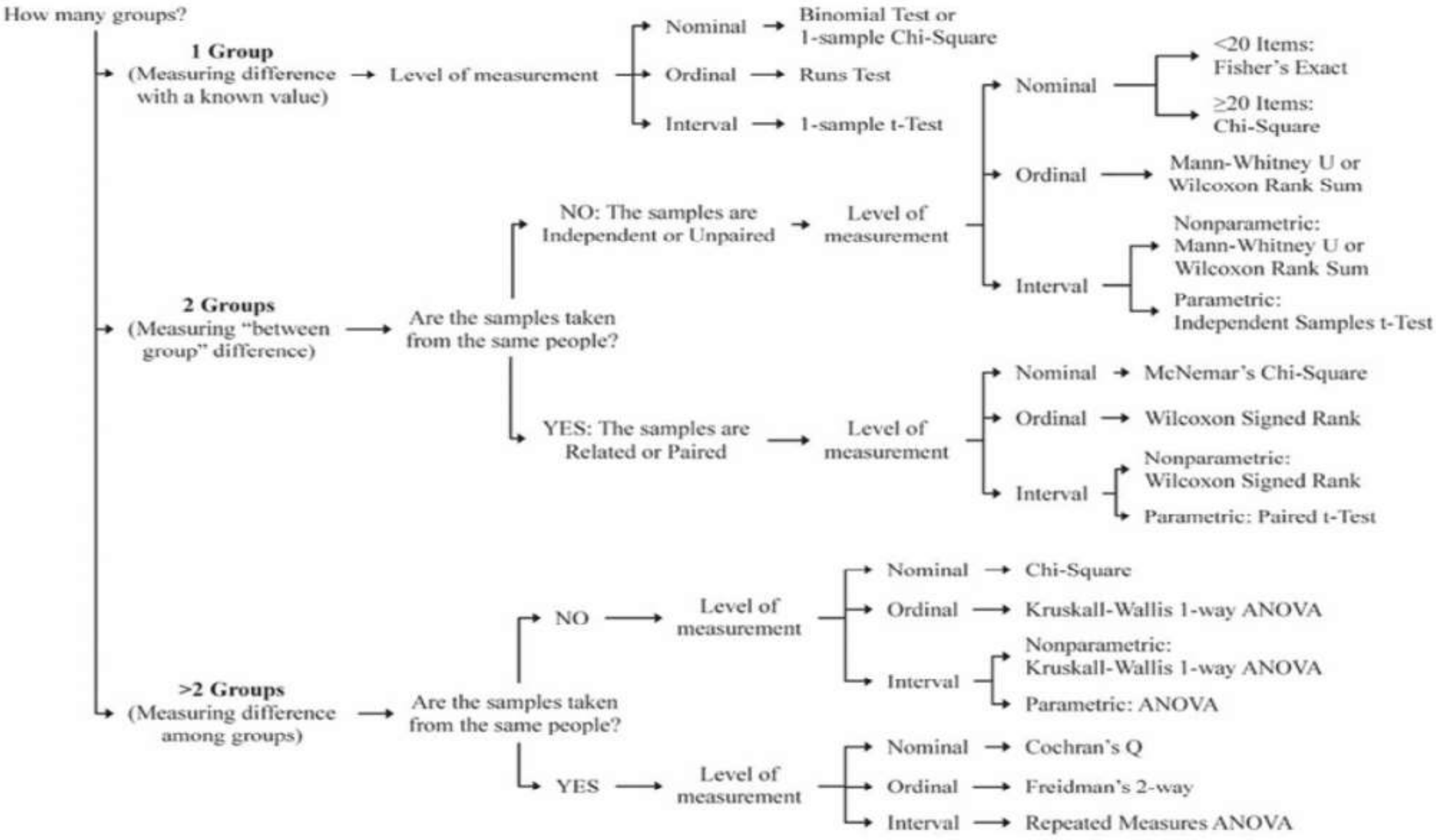
Medical Statistics with R

Dr. Gulser Caliskan
Prof. Giuseppe Verlato

Unit of Epidemiology and Medical Statistics
Department of Diagnostics and Public Health
University of Verona, Italy

LESSON 4 INDEX

1. Independent T-Test
2. Wilcoxon Rank-Sum (Mann-Whitney U Test)
3. Paired T-Test
4. Wilcoxon Sign Test
5. ANOVA (Analysis of Varyans)
6. Kruskal Wallis H Test



How many groups?

1 Group (Measuring difference with a known value) → Level of measurement

Nominal → Binomial Test or 1-sample Chi-Square

Ordinal → Runs Test

Interval → 1-sample t-Test

2 Groups (Measuring "between group" difference) → Are the samples taken from the same people?

NO: The samples are Independent or Unpaired → Level of measurement

Nominal

<20 Items: Fisher's Exact

≥20 Items: Chi-Square

Ordinal → Mann-Whitney U or Wilcoxon Rank Sum

Interval

Nonparametric: Mann-Whitney U or Wilcoxon Rank Sum

Parametric: Independent Samples t-Test

YES: The samples are Related or Paired → Level of measurement

Nominal → McNemar's Chi-Square

Ordinal → Wilcoxon Signed Rank

Interval

Nonparametric: Wilcoxon Signed Rank

Parametric: Paired t-Test

>2 Groups (Measuring difference among groups) → Are the samples taken from the same people?

NO → Level of measurement

Nominal → Chi-Square

Ordinal → Kruskall-Wallis 1-way ANOVA

Interval

Nonparametric: Kruskall-Wallis 1-way ANOVA

Parametric: ANOVA

YES → Level of measurement

Nominal → Cochran's Q

Ordinal → Freidman's 2-way

Interval → Repeated Measures ANOVA

Figure 2. *Decision tree to identify inferential statistics for measuring a difference.*

Classical Hypothesis Testing

Test of a ***null hypothesis*** against an ***alternative hypothesis***. There are five steps, the first four of which should be done before inspecting the data.

- Step 1. Declare the null hypothesis H_0 and the alternative hypothesis H_1.

Types Of Hypotheses

A hypothesis that completely specifies the parameters is called ***simple***. If it leaves some parameter undetermined it is composite. A hypothesis is ***one-sided*** if it proposes that a parameter is > some value or < some value; it is ***two-sided*** if it simply says the parameter is /= some value.

Types of Error

Rejecting H_0 when it is actually true is called a ***Type I Error***. In biomedical settings it can be considered a ***false positive***. (Null hypothesis says "nothing is happening" but we decide "there is disease".)

- Step 2. Specify an acceptable level of Type I error, α, normally 0.05 or 0.01. This is the threshold used in deciding to reject H_0 or not. If $\alpha = 0.05$ and we determine the probability of our data assuming H_0 is 0.0001, then we reject H_0.

The Test Statistic

➢Step 3. Select a test statistic.

This is a quantity calculated from the data whose value leads me to reject the null hypothesis or not. For matching sequences one choice would be the number of matches. For a contingency table compute Chi-squared. Normally compute the value of the statistic from the data assuming H_0 is true.

A great deal of theory, experience and care can go into selecting the right statistic.

The Critical Value or Region

- Step 4. Identify the values of the test statistic that lead to rejection of the null hypothesis.

Ensure that the test has the numerical value for type I error chosen in Step 2. For a one-sided alternative we normally find a value x so that only α = 0.05 values of the statistic are > x (or < x for an alternative in the other direction).

Obtain the Data and Execute

- Step 5. Obtain the data, calculate the value of the statistic assuming the null hypothesis and compare with the threshold.

P-Values (Substitute for Step 4)

Once the data are obtained calculate the null hypothesis probability of obtaining the observed value of the statistic or one more extreme. This is called the ***p-value***. If it is < the selected Type I Error threshold then we reject the null hypothesis.

