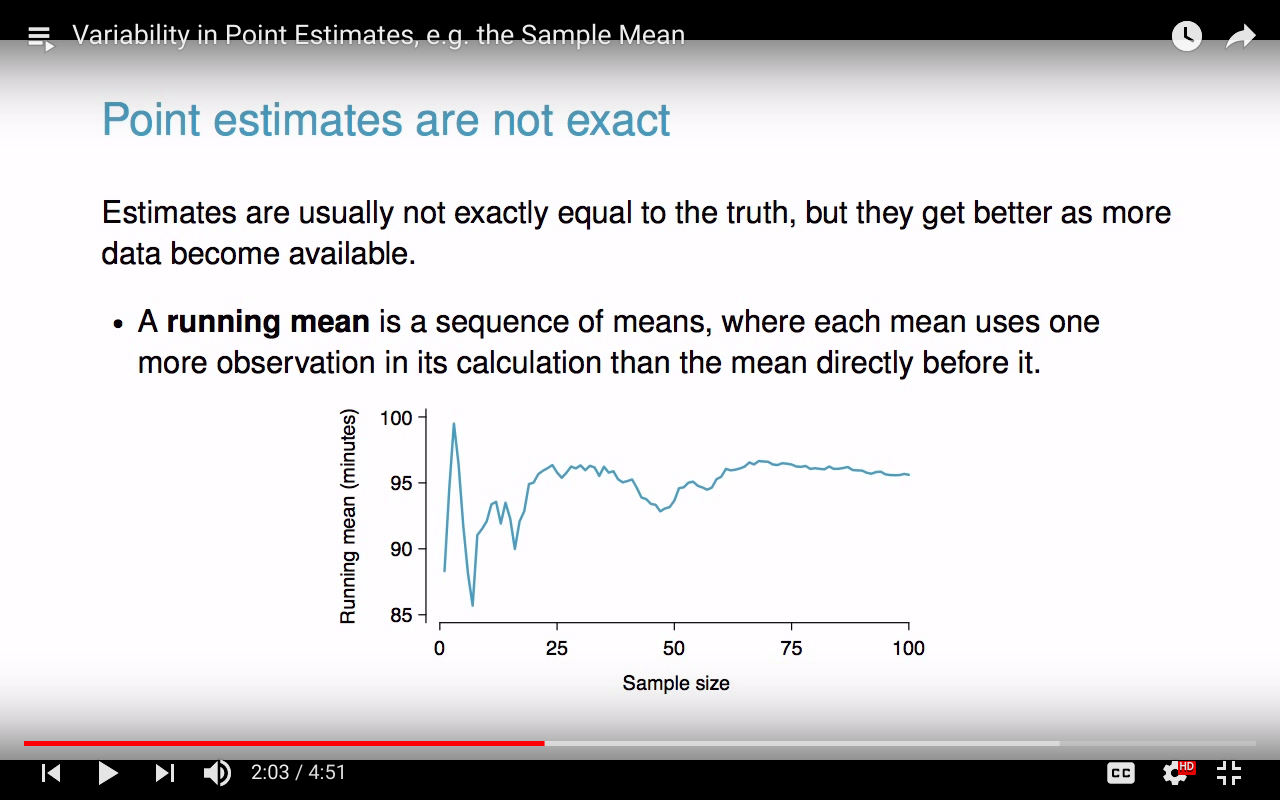
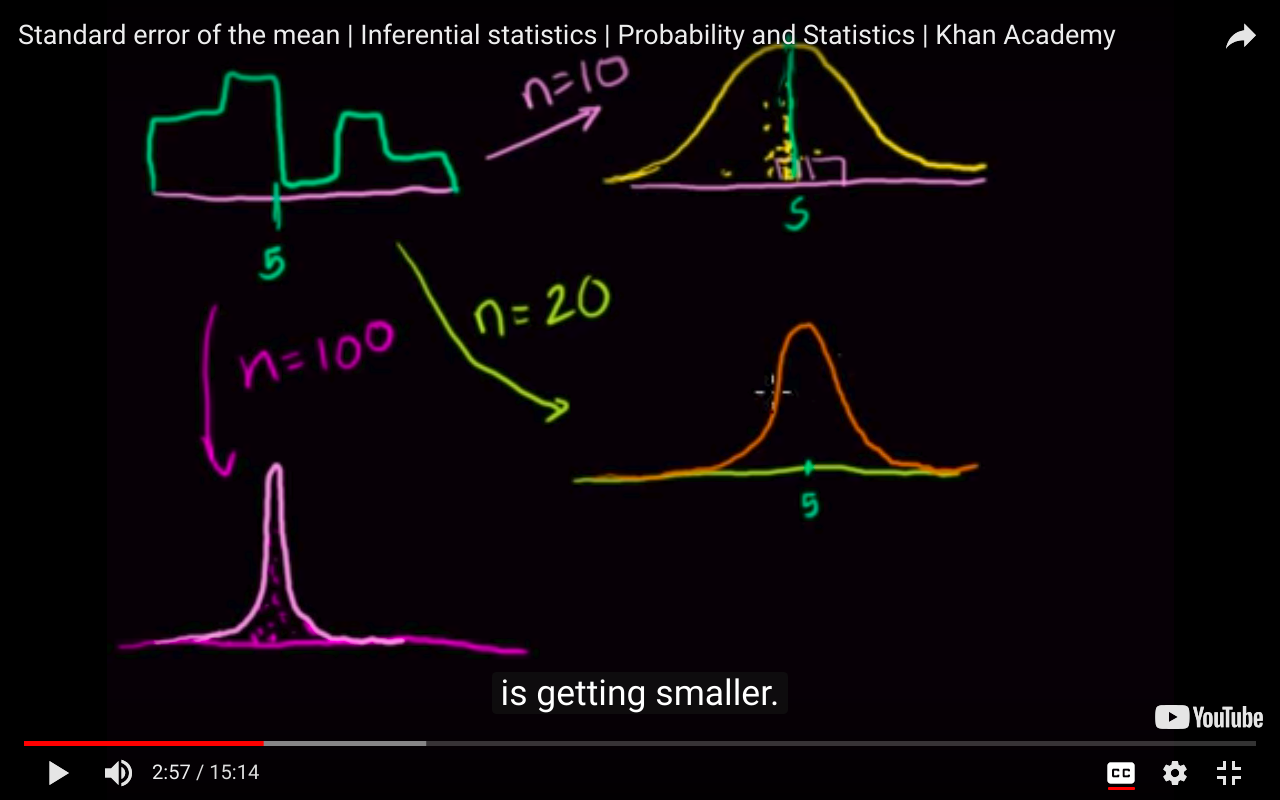
~~STD: std is used when the dist is approximately normal.~~

