PROBLEM SET 9

VINAY KUMAR RANGANATH BABU

Problem 1

heavy.a=c(233,291,312,250,246,197,268,224,239,239,254,276,234,181,248,252,202,218,212,325)

heavy.b=c(344,185,263,246,224,212,188,250,148,169,226,175,242,252,153,183,137,202,194,213)

1)

a) The experimental unit here is the middle aged heavy person

b) The populations were drawn from 2 samples (i.e. Type A - from middle aged heavy men with urgency, aggression and ambition and second Type B with noncompetitive, more relaxed and less hurried middle aged heavy men). This is a 2 sample problem.

c) 20 samples (cholesterol levels) were drawn from the Type A middle aged heavy men population which is one measurement. 20 samples were drawn from the Type B middle aged heavy men population which is second measurement.

d) The parameter of interest for this problem is difference in the average of cholesterol values between two men (i.e. delta=mean(Type A)-mean(Type B))
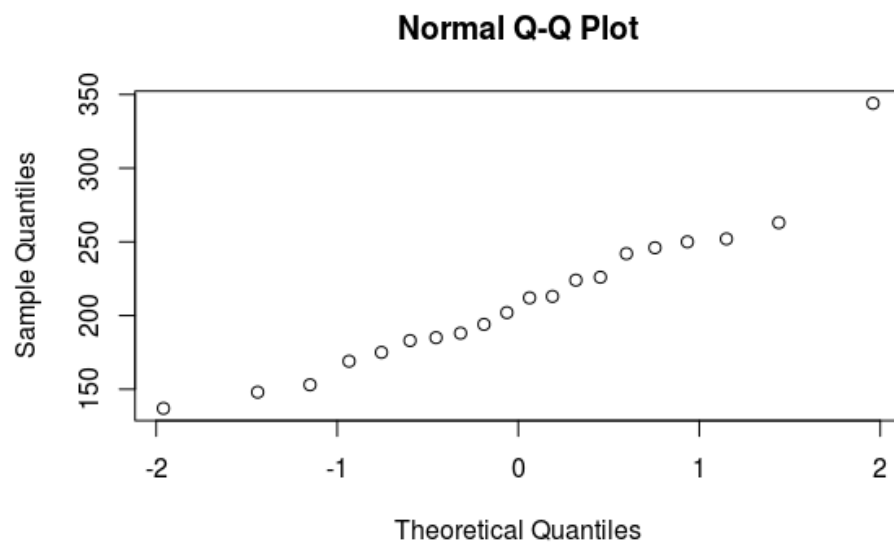
e) null hypothesis H0: delta<=0 and alternative H1: delta>0

2)

qqnorm(heavy.a) # Its almost a straight line. Hence we can assume that the samples were drawn from the normal distribution

Normal Q-Q Plot
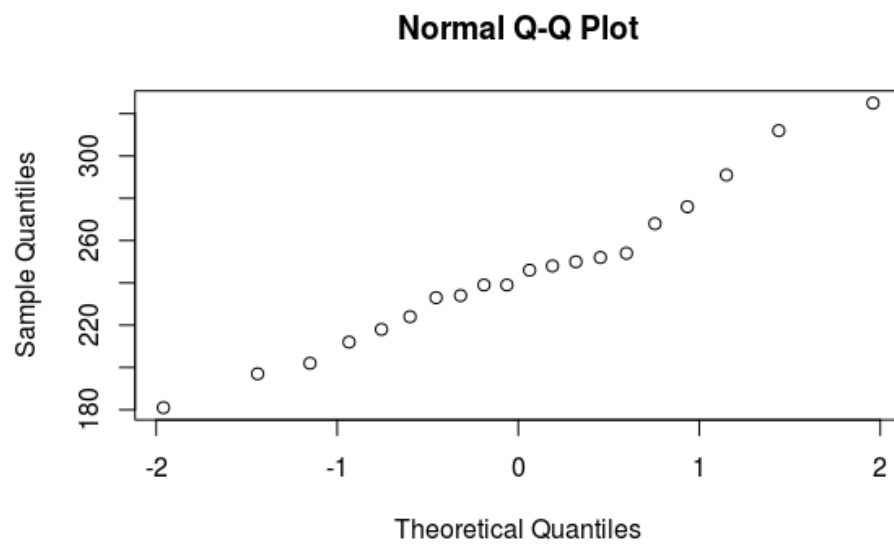
qqnorm(heavy.b) # Its almost a striaght line. Hence we can assume that the samples were drawn from the normal distribution

