**CH5020 Assignment 1**

**Mayur Vikas Joshi (ME16B148)**

**Wadavkar Sushant Uttam (ME16B172)**

**Instructions:**

1. Do in groups of 2
2. Submit as word document (Arial font, 11 font size, 1.5 line spacing)
3. Give Matlab code wherever appropriate
4. PLEASE do discuss within and outside your group if required but provide original answers
5. Due Date:  21st September 2020

**1.** A probability density function is given such that **[7]**

f(x) = Ax+b 0≤x≤1

= 0, elsewhere

P(0.25 < X < 0.5)=19/80

Find the constants A and B as well as the mean value.

**Answer:**

A = 2/5; b = 4/5

Mean = 0.533

MATLAB script:

syms A b x

pdf = A\*x + b

equation\_1 = int(pdf, x,0,1)

equation\_2 = int(pdf, x,0.25,0.5)

equations = [ equation\_1 == 1, equation2 == 19/80 ]

vars = [ A b ]

[Aval bval] = solve(equations,vars)

mean = vpa(subs(int(x\*pdf,x,0,1),[A b],[Aval bval]),3)

**2.** Experiments are subject to measurement errors. Suppose the error is treated as a random variable X whose continuous probability density function is given by

f(x) = q(3-x2) -1≤x≤1

= 0, elsewhere

**a.** What are the minimum & maximum values of the error? **b.** Find q such that f(x) is valid. **c.** What is the probability that the random error in a measurement is less than 0.5? **d.** If the magnitude of the error does NOT exceed 0.8, what is the probability of this outcome? **[8]**

**Answer:**

1. Maximum value of error = 1

Minimum value of error = 0 (By magnitude)

1. q = 3/16
2. 0.773
3. 0.836

MATLAB Script:

syms q x

pdf = q\*(3-x^2)

qval = solve(int(pdf,x,-1,1)==1, q)

pdf = subs(pdf,q,qval)

B = vpa(int(pdf,x,-1,0.5),3)

C = vpa(int(pdf,x,-0.8,0.8),3)

**3.**  If X is a random variable described by a continuous probability density function

                                      fx=x23             -1<x<2    and 0 elsewhere

Find the expected value of g(X) = 12X+5.                               **[4]**

**Ans:**

By definition, we have E[g(X)] =

After operation, we get E[g(X)] = 20

**4.**

**a.** In the age of fast computers, modeling and simulation encompassing different length and timescales, why is it still necessary to do experiments?

**b.** Why do repeats of experiments carried out even with state-of-the-art equipment and instruments give different results each time?

1. Experiments can provide us with new empirical data, computer simulations cannot. While it is true that computer simulations deliver results that may not have been known or expected by us beforehand, computer simulations can by their very nature only produce results that are implied by the premises on which the simulation is built. Another important difference is that some experiments operate directly on the target system, while computer simulations never do. Finally, experiments can be used for the testing of fundamental hypotheses, which, again, computer simulations cannot.
2. Replication shows patterns and trends in your results. This helps maintain integrity of data. If someone is to thoroughly review your work, then they would carry out the experiments again themselves. If someone wants to replicate an experiment, the first scientist should do everything possible to allow replicability. Being able to replicate experiments and the resulting data allows to check the extraneous variables. These are variables that you are not actually testing, but that may be influencing the results.

**[2+2]**

**5.** A government agency advertises in the newspaper inviting bids for some of its projects. It generally estimates what a reasonable bid would be and calls it as ‘b’. From experience, it has determined the probability density function of the winning (lowest) bids as follows

f(y) = 5/(8b) for 0.4 b ≤ y ≤ 2b

= 0, elsewhere

a. Prove this probability density function is correctly defined

b. Find the cumulative probability density function, F(y) and sketch it

c. Find the probability that the winning bid is less than the agency’s preliminary estimate ‘b’

**Answer:**

1. To prove that the probability density function is correct, we will check if .

MATLAB Script:

syms b y

f(b) = 5/(8\*b);

p = int(f(b),y,0.4\*b,2\*b)

p=1.

∴ The probability density function is correct.

