Análise de Experimentos em Algoritmos de Otimização

Análise de Experimentos em Algoritmos de Otimização

Conteúdo

Conteúdo

Conteúdo

Conteúdo

Conteúdo

Conteúdo

Avaliação de Algoritmos

Sucesso de Algoritmos

É medido por dois atributos: rapidez de obtenção de uma solução e sua qualidade.

Princípios Gerais de Avaliação Experimental

- *Implementação computacional*: *estrutura de dados e economia de cálculo*.
- *Fontes de instâncias*.
- *Medida relativa da qualidade da solução*.

Avaliação de Algoritmos

Sucesso de Algoritmos

É medido por dois atributos: rapidez de obtenção de uma solução e sua qualidade.

Princípios Gerais de Avaliação Experimental

- *Implementação computacional*: *estrutura de dados e economia de cálculo*.
- *Fontes de instâncias*.
- *Medida relativa da qualidade da solução.*

Avaliação de Algoritmos

Sucesso de Algoritmos

É medido por dois atributos: rapidez de obtenção de uma solução e sua qualidade.

Princípios Gerais de Avaliação Experimental

- *Implementação computacional*: *estrutura de dados e economia de cálculo*.
- *Fontes de instâncias*.
- *Medida relativa da qualidade da solução*.

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Avaliação de Algoritmos

- *Tempo computacional até critério de parada e tempo para encontrar a melhor solução.*
- *Robustez em relação ao desempenho em instâncias distintas.*
- *Escolha e análise de componentes do algoritmo.*
- *Descrição do método usado para configurar parâmetros.*
- *Algoritmo deve ser reprodutível.*
- *Escolha dos fatores que influenciam o desempenho do algoritmo.*
- *Comparação entre algoritmos.*
- *Relatório e análise de resultados com tabelas e gráficos.*
- *Extração de conclusões dos experimentos.*

Referências

Journal of Heuristics, 1: 9–32 (1995)

Designing and Reporting on Computational Experiments with Heuristic Methods

RICHARD S. BARR
Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275-0122, phone: (214) 768-2605, fax: (214) 768-3085, email: barr@seas.smu.edu

BRUCE L. GOLDEN
College of Business and Management, University of Maryland, College Park, MD 20742, email: bgolden@umdacc.umd.edu

JAMES P. KELLY
College of Business and Administration, University of Colorado at Boulder, Boulder, CO 80309, email: james.kelly@colorado.edu

MAURICIO G.C. RESENDE
Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill, NJ 07974-2040, email: mgcr@research.att.com

WILLIAM R. STEWART, JR.
School of Business Administration, The College of William and Mary, Williamsburg, VA 23185, email: wrstew@mail.wm.edu

Abstract

This article discusses the design of computational experiments to test heuristic methods and provides reporting guidelines for such experimentation. The goal is to promote thoughtful, well-planned, and extensive testing of heuristics, full disclosure of experimental conditions, and integrity in and reproducibility of the reported results.

Key Words: heuristics, algorithms, experimental design, computational testing

While heuristic methods have always been a part of human problem solving, the mathematical versions are growing in their range of application as well as their variety of approach. New heuristic technologies are giving operations researchers, computer scientists, and practitioners the ability to routinely solve problems that were too large or complex for previous generations of algorithms.

The effectiveness of any proposed methodology for solving a given class of problems can be demonstrated by theoretical analysis and empirical testing. This article focuses on the issues involved in designing computational experiments to test heuristic methods and gives guidelines for reporting on the experimentation. When a new heuristic is presented in the computational and mathematical sciences literature, its contributions should be evaluated scientifically and reported in an objective manner, and yet this is not always done.

We follow in the footsteps of those pioneers who have championed high-quality reporting of computational experiments with mathematical programming software. These efforts began in the late 1970s with Crowder, Dembo, and Mulvey (1980), Gilsinn et al. (1977), Jackson

Referências

Journal of Heuristics, 7: 261–304 (2001)

Experimental Evaluation of Heuristic Optimization Algorithms: A Tutorial

RONALD L. RARDIN AND REHA UZSOY
School of Industrial Engineering, Purdue University, West Lafayette, IN 47907-1287, USA

Abstract

Heuristic optimization algorithms seek good feasible solutions to optimization problems in circumstances where the complexities of the problem or the limited time available for solution do not allow exact solution. Although worst case and probabilistic analysis of algorithms have produced insight on some classic models, most of the heuristics developed for large optimization problem must be evaluated empirically—by applying procedures to a collection of specific instances and comparing the observed solution quality and computational burden.

This paper focuses on the methodological issues that must be confronted by researchers undertaking such experimental evaluations of heuristics, including experimental design, sources of test instances, measures of algorithmic performance, analysis of results, and presentation in papers and talks. The questions are difficult, and there are no clear right answers. We seek only to highlight the main issues, present alternative ways of addressing them under different circumstances, and caution about pitfalls to avoid.

Key Words: Heuristic optimization, computational experiments

1. Introduction

Heuristic optimization algorithms (heuristisics for short) seek good feasible solutions to optimization problems in circumstances where the complexity of the problem or the limited time available for its solution do not allow exact solution. The formal intractability, in the sense of *NP*-hardness (Garey and Johnson, 1979), of many commonly encountered optimization problems and the growing use of real-time control have made the development of heuristics a major area within the field of operations research.

Unlike exact algorithms, where time-efficiency is the main measure of success, there are two burning issues in evaluating heuristics: how fast can solutions be obtained and how close do they come to being optimal. One body of research, beginning with Graham's investigations of heuristics for parallel processor scheduling problems (Graham, 1969), demands polynomially bounded time and seeks provable limits on the worst-case or maximum deviation of a heuristic from the optimal solution value. Another, pioneered by Karp (1977), tries to compute the expected deviation from optimality, assuming polynomial time and a probabilistic structure for the problem instances considered.

Both these approaches have yielded significant insights into a number of heuristics and problems, and both have the appeal of mathematical certainty. Still, their analytical difficulty, which makes it hard to obtain results for most realistic problems and algorithms, severely limits their range of application. Also, a worst-case result, which is by definition

Referências

A Theoretician's Guide to the Experimental Analysis of Algorithms

David S. Johnson
AT&T Labs – Research
http://www.research.att.com/~dsj/

November 25, 2001

Abstract

This paper presents an informal discussion of issues that arise when one attempts to analyze algorithms experimentally. It is based on lessons learned by the author over the course of more than a decade of experimentation, survey paper writing, refereeing, and lively discussions with other experimentalists. Although written from the perspective of a theoretical computer scientist, it is intended to be of use to researchers from all fields who want to study algorithms experimentally. It has two goals: first, to provide a useful guide to new experimentalists about how such work can best be performed and written up, and second, to challenge current researchers to think about whether their own work might be improved from a scientific point of view. With the latter purpose in mind, the author hopes that at least a few of his recommendations will be considered controversial.

DRAFT of a paper that appeared in the Proceedings of the 5th and 6th DIMACS Implementation Challenges, Goldwasser, Johnson, and McGeoch, (eds), American Mathematical Society, 2002, 215-250.

Inferência Estatística

- Métodos para tomada de decisões ou para extrair conclusões de uma população de instâncias de um problema de otimização.
- Teste *t-student*, teste-*F*, chi-quadrado, análise de variância, estatística não paramétrica.
- Métodos usam informação de uma amostra da população.
- As duas classe principais de problemas de inferência estatística são estimação de parâmetros e teste de hipóteses.

Inferência Estatística

- Métodos para tomada de decisões ou para extrair conclusões de uma população de instâncias de um problema de otimização.
- Teste *t-student*, teste-*F*, chi-quadrado, análise de variância, estatística não paramétrica.
- Métodos usam informação de uma amostra da população.
- As duas classe principais de problemas de inferência estatística são estimação de parâmetros e teste de hipóteses.

Inferência Estatística

- Métodos para tomada de decisões ou para extrair conclusões de uma população de instâncias de um problema de otimização.
- Teste *t-student*, teste-*F*, chi-quadrado, análise de variância, estatística não paramétrica.
- Métodos usam informação de uma amostra da população.
- As duas classe principais de problemas de inferência estatística são estimação de parâmetros e teste de hipóteses.

Inferência Estatística

- Métodos para tomada de decisões ou para extrair conclusões de uma população de instâncias de um problema de otimização.
- Teste *t-student*, teste-*F*, chi-quadrado, análise de variância, estatística não paramétrica.
- Métodos usam informação de uma amostra da população.
- As duas classe principais de problemas de inferência estatística são estimação de parâmetros e teste de hipóteses.

Amostragem Aleatória

Definição

As variáveis aleatórias $X_1, X_2, \ldots, X_n$ são uma amostra aleatória de uma variável aleatória X desconhecida se são independentes e têm a mesma função densidade de probabilidade. A função densidade de probabilidade da amostra é

$$f(x_1, x_2, \cdots x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

- Seja X uma variável aleatória com média μ e variância σ^2.

Definição

Uma estatística é uma variável aleatória que é função das variáveis aleatórias correspondentes a uma amostra aleatória.
Exemplo: *Média e variância da amostra aleatória.*

Amostragem Aleatória

Definição

As variáveis aleatórias $X_1, X_2, \ldots, X_n$ são uma amostra aleatória de uma variável aleatória X desconhecida se são independentes e têm a mesma função densidade de probabilidade. A função densidade de probabilidade da amostra é

$$f(x_1, x_2, \cdots x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

- Seja X uma variável aleatória com média μ e variância σ^2.

Definição

Uma estatística é uma variável aleatória que é função das variáveis aleatórias correspondentes a uma amostra aleatória.
Exemplo: *Média e variância da amostra aleatória.*

Estimadores

- θ : parâmetro desconhecido associado com a distribuição de probabilidade da variável aleatória X.
- A estatística $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ é uma variável aleatória chamada estimador do ponto θ.
- Quando a amostra é selecionada, $\hat{\Theta}$ assume um valor $\hat{\theta}$ chamado de estimativa de θ.
- $\hat{\mu}$ = estimativa da média μ da variável aleatória X.
- $\hat{\sigma}^2$ = estimativa da variância σ^2 da variável aleatória X.

Estimadores

- θ : parâmetro desconhecido associado com a distribuição de probabilidade da variável aleatória X.
- A estatística $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ é uma variável aleatória chamada estimador do ponto θ.
- Quando a amostra é selecionada, $\hat{\Theta}$ assume um valor $\hat{\theta}$ chamado de estimativa de θ.
- $\hat{\mu}$ = estimativa da média μ da variável aleatória X.
- $\hat{\sigma}^2$ = estimativa da variância σ^2 da variável aleatória X.

Propriedades de Estimadores

Definição

$\hat{\Theta}$ é um estimador não tendencioso (unbiased) para o parâmetro θ se

$$E(\hat{\Theta}) = \theta$$

Se o estimador não é tendencioso, a diferença

$$E(\hat{\Theta}) - \theta$$

é chamada de tendência do estimador $\hat{\Theta}$.

Propriedades de Estimadores

Definição

$\hat{\Theta}$ é um estimador não tendencioso (unbiased) para o parâmetro θ se

$$E(\hat{\Theta}) = \theta$$

Se o estimador não é tendencioso, a diferença

$$E(\hat{\Theta}) - \theta$$

é chamada de tendência do estimador $\hat{\Theta}$.

Propriedades de Estimadores

Proposição

i) A média amostral $\bar{X} = \sum_{i=1}^{n} X_i$ é uma estimativa da média $\mu = E[X]$ da população:

$$\bar{X} = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu.$$

ii) A variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

é uma estimativa da variância σ^2 da população.

Propriedades de Estimadores

Proposição

i) A média amostral $\bar{X} = \sum_{i=1}^{n} X_i$ é uma estimativa da média $\mu = E[X]$ da população:

$$\bar{X} = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu.$$

ii) A variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

é uma estimativa da variância σ^2 da população.

Propriedades de Estimadores

$$S^2 = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2\bar{X}X_i)\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right].$$

- Como $E[X_i^2] = \mu^2 + \sigma^2$ e $E[\bar{X}^2] = (E[\bar{X}])^2 + \mathrm{Var}(\bar{X}) = \mu^2 + \sigma^2/n$

$$E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n)\right]$$

$$= \frac{1}{n-1}(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2) = \sigma^2.$$

Propriedades de Estimadores

$$S^2 = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2\bar{X}X_i)\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right].$$

- Como $E[X_i^2] = \mu^2 + \sigma^2$ e $E[\bar{X}^2] = (E[\bar{X}])^2 + \text{Var}(\bar{X}) = \mu^2 + \sigma^2/n$

$$E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n)\right]$$

$$= \frac{1}{n-1}(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2) = \sigma^2.$$

Conceito de Grau de Liberade

- A variância S^2 de uma amostra aleatória $X_1, X_2, \ldots, X_n$ é

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2}{n-1}$$

tal que

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

e o desvio

$$d_i = X_i - \bar{X}$$

- Como

$$\sum_{i=1}^{n} d_i = 0 \qquad (1)$$

- $n-1$ desvios estão "livres" para assumir qualquer valor, enquanto o $n-$ésimo desvio tem que assumir um valor tal que (1) seja satisfeita.
- Diz-se que existem $n-1$ graus de liberdade para a variância amostral, que refletem os $n-1$ desvios "livres".

Conceito de Grau de Liberade

- A variância S^2 de uma amostra aleatória $X_1, X_2, \ldots, X_n$ é

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2}{n-1}$$

tal que

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

e o desvio

$$d_i = X_i - \bar{X}$$

- Como

$$\sum_{i=1}^{n} d_i = 0 \qquad (1)$$

- $n-1$ desvios estão ”livres” para assumir qualquer valor, enquanto o $n-$ésimo desvio tem que assumir um valor tal que (1) seja satisfeita.

- Diz-se que existem $n-1$ graus de liberdade para a variância amostral, que refletem os $n-1$ desvios ”livres”.

Propriedades de Média e Variância

Transformação Linear

Seja X uma variável aleatória com média $E[X]$ e desvio padrão σ_X.
Para a transformação linear $Y = \frac{X - E[X]}{\sigma_X}$

$$E[Y] = 0, \quad \text{var}(Y) = 1$$

Variância Expressa por Momentos

Seja X uma variável aleatória com média $E[X]$ e variância $\text{var}(X)$.

$$\text{var}(X) = E[X^2] - (E[X])^2$$

Propriedades de Média e Variância

Transformação Linear

Seja X uma variável aleatória com média $E[X]$ e desvio padrão σ_X.
Para a transformação linear $Y = \frac{X - E[X]}{\sigma_X}$

$$E[Y] = 0, \quad \text{var}(Y) = 1$$

Variância Expressa por Momentos

Seja X uma variável aleatória com média $E[X]$ e variância $\text{var}(X)$.

$$\text{var}(X) = E[X^2] - (E[X])^2$$

Função Geradora de Momentos

Definição

Para uma dada variável aleatória X e um parâmetro s, a função geradora de momentos é dada por $M_X(s) = E[e^{sX}]$.

- Para uma variável aleatória contínua X com FDP $f_X(x)$

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$$

- Tomando a derivada com relação a s dos dois lados

$$\frac{d}{ds} M_X(s) = \frac{d}{ds} \int_{-\infty}^{\infty} e^{sx} f_X(x) dx = \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x) dx = \int_{-\infty}^{\infty} x e^{sx} f_X(x) dx$$

- A troca entre derivada e integral pode ser feita sob certas condições (ver teorema de**Leibniz**.

Função Geradora de Momentos

- Quando $s = 0$

$$\frac{d}{ds} M_X(s)|_{s=0} = \int_{-\infty}^{\infty} x f_X(x) dx = E[X]$$

- Se diferenciarmos n vezes

$$\frac{d^n}{ds^n} M_X(s)|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x) dx = E[X^n]$$

Soma de Variáveis Independentes

$$Z = X_1 + \cdots + X_n$$

$$M_Z(s) = E[e^{sZ}] = E[e^{s(X_1+\cdots+X_n)}] = E[e^{sX_1} \cdots e^{sX_n}]$$

$$= E[e^{sX_1}] \cdots E[e^{sX_n}] = M_{X_1}(s) \cdots M_{X_n}(s)$$

Função Geradora de Momentos

- Quando $s = 0$

$$\frac{d}{ds} M_X(s)|_{s=0} = \int_{-\infty}^{\infty} x f_X(x) dx = E[X]$$

- Se diferenciarmos n vezes

$$\frac{d^n}{ds^n} M_X(s)|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x) dx = E[X^n]$$

Soma de Variáveis Independentes

$$Z = X_1 + \cdots + X_n$$

$$M_Z(s) = E[e^{sZ}] = E[e^{s(X_1+\cdots+X_n)}] = E[e^{sX_1} \cdots e^{sX_n}]$$

$$= E[e^{sX_1}] \cdots E[e^{sX_n}] = M_{X_1}(s) \cdots M_{X_n}(s)$$

Função de Densidade de Probabilidade Normal

Definição

Uma variável aleatória contínua X é chamada normal (Abraham de Moivre, 1733) se tem a função de densidade de probabilidade (FDP)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

em que μ e $\sigma > 0$ são parâmetros que caracterizam a FDP.

- Veriquemos que

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}\, dx = 1.$$

- A transformação linear $Y = (X - \mu)/\sigma$ faz com que $E[Y] = 0$ e var(Y) = 1, e resulta na integral

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy = 1.$$

Função de Densidade de Probabilidade Normal

$$
\begin{aligned}
\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y)^2/2}\,dxdy \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2} r\,drd\theta \\
&= \int_{0}^{\infty} e^{-r^2/2} r\,dr \\
&= \int_{0}^{\infty} e^{-u}\,du \\
&= 1
\end{aligned}
$$

- A partir da mesma transformação linear, demonstra-se que se X é normal

$$E[X] = \mu, \quad \text{var}(X) = \sigma^2$$

Função Gamma

Definição

A função gamma é motivada (Euler, 1729?) pelo problema de encontrar uma curva suave que conecta os pontos (x, y) dados por $y = (x-1)!$ em pontos inteiros positivos de x, isto é, a função fatorial. A função gamma é dada por $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$.

- Para $\alpha > 1$ e integração por partes

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, d(e^{-x})$$

$$= x^{\alpha-1}(-e^{-x})|_0^\infty - \int_0^\infty (-e^{-x})(\alpha-1)x^{\alpha-2}\, dx$$

$$= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1} e^{-x}\, dx = (\alpha-1)\Gamma\left((\alpha-1)\right)$$

- Para $\alpha = 1$, $\int_0^\infty e^{-x} = 1$. Portanto,

$\Gamma(2) = 1 \cdot 1$, $\Gamma(3) = 2 \cdot 1$, $\Gamma(4) = 3 \cdot 2 \cdot 1$, e por indução, $\Gamma(n) = (n-1)!$.

Resultado útil : $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Função Gamma

Definição

A função gamma é motivada (Euler, 1729?) pelo problema de encontrar uma curva suave que conecta os pontos (x, y) dados por $y = (x-1)!$ em pontos inteiros positivos de x, isto é, a função fatorial. A função gamma é dada por $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$.

- Para $\alpha > 1$ e integração por partes

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, d(e^{-x})$$

$$= x^{\alpha-1}(-e^{-x})|_0^\infty - \int_0^\infty (-e^{-x})(\alpha-1)x^{\alpha-2}\, dx$$

$$= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1} e^{-x}\, dx = (\alpha-1)\Gamma\left((\alpha-1)\right)$$

- Para $\alpha = 1$, $\int_0^\infty e^{-x} = 1$. Portanto,

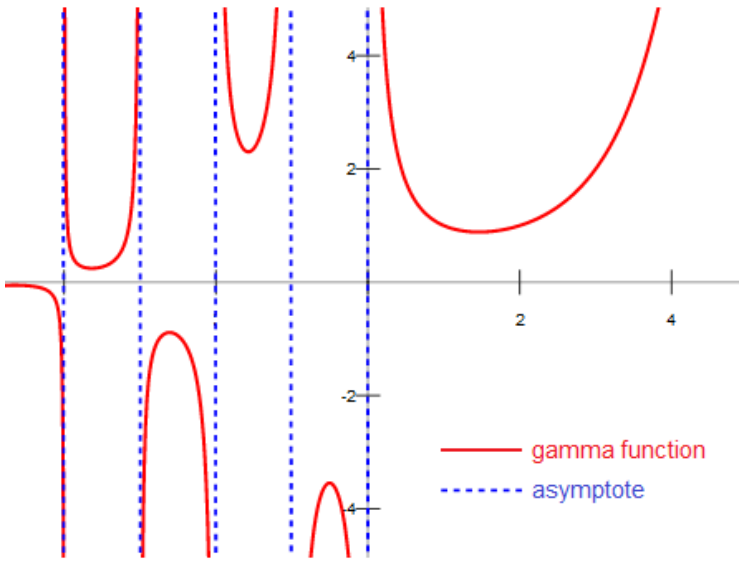
$$\Gamma(2) = 1 \cdot 1,\ \Gamma(3) = 2 \cdot 1,\ \Gamma(4) = 3 \cdot 2 \cdot 1,\ \text{e por indução, } \Gamma(n) = (n-1)!.$$

$$\text{Resultado útil : } \Gamma(\frac{1}{2}) = \sqrt{\pi}$$

Função Gamma

- Demonstra-se que a integral $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ diverge para $x \leq 0$.

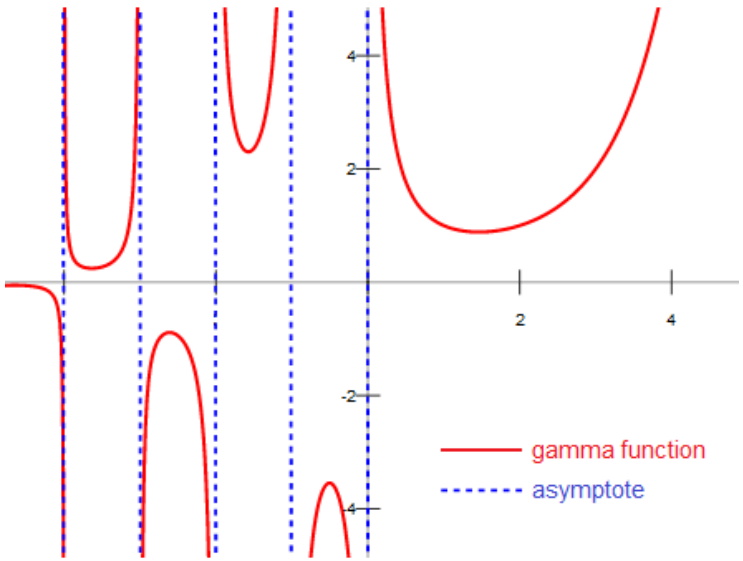
- A função $\frac{1}{\Gamma(z)}$ definida no plano complexo estende a função gamma com polos em $0, 1, 2, \ldots$



Função Gamma

- Demonstra-se que a integral $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ diverge para $x \leq 0$.

- A função $\frac{1}{\Gamma(z)}$ definida no plano complexo estende a função gamma com polos em $0, 1, 2, \ldots$



Função Densidade de Probabilidade Gamma

Definição

A FDP gamma com parâmetros α e β é dada por $\Gamma(\alpha, \beta) = \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}$.

- Ao se fazer $\alpha = 1$ e $\beta = \frac{1}{\lambda}$, obtém-se a FDP exponencial.
- Para α inteiro e positivo, a FDP gamma é chamada de FDP de Erlang ou processo de Poisson com parâmetro λ.
- A distribuição de Erlang é usada em sistemas de tráfego de sinais e em sistemas de filas com tempos de espera.

Função Densidade de Probabilidade Gamma

Definição

A FDP gamma com parâmetros α e β é dada por $\Gamma(\alpha, \beta) = \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}$.

- Ao se fazer $\alpha = 1$ e $\beta = \frac{1}{\lambda}$, obtém-se a FDP exponencial.
- Para α inteiro e positivo, a FDP gamma é chamada de FDP de Erlang ou processo de Poisson com parâmetro λ.
- A distribuição de Erlang é usada em sistemas de tráfego de sinais e em sistemas de filas com tempos de espera.

Relação entre Probabilidade Gamma e Processo de Poisson

Suponha que as chegadas seguem um processo de Poisson com parâmetro λ. Qual é o tempo até a $r-$ésima chegada? Seja

$F_r(t) = P(r-\text{ésima chegada} \leq t)$

$1 - F_r(t) = \sum_{k=0}^{r-1} \frac{1}{k!}(\lambda t)^k e^{-\lambda t}$

= probabilidade que o tempo de ocorrência da $r-$ésima chegada é maior que t.

= probabilidade que o número de chegadas no intervalo de tempo $[0, t]$ é menor que r.

$$\frac{d\,(1 - F_r(t))}{dt} = -\lambda e^{-\lambda t}\left[\sum_{k=0}^{r-1}\frac{1}{k!}(\lambda t)^k\right] + \lambda e^{-\lambda t}\sum_{k=0}^{r-1}\frac{k}{k!}(\lambda t)^{k-1}$$

$$= \lambda e^{-\lambda t}\left[\left[-\sum_{k=0}^{r-1}\frac{1}{k!}(\lambda t)^k\right] + \sum_{k=0}^{r-1}\frac{1}{(k-1)!}(\lambda t)^{k-1}\right]$$

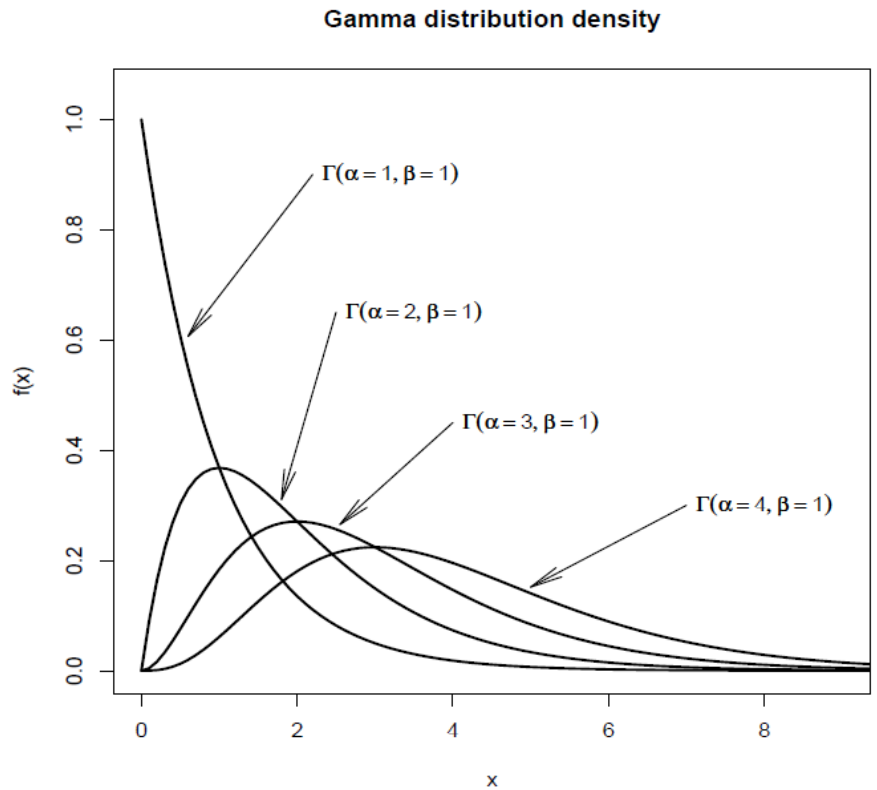
$$= \frac{-\lambda e^{-\lambda t}}{(r-1)!}(\lambda t)^{r-1}, \quad t \geq 0$$

Como desejamos $\frac{d(F_r(t))}{dt}$, a função densidade de probabilidade do tempo até a r-ésima chegada é

$$\frac{\lambda e^{-\lambda t}}{(r-1)!}(\lambda t)^{r-1}, \quad t \geq 0,$$

que é a função densidade de probabilidade gama para r inteiro.

Função Densidade de Probabilidade Gamma



Função geradora de momentos da variável aleatória $X \sim \Gamma(\alpha, \beta)$

Proposição

$$M_X(s) = \left(\frac{1}{1-\beta s}\right)^{\alpha}, \quad s < \frac{1}{\beta}$$

Demonstração

$$M_X(s) = E[e^{sX}] = \int_0^{\infty} e^{sx} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\, dx$$

$$= \int_0^{\infty} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/(\frac{\beta}{1-\beta s})}\, dx = \frac{\beta^{\alpha}}{(\beta-s)^{\alpha}} \int_0^{\infty} \frac{(\beta-s)^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/(\frac{1}{1-\beta s})}\, dx$$

$$M_X(s) = \left(\frac{\beta}{\beta-s}\right)^{\alpha} = \left(\frac{1}{1-\beta s}\right)^{\alpha} \quad \text{se } s < \frac{1}{\beta}.$$

Se $s \geq 1/\beta$, $1 - \beta s \leq 0$ e a integral é infinita.

Função Densidade de Probabilidade Chi-Quadrado

Esta distribuição foi descrita pela primeira vez pela estatístico alemão Friedrich Robert Helmert em artigos de 1875 e 1876, e independentemente redescoberta pela matemático inglês Karl Pearson em 1900.

A distribuição chi-quadrado é um caso especial da distribuição gama e é usada em:

- Qualidade da adequação de uma distribuição observada em relação a uma distribuição teórica.
- Confiança no intervalo de estimação para o desvio padrão de uma distribuição normal de um desvio padrão experimental.
- Uso em outros testes estatísticos, por exemplo, a análise da variança por *ranks*.

Função Densidade de Probabilidade Chi-Quadrado

Esta distribuição foi descrita pela primeira vez pela estatístico alemão Friedrich Robert Helmert em artigos de 1875 e 1876, e independentemente redescoberta pela matemático inglês Karl Pearson em 1900.

A distribuição chi-quadrado é um caso especial da distribuição gama e é usada em:

- Qualidade da adequação de uma distribuição observada em relação a uma distribuição teórica.
- Confiança no intervalo de estimação para o desvio padrão de uma distribuição normal de um desvio padrão experimental.
- Uso em outros testes estatísticos, por exemplo, a análise da variança por *ranks*.

Função Densidade de Probabilidade Chi-Quadrado

Esta distribuição foi descrita pela primeira vez pela estatístico alemão Friedrich Robert Helmert em artigos de 1875 e 1876, e independentemente redescoberta pela matemático inglês Karl Pearson em 1900.

