

100

STAT 8003, Homework 8

Group # 8

Members: Nooreen Dabbish, Yinghui Lu, Anastasia Vishnyakova

November 7, 2013

Problem 1. We want to know the mean percentage of butterfat in milk produced by a farm by sampling multiple loads of milk. Previous records indicate the average percent butterfat in milk is 3.35 and the standard deviation among loads is 0.15. Now we hope to detect a change of the percent butterfat in milk.

a). Find the rejection region at the significant level $\alpha = 0.05$.

Let Y_1, Y_2, \dots, Y_n represent the sample of mean percentage of butterfat in milk, $\mu_0 = 3.35$, $\hat{\sigma} = 0.15$. The test here is, $H_0: \mu = \mu_0$ versus $H_1: \mu = \mu_1$

Then we have,

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{D} N(0, 1)$$

Under H_0 , we have

$$Z_0 = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{D} N(0, 1)$$

Then,

$$\begin{aligned}\alpha &= P\left(|Z_0| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \\ &= P\left(\left|\frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma}\right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)\end{aligned}$$

So the rejection region is

$$\bar{Y} > \mu_0 + \frac{\Phi^{-1}(1 - \frac{\alpha}{2})\sigma}{\sqrt{n}} \cup \bar{Y} < \mu_0 - \frac{\Phi^{-1}(1 - \frac{\alpha}{2})\sigma}{\sqrt{n}}$$

Plugging in data, we have the rejection region is

$$\bar{Y} > 3.35 + \frac{0.294}{\sqrt{n}} \cup \bar{Y} < 3.35 - \frac{0.294}{\sqrt{n}}$$

b). Suppose 100 loads of milk are sampled. What is the power for the test for detecting a change of the mean to 3.40.

In this context, we have $n = 100$, $\mu_0 = 3.35$, $\mu_1 = 3.40$ and $\sigma = 0.15$. Then

$$\begin{aligned}1 - \beta &= P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right) \\ &\quad + P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} < \Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right) \\ &= P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} > 1.96 + \frac{-0.05 \times 10}{0.15}\right) + P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} < -1.96 + \frac{-0.05 \times 10}{0.15}\right) \\ &= \Phi(-5.293) + 1 - \Phi(-1.373) \\ &= 0.915\end{aligned}$$

Hence, the power for the test for detecting a change of the mean to 3.40 is 0.915.

c). Plot the power as a function of the absolute value of the change of the mean over the standard deviation (which is $|\mu_1 - \mu_0|/\sigma$).

Write $d = \frac{|\mu_1 - \mu_0|}{\sigma}$, then

When $\mu_0 > \mu_1$,

$$\begin{aligned}
1 - \beta &= P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \sqrt{nd}\right) + P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} < \Phi^{-1}\left(\frac{\alpha}{2}\right) + \sqrt{nd}\right) \\
&= \Phi(-\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - \sqrt{nd}) + \Phi(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \sqrt{nd}) \\
&= \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \sqrt{nd}\right) + \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \sqrt{nd}\right)
\end{aligned}$$

When $\mu_0 < \mu_1$

$$\begin{aligned}
1 - \beta &= P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - \sqrt{nd}\right) + P\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} < \Phi^{-1}\left(\frac{\alpha}{2}\right) - \sqrt{nd}\right) \\
&= \Phi(-\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \sqrt{nd}) + \Phi(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \sqrt{nd}) \\
&= \Phi(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \sqrt{nd}) + \Phi(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \sqrt{nd})
\end{aligned}$$

Combining the above two, we have

$$1 - \beta = \Phi(\Phi^{-1}(\frac{\alpha}{2}) + \sqrt{nd}) + \Phi(\Phi^{-1}(\frac{\alpha}{2}) - \sqrt{nd})$$

Since we are not given the value of n , we choose three cases with $n = 20$, $n = 50$ and $n = 100$ to demonstrate the power. Plot it in R, and the result is shown as Figure 1.

