UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

U-ESTATÍSTICAS: EXEMPLOS E APLICAÇÕES

AMEÓFIS DE PAULA VALE

Porto Alegre
2017

U-ESTATÍSTICAS: EXEMPLOS E APLICAÇÕES

AMEÓFIS DE PAULA VALE

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharelado em Estatística

Professor Dr. Marcio Valk (orientador)

Porto Alegre
2017

Agradecimentos

Agradeço a todos aqueles que de alguma forma contribuíram à realização desse trabalho e em especial à minha mãe Dorvalina, minha mulher Silvana e ao professor Marcio Valk. Também, não poderia deixar de agradecer a Deus por me oferecer força e coragem nos momentos de privação durante a graduação.

"No futuro, saiba que toda situação está sujeita a reviravoltas e que tudo o que acontece com qualquer pessoa também pode acontecer contigo".

(Sêneca)

Resumo

É notável que atualmente existe uma demanda por técnicas estatísticas que sejam menos restritivas, ou seja, que possuam uma menor exigência a respeito do comportamento dos dados. Por esse motivo, os métodos não-paramétricos sempre terão grande importância no contexto da Inferência Estatística, por justamente não necessitarem de algumas suposições a respeito da distribuição dos dados. Nesse cenário, a teoria de U-Estatísticas vem enriquecer o universo da Inferência Estatística, emprestando suas propriedades e facilitando a obtenção de resultados teóricos. Aqui, são apresentados conceitos importantes sobre as U-Estatísticas, exemplificando situações inicialmente mais complexas. Também, são introduzidos alguns teoremas que descrevem o comportamento das U-Estatísticas tanto num cenário de amostras finitas quanto em situações assintóticas. Apresenta-se, também, alguns procedimentos não-paramétricos para testes de hipóteses juntamente com sua versão U-Estatística. Explora-se o método dos Mínimos Quadrados Ordinários em sua versão tradicional e na versão U-Estatística. Estuda-se a construção de Intervalos de Confiança para os parâmetros do modelo de regressão em situações em que os dados não atendem as suposições do método de Mínimos Quadrados Ordinários. Com um procedimento *Bootstrap*, é possível corrigir o percentual de cobertura. O *software R* foi utilizando no desenvolvimento desse trabalho.

Palavras-chave: U-Estatísticas, não-paramétricos, *Bootstrap* e Mínimos Quadrados Ordinários.

Abstract

It is notable nowadays that there is a demand for statistical techniques that are less restrictive, that is, that they have a lower requirement about data behavior. For this reason, non-parametric methods will always have great importance in the context of Statistical Inference, because they do not need some assumptions about the data distribution. In this scenario, the U-Statistics theory enriches the universe of Statistical Inference, lending its properties and facilitating the achievement of theoretical results. Here, important concepts about U-Statistics are presented, exemplifying initially more complex situations. Also are introduced some theorems that describe the behavior of U-Statistics in both cases, a finite-sample scenario and asymptotic scenarios. It also presented some non-parametric procedures for testing hypotheses with its U-Statistics version. The Ordinary Least Squares method is explored in its traditional version and in the U-Statistics version. We study the construction of confidence intervals for the parameters of the regression model in situations where the data do not follow the assumptions of the method of Ordinary Least Squares. With a *Bootstrap* procedure is possible to correct the coverage percentage. The software R was used in the development of this work.

Keywords: U-Statistics, Non-Parametric, Bootstrap and Ordinary Least Squares.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Na Inferência Estatística, as ferramentas utilizadas dependem, em geral, de muitas suposições, tais como normalidade dos dados e independência das observações. Neste caso, é sabido que a obtenção de um Estimador Não-Viesado de Variância Uniformemente Mínima (ENVVUM) se torna mais difícil quando alguns desses pressupostos não são válidos. Além do mais, propriedades como a distribuição dos estimadores não são simples de se obter e, muitas vezes, nem são acessíveis. Uma das ferramentas úteis, que se pode utilizar nesses casos, são as U-Estatísticas. Contudo, sua aplicação é muito mais ampla, pois segundo (Lee, 1990): "A classe de U-Estatísticas é importante pelo menos por três razões.

Primeiro, uma grande variedade de estatísticas de uso comum são realmente membros dessa classe, de modo que a teoria fornece um modelo unificado para o estudo das propriedades distributivas de muitos testes estatísticos bem conhecidos e estimadores, particularmente no campo da Estatística Não-Paramétrica. Segundo, a estrutura simples das U-Estatísticas torna-as ideal para estudar processo de estimação em geral, tais como *Bootstrap* e *Jacknife*, e para generalizar aquelas partes da teoria assintótica preocupada com comportamento da sequências de médias amostrais. Terceiro, a aplicação da teoria frequentemente gera novas estatísticas úteis na prática para problemas de estimação".

O objetivo desse trabalho é apresentar o conceito de U-Estatísticas, demonstrando a aplicação de teoremas na obtenção de resultados importantes para a Inferência, desmitificando a teoria que não é vista na graduação e discutindo sua validade como ferramenta da Estatística/Matemática. No Capítulo 2, traça-se um comparativo entre os testes paramétricos e não-paramétricos, abordando as vantagens e desvantagens de alguns desses testes de forma genérica. No Capítulo 3, são apresentados conceitos importantes de estimação onde discorre-se um pouco sobre a definição de estimadores e suas propriedades. No Ca-

pítulo 4, mostra-se a relação das U-Estatísticas com os testes não-paramétricos e outros conceitos, tais como Método dos Momentos e Estimadores de Covariância. No Capítulo 5, é explorado o Método dos Mínimos Quadrados Ordinários e estuda-se a construção de Intervalos de Confiança para os parâmetros do modelo de regressão em situações em que os dados não atendem as suposições do método de Mínimos Quadrados Ordinários. Com um procedimento *Bootstrap* é possível corrigir o percentual de cobertura. O *software R* foi utilizando no desenvolvimento desse trabalho.

2 Estatística Paramétrica e Estatística Não-Paramétrica

Nesse capítulo, é feito um comparativo entre testes paramétricos e não-paramétricos, explorando as vantagens e desvantagens deles com alguns exemplos.

Embora, sejam amplamente utilizados os termos "Estatística Não-Paramétrica" e "Estatística Paramétrica", os mesmos podem não ser de simples definição. Segundo (Hoskin, 2014): "Uma definição precisa e universalmente aceitável do termo "não-paramétrico" não está atualmente disponível". Geralmente é mais fácil listar exemplos de cada tipo de procedimento (paramétrico e não-paramétrico) do que definir os termos em si. No entanto, para fins mais práticos, pode-se definir procedimentos estatísticos não-paramétricos como uma classe de procedimentos estatísticos que não se baseiam em suposições sobre a forma ou configuração da distribuição de probabilidade, a partir da qual os dados foram coletados. Esse campo da Estatística existe porque geralmente é impossível coletar dados de todos os indivíduos de interesse (população). A única solução é coletar dados de um subconjunto (amostra) dos indivíduos de interesse, mas o verdadeiro desejo é conhecer a "verdade" sobre a população.

Quantidades como médias, desvios-padrão e proporções são todas valores importantes e são chamados de "parâmetros" quando fala-se de uma população. Uma vez que normalmente não é possível obter dados de toda a população, não se pode conhecer os valores dos parâmetros para essa população. Pode-se, no entanto, calcular estimativas destas quantidades para uma amostra. Quando são calculadas a partir de dados da amostra, estas são chamadas "estatísticas". Os procedimentos estatísticos paramétricos baseiam-se em suposições sobre a forma da distribuição (assumir uma distribuição normal, por Ex.) na população subjacente e sobre o parâmetro (médias e desvios-padrão, por Ex.) da distribuição assumida. Os procedimentos estatísticos não-paramétricos baseiam-se em nenhuma ou em poucas hipóteses sobre os parâmetros da distribuição da população, a

partir da qual a amostra foi selecionada.

2.1 Testes paramétricos e procedimentos não-paramétricos análogos

Conforme já mencionado em (Hoskin, 2014), às vezes é mais fácil listar exemplos de cada tipo de procedimento do que definir os termos. A Tabela 2.1 contém os nomes de vários procedimentos estatísticos categorizados cada um como paramétrico ou não-paramétrico. Todos os paramétricos listados, na Tabela 2.1, baseiam-se em uma suposição de normalidade aproximada dos dados da amostra.

Tabela 2.1: Abordagens paramétrica e não-paramétrica para um mesmo problema.

Tipo de análise	Exemplo	Procedimento paramétrico	Procedimento não-paramétrico
Comparar as médias entre dois grupos distintos / independentes	A pressão arterial sistólica média (no início do estudo) para os pacientes com placebo é diferente da média para os pacientes do grupo de tratamento?	Teste t de duas amostras	Teste da soma-Wilcoxon
Comparar duas medidas tomadas do mesmo indivíduo	Houve uma alteração significativa na pressão arterial sistólica entre o início e os seis meses seguintes na medição do grupo de tratamento?	Teste t pareado	Teste de sinais-Wilcoxon
Comparar as médias entre três ou mais grupos distintos / independentes	Se o experimento tivesse três grupos (por exemplo: placebo, novo fármaco 1 e nova droga 2), pode-se querer saber se a pressão arterial sistólica medida no início do tratamento diferiu entre os três grupos.	Análise de Variância (ANOVA)	Teste de Kruskal-Wallis
Estimar o grau da associação entre duas variáveis quantitativas	A pressão arterial sistólica está associada a dor do paciente?	Coeficiente de correlação de Pearson	Coeficiente de correlação de Spearman

Fonte: (Hoskin, 2014)

Exemplo 2.1.1. Suponha que se tenha uma amostra de pacientes criticamente doentes. A amostra contém 20 pacientes do sexo feminino e 19 do sexo masculino. A variável

de interesse é a Duração da Estada Hospitalar (LOS) em dias e deseja-se comparar as mulheres e os homens. Os histogramas da variável LOS para homens e mulheres aparecem na figura 2.1. Vê-se que a distribuição para as mulheres tem uma forte inclinação para a direita. Observa-se que a média para as mulheres é de 60 dias, enquanto a mediana é de 31,5 dias. Para os homens, a distribuição é mais simétrica com média e mediana de 30,9 dias e 30 dias, respectivamente. Comparando os dois grupos, suas medianas são bastante semelhantes, mas suas médias são muito diferentes. Este é um caso em que a suposição de normalidade associada a um teste paramétrico provavelmente não é razoável. Um procedimento não-paramétrico seria mais apropriado.

Esta é a situação listada na primeira linha da Tabela 2.1 - comparar as médias entre dois grupos distintos. Assim, o procedimento não-paramétrico apropriado é o Teste soma-Wilcoxon. Este teste dá um p-valor de 0.63, e como é maior que 0.05, é uma indicação de um resultado não estatisticamente significativo. Assim, não há diferença significativa entre os sexos em relação à duração da estada com base no Teste da soma-Wilcoxon. Embora, os testes não-paramétricos tenham propriedades muito desejáveis, como a de fazer menos suposições sobre a distribuição de medidas na população da qual foi extraída a amostra, eles têm dois principais inconvenientes.