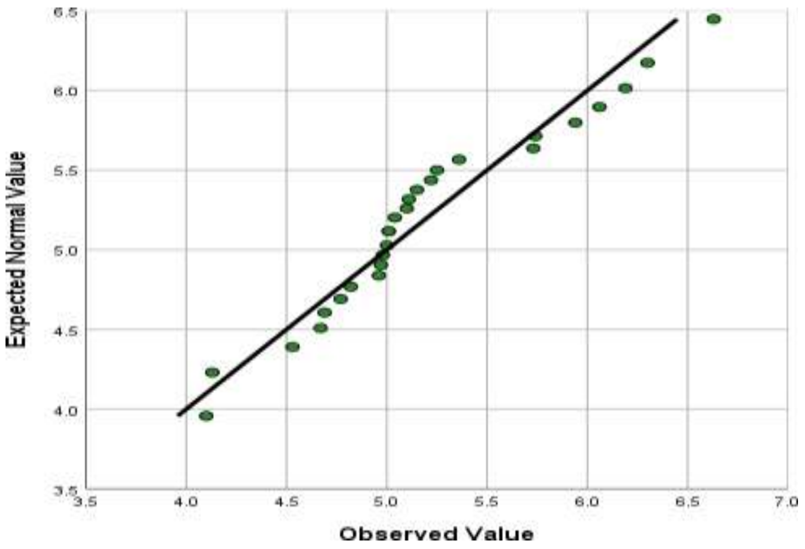
NORMALITY TESTS

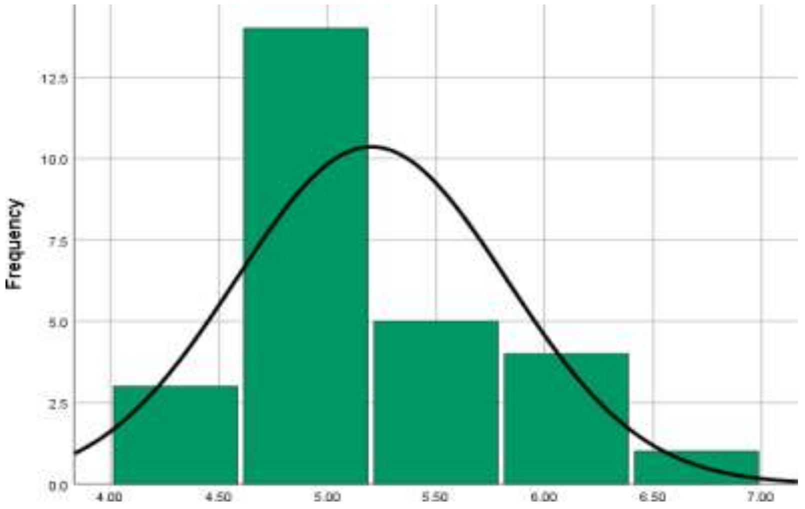
There are several methods of assessing whether data are normally distributed or not. They fall into two broad categories: *graphical* and *statistical*. The some common techniques are:

- ➢**Graphical** (Q probability plots, Cumulative frequency (P-P) plots, Histogram)
- ➢**Statistical** (W/S test, Jarque-Bera test, Shapiro-Wilks test, Kolmogorov-Smirnov test, D’Agostino test, Lilliefors test, Spiegelhalter’s T’ test, Anderson-Darling Test)

Different normality tests produce different probabilities. This is due to where in the distribution (central, tails) or what moment (skewness, kurtosis) they are examining.

Normality Test	Statistic	Probability	Results
W/S	4.05	> 0.05	Normal
Jarque-Bera	1.209	0.5463	Normal
D’Agostino	0.2734	> 0.05	Normal
Shapiro-Wilk	0.9428	0.1429	Normal
Kolmogorov-Smirnov	1.73	0.0367	Not-normal
Anderson-Darling	0.7636	0.0412	Not-normal
Lilliefors	0.1732	0.0367	Not-normal





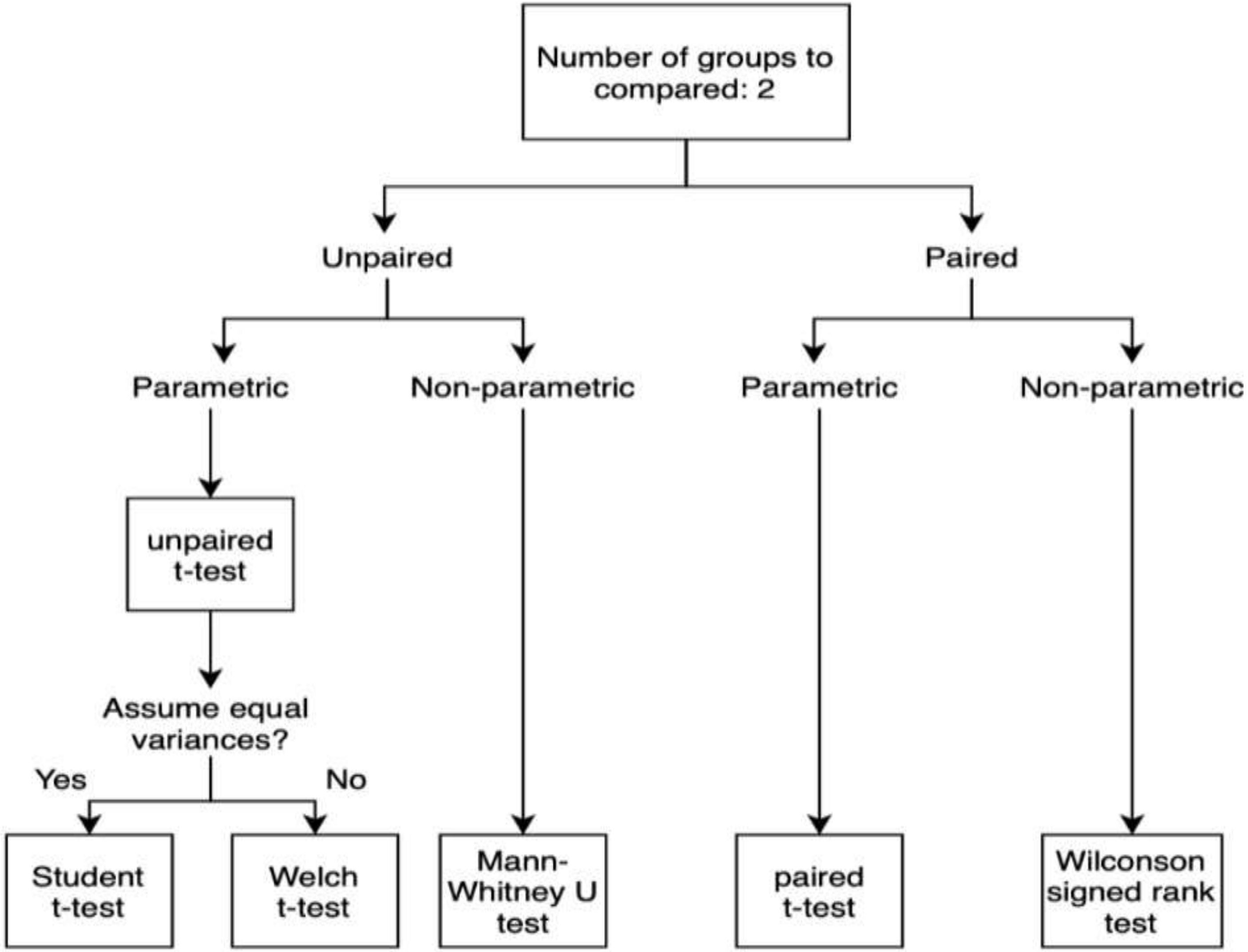
- **W/S or studentized range (q):**
    - Simple, very good for symmetrical distributions and short tails.
    - Very bad with asymmetry.
- **Shapiro Wilk (W):**
    - Fairly powerful omnibus test. Not good with small samples or discrete data.
    - Good power with symmetrical, short and long tails. Good with asymmetry.
- **Jarque-Bera (JB):**
    - Good with symmetric and long-tailed distributions.
    - Less powerful with asymmetry, and poor power with bimodal data.
- **D’Agostino (D or Y):**
    - Good with symmetric and very good with long-tailed distributions.
    - Less powerful with asymmetry.
- **Anderson-Darling (A):**
    - Similar in power to Shapiro-Wilk but has less power with asymmetry.
    - Works well with discrete data.
- **Distance tests (Kolmogorov-Smirnov, Lillifors, Chi^2):**
    - All tend to have lower power. Data have to be very non-normal to reject Ho.
    - These tests can outperform other tests when using discrete or grouped data.

Tests for normality calculate the probability that the sample was drawn from a normal population.

Null hypothesis H_0: Data follow a normal distribution

Alternative hypothesis H_1: Data do not follow a normal distribution

TWO SAMPLE COMPARISONS



Number of groups to compared: 2

Unpaired

Paired

Parametric

Non-parametric

Parametric

Non-parametric

unpaired t-test

Assume equal variances?

Yes

No

Student t-test

Welch t-test

Mann-Whitney U test

paired t-test

Wilconson signed rank test

INDEPENDENT T-TEST

The independent t-test, also called the two sample **t-test**, **independent-samples t-test** or **student's t-test**, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.

The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal:

$\mathbf{H_0: \mu_1 = \mu_2}$

In most cases, we are looking to see if we can show that we can reject the null hypothesis and accept the alternative hypothesis, which is that the population means are not equal:

$\mathbf{H_A: \mu_1 \neq \mu_2}$

To do this, we need to set a significance level (also called alpha) that allows us to either reject or accept the alternative hypothesis. Most commonly, this value is set at 0.05.

ASSUMPTIONS

Assumption Of Normality Of The Dependent Variable

The independent t-test requires that the dependent variable is approximately normally distributed within each group. You can test for this using a number of different tests, but the Shapiro-Wilks test of normality or a graphical method, such as a Q-Q Plot, are very common.

Assumption Of Homogeneity Of Variance

The independent t-test assumes the variances of the two groups you are measuring are equal in the population. If your variances are unequal, this can affect the Type I error rate. The assumption of homogeneity of variance can be tested using Barlett’s test or Levene's Test of Equality of Variances.

Overcoming A Violation Of The Assumption Of Homogeneity Of Variance

If the Levene's Test for Equality of Variances is statistically significant, which indicates that the group variances are unequal in the population, you can correct for this violation by not using the pooled estimate for the error term for the ***t*-statistic**, but instead using an adjustment to the degrees of freedom using the ***Welch-Satterthwaite method***.