A distribuição chi-quadrado é um caso especial da distribuição gama e é usada em:

- Qualidade da adequação de uma distribuição observada em relação a uma distribuição teórica.
- Confiança no intervalo de estimação para o desvio padrão de uma distribuição normal de um desvio padrão experimental.
- Uso em outros testes estatísticos, por exemplo, a análise da variança por *ranks*.

Função Densidade de Probabilidade Chi-Quadrado

Esta distribuição foi descrita pela primeira vez pela estatístico alemão Friedrich Robert Helmert em artigos de 1875 e 1876, e independentemente redescoberta pela matemático inglês Karl Pearson em 1900.

A distribuição chi-quadrado é um caso especial da distribuição gama e é usada em:

- Qualidade da adequação de uma distribuição observada em relação a uma distribuição teórica.
- Confiança no intervalo de estimação para o desvio padrão de uma distribuição normal de um desvio padrão experimental.
- Uso em outros testes estatísticos, por exemplo, a análise da variança por *ranks*.

Função Densidade de Probabilidade Chi-Quadrado

Teorema

Seja $Z \sim N(0,1)$. Se $X = Z^2$, então X tem a FDP chi-quadrado com 1 grau de liberdade, isto é, $X \sim \chi^2_1$.

Demonstração

$$P(X \leq x) = P(Z^2 \leq x)) = P(-x^{1/2} \leq Z \leq x^{1/2}) = \int_{-x^{1/2}}^{x^{1/2}} f_Z(z)dz$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$f_X(x) = \frac{dF_X(x)}{dx}$$

$$= \frac{d}{dx} \int_{-x^{1/2}}^{x^{1/2}} f_Z(z)dz$$

$$= f_Z\left(x^{1/2}\right) \frac{d\left(x^{1/2}\right)}{dx} - f_Z\left(-x^{1/2}\right) \frac{d\left(-x^{1/2}\right)}{dx}$$

Teorema de Leibniz novamente.

Função Densidade de Probabilidade Chi-Quadrado

$$= \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(x^{\frac{1}{2}}\right)^2\right)\frac{1}{2}x^{-1/2}$$

$$-\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(-x^{\frac{1}{2}}\right)^2\right)\left(-\frac{1}{2}x^{-1/2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\frac{1}{2}x^{-1/2}exp\left(-\frac{1}{2}x\right) + \frac{1}{\sqrt{2\pi}}\frac{1}{2}x^{-1/2}exp\left(-\frac{1}{2}x\right)$$

$$= \frac{1}{\sqrt{2\pi}}x^{-1/2}exp\left(-\frac{1}{2}x\right)$$

$$= \frac{2^{1/2}}{\Gamma(1/2)}x^{-1/2}exp\left(-\frac{1}{2}x\right)$$

Portanto,

$$f_X(x) = \begin{cases} \frac{2^{1/2}}{\Gamma(1/2)}x^{-1/2}exp\left(-\frac{1}{2}x\right), & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases}$$

que corresponde à função densidade de probabilidade gamma com $\alpha = \frac{1}{2}$ e $\beta = 2$.

Função Densidade de Probabilidade Chi-Quadrado

Teorema

Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com $Z_i \sim N(0,1), i = 1, 2, \ldots, n$. Se $X = \sum_{i=1}^{n} Z_i^2$, então X tem a FDP chi-quadrado com n graus de liberdade, isto é, $X \sim \chi_n^2$.

Demonstração

Como $Z_1, Z_2, \ldots, Z_n$ são independentes, então a função geradora de momentos de X é

$$M_X(s) = M_{Z_1^2}(s) \times M_{Z_2^2}(s) \times \cdots M_{Z_n^2}(s)$$

Como $Z_i^2 \sim \chi_1^2$, então $M_X(s) = \left(\frac{1}{1-2s}\right)^{\frac{-n}{2}}$.

- Esta é a função geradora de momentos de $\Gamma(\frac{n}{2}, 2)$, chamada distribuição chi-quadrado com n graus de liberdade.

Função Densidade de Probabilidade Chi-Quadrado

Teorema

Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com $Z_i \sim N(0,1), i = 1, 2, \ldots, n$. Se $X = \sum_{i=1}^{n} Z_i^2$, então X tem a FDP chi-quadrado com n graus de liberdade, isto é, $X \sim \chi_n^2$.

Demonstração

Como $Z_1, Z_2, \ldots, Z_n$ são independentes, então a função geradora de momentos de X é

$$M_X(s) = M_{Z_1^2}(s) \times M_{Z_2^2}(s) \times \cdots M_{Z_n^2}(s)$$

Como $Z_i^2 \sim \chi_1^2$, então $M_X(s) = \left(\frac{1}{1-2s}\right)^{\frac{-n}{2}}$.

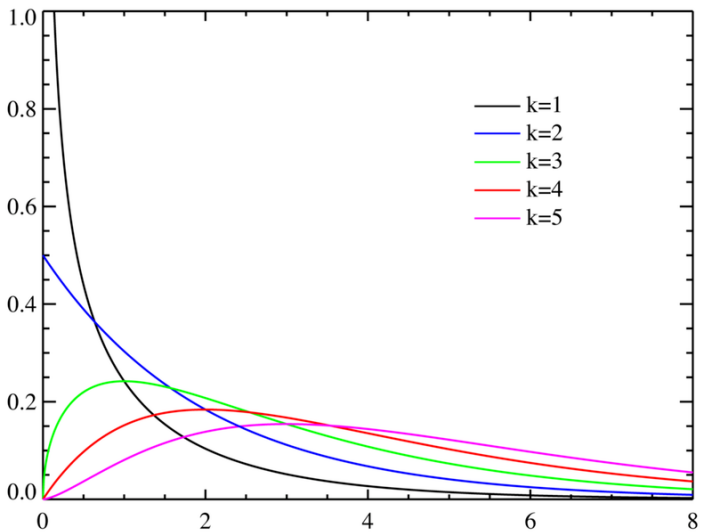
- Esta é a função geradora de momentos de $\Gamma(\frac{n}{2}, 2)$, chamada distribuição chi-quadrado com n graus de liberdade.

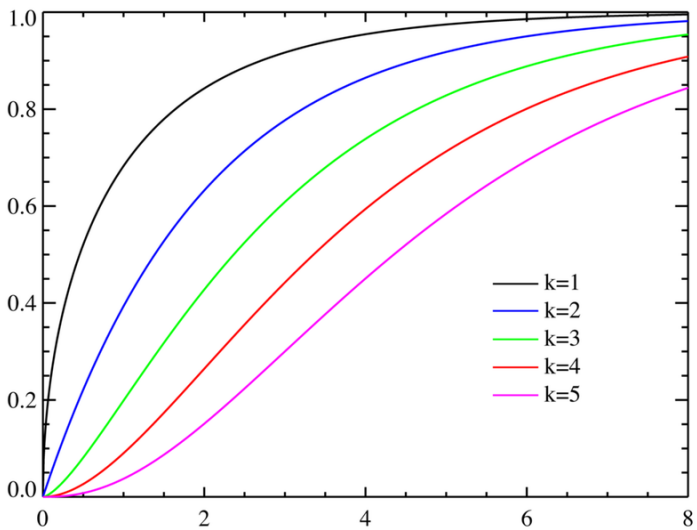
Função Densidade de Probabilidade Chi-Quadrado

Teorema

Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com $Z_i \sim N(\mu, \sigma), i = 1, 2, \ldots, n$. Do teorema anterior segue-se que se

$$X = \sum_{i=1}^{n} \left(\frac{Z_i - \mu}{\sigma} \right), \text{ então } X \sim \chi^2_n$$



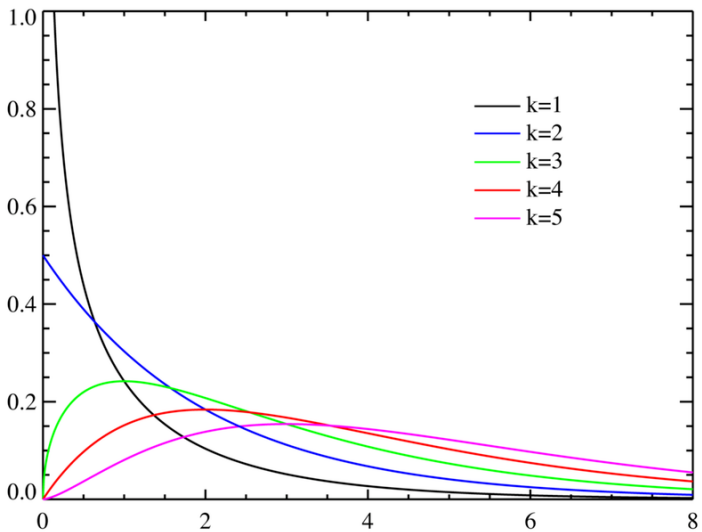


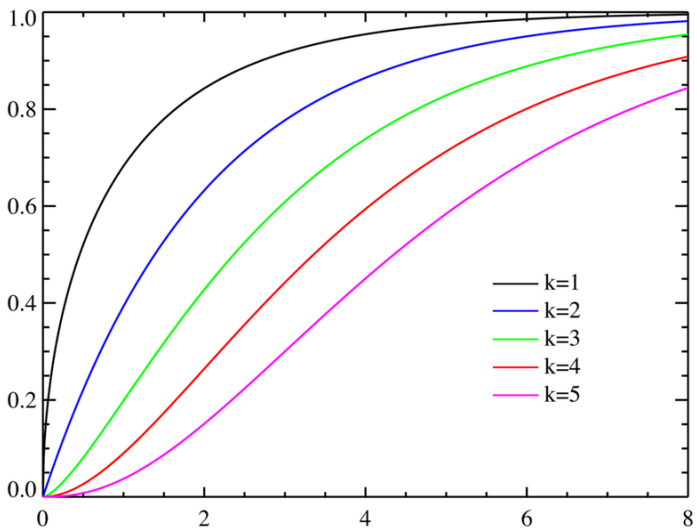
Função Densidade de Probabilidade Chi-Quadrado

Teorema

Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com $Z_i \sim N(\mu, \sigma), i = 1, 2, \ldots, n$. Do teorema anterior segue-se que se

$$X = \sum_{i=1}^{n} \left(\frac{Z_i - \mu}{\sigma} \right), \text{ então } X \sim \chi^2_n$$





Soma de Variáveis Aleatórias Normais Independentes

Meta: descobrir a média e a variância de uma amostra aleatória retirada de uma população normal.

Teorema

Se $X_1, X_2, \ldots, X_n$ são variáveis aleatórias normais independentes com médias $\mu_1, \mu_2, \ldots, \mu_n$ e variâncias $\sigma_1^2, \sigma_1^2, \ldots, \sigma_n^2$, então a combinação linear

$$Y = \sum_{i=1}^{n} c_i X_i$$

tem distribuição normal

$$N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right).$$

Demonstração

Depende da prova (não trivial) que a função geradora de momentos de variável aleatória normal $X \sim N(\mu, \sigma^2)$ é

$$M_X(s) = exp\left(\mu s + \frac{\sigma^2 s^2}{2}\right)$$

e, portanto,

$$M_Y(s) = exp\left[s\left(\sum_{i=1}^{n} c_i \mu_i\right) + \frac{s^2}{2}\left(\sum_{i=1}^{n} c_i^2 \sigma_i^2\right)\right]$$

Corolário

Se $X_1, X_2, \ldots, X_n$ são observações de uma amostra aleatória de tamanho n de uma população $N(\mu, \sigma^2)$, então a média

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

é normalmente distribuída com média μ e variância σ^2/n, isto é, a probabilidade da média amostrada é $N(\mu, \sigma^2/n)$.

Demonstração

Faça $c_i = 1/n, \mu_i = \mu, \sigma_i^2 = \sigma^2$ no teorema anterior. Então

$$M_{\bar{X}}(s) = exp\left[s\left(\frac{1}{n}\sum_{i=1}^{n}\mu\right) + \frac{s^2}{2}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right)\right] = exp\left[\mu s + \frac{s^2}{2}\left(\frac{\sigma^2}{n}\right)\right]$$

Corolário

Se $X_1, X_2, \ldots, X_n$ são observações de uma amostra aleatória de tamanho n de uma população $N(\mu, \sigma^2)$, então a média

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

é normalmente distribuída com média μ e variância σ^2/n, isto é, a probabilidade da média amostrada é $N(\mu, \sigma^2/n)$.

Demonstração

Faça $c_i = 1/n, \mu_i = \mu, \sigma_i^2 = \sigma^2$ no teorema anterior. Então

$$M_{\bar{X}}(s) = exp\left[s\left(\frac{1}{n}\sum_{i=1}^{n}\mu\right) + \frac{s^2}{2}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right)\right] = exp\left[\mu s + \frac{s^2}{2}\left(\frac{\sigma^2}{n}\right)\right]$$

Distribuição da Variância de uma Amostra Aleatória de uma População Normal

Teorema

Suponha que

- $X_1, X_2, \ldots, X_n$ *são observações de uma amostra aleatória de tamanho n de uma população* $N(\mu, \sigma^2)$.
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *é a média amostral das n observações.*
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *é a variância amostral das n observações.*

Então

(1) $\bar{X}$ *e* S^2 *são independentes.*

(2) $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.

Demonstração

Não trivial

Distribuição da Variância de uma Amostra Aleatória de uma População Normal

Teorema

Suponha que

- $X_1, X_2, \ldots, X_n$ *são observações de uma amostra aleatória de tamanho n de uma população* $N(\mu, \sigma^2)$.
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *é a média amostral das n observações.*
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *é a variância amostral das n observações.*

Então

(1) $\bar{X}$ *e* S^2 *são independentes.*

(2) $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.

Demonstração

Não trivial

Distribuição *t*-Student

Foi proposta pelo engenheiro alemão Friedrich Robert Helmert e pelo matemático alemão Jacob Lüroth em 1876.

- Na literatura inglesa a distribuição toma seu nome do artigo publicado em 1908 na revista Biometrika por William Sealy Gosset sob o pseudônimo ”Student”.
- Gosset trabalhava na Guinnes Brewery! que estava interessada em problemas de amostras pequenas, por exemplo, as propriedades químicas de cevada em que os tamanhos das amostras podiam ser tão pequenas como 3.
- O artigo ficou conhecido pelo trabalho de Ronald A. Fisher, que cunhou distribuição de $t-$Student com referência ao valor de t (proveniente de teste).
- Esta distribuição aparece na estimação da média de uma população normalmente distribuída em situações em que o tamanho da amostra é pequeno e a variancia é desconhecida.
- É usada em diversas análises estatísticas, incluindo o teste de *t*-Student para avaliar a significância estatística da diferença entre as médias de duas amostras.

Distribuição *t*-Student

Foi proposta pelo engenheiro alemão Friedrich Robert Helmert e pelo matemático alemão Jacob Lüroth em 1876.

- Na literatura inglesa a distribuição toma seu nome do artigo publicado em 1908 na revista Biometrika por William Sealy Gosset sob o pseudônimo ”Student”.
- Gosset trabalhava na Guinnes Brewery! que estava interessada em problemas de amostras pequenas, por exemplo, as propriedades químicas de cevada em que os tamanhos das amostras podiam ser tão pequenas como 3.
- O artigo ficou conhecido pelo trabalho de Ronald A. Fisher, que cunhou distribuição de $t-$Student com referência ao valor de t (proveniente de teste).
- Esta distribuição aparece na estimação da média de uma população normalmente distribuída em situações em que o tamanho da amostra é pequeno e a variancia é desconhecida.
- É usada em diversas análises estatísticas, incluindo o teste de *t*-Student para avaliar a significância estatística da diferença entre as médias de duas amostras.

Distribuição de t-Student como Teste Estatístico

Definição

Se $Z \sim N(0,1)$ e $U \sim \chi^2(n)$ são independentes, então a variável aleatória

$$T = \frac{Z}{\sqrt{U/n}}$$

tem uma distribuição de Student com n graus de liberdade, e escreve-se $T \sim t(n)$.

- Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma distribuição normal, sabemos que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

e

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}$$

- Além disso, Z e U são independentes. Portanto, da definição da variável T

$$T = \frac{\frac{\bar{X}-\mu}{\sigma\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/n-1}}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Distribuição de t-Student como Teste Estatístico

Definição

Se $Z \sim N(0,1)$ e $U \sim \chi^2(n)$ são independentes, então a variável aleatória

$$T = \frac{Z}{\sqrt{U/n}}$$

tem uma distribuição de Student com n graus de liberdade, e escreve-se $T \sim t(n)$.

- Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma distribuição normal, sabemos que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

e

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}$$

- Além disso, Z e U são independentes. Portanto, da definição da variável T

$$T = \frac{\frac{\bar{X}-\mu}{\sigma\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/n-1}}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Distribuição *t*-Student

- Considere a função densidade de probabilidade

$$f_X(x) = \int_0^\infty f_{X|\psi}(x) f_\Psi(\psi) d\psi$$

- $f_{X|\psi}(x)$ é a função densidade de probabilidade normal com média 0 e variança (inversa) desconhecida $\psi = \frac{1}{\sigma}$.

$$f_{X|\psi}(x) = \left(\frac{\psi}{2\pi}\right)^{1/2} exp\left(-\frac{\psi}{2}(x-\mu)^2\right)$$

- $f_\Psi(\psi)$ é a função densidade de probabilidade gama

$$f_\Psi(\psi) = \frac{\psi^{\alpha-1} e^{\frac{-\psi}{\beta}}}{\beta^\alpha \Gamma(\alpha)}$$

- Demonstra-se que a função densidade de probabilidade de Student é dada por

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Distribuição *t*-Student

- Considere a função densidade de probabilidade

$$f_X(x) = \int_0^\infty f_{X|\psi}(x) f_\Psi(\psi) d\psi$$

- $f_{X|\psi}(x)$ é a função densidade de probabilidade normal com média 0 e variança (inversa) desconhecida $\psi = \frac{1}{\sigma}$.

$$f_{X|\psi}(x) = \left(\frac{\psi}{2\pi}\right)^{1/2} exp\left(-\frac{\psi}{2}(x-\mu)^2\right)$$

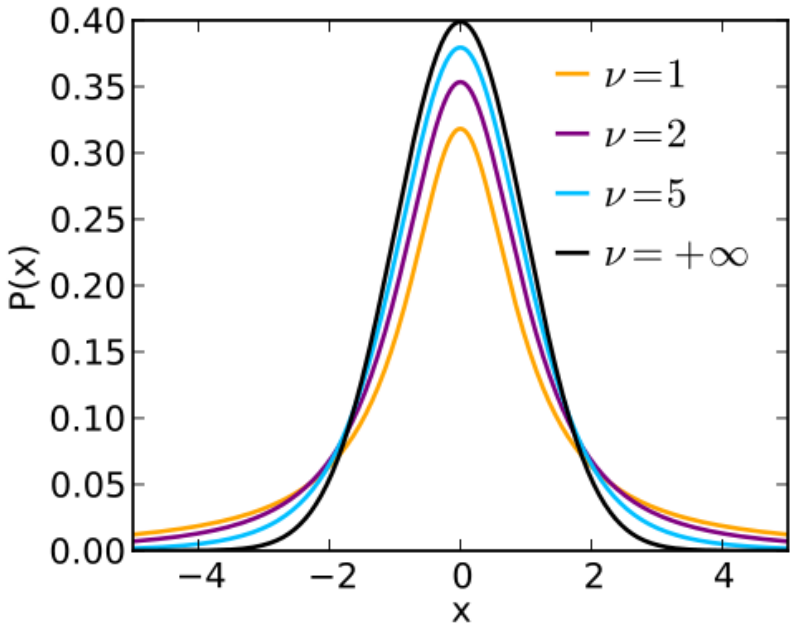
- $f_\Psi(\psi)$ é a função densidade de probabilidade gama

$$f_\Psi(\psi) = \frac{\psi^{\alpha-1} e^{\frac{-\psi}{\beta}}}{\beta^\alpha \Gamma(\alpha)}$$

- Demonstra-se que a função densidade de probabilidade de Student é dada por

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Distribuição *t*-Student



Teste de Hipótese - Exemplo

- Engenheiro considera a formulação de adicionar um polímero (borracha) derivado da emulsão de latex durante a mistura de cimento para testar o impacto no tempo de cura e resistência à força de tensão do cimento.
- São preparadas 10 amostras da formulação do cimento na forma original e 10 amostras na formulação modificada.
- As formulações são referidas como tratamentos ou como dois níveis de um fator.
- Técnica de inferência estatística chamada teste de hipótese é usada para comparar as duas formulações.

Teste de Hipótese - Exemplo

- Engenheiro considera a formulação de adicionar um polímero (borracha) derivado da emulsão de latex durante a mistura de cimento para testar o impacto no tempo de cura e resistência à força de tensão do cimento.
- São preparadas 10 amostras da formulação do cimento na forma original e 10 amostras na formulação modificada.
- As formulações são referidas como tratamentos ou como dois níveis de um fator.
- Técnica de inferência estatística chamada teste de hipótese é usada para comparar as duas formulações.

- Impressão visual na tabela abaixo é que que o cimento original suporta uma força maior que o modificado, o que é ”confirmado” pelos valores médios de força.

	Cimento Modificado	Cimento Original
j	y_{1j}	y_{2j}
1	16,85	16,62
2	16,40	16,75
3	17,21	17,37
4	16,35	17,12
5	16,52	16,98
6	17,04	16,87
7	16,96	17,34
8	17,04	17,02
9	16,59	17,08
10	16,57	17,27

- Valores médios de resistência à força de tensão:

$$\bar{y}_1 = 16,76 kgf/cm^2 \quad \bar{y}_2 = 17,04 kgf/cm^2$$

- Impressão visual na tabela abaixo é que que o cimento original suporta uma força maior que o modificado, o que é ”confirmado” pelos valores médios de força.

	Cimento Modificado	Cimento Original
j	y_{1j}	y_{2j}
1	16,85	16,62
2	16,40	16,75
3	17,21	17,37
4	16,35	17,12
5	16,52	16,98
6	17,04	16,87
7	16,96	17,34
8	17,04	17,02
9	16,59	17,08
10	16,57	17,27

- Valores médios de resistência à força de tensão:

$$\bar{y}_1 = 16,76 kgf/cm^2 \quad \bar{y}_2 = 17,04 kgf/cm^2$$

Teste de Hipótese

- $y_{11}, y_{12}, \ldots, y_{1n_1}$ representam n_1 observações do primeiro nível do fator.
- $y_{21}, y_{22}, \ldots, y_{2n_2}$ representam n_2 observações do segundo nível do fator.
- As duas amostras são retiradas de duas populações normais independentes.
- Modelo simples estatístico para o experimento:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2; \quad j = 1, \ldots, n$$

- y_{ij}: $j-$ésima observação associada ao nível i do fator.
- μ_i: média das observações associada ao nível i do fator.
- ϵ_{ij}: variável aleatória normal, chamada erro aleatório, associada com a $ij-$ésima observação, isto é $y_{ij} \sim N(\mu_i), \sigma_i^2, i = 1, 2$.

Teste de Hipótese

- $y_{11}, y_{12}, \ldots, y_{1n_1}$ representam n_1 observações do primeiro nível do fator.
- $y_{21}, y_{22}, \ldots, y_{2n_2}$ representam n_2 observações do segundo nível do fator.
- As duas amostras são retiradas de duas populações normais independentes.
- Modelo simples estatístico para o experimento:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2; \quad j = 1, \ldots, n$$

- y_{ij}: $j-$ésima observação associada ao nível i do fator.
- μ_i: média das observações associada ao nível i do fator.
- ϵ_{ij}: variável aleatória normal, chamada erro aleatório, associada com a $ij-$ésima observação, isto é $y_{ij} \sim N(\mu_i), \sigma_i^2, i = 1, 2$.

Hipótese Estatística

- Corresponde a uma afirmação (conjectura) sobre os parâmetros de uma distribuição de probabilidade ou parâmetros de um modelo.
- Para o exemplo do cimento, considere a tensão média e as hipóteses
- $H_0 : \mu_1 = \mu_2$, chamada hipótese nula.
- $H_1 : \mu_1 \neq \mu_2$, chamada hipótese alternativa de dois lados, pois é verdadeira para $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$.
- O procedimento geral do teste de hipótese é especificar o valor $\alpha = P(\text{ rejeitar} H_0 | H_0)$ é verdadeiro), chamado nível de significância.

Hipótese Estatística

- Corresponde a uma afirmação (conjectura) sobre os parâmetros de uma distribuição de probabilidade ou parâmetros de um modelo.
- Para o exemplo do cimento, considere a tensão média e as hipóteses
- $H_0 : \mu_1 = \mu_2$, chamada hipótese nula.
- $H_1 : \mu_1 \neq \mu_2$, chamada hipótese alternativa de dois lados, pois é verdadeira para $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$.
- O procedimento geral do teste de hipótese é especificar o valor $\alpha = P(\text{ rejeitar} H_0 | H_0)$ é verdadeiro), chamado nível de significância.

Teste-*t* Student para Duas Amostras

- Suponha que as variâncias são idênticas para as duas distribuições normais, $\sigma_1^2 = \sigma_2^2$.
- As duas amostras são extraídas de duas distribuições normais independentes.
- Portanto, a distribuição de $\bar{y}_1 - \bar{y}_2$ é $N\left[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\right]$.
- Se σ^2 fosse conhecido, e se H_0 fosse verdadeiro, então a distribuição

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

  seria $N(0, 1)$.

- No entanto, ao substituir σ em (2) pela sua estimativa

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Teste-*t* Student para Duas Amostras

- Suponha que as variâncias são idênticas para as duas distribuições normais, $\sigma_1^2 = \sigma_2^2$.
- As duas amostras são extraídas de duas distribuições normais independentes.
- Portanto, a distribuição de $\bar{y}_1 - \bar{y}_2$ é $N\left[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\right]$.
- Se σ^2 fosse conhecido, e se H_0 fosse verdadeiro, então a distribuição

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

  seria $N(0, 1)$.

- No entanto, ao substituir σ em (2) pela sua estimativa

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Teste-t para Duas Amostras

- a distribuição de Z_0 muda para a estatística $t-$Student

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- com $n_1 - n_2 - 2$ graus de liberdade, devido à informação adicional que $\sigma_1^2 = \sigma_2^2$.
- Em alguns problemas, deseja-se rejeitar H_0 se uma média é maior que a outra.
- A hipótese alternativa de um lado $H_1 : \mu_1 > \mu_2$ e H_0 é rejeitada se $t_0 > t_{\alpha, n_1+n_2-2}$.
- A hipótese alternativa de um lado $H_1 : \mu_1 < \mu_2$ e H_0 é rejeitada se $t_0 < -t_{\alpha, n_1+n_2-2}$.

Teste-t para Duas Amostras

- a distribuição de Z_0 muda para a estatística $t-$Student

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- com $n_1 - n_2 - 2$ graus de liberdade, devido à informação adicional que $\sigma_1^2 = \sigma_2^2$.

- Em alguns problemas, deseja-se rejeitar H_0 se uma média é maior que a outra.

- A hipótese alternativa de um lado $H_1 : \mu_1 > \mu_2$ e H_0 é rejeitada se $t_0 > t_{\alpha, n_1+n_2-2}$.

- A hipótese alternativa de um lado $H_1 : \mu_1 < \mu_2$ e H_0 é rejeitada se $t_0 < -t_{\alpha, n_1+n_2-2}$.

Teste-t para as Duas Amostras do Exemplo

- Para o exemplo do cimento temos

Cimento Modificado	Cimento Original
$\bar{y}_1 = 16,76 kgf/cm^2$	$\bar{y}_2 = 17,04 kgf/cm^2$
$S_1^2 = 0,100$	$S_2^2 = 0,061$
$S_1 = 0,316$	$S_2 = 0,248$
$n_1 = 10$	$n_2 = 10$

- Desvios padrão são semelhantes e, portanto, a hipótese de igualdade das variâncias é razoável.

- $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, e ao se escolher $\alpha = 0,05$, rejeitamos H_0 se $t_0 > t_{0,025,18} = 2,101$ ou se $t_0 < -t_{0,025,18} = -2,101$.

Teste-t para as Duas Amostras do Exemplo

- Para o exemplo do cimento temos

Cimento Modificado	Cimento Original
$\bar{y}_1 = 16,76 kgf/cm^2$	$\bar{y}_2 = 17,04 kgf/cm^2$
$S_1^2 = 0,100$	$S_2^2 = 0,061$
$S_1 = 0,316$	$S_2 = 0,248$
$n_1 = 10$	$n_2 = 10$

- Desvios padrão são semelhantes e, portanto, a hipótese de igualdade das variâncias é razoável.

- $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, e ao se escolher $\alpha = 0,05$, rejeitamos H_0 se $t_0 > t_{0,025,18} = 2,101$ ou se $t_0 < -t_{0,025,18} = -2,101$.

Teste-t para as Duas Amostras do Exemplo

- Para o exemplo do cimento temos

Cimento Modificado	Cimento Original
$\bar{y}_1 = 16,76 kgf/cm^2$	$\bar{y}_2 = 17,04 kgf/cm^2$
$S_1^2 = 0,100$	$S_2^2 = 0,061$
$S_1 = 0,316$	$S_2 = 0,248$
$n_1 = 10$	$n_2 = 10$

- Desvios padrão são semelhantes e, portanto, a hipótese de igualdade das variâncias é razoável.

- $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, e ao se escolher $\alpha = 0,05$, rejeitamos H_0 se $t_0 > t_{0,025,18} = 2,101$ ou se $t_0 < -t_{0,025,18} = -2,101$.

Teste-t para as Duas Amostras do Exemplo

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$= \frac{9(0,100) + 9(0,061)}{10 + 10 - 2} = 0,081$$

$$S_p = 0,284$$

- Estatística do teste:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{16,76 - 17,04}{0,284\sqrt{\frac{1}{10} + \frac{1}{10}}} = -2,20$$

- Como $t_0 = -2,20 < -t_{0,025,18} = -2,101$, rejeita-se H_0 e conclui-se que $\mu_1 < \mu_2$.

- Isto é confirmado na prática: tempo de cura do cimento modificado é menor, mas sua resistência à força de tensão é menor.

Teste-t para as Duas Amostras do Exemplo

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$= \frac{9(0,100) + 9(0,061)}{10 + 10 - 2} = 0,081$$

$$S_p = 0,284$$

- Estatística do teste:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{16,76 - 17,04}{0,284\sqrt{\frac{1}{10} + \frac{1}{10}}} = -2,20$$

- Como $t_0 = -2,20 < -t_{0,025,18} = -2,101$, rejeita-se H_0 e conclui-se que $\mu_1 < \mu_2$.

