Ulisses Cotta Cavalca

Uso de ferramentas estatísticas na gerência de performance de redes de computadores

Monografia de Pós-Graduação "*Lato Sensu*" apresentada ao Departamento de Ciência da Computação para obtenção do título de Especialista em "Administração em Redes Linux"

Orientador
Prof. Joaquim Quinteiro Uchôa

Lavras
Minas Gerais - Brasil
2011

Ulisses Cotta Cavalca

Uso de ferramentas estatísticas na gerência de performance de redes de computadores

Monografia de Pós-Graduação "*Lato Sensu*" apresentada ao Departamento de Ciência da Computação para obtenção do título de Especialista em "Administração em Redes Linux"

Aprovada em *30 de abril de 2011*

Prof. Sanderson Lincohn Gonzaga de Oliveira

Prof. Eric Fernandes de Mello Araújo

Prof. Joaquim Quinteiro Uchôa
(Orientador)

Lavras
Minas Gerais - Brasil
2011

À Viviane por todo o amor e companheirismo, e aos meus pais.

Agradecimentos

Agradeço ao CEFET-MG, por disponibilizar o ambiente e recursos necessários para a realização desse trabalho.

Aos amigos do Centro de Computação Científica, pelo compartilhamento de informações sobre gerência de redes na plataforma GNU/-Linux.

Aos colegas do curso de "Administração de Redes Linux", por toda a caminhada nesse curso de pós-graduação.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Com o crescimento significativo das redes de computadores é comum depararmos com ambientes cada vez mais heterogêneos, quanto à diversidade das formas de acessos e serviços disponíveis. Consequentemente, gerenciar essas estruturas sob o ponto de vista de performance tem sido o desafio cada vez maior para administradores de redes. Nesse contexto que o presente trabalho tem como proposta empregar ferramentas estatísticas para auxiliar na gestão de desempenho de redes de computadores. Em uma pesquisa multidisciplinar, com a abordagem do modelo FCAPS de gerência, protocolo SNMP, ciclo PDCA para planejamento de experimentos e técnicas estatísticas, o trabalho visa obter conclusões estatisticamente confiáveis perante análise descritiva de performance comumente feita nos ambientes de rede. Essas inferências contemplam a estimação de parâmetros de rede, análise de variância, problemas de correlação e regressão. No contexto desse trabalho serão analisados, de maneira objetiva, a carga de processamento, uso de memória, *throughput* e número de pacotes vazantes na infraestrutura do CEFET-MG - Campus II.

Palavras-Chave: Gerência de redes; Estatísticas; *Software* livre.

Capítulo 1

Introdução

1.1 Considerações iniciais

O crescente uso dos recursos de Tecnologia da Informação (TI), em especial o acesso à internet, é visível recentemente na sociedade sob vários aspectos. Inicialmente onde havia uma finalidade especificamente acadêmica, hoje temos aplicações envolvendo comércio eletrônico, governo eletrônico, educação a distância, entretenimento, *marketing*, dentre outros. Embora a sua utilização tenha crescido nos últimos anos, é pertinente ressaltar que em 2009 apenas 27% dos domicílios brasileiros possuem acesso à internet, como apontado na CETIC 2010 (CENTRO DE ESTUDOS SOBRE AS TECNOLOGIAS DA INFORMAÇÃO E DA COMUNICAÇÃO, 2010). No âmbito corporativo, a pesquisa CETIC 2010 revela ainda um aumento significativo no percentual de acesso através de redes sem fio. Isso sugere que as redes estejam mais heterogêneas, sob o ponto de vista da conectividade.

Para a democratização do acesso à internet no Brasil, o Plano Nacional de Banda Larga (PNBL) visa atender até 2014, 88% da população com acesso à rede mundial por conexão banda larga, Santos (2010). Adicionalmente, o PNBL estima uma redução em 70% do custo médio atualmente cobrado por este serviço. Na prática sugere um crescimento considerável da internet brasileira, que atinge domicílios, governos e empresas. Tal universalização implica em mais infraestrutura e mais mão de obra, não só na implantação do PNBL, como também na gerência e manutenção de redes de computadores emergentes e já existentes.

Considerando esse contexto e a crescente aplicação da internet na educação e pesquisa acadêmica, o Centro Federação Tecnológica de Minas Gerais (CEFET-MG)[1], como instituição federal de ensino, será o objeto de estudo do presente trabalho. É composto por 3 campi em Belo Horizonte/MG e mais 7 unidades interioranas, constituído por 34 cursos técnicos profissionalizantes, 13 cursos superiores, 7 programas de mestrado, além de cursos de pós-graduação latosensu e projetos de extensão. Conta com uma estrutura própria de redes de computadores (recursos, serviços e pessoal), ao qual a conexão à internet é provida pela Rede Nacional de Pesquisas (RNP[2]). O Campus II, local de aplicação desta pesquisa, tem uma conexão dedicada total de 6Mbps balanceadas por 3 modens de 2Mbps, além de um *link* de 2Mbps com o campus I que concentra grande parte dos serviços institucionais. Além disso atende a uma demanda hipotética de 650 a 700 máquinas clientes, inclusos dispositivos móveis, distribuídos em 16 pontos de fibra ótica e 4 pontos UTP.

1.2 Motivação

Dentro do que o modelo OSI FCAPS prevê em termos de gerência de rede (gerência de falhas, configuração, contabilização, performance e segurança), o desempenho de uma rede é atualmente verificado por ferramentas como Cacti[3] e Nagios[4]. Elas constroem, respectivamente, um histórico do comportamento de *links* e de servidores, e registros e alertas de qualquer anomalia no ambiente conforme limiares definidos pelos administradores da rede. Existem outras ferramentas de gerência, como MRTG[5], Zabbix[6], e Pandora FMS[7], com o mesmo propósito de funcionamento, porém com suas particularidades que não vem ao caso discuti-las neste trabalho. São úteis na solução de problemas e detecção de eventos significativos na rede, bem como a comprovação do funcionamento correto em casos mais críticos. De qualquer maneira essas ferramentas consistem em um monitoramento descritivo das informações quando analisadas matematicamente.

No entanto, a gerência de rede praticada atualmente na grande maioria dos casos se limita na leitura das informações geradas por essas ferramentas através

[1] CEFET-MG: http://www.cefetmg.br/
[2] RNP: http://www.rnp.br/
[3] Cacti: http://www.cacti.net/
[4] Nagios: http://www.nagios.org/
[5] MRTG: http://oss.oetiker.ch/mrtg/
[6] Zabbix: http://www.zabbix.com/
[7] Pandora FMS: http://pandorafms.org/

de gráficos e tabelas, aliada ao conhecimento teórico e do ambiente que administrador da rede possui. Isso representa, muitas vezes, uma análise subjetiva e até mesmo intuitiva do comportamento da rede. Essa falta de rigor, não só na obtenção de conclusões do funcionamento da rede como na solução de problemas, justifica o emprego de ferramentas estatísticas para a análise de desempenho de uma rede, com o propósito de se obter conclusões mais confiáveis. Concomitantemente, o emprego de técnicas matemáticas para análise de correlação entre as variáveis que relatam o funcionamento de uma rede de computadores resultará em uma abordagem científica na sua gerência.

Podemos citar como benefícios de uma abordagem científica na gerência de redes de computadores:

- a validação e comprovação, sob o ponto de vista matemático, da análise dos dados de um ambiente de rede e suas correlações;
- a obtenção de conclusões, estatisticamente válidas e confiáveis sob uma dada margem de erro, na questão da gerência de falhas e de performance;
- a fundamentação de futuras análises e pesquisas no âmbito de redes de computadores.

O uso desses procedimentos não será essencial ou obrigatório para o monitoramento e solução de problemas. Cabe ao administrador empregar ou não as técnicas estatísticas de maneira complementar às ferramentas de monitoramento atualmente aplicadas. Assim como associar os resultados obtidos ao conhecimento técnico já existente de seu funcionamento.

1.3 Necessidades

Observa-se na rede do CEFET-MG Campus II, o uso de quase 100% da banda disponível nos horários de maior demanda de conexão à internet, representando um congestionamento dos *links*. Consequentemente, o desempenho de alguns serviços institucionais, pesquisas no âmbito acadêmico, e navegação *web* são prejudicados por conta dessa problemática. Não serão considerados nesse momento possíveis gargalos de rede, provenientes de cascateamento sem o devido planejamento ou uso de equipamentos defasados tecnologiacamente, que na prática implica em queda de performance para o usuário final.

É pertinente observar variáveis como uso da CPU e memória do roteador, taxa de descarte de pacotes, fração de erros na entrada e na saída, tempo de resposta, latência, dentre outras. E relacioná-las com uso de técnicas estatísticas como intervalo de confiança, testes de hipóteses, análise de variância e correlação, para compor um diagnóstico do desempenho da rede. Isso auxilia na solução de problemas, conhecimento da rede, e conclusão de alguns questionamentos como:

- Se a lentidão percebida tem alguma relação com o histórico da rede ou não;
- Se existe diferença significativa no balanceamento dos *links*;
- Se a atual largura de banda atende a demanda de acesso à internet, observado nível de significância aplicado.

As análises dessa natureza complementarão as ferramentas de monitoramento Cacti e Nagios, atualmente aplicados pelos administradores da rede do CEFET-MG Campus II.

1.4 Objetivos e metas

O presente trabalho tem os seguintes objetivos gerais:

- Verificar cientificamente, através de técnicas estatísticas, análises e conclusões sobre o desempenho e comportamento de uma rede de computadores;
- Abordar o conceito de qualidade no contexto de administração de redes;
- Oferecer procedimentos e estratégias, de caráter auxiliar, na gerência de redes de computadores sob o ponto de vista estatístico;
- Complementar os recursos de gerência e monitoramento de uma administrador de redes.

O trabalho tem como metas:

- Verificar a correlação entre as variáveis que descrevem o comportamento e funcionamento de uma rede;

- Utilizar técnicas estatísticas como intervalo de confiança, testes de hipóteses e análises de variância na interpretação de dados no monitoramento de uma rede;

- Propor um procedimento para planejamento de um experimento de natureza estatística, além da análise dos resultados no contexto de uma rede de computadores.

1.5 Metodologia

A metodologia deste trabalho será embasada na teoria da qualidade, baseada na gerência e solução de problemas, com o foco em técnicas de planejamento e análise de experimentos. Dessa forma, a etapa inicial consistirá na identificação dos objetos de experimento, definição de variáveis e demais parâmetros que influenciam no estudo. Adicionalmente serão elencados possíveis erros e não conformidades que tendenciem os dados obtidos das amostras.

O próximo passo consiste no planejamento do procedimento experimental, ou seja, a sistematização das coletas das amostras. E a realização do experimento propriamente dito, com dados coletados a partir do protocolo SNMP[8].

Por fim os dados serão analisados de maneira descritiva e de forma que seja possível inferir estatisticamente, para que a interpretação dos resultados e conclusões sejam inseridos no contexto de gerência de redes de computadores.

1.6 Estrutura do trabalho

No capítulo 2 será levantado um referencial teórico sobre conhecimentos chaves para a elaboração deste trabalho. Inicialmente será abordado a questão da gerência de redes sob o ponto de vista teório, com sistemas e padrões empregados. Uma breve discussão sobre questões de qualidade aplicadas à TI, de forma a embasar o planejamento e a análise de experimentos. Em seguida será abordado o funcionamento do protocolo SNMP para aquisição de dados. E por fim uma revisão das técnicas estatísticas, como intervalo de confiança, testes de hipóteses, análise de variância e correlação linear.

[8]SNMP: `http://net-snmp.sourceforge.net/`

O capítulo 3 será a descrição de toda a metodologia e desenvolvimento do trabalho. É a definição do roteiro do experimento propriamente dito, além da descrição do seu desenvolvimento.

A pesquisa terá seus resultados, com as respectivas análises, apresentados no capítulo 4. Em um primeiro momento será feita a exibição dos dados de maneira descritiva em formato de gráficos e tabelas, seguidos pela aplicação das ferramentas estatísticas para interpretação dos resultados e embasamento da conclusão da pesquisa.

E o capítulo 5 trará a conclusão do trabalho, trazendo não só a inferência de toda a pesquisa no contexto de redes de computadores, como também sugestões para pesquisas futuras.

Capítulo 2

Revisão bibliográfica

2.1 Gerência de redes

Redes de computadores podem ser entendidas como a conexão de computadores e equipamentos, de modo a compartilhar serviços e informações. O conceito de internet se encaixa quando falamos de várias redes distintas, separadas geograficamente, conectadas entre si. Para que tal conexão venha de fato acontecer é preciso uma série de equipamentos ativos de redes, como *switches*, roteadores, *modems*, além de computadores no papel de servidores e clientes.

Nessa ótica, a gerência de uma rede de computadores é a tarefa de garantir ao usuário a troca de serviços e informações de maneira satisfatória. Gerência de redes, como definido por Goupta (2006), é o emprego de uma variedade de ferramentas, aplicações e dispositivos para auxiliar administradores no monitoramento e manutenção de redes. Adicionalmente Udupa[1], citado por Narang e Mittal (2000), apresenta a gerência de redes como o monitoramento e controle de recursos, conexão e comunicação de computadores e suas aplicações utilizadas.

A expansão das redes de computadores na decáda de 80 exigiu que algum modelo de gerência fosse criado. Como na ocasião não havia nenhuma estratégia clara de gestão de redes, o seu crescimento além de ser de forma não sistemática, afetava o funcionamento dos segmentos já em operação. Além disso, a manutenção tornava-se árdua e custosa sob o ponto de vista da produtividade.

[1] Udupa, Divakara K., Network Management System Essentials, McGraw-Hill, U.S.A., 1996.

A ITU-T[2] elaborou um conjunto de recomendações e práticas para a gerência de equipamentos de rede e de telecomunicações. O TMN (*Telecommunications Management Network*[3]), contemplado pela série M.3000 da ITU-T, consiste em um modelo genérico de rede que considera diferentes formas de tecnologia em diversos níveis de abragência. Isso sugere a gestão de uma rede com cabeamento estruturado, conexão sem fio, rede local virtual, em alcance local (LAN), amplo (WAN) ou metropolitano (MAN). O TMN dispõe de 4 camadas de aplicação, descritas a seguir, conforme Goyal, Mikkilineni e Ganti (2009).

- **Gerência de negócios** (*Business Management* - BML): relaciona aspectos relacionados a negócios, tendências, e governança de uma maneira geral;
- **Gerência de serviços** (*Service Management* - SML): relaciona funcionalidades de serviços, definições e administração;
- **Gerência de rede** (*Network Management* - NML): realiza a distribuição de recursos da rede, com a devida definição, controle e supervisão;
- **Gerência de elementos** (*Element Management* - EML): reune elementos individuais da rede como alarmes, backups, logs e manutenção de hardware.

O TMN apresenta 5 funcionalidades, dentre elas Falhas (***F**ault*), Configuração (***C**onfiguration*), Contabilização (***A**ccounting*), Performance (***P**erformance*) e Segurança (***S**ecurity*), ou simplesmente FCAPS que serão discutidas adiante. Segundo Santos (2004), estas funcionalidades foram integradas pela ISO como parte da especificação do modelo OSI. Além disso, o FCAPS pode ser empregado na gerência de cada uma das 4 camadas de aplicação definidas pela TMN, no âmbito da governança de TI conforme citadas anteriormente. Essa relação pode ser visualizada na Figura 2.1

Boutaba e Polyrakis (2001) discutem a gerência de redes de forma distribuída, aplicado ao contexto das funcionalidades do FCAPS. O emprego de agente móveis para aquisições de informações na rede faz-se necessário, introduzindo dessa forma o uso do protocolo SNMP.

O contexto desse trabalho explora as funcionalidades de falhas e performance, aplicado à gerência de redes na definição do modelo TMN.

[2]ITU-T: `http://www.itu.int`

[3]Conjunto de recomendações da ITU-T (série M.3000) para gerência da interconectividade e comunicação entre sistemas operacionais heterogêneos, e da comunicação entre redes.

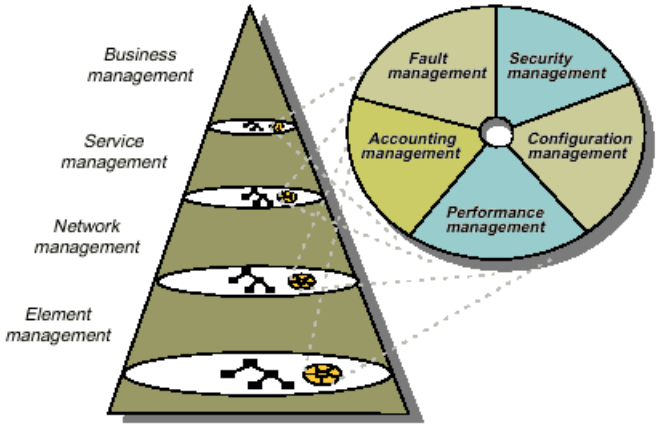

Figura 2.1: Relação entre modelo TMN e funcionalidades FCAPS. Fonte: (JAVVIN TECHNOLOGIES, 2010)

2.1.1 Gerência de falhas

A funcionalidade gerência de falhas atende pela detecção, alarme, análise e registro de falhas de serviços, anormalidades, ou eventos significativos.. Tais componentes compõem um conjunto de artifícios para a solução de problemas. Quando detectada uma falha na rede, Santos (2004) sugere a adoção dos seguintes passos:

1. **Localização** da falha;
2. **Isolamento** do problema;
3. **Reconfiguração** ou modificação de forma a minimizar o impacto causado;
4. **Correção** da falha.

Destaca-se a importância de identificação do impacto que a falha ocasionará no processo de sua localização. Isso faz com que o isolamento seja feito de maneira efetiva, onde apenas o segmento atingido será tratado e não a rede como um todo. Além disso, o uso do esquema de *bypass* (desvio) facilita o isolamento de partes da rede, de forma que outros segmentos não sofram os impactos da falha gerada.

Goyal, Mikkilineni e Ganti (2009) advertem que a gerência de falhas consiste também na interpretação de determinados alarmes e mensagens gerados, assim como os registros de *logs*. Uma situação seria a de erros transitórios ou erros

persistentes, que são anomalias ocorrentes e registradas em *log* porém não comunicados por um alarme justamente por não excederem a um *threshold* (limiar). Supondo que esse limiar seja atingido e a falha comunicada, no contexto de erros transitórios, o próprio sistema o corrige automaticamente. Exemplo seria uma perda momentânea de pacotes numa conexão ou uma alta latência, devido a uma problema no meio físico de conexão.

Outra interpretação seria em situações que as falhas geram algum tipo de efeito cascata. Mesmo que a causa raiz de uma falha F_0 de um recurso R_0 seja comunicada, serviços ou recursos $R_1, \ldots, R_n$ dependentes a este também sofrerão algum tipo de impacto e falhas $F_1, \ldots, F_n$, que serão registradas. A solução de problemas dessa natureza será conduzida em função de prioridades, onde anomalias consequentes não seriam tratadas em um primeiro momento, e sim a falha raiz. Para tal metodologia é necessário o conhecimento e a determinação da relação e interação dos serviços e recursos existentes.

Santos (2004) relaciona 3 problemas que podem dificultar no processo de identificação de uma falha.

- falhas não observadas, por conta de sua dificuldade. Por exemplo, processo em estado *zombie*, ou *deadlock* entre dois processos;
- falhas parcialmente observadas, onde as informações coletadas não são suficientes para a interpretação dos problemas;
- observação não confiável, decorrente de métricas inexatas ou incoerentes, além de outras informações incoerentes com a infraestrutura em questão.

Levando em consideração o tamanho da rede, os tipos de serviços nela existente e as falhas a serem verificadas, o uso de ferramentas automatizadas torna-se viável para a gerência de falhas, sob o ponto de vista da agilidade e praticidade. Nesse contexto citamos o Nagios[4] como um exemplo de serviço que monitora alguns parâmetros de rede e de servidores, ao qual registra os eventos que excedem um limiar aceitável e emite alarmes aos administradores de rede.

2.1.2 Gerência de configuração

A gerência de configuração é a organização que compõem o conjunto de configurações de rede, recursos e sistema. Segundo Goyal, Mikkilineni e Ganti (2009)

[4]Nagios: `http://www.nagios.org/`

essa gerência padroniza a ativação e a desativação de serviços ou segmentos de redes, em situações planejadas ou de emergência. Além disso existe o desafio de não só manter as informações atualizadas como também rastrear as modificações realizadas. Tal requisito pode ser provido por ferramentas e métodos automáticos de descoberta de serviços, mapeando continuamente configurações, componentes e suas dependências. Como exemplo dessas ferramentas, podemos citar o CACIC[5] para gerência do parque computacional de uma instituição.

Das atividades realizadas dentro do contexto da gerência de configurações, Santos (2004) destaca:

- a identificação dos elementos funcionais da infraestrutura;
- mapas de topologias, tanto físicas como lógicas;
- inventário de *harware* e *software*, deste último observando-se a questão das licenças de uso;
- base de dados, de acesso fácil e compartilhado, para disponibilização dessas informações;
- mecanismo para gerência de alteração de configuração, para que estas informações se mantenham de forma íntegra.

2.1.3 Gerência de contabilização

De uma maneira ampla, gerência de contabilização é o mecanismo de acompanhamento de como os usuários utilizam os recursos providos pela infraestrutura em questão, de acordo com as finalidades a que se aplicam. Leinwand e Conroy (1996) afirmam que informações oriundas desse tipo de gerência são úteis no processo de alocação de recursos de rede, como capacidade de armazenamento e processamento, e políticas de *backup*. Adicionalmente, a gerência de contabilização fundamenta a expansão e a configuração de redes de computadores de forma mais produtiva.

Assim como as outras funcionalidades, a gerência de contabilização precisa ser devidamente registrada, observada a devida integridade e importância das informações. A gerência de contabilização tem como metas:

[5]CACIC: `http://www.softwarepublico.gov.br/ver-comunidade?community_id=3585`, acessado em 21.ago.2010

- verificar uso excessivo de alguns recursos, por exemplo quotas de usuários, que possam comprometer o seu funcionamento;
- verificar uso abusivo, por exemplo tentativas de invasão ou fraudes de dados, que comprometam a integridade do sistema e da infraestrutura;
- subsidiar algum processo de cobrança de serviços, se cabível, como o número de acesso em uma página ou capacidade de armazenamento para caso de serviços de hospedagem, ou mesmo a quantidade de informação (pacotes por segundos) copiados tratando-se de provedores de acesso;
- fundamentar processos de auditoria que venham ocorrer, devido a uma violação de direitos ou mesmo verificação da eficiência da estrutura em funcionamento.

2.1.4 Gerência de performance

De todas as funcionalidades do FCAPS, a gerência de performance é a que mais se adequa ao estudo matemático e estatístico devido ao seu aspecto quantitativo. É responsável em prover dados sobre o desempenho da rede revelando sua qualidade, além de embasar análises de tendências de alguns aspectos e seu comportamento com um todo. Além disso é empregada em conjunto com a gerência de falhas, pois valores máximos aceitáveis de alguns parâmetros só serão verificados com a existência de coleta de dados.

Atualmente na maioria das redes de computadores, a coleta de dados em ativos de redes como roteadores e *switches*, e em *hosts*, é feita de maneira distribuída através do protocolo SNMP. Rodrigues (2009) relaciona alguns tipos de métricas aplicadas à gerência de performance:

interfaces de rede: taxa de utilização de interfaces por protocolo, quantidade de *bits* transmitidos e recebidos, quantidade de pacotes chaveados por segundo, taxa de pacotes com erros e descartados, MTU, número de colisões;

hosts: consumo de memória, utilização da CPU, carga média de processamento, uso de partições físicas e lógicas, quantidade de usuários no sistema;

roteadores: utilização da CPU, carga média de processamento, quantidade de memória livre, *swap*, disponibilidade, taxa de pacotes com erros ou descartados na entrada e na saída, quantidade de bits transmitidos e recebidos, quantidade de pacotes chaveados por segundo;

Tais dados permitem que o administrador encontre alguns indicadores que representam o desempenho da rede: latência, *jitter*, vazão, *throughput*, disponibilidade, carga de processamento, dentre outros. A análise desses indicadores em conjunto com outras informações da rede posibilita:

- verificar a capacidade da rede;
- planejar a expansão, tanto no aumento da largura de banda como no incremento de máquinas clientes;
- identificar gargalos na rede, ou por sua topologia física, ou por configuração lógica;
- dimensionar, de forma satisfatória, os recursos, equipamentos ativos de redes, e máquinas servidoras;
- avaliar os impactos dos indicadores de desempenho de rede;
- gerenciar, de maneira mais eficaz, possíveis congestionamentos;
- garantir a Qualidade de Serviços (QoS) aos usuários da rede, através de medidas como restrição à recursos e controle de tráfego.

Na gestão de redes de computadores, Tanenbaum (1997) alerta sobre alguns erros típicos sobre a análise de desempenho. Por exemplo, o monitoramento da vazão de uma rede será diferente em dias que algum segmento está realizando uma videoconferência, ou se um servidor está executando *backup* para outro ponto da rede. Outro erro típico é a realização de testes que não representam o problema em estudo. No caso de congestionamento, a coleta de dados para avaliar sua performance não terá nenhuma representatividade se ocorrer em horários que a rede não está operando em seu limite.

Para que a gerência de desempenho ocorra de forma sistemática e organizada, o uso de *software* para esse fim é amplamente adotado por administradores de redes. Ferramentas que fazem o acompanhamento da performance como o Cacti[6], Pandora FMS[7] e Zabbix[8], são viáveis sob o aspecto da gerência distribuída, além de garantir a integridade, disponibilidade e fácil acesso por meio de interfaces gráficos. Consequentemente, o acompanhamento do comportamento da rede pode ser feita quase que em tempo real, salvo as limitações das ferramentas como tempo

[6]Cacti: http://www.cacti.net/
[7]Pandora FMS: http://pandorafms.org/
[8]Zabbix: http://www.zabbix.com/

de aquisição de dados, acesso a banco, dentre outras. E até mesmo a eficácia de uma eventual modificação na configuração, ou na estrutura da rede, pode ser verificada a partir da comparação dos dados.

2.1.5 Gerência de segurança

Gerência de segurança provê a defesa em vários níveis para controle de acesso e utilização de serviços, mantendo a privacidade, confidencialidade e integridade das informações. É projetada para proteger serviços e dados que a estrutura comporta, prevenir contra códigos maliciosos, negligências e comportamentos abusivos de usuários autorizados ou não. Viabiliza o emprego efetivo de uma Política de Segurança, complementado por níveis de privilégios, *logs* de acesso, estratégias para auditorias, e alarmes de segurança (GOYAL; MIKKILINENI; GANTI, 2009). Esses alarmes são controlados por sistema de detecção de intrusos a nível de *host* (*Host Intrusion Detection System* - HIDS) e a nível de rede (*Network Intrusion Detection System* - NIDS).

As informações trafegadas, ou armazenadas, na rede de computadores deverão ter sua criticidade e sensibilidade muito bem definida. Isso facilita o projeto de sistemas de alta disponibilidade (uso de mecanismos de redundância), *backups*, e *firewalls*, tudo associado a uma Política de Segurança aprovada pela alta administração e amplamente divulgada entre os usuários atingidos por ela. É válido ressaltar também a importância da existência de um Plano de Continuidade de Negócios (PCN). Magalhães e Pinheiro (2007) definem o PCN como "regras bem-detalhadas, assim como responsabilidades, equipes e procedimentos relacionados com a recuperação do ambiente informatizado após a ocorrência de um desastre".

Na prática, o uso de certificados SSL[9] em servidores de páginas, autenticação, correio eletrônico, dentre outros serviços, elevam o nível de segurança da rede, consequentemente de todo o ambiente. Além disso é importante padronizar os algoritmos de criptografia a serem utilizados em toda a estrutura de rede. Atualmente, o conceito de Infraestrutura de Chave Pública (ICP) vem sendo implementado em instituições para que os três princípios básico da segurança sejam alcançados: confidencialidade, integridade e não repúdio (ESR/RNP, 2010). Uma ICP introduz o conceito hieráquico de Autoridades Certificadoras (AC) para uso de certificados digitais. Dessa forma, serviços de uma maneira geral poderão ter sua autenticidade garantida por meio de assinaturas digitais.

[9] SSL: Secure Socket Layer

2.2 Qualidade em TI

Atualmente no contexto da Tecnologia da Informação, existem amplas práticas de governança que relacionam ações estratégicas a níveis de negócios e gestão organizacional. Tais práticas, como exemplo o ITIL[10] e o COBIT[11], são voltadas para gestão de métodos, problemas, expansão, continuidade, dentre outros requisitos que garantam a qualidade e melhoria contínua dos serviços de TI.

O COBIT traz alguns métodos em sua estrutura de funcionamento que atendem as demandas de gerenciamento, controle, e medidas em TI, segundo Laurindo (2008). Desses métodos vale destacar elementos de medidas de desempenho, ao qual mais se aproximam às finalidades desse trabalho. Já o ITIL trata todos os recursos sob o formato de serviços ao qual também prevê, dentre outras funcionalidades, a gestão de problemas. Diferentemente do TMN e das funcionalidades FCAPS discutidos anteriormente que tratam da gestão direta de equipamentos de redes e de telecomunicações, o COBIT e o ITIL estão em patamares mais altos de gerência e de planejamento dos recursos de TI e dos negócios aos quais estão inseridos.

O gerenciamento de problemas é uma prática adota em várias áreas, não exclusivamente em TI. Segundo Magalhães e Pinheiro (2007), a busca contínua de causas e soluções para problemas reais ou possíveis garante a essência do conceito da melhoria contínua em serviços da instituição. Com o propósito de direcionar a solução de maneira mais efetiva foram desenvolvidos métodos para a gestão de problemas, tendo como característica a sua finalidade (generalista ou específica) e proteção (por direitos autorais e de domínio público). Desses métodos, relacionados em (MAGALHÃES; PINHEIRO, 2007), podemos citar: método científico, metodologia 5s, Método Análise e Solução de Problemas (MASP), Controle Estatístico do Processo, PDCA, dentre outros.

Todos os métodos para gerência de problemas estão contidos em um conceito maior de Gestão da Qualidade e melhoria contínua, assim como as ferramentas da qualidade tais como: estratificação, folha de coleta de dados, diagrama de pareto, histograma, diagrama de dispersão, carta de controle e diagrama de causa e efeito. Neste trabalho não serão discutidas a proposta, funcionamento e aplicação de métodos de gerência de problemas. Será abordado o ciclo PDCA para embasar a realização do experimento estatístico, (WERKEMA; AGUIAR, 1996), no âm-

[10] ITIL: http://www.itil-officialsite.com/
[11] COBIT: http://www.isaca.org/Knowledge-Center/COBIT

bito da gerência de redes sob o aspecto de desempenho, que auxilia na solução de problemas com foco na gestão da qualidade e melhoria contínua.

2.2.1 Ciclo PDCA

O PDCA é um método de solução de problemas aplicável em processos de melhorias contínuas. Seu funcionamento basea-se em um ciclo, com o objetivo de gerenciar problemas reais ou possíveis, e manter a qualidade alcançada em processos anteriores. Por se tratar de um ciclo pode ser aplicado continuamente, de maneira que se obtenha o máximo de performance. Consiste em: Planejar (***P****lan*), Executar (***D****o*), Verificar (***C****heck*) e Agir (***A****ction*). A Tabela 2.1 relaciona as 4 fases do ciclo PDCA e os 8 passos para a sua realização, bem como uma breve descrição de cada um.

Tabela 2.1: Descrição das fases e etapas do ciclo PDCA, segundo Qing-Ling *et al.* (2008)

Fase	Passos	Descrição
Planejar (*Plan*)	1	Análise das condições atuais e definição dos problemas existentes;
	2	Descrição das variáveis causadoras dos problemas existentes;
	3	Identificação dos fatores de maior relevância que influenciam no problema identificado;
	4	Elaboração de plano de trabalho para aplicação da solução do problema, de acordo comas condições atuais e fatores descritos;
Executar (*Do*)	5	Mensuração e aplicação do plano de trabalho proposto na etapa anterior;
Verificar (*Check*)	6	Verificação da implementação feita de acordo com o plano de trabalho;
Agir (*Action*)	7	Obter conclusões, sumarizar as experiências, e realizar o registro do trabalho feito;
	8	Elencar problemas que não puderam ser resolvidos e apontar a continuidade do trabalho, de forma a se obter a melhoria contínua.

Segundo Qing-Ling *et al.* (2008), o ciclo PDCA se aplica à variados tipos de atividades e diversos níveis de gestão, desde procedimentos operacionais à tarefas envolvendo governança. Embora a literatura referencie em sua maioria exemplos

de planejamento de experimentos e emprego das técnicas estatísticas relacionados à engenharia de produção, a proposta dessa metodologia é multidisciplinar e transversal.

2.2.2 Planejamento de experimentos

Baseado nas fases e etapas do PDCA, Werkema e Aguiar (1996) propõem 8 itens para a construção de roteiro para a realização de um experimento estatístico, no contexto da solução de problemas discutido anteriormente.

1. **Identificação dos objetivos**: Nessa etapa inicial é importante fazer uma coleção de idéias, colocações e hipóteses sobre o objeto em estudo. Em seguida definir claramente os objetivos do experimento, com base no conhecimento disponível sobre o problema, e elencar as principais informações quantitativas, que descrevam o problema atual;

2. **Seleção da variável resposta**: Essa etapa consiste na seleção da variável que irá representar o problema em estudo. Vale ressaltar que mais variáveis podem ser definidas, dependendo da abrangência do experimento. E por fim determinar alguns parâmetros como o método de medição da variável resposta e sua escala (exemplo, linear ou exponencial ou logarítmica);

3. **Escolha de fatores**: Esse momento consiste na identificação dos fatores que de alguma forma influenciam não só no objeto em estudo, como também na realização do experimento. É de grande importância o uso de conhecimentos técnicos no estudo e não meramente estatísticos, para que se determine os fatores com níveis variáveis, fatores constantes, e fatores independentes que não podem ser controlados. Além disso, identificar as faixas de variação, os níveis desejados e o mecanismo de medição no contexto do experimento;

4. **Planejamento do procedimento experimental**: O planejamento do procedimento é a fase mais crítica e elaborada de todo o experimento. Compreende aspectos como ações que minimizem, ou se possível eliminem, a influência de fatores não controláveis, relação entre os fatores, proposta de modelo matemático para o experimento, e determinação do tamanho da amostra para que os dados sejam devidamente representados. É válido elaborar uma sequência de trabalho a ser adotada durante a realização do experimento, reduzindo as chances de ocorrência de erros. Em função do tamanho do experimento e de sua abragência, essa etapa descreve orçamentos, cronogramas e recursos necessários para a pesquisa;

5. **Realização do experimento**: Consiste na execução do planejamento do experimento, desde que haja o cuidado de monitorar e registrar as informações que possam representar algum viés na análise de dados, ou alguma importância significativa na interpretação dos resultados;

6. **Análise e tratamento de dados**: Inicia-se a etapa de análise de dados com uma revisão do que foi coletado, com o propósito de averiguar possíveis erros ou omissões. Em seguida fazer uso da estatística descritiva, como gráficos, tabelas e diagramas, para visualização dos resultados. E por fim o emprego do modelo matemático definido na etapa de planejamento do experimento, com o objetivo de embasar a interpretação dos resultados;

7. **Interpretação dos resultados**: Com posse dos dados coletados durante o experimento, sua visualização de forma descritiva e seu tratamento conforme um modelo matemático é que se pode estabeler conclusões do experimento. Nessa etapa se faz necessário o conhecimento específico sobre o tema que a pesquisa se aplica, para que se possa avaliar a significância dos resultados e as probabilidades associadas no seu contexto. É importante registrar as limitações tantos dos dados coletados como do método utilizado na interpretação dos resultados;

8. **Elaboração de relatório**: A finalização da pesquisa consiste na elaboração do relatório final, que contempla a descrição e detalhamento do experimento como um todo. Deve-se ter a atenção na inserção das informações, tabelas e gráficos, de maneira suficiente a verificar os resultados e sua relação com a conclusão do experimento. Recomenda-se minimizar o uso de termos estatísticos que carregam a leitura, bem como o emprego de linguagem simples, além de descrever recomendações a partir das conclusões obtidas.

2.3 Protocolo SNMP

Durante o uso da ARPANET, a gerência de uma rede de computadores era feita de maneira básica e simplista, onde o objetivo da gerência consistia basicamente em verificar se um *host* estava ativo ou não na rede. Essa verificação era garantida pelo programa *ping*, que por sua vez possue funcionamento baseado no protocolo ICMP[12] (TANENBAUM, 1997). Como as redes de computadores tiveram sua expansão e sua complexidade aumentada, sua gerência precisou ser mais elaborada e eficiente com a aquisição de dados que melhor retrasse o comportamento da

[12]ICMP: Internet Control Message Protocol

rede. Dessa forma, em maio de 1990, foi publicada a RFC 1157 (CASE *et al.*, 1990) que definia o funcionamento da versão 1 do *Simple Network Management Protocol* (SNMP). O protocolo SNMP possibilita a aquisição de um conjunto de informações sobre equipamentos ativos de redes e servidores de forma sistêmica, incrementando de maneira significativa o processo de gerência de redes. O funcionamento do SNMP v1 (versão 1) basea-se na RFC 1155 (ROSE; MCCLOGHRIE, 1990) que define o funcionamento da estrutura de gerenciamento da informação (*Structure of Management Information* - SMI).

A RFC 1441 (CASE *et al.*, 1993b) e RFC 1452 (CASE *et al.*, 1993a) implementam a versão 2 do protocolo SNMP, onde respectivamente descrevem as novas funcionalidades da versão 2 e a coexistência entre as duas versões. A principal diferença consiste na versão 2 propor um *framework* de gerência de rede padrão aplicado à internet, trazendo além disso um conjunto de melhorias inclusive a operação com protocolos adicionais. Um dos *framework* que podemos citar é o SNMP v2c (CASE *et al.*, 1996), onde as mensagens enviadas pelo protocolo a partir do equipamento monitorado são associadas a uma comunidade. Isso permite que a configuração do protocolo possa ser segmentada, no âmbito do seu modelo de implementação, facilitando a organização e a divisão dos equipamentos gerenciados.

Algumas questões de segurança envolvendo o SNMP, das quais comprometem a integridade dos sistemas e da estrutura da rede, começam a entrar em foco. Dessas questões podemos citar a má configuração ou configuração insufuciente do protocolo, de maneira que informações de servidores e ativos de redes sejam obtidas por usuários indevidos. Nesse contexto, a RFC 2572 (CASE *et al.*, 1999) propõe a terceira versão do protocolo para processamento e expedição de mensagens SNMP. O SNMP v3 suporta não só criptografia para comunicação das informações como também mecanismos de autenticação entre agentes e gerentes.

2.3.1 Modelos de implementação

Segundo Tanenbaum (1997), o modelo de gerência do SNMP consiste de quatro componentes:

- **gerente**: pontos na estrutura da rede que se dedica no gerenciamento do protocolo;
- **ponto gerenciado (ou agente)**: na estrutura da rede seria o nó, ou estação, a ser gerenciado;

- **informações de gerenciamento**: consiste nos dados e métricas que representam o comportamento dos pontos gerenciados;

- **protocolo de gerenciamento**: uso na prática de um protocolo que viabiliza o envio das informações de gerenciamento, bem como a aquisição de cada uma delas e a gestão por parte do gerente;

Em termos de implementação do SNMP numa rede de computadores, podemos citar dois modelos de funcionamento: agente e gerente; e *trap*.

No modelo *agente e gerente* são disponibilizados na rede pontos que fazem o papel de gerente do protocolo, ao passo que é determinado quais agentes serão monitorados pelo protocolo SNMP. Seja uma rede de computadores com um gerente SNMP G_1, que controla pelo protocolo os agentes $A_1, A_2, \ldots, A_n$. G_1 pode enviar uma solicitação das informações de seu desempenho à A_1, onde este responde com os dados requisitados ao gerente pertinente. O processo pode se repetir entre G_1 até o agente A_n. Dependendo de como o SNMP estiver configurado entre o gerente e os agentes é possível que um gerente G_2, na instância de um usuário mal intensionado, consiga enviar uma requisição à A_1 e obter os dados que o protocolo gerencia. Tal falha de segurança é que o SNMP versão 3 pretende sanar através de mecanismo de autenticação. Aplicativos de gerência de rede, como o Cacti, tem em sua estrutura de funcionamento o gerente SNMP automatizado para enviar solicitação e obter dados do equipamento monitorado em um determinado ciclo de tempo. A Figura 2.2 ilustra o modelo de implementação *agente e gerente*.

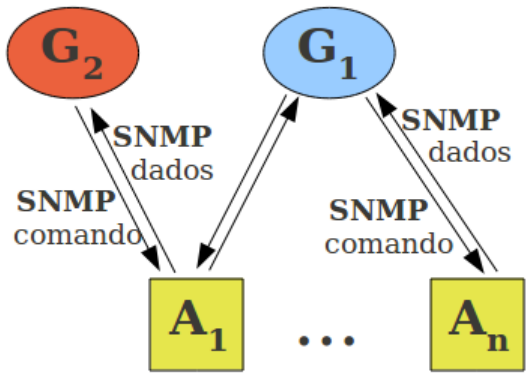

Figura 2.2: Diagrama do modelo de funcionamento do SNMP como agente e gerente

Seja uma rede com um gerente SNMP G_1 e um conjunto de agentes $A_1, \ldots, A_n$, o modelo *trap* permite que o agente envie dados apenas para o gerente que estiver

configurado. Adicionalmente, as informações são enviadas somente se um evento significativo ocorrer, tal como queda de um serviço ou estouro de um limiar para parâmetro de rede. Isso impede, pelo menos teoricamente, que um agente A_n envie dados em função de algum evento significativo para um gerente G_2. Nesse caso, um agente $A_n + 1$ enviaria dados para um gerente G_2 se sua configuração prever essa situação. O modelo *trap* pode ser visualizado na Figura 2.3.

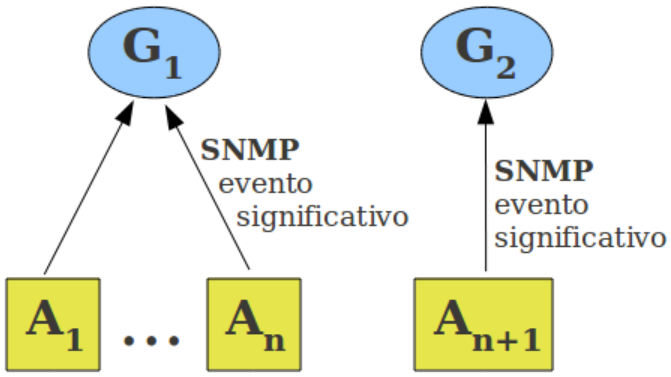

Figura 2.3: Diagrama do modelo de funcionamento do SNMP como trap

Vale ressaltar que cada ponto gerenciado irá se comportar como agente ou *trap* conforme sua configuração.

2.3.2 Organização de dados do SNMP

No contexto do protocolo SNMP, os valores das variáveis que descrevem o comportamento de um ativo de rede, *host*, ou qualquer outro equipamento monitorado pelo protocolo, são armazenados em objetos. Entretanto vale ressaltar que o conceito de objeto quando se trata de protocolo SNMP é diferente da concepção de programação orientada a objeto. Objetos, no âmbito do protocolo SNMP, apenas armazena valores e atributos de dispositivos na rede, e não possuem nenhum métodos de escrita e leitura de dados (TANENBAUM, 1997).

Embora a proposta de funcionamento do SNMP seja promissora sob o ponto de vista da gerência de rede, a maior dificuldade está em manter um padrão para armazenamento e leitura de seus objetos. Diferentes equipamentos de diferentes fabricantes devem disponibilizar as informações para qualquer ambiente de rede gerenciado pelo protocolo. Dessa forma, o SNMP segue o padrão ASN.1 (*Abstract Syntax Notation 1*) que trata-se de um padrão para notação flexível de descrição de

estrutura de dados. A Tabela 2.2, conforme Tanenbaum (1997), mostra os dados primitivos da ASN.1.

Tabela 2.2: Tabela com os tipos de dados do ASN.1, conforme Tanenbaum (1997)

Tipo primitivo	Significado
INTEGER	Inteiro de tamanho arbitrário
BIT STRING	Um *string* de 0 ou mais *bits*
OCTET STRING	Um *string* de 0 ou mais *bytes* sem sinal
NULL	Um marcador de lugar
OBJECT IDENTIFIER	Um tipo de dados oficialmente definido

Mesmo que a organização do SNMP esteja baseada no padrão ASN.1, os objetos do protocolo são definidos de acordo com a Estrutura de Informações Gerenciais (*Structure of Managemente Information* - SMI). O SMI descreve de uma maneira mais detalhada e burocrática, a partir do padrão ASN.1, os tipos de dados que cada objeto irá armazenar.

Dos tipos de dados definidos no SMI vale destacar o *Object Identifier*, referenciado pela sigla OID. Constite no mecanismo de identificação de um objeto baseado em uma estrutura de árvores, onde cada ramo desta árvore possa ser identificado de forma unívoca. O primeiro nível dessa árvore tem suas atribuições destinadas ao ITU-T (valor 0), ISO (valor 1), e cessão para o conjunto ITU-T e ISO (valor 2). A RFC 3061 (MEALLING, 2001) define a resolução de nomes uniformes (URN) para árvores de OIDs aplicados à internet, com o número raiz 1.3.6.1., distribuídos de duas maneiras: OIDs corporativos e uso genérico.

O número 1.3.6.1.4.1 é atribuído pelo IANA para gerência de OIDs de corporações. O IANA[13] (*Internet Assigned Number Authority*), operado pela ICANN[14] (*Internet Corporation for Assigned Names and Numbers*), é o orgão responsável pela alocação de IPs globais, zonas raízes de DNS e outras atribuições relacionadas com o protocolo de internet. É possível que uma instituição tenha seu cadastro na IANA, denominado como PEN[15] (*Private Enterprise Numbers*), de maneira que todo objeto para qualquer equipamento e *host* na rede tenha seu OID único. Supondo que uma corporação *A* tenha um PEN *X* no IANA, então o OID raiz para

[13]IANA: `http://www.iana.org/`
[14]ICANN: `http://www.icann.org/`
[15]PEN: `http://pen.iana.org/pen/PenApplication.page`

seus equipamentos será 1.3.6.1.4.1.*X*. Ramificações a partir desse número para dispositivos na rede são definidas e configuradas pelo administrador responsável.

Em situações que uma instituição não possui seu PEN junto ao IANA, porém utiliza o SNMP para gerência de dispostivos de rede, a RFC 3061 (MEALLING, 2001) define o número raiz 1.3.6.1.2.1.27 para uso com objetos do protocolo de maneira genérica, garantindo o monitoramento e funcionamento do protocolo SNMP.

O conjunto de objetos gerenciados pelo SNMP é definido como MIB *(Management Information Base*), ou base de informações gerenciais, que abrange todos os tipos de dispositivos gerenciados pelo protocolo. Atualmente está em uso o MIB-II, definido pela RFC 1213 (MCCLOGHRIE; ROSE, 1991), relativo SNMP v2 e a gerência de redes TCP/IP aplicado à internet. A Tabela 2.3 relaciona as 10 categorias de objetos gerenciados pelo SNMP.

Tabela 2.3: Tabela com as categorias MIB gerenciadas pelo SNMP, conforme Tanenbaum (1997)

Categoria	**Número de objetos**	**Descrição**
`System`	7	Nome, local e descrição do equipamento
`Interfaces`	23	Interfaces de rede e seu tráfego
`AT`	3	Conversão de endereço (obsoleto)
`IP`	42	Estatísticas de pacotes IP
`ICMP`	26	Estatísticas sobre as mensagens ICMP recebidas
`TCP`	19	Algoritmos TCP, parâmetros e estatísticas
`UDP`	6	Estatísticas de tráfego UDP
`EGP`	20	Estatísticas de tráfego de protocolo de *gateway* externo
`Transmission`	0	Reservado para MIDs de meios físicos específicos
`SNMP`	29	Estatísticas de tráfego SNMP

2.4 Estatística

Estatística é a ciência que estuda dados de fenômenos a partir da sua observação, de maneira que seja possível entender seu comportamento, obter conclusões confiáveis, realizar previsões e fundamentar tomadas de decisões. Como ciência exata, faz uso de fundamentos, teorias e artifícios matemáticos para a descrição, análise e intepretação dos dados. O seu uso em si não significa o entendimento e a solução do problema em estudo, onde é necessário o conhecimento teórico do contexto ao qual a estatística é empregada. Dessa forma, a Estatística tem sua importância na tomada de decisões,

> "...no fato de que ela não deve ser considerada como um fim em si própria, mas como um instrumento fornecedor de informações que subsidiarão, em consequência, a tomada de melhores decisões, baseadas em fatos e dados"(NETO, 2002).

O conjunto de todos os elementos definidos por pelo menos uma particularidade, sob o ponto de vista estatístico, é denominado *população* ou *universo*. Em outras palavras, são todos os elementos em análise identificados por ao menos uma característica. *Amostra* é todo subconjunto pertencente a uma população, sendo necessariamente um subconjunto finito. Limita-se na observação de uma parte da população em análise, porém com o objetivo de representá-la de forma significativa.

