Atividade 03 - Introdução ao Aprendizado Estatístico

Marcel Dantas de Quintela

Atividade apresentada como parte das avaliações da disciplina de Introdução ao Aprendizado Estatístico, ministrada pela ***Profª. Mariane Barros Alves*** para o curso de Especialização em Ciência e Dados do Instituto de Matemática da Universidade Federal do Rio de Janeiro

---

Instruções

Assuma que a variável aleatória Y denote custos de cancelamento de contratos (em milhares de Reais) registrados por uma empresa. O arquivo **custos.txt** contém 1000 registros de custos de cancelamento, selecionados aleatoriamente da base de dados da empresa. Na primeira coluna, estão os custos na escala original e, na segunda coluna do arquivo, tem-se o logaritmo natural de cada custo (Z). Os dados usados serão os log-custos (segunda coluna do arquivo de dados).

```
dados<-matrix(scan(file="custos.txt"),ncol=2,byrow=T) #Leitura do arquivo
dados<-as.data.frame(dados) #transformar em dataframe
names(dados)<-c("Custos","LogCustos")
```

Faremos a análise bayesiana com foco na média de um modelo Normal(θ, σ^2). para os log-custos (Z). Ao longo de toda a análise, assuma a variância σ^2 conhecida. Fixaremos seu valor na estimativa pontual obtida no trabalho 1, ou seja, assumiremos $\sigma^2 = 0.01$

Obs: Leia atentamente os enunciados abaixo. Você pode adaptar o código R: *Preditiva_Normal-Normal.R* para responder as questões.

ETAPA 1 (40%): *Estimação de θ - Obtenção da distribuição a posteriori e sua exploração*

a. Assuma que, a priori, θ siga uma distribuição Normal $(0, 100)$. Utilize os resultados do slide 92 (Conjugação Normal-Normal) e obtenha a distribuição a posteriori de θ, comentando a influência da priori e da verossimilhança sobre essa distribuição.

```
#Informações dos Log-Custos dos Dados
n<-length(dados$LogCustos)
yhat<-mean(dados$LogCustos)
sigma2<-0.01
phi<-1/sigma2

#Priori:
mu0<-0
phi0<-1/10^2
sd0<-sqrt(1/phi0)

#Posteriori:
phi1<-phi0+(n*phi)
mu1<-(1/phi1)*(phi0*mu0+n*phi*yhat)
sd1<-sqrt(1/phi1)
```

Table 1: Resumo Paramétrico das Distribuições

	Média	Sigma²
LogCustos	1.996	1.000000e-02
Priori	0.000	1.000000e+02
Posteriori	1.996	9.999999e-06

```
x<-data.frame("LogCustos"=c(yhat,0.01),
              "Priori"=c(mu0,sd0^2),
              "Posteriori"=c(mu1,sd1^2),
              row.names = c("Média","Sigma²"))

kable_classic(kable(t(x),digits = c(3,12),
                    caption = "Resumo Paramétrico das Distribuições"),
              full_width = T, html_font = "Cambria")
```

A variância da priori é muito alta, tornando a crença inicial vaga ou pouco informativa.

