Capítulo 5

Estatística em Informática na Educação

Patrícia L. Espinheira, Ranilson Paiva, Diego Dermeval e Ig I. Bittencourt

Abstract

The use of computers to support education started and evolved with the appearance of the digital computers. This led to the creation of different types of educational environments as the Learning Management Systems, Intelligent Tutoring Systems, Adaptive Hypermedia Systems, Computer-Supported Collaborative Learning Systems and, more recently, the Massive Open Online Courses. However, considering the publications in the Brazilian Symposium of Computers in Education, between 2001 and 2012, as a representative of the Brazilian publications in the Computers in Education field, we noticed that only 29.9% of them applied any method for empirical evaluation. In order to overcome this situation, we propose the use of statistics, a knowledge area responsible for studying the appropriate methods for collecting, representing and analyzing data. In this chapter, we present statistical techniques and methods for empirical evaluation. Besides that, we cement our approach to the subject, applying these methods and techniques in real data from studies in the Computer in Education field.

Resumo

O uso dos computadores para apoiar a educação surgiu, e evolui com o surgimento e a evolução dos próprios computadores digitais. Isso levou à criação de diferentes tipos de ambientes educacionais como os sistemas de gerenciamento de aprendizagem, sistemas tutores inteligentes baseados na web, sistemas de hipermídia adaptativos, sistemas de aprendizagem colaborativa apoiados por computador e, os cursos massivos abertos online. Entretanto, usando como referências as publicações no Simpósio Brasileiro de Informática na Educação, entre os anos 2001 e 2012, para representar as publicações em informática na educação no Brasil, percebemos que apenas 29.9% delas aplicaram alguns métodos de avaliação empírica. Para superar esta situação, propomos o uso da estatística, que é a área do conhecimento onde se estudam métodos apropriados para a obtenção, representação e análise de dados. Neste capítulo, apresentaremos técnicas e métodos estatísticos para a avaliação empírica. Além disso, consolidaremos nossa abordagem ao assunto através da aplicação da estatística em dados reais de trabalhos na área da Informática na Educação.

1. Introdução

O início do uso de computadores para apoiar a educação confunde-se com o próprio surgimento dos computadores digitais. Em meados da década de 50, onde os primeiros computadores começaram a ser comercializados, já haviam iniciativas de uso dos mesmos em contextos educacionais. A partir da década de 70, as tecnologias para apoiar a educação por meio de computadores foram evoluindo e sendo amplamente difundidas, em um primeiro momento, com o uso da instrução auxiliada por computador ou *Computer-Aided Instruction* (CAI). Em seguida, com o aparecimento dos microcomputadores pessoais na década de 80 houve a disseminação de computadores nas escolas e a ampla produção e diversificação de CAIs.

Com a evolução e massificação da Internet a partir da década de 90, a Informática na Educação passou a incluir também o uso da Internet para possibilitar: (i) o acesso a materiais de aprendizagem; (ii) interação com conteúdo, instrutores e outros aprendizes; e (iii) suporte durante o processo de aprendizagem visando adquirir conhecimento para construir significado pessoal e para crescer a partir da experiência de aprendizagem [Anderson, 2008]. A partir do acesso a Internet em larga escala e com a constante evolução dos computadores e técnicas computacionais, diferentes tipos de ambientes educacionais têm sido propostos e pesquisados como, por exemplo, os sistemas de gerenciamento de aprendizagem [Sclater, 2008], sistemas tutores inteligentes baseados na web [Ma et al, 2014], sistemas de hipermídia adaptativos [Brusilovsky,1998], sistemas de aprendizagem colaborativa apoiados por computador [Stahl et al., 2006] e, mais recentemente, os cursos massivos abertos online [Martin, 2012]. Por outro lado, diferentes métodos, técnicas (ex.: mineração de dados e gamificação) e ferramentas têm sido estudados tendo como inspiração diferentes áreas do conhecimento como, por exemplo, Computação, Educação e Psicologia.

No Brasil, um dos principais eventos da área é o Simpósio Brasileiro de Informática na Educação (SBIE), que este ano está na sua vigésima sexta edição. De fato, o contínuo interesse dos pesquisadores na área de Informática na Educação pode ser constatado pelo aumento crescente na quantidade de publicações ao longo das edições do SBIE. Por exemplo, em 2001 foram 57 artigos, já em 2012 foram 92 artigos publicados. Um mapeamento sistemático recente analisa os artigos publicados no SBIE entre os anos de 2001 e 2012 [Magalhães et al., 2013]. Magalhães et al. (2013) apresentam como um dos principais resultados do estudo a verificação de que um número muito elevado de artigos publicados no SBIE, considerando o período analisado, não utilizou nenhum método empírico na pesquisa realizada. Além disso, em muitos casos, os autores dos artigos considerados não realizaram nenhum tipo de validação ou avaliação dos resultados obtidos. Em resumo, pouco mais de 70% dos artigos não fizeram uso de algum método de avaliação empírica nos estudos publicados.

Por outro lado, considerando um contexto de pesquisa internacional, como o da comunidade de Inteligência Artificial na Educação, cujas principais conferências científicas (*Artificial Intelligence in Education – AIED e Intelligent Tutoring System – ITS*) ocorrem alternadamente a cada dois anos, pode-se notar um nível oposto de validação empírica dos trabalhos propostos. Segundo Blanchard (2012), os artigos empíricos da conferência ITS de 2002 e da AIED de 2003 representaram,

respectivamente, 50.5% e 57.5% do total de artigos publicados naquele ano. Enquanto que no ITS de 2010 e no AIED de 2011 estes valores foram 90.2% e 93.9%, respectivamente.

Ainda, segundo Magalhães et al. (2013), mesmo com poucos artigos validando empiricamente suas pesquisas e levando em conta a quantidade total de trabalhos publicados, há uma certa tendência de crescimento no uso de métodos empíricos nas edições do SBIE de 2001 a 2012. Seguindo a classificação de métodos empíricos para validação de software proposta por Easterbrook et al., (2008), Magalhães et al. (2013) classificam os métodos de pesquisa dos artigos do SBIE em: estudo de caso, experimento formal, *survey*, pesquisa-ação e etnografia. A utilização dos estudos de caso em IE representa 18.3% do total de artigos publicados no SBIE. Os experimentos formais (e quasi-experimentos) também apresentaram crescimento nas pesquisas da área, aparecendo com 8.7% das pesquisas. Os *surveys* com 1%, a pesquisa-ação com 0.7% e, por fim, as etnografias com apenas 0.2% dos artigos.

É importante ressaltar que as pesquisas de caráter avaliativo e empírico demonstram um maior nível de rigor científico e aumentam a confiabilidade dos resultados encontrados, além de favorecer a reprodutibilidade das mesmas. Uma das principais ferramentas para validar e construir modelos matemáticos que possam explicar a aplicação de modelos, técnicas e ferramentas computacionais no contexto da comunidade da IE é a estatística.

Estatística é a área do conhecimento onde se estudam métodos apropriados para a obtenção, representação e análise de dados. Os métodos estatísticos devem constituir o início de todo o processo de pesquisa empírica, desde que toda análise de dados requer um planejamento experimental estatístico que permitirá a qualidade e verossimilhança das conclusões obtidas. Desde do planejamento do estudo, que consiste por exemplo na definição do plano amostral que otimize a investigação empírica, até a construção de modelos matemáticos complexos capazes de explicar adequadamente o problema de interesse, a estatística é imprescindível também para a comunidade de Informática na Educação.

O objetivo deste capítulo é apresentar conceitos e métodos que podem ser aplicados no contexto de informática na educação. Para ilustrar o uso da estatística no contexto de IE serão utilizados três exemplos ao decorrer da apresentação dos conceitos neste capítulo. O primeiro exemplo envolve a avaliação de uma abordagem para classificar os estudantes de um ambiente online de aprendizagem, considerando suas interações com os recursos educacionais disponíveis. O segundo exemplo envolve os aspectos de avaliação empírica de uma ferramenta que utiliza correções feitas por pares de alunos.

Este capítulo está organizado da seguinte forma. A Seção 1.2 introduz a área de estatística. A Seção 1.3 apresenta conceitos básicos sobre indivíduos e variáveis. A Seção 1.4 apresenta as variáveis qualitativas através de um exemplo em IE. A Seção 1.5 discute os conceitos de variável aleatória e os seus tipos, e apresenta um exemplo na informática na educação. A Seção 1.6 apresenta conceitos acerca de medidas de posição. Já a Seção 1.7 apresenta conceitos de separatrizes. Na Seção 1.8, são discutidos conceitos básicos sobre as distribuições simétricas e assimétricas. A Seção 1.9 define intervalos de confiança. A Seção 1.10 apresenta conceitos de testes estatísticos

(normalidade, paramétricos e não paramétricos). A Seção 1.11 apresenta conceitos sobre modelos de regressão. Na Seção 1.12 são discutidas algumas diretrizes para conduzir bons experimentos. Por fim, na Seção 1.13 as conclusões são apresentadas.

2. Introdução a Estatística

A estatística é a ciência dos dados. É a área do conhecimento onde se estudam métodos apropriados para a obtenção, representação e análise de dados.

Os métodos estatísticos podem ser aplicados nos mais diversos setores de atividades, contribuindo de forma essencial para tomadas de decisões estratégicas. Mas, para isso é necessário coletar, organizar e interpretar dados. E é exatamente isso que faremos aqui. Mas, o volume de dados pode ser avassalador. Precisamos aprender a sumarizar esse volume de dados para obter o que nos interessa: informações úteis na prática.

Apesar de um dos passos iniciais de uma análise estatística ser a construção da amostra, para proceder esse passo precisamos de conhecimentos prévios de estatística. Como por exemplo a caracterização do tipo de variáveis que compõem o experimento e adicionalmente associar as distribuições de probabilidades mais adequadas para estas variáveis, além da usual distribuição de probabilidades normal.

3. Indivíduos e Variáveis

Todo conjunto de dados contém informações sobre um grupo de indivíduos ou de casos. Em geral, estamos interessados em investigar características desse grupo. Se essas características envolvem aleatoriedade, isto é, fatores que não podemos mensurar totalmente, elas são chamamos de Variáveis aleatórias. Tudo em Estatística se baseia nas propriedades matemáticas de uma variável aleatória (v.a.).

4. Variáveis Qualitativas

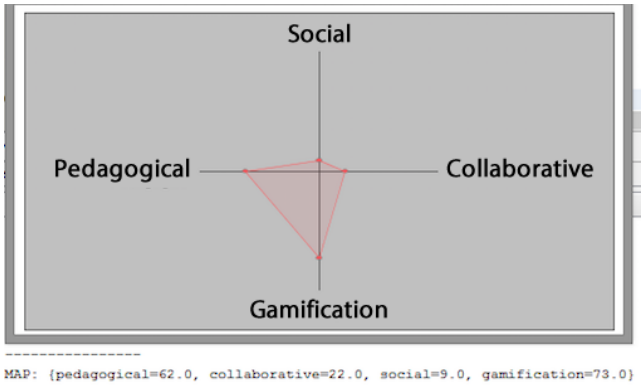
Estes tipos de variáveis aleatórias não assumem valores numéricos, mas apresentam categorias e por isso são chamadas de Variáveis Categóricas. Este tipo de variável posiciona os indivíduos em um dos diversos grupos ou categorias. Pode-se sumarizar as informações contidas em variáveis categóricas utilizando, por exemplo, porcentagens, gráficos de barra e os gráficos de pizza. A seguir apresentamos um exemplo no contexto de IE em que os dados são categóricos.

4.1 Exemplo na classificação de estudantes de um ambiente online de aprendizagem

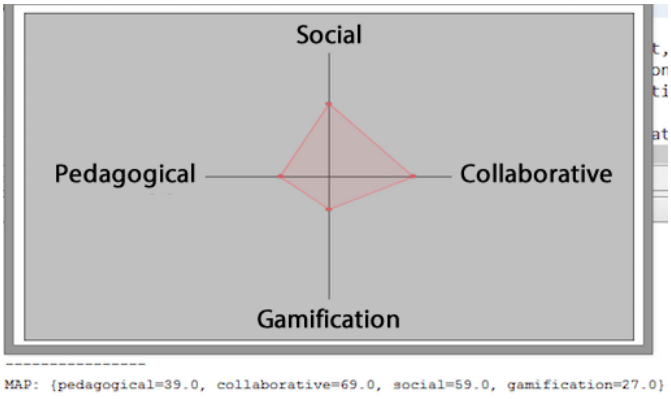
No trabalho de Paiva et. al. (2015) foi apresentada uma abordagem para classificar os estudantes de um ambiente online de aprendizagem considerando suas interações com os recursos educacionais disponíveis. O objetivo geral foi o de auxiliar professores/tutores na criação de missões personalizadas (um elemento de gamificação que contém tarefas para manter o estudante engajado). Para o ambiente de aprendizagem estudado (MeuTutor®), seus recursos educacionais foram mapeados em 4 categorias/perfis:

1. Colaborativo: interações com recursos educacionais destinados ao aprendizado colaborativo. Por exemplo, avaliação de vídeos, sugestões, problemas encontrados;
2. Gamificação: interação com recursos educacionais destinados à gamificação (troféus conquistados, pontos ganhos, missões concluídas);
3. Pedagógico: recursos educacionais destinados ao aprendizado individual (problemas resolvidos, testes resolvidos, vídeos assistidos);
4. Social: interação com recursos educacionais destinados à socialização; (convites de amizade enviados, interações por chat, conteúdo compartilhado);

Os pesquisadores acompanharam estudantes do sistema MeuTutor® e coletaram suas interações com o sistema. Com base nestas interações os pesquisadores selecionaram quatro grupos. Para os dois primeiros grupos foram construídos gráficos do tipo teia de aranha, os quais fornecem uma representação gráfica das categorias (perfis) de cada um dos dois grupos com base em suas interações. Esses gráficos vão representar o perfil dos que chamaremos aqui de Estudante 1 e Estudante 2, Figuras 1.1a e 1.1b, respectivamente.



(a) Estudante 1.



(b) Estudante 2.

Figura 2.1 Perfil dos Estudantes 1 e 2 segundo o gráfico de teia de aranha

Para os dois outros grupos apenas foi gerado o gráfico relatório do MeuTutor®, o qual fornece as interações do aluno com o sistema. Esse gráficos vão representar o comportamento de interações com o sistema que chamaremos aqui de Estudante 3 e Estudante 4, Figuras 1.2a. e 1.2b, respectivamente.

(a) Estudante 3

(b) Estudante 4

Figura 1.3. Perfil dos Estudantes 3 e 4 segundo o gráfico disponibilizado pelo MeuTutor

Em seguida, foi solicitado a 31 avaliadores, via questionário *online* que analisassem os quatro gráficos e atribuíssem um perfil predominante à cada tipo de estudante representado nos respectivos gráficos. Esses avaliadores foram selecionados segundo alguns critérios, entre os quais anos de experiência em ensino. Na Tabela 1.1 apresentamos um recorte de como foram armazenados parte dos resultados do experimento. Apenas selecionamos do experimento os "anos de experiência do avaliador com ensino" e sua respectiva classificação dos alunos 1-4 a partir dos gráficos disponibilizados.

