R Intermediate Short Course
Session 6 - Further Hypothesis Testing (Analysis of Variance Test (ANOVA))

The University of the West Indies, St Augustine

Saturday 23th July 2022
(5:00pm - 7:00pm)
(online)

Thus far we have compared population means and proportions for both one-sample and two-sample hypothesis testing.

Further to this, we can compare the population means of 3 or more samples using the methods:

1. One-way Analysis of Variance (ANOVA) *(normally distributed)*
2. Kruskal Wallis *(not normally distributed)*

The two tests above tell us whether or not the means are significantly different from each other.

If the means are equal then the test ends there. If they are different, then the next step is to determine which means differ by a comparison test.

One-way ANOVA

Assumptions of ANOVA Test (F-Test)

- Each group sample is drawn from a normally distributed population.
- All populations have a common variance.
- All samples are drawn independently of each other.

A one-way analysis of variance is a **parametric** test used to determine if there are any significant differences in means.

Let a be the number of treatments or groups being compared, then the general hypotheses are:

H_0: $\mu_1 = \mu_2 = ... = \mu_a$ versus
H_1: At least one μ_i is different

where the null hypothesis H_0 is rejected if,

1. The test statistic $>$ critical value
2. The p-value $<$ level of significance, α

This test is called the F-test.

The test statistic **F**, follows the f distribution which is the ratio of two independent chi-squared random variables. As a result the f distribution carries two sets of degrees of freedom.

The test statistic and the p-value are both calculated by means of an ANOVA (or ANalysis Of VAriance) table.

The general form of an ANOVA table is as follows:

Source	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Treatments	a - 1	SSTr	MSTr	MSTr/MSE
Residuals	N - a	SSE	MSE	
Total	N - 1	SST		

where n = number of observations in each treatment a

and N = $n \times a$ = number of observations in the data set

In R, the p-value is included as the last column in the ANOVA table and the last row of totals is excluded.

The critical value F_{α, df_1, df_2} is computed by:

```
qf(1 - alpha, df1, df2)
```

where df1 = a - 1 (treatment degree of freedom)
and df2 = N - a (residual/error degree of freedom)

Constructing ANOVA table in R

```
anova = aov(replicates ~ treatments)
anova = aov(data ~ labels)
summary(anova)
```

Here, *data* is a numeric vector and *labels* is a vector of labels that correspond to the *data* vector.

If an F test rejects H_0 concluding that at least one mean is different, we can perform a comparison test. This is done to determine which of the treatment means are different.

Tukey's Comparison test, (post hoc test when Reject Ho in ANOVA)

Let i and j be labels 1, 2, ... a to represent the treatments in a data set where μ_i = mean of treatment i and μ_j = mean of treatment j

Then the hypotheses being tested are,
H_0: $\mu_i = \mu_j$ versus H_1: $\mu_i \neq \mu_j$

We reject H_0 if

1. The confidence interval does not include 0
2. The adjusted p-value $<$ level of significance, α

Performing Tukey's Comparison test in R

```
TukeyHSD(anova, conf.level = CL)
```

where *anova* was defined previously in the F test.

The R output of tukey's test provides all possible comparisons between the treatment groups, the respective confidence intervals (lower and upper) and the adjusted p-values.

Example 1: The following data shows the price of vineyard properties for 3 consecutive years in California. The sale price is given in thousands of dollars.

1996:	30	34	36	38	40
1997:	30	35	37	38	40
1998:	40	41	43	44	50

Construct an ANOVA table to determine whether the mean price of vineyard land is the same for each of the 3 years. Conduct the test at a 0.01 level of significance.

Required to test
H_0: $\mu_1 = \mu_2 = \mu_3$ vs
H_1: At least one μ_i is different

STEP 1: Enter the price and year label as vectors in R

```
> price = c(30,34,36,38,40,30,35,37,38,40,40,41,
    43,44,50)
```