Entretanto, o estudo completo de todos os elementos de uma população é caracterizada como *censo* ou *recenseamento*. Contudo, conforme o tamanho da população a sua análise se torna inviável sob o aspecto financeiro, exiquibilidade e tempo. Considere como exemplo o caso de escolha de representantes por meio de eleições. O censo consistiria justamente na análise de 100% dos elementos desse universo, ou seja, todos os indivíduos com direito de voto. Verificar as intenções de votos a partir da população seria a eleição propriamente dita, que em termos práticos significa elevado custo financeiro e operacional. Dessa forma aplica-se a *amostragem*, que consiste na manipulação de uma amostra que represente a população significativamente. A viabilidade da amostragem em um curso espaço de tempo, com a obtenção dos dados simplificada e facilidade de tratamento dos resultados, apresenta seus riscos. Como a parte representa o todo, é necessário o uso do cálculo de probabilidades, que acarreta em erros uma vez que é aplicado o conceito de previsão. O uso de amostras que não representam de forma significativa a população invalida toda a análise estatística em construção. E a existência de viés

na coleta de dados onde os elementos da amostra teriam probabilidades diferentes, que torna a pesquisa tendenciosa e imparcial.

Entende-se como *Estatística Descritiva* a organização dos dados sob formas algébricas, gráficos, tabelas e diagramas. De sua nomenclatura, tem o propósito de apenas descrever os dados manipulados. Cavalca (2007) aborda probabilidade como o estudo matemático de leis baseadas no acaso, fundamentado na observação prévia e no rigor científico. Dessa forma, a *Estatística Indutiva* é constituída pela Estatística Descritiva, amostragem e estudo de probabilidade, conforme a Figura 2.4. Tem como objetivo referenciar toda análise e interpretação de dados, de maneira que atinja a proposta da ciência estatística. Neto (2002) cita os termos estatística inferencial, inferência estatística ou indução estatística para denominação de estatística indutiva.

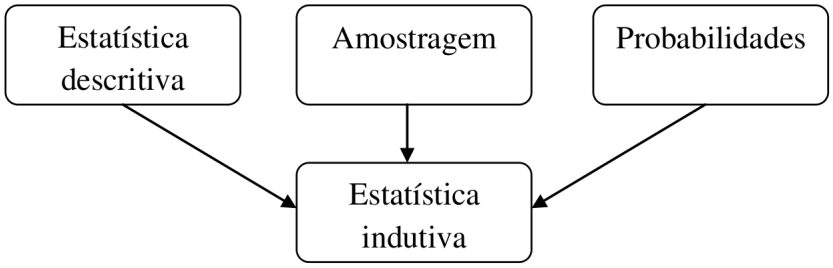

Figura 2.4: Estudo da Estatística, segundo Neto (2002).

O estudo estatístico que compõe a revisão bibliográfica deste trabalho será sintética, abordando apenas pontos chaves para o entendimento do desenvolvimento da pesquisa.

2.4.1 Distribuição normal

A distribuição normal, também conhecida como distribuição de Gauss ou gaussiana, é uma importante função densidade de probabilidade aplicada a diversos modelos físicos e financeiros que descrevem fenômenos da realidade. Tem característica simétrica em torno do parâmetro média (μ) complementada pelo desvio padrão (σ), definida pela equação 2.1. Dessa forma, os pontos $\mu - \sigma$ e $\mu + \sigma$ definem os pontos de inflexão da curva característica da distribuição, ilustrado pela Figura 2.5.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < x < \infty \tag{2.1}$$

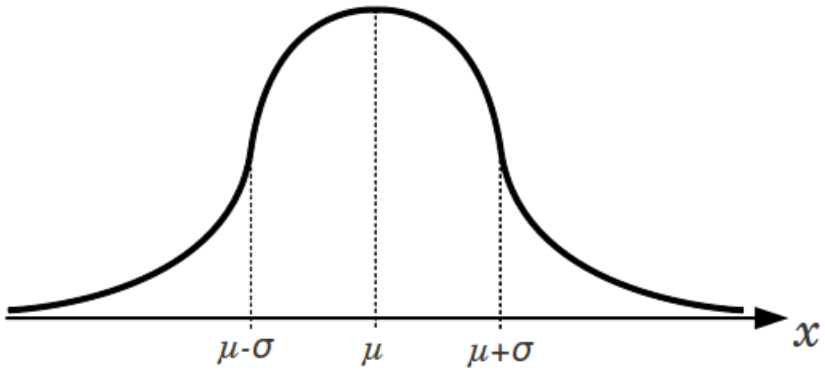

Figura 2.5: Curva característica da distribuição normal

É válido ressaltar a relação da distribuição normal com o *teorema do limite central* e o *teorema das combinações lineares*. Segundo Neto (2002), o teorema do limite central afirma que "uma variável aleatória, resultante de uma soma de n variáveis aleatórias independentes, no limite, quando n tende ao infinito, tem distribuição normal". Ao passo que o teorema das combinações lineares define que variáveis aleatórias normais independentes, combinadas linearmente, tem distribuição normal.

Com o propósito de simplificar o cálculo de probabilidade reduz-se a média para 0 e desvio padrão 1, originando na distribuição normal reduzida ou padronizada. Assim a variável x reduzida será denotada pela letra Z, que representa os valores de probabilidade da distribuição normal, conforme Tabela A.1. De x_0, valor originalmente proposto, obtém-se z_0 a partir de 2.2. Logo, a área definida pelo intervalo $[0, z_0]$ corresponde à probabilidade $P(0 \leq Z \leq z_0)$ ilustrada na Figura 2.6, sendo análogo para a distribuição normal $P(0 \leq X \leq x_0)$. A partir da característica de simetria da distribuição, é possível determinar demais probabilidades para qualquer valor de Z.

$$z_0 = \frac{x_0 - \mu}{\sigma} \tag{2.2}$$

2.4.2 Estimativa de parâmetro

Em grande parte dos casos de aplicação estatística, parâmetros populacionais como média, desvio padrão e variância são desconhecidos ou praticamente inviáveis de se obter. Existe, portanto, a necessidade de conhecê-los através da estimação de

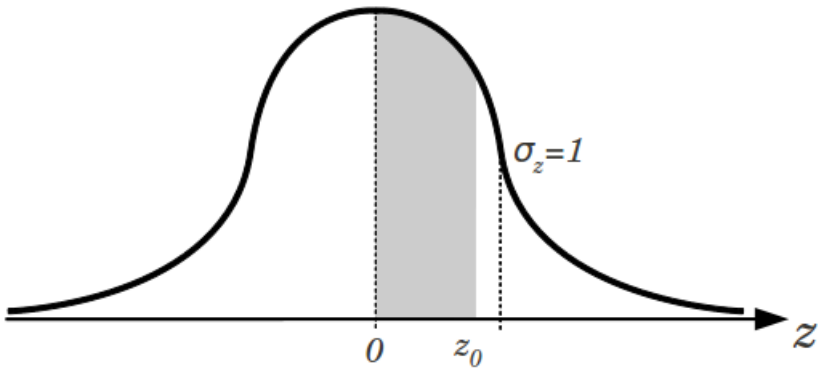

Figura 2.6: Distribuição normal padronizada

parâmetros para que a análise seja feita de fato. Para isso, chamamos de *estimador* a "variável aleatória caracterizada por uma distribuição de probabilidade e seus respectivos parâmetros", conforme Neto (2002), e *estimativa* os valores de cada parâmetro pertencente a um estimador. No contexto desse estudo, é de grande importância considerar as amostras como probabilísticas e o processo de amostragem como casual simples, onde cada elemento tem a mesma probabilidade.

Uma forma de realizar a estimação de parâmetros é através de pontos, onde cada valor será estimado por um único número contido no conjunto que representa a variável em estudo. Como os valores dos parâmetros são provenientes de variáveis aleatórias e na maioria das vezes contínuas, em outras palavras pertencentes a um intervalo real, um erro de estimação é praticamente garantido. Isso ocorre porque, sendo as variáveis aleatórias, haverão diferentes estimativas para diferentes amostras, mesmo com iguais números de elementos. Dessa forma, dado um nível de significância, um parâmetro será definido por meio de um intervalo. Ou seja, definida uma margem de erro, uma estimativa estará contida em um intervalo de confiança.

2.4.2.1 Intervalo de confiança da μ com σ conhecido

Uma maneira de estimar a média de uma população com o seu desvio padrão conhecido é através de um intervalo de confiança. A Figura 2.7 auxilia o entendimento da construção desse intervalo. Toda a área compreendida entre $\mu - e_0$ e $\mu + e_0$ representa o intervalo propriamente dito, onde o valor da média a ser estimada deverá estar contida.

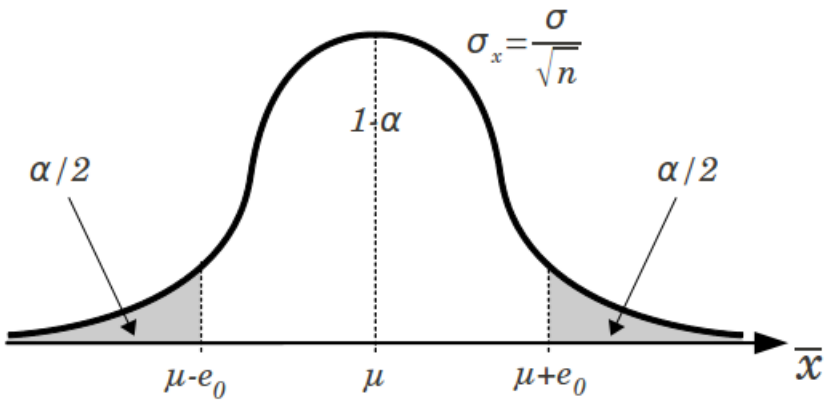

Figura 2.7: Intervalo de confiança de μ, Neto (2002)

Cabe a observação que toda a curva ilustrada na Figura 2.7 é simétrica em relação a média μ. Além disso, a área para valores maiores que $\mu + e_0$ e menores que $\mu - e_0$ representa a probabilidade da média não estar contida no intervalo construído. Em sequencia ao raciocínio da estimação de parâmetros, o intervalo de confiança será definido pela equação 2.3.

$$P(\mu - e_0 \leq \bar{x} \leq \mu + e_0) = 1 - \alpha \tag{2.3}$$

onde:

- μ: média da população;
- σ: desvio padrão conhecido da população;
- $\bar{x}$: média da amostra retirada;
- n: tamanho da amostra recolhida;
- e_0: semi-amplitude do intervalo de confiança;
- α: nível de significância considerado. Podemos também caracterizar $1 - \alpha$ como coeficiente de confiança, onde é representada a probabilidade de se obter o intervalo desejado (WERKEMA, 1996).

Da desigualdade definido em 2.3

$$\mu - e_0 \leq \bar{x} \qquad e \qquad \bar{x} \leq \mu + e_0$$

$$\therefore \quad \mu \leq \bar{x} + e_0 \qquad e \qquad \bar{x} - e_0 \leq \mu$$

$$\therefore \quad \bar{x} - e_0 \leq \mu \qquad e \qquad \mu \leq \bar{x} + e_0$$

$$\therefore \quad P(\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0) = 1 - \alpha \tag{2.4}$$

Da equação 2.4 verifica-se a média populacional μ a ser estimativa, contida no intervalo definido por $\bar{x} - e_0$ e $\bar{x} + e_0$ a um nível de significância α. Supondo que $\alpha = 5\%$, a média μ será estimada com 95% de certeza.

Resta definir o valor de e_0 para o cálculo do intervalo de confiança desejado. A semi-amplitude do intervalo será calculado a partir de da variável z da distribuição normal padronizada, conforme equação 2.2. Segundo Neto (2002), como $\bar{x}$ é uma distribuição amostral, o desvio padrão da amostra será o quociente entre o desvio padrão da população e a raiz quadrada do tamanho da amostra, como é definido em 2.5.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2.5}$$

Assim:

$$z_{\alpha/2} = \frac{x - \mu}{\sigma_{\bar{x}}}$$

$$z_{\alpha/2} = \frac{(\mu + e_0) - \mu}{\sigma/\sqrt{n}}$$

$$\therefore \quad e_0 = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{2.6}$$

Substituindo a semi-amplitude e_0 definida em 2.6 na construção do intervalo de confiança conforme 2.4, a estimativa da media será dada pela equação 2.7.

$$P\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{2.7}$$

2.4.2.2 Intervalo de confiança da μ com σ desconhecido

Na prática, a grande problemática da estimativa de parâmetros está no total desconhecimento de parâmetros populacionais. Diferentemente do item anterior, onde foi discutido a construção de intervalo de confiança para a média com o desvio padrão conhecido, a maioria dos problemas práticos tem como característica a não

determinação do desvio padrão da população. Uma solução seria a simples substituição do desvio padrão da população σ pelo da amostra s_x. Como s_x é uma estimativa para desvio padrão obtido a partir da amostra coletada, o grau de incerteza do intervalo de confiança será automaticamente incrementado.

Uma medida corretiva para minimizar essa subtituição é o emprego da distribuição t de Student com grau de liberdade $n-1$, definido pela equação 2.8 conforme Neto (2002). Os valores da distribuição de t de Student podem ser visualizados na Tabela A.2. Para amostras grandes, s_x se aproxima de σ, assim como a distribuição t de Student terá comportamento semelhante à distribuição normal. Empiricamente, segundo Werkema (1996), amostras com $n \geq 30$ garantem a estimativa da média com desvio padrão populacional desconhecido.

$$t_{n-1,\alpha/2} = z_{\alpha/2}\frac{\sigma}{s_x} \tag{2.8}$$

A partir do intervalo de confiança para média populacional com desvio padrão desconhecido definido na equação 2.7, temos:

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.9}$$

A expressão 2.9 pode ser escrita da seguinte maneira, sendo em seguida substituído por 2.8:

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{s_x}\frac{s_x}{\sqrt{n}}$$

$$\therefore \qquad \overline{x} \pm t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}} \tag{2.10}$$

De 2.10, a média populacional μ com desvio padrão σ pode ser estimada a partir do intervalo construído na equação 2.11

$$P\left(\overline{x} - t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}}\right) = 1-\alpha \tag{2.11}$$

2.4.2.3 Tamanho da amostra

Em algumas situações, deseja-se determinar o tamanho da amostra n necessária para a estimativa de parâmetros. Na verdade, a determinação do tamanho da amostra é de grande relevância para a validação do planejamento estatístico. Para isso são fixados os valores da semi-amplitude e_0 do intervalo, em outras palavras a

margem do intervalo, e o nível de significância α da estimativa. Da equação 2.6, basta isolar a variável n para se obter o tamanho da amostra. Assim:

$$n = \left(\frac{z_{\alpha/2}\ \sigma}{e_0} \right)^2 \tag{2.12}$$

Hipoteticamente, intervalos de confiança com semi-amplitude próximas a zero representam um intervalo curto, que na prática nos remetem a uma precisão muito alta. Entretanto a medida que $e_0 \to 0$, n tenderá a valores cada vez maiores que na prática pode se tornar inviável quando se fala na retirada da amostra, considerando um $z_{\alpha/2}$ constante. Por outro lado, para um nível de confiança próximo a 100%, $z_{\alpha/2} \to 0$ e e_0 assumirá valor grande o suficiente para que os intervalos definidos tenham grande amplitude. Logo, imprecisos e sem grande validade estatística.

Para intervalos de confiança para média populacional com desvio padrão desconhecido, a definição da distribuição t de Student em 2.8 pode ser escrita da seguinte maneira:

$$t_{n-1,\alpha/2} s_x = z_{\alpha/2}\ \sigma$$

Assim, pelo tamanho da amostra definido na expressão 2.12 :

$$n = \left(\frac{t_{n'-1,\alpha/2}\ s_x}{e_0} \right)^2 \tag{2.13}$$

O cálculo da distribuição t de Student exige que se tenha o valor de n, embora seja este o parâmetro a ser calculado. Neste caso, uma amostra piloto n' é tomada como base e aceita até que $n \leq n'$. Caso essa condição não seja satisfeita, uma nova amostra piloto deve ser recolhida, calculada a partir de 2.13 e verificada. A iteração se repete até que o tamanho da amostra seja menor ou igual ao tamanho da amostra piloto.

2.4.3 Teste de hipóteses

Dada a existência de uma determinada hipótese, a realização de testes para que esta seja validada ou não constitui no estudo de teste de hipóteses. Werkema (1996) define hipótese como "uma afirmação sobre os parâmetros de uma ou mais população". De fato, considera-se a hipótese existente H_0 a ser testada, e a hipótese alternativa H_1 que complementa H_0.

Feito o teste, aceitar H_0 implica em rejeitar H_1, da mesma forma que aceitar H_1 significa rejeitar H_0. O nível de significância corresponde a probabilidade de

ocorrer uma erro na realização de um teste. Em outras palavras existe a possibilidade de rejeição de H_0 sendo esta verdadeira, assim como a aceitar H_1 onde na realidade é falsa. Esses erros podem ser sintetizados como:

- nível de significância α - erro Tipo I: seja H_0 verdadeira, rejeita-se H_0;
- nível de significância β - erro Tipo II: seja H_1 falsa, aceita-se H_1.

Em testes de hipóteses, a faixa de valores que remetem a rejeição de H_0 é denominada como *região crítica (RC)*, ao passo que definimos como *região de aceitação* os valores restantes que implicam na aceitação de H_0.

Na prática, um teste de hipótese consiste verificar se a média $\bar{x}$ da amostra recolhida pertence ao intervalo de valores que definem a região crítica. Caso verdade, H_1 será aceita e H_0 automaticamente rejeitada. Outro caso é $\bar{x}$ estar contido na região de aceitação, que implica em aceitar H_0 e rejeitar H_1. Para tanto, inicialmente fixa-se um valor para o nível de significância α, por questões didáticas valores de 1% ou 5%. Na aplicação, a escolha do nível de significância depende do contexto ao qual o teste se aplica. Geralmente o nível de significância é fixado em 5%.

O próximo passo consiste na determinação do limite das regiões crítica e de aceitação, dito por $\overline{x_1}$. Este valor é calculado a partir:

- da média da população, sendo esta na grande maioria dos casos estimada através de um intervalo de confiança, segundo Neto (2002);
- do desvio padrão:
  - da população, caso conhecido, é descrito pela distribuição normal;
  - da amostra, caso desconhecido, é descrito pela distribuição de t de *Student*;
- do nível de significância fixado inicialmente;
- do tamanho da amostra.

Com a região crítica delimitada torna-se possível realizar o teste, verificando neste caso se a média da população ($\bar{x}$) é *menor* que o valor limite da região crítica ($\overline{x_1}$), o mesmo que afirmar $\mu < \mu_0$. Podemos dizer que existe H_0 igual a média estimada e deseja-se verificar se é aceitável afirmar que a média da amostra é

menor que a média da população. A Figura 2.8 ilustra a construção desse teste de hipótese. A construção deste teste da seguinte maneira:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$$

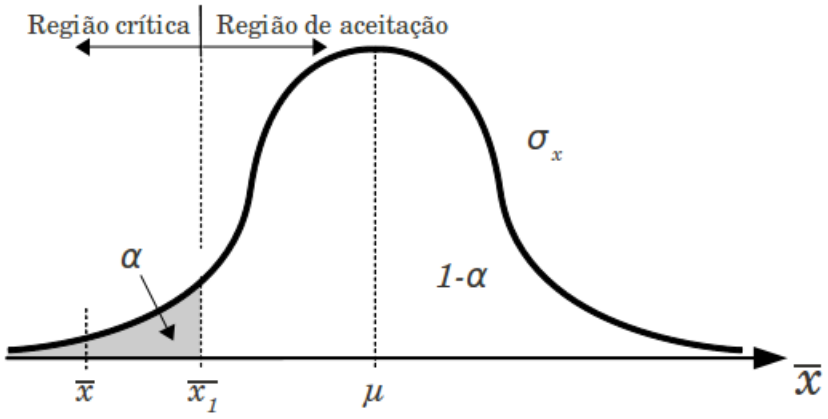

Figura 2.8: Construção de um teste de hipótese

Existem mais duas possíveis situações na construção e realização de testes de hipóteses. Uma seria a hipótese alternativa H_1 consistir na média da amostra ser *maior* que a média da população estimada, ou seja, $\mu > \mu_0$. O raciocínio é análogo, diferenciado apenas no valor de limite da região crítica $\bar{x}$ ser maior que a média da população, ou seja, $\bar{x}$ à direta de μ.

A outra situação seria a hipótese da média da amostra ser *diferente* da média da população, o mesmo que $\mu \neq \mu_0$. Este tipo de teste é denominado como bicaudal, onde há duas regiões críticas delimitadas por $\overline{x_1}$ e $\overline{x_2}$ localizados à esquerda e à direta de μ. A probabilidade de erro de cada uma das regiões críticas nesse teste corresponde à metade do nível de significância.

2.4.3.1 Teste de hipótese da μ com σ conhecido

A partir da construção de um teste de hipótese conforme a Figura 2.8, o limite da região crítica será definido pela diferença da média da populção estimada μ e a semi-amplitude da distribuição normal que se aplica, definido na equação 2.6. Nesse teste para a média da população, é conhecido o desvio padrão da população.

Dessa forma, podemos afirmar:

$$\overline{x_1} = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \tag{2.14}$$

Como a intenção desse teste em específico é verificar se a média $\bar{x}$ da amostra colhida é menor que $\overline{x_1}$, em outras palavras $H_1 : \mu < \mu_0$, temos:

$$\bar{x} < \overline{x_1} \tag{2.15}$$

Substituindo $\overline{x_1}$ definido em 2.14 na expressão 2.15:

$$\bar{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$\therefore \qquad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

$$\therefore \qquad z < -z_\alpha \tag{2.16}$$

A substituição de z em 2.16 foi deduzida a partir da construção do intervalo de confiança definido na expressão 2.7. De qualquer maneira, o valor de z pode ser calculado a partir dos parâmetros da amostra conforme explicitado em 2.17. A partir da condição definida em 2.16, caso verdade, H_1 é aceita e H_0 rejeitada.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{2.17}$$

A partir da expressão definida em 2.15, a realização do teste para a hipótese $H_1 : \mu > \mu_0$ será de maneira análoga, deferenciando apenas a condição inicial $\bar{x} > \overline{x_1}$. Simplificando os cálculos em função dos valores de z, nesse caso aceitar H_1 será possível se $z > z_\alpha$.

Para o caso de hipótese como $H_1 : \mu \neq \mu_0$, a condição inicial será definida por $\bar{x} < \overline{x_1}$ ou $\bar{x} > \overline{x_2}$. O condicional *ou* se aplica devido ao teste bicaudal devido H_1. Cabe lembrar que o nível de significância para cada limite de região crítica será a metade do valor inicial fixado. Aceitar H_1 será possível se $|z| > z_{\alpha/2}$.

Em resumo, a Tabela 2.4 sintetiza os possíveis casos de testes de hipóteses com a devida condição para aceitação de H_1.

Tabela 2.4: Testes de hipóteses para média com σ conhecido, conforme Neto (2002)

Hipóteses	Rejeição de H_0
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu < \mu_0$	$z < -z_\alpha$
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu > \mu_0$	$z > z_\alpha$
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu \neq \mu_0$	$\lvert z \rvert > z_{\alpha/2}$

2.4.3.2 Teste de hipótese da μ com σ desconhecido

Assim como visto na estimação de parâmetros através da construção de intervalos de confiança, a maioria dos casos práticos o desvio padrão da população é desconhecido. Isso se aplica ao estudo de testes de hipóteses e da mesma maneira, a distribuição normal descrita pelos valores de z pode ser representada através da distribuição de t de Student. Portanto, a partir da equação 2.17 para cálculos de valores para realização do teste, podemos reescrevê-la em função da distribuição de t de Student como na expressão 2.18.

$$t_{n-1} = \frac{\overline{x} - \mu_0}{s_x/\sqrt{n}} \tag{2.18}$$

Os mesmos casos de hipóteses deduzidos anteriomente são aplicados para estes tipos de teste, da mesma maneira que as condições para aceitar H_0 ou H_1 são análogas. A Tabela 2.5 resume os testes de hipóteses quando o desvio padrão da população é desconhecido.

Tabela 2.5: Testes de hipóteses para média com σ desconhecido, conforme Neto (2002)

Hipóteses	Rejeição de H_0
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu < \mu_0$	$t_{n-1} < -t_{n-1,\alpha}$
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu > \mu_0$	$t_{n-1} > t_{n-1,\alpha}$
$H_0: \quad \mu = \mu_0$ $H_1: \quad \mu \neq \mu_0$	$\lvert t_{n-1} \rvert > t_{n-1,\alpha/2}$

2.4.4 Comparação entre duas médias

Foi visto até agora testes de hipóteses com base em um conjunto amostral de uma única variável aleatória. É possível, entretanto, que situações que envolvam duas ou mais populações sejam averiguadas a partir desse raciocínio. Da mesma forma, para os casos vistos nas seções anteriores, os teste de hipóteses podem se extender para os parâmetros de média, variância, e proporção populacional. Nessa seção será abordada a comparação do estimador média para duas amostras.

Comparação entre duas médias, no contexto de teste de hipóteses, implica em verificarmos a diferença desse estimador para as duas amostras. Dessa forma, averiguamos a condição conforme 2.19. Neto (2002) ressalta a atenção para casos em que $\Delta = 0$, que pode ser escrito como $\mu_1 = \mu_2$.

$$H_0: \quad \mu_1 - \mu_2 = \Delta \tag{2.19}$$

A partir da hipótese principal, definida por H_0, constroe-se as demais comparações de média com base nos testes discutidos nas seções anteriores. Essas comparações estão listadas nas expressões 2.20, 2.21 e 2.22.

$$\begin{cases} H_0: & \mu_1 - \mu_2 = \Delta \\ H_1: & \mu_1 - \mu_2 > \Delta \end{cases} \tag{2.20}$$

$$\begin{cases} H_0: & \mu_1 - \mu_2 = \Delta \\ H_1: & \mu_1 - \mu_2 < \Delta \end{cases} \tag{2.21}$$

$$\begin{cases} H_0: & \mu_1 - \mu_2 = \Delta \\ H_1: & \mu_1 - \mu_2 \neq \Delta \end{cases} \tag{2.22}$$

Pode-se reduzir a comparação de duas médias considerando que os dados são *emparelhados* ou *não-emparelhados*. Diremos que os dados serão emparelhados quando as populações tiverem algum tipo de correlação, com base em algum critério. Como exemplo, seja um conjunto A que representa os alunos de uma escola onde a média de uma matéria qualquer é representada pela variável x. Podemos associar, dessa forma, cada média x_i a cada aluno a_i. Obtém-se, em um segundo momento, outra média dessa matéria representada pela variável y, onde permance a associação a cada aluno a_i através de y_i. Essa associação garante a correlação de cada elemento analisado com cada uma das duas médias, criando uma comparação do tipo "inicial"e "final", ou "anterior"e "posterior". Dessa forma dizemos que as populações são correlatas, logo, os dados são emparelhados.

Serão considerados testes com dados emparelhados somente populações que, de alguma maneira, podem ser correlacionadas. Isso implica dizer que nesse tipo de comparação o tamanho das duas amostras são iguais. Se alguma dessas duas situações não forem satisfeitas, associamos à comparação de dados não emparelhados. É possível, matematicamente, realizar a comparação de duas amostras não correlacionadas considerando que os dados sejam emparelhados. No entanto, segundo Neto (2002), isso implica em perda no poder do teste, o que torna indesejável sob o ponto de vista estatístico.

Os casos de dados não-emparelhados podem ser subdivididos quando:

- os desvios padrão σ_1 e σ_2 das populações são conhecidos;
- os desvios padrão σ_1 e σ_2 das populações são desconhecidos porém admitidos que são iguais, ou seja, $\sigma_1 = \sigma_2 = \sigma$;
- os desvios padrão σ_1 e σ_2 das populações são desconhecidos e diferentes.

No contexto dessa pesquisa não serão estabelecidas, na fase de planejamento do experimento, correlação entre duas variáveis aleatórias. Portanto esses dados não serão emparelhados. Da mesma forma que são desconhecidos os desvios padrão das populações manipuladas, implicando na exclusão do primeiro caso de dados não emparelhados. Mesmo com os desvios padrão desconhecidos, esses estimadores não serão considerados iguais. Este trabalho abordará a comparação entre duas médias apenas para dados não emparelhados, onde os desvios padrão da população são desconhecidos e, a princípio, diferentes.

Dessa forma, seja a comparação entre duas médias apresentado na expressão 2.20. Como os desvios padrão são desconhecidos, será empregado a distribuição t de Student. Como a comparação entre as médias tem o mesmo raciocínio dos testes de hipóteses, a Tabela 2.6 relacinoa os casos que a hipótese inicial H_0 é rejeitada.

A equação 2.24 define o valor de $t_{n_1+n_2-2}$, segundo Neto (2002).

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{2.23}$$

Como já foi discutido anteriormente, para amostras suficientemente grandes pode ser feita uma aproximação com a distribuição normal. Caso o tamanho da amostra não seja grande o suficiente, ou para que se obtenha uma maior precisão,

Tabela 2.6: Comparação entre média com σ desconhecido

Hipóteses	Rejeição de H_0
$H_0: \quad \mu_1 - \mu_2 = \Delta$ $H_1: \quad \mu_1 - \mu_2 < \Delta$	$t_{n_1+n_2-2} < -t_{v,\alpha}$
$H_0: \quad \mu_1 - \mu_2 = \Delta$ $H_1: \quad \mu_1 - \mu_2 > \Delta$	$t_{n_1+n_2-2} > t_{v,\alpha}$
$H_0: \quad \mu_1 - \mu_2 = \Delta$ $H_1: \quad \mu_1 - \mu_2 \neq \Delta$	$\lvert t_{n_1+n_2-2}\rvert > t_{v,\alpha/2}$

Neto (2002) segure o método de *Aspin-Welch*. Consiste encontrar um valor t crítico, representado por $t_{v,\alpha}$, em função do número de grau de liberdade definido a equação 2.24.

$$v = \frac{(w_1 + w_2)^2}{w_1^2/(n_1+1) + w_2^2/(n_2+1)} \tag{2.24}$$

onde w_1 e w_2 são definidos por:

$$w_1 = \frac{s_1^2}{n_1} \qquad e \qquad w_2 = \frac{s_2^2}{n_2}$$

2.4.5 Comparação entre várias médias

Na seção anterior foi discutida a comparação de médias de duas populações, representadas por duas variáveis aleatórias. Isso implica concluir se o estimador "média"entre as amostras tem ou não diferença significativa. O estudo se concentra agora na comparação entre várias médias, ao qual o método empregado para realizar essa inferência estatística é a denominada Análise de Variância, ou ANOVA. Historicamente, segundo Neto (2002), a análise de variância foi elaborada como ferramenta para análise de experimentos estatísticos, pelo estatístico britânico sir R. A. Fisher.

A comparação entre várias médias podem ser dividida em quatro situações, descritas a seguir:

- uma classificação, onde as amostras possuem o mesmo tamanho;
- uma classificação, onde as amostras possuem tamanhos distintos;
- duas classificações sem repetição;

- duas classificações com repetição.

Como nesse estudo são consideradas várias populações distintas, denominaremos como *observações* cada dado, ou conjunto de dados, coletado que constituem nas amostras. Feitas as observações, elas podem ser agrupadas conforme critérios definidos na fase de planejamento do experimento, também denotados como *classificação*. Casos em que a análise refere-se apenas à uma classificação, o número de observações podem ser tanto iguais quanto diferentes. Nos casos de duas classificações, as observações podem ser únicas (sem repetição) como replicadas (com repetição).

Será abordado com um detalhamento maior o caso de análise de variância de uma classificação com tamanho de amostras iguais. Os demais casos são análogos, não cabendo ao presente trabalho uma abordagem matemática mais abrangente.

2.4.5.1 Uma classificação com amostras do mesmo tamanho

Sejam k amostras, oriundas de k populações de tamanho n, com as respectivas médias $\mu_i (i = 1,2,\ldots,n)$. Consideraremos que todas as populações possuem a mesma variância, e que as variáveis aleatórias que representam cada uma das populações são distribuídas uniformemente. Embora matematicamente obter variâncias numericamente exatas seja inviável, o método de análise de variância garante uma boa aproximação entre as amostras, logo uma grande eficiência na comparação.

A premissa inicial é comparar todas as médias por meio de uma hipótese existente H_0, conforme a expressão 2.25.

$$\{H_0 : \quad \mu_1 = \mu_2 = \ldots \mu_n \tag{2.25}$$

Para prosseguir com a comparação entre várias médias, adotaremos a seguinte notação para o estudo de análise de variância, conforme Neto (2002). Cada valor será representado por $x_{ij} (i = 1,2,\ldots,k; j = 1,2,\ldots,n)$, onde i é uma amostra dentre das k amostras recolhidas, e j um dado de uma amostra de n elementos. Assim como:

- Soma dos valores da i-ésima amostra: $T_i = \sum_{j=1}^{n} x_{ij}$;
- Soma dos quadrados dos valores da i-ésima amostra: $Q_i = \sum_{j=1}^{n} x_{ij}^2$;
- Soma dos valores: $T = \sum_{i=1}^{n} T_i = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}$;

- Soma dos quadrados dos valores: $Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^2$;
- Média da i-ésima amostra: $\overline{x}_i = T_i/n$;
- Média de todos os valores: $\overline{\overline{x}} = T/nk$;

A ANOVA consiste em estimar a variância σ^2 através de 3 maneiras diferentes, considerando que a hipótese H_0 seja verdadeira.

A primeira delas é a **estimativa total** S_T^2, ao qual se unificam todas as k amostras analisadas. Em outras palavras, as amostras irão se comportar como uma única variável resultante, sendo possível obter uma estimativa única e total para a variância σ^2. Conforme Neto (2002), a estimativa total é obtida através da equação 2.26.

$$S_T^2 = \frac{Q - T^2/nk}{nk - 1} \tag{2.26}$$

Denominaremos o numerador de S_T^2 como a *soma de quadrados total* (SQT), ilustrado em 2.27:

$$SQT = Q - T^2/nk \quad \Longrightarrow \quad S_T^2 = \frac{SQT}{nk - 1} \tag{2.27}$$

Com base na situação anterior, o fato de todas as amostras serem unificadas permite dizer que a média total $\overline{\overline{x}}$ dessa amostra única é a somatória de cada uma das médias $\overline{x}_i$ das k amostras. Analogamente, a unificação da variância de cada uma das amostras é um bom estimador σ^2 para a variância total resultante, que implica na **estimativa entre amostras** S_E^2. A equação 2.28 ilustra o cálculo da estimativa entre amostras, segundo Neto (2002).

$$S_E^2 = \frac{\sum_{i=1}^{k} T_i^2/n - T^2/nk}{k - 1} \tag{2.28}$$

O numerador de S_E^2 denominaremos como a *soma de quadrado entre amostras* (SQE), conforme 2.29.

$$SQE = \sum_{i=1}^{k} T_i^2/n - T^2/nk \quad \Longrightarrow \quad S_E^2 = \frac{SQE}{k - 1} \tag{2.29}$$

A terceira maneira que estima o valor da variância pode ser obtida através da média aritmética de σ^2 de cada amostra. Essa média obtida representa, significativamente e independente da hipótese H_0 ser verdadeira ou não, a estimativa da

variância total do conjunto de amostras. Denominamos como **estimativa residual** S_R^2, calculada conforme a equação 2.30.

$$S_R^2 = \frac{Q - \sum_{i=1}^{k} T_i^2/n}{k(n-1)} \tag{2.30}$$

Chamaremos o numerador de *soma dos quadrados residual* (SQR), conforme a expressão 2.31.

$$SQR = Q - \sum_{i=1}^{k} T_i^2/n \quad \Longrightarrow \quad S_R^2 = \frac{SQR}{k(n-1)} \tag{2.31}$$

Das equações 2.26, 2.28 e 2.30 é possível visualizar que $SQT = SQR + SQE$, onde algebricamente SQR e SQE são independentes. Paralelamente, a partir de duas amostras s_1^2 e s_2^2, a distribuição F de Snedecor analisa justamente a relação s_1^2/s_2^2. Neto (2002) descreve com maiores detalhes o comportamento da distribuição F de Snedecor, suas características e comportamento.

Como SQR e SQE são independentes, verificamos o quociente da estimativa entre amostras com a estimativa residual para a inferência da análise de variância conforme a equação 2.32.

$$F = \frac{S_E^2}{S_R^2} \tag{2.32}$$

O valor crítico de F será determinado em função do número de graus de liberdade da estimativa entre amostras (S_E^2) e da estimativa residual (S_R^2), com os respectivos valores $k-1$ e $k(n-1)$. Denotamos o valor crítico da distribuição de F de Snedecor como $F_{k-1,k(n-1),\alpha}$, onde α é o nível de significância escolhido para a comparação de testes. Portanto, **rejeitaremos a hipótese** H_0 se a condição descrita em 2.33 for satisfeita.

$$F > F_{k-1,k(n-1),\alpha} \tag{2.33}$$

O processo de análise de variância em muitos casos é executado e auxiliado por *software* ou ferramentas computacionais. Entre os vários *software* livres disponíveis com recursos estatísticos, citamos o R-Project[16], GNU-Octave[17] e SciLab[18].

[16] **R-Project**: `http://www.r-project.org/`
[17] **GNU-Octave**: `http://www.gnu.org/software/octave/`
[18] **SciLab**: `http://www.scilab.org/`

Segundo Neto (2002), a comparação entre várias médias é usualmente sintetizada conforme a Tabela 2.7, onde inclusive é utilizada por diversas ferramentas computacionais.

Tabela 2.7: Síntese para comparação entre médias para uma classificação com amostras de mesmo tamanho

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	F_α
`Entre amostras`	$SQE = \sum_{i=1}^{k} \frac{T_i^2}{n} - \frac{T^2}{nk}$	$k-1$	$S_E^2 = \frac{SQE}{k-1}$	$F = \frac{S_E^2}{S_R^2}$	$F_{k-1,k(n-1),\alpha}$
`Residual`	$SQR = Q - \sum_{i=1}^{k} \frac{T_i^2}{n}$	$k(n-1)$	$S_R^2 = \frac{SQR}{k(n-1)}$		
`Total`	$SQT = Q - \frac{T^2}{nk}$	$n(k-1)$			

2.4.5.2 Uma classificação com amostras de tamanhos distintos

A análise de variância com uma única classificação, onde agora as amostras possuem tamanhos destintos, é análoga ao caso anteriormente discutido. O principal diferencial está na notação de j, que representa o elemento de cada i amostra de tamanho n_i. A Tabela 2.8 sintetiza a comparação para esse tipo de análise.

Tabela 2.8: Síntese para comparação entre médias para uma classificação com amostras de tamanhos diferentes

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	F_α
`Entre amostras`	$SQE = \sum \frac{T_i^2}{n_i} - \frac{T^2}{\sum n_i}$	$k-1$	$S_E^2 = \frac{SQE}{k-1}$	$F = \frac{S_E^2}{S_R^2}$	$F_{k-1,\sum n_i-k,\alpha}$
`Residual`	$SQR = Q - \sum \frac{T_i^2}{n_i}$	$\sum n_i - k$	$S_R^2 = \frac{SQR}{\sum n_i - k}$		
`Total`	$SQT = Q - \frac{T^2}{\sum n_i}$	$\sum n_i - 1$			

2.4.5.3 Duas classificações sem repetição

A comparação entre várias médias com duas classificações consiste em subdividir todos os elementos conforme critérios, pré estabelecidos na etapa de planejamento do experimento. Sejam todos os dados observados dispostos em uma matriz com k linhas e n colunas. Dados classificados em função de um primeiro critério, de maneira que temos k amostras de n elementos, são representados nas linhas da matriz. Da mesma maneira que a classificação dos dados conforme um segundo critério de n amostras com k elementos são dispostos nas colunas da matriz. A matriz em 2.34 ilustra a representação desses dados para análise de variância com duas classificações sem repetição.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kj} & \cdots & x_{kn} \end{bmatrix} \tag{2.34}$$

Com posse dos dois critérios que determinam as duas classificações, o teste que corresponde à comparação entre várias médias será definido pelas hipóteses iniciais H_{01} e H_{02} conforme a expressão 2.35. Analogamente ao caso de uma classificação, como em 2.25, temos:

$$\begin{cases} H_{01}: & \mu_1 = \mu_2 = \ldots \mu_k \\ H_{02}: & \mu_1 = \mu_2 = \ldots \mu_n \end{cases} \tag{2.35}$$

Os três mecanismos empregados para a estimativa da variância são mantidos, onde se altera apenas a estimativa entre amostras (S_E^2) para cada uma das classificações. Dessa forma, a estimativa ente amostras se dará entre linhas (S_L^2) e entre colunas (S_C^2), representadas pelas expressões 2.36 e 2.37 respectivamente. O numerador de S_L^2 será denominado como SQL, e de S_C^2 como SQC.

$$S_L^2 = \frac{\sum_{i=1}^{k} T_i^2/n - T^2/nk}{k-1} \quad \Rightarrow \quad SQL = \sum_{i=1}^{k} T_i^2/n - T^2/nk \quad \Rightarrow \quad S_L^2 = \frac{SQL}{k-1} \tag{2.36}$$

$$S_C^2 = \frac{\sum_{j=1}^{k} T_j^2/k - T^2/nk}{n-1} \quad \Rightarrow \quad SQC = \sum_{j=1}^{k} T_j^2/k - T^2/nk \quad \Rightarrow \quad S_C^2 = \frac{SQC}{n-1} \tag{2.37}$$

A dedução da soma de quadrados total (SQT) é análoga à comparação com apenas uma classificação, onde $SQT = SQR + SQL + SQC$, e SQR, SQL e SQC são independentes. Aplicando a distribuição F de Snedecor, temos:

$$F_L = \frac{S_L^2}{S_R^2} \qquad e \qquad F_C = \frac{S_C^2}{S_R^2}$$

onde F_L e F_C são independentes. Adicionalmente, a soma dos quadrados residual pode ser determinada através da diferença $SQR = SQT - SQL - SQC$.

O cálculo de F crítico será em função dos graus de liberdade de cada critério, ou seja, de $(k-1)$ para linhas e $(n-1)$ para colunas. O grau de liberdade para a estimativa residual é $(k-1)(n-1)$. Dessa forma, H_0 será rejeitada se a condição em 2.38 for satisfeita.

$$F_L > F_{k-1,(k-1)(n-1),\alpha} \quad ou \quad F_C > F_{n-1,(k-1)(n-1),\alpha} \tag{2.38}$$

Na análise de variância com duas classificações sem repetição, não concluímos o fato dos critérios terem algum tipo de relação. Inferimos de maneira que, a um nível de significância α, existe ou não diferença significativa entre as linhas, assim como entre as colunas.

A comparação entre várias médias pode ser sintetizada conforme a Tabela 2.9.

2.4.5.4 Duas classificações com repetições

Este caso de comparação entre várias médias herda grande parte das características do caso com duas classificações sem repetição. A partir da matriz em 2.34 que representa as observações dispostas por dois critérios, em linhas e em colunas, cada dado x_{ij} será replicado com diferentes valores dentro do contexto de cada observação. Obviamente que esses valores replicados são provenientes da coleta de dados do experimento realizado. Assim, teremos r repetições para cada uma das nk observações.

Tabela 2.9: Síntese para comparação entre médias para duas classificações sem repetição

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	F	F_α
`Entre linhas`	$SQL = \frac{\sum_{i=1}^{k} \frac{T_i^2}{n} - \frac{T^2}{nk}}{k-1}$	$k-1$	$S_L^2 = \frac{SQL}{k-1}$	$F_L = \frac{S_L^2}{S_R^2}$	$F_{k-1,(k-1)(n-1),\alpha}$
`Entre colunas`	$SQC = \frac{\sum_{j=1}^{k} \frac{T_j^2}{k} - \frac{T^2}{nk}}{n-1}$	$n-1$	$S_C^2 = \frac{SQC}{n-1}$	$F_C = \frac{S_C^2}{S_R^2}$	$F_{n-1,(k-1)(n-1),\alpha}$
`Residual`	$SQR = SQT - SQL - SQC$	$(n-1)(k-1)$	$S_R^2 = \frac{SQR}{(k-1)(n-1)}$		
`Total`	$SQT = Q - \frac{T^2}{nk}$	$n(k-1)$			

Entretanto, é pertinente a devida atenção na situação de interação entre os critérios. Isso significa que as linhas e as colunas podem se interagir, de maneira que quando são analisadas separadamente, dizemos que o teste que constitui a comparação entre as médias perde sua força. Em outras palavras, ao se realizar o comparação entre várias médias dessa natureza, uma das conclusões é a aceitação ou rejeição de interação entre os critérios I e II, além das diferenças significativas entre as linhas e as colunas.

O raciocínio da análise de variância para esse caso, onde se tem duas classificações com repetição, é semelhante ao anteriormente discutido. Será incluído agora a estimativa da variância de interação e tratamento entre os critérios. A Tabela 2.10 sintetiza o procedimento para a comparação entre várias médias para este caso em estudo.

Segundo Neto (2002), o procedimento que envolve a interação entre os critérios é recomendado somente quando a condição 2.39 for satisfeita. Caso contrário, a soma de quadrados de interação e o seu respectivo número de graus de liberdade devem ser somados à variação residual, e a sua relação representará o quadrado médio residual (S_R^2). Consequentemente a análise de variância será finalizada apenas com a comparação de F_L e F_C, semelhante ao processo apresentado na Tabela 2.9 para casos sem repetição.

$$F_I < 2F_{(k-1)(n-1),nk(r-1),50\%} \tag{2.39}$$

Tabela 2.10: Síntese para comparação entre médias para duas classificações com repetição

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	F	F_α
Entre linhas	$SQL = \dfrac{\sum_{i=1}^{k} \frac{T_i^2}{nr} - \frac{T^2}{nkr}}{k-1}$	$k-1$	$S_L^2 = \dfrac{SQL}{k-1}$	$F_L = \dfrac{S_L^2}{S_R^2}$	$F_{k-1,nk(r-1),\alpha}$
Entre colunas	$SQC = \dfrac{\sum_{j=1}^{k} \frac{T_j^2}{kr} - \frac{T^2}{nkr}}{n-1}$	$n-1$	$S_C^2 = \dfrac{SQC}{n-1}$	$F_C = \dfrac{S_C^2}{S_R^2}$	$F_{n-1,nk(r-1),\alpha}$
Interação	$SQI = SQTr - SQL - SQC$	$(k-1)(n-1)$	$S_I^2 = \dfrac{SQI}{(k-1)(n-1)}$	$F_I = \dfrac{S_I^2}{S_R^2}$	$F_{(k-1)(n-1),nk(r-1),\alpha}$
Entre tratamentos	$SQTr = \sum_{i=1}^{k}\sum_{j=1}^{n} \frac{T_{ij}^2}{r} - \frac{T^2}{nkr}$	$nk-1$	$S_{Tr}^2 = \dfrac{SQTr}{nk-1}$	$F_{Tr} = \dfrac{S_{Tr}^2}{S_R^2}$	$F_{nk-1,nk(r-1),\alpha}$
Residual	$SQR = SQT - STr$	$nk(r-1)$	$S_R^2 = \dfrac{SQR}{nk(r-1)}$		
Total	$SQT = Q - \dfrac{T^2}{nkr}$	$nkr-1$			

2.4.6 Correlação e regressão

Esta sessão reduz-se no estudo de correlação e regressão de duas ou mais variáveis aleatórias quantitativas, onde cada variável é representada por um conjunto de dados amostrais. Correlação pode ser entendida como a existência de alguma relação entre as variáveis, e o quão elas se relacionam entre si. Já regressão é a técnica de exploração e inferência de uma correlação, onde a partir de um modelo matemático são estimadas equações que a descrevem.

Para ambos os casos é pertinente a localização dos pontos, a partir do conjunto amostral, em um plano ou espaço cartesiano. Dessa forma é possível analisar a tendência da correlação entre as variáveis e até mesmo prever o comportamento entre os pontos. Por exemplo, a partir de um diagrama de dispersão verifica-se visualmente se os pontos podem ser modelados por uma função linear, polinomial, logarítmica, dentre outras.

2.4.6.1 Correlação linear

O estudo de correlação será limitado, neste momento, a apenas duas variáveis aleatórias X e Y. Sendo uma correlação linear, os dados deverão estar relacionados com base numa função linear ou de 1º grau.

Situações em que, para maiores valores X temos maiores valores para Y, definimos como *correlação linear positiva* por comportar-se como uma função linear crescente. Ao passo que, para maiores valores X temos menores valores para Y, definimos como *correlação linear negativa* devido a semelhança com uma função linear decrescente. Há situações em que, a partir dos dados não é possível estabelecer nenhum tipo de correlação por causa da não adequação à um modelo matemático, podendo ser chamado de *correlação linear nula*. Entretanto, casos em que valores distintos de X, implicam em valores maiores ou menores de Y. Essa situação denominamos como *correlação não linear*, onde os dados das variáveis X e Y não podem ser relacionados linearmente. A Figura 2.9 ilustra essas quatro situações dentro do contexto de correlação linear.

