THEORETICAL DISTIBUTIONS: DISCRETE AND CONTINUOUS

Ankur Bhatnagar[1]

Contents

1. LEARNING OBJECTIVES

We discuss 4 distributions here.

- Discrete distributions: Binomial, Poisson
- Continuous: Uniform, Normal

For each we need to be familiar with

- Probability mass function(for discrete distributions), probability distribution function(for continuous distributions)
- Cumulative distribution function
- Conditions for a distribution to hold
- Approximations (as applicable)
- Mean ,variance and standard deviation

[1] Associate Professor, Satyawati College, University of Delhi.

2. INTRODUCTION

What are theoretical distributions?

Theoretical distribution is based on mathematical formulae. They are derived from model or estimated from data, rather than conducting experiments physically or making a sample space. If certain conditions are fulfilled we can say that a variable follows a particular distribution.

2.1 DISCRETE DISTRIBUTIONS

2.1.1 BINOMIAL DISTRIBUTION

For any variable x to follow a binomial distribution the following conditions must be met.

1. There are n fixed and identical trials.
2. Each trial is independent of other trials, so that outcome of one trial does not affect the outcome of any other trial.
3. Each trial has **ONLY** two possible outcomes S (Success) and F (Failure).
4. P(Success) =P(S) is denoted as p and is constant in each trial. P(Failure) = q = 1 − p.

The pmf is given as the probability of r successes in n trials.

$P(X=r) = {}^{n}C_{r} * p^{r} q^{n-r} = n!/(r!(n-r)!)\ p^{r} q^{n-r}$

Mean=E(x) = n*p

Variance= V(x) = n*p*q

> TIP: Success and Failure are labelled in an arbitrary fashion. We can label any of the two events in the sample space as success. For example a girl/boy child can be labelled success or failure, without affecting the answer. However we need to be careful with the value of r.

Q Assume that a die is tossed 5 times. What is the probability of getting exactly 2 fours?

We use this example to show that any of the two events in a binomial experiment can be labelled as success.

Option 1:

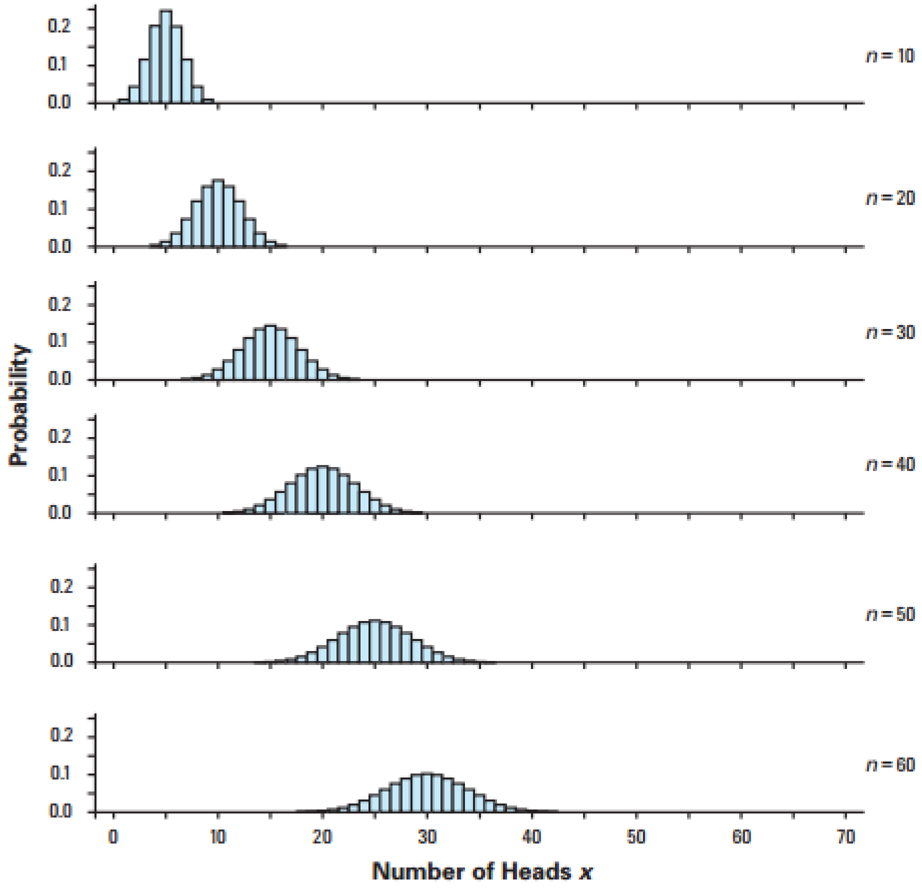
Define getting a four to be a success so that p=1/6. We want P(r=2). Use n= 5,p= 1/6 and r= 2 to get b(2; 5, 0.167) = ${}_{5}C_{2} * (0.167)^{2} * (0.833)^{3} = 0.161$.

