Charles Valentine
Midterm
10/30/2016

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

Problem 1a

The expected value and standard deviation for a random variable X that follows a distribution with pdf $f_X(x) = 2.469862\left(xe^{-x^2}\right), x = 1, 2, 3$; can be calculated as follows:

$$\mathrm{E}(X) = \sum_x xf(x)$$

$$\mathrm{E}(X) = 1 \cdot 0.9086 + 2 \cdot 0.0905 + 3 \cdot 0.0009 = 1.0923$$

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)} = \sqrt{\sum_x (x - \mu_x)^2 f(x)}$$

$$\mathrm{sd}(X) = \sqrt{0.0085 \cdot 0.9086 + 0.8239 \cdot 0.0905 + \ 3.6393 \cdot 0.0009} = 0.2926$$

The expected value and standard deviation for a random variable Y that follows a distribution with pdf $f_Y(y) = 2ye^{-y^2}, y > 0$; can be calculated as follows. We will use the integrate function in R to finish calculations:

$$\mathrm{E}(Y) = \int_0^\infty yf(y)dy = 0.8862$$

$$\mathrm{sd}(Y) = \sqrt{\mathrm{var(Y)}} = \sqrt{\int_0^\infty \left(y - \mu_y\right)^2 f(y)} = 0.4633$$

OUTPUT

```
Problem 1a
==========
E(X)  = 1.0923
sd(X) = 0.2926

E(Y)  = 0.8862
sd(Y) = 0.4633
```

Problem 1b

If X and Y are independent, then we can apply a formula for linear combinations of variables:

$$\mathrm{E}(2X - 3) = 2 \cdot 1.0923 - 3 \cdot 0.8862 = -0.4741$$

$$\mathrm{sd}(2X - 3Y) = 2 \cdot 0.2926 - 3 \cdot 0.4633 = -0.8046$$

OUTPUT

```
Problem 1b
==========
E(2X - 3Y)  = -0.4741
sd(2X - 3Y) = -0.8046
```

Problem 2

If X, following the standard normal distribution $\mathrm{N}(\mu = 0, \sigma = 1)$, and Y, following the Chi-square distribution $\mathrm{chisq}(m = 4)$, are independent, then we can estimate $\mathrm{E}({}^{X^2}/_{X^2 + Y})$ from the linear combination of $\mathrm{E}(X)$ and $\mathrm{E}(Y)$. The expected value or mean of a normal distribution is always μ and for a Chi-square distribution, m.

$$\mathrm{E}\left(\frac{X^2}{X^2 + Y}\right) = \left(\frac{\mu^2}{\mu^2 + m}\right) = \left(\frac{0^2}{0^2 + 4}\right) = 0.00$$

The expected value can be estimated using random draws from both distributions in R. After 2000 draws from both distributions we converge to $\mathrm{E}\left({}^{X^2}/_{X^2 + Y}\right) \approx 0.00$.

OUTPUT

```
Problem 2
=========
E(X^2/(X^2 + Y)) = 0.00
```

Problem 3

For $nsim = 1000$ datasets we calculated the true coverage of a 95% confidence level formula to be 92%. We can calculate the probability of our empirical coverage being greater than 94% by using the binomial distribution. The cumulative distribution $1 - \mathrm{pbinom}(q = 940, n = 1000, p = 0.92)$ will give us the correct probability. Finding the cumulative probability at point 940 gives us the probability that our empirical coverage is less than or equal to that of 94% of 1000, and one minus this probability is the probability that our empirical coverage is exclusively greater than 94%. This gives us a probability of 0.6617%.

OUTPUT

```
Problem 3
=========
P(X>940) = 0.6617 %
```

Problem 4

Minimizing the negative log likelihood of the distribution $\mathrm{N}(\mu = \theta, \sigma = \theta)$ using the `optimize` routine in R will solve for a value of θ.

OUTPUT

```
Problem 4
=========
MLE point estimator for theta: 2.4265
```

Problem 5

a. There are 502 genes in the Golub et al. (1999) data set with a mean expression value greater than 0.6. A FDR correction was applied for multi-hypothesis testing and an alpha level of 0.1 was chosen.

