Reference Material for PGPBABI-SMDM Residency I

A. Importance of Statistics in Business Analytics

If technology were the competitive edge for business during the later part of the 20th century, for 21st century it is going to be knowledge. E-commerce, e-business, e-transactions as well as less technology intensive method of doing business nowadays generate a plethora of data. Historical data are collected and examined for knowledge gain. This is now applied through predictive modeling for further development of business.

Data are observed facts and current status of the business. Consolidation of facts leads to information gain. All spheres of business are coming to realize the advantages of analytics. Hence analytics specialists are coming from all types of backgrounds and trainings. Reliance on software has increased manifolds. It is therefore mandatory that software outputs are to be interpreted accurately. At the same time, analysts should realize the pitfalls of blind dependence on software. After all, beyond a certain point, one must not let the computing machines take over and dictate human analytical capability!

B. How Statistical Methods helps in Decision Making

Statistics is the art and science of collecting, presenting, analyzing and interpreting data.

If data is not used for better decision making through quantitative techniques, the data is wasted and the business performs in a suboptimal manner. While putting money in the bank ATMs, the distribution of footfalls and money withdrawn is considered so that the machines do not easily run out of money. During peak seasons and vacation times hotels and airlines plan for a larger roster so that better service may be ensured.

All businesses depend on past data to understand current status and plan for future. Depending on the size of business and its maturity in data handling, application areas of statistics are manifold. Some of the most common areas of statistical applications are marketing, including digital marketing; financial risk analysis; quality control of assembly line production; sample audit etc. However, operations management and supply chain and human resource management areas also benefit by application of appropriate statistical methodology.

Following is an example of a small data set on Indian Premier League teams' revenues in 2009.

Table 1: Revenue of IPL Teams (2009)

	Amount and Sources of Revenue 2009				
	Broadcasting	*Sponsorships*	*Other Income*	*Prize Money*	*City*
Rajasthan Royals	67.5	24	14.2	0.7	Jaipur
Chennai Super Kings	67.5	24	18.5	1.2	Chennai
Kolkata Knight Riders	67.5	24	18.9	0.4	Kolkata
Kings XI Punjab	67.5	24	14.3	0.8	Chandigarh
Delhi Daredevils	67.5	24	14.7	1.2	Delhi
Mumbai Indians	67.5	24	14	0.5	Mumbai
Bangalore Royal Challengers	67.5	24	13.5	2.25	Bangalore
Deccan Chargers Hyderabad	67.5	24	13.5	4.5	Hyderabad

Typically data set is represented by a matrix each row of which represents an observation and each column represents a variable. The value in each cell of the matrix is an observed value or a value realized by the variable mentioned in the column heading. For example, the above data set contains 8 observations and each observation has data on 4 variables.

Scales of Measurements and Data Types

Data is collected in one of the following scales of measurements: **nominal, ordinal, interval** or **ratio**. Ratio scale is the most advanced scale of measurement and renders itself to the most extensive quantitative analysis. In Table 1 City is a nominal variable whereas the other 4 variables are all measured in ratio scale.

Data can also be classified as **qualitative** or **quantitative**. Ordinal, interval and ratio scale data are also known as quantitative data. Depending on the nature of data collection, data may also be known as **cross-sectional** or **time-series**. Table 1 is an example of cross-sectional data where all observations are made at the same point of time. However, if data is collected on the same variable at different points of time, it would be known as time-series data. Following table shows KKR's winning across years. This is an example of time-series data.