Option 2:

Now define any number except four to be a success so that p=5/6. We want 2 fours so that the number of success = 5-2=3 We now want P(r=3). Use n= 5, p= 5/6 and r= 3 to get

P(r=3) = ${}_{5}C_{3} * (0.833)^{3} * (0.161)^{2} = 0.161$.

We have illustrated that ‘labelling’ of success and failure have no impact on the answer as long as the number of success is chosen properly and correctly.



2.1.2 POISSON DISTRIBUTION

The Poisson distribution is a discrete probability distribution for the number of events that occur randomly in a given interval of time/period. This is unlike a binomial, hypergeometric or negative binomial distributions that are based on an experiment that uses trials/ draws to get probability of various outcomes.

Let X = the number of events in a given interval

λ = mean number of events per interval

The probability of observing r events in a given interval is given by the pdf.

$P(X=r) = e^{-\lambda}\lambda^r/r!$ Where r takes values 0,1,2,3,4.... and e=2.718282

Mean= variance= λ

NOTE: the rate/number of events is always in terms of a specified interval like per hour or minutes or days.

Q Historical data shows that there are 1.8 births per hour in a village. What is the probability that 4 babies will be born in any given hour here?

Let X= number of births, we need P(X=4) = $e^{-1.8}\ 1.8^4/4! = 0.0723$

What is the probability of having 2 or more births in an hour.

We need P(X ≥2). Since the value of number of births is infinite. We get probability as

$P(X \geq 2) = 1 - P(X=0) - P(X=1) = 1 - e^{-1.8} 1.8^0/0! - e^{-1.8}\ 1.8^1/1! = 0.537$

Q Let *X* equal the number of typos on a printed page with a mean of 3 typos per page.

a. What is the probability that a randomly selected page has **at least one typo** on it? We can find the requested probability directly from the pdf. The probability that *X* is at least one is: $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-3}\ 3^0/3! = 1 - 0.0498 = 0.9502$ That is, there is just over a 95% chance of finding at least one typo on a randomly selected page when the average number of typos per page is 3.

b. What is the probability that a randomly selected page has **at most one typo** on it?

The probability required is $P(X \leq 1) = P(X = 0) + P(X = 1)$

$= P(X \leq 1) = 1 - e^{-3}\ 3^0/0! + e^{-3}\ 3^1/1! = 0.1992.$

That is, there is just under a 20% chance of finding at most one typo on a randomly selected page when the average number of typos per page is 3.

Q. An office receives 20 faxed orders every two hours.

This is a Poisson distribution with average no of faxes= 10 faxes per hour.

a. What is the probability that it will receive 8 orders in the next hour?

P(x =8) = .1125

b. What is the probability that an order will be faxed within the next 9 minutes?

NOTE that the rate was in terms of hours whereas this question talks in minutes. So we convert minutes to hours. 9 minutes= 9/60 hours

We need P(x < 9/60) which has an exponential distribution.

$= 1 - e^{(-(9/60)*10)} = 1 - e^{(-3/2)} = .77687$

c. What is the probability that more than 12 minutes will elapse between faxed orders?

12 minutes= 12/60 hours= .2 hours

We need P(x > .2) which has an exponential distribution.

= $e^{(-(.2)*10)}$ = $e^{(-2)}$ = .135335

2.1.3 BINOMIAL APPROXIMATION TO POISSON

A binomial distribution approximates a Poisson distribution when n approaches infinity and p approaches 0. A simple thumb rule is that n >50 and np< 5. Let us see the approximation below:

Q The publisher of a medical journal claims that probability of an error is .005. the errors on each page are independent of each other. If a journal has 400 pages,

a. what is the probability that only 1 page has an error?

