Charles Valentine
Homework 6
10/16/2016

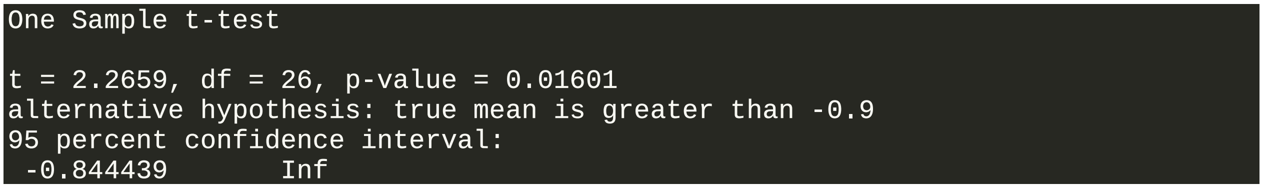
Note:
All code is in script hw06.R --- I have made use of cat and print statements to display information easily!

Problem 1

a) We will test the hypothesis (H_O) that the mean expression data for the *H4/j* gene in the ALL group is not greater than -0.9. We will use a one-sided t-test to test this hypothesis because we are working with a sampling distribution for the sample mean. See code for implementation.

   The output of this one-tailed t-test gives a p-value of 0.0160. This value is less than $\alpha = 0.05$ which indicates that we should reject H_O and accept that the mean expression data for the *H4/j* gene in the ALL group is greater than -0.9.

OUTPUT

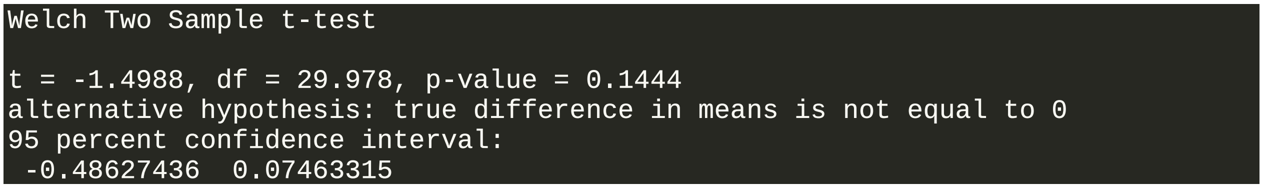


```
One Sample t-test

t = 2.2659, df = 26, p-value = 0.01601
alternative hypothesis: true mean is greater than -0.9
95 percent confidence interval:
 -0.844439       Inf
```

b) We will test the hypothesis (H_O) that the mean expression data for the *H4/j* gene in the ALL group is the same as the mean expression data for the *H4/j* gene in the AML data. A two-tailed t-test can be used to test this hypothesis. See code for implementation.

   The output of this two-tailed t-test gives a p-value of 0.1444. This value is greater than $\alpha = 0.05$ which indicates that we should accept H_O and reject that the mean expression data for the *H4/j* gene in the ALL group is different than the mean expression data for the *H4/j* gene in the AML data.

OUTPUT



```
Welch Two Sample t-test

t = -1.4988, df = 29.978, p-value = 0.1444
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.48627436  0.07463315
```

c) We will test the hypothesis (H_O) that the mean expression data for the *H4/j* gene in the ALL group is not less than the mean expression data for the *APS Prostate specific antigen* gene in the ALL group. A one-sided paired t-test can be used to test this hypothesis. See code for implementation.

The output of this one-tailed paired t-test gives a p-value of 0.0389. This value is less than $\alpha = 0.05$ which indicates that we should reject H_O and accept that the mean expression data for the *H4/j* gene in the ALL group is less than the mean expression data for the *APS Prostate specific antigen* gene in the ALL group.

OUTPUT

```
Welch Two Sample t-test

t = -1.8366, df = 26, p-value = 0.03886
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.02175309
```

d) We will test the hypothesis (H_O) that the proportion of patients for whom the *H4/j* gene expression values are lower than the *APS Prostate specific antigen* expression values in the ALL group p_{low} is not greater than 50%. A one-sided proportion test can be used to test this hypothesis. See code for implementation.

The output of this one-tailed proportion test gives a p-value of 0.0105. This value is less than $\alpha = 0.05$ which indicates that we should reject H_O and accept that the proportion of patients for whom the *H4/j* gene expression values are lower than the *APS Prostate specific antigen* expression values in the ALL group p_{low} is greater than 50%.

These results are expected as in 1c we illustrated that the two distributions were statistically significant in the difference of their sample means. Specifically, the difference described *H4/j* gene expression values as being less than *APS Prostate specific antigen* gene expression values. Therefore, more than half of the *H4/j* gene expression values should be less than that of the mean of the *APS Prostate specific antigen* gene expression values in the ALL group.

OUTPUT

```
1-sample proportions test with continuity correction

null probability 0.5
X-squared = 5.3333, df = 1, p-value = 0.01046
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5653018 1.0000000
```

e) We will test the hypothesis (H_O) that the proportion of patients for whom the *H4/j* gene expression values are lower than -0.6 in the ALL group p_{H4j} is not greater than 50%. A one-sided proportion test can be used to test this hypothesis. See code for implementation.

The output of this one-tailed proportion test gives a p-value of 0.1241. This value is greater than $\alpha = 0.05$ which indicates that we should accept H_O and reject that the proportion of patients for whom the *H4/j* gene expression values are lower than -0.6 in the ALL group p_{H4j} is not greater than 50%

OUTPUT

```
1-sample proportions test with continuity correction

null probability 0.5
X-squared = 1.3333, df = 1, p-value = 0.1241
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.5464797
```

f) We will test the hypothesis (H_O) that the proportion of p_{H4j} in the ALL group is the same as the proportion of p_{H4j} in the AML group. A two-sided two-proportions test can be used to test this hypothesis. See code for implementation.

