-4	16			
Total	-8	44	Total	0	20

$$\text{For sample I, } \bar{x}_1 = 18 + \bar{d}_1 = 18 + \frac{1}{8} \sum d_1 \\ = 18 + \frac{1}{8} \times (-8) = 17.$$

$$s_1^2 = \frac{1}{n_1} \sum d_1^2 - \left( \frac{1}{n_1} \sum d_1 \right)^2 \\ = \frac{1}{8} \times 44 - \left( \frac{1}{8} \times -8 \right)^2 = 4.5 \\ \therefore s_1 = 2.12.$$

$$\text{For sample II, } \bar{x}_2 = 16 + \bar{d}_2 = 16 + \frac{1}{7} \sum d_2 = 16 \\ s_2^2 = \frac{1}{n_2} \sum d_2^2 - \left( \frac{1}{n_2} \sum d_2 \right)^2 \\ = \frac{1}{7} \times 20 - \left( \frac{1}{7} \times 0 \right)^2 = 2.857 \\ \therefore s_2 = 1.69$$

## Solution (Continued):

Two-tailed test is to be used. Let LOS be 5 %

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left( \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{17 - 16}{\sqrt{\left( \frac{8 \times 4.5 + 7 \times 2.857}{13} \right) \left( \frac{1}{8} + \frac{1}{7} \right)}} \\ = 0.93$$

Also  $v = n_1 + n_2 - 2 = 13$ .

From the  $t$ -table,  $t_{5\%}$  ( $v = 13$ ) = 2.16

Since  $|t| < t_{5\%}$ ,  $H_0$  is accepted and  $H_1$  is rejected.

i.e. the two sample means do not differ significantly at 5% LOS

### **Exercise:**

The mean height and the S.D. height of eight randomly chosen soliders are 166.9 cm. and 8.29 cm. respectively. The corresponding values of six randomly chosen sailors are 170.3 cm and 8.50 cm. respectively. Based on this data, can we conclude that soldiers are, in general, shorter than sailors?

### **Example:**

The following data relate to the marks obtained by 11 students in two tests, one held at the beginning of a year and the other at the end of the year after intensive coaching. Do the data indicate that the students have benefited by coaching?

Test 1: 19, 23, 16, 24, 17, 18, 20, 18, 21, 19, 20

Test 2: 17, 24, 20, 24, 20, 22, 20, 20, 18, 22, 19

### **Note:**

The given data relate to the marks obtained in two tests by the same set of students. Hence the marks in the two tests can be regarded as correlated and so the *t*-test for paired values should be used.

### Solution:

Let  $d = x_1 - x_2$ ,

where  $x_1, x_2$  denote the marks in the two tests.

Thus the values of  $d$  are 2, -1, -4, 0, -3, -4, 0, -2, 3, -3, 1.

$$\Sigma d = -11 \quad \text{and} \quad \Sigma d^2 = 69$$

$$\therefore \bar{d} = \frac{1}{n} \Sigma d = \frac{1}{11} \times -11 = -1$$

$$s^2 = s_d^2 = \frac{1}{n} \Sigma d^2 - (\bar{d})^2 = \frac{1}{11} \times 69 - (-1)^2 = 5.27$$

$$\therefore s = 2.296$$

$H_0 : \bar{d} = 0$ , i.e. the students have not benefited by coaching;  $H_1 : \bar{d} < 0$  (i.e.  $\bar{x}_1 < \bar{x}_2$ ).

One-tailed test is to be used. Let LOS be 5%

$$t = \frac{\bar{d}}{s / \sqrt{n-1}} = \frac{-1}{2.296 / \sqrt{10}} = -1.38 \quad \text{and} \quad v = 10$$

$t_{5\%} (v=10)$  for one-tailed test =  $t_{10\%} (v=10)$  for two-tailed test = 1.81 (from  $t$ -table).

Now  $|t| < t_{10\%} (v=10)$

$\therefore H_0$  is accepted and  $H_1$  is rejected.

i.e. there is no significant difference between the two sets of marks.

i.e. the students have not benefitted by coaching.

# Snedecor's F-Distribution $F(\gamma_1, \gamma_2)$

A random variable  $F$  is said to follow snedecor's  $F$ -distribution or simply  $F$ -distribution, if its probability density function is given by

$$f(F) = \frac{(\nu_1 / \nu_2)^{\nu_1/2}}{\beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{\left(1 + \frac{\nu_1 F}{\nu_2}\right)^{(\nu_1+\nu_2)/2}}, \quad F > 0.$$

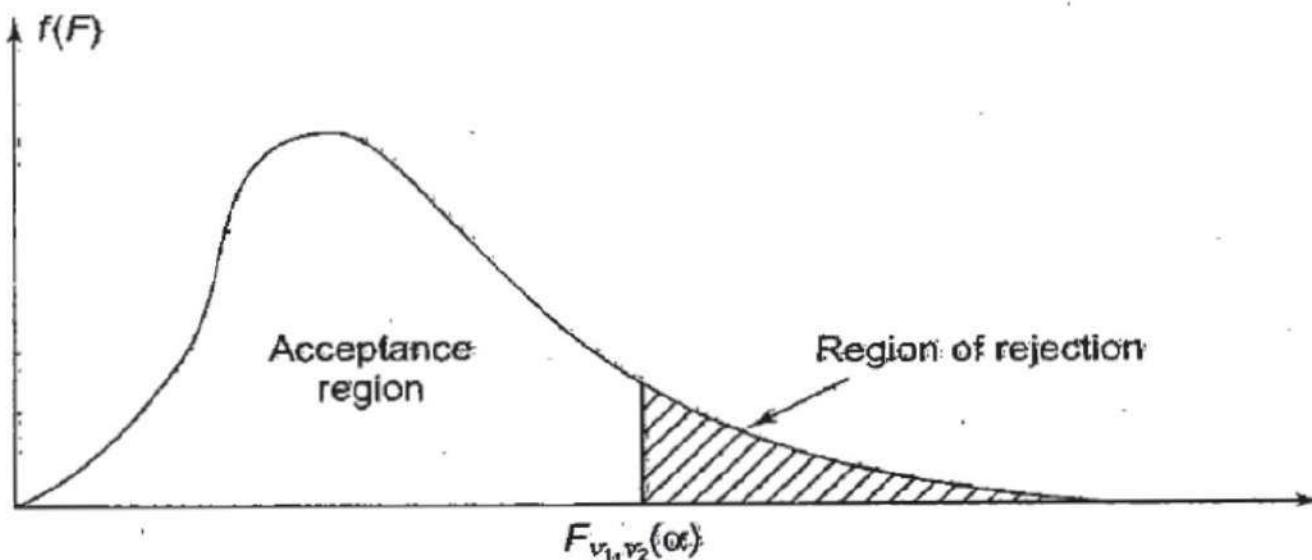
**Note**

(The mathematical variable corresponding to the random variable  $F$  is also taken as  $F$ .)  
 $\nu_1$  and  $\nu_2$  used in  $f(F)$  are the degrees of freedom associated with the  $F$ -distribution.

$\beta(m, n) \rightarrow$  Beta Function

## Properties of *F*-Distribution

1. The probability curve of the *F*-distribution is roughly sketched in Fig.



**Fig.**

2. The square of the *t*-variate with  $n$  degrees of freedom follows a *F*-distribution with 1 and  $n$  degrees of freedom.
3. The mean of the *F*-distribution is  $\frac{v_2}{v_2 - 2}$  ( $v_2 > 2$ ).
4. The variance of the *F*-distribution is

$$\frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad (v_2 > 4).$$

# Uses of *F*-Distribution

*F*-distribution is used to test the equality of the variance of the populations from which two small samples have been drawn.

## Table for F-Test

**F Distribution: Critical Values of F (5% significance level)**

$v_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
$v_2$															
<b>1</b>	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
<b>2</b>	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
<b>3</b>	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
<b>4</b>	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
<b>5</b>	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
<b>6</b>	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
<b>7</b>	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
<b>8</b>	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
<b>9</b>	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
<b>10</b>	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
<b>11</b>	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
<b>12</b>	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
<b>13</b>	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
<b>14</b>	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
<b>15</b>	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
<b>16</b>	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
<b>17</b>	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
<b>18</b>	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
<b>19</b>	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
<b>20</b>	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
<b>21</b>	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
<b>22</b>	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
<b>23</b>	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
<b>24</b>	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
<b>25</b>	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
<b>26</b>	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
<b>27</b>	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
<b>28</b>	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
<b>29</b>	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
<b>30</b>	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
<b>35</b>	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
<b>40</b>	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
<b>50</b>	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
<b>60</b>	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
<b>70</b>	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
<b>80</b>	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
<b>90</b>	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
<b>100</b>	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
<b>120</b>	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
<b>150</b>	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
<b>200</b>	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
<b>250</b>	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
<b>300</b>	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
<b>400</b>	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
<b>500</b>	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
<b>600</b>	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
<b>750</b>	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
<b>1000</b>	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58

**Table for F-Test (Continued)**

		F Distribution: Critical Values of F (5% significance level)									
<i>v<sub>1</sub></i>	<b>25</b>	<b>30</b>	<b>35</b>	<b>40</b>	<b>50</b>	<b>60</b>	<b>75</b>	<b>100</b>	<b>150</b>	<b>200</b>	
<b><i>v<sub>2</sub></i></b>											
<b>1</b>	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68	
<b>2</b>	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49	
<b>3</b>	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54	
<b>4</b>	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65	
<b>5</b>	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39	
<b>6</b>	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69	
<b>7</b>	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25	
<b>8</b>	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95	
<b>9</b>	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73	
<b>10</b>	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56	
<b>11</b>	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43	
<b>12</b>	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32	
<b>13</b>	2.41	2.38	2.36	2.34	2.31	2.30	2.28	2.26	2.24	2.23	
<b>14</b>	2.34	2.31	2.28	2.27	2.24	2.22	2.21	2.19	2.17	2.16	
<b>15</b>	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.10	2.10	
<b>16</b>	2.23	2.19	2.17	2.15	2.12	2.11	2.09	2.07	2.05	2.04	
<b>17</b>	2.18	2.15	2.12	2.10	2.08	2.06	2.04	2.02	2.00	1.99	
<b>18</b>	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.98	1.96	1.95	
<b>19</b>	2.11	2.07	2.05	2.03	2.00	1.98	1.96	1.94	1.92	1.91	
<b>20</b>	2.07	2.04	2.01	1.99	1.97	1.95	1.93	1.91	1.89	1.88	
<b>21</b>	2.05	2.01	1.98	1.96	1.94	1.92	1.90	1.88	1.86	1.84	
<b>22</b>	2.02	1.98	1.96	1.94	1.91	1.89	1.87	1.85	1.83	1.82	
<b>23</b>	2.00	1.96	1.93	1.91	1.88	1.86	1.84	1.82	1.80	1.79	
<b>24</b>	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.80	1.78	1.77	
<b>25</b>	1.96	1.92	1.89	1.87	1.84	1.82	1.80	1.78	1.76	1.75	
<b>26</b>	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76	1.74	1.73	
<b>27</b>	1.92	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72	1.71	
<b>28</b>	1.91	1.87	1.84	1.82	1.79	1.77	1.75	1.73	1.70	1.69	
<b>29</b>	1.89	1.85	1.83	1.81	1.77	1.75	1.73	1.71	1.69	1.67	
<b>30</b>	1.88	1.84	1.81	1.79	1.76	1.74	1.72	1.70	1.67	1.66	
<b>35</b>	1.82	1.79	1.76	1.74	1.70	1.68	1.66	1.63	1.61	1.60	
<b>40</b>	1.78	1.74	1.72	1.69	1.66	1.64	1.61	1.59	1.56	1.55	
<b>50</b>	1.73	1.69	1.66	1.63	1.60	1.58	1.55	1.52	1.50	1.48	
<b>60</b>	1.69	1.65	1.62	1.59	1.56	1.53	1.51	1.48	1.45	1.44	
<b>70</b>	1.66	1.62	1.59	1.57	1.53	1.50	1.48	1.45	1.42	1.40	
<b>80</b>	1.64	1.60	1.57	1.54	1.51	1.48	1.45	1.43	1.39	1.38	
<b>90</b>	1.63	1.59	1.55	1.53	1.49	1.46	1.44	1.41	1.38	1.36	
<b>100</b>	1.62	1.57	1.54	1.52	1.48	1.45	1.42	1.39	1.36	1.34	
<b>120</b>	1.60	1.55	1.52	1.50	1.46	1.43	1.40	1.37	1.33	1.32	
<b>150</b>	1.58	1.54	1.50	1.48	1.44	1.41	1.38	1.34	1.31	1.29	
<b>200</b>	1.56	1.52	1.48	1.46	1.41	1.39	1.35	1.32	1.28	1.26	
<b>250</b>	1.55	1.50	1.47	1.44	1.40	1.37	1.34	1.31	1.27	1.25	
<b>300</b>	1.54	1.50	1.46	1.43	1.39	1.36	1.33	1.30	1.26	1.23	
<b>400</b>	1.53	1.49	1.45	1.42	1.38	1.35	1.32	1.28	1.24	1.22	
<b>500</b>	1.53	1.48	1.45	1.42	1.38	1.35	1.31	1.28	1.23	1.21	
<b>600</b>	1.52	1.48	1.44	1.41	1.37	1.34	1.31	1.27	1.23	1.20	
<b>750</b>	1.52	1.47	1.44	1.41	1.37	1.34	1.30	1.26	1.22	1.20	
<b>1000</b>	1.52	1.47	1.43	1.41	1.36	1.33	1.30	1.26	1.22	1.19	