Tabela 1.1. Classificação de alunos de ambiente de aprendizagem online. Tempo de experiência do avaliador com ensino (Recorte do conjunto de dados original)

Tempo de experiência com ensino do avaliador	Perfil Aluno 1	Perfil Aluno 2	Perfil Aluno 3	Perfil Aluno 4
Mais de 6 anos	Colaborativo	Gamificação	Pedagógico	Social
Entre 1 e 3 anos	Social	Gamificação	Pedagógico	NPR
Mais de 6 anos	Colaborativo	Gamificação	Pedagógico	Gamificação
Menos de 1 ano	Não é possível responder(NPR)	NPR	Gamificação	Gamificação
Entre 1 e 3 anos	Social	Gamificação	Gamificação	NPR
Nenhuma	Colaborativo	Gamificação	Pedagógico	Social
Entre 1 e 3 anos	Colaborativo	Gamificação	Gamificação	NPR
Menos de 1 ano	Colaborativo	Gamificação	Pedagógico	Social
⋮	⋮	⋮	⋮	⋮

A partir dos dados apresentados na Tabela 1.1, destacam-se as variáveis envolvidas no estudo e quais os resultados possíveis destas variáveis. As variáveis são:

* A – Tempo de experiência com ensino do avaliador, com resultados possíveis: "0" – Nenhuma experiência com ensino; "-1" – Menos de 1 ano de experiência com ensino; "1 + - 6 – Mais de um 1 ano de ensino e menos que 6 anos; e "6+" – Mais de 6 anos de ensino);

* B – Perfil do Aluno 1; * C – Perfil do Aluno 2; * D – Perfil do Aluno 3; e * E – Perfil do Aluno 4, com resultados possíveis: "Colaborativo" (Col.), "Gamificação" (Game), "Pedagógico (Ped.), "Social" (Soc.) e "Não é Possível Responder" (NPR).

Note que as variáveis A, B, C, D e E não assumem valores numéricos; elas são variáveis categóricas (qualitativas). Desta forma, a Tabela 1.2 apresenta uma forma de extrair informações com bases nestas variáveis. Ela representa os dados alocando tanto a experiência do avaliador (A) quanto os indicadores dos alunos (B, C, D e E). De fato, na Tabela 1.2, as informações da variável A são cruzadas com as informações das variáveis B, C, D e E.

Tabela 1.2. Classificação de alunos de ambiente de aprendizagem online. Tempo de experiência do avaliador com ensino

Categorias	Aluno 1					Aluno 2					Aluno 3					Aluno 4				
	0	-1	1+-6	6+	Σ	0	-1	1+-6	6+	Σ	0	-1	1+-6	6+	Σ	0	-1	1+-6	6+	Σ
Col.	4	4	5	5	17	0	0	1	0	1	1	0	2	0	1	1	0	0	0	1
Game	0	0	0	0	0	3	4	6	5	16	0	2	2	4	8	0	1	0	2	3
Ped.	0	0	0	0	0	2	0	0	0	2	2	2	0	3	7	0	0	0	0	0
Soc.	0	0	2	0	2	0	0	0	0	0	1	0	0	1	2	3	1	0	1	5
NPR	1	1	1	0	2	0	1	1	0	2	0	1	1	0	3	1	3	8	2	14
Σ	5	5	8	5	21	5	5	8	5	21	4	5	5	8	21	5	5	8	5	23

Decidimos colocar os percentuais associados à Tabela 2 nos gráficos de pizza. Com base nas Figuras 1.3 (gráficos de barras), podemos notar que os quatro alunos apresentam perfis muito distintos. Os Alunos 1 e 2 apresentam perfis predominantemente colaborativos e de gamificação, respectivamente. Esses alunos foram classificados com base no gráfico de teia de aranha e nota-se que a classificação é de certa forma homogênea considerando os anos de experiência dos avaliadores. Por outro lado, ocorre o contrário para os alunos 3 (Figura 1.3c) e 4 (Figura 1.3d). A classificação do estudante muda consideravelmente de acordo com a experiência do avaliador. É notável a dificuldade de identificar o perfil do Aluno 4, dadas as ocorrências de respostas não possíveis. As assertivas anteriores são confirmadas a partir da observação dos gráficos de pizza, Figura 1.4.

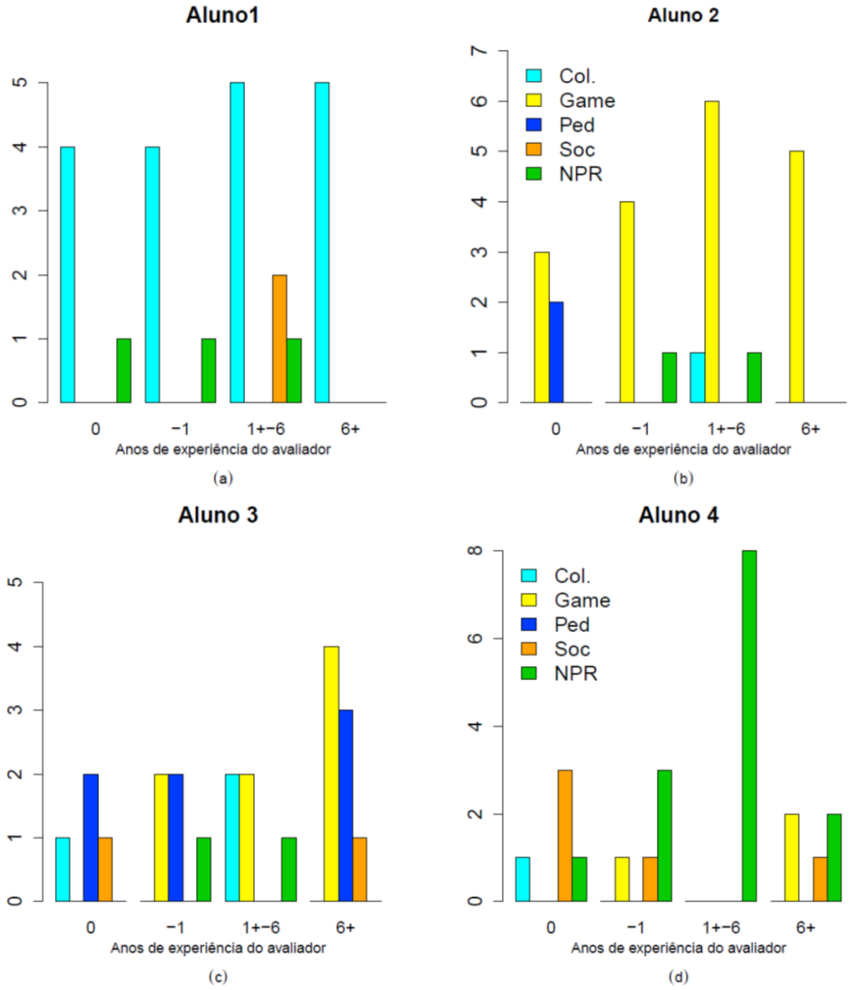


Figura 1.4. Gráficos de Barras. Perfis dos Alunos 1 – 4 segundo experiência do avaliador

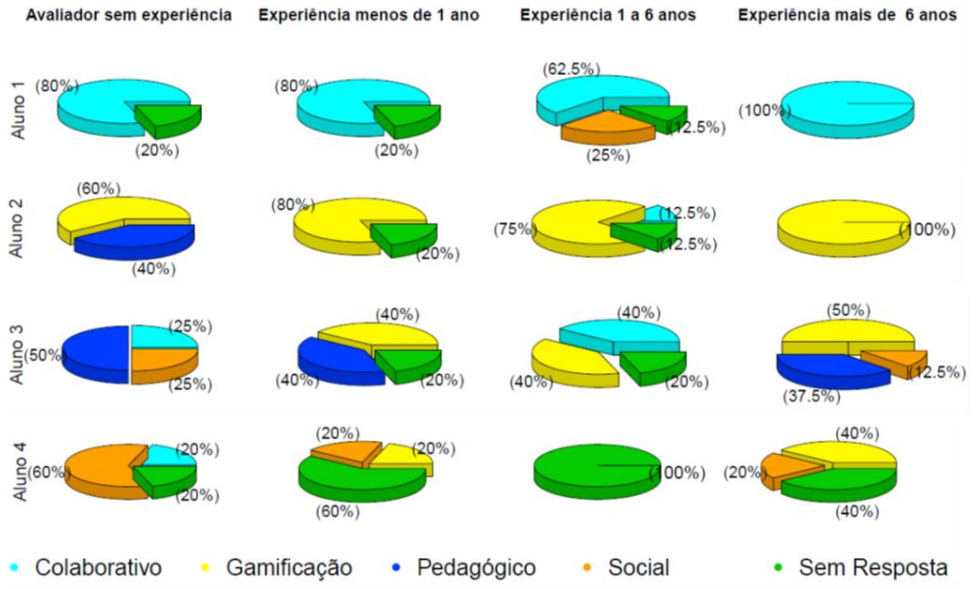


Figura 1.5. Gráficos de Pizzas. Perfis dos Alunos 1 – 4 segundo quatro categorias de experiência com ensino do avaliador

Adicionalmente, é interessante notar como 100% dos avaliadores com mais de seis anos de experiência com ensino apontam que o perfil do Aluno 1 é Colaborativo (Figura 1.4 – primeira linha, última coluna). Só entre os avaliadores com experiência entre um e seis anos é que 25% dos mesmos atribuem o perfil Social ao Aluno 1. Pode-se ver na Figura 1.4 que apenas 40% dos avaliadores sem experiência categorizaram o Aluno 2 como perfil pedagógico, enquanto 100% dos avaliadores com mais de seis anos de experiência identificaram que o Aluno 2 tem um perfil de Gamificação. Para os alunos 3 e 4 os avaliadores não entram em um consenso mesmo dentro da mesma categoria de experiência com ensino. Por exemplo, 50% dos avaliadores com mais de seis anos de experiência acreditam que o Aluno 3 tem um perfil de Gamificação, enquanto 37.5% atribuem a esse mesmo aluno um perfil de Gamificação e 12.5%, um perfil social. Ao que parece, a classificação do aluno sofre a influência da experiência do professor. Podemos investigar com base em um teste estatístico se existe a dependência entre o nível de experiência do avaliador e a sua respectiva classificação do aluno, isto será feito mais adiante.

5. Variável Aleatória

Em termos matemáticos, variável aleatória é uma variável cujo valor é um resultado numérico de um fenômeno aleatório. A distribuição de probabilidade de uma variável aleatória X nos diz quais valores X pode assumir e como atribuir probabilidades a esses valores. Usualmente, denotamos as variáveis aleatórias por letras maiúsculas, tais como X, Y, Z, W, T.

Observação Importante: Uma Variável aleatória X associada a uma distribuição de probabilidades é o que chamamos em Estatística de *POPULAÇÃO*. No entanto, na prática trabalhamos com amostras de nossa população X. Denotamos por $X1$ uma

amostra de tamanho *n* da população *X*. Temos ainda que cada *X1, X2, ..., Xn* também é variável aleatória} e funções destas amostras, como a soma ou a média aritmética, também são variáveis aleatórias. Isto implica dizer que essas funções, como são variáveis aleatórias, apresentam distribuições de probabilidades associadas a elas.

5.1 Tipos de Variáveis Aleatórias

As variáveis aleatórias podem ser divididas em dois tipos:

- Discretas: têm uma lista finita, ou infinita, enumerável de valores possíveis, por exemplo, $X = 0, 1, 2, 3$ (lista finita e enumerável) e $W = ..., -3, -2, -1, 0, 1, 2, 3,$ ... (lista infinita, porém enumerável)
- Contínuas: podem assumir qualquer valor em um intervalo, têm uma lista não enumerável de valores possíveis, por exemplo, $X = [0,1]$, $W = (-\infty,3]$, $Z = (-\infty,+\infty) = R$ (todos os conjuntos são não enumeráveis). Ou seja, $Z = 2.98$ é um possível valor para a variável aleatória Z. Enquanto que $X = 0,0001$ é um possível valor para variável aleatória *X*. Temos infinitos valores entre [0,1].

5.2 Variável Aleatória Quantitativa

Este tipo de variável aleatória assume valores numéricos com os quais faz sentido efetuar operações aritméticas, tais como adição e cálculo de médias. A seguir, descrevemos um exemplo deste tipo de variável na avaliação de uma ferramenta no contexto de IE.

Exemplo de avaliação de uma ferramenta criada para avaliação por pares conduzida por alunos para correções de questões discursivas em ambiente *online*.

Atualmente existem diversas tecnologias que apoiam a educação online, como por exemplo, os sistemas tutores inteligentes. Usualmente, esses sistemas oferecem soluções mais simples, como por exemplo, questões objetivas. Porém, ao serem incluídas avaliações escritas gera-se uma grande sobrecarga no professor. Neste sentido, é preciso a criação de algum mecanismo que possibilite a inclusão de avaliações escritas de tal maneira que não seja necessário aumentar o custo vinculado às correções destas e que estas correções sejam feitas de maneira (semi)automática e independente do número de alunos.

No trabalho de Tenório et al. (2015) foi proposto um mecanismo de avaliação por pares onde o sistema ficará então responsável pelo gerenciamento das correções que serão feitas de uma maneira semi-automatizada pelos próprios estudantes. Isto sem que seja necessário aumentar o custo vinculado as correções das atividades (**Custo**), sem interferir na qualidade da avaliação (**Nota**), nem no tempo para a execução das correções (**Tempo**).

Como pode ser visto na Tabela 1.3, variáveis quantitativas tipicamente assumem muitos valores. Precisamos sumarizar esses dados. Um passo inicial é construir uma tabela de distribuição de frequências, onde os dados são agrupados em classes. Em seguida, construir o histograma: uma representação gráfica desta tabela de frequência.

Pretendemos implementar algumas técnicas desenvolvidas neste curso no software estatístico R em sua versão 3.2.0 que se encontra disponível gratuitamente no endereço *http://www.r-project.org*. Apresentamos abaixo o script no R para a construção três primeiros gráficos (Histogramas) da Figura 1.5:

```
dados<-scan("C:\\Users\\patricia\\Patricia\\Patricia_2015\\Thyago\\CustoNotaTempo_T1T2.prn",
list(Custo_T1=0,Custo_T2=0, Nota_T1=0,Nota_T2=0,Tempo_T1=0,Tempo_T2=0))
attach(dados)
dados<-data.frame(dados)# Essas três linhas carregam os dados no R. Os dados têm que ter o mesmo comprimento.
dados1<-scan("C:\\Users\\patricia\\Patricia\\Patricia_2015\\Thyago\\CustoNotaTempo_T3.prn",
list(Custo_T3=0, Nota_T3=0,Tempo_T3=0))
attach(dados1)
dados1<-data.frame(dados1) #Como o grupo T3 (Professor_QuestOnLine) tem menos observações salvei em outro arquivo.
postscript("C:\\Users\\Patricia_2015\\HistCustoNota.eps") # Para salvar o gráfico. #Passo Inicial
par(mfrow=c(2,3), pty="s") #Para criar um painel com duas linhas de gráficos e três colunas.
hist(Custo_T1, freq = FALSE, main = "Professor_MeuTutor",
xlab ="Custo",ylab ="Frequência",cex.main=1.7,cex.lab=1.7,xlim = c(0,21))#freq=FALSE permite que não sejam as frequêcias mas sim as probabilidades das classes.
lines(density(Custo_T1), col=2) #quando usamos freq = FALSE podemos ajustar uma densidade ajustada. #Passo Final
# Repete o Passo Inicial e Final mais duas vezes.
dev.off()
```

Tabela 1.3. Custo, nota e tempo dos tipos de avaliações (professor ou pares de alunos) e tipos de ambiente (MeuTutor® ou Questionário online)

Avaliação Feita pelo professor No ambiente Meu Tutor			Avaliação Feita pelos pares de alunos No ambiente Meu Tutor			Avaliação Feita pelo professor Questionario on line		
Nota	Custo	Tempo	Nota	Custo	Tempo	Nota	Custo	Tempo
680	8.97	1009	626	1.49	450	6.07	440	840
480	8.04	905	573	1.23	270	3.03	440	420
720	7.63	858	680	1.49	450	4.33	640	600
480	8.00	900	520	1.49	450	3.90	520	540
480	3.98	448	573	1.19	240	4.77	560	660
400	5.33	600	380	1.27	300	3.57	400	494
640	10.4	1175	600	1.78	645	6.62	680	917
680	16.2	1833	520	1.64	555	4.77	600	660
760	4.73	532	546	1.71	600			
⋮	⋮	⋮	⋮		⋮	⋮		

Com base nos histogramas apresentados na Figura 1.5 percebemos que os custos apresentam distribuições de frequências distintas entre si, em especial o **Custo** das avaliações feita por pares de aluno no ambiente MeuTutor, inclusive na forma da densidade estimada. As variáveis **Nota** e **Tempo** apresentam distribuições de frequência bem similares quanto aos três métodos para fazer as correções das questões discursivas. Ainda exploraremos mais estes histogramas na seção 1.10.2.2.

