Capítulo

2

Delineamentos Experimentais em Informática na Educação

Eduardo Henrique da Silva Aranha (UFRN),
Thiago Reis da Silva (IFMA)

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar as diretrizes que devem ser seguidas para o planejamento de experimentos. Ao final da leitura deste capítulo, você deve ser capaz de:

- Entender os princípios estatísticos que suportam o delineamento de experimentos.
- Identificar e avaliar os efeitos de fatores que não estão sob investigação.
- Analisar o impacto da seleção dos materiais experimentais e participantes do experimento.
- Delinear experimentos que controlam até dois fatores de ruído.
- Entender a diferença entre experimentos e quasi-experimentos.
- Projetar experimentos com tratamentos compostos por mais de um fator.

Era uma vez... *um pesquisador chamado Ronaldo que durante seu doutorado analisou as limitações dos formatos de videoaulas mais utilizados para ensino à distância. Durante seu trabalho, ele propôs um novo formato de videoaula e realizou alguns estudos de caso com resultados promissores. Empolgado com os resultados alcançados, Ronaldo apresentou sua proposta de doutorado com segurança, mas seus estudos foram bastante criticados pelos membros da banca de avaliação de sua proposta. Ele foi questionado se os resultados positivos eram realmente devido ao uso de sua proposta ou devido a outros fatores, como conhecimento prévio dos estudantes, diferenças entre a qualidade das aulas gravadas, entre outros. Por sorte não era a banca de defesa e Ronaldo teve tempo para pensar em novos estudos. Buscou ajuda em um livro de metodologia de pesquisa e encontrou no capítulo de delineamento experimental uma forma de melhor analisar a real causa dos efeitos observados. Seus experimentos suportaram cientificamente suas afirmações sobre os benefícios de seu trabalho de doutorado, que foi posteriormente defendido e bastante elogiado pela banca de avaliadores.*

1 Introdução

A experimentação é uma importante ferramenta para você utilizar durante suas pesquisas em tecnologias educacionais. Ao experimentar, você obtém evidências sobre os reais benefícios e limitações dessas tecnologias na educação. O modo de experimentar, entretanto, pode fazer uma grande diferença na validade dos resultados observados e na sua capacidade de generalização para situações semelhantes.

De fato, esta não é uma preocupação só sua ou da área de informática educacional, mas de todos os pesquisadores, independentemente de área de atuação, e que precisam obter evidências cientificamente válidas a partir da aplicação prática de uma ou mais tecnologias. Esse é um problema antigo que foi sistematizado pela estatística e aplicado em diferentes áreas, como agronomia, medicina, engenharia, psicologia, computação e mais especificamente, no nosso caso, na Informática Educacional (IE).

Nas próximas seções deste capítulo, você terá contato com diferentes situações e delineamentos experimentais, cujos conhecimentos envolvidos podem ser transpostos para diferentes situações a serem enfrentadas nas suas pesquisas. Para isso, faremos uso da situação problema já apresentada. Além disso, o foco será no planejamento do experimento e não na análise dos dados, uma vez que isto é tema dos capítulos de introdução à estatística descritiva e à inferência estatística. A leitura desses capítulos é bastante relevante, uma vez que experimentos controlados são estudos com análises fortemente quantitativas.

2 Terminologia e Princípios Básicos

Em primeiro lugar, se você pensa em realizar experimentos, isso quer dizer que você tem algo a ser experimentado. Na medicina, o objeto de estudo geralmente é um tratamento para uma doença. Já na agronomia, pode ser um tratamento para combater pragas. Assim, a palavra **tratamento** acabou sendo convencionada para representar o que está sendo investigado. Na área educacional, são tratamentos as ferramentas, técnicas e métodos utilizadas no processo de ensino e aprendizagem. Por exemplo, se você quiser investigar o efeito de diferentes formatos de aula à distância, tanto novos formatos propostos por você quanto os formatos atualmente utilizados na elaboração das aulas serão chamados de tratamentos.

Como você pode ver na Figura 1, ao experimentar uma tecnologia ela geralmente precisará de entradas (**materiais experimentais**) e usualmente de **participantes** (alunos, professores e outros). Se você quiser investigar formatos de aulas, uma possível entrada desse experimento é o conteúdo a ser ministrado usando cada formato investigado. Além disso, ao aplicar os tratamentos você observará resultados geralmente relacionados ao processo de ensino e aprendizagem. De acordo com Seltman (2018), esses resultados são representados por variáveis de saída chamadas de **variáveis dependentes** ou **de resposta**, e cujos valores quantificam ou qualificam aspectos como o esforço requerido para uso da tecnologia, a qualidade do aprendizado gerado ou o nível de engajamento observado. No caso de se investigar formatos de aula, podemos pensar em variáveis como tempo

requerido para completar uma aula, taxa de resolução com sucesso das tarefas da aula, qualidade das tarefas submetidas e taxa de evasão em aulas à distância.

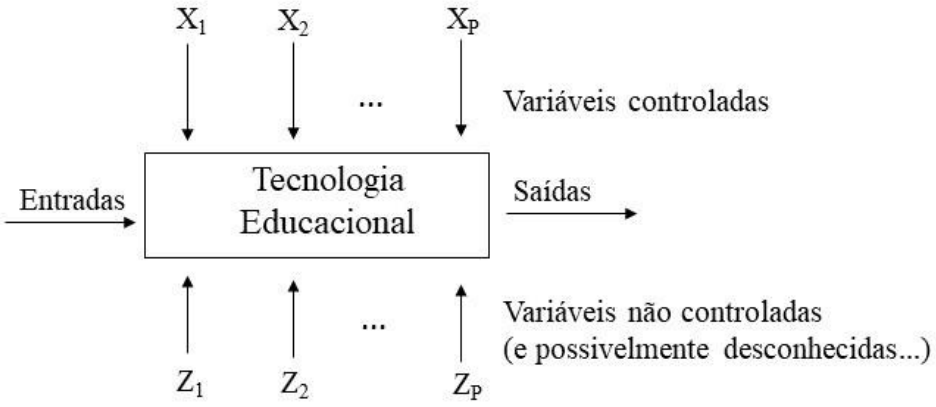


Figura 1. Visão geral de um experimento controlado na área de IE.

Ainda de acordo com Seltman (2018), os valores dessas variáveis de resposta (ou dependentes) são influenciados por um conjunto de variáveis manipuladas propositalmente (controladas) ou não pelo experimento, as quais são chamadas de **variáveis independentes** ou **explanatórias**. Uma dessas variáveis independentes se refere aos tratamentos investigados pelo experimento. As demais determinam o contexto de aplicação dos tratamentos investigados, ou seja, as condições nas quais os tratamentos foram aplicados e observados. No caso de se investigar diferentes formatos de aula, por exemplo, podemos ter variáveis independentes como a qualidade do material de aula produzido, a complexidade do assunto, e a experiência prévia dos alunos com o assunto da aula. Idealmente, você deve controlar no experimento o efeito das variáveis independentes que influenciem significativamente a variável de resposta. Na prática, porém, algumas dessas variáveis poderão ser de difícil controle ou até desconhecidas. A falta de controle sobre algumas destas variáveis não invalida necessariamente o experimento, mas cria ameaças, como discutido posteriormente.