Table 2: Performance of KKR (2008 – 14)

Kolkata Knight Rider (KKR) Results	
Year	Win
2008	50%

2009	29%
2010	50%
2011	53%
2012	71%
2013	43%
2014	69%

Data Summarization

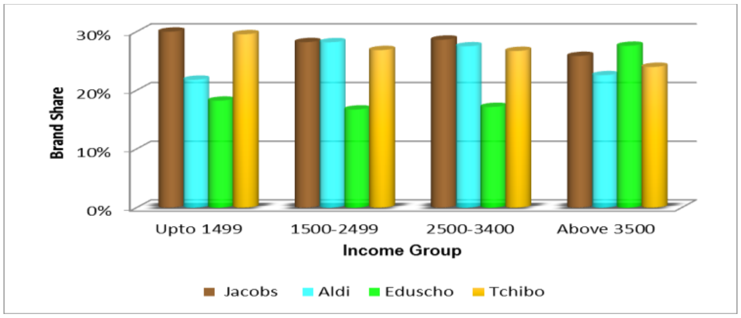
When data set is large, i.,e. data is collected on a number of observations and on many variables, summarization is paramount to extract any information out of the data. Two modes of summarization are primarily used

- Visual summarization – graphs charts etc
- Numerical summarization – measures of central tendencies and dispersion

Visual Summarization

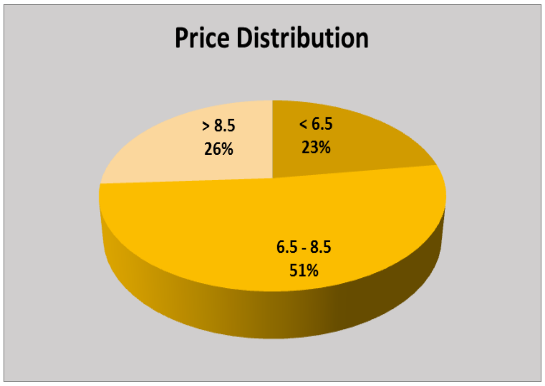
Data may be organized in various manner and the appropriate method of organization depends on the nature of the variable. Summary table presents tallied responses as **frequencies** or **relative frequencies** for each category . If the variable under consideration is a qualitative variable then a bar chart or a pie chart will be appropriate. An intelligently represented chart can impart a large amount of information at a glance. The following **bar diagram** is created on more than 1,30,000 coffee packet purchase records in former West Germany between Jan 1988 and Dec 1990. It shows coffee brand choices across income groups.

Figure 1: Coffee brand choice across income group among German consumers



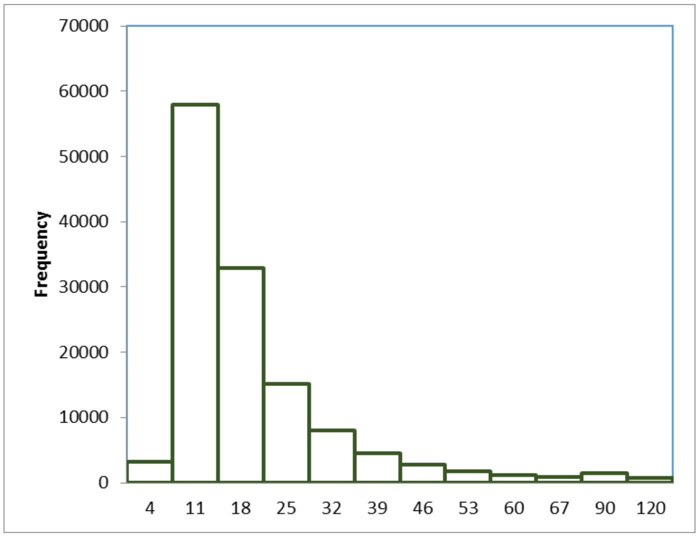
Pie charts are used to understand the relative distribution of different categories. Following chart looks at the price distribution of the coffee packets bought.

Figure 2: Price per packet of coffee purchased by German consumers



The most common visualization tool for continuous variables is the **histogram** – a grouped bar chart. On the X-axis is presented the class limits or the midpoints of the classes and on Y-axis is presented the frequencies of that class. The following histogram shows the purchase pattern of coffee by German consumers. On the X-axis is shown the number of days after which consumers have bought coffee in supermarket. About 3000 coffee purchases are made around 4 days whereas about 60000 purchases are made around 11 days.

