



JIGSAW ACADEMY
Analytics for Professionals

BASIC STATISTICAL CONCEPTS



INTRODUCTION

- What does Statistics cover?
- Population vs Sample
- Population vs Sample Statistics
- Probability Theory
- Probability Distribution Concepts
- Types of Distributions
- Tests of significance
- ANOVA



STATISTICS – WHAT IS IT?

Dictionary definition of statistics –

“the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.”

There are two parts to the definition

- the collection, classification, analysis and interpretation of numeric data
- AND
- the use of probability theory to impose order on aggregates of data

We will look at each of these elements in the following slides



STATISTICS - INTRODUCTION

In general, statistics deals with summarizing information about data in a meaningful and relevant way.

For example, let's look at the population of Bangalore.

If you are asked to describe the population of Bangalore, how would you do it?

- Population is nothing but the number of people in Bangalore.
 - You would start by saying the population in 2010 is 5.4 million
 - That is a statistic about Bangalore's population – the total (sum) of all full-time residents of Bangalore
 - What other statistics can you think of?
 - “Population density” “Median Age” “Distribution by Religion”
“Literacy Rate”
 - All of these are statistics used to summarize information, because talking about each data point is impossible



STATISTICS - INTRODUCTION

What are some commonly used statistics to summarize information?

- **Sum:** Total of all values in dataset
- **Mean :** The average of all values in the dataset
- **Median :** Mid value of sorted data
 - If even series?
- **Mode** – Most commonly occurring value in a series
- **Minimum** – Lowest value in Series
- **Maximum** – Highest value in Series

Data Series :

17,4,33,2,51,23,3,41,18,2,4,2

Mean = 16.67

$(17+4+33+2+51+23+3+41+18+2+4+2)/12$

Median = 10.5 $(4+17)/2$ – Why?

Mode = 2 – Why?

Minimum = 2

Maximum = 51



STATISTICS - INTRODUCTION

We can describe the series we looked at in the previous example as:

“Minimum of 2, Maximum of 51, average of 16.6.”

We were able to describe the series with 3 points, instead of all 12

All of the statistics we covered talk about the average value or central tendency

We also use statistics to discuss range or dispersion

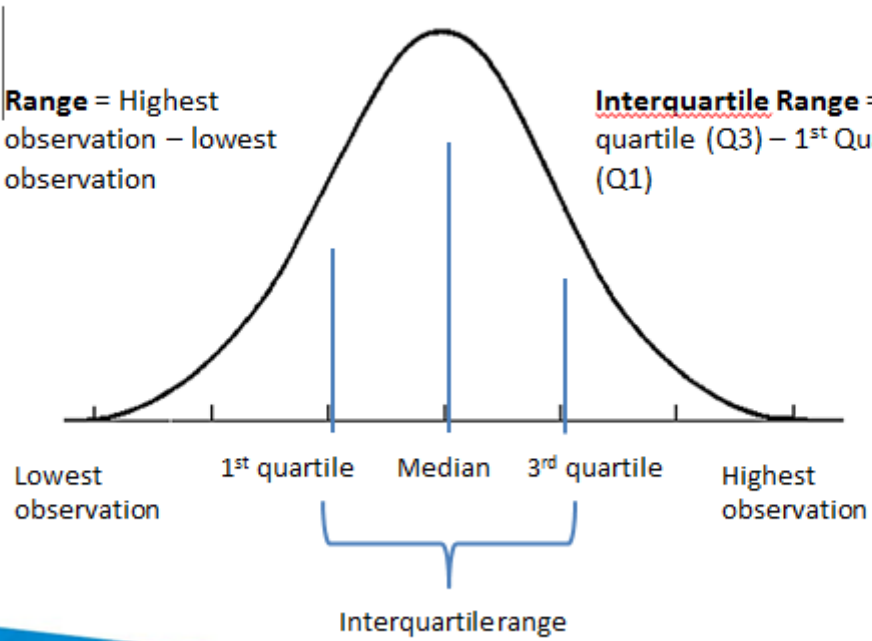
- For example: Range, Variance, Standard Deviation



RANGE

Range = Highest observation – lowest observation

Interquartile Range = 3rd quartile (Q3) – 1st Quartile (Q1)





VARIANCE

Customer Number	Average minute usage (monthly)	$x - \mu$	$(x - \mu)^2$
1	228	-6.8	46.24
2	260	25.2	635.04
3	252	17.2	295.84
4	298	63.2	3994.24
5	234	-0.8	0.64
6	50	-184.8	34151.04
7	264	29.2	852.64
8	230	-4.8	23.04
9	304	69.2	4788.64
10	228	-6.8	46.24

$$\text{Variance} = \sigma^2 = \sum(x - \mu)^2 / N$$

$$\text{Standard Deviation} = \sigma = \sqrt{\sigma^2}$$

x = observation

μ = population mean

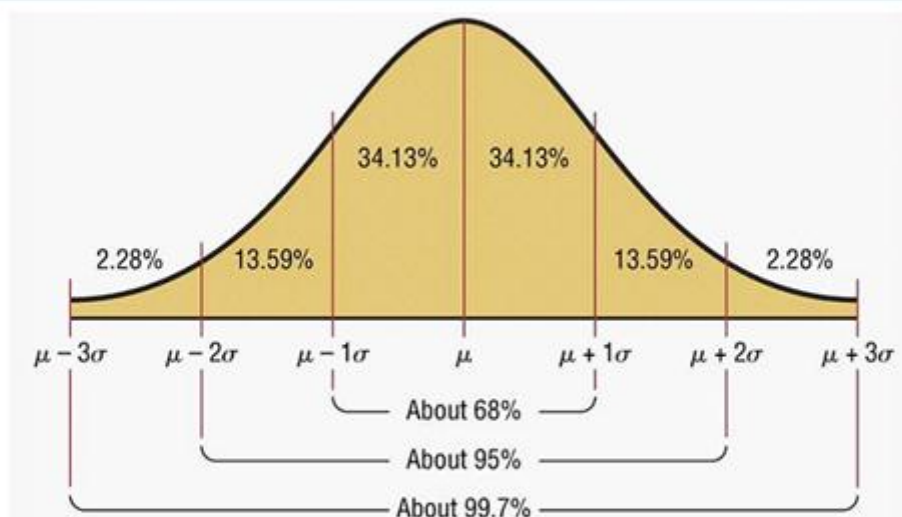
N = number of observations in the population