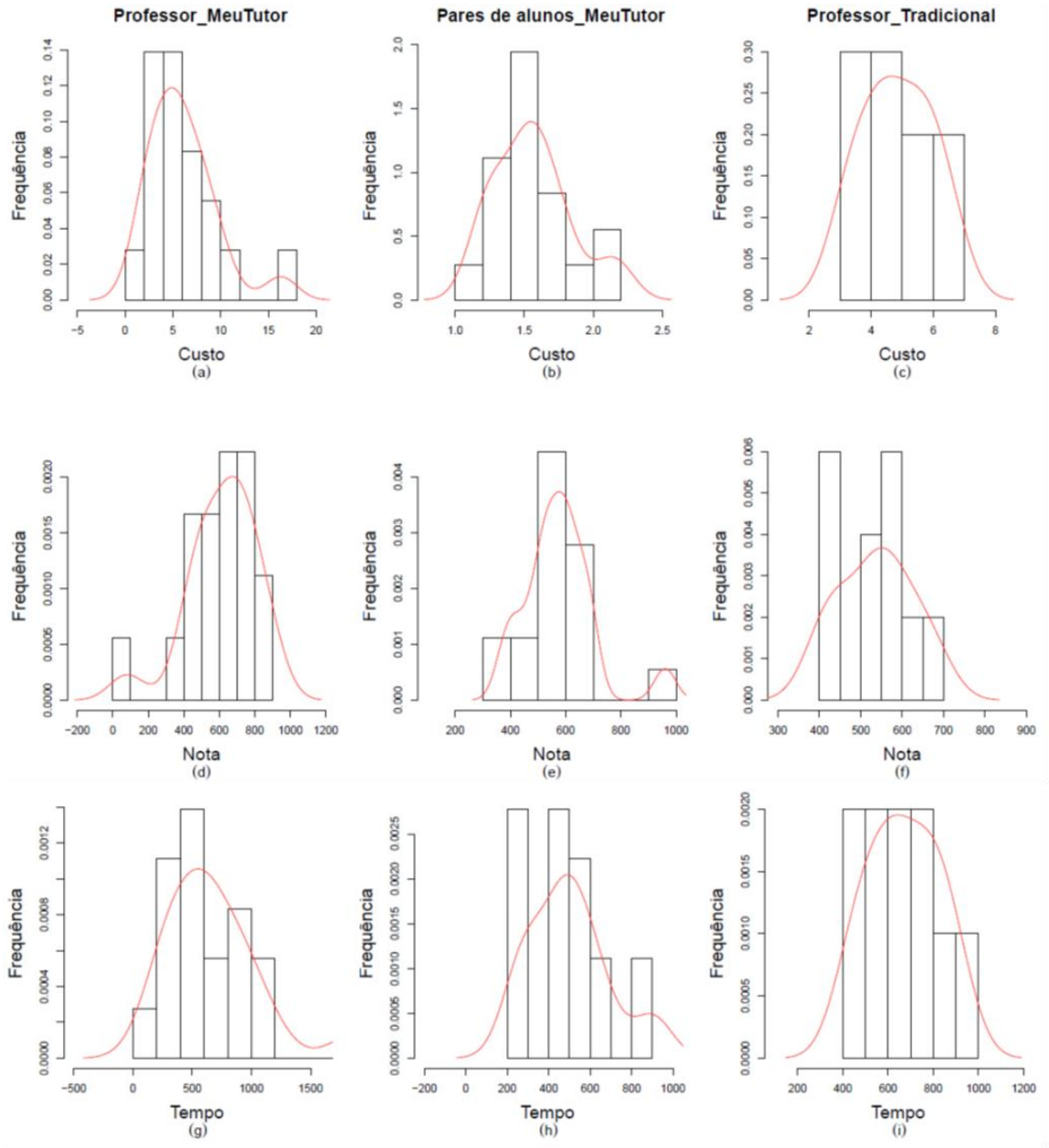


Figura 1.6. Histogramas das frequências das métricas do exemplo

6. Medidas de Posição

As medidas de posição mais importantes são as medidas de tendência central. Os dados observados tendem a se agruparem (se posicionarem) em torno dos valores centrais. Ou seja, ao redor dessas medidas encontram-se a maioria das observações. Quando nos referimos a média de uma variável aleatória *X*, ou seja a média da população *X*, usamos a notação $E(X) = \mu$, ou seja o E é o valor esperado}. Quando tratamos de uma amostra de *X, X1, ..., Xn* as medidas de tendência central mais utilizadas são a média amostral e a mediana amostral.

Seja X: as notas das avaliações dos professores no ambiente MeuTutor® (ver Seção 1.4.1). A amostra de tamanho *n* =18 desta população está apresentada na Tabela 1.3 e sua média amostral é calculada abaixo.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{18} X_i}{18} == \frac{X_1 + X_2 + \ldots + X_{18}}{18} = \frac{2.82 + 5.87 + \ldots + 6.13}{18} = 6.04$$

A mediana de um conjunto de valores ordenados crescentemente é o valor que o separa em dois subconjuntos de tamanhos iguais. Assim, metade dos valores são menores que mediana e a outra metade são maiores que a mediana. A mediana é muito útil, em especial quando há valores extremos que afetam de maneira acentuada a média aritmética.

7. Separatrizes

As separatrizes são medidas que se caracterizam por separar os dados. Estas medidas – os quartis, os decis e os percentis – são, juntamente com a mediana, conhecidas pelo nome genérico de separatriz, mas matematicamente são todos chamados de QUANTIS.

Considerando a amostra ordenada de forma crescente, os QUANTIS EMPÍRICOS são valores desta amostra ordenada que delimitam a porcentagem de valores menores ou iguais a eles. Vamos representar nossa amostra ordenada $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq X_{(4)} \leq \cdots \leq X_{(n)}$. Feito isto podemos obter o valor $x_{0.025}$ tal que $P\left(X_{(i)} \leq x_{(0.025)}\right) = 0.025$. $x_{0.025}$ é o quartil empírico 2,5% dos nossos dados. Em uma população normal padrão como veremos mais adiante, o quartil populacional 2,5% é igual a $z_{0.025} = -1.96$, implica dizer que 2,5% dos dados são menores que -1.96. Se $x_{(0.025)} \approx$ -1.96 é um sinal que nossa amostra pode ser de uma população normal padrão, isto com μ = 0 e$\sigma^2 = 1$.

Denominamos quartis os valores de uma série que a dividem em quatro partes iguais. Precisamos portanto de 3 quartis (Q1, Q2 e Q3) para dividir a série em quatro partes iguais. Assim, de 0 a 25% dos dados são inferiores à Q1; de 0 a 50% menores que Q2 e de 0 a 75% menores que Q3. Ou podemos dizer também, de 0 a 75% dos dados são maiores que Q1; de 0 a 50% maiores que Q2 e de 0 a 25% maiores que Q3. Q2 é sempre a mediana dos dados.

7.1 Variância, desvio padrão e coeficiente de variação

A variância de uma variável aleatória é uma medida da sua dispersão, indicando quão longe em geral os seus valores se encontram do valor esperado (valor médio). Quando o conjunto das observações é uma população, ela é chamada de variância populacional. Se o conjunto das observações é (apenas) uma amostra estatística, chamamos-lhe de variância amostral (ou variância da amostra).

A variância populacional de uma população Y, tal que Y_1 ... Y_n são TODOS os indivíduos que formam essa população, é dada por:

$$\sigma^2 = \frac{1}{N}\sum_{i=1|}^{N} (Y_i - \mu)^2,$$

onde μ é a média da população. Na prática, é quase sempre impossível achar o valor exato da variância populacional, devido ao tempo, custo e outras restrições de recursos.

Um método comum de estimar a variância da população é através de uma amostra desta população. Quando estimamos a variância da população usando uma amostra da população Y, tal que o tamanho da amostra é n, temos uma amostra do tipo $Y_1 \ldots Y_n$. A fórmula seguinte será um bom estimador para σ^2:

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

onde $\overline{Y}$ é a média da amostra.

O desvio padrão define-se como a raiz quadrada da variância. O coeficiente de variação é uma medida de dispersão que se presta à comparação de distribuições diferentes. O desvio-padrão é uma medida de dispersão relativa à média, e como duas distribuições podem ter médias diferentes, o desvio dessas duas distribuições não é comparável. A solução é usar o coeficiente de variação, que é igual ao desvio-padrão dividido pela média:

$$CV = \frac{\sigma}{\mu} \quad \text{e} \quad CV_{\text{amostral}} = \frac{S}{\overline{Y}}.$$

Algumas vezes, o coeficiente de variação é ainda multiplicado por 100, passando a ser expressado como percentagem. Abaixo apresentamos o script em R para obter as medidas que discutimos nas seções anteriores considerando as variáveis aleatórias: *X*, a qual representa a nota da avaliação por professor no ambiente MeuTutor; e *Y* que representa a nota da avaliação por pares conduzidas pelos alunos no ambiente MeuTutor.

```
X<-Custo_T1     Y<-Custo_T2
summary(X)  #A função summary calcula medidas como a média, a mediana e os quartis,
mas não calcula variância e desvio padrão.
#A função var() calcula a variância de uma variável aleatória de uma variável aleatória
#A função sd calcula o desvio padrão de uma variável aleatória
summary(cbind(X,Y)  #Aqui formamos um vetor através da função cbind(,)
e aplicamos a função summary ao vetor.
Ao abrir o arquivo Summary_XY.txt o resultado é o seguinte:
    " X"              "Y"
    "Min.   : 1.360  " "Min.   :1.190  " # mínimo
    "1st Qu.: 3.402  " "1st Qu.:1.325  " # Q1 - 25%
    "Median : 5.330  " "Median :1.525  " # Q2 - 50% (mediana)
    "Mean   : 6.038  " "Mean   :1.567  " # média
    "3rd Qu.: 7.907  " "3rd Qu.:1.692  " # Q3 - 75%
    "Max.   :16.290  " "Max.   :2.150  " # máximo
c1<-cbind(var(X), sd(X), ((sd(X))/mean(X)*100))  # No final temos o coeficiente de variação
```

Representamos os resultados obtidos acima na Tabela 1.4. Com base nesta tabela, podemos notar como o coeficiente de variação é útil para comparar a dispersão (variabilidade) de uma amostra de uma variável aleatória. Podemos perceber que o grupo com maior variabilidade é o associado ao método que utiliza o professor para a correção das redações associado ao MeuTutor. Se observássemos apenas o desvio padrão deste grupo para a variável custo, teríamos a impressão errônea que a dispersão do grupo é pequena. No entanto, o coeficiente de variação igual a 58,9% revela que essa é a amostra com maior variabilidade entre as nove variáveis estudadas (Tabela 1.4). Não

existe um limite para definir o quão dispersa é uma amostra, mas quanto menor seu coeficiente de variação (porém diferente de zero) mais informação a amostra contém sobre o problema estudado.

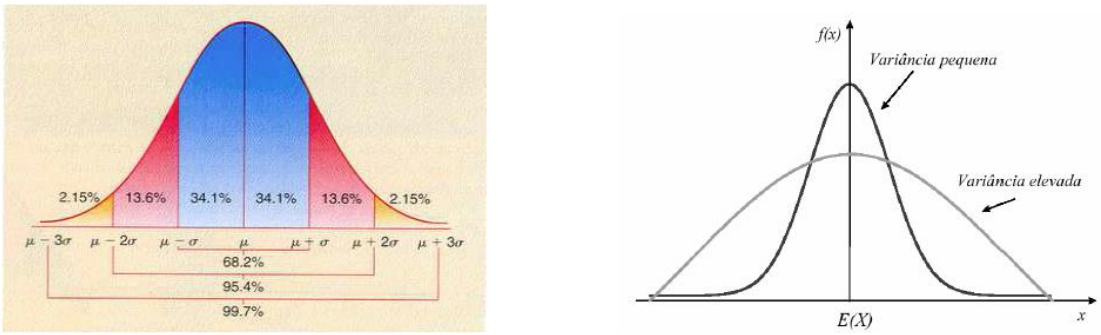
8. Distribuições Simétricas de Variáveis aleatórias

Toda variável aleatória apresenta uma distribuição de probabilidades associada a ela; a qual atribui probabilidades para valores (v.a. discreta) ou conjunto de valores (v.a. contínuas) que essa variável pode assumir. A seguir, apresentamos três tipos de distribuições para uma variável aleatória: Simétrica, assimétrica à direita e assimétrica à esquerda.

Tabela 1.4. Medidas de dispersão das amostras do experimento com pares de alunos

Grupos	Custo			Nota			Tempo		
	Var	Desvio	Coef. var.	Var	Desvio	Coef. var.	Var	Desvio	Coef. var.
Professor - MeuTutor	12.6	3.6	58.9%	37673.2	194.1	31.5%	159959.3	399.9	58.8%
Pares Alunos- MeuTutor	0.1	0.2	18.1%	17227.8	131.3	22.7%	37435.3	193.5	38.6%
Professor Trad.	1.3	1.2	23.9%	8248.9	90.8	16.9%	25508.5	159.7	23.9%

Em uma distribuição de probabilidades simétrica em torno de sua média μ, o seu gráfico à esquerda da média é um espelho do gráfico à direita, indicando que a forma como os valores se distribuem à esquerda da média é a mesma à direita. As caudas da distribuição possuem o mesmo formato. Um aspecto importante nas distribuições simétricas é que a média e a mediana são iguais. A mais importante distribuição simétrica é a Distribuição Normal (Figura 1.6a-b).



(a) Quantis (b) Dispersão

Figura 1.7. Distribuição normal

A variável aleatória com Distribuição Normal é contínua, tal que, $X = (-\infty, +\infty) = IR$. A variável aleatória X sempre representa uma população de dados. Esses dados tem uma média populacional que chamamos de μ e uma variância populacional que chamamos de σ^2. Se a variável aleatória X segue uma distribuição normal escreve-se: $X \sim N(\mu, \sigma^2)$. A maioria dos valores dessa variável deve se concentrar em torno de sua média, e seu desvio padrão σ^2 nos permite ter uma ideia da dispersão dessa variável aleatória (Figura 1.6). A distribuição normal padrão é representada por $X \sim N(0,1)$. E, de acordo com a Figura 1.6a temos que 95,4% dos

valores da variável estão entre os quantis -2 e 2, e que 95% dos valores da variável então entre os quantis -1.96 e 1.96.

8.1 Distribuições Assimétricas

Em uma distribuição assimétrica positiva, a cauda à direita é mais alongada que a cauda à esquerda, e a média, puxada pela cauda maior à direita, é maior que a mediana. Em uma distribuição assimétrica negativa, a cauda à esquerda é mais alongada que a cauda à direita, e a média, puxada pela cauda maior à esquerda, é menor que a mediana. Assimétrico à direita (assimetria positiva): maior frequência de valores pequenos. Assimétrico à esquerda (assimetria negativa): maior frequência de valores grandes (Figura 1.7).