Figure 3: Coffee purchase pattern among German consumers



From the above diagram it is easy to note that approximately 60,000 purchases are made at a gap of 2 weeks (11 days to be precise). Among the consumers studied, household purchase of coffee is made about once in 2 weeks. But there are a few households where coffee purchases are made extremely infrequently. Hence the shape of the purchase pattern distribution is not symmetric, but right-skewed.

Every visualization tells a story. There are many options for data visualization and there are no hard and fast rules regarding which visualization is to be applied in which situation. However there are some guiding principles. The most appropriate application must relate the method of visualization to the business question that needs to be answered. Visualization is an interactive process. It starts with answering simple questions and it should give rise to deeper questions and may even indicate the path to the proper analytical technique.

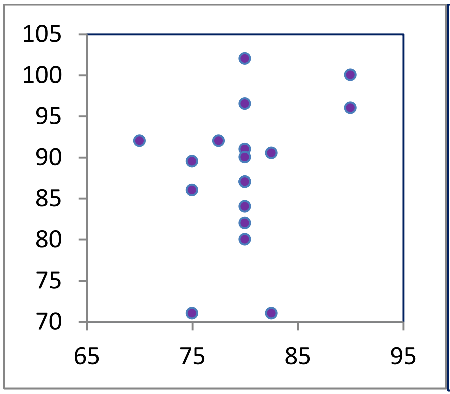
Very rarely we are interested in only one variable from each observation. When two (or more) variables are observed on each unit or individual in a sample, the data is known as **bivariate** (or **multivariate**) data. The following is an example of a bivariate data set.

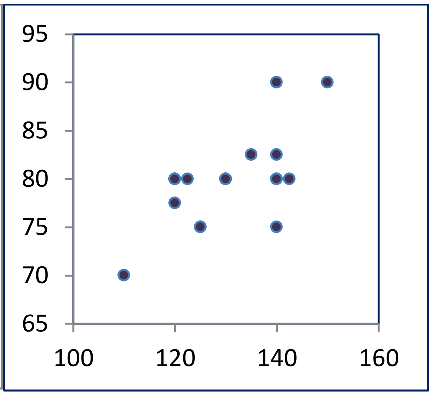
The managing director of a consulting group has the following monthly data on total overhead cost and professional labour hours to bill to clients (in rupees):

Total (1000)	340	400	435	477	529	587
Billable (1000)	3	4	5	6	7	8

For each month values on a pair of variables are observed. If the pairing is violated, the data will be different.

Scatterplot is a graph to visualize whether two variable X and Y are related in any manner. If X and Y have no dependency then the pattern will be completely random, i.e. all the data points will be uniformly distributed within a band. If X and Y have a positive dependency, i.e. if on the average Y increases with increasing X, then the points will form a band from left lower corner towards right upper corner. If X and Y have a negative dependency, i.e. if on the average Y decreases with increasing X, then the points will form a band from left upper corner towards right upper corner. Following are two typical diagrams showing (i) no relation (ii) positive relation.





Numerical Summarization

Visualization does not provide answer to all business questions. To understand the nature of the variables numerical summarizations are also necessary. The most important descriptive measure of a variable is its **central tendency**, or a single number that may be used as a representative of the whole. For qualitative variable the measure of central tendency is the **mode**, or the level with the highest frequency. For quantitative variable the measures are the **mean** (arithmetic average) or the **median** (the middlemost ranked value).