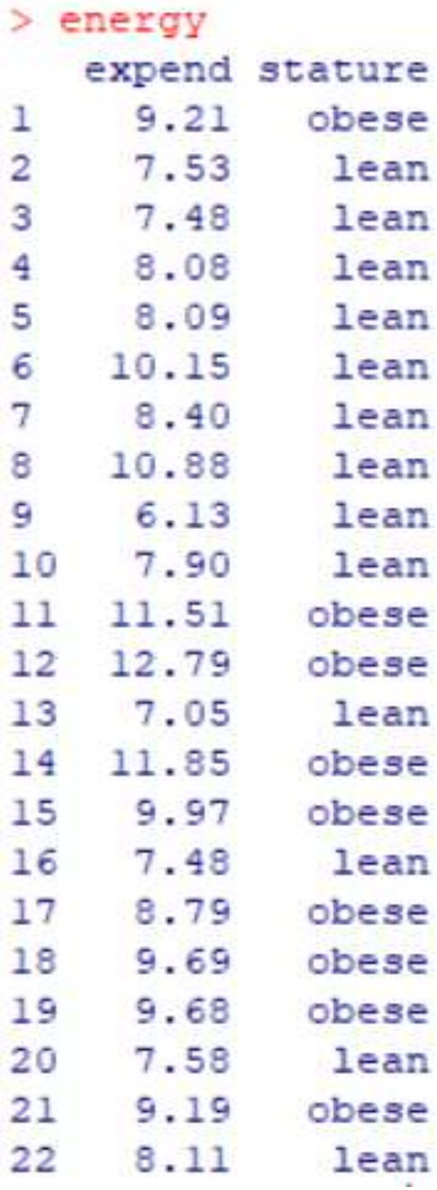
This test for homogeneity of variance provides an ***F*-statistic** and a significance value (***p*-value**).

We are primarily concerned with the significance value – if it is greater than **0.05** (i.e., ***p* > .05**), our group variances can be treated as equal.

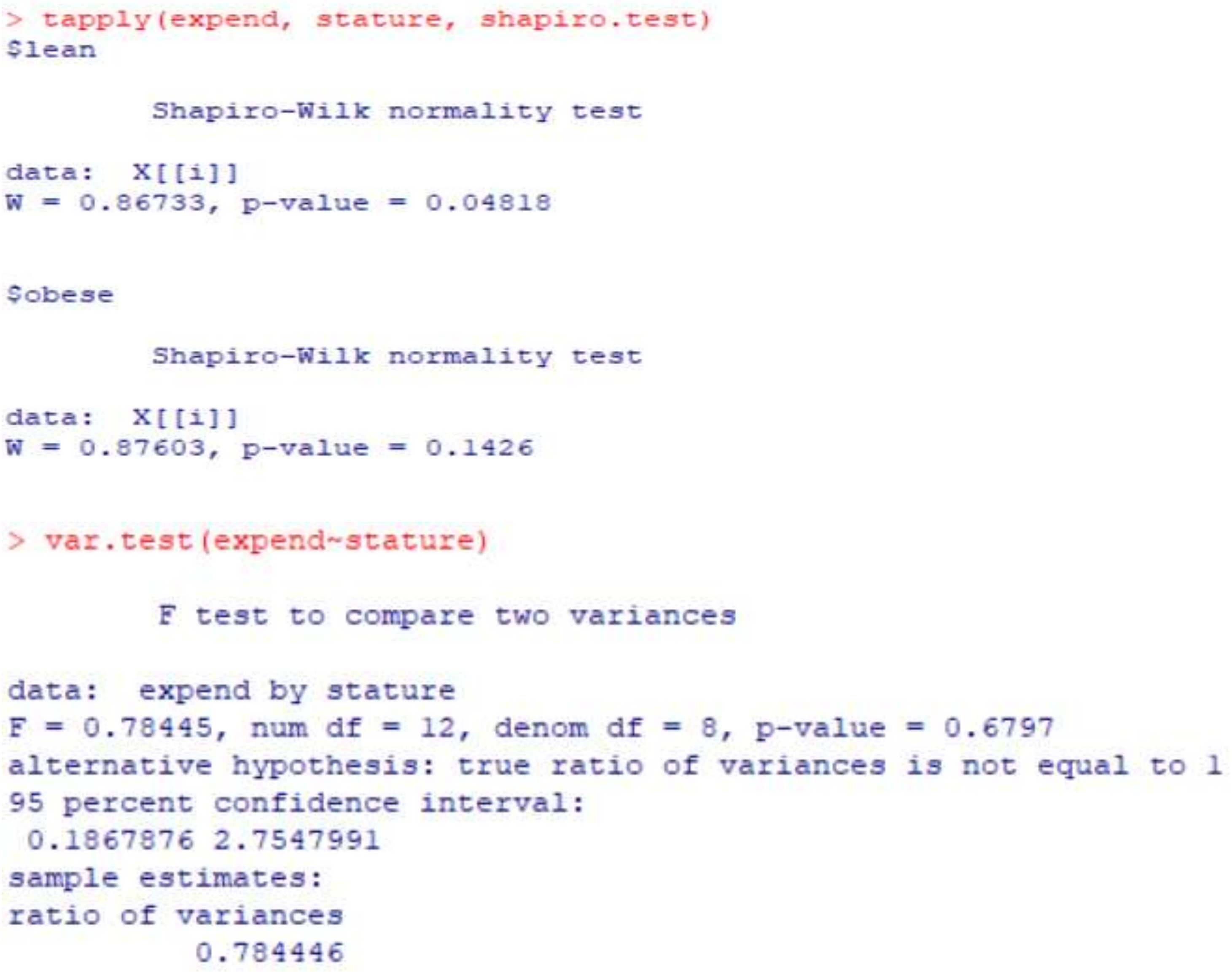
However, if ***p* < 0.05**, we have unequal variances and we have violated the assumption of homogeneity of variances.

In this case, we therefore do not accept the alternative hypothesis and accept that there are no statistically significant differences between means. This would not have been our conclusion had we not tested for homogeneity of variances.

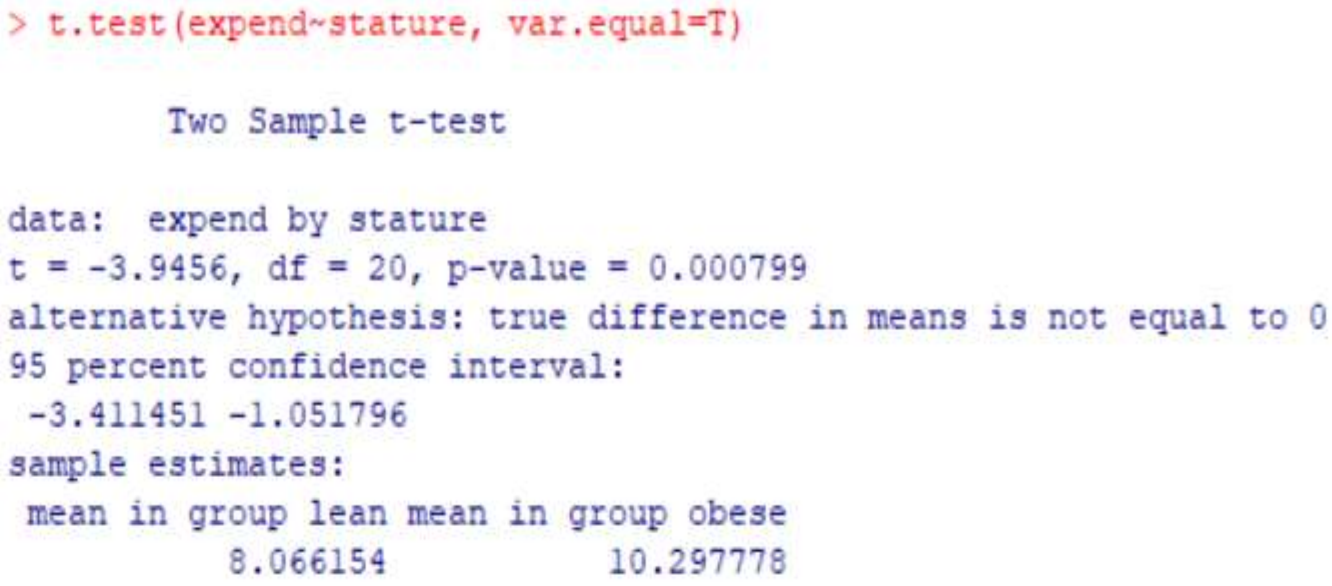
EXAMPLE:



