UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

Construção de Material de Divulgação de Probabilidade e Estatística

Coordenadora: **Prof. Dra. Laura Leticia Ramos Rifo**

Alunas: **Jordana Aparecida Oliveira**
Renata Lussier Spagnol

Fevereiro, 2008

Prefácio

O presente trabalho é fruto de uma inquietação pessoal perante a escassez de material de divulgação nas áreas de probabilidade e estatística dirigido a alunos de ensino médio ou início de graduação.

Esta apostila consiste basicamente na apresentação e discussão de problemas em planejamento estatístico, cálculo de probabilidades e inferência estatística, bem conhecidos na literatura, que foram escolhidos devido ao seu potencial para serem utilizados em divulgação.

A seleção dos tópicos foi feita ao longo de seis meses de reuniões semanais com as alunas Jordana e Renata, do Bacharelado em Estatística da Unicamp. A pesquisa bibliográfica foi bastante extensa, indo desde referências clássicas, como o livro de William Feller em probabilidade e Freedman et al. em estatística, a periódicos específicos da área, tais como a Revista do Professor de Matemática, American Statistician, Mathematical Monthly, entre outras.

Mesmo tentando ser o mais auto-contido possível, é suposto que o leitor esteja familiarizado com alguns conceitos básicos da teoria estatística e tenha conhecimento das regras básicas do cálculo de probabilidade. Para o leitor interessado, a lista de referências é bastante extensa e contempla diversos níveis de leitura.

É importante ressaltar que os exemplos aqui citados não entregam respostas definitivas, mas sim abrem as portas para uma discussão profunda e necessária sobre as diversas metodologias existentes em análise estatística e sua adequação aos problemas reais aos quais são aplicadas.

Parte deste material foi utilizado no Projeto MEC/Pró Reitoria de Pesquisa da Unicamp, para a construção de material multimídia em matemática para o Ensino Médio. Alguns dos problemas apresentados foram transformados em filmes de 10 minutos, atividades de softwares e experimentos para a sala de aula. Este material está disponível ao público no portal do MEC.

As autoras gostariam de agradecer o apoio oferecido pelo Serviço de Atendimento ao Estudante, SAE, da Unicamp, que disponibilizou bolsas às alunas, sem as quais este trabalho não teria sido viável.

Sumário

1 Planejamento de Experimentos **3**

1.1 Experimentos controlados 3

1.1.1 Vacina Salk 4

1.1.2 Desvio da veia cava 8

1.2 Estudos Observacionais 9

1.2.1 Pílula anticoncepcional e trombo-embolia 10

1.3 Amostragem 11

1.3.1 Eleição presidencial de 1946, EUA 12

2 Cálculo de Probabilidades **15**

2.1 Jogo dos dados Jankenpon 15

2.2 Jogo das três cartas e Jogo dos dois gatinhos 17

2.3 Paradoxo do tempo de espera 19

2.4 Jogo do anel 22

2.5 Jogo das amebas 24

2.6 Problema do campeão e do iniciante 25

2.7 Jogo do máximo 27

2.7.1 Máximo em 2 27

2.7.2 Máximo em 3 28

2.7.3 Minimax em 3 28

2.8 Jogo da soma em três dados 28

2.9 Jogo do tiro ao alvo 29

2.10 Problema das Três Portas 31
2.11 As duas moedas 32
2.12 Problemas dos dois rádios 35
2.13 Edwiges no mercado 36

3 Inferência Estatística **37**
3.1 Paradoxo de Simpson 37
3.2 Associação não é causalidade 40
3.2.1 Estudos observacionais 40
3.2.2 Pólio e refrigerante 41
3.2.3 Quanto mais cegonhas, mais bebês? 42
3.3 Correlação Ecológica 43
3.4 Falácia da regressão 45
3.4.1 Relação entre altura de pais e filhos 45
3.4.2 Teste e reteste em crianças de uma pré-escola 46
3.5 Um problema de decisão 47

Lista de Figuras

2.1 Os dados jankenpon. 16

2.2 Resultados possíveis entre os dados A e B. 16

2.3 Resultados possíveis entre os dados B e C. 16

2.4 Resultados possíveis entre os dados C e A. 17

2.5 Ilustração da primeira carta: frente e verso. 18

2.6 Ilustração dos gatos charmosos: um macho e uma fêmea 18

2.7 Gráfico de Frequências para o último lançamento. 21

2.8 Ilustração do Jogo do anel. Neste exemplo, o arco com o Pólo Norte perde. 23

2.9 A região destacada indica o evento em que a casa ganha. 23

2.10 Uma família de amebas. 24

2.11 Tiro ao alvo. 30

2.12 Figura de um quadrado com o disco. 30

2.13 Ilustração das portas 32

3.1 O gráfico do lado esquerdo representa os salários e nível educacional por indivíduo em três regiões geográficas, denominadas A, B e C. Cada indivíduo é marcado pela letra mostrando a região de sua residência. O gráfico do lado direito mostra os valores médios destas variáveis para cada região: a correlação entre as médias é quase 1. . 44

3.2 Gráfico de alturas de pais e filhos, obtidas no estudo de Sir Galton, em 1885. Extraído de Freedman et al. [7]. 45

3.3 Valores da função de utilidade de Edson. 48

Lista de Tabelas

1.1 Resultados do experimento da vacina Salk de 1954: tamanho dos grupos e taxas de casos de pólio por 100 000 em cada grupo (números arredondados). Fonte: T. Francis Jr. (1955) Am. J. of Public Health. . . . 7

1.2 Resultados dos tipos de planejamentos. 9

1.3 Previsões e Resultados 13

2.1 Tabela dos casos possíveis em três lançamentos e probabilidades condicionais. 33

3.1 Dados da admissão para os programas de graduação nas seis principais carreiras na Universidade da Califórnia Berkeley, outono, 1973. Fonte: Divisão de Graduação, Universidade da Califórnia, Berkeley. 38

3.2 Alunos aceitos por carreira, para as seis carreiras mais concorridas. 40

Capítulo 1

Planejamento de Experimentos

Neste primeiro capítulo, abordaremos exemplos de estudos com planejamento estatístico pouco cuidadoso e os diversos problemas que podem decorrer daí. O exemplo da vacina Salk mostra a importância de um estudo bem planejado.

A principal referência utilizada neste capítulo é Freedman et al.[7]. Outras referências clássicas no assunto são Box et al.[3], Snedecor and Cochran [20], e Yule and Kendall [21].

1.1 Experimentos controlados

A palavra controle é usada em estatística com dois sentidos:

- Um controle é um indivíduo que não recebe o tratamento, como nos exemplos anteriores.

- Dizemos que um experimento é controlado com aleatorização se os indivíduos participantes do estudo forem separados em dois grupos aleatoriamente: um grupo que receberá o tratamento a ser testado e um grupo de controle. O grupo de controle deve ser o mais semelhante possível ao grupo que receberá o tratamento, de modo que a única diferença seja o de não receber o tratamento. Deste modo, em estudos com muitos indivíduos, as diferenças observadas na resposta destes dois grupos são supostamente devidas ao efeito do tratamento. Usualmente, o grupo de controle recebe um tratamento padrão, se houver, ou placebo, se não.

1.1.1 Vacina Salk

Um novo medicamento pretende ser introduzido no mercado. Como deveria ser planejado um experimento para testar sua efetividade?

O método básico é a comparação. Um conjunto de pacientes-alvo é dividido em dois grupos: um chamado **grupo de tratamento**, e outro chamado **grupo de controle**. O grupo de tratamento recebe o medicamento novo a ser testado e o de controle recebe o tratamento já existente, se houver, ou placebo, caso contrário. Depois disso, as respostas dos dois grupos são comparadas.

Usualmente, os indivíduos para este tipo de estudo são escolhidos aleatoriamente, e o experimento é duplo cego: nem os indivíduos nem os médicos que medem as respostas devem saber quem está no grupo de tratamento nem quem está no grupo de controle. Estas idéias serão desenvolvidas no seguinte exemplo.

Nos anos 50, muitas vacinas contra a pólio foram descobertas. Uma, desenvolvida por Jonas Salk, parecia ser a mais promissora. Nos experimentos de laboratório, ela se mostrou efetiva e de fato provocou a produção de anticorpos contra a pólio. Um experimento a larga escala era necessário para ver se a vacina protegeria crianças contra a pólio fora do laboratório. Em 54, o serviço público de saúde decidiu organizar este experimento. Foram envolvidas aproximadamente dois milhões de crianças e em torno de meio milhão foi vacinada. Um outro meio milhão não permitiu ser vacinada. O experimento de campo foi feito sobre crianças nos grupos de idade mais vulneráveis - um, dois e três anos e foi levado a cabo nas regiões onde o risco de pólio acreditava-se ser maior.

Este é um exemplo do método de comparação. O tratamento (vacinação) é aplicado a alguns indivíduos, que formam o grupo de tratamento; os outros não recebem o tratamento e são usados como controle. As respostas dos dois grupos podem então ser comparadas para ver se o tratamento faz alguma diferença. Aqui, os grupos de tratamento e de controle foram de tamanhos diferentes, mas isso não era problema já que os pesquisadores estavam interessados em conhecer a taxa em que as crianças contraiam a pólio, em número de casos por mil. Olhar as taxas em vez dos números absolutos supera as diferenças de tamanho dos dois grupos.

Existe uma questão delicada de ética médica aqui. Não deveriam todas as crianças ter recebido a proteção da vacina? Uma resposta é que com novos medicamentos, mesmo depois de um extensivo período de teste de laboratórios, freqüentemente não é claro se os benefícios sobrepassam os riscos. Um experimento de campo é

necessário para encontrar o que o tratamento faz quando é usado na vida real. Por outro lado, dar a vacina a um grande número de crianças poderia parecer dar uma evidência decisiva, mesmo sem ter o grupo de controle. Por exemplo, se a incidência de pólio em 54 tivesse diferido marcadamente da de 53, isto pareceria uma prova da efetividade da vacina Salk. Mas na verdade não seria. A pólio é uma doença epidêmica, cuja incidência varia de ano para ano. Em 1952, houve em torno de 60000 casos, enquanto que em 53 houve a metade disso. Sem controles, uma baixa incidência em 54 poderia significar duas coisas: ou que a vacina funcionou ou que não houve epidemia esse ano.

A única forma de saber se a vacina funcionava era deixar algumas crianças sem vacinar. É claro que as crianças poderiam ser vacinadas somente com a aprovação dos pais. Assim, uma possível estratégia era esta: as crianças cujos pais consentissem formariam o grupo de tratamento e receberiam a vacina. As demais crianças formariam o grupo de controle. Mas era sabido que pais com mais alto nível sócio-econômico aceitariam o tratamento mais rapidamente que pais de nível mais baixo. Este fato teria criado um viés contra a vacina, já que crianças de nível econômico mais alto eram mais vulneráveis à pólio que as de nível mais baixo. Isto parece um paradoxo à primeira vista, mas pólio é uma doença de higiene. As crianças que vivem em um ambiente com pouca higiene tendem a contrair muitos casos leves de pólio no primeiro tempo de vida, enquanto ainda estão protegidas pelos anticorpos de sua mãe. Depois de serem infectadas, elas criam seus próprios anticorpos, que as protegem contra infecções mais sérias mais tarde. As crianças que vivem em um ambiente mais limpo estão menos predispostas a contrair estas infecções leves cedo, não desenvolvendo anticorpos, e são menos prováveis de estarem protegidas contra infecções mais sérias depois.

Portanto, do ponto de vista estatístico, para evitar o viés (erro sistemático), o grupo de tratamento e o grupo de controle deveriam ser o mais parecido possível, exceto pelo tratamento. Isto torna possível concluir que qualquer diferença na resposta entre os dois grupos é devida à presença do tratamento, mais que a qualquer outro fator. Se os dois grupos diferissem em algum fator além do tratamento, os efeitos deste outro fator poderiam ser confundidos (misturados) com os efeitos do tratamento. Efeitos misturados (confounding factors) são uma grande fonte de viés.

Voltando à pesquisa de campo da vacina Salk, várias estratégias foram propostas. A Fundação Nacional para a Paralisia Infantil (NFIP) queria vacinar todas as

crianças de dois anos cujos pais consentissem, deixando as crianças de um e três anos como controles. Esta proposta da NFIP foi aceita por muitos dos distritos; no entanto, tinha dois sérios defeitos. Primeiramente, a pólio é uma doença contagiosa, se espalha por contato. Assim, se a incidência fosse muito maior na faixa etária de dois anos, que nas outras duas, isto provocaria um viés no estudo contra a vacina. Ou se a incidência fosse muito menor nesta faixa etária, em relação com as outras duas, o estudo ficaria viesado a favor da vacina. Em segundo lugar, as crianças do grupo de tratamento (onde o consentimento dos pais era necessário) estavam limitadas a ter um histórico familiar diferente daquelas do grupo de controle (onde a autorização não era necessária). Com esta proposta, o grupo de tratamento incluiria muito mais crianças de famílias de maior nível econômico, tornando este grupo mais vulnerável à pólio que o grupo de controle. Isto daria novamente um viés contra a vacina.