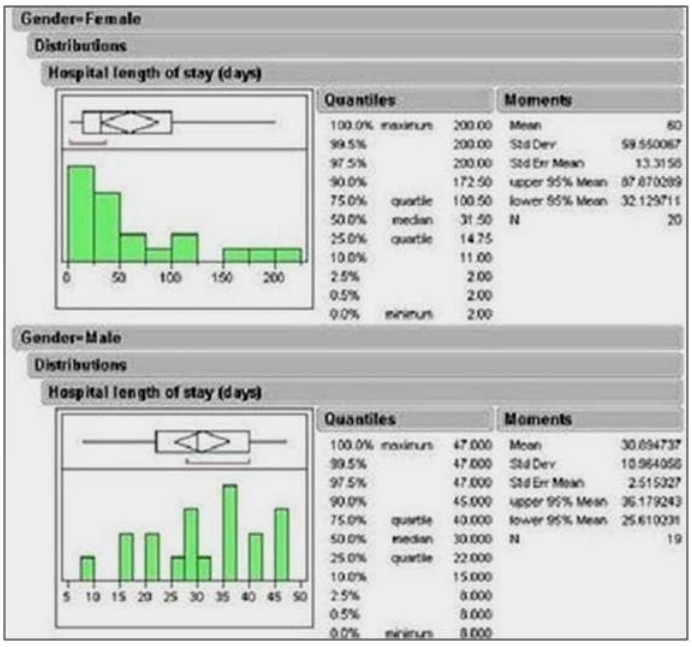

Figura 2.1: Histogramas da variável LOS para homens e mulheres. Fonte: Departamento Clínico de Ciências da Saúde de Mayo.

O primeiro é que eles geralmente são procedimentos estatísticos com menor poder

que procedimentos paramétricos análogos quando os dados são verdadeiramente normais. "Menor poder" significa que há uma menor probabilidade de que o procedimento nos diga que duas variáveis estão associadas umas com as outras quando estão realmente associadas. Ao se planejar um estudo na área da saúde, para tentar determinar quantos pacientes incluir, um teste não-paramétrico exigirá um tamanho de amostra um pouco maior para ter o mesmo poder que o teste paramétrico correspondente. A segunda desvantagem associada a testes não-paramétricos é que seus resultados são frequentemente mais difíceis de interpretar do que os resultados de testes paramétricos. Muitos testes não-paramétricos usam posições dos valores nos dados (estatística de ordem) ao invés de usar os dados reais. Saber, por exemplo, que a diferença entre as posições médias entre os dois grupos é de cinco não ajuda realmente a compreensão intuitiva dos dados. Por outro lado, saber que a pressão arterial sistólica média dos pacientes, que tomam o novo fármaco, era de 5 mmHg menor do que a pressão arterial sistólica média dos pacientes no tratamento padrão, é um tanto intuitivo e útil. Em suma, os procedimentos não-paramétricos são úteis em muitos casos e necessários em alguns, mas eles não são uma solução perfeita. Segundo *(Wikipédia, 2017a)*, entre os testes não-paramétricos mais usados, pode-se citar:

- Teste Anderson-Darling: testa se uma amostra é extraída de uma dada distribuição.
- Método *Bootstrap*: estima a precisão/distribuição de amostragem de uma estatística.
- Teste de Friedman: testa se k tratamentos em blocos casualizados têm efeitos idênticos.
- Kaplan-Meier: estima a função de sobrevivência a partir dos dados de vida, modelando a censura.
- O Tau de Kendall: mede a dependência estatística entre duas variáveis.
- Teste de Kolmogorov-Smirnov: verifica se uma amostra é extraída de uma dada distribuição ou se duas amostras são extraídas da mesma distribuição.
- Teste de Kruskal-Wallis: verifica se mais que duas amostras independentes são extraídas da mesma distribuição.

- Teste de Kuiper: verifica se uma amostra é extraída de uma dada distribuição, sensível a variações cíclicas como o dia da semana.

- Teste de logrank: compara as distribuições de sobrevivência de duas amostras censuradas à direita.

- Teste de Mann-Whitney ou Wilcoxon: verifica se duas amostras são extraídas da mesma distribuição, em comparação com uma dada hipótese alternativa.

- Teste de McNemar: verifica se, em tabelas de contingência 2×2 com um traço dicotômico e pares de indivíduos emparelhados, as frequências marginais de linha e coluna são iguais.

- Teste da mediana: testa se duas amostras são tiradas de distribuições com medianas iguais.

- Teste de sinal: testa se as amostras de pares correspondentes são extraídas de distribuições com medianas iguais.

- O coeficiente de correlação de Spearman: mede a dependência estatística entre duas variáveis usando uma função monotônica.

- Teste para homogeneidade de variâncias: testa igualdade de variância em duas ou mais amostras.

Algumas considerações, a respeito dessas técnicas, são feitas em (Hoskin, 2014):

- Paramétrico e não-paramétrico são duas grandes classificações de procedimentos estatísticos.

- Os testes paramétricos são baseados em suposições sobre a distribuição da população subjacente de onde a amostra foi colhida.

- Os testes não-paramétricos não se baseiam em suposições sobre a forma ou parâmetros da distribuição populacional subjacente.

- Se os dados se desviam fortemente dos pressupostos de um procedimento paramétrico, usando o procedimento paramétrico pode-se chegar a conclusões incorretas.

- O pressuposto paramétrico da normalidade é particularmente preocupante para amostras pequenas ($n < 30$). Testes não-paramétricos são muitas vezes uma boa opção para esses dados.

- Pode ser difícil decidir se deve-se usar um procedimento paramétrico ou não-paramétrico em alguns casos.

- Os procedimentos não-paramétricos geralmente têm menos poder para a mesma amostra que o procedimento paramétrico correspondente.

- A interpretação de procedimentos não-paramétricos também pode ser mais difícil que os procedimentos paramétricos.

Ainda, segundo (Callegari-Jacques, 2009), outra pressuposição nos testes clássicos é a homogeneidade de variâncias entre as populações que estão sendo comparadas. No entanto, muitas vezes as variâncias são heterogêneas e, mesmo transformando os dados, não se consegue homocedasticidade. Sendo assim, os testes não-paramétricos apresentam as seguintes vantagens em relação às técnicas clássicas:

- São os mais apropriados quando não se conhece os dados da população. São também úteis quando essa distribuição é assimétrica e não se deseja realizar uma transformação dos dados, quando existe heterogeneidade nas variâncias, ou ainda quando, na comparação entre tratamentos, a disposição é gaussiana em alguns grupos e assimétrica em outros. São, por isso, testes de aplicação mais ampla do que os paramétricos.

- São os indicados quando a variável é medida em escala ordinal. Também existem técnicas não-paramétricas para variáveis cujas categorias não são ordenáveis.

- Quando as exigências das técnicas clássicas não podem ser satisfeitas, os métodos não-paramétricos são mais eficientes do que os testes paramétricos (nas situações em que tais exigências são satisfeitas, os paramétricos são mais eficientes).

As desvantagens dos testes não-paramétricos são:

- Quando utilizados em dados que satisfazem as exigências dos testes clássicos, os métodos não-paramétricos apresentam uma eficiência menor. Isto equivale a dizer

que para se detectar uma diferença real entre duas populações por um teste não-paramétrico, o tamanho amostral deve ser um pouco maior do que seria necessário com um teste clássico. Por exemplo, em amostras de tamanho moderado, o Teste de Wilcoxon-Mann-Whitney (WMW) tem um poder de aproximadamente 95% quando comparado com o Teste t de Student. Assim, se o tamanho da amostra necessário para identificar uma diferença usando o Teste de WMW é de 100 indivíduos, usando-se o Teste t são necessários 95 indivíduos.

- Alguns autores afirmam que os testes não-paramétricos extraem menos informação do experimento porque são técnicas empregadas em dados mensurados em escalas não quantitativas (ou dados quantitativos reduzidos para uma escala qualitativa ordenável). Realmente, em muitos testes não-paramétricos o valor real medido é substituído pelo posto ocupado na ordenação dos valores obtidos; neste caso, há perda de informação relativa à variabilidade da característica (uma diferença numericamente grande pode representar apenas uma mudança para o posto seguinte).

- Uma análise não-paramétrica pode vir a ser uma operação tediosa, embora simples, se a quantidade de dados for grande. Tal problema não existe, caso se disponha de um computador que realize análises não-paramétricas.

3 Alguns aspectos da estimação

No Capítulo anterior, discutiu-se a diferença entre as abordagens paramétricas e não-paramétricas para inferência sobre parâmetros populacionais. Um aspecto primordial na Inferência é a estimação dos parâmetros. Nesse Capítulo, são apresentados conceitos e propriedades importantes dos estimadores e, em especial, a definição de Estimador Não-Viesado de Variância Uniformemente Mínima (ENVVUM).

3.1 Propriedades dos estimadores

Nesta seção, apresentamos algumas definições e propriedades de estimadores, uma vez que U-Estatísticas são estimadores com propriedades desejáveis no contexto de estimação. As definições e resultados apresentados na sequência são baseados em (Bussab and Morettin, 2010) e (Bolfarine and Sandoval, 2001).

Considere uma amostra $(X_1, X_2, \ldots, X_n)$ de uma v.a. que descreve uma característica de interesse de uma população. Seja θ um parâmetro que deseja-se estimar, como por exemplo a média $\mu = \mathbb{E}(X)$ ou a variância $\sigma^2 = \text{Var}(X)$.

Definição 3.1.1. Um estimador T do parâmetro θ é qualquer função das observações da amostra, ou seja, $T = g(X_1, \ldots, X_n)$. As principais qualidades de um estimador devem ser:

- Ausência de vício (estimador não-viciado)
- Consistência (estimador consistente)
- Eficiência (estimador de variância mínima)
- Suficiência (estimador suficiente)

Definição 3.1.2. O estimador T é não-viesado para θ se

$$\mathbb{E}(T) = \theta, \tag{3.1}$$

para todo θ.

Se (3.1) não valer, T diz-se viesado e a diferença $V(T) = \mathbb{E}(T) - \theta$ é chamado o viés de T.

Definição 3.1.3. Estimativa é o valor assumido pelo estimador em uma particular amostra.

Definição 3.1.4. Uma seqüência $\{T_n\}$ de estimadores de um parâmetro θ é consistente se, para todo $\varepsilon > 0$,

$$P\{|T_n - \theta| > \varepsilon\} \longrightarrow 0, \text{ quando } n \longrightarrow \infty.$$

Proposição 3.1.1. Uma seqüência $\{T_n\}$ de estimadores de θ é consistente se:

$$\lim_{n\to\infty} \mathbb{E}(T_n) = \theta$$

e

$$\lim_{n\to\infty} \text{Var}(T_n) = 0.$$

Definição 3.1.5. Se T e T' são dois estimadores não-viesados de um mesmo parâmetro θ, e ainda

$$Var(T) < Var(T'),$$

então T diz-se mais eficiente do que T'.

