Basic Statistics

1) Data Types – Continuous, Discrete, Nominal, Ordinal, Interval, Ratio, Random Variable, Probability, Probability Distribution
2) First, second, third & fourth moment business decisions
3) Graphical representation – Bar plot, Histogram, Boxplot, Scatter diagram
4) simple Linear Regression
5) Hypothesis Testing

SLIDE-13

Data types:

1) Continuous 2) Discrete

SLIDE-14

Data types: Preliminaries

Normal: Merely labels, no further information can be gleaned.

Ex: “coke” and “Pepsi”

Ordinal: Conveys only up to preference information. Direction alone.

Ex: “I prefer coffee to tea”

Interval: Conveys relative magnitude information, in addition to preference.

Ex: “I rate coke a 7 and Pepsi a 4 on a scale of 10.

Ratio: Conveys information on an absolute scale.

Ex: “I paid Rs11 for coke and Rs13 for Pepsi”.

NOMINAL	ORDINAL	INTERVAL	RATIO
Mode	Mode	Mode	Mode
Frequencies	Median	Median	Median
Percentages	Frequencies	Mean	Mean
	Percentages	Frequencies	Frequencies
	Some Statistical Analysis	Percentages	Percentages
		Variance	Variance
		Standard Deviation	Standard Deviation
		Most Statistical Analysis	Ratio of numbers
			All Statistical Analysis

SLIDE-15

Random Variable

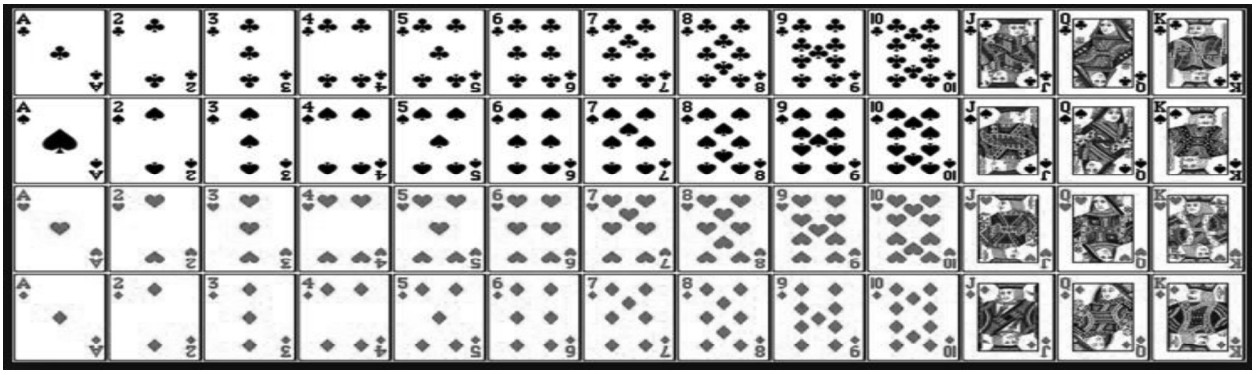
A random variable describes the probabilities for an uncertain future numerical outcome of a random process.

It is variable because it can take one of several possibilities.

It is a random because there is some chance associated with each possible value.

SLIDE-16

Poker cards example:



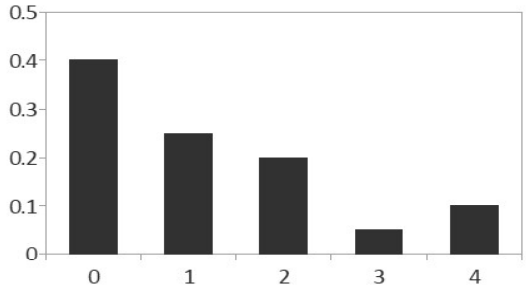
Suppose you have randomly picked a card from the card deck. What is the probability that this card will be?

- Bigger than 10?
- Equal to or Bigger than 10?
- Smaller than 3
- Greater than 4 and less than 8

SLIDE-17

The daily sales of large flat panel TVs at a store (X)

x	P(X=x)
0	0.40
1	0.25
2	0.20
3	0.05
4	0.10



What is the probability of a sale?

What is the probability of selling at least three TVs?

What is probability of sale?

What is the probability of selling at least 3 tv's?