variability in estimates: every time we take a new sample our estimates i.e mean,median e.t.c changes therefore point estimates aren't handy in estimating the parameter.





As u increase the sample size the std deviation becomes less and the sampling dist approaches normal

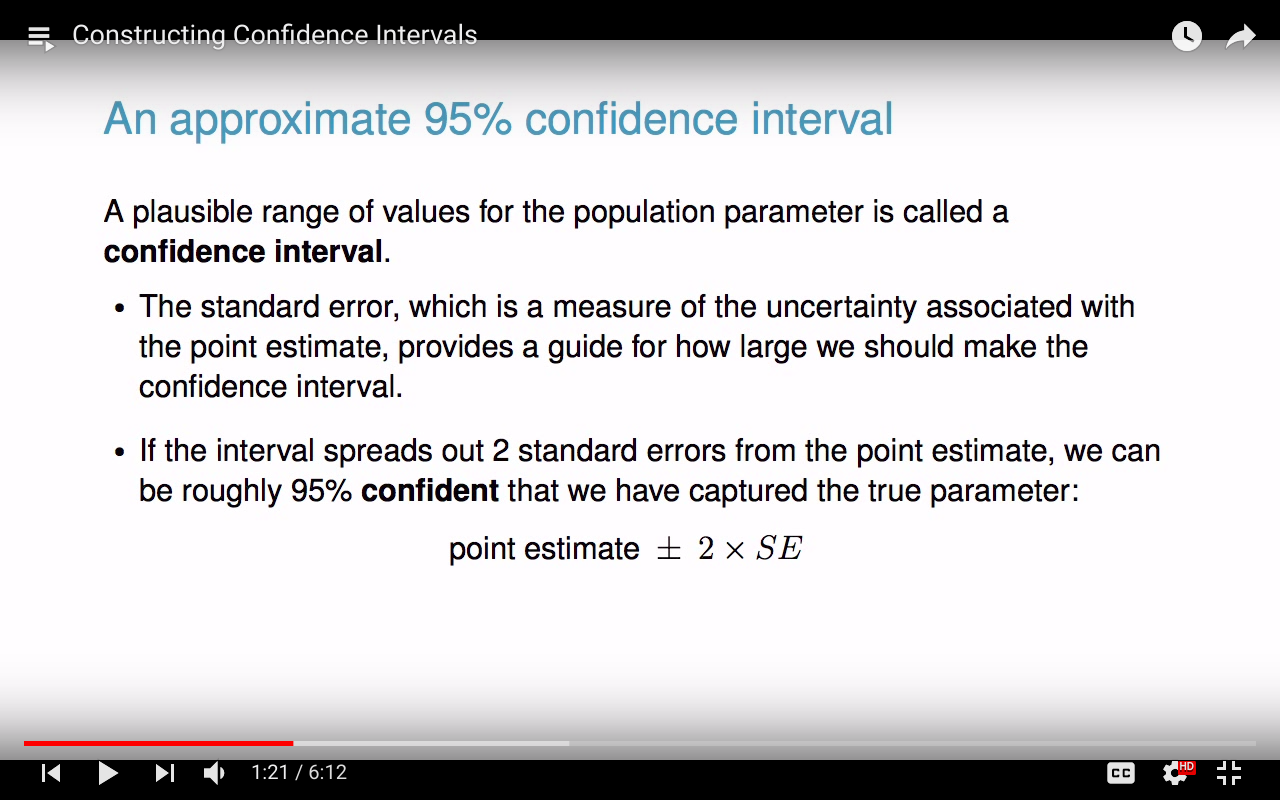
Khan academy tutorial :<https://youtu.be/J1twbrHel3o?t=630>

