HYPOTHESIS TESTING

Winter 2019

Hypothesis Testing

- Hypothesis testing is a decision making process for **evaluating claims about a population.**

- The researcher must:
  - Define the population under study
  - State the hypothesis that is under investigation
  - Give the significance level
  - Select a sample from the population
  - Collect the data
  - Perform the statistical test
  - Reach a conclusion

Population on Samples

- Population: the entire collection of individuals about which information is sought
- Sample: subset of a population, containing the individuals that are actually observed

Population and Samples

Give at least three examples of a population

1.
2.
3.

For the population listed in 1., give an example of a sample from the population

Can you make up some hypothesis about the population in 1.

Hypothesis Tests

- Examples of hypothesis tests include t-test, Chi-Square, and correlation analysis to name a few
- To use this tool properly, you must understand the statistics
- Applying an incorrect test to a given set of data will give incorrect results

Hypothesis Testing

- Hypothesis testing is the formal statistical technique of collecting data to answer questions through the use of a statistical model.

- “In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the **significance level**.”

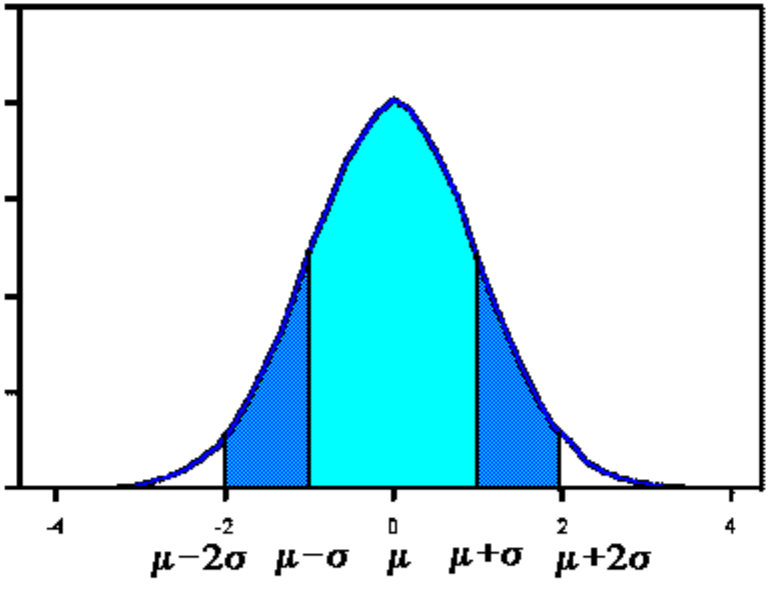
http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

NULL Hypothesis

- The Null Hypothesis refers to a general or default position – denoted H_0
- Null Hypothesis is assumed true until evidence indicate otherwise

The Normal Distribution

- The following Hypothesis Tests assume that the data is normally distributed.
- The standard normal curve in the picture has a mean of 0 and standard deviation of 1. A dataset with a normal distribution has about 68% of the observations within σ of the mean μ which in this case is (-1,1)



http://www.stat.yale.edu/Courses/1997-98/101/normal.htm

The Normal Distribution Continued

- About 95% of the observations will fall within 2 standard deviations of the mean (-2,2)
- About 99.7% of the observations will fall within 3 standard deviations of the mean
- Example: Consider 130 observations of body temperature with the results below. If the data is normal, what must be the case?

```
Variable          N     Mean   Median   StDev  Min      Max
BODY TEMP        130   98.249  98.300  0.733  96.300  100.800
```

Hypothesis Tests

- We will be using the following hypothesis tests in this course:
  - One sample t-test
  - Unpaired or independent samples t-test
  - Paired t-test
  - Correlation analysis

One-Sample T-Test

- This is the easiest of the statistical tests to understand

- Compare observed vs hypothesized mean
  - Observed: measured
  - Hypothesized: we choose this value to be meaningful

- T-Test determines the likelihood that the difference between the means occurs by chance

- The chance is reported as the p-value

p-value

- p-value: the probability that the difference occurs due to chance
  - A small p-value means that the difference is unlikely to be the result of chance
  - A large p-value means the difference is likely to be the result of chance
- What do we mean by random chance? Keep this question in mind and we will come back and give an answer.