The output of this two-tailed two-proportions test gives a p-value of 0.1010. This value is greater than $\alpha = 0.05$ which indicates that we should accept H_O and reject that the proportion of p_{H4j} in the ALL group differs from the proportion of p_{H4j} in the AML group.

OUTPUT

```
X-squared = 2.6901, df = 1, p-value = 0.101
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.74094690  0.02714219
```

Problem 2

a) With a rejection probability of 0.05 and 2000 trials we expect to see 100 rejections.
b) To calculate the probability of less than 90 rejections $P(X < 90)$ we can use the cumulative density function for the binomial in R. See code for implementation.

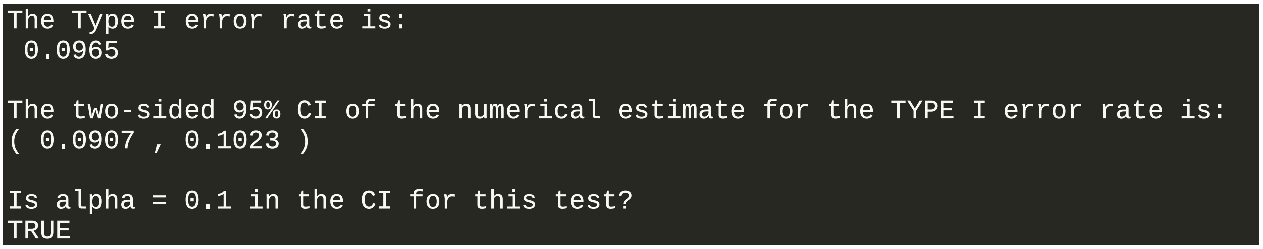
OUTPUT

```
Probability of less than 90 rejections P(X<90):
[1] 0.1649724
```

Problem 3

a) See code for implementation. This test is valid because the alpha level of 0.1 is within the 95% confidence interval for the numerically estimated Type I error rate.
b) This test appears to be valid given the parameters in the exercise. 10,000 estimations of the Type I error rate converge on the alpha level of 0.1 to within the 95% CI. However, this is useless for any simulated data not from a normal distribution with a mean of 3 and a standard deviation of 4, rendering this test only good for this one task.

OUTPUT

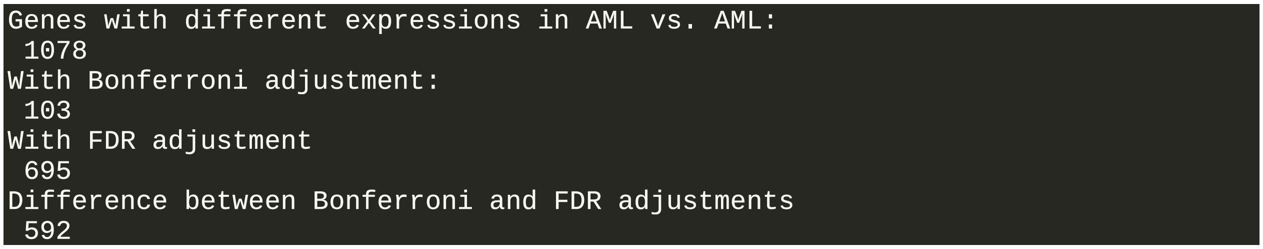


```
The Type I error rate is:
 0.0965

The two-sided 95% CI of the numerical estimate for the TYPE I error rate is:
( 0.0907 , 0.1023 )

Is alpha = 0.1 in the CI for this test?
TRUE
```

Problem 4

a) See code for implementation. There are 103 genes with different expressions between the AML and ALL groups using Bonferroni multiple hypothesis testing correction. When using the FDR adjustment for multiple hypothesis testing, there are 695 genes with different expressions between the AML and ALL groups. The difference between these two is 592 genes.

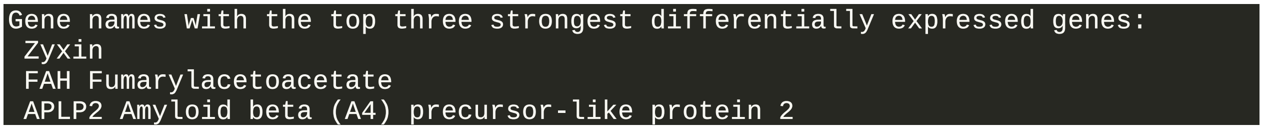
OUTPUT



```
Genes with different expressions in AML vs. AML:
 1078
With Bonferroni adjustment:
 103
With FDR adjustment
 695
Difference between Bonferroni and FDR adjustments
 592
```

b) See code for implementation.

OUTPUT



```
Gene names with the top three strongest differentially expressed genes:
 Zyxin
 FAH Fumarylacetoacetate
 APLP2 Amyloid beta (A4) precursor-like protein 2
```

Problem 5

a) See code for implementation.

b) Given a modest sample size (40) and probability (0.2) all three confidence interval estimations for the binomial proportion give similar results. It is apparent that Wald CI is, potentially, slightly worse for this combination of n and p. The following coverage proportions are calculated:

OUTPUT

```
Coverage for 95% Wald CI:
  0.9999
Coverage for 95% Agresti-Coull CI:
  1
Coverage for 95% Wilson CI:
  1
```