$$\text{Variance} = (44833.6/10) = \mathbf{4483}$$

$$\text{Standard deviation} = \sqrt{4483} = \mathbf{67}$$



USES OF STANDARD DEVIATION





SHAPE MEASURES

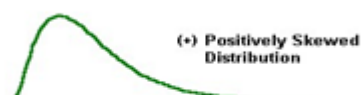
We have reviewed average measures and range measures. We can also look at “shape” measures to describe how the data changes in a particular set

For example, a simple concept is symmetry

- How does data fall on either side of the set mean? Is it symmetric or skewed?
- In the previous slide, when we showed the standard deviation example using a bell curve, was the data in the set symmetric?
- Is it possible to have symmetric distributions with more than one peak?

Skewness: Degree of Asymmetry

- Positive Skew; Long tail to the Right
- Negative Skew: Long tail to the Left

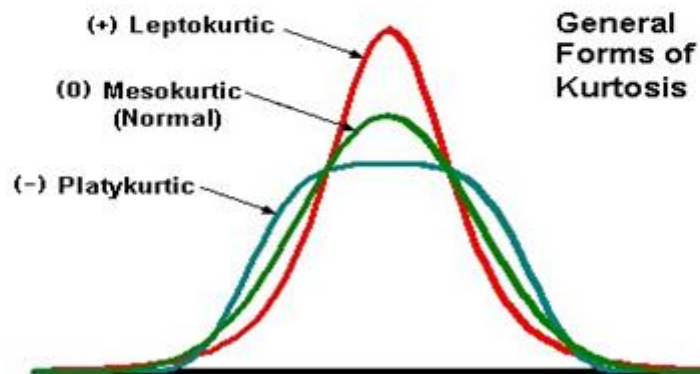




SHAPE MEASURES

Another measure of shape is Kurtosis, or the sharpness of the peak of the distribution

The following figure illustrates the various types of kurtosis:





SUMMARY STATISTICS

Why is it important for us to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

- A very simple example
 - You work for a credit card company, and you have data on credit card applications. You divide the data into applications from customers who have a great payment record, and applications from customers who have been late with payments at least 3 times in the last year
 - Average salary for Group 1: \$37,000. Standard Deviation: \$5,000
 - Average Salary for Group 2: \$26,000. Standard Deviation: \$9,000



STATISTICS - TYPES

Two types of statistics

- **Descriptive**
 - Provides a visual, tabular or numeric summary of large amounts of data to explain its key characteristics
 - Identify patterns in large amounts of data
 - Data Set is assumed as stand-alone
- **Inferential**
 - Uses a sample of data to make inferences about the general population
 - Assumes that sample is representative of a larger population
 - Draws conclusions about population based on the smaller sample



SAMPLE VS POPULATION

In the previous slide, we have introduced the concept of sample and population

Examples of Populations

- All applications received for credit cards from Bank XYZ
- All consumers of Product Y
- What others can you think of?

Examples of Samples:

- All applications received in the last 3 months
- Women consumers over the age of 45 that have bought Product Y in the last 6 months
- What else?



SAMPLE VS POPULATION

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyze data on the population
- It is easier to take a subset of the population, analyze the subset, and then make inferences about the population

The second point depends on a fundamental assumption – what is that?

Representativeness

- We have to find a sample that is representative of the population that it belongs to



SAMPLE VS POPULATION

Let's say we have a population of 10000 respondents to a survey, and we want to take a sample of 500. How many samples are possible?

Clearly, we can choose many samples from a population. A good sample is that which is chosen:

- Without bias :(not choosing only high income respondents)
- Full coverage: All segments in population are correctly represented
- Nonresponse inclusive: If 20% of your population are defaulters, your sample ideally should also have 20% defaulters



HOW TO CHOOSE A SAMPLE

So how do we select a sample?

– **Simple random sample :**

- All members in a population have an equal chance of being selected in the sample

– **Stratified random sampling:**

- Population members are first divided into groups or strata based on meaningfulness. Then random samples are taken from each strata

The important thing to remember while building a sample is representativeness. We assume that the sample represents the population, so any inferences we draw about the sample will be true of the population

STATISTICAL ESTIMATION



What kind of inferences can we make in statistics?

1. We can estimate unknown population parameters based on properties of a sample
2. We can test hypothesis about a population based on sample parameters



CASE STUDY

You work for Airline X as GM Sales. For any flight, let's assume for simplicity that there are always 100 seats. When selling flight tickets, you could always sell exactly 100 tickets. However, you know that in most flights, some people will not show up, so you could maximize sales by reselling that ticket.

How do you decide how many tickets you should oversell for each flight?

Remember, if you oversell by more than the number of people who don't show, you will lose money in terms of putting those people on other flights plus some compensation



TYPES OF DATA

Another critical basic concept that we need to cover is data types.

In general, two basic types of data

- Quantitative
- Qualitative

Quantitative		Qualitative	
Discrete Ex: Number of applications	Continuous Ex: Income of applicants	Nominal Gender: M,F	Ordinal Size: S,M,L

- All the summary statistics that we have discussed apply to what kind of data?