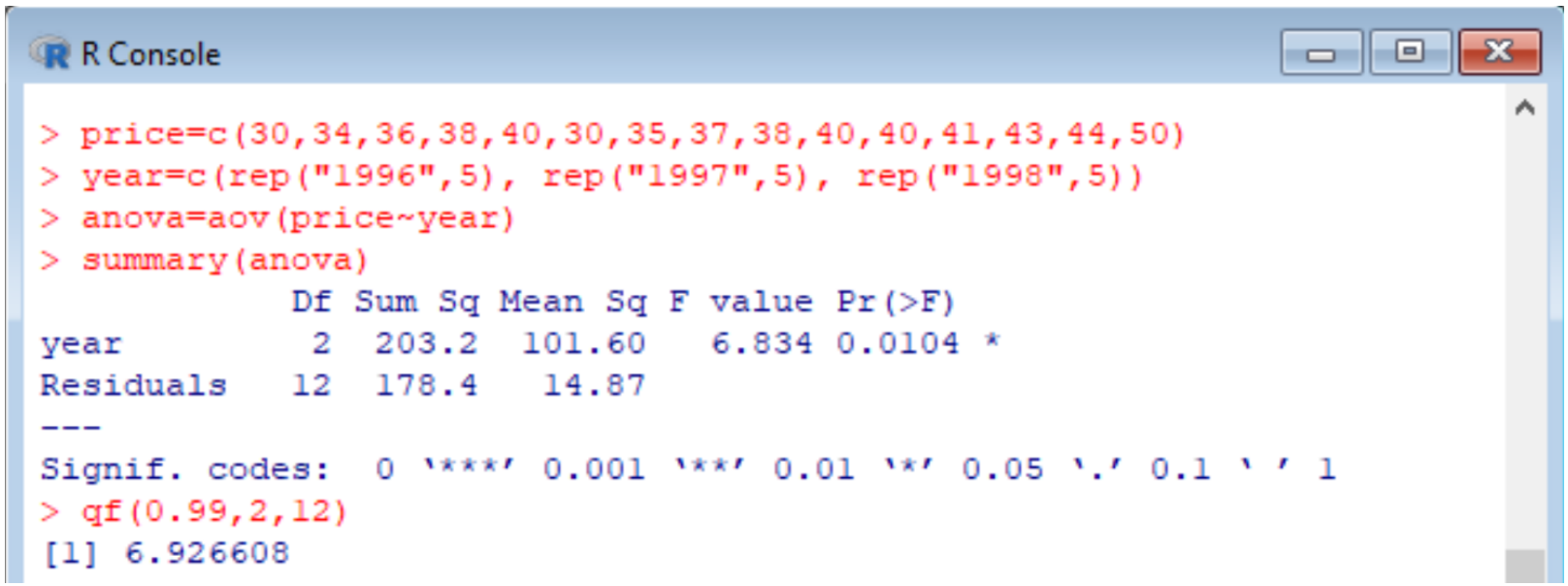
```
> year = c(rep("1996",5), rep("1997",5),
    rep("1998",5))
```

STEP 2: Perform and generate the ANOVA table

```
> anova = aov(price ~ year)
> summary(anova)
```

STEP 3: Compute the related critical value

```
> qf(0.99,2,12)
```



STEP 4: Interpret the results of the F test

1. The test statistic, F = 6.834 < critical value $F_{0.01,2,12} = 6.927$
2. The p-value (= 0.0104) > α (= 0.01)

We fail to reject H_0 at a 1% level of significance concluding that the mean price of vineyard land is the same for the 3 years.

Example 2: In an attempt to facilitate religious activities in a community, the following data was collected:

Hindu	Christian	Muslim	Other
16	20	17	3
20	21	15	5
18	25	19	10

Construct an ANOVA table and perform at $\alpha = 0.05$, the F test:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: At least one mean is different.
If necessary also determine which religious group/s are different.

STEP 1: Enter the data in R

```
> persons = c(16,20,18,20,21,25,17,15,19,3,5,10)
> religion = c(rep("h",3), rep("c",3), rep("m",3),
    rep("o",3))
```

