```r
#Daniel Williams
#21203054
#Intro to Data Analytics
#Assignment 2

#QUESTION 1 i
x1 <- c(22.67, 19.91, 24.77, 23.77, 24.41, 21.91, 21.37, 21.96)
x2 <- c(21.35, 19.19, 21.92, 24.25, 24.15, 21.61, 20.77, 20.71)
n1 <- 8
n2 <- 8
D_0 <- 0
alpha <- 0.05 #5% significance level
# H0  x1_bar - x2_bar <= 0
# H1  x1_bar - x2_bar > 0
# X1 normally distributed with mean mu_x1, and
# X2 normally distributed with mean mu_x2, and
# where both distributions are parameterised by the unknown variance varx.
# conduct a one sided, two sample t test
# interested in the right hand tail of the students t distribution with 8 + 8
- 2 = 14 DOF.
t_crit <- qt(1- alpha, df = 14)
# check with tables, with DOF 14, at 0.95 level, t_crit = 1.7613
# tables show 1.7 ~ 0.9444, 1.8 ~ 0.9533, R gives 1.7613

# now calculate the test statistic
x1_bar <- mean(x1)
# x1_bar is 22.59
# Calculated by (22.67+19.91+24.77+23.77+24.41+21.91+21.37+21.96) / 8
x2_bar <- mean(x2)
# x2_bar is 21.74
# Calculated by (21.35+19.19+21.92+24.25+24.15+21.61+20.77+20.71) / 8

var_x1 <- var(x1)
var_x2 <- var(x2)
# Variances calculated by the sum of the differences between the value of xi,
and x1_bar
# eg (22.67 - 22.59)^2 + (19.91 - 22.59)^2 + (24.77 - 22.59)^2 + ... etc for
rest of X1 elements
# sum of the square of the differences is 18.99
# 18.99 / n-1.  (18.99) / 7 = 2.71 for var_x1
# same process used to calculate var_x2 using vector x2, and x2_bar
# sum of the square of the differences is 20.81
# 20.81 / 7  = 2.97 for  var_x2

# Calculating the pooled sample variance
s_p <- sqrt(((n1 - 1) * (var_x1) + (n2 - 1) * (var_x2)) / (n1 + n2 - 2))
# pooled sample variance is s_p^2,  s_p is 1.686, formula shown above
# now calculate the test statistic for the observed data

t_star <- ((x1_bar - x2_bar - D_0) / s_p) * sqrt((n1 * n2) / (n1 + n2))
# t_star 1.01, forumla shown above

ifelse(t_star > t_crit, "Reject H_0", "Fail to reject H_0")
# t_star < t_crit therefore we fail to reject H0
```

```
# don't have sufficent evidence at the 5% significance level to conclude that
the
# average running times are lower after the series of training

P_val <- 1 - pt(t_star, df = 14)
