```r
# Testing means between groups.

# 1. Comparison between two groups: Two sample t test
# Data are now from two groups, x_1,...,x_n and y_1,...,y_T.
# Assume these two groups follow N(mu1,sigma1^2) and N(mu2,sigma2^2)
# H0: Two groups have the same mean (mu1=mu2)
# t = (x_bar - y_bar)/(SEDM), SEDM = sqrt(sigma1_hat^2/n + sigma2_hat^2/T)

data.set <- read.csv("county_data.csv", stringsAsFactors = FALSE)
head(data.set)

data_1 <- data.set[data.set$State == "California" | data.set$State == "Connecticut",]
data_1$State <- factor(data_1$State, levels=c("California", "Connecticut"))
boxplot(data_1$Unemployment~data_1$State)

t.test(data_1$Unemployment[data_1$State=="California"],data_1$Unemployment[data_1$State=="Connecticut"], var.equal=FALSE)

# In textbooks, it is usually assumed that the variances of the two groups are the same.

t.test(data_1$Unemployment[data_1$State=="California"],data_1$Unemployment[data_1$State=="Connecticut"], var.equal=TRUE)

# 2. Comparison among more than two groups.
# Analysis of variance (ANOVA)
# Let xgi denote observation no. i in group g.
# We can decompose the observations as xgi = x_bar + (xg_bar - x_bar) + (xgi - xg_bar)
# xg_bar - x_bar: deviation of group mean from the population mean
# xgi - xg_bar: deviation of observation from the group mean
# The corresponding model   Xgi = mu + alpha_g +egi,  egi~N(0,sigma^2)
# From xgi - x_bar = (xg_bar - x_bar) + (xgi - xg_bar)
# Total variation = sum_g sum_i (xgi - x_bar)^2
# Within variation = sum_g sum_i (xgi - xg_bar)^2
# between variation = sum_g sum_i ng(xg_bar - x_bar)^2
# MSw = sum_g sum_i (xgi - xg_bar)^2/(n-G) is an estimate of sigma^2
# H0 : all the group means are the same.
# If H0 is true, MSb = sum_g sum_i ng(xg_bar - x_bar)^2/(G-1) is also the estimate of sigma^2

# F = MSb / MSw ~F(G-1, n-G)

data_2 <- data.set[data.set$State %in% c("California", "Connecticut", "Alabama", "Ohio"),]
data_2$State <- factor(data_2$State, levels=c("Alabama", "California","Connecticut","Ohio"))
str(data_2)

boxplot(data_2$Unemployment~data_2$State)
anova(lm(data_2$Unemployment~data_2$State))

# In the outcome: Residual is the within group variation, data_2$State is the between group variation.

# You can also do this test based on the regression coefficients.
reg <- lm(data_2$Unemployment~data_2$State)   # Categorial variables (factors) are used as dummies.
summary(reg)

# Pairwise comparison. Which pair of states have different means?

pairwise.t.test(data_2$Unemployment, data_2$State, p.adj="bonferroni")
# In multiple testing, use this to be conservative.

# Exercise

college <- read.csv("College.csv")

# College data: Demographic characteristics, tuition, and more for USA colleges.
# Private: Public/private indicator
# Apps: Number of applications received
# Accept: Number of applicants accepted
# Enroll: Number of new students enrolled
# Top10perc: New students from top 10 % of high school class
# Top25perc: New students from top 25 % of high school class
# F.Undergrad: Number of full-time undergraduates
# P.Undergrad: Number of part-time undergraduates
# Outstate: Out-of-state tuition
# Room.Board: Room and board costs
# Books: Estimated book costs
# Personal: Estimated personal spending
# PhD: Percent of faculty with Ph.D.'s
# Terminal: Percent of faculty with terminal degree
# S.F.Ratio: Student/faculty ratio
# perc.alumni: Percent of alumni who donate
# Expend: Instructional expenditure per student
# Grad.Rate: Graduation rate

# 1. Compare the distributions of "personal" between private school and public school. For this, you can first draw box plots and do 2 sample t test.
# 2. Divide the colleges into three groups based on Top10perc. Make the group size to be the same with each other.
#    Compare the mean of "Grade.Rate" among these four groups. If you conclude there is any difference, identify which pair
#    of groups have different means.
```