Data Analysis Course

Probability distributions(version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

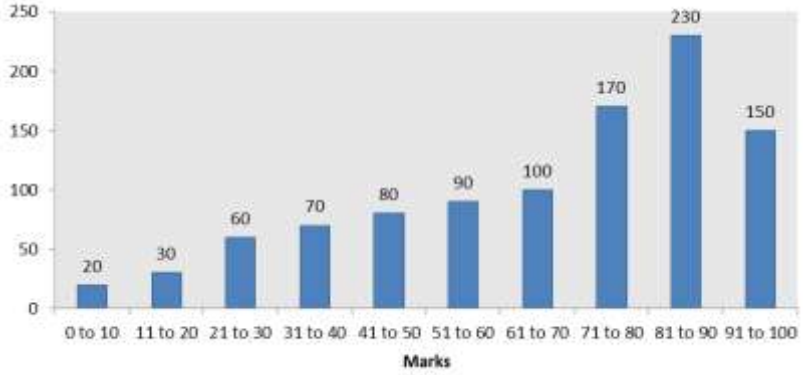
Contents

- What is probability distribution?
- Normal distribution
- Binomial distribution
- Sampling distributions

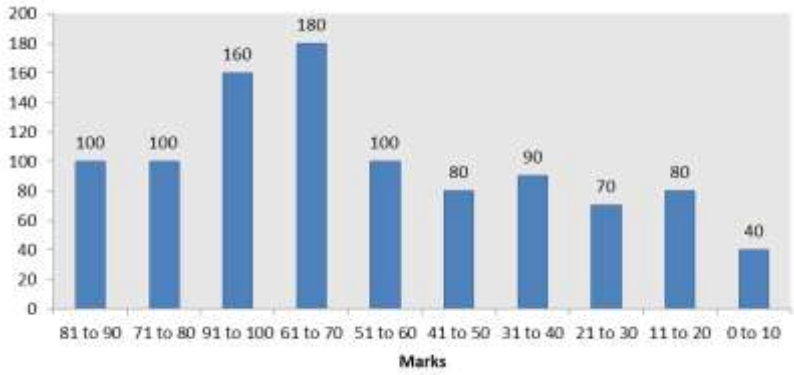
Distribution

Results of a mathematics test marks from last 10 years

School A



School B



- Which school is better?
- If we are going to conduct another test, how many students can be expected to score 91 to 100 from school A & from school B
- This is a frequency distribution of marks.
- What is the probability that a student will score more than 50 from school A & from School B?

What is the need of Probability Distributions?

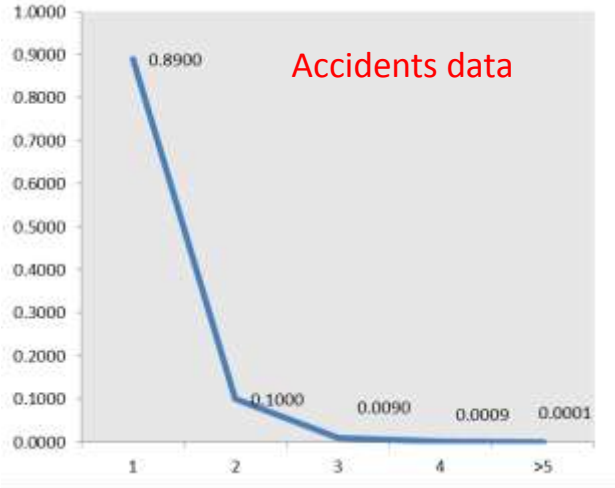
What is the probability that a person is alive after **4** bus accidents?

- =0.5
- >0.5
- <0.01
- <0.001

Toss a coin, what is the probability of heads?

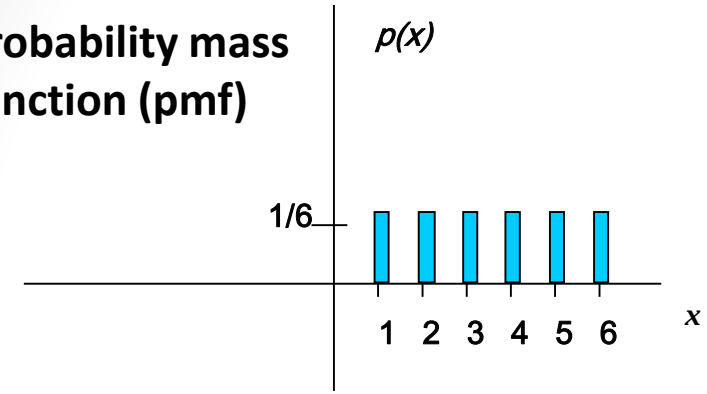
- =0.5
- > 0.5
- < 0.5
- Cant tell

- Did you do any calculation in above two examples? How can we tell the probability without calculating? Because we know their distributions
- If a variable follows a distribution we can find the probability without any experiment