Statistically Significant Difference

- The lower the p-value, the more certain that we can be that there is a **statistically significant** difference

- Most disciplines look for a p-value of 0.05 or less
  - if $p < 0.05$, reject the null hypothesis
  - if $p >= 0.05$, do not reject the null hypothesis

Problem 11.1

The file LipidData on Moodle represents a blood lipid screening of medical students.

1. Grab this Excel file, open it up in Excel.

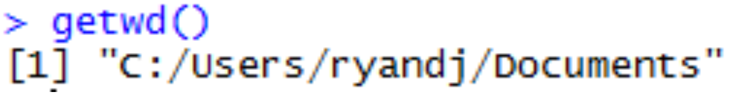
2. What is the mean Cholesterol value?

3. Is the cholesterol level significantly greater than 190? Can you tell by looking at the data? What do you think?

Problem 11.1

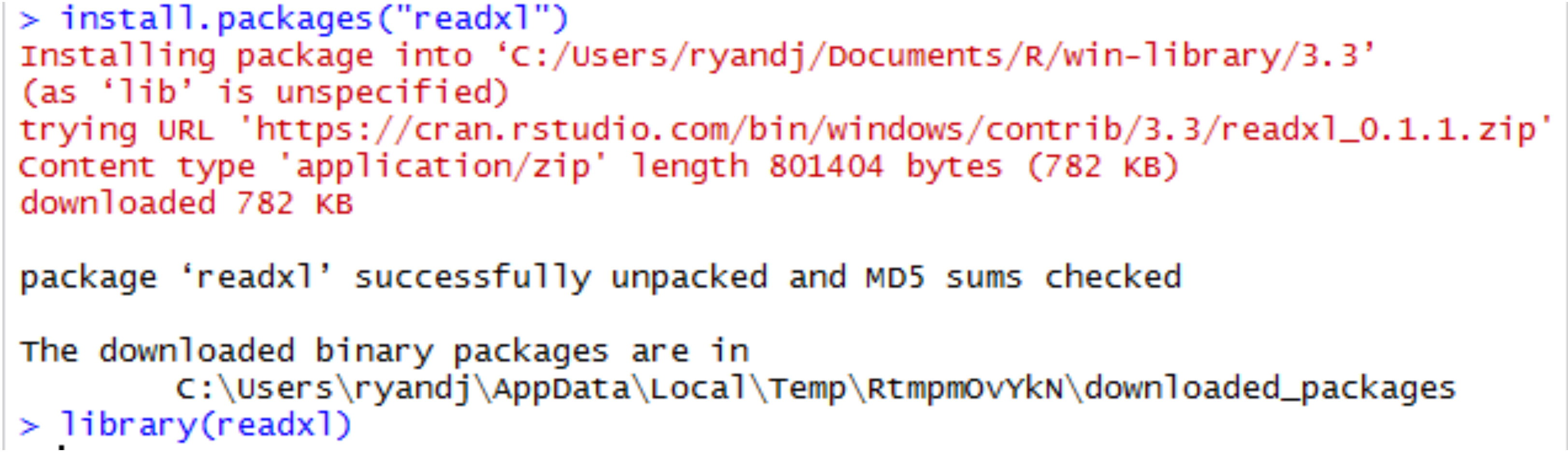
How to import Excel file into R

1. Prepare workspace rm(list=ls())
2. What directory are you working in getwd()



```
> getwd()
[1] "C:/Users/ryandj/Documents"
```

3. Install Readxl package and then activate the package



```
> install.packages("readxl")
Installing package into ‘C:/Users/ryandj/Documents/R/win-library/3.3’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.3/readxl_0.1.1.zip'
Content type 'application/zip' length 801404 bytes (782 KB)
downloaded 782 KB

package ‘readxl’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ryandj\AppData\Local\Temp\RtmpmOvYkN\downloaded_packages
> library(readxl)
```

