Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper B)

May 2023 Examination

INDICATIVE SOLUTION

Introduction
The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

Solution 1:

i) set.seed(052023) (1)
u<-runif(150)
round(mean(u),2) (1)
[1] 0.52

This shows sample mean ~0.52. **[2]**

ii) chi<-qchisq(u,2) [2]

iii) gam<-qgamma(u,1,1/2) (1)

sum(chi-gam) (0.5)
[1] 0

We know the property that if X ~ Gamma(α,ʎ) then 2ʎX has χ^2 distribution with 2α degrees of freedom. (1)

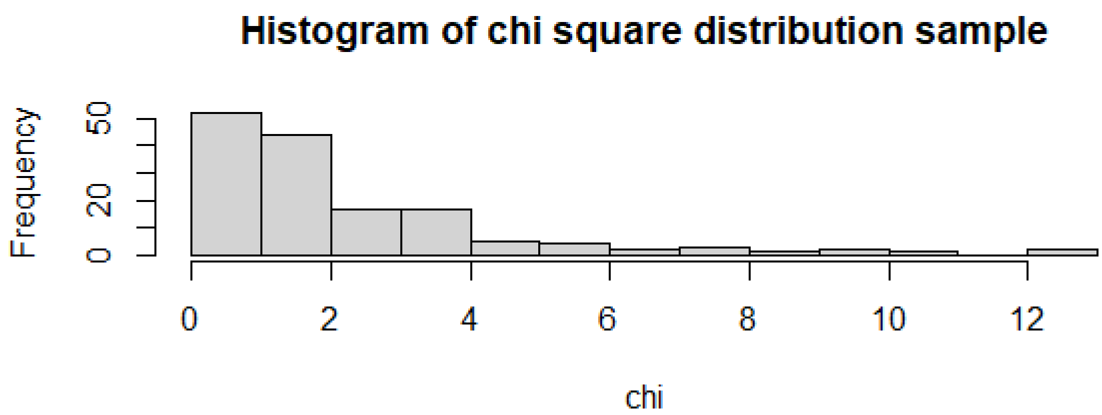
We have X ~ χ^2 with 2 degrees of freedom.
Above can be written as (2ʎX/2ʎ) ~ χ^2 with 2α degrees of freedom where α=1,ʎ =1/2 (0.5)
Thus,
X ~ Gamma(α,ʎ) with α=1,ʎ =1/2
X ~ Gamma(1,1/2) (0.5)

This is why both samples are same.

[Max 3]

iv)

a) hist(chi, main ="Histogram of chi square distribution sample") (1)
(1)



#Histogram shows positive skewed distribution (1)
[3]

b) > summary(chi) (1)

```
    Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
 0.006184  0.603187  1.534968  2.217946  2.983161 12.350562
```

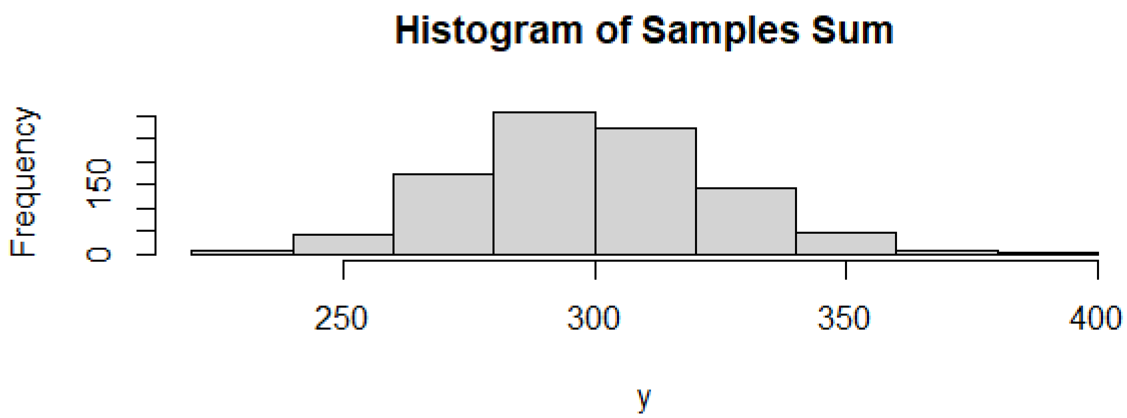
Alternate:
mean(chi) (0.5)
median(chi) (0.5)
#mean = 2.218
#median = 1.535