As variáveis independentes são geralmente chamadas de **fatores**, sendo um desses o fator tratamento (ex.: formatos de aula investigado). Os demais fatores são **fatores de ruído**, ou seja, fatores com influência indesejada nas variáveis de saída do experimento. Os valores dos fatores são categóricos ou tornados categóricos (ex.: níveis baixos, médio e alto de complexidade de aula) e chamados de **níveis dos fatores**. Já o **efeito de um fator** é a influência causada na variável de resposta pela mudança de níveis do fator. Por exemplo, podemos observar um aumento ou redução do tempo de aprendizagem ao se trocar uma aula por outra com diferente nível de complexidade.

Um dos princípios básicos que você deve utilizar para se delinear um experimento é o **controle local**, também conhecido como **blocagem**. O controle local consiste em tentar evitar que o fator de ruído exerça influência sobre as variáveis de saída. Imagine estar investigando dois formatos de aulas distintos com dois alunos, um que é especialista no tema da aula e outro que tem muita dificuldade no assunto. Provavelmente os valores das variáveis de saída serão mais dependentes do fator de ruído (experiência do

participante) do que do fator tratamento (formato de aula utilizado).

Olhando este problema de outro ângulo, de repente o aprendizado com o formato A pode até ser melhor do que com B, mas o participante mais experiente sempre se sairá melhor no experimento, independente do tratamento utilizado. Isso quer dizer que o resultado do seu experimento pode estar comprometido, levando a conclusões erradas sobre a causa (tratamento ou experiência do participante?) do melhor ou pior desempenho observado. Para evitar esse tipo de situação, o controle local é aplicado de forma a:

a) Idealmente eliminar o efeito dos fatores de ruído, evitando que o fator exerça qualquer influência sobre as variáveis de saída;
b) Reduzir o efeito de um determinado fator de ruído, fazendo com que sua influência seja irrelevante quando comparado ao efeito do tratamento;
c) Diluir o efeito do ruído de forma balanceada, mantendo justa a comparação entre tratamentos.

Os itens b e c (reduzir e diluir os efeitos) nos leva à questão de que você não precisa controlar todos os fatores de ruído, mas apenas aqueles que têm efeito significativo na saída, favorecendo um tratamento em detrimento de outro. Além disso, mesmo que um determinado tratamento não seja favorecido (efeito diluído igualmente entre tratamentos), esse ruído pode atrapalhar a análise estatística de dados e ocultar, por exemplo, diferenças entre tratamentos que não sejam de grande magnitude.

Um segundo princípio básico dos experimentos é a **replicação** interna. Este tipo de replicação nada mais é que aplicar um mesmo tratamento a diferentes situações, mesmo que semelhantes, dentro de um mesmo experimento. Cuidado para não confundir esse conceito de replicação com o de replicação externa, onde todo o experimento é replicado (usualmente por outro pesquisador) visando observar se chega aos mesmos resultados. Por exemplo, ao se investigar diferentes formatos de aula, você pode aplicar cada formato de aula investigado a 50 alunos de uma determinada escola, configurando assim uma replicação interna de magnitude 50. Esse mesmo experimento pode ser replicado por você ou por ainda um outro pesquisador em outra escola, seguindo os mesmos procedimentos, configurando assim uma replicação externa, cujos resultados podem ser comparados com o do primeiro experimento.

Para experimentar um dado tratamento, você terá que aplicá-lo. Mas quantas vezes precisamos aplicar cada tratamento? Uma vez só? É pouco, certo? Cada aplicação de um tratamento é chamada de uma **réplica** (interna), e quanto mais réplicas temos por tratamento, maior tende a ser a precisão da análise estatística dos dados. Para se conseguir mais réplicas, você pode envolver um maior número de participantes para cada tratamento ou, em alguns casos, fazer com que um mesmo participante experimente mais de um tratamento. Por exemplo, ter um participante que assista uma aula no formato A e uma outra usando o formato B. Entretanto, você deve estar ciente de que existirá um **efeito de aprendizado**.

Esse efeito representa a mudança de comportamento do participante após ter tido contato com o primeiro tratamento aplicado. Ao aplicar o segundo tratamento, ele pode ter desempenho melhor por ter aprendido algo na primeira experiência, ou pode se chatear por ter que fazer algo similar e ter desempenho pior. Algumas estratégias podem ser

utilizadas para minimizar esse efeito, como: (i) dar um tempo entre a aplicação dos tratamentos para que o participante "esqueça" o que foi feito da primeira vez; (ii) colocar materiais diferentes, como temas de aulas diferentes, para que o aprendizado anterior não seja tão relevante para a aplicação do segundo tratamento; entre outras possibilidades.

Além do controle local e replicação, o terceiro e provavelmente mais importante princípio é o de **aleatorização**. O delineamento experimental trata da alocação dos tratamentos investigados ao material experimental e participantes do experimento. Essa alocação em um experimento controlado deve ser aleatória, e suportada por software específico, como JMP, SAS e SPSS, ou programações específicas em R ou Python. Essa aleatorização é condição fundamental para você fazer uma análise de causa-efeito. Em outras palavras, a aleatorização reduzirá significativamente as chances de que o resultado observado tenha sido causado pelo acaso e não devido aos efeitos do tratamento. Por exemplo, considerando 20 participantes experientes e 20 inexperientes, em torno de 0,13% será a chance de se ter um sorteio onde o tratamento A seja alocado a todos os participantes experientes, e o tratamento B a todos os inexperientes. Embora possível, é muito improvável dessa situação acontecer, não concorda?

Quando você não controla fatores de ruído que possuem efeitos significativos na variável de resposta em relação ao efeito do fator tratamento, o mais comum de acontecer é seu **experimento não detectar significância** no efeito do tratamento. Por exemplo, imagine o caso de você investigar o efeito de se usar um ou outro formato de aula aplicando-os em estudantes com grandes diferenças em termos de experiência prévia sobre o tema. Provavelmente a aleatorização alocará o formato A para ser usado com estudantes experientes e inexperientes, assim como acontecerá com o formato B. Na análise estatística dos dados, um dos métodos comumente utilizados é a Análise de Variância (ANOVA) (FREUND *at. al*., 2010), a qual verifica se a mudança do formato A para o formato B causa um aumento ou redução no aprendizado do estudante. Se possível, olhe agora os capítulo de introdução à inferência estatística para você ter uma visão formal sobre esses tipos de análises.

Testes de hipóteses

Experimentos controlados são em geral fortemente embasados em análises quantitativas. Um dos principais tipos de análise é a realização de testes de hipóteses. De maneira geral, uma hipótese estatística é uma afirmação sobre a população dos dados testada a partir de uma amostra. A ANOVA é um teste estatístico que pode verificar se a média da variável de resposta é significativamente afetada pelo fator tratamento. Em caso positivo, significa dizer que faz diferença entre usar um tratamento ou outro de acordo com a variável de resposta considerada. Para mais detalhes sobre testes de hipóteses, consulte o capítulo de introdução à inferência estatística.