Definição 3.1.6. Dizemos que a estatística $T = T(X_1, \ldots, X_n)$ é suficiente para θ, quando a distribuição condicional de $X_1, \ldots, X_n$ dado T for independente de θ. Ou seja, a estatística a ser considerada deve, dentro do possível, conter toda a informação sobre θ presente na amostra. Em outras palavras, se pudermos usar uma estatística $T = T(X_1, \ldots, X_n)$ para extrairmos toda informação que a amostra $X_1, \ldots, X_n$ contem sobre θ, então dizemos que T (que pode ser um vetor) é suficiente para θ. Desse modo, o conhecimento apenas de T (e não necessariamente da amostra completa $X_1, \ldots, X_n$) é suficiente para que sejam feitas inferências sobre θ.

Definição 3.1.7. (Estimador Não-Viesado de Variância Uniformemente Mínima (ENVVUM)). Sejam $X_1, \ldots, X_n$ uma amostra aleatória da distribuição da variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$. Um estimador $T = t(X_1, \ldots, X_n)$ de $g(\theta)$ é dito ser um Estimador Não-Viesado de Variância Uniformemente Mínima de $g(\theta)$ se, somente se

1. $\mathbb{E}(T) = g(\theta)$, isto é um estimador não-viesado para $g(\theta)$, e

2. $\text{Var}(T) \leq \text{Var}(U)$ para qualquer outro estimador U não-viesado de $g(\theta)$

Definição 3.1.8. Chama-se Erro Quadrático Médio (EQM) do estimador T ao valor

$$EQM(T;\theta) = \mathbb{E}(e^2) = \mathbb{E}(T-\theta)^2, \tag{3.2}$$

onde $e = T - \theta$ é o erro amostral cometido ao estimar parâmetro θ da distribuição da v.a X pelo estimador $T = g(X_1, \ldots, X_n)$ baseado na amostra $(X_1, \ldots, X_n)$.

De (3.2) tem-se:

$$\begin{aligned} EQM(T;\theta) &= \mathbb{E}(T - \mathbb{E}(T) + \mathbb{E}(T) - \theta)^2 \\ &= \mathbb{E}(T - \mathbb{E}(T))^2 + 2\mathbb{E}[(T - \mathbb{E}(T))(\mathbb{E}(T) - \theta)] + \mathbb{E}(\mathbb{E}(T) - \theta)^2 \\ &= \mathbb{E}(T - \mathbb{E}(T))^2 + \mathbb{E}(\mathbb{E}(T) - \theta)^2 \\ &= \text{Var}(T) + V(T)^2. \end{aligned} \tag{3.3}$$

onde $\mathbb{E}(T) - \theta$ é uma constante, $\mathbb{E}(T - \mathbb{E}(T)) = 0$ e $V(T) = \mathbb{E}(T) - \theta$ é o viés de T.

Definição 3.1.9. Quantidade Pivotal.
Uma função Q da amostra $(X_1, \ldots, X_n)$ e do parâmetro θ, cuja distribuição de probabilidade não dependa do parâmetro θ é denominada quantidade pivotal. Desta forma, dado o nível de confiança $1 - \alpha$, toma-se:

$$1 - \alpha = P(q_1 \leq Q(X_1, \ldots, X_n; \theta) \leq q_2) \tag{3.4}$$

Se a quantidade pivotal é inversível, pode-se resolver a inequação acima em relação a θ e obter um Intervalo de Confiança (IC).

3.2 Definição de U-Estatística

Segundo *(Wikipédia, 2017b)*, na teoria estatística, uma U-Estatística é uma classe de estatísticas que é especialmente importante na teoria de estimação. A letra "U" significa "unbiased", traduzindo-se por não-viesado. Nas estatísticas elementares, as U-Estatísticas surgem naturalmente na produção de Estimadores Não-Viesados de Variância Uniformemente Mínima (ENVVUM). A teoria da U-Estatística permite que um Estimador Não- Viesado Uniformemente de Mínima Variância, seja derivado de cada estimador não-viesado de um parâmetro estimável (alternativamente, funcional estatístico) para grandes classes de distribuições de probabilidade. Um parâmetro estimável é uma função mensurável da distribuição de probabilidade acumulada da população: Por exemplo, para cada distribuição de probabilidade, a mediana da população é um parâmetro estimável. A teoria da U-Estatística aplica-se a classes gerais de distribuições de probabilidade. Muitas estatísticas, derivadas originalmente para famílias paramétricas particulares, foram reconhecidas como U-Estatísticas para distribuições gerais.

Na Estatística não-paramétrica, a teoria da U-Estatística é usada para estabelecer procedimentos estatísticos (como estimadores e testes) e estimadores relacionados à normalidade assintótica e à variância (em amostras finitas) de tais quantidades. A teoria tem sido usada para estudar estatísticas mais gerais, bem como processos estocásticos, gráficos aleatórios e para inferências em classificação e agrupamento (Valk and Pinheiro, 2012; Cybis et al., 2016).

Suponha que um problema envolva variáveis aleatórias independentes e distribuídas de forma idêntica e que a estimativa de um determinado parâmetro seja necessária. Suponha que uma estimativa simples e não- viesada possa ser construída com base em apenas algumas observações: isto define o estimador básico com base em um dado número de observações. Por exemplo, uma única observação é, ela própria, uma estimativa imparcial da média e um par de observações pode ser usado para derivar uma estimativa imparcial da variância. A U-Estatística baseada neste estimador é definida como a média (através de todas as seleções combinadas de dado tamanho a partir do conjunto completo de observações) do estimador básico aplicado à subamostras. (Kotz and Johnson, 2012) fornece uma revisão do artigo (Hoeffding, 1948), que introduziu a U-Estatística e estabeleceu a teoria relacionada a elas, e ao fazê-lo, esboça a importância que as U-Estatísticas têm na teoria Estatística.

Segundo (Kotz and Johnson, 2012): "O impacto de (Hoeffding, 1948) é esmagador no momento atual e é muito provável que continue nos próximos anos". Note-se que a teoria da U-Estatística não se limita ao caso de variáveis aleatórias independentes e identicamente distribuídas ou a variáveis aleatórias escalares.

Suponha que exista uma função ψ qualquer de uma a.a $X_1, X_2, \ldots, X_n$ i.i.d e um parâmetro qualquer tal que exista a seguinte relação: $\theta = \mathbb{E}(\psi(X_1, .., X_k))$, onde $k \leqslant n$. Então, a U-Estatística definida por:

$$U_n = \binom{n}{k}^{-1} \sum_{C_{n,k}} \psi(X_{i_1}, \cdots, X_{i_k}). \tag{3.5}$$

é um Estimador Não-Viesado de Variância Uniformemente Mínima (ENVVUM) de $\theta = \mathbb{E}(\psi(X_1, .., X_k))$, onde $C_{n,k}$ representa todas combinações de k elementos em n (Lee, 1990).

Exemplo 3.2.1. $\psi(X_i) = X_i$, núcleo (*kernel*) de ordem 1 ($k = 1$) e $\mathbb{E}(X_i) = \mu$ (média). Então,

$$U_n = \binom{n}{1}^{-1} \sum_{C_{n,1}} \psi(X_i) = \frac{1}{n} \sum X_i = \overline{X}. \tag{3.6}$$

Exemplo 3.2.2. Se a função da amostra for

$$\psi(X_i, X_j) = \frac{1}{2}(X_i - X_j)^2, \tag{3.7}$$

então, teremos um núcleo (*kernel*) de ordem 2 ($k = 2$). Observe que:

$$\mathbb{E}(\psi(X_i, X_j)) = E(\frac{1}{2}(X_i - X_j)^2) = \sigma^2. \tag{3.8}$$

Então,

$$\begin{aligned}
U_n &= \binom{n}{2}^{-1} \sum_{C_{n,2}} \psi(X_i, X_j) \\
&= \binom{n}{2}^{-1} \sum_{C_{n,2}} \frac{1}{2}(X_i - X_j)^2 \\
&= \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{2}({X_i}^2 - 2X_iX_j + {X_j}^2) \\
&= \binom{n}{2}^{-1} \frac{1}{2} \left(\sum_{i<j} {X_i}^2 - 2\sum_{i<j} X_iX_j + \sum_{i<j} {X_j}^2 \right) \\
&= \frac{1}{n(n-1)} \left((n-1)\sum_{i=1}^{n} {X_i}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j + \sum_{i=1}^{n} {X_i}^2 \right) \\
&= \frac{1}{n(n-1)} \left(n\sum_{i=1}^{n} {X_i}^2 - n^2\overline{X}^2 \right) \\
&= \frac{1}{(n-1)} \left(\sum_{i=1}^{n} {X_i}^2 - n\overline{X}^2 \right) \\
&= \frac{1}{(n-1)} \left(\sum_{i=1}^{n} {X_i}^2 - 2n\overline{X}^2 + n\overline{X}^2 \right) \\
&= \frac{1}{(n-1)} \left(\sum_{i=1}^{n} {X_i}^2 - 2n\overline{X}\frac{\sum_{i=1}^{n} X_i}{n} + \sum_{i=1}^{n} \overline{X}^2 \right) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} ({X_i}^2 - 2X_i\overline{X} + \overline{X}^2) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = s^2. \qquad (3.9)
\end{aligned}$$

Quando a distribuição de U_n é conhecida, é possível fazer inferências. Assim, por exemplo, se $\psi(X_i) = X_i$ e $\mu = \mathbb{E}(X_i) = \mathbb{E}(\psi(X_i))$, é possível mostrar que:

$$\frac{U_n - \mathbb{E}(U_n)}{\sqrt{\mathrm{Var}(U_n)}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{D} N(0,1) \qquad (3.10)$$

O qual é um resultado bem conhecido na Estatística, o Teorema Central do Limite. Esse particular resultado decorre de um resultado mais geral sobre U-Estatísticas genéricas.

