UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE CIÊNCIAS DA COMPUTAÇÃO
DISCIPLINA: Probabilidade e Estatística - INE5405
PROFESSOR: Pedro Alberto Barbetta

ANÁLISE ESTATÍSTICA DO VESTIBULAR DA UFSC 2010

Lucas Pereira, William Rodrigues

Florianópolis
Agosto de 2010

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nosso objetivo neste trabalho é aplicar os conhecimentos adquiridos na disciplina de Probabilidade e Estatística. Para tal fomos incumbidos de realizar a análise estatística de um arquivos de dados. A escolha dos dados poderia ser feita de duas formas: obter uma tabela de dados que já estive pronta e fosse relativa a um tema qualquer ou realizar uma pesquisa e dessa forma adquirir os dados. Escolhemos a primeira opção pela praticidade e pelo fato de que existem metodologias para a realização de pesquisas, metodologias essas que não temos conhecimento e não dispúnhamos de muito tempo para adquirir.

A partir dessa escolha nosso objetivo foi escolher um tema e buscar por uma tabela de dados relativa a esse tema. Pensamos inicialmente em trabalhar com o IDH (Índice de Desenvolvimento Humano) coletando os dados através do relatório do PNUD (Programa das Nações Unidas para o Desenvolvimento), porém logo no início da coleta decidimos mudar para um tema mais próximo da universidade e escolhemos então trabalhar com os dados estatísticos do Vestibular da UFSC 2010. O vestibular da UFSC consiste em uma prova realizada anualmente cujo objetivo é selecionar os melhores classificados para a universidade, uma vez que a demanda de vagas não é suficiente para todos. Os vestibulares em geral são na verdade um método paliativo para o preenchimento das vagas, já que o correto seria existirem vagas á todos

Adquirimos os dados através do site do Vestibular UFSC 2010 e coletamos 77 amostras, cada uma relativa a um curso oferecido pela universidade. Decidimos restringir nossas amostras apenas aos cursos oferecidos no campus Florianópolis, eliminando assim os cursos Engenharia de Energia, Tecnologia da Informação e Comunicação (campus Araranguá), Ciências Rurais (campus Curitibanos) e Engenharia de Mobilidade (campus Joinville).

2 COLETA E LEVANTAMENTO DOS DADOS

Foram coletadas 77 observações referentes aos cursos oferecidos no vestibular da UFSC 2010. A partir disso escolhemos trabalhar com as variáveis (quantitativas): candidatos por vaga, nota do primeiro classificado e nota do último classificado. Para a variável qualitativa escolhemos área do conhecimento, tendo como categorias: área de exatas, área de biológicas e área de humanas. Inicialmente a ideia foi utilizar o centro de cada curso como variável qualitativa, porém devido ao fato de existir 11 centros decidimos generalizar distribuindo assim por grandes áreas do conhecimento.

Para a manipulação dos dados utilizamos o software Open Office Fórmula.

2.1 Candidatos por Vaga

Essa variável é composta por duas outras informações: número de inscritos em determinado curso e número de vagas desse mesmo curso. O número de candidatos por vaga é obtido a partir da divisão do número de inscritos pelo número de vagas. Apesar de constar em nossa tabela de dados as informações de número de vagas e inscritos, não consideramos estas duas variáveis devido ao fato de termos já pronto no relatório do vestibular da UFSC a variável candidatos por vaga. Sendo assim, os números de vagas e inscritos apenas constam na base de dados como ilustração de como é calculado o número de candidatos por vaga.

2.2 Nota do Primeiro Classificado

Essa é uma variável continua que descreve a pontuação do primeiro classificado em determinado curso. Com ela poderemos verificar se existem grandes diferenças de notas em diferentes cursos e ainda ver se esta realmente tem relação com o número de candidatos por vaga. No vestibular da UFSC o intervalo de representação das notas vai de 0 a 100.

2.3 Nota do Último Classificado

Analogamente a nota do primeiro classificado podemos descrever a variável nota do último classificado.

2.4 Área do Conhecimento

Escolhemos área do conhecimento como variável qualitativa pois ela faz uma boa distinção entre as principais características de cada curso e dessa forma podemos fazer uma análise com mais qualidade e precisão. Assim ela foi dividida nas três categorias mães da área do conhecimento: exatas, biológicas e humanas. A distribuição das frequências de cada categoria é descrita abaixo.

Biológicas: 13 ocorrências.

Exatas: 24 ocorrências.

Humanas: 40 ocorrências.

2.5 Apresentação da Matriz de Dados

Tabela 1: Apresentação da matriz de dados (parcial).

Curso	Vagas	Inscritos	Candidatos por Vaga	Nota Primeiro Classificado	Nota do Último Classificado	Área do Conhecimento
Administração - Diurno	100	545	5,45	72,64	57,36	Humanas
Administração - Noturno	100	688	6,88	74,43	56,7	Humanas
Agronomia	110	379	3,45	74,55	45,93	Biológicas
Antropologia	25	53	2,12	70,34	42,03	Humanas
Arquitetura e Urbanismo	80	1182	14,78	87,45	68,15	Humanas
Arquivologia	60	23	0,38	54,08	37,45	Humanas
Artes Cênicas	30	158	5,27	71,35	45,87	Humanas
Biblioteconomia	80	148	1,85	59,17	39,05	Humanas
Ciência e Tecnologia Agroalimentar	70	106	1,51	67,52	39,16	Biológicas
Ciências Biológicas	60	586	7,33	77,99	61,75	Biológicas

3 APRESENTAÇÃO DE GRÁFICOS E DADOS ESTATÍSTICOS

Para o levantamento de gráficos e dados estatísticos utilizamos a linguagem R juntamente com o pacote Rcmdr.