## Table for F-Test (Continued)

**F Distribution: Critical Values of F (1% significance level)**

$v_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
$v_2$															
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6142.67	6170.10	6191.53	6208.73
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.75	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.08	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.77	9.68	9.61	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.64	2.56	2.50	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.35	2.27	2.20	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.29	2.21	2.14	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.23	2.15	2.09	2.03
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97
250	6.74	4.69	3.86	3.40	3.09	2.87	2.71	2.58	2.48	2.39	2.26	2.15	2.07	2.01	1.95
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92
600	6.68	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.11	2.03	1.96	1.91
750	6.67	4.63	3.81	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.21	2.11	2.02	1.96	1.90
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90

**Table for F-Test (Continued)**

F Distribution: Critical Values of F (1% significance level)

$v_1$	25	30	35	40	50	60	75	100	150	200
$v_2$										
<b>1</b>	6239.83	6260.65	6275.57	6286.78	6302.52	6313.03	6323.56	6334.11	6344.68	6349.97
<b>2</b>	99.46	99.47	99.47	99.47	99.48	99.48	99.49	99.49	99.49	99.49
<b>3</b>	26.58	26.50	26.45	26.41	26.35	26.32	26.28	26.24	26.20	26.18
<b>4</b>	13.91	13.84	13.79	13.75	13.69	13.65	13.61	13.58	13.54	13.52
<b>5</b>	9.45	9.38	9.33	9.29	9.24	9.20	9.17	9.13	9.09	9.08
<b>6</b>	7.30	7.23	7.18	7.14	7.09	7.06	7.02	6.99	6.95	6.93
<b>7</b>	6.06	5.99	5.94	5.91	5.86	5.82	5.79	5.75	5.72	5.70
<b>8</b>	5.26	5.20	5.15	5.12	5.07	5.03	5.00	4.96	4.93	4.91
<b>9</b>	4.71	4.65	4.60	4.57	4.52	4.48	4.45	4.41	4.38	4.36
<b>10</b>	4.31	4.25	4.20	4.17	4.12	4.08	4.05	4.01	3.98	3.96
<b>11</b>	4.01	3.94	3.89	3.86	3.81	3.78	3.74	3.71	3.67	3.66
<b>12</b>	3.76	3.70	3.65	3.62	3.57	3.54	3.50	3.47	3.43	3.41
<b>13</b>	3.57	3.51	3.46	3.43	3.38	3.34	3.31	3.27	3.24	3.22
<b>14</b>	3.41	3.35	3.30	3.27	3.22	3.18	3.15	3.11	3.08	3.06
<b>15</b>	3.28	3.21	3.17	3.13	3.08	3.05	3.01	2.98	2.94	2.92
<b>16</b>	3.16	3.10	3.05	3.02	2.97	2.93	2.90	2.86	2.83	2.81
<b>17</b>	3.07	3.00	2.96	2.92	2.87	2.83	2.80	2.76	2.73	2.71
<b>18</b>	2.98	2.92	2.87	2.84	2.78	2.75	2.71	2.68	2.64	2.62
<b>19</b>	2.91	2.84	2.80	2.76	2.71	2.67	2.64	2.60	2.57	2.55
<b>20</b>	2.84	2.78	2.73	2.69	2.64	2.61	2.57	2.54	2.50	2.48
<b>21</b>	2.79	2.72	2.67	2.64	2.58	2.55	2.51	2.48	2.44	2.42
<b>22</b>	2.73	2.67	2.62	2.58	2.53	2.50	2.46	2.42	2.38	2.36
<b>23</b>	2.69	2.62	2.57	2.54	2.48	2.45	2.41	2.37	2.34	2.32
<b>24</b>	2.64	2.58	2.53	2.49	2.44	2.40	2.37	2.33	2.29	2.27
<b>25</b>	2.60	2.54	2.49	2.45	2.40	2.36	2.33	2.29	2.25	2.23
<b>26</b>	2.57	2.50	2.45	2.42	2.36	2.33	2.29	2.25	2.21	2.19
<b>27</b>	2.54	2.47	2.42	2.38	2.33	2.29	2.26	2.22	2.18	2.16
<b>28</b>	2.51	2.44	2.39	2.35	2.30	2.26	2.23	2.19	2.15	2.13
<b>29</b>	2.48	2.41	2.36	2.33	2.27	2.23	2.20	2.16	2.12	2.10
<b>30</b>	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.13	2.09	2.07
<b>35</b>	2.35	2.28	2.23	2.19	2.14	2.10	2.06	2.02	1.98	1.96
<b>40</b>	2.27	2.20	2.15	2.11	2.06	2.02	1.98	1.94	1.90	1.87
<b>50</b>	2.17	2.10	2.05	2.01	1.95	1.91	1.87	1.82	1.78	1.76
<b>60</b>	2.10	2.03	1.98	1.94	1.88	1.84	1.79	1.75	1.70	1.68
<b>70</b>	2.05	1.98	1.93	1.89	1.83	1.78	1.74	1.70	1.65	1.62
<b>80</b>	2.01	1.94	1.89	1.85	1.79	1.75	1.70	1.65	1.61	1.58
<b>90</b>	1.99	1.92	1.86	1.82	1.76	1.72	1.67	1.62	1.57	1.55
<b>100</b>	1.97	1.89	1.84	1.80	1.74	1.69	1.65	1.60	1.55	1.52
<b>120</b>	1.93	1.86	1.81	1.76	1.70	1.66	1.61	1.56	1.51	1.48
<b>150</b>	1.90	1.83	1.77	1.73	1.66	1.62	1.57	1.52	1.46	1.43
<b>200</b>	1.87	1.79	1.74	1.69	1.63	1.58	1.53	1.48	1.42	1.39
<b>250</b>	1.85	1.77	1.72	1.67	1.61	1.56	1.51	1.46	1.40	1.36
<b>300</b>	1.84	1.76	1.70	1.66	1.59	1.55	1.50	1.44	1.38	1.35
<b>400</b>	1.82	1.75	1.69	1.64	1.58	1.53	1.48	1.42	1.36	1.32
<b>500</b>	1.81	1.74	1.68	1.63	1.57	1.52	1.47	1.41	1.34	1.31
<b>600</b>	1.80	1.73	1.67	1.63	1.56	1.51	1.46	1.40	1.34	1.30
<b>750</b>	1.80	1.72	1.66	1.62	1.55	1.50	1.45	1.39	1.33	1.29
<b>1000</b>	1.79	1.72	1.66	1.61	1.54	1.50	1.44	1.38	1.32	1.28

## F-Test for Differences of Population Variances

*F-test of significance of the difference between population variances and F-table.*

To test the significance of the difference between population variances, we shall first find their estimates,  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$  based on the sample variances  $s_1^2$  and  $s_2^2$  and then test their equality. It is known that  $\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$  with the number of degree of freedom  $v_1 = n_1 - 1$  and  $\hat{\sigma}_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$  with the number of degrees of freedom  $v_2 = n_2 - 1$ .

It is also known that  $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$  follows a *F*-distribution with  $v_1$  and  $v_2$  degrees of freedom. If  $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$ , then  $F = 1$ . Hence our aim is to find how far any observed value of  $F$  can differ from unity due to fluctuations of sampling.

Note:  $F > 1$

**Example:**

A sample of size 13 gave an estimated population variance of 3.0, while another sample of size 15 gave an estimate of 2.5. Could both samples be from populations with the same variance?

Soln/.

$$\begin{array}{c|c|c} n_1 = 13 & \sigma_1^2 = 3.0 & v_1 = n_1 - 1 \\ n_2 = 15 & \sigma_2^2 = 2.5 & v_2 = n_2 - 1 \end{array}$$

$$① H_0: \sigma_1^2 = \sigma_2^2$$

$$② H_1: \sigma_1^2 \neq \sigma_2^2$$

$$③ \alpha = 5\%$$

$$d = 5\%$$

$$F_{Tab} = F_{5\%}(12, 14) = ?$$

$$④ F_{cal} = \frac{\sigma_1^2}{\sigma_2^2} = ?$$

$$\begin{array}{c} \text{Computation \& Conclusion} \\ | F_{cal} \underset{?}{\underset{>}{\sim}} |F_{Tab}| ? \end{array}$$

### Solution:

$$n_1 = 13, \quad \hat{\sigma}_1^2 = 3.0 \quad \text{and} \quad v_1 = 12$$

$$n_2 = 15, \quad \hat{\sigma}_2^2 = 2.5 \quad \text{and} \quad v_2 = 14.$$

$H_0: \hat{\sigma}_1^2 = \hat{\sigma}_2^2$ , i.e. The two samples have been drawn from populations with the same variance.

$H_1: \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$ . Let L.O.S. be 5%

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{3.0}{2.5} = 1.2$$

$$v_1 = 12 \quad \text{and} \quad v_2 = 14.$$

$F_{5\%}(v_1 = 12, v_2 = 14) = 2.53$ , from the  $F$ -table.

$F < F_{5\%}$ .  $\therefore H_0$  is accepted

i.e. the two samples could have come from two normal populations with the same variance.

**Example:**

Two independent samples of eight and seven items respectively had the following values of the variable.

Sample 1 : 9, 11, 13, 11, 15, 9, 12, 14

Sample 2 : 10, 12, 10, 14, 9, 8, 10

Do the two estimates of population variance differ significantly at 5% level of significance?

Soln.

$$\begin{array}{c|c|c} n_1 = 8 & \bar{x}_1 = ? & s_1 = ? \\ n_2 = 7 & \bar{x}_2 = ? & s_2 = ? \end{array}$$

$$\therefore \sigma_1^2 = \frac{n_1 \cdot s_1^2}{(n_1 - 1)} \quad \& \quad \sigma_2^2 = \frac{n_2 \cdot s_2^2}{(n_2 - 1)}$$

① H<sub>0</sub>:  $\sigma_1^2 = \sigma_2^2$

② H<sub>1</sub>:  $\sigma_1^2 \neq \sigma_2^2$

③ Log  $\alpha = \alpha = 5\%$

F<sub>Tab</sub> = F<sub>α=5%</sub>(v<sub>1</sub>=7, v<sub>2</sub>=6) = ?

④ Test statistic

$$F_{cal} = \frac{\sigma_1^2}{\sigma_2^2} \text{ or } \frac{\sigma_2^2}{\sigma_1^2}$$

⑤ Comparison and Conclusion

$$|F_{cal}| < |F_{Tab}|$$

**Solution:**

For the first sample,  $\Sigma x_1 = 94$  and  $\Sigma x_1^2 = 1138$

$$\therefore s_1^2 = \frac{1}{n_1} \sum x_1^2 - \left( \frac{1}{n_1} \sum x_1 \right)^2$$

$$= \frac{1}{8} \times 1138 - \left( \frac{1}{8} \times 94 \right)^2 = 4.19$$

For the second sample,  $\Sigma x_2 = 73$  and  $\Sigma x_2^2 = 785$

$$\therefore s_2^2 = \frac{1}{n_2} \sum x_2^2 - \left( \frac{1}{n_2} \sum x_2 \right)^2$$

$$= \frac{1}{7} \times 785 - \left( \frac{1}{7} \times 73 \right)^2 = 3.39$$

$$\hat{\sigma}_1^2 = \frac{n_1}{n_1 - 1} s_1^2 = 4.79 \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{n_2}{n_2 - 1} s_2^2 = 3.96$$

since  $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$ ,  $v_1 = 7$  and  $v_2 = 6$

$$H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2 \quad \text{and} \quad H_1 : \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$$

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{4.79}{3.96} = 1.21$$

$F_{5\%}(v_1 = 7, v_2 = 6) = 4.21$ , from the  $F$ -table. Since  $F < F_{5\%}$ ,  $H_0$  is accepted.  
i.e.  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$  do not differ significantly at 5% level of significance.

**Example:**

Two samples of sizes nine and eight gave the sums of squares of deviations from their respective means equal to 160 and 91 respectively. Can they be regarded as drawn from the same normal population?

Soln.

$$\begin{array}{l} n_1 = 9 \\ n_2 = 8 \end{array}$$

$$\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 = 160$$

$$\Rightarrow n_1 s_1^2 = 160 \Rightarrow s_1^2 = \frac{160}{n_1-1}$$

$$\sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2 = 91$$

$$\Rightarrow n_2 s_2^2 = 91 \Rightarrow s_2^2 = \frac{91}{n_2-1}$$

①.  $H_0: \sigma_1^2 = \sigma_2^2$

②.  $H_1: \sigma_1^2 \neq \sigma_2^2$

③.  $\chi^2$ ,  $F_{tab}$

④.  $F_{cal} = \frac{\sigma_1^2}{\sigma_2^2} \text{ (or) } \frac{\sigma_2^2}{\sigma_1^2}$

⑤. Comparison and Conclusion

### Solution:

$$n_1 = 9, \quad \sum(x_i - \bar{x})^2 = 160, \quad \text{i.e. } n_1 s_1^2 = 160$$

$$n_2 = 8, \quad \sum(y_i - \bar{y})^2 = 91, \quad \text{i.e. } n_2 s_2^2 = 91$$

$$\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{1}{8} \times 160 = 20; \quad \hat{\sigma}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{1}{7} \times 91 = 13$$

Since  $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$ ,  $v_1 = n_1 - 1 = 8$  and  $v_2 = n_2 - 1 = 7$

$$H_0: \hat{\sigma}_1^2 = \hat{\sigma}_2^2 \quad \text{and} \quad H_1: \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2.$$

Let the LOS be 5%

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{20}{13} = 1.54$$

$F_{5\%}(v_1 = 8, v_2 = 7) = 3.73$ , from the  $F$ -table.