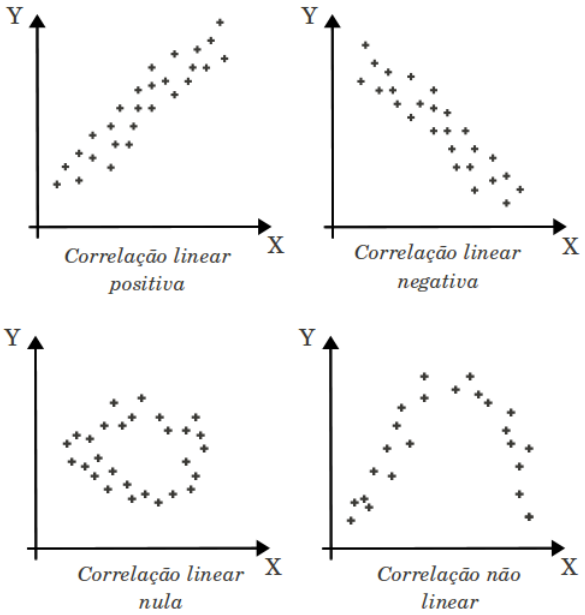

Figura 2.9: Casos de correlação linear, segundo Neto (2002)

Dessa forma, dizer que duas variáveis se correlacionam linearmente implica nos seus pontos estarem próximos ou distante a uma reta estimada. Em outras

palavras, verificar o quanto os pontos se aproximam da reta. Uma estimativa para esta verificação é a covariância entre as variáveis, definida pela equação 2.40.

$$S_{xy} = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{2.40}$$

Uma outra maneira de verificar uma correlação linear é através do coeficiente de *correlação linear de Pearson*, conforme Neto (2002). É definido pela relação da covariância entre as variáveis com o produto dos desvios padrões de ambas, como mostra a equação 2.41.

$$r = \frac{cov(x,y)}{S_x S_y} \tag{2.41}$$

onde:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \qquad e \qquad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

O *coeficiente de Pearson* é adimensional, compreendido no intervalo $-1 < r < 1$. Isso reduz os resultados a uma padronização, visto que os valores da covariância respeitam a mesma ordem de grandeza da amostra. Quanto $r \to -1$ implica em uma correlação negativamente perfeita, ao passo que $r \to 1$ corresponde a uma correlação positivamente perfeita. Em ambos os casos, r indica a tendência de correlação onde valores próximos a seus limitantes garante uma forte relação entre as variáveis.

2.4.6.2 Regressão linear

Uma regressão consiste na obtenção de um modelo que melhor se ajusta ao conjunto dos pontos amostrais de variáveis aleatórias, verificado se a correlação entre elas representam algum relação ou tendência. Nesse caso, regressão linear nada mais é que a estimação de uma reta que melhor se ajusta aos pontos resultantes de duas variáveis aleatórias. Deve ser assumido que uma das variáveis seja o domínio e a outra a imagem da função linear obtida. De vários métodos numéricos para a obtenção dessa reta. Neto (2002) ressalta o métodos dos quadrados mínimos, que tem como propósito reduzir a soma dos quadrados das distâncias de cada ponto à reta ajustada.

Teoricamente, uma regressão linear é obtida através da equação de reta conforme a expressão 2.42.

$$y = \alpha + \beta x \tag{2.42}$$

O valor de α será estimado pela variável a, assim como β pela variável b. Ainda de 2.42, os pontos que estarão contidos na reta obtida serão representados por $\widehat{y}$. Dessa forma, uma regressão linear será descrita conforme a expressão 2.43.

$$\widehat{y} = a + bx \tag{2.43}$$

Encontrar o modelo linear que melhor se ajusta aos pontos reduz-se, portanto, na determinação do coeficiente angular b e linear a da reta $\widehat{y}$, conforme 2.43. A dedução dos cálculos desses coeficientes são ilustrados em Neto (2002), não sendo o foco deste trabalho representar todas as demonstrações. Dessa forma seguem as definições de a e b na expressão 2.44, que resultarão na equação da reta estimada conforme 2.45.

$$b = \frac{S_{xy}}{S_{xx}} \qquad e \qquad a = \overline{y} - b\overline{x} \tag{2.44}$$

$$\therefore \qquad \widehat{y} = (\overline{y} - b\overline{x}) + \left(\frac{S_{xy}}{S_{xx}}\right) x$$

$$\therefore \qquad \widehat{y} = \left(\overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x}\right) + \left(\frac{S_{xy}}{S_{xx}}\right) x \tag{2.45}$$

2.4.6.3 Regressão linear múltipla

O estudo de regressão pode ser extendido para casos com mais de duas variáveis aleatórias. Temos portanto as variáveis independentes $X_1, X_2, \ldots, X_k$ e a variável dependente Y. Inicialmente, sejam 3 variáveis X_1, X_2 e Y. Com base na regressão linear simples vista anteriormente e representada pelo modelo de reta em 2.43, o modelo de $\widehat{y}$ para regressão linear múltipla pode ser definida conforme a equação 2.46.

$$\widehat{y} = a + b_1 x_1 + b_2 x_2 \tag{2.46}$$

Vale ressaltar que $\widehat{y}$ será uma função de duas variáveis, com domínio X_1 e X_2. Ao representar esse modelo simples não mais em um plano e sim num espaço cartesiano, a regressão linear obtida modelará um plano de estimativa. Os cálculos dos coeficientes que determinam a regressão são análogos aos apresentados no caso de regressão linear simples. Portanto, a será definido pela expressão 2.47:

$$a = \overline{y} - b_1\overline{x_1} - b_2\overline{x_2} \tag{2.47}$$

O coeficiente b descrito em 2.44 pode ser também definido como:

$$S_{x_1y} = S_{x_1x_1}b \quad \Rightarrow \quad S_{1y} = S_{11}b \tag{2.48}$$

De expressão 2.48 para o caso linear, em uma regressão múltipla teremos b_1 e b_2 para X_1 e X_2 respectivamente. Seus valores, portanto, serão definidos a partir da solução do sistema linear conforme 2.49.

$$\begin{cases} S_{1y} &= S_{11}b_1 + S_{12}b_2 \\ S_{2y} &= S_{21}b_1 + S_{22}b_2 \end{cases} \tag{2.49}$$

Ao generalizar uma regressão linear múltipla para k variáveis dependentes, o estimador $\widehat{y}$, a partir de 2.46, é definido pela expressão 2.50 a seguir:

$$\widehat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k \tag{2.50}$$

Analogamente, o coeficiente a pode ser generalizado conforme a equação 2.51.

$$a = \overline{y} - b_1\overline{x_1} - b_2\overline{x_2} - \ldots - b_k\overline{x_k} \tag{2.51}$$

O coeficiente b implicará em um sistema linear com k equações, com solução determinada por $(b_1, b_2, \ldots, b_k)$. O sistema linear resultante para o cálculo dos valores de b está descrito a seguir, onde pode ser resumido conforme a equação 2.52.

$$\begin{cases} S_{1y} &= S_{11}b_1 &+ S_{12}b_2 &+\cdots+ &S_{1k}b_k \\ S_{2y} &= S_{21}b_1 &+ S_{22}b_2 &+\cdots+ &S_{2k}b_k \\ \vdots &\quad\vdots &\quad\vdots &\ddots &\quad\vdots \\ S_{ky} &= S_{k1}b_1 &+ S_{k2}b_2 &+\cdots+ &S_{kk}b_k \end{cases}$$

$$S_{iy} = \sum_{i=1}^{k} S_{il}b_l, \quad (i = 1, 2, \ldots, k) \tag{2.52}$$

2.4.6.4 Correlação linear múltipla

Analogamente ao que foi feito com o estudo de regressão, é possível verificar a correlação entre k variáveis aleatórias. Tomando nesse momento a partir da regressão linear múltipla duas variáveis independentes X_1 e X_2, além da variável dependente Y, o coeficiente de correlação é determinado pela equação 2.53.

$$R = \sqrt{\frac{b_1S_{1y} + b_2S_{2y}}{S_{yy}}} \tag{2.53}$$

Cabe a análise de R estar compreendido entre $0 \leqslant R \leqslant 1$. Isso de deve ao fato da possibilidade de Y tem uma correlação positivamente perfeita com X_1 e uma correlação negativamente perfeita com X_2, ou vice-versa. Assim, dizer que $R < 0$ dificulta a análise da correlação entre as variáveis. Portanto, para uma correlação linear multipla, casos que $R \rightarrow 1$ implicará em forte correlação. A generalização do valor de R para k variáveis é definida pela equação 2.54.

$$R = \sqrt{\frac{b_1 S_{1y} + b_2 S_{2y} + \ldots + b_k S_{ky}}{S_{yy}}} \tag{2.54}$$

2.5 Comentários finais

Nesse capítulo foi possível consolidar os conhecimentos considerados fundamentais para a realização do presente trabalho. Embora seja de natureza teórica, a reflexão sobre o modelo de gerência FCAPS permitiu obtere uma visão panorâmica de toda a gestão de redes decomputadores, inclusive questões de performance. Adicionalmente, a discussão em torno do ciclo PDCA irá permitir, conforme será apresentado nos capítulos seguintes, a sistematização e definição das etapas da experimentação estatística.

No que tange o conhecimento técnico na adminstração de redes *Linux*, o protocolo *Simple Network Management Protocol* (SNMP) tem papel essencial na coleta de dados, e consequentemente a viabilidade da análise estatística no contexto de redes de computadores.

Por fim, toda a revisão em torno das ferramentas estatísticas faz-se necessário para a realização da análise proposta nesse trabalho. Além da ilustração de cada método estatístico, a revisão permitiu a consolidação do conhecimento teórico sobre o tema. Mesmo que nesse trabalho algumas técnicas não sejam utilizadas diretamente, o seu entendimento é pré requisito para o emprego de outras ferramentas estatísticas.

Capítulo 3

Metodologia e desenvolvimento

A metodologia e desenvolvimento do presente trabalho serão elaborados a partir da proposta do roteiro de planejamento de experimentos do ciclo PDCA, conforme discutido no capítulo anterior. Entretanto, serão abordados nesse momento as etapas de descrição geral do experimento, seleção da variável resposta, escolha de fatores, planejamento e execução do procedimento experimental.

3.1 Descrição geral do experimento

3.1.1 Coleção de idéias

O conceito de Redes da Próxima Geração (*Next Generation Network*) abrange o emprego de serviços de voz e vídeo, multimídia, compartilhamento de dados em formato texto ou gráfico, com razoável qualidade de serviço, segurança e custo reduzido (LIU; LIANG, 2009). É pertinente afirmar, complementarmente, que as tecnologias de acesso estão cada vez mais heterogêneas com distintas velocidades de conexão, ao qual representa outro fator na administração dessas redes de computadores.

Uma infraestrutura de rede é composta por diversos equipamentos ativos de rede, como *switches*, *switches* gerenciáveis, roteadores, pontos de acesso sem fio, além de servidores que garantem o funcionamento das diversas aplicações e serviços nesse ambiente. Para diferentes serviços e aplicações, diferentes métricas são analisadas e interpretadas de maneira que se garanta a gerência da performance.

Nessa discussão de como as variáveis podem se interagir, Leinwand e Conroy (1996) ilustram os seguintes casos:

- um servidor de arquivos, carga de processamento, percentual de uso do disco e utilização da placa de rede são informações pertinentes para análise de sua performance. Um processador com alta carga de processamento implica na lentidão na execução de processos do sistema, assim como leitura e escrita de dados em disco. Da mesma forma, a alta utilização de discos rígidos pode acarretar em queda da performance de acesso aos dados e risco de perda de informações;

- em dispositivos ou servidores encarregados de realizar o roteamento de conexões, métricas como carga de processamento, uso da memória, quantidade de pacotes trafegados, enfileirados e descartados permitem a análise e o entendimento de problemas em potencial, como congestionamento ou queda da performance da rede. A alta ou total utilização do *link* disponível para conexão implica em pacotes enfileirados, conforme a demanda do ambiente. Consequentemente, o sistema operacional responsável necessitará processar novamente o envio ou recepção desses pacotes. Isso exige maior recurso de processamento e memória do dispositivo, podendo inclusive ocasionar em perda de pacotes devido ao excessivo tempo que permaneceu enfileirado.

Em casos de servidores destinados a recursos de multimídia, é válida a análise das métricas de pacotes descartados e enfileirados, além da quantidade de informação trafegada. Para o serviço de videoconferência, por exemplo, não é interessante que pacotes cheguem com atrasos devido a um congestionamento na rede ou baixa performance de processamento de um dos dispositivos de roteamento. Isso, na prática, implica em visualização distorcidas e atrasos na voz e na imagem. Dessa forma é importante avaliar se não há perda de pacote devido à carga de processamento ou uso de memória, além da taxa de vazão de pacotes que representa a qualidade da videoconferência.

Leinwand e Conroy (1996) afirmam que, uma alta taxa de utilização do processador em um dispositivo aplicado à roteamento, não representa necessariamente queda de performance da conexão desde que o *link* de saída da rede não tenha grande taxa de utilização. Adicionalmente, analisar a carga de processamento e o uso de memória em dispositivos de roteamento nos leva à construção de algumas linhas de pensamento. A alta carga de processamento pode significar que o equipamento não esteja conseguindo tratar todas as conexões, seja pela elevada demanda da rede, seja por algum erro de configuração ou otimização. O consumo excessivo

da memória pode representar grande uso do *buffer*, ao qual implica em queda de performance.

Além da análise de como as variáveis se interagem em uma infraesutrura de rede, é plausível a discussão de outros pontos complementar à gestão de redes de computadores. Foi discutido no capítulo anterior a questão sobre performance, onde a gerência da rede era feita apenas pelo protocolo ICMP através do comando *ping*, durante o uso da ARPANET. O ICMP é um protocolo

> "usado pela implementação do protocolo IP de estações e roteadores para trocar informações de erro e controle, sinalizando situações especiais por meio de seus diversos tipos de mensagens"(ESR/RNP, 2005).

Na prática, o emprego do comando *ping* pelos administradores de rede está relacionado, na grande maioria dos casos, na verificação da resposta do *host* destino. Por questões de segurança, muitos equipamentos estão configurados de modo que não emitam todos os tipos de respostas previstas pelo protocolo ICMP. Essa prática de segurança inutiliza essa finalidade de uso do comando *ping*. Por outro lado, através de seus resultados como ilustrados na Figura 3.1, é possível obter algumas métricas estatísticas quanto ao desempenho da rede.

```
$ ping www.terra.com.br -c 5
PING www.terra.com.br (200.154.56.80) 56(84) bytes of data.
64 bytes from www.terra.com.br (200.154.56.80): icmp_seq=1 ttl=246 time=56.0 ms
64 bytes from www.terra.com.br (200.154.56.80): icmp_seq=2 ttl=246 time=56.1 ms
64 bytes from www.terra.com.br (200.154.56.80): icmp_seq=3 ttl=246 time=56.3 ms
64 bytes from www.terra.com.br (200.154.56.80): icmp_seq=4 ttl=246 time=57.3 ms
64 bytes from www.terra.com.br (200.154.56.80): icmp_seq=5 ttl=246 time=56.4 ms

--- www.terra.com.br ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4005ms
rtt min/avg/max/mdev = 56.014/56.469/57.336/0.506 ms
```

Figura 3.1: Saída do comando *ping*

Na última linha da resposta do comando *ping* temos os valores mínimo, médio, máximo e desvio padrão para o *round-trip-time* (RTT). ESR/RNP (2008) define o RTT como o tempo necessário para ida e volta do pacote ICMP, medido em milisegundos. Leinwand e Conroy (1996) ressalta que o *round-trip-time* é uma boa estimativa para o tempo total de resposta, devido ao fato da aplicação adicionar um tempo relativamente pequeno para processar o pacote do protocolo ICMP.

Entretanto, a abordagem do valor do tempo total de resposta não se aplica exclusivamente para o comando *ping*.

Observado o tempo total de resposta, definimos como latência todo o atraso agregado ao valor final, proveniente de aplicações e pontos de roteamento. Trivialmente percebe-se que toda comunicação terá uma latência, relacionando fortemente com a qualidade de uma conexão. Dessa forma a latência pode ser incrementada ou reduzida em função :

- do número de nós roteáveis entre os *hosts* remetente e destinatário;
- do *overhead* gerado pelas aplicações que fazem algum tipo de tratamento das conexões;
- da qualidade do meio físico que a conexão se propaga;
- do desempenho de servidores, equipamentos e dispositivos de redes;
- do planejamento incorreto da rede, como erro no endereçamento dos dispositivos ou cascateamento de ativos de rede, implicando em gargalos e colisão de pacotes.

Denominamos como *jitter* a variação dos valores de latência de uma conexão. Toda conexão possui *jitter* onde quanto menor for o valor, mais estável será a conexão estabelecida. Mais uma análise a partir da Figura 3.1 mostra que o desvio padrão dos valores de RTT, considerando que pode ser estimado como tempo total de resposta, representa o *jitter* de um conjunto de pacotes enviados. Adicionalmente uma conexão, como exemplo videoconferência, não terá grande problemas de funcionalidades caso possua uma latência alta desde que tenha o *jitter* estabilizado.

Tanenbaum (1997) relaciona ainda mais alguns elementos que auxiliam na gerência do desempenho de uma rede:

- probabilidade de falha no estabelecimento da conexão: chance de uma conexão não ser estabelecida dentro de um dado intervalo de tempo;
- *throughput*: quantidade de dados, em *bytes* ou *bits*, trafegados por segundo em um dado intervalo de tempo;
- taxa de erros residuais: percentual de mensagens perdidas ou com erros;
- prioridade: estabelecimento de ordem no tratamento das conexões;

- resiliência: autonomia à camada de transporte para encerramento de conexões oriundos de congestionamento ou problemas internos.

Em ambientes críticos com elevado grau de congestionamento, caracterizados principalmente pela saturação dos equipamentos que provêm conexão, adotam-se algoritmos para controle de congestionamento nativos no *kernel* GNU/Linux. Implementados pela variável *net.ipv4.tcp_congestion_control* do *sysctl*, os algoritmos que passíveis de implementação em um servidor GNU/Linux são: *reno*, *vegas*, HSTCP, STCP, *cubic*, *westwood* e *fast* TCP (SOUSA, 2007). Não é o foco do presente trabalho detalharmos e implementarmos um desses algoritmo visto a sua complexidade. Além de todo o estudo e análise da adoção desses algoritmos, é pertinente a previsão do impacto da sua implantação em todo o sistema operacional. Adicionalmente, a implementação desses algoritmos estão submetidos à demandas extremas, com total uso dos recursos dos equipamentos que garantem a conexão à *internet*.

3.1.2 Ambiente analisado

Conforme apresentado na introdução desse trabalho, o objeto de estudo será o Campus II do Centro Federal de Educação Tecnológica de Minas Gerais. Respeitadas as questões de sigilo e segurança quanto a topologia e configuração da rede de dados da instituição, a Figura 3.2 ilustra um breve esboço sobre o ambiente analisado nos experimentos estatísticos.

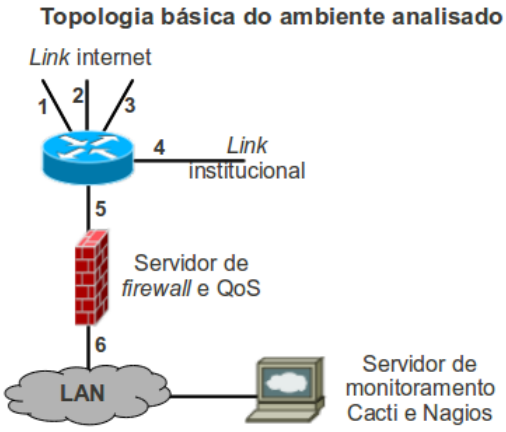

Figura 3.2: Topologia básica do ambiente analisado

O *link* de internet é composto no roteador por 3 seriais com 2 Mbps de largura de banda cada uma (conexões 1, 2 e 3 na Figura 3.2), o que integra um *link* total de 6 Mbps. A 4ª serial contempla o *link* de 1 Mbps com o Campus I (conexão 4), alocado exclusivamente para serviços institucionais. A união entre roteador e rede local (LAN) é garantida por uma conexão do tipo *ethernet*. No entanto, a colocação de um servidor nessa posição (conexão 5 e 6) permite que o ambiente esteja protegido por um *firewall*, além das adoções de controle do tráfego para uso racionalizado do *link* de dados e implementação de QoS (*Quality of Service*).

Na rede local, um dos servidores em produção está dedicado para o monitoramento da rede, sob o ponto de vista da gerência de falhas e performance. Na prática, as ferramentas Nagios e Cacti são responsáveis, respectivamente, pela detecção e alertas de situações anômalas de funcionamento e registro de dados sobre o desempenho de servidores, equipamentos e ativos de rede. Em especial o Cacti, a ferramenta utiliza o protocolo SNMP, versão 2c, para coleta de informações sobre a performance da rede. Analisando o modelo de funcionamento do SNMP a partir da topologia ilustrada na figura 3.2, o servidor de monitoramento atua como gerente do protocolo SNMP, enquanto que roteador, servidor de *firewall* e demais equipamentos estão configurados como agentes.

Sobre o funcionamento do servidor de *firewall*, cabe a observação quanto as interfaces de rede. Como as interfaces estão configuradas em modo *bridge*, a conexão 5 e 6 exibidas na Figura 3.2 são de certa forma equivalentes. Os valores analisados são idênticos, de maneira que o estudo de apenas uma interface seja suficiente.

3.1.3 Metas dos experimentos

Na introdução deste trabalho definimos como meta do trabalho a aplicação de ferramentas estatística como intervalo de confiança, teste de hipótese e análise de variância e verificação de possíveis correlações entre variáveis de rede. Adicionalmente, o trabalho como um todo consiste na meta de propor um procedimento para experimento estatístico no âmbito de redes de computadores.

No entanto, definimos como metas dos experimentos:

- estimar o parâmetro média, dentro de uma margem de erro, para variáveis de carga de processamento, uso de memória, *throughput* e número de pacotes. Essa estimação será aplicada tanto no servidor de *firewall* quanto no roteador

do ambiente analisado, levando em consideração a quantidade de interfaces de redes de cada um dos equipamentos;

- analisar a variância do *throughput*, número de pacotes, pacotes com erros e pacotes descartados, considerando classificações como sentido de tráfego (*download* e *upload*) e seriais do roteador;
- verificar a possibilidade de existência de correlação linear, validada por teste de hipótese, entre carga de processamento e uso de memória, e número de pacotes e *throughput*, tanto do roteador quanto do servidor de *firewall*. Além disso, verificar a correlação da carga de processamento e uso de memória entre roteador e servidor;
- obter regressões da carga de processamento e do uso de memória em função do número de pacotes e *throughput*. Concomitantemente verificar regressão da carga de processamento, uso de memória, número de pacotes e *throughput* em função de um dado intervalo de tempo.

3.2 Seleção da variável resposta

Quando se trata de experimentação estatística é de extrema importância termos em mente qual estimador de fato queremos analisar. Em outras palavras, para cada variável analisada devemos definir se a inferência será sobre a média, valor máximo ou desvio padrão. Analisar o valor máximo implica averiguar eventuais picos na utilização de recursos em um determinado intervalo de tempo. Manipular desvio padrão, em redes de computadores, significa obter conclusões em questões que envolvam latência e *jitter*. O estudo da média implica, dessa forma, avaliar todo o comportamento esperado de uma variável dada uma amostragem significativa.

Nos nossos experimentos consideraremos, de maneira geral, o valor médio de cada variável integralizada a cada 5 minutos. Na faremos nenhum análise, *a priori*, de outros estimadores como valor máximo, variância ou desvio padrão. Com base nas metas dos experimentos relacionadas na sessão anterior dividiremos os nossos experimentos em grupos conforme a Tabela 3.1.

3.2.1 Grupo 1: Estimativa de parâmetros

Novamente baseado nas metas dos experimentos relacionadas na sessão anterior, as Tabelas 3.2, 3.3, 3.4 e 3.5 apresentam a definição das variáveis para proble-

Tabela 3.1: Tabela dos grupos de experimentos

Grupo	Experimento
1	Estimativa de parâmetros
2	Análise de variância
3	Correlação
4	Regressão

mas de estimação de parâmetros envolvendo a média global de todas as amostras (grupo 1). No entanto, no âmbito da gerência de redes, é pertinente avaliarmos eventuais picos dos recursos através do valor máximo de cada variável. Para que possamos construir os intervalos de confiança para os valores máximos, estimaremos a média dessa variável para manutenção do seu comportamento conforme distribuição normal. As Tabelas 3.6, 3.7, 3.8 e 3.9 definem as variáveis dos valores máximos para carga de processamento, uso de memória, *throughput* e número de pacotes.

Cabe ressaltar que, conforme apresentado na topologia do ambiente analisado (Figura 3.2), não faremos a estimativa de parâmetros para cada um dos *links* seriais de internet. O que envolve interface de rede no equipamento de roteamento, analisaremos o *link* institucional e o somatório das 3 seriais que compõem o *link* de internet. Adicionalmente, consideraremos em nossos experimentos o tráfego de toda a LAN a partir do servidor de *firewall* e não da interface *ethernet* do roteador.

Tabela 3.2: Definição das variáveis para o grupo de experimentos 1: estimativa da média da carga de processamento

Experimento	Variável	Reamostragem por *bootstrapping*	Descrição
1-1	L_R	L_R^*	Média da carga de processamento do roteador integralizada a cada 5 minutos, contida em cada intervalo de hora
1-2	L_S	L_S^*	Média da carga de processamento do servidor de *firewall* integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.3: Definição das variáveis para o grupo de experimentos 1: estimativa da média do uso de memória

Experimento	**Variável**	**Reamostragem por *bootstrapping***	**Descrição**
1-3	M_R	M_R^*	Média do uso de memória do roteador integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-4	M_S	M_S^*	Média do uso de memória do servidor de *firewall* integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.4: Definição das variáveis para o grupo de experimentos 1: estimativa da média do *throughput*

Experimento	**Variável**	**Reamostragem por *bootstrapping***	**Descrição**
1-5	T_{Rd1}	T^*_{Rd1}	Média da soma da vazão de pacotes dos *links* de internet do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-6	T_{Ru1}	T^*_{Ru1}	Média da soma da vazão de pacotes dos *links* de internet do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-7	T_{Rd2}	T^*_{Rd2}	Média da vazão de pacotes do *link* institucional do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-8	T_{Ru2}	T^*_{Ru2}	Média da vazão de pacotes do *link* institucional do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-9	T_{Sd}	T^*_{Sd}	Média da vazão de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-10	T_{Su}	T^*_{Su}	Média da vazão de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.5: Definição das variáveis para o grupo de experimentos 1: estimativa da média do número de pacotes

Experimento	**Variável**	**Reamostragem por *bootstrapping***	**Descrição**
1-11	P_{Rd1}	P^*_{Rd1}	Média da soma do número de pacotes dos *links* de internet do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-12	P_{Ru1}	P^*_{Ru1}	Média da soma do número de pacotes dos *links* de internet do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-13	P_{Rd2}	P^*_{Rd2}	Média do número de pacotes do *link* institucional do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-14	P_{Ru2}	P^*_{Ru2}	Média do número de pacotes do *link* institucional do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-15	P_{Sd}	P^*_{Sd}	Média do número de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-16	P_{Su}	P^*_{Su}	Média do número de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.6: Definição das variáveis para o grupo de experimentos 1: estimativa da média dos valores máximos da carga de processamento

Experimento	Variável	Reamostragem por *bootstrapping*	Descrição
1-17	L_{Rmax}	L^*_{Rmax}	Média estimada dos máximos da carga de processamento do roteador integralizada a cada 5 minutos, contida em cada intervalo de hora
1-18	L_{Smax}	L^*_{Smax}	Média estimada dos máximos da carga de processamento do servidor de *firewall* integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.7: Definição das variáveis para o grupo de experimentos 1: estimativa da média dos valores máximos do uso de memória

Experimento	Variável	Reamostragem por *bootstrapping*	Descrição
1-19	M_{Rmax}	M^*_{Rmax}	Média estimada dos máximos do uso de memória do roteador integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-20	M_{Smax}	M^*_{Smax}	Média estimada dos máximos do uso de memória do servidor de *firewall* integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.8: Definição das variáveis para o grupo de experimentos 1: estimativa da média dos valores máximos do *throughput*

Experimento	**Variável**	**Reamostragem por *bootstrapping***	**Descrição**
1-21	$T_{Rd1-max}$	$T^*_{Rd1-max}$	Média estimada dos máximos da soma da vazão de pacotes dos *links* de internet do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-22	$T_{Ru1-max}$	$T^*_{Ru1-max}$	Média estimada dos máximos da soma da vazão de pacotes dos *links* de internet do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-23	$T_{Rd2-max}$	$T^*_{Rd2-max}$	Média estimada dos máximos da vazão de pacotes do *link* institucional do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-24	$T_{Ru2-max}$	$T^*_{Ru2-max}$	Média estimada dos máximos da vazão de pacotes do *link* institucional do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-25	T_{Sd-max}	T^*_{Sd-max}	Média estimada dos máximos da vazão de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-26	T_{Su-max}	T^*_{Su-max}	Média estimada dos máximos da vazão de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.

Tabela 3.9: Definição das variáveis para o grupo de experimentos 1: estimativa da média dos valores máximos do número de pacotes

Experimento	Variável	Reamostragem por *bootstrapping*	Descrição
1-27	$P_{Rd1-max}$	$P^*_{Rd1-max}$	Média estimada dos máximos da soma do número de pacotes dos *links* de internet do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-28	$P_{Ru1-max}$	$P^*_{Ru1-max}$	Média estimada dos máximos da soma do número de pacotes dos *links* de internet do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-29	$P_{Rd2-max}$	$P^*_{Rd2-max}$	Média estimada dos máximos do número de pacotes do *link* institucional do roteador, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-30	$P_{Ru2-max}$	$P^*_{Ru2-max}$	Média estimada dos máximos do número de pacotes do *link* institucional do roteador, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-31	P_{Sd-max}	P^*_{Sd-max}	Média estimada dos máximos do número de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *download* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.
1-32	P_{Su-max}	P^*_{Su-max}	Média estimada dos máximos do número de pacotes do tráfego *ethernet* no servidor de *firewall*, referente ao *upload* para a LAN, integralizada a cada 5 minutos, contida em cada intervalo de hora.

Todas as variáveis definidas para esse grupo de experimentos são aleatórias com população infinita. Veremos na sessão seguinte a discussão sobre a classificação das variáveis em intervalos de tempo, critérios para o período de medição e fatores externo que podem influenciar, de alguma maneira, na função densidade probabilística. Para esse último elemento, cabe a possibilidade das variáveis de rede não se apresentar conforme distribuição normal, o que na prática inviabilizaria a estimação de parâmetros conforme previsto nos experimentos.

Para que esses dados sejam corrigidos e ajustados para uma distribuição normal, aplicaremos em toda a amostra de dados coletada a técnica de *bootstrapping*. Na teoria, o *bootstrapping* é um procedimento estatístico empregado no ajuste de parâmetros para uma distribuição normal, em função ou do reduzido número de amostras, ou da não apresentação dos dados como variável Gaussiana. Matematicamente trata-se da aplicação do teorema do limite central e combinação linear para a reamostragem de um variável. Na prática, o *bootstrapping* equivale à repetição do experimento de maneira que mais dados sejam obtidos, devido à escassez do número de elementos que compõem a amostra.

Seja uma população $P = (p_1, p_2, \ldots, p_N)$ e uma amostra $A = (a_1, a_2, \ldots, a_n)$ com $n \ll N$. Retira-se uma nova amostra A_1 de n elementos com repetição a partir da amostra A, e desses elementos calcula-se o estimador desejado, no caso a média, conforme a equação 3.1. Calculado o estimador média e obtido o elemento a_1^*, repete-se esses procedimentos por B vezes. Dessa forma, a amostra ajustada será $A^* = (a_1^*, a_2^*, \ldots, a_B^*)$ (CARRANO; WANNER; TAKAHASHI, 2011).

$$a_i^* = \frac{a_{i1} + a_{i2} + \ldots + a_{in}}{n} \tag{3.1}$$

3.2.2 Grupo 2: Análise de variância

Dando continuidade a seleção e definição de variáveis respostas, classificaremos as informações de tráfego de internet, como *throughput*, vazão de pacotes, pacotes com erros e pacotes descartados, conforme a interface serial que está sendo tratada e o seu sentido de tráfego (*download* e *upload*). Isso nos permite avaliar a existência de diferenças significativas, para cada um dessas variáveis, conforme classificação apresentada.

Não utilizaremos a técnica de *bootstrapping* para ajuste das variáveis, visto que os valores não necessitam ter uma distribuição normal para este experimento. Dessa forma, faremos uso de dados colhidos através do Cacti para compor a nossa amostra.

A Tabela 3.10 apresenta a definição dos experimentos que envolvem análise de variância.

Tabela 3.10: Definição das variáveis para o grupo de experimentos 2: análise de variância

Experimento	Descrição
2-1	Análise da variância do *throughput* do *link* de internet do roteador, classificado pelas 3 interfaces seriais do roteador e sentido de tráfego (*download* e *upload*). Será selecionado o valor máximo de cada variável de cada dia, dentro do período analisado.
2-2	Análise da variância do número de pacotes do *link* de internet do roteador, classificado pelas 3 interfaces seriais do roteador e sentido de tráfego (*download* e *upload*). Será selecionado o valor máximo de cada variável de cada dia, dentro do período analisado.
2-3	Análise da variância do número de pacotes com erros do *link* de internet do roteador, classificado pelas 3 interfaces seriais do roteador e sentido de tráfego (*download* e *upload*). Será selecionado o valor máximo de cada variável de cada dia, dentro do período analisado.
2-4	Análise da variância do número de pacotes descartados do *link* de internet do roteador, classificado pelas 3 interfaces seriais do roteador e sentido de tráfego (*download* e *upload*). Será selecionado o valor máximo de cada variável de cada dia, dentro do período analisado.

3.2.3 Grupo 3: Correlação

Nos problemas que envolvem correlação, consideraremos as variáveis vazão de pacotes e número de pacotes como independentes. Embora estatisticamente definimos nos experimentos anteriores que as variáveis carga de processamento e uso de memória são variáveis aleatórias independentes, no contexto da correlação linear vamos encará-las como dependentes do *throughput* e número de pacotes. Em outras palavras, tanto o uso do processador como o consumo de memória do roteador e servidor de *firewall* são determinados pelo *throughput* e número de pacotes.

A Tabela 3.11 descreve não só o experimento que será realizado, como a definição das variáveis que serão correlacionadas. Será apresenta uma variável auxiliar t (tempo), ao qual fará o papel de índice para composição dos pares ordenados das variáveis em estudo de cada correlação.

Assim como as análises de variâncias, esse grupo de experimento não empregará o *bootstrapping*, pela mesma razão de não necessitar que as variáveis de rede tenham distriuição normal. Após analisar a correlação linear, cada experimento terá seu respectivo teste de hipótese com o intuito de validar a existência de correlação.

Uma correlação definida nesse grupo de experimentos, ao qual cabe destaque, é entre a carga de processamento e o uso de memória entre roteador e servidor de *firewall*. Como são entidades físicas distintas dentro do ambiente analisado, será possível analisar se a reação de cada equipamento são correlacionadas em um mesmo instante de tempo.

Tabela 3.11: Definição das variáveis para o grupo de experimentos 3: correlação

Experimento	Variáveis	Descrição
3-1	T_{Rd1} e P_{Rd1}	Correlação entre o *throughput* e número de pacotes do *link* de internet do roteador, considerando o sentido de tráfego externo-interno (*download*). Será considerada toda toda a amostra representativa do período analisado.
3-2	T_{Ru1} e P_{Ru1}	Correlação entre o *throughput* e número de pacotes do *link* de internet do roteador, considerando o sentido de tráfego interno-externo (*upload*). Será considerada toda toda a amostra representativa do período analisado.
3-3	T_{Rd2} e P_{Rd2}	Correlação entre o *throughput* e número de pacotes do *link* institucional do roteador, considerando o sentido de tráfego externo-interno (*download*). Será considerada toda toda a amostra representativa do período analisado.
3-4	T_{Ru2} e P_{Ru2}	Correlação entre o *throughput* e número de pacotes do *link* institucional do roteador, considerando o sentido de tráfego interno-externo (*upload*). Será considerada toda toda a amostra representativa do período analisado.
3-5	T_{Sd} e P_{Sd}	Correlação entre o *throughput* e número de pacotes do *link ethernet* do servidor de *firewall*, considerando o sentido de tráfego externo-interno (*download*). Será considerada toda toda a amostra representativa do período analisado.
3-6	T_{Su} e P_{Su}	Correlação entre o *throughput* e número de pacotes do *link ethernet* do servidor de *firewall*, considerando o sentido de tráfego interno-externo (*upload*). Será considerada toda toda a amostra representativa do período analisado.
3-7	L_R e M_R	Correlação entre a carga de processamento o uso de memória do roteador. Será considerada toda toda a amostra representativa do período analisado.
3-8	L_S e M_S	Correlação entre a carga de processamento o uso de memória do servidor de *firewall*. Será considerada toda toda a amostra representativa do período analisado.
3-9	L_R e L_S	Correlação da carga de processamento entre o roteador e o servidor de *firewall*. Será considerada toda toda a amostra representativa do período analisado..
3-10	M_R e M_S	Correlação do uso de memória entre o roteador e o servidor de *firewall*. Será considerada toda toda a amostra representativa do período analisado.

3.2.4 Grupo 4: Regressão

Conforme visto na revisão bibliográfica, os problemas de regressão linear podem ser aplicados nos casos simples, na obtenção de uma função linear com domínio composto por uma única variável. Adicionalmente, essa variável necessariamente não pode apresentar resíduos, ou seja, seus valores são estatisticamente não-aleatórios. Na prática seria pré estabelecermos valores para número de pacotes (P) e coletar os respectivos dados para carga de processamento (L), quando deseja-se encontrar, por exemplo, a regressão $L = f(P)$.

Dessa forma, para os casos de regressão simples, a variável X (abcissa) será representada pelo tempo (t) justamente pelo fato de não haver erros em virtude do mecanismo de coleta do Cacti. As variáveis de redes estudadas até então serão regredidas em função de t, tanto para o roteador como para o servidor de *firewall*. Embora não fora apresentado na revisão bibliográfica, faremos além da regressão linear a regressão exponencial e logarítmica através do *BrOffice.org*, e a regressão polinomial conforme Neto (2002).

Cabe ressaltar que para chegarmos aos modelos de regressão utilizaremos as médias de cada variável estimada a partir do grupo de experimentos 1 (estimativa de parâmetros), classificadas conforme intervalo de hora.

A Tabela 3.12 apresenta o conjunto de experimentos envolvendo problemas de regressão.

Tabela 3.12: Definição das variáveis para o grupo de experimentos 4: regressão simples

Experimento	Regressão	Descrição
4-1	$\overline{L}_R(t)$	Regressão da média da carga de processamento do roteador em função do tempo, durante todo o período diário de coleta de dados, com uso do parâmetro estimado L_R a cada intervalo de hora.
4-2	$\overline{M}_R(t)$	Regressão da média do uso de memória do roteador em função do tempo, durante todo o período diário de coleta de dados, com uso do parâmetro estimado M_R a cada intervalo de hora.
4-3	$\overline{C}_S(t)$	Regressão da média da carga de processamento do servidor de *firewall* em função do tempo, durante todo o período diário de coleta de dados, com uso do parâmetro estimado L_S a cada intervalo de hora.
4-4	$\overline{M}_S(t)$	Regressão da média do uso de memória do servidor de *firewall* em função do tempo, durante todo o período diário de coleta de dados, com uso do parâmetro estimado M_S a cada intervalo de hora.
4-5	$\overline{T}_{Sd}(t)$	Regressão da média do *throughput* do tráfego *ethernet* do servidor de *firewall*, sentido (*download*), em função do tempo durante todo o período diário de coleta de dados, com uso do parâmetro estimado T_{Sd} a cada intervalo de hora.
4-6	$\overline{T}_{Su}(t)$	Regressão da média do *throughput* do tráfego *ethernet* do servidor de *firewall*, sentido (*upload*), em função do tempo durante todo o período diário de coleta de dados, com uso do parâmetro estimado T_{Su} a cada intervalo de hora.
4-7	$\overline{P}_{Sd}(t)$	Regressão da média do número de pacotes do tráfego *ethernet* do servidor de *firewall*, sentido (*download*), em função do tempo durante todo o período diário de coleta de dados, com uso do parâmetro estimado P_{Sd} a cada intervalo de hora.
4-8	$\overline{P}_{Su}(t)$	Regressão da média do número de pacotes do tráfego *ethernet* do servidor de *firewall*, sentido (*upload*), em função do tempo durante todo o período diário de coleta de dados, com uso do parâmetro estimado P_{Su} a cada intervalo de hora.

3.3 Escolha de fatores e seus níveis

Um dos artifícios utilizados para a realização dos experimentos descritos na sessão anterior é a classificação das variáveis conforme intervalo de hora. Ao longo de um período de 24 horas, uma estrutura de rede de computadores possui diferentes comportamentos, determinados pela demanda que o ambiente atende. Dessa forma, para que melhor possamos analisar esses comportamentos, as variáveis serão classificadas em intervalos de hora.

A partir de uma análise descritiva do comportamento atual da rede, percebe-se o maior uso dos recursos dentre os horários de 7h e 21h. A Figura 3.3 evidencia esse comportamento, ilustrando de maneira descritiva o *throughput* da conexão *ethernet* a partir do roteador. Obviamente que em outros ambientes e outras situações, tal como provedor de acesso, o comportamento provavelmente será diferente. No entanto, por se tratar de uma instituição de ensino, esse intervalo corresponde à período de aula e realização de atividades administrativas e acadêmicas. Portanto, a classificação das variáveis respeitará os intervalos horários $(7,8],(8,9],\ldots,(20,21]$. No gráfico ilustrado em 3.3, a linha azul representa o *download* do *link* enquanto que a área em verde significa o *upload*.

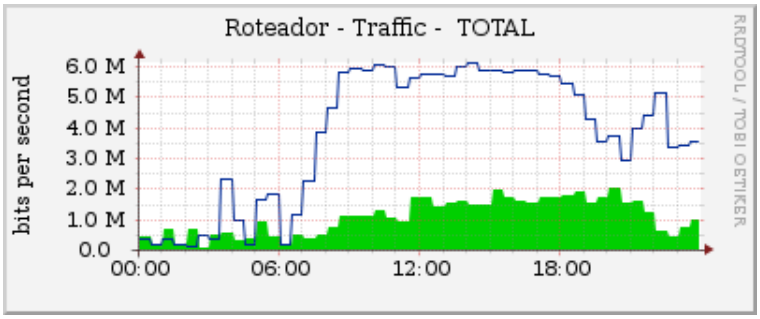

Figura 3.3: Panorama geral do comportamento da rede, a partir do *throughput*

Quanto ao período de medição, os dados coletados corresponderão às duas primeiras semanas letivas de 2011. Não faz sentido analisar, no nosso contexto de gerência de performance, o comportamento da rede durante as férias escolares. Nesse período, todo o ambiente não estará operando dentro de sua demanda total. O mesmo raciocínio se aplica na delimitação do nosso período de coleta apenas para os dias escolares, não havendo relevância a análise do ambiente de rede em finais de semana e feriados. Dessa forma, o período de medição que irá compor as amostras dos experimentos estatísticos corresponde entre os dias 6 e 19 de fevereiro de 2011, com exceção dos finais de semanas.

Outro assunto listado na sessão anterior, quanto a realização dos experimentos, refere-se aos fatores externos de uma rede de computadores. Em um ambiente de rede, diversos são os eventos inesperados que podem ocorrer durante o período de coleta de dados. Esses eventos correspondem desde uma maior solicitação de acesso de um *host*, ou de um nó na rede, até mesmo incidentes de segurança ou funcionamento inapropriado de equipamentos que, de alguma maneira, distorcem todo o comportamento do ambiente. Controlar eventos não previstos na rede não é uma tarefa simples, ao qual implicaria em mecanismos de segurança totalmente restritos, com o risco de prejudicar o desempenho e funcionamento de outras demandas já existentes na rede.

Estatisticamente esses fatores externos caracterizam a aleatoriedade de um ambiente de rede, representada pelas variáveis aleatórias. Por outro lado, existe o desafio em garantir que todas essas variáveis analisadas tenham distribuição normal. Afinal, as variáveis de ambiente de rede podem ser, de fato, totalmente aleatórias, ter distribuição distorcida em função dos fatores externos, e até mesmo ter a distribuição normal desejada para estimação de parâmetros. Angelis (2003) ilustra essa dificuldade em tratar as variáveis de rede como variáveis Gaussianas. Conforme citado anteriormente, utilizaremos no nosso trabalho a técnica de *bootstrapping* para ajuste dos dados da amostra.

3.4 Planejamento do procedimento experimental

A coleta de dados para realização dos experimentos será feita através do SNMP, protocolo apresentado no capítulo 2.3 desse trabalho. No entanto, a utilização nativa e isolada do SNMP não garante o tratamento e apresentação das amostras para análise estatística. Os objetos de interesse para o trabalho geridos pelo protocolo, conforme sua estrutura de gerenciamento de informação (SMI), são variáveis do tipo *Counter*. Obter esses dados simplesmente pelo comando *snmpwalk* ou *snmpget* implica em criação de *script* adicional para tratamento da variável, além da elaboração de mecanismos para sistematização dos períodos de coleta.

O ambiente analisado do CEFET-MG, onde ocorrerão os experimentos estatísticos, possui o Cacti como ferramenta para gerência de performance. Por a ferramenta já sistematizar a coleta, armazenamento e construção da base histórica de dados, conforme abordaremos na sessão 3.4.1, utilizaremos os valores geridos pelo Cacti para compor nosso conjunto amostral. Outro ganho seria a possibilidade de alinhar os resultados obtidos com os experimentos estatísticos junto ao

tratamento descritivo que a ferramenta faz, considerando o fato do Cacti ser de uso cotidiano entre os administradores da rede.

Como desvantagem nessa decisão citamos a redução da precisão dos valores integralizados. Embora seja comum a integralização das variáveis em intervalos de 5 minutos por várias ferramentas de gerência de performance, o adoção do *script* adicional para coleta de dados flexibiliza a minimização esse intervalo. Considerando que uma integralização para coleta de dados seja em um intervalo de 10 segundos, os valores obtidos representariam o comportamento da rede de maneira mais instantânea.

A seguir relacionamos as etapas para a coleta e realização das análises estatísticas, em conformidade com a definição das variáveis respostas apresentada na sessão 4.4, e com o método de coleta das amostras a partir do Cacti.

1. **Recuperação dos dados**: Essa etapa inicial consiste na recuperação dos dados armazenados pelo Cacti a partir dos mecanimos de *backup* do ambiente. A ferramenta armazena os 600 últimos registros em seu arquivo *.rra* das variáveis integralizadas a cada 5 minutos, o que corresponde a um período total de monitoramento de 50 horas, ou 2 dias e 2 horas. Como definimos na sessão 3.3 que o período de coleta terá 14 dias, faz-se necessário a recuperação dos demais arquivos *.rra*;

2. **Exportação dos dados**: Após a recuperação faremos a exportação dos dados armazenados nos arquivos *.rra* para o formato *.xml*, com o objetivo de acessibilizar a leitura dos valores. Essa exportação será feita a partir da própria ferramenta RRDtool[1] com o uso da função *xport*;

3. **Composição da amostra**: Nesse momento faremos a composição de toda a amostra a partir dos arquivos *.xml* para um banco de dados MySQL[2], a partir de um *script* próprio escrito em Perl[3] (*Practical Extraction and Report Language*). A escolha do armazenamento de todos os dados em um sistema do tipo SGBD como o MySQL fundamenta-se pela sistematização e facilidade de consulta e obtenção dos valores;

4. ***Bootstrapping***: A técnica de reamostragem *bootstrapping* será implementada no *software* científico para computação numérica Scilab[4], de maneira

[1] RRDtool: `http://oss.oetiker.ch/rrdtool/`
[2] MySQL: `http://www.mysql.com/`
[3] Perl: `http://www.perl.org/`
[4] Scilab: `http://www.scilab.org/`

que já constitua o ambiente para realização de grande parte das análises estatísticas a serem realizadas na próxima etapa;

5. **Análise estatística**: Essa etapa final consiste na aplicação efetiva das ferramentas estatísticas apresentadas na sessão 2.4 em todo o conjunto amostral tratado nas etapas anteriores. O Scilab será adotado para realização dos experimentos de estimação de parâmetros, análise de variância e correlação linear. Sobre os problemas de regressão será utilizadoalém do Scilab, o *BrOffice.org*.

3.4.1 Funcionamento básico do Cacti

Considerando o fato de que o Cacti proverá todos os dados para composição do conjunto amostral dos experimentos do trabalho é plausível apresentar brevemente o seu funcionamento. A ferramenta nada mais que é um *front-end* para exibição de dados descritivos de equipamentos e ativos de rede, com foco na gerência de performance. O Cacti é escrito predominatemente em PHP[5], ao qual é constituído pelo protocolo SNMP, banco de dados MySQL, aplicativo RRDtool, e servidor de arquivos Apache[6]. A Figura 3.4 ilustra de maneira simplifica o funcionamento da ferramenta e a integração dos seus componentes.