OUTPUT

```
Problem 5a
==========
Genes with mean expression values greater than 0.6:
 502
```

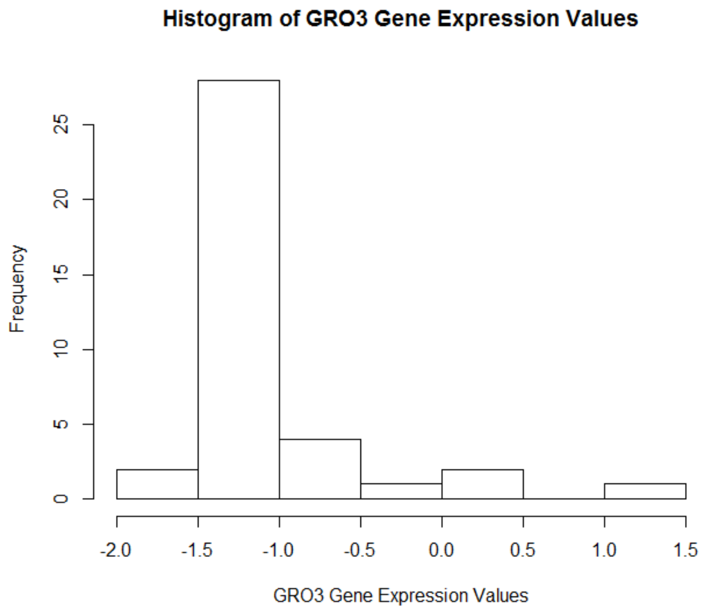
b. The genes were then sorted by significance and the top five were selected in descending order.

OUTPUT

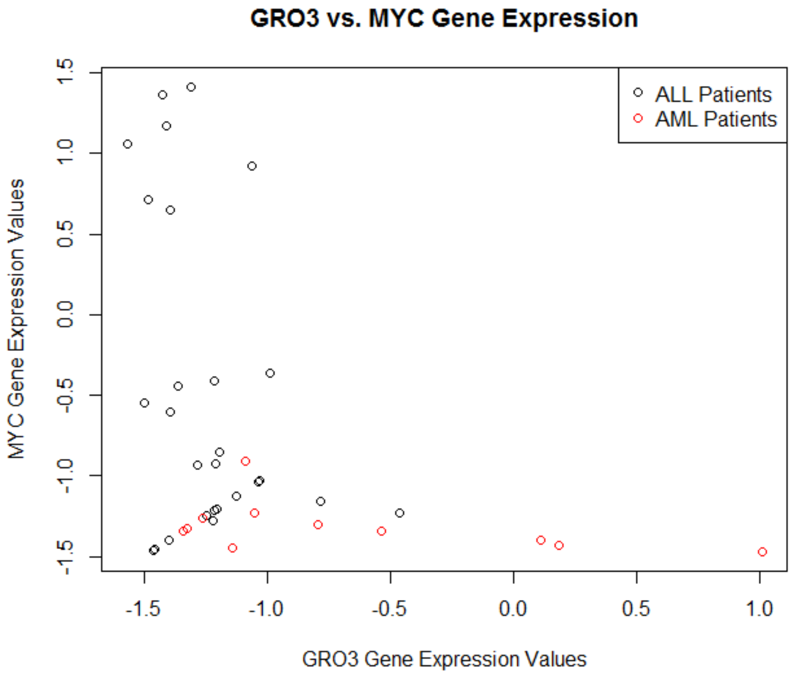
```
Problem 5b
==========
Top five genes with mean expression values > 0.6:
1.   HnRNP-E2 mRNA
2.   Ornithine decarboxylase antizyme, ORF 1 and ORF 2
3.   GB DEF = Polyadenylate binding protein II
4.   RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene
5.   GAPD Glyceraldehyde-3-phosphate dehydrogenase
```

Problem 6

a. Histogram for the gene expression values of the GRO3 GRO3 Oncogene.

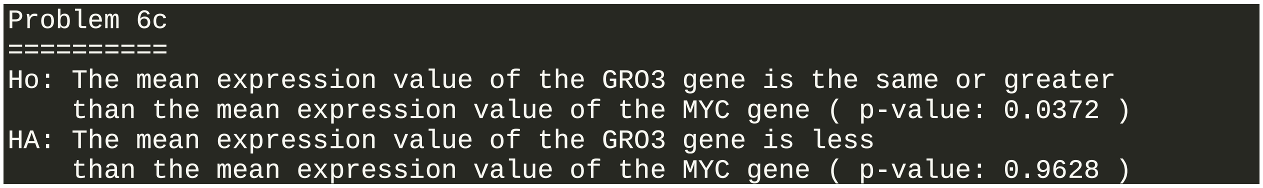


b. Scatterplot of the GRO3 GRO3 Oncogene gene expression values vs. the MYC V-myc avian myelocytomatosis viral oncogene homolog in ALL and AML patients.



c. To test the alternative hypothesis that the mean expression value of the GRO3 gene is less than the mean expression value of the MYC gene we will use a parametric paired T test. The p-value for the null hypothesis is 0.0372 which is less than 0.05 and we may reject the hypothesis. The mean expression value of the GRO3 gene is significantly less than the mean expression value of the MYC gene.