3.2.1 Momentos das U-Estatísticas

Teorema 3.2.1. Define-se para $c = 1, 2, \ldots, k$, a esperança condicional por $\psi_c(x_1, \ldots, x_c) = \mathbb{E}\{\psi(X_1, \ldots, X_c, X_{c+1}, \ldots, X_k)/X_1 = x_1, \ldots, X_c = x_c\}$ e suas variâncias por $\sigma_c^2 = \mathrm{Var}\{\psi_c(X_1, \ldots, X_c)\}$. Então:

$$\mathrm{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\sigma_c^2 \tag{3.11}$$

Pode-se comprovar o teorema com esses exemplos:
Como visto no exemplo 3.2.1: Média simples. Tem-se $\psi_1(x) = x$, e então $\sigma_1^2 = \sigma^2 = \mathrm{Var}(X_1)$. Logo, $\mathrm{Var}\overline{X} = \frac{\sigma^2}{n}$
Como visto no exemplo 3.2.2: Variância simples. No caso da variância simples, o grau do núcleo é $k = 2$, e $\psi_1(x) = \mathbb{E}\{\frac{1}{2}(x - X_2)^2\} = \frac{1}{2}(\sigma^2 + (x - \mu)^2)$. Então, escrevendo μ_v para $\mathbb{E}(X_1 - \mu)^v$, tem-se:

$$\begin{aligned}
\sigma_1 &= \mathrm{Var}\left\{\frac{1}{2}(\sigma^2 + (X_1 - \mu)^2)\right\} \\
&= \frac{1}{4}\mathrm{Var}\{(X_1 - \mu)^2\} \\
&= \frac{1}{4}\{\mathbb{E}(X_1 - \mu)^4 - [\mathbb{E}\{(X_1 - \mu)^2\}]^2\} \\
&= \frac{1}{4}(\mu_4 - \sigma^4).
\end{aligned} \tag{3.12}$$

Além disso, $\psi_2(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$, então obtêm-se que $\sigma_2^2 = \mathrm{Var}\{\frac{1}{2}(X_1 - X_2)^2\} = \frac{1}{2}(\mu_4 + \sigma^4)$ e, consequentemente, denota-se a variância simples por s_n^2 e obtêm-se, substituindo σ_1^2 e σ_2^2 em termos de μ_4 e σ^2:

$$\begin{aligned}
\mathrm{Var}(s_n^2) &= \binom{n}{2}^{-1}(2(n-2)\sigma_1^2 + \sigma_2^2) \\
&= \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}.
\end{aligned} \tag{3.13}$$

Esse exemplo é bastante útil em (3.10) no caso de $U_n = S_n^2$ e $\mathbb{E}(U_n) = \sigma^2 = \theta$.

Teorema 3.2.2. Se ${\sigma_1}^2 > 0$, então $\sqrt{n}(U_n - \theta)$ é assintoticamente normal com média zero e variância $k^2{\sigma_1}^2$. Ou seja:

$$\sqrt{n}(U_n - \theta) \xrightarrow{D} N(0, k^2\sigma_1^2). \tag{3.14}$$

Teorema 3.2.3. A Decomposição-H:

Para descrever a decomposição, introduzi-se os núcleos $h^{(1)}, h^{(2)}, \ldots, h^{(k)}$ de graus $1, 2, \ldots, k$ que estão definidos recursivamente pelas equações:

$$h^{(1)}(x_1) = \psi_1(x_1) - \theta. \tag{3.15}$$

e

$$h^{(c)}(x_1, \ldots, x_c) = \psi_c(x_1, \ldots, x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, \ldots, x_{i_j}) - \theta. \tag{3.16}$$

para $c = 2, 3, \ldots, k$. Fazendo-se $S_j\{i_1, \ldots, i_k\}$ igual a soma $\sum h^{(j)}(x_{v_1}, \ldots, x_{v_j})$ de todos os j-subconjuntos $\{v_1, \ldots, v_j\}$ de $\{i_1, \ldots, i_k\}$. Então, usando a relação:

$$\sum_{(n,k)} S_j\{i_1, \ldots, i_k\} = \binom{n-j}{k-j} \sum_{(n,j)} h^{(j)}(x_{v_1}, \ldots, x_{v_j}), \tag{3.17}$$

a identidade:

$$\binom{n}{k}^{-1} \binom{n-j}{k-j} = \binom{k}{j} \binom{n}{j}^{-1}. \tag{3.18}$$

e a relação (3.16) para $c = k$, pode-se escrever:

$$\begin{aligned} U_n &= \binom{n}{k}^{-1} \sum_{(n,k)} \psi(x_{i_1}, \ldots, x_{i_k}) \\ &= \binom{n}{k}^{-1} \sum_{(n,k)} (\sum_{j=1}^{k} S_j\{i_1, \ldots, i_k\} + \theta) \\ &= \theta + \binom{n}{k}^{-1} \sum_{j=1}^{k} \binom{n-j}{k-j} \sum_{(n,j)} h^{(j)}(x_{v_1}, \ldots, x_{v_j}) \\ &= \theta + \sum_{j=1}^{k} \binom{k}{j} {H_n}^{(j)}, \end{aligned} \tag{3.19}$$

onde ${H_n}^{(j)}$ é a U-Estatística de grau j baseado no núcleo $h^{(j)}$.

Pode-se exemplificar recursivamente a equação (3.16):

$$h^{(1)}(x_1) = \psi_1(x_1) - \theta. \tag{3.20}$$

$$h^{(2)}(x_1, x_2) = \psi_2(x_1, x_2) - h^{(1)}(x_1) - h^{(1)}(x_2) - \theta. \tag{3.21}$$

$$\begin{aligned} h^{(3)}(x_1,x_2,x_3) &= \psi_3(x_1,x_2,x_3) - h^{(2)}(x_1,x_2) - h^{(2)}(x_1,x_3) - h^{(2)}(x_2,x_3) \\ &\quad -h^{(1)}(x_1) - h^{(1)}(x_2) - h^{(1)}(x_3) - \theta. \end{aligned} \tag{3.22}$$

$$\begin{aligned} h^{(4)}(x_1,x_2,x_3,x_4) &= \psi_4(x_1,x_2,x_3,x_4) - h^{(3)}(x_1,x_2,x_3) - h^{(3)}(x_1,x_2,x_4) \\ &\quad -h^{(3)}(x_2,x_3,x_4) - h^{(3)}(x_1,x_3,x_4) - h^{(2)}(x_1,x_2) - \ldots \\ &\quad -h^{(1)}(x_1) - h^{1}(x_2) - h^{(1)}(x_3) - h^{(1)}(x_4) - \theta. \end{aligned} \tag{3.23}$$

Na equação (3.17), supondo $k = 4$, defini-se S_1, S_2, S_3 e S_4 da seguinte forma:

$$S_1\{i_1,\ldots,i_4\} = h^{(1)}(x_1) + h^{(1)}(x_2) + h^{(1)}(x_3) + h^{(1)}(x_4). \tag{3.24}$$

$$S_2\{i_1,\ldots,i_4\} = h^{(2)}(x_1,x_2) + \cdots + h^{(2)}(x_3,x_4). \tag{3.25}$$

$$S_3\{i_1,\ldots,i_4\} = h^{(3)}(x_1,x_2,x_3) + \cdots + h^{(3)}(x_2,x_3,x_4). \tag{3.26}$$

$$S_4\{i_1,\ldots,i_4\} = h^{(4)}(x_1,x_2,x_3,x_4). \tag{3.27}$$

Assim, por exemplo em (3.17), para $k = 4$ e $j = 1$:

$$\begin{aligned} \sum_{(n,4)} S_1\{i_1,\ldots,i_4\} &= h^{(1)}(x_1) + h^{(1)}(x_2) + h^{(1)}(x_3) + \cdots + h^{(1)}(x_n) \\ &= \binom{n-1}{4-1} \sum_{(n,1)} h^{(1)}(x_{v_1}). \end{aligned} \tag{3.28}$$

Se em vez de $k = 4$, supormos $k = 2$ e $n = 3$, teremos:

$$\begin{aligned} \sum_{(3,2)} S_1\{i_1,i_2\} &= h^{(1)}(x_1) + h^{(1)}(x_2) + h^{(1)}(x_1) + h^{(1)}(x_3) + h^{(1)}(x_2) + h^{(1)}(x_3) \\ &= \binom{3-1}{2-1} \sum_{(3,1)} h^{(1)}(x_{v_1}). \end{aligned} \tag{3.29}$$

Supondo $k = 2$ e $n = 2$, pode-se comprovar (3.19), pois sabe-se por (3.16) que:

$$h^{(1)}(x_1) = \psi_1(x_1) - \theta. \tag{3.30}$$

$$h^{(1)}(x_2) = \psi_1(x_2) - \theta. \tag{3.31}$$

$$\begin{aligned} h^{(2)}(x_1, x_2) &= \psi_2(x_1, x_2) - \sum_{(2,1)} h^{(1)}(x_{v_1}) - \theta \\ &= \psi_2(x_1, x_2) - h^{(1)}(x_1) - h^{(1)}(x_2) - \theta. \end{aligned} \tag{3.32}$$

Então:

$$\begin{aligned} S_1\{i_1, i_2\} + S_2\{i_1, i_2\} + \theta &= h^{(1)}(x_1) + h^{(1)}(x_2) + h^{(2)}(x_1, x_2) + \theta \\ &= h^{(1)}(x_1) + h^{(1)}(x_2) + \theta + \psi_2(x_1, x_2) - h^{(1)}(x_1) - h^{(1)}(x_2) - \theta \\ &= \psi_2(x_1, x_2) = \psi_1(x_1, x_2) = \psi(x_1, x_2). \end{aligned} \tag{3.33}$$

4 Testes não-paramétricos no contexto de U-Estatísticas

No capítulo anterior, abordou-se o tema estimação e definiu-se algumas propriedades dos estimadores para amostras finitas. Além disso, introduziu-se o conceito de U-Estatísticas e mostrou-se que algumas das estatísticas usuais podem ser escritas como U-Estatísticas. Ainda, explorou-se algumas propriedades das U-Estatísticas tanto para amostras finitas como para amostras infinitas. Nesse contexto, a Decomposição-H é essencial para obter resultados assintóticos.

Nesse capítulo, apresenta-se alguns dos testes mais conhecidos na literatura reescritos como U-Estatísticas. A vantagem disso é notada, por exemplo, quando se deseja calcular a variância de uma determinada estatística ou uma potência dela, às vezes derivadas de um cálculo complexo pelo método usual.