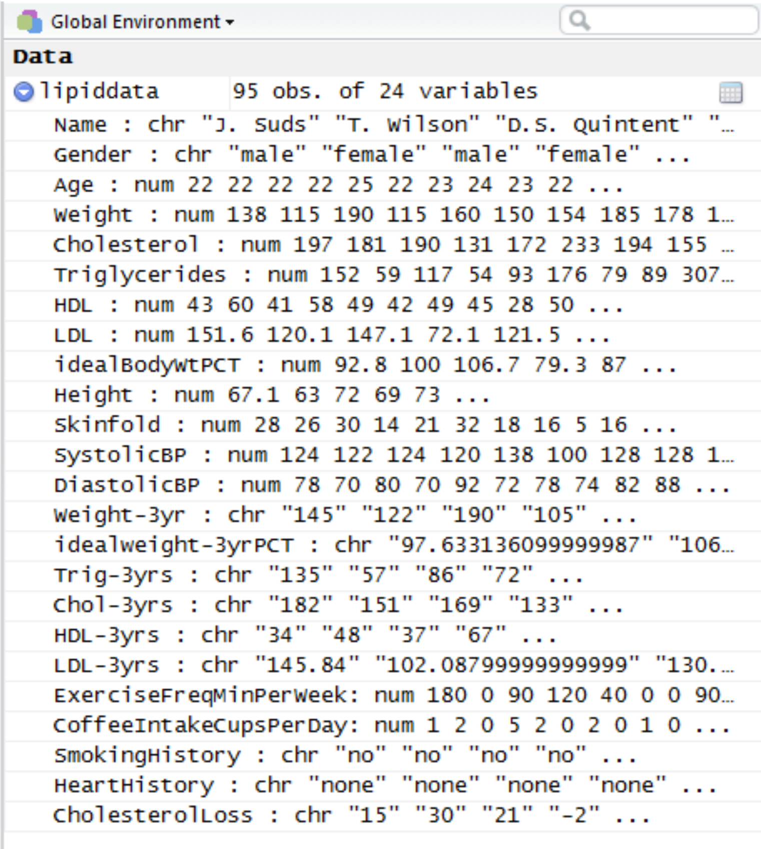
Problem 11.1

How to import Excel file into R

5. Copy the LipidData.xlsx from Moodle to your Documents

6. Import the data into R

```
> lipiddata=read_excel("LipidData.xlsx")
```

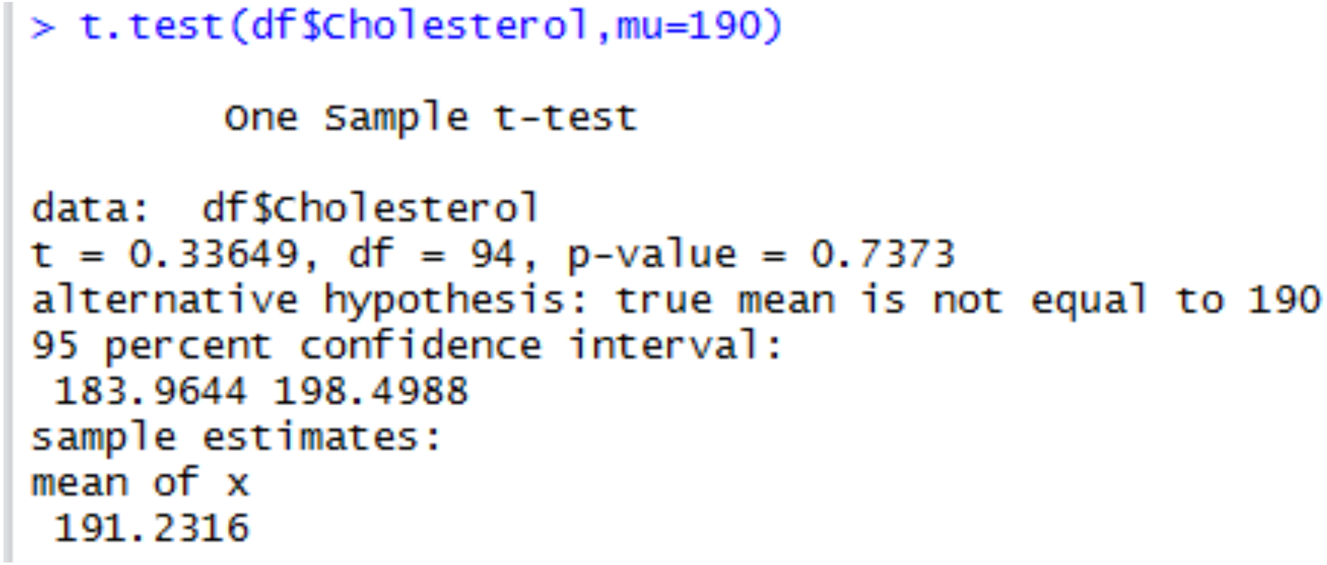


```
Data
lipiddata          95 obs. of 24 variables
    Name : chr "J. Suds" "T. Wilson" "D.S. Quintent" "…
    Gender : chr "male" "female" "male" "female" ...
    Age : num 22 22 22 22 25 22 23 24 23 22 ...
    Weight : num 138 115 190 115 160 150 154 185 178 1…
    Cholesterol : num 197 181 190 131 172 233 194 155 …
    Triglycerides : num 152 59 117 54 93 176 79 89 307…
    HDL : num 43 60 41 58 49 42 49 45 28 50 ...
    LDL : num 151.6 120.1 147.1 72.1 121.5 ...
    idealBodyWtPCT : num 92.8 100 106.7 79.3 87 ...
    Height : num 67.1 63 72 69 73 ...
    Skinfold : num 28 26 30 14 21 32 18 16 5 16 ...
    SystolicBP : num 124 122 124 120 138 100 128 128 1…
    DiastolicBP : num 78 70 80 70 92 72 78 74 82 88 ...
    Weight-3yr : chr "145" "122" "190" "105" ...
    idealWeight-3yrPCT : chr "97.633136099999987" "106…
    Trig-3yrs : chr "135" "57" "86" "72" ...
    Chol-3yrs : chr "182" "151" "169" "133" ...
    HDL-3yrs : chr "34" "48" "37" "67" ...
    LDL-3yrs : chr "145.84" "102.08799999999999" "130.…
    ExerciseFreqMinPerWeek: num 180 0 90 120 40 0 0 90…
    CoffeeIntakeCupsPerDay: num 1 2 0 5 2 0 2 0 1 0 ...
    SmokingHistory : chr "no" "no" "no" "no" ...
    HeartHistory : chr "none" "none" "none" "none" ...
    CholesterolLoss : chr "15" "30" "21" "-2" ...
```