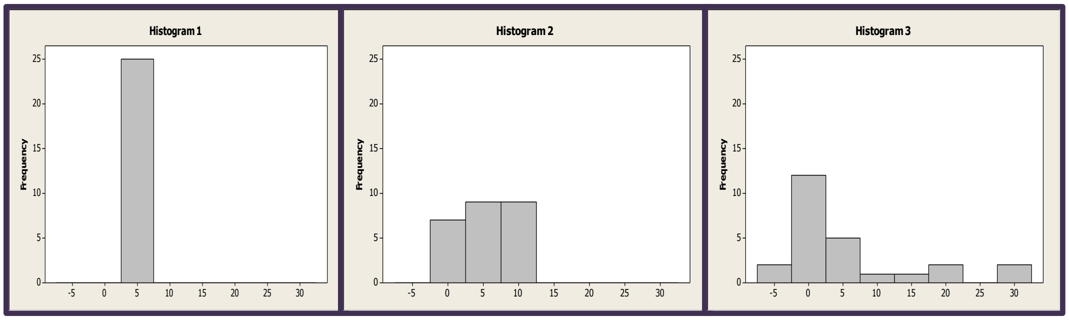
Example: A systems manager in charge of a company's network keeps track of the number of server failures that occur in a day. Following data represents the number of server failures in a day for the past two weeks:

No of days	0	3	0	3	30	2	7
No of failures	4	0	2	3	3	6	3

Mean number of failures is 4.7 and median number of failures is 3.

Relative position of Mean and Median helps to identify the **shape** or skewness of the distribution.

Measures of central tendency fail to tell the whole story. The spread or dispersion of the variable is also equally important to understand. The following fictitious example show three distinct distributions with number of observations 25 and identical mean 5.



The most commonly used measure of dispersion is the **variance** or its positive square root, the **standard deviation**. It gives an idea about the spread of the data from the central values. Therefore the first step in studying any continuous variable is to consider its mean and standard deviation jointly.

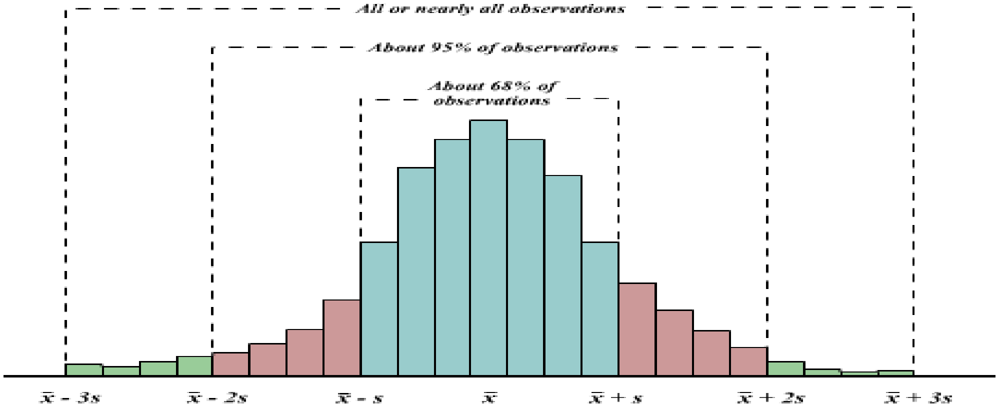
Empirical Rule: If data distribution (histogram) is symmetric with a single peak in the middle and tapering off gradually in the tails then

❑ The interval $\bar{x} \pm s$ contains approximately 68% of data

❑ The interval $\bar{x} \pm 2s$ contains approximately 95% of data

❑ The interval $\bar{x} \pm 3s$ contains 99.7% of data

Since many data distribution show approximate bell shaped distribution Empirical Rule has very wide applicability.



Bivariate Data: Correlation

Scatterplots show existence and direction of relationship between two variables but do not quantify the strength of relationship. To quantify the strength of relationship **Correlation Coefficient** between the variables is defined. Correlation coefficient may be positive or negative. It is a number always between +1 and − 1; 0 implies no linear relation between X and Y.

It is very important to realize that all statistical dependency is 'on the average'. Positive dependency of Y on X means that overall there is a relationship between Y and X so that with increasing X, Y also increases. That does not mean for each and every pair of (X_1, Y_1) and (X_2, Y_2) , $X_1 > X_2 \Rightarrow Y_1 > Y_2$.