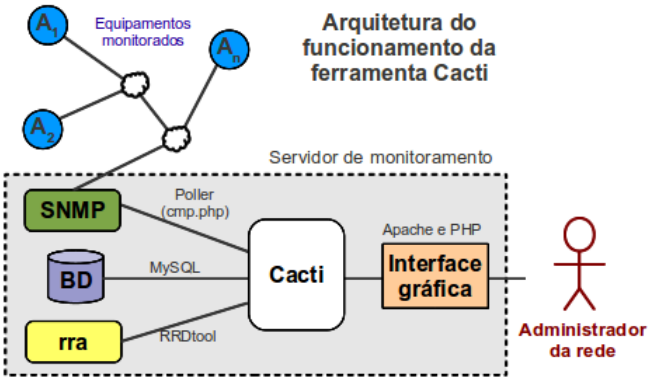

Figura 3.4: Arquitetura do funcionamento da ferramenta Cacti

Inicialmente é preciso ter em cada equipamentos monitorado o SNMP instalado e operando como agente. No servidor de monitoramento o SNMP está configurado como gerente, de modo que o arquivo *cmd.php* realiza todo o procedimento

[5]PHP: http://www.php.net/
[6]Apache: http://www.apache.org/

de coleta de dados pelo protocolo. Para o funcionamento do Cacti é indiferente a configuração do SNMP na versão 2c ou 3, o que na prática trata-se de uma decisão de projeto e administração da rede.

Com posse desses dados o Cacti utiliza o RRDtool (*Round Robin Database tool*) para duas finalidades essenciais: armazenamento de todos os dados coletados no formato *.rra* (*Round Robin Archives*), de maneira compacta e sistematizada; e renderização dos gráficos descritivos das variáveis de redes a partir dos arquivos *.rra*. O período de coleta e armazenamento dos dados, por padrão, é de 5 minutos podendo ser reajustado, preferenciamente, no momento de sua instalação.

A configuração de todo o ambiente, *data sources* de dispositivos monitorados, *templates* de gráficos, dentre outros, é armazenada no base de dados MySQL. Adicionalmente a ferramenta permite que seja configurada por uma interface gráfica escrita em PHP, suportado pelo servidor de página Apache. Essa versatilidade de acesso provida pela interface aos pelos administradores da rede, via navegador *web*, permite maior flexibilidade na gerência de performance da rede.

3.5 Realização do experimento

Nessa sessão abordaremos os passos necessários para a realização de todos os experimentos definidos nesse presente trabalho. Inicialmente trataremos a etapa de restauração dos arquivos *.rra*, extração dos dados do formato *.rra* para *.xml* com a ferramenta *rrdtool*, e a exportação da amostra para um banco de dados MySQL. Como grande parte dos experimentos depende da técnica de *bootstrapping*, exibiremos a função criada no Scilab para realização dessa reamostragem e seu efeito em comparação com a amostra original. Por fim, a ilustração dos *scripts* em Scilab para a realização dos experimentos estatísticos propriamente ditos.

3.5.1 Recuperação dos dados

Todo o mecanismo de *backup* do CEFET-MG é baseado no *software* Bacula[7], uma ferramenta *open source* capaz de gerenciar não só as cópias de segurança de um ambiente, como também a restauração e verificação da integridade de dados. O seu funcionamento baseado em rede permite que diferentes tipos de equipamentos em plataformas distintas sejam gerenciados.

[7]Bacula: `http://www.bacula.org/`

Por questões de segurança e sigilo não detalharemos maiores configurações sobre o Bacula dentro do contexto do CEFET-MG. A partir da aplicação própria do Bacula (*bconsole*) disponível no próprio sistema operacional GNU/Linux, a restauração dos dados será invocada a partir do comando *restore*. Dentro de um conjunto de formas para restauração de dados optamos pela opção número 6, responsável pela realização do *backup* de um cliente antes de um tempo especificado. Em seguida informamos o tempo referencial para a restauração no formato "YYYY-MM-DD HH:MM:SS".

O próximo passo consiste, a partir do comando *mark*, na marcação de todos os arquivos que se deseja realizar a restauração. É de interesse para o presente trabalho a restauração dos arquivos contidos dentro do diretório *var/www/cacti/rra/*, local onde o Cacti armazena os arquivos *.rra*. Vale ressaltar que essa marcação é realizada a partir da navegação na estrutura de diretórios e arquivos da máquina cliente. Essa estrutura, na arquitetura do Bacula, é também denominada "catálogo", ao qual é construído pelo próprio *software* durante a realização de cada cópia de segurança. Por fim basta confirmar a realização da restauração dos dados do cliente, a partir do dispositivo de armazenamento gerenciado pelo Bacula para o diretório */tmp/bacula-restores* residente na própria máquina cliente.

A Figura 3.5 ilustra de forma sintetizada o procedimento de restauração dos arquivos *.rra* para composição da amostra e realização dos experimentos.

Neste trabalho, a restauração dos dados foi realizada de forma iterativa para cada dia dentro do período de 06/02/2011 e 19/02/2011, definido em 3.3.

```
*restore

To select the JobIds, you have the following choices:
        ...
     6: Select backup for a client before a specified time
        ...
Select item:  (1-12): 6
The restored files will the most current backup
BEFORE the date you specify below.

Enter date as YYYY-MM-DD HH:MM:SS :2011-02-20 23:59:59
Defined Clients:
     1: monitor-fd
Select the Client (1-11): 1
...
cwd is: /
$ cd /var/www/cacti/rra
cwd is: /var/www/cacti/rra/
$ mark *
509 files marked.
$ done
Bootstrap records written to /var/lib/bacula/servidor-dir.restore.10.bsr
...
509 files selected to be restored.

Run Restore job
JobName:         monitor-RestoreFilesLinux
Bootstrap:       /var/lib/bacula/servidor-ccc-dir.restore.10.bsr
Where:           /tmp/bacula-restores
Replace:         always
FileSet:         monitor
Backup Client:   monitor-fd
Restore Client:  monitor-fd
Storage:         Storage
When:            2011-02-23 17:36:30
Catalog:         xxxxxxCCC
Priority:        10
OK to run? (yes/mod/no): yes
Job queued. JobId=2841
```

Figura 3.5: Principais mensagens do procedimento de restauração de arquivos do Bacula

3.5.2 Extração dos dados

Essa subsessão aborda o primeiro passo da aquisição dos dados para composição da amostra. Consiste no *script arl-extract.sh*, escrito em *Shell-Script*, para leitura e tratamento de um arquivo *.rra* em um conjunto de diretórios a partir do comando *rrdtool*. Nesse conjunto de diretórios, listados na Figura 3.6, cada pasta contém os arquivos *.rra* correspondentes a cada dia dentro do período de análise.

```
ulisses@cotta:~/Documentos/ARL/Monografia/Dados$ ls -d rra*
rra-06fev  rra-08fev  rra-10fev  rra-12fev  rra-14fev  rra-16fev  rra-18fev
rra-07fev  rra-09fev  rra-11fev  rra-13fev  rra-15fev  rra-17fev  rra-19fev
```

Figura 3.6: Estruturação do conjunto de diretórios dos arquivos *.rra*

Inicialmente são definidas as variáveis *diretorio* e *file*, responsáveis respectivamente pelo diretório de armazenamento dos resultados e nome do arquivo que armazenará os valores de cada variável de interesse para os experimentos. A variável *data_inicio* indica a data inicial da extração dos dados, ao qual é transformada em seguida para o formato *timestamp*.

Para o funcionamento do *script* é imprenscindível que esses diretórios listados na Figura 3.6 possuam alguma forma de ordenação cronológica. Neste caso o nome de cada pasta garante essa ordenação, porém outros artíficios como data de criação ou data de último acesso também podem ser adotados. Qualquer forma de ordenação dos diretórios adotada deve ser garantida na estrutura de repetição do *script arl-extract.sh* na linha 10, exibido na Figura 3.7. Isso se deve porque as variáveis de tempo do *script*, *dia_inf* e *dia_sup* responsáveis respectivamente pelos limites inferiores e superiores do intervalo de tempo da consulta pelo *rrdtool*, atuam de forma sincronizada e independente a cada iteração de diretório *rra** corrente. Conforme o *script* na Figura 3.7, as variáveis *dia_inf* e *dia_sup* são incrementadas em função do contador *dd*.

Na prática, para o diretório *rra-06fev* (variável *pasta*) será considerado o intervalo de tempo *2011-02-06 00:00:00* (variável *data_inf*) à *2011-02-06 23:59:59* (variável *data_sup*). Da mesma forma, para o diretório *rra-07fev* terá período compreendido entre *2011-02-07 00:00:00* a *2011-02-07 23:59:59*, e assim sucessivamente.

A exportação dos arquivos pelo comando *rrdtool*, localizado na linha 23 do *script arl-extract.sh* na Figura 3.7, é realizada pela opção *xport*. Adicionalmente,

os seguintes parâmetros compõem a exportação dos arquivos *.rra* de maneira completa e que atenda às necessidades dos experimentos:

- ***–start***: *Timestamp* inicial do intervalo de tempo para exportação dos valores;
- ***–end***: *Timestamp* final do intervalo de tempo para exportação dos valores;
- ***–enumds***: Gera marcadores enumerados para cada valor exportado no arquivo *.xml*;
- ***–step***: Intervalo de tempo igualmente espaçado entre cada coleta de dados. Por padrão, o valor de *step* está sincronizado com as configurações do Cacti, com valor igual a 300 segundos;
- ***–m***: Número máximo de linhas do arquivos *.xml*;
- **DEF**: Associa, a uma variável, os valores contidos no arquivo *.rrd* especificamente. É possível que dois ou mais arquivos *.rrd* sejam tratados simultaneamente;
- **CDEF**: Aplicação de cálculo aritmético, se necessário, da variável definida em DEF;
- **XPORT**: Exportação das variável definidas em DEF ou CDEF para o arquivo *.xml* propriamente dito.

Por fim, o *script* faz um pequeno tratamento nos arquivos *.xml* antes de sua finalização. Esse tratamento realiza, com a ajuda da linguagem *Awk*[8], a extração de cadeias de caracteres que não serão utilizadas no procedimento de exportação dos dados. O objetivo é facilitar procedimento escrito em Perl, que será discutido na sessão 3.5.3, para composição e disponibilização da amostra em um banco MySQL.

Vale ressaltar que a extração desses dados foi realizada para cada dia dentro do período de 06/02/2011 e 19/02/2011, definido em 3.3, considerando todas as variáveis de rede declaradas na sessão 4.4.

[8] Awk: `http://www.gnu.org/software/gawk/`

```
# ----  Parâmetros para coleta de dados
diretorio="resultado"; file="05-servidor_trg"; file_out=$diretorio/$file;

dia_inicio="2011-02-06"; dia_inicio=`date -d "$dia_inicio" "+\%s"`
step=300; rows=288; dd=0

# ---- Estrutura de repetição para os diretórios rra* contendo os arquivos de interesse
for pasta in `ls -d rra*`; do
    # -- Estrutura de repetição para os diretórios rra* contendo os arquivos de interesse
    dia_inf=`echo "$dia_inicio+\$dd*86400" | bc `
    dia_sup=`echo "$dia_inf+86400" | bc `
    dia=`date --date="1970-01-01 UTC $dia_inf seconds" +\%d-\%m-\%Y`

    # -- Impressão de mensagem informativa sobre processamento de diretório
    echo "Processando diretório \"$pasta\":  período de " `date --date="1970-01-01 UTC \
       $dia_inf seconds" +%d/%m/%Y-%H:%M:%S`  " à " `date --date="1970-01-01 UTC \
       $(expr $dia_sup - 1) seconds" +%d/%m/%Y-%H:%M:%S`

    # -- Execução do comando rrdtool para exportação de arquivos .rra para
    #    arquivos temporários .txt
    rrdtool xport --start $(expr $dia_inf - 1) --end $(expr $dia_sup - 1) --enumds \
       --step $step -m $rows \
            DEF:i=$pasta/bridge_traffic_in_383.rrd:traffic_in:AVERAGE \
            DEF:o=$pasta/bridge_traffic_in_383.rrd:traffic_out:AVERAGE \
            CDEF:ii=i,8,* \
            CDEF:oo=o,8,* \
            XPORT:ii:trafego-in \
            XPORT:oo:trafego-out > $diretorio/$file"--"$dia".txt"

    # -- Criação de arquivo .xml
    arquivo=$diretorio/$file"--"$dia".xml"; cat /dev/null > $arquivo;

    # -- Tratemento de arquivos .xml, para retirada de conteúdo desnecessário
    echo "<?xml version=\"1.0\" encoding=\"ISO-8859-1\"?>" >> $arquivo
    echo "" >> $arquivo
    echo "  <data>" >> $arquivo
    awk '/<row>/ {print $0}' $diretorio/$file"--"$dia".txt" >> $arquivo
    echo "  </data>" >> $arquivo

    dd=$(expr $dd + 1);   echo "Feito.";
done
echo "Finalizado!"
```

Figura 3.7: *Script arl-extract.sh* para extração de dados do formato *.rra* para *.xml*

3.5.3 Exportação dos dados

Nessa etapa é realizada, a partir do *script arl-export.pl*, a composição da amostra dos dados em um banco de dados MySQL. Os dados resultantes do tratamento feito pelo *rrdtool* anteriomente serão os parâmetros para o funcionamento desse *script*. Dessa forma, os arquivos analisados nesse momento estão armazenados no diretório informado pela variável *diretorio* no *script arl-extract.sh*.

Para a manipulação de arquivos *.xml* a partir de *scripts* Perl foi necessário a instalação do módulo *XML::Simple* (TECH REPUBLIC, 2004). Como a distribuição utilizada para a realização de todos os experimentos basea-se no Debian, a dependência desse módulo foi corrigida pela instalação do pacote *libxml-simpleobject-perl* via *apt-get*.

A Figura 3.8 ilustra o código do *script arl-export.pl*. As linhas 4, 7 e 8 representam o uso dos módulos *XML::Simple* e *Data::Dumper* como também a declaração de variáveis em Perl para manipulação de arquivos *.xml*. Em seguida é declarada a função *banco* responsável pela inclusão dos dados no banco de dados MySQL. Dada as variáveis contendo informações para conexão com a base de dados, tais como *hostname*, usuários, senha e *database*, a partir do próprio *script arl-export.pl* é invocado o comando *mysql* para execução da instrução SQL contida na variável *$mysql*.

O modelo Entidade-Relacionamento do banco de dados está representado na Figura 3.9. Contém basicamente duas entidades *variavel* e *dados*, responsáveis respectivamente pelo armazenamento do nome das variáveis de rede de interesse para o experimento e dos valores propriamente ditos. A disponibilização de todo o conjunto de dados em um banco de dados permite que as amostras sejam facilmente construídas a partir de instruções SQL.

Em continuidade do *script arl-export.pl*, uma estrutura de repetição faz-se necessária para o tratamento de cada arquivo *.xml* através da variável *$arq*. A variável *$data* assumirá todos os valores da *tag* <*row*> exemplificados na Figura 3.10. Consequentemente, mais uma estrutura de repetição é necessária para a leitura de cada valor da variável *$data*, com o objetivo de se extrair os valores das *tags* <*t*> e <*v0*>. Conforme a exportação dos dados a partir do *rrdtool* pelo *script arl-extract.sh*, novas *tags* $v0, v1, \ldots, vn$ são necessárias para o tratamento das variáveis de interesse para os experimentos.

Por fim, a função *banco* é chamada para a inclusão dos valores de rede no banco de dados MySQL. Cabe ressaltar que a execução do *script* foi repetido para

```
#!/usr/bin/perl

# use module
use XML::Simple;  use Data::Dumper;  use Switch;

# create object
$xml = new XML::Simple (KeyAttr=>[]);
my $xs = XML::Simple->new(ForceArray => 1, KeepRoot => 1);

sub banco {
        my $srv_host="localhost";   my $srv_db="dadosARL";
        my $srv_user="root";    my $srv_senha="root";
        my $mysql = "insert into dados \
              (timestamp,valor,variavel_idvariavel) \
              values (\\\"$_[0]\\\",$_[1],$_[2])";

        `mysql -h $srv_host -u $srv_user -p$srv_senha $srv_db \
            -e "$mysql" 2> /dev/null`;
}

# Exportação dos dados
my @arquivos = `ls resultado/*xml`;
for my $arq (@arquivos) {
       print "Processando arquivo: ", substr($arq,0,-1) ;
       my $data = $xml->XMLin( substr($arq,0,-1) );

       foreach $e (@{$data->{row}})
       {
               $dia=$e->{t};
               $dia=`date --date="1970-01-01 UTC $dia \
                  seconds" +%Y-%m-%d" "%H:%M:%S`;
               $dia=substr($dia,0,-1);

               $aux=$e->{v0}; $aux =~ s/,/./;
               &banco($dia,$aux,5);
               $aux=$e->{v1}; $aux =~ s/,/./;
               &banco($dia,$aux,6);
       }
}
```

Figura 3.8: *Script arl-export.pl* para exportação dos dados do formato *.xml* para base MySQL

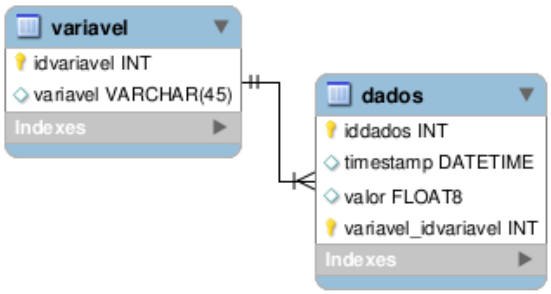

Figura 3.9: Arquitetura do funcionamento da ferramenta Cacti

```
  1 <?xml version="1.0" encoding="ISO-8859-1"?>
  2
  3   <data>
  4     <row><t>1296957600</t><v0>2.1813589333e+07</v0></row>
  5     <row><t>1296957900</t><v0>1.7530224640e+07</v0></row>
  6     <row><t>1296958200</t><v0>1.6269243733e+07</v0></row>
  7     <row><t>1296958500</t><v0>1.6180128427e+07</v0></row>
  8     <row><t>1296958800</t><v0>2.5422848000e+07</v0></row>
  9     <row><t>1296959100</t><v0>2.0750267733e+07</v0></row>
 10     <row><t>1296959400</t><v0>2.0793698987e+07</v0></row>
...
292   </data>
```

Figura 3.10: Exemplo de disponibilização de dados de um arquivo *.xml*

cada conjunto de arquivos *.xml* representando as variáveis de rede do presente trabalho.

3.5.4 *Bootstrapping*

Conforme foi apresentado na subsessão 3.2.1, o *bootstrapping* é uma técnica de re-amostragem com o propósito de ajustar as variáveis da amostra original conforme distribuição gaussiana. Adicionalmente, na prática, o *bootstrapping* implica na repetição do experimento para obtenção dos dados da amostra para se comportar conforme distribuição normal.

A Figura 3.11 apresenta a função desenvolvida para o *software* Scilab com o intuito de aplicar a técnica de *bootstrapping*. A função ilustrada recebe o vetor X

correspondente à amostra original dos dados, e b como representação do tamanho da amostra tratada. Em seguida é aplicada uma estrutura de repetição para que os valores de X sejam reamostrados, com repetição, b vezes, e o estimador média calculado como elemento da nova amostra Y. A função retorna o vetor Y como conjunto de dados reamostrados pela técnica.

```
function Y = boostrapping(X,b)
  S=size(X);  n=S(1,2);
  for i=1:b
    B=sample(n,X);
    Y(1,i)=mean(B);
    clear B;
  end
endfunction
```

Figura 3.11: Função em Scilab para reamostragem de uma amostra por *bootstrapping*

Para ilustrar a eficácia do *bootstrapping* foram gerados dois histogramas com número de classes igual a 25. As Figuras 3.12 e 3.13 trazem, respectivamente, dados da amostra original obtida pelo Cacti e conjunto reamostrado pela técnica de *bootstrapping*. No eixo das abcissas estão representados os percentuais de utilização da carga de processamento do roteador, ao passo que o eixo das ordenadas traz as frequências dos valores dentro de cada uma das classes. Nas representações ilustradas pelas Figuras 3.12 e 3.13 foram utilizados 120 elementos para a amostra original e o valor de $b = 2000$ para a reamostragem.

Na Figura 3.12 pode-se notar que a frequência tem valor máximo próximo de $0,3$, ao passo que os percentuais da carga de processamento estão distribuídos numa faixa entre 0% à aproximadamente 40%. Já na Figura 3.13 o valor máximo de frequência está próximo de 2, e os percentuais da carga de processamento esao concentrados em uma faixa estreita de $3,8\%$ à $5,6\%$. Isso se deve pelo fato do estimador utilizado no *bootstrapping* ser a média arimética, implicando na concentração dos dados da nova amostra em torno desse valor.

3.5.5 Análise estatística

Nessa última etapa serão aplicadas, de fato, as ferramentas estatísticas discutidas na sessão 2.4. Os *scripts* e funções apresentados a seguir foram escritos dentro do

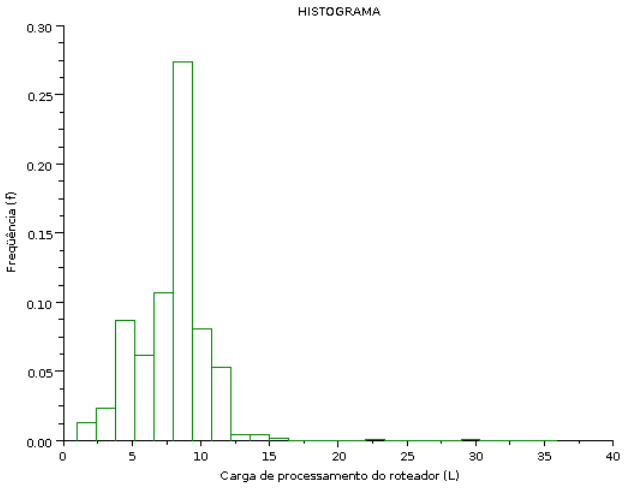

Figura 3.12: Amostra de dados original sem reamostragem

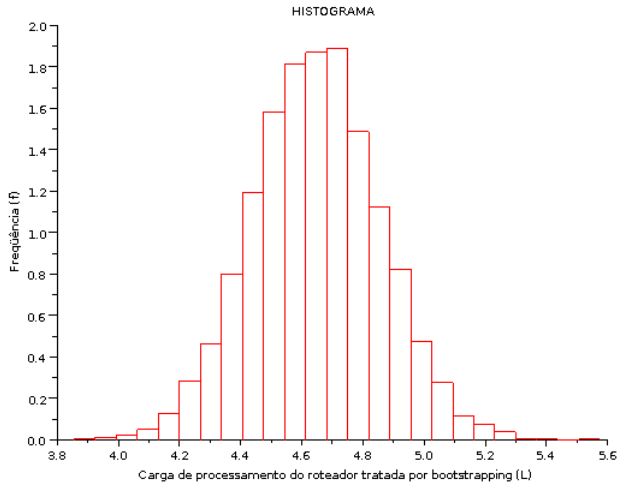

Figura 3.13: Amostra de dados original com reamostragem por *bootstrapping*

ambiente do Scilab, conforme previsto na etapa de planejamento dos experimentos.

A regressão polinomial foi o único experimento realizado de forma adicional aos demais previamente planejados e apresentados na sessão 3.4. Essa decisão foi tomada posteriormente com o intuito de aplicar mais uma ferramenta estatística em problemas de regressão, considerando os resultados, discutidos posteriormente

em 4.1, não satisfatórios dentro desse grupo de experimentos. Dessa forma, os experimentos envolvendo regressão polinomial foram tratados juntamente com os problemas de estimativa de parâmetro e intervalo de confiança, justamente por manipular os valores máximos médios das variáveis em estudo. A Figura 3.14 ilustra a função *reg_polinomial* responsável de fazer a regressão polinominal de um conjunto de valores ordenados, segundo Neto (2002).

```
function REG = reg_polinominal(X,Y)
  tam=size(X);  n=tam(1,2);  mediaX=mean(X);

  Sxy=0; Sxx=0; S1=0; S2=0;
  for i=1:n
    aux=(X(i)-mediaX)*Y(i);          Sxy=Sxy+aux;
    aux=(X(i)-mediaX)^2;             Sxx=Sxx+aux;
    aux=( (X(i)-mediaX)^2 ) * Y(i);  S1=S1+aux;
    aux=(X(i)-mediaX)^4;             S2=S2+aux;
  end

  A(1,1)=n; A(1,2)=Sxx; A(2,1)=Sxx;  A(2,2)=S2;
  B(1)=sum(Y);  B(2)=S1;
  R=inv(A)*B;

  a=R(1,1);  b=Sxy/Sxx;  c=R(2,1);

  REG(1)=a;   REG(2)=b;  REG(3)=c;  REG(4)=mediaX;
endfunction
```

Figura 3.14: Função "polinomial.sce"para regressão polinomial

A Figura 3.15 apresenta o *script* para a construção de intervalos de confiança das variáveis de performance de rede de interesse do presente trabalho. Inicialmente foram definidas as variáveis $b0$, n e t, correspondentes respectivamente ao tamanho da reamostragem por *bootstrapping*, tamanho da amostra original e valor de t na distribuição t de *Student*. Com $n = 120$ e um nível de significância de 5%, conforme a Tabela A.2 referente a distribuição t de *Student* temos que $t_{120,5\%} = 1,98$.

O comando *exec* dentro do *script* do Scilab realiza a inclusão das funções *function_bootstrapping* e *polinomial* apresentadas respectivamente nas Figuras 3.11 e 3.14. A função *read* do Scilab realiza a leitura dos arquivos *.txt* que compõem as amostras para cada variável, originários de consultas SQL a partir do modelo ER da Figura 3.9.

Em seguida, para cada um dos 14 intervalos de hora (07 : 00 às 08 : 00, 08 : 00 às 09 : 00, ..., 20 : 00 às 21 : 00), é realizado os seguintes passos:

- Reamostragem pelo *bootstrapping*: Além da reamostragem por *bootstrapping*, são tomados os valores ordenados e igualmente espaçados em n quantis ($n = 120$) para construção do intervalo de confiança (CESARIO; BARRETO, 2003);
- Construção do intervalo: Os valores de média, desvio padrão, semi-amplitude do intervalo (conforme equação 2.10) e limites do intervalos são de fato calculados a partir do conjunto Y reamostrado;
- Tamanho da amostra: Determinação do tamanho da amostra necessária para a construção do intervalo de confiança, com base em distribuições t de *Student*, conforme equação 2.13 e procedimento descrito em 2.4.2.3;

Como as amostras utilizadas para realização de todos os experimentos envolvendo intervalo de confiança foram de 120 elementos, dificilmente esse valor será insuficiente para estimativa de parâmetros. Grande parte da literatura na área de estatística sugere que a construção de intervalos de confiança para $n \geqslant 30$.

Por fim é realizada a regressão polinomial para os valores máximos médios das variáveis de interesse, conforme apresentado na sessão 3.4. Os valores que representam o domínio da função regredida variam de 7,5 à 20,5 a cada unidade, ou seja, $7,5, 8,5, \ldots, 20,5$. O acréscimo de 0,5 a cada limite inferior dos intervalos de tempo se deve à adequação à reta real do intervalo entre 07:00 à 21:00, período correpondente à regressão.

Para os problemas de análise de variância, o *script* ilustrado na Figura 3.16 tenta realizar e apresentar os dados de forma mais fiel possível os procedimentos de ANOVA apresentados na subsessão 2.4.5. Conforme o planejamento dos experimentos, a análise de variância realizada nesse trabalho compreende em duas classificações dos dados com repetição. Portanto, os resultados da ANOVA serão apresentados conforme a Tabela 2.10.

Inicialmente o *script* para construção da análise de variância define o número de colunas ($n = 2$) e de linhas ($k = 3$) da classificação, e o número de repetição ($r = 10$) dos dados. Em seguida, através do comando *read*, os dados são obtidos a partir de arquivos *.txt* construídos pelas consultas SQL no banco de dados MySQL.

O próximo passo consiste no cálculo da somatória das linhas e colunas, representado pelas variáveis Ti e $T\,j$ respectivamente, além do somatório do quadrado

dos dados das linhas e colunas, atribuido às variáveis Qi e Qj. Adicionalmente é realizada a soma dos quadrados do somatório dos dados conforme classificação de linha e coluna, associado à variável Tij.

O *script* termina com o cálculo das somas dos quadrados entre linhas (SQL), colunas (SQC), interação (SQI), entre tratamentos ($SQTr$), residual (SQR) e total (SQT), baseado na Tabela 2.10. Antes de se obter o cálculo do quadrado médio para cada uma das fontes de variação, S_L^2, S_C^2, S_I^2, S_{Tr}^2 e S_T^2, são obtidos os graus de liberdade $k-1$, $n-1$, $(k-1)(n-1)$, $nk-1$, $nk(r-1)$ e $nkr-1$. A finalização ocorre com a obtenção do valor de F calculado, para comparação de F crítico conforme distribuição F de *Snedecor*.

Para o grupo de experimentos envolvendo correlação linear foram desenvolvidas duas funções, uma para o cálculo da correlação linear em si e outra para realização do teste do coeficiente de correlação. A função *correlacao*, ilustrada na Figura 3.17, calcula o coeficiente de *correlação linear de Pearson*, obtido pela relação entre a covariância entre X e Y com o produto dos desvios padrão de X e Y segundo a equação 2.41. Já a função *testa_correlacao*, apresentada na Figura 3.18, realiza o teste do coeficiente de correlação para verifica a existência ou não de correlação linear, conforme Neto (2002).

O *script* para realização do cálculo do conjunto de correlações linear está ilustrado na Figura 3.19. Inicialmente as funções *correlacao* e *testa_correlacao* são carregados no ambiente do Scilab através do comando *exec*. Em seguida a função *read* do Scilab se encarrega de importar os valores dos arquivos *.txt* que contém os valores das amostras. Vale lembrar que nesse grupo de experimentos são correlacionadas variáveis distintas, sendo necessário informar os arquivos dos valores de amostras correspondente a cada variável de interesse.

O próximo passo consiste na realização da correlação linear propriamente dita através das funções *correlacao* e *testa_correlacao*. Em um primeiro instante é considerado cada intervalo de hora como período para a correlação e seu respectivo teste, garantido no *script* pela estrutura de repetição de 1 a 14. Logo a seguir são chamadas as mesmas funções, agora com os pares ordenados representando todo o período de tempo, das 07:00 às 21:00.

Quanto aos testes de correlação vale ressaltar que foram realizados a um nível de significância de 5% e 10%. Dessa forma, os valores de t na distribuição t de *Student* para $n-2$ graus de liberdade, são respectivamente $t_{118,5\%} = 1,657$ e $t_{118,10\%} = 1,289$. Para os experimentos envolvendo todo o período de tempo das 07:00 às 21:00 foram considerados 1678 graus de liberdade. Conforme a Tabela A.2 da distribuição t de *Student*, para graus de liberdade muito grandes o valor de t

tende a se estabilizar. Portanto, os mesmo valores de t foram utilizados para todos os testes de correlação linear, independente do número de graus de liberdade.

```
//Apaga variáveis, de forma a garantir a inicialização delas
clear; b0=1200; n=b0/10; t=1.98;

//Carrega a funcao 'bootstrapping' para o ambiente do SciLab
exec("function_bootstrapping.sce")
exec("polinomial.sce");

// Leitura dos dados .txt para matrix LR
// Arquivos estão compreendidos em uma faixa de LR-07.txt até LR-20.txt
A=read("LR-07.txt",1,n); A=samwr(n,1,A); X=[A];
...
A=read("LR-20.txt",1,n); A=samwr(n,1,A); X=[X A];
clear A;

for i=1:14
  bp=b0; //amostra b piloto
  c=1;
  while c>0
    //Reamostragem através do bootstrapping
    A=X(:,i);
    //Ordenação do vetor contendo os dados reamostrados
    Y=gsort(boostrapping(A',bp));  clear Z;
    // Captura dos valores reamostrados igualmente espaçados
    for j=0:(n-1)
      Z(1,j+1)=Y(1,j*10+1);
    end

    //Construção do intervalo de confiança
    m(i)=mean(Z); //(média da amostra)
    s(i)=st_deviation(Z); //(desvio padrão da amostra)
    e(i,4)=t*(s(i)/sqrt(n)); //(semi-amplitude do intervalo)
    e(i,1)=m(i)-e(i,4); //(construção do intervalo, limite inferior)
    e(i,2)=m(i);
    e(i,3)=m(i)+e(i,4); //(construção do intervalo, limite superior)

    //Cálculos para tamanho da amostra
    bc=round((t*s(i)/e(i,4))^2);
    B(i,1)=bc; B(i,2)=c; c=c+1;

    //Condicional para verificar sea amostra é suficiente
       // Se o tamanho da amostra calculado (bc) for menor igual que o tamanho da
       // amostra piloto (bp), (bp) é suficiente
    if bc<=bp
      c=0;
       // Se o tamanho da amostra calculado (bc) for maior que o tamanho da amostra
       // piloto, (bp) é incrementado em 25 unidades.
    else
      bp=bc+25;
    end
  end
end

//Regressão polinomial
H=7.5:20.5; E=e(:,2); E=E';
R=reg_polinominal(H,E);
```

Figura 3.15: *Script* em Scilab para construção dos intervalos de confiança

```
clear; n=2; k=3; r=10; sumTij=0;
A=read("AN-P1i-T.txt",1,r); B=[A]; A=read("AN-P2i-T.txt",1,r); B=[B, A];
A=read("AN-P3i-T.txt",1,r); B=[B, A]; D=[B'];
A=read("AN-P1o-T.txt",1,r); B=[A]; A=read("AN-P2o-T.txt",1,r); B=[B, A];
A=read("AN-P3o-T.txt",1,r); B=[B, A]; D=[D, B'];

//Eleva dados ao quadrado
for i=1:k*r
  for j=1:n
    D2(i,j)=D(i,j)^2;
  end
end

//Cálculo de Tj e Qj
Tj=sum(D,'r'); Qj=sum(D2,'r');
for i=1:n
   Tj(2,i)=Tj(1,i)^2;
end

//Cálculo de Ti e Qi
for i=1:k
  Ti(1,i)=sum(D(i*r-r+1:i*r,1:n));   Qi(1,i)=sum(D2(i*r-r+1:i*r,1:n));   Ti(2,i)=Ti(1,i)^2;
end

//Cálculo de Tij
for i=1:k
  for j=1:n
    Tij(i,j)=sum(D(i*r-r+1:i*r,j));  sumTij=sumTij+(Tij(i,j))^2;
  end
end

//Cálculo das variações (soma dos quadrados)
aux=(sum(Ti(1,:))^2)/(n*k*r);
SQL=sum(Ti(2,:))/(n*r)-aux; SQC=sum(Tj(2,:))/(k*r)-aux; SQT=sum(Qi)-aux; SQTr=(sumTij/r)-aux;

//Armazenamento somas dos quadrados
ANOVA(1,1)=SQL;    ANOVA(2,1)=SQC;       ANOVA(3,1)=SQTr-SQL-SQC;
ANOVA(4,1)=SQTr;   ANOVA(5,1)=SQT-SQTr;  ANOVA(6,1)=SQT;

//Cálculo e armazenamento dos graus de liberdade
ANOVA(1,2)=k-1;   ANOVA(2,2)=n-1;        ANOVA(3,2)=(k-1)*(n-1);
ANOVA(4,2)=n*k-1; ANOVA(5,2)=n*k*(r-1);  ANOVA(6,2)=n*k*r-1;

//Cálculo de armazenamento dos quadrados médios
for i=1:5
  ANOVA(i,3)=ANOVA(i,1)/ANOVA(i,2);
end
ANOVA(6,3)=0;

//Cálculo e armazenamento do valor de F
ANOVA(1,4)=ANOVA(1,3)/ANOVA(5,3); ANOVA(2,4)=ANOVA(2,3)/ANOVA(5,3); ANOVA(3,4)=ANOVA(3,3)/ANOVA(5,3);
ANOVA(4,4)=ANOVA(4,3)/ANOVA(5,3); ANOVA(5,4)=0;                     ANOVA(6,4)=0;

//Armazenamento de F_crítico
ANOVA(1,5)=3.17;  ANOVA(2,5)=4.02;  ANOVA(3,5)=3.17;  ANOVA(4,5)=2.39; ANOVA(5,5)=0; ANOVA(6,5)=0;
```

Figura 3.16: *Script* em Scilab para construção das análises de variância

```
function r = correlacao(X,Y)
  S=size(X);  n=S(1,1);  clear S;

  for i=1:n
    XY(i)=X(i)*Y(i);
  end

  Sxy=sum(XY)-(sum(X)*sum(Y)/n);
  Sxx=sum(X^2)-((sum(X))^2)/n;
  Syy=sum(Y^2)-((sum(Y))^2)/n;
  r=Sxy/(sqrt(Sxx*Syy));
endfunction
```

Figura 3.17: Função "correlacao"em Scilab para cálculo da correlação linear

```
function r = testa_correlacao(tt,n,r)
  tc=r*sqrt((n-2)/(1-r^2));

  if r>0 //Correlação positiva
    if tt<tc
      r=1; //Rejeitada  --> existe correlação linear
    else
      r=0; //Aceita --> não existe correlação linear
    end
  elseif r<0 //Correlação negativa
    if tt>tc
      r=1; //Rejeitada --> existe correlação linear
    else
      r=0; //Aceita --> não existe correlação linear
    end
  end
endfunction
```

Figura 3.18: Função "testa_correlacao"em Scilab para teste da correlação linear

```
  clear;

  exec("function_correlacao.sce"); exec("function_testa_correlacao.sce");

  n=120;
  // Leitura dos dados .txt para matrix PRD1
  // Arquivos estão compreendidos em uma faixa de PRD1-07.txt até PRD1-20.txt
  A=read("PRD1-07.txt",1,n); X=[A];
  ...
  A=read("PRD1-20.txt",1,n); X=[X; A];

  // Leitura dos dados .txt para matrix TRD1
  // Arquivos estão compreendidos em uma faixa de TRD1-07.txt até TRD1-20.txt
  A=read("TRD1-07.txt",1,n); Y=[A];
  ...
  A=read("TRD1-20.txt",1,n); Y=[Y; A];

  X=X';   Y=Y';

//CORRELAÇÃO PARA INTERVALOS DE TEMPO
for i=1:14
  R(i,1)=correlacao(X(:,i),Y(:,i));
  R(i,2)=testa_correlacao(1.289,n,R(i,1));
  R(i,3)=testa_correlacao(1.657,n,R(i,1));
end

//CORRELAÇÃO PARA TODAS A AMOSTRA
XX=[X(:,1)]; YY=[Y(:,1)];
for i=2:14
  XX=[XX; X(:,i)];  YY=[YY; Y(:,i)];
end
R(15,1)=correlacao(XX,YY);
R(15,2)=testa_correlacao(1.289,n,R(15,1));
R(15,3)=testa_correlacao(1.657,n,R(15,1));
```

Figura 3.19: *Script* em Scilab para correlação linear

3.6 Comentários finais

Com base nas etapas previstas no ciclo PDCA para realização de experimentos estatísticos, a etapa de apresentação e análise de resultados é considerada, muitas vezes, a mais importante de todo o trabalho de experimentação estatística. No entanto, a metodologia adotada e o desenvolvimento propriamente dito tem parcela significativa para o sucesso da pesquisa.

Nesse capítulo foi levantado um conjunto de idéias e conceitos sobre a performance de redes de computadores. A questão das redes heterogêneas, diversidade de serviços e funcionamento de alguns protocolos, permitiu melhor delineamento de todo a trabalho. Além disso, a seleção da variável resposta, escolha de fatores e planejamento de todos os experimentos possibilitaram o estabelecimento de limites na pesquisa, visto o potencial de exploração abrangente do tema.

Inicialmente havia a crença que a etapa de realização do experimento seria a etapa menos significativa de todo o processamento de experimentação. No entanto, para o presente trabalho que contempla a gerência de redes de computadores, a realização do experimento é a atividade mais importante para o sucesso de toda a pesquisa. Além do conhecimento estatístico necessário para a realização da análise dos dados propriamente dito, essa etapa exigiu considerável conhecimento técnico para a obtenção dos dados. Desses conhecimentos necessários podemos citar a automação de tarefa com as ferramentas *Shell-Script* e Perl, gerência de sistemas com o gerenciador de *backup* Bacula, e manipulação de dados a partir do SGBD MySQL.

Capítulo 4

Resultados e análises

Neste capítulo apresentaremos os resultados obtidos pela experimentação e tratamento estatístico das variáveis de rede conforme definido no capítulo anterior. Devido a extensão desse capítulo por conta do número de gráficos e tabelas, cada análise será feita a medida que os resultados forem apresentados. Adicionalmente, sempre que pertinente, gráficos descritivos obtidos da ferramenta Cacti que representam o ambiente de produção da infraestrutura de rede do CEFET-MG serão ilustrados para um comparação qualitativa com os resultados estatisticamente tratados. Na finalização desse capítulo será explanada uma breve análise geral de todos os experimentos, de maneira que sejam apresentadas possíveis relações entre experimentos e resultados distintos.

4.1 Grupo 1: Estimativa de parâmetros

Essa subsessão ilustra os resultados dos experimentos que contemplam a construção de intervalos de confiança. Os resultados desse grupo de experimentos serão exibidos através de um gráfico comparativo entre a média e a média máxima de cada variável, e duas tabelas contendo, respectivamente, os intervalos para a média e os intervalos para a média máxima.

Vale ressaltar que toda a estimativa de parâmetro foi calculada ao nível de significância de 5%. Dessa forma podemos afirmar, a partir de agora, que o estimador média de cada variável tem 95% de chance de estar contido no intervalo informado. Esses intervalos foram calculados com amostras de 120 elementos

($n = 120$), onde para todos os experimentos esse valor de n foi suficiente para a estimativa de parâmetros.

4.1.1 Carga de processamento do roteador

Nesse experimento foram analisadas a média da carga de processamento do roteador ($\overline{LR}$) e sua respectiva média dos valores máximos ($\overline{LR_{max}}$). Os intervalos de confiança construídos estão apresentados nas Tabelas 4.1 e 4.2, representando a estimativa do parâmetro média para as variáveis $\overline{LR}$ e $\overline{LR_{max}}$ respectivamente. Ao analisar a Figura 4.1 é importante verificar que os valores de $\overline{LR}$ e $\overline{LR_{max}}$ são menores nos extremos do período de tempo considerado. Na prática, nos intervalos de 07:00 às 09:00 e 19:00 às 21:00 temos percentuais menores de utilização da carga de processamento. Outro aspecto pertinente é sobre a diferença entre $\overline{LR_{max}}$ e $\overline{LR}$. Quanto maior a diferença, maior a variação dos valores da variável de interesse ao longo do período medido.

Analisando as Figuras 4.1 e 4.2 é plausível observamos a tendência dos valores da carga de processamento ao longo do período de tempo. Constata-se que em ambos os gráficos os valores são crescentes no início do período, apresentando uma ligeira estabilidade ora com variações suaves, ora com variações mais bruscas. No final do período, os valores de ambos os gráficos tendem a decrescer. Vale salientar que os valores de ambos os gráficos não são necessariamente iguais, visto que na Figura 4.1 ilustra-se o parâmetro média estimado e na Figura 4.2 dados reais da amostra.

Dado o valor da primeira linha da Tabela 4.1, período de 07:00 às 08:00, o intervalo $P(4,6099 \leqslant LR \leqslant 4,6876) = 95\%$ significa que a média da carga de processamento do roteador está compreendida entre $4,6099$ e $4,6876$ com a chance de 95% de acerto. Da mesma maneira, considerando a primeira linha da Tabela 4.2, afirmamos com 95% de certeza que a média do valores máximos da carga de processamento está contido entre o intervalo $7,0691$ e $8,1117$.

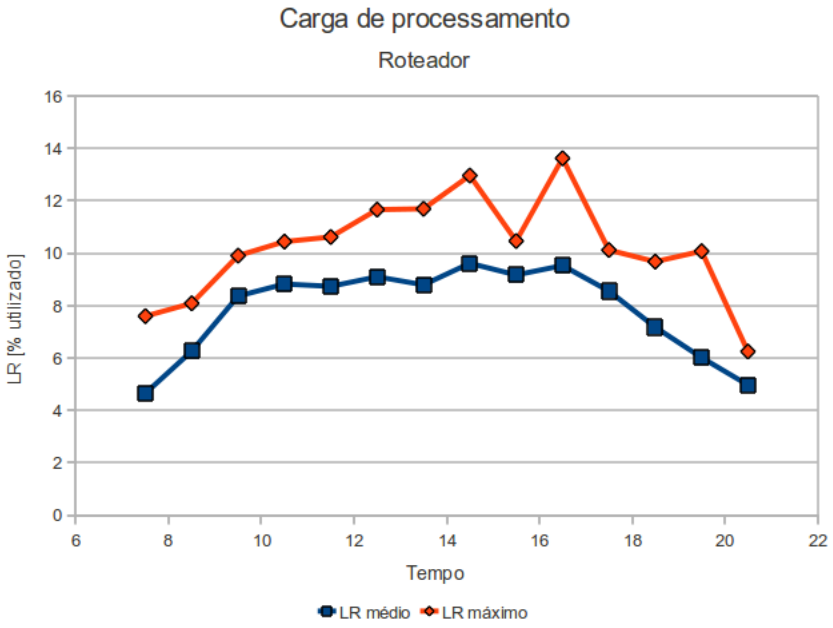

Figura 4.1: Carga de processamento do roteador

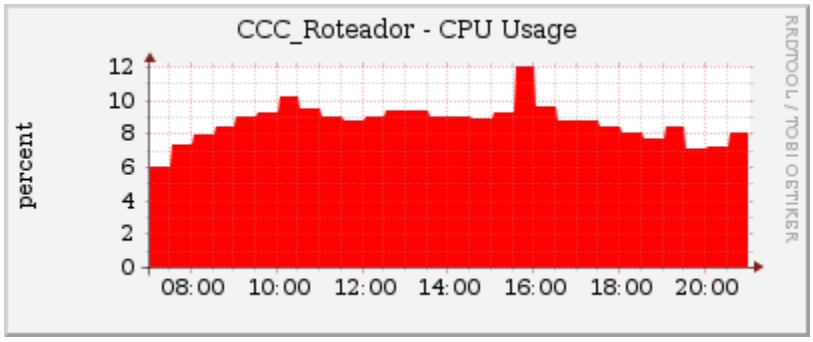

Figura 4.2: Carga de processamento do roteador, a partir da ferramenta Cacti

Tabela 4.1: Estimativa de parâmetro: média da carga de processamento do roteador

Período	Intervalo de confiança	$\overline{LR^*}$	$\overline{LR}$
07:00 - 08:00	$P(4,6099 \leqslant LR \leqslant 4,6876) = 95\%$	4,6487	4,6476
08:00 - 09:00	$P(6,2297 \leqslant LR \leqslant 6,2874) = 95\%$	6,2585	6,2560
09:00 - 10:00	$P(8,3330 \leqslant LR \leqslant 8,3805) = 95\%$	8,3567	8,3476
10:00 - 11:00	$P(8,7888 \leqslant LR \leqslant 8,8432) = 95\%$	8,8160	8,8078
11:00 - 12:00	$P(8,7035 \leqslant LR \leqslant 8,7525) = 95\%$	8,7280	8,7168
12:00 - 13:00	$P(9,0646 \leqslant LR \leqslant 9,1199) = 95\%$	9,0923	9,0932
13:00 - 14:00	$P(8,7459 \leqslant LR \leqslant 8,8224) = 95\%$	8,7841	8,7800
14:00 - 15:00	$P(9,5398 \leqslant LR \leqslant 9,6749) = 95\%$	9,6074	9,6027
15:00 - 16:00	$P(9,1442 \leqslant LR \leqslant 9,1827) = 95\%$	9,1635	9,1577
16:00 - 17:00	$P(9,4547 \leqslant LR \leqslant 9,5863) = 95\%$	9,5205	9,5099
17:00 - 18:00	$P(8,5235 \leqslant LR \leqslant 8,5748) = 95\%$	8,5491	8,5420
18:00 - 19:00	$P(7,1317 \leqslant LR \leqslant 7,2105) = 95\%$	7,1711	7,1695
19:00 - 20:00	$P(5,9456 \leqslant LR \leqslant 6,0746) = 95\%$	6,0101	6,0178
20:00 - 21:00	$P(4,9421 \leqslant LR \leqslant 4,9906) = 95\%$	4,9663	4,9586

Tabela 4.2: Estimativa de parâmetro: média dos valores máximos da carga de processamento do roteador

Período	Intervalo de confiança	$\overline{LR^*_{max}}$	$\overline{LR_{max}}$
07:00 - 08:00	$P(7,0691 \leqslant LR_{max} \leqslant 8,1117) = 95\%$	7,5904	7,3753
08:00 – 09:00	$P(7,6702 \leqslant LR_{max} \leqslant 8,4895) = 95\%$	8,0798	8,0000
09:00 – 10:00	$P(9,4791 \leqslant LR_{max} \leqslant 10,3197) = 95\%$	9,8994	9,7890
10:00 – 11:00	$P(10,0055 \leqslant LR_{max} \leqslant 10,8635) = 95\%$	10,4345	10,3850
11:00 – 12:00	$P(10,2512 \leqslant LR_{max} \leqslant 10,9702) = 95\%$	10,6107	10,4903
12:00 – 13:00	$P(11,2993 \leqslant LR_{max} \leqslant 12,0118) = 95\%$	11,6556	11,5873
13:00 – 14:00	$P(10,2027 \leqslant LR_{max} \leqslant 13,1658) = 95\%$	11,6843	11,4653
14:00 – 15:00	$P(11,4522 \leqslant LR_{max} \leqslant 14,4575) = 95\%$	12,9549	12,5847
15:00 – 16:00	$P(10,1695 \leqslant LR_{max} \leqslant 10,7414) = 95\%$	10,4555	10,3960
16:00 – 17:00	$P(10,9449 \leqslant LR_{max} \leqslant 16,2802) = 95\%$	13,6125	12,8857
17:00 – 18:00	$P(9,7385 \leqslant LR_{max} \leqslant 10,4859) = 95\%$	10,1122	10,0837
18:00 – 19:00	$P(9,0212 \leqslant LR_{max} \leqslant 10,3204) = 95\%$	9,6708	9,4913
19:00 – 20:00	$P(7,8267 \leqslant LR_{max} \leqslant 12,3106) = 95\%$	10,0686	9,4907
20:00 – 21:00	$P(5,9369 \leqslant LR_{max} \leqslant 6,5509) = 95\%$	6,2439	6,1943

4.1.2 Carga de processamento do servidor

A variável analisada nesse experimento foi a carga de processamento do servidor, com média representada por $\overline{LS}$ e a média dos valores máximos denotada por $\overline{LS_{max}}$. A Figura 4.3 ilustra o gráficos de dispersão dessas duas variáveis ao longo do tempo, da mesma forma que as Tabelas 4.3 e 4.4 apresentam os intervalos de confiança construídos.