A experiência prévia alta ou baixa do aluno no tema poderá estar anulando ou amplificando o efeito do tratamento, respectivamente, reduzindo a precisão no cálculo do efeito do tratamento no aprendizado do estudante. Isto porque alunos com alto conhecimento terão pouco para aprender, e com baixo conhecimento poderão aprender muito. Entretanto, se a quantidade de participantes for aumentada, ou seja, se você tiver uma maior quantidade de réplicas, a precisão da análise estatística inferencial aumenta e

talvez você consiga detectar o efeito do tratamento, mesmo sem ter controlado os fatores de ruído. Por isso, se seu experimento não indicar efeito estatisticamente significativo para o fator tratamento, não quer dizer que ele necessariamente não tenha efeito. Verifique a possibilidade de fatores de ruído terem influenciado o resultado e, se for o caso, recomende a realização de novos estudos controlando esses fatores e assim deixando mais claro para a análise estatística dos dados o efeito do tratamento.

Um caso particular é quando você tem um fator de ruído que se confunde com o fator tratamento. Este fator de ruído, chamado nesse caso de **fator de confundimento**, tem seu efeito distribuído de forma não aleatória aos grupos que recebem os diferentes tipos de tratamento. Por exemplo, se você investigar dois formatos diferentes de aula, sendo as aulas que seguem o formato A foram feitas por um ótimo professor e as do formato B foram feitas por um professor sem devida experiência, o efeito de mudar de A para B não é mais só o de mudar de formato de aula, mas também o de mudar de um professor exemplar para um sem experiência. Como o efeito do fator tratamento (formato de aula) foi combinado (confundido) com o fator experiência do professor, não será possível analisar de forma individual esses efeitos na análise estatística dos dados. Em resumo, não será possível saber se o resultado do experimento foi devido ao efeito do formato de aula, da experiência do professor, ou da interação entre os dois fatores (efeitos amplificados ou anulados).

Por fim, é importante você como pesquisador estar ciente de que os resultados de um experimento são restritos pelo contexto ao qual os tratamentos foram aplicados. Com técnicas de inferência estatística, porém, podemos generalizar os resultados, mas apenas para cenários semelhantes (por exemplo, pessoas com perfis semelhantes e temas de aula com complexidades semelhantes).

3 Seleção de Materiais Experimentais e Participantes

A seleção dos participantes de um experimento (ex.: alunos do ensino médio) e a seleção dos materiais a serem utilizados como entrada para os tratamentos investigados (ex.: temas de aulas) impactarão os resultados geralmente de duas formas:

- Inserindo ou controlando fatores de ruído;
- Limitando ou expandindo a capacidade de generalização dos resultados.

Isto porque se você usa material heterogêneo como entrada ou participantes com de diferentes perfis, essas diferenças poderão influenciar de forma indesejada o resultado do experimento. Por exemplo, a não ser que você queira investigar o efeito que o tema de estudo possua sobre o processo de aprendizado do aluno, inserir diferentes temas em um experimento estará em geral diminuindo a precisão de um experimento que quer avaliar apenas o efeito do formato de aula aplicado.

Apesar disso, muitas vezes você optará por ter maior diversidade nas entradas e perfil dos participantes, principalmente pelos seguintes motivos:

- Quantidade insuficiente de materiais experimentais homogêneos e de participantes com mesmo perfil;

- Necessidade de generalizar os resultados para uma maior diversidade de situações, uma vez que o efeito de determinados tratamentos pode mudar de acordo com outras variáveis de contexto (ex.: formato de aula que funciona bem com alunos engajados, mas não com alunos desmotivados). Isto se não for possível rodar vários experimentos, um para cada contexto diferente.

Em resumo, o controle de fatores de ruído pode ser feito no próprio processo de seleção dos participantes e do material experimental a ser utilizado. Se em um experimento que avalia o formato de aula você usar apenas um tema de estudo, estará eliminando o efeito que diferentes temas teriam sobre as variáveis de resposta do experimento. Caso selecione mais de um tema, mas com complexidades de aprendizado semelhantes, reduzirá o efeito deste fator de ruído de complexidade do tema. Por fim, caso você selecione temas com complexidades bem diferentes de aprendizado, inserirá um fator de ruído no experimento.

Mas é normal um experimento possuir fatores de ruído. O que você precisa observar é:

- Os efeitos desses ruídos são significativos comparados aos efeitos dos tratamentos?
- Esses efeitos podem ser controlados pelas técnicas estatísticas de delineamentos experimentais?

Existem alguns padrões da estatística utilizados para delinear experimentos de forma a controlar o efeito de alguns fatores de ruído. Isso é o que você vai ver nas próximas seções.

4 Delineamento Experimental

O delineamento experimental do ponto de vista estatístico está bastante relacionado ao processo de alocação dos tratamentos às unidades experimentais. Você pode considerar que uma **unidade experimental** representa aquilo que receberá o tratamento. Na medicina, por exemplo, a unidade experimental geralmente é o paciente. Na agronomia, pode ser uma faixa de terra, ou uma planta. Na informática educacional, pode ser um aluno que assistirá uma videoaula que segue determinado formato ou método pedagógico (tratamento), ou uma combinação de fatores, como o aluno e o tema de aula explorado.

A definição das unidades experimentais e da forma de alocação dos tratamentos a elas têm grande importância no experimento, pois pode controlar ou não fatores de ruído, além de determinar a forma na qual os dados são analisados. Você pode fazer esse delineamento experimental de diferentes maneiras, pois cada experimento pode requerer um delineamento diferente. Entretanto, assim como acontece na computação, você perceberá que existem soluções de delineamentos experimentais (ou planos experimentais) que se repetem, inclusive nas diferentes áreas de conhecimento (ex.: medicina, agronomia e computação). Essas soluções são planos experimentais que historicamente se mostraram eficientes no projeto de um grande número de experimentos

diferentes, independentes de área de conhecimento. Veja a seguir os planos mais comuns, e que são simples e bastante aplicáveis para o contexto da informática na educação.

4.1 Experimentos Completamente Aleatorizados

O delineamento experimental **completamente aleatorizado** considera que não há fatores de ruído relevantes para serem controlados. Isto porque esse plano não aplica nenhuma restrição na organização das unidades experimentais, as quais iremos daqui em diante chamar simplesmente de participantes. Imagine a situação onde você tenha 21 alunos à disposição para participar do experimento e avaliar diferentes formatos de videoaulas. Cada participante atenderá a uma videoaula elaborada segundo um dos três formatos de videoaula (A, B ou C) sob investigação. Para cada participante, de maneira individual, você sorteia qual formato de aula o aluno terá acesso. Fazendo desta forma, você está utilizando um delineamento completamente aleatorizado. Uma ilustração dessa aleatorização é apresentada na Figura 2, onde cada cor representa um formato de aula diferente.

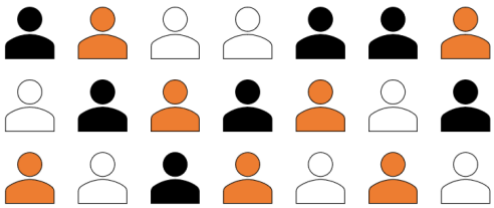


Figura 2. Aleatorização dos tratamentos (preto = A, laranja = B, branco = C).

Em geral, você como planejador do experimento irá elaborar uma planilha como a da Tabela 1, que mostra esse mesmo resultado de aleatorização da alocação dos tratamentos aos participantes.

Tabela 1. Alocação aleatória dos tratamentos.