C. Sample and Population – Moving from Descriptive to Inferential Statistics

Descriptive statistics is concerned about collection, summarization, visualization, presentation and overall description of data. Note that the data set that an analyst works on is a **random sample** from an underlying **population**. A population is the largest possible group with a definite set of characteristics.

A small subset of items or individuals from the population is known as a sample from the population. Sample size is typically only a fraction of the population size. One of the principal objectives of statistics is to understand the population through the sample. Descriptive statistics (Exploratory Data Analysis or EDA) is applicable on the sample. Extension of the knowledge gathered from the sample is applied to population with a certain degree of uncertainty. **Probability** is the study of uncertainty in the population. A numerical measure that characterizes a population is known as a parameter. A numerical measure that is computed from the sample is known as a statistic. Appropriate statistics are used to infer about population parameters, again through applications of probability distribution.

BASIC PROBABILITY

Probability is a quantitative measure of chance or uncertainty. It is a number between 0 and 1 (or 0% and 100%) and is always associated with an event, which is outcome of an experiment. An experiment may be any occurrence, which is repeatable and which may result in several possible outcomes. An event, which never occurs, has associated probability 0. An event, which always occurs, has associated probability 1. Example of experiments are as follows:

- Tossing a coin: Resulting outcomes: Head or Tail
- Setting up a new business : Breaks even within 1 year, 2 year or 3 year or after 3 year
- Sanctioning loan to an applicant: Bank makes a good decision or a bad decision (loan defaults)

Sample space is defined as the set of all possible outcomes of an experiment. All possible outcomes of an experiment are assumed to be **Equally Likely**. Two events are said to be **mutually exclusive** if occurrence of one precludes occurrence of the other. If A denotes the event that a new business breaks even within 2 years and B denotes an even that the same business breaks even after 4 years then A and B are mutually exclusive events. If C denotes an event that the business breaks even before 6 years then B and C are not mutually exclusive. Two events are **complementary** if together they cover the whole sample space and they are mutually exclusive. Complementary event of A is the event that the business will break even after 2 years.

Probability Rules

Probability of an event A is defined to be the relative frequency of that event. If an event A is defined as a collection of simpler events, then probability of A is the sum of the probabilities of the constituent events, if those events are mutually exclusive. If A and B are two events then the event A union B (A U B) occurs when A occurs or B occurs or A and B both occurs. The event A intersection B (A ∩ B) occurs when A and B both occur simultaneously.

- P(A U B) = P(A) + P(B) if A and B are mutually exclusive
- For any two events A and B, P(A U B) = P(A) + P(B) – P(A ∩ B)
- If A and B are mutually exclusive events P(A ∩ B) = 0

Example: The following table shows cross-classification of several students according to their employment status:

	Full-time	Part-time	Unemployed	All
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

P(A) = Probability of unemployment among females = 6/33

P(B) = Probability of employment among females = 3/33 + 24/33 = 27/33

Being unemployed and being employed are complementary events => P(B) = 1 – P(A)

Consider two other events : E and F where E : being employed and F : being female

E U F : either being employed or being female

E and F are not mutually exclusive events; being employed and being female can occur simultaneously

P(E U F) = P(E) + P(F) – P(E∩F) = (10+43)/62 + 33/62 - (3+24)/62 = 59/62

Conditional Probability and Independence

Conditional probability considers relative frequency of an event A *after* it has been known that the event B has already occurred. Probability of A given B:

$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, where P(B) ≠ 0

Continuing with the above example:

Probability that a randomly chosen student is a female P(F) = 33 / 62

Probability that a randomly chosen student is employed P(E) = 53 / 62

Probability that a randomly chosen student is a part-time employed male P(PTM) = 19 / 62

Probability that given a randomly chosen student is a female, she is full-time employed P(FTF | F)

= 3 / 33

Probability that given a randomly chosen student is part-time employed, the person is a male P(M | PTE) = 19 / 43

Note the differences between unconditional and conditional probabilities. Usually this difference is reflected in the denominator of the ratio.