R code

```

> n1=20
> n2=50
> n3=100
> d=seq(0,1,by=0.01)
> power1=pnorm(-1.96-sqrt(n1)*d)+pnorm(-1.96+sqrt(n1)*d)
> power2=pnorm(-1.96-sqrt(n2)*d)+pnorm(-1.96+sqrt(n2)*d)
> power3=pnorm(-1.96-sqrt(n3)*d)+pnorm(-1.96+sqrt(n3)*d)
> plot(d,power1,type='l',lwd=1,col='blue',xlab='d',ylab='Power')
> lines(d,power2,type='l',lwd=1,col='red')
> lines(d,power3,type='l',lwd=1,col='black')
> legend(0.6,0.4,c("n=20","n=50","n=100"),col=c("blue","red","black"),lty=c(2,2,2),
bty="n")

```

From Figure 1 we can see that with the sample size increasing, the power increases faster, which means given the same d ($|\mu_1 - \mu_0|/\sigma$), large sample size could give a more accurate result.

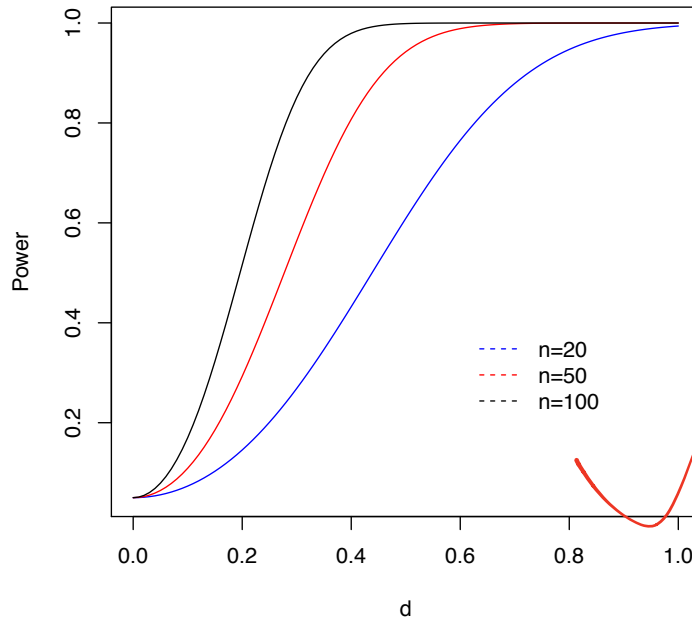


Figure 1: Power as a function of d

d). Now we hope to detect a change of the percent butterfat in milk to 3.40 with a power 0.8. How many loads do we need to sample?

With the information given, we have

$$d = \frac{|\mu_1 - \mu_0|}{\sigma} = \frac{|3.40 - 3.35|}{0.15} = \frac{1}{3}$$

And from part c), we know that

$$1 - \beta = \Phi(\Phi^{-1}(\frac{\alpha}{2}) + \sqrt{nd}) + \Phi(\Phi^{-1}(\frac{\alpha}{2}) - \sqrt{nd})$$

Plugging in data, then we have

$$\begin{aligned}
1 - \beta &= \Phi(-1.96 + \frac{1}{3}\sqrt{n}) + \Phi(-1.96 - \frac{1}{3}\sqrt{n}) \\
&\approx \Phi(-1.96 + \frac{1}{3}\sqrt{n}) \\
&= 80\%
\end{aligned}$$

Solving this equation, we have

$$n \approx 70.64$$

Hence, we need to sample at least 71 loads.

Problem 2. The Sydney-Hobart yacht race starts from Sydney Harbour on Boxing day (December 26) and finishes several days later in Hobart. It is a 630 nautical mile ocean race. The data give the winning times from 1945 to 1993, as they appeared in the Sydney Morning Herald on 24 December, 1994, plus the winning times for 1994 to 1997. The data file is a tab-delimited text, named as "yacht.txt", located in the "Data" folder on blackboard.

Variable	Description
Yacht	Name of winning yacht
Year	Year
Days	Days unit of winning time
Hours	Hours unit of winning time
Minutes	Minutes unit of winning time
Time	Winning time in minutes (should match time in Days, Hours and Minutes)

a). Plot histogram of Time and $\log(\text{Time} - 3100)$. Which one do you think are more likely to follow a normal distribution?

From the graph we can see that the distribution of $\log(\text{time}-3100)$ shows a more bell-like shape, while the original distribution of Time is skewed to the left. We then think that distribution of Time after transformation is more likely to follow a normal distribution.

