Statistical and Mathematical Methods

Statistical and Mathematical Methods for Data Science
DS5003

Dr. Nasir Touheed

Statistics

Uniform Distributions

- a random variable with any thinkable distribution can be generated from a Uniform random variable
- Uniform distribution is used in any situation when a value is picked "at random" from a given interval; that is, without any preference to lower, higher, or medium values.
- Whenever the probability is proportional to the length of the interval, the random variable is uniformly distributed

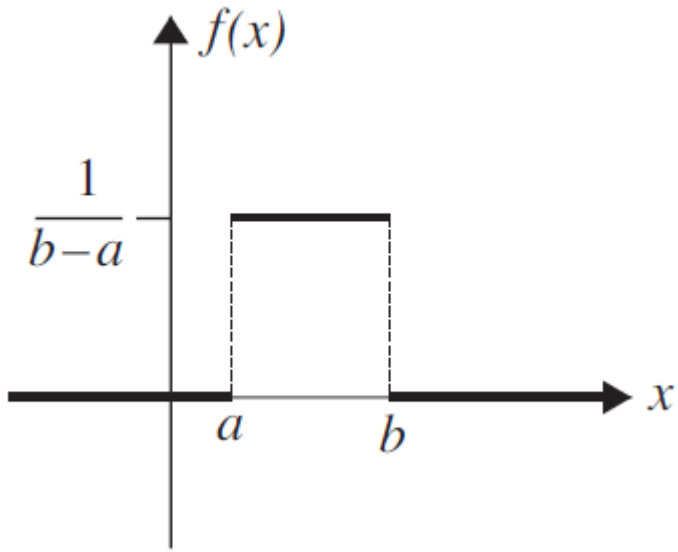
Uniform Distributions

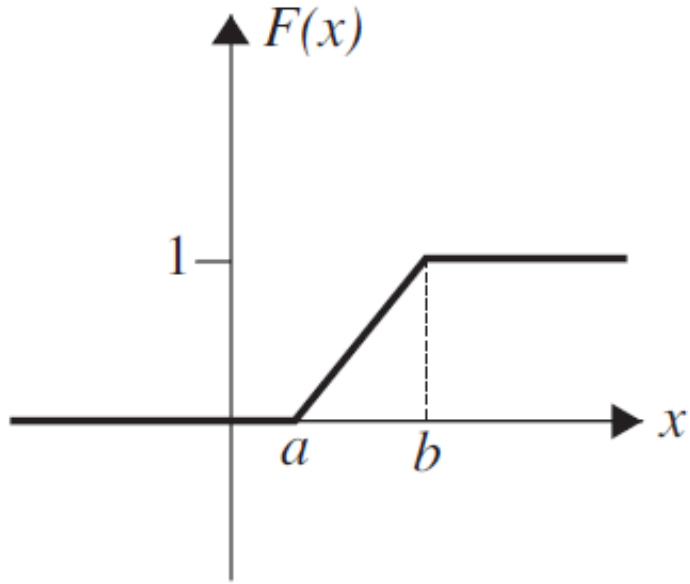
- Suppose that X is the value of the random point selected from an interval (a, b). Then X is called a uniform random variable over (a, b). Let F and f be distribution and probability density functions of X, respectively

$$F(t) = \begin{cases} 0 & t \leq a \\ \dfrac{t-a}{b-a} & a \leq t \leq b \\ 1 & t \geq b \end{cases}$$

$$f(t) = F'(t) = \begin{cases} \dfrac{1}{b-a} & a < t < b \\ 1 & \text{otherwise} \end{cases}$$

Uniform Distributions





Uniform Distributions

- In random selections of a large number of points from (a, b), we expect that the average of the values of the points will be approximately $\frac{a+b}{2}$, the midpoint of (a, b).