S.E: It is called an error because the standard deviation of the sampling distribution tells us how different a sample mean can be expected to be from the true mean. In other words, if we assume that the mean of our sample is always the true mean (even though it probably isn't) the standard deviation can tell us how likely we are to be wrong.

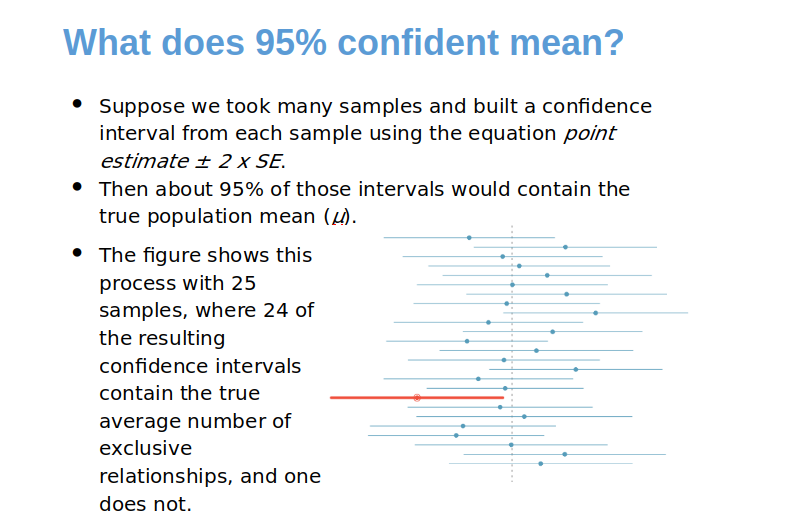
CONFIDENCE INTERVAL:

we construct it bcz we only have samples and not the whole population.

It helps us to identify how close is statistic to population parameter.



**we never construct CI for statistic.**



The most important part ( ^ ).

Calculating the value of Z for confidence interval.

CI formula: point estimate +- Z\*S.E

Z\*SE is margin of error

qnorm gives the z score

z score for 98% CI :

> qnorm(0.99)