Roll of a die

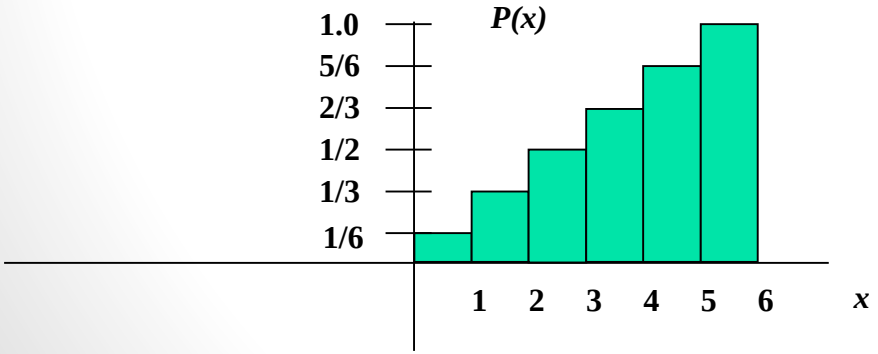
Probability mass function (pmf)



x	p(x)
1	p(x=1)=1/6
2	p(x=2)=1/6
3	p(x=3)=1/6
4	p(x=4)=1/6
5	p(x=5)=1/6
6	p(x=6)=1/6
	1.0

$$\sum_{\text{all } x} P(x)=1$$

Cumulative distribution function (CDF)



x	P(x≤A)
1	P(x≤1)=1/6
2	P(x≤2)=2/6
3	P(x≤3)=3/6
4	P(x≤4)=4/6
5	P(x≤5)=5/6
6	P(x≤6)=6/6

$$0 \le P(y) \le 1, \quad \Sigma P(y) = 1$$

Probability function

f(x)=0.5 for x= 'Heads'. 'Tails' in coin tossing

f(x)= 1/6 for x=1,2,3,4,5,6 for roll of dice

f(x)=0.2 for x=1,2,3,4,5 or more for accidents example?

Which of the following are probability functions?

a. *f(x)=.25* for x=9,10,11,12

b. *f(x)= (3-x)/2* for x=1,2,3,4

c. *f(x)= (x^2+x+1)/25* for x=0,1,2,3

Binomial Distribution

Binomial Distribution

- Suppose we flip a coin 2 times **H H HT T H T T**
- Sample space shows 4 possible outcomes or sequences. Each sequence is a permutation. Order matters.
- There are 2 ways to get a total of one heads (HT and TH). These are combinations. Order does NOT matter. **HH**, **HT**, **TH**, **TT**
- Suppose our interest is Heads. If the coin is fair, p(Heads) = .5; q = 1-p = .5.
- The probability of any permutation for 2 trials is ¼ = p*p, or p*q, or q*p, or q*q. All permutations are equally probable.

- The probability of 1 head in any order is 2/4 = .5 = **HT+TH/(HH+HT+TH+TT)**

Coin example - more flips

- 3 flips
  - HHH,
  - HHT,HTH,THH
  - HTT, THT, TTH
  - TTT
- All permutations equally likely = p*p*p = $(1/2)^3$ = 1/8.
- p(0 tail) = 1/8
- p(1 tail) = 3/8
- P(two tails) =??
- P(three tails) =??

Coin example

- 3 flips for count of number of tails
  - HHH, - Zero out of 3 (3**C**0?)
  - HHT,HTH,THH - one out of 3 (3**C**1?)
  - HTT, THT, TTH - two out of 3 (3**C2**?)
  - TTT - three out of 3 (3**C**3?)
- All permutations equally likely = p*p*p = $.5^3$ = .125 = 1/8.
- p(1 tail) = 3/8 =(3c1)(1/2)(1/4)
- P(two tails) =(3c1)(1/2)(1/4)
- P(three tails) =??

Binomial Distribution

- Black /White choose a color
- Out of 4 students , what is the probability that
  - 0 choose Black & 4 choose White
    - WWWW - Zero out of 4 (4**C**0?)
  - 1 choose Black & 3 choose White
    - BWWW, WBWW, WWBW, WWWB - - One out of 4 (4**C**1?)
  - 2 choose Black & 2 choose White
  - 3 choose Black & 1 choose White
  - 4 choose Black & 0 choose White

Which graph best describes the behavior of this count variable?

Binomial distribution function

$$P(x) = C_x^n p^x q^{n-x}$$

$$C_x^n = \frac{n!}{x!(n-x)!}$$

$$\mu = np$$

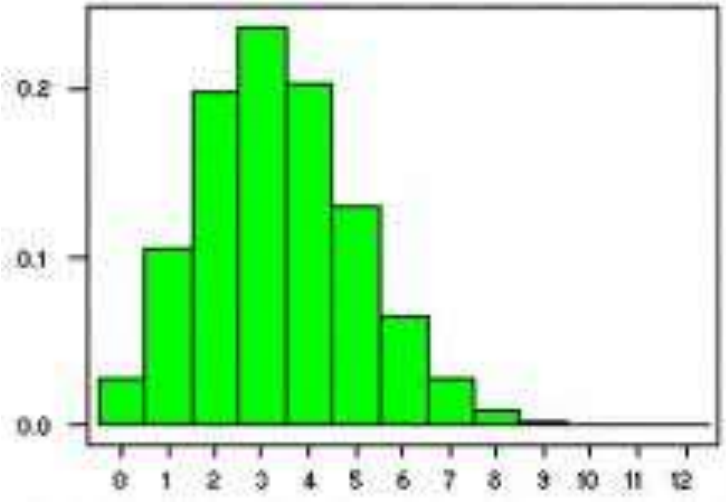
$$\sigma^2 = npq$$

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

Binomial distribution -Properties

The binomial distribution describes the behavior of a count variable X if the following conditions apply:

- The number of observations n is fixed.
- Each observation is independent.
- Each observation represents one of two outcomes ("success" or "failure").
- The probability of "success" p is the same for each outcome.



Mean of a distribution

- On an average, how many people will chose black?
- Like frequency distributions, probability distributions have descriptive measures, such as mean and standard deviation

$$\mu = E(Y) = \sum yP(y)$$

Calculate the mean for color example

Lab

- One player stand in the foul line to shoot free-throws 10 times. Suppose the probability that he makes it is 0.5
- Does this meet the criteria of a binomial distribution?
- What is the mean and variance?
- What is the probability that he get 6 out of 10? What is the mean and variance?
- What is the probability that he get 8 out of 10?
- What is the probability that he get 10 out of 10?