Code from R:

```

> data <- read.table("yacht.txt", sep="\t", header=T)
> hist(data$Time, breaks=6, col= "darkorange", xlab="Time", ylab='Density',
main="Histogram of Winning Time in Minutes", prob=TRUE)
> curve(dnorm(x, mean=mean(data$Time), sd=sd(data$Time)), col='blue', lty=2,
lwd=2, add=TRUE)

```

```

> legend(8000,0.00025,c("Normal fit"),col='blue',lty=2,lwd=2,bty='n')
> hist(log(data$Time-3100),breaks=6, col= "darkorange", xlab="log(Time-3100)",
ylab='Density', ylim=c(0,0.8),main="Histogram of log(Time-3100)", prob=TRUE)
> curve(dnorm(x,mean=mean(log(data$Time-3100)),sd=sd(log(data$Time-3100))),
col='blue',lty=2,lwd=2,add=TRUE)
> legend(8,0.7,c("Normal fit"),col='blue',lty=2,lwd=2,bty='n')

```

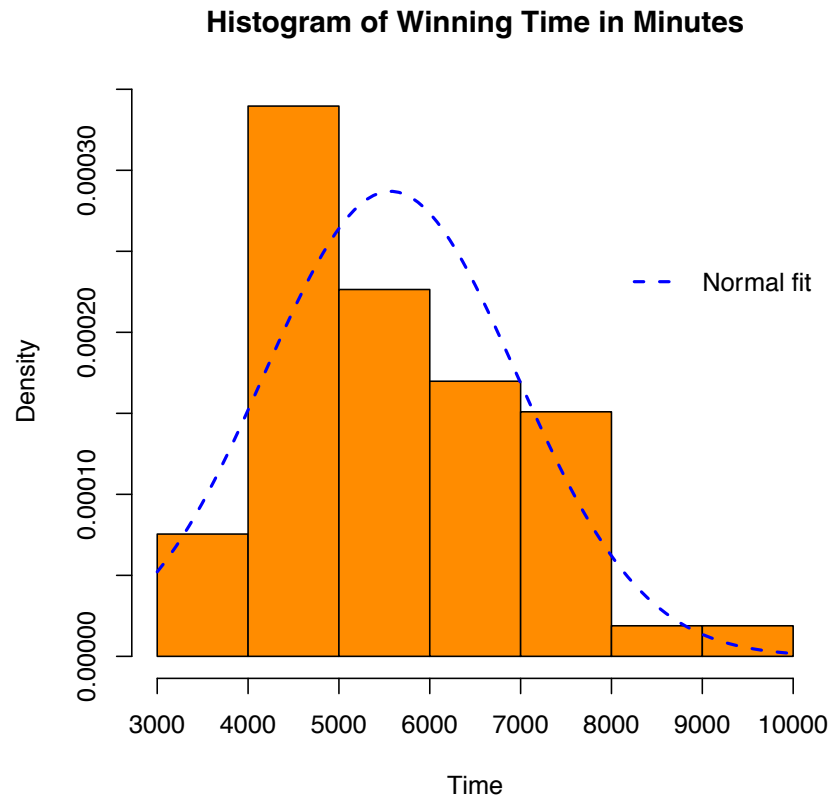


Figure 2: Comparison of distributions of Time before and after transformation

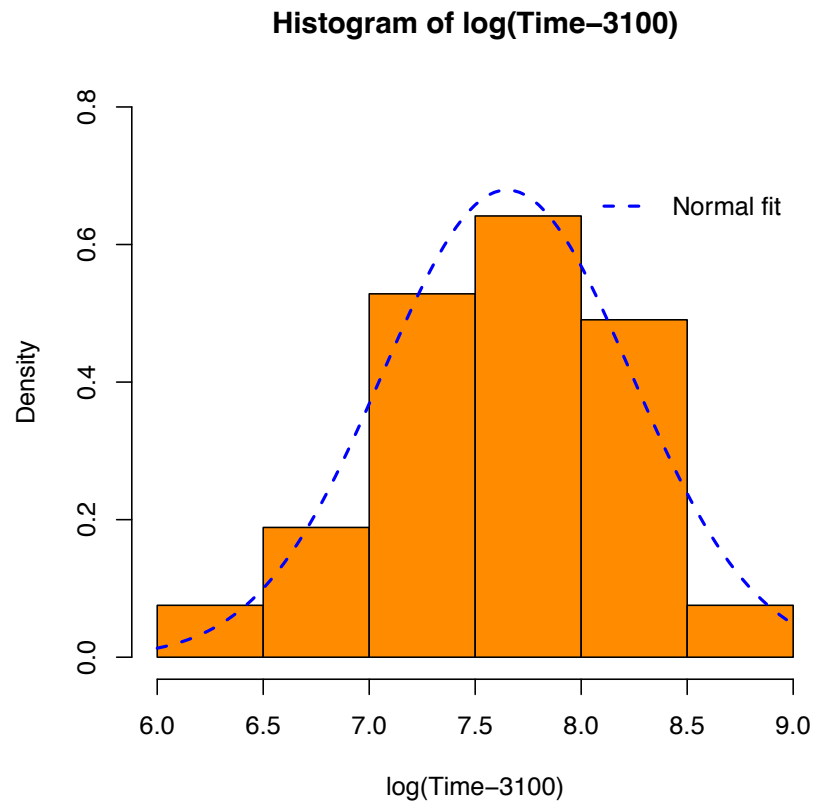


Figure 3: Comparison of distributions of Time before and after transformation

b). Plot a scatter plot $\log(\text{Time} - 3100)$ vs. Year. Do you see any trend?

We observe a negative relationship between transformed time variable and year. As years increase, transformed time seems to decrease, that is, it takes less time to finish the race. We are expecting a negative slope in the regression equation.

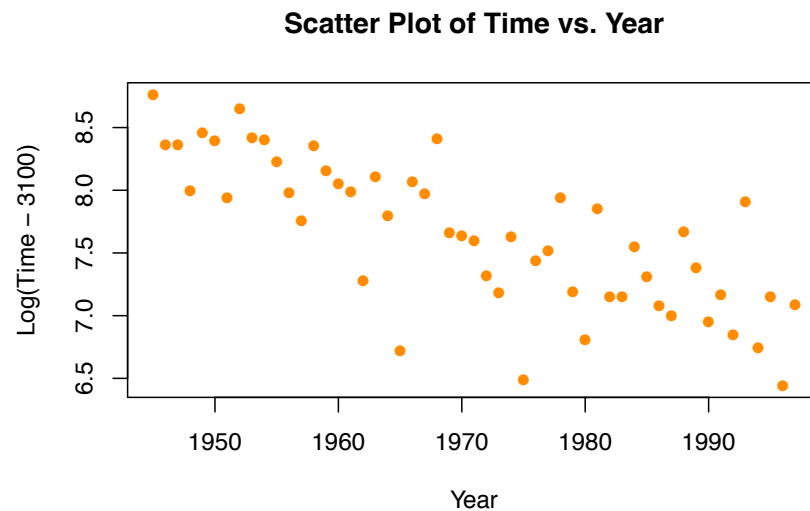


Figure 4: Scatter plot of transformed time variable vs. year

Code from R:

```
# Produce a scatterplot of log(Time - 3100) and Year
dev.new
pdf(file = "yacht_scatter.pdf", width=6, height=4)
plot(data$Year, log(data$Time-3100), pch=16, cex=1, col= "darkorange",