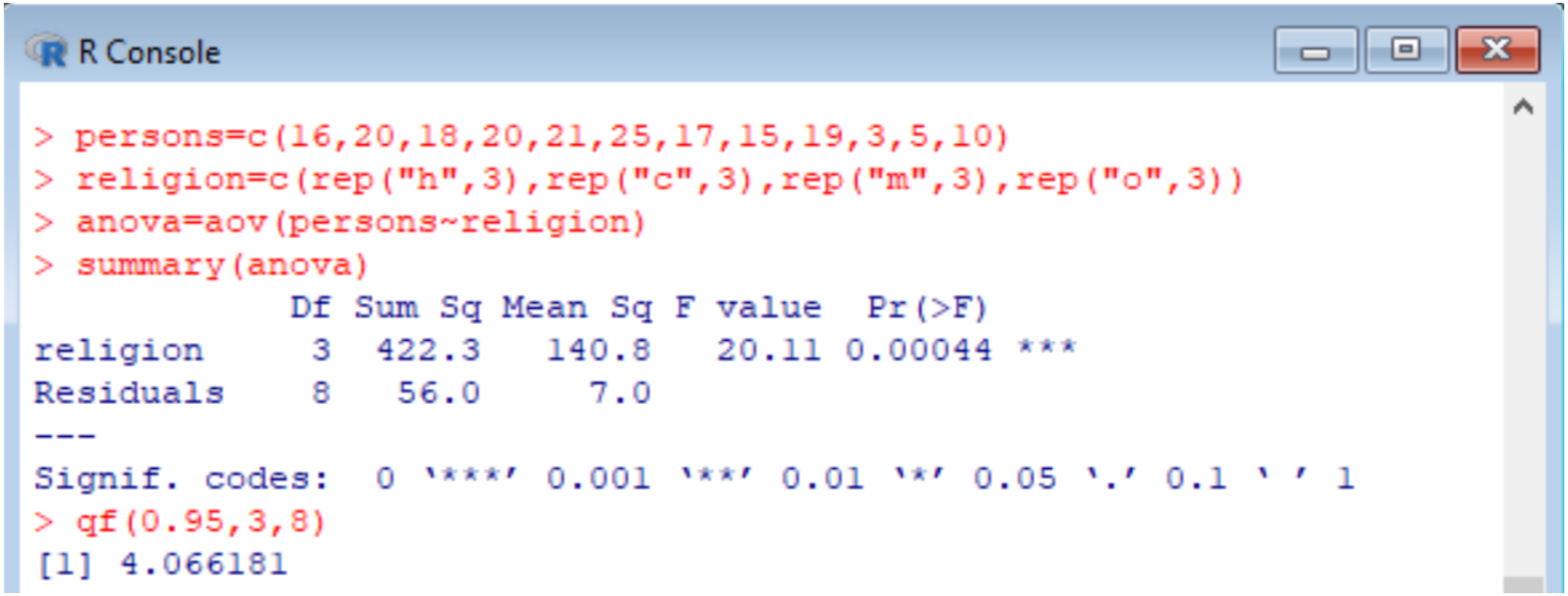
STEP 2: Perform and generate ANOVA table

```
> anova = aov(persons ~ religion)
> summary(anova)
```

STEP 3: Simulate the related critical value

```
> qf(0.95,3,8)
```

R Output (see next slide)



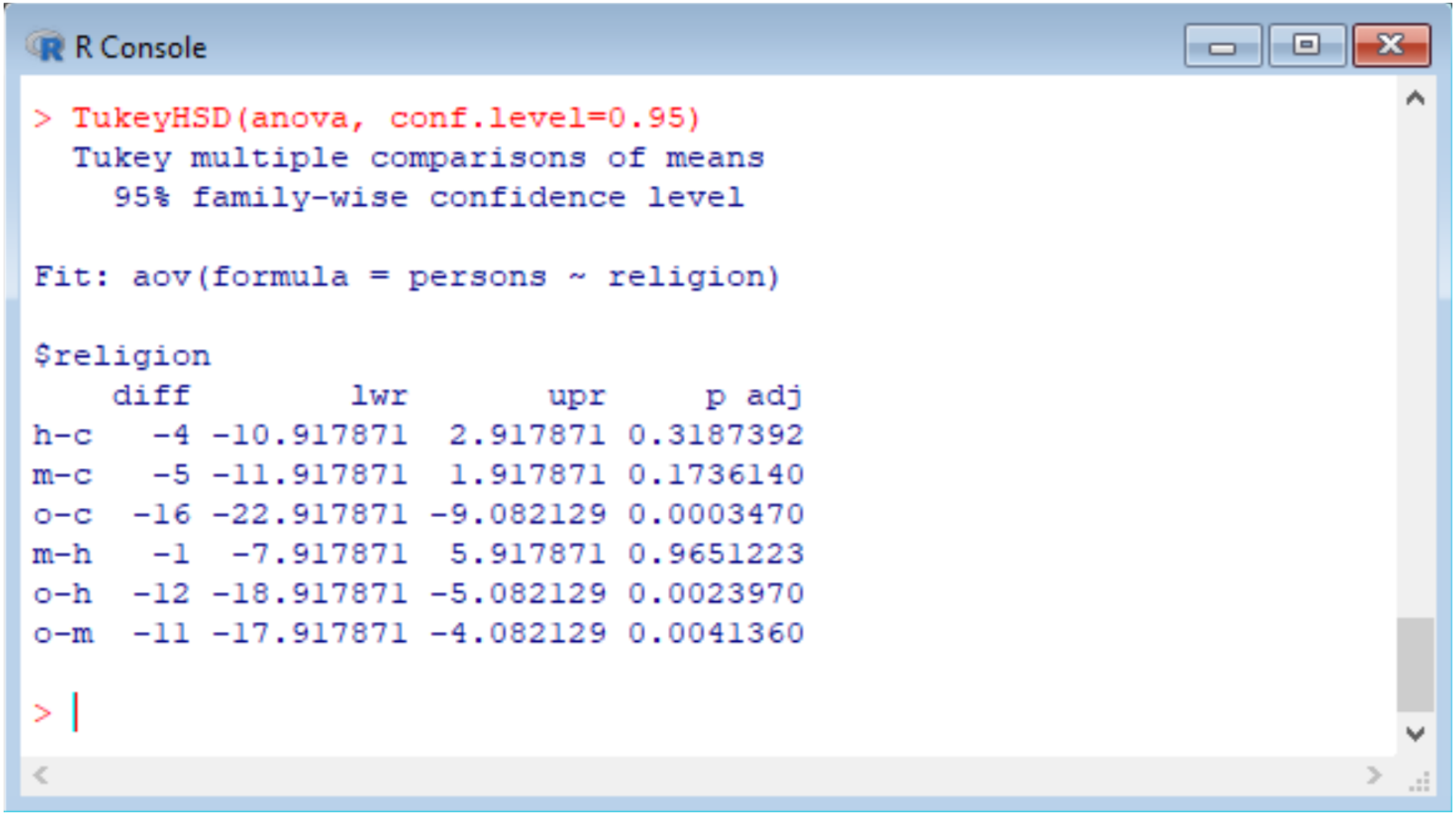
STEP 4: Interpret the results of the F test