Um aspecto pertinente é quanto a variação dos valores de $\overline{LS}$ e $\overline{LS_{max}}$ ao longo do período medido. No início os valores apresentam um considerável crescimento das 07:00 às 10:00, onde no período intermediário de medição apresentam alguns picos. A partir das 16:00 os valores do percentual de uso do servidor tem um queda muita acentuada, ao qual se estabiliza a partir das 20:00.

A partir das Tabelas 4.3 e 4.4 verificamos a amplitude dos intervalos construídos. Mesmo que $\overline{LS}$ e $\overline{LS_{max}}$ tenham sido tratada os por *bootstrapping*, a amplitude do intervalo de $\overline{LS_{max}}$ é maior que $\overline{LS}$. Isso ocorre porque a média dos valores máximos tem maior variação em comparação com a média da carga de processamento do servidor.

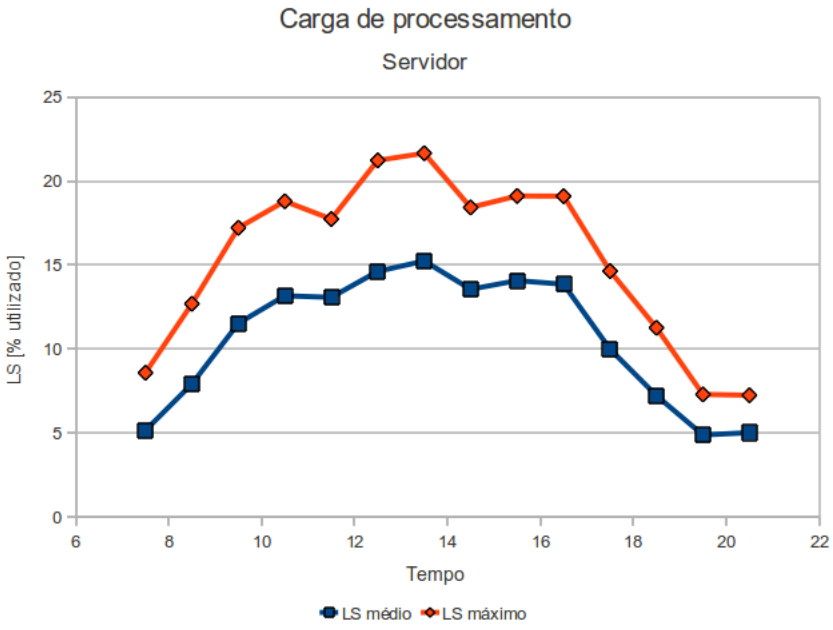

Figura 4.3: Estimativa de parâmetros: carga de processamento do servidor

Tabela 4.3: Estimativa de parâmetro: média da carga de processamento do servidor

Período	Intervalo de confiança	$\overline{LS^*}$	$\overline{LS}$
07:00 - 08:00	$P(5,0964 \leqslant LS \leqslant 5,1906) = 95\%$	5,1435	5,1161
08:00 – 09:00	$P(7,8471 \leqslant LS \leqslant 7,9725) = 95\%$	7,9098	7,8971
09:00 – 10:00	$P(11,4600 \leqslant LS \leqslant 11,5859) = 95\%$	11,5230	11,5132
10:00 – 11:00	$P(13,0988 \leqslant LS \leqslant 13,2242) = 95\%$	13,1615	13,1585
11:00 – 12:00	$P(13,0024 \leqslant LS \leqslant 13,1352) = 95\%$	13,0688	13,0712
12:00 – 13:00	$P(14,5329 \leqslant LS \leqslant 14,6946) = 95\%$	14,6137	14,6007
13:00 – 14:00	$P(15,1427 \leqslant LS \leqslant 15,3238) = 95\%$	15,2333	15,2159
14:00 – 15:00	$P(13,4918 \leqslant LS \leqslant 13,6115) = 95\%$	13,5517	13,5482
15:00 – 16:00	$P(14,0156 \leqslant LS \leqslant 14,1182) = 95\%$	14,0669	14,0538
16:00 – 17:00	$P(13,8002 \leqslant LS \leqslant 13,8985) = 95\%$	13,8493	13,8475
17:00 – 18:00	$P(9,9231 \leqslant LS \leqslant 10,0387) = 95\%$	9,9809	9,9596
18:00 – 19:00	$P(7,1715 \leqslant LS \leqslant 7,2507) = 95\%$	7,2111	7,2060
19:00 – 20:00	$P(4,8489 \leqslant LS \leqslant 4,8912) = 95\%$	4,8701	4,8715
20:00 – 21:00	$P(4,9814 \leqslant LS \leqslant 5,0337) = 95\%$	5,0075	5,0024

Tabela 4.4: Estimativa de parâmetro: média dos valores máximos da carga de processamento do servidor

Período	Intervalo de confiança	$\overline{LS^*_{max}}$	$\overline{LS_{max}}$
07:00 - 08:00	$P(7,5575 \leqslant LS_{max} \leqslant 9,5906) = 95\%$	8,5740	8,3490
08:00 – 09:00	$P(11,7754 \leqslant LS_{max} \leqslant 13,5913) = 95\%$	12,6833	12,2667
09:00 – 10:00	$P(16,2523 \leqslant LS_{max} \leqslant 18,1727) = 95\%$	17,2125	16,7333
10:00 – 11:00	$P(17,7286 \leqslant LS_{max} \leqslant 19,8592) = 95\%$	18,7939	18,3707
11:00 – 12:00	$P(16,9089 \leqslant LS_{max} \leqslant 18,5544) = 95\%$	17,7317	17,7750
12:00 – 13:00	$P(19,8815 \leqslant LS_{max} \leqslant 22,5756) = 95\%$	21,2285	20,6177
13:00 – 14:00	$P(19,7843 \leqslant LS_{max} \leqslant 23,5244) = 95\%$	21,6543	21,6547
14:00 – 15:00	$P(17,2937 \leqslant LS_{max} \leqslant 19,5598) = 95\%$	18,4268	18,1147
15:00 – 16:00	$P(18,4080 \leqslant LS_{max} \leqslant 19,8070) = 95\%$	19,1075	18,8917
16:00 – 17:00	$P(18,1254 \leqslant LS_{max} \leqslant 20,0698) = 95\%$	19,0976	19,0240
17:00 – 18:00	$P(14,0803 \leqslant LS_{max} \leqslant 15,1881) = 95\%$	14,6342	14,5167
18:00 – 19:00	$P(10,7646 \leqslant LS_{max} \leqslant 11,7454) = 95\%$	11,2550	11,0667
19:00 – 20:00	$P(7,0037 \leqslant LS_{max} \leqslant 7,5653) = 95\%$	7,2845	7,2147
20:00 – 21:00	$P(6,8618 \leqslant LS_{max} \leqslant 7,6115) = 95\%$	7,2367	7,1500

4.1.3 Uso de memória do roteador

Nessa subsessão são apresentados os resultados para a estimação do parâmetro média para o percentual de uso de memória do roteador, onde $\overline{MR}$ representa a sua média e $\overline{MR_{max}}$ a média dos valores máximos da variável de interesse. O gráfico de dispersão de $\overline{MR}$ e $\overline{MR_{max}}$ é apresentado na Figura 4.4, ao passo que os intervalos de confiança de $\overline{MR}$ e $\overline{MR_{max}}$ são exibidos nas Tabelas 4.5 e 4.6.

Sobre a Figura 4.4 cabe a observação quanto ao comportamento do uso de memória do roteador ao longo do período. Em um análise subjetiva, *a priori*, espera-se que tanto $\overline{MR}$ e $\overline{MR_{max}}$ tenham comportamento estável no período intermediário de medição. O que se observa é que há um grande crescimento e uso da memória no período da manhã, e no período da tarde há o decréscimo desses valores e a estabilidade de ambas as variáveis. Aplicando apenas o conhecimento técnico do perfil da rede sem fazer uso de técnicas ou ferramentas estatísticas, a tendência é $\overline{MR}$ e $\overline{MR_{max}}$ terem valores máximos durante o dia, e aumento e diminuição de valores no início e fim do intervalo medido respectivamente. Isso retrataria a premissa de maior utilização dos recursos ao longo do dia. No entanto, nos experimentos envolvendo correlação, verificaremos se o comportamento de $\overline{MR}$ e $\overline{MR_{max}}$ está relacionado com outras variáveis.

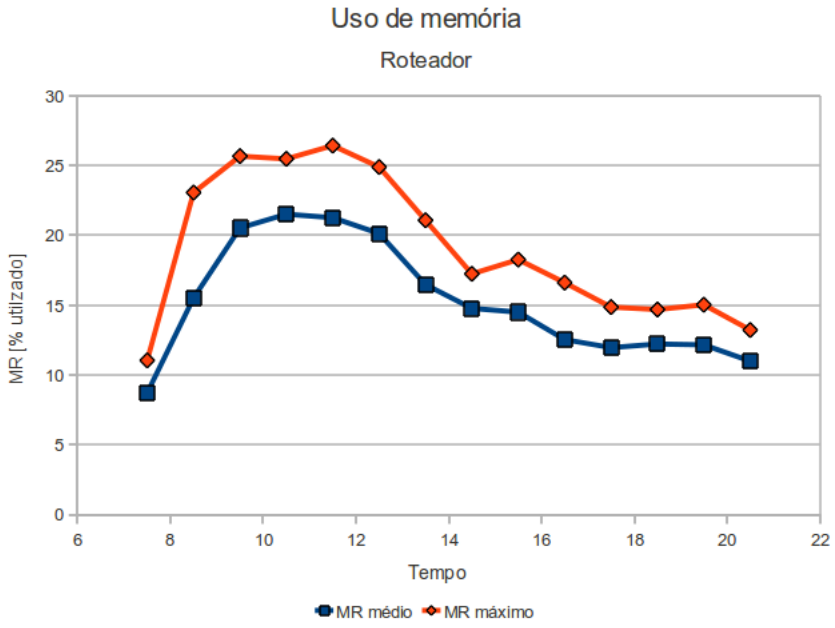

Figura 4.4: Estimativa de parâmetros: uso de memória do roteador

Tabela 4.5: Estimativa de parâmetro: média do uso de memória do roteador

Período	Intervalo de confiança	$\overline{MR^*}$	$\overline{MR}$
07:00 - 08:00	$P(8,6839 \leqslant MR \leqslant 8,7610) = 95\%$	8,7225	8,7153
08:00 – 09:00	$P(15,3802 \leqslant MR \leqslant 15,6550) = 95\%$	15,5176	15,5164
09:00 – 10:00	$P(20,3991 \leqslant MR \leqslant 20,6454) = 95\%$	20,5223	20,5197
10:00 – 11:00	$P(21,3944 \leqslant MR \leqslant 21,6207) = 95\%$	21,5075	21,4770
11:00 – 12:00	$P(21,1324 \leqslant MR \leqslant 21,3854) = 95\%$	21,2589	21,2373
12:00 – 13:00	$P(19,9906 \leqslant MR \leqslant 20,2639) = 95\%$	20,1273	20,1006
13:00 – 14:00	$P(16,3284 \leqslant MR \leqslant 16,6095) = 95\%$	16,4689	16,4398
14:00 – 15:00	$P(14,6425 \leqslant MR \leqslant 14,8936) = 95\%$	14,7681	14,7670
15:00 – 16:00	$P(14,3995 \leqslant MR \leqslant 14,5960) = 95\%$	14,4978	14,4949
16:00 – 17:00	$P(12,4347 \leqslant MR \leqslant 12,6314) = 95\%$	12,5331	12,5242
17:00 – 18:00	$P(11,8497 \leqslant MR \leqslant 12,0651) = 95\%$	11,9574	11,9245
18:00 – 19:00	$P(12,1108 \leqslant MR \leqslant 12,3200) = 95\%$	12,2154	12,2374
19:00 – 20:00	$P(12,0458 \leqslant MR \leqslant 12,2668) = 95\%$	12,1563	12,1397
20:00 – 21:00	$P(10,8870 \leqslant MR \leqslant 11,0968) = 95\%$	10,9919	10,9605

Tabela 4.6: Estimativa de parâmetro: média dos valores máximos do uso de memória do roteador

Período	Intervalo de confiança	$\overline{MR^*_{max}}$	$\overline{MR_{max}}$
07:00 - 08:00	$P(10,3181 \leqslant MR_{max} \leqslant 11,7727) = 95\%$	11,0454	10,9948
08:00 – 09:00	$P(21,2560 \leqslant MR_{max} \leqslant 24,8603) = 95\%$	23,0581	22,2415
09:00 – 10:00	$P(23,8891 \leqslant MR_{max} \leqslant 27,4311) = 95\%$	25,6601	25,2472
10:00 – 11:00	$P(23,9865 \leqslant MR_{max} \leqslant 26,9502) = 95\%$	25,4684	25,1765
11:00 – 12:00	$P(24,5092 \leqslant MR_{max} \leqslant 28,3299) = 95\%$	26,4195	25,7893
12:00 – 13:00	$P(22,7668 \leqslant MR_{max} \leqslant 27,0014) = 95\%$	24,8841	24,1705
13:00 – 14:00	$P(18,3875 \leqslant MR_{max} \leqslant 23,7568) = 95\%$	21,0722	20,2566
14:00 – 15:00	$P(15,1502 \leqslant MR_{max} \leqslant 19,3229) = 95\%$	17,2366	16,7273
15:00 – 16:00	$P(16,8758 \leqslant MR_{max} \leqslant 19,6416) = 95\%$	18,2587	17,7627
16:00 – 17:00	$P(14,7771 \leqslant MR_{max} \leqslant 18,4411) = 95\%$	16,6091	15,9702
17:00 – 18:00	$P(13,0172 \leqslant MR_{max} \leqslant 16,6941) = 95\%$	14,8557	14,0661
18:00 – 19:00	$P(12,4063 \leqslant MR_{max} \leqslant 16,9639) = 95\%$	14,6851	14,1949
19:00 – 20:00	$P(12,8570 \leqslant MR_{max} \leqslant 17,2006) = 95\%$	15,0288	14,2726
20:00 – 21:00	$P(11,3752 \leqslant MR_{max} \leqslant 15,0625) = 95\%$	13,2189	12,4906

4.1.4 Uso de memória do servidor

Nesse experimento foram analisadas a média do uso de memória do servidor ($\overline{MS}$) e sua respectiva média dos valores máximos ($\overline{MS_{max}}$). Os intervalos de confiança construídos estão apresentados nas Tabelas 4.7 e 4.8, representando a estimativa do parâmetro média para as variáveis $\overline{MS}$ e $\overline{MS_{max}}$ respectivamente.

A partir da observação da Figura 4.5, unicamente da dispersão apresentada no gráfico, verificamos que os dados relativos ao uso da memória do servidor apresenta certa instabilidade e variação ao longo do período medido. No entanto, em uma análise quantitativa das médias estimadas, verificamos que tanto a média quanto a média dos valores máximos estão compreendidos entre 36,25% e 36,45%. Na prática, esses percentuais significam o funcionamento estável do servidor analisado. Em uma conclusão matemática, a diferença entre $\overline{MS_{max}}$ e $\overline{MS}$ não ultrapassa dois décimos, o que significa que a média dos picos registrados para o uso da memória do servidor está próximo a sua média aritmética.

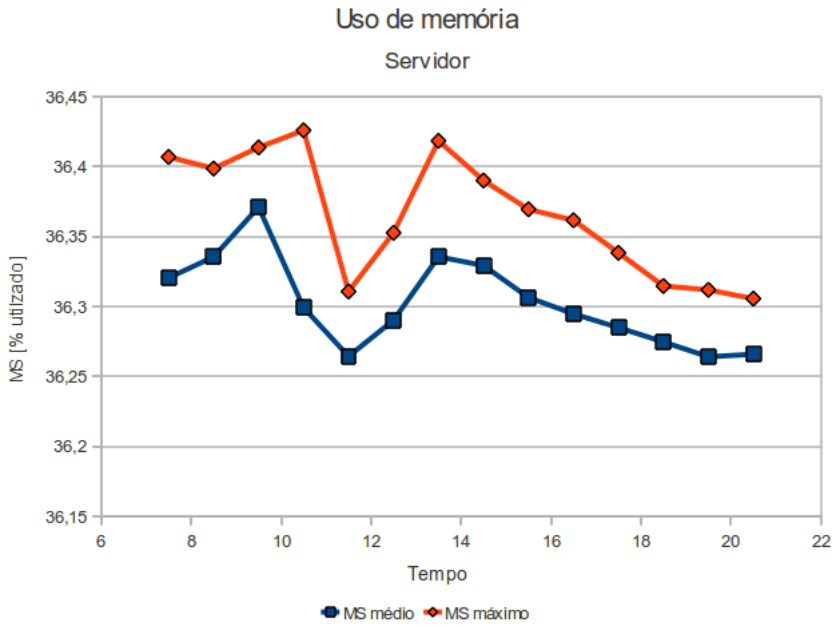

Figura 4.5: Estimativa de parâmetros: Uso de memória do servidor

Tabela 4.7: Estimativa de parâmetro: média do uso de memória do servidor

Período	Intervalo de confiança	$\overline{MS^*}$	$\overline{MS}$
07:00 - 08:00	$P(36,3153 \leqslant MS \leqslant 36,3254) = 95\%$	36,3203	36,3194
08:00 – 09:00	$P(36,3303 \leqslant MS \leqslant 36,3405) = 95\%$	36,3354	36,3342
09:00 – 10:00	$P(36,3659 \leqslant MS \leqslant 36,3760) = 95\%$	36,3709	36,3695
10:00 – 11:00	$P(36,2956 \leqslant MS \leqslant 36,3026) = 95\%$	36,2991	36,2990
11:00 – 12:00	$P(36,2623 \leqslant MS \leqslant 36,2662) = 95\%$	36,2643	36,2639
12:00 – 13:00	$P(36,2880 \leqslant MS \leqslant 36,2921) = 95\%$	36,2900	36,2898
13:00 – 14:00	$P(36,3329 \leqslant MS \leqslant 36,3380) = 95\%$	36,3354	36,3352
14:00 – 15:00	$P(36,3267 \leqslant MS \leqslant 36,3316) = 95\%$	36,3291	36,3283
15:00 – 16:00	$P(36,3041 \leqslant MS \leqslant 36,3086) = 95\%$	36,3064	36,3054
16:00 – 17:00	$P(36,2925 \leqslant MS \leqslant 36,2967) = 95\%$	36,2946	36,2943
17:00 – 18:00	$P(36,2827 \leqslant MS \leqslant 36,2869) = 95\%$	36,2848	36,2846
18:00 – 19:00	$P(36,2725 \leqslant MS \leqslant 36,2767) = 95\%$	36,2746	36,2749
19:00 – 20:00	$P(36,2619 \leqslant MS \leqslant 36,2659) = 95\%$	36,2639	36,2640
20:00 – 21:00	$P(36,2637 \leqslant MS \leqslant 36,2678) = 95\%$	36,2657	36,2654

Tabela 4.8: Estimativa de parâmetro: média dos valores máximos do uso de memória do servidor

Período	Intervalo de confiança	$\overline{MS^*_{max}}$	$\overline{MS_{max}}$
07:00 - 08:00	$P(36,3303 \leqslant MS_{max} \leqslant 36,4832) = 95\%$	36,4068	36,3698
08:00 – 09:00	$P(36,3125 \leqslant MS_{max} \leqslant 36,4846) = 95\%$	36,3985	36,3783
09:00 – 10:00	$P(36,3388 \leqslant MS_{max} \leqslant 36,4882) = 95\%$	36,4135	36,4091
10:00 – 11:00	$P(36,3503 \leqslant MS_{max} \leqslant 36,5012) = 95\%$	36,4258	36,4126
11:00 – 12:00	$P(36,2797 \leqslant MS_{max} \leqslant 36,3415) = 95\%$	36,3106	36,3039
12:00 – 13:00	$P(36,3260 \leqslant MS_{max} \leqslant 36,3791) = 95\%$	36,3526	36,3433
13:00 – 14:00	$P(36,3646 \leqslant MS_{max} \leqslant 36,4719) = 95\%$	36,4182	36,4047
14:00 – 15:00	$P(36,3543 \leqslant MS_{max} \leqslant 36,4257) = 95\%$	36,3900	36,3756
15:00 – 16:00	$P(36,3293 \leqslant MS_{max} \leqslant 36,4093) = 95\%$	36,3693	36,3650
16:00 – 17:00	$P(36,3340 \leqslant MS_{max} \leqslant 36,3891) = 95\%$	36,3615	36,3556
17:00 – 18:00	$P(36,3085 \leqslant MS_{max} \leqslant 36,3680) = 95\%$	36,3383	36,3223
18:00 – 19:00	$P(36,2836 \leqslant MS_{max} \leqslant 36,3455) = 95\%$	36,3146	36,3121
19:00 – 20:00	$P(36,2767 \leqslant MS_{max} \leqslant 36,3468) = 95\%$	36,3117	36,2963
20:00 – 21:00	$P(36,2800 \leqslant MS_{max} \leqslant 36,3310) = 95\%$	36,3055	36,2989

4.1.5 *Throughput* do *link* de *internet*, *download*

Nessa subsessão são estudadas as médias do *throughput* do *link* de *internet* (*download*) e a média dos seus valores máximos, representados respectivamente pelas variáveis $\overline{TRD1}$ e $\overline{TRD1_{max}}$.

A observação da Figura 4.6, que representa o gráfico de dispersão de $\overline{TRD1}$ e $\overline{TRD1_{max}}$, nos remete ao perfil de rede congestionado e já identificado nos capítulos anteriores. A largura de banda para o *throughput* do *link* de *internet* tem tamanho de 6Mbps, de maneira que tanto $\overline{TRD1}$ quanto $\overline{TRD1_{max}}$ estejam próximos a esse limite. A única ressalva seria quanto ao períodos iniciais e finais de medição, onde o *throughput* do *link* de *internet* (*download*) apresenta maior variação.

O raciocínio apresentado para as variáveis $\overline{TRD1}$ e $\overline{TRD1_{max}}$ no gráfico de dispersão Figura 4.6 também é refletido nos intervalos de confiança construídos. Em uma análise mais cuidadosa, vemos nas Tabelas 4.9 e 4.10 que a amplitude dos intervalos (diferença entre limite superior e inferior) é menor nos períodos intermediários de medição. Em outras palavras, embora a chance de acerto esteja fixada em 95%, $\overline{TRD1}$ e $\overline{TRD1_{max}}$ estão contidos em faixas de menores amplitudes. Na prática, essa análise permite inferir quanto a estabilidade do *throughput* do *link* de *internet* (*download*).

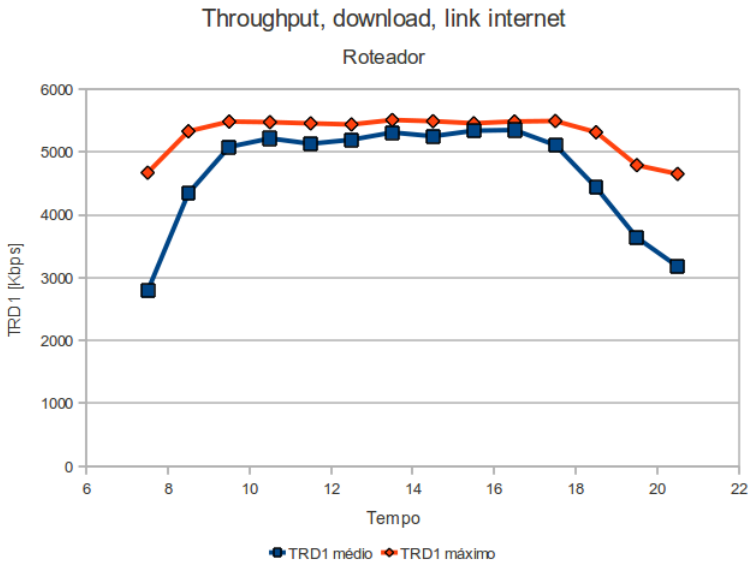

Figura 4.6: Estimativa de parâmetro: *throughput*, *download*, *link internet*

Tabela 4.9: Estimativa de parâmetro: média do *throughput*, *download*, *link internet*

Período	Intervalo de confiança	$\overline{TRD1^*}$	$\overline{TRD1}$
07:00 - 08:00	$P(2.762,11 \leqslant TRD1 \leqslant 2.817,15) = 95\%$	2.789,63	2.792,30
08:00 – 09:00	$P(4.318,61 \leqslant TRD1 \leqslant 4.364,55) = 95\%$	4.341,58	4.339,76
09:00 – 10:00	$P(5.065,55 \leqslant TRD1 \leqslant 5.091,76) = 95\%$	5.078,66	5.080,59
10:00 – 11:00	$P(5.205,44 \leqslant TRD1 \leqslant 5.221,86) = 95\%$	5.213,65	5.213,62
11:00 – 12:00	$P(5.119,90 \leqslant TRD1 \leqslant 5.137,49) = 95\%$	5.128,69	5.124,97
12:00 – 13:00	$P(5.183,36 \leqslant TRD1 \leqslant 5.201,82) = 95\%$	5.192,59	5.190,76
13:00 – 14:00	$P(5.299,31 \leqslant TRD1 \leqslant 5.317,08) = 95\%$	5.308,20	5.306,41
14:00 – 15:00	$P(5.235,84 \leqslant TRD1 \leqslant 5.261,84) = 95\%$	5.248,84	5.245,58
15:00 – 16:00	$P(5.330,63 \leqslant TRD1 \leqslant 5.342,34) = 95\%$	5.336,48	5.335,78
16:00 – 17:00	$P(5.342,63 \leqslant TRD1 \leqslant 5.351,25) = 95\%$	5.346,94	5.346,80
17:00 – 18:00	$P(5.098,52 \leqslant TRD1 \leqslant 5.117,14) = 95\%$	5.107,83	5.110,12
18:00 – 19:00	$P(4.413,45 \leqslant TRD1 \leqslant 4.454,16) = 95\%$	4.433,80	4.433,52
19:00 – 20:00	$P(3.614,77 \leqslant TRD1 \leqslant 3.657,32) = 95\%$	3.636,04	3.630,88
20:00 – 21:00	$P(3.154,18 \leqslant TRD1 \leqslant 3.201,33) = 95\%$	3.177,75	3.179,41

Tabela 4.10: Estimativa de parâmetro: média dos valores máximos do *throughput*, *download*, *link internet*

Período	Intervalo de confiança	$\overline{TRD1^*_{max}}$	$\overline{TRD1_{max}}$
07:00 - 08:00	$P(4.466,98 \leqslant TRD1_{max} \leqslant 4.870,62) = 95\%$	4.668,80	4.625,42
08:00 – 09:00	$P(5.246,35 \leqslant TRD1_{max} \leqslant 5.407,96) = 95\%$	5.327,15	5.312,32
09:00 – 10:00	$P(5.466,25 \leqslant TRD1_{max} \leqslant 5.497,23) = 95\%$	5.481,74	5.480,11
10:00 – 11:00	$P(5.444,63 \leqslant TRD1_{max} \leqslant 5.499,55) = 95\%$	5.472,09	5.462,49
11:00 – 12:00	$P(5.423,98 \leqslant TRD1_{max} \leqslant 5.480,46) = 95\%$	5.452,22	5.441,61
12:00 – 13:00	$P(5.407,51 \leqslant TRD1_{max} \leqslant 5.464,01) = 95\%$	5.435,76	5.424,24
13:00 – 14:00	$P(5.504,11 \leqslant TRD1_{max} \leqslant 5.510,18) = 95\%$	5.507,14	5.505,18
14:00 – 15:00	$P(5.471,77 \leqslant TRD1_{max} \leqslant 5.502,95) = 95\%$	5.487,36	5.485,31
15:00 – 16:00	$P(5.415,94 \leqslant TRD1_{max} \leqslant 5.492,27) = 95\%$	5.454,11	5.439,80
16:00 – 17:00	$P(5.462,05 \leqslant TRD1_{max} \leqslant 5.504,74) = 95\%$	5.483,39	5.479,37
17:00 – 18:00	$P(5.478,19 \leqslant TRD1_{max} \leqslant 5.502,76) = 95\%$	5.490,48	5.487,44
18:00 – 19:00	$P(5.262,54 \leqslant TRD1_{max} \leqslant 5.362,44) = 95\%$	5.312,49	5.293,84
19:00 – 20:00	$P(4.591,85 \leqslant TRD1_{max} \leqslant 4.982,94) = 95\%$	4.787,40	4.726,05
20:00 – 21:00	$P(4.425,92 \leqslant TRD1_{max} \leqslant 4.872,63) = 95\%$	4.649,28	.603,66

4.1.6 *Throughput* do *link* de *internet*, *upload*

Nesse experimento são estudadas as médias do *throughput* do *link* de *internet* (*upload*) e a média dos seus valores máximos, representados respectivamente pelas variáveis $\overline{TRU1}$ e $\overline{TRU1_{max}}$. O gráfico de dispersão dessas variáveis está ilustrado na Figura 4.7, assim como os intervalos de confiança de $\overline{TRU1}$ e $\overline{TRU1_{max}}$ são apresentados nas Tabelas 4.11 e 4.12 respectivamente.

Ao observar o gráfico de dispersão na Figura 4.7 podemos averiguar que existe uma tendência do *throughput* do *link* de *internet* (*upload*) ser maior nos intervalos intermediários de medição, o que atende a crença inicial de utilização dos recursos nesse período de tempo. Adicionalmente, $\overline{TRU1}$ e $\overline{TRU1_{max}}$ apresentam mesmo comportamento, o que implica em estabilidade ao longo do tempo dessa variável.

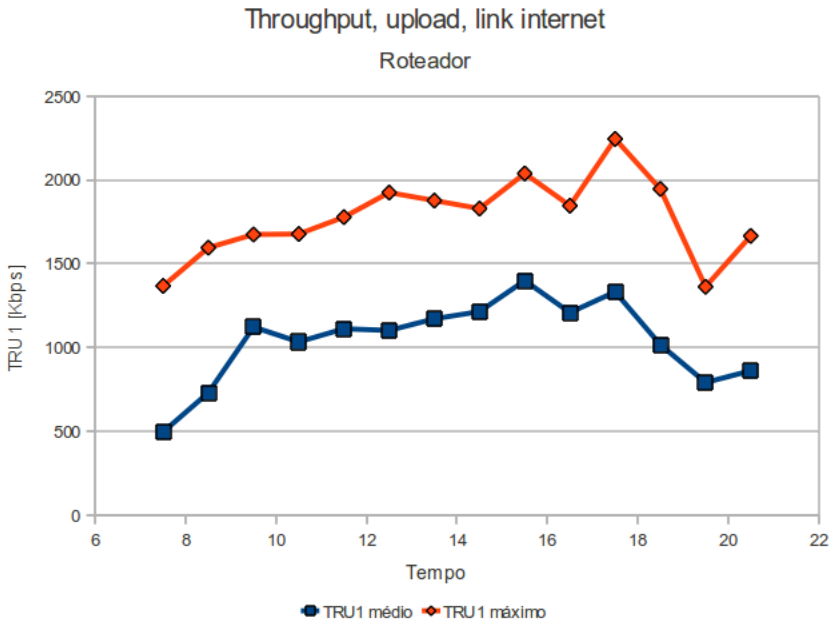

Figura 4.7: Estimativa de parâmetros: *throughput*, *upload*, *link internet*

Tabela 4.11: Estimativa de parâmetro: média do *throughput*, *upload*, *link internet*

Período	Intervalo de confiança	$\overline{TRU1^*}$	$\overline{TRU1}$
07:00 - 08:00	$P(488,39 \leqslant TRU1 \leqslant 504,08) = 95\%$	496,23	495,72
08:00 – 09:00	$P(723,32 \leqslant TRU1 \leqslant 737,14) = 95\%$	730,23	731,29
09:00 – 10:00	$P(1.113,12 \leqslant TRU1 \leqslant 1.130,82) = 95\%$	1.121,97	1.119,74
10:00 – 11:00	$P(1.027,68 \leqslant TRU1 \leqslant 1.039,32) = 95\%$	1.033,50	1.032,08
11:00 – 12:00	$P(1.104,61 \leqslant TRU1 \leqslant 1.118,76) = 95\%$	1.111,69	1.110,69
12:00 – 13:00	$P(1.094,55 \leqslant TRU1 \leqslant 1.109,05) = 95\%$	1.101,80	1.098,94
13:00 – 14:00	$P(1.164,94 \leqslant TRU1 \leqslant 1.177,79) = 95\%$	1.171,37	1.170,28
14:00 – 15:00	$P(1.207,43 \leqslant TRU1 \leqslant 1.221,11) = 95\%$	1.214,27	1.211,98
15:00 – 16:00	$P(1.390,73 \leqslant TRU1 \leqslant 1.407,60) = 95\%$	1.399,17	1.396,06
16:00 – 17:00	$P(1.200,71 \leqslant TRU1 \leqslant 1.216,47) = 95\%$	1.208,59	1.206,91
17:00 – 18:00	$P(1.320,52 \leqslant TRU1 \leqslant 1.343,16) = 95\%$	1.331,84	1.330,12
18:00 – 19:00	$P(1.006,96 \leqslant TRU1 \leqslant 1.026,09) = 95\%$	1.016,52	1.016,99
19:00 – 20:00	$P(783,80 \leqslant TRU1 \leqslant 78,43) = 95\%$	791,12	90,58
20:00 – 21:00	$P(852,21 \leqslant TRU1 \leqslant 87,54) = 95\%$	861,37	861,12

Tabela 4.12: Estimativa de parâmetro: média dos valores máximos do *throughput*, *upload*, *link internet*

Período	Intervalo de confiança	$\overline{TRU1^*_{max}}$	$\overline{TRU1_{max}}$
07:00 - 08:00	$P(1.177,19 \leqslant TRU1_{max} \leqslant 1.557,54) = 95\%$	1.367,36	1.320,66
08:00 – 09:00	$P(1.467,05 \leqslant TRU1_{max} \leqslant 1.723,89) = 95\%$	1.595,47	1.539,73
09:00 – 10:00	$P(1.537,33 \leqslant TRU1_{max} \leqslant 1.811,78) = 95\%$	1.674,55	1.626,99
10:00 – 11:00	$P(1.589,01 \leqslant TRU1_{max} \leqslant 1.765,97) = 95\%$	1.677,49	1.665,74
11:00 – 12:00	$P(1.656,03 \leqslant TRU1_{max} \leqslant 1.900,53) = 95\%$	1.778,28	1.743,10
12:00 – 13:00	$P(1.836,88 \leqslant TRU1_{max} \leqslant 2.012,52) = 95\%$	1.924,70	1.896,64
13:00 – 14:00	$P(1.767,79 \leqslant TRU1_{max} \leqslant 1.986,22) = 95\%$	1.877,00	1.847,71
14:00 – 15:00	$P(1.711,98 \leqslant TRU1_{max} \leqslant 1.946,12) = 95\%$	1.829,05	1.811,13
15:00 – 16:00	$P(1.907,08 \leqslant TRU1_{max} \leqslant 2.170,50) = 95\%$	2.038,79	2.055,55
16:00 – 17:00	$P(1.638,82 \leqslant TRU1_{max} \leqslant 2.053,20) = 95\%$	1.846,01	1.800,94
17:00 – 18:00	$P(.105,64 \leqslant TRU1_{max} \leqslant 2.384,21) = 95\%$	2.244,92	2.182,47
18:00 – 19:00	$P(1.786,80 \leqslant TRU1_{max} \leqslant 2.104,99) = 95\%$	1.945,90	1.908,13
19:00 – 20:00	$P(1.227,19 \leqslant TRU1_{max} \leqslant 1.498,40) = 95\%$	1.362,80	1.301,41
20:00 – 21:00	$P(1.483,79 \leqslant TRU1_{max} \leqslant 1.848,25) = 95\%$	1.666,02	1.576,03

4.1.7 *Throughput* do *link* institucional, *download*

Esse experimento consiste na análise do estimador média para o parâmetro *throughput* do *link* institucional (*download*). O *link* em questão, conforme abordado no subsessão 3.1.2 sobre a descrição do ambiente analisado, contém conexões de serviços institucionais do CEFET-MG, tais como *email*, banco de dados e páginas *web*. As variáveis $\overline{TRD2}$ e $\overline{TRD2_{max}}$ representam, respectivamente, a média e a média dos valores máximos do *throughput* do *link* institucional (*download*).

A partir da inferência da Figura 4.8 observamos grande variação da dispersão entre $\overline{TRD2}$ e $\overline{TRD2_{max}}$ nos intervalos intermediários de medição. Embora o comportamento de $\overline{TRD2}$ acompanhe a premissa inicial de utilização de recursos nesse período, a disposição de $\overline{TRD2_{max}}$ permite concluir sobre o uso desse *link* conforme horários de expediente da instituição. Durante o intervalo de 11:00 e 13:00 a média máxima do *throughput* do *link* institucional (*download*) tem visível queda em comparação aos períodos adjacentes.

A Tabela 4.13 mostra os intervalos de confiança construídos para $\overline{TRD2}$, assim como a Tabela 4.14 apresenta as estimavas de parâmetros para $\overline{TRD2_{max}}$.

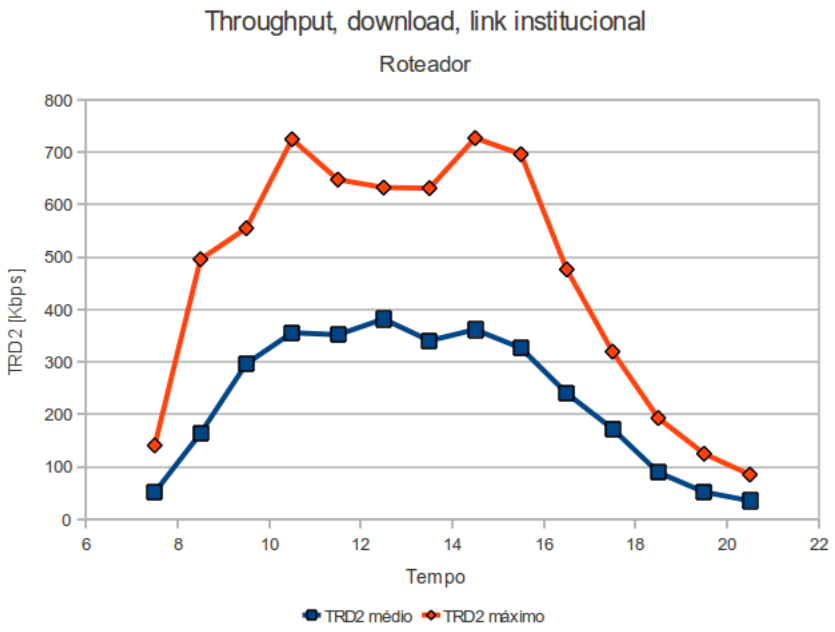

Figura 4.8: Estimativa de parâmetros: *throughput*, *download*, *link* institucional

Tabela 4.13: Estimativa de parâmetro: média do *throughput*, *download*, *link* institucional

Período	Intervalo de confiança	$\overline{TRD2^*}$	$\overline{TRD2}$
07:00 - 08:00	$P(50,28 \leqslant TRD2 \leqslant 52,18) = 95\%$	51,23	51,18
08:00 – 09:00	$P(160,75 \leqslant TRD2 \leqslant 166,61) = 95\%$	163,68	163,71
09:00 – 10:00	$P(292,81 \leqslant TRD2 \leqslant 300,21) = 95\%$	296,51	295,55
10:00 – 11:00	$P(351,35 \leqslant TRD2 \leqslant 359,81) = 95\%$	355,58	355,00
11:00 – 12:00	$P(347,72 \leqslant TRD2 \leqslant 355,81) = 95\%$	351,77	350,91
12:00 – 13:00	$P(376,46 \leqslant TRD2 \leqslant 386,93) = 95\%$	381,70	380,00
13:00 – 14:00	$P(334,96 \leqslant TRD2 \leqslant 345,78) = 95\%$	340,37	340,76
14:00 – 15:00	$P(356,83 \leqslant TRD2 \leqslant 366,19) = 95\%$	361,51	361,54
15:00 – 16:00	$P(322,22 \leqslant TRD2 \leqslant 330,63) = 95\%$	326,42	326,25
16:00 – 17:00	$P(237,05 \leqslant TRD2 \leqslant 242,93) = 95\%$	239,99	240,17
17:00 – 18:00	$P(169,88 \leqslant TRD2 \leqslant 175,08) = 95\%$	172,48	171,74
18:00 – 19:00	$P(88,16 \leqslant TRD2 \leqslant 90,19) = 95\%$	89,17	89,12
19:00 – 20:00	$P(51,14 \leqslant TRD2 \leqslant 52,49) = 95\%$	51,81	51,65
20:00 – 21:00	$P(34,69 \leqslant TRD2 \leqslant 35,57) = 95\%$	35,13	34,96

Tabela 4.14: Estimativa de parâmetro: média dos valores máximos do *throughput*, *download*, *link* institucional

Período	Intervalo de confiança	$\overline{TRD2^*_{max}}$	$\overline{TRD2_{max}}$
07:00 - 08:00	$P(120,18 \leqslant TRD2_{max} \leqslant 161,23) = 95\%$	140,71	135,17
08:00 – 09:00	$P(432,90 \leqslant TRD2_{max} \leqslant 558,34) = 95\%$	495,62	482,94
09:00 – 10:00	$P(494,98 \leqslant TRD2_{max} \leqslant 614,69) = 95\%$	554,83	545,88
10:00 – 11:00	$P(635,12 \leqslant TRD2_{max} \leqslant 813,82) = 95\%$	724,47	701,09
11:00 – 12:00	$P(582,90 \leqslant TRD2_{max} \leqslant 712,39) = 95\%$	647,64	643,27
12:00 – 13:00	$P(554,98 \leqslant TRD2_{max} \leqslant 709,98) = 95\%$	632,48	631,01
13:00 – 14:00	$P(524,25 \leqslant TRD2_{max} \leqslant 738,04) = 95\%$	631,14	611,92
14:00 – 15:00	$P(595,33 \leqslant TRD2_{max} \leqslant 858,21) = 95\%$	726,77	690,63
15:00 – 16:00	$P(598,18 \leqslant TRD2_{max} \leqslant 793,67) = 95\%$	695,93	647,03
16:00 – 17:00	$P(425,44 \leqslant TRD2_{max} \leqslant 27,11) = 95\%$	476,28	461,46
17:00 – 18:00	$P(271,69 \leqslant TRD2_{max} \leqslant 367,34) = 95\%$	319,51	305,04
18:00 – 19:00	$P(173,47 \leqslant TRD2_{max} \leqslant 212,13) = 95\%$	192,80	189,94
19:00 – 20:00	$P(106,58 \leqslant TRD2_{max} \leqslant 143,15) = 95\%$	124,87	122,01
20:00 – 21:00	$P(77,34 \leqslant TRD2_{max} \leqslant 92,65) = 95\%$	85,00	81,89

4.1.8 *Throughput* do *link* institucional, *upload*

O intervalo de confiança construído nessa sessão contempla a análise do *throughput* do *link* institucional (*upload*). A variável $\overline{TRU2}$ representa a média do parâmetro em questão, assim como $\overline{TRU2_{max}}$ a média dos valores máximos. As Tabelas 4.15 e 4.16 ilustram, respectivamente, as estimativas de parâmetros de $\overline{TRU2}$ e $\overline{TRU2_{max}}$.

Na Figura 4.1.14 verificamos que $\overline{TRU2}$ e $\overline{TRU2_{max}}$ são maiores nos intervalos intermediários de medição, comportamento do qual se assemelha à crença já descrita sobre utilização dos recursos nesse período. No entanto, os intervalos iniciais e finais merecem atenção quanto à sua apresentação no gráfico de dispersão. No intervalo entre 07:00 e 08:00, $\overline{TRU2}$ e $\overline{TRU2_{max}}$ apresentam valores muito baixos e próximos entre si, o que representa considerável inatividade do *link* nesses horários. Já no intervalo de 20:00 às 21:00 é notado um alto valor para $\overline{TRU2_{max}}$, embora $\overline{TRU2}$ mantenha-se relativamente baixo. No contexto da instituição isso pode representar maior atividade de setores que funcionam no período noturno, como exemplo a biblioteca, na necessidade do envio de requisições no uso do seu sistema de reserva de livros.