Aluno	Trat.	Aluno	Trat.	Aluno	Trat.	Aluno	Trat.	Aluno	Trat.
1	A	6	A	11	A	16	C	21	C
2	B	7	B	12	B	17	A		
3	C	8	C	13	C	18	B		
4	C	9	A	14	A	19	C		
5	A	10	B	15	B	20	B		

Note que esse experimento é balanceado, ou seja, todos os tratamentos têm o

mesmo número de réplicas. Isto facilita a análise estatística dos resultados, que avaliará a significância do efeito dos tratamentos em relação ao efeito dos demais fatores de ruído (erro experimental). E se o efeito do tratamento for similar ou menor que o efeito do erro experimental, então ele também não é significativo, concorda?

Note que ao usar um plano experimental completamente aleatorizado, é como se você estivesse dizendo que o efeito dos fatores de ruído não é significativo ao ponto de precisarem ser controlados. Se existem fatores de ruído com efeitos significativos, eles devem ser considerados e controlados pelo delineamento do experimento. Caso contrário, a precisão dos cálculos de análise pode não ser suficiente para detectar efeitos significativos de tratamentos, a não ser talvez se você tiver uma quantidade de réplicas bem maior do que seria necessário se tivesse controlado os fatores de ruído. Aumentar a quantidade de réplicas aumenta a precisão da análise dos dados, mas não supre totalmente a falta de controle de fatores de ruído em um experimento.

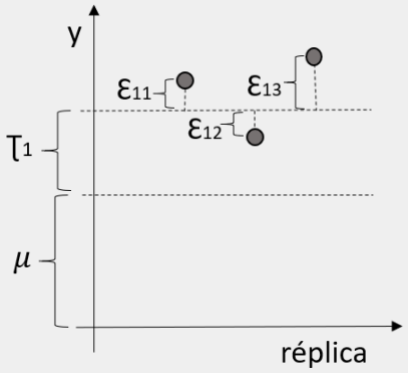
Efeitos considerados no delineamento completamente aleatorizado

O seguinte modelo linear explica a relação entre a variável de resposta (y) e as variáveis preditoras em um experimento completamente aleatorizado:

$$y_{ir} = \mu + \tau_i + \varepsilon_{ir}$$

Observe que o plano experimental explicado pela equação considera que o valor coletado na r-ésima aplicação (réplica) do i-ésimo tratamento investigado (y_{ir}) é determinado por três variáveis preditoras: uma constante μ; o efeito proporcionado pela aplicação do i-ésimo tratamento (τ_i); um erro experimental ε_{ir}, que representa o efeito conjunto de todos os fatores de ruído não controlados e atuantes (experiência do aluno, complexidade do tema da aula, entre outros) na r-ésima aplicação do i-ésimo tratamento.

Para você entender melhor essa decomposição do valor de cada valor y coletado, observe a ilustração a seguir, onde os valores relacionados a três aplicações (réplicas) do tratamento 1 são apresentados no gráfico. Note que a causa da variação do valor entre as três réplicas é atribuída ao erro experimental, já que tanto o valor médio μ e o efeito do tratamento 1 (τ_1) são valores fixos calculados na análise estatística.



Perceba também que nesta ilustração os valores de influência dos erros

experimentais ε_{12}, ε_{12}, e ε_{13} são menores que o efeito de se usar o tratamento 1. Se este padrão se mantiver em relação às demais réplicas do tratamento 1 e a do tratamento 2, a análise estatística deverá concluir que o fator tratamento possui um efeito estatisticamente significativo. Ou seja, faz diferença entre usar o tratamento 1, o tratamento 2, ou o tratamento N investigado.

4.2 Controlando um Fator de Ruído

Na seção anterior, você viu que delinear um experimento completamente aleatorizado não restringe a forma de alocação dos tratamentos aos participantes do experimento. Tal procedimento não faz nenhum controle específico sobre fatores de ruído, e é apropriado quando as unidades experimentais (nos exemplos aqui discutidos, participantes) são similares, homogêneas. Entretanto, nem sempre será possível você obter unidades experimentais com essa característica.

Imagine novamente a situação onde você tenha 21 alunos à disposição para avaliarem diferentes formatos de videoaulas. Considere também que todas as aulas tratarão sobre o mesmo tema e serão gravadas com a mesma qualidade. Entretanto, considere ainda que os 21 participantes têm diferentes níveis de conhecimento prévio sobre o tema das aulas. Percebe que essa diferença de experiência é um fator de ruído que pode ter efeito significativo no resultado do experimento? Mas se você tiver como identificar e agrupar os estudantes de acordo com seus níveis de experiência, esse fator pode ser controlado pela própria forma de delineamento experimental.

Considere que após você realizar um pré-teste cada participante possa ser classificado como experiente ou inexperiente. Digamos que 9 participantes sejam classificados como experientes e 12 como inexperientes. Uma primeira solução para controlar o efeito do fator experiência seria rodar dois experimentos completamente aleatorizados, um apenas com os experientes e outro apenas com os inexperientes. Dessa forma, em cada experimento o fator de ruído experiência foi controlado pelo processo seletivo dos participantes. Entretanto, os dois experimentos terão uma menor quantidade de réplicas/participantes, o que pode ameaçar a relevância do experimento.

Como alternativa, você pode rodar apenas um experimento seguindo o delineamento **aleatorizado em blocos completos**. Nesse tipo de experimento, um bloco é um conjunto de unidades experimentais homogêneas. No exemplo dos 21 estudantes, existem dois blocos, um com os participantes experientes e outro com os inexperientes. Em cada bloco, o nível de conhecimento prévio dos participantes é similar. A alocação dos tratamentos nesse caso é feita por bloco. Por exemplo, no bloco com 9 participantes experientes os tratamentos (formatos de videoaula) serão alocados de forma aleatória, mantendo balanceado: três réplicas para cada um dos três formatos de videoaulas investigados (A, B e C). Em seguida, de forma independente, procedimento similar é feito para o bloco de 12 participantes inexperientes, como ilustrado na Figura 3.

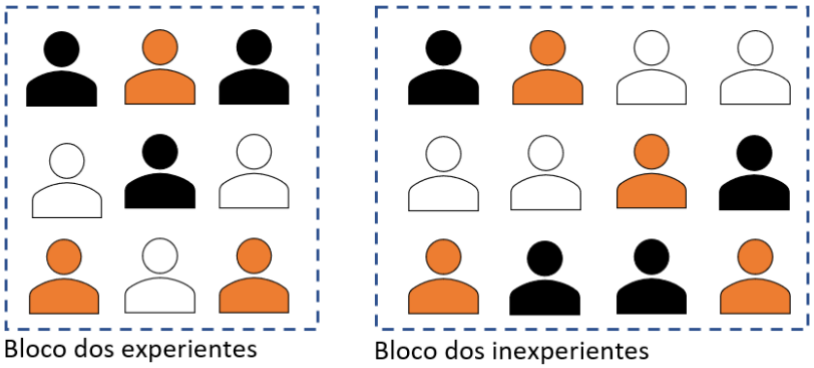


Figura 3. Aleatorização dos tratamentos feita de maneira individual em cada bloco.

Note que a quantidade de participantes por nível de conhecimento prévio (bloco) é um múltiplo da quantidade de tratamentos. Por exemplo, se você investigar três tratamentos diferentes, terá blocos com uma quantidade de participantes igual a 6, 9, 12, ou algum outro número maior e múltiplo de três, garantindo uma mesma quantidade de aplicações para cada tratamento investigado. Não é problema se o tamanho de cada bloco for diferente, como 9 no bloco dos experientes, e 12 no bloco dos inexperientes. O importante é que haja replicação por tratamento dentro de cada bloco.