Normal distribution

Why are normal distributions so important?

- The normal distribution is one of the most important distributions in statistics.
  - Many dependent variables are commonly assumed to be normally distributed in the population
  - If a variable is approximately normally distributed we can make inferences about values of that variable
- Many measured quantities in the natural sciences follow a normal distribution.
- Example: Sampling distribution of the mean

Normal Distribution

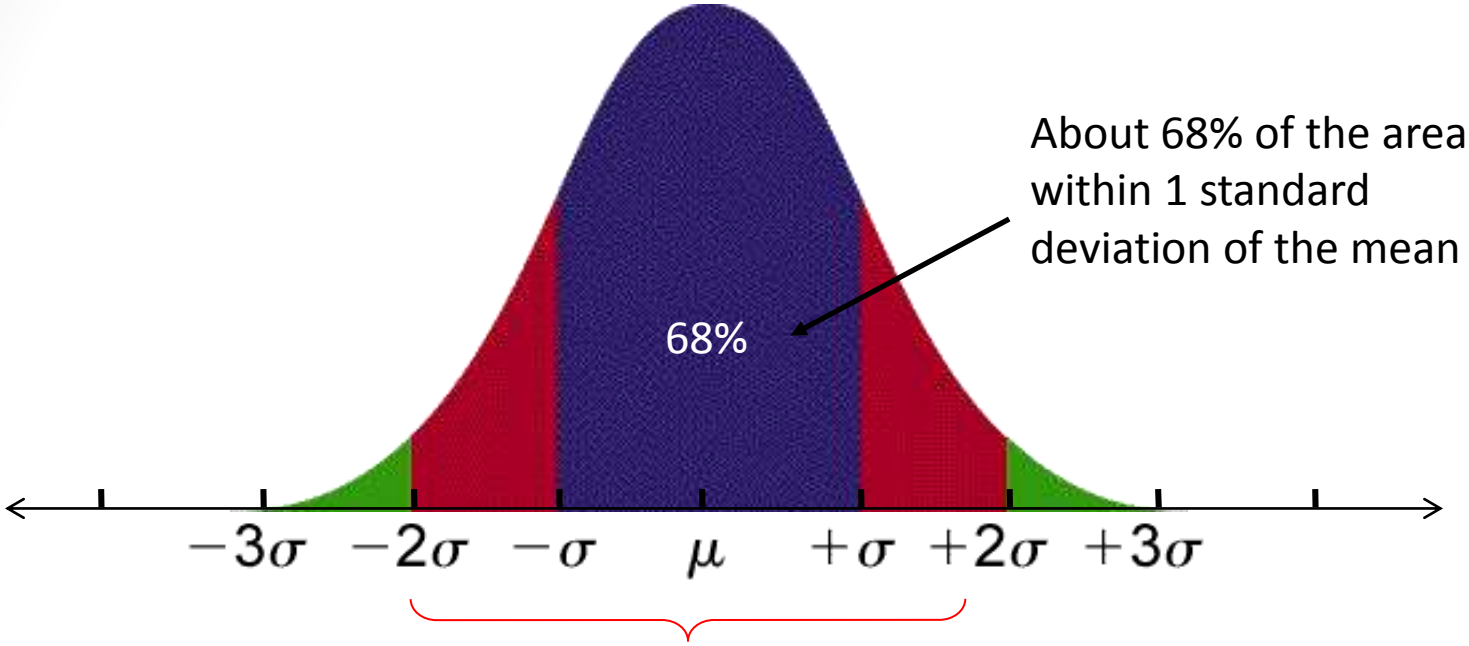
- Symmetrical, bell-shaped curve
- Also known as Gaussian distribution
- Point of inflection = 1 standard deviation from mean
- Mathematical formula

$$f(X) = \frac{1}{\sigma \cdot \sqrt{2\pi}} (e)^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Properties of Normal Distribution

- The mean, median, and mode are equal
- Bell shaped and is symmetric about the mean
- The total area that lies under the curve is one or 100%
- As the curve extends farther and farther away from the mean, it gets closer and closer to the x-axis but never touches it.

Empirical Rule



About 95% of the area lies within 2 standard deviations

About 99.7% of the area lies within 3 standard deviations of the mean

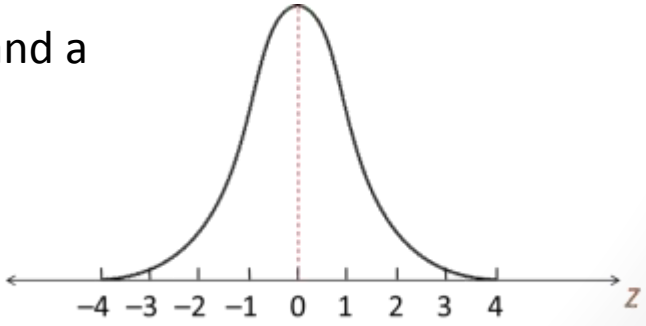
The Standard Normal Distribution

- Using the normal probability distribution function, calculate the probity of X >26 when mean is 20 and standard deviation is 6

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

  - Using the normal probability distribution function, calculate the probity of X >1 when mean is 0 and standard deviation is 1 –Use previous slide
  - Is it same as Z> (26-20/6)

- The standard normal distribution has a mean of 0 and a standard deviation of 1.
- Using z-scores any normal distribution can be transformed into the standard normal distribution.