N=400 p= .005 so that np= 2.

Using a Poisson distribution, P(X=1) = $e^{-2}2^1/1!$ = .270671

If we use a binomial then, the answer is .270669. The answers are close to each other, proving that approximation holds.

b. what is the probability that at most 2 pages have an error?

P(X=0)+P(X=1) +P(X=3) = $e^{-2}2^0/0! + e^{-2}2^1/1! + e^{-2}2^2/2!$

= .135226+.270671+.270671= .676653

2.2 CONTINUOUS DISTRIBUTIONS

2.2.1 NORMAL DISTRIBUTION:

The probability distribution function of a normal variable x, is given as:

f(x) = { 1/[σ √2π] } *$e^{-(x - \mu)^2/2\sigma^2}$

μ is the mean, σ is the standard deviation, π is approximately 3.14159, and e is approximately 2.71828. Each normal distribution with its own values of μ and σ would need its own calculation of the area under various points on the curve.

Using this function of this distribution is very cumbersome as it involves an exponential term.

We transform the x variable to another variable, named ‘z’ which is easier to use.

The transformation involves changing the scale and origin of x as follows:

Z=(x- mean of x)/ sdv of x= (x - μ) / σ

Z is also called the **standard normal variable**.

The mean of z= 0 and sd =1. Let us see how.(sd stands for standard deviation)

Note E(z)= E((x- mean of x)/ sd of x).

Using the rules of expectations,

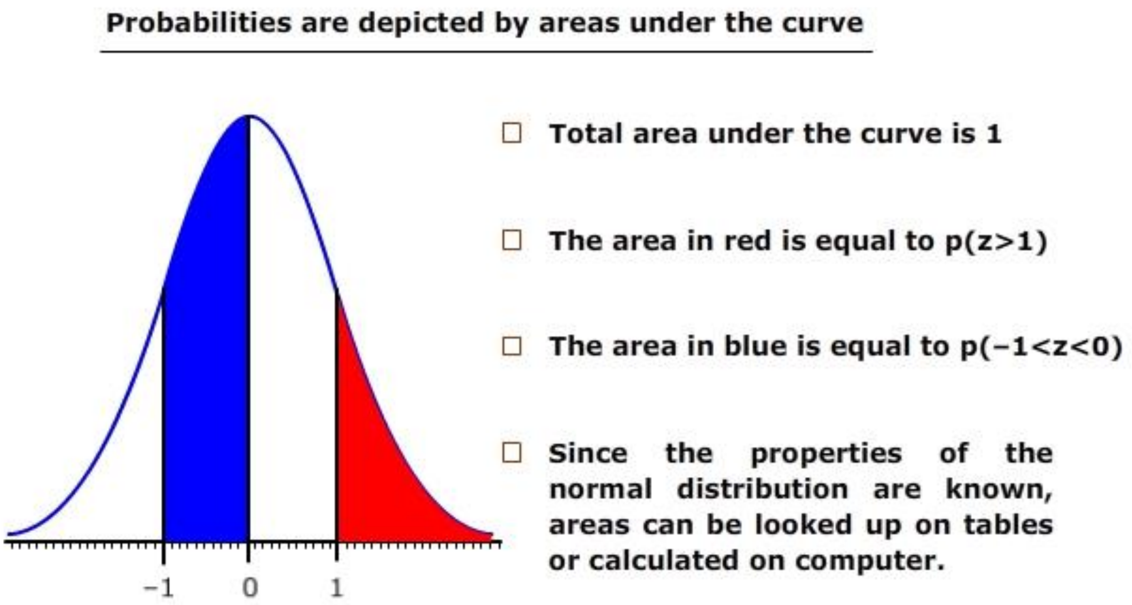
E(z)= (E(x) –E(x))/sd of x= 0 (sdv of x and E(X) are constants)

V(z) = V((x - μ) / σ) = V(x) / σ^2= σ^2 /σ^2 = 1 ($sd^2(x) = \sigma^2$=V(x))

The z distribution is also bell shaped. The area under this curve provides the cumulative probability of a given z value. Since we are dealing with a continuous variable, the probability of z taking a single value is zero. The area under the curve and between a given value of z gives us the probability that the variable X (that corresponds to z chosen)takes a value less than specified value