E caso não seja possível o balanceamento dos tratamentos em cada bloco, como no caso de um bloco com 10 participantes experientes, um tratamento seria aplicado uma vez a mais que os outros dois. Entretanto, esse delineamento é chamado agora de aleatorizado em blocos incompletos e sua análise estatística de dados é um pouco diferente do usual. Para se manter como um experimento aleatorizado em blocos completos, você poderia rodar o experimento com apenas 9 dos 10 alunos disponíveis.

Assim como acontece no delineamento experimental completamente aleatorizado, aqui a análise estatística também avaliará a significância do efeito dos tratamentos em relação ao efeito do erro experimental. A diferença é que o efeito do fator bloco (experiência prévia do participante) será calculado e isolado, não inflacionando o valor atribuído ao erro experimental e assim deixando mais precisa a análise de significância do efeito dos tratamentos.

Efeitos considerados no delineamento aleatorizado em blocos completos

Você pode perceber como é considerada a relação entre variáveis de resposta e preditoras utilizando o seguinte modelo linear:

$$y_{ijr} = \mu + \beta_j + \tau_i + \varepsilon_{ijr}$$

Observe que a equação considera que o r-ésimo valor de resposta coletado para o i-ésimo tratamento e j-ésimo bloco (y_{ijr}) é determinado por quatro variáveis preditoras: uma constante μ; o efeito do j-ésimo bloco (β_j); o efeito do i-ésimo tratamento (τ_i), que por sua vez foi aplicado no j-ésimo bloco pela r-ésima vez; um erro experimental ε_{ijr}, que representa o efeito conjunto de todos os fatores de ruído atuantes naquela aplicação não controlados pelo bloco (pequenas diferenças entre conhecimentos prévios de alunos classificados em um mesmo bloco, alguma diferença

na qualidade de gravação das aulas, e outros).

Um detalhe que tem que ser observado é se existe interação entre bloco e tratamento. Isto acontece, por exemplo, quando um determinado formato de videoaula influencia positivamente no bloco de inexperientes e influencia negativamente no bloco de experientes. Em outras palavras, se for um formato de aula que facilita para novatos, mas que atrapalha quem já tem experiência. Esta verificação é feita através de um teste estatístico de não aditividade (FREUND *et. al*. 2010). Se esse teste constatar a interação, a análise dos dados deve ser feita baseada em uma equação que considere esse fator de interação ($\beta_j\tau_i$):

$$y_{ijr} = \mu + \beta_j + \tau_i + \beta_j\tau_i + \varepsilon_{ijr}$$

4.3 Controlando Dois Fatores

Em algumas situações, você vai querer controlar dois fatores de ruído. Na seção anterior, você considerou que todas as aulas do experimento sobre formato de videoaulas trataram sobre o mesmo tema e foram gravadas com o mesmo nível de qualidade. O fator controlado foi o nível de conhecimento prévio dos participantes sobre o tema das aulas.

Imagine então que agora você vai rodar seu experimento não só com estudantes com diferentes níveis de experiência, mas também com aulas que tratam de temas diferentes. Para controlar os fatores nível de experiência prévio do participante e o nível de complexidade do tema de aula, podemos fazer uso de um delineamento experimental chamado de **Quadrado Latino**. Nesse tipo de delineamento, os dois fatores controlados e o fator tratamento são organizados em quadrados, como apresentado na Figura 4, que mostra um delineamento experimental que avalia apenas dois formatos de aula, A e B.

Os quadrados latinos são as partes cinzas da Figura 4. Cada quadrado é uma réplica dentro do experimento, como será discutido mais adiante. O conteúdo de cada uma das quatro células de um quadrado indica qual tratamento (formato de videoaula) será aplicado no contexto daquela célula. Esse contexto é determinado pelos índices de linha e coluna do quadrado. No caso da célula superior à esquerda do primeiro quadrado, ela indica que o formato A de videoaula será aplicado considerando o tema 1 de aula (índice da linha) e o aluno 1 como participante (índice da coluna).

	Aluno 1	**Aluno 2**		**Aluno 3**	**Aluno 4**		**Aluno 5**	**Aluno 6**
Tema 1	Formato A	Formato B	**Tema 1**	Formato B	Formato A	**Tema 1**	Formato A	Formato B
Tema 2	Formato B	Formato A	**Tema 2**	Formato A	Formato B	**Tema 2**	Formato B	Formato A

Figura 4. Organização de Quadrados Latinos de dimensão 2x2, fixando-se as linhas.

Você pode perceber através da Figura 4 que o aluno 1 testará os dois formatos de

aula, A e B, mas cada um com tema de aula diferente. Esta mudança é importante para reduzir efeitos de aprendizado. Já o aluno 2 também testará os dois formatos A e B, mas usando temas de aula de forma invertida em relação ao aluno 1. Essa inversão garante em cada quadrado que os tratamentos (formatos de aula A e B) sejam aplicados com dois alunos (fator de controle coluna) e com os dois temas de aula (fator de controle linha). Isto torna mais justa a comparação, evitando que um determinado formato de aula seja aplicado apenas com um aluno experiente ou com um tema de aula mais complexo.

Para usar um delineamento como esse, você precisará de várias réplicas de quadrados. Com isso, a quantidade de aplicações de tratamentos com quadrados latinos de dimensão 2 será sempre um número múltiplo de 4. Um detalhe importante: para replicar um quadrado, além de fazer um novo sorteio para alocação dos tratamentos, você deve mudar só os índices das linhas, só o das colunas, ou de ambos ao mesmo tempo. Por exemplo, veja que no segundo quadrado da Figura 4 tem-se os alunos 3 e 4, no terceiro quadrado os alunos 5 e 6, e assim por diante. Isto quer dizer que para cada réplica adicional do quadrado você precisará de dois novos participantes. A quantidade de réplicas dependerá então da quantidade de participantes que você terá à disposição.

Caso você prefira manter apenas dois alunos no experimento e mudar os temas de aula, isso pode ser feito também, mas cuidado redobrado com efeitos de aprendizado. Procure usar temas de aulas não relacionados ou espaçar os momentos de aula. Um terceiro cenário de replicação de quadrado é você mudar em cada réplica tanto os alunos como os temas de aula, o que pode reduzir um pouco a precisão da análise dos dados pelo aumento da variabilidade de cenários utilizados.

Sobre a dimensão do quadrado da Figura 4, ele é 2x2. Porém, o quadrado latino delineado em um experimento por ter dimensão 3x3, 4x4, ou de maneira geral NxN. Essa dimensão é determinada pela quantidade de tratamentos investigados no experimento. Isto porque a ideia é que cada tratamento seja aplicado em um nível diferente de cada um dos fatores controlados.

De fato, em um quadrado latino, não haverá repetição de tratamento tanto na linha como na coluna. Este processo de alocação do tratamento nas células do quadrado latino é aleatório e deve ser feito preferencialmente com o auxílio de ferramentas estatísticas que possuem funções próprias para isto. Para fazer essa aleatorização, comece sorteando o tratamento a ser aplicado na célula superior mais à esquerda. Se a dimensão do quadrado for 2x2, a alocação dos demais tratamentos é determinada automaticamente, já que não é possível repetir tratamentos em uma mesma linha ou coluna. Quando a dimensão do quadrado é maior, o procedimento utilizado no sorteio é mais elaborado e deve ser consultado em um livro de estatística (BOX *et. al.* 2005).