1. The test statistic, F = 20.11 > critical value $F_{0.05,3,8} = 4.066$
2. The p-value (= 0.0004) < α (= 0.05)

Since the conditions are satisfied we reject H_0 at a 5% level of significance. Therefore at least one religious body has a significantly different mean.

STEP 5: Conduct Tukey's comparison test

```
> TukeyHSD(anova, conf.level=0.95)
```



By observing the confidence intervals and p-values we find the following:

Comparison	Confidence Interval	P-value	Decision
h vs c	contains 0	> 0.05	do not reject H_0
m vs c	contains 0	> 0.05	do not reject H_0
o vs c	doesn't contain 0	< 0.05	reject H_0
m vs h	contains 0	> 0.05	do not reject H_0
o vs h	doesn't contain 0	< 0.05	reject H_0
o vs m	doesn't contain 0	< 0.05	reject H_0

At a 5% level of significance the religious group *other* is significantly different from the religious groups *hindu*, *christian* and *muslim*.

NB: Explain Mean difference, O and H, say

Kruskal Wallis (non-parametric distribution)

Perform the Shapiro-Wilk Test before to determine whether the data comes from a normal distribution.
The kruskal wallis test is a **non-parametric** test used to determine if there are any significant differences in means.

The general hypotheses are (as before):

H_0: $\mu_1 = \mu_2 = ... = \mu_a$ versus H_1: At least one μ_i is different

where a is the number of treatments or groups.

We reject the null hypothesis H_0 when,

1. The test statistic $>$ critical value
2. The p-value $< \alpha$

The Kruskal Wallis test uses the process of ranking to compute the related test statistic where ranks are obtained by assigning a value to each observation in the data set arranged in ascending order.

The test is approximated by a chi-square distribution with $a - 1$ degrees of freedom (df = a-1).

The critical value $\chi^2_{\alpha,df}$ is simulated by:

`qchisq(1 - α, df = a - 1)`

Performing Kruskal Wallis test in R

`kruskal.test(data ~ labels)`

where *data* is a numeric vector and *labels* is a vector of labels corresponding to the *data* vector.

Similarly, if the kruskal wallis test rejects H_0 concluding that at least one mean is different a comparison test can be performed to determine which of the treatment means are different.

Tukey's Comparison test in R

`TukeyHSD(aov(data ~ labels), conf.level = CL)`

where the hypothesis test H_0: $\mu_i = \mu_j$ vs H_1: $\mu_i \neq \mu_j$ rejects H_0 if the confidence interval does not include 0 and the adjusted p-value $< \alpha$.