Muitos distritos perceberam estas falhas na proposta da NFIP e adotaram portanto outra estratégia. Para fazer uma comparação válida, o grupo de controle deveria ser escolhido da mesma população que o grupo de tratamento: crianças cujos pais consentissem a vacinação. De outra forma, os efeitos dos antecedentes familiares poderiam ser confundidos com os efeitos da vacina. A próxima questão era como escolher as crianças para o tratamento e para o controle. Deveria ser um julgamento humano muito equilibrado para formar o grupo de tratamento e o de controle o mais similar possível com relação a variáveis relevantes, como a renda dos pais, o estado geral de saúde da criança, personalidade e hábitos sociais. Tais julgamentos poderiam resultar em um viés substancial. Foi portanto considerado mais adequado usar um procedimento aleatório cuidadosamente planejado. Para o experimento Salk, o procedimento usado foi equivalente a lançar uma moeda para cada criança, que então tinha um 50% de chance de estar em tratamento e um 50% de chance de estar no controle. Tal procedimento é imparcial e as leis probabilísticas garantem que, com um número suficiente de indivíduos, ambos os grupos permaneceriam muito próximos um do outro com relação a todas as variáveis importantes, estejam estas bem identificadas ou não.

> Quando um procedimento aleatório imparcial é usado para determinar os indivíduos do grupo de tratamento e do grupo de controle, o experimento se diz **controlado com aleatorização**.

Outra precaução básica foi o uso de um placebo. As crianças do grupo de controle receberam uma injeção de sal dissolvida em água, de modo que durante o

experimento, elas não soubessem se estavam em tratamento ou em controle. Isto assegurou que suas respostas se devessem à vacina e não à idéia do tratamento. Parece ser pouco provável que crianças possam estar protegidas contra a pólio somente pela força do pensamento, mas em inúmeros estudos feitos em pacientes que sofriam de uma forte dor pós-operatória foi dada a eles uma pílula feita de uma substância completamente neutra e descrita como um excelente analgésico contra a dor. Em torno de um terço dos pacientes obtiveram uma pronta melhora.

Ainda uma outra precaução do mesmo tipo: diagnósticos tinham que decidir se as crianças contraíram pólio durante esse ano. Muitas formas de pólio são difíceis de diagnosticar, e, em casos duvidosos, o diagnóstico pode ser facilmente afetado pelo conhecimento de se a criança foi vacinada contra a pólio ou não. Assim, os médicos que realizariam o diagnóstico não deveriam ser informados se a criança estava em tratamento ou era um controle. Este tipo de experimento é chamado duplo cego: nem o paciente e nem aquele que avalia a resposta ao tratamento sabem se ele recebeu o tratamento. Esta parte do experimento da vacina Salk foi do tipo **duplo cego**.

A Tabela 1.1 mostra o número total de indivíduos e a taxa de casos de pólio por cem mil indivíduos para os grupos de tratamento, de controle e não-autorizados. A taxa é muito menor para o grupo de tratamento, e esta é a prova decisiva da efetividade da vacina Salk.

	grupo de tratamento	*grupo de controle*	*não autoriza a participação*
experimento controlado com aleatorização duplo cego	200 000 - 28	200 000 - 71	350 000 - 46
proposta NFIP	225 000 - 25	725 000 - 54	125 000 - 44

Tabela 1.1: Resultados do experimento da vacina Salk de 1954: tamanho dos grupos e taxas de casos de pólio por 100 000 em cada grupo (números arredondados). Fonte: T. Francis Jr. (1955) Am. J. of Public Health.

Com o planejamento da NFIP, os resultados estão afetados por muitos fatores que (para os pesquisadores) são aleatórios: quais são as famílias voluntárias, quais crianças estão na faixa etária de dois anos, e assim por diante. No entanto, os pesquisadores não têm a mínima informação para estimar a chance de qualquer

um destes fatores, portanto não conseguem estimar o quanto esta diferença nas taxas de pólio é devida a eles.

Com o experimento controlado com aleatorização, a chance entra de uma forma simples e planejada ao determinar o tratamento ou o controle. Para explicar isto, a hipótese de advogado do diabo é que a vacina Salk não faz efeito nenhum. Sob esta hipótese, a determinação ao tratamento ou controle não tem nada a ver com a tendência de contrair pólio: cada criança tem uma chance de 50% de ficar em tratamento ou em controle, somente dependendo do lançamento de uma moeda. Assim, cada criança com maior tendência a contrair pólio tem uma chance meio a meio de estar em tratamento ou em controle. Portanto, o número de casos de pólio em ambos os grupos deve ser aproximadamente o mesmo, qualquer diferença é devida à variabilidade do lançamento da moeda. Esta variabilidade está bem estudada estatisticamente e é possível então calcular a chance de ter acontecido a variabilidade observada, que, neste caso, é astronômica: um em um bilhão.

Comparando os resultados dos dois planejamentos (veja Tabela 1.1), o NFIP apresentou um viés contra a vacina. O experimento controlado aleatorizado mostra que a vacina diminui a taxa de pólio de 71 a 28 por cem mil, uma redução da ordem de 60%. A redução aparente no modelo NFIP, de 54 a 25 por cem mil, é de apenas 54%. O impacto deste viés não parece ser muito sério; no entanto existem casos em que planejamentos mal feitos podem deixar os pesquisadores sem uma conclusão satisfatória.

1.1.2 Desvio da veia cava

Em alguns casos agudos de cirrose hepática, o paciente pode apresentar hemorragia, podendo chegar inclusive à morte. Um dos novos tratamentos surgidos nos anos 60 consistia em redirecionar o fluxo sangüíneo mediante o desvio da veia cava. Esta cirurgia era demorada e com altos riscos.

A pergunta natural do paciente é: os benefícios deste tratamento compensam os riscos?

Um estudo foi realizado analisando 51 estudos sobre os efeitos deste tratamento, relacionando **tipo de planejamento** adotado e **grau de entusiasmo** a favor da cirurgia. A tabela 1.2 resume os resultados:

Discussão. Observemos que dos 32 estudos sem controle, 75% foram entusiastas a respeito da cirurgia, dos 15 com controle não aleatorizado, 67% o foram, e

tipo de planejamento	*entusiasta*	*moderado*	*nenhum*
sem controle	24	7	1
controlado, sem aleatorização	10	3	2
controlado, com aleatorização	0	1	3

Tabela 1.2: Resultados dos tipos de planejamentos.

dos 4 com controle aleatorizado, nenhum foi. **Os planejamentos descuidados exageraram o valor desta cirurgia de risco**.

Em um experimento controlado com aleatorização, primeiramente devemos definir quem será a população a ser amostrada. Assim, por exemplo, haverá pacientes que não são selecionáveis para participar do estudo: pacientes muito doentes, que apresentem outras doenças, que não queiram participar do estudo, etc. Dentre os selecionáveis, os grupos de tratamento e controle são escolhidos aleatoriamente. Desta forma, a comparação somente será feita entre pacientes que podem de fato receber o tratamento.

Em experimentos precariamente controlados, é possível que pacientes não selecionáveis façam parte do grupo de controle, ou que dentre os pacientes selecionáveis para a cirurgia, o médico opte por operar somente os pacientes mais saudáveis, deixando os mais doentes como controle. Desta forma, as conclusões tiradas a partir da amostra apresentam um viés devido à inadequada seleção dos indivíduos.

De fato, isto é o que aconteceu nos estudos apresentados na tabela. Tanto nos estudos bem controlados quanto nos precariamente controlados, a taxa de sobrevivência a 3 anos no grupo de tratamento foi de 60%. Esta taxa foi a mesma para o grupo de controle, nos estudos com aleatorização. No entanto, a taxa de sobrevivência a 3 anos do grupo de controle, nos estudos sem aleatorização, foi apenas de 45%, indicando a tendência a escolher os casos mais graves como controle.

1.2 Estudos Observacionais

Em um estudo observacional, os pesquisadores não determinam quais os indivíduos que serão submetidos ao tratamento ou que serão controles. Por exemplo, os estudos sobre os efeitos do cigarro são necessariamente observacionais: ninguém vai fumar durante dez anos somente para agradar um estatístico. Mesmo assim, a terminologia tratamento-controle é ainda usada. Os pesquisadores comparam um

grupo de fumantes (o grupo de tratamento) com um grupo de não-fumantes (o grupo de controle) para determinar os efeitos do cigarro (o "'tratamento"').

Os fumantes geralmente se dão mal nesta comparação: doenças cardíacas, câncer de pulmão, acidentes automobilísticos e suicídios são mais comuns entre fumantes que entre não-fumantes. De modo que existe uma forte associação entre o hábito de fumar e doenças cardíacas, câncer de pulmão e o resto. Isto implica que existe uma relação de causalidade? Ou seja, o cigarro provoca câncer? O culpado é o ato de fumar? Fumar provavelmente não é bom, mas o veredito deve ser: não está provado. Poderia haver um fator hereditário que predispõe as pessoas a ambos: fumar e ter câncer. Isto poderia explicar a associação, os efeitos do fator hereditário estariam sendo confundidos com os efeitos do cigarro. E nesse caso, não haveria nenhum motivo para deixar de fumar, já que isso não alteraria o fator hereditário e nem reduziria o risco de câncer (não estamos sugerindo isto!!!).

Este argumento é hipotético, e de fato, há evidências de estudos com gêmeos univitelinos que o refutam. O importante é ressaltar que muitos problemas podem ser estudados somente mediante a observação e que em estudos observacionais devemos lidar com a possiblidade da mistura de efeitos. Portanto a escolha do grupo de controle deverá ser feita de modo que as diferenças entre os dois grupos possam ser creditados somente ao tratamento. Se houverem outros fatores cujos efeitos possam ser confundidos com os efeitos do tratamento, é importante encontrar quais ajustamentos poderiam ser feitos para superar este problema e definir se estes ajustamentos são razoáveis.

1.2.1 Pílula anticoncepcional e trombo-embolia

Em 1967, na Inglaterra, notou-se que as doenças trombo-embólicas tinham uma incidência muito maior entre as mulheres que usavam a pílula anticoncepcional do que entre as que não usavam.

Para estudar uma possível relação, Vessey e Doll (1969) consideraram os casos já ocorridos de tais doenças ao analisar os dados. Tal tipo de estudo é chamado **estudo retrospectivo**.

O estudo começou identificando mulheres entre 16 e 40 anos que tinham ingressado com um diagnóstico de trombose arterial, embolia pulmonar, trombose cerebral ou trombose coronária, em um dos 19 hospitais analisados. A idéia do estudo era entrevistar os casos para ver se a proporção de usuárias da pílula era maior

do que se poderia "esperar". Os pesquisadores precisavam saber de quanto seria esta proporção esperada, supondo que o uso da pílula não tivesse relação com a predisposição a uma trombose.

Estudando ao mesmo tempo um grupo de mulheres comparáveis aos casos, mas sem o quadro de trombo-embolia, a proporção de usuárias da pílula neste grupo de controle é usada como uma proporção padrão.

Duas mulheres de controle foram escolhidas por cada caso: ambas tinham sido submetidas a uma cirurgia aguda ou estavam em condições clínicas semelhantes, e tinham a mesma idade, data de entrada no hospital e paridade (número de filhos) que o caso. Os controles foram escolhidos de modo a não ter qualquer predisposição para doenças trombo-embólicas. Se não houvesse nenhuma relação entre o uso da pílula e essas doenças, então a proporção de usuárias no grupo de casos não deveria ser maior do que no grupo de controle.

No presente estudo, 42 de 84 casos (50%) usavam a pílula, enquanto que apenas 23 dos 168 controles (14%) eram usuárias.

Após decidir, que esta diferença é improvável de acontecer por mero acaso, os autores concluíram que existe uma relação entre o uso da pílula e a predisposição para doenças trombo-embólicas.

Este estudo é um exemplo de **estudo caso-controle**. O objetivo de tal estudo é examinar potenciais fatores de risco para uma doença. O estudo começa identificando os casos com a doença específica e depois, selecionando um grupo de controle: um grupo de indivíduos comparáveis aos casos, cuja única diferença seja, além da ausência da doença, a possibilidade de ausência dos fatores de risco. Os dois grupos são então examinados para ver se os fatores de risco são mais freqüentes nos casos do que nos controles.

1.3 Amostragem

Em uma eleição, onde temos milhões de eleitores, desejamos saber quem será o novo presidente da república, ou se haverá segundo turno. É possível obter uma resposta satisfatória antes da contagem total dos votos?

Se a população de interesse for grande, (126 milhões de eleitores no Brasil, 28 milhões no estado de São Paulo, por exemplo) é claramente inviável entrevistar cada um deles para conhecer sua intenção de voto em uma dada eleição.

Podemos então estimar a proporção de votantes em cada candidato a partir da proporção observada em uma amostra da população total de eleitores. Se a amostra for representativa da população de interesse, então os resultados observados deveriam ser próximos dos valores na população.