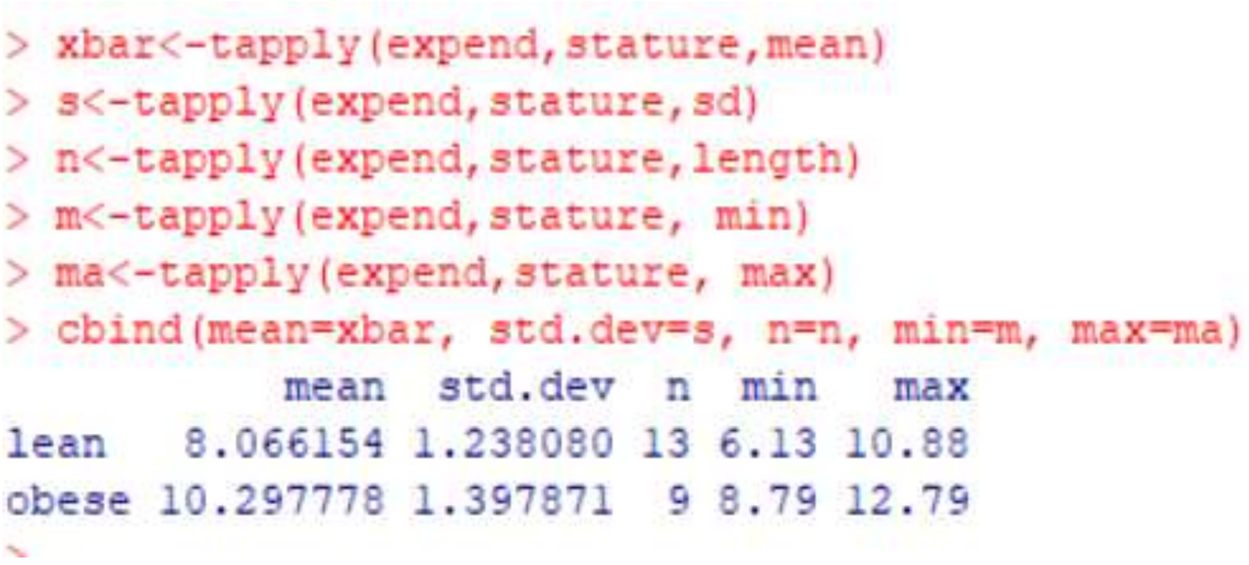
```
> energy
   expend stature
1    9.21   obese
2    7.53    lean
3    7.48    lean
4    8.08    lean
5    8.09    lean
6   10.15    lean
7    8.40    lean
8   10.88    lean
9    6.13    lean
10   7.90    lean
11  11.51   obese
12  12.79   obese
13   7.05    lean
14  11.85   obese
15   9.97   obese
16   7.48    lean
17   8.79   obese
18   9.69   obese
19   9.68   obese
20   7.58    lean
21   9.19   obese
22   8.11    lean
```



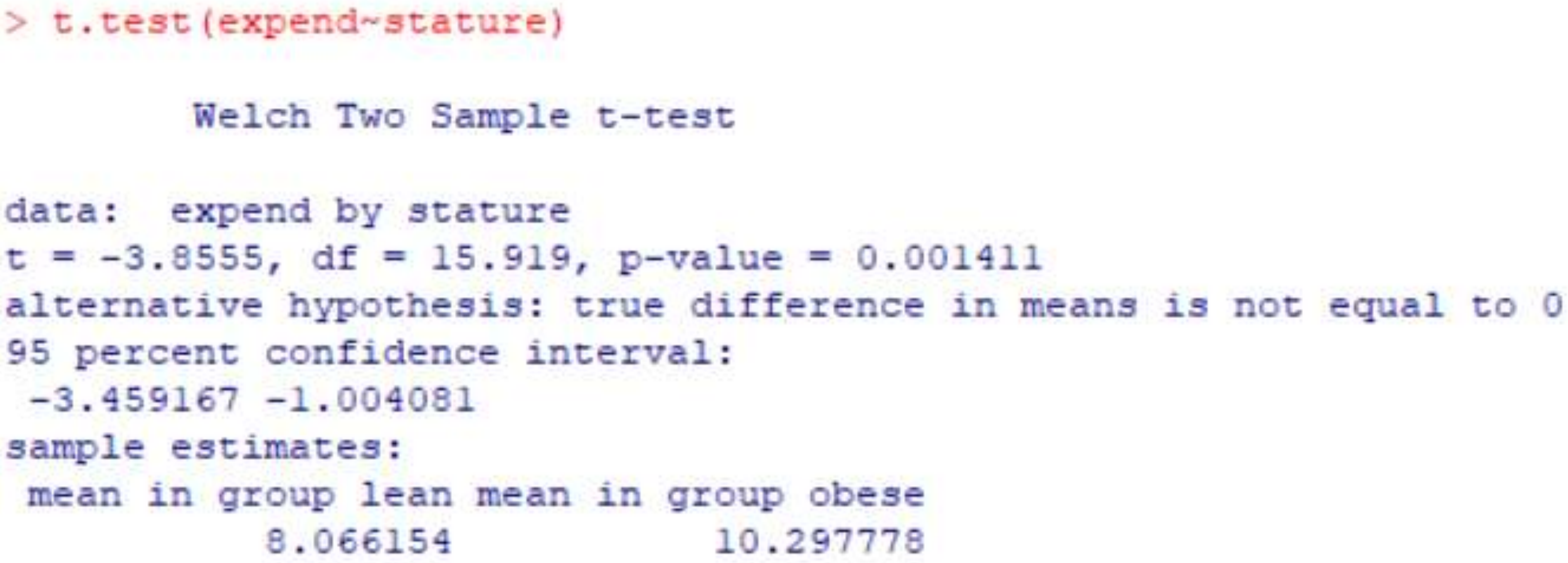
```
> tapply(expend, stature, shapiro.test)
$lean

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.86733, p-value = 0.04818


$obese

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87603, p-value = 0.1426


> var.test(expend~stature)

        F test to compare two variances

data:  expend by stature
F = 0.78445, num df = 12, denom df = 8, p-value = 0.6797
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1867876 2.7547991
sample estimates:
ratio of variances
          0.784446
```



```
> t.test(expend~stature, var.equal=T)

        Two Sample t-test

data:  expend by stature
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```



```
> xbar<-tapply(expend,stature,mean)
> s<-tapply(expend,stature,sd)
> n<-tapply(expend,stature,length)
> m<-tapply(expend,stature, min)
> ma<-tapply(expend,stature, max)
> cbind(mean=xbar, std.dev=s, n=n, min=m, max=ma)
           mean  std.dev  n  min   max
lean   8.066154 1.238080 13 6.13 10.88
obese 10.297778 1.397871  9 8.79 12.79
>
```



```
> t.test(expend~stature)

        Welch Two Sample t-test

data:  expend by stature
t = -3.8555, df = 15.919, p-value = 0.001411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.459167 -1.004081
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

THE MANN-WHITNEY *U* TEST

The Mann-Whitney *U* test (also called the **Mann–Whitney–Wilcoxon** (**MWW**), **Wilcoxon rank-sum test**, or **Wilcoxon–Mann–Whitney test**) is a nonparametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed.

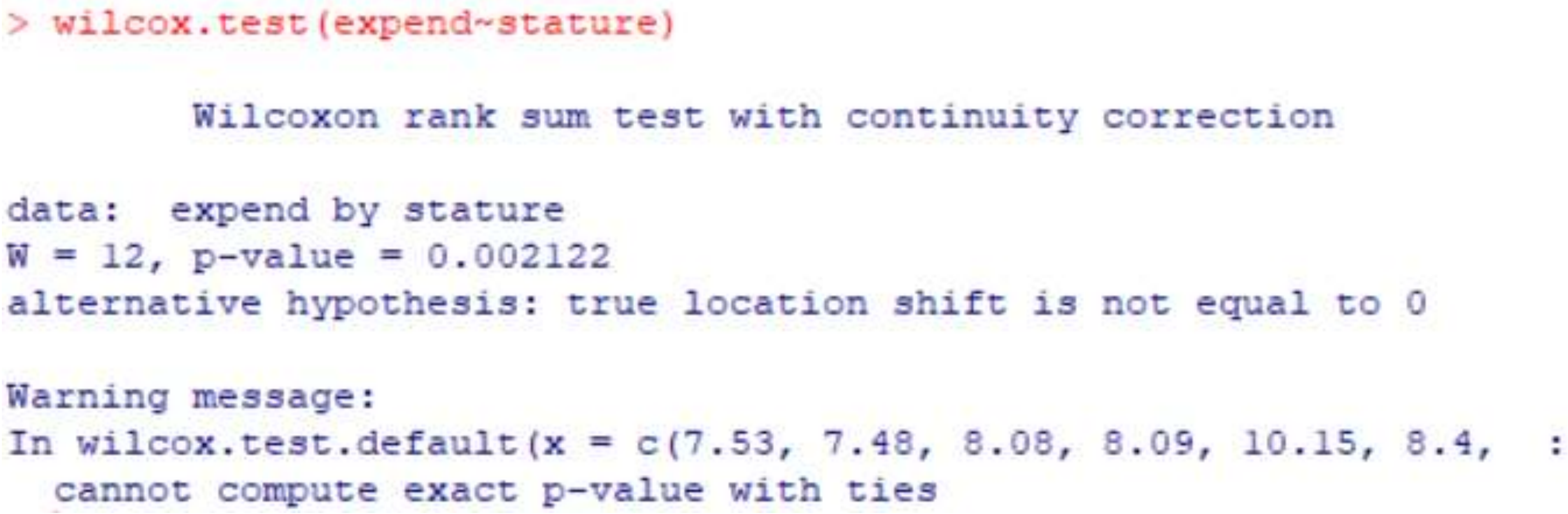
Requirements:

- Two random, independent samples
- The data is continuous - in other words, it must, in principle, be possible to distinguish between values at the nth decimal place
- Scale of measurement should be ordinal, interval or ratio
- For maximum accuracy, there should be no ties, though this test - like others - has a way to handle ties

Null Hypothesis:

The null hypothesis asserts that the *medians* of the two samples are identical.

