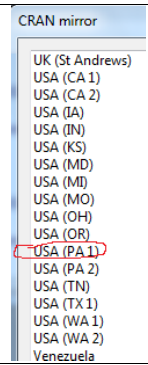
How to run T-tests using R

Presenters: Kyle Ward and Jon Wayland

Spring 2013

R is a free software downloadable at http://www.r-project.org/

Notes:	Code and Output:
1. R Console Setup: > prompts you for formula or function. The result appears on the next line(s).	
2. Comments begin with # Anything in the line following a # is a comment.	# This is a comment!
3. Installing a Package Many functions and data sets are available in packages that be downloaded from a CRAN site. We generally use PA 1 (Carnegie Mellon) We will be using a function called leveneTest( ) which is the package called **"car"**. 1) Select "**Install Packages**" in the dropdown menu "**Packages**" at the top of the screen. 2) Select the country, and state that is nearest you. 3) Select the package "**car**" and press "**ok**". 4) Activate the package using the library command.	 # In step 1) you can also use the command **>install.packages("car")** **> library(car)**
4. Creating a Dataset We will look at the differences between smokers and nonsmokers in terms of their scores on a memory test.	**> nonsmokers = c(18,22,21,17,20,17,23,20,22,21)** **> smokers = c(16,20,14,21,20,18,13,15,17,21)**
5. Alternate Form of the Data Rather than the scores being in separate vectors, data for t-tests is sometimes in this format: i) One vector with all scores (smokers and nonsmokers) ii) One vector identifying which group the individual belongs in.	**>scores = c(nonsmokers, smokers)** **>status=c(rep("no", length(nonsmokers)), rep("yes", length(smokers)))** **>data.frame(status, scores)** status scores 1 no 18 2 no 22 3 no 21 4 no 17 5 no 20 6 no 17 7 no 23 8 no 20 9 no 22 10 no 21 11 yes 16 12 yes 20 13 yes 14 14 yes 21 15 yes 20 16 yes 18 17 yes 13 18 yes 15 19 yes 17 20 yes 21

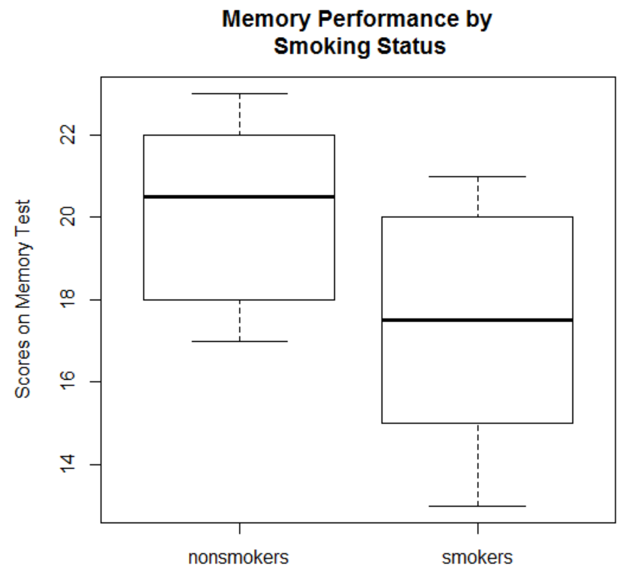
6. Boxplots

Boxplots are a useful graphical method for comparing multiple groups. It is important to keep in mind that boxplots are median oriented graphics, while the t-test is comparing means.

ylab is the label given to the y axis

the \n indicates that you want the main label to split onto a second line

```
> boxplot(nonsmokers, smokers, ylab="Scores on Memory Test", range=1.5, names=c("nonsmokers","smokers"), main="Memory Performance by\n Smoking Status")
```



7. Descriptive Statistics

Running the Mean and Standard Deviation of each group's scores gives you information to use during your interpretation of the t-test.

```
> mean(nonsmokers)
[1] 20.1
> mean(smokers)
[1] 17.5
> sd(nonsmokers)
[1] 2.131770
> sd(smokers)
[1] 2.953341
```

8. Independent-Samples T-test

We will be running an independent sample t-test comparing mean scores between two independent groups.

9. t-test() command

To the right is the general function for the t-test

It shows the **default** settings for a t-test run in R if you were to simply type in t.test(x,y)

```
# This is the general formula for a t-test

> t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

10. Checking Assumptions: Homogeneity of Variance

To check the assumption of equal variances, we will run a Levene's test in R

Levene's Test

In running a Levene's test in R, place the outcome variable first and the grouping variable second.

- A p-value less than .05 violates the assumption of homogeneity of variance
- We do not want a p-value less than .05

```
> # Be sure to do step 3. before using the levelTest() command.
> leveneTest(scores, status)

Levene's Test for Homogeneity of Variance (center = median)