Como escolher uma amostra representativa pode se basear no conhecimento prévio da população ou, caso este não esteja disponível, em alguns métodos que minimizam a chance de viés, chamados métodos probabilísticos de amostragem.

O seguinte é um exemplo de uma amostragem feita de forma incorreta, levando a conclusões erradas.

1.3.1 Eleição presidencial de 1946, EUA

Na eleição presidencial dos EUA em 1946, Roosevelt era o candidato democrata e Landon era o candidato republicano. Naquela época, o país estava lutando para se recuperar da Grande Depressão da década de 30, e havia cerca de 9 milhões de desempregados. Ambos os candidatos centralizaram suas campanhas nesta crise.

Uma revista de prestígio, que já havia previsto o candidato vencedor em eleições anteriores, indicou Landon como vencedor com uma esmagadora maioria de votos, 57% a 43%. No entanto, Roosevelt foi eleito com 62% dos votos.

Onde a revista errou?

O número amostrado foi mais do que o suficiente, porém o procedimento como foi feita a amostra estava errada, a revista colheu a amostra da seguinte maneira: enviaram questionários para 10 milhões de pessoas, eles obtiveram 2,4 milhões de respostas. Os nomes e endereços dessas pessoas foram retirados da lista telefônica, sendo que em 1946 haviam apenas 11 milhões de telefones residenciais e 9 milhões de desempregados, com isso eles excluíram os mais pobres. Portanto houve um grande erro na seleção da amostra, pois quando há um método sistemático de amostragem ou quando se exclui um tipo de pessoa, a amostra é chamada tendenciosa e não nos fornece informação significativa. Em outras eleições essa tendência não tinha afetado o resultado pois os pobres e ricos votavam em uma linha similar, porém devido a crise houve uma ruptura nessa linha e as pessoas mais pobres em sua maioria votaram em Roosevelt. Quando o método de seleção é tendencioso, tomarmos uma amostra grande não ajuda. Isto apenas repete o erro básico em uma grande escala. Outro erro cometido pela revista foi o fato de enviarem 10 milhões

de questionário e receberem 2,4 milhões de respostas, isso causa o chamado viés de não resposta.

Esse viés ocorre devido as classes sociais, sendo que o número de pessoas que não respondem varia dependendo da classe. E o fato de ter enviado questionários ao invés da pesquisa ser feita por entrevistas pessoalmente, aumentou o viés por não resposta pois as pessoas tendem a responder mais quando a pesquisa é feita pessoalmente.

Amostragem por cota

Em outra eleição presidencial nos EUA, três pessoas fizeram uma pesquisa para dizer quem seria o novo presidente dos EUA. Os três pesquisadores disseram que Dewer seria o vencedor com uma vantagem de 5 pontos percentuais, porém quem realmente ganhou a eleição foi Truman. A tabela 1.3 mostra a previsão dos pesquisadores e o resultado do dia da eleição.

	Previsões			
Os candidatos	*Crossley*	*Gallup*	*Roper*	*Os resultados*
Truman	45	44	38	50
Dewey	50	50	53	45
Thurmond	2	2	5	2
Wallace	3	4	4	2

Tabela 1.3: Previsões e Resultados

Fonte: F. Mosteller e outros, **A pesquisa pré-eleitoral de 1948**. Nova Yorque: Conselho de Pesquisa e Ciência Social.

O método escolhido foi amostragem por cota. Neste método, cada entrevistador entrevista uma cota fixa de pessoas classificas por categorias. Geralmente estas categorias são definidas de acordo com endereço, idade, classe social, sexo e raça, sendo o entrevistador livre para escolher o entrevistado dentro de cada categoria. O censo, por exemplo, pode ser utilizado para definir esta divisão.

Aparentemente a amostragem parece representar bem a população e nos fornecer bons resultados. Então qual foi o erro cometido pelos três pesquisadores?

Um dos erros que podemos apontar (posteriormente ao erro cometido) pode ter surgido pelo fato de que nos EUA há dois partidos, o republicano e o democrata. As entidades de pesquisa querem uma amostra que represente a opinião da população, da nação política. Entretanto, e é este o primeiro ponto a ser observado, não há

cotas que podem ser configuradas sobre votos republicanos e democratas, já que a distribuição da opinião política na nação é precisamente o que as organizações de pesquisas desconhecem e estão tentando descobrir.

Um segundo erro cometido, e possivelmente o mais grave, foi que o entrevistador tinha o livre arbítrio para escolher o entrevistado. Isto deixa muito para escolha humana e esse fato pode causar viés na amostra: o entrevistador poderia preferir entrevistar certo tipo de pessoa, ou aquelas de mais fácil acesso.

Nessa pesquisa os entrevistadores escolheram muitos republicanos para entrevistarem, porque os republicanos em geral eram mais educados, mais bem humorados e gentis do que os democratas.

Os entrevistadores deveriam ter escolhido os indivíduos por critérios pré-determinados para escolher o entrevistado, sem levar em conta a sua preferência pessoal, por exemplo: entrevistar uma pessoa a cada 15 entre as que passam por um certo ponto, ou, se o entrevistador for de casa em casa, entrevistar uma casa a cada 20 casas, etc.

Boas referências na área de amostragem são os livros clássicos de Kish [12] e de Cochran [4], onde são discutidos procedimentos sobre métodos de amostragem, estratégias para diminuir não-resposta e viés na amostragem.

Capítulo 2

Cálculo de Probabilidades

Este capítulo tratará de diversos problemas de cálculo de probabilidades, muitos deles apresentados sob a forma de jogos entre dois adversários: um jogador (Você) e a casa.

As referências utilizadas neste capítulo são alguns dos livros clássicos neste tema, como Sheldon Ross [18], Feller [8], Mosteller [14].

Antes de continuar, lembremos de algumas definições básicas.

Definição 1 (Experimento aleatório) *Chamamos de experimento aleatório qualquer fenômeno ou observação experimental cujo resultado não seja exatamente conhecido.*

Observemos que esta definição é bastante ampla, dando margem a fenômenos por ocorrer ou já ocorridos, cujo resultado não conhecemos.

Definição 2 (Espaço amostral) *O espaço amostral associado a um experimento é o conjunto de todos os resultados possíveis de serem observados.*

Definição 3 (Evento) *Um evento é um subconjunto observável do espaço amostral.*

2.1 Jogo dos dados Jankenpon

Material. um dado vermelho A com duas faces 10 e quatro faces 1; um dado verde B com as seis faces iguais a 4; um dado azul C com duas faces 0 e quatro faces 6 como mostra a Figura 2.1.

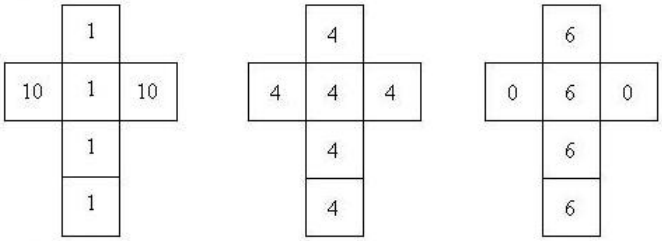

Figura 2.1: Os dados jankenpon.

Variantes: as faces dos dados podem ser diferentes das sugeridas, mas observe que a média para os três dados é a mesma, no caso, igual a 4; o número de dados pode ser maior.

Jogo. Você escolhe um dos dados e depois a casa escolhe um dos outros dois. Ganha o melhor de cinco lançamentos.

Pergunta. Qual dos dados Você prefere escolher?

Discussão. Com uma rápida olhada, Você provavelmente ficará na dúvida em quanto ao melhor dado, como usualmente ocorre neste jogo.

Observe que entre A e B, B tem duas vezes mais chance de ganhar, como mostra a Tabela 2.2. E entre B e C, C tem chances de 2 a 1 de ganhar, como mostra a Tabela 2.3.

A \ B	0	0	6	6	6	6
4	A	A	B	B	B	B

Figura 2.2: Resultados possíveis entre os dados A e B.

C \ B	0	0	6	6	6	6
1	C	C	B	B	B	B
1	C	C	B	B	B	B
1	C	C	B	B	B	B
1	C	C	B	B	B	B
1	C	C	C	C	C	C
1	C	C	C	C	C	C

Figura 2.3: Resultados possíveis entre os dados B e C.

Portanto, C é melhor do que B e B é melhor do que A. No entanto, entre A e C, C tem chances de 4 a 5 de ganhar, o que ocorre somente se A mostrar 1 e C, 6, como mostra a Tabela 2.4. Assim A é melhor do que C.

A \ C	1	1	1	1	10	10
4	A	A	A	A	C	C

Figura 2.4: Resultados possíveis entre os dados C e A.

Com isto, não existe "o" melhor dado, o que explica a "gentileza" da casa ao permitir que Você escolha primeiro.

Este jogo é conhecido na literatura como dados de Efron ou dados não-transitivos, já que a noção de transitividade de conjuntos completamente ordenados se perde. Na verdade o que está sendo comparado são probabilidades calculadas em espaços diferentes, portanto não há motivo para esperar uma ordem total entre os dados.

Mais exemplos de dados podem ser encontradas no livro de Ivars Peterson (2002) e nas referências aí citadas. Uma generalização para mais de dois jogadores pode ser encontrada na página de Ed Pegg Jr. [15].

2.2 Jogo das três cartas e Jogo dos dois gatinhos

Estes dois jogos, inspirados em Smullyan [19], são muito similares, já que em ambos devemos atualizar nossa distribuição de probabilidade a partir de uma informação dada, lidando portanto com probabilidades condicionais.

Ambos os jogos podem ser encontrados em livros introdutórios de probabilidade, como Berry [2], pp.132 e 138. O primeiro também é conhecido como o problema das três panquecas.

Material para as três cartas: três cartas de mesmo tamanho com uma figura de cada lado, a primeira com um rubi e uma esmeralda, como na Figura 2.5, a segunda com dois rubis e a terceira com duas esmeraldas.

Jogo. As cartas são embaralhadas e Você escolhe uma delas, observando a pedra em uma das faces. Você pode apostar em que a pedra da outra face é diferente ou em que a pedra da outra face é a mesma. Se Você perder, a casa ganha.

Pergunta. Em qual resultado Você aposta?

Discussão. Pela simetria da disposição das figuras nas cartas, a chance da casa ganhar é a mesma tanto se a face selecionada for esmeralda quanto se for rubi. Suponhamos então que foi rubi.

Deste modo, Você selecionou ou a face rubi da carta com pedras diferentes, ou qualquer uma das duas faces da carta com dois rubis. Em outras palavras, Você tem o dobro de chance de ter selecionado a carta com dois rubis. Assim, a probabilidade do resultado com faces diferentes é apenas 1/3.

Observamos, no entanto, que o pensamento mais comum é que já que uma das duas cartas foi selecionada, então a chance de ser a carta com pedras diferentes é 1/2, sem levar em conta que a carta com dois rubis é mais verossímil de ter sido a escolhida.

Figura 2.5: Ilustração da primeira carta: frente e verso.

Material para os dois gatinhos: dois pares de cartas do mesmo tamanho, nos quais uma carta tem um gato macho desenhado e a outra um gato fêmea, como na Figura 2.6.

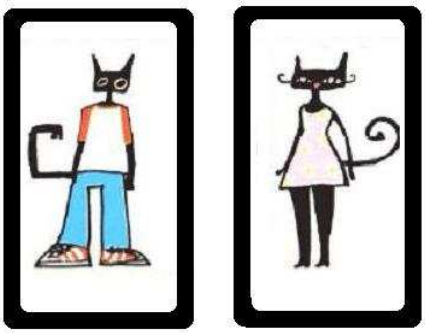

Figura 2.6: Ilustração dos gatos charmosos: um macho e uma fêmea

Jogo. Escolha ao acaso uma carta de cada par, sem olhar o resultado. Depois escolha ao acaso uma das duas cartas selecionadas e olhe a figura, observando se o gato é fêmea ou macho. Você pode apostar que o gatinho da outra carta é do mesmo sexo, ou não. Se Você perder, a casa ganha.

Pergunta. Em qual resultado Você aposta?

Discussão. Novamente, pela simetria do problema, a chance da casa ganhar é a mesma se o gatinho observado primeiro for macho ou fêmea. Suponhamos que foi macho.

Deste modo, na primeira seleção, Você selecionou ou duas cartas iguais (dois machos) ou um macho do primeiro par e uma fêmea do segundo, ou uma fêmea do primeiro par e um macho do segundo. Em outras palavras, há o dobro de chance de ter selecionado dois gatos de diferente sexo, e assim, a probabilidade de obter gatos do mesmo sexo é 1/3.

Neste problema, também é comum pensar que, como o sexo do primeiro gatinho é independente do sexo do segundo, a informação sobre um deles não afeta a probabilidade de que o outro seja macho ou fêmea. Este raciocínio, no entanto, não leva em conta o fato de que selecionar gatos de sexo diferente tem o dobro de chance do que selecionar gatos de mesmo sexo, depois de conhecer o sexo de um dos gatos.