SLIDE-18

Sampling Funnel:

1) Population 2) Sampling frame 3) SRS 4) Sample

SLIDE-19

Measures of central tendency

First moment Business decision:

Population –Mean or Average (μ) = $(\sum (x_i))/N$

Sample-Mean or Average ($\overline{X}$) = $(\sum (x_i))/n$

Median - Middle value of the data

Mode - Most occurring value in the data

SLIDE-20

Measures of Dispersion

Second moment Business decision:

Range= Max-Min

Population variance = $\sigma^2 = (\sum(X-\mu)^2)/N$

Population standard deviation =sqrt $((\sum (x_i\text{-population mean})^2)/N)$

Sample variance

$(\sum(x-\overline{x})^2)/ (n-1)$

Sample standard deviation = sqrt $((\sum (x_i\text{-sample mean})^2)/(n-1))$

SLIDE-21

Expected Value

For a probability distribution, the mean of the distribution is known as the expected value

The expected value intuitively refers to what one would find if they repeated the experiment an infinite number of times and took the average of all of the outcomes

Mathematically, it is calculated as the weighted average of each possible value

The formula for calculating the expected value for a discrete random variable X, denoted by μ, is:

$\sum Xp(X)$

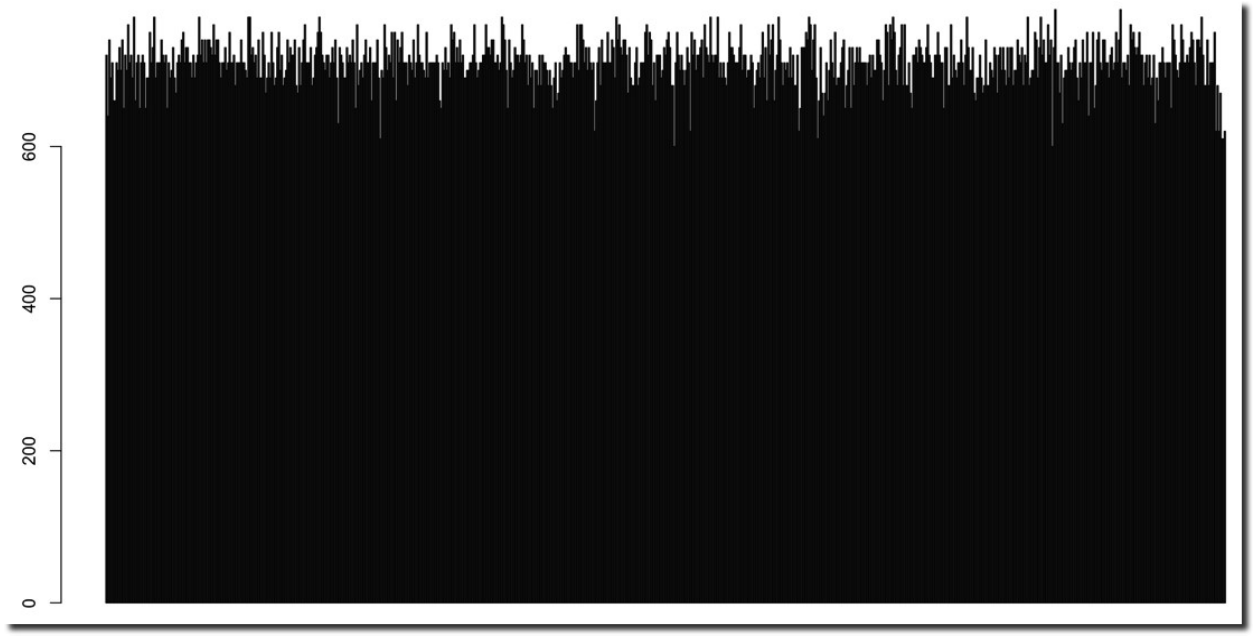
The variance of a discrete random variable X, denoted by σ2 is