Efeitos considerados no delineamento Quadrado Latino

A decomposição dos efeitos considerados em um experimento do tipo quadrado latino é apresentada a seguir:

$$y_{ijkr} = \mu + \alpha_k + \beta_j + \tau_i + \varepsilon_{ijkr}$$

Essa equação é similar a do aleatorizado em blocos completos, adicionando-se

o efeito de um segundo bloco (α_k). Nesse caso, é necessário verificar a inexistência de interação entre o tratamento e cada um dos fatores controlados (linha e coluna do quadrado). A forma estatística de se analisar muda um pouco dependendo da maneira que se replica os quadrados latinos (mudança de linhas, colunas, ou linhas e colunas). Você pode ver mais detalhes sobre a forma de se analisar em livros estatísticos, como na obra de Box *et. al.*, (2005).

Você deve ter notado que no delineamento do quadrado latino, diferentemente do que acontece no experimento completamente aleatorizado, para cada participante são coletadas duas ou mais observações (dados). Essa é uma característica dos **experimentos com medidas repetidas**, nos quais um único recurso (ex.: participante) pode gerar várias observações, aumentando assim o poder estatístico da análise.

Em um experimento de maior duração, por exemplo, você poderia aplicar a uma mesma turma de alunos diferentes tratamentos ao longo do tempo (***crossover design***), ou aplicar um único tratamento, mas realizar várias medições do aprendizado ao longo do período letivo. Nesses tipos de experimentos, uma preocupação recorrente é a influência da ordem de aplicação dos tratamentos (efeitos de aprendizado). O quadrado latino é um bom tipo de delineamento para essa situação, quando pode-se controlar dois fatores, ou o aleatorizado em bloco completo, cada participante representando um bloco que receberá vários tratamentos ao longo do tempo.

4.4 Outros Delineamentos Experimentais

Neste capítulo, você teve contato com três dos mais comuns delineamentos experimentais utilizados na prática, não só na área de informática educacional, mas de maneira geral em todas as áreas de pesquisa. Porém, assim como acontece nas outras áreas, você de informática educacional também poderá se deparar com situações específicas que requeiram delineamentos específicos para o contexto de sua pesquisa.

Nesses casos, o ideal é você consultar livros que apresentem outros tipos de delineamentos experimentais e verificar se algum deles se encaixa no tipo de delineamento que você precisa. São exemplos de outros delineamentos o plano em blocos incompletos balanceados (nem todos os tratamentos são aplicados em todos os blocos), o retângulo latino (quantidade de linhas diferente da de colunas), o quadrado greco-latino (controle de 3 fatores de ruído) e o quadrado hiper-greco-latino (controle de 4 fatores de ruído) (BOX *et. al.* 2005; KUEHL, 2000).

Caso seja algo bem específico, você também pode consultar estatísticos ou pesquisadores mais experientes para lhe ajudar a delinear seu experimento e analisá-lo com segurança.

5 Considerações sobre o tamanho da amostra

Você pode estar se perguntando quantas réplicas por tratamento você precisa ter em seus experimentos para que eles sejam válidos. Essa é uma pergunta que pode ser respondida através de cálculos estatísticos ou de acordo com suas restrições de recursos. Na prática, em geral o tempo que você tem disponível para rodar seus estudos e os recursos que você tem disponíveis determinarão a quantidade máxima de réplicas que você poderá ter nos seus experimentos.

Existem alguns aspectos que você pode estar observando. Quando maior a heterogeneidade nas unidades experimentais do seu experimento (efeitos não controlados), mais importante é considerar uma maior quantidade de réplicas. Isto reduzirá as chances de se ter resultados ao acaso (efeitos de fatores de ruído, entre outros.). Você pode obter mais informações sobre esse assunto no capítulo de inferência estatística.

Outro aspecto é rodar uma quantidade maior de experimentos menores ao invés de um único experimento grande. Pesquisa é um processo de aprendizagem, então pode ser interessante não usar todos os seus recursos em um único experimento, mas em uma sequência de experimentos na qual você tenha um aprendizado incremental e oportunidades de fazer ajustes nos experimentos seguintes a partir desse aprendizado. Entretanto, cuidado para ter uma quantidade de dados em cada um dos experimentos que seja suficiente para se realizar análises estatísticas e generalizar resultados.

Além disso, você pode trabalhar melhor seu material experimental para não desperdiçar oportunidades de se conseguir mais dados, e dados mais precisos, com quantidade similar de recursos. Por exemplo, se você for rodar um curso ao longo de uma semana, ao invés de sortear para um estudante um formato de aula para ser usado ao longo de todo o curso, você pode sortear um formato para cada dia de curso, mas apenas se as aulas de cada dia forem independentes das anteriores.

Análise de poder estatístico

Essa técnica de análise estatística pode ser utilizada para se estimar o quão grande o tamanho de sua amostra precisa ser para que sua análise tenha poder suficiente para detectar efeitos estatisticamente significativos. Entretanto, ela vai depender de suposições sobre a magnitude dos efeitos dos tratamentos e do desvio padrão, como pode ser visto nos métodos de análise estatística apresentados por (FREUND *et. al.*, 2010).

6 Tratamentos Fatoriais

O delineamento do seu experimento pode envolver não só o planejamento de como serão alocados os tratamentos às unidades experimentais, mas também a definição de quem são os tratamentos. Isto porque você pode estar querendo investigar tratamentos compostos. Para entender isso melhor, imagine que você supõe que o formato A de videoaula é mais apropriado para temas simples de aula, e o formato B para temas mais complexos. Em outras palavras, que o benefício do formato de aula depende das

características das aulas (complexidade do tema).

Note que nesse caso o efeito da complexidade do tema da aula não deve ser evitado, mas manipulado de forma a investigarmos seu efeito. Em outras palavras, um tratamento nesse experimento não é mais apenas o formato de videoaula, mas uma combinação do formato com a complexidade do tema. A combinação desses dois fatores forma o fator tratamento do experimento.

Se os níveis do fator formato de aula são A e B, e os níveis do fator complexidade do tema são Baixa e Alta, você pode dizer que seu experimento possui tratamento fatorial 2x2 (dois por dois). O fator tratamento possui então quatro valores, que são as quatro possíveis combinações de formatos de aula e complexidades de tema: (A, Baixa), (A, Alta), (B, Baixa), (B, Alta). Definido isso, o restante do experimento é planejado de forma similar a como se houvesse um fator tratamento simples.

Entretanto, na análise dos resultados do experimento você poderá observar não só o desempenho que se obtém com a aplicação de cada combinação específica, mas também o efeito individual de cada fator que compõe o tratamento e suas interações. Na Figura 5 a seguir, é possível ver dois cenários. No primeiro, à esquerda, você pode visualizar que o efeito de mudar de formato de aula de A para B quando a complexidade do tema de aula é alta (linha azul) ou quando é baixa (linha vermelha). As linhas estão praticamente paralelas, concorda? Independente da complexidade, a taxa de aprendizado (eixo Y) aumentará na mesma proporção, pois não há interação entre os fatores formato de aula e a complexidade de tema de aula.