2.3 Paradoxo do tempo de espera

Material. um dado balanceado de seis faces numeradas de 1 a 6; uma trilha com 8 casas; um peão para avançar pela trilha.

Variantes: o dado pode ter qualquer número F de faces a partir de duas, a trilha deve ter um número de casas maior que F, pode haver mais de um peão percorrendo a trilha simultaneamente.

Jogo. Você avança com o peão pela trilha de acordo com o lançamento do dado. O último lançamento é aquele que permite alcançar (ou passar) a última casa. Você ganha se, neste último lançamento, obtiver uma das faces 1, 2 ou 3; a casa ganha se for obtida uma das faces 4, 5 ou 6.

Explicação 1 (intuitiva).

Se depois de lançar o dado, o peão se encontrar na casa 1, então nem Você nem a casa podem ganhar no próximo lançamento, já que ainda restam sete casas.

Se o peão se encontrar na casa 2, então a casa pode ganhar no próximo lançamento se sair a face 6, com probabilidade 1/6, mas você não tem chance nenhuma de ganhar com mais um lançamento.

Se o peão estiver na casa 3, a casa pode ganhar com mais um lançamento se sair face 6 ou 5, com probabilidade 2/6. Do mesmo modo, se o peão estiver na casa 4, ainda a casa pode ganhar com somente mais um lançamento, se sair 6, 5 ou 4, com probabilidade 1/2.

Somente se o peão se encontrar na casa 5, Você poderá ganhar com mais um lançamento, se sair face 3, com probabilidade 1/6, e a casa, se sair 4, 5, ou 6, com probabilidade 1/2. Novamente, se o peão estiver na casa 6, você tem a possibilidade de ganhar no próximo lançamento se você obtiver 2 ou 3, com probabilidade 2/6, enquanto que a casa mantém probabilidade 1/2 de ganhar no próximo lançamento.

Finalmente, se o peão estiver na casa 7, então somente neste caso ambos, casa e Você, têm mesma chance de ganhar o jogo.

Portanto, a casa tem mais chances de ganhar o jogo que Você.

Explicação 2 (mais formal)

Denotando as trajetórias como um vetor, uma possível trajetória para a corrida é a sequência $(5, 1, 4)$, na qual o primeiro lançamento é "5", o segundo é "1" e o último lançamento é "4", totalizando 10 passos.

Seja N o número de lançamentos necessários pra terminar a corrida, e X_N o valor que aparecer no último lançamento. Denotemos a posição do jogador imediatamente antes do seu último lançamento (que é aquele em que o jogador conclui a corrida) por S_{N-1}.

A variável aleatória S_{N-1} pode assumir somente valores no conjunto $\{K-1,\ K-2,\ K-3,\ K-4,\ K-5,\ K-6\}$. Se $S_{N-1} = K-1$, então o jogador irá terminar a corrida no próximo lançamento do dado, não importando o seu valor.

Se $S_{N-1} = K-2$, então $X_N \in \{2, 3, 4, 5, 6\}$. Se $S_{N-1} = K-3$, então $X_N \in \{3, 4, 5, 6\}$, e assim por diante até o caso $S_{N-1} = K-6$, no qual a corrida termina com $X_N = 6$.

Segue que $X_N = 1$ com probabilidade 1/6 quando $S_{N-1} = K-1$; $X_N = 2$ com probabilidade 1/6 quando $S_{N-1} = K-1$ e $X_N = 2$ com probabilidade 1/5 quando $S_{N-1} = K-2$, e assim sucessivamente.

Assim

$$
\begin{aligned}
Pr(X_N = 1) &= \frac{1}{6}\, Pr(S_{N-1} = K-1) \qquad (2.1)\\
Pr(X_N = 2) &= Pr(X_N = 1) + \frac{1}{5}\, Pr(S_{N-1} = K-2)\\
Pr(X_N = 3) &= Pr(X_N = 2) + \frac{1}{4}\, Pr(S_{N-1} = K-3)\\
Pr(X_N = 4) &= Pr(X_N = 3) + \frac{1}{3}\, Pr(S_{N-1} = K-4)\\
Pr(X_N = 5) &= Pr(X_N = 4) + \frac{1}{2}\, Pr(S_{N-1} = K-5)\\
Pr(X_N = 6) &= Pr(X_N = 5) + \frac{1}{1}\, Pr(S_{N-1} = K-6)
\end{aligned}
$$

e portanto

$$Pr(X_N = 1) < Pr(X_N = 2) < \cdots < Pr(X_N = 6)$$

sempre que $K > 6$. Em outras palavras, valores maiores tem maiores chances de serem obtidos no último lançamento.

Conseqüentemente, a probabilidade de obter 1, 2 ou 3 no último lançamento é menor que 1/2. O valor correto é em torno de 30%.

As combinatórias necessárias podem ser facilmente dominadas por estudantes de graduação. (Dica: obtenha as probabilidades de S_{N-1} nas equações (2.1).)

O Grafico 2.7 mostra a frequência dos valores obtidos no último lançamento em uma simulação de 40 realizações da corrida: a face "6" aparece 11 vezes, enquanto que a face "1" aparece somente duas vezes.

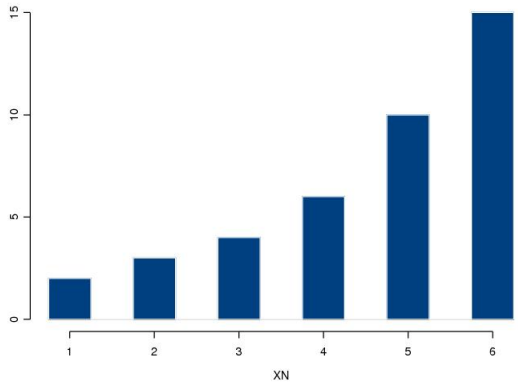

Figura 2.7: Gráfico de Frequências para o último lançamento.

Esta é uma versão discreta do conhecido Paradoxo do Tempo de Espera, originalmente estabelecido no contexto de processo de renovação em tempo contínuo. No entanto, não há paradoxo.

Em nosso exemplo, para cada valor fixo de n, a variável aleatória X_n tem distribuição uniforme em $\{1, \ldots, 6\}$. Entretanto, esse não é o caso da variável aleatória X_N. Devemos manter em mente que N, o número total de passos, também é uma variável aleatória. É intuitivo que um salto maior (digamos 6 passos) tenha mais chances de cobrir o fim da corrida do que um salto menor (digamos 1 passo).

O paradoxo pode ser facilmente encontrado na literatura. Feller [9], p.12, por exemplo, define o problema para o Processo de Poisson, obtendo o exemplo mais conhecido de Tempo de Espera para a chegada do ônibus em um ponto do trajeto. Uma versão para um dado com dois lados pode ser encontrada em Rifo [17].

David [5] discute o efeito do paradoxo no planejamento de experimentos para testar tempo de vida útil, que, para um amostrador descuidado, pode levar à superestimação do tempo médio de vida. Os rendimentos de um supermercado podem ser superestimados ao realizar uma amostragem aleatória simples do valor gasto por clientes selecionados em um instante determinado do dia: clientes que demoram mais tempo no caixa (e que provavelmente gastam mais) tendem a aparecer mais na amostra.

2.4 Jogo do anel

Material. círculo com uma marca no Pólo Norte; bolinhas de bingo enumeradas de 1 a 60.

Jogo. Dois pontos serão sorteados ao acaso no círculo, dividindo-o em dois arcos. Antes do sorteio, Você pode escolher o arco com o Pólo Norte ou o arco complementar. Você ganha se seu arco for o maior depois do sorteio dos pontos.

O sorteio pode ser feito com as bolinhas de bingo, indicando os pontos sorteados como nos ponteiros de um relógio.

Pergunta. Qual arco você escolhe: com ou sem Pólo Norte?

Discussão. Suponhamos sem perda de generalidade que o primeiro ponto, P_1, cai na metade leste do círculo, isto é, é um dos valores de 1 a 30. Então, se o segundo ponto, P_2, cair no intervalo $(P_1, P_1 + 30)$, o Pólo Norte ganha. Mas, o Pólo Norte

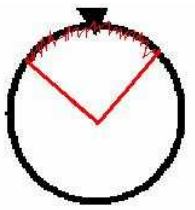

Figura 2.8: Ilustração do Jogo do anel. Neste exemplo, o arco com o Pólo Norte perde.

ganha também se P_2 estiver em $(0, P_1)$. Portanto, dado P_1, a casa tem chance maior que $1/2$ de ganhar, para todo valor possível de P_1.

Observe que com a discretização, a chance de empate não é nula. Neste caso, o sorteio deve ser feito novamente.

Para calcular exatamente a probabilidade do Pólo Norte ganhar, cortemos o círculo no Pólo Norte, identificando-o com o intervalo $[0, 1]$, de modo que os extremos 0 e 1 representem o Pólo Norte.

Selecionar dois pontos ao acaso, corresponde a definir duas variáveis aleatórias uniformes em $[0, 1]$, U_1 e U_2, independentes entre si. A casa ganha se $|U_1 - U_2| < 1/2$, o que ocorre com probabilidade $3/4$, como indicado na Figura 2.9.

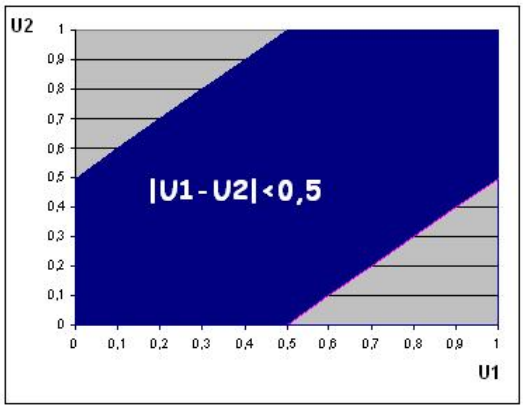

Figura 2.9: A região destacada indica o evento em que a casa ganha.

Este é um caso do Paradoxo do Tempo de Espera, começando na casa aleatória U_1, e fixando o número de passos igual a dois, com tamanhos $|U_1 - U_2|$ e $1 - |U_1 - U_2|$. Novamente, é mais fácil cobrir um ponto fixado com o intervalo maior.

2.5 Jogo das amebas

Material. Uma moeda balanceada.

Jogo. Você tem uma ameba que, com probabilidade $\frac{1}{2}$ morre, e com probabilidade $\frac{1}{2}$ se divide em duas outras amebas, após um certo tempo. Cada uma destas amebas tem o mesmo comportamento da primeira. Chame a primeira ameba de geração 0. Você pode apostar que na quinta geração existirá pelo menos uma ameba ou que na quinta geração não existirá nenhuma.

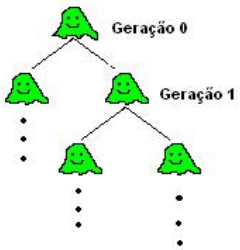

Figura 2.10: Uma família de amebas.

Pergunta. Em qual resultado Você prefere apostar?

Discussão. Observe que haverá amebas na quinta geração se e somente se a primeira ameba se dividir E pelo menos uma das amebas da geração 1 se dividir E pelo menos uma das amebas de cada geração seguinte se dividir. Se, no entanto, a primeira ameba morrer, o que ocorre com probabilidade 1/2, então não haverá amebas na quinta geração.

Ou seja, a probabilidade de que não haja nenhuma ameba na quinta geração é maior que 1/2.

Denotemos por X_n a variável que indica o número de amebas na n-ésima geração.

Assim, pelo enunciado do problema, temos que X_1 pode assumir os valores 0 ou 2, com $P(X_1 = 0) = 1/2 = P(X_1 = 2)$.

Para a segunda geração, temos que X_2 pode ser 0, 2, 4, com

$$\begin{aligned}
P(X_2 = 0) &= P(X_1 = 0) + P(X_1 = 2)P(X_2 = 0|X_1 = 2) = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}\right)^2 \\
P(X_2 = 2) &= P(X_1 = 2)P(X_2 = 2|X_1 = 2) = \frac{1}{2}\,2\left(\frac{1}{2}\right)^2 = 2\left(\frac{1}{2}\right)^3 \\
P(X_2 = 4) &= P(X_1 = 2)P(X_2 = 4|X_1 = 2) = \frac{1}{2}\left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^3,
\end{aligned}$$

de modo que $P(X_2 = 0) = 0.625$.

Para a terceira geração temos que X_3 pode ser 0, 2, 4, 6 ou 8, com

$$\begin{aligned} P(X_3 = 0) &= P(X_2 = 0) + P(X_2 = 2)\, P(X_3 = 0|X_2 = 2) \\ &\quad + P(X_2 = 4)\, P(X_3 = 0|X_2 = 4) \\ &= \frac{1}{2} + \left(\frac{1}{2}\right)^3 + 2\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^4 \\ &= \frac{1}{2} + \left(\frac{1}{2}\right)^3 + 2\left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^7 = 0.695. \end{aligned}$$

Continuando desta maneira, obtemos que $P(X_5 = 0) \approx 0.735$, quase três quartos de chance para o primeiro resultado.