- Isto é confirmado na prática: tempo de cura do cimento modificado é menor, mas sua resistência à força de tensão é menor.

Valor-P em Teste de Hipótese

- Hipótese nula H_0 é ou não rejeitada em um nível especificado de valor-α ou nível de significância fixo.
- Valor calculado da estatística do teste está próximo ou distante do limiar da região de rejeição?
- Enfoque do valor-P : probabilidade que o valor da estatística do teste assuma o menor nível de significância, quando a hipótese nula é verdadeira.
- Como $|t_0| = 2,20 > t_{0.025,18} = 2,101$, então o valor-$P$ é < 0.05.
- Como $t_{0.01,18} = 2,552 > |t_0| = 2,20$, então o valor-$P$ está entre 0.05 e $2(0,01) = 0,02$.
- Softwares de estatística fornecem o valor-P exato ou na forma $< 0,001$. Para este exemplo, o software Minitab fornece o valor-P igual a 0.042.

Valor-P em Teste de Hipótese

- Hipótese nula H_0 é ou não rejeitada em um nível especificado de valor-α ou nível de significância fixo.
- Valor calculado da estatística do teste está próximo ou distante do limiar da região de rejeição?
- Enfoque do valor-P : probabilidade que o valor da estatística do teste assuma o menor nível de significância, quando a hipótese nula é verdadeira.
- Como $|t_0| = 2,20 > t_{0.025,18} = 2,101$, então o valor-$P$ é < 0.05.
- Como $t_{0.01,18} = 2,552 > |t_0| = 2,20$, então o valor-$P$ está entre 0.05 e $2(0,01) = 0,02$.
- Softwares de estatística fornecem o valor-P exato ou na forma $< 0,001$. Para este exemplo, o software Minitab fornece o valor-P igual a 0.042.

Intervalo de Confiança

- Em diversas situações, o teste nulo de hipótese de médias $\mu_1 = \mu_2$ é de pouco interesse.
- Mais importante: de quanto diferem? Resposta: intervalo de confiança.
- Seja θ um parâmetro desconhecido. Encontre duas estatísticas L e U tal que

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{3}$$

é verdadeiro.

- O intervalo $L \leq \theta \leq U$ é chamado um intervalo de $100(1-\alpha)\%$ de confiança para o parâmetro θ.
- O intervalo de confiança tem um interpretação frequentista, isto é, o método usado para produzir o intervalo de confiança gera afirmativas corretas $100(1-\alpha)\%$ das vezes.

Intervalo de Confiança

- Em diversas situações, o teste nulo de hipótese de médias $\mu_1 = \mu_2$ é de pouco interesse.
- Mais importante: de quanto diferem? Resposta: intervalo de confiança.
- Seja θ um parâmetro desconhecido. Encontre duas estatísticas L e U tal que

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{3}$$

  é verdadeiro.

- O intervalo $L \leq \theta \leq U$ é chamado um intervalo de $100(1-\alpha)\%$ de confiança para o parâmetro θ.
- O intervalo de confiança tem um interpretação frequentista, isto é, o método usado para produzir o intervalo de confiança gera afirmativas corretas $100(1-\alpha)\%$ das vezes.

Intervalo de Confiança

- A estatística

$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$P\left(-t_{\alpha/2,n_1+n_2-2} \leq \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2}\right) = 1 - \alpha$$

$$P\left(\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2\right.$$

$$\left. \leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha \qquad (4)$$

- Comparando (3) e (4)

Intervalo de Confiança

$$\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2$$

$$\leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- é um intervalo de confiança 100(α)% para $\mu_1 - \mu_2$.
- Para o exemplo do cimento com uma estimativa de 95% para o intervalo de confiança

$$16,76 - 17,04 - (2,10)0,284\sqrt{\frac{1}{10} + \frac{1}{10}} \leq \mu_1 - \mu_2 \leq 16,76 - 17,04 + (2,10)0,284\sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$-0,55 \leq \mu_1 - \mu_2 \leq -0,101$$

- o intervalo de confiança 95% é $\mu_1 - \mu_2 = -0,28 \pm 0,27$, ou a diferença na média é -0,28 e a precisão da estimativa é $\pm 0,27$.
- Como $\mu_1 - \mu_2 = 0$ não está incluído no intervalo de confiança, os dados não suportam a hipótese que $\mu_1 = \mu_2$ no nível de significância 0.05.

Intervalo de Confiança

$$\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2$$

$$\leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- é um intervalo de confiança $100(\alpha)\%$ para $\mu_1 - \mu_2$.
- Para o exemplo do cimento com uma estimativa de 95% para o intervalo de confiança

$$16,76 - 17,04 - (2,10)0,284\sqrt{\frac{1}{10} + \frac{1}{10}} \leq \mu_1 - \mu_2 \leq 16,76 - 17,04 + (2,10)0,284\sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$-0,55 \leq \mu_1 - \mu_2 \leq -0,101$$

- o intervalo de confiança 95% é $\mu_1 - \mu_2 = -0,28 \pm 0,27$, ou a diferença na média é -0,28 e a precisão da estimativa é $\pm 0,27$.
- Como $\mu_1 - \mu_2 = 0$ não está incluído no intervalo de confiança, os dados não suportam a hipótese que $\mu_1 = \mu_2$ no nível de significância 0.05.

Teste-*t* para Duas Amostras: Tópicos Adicionais

- Desvios moderado da hipótese de distribuição normal das duas populações não afeta muito o desempenho do teste-*t*.
- Escolha do tamanho da amostra.
- O caso $\sigma_1^2 \neq \sigma_2^2$.
- O caso σ_1^2 e σ_2^2 conhecidos.

Distribuição-F

- A distribuição-F é conhecida como distribuição de Snedecor ou distribuição de Fisher-Snedecor.
- Aparece frequentemente como uma distribuição nula de um teste estatístico, principamente na análise das variâncias de duas amostras de duas populações com distribuição normal.

Definição

Sejam U e V duas variáveis aleatórias independentes com distribuição chi-quadrado com m e n graus de liberdade, respectivamente. Então a variável aleatória $X = \frac{U/m}{V/n}$ define a estatística $F_{m,n}$.

Proposição

A função densidade associada à estatística $X = \frac{U/m}{V/n}$ é dada por

$$f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2} x^{(m/2)-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)(n+mx)^{(m+n)/2}}, \quad x > 0$$

A demonstração é complexa.

Distribuição-F

- A distribuição-F é conhecida como distribuição de Snedecor ou distribuição de Fisher-Snedecor.
- Aparece frequentemente como uma distribuição nula de um teste estatístico, principamente na análise das variâncias de duas amostras de duas populações com distribuição normal.

Definição

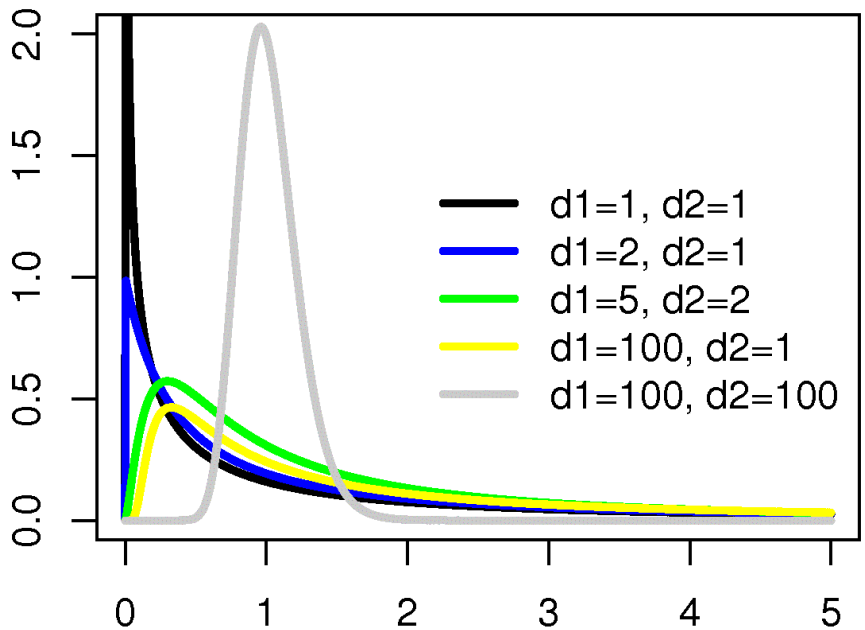
Sejam U e V duas variáveis aleatórias independentes com distribuição chi-quadrado com m e n graus de liberdade, respectivamente. Então a variável aleatória $X = \frac{U/m}{V/n}$ define a estatística $F_{m,n}$.

Proposição

A função densidade associada à estatística $X = \frac{U/m}{V/n}$ é dada por

$$f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2} x^{(m/2)-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)(n+mx)^{(m+n)/2}}, \quad x > 0$$

A demonstração é complexa.


0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

Variâncias de Duas Populações Normais

- Meta: comparar variâncias de duas populações normais com variâncias σ_1^2 e σ_2^2, respectivamente.
- Tome uma amostra de tamanho n_1 de uma população e outra amostra de tamanho n_2 da outra.

Proposição

$s_1^2/s_2^2 \sim F(n_1 - 1, n_2 - 1)$, se $\sigma_1^2 = \sigma_2^2$.

Demonstração

$$W_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

$$W_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

$$F = \frac{W_1/n_1 - 1}{W_2/n_2 - 1} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/n_1 - 1}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/n_2 - 1} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

Variâncias de Duas Populações Normais

- Meta: comparar variâncias de duas populações normais com variâncias σ_1^2 e σ_2^2, respectivamente.
- Tome uma amostra de tamanho n_1 de uma população e outra amostra de tamanho n_2 da outra.

Proposição

$s_1^2/s_2^2 \sim F(n_1 - 1, n_2 - 1)$, se $\sigma_1^2 = \sigma_2^2$.

Demonstração

$$W_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

$$W_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

$$F = \frac{W_1/n_1 - 1}{W_2/n_2 - 1} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/n_1 - 1}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/n_2 - 1} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

Teste Chi-Quadrado de Qualidade de Ajuste

- Considere r caixas e jogue n bolas $X_1, \ldots, X_n$ nas caixas independentemente entre estas com probabilidades

$$P_{X_i} \in B_i = p_i, \quad i = 1, \ldots, r$$

$$p_1 + \cdots + p_r = 1$$

- Seja $\nu_j =$ número de bolas $\{ X_1, \ldots, X_n$ na Caixa $B_j\} = \sum_{l=1}^{n} I(X_l \in B_j)$ (I:função indicadora)

$$E\nu_j = \sum_{l=1}^{n} I(X_l \in B_j) = \sum_{l=1}^{n} P(X_l \in B_j) = np_j$$

- A dificuldade na demonstração do teorema de Pearson (1900) a seguir, é que as variáveis ν_j não são independentes, pois

$$\nu_1 + \cdots + \nu_r = n$$

Teorema (Pearson)

A variável aleatória

$$\sum_{j=1}^{r} = \frac{(\nu_j - np_j)^2}{np_j} \to^d \chi^2_{r-1}$$

converge em distribuição para a distribuição χ^2_{r-1} com $r-1$ graus de liberdade.

Teste Chi-Quadrado de Qualidade de Ajuste

- Considere r caixas e jogue n bolas $X_1, \ldots, X_n$ nas caixas independentemente entre estas com probabilidades

$$P_{X_i} \in B_i = p_i, \quad i = 1, \ldots, r$$

$$p_1 + \cdots + p_r = 1$$

- Seja $\nu_j =$ número de bolas $\{ X_1, \ldots, X_n$ na Caixa $B_j\} = \sum_{l=1}^{n} I(X_l \in B_j)$ (I:função indicadora)

$$E\nu_j = \sum_{l=1}^{n} I(X_l \in B_j) = \sum_{l=1}^{n} P(X_l \in B_j) = np_j$$

- A dificuldade na demonstração do teorema de Pearson (1900) a seguir, é que as variáveis ν_j não são independentes, pois

$$\nu_1 + \cdots + \nu_r = n$$

Teorema (Pearson)

A variável aleatória

$$\sum_{j=1}^{r} = \frac{(\nu_j - np_j)^2}{np_j} \to^d \chi^2_{r-1}$$

converge em distribuição para a distribuição χ^2_{r-1} com $r-1$ graus de liberdade.

Teste Chi-Quadrado de Qualidade de Ajuste

- Considere r caixas e jogue n bolas $X_1, \ldots, X_n$ nas caixas independentemente entre estas com probabilidades

$$P_{X_i} \in B_i = p_i, \quad i = 1, \ldots, r$$

$$p_1 + \cdots + p_r = 1$$

- Seja $\nu_j =$ número de bolas $\{ X_1, \ldots, X_n$ na Caixa $B_j\} = \sum_{l=1}^{n} I(X_l \in B_j)$ (I:função indicadora)

$$E\nu_j = \sum_{l=1}^{n} I(X_l \in B_j) = \sum_{l=1}^{n} P(X_l \in B_j) = np_j$$

- A dificuldade na demonstração do teorema de Pearson (1900) a seguir, é que as variáveis ν_j não são independentes, pois

$$\nu_1 + \cdots + \nu_r = n$$

Teorema (Pearson)

A variável aleatória

$$\sum_{j=1}^{r} = \frac{(\nu_j - np_j)^2}{np_j} \rightarrow^d \chi^2_{r-1}$$

converge em distribuição para a distribuição χ^2_{r-1} com $r-1$ graus de liberdade.

Exemplo do *chi-squared goodness-of-fit test*

- Usar o teste para verificar se os dados ajustam-se a uma distribuição contínua $F_0(w)$ conhecida, isto é,

$$H_0 : F(w) = F_0(w)$$

- Divida o intervalo contínuo em k categorias (caixas), chamadas $A_1, \ldots, A_k$ em que os dados observados podem situar-se.
- $X_i =$ número de vezes que um valor observado de W pertence à categoria A_i com probabilidade p_i.
- O teste de hipótese acima é modificado para

$$H_0' : p_i = p_{i0}, \quad i = 1, \ldots, k$$

- A hipótese é rejeitada se o valor do teste chi-quadrado

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

é maior ou igual a $\chi^2_{k-1,\alpha}$.

- O_i e E_i representam, o valor observado e o valor esperado de X_i, respectivamente.
- Se a hipótese $H_0' : p_i = p_{i0}, \ i = 1, \ldots, k$ não é rejeitada, então a hipótese original $H_0 : F(w) = F_0(w)$ não é rejeitada.

Exemplo do *chi-squared goodness-of-fit test*

- Usar o teste para verificar se os dados ajustam-se a uma distribuição contínua $F_0(w)$ conhecida, isto é,

$$H_0 : F(w) = F_0(w)$$

- Divida o intervalo contínuo em k categorias (caixas), chamadas $A_1, \ldots, A_k$ em que os dados observados podem situar-se.
- $X_i =$ número de vezes que um valor observado de W pertence à categoria A_i com probabilidade p_i.
- O teste de hipótese acima é modificado para

$$H_0' : p_i = p_{i0}, \quad i = 1, \ldots, k$$

- A hipótese é rejeitada se o valor do teste chi-quadrado

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

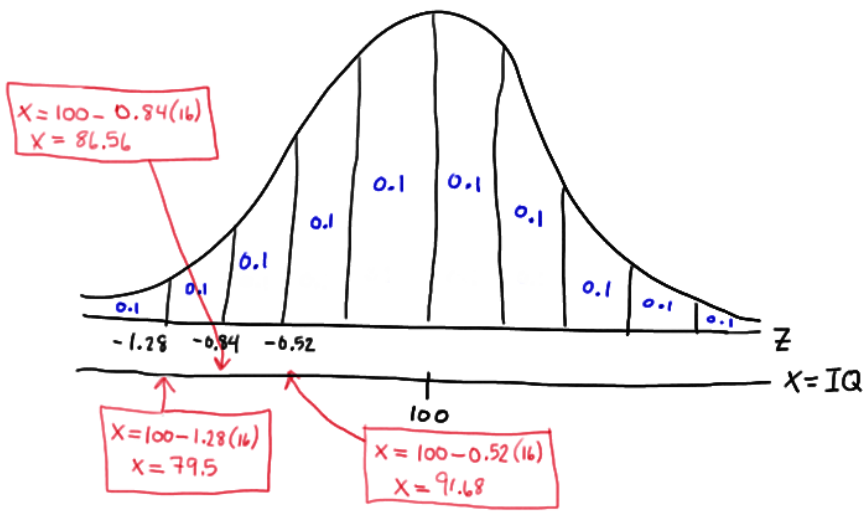
é maior ou igual a $\chi^2_{k-1,\alpha}$.

- O_i e E_i representam, o valor observado e o valor esperado de X_i, respectivamente.
- Se a hipótese $H_0' : p_i = p_{i0},\ i = 1, \ldots, k$ não é rejeitada, então a hipótese original $H_0 : F(w) = F_0(w)$ não é rejeitada.

Exemplo do *chi-squared goodness-of-fit test*

- São dados os QI's de 100 pessoas. Teste a hipótese nula que os dados provêm de uma distribuição normal X com média $E[X] = 100$ e desvio padrão $\sigma_X = 16$.
- Defina as categorias pela divisão dos QI's em $k = 10$ conjuntos com igual probabilidade $1/k = 1/10$.
- A tabela mostra os QI's e a figura ilustra os $k = 10$ intervalo com igual probabilidade. Além disso, é mostrado os valores Z associados aos QI's que correspondem às $k = 10$ probabilidades cumulativas $0,1; 0,2; 0,3$; etc., e os valores $X = E[X] + \sigma_X Z$.

54	66	74	74	75	78	79	80	81	82
82	82	83	84	87	88	88	88	88	89
89	89	89	89	90	90	90	91	92	93
93	93	94	96	96	97	97	98	98	99
99	99	99	99	100	100	100	102	102	102
102	102	103	103	104	104	104	105	105	105
105	106	106	106	107	107	108	108	108	109
109	109	110	111	111	111	111	112	112	112
114	114	115	115	115	116	118	118	120	121
121	122	123	125	126	127	127	131	132	139



- O número de graus de liberdade é $10 - 1 - 2 = 7$, devido ao conhecimento da média e desvio padrão da distribuição normal.

- A tabela abaixo mostra as categorias associados a intervalos ou classes, valores observados, valor esperado $(100 \times 0,1)$ e a contribuição ao valor do teste Q_7.

$$Q_9 = 8,2 < \chi^2_{10-1;0,05} = \chi^2_{7;0,05} = 14,07$$

- A hipótese nula não é rejeitada no nível $0,05$.

Category	Class	Obs'd	Exp'd	Contribution to Q
1	$(-\infty, 79.5]$	7	10	$(7-10)^2/10 = 0.9$
2	$(79.5, 86.5]$	7	10	$(7-10)^2/10 = 0.9$
3	$(86.5, 91.6]$	14	10	$(14-10)^2/10 = 1.6$
4	$(91.6, 95.9]$	5	10	$(5-10)^2/10 = 2.5$
5	$(95.9, 100.0]$	14	10	$(14-10)^2/10 = 1.6$
6	$(100.0, 104.1]$	10	10	$(10-10)^2/10 = 0.0$
7	$(104.1, 108.4]$	12	10	$(12-10)^2/10 = 0.4$
8	$(108.4, 113.5]$	11	10	$(11-10)^2/10 = 0.1$
9	$(113.5, 120.5]$	9	10	$(9-10)^2/10 = 0.1$
10	$(120.5, \infty]$	11	10	$(11-10)^2/10 = 0.1$
		$n = 100$	$n = 100$	$Q_9 = 8.2$

Regras de Parada Probabilística para GRASP

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 20 (2013) 301–323
DOI: 10.1111/itor.12010

Probabilistic stopping rules for GRASP heuristics and extensions

Celso C. Ribeiro[a], Isabel Rosseti[a] and Reinaldo C. Souza[b]

[a]*Department of Computer Science, Universidade Federal Fluminense, Rua Passo da Pátria 156, Niterói, RJ 24210-240, Brazil*
[b]*Department of Electrical Engineering, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ 22453-900, Brazil*
E-mail: celso@ic.uff.br [Ribeiro]; rosseti@ic.uff.br [Rosseti]; reinaldo@ele.puc-rio.br [Souza]

Received 24 January 2013; received in revised form 28 January 2013; accepted 29 January 2013

Abstract

The main drawback of most metaheuristics is the absence of effective stopping criteria. Most implementations of such algorithms stop after performing a given maximum number of iterations or a given maximum number of consecutive iterations without improvement in the best-known solution value, or after the stabilization of the set of elite solutions found along the search. We propose effective probabilistic stopping rules for randomized metaheuristics such as GRASP (Greedy Randomized Adaptive Search Procedures). We show how the probability density function of the solution values obtained along the iterations of such algorithms can be used to implement stopping rules based on the tradeoff between solution quality and the time needed to find a solution that might improve the best solution found. We show experimentally that, in the particular case of GRASP heuristics, the solution values obtained along its iterations fit a normal distribution that may be used to give an online estimation of the number of solutions obtained in forthcoming iterations that might be at least as good as the incumbent. This estimation is used to validate the stopping rule based on the tradeoff between solution quality and the time needed to find a solution that might improve the incumbent. The robustness of this strategy is illustrated and validated by a thorough computational study reporting results obtained with GRASP implementations to four different combinatorial optimization problems.

Keywords: applied probability; artificial intelligence; combinatorial optimization; experimental results; heuristics; local search; metaheuristics

1. Introduction and motivation

Metaheuristics are general high-level procedures that coordinate simple heuristics and rules to find good approximate solutions to computationally difficult combinatorial optimization problems. Among them, we find simulated annealing, tabu search, GRASP, VNS (Variable Neighborhood

Greedy Randomized Adaptive Search Procedures - GRASP

Procedimento GRASP (Max-Iterações, Semente)
1. Faça $f* \leftarrow \infty$;
2. para $k = 1, \ldots,$ Max-Iterações faça
3. $x \leftarrow$ Greedy Randomized Construção(Semente);
4. $x \leftarrow$ Busca Local (Solução);
5. se $f(x) < f*$ então
6. $x* \leftarrow x$;
7. $f* \leftarrow f(x)$;
8. fim;
9. fim;
10. retorne $x*, f*$;

fim.

Fase Construtiva de GRASP

Procedimento Greedy-Randomized-Construction(Semente)
1. Solução $\leftarrow \emptyset$;
2. Avalie o custo incremental dos elementos candidatos;
3. enquanto Solução não está completa faça;
4. Construa a lista restrita de candidatos (LRC);
5. Selecione elemento s de LRC aleatoriamente;
6. Solução $\leftarrow$ Solução $\cup\{s\}$;
7. Reavalie os custos incrementais;
8. fim;
9. retorne Solução;

fim.

4 Instâncias de 4 Problemas Abordados

- The 2-path network design problem
- The $p-$median problem
- The quadratic assignment problem
- The set $k-$covering problem

Problema das $p-$ Medianas

Table 1
Test instances of the 2-path network design problem

Instance	$\|V\|$	$\|E\|$	K
2pndp50	50	1225	500
2pndp70	70	2415	700
2pndp90	90	4005	900
2pndp200	200	19,900	2000

problem consists in finding a minimum weighted subset of edges containing a path formed by at most two edges between every origin–destination pair. Applications can be found in the design of communication networks, in which paths with few edges are sought to enforce high reliability and small delays. Its decision version was proved to be NP-complete by Dahl and Johannessen (2004). The GRASP heuristic that has been used in the computational experiments with the 2-path network design problem was originally presented in Ribeiro and Rosseti (2002, 2007). The main characteristics of the four instances involved in the experiments are summarized in Table 1.

3.3. The p-median problem

Given a set F of m potential facilities, a set U of n customers, a distance function $d : U \times F \to \mathbb{R}$, and a constant $p \leq m$, the p-median problem consists in determining which p facilities to open so as to minimize the sum of the distances from each costumer to its closest open facility. It is a well-known NP-hard problem (Kariv and Hakimi, 1979), with numerous applications to location Tansel et al. (1983) and clustering (Rao, 1971; Vinod, 1969) problems. The GRASP heuristic that has been used in the computational experiments with the p-median problem was originally presented in Resende and Werneck (2004). The main characteristics of the four instances involved in the experiments are summarized in Table 2.

3.4. The quadratic assignment problem

Given n facilities and n locations represented, respectively, by the sets $F = \{f, \ldots, f_n\}$ and $L = \{\ell_1, \ldots, \ell_n\}$, the quadratic assignment problem proposed by Koopmans and Beckmann (1957) consists in determining to which location each facility must be assigned. Let $A^{n\times n} = (a_{ij})$ be a

Table 2
Test instances of the p-median problem

Instance	m	n	p
pmed10	200	800	67
pmed15	300	1800	100
pmed25	500	5000	167
pmed30	600	7200	200

Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

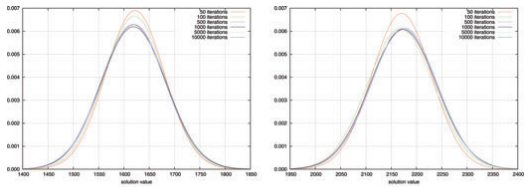
Aproximação da Normal por Iterações GRASP

- $f_1, \ldots, f_N$: amostra de valores de soluções obtidas por GRASP em N iterações.
- H_0 : a amostra $f_1, \ldots, f_N$ segue uma distribuição normal.
- H_1 : a amostra $f_1, \ldots, f_N$ não segue uma distribuição normal.
- m e S representam a média e desvio padrão da amostra.
- A amostra normalizada $f'_i = (f_i - m)/S$ implica em uma normal $N(0, 1)$.
- Nos experimentos, $\alpha = 0,1$ e número de intervalos igual a 14:

$$(-\infty, -3,0), [-3; -2,5], [-2,5; -2,0], \cdots, [2,0; 2,5], [2,5; 3,0], [3,0; \infty).$$

- Ajuste da normal é ilustrada para $N = 50, 100, 500, 1000, 5000$, e 10.000 iterações de GRASP.

Aproximação da Normal por Iterações GRASP



(a) Instance pmed10 with $p = 67$ (b) Instance pmed15 with $p = 100$

(c) Instance pmed25 with $p = 167$ (d) Instance pmed30 with $p = 200$

Fig. 3. Normal distributions: fitted probability density functions for the p-median problem.

Table 9
Chi-square test for $1 - \alpha = 90\%$ confidence level: quadratic assignment problem

Instance	Iterations	D	$\chi^2_{[1-\alpha;k-3]}$
tai30a	50	0.127260	17.275000
tai35a	50	0.213226	17.275000
tai40a	50	0.080164	17.275000
tai50a	50	0.075752	17.275000

fittings to the solution values obtained along the iterations of the GRASP heuristic for the p-median problem.

Results obtained with the GRASP heuristic for the quadratic assignment problem are reported in Tables 9 and 10 and in Fig. 4. The same statistics and plots provided for the previous problems lead to similar findings: they illustrate the robustness of the normal fittings to the solution values obtained along the iterations of the GRASP heuristic for the quadratic assignment problem.

Finally, we report in Tables 11 and 12 and in Fig. 5 the results obtained with the GRASP heuristic for the set k-covering problem. The same statistics and plots already given to the other problems show that also for the set k-covering problem the normal fittings to the solution values obtained along the iterations of the GRASP heuristic are very robust.

We conclude this section by observing that the null hypothesis cannot be rejected with $1 - \alpha = 90\%$ of confidence. Therefore, we may approximate the solution values obtained along N iterations of a GRASP heuristic by a normal distribution that can be progressively fitted and improved as

Aproximação da Normal por Iterações GRASP

Table 7
Chi-square test for $1 - \alpha = 90\%$ confidence level: p-median problem

Instance	Iterations	D	$\chi^2_{[1-\alpha;k-3]}$
pmed10	50	0.196116	17.275000
pmed15	50	0.167526	17.275000
pmed25	50	0.249443	17.275000
pmed30	50	0.160131	17.275000

Table 8
Statistics for normal fittings: p-median problem

Instance	Iterations	Mean	Standard deviation	Skewness	Kurtosis
	50	1622.020000	57.844097	−0.179163	3.255009
	100	1620.890000	59.932611	−0.364414	3.304588
pmed10	500	1620.332000	63.484721	0.111186	3.142248
$p = 67$	1000	1619.075000	64.402076	0.074091	2.964164
	5000	1617.875200	63.499795	0.043152	2.951273
	10,000	1618.415400	63.415181	0.087909	2.955408
	50	2170.500000	58.880642	−0.041262	1.949923
	100	2168.450000	65.313609	0.270892	2.693553
pmed15	500	2173.060000	65.881958	0.202400	2.828056
$p = 100$	1000	2173.484000	65.590272	0.129234	2.784433
	5000	2174.860000	64.639604	0.086450	2.940204
	10,000	2175.651600	65.101495	0.096328	2.954639
	50	2277.780000	54.782220	0.330959	3.028905
	100	2279.610000	58.034799	0.360133	3.466265
pmed25	500	2271.546000	56.029848	0.219415	3.311486
$p = 167$	1000	2274.182000	56.915366	0.081878	3.068963
	5000	2276.305200	56.985195	−0.041096	3.108109
	10,000	2277.151600	57.583524	−0.041570	3.073374
	50	2434.660000	57.809899	−0.130383	2.961249
	100	2446.560000	57.292464	−0.259531	2.667470
pmed30	500	2444.638000	56.109134	−0.189935	2.691882
$p = 200$	1000	2441.465000	57.265005	−0.053183	2.858399
	5000	2441.340400	54.941836	−0.013377	3.054188
	10,000	2441.277700	54.978827	0.006407	3.066879

Table 8 gives the main statistics for each instance of the p-median problem and for increasing values of the number $N = 50$, 100, 500, 1000, 5000, and 10,000 of GRASP iterations: mean, standard deviation, skewness, and kurtosis. As for the previous problem, we notice that the mean value converges or oscillates very slightly when the number of iterations increases. Furthermore, the mean after 50 iterations is already very close to that of the normal fitting after 10,000 iterations. Once again, the skewness values are consistently very close to 0, while the measured kurtosis of the sample is always close to 3.

