

R part II

Data Analysis with R

Dr. Irene Vrbik

University of British Columbia Okanagan
`irene.vrbik@ubc.ca`

Term 1, 2018

Vector

A **vector** is an indexed list of data of any type.

Create vectors using a colon or `seq()` (R's version of range)

```
> 1:10
```

```
[1] 1 2 3 4 5 6 7 8 9 10
```

```
> seq(5, 1, by = -0.5 )
```

```
[1] 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0
```

Create an empty vector with `c()` or fill it by specifying elements.

```
> c()
```

```
NULL
```

```
> c(4, 3, 5, 'a', 'd')
```

```
[1] "4" "3" "5" "a" "d"
```

Vector

Access elements in a vector using []

```
> myVector = c(4, 3, 5, 'a', 'd')
```

```
> myVector[i]  #Returns the ith element of myVector
```





```
> myVector[1]
```

```
[1] "4"
```

Vectors in R

Question

Which (if any) of the following statements are true?

1. Vectors in R are indexed from 0. 
2. `1:10` creates a vector of ten numbers. 
3. A vector may have data value of different types. 
4. If `data <- 1:5` then `data[2] + data[3] = 3`. 

Matrices

A **matrix** is a structure of rows and columns where each data value is the same data type. All rows must have the same length. All columns must have the same length.

Create a matrix from the vector `x` using `matrix()`.

```
> matrix(x, nrow=5, ncol=3, byrow=False)
# Starts at [1,1] and fills the column first before going to
# next column.
# Need to only specify ncol or nrow.
```

Access elements using `[row, col]`. Leaving one of them blank returns the whole row or column.

```
> myMatrix[i, j] #Returns ith row and jth column.
```

Matrices and Vectors

Append a vector to a matrix as a row using `rbind()`

```
> myMatrix = rbind(myMatrix, vec)
```

Append a vector to a matrix as a column using `cbind()`

```
> myMatrix = cbind(myMatrix, vec)
```

Lists

A **list** is an ordered collection of objects of **any** type.

Create a list using `list()`. Specify names of elements by using `name =` inside the brackets.

```
> myList = list(x=1:4, y=c('a', 'b'))  
#Creates a list with two elements x and y
```

Access elements using the double square brackets.

```
> myList[[2]]    # Returns the 2nd item of list (y)  
> myList[['x']]  #Returns item with the name x.  
> myList$x      #Returns item with the name x (no quotes)
```

Lists and Matrices in R

Question

Which (if any) of the following statements are true?

1. Data values in a list may be of different types. **X**
2. In a matrix, the number of rows and columns must be the same. **X**
3. Given matrix `m`, `m[2]` would return an error. **X**
4. Given matrix `m`, `m[,2]` would return all data in column 2. **X**

Question

Create a list called `grades` and add the following elements

1. Name (first name and last name),
2. Student number,
3. Assignment grades (multiple entries), and
4. Midterm grade.

Data Frame

A **data frame** is similar to a matrix but the columns can have different data types. Note these data frames are quite common and have uniform length of rows and columns.

To create a data frame by using `data.frame()`, specify names of variables within the brackets:

```
> myDF = data.frame(x = c(1:3), y = (2:4))
```

To convert a matrix into a data frame using `as.data.frame()`.

```
> myDF = as.data.frame(myMatrix)
```

Accessing Data in Data Frames

Access elements using `[row, col]` or `$variable_name`.

```
> myDF[i, j]           #ith row and jth column  
> myDF$x               #returns the column labelled x
```

Add new columns called `vec` into the data frame use in `$`.

```
> myDF$new_col = vec   #Adds vec as new_col
```

Factors

Factors are used for qualitative groups/categories (i.e. male/female). Use `as.factor()` to turn a vector or `data.frame` column into a factor.

```
mFyFactor = as.factor(x)
myDF$x = as.factor(myDF$x)
```

Access elements using `[]`:

```
myFactor[i]      #Returns ith element
```

Can use `class()` or `str()` to gain information about the type and/or structure of your variable/data. `str()` gives me detail.

Question

Which (if any) of the following are true?

1. Matrices must have the same number of rows as columns. **X**
2. Vectors must contain only one data type. **X**
3. A factor can contain only characters.
4. A data **frame's** columns can be of varying length. **X**

Subsets

Subsetting is used to extract data with particular values.