4.1 Tau de Kendall

Dois pontos P_1 e P_2 no plano são ditos ser concordantes, se a linha que os liga tem inclinação positiva e discordantes, se a inclinação é negativa. Se ζ é o conjunto de todas as funções de distribuição bivariadas absolutamente contínuas dos vetores aleatórios (X,Y), então a medida da associação entre X e Y é a funcional τ definido em ζ por:

$$\tau = Pr(P_1 \text{ e } P_2 \text{ são concordantes }) - Pr(P_1 \text{ e } P_2 \text{ são discordantes}),$$

onde P_1 e P_2 são dois pontos independentes distribuídos como (X,Y). A funcional τ é chamada *coeficiente de concordância de Kendall* ou *Tau de Kendall* e satisfaz as propriedades usuais de correlação, assume valores entre [-1;1], sendo zero quando X e Y são independentes e igual a ± 1 sempre que $Y = f(X)$ para alguma função monótona f. Se

definido um núcleo κ_τ por:

$$\kappa_\tau(P_1, P_2) = \begin{cases} 1, & \text{se } P_1 \text{ e } P_2 \text{ são concordantes;} \\ -1, & \text{se } P_1 \text{ e } P_2 \text{ são discordantes;} \end{cases}$$

Então, $\kappa_\tau(P_1, P_2)$= $sgn(X_1 - X_2)(Y_1 - Y_2)$, τ=$\mathbb{E}(\kappa_\tau(P_1, P_2))$ e o estimador U-Estatística para estimar o parâmetro τ é:

$$\hat{\tau}_n = \binom{n}{2}^{-1} \sum_{(n,2)} \kappa_\tau(P_i, P_j), \tag{4.1}$$

que é a proporção de pares de pontos concordantes na amostra $(P_1, \ldots, P_n)$ menos a proporção de pares de pontos que são discordantes. O estimador (4.1) é obviamente não-viciado e a equação 3.11 diz que a variância de $\hat{\tau}_n$ é dada por:

$$\text{Var}(\hat{\tau}_n) = \binom{n}{2}^{-1} \{2(n-2)\sigma_1^2 + \sigma_2^2\}. \tag{4.2}$$

onde $\sigma_1^2 = \text{Var}(\kappa_1(X_1, Y_1))$ e $\sigma_2^2 = 1 - \tau^2$. A esperança condicional $\kappa_1(x_1, y_1)$ é dada por:

$$\begin{aligned} \kappa_1(x, y) &= \mathbb{E}\{\kappa_\tau((x, y), (X_1, Y_1))\} \\ &= Pr((x - X_1)(y - Y_1) > 0) - Pr((x - X_1)(y - Y_1) < 0) \\ &= Pr((X_1 > x \wedge Y_1 > y) \vee (X_1 < x \wedge Y_1 < y)) \\ &\quad -Pr((X_1 > x \wedge Y_1 < y) \vee (X_1 < x \wedge Y_1 > y)) \\ &= 1 - 2F(x, \infty) - 2F(\infty, y) + 4F(x, y) \\ &= (1 - 2F_1(x))(1 - 2F_2(y)) + 4(F(x, y) - F_1(x)F_2(y)). \end{aligned} \tag{4.3}$$

onde F_1 e F_2 são as funções de distribuição marginais de X_1 e Y_1. Sob a independência de X_1 e Y_1, $F(x, y)$=$F_1(x)F_2(y)$ e então $\kappa_1(x, y) = (1 - 2F_1(x))(1 - 2F_2(y))$. As variáveis aleatórias U e V dadas por $U = 1 - 2F_1(X)$ e $V = 1 - 2F_2(Y)$ são independentes e

uniformemente distribuídas em [-1;1] de tal modo que:

$$\begin{aligned} \mathrm{Var}(\kappa_1(X,Y)) &= \mathrm{Var}(UV) = \mathbb{E}(U^2)\mathbb{E}(V^2) - (\mathbb{E}(U))^2(\mathbb{E}(V))^2 \\ &= \left(\frac{1}{2}\int_{-1}^{-1} u^2 du\right)^2 \\ &= \frac{1}{9}. \end{aligned} \tag{4.4}$$

Sob a independência:

$$\begin{aligned} \mathrm{Var}(\hat{\tau}_n) &= \binom{n}{2}^{-1}\left(\frac{2(n-2)}{9}+1\right) \\ &= \frac{2(2n+5)}{9n(n+1)}. \end{aligned} \tag{4.5}$$

4.2 Método do Momentos

Sejam $Y_1, Y_2, \ldots, Y_n$ v.a i.i.d de uma população com média μ e variância $\sigma^2 (1 \leqslant i \leqslant n)$. Seja $h(Y_1, Y_2) = Y_1 Y_2$. Deseja-se estimar o quadrado da média μ^2. Então,

$$\mathbb{E}[h(Y_1, Y_2)] = \mathbb{E}(Y_1 Y_2) = \mathbb{E}(Y_1)\mathbb{E}(Y_2) = \mu^2. \tag{4.6}$$

o parâmetro de interesse. A U-Estatística baseada nesse *kernel* é:

$$U_n = \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} h(Y_i, Y_j) = \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} Y_i Y_j. \tag{4.7}$$

Novamente, U_n é um estimador não-viesado de μ^2.

4.3 Estimadores da Covariância

Sejam $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, um par de variáveis aleatórias i.i.d. Considerando a estimativa para a covariância entre X_i e Y_i, $\theta = \mathrm{Cov}(X_i, Y_i)$. Seja

$$h(Z_i, Z_j) = \frac{1}{2}(X_i - X_j)(Y_i - Y_j). \tag{4.8}$$

Então,

$$\begin{aligned}\mathbb{E}[h(Z_i, Z_j)] &= \frac{1}{2}[\mathbb{E}(X_iY_i) - \mathbb{E}(X_i)\mathbb{E}(Y_i)] + \frac{1}{2}[\mathbb{E}(X_jY_j) - \mathbb{E}(X_j)\mathbb{E}(Y_j)] \\ &= \frac{1}{2}\mathrm{Cov}(X_i, Y_i) + \frac{1}{2}\mathrm{Cov}(X_j, Y_j) = \theta. \end{aligned} \tag{4.9}$$

Portanto, a U-Estatística abaixo:

$$U_n = \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} h(Z_i, Z_j). \tag{4.10}$$

é um estimador não-viesado de $\theta = \mathrm{Cov}(X, Y)$. Também, pode ser expressa como uma média amostral sobre sujeitos simples ou funções de médias amostrais. Por exemplo, a U_n em (4.10) pode ser representada numa forma familiar como:

$$\begin{aligned} U_n &= \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} \frac{1}{2}(X_i - X_j)(Y_i - Y_j) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n). \end{aligned} \tag{4.11}$$

Essa igualdade pode ser verificada com o exemplo da programação em R em Listing 4.1:

Listing 4.1: Cálculo da Covariância "via U-Estatística" com R.

```
n=10
x=rt(n,5)
y=rnorm(n)
cvu=vector()
r=0
for (i in 1:(n-1)){
  for (j in ((i+1):n)){
    r=r+1
    cvu[r]=(x[i]-x[j])*(y[i]-y[j])
  }
}
covU=1/(n*(n-1))*sum(cvu)
############################ Forma padrão de estimar a covariância ##############
cvp=vector()
for (i in 1:n){
    cvp[i]=(x[i]-mean(x))*(y[i]-mean(y))
}
cp=1/((n-1))*sum(cvp)
######################### Pode-se observar que os resultados são equivalentes #######
cp                                          nos dois métodos
[1] -0.3530171
covU
[1] -0.3530171
```

4.4 Teste de Wilcoxon

O Teste estatístico de Wilcoxon para uma única amostra, cuja estatística do teste é representada por W_n^+, para testar a hipótese nula de que a média é igual a zero, é definido por:

$$W_n^+ = \sum_{i=1}^{n} I_{\{Y_i>0\}} R_i. \tag{4.12}$$

Essa estatística dista de 0 a $\frac{n(n+1)}{2}$. Sob $H_0 : \mu = \mathbb{E}(Y_i) = \mu_0 = 0$, aproximadamente metade de $Y_i s$ são negativos e seus postos são comparados com os daqueles $Y_i s$ positivos por causa da simetria. Segue que W_n^+ tem média $\frac{n(n+1)}{4}$, metade da distância $\frac{n(n+1)}{2}$. W_n^+ também pode ser expresso por:

$$\begin{aligned} W_n^+ = \sum_{i=1}^{n} I_{\{Y_i>0\}} R_i &= \sum_{i=1}^{n}\sum_{j=1}^{i} I_{\{Y_{(i)}+Y_{(j)}>0\}} = \sum_{i=1}^{n}\sum_{j=1}^{i} I_{\{Y_i+Y_j>0\}} \\ &= \sum_{i=1}^{n} I_{\{Y_i>0\}} + \sum_{1\leqslant i<j\leqslant n} I_{\{Y_i+Y_j>0\}}. \end{aligned} \tag{4.13}$$

A segunda igualdade pode ser verificada com o seguinte exemplo:

Exemplo 4.4.1.

Y_i	Y_1	Y_2	Y_3	Y_4
Valor	0	2	1	3
R_i	1	3	2	4

$Y_{(i)}$	$Y_{(1)}$	$Y_{(2)}$	$Y_{(3)}$	$Y_{(4)}$
Valor	0	1	2	3

Assim:

$I_{\{Y_{(i)}+Y_{(j)}\}}$	$Y_{(1)}$	$Y_{(2)}$	$Y_{(3)}$	$Y_{(4)}$
$Y_{(1)}$	0	1	1	1
$Y_{(2)}$	1	1	1	1
$Y_{(3)}$	1	1	1	1
$Y_{(4)}$	1	1	1	1

Então:

$$W_n^+ = \sum_{i=1}^{n} I_{\{Y_i>0\}} R_i = 1*3+1*2+1*4 = 9*1 = \sum_{i=1}^{n}\sum_{j=1}^{i} I_{\{Y_{(i)}+Y_{(j)}>0\}}. \tag{4.14}$$

Segue de (4.13) que:

$$\begin{aligned}
\binom{n}{2}^{-1} W_n^+ &= \binom{n}{2}^{-1} \sum_{i=1}^{n} I_{\{Y_i>0\}} + \binom{n}{2}^{-1} \sum_{1\leqslant i<j\leqslant n} I_{\{Y_i+Y_j>0\}} \\
&= \frac{2}{n-1}\frac{1}{n}\sum_{i=1}^{n} h_1(Y_i) + \binom{n}{2}^{-1} \sum_{1\leqslant i<j\leqslant n} h_2(Y_i,Y_j) \\
&= \frac{2}{n-1} U_{n1} + U_{n2}.
\end{aligned} \tag{4.15}$$

Acima, U_{n1} é a U-Estatística para a média simples da amostra e U_{n2} é também uma U-Estatística definida pelo núcleo simétrico $h_2(Y_1,Y_2) = I_{\{Y_i+Y_j>0\}}$. $\binom{n}{2}^{-1} W_n^+$ e U_{n2} tem a mesma distribuição assintótica e então, a inferência para H_0 pode ser executada equivalentemente somente pelo uso da distribuição assintótica de U_{n2} para grandes amostras. A U-Estatística U_{n2} é um estimador não-viesado de $\theta = \mathbb{E}(h_2(Y_i,Y_j)) = Pr(Y_i+Y_j>0)$. Tomando a esperança de ambos os lados na igualdade (4.15), abaixo sob $H_0 : \mu_0 = 0$, tem-se:

$$\binom{n}{2}^{-1} \mathbb{E}(W_n^+) = \binom{n}{2}^{-1} \frac{n(n+1)}{4} = \frac{1}{2}, \tag{4.16}$$

$$\begin{aligned}
\frac{2}{n-1}\mathbb{E}(U_{n1}) + \mathbb{E}(U_{n2}) &= \frac{2}{n-1} Pr(Y_i>0) + Pr(Y_i+Y_j>0) \\
&= \frac{1}{n-1} + \theta.
\end{aligned} \tag{4.17}$$

Como $n \longrightarrow \infty$, segue que:

$$\theta = \frac{1}{2} - \frac{1}{n-1} \longrightarrow \frac{1}{2}. \tag{4.18}$$

Então, pode-se expressar a hipótese nula $H_0 : \mu_0 = 0$ em termos de θ, como:

$$H_0 : \theta = Pr(Y_i + Y_j \leqslant 0) = \frac{1}{2}. \tag{4.19}$$

Para ilustrar o Teste de Wilcoxon, tem-se o exemplo didático do portal *"www.portalaction.com.br"*: Considere a seguinte amostra:

Exemplo 4.4.2.