Figure 3 displays the normal distribution fitted for each instance and for each number of iterations. Once again, the statistics and plots in these tables and figure illustrate the robustness of the normal

Estimativa da Melhor Solução

- Seja $UB^k = \min\{f_1, \ldots, f_k\}$, média m^k e desvio padrão S^k
- A aproximação da normal é dada por

$$f_X^k(x) = \frac{1}{S^k\sqrt{2\pi}} exp\left(\frac{-(x-m^k)^2}{2S_k^2}\right).$$

- Probalidade de de encontrar uma solução com valor menor ou igual a UB^k na próxima iteração :

$$F_X^k(UB^k) = \int_{-\infty}^{UB^k} f_X^k(\tau)d\tau$$

- Estimativa melhor com limitante inferior $\underline{I}$ para o valor de qualquer solução :

$$\hat{F}_X^k(UB^k) = \int_{\underline{I}}^{UB^k} \hat{f}_X^k(\tau)d\tau$$

- $\hat{f}_X^k$: função densidade de probabilidade da normal truncada

$$\hat{f}_{X|\ \underline{I}\leq X\leq UB^k}(x|(\underline{I} \leq x \leq UB^k) = \frac{\hat{f}(x)_X^k}{\hat{F}_X^k(UB^k) - \hat{F}_X^k(\underline{I})}$$

Exemplos de Truncagem

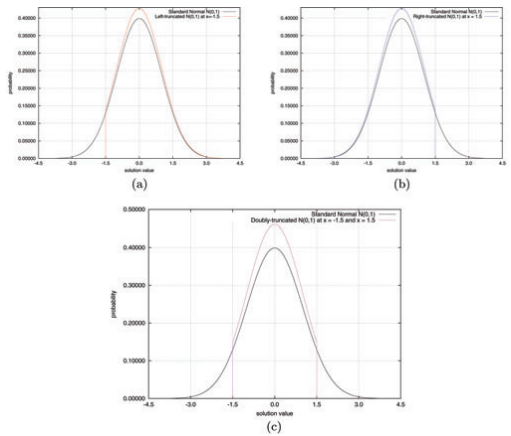


Fig. 6. Probability density functions of the truncated normal distributions. (a) Left-truncated normal (0,1) at $\underline{\ell} = -1.5$; (b) right-truncated normal (0,1) at $\overline{u} = 1.5$; (c) doubly truncated normal (0,1) at $\underline{\ell} = -1.5$ in the left and $\overline{u} = 1.5$ in the right.

Next, we perform N additional iterations and we count the number $N^{\leq}$ of solutions whose value is smaller than or equal to $\hat{F}_X^{\overline{k}}(\overline{UB})$.

The computational results displayed in Tables 14 and 15 show that $\hat{N}^{\leq}$ is a good estimation for the number $N^{\leq}$ of solutions found after N additional iterations whose value is smaller than or equal to the best solution value at the time the algorithm would stop for each threshold value β. Using the threshold $\beta = 10^{-1}$ is not appropriate, since at this point we are usually still very far from the optimal value and $\hat{F}_X^{\overline{k}}(\overline{UB})$ does not give a good estimate of the probability of finding a solution at least as good as the best known at this time. We also observe that for both the quadratic assignment and the set k-covering problems, whose results are depicted in Table 15, it has not been possible to reach a solution satisfying the threshold $\beta = 10^{-4}$ for any of their instances.

Therefore, the probability $\hat{F}_X^{\overline{k}}(UB^k)$ may be used to estimate the number of iterations that must be performed by the algorithm to find a new solution at least as good as the currently best one. The threshold β used to implement the stopping criterion may either be fixed a priori as a parameter or iteratively computed. In the last case, since the user is able to account for the average time taken by each GRASP iteration, this threshold can be determined online so as to limit the computation time when the probability of finding improving solutions becomes very small and the time needed to find improving solutions could become very large.

Critério de Parada e Validação

- Critério de parada: para um dado limiar β, pare GRASP quando $\hat{F}_X^k(UB^k) \leq \beta$.
- Validação: para cada valor de β, GRASP é executado até que $\hat{F}_X^k(UB^k) \leq \beta$.
- Seja $\bar{k}$ a iteração em que a condição é satisfeita e $\overline{UB}$ o valor da melhor solução.
- Neste ponto estima-se por $\hat{N}^{\leq} = \lfloor N \times \hat{F}_X^k(\overline{UB}) \rfloor$.
- $\hat{N}^{\leq}$: número de soluções com valor pelo menos tão bom quanto $\overline{UB}$ se N iterações adicionais de GRASP são executadas.
- A seguir as N iterações (empiricamente $N = 1.000.000$) são executadas e conta-se o número $\hat{N}^{\leq}$ de soluções com valor menor ou igual a $\overline{UB}$.

Critério de Parada e Validação

- Critério de parada: para um dado limiar β, pare GRASP quando $\hat{F}_X^k(UB^k) \leq \beta$.
- Validação: para cada valor de β, GRASP é executado até que $\hat{F}_X^k(UB^k) \leq \beta$.
- Seja $\bar{k}$ a iteração em que a condição é satisfeita e $\overline{UB}$ o valor da melhor solução.
- Neste ponto estima-se por $\hat{N}^{\leq} = \lfloor N \times \hat{F}_X^k(\overline{UB}) \rfloor$.
- $\hat{N}^{\leq}$: número de soluções com valor pelo menos tão bom quanto $\overline{UB}$ se N iterações adicionais de GRASP são executadas.
- A seguir as N iterações (empiricamente $N = 1.000.000$) são executadas e conta-se o número $\hat{N}^{\leq}$ de soluções com valor menor ou igual a $\overline{UB}$.

Critério de Parada e Validação

- Critério de parada: para um dado limiar β, pare GRASP quando $\hat{F}_X^k(UB^k) \leq \beta$.
- Validação: para cada valor de β, GRASP é executado até que $\hat{F}_X^k(UB^k) \leq \beta$.
- Seja $\bar{k}$ a iteração em que a condição é satisfeita e $\overline{UB}$ o valor da melhor solução.
- Neste ponto estima-se por $\hat{N}^{\leq} = \lfloor N \times \hat{F}_X^k(\overline{UB}) \rfloor$.
- $\hat{N}^{\leq}$: número de soluções com valor pelo menos tão bom quanto $\overline{UB}$ se N iterações adicionais de GRASP são executadas.
- A seguir as N iterações (empiricamente $N = 1.000.000$) são executadas e conta-se o número $\hat{N}^{\leq}$ de soluções com valor menor ou igual a $\overline{UB}$.

- Para as instâncias das $p-$medianas a estimativa é melhor para os valores de β iguais a 10^{-4} e 10^{-5}.

- O valor de *beta* pode ser fixo a priori ou calculado iterativamente.

Table 14
2-path network design and p-median problems: stopping criterion vs. estimated and counted number of solutions at least as good as the incumbent after $N = 1,000,000$ additional iterations

Problem	Instance	Threshold β	Iteration $\overline{k}$	Probability $\hat{F}_X^{\overline{k}}(\overline{UB})$	Estimation $\hat{N}^{\leq}$	Count $N^{\leq}$
		10^{-1}	3	0.079046	79,046	1843
		10^{-2}	25	0.009970	9970	1843
	2pndp50	10^{-3}	318	0.000757	757	738
		10^{-4}	4778	0.000001	1	0
		10^{-5}	4778	0.000001	1	0
		10^{-1}	3	0.078669	78,669	148,028
		10^{-2}	102	0.008923	8923	9537
2pndp	2pndp70	10^{-3}	1870	0.000643	643	465
		10^{-4}	32,771	0.000036	36	26
		10^{-5}	49,633	0.000005	5	4
		10^{-1}	4	0.085933	85,933	2066
		10^{-2}	41	0.009257	9257	2066
	2pndp90	10^{-3}	722	0.000326	326	190
		10^{-4}	5209	0.000015	15	7
		10^{-5}	270,618	0.000001	1	0
		10^{-1}	23	0.028989	28,989	32,151
		10^{-2}	232	0.001821	1821	1539
	2pndp200	10^{-3}	556	0.000566	566	503
		10^{-4}	5377	0.000100	100	95
		10^{-5}	77,448	0.000001	1	1
		10^{-1}	4	0.060647	60,647	79,535
		10^{-2}	21	0.008542	8542	7507
	pmed14	10^{-3}	608	0.000786	787	215
		10^{-4}	217,169	6.93×10^{-5}	69	5
		10^{-5}	437,422	5.55×10^{-6}	6	0
		10^{-1}	5	0.069694	69,694	117,054
		10^{-2}	56	0.009214	9214	16,968
p-Median	pmed15	10^{-3}	3533	0.000626	626	311
		10^{-4}	10,264	6.36×10^{-5}	63	26
		10^{-5}	235,853	9.99×10^{-6}	10	3
		10^{-1}	3	0.089011	89,011	12,428
		10^{-2}	34	0.009309	9309	4176
	pmed25	10^{-3}	1060	0.000998	998	1232
		10^{-4}	2760	2.82×10^{-5}	28	38
		10^{-5}	81,382	4.84×10^{-6}	5	4
		10^{-1}	4	0.089941	89,941	120,598
		10^{-2}	40	0.004635	4635	1426
	pmed30	10^{-3}	320	0.000992	992	1133
		10^{-4}	29,142	2.86×10^{-6}	3	1
		10^{-5}	29,142	2.86×10^{-6}	3	1

GRASP com Critério de Parada Probabilístico

Procedimento GRASP (β, Semente)

1. Faça $f* \leftarrow \infty$;
2. Faça $k \leftarrow 0$;
3. repita
4. $x \leftarrow$ Greedy Randomized Construção(Semente);
5. $x \leftarrow$ Busca Local (Solução);
6. se $f(x) < f*$ então
7. $x* \leftarrow x$;
8. $f* \leftarrow f(x)$;
9. fim;
10. $k; \leftarrow k + 1$;
11. $f_k \leftarrow f(x)$;
12. $UB^k \leftarrow f*$;
13. Atualize a média m^k e o desvio padrão S^k de $f_1, \ldots, f_k$;
14. Calcule a estimativa $\hat{F}_X^k(f*) = \hat{F}_X^k(UB^k) = \int_{-\infty}^{f*} \hat{f}_X^k(\tau) d\tau$;
15. até $\hat{F}_X^k(f*) < \beta$;
16. retorne $x*, f*$;

fim.

Análise de Variância: Introdução

- Característica comum: variação da observação física de medidas científicas, causada, por exemplo, a condições externas não controláveis.
- Metodologia análise de variança - ANOVA (analysis of variance) investiga variações da observação física associadas com fatores distintos.
- Envolve a divisão da variação total nos dados em componentes individuais atribuídos a vários fatores e aqueles devidos a erros aleatórios.
- Realiza testes de significância para determinar quais fatores influenciam o experimento.
- Metodologia desenvolvida por Sir Ronald A. Fisher (1918, 1925, 1935) que batizou-a análise de variança, ferramenta mais usada em estatística moderna (pós 1950) nas mais diversas áreas tais como, biologia, psicologia, medicina, sociologia, educação, agricultura e engenharia.
- Modelos de análise de variança são largamente usados para analisar o efeito de variáveis independentes em variáveis dependentes ou medida de resposta de interesse.

Análise de Variância: Introdução

- Característica comum: variação da observação física de medidas científicas, causada, por exemplo, a condições externas não controláveis.
- Metodologia análise de variança - ANOVA (analysis of variance) investiga variações da observação física associadas com fatores distintos.
- Envolve a divisão da variação total nos dados em componentes individuais atribuídos a vários fatores e aqueles devidos a erros aleatórios.
- Realiza testes de significância para determinar quais fatores influenciam o experimento.
- Metodologia desenvolvida por Sir Ronald A. Fisher (1918, 1925, 1935) que batizou-a análise de variança, ferramenta mais usada em estatística moderna (pós 1950) nas mais diversas áreas tais como, biologia, psicologia, medicina, sociologia, educação, agricultura e engenharia.
- Modelos de análise de variança são largamente usados para analisar o efeito de variáveis independentes em variáveis dependentes ou medida de resposta de interesse.

Análise de Variância: Introdução

- Característica comum: variação da observação física de medidas científicas, causada, por exemplo, a condições externas não controláveis.
- Metodologia análise de variança - ANOVA (analysis of variance) investiga variações da observação física associadas com fatores distintos.
- Envolve a divisão da variação total nos dados em componentes individuais atribuídos a vários fatores e aqueles devidos a erros aleatórios.
- Realiza testes de significância para determinar quais fatores influenciam o experimento.
- Metodologia desenvolvida por Sir Ronald A. Fisher (1918, 1925, 1935) que batizou-a análise de variança, ferramenta mais usada em estatística moderna (pós 1950) nas mais diversas áreas tais como, biologia, psicologia, medicina, sociologia, educação, agricultura e engenharia.
- Modelos de análise de variança são largamente usados para analisar o efeito de variáveis independentes em variáveis dependentes ou medida de resposta de interesse.

Modelo Linear com um Fator (One-Way Classification)

Modelo simples, mas muito usado.

- a níveis distintos de um único fator.
- n observações em cada nível, totalizando $N = an$ observações.
- y_{ij} : valor associado à $j-$ésima observação no $i-$ésimo nível do fator.

Modelo Matemático

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, n \tag{5}$$

- μ : média global sobre todas as N observações.
- τ_i : efeito (um parâmetro) associado ao $i-$ésimo nível.
- e_{ij} : erro aleatório associado com $j-$ésima observação no $i-$ésimo nível.

Modelo Linear com um Fator (One-Way Classification)

Modelo simples, mas muito usado.

- a níveis distintos de um único fator.
- n observações em cada nível, totalizando $N = an$ observações.
- y_{ij} : valor associado à $j-$ésima observação no $i-$ésimo nível do fator.

Modelo Matemático

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, n \tag{5}$$

- μ : média global sobre todas as N observações.
- τ_i : efeito (um parâmetro) associado ao $i-$ésimo nível.
- e_{ij} : erro aleatório associado com $j-$ésima observação no $i-$ésimo nível.

Análise de Modelos de Variança

- Dois tipos de efeitos: Sistemáticos ou Fixos e Aleatórios.
- Modelo de efeitos fixos: τ_i é constante.
  - O experimento consiste em analisar os níveis pré-determinados de cada parâmetro
  - Objetivo: fazer inferências sobre os parâmetros μ, τ_i, σ^2.
- Modelo de efeitos aleatórios: τ_i são variáveis aleatórias com médias nulas e variâncias σ_τ e independentes de ϵ_{ij}.
  - O experimento consiste em analisar amostras de populações infinitas associadas com variáveis aleatórias.
  - Objetivo: fazer inferências sobre $\mu, \tau_i, \sigma^2, \sigma_\tau^2$.
- Modelos mistos: contêm paramêtros constantes e variáveis aleatórias.
- Outros modelos: população finita e outras distribuições sobre o erro aleatório.

Modelo com Efeito Fixo: Hipóteses

- Os erros e_{ij} são independentes com distribuição normal e variância σ^2, constante para todos níveis do fator.
- Os erros associados com qualquer par de observações são não correlacionados :

$$E(e_{ij}e_{i'j'}) = 0 \begin{cases} i \neq i', j \neq j' \\ i = i', j \neq j' \end{cases}$$

- Assume-se igualdade de todos os níveis, isto é,

$$E(y_{ij}) = \mu + \tau_i, i = 1, 2, \ldots, a$$

- Como μ é a média global

$$\frac{1}{a}\sum_{i=1}^{a} E(y_{ij}) = \mu + \sum_{i=1}^{a} \tau_i \qquad \Rightarrow \sum_{i=1}^{a} \tau_i = 0$$

Modelo com Efeito Fixo: Hipóteses

- Os erros e_{ij} são independentes com distribuição normal e variância σ^2, constante para todos níveis do fator.

- Os erros associados com qualquer par de observações são não correlacionados :

$$E(e_{ij}e_{i'j'}) = 0 \begin{cases} i \neq i', j \neq j' \\ i = i', j \neq j' \end{cases}$$

- Assume-se igualdade de todos os níveis, isto é,

$$E(y_{ij}) = \mu + \tau_i, i = 1, 2, \ldots, a$$

- Como μ é a média global

$$\frac{1}{a}\sum_{i=1}^{a} E(y_{ij}) = \mu + \sum_{i=1}^{a}\tau_i \qquad \Rightarrow \sum_{i=1}^{a}\tau_i = 0$$

Partição da Soma Total dos Quadrados

- Notação: barra no topo e um ponto como sufixo, indicam indicam uma média sobre o sufixo, como indicado na identidade

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \tag{6}$$

tal que

$$\bar{y}_{.} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$$

$$\bar{y}_{i..} = \frac{1}{an} \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$$

- Tomando o quadrado em (6) e somando sobre i e j

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{.} - \bar{y}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 + 2 \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.}) \tag{7}$$

Partição da Soma Total dos Quadrados

O produto cruzado se anula, pois

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{\cdot} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{i\cdot}) = \sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot}) = 0$$

e

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

A expressão (7) é simplificada

Partição da Soma Total dos Quadrados

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \tag{8}$$

- Equation 8: soma dos desvios quadráticos de observações individuais em relação à média global ou soma total dos quadrados SS_T.
- Soma dos desvios quadráticos das médias de cada grupo (fator) em relação à média global ou soma dos quadrados inter-grupos SS_B.
- Soma dos desvios quadráticos das observações em relação às médias dos grupos ou soma de quadrados intra-grupos SS_W.

Graus de Liberdade para Somas Quadráticas

- A soma total de quadrados SS_T é baseada em an desvios $y_{ij} - \bar{y}$ com restrições

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..}) = 0$$

  e, portanto, $a(n-1)$ graus de liberdade.

- De modo análogo a soma dos quadrados inter-grupos SS_B tem a desvios com restrições

$$\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..}) = 0$$

  e, portanto, $a-1$ graus de liberdade.

- No caso da soma dos quadrados intra-grupos SS_W, considere a componente corresponde ao $i-$ésimo fator, isto é,

$$\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

  existem $n-1$ graus de liberdade. Como existem a componentes, SS_W tem $a(n-1)$ graus de liberdade.

Esperança de Somas Quadráticas Médias

Modelo Matemático

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, n \quad (5)$$

Soma Total dos Quadrados

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i})^2 \quad (8)$$

- Tomando a esperança, obtém-se

$$\bar{y}_{i} = \frac{1}{n}\sum_{j=1}^{n}(\mu + \tau_i + e_{ij}) = \mu + \tau_i + \bar{e}_{i.} \quad (9)$$

$$\bar{y}_{i..} = \frac{1}{a}\sum_{i=1}^{a}(\mu + \tau_i + \bar{e}_{i.}) = \mu + \tau_{.} + \bar{e}_{..} \quad (10)$$

Esperança de Somas Quadráticas Médias

- Como os e_{ij}'s são não correlatos com média zero e variança σ^2 segue-se que

$$E(e_{ij}^2) = \sigma^2 \tag{11}$$

$$E(\bar{e}_{i.}^2) = \sigma^2/n \tag{12}$$

$$E(\bar{e}_{..}^2) = \sigma^2/an \tag{13}$$

- Substituindo as expressões (5), (9) e (10) no segundo e terceiro termos de (8), obtém-se

$$SS_W = \sum_{i=1}^{a}\sum_{j=1}^{n}(e_{ij} - \bar{e}_{i.})^2 \tag{14}$$

$$SS_B = n\sum_{i=1}^{a}(\tau_i - \bar{\tau}_. + \bar{e}_{i.} - \bar{e}_{..})^2 \tag{15}$$

- A partir de (9)-(13) obtém-se

$$E(MS_W) = E\left(\frac{SS_W}{a(n-1)}\right) = \sigma^2$$

$$E(MS_B) = \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2 + \sigma^2$$

Distribuição Amostral das Médias Quadráticas

- As médias quadráticas são funções de observações amostrais e, portanto, devem ter distribuições amostrais.
- Do teorema (pag.31) segue-se que em um estimador não tendencioso $\hat{\sigma}^2$ de σ^2

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2[\nu]}{\nu}$$

- Como

$$E(MS_W) = \sigma^2$$

então

$$\frac{MS_W}{\sigma^2} \sim \frac{\chi^2[a(n-1)]}{a(n-1)}$$

- Como

$$E(MS_B) = \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2 + \sigma^2$$

- então, MS_B é um estimador não tendencioso de σ^2 se $\tau_i = 0, i = 1, \ldots, a$, isto é

$$\frac{MS_B}{\sigma^2} \sim \frac{\chi^2[a-1]}{a-1}$$

Teste de Hipótese: Teste F para Análise de Variança

- Teste de hipótese nula: todos os níveis do fator têm o mesmo efeito, isto é

$$H_0 : \tau_1 = \tau_2 = \cdots \tau_a = 0$$

- A alternativa é

$$H_1 : \text{ os } \tau_i's \text{ não são todos nulos}$$

- Vimos que quando H_0 é verdadeiro

$$E(MS_W) = E(MS_B) = \sigma^2.$$

  e quando H_0 é falso

$$E(MS_B) > E(MS_W)$$

- Quando H_0 é verdadeiro, segue-se que a razão

$$F = \frac{MS_B/\sigma^2}{MS_W/\sigma^2} = \frac{MS_B}{MS_W} \tag{16}$$

- é distribuída como uma variável F com $a-1$ e $a(n-1)$ graus de liberdade.

- Por exemplo, se $\alpha = 0,05$ e se o valor calculado em (14) é maior que o ponto correspondente a 95% da distribuição de F, pode-se concluir que a hipótese H_0 é falsa no nível $\alpha = 0,05$ de significância.

Tabela para Análise da Variança

Análise da Variança com Efeito Fixo

Fonte de Variação	Graus de Liberdade	Soma dos Quadrados	Média Quadrática	Valor F
Inter	$a-1$	SS_B	MS_B	MS_B/MS_W
Intra	$a(n-1)$	SS_W	MS_W	
Total	$an-1$	SS_W		

Wafer de Circuito Integrado

- Circuitos integrados construídos sobre um disco de material semicondutor chamado wafer (bolacha) com dimensão típica de 5 a 8 polegadas.
- Máscara e um processo de aplicação de plasma são usados para criação de padrões de circuitos em que deposita-se alumínio ou cobre.
- Plasma é um gás parcialmente ionizado com mesmo número de cargas positivas e negativas, bem como partículas de gás não ionizado, por exemplo, fluorocarbono.
- Plasma é obtido por um gerador de radiofrequência (RF).
- O objetivo do experimento é modelar a relação entre a taxa de remoção (angstroms/min) do semicondutor e a potência RF.

Wafer de Circuito Integrado

Taxa de Remoção Observada							
Potência RF (W)	1	2	3	4	5	Totais	Médias
160	575	542	530	539	570	2756	551,2
180	565	593	590	579	610	2937	587,4
200	600	651	610	637	629	3127	625,4
220	725	700	715	685	710	3535	707,0
						$y_{..} = 12.355$	$\bar{y}_{..} = 617,75$

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - y_{..}^2$$

$$= (575)^2 + (542)^2 + \cdots + (710)^2$$

$$= 72.209$$

$$SS_B = \frac{1}{n}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{an}$$

$$= \frac{1}{5}[(2756)^2 + \cdots + (3535)^2] - \frac{(12.355)^2}{20}$$

$$= 66.870,55$$

$$SS_W = SS_T - SS_B$$

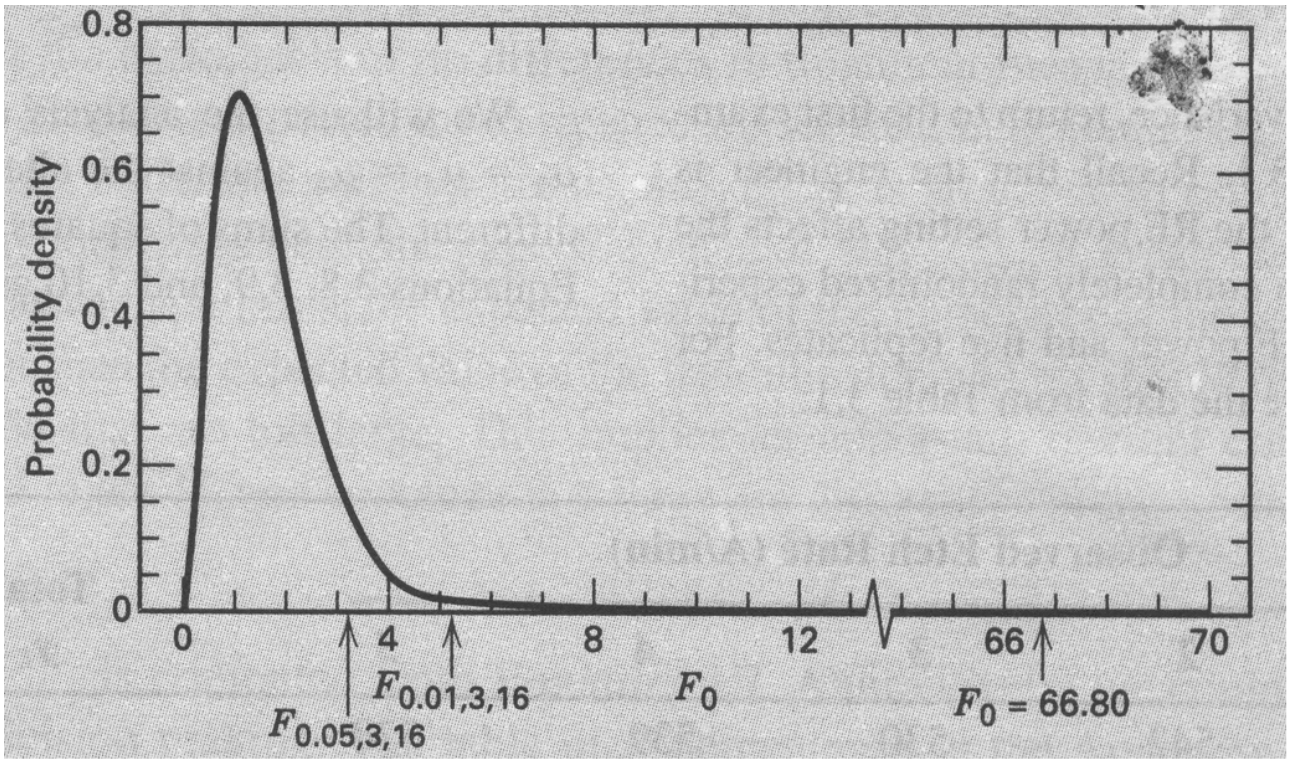
$$= 72.209,75 - 66.870,55 = 5339,20$$

Wafer de Circuito Integrado

Fonte de Variação	Graus de Liberdade	Soma dos Quadrados	Média Quadrática	Valor F_0	Valor-P
Potência RF	3	66.870	22.290,18	66,80	$< 0,01$
Intra	16	5339,20	333,70		
Total	19	72.209,75			

- $F_0 = 22.290,18/333,70 = 66,80$
- Para $\alpha = 0,05$, $F_{0,05;3;16} = 3,24$. Como $F_0 = 66,80 > 3,24$, a hipótese H_0 é rejeitada, isto é, a potência RF afeta significativamente a média da taxa de remoção.
- Como $F_{0,01;3;16} = 5,29$, e $F_0 > 5,29$, um limitante superior para o valor$-P$ é 0, 01 (o valor exato do valor$-P$ é $2,88 \times 10^{-9}$).

Wafer de Circuito Integrado



Verificação da Adequação do Modelo

- As seguintes hipóteses foram feitas para o modelo

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, n$$

(i) e_{ij}'s são normalmente distribuídos.

(ii) e_{ij}'s têm a mesma variança σ^2.

(iii) e_{ij}'s são não correlacionados.

- Na prática estas hipóteses não são completamente satisfeitas.
- Qual o efeito de desvios das hipóteses na inferências?
- Desvios relativamente pequenos em (i) e (ii) causam efeito pequeno em relação ao valor de F, enquanto desvios em (iii) afetam bastante F.

ANOVA : Problema de Cobertura Total

- Considere o atendimento de m por facilidades que podem ser instaladas em n locais. O coeficiente de cobertura para o cliente i e local j, $c_{ij} = 1$ se a distância d_{ij} entre um cliente i e um local j é limitada por $d_{\max}$, e $c_{ij} = 0$, caso contrário

- Seja $x_j = 1$ se uma facilidade é designada ao local j, $x_j = 0$, caso contrário.

iBusiness, 2010, 2, 156-167
doi:10.4236/ib.2010.22019 Published Online June 2010 (http://www.SciRP.org/journal/ib)

Scientific Research

Comparison of GA Based Heuristic and GRASP Based Heuristic for Total Covering Problem

Chandragiri Narashimhamurthy Vijeyamurthy[1], Ramasamy Panneerselvam[2]

[1]Network Coordination, BSNL, Chennai, India; [2]Department of Management Studies, School of Management, Pondicherry University, Pondicherry, India.
Email: vijeyamurthy@gmail.com, Panneer_dms@yahoo.co.in

Received January 25th, 2010; revised March 3rd, 2010; accepted April 9th, 2010.

ABSTRACT

This paper discusses the comparison of two different heuristics for total covering problem. The total covering problem is a facility location problem in which the objective is to identify the minimum number of sites among the potential sites to locate facilities to cover all the customers. This problem is a combinatorial problem. Hence, heuristic development to provide solution for such problem is inevitable. In this paper, two different heuristics, viz., GA based heuristic and GRASP based heuristic are compared and the best is suggested for implementation.

Keywords: *Genetic Algorithm, GRASP, Total Covering Problem, Boolean Operators, Care and Share Operator*

1. Introduction

Consider a sales region of a product, which in turn has different customer regions. The company selling the product should fix necessary number of dealer points such that the customers in that sales region are fully served. The company may fix a maximum of say 5 km of distance for the customers to reach a given dealer point from his/her region (customer region). In this process, a given customer may be covered by more than one dealer point. The objective is to locate the minimum number of dealer points in the sales region under consideration such that those dealer points cover all the customers regions.

Here, the process of serving a customer region by a dealer point is called as covering that customer region by that dealer point. The word "cover" means that the location of a customer is well within the given upper limit for the distance from a facility-location from where that customer will be served. Here, the objective is to locate facilities at minimum number of sites to cover all the customers. Such problem is known as *total covering problem* [1].

means that no region in the sales region is beyond 5 km from a site where a dealer is located. Based on this assumption, a covering coefficient is defined as shown below.

Let, $c_{ij} = 1$, *if* $d_{ij} \le 5\,km$
$= 0$, otherwise

where, c_{ij} is the covering coefficient for the customer region i and the potential dealer point j and d_{ij} is the distance between the customer region i and the potential dealer point j. Based on this definition of covering coefficient, the corresponding covering coefficient matrix is shown in **Table 2**.