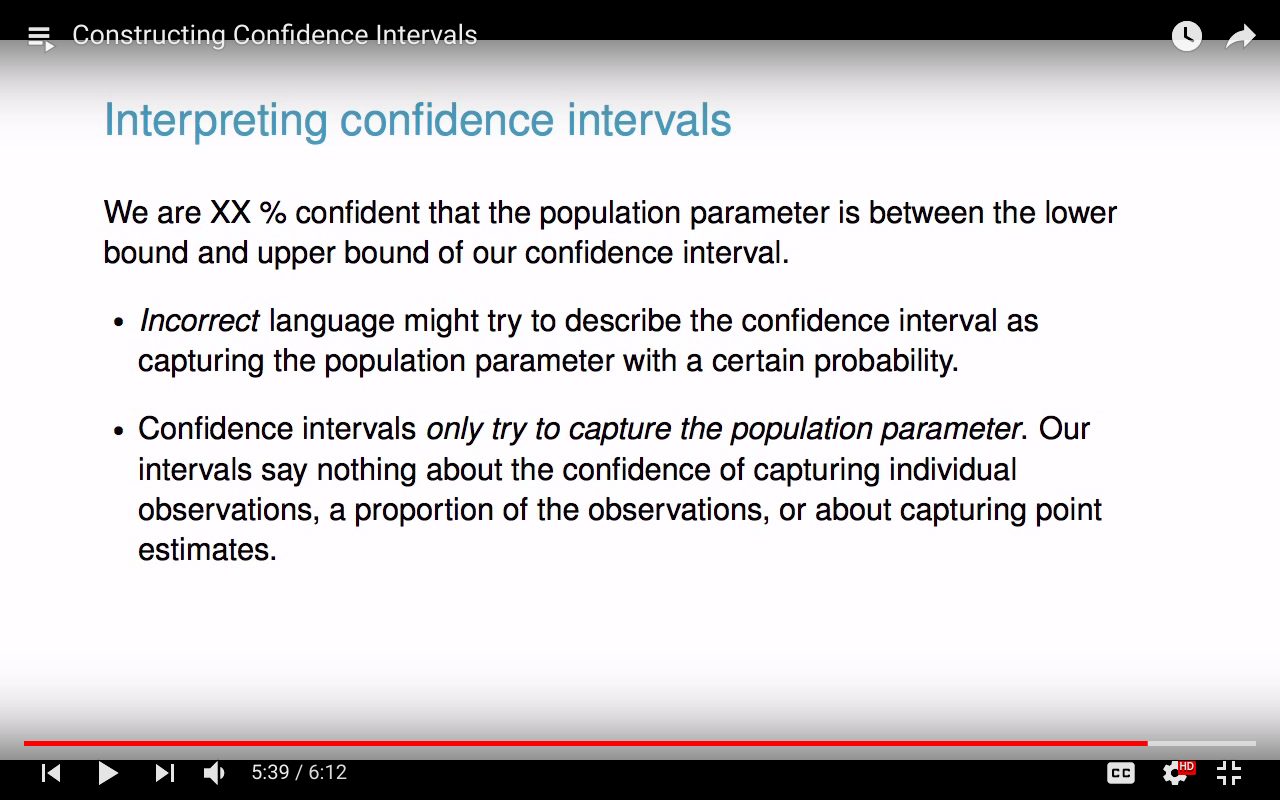
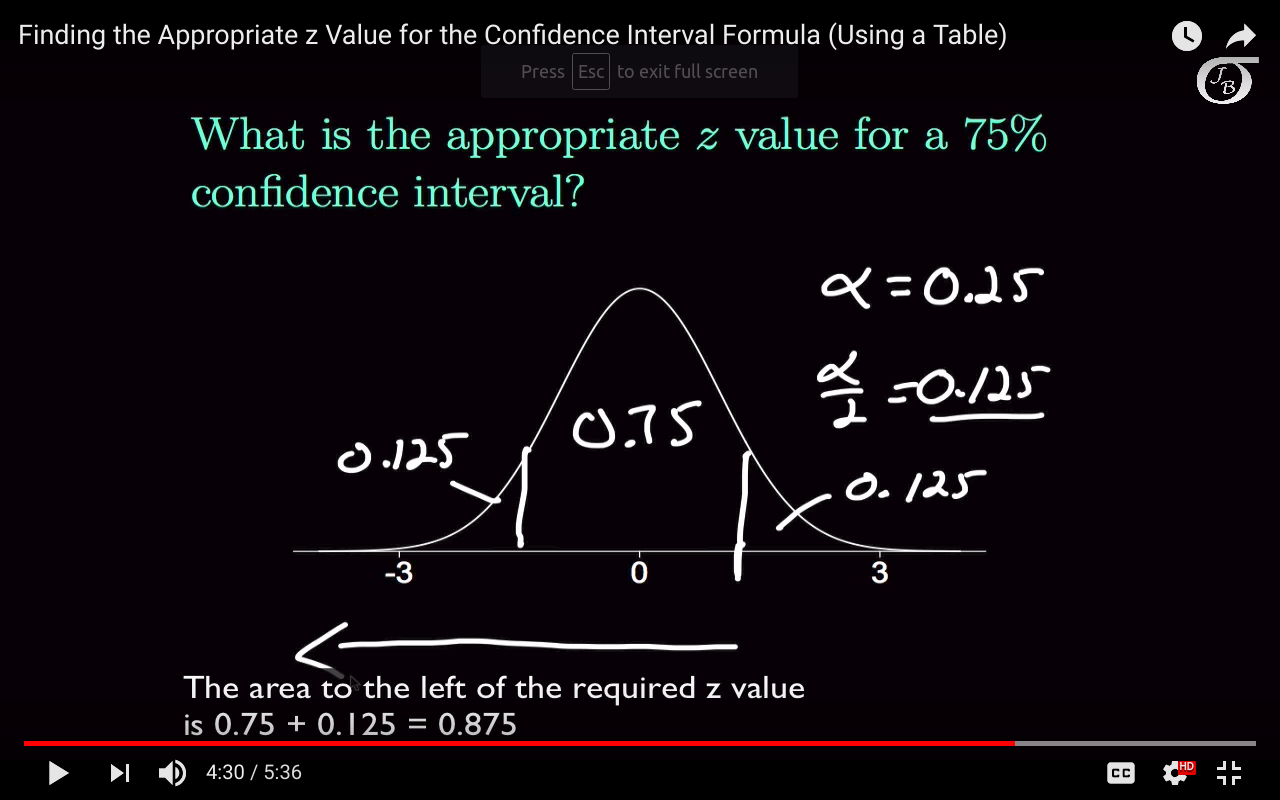
[1] 2.326348