SUMMARY STATISTICS FOR QUALITATIVE DATA

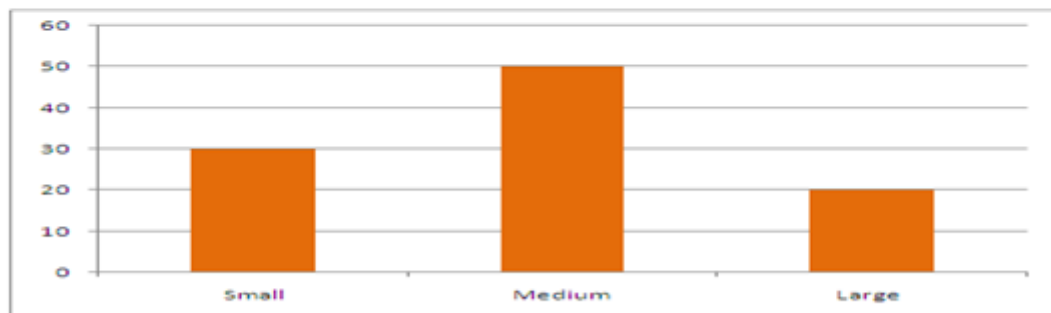


Frequency Distributions

Number of observations of each level displayed

Example:

- Small: 30%, Medium, 50%, Large: 20%
- High Income: 25%, Low Income: 75%





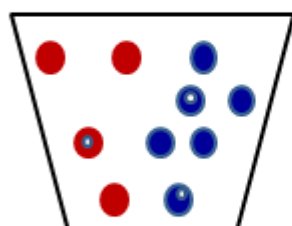
PROBABILITY THEORY

1. Probability is very important to understand statistical inference, which is the foundation of analytical methods
2. In general, it is not possible to truly know a population because it is too large and difficult to collect data. So we work with samples of data.
3. In order to get from sample inferences to population (which is not known), we need to understand probability theory and its application



PROBABILITY – THE BASICS

Probability of event A happening = $P(A)$ = number of outcomes where event A occurs / total no. of possible outcomes

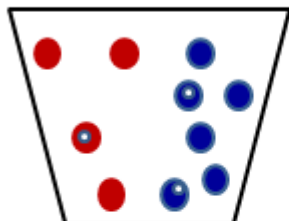


Probability of drawing a red ball = $P(A) = 4/10 = .4$

Probability of drawing a dotted ball = $P(B) = 3/10 = .3$

Probability of drawing a red dotted ball = $P(AB) = 1/10 = .1$

CONDITIONAL PROBABILITY



Probability of a ball being dotted given that it is

$$\text{red} = P(B | A) = P(BA) / P(A) = .1 / .4 = .25$$

Probability of a ball NOT being dotted given that

$$\text{it is red} = 1 - P(B | A) = 1 - P(BA) / P(A) = .75$$

Probability of a ball being blue given that it is dotted?



CASE STUDY

You work for Airline X as GM Sales. For any flight, let's assume for simplicity that there are always 100 seats. When selling flight tickets, you could always sell exactly 100 tickets. However, you know that in most flights, some people will not show up, so you could maximize sales by reselling that ticket.

How do you decide how many tickets you should oversell for each flight?

Remember, if you oversell by more than the number of people who don't show, you will lose money in terms of putting those people on other flights plus some compensation

RANDOM VARIABLES AND PROBABILITY



We now should spend some time understanding the concept of random variables and how they apply to statistical theory

Let's say you are taking a math test. If you answer all the questions correctly, you will get 100%. This is not "random"

But let's say you are in a class of 50 students. How many students will score 100%? That is a random variable

A random variable is one that takes a numerical value whose outcome is determined by an experiment.

In the student example, the test is the experiment

ANOTHER EXAMPLE



Let's look at a very commonly used example: the coin flip.

- Suppose you flip a coin 10 times. How many times will you get Heads?

Use the probability theory example that we have reviewed:

- Probability of heads in any coin flip = ?
- The number of heads is a random variable because it is possible to get a different number each time the experiment is repeated:
 - Trial 1: 10 Flips: 7 Heads
 - Trial 2: 10 Flips: 3 Heads
 - And so on



EXPECTED VALUE OF A RANDOM VARIABLE



In the coin flip experiment, if we are asked to provide the expected number of heads in 100 in 10 coin flips, could we provide one number?

The Expected Value is simply the sum of product of each outcome value with the probability of that outcome

Number of Heads in 10 Flips	Frequency	Relative Frequency (Probability)	Expected Value Calculation
0	1	0.01	0
1	2	0.02	0.02
2	3	0.03	0.06
3	7	0.07	0.21
4	18	0.18	0.72
5	30	0.3	1.5
6	20	0.2	1.2
7	12	0.12	0.84
8	5	0.05	0.4
9	1	0.01	0.09
10	1	0.01	0.1
Expected Value			5.14



CASE STUDY

In the airlines examples, the number of people who don't show for each flight is a random variable

Let's assume there are 10 flights per day. If we collect data on no-shows for each day for 100 days, we will get a distribution of outcomes, with no show percentages ranging from ?

