SMP Exam - Cheat Sheet

Gergo Nador[1], **Richard Brooks**[2], and **ChatGPT**[3]

[1]The human author
[2]The human whose notes this cheat sheet is based on
[3]The not so human author

June 3, 2024

Contents

0 Useful Tools

0.1 Summation Formulae

Sum	**Closed Form**
$\sum_{k=0}^{n} ar^k \quad (r \neq 0)$	$\frac{ar^{n+1}-a}{r-1}, \quad r \neq 1$
$\sum_{k=1}^{n} k$	$\frac{n(n+1)}{2}$
$\sum_{k=1}^{n} k^2$	$\frac{n(n+1)(2n+1)}{6}$
$\sum_{k=1}^{n} k^3$	$\left(\frac{n(n+1)}{2}\right)^2$
$\sum_{k=0}^{\infty} x^k, \quad \|x\| < 1$	$\frac{1}{1-x}$
$\sum_{k=1}^{\infty} kx^{k-1}, \quad \|x\| < 1$	$\frac{1}{(1-x)^2}$

0.2 Z-Score table - Negative

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table 1: Z-Scores Table - Negative

0.3 Z-Score table - Positive

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 2: Z-Scores Table - Positive

0.4 T Table (T-Distribution)

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Confidence Level	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%

Table 3: t-Distribution Table

0.5 Chi-Square Table

Degrees of freedom (df)	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Table 4: Chi-squared Distribution Table

0.6 Table of Integrals

Table of Integrals

BASIC FORMS

(1) $\int x^n dx = \frac{1}{n+1}x^{n+1}$

(2) $\int \frac{1}{x}dx = \ln x$

(3) $\int udv = uv - \int vdu$

(4) $\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx$

RATIONAL FUNCTIONS

(5) $\int \frac{1}{ax+b}dx = \frac{1}{a}\ln(ax+b)$

(6) $\int \frac{1}{(x+a)^2}dx = \frac{-1}{x+a}$

(7) $\int (x+a)^n dx = (x+a)^n\left(\frac{a}{1+n} + \frac{x}{1+n}\right),\ n \neq -1$

(8) $\int x(x+a)^n dx = \frac{(x+a)^{1+n}(nx+x-a)}{(n+2)(n+1)}$

(9) $\int \frac{dx}{1+x^2} = \tan^{-1} x$

(10) $\int \frac{dx}{a^2+x^2} = \frac{1}{a}\tan^{-1}(x/a)$

(11) $\int \frac{xdx}{a^2+x^2} = \frac{1}{2}\ln(a^2+x^2)$

(12) $\int \frac{x^2dx}{a^2+x^2} = x - a\tan^{-1}(x/a)$

(13) $\int \frac{x^3dx}{a^2+x^2} = \frac{1}{2}x^2 - \frac{1}{2}a^2\ln(a^2+x^2)$

(14) $\int (ax^2+bx+c)^{-1}dx = \frac{2}{\sqrt{4ac-b^2}}\tan^{-1}\left(\frac{2ax+b}{\sqrt{4ac-b^2}}\right)$

(15) $\int \frac{1}{(x+a)(x+b)}dx = \frac{1}{b-a}\left[\ln(a+x) - \ln(b+x)\right],\ a \neq b$

(16) $\int \frac{x}{(x+a)^2}dx = \frac{a}{a+x} + \ln(a+x)$

(17) $$\int \frac{x}{ax^2+bx+c}dx = \frac{\ln(ax^2+bx+c)}{2a} - \frac{b}{a\sqrt{4ac-b^2}}\tan^{-1}\left(\frac{2ax+b}{\sqrt{4ac-b^2}}\right)$$

INTEGRALS WITH ROOTS

(18) $\int \sqrt{x-a}\,dx = \frac{2}{3}(x-a)^{3/2}$

(19) $\int \frac{1}{\sqrt{x\pm a}}dx = 2\sqrt{x\pm a}$

(20) $\int \frac{1}{\sqrt{a-x}}dx = 2\sqrt{a-x}$

(21) $\int x\sqrt{x-a}\,dx = \frac{2}{3}a(x-a)^{3/2} + \frac{2}{5}(x-a)^{5/2}$

(22) $\int \sqrt{ax+b}\,dx = \left(\frac{2b}{3a} + \frac{2x}{3}\right)\sqrt{b+ax}$

(23) $\int (ax+b)^{3/2}dx = \sqrt{b+ax}\left(\frac{2b^2}{5a} + \frac{4bx}{5} + \frac{2ax^2}{5}\right)$

(24) $\int \frac{x}{\sqrt{x\pm a}}dx = \frac{2}{3}(x \pm 2a)\sqrt{x\pm a}$

(25) $\int \sqrt{\frac{x}{a-x}}dx = -\sqrt{x}\sqrt{a-x} - a\tan^{-1}\left(\frac{\sqrt{x}\sqrt{a-x}}{x-a}\right)$

(26) $\int \sqrt{\frac{x}{x+a}}dx = \sqrt{x}\sqrt{x+a} - a\ln\left[\sqrt{x} + \sqrt{x+a}\right]$

(27) $\int x\sqrt{ax+b}\,dx = \left(-\frac{4b^2}{15a^2} + \frac{2bx}{15a} + \frac{2x^2}{5}\right)\sqrt{b+ax}$

(28) $$\int \sqrt{x}\sqrt{ax+b}\,dx = \left(\frac{b\sqrt{x}}{4a} + \frac{x^{3/2}}{2}\right)\sqrt{b+ax} - \frac{b^2\ln\left(2\sqrt{a}\sqrt{x} + 2\sqrt{b+ax}\right)}{4a^{3/2}}$$

(29) $$\int x^{3/2}\sqrt{ax+b}\,dx = \left(-\frac{b^2\sqrt{x}}{8a^2} + \frac{bx^{3/2}}{12a} + \frac{x^{5/2}}{3}\right)\sqrt{b+ax} - \frac{b^3\ln\left(2\sqrt{a}\sqrt{x} + 2\sqrt{b+ax}\right)}{8a^{5/2}}$$

(30) $\int \sqrt{x^2 \pm a^2}\,dx = \frac{1}{2}x\sqrt{x^2 \pm a^2} \pm \frac{1}{2}a^2\ln\left(x + \sqrt{x^2 \pm a^2}\right)$

(31) $\int \sqrt{a^2 - x^2}\,dx = \frac{1}{2}x\sqrt{a^2-x^2} - \frac{1}{2}a^2\tan^{-1}\left(\frac{x\sqrt{a^2-x^2}}{x^2-a^2}\right)$

(32) $\int x\sqrt{x^2 \pm a^2} = \frac{1}{3}(x^2 \pm a^2)^{3/2}$

(33) $\int \frac{1}{\sqrt{x^2 \pm a^2}}dx = \ln\left(x + \sqrt{x^2 \pm a^2}\right)$

(34) $\int \frac{1}{\sqrt{a^2-x^2}} = \sin^{-1}\frac{x}{a}$

(35) $\int \frac{x}{\sqrt{x^2 \pm a^2}} = \sqrt{x^2 \pm a^2}$

(36) $\int \frac{x}{\sqrt{a^2-x^2}}dx = -\sqrt{a^2-x^2}$

(37) $\int \frac{x^2}{\sqrt{x^2 \pm a^2}}dx = \frac{1}{2}x\sqrt{x^2 \pm a^2} \mp \frac{1}{2}\ln\left(x+\sqrt{x^2 \pm a^2}\right)$

(38) $\int \frac{x^2}{\sqrt{a^2-x^2}}dx = -\frac{1}{2}x\sqrt{a-x^2} - \frac{1}{2}a^2\tan^{-1}\left(\frac{x\sqrt{a^2-x^2}}{x^2-a^2}\right)$

(39)
$$\int \sqrt{ax^2+bx+c}\ dx = \left(\frac{b}{4a}+\frac{x}{2}\right)\sqrt{ax^2+bx+c} + \frac{4ac-b^2}{8a^{3/2}}\ln\left(\frac{2ax+b}{\sqrt{a}}+2\sqrt{ax^2+bc+c}\right)$$

(40)
$$\int x\sqrt{ax^2+bx+c}\ dx = \left(\frac{x^3}{3}+\frac{bx}{12a}+\frac{8ac-3b^2}{24a^2}\right)\sqrt{ax^2+bx+c} - \frac{b(4ac-b^2)}{16a^{5/2}}\ln\left(\frac{2ax+b}{\sqrt{a}}+2\sqrt{ax^2+bc+c}\right)$$

(41) $\int \frac{1}{\sqrt{ax^2+bx+c}}dx = \frac{1}{\sqrt{a}}\ln\left[\frac{2ax+b}{\sqrt{a}}+2\sqrt{ax^2+bx+c}\right]$

(42)
$$\int \frac{x}{\sqrt{ax^2+bx+c}}dx = \frac{1}{a}\sqrt{ax^2+bx+c} - \frac{b}{2a^{3/2}}\ln\left[\frac{2ax+b}{\sqrt{a}}+2\sqrt{ax^2+bx+c}\right]$$

LOGARITHMS

(43) $\int \ln x dx = x\ln x - x$

(44) $\int \frac{\ln(ax)}{x}dx = \frac{1}{2}(\ln(ax))^2$

(45) $\int \ln(ax+b)dx = \frac{ax+b}{a}\ln(ax+b)-x$

(46) $\int \ln(a^2x^2 \pm b^2)dx = x\ln(a^2x^2 \pm b^2) + \frac{2b}{a}\tan^{-1}\left(\frac{ax}{b}\right)-2x$

(47) $\int \ln(a^2-b^2x^2)dx = x\ln(a^2-b^2x^2) + \frac{2a}{b}\tan^{-1}\left(\frac{bx}{a}\right)-2x$

(48)
$$\int \ln(ax^2+bx+c)dx = \frac{1}{a}\sqrt{4ac-b^2}\tan^{-1}\left(\frac{2ax+b}{\sqrt{4ac-b^2}}\right) - 2x + \left(\frac{b}{2a}+x\right)\ln\left(ax^2+bx+c\right)$$

(49) $\int x\ln(ax+b)dx = \frac{b}{2a}x - \frac{1}{4}x^2 + \frac{1}{2}\left(x^2-\frac{b^2}{a^2}\right)\ln(ax+b)$

(50) $\int x\ln(a^2-b^2x^2)dx = -\frac{1}{2}x^2 + \frac{1}{2}\left(x^2-\frac{a^2}{b^2}\right)\ln(a^2-bx^2)$

EXPONENTIALS

(51) $\int e^{ax}dx = \frac{1}{a}e^{ax}$

(52) $\int \sqrt{x}e^{ax}dx = \frac{1}{a}\sqrt{x}e^{ax} + \frac{i\sqrt{\pi}}{2a^{3/2}}\mathrm{erf}\left(i\sqrt{ax}\right)$ where $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$

(53) $\int xe^x dx = (x-1)e^x$

(54) $\int xe^{ax}dx = \left(\frac{x}{a}-\frac{1}{a^2}\right)e^{ax}$

(55) $\int x^2e^x dx = e^x(x^2-2x+2)$

(56) $\int x^2e^{ax}dx = e^{ax}\left(\frac{x^2}{a}-\frac{2x}{a^2}+\frac{2}{a^3}\right)$

(57) $\int x^3e^x dx = e^x(x^3-3x^2+6x-6)$

(58) $\int x^n e^{ax}dx = (-1)^n\frac{1}{a}\Gamma[1+n,-ax]$ where $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$

(59) $\int e^{ax^2}dx = -i\frac{\sqrt{\pi}}{2\sqrt{a}}\mathrm{erf}\left(ix\sqrt{a}\right)$

TRIGONOMETRIC FUNCTIONS

(60) $\int \sin x dx = -\cos x$

(61) $\int \sin^2 x dx = \frac{x}{2} - \frac{1}{4}\sin 2x$

(62) $\int \sin^3 x dx = -\frac{3}{4}\cos x + \frac{1}{12}\cos 3x$

(63) $\int \cos x dx = \sin x$

(64) $\int \cos^2 x dx = \frac{x}{2} + \frac{1}{4}\sin 2x$

(65) $\int \cos^3 x dx = \frac{3}{4}\sin x + \frac{1}{12}\sin 3x$

(66) $\int \sin x\cos x dx = -\frac{1}{2}\cos^2 x$

(67) $\int \sin^2 x\cos x dx = \frac{1}{4}\sin x - \frac{1}{12}\sin 3x$

(68) $\int \sin x\cos^2 x dx = -\frac{1}{4}\cos x - \frac{1}{12}\cos 3x$

(69) $\int \sin^2 x\cos^2 x dx = \frac{x}{8} - \frac{1}{32}\sin 4x$

(70) $\int \tan x dx = -\ln\cos x$

(71) $\int \tan^2 x dx = -x + \tan x$

(72) $\int \tan^3 x dx = \ln[\cos x] + \frac{1}{2}\sec^2 x$

(73) $\int \sec x dx = \ln|\sec x + \tan x|$

(74) $\int \sec^2 x dx = \tan x$

(75) $\int \sec^3 x dx = \frac{1}{2}\sec x\tan x + \frac{1}{2}\ln|\sec x\tan x|$

(76) $\int \sec x\tan x dx = \sec x$

(77) $\int \sec^2 x\tan x dx = \frac{1}{2}\sec^2 x$

(78) $\int \sec^n x\tan x dx = \frac{1}{n}\sec^n x, \quad n \neq 0$

(79) $\int \csc x dx = \ln|\csc x - \cot x|$

(80) $\int \csc^2 x dx = -\cot x$

(81) $\int \csc^3 x dx = -\frac{1}{2}\cot x\csc x + \frac{1}{2}\ln|\csc x - \cot x|$

(82) $\int \csc^n x\cot x dx = -\frac{1}{n}\csc^n x, \quad n \neq 0$

(83) $\int \sec x\csc x dx = \ln\tan x$

TRIGONOMETRIC FUNCTIONS WITH x^n

(84) $\int x\cos x dx = \cos x + x\sin x$

(85) $\int x\cos(ax) dx = \frac{1}{a^2}\cos ax + \frac{1}{a}x\sin ax$

(86) $\int x^2\cos x dx = 2x\cos x + (x^2-2)\sin x$

(87) $\int x^2\cos ax dx = \frac{2}{a^2}x\cos ax + \frac{a^2x^2-2}{a^3}\sin ax$

(88) $\int x^n\cos x dx = -\frac{1}{2}(i)^{1+n}\left[\Gamma(1+n,-ix) + (-1)^n\Gamma(1+n,ix)\right]$

(89) $\int x^n\cos ax dx = \frac{1}{2}(ia)^{1-n}\left[(-1)^n\Gamma(1+n,-iax) - \Gamma(1+n,iax)\right]$