Normal Q-Q Plot

3) a)

n1= length(heavy.a)

> [1] 20

n2= length(heavy.b)

> [1] 20

mean.a=mean(heavy.a)

> [1] 245.05

sd.a=sd(heavy.a)

> [1] 36.63831

mean.b=mean(heavy.b)

> [1] 210.3

sd.b=sd(heavy.b)

> [1] 48.33991

delta=mean.a-mean.b

> [1] 34.75

stats=delta/sqrt(((sd.a*sd.a)/20)+((sd.b*sd.b)/20))

> [1] 1.105045

The degrees of freedom calculated comes out to be

nu=((((sd.a*sd.a)/n1)+((sd.b*sd.b)/n1))^2)/(((((sd.a*sd.a)/n1)^2)/(n1-1))+((((sd.b*sd.b)/n2)^2)/(n2-1)))

> [1] 35.41308

p.value=(1-pt(2.56,df=nu))

> [1] 0.007

Since we have the p value to be 0.007 and our alpha is 0.05, hence we can reject the null hypothesis. Hence type A men have high cholesterol.

b) t.test(heavy.a,heavy.b,conf.level = 0.90)

Gives us the confidence interval between 11.8 and 57.6

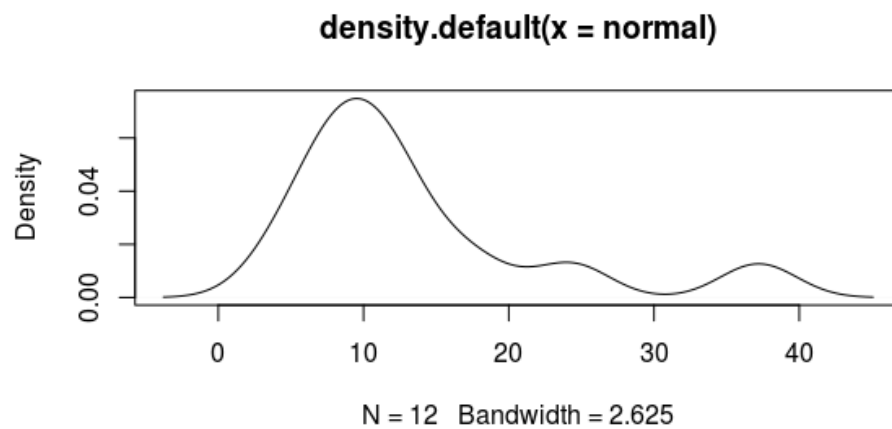# Question 2

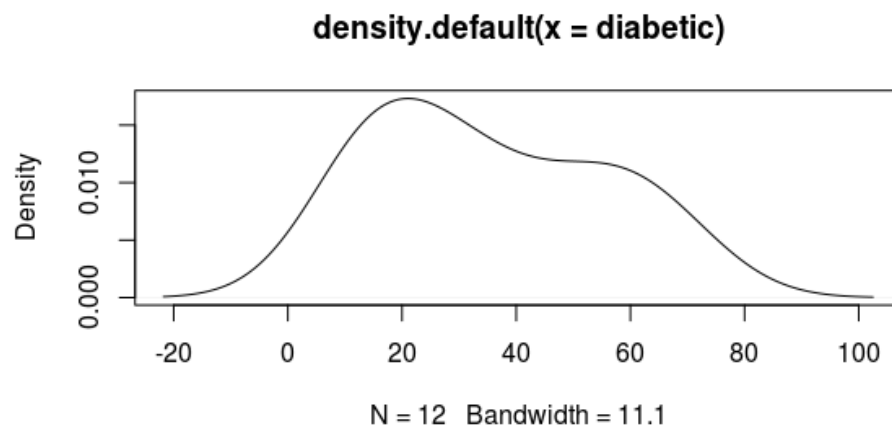normal=c(4.1,6.3,7.8,8.5,8.9,10.4,11.5,12.0,13.8,17.6,24.3,37.2)

diabetic=c(11.5,12.1,16.1,17.8,24.0,28.8,33.9,40.7,51.3,56.2,61.7,69.2)

