

# Significance Testing Using R

In the following handout words and symbols in **bold** indicate R functions and words and symbols in *italics* indicate entries supplied by the user; underlined words and symbols are optional entries. Shaded text represents examples from an R session. Examples use the objects in the file “StatExI.RData.”

## F-Test

In R the function **var.test** allows for the comparison of two variances using an F-test. Although it is possible to compare values of  $s^2$  for two samples, there is no capability within R for comparing the variance of a sample,  $s^2$ , to the variance of a population,  $\sigma^2$ . The syntax for the testing variances is

```
var.test(X, Y, ratio = 1, alternative = “two.sided”, conf.level = 0.95)
```

where  $X$  and  $Y$  are vectors containing the two samples. The optional command *ratio* is the null hypothesis; the default value is 1 if not specified. The command *alternative* gives the alternative hypothesis should the experimental F-ratio is found to be significantly different than that specified by *ratio*. The default for *alternative* is “two-sided” with the other possible choices being “less” or “greater.” The command *conf.level* gives the confidence level to be used in the test; the default value of 0.95 is equivalent to  $\alpha = 0.05$ . Here is a typical result using the objects *std.method* and *new.method*.

```
> std.method; new.method    # Data from problem 23 in Chapter 4
```

```
[1] 21.62  22.20  24.27  23.54  24.25  23.09  21.01  
[1] 21.54  20.51  22.31  21.30  24.62  25.72  21.54
```

```
> var(std.method); var(new.method)
```

```
[1] 1.638495  
[1] 3.690329
```

```
> var.test(std.method, new.method)    #  $H_0: (s_{\text{std}})^2 = (s_{\text{new}})^2$ ;  $H_A: (s_{\text{std}})^2 \neq (s_{\text{new}})^2$   
                                         #  $\alpha = 0.05$ 
```

F test to compare two variances

data: std.method and new.method

F = 0.444, num df = 6, denom df = 6, p-value = 0.3462

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.07629135 2.58395513

sample estimates:

ratio of variances

0.4439971

There are two ways to interpret the results provided by R. First, the p-value provides the smallest value of  $\alpha$  for which the F-ratio is significantly different from the hypothesized value. If this value is larger than the desired  $\alpha$ , then there is insufficient evidence to reject the null hypothesis; otherwise, the null hypothesis is rejected. Second, R provides the desired confidence interval for the F-ratio; if the calculated value falls within the confidence interval, then the null hypothesis is retained. For this example, the null hypothesis is retained and we find no evidence for a difference in the variances for the objects `std.method` and `new.method`. Note that R does not restrict the F-ratio to values greater than 1.

## t-Test

In R the function `t.test` provides for the comparison of an experimental mean,  $\bar{X}$ , to a theoretical mean,  $\mu$ , or between two experimental means. The function can handle both paired and unpaired data, as well as equal or unequal variances. The general syntax is

`t.test(X, Y, mu = 0, alternative = "two.sided", paired = FALSE, var.equal = FALSE, conf.level = 0.95)`

where  $X$  and  $Y$  are vectors containing the two samples. Note that vector  $Y$  is optional and is not used when comparing  $\bar{X}$  to  $\mu$ . The optional command `mu` gives the value of  $\mu$  when comparing  $\bar{X}$  to  $\mu$ , or the hypothesized difference between  $\bar{X}_1$  and  $\bar{X}_2$  when comparing two mean values; the default value is 0. The command `alternative` gives the alternative hypothesis should the difference between  $\bar{X}$  and  $\mu$ , or  $\bar{X}_1$  and  $\bar{X}_2$  be found significant. The default for `alternative` is "two-sided" with the other possible choices being "less" or "greater." The command `paired` indicates whether the data are paired or unpaired and defaults to FALSE or unpaired data if not specified. The command `var.equal` specifies whether the variances associated with  $\bar{X}_1$  and  $\bar{X}_2$  are equal or unequal, defaulting to unequal if not specified. Finally, the command `conf.level` gives the confidence level to be used in the t-test; the default value of 0.95 is equivalent to  $\alpha = 0.05$ . Shown below are several examples:

```
> abs    # Data from problem 18 in Chapter 4
```

```
[1] 0.639 0.638 0.640 0.639 0.640 0.639 0.638
```

```
> t.test(abs, mu=0.640, conf.level=0.99)      # H0:  $\bar{X} = 0.640$ ; HA  $\bar{X} \neq 0.640$   
                                              #  $\alpha = 0.01$ 
```