(90) $\int x\sin x dx = -x\cos x + \sin x$

(91) $\int x\sin(ax) dx = -\frac{x}{a}\cos ax + \frac{1}{a^2}\sin ax$

(92) $\int x^2\sin x dx = (2-x^2)\cos x + 2x\sin x$

(93) $\int x^3\sin ax dx = \frac{2-a^2x^2}{a^3}\cos ax + \frac{2}{a^3}x\sin ax$

(94) $\int x^n\sin x dx = -\frac{1}{2}(i)^n\left[\Gamma(n+1,-ix) - (-1)^n\Gamma(n+1,-ix)\right]$

TRIGONOMETRIC FUNCTIONS WITH e^{ax}

(95) $\int e^x\sin x dx = \frac{1}{2}e^x\left[\sin x - \cos x\right]$

(96) $\int e^{bx}\sin(ax) dx = \frac{1}{b^2+a^2}e^{bx}\left[b\sin ax - a\cos ax\right]$

(97) $\int e^x\cos x dx = \frac{1}{2}e^x\left[\sin x + \cos x\right]$

(98) $\int e^{bx}\cos(ax) dx = \frac{1}{b^2+a^2}e^{bx}\left[a\sin ax + b\cos ax\right]$

TRIGONOMETRIC FUNCTIONS WITH x^n AND e^{ax}

(99) $\int xe^x\sin x dx = \frac{1}{2}e^x\left[\cos x - x\cos x + x\sin x\right]$

(100) $\int xe^x\cos x dx = \frac{1}{2}e^x\left[x\cos x - \sin x + x\sin x\right]$

HYPERBOLIC FUNCTIONS

(101) $\int \cosh x dx = \sinh x$

(102) $\int e^{ax}\cosh bx dx = \frac{e^{ax}}{a^2-b^2}\left[a\cosh bx - b\sinh bx\right]$

(103) $\int \sinh x dx = \cosh x$

(104) $\int e^{ax}\sinh bx dx = \frac{e^{ax}}{a^2-b^2}\left[-b\cosh bx + a\sinh bx\right]$

(105) $\int e^x\tanh x dx = e^x - 2\tan^{-1}(e^x)$

(106) $\int \tanh ax dx = \frac{1}{a}\ln\cosh ax$

(107) $\int \cos ax\cosh bx dx = \frac{1}{a^2+b^2}\left[a\sin ax\cosh bx + b\cos ax\sinh bx\right]$

(108) $$\int \cos ax \sinh bx dx = \frac{1}{a^2+b^2}\left[b\cos ax\cosh bx + a\sin ax\sinh bx\right]$$

(109) $$\int \sin ax \cosh bx dx = \frac{1}{a^2+b^2}\left[-a\cos ax\cosh bx + b\sin ax\sinh bx\right]$$

(110) $$\int \sin ax \sinh bx dx = \frac{1}{a^2+b^2}\left[b\cosh bx\sin ax - a\cos ax\sinh bx\right]$$

(111) $$\int \sinh ax \cosh ax dx = \frac{1}{4a}\left[-2ax + \sinh(2ax)\right]$$

(112) $$\int \sinh ax \cosh bx dx = \frac{1}{b^2-a^2}\left[b\cosh bx\sinh ax - a\cosh ax\sinh bx\right]$$

0.7 Guide on Choosing Degrees of Freedom (df)

Degrees of freedom generally refer to the number of independent pieces of information that are available to estimate another piece of information. The choice of degrees of freedom depends on the specific statistical test being performed.

0.7.1 1. t-Tests

a. One-Sample t-Test Used to determine if the mean of a single sample is different from a known or hypothesized population mean.

- **Degrees of Freedom**: $df = n - 1$
- **Reason**: One parameter (the sample mean) is estimated from the data.

b. Two-Sample t-Test (Independent Samples) Used to compare the means of two independent groups.

- **Degrees of Freedom**: $df = n_1 + n_2 - 2$
- **Reason**: Two parameters (the means of the two groups) are estimated.

c. Paired t-Test Used to compare means from the same group at different times (e.g., before and after treatment).

- **Degrees of Freedom**: $df = n - 1$
- **Reason**: One parameter (the mean difference) is estimated from the data.

0.7.2 2. ANOVA (Analysis of Variance)

Used to compare means across three or more groups.

- **Between-Group Degrees of Freedom**: $df = k - 1$
  - k is the number of groups.
- **Within-Group Degrees of Freedom**: $df = N - k$
  - N is the total number of observations across all groups.
  - **Reason**: Each group mean provides information, and we subtract the number of groups to find the within-group degrees of freedom.

0.7.3 3. Chi-Square Test

Used for testing relationships between categorical variables.

a. Goodness-of-Fit Test

- **Degrees of Freedom**: $df = k - 1$
  - k is the number of categories.
  - **Reason**: The constraint that the sum of the observed frequencies must equal the sum of the expected frequencies reduces the degrees of freedom by 1.

b. Test of Independence

- **Degrees of Freedom**: $df = (r - 1) \times (c - 1)$
  - r is the number of rows.
  - c is the number of columns.
  - **Reason**: The constraints on the row and column totals reduce the degrees of freedom by the number of rows and columns minus 1 each.

0.7.4 4. Regression Analysis

a. Simple Linear Regression Used to model the relationship between a dependent variable and a single independent variable.

- **Degrees of Freedom for the Residuals**: $df = n - 2$
- **Reason**: Two parameters (the intercept and the slope) are estimated from the data.

b. Multiple Regression Used to model the relationship between a dependent variable and multiple independent variables.

- **Degrees of Freedom for the Residuals**: $df = n - k - 1$
  - k is the number of independent variables.
  - **Reason**: Each independent variable's coefficient is estimated, and the intercept, totaling $k+1$ parameters.

0.7.5 Summary Table

Test Type	**Degrees of Freedom (df)**
One-Sample t-Test	$n - 1$
Two-Sample t-Test	$n_1 + n_2 - 2$
Paired t-Test	$n - 1$
ANOVA (Between-Groups)	$k - 1$
ANOVA (Within-Groups)	$N - k$
Chi-Square Goodness-of-Fit	$k - 1$
Chi-Square Test of Independence	$(r - 1) \times (c - 1)$
Simple Linear Regression	$n - 2$
Multiple Regression	$n - k - 1$

1 Linear Algebra Vectors and Matrices

Linear algebra is a branch of mathematics that studies vectors, vector spaces (or linear spaces), linear transformations, and systems of linear equations. It is a foundational field with applications across various scientific disciplines, including physics, computer science, engineering, and economics.

1.1 Vectors

A vector is a mathematical object that has both magnitude and direction. Vectors are often represented as arrows in a geometric space, but they can also be represented as ordered lists of numbers, which are called components.

1.1.1 Vector Notation

Vectors are typically denoted by boldface letters (e.g., $\mathbf{v}$) or with an arrow above the letter (e.g., $\vec{v}$). In component form, a vector in $\mathbb{R}^n$ is written as:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

where v_i represents the i-th component of the vector $\mathbf{v}$.

1.1.2 Addition and Subtraction

If $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$, their sum $\mathbf{u} + \mathbf{v}$ is defined as:

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{pmatrix}$$

1.1.3 Scalar Multiplication

If $\mathbf{v}$ is a vector in $\mathbb{R}^n$ and c is a scalar, the scalar multiplication $c\mathbf{v}$ is defined as:

$$c\mathbf{v} = c \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_n \end{pmatrix}$$

1.1.4 Dot Product

The dot product (or scalar product) of two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ is defined as:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

1.2 Matrices

A matrix is a rectangular array of numbers arranged in rows and columns. Matrices are used to represent linear transformations and to solve systems of linear equations.

1.2.1 Matrix Notation

Matrices are typically denoted by uppercase boldface letters (e.g., $\mathbf{A}$). An $m \times n$ matrix $\mathbf{A}$ with m rows and n columns is written as:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

where a_{ij} represents the element in the i-th row and j-th column of $\mathbf{A}$.

1.2.2 Matrix Addition

If $\mathbf{A}$ and $\mathbf{B}$ are matrices of the same dimension, their sum $\mathbf{A} + \mathbf{B}$ is obtained by adding their corresponding elements:

$$\mathbf{A}+\mathbf{B} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

1.2.3 Scalar Multiplication

If $\mathbf{A}$ is a matrix and c is a scalar, the scalar multiplication $c\mathbf{A}$ is defined as:

$$c\mathbf{A} = c\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{21} & ca_{22} & \cdots & ca_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{m1} & ca_{m2} & \cdots & ca_{mn} \end{pmatrix}$$

1.2.4 Matrix Multiplication

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is an $n \times p$ matrix, their product $\mathbf{AB}$ is an $m \times p$ matrix defined as:

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

for $1 \leq i \leq m$ and $1 \leq j \leq p$. Here is a step-by-step example to illustrate the process:

Let $\mathbf{A}$ be a 2×3 matrix and $\mathbf{B}$ be a 3×2 matrix:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

To find the product $\mathbf{C} = \mathbf{AB}$, where $\mathbf{C}$ is a 2×2 matrix, we compute each element c_{ij} of $\mathbf{C}$ as follows:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

with

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ c_{12} &= a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ c_{22} &= a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{aligned}$$

Thus, the resulting matrix $\mathbf{C}$ is:

$$\mathbf{C} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

Matrix multiplication is not commutative, meaning that $\mathbf{AB} \neq \mathbf{BA}$ in general.

1.2.5 Multiplying a Matrix and a Vector

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{x}$ is a vector in $\mathbb{R}^n$, their product $\mathbf{Ax}$ is a vector in $\mathbb{R}^m$.

Given an $m \times n$ matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

and a vector $\mathbf{x}$ in $\mathbb{R}^n$:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

The product $\mathbf{Ax}$ is an m-dimensional vector:

$$\mathbf{Ax} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}$$

For example, let $\mathbf{A}$ be a 2×3 matrix and $\mathbf{x}$ be a vector in $\mathbb{R}^3$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

The product $\mathbf{Ax}$ is a vector in $\mathbb{R}^2$:

$$\mathbf{Ax} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{pmatrix}$$

In this multiplication, each element of the resulting vector is a linear combination of the elements of the vector $\mathbf{x}$ weighted by the corresponding row elements of the matrix $\mathbf{A}$.

1.2.6 Transpose of a Matrix

The transpose of a matrix $\mathbf{A}$, denoted by $\mathbf{A}^T$, is obtained by swapping its rows and columns. If $\mathbf{A}$ is an $m \times n$ matrix, then $\mathbf{A}^T$ is an $n \times m$ matrix defined as:

$$(\mathbf{A}^T)_{ij} = a_{ji}$$

The transpose of $\mathbf{A}$, denoted $\mathbf{A}^T$, is a 3×2 matrix given by:

$$\mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}$$

Visualizing the transposition process, we swap the rows and columns of $\mathbf{A}$:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \xrightarrow{\text{Transpose}} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}$$

2 Introduction to Probability

2.1 Why Probability and Statistics

- **Stochastic** = Random
- **Natural Process**: Variability and uncertainty
  - Statistics is the study of how to deal with uncertainty
  - Probability is the unit in which we measure uncertainty

2.2 Samples and Populations

- **Population**: The entire group being studied (e.g., all Android devices)
- **Sample**: A representative subset of the population
- **Sample Size**: Important in Big Data

2.3 Axioms of Probability

1. For any event A, $P(A) \geq 0$.
2. $P(S) = 1$.
3. If $A_1, A_2, A_3, \ldots$ are disjoint,

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

2.4 Notation

- $P(A \cap B) = P(A \text{ and } B) = P(A|B)$
- $P(A \cup B) = P(A \text{ or } B)$
- $P(\overline{A}) = P(A') = P(A^c) = P(\text{not } A)$

2.5 Rules of Probability

(a) $P(A^c) = 1 - P(A)$

(b) $P(\emptyset) = 0$

(c) $P(A) \leq 1$

(d) $P(A - B) = P(A) - P(A \cap B)$

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

(f) $A \subseteq B \implies P(A) \leq P(B)$ (note: $A \subseteq B$ means A is a subset of B)

3 Discrete Probability Models

If a sample space is countable, we use a discrete probability model.

If $A \subseteq S$:

$$P(A) = P\left(\bigcup_{s_j \in A} \{s_j\}\right) = \sum_{s_j \in A} P(s_j) \tag{1}$$

This formula states that if A is a subset of the sample space S, then the probability of A can be found by summing the probabilities of the individual elements s_j in A. In other words, for a countable sample space, the probability of an event A is the sum of the probabilities of all the outcomes that constitute A.

3.1 Conditional Probability

Conditional probability is used when we want to determine the probability of an event A given that another event B has occurred. It is particularly useful in situations where the occurrence of one event affects the likelihood of another. For example, in medical testing, the probability of having a disease (event A) given a positive test result (event B) can be determined using conditional probability. It is also widely used in fields such as finance, engineering, and computer science for tasks such as risk assessment, reliability analysis, and machine learning.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0 \tag{2}$$

3.1.1 Some Important Formulas

- $P(A^c|B) = 1 - P(A|B)$
- $P(\emptyset|B) = 0$
- If $B \subseteq A$, $P(A|B) = 1 = \frac{P(A \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$
- If $A \subseteq B$, $P(A|B) = \frac{P(A)}{P(B)}$
- $P(A \mid B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{P(A \cap B^c)}{1 - P(B)} = \frac{P(A) - P(A \cap B)}{1 - P(B)}$

3.2 Chain Rule

The Chain Rule of probability is used to decompose the probability of a sequence of events happening together into a product of conditional probabilities. This rule is particularly useful in complex probabilistic models such as Bayesian networks, where the probability of a sequence of dependent events needs to be calculated. For example, in medical diagnosis, the probability of a combination of symptoms can be determined by considering the conditional probabilities of each symptom given the presence of other symptoms.

$$P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1, A_2) \cdot \ldots \cdot P(A_k \mid A_{k-1}, A_{k-2}, \ldots, A_1)$$

3.3 Independence

Two events A and B are independent if the occurrence of one event does not affect the probability of the other event. This means that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. The independence of events is particularly useful in simplifying the calculation of joint probabilities. For instance, in scenarios where events are independent, the probability of both events occurring together is simply the product of their individual probabilities, i.e., $P(A \cap B) = P(A) \cdot P(B)$.

$$P(A|B) = P(A) \quad \text{iff} \quad A \text{ and } B \text{ are independent}$$

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

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

3.4 Law of Total Probability

The Law of Total Probability provides a way to compute the probability of an event A by considering all possible scenarios (events B_i) that can lead to A. Specifically, it states that the probability of A can be found by summing the probabilities of A occurring in conjunction with each B_i, weighted by the probability of B_i. This is particularly useful when A can occur due to several different, mutually exclusive events B_i.

$$P(A) = \sum_i P(A \cap B_i)$$

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots$$

$$P(A) = \sum_i P(A \mid B_i) \cdot P(B_i)$$

3.5 Bayes' Rule

Bayes' Rule is a fundamental theorem in probability theory that relates the conditional probability of two events. It provides a way to update the probability of a hypothesis B given new evidence A. This is particularly useful in various fields such as statistics, machine learning, and medical diagnosis, where it is necessary to update the probability of a hypothesis as more information becomes available. Bayes' Rule can be used to calculate the posterior probability $P(B|A)$ from the prior probability $P(B)$, the likelihood $P(A|B)$, and the marginal probability $P(A)$.

$$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{\sum_i P(A|B_i) \cdot P(B_i)}$$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)}$$

3.6 Conditional Independence

Conditional Independence means that two events A and B are independent of each other when conditioned on a third event C. This property simplifies the computation of joint probabilities and is widely used in Bayesian networks, where the dependencies between variables are modeled. Understanding conditional independence helps in decomposing complex probabilistic models into simpler components that are easier to analyze and compute.

