R Tutorial sheet 1 - two group tests

# R Tutorial sheet 1 - two group tests

# Instructions

Please use RStudio Cloud, or RStudio on your computer to complete this exercise sheet. The below **tasks** indicate where you need to run some code in R. All the code you need is provided, but you will need to copy it into your own script (in order) for it to run. Based on the output from R, please try to answer all questions.

The analysis is based on analysis of a research question, which is first introduced.

# Scenario

Based on: Altered Microenvironment Promotes Progression of Preinvasive Breast Cancer: Myoepithelial Expression of αvβ6 Integrin in DCIS Identifies High-risk Patients and Predicts Recurrence, Michael D. Allen, Gareth J. Thomas, Sarah Clark, Marwa M. Dawoud, Sabarinath Vallath, Sarah J. Payne, Jennifer J. Gomm, Sally A. Dreger, Sarah Dickinson, Dylan R. Edwards, Caroline J. Pennington, Ivana Sestak, Jack Cuzick, John F. Marshall, Ian R. Hart and J. Louise Jones. Clin Cancer Res January 15 2014 (20) (2) 344-357; DOI: 10.1158/1078-0432.CCR-13-1504

## Research question

A working hypothesis of the authors was that upregulation of αvβ6 may be a key marker of transition of DCIS to invasive disease.

Like most solid tumours, breast cancers require new blood vessel growth in order to grow more than a few millimetres in diameter. The degree of angiogenesis in association with breast cancer has been shown to be an independent predictor of prognosis. There is a correlation between tumour microvessel density (MVD) and the presence of axillary lymph node, and distant metastases. Some evidence from a limited number of studies also suggests that angiogenesis is associated with more aggressive ductal carcinoma in situ, which is a non-invasive pre-cursor to invasive breast cancer.

To help assess the overall research hypothesis, we will investigate whether expression of avb6 in DCIS is associated with angiogenesis.

## Analysis aims

1. Determine the association between avb6 and grade of DCIS
2. Assess the relationship between avb6 on myoepithelial cells in DCIS and angiogenesis

## Study design

140 patients with histologically confirmed DCIS were identified from the hospital database at St. Bartholomew’s Hospital (N=90) or from the International Breast Cancer Intervention Study II (IBIS-II) clinical trial database (N=50).

## Methods

Myoepithelial cell staining was categorised homogenous, heterogenous or negative in relation to the expression of the avb6 integrin. The pathological grade of the DCIS was scored according to nuclear grade, as used for NHS breast screening program: low, intermediate, high grade. Micro vessel density was assessed using the Chalkley grid method.

# Analysis

## Data preparation

### Task: Load the data

*Code*

mydta=read.csv("181008-classdta.csv")

### Task: Print the first few rows of the data, and obtain summary statistics for each column

*Code*

head(mydta)  
summary(mydta)

#### Question: List all the variables and describe their data type

The R output will show you that the data for our analysis has four columns

1. **B6**: a binary (TRUE/FALSE) indicator variable, for *avb6 positivity*
2. **B6t**: a three level variable for *avb6 level* (0 = negative, 1 = heterogeneous, 2 = positive)
3. **CHALKLEY** the main variable of interest: *Chalkley count*
4. **GRADE** histological *grade* of the cancer (coded as 0,1,2)

#### Question: Which variables will you use for analysis to look each of the two main aims above?

## Descriptive analysis

In an analysis it is important to first have a look at fundamental aspects such as the number of samples overall and in different groups, as well as overall summary statistics mean, median, inter-quartile range. This is partly because it will help to identify if there are issues such as missing data, or something has gone wrong with the data import.

### Task: Determine how many samples are included in this analysis

*Code*

nrow(mydta)

Note that this is less that the number of samples obtained (see methods above).

### Task: What percentage of samples are excluded from this analysis

*Code*

round(nrow(mydta) / 140 \*100,1)

### Task: How many are avb6 positive?

*Code*

table(mydta$B6)

### Task: Of those avb6 positive, how many are heterogeneous?

*Code*

table(mydta$B6t)

You should find that 48/62 were heterogenous for avb6.

### Task: What is the range of Chalkely count, mean, overall?

*Code*

summary(mydta$CHALKLEY)

### Task: How many are in the different tumour grade categories?

*Code*

table(mydta$GRADE)

## Comparing independent samples

We next consider some statistical analysis methods to compare independent samples, in independent groups, to assess whether there are differences between the distribution of (the same) numerical variable between the groups.

Recall **Independence** between two units of analysis occurs when the outcome for each does not depend on the outcome of the other. For example, if you roll a die twice the value of the second roll does not depend on the first: the two are independent. An example where this is not true is when measurements are made on the same unit more than once. For example, we measure blood pressure on John today and tomorrow; an example of paired data.

### Task: create a boxplot of Chalkely count by avb6 positivity

*Code*

## Chalkley by avb6 positivity  
boxplot(split(mydta$CHALKLEY, mydta$B6), xlab="avb6 positivity", ylab="Chalkley count")

![](data:image/png;base64;base64,)

#### Question: Looking at the box plots, what are the approximate median and inter-quartile range for the two groups?

### Task: Calculate summary statistics by avb6 positivity

*Code*

tapply(mydta$CHALKLEY, mydta$B6, summary)

You should find, for example, the mean is 5.43 in the avb6 positive group, vs 2.96 in the avb6 negative group.

### Task: repeat previous two tasks analysis by 3-category avb6

*Code* plot:

## Chalkley by avb6 trend  
boxplot(split(mydta$CHALKLEY, mydta$B6t), names=c("Neg", "Het", "Pos"), ylab="Chalkley count", xlab="avb6 category")

![](data:image/png;base64;base64,)

*Code* statistics: adapt code from above to find out.