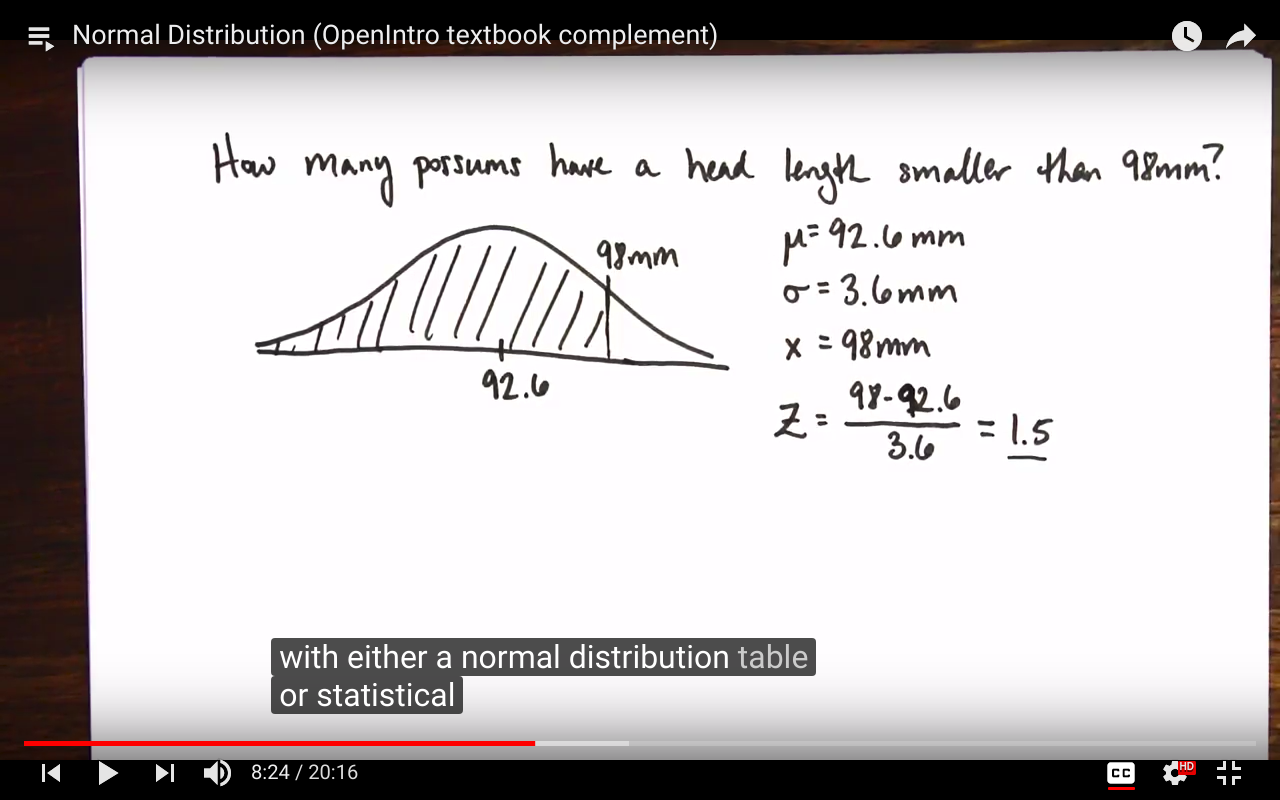
> qnorm(0.01)