The Standard Score

The **standard score**, or ***z*-score**, represents the number of standard deviations a random variable *x* falls from the mean.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

The test scores for a civil service exam are normally distributed with a mean of 152 and a standard deviation of 7. Find the standard *z*-score for a person with a score of:

(a) 161 **(b) 148** **(c) 152**

(a) $z = \frac{161 - 152}{7}$

$z = 1.29$

(b) $z = \frac{148 - 152}{7}$

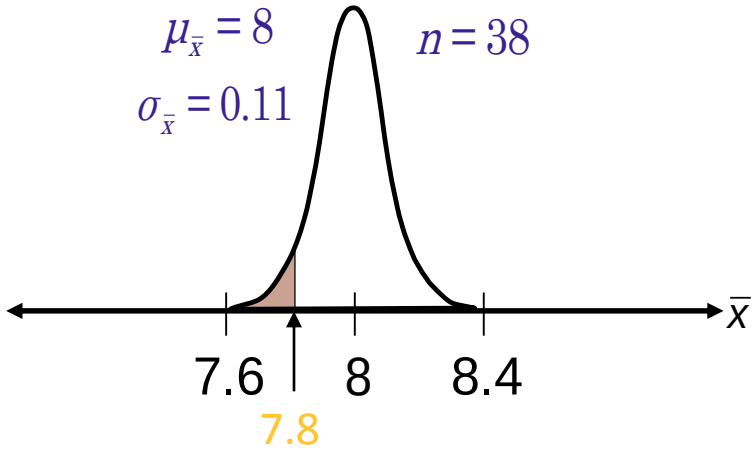
$z = -0.57$

(c) $z = \frac{152 - 152}{7}$

$z = 0$

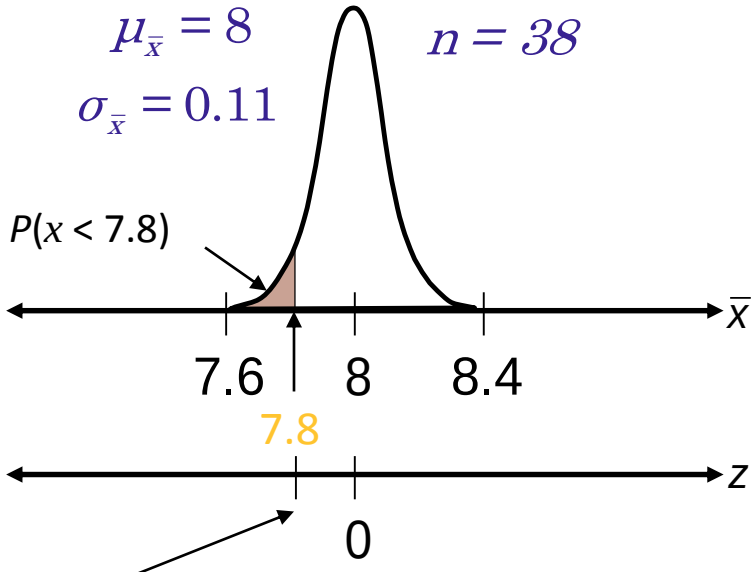
Example-1

- The heights of fully grown magnolia bushes have a mean height of 8 feet and a standard deviation of 0.7 feet. 38 bushes are randomly selected from the population, and the mean of each sample is determined.
- The mean of the sampling distribution is 8 feet, and the standard error of the sampling distribution is 0.11 feet.
- Find the probability that the mean height of the 38 bushes is less than 7.8 feet.



Finding Probabilities

Example continued:
Find the probability that the mean height of the 38 bushes is less than 7.8 feet.



$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$= \frac{7.8 - 8}{0.11}$$

$$= -1.82$$

$P(x < 7.8) = P(z < -1.82) = 0.0344$

The probability that the mean height of the 38 bushes is less than 7.8 feet is 0.0344.

Example2

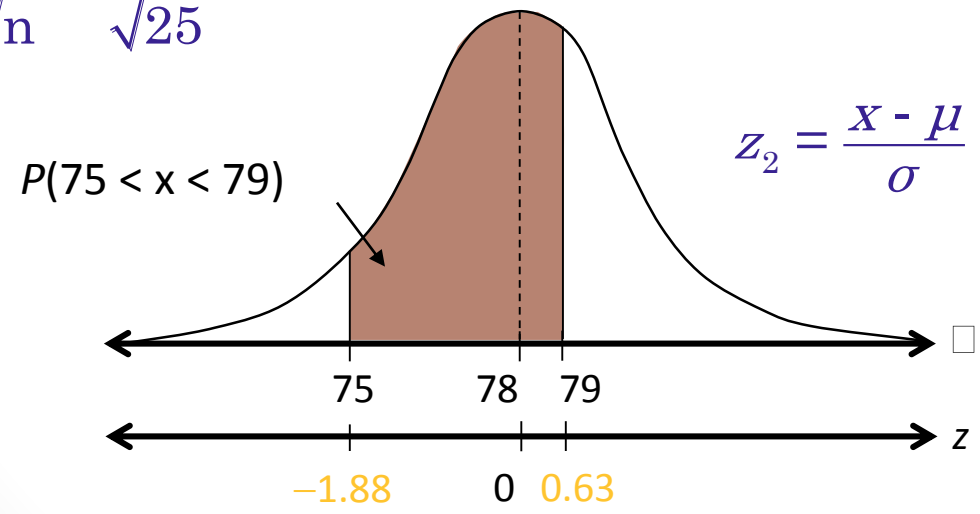
- The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that the mean score of 25 randomly selected students is between 75 and 79.

