The University of Sydney
STAT2012 Statistical Tests

Semester 2	Computer Practice Week 5	2015

Useful R commands

- The R command to perform a two samples t-test for the hypotheses: $H_0: \mu_x = \mu_y$ against $H_1: \mu_x \neq \mu_y$ with the *equality of variance* assumption is

  `t.test(x,y,alt="two.sided",mu=0,paired=F,var.equal=T,0.95)`.

  The alternative can be `greater`, `less` and `two.sided`. Without `mu=0`, the default `mu=0` is used. The assumption on the two population variances can be `var.equal=T` and `var.equal=F`.

  This command produces the test statistic, the p-value and a 95% CI for μ.

- The R command to perform the Wilcoxon rank sum test for the hypotheses: $H_0: \mu_x = \mu_y$ against $H_1: \mu_x \neq \mu_y$ is

  `wilcox.test(x,y,alternative="two.sided",mu=0,exact=T,correct=F)` .

  This gives exact p-value. If there are ties, you may set `exact=F` and normal approximation will be used.

Important points

- You will perform the two sample t test and the Wilcoxon rank sum test to compare the means of two populations.
- Only some of the R codes are provided.

Practice Problems

Open the data set `survey` containing measurements of the following variables from 95 students:

sex	1=male; 2=female
age	Year
height:	Inches
credit:	Number of credit cards in possession
pulse:	Number of heartbeats in one minute
pulse.ex:	Number of heartbeats in one minute after regular exercise over a period
exercise:	Number of hours during last week
smoke:	1=yes; 2=no
hand:	1=left-handed; 2=right-handed; 3=ambidextrous

Read the data `survey`. Set `pulse.sf` to contain the `pulse` among female students who smoke and `pulse.nf` to contain the `pulse` among female students who do not smoke.

```
survey=read.csv("http://www.maths.usyd.edu.au/u/UG/IM/STAT2012/r/survey.csv")
attach(survey)
pulse.sf=pulse[smoke==1 & sex==2]
pulse.sf
pulse.nf=pulse[smoke==2 & sex==2]
pulse.nf
```

Note that you can use this link to read the data set at home.

1. Test if the pulses of female student smokers and nonsmokers are *different* using the *two sample t-test*.

   (a) State the null and alternative hypotheses.

   (b) Draw boxplots of each data and qq-plot of combined data and comment the equal-of-variance and normality assumptions respectively.

   (c) Perform the test assuming normality and equal-of-variance for the populations of female smokers and nonsmokers. State the test statistic, p-value and confidence interval for the difference in pulse between female smokers and nonsmokers. Draw your conclusion on H_0 using *both the p-value and confidence interval*.

   (d) Drop the assumption of equality-of-variance and perform the test in (c) again. Comment any differences in test statistic, degrees of freedom and p-values between this test and the test in (c).

2. Repeat the test in Question 1 using the Wilcoxon rank sum test.

   (a) Find the ranks for the samples of female smokers and non-smokers, their sums of ranks and hence the test statistic W. Are there any ties? Should normal approximation to the distribution of test statistic W be used?

   (b) Perform the test and report the test statistic and p-value. Draw your conclusion about H_0 based on the p-value.

   Show that the reported test statistic is $w_0 = w - \min(w) = w - m(m+1)/2$ by calculating

   ```
   w0=w-m*(m+1)/2
   w0
   ```

   Note that `m` is the sample size of `pulse.sf` and you need to change it if you use another name.

   (c) Find the mean and variance of W. Hence verify the p-value of the test.