QSCI 482 Story 2

Simon Hans Edasi

2023-10-16

Question 1

a. Read in the data, extract the appropriate column containing the diameters, and make sure that the data you have read in seem reasonable before proceeding.

```
firs <- read.csv('DouglasFirs 2022.csv')
diameters <- firs$diameters
# print(firs)
# plot(firs$treecode, firs$diameters)
# hist(diameters)
# plot(density(diameters))
```

b. Find the sample mean, sample standard deviation, and sample size.

```
dia_mean <- signif(mean(diameters),3)
dia_std <- signif(sd(diameters), 3)
dia_n <- signif(length(diameters), 3)
library(glue)
glue('Mean: {dia_mean}; STD: {dia_std}; Sample Size: {dia_n}')
```

```
## Mean: 95.9; STD: 26.1; Sample Size: 65
```

c. Assuming that the known population SD applies to the sample of trees, calculate the standard error, the 90% confidence interval, 95% confidence interval, and 99% confidence interval for the sample.

```
dia_std <- 20.8

dia_se <- dia_std / sqrt(dia_n)
Z <- qnorm(0.95)
ci190 <- signif(dia_mean - Z*dia_se, 3)
ciu90 <- signif(dia_mean + Z*dia_se, 3)

Z <- qnorm(0.975)
ci195 <- signif(dia_mean - Z*dia_se, 3)
ciu95 <- signif(dia_mean + Z*dia_se, 3)

Z <- qnorm(0.995)
ci199 <- signif(dia_mean - Z*dia_se, 3)
ciu99 <- signif(dia_mean + Z*dia_se, 3)
glue('90% CI [{ci190},{ciu90}]; 95% CI [{ci195},{ciu95}]; 99% CI [{ci199},{ciu99}]')
```

```
## 90% CI [91.7,100]; 95% CI [90.8,101]; 99% CI [89.3,103]
```

d. Conduct a hypothesis test that calculates the probability that the sample comes from a population of 100-year-old trees. Specify your null and alternative hypotheses, outline the test, note whether the test is one-tailed or two-tailed, calculate the test statistic, and report the resulting p-value.

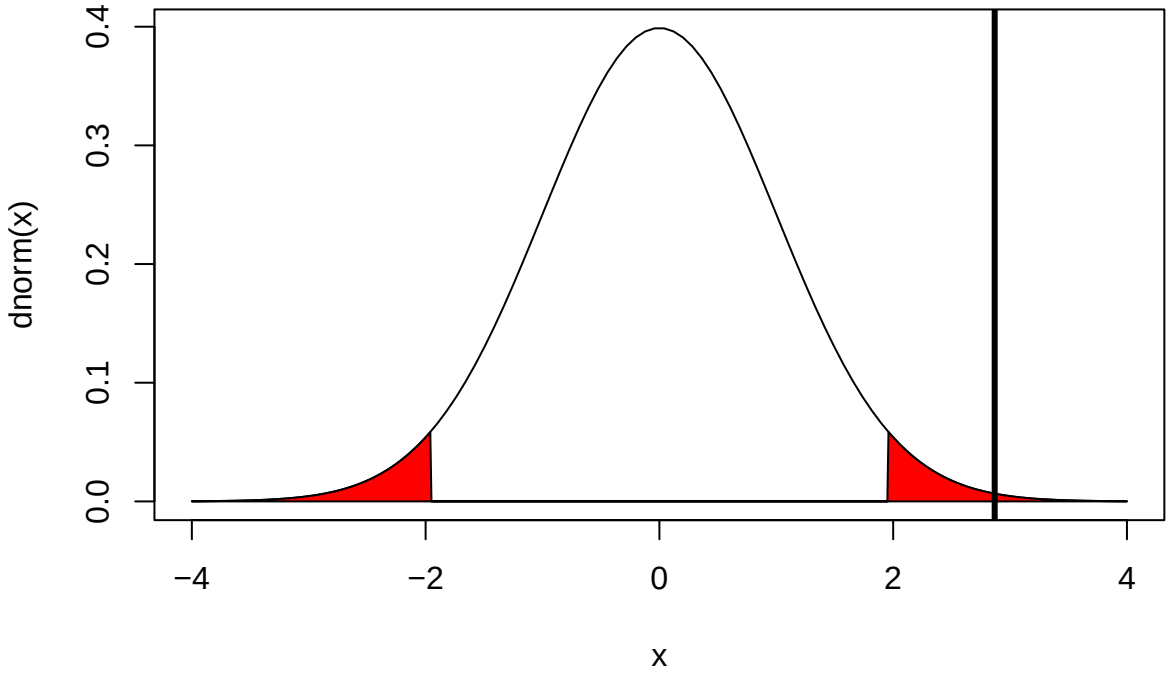
We want to test if our trees are 100 years old which is determined by a diameter of 88.5 cm.