$$\mu_{\bar{x}} = 78$$

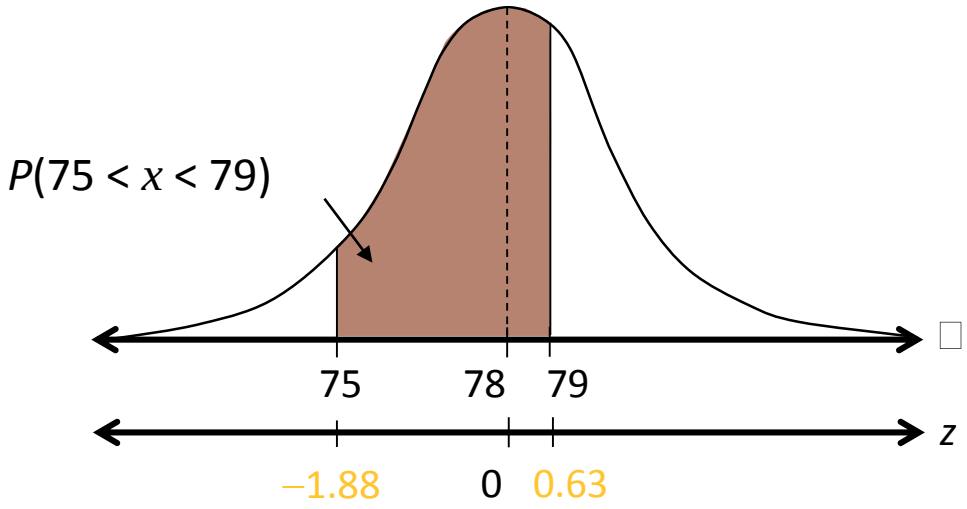
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{25}} = 1.6$$

$$z_1 = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{75 - 78}{1.6} = -1.88$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{79 - 78}{1.6} = 0.63$$



Finding the probability

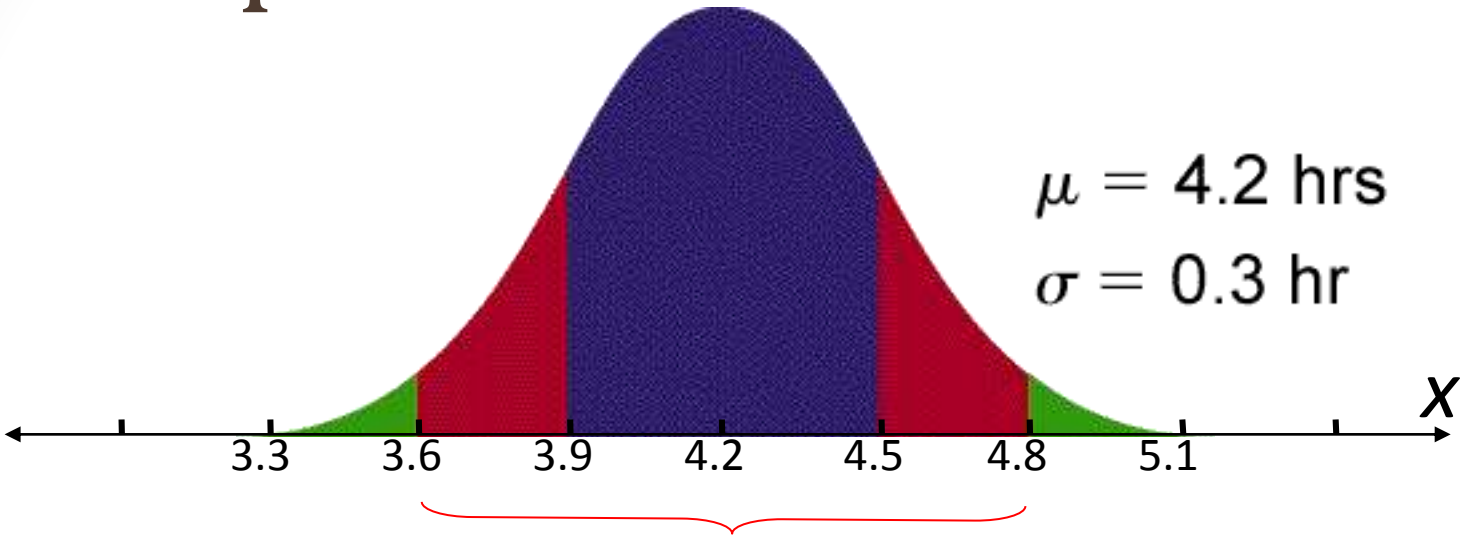


$P(75 < x < 79) = P(-1.88 < z < 0.63) = P(z < 0.63) - P(z < -1.88)$

$= 0.7357 - 0.0301 = 0.7056$

Approximately 70.56% of the 25 students will have a mean score between 75 and 79.

Example 3



An instruction manual claims that the assembly time for a product is normally distributed with a mean of 4.2 hours and standard deviation 0.3 hour. Determine the interval in which 95% of the assembly times fall.

95% of the data will fall within 2 standard deviations of the mean.

4.2 – 2 (0.3) = 3.6 and 4.2 + 2 (0.3) = 4.8.

95% of the assembly times will be between 3.6 and 4.8 hrs.