2.6 Problema do campeão e do iniciante

Este problema foi retirado de Mosteller [14].

Um jovem jogador com uma carreira promissora tem que disputar três jogos, podendo optar por uma das duas seguintes seqüências.

Poderá começar jogando com um jogador iniciante (I), depois com o campeão do clube (C) e novamente com o iniciante (seqüência ICI). Ou ele poderá disputar o primeiro jogo com o campeão do clube, depois com o iniciante e novamente com o campeão (seqüência CIC).

Para passar à fase seguinte, o jovem jogador deve ganhar no mínimo duas partidas consecutivas.

Pergunta. Qual sequência o jovem jogador deveria escolher?

Este problema pode ser transformado na seguinte atividade.

Material. Dois dados com as faces $C = (V, V, D, D, D, D)$ e $I = (V, V, V, V, V, D)$.

Dica: Você poderá escolher as faces 1 e 2 de um dado balanceado para representar V no primeiro dado e para o segundo dado poderá escolher as faces 1,2,3,4,5 de outro dado balanceado.

Jogo. Cada dado representa um dos possíveis adversários, com V indicando vitória e D, derrota na partida. Você pode escolher uma das duas seqüências ICI ou CIC, e a casa fica com a outra seqüência. Ambos, a casa e Você, realizam os lançamentos das seqüências escolhidas alternadamente. Ganha quem obtiver pelo menos duas vitórias consecutivas nos 3 lançamentos.

Pergunta. Qual sequência Você prefere?

Discussão. Denotemos por p a probabilidade de Você ganhar como o dado C e por k, a probabilidade de Você ganhar com o dado I.

As tabelas abaixo mostram os resultados favoráveis para cada uma das seqüências e suas respectivas probabilidades.

I	C	I	Probabilidade
V	V	V	$kp\,k$
V	V	D	$kp\,(1-k)$
D	V	V	$(1-k)p\,k$
Total			$kp\,(2-k)$

C	I	C	Probabilidade
V	V	V	$p\,kp$
V	V	D	$p\,k(1-p)$
D	V	V	$(1-p)kp$
Total			$kp\,(2-p)$

Assim, para a sequência ICI, a probabilidade de se obter duas vitórias consecutivas é $kp\,(2-k) = (5/6) * (2/6) * 2 - (5/6) = 0,32$, e para a sequência CIC, a probabilidade é $kp\,(2-p) = (5/6)*(2/6)*2 - (2/6) = 0,46$. Portanto, a seqüência CIC é mais vantajosa.

Intuitivamente, para obter duas vitórias consecutivas, é necessário obter vitória no segundo lançamento, que é mais provável de obter com o dado I na segunda jogada.

Denotemos por p_a a probabilidade de ganhar com ICI, e por p_b, a probabilidade de ganhar com CIC, e suponhamos que Você escolhe a seqüência CIC, lembrando que a casa começa a jogar.

Assim teremos que a probabilidade de Você ganhar é

$$\begin{aligned}
P(CIC) &= (1-p_a)p_b + (1-p_a)^2(1-p_b)p_b + (1-p_a)^3(1-p_b)^2p_b + \dots \\
&= p_b(1-p_a)\left[1 + (1-p_a)(1-p_b) + (1-p_a)^2(1-p_b)^2 + \dots\right] \\
&= \frac{p_b(1-p_a)}{1-(1-p_a)(1-p_b)} \\
&= \frac{(0,46)(1-0,32)}{1-(1-0,32)(1-0,46)} = 0,495.
\end{aligned}$$

Se Você escolher a seqüência ICI, então a probabilidade de Você ganhar é

$$\begin{aligned} P(ICI) &= (1-p_b)p_a + (1-p_b)^2(1-p_a)p_a + (1-p_b)^3(1-p_a)^2p_a + \cdots = \\ &= p_a(1-p_b)\left[1 + (1-p_b)(1-p_a) + \ldots\right] \\ &= \frac{p_a(1-p_b)}{1-(1-p_b)(1-p_a)} \\ &= \frac{(0,32)(1-0,46)}{1-(1-0,46)(1-0,32)} = 0,273. \end{aligned}$$

Portanto, se Você escolher a sequência ICI, Você terá apenas 27,3% de chances de ganhar o jogo, enquanto que se Você escolher a sequência CIC, Você terá 49,5% de chances de ganhar o jogo.

2.7 Jogo do máximo

Estes jogos foram criados a partir de Feller [8], capítulo XV, exercício 3.

2.7.1 Máximo em 2

Material. dois dados comuns de 6 faces.

Jogo. Você lança os dados, observando os valores obtidos. A casa ganha se o máximo observado for 4, 5 ou 6; caso contrário, você ganha.

Pergunta. Qual é a probabilidade da casa vencer?

Discussão. Seja A o evento "a maior face é igual 4, 5 ou 6".

Podemos representar todos os resultados possíveis para o valor máximo em dois lançamentos como na seguinte tabela.

faces	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

Assim, a probabilidade do evento A é igual a

$$P(A) = \frac{25}{36} = 0,694.$$

2.7.2 Máximo em 3

Material. três dados comuns de 6 faces.

Jogo. Você lança os dados, observando a maior valor obtido. Você pode apostar que o máximo observado será 5 ou 6, ou que será um dos valores 1, 2, 3 ou 4.

Pergunta. Qual aposta Você prefere?

Discussão. Denotemos por B o evento "a maior face observada é igual a 1, 2, 3 ou 4".

Observe que o evento B ocorre se não for observado nem 5 nem 6, em nenhum dos três lançamentos. Portanto,

$$P(A) = 1 - \left(\frac{4}{6}\right)^3 = 0,704.$$

2.7.3 Minimax em 3

Material. três dados comuns de 6 faces.

Jogo. Você lança os dados, observando os valores obtidos. A casa ganha se o mínimo for 1 ou o máximo for 6. Em qualquer outro caso, Você ganha.

Discussão. Denotemos por C o evento "a face mínima é 1 ou a face máxima é 6". Novamente, observemos que o evento C não ocorre se e somente se não for observado nem 1 nem 6 em nenhum dos três lançamentos. Portanto,

$$P(C) = 1 - \left(\frac{4}{6}\right)^3 = 0,704.$$

2.8 Jogo da soma em três dados

Material. três dados comuns de 6 faces.

Jogo. Você lança os dados, observando os valores obtidos. Você pode apostar que a soma das três faces obtidas será um dos valores $\{7, 8, 9, 10, 11, 12, 13, 14\}$ ou que será um dos valores $\{3, 4, 5, 6, 15, 16, 17, 18\}$.

Pergunta. Em qual opção você prefere apostar?

Discussão. Considere o lançamento de três dados balanceados, sabendo que os valores possíveis para soma S obtida nas faces dos dados são: 3, 4, 5 ,6 ,7 ,8 ,9 ,10

, 11, 12, 13, 14, 15, 16, 17 ou 18. Denotemos por ijk cada um dos 6^3 resultados possíveis e por $\pi(ijk)$ o total de permutações dos valores ijk.

Com esta notaçao, a probabilidade de cada um destes resultados é

$$
\begin{aligned}
P(S=3) &= P(111) = (1/6)^3 \\
P(S=4) &= P(\pi(211)) = 3\,(1/6)^3 \\
P(S=5) &= P(\pi(311), \pi(221)) = 6\,(1/6)^3 \\
P(S=6) &= P(\pi(411), 222, \pi(321)) = 10\,(1/6)^3 \\
P(S=7) &= P(\pi(511), \pi(412), \pi(331), \pi(322)) = 15\,(1/6)^3 \\
P(S=8) &= P(\pi(611), \pi(512), \pi(431), \pi(422), \pi(332)) = 21(1/6)^3 \\
P(S=9) &= P(\pi(612), \pi(531), \pi(522), \pi(441), \pi(432), 333) = 25(1/6)^3 \\
P(S=10) &= P(\pi(622), \pi(613), \pi(541), \pi(532), \pi(442), \pi(433)) = 27(1/6)^3 \\
P(S=11) &= P(\pi(641), \pi(632), \pi(551), \pi(542), \pi(533), \pi(443)) = 27(1/6)^3 \\
P(S=12) &= P(\pi(651), \pi(642), \pi(633), \pi(552), \pi(543), 444) = 25(1/6)^3 \\
P(S=13) &= P(\pi(661), \pi(652), \pi(643), \pi(553), \pi(544)) = 21(1/6)^3 \\
P(S=14) &= P(\pi(662), \pi(653), \pi(644), \pi(554)) = 15(1/6)^3 \\
P(S=15) &= P(\pi(663), \pi(654), 555) = 10(1/6)^3 \\
P(S=16) &= P(\pi(664), \pi(655)) = 6(1/6)^3 \\
P(S=17) &= P(\pi(665)) = 3(1/6)^3 \\
P(S=18) &= P(666) = (1/6)^3
\end{aligned}
$$

A soma é igual a qualquer valor entre 7 e 14 com probabilidade igual à soma das probabilidades destes resultados, $0,814$. Portanto, a primeira opção é a melhor.

2.9 Jogo do tiro ao alvo

Material. três quadrados com lados 40 cm, 30 cm e 20 cm (cada quadrado deve ser reticulado em cem quadradinhos de tamanhos iguais numerados de 1 a 100); um círculo de raio de 3 cm (para ser o disco); alfinetes ou fita adesiva para fixar o círculo sobre os quadrados; números de 1 a 100 para sorteio (tabela de números aleatórios ou papéis numerados). Os quadrados e o círculo poderão ser feitos de E.V.A ou cortiça.

Jogo. O jogador deve "lançar" aleatoriamente o disco em cada um dos quadrados.

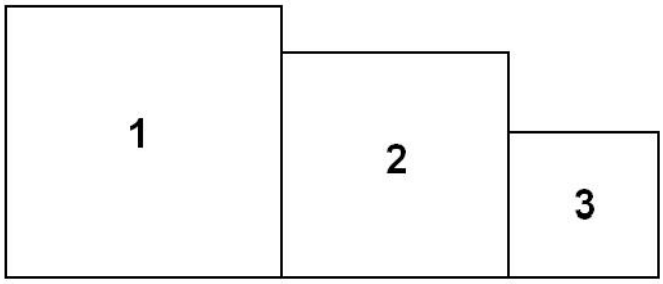

Figura 2.11: Tiro ao alvo.

Ganha o jogo se o disco estiver completamente incluído dentro do quadrado, em cada um dos três lançamentos.

Os "lançamentos aleatórios" são definidos pelo sorteio de 3 números entre 1 e 100 indicando o quadradinho numerado onde deve ficar o centro do disco.

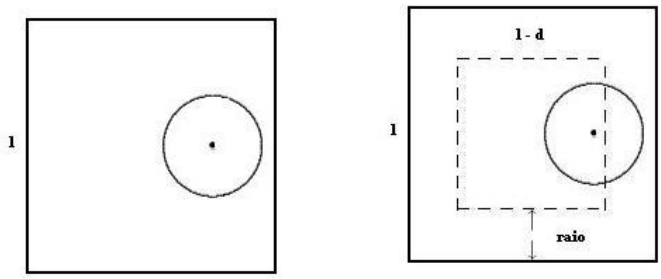

Figura 2.12: Figura de um quadrado com o disco.

Discussão. Denotemos por l o lado do quadrado e por r o raio do círculo.

Observemos que o jogador acerta o lançamento se e somente se o centro do círculo estiver a uma distância maior do que r de qualquer um dos lados do quadrado, isto é, se o centro do círculo ficar dentro do quadrado menor da Figura 2.12. Este quadrado menor tem lado $l - 2r$.

Como o centro pode ser sorteado em qualquer lugar do quadrado, a probabilidade de que caia dentro do quadrado menor é igual à proporção entre as áreas do quadrado menor e do maior

$$P(\text{acerto em um lançamento}) = \frac{(l-2r)^2}{l^2}.$$

Assim para cada um dos quadrados, temos probabilidades de acerto iguais a

$$\begin{aligned} p_1 &= \frac{(40-6)^2}{40^2} &= \frac{34^2}{40^2} = 0,7225 \\ p_2 &= \frac{(30-6)^2}{30^2} &= \frac{24^2}{300^2} = 0,64 \\ p_3 &= \frac{(20-6)^2}{20^2} &= \frac{14^2}{20^2} = 0,49. \end{aligned}$$

Finalmente, o jogador ganha o jogo se acertar em todos os lançamentos. Isto ocorre com probabilidade igual a

$$p = p_1\, p_2\, p_3 \; = 0,7225\, 0,64\, 0,49 = 0,227.$$

Podemos perceber que a probabilidade do jogador acertar apenas um dos quadrados é alta, porém a probabilidade dele acertar os três quadrados é baixa, já que é o produto de valores menores que um.

2.10 Problema das Três Portas

Material. três cartas de um baralho, duas cartas do mesmo naipe e a terceira de naipe diferente.