     xlab="Year", ylab="Log(Time - 3100)", main="Scatter Plot of Time vs. Year")
dev.off()
```

Write out a linear model to study the relationship between $\log(\text{Time} - 3100)$ and Year. Interpret your two parameters in the model.

Let

$$X = \text{year}$$

$$Y = \log(\text{Winning time in minutes on X year} - 3100)$$

Where Y is the depended variable and X is the predictor variable.

Linear model is then:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

First, use method of least squares to find slope and intercept for the model equation

using estimates

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$; $S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$;

```
> #Do calculations for the model:
> S_xx <- sum((data$Year-mean(data$Year))^2)
> S_xy <- sum((data$Year-mean(data$Year))*(log(data$Time-3100)+\\
-mean(log(data$Time-3100))))
> #slope
> S_xy/S_xx
[1] -0.02942316
> #intercept
> mean(log(data$Time-3100)) - (S_xy/S_xx)*mean(data$Year)
[1] 65.6426
```

Thus, we find the estimate of β_0 is $\hat{\beta}_0 = 65.64$, the estimate of β_1 is $\hat{\beta}_1 = -0.029$.

Then, we use the projection matrix to solve the problem.

Write matrix A as

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_{53} \end{pmatrix}$$

where $x_1 = 1945$ and $x_{53} = 1997$.

Vector $\vec{\hat{\beta}}$ of fitted values as

$$\vec{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

And \vec{y} as

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{53} \end{pmatrix}$$

Where y_i are $\log(\text{Winning time in minutes on the } x_i \text{ year} - 3100)$.