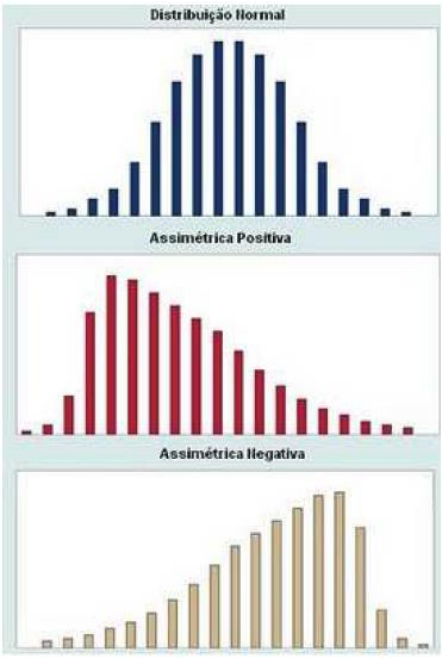


Figura 1.8. Distribuições simétricas e assimétricas

9. Intervalos de confiança baseados em populações normais

Dados amostrais são representações de uma variável aleatória X, cuja distribuição pertence a uma família de distribuições tipicamente conhecida, mas indexada por parâmetros desconhecidos θ, μ, σ^2, etc. O objetivo da inferência estatística é, com base na amostra observada, encontrar, definir, valores plausíveis para esses parâmetros.

9.1 Exemplo: Família Normal de Distribuições

Seja $X \sim N(\mu, \sigma^2)$, temos que $\theta = (\mu, \sigma^2) \in \theta \subset IR^2$, tal que as distribuições na família normal tem densidade na forma:

$$f(\mu, \sigma^2/x) = (2\pi\sigma^2)^{-1/2} \times \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad \mu \in IR, \ X \in IR, \ \sigma^2 > 0,$$

Note que para cada valor de $E(X) = \mu \in (-\infty, +\infty)$ e $Var(X) = \sigma^2 \in (0, \infty)$ temos uma distribuição de probabilidades diferente, pois temos uma densidade $f_\theta(x)$

diferente. Como não conhecemos μ e σ^2, precisamos estimá-los. Bons estimadores de μ e σ^2, respectivamente, são:

$$\widehat{\mu} = \overline{X} \quad \text{e} \quad \widehat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Mas, estes são estimadores pontuais que fornecem um único valor como estimativas para os parâmetros desconhecido. Estimadores intervalares, por outro lado, fornecem um intervalo de valores possíveis para o parâmetro e ainda associam confiança para esta estimativa intervalar.

9.2 Calculando Intervalos de Confiança em uma Distribuição Normal

Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n de $X \sim N(\mu, \sigma^2)$, com $\sigma^2 > 0$ conhecido. Temos que $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ que o melhor estimador intervalar para n com nível de confiança $1 - \alpha, \alpha \in (0,1)$ é

$$IC_\mu = \left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \quad \text{e} \quad \mu \in \left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \quad \text{com confiança} \quad (1-\alpha).$$

Note que α é pequeno pois $1 - \alpha$ deve ser grande. Tipicamente, é utilizado $\alpha =$ 0.05 que implica em intervalos de 0.95 de confiança. Ou intervalos de confiança de 95%. Na Figura 1.8 representamos graficamente como um intervalo de confiança $1 - \alpha$ é construído com base na população normal. A região cinza implica dizer que entre os pontos (quantis) $-Z\alpha_{/2}$ e $Z\alpha_{/2}$temos uma probabilidade $1 - \alpha$. Quando $\alpha = 0.05$, $-Z\alpha_{/2} = -1.96$ e $Z\alpha_{/2} = 1.96$, isto devido a simetria da normal.

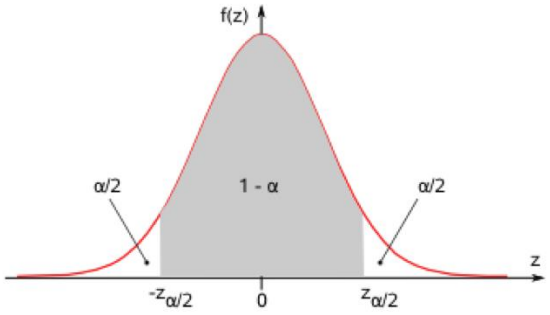


Figura 1.9. Gráfico da densidade de uma população normal segundo a probabilidade $1 - \alpha$

Quando σ^2 é desconhecido mas a amostra $X_1, \ldots, X_n$ é retirada de uma POPULAÇÃO X normal optamos por utilizar estimador usual de σ^2 que é $S^2 = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{(n-1)}$ e neste caso para a construção dos intervalos de confiança utilizamos os quantis da distribuição t de Student com n-1 graus de liberdade.

$$\left(\overline{X} - t_{(n-1);\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + t_{(n-1);\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \approx \left(\overline{X} - 2.1\frac{S}{\sqrt{n}}, \overline{X} + 2.1\frac{S}{\sqrt{n}}\right), \quad \alpha = 5\%; n \geq 15.$$

A distribuição t de Student é simétrica, bem aproximada pela distribuição normal, no entanto, com caudas mais pesadas. Como a distribuição t de Student depende do tamanho da amostra (n), pois seus graus de liberdade estão associados com n, segue

que quanto maior o valor de n mais próxima a distribuição t de Student se aproxima da normal. De fato para n próximo de 30 a aproximação da distribuição t de Student pela distribuição normal é consideravelmente boa. Note ainda que para $\alpha = 0.05$ e para um tamanho de amostra $n \geq 15$ o ERRO em torno de $\bar{X}$ pode ser aproximado por

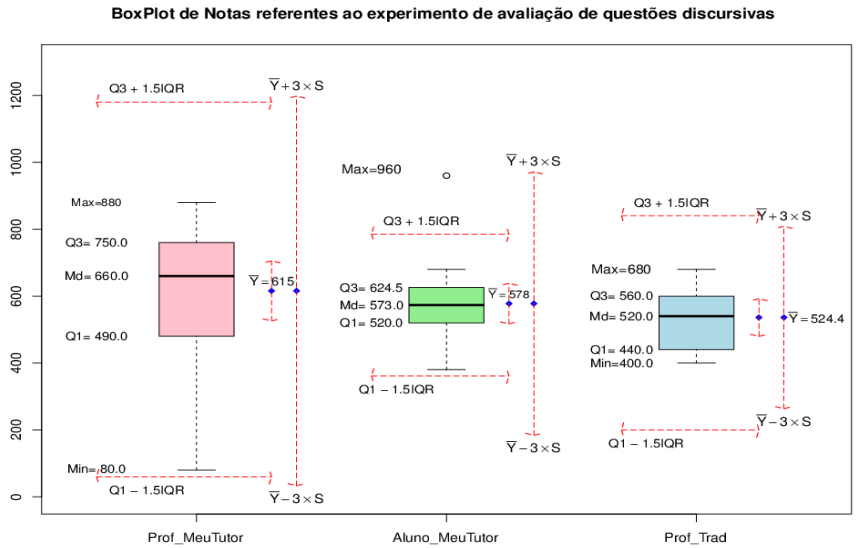
$$\delta_{\bar{X}} = 2.1\frac{S}{\sqrt{n}}.$$

9.3 Investigando a distribuição de probabilidades da população de onde retiramos a amostra

Como vimos na seção anterior, precisamos saber se a população de interesse é normal para construir os usuais intervalos de confiança. Um dos gráficos que nos permite ter uma noção da distribuição de onde os dados são retirados é o *boxplot*.

Gráfico *boxplot*

O *boxplot* é um gráfico que possibilita estimar a distribuição de um conjunto de dados numéricos com base em algumas de suas estatísticas descritivas, a saber: a mediana (Q2), o quartil inferior (Q1), o quartil superior (Q3) e do intervalo interquartil (IQR = Q3 - Q1), que é também uma medida de dispersão. Para exemplificar apresentamos na separadamente abaixo um recorte da Figura 1.9, que representa o *boxplot* das notas referentes ao experimento de avaliação por pares de questões discursivas do MeuTutor® .



A linha central da caixa marca a mediana do conjunto de dados. A parte inferior da caixa é delimitada pelo quartil inferior (Q1) e a parte superior pelo quartil superior (Q3). As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior a Q1 - 1.5IQR e do quartil superior até o maior valor não superior a Q3 + 1.5IQR. As quantidades Q1 - 1.5IQR e Q3 + 1.5IQR delimitam,

respectivamente, as cercas inferior e superior e constituem limites para além dos quais os pontos passam a ser considerados *outliers* (discrepantes).

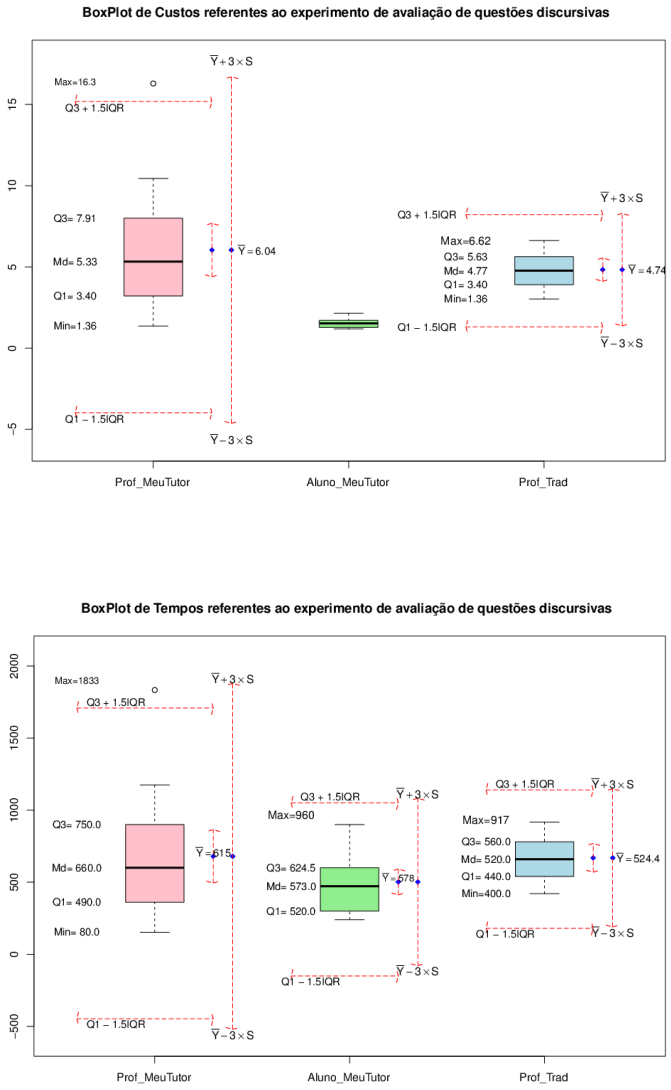


Figura 1.10. Gráficos boxplots para as variáveis Custo, Nota e Tempo do experimento da ferramenta de avalições escritas por pares

Na Figura 1.9 apresentamos os boxplots das amostras das variáveis custo, notas e tempo, considerando os três métodos de avaliação, professor utilizando o MeuTutor®, pares de alunos utilizando o MeuTutor® e professor usando formulário *online* disponibilizado pelos organizadores do experimento. Adicionamos ao usual gráfico de *boxplot* fornecido pelo R algumas medidas interessantes como um intervalo de confiança "supondo" normalidade. Nota-se como o grupo dos professores usando MeuTutor® apresenta para todas as variáveis, Custo, Nota e Tempo a maior dispersão entre todos os grupos. Também com base no boxplot notamos como o grupo pares de aluno MeuTutor® apresenta uma variabilidade muito baixa quanto ao Custo. Os

intervalos supondo normalidade e considerando um nível de confiança 99.7%, $[\bar{Y} - 3S; \bar{Y} + 3S]$ deixam claro que apesar dos boxplots terem detectado bem os *outliers*, esses pontos ainda são plausíveis para os dados.

O script R abaixo ilustra um exemplo de como criar um gráfico boxplot.

```
faixa <- range(-6,18)       cn1<-nrow(custo_T1)     mn.t <-(mean(Custo_T1))  sd.t <-(sd(Custo_T1))
er<-1.96*(sd.t/sqrt(cn1))   cn2<-nrow(custo2)       mn.t2 <-(mean(Custo_T2))  sd.t2 <-(sd(Custo_T2))
er2<-1.96*(sd.t2/sqrt(cn2)) cn3<-nrow(custo3)       mn.t3 <-(mean(Custo_T3))  sd.t3 <-(sd(Custo_T3))
er3<-1.96*(sd.t3/sqrt(cn3))
rb<-boxplot(Custo_T1,Custo_T2,Custo_T3,
main="BoxPlot de Custos referentes ao experimento de avaliação de questões discursivas "
,names=c("Prof_MeuTutor", "Aluno_MeuTutor", "Prof_Trad"),
col=cbind("pink", "lightgreen","lightblue"), ylim=faixa, boxwex=0.3)
xi <- 1 + 0.3      yi<-  1 + 0.4      zi<-  1-0.4     wi<-  1-0.4
points(xi, mn.t, col = "blue", pch = 18, cex=1.1)
arrows(xi, mn.t - er, xi, mn.t + er,
code = 3, col = "red", angle = 75, length = .1, lty=5 )
points(yi, mn.t, col = "blue", pch = 18, cex=1.1)
arrows(yi, mn.t - 3*sd.t, yi, mn.t + 3*sd.t,
code = 3, col = "red", angle = 75, length = .1, lty=5)
text(yi, (mn.t - 3*sd.t)-1, (expression(bar(Y) -  3%*%S) ) )
text(yi,(mn.t + 3*sd.t) + 1, (expression(bar(Y) +  3%*%S) ) )
text(zi, rb$stats[1,1], "Min=1.36", cex=0.9)     text(zi, rb$stats[2,1], "Q1= 3.40", cex=0.9)
text(zi, rb$stats[3,1], "Md= 5.33", cex=0.9)     text(zi, rb$stats[4,1], "Q3= 7.91", cex=0.9)
text(zi, 16.4, "Max=16.3", cex=0.8)              text(xi+0.23,  mn.t, (expression(bar(Y) == 6.04)), cex=0.9)
IQR = rb$stats[4,1] - rb$stats[2][1]
LI = rb$stats[2,1] - 1.5*IQR    LS = rb$stats[4,1] + 1.5*IQR
arrows(zi,LI,xi,LI,
code = 3, col = "red", angle = 75, length = .1, lty=5)
arrows(zi,LS,xi,LS,
code = 3, col = "red", angle = 75, length = .1,lty=5)
text(wi+0.1,  LI-0.4, "Q1 - 1.5IQR", cex=0.9)     text(zi+0.1,  LS -0.4, "Q3 + 1.5IQR" , cex=0.9)
```

10. Testes Estatísticos

Uma pesquisa parte de uma interrogação. Hipóteses são levantadas e a pesquisa pode invalidar ou confirmar as mesmas. Os testes estatísticos fornecem a possibilidade da verificação, através de procedimentos científicos das hipóteses levantadas.

Definição: Uma hipótese estatística é uma conjectura a respeito da distribuição de probabilidades de uma ou mais variáveis aleatórias. Para cada teste estatístico são formulados duas Hipóteses:

- A Hipótese Nula H_0 que é uma hipótese de que não haja diferenças, sempre;
- A Hipótese Alternativa H_1 contradiz de alguma forma a hipótese nula.