$\sigma^2 = \sum [(x-\mu/\sigma)]^2 = \sum (x- \mu)^2 p(x)$

SLIDE-22

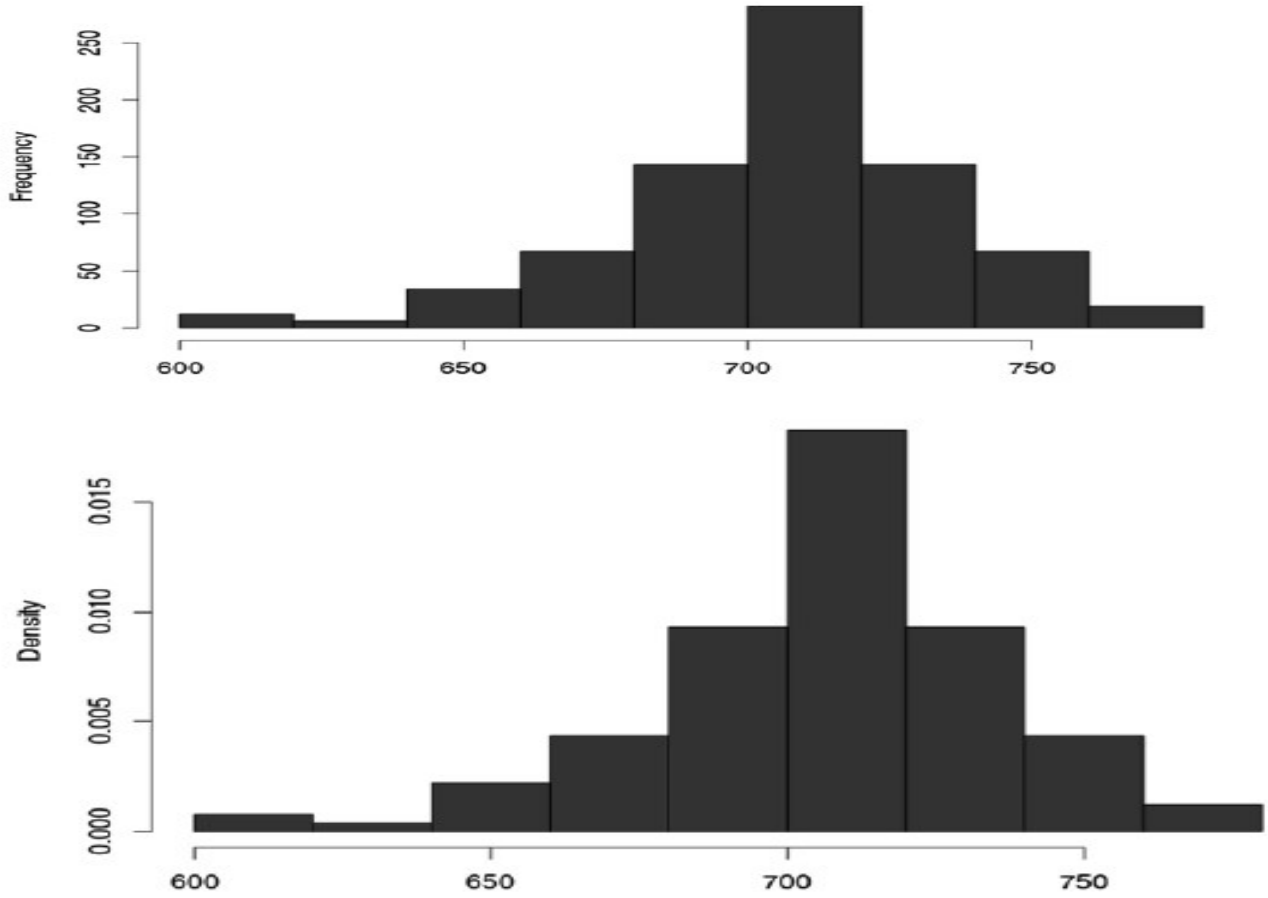
Graphical techniques:

1) Bar plot : plotting each point in bar shape



SLIDE-23

Histogram: Represents frequency distribution of data, how many observations of take the value within certain interval.



SLIDE-24

Third Business Moment: Skewness

4rth Business Moment: Kurtosis

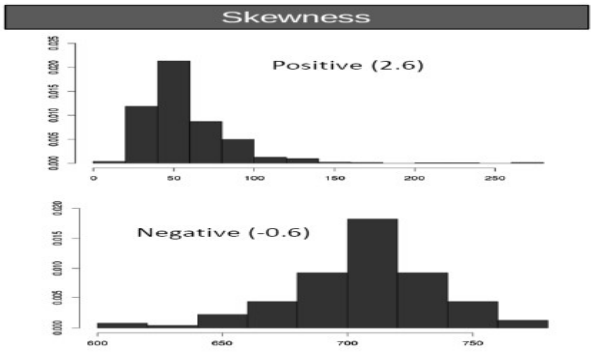
Skewness

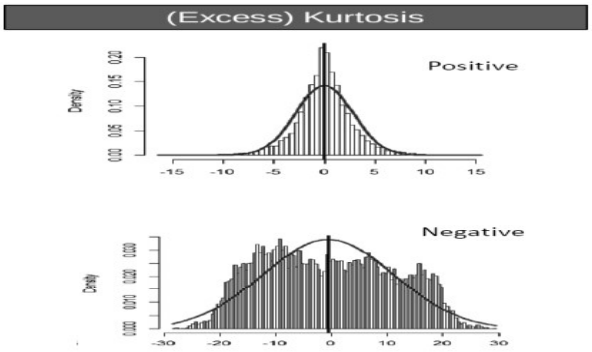
- A measure of asymmetry in the distribution
- Mathematically it is given by: $E\ [(x-\mu/\sigma)]^{3}$

- Negative skewness implies mass of the Distribution is concentrated on the Right

Kurtosis

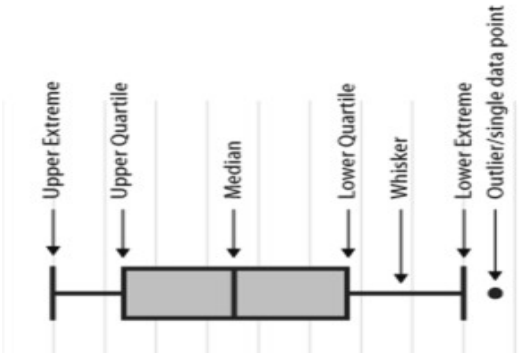
- A measure of the “Peakedness” of the distribution
- Mathematically it is given by $E[(x-\mu/\sigma)]^4 - 3$
- For Symmetric distributions, negative Kurtosis implies wider peak and thinner tails





SLIDE-25

Boxplot:



- Range (IQR): The middle half of a data set falls within the inter-quartile range. – Inter Quartile Range.