Since  $F < F_{5\%}$ ,  $H_0$  is accepted.

i.e. the two samples could have come from two normal populations with the same variance.

## Exercise:

Two random samples gave the following data:

	Size	Mean	Variance
Sample I	8	9.6	1.2
Sample II	11	16.5	2.5

Can we conclude that the two samples have been drawn from the same normal population?

### **Exercise:**

The nicotine contents in two random samples of tobacco are given below.

Sample I : 21    24    25    26    27

Sample II : 22    27    28    30    31    36.

Can you say that the two samples came from the same population?

# Chi-square Distribution

$\chi^2(\gamma)$

Definition:

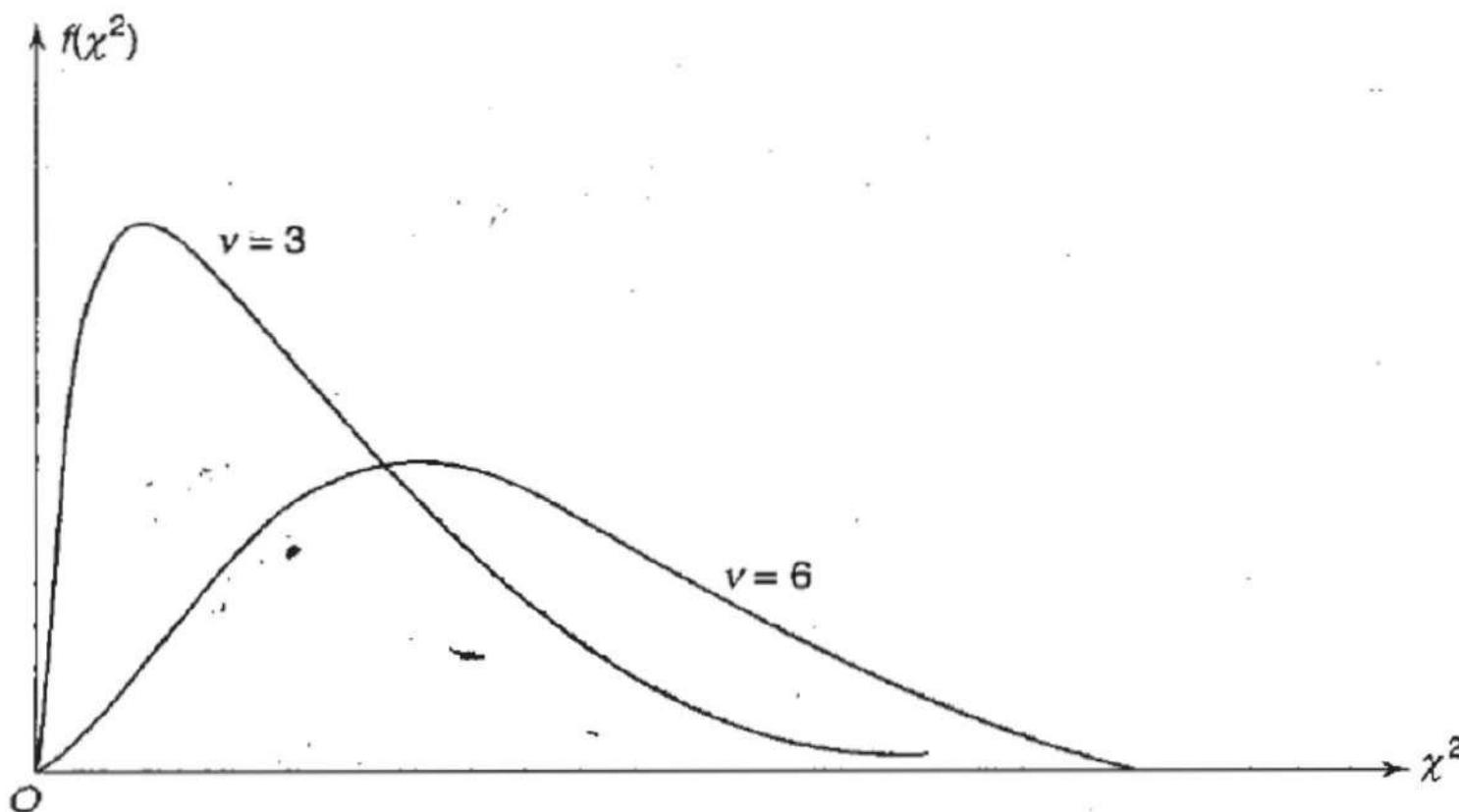
The pdf is,

$$f(\chi^2) = \frac{1}{2^{v/2} \sqrt{\left(\frac{v}{2}\right)}} \cdot (\chi^2)^{v/2 - 1} e^{-\chi^2/2}$$

$0 < \chi^2 < \infty$ , where  $v$  is the number of degrees of freedom.

## Properties of $\chi^2$ -Distribution

1. A rough sketch of the probability curve of the  $\chi^2$ -distribution for  $v=3$  and  $v=6$  is given in Fig.
2. As  $v$  becomes smaller and smaller, the curve is skewed more and more to the right. As  $v$  increases, the curve becomes more and more symmetrical.
3. The mean and variance of the  $\chi^2$ -distribution are  $v$  and  $2v$  respectively.



**Fig.**

4. As  $n$  tends to  $\infty$ , the  $\chi^2$ -distribution becomes a normal distribution.

## **Uses of $\chi^2$ -Distribution**

1.  $\chi^2$ -distribution is used to test the goodness of fit. i.e., it is used to judge whether a given sample may be reasonably regarded as a simple sample from a certain hypothetical population.
2. It is used to test the independence of attributes. i.e. If a population is known to have two attributes (or traits), then  $\chi^2$ -distribution is used to test whether the two attributes are associated or independent, based on a sample drawn from the population.

## Table for Chi-square Test

$df$	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

## $\chi^2$ -Test of Goodness of Fit

On the basis of the hypothesis assumed about the population, we find the expected frequencies  $E_i (i = 1, 2, \dots, n)$ , corresponding to the observed frequencies

$O_i (i = 1, 2, \dots, n)$  such that

$$\sum E_i = \sum O_i$$

It is known that

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

follows approximately a  $\chi^2$ -distribution with degrees of freedom equal to the number of independent frequencies. In order to test the goodness of fit, we have to determine how far the differences between  $O_i$  and  $E_i$  can be attributed to fluctuations of sampling and when we can assert that the differences are large enough to conclude that the sample is not a simple sample from the hypothetical population. In other words, we have to determine how large a value of  $\chi^2$  we can get so as to assume that the sample is a simple sample from the hypothetical population.

Note:  $v = n - 1$

If the calculated  $\chi^2 < \chi^2_v(\alpha)$ , we will accept the null hypothesis  $H_0$  which assumes that the given sample is one drawn from the hypothetical population, i.e. we will conclude that the difference between the observed and expected frequencies is not significant at  $\alpha$  % LOS If  $\chi^2 > \chi^2_v(\alpha)$ , we will reject  $H_0$  and conclude that the difference is significant.

## **Conditions for the Validity of $\chi^2$ -Test**

1. The number of observations  $N$  in the sample must be reasonably large, say  $\geq 50$ .
2. Individual frequencies must not be too small, i.e.  $O_i \geq 10$ . In case  $O_i < 10$ , it is combined with the neighbouring frequencies, so that the combined frequency is  $\geq 10$ .
3. The number of classes  $n$  must be neither too small nor too large i.e.,  $4 \leq n \leq 16$ .

## $\chi^2$ -Test of Independence of Attributes

If the population is known to have two major attributes  $A$  and  $B$ , then  $A$  can be divided into  $m$  categories  $A_1, A_2, \dots, A_m$  and  $B$  can be divided into  $n$  categories  $B_1, B_2, \dots, B_n$ . Accordingly the members of the population and hence those of the sample can be divided into  $mn$  classes. In this case, the sample data may be presented in the form of a matrix containing  $m$  rows and  $n$  columns and hence  $mn$  cells and showing the observed frequencies  $O_{ij}$ , in the various cells, where  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ .  $O_{ij}$  means the number of observed frequencies possessing the attributes  $A_i$  and  $B_j$ . The matrix or tabular form of the sample data, called an  $(m \times n)$  contingency table is given below:

$A \setminus B$	$B_1$	$B_2$	-	$B_j$	-	$B_n$	Row Total
$A_1$	$O_{11}$	$O_{12}$	-	$O_{1j}$	-	$O_{1n}$	$O_{1*}$
$A_2$	$O_{21}$	$O_{22}$	-	$O_{2j}$	-	$O_{2n}$	$O_{2*}$
:	-	-	-	-	-	-	-
$A_i$	$O_{i1}$	$O_{i2}$	-	$O_{ij}$	-	$O_{in}$	$O_{i*}$
:	-	-	-	-	-	-	-
$A_m$	$O_{m1}$	$O_{m2}$	-	$O_{mj}$	-	$O_{mn}$	$O_{m*}$
Column Total	$O_{*1}$	$O_{*2}$	-	$O_{*j}$	-	$O_{*n}$	$N$

Now, based on the null hypothesis  $H_0$  i.e. the assumption that the two attributes  $A$  and  $B$  are independent, we compute the expected frequencies  $E_{ij}$  for various cells, using the following formula  $E_{ij} = \frac{O_{i*} \cdot O_{*j}}{N}$ ,  $i = 1, 2, \dots, m$ ; and  $j = 1, 2, \dots, n$

i.e.

$$E_{ij} = \left\{ \begin{array}{l} \left( \text{Total of observed frequencies in the } i^{\text{th}} \text{ row} \right) \times \\ \left( \text{total of observed frequencies in the } j^{\text{th}} \text{ column} \right) \\ \hline \text{Total of all cell frequencies} \end{array} \right\}$$

Then we compute  $\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \left\{ \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right\}$

The number of degrees of freedom for this  $\chi^2$  computed from the  $(m \times n)$  contingency table is  $v = (m - 1)(n - 1)$ .

If  $\chi^2 < \chi^2_v(\alpha)$ ,  $H_0$  is accepted at  $\alpha$  % LOS i.e. the attributes  $A$  and  $B$  are independent.

If  $\chi^2 > \chi^2_v(\alpha)$ ,  $H_0$  is rejected at  $\alpha$  % LOS i.e.  $A$  and  $B$  are not independent.

**Example:**

Table gives the number of air-craft accidents that occurred during the various days of a week. Test whether the accidents are uniformly distributed over the week.

Day:	Mon	Tues	Wed	Thu	Fri	Sat	Total
No. of accidents:	15	19	13	12	16	15	90

Soln.  $N = 90 \times n = 6$ .

- ①.  $H_0$ : The accidents are distributed uniformly.
- ②.  $H_1$ : Not uniformly distributed.
- ③.  $\lambda_{0.05} = \alpha = 2.7 = 0.02$

$$\chi^2_{Tab} = \chi^2_{5\%}(V=5) = ?$$

- ④. Test Statistic:

$$\chi^2_{Cal} = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i}$$

$$E_i = \frac{90}{6} = 15; \quad i=1, 2, \dots, 6$$

[Since uniform distribution]

- ⑤. Comparison and Conclusion

### Solution:

$H_0$ : Accidents occur uniformly over the week.

Total number of accidents = 90

Based on  $H_0$ , the expected number of accidents on any day =  $\frac{90}{6} = 15$ .

$O_i$	:	15	19	13	12	16	15
$E_i$	:	15	15	15	15	15	15

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{1}{15} (0 + 16 + 4 + 9 + 1 + 0) = 2.$$

Since  $\sum E_i = \sum O_i$ ,  $v = 6 - 1 = 5$

From the  $\chi^2$ -table,  $\chi^2_{5\%} (v = 5) = 11.07$ .

Since  $\chi^2 < \chi^2_{5\%}$ ,  $H_0$  is accepted.

i.e. accidents may be regarded to occur uniformly over the week.

## **Example:**

Theory predicts that the proportion of beans in four groups  $A$ ,  $B$ ,  $C$ ,  $D$  should be  $9 : 3 : 3 : 1$ . In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experiment support the theory?

### Solution:

$H_0$ : The experiment supports the theory, i.e. the numbers of beans in the four groups are in the ratio 9 : 3 : 3 : 1

Based on  $H_0$ , the expected numbers of beans in the four groups are as follows

$$E_i : \frac{9}{16} \times 1600, \quad \frac{3}{16} \times 1600, \quad \frac{3}{16} \times 1600, \quad \frac{1}{16} \times 1600$$

i.e.  $E_i : 900, \quad 300, \quad 300, \quad 100$

$$O_i : 882, \quad 313, \quad 287, \quad 118$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{18^2}{900} + \frac{13^2}{300} + \frac{13^2}{300} + \frac{18^2}{100} = 4.73$$

Since  $\sum E_i = \sum O_i, v = 4 - 1 = 3$

From the  $\chi^2$ -table,  $\chi^2_{5\%}(v=3) = 7.82$

Since  $\chi^2 < \chi^2_{5\%}$ ,  $H_0$  is accepted.

i.e. the experimental data support the theory.

### **Example:**

Fit a binomial distribution for the following data and also test the goodness of fit.

$x:$	0	1	2	3	4	5	6	Total
$f:$	5	18	28	12	7	6	4	80

### Solution:

To find the binomial distribution  $N(q + p)^n$ , which fits the given data, we require  $p$ .

We know that the mean of the binomial distribution is  $np$ , from which we can find  $p$ . Now the mean of the given distribution is found out and is equated to  $np$ .

$x$ :	0	1	2	3	4	5	6	Total
$f$ :	5	18	28	12	7	6	4	80
$fx$ :	0	18	56	36	28	30	24	192

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{192}{80} = 2.4$$

i.e.  $np = 2.4$  or  $6p = 2.4$ , since the maximum value taken by  $x$  is  $n$ .

$$\therefore p = 0.4 \text{ and hence } q = 0.6$$

$\therefore$  The expected frequencies are given by the successive terms in the expansion of  $80(0.6 + 0.4)^6$ .