For the most part, you would expect very few flights with 0% no show, 100% no shows. Most likely the most frequent % is somewhere between 2 - 10%

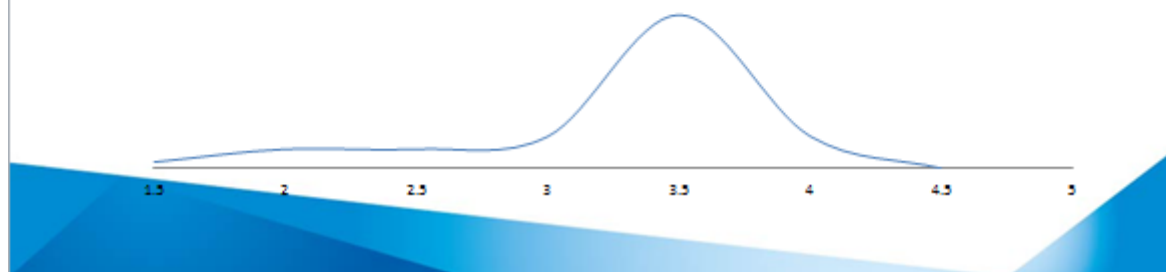
PROBABILITY DISTRIBUTION



Let's look at a database of baby birth weights

- If we plot the birth weights what would we expect to see?
- Fewer babies with very low or very high weights, and most babies within a fairly narrow range of weights
- Here is what the actual plot looks like. In order to create the plot, we have to essentially bin the weights and then calculate frequency for each range of weights

Baby Weight Plot



PROBABILITY DISTRIBUTION FUNCTION



A probability distribution therefore includes all possible outcomes of an experiment repeated n times

A **discrete probability distribution** is a list of discrete outcomes, and a **continuous distribution** is a product of an experiment with continuous possible outcomes

The function $f(x)$ is called a **probability density function** for the continuous random variable X where the total area under the curve bounded by the x axis is equal to 1

The area under the curve between any two ordinates $x = a$ and $x = b$ is the probability that X lies between a and b .

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION



The Binomial Distribution is an example of a probability distribution of a discrete random variable

The coin toss example is most widely used, but there are many other examples:

1. Gender of babies delivered in a hospital
2. Number of fatal side effect deaths for a Schedule H drug

The common feature across all these examples is:

1. There are only two possible outcomes: Win or Lose, 1 or 0, Male or Female
2. There are no external factors influencing the probability of each outcome over time
3. The chances of each outcome are independent of previous results

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION



You are the Finance Manager for Company X, and you are looking at your AR balances. Based on past data, you know that on average 40% of your customers are more than 60 days late with payments

Let's say you have a total of 5 customers, and you want to create a contingency for situations where >50% of your customers are late

Probability of 0 customer being late?

Overall probability of being late (p): 0.4

Probability of not being late (1-p): 0.6

Total number of trials (n) : 5

Total successes (x); 0

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

If you calculate this you will get: 0.0776

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION



- If you recalculate for 1 late customer out of 5:
 - $p = 0.2592$
- 2 Late customers:
 - $P = 0.3456$
- 3 late customers:
 - $P = 0.2304$
- 4 late customers:
 - $P = 0.0768$
- 5 late customers:
 - $P = 0.0120$

So p of $> 50\%$ customers being late = $0.2304 + 0.0768 + 0.0120 = \mathbf{0.3192}$

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION



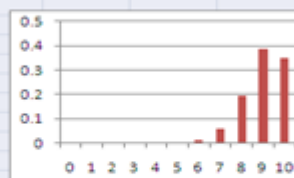
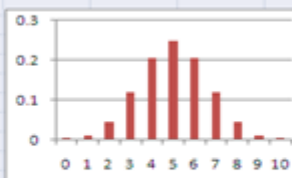
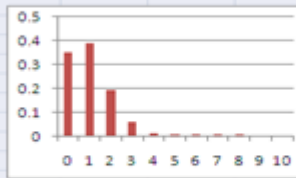
Things to note:

1. Irrespective of the actual experiment, as long as number of trials are 5 and probability of success is 0.4, $p=0$ is always 0.0776
2. If probability of success is 0.4, it makes intuitive sense that probability of success = 0 is a low number relative to probability of success, as is probability of success = 5
3. We can calculate the probabilities for a family of binomials, where $n = 5$ but p changes from 0.1 to 0.9
4. We can also calculate probabilities for any n and success combination, and these will be fixed values

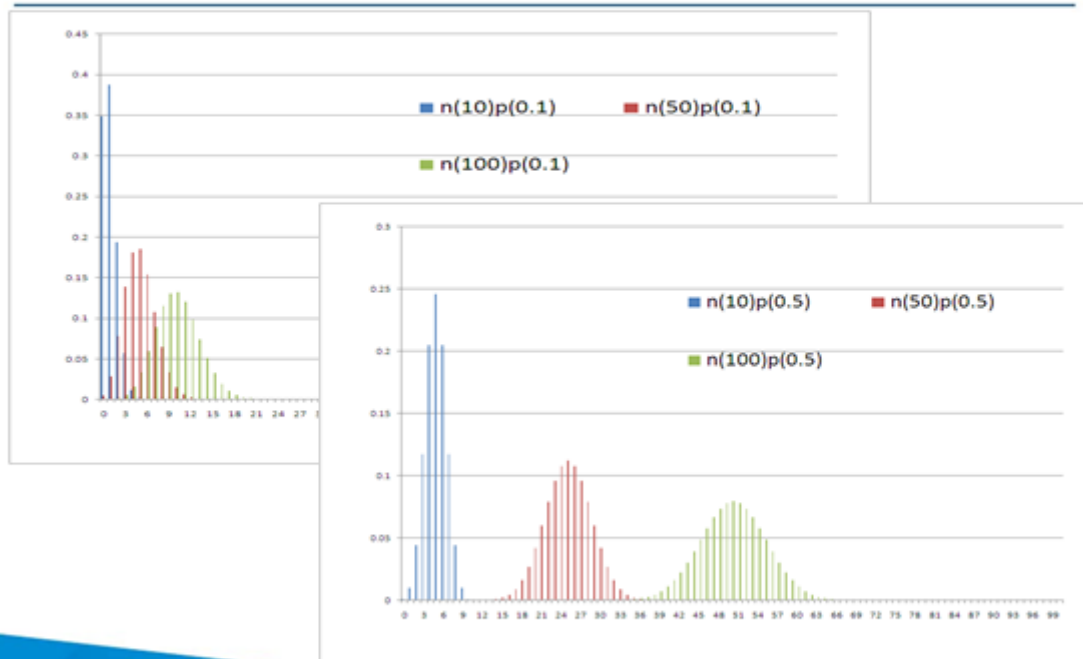
DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3		Binomial(10, p)			Binomial(10, p)			Binomial(10, p)				
4		n	p		n	p		n	p			
5		10	0.1		10	0.5		10	0.9			
6												
7		k	P(X=k)		k	P(X=k)		k	P(X=k)			
8		0	0.348678		0	0.000977		0	1E-10			
9		1	0.38742		1	0.009766		1	9E-09			
10		2	0.19371		2	0.043945		2	3.64E-07			
11		3	0.057396		3	0.117188		3	8.75E-06			
12		4	0.01116		4	0.205078		4	0.000138			
13		5	0.001488		5	0.246094		5	0.001488			
14		6	0.000138		6	0.205078		6	0.01116			
15		7	8.75E-06		7	0.117188		7	0.057396			
16		8	3.65E-07		8	0.043945		8	0.19371			
17		9	9E-09		9	0.009766		9	0.38742			
18		10	1E-10		10	0.000977		10	0.348678			
19		total	1		total	1		total	1			
20												
21												
22												
23												
24												
25												
26												
27												
28												
29												
30												