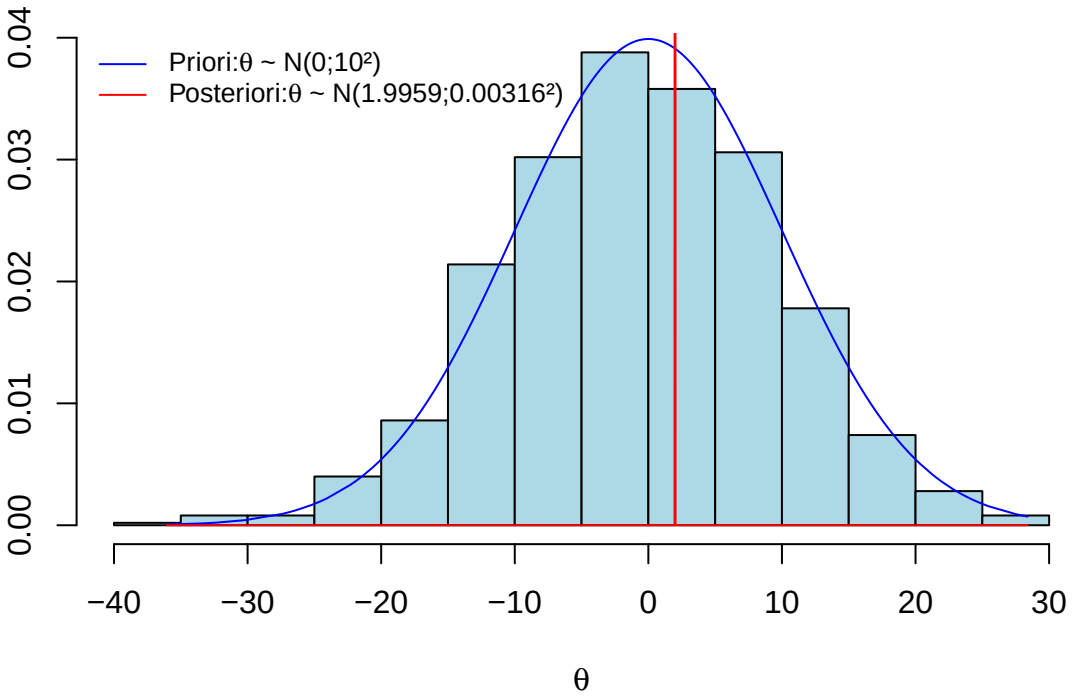
Aliado ao fato de termos uma amostra consideravelmente grande, fez com que a função de verossimilhança tivesse muito peso na construção da distribuição a posteriori de θ.

b. Esboce o gráfico da distribuição a priori e da distribuição a posteriori de θ. A amostra observada parece ter modificado as crenças a priori sobre θ? Comente.

```
set.seed(125)
prio.theta<-rnorm(1000,mu0,sd0)
post.theta<-rnorm(1000,mu1,sd1)
```

```
hist(prio.theta,
     breaks=20,
     prob=T,
     col="light blue",
     xlab=expression(theta),
     ylab="",
     main = "")
lines(sort(prio.theta), dnorm(sort(prio.theta),mu0, sd0),
      lty=1, col="blue")
lines(sort(prio.theta), dnorm(sort(prio.theta),mu1, sd1),
      lty=1,lwd=1,col="red")

legend("topleft",
       legend=c(expression(paste("Priori:",theta," ~ N(0;10²)")),
                expression(paste("Posteriori:",theta," ~ N(1.9959;0.00316²)"))),
       lty=c(1,1),col=c("blue","red"),bty = "n", cex=0.8)
```