Thus  $E_i$ : 3.73, 14.93, 24.88, 22.12, 11.06, 2.95, 0.33

Converting the  $E_i$ 's into whole number such that  $\sum E_i = \sum O_i = 80$ , we get

$E_i$ : 4 15 25 22 11 3 0

Let us now proceed to test the goodness of binomial fit.

$O_i$ : 5 18 28 12 7 6 4

## Solution (Continued):

The first class is combined with the second and the last two classes are combined with the last but second class in order to make the expected frequency in each class greater than or equal to 10. Thus, after regrouping, we have,

$$E_i : \quad 19 \quad \quad 25 \quad \quad 22 \quad \quad 14$$

$$O_i : \quad 23 \quad \quad 28 \quad \quad 12 \quad \quad 17$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{4^2}{19} + \frac{3^2}{25} + \frac{10^2}{22} + \frac{3^2}{14} = 6.39$$

We have used the given sample to find

$$\sum E_i (= \sum O_i) \text{ and } p \text{ through its mean.}$$

Hence

$$v = n - k$$

$$= 4 - 2 = 2$$

$$\chi^2_{5\%} (v = 2) = 5.99, \text{ from the } \chi^2\text{-table.}$$

Since  $\chi^2 > \chi^2_{5\%}$ ,  $H_0$ , which assumes that the given distribution is approximately a binomial distribution, is rejected. i.e., the binomial fit for the given distribution is not satisfactory.

### **Exercise:**

The following table shows the distribution of digits in the numbers chosen at random from a telephone directory:

**Table**

Digit:	0	1	2	3	4	5	6	7	8	9	Total
Frequency:	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory.

### Exercise:

According to genetic theory, children having one parent of blood type  $M$  and the other of blood type  $N$  will always be one of the three types- $M$ ,  $MN$  and  $N$  and the average proportions of these types will be  $1 : 2 : 1$ . Out of 300 children, having one  $M$  parent and one  $N$  parent, 30 per cent were found to be of type  $M$ , 45 per cent of type  $MN$  and the remaining of type  $N$ . Test the genetic theory by  $\chi^2$ -test.

### Exercise:

Fit a Poisson distribution for the following distribution and also test the goodness of fit.

$x$ :	0	1	2	3	4	5	Total
$f$ :	142	156	69	27	5	1	400

### **Example:**

The following data are collected on two characters

	<i>Smokers</i>	<i>Non-smokers</i>
Literates :	83	57
Illiterates :	45	68

Based on this, can you say that there is no relation between smoking and literacy?

**Solution:**

$H_0$ : Literacy and smoking habit are independent

	Smokers	Non-smokers	Total
Literates	83	57	140
Illiterates	45	68	113
Total	128	125	253

$O$	$E$	$E$ (rounded)	$(O - E)^2/E$
83	$\frac{128 \times 140}{253} = 70.83$	71	$122/71 = 2.03$
57	$\frac{125 \times 140}{253} = 69.17$	69	$122/69 = 2.09$
45	$\frac{128 \times 113}{253} = 57.17$	57	$122/57 = 2.53$
68	$\frac{125 \times 113}{253} = 55.83$	56	$122/56 = 2.57$
$\chi^2 = 9.22$			

$$v = (m - 1)(n - 1) \\ = (2 - 1)(2 - 1) = 1.$$

From the  $\chi^2$ -table,  $\chi^2_{5\%}(v=1) = 3.84$

Since  $\chi^2 > \chi^2_{5\%}$ ,  $H_0$  is rejected.

i.e. there is some association between literacy and smoking.

## Example:

A total number of 3759 individuals were interviewed in a public opinion survey on a political proposal. Of them, 1872 were men and the rest women. 2257 individuals were in favour of the proposal and 917 were opposed to it. 243 men were undecided and 442 women were opposed to the proposal. Do you justify or contradict the hypothesis that there is no association between sex and attitude?

**Table:**

	<i>Favoured</i>	<i>Opposed</i>	<i>Undecided</i>	<i>Total</i>
Men	1154	475	243	1872
Women	1103	442	342	1887
Total	2257	917	585	3759

**Solution:**

$H_0$  : Sex and attitude are independent, i.e. there is no association between sex and attitude.

$O$	$E$ (rounded $E$ )	$(O - E)^2/E$
1154	$\frac{1872 \times 2257}{3759} \approx 1124$	$302/1124 = 0.80$
475	$\frac{1872 \times 917}{3759} \approx 457$	$182/457 = 0.71$
243	$\frac{1872 \times 585}{3759} \approx 291$	$482/291 = 7.92$
1103	$\frac{1887 \times 2257}{3759} \approx 1133$	$302/1133 = 0.79$
442	$\frac{1887 \times 917}{3759} \approx 460$	$182/460 = 0.70$
342	$\frac{1887 \times 585}{3759} \approx 294$	$482/294 = 7.84$
$v = (3 - 1)(2 - 1) = 2$		$\chi^2 = 18.76$

From the  $\chi^2$ -table,  $\chi^2_{5\%}(v = 2) = 5.99$

Since  $\chi^2 > \chi^2_{5\%}$ ,  $H_0$  is rejected.

That is, sex and attitude are not independent i.e. there is some association between sex and attitude.

## **Exercise:**

A survey of radio listeners' preference for two types of music under various age groups gave the following information.

**Table**

<i>Type of music</i>	<i>Age group</i>		
	19-25	26-35	Above 36
Carnatic music :	80	60	90
Film music :	210	325	44
Indifferent :	16	45	132

Is preference for type of music influenced by age?

# Design of Experiments

**B**y 'experiment', we mean collection of data (which usually consist of a series of measurement of some feature of an object) for a scientific investigation, according to certain specified sampling procedures. Statistics provides not only the principles and the basis for the proper planning of the experiments but also the methods for proper interpretation of the results of the experiment.

In the beginning, the study of the design of experiments was associated only with agricultural experimentation. The need to save time and money has led to a study of ways to obtain maximum information with the minimum cost and labour. Such motivations resulted in the subsequent acceptance and wide use of the design of experiments and the related analysis of variance techniques in all fields of scientific experimentation. In this chapter we consider some aspects of experimental design briefly and analysis of data from such experiments using analysis of variance techniques.

# Aim of the Design of Experiments

A statistical experiment in any field is performed to verify a particular hypothesis. For example, an agricultural experiment may be performed to verify the claim that a particular manure has got the effect of increasing the yield of paddy. Here the quantity of the manure used and the amount of yield are the two variables involved directly. They are called *experimental variables*. Apart from these two, there are other variables such as the fertility of the soil, the quality of the seed used and the amount of rainfall, which also affect the yield of paddy. Such variables are called *extraneous variables*. The main aim of the design of experiments is to control the extraneous variables and hence to minimise the experimental error so that the results of the experiments could be attributed only to the experimental variables.

# **Basic Principles of the Experimental Design**

In order to achieve the objective mentioned above, the following three principles are adopted while designing the experiments— (1) randomisation, (2) replication and (3) local control.

## 1. Randomisation

As it is not possible to eliminate completely the contribution of extraneous variables to the value of the response variable (the amount of yield of paddy), we try to control it by randomisation. The group of experimental units (plots of the same size) in which the manure is used is called the *experimental group* and the other group of plots in which the manure is not used and which will provide a basis for comparison is called the *control group*. If any information regarding the extraneous variables and the nature and magnitude of their effect on the response variable in question is not available, we resort to randomisation. That is, we select the plots for *the* experimental and control groups in a random manner, which provides the most effective way of eliminating any unknown bias in the experiment.

## **2. Replication**

In a comparative experiment, in which the effects of different manures on the yield are studied, each manure is used in more than one plot. In other words, we resort to replication which means repetition. It is essential to carry out more than one test on each manure in order to estimate the amount of the experimental error and hence to get some idea of the precision of the estimates of the manure effects.

### 3. Local Control

To provide adequate control of extraneous variables, another essential principle used in the experimental design is the local control. This includes techniques such as grouping, blocking and balancing of the experimental units used in the experimental design. By *grouping*, we mean combining sets of homogeneous plots into groups, so that different manures may be used in different groups. The number of plots in different groups need not necessarily be the same. By *blocking*, we mean assigning the same number of plots in different blocks. The plots in the same block may be assumed to be relatively homogeneous. We use as many manures as the number of plots in a block in a random manner. By *balancing*, we mean adjusting the procedures of grouping, blocking and assigning the manures in such a manner that a balanced configuration is obtained.

# **Basic Designs of Experiment**

## **1. Completely Randomised Design (CRD)**

**Analysis of Variance (ANOVA) for One Way Classification**

## **2. Randomised Block Design (RBD)**

**Analysis of Variance (ANOVA) for Two Way Classification**

## **3. Latin Square Design (LSD)**

**Analysis of Variance (ANOVA) for Three Way Classification**

# 1. Completely Randomised Design (CRD)

Let us suppose that we wish to compare ' $h$ ' treatments (use of ' $h$ ' different manures) and there are  $n$  plots available for the experiment.

Let the  $i$ th treatment be replicated (repeated)  $n_i$  times, so that  $n_1 + n_2 + \dots + n_h = n$ .

The plots to which the different treatments are to be given are found by the following randomisation principle. The plots are numbered from 1 to  $n$  serially.  $n$  identical cards are taken, numbered from 1 to  $n$  and shuffled thoroughly. The numbers on the first  $n_1$  cards drawn randomly give the numbers of the plots to which the first treatment is to be given. The numbers on the next  $n_2$  cards drawn at random give the numbers of the plots to which the second treatment is to be given and so on. This design is called a completely randomised design, which is used when the plots are homogeneous or the pattern of heterogeneity of the plots is unknown.

## 2. Randomised Block Design (RBD)

Let us consider an agricultural experiment using which we wish to test the effect of ' $k$ ' fertilizing treatments on the yield of a crop. We assume that we know some information about the soil fertility of the plots. Then we divide the plots into ' $h$ ' blocks, according to the soil fertility, each block containing ' $k$ ' plots. Thus the plots in each block will be of homogeneous fertility as far as possible.

Within each block, the ' $k$ ' treatments are given to the ' $k$ ' plots in a perfectly random manner, such that each treatment occurs only once in any block. But the same ' $k$ ' treatments are repeated from block to block. This design is called Randomised Block Design.

### 3. Latin Square Design (LSD)

We consider an agricultural experiment, in which  $n^2$  plots are taken and arranged in the form of an  $n \times n$  square, such that the plots in each row will be homogeneous as far as possible with respect to one factor of classification, say, soil fertility and plots in each column will be homogeneous as far as possible with respect to another factor of classification, say, seed quality.

Then  $n$  treatments are given to these plots such that each treatment occurs only once in each row and only once in each column. The various possible arrangements obtained in this manner are known as Latin squares of order  $n$ . This design of experiment is called the Latin Square Design.

## Comparison of RBD and LSD

1. The number of replications of each treatment is equal to the number of treatments in LSD, whereas there is no such restrictions on treatments and replication in RBD.
2. LSD can be performed on a square field, while RBD can be performed either on a square field or a rectangular field.
3. LSD is known to be suitable for the case when the number of treatments is between 5 and 12, whereas RBD can be used for any number of treatments.
4. The main advantage of LSD is that it controls the effect of two extraneous variables, whereas RBD controls the effect of only one extraneous variable. Hence the experimental error is reduced to a larger extent in LSD than in RBD.

# **Analysis of Variance (ANOVA)**

The analysis of variance is a widely used technique developed by R.A. Fisher. It enables us to divide the total variation (represented by variance) in a group into parts which are ascribable to different factors and a residual random variation which could not be accounted for by any of these factors. The variation due to any specific factor is compared with the residual variation for significance by applying the F-test, with which the reader is familiar. The details of the procedure will be explained in the sequel.