Syntax in R

```
subset(data, condition)
```

Example in R

```
cars_bc = subset(cars, prov == 'BC')
```

Question

1. Create a **data frame** `mydata` with the following column names/data:
 - 1.1 `id` – numbers one to 5.
 - 1.2 `location` – “BC”, “BC”, “AB”, “MB”, “BC”.
 - 1.3 `value` – 10, 20, 30, 40, 50.
 - 1.4 Make `location` a factor.
2. Add one more column to your data frame that is a factor
 - 2.1 `success` – “Y”, “N”, ‘N’, ‘N’, “Y”.
3. Display only the data from BC and $\text{value} \geq 20$.

Visualizing Data in R

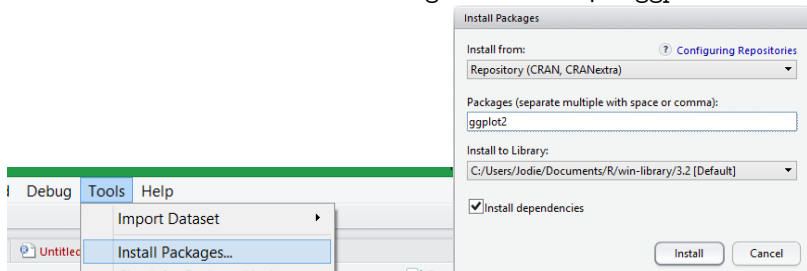
R supports several graphing libraries to produce graphs for qualitative and quantitative data including bar charts, histograms, and box plots.

We will use the package `ggplot2`. `gg` stands for Grammar of Graphics.

See the [ggplot cheatsheet](#)

Visualizing Data in R

To install Tools → Install Packages... Then input `ggplot2`.



Or (and this would be my preferred method) type the command:

```
install.packages("ggplot2")
```

To load that library into your working session type:

```
load("ggplot2") # quotes not necessary
```

Graphs for Qualitative Data: Frequency Tables

Frequency tables summarize the number of observations in each group.

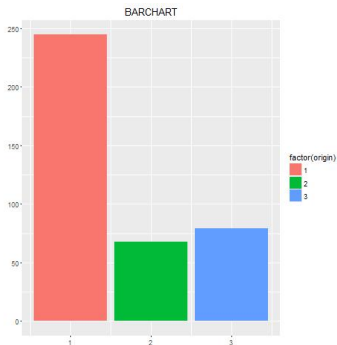
Frequency Table

```
1 > table(Auto$origin)
2     1     2     3
3 245   68   79
```

Graphs for Qualitative Data: Bar Charts

Bar charts have each group along the x-axis and a vertical bar with the height representing the number of observations of each group.

Using the dataset `Auto` in the `ISLR` package.



Frequency Table

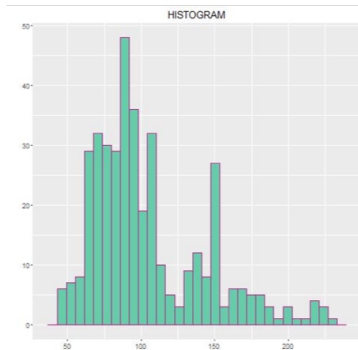
```
1 ggplot(Auto, aes(x=origin))
2 + geom_bar(aes(fill=factor(origin)))
3 + xlab("") + ylab("") + ggtitle("BARCHART")
```

Graphs for Quantitative Data: Histogram

A **histogram** is similar to a bar chart, but the x -axis is divided into bins.

The variable of interest is on the x -axis and the y -axis represents count of observations within each bin.

Visualizes the data distribution.



Histogram Example

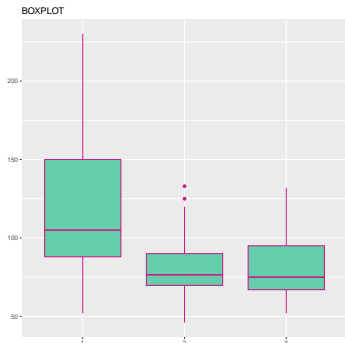
```
ggplot(Auto, aes(x=horsepower))  
+ geom_histogram(color='mediumvioletred', fill='mediumaquamarine')  
+ xlab("") + ylab("") + ggtitle("HISTOGRAM")
```

Graphs for Quantitative Data: Boxplot

A **boxplot** is a visualization of the five number summary.