1. 3/8.

MATLAB Script:

syms b y

f(b) = 5/(8\*b);

v = [0.4\*b: 0.01\*b: 2\*b];

p = int(f(b),y,0.4\*b,b)

**[5]**

**6.** In the annual book exhibition held in YMCA grounds, Chennai, over the past several years, during Pongal festival, there is a steady stream of visitors. They visit the various stalls and exit the exhibition at different times. Depending on the path they take – a few bored ones may bypass many stalls and exit practically instantaneously, some may follow the straight path, some may keep circling - unwilling to leave etc. Suppose the duration of their visit is described in terms of a random variable T and the associated continuous probability density function is given by

Here is a parameter of the distribution with units of time.

1. What are the expected bounds of the probability distribution function?
2. Show that the above function is a legitimate probability distribution function.
3. What is the mean duration of the visit to the books exhibition?

**[5]**

**Answer:**

1. For t=0, f(t) = 1/τ

For t=∞, f(t) = 0

∴

1. To check the legitimacy of the above pdf, .

∴

⇒ The above pdf is legitimate.

1. Finding the expected value function using MATLAB, we get:

E[f(t)] = τ

MATLAB Script:

syms t tau

f(t) = (1/tau)\*exp((-t)/tau);

F(t) = t\*f(t);

p = int(F(t),t)

**[5]**

1. The random variable X is described by the following probability density function, f(x) (-∞ <x< +∞)

What are the mean and variance of this distribution? What is the maximum value of this distribution?

**Answer:**

Comparing f(x) with the normal distribution formula,

By comparison, it can be observed that:

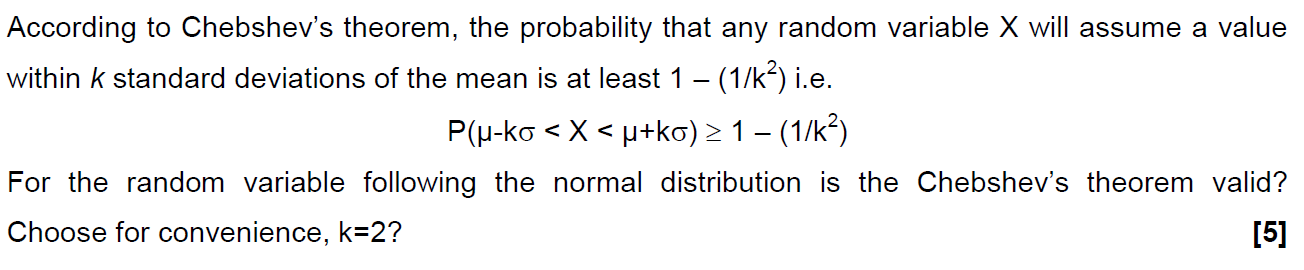
Mean (μ) = 5

Variance (σ2) = 9

Maximum value will be at x = μ, and it is **0.133.**

**[4]**

**Problem 8**

****

1. The equation for standard deviation often results in a relatively broad range of values, it is useful in that it only requires knowledge of the mean and standard deviation, both of which are easily calculated from any sample or data population. The theorem provides what might be called a worst-case look at data dispersion within any data distribution.

In order to validate this theorem, let's first compare the calculations to the 68-95-99.7 rule of thumb for normal distributions. Since those numbers represent the data lying inside the bounds, we use Chebyshev's inequality for data inside the bounds:

Probability = 1 - (1 / k2)

Mathematically, values less than or equal to 1 are not valid for this computation. However, plugging in the k values for 2 and 3 is relatively simple:

P(K=2): 1 - (1 / 22) = 1 - 0.25 = 0.75 (75%)

P(K=3): 1 - (1 / 32) = 1 - 0.11 = 0.89 (89%)

In these cases, Chebyshev's inequality states that at least 75% of the data will fall within 2 standard deviations of the mean, and 89% of the data is expected to fall within 3 standard deviations of the mean. This is less precise than the 95% and 99.7% values that can be used for a known normal distribution; however, Chebshyev's inequality is true for all data distributions, not just a normal distribution.