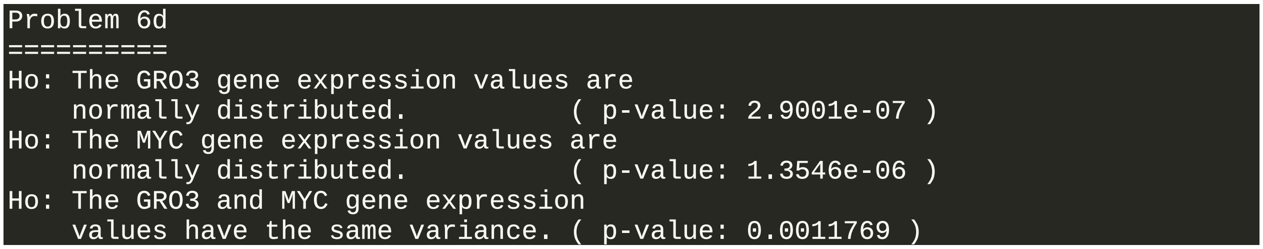
OUPUT



```
Problem 6c
==========
Ho: The mean expression value of the GRO3 gene is the same or greater
    than the mean expression value of the MYC gene ( p-value: 0.0372 )
HA: The mean expression value of the GRO3 gene is less
    than the mean expression value of the MYC gene ( p-value: 0.9628 )
```

d. To test whether the parametric T-test was appropriate to use for the previous hypothesis we will check a few of the assumptions of the data that are needed before we should have proceeded. The normality tests were conducted with the Shapiro-Wilkes test and the variance test were conducted with a F test to compare two variances. All p-values are less than 0.05 so we may reject that both gene expression sets are normally distributed and we may also reject that they have equal variances. The parametric T-test was **not** the appropriate test choice because of this.

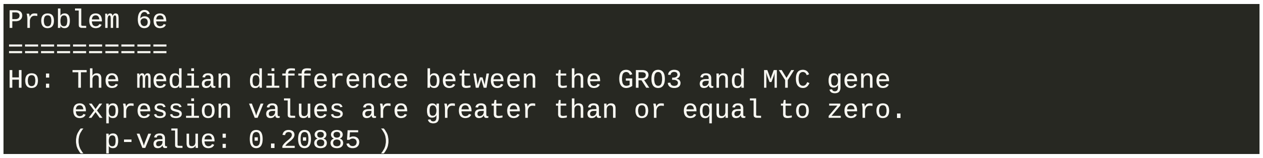
OUTPUT



```
Problem 6d
==========
Ho: The GRO3 gene expression values are
    normally distributed.          ( p-value: 2.9001e-07 )
Ho: The MYC gene expression values are
    normally distributed.          ( p-value: 1.3546e-06 )
Ho: The GRO3 and MYC gene expression
    values have the same variance. ( p-value: 0.0011769 )
```

e. To find a one-sided median difference between the GRO3 and MYC gene expression values we can use a rank text, specifically the Wilcoxon rank sum test. We will test the alternate hypothesis that the median difference between the GRO3 and MYC gene expression values is greater than 0. With a p-value of 0.20885 for the null hypothesis we cannot accept the alternative hypothesis.

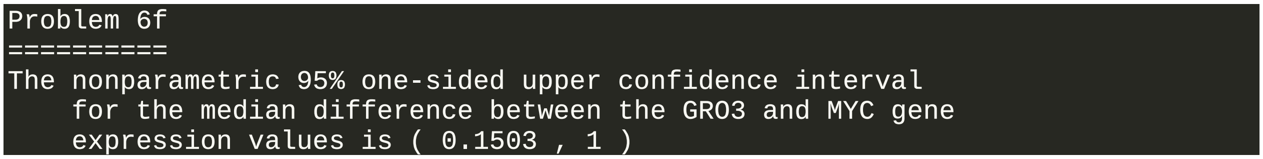
OUTPUT



```
Problem 6e
==========
Ho: The median difference between the GRO3 and MYC gene
    expression values are greater than or equal to zero.
    ( p-value: 0.20885 )
```

f. Two quickly solve the 95% nonparametric one sided upper CI for the median difference between the expression values of GRO3 gene and of MYC gene we take the 90% confidence interval for the two-tailed binomial test and use the lower bound as the lower bound for our 95% interval.

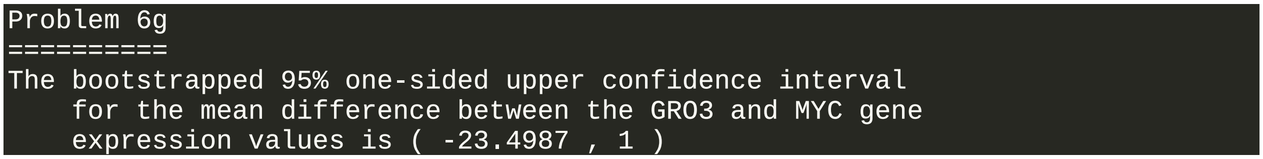
OUTPUT



```
Problem 6f
==========
The nonparametric 95% one-sided upper confidence interval
    for the median difference between the GRO3 and MYC gene
    expression values is ( 0.1503 , 1 )
```

g. To bootstrap a nonparametric 95% one-sided upper confidence interval for the mean difference between expression values of GRO3 gene and MYC gene we simulate 2000 random samples, with replacement, from both sample distributions and calculate a pseudo-statistic.