Jogo. Um apresentador de um programa de TV mostra a um candidato três portas fechadas. Ele informa que atrás de uma delas há um carro e nas outras não há prêmios. O candidato, então escolhe uma das portas mas não a abre. O apresentador, que sabe onde está o carro, abre uma das portas que não foi escolhida e mostra que atrás dela não há prêmio nenhum. Veja figura 2.13. Nesse momento, o apresentador pergunta para o candidato se ele deseja mudar de porta. É vantajoso para o candidato mudar de porta?

Discussão. Aparentemente esse problema parece bem simples. Considere três portas, A, B e C. Quando o jogador escolhe uma delas, digamos a A, a chance de que ela seja a premiada é de 1/3, contra 2/3 de chance dele errar. Porém quando o apresentador (que sabe onde está o prêmio) abre uma das portas, suponha que seja a porta B.

Quando o apresentador pergunta ao jogador se ele deseja mudar, ele intuitivamente deve pensar que se agora só restaram duas portas a chance do prêmio estar em uma das portas é 1/2, então tanto faz ele mudar ou não. Porém essa resposta está errada, o pensamento intuitivo falha nesse caso devido o jogador pensar que o ato

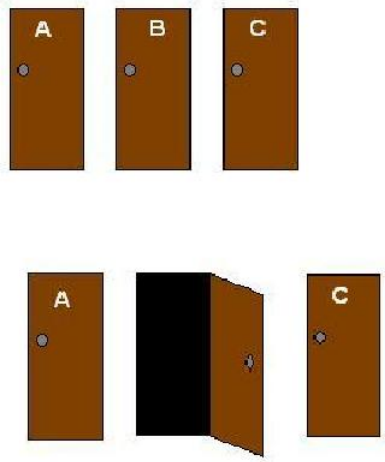

Figura 2.13: Ilustração das portas

do apresentador abrir uma porta, está sendo criado um jogo todo novo, mas não é isso que ocorre, o apresentador sabe desde o começo onde está o prêmio (ele nunca abrirá uma porta premiada), ao abrir a porta ele nos fornece uma informação importante. Quando o jogador escolheu a porta A, a chance de ele ter escolhido a porta sem o prêmio é mais provável (2/3), então considerando que o prêmio não estava na porta que ele escolheu, o carro estará nas portas (B ou C) que tinham em conjunto 2/3 de probabilidade de ter o carro, e como o apresentador abriu a porta B, a porta NÃO escolhida que continua fechada passa a ter 2/3 de probabilidade de ser a porta do carro, ou seja a porta C.

Esse problema também é conhecido como Problema de Monty Hall ou dos Três Prisioneiros, e foi discutido em Esteves et al. [6] e Morgado [13].

2.11 As duas moedas

Problema. Consideremos duas moedas, assumindo que a moeda A tem probabilidade 0,7 de resultar em cara quando lançada e a moeda B tem probabilidade 0,6 de resultar cara quando lançada.

As moedas serão lançadas consecutivamente de acordo com a seguinte regra: se a

moeda lançada no instante n resultar em cara, então a moeda A será lançada no instante $n+1$, e se resultar coroa, a moeda B será lançada no instante $n+1$.

Se a moeda inicialmente lançada tem mesma probabilidade de ser a moeda A ou a moeda B, qual é a probabilidade de que a moeda utilizada no terceiro lançamento seja a moeda A?

Discussão. Se a face da moeda lançada no instante n resultar cara, então lançaremos a moeda A que no próximo lançamento resultará em cara com probabilidade 0,7 e coroa com probabilidade 0,3. Porém se sair coroa no lançamento n, escolheremos a moeda B para o lançamento $n+1$, que poderá resultar em cara com probabilidade 0,6 e coroa com probabilidade 0,4.

Para $n \geq 1$, definamos as variáveis aleatórias X_n como

$$X_n = \begin{cases} 0 & \text{se sair coroa no } n\text{-ésimo lançamento} \\ 1 & \text{se sair cara no } n\text{-ésimo lançamento .} \end{cases}$$

A Tabela 2.1 mostra os possíveis resultados em três lançamentos.

X_1	Moeda	X_2	Probabilidade condicional	Moeda	X_3	Probabilidade condicional
1	A	1	0,7	A	1	0,7
					0	0,3
		0	0,3	B	1	0,6
					0	0,4
0	B	1	0,6	A	1	0,7
					0	0,3
		0	0,4	B	1	0,6
					0	0,4

Tabela 2.1: Tabela dos casos possíveis em três lançamentos e probabilidades condicionais.

Denotemos por A_i usar a moeda A no i-ésimo lançamento, e por B_i usar a moeda B no i-ésimo lançamento.

Se escolhermos a moeda A ou B com mesma probabilidade para o mesmo lançamento, observemos que então a probabilidade de obter cara no primeiro lançamento é

$$\begin{aligned} P(X_1 = 1) &= P(A_1)P(X_1 = 1|A_1) + P(B_1)P(X_1 = 1|B_1) \\ &= 0,5 \times 0,7 + 0,5 \times 0,6 = 0,65 \end{aligned}$$

e a de obter coroa, analogamente, é

$$\begin{aligned} P(X_1 = 0) &= P(A_1)P(X_1 = 0|A_1) + P(B_1)P(X_1 = 0|B_1) \\ &= 0,5 \times 0,3 + 0,5 \times 0,4 = 0,35. \end{aligned}$$

Notemos que os valores que aparecem nas expressões acima podem ser escritos como o resultado do produto matricial

$$\begin{pmatrix} 0,5 & 0,5 \end{pmatrix} \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix}.$$

Do mesmo modo, a probabilidade de obter cara no segundo segundo lançamento é

$$\begin{aligned} P(X_2 = 1) &= P(A_2)P(X_2 = 1|A_2) + P(B_2)P(X_2 = 1|B_2) \\ &= P(X_1 = 1)P(X_2 = 1|A_2) + P(X_1 = 0)P(X_2 = 1|B_2) \\ &= 0,65 \times 0,7 + 0,35 \times 0,6 = 0,665. \end{aligned}$$

Observemos novamente que o resultado acima corresponde à primeira coordenada do produto matricial

$$\begin{pmatrix} 0,65 & 0,35 \end{pmatrix} \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix} = \begin{pmatrix} 0,5 & 0,5 \end{pmatrix} \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix} \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix}.$$

Repetindo este processo, obtemos que as probabilidades de X_n podem ser escritas como

$$\begin{pmatrix} P(X_n = 1) \\ P(X_n = 0) \end{pmatrix} = \begin{pmatrix} 0,5 & 0,5 \end{pmatrix} \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix}^n. \tag{2.2}$$

Esta representação matricial é possível porque o experimento determina que a probabilidade dos eventos do lançamento $(n+1)$ depende apenas do conhecimento do evento ocorrido no lançamento n e das probabilidades de cada moeda.

A seqüência de variáveis aleatórias $X_1, X_2, \ldots$ com esta propriedade é chamada cadeia de Markov e a matriz

$$P = \begin{pmatrix} 0,7 & 0,3 \\ 0,6 & 0,4 \end{pmatrix}$$

é chamada matriz de transição da cadeia. As linhas desta matriz indicam a probabilidade condicional de cada uma dos possíveis eventos (0 ou 1) no passo $(n+1)$, dado cada possível estado no passo n.

Assim, a probabilidade de obter cara no terceiro lançamento corresponde à primeira coordenada do produto (2.2), para $n = 3$, igual a 0,6665.

2.12 Problemas dos dois rádios

Este problema foi retirado de Feller [8], capítulo XV.

Consideremos dois rádios, que podem estar ligados ou desligados durante um dia qualquer. No dia n cada rádio estará ligado, independentemente do estado do outro rádio, com probabilidade

$$\frac{1+\text{ número de rádios ligados durante o dia } (n-1)}{4}.$$

Por exemplo, se ambos os rádios estiverem ligados durante o dia $(n-1)$, então cada um deles estará ligado, independentemente do estado do outro, durante o dia n com probabilidade $3/4$.

Que proporçao de dias ambos os rádios estarão ligados durante um ano? E que proporçao dos dias ambos os rádios estarão desligados?

Discussão.

Denotemos por X_n o número de rádios ligados durante o dia $(n-1)$. Assim, o enunciado nos diz que

X_{n-1}	probabilidade de cada rádio ligado no dia n
0	1/4
1	2/4
2	3/4

A informação dada pelo problema pode ser representada na matriz da transição P, onde cada linha corresponde a um dos possíveis estados de X_{n-1}, 0, 1 ou 2, e cada coluna entrega a distribuição de probabilidade condicional de X_n, dado X_{n-1},

$$P = \begin{pmatrix} (3/4)(3/4) & 2(1/4)(3/4) & (1/4)(1/4) \\ (1/2)(1/2) & 2(1/2)(1/2) & (1/2)(1/2) \\ (1/4)(1/4) & 2(1/4)(3/4) & (3/4)(3/4) \end{pmatrix} = \begin{pmatrix} 9/16 & 6/16 & 1/16 \\ 1/4 & 1/2 & 1/4 \\ 1/16 & 6/16 & 9/16 \end{pmatrix}.$$

Denotemos por π_i a probabilidade de que em um dia qualquer $i-1$ rádios permaneçam ligados, $i \in \{1,2,3\}$. Esta probabilidade pode ser interpretada como a proporção das vezes em que $X_n = i-1$, quando n tende a ∞.

Por um resultado da teoria de cadeias de Markov (cf. Feller [8]), podemos provar que o vetor definido por $\pi = (\pi_1, \pi_2, \pi_3)$ deve satisfazer a seguinte equação

$$\pi P = \pi.$$

Obtemos assim um sistema de equações lineares em π_i, cuja solução, para nosso exemplo, é $\pi = (2/7, 3/7, 2/7)$. Em particular, a proporção do tempo em que ambos os rádios permanecem desligados é $\pi_1 = 2/7$.

2.13 Edwiges no mercado

Edwiges deixou seu carro no estacionamento e foi ao mercado. Ao retornar ao estacionamento, percebeu que estava sem as chaves do carro.

Ela imagina que a chave pode estar em alguma das lojas do mercado com probabilidade p e fora do mercado, na banca de jornal, com probabilidade $1-p$.

Dentro do mercado há seis lojas nas que Edwiges passou e, de acordo com sua opinião, a probabilidade da chave estar em qualquer uma das seis lojas é a mesma.

Edwiges voltou ao mercado em busca da sua chave. Entrou na loja 1 e não a encontrou. Foi na loja 2 e também não. Passou pela loja 3 e a chave também não estava lá. Assim ela fez passando pela loja 4 e 5, sem sucesso.

Pergunta. Qual é a probabilidade de que a chave esteja na loja 6?

Discussão. De acordo com a opinião inicial de Edwiges, a probabilidade de que a chave esteja em uma das lojas do mercado é $p/6$, enquanto que a chance de estar na banca de jornal, fora do mercado, é igual a $(1-p)$.

Depois de Edwiges ter descartado cinco das seis lojas do mercado, a probabilidade de que a chave esteja na loja 6 deve ser determinada condicionalmente sobre esta nova informação,

$$\begin{aligned}
&P(\text{a chave está na loja 6} \,|\, \text{a chave está na loja 6 ou fora do mercado}) \\
&= \frac{P(\text{a chave está na loja 6})}{P(\text{a chave está na loja 6 ou fora do mercado})} \\
&= \frac{\frac{p}{6}}{\frac{p}{6} + (1-p)}.
\end{aligned}$$

Supondo, por exemplo, $p = 0,6$, teremos que inicialmente a probabilidade de estar na loja 6 era

$$\frac{0,6}{6} = 0,1,$$

e depois que Edwiges constatou que a chave não estava nas lojas 1 a 5, a probabilidade de estar na loja 6 passou para

$$\frac{0,1}{0,1+0,4} = 0,2.$$

Assim, a informação de que a chave não está nas lojas 1, 2, 3 ,4 e 5 aumenta a probabilidade de estar na loja 6 de 10% para 20%.

Capítulo 3

Inferência Estatística

As referências utilizadas neste capítulo são principalmente Freedman et al. [7] e Bekman e Costa Neto [1].

3.1 Paradoxo de Simpson

Discriminação por sexo em admissões na graduação.

Uma fonte de problemas em estudos observacionais são as diferenças entre os indivíduos da amostra que poderiam afetar as conclusões do estudo, sem serem percebidas em uma primeira análise. Algumas vezes estas diferenças podem ser identificadas depois, permitindo realizar ajustamentos comparando subgrupos menores e mais homogêneos. Este processo é chamado controle do fator de confusão.

A seguir apresentamos um exemplo clássico de dados que rapidamente analisados podem levar a conclusões errôneas (como, de fato, levaram), mas que olhados mais cuidadosamente nos mostram a eficiência de um estudo bem feito.

Alertados sobre uma possível discriminação por gênero na admissão à graduação na Universidade da Califórnia, Berkeley, no semestre de outono de 1973, uma análise dos ingressantes foi feira.