Another very important concept is **independence** of two events. Two events A and B are said to be independent, if, whether or not A occurs has no impact on the occurrence of B. This concept is very difference from mutual exclusive events. In fact mutually exclusive events are not independent, because the fact that one of them occurs implies that (or impacts) the other will not happen (occurrence of the other).

A and B are independent events if P(A | B) = P(A) or if P(B | A) = P(B) or in other words, if

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

Example of independence: It has been seen that vernacular dailies are read by 80% of urban literate population. What is the probability that two randomly chosen persons have read vernacular dailies today? Here we have no reason to believe that one person's behaviour will have any impact on the other person. Hence P(Both have read) = (0.8)(0.8) = 0.64 = 64%

Random Variable and Probability Distribution

Random variables are at the basis of every statistical technique. Each variable, in statistical parlance, is a random variable. A random variable can take only numerical values and each value of a random variable corresponds to one or more outcomes of an experiment. Since events have different probabilities of occurrence, a random variable takes different values with different probabilities. Following is an example of random variables associated with experiments:

Table 1: Numerical description of outcome of an experiment

Experiment	Random Variable	Type of Variable
Taking loan	Loan amount	Continuous
Satisfying customer	Level of satisfaction	Categorical, Ordinal
Choosing a TV channel	Time spent watching	Continuous
Performance of a country's economy	S&P Credit Rating	Categorical, Ordinal
Choice of a soap	Listing of brands	Categorical, Nominal
Choice of generic drug	Choice of generic drug	Binary

A random variable may be either **discrete** or **continuous**. Some examples of a discrete random variable are Choice of soap, Preference of TV News Channel, Defaults on a loan, Number of declared / non-declared bank accounts, etc. Some examples of continuous random variable are Amount of loan, Duration of loan, Interest rate, Customer satisfaction on a scale of 1 to 10, etc. The collection of values a random variable (r. v.) takes and the associated probabilities is known as the **probability distribution**. Probability distribution of a random variable is often characterized by one or more parameters, e.g. Mean, Variance, Proportion etc.

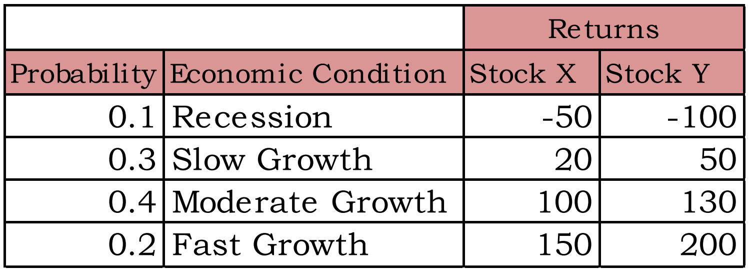
The two most important population parameters are its **expectation** and **variance**. The expectation or mean of a probability distribution is a measure of its centrality or location. Variance of a random variable is the expected squared deviation from the mean. Typically population parameters are denoted by Greek alphabets and sample statistics by Latin alphabets. Notationally

Mean : $\mu = E(X) = \sum x\ P(x)$

Variance: $\sigma 2 = Var(X) = \sum(x - \mu)2\ P(x)$; Standard deviation: $\sigma = \sqrt{Var(X)}$

Example of A Probability Distribution

Consider a random variable depicting condition of an economy. This random variable can take 4 values with different probabilities and according to the economic condition two stocks perform differently. Following table shows the returns:



		Returns	
Probability	Economic Condition	Stock X	Stock Y
0.1	Recession	-50	-100
0.3	Slow Growth	20	50
0.4	Moderate Growth	100	130
0.2	Fast Growth	150	200

A portfolio manager would like to advise her client on investment strategy. If the portfolio manager's advice is based solely on relative performance of the two stocks, then she would like to know how the stocks are expected to perform and what would be deviations from the expectations.

