# Solutions to Module 4 Homework

**1. (20 points)**

$X_1, ..., X_5$ are independent random samples from a distribution with mean 5 and standard deviation 3. Complete the following:

**(a)** For the sample mean $\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i$, find its mean $E(\bar{X})$ and standard deviation $sd(\bar{X})$.

**(b)** Can you find the $P(2 < \bar{X} < 5.1)$ approximately? If yes, what is your estimate for $P(2 < \bar{X} < 5.1)$? If no, why not?

**Solution:**

**a)** Treating $\bar{X}$ as the linear combination

$$\bar{X} = \frac{1}{5}\sum_{i=1}^{5} Xi$$

From central limit Theoerm the properties of linear combination of random variables , $\bar{X}$ has mean

$$\sum_{i=1}^{5} \mu = \mu = 5$$

Therefore

$$E(\bar{X}) = 5$$

The variance has property

$$(\frac{1}{5})^2 \sum_{i=1}^{5} \sigma^2 = \frac{(\sigma)\text{^}2}{5} = \frac{9}{5} = 1.8$$

Therefore, $Var(\bar{X}) = 1.8$

Since standars deviation = $sd(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{1.8} \approx 1.34$

**b) No,** I don't think its possible to find $P(2 < \bar{X} < 5.1)$ as the sample size n = 5 for for $\sqrt{n}(\bar{X} - \mu)$ to cover distribution .

# Solutions to Module 4 Homework

**2. (20 points)**
Suppose that for certain microRNA of size 20 the probability of a purine is binomially distributed with probability 0.7. Say there are 100 such microRNAs, each independent of the other. Let Y denote the average number of purine in these microRNAs. Find the probability that Y is great than 15. Please give a theoretical calculation, do NOT use Monte Carlo simulation to approximate. Show all the steps and formulas in your calculation.

Solution: As each microRNA follows a binomial distribution with the probability of the purine being 0.7 . We can take that as the probability of success p=0.7 . Each microRNA is of size 20 , so n=20. We can now calculate E(X) using binomial distribution E(X) = np and variance Var(X) = np(1-p)

Substituting the values in the respective equation v get :

$$E(X) = 20 * 0.7 = 14$$

$$Var(X) = 20 * 0.7 *(1 – 0.7) = 4.2$$

Therefore, the standard deviation is $sd(\bar{X}) = \sqrt{Var(\bar{X})} = 2.04939$

When each microRNA is treated as an independent random variable $X_i$, we get $\{X_1,\ldots\ldots,X_{100}\}$. Y is the average number of purines in microRNA, we can write this formula as

$$Y = \frac{1}{100}\sum_{i=1}^{100} X_i$$

$\{X_1,\ldots\ldots,X_{100}\}$ are independent random samples drawn from a distribution with mean = 14 and sd = 2.04939. Here we can use CLT to calculate probability of calculate probability of events about the sample mean Y ,

$$P( Y > 15 ) = 1 – P( Y \leq 15 )$$

Therefore, P( Y > 15 ) can be approximately calculated in R as:

```
> 1-pnorm(15, mean=14,sd=2.04393/sqrt(100))
[1] 4.977259e-07
```

Therefore, $P( Y > 15 ) \approx 0$. This is the probability that the average number of purines in the microRNA is greater than 15.

# Solutions to Module 4 Homework

**3. (20 points)**
Two genes' expression values follow a bivariate normal distribution. Let X and Y denote their expression values respectively. Also assume that X has mean 9 and variance 3; Y has mean 10 and variance 5; and the covariance between X and Y is 2.

In a trial, 50 independent measurements of the expression values of the two genes are collected, and denoted as $(X_1, Y_1)$, …, $(X_{50}, Y_{50})$. We wish to find the probability $P(\bar{X} + 0.5 < \bar{Y})$, that is, the probability that the sample mean for the second gene exceeds the sample mean of the first gene by more than 0.5.

Conduct a Monte Carlo simulation to approximate this probability, providing a 95% confidence interval for your estimation. Submit your R script for the Monte Carlo simulation, and a brief summary of the actual simulation results.

(**Extra bonus**: Provide a theoretical calculation of this probability. While the formula has not been given in the course lecture, it can be calculated from a bivariate normal distribution. You do not have to do this theoretical calculation if have no idea. You will get extra bonus points for doing it correctly.)

**Solution:** As the two random variables X and Y are not independent, we can model them via Monte Carlo simulation in R using the mvtnorm package. The code can be summarized as:

For this code we need to install mvtnorm so that we can use rmvnorm to generate realizations of X and Y.