[1] -2.326348

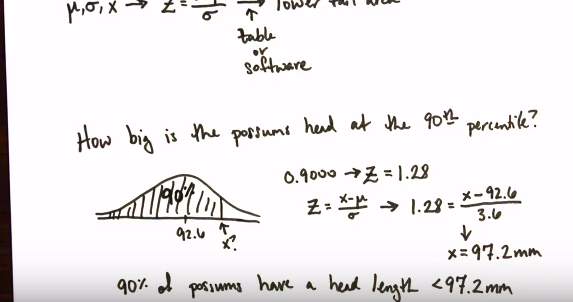
Question: if the 95% CI is (128,147) then determine the margin of error and sample mean.

The sample mean is the avg of CI i.e (128/147)/2 while margin of error is half the wiidth of the interval i.e (147-128)/2





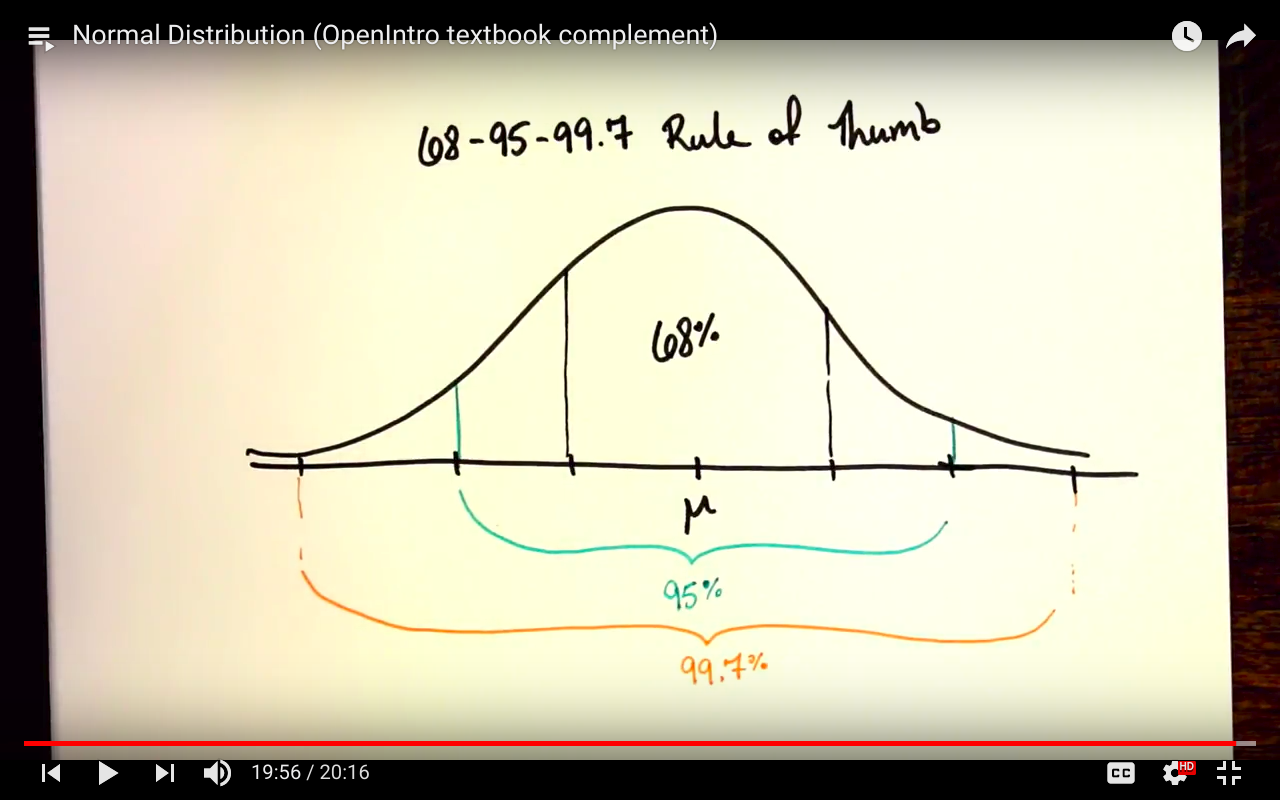
It can be computed using the z table or R command i.e > pnorm(98,92.6,3.6)



