

# Estatística Aplicada

Estatística é a ciência que se utiliza das teorias probabilísticas para explicar a frequência da ocorrência de eventos, tanto em estudos observacionais quanto em experimento modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros, conforme o caso. A estatística tem como objetivo fornecer informação (conhecimento) utilizando quantidades numéricas. Seguindo este raciocínio, a estatística divide o estudo e a análise dos dados (factos numéricos) em três fases:

1. Obtenção dos dados
2. Descrição, classificação e apresentação dos dados
3. Conclusões a tirar dos dados.

## Objetivo do Projeto – Prof. André

Tratamento dos dados gerados pelo Servidor de Web a partir do arquivo de logs, de acordo com o roteiro abaixo:

1. Fazer análise do arquivo de log de acesso do servidor apache e identificar os dados dos quais farão o tratamento;
2. Realizar o tratamento dos dados (construção de gráficos, cálculo de média de acessos, comparações e análises na Estatística Inferencial).

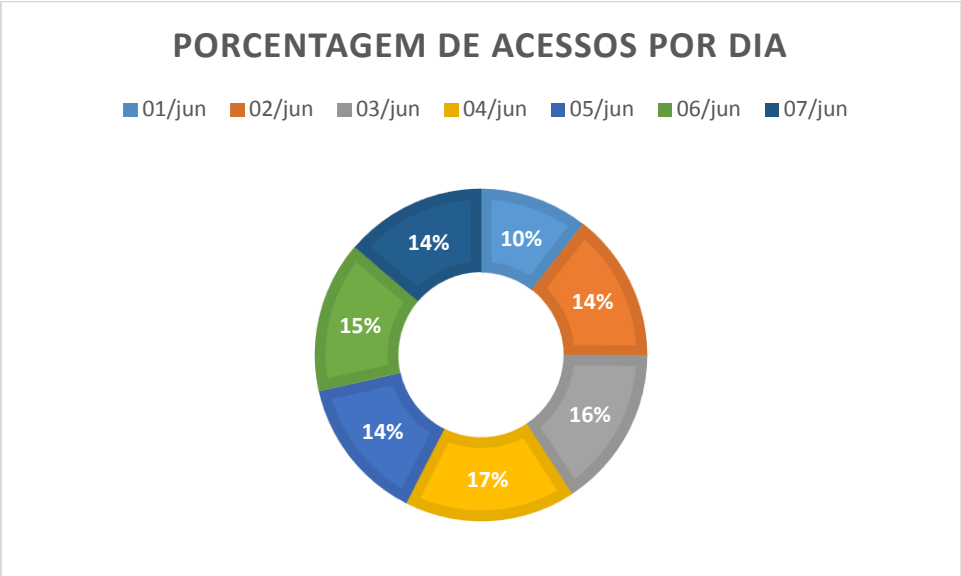
## Análise Estatística:

Analisando o log de acessos, obtemos resultados de um dia de acesso ao site. Temos diferentes acessos a cada dia ao site. Assim obtivemos IP e o número de acesso diário, bem como o total de acessos ao site. Obtivemos os seguintes resultados.

| Dias                  | Qtd de Acesso | % Acesso    |
|-----------------------|---------------|-------------|
| 01/06/2015            | 43            | 10%         |
| 02/06/2015            | 60            | 14%         |
| 03/06/2015            | 64            | 16%         |
| 04/06/2015            | 68            | 17%         |
| 05/06/2015            | 57            | 14%         |
| 06/06/2015            | 60            | 15%         |
| 07/06/2015            | 56            | 14%         |
| <b>Total de cesso</b> | <b>408</b>    | <b>100%</b> |

Com base no total de acessos que foi obtido conseguiremos fazer uma apresentação de dados.

E estimar a porcentagem de acessos de cada dia:



Continuando, podemos definir a média, moda e a mediana.

A média de um conjunto de dados numéricos obtém-se somando os valores de todos os dados e dividindo a soma pelo número de dados.

|       |      |
|-------|------|
| MÉDIA | 58,2 |
|-------|------|

A moda é o valor mais frequente de um conjunto de dados:

|      |    |           |    |    |    |           |    |
|------|----|-----------|----|----|----|-----------|----|
| MODA | 43 | <u>60</u> | 64 | 68 | 57 | <u>60</u> | 56 |
|------|----|-----------|----|----|----|-----------|----|

A mediana, depois de ordenados os valores por ordem crescente ou decrescente, a mediana é o valor que ocupa a posição central, se a quantidade desses valores for ímpar, ou média dos dois valores centrais, se a quantidade desses valores for par.

|         |    |    |    |           |    |    |    |
|---------|----|----|----|-----------|----|----|----|
| MEDIANA | 43 | 56 | 57 | <u>60</u> | 60 | 64 | 68 |
|---------|----|----|----|-----------|----|----|----|

As medidas de dispersão verificam o grau de afastamento dos elementos em relação à média. Assim temos a variância, desvio padrão, coeficiente de variação:

Quanto menor é a variância, mais próximos os valores estão da média; mas quanto maior ela é, mais os valores estão distantes da média.

|                  |       |
|------------------|-------|
| <b>VARIÂNCIA</b> | 53,29 |
|------------------|-------|

O desvio padrão representa a distância média que cada elemento se encontra em relação à média.

|                      |      |
|----------------------|------|
| <b>DESVIO PADRÃO</b> | 7,30 |
|----------------------|------|

O coeficiente da variação verifica de forma relativa o grau de dispersão dos elementos em torno da média.

Existe uma regra básica, se o coeficiente de variação for igual ou menor que 0,15%, indica uma baixa dispersão, esse resultado seria bom, pois seriam mais lineares, entre 0,15% e 0,30%, teriam uma média dispersão, porém se o coeficiente de variação, for maior que 0,30%, isso representa uma alta dispersão, ou seja que os dados analisados acima não possuem uma uniformidade, e não é muito confiável.

Como podemos ver temos uma baixa dispersão em torno da média.

|                                |       |
|--------------------------------|-------|
| <b>COEFICIENTE DE VARIAÇÃO</b> | 0,12% |
|--------------------------------|-------|

Observamos que existe um equilíbrio de acessos durante os dias acompanhados. A variação está próxima da média, o desvio padrão mostra a distância que cada elemento se encontra da média, sendo assim um dado confiável. O coeficiente de variação indicou uma baixa dispersão, ou seja, mostrando ser mais linear com relação aos elementos da média.

### Exemplo:

Na amostra, analisamos o log de acessos e percebemos no dia 04/05 houve 68 acessos/dia e no geral teve uma média de 58,2 acessos/dia, e o mesmo possui desvio de 7,30 com relação à média geral. Sabendo disso, qual a probabilidade desse dia ter mais de 68 acessos num dia qualquer. Será demonstrar abaixo:

Formula:  $Z = \frac{X - M}{D}$

$D$

$$P(X > 68) = Z = \frac{68 - 58,2}{7,30} = 1,34 = 0,4099 \quad Z = 0,5 - 0,4099 = 0,0901$$

$$P(X > 68) = 9,01\%$$