1. If K = 2

P(Given) = P(Z<k) - P(Z<-k)

= P(Z<2) - P(Z<-2)

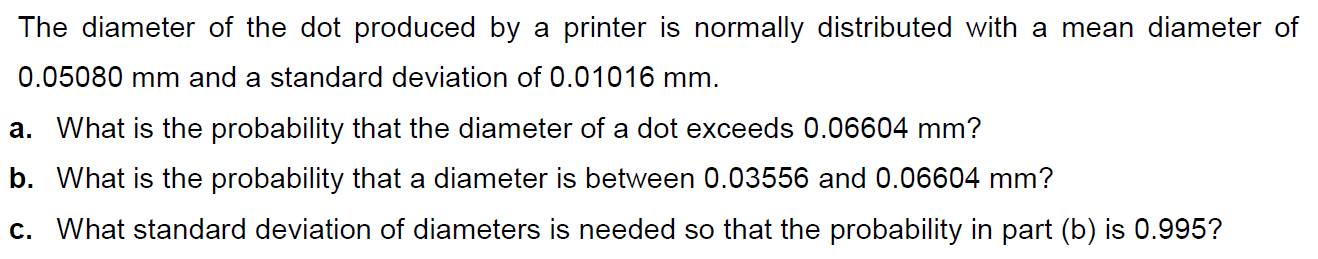
= 0.9821 - 0.0179

= 0.9642

The above inequality can be written as:

For k = 2, verifying the above inequality for normal distribution:

**Problem 9**

****

1. 0.067
2. 0.87
3. 0.00543

MATLAB Script:

syms X sigma mu

mu = 0.05080;

sigma = 0.01016;

f(X) = normpdf(X,mu,sigma);

p1 = vpa(int(f(X),X,0.06604,inf),2);

p2 = vpa(int(f(X),X,0.03556, 0.06604),2);

syms X sigma mu

mu = 0.05080;

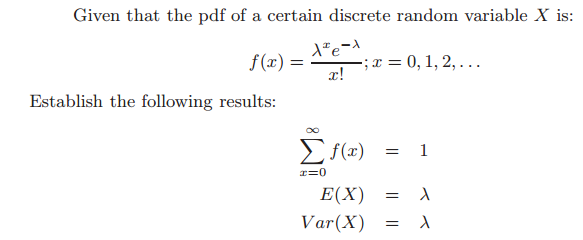
f(X) = (1/(((2\*pi)^0.5)\*sigma))\*exp(-0.5\*((X-mu)/sigma)^2);

p = int(f(X),X,0.03556, 0.06604);

sigma = solve(p-0.995);

vpa(sigma,3)

**[7]**

**Problem 10**

**Answer:**

Part 1) ,

Part 2) For E(X) = λ,

Part 3) For Var(X) = λ,

**[8]**

**Problem 11** Covid (S)care

 It is known that a *healthy* human body has an average temperature of 98.6oF, with a standard deviation of 0.95oF. **Sixty** healthy humans are selected at random.

**a.** What can you say about the probability distribution of the population and about the sampling distribution? Justify your answers.

**b.** What is the probability that their temperatures average at least 99.1oF? If very low/high, explain.

**c.** As many sampled humans actually turned out to be suffering from fever, the sample size had to be reduced to only six, what condition has to be satisfied to re-estimate the probability required in part b? Find the new probability with the reduced sample size and again comment on the magnitude of the probability value now.

μ= 98.6 F

= 0.95 F

n = 60

1. The Central Limit Theorem states that for the samples of size 30 or more, the mean is nearly equal to the normally distributed with mean being μ and standard deviation being . Further, the probability distribution of the given population need not be normally distributed it could take any form. The CLT is valid for samples less than 30 when the population is also normally distributed.
2. P(T> 99.1) = P[Z > ] = P[Z > 4.077] = 1- P[Z < 4.077]

P[Z < 4.077] = 0.99997 then P(T>99.1) = 0.00003

1. The probability should be normally distributed to satisfy the condition for Central Limit Theorem.

P(T > 99.1) = P[Z > = P[Z > 1.2892]

= 1- P[Z < 1.2892] = 1-0.90147 = 0.09853

**[11]**

**Problem 12**

The Weibull probability distribution is a two parameter distribution defined for x>0 by

1. If δ and β are both unity, what will be the general shape of the distribution?
2. when special case a holds, for what value of X, say q, will P(X>q) =P(X<q)?
3. What is the equation of the cumulative probability distribution, for the special case a and for the general case?