Our null hypothesis is $H_0 : \bar{X} = 88.5$

Our alternative hypothesis is $H_A : \bar{X} \neq 88.5$, which will involve a two-tailed test where because we will reject the null if the mean is higher or lower than our population. We don't know the population standard deviation so we will need to use a t-test instead of a Z-test.

Sticking with a standard confidence of 95%, choose $\alpha = 0.05$.

```
alpha = 0.05
# define rejection regions
rej_left <- qnorm(alpha/2, 0,1)
rej_right <- qnorm(alpha/2, 0,1, lower.tail = FALSE)
#plot the curve with rejection regions
x <- seq(-4,4,length.out = 100)
plot(
  x,
  y = dnorm(x),
  type = 'l')

## x and y for the whole area
xReject <- c(seq(-4,4,by=0.01))
yReject <- dnorm(xReject, 0, 1)

yReject[xReject > rej_left & xReject < rej_right] <- 0

polygon(c(xReject,xReject[length(xReject)],xReject[1]),
        c(yReject,0, 0), col='red')

mu <- 88.5                              # null hyp pop mean
dia_se <- dia_std / sqrt(dia_n)         # standard error
dia_Z <- (dia_mean - mu) / dia_se       # Z statistic
abline(v = dia_Z, col = 'black', lwd = 3)
```



What is our p-value? Probability of observing our Z_obs or a more extreme value.

```
p_left <- pnorm(dia_Z,0,1,lower.tail = FALSE)      # left rejection region

p <- 2 * p_left                            # p value for a two tail test
print(p)
```

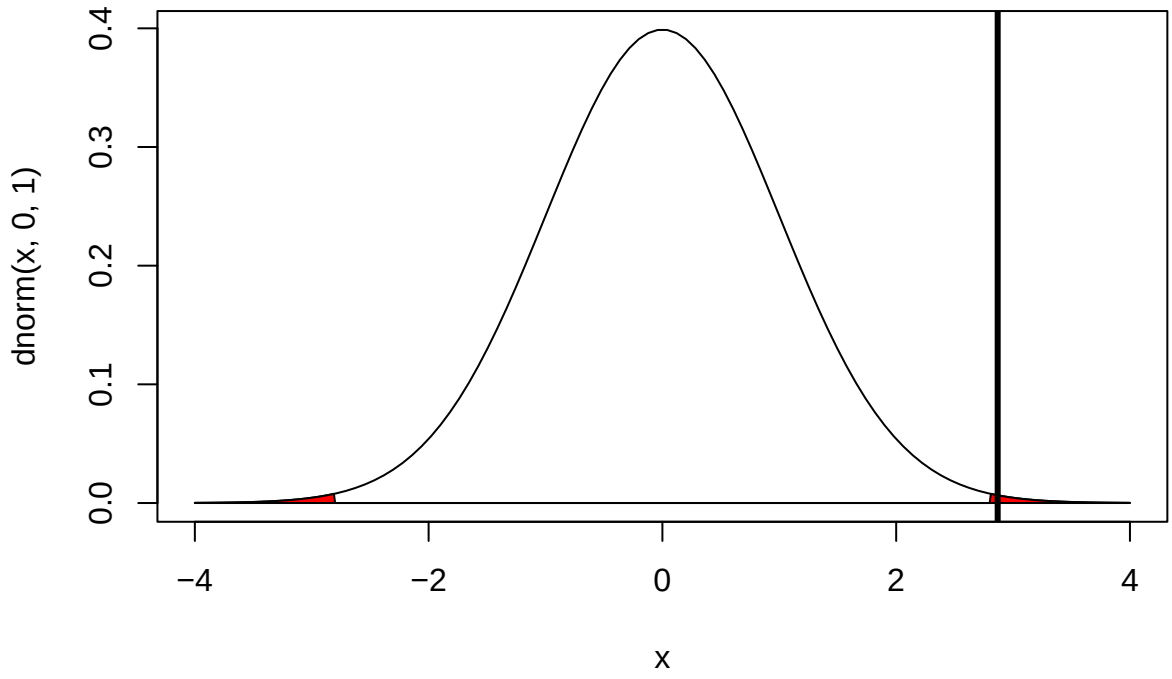
```
## [1] 0.004126798
```

e. If the significance level, α, is 0.05, would you reject the null hypothesis? What if the chosen significance level is 0.005? Given the test results, are the trees on the property owner's land about 100 years old, or are they significantly older or younger?