# p_val is 0.164 > alpha (0.05), backs up fail to reject H0 decision.
# check with tables at t=1 DOF 14, yes P value 0.832 (1-0.833) = 0.167



# QUESTION 1 ii

# conducting a paired t test
n <- 8
x1 <- c(22.67, 19.91, 24.77, 23.77, 24.41, 21.91, 21.37, 21.96)
x2 <- c(21.35, 19.19, 21.92, 24.25, 24.15, 21.61, 20.77, 20.71)
alpha <- 0.05
d <- x1 - x2
# we assume that d is normally distributed with mean d_bar, and variance var_d
# Hypothesis H0: x1_bar - x2_bar = d_bar <= 0
# Hypothesis H1: x1_bar - X2_bar = d_bar > 0

d_bar <- mean(d)
# takes the elements of the vector d, and divides by the total number of
readings
# (1.32 + 0.72 + 2.85 + -0.48 + 0.26 + 0.30 + 0.60 + 1.25) / 8
# (6.82 / 8) = 0.8525   = d_bar
d_sd <- sd(d)
# standard deviation calculated by the square room of the sum of the
differences between the value of d, and d_bar
# eg (1.32 - 0.8525)^2 + (0.72 - 0.8525)^2 + (2.85 - 0.8525)^2 + ... etc for
rest of d elements
# sum of the square of the differences is 6.879
# 6.879 / n-1.  (6.879) / 7 = 0.982 for var(d)
# square root 0.982 = 0.991 for d_sd

# defining the rejection region, one tailed test
t_crit <- qt(1 - alpha, df = 7)
# check with tables, students t distribution, 0.95, DOF 7 (n-1), t_crit =
1.895

# calculating the observed test statistic
t_star <- sqrt(n) * (d_bar / d_sd)
#t_star 2.43, formula shown above

ifelse(t_star > t_crit, "Reject H_0", "Fail to reject H_0")
# t_star > t_crit so we reject H0

p_val1b <- (1 - pt(t_star, df = 7))
# p_val1b is 0.0226, less than alpha = 0.05, backs up reject H0
# validate from tables with 2.43, and DOF 7  value is 0.9774  (1-0.9774), p
value 0.0226
# yes this differs from q1i because we have now assumed that the readings are
paired, our
```

```
# DOF have reduced from 14 to 7, and different test statistic is used. The
samples are now considered dependent
# pooled variance is larger than sample variance for the differences, pairing
reduces variability that may
# otherwise obscure small mean differences, paired design is more efficient
that independent samples design.


# QUESTION 1 iii
# confidence interval at 95% level for d
# formula shown below, students t distribution with 0.95, and DOF 7 (n-1  8-1)
# d_bar +_  t(7, 0.95) (variance of d / n)^0.5

Upper <- d_bar + (t_crit * ((d_sd^2) / 8)^0.5)
# Upper = (0.8525) + (1.894 * ((0.991^2) / 8)^0.5)
# Upper = 1.516
# all values calculated in part ii

Lower <-  d_bar - (t_crit * ((d_sd^2) / 8)^0.5)
# Lower = (0.8525) - (1.894 * ((0.991^2) / 8)^0.5)
# Lower = 0.188

# interval [0.188, 1.516] we are 95% confident that this interval covers the
mean difference
# in 5k times after 6 weeks of training
# note this does not include 0, which backs up my answer to 1ii


# QUESTION 2....

x1 <- c(18.19, 16.98, 19.97, 16.98, 18.19, 15.99, 13.79, 15.9, 15.9, 15.9,
15.9, 15.9, 19.97, 17.72)
x2 <- c(10.5, 12, 9.54, 10.55, 11.99, 9.3, 10.59, 10.5, 10.01, 11.89, 11.03,
9.52, 15.49, 11.02)

x1_bar <- mean(x1)
# x1_bar =
(18.19+16.98+19.97+16.98+18.19+15.99+13.79+15.9+15.9+15.9+15.9+15.9+19.97+17.72)/
14
# x1_bar = (237.28 / 14) = 16.949

x2_bar <- mean(x2)
# x2_bar =
(10.5+12+9.54+10.55+11.99+9.3+10.59+10.5+10.01+11.89+11.03+9.52+15.49+11.02) /
14
# x2_bar = (153.93 / 14) = 10.995

var_x1 <- var(x1)
var_x2 <- var(x2)
# Variances calculated by the sum of the differences between the value of xi,
and x1_bar
# eg (18.19 - 16.949)^2 + (16.98 - 16.949)^2 + (19.97 - 16.949)^2 + ... etc
for rest of X1 elements
# sum of the square of the differences is 38.33
```

```
# 38.33 / n-1.  (38.33) / 13 = 2.948 for var_x1
# same process used to calculate var_x2 using vector x2, and x2_bar
# sum of the square of the differences is 31.995
# 31.995 / 13  = 2.461 for  var_x2

sd_x1 <- sqrt(var_x1)
sd_x2 <- sqrt(var_x2)
# to find the standard deviations of each dataset, we take the square root of
the variance
# sd_x1 = 1.717
# sd_x2 = 1.569

cv_x1 <-  sd_x1 / x1_bar
cv_x2 <-  sd_x2 / x2_bar
# to find the coefficient of variation for a dataset we divide the standard
deviation by the mean
# cv_x1 = 1.717 / 16.949 = 0.101 can do this x100 to give 10.1
# cv_x2 = 1.569 / 10.995 = 0.143 can do this x100 to give 14.3


# Question 2 ii
# one sample hypothesis test for the population mean, assuming data is normal
# testing normal population with unknown variance
# let X1 be the sales price (online - no auction), for the book
# assume that X is normally distributed with mean mu, and standard deviation
sigma.