1) To check for symmetric distribution

plot(density(normal))

density.default(x = normal)

N = 12 Bandwidth = 2.625

plot(density(diabetic))

density.default(x = diabetic)

N = 12 Bandwidth = 11.1

As we can see in the plots that the data is skewed to the right. Hence the samples are not from the symmetric distribution

2)

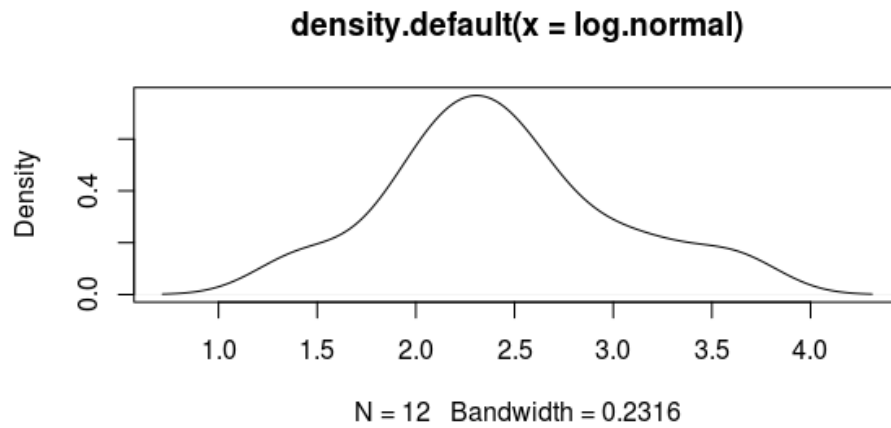Applying logs to the data gives us the following

log.normal=log(normal)

> [1] 1.410987 1.840550 2.054124 2.140066 2.186051 2.341806 2.442347 2.484907 2.624669 2.867899[11] 3.190476 3.616309
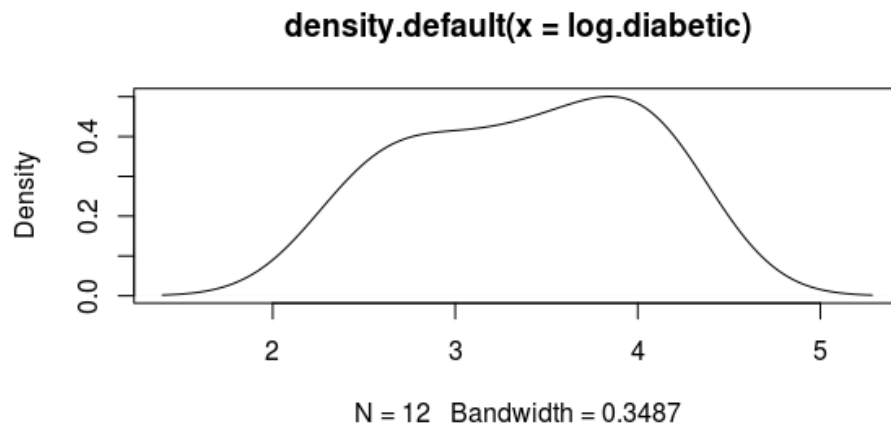
log.diabetic=log(diabetic)

> [1] 2.442347 2.493205 2.778819 2.879198 3.178054 3.360375 3.523415 3.706228 3.937691 4.028917[11] 4.122284 4.237001

plot(density(log.normal))

density.default(x = log.normal)

N = 12 Bandwidth = 0.2316

plot(density(log.diabetic))

density.default(x = log.diabetic)

N = 12 Bandwidth = 0.3487

Applying square roots to the data gives us the following
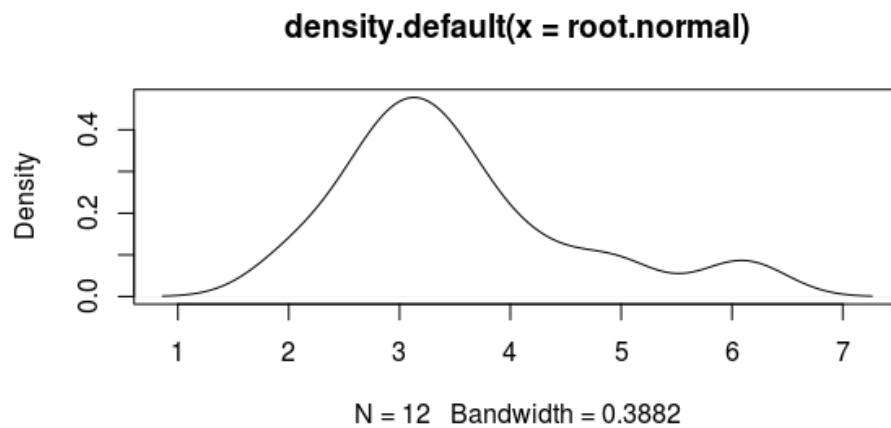
root.normal=sqrt(normal)