$$E(X) = \int_a^b x \, \frac{1}{b-a} dx$$

$$= \frac{1}{b-a} \left[\frac{1}{2} x^2 \right]_a^b$$

$$= \frac{1}{b-a} \left(\frac{1}{2} b^2 - \frac{1}{2} a^2 \right)$$

$$= \frac{(b-a)(b+a)}{2(b-a)}$$

Uniform Distributions

- To find $Var(X)$, we have

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx$$

$$= \frac{1}{3} \frac{b^3 - a^3}{b-a}$$

$$= \frac{1}{3}(a^2 + ab + b^2)$$

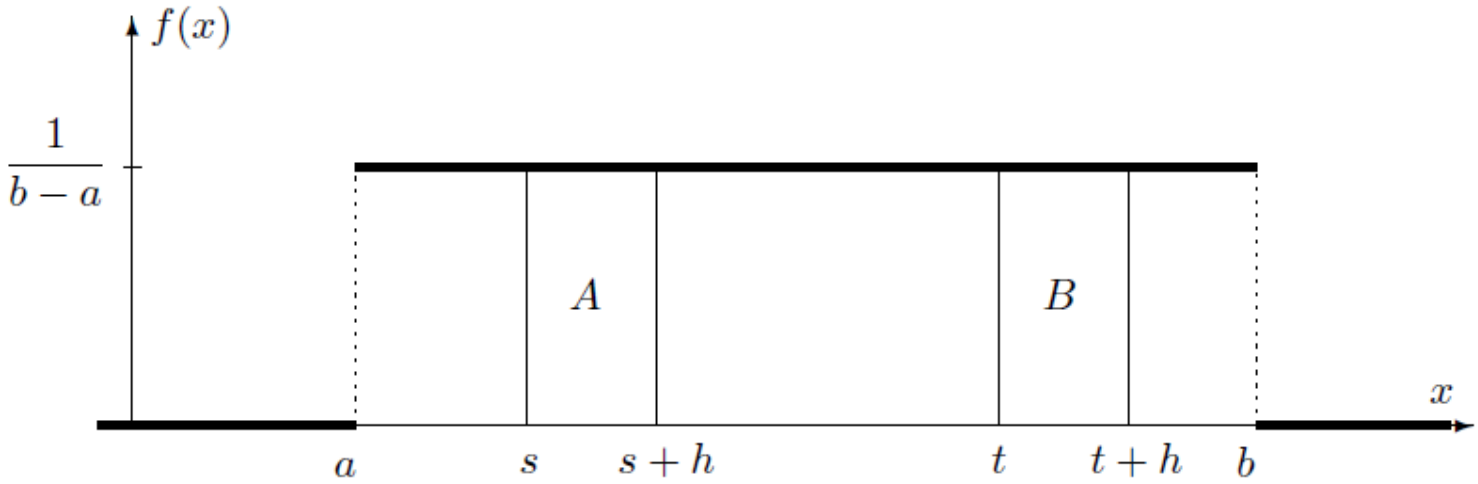
Hence

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

$$= \frac{1}{3}(a^2 + ab + b^2) - \left(\frac{a+b}{2}\right)^2$$

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

Uniform Distributions



$$
\begin{aligned}
(a, b) &= \text{range of values} \\
f(x) &= \frac{1}{b-a}, \quad a < x < b \\
\mathbf{E}(X) &= \frac{a+b}{2} \\
\text{Var}(X) &= \frac{(b-a)^2}{12}
\end{aligned}
$$

Uniform Distributions

- Starting at 5:00 A.M., every half hour there is a flight from Karachi International airport to Islamabad International airport. Suppose that none of these planes is completely sold out and that they always have room for passengers. A person who wants to fly to L.A. arrives at the airport at a random time between 8:45 A.M. and 9:45 A.M. Find the probability that he waits
  a. at most 10 minutes
  b. at least 15 minutes

Uniform Distributions

- Starting at 5:00 A.M., every half hour there is a flight from Karachi International airport to Islamabad International airport. Suppose that none of these planes is completely sold out and that they always have room for passengers. A person who wants to fly to L.A. arrives at the airport at a random time between 8:45 A.M. and 9:45 A.M. Find the probability that he waits
- at most 10 minutes
- The passenger arrive at the airport X minutes past 8:45.
- Then X is a uniform random variable over the interval $(0, 60)$.
- Hence the probability density function of X is given by f

$$f(x = \begin{cases} \frac{1}{60} & 0 < x < 60 \\ 1 & \text{otherwise} \end{cases}$$

Uniform Distributions

- Starting at 5:00 A.M., every half hour there is a flight from Karachi International airport to Islamabad International airport. Suppose that none of these planes is completely sold out and that they always have room for passengers. A person who wants to fly to L.A. arrives at the airport at a random time between 8:45 A.M. and 9:45 A.M. Find the probability that he waits
- at most 10 minutes
- Now the passenger waits at most 10 minutes if she arrives between 8:50 and 9:00 or 9:20 and 9:30; that is, if
  $5 < X < 15$ *or* $35 < X < 45$

$$P(5 < X < 15) + P(35 < X < 45) = \int_5^{15} \frac{1}{60} + \int_{35}^{45} \frac{1}{60}$$