# Hypothesis  H0:  mu = 16.95    H1: mu != 16.95
# significance level alpha = 0.05
# finding the critical values, two tailed test
n <- 14
alpha <- 0.05
t_crit <- qt(1-(alpha/2), df = n-1)
# t critical is 2.16, checking with tables degrees of freedom 13 (n-1), P =
2.5. Yes 2.16
# we will reject H0 if the modulus of test statistic is greater than this
critical value
# calculating the test statistic , take the x1_bar and sd_x1 from q2i , mu0
from the question
mu0 = 16.95
t_star <- sqrt(n) * ((x1_bar - mu0) / sd_x1)
# t star = -0.003,
# we find mod t star = 0.003, which is not greater than t_crit. therefore we
fail to reject H0.
pval <- pt(t_star, df = n-1)
# p value is 0.49, greater than alpha, backs up decision - fail to reject H0.
# there is insufficient evidence to state at 5% significance level
# that at the population mean is not 16.95


# Question 2 iii
# using formula from week 6, CI with pop sigma unknown
# taking x1_bar, alpha, n, sd_x1 - from the questions above

Upper <- x1_bar + qt(1-(alpha/2), df = n-1) * sd_x1 / sqrt(n)
```

```
Lower <- x1_bar - qt(1-(alpha/2), df = n-1) * sd_x1 / sqrt(n)

# This gives the 95% confidence interval for mu as [15.96, 17.94]
# we are 95% confident that the true average sales price lies between [15.96,
17.94]
# in Q2ii, we were testing WRT to mu0 being 16.95, this lies within the
confidence interval
# therefore we were correct at the alpha significance level to fail to reject
H0.



# Question 2 iv
# Test auction, mu0 = 16.95
# one sample hypothesis test for population mean, assuming data is normal
# testing normal population with unknown variance
# let X2 be the sales price (online - auction), for the book
# assume that X is normally distributed with mean mu, and standard deviation
sigma.
# Hypothesis  H0:  mu >= 16.95    H1: mu < 16.95
# significance level alpha = 0.05
# finding the critical values, one tailed test

t_crit2 <- qt(alpha, df = n-1)
# t criticial is -1.77, checking with tables, DOF 13, P = 5,  yes -1.77, with
symmetry around 0
# we will reject H0 if observed test statistic is less than this value
# calculating the observed test statstic, taking values from previous
questions
# n = 14, x2_bar = 10.995, mu0 = 16.95, sd_x2 = 1.568

t_star2 <- sqrt(n) * ((x2_bar - mu0) / sd_x2)
# t_star2 is -14.20, which is less than t_crit2 (-1.77), therefore at the 0.05
significance level
# we reject H0, the mean price online (auction), is less that 16.95

pval2 <- pt(t_star2, df = n-1)
# pval2 is 1.35e-9, which is less than 0.05, backing up our rejection of H0.
# We can't check this value in the tables



# Question 2v
# Assume variances are equal, two sample test, normally distributed
populations with
# unknown equal variance.
# Assume x1 (no auction) is normally distributed with mean mu_x1, and standard
deviation sigma
# Assume x2 (auction) is normally distributed with mean mu_x2, and standard
deviation sigma
# H0:  mu_x1 <= mu_x2       H1:  mu_x1 > mu_x2
# that is, that the alternate hypothesis is the no-auction prices are higher
than the auction prices
# conducting a one sided test, case (b) in notes with D0 = 0
```

```
D0 <- 0
n1 <- 14
n2 <- 14
# significance level alpha = 0.05
# defining rejection region
t_crit3 <- qt(1-alpha, df = n1 + n2 - 2)