OUTPUT

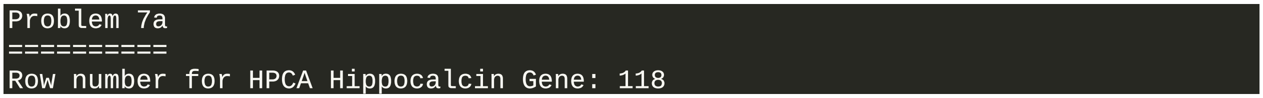


```
Problem 6g
==========
The bootstrapped 95% one-sided upper confidence interval
    for the mean difference between the GRO3 and MYC gene
    expression values is ( -23.4987 , 1 )
```

Problem 7

a. The row number of the “HPCA Hippocalcin” gene is: 118

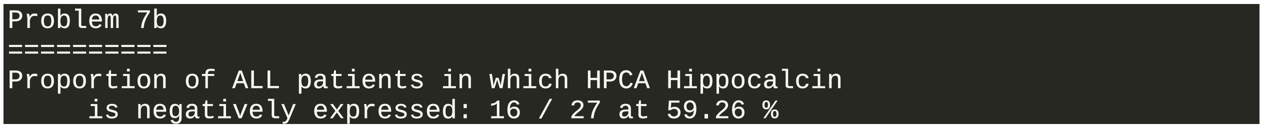
OUTPUT



```
Problem 7a
==========
Row number for HPCA Hippocalcin Gene: 118
```

b. The proportion of ALL patients that have the “HPCA Hippocalcin” gene negatively expressed is 16/27 at 59.26%.

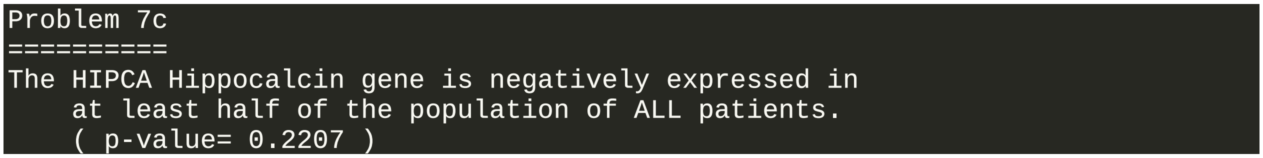
OUTPUT



```
Problem 7b
==========
Proportion of ALL patients in which HPCA Hippocalcin
     is negatively expressed: 16 / 27 at 59.26 %
```

c. The null hypothesis Ho is that the “HPCA Hippocalcin” gene is negatively expressed in at least half of the population of ALL patients. The alternative of this hypothesis is that the “HPCA Hippocalcin” gene is negatively expressed in less than half of the population of ALL patients. A 1-sample proportions test with continuity correction was applied and a p-value of 0.2207 was calculated. We should accept the null hypothesis that the “HIPCA Hippocalcin” gene is negatively expressed in at least half of the population of ALL patients.

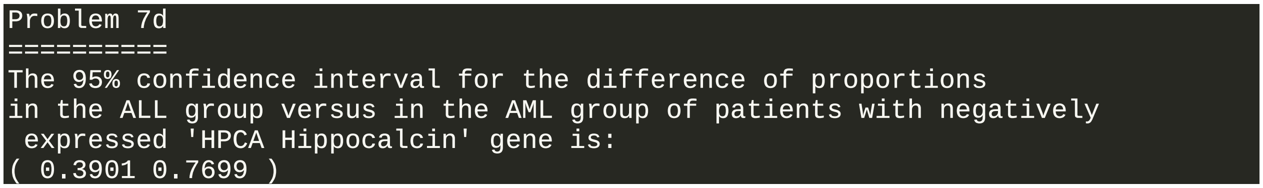
OUTPUT



```
Problem 7c
==========
The HIPCA Hippocalcin gene is negatively expressed in
    at least half of the population of ALL patients.
    ( p-value= 0.2207 )
```

d. See output

OUTPUT



```
Problem 7d
==========
The 95% confidence interval for the difference of proportions
in the ALL group versus in the AML group of patients with negatively
 expressed 'HPCA Hippocalcin' gene is:
( 0.3901 0.7699 )
```