A careful examination of the **Table 2** reveals that if the potential site 1 and the potential site 3 are assigned with dealers, all the six customer regions in the sales regions will be fully covered. As per this coverage, the potential site 1 will cover the customer regions 1, 2 and 4 and the potential site 3 will cover the customer regions

Table 1. Distance matrix of locating dealer points (distance in km)

Fatores para o Problema de Cobertura Total

- Fator A : esparsidade do coeficiente de cobertura c_{ij} (porcentagem de $1's$).
  - Níveis : $16\%, 18\%, 20\%, 22\%, 24\%$
- Fator B: tamanho da instância
  - Níveis : $30 \times 30, 40 \times 40, 50 \times 50, 60 \times 60, 70 \times 70, 80 \times 80$.
- Fator C : Algoritmos
  - Níveis: Algoritmo Genético (Alg_1) e GRASP (Alg_2) .

Modelo ANOVA

$$Y_{ijkl} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + e_{ijkl}$$

- Y_{ijk} : número de facilidades instaladas.
- μ : média global.
- Índices i, j, k representam os níveis dos fatores A, B, C.
- Índice l corresponde à $l-$ésima replicação para o $i-$ésimo nível do fator A, $j-$ésimo nível do fator B, $k-$ésimo nível do fator C.
- Para cada combinação i, j, k, tem-se 5 replicações.
- e_{ijkl} : erro aleatório associado aos índices i, j, k, l.
- Demais componentes do modelo representam os fatores A, B, C, e interação entre eles.

Resultados dos Algoritmos e ANOVA

- Valor Médio : $Alg_1 = 6,81 \quad Alg_2 = 15,79.$

Table 5. Responses (minimum number of sites assigned with facilities) of factorial experiment for comparison of GA based heuristic and GRASP based heuristic

Sparsity (%) (Factor A)	Replication	Problem size (Factor B)											
		30 × 30		40 × 40		50 × 50		60 × 60		70 × 70		80 × 80	
		Algorithm (Factor C)											
		Alg_1	Alg_2	Alg_1	Alg_2	Alg_1	Alg_2	Alg_1	Alg_2	Alg_1	Alg_2	Alg_1	Alg_2
16%	1	7	13	5	14	9	16	9	15	5	20	10	19
	2	7	11	8	14	8	14	7	17	8	19	9	20
	3	6	9	7	14	7	14	9	18	9	19	9	19
	4	6	16	7	13	7	12	8	19	9	18	9	18
	5	7	10	8	16	8	17	7	19	8	18	9	18
18%	1	8	12	7	14	7	13	7	15	7	16	9	15
	2	7	10	7	12	8	16	6	19	8	18	9	19
	3	6	10	8	17	6	18	7	19	9	20	7	18
	4	6	9	7	16	6	14	6	18	6	18	8	20
	5	8	12	6	14	7	13	7	17	7	20	9	18
20%	1	6	13	7	13	7	15	7	17	7	20	9	15
	2	5	10	6	13	6	16	8	20	9	19	7	18
	3	6	15	6	17	7	15	7	18	8	18	7	18
	4	4	10	6	13	7	13	7	15	7	17	8	18
	5	5	10	6	12	6	13	7	18	6	18	7	18
22%	1	5	13	6	13	7	16	6	18	8	18	7	16
	2	5	10	5	12	7	16	6	18	7	19	7	17
	3	5	10	6	12	7	17	6	19	7	22	7	19
	4	6	9	7	13	8	16	6	20	6	19	8	21
	5	6	11	5	14	7	16	7	19	8	19	8	18
24%	1	5	10	3	11	6	18	7	16	7	21	7	18
	2	5	9	6	11	6	18	7	23	6	18	6	19
	3	5	10	3	7	7	18	4	22	6	21	7	16
	4	6	11	6	13	6	15	6	15	6	19	7	19
	5	4	8	6	16	6	15	6	15	7	22	7	20

Table 6. ANOVA results for comparison of GA based heuristic with GRASP based heuristic

Source of Variation	Degrees of Freedom	Sum of squares	Mean Sum of Squares	Calculated F Ratio	Table F value at α = 0.05	Inference
A_i	4	34.7500	8.6875	4.2972	2.37	Significant
B_j	5	933.4571	186.6914	92.3345	2.21	Significant
AB_{ij}	20	61.2929	3.06465	1.5159	1.57	Not Significant
C_k	1	6048.0320	6048.0320	2991.6110	3.84	Significant
AC_{ik}	4	29.9805	7.4952	3.7074	2.37	Significant
BC_{jk}	5	421.1914	84.2383	41.6678	2.21	Significant
ABC_{ijk}	20	40.6953	2.0348	1.0065	1.57	Not Significant
e_{ijk}	240	485.1992	2.0217			
Total	299	8054.5980				

Comparação Valor Ótima Solução versus AG

mathematical model as presented in Section 3, which gives the optimal solution. It is well known that solving any mathematical model will be limited by its number of variables and the number of constraints, because of the software which will be used to solve the problem of interest. So, in this section, problems of limited sizes, viz., 32 × 32, 34 × 34, 36 × 36, 38 × 38 and 40 × 40, each with five replications are considered for comparing the performance of the GA based heuristic (Method 1) and that of the model (Method 2). The Problem Sizes are assumed to be the levels of "Factor A" and the Methods are assumed to be the levels of "Factor B".

The corresponding ANOVA model is shown below.

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk}$$

where,

Y_{ijk} is the response (number of sites assigned with facilities) of the k^{th} replication for the i^{th} treatment of the Factor A and the j^{th} treatment of the Factor B.

μ is the overall mean of the response.

A_i is the effect on the response due to the i^{th} treatment of the Factor A

B_j is the effect on the response due to the j^{th} treatment of the Factor B

AB_{ij} is the effect on the response due to the i^{th} treatment of the Factor A and the j^{th} treatment of the Factor B

e_{ijk} is the random error associated with the k^{th} replication under the i^{th} treatment of the Factor A and j^{th} treatment of the Factor B.

The different hypotheses relating to this model are as listed below.

Factor A (Problem Size)

H_0: There is no significant difference in terms of solution between different pairs of treatments of the Factor A (Problem Size).

H_1: There is significant difference in terms of solution between different pairs of treatments of the Factor A (Problem Size).

Factor B (Method)

H_0: There is no significant difference in terms of solution between different pairs of treatments of the Factor B (Method).

H_1: There is significant difference in terms of solution between different pairs of treatments of the Factor B (Method).

Factor A × Factor B: (AB_{ij})

H_0: There is no significant difference in terms of solution between the different pairs of interaction between Factor A and Factor B.

H_1: There is significant difference in terms of solution between different pairs of interaction between Factor A and Factor B.

A factorial experiment as per the above design was carried out to find the minimum number of potential sites which are to be assigned facilities under each experimental combination and such results (minimum number of potential sites assigned with facilities to cover all the customers) are summarized in **Table 7**. The results of ANOVA for the given factorial experiment are summarized in **Table 8**. From the **Table 8**, it is clear that all the calculated *F* ratios are less than the respective table *F* values at a significance level of 5%. Our prime concern is to check the significance of the effect of the Factor B (Method) on the response variable. The calculated *F* value for the Factor B is 1.5283 as against the table *F* value of 4.09. Hence, the corresponding null hypothesis is to be accepted and its alternate hypothesis is to be rejected. This means that there is no significant difference between the methods (Method 1 and Method 2) in terms of providing the minimum number of potential sites, which are assigned with facilities to cover all the customers. So, the performance of the GA based heuristic is equivalent to that of the model for the assumed problems of limited sizes.

Table 7. Results of GA based heuristic and model

Factor A (Problem Size)	Replication	Factor B (Method)	
		GA Based Heuristic (Method 1)	Model (Method 2)
32 × 32	1	4	4
	2	5	5
	3	4	4
	4	5	5
	5	5	5
34 × 34	1	5	4
	2	6	5
	3	4	4
	4	5	3
	5	5	5
36 × 36	1	5	4
	2	3	3
	3	4	4
	4	5	4
	5	5	5
38 × 38	1	5	5
	2	5	5
	3	4	3
	4	7	6
	5	6	6
40 × 40	1	6	6
	2	5	5
	3	3	2
	4	6	6
	5	6	6

Comparação ANOVA Solução Ótima versus AG

- Fator Método B_j : Hipótese nula é aceita, isto é, não há diferença significativa entre os métodos.

Table 8. ANOVA results of comparison of GA based heuristic and model

Source of Variation	Sum of squares	Degrees of Freedom	Mean Sum of Squares	Calculated F Ratio	Table F value at α = .05	Inference
A_i	6.72	4	1.68	1.5849	2.61	Not Significant
B_j	1.62	1	1.62	1.5283	4.09	Not Significant
AB_{ij}	0.88	4	0.22	0.2076	2.61	Not Significant
e_{ijkl}	42.40	40	1.06			
Total	51.62	49				

7. Conclusions

The total covering problem under facility location problem is an important problem to determine the minimum number of sites to locate the facilities to cover all the customers. Since, this problem comes under combinatorial category, in this paper, an attempt has been made to develop heuristics and compare them in terms of their performance. In the first phase, the design of GA based heuristic is given and it is followed by the design of GRASP based heuristic. Later, a complete factorial experiment has been conducted to compare the performance of the two heuristics by assuming three factors, Factor A (Percentage Sparsity), Factor B (Problem Size) and Factor C (Algorithms). The Factor *A* is assumed with 5 levels, which are viz. 16%, 18%, 20%, 22% and 24%. The Factor *B* is assumed with 6 levels, which are, viz. 30 × 30, 40 × 40, 50 × 50, 60 × 60, 70 × 70 and 80 × 80. The Factor *C* is assumed with 2 levels, which are viz. Alg_1 and Alg_2. For each experimental combination, 5 replications are carried out. Through ANOVA, it is found that the GA based heuristic performs better than the GRASP based heuristic in terms of providing the solution for the total covering problem.

After having concluded that the GA based heuristic performs better than the other heuristic, in the next phase, a comparison is done between the solution of the GA based heuristic and that of the mathematical model presented in Section 3 for the total covering problem through a two factor complete factorial experiment. In this experiment five different problem sizes (32 × 32, 34 × 34, 36 × 36, 38 × 38 and 40 × 40) are considered. For each problem size, five replications are considered. By taking the problem size as one factor and the methods of solving the total covering problem (*GA based heuristic* and *Model*) as another factor, a factorial experiment was conducted and found that the there is no difference between the methods, viz., *GA based heuristic* and *model* in terms of providing solution for the total covering problem. Hence, it is concluded that the performance of the GA based heuristic can be equated to that of the model which gives number of variables and the number of constraints of a model that can be handled by software. Finally, it is concluded that the GA based heuristic performs better than the GRASP based heuristic to solve the total covering problem. Further, there is no significant difference between the GA based heuristic and the mathematical model, in terms of providing solution for the total covering problem for moderate size problems.

REFERENCES

[1] R. Panneerselvam, "Production and Operations Management," 2nd Edition, Prentice-Hall India (P) Ltd., New Delhi, 2005.

[2] C. Toregas, R. Swain, C. Revelle and L. Bergman, "The Location of Emergency Service Facilities," *Operations Research*, Vol. 19, No. 6, 1971, pp. 1363-1373.

[3] N. R. Patel, "Location of Rural Social Service Centers in India," *Management Science,* Vol. 25, No. 1, 1979, pp. 22-30.

[4] T. D. Klastorin, "On the Maximal Covering Location Problem and the Generalized Assignment Problem," *Management Science*, Vol. 25, No.1, 1979, pp. 107-111.

[5] O. Saatcioglu, "Mathematical Programming Model for Airport Site Selection," *Transportation Research-B*, Vol. 16B, No. 6, 1982, pp. 435-447.

[6] A. W. Neebe, "A Procedure for Locating Emergency-Service Facilities for All Possible Response Distances," *Journal of Operational Research Society*, Vol. 39, No. 8, 1988, pp. 743-748.

[7] R. Panneerselvam, "A Heuristic Algorithm for Total Covering Problem," *Industrial Engineering Journal*, Vol. 19, No. 2, 1990, pp. 1-10.

[8] G. Rajkumar and R. Panneerselvam, "An Improved Heuristic for Total Covering Problem," *Industrial Engineering Journal*, Vol. 20, No. 8, 1991, pp. 4-7.

[9] R. Panneerselvam, "Efficient Heuristic for Total Covering Problem," *Productivity*, Vol. 36, No. 4, 1996, pp. 649-657.

[10] M. E. O'Kelly, "The Location of Interacting Hub Facilities," *Transportation Science*, Vol.20, No.2, 1986, pp. 92-106.

Scheduling a Dynamic Job Shop with Sequence Dependent Setups

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Robotics and Computer-Integrated Manufacturing 24 (2008) 435–449

Robotics and Computer-Integrated Manufacturing

www.elsevier.com/locate/rcim

Scheduling a dynamic job shop production system with sequence-dependent setups: An experimental study

V. Vinod[a], R. Sridharan[b,*]

[a]Department of Mechanical Engineering, N.S.S. College of Engineering , P.O. 678008, Palakkad, Kerala, India
[b]Department of Mechanical Engineering, National Institute of Technology Calicut, N. I. T. Campus, P.O. 673601, Calicut, Kerala, India

Received 20 September 2006; received in revised form 1 March 2007; accepted 7 May 2007

Abstract

This paper presents the salient aspects of a simulation-based experimental study of scheduling rules for scheduling a dynamic job shop in which the setup times are sequence dependent. A discrete event simulation model of the job shop system is developed for the purpose of experimentation. Seven scheduling rules from the literature are incorporated in the simulation model. Five new setup-oriented scheduling rules are proposed and implemented. Simulation experiments have been conducted under various experimental conditions characterized by factors such as shop load, setup time ratios and due date tightness. The results indicate that setup-oriented rules provide better performance than ordinary rules. The difference in performance between these two groups of rules increases with increase in shop load and setup time ratio. One of the proposed rules performs better for mean flow time and mean tardiness measures.

Keywords: Dynamic job shop scheduling; Sequence-dependent setup; Simulation; Scheduling rules; Regression metamodel

1. Introduction

Scheduling is the allocation of resources for performing a set of tasks [1]. Resources may be machines in a shop floor, runways in an airport, crews at a construction site or processing units in a computing environment [2]. Tasks may be operations in a shop floor, takeoffs and landing in an airport, stages in a construction project or computer programs to be executed. Proper scheduling leads to increased efficiency and capacity utilization, reduced time required to complete tasks and consequently increased profitability of an organization.

The dynamic job shop scheduling problem (DJSSP) is described as follows [3]. The job shop consists of M machines (work stations) and jobs arrive continuously over time. Each job requires a specific set of operations that need to be performed in a specified sequence (routing) on the machines and involves certain amount of processing time. The job shop becomes a queuing system: a job leaves one machine and proceeds on its route to another machine for the next operation, only to find other jobs already waiting for the machine to complete its current task, so that a queue of jobs in front of that machine is formed. Hence, DJSSP essentially involves deciding the order or priority for the jobs waiting to be processed at each machine to achieve the desired objectives. Scheduling rules or dispatching rules are used for this purpose. Blackstone et al. [4] have presented a review of dispatching rules that are used in job shop scheduling.

One of the standard assumptions in DJSSP is that setup times are included in the processing times. Setup involves the activities such as preparing a machine or workstation to perform the next machining operation. Setups may depend upon the type of job, the type of machine or both. Setup time is defined as the time interval between the end of processing of the current job and the beginning of processing of the next job. Setup time is encountered in manufacturing firms such as printing, plastics manufacturing, metal and chemical processing, paper industry, etc. The typical case in DJSSP is sequence-dependent setup times, where the setup time depends on the job previously processed. A typical example

*Corresponding author. Tel.: +91 495 2286406; fax: +91 495 2287250.
E-mail address: sreedhar@nitc.ac.in (R. Sridharan).

0736-5845/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.rcim.2007.05.001

Scheduling a Dynamic Job Shop with Sequence Dependent Setups

- O *job shop* consiste de M máquinas e tarefas (jobs) chegam continuamente ao longo do tempo de acordo com um processo de Poisson.
- Cada tarefa tem um roteiro de operações realizadas em uma sequência de máquinas.
- Tempos de preparação de tarefa são dependentes da sequência de tarefas e das máquinas.
- 10 tipos de tarefas são processadas com probabilidade igual de ser designada como a tarefa que chega.
- O número de operações para cada tarefa é uniformemente distribuído no intervalo $[5, 8]$. O roteiro é aleatório, e uma máquina não é revisitada.
- Tempos de setup, datas de entrega, e média de tempo entre chegadas de tarefas é variado.

Regras de Despacho

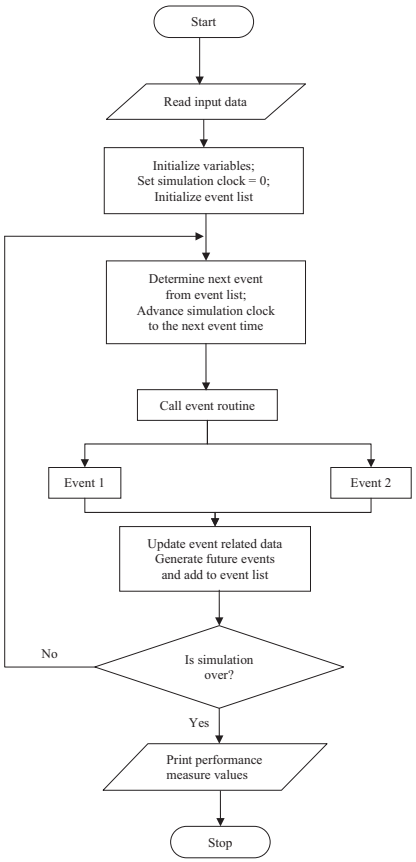


Fig. 1. Flow chart of simulation logic.

3.1.1. Event routine module

This module contains the subroutines, which deal with the following events that characterize the operation of the system.

(1) Arrival of a job to the shop.
(2) Departure of a job from a machine.

3.1.2. Job scheduling module

This module contains subroutines to deal with the scheduling of jobs on machines using various scheduling rules. Using a scheduling rule, for each of the machines, the jobs to be processed are scheduled as follows [3]. When a machine becomes free, it has to be decided which of the waiting jobs (if there is any in the queue of the machine) is to be processed on the machine. For making this decision, a scheduling rule is used to assign to each of the waiting jobs, a priority value. The job having the highest priority, which is defined by either the smallest or the largest priority value is selected for processing next. In the present study, seven existing rules from the literature are used. Five new scheduling rules for the sequence-dependent setup time environment of the job shop operation have also been proposed in the present study.

The following notations are used for the description of the scheduling rules:

m: index of the machine for which the job to be processed next has to be selected;
t: time at which the priority values are calculated;
i: index of the job for which the priority values are calculated;
j: index of the operation of job i;
k_i: set of operations to be performed on machines according to the routing of job i;
r_i^m: arrival time of job; i at machine m;
N_m^t: set of jobs waiting for processing in the queue of machine m at time t;
s_{ij}^m: setup time of operation j of job i on machine m;
p_{ij}^m: processing time of operation j of job i on machine m;
d_i: due date of job i;
Z_i^t: priority value of job i at time t.

The Scheduling rules are described as follows:
Existing rules:

(1) FIFO: First In First Out.
$Z_i^t = r_i^m$ where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_i^m\}$. Using the FIFO rule, the jobs are processed in the order they arrive at the machine.
(2) SPT: Shortest Processing Time.
$Z_i^t = p_{ij}^m$ where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the shortest processing time for the imminent operation is selected.
(3) EDD: Earliest Due Date.
$Z_i^t = d_i$ where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the smallest due date is selected.
(4) EMDD: Earliest Modified Due Date [4].
$Z_i^t = \max\{0, d_i - t - \sum_{q=j}^{Ki} piq\}$, where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the smallest modified due date is selected.
(5) CR: Critical Ratio.
$Z_i^t = (d_i - t)/\sum_{q=j}^{Ki} piq$, where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the smallest critical ratio is selected.

Regras de Despacho

(6) SIMSET: SIMilar SETup [15].
$Z_i^t = s_{ij}^m$, where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the smallest setup time is selected.
(7) JCR: Job with similar setup and Critical Ratio [15].

Select a job identical to the job that just finishes processing on the machine. When there is no identical job, select a job with the smallest critical ratio.

The following are the new setup-oriented scheduling rules proposed in the present study:

(1) SSPT: Shortest (Setup time + Processing Time).
$Z_i^t = p_{ij}^m + s_{ij}^m$, where the highest priority is given to the job i^* with $Z_{i*}^t = \min\{Z_i^t | i \in N_m^t\}$, i.e., the job with the smallest value of the sum of setup time and processing time is selected.
(2) JSPT: Job with similar setup and Shortest Processing Time.
Select a job identical to the job that just finishes processing on the machine.
When there is no identical job, select a job with the smallest processing time for the imminent operation.
(3) JEDD: Job with similar setup and Earliest Due Date.
Select a job identical to the job that just finishes processing on the machine. When there is no identical job, select a job with the earliest due date.
(4) JEMDD: Job with similar setup and Earliest Modified Due Date.
Select a job identical to the job that just finishes processing on the machine. When there is no identical job, select a job with the earliest modified due date.
(5) JSSPT: Job with similar setup and Shortest (Setup time + Processing Time).

Select a job identical to the job that just finishes processing on the machine. When there is no identical job, select a job with the smallest value of the sum of setup time and processing time for the imminent operation.

3.1.3. Report generation module

This module performs the task of consolidating the output of the simulation model to present results for the performance measures such as mean flow time, mean tardiness, mean setup time and mean number of setups. These performance measures are described as follows:

(1) mean flow time, F: it is the average time a job spends in the shop

$$F = [1/n]\left[\sum_{i=1}^{n} F_i\right];$$

(2) mean tardiness, T: it is the average tardiness of a job

$$T = [1/n]\left[\sum_{i=1}^{n} Ti\right];$$

(3) mean setup time: it is the average time spent by a job for the setup;
(4) mean number of setups: it is the average number of setups encountered by a job during its processing through various machines in the shop;

where C_i is the completion time of job i; a_i the arrival time of job i; d_i the due date of job i; n the number of jobs completed during the time interval from steady state period to simulation ending time; F_i the flow time of job i; $F_i = C_i - a_i$; T_i the tardiness of job i; $T_i = \max\{0, L_i\}$ (L_i = lateness of job i; $L_i = C_i - d_i$).

These performance measures are determined using the simulation output after the shop reaches steady state. Welch's method described by Law and Kelton [24] is used for identifying the steady state. The simulation output corresponding to the initial transient period is not considered for the computation of performance measures.

3.2. Verification and validation of the simulation model

Since the present study involves a conceptual job shop system, a multi-level verification exercise was performed to ensure correct programming and implementation of the conceptual model using the following steps.

(a) Debugging the program.
(b) Checking the internal logic of the modules of the model.
(c) Comparing the model output with the information obtained from a manual simulation using the same data.
(d) Running the model under different settings of the input parameters and checking whether the model behaves in a plausible manner.

4. Experimentation

Using the simulation model as an engine for experimentation, a number of experiments have been conducted. The objective of the experimentation is to investigate the performance of scheduling rules in a job shop system when setup times are sequence dependent. The details of experimentation are provided in the following sections.

4.1. Identification of steady state

The first stage in the simulation experimentation is determining the end of the initial transient period (identification of the steady state). For this purpose, Welch's procedure described in Law and Kelton [24] is

Cenários de Experimento

used. It is a graphical procedure consisting of plotting moving averages for the output performance measures. The end of the initial transient period is the time at which the moving averages approach a level value. For this purpose, a pilot simulation study was conducted in the present study. Ten replications were made. Each replication simulated the operation of the system for the completion of 1000 jobs. The experimental setting used was as follows. Mean interarrival time of jobs: 27 min; due date tightness factor: 5; ratio of mean setup time to mean processing time: 30%; scheduling rule: FIFO. The performance measures such as mean flow time, mean tardiness, mean number of setups and mean setup time were determined. It was found that the moving averages for all the performance measures approached a level value when 250 jobs were completed.

4.2. Identifying different scenarios for analysis

In the present simulation study, the first experimentation (scenario 1) involves the following settings: Mean interarrival time of jobs, a: 27 min; due date tightness factor, k: 5; ratio of mean setup time to mean processing time, s: 30%; scheduling rules: seven scheduling rules from the literature and five new setup-oriented scheduling rules described in Section 3.1.6. The performance of scheduling rules has also been investigated for three other scenarios. The experimental settings for all the four scenarios are summarized in Table 1.

Ten replications are performed for each experimental setting. The simulation for each replication is run for 1250 job completions. Jobs are numbered on arrival at the system and the simulation output from jobs numbering 1–250 is discarded. The outputs for the remaining 1000 jobs (jobs numbering 251–1250) are used for the computation of the performance measures.

5. Results and discussion

For each scenario the simulation results are subjected to statistical analysis using the analysis of variance (ANOVA) procedure in order to study the effect of experimental factors on the performance measures.

In scenario 1, the scheduling rule is the only factor and hence, one-way ANOVA has been carried out. For scenarios 2–4, two-factor ANOVA method is adopted. In performing statistical analysis, the simulation results pertaining to each replication have been accommodated in each treatment combination (cell). ANOVA-F test has been carried out to determine whether the treatment means are significantly different from each other. The least significant difference (LSD) method was used for performing pairwise comparisons in order to determine the means that differ from other means. The null hypothesis (H_0) is that all means are equal. The alternate hypothesis (H_1) is that at least two means are significantly different. All the tests were conducted at 5% level of significance. Values that are not significantly different are grouped. The results obtained and their analyses are presented in the following sections.

5.1. Results and discussion for scenario 1

Scenario 1 represents the base case wherein the purpose of analysis is to investigate the performance of scheduling rules in sequence-dependent job shop environment. At first, the average values of the performance measures are analyzed. Then, statistical analysis using ANOVA and means test are presented.

5.1.1. Analysis of means

The experimental settings for the scenario 1 are as follows: mean interarrival time of jobs, $a = 27$ min, due date tightness factor, $k = 5$ and setup time ratio $s = 30\%$. For each of the 12 scheduling rules, the simulation output for the 10 replications is averaged. These average values are presented in Figs. 2–5.

5.1.1.1. Mean flow time. Fig. 2 shows the simulation results for the mean flow time measure when different scheduling rules are used. It is found that SPT and SSPT rules provide smaller values for mean flow time. It is reported in the literature [21] that SIMSET performs best for mean flow time when the setup time of jobs are sequence dependent. But, the present study shows that SSPT is better then SIMSET for mean flow time.

Table 1
Experimental settings for the scenarios

Scenario	Experimental setting			Purpose of investigation
	Mean interarrival time	Setup time ratio (%)	Due date tightness factor	
1	27	30	5	Base case—analyze the performance of scheduling rules
2	27	20, 30, 40	5	Analyze the effect of changing setup time ratio
3	27	30	3, 5, 7	Analyze the effect of due date tightness factor
4	27, 28	30	5	Analyze the effect of changing shop load (i.e., changing mean interarrival time of jobs)

Cenário 1 de Experimento - Fator: Regra de Despacho

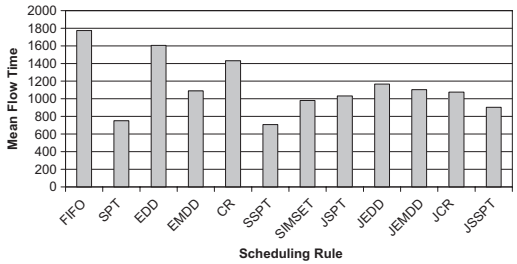


Fig. 2. Mean flow time ($k = 5$, $s = 30\%$).

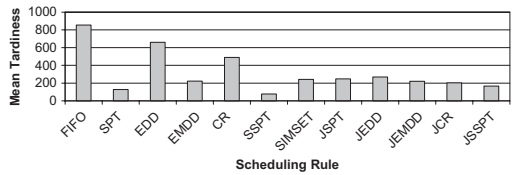


Fig. 3. Mean tardiness ($k = 5$, $s = 30\%$).

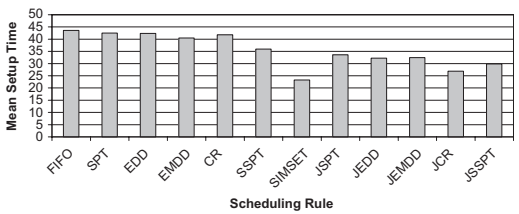


Fig. 4. Mean setup time ($k = 5$, $s = 30\%$).

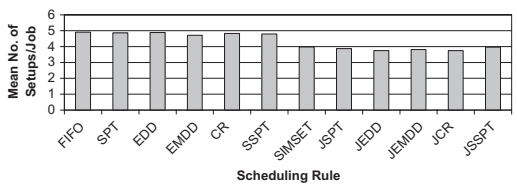


Fig. 5. Mean number of setups/job ($k = 5$, $s = 30\%$).

It can be observed that the setup-oriented rules such as JEDD, JEMDD and JCR perform better than ordinary rules such as EDD, EMDD and CR. Also, the modified due date based rules (EMDD, JEMDD) perform better than their counter parts EDD, JEDD which do not update the due dates dynamically.

Cenário 1 - Resultados de ANOVA e LSD

- Testes conduzidos em um nível de significância de 5%.
- Valores que não são significantemente diferentes são agrupados.
- Tabela 2 : 10 replicações da simulação - ANOVA, para cada medida de desempenho, determina se as médias são significantemente diferentes entre elas.
- Em todos os casos, existe uma diferença estatística entre médias de medidas de desempenho de uma regra de despacho para outra a um nível de confiança de 95%.
- Para determinar as médias que são significantemente diferentes usa-se na Tabela 3 o método de comparação entre pares da diferença estatística mínima (*least significant difference* - LSD).

Diferença Estatística Mínima (Fisher, 1935) - LSD

- A identificação de um resultado significativo por ANOVA indica que pelo menos um grupo difere dos outros grupos. Mas, este teste ”ônibus” não indica quem são os grupos distintos.
- ANOVA é então seguido por comparação entre pares de grupos.
- Primeiro teste (LSD) entre pares de grupos proposto por Fisher em 1935, usado quando ANOVA F é significante.
- A LSD calcula a diferença entre médias de todos os pares de grupos com o teste t, e declara significante qualquer diferença maior que LSD.

Diferença Estatística Mínima - LSD

- Considere A grupos, $i = 1, \ldots, A$.
- n_i : número de observações do grupo i. Se todos os grupos tem o mesmo tamanho, este número é denotado S.
- N : número total de observações.
- M_i : média do grupo i.
- MS_W : média do erro dentro do grupo.
- MS_B : média do erro entre grupos.