> [1] 2.024846 2.509980 2.792848 2.915476 2.983287 3.224903 3.391165 3.464102 3.714835 4.195235[11] 4.929503 6.099180
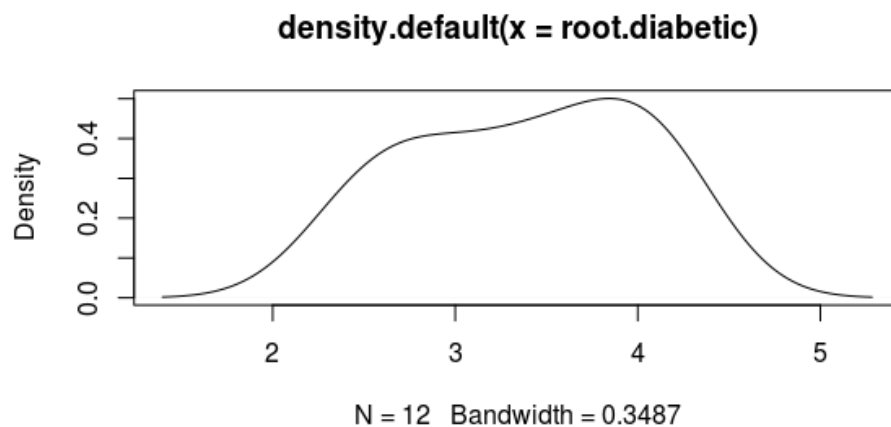
root.diabetic=log(diabetic)

> [1] 2.442347 2.493205 2.778819 2.879198 3.178054 3.360375 3.523415 3.706228 3.937691 4.028917[11] 4.122284 4.237001

plot(density(root.normal))

density.default(x = root.normal)

N = 12 Bandwidth = 0.3882

plot(density(root.diabetic))

density.default(x = root.diabetic)


N = 12 Bandwidth = 0.3487

After applying the transformations, we see that the skewness to the samples are somewhat reduced and we can now say the sample can be assumed to be taken from symmetric distribution. The symmetry is more clear when we apply log. Hence we prefer to take log transformation
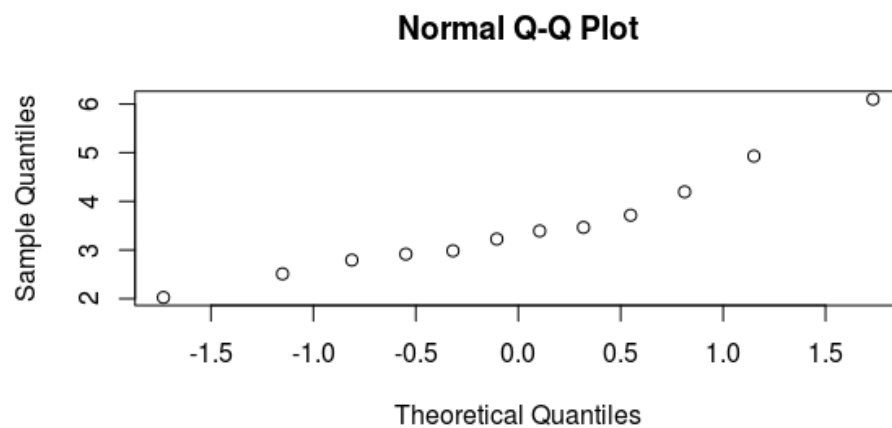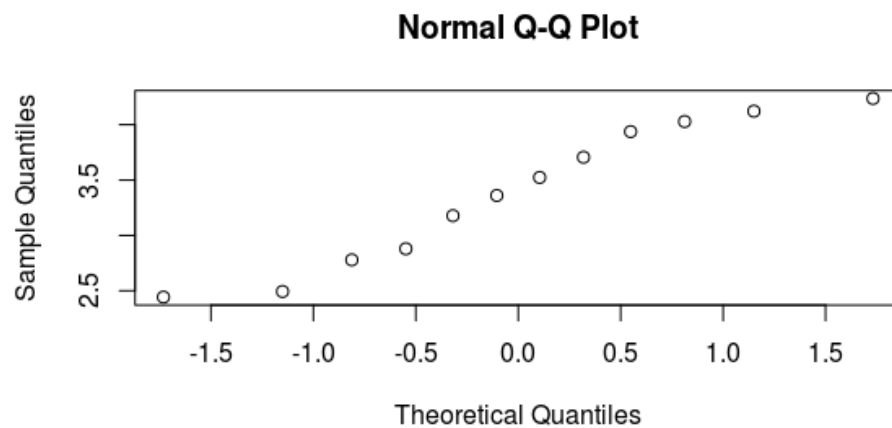
3)