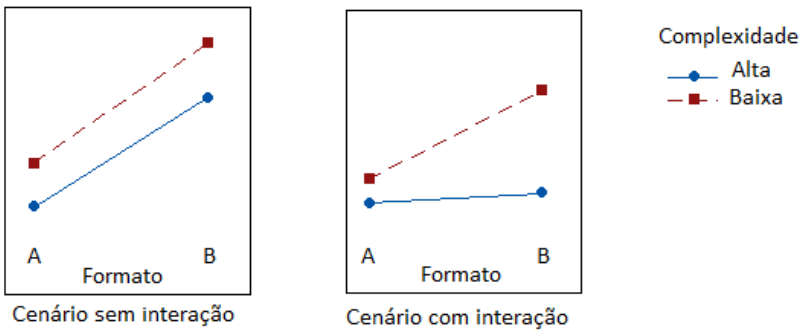


Figura 5. Fatores com e sem interação.

Já no segundo cenário, a linha azul está praticamente na horizontal. Isto significa que o efeito na taxa de aprendizado é quase inexistente quando se muda do formato A para o formato B com aulas de temas complexos. Por outro lado, o efeito é perceptível em se tratando de aulas com temas de complexidade baixa (linha vermelha). Mais detalhes podem ser vistos em Kuehl (2010).

7 Quasi-Experimentos

Ao delinear um experimento, você se preocupará em tentar garantir que os

tratamentos investigados sejam a real causa dos efeitos observados (resultados). A aplicação dos princípios de controle local, replicação e aleatorização lhe proporciona um suporte para isso. Entretanto, em alguns casos não é possível aleatorizar a alocação dos tratamentos às unidades experimentais. Os delineamentos que não aplicam a aleatorização, mas que possuem objetivo similar ao dos experimentos e que seguem princípios como controle local e replicação são chamados de quasi-experimentos (SHADISH *et. al.* 2002).

O não uso da aleatorização nos quasi-experimentos normalmente se dá por restrições no uso de recursos. Por exemplo, você pode querer comparar dois formatos de videoaulas coletando dados em cursos à distância que já fazem uso deles. Ou seja, provavelmente não tem como um determinado curso refazer seu material didático usando outro formato de videoaula apenas para que você realize seu experimento. Nem terá condições de solicitar que estudantes saiam de um curso e entrem em outro para que, como unidades experimentais, recebam um tratamento diferente definido pelo sorteio. Sem aleatorizar o tratamento, o quasi-experimento podem fazer uso de testes estatísticos como o ANOVA para tentar mostrar diferenças entre tratamentos, mas existe sempre a ameaça do resultado ter sido por algum outro fator que não o tratamento.

Para mitigar esse risco, ao delinear um quasi-experimento você deve analisar fatores não controlados para tentar garantir que não haja nenhuma alternativa plausível sobre a causa dos efeitos observados que não a do tratamento investigado.

8 Ameaças à validade

Ao delinear seu experimento, por mais cuidado que você tenha, sempre existirão ameaças à validade do resultado obtido. De fato, ao usar os delineamentos apresentados neste capítulo, várias ameaças são controladas, ou até eliminadas. Você pode avaliar as ameaças relacionadas ao seu experimento de acordo com o tipo dela. Dentre os tipos que podem ser encontrados na literatura, os seguintes são os mais aceitos e utilizados na academia:

Validade interna. As ameaças relacionadas à validade interna dizem questão ao risco de se estar observando efeitos nas variáveis de resposta que sejam atribuídos ao fator tratamento, sendo na verdade efeitos produzidos por outros fatores não controlados. Não controlar um determinado fator de ruído ou realizar um controle inadequado são situações que levam a geração de ameaças à validade interna. De maneira mais concreta, são exemplos de situações que geram essas ameaças:

- A coleta de dados através de instrumentos de baixa precisão (participante ser responsável por registrar dados, sendo que ele pode se equivocar);
- Não controlar fatores relevantes como a experiência dos participantes, capacidade de aprender (interação seleção-testagem) durante o experimento, e resistência física para participar do experimento (seleção-maturação);
- Abandono desbalanceado ou com viés de participantes (mortalidade seletiva), levando a se coletar dados apenas de grupos específicos;
- Comunicação entre participantes alocados a diferentes tratamentos

(contaminação);

- Sentimento de competição, de desmoralização (ou desvalorização) e de compensação (prêmio oferecido), dependendo do tratamento alocado, que pode levar a mudança seletiva de comportamento dos participantes;
- Influência do pesquisador nos participantes, mesmo que não intencional, de crenças e suposições que possam afetar avaliação final realizada pelos participantes;
- Efeito placebo, onde o fato de ser observado faz com que o participante crie uma expectativa de melhora, mesmo quando recebido o tratamento de controle (método atual ao invés do melhorado).

Procedimentos de aleatorização podem ser utilizados na seleção de participantes, por exemplo, para se tentar minimizar os riscos de seleção tendenciosa ou que não represente adequadamente a população de origem, como no caso de se selecionar apenas os melhores alunos de uma determinada escola.

Validade externa. Os experimentos em geral têm como objetivo realizar análises de dados amostrais, mas que possam ser generalizados para situações semelhantes (população). Com isto, o processo de seleção de materiais experimentais e de participantes cria naturalmente uma ameaça à validade externa do experimento. Isto porque os resultados serão válidos apenas para situações semelhantes em termos de perfil de participantes e de material utilizado. No caso da educação, é possível realizar estudos na educação infantil e querer usar os resultados como evidências para afirmações sobre a educação no ensino médio? Teorias podem até ser geradas nesse contexto, mas demandam experimentos que pudessem confirmar os mesmos resultados considerando agora como participantes o público do ensino médio. O uso de ambientes artificiais que sejam muito diferentes dos ambientes do mundo real da educação também são ameaças à validade externa, uma vez que geram o risco dos resultados só se aplicarem para ambientes com essas características artificiais, e não na sala de aula vivenciada pelos alunos do mundo real.

De maneira geral, todo experimento terá ameaças e o importante é você como pesquisador reconhecê-las e reportá-las explicitamente, de forma que o cliente ou leitor de seus trabalhos não se sinta que está sendo enganado ou que você não se atentou a essas limitações. O fato de possuir ameaças também não invalida os resultados do experimento, apenas limita a relevância ou aplicabilidade dos resultados, a não ser que sejam ameaças como uma coleta equivocada de dados que não tenha como ser corrigida (ex.: bug em sistema de prova online com troca irrastreável de gabaritos entre participantes). Nesse caso drástico, seria realmente necessário refazer o experimento.

Para um maior aprofundamento sobre ameaças à validade, seja sobre as citadas aqui ou sobre outros tipos existentes, é recomendado que você realize a leitura de textos como (WAINER, 2012) e (WOHLIN *et. al.* 2012).

9 Cenário Ilustrativo

Esta seção apresenta um exemplo real de planejamento de experimento controlado

cujo objetivo é o de comparar diferentes formatos de videoaulas (**fator tratamento**) para o ensino de programação de jogos digitais. Os **tratamentos** investigados são os seguintes: Formato Tutor Virtual (FTV), baseado em um sistema proposto de tutoria virtual; Formato Videoaula Gravada (FVG), representando o formato tradicional de videoaula atualmente utilizado em um curso.