Example 1: The R built-in data set *chickwts* measures the effectiveness of different feed supplements on the weight of newly hatched chicks.

(i) By using a non-parametric approach at a 10% level of significance, determine whether feed type has a significant effect on weight.

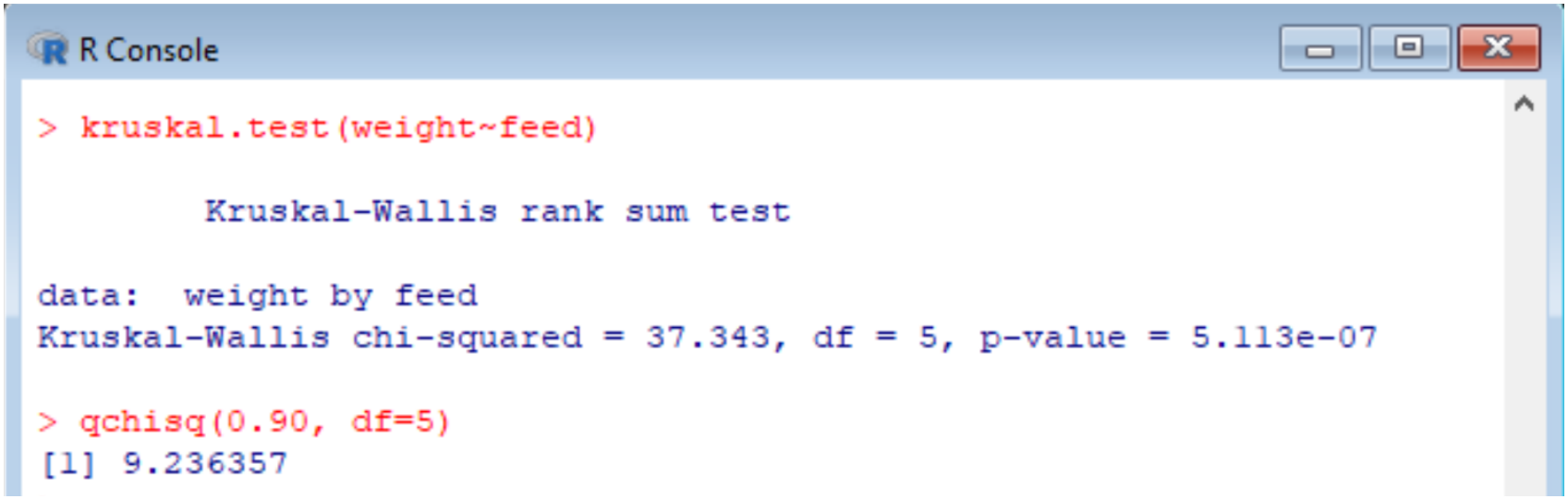
(ii) At the same level of significance, deduce which feeds are different.

(i) Required to test,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ vs H_1: At least one μ_i is different (there are 6 feed types, a = 6)

Observing data set and performing kruskal wallis test:

```
> attach(chickwts)
> kruskal.test(weight ~ feed)
> qchisq(0.90, df=5)
```



The test statistic ($= 37.343$) $>$ critical value ($\chi^2_{0.10,5} = 9.236$) and the p-value ($= 5.113 \times 10^{-7}$) < 0.10. We can reject H_0 concluding that at least one feed type has a different mean.

(ii) Tukey's comparison test:

```
> TukeyHSD(aov(weight ~ feed), conf.level=0.90)
```

R Console

```
> TukeyHSD(aov(weight~feed), conf.level=0.90)
  Tukey multiple comparisons of means
    90% family-wise confidence level

Fit: aov(formula = weight ~ feed)

$feed
                          diff         lwr         upr     p adj
horsebean-casein    -163.383333 -225.614094 -101.15257 0.0000000
linseed-casein      -104.833333 -164.168035  -45.49863 0.0002100
meatmeal-casein      -46.674242 -107.342475   13.99399 0.3324584
soybean-casein       -77.154762 -134.331112  -19.97841 0.0083653
sunflower-casein       5.333333  -54.001369   64.66804 0.9998902
linseed-horsebean     58.550000   -3.680761  120.78076 0.1413329
meatmeal-horsebean   116.709091   53.205586  180.21260 0.0001062
soybean-horsebean     86.228571   26.052200  146.40494 0.0042167
sunflower-horsebean  168.716667  106.485906  230.94743 0.0000000
meatmeal-linseed      58.159091   -2.509142  118.82732 0.1276965
soybean-linseed       27.678571  -29.497778   84.85492 0.7932853
sunflower-linseed    110.166667   50.831965  169.50137 0.0000884
soybean-meatmeal     -30.480519  -89.039571   28.07853 0.7391356
sunflower-meatmeal    52.007576   -8.660657  112.67581 0.2206962
sunflower-soybean     82.488095   25.311746  139.66444 0.0038845
```