# Analysis of Variance (ANOVA) for One Way Classification

*ANOVA table for one factor of classification*

<i>Source of variation</i> (S.V.)	<i>Sum of squares</i> (S.S.)	<i>Degree of freedom</i> (d.f.)	<i>Mean square</i> (M.S.)	<i>Variance ratio</i> (F)
Between classes	$Q_1$	$h - 1$	$Q_1 / (h - 1)$	$\frac{Q_1 / (h - 1)}{Q_2 / (N - h)}$ (OR)
Within classes	$Q_2$	$N - h$	$Q_2 / (N - h)$	$\frac{Q_2 / (N - h)}{Q_1 / (h - 1)}$
<i>Total</i>	$Q$	$N - 1$	-	-

**Note** For calculating  $\mathcal{Q}$ ,  $\mathcal{Q}_1$ ,  $\mathcal{Q}_2$ , the following computational formulas may be used:

$$\begin{aligned}\mathcal{Q} &= N \left\{ \frac{1}{N} \sum \sum x_{ij}^2 - \bar{x}^2 \right\} \\ &= N \left\{ \frac{1}{N} \sum \sum x_{ij}^2 - \left( \frac{1}{N} \sum \sum x_{ij} \right)^2 \right\} \\ &= \sum \sum x_{ij}^2 - \frac{T^2}{N}, \text{ where } T = \sum \sum x_{ij}\end{aligned}$$

Similarly, for the  $i$ th class,

$$\sum_j (x_{ij} - \bar{x}_i)^2 = \sum_j x_{ij}^2 - \frac{T_i^2}{n_i}, \text{ where } T_i = \sum_j x_{ij}.$$

$$\mathcal{Q}_2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \sum_i \sum_j x_{ij}^2 - \sum_i \frac{T_i^2}{n_i}$$

Hence

$$\mathcal{Q}_1 = \mathcal{Q} - \mathcal{Q}_2$$

$$= \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

# Analysis of Variance (ANOVA) for Two Way Classification

*The ANOVA table for the two factors of classifications*

S.V.	S.S.	d.f.	M.S.	F
Between rows	$Q_1$	$h - 1$	$Q_1 / (h - 1)$	$\left[ \frac{Q_1 / (h - 1)}{Q_3 / (h - 1)(k - 1)} \right]^{\pm 1}$
Between columns	$Q_2$	$k - 1$	$Q_2 / (k - 1)$	$\left[ \frac{Q_2 / (k - 1)}{Q_3 / (h - 1)(k - 1)} \right]^{\pm 1}$
Residual	$Q_3$	$(h - 1)(k - 1)$	$Q_3 / (h - 1)(k - 1)$	-
Total	$Q$	$hk - 1$	-	-

**Note**

The following working formulas that can be easily derived may be used to compute  $\mathcal{Q}$ ,  $\mathcal{Q}_1$ ,  $\mathcal{Q}_2$  and  $\mathcal{Q}_3$ :

$$1. \mathcal{Q} = \sum \sum x_{ij}^2 - \frac{T^2}{N}, \text{ where } T = \sum \sum x_{ij}$$

$$2. \mathcal{Q}_1 = \frac{1}{k} \sum T_i^2 - \frac{T^2}{N}, \text{ where } T_i = \sum_{j=1}^k x_{ij}$$

$$3. \mathcal{Q}_2 = \frac{1}{h} \sum T_j^2 - \frac{T^2}{N}, \text{ where } T_j = \sum_{i=1}^h x_{ij}$$

$$4. \mathcal{Q}_3 = \mathcal{Q} - \mathcal{Q}_1 - \mathcal{Q}_2$$

It may be verified that  $\sum_i T_i = \sum_j T_j = T$ .

# Analysis of Variance (ANOVA) for Three Way Classification

*The ANOVA table for three factors of classification*

S.V.	S.S.	d.f.	M.S.	F
Between rows	$Q_1$	$n - 1$	$Q_1 / (n - 1) = M_1$	$\left(\frac{M_1}{M_4}\right)^{\pm 1}$
Between columns	$Q_2$	$n - 1$	$Q_2 / (n - 1) = M_2$	$\left(\frac{M_2}{M_4}\right)^{\pm 1}$
Between letters	$Q_3$	$n - 1$	$Q_3 / (n - 1) = M_3$	$\left(\frac{M_3}{M_4}\right)^{\pm 1}$
Residual	$Q_4$	$(n - 1)(n - 2)$	$Q_4 / (n - 1)(n - 2) = M_4$	-
Total	$Q$	$n^2 - 1$	-	-

**Note**

The following working formulas may be used to compute the  $\mathcal{Q}$ 's:

$$1. \mathcal{Q} = \sum \sum x_{ij}^2 - \frac{T^2}{n^2}, \text{ where } T = \sum \sum x_{ij}$$

$$2. \mathcal{Q}_1 = \frac{1}{n} \sum T_i^2 - \frac{T^2}{n^2}, \text{ where } T_i = \sum_{j=1}^n x_{ij}$$

$$3. \mathcal{Q}_2 = \frac{1}{n} \sum T_j^2 - \frac{T^2}{n^2}, \text{ where } T_j = \sum_{i=1}^n x_{ij}$$

$$4. \mathcal{Q}_3 = \frac{1}{n} \sum T_k^2 - \frac{T^2}{n^2}, \text{ where } T_k \text{ is the sum of all } x_{ij}'s \text{ receiving the } k^{th} \text{ treatment.}$$

$$5. \mathcal{Q}_4 = \mathcal{Q} - \mathcal{Q}_1 - \mathcal{Q}_2 - \mathcal{Q}_3.$$

$$\text{Also } T = \sum_i T_i = \sum_j T_j = \sum_k T_k$$

## **Example:**

A completely randomised design experiment with 10 plots and 3 treatments gave the following results:

Plot No.	:	1	2	3	4	5	6	7	8	9	10
Treatment	:	A	B	C	A	C	C	A	B	A	B
Yield	:	5	4	3	7	5	1	3	4	1	7

Analyse the results for treatment effects.

## Solution:

Rearranging the data according to the treatments, we have the following table:

Treatment	Yield from plots ( $x_{ij}$ )				$T_i$	$T_i^2$	$n_i$	$\frac{T_i^2}{n_i}$
A	5	7	3	1	16	256	4	64
B	4	4	7	-	15	225	3	75
C	3	5	1	-	9	81	3	27
	<i>Total</i>		$T = 40$		$N = 10$		166	

$$\begin{aligned}\sum \sum x_{ij}^2 &= (25 + 49 + 9 + 1) + (16 + 16 + 49) + (9 + 25 + 1) \\ &= 84 + 81 + 35 = 200\end{aligned}$$

$$Q = \sum \sum x_{ij}^2 - \frac{T^2}{N} = 200 - \frac{40^2}{10} = 200 - 160 = 40$$

$$Q_1 = \sum \frac{T_i^2}{n_i} - \frac{T^2}{N} = 166 - 160 = 6$$

$$Q_2 = Q - Q_1 = 40 - 6 = 34$$

## Solution (Continued):

*ANOVA table*

<i>S.V.</i>	<i>S.S.</i>	<i>d.f.</i>	<i>M.S.</i>	<i>F<sub>0</sub></i>
Between classes (treatments)	$Q_1 = 6$	$h - 1 = 2$	3.0	$\frac{4.86}{3.0}$
Within classes	$Q_2 = 34$	$N - h = 7$	4.86	= 1.62
<i>Total</i>	$Q = 40$	$N - 1 = 9$	-	-

From the *F*-table,  $F_{5\%} (v_1 = 2, v_2 = 7) = 19.35$

We note that  $F_0 < F_{5\%}$

Let  $H_0$  : The treatments do not differ significantly.

∴ The null hypothesis is accepted.

i.e., the treatments are not significantly different.

### **Example:**

Three varieties of a crop are tested in a randomised block design with four replications, the layout being as given below: The yields are given in kilograms. Analyse for significance

C48	A51	B52	A49
A47	B49	C52	C51
B49	C53	A49	B50

## Solution:

Rewriting the data such that the rows represent the blocks and the columns represent the varieties of the crop (as assumed in the discussion of analysis of variance for two factors of classification), we have the following table:

*Crops*

<i>Blocks</i>	<i>A</i>	<i>B</i>	<i>C</i>
1	47	49	48
2	51	49	53
3	49	52	52
4	49	50	51

We shift the origin to 50 and work out with the new values of  $x_{ij}$ .

### Solution (Continued):

*Crops*

Blocks	A	B	C	$T_i$	$T^2_i / k \sum_j x_{ij}^2$
1	-3	-1	-2	-6	$36/3 = 12$
2	1	-1	3	3	$9/3 = 3$
3	-1	2	2	3	$9/3 = 3$
4	-1	0	1	0	$0/3 = 0$
$T_j$	-4	0	4	$T = 0$	$\sum \frac{T_i^2}{k} = 18$
$T^2_j / h$	$\frac{16}{4} = 4$	$\frac{0}{4} = 0$	$\frac{16}{4} = 4$	$\sum \frac{T_j^2}{h} = 8$	
$\sum_i x_{ij}^2$	12	6	18	36	

$$Q = \sum \sum x_{ij}^2 - \frac{T^2}{N} = 36 - \frac{0^2}{12} = 36$$

$$Q_1 = \frac{1}{k} \sum T_i^2 - \frac{T^2}{N} = 18 - 0 = 18$$

$$Q_2 = \frac{1}{h} \sum T_j^2 - \frac{T^2}{N} = 8 - 0 = 8$$

$$Q_3 = Q - Q_1 - Q_2 = 36 - 18 - 8 = 10$$

## Solution (Continued):

*ANOVA table*

S.V.	S.S.	d.f.	M.S.	$F_0$
Between rows (blocks)	$Q_1 = 18$	$h - 1 = 3$	6	$\frac{6}{1.67} = 3.6$
Between columns (crops)	$Q_2 = 8$	$k - 1 = 2$	4	$\frac{4}{1.67} = 2.4$
Residual	$Q_3 = 10$	$(h - 1)(k - 1) = 6$	1.67	-
Total	$Q = 36$	$hk - 1 = 11$	-	-

From  $F$ -tables,  $F_{5\%}$  ( $v_1 = 3, v_2 = 6$ ) = 4.76 and  $F_{5\%}$  ( $v_1 = 2, v_2 = 6$ ) = 5.14  
 Considering the difference between rows, we see that  $F_0 (= 3.6) < F_{5\%} (= 4.76)$   
 Hence the difference between the rows is not significant. ( $H_0$  is accepted) viz.,  
 the blocks do not differ significantly with respect to the yield.

Considering the difference between columns, we see that  $F_0 (= 2.4) < F_{5\%}$   
 ( $= 5.14$ )

Hence the difference between the columns is not significant. ( $H_0$  is accepted)  
 viz., the varieties of crop do not differ significantly with respect to the yield.

### **Exercise:**

The following table shows the lives in hours of four brands of electric lamps:

*Brand*

A : 1610, 1610, 1650, 1680, 1700, 1720, 1800

B : 1580, 1640, 1640, 1700, 1750

C : 1460, 1550, 1600, 1620, 1640, 1660, 1740, 1820

D : 1510, 1520, 1530, 1570, 1600, 1680

Perform an analysis of variance and test the homogeneity of the mean lives of the four brands of lamps.

### **Exercise:**

In order to determine whether there is significant difference in the durability of makes of computers, samples of size 5 are selected from each make and the frequency of repair during the first year of purchase is observed. The results are as follows :

<i>Makes</i>		
A	B	C
5	8	7
6	10	3
8	11	5
9	12	4
7	4	1

In view of the above data, what conclusion can you draw?

### **Exercise:**

The following data represent the number of units of production per day turned out by 5 different workers using 4 different types of machines:

Workers:	<i>Machine Type</i>			
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

- (a) Test whether the five men differ with respect to mean productivity.
- (b) Test whether the mean productivity is the same for the four different machine types.

## Exercise:

Four doctors each test four treatments for a certain disease and observe the number of days each patient takes to recover. The results are as follows (recovery time in days)

Doctor	Treatment			
	1	2	3	4
A	10	14	19	20
B	11	15	17	21
C	9	12	16	19
D	8	13	17	20

Discuss the difference between (a) doctors and (b) treatments.



# **MAT2001**

## **Statistics for Engineers**

### **Module 7**

### **Reliability**

## **Syllabus**

### **Reliability:**

Basic concepts- Hazard function-Reliabilities of series and parallel systems- System Reliability - Maintainability-Preventive and repair maintenance- Availability.

# Reliability

## Definition

The definitions given below with respect to a component hold good for a system or a device also.

If a component is put into operation at some specified time, say  $t = 0$ , and if  $T$  is the time until it fails or ceases to function properly,  $T$  is called the *life length* or *time to failure* of the component. Obviously  $T (\leq 0)$  is a continuous random variable with some probability density function  $f(t)$ . Then the *reliability* or *reliability function* of the component at time ' $t$ ', denoted by  $R(t)$ , is defined as

$$\begin{aligned} R(t) &= P(T > t) \text{ or } 1 - P(T \leq t) \\ &= 1 - F(t), \end{aligned} \tag{1}$$

where  $F(t)$  is the cumulative distribution function of  $T$ , given by

$$F(t) = \int_0^t f(t) dt$$

Thus

$$R(t) = 1 - \int_0^t f(t) dt = \int_t^\infty f(t) dt \tag{2}$$

## Reliability

### Definition

$T$  - CRV represents the life-time of a system/comp.  
 $0 \leq T < \infty$  with pdf  $f(t)$ .

Reliability (Reliability Function)

$$\underline{R(t)} = P(T > t) = \int_t^{\infty} f(t) dt$$

$$R(t) = 1 - P(T \leq t)$$

$$R(t) = 1 - \int_0^t f(t) dt$$

$$\frac{d(R(t))}{dt} = R'(t) = 0 - f(t)$$

$$\Rightarrow f(t) = -R'(t)$$

**Note:**

Since  $F(0) = 0$  and  $F(\infty) = 1$  by the property of cdf,  $R(0) = 1$  and  $R(\infty) = 0$  i.e.,  $0 \leq R(t) \leq 1$ . Also since  $\frac{d}{dt} F(t) = f(t)$ ,

we get  $f(t) = -\frac{dR(t)}{dt}$  (3)

# Hazard Function (Instantaneous Failure Rate Function)

the *instantaneous failure rate* or *hazard function* of the component, denoted by  $\lambda(t)$ .