3.1 Gráficos de Distribuição de Frequências

Trabalhamos nessa parte com dois tipos te gráficos. O gráfico em forma de histograma para as variáveis quantitativas e gráfico de barras para as variáveis qualitativas. No eixo x do gráfico é mostrada a variável em análise, dividida em classes se necessário e no eixo y se encontram as frequências. No caso da variável qualitativa o eixo x é subdividido por categorias, onde cada barra representa uma dessas categorias.

3.1.1 Gráfico de distribuição de frequências - Candidatos por vaga

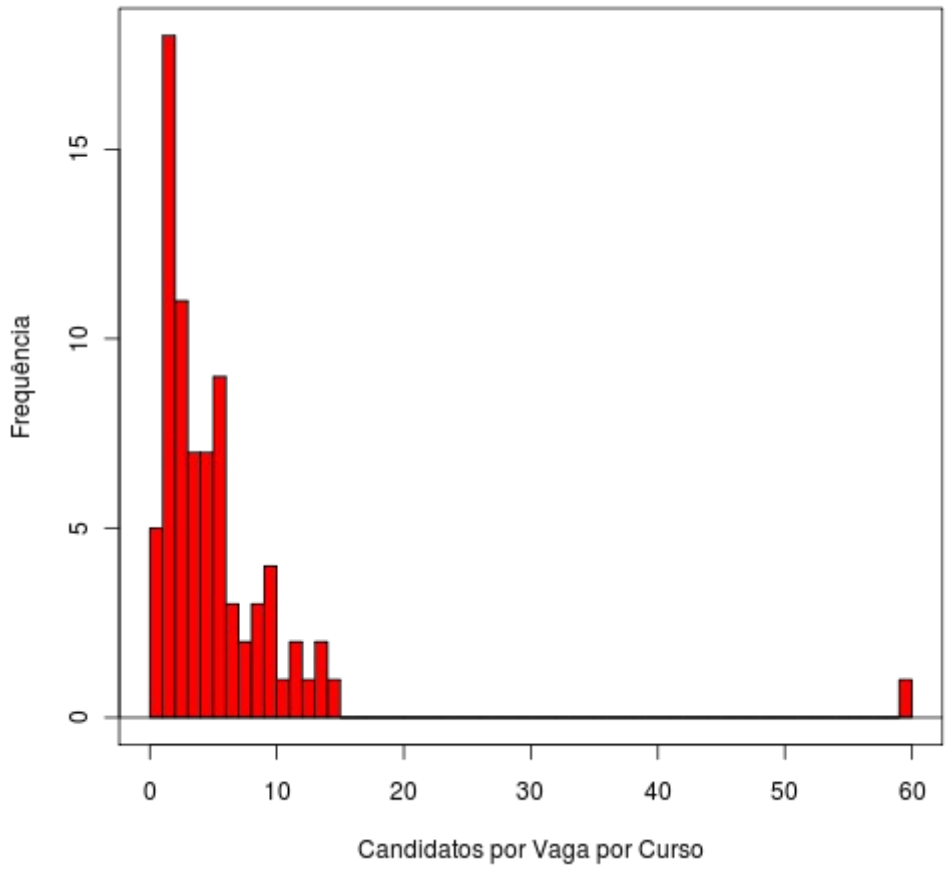

Figura 1: Gráfico de distribuição de frequências - Candidatos por vaga

3.1.2 Gráfico de distribuição de frequências - Nota do primeiro classificado

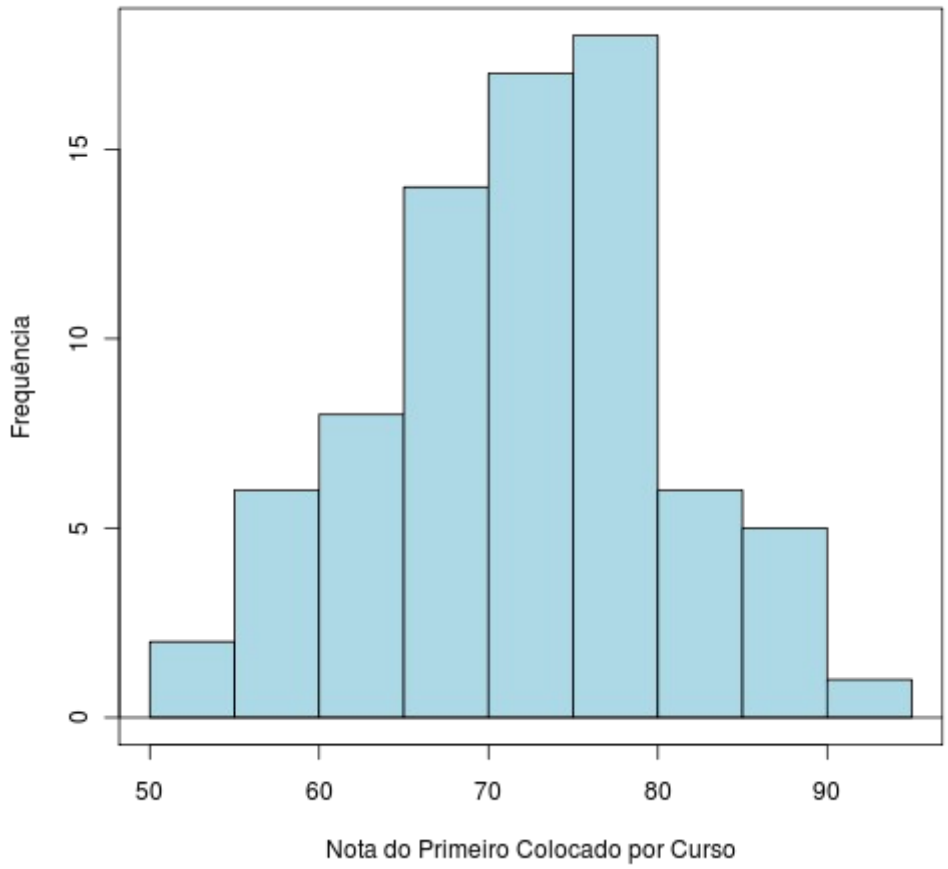

Figura 2: Gráfico de distribuição de frequências - Nota do primeiro classificado

3.1.3 Gráfico de distribuição de frequências - Nota do último classificado