#Mean is greater than median since it is positively skewed distribution (1)
[2]

v)
```
set.seed(052023)
y = rep(0,1000)
for(i in 1:1000){
  y[i] = sum(rchisq(150,2))
}
```
(0.5)
(1)
(1)
(2)

[Max 4]

vi)
```
hist(y, main ="Histogram of Samples Sum")
```
(0.5)
(1)



#The distribution of sample sums is roughly symmetrical. (1)
#This displays Central Limit Theorem property. As the sample size (1)
#gets large , distribution move towards normality. (1)

[Max 3]
[19 Marks]

Solution 2:

i) Sales<-read.csv(<>) (1)

```
 > mean(Sales$Value)
[1] 1307.167
```
(1)

[2]

ii)

a)
```
cor(Sales$Value,Sales$City,method = "kendall")
[1] -0.2130327
```
[1]

b)
```
r<-cor(Sales$Value,Sales$Age)
> r*sd(Sales$Value)*sd(Sales$Age)
[1] -278.5169
```
(1)
(1)

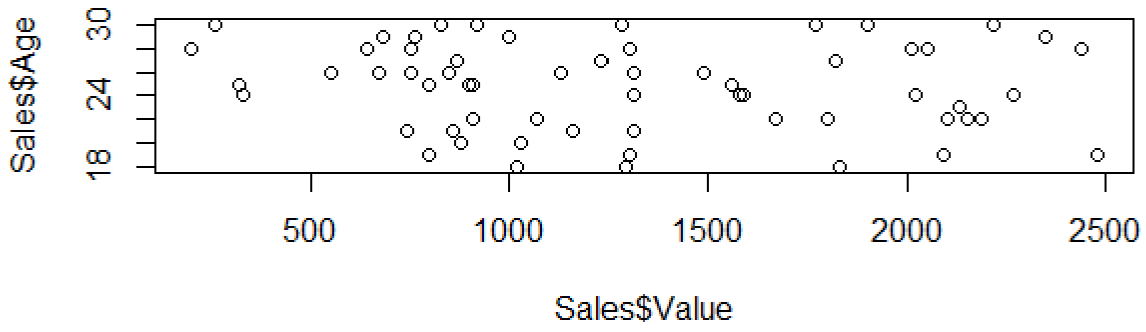
Alternate:
```
cov(sales$Value,sales$Age)
[1] -278.5169
```
(1.5)
(0.5)

Credit is given if Kendall covariance is computed.

[2]

iii)

a) plot(Sales$Value,Sales$Age) (1)



(1)

No trend (showing linear relationship) is visible from the scatter plot. (1)
Most likely it indicates that correlation is zero.

[3]

b) > cor.test(Sales$Value,Sales$Age) (1)

```
Pearson's product-moment correlation

data:  Sales$Value and Sales$Age
t = -0.95277, df = 58, p-value = 0.3447
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3665088  0.1340120
sample estimates:
      cor
-0.1241369
```

(1) for CI

Confidence Interval is (-0.367,0.134)

Since 0 lies in the confidence interval, we cannot reject the hypothesis that correlation coefficient = 0. (1)

[3]

iv)

a) Since correlation between Value and Age is (close to) 0, age can be excluded. (2)

b) > model1<-lm(data = Sales,Value~Device+City+Age) (1)
> summary(model1) (1)

```
Call:
lm(formula = Value ~ Device + City + Age, data = Sales)

Residuals:
   Min      1Q  Median      3Q     Max
-768.21 -239.97  -19.05  236.74  959.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2887.26     378.29   7.632 3.12e-10 ***
DeviceMobile -1022.97     110.18  -9.284 6.28e-13 ***
City         -135.98     100.03  -1.359   0.1795
Age           -26.25      13.41  -1.958   0.0552 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 377 on 56 degrees of freedom
Multiple R-squared:  0.6388,	Adjusted R-squared:  0.6195
F-statistic: 33.01 on 3 and 56 DF,  p-value: 2.019e-12
```

Device = 1 for Mobile and 0 for Laptop. Parameter for this variable is significant.
Alternate : Device is significant (0.5)