1. Groups along the x -axis.
2. Data values along the y -axis.
3. Lowest and highest points are the min and max of the data respectively.
4. Bottom of box is Q1 and top is Q3.
5. Median is represented as the bar inside the box.

6. Single points represent outliers.



Boxplot Example Code

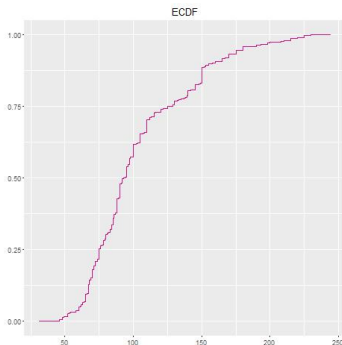
Boxplot Example

```
ggplot(Auto, aes(x=factor(origin), y=horsepower))  
+ geom_boxplot(color='mediumvioletred', fill='mediumaquamarine')  
+ xlab("") + ylab("") + ggtitle("BOXPLOT")
```

Graphs for Quantitative Data: ECDF

An **empirical cumulative distribution function (ECDF)** plot shows values along the x -axis and quantiles along the y -axis.

Each data point is plotted along with its corresponding quantile.







Boxplot Example

```
ggplot(Auto, aes(x=horsepower))  
+ stat_ecdf(color='mediumvioletred')  
+ xlab("") + ylab("") + ggtitle("ECDF")
```

Question

Which (if any) of the following are true?

1. Bar charts and histograms will work for the same variables. 
2. Boxplots show a five number summary. 
3. Variables type does not matter, any graph can be used. 
4. Histograms can give an idea of the distribution of a variable. 

Confidence Intervals

Consider the following statement

*“62% of US college students miss a class due to excessive drinking.
The result is accurate with 1.7 percentage points 19 times out of 20.”*

Unpacking this statement:

1. 62 is the estimated percentage.
2. 1.7 is the margin of error.
3. 19 times out of 20 is the stated confidence and $100\% \frac{19}{20} = 95\%$.

We call this a 95% confidence interval: (60.3, 63.7)

General Form of Confidence Interval

$$(\mu - me, \mu + me)$$

Interpretation of a 95% confidence interval: we are 95% confident that the interval will contain the true value of the parameter.

Hypothesis Testing

Hypothesis testing is used to determine if a relationship exists between two sets of data and make decisions/conclusions about that relationship.

Hypothesis testing is useful for

1. **business** in determining effectiveness of marketing, identifying customer buying properties, online advertising optimization.
2. **science and social science** in determining if data sets match a model, understanding scientific process based on collected data values, analysis of study data.

Probability Distribution Functions

`rnorm` Random Normal variables.

`dnorm` Evaluate normal probability density/

`pnorm`

`qnorm` Default distribution is $\text{mean} = 0$ and $\text{sd} = 1$.

`d` Density.

`r` Random number generation.

`p` Cumulative distribution.

`q` Quantile function.

Hypothesis Testing Steps

1. Declare hypotheses statement and null hypothesis.
2. Decide on a test statistic.
3. Use P-value and/or confidence interval to make decision/conclusion.
 - 3.1 A p-value of 0.05 “signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed.”
4. Data is used as evidence. Perform a test in order to make a decision: reject the null hypothesis or fail to reject the null hypothesis.

Note: We cannot **prove** that the null hypothesis is true or false. We can only show that there is evidence to suggest one conclusion or another.

Assumptions

There are assumptions that need to be met before performing statistical tests.

For the one sample:

1. Population of interest is normally distributed.
2. Independent random samples are taken.

For the two sample case:

1. The two samples are independent.
2. Populations of interest are normally distributed.

One Sample Test

A **one sample test** is used when a sample is compared to a model or known population/estimate.

As an example, using the car data test if the average mileage is different than 10km/L.

One Sample Test: Hypotheses Statements

Null hypothesis (H_0) always contains a statement of no change (=).

Alternatively hypothesis (H_A) can be one sided (< or >) or two sided (\neq).

1. $H_0: \mu = \text{test_number}$,
2. $H_A: \mu \neq \text{test_number}$.

Car mileage example

1. $H_0: \mu = 10$,
2. $H_A: \mu \neq 10$.