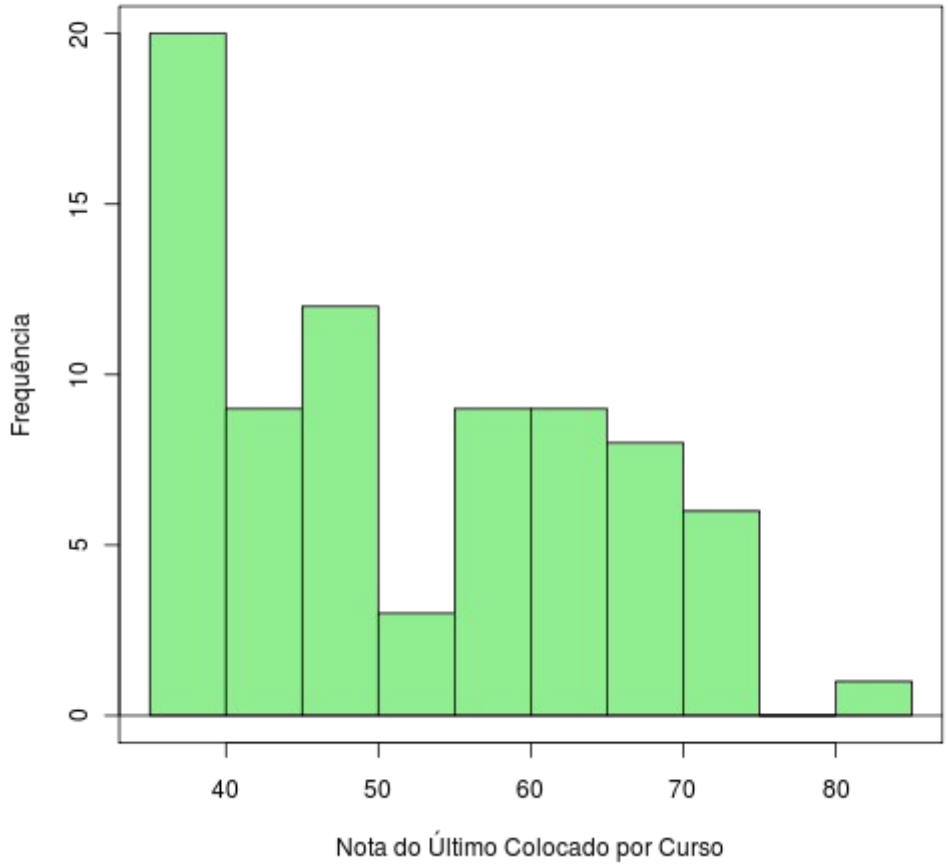

Figura 3: Gráfico de distribuição de frequências - Nota do último classificado

3.1.4 Gráfico em Barras - Área do Conhecimento

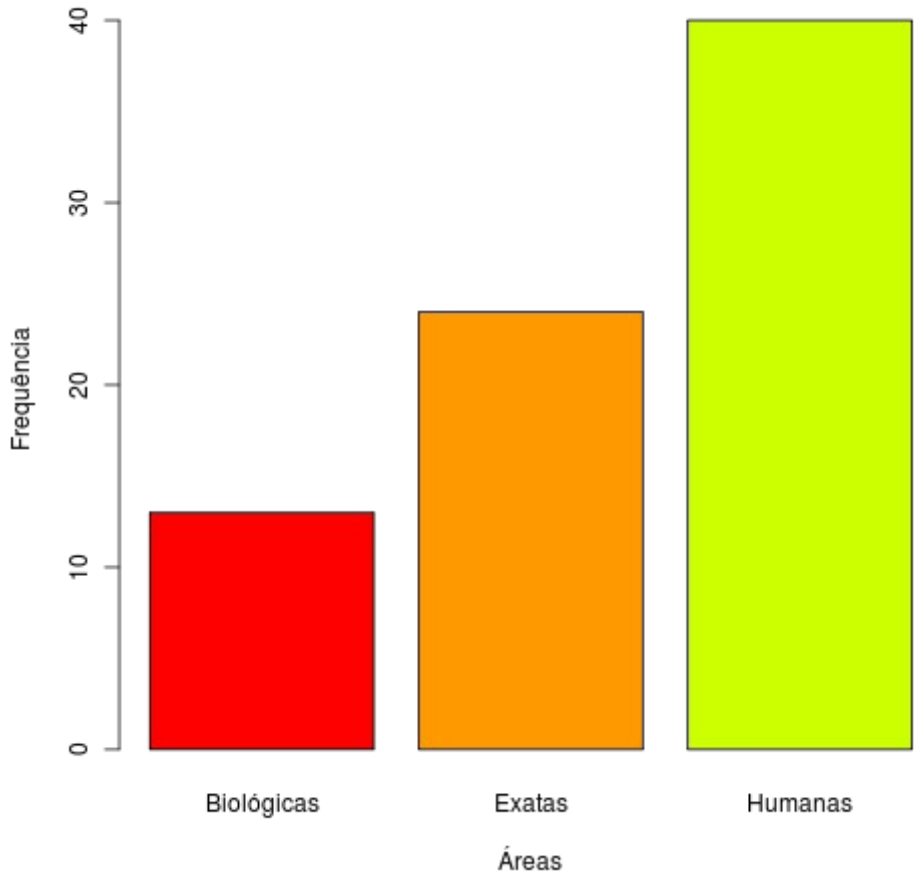

Figura 4: Gráfico em barras - Área do conhecimento

3.2 Medidas Descritivas

Nessa etapa calculamos as medidas descritivas referentes a cada variável. Fizemos tanto o cálculo geral quanto o cálculo com as variáveis separadas por categorias. Desta forma podemos ter uma melhor interpretação dos dados. Foram calculados: média, desvio padrão, quartil inferior, mediana e quartil superior.

3.2.1 Medidas descritivas de candidatos por vaga

Tabela 2: Medidas descritivas de candidatos por vaga

	Médias	Desvio Padrão	Quartil Inferior	Mediana	Quartil Superior
Geral	5,37	7,19	1,95	3,45	6,22
Biológicas	8,91	15,52	2,66	4,47	7,33
Exatas	4,84	3,69	1,86	4,76	5,98
Humanas	4,54	3,67	1,96	3,1	6,17

3.2.2 Medidas descritivas da nota do primeiro classificado

Tabela 3: Medidas descritivas da nota do primeiro classificado

	Médias	Desvio Padrão	Quartil Inferior	Mediana	Quartil Superior
Geral	72,30	8,69	65,94	72,36	78,38
Biológicas	71,77	7,90	66,19	72,06	74,90
Exatas	74,99	7,71	69,48	77,15	79,48
Humanas	70,86	9,29	64,15	71,55	76,52

3.2.3 Medidas descritivas da nota do último classificado

Tabela 4: Medidas descritivas da nota do último classificado

	Médias	Desvio Padrão	Quartil Inferior	Mediana	Quartil Superior
Geral	52,02	12,24	39,91	49,13	61,75
Biológicas	50,76	12,97	39,90	48,25	56,19
Exatas	57,23	13,11	41,51	61,11	67,52
Humanas	49,3	10,67	39,68	47,04	57,32

3.3 Diagrama em Caixas

Fizemos o diagrama em caixas distribuídos por categoria para cada variável quantitativa. No eixo y se encontram os valores que as variáveis podem assumir e no eixo x estão as categorias.