126	142	156	228	245	246	370	419	433	454	478	503

Supondo que os dados da amostra são distribuídos simetricamente em torno da mediana $\mu_0 = 220$ e subtraindo o valor 220 de cada valor da amostra, tem-se um novo conjunto de dados:

-94	-78	-64	8	25	26	150	199	213	234	258	283

Colocando em ordem crescente os valores absolutos, associamos à cada valor o posto correspondente R_i e a função indicadora $I_{\{Y_i>0\}}$:

Valor	8	25	26	-64	-78	-94	150	199	213	234	258	283
Posto	1	2	3	4	5	6	7	8	9	10	11	12
$I_{\{Y_i>0\}}$	1	1	1	0	0	0	1	1	1	1	1	1

Neste caso, a estatística W_n^+ é a soma dos postos positivos, isto é: $W_n^+ = \sum_{i=1}^{12} I_{\{Y_i>0\}} R_i = 1 + 2 + 3 + 7 + 8 + 9 + 10 + 11 + 12 = 63$.

1. Hipóteses: $\begin{cases} H_0 : \mu = 220; \\ H_1 : \mu \neq 220. \end{cases}$

O código em R, testando as hipóteses, pode ser visto em Listing 4.2:

Listing 4.2: Exemplo do Teste Wilcoxon com R.

```
y=c(126,142,156,228,245,246,370,419,433,454,478,503) ##simulação do exemplo do ##
 n=length(y)                                           ## portal action ##
 nr=y-220                                              ## passo a passo ###
 ans=matrix(rep(0,n*5),ncol=n,nrow=5)
 absord=sort(abs(nr))
 ord=1:n
 for (i in 1:n){
   if (nr[i]<=0){
     p1=which(absord==-nr[i])[1]
     print(p1)
     absord[p1]=-absord[p1]
   }
 }
[1] 6
[1] 5
[1] 4
 a=(absord>0)*1
 ans[1,]=y
 ans[2,]=nr
 ans[3,]=absord
 ans[4,]=ord
 ans[5,]=a
 ans
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,]  126  142  156  228  245  246  370  419  433   454   478   503
[2,]  -94  -78  -64    8   25   26  150  199  213   234   258   283
[3,]    8   25   26  -64  -78  -94  150  199  213   234   258   283
[4,]    1    2    3    4    5    6    7    8    9    10    11    12
[5,]    1    1    1    0    0    0    1    1    1     1     1     1

W= sum(ans[4,]*ans[5,])
 somai=rep(0,n)
 for (i in 1:(n-1)){
   somaj=rep(0,n)
   for ( j in (i+1):n){
     somaj[j]=((nr[i]+nr[j])>0)*1
   }
   somai[i]=sum(somaj)
 }
 WU=sum((nr>0)*1) + sum(somai)
 WU
[1] 63
########################################Usando a rotina pré-programada do R##############
 wR1=wilcox.test(y,mu=220)
 wR1

        Wilcoxon signed rank test

data:  y
V = 63, p-value = 0.06396
alternative hypothesis: true location is not equal to 220
```

4.5 Teste de Independência para Duas Variáveis Binárias

Sejam duas variáveis categóricas ordinais U e V. Suponha que U tem K e V tem M níveis indexados por k e m, respectivamente. Assume-se que U e V tem ambos valores binários. Por conveniência, codifica-se U e V de modo que todos tenham valores 0 e 1. A independência entre U e V é equivalente a seguinte condição:

$$p_{UV} = Pr(UV = 1) = Pr(U = 1)Pr(V = 1) = p_U p_V. \tag{4.20}$$

Seja $h(Z_i, Z_j) = \frac{1}{2}(U_i - U_j)(V_i - V_j)$. Define a seguinte U-Estatística:

$$U_n = \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} h(Z_i, Z_j) = \binom{n}{2}^{-1} \sum_{(i,j)\in C_{n,2}} \frac{1}{2}(U_i - U_j)(V_i - V_j). \tag{4.21}$$

A U-Estatística é um estimador não-viesado de $\sigma = \mathbb{E}(h(Z_i, Z_j)) = p_{UV} - p_U p_V$. A condição para independência em (4.20) é equivalente a nulidade $H_0 : \sigma = 0$. Então, pode-se usar a distribuição de U_n para testar a independência entre U e V.

5 Inferência em Mínimos Quadrados Ordinários via U-Estatísticas

No capítulo anterior, mostrou-se que alguns dos métodos não-paramétricos mais conhecidos na literatura estatística podem ser escritos como uma U-Estatística. A vantagem de se fazer essa correspondência é que ganha-se "de graça" as propriedades assintóticas das U-Estatísticas. Nesse capítulo, explora-se os recursos da simulação com a programação em R, expandido o conceito do Método do Mínimo Quadrados Ordinários com a abordagem do conceito de U-Estatística.

Sejam $Z_1, \ldots, Z_n$ variáveis aleatórias independentes e identicamente distribuídas. Seja $\theta \in R^p$ um vetor de parâmetros e suponha que a estimação de θ é baseada numa função estimadora não-viciada

$$S(\theta) = \binom{n}{K}^{-1} \sum_{1 \leqslant i_1 \leqslant i_2 \cdots \leqslant i_K \leqslant n} h(Z_{i_1}, \ldots, Z_{i_K}; \theta), \tag{5.1}$$

que tem a estrutura da U-Estatística de grau K. Então, $\mathbb{E}\{h(Z_1, \ldots, Z_K; \theta)\} = 0$ onde 0 é o vetor de zeros em R^p e h é simétrica nos argumentos $Z_{i_1}, \ldots, Z_{i_K}$ ((Jiang and Kalbfleisch, 2012)). Nota-se que, em geral, no contexto de U-Estatísticas, temos que $\mathbb{E}\{h(Z_1, \ldots, Z_K; \theta)\} = 0$, mas nesse caso, a função de estimação $S(\theta)$ surge quando o objetivo é minimizar uma função objetiva,

$$U(\theta) = \binom{n}{K}^{-1} \sum_{1 \leqslant i_1 \leqslant i_2 \cdots \leqslant i_K \leqslant n} H(Z_{i_1}, \ldots, Z_{i_K}; \theta), \tag{5.2}$$

onde $h = \dfrac{\partial H}{\partial \theta}$. O objetivo aqui é estudar o método para construção de Intervalos de Confiança para θ (ou componentes de θ) através de procedimentos *Bootstrap* para a função

de estimação, a qual denota-se por EFB (Estimating Function Bootstrap). Um exemplo de uma função estimadora, com $K = 2$ na classe (5.1), surge da minimização da função objetiva

$$U(\beta) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 \leqslant i_2 \leqslant n} |Y_{i_1} - Y_{i_2} - (X_{i_1} - X_{i_2})^T \beta|^{\alpha}, \tag{5.3}$$

com respeito a regressão para o parâmetro $\beta \in R^p$.

Nesse caso, $1 \leqslant \alpha \leqslant 2$, $Y_i = \gamma + {X_i}^T \beta + e_i$, para a constante γ, e $e_i, i = 1, \ldots, n$, são independentes e identicamente distribuídas (i.i.d) com erros de $\mathbb{E}(e_i) = 0$. Uma estimativa para o parâmetro de regressão $\hat{\beta}$, pode ser obtido resolvendo a equação $S(\beta) = 0$, onde a função de estimação é dada por:

$$S(\beta) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 \leqslant i_2 \leqslant n} sign\{Y_{i_1} - Y_{i_2} - (X_{i_1} - X_{i_2})^T \beta\}(X_{i_1} - X_{i_2}) |Y_{i_1} - Y_{i_2} - (X_{i_1} - X_{i_2})^T \beta|^{\alpha - 1}. \tag{5.4}$$

Caso $a = 1$, a função de estimação 5.4 remete ao método "Regressão de postos de Wilcoxon"([Hettmans-perger, 1984]). Segundo ((Jiang and Kalbfleisch, 2012)), as derivadas de $S(\beta)$ não são comportadas para $1 \leqslant \alpha \leqslant 2$. Por exemplo, se $p = 1$ e $1 \leqslant \alpha \leqslant 2$, as derivadas de $S(\beta)$ não estão definidas se β assume valor $\frac{(Y_{i_1} - Y_{i_2})}{(X_{i_1} - X_{i_2})}$ para $1 \leqslant i_1 \leqslant i_2 \leqslant n$. Consequentemente, é difícil aplicar a Inferência tradicional como um estimador "sandwich" de variâncias ou um Teste de Wald baseado em $\hat{\beta}$. ((Jiang and Kalbfleisch, 2012)) argumenta que em geral $S(\theta)$ não é uma quantidade pivotal e propõe uma versão "estudantizada" de $S(\theta)$, denotada por $S_t(\theta)$. O procedimento de "estudantização" da $S(\theta)$ será feito para $k = 2$. Neste caso,

$$S(\theta) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 \leqslant i_2 \leqslant n} h(Z_{i_1}, Z_{i_2}; \theta), \tag{5.5}$$

que é um caso especial de (5.1) e é uma U-Estatística de grau 2. Seguindo (Sen, 1960), com $k = 2$, define-se em R^p:

$$q_i(\theta) = \frac{1}{n-1} \sum_{j:j \neq i} h(Z_i, Z_j; \theta), \tag{5.6}$$

para $i = 1, \ldots, n$. Os $q_i s$ são identicamente distribuídos e $S(\theta) = \frac{1}{n} \sum_{i=1}^{n} q_i(\theta)$. Seja ${S_q}^2$ a matriz de variâncias e covariâncias amostrais de $q_1, \ldots, q_n$. O estimador para a variância

de S é definido por:

$$V(\theta) = \frac{4}{n}{S_q}^2 = \frac{4}{n^2(n-1)} \sum_{1 \leqslant i \leqslant j \leqslant n} \{q_i(\theta) - q_j(\theta)\}^{\otimes 2}, \tag{5.7}$$

em que $a^{\otimes 2} = a^T a$ para um vetor coluna a.