Example



```
> wilcox.test(expend~stature)

        Wilcoxon rank sum test with continuity correction

data:  expend by stature
W = 12, p-value = 0.002122
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4,   :
  cannot compute exact p-value with ties
```

PAIRED T TEST

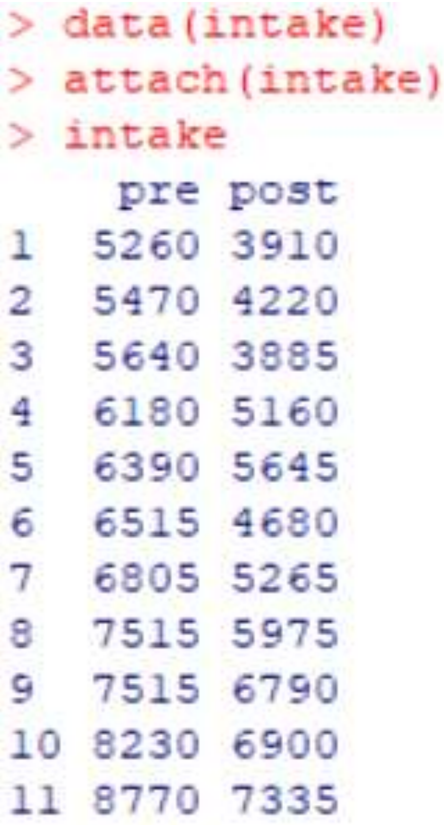
The ***paired sample t-test***, sometimes called the ***dependent sample t-test***, is used when there are two measurements on the same experimental unit. In a paired sample *t*-test, each subject or entity is measured twice, resulting in *pairs* of observations.

Common applications of the ***paired sample t-test*** include repeated-measures designs. One approach you might consider would be to measure the performance of a sample of patients before and after completing the treatment and analyze the differences using a paired sample t-test.

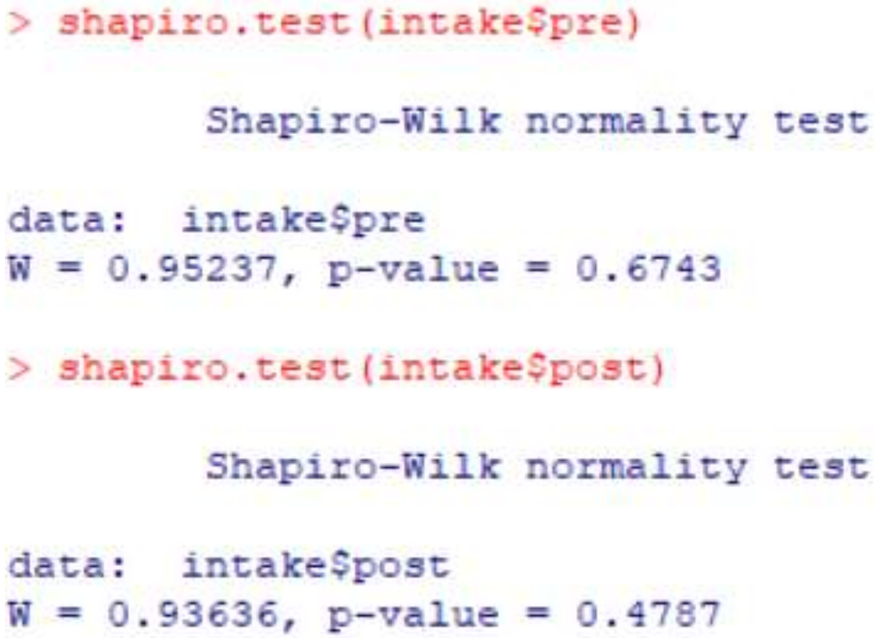
The ***paired sample t-test*** has four main assumptions:

- The dependent variable must be continuous (interval/ratio).
- The observations are independent of one another.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

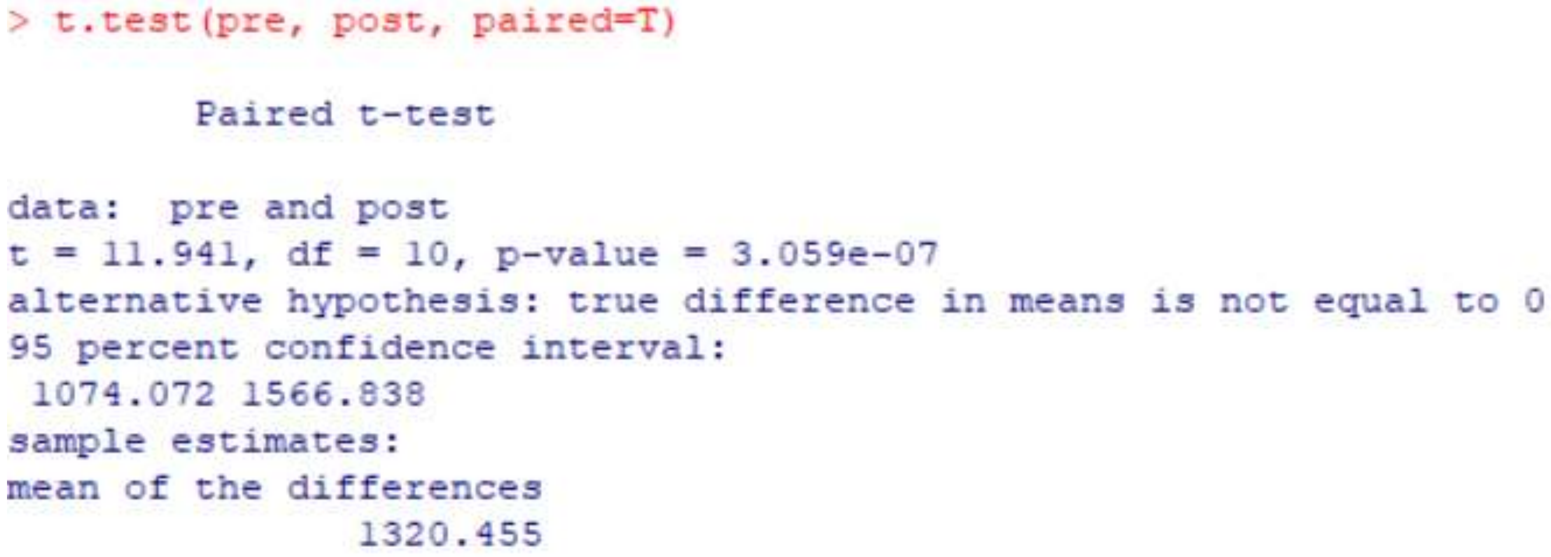
Example



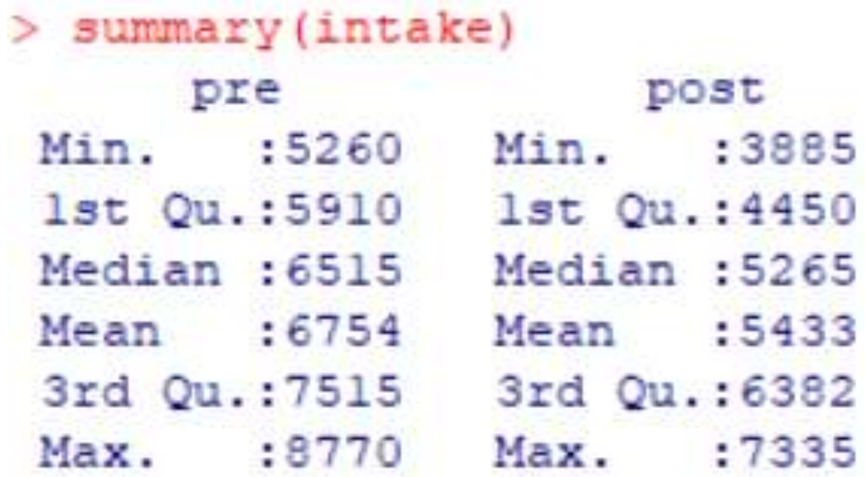
```
> data(intake)
> attach(intake)
> intake
    pre post
1  5260 3910
2  5470 4220
3  5640 3885
4  6180 5160
5  6390 5645
6  6515 4680
7  6805 5265
8  7515 5975
9  7515 6790
10 8230 6900
11 8770 7335
```



```
> shapiro.test(intake$pre)

        Shapiro-Wilk normality test

data:  intake$pre
W = 0.95237, p-value = 0.6743

> shapiro.test(intake$post)

        Shapiro-Wilk normality test

data:  intake$post
W = 0.93636, p-value = 0.4787
```



```
> t.test(pre, post, paired=T)

        Paired t-test

data:  pre and post
t = 11.941, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
               1320.455
```



```
> summary(intake)
      pre             post
 Min.   :5260   Min.   :3885
 1st Qu.:5910   1st Qu.:4450
 Median :6515   Median :5265
 Mean   :6754   Mean   :5433
 3rd Qu.:7515   3rd Qu.:6382
 Max.   :8770   Max.   :7335
```