Two events A and B are said to be conditionally independent given an event C if

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C), \quad P(C) > 0$$

3.6.1 Understanding Conditional Independence

$$P(A \mid B, C) = A \text{ given } (B \text{ and } C)$$

If A and B are conditionally independent, then $P(A \mid B, C)$ does not depend on B. We get

$$P(A \mid B, C) = P(A \mid C)$$

and also

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

i.e., A and B are independent given C.

3.6.2 Example

- **Medical Diagnosis**: Let A be the event that a patient has a fever, B be the event that a patient has a cough, and C be the event that a patient has the flu. Given that the patient has the flu (event C), the events of having a fever (event A) and having a cough (event B) are conditionally independent. This means that knowing the patient has a cough does not change the probability that they have a fever if we already know they have the flu.

4 Random Variables

A function that assigns a real number to each outcome:

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

- Usually a random variable (R.V.) is denoted by X, Y, Z (upper case).
- The outcomes are denoted by x, y, z (lower case).

For example:

$$\begin{aligned} &\mathrm{P}(X = x) \quad \text{if} \quad x = 2 \quad (\text{or } \mathrm{P}(X = 2)) \\ &\mathrm{P}(Y = y) \quad \text{if} \quad y = 3 \quad (\text{or } \mathrm{P}(Y = 3)) \end{aligned}$$

4.1 Expectation in General

In probability theory and statistics, the expectation (or expected value) of a random variable is a measure of the central tendency of the possible outcomes of the random variable. It is essentially a weighted average of all possible values the random variable can take on, with the weights being the probabilities of those values.

Mathematically, the expectation of a discrete random variable X, which can take on values $x_1, x_2, \ldots, x_n$ with probabilities $P(X = x_1), P(X = x_2), \ldots, P(X = x_n)$ respectively, is defined as:

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

For a continuous random variable X with probability density function $f(x)$, the expectation is given by:

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

4.1.1 Linearity of Expectation

For any random variables X and Y and constants a and b:

$$E(aX + bY) = aE(X) + bE(Y)$$

4.1.2 Expectation of a Constant

For a constant c:

$$E(c) = c$$

4.1.3 Expectation of the Sum of Random Variables

For any random variables X and Y:

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

This property extends to any finite number of random variables.

4.1.4 Multiplication by a Constant

For a constant a and a random variable X:

$$E(aX) = aE(X)$$

4.1.5 Expectation of the Product of Independent Random Variables

If X and Y are independent random variables:

$$E(XY) = E(X)E(Y)$$

4.1.6 Non-Negativity

If X is a non-negative random variable (i.e., $X \geq 0$):

$$E(X) \geq 0$$

4.1.7 Iterated Expectation (Law of Total Expectation)

If X and Y are random variables, then:

$$E(X) = E(E(X|Y))$$

where $E(X|Y)$ is the conditional expectation of X given Y.

4.1.8 Jensen's Inequality

If g is a convex function and X is a random variable:

$$g(E(X)) \leq E(g(X))$$

If g is concave, the inequality is reversed:

$$g(E(X)) \geq E(g(X))$$

4.1.9 Variance and Expectation

The variance of X, denoted $\mathrm{Var}(X)$, is related to the expectation by:

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

5 Discrete Random Variables

Let X be a discrete random variable (R.V.) with $\mathrm{P}_X = \{x_1, x_2, x_3, \ldots\}$.

The function

$$\mathrm{P}_X(x_k) = \mathrm{P}(X = x_k), \quad \text{for } k = 1, 2, 3, \ldots$$

is called the probability mass function (PMF) of X.

5.1 Properties:

1. $f(X_x) = \mathrm{P}_X(x_k)$
2. $0 \leq f(X_x) \leq 1$
3. $\sum_{k=1}^{n} f(X_x) = 1$

5.2 Independent Random Variables

Consider X and Y. We say X and Y are independent if:

$$\mathrm{P}(X = x) \cap \mathrm{P}(Y = y) = \mathrm{P}(X = x) \cdot \mathrm{P}(Y = y)$$

It follows:

$$\mathrm{P}(Y = y \mid X = x) = \mathrm{P}(Y = y)$$

Two random variables X and Y are said to be independent if the occurrence of one does not affect the probability distribution of the other. This implies that the joint probability distribution of X and Y can be expressed as the product of their individual probability distributions. In other words, knowing the value of X provides no additional information about the value of Y, and vice versa. Therefore, the conditional probability of Y given X is just the probability of Y.

5.3 Key Concepts in Probability Distributions

5.3.1 Probability Mass Function (PMF)

The Probability Mass Function (PMF) of a discrete random variable X gives the probability that X takes on a specific value. It is denoted as:

$$\mathrm{P}(X = x) = f(x)$$

Other Rules:

$$\sum_{x_k \in R_X} \mathrm{P}(X = x_k) = 1$$

5.3.2 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a discrete random variable X gives the probability that X takes on a value less than or equal to x. It is denoted as:

$$F_X(x) = \sum_{x_k \leq x} \mathrm{P}_X(x_k)$$

Other Rules:

$$0 \leq F(X) \leq 1$$
$$\mathrm{P}(a < X \leq b) = F(b) - F(a)$$
$$\mathrm{P}(a \leq X \leq b) = F(b) - F(a - 1)$$

5.3.3 Survival Function (SF)

The Survival Function (SF), also known as the complementary cumulative distribution function (CCDF), provides the probability that a random variable X takes on a value greater than a specific value x. It is denoted as $\mathrm{SF}(x)$ or sometimes $\overline{F}(x)$.

The survival function is defined as:

$$\mathrm{SF}(x) = \mathrm{P}(X > x) = 1 - F_X(x)$$
$$\mathrm{P}(X \geq k) = \mathrm{SF}(k - 1)$$

where $F_X(x)$ is the cumulative distribution function (CDF) of X.

Properties of the survival function:

- $0 \leq \mathrm{SF}(x) \leq 1$
- $\mathrm{SF}(x)$ is a non-increasing function
- $\mathrm{SF}(x)$ and $F_X(x)$ are complementary: $\mathrm{SF}(x) + F_X(x) = 1$

5.3.4 Expected Value (Mean)

The Expected Value (mean) of a random variable X is a measure of the central tendency, representing the average value of X. It is denoted as:

$$\mu = E[X] = \sum_{x_k \in R_X} x_k \cdot \mathrm{P}(X = x_k)$$

Expectation is linear:

$$\mathrm{E}(aX + b) = a \cdot \mathrm{E}(X) + b$$

5.3.5 Variance

The Variance of a random variable X measures the spread or dispersion of the values of X around the mean. It is denoted as:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = E[X^2] - (E[X])^2$$

General formula:

$$\mathrm{Var}(X) = \sum_{k=1}^{n} x_k^2 \cdot \mathrm{P}(X = x_k) - (E[X])^2$$

Variance is not linear:

$$\mathrm{Var}(aX + b) = a^2 \cdot \mathrm{Var}(X)$$

5.3.6 Standard Deviation

The Standard Deviation of a random variable X is the square root of the variance and provides a measure of the average distance of the values of X from the mean. It is denoted as:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

5.3.7 Common Notations

- $f(x)$: PMF
- $F_X(x)$: CDF
- $\mathrm{SF}(x)$: Survival Function
- $E[X]$: Expected Value (Mean)
- $\mathrm{Var}(X)$: Variance
- σ_X: Standard Deviation

5.4 Distributions

5.4.1 Bernoulli Distribution

A Bernoulli distribution is a discrete distribution with only two possible outcomes: success (usually denoted by 1) and failure (usually denoted by 0). The probability of success is p and the probability of failure is $1 - p$. This distribution is used to model scenarios with binary outcomes, such as a coin flip or a yes/no question.

A Bernoulli random variable (R.V.) can only take two values:

$$\begin{cases} 1 : \text{Success} \\ 0 : \text{Failure} \end{cases}$$

Probability Mass Function (PMF):

$$\mathrm{P}_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - p & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

Expected Value (mean) $E[X]$:

$$E[X] = p$$

Variance $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = p(1 - p)$$

```python
from scipy.stats import bernoulli

p = 0.6  # Probability of success

bernoulli_dist = bernoulli(p)
```

5.4.2 Binomial Distribution

The Binomial distribution describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. It models scenarios where there are multiple independent attempts at achieving a success, such as flipping a coin a certain number of times and counting the number of heads.

Probability Mass Function (PMF):

$$\mathrm{P}_X(x) = f(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where

$$\begin{aligned} X &= \text{number of successes} \\ p &= \text{probability of success} \\ n &= \text{number of trials/experiments} \\ \binom{n}{x} &= \text{Binomial coefficient} = \frac{n!}{x!(n-x)!} \end{aligned}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

Expected Value (mean) $E[X]$:

$$E[X] = np$$

Variance $\text{Var}(X)$:

$$\text{Var}(X) = np(1-p)$$

```python
from scipy.stats import binom

n = 10   # Number of trials
p = 0.6   # Probability of success

binom_dist = binom(n, p)
```

5.4.3 Number of Trials until r Successes (from notes)

Let X denote the number of trials until r successes. Then X has a negative binomial probability mass function (PMF):

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

where

$$\begin{aligned} x &= \textbf{number of trials} \\ r &= \text{number of successes} \\ p &= \text{probability of success} \\ 1-p &= \text{probability of failure} \\ \binom{x-1}{r-1} &= \frac{(x-1)!}{(r-1)!(x-r)!} \quad \text{(binomial coefficient)} \end{aligned}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \sum_{k=r}^{x} \binom{k-1}{r-1} (1-p)^{k-r} p^r$$

Expected Value (mean) $E[X]$:

$$E[X] = \frac{r}{p}$$

Variance $\text{Var}(X)$:

$$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

5.4.4 Number of Failures until r Successes (from Wikipedia)

Let Y denote the number of failures until r successes. Then Y has a negative binomial probability mass function (PMF):

$$f(y) = \binom{y+r-1}{r-1} (1-p)^r p^y \text{ (Scipy docs)}$$

or

$$f(y) = \binom{y+r-1}{y}(1-p)^r p^y \text{ (Wikipedia)}$$

where

$$\begin{aligned} y &= \textbf{number of failures} \\ r &= \text{number of successes} \\ p &= \text{probability of success} \\ 1-p &= \text{probability of failure} \\ \binom{y+r-1}{r-1} &= \frac{(y+r-1)!}{(r-1)!y!} \quad \text{(binomial coefficient)} \end{aligned}$$

Cumulative Distribution Function (CDF):

$$F_Y(y) = \sum_{k=0}^{y} \binom{k+r-1}{r-1}(1-p)^r p^k$$

Expected Value (mean) $E[Y]$:

$$E[Y] = r\left(\frac{1-p}{p}\right)$$

Variance $\mathrm{Var}(Y)$:

$$\mathrm{Var}(Y) = r\left(\frac{1-p}{p^2}\right)$$

```python
from scipy.stats import nbinom

r = 5   # Number of successes
p = 0.6 # Probability of success

nbinom_dist = nbinom(r, p)
```

5.4.5 Geometric Distribution

A Geometric distribution describes the number of trials until the first success. This distribution models scenarios where we are interested in the number of attempts needed to achieve the first success, such as the number of coin flips required to get the first head.

Let X denote the number of trials until the first success. Then X has a geometric probability mass function (PMF):

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

Relation to Negative Binomial Distribution:

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

which is the same as the negative binomial distribution with $r = 1$.

Cumulative Distribution Function (CDF):

$$F_X(x) = 1 - (1-p)^x$$

Expected Value (mean) $E[X]$:

$$E[X] = \frac{1}{p}$$

Variance $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = \frac{1-p}{p^2}$$

```python
from scipy.stats import geom

p = 0.1 # Probability of success

geom_dist = geom(p)
```

5.4.6 Hypergeometric Distribution

The Hypergeometric distribution describes the probability of x successes in a sample of n draws from a finite population of size N that contains r successes, without replacement. This distribution is used to model scenarios where sampling is done without replacement, such as drawing cards from a deck or selecting individuals from a group.

Given:

- A pool of size N
- With exactly r successes
- And a random sample of size n

Now, let X denote the number of successes in n. Then X is a Hypergeometric random variable (R.V.) with PMF:

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \sum_{k=0}^{x} \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

Expected Value (mean) $E[X]$:

$$E[X] = n\frac{r}{N}$$

Variance $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1}$$

```python
from scipy.stats import hypergeom

N = 50  # Total population size
r = 10  # Number of successes in the population
n = 5   # Number of draws (sample size)

hypergeom_dist = hypergeom(N, r, n)
```

5.4.7 Poisson Distribution

The Poisson distribution is used to model the number of events occurring within a specific time/space interval. This distribution is appropriate for scenarios where events happen independently, and the rate at which they occur is constant.

The Poisson random variable (R.V.) has a probability mass function (PMF):

$$f(x) = \mathrm{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where λ indicates the average number of events in a given interval.

Cumulative Distribution Function (CDF):

$$F_X(x) = e^{-\lambda} \sum_{k=0}^{x} \frac{\lambda^k}{k!}$$

Expected Value (mean) $E[X]$:

$$E[X] = \lambda$$

Variance $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = \lambda$$

The parameter λ is crucial as it represents the average number of events in the given interval. This is important for understanding the distribution's behavior and for making predictions about event occurrences.

```python
from scipy.stats import poisson

lambda_ = 3  # Average number of events in a given interval

poisson_dist = poisson(lambda_)
```

6 Continuous Random Variables

A continuous random variable (R.V.) has only continuous values, i.e., values that are uncountable and are related to real numbers, $\mathbb{R}$.

6.1 Main Difference to Discrete Random Variables

- Discrete random variables (DRV) are measured on exact values.
- Continuous random variables (CRV) are measured on intervals.