Uma vez que as hipóteses são formuladas, o próximo passo consiste em especificar o nível de significância: α. A seguir, apresentamos as probabilidades de rejeitar a hipótese nula e os dois tipos de erros:

- $\alpha = P(rejeitar H_0 / H_0 \text{ é } verdadeira)$: é a probabilidade de cometer o erro do tipo I, que é a probabilidade de rejeitar H_0 dado que H_0 é verdadeira. Como é um erro, tal probabilidade deve ser fixada em um valor bem pequeno, por exemplo, 0.01, 0.05, no máximo 0.1.

- $\beta = P(\text{não} rejeitar H_0 / H_0 \text{ é } falsa)$: é a probabilidade do erro do tipo II que deve também ser pequena, pois é a probabilidade de cometer um erro de não rejeitar H_0 dado que H_0 é falsa.

Note que $1 - \beta = P(rejeitar H_0 / H_0 \text{ é } falsa)$ é a probabilidade de tomar a decisão correta. Probabilidade de rejeitar H_0 quando H_0 **tem que ser rejeitada** esse é o poder do teste. Uma das características mais desejáveis de um teste é que ele seja altamente poderoso.

Deve existir um equilíbrio entre α e β. Pois se diminuirmos α, aumentamos β e consequentemente reduzimos o poder do teste (isto é muito grave). Matematicamente, o valor $\alpha = 0.05$ implica em um valor aceitável para β. Por isso que tipicamente utilizamos $\alpha = 5\%$ nas pesquisas. Se quisermos reduzir α e β, devemos aumentar n, que é o tamanho da amostra.

10.1 Testes de Hipóteses Paramétricos

Uma hipótese estatística paramétrica é uma conjectura a respeito dos parâmetros que definem as distribuições de probabilidades em questão. A seguir, definimos testes de hipóteses para diferença de médias de duas populações normais independentes e também descrevemos o p-Valor.

Teste de hipóteses para diferença de médias de duas populações normais independentes com Variâncias Conhecidas

Sejam $X_1, \ldots, X_n$ uma amostra aleatória de tamanho de $X \sim N(\mu, \sigma^2)$ e $Y_1, \ldots, Y_m$ uma amostra aleatória de $Y \sim N(\mu, \sigma_2^2)$. Aqui, X e Y são variáveis aleatórias independentes com $\sigma_1^1 > 0$ e $\sigma_2^2 > 0$ conhecidas. Antes de qualquer coisa precisamos saber o que queremos testar. Por exemplo, digamos que estamos interessados em construir um teste de hipóteses para testar se $\mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = 0$, ou seja, se a média da população X não difere da média da população Y. Para isso vamos considerar o estimador da diferença $\mu_1 - \mu_2$, dado por $\bar{X} - \bar{Y}$. Temos que $\bar{X} \sim N(\mu_1, \sigma_1^2/n)$ e $\bar{Y} \sim N(\mu_1, \sigma_2^2/m)$

$$\text{Assim,} \quad \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_1^2}{m}\right). \quad \text{Logo,} \quad Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

Então as hipóteses que podemos construir são

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad \begin{cases} a) & H_1 : \mu_1 \neq \mu_2 \quad \text{ou} \\ b) & H_1 : \mu_1 > \mu_2 \quad \text{ou} \\ c) & H_1 : \mu_1 < \mu_2. \end{cases}$$

A opção mais comum é a primeira (a) e é conhecida como teste bicaudal, pois se $\mu_1 \neq \mu_2$, então $\mu_1 > \mu_2$ ou $\mu_1 < \mu_2$. As áreas do teste bicaudal aparecem na Figura 1.12. E rejeitamos H_0 se $Z_{cal} > z_{\alpha/2}$ ou se $Z_{cal} < -z_{\alpha/2}$. Temos que Z_{cal} é o valor da estatística Z definida acima calculado com base nos valores amostrais de X e Y. É importante ressaltar que Z_{cal} é calculado considerando H_0 verdadeira. Então,

$$Z_{cal} = \frac{(\overline{X}_{cal} - \overline{Y}_{cal})}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}, \text{ em que } \overline{X}_{cal}, \overline{Y}_{cal} \text{ são valores obtidos com base nos dados (amostras).}$$

Quando não conhecemos as variâncias das duas populações a alternativa é usar os estimadores dessas variâncias. Surge então o famoso **Teste t-Student**; Assim, a estatística do teste é dada por

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \text{t-Student}_{(n+m-2)}, \quad \text{onde}$$

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}; \quad S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \quad \text{e} \quad S_y^2 = \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m-1}.$$

Em que t-$Student_{(n+m-2)}$ refere-se à distribuição t de Student com n+m-2 graus de liberdade.

p-Valor

O p-Valor é definido como $P(Z > Z_{cal}) + P(Z < -Z_{cal})$ no caso da distribuição normal e de testes bilaterais. E no caso do teste t-Student p-Valor $P(T > T_{cal}) + P(Z < -T_{cal})$. E rejeitamos H_0 SEMPRE quando p-Valor $< \alpha$. No limite se p-Valor $\cong \alpha$ e conhecemos bem se o teste é PODEROSO decidimos por rejeitar H_0. Exemplo: digamos que p-Valor =0.05202 e o teste é reconhecidamente PODEROSO e ROBUSTO (não sofre muito com pontos atípicos e variações no dados), podemos decidir por REJEITAR H_0. Note que os testes acima são totalmente dependentes da suposição que as populações comparadas são normais. Na prática não temos as populações, mas podemos usar as respectivas amostras para verificar se a suposição de normalidade é pelo menos plausível.

10.2 Testes de Hipóteses Não Paramétricos

Os testes Não Paramétricos são aqueles cujos modelos teóricos não especificam condições sobre os parâmetros da população de onde os dados foram obtidos. Quando não são feitas conjecturas sobre os parâmetros. Podem ser aplicados a dados discretos, contínuos, classificados em classes (ou categorias). A seguir descrevemos os testes de normalidade e a gráfico do tipo QQPlot.

Testes de Normalidade

Nesta seção apresentamos um teste não paramétrico que apresenta excelente desempenho quanto ao poder e a robustez, tanto para amostras grandes quanto para amostras pequenas (Thas, 2010).

Teste Anderson-Darling (ad): Seja $F(x) = P(X \leq x)$ a função de distribuição acumulada de uma população normal. Seja $G(x)$ a função de distribuição empírica dos dados (EDF em inglês), que pode ser definida como a função de distribuição acumulada das frequências relativas. O teste **ad** avalia se $G(x) \approx F(x)$. Então, nossas hipóteses são:

- H_0: os dados são provenientes de uma distribuição normal;
- H_1: os dados não são provenientes de uma distribuição normal;

A seguir é apresentado um script em R que ilustra como realizar este teste no conjunto de dados do exemplo da ferramenta de avaliações escritas feita por pares. O respectivo p-Valor resultante também é apresentado ao lado de cada teste.

```
> ad.test(Custo_T1)  p-value = 0.1409
> ad.test(Custo_T2) p-value = 0.1609
> ad.test(Custo_T3) p-value = 0.9002
> ad.test(Nota_T1) p-value = 0.3349
> ad.test(Nota_T2) p-value = 0.18
> ad.test(Nota_T3) p-value = 0.7965
> ad.test(Tempo_T1) p-value = 0.1412
> ad.test(Tempo_T2) p-value = 0.1745
> ad.test(Tempo_T3) p-value = 0.8979
```

Vemos acima que todos os p-Valores são maiores que 5%. Não rejeitamos a possibilidade das populações, considerando todas as variáveis do estudo, serem normais. Então, podemos realizar o teste t-Student para testar: $H_0 = \mu_1 - \mu_2 = 0$ contra $H_1 = \mu_1 \neq \mu_2$. A seguir realizamos o teste t-Student para os mesmos dados acima. Segue o script em R:

```
> t.test(Custo_T1,Custo_T3) p-value = 0.2002
> t.test(Nota_T1,Nota_T2)   p-value = 0.498
> t.test(Nota_T1,Nota_T3)   p-value = 0.153
> t.test(Nota_T2,Nota_T3)   p-value = 0.3333
> t.test(Tempo_T1,Tempo_T2)  p-value = 0.1024
> t.test(Tempo_T1,Tempo_T3) p-value = 0.9246
> t.test(Tempo_T2,Tempo_T3) p-value = 0.02222
> t.test(Custo_T1,Custo_T2) p-value = 5.405e-05
> t.test(Custo_T2,Custo_T3)  p-value = 6.601e-06
```

Notem que apenas as médias dos custos são diferentes. Se olharmos os dados, o custo da correção por pares de alunos (Custo_T2) parece ser menor que o custo da correção do professor pelo MeuTutor® e que o custo da correção do Professor pelo questionário *online*. Podemos realizar um teste t-Student unilateral para testar $H_0: \mu_{CustoT2} = \mu_{CustoT1}$ contra $H_1: \mu_{CustoT2} < \mu_{CustoT1}$ e $H_0: \mu_{CustoT2} = \mu_{CustoT3}$ contra $H_1: \mu_{CustoT2} < \mu_{CustoT3}$. Como vemos abaixo (script em R), rejeitamos as duas $H_0's$. O custo do método de correção por pares de alunos é menor que os custos dos demais métodos.

```
t.test(Custo_T2,Custo_T1, alternative = c("less")) p-value = 2.702e-05
t.test(Custo_T2,Custo_T3, alternative = c("less")) p-value = 3.301e-06
```

Podemos realizar testes não paramétricos para tentar chegar às mesmas conclusões acima. No entanto, as hipóteses são bem diferentes.

Testes de Kruskal-Wallis para várias amostras:

⋆ $H_0 : F_1(x) = F_2(x) = F_2(x) = F_3(x) \ldots F_k(x)$. As K amostras são provenientes da mesma distribuição;

⋆ $H_1 : F_t(x) \neq F_j(x), t \neq j$. Pelo menos duas amostras NÃO são provenientes da mesma distribuição;

Segue um comando em R para este teste:

```
kruskal.test(list(Custo_T1, Custo_T2, Custo_T3)) p-value = 4.953e-07
kruskal.test(list(Custo_T1, Custo_T3))           p-value = 0.5327
kruskal.test(list(Custo_T1, Custo_T2))           p-value = 2.356e-06
kruskal.test(list(Custo_T3, Custo_T2))           p-value = 1.52e-05
Wilcoxon rank sum test with continuity correction
wilcox.test(Custo_T2,Custo_T3, alternative = c("less")) p-value = 1.273e-06
wilcox.test(Custo_T2,Custo_T3, alternative = c("less")) p-value = 8.471e-06
```

No teste de wilcox acima temos que:

$$\star H_0 : F_{Custot2}(x) = F_{Custot1}(x) \text{ contra } H_1 : F_{Custot2}(x) < F_{Custot1}(x).$$

QQplot

Ainda podemos usar a ferramenta gráfica QQplot do R para avaliar aderência dos dados com a distribuição normal. Esse gráfico cruza os quantis empíricos da amostra com os verdadeiros quantis da distribuição normal. Se os pontos se ajustam bem à reta é sinal de uma boa aderência à distribuição normal. Esse gráfico associado ao histograma com a densidade estimada pode ser muito útil para investigar normalidade. Segue um comando em R para gerar gráficos QQplot.

```
qqnorm(Custo_T1,xlab="quantis normais",ylab=" ", main = "Professor_MeuTutor", pch=20,cex.axis=0.7)
qqline(Custo_T1, distribution = qnorm,  col = 2,  probs = c(0.089, 0.811))
mtext(expression(Custo),line=2.7,side=2,cex=1.2)
mtext(expression((a)),line=4.5,side=1,cex=0.8)
```

De fato, todos os teste estatísticos se baseiam em suposições matemáticas rigorosas, em especial, os teste para avaliar normalidade dos dados. Esses testes, sobre normalidade, são muito sensíveis à presença de ouliers, por exemplo. Assim, a conclusão sobre normalidade dos dados pode ser duvidosa. As ferramentas gráficas surgem exatamente como um suporte para validar o resultado dos testes.

Consideremos o exemplo de avaliação por pares. Com base nos testes estatísticos não rejeitamos as hipóteses de distribuição normal considerando todas as variáveis envolvidas no problema. Mas, desde da visualização do histograma já era possível notar desvios em relação à suposição de normalidade das variáveis. Note que as três variáveis: Custo, Nota e Tempo são positivas, mas a densidade normal estimada pelo Histograma (Figuras 1.5a e 1.5d) sugere a possibilidade de ocorrerem valores negativos, e os QQplots (Figuras 1.10a,d) apresentam desvios da normalidade. De fato, os gráficos, em especial os histogramas, apontam para uma distribuição levemente assimétrica positiva, que poderia ser a gama, uma distribuição interessante para competir com a normal, uma vez que a distribuição gama é própria para modelar dados assimétricos positivos, como veremos na seção 1.11.

Em relação aos QQplots espera-se uma aderência dos pontos à reta em vermelho, quando a variável aleatória apresenta distribuição normal. Se essa distribuição normal for a padrão, isto é, com média “0” e variância “1”, então a reta vermelha deve ser a primeira bissetriz. Com base nas Figuras 1.10 é possível perceber que nenhum dos dois comportamentos esperados acontecem para as variáveis relacionadas com as correções feitas pelo professor utilizando MeuTutor® e alunos

usando MeuTutor® (avaliação por pares), evidenciado que a suposição de normalidade para estas variáveis, pode ser contestada, até porque os dados são positivos.

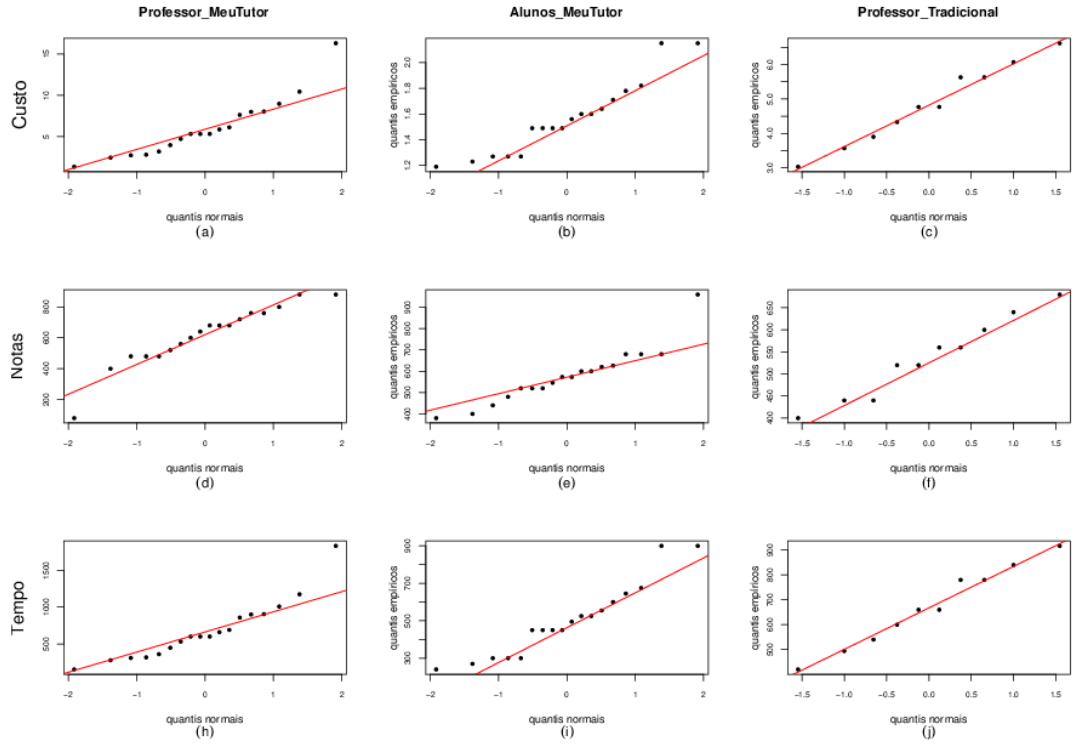


Figura 1.11. QQplots para os dados do experimento que avalia a ferramenta de avaliação por pares