One Sample t-test

```
data: abs
```

```
t = -3.2404, df = 6, p-value = 0.01768      # this is texp
```

```
alternative hypothesis: true mean is not equal to 0.64
```

```
99 percent confidence interval:
0.6378559 0.6401441
sample estimates:
mean of x
0.639
```

There are two ways to interpret the results provided by R. First, the p-value provides the smallest value of  $\alpha$  for which the null hypothesis can be rejected. If this value is larger than the desired  $\alpha$ , then there is insufficient evidence to reject the null hypothesis and it must be retained. Second, R provides the desired confidence interval for  $\mu$ ; if the theoretical value falls within the confidence interval, then the null hypothesis is retained. For this example, the null hypothesis is retained and we find no evidence for a difference the mean absorbance of 0.639 for the data in the object *abs* and the theoretical value of 0.640 given by  $\mu$ .

```
> t.test(std.method, new.method, var.equal=TRUE) # Data from problem 23, Chapter 4
# variances taken as equal
# data are unpaired
#  $\bar{X}_1 = \bar{X}_2$ ;  $\bar{X}_1 \neq \bar{X}_2$ 
#  $\alpha = 0.05$ 

Two Sample t-test

data: std.method and new.method
t = 0.3995, df = 12, p-value = 0.6965
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.552448 2.249591
sample estimates:
mean of x mean of y
22.85429 22.50571
```

Here we see that there is no significant difference between the mean values for the data in *std.method* and that in *new.method*.

```
> air.water; sed.water # Data from problem 28 in Chapter 4

[1] 0.430 0.266 0.567 0.531 0.707 0.716
[1] 0.415 0.238 0.390 0.410 0.605 0.609

> t.test(air.water, sed.water, paired=TRUE) # data are paired
#  $H_0: d = 0$ ;  $H_A: d \neq 0$ 
#  $\alpha = 0.05$ 

Paired t-test

data: air.water and sed.water
t = 3.6998, df = 5, p-value = 0.01400
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.02797823 0.15535510
sample estimates:
```

mean of the differences  
0.09166667

This time we see that there is a significant difference between the mean values for the objects air.water and sed.water.

## Significance Tests When Data Are Not Normally Distributed

The significance tests described above apply only to data that are normally distributed. When data are not normally distributed, alternative statistical methods are needed that use the median in place of the mean. The equivalent to a t-test in this case is the Wilcoxon Signed Rank test (for comparing results to a theoretical median or for comparing paired data) and the Wilcoxon Rank Sum test for comparing two independent, unpaired sets of data. These test are considered robust because they make no underlying assumption about a distribution's shape.

### Wilcoxon Signed Rank Test

In the signed rank test, the individual values in a data set are compared to a hypothesized median by first subtracting the median from each data point. These differences are ordered, without attention to sign, from smallest-to-largest and then assigned a rank from 1 to  $n$ , where  $n$  is the number of values in the data set (for ties, each is assigned an average ranking). These rankings include the associated signs. The smaller of the sum of positive rankings or negative rankings is then compared to a critical value, whose value depends upon the desired confidence level and the number of values in the data set. Since the sum for the positive rankings should be similar to the sum of negative rankings, the critical value to which a comparison is made is the largest value for which a significant difference is possible. For example, suppose the following levels of Pb (in pg/mL) are reported for blood drawn from seven individuals