3.3.1 Diagrama em caixa de candidatos por vaga

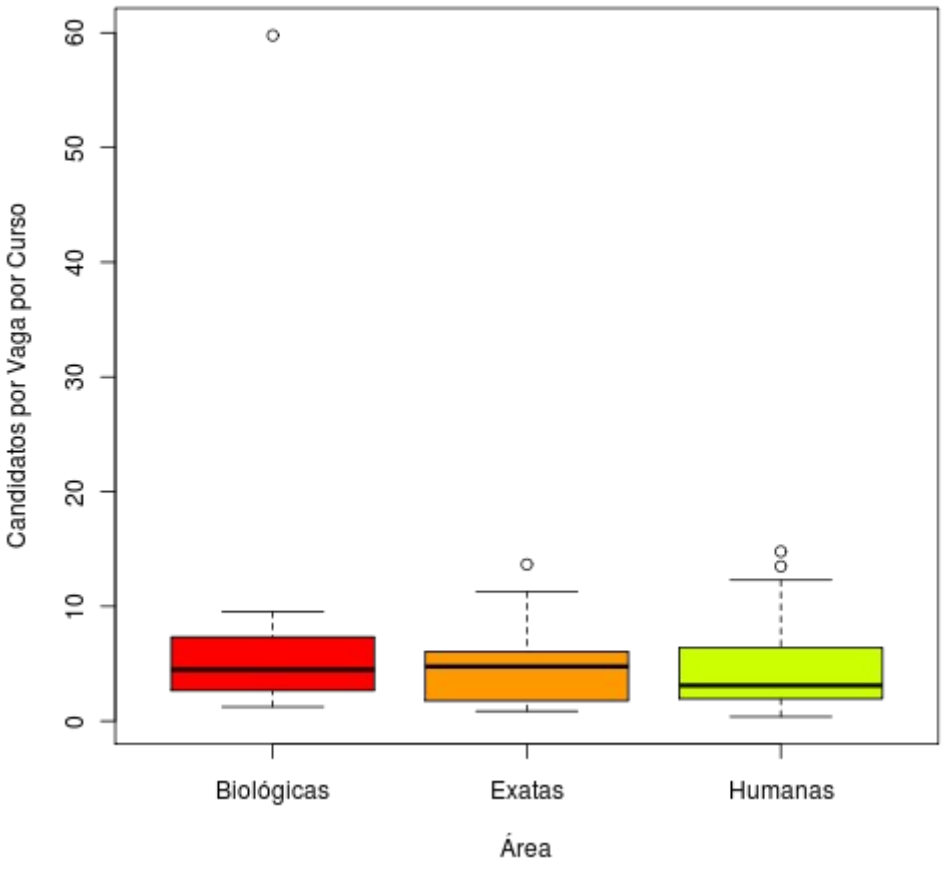

Figura 5: Diagrama em caixa de candidatos por vaga

3.3.2 Diagrama em caixa da nota do primeiro classificado

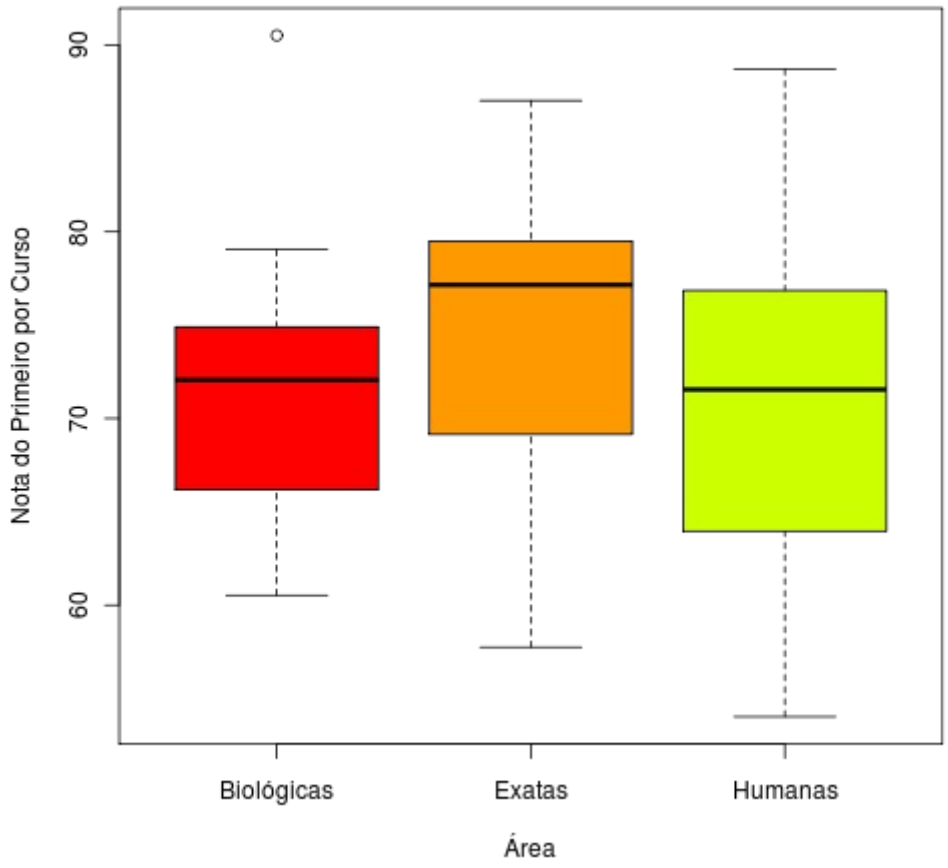

Figura 6: Diagrama em caixa da nota do primeiro classificado

3.3.3 Diagrama em caixa da nota do último classificado

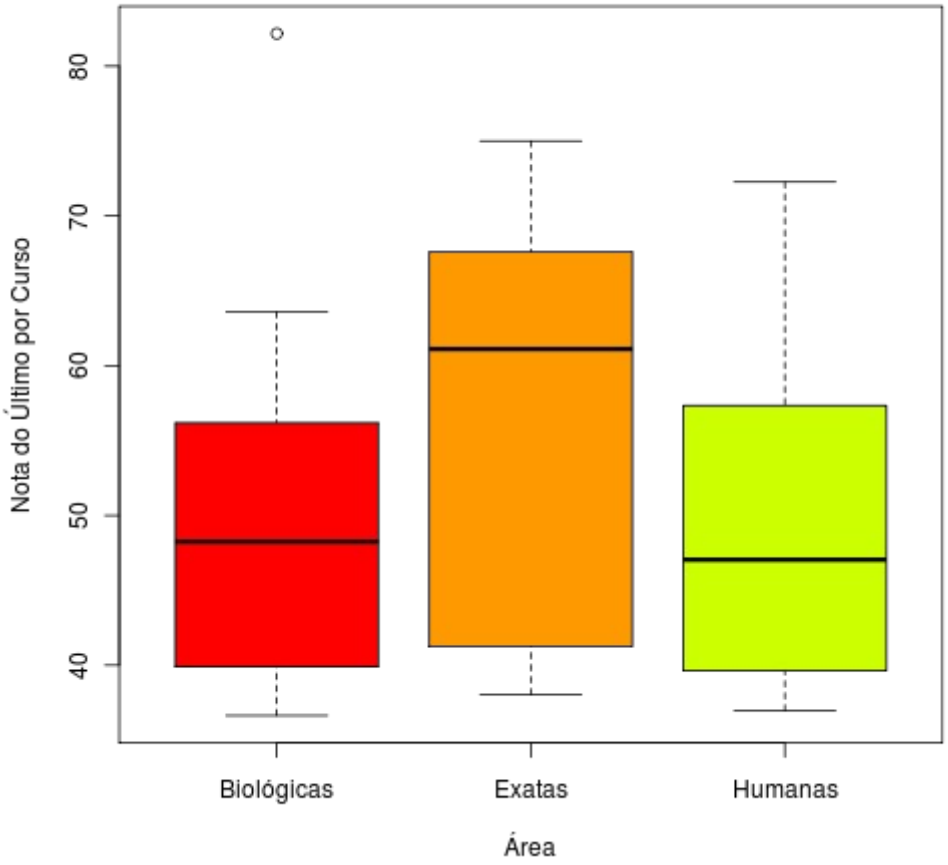

Figura 7: Diagrama em caixa da nota do último classificado

3.4 Diagrama de Dispersão