Thus

$$\lambda(t) = \frac{f(t)}{R(t)}$$

Now, using (3) in (4), we have

$$-\frac{R'(t)}{R(t)} = \lambda(t)$$

Note:

$$R(t) = e^{-\int_0^t \lambda(t) dt}$$

$$f(t) = \lambda(t) e^{-\int_0^t \lambda(t) dt}$$

$$-\int f(t) dt = -R'(t) \quad (4)$$

(5)

## Mean and Variance

### Mean Time To Failure (MTTF)

$$\text{MTTF} = E(T) = \int_0^{\infty} t f(t) dt.$$

$$\text{MTTF} = \int_0^{\infty} R(t) dt$$

### Variance

$$\text{Var}(T) = \sigma_T^2 = E\{T - E(T)\}^2 \text{ or } E(T^2) - \{E(T)\}^2$$

$$= \int_0^{\infty} t^2 f(t) dt - (\text{MTTF})^2$$

# Conditional Reliability of a System or Component

For a Wear-in Period / Burn-in Period / After to Warranty Period:

It is defined as

$$R(t/T_0) = P\{T > T_0 + t | T > T_0\}$$

$$= \frac{P\{T > T_0 + t\}}{P\{T > T_0\}} = \frac{R(T_0 + t)}{R(T_0)}$$

$$= \frac{e^{-\int_0^{T_0+t} \lambda(t) dt}}{e^{-\int_0^{T_0} \lambda(t) dt}} = e^{-\left[ \int_0^{T_0+t} \lambda(t) dt - \int_0^{T_0} \lambda(t) dt \right]}$$

$$= e^{-\int_{T_0}^{T_0+t} \lambda(t) dt} = e^{-\int_{T_0}^{T_0+t} \lambda(t) dt}$$

# Special Failure Probability Distributions

## 1. Exponential Distribution

If the time to failure  $T$  follows an exponential distribution with parameter  $\lambda$ , then its pdf is given by

$$f(t) = \lambda e^{-\lambda t}, t \geq 0$$

$$\underline{R(t)} = \int_t^{\infty} \lambda e^{-\lambda u} dt = [-e^{-\lambda u}]_t^{\infty} = \underline{e^{-\lambda t}}$$

$$\underline{\lambda(t)} = \frac{f(t)}{R(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \underline{\lambda}$$

This means that when the failure distribution is an exponential distribution with parameter  $\lambda$ , the failure rate at any time is a constant, equal to  $\lambda$ . Conversely when  $\lambda(t) =$  a constant  $\lambda$ , we get

$$f(t) = \lambda \cdot e^{-\int_0^t \lambda dt} = \lambda e^{-\lambda t}, t \geq 0$$

Due to this property, the exponential distribution is often referred to as constant failure rate distribution in reliability contexts.

## 1. Exponential Distribution

$$\text{MTTF} = E(T) = \frac{1}{\lambda}$$

$$\text{Var}(T) = \sigma_T^2 = \frac{1}{\lambda^2}$$

$$R(t/T_0) = \frac{R(T_0 + t)}{R(T_0)} = \frac{e^{-\lambda(T_0 + t)}}{e^{-\lambda T_0}}$$

$$= e^{-\lambda t} = e^{-\lambda t}$$

$$R(t) = e^{-\lambda t}$$

This means that the time to failure of a component is not dependent on how long the component has been functioning. In other words the reliability of the component for the next 1000 hours, say, is the same regardless of whether the component is brand new or has been operating for several hours. This property is known as the memoryless property of the constant failure rate distribution.

## 2. Weibull Distribution

The *pdf* of the Weibull distribution was defined as

$$f(t) = \alpha \beta t^{\beta-1} e^{-\alpha t^\beta}, t \geq 0$$

An alternative form of Weibull's *pdf* is

$$f(t) = \frac{\beta}{\theta} \left( \frac{t}{\theta} \right)^{\beta-1} \exp \left[ -\left( \frac{t}{\theta} \right)^\beta \right], \theta > 0, \beta > 0, t \geq 0$$

by putting  $\alpha = \frac{1}{\theta^\beta}$   $\beta$  is called the shape parameter

and  $\theta$  is called the characteristic life or scale parameter of the Weibull's distribution

$$R(t) = \int_t^\infty \frac{\beta}{\theta} \cdot \left( \frac{t}{\theta} \right)^{\beta-1} \exp \left[ -\left( \frac{t}{\theta} \right)^\beta \right] dt$$

$$= \int_x^\infty e^{-x} dx, \text{ on putting } \left( \frac{t}{\theta} \right)^\beta = x$$

$$= e^{-x} = \exp \left[ -\left( \frac{t}{\theta} \right)^\beta \right]$$

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{\beta}{\theta} \cdot \left( \frac{t}{\theta} \right)^{\beta-1}$$

## 2. Weibull Distribution

$$\text{MTTF} = E(T) = \theta \Gamma\left(1 + \frac{1}{\beta}\right)$$

$$\text{Var}(T) = \sigma^2_T = \theta^2 \left\{ \Gamma\left(1 + \frac{2}{\beta}\right) - \left[ \Gamma\left(1 + \frac{1}{\beta}\right) \right]^2 \right\}$$

$$R(t/T_0) = \frac{R(t + T_0)}{R(T_0)}$$

$$= \frac{\exp\left[-\left(\frac{t + T_0}{\theta}\right)^\beta\right]}{\exp\left[-\left(\frac{T_0}{\theta}\right)^\beta\right]}$$

$$= \exp\left[-\left(\frac{t + T_0}{\theta}\right)^\beta + \left(\frac{T_0}{\theta}\right)^\beta\right]$$

### 3. Normal Distribution

$$N(\mu, \sigma)$$

If the time to failure  $T$  follows a normal distribution  $N(\mu, \sigma)$  its pdf is given by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right], -\infty < t < \infty$$

In this case,  $MTTF = E(T) = \mu$  and

$$\text{Var}(T) = \sigma_T^2 = \sigma^2.$$

$R(t) = \int_t^\infty f(t) dt$  is found out by expressing the integral in terms of standard

normal integral and using the normal tables.

then found out  $\lambda(t)$

#### 4. Log-Normal Distribution $L-N(t_M, S)$

If  $X = \log T$  follows a normal distribution  $N(\mu, \sigma)$ , then  $T$  follows a lognormal distribution whose *pdf* is given by

$$f(t) = \frac{1}{st\sqrt{2\pi}} \exp\left[-\frac{1}{2s^2}\left\{\log\left(\frac{t}{t_M}\right)\right\}^2\right], t \geq 0$$

where  $s = \sigma$  is a *shape parameter* and  $t_M$ , the *median time to failure* is the location parameter, given by  $\log t_M = \mu$ .

It can be proved that

$$\text{MTTF} = E(T) = t_M \exp\left(\frac{s^2}{2}\right)$$

$$\text{Var}(T) = \sigma^2_T = t_M^2 \exp(s^2) [\exp(s^2) - 1]$$

## Example:

The density function of the time to failure in years of the gizmos (for use in widgets) manufactured by a certain company is given by  $f(t) = \frac{200}{(t+10)^3}$ ,  $t \geq 0$ .

- (a) Derive the reliability function and determine the reliability for the first year of operation.
- (b) Compute the MTTF.
- (c) What is the design life for a reliability 0.95?
- (d) Will a one-year burn-in period improve the reliability in part (a)? If so, what is the new reliability?

## Solution:

(a)  $f(t) = \frac{200}{(t+10)^3}, t \geq 0$

$$R(t) = \int_t^{\infty} f(t) dt = \left[ \frac{-100}{(t+10)^2} \right]_t^{\infty} = \frac{100}{(t+10)^2}$$

$$R(1) = \frac{100}{(1+10)^2} = 0,8264.$$

(b) MTTF =  $\int_0^{\infty} R(t) dt = \int_0^{\infty} \frac{100}{(t+10)^2} dt$

$$= \left( \frac{-100}{t+10} \right)_0^{\infty} = 10 \text{ years.}$$

### Solution (Continued):

(c) Design life is the time to failure ( $t_D$ ) that corresponds to a specified reliability. Now it is required to find  $t_D$  corresponding to  $R = 0.95$

$$\frac{100}{(t_D + 10)^2} = 0.95$$

$$\text{i.e., } (t_D + 10)^2 = 100.2632$$

$$t_D = 0.2598 \text{ year or 95 days}$$

(d)  $R(t/1) = \frac{R(t+1)}{R(1)} = \frac{100}{(t+11)^2} \div \frac{100}{(t+10)^2} = \frac{121}{(t+11)^2}$

Now  $R(t/1) > R(t)$ , if  $\frac{121}{(t+11)^2} > \frac{100}{(t+10)^2}$

if  $\frac{(t+10)^2}{(t+11)^2} > \frac{100}{121}$

if  $\frac{t+10}{t+11} > \frac{10}{11}$

then  $11t + 110 > 10t + 110$ , which is true, as  $t \geq 0$

$\therefore$  One year burn-in period will improve the reliability.

Now  $R(1/1) = \frac{121}{(1+11)^2} = 0.8403 > 0.8264$ .

## Example:

The time to failure in operating hours of a critical solid-state power unit has the hazard rate function  $\lambda(t) = 0.003 \left(\frac{t}{500}\right)^{0.5}$ , for  $t \geq 0$ .

- (a) What is the reliability if the power unit must operate continuously for 500 hours?
- (b) Determine the design life if a reliability of 0.90 is desired.
- (c) Compute the MTTF.
- (d) Given that the unit has operated for 50 hours, what is the probability that it will survive a second 50 hours of operation?

## Solution:

(a)  $R(t) = \exp \left[ - \int_0^t \lambda(t) dt \right]$

$$R(50) = \exp \left[ - \int_0^{50} 0.003 \left( \frac{t}{500} \right)^{0.5} dt \right]$$

$$= \exp \left[ - \frac{0.003}{\sqrt{500}} \cdot \frac{2}{3} t^{3/2} \Big|_0^{50} \right]$$

$$= \exp \left[ - \frac{0.003}{\sqrt{500}} \times \frac{2}{3} \times 50\sqrt{50} \right]$$

$$= \exp [-0.03162]$$

$$= 0.9689.$$

## Solution (Continued):

(b)  $R(t_D) = 0.90$

$$\exp \left[ - \int_0^{t_D} 0.003 \left( \frac{t}{500} \right)^{0.5} dt \right] = 0.90$$

$$- \int_0^{t_D} \frac{0.003}{\sqrt{500}} t^{1/2} dt = -0.10536$$

$$\frac{0.003}{\sqrt{500}} \times \frac{2}{3} t_D^{3/2} = 0.10536$$

$$t_D = \left\{ \frac{3 \times \sqrt{500} \times 0.10536}{2 \times 0.003} \right\}^{2/3} = 111.54 \text{ hours.}$$

### Solution (Continued):

$$\begin{aligned}(c) \quad \text{MTTF} &= \int_0^{\infty} R(t) dt \\&= \int_0^{\infty} e^{-\left(\frac{0.003}{\sqrt{500}} \times \frac{2}{3} \times t^{3/2}\right)} dt \\&= \int_0^{\infty} e^{-at^{3/2}} dt, \text{ where } a = \frac{0.003 \times 2}{3 \times \sqrt{500}} \\&= \int_0^{\infty} e^{-x} \cdot \frac{2}{3a^{2/3}} x^{-1/3} dx, \text{ on putting } x = at^{3/2} \\&= \frac{2}{3a^{2/3}} \Gamma(2/3) = \frac{2}{3a^{2/3}} \frac{3}{2} \Gamma(5/3) \\&= \frac{0.9033}{a^{2/3}}, \text{ from the table of values of Gamma function.} \\&= 451.65 \text{ hours.}\end{aligned}$$

## Solution (Continued):

$$(d) P(T \geq 100 / T \geq 50) = \frac{P(T \geq 100)}{P(T \geq 50)} = \frac{R(100)}{R(50)}$$

$$= \exp \left[ - \int_{50}^{100} \lambda(t) dt \right]$$

$$= \exp \left[ \left\{ -\frac{0.002}{\sqrt{500}} \times 100^{3/2} \right\} - \left\{ -\frac{0.002}{\sqrt{500}} \times 50^{3/2} \right\} \right]$$

$$= \exp [ \{-0.08944\} - \{-0.03162\} ]$$

$$= 0.9438$$

## Exercise:

The reliability of a turbine blade is given by  $R(t) = \left(1 - \frac{t}{t_0}\right)^2$ ,  $0 \leq t \leq t_0$ , where  $t_0$  is the maximum life of the blade.

- (a) Show that the blades are experiencing wear out.
- (b) Compute MTTF as a function of the maximum life.
- (c) If the maximum life is 2000 operating hours, what is the design life for a reliability of 0.90?

## Example:

A manufacturer determines that, on the average, a television set is used 1.8 hours per day. A one-year warranty is offered on the picture tube having a MTTF of 2000 hours. If the distribution is exponential, what percentage of the tubes will fail during the warranty period?

## Solution:

Since the distribution of the time to failure of the picture tube is exponential;

$$R(t) = e^{-\lambda t}, \text{ where } \lambda \text{ is the failure rate}$$

Given that MTTF = 2000 hours

i.e.,  $\int_0^{\infty} e^{-\lambda t} dt = 2000$

i.e.  $\frac{1}{\lambda} = 2000 \text{ or } \lambda = 0.0005/\text{hour}$

$$P(T \leq 1 \text{ year}) = P(T \leq 365 \times 1.8 \text{ hours}) [\because \text{the T.V. is operated for 1.8 hours/day}]$$

$$= 1 - P\{T > 657\}$$

$$= 1 - R(657)$$

$$= 1 - e^{-0.0005 \times 657}$$

$$= 0.28$$

i.e., 28% of the tubes will fail during the warranty period.

## Example:

A cutting tool wears out with a time to failure that is normally distributed. It is known that about 34.5% of the tools fail before 9 working days and about 78.8% fail before 12 working days.