6.2 Key Concepts in Probability Distributions for Continuous Random Variables

6.2.1 Probability Density Function (PDF)

The Probability Density Function (PDF) of a continuous random variable X describes the likelihood of X taking on a specific value. Unlike the PMF for discrete variables, the PDF is used for continuous variables and does not give probabilities directly. Instead, the area under the PDF curve over an interval gives the probability:

$$f_X(x) \quad \text{such that} \quad \mathrm{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \frac{dF(x)}{dx}$$

The PDF must satisfy:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
3. $F(x) = \int_{-\infty}^{x} f(u)\,du$
4. $\int_{x_1}^{x_2} f(x)\,dx = P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$
5. $P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$

6.2.2 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a continuous random variable X gives the probability that X takes on a value less than or equal to x. It is denoted as:

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,du$$

The CDF must satisfy:

1. $0 \leq F(x) \leq 1$
2. $F(-\infty) = 0, \quad F(\infty) = 1 \qquad (\text{where } -\infty = \text{lower bound} \quad \infty = \text{upper bound})$
3. $F(x)$ is non-decreasing as x increases
4. $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$
5. $P(X \leq x_1) = P(X < x_1)$
6. $P(X > x_1) = 1 - P(X \leq x_1) = 1 - F(x_1)$

6.2.3 Survival Function (SF)

The Survival Function (SF), also known as the complementary cumulative distribution function (CCDF), provides the probability that a continuous random variable X takes on a value greater than a specific value x. It is denoted as:

$$\mathrm{SF}(x) = \mathrm{P}(X > x) = 1 - F_X(x)$$

Properties of the survival function:

- $0 \leq \mathrm{SF}(x) \leq 1$
- $\mathrm{SF}(x)$ is a non-increasing function
- $\mathrm{SF}(x)$ and $F_X(x)$ are complementary: $\mathrm{SF}(x) + F_X(x) = 1$

6.2.4 Expected Value (Mean)

The Expected Value (mean) of a continuous random variable X is a measure of the central tendency, representing the average value of X. It is denoted as:

$$\mu = E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

Expectation is linear:

$$E(aX + b) = a \cdot E(X) + b$$

6.2.5 Variance

The Variance of a continuous random variable X measures the spread or dispersion of the values of X around the mean. It is denoted as:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = E[X^2] - (E[X])^2$$

General formula:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - (E[X])^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2\, f_X(x)\, dx$$

Variance is not linear:

$$\mathrm{Var}(aX + b) = a^2 \cdot \mathrm{Var}(X)$$

6.2.6 Standard Deviation

The Standard Deviation of a continuous random variable X is the square root of the variance and provides a measure of the average distance of the values of X from the mean. It is denoted as:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

6.2.7 Law of the Unconscious Statistician (LOTUS)

The Law of the Unconscious Statistician (LOTUS) states that for a continuous random variable X with probability density function (PDF) $f_X(x)$, and a function $g(x)$, the expected value of $g(X)$ can be calculated as:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

6.2.8 Common Notations

- $f_X(x)$: PDF
- $F_X(x)$: CDF
- $\mathrm{SF}(x)$: Survival Function
- $E[X]$: Expected Value (Mean)
- $\mathrm{Var}(X)$: Variance
- σ_X: Standard Deviation

6.3 Distributions

6.3.1 Uniform Distribution

A Uniform distribution, denoted as $X \sim U(a, b)$, is a continuous probability distribution where all intervals of the same length within the range $[a, b]$ are equally probable. This distribution models scenarios where every outcome in the interval $[a, b]$ has an equal probability.

Let X be a random variable that follows a uniform distribution. Then X has a uniform probability density function (PDF):

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases}$$

Expected Value (mean) $E[X]$:

$$E[X] = \frac{a+b}{2}$$

Variance $\text{Var}(X)$:

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

```python
from scipy.stats import uniform

a = 0  # Lower bound
b = 10 # Upper bound

uniform_dist = uniform(a, b - a)
```

6.3.2 Exponential Distribution

An Exponential distribution is often used to model the time between events. It is defined by a single parameter $\lambda > 0$, known as the rate parameter. This distribution is commonly used in scenarios such as the time until a radioactive particle decays or the time between arrivals at a service point.

Let X be a continuous random variable that follows an exponential distribution with parameter λ. Then X has an exponential probability density function (PDF):

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

Cumulative Distribution Function (CDF):

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x \geq 0 \end{cases}$$

Expected Value (mean) $E[X]$:

$$E[X] = \frac{1}{\lambda}$$

Variance $\text{Var}(X)$:

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

```python
from scipy.stats import expon

lambda_ = 2  # Rate parameter

expon_dist = expon(scale=1/lambda_)
```

6.3.3 Normal Distribution

The Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by its symmetric, bell-shaped curve. It is widely used in statistics and natural sciences to represent real-valued random variables with unknown distributions.

Let X be a continuous random variable that follows a normal distribution with mean μ and variance σ^2. The normal distribution is denoted as $X \sim N(\mu, \sigma^2)$.

Probability Density Function (PDF):

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cumulative Distribution Function (CDF):

$$F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

where erf is the error function:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

Expected Value (mean) $E[X]$:

$$E[X] = \mu$$

Variance $\text{Var}(X)$:

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

Additional Information:

- The normal distribution is defined for all real numbers ($-\infty < X < \infty$).
- About 68.27% of all cases lie within one standard deviation ($\mu \pm \sigma$) of the mean.
- About 95.45% of all cases lie within two standard deviations ($\mu \pm 2\sigma$) of the mean.
- About 99.74% of all cases lie within three standard deviations ($\mu \pm 3\sigma$) of the mean.

```python
from scipy.stats import norm

mu = 0       # Mean
sigma = 1    # Standard deviation

norm_dist = norm(mu, sigma)
```

6.3.4 Standard Normal Distribution

The Standard Normal distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma^2 = 1$, i.e., $X \sim N(0, 1)$.

1. Convert the problem to a standardized normal variable, z-score:

$$z = \frac{X-\mu}{\sigma} \sim N(0, 1)$$

2. A table of z-scores exists, which allows us to find probabilities associated with standard normal variables.
3. We can convert back to the original variable:

$$X = z \cdot \sigma + \mu$$

6.3.5 Standard Normal Distribution - Rules

Here are some important rules for working with the standard normal distribution, denoted as $Z \sim N(0, 1)$:

1. **Cumulative Probability:**

$$P(Z \leq a) = F(a)$$

where $F(a)$ is the cumulative distribution function (CDF) of the standard normal distribution at a. This represents the probability that the standard normal variable Z is less than or equal to a.

2. **Complementary Cumulative Probability:**

$$P(Z \geq a) = 1 - F(a) = F(-a)$$

This rule states that the probability that Z is greater than or equal to a is equal to one minus the CDF at a, which is also equal to the CDF at $-a$.

3. **Probability between Two Values:**

$$P(a \leq Z \leq b) = F(b) - F(a) \quad \text{for } b \geq a$$

This rule provides the probability that the standard normal variable Z lies between two values a and b. It is calculated as the difference between the CDF at b and the CDF at a.

7 Multivariate Random Variables

A multivariate or joint random variable involves two or more random variables considered together. The joint probability distribution of these random variables provides the probabilities that each of the variables falls within any particular range or set of ranges.

7.1 Two Discrete Random Variables

7.1.1 Joint PMF

$$P_{X,Y}(x_i, y_j) = f_{XY}(x_i, y_j) = P(X = x_i, Y = y_j) \tag{3}$$
$$\mathbb{R}_{XY} = \{(x_i, y_j) : f_{XY}(x_i, y_j) > 0\} \tag{4}$$
$$\sum_{(x_i,y_j)\in R_{XY}} f_{XY}(x_i, y_j) = 1 \tag{5}$$

In General:

$$P((x, y) \in A) = \sum_{(x_i,y_j)\in A\cap R_{XY}} f_{XY}(x_i, y_j) \tag{6}$$

7.1.2 Marginal PMF

The marginal probability mass function (PMF) provides the probabilities of individual outcomes of a subset of random variables, summing over the possible values of the other variables. For two discrete random variables X and Y, the marginal PMFs are calculated by summing the joint PMF over all possible values of the other variable.

$$f_X(x) = \sum_{y_j} f_{XY}(x, y_j) \tag{7}$$
$$f_Y(y_j) = \sum_{x_i} f_{XY}(x_i, y_j) \tag{8}$$

(9)

7.1.3 Joint CDF

The joint cumulative distribution function (CDF) of two random variables X and Y is given by:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \tag{10}$$

The joint CDF satisfies the following properties:

$$0 \leq F_{XY}(x, y) \leq 1 \tag{11}$$

(12)

7.1.4 Joint CDF for Given Ranges

The probability that X lies between x_1 and x_2, and Y lies between y_1 and y_2 is given by:

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1) \tag{13}$$

7.1.5 Marginal CDF

The marginal CDFs of X and Y can be derived from the joint CDF by taking limits:

$$F_X(x) = F_{XY}(x, \infty) = \lim_{y \to \infty} F_{XY}(x, y) \tag{14}$$

$$F_Y(y) = F_{XY}(\infty, y) = \lim_{x \to \infty} F_{XY}(x, y) \tag{15}$$

7.1.6 Expected Value (Mean)

The expected value (mean) of a discrete random variable X given a joint distribution with Y is given by:

$$\begin{aligned} E[X] &= \sum_{x_i} x_i \, P(X = x_i) && (16) \\ &= \sum_{x_i} x_i \, f_X(x_i) && (17) \\ &= \sum_{x_i} \sum_{y_j} x_i \, f_{XY}(x_i, y_j) && (18) \end{aligned}$$

7.1.7 Conditional Expected Value - Steps

Step 1: Define the Joint PDF Assume the joint PDF of X and Y is given by:

$$f_{XY}(x, y)$$

with the support of X and Y defined over specific intervals.

Step 2: Calculate the Marginal PDF of Y The marginal PDF of Y, $f_Y(y)$, is obtained by integrating the joint PDF over the range of x:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx$$

Step 3: Evaluate the Marginal PDF at $Y = y$ Evaluate the marginal PDF of Y at the specific value y:

$$f_Y(y) = \int_{x_{\min}}^{x_{\max}} f_{XY}(x, y) \, dx$$

Step 4: Calculate the Conditional PDF $f_{X|Y}(x|y)$ The conditional PDF $f_{X|Y}(x|y)$ is given by:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Step 5: Calculate the Conditional Expectation $E[g(X) \mid Y = y]$ Finally, calculate the conditional expectation:

$$E[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) \, dx$$

Substitute the conditional PDF:

$$E[g(X) \mid Y = y] = \int_{x_{\min}}^{x_{\max}} g(x) \left(\frac{f_{XY}(x, y)}{f_Y(y)} \right) dx$$

7.1.8 Variance

The variance of a discrete random variable X given a joint distribution with Y is given by:

$$V[X] = \sum_{x_i} x_i^2 \, f_X(x_i) - \left(\sum_{x_i} x_i \, f_X(x_i) \right)^2 \tag{19}$$

$$= \sum_{x_i} \sum_{y_j} x_i^2 \, f_{XY}(x_i, y_j) - \left(\sum_{x_i} \sum_{y_j} x_i \, f_{XY}(x_i, y_j) \right)^2 \tag{20}$$

$$= E[X^2] - (E[X])^2 \tag{21}$$

7.1.9 Independence

Two random variables X and Y are independent if the conditional PMF of X given $Y = y_j$ is equal to the marginal PMF of X:

$$f_{X|Y}(x_i \mid y_j) = f_X(x_i) \tag{22}$$

This implies:

$$\frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} = f_X(x_i) \tag{23}$$

Therefore, the joint PMF can be expressed as the product of the marginal PMFs:

$$f_{XY}(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j) \tag{24}$$

7.1.10 Conditioning

The conditional PMF of X given $Y = y_j$ is defined as:

$$f_{X|Y}(x_i \mid y_j) = P(X = x_i \mid Y = y_j) \tag{25}$$

$$= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \tag{26}$$

$$= \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} \tag{27}$$

7.1.11 Conditional Expectation

The conditional expectation of X given $Y = y_j$ is the weighted average of the possible values of X, where the weights are the conditional probabilities. Mathematically, it is defined as:

$$E(X \mid Y = y_j) = \sum_{x_i \in A_X} x_i \cdot P(X = x_i \mid Y = y_j) \tag{28}$$

Since $P(X = x_i \mid Y = y_j)$ is the same as the conditional PMF $f_{X|Y}(x_i \mid y_j)$, the formula can also be written as:

$$E(X \mid Y = y_j) = \sum_{x_i \in A_X} x_i \cdot f_{X|Y}(x_i \mid y_j) \tag{29}$$

7.1.12 Total Expectation

The law of total expectation states that the expected value of a random variable can be computed as the weighted average of the expected values of that variable given a partition of the sample space.

1. Given a partition $\{B_1, B_2, \ldots\}$ of the sample space S:

$$E[X] = \sum_i E[X \mid B_i] \cdot P(B_i) \tag{30}$$

2. For a random variable Y taking values y_j in the range R_Y:

$$E[X] = \sum_{y_j \in R_Y} E[X \mid Y = y_j] \cdot P(Y = y_j) \tag{31}$$

3. The total expectation can also be expressed as the expectation of the conditional expectation:

$$E[X] = E[E[X \mid Y]] \tag{32}$$

7.2 Two Continuous Random Variables

7.2.1 Joint PDF

The joint probability density function (PDF) of two continuous random variables X and Y is given by:

$$P_{XY}((X, Y) \in A) = \iint_A f_{XY}(x, y)\, dx\, dy \tag{33}$$

If $A = \mathbb{R}^2$:

$$P_{XY}((X, Y) \in \mathbb{R}^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx\, dy \tag{34}$$

7.2.2 Marginal PDF

The marginal probability density functions (PDFs) of X and Y can be obtained by integrating the joint PDF over the other variable.

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \tag{35}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \tag{36}$$

7.2.3 Joint CDF

The joint cumulative distribution function (CDF) of X and Y is given by:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \tag{37}$$

$$= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x, y)\, dx\, dy \tag{38}$$

The relationship between the joint PDF and joint CDF is:

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \tag{39}$$

The CDF must satisfy:

$$F_{XY}(\infty, \infty) = 1 \tag{40}$$

7.2.4 Join CDF for Given Ranges

The probability that X lies between x_1 and x_2, and Y lies between y_1 and y_2 is given by:

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1) \tag{41}$$

7.2.5 Marginal CDF

The marginal cumulative distribution functions (CDFs) of X and Y can be derived from the joint CDF $F_{XY}(x, y)$. They represent the probabilities that each variable takes on a value less than or equal to a certain value, regardless of the value of the other variable.

1. The marginal CDF of X is given by:

$$F_X(x) = F_{XY}(x, \infty) \tag{42}$$

2. The marginal CDF of Y is given by:

$$F_Y(y) = F_{XY}(\infty, y) \tag{43}$$

7.2.6 Expected Value (Mean)

The expected value (mean) of X is given by:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx && (44) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{XY}(x, y)\, dy\, dx && (45) \end{aligned}$$

7.2.7 Variance

The variance of X is given by:

$$\begin{aligned} V[X] &= \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx - \left(\int_{-\infty}^{\infty} x\, f_X(x)\, dx \right)^2 && (46) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2\, f_{XY}(x, y)\, dy\, dx - \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{XY}(x, y)\, dy\, dx \right)^2 && (47) \\ &= E[X^2] - (E[X])^2 && (48) \end{aligned}$$

$$Var(XY) = E[X^2] \cdot E[Y^2] - (E[X] \cdot E[Y])^2 \tag{49}$$

7.2.8 Independence

Same as the Discrete case.