In short calculating probability for a normal distribution involves the following steps:

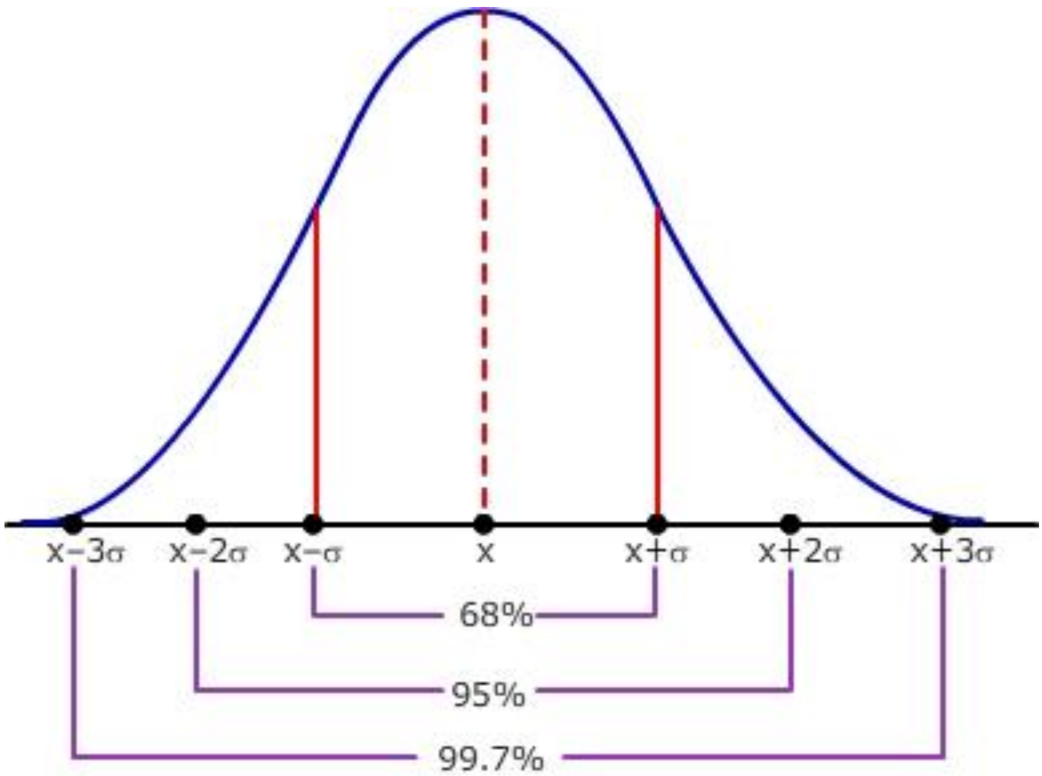
- Choose a value of X (say x1) for which we need P(X < x1)
- Find the value of z that corresponds to x1, and call it z1= (x1=mean)/sd
- From the tables find the row that corresponds to z1 and get the associated value given in table. This value is the probability of z < z1 or P(x <x1) .
- Now suppose we want P(X>x1) . we find this as 1-P(z<z1)
- Next we want P(x1<x<x2). We get z1= (x1-mean)/sd and z2=(x2-mean)/sd
- P(x1<X<x2)=P(z1<Z<z2) =P(z< z2) –P(z < z1)
- = table entry that corresponds to z2 – value that corresponds to z1
  This is shown in the following diagram:



EMPIRICAL RULE:

The *Empirical Rule* is based on the above concept of z distribution. It states that if a data set is normally distributed with population mean μ and standard deviation σ, then the following are true:

- About 68% of the values lie within 1 standard deviation of the mean In statistical notation, this is represented as $\mu \pm \sigma$
- About 95% of the values lie within 2 standard deviations of the mean .The statistical notation for this is $\mu \pm 2\sigma$
- About 99.7% of the values lie within 3 standard deviations of the mean or between - $\mu \pm 3\sigma$.



Consider an example:

Q The Bulb Co, Ltd finds that its average CFL lasts 1000 hours with a standard deviation of 100 hours. Assume that CFL life is normally distributed.