10.4 Testes de Independência

Para ilustrar este conceito de independência vamos retomar nosso exemplo apresentado na Seção 1.4.1: Queremos saber se os anos de experiência com ensino dos avaliadores influencia sua caracterização de perfil do aluno. Isto poderia ser feito com base no teste não paramétrico de independência Qui Quadrado, ($test\,\chi^2$), O teste compara as frequências observadas das frequências esperadas. No entanto, o teste só pode ser realizado se a frequência esperada de cada célula for maior que 5% [Barbetta, Reis e Bornia, 2008], caso contrário o teste fica comprometido e o resultado pode ser inválido, como vemos abaixo, no script em R:

```
dados<-scan("C:\\Users\\patricia\\Patricia\\Patricia_2015\\Minicurso\\Quest.txt",
list(C1=0,C2=0,C3=0,C4=0,C5=0,C6=0,C7=0,C8=0,C9=0,C10=0,C11=0,C12=0,C13=0,C14=0,C15=0,C16=0))
attach(dados)
dados<-data.frame(dados)
Zero<-cbind(C1,C5,C9,C13)
Um<-cbind(C2,C6,C10,C14)
Seis<-cbind(C3,C7,C11,C15)
MSeis<-cbind(C4,C4,C12,C16)
```

```
S_Zero<-rowSums(Zero)
S_Um<-rowSums(Um)
S_Seis<-rowSums(Seis)
S_Mseis<-rowSums(MSeis)
M1<- as.table(rbind(S_Zero,S_Um,S_Seis,S_Mseis))
dimnames(M1) <- list(Experiência = c("0","-1","1+-6","6+"),Perfil = c("Col.","Game","Ped","Soc","NPR"))
> M1
Perfil
Experiência Col. Game Ped Soc NPR
0       6    3  4   4    2
-1      4    7  2   1    6
1+-6    8    8  0   2   11
6+     10    6  3   2    2
> chisq.test(M1)  # Prints test summary    p-value = 0.09702
Warning message:
In chisq.test(M1) : Chi-squared approximation may be incorrect

> chisq.test(M1, simulate.p.value=T, B=9999) ##Pearson's Chi-squared test with simulated p-value (based on 9999 replicates) p-value = 0.0915
```

Se usássemos o p-Valor fornecido pelo R, diríamos que a definição do perfil do estudante NÃO depende da experiência do avaliador, pois NÃO rejeitamos a hipótese nula de independência. No entanto, os gráficos sugerem outra coisa, uma vez que as classificações de perfis atribuídas pelos avaliadores com mais de seis anos de experiência em ensino se diferenciam das classificações de perfis atribuídas por avaliadores com menor tempo de experiência, em especial quando os alunos 3 e 4 são avaliados . Esse é um exemplo de que o desenho de experimento não facilitou análises mais profundas.

Podemos chegar às conclusões acima com informações adicionais muito relevantes utilizando modelos de regressão.

11. Modelos de Regressão

Em muitos casos, temos conhecimento limitado sobre a relação entre variáveis envolvidas em um problema de interesse. Se visualizamos os valores observados de tais variáveis como os resultados de um experimento devemos ter, então, uma ferramenta teórica, um modelo matemático através do qual estas variáveis estejam relacionadas, para atuar como base do processo gerador de dados. No entanto, todos os modelos são inevitavelmente simplificações da realidade e além das variáveis possíveis de serem mensuradas existem fatores que não podem ser controlados, ou que são desconhecidos, os quais podem ser considerados através de uma componente casual, representada pelos erros aleatórios. E é neste contexto que se inclui o modelo de regressão linear. Considere o seguinte modelo linear:

$$y = X\beta + \epsilon = \tag{1.1}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}}_{\begin{pmatrix} x_1 & x_2 & \dots & x_k \end{pmatrix}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Podemos ver que:

- y é um vetor de n observações da variável aleatória dependente, ou ainda, $y_1, \dots, y_n$ é uma amostra da população de interesse;
- X é uma matriz $n \times k$ formada pelas covariadas. Note que cada coluna de X é um conjunto de n observações da covariada $x_t, t = 1, \dots, k$. É importante ressaltar que X não é variável aleatória, ela é observada e fixa;
- Ainda temos que β é um vetor de k parâmetros também fixos e desconhecidos (não são variáveis aleatórias);
- Finalmente, ϵ é um vetor de n erros aleatórios com média zero $(E)(\epsilon_i) = 0$ e variância constante ao longo das observações, isto é, $var(\epsilon_i) = \sigma^2$ para todo $i = 1, \dots, n$.

Em um modelo de regressão, tentamos que nosso modelo matemático explique ao máximo possível o carácter aleatório de nossa resposta de forma que, o que não consigamos explicar esteja contido no erro aleatório ϵ. Como queremos explicar o máximo com base no modelo que envolve as covariadas e os parâmetros desconhecidos, o erro deve ser bem pequeno, ou seja, ESPERAMOS que ele seja zero. Por isso, uma das principais suposições de modelos lineares de regressão é que

$$\mathrm{E}(\epsilon) = \mu_\epsilon = 0 \tag{1.2}$$

Este fato nos traz consequências muito importantes, vamos calcular o valor esperado da expressão em (1.1) com base na suposição em (1.2).

$$\mathrm{E}(y) = \mathrm{E}(X\beta) + \mathrm{E}(\epsilon) \Leftrightarrow \mathrm{E}(y) = X\beta \Leftrightarrow \mu = X\beta. \tag{1.3}$$

Então, como vemos em (1.3) na prática, o nosso modelo final é

$$\mu = X\beta \tag{1.4}$$

Note em (1.3) que $E(X\beta) = X\beta$ por que nem X nem β são variáveis aleatórias (**o valor esperado de uma constante é uma constante**). Vamos ver o que ocorre com a variância,

$$\mathrm{var}(y) = \underbrace{\mathrm{var}(X\beta)}_{0} + \underbrace{\mathrm{var}(\epsilon)}_{\sigma^2} \Leftrightarrow \mathrm{var}(y) = \sigma^2 \tag{1.5}$$

Pois é, $var(X\beta) = 0$ por que nem X nem β são variáveis aleatórias (**a variância de uma constante é ZERO, e se a variância é zero é porque é uma constante, não é variável aleatória**). Podemos então representar nosso modelo considerando a i-ésima observação como:

$$\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Para conhecermos de fato o modelo acima precisamos estimar $\beta_1, \beta_2 \dots, \beta_k$, **que são conhecidos como coeficientes do modelo de regressão.** Tipicamente, fazemos isto usando o método de máxima verossimilhança [Lehmann e Casella, 1998]. Assim, obtemos $\hat{\mu}_i$ quando obtemos $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ tal que,

$$\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n.$$

Agora podemos obter estimativas para μ_i. Diante do que apresentamos, começamos a entender que, o que realmente nos interessa é a distribuição de *y*, sua média, sua variância e o tipo de distribuição de probabilidades que a variável aleatória *y* (nossa resposta) segue. Pois bem, a distribuição mais conhecida é a distribuição normal. No entanto, na prática essa distribuição não é adequada para diversos tipos de variáveis aleatórias:

- Se μ é a média da variável resposta pode assumir tanto valores positivos quanto valores negativos e a curva de densidade de *y* é próxima da forma de sino, então se justifica pensar na distribuição normal;
- Se μ só pode assumir positivos e a curva de densidade de *y* é simétrica positiva, devemos pensar na distribuição gama;
- Se a variável aleatória y, nossa resposta, representa dados de contagens, por exemplo, $y = 0, 1, 2, 3 ...$, v.a. discreta, podemos pensar na distribuição binomial;
- Se $y \in (0,1)$ podemos pensar na distribuição beta ou na distribuição simplex.

Este último exemplo é particularmente útil nesta área, pois no contexto de diversos problemas da informática na educação muitos escores, notas, desempenho, etc., são tratados, os quais podem ser taxas, proporções, índices, por exemplo. Neste caso se encontram em um intervalo do tipo (a, b) que pode ser transformado facilmente para o intervalo $(0,1)$. Neste ponto, precisamos generalizar nosso modelo linear com o objetivo de permitir o uso de diversas distribuições além da normal. Fazemos isso considerando a expressão abaixo.

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}$$

Onde $g(\mu_i)$ é uma função de ligação que conecta a média da variável resposta e o modelo envolvendo as covariadas e os $\beta's$.

Por que usamos essa $g(\mu_i)$? Quando $y_i \sim N(\mu_i, \sigma^2)$, temos que $g(\mu_i) = \mu_i$, ou seja *g* é o que chamamos de função identidade. Isto acontece no modelo normal porque assim como a resposta que pertence a todos os reais, isto é, $y \in (-\infty, +\infty)$, o mesmo ocorre com sua média $\mu \in (-\infty, +\infty)$ = IR. Quando obtemos $\hat{\mu}_i$ a mesma coisa deve acontecer, ou seja, $\hat{\mu}_i \in (-\infty, +\infty) = IR$. Daí, note que $\hat{\mu}_i$ pode assumir qualquer valor Real. Assim, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ estão LIVRES para também assumir qualquer valor.

Isto não acontece, por exemplo, se a variável resposta segue uma distribuição gama. Porque, assim como a resposta que assume valores reais positivos, isto é, $y \in (0, +\infty)$, o mesmo ocorre com sua média $\mu \in (0, +\infty) = IR^+$ e DEVE acontecer com $\hat{\mu}$, ou seja, tem que acontecer: $\hat{\mu} \in (0, +\infty) = IR^+$. Com esta restrição os $\beta's$ não estão livres, pois temos que garantir que $X\hat{\beta}$ só assuma valores reais positivos. Em *X* não podemos mexer (fixa e conhecida), então teríamos que realizar um processo de estimação do $\beta's$ com restrição, para garantir que $X\hat{\beta} \in (0, +\infty) = IR^+$. Este processo pode ser bastante complicado.

ENTÃO a alternativa é aplicar uma função *g* em μ_i de forma que $g(\mu_i) \in (-\infty, +\infty) = IR$. Daí então os $\hat{\beta}'s$ estão LIBERADOS. Vamos ao exemplo da gama, seja *y* uma variável aleatória com distribuição gama, aqui denotada por $Y \sim G(\mu, \phi)$, tal que

$$f_{(\mu,\phi)}(y) = \frac{1}{\Gamma(\phi)}\left(\frac{\phi x}{\mu}\right)^{\phi}\exp\left(-\frac{\phi y}{\mu}\right)\frac{1}{y},\ y \geq 0, \mu > 0, \phi > 0, \Gamma(\phi) = \int_0^{\infty} t^{(\phi-1)}e^{-t}dt.$$

Se nossa variável resposta pertence aos reais positivos e decidimos usar a distribuição gama, um modelo de regressão adequado seria

$$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, i = 1, \ldots, n.$$

Isto porque o logaritmo só pode ser calculado para valores reais positivos e o seu resultado assume valores em todos os reais, o que libera os $\hat{\beta}'s$. Note que $\log(u) \in (-\infty, +\infty)$ para todo $u \in (0, +\infty) = IR^{+}$.

Vamos a um exemplo: Experimento de correção de provas discursivas por pares no ambiente MeuTutor. Testamos o modelo:

$$\log(\mu_i) = \beta_1 + \beta_2 IndicadoraCustoT2$$

em que μ_i é a média do Custo considerando todos os três grupos e IndicadoraCustoT2 é uma variável que assume o valor "1" quando o grupo é o de pares de alunos usando MeuTutor e "0" para os outros dois grupos. Segue um script em R tratando ilustrando a construção deste modelo.

```
>fit <- glm(t(Custo)~t(Ind_T2), family=Gamma(link=log))# função para glm - generalized linear model -
modelos lineares generalizados - supondo distribuição gama para a resposta e usando a função de ligação log.
> summary(fit)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7240     0.0809   21.31  < 2e-16 ***
t(Ind_T2)    -1.2751     0.1293   -9.86 1.03e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gamma family taken to be 0.1832411)
Null deviance: 22.5610  on 45  degrees of freedom
Residual deviance:  7.1323  on 44  degrees of freedom
AIC: 157.56
```

Então, com base nos resultados acima temos que a covariável indicadora do grupo pares de aluno no MeuTutor® é altamente significativa, a um nível maior que 0.00001, ou seja, rejeitamos H_0: $\beta_2 = 0$. De fato, o método proposto de correções por pares tem um custo diferente do custo dos outros dois grupos.

Até já sabíamos isso, também sabíamos que era um custo menor, o que é confirmado pelo sinal negativo da estimativa $\hat{\beta}_2 = -1.2751$. Mas, agora temos uma informação adicional com base no modelo de regressão. Temos que $\{1 - \exp(\hat{\beta}_2)\} \times 100$ representa a redução percentual ao implantarmos o sistema de correção por pares de aluno. Pois bem, esse valor é igual a 72%. A fórmula acima ocorre pois usamos como função de ligação o logaritmo, então para obter a interpretação das estimativas dos coeficientes da regressão aplicamos a função inversa, neste caso a exponencial. Se o resultado for menor que "1", como ocorreu no nosso caso, então subtraímos: "1" – (menos) essa exponencial, e depois multiplicamos por 100 para transformar em um percentual. Se o resultado da exponencial for maior que "1", então fazemos a

exponencial menos "1". Para mais detalhes sobre modelos lineares generalizados, acesse https://www.ime.usp.br/~giapaula/cursospos.htm.

Agora vamos a um exemplo de dados categóricos usando Modelos Lineares Generalizados: os dados de perfil dos alunos. Queremos saber se a experiência do avaliador ou o aluno avaliado (que depende do tipo de gráfico utilizado) interferem na porcentagem atribuída a cada uma das cinco possibilidades de respostas sobre perfil, a saber: colaborativo, gamificação, pedagógico, social, não há resposta possível. Esses percentuais estão demostrados no gráfico 1.4. Por exemplo, tivemos cinco avaliadores com zero experiência para o aluno 1 e quatro consideraram que o aluno 1 tem perfil colaborativo, tem quatro sucessos para colaborativo em cinco tentativas (cinco respostas possíveis), igual a uma probabilidade de sucesso de 0.8. Vamos tentar identificar se a probabilidade de ser colaborativo depende da experiência do avaliador e\ou do aluno avaliado.

O Número de sucessos em n tentativas caracteriza a DISTRIBUIÇÃO DISCRETA BINOMIAL, se $y \sim binomial(n,p)$, temos que $E(y) = \mu_y = np$. De fato, o que temos inicialmente é, por exemplo, a quantidade de avaliadores com nenhuma experiência que classificam o aluno 1 no perfil colaborativo, denotada aqui por y_{111}. O código 111 significa: 1 – avaliador sem experiência; 1- perfil colaborativo; 1- Aluno 1.