7.2.9 Conditioning

The same principles apply as in the discrete case. Most importantly:

1. The conditional PDF of X given $Y = y$ is:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} \tag{50}$$

2. The probability that X is in set A given $Y = y$ is:

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y)\, dx \tag{51}$$

3. The conditional CDF of X given $Y = y$ is:

$$F_{X|Y}(x \mid y) = P(X \leq x \mid Y = y) \tag{52}$$
$$= \int_{-\infty}^{x} f_{X|Y}(x \mid y)\, dx \tag{53}$$

4. The conditional expectation of X given $Y = y_j$ is:

$$E[X \mid Y = y_j] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y_j)\, dx \tag{54}$$

5. The conditional variance of X given $Y = y_j$ is:

$$V[X \mid Y = y_j] = E[X^2 \mid Y = y_j] - (E[X \mid Y = y_j])^2 \tag{55}$$

7.2.10 Law of the Unconscious Statistician (LOTUS)

The Law of the Unconscious Statistician (LOTUS) still applies for continuous random variables. It allows us to compute the expected value of a function of random variables.

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f_{XY}(x, y)\, dy\, dx \tag{56}$$

7.3 Covariance and Correlation

The covariance gives information about how X and Y are statistically related. It measures the degree to which two random variables change together.

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{57}$$
$$= E[XY] - (E[X])(E[Y]) \tag{58}$$

7.3.1 Properties

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
2. If X and Y are independent, $\mathrm{Cov}(X, Y) = 0$
3. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
4. $\mathrm{Cov}(\alpha X, Y) = \alpha\, \mathrm{Cov}(X, Y)$
5. $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$
6. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
7. $\mathrm{Cov}(X + Y, Z + W) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(X, W) + \mathrm{Cov}(Y, Z) + \mathrm{Cov}(Y, W)$
8. More generally:

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\, \mathrm{Cov}(X_i, Y_j) \tag{59}$$

If $Z = X + Y$:

9. $\mathrm{Var}(Z) = \mathrm{Cov}(Z, Z) = \mathrm{Cov}(X + Y, X + Y)$

$$= \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y) \tag{60}$$
$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X, Y) \tag{61}$$

10. More generally:

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab \cdot \mathrm{Cov}(X, Y) \tag{62}$$

7.3.2 Correlation Coefficient

The correlation coefficient, ρ_{XY} (greek letter Rho) or $\rho(X,Y)$, is obtained by normalizing the covariance. More specifically, we use the standardized versions of X and Y:

$$U = \frac{X - E[X]}{\sigma_X} \tag{63}$$

$$V = \frac{Y - E[Y]}{\sigma_Y} \tag{64}$$

The correlation coefficient is defined as:

$$\begin{aligned}
\rho_{XY} = Corr(X,Y) &= \text{Cov}(U,V) && (65)\\
&= \text{Cov}\left(\frac{X - E[X]}{\sigma_X}, \frac{Y - E[Y]}{\sigma_Y}\right) && (66)\\
&= \text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \quad \text{(property 5)} && (67)\\
&= \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} \quad \text{(property 4)} && (68)
\end{aligned}$$

8 Point Estimation and Sampling

8.1 Descriptive Statistics

Descriptive statistics summarizes data from a sample using parameters such as the mean or standard deviation. It also includes tables, graphs, etc.

8.2 Inferential Statistics

Inferential statistics infers predictions about a larger population than the sample represents. It uses patterns in the sample data to draw inferences about the population, accounting for randomness. These inferences may take the form of:

- Answering yes/no questions about the data (hypothesis testing)
- Estimating numerical characteristics of the data (estimation)
- Describing associations within the data (correlation)
- Modeling relationships within the data (regression analysis)

8.3 Population and Sample

8.3.1 Population

In statistics, a population is a complete set of items that share at least one property in common that is the subject of a statistical analysis. Populations can be diverse, such as “all persons living in a country” or “every atom composing a crystal.” In the cup filling example, the population is “all cups filled by the machine.”

8.3.2 Sample

A sample is a subset of the population used to make inferences about the entire population. Samples should be independent and identically distributed (i.i.d.).

8.3.3 Statistical Model

A statistical model represents the data generation process and the relationship between variables. It includes assumptions about the form of the data distribution and the relationships between variables.

8.4 Point Estimation

A point estimate is a reasonable value of a population parameter. Point estimates are random variables and functions of these random variables, called statistics, which have their unique distributions known as sampling distributions.

8.4.1 Some Parameters and Their Statistics

To estimate the mean of a population, we could choose:

- Sample mean
- Sample median
- Average of the largest and smallest observations in the sample

8.5 Maximum-Likelihood Estimation (MLE)

Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. For a given distribution with parameter θ and the pdf f, the likelihood function is:

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta) \tag{69}$$

Important points:

- The likelihood function $L(\theta)$ is a function of the parameter estimate θ and not x.
- The likelihood function $L(\theta)$ is the probability density of observing x if the true parameter is θ.
- The optimum (MLE) estimate of the parameter is the one that maximizes $L(\theta)$.

8.5.1 Maximum-Likelihood Estimate of the Mean

For i.i.d. Gaussian data, the log-likelihood function is used:

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{70}$$

Note: i.i.d. Gaussian data refers to a set of data points that are independent and identically distributed (i.i.d.), and each data point follows a Gaussian (or normal) distribution. Independence means that each data point does not influence any other, and identically distributed means that each data point is drawn from the same probability distribution, in this case, a normal distribution characterized by its mean and variance.

8.5.2 MLE for Binomial Data

For the binomial distribution, the likelihood function is:

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{71}$$

The MLE for p maximizes:

$$\log L(p) = k \log p + (n-k)\log(1-p) \tag{72}$$

8.6 Bias

Let $\hat{\theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for θ. The bias of point estimator $\hat{\theta}$ is defined by

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta. \tag{73}$$

Let $\hat{\theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for a parameter θ. We say that $\hat{\theta}$ is an **unbiased** estimator of θ if

$$B(\hat{\theta}) = 0, \quad \text{for all possible values of } \theta. \tag{74}$$

8.7 Popular Estimators

8.7.1 Sample Mean

The sample mean $\bar{X}$ is an estimator for the population mean μ:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{75}$$

The sample mean is an **unbiased** estimator of the population mean μ:

$$E[\bar{X}] = \mu. \tag{76}$$

8.7.2 Sample Variance

The sample variance s^2 is an estimator for the population variance σ^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{77}$$

The sample variance is an **unbiased** estimator of the population variance σ^2:

$$E[s^2] = \sigma^2. \tag{78}$$

8.7.3 Sample Proportion

The sample proportion $\hat{p}$ is an estimator for the population proportion p:

$$\hat{p} = \frac{k}{n}, \tag{79}$$

where k is the number of successes in the sample. The sample proportion is an **unbiased** estimator of the population proportion p:

$$E[\hat{p}] = p. \tag{80}$$

8.7.4 Maximum Likelihood Estimator (MLE) for Exponential Distribution

For a sample from an exponential distribution with rate parameter λ, the MLE for λ is:

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{X}}. \tag{81}$$

The MLE for λ is a **biased** estimator of the true rate parameter λ.

8.7.5 Maximum Likelihood Estimator (MLE) for Normal Distribution

For a sample from a normal distribution with mean μ and variance σ^2, the MLE for σ^2 is:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{82}$$

The MLE for σ^2 is a **biased** estimator of the true variance σ^2.

8.8 Sample Variance

The sample variance is an unbiased estimator of the population variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{83}$$

8.9 Standard Error

The standard error is a measure of the statistical accuracy of an estimate.

8.10 Central Limit Theorem (CLT)

The CLT states that the distribution of the sample mean approaches a normal distribution as the sample size becomes large.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with expected value $EX_i = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \tag{84}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is,

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x \in \mathbb{R}, \tag{85}$$

where $\Phi(x)$ is the standard normal CDF.

8.11 How to Apply The Central Limit Theorem (CLT)

Here are the steps that we need in order to apply the CLT:

1. Write the random variable of interest, Y, as the sum of n i.i.d. random variables X_i's:
$$Y = X_1 + X_2 + \cdots + X_n. \tag{86}$$
2. Find $E[Y]$ and $\mathrm{Var}(Y)$ by noting that
$$E[Y] = n\mu, \quad \mathrm{Var}(Y) = n\sigma^2, \tag{87}$$
where $\mu = EX_i$ and $\sigma^2 = \mathrm{Var}(X_i)$.
3. According to the CLT, conclude that
$$\frac{Y - E[Y]}{\sqrt{\mathrm{Var}(Y)}} = \frac{Y - n\mu}{\sqrt{n\sigma^2}} \tag{88}$$
is approximately standard normal; thus, to find $P(y_1 \leq Y \leq y_2)$, we can write
$$P(y_1 \leq Y \leq y_2) = P\left(\frac{y_1 - n\mu}{\sqrt{n\sigma^2}} \leq \frac{Y - n\mu}{\sqrt{n\sigma^2}} \leq \frac{y_2 - n\mu}{\sqrt{n\sigma^2}}\right) \tag{89}$$
$$\approx \Phi\left(\frac{y_2 - n\mu}{\sqrt{n\sigma^2}}\right) - \Phi\left(\frac{y_1 - n\mu}{\sqrt{n\sigma^2}}\right). \tag{90}$$

8.12 Skewness and Kurtosis

- Positive (right) skewed: skewness > 0
- Negative (left) skewed: skewness < 0
- Symmetrical: skewness $= 0$
- Mesokurtic: $-0.5 <$ kurtosis < 0.5
- Platykurtic: kurtosis < -0.5
- Leptokurtic: kurtosis > 0.5

8.13 Kernel Density Estimate (KDE)

KDE is a way to estimate the probability density function (PDF) of the random variable that "underlies" the sample.

8.14 Mean Squared Error of an Estimator

The mean squared error (MSE) is equal to the variance of the estimator plus the squared bias:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2 \tag{91}$$

The **mean squared error (MSE)** of a point estimator $\hat{\theta}$, shown by $MSE(\hat{\theta})$, is defined as

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]. \tag{92}$$

9 Statistical Intervals

9.1 Types of Interval Estimates

- Confidence intervals: An interval estimate that provides a range within which the true parameter is expected to lie with a certain level of confidence.
- Prediction intervals: An interval estimate that predicts the range within which future observations will fall with a specified probability.
- Tolerance interval: An interval estimate that covers a specified proportion of the population with a certain level of confidence.

9.2 Confidence Interval

An interval within which the true parameter is believed to lie. Our belief is associated with a degree of belief. The interval depends on:

1. The parameter of interest (μ, σ, p)
2. What is known (e.g., σ)

9.2.1 CI for Mean, True σ is Known, Data Normal

$$\text{CI: } \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{93}$$

$$\alpha = \text{error} \tag{94}$$

$$\text{Confidence level} = 1 - \alpha \tag{95}$$

$$\text{Upper: } \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{S.E.} \tag{96}$$

$$\text{Lower: } \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{97}$$

9.2.2 Z-scores for Different CI Levels

For a given confidence level, the Z-score defines the boundaries within which we expect a certain percentage of our data to fall. Here are common confidence levels and their corresponding Z-scores:

$$99\% : \quad Z_{1-\alpha/2} = 2.58 \tag{98}$$

$$95\% : \quad Z_{1-\alpha/2} = 1.96 \tag{99}$$

$$90\% : \quad Z_{1-\alpha/2} = 1.64 \tag{100}$$

This means that 90% of the data will fall within 1.64 standard deviations from the mean.

$$E = \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 \tag{101}$$

$$Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} : \text{margin of error} \tag{102}$$

$$\frac{\sigma}{\sqrt{n}} : \text{standard error} \tag{103}$$

$$2\left[Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] : \text{length of error} \tag{104}$$

9.2.3 Errors in Confidence Intervals

Margin of Error (E):

The margin of error represents the range above and below the sample mean in which we expect the true population parameter to lie.

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{105}$$

Example: For a 95% confidence interval with $\sigma = 10$, $n = 100$, and $Z = 1.96$:

$$E = 1.96 \times \frac{10}{\sqrt{100}} = 1.96 \times 1 = 1.96 \tag{106}$$

This means we expect the true population parameter to lie within 1.96 units of the sample mean.

Standard Error (SE):

The standard error measures the variability of the sample mean.

$$SE = \frac{\sigma}{\sqrt{n}} \tag{107}$$

Example: With $\sigma = 10$ and $n = 100$:

$$SE = \frac{10}{\sqrt{100}} = 1 \tag{108}$$

This is the standard deviation of the sampling distribution of the sample mean.

Length of Error (LE):

The length of error is twice the margin of error, covering both sides of the mean.

$$LE = 2\left[Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \tag{109}$$

Example: Continuing from the margin of error example, the length of error is:

$$LE = 2 \times 1.96 = 3.92 \tag{110}$$

This means the total width of the confidence interval is 3.92 units.

To construct a confidence interval for a mean when the population standard deviation (σ) is known and the data is normally distributed, you follow these steps:

1. Determine the desired confidence level (e.g., 95%).
2. Find the corresponding Z-score (e.g., 1.96 for 95%).
3. Calculate the standard error (SE).
4. Multiply the Z-score by the standard error to get the margin of error (E).
5. Add and subtract the margin of error from the sample mean to get the confidence interval.

Example: If your sample mean is 50, $\sigma = 10$, $n = 100$, and you want a 95% confidence interval:

$$E = 1.96 \times \frac{10}{\sqrt{100}} = 1.96 \tag{111}$$

$$\text{Confidence Interval} = 50 \pm 1.96 = (48.04, 51.96) \tag{112}$$

This means we are 95% confident that the true population mean lies between 48.04 and 51.96.

9.2.4 Finding n

Assume you need to find a sample size for some test/research.

1. Determine a confidence level ($Z_{\alpha/2}$)
2. Determine an allowed margin of error E

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 \tag{113}$$

9.2.5 One-Sided Confidence Bounds

One-sided confidence bounds are used when we are interested in estimating a boundary on only one side (either upper or lower) of the parameter. This can be useful in scenarios where we need to ensure that a parameter does not exceed a certain value (upper bound) or does not fall below a certain value (lower bound).

A $100(1-\alpha)\%$ **upper**-confidence bound for μ is

$$\mu \leq \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{114}$$

This formula calculates the upper bound of the confidence interval, indicating that with $100(1-\alpha)\%$ confidence, the true mean μ is less than or equal to $\bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$.

A $100(1-\alpha)\%$ **lower**-confidence bound for μ is

$$\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \tag{115}$$

This formula calculates the lower bound of the confidence interval, indicating that with $100(1-\alpha)\%$ confidence, the true mean μ is greater than or equal to $\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$.