Decidimos trabalhar com dois diagramas de dispersão: candidatos por vaga juntamente com notas do primeiro classificado e candidatos por vaga com nota do último classificado. Além disso realizamos o cálculo da correlação destas variáveis.

Tabela 5: Correlação entre as variáveis

	Correlação
Candidatos por Vaga e Nota Primeiro Classificado	0,54
Candidatos por Vaga e Nota Último Classificado	0,63

3.4.1 Diagrama de dispersão de candidatos por vaga e nota do primeiro classificado

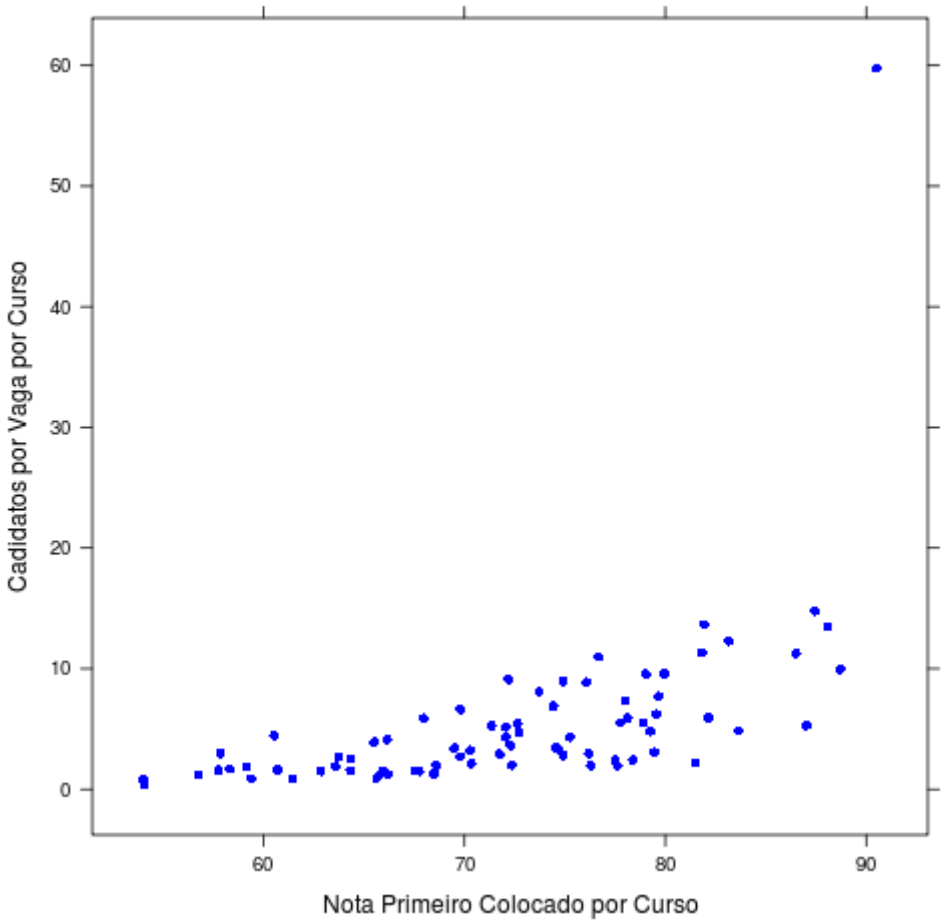

Figura 8: Diagrama de dispersão de candidatos por vaga e nota primeiro classificado