With $\alpha = 0.05$ we should reject the null hypothesis. For a significance level of 0.005 we can easily repeat the analysis and see that the test statistic falls outside of the rejection region.

```
alpha = 0.005
# define rejection regions
rej_left <- qnorm(alpha/2, 0,1)
rej_right <- qnorm(alpha/2, 0,1, lower.tail = FALSE)
mu <- 88.5                                 # null hyp pop mean
dia_se <- dia_std / sqrt(dia_n)            # standard error
dia_Z <- (dia_mean - mu) / dia_se          # Z statistic

#plot the curve with rejection regions
x <- seq(-4,4,length.out = 100)
plot(
  x,
```

```
  y = dnorm(x,0,1),
  type = 'l')
## x and y for the whole area
xReject <- c(seq(-4,4,by=0.01))
yReject <- dnorm(xReject,0,1)

yReject[xReject > rej_left & xReject < rej_right] <- 0

polygon(c(xReject,xReject[length(xReject)],xReject[1]),
        c(yReject,0, 0), col='red')

# plot test statistic
abline(v = dia_Z, col = 'black', lwd = 3)
```




```
# p_left <- pt(-dia_t, df = dia_dof)      # left rejection region
p_left <- pnorm(dia_Z, 0, 1, lower.tail =FALSE )
p <- 2 * p_left                          # p value for a two tail test
print(signif(p,2))
```

```
## [1] 0.0041
```

Given these results, I would conclude the trees are significantly older than 100 years. We have a positive Z value so we know our mean is higher than the population mean.

With a lower p-value we still reject the null hypothesis.

Question 2

Assume now that you do not know the population standard deviation of diameters for secondary growth trees, nor is my sample size large enough that the standard deviation of my sample can be assumed to be the same as the standard deviation of the population.

Conduct a statistical test, using a similar sequence of steps to question 1, to test whether the sample of trees on my property could be secondary growth, if secondary growth trees (the result of a single logging event) have a mean diameter of 88.5 cm. Explain your choice of statistical test, outline the steps involved, highlight your assumptions, report the 95% confidence interval from this sample, test statistic, p-value, and whether you accepted or rejected the null hypothesis, and finally answer the question above.

We do not know the population standard deviation and our sample size is small so we will need to use a t-test to test our null hypothesis.

$H_0 : \mu = 88.5$

$H_A : \mu \neq 88.5$

We will use a two tail test because we are determining if the means are equal. We will calculate a t-statistic for our sample called t_obs, and compare it to a critical value defined by $\alpha = 0.05$. To use a t distribution we will need to calculate degrees of freedom.

```
rm(list = ls())
diameters <- read.csv('MyDouglasFirs.csv')$diameter
```

```
dia_mean <- signif(mean(diameters))
dia_std <- (sd(diameters))
dia_n <- (length(diameters))
dia_se <- (dia_std / sqrt(dia_n))
dia_dof <- (dia_n - 1)
library(glue)
t <- qt(0.975,df = dia_dof)
cil95 <- signif(dia_mean - t*dia_se, 3)
ciu95 <- signif(dia_mean + t*dia_se, 3)

glue(' 95% CI [{cil95},{ciu95}]')
```