Age and City are not significant … (1)
…….since p-value > 0.05. (0.5)
Age is expected to insignificant per the earlier analysis. (0.5)
[Max 4]

c)

```
model2<-lm(data = Sales,Value~Device)
>
> anova(model1,model2)
Analysis of Variance Table

Model 1: Value ~ Device + City + Age
Model 2: Value ~ Device
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     56 7958529
2     58 8716151 -2   -757622 2.6655 0.07838 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1) (1)

H0: β (City) = β (Age) =0 against H1: atleast one of β (City) or β (Age) is non-zero. (1)
In model 2, there are 2 less parameters thus -2 degrees of freedom in Anova analysis (1)
p-value >0.05 showing we can't reject H0: β (City) = β (Age) =0. (1)
This indicates neither of the covariates have strong relationship with Order value. (0.5)
[Max 5]

d)

```
summary(model2)

Call:
lm(formula = Value ~ Device, data = Sales)

Residuals:
   Min     1Q Median     3Q    Max
-810.9 -240.1   -8.7  271.6  889.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2056.47      94.02  21.873  < 2e-16 ***
DeviceMobile -1045.54     111.06  -9.414 2.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 387.7 on 58 degrees of freedom
Multiple R-squared:  0.6044,	Adjusted R-squared:  0.5976
F-statistic: 88.62 on 1 and 58 DF,  p-value: 2.77e-13
```

(1)

> # Value = 2056.47 - 1045.54 X Device, (1.5)
where Device= 1 for Mobile else 0 (0.5)
Alternate:
> # Value = 2056.47 - 1045.54 X Device Mobile (2)

[Max 3]

v)

a) confint(model2,level=.95) (1)

```
              2.5 %    97.5 %
(Intercept)   1868.268 2244.6737
DeviceMobile -1267.855 -823.2257
> # C.I. (-1267.8,-823.2)
```
(1)

[2]

b)

```
> residual<-model2$residuals                         (1)
> n<-length(residual)                                (0.5)
> varhat<-var(residual)                              (1.5)
> (n-2)*varhat/qchisq(c(0.975,.025),58)              (2)
[1] 105867.1 220588.2
> # C.I. (105867,220588)                             (1)
```

[Max 5]

vi)

a)

```
m<-mean(Sales$Order)                                 (1)
> m
[1] 2.383333                                         (1)
 Mu hat = 2.383
```

[2]

b) table(Sales$Order) (1)

```
 0  1  2  3  4  5  7
 8 13 13  8 12  5  1
```
(1)

[2]

c)

```
a<-as.numeric(table(Sales$Order))                    (1)
#using above table, combine order 5 and 5+
a[6]=sum(a[6:7])                                     (1)
a<-a[-7]  #to remove 5+ as combine above             (0.5)

e<-dpois(c(0:4),m)                                   (1)
sum(e)                                               (0.5)
[1] 0.9062099
e[6]<-1 - sum(e)                                     (1)
sum(e)                                               (1)
[1] 1

chisq.test(x=a,p=e)                                  (2)
        Chi-squared test for given probabilities
data:  a
X-squared = 6.0026, df = 5, p-value = 0.306
```

Since p-value is greater than 0.05 , we can reject the hypothesis that number of order follows poisson distribution. (1)

[Max 8]

vii)

a)

```
glm2<-glm(data=Sales,Order~Device,family=poisson(link="log"))   (2)
> summary(glm2)                                                 (1)

Call:
glm(formula = Order ~ Device, family = poisson(lin = "log"),
```

```
data = Sales)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.4495  -0.6149   0.0000   0.5491   1.9652  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -0.1942     0.2673  -0.726    0.468    
DeviceMobile   1.2928     0.2814   4.594 4.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 81.185  on 59  degrees of freedom
Residual deviance: 51.570  on 58  degrees of freedom
AIC: 199.88

Number of Fisher Scoring iterations: 5
```

[3]

b) Log link function is used in above model. (0.5)

Log link implied log mu = linear predictor. Inverting to mu leads to mu = exp(linear predictor). Exponent will make sure mu always remain greater than 0, an essential feature for poisson distribution with mean mu. (1.5)

[2]

viii)

a)

```
Customers<- data.frame(Device =c("Mobile","Mobile"),Age =c(18,28), City = c(1,2))   (1.5)
> predict.glm(glm2,newdata = Customers,type= "response")                            (2)
1 2
3 3                                                                                 (0.5)
```

[4]

b) Only device is used in the model and for both customers, device is same and thus, the predicted value is same for both customers. [2]

ix)

```
Channel <-data.frame(Device =c("Mobile","Laptop"))                        (1.5)
> pred_order <- predict.glm(glm2,newdata = Channel,type= "response")      (1.5)
> pred_value <- predict(model2,newdata = Channel)                         (1)
>
> totalvalue<-pred_order * pred_value                                     (1.5)
> totalvalue
       1        2 
3032.791 1693.564                                                         (1)
> #total value for Mobile = 3032.8
> # and for Laptop = 1693.6
```

[Max 6]

[61 Marks]

Solution 3:

i)