Diferença Estatística Mínima - LSD

- Quando a hipótese nula é verdadeira, o valor da estatística t entre grupos i e j

$$t = \frac{M_i - M_j}{\sqrt{MS_W \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad i \neq j$$

segue uma distribuição de t-Student com $\nu = N - A$ graus de liberdade.

- A diferença entre os grupos i e j é significante em um nível α se

$$|M_i - M_j| > LSD = t_{\nu,\alpha} \sqrt{MS_W \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Quando o número de observações é igual a S para todos os grupos

$$LSD = t_{\nu,alpha} \sqrt{MS_W \frac{2}{S}}$$

- Este procedimento é repetido por $\frac{A(A-1)}{2}$ comparações.

Diferença Estatística Mínima Modificada - MLSD

- O nível de α não é corrigido para múltiplas comparações entre pares.
- Pode então indicar uma diferença significante entre pares, quando isto não é verdade.
- Versão revisada de Hayter (1986) corrige isto.

$$MLSD = q_{\alpha,A-1}\sqrt{\frac{MS_W}{S}}$$

- $q_{\alpha,A-1}$: nível α para a distribuição de Student com faixa $A-1$ e $\nu = N - A$ graus de liberdade.

Cenário 1 - Resultados de ANOVA e LSD

5.1.1.2. Mean tardiness. Mean tardiness is a due date related performance measure and hence it has implications on average customer delivery performance. The performance of various scheduling rules for the mean tardiness measure is shown in Fig. 3. The SSPT rule performs best. As observed for the mean flow time measure, it is found that the SSPT rule outperformed SIMSET for the mean tardiness measure also.

Among the due date based rules, JCR provides lower values for mean tardiness, since it makes use of the due date data in addition to setup time data.

5.1.1.3. Mean setup time. This measure is a setup related measure. It denotes the average time incurred for setup activities in processing a job. Fig. 4 depicts the performance of the scheduling rules for the mean setup time measure. It is found that SIMSET rule outperforms all other rules. The second best rule in many cases is JCR rule followed by JSSPT rule. As expected, the non-setup-oriented rules such as FIFO, SPT, EDD, EMDD and CR lead to higher values for mean setup time of a job.

5.1.1.4. Mean number of setups. Fig. 5 shows the results for the mean number of setups per job when different rules are used for the scheduling decision. It is found that JCR rule provides smallest value for mean number of setups per job. Setup-oriented rules such as SIMSET, JSPT, JEDD, JEMDD, JCR and JSSPT rules perform better than non-setup-oriented rules.

5.1.2. ANOVA results

Using simulation results for 10 replications, ANOVA-*F* test has been carried out for each performance measure to determine whether the means are significantly different from each other. These results are shown in Table 2 for the performance measures such as mean flow time, mean tardiness, mean setup time, and mean number of setups. In all cases, since the *P*-value of the *F*-test is less than 0.05, there is a statistically significant difference between the mean performance measures from one scheduling rule to another at the 95% confidence level. To determine the means that are significantly different from other means, the LSD method of multiple comparison test is used. The results obtained using the LSD test are shown in Table 3.

The LSD test groups the results into five significantly different groups labeled a, b, c, d, e for mean flow time, six groups labeled a, b, c, d, e, f for mean tardiness, eight groups labeled a, b, c, d, e, f, g, h for mean setup time and five groups

Table 2
ANOVA results for base case

Performance measure	Source of variation	Sum of squares	Mean square	*F*-ratio	*P*-value
Mean flow time	Between groups	6.18704E6	562458.0	24.22*	0.0000
	Within groups	2.50792E6	23221.5		
Mean tardiness	Between groups	5.97862E6	543511.0	33.96*	0.0000
	Within groups	1.72854E6	16005.0		
Mean setup time	Between groups	5073.11	461.191	158.27*	0.0000
	Within groups	314.698	2.91387		
Mean number setups/job	Between groups	29.6142	2.6922	119.25*	0.0000
	Within groups	2.43829	0.0225768		

*Denotes *f* ratios significant at 5% significance level.

Table 3
Results for multiple regression base case (scenario 1)

Scheduling rule	Mean flow time	Mean tardiness	Mean setup time	Mean number of setups
FIFO	1372.610^{e}	855.792^{f}	43.5942^{h}	4.90731^{e}
SPT	613.379^{a}	128.315^{a,b}	42.5244^{g,h}	4.86445^{e}
EDD	1281.990^{d,e}	659.333^{e}	42.3745^{g,h}	4.88306^{e}
EMDD	1017.780^{c}	223.063^{b,c}	40.464^{f}	4.71229^{d}
CR	1208.380^{d}	489.325^{d}	41.8205^{f,g}	4.8289^{d,e}
SSPT	597.756^{a}	76.4587^{a}	35.9372^{e}	4.79169^{d,e}
SIMSET	919.306^{b,c}	241.362^{c}	23.2358^{a}	3.97076^{c}
JSPT	905.452^{b,c}	247.457^{c}	33.6076^{d}	3.87475^{b,c}
JEDD	1002.170^{c}	268.115^{c}	32.219^{d}	3.74585^{a,b}
JEMDD	954.406^{c}	219.662^{b,c}	32.3929^{d}	3.81382^{a,b}
JCR	1007.170^{c}	202.155^{b,c}	26.9052^{b}	3.73854^{a}
JSSPT	811.125^{b}	165.866^{a,b,c}	29.8227^{c}	3.96113^{c}

For each performance measure, values with the same letter are not found significantly different from each other by statistical test.

Cenário 2 de Experimento - Fatores: Regra de Despacho e Tempos de Setup

Table 4
ANOVA results for two-way analysis for scenario 2

Source of variation	*F*-ratio for performance measures			
	Mean flow time	Mean tardiness	Mean setup time	Mean number of setups/job
Main effects				
A: scheduling rule	130.75*	91.94*	359.07*	340.42*
B: setup time ratio	130.58*	108.32*	4030.68*	34.63*
Interaction AB	4.56*	7.31*	21.29*	1.02*

*Denotes *f* ratios significant at 5% significance level.

labeled a, b, c, d, e for mean number of setups. For the mean flow time measure, SSPT and SPT rules form a unique group labeled 'a'. Though there is no statistically significant difference among SSPT, SPT and JSSPT rules for the mean tardiness measures, SSPT rule provides the smallest value. The SIMSET rule forms a unique group labeled 'a' that denotes its superior performance for the mean setup time measure. The setup-oriented rules such as JCR, JEDD and JEMDD provide smaller values for mean number of setups.

5.2. Results and analysis for scenario 2

In this scenario, three different setup time matrices for each of the eight machines in the shop are used to investigate how the system performance is affected when the ratio of mean setup time to mean processing time changes. Simulation results are obtained for the two-factor experiments wherein the 12 scheduling rules form the first factor and the three levels of setup time ratio ($s = 20\%$, 30% and 40%) form the second factor. Ten replications are made for each of the 36 simulation experiments arising out of the combination of 12 scheduling rules and three setup time ratios. The results of two-factor ANOVA are shown in Table 4.

The main effects (scheduling rule, setup time) are significant for all the performance measures. The interaction effects are significant for the measures such as mean flow time, mean tardiness and mean setup time. The interaction plots are obtained for all the measures. However, due to space limitations, the plots for mean flow time and mean tardiness are shown in Figs. 6 and 7, respectively.

As evident from these figures, there is an increase in the performance measure values when the setup time ratio is increased. However, it is found that the rate of increase is smaller for the setup-oriented rules when compared to ordinary rules. The proposed rule, SSPT performs better than SPT rule for the measures such as mean flow time and mean tardiness when the setup time is fixed at 30% or 40%.

5.3. Results and analysis for scenario 3

The total work content method has been used in the present study for setting the due dates of jobs. The due date

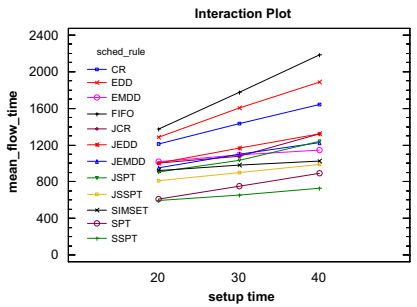


Fig. 6. Interaction plot for scenario 2—mean flow time.

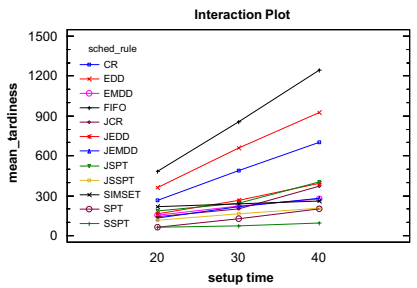


Fig. 7. Interaction plot for scenario 2—mean tardiness.

of each job is set equal to the sum of the arrival time and a multiple (due date factor, *k*) of the total processing time. In the base case (scenario 1), the due date factor is set equal to 5. In order to investigate the effect of due date tightness, the due date factor has been set at 3 and 7 to represent tight and loose due dates, respectively. Simulation experiments are conducted using a two-factor full factorial design. The experimental factors are job-scheduling rules (12 rules) and

Modelo de Regressão para as Três Melhores Regras

- Coeficiente de determinação R^2

formulating metamodel. Hence, the three scheduling rules are modeled by two-indicator variables x_1 and x_2. These variables take values 0 or 1 as defined below.

x_1	x_2	
0	1	If the observation is from SSPT rule
1	0	If the observation is from SPT rule
1	1	If the observation is from JCR rule

The setup time ratio is a quantitative variable and it is modeled by the variable $x3$. Significant interactions between scheduling rule and setup time ratio have been observed as described in Section 5.2. Hence, the cross product terms involving the indicator variables and the quantitative variables are also defined to represent the interaction effects. Taking these into consideration, the metamodel has been formulated as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x3 + \beta_5 x_2 x_3 + e, \quad (3)$$

where Y is the performance measure; β_0 the constant or intercept; β_1, β_2 the coefficients corresponding to the main effect of scheduling rules; β_3 the coefficients corresponding to the main effect of setup time ratios; β_4, β_5 the coefficients corresponding to the interaction effect of scheduling rules and setup time ratios; e the error.

There are nine simulation experiments arising out of the combinations of scheduling rules and setup time ratios (3 scheduling rules × 3 setup time ratios). For each combination, simulation results are available for each of the five performance measures (simulation results from scenario 2). Multiple linear regression analysis has been carried out using the simulation results for getting a set of five metamodels, one corresponding to each performance measure based on Eq. (3). The ANOVA results for the metamodels are shown in Table 8.

5.6.1. Results and discussion

Table 8 provides the ANOVA results for the metamodels. The explanatory power of the metamodels can be inferred from the value of the coefficient of determination R^2 obtained from the ANOVA for the whole model. This analysis also provides the P-value (probability value) for the model, from which the significance of the postulated metamodel can be known. Multiple linear regression analysis of the simulation results provides the estimates of the regression coefficients of the independent variables in the metamodels. The following observations are made:

(1) The coefficient of determination R^2 for the metamodels for the five performance measures has a high value. Hence, a larger proportion of the variation in the performance measures is explained by independent variables, namely the scheduling rules and the setup time ratios.
(2) The regression metamodels are also found to be highly significant since the P-value (0.0000) is less than the significance level, 0.05.
(3) The coefficient of determination R^2 and the adjusted R^2 values are very close implying that the model has not been over specified by including terms that do not contribute meaningfully to the fit.
(4) The adequacy of the models has been verified using the residual plots. Plots of the residual values versus the corresponding fitted values were made. These plots were found to be having no patterns, implying that there are no obvious model defects.
(5) Similar inferences have been made by plotting the residual values against the corresponding values of the independent variables and by plotting the residual values on the normal probability paper.

These inferences reveal that the metamodels developed adequately model the simulation model and thus can be of considerable interest for further application.

The metamodels obtained using the estimates of regression coefficients are as follows:

$$\begin{aligned}\text{mean flow time} = {} & 466.6330 - 132.6820x_1 \\ & + 192.2190x_2 + 647.5970x_3 \\ & + 746.5180x_1x_3 + 942.0750x_2x_3, \end{aligned} \quad (4)$$

$$\begin{aligned}\text{mean tardiness} = {} & 36.4485 - 111.3120x_1 \\ & - 141.8670x_2 + 142.5250x_3 \\ & + 543.8900x_1x_3 + 1008.6000x_2x_3, \end{aligned} \quad (5)$$

Table 8
Results of analysis of variance for the metamodels

Performance measure	Source of variation	Sum of squares	Mean squares	F-ratio	P-value	Coefficient of determination R^2	Adjusted R^2
Mean flow time	Model	4.7873E6	957459.0	71.84*	0.0000	81.0471	79.9189
	Error	1.11951E6	13327.5				
Mean tardiness	Model	7666990.0	153398.0	18.36*	0.0000	82.212	81.3674
	Error	702003.0	8357.18				
Mean setup time	Model	9715.82	1943.16	438.90*	0.0000	96.3134	96.0939
	Error	371.896	4.42733				
Mean number of setups/job	Model	24.762	4.95241	331.41*	0.0000	95.1753	94.8881
	Error	1.25526	0.0149435				

Validação do Modelo de Regressão

- Desvio dos valores preditos em relação ao resultado de simulação está dentro de 5%.

$$\text{mean setup time} = 5.6839 - 4.0929x_1 - 4.7249x_2 + 99.8487x_3 + 35.5280x_1x_3 - 2.2044x_2x_3, \quad (6)$$

$$\text{mean number of setups} = 4.9733 + 0.0905x_1 - 0.8511x_2 - 0.6179x_3 - 0.0532x_1x_3 - 0.7027x_2x_3. \quad (7)$$

5.6.2. Validation of metamodels

In order to test the validity of the metamodels developed input values for the independent variables that fall within the domain of definition of Eqs. (4)–(7) are used. For example the setup time ratio, s is fixed at 0.20.0.30 and 0.40 for constructing the metamodels denoted by Eqs. (4)–(7). Different values of setup time ratio in the interval 0.2–0.4 are chosen and used as inputs to the simulation model when the three scheduling rules are used. The performance

Table 9
Validation results of the metamodels

Scheduling rule	Performance measure	Setup time ratio (%)	Simulation results	Metamodel results	Error deviation (%)
SSPT	Mean flow time	24	618.443	622.06	−0.005
		28	611.6263	647.96	−0.032
		32	669.4373	673.86	−0.006
		36	680.3753	699.79	−0.028
	Mean tardiness	24	73.59988	70.65	0.040
		28	73.08341	76.36	−0.044
		32	84.86926	82.06	0.033
		36	84.3331	87.76	−0.040
	Mean setup time	24	30.23413	29.65	0.019
		28	34.51634	33.64	0.025
		32	38.44943	37.64	0.021
		36	42.30409	41.63	0.015
	Mean number of setups	24	4.878553	4.83	0.009
		28	4.842377	4.80	0.008
		32	4.811	4.78	0.006
		36	4.802141	4.75	0.010
SPT	Mean flow time	24	640.414	668.54	−0.043
		28	689.4014	724.30	−0.050
		32	766.7513	780.07	−0.017
		36	815.66	835.83	−0.025
	Mean tardiness	24	85.64193	89.88	−0.049
		28	113.77989	117.33	−0.031
		32	138.71716	144.78	−0.043
		36	165.3017	172.25	−0.042
	Mean setup time	24	34.00399	34.08	−0.002
		28	39.27447	39.50	−0.005
		32	45.88813	44.91	0.021
		36	51.14744	50.33	0.015
	Mean number of setups	24	4.933924	4.90	0.006
		28	4.904024	4.88	0.004
		32	4.896272	4.85	0.009
		36	4.854559	4.82	0.007
JCR	Mean flow time	24	1052.597	1040.37	0.011
		28	1130.595	1103.96	0.023
		32	1171.223	1167.55	0.003
		36	1267.891	1231.14	0.028
	Mean tardiness	24	163.3351	170.85	−0.046
		28	208.2177	216.90	−0.041
		32	257.8589	262.94	−0.019
		36	295.6834	308.98	−0.044
	Mean setup time	24	25.10198	24.39	0.028
		28	29.33342	28.30	0.035
		32	33.53766	32.21	0.039
		36	37.90635	36.11	0.047
	Mean number of setups	24	4.396604	4.57	−0.039
		28	4.443079	4.51	−0.015
		32	4.361366	4.46	−0.022
		36	4.213295	4.40	−0.044

Testes não Paramétricos (Livres de Distribuição)

Vantagens

- Requerem poucas hipóteses da população.
- São aplicáveis em muitas situações que procedimentos da teoria normal não podem ser utilizados.
- Muitos procedimentos não paramétricos requerem somente a classificação (*rank*) das observações e dispensam a magnitude das observações, que é requerida por procedimentos paramétricos.
- Enfoques não paramétricos podem ser usados em situações muito complicadas em que a teoria de distribuição necessária para apoiar métodos paramétricos é intratável.

Teste de Wilcoxon com *Ranks* Positivos ou Negativos

- Seja X uma variável aleatória contínua com distribuição simétrica.
- Seja M a mediana de X com $P(X \leq M) = P(X \geq M) = 1/2$.
- Se a função densidade de probabilidade é simétrica, a mediana é igual à média. Se não é simétrica, usualmente, não são iguais.
- Vamos apresentar um teste para

$$H_0 : M = M_0 \quad \text{versus} \quad H_1 : M \neq M_0 \quad (\text{ou } M < M_0 \text{ ou } M > M_0)$$

  a partir de uma amostra $X_1, \ldots, X_n$.

- Exemplo: Se U e V têm a mesma distribuição, então $X = U - V$ tem uma distribuição simétrica centrada na mediana com valor 0.

Cálculo do Teste Estatístico de Wilcoxon

- Seja uma amostra $X_1 \ldots X_n$: $1,1;\ 8,2;\ 2,3;\ 4,4;\ 7,5,;\ 9,6$. A mediana $M_0 = 5$ é plausível para esta amostra?
- Calcule $X_i - M_0, i = 1, \ldots, n$.
- Ordene $|X_i - M_0|$ do menor para o maior valor e associe *ranks* $1, 2, \ldots, n$.

  Seja R_i o *rank* de $|X_i - M_0|$ e defina $Z_i = \begin{cases} 0 & \text{se } X_i - M_0 < 0 \\ 1 & \text{se } X_i - M_0 > 0 \end{cases}$
- Calcule o teste estatístico $W = Z_1 R_1 + \ldots + Z_n R_n$.

i	X_i	$X_i - M_0$	R_i	Z_i
1	$1,1$	$-3,9$	5	0
2	$8,2$	$3,2$	4	1
3	$2,3$	$-2,7$	3	0
4	$4,4$	$-0,6$	1	0
5	$7,5$	$2,5$	2	1
6	$9,6$	$4,6$	6	1

$$n = 6$$

$|X_i - M_0|$ ordenado do menor para o maior valor

$$0,6;\ 2,5;\ 2,7;\ 3,2;\ 3,9;\ 4,6$$

$$W = 4 + 2 + 6 = 12$$

Teste de Wilcoxon com *Ranks* Positivos ou Negativos

- **Objetivo:** Comparação entre duas heurísticas aplicadas a n instâncias, totalizando $2n$ observações.

- O valor da função objetivo da instância i obtido pelas heurísticas 1 e 2 é representado por X_i e $Y_i, i = 1, 2, \ldots n$.

Instâncias	Heurística 1	Heurística 2
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

- **Hipóteses:**

- Seja $Z_i = Y_i - X_i$. As variáveis aleatórias $Z_i, i = 1, 2, \ldots n$ são independentes.

- Seja $Z_i, i = 1, 2, \ldots n$ uma amostra da função densidade de probabilidade $f_Z(z)$ que é contínua e simétrica com mediana zero.

Teste de *Rank* Wilcoxon

- Deseja-se testar

$$H_0 : M_0 = 0$$

versus

$$H_1 : M_0 \neq 0$$

- A estatística de *rank* com sinal de Wilcoxon é baseada nos desvios de $|Y_i - X_i|$ em relação a zero, e não depende da magnitude dos desvios.

- O menor valor de $|Y_i - X_i|$ tem *rank* 1, o segundo menor tem *rank* 2, e assim por diante até n.

- Associado ao valor de cada *rank* R_i, defina um indicador de sinal $Z_i = 0$ se $Y_i - X_i < 0$ e $Z_i = 1$ se $Y_i - X_i > 0$.

- A estatística de rank com sinal de Wilcoxon é definida como

$$W = \sum_{i=1}^{n} Z_i R_i$$

Teste de *Rank* com Sinal de Wilcoxon

Proposição

A distribuição de probabilidade de W quando H_0 verdadeira é dada por

$$P(W = w) = f_W(w) = \frac{1}{2^n} c(w)$$

Demonstração

Quando H_0 é verdadeira, a distribuição de $W = \sum_{i=1}^{n} Z_i R_i$ é equivalente à distribuição de $U = \sum_{i=1}^{n} U_i$ em que $U_i = 0$ com probabilidade 1/2 e $U_i = 1$ com probabilidade 1/2.

Portanto,

$$P(W = w) = f_W(w) = \frac{1}{2^n} c(w)$$

em que $c(w)$ é o número de formas de designar $1's$ e zeros aos n inteiros de forma que $\sum_{i=1}^{n} Z_i R_i = w$.

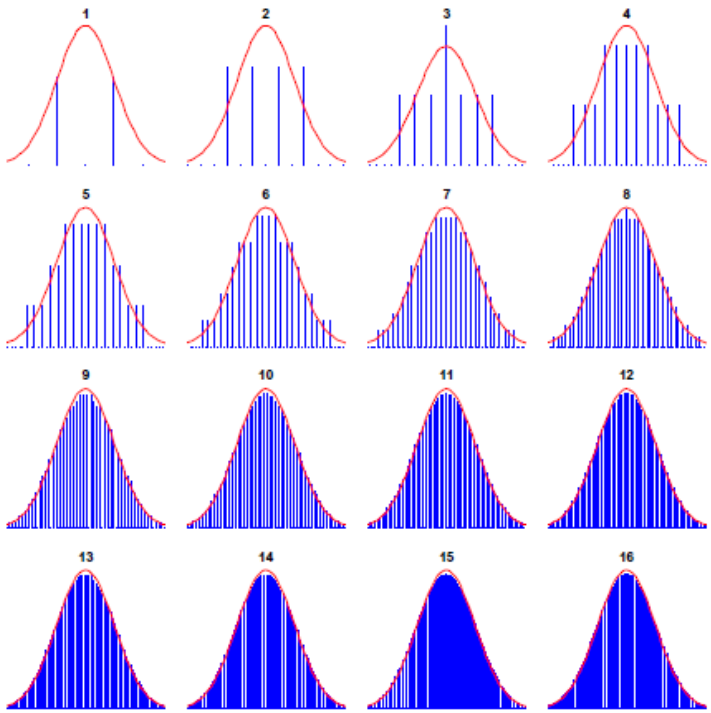
Distribuição da Estatística do Teste de Wilcoxon

Distribuição de Probabilidade de W para $n = 4$

		r_i			
w	$f_W(w) = P(W = w)$	1	2	3	4
0	$\frac{1}{16}$	0	0	0	0
1	$\frac{1}{16}$	+	0	0	0
2	$\frac{1}{16}$	0	+	0	0
3	$\frac{2}{16}$	+	+	0	0
		0	0	+	0
4	$\frac{2}{16}$	+	0	+	0
		0	0	0	+
5	$\frac{2}{16}$	+	0	0	+
		0	+	+	0
6	$\frac{2}{16}$	+	+	+	0
		0	+	0	+
7	$\frac{2}{16}$	+	+	0	+
		0	0	+	+
8	$\frac{1}{16}$	+	0	+	+
9	$\frac{1}{16}$	0	+	+	+
10	$\frac{1}{16}$	+	+	+	+
	1				

Distribuição da Estatística do Teste de Wilcoxon

Distribuição de Probabilidade de W para $n = 1, 2, \ldots, 16$



Média e Variança de Wilcoxon

Proposição

- Quando H_0 é verdadeira, a média e a variância da distribuição de probabilidade de W são dados por

$$E(W) = \frac{n(n+1)}{4}$$

$$\text{Var}(W) = \frac{n(n+1)(2n+1)}{24}$$

- Para $n > 12$, a distribuição de

$$W' = \frac{W - E(W)}{\sqrt{\text{Var}(w)}}$$

aproxima-se da distribuição normal padrão $N(0,1)$, ver Hogg e Craig, 1970.

- Quando ocorrem g empates em um desvio não zero, o desvio padrão da aproximação da normal é

$$\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum g^3 - \sum g}{48}}$$

(Estou à caça da demonstração deste resultado.)

Teste de *Rank* de Wilcoxon

Demonstração

$$E(W) = E\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} E(U_i)$$

$$= \sum_{i=1}^{n} \left(0 \cdot \frac{1}{2} + i \cdot \frac{1}{2}\right) = \sum_{i=1}^{n} \frac{i}{2} = \frac{n(n+1)}{4}$$

$$Var(W) = Var(U) = \sum_{i=1}^{n} Var(U_i)$$

$$Var(U_i) = E(U_i^2) - [E(U_i)]^2 = \frac{i^2}{2} - \left(\frac{i}{2}\right)^2 = \frac{i^2}{4}$$

$$Var(W) = \sum_{i=1}^{n} \frac{i^2}{4} = \frac{1}{4}\left[\frac{n(n+1)(2n+1)}{6}\right]$$

Exemplo:

Instâncias	Heurística 1 X_i	Heurística 2 Y_i
1	30	39
2	26	11
3	31	49
4	44	20
5	56	72
6	37	10
7	39	70
8	75	88
9	20	20
10	47	12
11	8	35
12	12	42
13	22	49

Teste de *Rank* de Wilcoxon

Exemplo

Diferença	|Diferença| $\lvert Y_i - X_i \rvert$	*Rank* R_i	*Rank* com Sinal	Z_i	$R_i Z_i$
9	9	1	1	1	1
−15	15	3	−3	0	0
18	18	5	5	1	5
−24	24	6	−6	0	0
16	16	4	4	1	4
−27	27	8	−8	0	0
31	31	11	11	1	11
13	13	2	2	1	2
0	0	Ignore	–	–	–
−35	35	12	−12	0	0
27	27	8	8	1	8
30	30	10	10	1	10
27	27	8	8	1	8
					$\overline{49}$

Teste de *Rank* de Wilcoxon

- Note que a diferença $|Y_i - X_i| = 27$ ocorre 3 vezes no sétimo lugar, após o *rank* 6. O *rank* associado a estes empates é $(7 + 8 + 9)/3 = 8$.

- Desta forma garantimos que a soma dos *ranks* é $n(n+1)/2 = 78$, que é o valor da soma *ranks* $1 + 2 + \cdots + 12$ correspondente ao valor máximo de $W = \sum_{i=1}^{12} R_i Z_i$, quando todos os valores de Z_i são iguais a 1.

- Para o exemplo, $W = \sum_i R_i Z_i = 49$.

- Para um nível de significância exato $\alpha = 0.052$ a região crítica da distribuição de W em que a hipótese H_0 é rejeitada é $C = \{w : w \leq 14 \text{ ou } w \geq 64\}$. Como $W = 49$, a hipótese nula H_0 não pode ser rejeitada.

Teste de *Rank* de Wilcoxon

- Seja $W' = \frac{W - E(W)}{\sqrt{\text{Var}(w)}}$, e considere o uso da distribuição normal.
- Em um problema de minimização, a heurística A é melhor que B se
- $H_0 : A - B \geq 0 \qquad H_1 : A - B < 0$
- H_0 é rejeitado se $P_{W'} \leq -0,05$.
- Em um problema de maximização, a heurística A é melhor que B se
- $H_0 : A - B \leq 0 \qquad H_1 : A - B > 0$
- H_0 é rejeitado se $P_{W'} \geq 0,95$.
- Para este exemplo, o valor-P é $0,001 << 0,05$.

Comparação de Múltiplos Algoritmos

Swarm and Evolutionary Computation 1 (2011) 3–18

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

journal homepage: www.elsevier.com/locate/swevo

Invited paper

A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms

Joaquín Derrac [a,*], Salvador García [b], Daniel Molina [c], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, 18071 Granada, Spain

[b] Department of Computer Science, University of Jaén, 23071 Jaén, Spain

[c] Department of Computer Engineering, University of Cadiz, 11003 Cadiz, Spain

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form
22 December 2010
Accepted 8 February 2011
Available online 18 February 2011

Keywords:
Statistical analysis
Nonparametric statistics
Pairwise comparisons
Multiple comparisons
Evolutionary algorithms
Swarm intelligence algorithms

ABSTRACT

The interest in nonparametric statistical analysis has grown recently in the field of computational intelligence. In many experimental studies, the lack of the required properties for a proper application of parametric procedures – independence, normality, and homoscedasticity – yields to nonparametric ones the task of performing a rigorous comparison among algorithms.

In this paper, we will discuss the basics and give a survey of a complete set of nonparametric procedures developed to perform both pairwise and multiple comparisons, for multi-problem analysis. The test problems of the CEC'2005 special session on real parameter optimization will help to illustrate the use of the tests throughout this tutorial, analyzing the results of a set of well-known evolutionary and swarm intelligence algorithms. This tutorial is concluded with a compilation of considerations and recommendations, which will guide practitioners when using these tests to contrast their experimental results.

1. Introduction

In recent years, the use of statistical tests to improve the evaluation process of the performance of a new method has become a widespread technique in computational intelligence. Usually, they are employed inside the framework of any experimental analysis to decide when one algorithm is considered better than another. This task, which may not be trivial, has become necessary to confirm whether a new proposed method offers a significant improvement, or not, over the existing methods for a given problem.

Statistical procedures developed to perform statistical analyses can be categorized into two classes: parametric and nonparametric, depending on the concrete type of data employed [1]. Parametric tests have been commonly used in the analysis of experiments in computational intelligence. Unfortunately, they are based on assumptions which are most probably violated when analyzing the performance of stochastic algorithms based on computational intelligence [2,3]. These assumptions are known as independence, normality, and homoscedasticity. To overcome this problem, our interest is focused on nonparametric statistical procedures, which provide to the researcher a practical tool to use when the previous assumptions cannot be satisfied, especially in multi-problem analysis.

In this paper, the use of several nonparametric procedures for pairwise and multiple comparison procedures is illustrated. Our objectives are as follows.