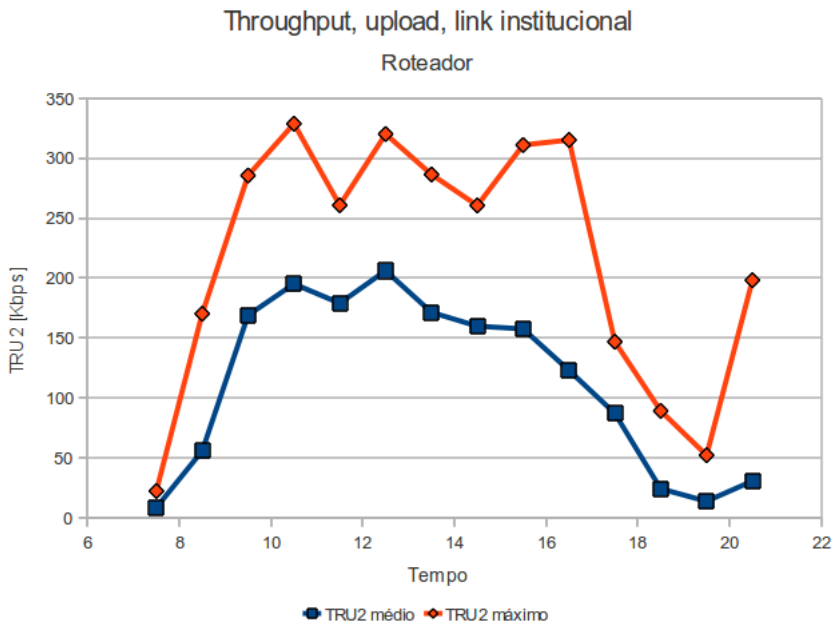

Figura 4.9: Estimativa de parâmetros: *throughput*, *upload*, *link* institucional

Tabela 4.15: Estimativa de parâmetro: média do *throughput*, *upload*, *link* institucional

Período	Intervalo de confiança	$\overline{TRU2^*}$	$\overline{TRU2}$
07:00 - 08:00	$P(8,27 \leqslant TRU2 \leqslant 8,50) = 95\%$	8,39	8,35
08:00 – 09:00	$P(53,85 \leqslant TRU2 \leqslant 57,49) = 95\%$	55,67	55,78
09:00 – 10:00	$P(165,80 \leqslant TRU2 \leqslant 172,02) = 95\%$	168,91	168,42
10:00 – 11:00	$P(191,45 \leqslant TRU2 \leqslant 198,65) = 95\%$	195,05	195,33
11:00 – 12:00	$P(175,38 \leqslant TRU2 \leqslant 182,53) = 95\%$	78,95	178,36
12:00 – 13:00	$P(201,02 \leqslant TRU2 \leqslant 211,21) = 95\%$	206,11	204,11
13:00 – 14:00	$P(167,47 \leqslant TRU2 \leqslant 175,15) = 95\%$	171,31	171,33
14:00 – 15:00	$P(156,42 \leqslant TRU2 \leqslant 163,25) = 95\%$	159,84	158,44
15:00 – 16:00	$P(154,30 \leqslant TRU2 \leqslant 160,71) = 95\%$	157,51	158,10
16:00 – 17:00	$P(120,19 \leqslant TRU2 \leqslant 125,67) = 95\%$	122,93	122,58
17:00 – 18:00	$P(84,80 \leqslant TRU2 \leqslant 89,22) = 95\%$	87,01	86,41
18:00 – 19:00	$P(23,41 \leqslant TRU2 \leqslant 24,68) = 95\%$	24,04	23,98
19:00 – 20:00	$P(13,18 \leqslant TRU2 \leqslant 14,11) = 95\%$	13,65	13,65
20:00 – 21:00	$P(27,61 \leqslant TRU2 \leqslant 33,37) = 95\%$	30,49	30,08

Tabela 4.16: Estimativa de parâmetro: média dos valores máximos do *throughput*, *upload*, *link* institucional

Período	Intervalo de confiança	$\overline{TRU2^*_{max}}$	$\overline{TRU2_{max}}$
07:00 - 08:00	$P(18,03 \leqslant TRU2_{max} \leqslant 26,30) = 95\%$	22,16	20,79
08:00 – 09:00	$P(127,57 \leqslant TRU2_{max} \leqslant 212,73) = 95\%$	170,15	160,75
09:00 – 10:00	$P(230,61 \leqslant TRU2_{max} \leqslant 340,50) = 95\%$	285,56	268,97
10:00 – 11:00	$P(260,52 \leqslant TRU2_{max} \leqslant 397,10) = 95\%$	328,81	296,07
11:00 – 12:00	$P(202,74 \leqslant TRU2_{max} \leqslant 318,84) = 95\%$	260,79	250,14
12:00 – 13:00	$P(213,54 \leqslant TRU2_{max} \leqslant 426,86) = 95\%$	320,20	294,75
13:00 – 14:00	$P(199,64 \leqslant TRU2_{max} \leqslant 373,29) = 95\%$	286,47	277,33
14:00 – 15:00	$P(187,90 \leqslant TRU2_{max} \leqslant 333,19) = 95\%$	260,54	242,98
15:00 – 16:00	$P(251,90 \leqslant TRU2_{max} \leqslant 370,15) = 95\%$	311,03	284,85
16:00 – 17:00	$P(234,22 \leqslant TRU2_{max} \leqslant 396,27) = 95\%$	315,24	283,73
17:00 – 18:00	$P(107,72 \leqslant TRU2_{max} \leqslant 185,79) = 95\%$	146,75	142,85
18:00 – 19:00	$P(68,59 \leqslant TRU2_{max} \leqslant 109,59) = 95\%$	89,09	86,42
19:00 – 20:00	$P(31,97 \leqslant TRU2_{max} \leqslant 72,13) = 95\%$	52,05	51,27
20:00 – 21:00	$P(93,13 \leqslant TRU2_{max} \leqslant 303,12) = 95\%$	198,13	159,66

4.1.9 *Throughput* do *link ethernet*, *download*

Nessa subsessão analisaremos o *throughput* do *link ethernet* (*download*) onde as variáveis $\overline{TSD}$ e $\overline{TSD_{max}}$ representam a média e a média dos valores máximos do parâmetro em questão. Vale ressaltar que, conforme a apresentação do ambiente analisado na subsessão 3.1.2, o *link ethernet* em estudo nesse momento representa o *link* de *internet* e o *link* institucional do CEFET-MG. No entanto, essa conexão não é mais efetivada pelo roteador e sim pelo servidor de *firewall* da instituição.

Na prática, como o *link* de internet é predominante sobre o *link* institucional, o *link ethernet* aqui analisado terá comportamento semelhante ao apresentado na subsessão 4.1.5. Essa conclusão pode ser facilmente verificada ao analisar o gráfico de dispersão das variáveis $\overline{TSD}$ e $\overline{TSD_{max}}$ na Figura 4.10, e o gráfico na Figura 4.6 .

Os intervalos de confiança das variáveis $\overline{TSD}$ e $\overline{TSD_{max}}$ estão apresentados, respectivamente, nas Tabelas 4.17 e 4.18.

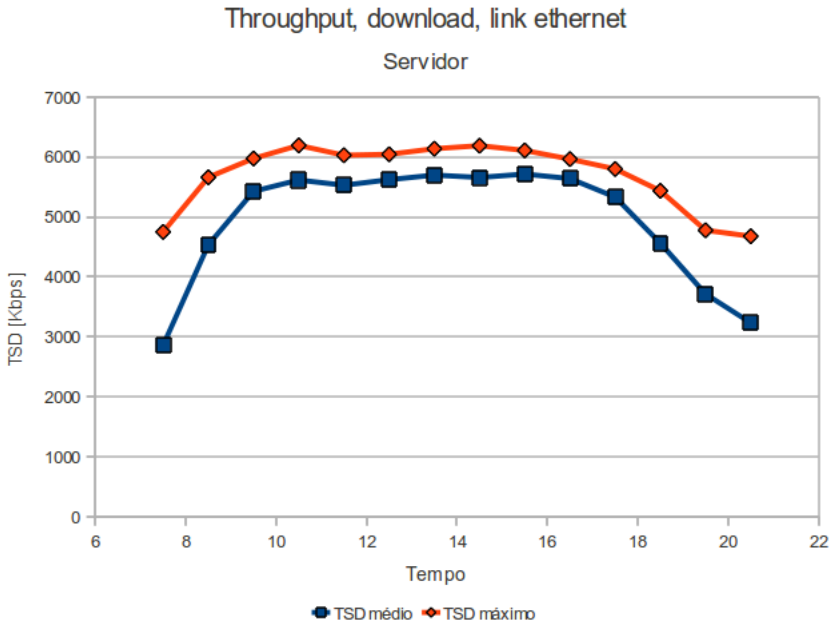

Figura 4.10: Estimativa de parâmetros: *throughput*, *download*, *link ethernet*

Tabela 4.17: Estimativa de parâmetro: média do *throughput*, *download*, *link ethernet*

Período	Intervalo de confiança	$\overline{TSD^*}$	$\overline{TSD}$
07:00 - 08:00	$P(2.830,74 \leqslant TSD \leqslant 2.887,04) = 95\%$	2.858,89	2.855,16
08:00 – 09:00	$P(4.511,77 \leqslant TSD \leqslant 4.555,55) = 95\%$	4.533,66	4.533,49
09:00 – 10:00	$P(5.412,70 \leqslant TSD \leqslant 5.439,99) = 95\%$	5.426,35	5.422,63
10:00 – 11:00	$P(5.604,05 \leqslant TSD \leqslant 5.624,95) = 95\%$	5.614,50	5.616,44
11:00 – 12:00	$P(5.517,04 \leqslant TSD \leqslant 5.537,79) = 95\%$	5.527,42	5.527,23
12:00 – 13:00	$P(5.609,21 \leqslant TSD \leqslant 5.632,59) = 95\%$	5.620,90	5.618,42
13:00 – 14:00	$P(5.686,49 \leqslant TSD \leqslant 5.707,96) = 95\%$	5.697,23	5.698,01
14:00 – 15:00	$P(5.643,01 \leqslant TSD \leqslant 5.672,58) = 95\%$	5.657,80	5.657,79
15:00 – 16:00	$P(5.706,67 \leqslant TSD \leqslant 5.721,93) = 95\%$	5.714,30	5.713,80
16:00 – 17:00	$P(5.634,11 \leqslant TSD \leqslant 5.644,38) = 95\%$	5.639,25	5.639,30
17:00 – 18:00	$P(5.321,39 \leqslant TSD \leqslant 5.339,63) = 95\%$	5.330,51	5.327,53
18:00 – 19:00	$P(4.543,25 \leqslant TSD \leqslant 4.580,37) = 95\%$	4.561,81	4.562,30
19:00 – 20:00	$P(3.691,11 \leqslant TSD \leqslant 3.735,07) = 95\%$	3.713,09	3.708,30
20:00 – 21:00	$P(3.208,04 \leqslant TSD \leqslant 3.254,37) = 95\%$	3.231,20	3.235,00

Tabela 4.18: Estimativa de parâmetro: média dos valores máximos do *throughput*, *download*, *link ethernet*

Período	Intervalo de confiança	$\overline{TSD^*_{max}}$	$\overline{TSD_{max}}$
07:00 - 08:00	$P(4.528,07 \leqslant TSD_{max} \leqslant 4.971,44) = 95\%$	4.749,75	4.715,58
08:00 – 09:00	$P(5.509,30 \leqslant TSD_{max} \leqslant 5.812,19) = 95\%$	5.660,75	5.608,84
09:00 – 10:00	$P(5.890,90 \leqslant TSD_{max} \leqslant 6.058,09) = 95\%$	5.974,49	5.968,86
10:00 – 11:00	$P(6.101,70 \leqslant TSD_{max} \leqslant 6.284,14) = 95\%$	6.192,92	6.149,59
11:00 – 12:00	$P(5.927,59 \leqslant TSD_{max} \leqslant 6.128,75) = 95\%$	6.028,17	6.010,85
12:00 – 13:00	$P(5.919,20 \leqslant TSD_{max} \leqslant 6.169,91) = 95\%$	6.044,55	6.027,63
13:00 – 14:00	$P(6.032,75 \leqslant TSD_{max} \leqslant 6.241,05) = 95\%$	6.136,90	6.112,61
14:00 – 15:00	$P(6.082,74 \leqslant TSD_{max} \leqslant 6.291,99) = 95\%$	6.187,36	6.175,38
15:00 – 16:00	$P(6.008,06 \leqslant TSD_{max} \leqslant 6.208,24) = 95\%$	6.108,15	6.069,95
16:00 – 17:00	$P(5.905,07 \leqslant TSD_{max} \leqslant 6.022,12) = 95\%$	5.963,59	5.959,11
17:00 – 18:00	$P(5.778,81 \leqslant TSD_{max} \leqslant 5.819,07) = 95\%$	5.798,94	5.795,74
18:00 – 19:00	$P(5.378,26 \leqslant TSD_{max} \leqslant 5.491,04) = 95\%$	5.434,65	5.418,87
19:00 – 20:00	$P(4.577,25 \leqslant TSD_{max} \leqslant 4.975,92) = 95\%$	4.776,58	4.754,15
20:00 – 21:00	$P(4.456,50 \leqslant TSD_{max} \leqslant 4.896,26) = 95\%$	4.676,38	4.592,58

4.1.10 *Throughput* do *link ethernet*, *upload*

Nessa sessão abordaremos o *throughput* do *link ethernet* (*upload*), onde a variável $\overline{TSU}$ representa a média do parâmetro e $\overline{TSU_{max}}$ a média máxima. Pelas mesmas razões apresentadas no experimento anterior (subsessão 4.1.9), o *throughput* do *link ethernet* (*upload*) terá comportamento semelhante ao *throughput* do *link* de *internet* (*upload*).

A Figura 4.11 ilustra o gráfico de dispersão das variáveis $\overline{TSU}$ e $\overline{TSU_{max}}$. As Tabelas 4.19 e 4.20 apresentam, respectivamente, os intervalos de confiança construídos para as variáveis $\overline{TSU}$ e $\overline{TSU_{max}}$.

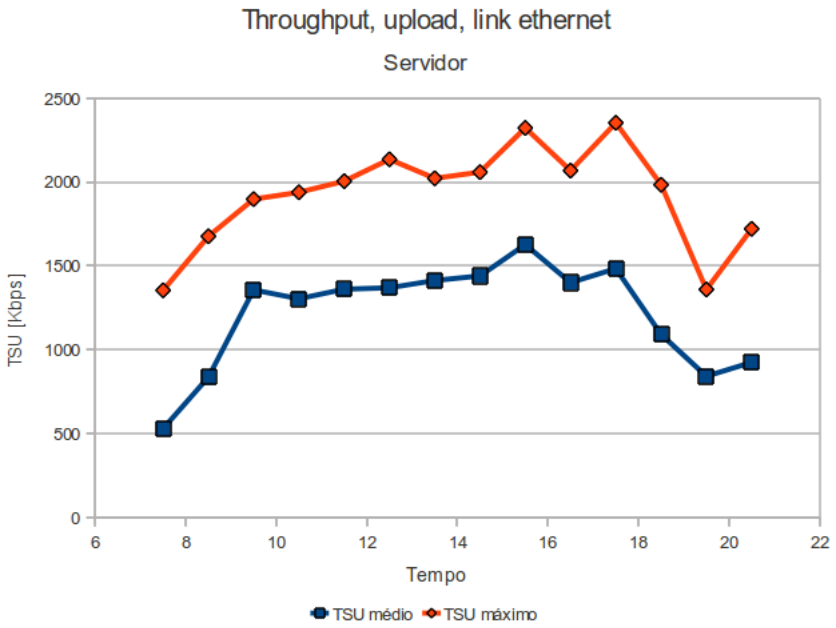

Figura 4.11: Estimativa de parâmetros: *throughput*, *upload*, *link ethernet*

Tabela 4.19: Estimativa de parâmetro: média do *throughput*, *upload*, *link ethernet*

Período	Intervalo de confiança	$\overline{TSU^*}$	$\overline{TSU}$
07:00 - 08:00	$P(521,20 \leqslant TSU \leqslant 537,66) = 95\%$	529,43	527,92
08:00 – 09:00	$P(828,25 \leqslant TSU \leqslant 843,05) = 95\%$	835,65	833,42
09:00 – 10:00	$P(1.346,88 \leqslant TSU \leqslant 1.365,50) = 95\%$	1.356,19	1.354,22
10:00 – 11:00	$P(1.296,10 \leqslant TSU \leqslant 1.311,28) = 95\%$	1.303,69	1.305,35
11:00 – 12:00	$P(1.352,30 \leqslant TSU \leqslant 1.368,64) = 95\%$	1.360,47	1.357,76
12:00 – 13:00	$P(1.360,74 \leqslant TSU \leqslant 1.377,45) = 95\%$	1.369,09	1.367,67
13:00 – 14:00	$P(1.405,99 \leqslant TSU \leqslant 1.418,93) = 95\%$	1.412,46	1.411,02
14:00 – 15:00	$P(1.432,52 \leqslant TSU \leqslant 1.446,93) = 95\%$	1.439,72	1.439,53
15:00 – 16:00	$P(1.617,60 \leqslant TSU \leqslant 1.635,53) = 95\%$	1.626,57	1.624,54
16:00 – 17:00	$P(1.392,27 \leqslant TSU \leqslant 1.408,13) = 95\%$	1.400,20	1.399,84
17:00 – 18:00	$P(1.471,11 \leqslant TSU \leqslant 1.495,57) = 95\%$	1.483,34	1.479,64
18:00 – 19:00	$P(1.081,75 \leqslant TSU \leqslant 1.100,26) = 95\%$	1.091,00	1.089,38
19:00 – 20:00	$P(833,75 \leqslant TSU \leqslant 849,30) = 95\%$	841,53	839,83
20:00 – 21:00	$P(915,28 \leqslant TSU \leqslant 933,33) = 95\%$	924,30	922,33

Tabela 4.20: Estimativa de parâmetro: média dos valores máximos do *throughput*, *upload*, *link ethernet*

Período	Intervalo de confiança	$\overline{TSU^*_{max}}$	$\overline{TSU_{max}}$
07:00 - 08:00	$P(1.201,27 \leqslant TSU_{max} \leqslant 1.507,20) = 95\%$	1.354,24	1.317,69
08:00 – 09:00	$P(1.574,19 \leqslant TSU_{max} \leqslant 1.777,75) = 95\%$	1.675,97	1.648,48
09:00 – 10:00	$P(1.739,15 \leqslant TSU_{max} \leqslant 2.056,67) = 95\%$	1.897,91	1.873,91
10:00 – 11:00	$P(1.836,45 \leqslant TSU_{max} \leqslant 2.040,53) = 95\%$	1.938,49	1.913,65
11:00 – 12:00	$P(1.907,75 \leqslant TSU_{max} \leqslant 2.102,07) = 95\%$	2.004,91	1.970,18
12:00 – 13:00	$P(2.048,02 \leqslant TSU_{max} \leqslant 2.222,29) = 95\%$	2.135,16	2.100,76
13:00 – 14:00	$P(1.951,10 \leqslant TSU_{max} \leqslant 2.092,87) = 95\%$	2.021,98	2.008,71
14:00 – 15:00	$P(1.956,95 \leqslant TSU_{max} \leqslant 2.161,08) = 95\%$	2.059,01	2.019,89
15:00 – 16:00	$P(2.173,13 \leqslant TSU_{max} \leqslant 2.471,59) = 95\%$	2.322,36	2.269,93
16:00 – 17:00	$P(1.880,83 \leqslant TSU_{max} \leqslant 2.254,93) = 95\%$	2.067,88	1.997,86
17:00 – 18:00	$P(2.172,74 \leqslant TSU_{max} \leqslant 2.533,08) = 95\%$	2.352,91	2.322,54
18:00 – 19:00	$P(1.837,39 \leqslant TSU_{max} \leqslant 2.129,98) = 95\%$	1.983,69	1.961,30
19:00 – 20:00	$P(1.220,20 \leqslant TSU_{max} \leqslant 1.497,63) = 95\%$	1.358,92	1.337,83
20:00 – 21:00	$P(1.589,80 \leqslant TSU_{max} \leqslant 1.851,63) = 95\%$	1.720,72	1.714,85

4.1.11 Número de pacotes do *link* de *internet*, *download*

A partir desse experimento até o último que encerra o estudo de estimação de parâmetros de variáveis de rede, abordaremos o número de pacotes passantes nos *links* ativos do roteador e *link* do servidor de *firewall* do CEFET-MG. *A priori*, podemos observar a semelhança do comportamento das variáveis de cada um desses seis experimentos sobre vazão de pacotes por segundo, com os estudos anteriormente apresentados que contemplam a análise do *throughput*. Essa semelhança foi observada apenas para o sentido *download*, ou seja, externo-interno. Na subsessão 4.3 desse presente capítulo, ao apresentar os resultados das correlações entre variáveis verificaremos porque o *throughput* e o número de pacotes se assemelham.

A média do número de pacotes do *link* de *internet* (*download*) é representada pela variável $\overline{PRD1}$, da mesma forma que a média dos valores máximos desse parâmetro é denotada pela variável $\overline{PRD1}_{max}$. Na Figura 4.12 os valores de $\overline{PRD1}$ e $\overline{PRD1}_{max}$ estão dispostos em um gráfico de dispersão. A construção dos intervalos de confiança para esse experimento é apresentado nas Tabelas 4.21 e 4.22. A estimativa de parâmetro do número de pacotes do *link* de *internet* (*download*) se assemelham com o experimento já discutido na sessão 4.1.5.

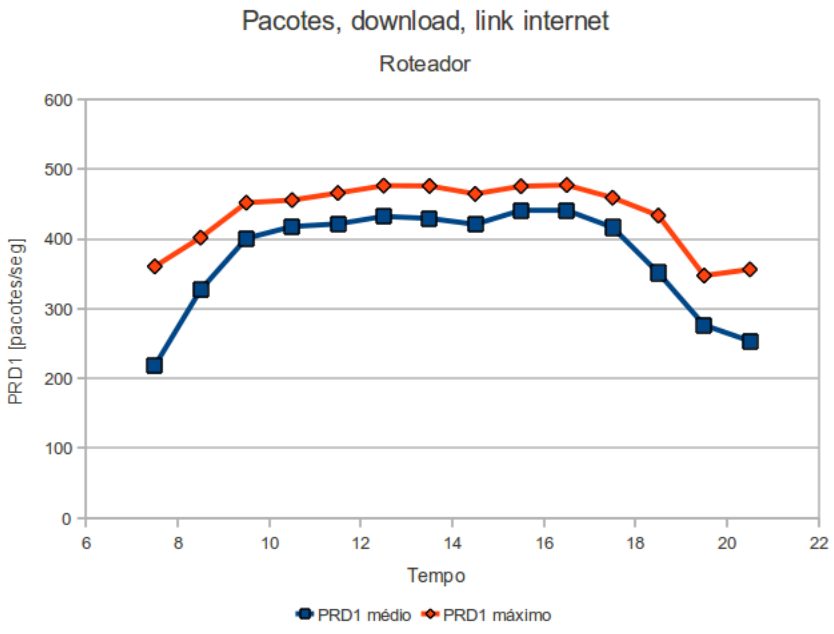

Figura 4.12: Estimativa de parâmetros: pacotes, *download*, *link internet*

Tabela 4.21: Estimativa de parâmetro: média do número de pacotes, *download*, *link internet*

Período	Intervalo de confiança	$\overline{PRD1^*}$	$\overline{PRD1}$
07:00 - 08:00	$P(216,31 \leqslant PRD1 \leqslant 219,98) = 95\%$	218,14	218,00
08:00 – 09:00	$P(325,44 \leqslant PRD1 \leqslant 328,37) = 95\%$	326,91	326,69
09:00 – 10:00	$P(399,89 \leqslant PRD1 \leqslant 401,46) = 95\%$	400,68	400,30
10:00 – 11:00	$P(416,93 \leqslant PRD1 \leqslant 418,17) = 95\%$	417,55	417,55
11:00 – 12:00	$P(420,74 \leqslant PRD1 \leqslant 422,02) = 95\%$	421,38	421,26
12:00 – 13:00	$P(431,42 \leqslant PRD1 \leqslant 432,80) = 95\%$	432,11	431,94
13:00 – 14:00	$P(428,48 \leqslant PRD1 \leqslant 429,95) = 95\%$	429,21	429,22
14:00 – 15:00	$P(419,73 \leqslant PRD1 \leqslant 421,79) = 95\%$	420,76	420,68
15:00 – 16:00	$P(440,40 \leqslant PRD1 \leqslant 441,12) = 95\%$	440,76	440,64
16:00 – 17:00	$P(440,26 \leqslant PRD1 \leqslant 441,21) = 95\%$	440,74	440,59
17:00 – 18:00	$P(415,64 \leqslant PRD1 \leqslant 417,08) = 95\%$	416,36	416,14
18:00 – 19:00	$P(349,93 \leqslant PRD1 \leqslant 352,60) = 95\%$	351,27	351,05
19:00 – 20:00	$P(275,09 \leqslant PRD1 \leqslant 277,69) = 95\%$	276,39	276,05
20:00 – 21:00	$P(252,16 \leqslant PRD1 \leqslant 255,11) = 95\%$	253,63	253,33

Tabela 4.22: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *download*, *link internet*

Período	Intervalo de confiança	$\overline{PRD1^*_{max}}$	$\overline{PRD1_{max}}$
07:00 - 08:00	$P(342,38 \leqslant PRD1_{max} \leqslant 378,50) = 95\%$	360,44	352,54
08:00 – 09:00	$P(392,90 \leqslant PRD1_{max} \leqslant 410,68) = 95\%$	401,79	397,90
09:00 – 10:00	$P(443,69 \leqslant PRD1_{max} \leqslant 459,82) = 95\%$	451,76	448,98
10:00 – 11:00	$P(450,85 \leqslant PRD1_{max} \leqslant 460,01) = 95\%$	455,43	454,29
11:00 – 12:00	$P(460,83 \leqslant PRD1_{max} \leqslant 470,37) = 95\%$	465,60	464,11
12:00 – 13:00	$P(472,44 \leqslant PRD1_{max} \leqslant 480,01) = 95\%$	476,22	476,00
13:00 – 14:00	$P(471,92 \leqslant PRD1_{max} \leqslant 479,68) = 95\%$	475,80	474,21
14:00 – 15:00	$P(460,99 \leqslant PRD1_{max} \leqslant 467,69) = 95\%$	464,34	463,11
15:00 – 16:00	$P(472,93 \leqslant PRD1_{max} \leqslant 477,81) = 95\%$	475,37	474,55
16:00 – 17:00	$P(469,11 \leqslant PRD1_{max} \leqslant 484,71) = 95\%$	476,91	474,17
17:00 – 18:00	$P(453,40 \leqslant PRD1_{max} \leqslant 464,56) = 95\%$	458,98	458,69
18:00 – 19:00	$P(424,71 \leqslant PRD1_{max} \leqslant 441,94) = 95\%$	433,33	429,80
19:00 – 20:00	$P(331,07 \leqslant PRD1_{max} \leqslant 363,51) = 95\%$	347,29	343,26
20:00 – 21:00	$P(340,81 \leqslant PRD1_{max} \leqslant 371,46) = 95\%$	356,14	349,28

4.1.12 Número de pacotes do *link* de *internet*, *upload*

Nessa sessão abordaremos a estimativa de parâmetro do número de pacotes do *link* de *internet* (*upload*), onde a média é representada por $\overline{PRU1}$ e a média máxima por $\overline{PRU1_{max}}$. A Figura 4.13 ilustra o gráfico da dispersão de $\overline{PRU1}$ e $\overline{PRU1_{max}}$, ao passo que as Tabelas 4.23 e 4.24 apresentam os resultados dos intervalos de confiança para ambas as variáveis, respectivamente.

A partir da Figura 4.13 verificamos a proximidade de $\overline{PRU1}$ e $\overline{PRU1_{max}}$ nos períodos intermediários. Além disso foi verificado seu crescimento e queda nos respectivos intervalos iniciais e finais do período de medição. Isso vai de encontro com a idéia inicial de maior uso dos recusos da rede em intervalos intermediários. Do raciocínio apresentado sobre a semelhança do comportamento *throughput* e número de pacotes de *download* na subsessão 4.1.11, não foi observado o mesmo pensamento para o sentido *upload*, com base nos resultados do experimento apresentados em 4.1.6.

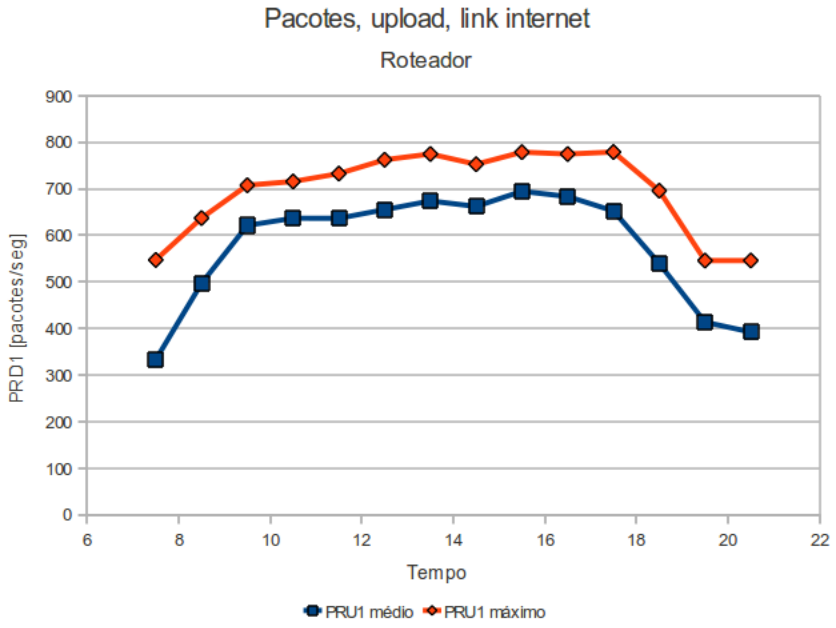

Figura 4.13: Estimativa de parâmetros: pacotes, *upload*, *link internet*

Tabela 4.23: Estimativa de parâmetro: média do número de pacotes, *upload*, *link internet*

Período	Intervalo de confiança	$\overline{PRU1^*}$	$\overline{PRU1}$
07:00 - 08:00	$P(329,91 \leqslant PRU1 \leqslant 335,59) = 95\%$	332,75	332,17
08:00 – 09:00	$P(494,18 \leqslant PRU1 \leqslant 499,27) = 95\%$	496,73	496,27
09:00 – 10:00	$P(619,66 \leqslant PRU1 \leqslant 622,58) = 95\%$	621,12	620,96
10:00 – 11:00	$P(635,24 \leqslant PRU1 \leqslant 637,83) = 95\%$	636,53	636,38
11:00 – 12:00	$P(635,06 \leqslant PRU1 \leqslant 637,74) = 95\%$	636,40	636,50
12:00 – 13:00	$P(653,64 \leqslant PRU1 \leqslant 656,45) = 95\%$	655,04	654,88
13:00 – 14:00	$P(672,31 \leqslant PRU1 \leqslant 675,06) = 95\%$	673,69	673,66
14:00 – 15:00	$P(660,31 \leqslant PRU1 \leqslant 663,57) = 95\%$	661,94	661,62
15:00 – 16:00	$P(693,48 \leqslant PRU1 \leqslant 695,25) = 95\%$	694,37	694,43
16:00 – 17:00	$P(682,06 \leqslant PRU1 \leqslant 684,27) = 95\%$	683,16	682,91
17:00 – 18:00	$P(650,45 \leqslant PRU1 \leqslant 654,14) = 95\%$	652,30	652,55
18:00 – 19:00	$P(537,24 \leqslant PRU1 \leqslant 541,61) = 95\%$	539,43	538,99
19:00 – 20:00	$P(411,32 \leqslant PRU1 \leqslant 415,65) = 95\%$	413,49	412,97
20:00 – 21:00	$P(389,83 \leqslant PRU1 \leqslant 394,46) = 95\%$	392,14	392,46

Tabela 4.24: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *upload*, *link internet*

Período	Intervalo de confiança	$\overline{PRU1^*_{max}}$	$\overline{PRU1_{max}}$
07:00 - 08:00	$P(511,85 \leqslant PRU1_{max} \leqslant 582,28) = 95\%$	547,07	536,44
08:00 – 09:00	$P(611,64 \leqslant PRU1_{max} \leqslant 662,26) = 95\%$	636,95	631,55
09:00 – 10:00	$P(695,68 \leqslant PRU1_{max} \leqslant 719,07) = 95\%$	707,37	704,53
10:00 – 11:00	$P(700,54 \leqslant PRU1_{max} \leqslant 730,29) = 95\%$	715,42	709,25
11:00 – 12:00	$P(722,71 \leqslant PRU1_{max} \leqslant 742,08) = 95\%$	732,39	728,56
12:00 – 13:00	$P(750,55 \leqslant PRU1_{max} \leqslant 773,51) = 95\%$	762,03	759,43
13:00 – 14:00	$P(764,07 \leqslant PRU1_{max} \leqslant 785,09) = 95\%$	774,58	773,40
14:00 – 15:00	$P(743,09 \leqslant PRU1_{max} \leqslant 761,69) = 95\%$	752,39	749,70
15:00 – 16:00	$P(766,44 \leqslant PRU1_{max} \leqslant 790,64) = 95\%$	778,54	776,38
16:00 – 17:00	$P(750,73 \leqslant PRU1_{max} \leqslant 798,03) = 95\%$	774,38	769,72
17:00 – 18:00	$P(761,31 \leqslant PRU1_{max} \leqslant 796,26) = 95\%$	778,78	773,02
18:00 – 19:00	$P(681,13 \leqslant PRU1_{max} \leqslant 710,21) = 95\%$	695,67	689,61
19:00 – 20:00	$P(516,98 \leqslant PRU1_{max} \leqslant 574,30) = 95\%$	545,64	540,26
20:00 – 21:00	$P(517,08 \leqslant PRU1_{max} \leqslant 574,39) = 95\%$	545,74	531,65

4.1.13 Número de pacotes do *link* institucional, *download*

A estimativa de parâmetro média do número de pacotes do *link* institucional (*upload*) está representada nesse experimento pela variável $\overline{PRD2}$. Da mesma forma, a média dos valores máximos do mesmo parâmetro é representado pela varíavel $\overline{PRD2_{max}}$.

As Tabelas 4.25 e 4.26 apresentam os resultados da estimativa de parâmetros para as variáveis $\overline{PRD2}$ e $\overline{PRD2_{max}}$ respectivamente. A Figura 4.14 ilustra a dispersão de ambas as variáveis supra citadas. Podemos confirmar, a partir da visualização dos gráficos na Figura 4.14 e 4.8, o raciocínio sobre a semelhança entre *throughput* e número de pacotes para *download* conforme apresentado na subsessão 4.1.11.

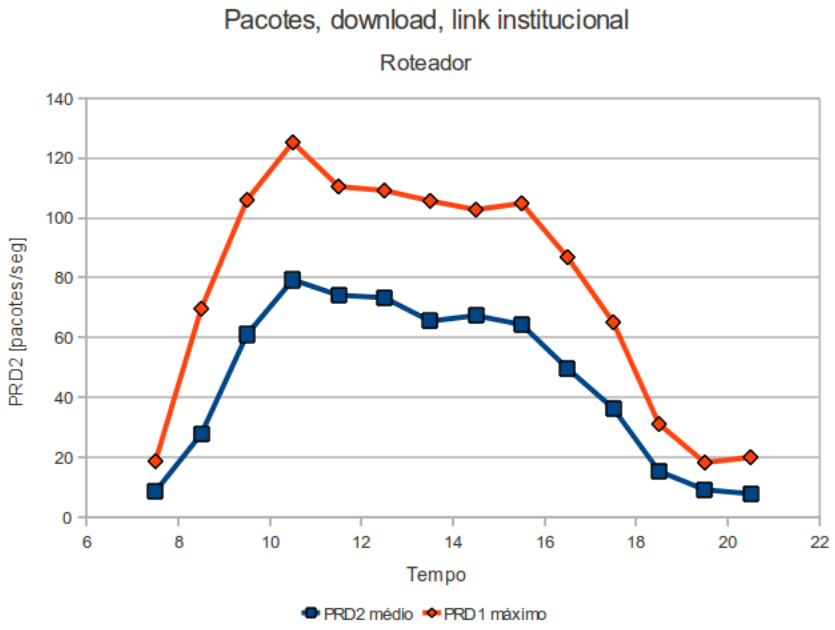

Figura 4.14: Estimativa de parâmetros: pacotes, *download*, *link* institucional

Tabela 4.25: Estimativa de parâmetro: média do número de pacotes, *download*, *link* institucional

Período	Intervalo de confiança	$\overline{PRD2^*}$	$\overline{PRD2}$
07:00 - 08:00	$P(8,47 \leqslant PRD2 \leqslant 8,68) = 95\%$	8,57	8,55
08:00 – 09:00	$P(27,29 \leqslant PRD2 \leqslant 28,15) = 95\%$	27,72	27,61
09:00 – 10:00	$P(60,20 \leqslant PRD2 \leqslant 61,84) = 95\%$	61,02	60,83
10:00 – 11:00	$P(78,36 \leqslant PRD2 \leqslant 80,22) = 95\%$	79,29	79,09
11:00 – 12:00	$P(73,33 \leqslant PRD2 \leqslant 74,97) = 95\%$	74,15	73,72
12:00 – 13:00	$P(72,32 \leqslant PRD2 \leqslant 74,28) = 95\%$	73,30	73,24
13:00 – 14:00	$P(64,62 \leqslant PRD2 \leqslant 66,36) = 95\%$	65,49	65,49
14:00 – 15:00	$P(66,56 \leqslant PRD2 \leqslant 68,09) = 95\%$	67,32	67,34
15:00 – 16:00	$P(63,63 \leqslant PRD2 \leqslant 65,03) = 95\%$	64,33	64,34
16:00 – 17:00	$P(49,16 \leqslant PRD2 \leqslant 50,14) = 95\%$	49,65	49,43
17:00 – 18:00	$P(35,80 \leqslant PRD2 \leqslant 36,83) = 95\%$	36,31	36,35
18:00 – 19:00	$P(15,16 \leqslant PRD2 \leqslant 15,46) = 95\%$	15,31	15,33
19:00 – 20:00	$P(9,03 \leqslant PRD2 \leqslant 9,19) = 95\%$	9,11	9,11
20:00 – 21:00	$P(7,62 \leqslant PRD2 \leqslant 7,94) = 95\%$	7,78	7,75

Tabela 4.26: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *download*, *link* institucional

Período	Intervalo de confiança	$\overline{PRD2^*_{max}}$	$\overline{PRD2_{max}}$
07:00 - 08:00	$P(16,68 \leqslant PRD2_{max} \leqslant 20,65) = 95\%$	18,66	18,21
08:00 – 09:00	$P(59,54 \leqslant PRD2_{max} \leqslant 79,73) = 95\%$	69,63	67,23
09:00 – 10:00	$P(89,02 \leqslant PRD2_{max} \leqslant 122,89) = 95\%$	105,95	99,72
10:00 – 11:00	$P(108,73 \leqslant PRD2_{max} \leqslant 141,71) = 95\%$	125,22	119,31
11:00 – 12:00	$P(93,43 \leqslant PRD2_{max} \leqslant 127,50) = 95\%$	110,46	107,20
12:00 – 13:00	$P(93,77 \leqslant PRD2_{max} \leqslant 124,50) = 95\%$	109,14	103,69
13:00 – 14:00	$P(89,44 \leqslant PRD2_{max} \leqslant 121,92) = 95\%$	105,68	104,01
14:00 – 15:00	$P(89,25 \leqslant PRD2_{max} \leqslant 116,23) = 95\%$	102,74	101,56
15:00 – 16:00	$P(90,31 \leqslant PRD2_{max} \leqslant 119,40) = 95\%$	104,86	99,80
16:00 – 17:00	$P(77,36 \leqslant PRD2_{max} \leqslant 96,37) = 95\%$	86,86	83,96
17:00 – 18:00	$P(55,74 \leqslant PRD2_{max} \leqslant 74,48) = 95\%$	65,11	62,56
18:00 – 19:00	$P(25,84 \leqslant PRD2_{max} \leqslant 36,34) = 95\%$	31,09	29,33
19:00 – 20:00	$P(16,61 \leqslant PRD2_{max} \leqslant 19,79) = 95\%$	18,20	17,48
20:00 – 21:00	$P(14,82 \leqslant PRD2_{max} \leqslant 25,15) = 95\%$	19,99	18,40

4.1.14 Número de pacotes do *link* institucional, *upload*

Nesse experimento serão construídos intervalos de confiança para a média do número de pacotes do *link* institucional (*upload*), representado pela variável $\overline{PRU2}$, e média dos valores máximos do mesmo parâmetro, associado à variável $\overline{PRU2_{max}}$. A Figura 4.15 ilustra o gráfico de dispersão entre as variáveis $\overline{PRU2}$ e $\overline{PRU2_{max}}$. Nesse gráfico podemos inferir sobre o comportamento contínuo e elevado de ambas as variáveis, em relação aos intervalos de tempo nas extremidades do período medido. Em comparação ao experimento sobre o *throughput* do mesmo *link*, $\overline{PRU2_{max}}$ não apresenta o mesmo pico observado na variável $\overline{TRU2_{max}}$ no intervalo final, conforme representado pela Figura .

As Tabelas 4.27 e 4.28 apresentam respectivamente os intervalos de confiança de $\overline{PRU2}$ e $\overline{PRU2_{max}}$.

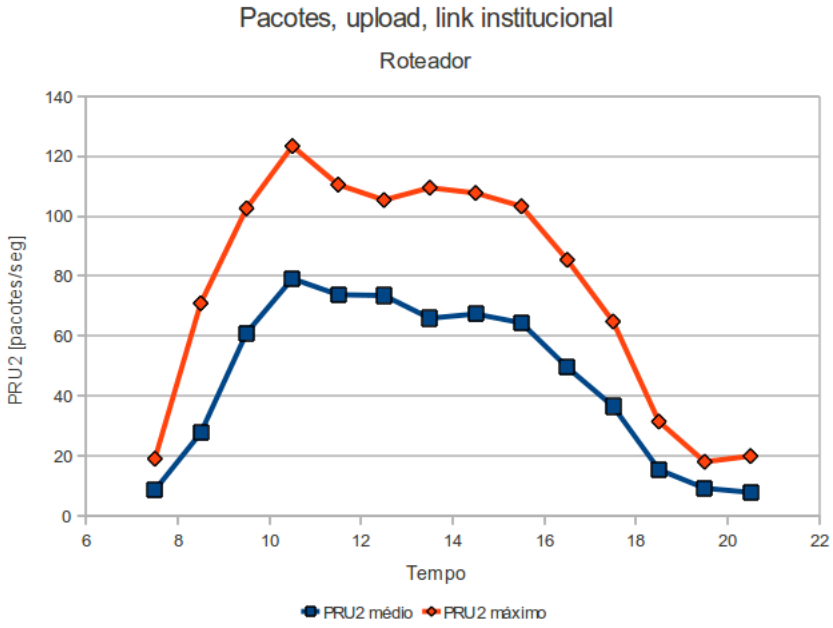

Figura 4.15: Estimativa de parâmetros: pacotes, *upload*, *link* institucional

Tabela 4.27: Estimativa de parâmetro: média do número de pacotes, *upload*, *link* institucional

Período	Intervalo de confiança	$\overline{PRU2^*}$	$\overline{PRU2}$
07:00 - 08:00	$P(8,43 \leqslant PRU2 \leqslant 8,65) = 95\%$	8,54	8,55
08:00 – 09:00	$P(27,16 \leqslant PRU2 \leqslant 28,01) = 95\%$	27,59	27,61
09:00 – 10:00	$P(60,14 \leqslant PRU2 \leqslant 61,81) = 95\%$	60,97	60,83
10:00 – 11:00	$P(78,18 \leqslant PRU2 \leqslant 80,05) = 95\%$	79,11	79,09
11:00 – 12:00	$P(72,90 \leqslant PRU2 \leqslant 74,55) = 95\%$	73,73	73,72
12:00 – 13:00	$P(72,54 \leqslant PRU2 \leqslant 74,44) = 95\%$	73,49	73,24
13:00 – 14:00	$P(64,98 \leqslant PRU2 \leqslant 66,77) = 95\%$	65,88	65,49
14:00 – 15:00	$P(66,54 \leqslant PRU2 \leqslant 68,04) = 95\%$	67,29	67,34
15:00 – 16:00	$P(63,61 \leqslant PRU2 \leqslant 65,07) = 95\%$	64,34	64,34
16:00 – 17:00	$P(48,98 \leqslant PRU2 \leqslant 50,02) = 95\%$	49,50	49,43
17:00 – 18:00	$P(35,97 \leqslant PRU2 \leqslant 37,01) = 95\%$	36,49	36,35
18:00 – 19:00	$P(15,20 \leqslant PRU2 \leqslant 15,51) = 95\%$	15,36	15,33
19:00 – 20:00	$P(9,04 \leqslant PRU2 \leqslant 9,20) = 95\%$	9,12	9,11
20:00 – 21:00	$P(7,60 \leqslant PRU2 \leqslant 7,90) = 95\%$	7,75	7,75

Tabela 4.28: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *upload*, *link* institucional

Período	Intervalo de confiança	$\overline{PRU2^*_{max}}$	$\overline{PRU2_{max}}$
07:00 - 08:00	$P(16,85 \leqslant PRU2_{max} \leqslant 21,07) = 95\%$	18,96	18,21
08:00 – 09:00	$P(61,21 \leqslant PRU2_{max} \leqslant 80,59) = 95\%$	70,90	67,23
09:00 – 10:00	$P(86,77 \leqslant PRU2_{max} \leqslant 118,37) = 95\%$	102,57	99,72
10:00 – 11:00	$P(103,30 \leqslant PRU2_{max} \leqslant 143,38) = 95\%$	123,34	119,31
11:00 – 12:00	$P(94,70 \leqslant PRU2_{max} \leqslant 126,24) = 95\%$	110,47	107,20
12:00 – 13:00	$P(89,53 \leqslant PRU2_{max} \leqslant 121,17) = 95\%$	105,35	103,69
13:00 – 14:00	$P(91,85 \leqslant PRU2_{max} \leqslant 127,06) = 95\%$	109,46	104,01
14:00 – 15:00	$P(87,66 \leqslant PRU2_{max} \leqslant 127,69) = 95\%$	107,68	101,56
15:00 – 16:00	$P(89,38 \leqslant PRU2_{max} \leqslant 117,15) = 95\%$	103,26	99,80
16:00 – 17:00	$P(75,11 \leqslant PRU2_{max} \leqslant 95,65) = 95\%$	85,38	83,96
17:00 – 18:00	$P(54,73 \leqslant PRU2_{max} \leqslant 74,86) = 95\%$	64,79	62,56
18:00 – 19:00	$P(26,70 \leqslant PRU2_{max} \leqslant 36,09) = 95\%$	31,40	29,33
19:00 – 20:00	$P(16,05 \leqslant PRU2_{max} \leqslant 19,88) = 95\%$	17,96	17,48
20:00 – 21:00	$P(15,13 \leqslant PRU2_{max} \leqslant 24,56) = 95\%$	19,84	18,40

4.1.15 Número de pacotes do *link ethernet*, *download*

Essa subsessão contempla o experimento de estimação de parâmetro para o número de pacotes do *link ethernet* (*download*). A média do parâmetro é representada pela variável $\overline{PSD}$ e a média dos valores máximos do mesmo parâmetro pela variável $\overline{PSD_{max}}$. A Figura 4.16 ilustra a dispersão dos valores de ambas as variáveis, sendo válido ressaltar a semelhança do comportamento do *throughput* com o número de pacotes por segundo com o gráfico da Figura 4.10.

As Tabelas 4.29 e 4.30 apresentam os intervalos de confiança construídos para as variáveis $\overline{PSD}$ e $\overline{PSD_{max}}$ respectivamente.

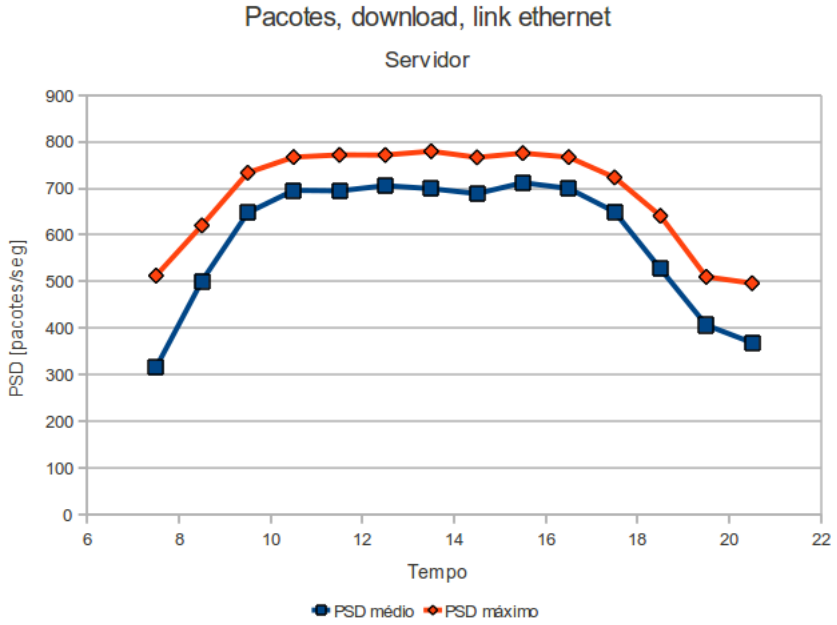

Figura 4.16: Estimativa de parâmetros: pacotes, *download*, *link ethernet*

Tabela 4.29: Estimativa de parâmetro: média do número de pacotes, *download*, *link ethernet*

Período	Intervalo de confiança	$\overline{PSD^*}$	$\overline{PSD}$
07:00 - 08:00	$P(312,53 \leqslant PSD \leqslant 318,03) = 95\%$	315,28	315,62
08:00 – 09:00	$P(498,48 \leqslant PSD \leqslant 503,31) = 95\%$	500,89	500,75
09:00 – 10:00	$P(645,95 \leqslant PSD \leqslant 649,11) = 95\%$	647,53	647,38
10:00 – 11:00	$P(694,04 \leqslant PSD \leqslant 697,08) = 95\%$	695,56	695,66
11:00 – 12:00	$P(693,00 \leqslant PSD \leqslant 696,00) = 95\%$	694,50	694,42
12:00 – 13:00	$P(704,23 \leqslant PSD \leqslant 707,31) = 95\%$	705,77	705,71
13:00 – 14:00	$P(697,82 \leqslant PSD \leqslant 700,70) = 95\%$	699,26	699,17
14:00 – 15:00	$P(686,88 \leqslant PSD \leqslant 690,27) = 95\%$	688,58	688,47
15:00 – 16:00	$P(711,11 \leqslant PSD \leqslant 712,98) = 95\%$	712,04	711,96
16:00 – 17:00	$P(698,87 \leqslant PSD \leqslant 700,63) = 95\%$	699,75	699,65
17:00 – 18:00	$P(648,33 \leqslant PSD \leqslant 650,87) = 95\%$	649,60	649,78
18:00 – 19:00	$P(525,38 \leqslant PSD \leqslant 529,16) = 95\%$	527,27	526,73
19:00 – 20:00	$P(404,33 \leqslant PSD \leqslant 408,44) = 95\%$	406,39	406,06
20:00 – 21:00	$P(366,45 \leqslant PSD \leqslant 370,49) = 95\%$	368,47	368,28

Tabela 4.30: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *download*, *link ethernet*

Período	Intervalo de confiança	$\overline{PSD^*_{max}}$	$\overline{PSD_{max}}$
07:00 - 08:00	$P(487,90 \leqslant PSD_{max} \leqslant 537,10) = 95\%$	512,50	500,92
08:00 – 09:00	$P(603,62 \leqslant PSD_{max} \leqslant 637,04) = 95\%$	620,33	619,38
09:00 – 10:00	$P(708,88 \leqslant PSD_{max} \leqslant 758,01) = 95\%$	733,44	728,34
10:00 – 11:00	$P(743,01 \leqslant PSD_{max} \leqslant 791,16) = 95\%$	767,08	759,72
11:00 – 12:00	$P(751,90 \leqslant PSD_{max} \leqslant 791,74) = 95\%$	771,82	765,17
12:00 – 13:00	$P(756,00 \leqslant PSD_{max} \leqslant 786,93) = 95\%$	771,46	765,11
13:00 – 14:00	$P(761,97 \leqslant PSD_{max} \leqslant 797,38) = 95\%$	779,68	775,78
14:00 – 15:00	$P(754,60 \leqslant PSD_{max} \leqslant 778,25) = 95\%$	766,42	761,97
15:00 – 16:00	$P(765,35 \leqslant PSD_{max} \leqslant 785,72) = 95\%$	775,53	771,41
16:00 – 17:00	$P(753,68 \leqslant PSD_{max} \leqslant 780,19) = 95\%$	766,94	761,50
17:00 – 18:00	$P(710,03 \leqslant PSD_{max} \leqslant 736,66) = 95\%$	723,34	721,50
18:00 – 19:00	$P(629,92 \leqslant PSD_{max} \leqslant 651,92) = 95\%$	640,92	637,99
19:00 – 20:00	$P(486,97 \leqslant PSD_{max} \leqslant 532,84) = 95\%$	509,90	503,67
20:00 – 21:00	$P(476,01 \leqslant PSD_{max} \leqslant 516,63) = 95\%$	496,32	485,46

4.1.16 Número de pacotes do *link ethernet*, *upload*

Esse último experimento do grupo de estimativa de parâmetros trata o número de pacotes do *link ethernet* (*upload*). A média e a média dos valores máximos do parâmetro em análise são representadas, respectivamente, pelas variáveis $\overline{PSU}$ e $\overline{PSU_{max}}$.