We find the rvnorm , then the mean and variance for both x and y

Finally v find rnorm

# Solutions to Module 4 Homework

```
rm(list=ls())
require(mvtnorm)
data<-rmvnorm(50,mean=c(9,10),sigma=matrix(c(3,2,2,5),nrow=2))
meanxy<-apply(data,1,mean)
varxy<-apply(data,1,var)
mean(meanxy) + c(-1,1)*1.96*sqrt(var(meanxy)/50) #95%CI for mean
mean(varxy) + c(-1,1)*1.96*sqrt(var(varxy)/50) #95%CI for variance
sqrt(1.912846)
sqrt(4.173590)
rnorm(10000, mean=c(9.27048, 10.40593), sd=c(1.383057, 4.173590))
```

# Solutions to Module 4 Homework

**4. (20 points)**

Assume there are three independent random variables $X_1 \sim chisq(df = 10)$, $X_2 \sim Gamma(\alpha = 1, \beta = 2)$, $X_3 \sim$ t-distribution with m=3 degrees of freedom.

Define a new random variable Y as $Y = \sqrt{X_1} X_2 + 4(X_3)^2$.

Use Monte Carlo simulation to find the mean of Y. Submit your R script for the Monte Carlo simulation, and a brief summary of the actual simulation results.

Solution:

The random sampling of the distribution for $X_1 \sim chisq(df = 10)$ was modeled using rchisq. The random sampling of the distributioin for $X_2 \sim Gamma(\alpha = 1, \beta = 2)$ was modeled using rgamma and $X_3 \sim$ t-distribution with m=3 degrees of freedom. was modeled using rt .

The mean for variable Y as $Y = \sqrt{X_1} X_2 + 4(X_3)^2$ can be computed as

```
> rm(list=ls())
> #chi-square distribution
> X1 <- rchisq(10000, df=10)
> # gamma distribution
> X2 <- rgamma(10000, shape=1, scale=2)
> # t distribution
> X3 <- rt(10000, df=3)
> # calculate mean E(Y)
> Y <- sqrt(X1)*(X2)+4*(X3^2)
> meanY<- mean(Y)
> # Print mean E(Y)
> print (meanY)
[1] 18.27678
```

The computed sampling to find the mean of Y was:

Mean(Y) = 18.27678

# Solutions to Module 4 Homework

**5. (20 points)**
Complete exercise 10 in Chapter 3 of *Applied Statistics for Bioinformatics using R* (page 45-46). Submit the plot, and a brief explanation of your observation.
The problem refers to the density function of extreme value distribution in another book. You do not have to look for the other book, the density function is $f(x) = (e^{-x})e^{-e^{-x}}$.

Solution: The density function is $f(x) = (e^{-x})e^{-e^{-x}}$

Taking 1000 random numbers from normal distribution N(mean=0,sd= 1) using rnorm , selecting maxima and performing it 1000 times to get 1000 maxima from normal distribution.

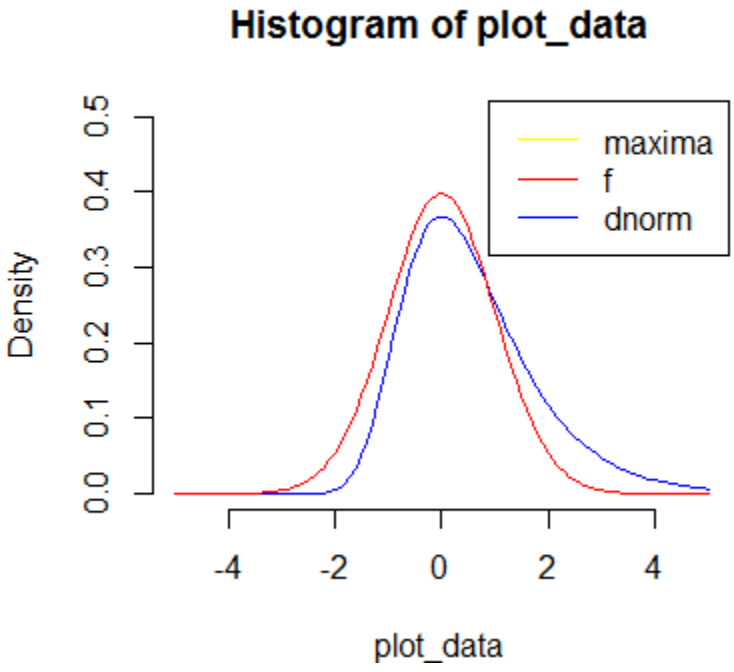
We then use the functions

an <- sqrt(2*log(n)) - 0.5*(log(log(n))+log(4*pi))*(2*log(n))^(-1/2)

bn <- (2*log(n))^(-1/2)

to get the set of maxim M = ( M – a )/b

Now we plot the density from the normalized maxima and density of normal distribution to get a histogram.



# Solutions to Module 4 Homework