Mas, o que calculamos é y_{111}/n_{11}, percentual de classificação como colaborativo em um total de n_{11}: quantidades de avaliadores sem experiência para o Aluno 1. Assim, nossa população pode ser descrita como $y/n \sim binomial(\mu)$ pois, $E(y/n) = E(y)/n = np/n = p = \mu \in (0,1)$. Como $\mu \in (0,1)$, uma função de ligação adequada é a logito descrita abaixo:

$$\mu_i \in (0,1) \leftrightarrow \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\} \in (-\infty, +\infty) = IR.$$

Assim, nosso modelo é definido como:

$$log\left\{\frac{\mu}{(1-\mu)}\right\} = \beta_1 + \beta_2\, Exper2 + \beta_3\, Exper3 + \beta_4\, Exper4 + \beta_5 Aluno2 + \beta_6\, Aluno3 + \beta_7 Aluno4.$$

As covariáveis do modelo acima são conhecidas como variáveis Dummy. Elas só assumem valores zero ou um. Por exemplo, a variável Experiência contém quatro categorias, que vamos codificar em: categorias 1, 2, 3 e 4 e vamos usar variáveis dummies para representa-las. Mas aqui temos um problema. Se a variável qualitativa assume k categorias, então, só podemos criar (k -1) variáveis dummies devido a uma propriedade matemática de matrizes. No nosso exemplo, Exper2: é "1" se o avaliador tem menos de um ano de experiência e "0" caso contrário. Exper3: "1" se o avaliador tem mais de um ano e menos de seis anos de experiência, e "0" caso contrário. Exper4: "1" se o avaliador tem mais de zero ano de experiência, e "0" caso contrário. E nenhuma experiência? Ocorre quando Exper2=Exper3=Exper4=0. Agora, a estimativa do coeficiente de nenhuma experiência vai depender também do valor que assume Aluno2, Aluno3, Aluno4. Por exemplo, se Exper2=Exper3=Exper4=Aluno2=

Aluno3=Aluno4=0, então β_1 capta a influência do avaliador sem experiência para o Aluno 1 (y_{11}). Veja o código do R abaixo.

```
>dados<-scan("C:\\Users\\patricia\\Patricia\\Patricia_2015\\Minicurso\\PerfisAlunos.txt",
+ list(Exper=0, Aluno=0, Perc=0))
Read 80 records
> attach(dados)
#Aqui estão parte dos dados, agora com as variáveis Experiência codificada em quatro categorias (1, 2, 3, 4) e Aluno
Também em quatro categorias. Estamos tentando modelar o percentual dos perfis segundo as Variáveis Alunos e
Experiência.


1  1     0.8
1  1     0.0
1  1     0.0
1  1     0.0
1  1     0.2
2  1     0.8
2  1     0.0
2  1     0.0
2  1     0.0
2  1     0.2
3  1     0.625
3  1     0.000
3  1     0.000
3  1     0.250
3  1     0.125
4  1     1

> fnames<-list(Exper=c("1","2","3","4"))
> Exper<-factor(Exper)
> Exper<-C(Exper,treatment) # Aqui estamos avisando que 1, 2, 3 e 4 não são valores são categorias
> fnames<-list(Perfil=c("1","2","3","4","5"))
> Aluno<-factor(Aluno)
> Aluno<-C(Aluno,treatment)
> fit1<-glm(Perc~Exper+Aluno,family=binomial) # Aqui ajustamos o modelo binomial

Warning message: # Essa mensagem deve-se ao fato de haver muitos percentuais iguais a zero.
In eval(expr, envir, enclos) : non-integer #successes in a binomial glm!

> summary(fit1)
Call:
glm(formula = Perc ~ Exper + Aluno, family = binomial(link= logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.386e+00  7.395e-01  -1.875   0.0608 .

Exper2      -5.829e-13  7.906e-01   0.000   1.0000
Exper3      -9.591e-13  7.906e-01   0.000   1.0000

Exper4      -1.042e-12  7.906e-01   0.000   1.0000
Aluno2       2.471e-15  7.906e-01   0.000   1.0000
Aluno3       3.649e-12  7.906e-01   0.000   1.0000
Aluno4       8.531e-14  7.906e-01   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
```

Null deviance: 40.662 on 79 degrees of freedom
Residual deviance: 40.662 on 73 degrees of freedom
AIC: 80.201.

O resultado acima, demostra que nenhum dos fatores de experiência com ensino e nenhum dos fatores de qual aluno está sendo avaliado interferem nos percentuais de classificação dos alunos nas possíveis respostas sobre o perfil, pois os p-Valores referentes a todos os fatores das duas variáveis categóricas são iguais a UM. Ou seja, com certeza, NÃO rejeitamos as hipóteses: $H_0: \beta_2 = 0$, (Exper2); $H_0 : \beta_3 = 0$, (Exper3); $H_0 : \beta_4 = 0$ (Exper4); $H_0 : \beta_5 = 0$, $H_0 : \beta_6 = 0$, $H_0 : \beta_7 = 0$ (Aluno 4).

Mas, esse resultado não é necessariamente verdadeiro, pois o ajuste do modelo foi prejudicado pela baixa frequência de alguns perfis dentro dos fatores investigados.

12. Planejamento de Experimento

A seguir apresentamos alguns passos para a construção de um bom experimento. Primeiramente, alguns questionamentos devem ser feitos:

1. O que queremos investigar?
2. Por que queremos investigar esse tema?
3. Que características deste tema queremos mensurar?

Após isso, algumas questões relacionadas com os métodos estatísticos devem ser consideradas:

- Que tipos de testes estatísticos queremos realizar? Verificar as suposições desses testes e tentar cumpri-las. Por exemplo, o caso do perfil de alunos e dos anos de experiência do avaliador (tamanho da amostra, repetir o experimento até que não haja valores nulos em cada célula e que a frequência esperada seja maior ou igual a 5%);
- É interessante também ter em mente um modelo matemático que estabeleça funções que relacionem as diversas variáveis influentes no processo. A sugestão de um modelo matemático, mesmo que aproximado, possibilita um planejamento experimental mais dirigido, definindo-se valores de estudo adequados para as variáveis, reduzindo desta forma o número de ensaios (o tamanho da amostra);
- Escolha adequada da variável de resposta, de modo que se garanta a objetividade na análise dos resultados obtidos;
- Após definir a variável de resposta de interesse, investigar principalmente na literatura quais variáveis podem estar relacionadas com a resposta;
- A avaliação intensiva de diversas variáveis pode ser necessária quando o estudo encontra-se em seus estágios iniciais e não se detém uma experiência anterior, exigindo a avaliação das variáveis em diversos níveis;
- A experimentação é um processo iterativo. Principalmente em processos complexos, com diversas variáveis influentes, não se deve partir de um conjunto extenso de experimentos, que envolva um grande número de variáveis;
- É mais produtivo estabelecer um conjunto inicial e ir aprendendo sobre o processo aos poucos, acrescentar novas variáveis ou excluir variáveis;
- O conhecimento técnico específico, não estatístico, sobre o problema deve ser usado.

Ainda, temos que ter em mente quando realizamos um experimento, as seguintes informações:

- Os dados que coletamos já são uma amostra. Porque a verdadeira população na prática é sempre desconhecida. Não precisamos criar nenhum critério para extrair outro conjunto dos dados do conjunto de dados originalmente obtido para caracterizar um processo de amostragem;
- Não devemos criar critérios para a exclusão de elementos da amostra, ou para "dividir" a amostra de forma simplista. Uma vez que definimos nossa variável de interesse e quais os valores possíveis para essa variável, só excluímos casos que estejam fora do conjunto dos possíveis valores da nossa variável de interesse;
- Mas, devemos refazer a coleta de dados se percebermos características não aleatórias na amostra. Por exemplo, se a variabilidade dos dados da amostra coletada é zero, ou muito próximo de zero, devemos refazer o ensaio, ou redefinir a construção desta variável. Caso contrário, não temos uma variável, temos uma constante.

12.1 Exemplo do teste de habilidade em leitura em crianças da terceira série do primeiro grau

Definição da variável de interesse: o escore em um teste de leitura que varia entre 0 e 1; Uma questão inerente ao problema é que a habilidade em leitura de crianças da terceira série não tem um caráter homogêneo. Ou seja, não é comum que todas as crianças apresentem o mesmo nível de habilidade em leitura.

Então, poderíamos avaliar inicialmente um grupo pequeno de crianças, digamos 15 crianças, e obter a média e o desvio padrão dos escores destes alunos. Por exemplo, digamos que $\bar{X}$= 0.62 e $S = 0.1$. Vamos admitir um erro em torno da média de 2%. Isto quer dizer que $\delta_{\bar{X}} = 0.02 \times 0.62 = 0.0124$. Então pela expressão temos que

$$\delta_{\overline{X}} = 2.1\frac{S}{\sqrt{n}} \rightarrow 0.0124 = 2.1\frac{0.1}{\sqrt{n}} \rightarrow n = \left(2.1 \times \frac{0.1}{0.0124}\right)^2 = 42.38$$

Assim, se considerarmos uma amostra de 42 alunos, estamos controlando o erro em torno da média em aproximadamente 2%. Agora, digamos que $\bar{X}$= 0.62, mas $S = 0.2$, a dispersão dessa amostra de ensaio é maior que a amostra anterior. Vamos manter o erro em torno da média em 2%.

$$n = \left(2.1 \times \frac{S}{\delta}\right)^2 = \left(2.1 \times \frac{0.2}{0.0124}\right)^2 = 169.5$$

Veja como a dispersão interfere no tamanho da amostra. Agora precisaríamos de 170 alunos para realizar um estudo com o erro em torno da média controlado em 2%. Fica fácil notar também que se diminuímos o erro aceitável, ou seja, somos MAIS rigorosos, necessitamos de uma amostra maior.

Mas, vamos nos deter no primeiro caso. Avaliamos a nota (*escore*) de 42 alunos da terceira série desta escola. Se tem mais de uma turma da terceira série, tentamos selecionar quantidades iguais de alunos de cada turma até obtermos 42 alunos (ou mais). Um cuidado que se deve ter é não selecionar alunos só com bons escores ou só com

baixos escores. Faça, por exemplo, um sorteio pelo nome do aluno. Mas, depois analise os resultados e, se foram sorteados alunos só com bons escores, por exemplo, refaça o sorteio novamente com TODOS os alunos da turma.

Após essa etapa verifique se ocorreu por erro, por exemplo, valores de escores menores que zero, ou maiores que 1. Caso isto ocorra aplique o teste ao aluno em questão, se isto não for possível, exclua esse dado.

Após a crítica dos dados coletados, essa é a sua amostra. Pode realizar com essa amostra testes antes e depois de alguma nova técnica de ensino. Pode tentar modelar essa habilidade em leitura com base em outras variáveis, por exemplo QI da criança e possibilidade de dislexia. Mas, se não coletamos dados sobre o QI e dislexia desses alunos não dá para incluir no experimento.O que enfatiza a importância de conhecer bem o problema que queremos investigar e as variáveis que podem interferir neste problema.

12.2 Exemplo Final usando Modelo de regressão beta

Neste exemplo, um experimento com 37 estudantes de um ambiente *online* de aprendizagem gamificado, denominado MeuTutor®, foi realizado. O objetivo foi melhorar as interações desses estudantes com os recursos educacionais disponíveis no ambiente. Para isso, as interações atuais foram armazenadas e os estudantes foram classificados (perfil de interação) considerando a forma como estes interagiam com as diferentes categorias de recursos educacionais (Paiva, 2015). Em seguida, essa informação foi utilizada para a criação de missões personalizadas para os mesmos 37 alunos, focando nos tipos de interação que os estudantes precisavam aprimorar. O ambiente MeuTutor® tem implementado um módulo de geração, entrega e monitoramento de missões, o que ajudou nesta parte do experimento. Para cada missão foi definido o que o estudante deveria fazer, a quantidade de pontos a receber após a conclusão e o tempo para a realização da mesma. Nosso objetivo é verificar se as missões sugeridas afetaram de alguma forma o aluno, para isto decidimos avaliar o desempenho do aluno, que é a proporção de perguntas corretamente resolvidas entre todas as questões resolvidas. Outras variáveis foram também levantadas e apresentamos na Tabela 1.5 uma parte dos dados. As variáveis são: número de amigos (Amigos); quantidade de vídeos assistidos (Vídeo); registro de número de questões resolvidas (Resolvidas); registro de número de questões resolvidas corretamente (Certas).

Tabela 1.5. Desempenho do aluno no ambiente MeuTutor® antes e depois das missões sugeridas

Controle						Caso					
Obs	Desemp.	Amig.	Vídeos	Res.	Certas	Obs	Desemp.	Amig.	Vídeos	Resp.	Certas
01	0.27	0.83	0.00	242.16	65.83	38	1.00	1.00	1.0	165.00	165
02	0.46	0.16	0.00	94.00	43.50	39	0.73	1.00	0.0	121.00	89
03	0.62	0.16	6.00	71.50	44.83	40	0.63	0.00	6.0	131.00	83
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	0.71	1.50	4.50	98.33	70.83	27.50	74	0.74	2.00	160	120

A distribuição beta é comumente usada para modelar variáveis aleatórias que assumem valores no intervalo de (0,1), tais como taxas, porcentagens e proporções como no caso do experimento acima. A densidade beta pode assumir formas diferentes dependendo da combinação de valores de parâmetros (ver Figura 1.11). Seja $y_1, \ldots, y_n$

uma amostra de v.a.'s independentes tal que cada $y_i, i = 1, \ldots, n$, segue uma distribuição beta com densidade:

$$f(y; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y)^{(1-\mu_i)\phi_i - 1}, \quad 0 < y < 1,$$

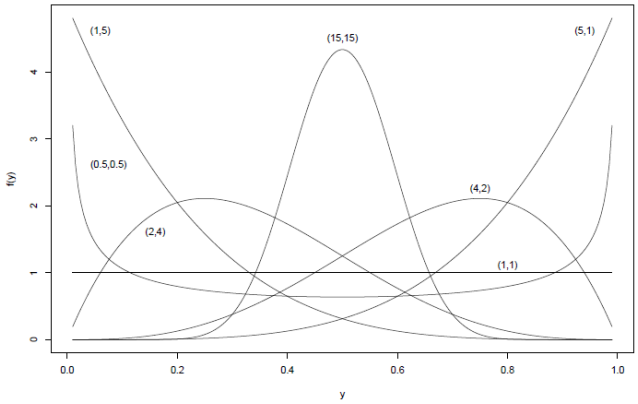


Figura 1.12. Diferentes formas das distribuições beta

Onde $0 < \mu_i < 1$ e $\phi > 0$. Aqui, $E(y_i) = \mu_i$ e $var(y_i) = (\mu_i(1-\mu_i))/(1+\phi_i)$. Note que ϕ pode ser visto como uma parâmetro de precisão, pois quanto maior ϕ menor a variância de y_i, por outro lado, $\phi - 1$ é um parâmetro de dispersão; Ferrari e Cribari-Neto (2004) propõem que a média da variável resposta y_i, ou seja, μ_i possa ser escrita como

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Sabemos que utilizamos a função de ligação para liberar os possíveis valores que os $\beta's$ podem assumir. Neste caso como $\mu_i \in (0,1)$ um função de ligação conduz essa média a todos os Reais é a função de ligação logito, dada por

$$\mu_i \in (0,1) \leftrightarrow \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\} \in (-\infty, +\infty) = IR.$$

Assim como tentamos explicar a média da variável resposta μ_i, que é um parâmetro desconhecido, podemos também tentar modelar sua variância, mas neste caso implica em modelar o parâmetro ϕ, pois a variância depende de μ_i (que já está sendo modelada) e de ϕ_i. De fato, sugerimos um modelo para ϕ quando suspeitamos que a dispersão não é constante para os dados, ou que é possível haver grupos com dispersões diferentes. Assim, Smithson e Verkuilen (2006) propõem um modelo de regressão beta em que

$$h(\phi_i) = \gamma_1 + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \ldots + \gamma_q z_{iq}, \quad i = 1, \ldots, n.$$

Neste caso, como $\phi > 0$ uma função de ligação adequada é $\log\phi \in IR$. Precisamos estimar os $\beta_t's$ e os $\gamma_j's$ para estimarmos ϕ_i e μ_i. Isto é feito utilizando o método de máxima verossimilhança[25]. Pesquisadores podem usar o pacote *betareg* disponível no software estatístico R.