3.4.2 Diagrama de dispersão de candidatos por vaga e nota do último classificado

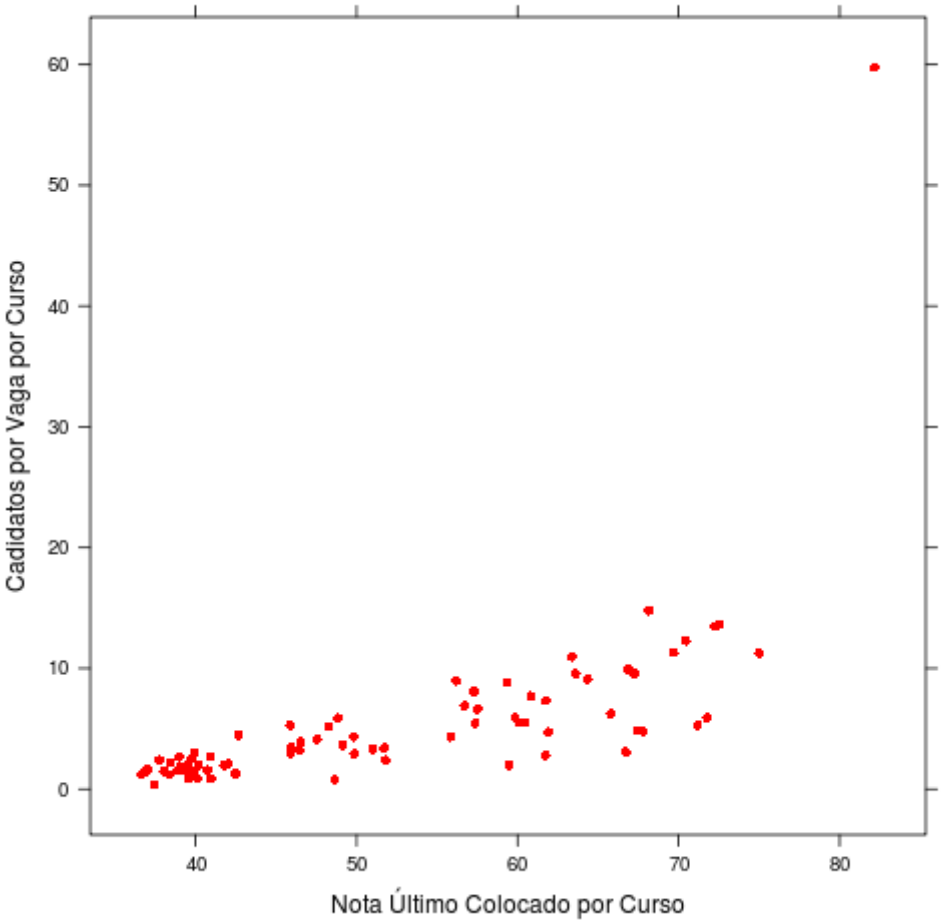

Figura 9: Diagrama de dispersão de candidatos por vaga e nota último classificado

4 INTERPRETAÇÃO DOS RESULTADOS

Iniciamos a interpretação com os histogramas para a partir dele fazer comparações com os diagramas e medidas descritivas coletadas. Inicialmente a primeira observação percebida é a grande discrepância no gráfico de candidatos por vaga (figura 1). Já percebemos nesse momento o quão difícil é realizar relatórios estatísticos pelo fato de poder haver inconsistências durante a coleta dos dados ou manipulação dos mesmo. Isso porque ao analisar aquela discrepância logo imaginamos se tratar do curso de medicina, o mais concorrido em toda universidade. Mas a observação importante a ser feita nesse ponto é que apenas conseguimos identificar essa discrepância e saber o porque dela existir devido ao conhecimento prévio que tínhamos sobre o vestibular, porém, e se não soubéssemos nada sobre esse ou qualquer outro vestibular? Tentamos imaginar que isso pudesse acontecer com qualquer outro tipo de pesquisa, o que nos levou a pensar na dificuldade em se analisar essas discrepâncias uma vez que provavelmente a primeira coisa imaginada é que o dado foi coletado erroneamente. Imaginamos também se a amostra fosse consideravelmente maior, então a dificuldade em identificar essa discrepância seria igualmente maior.

Ainda na figura 1 vemos uma certa assimetria. Assimetria esta que fica confirmada ao analisar as medidas descritivas na tabela 2. Em todas as analises descritivas (geral, biológicas, exatas e humanas) a média é maior que a mediana. Isso se intensifica na área de biológicas, mostrando existir uma longa cauda voltada para a direita na variável candidatos por vaga. Nota-se também na área de biológicas um desvio padrão um tanto quanto alto, diferente das duas outras áreas. Fato que pode ser justificado pela discrepância existente. Pode-se dizer ainda que na área de exatas praticamente não existe assimetria, pois apesar da mediana ser ligeiramente menor que a média, essa é uma diferença desprezível. Por fim em humanas percebemos uma assimetria, não tão grande como ocorre na área de biológicas, mas considerável. Percebemos que no computo geral também existe assimetria o que já era esperado uma vez que essa assimetria existe fortemente em duas áreas. Porém, o importante a se notar aqui é que a separação das medidas descritivas por área torna a compreensão muito mais desmistificada dando uma

melhor interpretação aos dados.