One Sample Test: Calculate Test Statistic

For the one sample test the t -test statistic is calculated as:

$$t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

\bar{y} is a sample mean, s is sample standard deviation, n is sample size, μ_0 is hypothesized mean value

— R code —

```
car_data <- read.csv("car_data.csv")  
t.test(x=car_data$km.L,  
       alternative=c("two.sided"), mu=10)
```

One Sample Test: Decision and Conclusion (using P -value)

If $p\text{-value} > 0.05$, the probability of seeing a sample mean more extreme is not that unlikely.

1. Fail to reject the null hypothesis.
2. There is no evidence to suggest that the mean value is less than, greater than, or different than the test value.

If $p\text{-value} < 0.05$,

1. Reject the null hypothesis.
2. There is statistically significant evidence to suggest that the mean value is less than that, greater than, or different than the test value.

One Sample Test: Decision and Conclusions (using P -value)

R code

```
> t.test(x=car_data$km.L, alternative=c("two.sided"),mu=10)
```

One Sample t-test

data: car_data\$km.L

t = 1.608, df = 29, p-value = 0.1187

alternative hypothesis: true mean is not equal to 10

95 percent confidence interval:

9.90338 10.80729

sample estimates:

mean of x

10.35533

One Sample Test: Decision and Conclusions (using P -value)

$P\text{-value} = 0.1187 > 0.05 \implies$ fail to reject the null hypothesis.

There is no evidence to suggest that the mean mileage is no 10 km/L.

Note: Unable to claim that either the null or alternative hypothesis is true. Can only reject or fail to reject the null hypothesis.

One Sample Test: Decision and Conclusions (using CI)

R code

```
> t.test(x=car_data$km.L, alternative=c("two.sided"),mu=10)
```

```
One Sample t-test
```

```
data: car_data$km.L
```

```
t = 1.608, df = 29, p-value = 0.1187
```

```
alternative hypothesis: true mean is not equal to 10
```

```
95 percent confidence interval:
```

```
9.90338 10.80729
```

```
sample estimates:
```

```
mean of x
```

```
10.35533
```

Can also make a conclusion (reject or fail to reject) based on the confidence interval.
We are 95% confident that the true mean mileage of the car lies within those bounds.
Since 10 km/L is within those bounds, fail to reject the null hypothesis.

Two Sample Unpaired

An unpaired (independent) **two sample test** compares two independent samples to determine if there is a difference between the groups.

Example

1. Compare effectiveness of two different drugs tested on two sets of patients.
2. Experiment versus control samples.

Two Sample Unpaired Example Hypothesis Statement

Using the `beaver2` dataset in R, test the hypothesis that there is no difference between the mean active temperature and the mean non-active temperatures.

$$H_0 : \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$$

Two Sample Unpaired Example Test Statistic

Use t -test statistics.

Using P-value

```
1 # Need to set active to be a factor first.  
2 beaver2$activ = as.factor(beaver2$activ)  
3 # Perform unpaired test  
4 t.test(temp~activ, data=beaver2,  
5         alternative=c("two.sided"), mu=0,  
6         paired=FALSE)
```


Two Sample Unpaired Example Decision and Conclusions

Using CI-value

```
> t.test(temp-activ, data=beaver2, alternative=c("two.sided"), mu=0,
paired=FALSE)

      welch Two Sample t-test

data:      temp by activ
t = -18.548, df = 80.852, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent condifence interval:
 -0.8927106 -0.7197342
sample estimates:
mean in group 0 mean in group 1
      37.09684      37.90306
```

The p-value $\ll 0.05$.

Reject the null hypothesis. There is vert strong evidence to suggest that there is a difference between active and non-active temperatures.

The two sample case tests a DIFFERENCE between the groups ($\mu_1 - \mu_2 \neq 0$).

The CI stated on the previous slide is the CI for the difference, $\mu_1 - \mu_2$

We reject the null hypothesis because 0 is not contained in the interval.

If 0 was contained we would fail to reject the null hypothesis.

Two Sample Paired Test

A **paired (dependent) two sample test** compares two dependent samples to see if there is a difference between the groups.

1. This test typically uses multiple measurements on one subject.
2. Also called a "repeated measures" test.

Examples

1. Affect of treatment on a patient (before and after)
2. Apply something to test subjects to see if there is an effect
3. Car example: Do cars get better mileage with different grades of gasoline?

Two Sample Paired Test Example Hypothesis Statement

The `athlete.csv` dataset contains data on ten athletes and their speeds for the 100m dash before training (Training = 0) and after (Training = 1).

Test the hypothesis that the training has no effect on the times of the athletes.