- **Box Plot:** This graph shows the distribution of data by dividing the data into four groups with the same number of data points in each group. The box contains the middle 50% of the data points and each of the two whiskers contain 25% of the data points. It displays two common measures of the variability or spread in a data set
- **Range:** It is represented on a box plot by the distance between the smallest value and the largest value, including any outliers. If you ignore outliers, the range is illustrated by the distance between the opposite ends of the whiskers

SLIDE-26

Normal Distribution

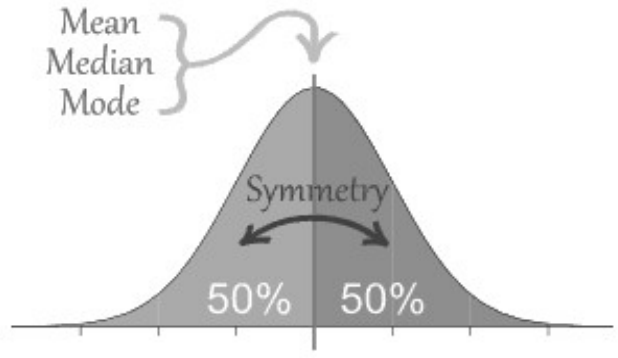
The normal random variable takes values from -∞ to +∞

The Probability associated with any single value of a random variable is always zero

Area under the entire curve is always equal to 1.

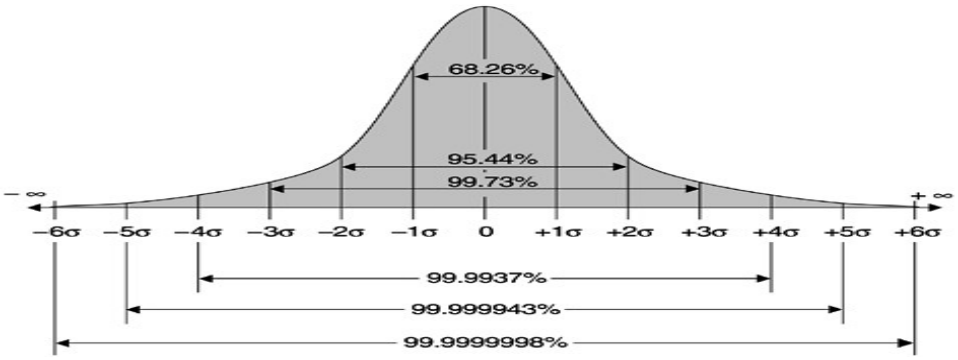
SLIDE-27

. Characterized by bell shaped



Properties:

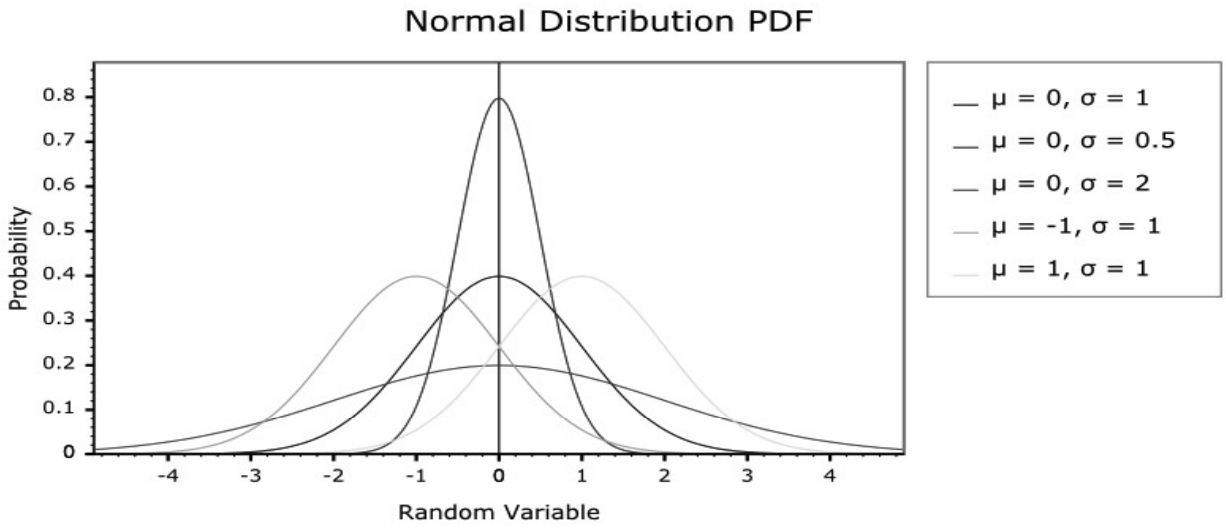
- 68.26% of values lie within ±1 σ from the mean
- 95.46% of the values lie within ±2 σ from the mean
- 99.73% of the values lie within ± 3σ from the mean