Naquele ano, inscreveram-se 8442 homens e 4321 mulheres para a admissão na graduação. Em torno de 44% dos homens e 35% das mulheres foram admitidos. Considerar a porcentagem ajusta a diferença entre a quantidade de homens e mulheres inscritos no processo.

Assumindo que homens e mulheres estão igualmente qualificados (e não há nenhuma evidência em contra), a diferença na taxa de admissão parece ser uma forte

evidência para mostrar que homens e mulheres são tratados de forma diferente no processo de admissão: em um grupo de homens e mulheres igualmente qualificados, a universidade parece preferir os homens por 44 a 35.

As admissões à graduação podem ser separadas por carreira, e olhando cada uma delas separadamente, deveria ser possível identificar as que discriminam mulheres em favor dos homens. Mas, carreira por carreira, não parecia haver nenhum viés contra as mulheres; algumas favoreciam os homens e outras as mulheres. No total, se houvesse um viés parecia ser contra os homens. O que estava acontecendo?

Analisar cada uma das carreiras, mais de cem no total, não acrescentaria muita informação na análise, por isso foram escolhidas as seis carreiras mais concorridas, contabilizando mais de um terço do número total de candidatos. O perfil deles era típico para todas as carreiras, portanto a análise destas seis carreiras poderia ser estendida às demais.

A Tabela 3.1 mostra o número de homens e mulheres candidatos, e as porcentagens de admissão para cada uma destas seis carreiras principais.

	Homens		**Mulheres**	
Carreira	Número de Candidatos	Porcentagem Admitidos	Número de Candidatos	Porcentagem Admitidos
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Tabela 3.1: Dados da admissão para os programas de graduação nas seis principais carreiras na Universidade da Califórnia Berkeley, outono, 1973. Fonte: Divisão de Graduação, Universidade da Califórnia, Berkeley.

Podemos observar que em cada carreira, a porcentagem de candidatas que foram admitidas é muito similar à dos homens.

A exceção mais marcante é a carreira A, que parece desfavorecer os homens: 82% das mulheres foram admitidas, e apenas 62% dos homens.

A carreira que parece discriminar mais contra as mulheres é a E, que admitiu 28% dos homens e 24% das mulheres; uma diferença de apenas 4%.

No entanto, quando todas as carreiras são tomadas em conjunto, elas admitiram 44% dos homens e somente 30% das mulheres: uma diferença de 14%!!!

Isto de fato parece um paradoxo, mas existe uma explicação.

Uma observação mais detalhada mostra que as primeiras duas carreiras têm taxa de admissão mais elevada, ou seja, são mais fáceis de entrar, e que mais de 50% dos homens se candidataram a elas.

As outras quatro carreiras têm uma taxa de admissão bem menor, são mais difíceis de entrar, e mais de 90% das mulheres se candidataram nestas.

Assim, podemos ver que mais homens se inscreveram nas carreiras mais fáceis, e mais mulheres nas mais difíceis.

Existe um efeito da escolha da carreira nas taxas de admissão confundido com o efeito do gênero. Quando a escolha da carreira é controlada como na tabela anterior, existe pouca diferença nas taxas de admissão para homens e mulheres.

Em muitos estudos observacionais, é possível controlar os fatores de confusão por um processo similar: fazendo comparações separadas em subgrupos homogêneos.

A conclusão principal deste exemplo é que uma análise precipitada pode levar a conclusões incorretas.

Se as admissões são feitas por carreira, é inapropriado comparar as taxas de admissão para homens e mulheres no campus como um todo. Uma comparação cuidadosa deve controlar a escolha da carreira, dado que esta característica tem um efeito de confusão.

Para a pergunta sobre a existência de discriminação por sexo nas admissões para a graduação em Berkeley, a resposta é de que não há evidências que a confirmem. Se existir discriminação, ela é mais sutil. Para comprová-lo deveria mostrar-se que as candidatas mulheres são no total mais qualificadas que os candidatos homens, ou então que as taxas de admissão são deliberadamente mantidas mais baixas nas carreiras onde há maior taxa de candidatas mulheres. Mas os dados apresentados não permitem concluir nenhuma dastas teorias.

A Tabela 3.1 poderia ser resumida, calculando uma taxa de admissão geral para homens e outra para mulheres, ajustando-as à diferença por sexo nas taxas de aplicação em cada carreria. O procedimento é o de tomar uma espécie de média ponderada, separadamente para homens e mulheres, onde os pesos são o número total de candidatos por carreira sobre o total de candidatos, apresentados na Tabela 3.2.

Carreira	Número total de candidatos
A	933
B	585
C	918
D	792
E	584
F	714
	4,526

Tabela 3.2: Alunos aceitos por carreira, para as seis carreiras mais concorridas.

Sendo assim, a média ponderada para a taxa de admissão dos homens é

$$\frac{(933)(62) + (585)(63) + (918)(37) + (793)(33) + (584)(28) + (714)(6)}{4526},$$

em torno de 39%.

Similarmente, a taxa de admissão das mulheres é

$$\frac{(933)(82) + (585)(68) + (918)(34) + (793)(35) + (584)(24) + (714)(7)}{4526}$$

igual a 43%. Nestas fórmulas os pesos são os mesmos para homens e mulheres, extraídos do número total de candidatos, e as taxas de admissão são as extraídas da Tabela 3.1.

A comparação final é portanto uma taxa de admissão para os homens de 39% e para as mulheres de 43%.

3.2 Associação não é causalidade

3.2.1 Estudos observacionais

Geralmente quando se quer estudar o quanto uma variável influencia em outra, o pesquisador altera a variável independente e observa qual é o efeito que a variável dependente sofreu.

Por exemplo, suponhamos que um pesquisador deseja determinar qual é a relação entre a espessura de uma esponja e o peso de uma carga colocada nela. Ele utiliza uma esponja sintética e sobre ela coloca placas de pesos diferentes, verificando que

quanto mais pesada a placa, menor é a espessura da esponja, e quanto mais leve a placa, menor é a alteração na espessura original da esponja.

Podemos ver o peso da placa de metal como a variável independente, que o pesquisador pode variar dentro de certos limites, e a espessura da esponja como a variável dependente, pois neste caso o pesquisador não escolhe os valores da espessura, apenas observa como ela se comporta para os deversos valores de peso da placa.

Já que o peso é controlado pelo pesquisador, ele pode verificar a influência do peso da placa sobre a espessura da esponja.

Em outro estudo, por exemplo, um pesquisador deseja saber como o peso depende da altura em adolescentes do sexo feminino de uma certa idade. Neste caso, podemos considerar o peso como a variável dependente e a altura, como a variável independente, já que geralmente consideramos a altura influenciando o peso, e não o contrário.

Se o pesquisador pudesse controlar a variável independente, fixando-a em certos valores escolhidos por ele, então uma forte associação argumentaria a favor de uma possível relação de causalidade. Se o pesquisador apenas observar os valores das variáveis independentes e dependentes no estudo realizado, a associação observada não permite estabelecer uma relação de causalidade.

Um exemplo disto é o caso que veremos a seguir referente ao consumo de refrigerante e a prevalência de pólio, nos dados coletados nos anos 50 nos Estados Unidos. O segundo exemplo descreve uma análise relacionando cegonhas e nascimentos de bebês.

3.2.2 Pólio e refrigerante

Antes da criação da vacina contra pólio, nos anos 50 (cf. Seção 1.1.1), alguns estudos mostraram uma correlação positiva entre a incidência da pólio e o número de refrigerantes vendidos, ou seja, em períodos onde a venda de refrigerantes era maior, havia também mais novos casos de pólio relatados, e em período em que as vendas de refrigerante caíam, havia menos novos casos de pólio.

Podemos então concluir que refrigerante causa a pólio? Se assim fosse, e a venda do refrigerante fosse proibida, haveria uma redução na incidência de novos casos da doença?

Olhando retrospectivamente, sabemos que não há relação de causalidade entre consumo de refrigerante e a doença da poliomelite. Mas na época, a informação sobre as possíveis causas e comportamento epidemiológico dessa doença ainda estavam sendo estudados.

O mesmo ocorre em estudos atuais em que a informação sobre o assunto em estudo ainda está sendo obtida, seja em estudos médicos, em estudos da área de ciências sociais ou de ciências físicas, por exemplo.

Neste caso, a pergunta é: qual é a variável de confusão correlacionada tanto com número de casos de pólio quanto com venda de refrigerante?

Os cientistas conseguiram perceber que ambas as variáveis são afetadas pelo fator de confusão "estação do ano" ou, mais precisamente, "temperatura".

Sendo assim no verão, em que as condições de higiêne são mais afetadas, a incidência de pólio aumenta, assim como a venda de refrigerantes. Por isto, ambas as variáveis aparesentavam uma alta correlação positiva.

Correlação mede associação. Mas associação não é o mesmo que causalidade.

3.2.3 Quanto mais cegonhas, mais bebês?

Em uma famosa conferência feita em 1952 pelo estatístico Jerzy Neyman, foi dado mais este exemplo mostrando que correlação não é causalidade, e que procura pelo fator de confusão e por explicações plausíveis deve guiar de maneira geral estudos observacionais.

Na conferência, ele conta a história de um amigo que recolheu dados sobre a população de cegonhas e o número de nascimentos em 54 condados dos Estados Unidos.

A correlação entre o número de cegonhas e o número de bebés nascidos durante o período considerado era próxima do valor máximo 1, indicando que em condados com mais cegonhas havia também mais nascimentos.

Diante desta evidência, o amigo sugeria que, como a presença de cegonhas claramente influenciava os nascimentos, fosse realizada uma experiência de controle de natalidade, removendo todas as cegonhas de um subconjunto de condados selecionados aleatoriamente.

O problema do amigo de Neyman era não considerar a variável de confusão "número de mulheres" nos condados analisados. Esta variável está correlacionada com o

número de nascimentos, já que se uma região tem mais mulheres, então, em média, deveria ter mais nascimentos. E está também correlacionada com o número de cegonhas no condado, já que um grande número de mulheres é indicativo de uma população maior, e portanto de mais casas onde cegonhas podem construir seus ninhos.

Ao dividirmos os condados por grupos que tenham aproximadamente o mesmo número de mulheres, podemos observar que dentro de cada um desses grupos a correlação entre cegonhas e nascimentos é praticamente nula.

Condicionando a análise por uma variável de controle, no caso o número de mulheres, que está relacionada tanto com a suposta variável independente quanto com a variável dependente, podemos eliminar a influência dessa variável, estimando assim um valor menos viesado para a correlação entre elas.

3.3 Correlação Ecológica

Anos atrás um pesquisador publicou um artigo sobre a relação entre o uso do cigarro e o câncer de pulmão. Uma importante evidência usada foi um diagrama ter mostrado a relação entre a taxa do uso do cigarro (por pessoa) e a taxa de câncer no pulmão, em onze países. A correlação entre esses onze pares de taxas foi de 0,7, que é considerada uma forte correlação.

No entanto, não é o país que fuma e adquire o câncer, e sim as pessoas. Para medir corretamente a correlação, é necessário ter dados relacionando o cigarro e a doença com cada indivíduo. Quando a pesquisa foi realizada não era possível ter esta informação, por isso o pesquisador usou taxas para países - era tudo o que ele podia obter.

Os coeficientes de correlação baseados em taxas ou médias são frequentemente enganosas, pois mostram algo que na realidade não ocorre.

O seguinte estudo é um exemplo disto.

A partir dos dados do censo de 1970, é possível calcular a correlação entre o salário e o nível educacional, para homens com idade entre 35 e 54 anos nos EUA. Essa correlação era aproximadamente de 0,4, o que mostra que os dados são pouco correlacionados, como mostra a Figura 3.1, lado esquerdo.

O censo dividiu os Estados Unidos em 9 regiões geográficas. Para cada região, é

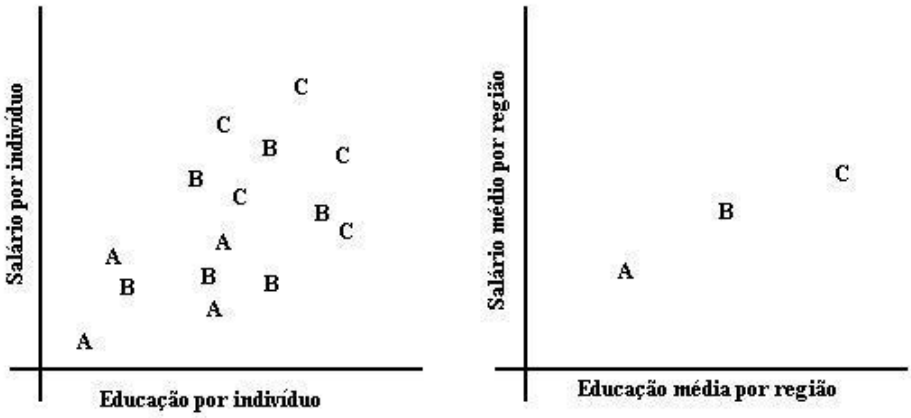

Figura 3.1: O gráfico do lado esquerdo representa os salários e nível educacional por indivíduo em três regiões geográficas, denominadas A, B e C. Cada indivíduo é marcado pela letra mostrando a região de sua residência. O gráfico do lado direito mostra os valores médios destas variáveis para cada região: a correlação entre as médias é quase 1.

possível calcular a média dos salários e a média do nível educacional para homens que vivem naquela região, mostrados na Figura 3.1, lado direito.