**Answer:**

1. For, β = 1 and δ = 1:
2. For, P(X>q):

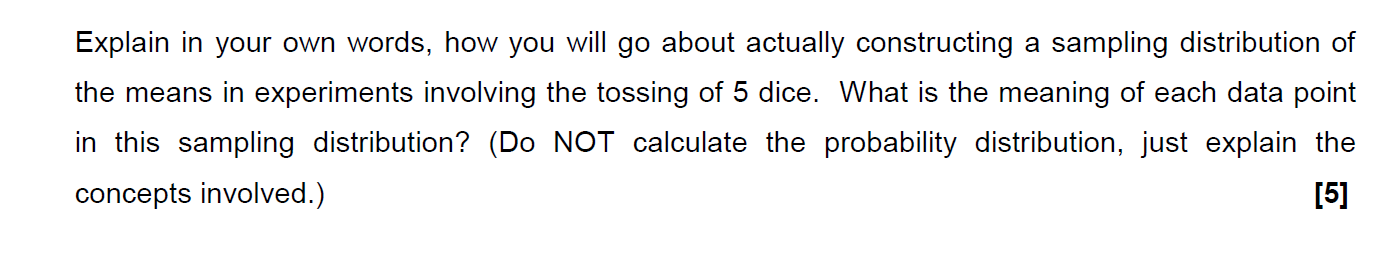
….. (1)

…… (2)

1. For special case,

For normal case,

**[5]**

**Problem 13**

Ans:

In a sampling distribution, each dot represents a sample from the population and a mean calculated from that sample.

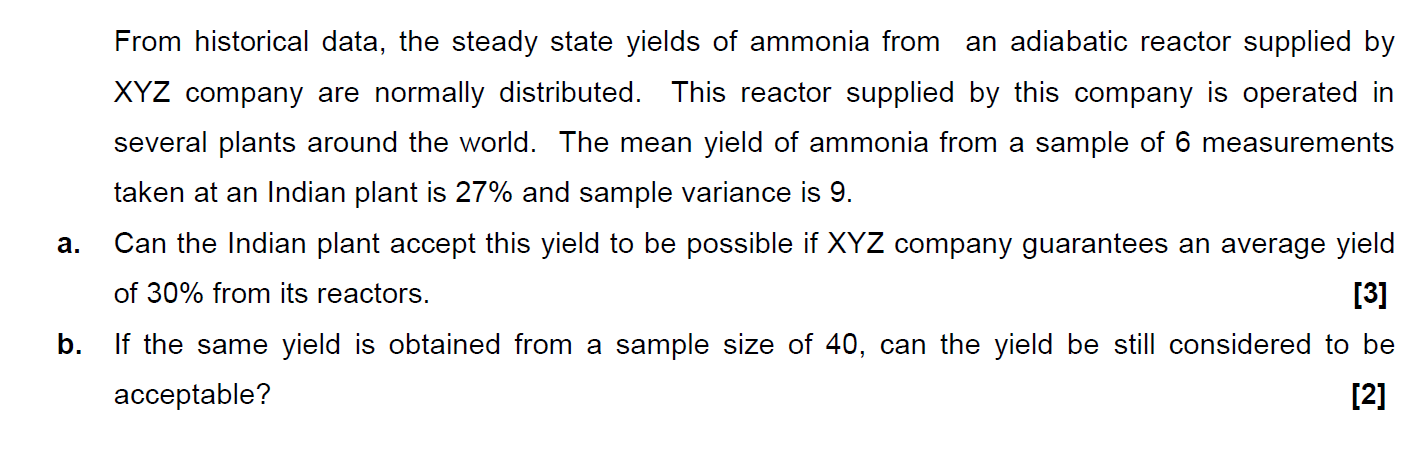
In a sampling distribution, each dot represents a sample from the population and a mean calculated from that sample.