Uniform Distributions

- A person arrives at a bus station every day at 7:00 A.M. If a bus arrives at a random time between 7:00 A.M. and 7:30 A.M., what is the average time spent waiting
- It takes a professor a random time between 20 and 27 minutes to walk from his home to school every day. If he has a class at 9:00 A.M. and he leaves home at 8:37 A.M. , find the probability that he reaches his class on time.
- The time at which a bus arrives at a station is uniform over an interval (a, b) with mean 2:00 P.M. and standard deviation $\sqrt{12}$ minutes. Determine the values of a and b

Normal Distributions

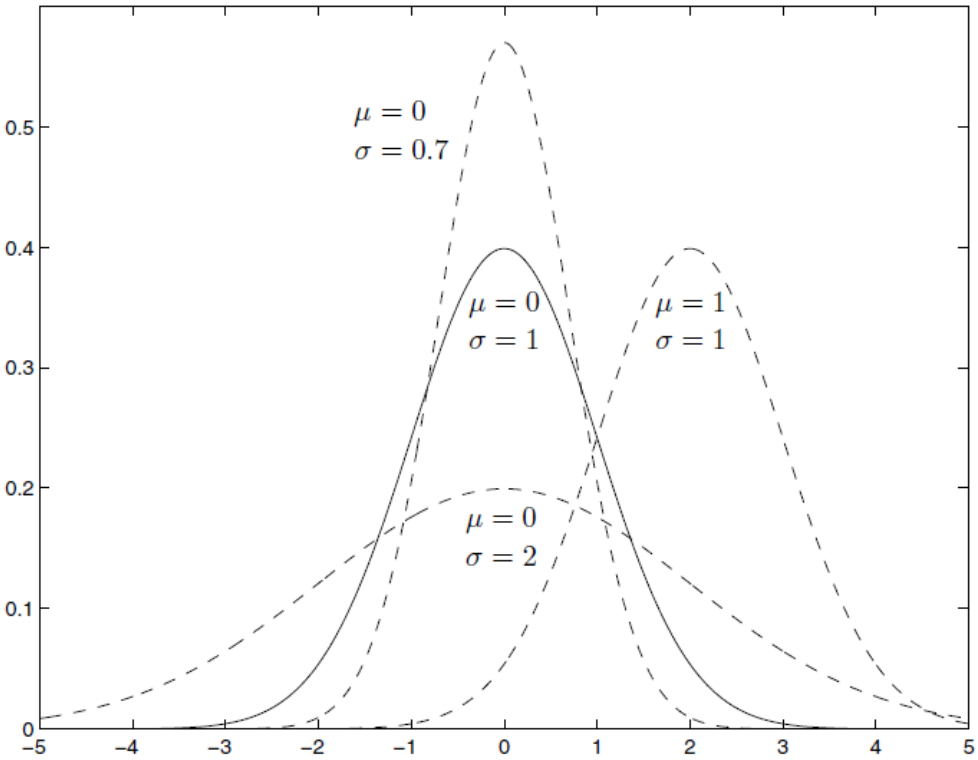
- Normal distribution is often found to be a good model for physical variables like weight, height, temperature, voltage, pollution level, and for instance, household incomes or student grades.
- Normal distribution has a density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

- where parameters μ and σ have a simple meaning of the expectation $E(X)$ and the standard deviation Std(X).
- This density is known as the bell-shaped curve, symmetric and centered at μ, its spread being controlled by σ.