9.2.6 Large Sample Confidence Interval (n $\geq$ 30 or $n \geq 40$)

For large samples, where $n \geq 30$ or $n \geq 40$, the confidence interval for the mean μ is given by:

$$\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{116}$$

This formula calculates the range within which the true mean μ is expected to lie with a specified level of confidence, taking into account the sample mean ($\bar{X}$), sample standard deviation (s), and sample size (n).

Similar considerations apply for one-sided intervals, meaning the approach to calculating the upper or lower bounds is analogous to the one-sided confidence bounds discussed previously.

9.2.7 CI on Mean, σ Unknown, Data Normal, n Small

When constructing a confidence interval for the mean with a small sample size (n is small), unknown population standard deviation (σ), and normally distributed data, we use a more conservative normal distribution called the t-distribution. The standard deviation is calculated from the sample.

The confidence interval is given by:

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{117}$$

where:

- $\bar{X}$ is the sample mean.
- $t_{\alpha/2,n-1}$ is the critical value from the t-distribution with $n-1$ degrees of freedom (T-Table).
- s is the sample standard deviation.
- n is the sample size.

9.2.8 Degree of Freedom

Definition: The degrees of freedom (df or dof) indicate the number of independent values that can vary in an analysis without breaking any constraints.

Since $\bar{X}$ was estimated and is used to estimate s, we have $n-1$ degrees of freedom (dof).

Note: The sample standard deviation S is calculated as:

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \tag{118}$$

However, NumPy does:

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N}} \tag{119}$$

Must pass `ddof=1` to get the correct sample standard deviation.

9.2.9 Proportions

In statistics, a proportion refers to the fraction of the total that exhibits a particular characteristic or attribute. It is a type of ratio that compares the part to the whole. Proportions are often used to describe the prevalence or frequency of an event within a sample or population.

Example: If you have a sample of 100 students and 40 of them are female, the proportion of female students in the sample is calculated as:

$$\hat{p} = \frac{\text{Number of females}}{\text{Total number of students}} = \frac{40}{100} = 0.4$$

9.2.10 Confidence Interval for Proportions:

The confidence interval for a proportion is used to estimate the range within which the true population proportion lies, based on a sample proportion. The formula for the confidence interval for a proportion is:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where:

- $\hat{p}$ is the sample proportion.
- $Z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.
- n is the sample size.

Conditions:
For the formula to be valid, certain conditions must be met:

$$n\hat{p} \geq 5 \quad \text{or} \quad n(1-\hat{p}) \geq 5 \tag{120}$$

These conditions ensure that the sample size is large enough for the normal approximation to the binomial distribution to be appropriate.

Practical Use:
Proportions are widely used in surveys, quality control, public health studies, and many other fields to quantify and analyze the frequency of specific outcomes or characteristics within a population. For example, proportions can be used to estimate the percentage of voters who support a particular candidate, the proportion of defective items in a batch, or the prevalence of a disease in a community.

9.2.11 CI for Variance/Standard Deviation

To construct a confidence interval for the population variance (σ^2), we use the chi-squared distribution (χ^2). The chi-squared distribution is asymmetric and provides critical values needed for the interval.

The confidence interval for the population variance σ^2 is calculated using the formula:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{121}$$

where:

- $(n-1)s^2$ is the degrees of freedom $(n-1)$ multiplied by the sample variance (s^2).
- $\chi^2_{\alpha/2,n-1}$ is the critical value from the chi-squared distribution for the upper tail with $n-1$ degrees of freedom (Chi-Squared Table).
- $\chi^2_{1-\alpha/2,n-1}$ is the critical value from the chi-squared distribution for the lower tail with $n-1$ degrees of freedom (Chi-Squared Table).

If you need the confidence interval for the population standard deviation (σ), you can obtain it by taking the square root of the bounds of the variance confidence interval:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}} \tag{122}$$

10 Hypothesis Testing

10.1 What is a hypothesis?

A hypothesis is a statement about the world that has a definite truth value:

- *Normative:*
  - "It is wrong to kill another human being."
  - "The prettiest city in Europe is Prague."
- *Descriptive:*
  - "The capital of Denmark is Horsens."
  - "The accused is guilty."
  - "The accused is innocent."

A hypothesis can either be rejected or not rejected.
A hypothesis can either be true or false.

10.2 Decision Making in Hypothesis Testing

Decision/Actual	**Hypothesis is True**	**Hypothesis is False**
Fail to reject (accept)	Correct True Positive (TP) $1-\alpha$ (Confidence level)	Error (Type II error) False Negative (FN) β
Reject	Error (Type I error) False Positive (FP) α (Level of significance)	Correct True Negative (TN) $1-\beta$ (Power)

$$P(\text{Reject} H_0 \mid H_0 \text{ is false}) = 1 - \beta \tag{123}$$

$$P(\text{Not reject} H_0 \mid H_0 \text{ is false}) = \beta \tag{124}$$

$$P(\text{Reject} H_0 \mid H_0 \text{ is true}) = \alpha \tag{125}$$

$$P(\text{Not reject} H_0 \mid H_0 \text{ is true}) = 1 - \alpha \tag{126}$$

10.3 Null and Alternative Hypotheses

10.3.1 Null Hypothesis

Null hypothesis (H_0): is a statement that there is no effect or no difference, and it serves as the default or starting assumption. It is the hypothesis that we attempt to disprove or reject through our statistical test.

10.3.2 Alternative hypothesis

Alternative hypothesis (H_1): is a statement that indicates the presence of an effect or a difference. It is what you want to prove or support through your statistical analysis. True if H_0 is false.

10.3.3 Example

$$H_0 : \mu_t \leq \mu_m \tag{127}$$

$$H_1 : \mu_t > \mu_m \tag{128}$$

10.4 Overall Method

1. State H_0 and H_1 and try to disprove H_0.
2. Determine α (Researcher decides).
3. Calculate the test statistic.
4. Based on the test statistic, reject or fail to reject H_0.
5. Conclude.

You cannot prove H_0, only reject or fail to reject.

10.5 Z-Test

The Z-test is a statistical test used to determine whether there is a significant difference between the sample mean and the population mean. It is used when the sample size is large ($n > 30$) and the population standard deviation is known.

10.5.1 Steps in Z-Test

1. **State the hypotheses:**
   - Null hypothesis (H_0): $\mu = \mu_0$
   - Alternative hypothesis (H_1): $\mu \neq \mu_0$ (for a two-tailed test)
2. **Determine the significance level** (α): This is the probability of rejecting the null hypothesis when it is true. Common choices are 0.05, 0.01, or 0.10.
3. **Calculate the test statistic**:
$$Z_{\text{test}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{129}$$
   Where $\overline{X}$ is the sample mean, μ_0 is the population mean, σ is the population standard deviation, and n is the sample size.
4. **Determine the critical value** (Z_{crit}) based on α: This value can be found in Z-tables.
5. **Make a decision**:
   - For a two-tailed test: If $|Z_{\text{test}}| > Z_{\text{crit}}$, reject H_0; otherwise, fail to reject H_0.
   - For a right-tailed test: If $Z_{\text{test}} > Z_{\text{crit}}$, reject H_0; otherwise, fail to reject H_0.
   - For a left-tailed test: If $Z_{\text{test}} < -Z_{\text{crit}}$, reject H_0; otherwise, fail to reject H_0.
6. **Conclude**: Based on the decision, conclude whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

10.5.2 Formulas

- **Two-tailed test:**
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$
- **Right-tailed test:**
$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$
- **Left-tailed test:**
$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

10.5.3 Rejection Criteria

- Two-tailed test: Reject H_0 if $|Z_{\text{test}}| > Z_{\text{crit}}$
- Right-tailed test: Reject H_0 if $Z_{\text{test}} > Z_{\text{crit}}$
- Left-tailed test: Reject H_0 if $Z_{\text{test}} < -Z_{\text{crit}}$

10.6 P-Value in Hypothesis Testing

The p-value is a measure of the probability that an observed difference could have occurred just by random chance. It helps to determine the significance of the results obtained in a hypothesis test.

10.6.1 Steps involving the p-value

1. **Calculate the test statistic**:

$$Z_{\text{test}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{130}$$

2. **Find the p-value from the test statistic**:

   - For a two-tailed test:

$$\text{p-value} = 2 \cdot P(Z > |Z_{\text{test}}|) = 2 \cdot (1 - P(Z < |Z_{\text{test}}|)) \tag{131}$$

   - For a right-tailed test:

$$\text{p-value} = P(Z > Z_{\text{test}}) = 1 - P(Z < Z_{\text{test}}) \tag{132}$$

   - For a left-tailed test:

$$\text{p-value} = P(Z < Z_{\text{test}}) \tag{133}$$

   - The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

3. **Compare the p-value to the significance level** (α):

   - If the p-value $< \alpha$, reject H_0.
   - If the p-value $\geq \alpha$, fail to reject H_0.

10.6.2 Decision Rule

- Reject H_0 if p-value $< \alpha$.
- Fail to reject H_0 if p-value $\geq \alpha$.

The p-value provides a method to quantify the strength of the evidence against the null hypothesis. Lower p-values indicate stronger evidence against the null hypothesis.

$$\text{p-value} = P(\text{observed or more extreme data} \mid H_0 \text{ is true}) \tag{134}$$

10.7 Types of Tests

1. **Independent samples:**

   - This test is used to compare the means of two independent groups to determine if there is statistical evidence that the associated population means are significantly different.
   - Example: Comparing the test scores of two different classes to see if one class performs better than the other.

2. **Paired test:**

   - This test is used to compare the means of the same group at different times. It is also known as the dependent sample test.

  - Example: Measuring the weight of patients before and after a treatment to see if there is a significant change.

3. **One sample test:**
   - This test is used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean.
   - Example: Testing if the average IQ score of a class is different from the national average IQ score.

4. **Two variables:**
   - This test is used to determine if there is a significant association between two categorical variables. It is often conducted using the chi-square test for independence.
   - Example: Examining if there is an association between gender (male/female) and preference for a new product (like/dislike).

10.8 What Can We Test?

10.8.1 One Sample

- **Mean:**
  - Use the Z-test or t-test to compare the sample mean to a known population **mean**.
- **Proportion:**
  - Use the Z-test to compare the sample **proportion** to a known population proportion.
- **Variance/Standard Deviation:**
  - Use the chi-square (χ^2) test to compare the sample **variance** to a known population variance.

10.8.2 Two Samples

- **Compare Means:**
  - Compare **means** of two **distinct groups**.
  - Compare **means** of the **same group** at different times.
- **Compare Proportions:**
  - Compare the **proportions** of **two distinct** groups.
- **Compare Variances/Standard Deviations:**
  - Compare the **variances** or standard deviations of **two distinct groups**.

10.8.3 Two Variables

- **Compare Independence:**
  - Test the **independence** between two categorical variables using tests such as the chi-square test for independence.

10.9 Test for Independence

The test for independence is used to determine whether two categorical variables are independent of each other. This can be done using the chi-square test for independence.

10.9.1 Discrete Probability Mass Function (PMF)

Consider two discrete random variables X and Y with observed values represented in a contingency table:

	$y=0$	$y=1$	$y=2$	$f_X(x)$
$x=0$	a_{11}	a_{12}	a_{13}	$f_X(0)$
$x=1$	a_{21}	a_{22}	a_{23}	$f_X(1)$
$x=2$	a_{31}	a_{32}	a_{33}	$f_X(2)$
$f_Y(y)$	$f_Y(0)$	$f_Y(1)$	$f_Y(2)$	

Table 5: Observed Values

10.9.2 Expected Values

Based on the marginal probabilities, we can compute the expected values (E_{ij}) under the null hypothesis (H_0) that X and Y are independent:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}} \tag{135}$$

For example:

$$E_{11} = \frac{f_X(0) \times f_Y(0)}{\text{Grand Total}}$$
$$E_{12} = \frac{f_X(0) \times f_Y(1)}{\text{Grand Total}}$$
$$E_{33} = \frac{f_X(2) \times f_Y(2)}{\text{Grand Total}}$$

10.9.3 Hypotheses

- Null hypothesis (H_0): X and Y are independent.
- Alternative hypothesis (H_1): X and Y are dependent.

10.9.4 Chi-Square Test Statistic

The chi-square test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{136}$$

Where O_{ij} represents the observed frequencies and E_{ij} represents the expected frequencies under H_0.

10.9.5 Determine the Critical Value

To determine the critical value for the chi-square test, use the chi-square distribution table:

1. Calculate the degrees of freedom: $df = (r-1) \times (c-1)$
   - Where r is the number of rows and c is the number of columns.
2. Choose the significance level (α), commonly 0.05.
3. Look up the critical value in the chi-square distribution table using the calculated degrees of freedom and the chosen significance level (Chi-Squared Table).

10.9.6 Decision Rule

Compare the calculated chi-square statistic to the critical value from the chi-square distribution table with appropriate degrees of freedom:

- If the calculated chi-square is greater than the critical value, reject H_0.
- If the calculated chi-square is less than or equal to the critical value, fail to reject H_0.

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{if } A \text{ and } B \text{ are independent} \tag{137}$$

11 Linear Regression

Linear regression is a statistical method used to **model the relationship** between a dependent variable and one or more independent variables. The goal of linear regression is to **find the best-fitting linear equation that predicts the dependent variable based on the independent variables**. This method assumes a linear relationship between the variables and is widely used for predictive analysis, understanding the strength of associations, and determining the impact of independent variables on a dependent variable.

11.1 Contingency Table Tests

Contingency Table Tests allow us to explore the **association between two categorical variables**. These tests help us understand if there is a significant relationship between the variables. For example, we might want to see if there is an association between gender and preference for a certain color:

$$R_Y = \{\text{Blue, Green, Red, Male, Turkish}\}$$

By analyzing the frequencies of occurrences in different categories, we can determine if the variables are independent or related.

11.2 Regression Analysis

Regression analysis allows us to explore the **association between** two variables, but instead of categorical variables, we use **numeric variables**. This method helps in predicting the value of a dependent variable based on the value(s) of one or more independent variables. For example:

$$Y = \{2.18, 4.28, 7.92, \ldots\}$$

$$X = \{1.1, 1.2, 1.4, \ldots\}$$

By fitting a linear equation to the observed data, we can make predictions and infer relationships between the variables.

11.3 Random Variables

In statistical analysis, variables can be classified as either discrete or continuous, depending on the nature of the data.

- **Discrete:** Contingency Tables - Used for analyzing the association between categorical variables. Discrete random variables take on a countable number of distinct values.
- **Continuous:** Regression - Used for analyzing the relationship between continuous variables. Continuous random variables can take any value within a given range.

11.4 Linear Regression

11.4.1 Linear Regression Equation

The general form of a simple linear regression equation is:

$$y = ax + b$$

where y is the dependent variable, x is the independent variable, a is the slope of the line, and b is the y-intercept.

For multiple linear regression, the equation extends to:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots$$

where $\hat{y}$ is the predicted value, and $\hat{\beta}_i$ are the estimated coefficients for each independent variable x_i.