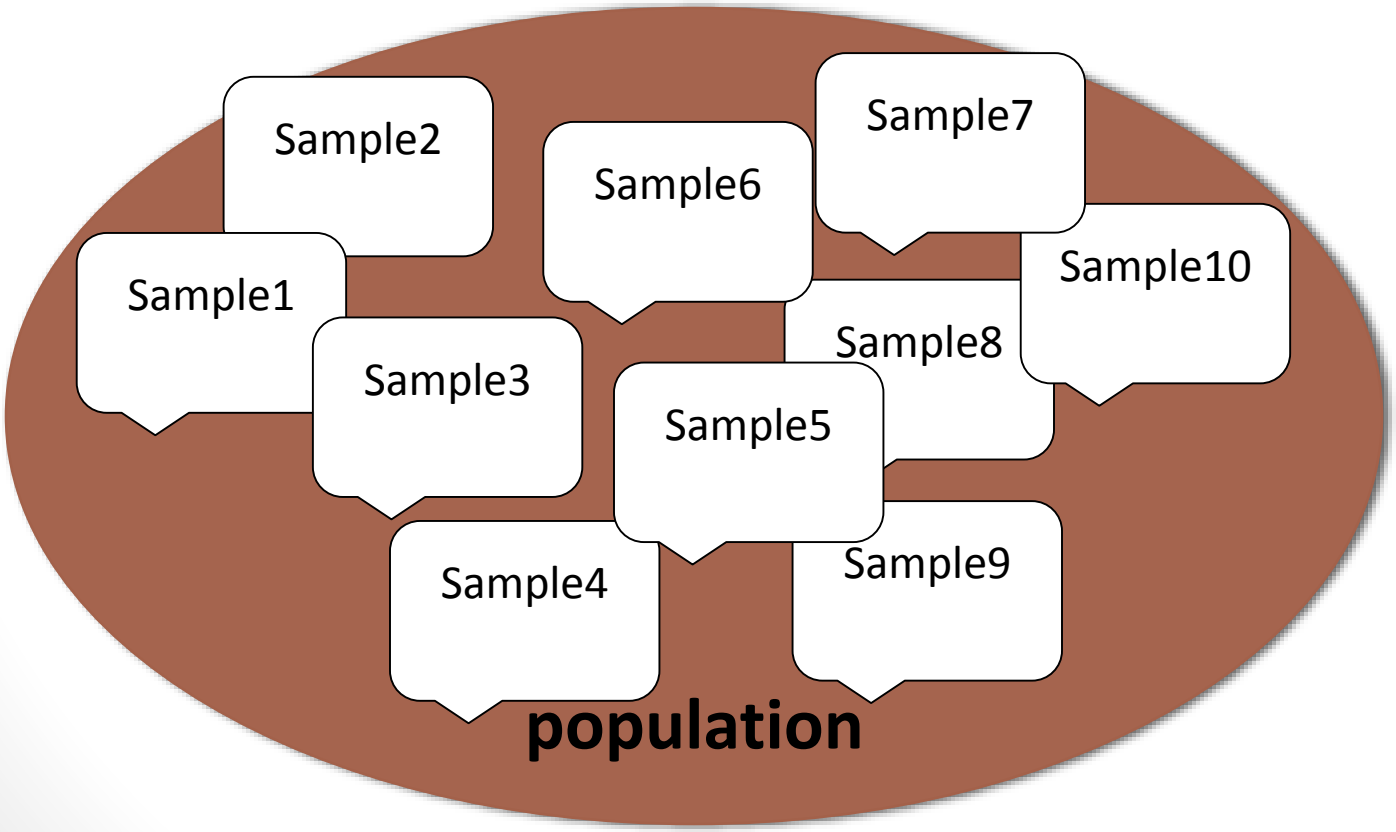
Lab

- If birth weights in a population are normally distributed with a mean of 3 kg and a standard deviation of 0.3 kg
  - What is the chance of obtaining a birth weight of 5 kg or heavier when sampling birth records at random?
  - What is the chance of obtaining a birth weight of 2kg or lighter?
  - What is the chance of obtaining a birth weight of 10 kg or heavier?
  - What is the chance of obtaining a birth weight of 1kg or lighter?
- In the instruction manual example(example-3) what is the probability that the assembly time is
  - More than 4 hours?
  - More than 6 hours?
  - Less than 3 hours?
  - Less than 4.2 hours?

Central Limit Theorem

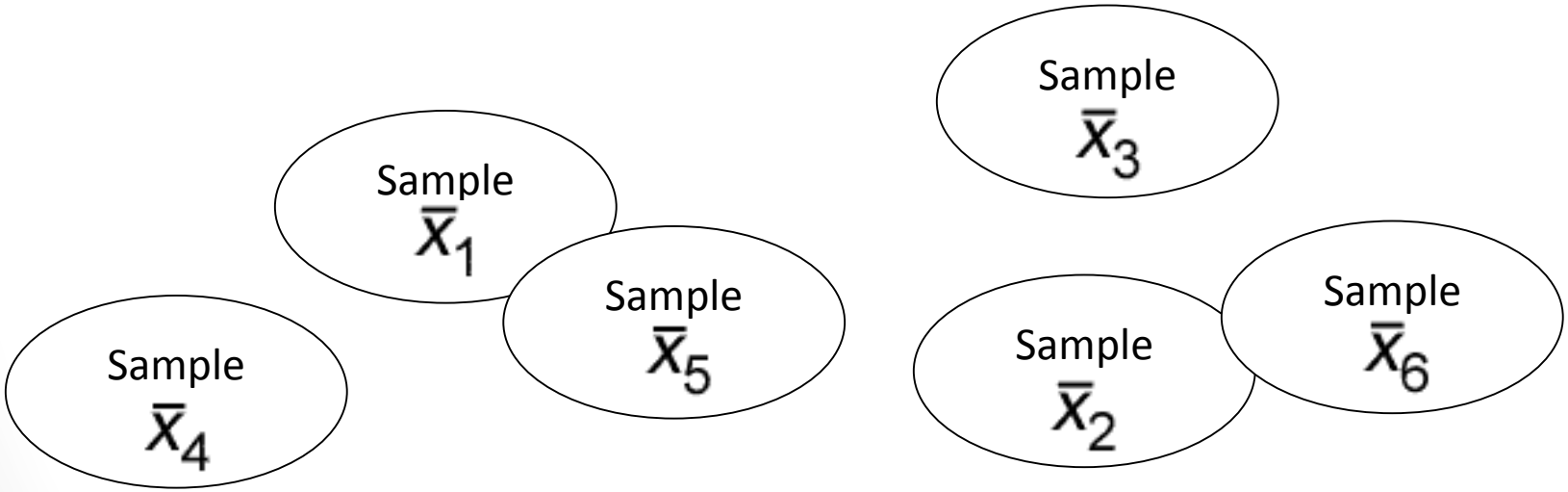
Sampling Distributions

- A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population.