```
#m1x0 = E[X] and                                                          (0.5)
#m2x0=E[x^2] and var = E[x^2] - E[x]^2 >> m2x0 = var + E[x]^2             (0.5)

prior_mean=60
prior_sd=5
```

```
m1x0= prior_mean                                   (0.5)
m2x0 = prior_sd^2 + prior_mean^2                   (1)
> m1x0                                             (0.5)
[1] 60
> m2x0
[1] 3625                                           (0.5)
```
[Max 3]

ii)

a)
```
# theta follows N(prior mean,prior variance)
# Random Variable X follows N(theta,variance)
# posterior distribution of theta follows Normal with                        (0.5)
# post mean = (n*sample mean/variance +prior mean/prior variance)/          (1.5)
#                   (n/variance + 1/prior variance)
#post variance = 1/(n/variance + 1/prior variance)                           (1)
Max 2
n<-5
sample_mean<- 340/5
sdev<-20

post_mean = (n*sample_mean/sdev^2 + prior_mean/prior_sd^2)/(n/sdev^2 + 1/prior_sd^2)   (2)
post_var= 1/(n/sdev^2 + 1/prior_sd^2)                                                   (1)

> post_mean                                        (0.5)
[1] 61.90476
> post_var                                         (0.5)
[1] 19.04762
> sqrt(post_var)                                   (0.5)
[1] 4.364358
```
[Max 6]

b)
```
sample2_mean<-3400/50                              (0.5)
n2<-50
post2_mean = (n2*sample2_mean/sdev^2 + prior_mean/prior_sd^2)/(n2/sdev^2 + 1/prior_sd^2)   (1.5)

post2_var= 1/(n2/sdev^2 + 1/prior_sd^2)            (0.5)
> post2_mean                                       (0.5)
[1] 66.06061
> post2_var                                        (0.5)
[1] 6.060606
> sqrt(post2_var)
[1] 2.46183
```
[Max 3]

iii)

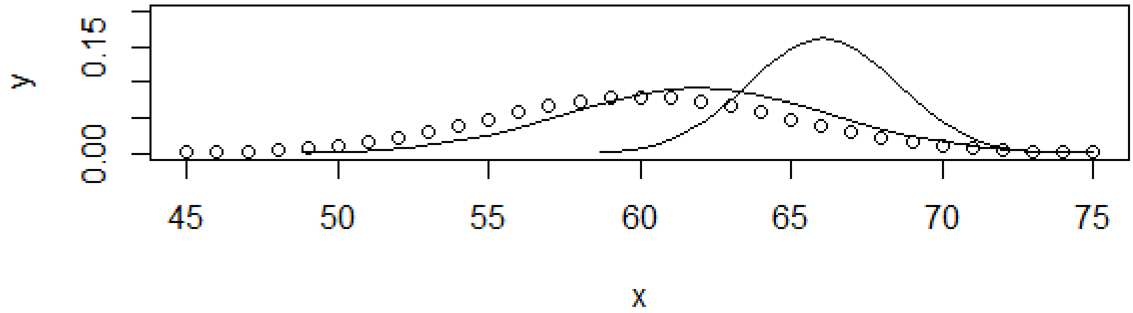
a)
```
x<-60+seq(-3,3,by=0.2)*5
y<-dnorm(x,mean=60,sd=5)
plot(x,y,ylim=c(0,.2))                             [1]
```

b)
```
px1<-post_mean+seq(-3,3,by=0.2)*sqrt(post_var)     (0.5)
py1<-dnorm(px1,mean=post_mean,sd=sqrt(post_var))   (0.5)
lines(px1,py1)                                     (1)
```
[2]

c)
```
px2<-post2_mean+seq(-3,3,by=0.2)*sqrt(post2_var)     (0.5)
py2<-dnorm(px2,mean=post2_mean,sd=sqrt(post2_var))   (0.5)
lines(px2,py2)                                       (1)
```

[2]

d)



(1)

The posterior distribution with sample size =5 is close to prior distribution. There is slight shift to mean towards sample mean and similar dispersion. (1)

When the sample size increased, the posterior distribution moves towards sample mean and dispersion. (1)

More weight is given to sample where sample is big. Further, the variation reduced with larger sample size. (1)

[Max 3]

[20 Marks]

****