Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Programa FRH-Analytics UFCG/HP
Disciplina: Fundamentos de Analytics

Lab02 - Estatística Descritiva 1 - 16/02/2011

Exercício

Para este lab, usaremos dados de ociosidade de máquinas de alguns laboratórios do DSC, sendo dois laboratórios de pesquisa (GMF e LSD) e dois laboratórios de alunos de graduação (LCC e LCC2). A máquina é considerada ociosa quando não há atividade do usuário interagindo na mesma, ou seja, o teclado e o *mouse* não são utilizados por um certo período de tempo.
Os dados possuem as seguintes colunas:

- **intervalo** *[integer]*: tempo em segundos em que a máquina passou em um determinado estado (ociosa ou ocupada).
- **ociosa** *[logical]*: indica o estado da máquina correspondente ao intervalo de tempo indicado, sendo *TRUE* quando a máquina está ociosa ou *FALSE* caso contrário.
- **maquina** *[character]*: identificador da máquina.
- **laboratorio** *[character]*: identificador do laboratório ao qual a máquina pertence.

Os dados podem ser acessados em:
http://www2.lsd.ufcg.edu.br/~marcus/analytics/dados/atividade-maquinas-dsc.txt

Nesta lab, teremos dois tipos de resposta. Uma em formato de *script R* e outra em formato texto *pdf*. Para cada questão, os arquivos de resposta devem ter os seguintes nomes:

- *seunome-lab01-questaoX.R* para respostas no formato de scripts *R*
- *seunome-lab01-questaoX.pdf* para respostas discussivas em formato texto.

Questões:

1) Escreva um script *R* que calcula as seguintes estatísticas: **média** (que você julgar mais adequada); **mediana**; **mínimo** e **máximo**; **1o e 3o quartis**; **5-percentil** e **95-percentil**; **desvio padrão**; **IQR**. O script também deve gerar gráficos para: **histograma** e **boxplot**.
Os dados dos quais as estatísticas devem ser calculadas e o gráficos gerados são:

- Duração do intervalo de tempo em que as máquinas estiveram ociosas, agrupadas por laboratório.

- Duração do intervalo de tempo em que as máquinas estiveram ocupadas, agrupadas por laboratório.
- Proporção do tempo em que as máquinas estiveram ocupadas (considerando os intervalos medidos), agrupadas por laboratório.
- Quantidade de vezes que as máquinas mudaram de estado, agrupadas por laboratório.

Devem ser gerados na execução do script **três** arquivos de saída para cada um dos 4 dados listados acima: (i) um arquivo texto contendo a tabela de estatísticas, onde cada linha é um laboratório e, nas colunas, primeiro deve ter o nome do laboratório e em seguida cada coluna deve ter uma das estatísticas relacionadas a ele, resultando 11 colunas no total; (ii) um arquivo de imagem, no formato *png*, com os histogramas; e (iii) um arquivo de imagem, no mesmo formato, com os *boxplots*.
Obs: *consultar os comandos ?png e ?dev.off para salvar as imagens.*

2) Usando como base as estatísticas e gráficos extraídos na questão anterior, discuta qual a melhor maneira de sumarizar cada um dos 4 dados apresentados, com relação a índices de tendência central, índices de dispersão, além de uma análise de outliers.
Apresente a discussão em um arquivo no formato *pdf*. É fortemente recomendado o uso de números e gráficos na elaboração dos seus argumentos.

3) Escreva um script *R* que indica quais as 10 máquinas que passaram mais tempo ocupadas, baseado na proporção do tempo em que elas tiveram ocupadas, ordenadas do maior para o menor valor. Se você julgar que houve erro na medição e que é preciso fazer remoção de *outliers*, isto pode ser adicionado no script, contanto que seja adicionado um comentário argumentando o motivo da remoção.

O arquivo de saída deve ter: na 1a coluna o nome da máquina, na 2a coluna o nome do laboratório a qual pertence e na 3a coluna o valor da proporção de tempo que ela esteve ocupada.

4) Suponha que você deve fazer uma análise de qual laboratório as pessoas mais trabalham / estudam (ou seja, as máquinas ficam mais tempo ocupadas). Qual laboratório você indicaria? Discuta como você chegou à sua conclusão, apresentando o resultado no mesmo formato da questão 2.

Desafio (opcional):

1 Descrição

Esse desafio consiste em duas partes: (i) implementação em R da solução de um problema que usa medidas estatísticas aprendidas em aula; (ii) discussão das decisões estatísticas definidas no algoritmo.

1.1 Implementação

Você está procurando um computador do DSC para rodar um serviço de monitoramento de mensagens no Twitter. Você avalia o computador sob dois critérios: disponibilidade e volatilidade. A disponibilidade é o total de tempo em que o computador permanece com o atributo “ociosa” com o valor “TRUE”, quanto mais disponível a máquina for mais tráfego de mensagens ela poderá monitorar. Por outro lado, a volatilidade é a quantidade de vezes em que o computador mudou o atributo “ociosa”, quanto mais volátil a máquina for mais vezes você precisará acessá-la para verificar manualmente o estado do monitoramento e reiniciar o serviço.

Primeiramente, aplique a *Rule of Thumb* para eliminar computadores com disponibilidade e volatilidade destoantes, i.e, possíveis outliers. Após isso, em uma fase de pre-processamento realizada em cada laboratório, elimine os computadores que apresentem ambos disponibilidade menor que o 55 percentil e volatilidade maior que o 55 percentil. Do conjunto de todos os computadores que satisfazem essa restrição, você está interessado em identificar aquele que apresenta maior disponibilidade e calcular quanto porcento ele é mais disponível e quanto porcento ele é menos volátil que a média e a mediana dos dados filtrados.

O desafio será avaliado pela corretude da resposta gerada segundo a descrição acima.

1.2 Discussão das decisões

Além do script, você deve discutir sobre: (i) o que você mudaria no algoritmo para que ele fizesse uma melhor escolha e (ii) se você acha que esse algoritmo pode ser generalizado para outras bases de dados.

2 Instruções para entrega

Você deve entregar um script *R* que recebe como entrada um arquivo CSV de atividade das máquinas (no formato dos dados de exemplo do lab). Exemplo de execução do script: $ Rscript meuscript.R dados.csv. A implementação deve seguir a descrição do problema e gerar como saída em um arquivo em formato texto com o nome da máquina escolhida, o laboratório em que ela está e o ganho que ela proporciona em termos de volatilidade e disponibilidade em relação à média e a mediana. Além disso, deve ser entregue um arquivo *pdf* com a discussão das decisões.