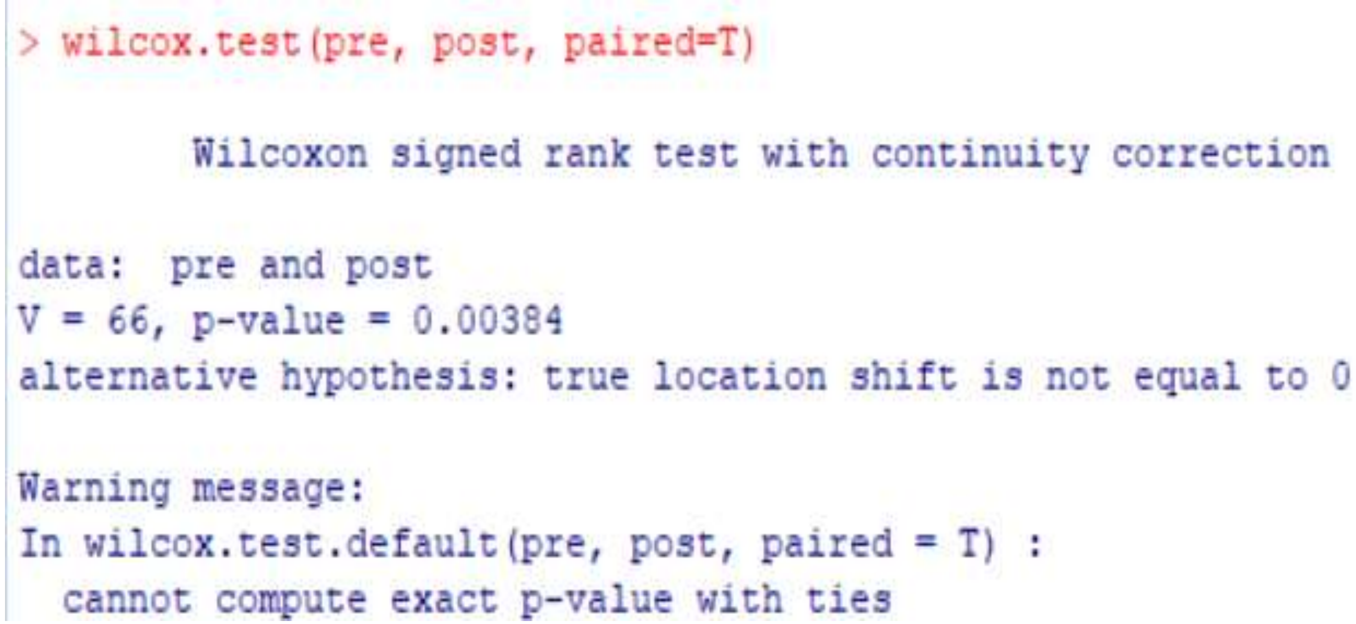
THE WILCOXON SIGNED RANK TEST

The Wilcoxon signed rank test is the non-parametric of the dependent t-test. Because the dependent samples t-test analyzes if the average difference of two repeated measures is zero, it requires metric (interval or ratio) and normally distributed data; the Wilcoxon Sign Test uses ranked or ordinal data; thus, it is a common alternative to the dependent samples t-test when its assumptions are not met.

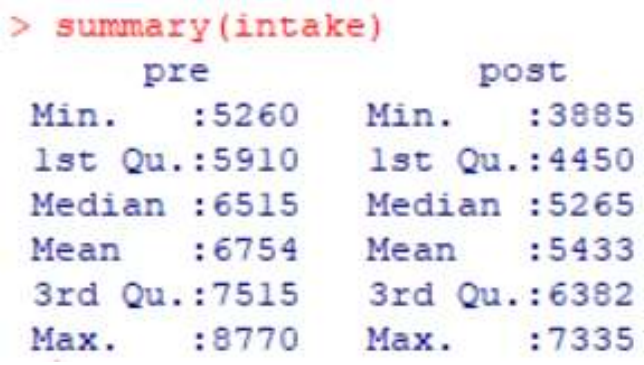
The **Wilcoxon signed-rank test** has three main assumptions:

- The dependent variable should be measure the ordinal or continuous level
- Data are paired and come from the same population.
- Each pair is chosen randomly and independently

Example



```
> wilcox.test(pre, post, paired=T)

        Wilcoxon signed rank test with continuity correction

data:  pre and post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(pre, post, paired = T) :
  cannot compute exact p-value with ties
```



```
> summary(intake)
      pre              post     
 Min.   :5260   Min.   :3885  
 1st Qu.:5910   1st Qu.:4450  
 Median :6515   Median :5265  
 Mean   :6754   Mean   :5433  
 3rd Qu.:7515   3rd Qu.:6382  
 Max.   :8770   Max.   :7335  
```

One-way ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

where ***μ* = group mean** and ***k* = number of groups**. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (H_A), which is that there are at least two group means that are statistically significantly different from each other.

At this point, it is important to realize that the one-way ANOVA is an **omnibus** test statistic and cannot tell you which specific groups were statistically significantly different from each other, only that at least two groups were.

To determine which specific groups differed from each other, you need to use a **Post Hoc Test**.

The Three Assumptions of ANOVA

1.) Assumption of independence

ANOVA assumes that the observations are random and that the samples taken from the groups are independent of each other. One event should not depend on another; that is, the value of one observation should not be related to any other observation.

2-) Assumption of homogeneity of variance

ANOVA assumes that the variances of the distributions in the groups are equal. Remember, the purpose of the ANOVA test is to determine the plausability of the null hypothesis, where the null hypothesis says that all observations come from the same underlying group with the same degree of variability.

Therefore, if the variances of each group differ from the outset, then the null hypothesis will be rejected (within certain limits) and thus there is no point in using ANOVA in the first place.

3-)Assumption of normality

ANOVA is based on the F-statistic, where the F-statistic requires that the dependent variable is normally distributed in each group. Thus, ANOVA requires that the dependent variable is normally distributed in each group.

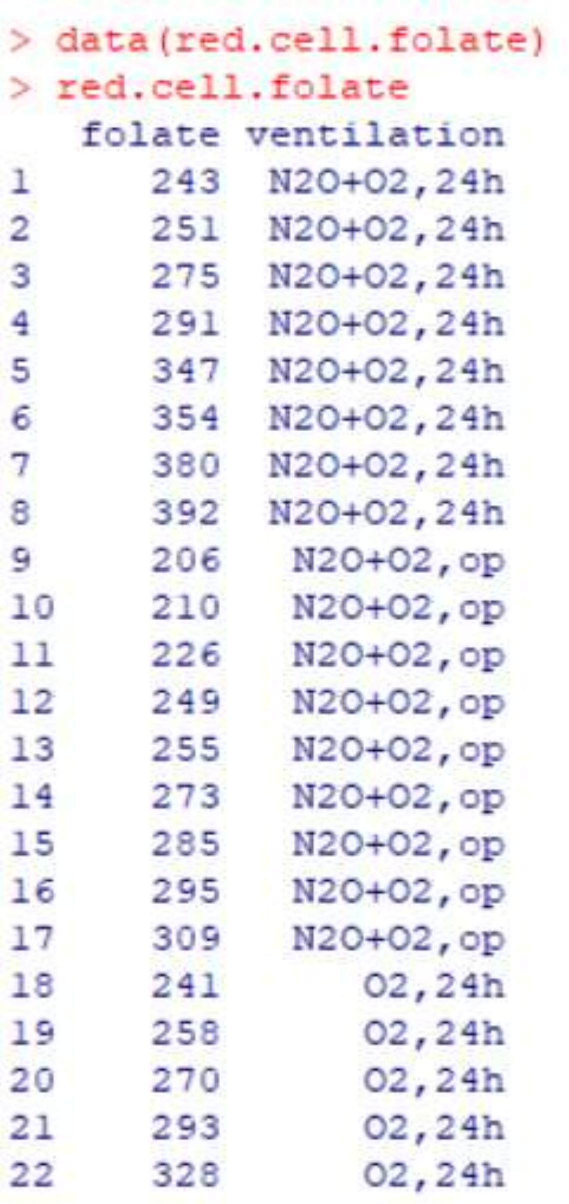
Pairwise Comparisons And Multiple Testing

If the F test shows that there is a difference between groups, the question quickly arises of wherein the difference lies. It becomes necessary to compare the individual groups.

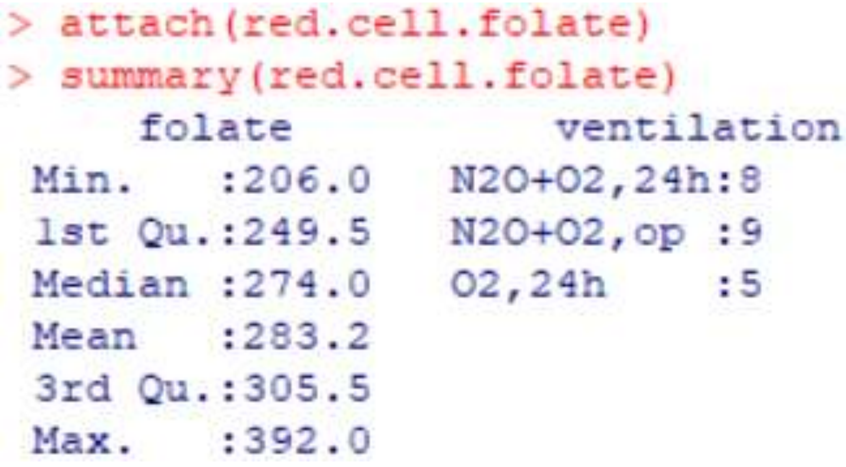
A function called ***pairwise.t.test*** computes all possible two-group comparisons. A common adjustment method is the **Bonferroni** correction, which is based on the fact that the probability of observing at least one of n events is less than the sum of the probabilities for each event.