SLIDE-28

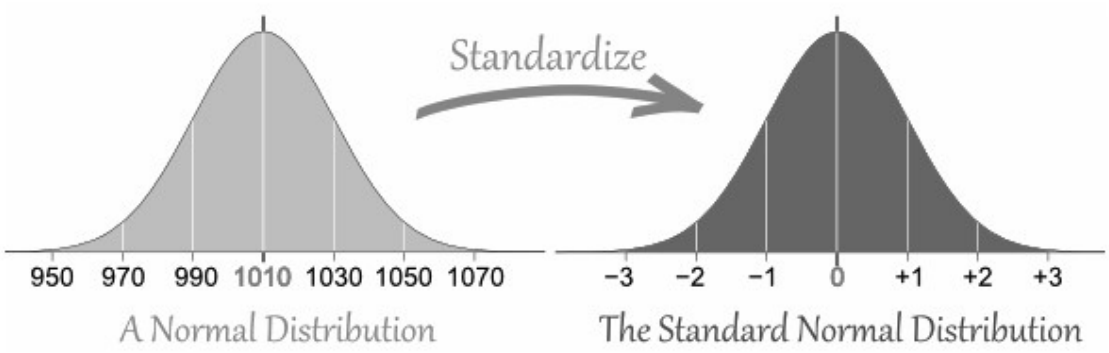
X~N(μ,σ)

Characterized by mean, μ, and standard deviation, σ



SLIDE-29

Z scores, Standard Normal Distribution:



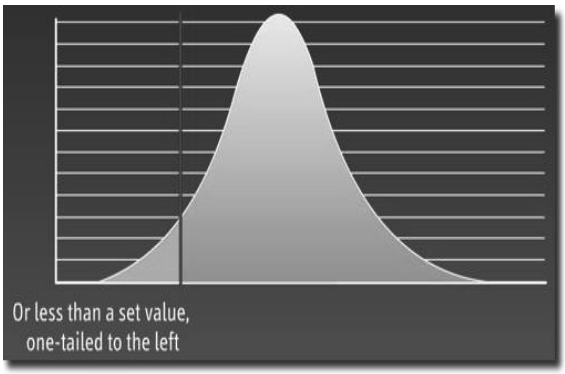
- For every value (x) of the random variable X, we can calculate Z score : $Z = (X-\mu)/\sigma$
- Interpretation − How many standard deviations away is the value from the mean?

SLIDE-30

Calculating Probability from Z distribution

Suppose GMAT scores can be reasonably modelled using a normal distribution

- $\mu = 711 \quad \sigma = 29$



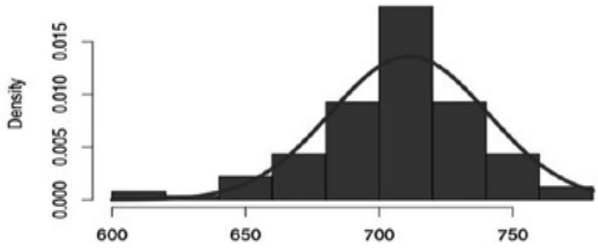
What is $p(x \leq 680)$?

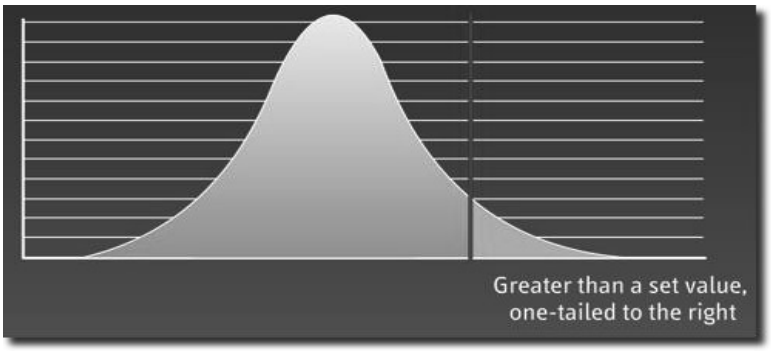
Step 1: Calculate Z score corresponding to 680