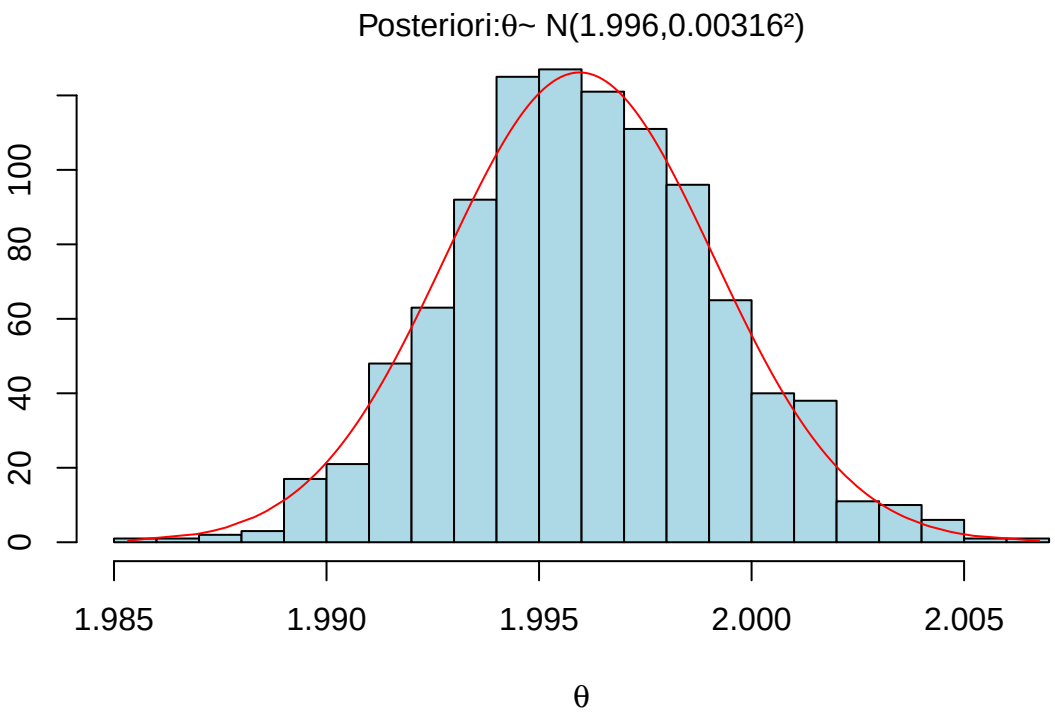
A amostra observada modificou as crenças existentes sobre o parâmetro investigado.

É possível observar na amostra coletada que a posteriori é tão estreita que quando plotada dentro do intervalo paramétrico da priori ela tem aparência de um "T" invertido.

A distribuição à posteriori apresenta caldas tendendo a zero e miolo sendo praticamente uma reta vertical centrada em sua média $\mu_1 \cong 1.99595$.

Esta representação indica altíssima precisão da posteriori resultante da predominância exercida pela verossimilhança dada a amostra grande coletada e pela falta de informação da priori.

```
hist(post.theta,
     breaks=30,
     prob=T,
     col="light blue",
     xlab=expression(theta), ylab="",
     main=mtext(bquote(plain("Posteriori:")*theta*plain("~ N(")*.(round(mu1,3))*
                       plain(",")*.(round(sqrt(1/phi1),5))*plain("²)"))))
lines(sort(post.theta), dnorm(sort(post.theta),mu1, sd1),
      lty=1, col="red")
```



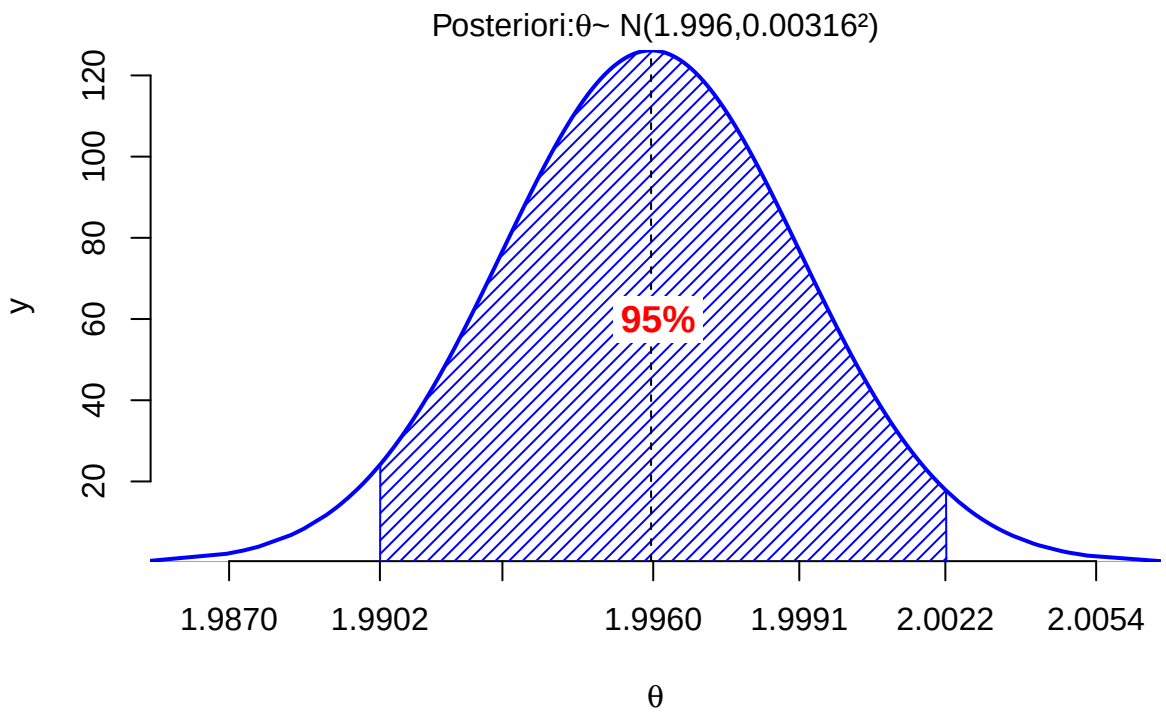
A representação da distribuição à Posteriori dentro de seu espaço paramétrio nos permite observar o alto grau de precisão existente. Dada sua baixíssima variância, o domínio da posteriori passeia por valores numa escal milessimal, concentrando sua massa em torno de 1,996 (média).