E(X) = 71;   Var(X) = 3829;

E(Y) = 97;   Var(Y) = 7101

Based on the above observations it is clear that expected performance of Stock Y is higher than that of Stock X. But variance of Stock Y is also higher, hence it may be more risky investment than Stock X.

A Common Discrete Distribution : Binomial

Suppose an experiment is repeated n times and each of the n trials is independent. The experiment has only two outcomes – success and failure, and in each trial the probability of success is fixed. If X is the random variable counting the number of successes in n trials, then X is said to follow a binomial distribution. The parameters of a binomial distribution are

n: the number of trials and π=prob(success).

Expected value of X is nπ and variance of X is nπ(1- π).

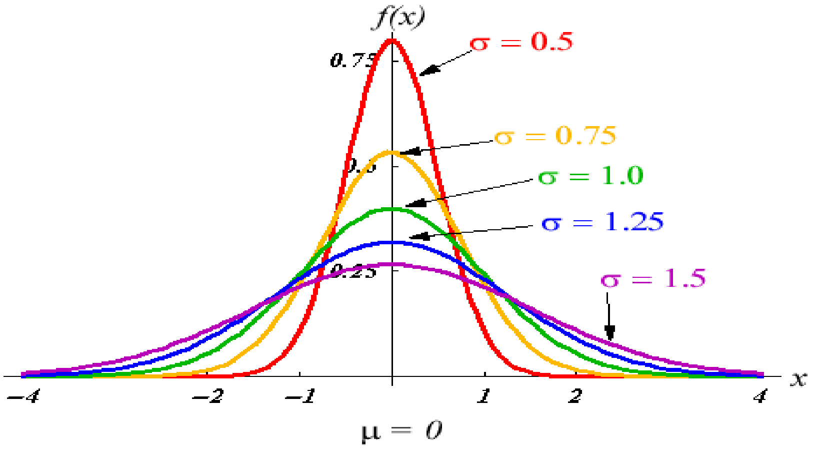
An example: My favourite food joint has the reputation of filling 90% of orders correctly the first time. If this joint fills 125 orders on a given day, what is the average number of orders filled correctly?

Note that n = 125 (the number of trials) and the trials are independent – there is no reason to believe that correct or incorrect filling of one order has any impact on the next one. Probability that each order is filled correctly (success) is constant (90%). Hence expected number of orders filled correctly in the given day is E(X) = 125(0.9) = 112.5. (This need not be rounded up. Why?).

Binomial probability distribution is completely characterized by its parameters n and π. This means that, if the number of trials is known and the success probability of each trial, which is a constant, is known, then one can find out the probability of all the events.

A Common Continuous Distribution : Normal

The most common continuous probability distribution is called the normal distribution which is symmetric and bell-shaped. Many continuous variables that are commonly encountered follow normal distribution and many others may easily be approximated by normal distribution. Normal distribution is characterized by its mean μ and standard deviation σ. The following graph shows the dependency of the spread of the curve according to the values of σ. The smaller σ is, the more concentrated and peaked is the curve around the mean value. Note that for normal distribution mean, median and mode of the distribution are identical and equal to μ.