#### Question: Does splitting up avb6 positivity help to assess the hypothesis of a difference in chalkely count according to avb6? Explain

## Difference in average between two independent samples?

So far the plots and summary statistics indicate some differences in Chalkley count by avb6 status. There are n=32 in the negative and n=64 in the positive avb6 groups, so you might well think these are real. We may check the intuition using some common statistical analysis methods.

We first look at some statistical methods to estimate the potential size of the mean difference (confidence intervals), and conduct hypothesis tests to help quantify on a commonly used scale (p-values) whether the differences observed could be due to chance. The focus is on differences between avb6 positive and negative samples.

Standard t-tests in statistical software require the following assumptions:

1. The groups are independent [**A1**]
2. The observations within each group are independent [**A2**]
3. The distribution of data in both groups is approximately normal [**A3**]
4. The two groups have the same variance [**A4** - OPTIONAL]

### Task: Apply a t-test under assumptions [A1]-[A4] for the difference in mean chalkey count between the avbg positive (homogeneous or heterogenous) vs negative

*Code*

##classical t-test - assuming constant variance, t-distribution reference  
myttest=t.test(mydta$CHALKLEY[mydta$B6==0], mydta$CHALKLEY[mydta$B6==1], var.equal=TRUE)  
myttest

#### Question: How strong is the evidence for a difference between the groups based on these results?

The results provide the t-statistic (you should get t = -4.2) and associated p-value, as well as an estimated 95% confidence interval for the mean difference (you should get 95%CI -3.6 to -1.3). The confidence interval is obtained here under the same assumptions [A1]-[A4]. They show there is strong evidence of a difference based on this test: the p-value is very small, and the confidence interval is some way from 0 suggesting the differences are real.

#### Question: Are the results robust to the assumption [A4] or constant variance between the two groups?

If you inspect the boxplots and summary statistics above you might not be happy to accept a constant variance assumpution.

### Task: Repeat the t-test, but without making assumption A4

*Code*

##t-test that allows different variance between groups (Welch t-test)  
t.test(mydta$CHALKLEY[mydta$B6==0], mydta$CHALKLEY[mydta$B6==1], var.equal=FALSE)

#### Question: Does this new analysis change your interpretation?

### Non-parametric tests

We might also not be certain about assumption [A3] - that the data are normal. For example, Chalkley count cannot be less than 0, but the normal distribution does not have a lower bound. Do we want to make this assumption?

One alternative method to test for differences in a hypothesis testing frameword is a *Wilcoxon ‘non-parametric’ test*. This just assumes - independent groups [A1] - independent observations within each group [A2].

It does not makes any assumption about the distribution (such as A3 and A4 above). It is called a *non-parametric* test because there is no parametric distributional assumption. It works by transforming the data to their *ranking*, whereby the smallest value is 1 and the largest value in the data is equal to the sample size, and then making comparisons based on summations of that data in the two groups (hence the term “rank sum” below).

### Task: Run a non-parametric test of a difference in Chalkley count between avb6 positive and negative groups (Wilcoxon test)

*Code*

wilcox.test(mydta$CHALKLEY[mydta$B6==0], mydta$CHALKLEY[mydta$B6==1])

#### Question: Do findings change from the t-tests?

## Difference in proportions between two groups?

When the outcome in a binary variable such as yes / no, or true / false, or black / white one does not use a t-test or a Wilcoxon test for differences. The standard method for a difference in two proportion follows the same basic idea as a t-test: it divides the difference by an estimate of the standard error of the difference.

### Task: Test whether the proportion avb6+ differs between grade <2 and grade = 2, using a two-sample z-test

*Code*

crosstab=table(mydta$GRADE==2,mydta$B6==1)  
crosstab  
ngrade=table(mydta$GRADE==2)  
npos=crosstab[,2]  
prop.test(npos, ngrade)

#### Question : Does the 95%CI for the difference overlap with zero? Is this consistent with the p-value? Explain

# Appendix: Some technical details

## t-tests and related confidence intervals

A t-test is a test for the mean difference between two groups.

The t statistic is the sample mean difference divided by an estimate of the standard error of the mean difference. For example - t=2 means that the difference in means is twice the standard error of the difference. - t=3 is three times the standard error of the difference, etc.

One way to interpret the t-statistic is that it *standardises* the mean difference. For example, you can interpret the t-statistic whether or not you are comparing the height in cm between two groups, or the height in metres between two groups, or the weight in kg between two groups. If t=2 for the difference in height in cm, it will equal 2 if height is measured in metres. If t=2 for difference in weight then it has the same interpretaion as when t=2 for differences in height: the mean difference is twice the estimated standard error of the difference.

In a hypothesis-testing framework one needs to compare the t-test statistic to a *reference distribution*, which gives an estimate of the t-test statistic *under the null hypothesis of no difference, and all other model assumptions*. This reference distribution may be calculated theoretically based on assumptions **A1, A2 and A3**, and optionally **A4**. Unless you instruct it differently, the standard t.test() command in R only uses **A1, A2 and A3**.

## What is best, a parametric or non-parametric test?

The justification for making more assumptions in the analysis is that it buys the analyst *power*. The more power, the more chance to reject the null hypothesis. However, if the assumptions are not met it can lead to erroneous conclustions. For this reason several methodologies have been developed to avoid making *parametric* assumptions such as the data being from a normal distribution.

The choice between parametric and non-parametric tests will depend on context. Sometimes it may be reasonable to borrow evidence from earlier experiments that suggest approximate normality is reasonable, particularly when sample size is small and a non-parametric analysis may be underpowered to show any difference, even if there are striking differences in the mean difference as an absolute value. On the other hand, if sample size is moderate to large, little is lost in power from a non-parametric test, but there can be a gain in robustness.