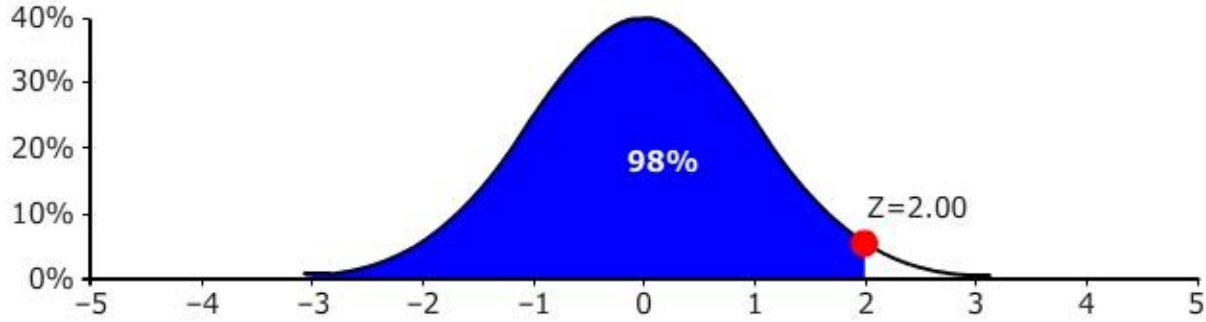
a. What is the probability that a randomly selected CFL will burn out in 1200 hours or less?

Let x be the life of CFL in hours:

E(x) = 1000. standard deviation(x)=100.We want P(x <1200)

Let us transform x to z: Z=(1200-1000)/100= 2

So P(x <1200)=P(z <2). Looking at normal tables we find the answer to be 0.977. Thus, there is a 97.7% probability that a CFL will burn out within 1200 hours. The diagram below shows the probability as rounded off figure. The blue area is the required probability.



b. What is the probability that a randomly selected CFL will last more than 1200 hours?

We want P(x >1200).Let us transform x to z:

Z=(1200-1000)/100= 2

So P(x >1200)=P(z >2). Looking at normal tables we find that P(z<2) =0.977.

Since area under the curve is 1 we find P(z>2) = 1-.977= 0.023. Thus, there is a 2.3% probability that a CFL will last more than 1200 hours.

c. What is the probability that a randomly selected CFL will last between 1100 and 1200 hours?

P(1100<x<1200) = P(1< z <2)

When x= 1100, z =(1100-1200/100)= 1
Since the table gives us cumulative probabilities, we find that P(z <1100) = 0.841 and P(z< 1200)= .977.
So P(1<z<2) = .977-.841 = 0.136.

Q The mean of a normal probability distribution is 60; the standard deviation is 5.

a. About what percent of the observations lie between 55 and 65?
P(55<x <65) = p[(55-60)/5 < z <(65-60)/5] = P(-1<z <1) = .6826 or 68.26%

b. About what percent of the observations lie between 50 and 70?
P(50 < x < 60) = p[(50-60)/5 < z <(60-60)/5] = P(-2<z <2) = .9545 or 95.45%

c. About what percent of the observations lie between 45 and 75?
P(45<x<75) = p[(45-60)/5 < z <(75-60)/5] = P(-3<z <3) = .9973 or 99.73%

Q The GMAT is used to enter USA for education. Assume that scores are based on a normal distribution with a mean of 1500 and a standard deviation of 300. New York College would like to offer a scholarship to students who score in the top 10 percent of this test. What is the minimum score that qualifies for the scholarship?

We want x1 such that $P(x > x1) = 0.1$

$P(z > z1) = .1$

From the tables we can $z1 = 1.282$

$1.282 = (x1-1500)/300$

$X1 = 1884.6$ is the minimum score that qualifies for the scholarship

Q Chemical Company claims that its chemical X contains on the average 4.0 fluid ml of caustic materials per liter. It further states that the distribution of caustic materials per liter is normal and has a standard deviation of 1.3 fluid ml. What proportion of the individual liter containers for this product will contain more than 5.0 fluid ml of X?

$P(x > 5) = P(z > (5-4)/1.3) = P(Z > .769) = 0.220947$ or 22.0947% of the individual liter containers for this product will contain more than 5.0 fluid ml of X.