Podemos perceber que a correlação fica perto de um ao considerar apenas os 9 pares de médias; de fato, o coeficiente de correlação linear neste caso é aproximadamente 0,7.

A razão desta diferença é a grande dispersão e heterogeneidade dentre de cada região, o que não permite que médias sejam bons indicadores das variáveis em questão.

Ao substituir os valores individuais pelas médias, eliminamos a dispersão, provocando a falsa impressão de forte correlação entre os dados. De fato, se tivéssemos apenas duas regiões, a correlação seria igual a 1.

Correlações baseadas em taxas ou médias são chamadas correlações ecológicas, e são normalmente utilizadas em ciências políticas e sociais.

3.4 Falácia da regressão

3.4.1 Relação entre altura de pais e filhos

A palavra **regressão** foi criada pelo estatístico Sir Francis Galton em um estudo realizado em 1885, relacionado diversas variáveis biométricas entre pais e filhos.

O Gráfico 3.2 mostra as alturas obtidas para pais e filhos (em polegadas), indicando uma clara correlação positiva entre estas variáveis, ou seja, pais mais altos têm em média filhos mais altos.

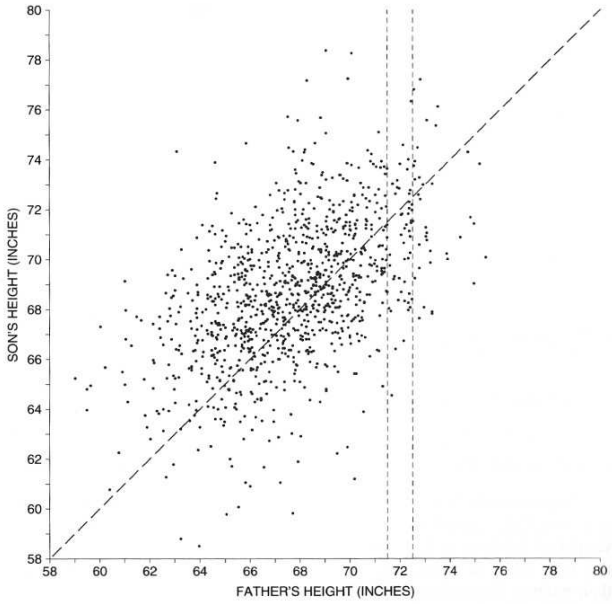

Figura 3.2: Gráfico de alturas de pais e filhos, obtidas no estudo de Sir Galton, em 1885. Extraído de Freedman et al. [7].

A altura média de pais e filhos, neste conjunto de dados, é aproximadamente 68 polegadas.

Consideremos o conjunto dos pais com altura em torno de 72 polegadas, marcado no gráfico pelos dois segmentos horizontais. Observemos que, neste conjunto, a altura média dos filhos não é igual a 72 polegadas; a altura média dos filhos é menor que 72 polegadas, ficando mais próxima da altura média dos filhos.

Esta mesma constatação, pode ser feita para alturas maiores dos pais, e para alturas dos pais menores que a média: a altura média dos filhos em cada uma dessas regiões tende a ficar mais próxima da média do que a altura dos pais. É como se as alturas dos filhos regredissem à altura média.

Esta característica é chamada efeito de regressão.

3.4.2 Teste e reteste em crianças de uma pré-escola

O seguinte exemplo foi extraído de Freedman et al. [7].

Um programa de pré-escola foi desenvolvido para melhorar o coeficiente de inteligência (IQ) das crianças. Basicamente, o procedimento consistia em que as crianças fizessem um teste no momento de entrar no programa (pré-teste) e depois quando elas deixassem o programa (pós-teste).

Em ambos os testes, a média dos pontos das crianças ficou próximo de 100 com um desvio-padrão de aproximadamente 15 unidades, de modo que aparentemente o programa não tinha surtido nenhum efeito.

Analisando mais profundamente os dados, no entanto, percebeu-se um fato inesperado. As crianças que estavam abaixo da média no pré-teste apresentaram em média 5 pontos a mais no pós-teste. Por outro lado, aquelas crianças que estavam acima da média no pré-teste apresentaram perda de 5 pontos em média no pós-teste.

O que isso prova? O programa opera de maneira a igualar as inteligências? Ou que crianças menos inteligentes ficam mais inteligentes depois do programa? Mas que crianças mais inteligentes ficam menos inteligentes depois do programa?

O que poderia explicar isso? Quando crianças mais "brilhantes" brincam com as crianças menos interessadas, a diferença entre os dois grupos tendem a diminuir?

Todas estas perguntas podem ser muito interessantes para um educador, mas a análise estatísticas destes dados não permite responder a nenhuma delas. O que foi constatado acima é conseqüência simplesmente do efeito da regressão.

Intuitivamente, é claro que não podemos que todas as crianças obtenham exatamente a mesma pontuação nos dois testes. Assim, considerando o caso extremo de crianças que tiveram nota máxima no primeiro teste, é razoável que a média no segundo teste seja menor, já que nem todas estas crianças conseguirão nota máxima no segundo. Do mesmo modo, das que tiraram nota mínima no primeiro teste, dificilmente todas elas tirarão nota mínima também no segundo teste.

Como vimos no exemplo das alturas de pais e filhos, ao condicionar a nota do primeiro teste, obtemos que a nota média do segundo teste, neste grupo, se aproxima da nota média total do segundo teste.

3.5 Um problema de decisão

Este divertido exemplo de teoria de decisão foi extraído do excelente livro de Bekman e Costa Neto [1].

Edson é um torcedor fanático da seleção brasileira, que está tentando resolver um problema dias antes da final da copa do mundo, entre a seleção do Brasil e a da Argentina, que será no Maracanã.

Como a televisão dele quebrou, ele pode ficar em casa sozinho ouvindo o jogo pelo rádio a pilha que ele tem, ou ele pode assistir o jogo na TV de plasma de 84 polegadas na casa de seu grande amigo Armando, que é argentino, ou ele pode comprar passagens e ingresso de última hora para ir ao Maracanã.

O que ele deve fazer?

Edson pode optar por uma das três possíveis ações:

$$\begin{aligned} A_1 &= \text{assistir o jogo no estádio} \\ A_2 &= \text{assistir na super TV de seu amigo argentino} \\ A_3 &= \text{ouvir o jogo sozinho no rádio a pilha} \end{aligned}$$

Analisando os possíveis resultados do jogo, temos apenas duas opções, ou, na terminologia da Teoria da Decisão, dois estados da natureza:

$$\begin{aligned} A &= \text{Derrota da seleção brasileira} \\ B &= \text{Vitória da seleção brasileira} \end{aligned}$$

Cada uma das ações tem diversos aspectos a serem considerados.

Por exemplo, a ação A_1 envolve o desconforto e os custos de ir para outra cidade assistir o jogo: fila para comprar o ingresso, fila para entrar no estádio, custo das passagens, preço do ingresso (provavelmente no mercado informal), e assim por diante. Mas, se o Brasil ganhar, todos esses contratempos não terão nenhuma importância para ele, já que ele terá muita alegria em estar lá, assistindo todo espetáculo bem na sua frente com várias pessoas festejando junto. Em sua opinião, a escolha de ir ao estádio será a ação mais acertada em caso de vitória da seleção brasileira. Por outro lado, se a Argentina ganhar ele vai ficar muito triste por ter gasto todas as suas economias para assistir a vitória da Argentina. Isto, em sua opinião, seria o quadro mais desagradável.

A ação A_2 envolve o elemento conforto de assistir o jogo pela TV, junto com amigos, mesmo sem as emoções de estar vendo o jogo no estádio. Como seu amigo é argentino, caso o Brasil ganhe seria muito bom para ele ver a reação do amigo tão de perto. Porém se a Argentina ganhar, ele sabe que seu amigo não o livrará de típicas gozações, o que para Edson será muito desagradável.

Finalmente, em caso de vitória do Brasil, a ação A_3 deixaria Edson com um enorme remorso e nunca iria se perdoar por não ter ido ao estádio ou mesmo na casa do amigo. Mas se a Argentina ganhasse não sofreria tanto quanto nos outros dois casos e ainda ficaria livre do vexame.

Com base nestas considerações, qual é a melhor ação a ser tomada?

A resposta dessa questão deve levar em conta cada uma das ações, com suas vantagens e desvantagens a serem pontuadas com um grau maior ou menor por Edson. Cada um dos quadros possíveis (ação e resultado do jogo) pode receber um valor, indicando grau de satisfação, ou custo em dinheiro, de modo que eles possam ser comparados em uma mesma escala.

A Tabela 3.3 mostra um exemplo de possíveis valores que Edson poderia atribuir a cada um dos quadros, onde valores negativos indicam situações desfavoráveis.

Ação / Natureza	A_1	A_2	A_3
B	1000	600	-300
A	-500	-100	200

Figura 3.3: Valores da função de utilidade de Edson.

Assim, por exemplo, em uma escala de satisfação, o valor 1000 é o que Edson considera equivalente ao prazer de assistir o Brasil ganhar da Argentina no estádio. Estes valores definem a chamada função de utilidade de Edson, definida para cada uma das ações possíveis e para cada um dos estados da natureza.

Suponhamos agora que Edson atribua probabilidade 0,6 à vitória da seleção brasileira, e probabilidade 0,4 à derrota.

Com esta informação, podemos calcular o valor esperado da função de utilidade para cada uma das decisões possíveis. Denotemos por $X_i(\theta)$, o valor atribuído à decisão A_i se ocorrer o resultado θ, $\theta \in \{A, B\}$.

Para cada decisão, temos

$$\begin{aligned}
E(X_1) &= X_1(A) \times P(A) + X_1(B) \times P(B) \\
&= 1000 \times 0,6 + (-500) \times 0,4 = 400 \\
E(X_2) &= X_2(A) \times P(A) + X_1(B) \times P(B) \\
&= 600 \times 0,6 + (-100) \times 0,4 = 320 \\
E(X_3) &= X_3(A) \times P(A) + X_1(B) \times P(B) \\
&= (-300) \times 0,6 + 200 \times 0,4 = -100
\end{aligned}$$

Podemos considerar a melhor decisão como aquela que maximiza a utilidade esperada. De acordo com este critério, a decisão ótima que Edson pode tomar é A_1.

Referências Bibliográficas

[1] O.R. Bekman, P.L.O. Costa Neto (2002) *Análise Estatística da Decisão*. Ed. Edgard Blücher.

[2] D. Berry (1995) *Statistics: a Bayesian perspective*. Duxbury.

[3] G.E.P. Box, W.G. Hunter, J.S. Hunter (2005) *Statistics for Experimenters*. Wiley.

[4] W.G. Cochran (1953) *Sampling Techniques*. Wiley.

[5] H.A. David (1973) Waiting time paradoxes and order statistics, *J. Amer. Statist. Assoc.* 68, 743–745.

[6] L.G. Esteves, S. Wechsler, A. Simonis, C.M. Peixoto (2005) Indifference, Neutrality and Noninformativeness: Generalizing the Three Prisoners Paradox. *Synthese* 143, 255–272.

[7] D. Freedman, R. Pisani, R. Purves (1978) *Statistics*. Norton.

[8] W. Feller (1971) *An Introduction to Probability Theory and its Applications*. Vol. I, John Wiley & Sons.

[9] W. Feller (1991) *An Introduction to Probability Theory and its Applications*. Vol. II, John Wiley & Sons.

[10] L. Gonick, W. Smith (1993) *Cartoon Guide to Statistics*. Collins.

[11] Jogos de matemática (diversos) *http://www.mathpuzzle.com/*

[12] L. Kish (1995) *Survey Sampling*. Wiley.

[13] A.C. Morgado, Os dois bodes *Revista do Professor de Matemática* 33, p.26.

[14] F. Mosteller (1987) *Fifty Challenging Problems in Probability with Solutions*. Dover.

[15] E. Pegg (2005) Tournament dice. Disponível em *http://www.maa.org/editorial/mathgames/mathgames_07_11_05.html*

[16] I. Peterson (2000) *Math Trek - Aventuras no Parquemático*. Gradiva.

[17] L.L.R. Rifo (2007) A outra face da moeda honesta. *Revista do Professor de Matemática* 65, 5–7.

[18] S.M. Ross (2003) *Introduction to Probability Models*. Academic Press.

[19] R. Smullyan (1998) *O Enigma de Sherazade*. Jorge Zahar.

[20] G.W. Snedecor, W.G. Cochran (1989) *Statistical Methods*. Iowa State University Press.

[21] G.U. Yule, M.G. Kendall (1956) *An Introduction to the Theory of Statistics*.