A partir da Figura 4.17, que ilustra a disposição dos valores de $\overline{PSU}$ e $\overline{PSU_{max}}$, notamos a semelhança entre o *throughput* do mesmo *link* com o número de pacotes através do gráfico na Figura 4.13. Conforme apresentado anteriormente, o *link* de *internet* tem forte semelhança com o *link ethernet*, o que justifica a mesma disposição desses parâmetros.

As Tabelas 4.31 e 4.32 apresentam os intervalos de confiança construídos para as variáveis $\overline{PSU}$ e $\overline{PSU_{max}}$ respectivamente.

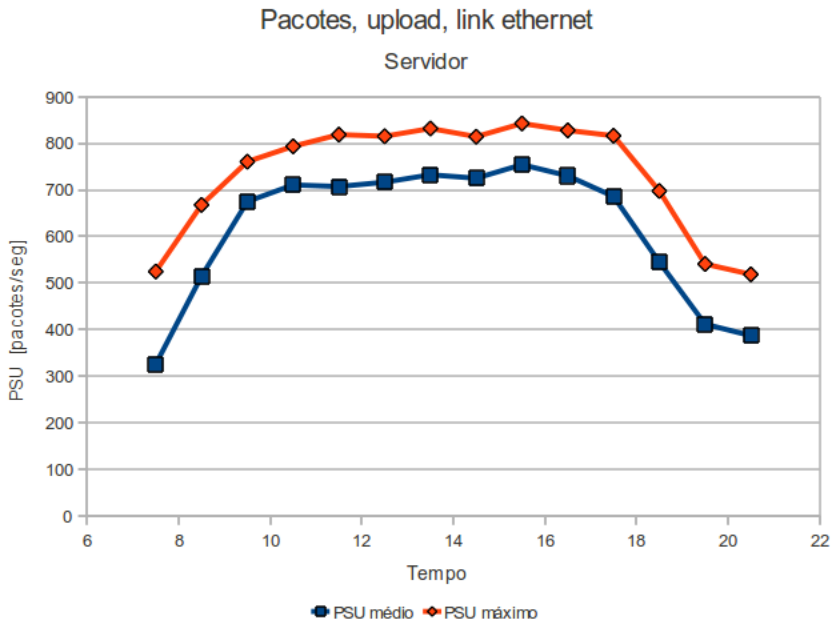

Figura 4.17: Estimativa de parâmetros: pacotes, *upload*, *link ethernet*

Tabela 4.31: Estimativa de parâmetro: média do número de pacotes, *upload*, *link ethernet*

Período	Intervalo de confiança	$\overline{PSU^*}$	$\overline{PSU}$
07:00 - 08:00	$P(322,83 \leqslant PSU \leqslant 328,47) = 95\%$	325,65	325,45
08:00 – 09:00	$P(511,99 \leqslant PSU \leqslant 517,25) = 95\%$	514,62	513,77
09:00 – 10:00	$P(673,22 \leqslant PSU \leqslant 676,61) = 95\%$	674,92	674,64
10:00 – 11:00	$P(709,01 \leqslant PSU \leqslant 712,55) = 95\%$	710,78	710,67
11:00 – 12:00	$P(705,53 \leqslant PSU \leqslant 709,14) = 95\%$	707,34	707,15
12:00 – 13:00	$P(714,69 \leqslant PSU \leqslant 718,61) = 95\%$	716,65	716,62
13:00 – 14:00	$P(730,98 \leqslant PSU \leqslant 734,14) = 95\%$	732,56	732,50
14:00 – 15:00	$P(723,41 \leqslant PSU \leqslant 726,78) = 95\%$	725,09	724,75
15:00 – 16:00	$P(753,13 \leqslant PSU \leqslant 755,50) = 95\%$	754,31	754,14
16:00 – 17:00	$P(729,18 \leqslant PSU \leqslant 731,63) = 95\%$	730,41	730,28
17:00 – 18:00	$P(683,95 \leqslant PSU \leqslant 688,00) = 95\%$	685,97	685,24
18:00 – 19:00	$P(542,75 \leqslant PSU \leqslant 547,05) = 95\%$	544,90	544,88
19:00 – 20:00	$P(409,38 \leqslant PSU \leqslant 413,75) = 95\%$	411,56	411,19
20:00 – 21:00	$P(384,67 \leqslant PSU \leqslant 389,15) = 95\%$	386,91	386,51

Tabela 4.32: Estimativa de parâmetro: média dos valores máximos do número de pacotes, *upload*, *link ethernet*

Período	Intervalo de confiança	$\overline{PSU^*_{max}}$	$\overline{PSU_{max}}$
07:00 - 08:00	$P(491,88 \leqslant PSU_{max} \leqslant 557,61) = 95\%$	524,75	518,50
08:00 – 09:00	$P(641,43 \leqslant PSU_{max} \leqslant 694,31) = 95\%$	667,87	660,87
09:00 – 10:00	$P(737,54 \leqslant PSU_{max} \leqslant 783,79) = 95\%$	760,66	754,71
10:00 – 11:00	$P(773,39 \leqslant PSU_{max} \leqslant 814,04) = 95\%$	793,71	786,11
11:00 – 12:00	$P(796,38 \leqslant PSU_{max} \leqslant 841,16) = 95\%$	818,77	809,60
12:00 – 13:00	$P(793,11 \leqslant PSU_{max} \leqslant 837,50) = 95\%$	815,31	808,66
13:00 – 14:00	$P(816,68 \leqslant PSU_{max} \leqslant 847,36) = 95\%$	832,02	827,26
14:00 – 15:00	$P(800,63 \leqslant PSU_{max} \leqslant 828,25) = 95\%$	814,44	811,53
15:00 – 16:00	$P(825,72 \leqslant PSU_{max} \leqslant 859,63) = 95\%$	842,68	835,88
16:00 – 17:00	$P(805,90 \leqslant PSU_{max} \leqslant 850,03) = 95\%$	827,96	819,43
17:00 – 18:00	$P(793,84 \leqslant PSU_{max} \leqslant 838,80) = 95\%$	816,32	809,18
18:00 – 19:00	$P(680,02 \leqslant PSU_{max} \leqslant 715,48) = 95\%$	697,75	690,74
19:00 – 20:00	$P(514,87 \leqslant PSU_{max} \leqslant 566,63) = 95\%$	540,75	530,43
20:00 – 21:00	$P(490,77 \leqslant PSU_{max} \leqslant 546,51) = 95\%$	518,64	514,30

4.2 Grupo 2: Análise de variância

Nesse grupo de experimentos serão apresentados os resultados das análises de variâncias (ANOVA) para o *throughput*, número de pacotes, taxa de pacotes com erro e taxa de pacotes descartados. As quatro variáveis analisadas nessa subsessão foram agrupadas conforme classificação em k linhas e n colunas: *link* físico ($k = 3$) e sentido *download* e *upload* ($n = 2$). O número de elementos de cada classificação agrupada é de 10 elementos ($r = 10$). Dessa forma, temos a análise de variância com duas classificações com repetição, onde dentro das terminologias e conceitos estatísticos esse teste é também denominado por alguns autores como *two-way*. Sendo a ANOVA uma comparação entre várias médias populacionais, as análises de variâncias dos quatro experimentos desse grupo foram realizadas ao nível de 5% de significância.

O objetivo desse experimento conforme apresentado na sessão 4.4 é de verificar a existência de diferenças significativas para os parâmetros *throughput*, número de pacotes, taxa de pacotes com erro e taxa de pacotes descartadas, considerando as classificações de *link* físico (*Link1*, *Link2* e *Link3*) e sentido (*download* e *upload*). Dessa forma, a comparação dessas médias pode ser esquematizada na Tabela 4.33 a seguir.

Tabela 4.33: Esquematização do experimento de análise de variância

	Download	*Upload*
Link1	amostra com 10 elementos	amostra com 10 elementos
Link2	amostra com 10 elementos	amostra com 10 elementos
Link3	amostra com 10 elementos	amostra com 10 elementos

4.2.1 Comparação entre médias do *throughput*

A proposta dessa análise de variância, além de verificar a existência de diferenças significativas do *throughput* para cada classificação, é averiguar se o roteador está tratando com igualdade a vazão de dados entre os *links* e entre os sentidos *download* e *upload*. Nesse momento analisaremos se toda a vazão de informação entre os três *links* está sendo feita de forma balanceada, validando estatisticamente a sua operação de forma igualitária.

Conforme apresentado na revisão bibliográfica na sessão 2.4, verificamos inicial se existe evidência de interação entre as classificações através da desigualdade $F_I > F_{crtico}$. Se existe evidência de interação, F_L (linha) e F_C (coluna) serão calculados em função de S_I^2 (quadrado médio da interação), ou seja, $F_L = S_L^2/S_I^2$ e $F_C = S_C^2/S_I^2$. Caso contrário, F_L (linha) e F_C (coluna) serão determinados a partir de S_R^2 (quadrado médio residual), em outras palavras, $F_L = S_L^2/S_R^2$ e $F_C = S_C^2/S_R^2$.

A Tabela 4.34 apresenta os resultados da ANOVA para o *throughput*, conforme modelo apresentado em 2.10.

Tabela 4.34: Resultado da comparação entre médias do *throughput*

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	$F_{\alpha=5\%}$
Entre linhas	1.863.653,8	2	931.826,91		
Entre colunas	3.659.082,6	1	3.659.082,6		
Interação	1.521.701,6	2	760.850,78	13.16	3.17
Entre tratamentos	7.044.437,9	5	1.408.887,6	24.37	2.39
Residual	3.122.024,6	54	57.815,269		
Total	10.166.462	59			

Desses resultados vemos que há evidência de interação entre as classificações, pois:

$$F_I \; e \; F_{2,54,5\%} \quad \Rightarrow \quad 13,16 > 3,17$$

Logo teremos F_L e F_C em função do quadrado médio da interação, ou seja:

$$F_L \; e \; F_{2,2,5\%} \quad \Rightarrow \quad 1,22 < 19,00 \tag{4.1}$$

$$F_C \; e \; F_{1,2,5\%} \quad \Rightarrow \quad 4,81 < 18,51 \tag{4.2}$$

Da expressão 4.1, como F_L é menor que $F_{critico}$ concluímos que não existe diferença significativa entre os 3 *links* que compõem a conexão de *internet* do CEFET-MG, o que na prática significa a operação dos meios físicos de forma balanceada. Da desigualdade exibida em 4.2, como F_L é menor que $F_{critico}$ também inferimos na inexistência de diferença significativa entre os sentidos *download* e *upload*.

4.2.2 Comparação entre médias do número de pacotes

A análise de variância do números de pacotes dos dados classificados conforme classificação por *link* e sentido de conexão tem por objetivo complementar a comparação de médias do *throughput*. Esse complemento foi considerado a partir do fato dessas variáveis serem da mesma natureza, desconhecendo qualquer tipo de relação ou conexão entre elas.

Toda a discussão sobre interação entre as classificações e cálculo do valor de F apresentada na subsessão 4.2.1 permanecem válidas. A Tabela 4.35 apresenta os resultados para a comparação entre médias do número de pacotes dos *links* de internet.

Tabela 4.35: Resultado da comparação entre médias do número de pacotes

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	$F_{\alpha=5\%}$
Entre linhas	12.596,1	2	6.298,1	10,82	
Entre colunas	220.021,7	1	220.021,8	378,09	
Interação	7.683,9	2	3.841,9	6,60	3,17
Entre tratamentos	240.301,8	5	48.060,4	82,58	2,39
Residual	31.423,8	54	581,9		
Total	271.725,7	59			

Desses resultados confirmamos a evidência de interação entre as classificações, pois:

$$F_I \; e \; F_{2,54,5\%} \quad \Rightarrow \quad 6,6 > 3,17$$

Logo teremos F_L e F_C em função do quadrado médio da interação, ou seja:

$$F_L \; e \; F_{2,2,5\%} \quad \Rightarrow \quad 1,63 < 19,00 \tag{4.3}$$

$$F_C \; e \; F_{1,2,5\%} \quad \Rightarrow \quad 57,26 > 18,51 \tag{4.4}$$

Da expressão 4.3, como F_L é menor que $F_{critico}$ concluímos que não existe diferença significativa entre os 3 *links* físicos que compõem a conexão de *internet* do CEFET-MG, considerado o número de pacotes. Por outro lado, a partir da expressão 4.4 vemos que F_L é maior que $F_{critico}$. Portanto existe diferença significativa para o número de pacotes com classificação conforme sentido de conexão.

4.2.3 Comparação entre médias do número de pacotes com erro

Nessa sessão abordaremos os resultados do experimento da análise de variância do número de pacotes com erro. Na prática, esse tipo de comparação permite dizer se algum *link* é responsável ou não pela alta taxa de pacotes com erros, devido a uma inconformidade do meio o qualquer outro motivo. Além disso, a análise da variância dessa variável possibilita averiguar qual sentido de conexão apresenta diferença significativa quanto ao número de pacotes com erro. No contexto da administração de uma estrutura de rede, esse experimento permite gerenciar e validar manutenções no meio físico responsável pela transmissão de dados.

A Tabela 4.36 apresenta os resultados para a comparação entre médias do número, ou taxa de erro, de pacotes dos *links* de *internet*. Vale destacar a permanência do conceito de interação entre as classificações e o respectivo cálculo de F, conforme citado na subsessão 4.2.1.

Tabela 4.36: Resultado da comparação entre médias do número de pacotes com erro

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	$F_{\alpha=5\%}$
Entre linhas	2.295,4	2	1.147	1,16	
Entre colunas	3.914,6	1	3.914,6	3,78	
Interação	2.295,4	2	1.147,7	1,11	3.17
Entre tratamentos	8.505,4	5	1.701,1	1,65	2.39
Residual	55.849,9	54	1.034,3		
Total	64.355,3	59			

Desses resultados vemos que não há evidência de interação entre as classificações, pois:

$$F_I \; e \; F_{2,54,5\%} \quad \Rightarrow \quad 1,11 < 3,17$$

Nesse caso, a fonte de variação "interação"foi descartada simplesmente por não haver evidência de sua existência. A soma dos quadrados e os graus de liberdade foram incluídos na fonte de variação "residual". Assim, os resultados da Tabela 4.36 podem ser simplificados conforme os dados na Tabela 4.37.

Logo teremos F_L e F_C em função do quadrado médio residual, ou seja:

$$F_L \; e \; F_{2,56,5\%} \quad \Rightarrow \quad 1,16 < 3,16 \tag{4.5}$$

Tabela 4.37: Resultado simplificado da comparação entre médias do número de pacotes com erro

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	$F_{\alpha=5\%}$
Entre linhas	2.295,4	2	1147,7	1,16	3,16
Entre colunas	3.914,5	1	3914,6	3,78	4,01
Residual	58.145,3	56	1038,3		
Total	64.355,2	59			

$$F_C \; e \; F_{1,56,5\%} \quad \Rightarrow \quad 3,78 < 4,01 \tag{4.6}$$

De ambas as desigualdades apresentadas nas expressões 4.5 e 4.5 conclui-se que não há diferença significativa entre os 3 *links* físicos que compõem a conexão de *internet* e o sentido de conexão, para o número de pacotes descartados.

4.2.4 Comparação entre médias do número de pacotes descartados

Essa análise de variância, que contempla a comparação das médias do número de pacotes descartados, finaliza o grupo de experimentos envolvendo ANOVA. A consideração desses resultados complementam o estudo feito na subsessão 4.2.3, devido a proximidade conceitual sobre pacotes com erros e pacotes descartados em dispositivos de rede.

A Tabela 4.38 apresenta os resultados para a comparação entre médias do número, ou taxa de descarte, de pacotes dos *links* de *internet*.

Desses resultados vemos que há evidência de interação entre as classificações, pois:

$$F_I \; e \; F_{2,54,5\%} \quad \Rightarrow \quad 3,73 > 3,17$$

Logo teremos F_L e F_C em função do quadrado médio da interação entre as classificações, ou seja:

$$F_L \; e \; F_{2,2,5\%} \quad \Rightarrow \quad 0,97 < 19 \tag{4.7}$$

$$F_C \; e \; F_{1,2,5\%} \quad \Rightarrow \quad 19 > 18,51 \tag{4.8}$$

Tabela 4.38: Resultado da comparação entre médias do número de pacotes descartados

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado médio**	F	$F_{\alpha=5\%}$
Entre linhas	1,5	2	0,75	3,67	
Entre colunas	14,64	1	14,63	71,45	
Interação	1,53	2	0,77	3,73	3,17
Entre tratamentos	17,68	5	3,53	17,25	2,39
Residual	11,06	54	0,2		
Total	28,73	59			

Da desigualdade apresentada na expressão 4.7 concluímos que não existe diferença significativa entre os meios físicos. Por outro lado, a expressão 4.8 revela a existência de diferença para os sentidos de conexão *download* e *upload*.

4.3 Grupo 3: Correlação

Os problemas de correlação serão tratados nessa sessão com o intuito de verificar possíveis relações e associações entre as variáveis estudadas até então. Durante o planejamento dos experimentos procurou-se levar em consideração, de forma subjetiva, as correlações que tivessem maior representatividade dentro do ambiente analisado. Dessa forma, além da carga de processamento e uso de memória do roteador e servidor, foram considerados o *throughput* e o número de pacotes do *link ethernet* para ambos os sentidos de conexão.

O presente trabalho aborda apenas a correlação linear no conjunto de experimentos dessa natureza. No entanto, o fato de não existir correlação linear não significa que os dados não se correlacionam de algum forma. O gráfico de dispersão nos permite visualizar se essa correlação será linear, exponencial, polinomial, ou adere ao comportamento de qualquer outra função matemática. Dessa forma, a visualização descritiva dos pares ordenados de cada dupla de variável é essencial para o entendimento de toda a problemática em estudo.

As correlações estão representadas pelo seu respectivo *coeficiente de correlação linear de Pearson*, apresentadas em tabelas junto com seus testes do coeficiente de correlação para os níveis de 5% e 10%. Quanto ao teste do coeficiente de correlação, construído na expressão 4.9 e calculado pela equação 4.10 (NETO, 2002), vale ressaltar que os valores de t calculado estão em geral altos em comparação com os valores da distribuição t de *Student*. Consequentemente, o teste apontará para a existência de correlação linear em grande parte dos casos devido ao elevado valor de n (tamanho da amostra), mesmo que visualmente se verifique uma fraca correlação linear.

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases} \tag{4.9}$$

$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}} \tag{4.10}$$

Ainda sobre os testes do coeficiente de correlação (expressão 4.9), quando dizemos que este foi rejeitado estamos nos referindo ao descarte da hipótese principal. A hipótese principal equivale dizer que o coeficiente de correlação é nulo. Portanto, rejeitar um teste implica inferir na existência de correlação linear, ao passo que aceitar o teste significa crer na não existência de correlação.

4.3.1 *Throughput* e número de pacotes do *link* de *internet*, *download*

Esse experimento visa relacionar o *throughput* e número de pacotes do *link* de *internet*. A Tabela 4.39 ilustra todas as correlações para cada intervalo de hora, além do período total de medição.

Tabela 4.39: Correlação linear: *throughput* e número de pacotes do *link* de *internet*, *download*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,970	Rejeitado	Rejeitado
08:00 – 09:00	0,970	Rejeitado	Rejeitado
09:00 – 10:00	0,851	Rejeitado	Rejeitado
10:00 – 11:00	0,810	Rejeitado	Rejeitado
11:00 – 12:00	0,776	Rejeitado	Rejeitado
12:00 – 13:00	0,697	Rejeitado	Rejeitado
13:00 – 14:00	0,853	Rejeitado	Rejeitado
14:00 – 15:00	0,895	Rejeitado	Rejeitado
15:00 – 16:00	-0,140	Rejeitado	Rejeitado
16:00 – 17:00	0,178	Rejeitado	Rejeitado
17:00 – 18:00	0,628	Rejeitado	Rejeitado
18:00 – 19:00	0,907	Rejeitado	Rejeitado
19:00 – 20:00	0,962	Rejeitado	Rejeitado
20:00 – 21:00	0,959	Rejeitado	Rejeitado
07:00 – 21:00	0,939	Rejeitado	Rejeitado

Como todos os testes foram rejeitados podemos concluir a existência de correlação linear para cada intervalo de tempo. Baseado no que foi citado na introdução dessa sessão, alguns teste apontam a existência de correlação mesmo com valor de r representando uma fraca correlação linear. É o caso do período das 15:00 às 16:00, onde foi obtida uma correlação linear negativa com $r = -0,14$.

De um modo geral para esse experimento, os resultados do coefiente de *Pearson* indicam a existência de forte correlação linear entre o *throughput* e número de pacotes. A Figura 4.18 ilustra o gráfico de dispersão de cada par ordenado.

Adicionalmente nesse gráfico foram incluídos uma linha de tendência linear e seu respectivo *coeficiente de determinação* (r^2). O coeficiente de determinação nada mais é que o valor do coefiente de correlação de *Pearson* elevado ao quadrado, empregado para validar a aderência de valores à linhas de tendências e

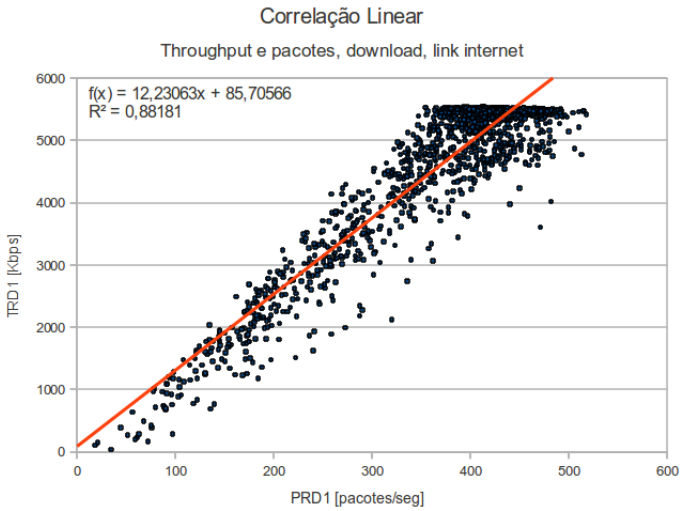

Figura 4.18: Correlação linear: *throughput* e número de pacotes do *link* de *internet*, *download*

regressões. Sendo $r^2 \in [0,1]$, quanto mais próximo de 1 melhor a aderência, ao passo que próximo de 0 pior a regressão encontrada.

Como na Figura 4.18 foi incluída uma reta como tendência dos valores, o seu coefiente angular representa a relação entre o *throughput* e o número de pacotes vazantes. Essa relação significa justamente o tamanho médio dos pacotes em *Kb*, ao qual podemos aliar nesse momento o conhecimento técnico de redes para os valores de MTU. Transformando o coeficiente angular apresentado nesse experimento de *Kbit* para *bytes* temos:

$$\frac{12,23 * 1024}{8} = 1565,44$$

Grande parte dos equipamentos de rede trazem como valor padrão o $MTU = 1500$ *bytes*. O MTU calculado a partir do coeficiente angular da reta de tendência é de $1565,44$, valor maior que 1500 devido ao fato do número de pacotes monitorados consistirem apenas em pacotes *unicast*. Dessa forma, ao considerar uma dada quantidade de pacotes *non-unicast*, naturalmente o número de pacotes aumentará e o provavelmente o coeficiente angular diminuirá a ponto de ser menor que 1500.

Embora o Cacti realize o monitoramento de pacotes *non-unicast*, a MIB que realiza a gerência desse objetivo foi descontinuado, conforme Net SNMP (2009). Na prática, os valores obtidos através da ferramenta de monitoramento não condizem com todo o ambiente de rede analisado.

4.3.2 *Throughput* e número de pacotes do *link* de *internet*, *upload*

Nesse segundo experimento é averiguada a correlação entre o *throughput* e o número de pacotes para o *upload* do *link* de *internet*. As correlações resultantes, conforme Tabela 4.40, não apresentaram grande força quando comparados com o experimento de correlação na subsessão 4.3.1. Em geral, os valores estão contidos entre $0,6$ e $0,7$, e o teste de correlação confirme a existência de correlação linear.

Tabela 4.40: Correlação linear: *throughput* e número de pacotes do *link* de *internet*, *upload*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,425	Rejeitado	Rejeitado
08:00 – 09:00	0,629	Rejeitado	Rejeitado
09:00 – 10:00	0,284	Rejeitado	Rejeitado
10:00 – 11:00	0,578	Rejeitado	Rejeitado
11:00 – 12:00	0,586	Rejeitado	Rejeitado
12:00 – 13:00	0,617	Rejeitado	Rejeitado
13:00 – 14:00	0,608	Rejeitado	Rejeitado
14:00 – 15:00	0,656	Rejeitado	Rejeitado
15:00 – 16:00	0,689	Rejeitado	Rejeitado
16:00 – 17:00	0,798	Rejeitado	Rejeitado
17:00 – 18:00	0,840	Rejeitado	Rejeitado
18:00 – 19:00	0,641	Rejeitado	Rejeitado
19:00 – 20:00	0,674	Rejeitado	Rejeitado
20:00 – 21:00	0,501	Rejeitado	Rejeitado
07:00 – 21:00	0,647	Rejeitado	Rejeitado

A Figura 4.19 ilustra a dispersão dos pontos no plano cartesiano. Em um primeiro momento foi inserida a linha de tendência linear para melhor análise da correlação dos dados, com coeficiente de determinação igual a $0,41$. Como seu comportamento se assemelha à uma função exponencial foi inserida a linha de tendência com essa característica apresentada na Figura 4.20. De maneira elementar, podemos concluir que nesse caso a regressão exponencial melhor se aplica para esse experimento.

Na prática esse comportamento é pertinente, pois o *throughput* será baixo mesmo que o número de pacotes aumente. Esse momento representa a ação de requisições no sentido interno - externo (*upload*). Em continuidade, o *throughput* terá um aumento súbito para um número maior de pacotes, o que representa o

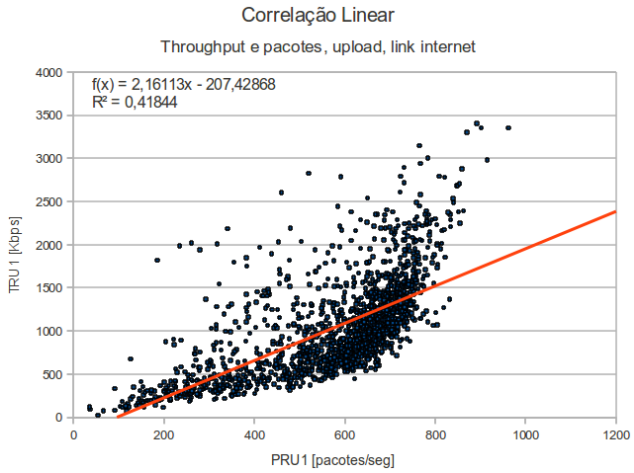

Figura 4.19: Correlação linear: *throughput* e número de pacotes do *link* de *internet*, *upload* (modelo linear)

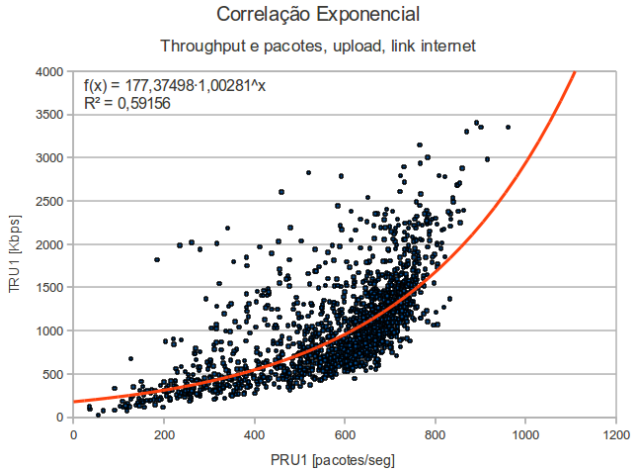

Figura 4.20: Correlação linear: *throughput* e número de pacotes do *link* de *internet*, *upload* (modelo exponencial)

tráfego de informações dos servidores alocados internamente no ambiente para clientes requisitantes fora da instituição. Para essa análise basta considerar que a cópia de um arquivo (*download*) utiliza todo o tamanho disponível no datagrama TCP/IP para armazenamento de dados, ao passo que uma requisição (*upload*) nem sempre utiliza todo o espaço disponível em um pacote para dados.

4.3.3 *Throughput* e número de pacotes do *link* institucional, *download*

Nesse experimento será verificada a correlação entre *throughput* e número de pacotes do *link* institucional, no sentido *download*. Os coeficientes de correlação calculados, conforme exibição dos resultados na Tabela 4.41, apresentam razoável correlação linear mesmo com todos os testes sugerindo a sua existência.

Tabela 4.41: Correlação linear: *throughput* e número de pacotes do *link* institucional, *download*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,425	Rejeitado	Rejeitado
08:00 – 09:00	0,629	Rejeitado	Rejeitado
09:00 – 10:00	0,284	Rejeitado	Rejeitado
10:00 – 11:00	0,578	Rejeitado	Rejeitado
11:00 – 12:00	0,586	Rejeitado	Rejeitado
12:00 – 13:00	0,617	Rejeitado	Rejeitado
13:00 – 14:00	0,608	Rejeitado	Rejeitado
14:00 – 15:00	0,656	Rejeitado	Rejeitado
15:00 – 16:00	0,689	Rejeitado	Rejeitado
16:00 – 17:00	0,798	Rejeitado	Rejeitado
17:00 – 18:00	0,840	Rejeitado	Rejeitado
18:00 – 19:00	0,641	Rejeitado	Rejeitado
19:00 – 20:00	0,674	Rejeitado	Rejeitado
20:00 – 21:00	0,501	Rejeitado	Rejeitado
07:00 – 21:00	0,647	Rejeitado	Rejeitado

Por se tratar do sentido *download*, espera-se que o comportamento seja semelhante ao apresentado no experimento de correlação sobre o *download* do *link* de *internet* na subsessão 4.3.1. A Figura 4.21 ilustra o gráficos dos dados desse experimento.

No entanto, mesmo que os pares ordenados apresentem relativa correlação linear, os pontos estão concentrados numa área com baixa vazão de pacotes e baixo *throughput*. Isso significa que existe uma grande quantidade de pacotes de tamanhos pequenos e uma quantidade razoável de pacotes médios e grandes. Na prática, tratando-se de um *link* de caráter meramente institucional, isso implica em serviços heterogêneos como transferência de arquivos via protocolo *HTTP*, DNS, banco de dados, dentre outros.

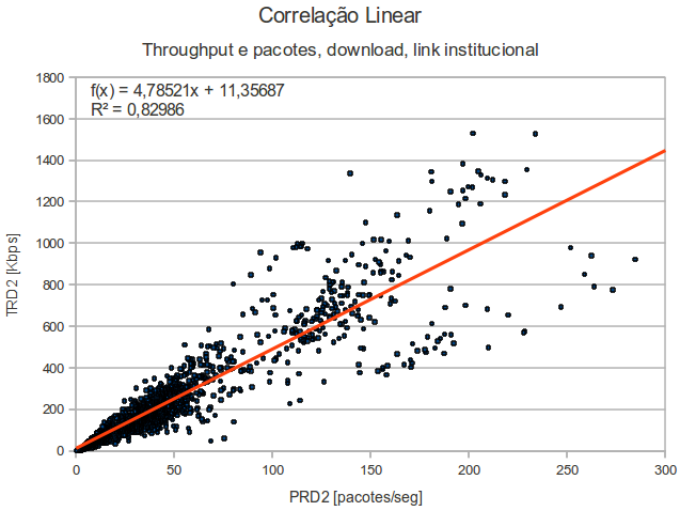

Figura 4.21: Correlação linear: *throughput* e número de pacotes do *link* institucional, *download*

Em uma breve comparação com o experimento apresentado na subsessão 4.3.1, caso o *link* de *internet* (*download*) tivesse parte desse comportamento, poderíamos inferir sobre um elevado número de requisições. Essas requisições podem significar desde a atividade de um servidor na rede interna, sem a ciência dos administradores da rede, até a violação da segurança da rede com ataques do tipo DDoS ou força bruta.

4.3.4 *Throughput* e número de pacotes do *link* institucional, *upload*

Nesse experimento abordamos o *throughput* e número de pacotes do *link* institucional, *upload*. A Tabela 4.42 apresenta os valores dos coeficientes de correlação do experimento. De uma maneira geral foram encontradas desejáveis correlações lineares ao longo dos intervalos de tempo, à exceção de intervalos de 18:00 às 19:00 e 19:00 às 20:00 com r igual a $0,401$ e $0,557$ respectivamente.

Tabela 4.42: Correlação linear: *throughput* e número de pacotes do *link* institucional, *upload*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,702	Rejeitado	Rejeitado
08:00 – 09:00	0,838	Rejeitado	Rejeitado
09:00 – 10:00	0,878	Rejeitado	Rejeitado
10:00 – 11:00	0,864	Rejeitado	Rejeitado
11:00 – 12:00	0,940	Rejeitado	Rejeitado
12:00 – 13:00	0,945	Rejeitado	Rejeitado
13:00 – 14:00	0,874	Rejeitado	Rejeitado
14:00 – 15:00	0,899	Rejeitado	Rejeitado
15:00 – 16:00	0,895	Rejeitado	Rejeitado
16:00 – 17:00	0,927	Rejeitado	Rejeitado
17:00 – 18:00	0,912	Rejeitado	Rejeitado
18:00 – 19:00	0,401	Rejeitado	Rejeitado
19:00 – 20:00	0,557	Rejeitado	Rejeitado
20:00 – 21:00	0,943	Rejeitado	Rejeitado
07:00 – 21:00	0,880	Rejeitado	Rejeitado

Quanto à disposição dos dados no plano cartesiano na Figura 4.22, podemos verificar que existe uma leve semelhança de comportamento dos dados em relação ao experimento anterior. É possível perceber que existe a tendência da grande quantidade de pontos que representam pequena vazão de pacotes manter o *throughtput* baixo. Mais uma vez, isso retrata a predominância de pacotes referentes à solicitações nesse sentido de conexão.

4.3.5 *Throughput* e número de pacotes do *link ethernet*, *download*

A correlação linear entre *throughput* e número de pacotes do *link ethernet* (*download*) nos leva a resultados semelhantes ao mesmo experimento envolvendo apenas

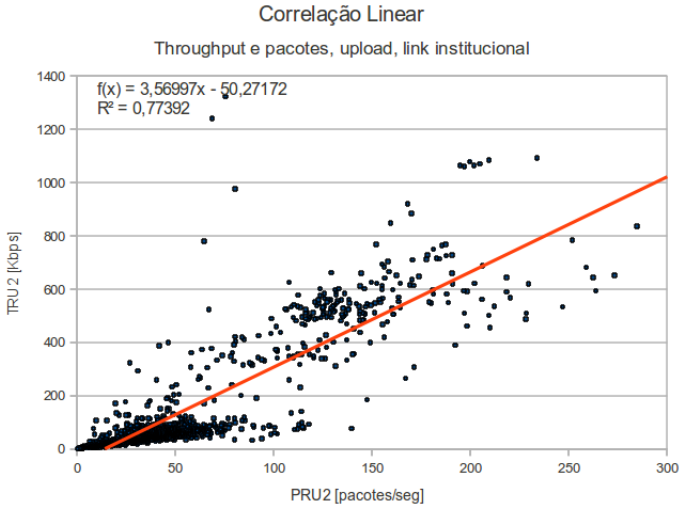

Figura 4.22: Correlação linear: *throughput* e número de pacotes do *link* institucional, *upload* (modelo linear)

o *link* de *internet*, apresentado na subsessão 4.3.1. A Tabela 4.43 ilustra os valores do coeficiente de correlação de cada intervalo de tempo. Ao analisar esses dados podemos verificar que há uma grande correlação linear entre as variáveis em estudo, para grande parte dos intervalos de tempo.

A partir da Figura 4.23 podemos observar que os dados estão aderidos ao longo de toda a reta regredida. Os valores de *throughput* e número de pacotes da interface de rede do servidor contemplam o *link* de *internet* e o *link* institucional. Dessa forma é possível encontrar pares ordenados com baixos valores, o que em sua maioria representam o *link* institucional. Na mesma linha de raciocínio verifica-se a predominância dos pares ordenados com altos valores para ambas as variáveis, conforme Figura 4.23. Isso é justificado na prática pela maior demanda de *download* do *link* de *internet*, como já é esperado a partir da análise de correlação desse *link* feita anteriormente.

Tabela 4.43: Correlação linear: *throughput* e número de pacotes do *link ethernet*, *download*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,983	Rejeitado	Rejeitado
08:00 – 09:00	0,980	Rejeitado	Rejeitado
09:00 – 10:00	0,872	Rejeitado	Rejeitado
10:00 – 11:00	0,805	Rejeitado	Rejeitado
11:00 – 12:00	0,834	Rejeitado	Rejeitado
12:00 – 13:00	0,850	Rejeitado	Rejeitado
13:00 – 14:00	0,888	Rejeitado	Rejeitado
14:00 – 15:00	0,920	Rejeitado	Rejeitado
15:00 – 16:00	0,715	Rejeitado	Rejeitado
16:00 – 17:00	0,570	Rejeitado	Rejeitado
17:00 – 18:00	0,641	Rejeitado	Rejeitado
18:00 – 19:00	0,922	Rejeitado	Rejeitado
19:00 – 20:00	0,978	Rejeitado	Rejeitado
20:00 – 21:00	0,973	Rejeitado	Rejeitado
07:00 – 21:00	0,948	Rejeitado	Rejeitado

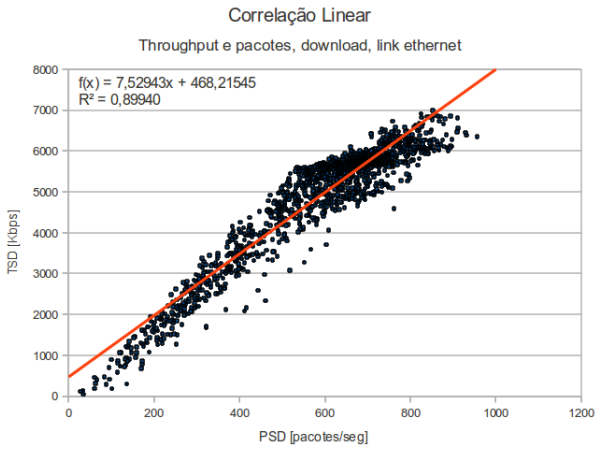

Figura 4.23: Correlação linear: *throughput* e número de pacotes do *link ethernet*, *download*

4.3.6 *Throughput* e número de pacotes do *link ethernet*, *upload*

Esse experimento conclui a análise de correlação entre o *throughput* e o número de pacotes proposta para as interfaces de rede do ambiente analisado. Nesse momento é verificada a correlação linear dessas variáveis para o *link ethernet* (*upload*). A Tabela 4.44 traz os resultados para os coeficientes de correlação, sugerindo razoável correlação linear entre o *throughput* e o número de pacotes.

Tabela 4.44: Correlação linear: *throughput* e número de pacotes do *link ethernet*, *upload*

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,472	Rejeitado	Rejeitado
08:00 – 09:00	0,711	Rejeitado	Rejeitado
09:00 – 10:00	0,399	Rejeitado	Rejeitado
10:00 – 11:00	0,751	Rejeitado	Rejeitado
11:00 – 12:00	0,707	Rejeitado	Rejeitado
12:00 – 13:00	0,738	Rejeitado	Rejeitado
13:00 – 14:00	0,708	Rejeitado	Rejeitado
14:00 – 15:00	0,707	Rejeitado	Rejeitado
15:00 – 16:00	0,699	Rejeitado	Rejeitado
16:00 – 17:00	0,792	Rejeitado	Rejeitado
17:00 – 18:00	0,876	Rejeitado	Rejeitado
18:00 – 19:00	0,646	Rejeitado	Rejeitado
19:00 – 20:00	0,669	Rejeitado	Rejeitado
20:00 – 21:00	0,543	Rejeitado	Rejeitado
07:00 – 21:00	0,727	Rejeitado	Rejeitado

Nas subsessões anteriores vimos a semelhança entre os experimentos de correlação envolvendo *download* do *link* de *internet* e *link ethernet*, ambos apresentados respectivamente nas subsessões 4.3.1 e 4.3.5. Nesse experimento espera-se que o comportamento do *throughput* e número de pacotes do *upload* do *link ethernet* também seja semelhante em relação à correlação do *upload* do *link* de *internet*, conforme estudado na subsessão 4.3.2.

A Figura 4.24 ilustra o gráfico de dispersão dos pares ordenados do *throughput* e número de pacotes do *upload* do *link ethernet*. Nesse gráfico foi inserida uma função de tendência linear com $r^2 = 0,52$. Percebe-se a disposição dos pontos conforme função exponencial, ao qual a Figura 4.25 traz uma regressão exponencial

com $r^2 = 0,65$. Dessa forma, para esse experimento, a correlação linear não é o melhor método para se estabelecer correlação entre as variáveis em estudo.

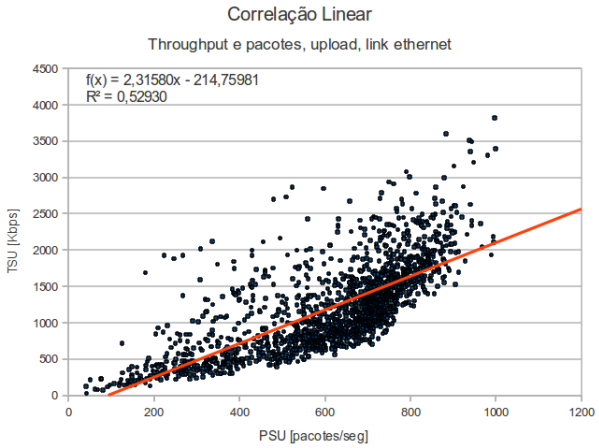

Figura 4.24: Correlação linear: *throughput* e número de pacotes do *link ethernet*, *upload* (modelo linear)

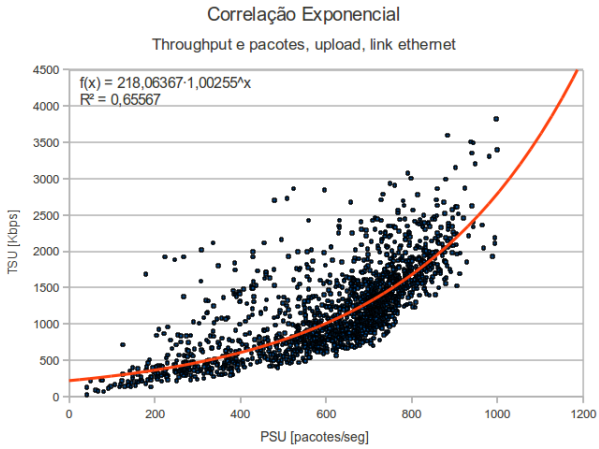

Figura 4.25: Correlação linear: *throughput* e número de pacotes do *link ethernet*, *upload* (modelo exponencial)

Na prática, o conjunto de valores altos de *throughput* e número de pacotes referem-se ao atendimento de requisições alocados internamente à instituição. Da mesma forma, o conjunto de valores baixos para essas variáveis sugerem não só às requisições de navegação à *internet*, como também a operação de serviços do CEFET-MG a partir do *link* institucional.

4.3.7 Carga de processamento e uso de memória do roteador

Nesses últimos quatro experimentos serão analisados a carga de processamento e o uso de memória combinados entre o roteador e servidor de *firewall* do CEFET-MG. Esse experimento em específico trata a correlação linear entre os percentuais de uso da carga de processamento e memória do roteador, afim de se estabelecer relação entre essas variáveis.

No entanto, a partir da Tabela 4.45, podemos observar que a correlação linear nesse experimento não é boa. Os melhores coeficientes de correlação calculados estão próximos de $0,49$. Além dos resultados de r próximos de zero, alguns testes sobre a correlação foram aceitos mesmo com o elevado valor de n ($n = 120$), o que sugere a inexistência de correlação linear.

Tabela 4.45: Correlação linear: carga de processamento e uso de memória do roteador

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,149	Rejeitado	Aceito
08:00 – 09:00	0,496	Rejeitado	Rejeitado
09:00 – 10:00	0,490	Rejeitado	Rejeitado
10:00 – 11:00	0,483	Rejeitado	Rejeitado
11:00 – 12:00	0,446	Rejeitado	Rejeitado
12:00 – 13:00	0,419	Rejeitado	Rejeitado
13:00 – 14:00	0,304	Rejeitado	Rejeitado
14:00 – 15:00	0,398	Rejeitado	Rejeitado
15:00 – 16:00	-0,057	Rejeitado	Rejeitado
16:00 – 17:00	0,022	Aceito	Aceito
17:00 – 18:00	-0,226	Rejeitado	Rejeitado
18:00 – 19:00	-0,029	Rejeitado	Rejeitado
19:00 – 20:00	-0,087	Rejeitado	Rejeitado
20:00 – 21:00	-0,212	Rejeitado	Rejeitado
07:00 – 21:00	0,289	Rejeitado	Rejeitado

Podemos concluir, para este experimento, que não é possível estabelecer uma correlação linear aceitável para a carga de processamento e uso de memória para o roteador. Para visualização dessa conclusão, a Figura 4.26 apresenta os pares ordenados das variáveis em estudo dispostos em um gráfico de dispersão. Vale ressaltar, a partir da visualização do gráfico, que dificilmente será estabelecida qualquer outra correlação matemática para esses valores.

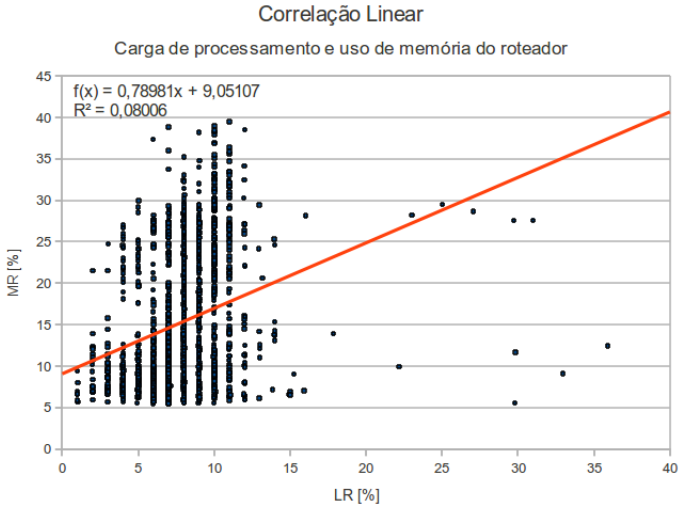

Figura 4.26: Correlação linear: carga de processamento e uso de memória do roteador

4.3.8 Carga de processamento e uso de memória do servidor

Esse experimento realiza, da mesma forma que o experimento apresentado em 4.3.7, a análise de correlação linear da carga de processamento e uso de memória, agora para o servidor de *firewall* da instituição. A Tabela 4.46 traz os coeficientes de correlação calculados para cada intervalo de tempo, juntamente com seus teste de correlação. Em comparação à correlação apresentada na subsessão 4.3.7, com estudo das mesmas variáveis porém aplicadas ao roteador, o experimento atual trouxe resultados piores. Os coeficientes de correlação apresentam valores mais próximos de zero, da mesma forma que há maior número de testes de correlação que foram aceitos.