qqnorm(log.normal)

Normal Q-Q Plot

qqnorm(log.diabetic)

Normal Q-Q Plot
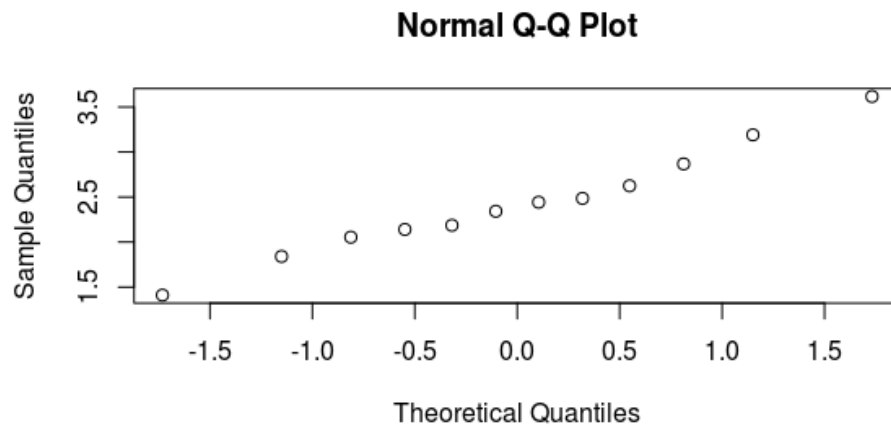
qqnorm(root.normal)

Normal Q-Q Plot
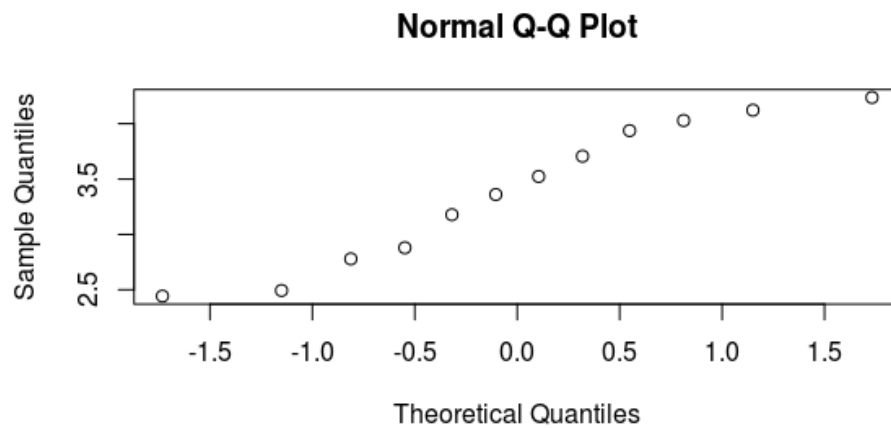
qqnorm(root.diabetic)

Normal Q-Q Plot

After applying the transformations, the normal curve is almost a straight line. Hence we can assume the sample data comes from the normal distribution

4) t.test(log.diabetic,log.normal,alt="greater")

Welch Two Sample t-testdata: log.diabetic and log.normalt = 3.8041, df = 21.9, p-value = 0.0004888alternative hypothesis: true difference in means is greater than 095 percent confidence interval: 0.5250797   Infsample estimates:mean of x mean of y 3.390628 2.433349