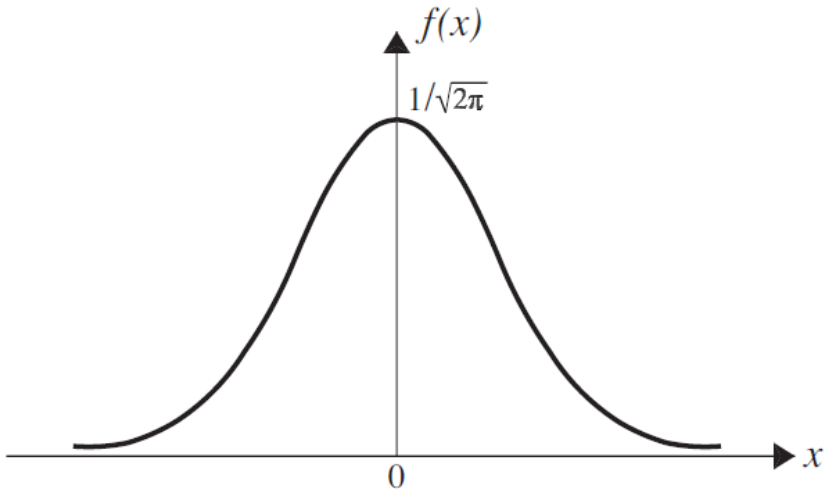
Normal Distributions

- A normal probability distribution, when plotted gives a bell shaped curve such that
  - The total area under the curve is 1
  - The curve is symmetric about mean
  - The two tails of the curve extend indefinitely



Normal Distributions

- Normal distribution with "standard parameters" $\mu = 0$ and $\sigma = 1$ is called Standard Normal distribution.



$$\begin{aligned} Z &= \text{Standard Normal random variable} \\ \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \text{ Standard Normal pdf} \\ \Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \text{ Standard Normal cdf} \end{aligned}$$

Normal Distributions

- Suppose that of all the clouds that are seeded with silver iodide, 58% show splendid growth. If 60 clouds are seeded with silver iodide, what is the probability that exactly 35 show splendid growth

Normal Distributions

- Suppose that of all the clouds that are seeded with silver iodide, 58% show splendid growth. If 60 clouds are seeded with silver iodide, what is the probability that exactly 35 show splendid growth?
- Let X be the number of cloud showed splendid growth
- $E(x) = np = 34.80$
- $\sigma_X = \sqrt{npq} = 3.82$
- $P(X = 35) \approx P(34.5 < X < 35.5)$

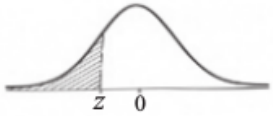
$$\begin{aligned} P(X = 35) \approx & P(34.5 < X < 35.5) \\ = & P\left(\frac{34.5 - 34.8}{3.82} < \frac{X - 34.8}{3.82} < \frac{35.5 - 34.8}{3.82}\right) \end{aligned}$$

Normal Distributions

$$
\begin{aligned}
P(X = 35) \approx & P(34.5 < X < 35.5) \\
= & P\left(\frac{34.5 - 34.8}{3.82} < \frac{X - 34.8}{3.82} < \frac{35.5 - 34.8}{3.82}\right) \\
= & P\left(-0.08 < \frac{X - 34.8}{3.82} < 0.18\right) \\
= & \frac{1}{\sqrt{2\pi}} \int_{-0.08}^{0.18} exp(-\frac{x^2}{2}) dx \\
= & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.18} exp(-\frac{x^2}{2}) dx - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0.08} exp(-\frac{x^2}{2}) dx \\
= & \Phi(0.18) - \Phi(-.08)
\end{aligned}
$$

Table 1 Area under the Standard Normal Distribution to the Left of z: Negative z

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$$



Note that for $z \le -3.90$, $\Phi(z) = P(Z \le z) \approx 0$

z	0	1	2	3	4	5	6	7	8	9
−3.8	.0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
−3.7	.0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
−3.6	.0002	0002	0001	0001	0001	0001	0001	0001	0001	0001
−3.5	0002	0002	0002	0002	0002	0002	0002	0002	0002	0002
−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 2 Area under the Standard Normal Distribution to the Left of z: Positive z

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$$

z	0	1	2	3	4	5	6	7	8	9
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.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	.9889	.9889	.9889	.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
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Note that for $z > 3.89$, $\Phi(z) = P(Z \le z) \approx 1$.

Normal Distributions

- Suppose that the average household income in some country is 900 coins, and the standard deviation is 200 coins. Assuming the Normal distribution of incomes, compute the proportion of "the middle class," whose income is between 600 and 1200 coins.
- The government of the country decides to issue food stamps to the poorest 3% of households. Below what income will families receive food stamps?