DISCRETE PROBABILITY DISTRIBUTION FUNCTION: BINOMIAL DISTRIBUTION





BINOMIAL DISTRIBUTION TABLE

$$B(x; n, p) = \sum_{y=x}^n b(y; n, p)$$

The values of $B(x; n, p)$ for $0.5 < p < 1.0$ are obtained by using the formula

$$B(x; n, 1 - p) = 1 - B(n - 1 - x; n, p)$$

n	x	p									
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
5	0	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031
	1	0.977	0.919	0.835	0.737	0.633	0.528	0.428	0.337	0.256	0.188
	2	0.999	0.991	0.973	0.942	0.896	0.837	0.765	0.683	0.593	0.500
	3	1.000	1.000	0.998	0.993	0.984	0.969	0.946	0.913	0.869	0.813
	4	1.000	1.000	1.000	1.000	0.999	0.998	0.995	0.990	0.982	0.969
	5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001
	1	0.914	0.736	0.544	0.376	0.244	0.149	0.086	0.046	0.023	0.011
	2	0.988	0.930	0.820	0.678	0.526	0.383	0.262	0.167	0.100	0.055
	3	0.999	0.987	0.950	0.879	0.776	0.650	0.514	0.382	0.266	0.172
	4	1.000	0.998	0.990	0.967	0.922	0.850	0.751	0.633	0.504	0.377
	5	1.000	1.000	0.999	0.994	0.980	0.953	0.905	0.834	0.738	0.623
	6	1.000	1.000	1.000	0.999	0.996	0.989	0.974	0.945	0.898	0.828
	7	1.000	1.000	1.000	1.000	1.000	0.998	0.995	0.988	0.973	0.945
	8	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.995	0.989
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	0	0.463	0.206	0.087	0.035	0.013	0.005	0.002	0.000	0.000	0.000
	1	0.829	0.549	0.319	0.167	0.080	0.035	0.014	0.005	0.002	0.000
	2	0.964	0.816	0.604	0.398	0.236	0.127	0.062	0.027	0.011	0.004
	3	0.995	0.944	0.823	0.648	0.461	0.297	0.173	0.091	0.042	0.018
	4	0.999	0.987	0.938	0.836	0.686	0.515	0.352	0.217	0.120	0.059
	5	1.000	0.998	0.983	0.939	0.852	0.722	0.564	0.403	0.261	0.151
	6	1.000	1.000	0.996	0.982	0.943	0.869	0.755	0.610	0.452	0.304
	7	1.000	1.000	0.999	0.996	0.983	0.950	0.887	0.787	0.654	0.500
	8	1.000	1.000	1.000	0.999	0.996	0.985	0.958	0.905	0.818	0.696
	9	1.000	1.000	1.000	1.000	0.999	0.996	0.988	0.966	0.923	0.849
	10	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.991	0.975	0.941
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.994	0.982
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.996
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

BINOMIAL DISTRIBUTION MEAN AND STD DEVIATION



The Mean (or Expected Value) of a Binomial Distribution: **np**

The Std Deviation is:

where, $q = 1-p$

$$\sigma = \sqrt{npq}$$

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: NEGATIVE BINOMIAL DISTRIBUTION



In a negative binomial distribution, we look at the probability of r successes in n independent trials, where probability of success is constant across the trials

Examples:

1. Tossing a coin repeatedly till you get to 4 heads
2. Number of customers that need to walk in to get 20 sales in a store

The probability distribution here is the probability of seeing

- 4 heads in 4 throws
- 4 heads in 5 throws
- 4 heads in 6 throws and so on

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: NEGATIVE BINOMIAL DISTRIBUTION



$$P(X = r) = {}_{n-1}C_{r-1} p^r (1-p)^{n-r}$$

where,

n = Number of events.

r = Number of successful events.

p = Probability of success on a single trial.

$${}_{n-1}C_{r-1} = \frac{(n-1)!}{((n-1)-(r-1))! (r-1)!}$$

$1-p$ = Probability of failure.