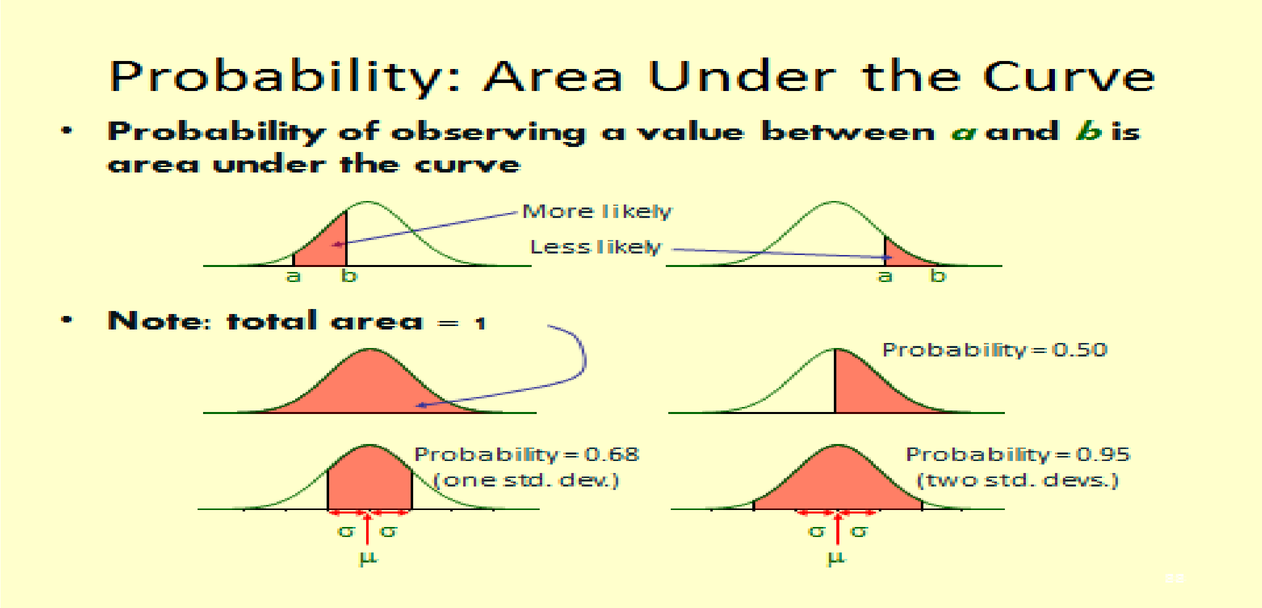


A special case of normal distribution is when μ = 0 and σ = 1. This is known as the **standard normal** distribution and the random variable is denoted by Z. Any normal distribution may be converted to standard normal distribution by the following rule

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

This is called a Z-score and has many applications.

The reason why normal distribution is so commonly used, is that, it is very easy to evaluates the probability under a normal curve, which is commonly known as the area under the curve. If it is known that the distribution is normal, then probability of events like [a < X < b] for any arbitrary (a, b) is also known for any μ and σ. It is also very easy to find normal percentile points, i.e. a, such that Pr[X > a] = p.



Probability: Area Under the Curve

- Probability of observing a value between *a* and *b* is area under the curve
- Note: total area = 1

Sampling and Sampling Distribution

Recall that the population is always unknown and often infinite. Numerical values of parameters that characterize a population is never known for sure, but are to be estimated through sample statistics. If a sample is a representative sample of the population, the sample statistics will be good or accurate estimates for the population parameters. Note, however, that numerical values of sample statistics will change from one sample to the next. Hence any statement regarding population parameter made on the basis of a sample statistics must have a probability attached to it.

Sample mean and sample standard deviation are estimates of μ and σ respectively. Observed proportion of success in n repetitions of trials is an estimate of π. Accuracy of sample measurements can be estimated, never known. Good sampling procedure minimizes inaccuracy, never eliminates it.

Central Limit Theorem

Suppose n sample observations are taken from a population with mean μ and standard deviation σ. Let $\bar{x}$ and s denote sample average and sample standard deviation respectively. If the sample size is large enough ($n \geq 30$) then, $\bar{x}$ follow a normal distribution with mean μ and standard deviation $^{\sigma}/_{\sqrt{n}}$.

This is one of the cornerstone of statistics and emphasises the fact that even if the parent population is not normal, sample average follows a normal distribution with the population mean as its mean and standard deviation much smaller than the population standard deviation.