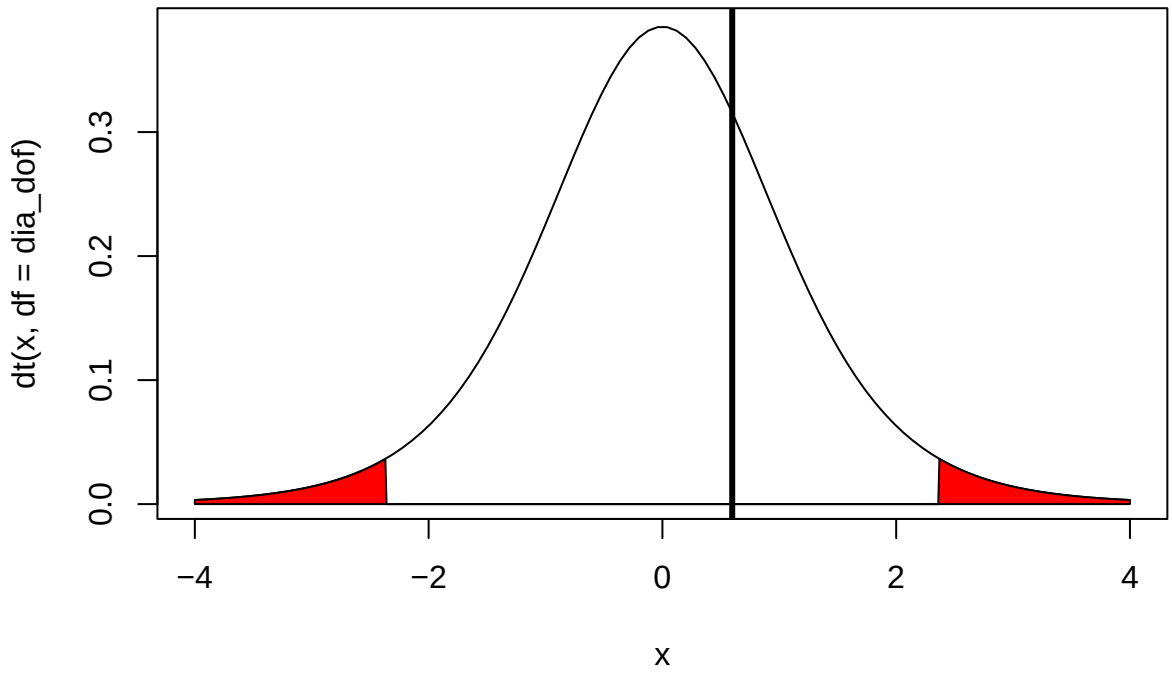
```
##  95% CI [63.7,130]
```

```
glue('Mean: {signif(dia_mean,3)}; STD: {signif(dia_std,3)}; Sample Size: {signif(dia_n,3)}, se: {signif
```

```
## Mean: 96.9; STD: 39.7; Sample Size: 8, se: 14 dof: 7
```

```
alpha = 0.05
# define rejection regions
rej_left <- qt(alpha/2,df = dia_dof)
rej_right <- qt(alpha/2,df = dia_dof,lower.tail = F)
#plot the curve with rejection regions
x <- seq(-4,4,length.out = 100)
plot(
  x,
  y = dt(x,df = dia_dof),
  type = 'l')

## x and y for the whole area
xReject <- c(seq(-4,4,by=0.01))
yReject <- dt(xReject, df = dia_dof)
```

```
yReject[xReject > rej_left & xReject < rej_right] <- 0

polygon(c(xReject,xReject[length(xReject)],xReject[1]),
        c(yReject,0, 0), col='red')

mu <- 88.5                              # null hyp pop mean
dia_se <- dia_std / sqrt(dia_n)         # standard error
t <- (dia_mean - mu) / dia_se       # Z statistic
abline(v = t, col = 'black', lwd = 3)
```




```
print(t)
```

```
## [1] 0.5972662
```

Calculate p-value, the probability of observing this value or a more extreme t_obs.

```
# p_left <- pt(-dia_t, df = dia_dof)      # left rejection region
p_left <- pt(t, df = dia_dof, lower.tail = F )
# p <- 2 * p_left                         # p value for a two tail test
print(signif(p_left,2))
```

```
## [1] 0.28
```

p $>$ α, so we do not reject the null hypothesis and accept that this sample is from an old growth forest. Yes, the trees are 100 years old.

Question 3

a. Calculate the 95% CI for the population SD, from the sample of data.

```
rm(list = ls())
weights <- read.csv('scalereadings.csv')$Weight

samp_n <- length(weights)
samp_mean <- mean(weights)
samp_std <- sd(weights)
samp_var <- var(weights)

samp_dof <- samp_n - 1

pop_sig <- 0.2
alpha <- 0.05


chisq_l <- qchisq(alpha/2,df = samp_dof,lower.tail = F)
chisq_u <- qchisq(1-alpha/2,df = samp_dof,lower.tail = F)
chisqci_l <- (samp_dof * samp_var) / chisq_l
chisqci_u <- (samp_dof * samp_var) / chisq_u

glue('95% CI for sample STD [{round(sqrt(chisqci_l),3)},{round(sqrt(chisqci_u),3)}]')
```

```
## 95% CI for sample STD [0.225,0.45]
```

b. Conduct a statistical hypothesis test to detect whether the SD of measurements is equal to the SD promised by the manufacturer, for $\alpha = 0.05$. Outline the null and alternative hypotheses, the choice of test and whether it is one-tailed or two-tailed, and report the sample SD, the value of the test statistic, the p-value, and whether the null hypothesis is accepted or rejected. Does the scale conform to the manufacturer's promises of accuracy?