No caso em que $p = 1$, $V(\theta) = \frac{4}{n^2(n-1)} \sum_{i \leqslant j} \{q_i(\theta) - q_j(\theta)\}^2$. Em geral, define-se a U-Estatística de grau k, $q_i(\theta) = \binom{n-1}{k-1}^{-1} \sum C_i \times h(Z_i, Z_{l_1}, \ldots, Z_{l_{k-1}}; \theta)$ onde $C_i = \{(l_1, \ldots, l_{k-1}) : 1 \leqslant l_1 \leqslant \cdots \leqslant l_{k-1} \leqslant n$, e $l_1, \ldots, l_{k-1} \neq i\}$. Como acima $S(\theta) = \frac{1}{n} \sum_{i=1}^n q_i(\theta)$, então o resultado da estimação da variância de $S(\theta)$ é:

$$V(\theta) = \frac{k^2}{n^2(n-1)} \sum_{1 \leqslant i \leqslant j \leqslant n} \{q_i(\theta) - q_j(\theta)\}^{\otimes 2}, \tag{5.8}$$

A versão "estudantizada" de $S(\theta)$ é dada por:

$$S_t(\theta) = \{V(\theta)\}^{-\frac{1}{2}} S(\theta). \tag{5.9}$$

((Jiang and Kalbfleisch, 2012)) mostra que $S_t(\theta)$ converge assintoticamente para uma normal padrão. Aproximações assintóticas podem ser usadas para construir os Intervalos de Confiança ou, pode-se usar métodos de reamostragem para obter a distribuição de S_t. A Regressão de Mínimos Quadrados Ordinários pode ser simulada com o uso do *Bootstrap*. No presente caso, o parâmetro de regressão β (Regressão Linear Simples) deve ser estimado a partir da função estimadora:

$$S(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - {X_i}^T \beta) X_i, \tag{5.10}$$

que é a U-Estatística de grau 1. Gerou-se os dados dos seguintes modelos:

- Caso 1 (Erros normais homogêneos): Dados são gerados de $Y_i = \beta X_i + e_i, i = 1, \ldots, n$, $n = 12$, os $e_i's$ são i.i.d da $N(0, 8)$, $X_i = -1.6, \ldots, -1.1, 1.1, \ldots, 1.6$ e $\beta = 1$.

- Caso 2 (Erros normais heterogêneos): Dados são gerados de $Y_i = \beta X_i + W_i e_i, i = 1, \ldots, n$, onde $W_i = X_i^2 / \sqrt{\sum_{i=1}^n {X_i}^4 / n}$, e outros valores são o mesmo caso que 1.

- Caso 3 (Erros assimétricos): Os $e_i's$ são i.i.d de $\chi_4^2 - 4$ onde χ_4^2 é definida pela

distribuição Qui-quadrado com 4 graus de liberdade.

- Caso 4 (Normais assimétricas): O tamanho da amostra é $n = 16$. As covariáveis $X_1, \ldots, X_8$ é uma amostra obtida da $N(0, 1)$ e $X_9, \ldots, X_{16}$ é uma amostra obtida da $N(1, 0.25)$.

Os resultados do *Bootstrap*, para diferentes valores de α e para as quatro situações consideradas, estão sintetizados na Tabela 5.1 e o código em R, para $\alpha = 2.5\%$, pode ser visto entre Listing 5.1 e Listing 5.4. Assim, por exemplo, o percentual de cobertura de 2% indica que somente 2% dos Intervalos de Confiança, obtidos via *Bootstrap*, tiveram seu limite inferior maior que $\beta = 1$. Assim, para $\alpha = 2.5\%$, o erro do tipo I estaria controlado. Nota-se que há uma correspondência entre o IC inferior a 95% e o IC superior a 5%.

Os resultados da Tabela 5.1 mostram que, embora os intervalos obtidos via *Bootstrap* sejam generosos, eles atendem a propriedade de que o erro tipo I seja de no máximo α, enquanto o IC obtido pela Inferência tradicional do Método de Mínimos Quadrados Ordinários, padrão da saída da função *"lm"* do R, tem uma taxa de cobertura muito fora do que seria esperado. Por exemplo, para $\alpha = 0.05$, a taxa de cobertura, para os quatro casos, fica em torno de 50%. Isso quer dizer que metade dos intervalos construídos nas 1000 replicações não continham o verdadeiro parâmetro. Embora, o erro padrão do método *Bootstrap* seja ligeiramente superior, os resultados apontam para melhores propriedades desse método.

Tabela 5.1: Intervalos de Confiança inferior para $\beta = 1$ (Lim Inf) obtidos via *Bootstrap* (boot) e via Mínimos Quadrados Ordinários (lm), taxa de cobertura (% de vezes em que o Lim Inf foi maior que $\beta = 1$ em 1000 replicações) e Erro padrão dos estimadores, considerando $\alpha = 0.025, 0.05, 0.95$ e 0.975.

Casos	Indicadores	2.5%		5%		95%		97.5%	
		boot	lm	boot	lm	boot	lm	boot	lm
	% de cobertura	0.02	0.49	0.03	0.47	0.96	0.25	0.97	0.23
Caso 1	Erro padrão	0.68	0.60	0.66	0.61	0.65	0.59	0.66	0.58
	Lim Inf	-0.43	1.00	-0.20	0.97	2.18	0.59	2.35	0.56
	% de cobertura	0.02	0.49	0.03	0.48	0.96	0.26	0.97	0.25
Caso 2	Erro padrão	0.71	0.63	0.69	0.64	0.69	0.64	0.69	0.61
	Lim Inf	-0.40	1.00	-0.18	0.98	2.17	0.60	2.34	0.56
	% de cobertura	0.01	0.46	0.04	0.50	0.97	0.27	0.98	0.25
Caso 3	Erro padrão	0.73	0.64	0.71	0.62	0.69	0.60	0.77	0.60
	Lim Inf	-0.41	0.95	-0.15	1.00	2.16	0.64	2.41	0.63
	% de cobertura	0.03	0.52	0.05	0.47	0.94	0.27	0.96	0.27
Caso 4	Erro padrão	0.72	0.68	0.71	0.74	0.73	0.70	0.76	0.70
	Lim Inf	-0.27	1.01	-0.12	0.94	2.10	0.58	2.32	0.58

Listing 5.1: *Bootstrap* com α igual a 2.5%

```
#######################################################
# Caso 1
# Erros homogeneos normais

#n=12
x= c(-1.6,-1.5,-1.4,-1.3,-1.2,-1.1,1.1,1.2,1.3,1.4,1.5,1.6)
n=length(x)

Rep=1000
q_inf1=vector(); q_sup1=vector()
q_inf2=vector(); q_sup2=vector()
q_inf3=vector(); q_sup3=vector()
q_inf4=vector(); q_sup4=vector()
IClm1=vector(); IClm2=vector(); IClm3=vector();IClm4=vector()
ICSuplm1=vector();ICSuplm2=vector();ICSuplm3=vector();ICSuplm4=vector()
alpha=0.025

for(r in 1:Rep){

  beta=1 # verdadeiro parametro para todas as simulacoes

  y=vector()
  e=rnorm(n,0,sd=sqrt(8)) #e rros com media zero e variancia 8
  for (i in 1:n){
    y[i]=beta*x[i]+e[i]
  }
  dados=data.frame(x,y)
#############################################################
  mqo1 <- function(data, par) {
    with(data, sum((y- par * x )^2)) # U(beta)
  }

  result <- optim(par = 1, mqo1, method = "Brent",data = dados, lower=-1000,upper=1000)
  beta_chapeu=result$par
#############################################################
mod1=lm(y~x-1)
IClm1[r]=confint(mod1,level = alpha/2)[1]
ICSuplm1[r]=confint(mod1,level = alpha/2)[2]

B=999 # numero de repeticoes do bootstrap
Ss=vector()
for( b in 1:B){
   z=vector()
   for (i in 1:n){
      z[i] = y[i]-x[i]*beta_chapeu
   }
   zs=sample(z,n,replace=TRUE) #z*
   Ss[b]=mean(zs) # S* do Kalbfleish
}
  q_inf1[r]= beta_chapeu+qnorm(alpha)*sd(Ss)
  q_sup1[r]= beta_chapeu+qnorm(1-alpha)*sd(Ss)
#q_inf1[r]=as.numeric(quantile(Ss,0.025))
#q_sup1[r]=as.numeric(quantile(Ss,0.975))
#######################################################
# Caso 2

# Erros heterogeneos normais
y=vector()
for (i in 1:n){
  w=x[i]^2/(sqrt(mean(x^4)))
  y[i]=beta*x[i]+w*e[i]
}
dados=data.frame(x,y)
mqo1 <- function(data, par) {
  with(data, sum((y- par * x )^2)) # U(beta)
}
result <- optim(par = 1, mqo1, method = "Brent",data = dados, lower=-1000,upper=1000)
beta_chapeu=result$par
```

Listing 5.2: (continuação) *Bootstrap* com α igual a 2.5%

```
 mod2=lm(y~x-1)
  IClm2[r]=confint(mod2,level = alpha/2)[1]
  ICSuplm2[r]=confint(mod2,level = alpha/2)[2]

B=999 # numero de repeticoes do bootstrap
Ss=vector()
for( b in 1:B){
  z=vector()
  for (i in 1:n){
    z[i] = y[i]-x[i]*beta_chapeu
  }
  zs=sample(z,n,replace=TRUE) #z*
  Ss[b]=mean(zs) # S* do Kalbfleish
}

  q_inf2[r]= beta_chapeu+qnorm(alpha)*sd(Ss)
  q_sup2[r]= beta_chapeu+qnorm(1-alpha)*sd(Ss)
# q_inf2[r]=as.numeric(quantile(Ss,0.025))
# q_sup2[r]=as.numeric(quantile(Ss,0.975))
#####################################################
# Caso 3

# Erros assimetricos
  e=rchisq(n,4)-4
  y=vector()
  for (i in 1:n){
    y[i]=beta*x[i]+e[i]
  }
  dados=data.frame(x,y)
  ###########################################################
  mqo1 <- function(data, par) {
    with(data, sum((y- par * x )^2)) # U(beta)
  }

  result <- optim(par = 1, mqo1, method = "Brent",data = dados, lower=-1000,upper=1000)
  beta_chapeu=result$par
  ###########################################################
  mod3=lm(y~x-1)
  IClm3[r]=confint(mod3,level = alpha/2)[1]
  ICSuplm3[r]=confint(mod3,level =alpha/2)[2]

  B=999 # numero de repeticoes do bootstrap
  Ss=vector()
  for( b in 1:B){
    z=vector()
    for (i in 1:n){
      z[i] = y[i]-x[i]*beta_chapeu
    }
    zs=sample(z,n,replace=TRUE) #z*
    Ss[b]=mean(zs) # S* do Kalbfleish
  }

  q_inf3[r]= beta_chapeu+qnorm(alpha)*sd(Ss)
  q_sup3[r]= beta_chapeu+qnorm(1-alpha)*sd(Ss)
#   q_inf3[r]=as.numeric(quantile(Ss,0.025))
#   q_sup3[r]=as.numeric(quantile(Ss,0.975))
#########################################################
# Caso 4 (Normais Assimetricas)

  y=vector()
  n4=16
  e1=rnorm(n4/2,0,sd=1) #erros com media zero e variancia1
  e2=rnorm(n4/2,1,sd=sqrt(0.25)) #erros com media 1 e variancia 0.25
  e=rnorm(n4,0,sd=sqrt(8)) #e rros com media zero e variancia 8
  #e=rchisq(n4,4)-4
  for (i in 1:n4/2){
    y[i]=beta*e1[i]+e[i]
  }
```