Z = (680-711)/29 = -1.06

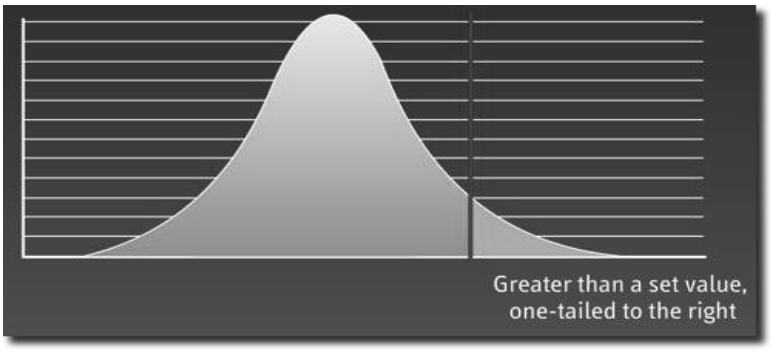
Step 2: Calculate the probabilities using Z – Tables

- P (Z ≤ -1) = 0.14





SLIDE-31

- What is P (697≤ X ≤ 740)?
- Step 1 : Use P(x1 ≤ X ≤ x2) = Use P(X ≤ x2) − P(X ≤ x1)
- 

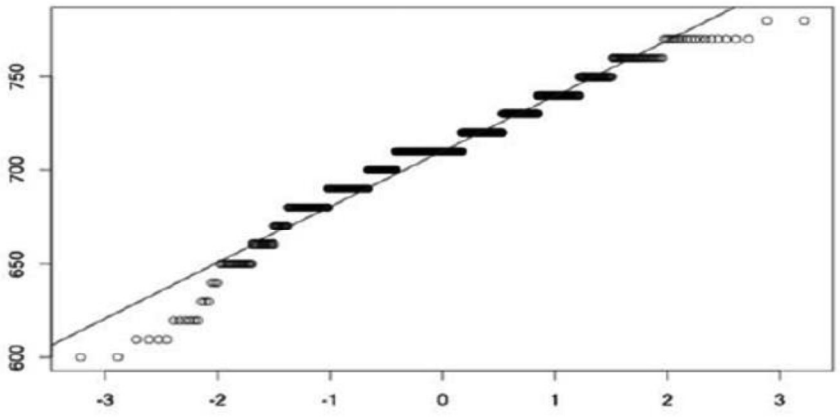
- Step 2 : Calculate P(X ≤ x2) and P(X ≤ x1) as before

$$P(X \leq 740) = P(Z \leq 1) = 0.84;\ P(X \leq 697) = P(Z \leq -0.5) = 0.31$$

- Step 3 : Calculate $P(697 \leq X \leq 740) = 0.84 - 0.31 = 0.53$

SLIDE-32

Normal Quantile plot (Q-Q plot):



To check whether the data is normally distributed

If plot is straight line (do not have to be absolute straight line) then we say data is normally distributed

If not then they are not normally distributed.