Each point would act as a mean of 5 randomly selected outcomes of the dice and that would form a set which is a point in that distribution.

There are 6 possible outcomes from a dice.

The values of means would range from 1 to 6 as lowest and highest valued sets are {1,1,1,1,1} & {6,6,6,6,6}

The distribution won’t be a continuous distribution as the means would have definite values and they are bounded.

**Problem 14**

Matlab Code:

N = 6

yMean = 27

ySEM = 3/sqrt(N)

CI95 = tinv([0.025 0.975], N-1); % Calculate 95% Probability Intervals Of t-Distribution

yCI95 = bsxfun(@times, ySEM, CI95(:));

figure

plot(x, yMean) % Plot Mean Of All Experiments

hold on

plot(x, yCI95+yMean) % Plot 95% Confidence Intervals Of All Experiments

hold off

grid

1. 30% does lie in the 95% confidence interval. Hence it can be accepted.
2. N = 40;

yMean = 27

ySEM = 0.4743

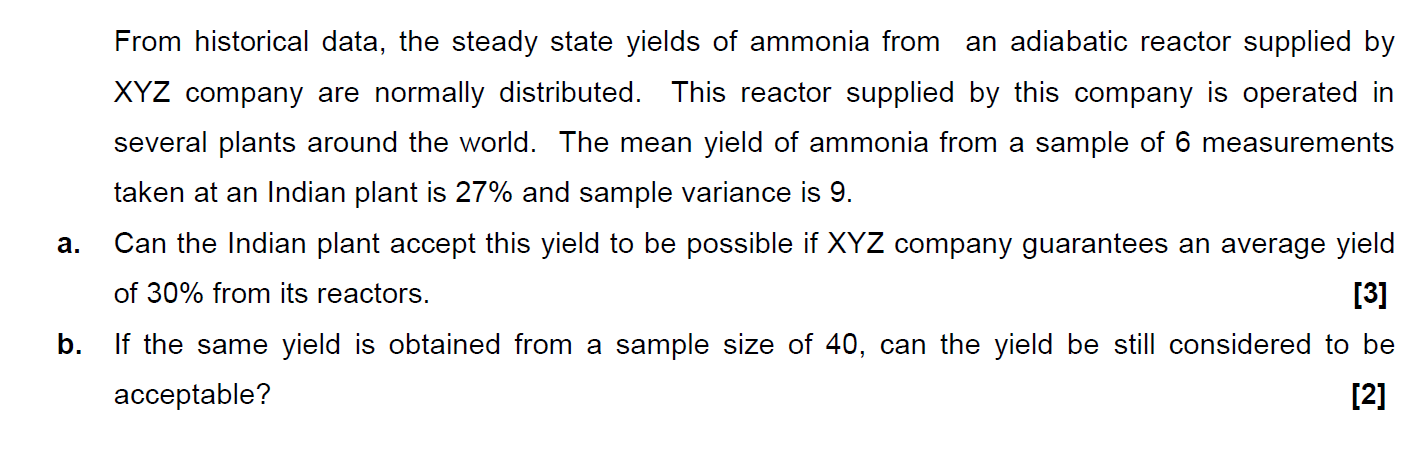
>> CI95

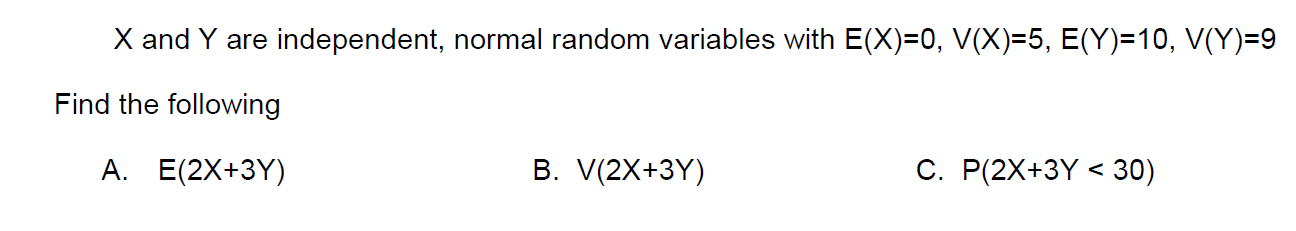
CI95 = -2.0227 2.0227

>> yCI95

yCI95 = -0.9594 0.9594

As the sample size increased, 30% does not lie inside the 95% confidence interval. Hence it can not be accepted.