P-Value is 0.0005. there's strong evidence that the (population) mean of the logged diabetic thromboglobulin exceeds the mean of the logged diabetic thromboglobulin.
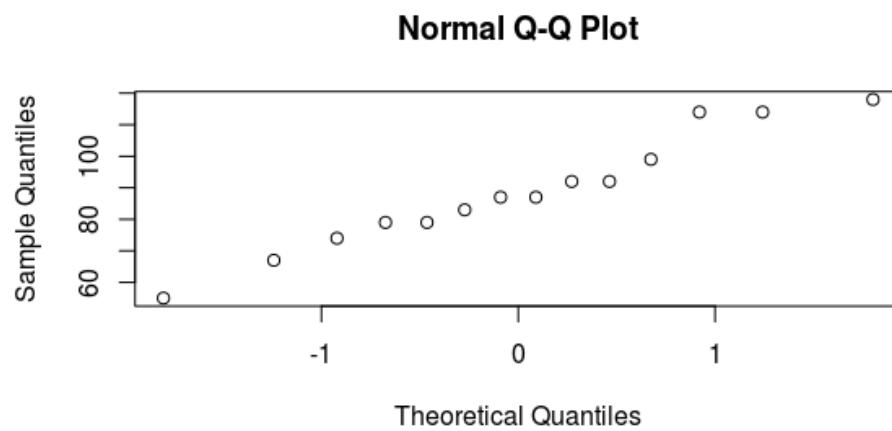
Question 3
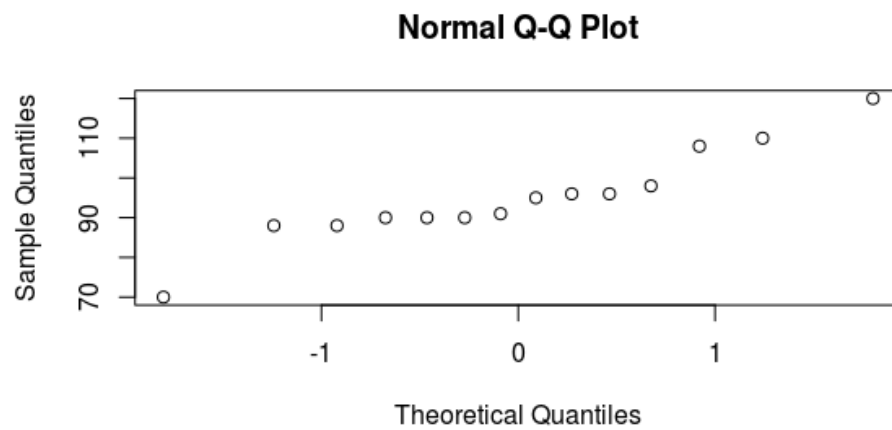
old=c(74,114,114,87,92,55,67,118,79,83,79,92,99,87)

new=c(70,98,90,95,88,108,110,96,91,88,120,96,90,90)

lets check if the sample data follows normality

qqnorm(old)

Normal Q-Q Plot

qqnorm(new)

Normal Q-Q Plot

Since there are 2 samples with unknown variances, we can compute Welche's test

mean.old=mean(old)

> [1] 88.57143

mean.new=mean(new)

> [1] 95

sd.old=sd(old)

> [1] 18.21966

sd.new=sd(new)

> [1] 11.90992

meandiff=mean.new-mean.old

> [1] 6.428571

stats=meandiff/sqrt(((sd.old*sd.old)/length(old)+(sd.new*sd.new)/length(new)))

> [1] 1.105045

nu1=((((sd.old*sd.old)/n3)+((sd.new*sd.new)/n4))^2)/(((((sd.old*sd.old)/n3)^2)/(n3-1))+((((sd.new*sd.new)/n4)^2)/(n4-1)))

> [1] 22.39456

p.value=(1-pt(stats,df=nu1))

> [1] 0.1404357

t.test(new,old,alt="greater") also gives us the same result.

Welch Two Sample t-testdata: new and oldt = 1.105, df = 22.395, p-value = 0.1404alternative hypothesis: true difference in means is greater than 095 percent confidence interval: -3.553149 Infsample estimates:mean of x mean of y 95.00000 88.57143

Hence the null hypothesis is consistent saying we discourage the film buff's impression that the movies are shorter than they used to be.