Sampling Distributions

- A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population.
- If the sample statistic is the sample mean, then the distribution is the sampling distribution of sample means.

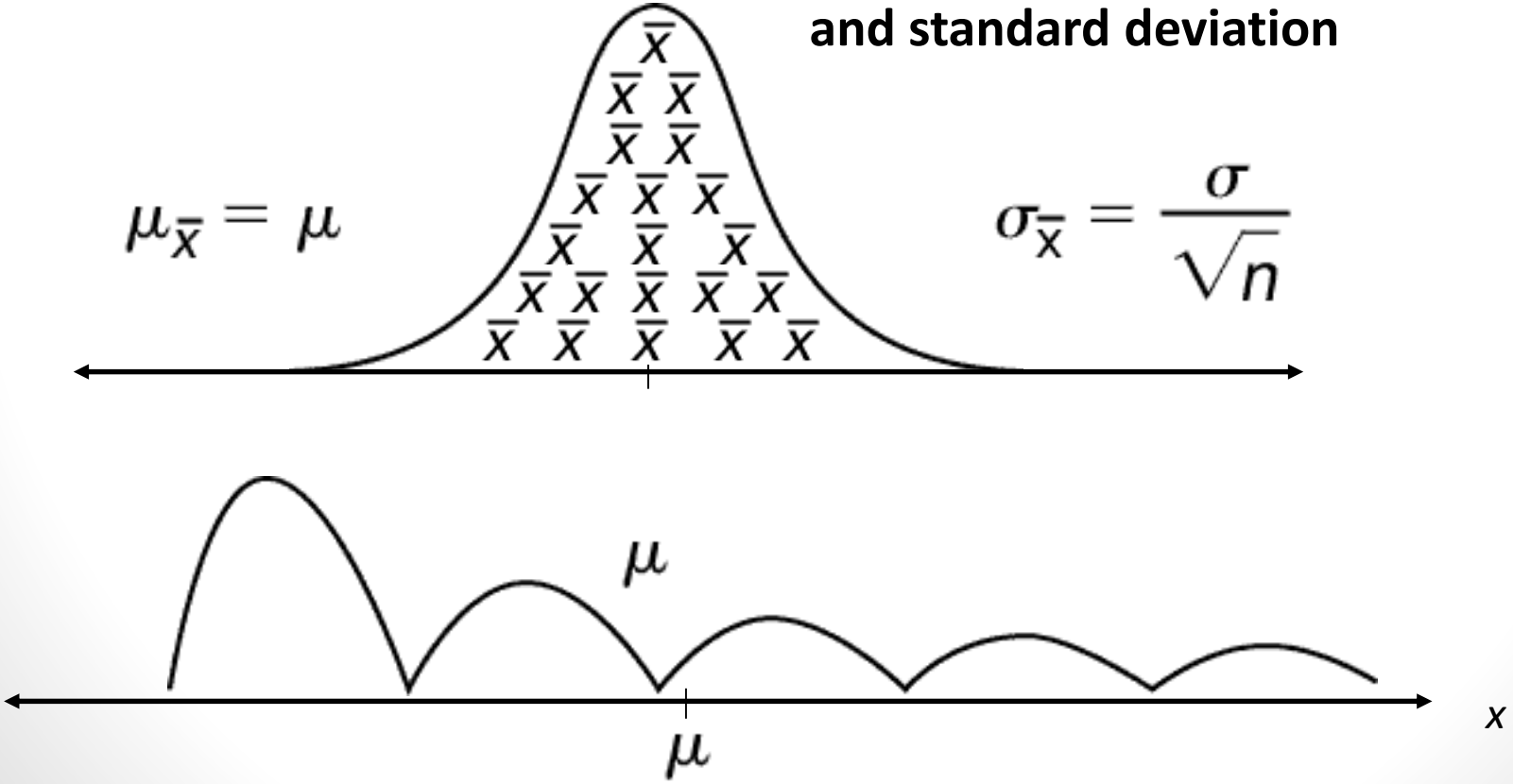


The sampling distribution consists of the values of the sample means, $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5, \bar{x}_6, \ldots$

Central Limit theorem

If a sample *n* (30) is taken from a population with ***any type distribution*** that has a mean = μ and standard deviation = σ

the ***sample means*** will have a **normal distribution and standard deviation**



Lab

- Open excel
- Fill column a with random numbers (use ran between function)
- In column B, find the mean of first 30 observations from A and drag it
- Draw a histogram of B.