Analisando o diagrama em caixa (figura 5) percebemos a confirmação do que foi dito anteriormente em respeito a assimetria e a discrepância da variável candidatos por vaga. Notamos também as discrepâncias existentes na área de humanas e uma discrepância não tão grande em exatas. É interessante notar que se eliminássemos as discrepâncias da amostra de dados teríamos uma situação bem diferente.

Passando para o gráfico da figura 2 que mostra a nota do primeiro classificado por frequências, podemos notar inicialmente uma boa simetria. Ao analisar as medidas descritivas (tabela 3) confirmamos justamente isso. Ao calcular a média e a mediana geral percebesse mais claramente essa simetria, tendo as duas valores muito próximos uma da outra. Pelas medidas descritivas a categoria com uma menor simetria (embora exista) é a categoria de exatas. O desvio padrão para todas as categorias é bem similar e não muito alto.

Analisando o diagrama em caixas da figura 6 percebemos uma discrepância na área de biológicas. O motivo dela pode ser explicado pela discrepância na variável candidatos por vaga, uma vez que como veremos mais a frente a tendência é que quanto maior o número de candidatos por vaga tão maior será a nota do primeiro/último classificado.

A figura 3 demonstra uma certa assimetria em seu gráfico tendo a cauda voltada para a direita. O que percebesse ai é que conforme a nota do último colocado vai aumentando o número de cursos onde a nota em questão ocorreu diminui ou seja são inversamente proporcionais.

Considerando a tabela 4 verificamos que essa assimetria realmente existe e que a cauda está virada para a direita, porém percebemos que o oposto ocorre com a categoria de exatas onde a cuada fica virada para a esquerda. Analogamente ao que acontece com a nota do primeiro candidato percebemos que na área de exatas essas duas notas são em geral mais altas. É notado também que o oposto ocorre com a área de humanas onde as notas são razoavelmente menores.

Já o desvio padrão da variável nota do último classificado é consideravelmente grande principalmente se comparado com a variável nota do primeiro classificado. Outra diferença entre ambas é a dispersão. Ao ver os diagramas em caixa dessas duas variáveis (figura 6 e 7) notamos uma maior amplitude inter-quartis na variável nota do último classificado. Isso mostra portanto

uma maior dispersão nessa variável comparada com a nota do primeiro classificado. É também na figura 7 que notamos a discrepância existente na categoria biológicas.

4.1 Correlação

Queríamos descobrir se a variável candidatos por vaga realmente influencia nas notas do primeiro e último classificados de cada curso. Para isso usamos o diagrama de dispersão para duas variáveis e chegamos a reposta que já esperávamos: sim, o número de candidatos por vaga de cada curso está diretamente relacionado com a nota dos seus classificados. Apenas na analise dos gráficos das figuras 8 e 9 notamos que essa afirmação e condizente, porém também fizemos o cálculo e conforme mostrado na tabela 5 a correlação é um tanto quanto alta e positiva tendo para as variáveis nota do primeiro classificado e nota do último classificado os valores 0,54 e 0,63 respectivamente.

CONCLUSÃO

O principal ponto que notamos ao decorrer da interpretação dos resultados nesse trabalho foi a importância de analisar esses dados de diversas formas diferentes para obter um melhor resultado. Notamos que ao analisar um gráfico individualmente não temos subsídios suficientes para ter uma analise adequada podendo até mesmo cometer erros graves. Além disso interpretar os dados de forma geral e separados por categoria é de enorme ajuda uma vez que é nesse momento onde conseguimos identificar os pontos particulares de cada uma dessas categorias.

Um fator que nos deu muito trabalho para compreender melhor os dados e principalmente na observação dos gráficos foram as discrepâncias pois elas confundem a amostra e se a analise não for bem feita podem deixar a interpretação com um sentido que não é o real.

Para a observação dos gráficos o que mais deu trabalho foram os diagramas em caixas que se mostraram um tanto quanto complicados de se entender. Essa dificuldade na interpretação se agravou ainda mais com as discrepâncias, porém a partir do momento em que passamos a entender melhor como este tipo de diagrama funciona pudemos ter uma visão melhor de todo o projeto.

Vimos assim que trabalhar com estatística é algo bastante complicado em todas as etapas do processo. Um dos maiores problemas é que pequenos erros botam em risco todo o trabalho da analise. Erros que vão desde a coleta de dados até a manipulação e interpretação destes. Uma virgula em uma casa decimal diferente pode tornar o resultado da amostra completamente distorcido, assim como, uma interpretação mal feita de um gráfico também. Desta forma ao se trabalhar com dados estatísticos o melhor a se fazer é ter o máximo de cuidado possível e dividir o processo de interpretação em etapas para que a analise geral como um todo fique mais clara.

REFERÊNCIAS

UNIVERSIDADE FEDERAL DE SANTA CATARINA. **Relatório Oficial do Vestibular da UFSC 2010**. Disponível em: <http://www.vestibular2010.ufsc.br/relatorio/>. Acesso em: 26 de agosto de 2010.

WIKIPÉDIA. **Anexo: Lista de Cursos Superiores**. Disponível em: <http://pt.wikipedia.org/wiki/Anexo:Lista_de_cursos_superiores>. Acesso em: 28 de agosto de 2010.