No contexto da correlação linear, esses resultados ruins podem ser visualizados na Figura 4.27. Cabe ressaltar novamente que dificilmente uma função matemática será capaz de correlacionar os valores da carga de processamento e uso de memória do servidor, com seu respectivo coeficiente de correlação aceitável. Como conclusão, da mesma forma que essas variáveis não se correlacionam no âmbito do funcionamento do roteador, a correlação linear no servidor de *firewall* envolvendo percentuais de uso de processador e memória também não apresentam qualquer relação entre si.

Tabela 4.46: Correlação linear: carga de processamento e uso de memória do servidor

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	0,372	Rejeitado	Rejeitado
08:00 – 09:00	0,472	Rejeitado	Rejeitado
09:00 – 10:00	0,457	Rejeitado	Rejeitado
10:00 – 11:00	0,425	Rejeitado	Rejeitado
11:00 – 12:00	0,234	Rejeitado	Rejeitado
12:00 – 13:00	0,099	Aceito	Aceito
13:00 – 14:00	0,472	Rejeitado	Rejeitado
14:00 – 15:00	0,069	Aceito	Aceito
15:00 – 16:00	0,084	Aceito	Aceito
16:00 – 17:00	0,178	Rejeitado	Rejeitado
17:00 – 18:00	-0,070	Rejeitado	Rejeitado
18:00 – 19:00	-0,054	Rejeitado	Rejeitado
19:00 – 20:00	-0,060	Rejeitado	Rejeitado
20:00 – 21:00	0,039	Aceito	Aceito
07:00 – 21:00	0,376	Rejeitado	Rejeitado

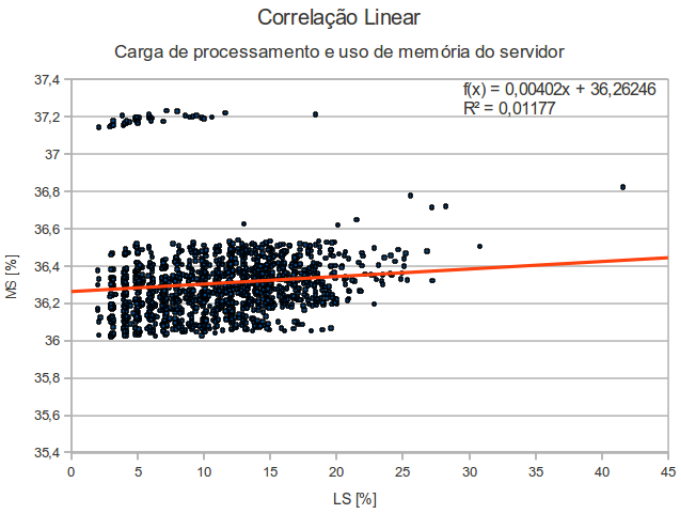

Figura 4.27: Correlação linear: carga de processamento e uso de memória do servidor

4.3.9 Carga de processamento entre roteador e servidor

Nessa subsessão será analisada a correlação linear da carga de processamento entre o roteador e o servidor de *firewall*. No experimento de correlação do *throughput* e número de pacotes há, conforme já apresentado, comportamento semelhante entre as interfaces de rede que representam o *link* de *internet* e *link ethernet*. Dessa forma é pertinente acreditar que baseada nessa semelhança do *throughput* e número de pacotes para ambos os *links* dos dois equipamentos, a carga de processamento também seja semelhante entre o roteador e servidor de *firewall*.

Contudo, a Tabela 4.47 não representa essa suposição, com valores do coeficiente de correlação linear totalmente variantes para cada intervalo de tempo e alguns testes de correlação aceitos. Isso implica, mais uma vez, na observância de uma baixíssima correlação linear da carga de processamento entre o roteador e servidor de *firewall* da instituição.

Tabela 4.47: Correlação linear: carga de processamento entre roteador e servidor

Período	**Correlação linear**	**Teste da correlação com $\alpha = 10\%$**	**Teste da correlação com $\alpha = 5\%$**
07:00 – 08:00	0,390	Rejeitado	Rejeitado
08:00 – 09:00	0,699	Rejeitado	Rejeitado
09:00 – 10:00	0,331	Rejeitado	Rejeitado
10:00 – 11:00	0,484	Rejeitado	Rejeitado
11:00 – 12:00	0,326	Rejeitado	Rejeitado
12:00 – 13:00	0,107	Aceito	Aceito
13:00 – 14:00	0,296	Rejeitado	Rejeitado
14:00 – 15:00	0,025	Aceito	Aceito
15:00 – 16:00	0,139	Rejeitado	Aceito
16:00 – 17:00	-0,049	Rejeitado	Rejeitado
17:00 – 18:00	0,280	Rejeitado	Rejeitado
18:00 – 19:00	0,507	Rejeitado	Rejeitado
19:00 – 20:00	0,122	Rejeitado	Aceito
20:00 – 21:00	0,344	Rejeitado	Rejeitado
07:00 – 21:00	0,505	Rejeitado	Rejeitado

A Figura 4.28 corrobora a conclusão sobre a baixíssima correlação linear entre as variáveis estudadas esse experimentos. Mais uma vez cabe ressaltar a dificuldade de se encontrar alguma função matemática que correlacione os valores representativos da carga de processamento do roteador e do servidor.

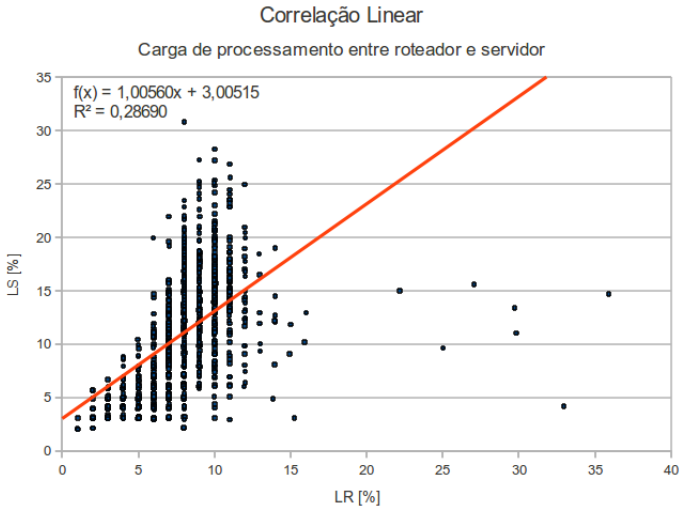

Figura 4.28: Correlação linear: carga de processamento entre roteador e servidor

4.3.10 Uso de memória entre roteador e servidor

Nesse último experimento será avaliada a correlação linear do percentual de uso de memória do roteador e do servidor. A Tabela 4.48 apresenta os valores do coeficiente de correlação, dos quais pode-se notar a tendência, em alguns intervalos de tempo, de haver correlação linear negativa. Entretanto, independente do sinal de r que indica uma correlação positiva ou negativa, os valores em módulo de r não ultrapassam o valor aproximado de $0,53$, o que indica uma baixa correlação linear. Adicionalmente, pode-se notar uma grande variação do valor de r calculado para cada intervalo de tempo.

A Figura 4.29 ilustra os valores dispostos em um gráfico de dispersão. Pode-se notar a péssima correlação linear de todos os dados compreendidos entre o intervalo de tempo de 07:00 às 21:00, onde em alguns momentos pode representar uma correlação nula dos dados.

Tabela 4.48: Correlação linear: uso de memória entre roteador e servidor

Período	Correlação linear	Teste da correlação com $\alpha = 10\%$	Teste da correlação com $\alpha = 5\%$
07:00 – 08:00	-0,064	Rejeitado	Rejeitado
08:00 – 09:00	-0,191	Rejeitado	Rejeitado
09:00 – 10:00	-0,515	Rejeitado	Rejeitado
10:00 – 11:00	-0,164	Rejeitado	Rejeitado
11:00 – 12:00	0,112	Aceito	Aceito
12:00 – 13:00	-0,011	Rejeitado	Rejeitado
13:00 – 14:00	0,231	Rejeitado	Rejeitado
14:00 – 15:00	0,012	Aceito	Aceito
15:00 – 16:00	-0,186	Rejeitado	Rejeitado
16:00 – 17:00	-0,452	Rejeitado	Rejeitado
17:00 – 18:00	-0,451	Rejeitado	Rejeitado
18:00 – 19:00	-0,477	Rejeitado	Rejeitado
19:00 – 20:00	-0,539	Rejeitado	Rejeitado
20:00 – 21:00	-0,265	Rejeitado	Rejeitado
07:00 – 21:00	-0,140	Rejeitado	Rejeitado

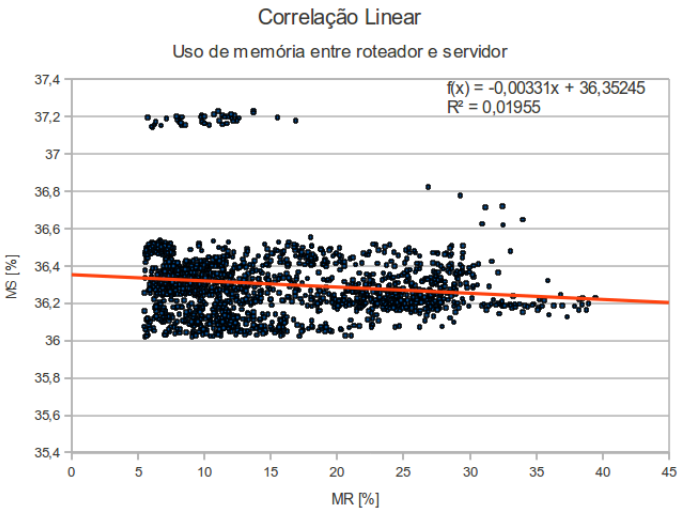

Figura 4.29: Correlação linear: uso de memória entre roteador e servidor

4.4 Grupo 4: Regressão

Os experimentos envolvendo regressão linear buscam, na maioria dos casos, encontrar uma função matemática analítica para todo o conjunto de valores e variáveis de interesse analisado sob um contexto. Até o momento foram abordados problemas envolvendo estimativa de parâmetros, análise de variância e correlação linear. Dessa forma, a partir do momento que são obtidas as função analíticas de cada variável por meio de regressão, novos problemas matemáticos podem ser avaliados e inseridos dentro do contexto da gerência de performance de redes de computadores. Dentre esses problemas podemos citar o estudo de pontos máximos e mínimos, e diferenciação. Cabe ressaltar que esses problemas não serão necessariamente de ordem estatística, de maneira que são mantidos o caráter e a natureza acadêmica da análise como um todo.

Cabe ressaltar nesse grupo de experimentos foram regredidas as médias dos valores máximos de cada variável. Para que a regressão seja realizada sem erros ou viéses, os valores do eixo X (abscissa) foram definidos dentro do intervalo $[7,5,20,5]$ com variação de 1 unidade, com o objetivo de representar os intervalos de tempo.

Reduziremos a análise de regressão, conforme sessão sobre definição de variáveis, apenas à carga de processamento e uso de memória do roteador e do servidor, *throughput* e número de pacotes do *link ethernet*. A Tabela 4.49 apresenta os coeficientes de determinação calculados a partir do *BrOffice.org* para as regressões lineares, logarítmicas e exponenciais.

A primeira conclusão que podemos obter é quanto ao coeficiente de determinação (r^2) de todas as variáveis, exceto uso de memória do roteador e servidor. O *throughput* e o número de pacotes do servidor apresentam valores de r^2 baixíssimos, menores que $0,091$. Isso significa um fraco poder nas regressões obtidas, de forma que as funções lineares, logarítmicas e exponenciais regredidas não representam, analiticamente, os valores médios máximos estimados de cada variável.

Outra conclusão que podemos chegar a partir da análise dos dados na Tabela 4.49 é quanto ao relativo aumento dos valores de r^2. O coeficiente de determinação do uso de memória do roteador está entre $0,2$ e $0,3$, sendo a regressão linear a melhor função obtida com $r^2 = 0,2904$. Quanto ao uso de memória do servidor, os coeficientes de determinação tiveram melhora considerável, dos quais pertencem à faixa de $0,49$ e $0,53$. Da mesma forma que a regressão do uso de memória do roteador, a função linear é a melhor regressão para o uso de memória para o servidor. No entanto, mesmo que os coeficientes de regressão tenham maiores

Tabela 4.49: Coeficientes de determinação das regressões lineares, logarítmicas e exponenciais

Variável	Regressão linear	Regressão logarítmica	Regressão exponencial
Carga de processamento do roteador	0,0035	0,0296	0,0004
Carga de processamento do servidor	0,0866	0,0275	0,1092
Uso de memória do roteador	0,2904	0,2066	0,2281
Uso de memória do servidor	0,5347	0,4963	0,5346
Throughput, *download*, servidor	0,0907	0,0304	0,0903
Throughput, *upload*, servidor	0,0317	0,0809	0,0262
Número de pacotes, *download*, servidor	0,0467	0,0073	0,0499
Número de pacotes, *upload*, servidor	0,0178	0,0001	0,0215

valores quando comparados à carga de processamento, *throughput* e números de pacotes, a regressão linear obtida ainda não é desejável levando em consideração que o valor máximo de r^2 é igual 1.

Adicionalmente, outra observação pertinente reside nos valores semelhantes dos coeficiente de determinação das regressões lineares e exponenciais do uso de memória do servidor. Os valores r^2 são tão próximo que, na prática, podemos concluir que a regressão exponencial tem comportamento semelhante à uma reta.

Sobre a fraca regressão obtida para a carga de processamento, *throughput* e números de pacotes, as Figuras 4.30 e 4.30 ilustram o posicionamento dos pontos das médias máximas estimadas de cada variável e a respectiva regressão linear. Os gráficos representam, respectivamente, a carga de processamento e o *throughput* (*download*) do servidor. Como os coeficientes de regressão são ruins para essas 3 variáveis, a partir da análise de dados da Tabela 4.49, não faz-se necessária a apresentação dos gráficos para a carga de processamento do servidor, *throughput* (*upload*), e número de pacotes do servidor.

Com base nos resultados nos experimentos da sessão 4.1, a carga de processamento, *throughput* e números de pacotes têm considerável crescimento nos

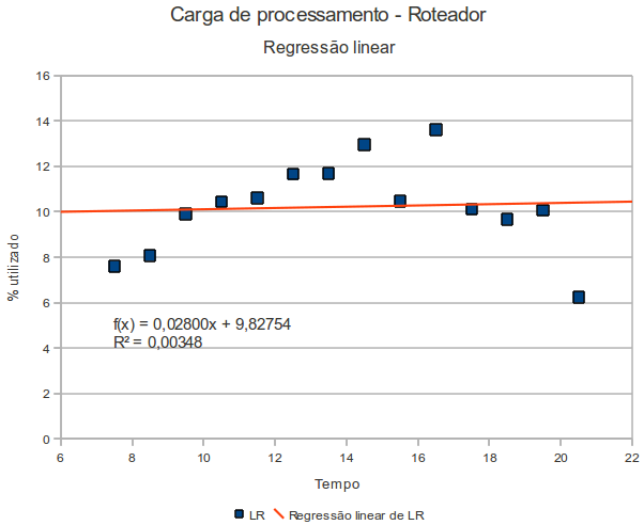

Figura 4.30: Regressão linear: carga de processamento do roteador

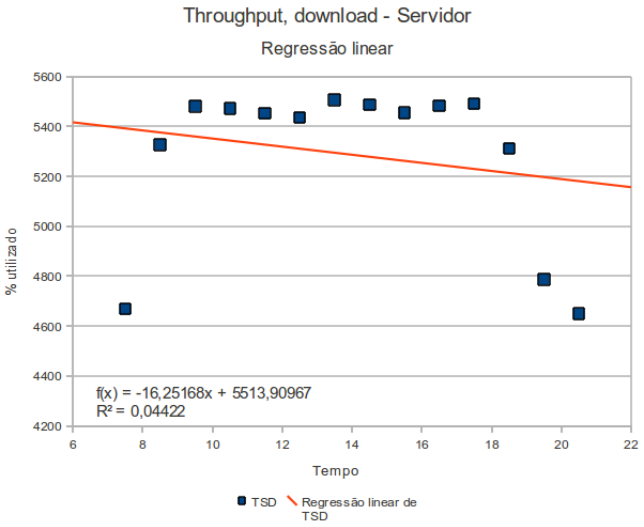

Figura 4.31: Regressão linear: *throughput*, *download*, do servidor

intervalos de tempo iniciais, estabilidade ao longo do período analisado, e decréscimo nos intervalos de tempo finais. A partir desse comportamento já averiguado, encontrar uma regressão para esse caso significa obter uma função polinomial. Nesse grupo de experimentos, por questão de simplificação, serão obtidas funções

polinomiais de grau 2. Na prática serão obtidas parábolas com o propósito de se encontrar a melhor regressão para as variáveis em questão. A Tabela 4.50 apresenta as funções polinomiais de grau 2 regredidas para os 3 conjuntos de variáveis.

Tabela 4.50: Regressões polinomiais de grau 2

Variável	**Função quadráticas regredida**
Carga de processamento do roteador [LR]	$f(x) = -0,1159^2 + 3,2666x - 10,8753$
Carga de processamento do servidor [LS]	$f(x) = -0,3133^2 + 8,4212x - 36,0049$
Throughput, *download*, servidor [TSD]	$f(x) = -34,5968^2 + 927,6496x + 50,2903$
Throughput, *upload*, servidor [TSU]	$f(x) = -15,417^2 + 445,2689x - 1019,0326$
Número de pacotes, *download*, servidor [PSD]	$f(x) = -6,993^2 + 189,9304x - 485,6241$
Número de pacotes, *upload*, servidor [PSU]	$f(x) = -7,5015^2 + 101,6363x - 566,6902$

As Figuras 4.4, 4.4 e 4.4 a seguir ilustram a regressão polinomial de cada conjunto de variáveis perante os valor médios máximos estimados.

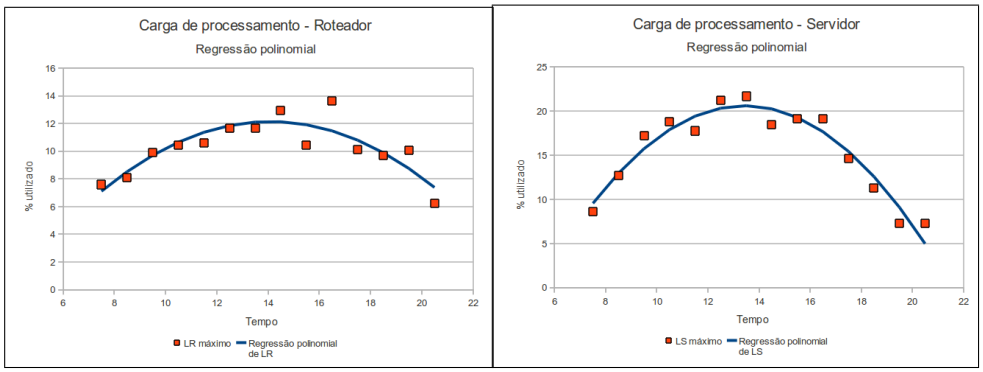

Figura 4.32: Regressão polinomial da carga de processamento do roteador e do servidor

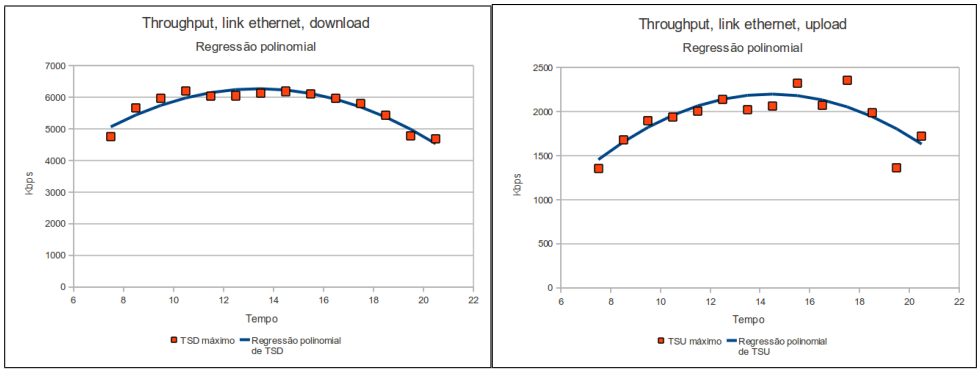

Figura 4.33: Regressão polinomial do *throughput*, *download* e *upload*, do servidor

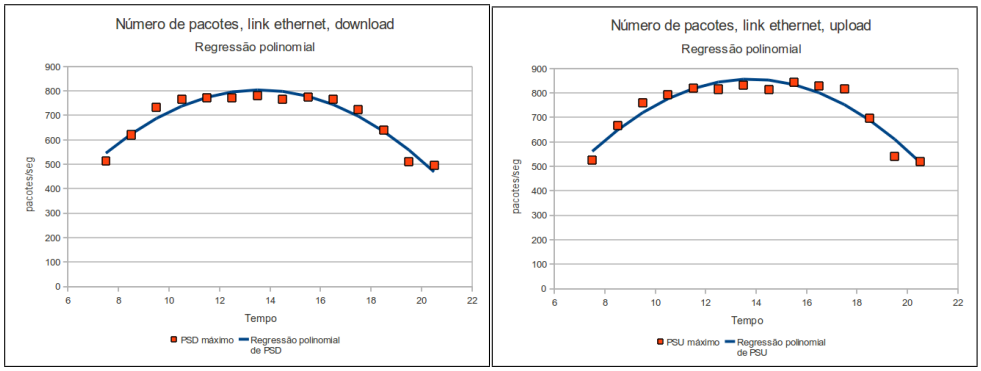

Figura 4.34: Regressão polinomial do número de pacotes, *download* e *upload*, do servidor

4.5 Comentários finais

Nessa sessão serão apresentados os comentários e conclusões finais sobre os resultados finais, de forma que seja possível encontrar possíveis relações e associações entre os diversos experimentos. A construção desses comentários foi possível mediante comparações e analogias entre os resultados de cada experimento, e observações descritivas do ambiente analisado a partir de ferramentas de monitoramento.

Inicialmente, ao se aplicar a técnica de *bootstrapping*, esperou-se que o método ajustasse todos os dados dentro da faixa de valores determinada pela amostragem original. No entanto, após repetitivos cálculos e revisões no procedimento de reamostragem, concluiu-se a polarização dos valores em torno do estimador média. Na prática os resultados foram satisfatórios, considerando a proximidade entre o estimador média da amostra original e da amostra ajustada, e a sua respectiva inclusão no intervalo de confiança.

Outra questão envolvendo estimativa de parâmetro é quanto o significado dos estimadores propriamente dito. A determinação do estimador média significa, dentro do contexto da gerência de rede de computadores, o valor esperado e encarado como normal. Outros tipos de análises, como Controle Estatístico de Processos (CEP) não abordado nesse trabalho, podem ser feitas a partir dessa estimação de parâmetro. Adicionalmente, conforme já dito anteriormente, a média dos valores máximos objetiva estimar picos e prever situações que todo o ambiente funcionará sob máxima demanda, para qualquer variável de rede analisada. Em função dessa espectativa de demanda, a média dos valores máximos é o estimador que mais interessa ao administrador de redes de computadores.

Embora esta pesquisa apresente um modelo simples de representação e caracterização de um ambiente de rede, a construção de intervalos de confiança a partir da distribuição normal é viável a partir do momento que é trabalhado o estimador média. O teorema do limite central e das combinações lineares garantem que uma estimativa obtida a partir da média de outras variáveis terá distribuição normal. Outros modelos como o de Monte Carlo e distribuições propabilísticas como a de *Poison* podem ser empregados para descrever o comportamento de uma rede de computadores, considerando a sua caracterização como uma série temporal.

Nos experimentos de análise de variância, o resultado que vale a pena ser discutido é quanto à busca de diferença significativa entre o *throughput* e o número de pacotes. No grupo de experimento envolvendo correlação linear foi verificado a forte correlação entre ambas as variáveis. Seguindo essa mesma linha de racio-

cínio espera-se que os mesmos resultados obtidos para a ANOVA do *throughput* também se repitam na ANOVA para o número de pacotes. Os resultados não retrataram essa lógica, de maneira que para o *throughput* não há diferença significativa enquanto que para o número de pacotes existe diferença significativa entre colunas (*download* e *upload*). Isso significa que, mesmo com a evidência de igualdade entre o *throughput* para o sentido de conexão, existe a diferença significativa quanto ao número de pacotes, onde na prática pode representar um elevado número de requisições.

Um detalhe importante quanto a construção de experimentos de análise de variância consiste na dependência entre as classificações. No contexto desse estudo, a ANOVA construída considera que os *links* de dados (classificação por linha) são dependentes, como por exemplo, balanceamento entre *links*. Caso essa classificação fosse independente é recomendável que a análise de variância de uma classificação com repetição fosse adotado. A principal diferença reside no número de testes realizados em cada análise. Na comparação com uma classificação o número de teste é menor, ao passo que a comparação com duas ou mais classificações o número de testes é maior devido a possibilidade de interação entre linhas e colunas. Esse fato impacta significamente no poder do teste, devido ao acúmulo de erros em cada comparação.

De uma maneira geral, analisar e encontrar uma correlação entre duas variáveis implicar dizer que ambas são dependentes entre si. Obviamente que essa correlação não necessitar ser exclusivamente linear, e sim logarítmica, exponencial, polinomial, trigonométrica, ou conforme qualquer outra função matemática. A ausência de correlação, seja qual for a função matemática analisada em conjunto, implica na independência entre as variáveis em estudo, na maioria dos casos.

Quanto aos experimentos que envolvem correlação das variáveis *throughput* e número de pacotes, o planejamento e a análise do resultado permitiu a observação do comportamento da rede a partir de quadrantes no plano cartesiano:

1. baixos valores de *throughput* e número de pacotes;
2. altos valores de *throughput* e baixo número de pacotes;
3. baixos valores de *throughput* e alto número de pacotes;
4. altos valores de *throughput* e número de pacotes.

A Figura 4.35 a seguir ilustra essa representação de quadrantes do plano cartesiano para as variáveis *throughput* e número de pacotes.

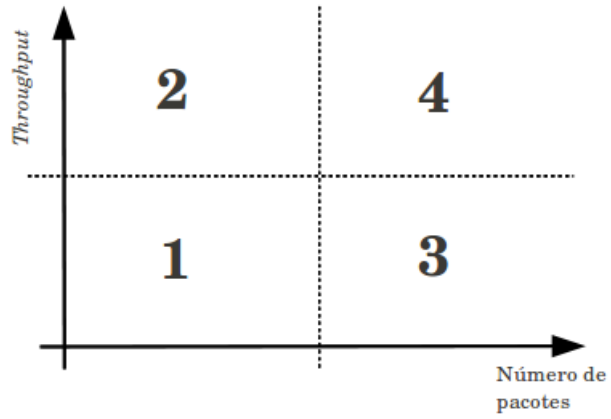

Figura 4.35: Quadrante para análise de correlação linear entre *throughput* e número de pacotes

A primeira e a quarta situação indica uma situação normal na rede, pois a vazão de dados comporta-se de forma proporcional ao número de pacotes entrantes e saintes na rede. Tal relação é comprovada, conforme já descrito nesse trabalho, nos experimentos de correlação linear. A terceira situação pode refletir uma situação anômala de segurança ou o envio demasiado de requisições, pois uma grande quantidade de pacotes na rede está relacionada com um quantidade muito baixa de informações. Já o segundo caso é pouco provável de acontecer. Como existe a forte correlação linear entre essas variáveis, altos valores de *throughput* associados a um baixo número de pacotes representa, a partir do coeficiente angular da reta de tendência, valor de MTU maior que $1500 bytes$. Isso na prática é inviável devido à maioria dos equipamentos adotar esse valor como padrão. Vale lembrar que esse raciocínio é aplicado à diversos contextos, independentemente da largura de banda disponível par ao *link* de internet.

Outra conclusão inferida está quanto ao comportamento das variáveis de uso de memória e carga de processamento. Partindo do raciocínio da dependência de variáveis e sua correlação, o resultado dos experimentos envolvendo o uso de memória e carga de processamento sugerem grande independência dessas variáveis. Não foi possível identificar nenhuma função matemática que representasse a relação entre essas variáveis, tampouco que descrevesse a sua aderência analiticamente.

De uma maneira geral, embora exista toda a aleatoriedade ao redor dos problemas de estimação de parâmetros, com cálculo de probabilidades e comportamento de variáveis conforme distribuição gaussiana, os experimentos que envolvem cor-

relação permitiu a obtenção de inferências mais significativas na gerência de rede. Como dito anteriormente, a estimação de parâmetros retrata a esperança de valores dada a amostragem e uma probabilidade de acerto. Já problemas de correlação permitem que variáveis sejam comparadas, de maneira que sejam identificadas, por meio de ferramentas matemáticas e estatísticas, perfis e comportamentos na rede gerenciada. Conforme esperado no objetivo principal desse trabalho, as ferramentas estatísticas auxiliam no entendimento da rede de computadores em questão.

Capítulo 5

Conclusão

De uma maneira geral, o entendimento do perfil da rede a partir desses resultados estatísticos é facilitado pela visualização dos gráficos de dispersão. Esse conhecimento da performance de rede, conforme proposto no capítulo introdutório desse trabalho, está agora baseado em informações estatisticamente tratadas. A comparação proposta de dados estatísticos de redes concomitante à gráficos e tabelas descritivos de ferramenta sendimentam o conhecimento do perfil do ambiente da rede em questão.

Outro aspecto que vale a pena destacar é a proximidade do estimador média de cada variável a partir das amostras originais e das amostras tratadas por *bootstrapping*. Isso nos permite chegar nas seguintes conclusões:

- O *bootstrapping* é uma técnica também aplicável para ajustes de dados de rede conforme distribuição normal, visto não só a semelhança do resultado como os estimadores destes estarem contidos nos intervalos de confiança;

- Adicionalmente, a partir da teoria estatística para a construção de intervalo de confiança, amostras com mais de 30 elementos garantem a estimação de parâmetros. Não é necessário a coleta de um número maior de dados para composição da amostra, tal como 200, 500 ou 1000 elementos. Amostras com tamanho nessa grandeza, ou maior, oferecem resultados muito próximos entre si. Se por ventura esse número de elementos ainda não for suficiente, a reamostragem por *bootstrapping* garante não só o tamanho necessário como o seu comportamento como variável gaussiana;

- Tanenbaum (1997) propõe que para variáveis de redes, milhares de valores sejam lidos para então se calcular a média. Estatisticamente, o cálculo de estimadores a partir de uma amostra com um número muito grande de elementos não implica em efeito prático na sua indução de seu valor.

A idéia de se calcular estimadores a partir de amostras com grande número de elementos está associada, muitas vezes, com ojetivo de se prever todo o comportamento da rede. Entretanto, alguns fatores devem ser levados em consideração, tais como:

- predição de tráfego de rede, consequentemente desempenho de ativos de rede em função dessas variáveis como número de pacotes, *throughput* e número de solicitações, não é uma tarefa elementar. A predição de tráfego depende, dentre diversos fatores, da estrutura e topologia física da rede, da qualidade dos equipamentos ativos no ambiente, da complexidade do protocolo TCP/IP gerando perfis não previstos, do uso atípico acarretando em situações anômala, dentre outros fatores;
- o administrador da rede que está realizando o estudo da performance da rede, seja por análise descritiva ou inferência estatística, necessita do conhecimento teórico e prático do perfil de toda infraestrutura. Isso auxilia na pré-visualização de perfis, períodos de maior uso dos recursos e entendimento do funcionamento dos equipamentos.

Esses fatores, no âmbito da estatística, podem ser caracterizados como viéses da amostra. A sua difícil predição durante a realização de um experimento estatístico, dentro do contexto de rede de computadores, é um risco considerável para o sucesso do experimento. A amostra original pode estar tão enviesada e viciada que, mesmo com a técnica de *bootstrapping*, pode não ter comportamento normal. Isso dificulta todo o estudo envolvendo a construção de intervalos de confiança, teste de hipóteses e análise de variância. Na pior das situações, a amostra obtida pode não representar, de fato, o perfil da rede estudado por conta de situações anômalas na rede.

Outra aplicação de ferramentas estatísticas que não foi abordado com mais detalhes é quanto ao teste de hipóteses. No caso de teste de hipóteses podemos averiguar se, em um dado momento de anormalidade ou qualquer outro evento representativo na rede, os estimadores da variável de interesse tem ou não diferença significativa. Isso permite gerar, de forma eficaz e com validade estatística, alertas calculados dinamicamente em função da base histórica da rede a um dado nível de

significância. Em outras palavras, a gerência de falhas dentro do modelo FCAPS pode ser garantida à uma dada probabilidade de acerto. Por exemplo, seja um latência L estimada a um nível de significância α, e uma latência momentânea L_i calculada em tempo de execução ou num intervalo de tempo. O teste de hipótese validaria estatisticamente se a latência momentânea média L_i é diferente ou não da latência histórica média L estimada.

Na linha dos testes paramétricos, a comparação de duas médias validaria uma mudança na estrutura, topologia ou configuração da rede com impacto representativo em todo o ambiente. Por exemplo, vamos considerar que a topologia de um determinado segmento de rede foi alterada, inclusive com a substituição de equipamentos de melhor performance. Com a comparação das médias anterior e atual do parâmetro latência podemos validar estatisticamente se essa ação foi de fato válida ou não, a um dado nível de significância.

Outra problemática verificada no presente trabalho é a questão das variáveis de performance de rede. Por uma questão de limitação da abrangência da presente pesquisa, vários parâmetros não foram verificados tais como: número de requisições do protocolo *tcp* ou *udp*; número de máquinas clientes ativas na rede; disponibilidade de serviço (dado em percentual); número máximo de conexões TP abertas; latência; *jitter*; dentre outras. Novos intervalos de confiança e novas correlações podem ser averiguadas entre latência, *jitter* e número de máquinas clientes.

Até o momento encaramos as variáveis como contínuas, matematicamente podendo assumir qualquer valor na reta númerica no conjunto dos números reais. No entanto é possível analisar variáveis como número de requisições, disponibilidade de serviço e número de máquinas clientes sob o ponto de vista da matemática discreta. Para essas variáveis supra citadas poder-se-ia encontrar função densidade de probabilidade discretas, tais como *Poisson* ou Binomial. Dessa forma, a partir de uma curva probabilística discreta podemos prever algumas situações no ambiente de rede com aceitável conclusão estatística.

Sob o ponto de vista do modelo OSI de rede é pertinente a aplicação de ferramentas estatísticas em elementos na camada 7 de aplicações. Em outras palavras, é válido a identificação de distribuições probabilidade, construção de intervalos, testes de hipóteses e cálculo de correlação para variáveis (número de conexões, *throughput*) que retratam a atividade de aplicações na rede. Isso na prática implica em ganhos para a gerência do controle de tráfego, consequentemente a efetiva gerência do conceito de Qualidade de Serviço (QoS).

Além de todas as questões citadas anteriormente, podemos sugerir como trabalhos futuros a partir da presente pesquisa:

- Aquisição de dados a partir do uso nativo (via *script*) do protocolo SNMP, ou de outro mecanismo de coleta de dados de performance dos próprios equipamentos monitorados;
- Composição automatizada e dinâmica do conjunto amostral das variáveis de rede, de maneira que se contemple um maior número de equipamentos, ativos de rede e servidores na análise estatística da performance;
- Criação de *framework* para realização dos experimentos estatísticos, seja *plugin* para a ferramenta Cacti, seja sistema de informação dedicado para essa finalidade. Essa aplicação possuiria conexão automatizada com a base de dados garantida pelos itens anteriores;
- Ampliação do uso das ferramentas estatísticas para exploração de outras variáveis de rede, consequentemente maior conhecimento do ambiente de rede em estudo com base em conclusões estatística;
- Emprego de ferramentas estatísticas para a determinação de Acordos de Níveis de Serviços (SLA, *Service Level Agreement*), úteis na determinação de limiares de níveis de alerta e de criticidade na gerência de falhas. Além disso, esses SLAs podem também ser aplicados na obtenção de parâmetros para a gerência de contabilização de uso dos recursos da rede.

Uma das grandes dificuldades encontradas no presente trabalho foi o tratamento do conjunto de dados, sob a premissa de terem comportamento baseado na variável gaussiana. Variáveis de redes apresentam uma grande aleatoriedade, considerando os inúmeros eventos que podem ocorrer durante o período de medição. Devido a esse fato, *throughput*, carga de processamento, latência, dentre outras, podem assumir qualquer outra distribuição diferente da normal. Isso acarretaria na anulação da estimação de parâmetros, teste de hipótese e análise de variância, e outas ferramentas para inferência estatística baseadas na distribuição normal. Caso isso ocorra, na prática, toda a análise poderia se resumir nos procedimentos de Controle Estatísticos de Processo (CEP). Além de indicadores que representam algum nível de qualidade ou fator de desempenho com base nos dados coletados, seriam gerados gráficos da média, ou qualquer outro estimador, com seus respectivos limitantes a $\pm 1\sigma$, $\pm 2\sigma$ e $\pm 3\sigma$. Obviamente que esses artefatos também podem ser utilizados na gerência de performance de rede de computadores. No entanto, o objetivo inicial do trabalho de realizar inferência estatística em variáveis de rede não seria atingido.

Outra dificuldade, podendo no entanto também ser assumido como tema para trabalhos futuros, é a determinação de níveis de significância. Na literatura é co-

mum o emprego de níveis de significância de 1% e 5%, e em alguns casos nível de 10%. Neto (2002) e Werkema (1996), inclusive, apresentam esses percentuais na explanação didática do estudo da Estatística, adotando na maioria das vezes o nível de significância de 5% por uma questão de simplificação. No entanto, os seguintes questionamentos residem na real determinação desses níveis de significância:

- dada a aleatoriedade, o número de eventos significativas, anômolos e/ou não previstos na rede, qual nível de significância adotar no processo de inferência estatística?
- considerando outros viéses já previstos na infraestrutura de rede, como o *overhead* de protocolos, aplicações e equipamentos, qual probabilidade deve ser associada às ferramentas estatística?
- atualmente, em redes cada vez mais heterogêneas conforme forma de conexão, aplicações e serviços, qual o nível de significância deve ser adotado?

Sob o ponto de vista macro de todo o trabalho, a realização dessa pesquisa permitiu a sedimentação do conhecimento envolvendo gerência de redes. Além disso, o aspecto multidisplinar que abrange conhecimentos de qualidade com o ciclo PDCA para planejamento de experimentos, protocolo SNMP, e modelos de gestão de redes de computadores permitiu a ampliação da base teórica para a gerência de performance em redes de dados. Adicionalmente, o emprego de ferramentas estatísticas aplicadas à redes de computadores fortaleceu a formação matemática e a proeficiência estatística do aluno.

Sobre a exequibilidade, o trabalho foi possível graças à diversas ferramentas livres, tais como Scilab[1], Bacula[2], Cacti[3], BrOffice.org[4], RRDtool[5] e MySQL[6]. Além do fato de todo o ambiente analisado estar baseado na plataforma livre, em específico GNU/Linux, existe uma gama de dados sobre performance gerados por todos equipamentos ativos de redes e servidores, dos quais cabem análise estatística.

Quanto à aplicabilidade da pesquisa é importante ressaltar que o uso das ferramentas estatísticas na gerência de desempenho de redes de computadores tem o

[1]Scilab: `http://www.scilab.org/`
[2]Bacula: `http://www.bacula.org/`
[3]Cacti: `http://www.cacti.net/`
[4]BrOffice.org: `http://broffice.org/`
[5]RRDtool: `http://oss.oetiker.ch/rrdtool/`
[6]MySQL: `http://www.mysql.com/`

caráter adicional, não havendo o intuito de substituir a gestão descritiva da performance atualmente feita por diversas ferramentas de monitoramento. A motivação maior do trabalho residiu na busca de conclusões estatisticamente confiáveis, além de um meio paralelo para entendimento do perfil da rede gerenciada no CEFET-MG. No que envolve os trabalho futuros, o emprego da estatística em ambientes de rede é fortalecida pela validação de mudanças e expansão da estrutura, além de trabalhos de cunho acadêmico que contemplam várias abordagens desse assunto, dentre eles, a predição de tráfego.

Referências Bibliográficas

ANGELIS, A. F. de. Tese (dissertação em física computacional), *Um modelo de tráfego de rede para aplicação de técnicas de Controle Estatístico de Processos*. São Carlos, São Paulo: [s.n.], 2003.

BOUTABA, R.; POLYRAKIS, A. Projecting fcaps to active networks. In: *Enterprise Networking, Applications and Services Conference Proceedings, 2001*. [S.l.: s.n.], 2001. p. 97 –104.

CARRANO, E. G.; WANNER, E. F.; TAKAHASHI, R. H. C. A multicriteria statistical based comparison methodology for evaluating evolutionary algorithms. *Evolutionary Computation, IEEE Transactions on*, 2011.

CASE, J.; FEDOR, M.; SCHOFFSTALL, M.; DAVIN, J. *A Simple Network Management Protocol (SNMP)*. RFC1157, maio 1990. 36 p. Disponível em: <http://www.ietf.org/rfc/rfc1157.txt>.

CASE, J.; HARRINGTON, D.; PRESUHN, R.; WIJNEN, B. *Message Processing and Dispatching for the Simple Network Management Protocol (SNMP)*. RFC2572, abr. 1999. 44 p. Disponível em: <http://www.ietf.org/rfc/rfc2572.txt>.

CASE, J.; MCCLOGHRIE, K.; ROSE, M.; WALDBUSSER, S. *Coexistence between version 1 and version 2 of the Internet-standard Network Management Framework*. RFC1452, abr. 1993. 17 p. Disponível em: <http://www.ietf.org/rfc-/rfc1452.txt>.

CASE, J.; MCCLOGHRIE, K.; ROSE, M.; WALDBUSSER, S. *Introduction to version 2 of the Internet-standard Network Management Framework*. RFC1441, abr. 1993. 13 p. Disponível em: <http://www.ietf.org/rfc/rfc1441.txt>.

CASE, J.; MCCLOGHRIE, K.; ROSE, M.; WALDBUSSER, S. *Introduction to Community-based SNMPv2*. RFC1901, jan. 1996. 8 p. Disponível em: <http://www.ietf.org/rfc/rfc1901.txt>.

CAVALCA, U. C. Monografia (graduação em licenciatura em matemática), *Ferramenta web para avaliação de pesquisas com análise estatística dos dados*. Guaratinguetá, São Paulo: [s.n.], 2007.

CENTRO DE ESTUDOS SOBRE AS TECNOLOGIAS DA INFORMAÇÃO E DA COMUNICAÇÃO. *Pesquisa sobre o uso das Tecnologias da Informação e da Comunicação no Brasil 2009*. São Paulo: Núcleo de Informação e Coordenação do Ponto BR, 2010. Disponível em: <http://www.cetic.br/tic/2009/index.htm>.

CESARIO, L. C.; BARRETO, M. C. M. Um estudo sobre o desempenho de intervalos de confiança bootstrap para a média de uma distribuição normal usando amostragem por conjuntos ordenados perfeitamente. *Revista Matemática e Estatística*, 2003.

ESR/RNP. *Arquitetura e protocolos de redes TCP-IP*. Rio de Janeiro, 2005.

ESR/RNP. *Administração de sistemas Linux: redes e segurança*. Rio de Janeiro, 2008. 256 p.

ESR/RNP. *Introdução a Infraestrutura de Chaves Públicas e Aplicações*. Brasília, 2010. 216 p. Disponível em: <http://esr.rnp.br/leitura/seguranca/icpedu>.

GOUPTA, A. Network management: Current trends and future perspectives. In: *Journal of Network and Systems Management*. [S.l.: s.n.], 2006. v. 14, n. 4, p. 483 – 491.

GOYAL, P.; MIKKILINENI, R.; GANTI, M. Fcaps in the business services fabric model. In: *Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE '09. 18th IEEE International Workshops on*. [S.l.: s.n.], 2009. p. 45 –51. ISSN 1524-4547.

JAVVIN TECHNOLOGIES. *FCAPS: Network Management Functional Model*. [S.l.], 2010. Disponível em: <http://www.networkdictionary.com/networking%-/FCAPS.php>.

LAURINDO, F. J. B. *Tecnologias da Informação, planejamento e gestão de estratégias*. São Paulo: Atlas, 2008.

LEINWAND, A.; CONROY, K. F. *Network Management: A pratical perspective*. 2nd edition. ed. United States of America: Addison Wesley, 1996. (Unix and OpenSystems series).

LIU, Y.; LIANG, X. New regulations to the next generation network. In: *Communications and Mobile Computing, 2009. CMC '09. WRI International Conference on*. [S.l.: s.n.], 2009. v. 2, p. 172 –174.

MAGALHÃES, I. L.; PINHEIRO, W. B. *Gerenciamento de Serviços de TI na Prática*. São Paulo: Novatec, 2007.

MCCLOGHRIE, K.; ROSE, M. *Management Information Base for Network Management of TCP/IP-based internets: MIB-II*. RFC1213, mar. 1991. 70 p. Disponível em: <http://www.ietf.org/rfc/rfc1213.txt>.

MEALLING, M. *A URN Namespace of Object Identifiers*. RFC3061, fev. 2001. 6 p. Disponível em: <http://www.ietf.org/rfc/rfc3061.txt>.

NARANG, N.; MITTAL, R. Network management for next generation. In: *8th International Conference on Advanced Computing and Communications*. [S.l.: s.n.], 2000.

NET SNMP. *Net SNMP distributed MIBs*. [S.l.], 2009. Disponível em: <http://www.net-snmp.org/docs/mibs/>.

NETO, P. L. de O. C. *Estatística*. 2 edição. ed. São Paulo: Edgard Blücher Ltda, 2002.

QING-LING, D.; SHU-MIN, C.; LIAN-LIANG, B.; JUN-MO, C. Application of pdca cycle in the performance management system. In: . [S.l.: s.n.], 2008. p. 1–4.

RODRIGUES, R. A. B. Monografia (especialização em Administração de Redes Linux), *Métricas e ferramentas livres para análise de capacidade em servidores Linux*. Lavras, Minas Gerais: [s.n.], 2009.

ROSE, M.; MCCLOGHRIE, K. *Structure and Identification of Management Information for TCP/IP-based Internets*. RFC1155, maio 1990. 22 p. Disponível em: <http://www.ietf.org/rfc/rfc1155.txt>.

SANTOS, F. J. J. dos. Monografia (especialização em Administração de Redes Linux), *Sistema de Gerenciamento de Redes Baseado em Conhecimento*. Lavras, Minas Gerais: [s.n.], 2004.

SANTOS, R. S. dos. Plano nacional poderá levar banda larga a 88% da população brasileira. *Pesquisa sobre o uso das Tecnologias da Informação e da Comunicação no Brasil 2009*, p. 53–57, 2010. Disponível em: <http://www.cetic.br/tic/2009/index.htm>.

SOUSA, L. S. de. Dissertação (mestrado em computação), *Avaliação e implementação de uma variação do protocolo TCP, projetada para redes de alto desempenho, visando à distribuição de objetos multimídia nas unidades de armazenamento do Servidor RIO*. Niterói, Rio de Janeiro: [s.n.], 2007.

TANENBAUM, A. S. *Redes de computadores*. 3 edição. ed. Rio de Janeiro: Campus, 1997.

TECH REPUBLIC. *Parsing XML documents with Perl's XML::Simple*. [S.l.], 2004. Disponível em: <http://www.techrepublic.com/article/parsing-xml-documents-with-perls-xmlsimple/5363190>.

WERKEMA, M. C. C. *Como estabelecer conclusões com confiança: entendendo inferência estatística*. 1 edição. ed. Belo Horizonte, Minas Gerais: Fundação Christiano Ottoni, 1996.

WERKEMA, M. C. C.; AGUIAR, S. *Planejamento e análise de experimentos: Como identificar as principais variáveis influentes em um processo*. 1 edição. ed. Belo Horizonte, Minas Gerais: Fundação Christiano Ottoni, 1996.

Apêndice A

Distribuições probabilísticas

A.1 Distribuiçao Z

Tabela A.1: Distribuição normal padronizada, valores de $P(0 \leq Z \leq z_0$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4965	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4983	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

A.2 Distribuiçao t de Student

Tabela A.2: Distribuição t de Student, valores de $t_{v,P}$ onde $P = P(t_v \geq t_{v,P})$

v/P	0,10	0,05	0,025	0,01	0,005
01	3,078	6,314	12,706	31,821	63,657
02	1,886	2,920	4,303	6,965	9,925
03	1,638	2,353	3,182	4,541	5,541
04	1,533	2,132	2,776	3,747	4,604
05	1,476	2,015	2,571	3,365	4,032
06	1,440	1,943	2,447	3,143	3,707
07	1,415	1,895	2,365	2,365	3,499
08	1,397	1,860	2,306	2,896	3,355
09	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,578
60	1,296	1,671	2,000	2,390	2,660
80	1,292	1,664	1,990	2,374	2,639
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576