H_0 : Sample STD $= 0.2$

H_A : Sample STD $\neq 0.2$

We will be using a two tailed χ^2 test to determine if our variance is equal to the population.

```
alpha = 0.05
# define rejection regions
chi_crit_left <- qchisq(alpha/2,df = samp_dof)
chi_crit_right <- qchisq(1-alpha/2,df = samp_dof)
glue('{round(chi_crit_left,2)}, {round(chi_crit_right,2)}')
```

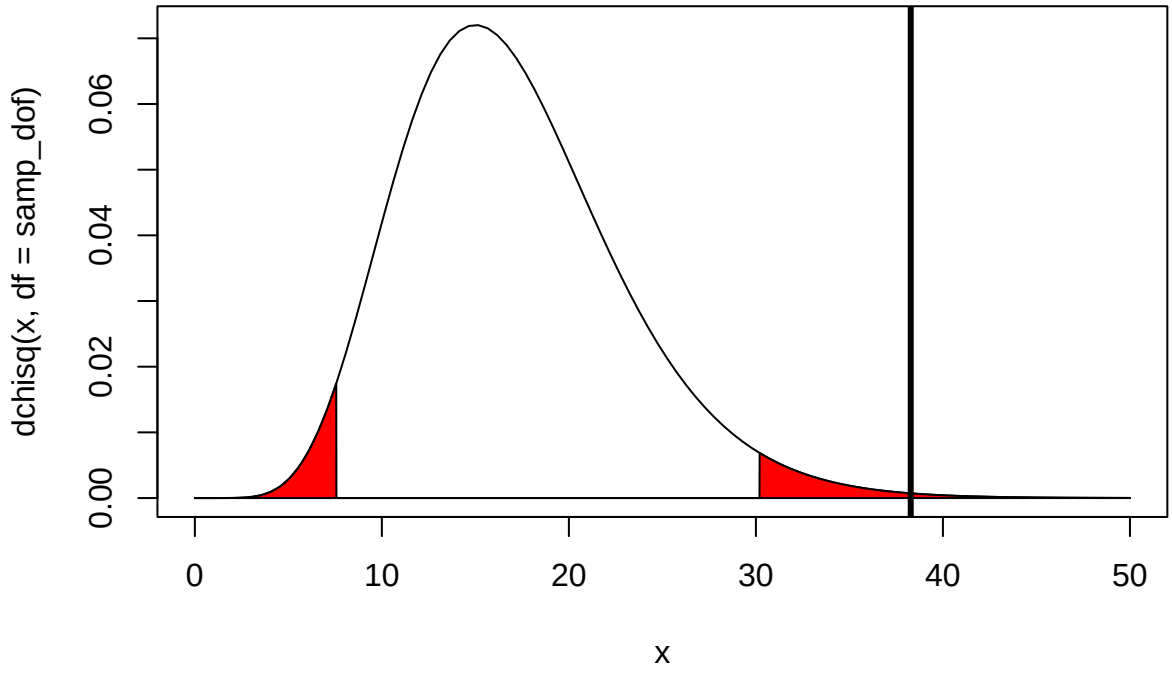
```
## 7.56, 30.19
```

```
#plot the curve with rejection regions
x <- seq(0,50,length.out = 100)
plot(
  x,
  y = dchisq(x,df = samp_dof),
  type = 'l')

## x and y for the whole area
xReject <- c(seq(0,50,by=0.01))
yReject <- dchisq(xReject, df = samp_dof)
```

```
yReject[xReject > chi_crit_left & xReject < chi_crit_right] <- 0

polygon(c(xReject,xReject[length(xReject)],xReject[1]),
        c(yReject,0, 0), col='red')
chi_sq <- (samp_dof * samp_var) / pop_sig^2

# SS <- sum(weights - mean(weights))^2
#
# chi_sq <- SS / pop_sig^2

print(chi_sq)
```

```
## [1] 38.27778
```

```
abline(v = chi_sq, col = 'black', lwd = 3)
```




```
# calculate p-value
# what is the probability of observing our statistic or more extreme value?
p_left <- 1 - pchisq(chi_sq,df = samp_dof)
p_right <- 1 - pchisq(chi_sq, df = samp_dof,lower.tail = F)
print(p_left)
```

```
## [1] 0.002250574
```

```
print(p_right)
```

```
## [1] 0.9977494
```

```
glue('alpha = 0.05, p-value = {round(2*p_left,3)}')
```

```
## alpha = 0.05, p-value = 0.005
```

Our χ^2 value falls within our rejection region and our p-vaue is less than our selected α. We reject the null hypothesis that the scale variance is similar to the manufacturer reported variance. No, the scale does not conform to manufacturer variance standards.