104, 79, 98, 150, 87, 136, 101

and the hypothesized median is 95. Subtracting the median from each value gives

9, -16, 3, 55, -8, 41, 6

Ordering these results from smallest-to-largest, without attention to signs gives

3, 6, -8, 9, -16, 41, 55

The ranked values, therefore, are

1, 2, -3, 4, -5, 6, 7

and the sum of positive rankings is 20 and the sum of negative rankings is 8. For seven measurements and  $\alpha = 0.05$ , the critical sum is 2. Since the actual sum is larger, there is no evidence that the hypothesized median is incorrect.

This test also can be used with paired data. In this case, the differences between the paired data are compared to an expected median, which typically is zero.

In R, the syntax for the Wilcoxon signed rank test is

```
wilcox.test(X, Y, mu = 0, alternative = "two.sided", paired = FALSE, conf.level = 0.95)
```

where the terms (and their default values) are the same as defined for the t-test. When comparing a single data set to a median, the object *Y* is omitted. For paired data, the optional object *Y* is included and *paired* is set to TRUE.

```
> blood
```

```
[1] 104 79 98 150 87 136 101
```

```
> wilcox.test(blood, mu=95)      # H0: median = 95; HA: median ≠ 95; α = 0.05
```

Wilcoxon signed rank test

data: blood

V = 20, p-value = 0.375

alternative hypothesis: true mu is not equal to 95

The interpretation here is the same as for the t-test. Because the p-value is larger than that specified by the confidence level ( $\alpha = 0.05$ ), there is no evidence that the median is not 95; the null hypothesis, therefore, is retained.

## Wilcoxon Rank Sum Test (aka – Mann-Whitney Test)

Comparing two sets of data uses a variation of the Wilcoxon signed rank test. In this case, the values for each set are merged together and ranked from smallest-to-largest and then to assign each value equal to its position in the ranking (with ties each assigned an average ranking). The sum of the rankings for each set are converted into test statistics

$$T_1 = S_1 - n_1(n_1 + 1)/2 \quad T_2 = S_2 - n_2(n_2 + 1)/2$$

where *S* is the sum of the rankings for a data set and *n* is the number of values in the data set. The smaller of  $T_1$  and  $T_2$  is then compared in a manner similar to that described for the Wilcoxon signed rank test. For example, suppose a standard method and a new method give the following results

standard method:	21.62	22.20	24.27	23.54	24.25	23.09	21.01
new method:	<u>21.54</u>	<u>20.51</u>	<u>22.31</u>	<u>21.30</u>	<u>24.62</u>	<u>25.72</u>	<u>21.54</u>

where we underline the new method's values as a means for keeping track of their position in the rankings. Ordering from smallest-to-largest gives

20.51 21.01 21.30 21.54 21.54 21.62 22.20

22.31 23.09 23.54 24.25 24.27 24.62 25.72

and the ranking becomes

1 2 3 4.5 4.5 6 7 8 9 10 11 12 13 14

This gives  $S_{\text{std}} = 61.5$  and  $S_{\text{new}} = 43.5$ , and  $T_1 = 33.5$  and  $T_2 = 15.5$ . The smaller of these two values, which is 15.5, is compared to a critical value of 8, which is the largest value for which we would reject the null hypothesis that the two methods are equivalent.

The syntax for this test in R is the same as for the Wilcoxon signed rank test; thus

```
> wilcox.test(std.method, new.method)
```

Wilcoxon rank sum test with continuity correction

data: std.method and new.method

W = 29, p-value = 0.6089

alternative hypothesis: true mu is not equal to 0

Warning message:

cannot compute exact p-value with ties in: wilcox.test.default(std.method,  
new.method)

Note that R cannot handle the presence of ties, thus the calculated p-value is not exact; nevertheless, our interpretation here is the same as that from a t-test.