Of course, you can just use the NEGBINOMDIST function in Excel

=NEGBINOMDIST(f,s,x), WHERE

f = Number of failures

s = Total number of successes

x = Number of Trials

H4		f_x	=NEGBINOMDIST(G4-2,2,0.5)
F	G	H	I
	Negative Binomial : 2 heads in number of throws		
	Number of Throws	Probability of 2 heads	
	2	0.25	
	3	0.25	
	4	0.1875	
	5	0.125	
	6	0.078125	
	7	0.046875	
	8	0.02734375	
	9	0.015625	
	10	0.008789063	
	11	0.004882813	
	12	0.002685547	

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: GEOMETRIC DISTRIBUTION



The Geometric Distribution Is a special case of the negative binomial distribution, and computes the probability of seeing the first success in r trials

Example:

1. Toss a coin till it lands on heads
2. Customers walk-in till first purchase

Geometric Probability Formula.

Suppose a negative binomial experiment consists of x trials and results in one success.

If the probability of success on an individual trial is P , then the geometric probability is:

$$g(x; P) = P * Q^{x-1}, \text{ where } Q = (1-P)$$

DISCRETE PROBABILITY DISTRIBUTION FUNCTION: GEOMETRIC DISTRIBUTION



K5		f_x	=NEGBINOMDIST(J5-1,1,0.5)	
	J	K	L	M
	Geometric: First head in number of throws			
	Number of Throws	Probability of first head		
2	2	0.25		
3	3	0.125		
4	4	0.0625		
5	5	0.03125		
6	6	0.015625		
7	7	0.0078125		
8	8	0.00390625		
9	9	0.001953125		
10	10	0.000976563		
11	11	0.000488281		
12	12	0.000244141		



POISSON DISTRIBUTION

Another discrete probability distribution, that is used to model number of events occurring in a time frame

Examples include:

1. Number of insurance claims in a month
2. Disease spread in a day
3. Number of telephone calls in an hour
4. Number of patients needing emergency services in a day

The following conditions apply to correctly use a Poisson Distribution:

- Events have to be counted as whole numbers
- Events are independent: so if one event occurs, it does not impact the chances of the second event occurring
- Avg frequency of occurrence for the given time period is known
- Number of events that have already occurred can be counted



POISSON DISTRIBUTION

Poisson Probabilities are calculated as:

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

Where, Lambda is the mean number of occurrences in a given interval of time .

Notice here that there is no n (sample size) impact

Similar to how the Binomial probabilities are calculated, for every X and Lambda we can calculate the Poisson Probabilities

Mean of a Poisson Dist is Lambda; Std Deviation is Sq Root of Lambda



POISSON DISTRIBUTION

You work as a manager in a call center. You have a staff of 55 people, who on average handle 330 calls in an hour. A major holiday is coming up and 5 resources want leave. You estimate that the 50 remaining resources can manage 20% greater calls, but want to plan for the chances if greater than 20% increased call volume.

What are the chances that the number of calls on that day will go up by more than 20%?

$r = (330)/55 = 6$ calls an hour;

20% greater calls with 5 less resources = $(330 \times 1.2)/50 = 8$ calls an hour

So we need Probability of seeing 8 or more calls an hour when average is 6.

You can either calculate it manually or look it up in the Poisson Distribution table

POISSON DISTRIBUTION TABLE



Microsoft Excel Ribbon: Home, Insert, Page Layout, Formulas, Data, Review, View							
Formulas Bar: =POISSON(A3,\$B\$1,FALSE)							
	A	B	C	D	E	F	G
1	mean, rate, r	P(exactly r calls)	P(r or fewer arrivals)				
2							
3	0	0.002478752	0.0024788				
4	1	0.014872513	0.0173513				
5	2	0.044617539	0.0619688				
6	3	0.089235078	0.1512039				
7	4	0.133852618	0.2850565				
8	5	0.160623141	0.4456796				
9	6	0.160623141	0.6063028				
10	7	0.137676978	0.7439798				
11	8	0.103257734	0.8472375				
12	9	0.068838489	0.916076				
13	10	0.041303093	0.9573791				
14	11	0.02252896	0.979908				
15	12	0.01126448	0.9911725				
16	13	0.005198991	0.9963715				
17	14	0.002228139	0.9985996				
18	15	0.000891256	0.9994909				
19	16	0.000334221	0.9998251				
20	17	0.00011796	0.9999431				
21	18	3.93201E-05	0.9999824				

$P(r \leq 8) = 0.84$
Therefore $P(r > 8) = ?$

CONTINUOUS PROBABILITY DISTRIBUTIONS



Continuous distributions are applicable when an event can take on any value within a given range

Examples include:

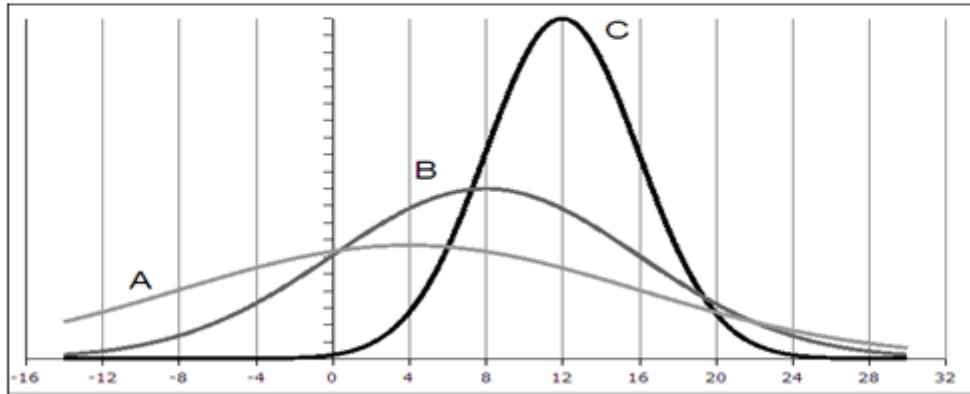
1. Height of males in Bangalore
2. Average waiting time per patient at a hospital
3. Per Capita Income

The most common kind of a continuous probability distribution is the Normal Distribution, on which we will spend some time because of its useful applications in statistics

NORMAL PROBABILITY DISTRIBUTION



1. Symmetric about the (single) mean
2. Mean = Median = Mode
3. The two tails extend indefinitely and never touch the axis



What are the differences between the curves?