# t_crit3 1.705, check using tables, P 5, v 26, 1.706. Ok verified.
# we will reject H0 if the observed test statistic is greater than this

# constructing test statistic - must estimate pop variance from pooled sample
variance
# taking variances for x1, and x2, from questions above
# var_x1 = 2.94,   var_x2 = 2.46

var_pool <- (((n1 - 1) * (var_x1) + (n2 - 1) * (var_x2)) / (n1 + n2 - 2))
sd_pool <- sqrt(var_pool)

# var_pool is 2.7, sd_pool is 1.644

#calculating observed test statistic, values from above
t_star3 <- ((x1_bar - x2_bar - D0) / sd_pool) * sqrt((n1 * n2) / (n1 + n2))
# t_star3 is 9.57, we find t_star3 > t_crit3,  9.57 > 1.705
# therefore we reject H0, at the 95% confidence level the mean non-auction
price
# is higher than the auction price.

p_val3 <- (1-pt(t_star3, df = n1+n2-2))
# p_val3 is 2.5e-10, therefore less than alpha =0.05, backs up reject H0.
# can't check this value on tables, too large t


# QUESTION 3.
# Test using chi sq goodness-of-fit.
# We have 5 categories - test for uniform distribution therefore
# pw = pr = po = py = pg = 0.2,  k=5 categories

# Set hypothesis H0: pw = pr = po = py = pg = 0.2  H1: that H0 is false
# Define variables, and actual results recorded

k <- 5
pw <- 0.2
pr <- 0.2
po <- 0.2
py <- 0.2
pg <- 0.2

w <- 222
r <- 279
o <- 251
y <- 232
g <- 266

#Need to calculate total jelly beans, to then calculated er (expected red),
etc..
```

```
total <- w + r + o + y + g
# total is 1,250 beans

ew <- total * pw
er <- total * pr
eo <- total * po
ey <- total * py
eg <- total * pg

# all expected values are 250 beans, ew = er = eo = ey = eg = 250
# all n(i), are greater than or equal to 5, which satisfies the large sample
req
# for the chi sq goodness of fit test
# calculating the chi-sq values for each category
# observed value minus expected value, squared, divided by expected value

Chi_w <- (w - ew)^2 / ew
Chi_r <- (r - er)^2 / er
Chi_o <- (o - eo)^2 / eo
Chi_y <- (y - ey)^2 / ey
Chi_g <- (g - eg)^2 / eg

#chi_w = 3.13, chi_r = 3.36, chi_o = 0.004, chi_y = 1.30, chi_g = 1.02

#adding the chi individual elements together to get to a total figure
chi_tot <- Chi_w + Chi_r + Chi_o + Chi_y + Chi_g
#chi_tot = 8.82

# Need to optain a critical value for chi, using significance level of 0.05
# chi-sq crit with k-1 DOF (4), p = 5, from tables 9.488
chi_crit <- qchisq(0.95, df = 4)
# validated in R, chi_crit = 9.488

# we find  chi_tot < chi_crit  , therefore we fail to reject H0
# we do not have evidence at 5% significance level that JB are not filled
equally

# getting a P value, require chi <= 8.82, dof = 4, tables
# <8.5 is 0.925,   <9 is 0.939, both <0,95,
# 1-0.925 = 0.075    1-0.939 = 0.061, both greater than alpha = 0.05
# hence backs up decision to
# fail to reject H0 at the alpha = 0.05 significance level, check with R
p_chi <- 1 - pchisq(chi_tot, df = 4)
# p_chi is 0.065 > 0.05 alpha, therefore agrees fail to reject H0.
# advice for Sheila
# at the 95% confidence level we have failed to prove that the JB are not
filled equally
# At this confidence level level, I do not recommend that you complain


# you could lower your confidence level & that would statistically suggest you
have a point
chi_crit2 <- qchisq(0.93, df = 4)
```

```
# at the 93% confidence level we would get chi_crit2 to be 8.66, so we would
reject H0 at
# the 93% confidence level, so perhaps you should complain based on that
statistic.
```