**Problem 15**

****

**[6]**

1. E(2X+3Y)

2E(X)+3E(Y) = 30

1. V(2X+3Y)

V(2X) + V(3Y)

2\*2V(X) + 3\*3V(Y)

4V(X)+9V(Y)  = 101

c. P(2X+3Y < 30)

P(Z< ()

P(Z< ()

P(Z<0)

0.5

**Problem 16**

A tobacco company claims that the amount of nicotine in cigarettes is a random variable with mean 2.2 mg and standard deviation 0.3 mg. However, the sample mean nicotine content of 100 randomly chosen cigarettes was 3.1 mg. What is the approximate probability that the sample mean would have been as high or higher than 3.1 mg if the company’s claim was true?

**[6]**

Let X̅ be the average nicotine content of 100 randomly chosen cigarettes

Sample mean (X̅) = 2.2 mg

Sample standard deviation = = (0.3/10) = 0.03 mg

Therefore, if the company’s claims are true,

P (X̅ > 3.1) = 1 – P (X̅ < 3.1) = 1 – 1 = 0

The probability is approximately equal to zero.

MATLAB CODE:

p = 1 – normcdf (3.1,2.2,0.03)

**Problem 17**

A city installs 2000 electric lamps for street lighting. These lamps have a mean burning life of 1000 hours with a standard deviation of 200 hours. The normal distribution is a close approximation to this case.

1. What is the probability that a lamp will fail in the first 700 burning hours?
2. What is the probability that a lamp will fail between 900 and 1300 burning hours?
3. How many lamps are expected to fail between 900 and 1300 burning hours?
4. What is the probability that a lamp will burn for exactly 900 hours?

Ans:

Let X be the random variable denoting the burning life We know that , X~N(1000,200^2)

(A) P(X<700) Standardizing: P(Z<(700-1000) /200) =P(Z<-1.5) = 0.067

(B) P(900<X<1300) = P((900-1000) /200 <Z< (1300-1000)/200)) =P(-0.5<Z<1.5) = 0.6247

(C) The expected number of failures is given by the total number of lamps multiplied by the probability of failure in that interval.

Thus, The expected number of failures = (2000) (0.6247) ~ 1250 lamps

(D) Since the burning life is a continuous random variable, the probability of a life of exactly 900 burning hours (not 900.1 hours or 900.01 hours or 900.001 hours, etc.) is zero

**[10]**

**Problem 18**

An engineer decides to buy four new snow tires for his car. He finds that Retailer A is offering a special cash rebate, which depends on how much snow falls during the first winter. If this snowfall is less than 50% of the mean annual snowfall for his city, his rebate will be 50% of the list price. If the snowfall that winter is more than 50% but less than 75% of the mean annual snowfall, his rebate will be 25% of the list price. If the snowfall is more than 75% of the mean annual snowfall, he will receive no rebate. The engineer finds from a reference book that the annual snowfall for his city has a mean of 80 cm and standard deviation of 20 cm and approximates a normal distribution. The list price for the brand and size of tires he wants is $80.00 per tire.

The engineer checks other retailers and finds that Retailer B sells the same brand and size of tires with the same warranty for the same list price but offers a discount of 5% of the list price regardless of snowfall that year.

1. Compare the expected costs of the two deals. Which expected cost is less?
2. How much is the difference for four new snow tires? Neglect the relative advantages of a cash rebate as compared to a discount.