Q The Federal Government is stepping up efforts to reduce average response times of fire departments to fire calls. The distribution of mean response times to fire calls follows a normal distribution with a mean of 12.8 minutes and a standard deviation of 3.7 minutes.

a. Find the probability that a randomly selected response time is less than 15 minutes.

$P(x < 15) = P(z < (15-12.8)/3.7) = P(z < 0.5946) = 0.72394$

b. Find the probability that a randomly selected response time is between 13 minutes and 15 minutes.

$P(13 < x < 15) = P((13-12.8)/3.7 < z < (15-12.8)/3.7)) = P(-0.054 < z < 0.5946)$

0.42934

c. The fastest 20% of fire departments will be singled out for a special safety award. How fast must a fire department be in order to qualify for the special safety award?

We need z1 such that $P(z > z1) = .2$

$Z1 = 0.842$

$X1 = .842*3.7 + 12.8 = 15.9154$

TIP: we need to understand the relation between x , z and area under normal curve. Given any one of them we must be able to find the other two.

X ⟶ Z ⟶ AREA

The z table can be understood like this. The first column gives the value of z, while the other columns contain the area to the left of a given z value. From the values shown below:

i. P(z < .4) = .6554
ii. P(z <.46)= .6772
iii. P(z > .4) = 1-P(z <.4) = 1-.6554= .3446
iv. P(.4< z< 1.1) = .8643 -.6554= .2089

Standard Normal Probabilities

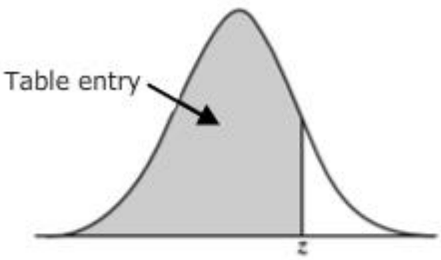


Table entry for *z* is the area under the standard normal curve to the left of *z*.

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

2.2.2 STANDARD NORMAL VARIATE (z) IN TERMS OF PERCENTILES

We can interpret z_α as the $100(1-\alpha)^{th}$ percentile of the standard normal variable. In this sense, choosing α to be . 05 gives us the 95^{th} percentile corresponds to a value of z (say z1) such that P(z < z1) = .95. From the tables this value of z is 1.645.

So the area to the left of 1.645 is .95(or 95%) while 5% area lies to the right of 1.645.

Take some examples:

α	percentile	z_α value	area to left of z_α	area to right of z_α
0.25	75	0.67449	75%	25%
0.45	55	0.125661	55%	45%
0.36	64	0.358459	64%	36%
0.85	15	-1.03643	15%	85%
0.9	10	-1.28155	10%	90%

Note that due to symmetry of normal bell shaped curve, $z_{\alpha} = - z_{\alpha}$.

Note that for some special and more frequently used values of α, we refer to them as 'critical' value of z. These are provided below:

α	percentile	z_α value	area to left of z_α	area to right of z_α
0.1	90	1.281552	90%	10%
0.05	95	1.644854	95%	5%
0.025	97.5	1.959964	97.50%	2.50%
0.01	99	2.326348	99.00%	1.00%
0.005	99.5	2.575829	99.50%	0.50%
0.001	99.9	3.090232	99.90%	0.10%
0.0005	99.95	3.290527	99.95%	0.05%

We can now move to calculate percentiles for non standard distributions (those that do not have 0 mean and 1 sd). If a data set X, is normally distributed with population mean μ and standard deviation σ, then

$100(1-\alpha)^{th}$ percentile for X is given as μ +($100(1-\alpha)^{th}$ percentile for z distribution) *σ.

Take an example:

Q. The CAT exam is used to enter prestigious IIMs in India. Assume that scores are based on a normal distribution with a mean of 1500 and a standard deviation of 300. IIM Bangalore offers an interview to those who are in the top 3%, while IIM Guwahati offers an interview to top 8%. How much do you need to score to guarantee an interview in both places?

Let scores obtained in CAT be denoted by X.

For Bangalore: We want a minimum 97^{th} percentile here to qualify

Using α = .03 we get z_α as 1.88 according to the standard normal distribution (z distribution).

Converting this to X distribution implies that 97^{th} percentile for X is 1500 +1.88*300= 2064. So we need a score of 2064 to get into IIM Bangalore.