c. Utilize a distribuição a posteriori para obter uma estimativa intervalar, ao nível de credibilidade 95%, para a média dos log-custos, θ.

```
x<-sort(post.theta)
y<-dnorm(x,mu1,sd1)

a<-quantile(post.theta,probs=c(0.025,0.975))
# qnorm(0.025,mu1,sd1); # qnorm(0.975,mu1,sd1) # qnorm usa Z-score
i<-x>=a[[1]]&x<=a[[2]]

plot(x,y,type = "l",col="blue",lwd=2,xaxs="i",yaxs="i",axes=F,
     xlab=expression(theta), bty = "n",
     main=mtext(bquote(plain("Posteriori:")*theta*plain("~ N(")*.(round(mu1,3))*
                         plain(",")*.(round(sqrt(1/phi1),5))*plain("²)"))))
axis(1, round(c(a[[1]]-sd1,a[[1]],mu1-sd1,mu1,mu1+sd1,a[[2]],a[[2]]+sd1),4))
axis(2)
polygon(c(a[[1]],x[i],a[[2]]),c(0,y[i],0),col="blue", density = 20, border="blue")
abline(h=0)
abline(v=mu1,lty=2)
mark(labels = "95%", x =mu1-0.0008 , y = 60, col="red", col_bg = "white", cex=1.2)
```




```
cat("Considerando a amostra coletada (n=1000) é possível afirmar,\n",
    "com credibilidade de 95%, que a média dos log-custos está\nentre ",
    prettyNum(a[[1]],digits=5)," e ",
    prettyNum(a[[2]], digits=5),".", sep="")
```