R command : > qnorm(0.9,92.6,3.6)

pnorm gives us the area under the curve which is probability

We can also give the zscore in pnorm to find the area. pnrom(zscore)=area



ztable gives us the same thing as pnorm the values in z table are the prcentiles(% of observations that fall below a given data point)

> pnorm(1800,1500,300)-pnorm(1200,1500,300)

[1] 0.6826895

this is validating that 68% of the data is withing 1 std

qnorm gives us the cutoff value at a given percentiles

Qnorm is used for cutoff value we give it percentile we are interested in and then pass the mean and std.

Question:shuru k 95% test scores or the cutoff values at 2 sd on either side.

> qnorm(0.975,1500,300)

[1] 2087.989

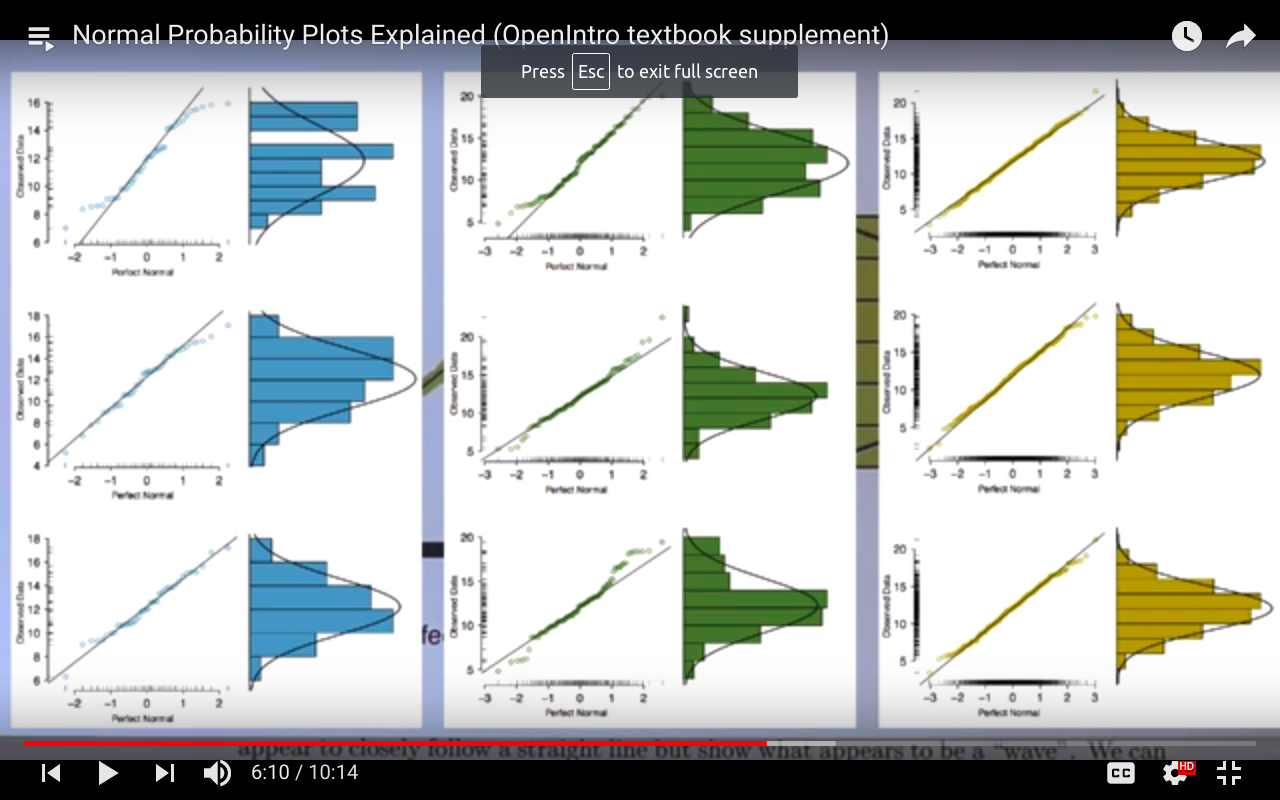
> qnorm(0.025,1500,300)