- To give a comprehensive and useful tutorial about the use of nonparametric statistical tests in computational intelligence, using tests already proposed in several papers of the literature [2–5]. Through several examples of application, we will show their properties, and how the use of this complete framework can improve the way in which researchers and practitioners contrast the results achieved in their experimental studies.
- To analyze the lessons learned through their use, providing a wide list of guidelines which may guide users of these tests when selecting procedures for a given case of study.

For each kind of test, a complete case of application is shown. A contest held in the CEC'2005 special session on real parameter optimization defined a complete suite of benchmarking functions (publicly available; see [6]), considering several well-known domains for real parameter optimization. These benchmark functions will be used to compare several evolutionary and swarm intelligence continuous optimization techniques, whose differences will be contrasted through the use of nonparametric procedures.

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: jderrac@decsai.ugr.es (J. Derrac), sglopez@ujaen.es (S. García), daniel.molina@uca.es (D. Molina), herrera@decsai.ugr.es (F. Herrera).

2210-6502/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.swevo.2011.02.002

Benchmark e Algoritmos

- Benchmark com 25 instâncias com 10 variáveis em otimização contínua e 9 algoritmos heurísticos

To do so, this paper is organized as follows. Section 2 shows the experimental framework considered for the application of the statistical methods and gives some preliminary background. Section 3 describes the nonparametric tests for pairwise comparisons. Section 4 deals with multiple comparisons by designating a control method, whereas Section 5 deals with multiple comparisons among all methods. Section 6 surveys several recommendations and considerations on the use of nonparametric tests. Finally, Section 7 concludes this tutorial.

2. Preliminaries

In this section, the benchmark functions (Section 2.1) and the evolutionary and swarm intelligence algorithms considered for our case of study (Section 2.2) are presented. Furthermore, some basic concepts on inferential statistics are introduced (Section 2.3), providing the necessary background for properly presenting the statistical procedures included in this tutorial.

2.1. Benchmark functions: CEC'2005 special session on real parameter optimization

Thorough this paper, the results obtained in a experimental study regarding 9 well-known algorithms and 25 optimization functions will be used, illustrating the application of the different statistical methodologies considered. The nonparametric tests will be used to show significant statistical differences among the different algorithms of the study.

As benchmark suite, we have selected the 25 test problems of dimension 10 that appeared in the CEC'2005 special session on real parameter optimization [6]. This suite is composed of the following functions.

- 5 unimodal functions
  - F1: Shifted Sphere Function.
  - F2: Shifted Schwefel's Problem 1.2.
  - F3: Shifted Rotated High Conditioned Elliptic Function.
  - F4: Shifted Schwefel's Problem 1.2 with Noise in Fitness.
  - F5: Schwefel's Problem 2.6 with Global Optimum on Bounds.
- 20 multimodal functions
  - 7 basic functions.
    - F6: Shifted Rosenbrock's Function.
    - F7: Shifted Rotated Griewank Function without Bounds.
    - F8: Shifted Rotated Ackley's Function with Global Optimum on Bounds.
    - F9: Shifted Rastrigin's Function.
    - F10: Shifted Rotated Rastrigin's Function.
    - F11: Shifted Rotated Weierstrass Function.
    - F12: Schwefel's problem 2.13.
  - 2 expanded functions.
    - F13: Expanded Extended Griewank's plus Rosenbrock's Function (F8F2)
    - F14: Shifted Rotated Expanded Scaffers F6.
  - 11 hybrid functions. Each one (F15 to F25) has been defined through compositions of 10 out of the 14 previous functions (different in each case).

All functions are displaced in order to ensure that their optima can never be found in the center of the search space. In two functions, in addition, the optima cannot be found within the initialization range, and the domain of search is not limited (the optimum is out of the range of initialization).

2.2. Evolutionary and swarm intelligence algorithms

Our main case of study consist of the comparison of performance between 9 continuous optimization algorithms. Their main

- **PSO**: A classic Particle Swarm Optimization [7] model for numerical optimization has been considered. The parameters are $c1 = 2.8$, $c2 = 1.3$, and w from 0.9 to 0.4. Population is composed by 100 individuals.
- **IPOP-CMA-ES**: IPOP-CMA-ES is a restart Covariant Matrix Evolutionary Strategy (CMA-ES) with Increasing Population Size [8]. This CMA-ES variation detects premature convergence and launches a restart strategy that doubles the population size on each restart; by increasing the population size, the search characteristic becomes more global after each restart, which empowers the operation of the CMA-ES on multi-modal functions. For this algorithm, we have considered the default parameters. The initial solution is uniform randomly chosen from the domain, and the initial distribution size is a third of the domain size.
- **CHC**: The key idea of the CHC algorithm [9] concerns the combination of a selection strategy with a very high selective pressure and several components inducing a strong diversity. In [10], the original CHC model was extended to deal with real-coded chromosomes, maintaining its basis as much as possible. We have tested it using a real-parameter crossover operator, BLX-α (with $\alpha = 0.5$), and a population size of 50 chromosomes.
- **SSGA**: A real-coded Steady-State Genetic Algorithm specifically designed to promote high population diversity levels by means of the combination of the BLX-α crossover operator (with $\alpha = 0.5$) and the negative assortative mating strategy [11]. Diversity is favored as well by means of the BGA mutation operator [12].
- **SS-arit & SS-BLX**: Two instances of the classic Scatter Search model [13] have been included in the study: the original model with the arithmetical combination operator, and the same model using the BLX-α crossover operator (with $\alpha = 0.5$) [14].
- **DE-Exp & DE-Bin**: We have considered a classic Differential Evolution model [15], with no parameter adaptation. Two classic crossover operators proposed in the literature, *Rand/1/exp*, and *Rand/1/bin*, are applied. The F and CR parameters are fixed to 0.5 and 0.9, respectively, and the population size to 100 individuals.
- **SaDE**: Self-adaptive Differential Evolution [16] is a Differential Evolution model which can adapt its CR and F parameters for enhance its results. In this model, the population size has been fixed to 100 individuals.

All the algorithms have been run 50 times for each test function. Each run stops either when the error obtained is less than 10^{-8}, or when the maximal number of evaluations (100 000) is achieved. Table 1 shows the average error obtained for each one over the 25 benchmark functions considered.

2.3. Some basic concepts on inferential statistics

Single-problem and multi-problem analyses can usually be found contrasting the results of computational intelligence experiments, both in isolation [17] and simultaneously [18]. The first kind, single-problem analysis, deals with results obtained over several runs of the algorithms over a given problem, whereas multi-problem analysis considers a result per algorithm/problem pair.

Inside the field of inferential statistics, hypothesis testing [19] can be employed to draw inferences about one or more populations from given samples (results). In order to do that, two hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1, are defined. The null hypothesis is a statement of no effect or no difference, whereas the alternative hypothesis represents the presence of an effect or a difference (in our case, significant differences between algorithms). When applying a statistical procedure to reject a hypothesis, a level of significance α is used to determine at which

Erros Médios Relativos

Table 1
Average error obtained in the 25 benchmark functions.

Function	PSO	IPOP-CMA-ES	CHC	SSGA	SS-BLX	SS-Arit	DE-Bin	DE-Exp	SaDE
F1	$1.234 \cdot 10^{-4}$	0.000	2.464	$8.420 \cdot 10^{-9}$	$3.402 \cdot 10$	1.064	$7.716 \cdot 10^{-9}$	$8.260 \cdot 10^{-9}$	$8.416 \cdot 10^{-9}$
F2	$2.595 \cdot 10^{-2}$	0.000	$1.180 \cdot 10^{2}$	$8.719 \cdot 10^{-5}$	1.730	5.282	$8.342 \cdot 10^{-9}$	$8.181 \cdot 10^{-9}$	$8.208 \cdot 10^{-9}$
F3	$5.174 \cdot 10^{4}$	0.000	$2.699 \cdot 10^{5}$	$7.948 \cdot 10^{4}$	$1.844 \cdot 10^{5}$	$2.535 \cdot 10^{5}$	$4.233 \cdot 10$	$9.935 \cdot 10$	$6.560 \cdot 10^{3}$
F4	2.488	$2.932 \cdot 10^{3}$	$9.190 \cdot 10$	$2.585 \cdot 10^{-3}$	6.228	5.755	$7.686 \cdot 10^{-9}$	$8.350 \cdot 10^{-9}$	$8.087 \cdot 10^{-9}$
F5	$4.095 \cdot 10^{2}$	$8.104 \cdot 10^{-10}$	$2.641 \cdot 10^{2}$	$1.343 \cdot 10^{2}$	2.185	$1.443 \cdot 10$	$8.608 \cdot 10^{-9}$	$8.514 \cdot 10^{-9}$	$8.640 \cdot 10^{-9}$
F6	$7.310 \cdot 10^{2}$	0.000	$1.416 \cdot 10^{6}$	6.171	$1.145 \cdot 10^{2}$	$4.945 \cdot 10^{2}$	$7.956 \cdot 10^{-9}$	$8.391 \cdot 10^{-9}$	$1.612 \cdot 10^{-2}$
F7	$2.678 \cdot 10$	$1.267 \cdot 10^{3}$	$1.269 \cdot 10^{3}$	$1.271 \cdot 10^{3}$	$1.966 \cdot 10^{3}$	$1.908 \cdot 10^{3}$	$1.266 \cdot 10^{3}$	$1.265 \cdot 10^{3}$	$1.263 \cdot 10^{3}$
F8	$2.043 \cdot 10$	$2.001 \cdot 10$	$2.034 \cdot 10$	$2.037 \cdot 10$	$2.035 \cdot 10$	$2.036 \cdot 10$	$2.033 \cdot 10$	$2.038 \cdot 10$	$2.032 \cdot 10$
F9	$1.438 \cdot 10$	$2.841 \cdot 10$	5.886	$7.286 \cdot 10^{-9}$	4.195	5.960	4.546	$8.151 \cdot 10^{-9}$	$8.330 \cdot 10^{-9}$
F10	$1.404 \cdot 10$	$2.327 \cdot 10$	7.123	$1.712 \cdot 10$	$1.239 \cdot 10$	$2.179 \cdot 10$	$1.228 \cdot 10$	$1.118 \cdot 10$	$1.548 \cdot 10$
F11	5.590	1.343	1.599	3.255	2.929	2.858	2.434	2.067	6.796
F12	$6.362 \cdot 10^{2}$	$2.127 \cdot 10^{2}$	$7.062 \cdot 10^{2}$	$2.794 \cdot 10^{2}$	$1.506 \cdot 10^{2}$	$2.411 \cdot 10^{2}$	$1.061 \cdot 10^{2}$	$6.309 \cdot 10$	$5.634 \cdot 10$
F13	1.503	1.134	$8.297 \cdot 10$	$6.713 \cdot 10$	$3.245 \cdot 10$	$5.479 \cdot 10$	1.573	$6.403 \cdot 10$	$7.070 \cdot 10$
F14	3.304	3.775	2.073	2.264	2.796	2.970	3.073	3.158	3.415
F15	$3.398 \cdot 10^{2}$	$1.934 \cdot 10^{2}$	$2.751 \cdot 10^{2}$	$2.920 \cdot 10^{2}$	$1.136 \cdot 10^{2}$	$1.288 \cdot 10^{2}$	$3.722 \cdot 10^{2}$	$2.940 \cdot 10^{2}$	$8.423 \cdot 10$
F16	$1.333 \cdot 10^{2}$	$1.170 \cdot 10^{2}$	$9.729 \cdot 10$	$1.053 \cdot 10^{2}$	$1.041 \cdot 10^{2}$	$1.134 \cdot 10^{2}$	$1.117 \cdot 10^{2}$	$1.125 \cdot 10^{2}$	$1.227 \cdot 10^{2}$
F17	$1.497 \cdot 10^{2}$	$3.389 \cdot 10^{2}$	$1.045 \cdot 10^{2}$	$1.185 \cdot 10^{2}$	$1.183 \cdot 10^{2}$	$1.279 \cdot 10^{2}$	$1.421 \cdot 10^{2}$	$1.312 \cdot 10^{2}$	$1.387 \cdot 10^{2}$
F18	$8.512 \cdot 10^{2}$	$5.570 \cdot 10^{2}$	$8.799 \cdot 10^{2}$	$8.063 \cdot 10^{2}$	$7.668 \cdot 10^{2}$	$6.578 \cdot 10^{2}$	$5.097 \cdot 10^{2}$	$4.482 \cdot 10^{2}$	$5.320 \cdot 10^{2}$
F19	$8.497 \cdot 10^{2}$	$5.292 \cdot 10^{2}$	$8.798 \cdot 10^{2}$	$8.899 \cdot 10^{2}$	$7.555 \cdot 10^{2}$	$7.010 \cdot 10^{2}$	$5.012 \cdot 10^{2}$	$4.341 \cdot 10^{2}$	$5.195 \cdot 10^{2}$
F20	$8.509 \cdot 10^{2}$	$5.264 \cdot 10^{2}$	$8.960 \cdot 10^{2}$	$8.893 \cdot 10^{2}$	$7.463 \cdot 10^{2}$	$6.411 \cdot 10^{2}$	$4.928 \cdot 10^{2}$	$4.188 \cdot 10^{2}$	$4.767 \cdot 10^{2}$
F21	$9.138 \cdot 10^{2}$	$4.420 \cdot 10^{2}$	$8.158 \cdot 10^{2}$	$8.522 \cdot 10^{2}$	$4.851 \cdot 10^{2}$	$5.005 \cdot 10^{2}$	$5.240 \cdot 10^{2}$	$5.420 \cdot 10^{2}$	$5.140 \cdot 10^{2}$
F22	$8.071 \cdot 10^{2}$	$7.647 \cdot 10^{2}$	$7.742 \cdot 10^{2}$	$7.519 \cdot 10^{2}$	$6.828 \cdot 10^{2}$	$6.941 \cdot 10^{2}$	$7.715 \cdot 10^{2}$	$7.720 \cdot 10^{2}$	$7.655 \cdot 10^{2}$
F23	$1.028 \cdot 10^{3}$	$8.539 \cdot 10^{2}$	$1.075 \cdot 10^{3}$	$1.004 \cdot 10^{3}$	$5.740 \cdot 10^{2}$	$5.828 \cdot 10^{2}$	$6.337 \cdot 10^{2}$	$5.824 \cdot 10^{2}$	$6.509 \cdot 10^{2}$
F24	$4.120 \cdot 10^{2}$	$6.101 \cdot 10^{2}$	$2.959 \cdot 10^{2}$	$2.360 \cdot 10^{2}$	$2.513 \cdot 10^{2}$	$2.011 \cdot 10^{2}$	$2.060 \cdot 10^{2}$	$2.020 \cdot 10^{2}$	$2.000 \cdot 10^{2}$
F25	$5.099 \cdot 10^{2}$	$1.818 \cdot 10^{3}$	$1.764 \cdot 10^{3}$	$1.747 \cdot 10^{3}$	$1.794 \cdot 10^{3}$	$1.804 \cdot 10^{3}$	$1.744 \cdot 10^{3}$	$1.742 \cdot 10^{3}$	$1.738 \cdot 10^{3}$

Instead of stipulating a priori a level of significance α, it is possible to compute the smallest level of significance that results in the rejection of H_0. This is the definition of the *p*-value, which is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that H_0 is true. It is a useful and interesting datum for many consumers of statistical analysis. A *p*-value provides information about whether a statistical hypothesis test is significant or not, and it also indicates something about *how significant* the result is: the smaller the *p*-value, the stronger the evidence against H_0. Most importantly, it does this without committing to a particular level of significance [20].

Parametric tests have been commonly used in the analysis of experiments in computational intelligence. For example, a common way to test whether the difference between the results of two algorithms is non-random is to compute a paired t-test, which checks whether the average difference in their performance over the problems is significantly different from zero. When comparing a set of multiple algorithms, the common statistical method for testing the differences between more than two related sample means is the repeated-measures ANOVA (or within-subjects ANOVA) [21].

Nonparametric tests, besides their original definition for dealing with nominal or ordinal data, can be also applied to continuous data by conducting ranking-based transformations, adjusting the input data to the test requirements [20]. They can perform two classes of analysis: pairwise comparisons and multiple comparisons. Pairwise statistical procedures perform individual comparisons between two algorithms, obtaining in each application a *p*-value independent from another one. Therefore, in order to carry out a comparison which involves more than two algorithms, multiple comparisons tests should be used. In $1 \times N$ comparisons, a control method is highlighted (the best performing algorithm) through the application of the test. Then, all hypotheses of equality between the control method and the rest can be tested by the application of a set of post-hoc procedures. $N \times N$ comparisons, considering the hypotheses of equality between all existing pairs of algorithms, are also possible, with the inclusion of specific post-hoc procedures for this task.

In this tutorial, we describe the use of several pairwise and multiple comparison procedures. Tables 2 and 3 enumerates the statistical tests and the post-hoc procedures considered, respectively. Furthermore, we present here some common notation that is used.

Table 2
Nonparametric statistical procedures considered in this tutorial.

Type of comparison	Procedures	Section
Pairwise comparisons	Sign test	3.1
	Wilcoxon test	3.2
Multiple comparisons ($1 \times N$)	Multiple sign test	4.1
	Friedman test	4.2
	Friedman Aligned ranks	4.2
	Quade test	4.2
	Contrast Estimation	4.4
Multiple comparisons ($N \times N$)	Friedman test	5

Table 3
Associated post-hoc procedures.

Type of comparison	Procedures	Section
Multiple comparisons ($1 \times N$)	Bonferroni	4.3
	Holm	4.3
	Hochberg	4.3
	Hommel	4.3
	Holland	4.3
	Rom	4.3
	Finner	4.3
	Li	4.3
Multiple comparisons ($N \times N$)	Nemenyi	5
	Holm	5
	Shaffer	5
	Bergmann	5

- n is the number of problems considered. i is its associated index.
- k is the number of algorithms included in the comparison. j is its associated index.
- d denotes the difference of performance between two algorithms in a given problem.

This notation will be employed throughout the study, unless a particular case is stated explicitly.

Comparação entre Pares de Algoritmos - Teste do Sinal

- Tabela 4 ilustra o teste do sinal, que conta quantas vezes SADE foi melhor que outro algoritmo.
- Tabela 5 destaca os casos em que uma diferença significativa foi detetada.

Table 4
Critical values for the two-tailed sign test at $\alpha = 0.05$ and $\alpha = 0.1$. An algorithm is significantly better than another if it performs better on at least the cases presented in each row.

#Cases	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$\alpha = 0.05$	5	6	7	7	8	9	9	10	10	11	12	12	13	13	14	15	15	16	17	18	18
$\alpha = 0.1$	5	6	6	7	7	8	9	9	10	10	11	12	12	13	13	14	14	15	16	16	17

Table 5
Example of Sign test for pairwise comparisons. SaDE shows a significant improvement over PSO, CHC, and SSGA, with a level of significance $\alpha = 0.05$, and over SS-Arit, with a level of significance $\alpha = 0.1$.

SaDE	PSO	IPOP-CMA-ES	CHC	SSGA	SS-BLX	SS-Arit	DE-Bin	DE-Exp
Wins (+)	20	15	20	18	16	17	13	9
Loses (−)	5	10	5	7	9	8	12	16
Detected differences	$\alpha = 0.05$	–	$\alpha = 0.05$	$\alpha = 0.05$	–	$\alpha = 0.1$	–	–

3. Pairwise comparisons

Pairwise comparisons are the simplest kind of statistical tests that a researcher can apply within the framework of an experimental study. Such tests are directed to compare the performance of two algorithms when applied to a common set of problems. In multi-problem analysis, a value for each pair algorithm/problem is required (often an average value from several runs).

In this section, first we focus our attention on a quick and easy, yet not very powerful, procedure, which can provide a first snapshot about the comparison: the Sign test (Section 3.1). Then, we will introduce the use of the Wilcoxon signed ranks test (Section 3.2), as a example of a simple, yet safe and robust, nonparametric test for pairwise statistical comparisons. Examples thorough this section will focus in characterizing the behavior of SaDE, in 1×1 comparisons with the rest of algorithms considered.

3.1. A simple first-sight procedure: the Sign test

A popular way to compare the overall performances of algorithms is to count the number of cases on which an algorithm is the overall winner. Some authors also use these counts in inferential statistics, with a form of two-tailed binomial test that is known as the Sign test [22]. If both algorithms compared are, as assumed under the null hypothesis, equivalent, each should win on approximately $n/2$ out of n problems.

The number of wins is distributed according to a binomial distribution; for a greater number of cases, the number of wins is under

populations? It is a nonparametric procedure employed in hypothesis testing situations, involving a design with two samples. This is analogous to the paired t-test in nonparametric statistical procedures; therefore, it is a pairwise test that aims to detect significant differences between two sample means, that is, the behavior of two algorithms.

Wilcoxon's test is defined as follows. Let d_i be the difference between the performance scores of the two algorithms on ith out of n problems (if these performance scores are known to be represented in different ranges, they can be normalized to the interval [0, 1], in order to not prioritize any problem; see [23]). The differences are ranked according to their absolute values; in case of ties, the practitioner can apply one of the available methods existing in the literature [24] (ignore ties, assign the highest rank, compute all the possible assignments and average the results obtained in every application of the test, and so on), although we recommend the use of average ranks for dealing with ties (for example, if two differences are tied in the assignation of ranks 1 and 2, assign rank 1.5 to both differences).

Let R^+ be the sum of ranks for the problems in which the first algorithm outperformed the second, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i)$$

Comparação entre Pares de Algoritmos - Teste de Wilcoxon

- Tabela 6 mostra que SADE atinge melhoria significativa sobre PSO, CHC e SSGA com nível de significância $\alpha = 0,01$, e sobre IPOP, CMA, ES e SS=Arit com $\alpha = 0,05$.

Table 6
Wilcoxon signed ranks test results. SaDE shows an improvement over PSO, CHC, and SSGA, with a level of significance $\alpha = 0.01$, over IPOP-CMA-ES and SS-Arit, with $\alpha = 0.05$, and over SS-BLX, with $\alpha = 0.1$.

Comparison	R^+	R^-	p-value	Comparison	R^+	R^-	p-value
SaDE versus PSO	261	64	0.00673	SaDE versus SS-BLX	232	93	0.06262
SaDE versus IPOP-CMA-ES	239	86	0.03934	SaDE versus SS-Arit	243	82	0.02958
SaDE versus CHC	287	38	0.00038	SaDE versus DE-Bin	176	149	>0.2
SaDE versus SSGA	260	65	0.00737	SaDE versus DE-Exp	119	206	>0.2

Example 2. When using Wilcoxon's test in our experimental study, the first step is to compute the R^+ and R^- related to the comparisons between SaDE and the rest of algorithms. Once they have been obtained, their associated p-values can be computed. Note that, for every comparison, the property $R^+ + R^- = n \cdot (n+1)/2$ must be true.

Table 6 shows the R^+, R^-, and p-values computed for all the pairwise comparisons concerning SaDE (the p-values have been computed by using SPSS). As the table states, SaDE shows a significant improvement over PSO, CHC, and SSGA, with a level of significance $\alpha = 0.01$, over IPOP-CMA-ES and SS-Arit, with $\alpha = 0.05$, and over SS-BLX, with $\alpha = 0.1$.

4. Multiple comparisons with a control method

One of the most frequent situations where the use of statistical procedures is requested is in the joint analysis of the results achieved by various algorithms. The groups of differences between these methods (also called blocks) are usually associated with the problems met in the experimental study. For example, in a multiple problem comparison, each block corresponds to the results offered over a specific problem. When referring to multiple comparisons tests, a block is composed of three or more subjects or results, each one corresponding to the performance evaluation of the algorithm over the problem.

In pairwise analysis, if we try to extract a conclusion involving more than one pairwise comparison, we will obtain an accumulated error coming from its combination. In statistical terms, we are losing the control on the Family-Wise Error Rate (FWER), defined as the probability of making one or more false discoveries among all the hypotheses when performing multiple pairwise tests. The true statistical significance for combining pairwise comparisons is given by

$$\begin{aligned} p &= P(Reject H_0 | H_0 true) \\ &= 1 - P(Accept H_0 | H_0 true) \\ &= 1 - P(Accept A_k = A_i,\ i = 1, \ldots, k-1 | H_0 true) \\ &= 1 - \prod_{i=1}^{k-1} P(Accept A_k = A_i | H_0 true) \\ &= 1 - \prod_{i=1}^{k-1} [1 - P(Reject A_k = A_i | H_0 true)] \\ &= 1 - \prod_{i=1}^{k-1} (1 - p_{H_i}). \end{aligned}$$

Therefore, a pairwise comparison test, such as Wilcoxon's test, should not be used to conduct various comparisons involving a set of algorithms, because the FWER is not controlled.

This section is devoted to describing the use of several procedures for multiple comparisons considering a control method. In

- First, we will introduce the use of the Sign test for multiple comparisons. This Multiple Sign test (Section 4.1) is a not very powerful method for detecting significant differences between algorithms, but it is still a quick and easy procedure which can be interesting for a first glance at the results.
- The best-known procedure for testing the differences between more than two related samples, the Friedman test, will be introduced in Section 4.2. In that section, we will also include the use of its extension, the Iman–Davenport test, and two advanced versions: the Friedman Aligned Ranks test and the Quade test.
- In Section 4.3, we will illustrate the use of a family of post-hoc procedures, as a suitable complement for the Friedman-related tests. Given a control method and the ranks of the Friedman (or any related) test, these post-hoc methods allow us to determine which algorithms are significantly better/worse than it.
- Finally, in Section 4.4, we present a procedure to estimate the differences between several algorithms: the Contrast Estimation of medians. This method is very recommendable if we assume that the global performance is reflected by the magnitudes of the differences among the performances of the algorithms.

4.1. Multiple Sign test

Given a control labeled algorithm, the Sign test for multiple comparisons allows us to highlight those ones whose performances are statistically different when compared with the control algorithm. This procedure, proposed in [26,27], proceeds as follows.

1. Represent by $x_{i,1}$ and x_{ij} the performances of the control and the jth algorithm in the ith problem.
2. Compute the signed differences $d_{i,j} = x_{i,j} - x_{i,1}$. That is, pair each performance with the control and, in each problem, subtract the control performance from the performance of the jth algorithm.
3. Let r_j equal the number of differences, $d_{i,j}$, that have the less frequently occurring sign (either positive or negative) within a pairing of an algorithm with the control.
4. Let M_1 be the median response of a sample of results of the control algorithm and M_j be the median response of a sample of results of the jth algorithm. Apply one of the following decision rules.
   - For testing H_0: $M_j \geq M_1$ against H_1: $M_j < M_1$, reject H_0 if the number of minus signs is less than or equal to the critical value of R_j appearing in Table A.21 in Appendix A for $k-1$ (number of algorithms excluding control), n (number of problems), and the chosen experimentwise error rate.
   - For testing H_0: $M_j \leq M_1$ against H_1: $M_j > M_1$, reject H_0 if the number of plus signs is less than or equal to the critical value of R_j appearing in Table A.21 in Appendix A for $k-1$, n, and the chosen experimentwise error rate.

Erros em Comparação de Múltiplos Algoritmos

	Situação Atual "Verdadeiro"	
Decisão	H_0 Verdadeiro	H_0 Falso
Não Rejeite H_0	Decisão Correta $1 - \alpha$	Decisão Incorreta Erro Tipo II β
Rejeite H_0	Decisão Incorreta Erro Tipo I α	Decisão Correta $1 - \beta$

Erros em Comparação de Múltiplos Algoritmos

- **Taxa de Erro em Família (*Familywise Error Rate*)**
- Considere o teste de $H_1, \ldots, H_m$.
- $m_0 =$ número de hipóteses verdadeiras; $R =$ número de hipóteses rejeitadas.

	Taxa de Descoberta de Falsos		
	H_0 Verdadeiro	H_1 Verdadeiro	Total
Não Significativos	U	T	$m - R$
Significativos	V	S	R
	m_0	$m - m_0$	m

- $V =$ number of erros falso positivos tipo I.
- Taxa de Descoberta de Falsos é projetada para controlar a proporção de falsos positivos dentre o conjunto de hipóteses rejeitadas $\frac{V}{R}$.

Erros em Comparação de Múltiplos Algoritmos

- Ao se fazer m comparações entre duas hipóteses:

$$P(\text{fazer um erro em um teste}) = \alpha$$

$$P(\text{não fazer um erro em um teste}) = 1 - \alpha$$

$$P(\text{não fazer um erro em } m \text{ testes}) = (1 - \alpha)^m$$

$$P(\text{fazer pelo menos um erro em } m \text{ testes}) = 1 - (1 - \alpha)^m$$

- Para $\alpha = 0,05$, $m = 3$, $1 - (1 - \alpha)^3 = 0,1426$

- Dizer que o valor-P está sendo ajustado para múltiplas hipóteses é equivalente a controlar a taxa de erro do tipo I.

- Existem muitos métodos com enfoques distintos para este tipo de controle.

Teste de *Rank* de Friedman

- É um teste não paramétrico desenvolvido pelo economista Milton Friedman (ele mesmo, o grande monetarista da Universidade de Chicago!!)
- Considere n instâncias de um problema, k heurísticas, e X_{ij} o valor da função objetivo da instância i e heurística j. Seja η_j a mediana dos valores obtidos pela heurística j.
- Suponha que as variáveis aleatórias associadas às k heurísticas são independentes, isto é, para uma dada instância, os resultados obtidos pelas heurísticas são independentes.
- Hipótese nula $H_0 : \eta_1 = \eta_2 = \cdots = \eta_k$
- Hipótese alternativa: as medianas não são iguais.

Instâncias	Heurística 1	Heurística 2	$\cdots$	Heurística k
1	X_{11}	X_{12}		X_{1k}
2	X_{21}	X_{22}		X_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	X_{n1}	X_{n2}		X_{nk}

Teste de *Rank* de Friedman

- Para um problema de minimização e uma dada instância (linha) i associe cada valor obtido pela heurística (coluna) j e atribua ranks $1, 2, \ldots, k$, em que 1 corresponde ao menor valor, 2 ao segundo menor valor, etc.

- Independência das heurísticas: para cada instância o *rank* atribuído é independente da heurística, e, portanto, o conjunto de *ranks* em cada coluna representa uma amostra aleatória dos ranks $1, 2, \ldots, p$.

- A soma dos *ranks* $1, 2, \ldots, p$ é $\frac{1}{2}p(p+1)$, e, portanto, o valor médio do rank da $j-$ésima coluna é $\frac{1}{2}(p+1)$.

- A soma dos quadrados dos ranks $1, 2, \ldots, p$ é $p(p+1)(2p+1)/6$.