      Df  F value   Pr(>F)
group  1  1.9459     0.18
      18
```

11. Checking Assumptions: Normality

Shapiro-Wilk Test

The Shapiro-Wilk test compares the scores in your sample to a normally distributed set of scores with the same mean and standard deviation.

- If the test is not significant (p > .05) it tells us that our distribution is not statistically different from the normal distribution.
- Like the Levene's test, we do not want this test to be statistically significant

>shapiro.test(scores)

Shapiro-Wilk normality test

data: scores

W = 0.9369, p-value = 0.2098

12. Checking Assumptions: Normality

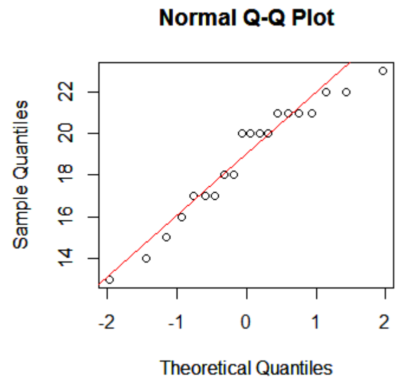
Q-Q plots

The Q-Q chart plots the values you would expect to get if the distribution were normal (theoretical values) against the values actually seen in our dataset (sample values).

- If the data were normally distributed, the data would fall along a straight line.
- Any deviation from a straight line represents a deviation from normality.

> qqnorm(scores)

>qqline(scores, col= "red ")



13. Independent Samples T-Test

To run an independent samples, two-tailed t-test, simply input the variable names in the t-test command. This command assumes the default that it is a two-sample test, it is a two-tailed test, equal variances are not assumed, and the confidence level is set at .95.

Note: We are using the default assumptions: alternative =c("two.sided"), mu = 0, paired = FALSE, var.equal = FALSE, conflevel = .095

>t.test(nonsmokers,smokers, var.equal = TRUE)

Two Sample t-test

data: nonsmokers and smokers

t = 2.2573, df = 18, p-value = 0.03665

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1801366 5.0198634

sample estimates:

mean of x mean of y

20.1   17.5

14. Paired Sample T-Test

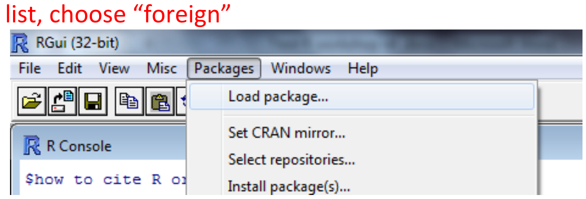
Do Male employees tend to earn more than female employees?

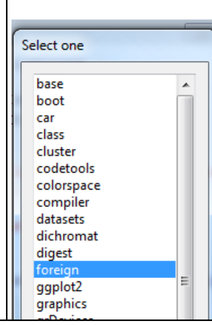
The following analyses uses the SPSS data set: **Salaries.sav**

The data set contains weekly salaries (in $) for pairs of 100 male and female employees. The individuals were matched by an indicator of salary potential and represent the entire spectrum of earnings. The data is not real but was generated to reflect actual 2011 earning distributions published by the Bureau of Labor Statistics. http://www.bls.gov/cps/cpswom2011.pdf

to load spss file, select "packages" in the top menu and choose "Load Packages"

from the package list, choose "foreign"





> library(foreign)

>salaries=read.spss (file.choose(), to.data.frame=TRUE)

select file from the folder it has been saved

>attach(salaries)

15. Looking at your data

Using the **head( )** function in R allows you to view 6 respondents. This is an easy way to see what the variables are named and how the data has been recorded.

```
>head(salaries)

 PairNo SalaryF SalaryM
1    1     287     256
2    2     291     314
3    3     180     257
4    4     177     243
5    5     304     276
6    6     322     268
```

16. Descriptive Statistics

```
> mean(SalaryM)
[1] 1004.98
> sd(SalaryM)
[1] 604.1701
> mean(SalaryF)
[1] 840.38
> sd(SalaryF)
[1] 520.951
```

17. Boxplots

Like the example above comparing smokers and nonsmokers on a memory task, we want to look at the distribution of the data using a boxplot.

Note: The range=1.5 tells R to mark any points that are further than 1.5*IQR from the box as outliers.
where
IQR=Interquartile range = Q_3-Q_1

```
> boxplot(SalaryM,SalaryF,ylab="Weekly Salary",
names=c("Males","Females"), main="Weekly Salaries by Gender",
range=1.5)
```

18. Paired Sample T-Test

- Because the equation runs left to right, the original measurement should be placed first.
- We will make the alternative hypothesis “greater” because we are testing if males have a larger weekly income compared to a paired group of females
- We change “paired” to true to change it to a paired sample t-test
- Based on the results, the p-value is less than .05, meaning that men have a significantly higher weekly income when compared to women.

```
> t.test(SalaryM,SalaryF, alternative = c("greater"), paired = TRUE,
var.equal = FALSE, conf.level = 0.95)


Paired t-test

data:  SalaryM and SalaryF
t = 9.2209, df = 99, p-value = 2.738e-15

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:
134.9607     Inf

sample estimates:
mean of the differences
164.6
```

19. Histogram of the Differences

```
> hist(SalaryM-SalaryF, main="Distribution of Differences in Weekly Salaries\nMale-Female",  ylab="Number of Pairs of Employees ", xlab=" Salaries(USD) ", col="light blue")
> abline(v=0, col="red", lwd=3)
```