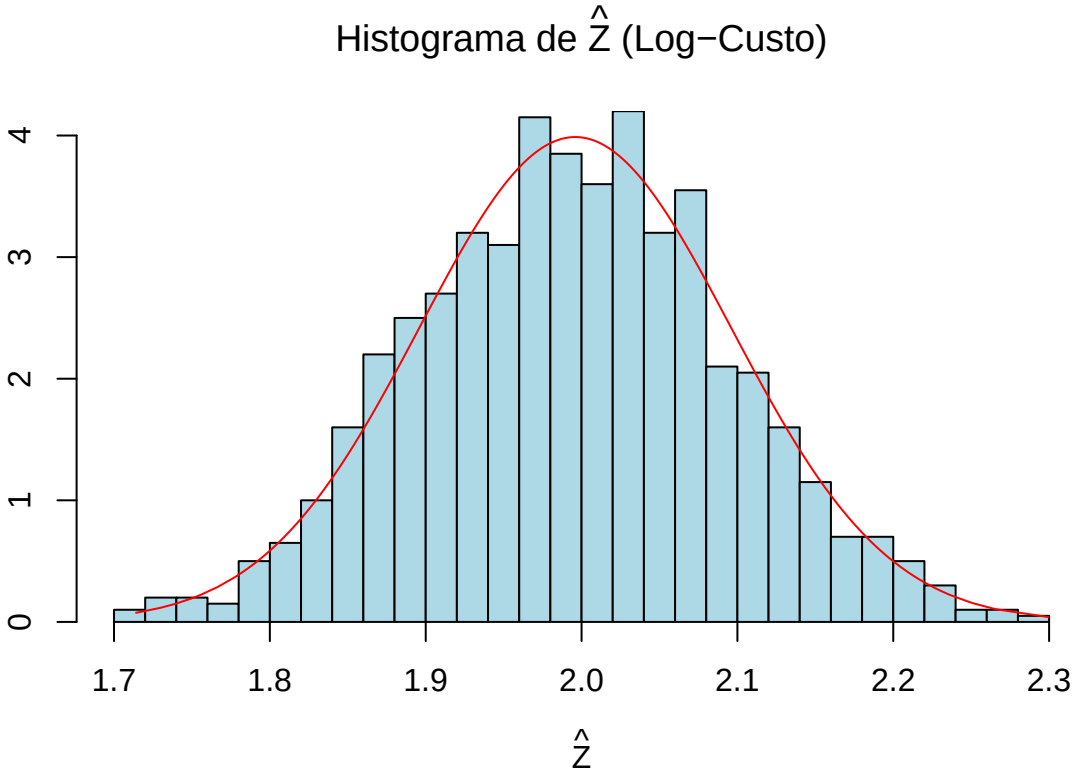
```
## Considerando a amostra coletada (n=1000) é possível afirmar,
## com credibilidade de 95%, que a média dos log-custos está
## entre 1.9902 e 2.0022.
```

ETAPA 2 (40 %): *Exploração do Comportamento de uma observação futura Y (log-custo) – obtenção de distribuição preditiva.*

d. Obtenha, por amostragem, uma aproximação para a distribuição preditiva de $Z(log - custo)$.Faça um histograma da amostra da distribuição preditiva.

```
set.seed(524)
z.pred<-NULL
for (i in 1:length(post.theta)){
z.pred[i]<-rnorm(1,post.theta[i],sqrt(1/phi))
}
```