By observing the confidence intervals and p-values of the comparisons made in the R output above, the following is a list of the pairs of feed types that have significantly different means at a 10% level of significance.

1. horsebean and casein
2. linseed and casein
3. soybean and casein
4. meatmeal and horsebean
5. soybean and horsebean
6. sunflower and horsebean
7. sunflower and linseed
8. sunflower and soybean

where different means are a result of the confidence interval not including 0 and the p-value < 0.10.

Example 2: The North Regional Health Authority (NRHA) in Trinidad has a policy whereby any patient admitted of suspected chronic obstructive pulmonary disease (COPD) is automatically placed in the ICU. The data below gives the number of hours spent in the ICU by such patients in 4 hospitals in the region.

Hospital 1	59	60	59	55	50
Hospital 2	57	57	46	54	51
Hospital 3	45	53	49	51	46
Hospital 4	44	45	39	58	44

Use a non parametric approach for analysis of variance at a 1% level of significance.

Required to test

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: At least one μ_i mean is different

STEP 1: Enter the data in R.
This can be done manually or in a data file which is then imported into R.

Option 1: Manually

```
> hours = c(59,60,59,55,50,57,57,46,54,51,45,
    53,49,51,46,44,45,39,58,44)
> hospital = c(rep("1",5),rep("2",5),rep("3",5),
    rep("4",5))
```

Notice that the labels require quotation " " even if they are numbers.