Thus, by dividing the significance level by the number of tests or, equivalently, multiplying the p-values, we obtain a conservative test where the probability of a significant result is less than or equal to the formal significance level.

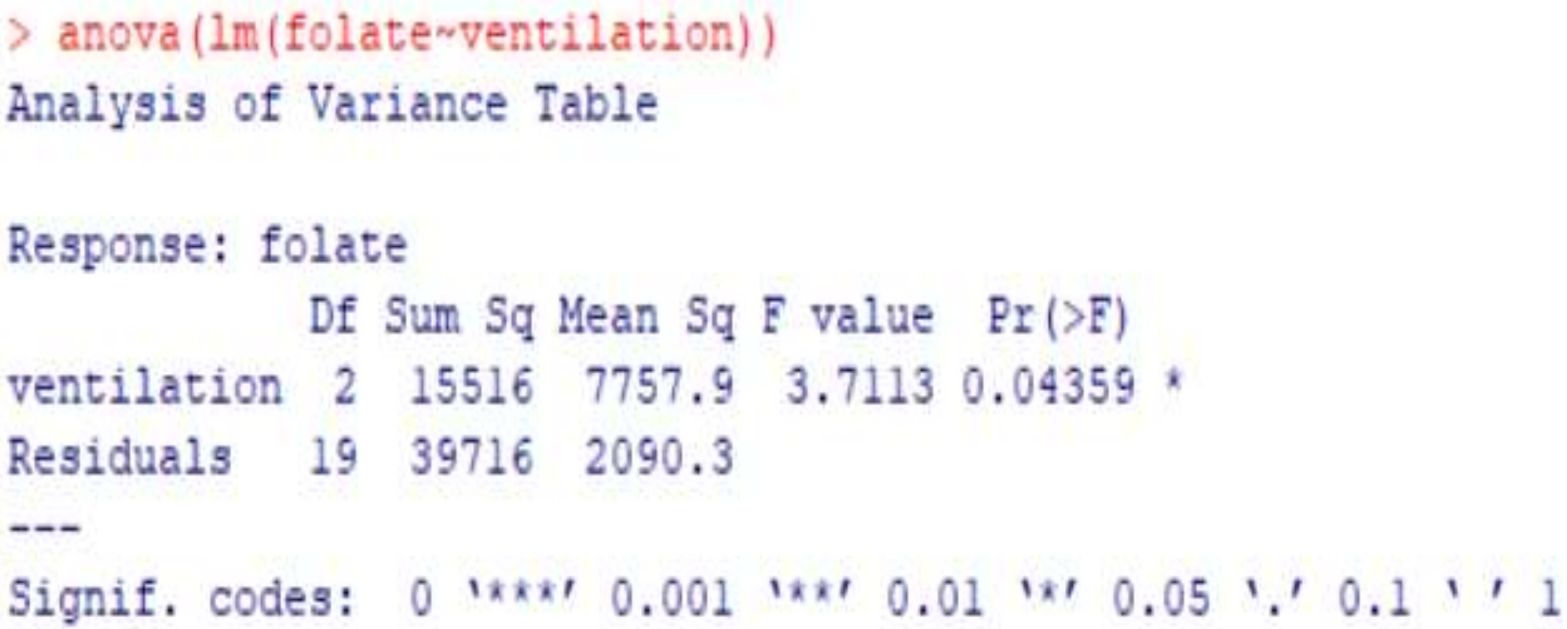
Example:



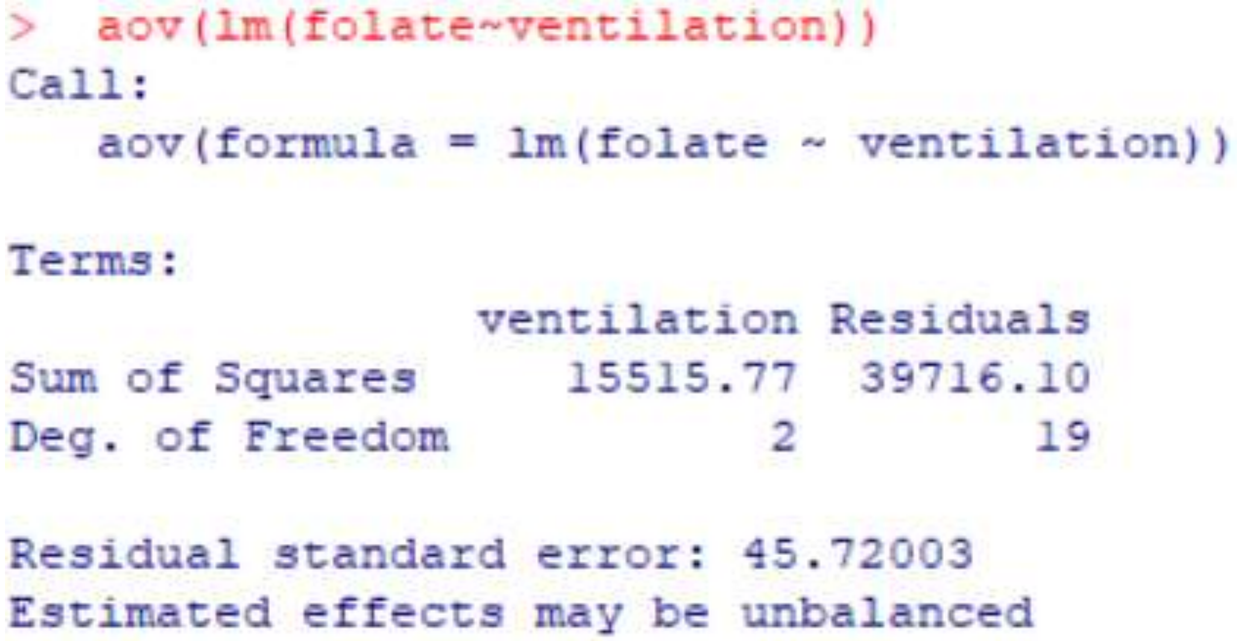
```
> data(red.cell.folate)
> red.cell.folate
   folate ventilation
1     243  N2O+O2,24h
2     251  N2O+O2,24h
3     275  N2O+O2,24h
4     291  N2O+O2,24h
5     347  N2O+O2,24h
6     354  N2O+O2,24h
7     380  N2O+O2,24h
8     392  N2O+O2,24h
9     206   N2O+O2,op
10    210   N2O+O2,op
11    226   N2O+O2,op
12    249   N2O+O2,op
13    255   N2O+O2,op
14    273   N2O+O2,op
15    285   N2O+O2,op
16    295   N2O+O2,op
17    309   N2O+O2,op
18    241      O2,24h
19    258      O2,24h
20    270      O2,24h
21    293      O2,24h
22    328      O2,24h
```



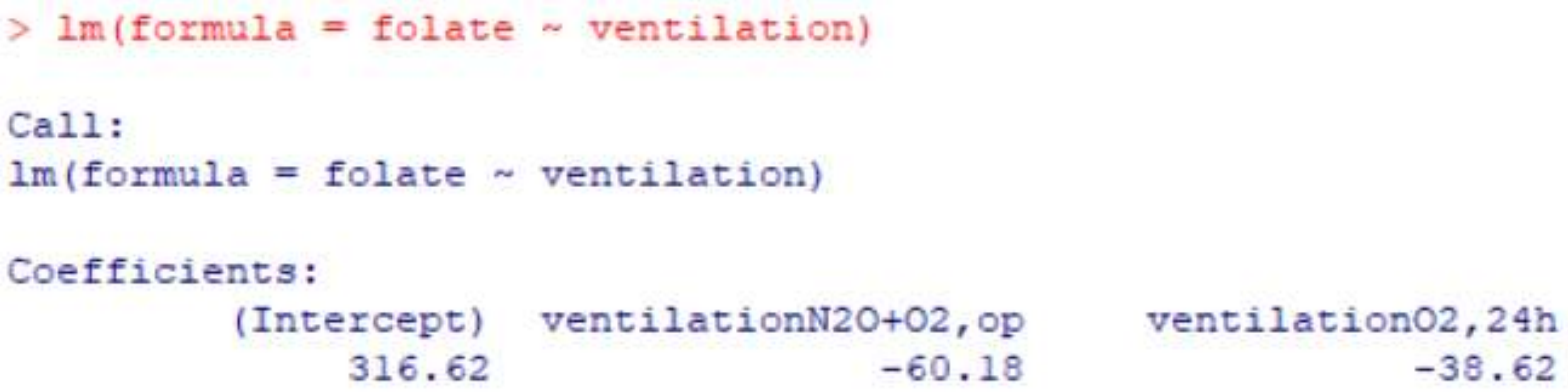
```
> attach(red.cell.folate)
> summary(red.cell.folate)
     folate            ventilation
 Min.   :206.0   N2O+O2,24h:8
 1st Qu.:249.5   N2O+O2,op :9
 Median :274.0   O2,24h    :5
 Mean   :283.2
 3rd Qu.:305.5
 Max.   :392.0
```