For Guwahati: We want a minimum 92^{th} percentile here to qualify

Using α = .08 we get z_α as 1.405 according to the standard normal distribution (z distribution).

Converting this to X distribution implies that 92^{th} percentile for X is 1500 +1.405*300= 1921.5. So we need a score of 1921.5 to get into IIM Guwahati.

2.2.3 BINOMIAL APPROXIMATION TO NORMAL

When the mean of a binomial distribution exceeds 5 we can approximate a binomial distribution with a normal distribution where μ =n*p and σ =√npq.

Let X be a binomial random variable, based on n trails and p as probability of success.

Assuming that np>5,

P(X≤ x)= area under normal curve that lies between x-.5 and x+.5. consider an example:

Q Assume a binomial probability distribution with n = 40 and = 0.55.

a. The mean and standard deviation of the random variable.

Mean = np= 40*.55 =22

Std dev = 3.1464.

Since np >5 we can approximate binomial probabilities with normal probabilities.

b. The probability that X is 25 or greater

P(X > 25) = P(x > ((25-22)/3.146) =.P(x > .95) = 0.17106

c. The probability that X is 15 or less.

P(X < 15) = P(x < ((15-22)/3.146) =.P(x< -2.22) =0.01321

d. The probability that X is between 15 and 25, inclusive

P(15< X < 25) = P(-2.22< z < .95) = 0.81519

Q. The probability of recovery from a rare TB virus in known to be 0.4. If 100 people contract this virus, what is probability that less than 30 will survive?

Let the no of patients that survive be X. We want P(X<30). We can use binomial distribution with n=100 and p= 0.4 and r=30. We can also approximate this toa normal distribution where mean=100*.4=4 and variance= 100*.4*.6=24

We now want P(29.5<X<30.5) = P(z1< z<z2)

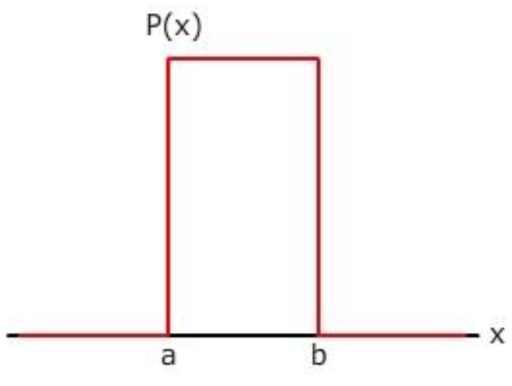
z1= (29.5-40)/√24 = -2.14

z2 = (30.5-40)/√24 = 2.14

P(-2.14<z<2.14) = 0.0162.

2.2.4 UNIFORM DISTRIBUTION

A uniform distribution, (or a rectangular distribution), is a distribution that has constant probability.



Pdf: f(x) = 0 for x<a

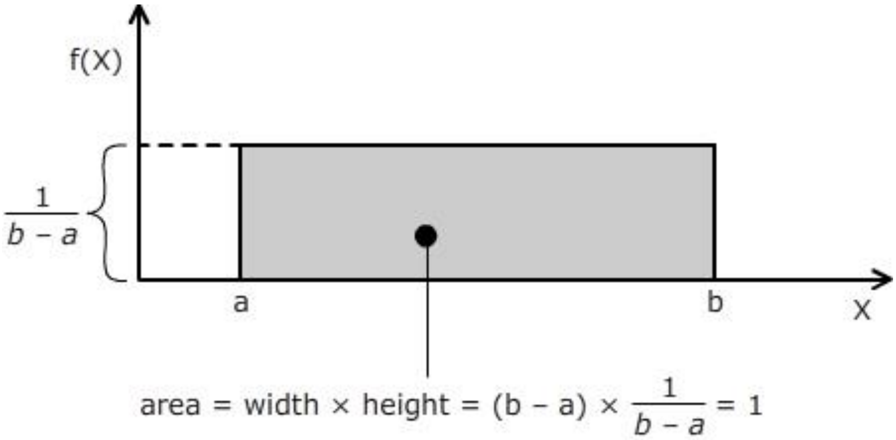
= 1/(b-a) for a<x<b

= 0 for x >b

cdf: F(x) = 0 for x<a

= (x-a)/(b-a) for a<x<b

= 0 for x >b



Mean= E(x) = (a+b)/2

Variance=V(x) = (b-a)2/12 .