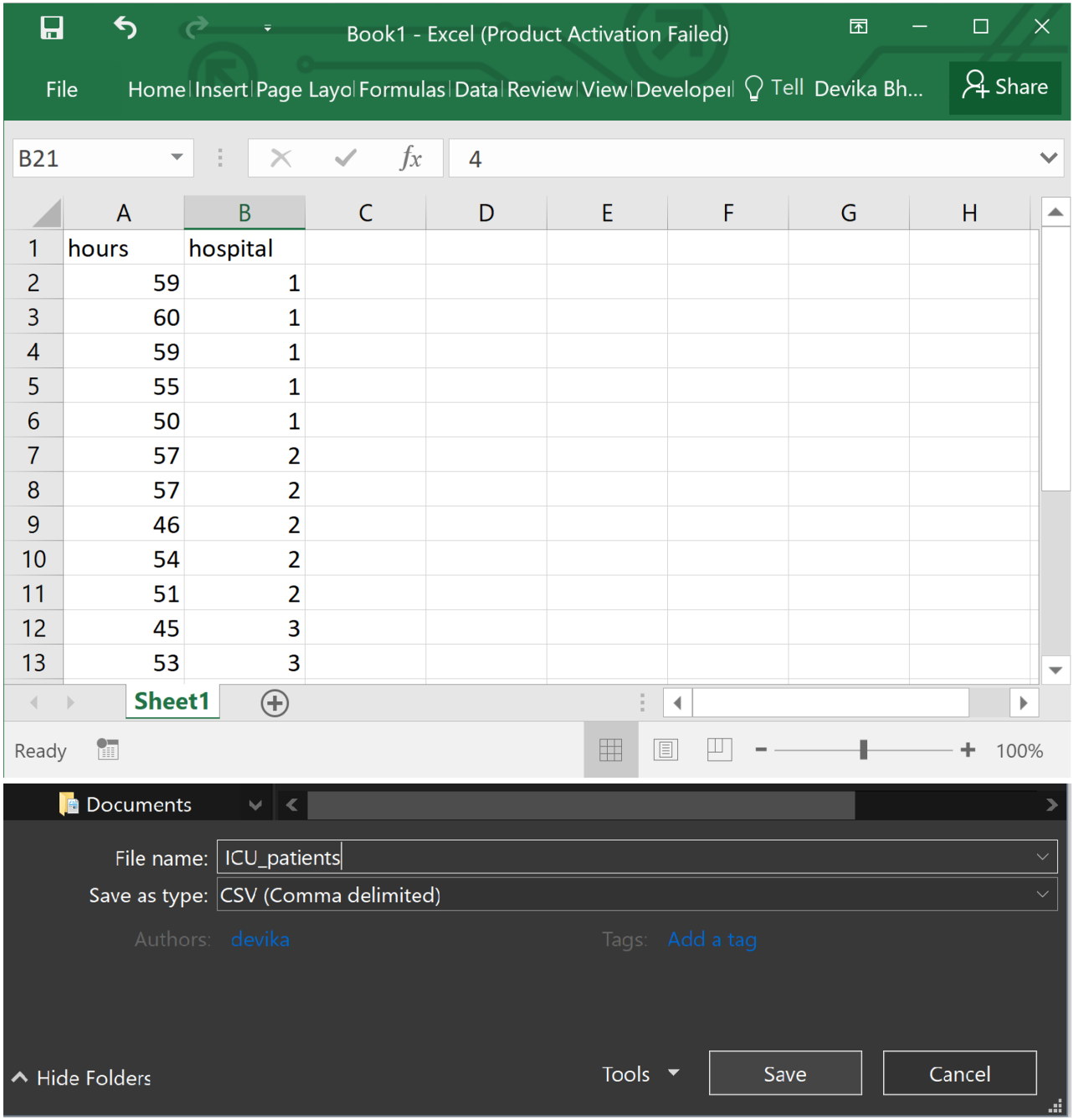
Option 2: Data file - example csv

Enter the data as two columns in Excel and save as a *csv* file.

(see next slide for demonstration)

Then in R, change the directory to the location of the file and import:

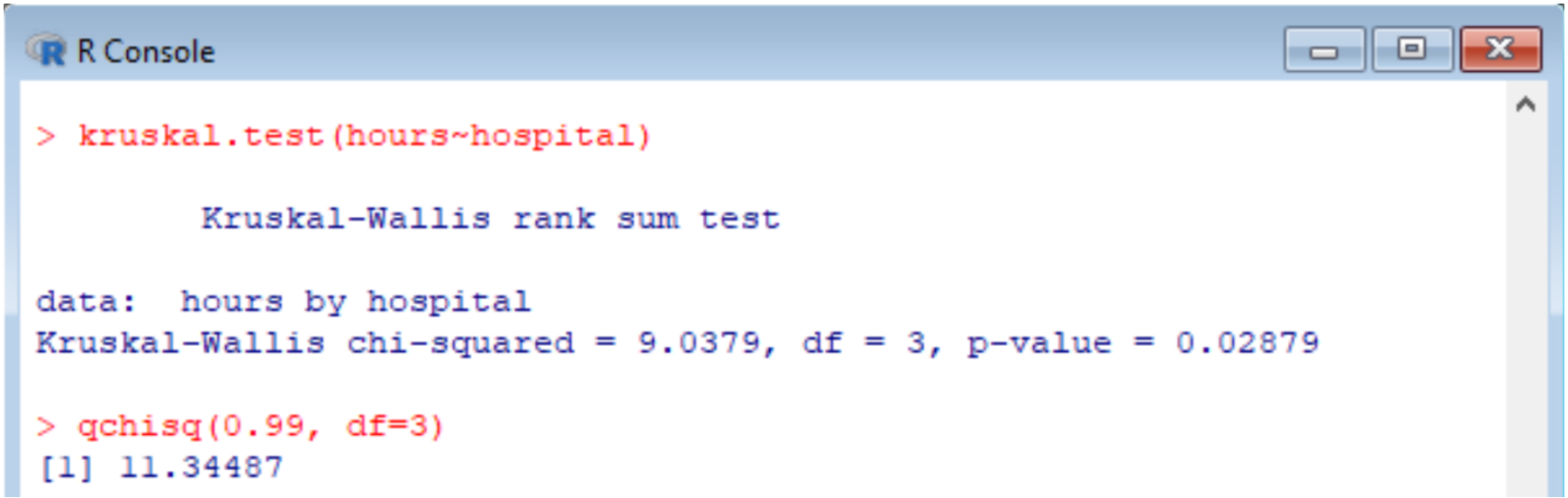
```
> icu = read.csv("ICU_patients.csv", header=T)
> attach(icu)
```



	A	B	C	D	E	F	G	H
1	hours	hospital						
2	59	1						
3	60	1						
4	59	1						
5	55	1						
6	50	1						
7	57	2						
8	57	2						
9	46	2						
10	54	2						
11	51	2						
12	45	3						
13	53	3						

STEP 2: Perform the kruskal wallis test at $\alpha = 0.01$

```
> kruskal.test(hours ~ hospital)
> qchisq(0.99, df=3)
```



Since test statistic 9.038 $<$ critical value 11.345 and p-value ($= 0.029$) > 0.01 we fail to reject H_0 at a 1% level of significance. Therefore the mean number of hours spent in ICU is the same in all 4 hospitals.