Answer

1. For Retailer A:

Let X be the random variable which represents annual snowfall

Then the expected price from Retailer A is:

% Rebate = (50%)(0.0228) + (25%)(0.1359) = 4.5375 % of list price

Retailer A price = $ 76.37

Discount given by Retailer B is 5% on list price.

Retailer B price = $ 76

The expected cost of buying from Retailer A is **slightly more** than expected cost of buying from Retailer B.

1. Cost of four snow tires is as follows:

List price = 4 \* 80 = $320

Expected cost of buying from Retailer A = (1-0.045375)(320) = $305.48

Expected cost of buying from Retailer A = (1-0.05)(320) = $304.00

The difference between expected cost = 305.48 – 304.00 = $1.48

**Problem 19**

An assembly plant has a bin full of steel rods, for which the diameters follow a normal distribution with a mean of 7.00 mm and a variance of 0.100 mm2, and a bin full of sleeve bearings, for which the diameters follow a normal distribution with a mean of 7.50 mm and a variance of 0.100 mm2 . What percentage of randomly selected rods and bearings will not fit together?

If for any selection of one rod and one bearing, the difference between the bearing diameter and the rod diameter is positive, the pair will fit.

If not, they won’t. (That may be a little oversimplified, but let us neglect consideration of clearance.)

Let d be the difference between the bearing diameter and the rod diameter.

Because both the diameters of bearings and the diameters of rods follow normal distributions, the difference will also follow a normal distribution.

The mean difference will be 7.50 mm – 7.00 mm = 0.50 mm.

The variance of the differences will be the sum of the variances of bearings and rods:

Sigma(d) 2 = 0.100 + 0.100 = 0.200 mm2 .

Then,

Sigma(d) = 0.200 = 0.447 mm.

Z1 = (0 - 0.5) / 0.447  = − 1.118

Phi(z1) = Phi(–1.118) = 0.1314 (from normal distribution table with z taken to two decimals) or 0.1318 (from the Excel function NORMSDIST).

Therefore, 13.2% or 13.1% of randomly selected sleeves and rods will not fit together.

**Problem 20**

A certain dimension is measured on four successive items coming off a production line. This sample gives x̅ = 2.384 and s = 0.048.

(a) On the basis of this sample, what is the 95% confidence interval for the population mean?

(b) If instead of estimating the standard deviation from a sample, we knew the true standard deviation was 0.048, what then would be the 95% confidence interval for the population mean?

Answer

a)    x̅ = 2.384 and s = 0.048

95% confidence interval,

MATLAB code:

x\_bar=2.348; s=0.048; n=4; alpha=0.05

t=tinv(1-(alpha/2), n-1)

S=s/sqrt(n)

Lower = x\_bar-(t\*S)

Upper = x\_bar+(t\*S)

Therefore, Lower bound = 2.2716 ; Upper bound = 2.4244

b)    True standard deviation (𝝈) = 0.048

MATLAB code:

x\_bar=2.348; s=0.048; n=4; alpha=0.05

z=norminv(1-(alpha/2), n-1)

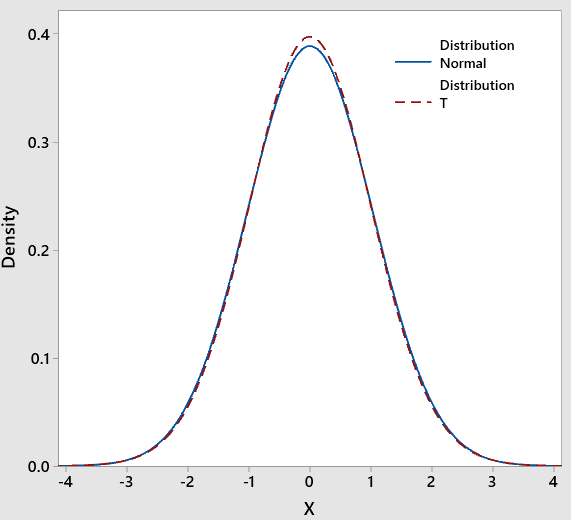
S=s/sqrt(n)

Lower = x\_bar-(z\*S)

Upper = x\_bar+(z\*S)

Therefore, Lower bound =  2.3010 ; Upper bound = 2.3950

**Problem 21               [4]**

.