NORMAL PROBABILITY DISTRIBUTION



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

- where:
- y = vertical height of a point on the normal distribution
 - x = distance along the horizontal axis
 - σ = standard deviation of the data distribution
 - μ = mean of the data distribution
 - e = exponential constant = 2.71828...
 - π = pi = 3.14159....

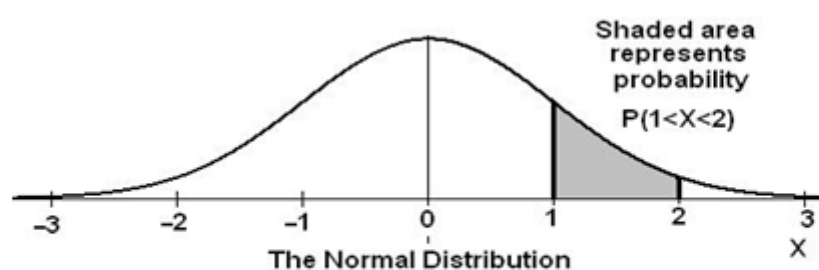
http://www.pavementinteractive.org/index.php?title=Normal_Distribution



NORMAL DISTRIBUTION

Area under the curve:

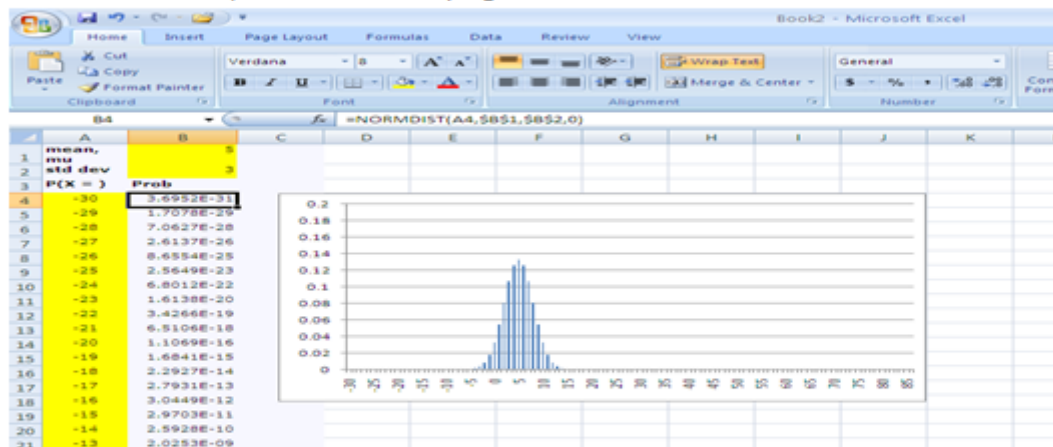
The total area under a normal probability curve is always 1. This property allows us to think of the area as probability, and therefore we can compute probability two values on the curve



NORMAL DISTRIBUTION



We can calculate probabilities of any X given mean and std deviation

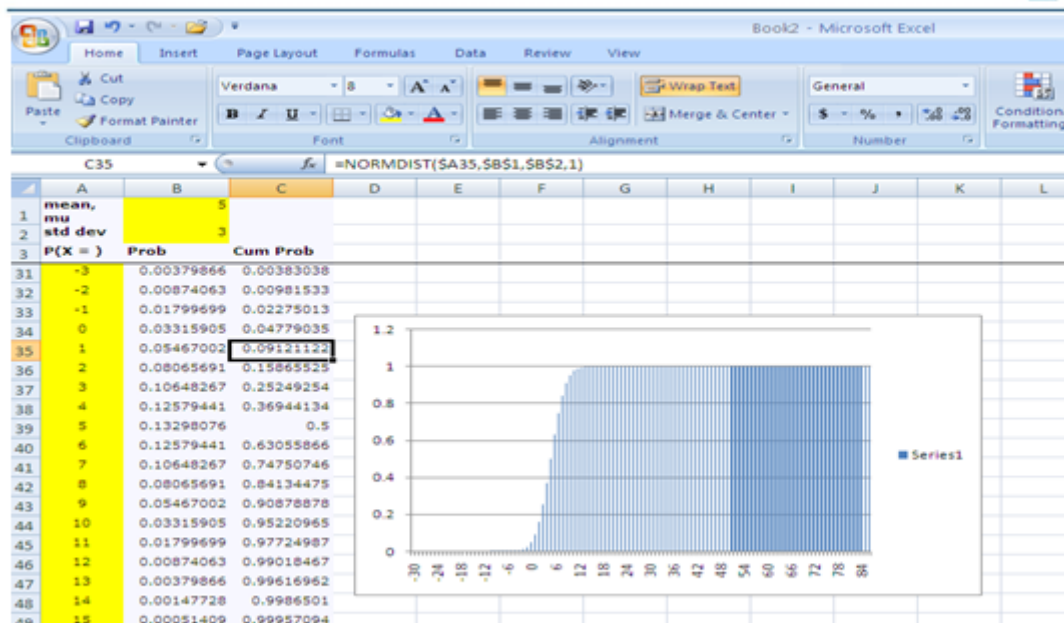


If average years of experience in company x is 5 years, with a standard deviation of 3, what is the probability that a person chosen randomly will have 2 years of experience?

If you actually look up the table, you will see it equals 0.08

What about probability that it is 2 or less?