Example 3: The *cereal.csv* data set provided shows the properties of 16 popular brands of cereal.

(i) Conduct the kruskal wallis test at $\alpha = 0.05$ to determine if the mean cost of cereal per 100g is different for the 4 shelf locations.

(ii) If necessary, perform a comparison test to determine if the mean cost of cereal per 100g differs with shelf number.

(i) Required to test,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: At least one μ_i mean is different

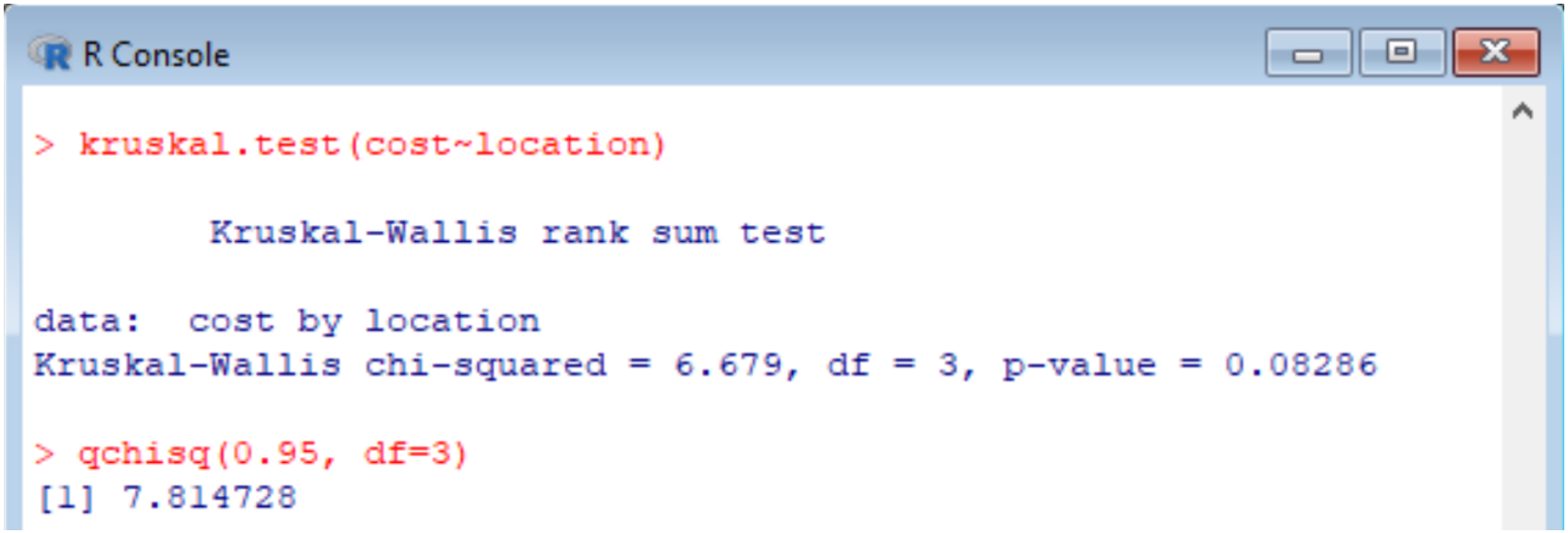
Read and observe the data in R then perfom the kruskal wallis test.

Consider using the square brackets notation to extract the relevant data and labels columns.

```
> cereal = read.csv("cereal.csv", header=T)
> cost = cereal[,2]
> location = cereal[,9]
```

We can also use the attach command or $ to extract the columns.

```
> kruskal.test(cost ~ location)
> qchisq(0.95, df=3)
```



Since test statistic 6.679 $<$ critical value 7.815 and p-value ($= 0.083$) > 0.05 we fail to reject H_0 at a 5% level of significance. The mean cost of cereal per 100g is the same for all 4 shelf locations.

(ii) Since we failed to reject H_0, a comparison test is not necessary. Note that performing this test will result in all mean comparisons being equal.