Application Central Limit Theorem

- During a certain week the mean price of gasoline in California was $1.164 per gallon. What is the probability that the mean price for the sample of 38 gas stations in California is between $1.169 and $1.179? Assume the standard deviation = $0.049.

mean $\mu_{\bar{x}} = \mu_{\bar{x}} = 1.164$

standard deviation $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.049}{\sqrt{38}} = 0.0079$

Calculate the standard *z*-score for sample values of $1.169 and $1.179.

$$z = \frac{1.169 - 1.164}{0.0079} = 0.63 \qquad z = \frac{1.179 - 1.164}{0.0079} = 1.90$$

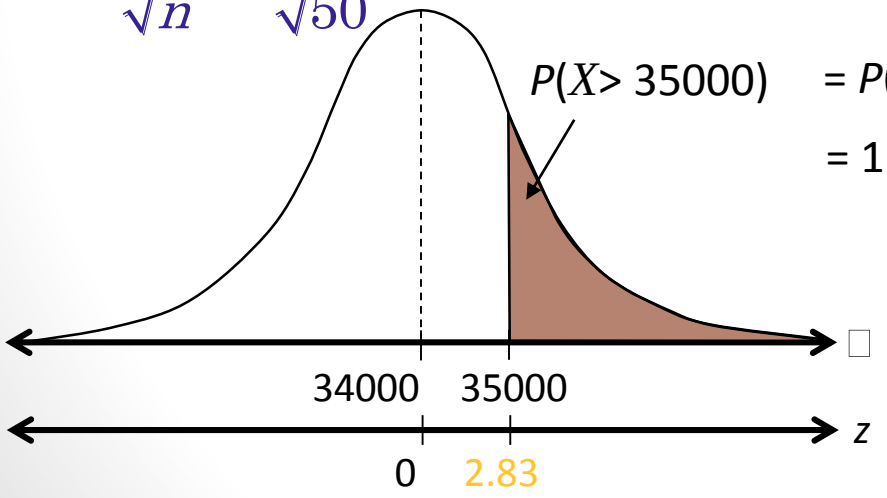
Probabilities of x and $\bar{x}$

The population mean salary for auto mechanics is $34,000 with a standard deviation of $2,500. Find the probability that the mean salary for a randomly selected sample of 50 mechanics is greater than $35,000.

$$\mu_{\bar{x}} = 34000$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2500}{\sqrt{50}} = 353.55$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{35000 - 34000}{353.55} = 2.83$$



$P(X > 35000) = P(z > 2.83) = 1 - P(z < 2.83)$

$= 1 - 0.9977 = 0.0023$

The probability that the mean salary for a randomly selected sample of 50 mechanics is greater than $35,000 is 0.0023.

Probabilities of x and $\bar{x}$

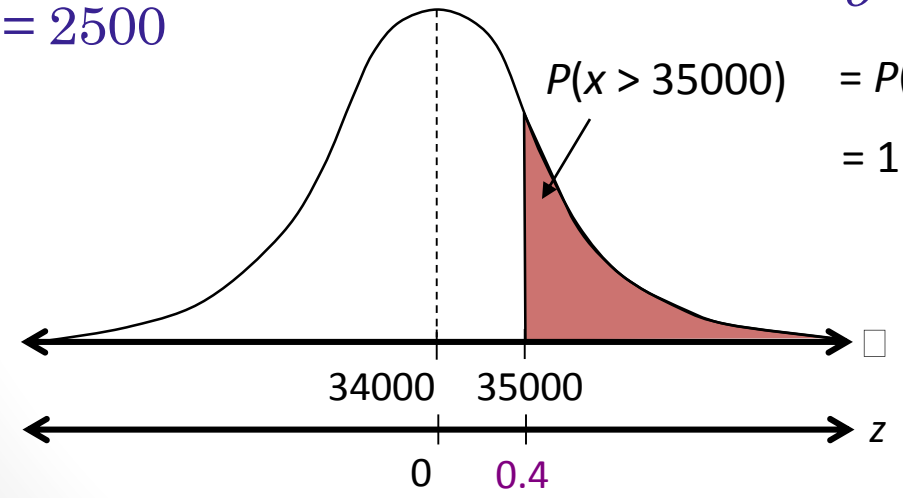
The population mean salary for auto mechanics is μ = \$34,000 with a standard deviation of σ = \$2,500. Find the probability that the salary for <u>one</u> randomly selected mechanic is greater than \$35,000.

(Notice that the Central Limit Theorem does not apply.)

$\mu = 34000$

$\sigma = 2500$

$$z = \frac{x - \mu}{\sigma} = \frac{35000 - 34000}{2500} = 0.4$$



$P(x > 35000) = P(z > 0.4) = 1 - P(z < 0.4)$

$= 1 - 0.6554 = 0.3446$

The probability that the salary for one mechanic is greater than \$35,000 is 0.3446.

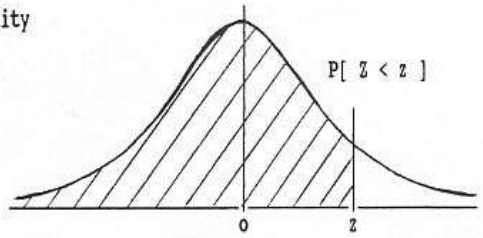
Looking up probabilities in the standard normal table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}Z^2)\ dZ$



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	0.5159	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	0.7854
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	0.8804	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

What is the area to the left of Z=1.51 in a standard normal curve?

Z=1.51

Area is 93.45%

Z=1.51

Normal probabilities in SAS

```
data _null_;
  theArea=probnorm(1.5);
  put theArea;
run;
0.9331927987
```

The "probnorm(Z)" function gives you the probability from negative infinity to Z (here 1.5) in a standard normal curve.

And if you wanted to go the other direction (i.e., from the area to the Z score (called the so-called "Probit" function →

```
data _null_;
  theZValue=probit(.93);
  put theZValue;
run;
1.4757910282
```

The "probit(p)" function gives you the Z-value that corresponds to a left-tail area of p (here .93) from a standard normal curve. The probit function is also known as the inverse standard normal function.

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879