Problem 11.1 Continued

- Our first objective is to perform a one-sample t-test on data from blood lipid screening of medical students. Specifically, we will test whether the mean cholesterol level is different than 190 in a *statistically significant* way, the point at which cholesterol levels may be unhealthy.
- What is the NULL hypothesis?
- What is the alternative hypothesis?

Problem 11.1 Results



```
> t.test(df$Cholesterol,mu=190)

        One Sample t-test

data:  df$Cholesterol
t = 0.33649, df = 94, p-value = 0.7373
alternative hypothesis: true mean is not equal to 190
95 percent confidence interval:
 183.9644 198.4988
sample estimates:
mean of x 
 191.2316 
```

Problem 11.1 Results

- The mean is slightly higher than 190; however, this difference is well within the range of sampling variance.
- A significance level of .737 indicates you would see a difference of this magnitude by chance more than 73% of the time
- Thus the cholesterol level is not significantly different than 190

Paired T-Test

- The most common use of the paired t-test is the comparison of two measurements (typically one measurement occurs “before” a treatment and the other “after” a treatment from the **same** individual or group.

- This test can determine if the treatment had a statistically significant effect.

- The p-value is the primary statistic of concern and the interpretation of the p-value is the same as for the one-sample t-test

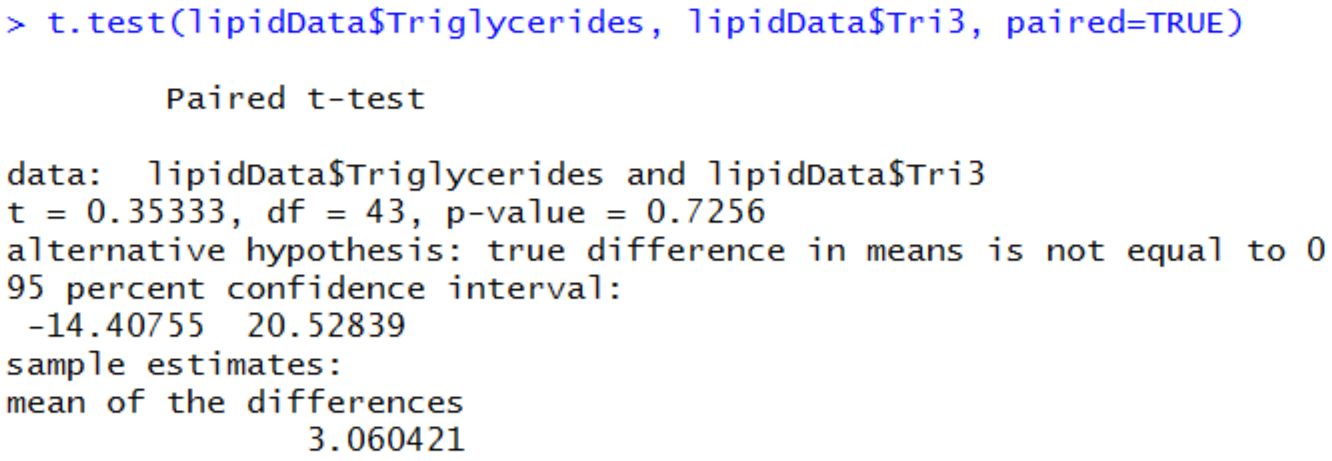
Problem 11.2

- Using the LipidData

1. What is the mean for Triglycerides?
2. What is the mean for Trig-3yrs?
3. Does it look like there is a statistically significant difference between Triglycerides and Trig-3yrs?

Problem 11.2 Continued

- Perform the paired t-test using the LipidData file
- State the Null Hypothesis and the alternative hypothesis
- There are only 43 students that have a before and after. How to we create tri43 (the before students) and tri433yrs (the after students)? Notice: These variables are not part of the data frame



```
> t.test(lipidData$Triglycerides, lipidData$Tri3, paired=TRUE)

        Paired t-test

data:  lipidData$Triglycerides and lipidData$Tri3
t = 0.35333, df = 43, p-value = 0.7256
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.40755  20.52839
sample estimates:
mean of the differences 
               3.060421 
```

- Should we accept the Null Hypothesis? Why?
- State your conclusion

Unpaired T-Test

- One measurement per individual

- Break our population into two natural subgroups
  - Male/Female; Smoker/Non-Smoker; Oak/Maple
  - Do the groups have a difference in measurement?

- Our primary statistic of concern is the p-value
  - How likely to occur by chance?

Problem 11.3

Question: Are the prices of houses near the Charles River more expensive than the prices of houses away from the Charles River.

The file BostonHousingData on Moodle contains information about Boston houses.

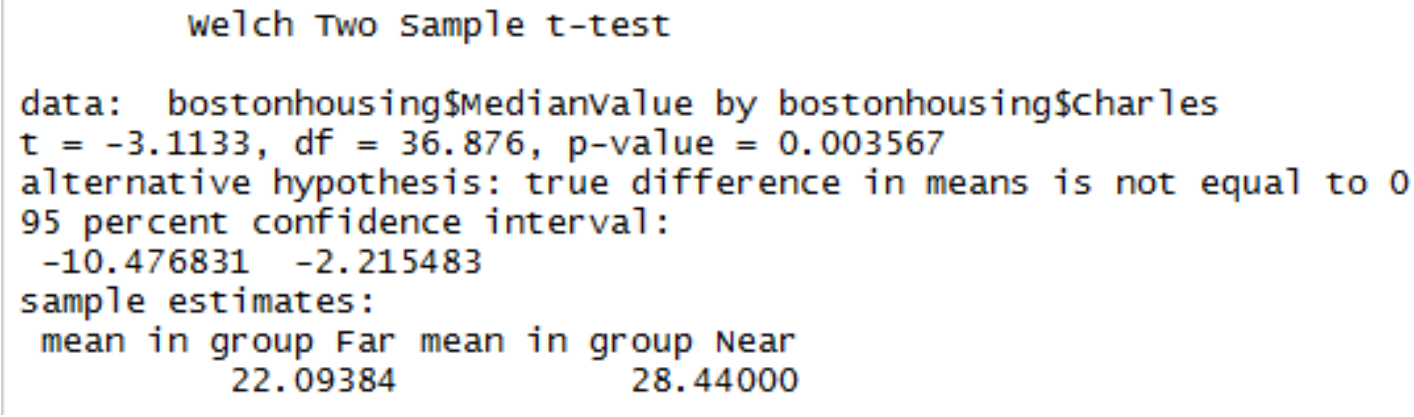
1. Grab this Excel file, open it up in R
2. State the Null Hypothesis and the alternative hypothesis
3. Perform an unpaired t-test

Problem 11.3

- What is the test variable? Why?