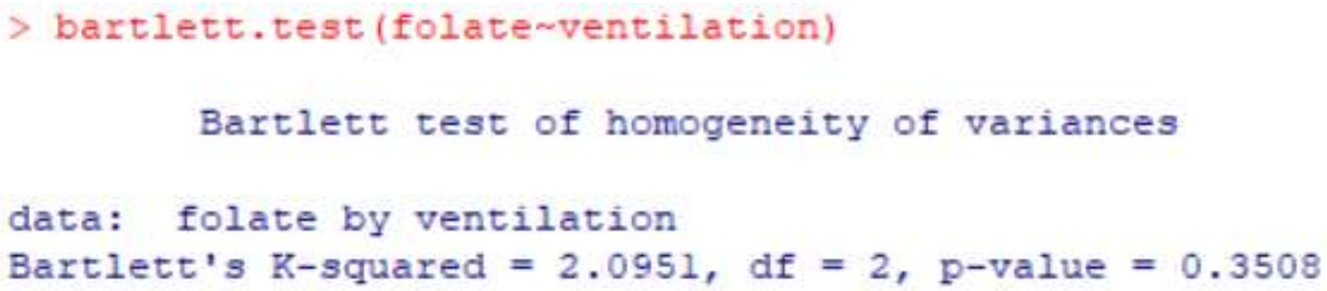
```
> anova(lm(folate~ventilation))
Analysis of Variance Table

Response: folate
            Df Sum Sq Mean Sq F value  Pr(>F)
ventilation  2  15516  7757.9  3.7113 0.04359 *
Residuals   19  39716  2090.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



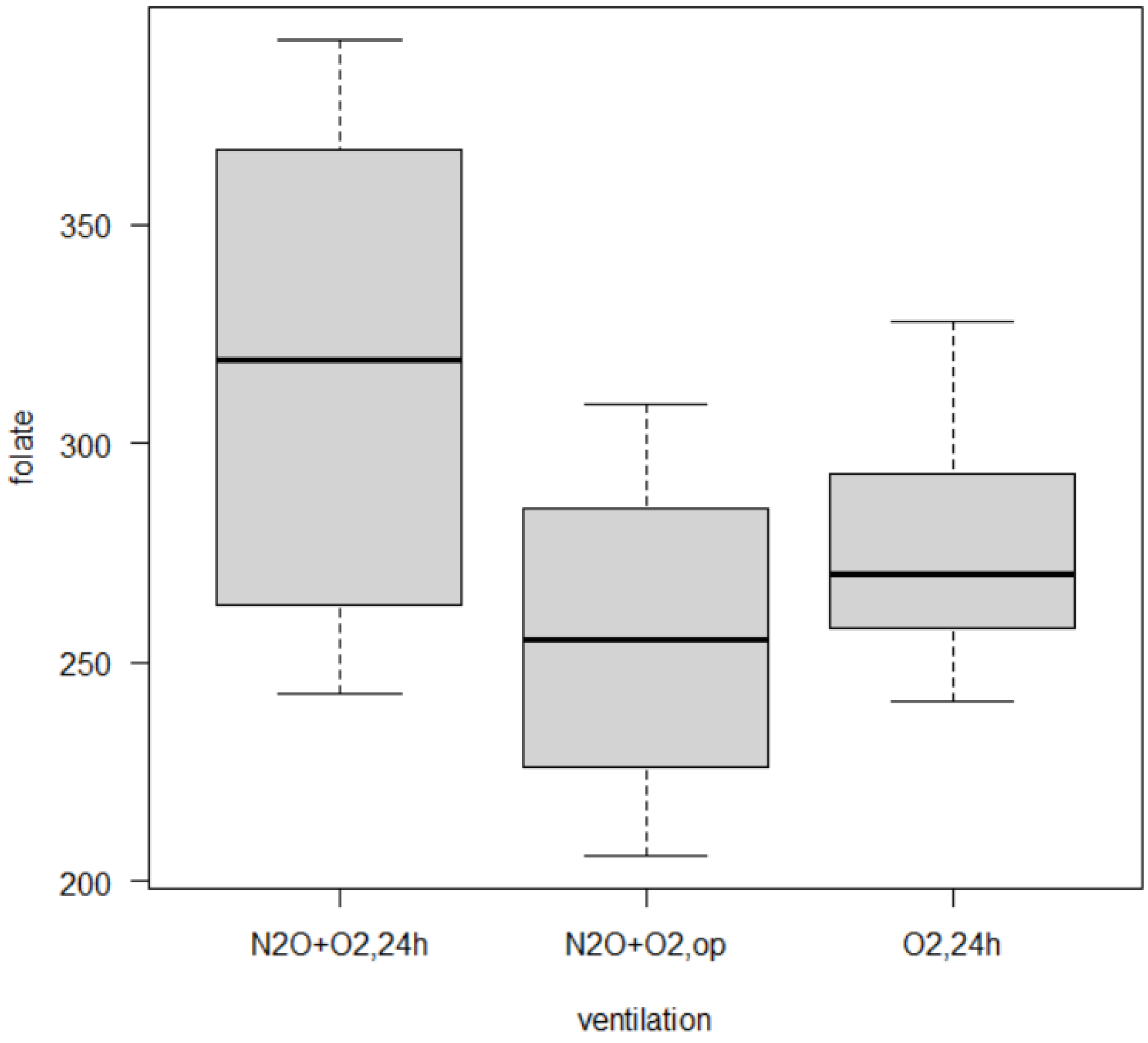
```
>  aov(lm(folate~ventilation))
Call:
   aov(formula = lm(folate ~ ventilation))

Terms:
                ventilation Residuals
Sum of Squares     15515.77  39716.10
Deg. of Freedom           2        19

Residual standard error: 45.72003
Estimated effects may be unbalanced
```



```
> lm(formula = folate ~ ventilation)

Call:
lm(formula = folate ~ ventilation)

Coefficients:
         (Intercept)  ventilationN2O+O2,op     ventilationO2,24h
              316.62                -60.18                -38.62
```



```
> bartlett.test(folate~ventilation)

        Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```



350
300
250
200
folate
N2O+O2,24h
N2O+O2,op
O2,24h
ventilation

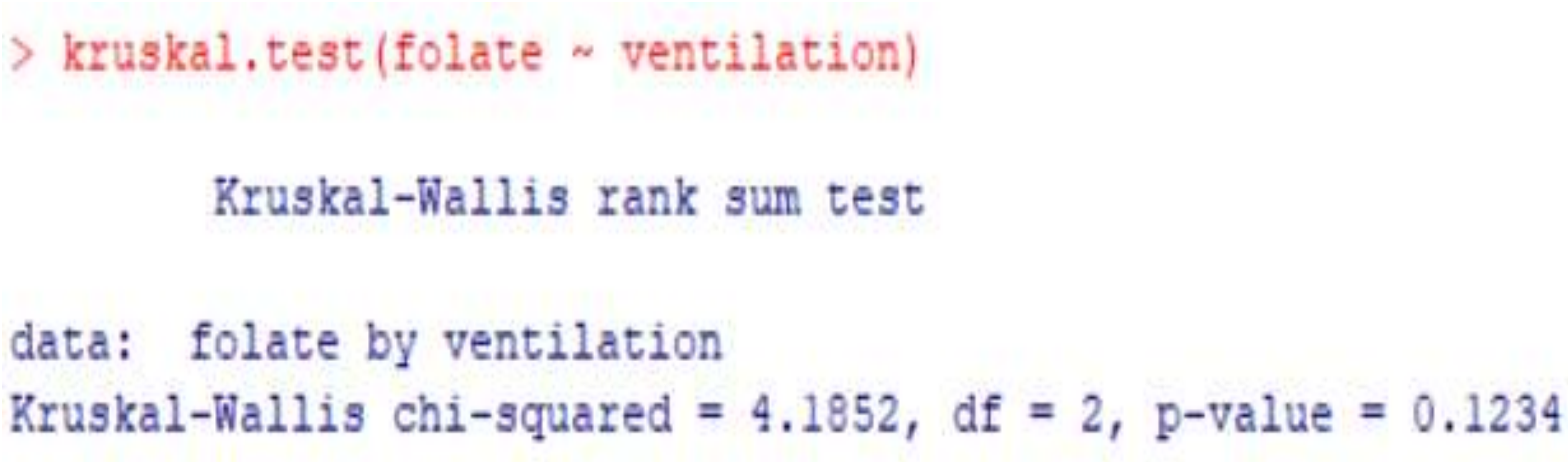
KRUSKAL WALLIS H TEST

The Kruskal Wallis test is the non parametric alternative to the One Way ANOVA. Non parametric means that the test doesn’t assume your data comes from a particular distribution. The H test is used when the assumptions for ANOVA aren’t met (like the assumption of normality).

It is sometimes called the *one-way ANOVA on ranks*, as the ranks of the data values are used in the test rather than the actual data points.

The Kruskal Wallis test will tell you if there is a significant difference between groups. However, it won’t tell you *which* groups are different. For that, you’ll need to run a Post Hoc test.

Example:



```
> kruskal.test(folate ~ ventilation)

        Kruskal-Wallis rank sum test

data:  folate by ventilation
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```