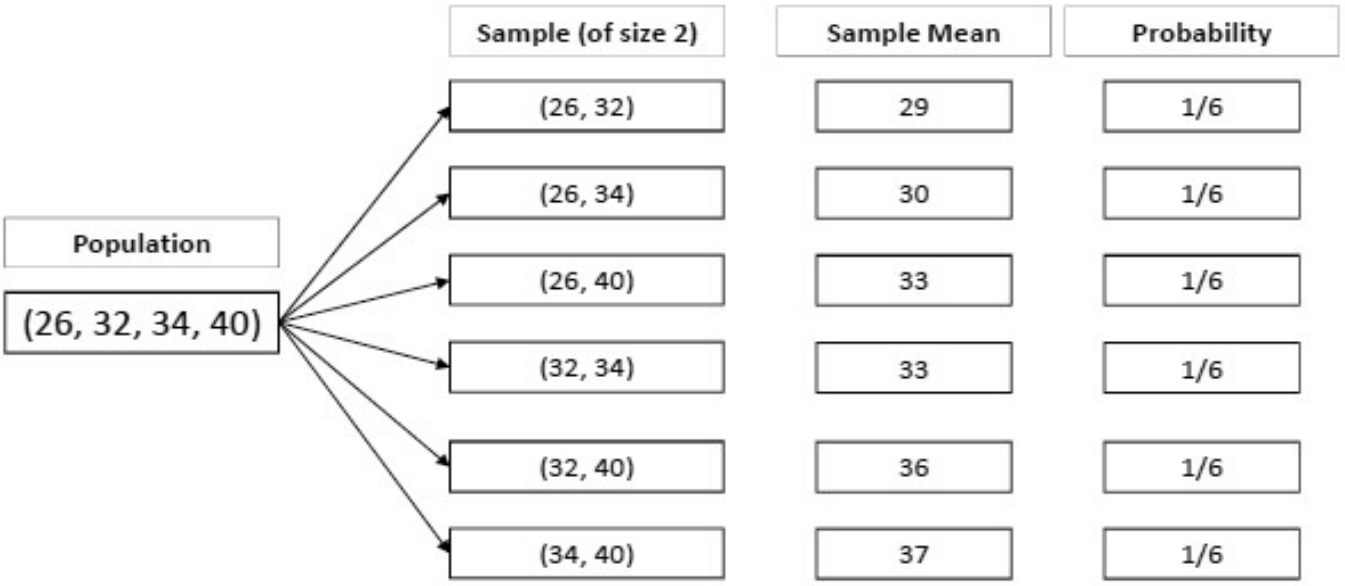
X-axis ->theoretical Quantiles

Y-axis ->Sample Quantiles

SLIDE-33

Sampling variation

- Sample mean varies from one sample to another.
- Sample mean can be (and most likely is) different from the population mean.
- Sample mean is a random variable.



SLIDE-34

Central Limit Theorem

The Distribution of the sample mean

- will be normal when the distribution of data in the population is normal

- will be approximately normal even if the distribution of data in the population is not normal if the “sample size” is fairly large

Mean (X) = μ (the same as the population mean of the raw data)

Standard Deviation (X) = $\sigma / \sqrt{n}$, where σ is the population standard deviation and n is the sample size

- This is referred to as standard error of mean.

The standard error of the mean estimates the variability between samples whereas the standard deviation measures the variability within a single sample.

SLIDE-35

Sample Size Calculation

A Sample Size of 30 is considered large enough, but that may /may not be adequate

More Precise conditions

- $n > 10(K_3)^2$, where (K_3) is sample skewness and
- $n > 10(K_4)$, where (K_4) is sample kurtosis

SLIDE-36

Confidence Interval

- What is the Probability of tomorrow's temperature being 42 degrees?
- Probability is '0'
- Can it be between [-50^{o}C & 100^{o}C]?

SLIDE-37

Case Study: Confidence Interval

- A University with 100,000 alumni is thinking of offering a new affinity credit card to its alumni.
- Profitability of the card depends on the average balance maintained by the card holders.
- A Market research campaign is launched, in which about 140 alumni accept the card in a pilot launch.

- Average balance maintained by these is $1990 and the standard deviation is $2833. Assume that the population standard deviation is $2500 from previous launches.
- What we can say about the average balance that will be held after a full−fledged market launch?

SLIDE-38

Interval estimates of parameters

- Based on sample data
  - − The point estimate for mean balance = $1990
  - − Can we trust this estimate?
- What do you think will happen if we took another random sample of 140 alumni?
- Because of this uncertainty, we prefer to provide the estimate as an interval (range) and associate a level of confidence with it
- Interval Estimate = Point Estimate ± Margin of Error

SLIDE-39

Confidence Interval for the Population Mean

Start by choosing a confidence level (1-α) % (e.g. 95%, 99%, 90%)

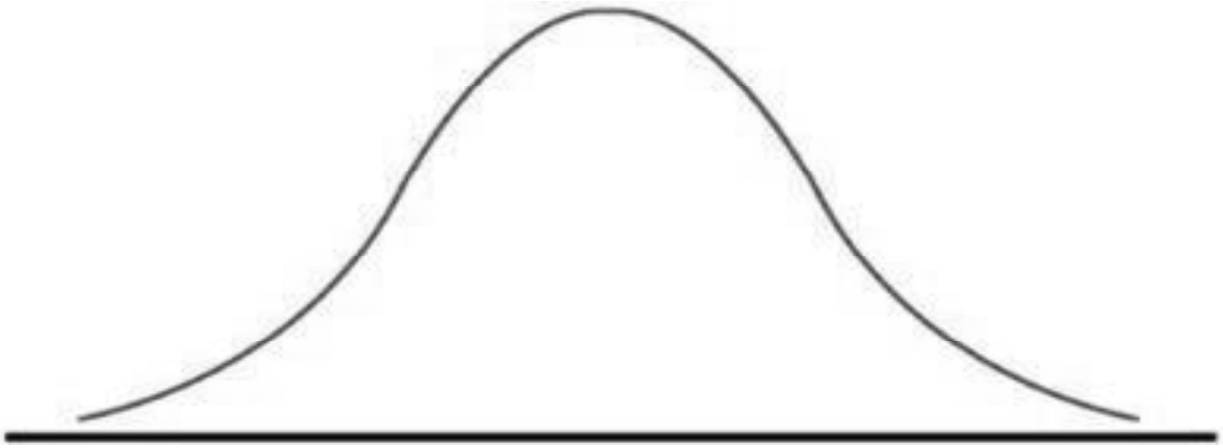
Then, the population mean will be with in

$X \pm Z_{1-\alpha}\, \sigma/\sqrt{n}$ where $Z_{1-\alpha}$ satisfies $p(-Z_{1-\alpha} \le Z \le Z_{1-\alpha}) = 1-\alpha$

Margin of error depends on the underlying uncertainty, confidence level and sample size.