[25] https://www.ime.usp.br/~giapaula/cursospos.htm

Análise de Diagnóstico

Verificar se um determinado modelo é uma representação adequada dos dados é um passo importante da análise estatística. A construção de um modelo de regressão envolve a definição da distribuição da variável de resposta, a escolha da função de ligação, a escolha das covariáveis. Vários fatores podem levar um modelo ajustado pobre: é importante verificar se alguma escolha errada da distribuição da variável resposta foi feita, se há pontos influentes, se houve especificação incorreta da variância, entre outros fatores.

Ou seja, modelos estatísticos estão baseados em certas suposições. A fim de ter confiança na análise devemos verificar se os pressupostos associados são válidos. Isso pode ser alcançado por meio de análise de diagnóstico. Normalmente, esses diagnósticos são construídos em torno de resíduos e critérios de seleção como o R^2.

A maior parte dos resíduos é baseada nas diferenças entre o respostas observadas (y) e a média estimada $\hat{\mu}$. Por exemplo $r_i = y_i - \hat{\mu}_i$, ou seja, o resíduo é uma medida de discrepância entre os dados reais e o modelo ajustado. Aqui vamos utilizar o resíduo proposto por Espinheira et al (2015), $r_{p,i}^{\beta\gamma}$ denominado resíduo combinado e baseado na diferença

$$(y_i^* - \widehat{\mu}_i^*), \quad \text{em que} \quad y_i^* = \log\left\{\frac{y_i}{(1-y_i)}\right\} \quad \text{e} \quad \mu_i^* = \mathrm{E}(y_i^*).$$

Os gráficos de resíduos versus índices das observações ou versus valores preditos ($\hat{\mu}_i$) são os mais básicos. Se um modelo está especificado corretamente, então estes gráficos não devem apresentar nenhuma tendência, os resíduos devem estar aleatoriamente distribuídos em torno do zero. A presença de quaisquer características sistemáticas tipicamente implica uma falha de um ou mais pressupostos do modelo. Outro gráfico de resíduos importante é o gráfico de probabilidade normal com envelope simulado, que pode ser usada mesmo quando as distribuições empíricas dos resíduos não são normais. Se o modelo está adequado aos dados, esperamos que a maioria dos resíduos estejam aleatoriamente distribuídos dentro das bandas do envelope.

Modelando os dados

Nosso objetivo é modelar a proporção de perguntas corretamente resolvidas entre todas as questões resolvidas, aqui chamada de **desempenho**. O desempenho de estudantes assume valores em (0,1), o que justifica o modelo de regressão beta. Para enfatizar nossa decisão, veja a Figura 1.12. Na Figura 1.12 observa-se que a distribuição beta se encaixa melhor aos dados do que a distribuição normal. Na verdade, se utilizarmos a distribuição normal para estimar a média de desempenho, ocorreriam valores maiores do que valores negativos; o que seria impossível para este tipo de variável aleatória. As covariáveis do modelo são: x_2 – exponencial do número de amigos (**Amigos**); x_3 – quantidade de vídeos assistidos (**Vídeos**); x_4 – registro de número de questões resolvidas (**Resolvidas**); e x_5 – um indicador de alunos antes e depois das intervenções (missões sugeridas), de tal forma que x_5 (**Indicador**) assume o valor "0" antes da intervenção ("missões") e "1" depois da intervenção. Consideramos exp(*número de amigos*) = exp(*amigos*), porque a variável original: número de amigos assume valores muito pequenos e, por outro lado, log(*questões resolvidas*) = log(*resolvidos*) é melhor

para o modelo, porque a variável original assume valores muito grandes. No início, consideramos o modelo de regressão beta com dispersão constante definida como

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + \beta_5\, x_{i5} = \beta_1 + \beta_2 \exp(\text{Amigos})_i + \beta_3 \text{Vídeos}_i + \beta_4 \log(\text{Resolvidas})_i + \beta_5 \text{Indicador}_i, \quad i = 1, \ldots, n; \tag{1.6}$$

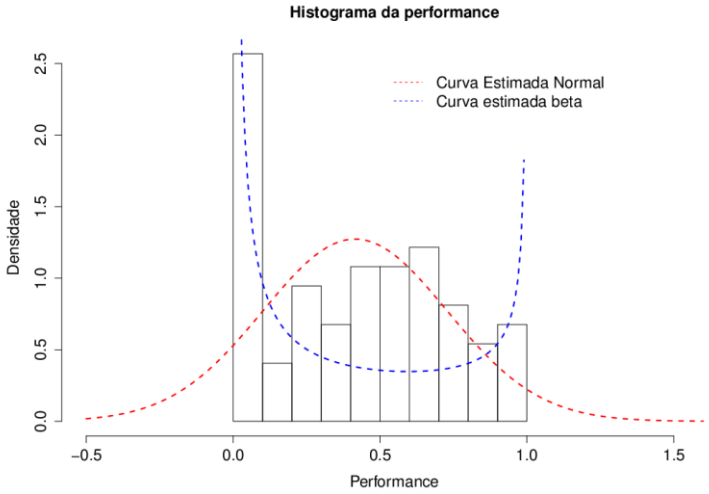


Figura 1.13. Histograma da performance dos estudantes e densidades estimadas considerando a distribuição normal e a distribuição beta

A análise de residual deste modelo é apresentada na Figura 1.13. A Figura 1.13a apresenta o gráfico do resíduo combinado contra os índices das observações do modelo em (1.6). A figura indica que a dispersão dos resíduos não é constante para todas as observações. De fato, há mais dispersão para o grupo caso (depois de missões) do que para o grupo controle. O gráfico de probabilidade normal com envelopes simulados apresenta a mesma característica incomum (Figura 1.13b), a maioria dos pontos estão fora do envelope.

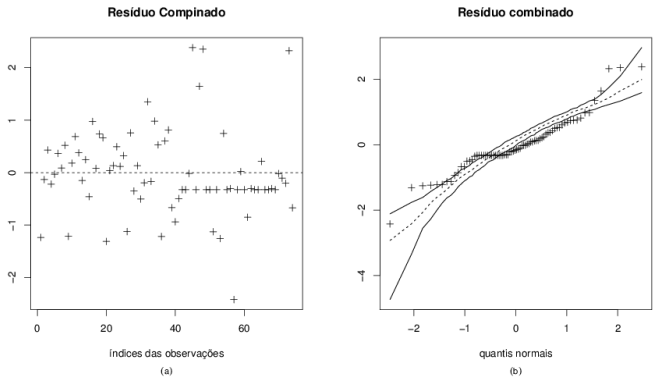


Figura 1.14. Gráfico de Resíduos

De fato, os resíduos estão cruzando o envelope, sendo uma forte evidência de dispersão não constante, e uma estratégia de modelação de ϕ é necessária. Assim, devemos considerar um modelo de regressão beta com dispersão variável, conforme abaixo descrito.

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + \beta_5\, x_{i5} = \\ \beta_1 + \beta_2 \exp(\text{Amigos})_i + \beta_3 \text{Vídeos}_i + \beta_4 \log(\text{Resolvidas})_i + \beta_5\ \text{Indicador}_i, \quad i = 1, \ldots, n \tag{1.7}$$

e

$$h(\phi_i) = \log(\phi_i) = \gamma_1 + \gamma_2\, x_{i2} + \gamma_3\, x_{i4} + \gamma_4\, x_{i5} = \\ \gamma_1 + \gamma_2 \exp(\text{Amigos})_i + \gamma_3 \log(\text{Resolvidas})_i + \gamma_4\ \text{Indicador}_i. \tag{1.8}$$

As estimativas dos parâmetros e os correspondentes p-Valores são apresentados na Tabela 1.6. Note na Tabela 1.6 que todas as covariáveis são significativas ao nível nominal 5%. Na Figura 1.14, apresentam-se os gráficos de resíduos do modelo definido em (1.7) e (1.8).

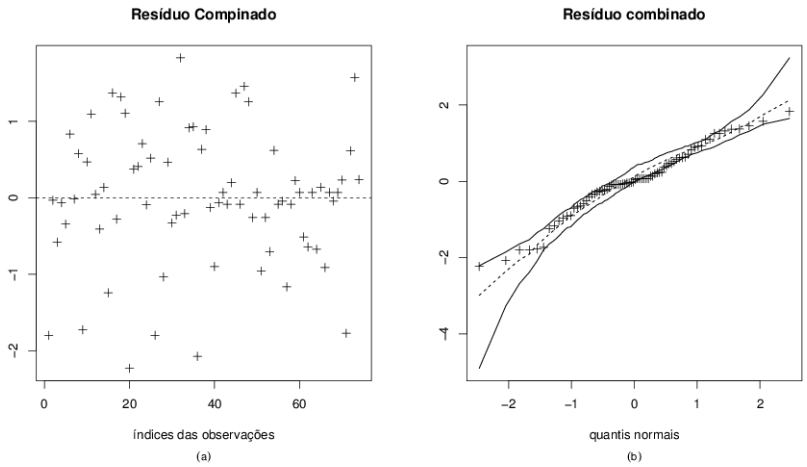


Figura 1.15. Gráfico de Resíduos

O gráfico dos resíduos contra os índices das observações sugere que os resíduos estão espalhados aleatoriamente em torno de zero (sem características sistemáticas); veja a Figura 1.14a. No gráfico de probabilidade normal não existe nenhuma evidência de má especificação, a maioria dos resíduos estão espalhados aleatoriamente dentro das bandas de envelopes; veja a Figura 1.14b. Assim, a análise de resíduos indica que os dados foram ajustados bem pelo modelo de regressão beta.

Tabela 1.6. Estimativas dos parâmetros e p–Valores.

Descrição das variáveis	Parâmetros							
	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	γ_3
	Constante	Amigos	Vídeos	Resolvidas	Indicadora	Constante	Amigos	Indicadora
estimativa	−1.28	−0.06	0.17	0.29	0.95	1.88	0.16	−1.85
p–Valor	0.0000	0.0000	0.0079	0.0000	0.0015	0.0000	0.0000	0.0000

No geral, nota-se que o modelo de regressão beta com dispersão variável é uma ferramenta útil para modelar o desempenho dos alunos (ver as estimativas dos parâmetros Tabela 1.6). Devemos salientar o efeito das covariáveis sobre o desempenho médio. Temos que, se a estimativa de um β associado com determinada variável é negativa, quanto maior o valor da estimativa deste β maior é a queda no desempenho do aluno. Por outro lado, se o β estimado é positivo, mais importante é o aumento no desempenho do estudante quanto maior for esta estimativa. Uma vez que $\hat{\beta}_2 = -0.06$, conclui-se que a quantidade de amigos contribui para a redução do desempenho do aluno. Por outro lado, a quantidade de vídeos assistidos contribui positivamente sobre o desempenho do estudante, $\hat{\beta}_3 = 0.17$. De fato, a cada vídeo adicional que o aluno assiste, a chance de melhorar seu desempenho cresce em 30%. Além disso, cada pergunta adicional que o aluno responde aumenta consideravelmente seu desempenho, $\hat{\beta}_4 = 0.29$. A conclusão mais importante que temos alcançado com base no modelo beta é a forte influência positiva das intervenções (missões) sobre o desempenho dos alunos ($\hat{\beta}_5 = 0.95$)

Outra informação importante que o modelo estimado fornece são os valores estimados para os desempenhos médios. Assim, o modelo pode ser usado para outro conjunto de dados ou por outros investigadores. Para chegar a isso, é necessário apenas fornecer valores para as variáveis Amigos, Resolvidas, Vídeos e Indicadores (0 - nenhum missão dada; 1 - missões dadas) e com base na fórmula abaixo obteremos $\hat{\mu}_i$, que representa a estimativa de desempenho do estudante. Como nós usamos a função de ligação logística, com base em (1.7), temos que

$$\widehat{\mu}_i = \frac{\exp\{-1.3 - 0.06\exp(\text{amigos})_i + 0.17\text{Vídeos}_i + 0.29\log(\text{resolvido})_i + 0.95\text{Indic}_i\}}{1 + \exp\{-1.3 - 0.06\exp(\text{amigos})_i + 0.17\text{Vídeos}_i + 0.29\log(\text{resolvidas})_i + 0.95\text{Indic}_i\}}.$$

13. Conclusões

Neste minicurso podemos mostrar com base em exemplos de pesquisas reais na área de informática em educação como a estatística é uma ferramenta indispensável em qualquer trabalho de cunho científico. Evidenciamos como o delineamento do experimento pode comprometer análises mais aprofundadas dos dados. De fato, exploramos desde aspectos básicos da estatística até métodos como os modelos lineares generalizados e seus similares. Esses métodos permitem uma modelagem ampla dos dados, considerando desde a usual distribuição normal até modelos mais avançados, como o modelo de regressão beta, extremamente útil na modelagem de proporções, índices, escores etc. Deste modo, elucidamos como uma análise estatística mais aprofundada pode revelar informações sofisticadas, decisivas em tomadas de decisões, contribuindo de forma concreta para a comprovação de métodos propostos e que impactam direta e positivamente a área da informática na educação.

Referências

Anderson, T. (2008). "The theory and practice of online learning". Athabasca University Press.

Barbetta, A.P., Reis, M.M., Bornia, A.C. (2008). "Estatística Para Cursos De Engenharia e Informática". 2nd ed. São Paulo: Atlas S.A..

Blanchard, E. G. (2012). "On the WEIRD Nature of ITS/AIED Conferences". In: Intelligent Tutoring Systems (pp. 280-285). Springer Berlin Heidelberg.

Brusilovsky, P. (1998). "Methods and techniques of adaptive hypermedia". In: Adaptive hypertext and hypermedia (pp. 1-43). Springer Netherlands.

Espinheira, P.L., Silva, C.M. Silva, A.O. (2015). "Prediction Measures in Beta Regression Models". arXiv:1501.04830.

Easterbrook, S., Singer, J., Storey, M. A., Damian, D. (2008). "Selecting empirical methods for software engineering research". In: Guide to advanced empirical software engineering (pp. 285-311). Springer London.

Ferrari, S.L.P.; Cribari-Neto, F. (2004) "Beta regression for modelling rates and proportions", Journal of Applied Statistics, 31, 799-815

Lehmann, E. L., Casella, E. (1998) "Theory of Point Estimation", 2nd ed. New York: Springer-Verlag,

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). "Intelligent tutoring systems and learning outcomes: A meta-analysis". Journal of Educational Psychology, 901–918

Magalhães, C. V., Santos, R. E., da Silva, F. Q., & Gomes, A. S. (2013). "Caracterizando a Pesquisa em Informática na Educação no Brasil: Um Mapeamento Sistemático das Publicações do SBIE". In: Anais do Simpósio Brasileiro de Informática na Educação (Vol. 24, No. 1).

Martin, F. G. (2012). "Will massive open online courses change how we teach?". Communications of the ACM, 55(8), 26-28.

Paiva, R. O. (2015) Araújo et al. "Improving Pedagogical Recommendations by Classifying Students According to their Interactional Behavior in a Gamified Learning Environment". Proceedings of the ACM Symposium On Applied Computing (inpress).

Sclater, N. (2008). "Web 2.0, personal learning environments, and the future of learning management systems". Research Bulletin, 13(13), 1-13.

Smithson, M., & Verkuilen, J. (2006). "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables". Psychological methods, 11(1), 54.

Stahl, G., Koschmann, T., & Suthers, D. (2006). "Computer-supported collaborative learning: An historical perspective". Cambridge handbook of the learning sciences, 2006, 409-426.

Tenório, T. (2015) "Um Modelo de Avaliação por Pares Gamificado para Avaliações Escritas em Ambientes Educacionais Online". Dissertação de Mestrado defendida no Instituto de Computação da Universidade Federal de Alagoas.

Thas, O. (2010). "Comparing Distribuitions", 1nd ed. New York: Springer.