Listing 5.3: (continuação) *Bootstrap* com α igual a 2.5%

```
    for (i in 1:n4/2){
    y[i]=beta*e1[i]+e[i]
  }
  for (i in (n4/2+1):n4){
    y[i]=beta*e2[(i-n4/2)]+e[i]
  }
  x4=c(e1,e2)
  dados=data.frame(x4,y)
  ############################################################
  mqo1 <- function(data, par) {
    with(data, sum((y- par * x4 )^2)) # U(beta)
  }

  result <- optim(par = 1, mqo1, method = "Brent",data = dados, lower=-1000,upper=1000)
  beta_chapeu=result$par
  ############################################################
  mod4=lm(y~x4-1)
  IClm4[r]=confint(mod4,level = alpha/2)[1]
  ICSuplm4[r]=confint(mod4,level = alpha/2)[2]

  B=999 # numero de repeticoes do bootstrap
  Ss=vector()
  for( b in 1:B){
    z=vector()
    for (i in 1:n4){
      z[i] = y[i]-x4[i]*beta_chapeu
    }
    zs=sample(z,n4,replace=TRUE) #z*
    Ss[b]=mean(zs) # S* do Kalbfleish
  }
  q_inf4[r]= beta_chapeu+qnorm(alpha)*sd(Ss)
  q_sup4[r]= beta_chapeu+qnorm(1-alpha)*sd(Ss)
  #q_inf1[r]=as.numeric(quantile(Ss,0.025))
  #q_sup1[r]=as.numeric(quantile(Ss,0.975))

}

 # CP (%) Percentual de cobertura
 mean(q_inf1>1)
[1] 0.02
 mean(q_inf2>1)
[1] 0.023
 mean(q_inf3>1)
[1] 0.013
 mean(q_inf4>1)
[1] 0.034

 mean(IClm1>1)
[1] 0.493
 mean(IClm2>1)
[1] 0.492
mean(IClm3>1)
[1] 0.464
mean(IClm4>1)
[1] 0.525

 # Avg.CI  media dos limites inferiores dos intervalos
 mean(q_inf1) # IC inf a (1-alpha)*100%  para o Caso 1
[1] -0.4282888
 mean(q_inf2) # IC inf a (1-alpha)*100%  para o Caso 2
[1] -0.4018546
 mean(q_inf3) # IC inf a (1-alpha)*100%  para o Caso 3
[1] -0.4147937
 mean(q_inf4) # IC inf a (1-alpha)*100%  para o Caso 4
[1] -0.2742044

> mean(IClm1)  # IC a (1-alpha)*100%  para o Caso 1 usando o MQO padrao (lm)
[1] 0.9962556
```

Listing 5.4: (continuação) *Bootstrap* com α igual a 2.5%

```
> mean(IClm2)  # IC a (1-alpha)*100%  para o Caso 2 usando o MQO padrao (lm)
[1] 0.9989974
> mean(IClm3)  # IC a (1-alpha)*100%  para o Caso 3 usando o MQO padrao (lm)
[1] 0.9556362
> mean(IClm4)  # IC a (1-alpha)*100%  para o Caso 4 usando o MQO padrao (lm)
[1] 1.014451

 # SE.CI Erro padrao
 sd(q_inf1)
[1] 0.6799746
 sd(q_inf2)
[1] 0.712792
 sd(q_inf3)
[1] 0.7312936
 sd(q_inf4)
[1] 0.7235925

 sd(IClm1)
[1] 0.5960478
 sd(IClm2)
[1] 0.6268093
 sd(IClm3)
[1] 0.5930314
 sd(IClm4)
[1] 0.6834764
```

5.1 Reamostragem Para a Função de Estimação

Para o processo de reamostragem, considerando a função de estimação (5.5), o procedimento consiste em repassar θ por $\hat{\theta}$ em (5.5) e estimar a distribuição de S ou versão "estudantizada" de S_t, substituindo seus termos por suas versões estimadas. Pode-se descrever o método EF_t para S_t, com $k = 2$, da seguinte forma:

1. Gera $(V_1, \ldots, V_n)$ de uma distribuição multinomial $(n, \frac{1}{n}, \ldots, \frac{1}{n})$

2. Seja $S^* = \frac{1}{n}\sum_{i=n}^{n} V_i \tilde{q}_i^*$ em que:

$$\tilde{q}_i^* = \frac{1}{n-1}\sum_{l:l\neq i} V_l h(Z_i, Z_l; \hat{\theta}). \tag{5.11}$$

   De (5.8) segue que a variância de S^* é:

$$V^* = \frac{k^2}{n^2(n-1)}\sum_{i\leqslant j} V_i V_j (\tilde{q}_i^* - \tilde{q}_j^*)^{\otimes 2}. \tag{5.12}$$

3. Finalmente, defina:

$$S_t^* = V^{*-\frac{1}{2}} S^*. \tag{5.13}$$

O procedimento acima é repetido B vezes (um número grande de vezes). (Jiang and Kalbfleisch, 2012) mostra que a distribuição empírica de S_t^* tem a mesma distribuição assintótica de S_t, ou seja:

$$S_t^* \longrightarrow N_p(0, \mathbb{I}), \tag{5.14}$$

em que $\mathbb{I}$ é um vetor de tamanho p (considera-se p=1 nesse trabalho). A utilização dos pesos multinomiais $(V_1, \ldots, V_n)$, no procedimento de reamostragem, deve produzir resultados idênticos ao procedimento de *Bootstrap* padrão, segundo (Jiang and Kalbfleisch, 2012), em que $({Z_1}^*, \ldots, {Z_n}^*)$ são sorteadas com reposição da amostra $(Z_1, \ldots, Z_n)$. A vantagem dessa abordagem é que ela possibilita a utilização de pesos de qualquer outra distribuição. Se θ é um escalar, define-se um intervalo unilateral com $100(1-\alpha)\%$ de confiança para θ com o $\{\theta : S_t(\theta) > \hat{S}_{t_\alpha}\}$ em que $\hat{S}_{t_\alpha}$ é uma estimativa do quantil α de S_t. Pode-se também utilizar $\hat{S}_{t_\alpha} = Z_\alpha$, de acordo com a propriedade assintótica de S_t, em que Z_α é o α-ésimo quantil da normal padrão, ou pode-se tomar $\hat{S}_{t_\alpha} = {S^*}_{t_\alpha}$, em que ${S^*}_{t_\alpha}$ é o α-ésimo quantil empírico obtido a partir das replicações do procedimento EF_t.

Se $S_t(\theta)$ é monótona não-crescente em θ, uma aproximação para o $IC\ 100(1-\alpha)\%$ é dada por $(-\infty, \hat{\theta}_{1-\alpha}]$ em que $\hat{\theta}_{1-\alpha}$ é a solução de $S_t(\theta) = \hat{S}_{t_\alpha}$.

6 Conclusão

Neste trabalho, procurou-se introduzir o conceito de U-Estatísticas com o objetivo de desmistificar o assunto, o qual não é abordado no universo da graduação, e mostrar a sua versatilidade principalmente no campo das Estatísticas Não-Paramétricas. A teoria é muito abrangente, pode-se perceber a matemática extensa envolvida nas definições e resultados. Cabe ressaltar que a literatura é limitada, o que torna mais difícil o desafio de apresentar o tema de forma compreensível. Além do mais, as publicações existentes são um tanto quanto densas e requerem uma boa base matemática para o entendimento.

Um dos objetivos desse trabalho é justamente esse, "abrir" resultados já prontos para facilitar a compreensão geral da teoria desenvolvida. Assim, foi no exemplo da U-Estatística para um *kernel* de ordem $k = 2$, no caso da variância populacional, onde definições aparentemente simples, terminaram em cálculos trabalhosos. Também, a Decomposição-H não é de compreensão elementar e demandou bastante empenho na análise. Em outros momentos, aproveitou-se a programação em R para atestar alguns resultados, como por exemplo no caso da U-Estatística para a covariância. Do mesmo modo, foi interessante a simulação do Teste de Wilcoxon com o enfoque da U-Estatística para comprovar as demonstrações que envolviam muitas funções indicadoras.

Para exemplificar os diversos campos de aplicações das U-estatísticas, considerou-se uma abordagem em que o objetivo era melhorar os Intervalos de Confiança para os parâmetros de um modelo de regressão. Mostrou-se que a função objetiva do Método de Mínimos Quadrados pode ser escrito no contexto de U-Estatísticas. Também, comparou-se um método baseado em *Bootstrap* e o Método de Mínimos Quadrados quanto a performance na construção de limites inferiores para os Intervalos de Confiança. Observou-se que para os casos considerados, a abordagem via U-Estatísticas foi muito mais eficiente. Esse estudo apresentou uma programação computacional extensa e optou-se em mostrar o código em

R para $\alpha = 2.5\%$, citando os demais resultados para α. Os resultados obtidos em todas as simulações foram compatíveis com a teoria subjacente, o que indica a validade da U-Estatística nesses casos.

Sem dúvida, a U-Estatística é um dos ramos mais promissores da Estatística/Matemática e oferece a Estatística muitas ferramentas importantes. Aqueles que se interessarem pelo tema vão encontrar aplicações em Processos Estocásticos, Séries Temporais e Análise Multivariada, por exemplo. Há muito conteúdo disponível para analisar e desvendar na teoria de U-Estatística e esse trabalho espera assim ter contribuído.

Referências

Bolfarine, H. and Sandoval, M. C. (2001). *Introdução à inferência estatística*, volume 2. SBM.

Bussab, W. d. O. and Morettin, P. A. (2010). *Estatística básica*. Saraiva.

Callegari-Jacques, S. M. (2009). *Bioestatística: princípios e aplicações*. Artmed Editora.

Cybis, G. B., Valk, M., and Lopes, S. R. C. (2016). Clustering and classification of genetic data through u-statistics. *arXiv preprint arXiv:1606.03376*.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325.

Hoskin, T. (2014). Parametric and nonparametric: Demystifying the terms. *Mayo Clinic CTSA BERD Resource Retrieved from http://www mayo edu/mayo-edudocs/center-fortranslational-science-activities-documents/berd-5-6 pdf*.

Jiang, W. and Kalbfleisch, J. D. (2012). Bootstrapping u-statistics: applications in least squares and robust regression. *Sankhya B*, 74(1):56–76.

Kotz, S. and Johnson, N. L. (2012). *Breakthroughs in Statistics: Foundations and basic theory*. Springer Science & Business Media.

Lee, J. (1990). U-statistics: Theory and practice.

Sen, P. K. (1960). On some convergence properties of ustatistics. *Calcutta Statistical Association Bulletin*, 10(1-2):1–18.

Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*, 33(4):608–619.

Wikipédia (2017a). Bibtex — wikipédia, a enciclopédia livre. [Online; accessed 06-Maio-2017].

Wikipédia (2017b). Bibtex — wikipédia, a enciclopédia livre. [Online; accessed 06-Maio-2017].