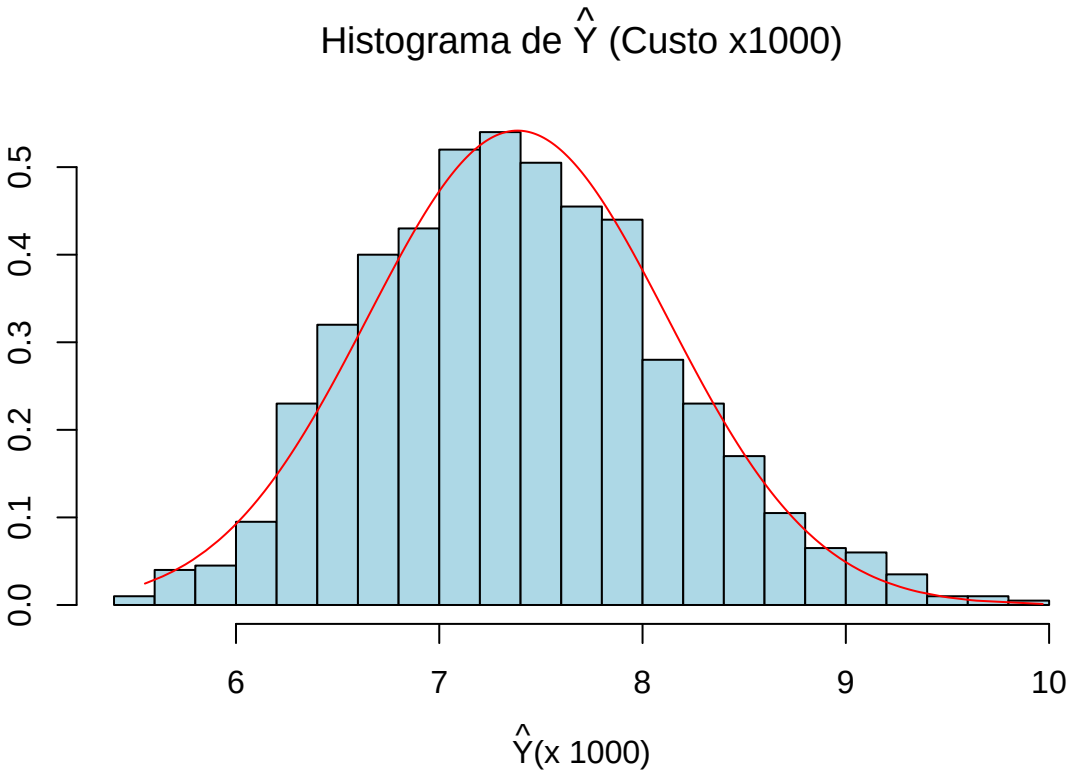
```
hist(z.pred,
     breaks=30,
     main=expression(paste("Histograma de ",hat(Z)," (Log-Custo)")),
     ylab="",
     xlab=expression(hat(Z)),
     prob=T,
     yaxs="i",
     col="light blue")
phi.pred<-phi*phi1/(phi1+phi) #Atualização da variância pela posteriori
lines(sort(z.pred),dnorm(sort(z.pred),mu1, sqrt(1/phi.pred)),
      lty=1, col="red")
```



e. A partir da amostra da distribuição preditiva de $Z(log - custo)$ gere uma amostra da distribuição preditiva de $Y = exp(Z)$. Faça um histograma da distribuição preditiva de $Y(custo)$ e teça comentários sobre o comportamento de custos de cancelamento futuros.

```
y.pred<-exp(z.pred)
hist(y.pred,
     breaks=30,
     main=expression(paste("Histograma de ",hat(Y)," (Custo x1000)")),
     ylab="",
     xlab=expression(paste(hat(Y),"(x 1000)")),
```

```
     prob=T,
     col="light blue")
lines(sort(y.pred),dnorm(sort(y.pred),mean(y.pred), sd(y.pred)),
      lty=1, col="red")
```



Retomando a escala nominal dos custos, a distribuição preditiva de Y volta ter comportamento semelhante os custos coletados na amostra. Apresentando leve assimetria a direita, fato pelo qual decidiu-se anteriormente pela transformação logarítmica antes de prosseguir com aos estudos numa escala Log-Normal.

f. Obtenha, a partir da amostra da preditiva de $Y(custos)$:

- a probabilidade de que um custo de cancelamento futuro ultrapasse 9 (mil Reais);

```
print(paste("P(Y>9) =",length(y.pred[y.pred>9])/length(y.pred)))
```

```
## [1] "P(Y>9) = 0.024"
```

- o custo esperado de cancelamento;

```
cat("O custo esperado de cancelamento é de \nR$",
    prettyNum(mean(y.pred)*1000, big.mark=".",
              decimal.mark=",", digits=6),".",sep="")
```

```
## O custo esperado de cancelamento é de
## R$7.384,97.
```

- estimativa intervalar, ao nível de credibilidade 95%, para um custo futuro de cancelamento.

```
q<-quantile(y.pred,probs=c(0.025,0.975))
cat("Pode-se precisar que 95% dos custos de cançelamento\nestarão entre R$",
    prettyNum(q[[1]]*1000, big.mark=".",decimal.mark=",", digits=6),
    " e R$",
    prettyNum(q[[2]]*1000, big.mark=".",decimal.mark=",", digits=6),
    ".",sep="")
```

```
## Pode-se precisar que 95% dos custos de cançelamento
## estarão entre R$6.071,82 e R$8.988,29.
```