[1] 912.0108

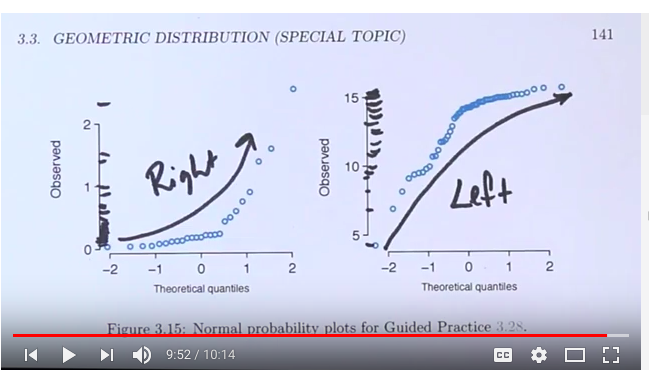
We can also pass the percentile in qnorm to get the zscore: qnorm(percentile)=zscore.

qnorm me agar sirf percentile ka argument hoga to it will assume it as std normal

*Deviations matters a lot in normal dist.*



The curvature of the normal probability plot shows weather it is either right/positive or left/negative skewed.



sampling dist ka std = S.E

deviation of mean of sampling dsit from pop mean #we don't know he mean of

> nosim=1000

> n=50

> sd(rowMeans(matrix(rnorm(nosim\*n),nrow=nosim)))

[1] 0.1443383

> 1/sqrt(50) #s.e=std of pop/sqrt(n)

[1] 0.1414214

CLT= if the sample size is big enough then the sampling dist of the sample mean will follows normal dist

library(openintro)

library(statsr)

data('yrbss)

pop\_mean=mean(yrbss$physically\_active\_7d,na.rm=T)

pop\_sd=sd(yrbss$physically\_active\_7d,na.rm=T)

samp\_param=yrbss%>% rep\_sample\_n(size=100,reps = 15000,replace=T) %>% summarise(avgactivity=mean(physically\_active\_7d ,na.rm=T),sigma=sd(physically\_active\_7d ,na.rm=T))

write the code in coding area and then highlight it and then press ctrl+enter

samp\_params

mean(samp\_params$avgactivity)

ggplot(data=samp\_params,aes(x=avgactivity))+geom\_histogram()

hypothesis testing:

for hypothesis testing we use p value , first we assume that the null hypothesis is true then if the mean is given let’s

we dont have to create CI everytime for hypothesis testing p value is enough.

4.23 Nutrition labels. The nutrition label on a bag of potato chips says that a one ounce

(28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams

of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard

deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate

measure of calories in the bags of potato chips? We have verified the independence, sample size,

and skew conditions are satisfied.

Solution in R:

H 0 : μ = 130. H A : μ 6 = 130

> (134-130)/(17/sqrt(35))

[1] 1.392019

> pnorm(1.39)

[1] 0.9177356

> 1-pnorm(1.39)

[1] 0.08226444

> 2\*0.082

[1] 0.164

The data do not provide convincing evidence that the true average calorie content in bags of potato chips is di↵erent than 130 calories.

4.25 Waiting at an ER, Part III. The hospital administrator mentioned in Exercise 4.13

randomly selected 64 patients and measured the time (in minutes) between when they checked in

to the ER and the time they were first seen by a doctor. The average time is 137.5 minutes and

the standard deviation is 39 minutes. She is getting grief from her supervisor on the basis that

the wait times in the ER has increased greatly from last year’s average of 127 minutes. However,

she claims that the increase is probably just due to chance.

(a) Are conditions for inference met? Note any assumptions you must make to proceed.

(b) Using a significance level of ↵ = 0.05, is the change in wait times statistically significant? Use

a two-sided test since it seems the supervisor had to inspect the data before she suggested an

increase occurred.

(c) Would the conclusion of the hypothesis test change if the significance level was changed to

↵ = 0.01?

**solution in R:**

**H 0 : μ = 127. H A : μ != 127.**

> (137.5-127)/(39/8)

[1] 2.153846

> pnorm((137.5-127)/(39/8))

[1] 0.9843739

> 1-0.9843739

[1] 0.0156261

> 2\*0.0156

[1] 0.0312

**CI approach for above question:**

ci=137.5+c(1,-1)\*1.96\* 39/sqrt(64)

> ci

[1] 147.055 127.945

the ci doesnt contain 127.5 therefore we can reject our null hypothesis.