Take an example:

Q Assume that the amount of petrol sold every day at a petrol pump is uniformly distributed. The least amount sold is 1000 litres and maximum sold is 3000 litres.

a. What is probability of selling more than 2500 litres on a day?

= (3000-2500)/(3000-1000)= 5000/2000=0.25 or 25%

b. What is probability of selling a maximum of 1200 litres on a day?

= (1200-1000)/(3000-1000)= 200/2000=0.or 10%.

c. What is the probability that the pump will sell between 1500 and 2500 litres?

=(2500-1500)/(3000-1000) = 1000/2000= 0.5 or 50%

d. What is the average amount sold?

= (3000+1000)/2= 2000 litres.

Q: According to the Insurance Institute, a family of four spends between Rs 400 and Rs 3,800 per year on its insurance. Assume that this money spent is uniformly distributed

What is the probability a family spends more than 3,000 per year?

P(X > 3000) = P(3000< x < 3800) = (3800-3000)/(3800 -400) = .25 or 25%

Q. The random variable X is continuous and uniform in [-1, 1] Answer the following questions:

a. What is probability density function of x?

f(x)=1/(2 -1)= 1 for -1≤x≤<1

b. Consider the variable Y such that $Y = 2X^2 - X$. Determine the sample space of Y; when X= 1, Y = ($2*1^2$-1) = 1 and when X= -1 then Y=(2*-1*-1 –(-1))= 3. So Y ranges from 1 to 3

X	Y
-1	3
-0.5	1
0	0
0.5	0
1	1

c. compute the mean and variance of X

Mean (1-(-1))/2= 1

Variance = ($1-(-1)^2$)/12= 1/6

Q: The closing price of Sport Goods Ltd is uniformly distributed between Rs15 and Rs33 per share.

What is the probability that the stock price will be:

a. More than Rs 28? =.27778

b. Less than or equal to Rs20? =.27778

3. USEFUL LINKS

http://mathworld.wolfram.com

http://www.stat.purdue.edu/~zhanghao/STAT511/handout/Stt511%20Sec3.5.pdf

http://www.stattrek.com

4. EXERCISES

Here is list of old questions with the answers provided.

Q1. Ina normal distribution 31% of the observations are under 45 and 8% are above 64. What is the mean and variance of X.

We are given P(X >64) =.08 and P(X <45) = .31

P(X >64) = P(z > z1) =0.08. from the normal tables z1= +1.405

z1= (64-mean)/ sd = 1.405 eq1

P(X <45) = P(z <z2) =0.31. from the normal tables z2= -.496

z2= (45-mean)/ sd = -0.496 eq2

Solve eq1 and 2 together to get mean = 49.95739 and sd = 9.99474 (sd is standard deviation)

Q2. The average time taken to finish a project by L&T is 11 months with sd deviation=2.4 months. If the firm has 19 projects in the pipeline how many can be expected to be completed in less than 1 year?

P(X <12) = .6628

E(X)=12.59 so 12 projects.

Q3. The time for waiting for playing ground in a local tennis club ranges uniformly between 23.5 to 40.5 minutes. If the probability that Harsh has to wait for more than 30 minutes is 60%, he will rather play badminton. Should game will he choose?

From uniform distribution, P(X>30) = (40.5-30)/(40.5-23.5) = 10.5/17= 0.617 or 61.7%. Since we get a value >60% he must choose badminton.

Q4. JK tyres claims an average life of 45000 km for its tyres with standard deviation of 2000kms. Bharat buys 4 tyres for his old car. What is the chance that all 4 tyres will last at least 46000 kms, assuming life of each tyre as independent of all other tyres in the car?

Probability of 1 tyre lasting more than 46000 kms is P(X≥46000) = P(z ≥.5)= .3085

P (all 4 tyres have life of at least 46000kms)=$.3085^4$ =0.0091.

Q5. Let X be normally distributed with mean=30 and variance=49. Find C such that P((X -30) <C) =.9545.

P ((X -30) <C)=P(z< (C/7))=.9545

From the normal tables P (z < 1.69) =.9545

So C/7 = 1.69

C= 7*1.69= 11.83