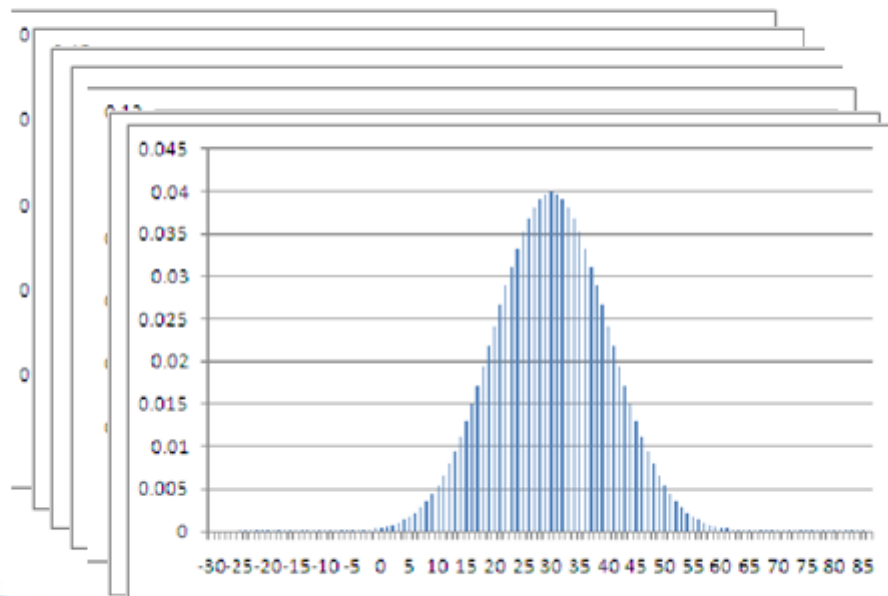
NORMAL DISTRIBUTION – CUMULATIVE P





NORMAL DISTRIBUTION

We can compute the probabilities for any Mean, Std Dev and an X





STANDARD NORMAL DISTRIBUTION

Instead of calculating probabilities for every combination of Mean and Std Deviation, we can standardize – express any X as number of deviations from the mean

So for example, if Mean = 5, Std Deviation = 3, then

$$2 = (2-5)/3 = -1 \text{ std deviation from the Mean}$$

$$3 = (3-5)/3 = -0.66 \text{ std deviation from the Mean}$$

$$4 = (4-5)/3 = -0.33 \text{ std deviation from the mean}$$

$$5 = (5-5)/3 = 0 \text{ std deviations from the mean}$$

$$6 = (6-5)/3 = 0.33 \text{ std deviations from the mean}$$

$$7 = (7-5)/3 = 0.66 \text{ std deviations from the mean}$$

$$8 = (8-5)/3 = 1 \text{ std deviation from the mean}$$

We already know what the probability is for X = 2 when Mean = 5 and Std Dev = 2. So probability of -1 std deviation from Mean for this distribution is known

What will be probability of -1 std deviation from Mean for dist where Mean = 6 and Std Dev = 3?

STANDARD NORMAL DISTRIBUTION



What does this imply? As long as something is -1 std deviation from any mean, p values will be the same. So you will not need to calculate the probability tables. You can just use this conversion

	mean	mu	std de	P(X =
35	1			
36	2			
37	3			
38	4			
39	5			
40	6			
41	7			
42	8			
43	9			
44	10			
45	11			
46	12			
47	13			
48	14			
49	15			
50	16	0.00051409	0.99957094	
51	17	0.00016009	0.99987713	
52	18	4.461E-05	0.99996833	
53	19	1.1124E-05	0.99999266	

STANDARD NORMAL DISTRIBUTION



The standard normal distribution is a special case of the normal distribution with a mean of 0 and a std deviation of 1 (why is mean 0?)

All normal distributions can be converted to std normal by the following formula

$$Z = \frac{X - \mu}{\sigma}$$

Where X is the value of the random variable in the original normal distribution, μ is the mean, and σ is the std. deviation of the original normal distribution

$$y_z = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

where: y_z = vertical height on the standard normal distribution

z = as defined above




STD. NORMAL DISTRIBUTION TABLE

How to read a Distribution Table:

Tables of the Normal Distribution

Probability Content
from $-\infty$ to Z



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



STANDARD NORMAL DISTRIBUTION

Case Study 1

You are the training manager for your company and are developing online modules for specific business modules. Once participants finish the content, they need to take a test consisting of 20 questions. You are trying to decide how many minutes to time the test.

You ask 5 colleagues chosen at random to take the test, and average time across the 5 is 45 minutes, with standard deviation of 7 minutes.

Which option would you choose for the test timing:

1. Average time taken is 45 minutes, so 45 minutes
2. 45 plus some buffer: maybe 50 minutes?
3. 45 minutes plus lots of buffer, maybe 80 minutes?

STANDARD NORMAL DISTRIBUTION



Case Study 2:

You are the general manager for retail chain A. You are looking at optimizing your inventory costs. For Brand A, an FMCG product, you are looking at weekly sales at stores in large metros. You find that on average weekly sales for Brand A are 600 units, with a std deviation of 150 units, and that average stock in any store at beginning of week is 400 units

Your minimum order quantity is 300. You do not want to be in a situation where in a particular week total stock at hand is to go below 300. What do you think the chances are of that situation happening – high or low?



Conditional Probability:

For two events A and B, conditional probability of A given B is:

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

Independent Events:

Two events A and B are independent if

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

Binomial Coefficient:

$$P(nCk) = n! / k! \cdot (n-k)!; \text{ where } k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 1;$$

This is the number of ways of choosing k elements from a set of n elements

Expected Value:

The Expected value of a discrete random variable is:

$$E(X) = \sum_{k=1}^{\infty} P(X = k) \cdot k$$

The expected value of a continuous random variable is:

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



JIGSAW ACADEMY

Analytics for Professionals

www.jigsawacademy.com