- (a) Compute the MTTF
  - (b) Determine its design life for a reliability of 0.99.
  - (c) Determine the probability that the cutting tool will last one more day given that it has been in-use for 5 days.
- (a) Let  $T$  follow a  $N(\mu, \sigma)$

**Solution:**

$$\text{Given } \int_{-\infty}^9 f(t) dt = 0.345$$

$$\text{i.e., } \int_{-\infty}^{\frac{9-\mu}{\sigma}} \phi(z) dz = 0.345, \text{ on putting } z = \frac{t-\mu}{\sigma}$$

$$\text{i.e., } \int_0^{\frac{\mu-9}{\sigma}} \phi(z) dz = 0.155$$

$$\therefore \frac{\mu-9}{\sigma} = 0.4$$

using the normal tables.

$$\text{Also } \int_{-\infty}^{12} f(t) dt = 0.788$$

$$\text{i.e., } \int_{-\infty}^{\frac{12-\mu}{\sigma}} \phi(z) dz = 0.788$$

$$\text{i.e., } \int_0^{\frac{12-\mu}{\sigma}} \phi(z) dz = 0.288$$

**Solution (Continued):**

$$\therefore \frac{12 - \mu}{\sigma} = 0.8$$

using the normal tables

Solving equations (1) and (2), we get

$$\mu = 10 \text{ and } \sigma = 2.5$$

i.e., M.T.T.F = 10 days.

(b) Let  $t_R$  be the required design life for  $R = 0.99$

$$\therefore \int_{t_R}^{\infty} f(t) dt = 0.99 \quad \text{or} \quad \int_{\frac{t_R-10}{2.5}}^{\infty} \phi(z) dz = 0.99$$

$$\therefore \int_0^{\frac{10-t_R}{2.5}} \phi(z) dz = 0.49$$

$$\therefore \frac{10 - t_R}{2.5} = 2.32, \text{ using the normal tables.}$$

$$\therefore t_R = 4.2 \text{ days}$$

$$(c) P(T \geq 6 / T > 5) = \frac{P(T \geq 6)}{P(T > 5)} = \frac{\int_6^{\infty} f(t) dt}{\int_5^{\infty} f(t) dt}$$

$$= \frac{\int_{-1.6}^{\infty} \phi(z) dz}{\int_{-2}^{\infty} \phi(z) dz} = \frac{0.5 + \int_0^{1.6} \phi(z) dz}{0.5 + \int_0^2 \phi(z) dz}$$

$$= \frac{0.94520}{0.97725} = 0.9672.$$

### **Exercise:**

A one-year guarantee is given based on the assumption that no more than 10% of the items will be returned. Assuming an exponential distribution, what is the maximum failure rate that can be tolerated?

### **Exercise:**

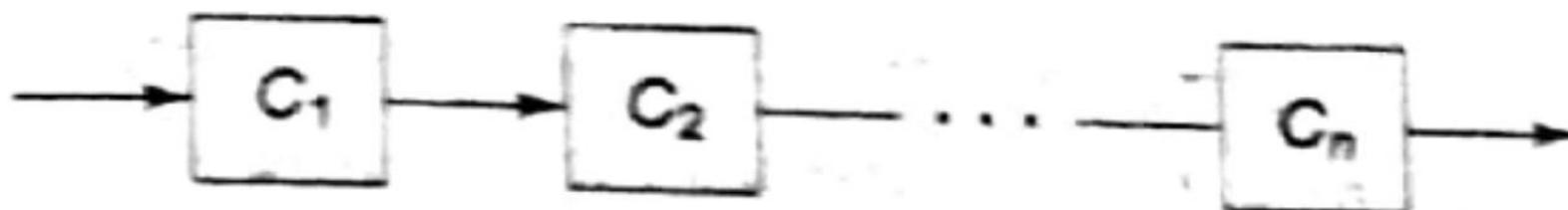
A component experiences failures at a constant rate (CFR) with an MTTF of 1100 hours. Find the reliability for a 200-hr mission. Please round your answer to 2 decimals.

# System Reliability

## Serial or Non-redundant Configuration

Series or nonredundant configuration is one in which the components of the system are connected in series (or serially) as shown in the following reliability block diagram.

Each block represents a component.



## Serial or Non-redundant Configuration

In series configuration, all components must function for the system to function. In other words the failure of any component causes system failure.

Let  $R_1(t)$ ,  $R_2(t)$  and  $R_s(t)$  be the reliabilities of the components  $C_1$  and  $C_2$  and the system (assuming that there are only 2 components in series),

Then  $R_1 = P(C_1)$  = probability that  $C_1$  functions

and  $R_2 = P(C_2)$  = probability that  $C_2$  functions

Now  $R_s$  = probability that both  $C_1$  and  $C_2$  function

$= P(C_1 \cap C_2) = P(C_1)P(C_2)$ , assuming that  $C_1$  and  $C_2$  function

independently.

$$= R_1 \times R_2$$

## Serial or Non-redundant Configuration

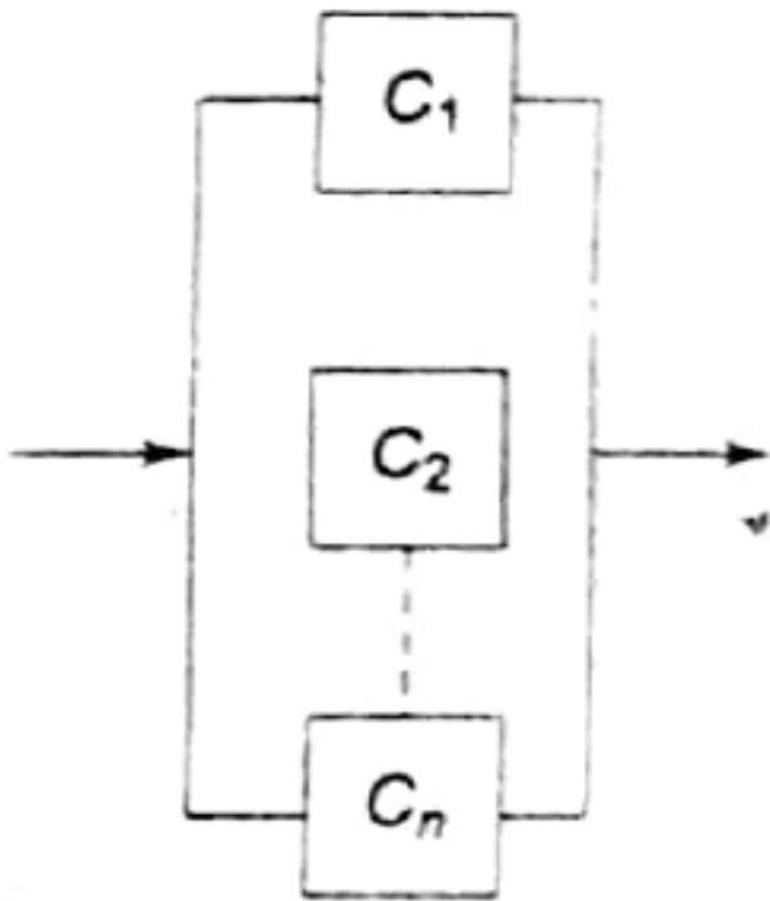
This result may be extended. If  $C_1, C_2, \dots, C_n$  be a set of  $n$  independent components in series with reliabilities  $R_1(t), R_2(t), \dots, R_n(t)$ , then

$$R_s(t) = R_1(t) \times R_2(t) \times \cdots \times R_n(t)$$
$$\leq \min\{R_1(t), R_2(t), \dots, R_n(t)\} \quad [\because 0 < R_i(t) < 1]$$

i.e., the system reliability will not be greater than the smallest of the component reliabilities.

## Parallel or Redundant Configuration

*Parallel* or redundant configuration is one in which the components of the system are connected in parallel as shown in the following reliability block diagram.



## Parallel or Redundant Configuration

In parallel configuration, all components must fail for the system to fail. This means that if one or more components function, the system continues to function.

Taking  $n = 2$  and denoting the system reliability by  $R_p$  ('p' for parallel configuration), we have

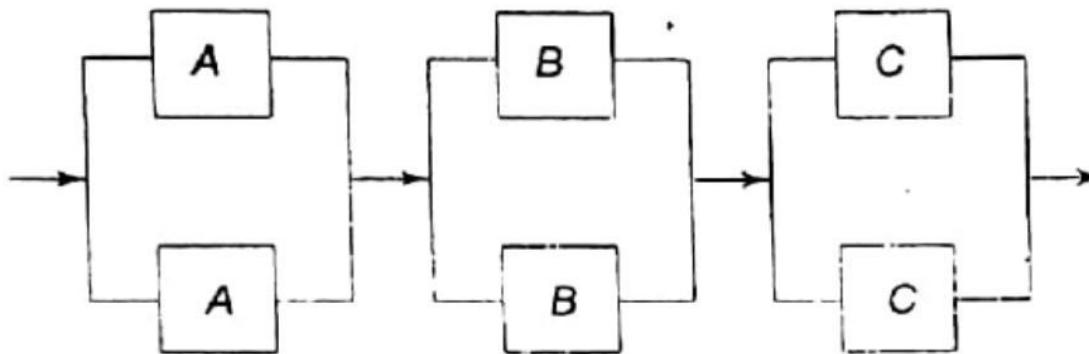
$$\begin{aligned} R_p &= P(C_1 \text{ or } C_2 \text{ or both function}) \\ &= P(C_1 \cup C_2) \\ &= P(C_1) + P(C_2) - P(C_1 \cap C_2) \\ &= P(C_1) + P(C_2) - P(C_1)P(C_2), \text{ since } C_1 \text{ and } C_2 \text{ are independent} \\ &= R_1 + R_2 - R_1 R_2 = 1 - (1 - R_1)(1 - R_2) \end{aligned}$$

Extending to  $n$  components, we have

$$\begin{aligned} R_p &= 1 - (1 - R_1)(1 - R_2) \cdots (1 - R_n) \\ &\geq \text{Max } \{R_1, R_2, \dots, R_n\} \end{aligned}$$

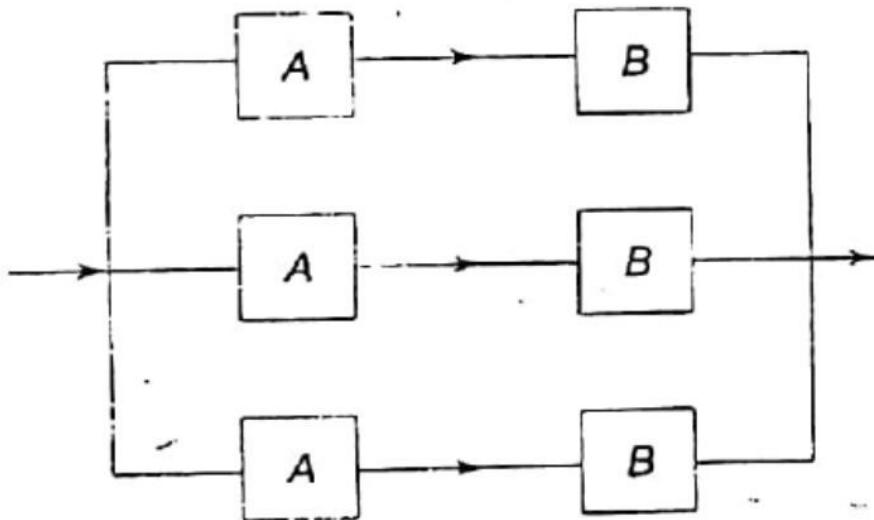
## Parallel Series Configuration

A system, in which  $m$  subsystems are connected in series where each subsystem has  $n$  components connected in parallel as in shown Fig. It is said to be in parallel series configuration or *low-level redundancy*.



## Series Parallel Configuration

A system, in which  $m$  subsystems are connected in parallel where each subsystem has  $n$  components connected in series as in Fig. , is said to be in series-parallel configuration or *high level redundancy*.



## Example:

An electronic circuit consists of 5 silicon transistors, 3 silicon diodes, 10 composition resistors and 2 ceramic capacitors connected in series configuration. The hourly failure rate of each component is given below.

Silicon transistor

$$\lambda_t = 4 \times 10^{-5}$$

Silicon diode

$$\lambda_d = 3 \times 10^{-5}$$

Composition resistor

$$\lambda_r = 2 \times 10^{-4}$$

Ceramic capacitor

$$\lambda_c = 2 \times 10^{-4}$$

Calculate the reliability of the circuit for 10 hours, when the components follow exponential distribution.

## Solution:

Since the components are connected in series, the system (circuit) reliability given by

$$\begin{aligned}R_s(t) &= R_1(t) \cdot R_2(t) \cdot R_3(t) \cdot R_4(t) \\&= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} \cdot e^{-\lambda_3 t} \cdot e^{-\lambda_4 t} \\&= e^{-(5\lambda_1 + 3\lambda_d + 10\lambda_r + 2\lambda_c)t}\end{aligned}$$

$$\begin{aligned}R_s(10) &= e^{-(20 \times 10^{-5} + 9 \times 10^{-5} + 20 \times 10^{-4} + 4 \times 10^{-4}) \times 10} \\&= e^{-(20 + 9 + 200 + 40) \times 10^{-4}} \\&= e^{-0.0269} = 0.9735\end{aligned}$$

## Example:

There are 16 components in a non-redundant system. The average reliability of each component is 0.99. In order to achieve at least this system reliability using a redundant system with 4 identical new components, what should be the least reliability of each new component?

## Solution:

For the non-redundant system,

$$R_s = R^{16} = (0.99)^{16} \approx 0.85$$

Let the new components have a reliability of  $R'$  each.

Then for the redundant system with 4 components,  $R_p \geq 0.85$

i.e.,  $1 - (1 - R')^4 \geq 0.85$

i.e.,  $(1 - R')^4 \leq 0.15$

$$1 - R' \leq (0.15)^{\frac{1}{4}} \text{ or } 0.62$$

$$R' \geq 0.38$$

i.e., the reliability of each of the new components should be at least 0.38.

## Example:

Thermocouples of a particular design have a failure rate of 0.008 per hour. How many thermocouples must be placed in parallel if the system is to run for 100 hours with a system failure probability of no more than 0.05? Assume that all failures are independent.