LAWS OF LARGE NUMBERS

- Let A be an event of some experiment that can be repeated
- the mathematicians of the eighteenth and nineteenth centuries observed that, in a series of sequential or simultaneous repetitions of the experiment, the proportion of times that A occurs approaches a constant.
- they were motivated to define the probability of A to be the number $p = \lim_{n\to\infty} \frac{n(A)}{n}$, where n(A) is the number of times that A occurs in the first n repetitions.
- this relative frequency interpretation of probability, which to some extent satisfies one's intuition, is mathematically problematic and cannot be the basis of a rigorous probability theory.

LAWS OF LARGE NUMBERS

- for repeatable experiments, the relative frequency interpretation is valid and is the special case of one of the most celebrated theorems of probability and statistics: the strong law of large numbers
- For sufficiently large n, it is very likely that

$$| \frac{n(A)}{n} - P(A) |$$

is very small.

LAWS OF LARGE NUMBERS

(Weak Law of Large Numbers) *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with $\mu = E(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i) < \infty, i = 1, 2, \ldots$. Then $\forall \varepsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

(Strong Law of Large Numbers) *Let $X_1, X_2, \ldots$ be an independent and identically distributed sequence of random variables with $\mu = E(X_i)$, $i = 1, 2, \ldots$. Then*

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right) = 1.$$

LAWS OF LARGE NUMBERS

Example At a large international airport, a currency exchange bank with only one teller is open 24 hours a day, 7 days a week. Suppose that at some time $t = 0$, the bank is free of customers and new customers arrive at random times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, ..., where $T_1, T_2, T_3, \ldots$ are identically distributed and independent random variables with $E(T_i) = 1/\lambda$. When the teller is free, the service time of a customer entering the bank begins upon arrival. Otherwise, the customer joins the queue and waits to be served on a first-come, first-served basis. The customer leaves the bank after being served. The service time of the ith new customer is S_i, where $S_1, S_2, S_3, \ldots$ are identically distributed and independent random variables with $E(S_i) = 1/\mu$. Therefore, new customers arrive at the rate λ, and while the teller is busy, they are served at the rate μ. Show that if $\lambda < \mu$, that is, if new customers are served at a higher rate than they arrive, then with probability 1, eventually, for some period, the bank will be empty of customers again.

Central Limit Theorem

(Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with expected value μ and variance σ^2.*

Then the distribution of

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges to the distribution of a standard normal random variable. That is,

$$\lim_{n\to\infty} P(Z_n \leq x) = \lim_{n\to\infty} P\Big(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\Big)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

Central Limit Theorem

Central Limit Theorem Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with expected value μ and variance σ^2. Then

$$\lim_{n\to\infty} P\Big(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le x\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

Let $\bar{X}_n$ be the mean of the random variables $X_1, X_2, \ldots, X_n$. The central limit theorem is equivalent to

$$\lim_{n\to\infty} P\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le x\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

Central Limit Theorem

The lifetime of a TV tube (in years) is an exponential random variable with mean 10. What is the probability that the average lifetime of a random sample of 36 TV tubes is at least 10.5?

The parameter of the exponential density function of the lifetime of a tube is

$$\lambda = 1/10.$$

For $1 \leq i \leq 36$,

let X_i be the lifetime of the ith TV tube in the sample.

for $1 \leq i \leq 36$,

$$E(X_i) = 1/\lambda = 10 \text{ and } \sigma_{X_i} = 1/\lambda = 10.$$

By central limit theorem

$$P(\bar{X} \geq 10.5) = P\Big(\frac{\bar{X} - 10}{10/\sqrt{36}} \geq \frac{10.5 - 10}{10/\sqrt{36}}\Big) = P\Big(\frac{\bar{X} - 10}{10/\sqrt{36}} \geq 0.30\Big)$$

$$\approx 1 - \Phi(0.30) = 1 - 0.6179 = 0.3821.$$

Central Limit Theorem

Example (ELEVATOR). You wait for an elevator, whose capacity is 2000 pounds. The elevator comes with ten adult passengers. Suppose your own weight is 150 lbs, and you heard that human weights are normally distributed with the mean of 165 lbs and the standard deviation of 20 lbs. Would you board this elevator or wait for the next one?

Solution. In other words, is overload likely? The probability of an overload equals

$$\begin{aligned} \boldsymbol{P}\{S_{10}+150>2000\} &= \boldsymbol{P}\left\{\frac{S_{10}-(10)(165)}{20\sqrt{10}} > \frac{2000-150-(10)(165)}{20\sqrt{10}}\right\} \\ &= 1-\Phi(3.16)=0.0008. \end{aligned}$$