The figure shows the t-distribution (dashed line) with 80 degrees of freedom overlaid on a normal distribution (solid line).  The mean and variance of both these distributions are identical.  Find these two parameters and explain the almost identical overlap between the two distributions.

The t distribution has the following properties:

* The mean of the distribution is equal to 0 .
  + Mean = 0
* The [variance](https://stattrek.com/statistics/dictionary.aspx?definition=Variance) is equal to v / ( v - 2 ), where v is the degrees of freedom (see last section) and v > 2.
  + Variance = 80/78 = 1.025641
* The [variance](https://stattrek.com/statistics/dictionary.aspx?definition=Variance) is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the [standard normal distribution](https://stattrek.com/statistics/dictionary.aspx?definition=Standard_normal_distribution).
  + As DoF = 80, which is comparatively larger number, the variance approaches to 1
  + Hence, we are seeing the nature of the curves being identically distributed or rather overlapping with each other

**Problem 22 [7]**

Proof:

- = - + -

(- ) = [( - ) + ( - )]

(- ) = (- )+ n()

(- )) -

) -

=

= + + )

=

) -

Hence proved.

**Problem 23**

Thermo-mechanically treated iron (TMT) rods used in construction are to have area specific earthquake resistance.  The rods prepared from a company have mean yield strength of 500 N/mm2 and standard deviation of 50 N/mm2.                                          **[7]**

1. Will 95% of the bars chosen from this population have the required yield strength of at least 470 N/mm2?
2. If not, what process modification may be suggested (maintaining the same mean yield strength) to achieve the aim stated in part **a**?
3. Let X represent the yield strength of TMT rods

Hence, only 72.57% of bars chosen from this population have the yield strength of at least 470 N/mm2

1. The standard deviation can be changed to attain the required result,

Using MATLAB;

**Problem 24**

A maker of a certain brand of low-fat cereal bars claims that their average saturated fat content is 0.5g.  In a random sample of 8 cereal bars of this brand the saturated fat content was 0.6,0.7,0.7,0.3,0.4.0.5,0.4 and 0.2.  Would you agree with the claim?  Assume a normal distribution.

**[6]**

Answer:

Sample mean = 0.4750

Sample standard deviation = 0.1714

Calculating 95% confidence interval (CI) using the sample mean and sample standard deviation:

MATLAB Code

x = [0.6 0.7 0.7 0.3 0.4 0.5 0.4 0.2]

sam\_mean = mean(x)

std\_dev = std(x,1)

n = 8

alpha = 0.05

z=tinv(1-(alpha/2), n-1)

S=std\_dev/sqrt(n)

L=sam\_mean - (t\*S)

U=sam\_mean + (t\*S)

 L = 0.3317, U = 0.6183

The average saturated fat content claimed is within 95% confidence interval so therefore we would agree to the claim under this confidence interval.

**Problem 25**

The chemical benzene is highly toxic to humans and it causes cancer. In any production process Involving benzene, the water in the output of the process must not exceed 7950 ppm because of Government Regulations. For a particular process, the water sample was collected by a manufacturer 25 times randomly and the sample average was 7960 ppm. It is known from historical data that the standard deviation σ is 100 ppm.

a. What is the probability that the sample average in this experiment would exceed the government limit if the population mean is equal to the limit? Use the central limit theorem.

*Assume the distribution of benzene concentrations to be normal.*

**[5]**

Answer:

Let X be a random variable that represents the amount of benzene, in ppm.

Assume that the population mean is Mu\_x  = 7950, we need to calculate the probability,

P(X\_hat > = Mu\_x)

1. Sample Mean: Mu\_x\_hat = Mu\_x  = 7950
2. Standard deviation: 100/sqrt(25)  = 20

We know that random variable Z has a normal distribution with mean 0 & standard deviation 1

Hence, we have

P(X\_hat > Mu\_x) = P(X\_hat > 7950)

= P(Z > (7950-7950)/20)

= P(Z > 0) = 0.5

Final answer: 0.5