11.4.2 Design Matrix

The association between X and $\hat{y}$ can be represented using a design matrix X:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

where each row represents an observation and each column represents a variable, including a column of 1s for the intercept term.

The regression coefficients can be represented as a vector β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}$$

The predicted values $\hat{y}$ can then be calculated as:

$$\hat{y} = X\beta$$

11.5 Performing Linear Regression - Step by Step

In this section, we will go through the step-by-step process of performing a linear regression analysis. This includes visual inspection of data, hypothesis testing, and fitting the appropriate model to understand the relationship between the variables.

11.5.1 Scatter Plot of X and Y to Visually Inspect Relationship

Creating a scatter plot of X and Y allows us to visually inspect the relationship between the two variables. This helps in identifying the type of relationship and choosing the appropriate model for regression.

- **Quadratic Relationship:**

$$H_0 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \quad \hat{\beta}_2 > 0$$

- **Positive Linear Relationship:**

$$H_0 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \hat{\beta}_1 > 0$$

- **Negative Linear Relationship:**

$$H_0 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \hat{\beta}_1 < 0$$

- **Exponential Relationship:**

$$H_0 : \hat{y} = \hat{\beta}_0 \cdot e^{\hat{\beta}_1 x}$$

11.5.2 Remove Outliers

Removing outliers is an important step in regression analysis. Outliers can disproportionately affect the results of the regression model, leading to inaccurate predictions. In regression, this is often done qualitatively by visually inspecting the scatter plot and identifying points that do not fit the overall pattern of the data.

11.5.3 Determine Regression Equation, i.e., Estimate β_0 and β_1

The regression equation is given by:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ represents the error term, which accounts for the variability in Y that cannot be explained by the linear relationship with X.

To estimate the coefficients β_0 and β_1, we use the following formulas:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The expected value of Y is:

$$E[Y] = \beta_0 + \beta_1 E[X] + E[\epsilon]$$

the error term ϵ represents the deviation of the observed values from the values predicted by the model. The assumption that $E[\epsilon] = 0$ means that, on average, the errors cancel out. This implies that the model's predictions are unbiased, meaning there is no systematic overestimation or underestimation of the dependent variable Y.

Since $E[\epsilon] = 0$, this simplifies to:

$$E[Y] = \beta_0 + \beta_1 E[X]$$

Solving for β_0:

$$\beta_0 = E[Y] - \beta_1 E[X]$$

The covariance between X and Y is given by:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, \beta_0 + \beta_1 X + \epsilon) && (138)\\
&= \mathrm{Cov}(X, \beta_0) + \mathrm{Cov}(X, \beta_1 X) + \mathrm{Cov}(X, \epsilon) && (139)\\
&= \beta_0 \mathrm{Cov}(X, 1) + \beta_1 \mathrm{Cov}(X, X) + 0 && (140)\\
&= \beta_1 \mathrm{Var}(X) && (141)
\end{aligned}$$

Solving for β_1:

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

To estimate the coefficients β_0 and β_1, we start with the following definitions and formulas:

The sample means $\bar{x}$ and $\bar{y}$ are defined as:

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$$

The covariance between X and Y and the variance of X are given by:

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Var}(X) = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

Let's factor out the $\frac{1}{1-n}$, we introduce the following terms:

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

Thus, the slope coefficient β_1 can be estimated as:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{142}$$
$$= \frac{\sum(y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{143}$$
$$= \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - \left(\frac{(\sum x_i)^2}{n}\right)} \tag{144}$$
$$= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \tag{145}$$

Finally, the intercept β_0 can be estimated as:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

11.5.4 Check Assumptions that Errors are Normally Distributed

To ensure the validity of the regression model, we check if the errors (residuals) are normally distributed. This can be done using a normal probability plot.

The residual (error) for each observation is defined as:

$$e_i = y_i - \hat{y}_i$$

where

- y_i: observed value
- $\hat{y}_i$: predicted value
- $\bar{y}$: average value

The total deviation can be broken down into:

$$\text{Total Deviation} = \text{Unexplained Deviation} + \text{Explained Deviation} \tag{146}$$
$$\text{Total Deviation} = y_i - \bar{y} = \text{obs} - \text{avg} \tag{147}$$
$$\text{Explained Deviation} = \hat{y}_i - \bar{y} = \text{pred} - \text{avg} \tag{148}$$
$$\text{Unexplained Deviation} = y_i - \hat{y}_i = \text{obs} - \text{pred} \tag{149}$$
$$\tag{150}$$

11.5.5 Sum of Squares

We calculate the following sum of squares to understand the variability in the data:

Total Sum of Squares (SS_T):

$$SS_T = \sum(y_i - \bar{y})^2 = \sum(\text{Total Deviation})^2 \tag{151}$$

Regression Sum of Squares (SS_R):

$$SS_R = \sum(\hat{y}_i - \bar{y})^2 = \sum(\text{Explained Deviation})^2 \tag{152}$$

Residual Sum of Squares (SS_E):

$$SS_E = \sum(y_i - \hat{y}_i)^2 = \sum(\text{Unexplained Deviation})^2 \tag{153}$$

The goal is to minimize the Residual Sum of Squares (SS_E) to improve the accuracy of the regression model.

11.5.6 Assess Adequacy of Model

a) Hypothesis Testing To assess the adequacy of the model, we first test the hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The test statistic is given by:

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

where

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

b) Determine Correlation To determine the correlation coefficient r, we can use different methods. Here are some of them:

1) Using z scores of all X and Y:

$$r = \frac{\sum (z_{x_i} z_{y_i})}{n-1}$$

2) Direct calculation using observed values:

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

3) Using the sum of squares:

$$r = \sqrt{\frac{SS_R}{SS_T}} = \sqrt{1 - \frac{SS_E}{SS_T}}$$

4) Direct calculation using observed values:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

This formula simplifies to:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Where:

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \tag{154}$$
$$S_{xx} = \sum (x_i - \bar{x})^2 \tag{155}$$
$$S_{yy} = \sum (y_i - \bar{y})^2 \tag{156}$$

The strength of the correlation can be assessed as:

$$|r| > 0.8 : \text{High} \tag{157}$$
$$0.6 < |r| \leq 0.8 : \text{Good} \tag{158}$$
$$0.45 < |r| \leq 0.6 : \text{OK} \tag{159}$$
$$\tag{160}$$

c) Find Correlation of Determination The correlation of determination r^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variable. It is calculated as the square of the correlation coefficient r:

$$r^2 = (\text{correlation coefficient})^2$$

Alternatively, it can be expressed in terms of the sum of squares:

$$r^2 = \frac{SS_R}{SS_T}$$

Where:

$$SS_R = \text{Regression Sum of Squares (Explained variance)}$$

$$SS_T = \text{Total Sum of Squares (Total variance)}$$

The coefficient of determination r^2 represents the amount of variance in the dependent variable that the model is able to explain.

11.5.7 Find Confidence Intervals for β_0 and β_1

To determine the reliability of the estimated coefficients β_0 and β_1, we calculate their confidence intervals.

Error of Slope The error of the slope E_S is given by:

$$E_S = t_{\alpha/2,n-2} \cdot \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Where:

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

$$t_{\alpha/2,n-2}$$

is the critical value from the t-distribution (T-Table) with $n-2$ degrees of freedom.

The confidence interval for β_1 is:

$$[\hat{\beta}_1 - E_S, \hat{\beta}_1 + E_S]$$

Error of Intercept The error of the intercept E_I is given by:

$$E_I = t_{\alpha/2,n-2} \cdot \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

Where:

$$\bar{x} = \frac{\sum x_i}{n}$$

The confidence interval for β_0 is:

$$[\hat{\beta}_0 - E_I, \hat{\beta}_0 + E_I]$$

11.5.8 Prediction Intervals

To predict the interval within which a new observation will fall, we use the prediction interval formula.

The error of prediction E_P is given by:

$$E_P = t_{\alpha/2,n-2} \cdot \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Where: x_0 is the x-value we want to find $\hat{y}$ at.

The prediction interval for a given x_0 is:

$$[x_0 - E_P, x_0 + E_P]$$

12 Stochastic Processes

A stochastic process (or random process) is a collection of random variables indexed by time (or space). Stochastic processes can be categorized into two types based on the nature of the index set: discrete and continuous.

12.1 Types of Stochastic Processes

- **Discrete Stochastic Processes:** The random variables are indexed by discrete time points, such as $X_0, X_1, X_2, \ldots, X_n$.
- **Continuous Stochastic Processes:** The random variables are indexed by continuous time, denoted as $\{X_t\}_{t\geq 0}$.

A random process is **recurrent** if we are guaranteed to return to the origin and is **transient** if the probability of not returning is greater than zero.

12.2 Counting Process

A counting process is a type of stochastic process $\{N(t), t \in [0, \infty)\}$ that tracks the number of events $N(t)$ occurring over time t. It is characterized by the following key properties:

1. $N(0) = 0$: The count starts at zero.
2. $N(t) \in \{0, 1, 2, \ldots\}$ for all $t \in [0, \infty)$: The number of events at any time tt is a non-negative integer.
3. For $0 \leq s \leq t$, $N(t) - N(s)$ shows the number of events that occur in the interval $(s, t]$, i.e., from but not including s up to and including t.

12.2.1 Types of Increments

- A counting process exhibits **independent increments** if the number of events occurring in non-overlapping intervals are independent of each other.
- A counting process exhibits **stationary increments** if the distribution of the number of events occurring between any two time points depends only on the length of the interval, not on the specific starting and ending times. Mathematically, $N(t_2) - N(t_1)$ has the same distribution as $N(t_2 - t_1)$ for $t_2 \geq t_1 \geq 0$.

12.3 Poisson Process

A counting process $\{N(t), t \in [0, \infty)\}$ is called a Poisson process with rate $\lambda > 0$ if it satisfies the following conditions:

1. $N(0) = 0$: The process starts with zero events at time zero.
2. $N(t)$ has independent increments: The number of events occurring in non-overlapping intervals are independent of each other.
3. The number of events occurring in any interval of length $\tau > 0$ follows a Poisson distribution with parameter $\lambda\tau$:

$$P(X = x) = \frac{e^{-\lambda\tau}(\lambda\tau)^x}{x!} \tag{161}$$

This formula is used to calculate the probability of observing exactly x events in a given time interval τ for a Poisson process with rate λ.

Here, λ represents the average rate at which events occur per unit time, and τ represents the length of the time interval under consideration. The product $\lambda\tau$ gives the expected number of events occurring in the interval of length τ. This relationship indicates that the probability of observing a certain number of events x in a given interval is governed by the Poisson distribution with parameter $\lambda\tau$.

12.3.1 Interarrival Times

Given a Poisson process $N(t)$, the times between successive events $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables. Specifically, each X_i follows an exponential distribution with rate parameter λ.

The probability that the time until the first event X_1 is greater than t is given by the probability that no events occur in the interval $(0, t]$:

$$\begin{aligned} P(X_1 > t) &= P(\text{No event in } (0, t]) && (162) \\ &= e^{-\lambda t} && (163) \end{aligned}$$

We conclude that

$$F_{X_i}(t) = \begin{cases} 1 - e^{-\lambda t}, & t > 0 \\ 0, & \text{otherwise} \end{cases} \tag{164}$$

is the cumulative distribution function (CDF) of the interarrival time X_i, denoted by $F_{X_i}(t)$.

The probability density function (PDF) of the interarrival time X_i, denoted by $f_{X_i}(t)$, is given by:

$$f_{X_i}(t) = \lambda e^{-\lambda t}, \quad t > 0 \tag{165}$$

The PDF $f_{X_i}(t)$ calculates the likelihood of the time between successive events being exactly t. The CDF $F_{X_i}(t)$ calculates the probability that the time between successive events is less than or equal to t. In other words, the CDF gives the probability that an event occurs within a specified time interval t.

12.3.2 Memoryless Property

The memoryless property of an exponential distribution states that the probability of an event occurring after a certain time a, given that it has not occurred up to time a, is the same as the probability of the event occurring after time x from the start. Mathematically, this is expressed as:

$$P(X > x + a \mid X > a) = P(X > x), \quad a, x \geq 0 \tag{166}$$

This property indicates that the exponential distribution does not "remember" how much time has already passed. The probability of waiting an additional x units of time is the same regardless of how much time has already elapsed.

12.3.3 Merging Independent Poisson Processes

Let $N_1(t), N_2(t), \ldots, N_m(t)$ be m independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_m$. Define $N(t)$ as the sum of these independent Poisson processes:

$$N(t) = N_1(t) + N_2(t) + \ldots + N_m(t), \quad t \in [0, \infty) \tag{167}$$

Then, $N(t)$ is a Poisson process with rate $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_m$.

This means that if we have multiple independent Poisson processes each counting events at their own rates, the combined process, which counts all events from the individual processes together, will also be a Poisson process. The rate of the combined process is simply the sum of the individual rates. This property is useful in scenarios where multiple independent sources of events are being considered together.

12.3.4 Example: Merging Independent Poisson Processes

Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates $\lambda_1 = 1$ and $\lambda_2 = 2$, respectively.

a. Find $N(1) = 2$ and $N(2) = 5$. The probability that $N(1) = 2$ and $N(2) = 5$ can be found as follows:

$$\begin{aligned}
P(N(1) = 2 \text{ and } N(2) = 5) &= P(\text{Two events in } (0,1] \text{ and five events in } (0,2]) && (168)\\
&= P(\text{Two events in } (0,1] \text{ and three events in } (1,2]) && (169)\\
&= P(N(1) = 2) \cdot P(N(2) = 5) && (170)
\end{aligned}$$

Using the Poisson distribution formula for each interval, we get:

$$\begin{aligned}
P(N(1) = 2) &= \frac{e^{-3} \cdot 3^2}{2!} && (171)\\
P(N(2) = 5) &= \frac{e^{-3} \cdot 3^5}{5!} && (172)
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(N(1) = 2) \cdot P(N(2) = 5) &= \frac{e^{-3} \cdot 3^2}{2!} \cdot \frac{e^{-3} \cdot 3^5}{5!} && (173)\\
&\approx 0.05 && (174)
\end{aligned}$$

b. Calculate $P(N_1(1) = 1 \mid N(1) = 2)$ Determine the probability that exactly one event is from N_1 in the interval $(0, 1]$, given that there are exactly two total events from both N_1 and N_2 (since $N = N_1 + N_2$) in $(0, 1]$:

This analysis looks at the likelihood that out of the two total events (from the combined process N), one event specifically originates from N_1, which has a lower event rate compared to N_2.

$$\begin{aligned}
P(N_1(1) = 1 \mid N(1) = 2) &= \frac{P(N_1(1) = 1 \text{ and } N(1) = 2)}{P(N(1) = 2)} && (175)\\
&= \frac{P(N_1(1) = 1) \cdot P(N_2(1) = 1)}{P(N(1) = 2)} && (176)
\end{aligned}$$

We use the Poisson distribution to compute the individual probabilities:

$$\begin{aligned}
P(N_1(1) = 1) &= \frac{e^{-1} \cdot 1^1}{1!} = e^{-1} && (177)\\
P(N_2(1) = 1) &= \frac{e^{-2} \cdot 2^1}{1!} = 2e^{-2} && (178)\\
P(N(1) = 2) &= \frac{e^{-3} \cdot 3^2}{2!} = \frac{9e^{-3}}{2} && (179)
\end{aligned}$$

Thus,

$$\begin{aligned}
P(N_1(1) = 1 \mid N(1) = 2) &= \frac{e^{-1} \cdot 2e^{-2}}{\frac{9e^{-3}}{2}} && (180)\\
&= \frac{2e^{-3}}{\frac{9e^{-3}}{2}} && (181)\\
&= \frac{2}{\frac{9}{2}} && (182)\\
&= \frac{4}{9} && (183)
\end{aligned}$$

This result indicates that if there are two events occurring in the interval $(0, 1]$, the probability that exactly one of these events is from N_1 is $\frac{4}{9}$.