## Solution:

If  $T$  is the time to failure of the system, it is required that

$$P(T \leq 100) \leq 0.05$$

i.e.,  $1 - R_p(100) \leq 0.05$

Let the number of thermocouples to be connected in parallel be  $n$ .

Then  $R_p(t) = 1 - (1 - R)^n$

where  $R$  is the reliability of each couple.

The failure rate of each couple = 0.008 (constant)

$$R = e^{-0.008t}$$

Using (3) in (2), we have

$$1 - R_p(t) = (1 - e^{-0.008t})^n$$

$$\therefore 1 - R_p(100) = (1 - e^{-0.8})^n$$

Using (4) in (1), we have

$$(1 - e^{-0.8})^n \leq 0.05$$

i.e.,  $(0.55067)^n \leq 0.05$

By trials, we find that (5) is not satisfied when  $n = 0, 1, 2, 3, 4$  and  $5$ .

When  $n = 6$ ,  $(0.55067)^6 = 0.02788 < 0.05$

Hence 6 thermocouples must be used in the parallel configuration.

# Maintainability

No equipment (system) can be perfectly reliable in spite of the utmost care and best effort on the part of the designer and manufacturer. In fact, very few systems are designed to operate without maintenance of any kind. For a large number of systems, maintenance is a must, as it is one of the effective ways of increasing the reliability of the system.

Usually, two kinds of maintenance are adopted. They are preventive maintenance and corrective or repair maintenance. Preventive maintenance is maintenance done periodically before the failure of the system, so as to increase the reliability of the system by removing the ageing effects of wear, corrosion, fatigue and related phenomena. On the other hand, repair maintenance is performed after the failure has occurred so as to return the system to operation as soon as possible.

The amount and type of maintenance that is used depends on the respective costs and safety consideration of system failure. It is generally assumed that a preventive maintenance action is less costly than a repair maintenance action.

## Reliability Under Preventive Maintenance

Let  $R(t)$  and  $R_M(t)$  be the reliability of a system without maintenance and with maintenance.

Let the preventive maintenance be performed on the system at intervals of  $T$ . Let the preventive maintenance be performed on the system at intervals of  $T$ . Since  $R_M(t) = P\{\text{the maintained system does not fail before } t\}$ , we have

$$R_M(t) = R(t), \text{ for } 0 \leq t < T$$

$$= R(T), \text{ for } t = T.$$

After performing the first maintenance operation at  $T$ , the system becomes as good as new.

Hence, if  $T \leq t < 2T$ ,

$R_M(t) = P\{\text{the system does not fail up to } T \text{ and it survives for a time } (t - T) \text{ without failure}\}$

$$= R(T) \cdot R(t - T), \text{ for } T \leq t < 2T$$

Similarly after two maintenance operations,

$R_M(t) = \{R(T)\}^2 \cdot R(t - 2T), \text{ for } 2T \leq t < 3T$

Proceeding like this, we get in general,

$R_M(t) = \{R(T)\}^n \cdot R(t - nT), \text{ for } nT \leq t < (n + 1)T$

$$(n = 0, 1, 2)$$

## Reliability Under Preventive Maintenance

$$R_M(t) = \{R(\tau)\}^n \times R(t-n\bar{T})$$

for

$\tau \rightarrow$  Period

$$n\bar{T} < t < (n+1)\bar{T}$$

Period wise

$n = \text{No. of Maintenance}$

,  $n = 0, 1, 2, \dots$

## MTTF of a System with Preventive Maintenance

$$\text{MTTF} = \int_0^{\infty} R_M(t) dt$$

$$= \sum_{n=0}^{\infty} \int_{nT}^{(n+1)T} R_M(t) dt, \text{ by dividing the range into intervals of length } T$$

$$= \sum_{n=0}^{\infty} \int_{nT}^{(n+1)T} \{R(T)\}^n R(t - nT) dt$$

$$= \sum_{n=0}^{\infty} \{R(T)\}^n \int_0^T R(t') dt', \text{ on putting } t - nT = t'$$

$$\text{MTTF} = \frac{\int_0^T R(t) dt}{1 - R(T)}$$

# Reliability Under Repair or Corrective Maintenance

## Maintainability

A measure of how fast a component (system) may be repaired following failure is known as maintainability. Repairs require different lengths of time and even the time to perform a given repair is uncertain (random), because circumstances, skill level, experience of maintenance personnel and such other factors vary. Hence the time  $T$  required to repair a failed component (system) is a continuous R.V.

Maintainability is mathematically defined as the cumulative distribution function (*cdf*) of the R.V.  $T$ , representing the time to repair and denoted as  $M(t)$

i.e.,

$$M(t) = P\{T \leq t\} = \int_0^t m(t) dt$$

where  $m(t)$  is the *pdf* of  $T$ .

Pdf of  $T$   
CRV -  $\bar{T}$  ↓  
time to Repair

## MTTR and Repair Rate Function

The expected value of repair time  $T$  is called *the mean time to repair (MTTR)* and is given by

$$\text{MTTR} = E(T) = \int_0^t t \times m(t) dt$$

If the conditional probability that the (component) system will be repaired (made operational) between  $t$  and  $t + \Delta t$ , given that it has failed at  $t$  and the repair starts immediately, is  $\mu(t) \Delta t$ , then  $\mu(t)$  is called the instantaneous repair rate or simply the repair rate and denotes the number of repairs in unit time.

$$\text{i.e., } \mu(t) \Delta t = \frac{P\{t \leq T \leq t + \Delta t\}}{P(T > t)}$$

$$\mu(t) = \frac{m(t)}{1 - M(t)}$$

$$m(t) = \frac{d}{dt} M(t)$$

$$\mu(t) = \frac{M'(t)}{1 - M(t)}$$

$$M(t) = 1 - e^{- \int_0^t \mu(t) dt}$$

$$m(t) = \mu(t) \cdot e^{- \int_0^t \mu(t) dt}$$

## Example:

If a device has a failure rate of

$$\lambda(t) = (0.015 + 0.02t)/\text{year}, \text{ where } t \text{ is in years,}$$

- (a) Calculate the reliability for a 5 year design life, assuming that no maintenance is performed.
- (b) Calculate the reliability for a 5 year design life, assuming that annual preventive maintenance restores the device to an as-good as new condition.
- (c) Repeat part (b) assuming that there is a 5% chance that the preventive maintenance will cause immediate failure.

## Solution:

(a)

$$R(t) = e^{-\int_0^t \lambda(t) dt}$$
$$= e^{-\int_0^5 (0.015 + 0.02t) dt}$$
$$R(5) = e^{-(0.015 \times 5 + 0.01 \times 25)}$$
$$= e^{-0.325} = 0.7225$$

Since annual preventive maintenance is performed, there will be 4 preventive maintenances in the first 5 years.

$$R_M(t) = \{R(T)\}^n \times R(t - nT), \text{ after } n \text{ maintenances}$$

Here

$$t = 5, T = 1 \text{ and } n = 4$$

$$\therefore R_M(5) = \{R(1)\}^4 \times R(5 - 4)$$
$$= \{R(1)\}^5$$
$$= \{e^{-0.025}\}^5, \text{ using (1)}$$
$$= 0.8825.$$

## Solution (Continued):

(c)  $P\{\text{preventive maintenance causes immediate failure}\} = 0.05$   
 $\therefore P\{\text{the device survives after each preventive maintenance}\} = 0.95$   
As there are 4 maintenances,

$$\begin{aligned}R_M(5) &= R_M(5) \text{ without breakdown} \times \text{probability of no} \\&\quad \text{breakdown in 5 years} \\&= 0.8825 \times (0.95)^4 \\&= 0.7188.\end{aligned}$$

## Example:

The time to repair a power generator is best described by its *pdf*

$$m(t) = \frac{t^2}{333}, \quad 1 \leq t \leq 10 \text{ hours}$$

- (a) Find the probability that a repair will be completed in 6 hours.
- (b) What is the MTTR?
- (c) Find the repair rate.

(a).  $P(T \leq 6) = ?$

$M(t) = P(T \leq t)$

$$= \int_0^t m(t) dt$$

$M(6) = ?$

**Solution:**(a)  $P(T < 6) = P(1 \leq T < 6)$ , where  $T$  is the time to repair

$$M(6) = \int_1^6 m(t) dt$$

$$= \int_1^6 \frac{t^2}{333} dt = \left( \frac{t^3}{999} \right)_1^6 = 0.2152$$

$$(b) \text{ MTTR} = \int_0^\infty tm(t) dt = \int_1^{10} \frac{t^3}{333} dt = \left( \frac{t^4}{4 \times 333} \right)_1^{10} \\ = 7.5 \text{ hours}$$

$$(c) \text{ Repair rate} = \mu(t) = \frac{m(t)}{1 - M(t)}$$

$$1 - M(t) = \frac{t^2 / 333}{\int_1^{10} \frac{t^2}{333} dt} = \frac{t^2 / 333}{\frac{1}{999}(10^3 - t^3)}$$

$$\mu(t) = \frac{3t^2}{1000 - t^3} \text{ per hour.}$$

$$\begin{aligned} 1 - M(t) &= \frac{t^2 / 333}{\int_1^{10} \frac{t^2}{333} dt} = \frac{t^2 / 333}{\frac{1}{999}(10^3 - t^3)} \\ &\approx 1 - \int_0^t m(t) dt \\ &\approx \int_t^\infty m(t) dt \end{aligned}$$

## Exercise:

A reliability engineer has determined that the hazard rate function for a milling machine is  $\lambda(t) = 0.0004521t^{0.8}$ ,  $t \geq 0$ , where  $t$  is measured in years. Determine which of the following options will provide the greatest reliability over the machine's 20 years operating life.

*Option A* : Do nothing-operate the machine until it fails.

*Option B*: An annual preventive maintenance program (with no maintenance-induced failures)

*Option C*: Operate a second machine in parallel with the first (active redundant).

# Availability

Closely associated with the reliability of repairable (maintained) systems is the concept of *availability*. Like reliability and maintainability, availability is also a probability.

*Availability* is defined as the probability that a component (or system) is performing its intended function at a given time ' $t$ ' on the assumption that it is operated and maintained as per the prescribed conditions. This is referred to as *point availability* and denoted by  $A(t)$ .

It is to be observed that reliability is concerned with failure-free operation up to time  $t$ , whereas availability is concerned with the capability to operate at the point of time  $t$ .

If  $A(t)$  is the point availability of a component (or system), then

$$A(t_2 - t_1) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} A(t) dt$$

This is called the *interval availability* or *mision availability*.

In particular, the interval availability over the interval  $(0; T)$  is

$$A(T) = \frac{1}{T} \int_0^T A(t) dt$$

Now  $\lim_{T \rightarrow \infty} A(t)$  is called the *steady-state* or *asymptotic* or *long-run availability* and denoted by  $A$  or  $A(\infty)$ .

# Availability Function of a Single Component (or System)

## Point Availability

$$A(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

Constants

$\lambda \rightarrow$  Hazard Function

$\mu \rightarrow$  Repair Rate

Function

## Interval Availability Over $(0, T)$

$$A(T) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{(\lambda + \mu)^2} \times T \{1 - e^{-(\lambda + \mu)T}\}$$

## Steady-State Availability

$$A(\infty) = \frac{1/\lambda}{1/\lambda + 1/\mu} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}$$

## Example:

Reliability testing has indicated that a voltage inverter has a 6 month reliability of 0.87 without repair facility. If repair facility is made available with an MTTR of 2.2 months, compute the availability over the 6-month period. (Assume constant failure and repair rates)

Soln.

ED( $\lambda$  or  $\mu$ )

$$R(6) = 0.87 = e^{-\lambda t} = e^{-\lambda 6} \Rightarrow \lambda = ?$$

$$MTTR = \frac{1}{\mu} = 2.2 \Rightarrow \mu = \frac{1}{2.2}$$

$$A(0,6) = A(T=6) = ?$$

## Solution:

For constant failure rate  $\lambda$ , reliability is given by  $R(t) = e^{-\lambda t}$ .

As  $R(6) = 0.87, e^{-6\lambda} = 0.87$

$$\therefore \lambda = 0.0232/\text{month}$$

$$\text{MTTR} = \frac{1}{\mu} = 2.2 \therefore \mu = 0.4545/\text{month}$$

Interval availability over  $(0, T)$  is given by

$$A(T) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{(\lambda + \mu)^2 T} \{1 - e^{-(\lambda + \mu)T}\}$$

$$A(6) = \frac{0.4545}{0.4777} + \frac{0.0232}{(0.4777)^2 \times 6} \{1 - e^{-0.4777 \times 6}\}$$

$$= 0.967$$

## Example:

A new computer has a constant failure rate of 0.02 per day (assuming continuous use) and a constant repair rate of 0.1 per day.

Compute the interval availability for the first 30 days and the steady-state availability.

Solution:

$$\lambda = 0.02, \mu = 0.1 \quad \& \quad T = 30$$

$$A_I(T) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{(\lambda + \mu)^2 \times T} \{1 - e^{-(\lambda + \mu)T}\}$$

$$A_I(30) = \frac{0.1}{0.12} + \frac{0.02}{(0.12)^2 \times 30} \{1 - e^{-0.12 \times 30}\}$$
$$= 0.8784$$

$$A(\infty) = \frac{\mu}{\lambda + \mu} = \frac{0.1}{0.12} = 0.8333$$

## Exercise:

The distribution of the time to failure of a component is Weibull with  $\beta = 2.4$  and  $\theta = 400$  hours and the repair distribution is lognormal with  $t_M = 4.8$  hours and  $s = 1.2$ . Find the steady-state availability.