Test to see if the mean of the difference is different than 0.

$$H_0 : d = 0$$

$$H_A : d \neq 0$$

Two Sample Paired Test Example Test Statistic – R Code

Using CI-value

```
# Read the data
athlete = read.csv("athlete.csv", header=TRUE)

# Perform paired test
t.test(Time~Training, data=athlete,
        alternative=c("two.sided"), my=0, paired=TRUE)
```

Two Sample Paired Test Example Decision and Conclusion

Using P-value

```
> t.test(Time~Training, data=athlete, alternative=c("two.sided"),
mu=0, paired=TRUE)

Paired t-test
data:    Time by Training
t = -0.12031, df = 9, p-value = 0.9069
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.5544647  0.4984647
sample estimates:
mean of the differences
-0.028
```

The p -value $>> 0.05$.

Fail to reject the null hypothesis. There is no evidence to suggest that there is a difference between pre and post training times.

Two Sample Paired Test Example Decision and Conclusion

Using CI

```
> t.test(Time-Training, data=athlete, alternative=c("two.sided"),
mu=0, paired=TRUE)





Paired t-test
data:    Time by training
t = -0.12031, df = 9, p-values = 0.9069
alternative hypothesis true difference in means is not equal to 0
-0.5544647    0.4984647
sample estimates:
mean of the differences
               -0.028
```

The two sample case tests for a difference between the groups ($d \neq 0$). The CI is for the difference.

Fail to reject the null hypothesis because 0 is contained in the confidence interval.

Question

Which (if any) of the following are true?

1. Paired and unpaired t -tests are the same thing. 
2. Confidence intervals can be of any level of confidence (not just 95%). 
3. Confidence intervals can be used to make a conclusion about a hypothesis test. 
4. Confidence intervals can be used to prove that the null hypothesis is false. 

Question

Which (if any) of the following are true?

1. Unpaired t-tests test the difference between two means μ_1 and μ_2 .
2. Paired t-tests can be used to compare the difference between two measurements on the same subject.
3. In both the paired and unpaired two sample cases, a confidence interval containing 0 would result in a decision of: fail to reject the null hypothesis.
4. In the one sample t -test, a confidence interval containing 0 would result in a decision of: fail to reject the null hypothesis.

Question

Which is the most appropriate test for each of the following?

1. Is the average student mark in courses 70%? one sample
2. Does a student's mark improve after studying? two sample paired (same student)
3. Has the average student height increased since 1990? two sample unpaired (two distinct student populations)

Question

Which is the most appropriate test for each of the following?

1. Does radiation reduce the size of tumours when used to treat patients? two sample paired (same patient), although could argue against control/experiment groups in which case two sample unpaired
2. Is aspirin more effective than Tylenol for treating headaches? two sample unpaired
3. Are college graduates better than high school graduates at standardized tests? two sample unpaired

Question

Using the car data, test the hypothesis that the mean distance at each fill up is less than 450km.

Question

Use the car data to see if the mean distance for Alberta fill ups is different than the mean distance for B.C fill ups.

```
t.test(x = car_data$Distance, alternative=c("less"), mu=450)
```

```
t.test(Distance~prov, data=car_data, alternative=c("two.sided"), mu=
```

Linear models in R

A linear model is an equation that relates a response variable (y) to some explanatory variables (x 's). The general form of the model is:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

Not all of the data points can fall on this line so the full equation is

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_nx_{ni} + \varepsilon_i$$

Where ε_i denotes the error term associated with observation i .

Linear models in R

We are assuming

1. Residuals are independent.
2. Residuals are normally distributed.
3. Residuals have a mean of 0 for all X .
4. Residuals have constant variance.

Fitting a Linear Model

Using P-value

```
> lm(km.L~Litres+Distance, data=car_data)
> model

call:
lm(formula = km.L ~ Litres + Distance, data = car_data)

Coefficients:
(Intercept)      Litres      Distance
  10.35447    -0.33295     0.03251
```

The formula can be then be created using the values stored in
`model$coefficients`

$\text{Km.L} = 10.35447 - 0.33295 \cdot \text{Litres} + 0.03251 \cdot \text{Distance}$

Conclusion

1. R is a free and open source programming language for statistical computing and graphics.
2. R contains many useful features for data analysis including data structures such as vectors and data frames that make it easy to perform statistical analysis and visualization.
3. R is often used for hypothesis testing and understanding how to properly setup and interpret a test is an important skill.