SLIDE-40

Calculate Z value - 90%, 95% & 99%



SLIDE-41

Confidence Interval Calculation

- Based on the survey and past data
- – n = 140; σ = $2500; X = $ 1990

  $\sigma_{\overline{x}} = \sigma/\sqrt{n} = 2500/\sqrt{(140)} = 211.29$
- Construct a 95% confidence interval for the mean card balance and interpret it?
- Construct a 90% confidence interval for the mean card balance and interpret it?

SLIDE-42

Confidence Interval Interpretation

Consider the 95% Confidence interval for the mean income: [$1576, $2404]

Does this mean that?

- The mean balance of the population lies in the range?

- The mean balance is in this range 95% of the time?
- 95% of the alumni have balance in this range?

Interpretation 1 : Mean of the population has a 95% chance of being in this range for a random sample

Interpretation 2 : Mean of the population will be in this range for 95% of the random samples

SLIDE-43

What if we don't know Sigma?

- Suppose that the alumni of this university are very different and hence population standard deviation from previous launches cannot be used

  We replace $\boldsymbol{\sigma}$ with our best guess (point estimate) **s**, which is the standard deviation of the sample:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

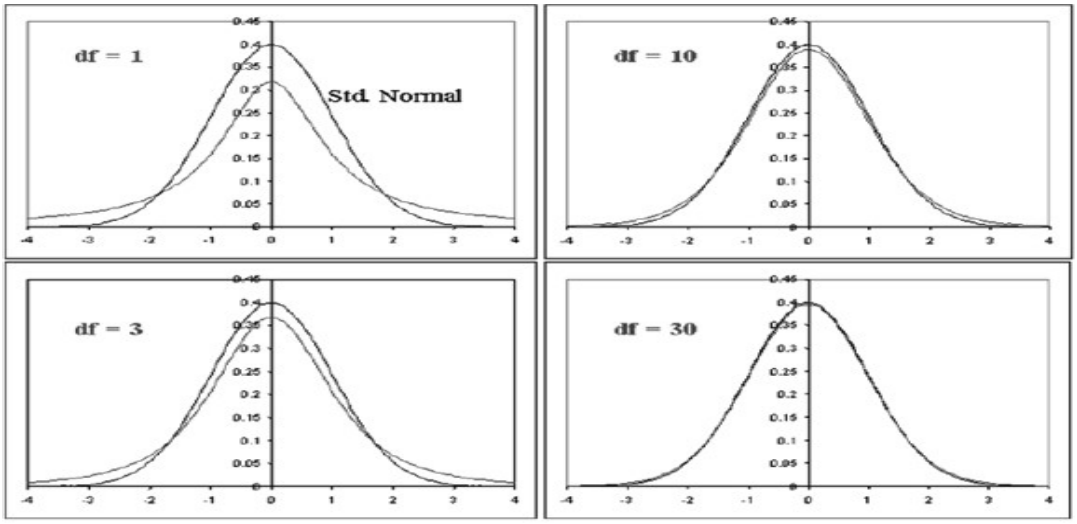
  Calculate:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- If the underlying population is normally distributed , T is a random variable distributed according to a t-distribution with n-1 degrees of freedom T_{n-1}
- Research has shown that the t-distribution is fairly robust to deviation of the population of the normal model

SLIDE-44

Student's t-distribution



As n -> ∞

t_n -> N(0,1)

i.e., as the degrees of the freedom increase, the t-distribution approaches the standard normal distribution.

Slide-45

Confidence Interval for mean with unknown Sigma

$\overline{x} \pm Z_{1-\alpha}\, \sigma/\sqrt{n}$ where $Z_{1-\alpha}$ satisfies $p(-Z_{1-\alpha} \le Z \le Z_{1-\alpha}) = 1-\alpha$

Instead of above equation we can use the below t distribution equation

$\overline{x} \pm t_{1-\alpha,\, n-1}\, s/\sqrt{n}$ where $t_{1-\alpha,\, n-1}$ satisfies $p(-t_{1-\alpha,\, n-1} \le T_{n-1} \le t_{1-\alpha,\, n-1}) = 1-\alpha$

Slide-46

Calculating t-value

- Construct a 95% confidence interval for the mean card balance and interpret it?

  n = 140; σ = $2500; $\overline{x}$ = $ 1990

  $\sigma_{\overline{x}}$ = 2833/sqrt(140) = 239.46

  Calculate $t_{0.95,\ 139}$ = 1.98

  Then the 95% confidence interval for balance is [$1516, $2464]