- Portanto, a variância do valor do rank da $j-$ésima coluna é

$$p(p+1)(2p+1)/6 - (p+1)^2/4 = (p^2-1)/12.$$

Teste de *Rank* de Friedman

- Se $\bar{r}_j$ é o rank médio da $j-$ésima coluna, então a distribuição amostral das médias dos ranks tem valor médio $(p+1)/2$ e uma variância com valor $\sigma^2 = (p^2-1)12n$.

- A estatística do teste de Friedman é baseada na hipótese que a amostra dos ranks médios provêm de uma única população normal $N(\frac{12}{(p+1)}, \sigma^2)$ (ver teorema da distribuição da distribuição chi-quadrado da variância de uma amostra de uma distribuição normal).

$$\chi_F^2 = \frac{p-1}{p\sigma^2}\sum[\bar{r}_j - \frac{1}{2}(p+1)]^2 = \frac{12n}{p(p+1)}\sum[\bar{r}_j - \frac{1}{2}(p+1)]^2.$$

- Friedman demonstrou que para $p > 2$ e $n \to \infty$ então χ_F^2 converge para a distribuição χ^2 com $p-1$ graus de liberdade.

- Se o número de linhas e colunas não é muito pequeno, então χ_F^2 tem uma distribuição χ^2 com $p-1$ graus de liberdade. A perda de um grau de liberdade, deve-se ao fato que a soma dos ranks médios das p colunas é igual a $\frac{1}{2}p(p+1)$.

Teste de *Rank* de Friedman

- Vimos anteriormente que se $Z_1, Z_2, \ldots, Z_n$ são variáveis aleatórias independentes com $Z_i \sim N(\mu, \sigma), i = 1, 2, \ldots, n$. Se

$$X = \sum_{i=1}^{n} \left(\frac{Z_i - \mu}{\sigma} \right), \text{então } X \sim \chi^2_n.$$

- Não existe nenhum critério para escolher $p - 1$ desvios em χ^2_F. Por este motivo, calcula-se o desvio quadrático médio e multiplica-se por $p - 1$.

- Se $\bar{r}_j = 0$ para todo j, então $\chi^2_F = 0$ o que implica que H_0 é verdadeiro.

- Uma expressão mais adequada para o cálculo de χ^2_F em termos dos ranks r_{ij} é dada por

$$\chi^2_F = \frac{12}{np(p+1)} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} r_{ij} \right)^2 - 3np(p+1).$$

Exemplo do Teste de *Rank* de Friedman

- Quatro instâncias são avaliadas por quatro algoritmos, como mostra a tabela abaixo.

	Erros Relativos					Ranks de Friedman			
Erro	A	B	C	D	Friedman	A	B	C	D
1	2,711	3,147	2,515	2,612	1	3	4	1	2
2	7,832	9,828	7,832	7,921	2	1,5	4	1,5	3
3	0,012	0,532	0,122	0,005	3	2	4	3	1
4	3,431	4,111	3,401	3,401	4	3	4	1,5	1,5
					Média	2,375	4	1,250	1,875

Teste de *Rank* de Friedman Alinhado

- Para cada instância calcule o valor médio obtido por todos os algoritmos.
- Para cada célula instância $\times$ algoritmo, calcule a diferença entre o valor da célula e o valor médio.
- Designe $p \times n$ ranks para todas as células como no método de Friedman, atribuindo rank 1 à diferenca de menor valor, rank 2 à diferença de segundo menor valor, etc.
- Estes ranks associados a observações alinhadas são chamados ranks alinhados.

Exemplo do Teste *Rank* Alinhado de Friedman

Diferenças

Friedman	A	B	C	D
P1	-0,035	0,401	-0,231	-0,134
P2	-0,525	1,475	-0,521	-0,432
P3	-0, 156	0,365	-0,045	-0,162
P4	-0,155	0,525	-0,185	-0,185

Ranks de Friedman Alinhados

Friedman Alinhados	A	B	C	D
P1	12	14	4	10
P2	1,5	16	1,5	3
P3	8	13	11	7
P4	9	15	5,5	5,5
Média	5,625	15,5	5,5	6,375

Teste de Quade

Algumas instâncias são mais **difíceis** ou a **qualidade de suas soluções difere** muito entre algoritmos.

Rank para cada instância é **ponderado de acordo as diferenças observadas no desempenho dos algoritmos**.

- Para cada instância i e algoritmo j atribua o rank r_{ij} como no teste de Friedman.
- Para cada instância i calcule o intervalo $\max_j x_{ij} - \min_j x_{ij}$.
- Designe rank 1 ao problema de menor intervalo, rank 2 ao problema de segundo menor intervalo, etc.
- Sejam $Q_1, Q_2, \ldots, Q_n$ os ranks associados às instâncias $1, 2, \ldots, n$.

Teste de Quade

- Calcule o rank médio entre problemas $(k+1)/2$.
- Calcule $S_{ij} = Q_i \left(r_{ij} - \frac{k+1}{2} \right)$, o tamanho ajustado de cada algoritmo em cada instância.
- Calcule $W_{ij} = Q_i r_{ij}$, o tamanho não ajustado para estabelecer uma comparação com o teste de Friedman.
- Calcule o rank médio do algoritmo j, $T_j = \frac{W_j}{n(n+1)/2}$, em que $W_j = \sum_{i=1}^{j} W_{ij}$.

Ranks de Quade $S_{ij}(W_{ij})$

Quade	A	B	C	D	
P1	1(6)	3(8)	-3(2)	-1(4)	
P2	-4(6)	6(16)	-4(6)	2(12)	←Difícil
P3	-0,5(2)	1,5(4)	0,5(3)	-1,5(1)	←Fácil
P4	1,5 (9)	4,5(12)	-3(4,5)	-3(4,5)	
T_j	2,3	4	1,55	2,15	

Teste de Quade

Erros Relativos

Erro	A	B	C	D
1	2,711	3,147	2,515	2,612
2	7,832	9,828	7,832	7,921
3	0,012	0,532	0,122	0,005
4	3,431	4,111	3,401	3,401

Ranks de Friedman

Friedman	A	B	C	D
1	3	4	1	2
2	1,5	4	1,5	3
3	2	4	3	1
4	3	4	1,5	1,5
Média	2,375	4	1,250	1,875

- $\max_j x_{1j} - \min_j x_{1j} = 3,147 - 2,612 = 0,535$
- $\max_j x_{2j} - \min_j x_{2j} = 9,828 - 7,832 = 1,996$
- $\max_j x_{3j} - \min_j x_{3j} = 0,522 - 0,005 = 0,517$
- $\max_j x_{4j} - \min_j x_{4j} = 4,111 - 3,401 = 0,71$

- $Q_1 = 2;\ Q_2 = 4;\ Q_3 = 1;\ Q_4 = 3$

- $S_{11} = 2(3 - 2,5) = 1;\ W_{11} = 6$
- $S_{12} = 2(4 - 2,5) = 3;\ W_{12} = 8$
- $S_{13} = 2(1 - 2,5) = -3;\ W_{13} = 2$
- $S_{14} = 2(2 - 2,5) = -1;\ W_{11} = 4$

Resultados Computacionais

Table 16
Adjusted p-values for the Friedman test (DE-Exp is the control method).

Friedman	Unadjusted	Bonferroni	Holm	Hochberg	Hommel	Holland	Rom	Finner	Li
PSO	0.000006	0.000050	0.000050	0.000050	0.000050	0.000050	0.000047	0.000050	0.000018
CHC	0.000332	0.002656	0.002324	0.002324	0.002324	0.002322	0.002210	0.001327	0.000978
SSGA	0.009823	0.078586	0.058940	0.058940	0.049116	0.057511	0.056042	0.025981	0.028137
SS-Arit	0.014171	0.113371	0.070857	0.070857	0.070857	0.068877	0.067384	0.028142	0.040093
IPOP-CMA-ES	0.083642	0.669139	0.334569	0.334569	0.282186	0.294885	0.319017	0.130431	0.197766
SS-BLX	0.141093	1.0	0.423278	0.423278	0.423278	0.366366	0.423278	0.183552	0.293707
DE-Bin	0.518605	1.0	1.0	0.660706	0.660706	0.768259	0.660706	0.566345	0.604506
SaDE	0.660706	1.0	1.0	0.660706	0.660706	0.768259	0.660706	0.660706	0.660706

Table 17
Adjusted p-values for the Friedman Aligned test (DE-Exp is the control method).

Friedman Aligned	Unadjusted	Bonferroni	Holm	Hochberg	Hommel	Holland	Rom	Finner	Li
CHC	0.000079	0.000635	0.000635	0.000635	0.000635	0.000635	0.000604	0.000635	0.000907
PSO	0.003300	0.026401	0.023101	0.023101	0.023101	0.022873	0.021963	0.013135	0.036400
SSGA	0.015888	0.127104	0.095328	0.095328	0.095328	0.091621	0.090642	0.041809	0.153880
IPOP-CMA-ES	0.088320	0.706559	0.441599	0.441599	0.353280	0.370186	0.419957	0.168839	0.502727
SS-BLX	0.208043	1.0	0.832172	0.631221	0.624129	0.606625	0.631221	0.311471	0.704264
SS-Arit	0.210407	1.0	0.832172	0.631221	0.631221	0.606625	0.631221	0.311471	0.706612
DE-Bin	0.847534	1.0	1.0	0.912638	0.912638	0.976754	0.912638	0.883457	0.906555
SaDE	0.912638	1.0	1.0	0.912638	0.912638	0.976754	0.912638	0.912638	0.912638

Table 18
Adjusted p-values for the Quade test (DE-Exp is the control method).

Quade	Unadjusted	Bonferroni	Holm	Hochberg	Hommel	Holland	Rom	Finner	Li
CHC	0.021720	0.173762	0.173762	0.173762	0.173762	0.161111	0.165195	0.161111	0.231846
PSO	0.052904	0.423235	0.370330	0.370330	0.369115	0.316471	0.352093	0.195409	0.423683
SSGA	0.118631	0.949049	0.711787	0.711787	0.593156	0.531245	0.676797	0.285908	0.622427
SS-Arit	0.158192	1.0	0.790962	0.790962	0.632769	0.577269	0.752197	0.29136	0.687327
SS-BLX	0.259289	1.0	1.0	0.928037	0.777867	0.69898	0.928037	0.38136	0.782754
IPOP-CMA-ES	0.357754	1.0	1.0	0.928037	0.928037	0.735086	0.928037	0.445882	0.832533
DE-Bin	0.803179	1.0	1.0	0.928037	0.928037	0.961261	0.928037	0.843964	0.917769
SaDE	0.928037	1.0	1.0	0.928037	0.928037	0.961261	0.928037	0.928037	0.928037

4.4. Contrast Estimation

Contrast Estimation based on medians [45,46] can be used to estimate the difference between the performance of two algorithms. It assumes that the expected differences between performances of algorithms are the same across problems. Therefore, the performance of algorithms is reflected by the magnitudes of the differences between them in each domain.

The interest of this test lies in estimating the contrast between medians of samples of results considering all pairwise comparisons. The test obtains a quantitative difference computed through medians between two algorithms over multiple problems, proceeding as follows.

1. For every pair of k algorithms in the experiment, compute the difference between the performances of the two algorithms in each of the n problems. That is, compute the differences

$$D_{i(u,v)} = x_{iu} - x_{iv}, \tag{16}$$

where $i = 1, \ldots, n$; $u = 1, \ldots, k$; $v = 1, \ldots, k$. (Consider only performance pairs where $u < v$.)

2. Find the median of each set of differences (Z_{uv}, which can be regarded as the *unadjusted estimator* of the medians of the algorithms u and v, $M_u - M_v$). Since $Z_{uv} = Z_{vu}$, it is only required to compute Z_{uv} in those cases where $u < v$. Also, note that $Z_{uu} = 0$.

3. Compute the mean of each set of unadjusted medians having the same first subscript, m_u:

$$m_u = \frac{\sum_{j=1}^{k} Z_{uj}}{k}, \quad u = 1, \ldots, k. \tag{17}$$

4. The estimator of $M_u - M_v$ is $m_u - m_v$, where u and v range from 1 through k. For example, the difference between M_1 and M_2 is estimated by $m_1 - m_2$.

These estimators can be understood as an advanced global performance measure. Although this test cannot provide a probability of error associated with the rejection of the null hypothesis of equality, it is especially useful to estimate by how far an algorithm outperforms another one.

An implementation of the Contrast Estimation procedure can be found in the CONTROLTEST package, which can be obtained at the SCI2S thematic public website *Statistical Inference in Computational Intelligence and Data Mining* (see footnote 1).

Example 8. In our experimental analysis, we can compute the set of estimators of medians directly from the average error results. Table 19 shows the estimations computed for each algorithm.

Focusing our attention in the rows of the table, we may highlight the performance of SaDE (all its related estimators are negative;, that is, it achieves very low error rates considering median estimators) and the Scatter Search-based approaches; on the other hand, CHC and PSO achieve higher error rates in our experimental study.

Testes Post-hoc

Table 19
Contrast estimation results. The estimators highlight SaDE, SS-BLX, and SS-Arit as the best performing algorithms.

Estimation	PSO	IPOP-CMA-ES	CHC	SSGA	SS-BLX	SS-Arit	DE-Bin	DE-Exp	SaDE
PSO	0	11.172	−23.671	10.495	24.010	21.150	15.115	17.631	25.035
IPOP-CMA-ES	−11.172	0	−34.843	−0.677	12.838	9.978	3.943	6.459	13.863
CHC	23.671	34.843	0	34.166	47.681	44.821	38.786	41.302	48.706
SSGA	−10.495	0.677	−34.166	0	13.514	10.655	4.620	7.136	14.539
SS-BLX	−24.010	−12.838	−47.681	−13.514	0	−2.859	−8.895	−6.378	1.025
SS-Arit	−21.150	−9.978	−44.821	−10.655	2.859	0	−6.036	−3.519	3.884
DE-Bin	−15.115	−3.943	−38.786	−4.620	8.895	6.036	0	2.516	9.920
DE-Exp	−17.631	−6.459	−41.302	−7.136	6.378	3.519	−2.516	0	7.403
SaDE	−25.035	−13.863	−48.706	−14.539	−1.025	−3.884	−9.920	−7.403	0

5. Multiple comparisons among all methods

Friedman's test is an omnibus test which can be used to carry out these types of comparison. It allows us to detect differences considering the global set of algorithms. Once Friedman's test rejects the null hypothesis, we can proceed with a post-hoc test in order to find the concrete pairwise comparisons which produce differences. In the previous section, we focused on procedures that control the FWER when comparing with a control algorithm, arguing that the objective of a study is to test whether a newly proposed algorithm is better than the existing ones. For this reason, we have described and studied procedures such as the Bonferroni–Dunn, Holm and Hochberg methods.

When our interest lies in carrying out a multiple comparison in which all possible pairwise comparisons need to be computed ($N \times N$ comparison), two classic procedures that can be used are the Holm test (the same as was described in Section 4.3) and the Nemenyi procedure [47]. This procedure adjusts the value of α in a single step by dividing it by the number of comparisons performed, $m = k(k-1)/2$. It is the simplest of this family, but it also has little power.

The hypotheses being tested belonging to a family of all pairwise comparisons are logically interrelated; thus not all combinations of true and false hypotheses are possible. As a simple example of such a situation, suppose that we want to test the three hypotheses of pairwise equality associated with the pairwise comparisons of three algorithms M_i, $i = 1, 2, 3$. It is easily seen from the relations among the hypotheses that, if any one of them is false, at least one other must be false. For example, if M_1 is better/worse than M_2, then it is not possible that M_1 has the same performance as M_3 and M_2 has the same performance as M_3. M_3 must be better/worse than M_1 or M_2 or the two algorithms at the same time. Thus, there cannot be one false and two true hypotheses among these three.

Based on this argument, Shaffer proposed two procedures which make use of the logical relation among the family of hypotheses for adjusting the value of α [48].

- Shaffer's static procedure: following Holm's step-down method, at stage j, instead of rejecting H_i if $p_i \leq \alpha/(m - i + 1)$, reject H_i if $p_i \leq \alpha/t_i$, where t_i is the maximum number of hypotheses which can be true given that any $(i, \ldots, 1)$ hypotheses are false. It is a static procedure; that is, $t_1, \ldots, t_m$ are fully determined for the given hypotheses $H_1, \ldots, H_m$, independent of the observed p-values. The possible numbers of true hypotheses, and thus the values of t_1 can be obtained from the recursive formula

$$S(k) = \bigcup_{j=1}^{k} \left\{ \binom{j}{2} + x : x \in S(k-j) \right\}, \tag{18}$$

where $S(k)$ is the set of possible numbers of true hypotheses with k algorithms being compared, $k \geq 2$, and $S(0) = S(1) = \{0\}$.

Input: $C = \{c_1, c_2, \ldots, c_k\}$:list of classifiers
Let $E = \emptyset$
$E = E \cup$ {set of all possible and distinct pairwise comparisons using C}
if $E == \emptyset$ **then**
  return E
foreach *possible divisions of C into two subsets C_1 and C_2, $c_k \in C_2$ and $C_1 \neq \emptyset$* **do**
  E_1 =obtainExhaustive(C_1)
  E_2 =obtainExhaustive(C_2)
  $E = E \cup E_1$
  $E = E \cup E_2$
  foreach *family of hypotheses e_1 of E_1* **do**
    foreach *family of hypotheses e_2 of E_2* **do**
      $E = E \cup (e_1 \cup e_2)$
return E

Fig. 2. obtainExhaustive(C). Algorithm for obtaining all exhaustive sets in Bergmann's procedure.

- Shaffer's dynamic procedure: this increases the power of the first by substituting $\alpha = t_i$ at stage i by the value $\alpha = t_i^*$, where t_i^* is the maximum number of hypotheses that could be true, given that the previous hypotheses are false. It is a dynamic procedure, since t_i^* depends not only on the logical structure of the hypotheses, but also on the hypotheses already rejected at step i. Obviously, this procedure has more power than the first one. However, we will not use this second procedure, given that it is included in an advanced procedure which we will describe in the following.

In [49], a procedure was proposed based on the idea of finding all elementary hypotheses which cannot be rejected. In order to formulate Bergmann–Hommel's procedure, we need the following definition.

Definition 1. An index set of hypotheses $I \subseteq \{1, \ldots, m\}$ is called exhaustive if exactly all $H_j, j \in I$, could be true.

Under this definition, the Bergmann–Hommel procedure works as follows.

- Bergmann and Hommel procedure: reject all H_j with $j \notin A$, where the acceptance set A, given as

$$A = \bigcup \{I : I \text{ exhaustive }, min\{P_i : i \in I\} > \alpha/|I|\}, \tag{19}$$

is the index set of null hypotheses which are retained.

For this procedure, one has to check for each subset I of $\{1, \ldots, m\}$ if I is exhaustive, which leads to intensive computation. Due to this fact, we will obtain a set, named E, which will contain all the possible exhaustive sets of hypotheses for a certain comparison. A rapid algorithm which was described in [50] allows a substantial reduction in computing time. Once the E set is obtained, the hypotheses that do not belong to the A set are rejected.

Fig. 2 shows a valid algorithm for obtaining all the exhaustive sets of hypotheses, using as input a list of algorithms C. E is a set

Testes Post-hoc

of families of hypotheses; likewise, a family of hypotheses is a set of hypotheses. The most important step in the algorithm is step 6. It performs a division of the algorithms into two subsets, in which the last algorithm k always is inserted in the second subset and the first subset cannot be empty. In this way, we ensure that a subset yielded in a division is never empty and no repetitions are produced.

Finally, we will explain how to compute the APVs for the three post-hoc procedures described above, following the indications given in [51].

- Nemenyi APV_i : $\min\{v : 1\}$, where $v = m \cdot p_i$.
- Holm APV_i (using it in all pairwise comparisons): $\min\{v : 1\}$, where $v = \max\{(m-j+1)p_j : 1 \leq j \leq i\}$.
- Shaffer static APV_i : $\min\{v : 1\}$, where $v = \max\{t_j p_j : 1 \leq j \leq i\}$.
- Bergmann–Hommel APV_i : $\min\{v : 1\}$, where $v = \max\{\|I\| \cdot min\{p_j, j \in I\} : I$ exhaustive ; $i \in I\}$.

where m is the number of possible comparisons in an all pairwise comparisons design; that is, $m = \frac{k \cdot (k-1)}{2}$.

An implementation of the Friedman Test for multiple comparisons, with all its related post-hoc procedures, can be found in the MULTIPLETEST package, which can be obtained at the SCI2S thematic public website *Statistical Inference in Computational Intelligence and Data Mining* (see footnote 1).

Example 9. Starting from the analysis performed by the Friedman test over our experimental results (see Example 7), we can raise the 36 hypotheses of equality among the 9 algorithms of our study, and apply the above-mentioned methods to contrast them. Table 20 lists all the hypotheses and the p-values achieved.

Using a level of significance $\alpha = 0.1$, only six hypotheses are rejected by the Nemenyi method. These hypotheses show the improvement of DE-Exp and SaDE over PSO and CHC, and DE-Bin and SS-BLX over PSO. The Holm and Shaffer methods reject an additional hypothesis, thus confirming the improvement of DE-Bin over CHC. Finally, the Bergmann procedure rejects eight hypotheses, the last one being the equality between PSO and IPOP-CMA-ES. None of the remaining 28 hypotheses can be rejected using these procedures.

6. Considerations and recommendations on the use of nonparametric tests

This section notes some considerations and recommendations concerning the nonparametric tests presented in this tutorial. Their characteristics as well as suggestions on some of their aspects and details of the multiple comparisons tests are presented. With this aim, some general considerations and recommendations are given first (Section 6.1). Then, some advanced guidelines for multiple comparisons with a control method (Section 6.2) and multiple comparisons among all methods (Section 6.3) are provided.

6.1. General considerations

- By using nonparametric statistical procedures, it is possible to analyze any unary performance measure (that is, associated to a single algorithm) with a defined range. This range does not have to be limited; thus, comparisons considering running times, memory requirements, and so on, are feasible.
- Being able to be applied in multi-domain comparisons, nonparametric statistical procedures can compare both deterministic and stochastic algorithms simultaneously, providing that their results are represented as a sample for each pair of algorithm/domain.
- For the application of these methods, only a result for each pair of algorithm/domain is required. A known and standardized procedure must be followed to gather them, using average results from several executions when considering stochastic algorithms.
- An appropriate number of algorithms in contrast with an appropriate number of case problems are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of case problems. The previous statement may not be true for the Wilcoxon test. The influence of the number of case problems used is more noticeable in multiple comparison procedures than in Wilcoxon's test [2,3].
- Although Wilcoxon's test and the post-hoc tests for multiple comparisons are nonparametric statistical tests, they operate in a different way. The main difference lies in the computation of the ranking. Wilcoxon's test computes a ranking based on differences between case problems independently, whereas the Friedman test and its derivative procedures compute the ranking between algorithms [2,3].
- In relation to the sample size (number of case problems when performing Wilcoxon's or Friedman's tests in a multi-problem analysis), there are two main aspects to be determined. First, the minimum sample size considered acceptable for each test needs to be stipulated. There is no established agreement about this specification. Statisticians have studied the minimum sample size when a certain power of the statistical test is expected. In our case, the employment of a sample size as large as possible is preferable because the power of the statistical tests (defined as the probability that the test will reject a false null hypothesis) will increase. Moreover, in a multi-problem analysis, the increase of the sample size depends on the availability of new case problems (which should be well known in computational intelligence or data mining field). Second, we have to study how the results are expected to vary if there were a larger sample size available. In all statistical tests used for comparing two or more samples, an increase of the sample size benefits the power of the test. In the following items, we will state that Wilcoxon's test is less influenced by this factor than Friedman's test. Finally, as a rule of thumb, the number of case problems in a study should be $n = a \cdot k$, where $a \geq 2$ [2,3].
- Although there is not a theoretical maximum number of domains to use in a comparison, it can be derived from the central limit theorem that, if this number is too high, the results may be unreliable. If the number of domains grows too much, statistical tests can lose credibility, as they may start highlighting true insignificant hypotheses as significant ones. For the Wilcoxon's test, a maximum of 30 domains is suggested [4]. For multiple comparisons, a value of $n \geq 8 \cdot k$ could be too high, obtaining no significant comparisons as a result [2,3].
- Taking into account the previous observation and knowing the operations performed by the nonparametric tests, we can deduce that Wilcoxon's test is influenced by the number of case problems used. On the other hand, both the number of algorithms and case problems are crucial when we refer to multiple comparisons tests (such as Friedman's test), given that all the critical values depend on the value of n (see the expressions above). However, the increasing/decreasing of the number of case problems rarely affects the computation of the ranking. In these procedures, the number of functions used is an important factor to be considered when we want to control the FWER [2,3].
- Another interesting procedure considered in this paper is related to Contrast Estimation based on medians between two samples of results. Contrast Estimation in nonparametric statistics is used for computing the real differences between two algorithms, considering the median measure the most important.

Testes Post-hoc

Table 20
Adjusted p-values for tests for multiple comparisons among all methods.

i	Hypothesis	Unadjusted p	Nemenyi	Holm	Shaffer	Bergmann
1	PSO versus DE-Exp	0.000006	0.000224	0.000224	0.000224	0.000224
2	PSO versus SaDE	0.000045	0.001624	0.001579	0.001263	0.001263
3	PSO versus DE-Bin	0.000108	0.00387	0.003655	0.00301	0.002365
4	CHC versus DE-Exp	0.000332	0.011952	0.010956	0.009296	0.009296
5	CHC versus SaDE	0.001633	0.058772	0.052242	0.045712	0.034284
6	PSO versus SS-BLX	0.002313	0.08328	0.071713	0.064773	0.04164
7	CHC versus DE-Bin	0.003246	0.116841	0.097367	0.090876	0.051929
8	PSO versus IPOP-CMA-ES	0.005294	0.190602	0.15354	0.148246	0.095301
9	SSGA versus DE-Exp	0.009823	0.353638	0.275052	0.275052	0.216112
10	SS-Arit versus DE-Exp	0.014171	0.51017	0.382627	0.311771	0.255085
11	SSGA versus SaDE	0.032109	1.0	0.834835	0.706398	0.513744
12	CHC versus SS-BLX	0.03424	1.0	0.856006	0.753286	0.513744
13	PSO versus SS-Arit	0.038867	1.0	1.0	0.855076	0.621874
14	SS-Arit versus SaDE	0.044015	1.0	1.0	0.968322	0.621874
15	SSGA versus DE-Bin	0.052808	1.0	1.0	1.0	0.63369
16	PSO versus SSGA	0.052808	1.0	1.0	1.0	0.686498
17	IPOP-CMA-ES versus CHC	0.063023	1.0	1.0	1.0	0.756271
18	SS-Arit versus DE-Bin	0.070701	1.0	1.0	1.0	0.756271
19	IPOP-CMA-ES versus DE-Exp	0.083642	1.0	1.0	1.0	1.0
20	SS-BLX versus DE-Exp	0.141093	1.0	1.0	1.0	1.0
21	IPOP-CMA-ES versus SaDE	0.196706	1.0	1.0	1.0	1.0
22	CHC versus SS-Arit	0.255925	1.0	1.0	1.0	1.0
23	SSGA versus SS-BLX	0.266889	1.0	1.0	1.0	1.0
24	IPOP-CMA-ES versus DE-Bin	0.278172	1.0	1.0	1.0	1.0
25	SS-BLX versus SaDE	0.3017	1.0	1.0	1.0	1.0
26	CHC versus SSGA	0.313946	1.0	1.0	1.0	1.0
27	SS-BLX versus SS-Arit	0.326516	1.0	1.0	1.0	1.0
28	PSO versus CHC	0.352622	1.0	1.0	1.0	1.0
29	IPOP-CMA-ES versus SSGA	0.394183	1.0	1.0	1.0	1.0
30	SS-BLX versus DE-Bin	0.40867	1.0	1.0	1.0	1.0
31	IPOP-CMA-ES versus SS-Arit	0.469706	1.0	1.0	1.0	1.0
32	DE-Bin versus DE-Exp	0.518605	1.0	1.0	1.0	1.0
33	DE-Exp versus SaDE	0.660706	1.0	1.0	1.0	1.0
34	IPOP-CMA-ES versus SS-BLX	0.796253	1.0	1.0	1.0	1.0
35	DE-Bin versus SaDE	0.836354	1.0	1.0	1.0	1.0
36	SSGA versus SS-Arit	0.897279	1.0	1.0	1.0	1.0

Taking into account that the samples of results in computational intelligence experiments rarely fulfill the needed conditions for a safe use of parametric tests, the computation of nonparametric contrast estimation through the use of medians is very useful. For example, one could provide, apart from the average values of accuracies over various problems reported by the methods compared, the contrast estimation between them over multiple problems, which is a safer metric in multi-problem environments [46].

- Finally, we want to remark that the choice of any of the statistical procedures presented in this paper for conducting an experimental analysis should be justified by the researcher. The use of the most powerful procedures does not imply that the results obtained by a given proposal will be better. The choice of a statistical technique is ruled by a trade-off between its power and its complexity when it comes to being used or explained to non-expert readers in statistics [46].

6.2. Multiple comparisons with a control method

- A multiple comparison of various algorithms must be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. Once this test rejects the hypothesis of equivalence of means, the detection of the concrete differences among the algorithms can be done with the application of post-hoc statistical procedures, which are methods used for comparing a control algorithm with two or more algorithms [2,3].
- An appropriate number of algorithms in contrast with an appropriate number of case problems are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of case problems. In general, p-values are lower on increasing the number of case problems used in multiple comparison procedures (so long as this number does not exceed $n \geq 8 \cdot k$); therefore, the differences among the algorithms are more detectable [2,3].
- As we have suggested, multiple comparisons tests must be used when we want to establish a statistical comparison of the results reported among various algorithms. We focus on cases when a method is compared against a set of algorithms. It could be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. There are three alternatives: the Friedman test with the Iman–Davenport extension, the Friedman Aligned Ranks test, and the Quade test. Once one of these tests rejects the hypothesis of equivalence of medians, the detection of the specific differences among the algorithms can be made with the application of post-hoc statistical procedures, which are methods used for specifically comparing a control algorithm with two or more algorithms [46].
- In this kind of test, it is possible to use just the rankings obtained when establishing a classification between the algorithms, and even employ them to measure their performance differences. However, this cannot be used to conclude that a given proposal outperform the rest, unless the null hypothesis is rejected.
- Although, by definition, post-hoc statistical procedures can be applied in an independent way from the rejection of the null hypothesis, it is advisable to check this rejection firstly.