- What is the grouping variable? Why?

- Is the grouping variable in the data set a Factor? If not, make it a factor.



```
        Welch Two Sample t-test

data:  bostonhousing$MedianValue by bostonhousing$Charles
t = -3.1133, df = 36.876, p-value = 0.003567
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.476831  -2.215483
sample estimates:
 mean in group Far mean in group Near 
          22.09384           28.44000 
```

Problem 11.3

- Do you reject the Null Hypothesis? Why?
- State your conclusion

Correlation Analysis

- Correlation Analysis addresses the following: Is there a statistically significant association between variable X and variable Y?

- Interpreting the Pearson Correlation Coefficient is not an exact science. We might use the following interpretation:

  - -1.0 to -0.7 strong negative association
  - -0.7 to -0.3 weak negative association
  - -0.3 to +0.3 little or no association
  - +0.3 to +0.7 weak positive association
  - +0.7 to +1.0 strong positive association

Correlation Analysis Visual

- Use Scattergrams (Scatterplots) to visually display data analyzed with this test.
- You can also produce a correlation matrix of the relationship of all variables in the matrix.

http://www.statmethods.net/stats/correlations.html

Correlations

```
cor(dataframe)
cor(mtcars, use=“complete.obs”)
install.packages(“Hmisc”)
library(Hmisc)
rcorr(as.matrix(mtcars))
x <- mtcars[1:3]
y <- mtcars[4:6]
print(x)
cor(x,y)
rcorr(as.matrix(mtcars[1:6]))
```

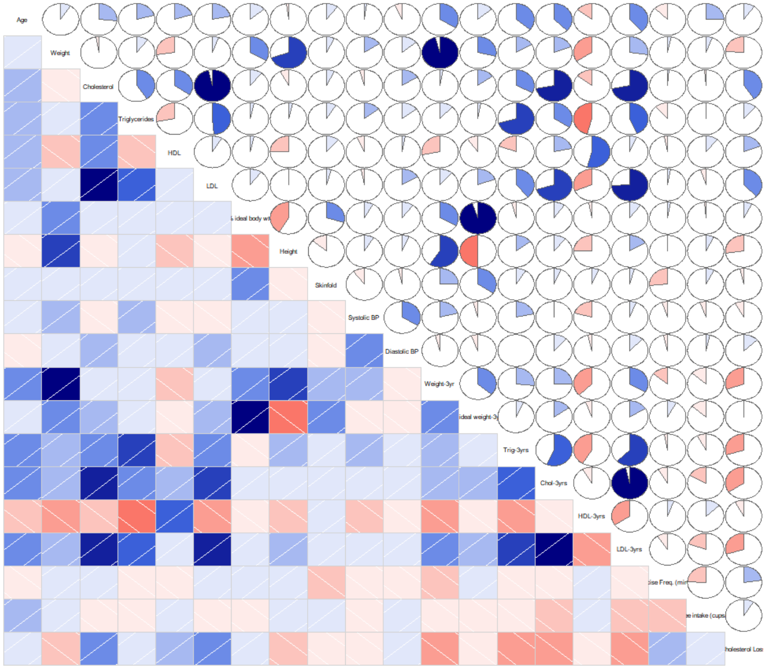
http://www.statmethods.net/stats/correlations.html

Problem 11.4

- What is the correlation between Cholesterol and Triglycerides?
- What is the correlation between Cholesterol and LDL?
- How would you graph either of these relationships?

Problem 13.4: Corrgram

```
install.packages("corrgram")
library(corrgram)
corrgram(mtcars, lower.panel=panel.shade,
upper.panel=panel.pie)
```


Age
Weight
Cholesterol
Triglycerides
HDL
LDL
% ideal body wt
Height
Skinfold
Systolic BP
Diastolic BP
Weight-3yr
ideal weight-3y
Trig-3yrs
Chol-3yrs
HDL-3yrs
LDL-3yrs
ise Freq. (mir
e intake (cups
holesterol Los