The solution is to find $\vec{\hat{\beta}}$, where $A\vec{\hat{\beta}}$ is the projection of \vec{y} onto the $\text{col}(A)$. Then,

$$\begin{aligned} A^T A \vec{\hat{\beta}} &= A^T \vec{y} \\ \vec{\hat{\beta}} &= (A^T A)^{-1} A^T \vec{y} \end{aligned}$$

We solve this using R:

Code from R:

```
> A <- cbind(data$Year, rep(1, nrow(data)))
> b <- solve(t(A) %*% A) %*% t(A) %*% log(data$Time - 3100)
> #Slope
> b[1]
[1] -0.02942316
> #Intercept
> b[2]
[1] 65.6426
```

Thus, we got the same solution as above using the projection matrix.

We interpret the two parameters in the following way:

β_1 is the difference in average transformed time ($\log(\text{Time} - 3100)$) to complete the race needed between the $(x+1)$ year and the x year. We observe that as we moved forward on the timeline (from 1945 to 1997), yachts needed less time to win the race. And we find the estimate for β_1 is -0.029.

β_0 is the expected value of $\log(\text{Time} - 3100)$ for the year of '0'. And we find the estimate for β_0 is 65.64.

c). Use the `lm` function in R to fit the model. Plot the regression line on the scatterplot.

Code from R:

```

> #Fit a model in R
> fit <- lm(log(data$Time-3100) ~ data$Year)
> fit
Call:
lm(formula = log(data$Time - 3100) ~ data$Year)
Coefficients:
(Intercept)      data$Year
  65.64260      -0.02942

```

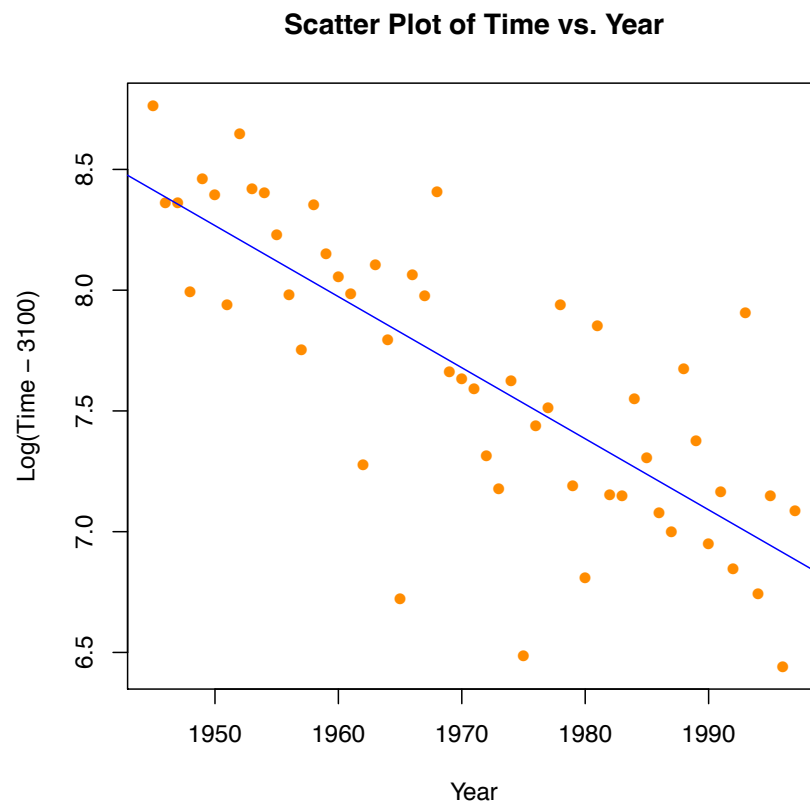


Figure 5: Scatter plot of transformed time variable vs. year with regression line

Problem 3. Some matrix practice.

a). Calculate AB by hand and check your result by R, where

$$A = \begin{pmatrix} 3 & 0 & 2 \\ -1 & 2 & 2 \\ 1 & 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -2 \\ 0 & -2 \\ -1 & 1 \end{pmatrix}.$$

$$\begin{aligned}
 AB &= \begin{pmatrix} 3 & 0 & 2 \\ -1 & 2 & 2 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & -2 \\ -1 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 3+0-1 & -6+0+2 \\ -1+0-2 & 2-4+2 \\ 1+0+1 & -2+0-1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & -4 \\ -3 & 0 \\ 2 & -3 \end{pmatrix}
 \end{aligned}$$

