ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

INTRODUÇÃO AO APRENDIZADO ESTATÍSTICO

TRABALHO 3 (PESO: 30% DA AVALIAÇÃO TOTAL)

Neste trabalho, retomaremos os dados usados no trabalho 1.

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).

Obs: Você pode ler o arquivo de dados e conferir se a leitura foi correta usando as linhas de comando R:

```
dados<-matrix(scan(file="custos.txt"),ncol=2,byrow=T) #Leitura do arquivo
head(dados) #Visualização das primeiras linhas do arquivo lido
dim(dados) #Conferindo a dimensão dos do arquivo
```

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

ETAPA 2 (60 %): 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.

(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.

(f) Obtenha, a partir da amostra da preditiva de Y (custos):
- a probabilidade de que um custo de cancelamento futuro ultrapasse 9 (mil Reais);
- o custo esperado de cancelamento;
- estimativa intervalar, ao nível de credibilidade 95%, para um custo futuro de cancelamento.