13 Markov Chains

Markov chains are mathematical systems that undergo transitions from one state to another on a state space. Two important concepts and properties are associated with Markov chains:

13.1 Concepts

1. **States**: The possible conditions or states in which the system can be.
2. **Transition Probabilities**: The probabilities of moving from one state to another. These can be represented using:
   - State transition matrix
   - State transition diagram

13.2 Properties

13.2.1 Memoryless (Markov Property)

The future state depends only on the current state and not on the sequence of events that preceded it. Mathematically, this is expressed as:

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots) = P(X_{n+1} = j \mid X_n = i) \tag{184}$$

13.2.2 Time Homogeneous

The transition probabilities are independent of time. This means that the probability of transitioning from state i to state j is the same at all times. Mathematically, this is expressed as:

$$p_{ij} = P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i) = P(X_2 = j \mid X_1 = i) \tag{185}$$

13.3 Special Markov Chains

13.3.1 Irreducible

Every state in the Markov chain can be reached from every other state in a finite number of steps. This means there is a non-zero probability of transitioning from any state to any other state, possibly over multiple steps.

13.3.2 Aperiodic

The Markov chain does not cycle in a fixed period. For every state, the greatest common divisor of the lengths of the paths that return to that state is 1. This ensures that the chain does not get trapped in periodic behavior.

13.3.3 Positive Recurrence

The expected return time to any state is finite.

13.3.4 Ergodic

An irreducible, aperiodic Markov Chain with positive recurrence

13.4 Initial State: Row Vector

In a Markov chain, the initial state distribution is represented as a row vector. This vector indicates the probabilities of the system being in each possible state at the initial time (usually $t = 0$).

$$\pi^{(0)} = \begin{bmatrix} P(X_0 = 1) & P(X_0 = 2) & \ldots & P(X_0 = r) \end{bmatrix} \tag{186}$$

Here, $\pi^{(0)}$ is the initial state distribution vector, and $P(X_0 = i)$ represents the probability that the system is in state i at time $t = 0$.

The state distribution after n steps, denoted as $\pi^{(n)}$, can be computed by multiplying the initial state vector $\pi^{(0)}$ by the n-th power of the state transition matrix P:

$$\pi^{(n)} = \pi^{(0)} \cdot P^n \tag{187}$$
$$= \begin{bmatrix} P(X_n = 1) & P(X_n = 2) & \dots & P(X_n = r) \end{bmatrix} \tag{188}$$

This equation allows us to determine the probabilities of being in each state after n transitions, starting from the initial state distribution. The matrix P contains the transition probabilities between states, and P^n represents the transition probabilities over n steps.

13.5 n-Step Transition Matrix

In a Markov chain, the n-step transition matrix, denoted as $P^{(n)}$, represents the probabilities of transitioning from one state to another in n steps. This matrix is obtained by raising the one-step transition matrix P to the power of n:

$$P^{(n)} = P^n = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{bmatrix}^n \tag{189}$$

Each element $p_{ij}^{(n)}$ of the matrix $P^{(n)}$ represents the probability of transitioning from state i to state j in exactly n steps.

13.6 Types of Calculations in Markov Chains

In Markov chains, there are three primary types of calculations that help determine the probabilities of various state transitions:

13.6.1 Probability of being in a specific state after n steps

$$P(X_n = j) = \pi^{(0)} \cdot P^n \tag{190}$$

This calculation determines the probability of the system being in state j after n steps, starting from the initial state distribution $\pi^{(0)}$.

13.6.2 Conditional probability of being in a specific state after $m + n$ steps

$$P(X_{m+n} = j \mid X_m = i) = P^n \tag{191}$$

This calculation determines the probability of the system being in state j after $m + n$ steps, given that it was in state i after m steps. This uses the n-step transition matrix.

13.6.3 Joint probability of being in specific states at different times

$$P(X_{t_1} = i_1, X_{t_2} = i_2, \dots, X_{t_k} = i_k) \tag{192}$$

Assume $t_1 > t_2 > \dots > t_k$:

$$P(X_{t_1} = i_1, X_{t_2} = i_2, \dots, X_{t_k} = i_k) = \pi_{i_k}^{(0)} \cdot p_{i_k i_{k-1}} \cdot p_{i_{k-1} i_{k-2}} \cdot \dots \cdot p_{i_2 i_1} \tag{193}$$

This calculation determines the joint probability of the system being in state i_k at time t_k, state i_{k-1} at time t_{k-1}, and so on, until state i_1 at time t_1. It involves multiplying the initial state probability by the appropriate transition probabilities.

13.7 Communication Relation

The communication relation between states in a Markov chain satisfies the properties of an equivalence relation:

1. $i \leftrightarrow$ for every i (reflexive)
2. If $i \leftrightarrow j$, then $j \leftrightarrow i$ (symmetric)
3. If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$ (transitive)

13.7.1 Communicating Classes

In a Markov chain, states can be grouped into communicating classes based on their accessibility. If you can get from state i to state j and back, we say that i and j belong to the same class.

- State j is **accessible** from state i if: $i \to j$
- States i and j are said to **communicate** if: $i \leftrightarrow j$

13.7.2 Special Types of Communicating Classes

- If the entire state space S forms one communicating class, the Markov chain is called **irreducible**.
- A communicating class is **closed** if no state outside the class is accessible from any state within the class.
- If a state i forms a communicating class by itself and that class is closed, then state i is called **absorbing**.

13.8 Periodicity

In a Markov chain, a state i is said to have a period d_i, defined as:

$$d_i = \gcd\{n \in \mathbb{N} \mid p_{ii}^{(n)} > 0\} \tag{194}$$

This means that d_i is the greatest common divisor of all step counts n for which there is a positive probability of returning to state i after n steps.

- If $d_i > 1$, the state i is called **periodic**. This indicates that the state can only be revisited at multiples of d_i steps.
- If $d_i = 1$, the state i is called **aperiodic**. This means that the state can be revisited at irregular intervals.

13.8.1 Communication and Periodicity

If states i and j communicate ($i \leftrightarrow j$), then their periods are equal ($d_i = d_j$). This property ensures that states within the same communicating class share the same periodic behavior.

13.8.2 Aperiodicity Condition

A state i is aperiodic if the greatest common divisor of the lengths of two returning paths is 1. Formally, if:

$$\gcd(\ell, m) = 1, \quad \text{and } p_{ii}^{(\ell)} > 0 \text{ and } p_{ii}^{(m)} > 0 \tag{195}$$

This condition implies that state i can be revisited at irregular intervals.

13.8.3 Self-Transition and Aperiodicity

Any state with a self-transition (i.e., $p_{ii} > 0$) is aperiodic. This is because:

$$\gcd(1, a) = 1 \tag{196}$$

Thus, the state can be revisited in a single step, ensuring aperiodicity.

13.9 Absorption

In Markov chains, an absorbing state is a state that, once entered, cannot be left.

13.9.1 Absorption Probabilities

Let a_i denote the absorption probability in state 1 if the process starts from state i. Formally, we define:

$$a_i = P(\text{absorption in state } 1 \mid X_0 = i) \tag{197}$$

13.9.2 Calculating Absorption Probabilities

To find the absorption probabilities $a_0, a_1, \ldots$, we can apply the law of total probability recursively. The absorption probability a_i can be calculated using the following formula:

$$a_i = \sum_k a_k \cdot p_{ik}, \quad \text{for } i = 0, 1, 2, \ldots \tag{198}$$

where p_{ik} is the transition probability from state i to state k.

This formula states that the absorption probability a_i is the sum of the absorption probabilities of all possible next states, weighted by the probabilities of transitioning to those states from i. The recursive nature of this formula allows us to compute a_i based on known absorption probabilities of other states.

13.9.3 Boundary Conditions

The boundary conditions are defined based on the properties of the absorbing states:

$$a_i = 1 \quad \text{if state } i \text{ is the destination absorbing state} \tag{199}$$

$$a_i = 0 \quad \text{if state } i \text{ can never reach the destination absorbing state} \tag{200}$$

13.10 Mean Hitting Times

Let $A \subset S$ be a set of states in the Markov chain, and let T be the first time the chain visits A. The mean hitting time t_i is the expected number of steps needed for the chain to reach the set A starting from state i.

The mean hitting time t_i is given by:

$$t_i = E[T \mid X_0 = i] = 1 + \sum_k p_{ik} t_k \tag{201}$$

where t_i is the number of steps needed until the chain hits the given state or set of states, and p_{ik} is the transition probability from state i to state k.

13.10.1 Boundary Conditions

The boundary conditions for the mean hitting times are as follows:

$$t_i = 0 \quad \text{if state } i \in A \tag{202}$$
$$t_i = 0 \quad \text{if state } i \text{ is absorbing and } i \notin A \tag{203}$$

If the initial state i is already in the target set A, the mean hitting time t_i is zero because no steps are needed to reach A. Similarly, if state i is an absorbing state and not in A, the mean hitting time t_i is zero because once i is reached, it cannot transition to any other state, including the states in A.

13.11 Mean Return Times

Similar to the considerations for mean hitting times, let r_i denote the expected number of steps needed to return to state i for the first time, starting from state i.

The mean return time r_i is given by:

$$r_i = E(R_i \mid X_0 = i) = 1 + \sum_k p_{ik} t_k \tag{204}$$

where R_i is defined as:

$$R_i = \min\{n \geq 1 \mid X_n = i\} \tag{205}$$

This equation states that the expected return time r_i is the sum of one step plus the expected hitting times from the next states, weighted by the probabilities of transitioning to those states. In this context, t_k represents the mean hitting time for state k.

13.12 State Distribution

The state distribution of a Markov chain at a particular time step represents the probabilities of being in each possible state at that time. It's essentially a snapshot of where the Markov chain is likely to be at a specific point in time.

In a Markov chain, we are interested in understanding the long-term behavior of the state probabilities. Specifically, we look at what happens to the state distribution $\pi^{(n)}$ as n approaches infinity.

A limiting distribution, when it exists, is always a stationary distribution, but the converse is not true. There may exist a stationary distribution but no limiting distribution.

13.12.1 Stationary Distribution

The stationary distribution of a Markov chain is a special probability distribution that remains unchanged as the chain evolves over time. Let $\{X_n, n \geq 0\}$ be an irreducible and aperiodic Markov chain. If there exists a probability vector π such that:

$$\pi P = \pi, \tag{206}$$

then π is the stationary distribution of the Markov chain.

1. The sum of the probabilities in the stationary distribution is 1:

$$\sum_{j \in S} \pi_j = 1 \tag{207}$$

2. The stationary distribution π_j is the limiting distribution of the Markov chain:

$$\pi_j = \lim_{n \to \infty} P(X_n = j \mid X_0 = i) \tag{208}$$

3. The mean return time r_j to state j is given by:

$$r_j = \frac{1}{\pi_j}, \quad r_j \text{ is the mean return time to state } j \tag{209}$$

To find a distribution π such that:

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, \quad \text{for all } j \in S, \tag{210}$$

we must set up a system of equations. This means we solve for the probabilities π_j such that the sum of the probabilities is 1:

$$\sum_{j \in S} \pi_j = 1 \tag{211}$$

Assuming $S = \{1, 2, 3, \ldots, r\}$, we solve the system of linear equations to find the stationary distribution.

13.12.2 Finding the Stationary Distribution

To determine the stationary distribution of a Markov chain, we follow a systematic procedure involving linear algebra techniques. The stationary distribution π is a probability vector that remains unchanged as the Markov chain evolves over time. Mathematically, it satisfies the equation $\pi P = \pi$, where P is the transition matrix. The following steps outline the method to find π:

Steps:

1. **Transpose P**:
   - Transpose the transition matrix P to facilitate setting up the system of linear equations. This step helps rewrite the equation in a standard linear algebra form.
2. **Subtract 1 from the Diagonal**:
   - Subtract 1 from each diagonal element of the transposed matrix. This transforms the problem into finding the null space of the matrix $P^T - I$, where I is the identity matrix.
3. **Drop Row n**:
   - Drop one row (say, row n) from the matrix to reduce the system to a consistent set of equations.
4. **Add Column of Zeros**:
   - Add a column of zeros to the matrix. This maintains the structure of the equations and helps in setting up the augmented matrix for solving the linear system.
5. **Add Row of Ones**:
   - Add a row of ones to the matrix. This step enforces the probability constraint that the sum of the probabilities in the stationary distribution equals 1.
6. **Formulate the Augmented Matrix**:
   - Construct the augmented matrix from the modified transition matrix and the additional constraints. The augmented matrix will help in solving the system of linear equations.
7. **Solve the System of Linear Equations**:
   - Use a method such as Gaussian elimination, LU decomposition, or matrix inversion to solve the augmented matrix. The solution will give the stationary distribution π.
8. **Normalize the Solution**:
   - Ensure that the solution vector π is normalized, meaning the sum of its elements equals 1. This may involve dividing each element of π by the sum of all elements.

These steps collectively allow us to solve for the stationary distribution, which provides insight into the long-term behavior of the Markov chain.

13.12.3 Limiting Distribution

The limiting distribution describes the long-term behavior of the Markov chain, showing where the chain will end up regardless of the initial state. If the Markov chain is irreducible and aperiodic, the limiting distribution π will coincide with the stationary distribution π.

$$\pi^{(n)} = \begin{bmatrix} P(X_n = 0) & P(X_n = 1) & \dots \end{bmatrix} \tag{212}$$

As $n \to \infty$, the distribution $\pi^{(n)}$ may converge to a limiting distribution. This is represented as:

$$\pi_j = \lim_{n \to \infty} P(X_n = j \mid X_0 = i) \tag{213}$$

$$\lim_{k \to \infty} \pi^{(0)} P^k = \pi \tag{214}$$

for all states j. The limiting distribution satisfies the condition:

$$\sum_j \pi_j = 1 \tag{215}$$

Note A finite irreducible Markov chain will only consist of recurrent states. In such cases, the Markov chain will exhibit well-defined limiting behavior if it is aperiodic.