Check this result with R:

```
A = matrix(
  c(3, 0, 2, -1, 2, 2, 1, 0, -1),
  nrow=3,
  ncol=3,
  byrow=TRUE)
```

```
B = matrix(
  c(1, -2, 0, -2, -1, 1),
  nrow=3,
  ncol=2,
  byrow = TRUE)
```

```
C = A%%B
```

```
C
```

```
##      [,1] [,2]
## [1,]    1  -4
## [2,]   -3    0
## [3,]    2   -3
```

b). For $a, b > 0$, suppose

$$A = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

What are the eigenvalues of A ? What are its eigenvectors? What is the trace? What is the determinant?

To solve for the eigenvalues, find the determinant of $|A - \lambda I|$:

$$\begin{aligned}
 |A - \lambda I| &= \begin{vmatrix} a - \lambda & b \\ b & a - \lambda \end{vmatrix} \\
 &= (a - \lambda)(a - \lambda) - b^2 \\
 &= \lambda^2 - 2\lambda ab + a^2 - b^2 \\
 &= (\lambda - (a + b))(\lambda - (a - b)) \\
 \lambda_1 &= a + b, \quad \lambda_2 = a - b
 \end{aligned}$$

To find corresponding eigenvectors, solve $A\vec{x} = \lambda\vec{x}$

$$\begin{aligned}
 A\vec{x} &= \lambda\vec{x} \\
 \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= (a + b) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
 \begin{pmatrix} ax_1 + bx_2 \\ bx_1 + ax_2 \end{pmatrix} &= \begin{pmatrix} ax_1 + bx_1 \\ ax_2 + bx_2 \end{pmatrix} \\
 \implies x_1 &= x_2,
 \end{aligned}$$

One eigenvector is $\vec{x} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$

Repeat to find an eigenvalue for the second eigenvalue, $\lambda = a - b$

$$\begin{aligned}
 \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= (a - b) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
 \begin{pmatrix} ay_1 + by_2 \\ by_1 + ay_2 \end{pmatrix} &= \begin{pmatrix} ay_1 - by_1 \\ ay_2 - by_2 \end{pmatrix} \\
 \implies y_1 &= -y_2,
 \end{aligned}$$

Another eigenvector is $\vec{y} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$

Trace: sum of elements on the main diagonal $tr(A) = \sum_{i=1}^n a_{ii}$


$$tr(A) = a + a = 2a$$

Determinant:

$$|A| = a^2 - b^2$$

- c). Follow Problem 3b, Let $a = 2$, $b = 1$. Calculate A^{-1} by hand and then check your result by R.

By hand (using Gauss-Jordan elimination)–

$$\begin{aligned} & \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array} \right) \\ & \sim \left(\begin{array}{cc|cc} 1 & 1/2 & 1/2 & 0 \\ 1 & 2 & 0 & 1 \end{array} \right) \\ & \sim \left(\begin{array}{cc|cc} 1 & 1/2 & 1/2 & 0 \\ 0 & 3/2 & -1/2 & 1 \end{array} \right) \\ & \sim \left(\begin{array}{cc|cc} 1 & 1/2 & 1/2 & 0 \\ 0 & 1 & -1/3 & 2/3 \end{array} \right) \\ & \sim \left(\begin{array}{cc|cc} 1 & 0 & 2/3 & -1/3 \\ 0 & 1 & -1/3 & 2/3 \end{array} \right) \\ & \text{So } A^{-1} = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} \end{aligned}$$


Check using R:

```
A = matrix(
  c(2, 1, 1, 2),
  nrow=2,
  ncol=2,
  byrow = TRUE)

AI = matrix(
  c(2/3, -1/3, -1/3, 2/3),
  nrow=2,
  ncol=2,
  byrow = TRUE)

B= A%%AI
B

##      [,1] [,2]
## [1,]  1   0
```

```
## [2,] 0    1

#Another solution from R:
> A <- matrix(c(2,1,1,2), nrow = 2)
> #Calculate A inverse
> solve(A)
      [,1] [,2]
[1,] 0.6666667 -0.3333333
[2,] -0.3333333 0.6666667
```