A análise será realizada observando-se o **efeito** de se mudar os **níveis do fator tratamento** (FTV e FVG) causado em **variáveis de resposta**, como eficiência (tempo requerido de aula) e efetividade (aprendizado do aluno). Para isso, aulas serão criadas usando os formatos de aula investigados, com a temática ensino de programação de dois jogos digitais (**materiais experimentais**) e aplicadas a estudantes do ensino médio (**participantes**). Além do fator tratamento, outras duas **variáveis explanatórias** precisam ser controladas nesse experimento: os conhecimentos prévios dos participantes e a complexidade de criação dos jogos. Pré-testes e outros tipos de análise poderiam ser utilizados para se verificar a real necessidade desse controle, ou até para efetuar algum tipo de controle. Para o cenário aqui ilustrado, utilizaremos o delineamento experimental **Quadrado Latino** para efetuar o controle desses dois **fatores de ruído**.

Considerando a disponibilidade de 21 alunos disponíveis, temos como criar 10 réplicas de quadrados. Dessa forma, um dos alunos participantes pode participar do experimento, mas não terá seus dados analisados pela ANOVA. A Figura 6 ilustra algumas réplicas desse delineamento. Note nas linhas dos quadrados o uso de dois jogos digitais, um estilo Arkanoid e outro do tipo Nave Estelar. Por questões logísticas, todos os participantes terão como primeira aula a baseada no jogo estilo Arkanoid, cada um utilizando o tratamento para o qual foi sorteado no processo de **aleatorização** de tratamentos executado para cada quadrado.

	João	**Pedro**
Arkanoid	FTV	FVG
Nave	FVG	FTV

	Maria	**Lucas**
Arkanoid	FVG	FTV
Nave	FTV	FVG

	Rodolfo	**Ricardo**
Arkanoid	FTV	FVG
Nave	FVG	FTV

Figura 6. *Layout* do *Design* Experimental.

Visando reduzir o **efeito de aprendizado** entre as aulas, o experimento conta com uma sessão inicial de treinamento na tecnologia utilizada para se criar os jogos. Essa sessão também trata de explanar os procedimentos a serem seguidos durante a execução do experimento.

10 Resumo

Este capítulo tratou do delineamento de experimentos controlados. Nesta seção, você poderá rapidamente revisar os principais conceitos trabalhados neste capítulo. São três os princípios básicos dos experimentos controlados: aleatorização, replicação e controle local. Quando não é possível realizar a aleatorização dos tratamentos às unidades experimentais, o estudo é chamado de quasi-experimento.

Todo experimento possui variáveis dependentes e independentes. As variáveis independentes são chamadas de fatores, sendo alguns destes indesejáveis (fatores de ruído ou confundimento). O fator sob investigação é chamado de tratamento e é aplicado nas unidades experimentais. Um tratamento é dito como fatorial se ele é composto por uma combinação de fatores diferentes investigados.

Existem vários tipos de delineamentos possíveis. Abordamos três deles neste capítulo: o completamente aleatorizado, o aleatorizado em blocos completos e o quadrado latino, controlando respectivamente nenhum, um ou dois fatores de ruído. Você precisa ter algumas preocupações, como o tamanho da amostra, e se haverá algum efeito de aprendizado entre aplicações.

Um resumo visual é apresentado na Figura 7, através de um mapa mental.

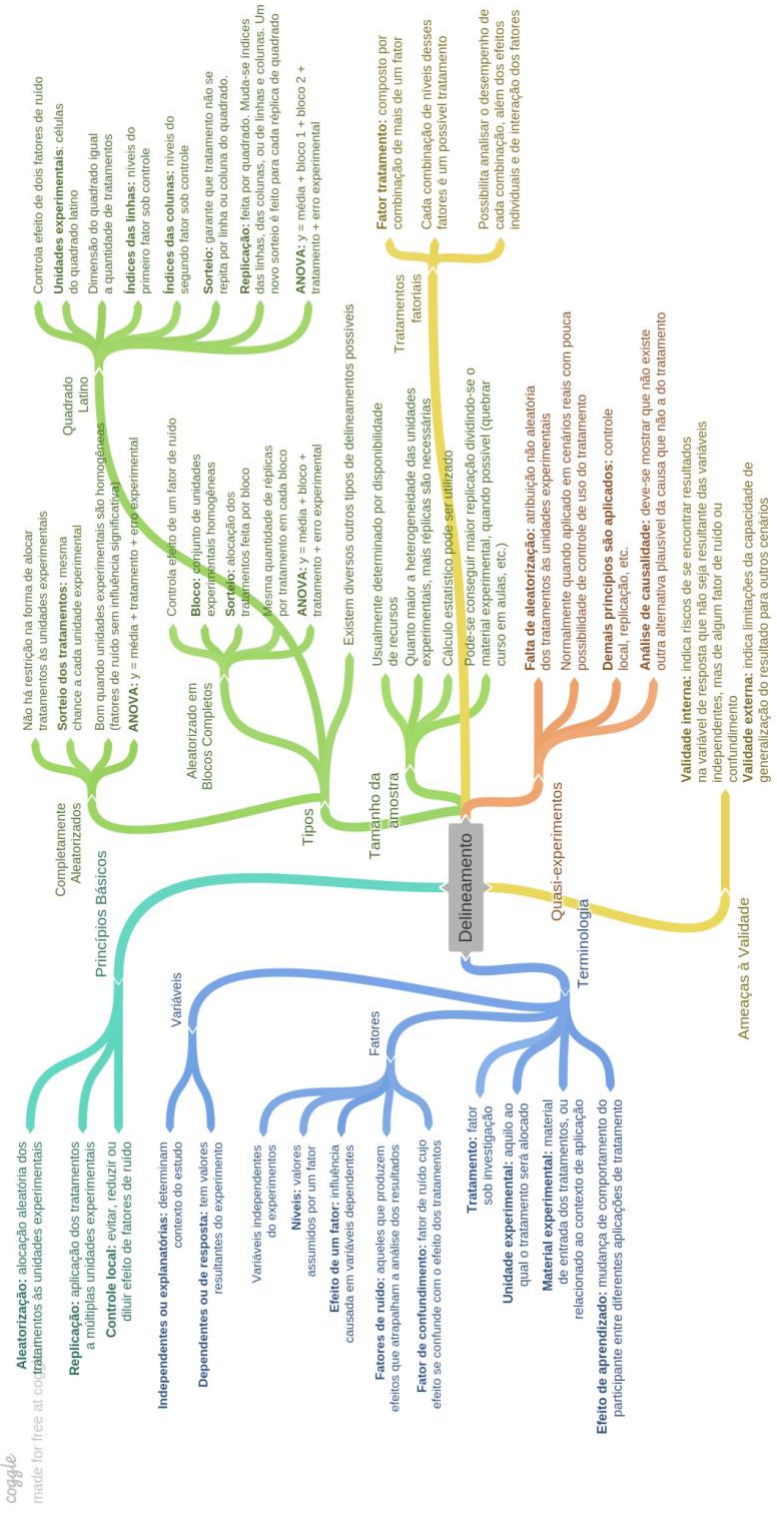


Figura 7. Mapa mental dos conceitos trabalhados neste capítulo.

11 Leituras Recomendadas

- **Statistics for Experimenters: Design, Innovation, and Discovery** (BOX *et. al.* 2005). Este livro de estatística contém vários exemplos práticos de delineamentos experimentais.
- **Basics of Software Engineering Experimentation**. (JURISTO, MORENO, 2001). Esta obra traz o tema de experimentação para software, sendo em parte uma adaptação da obra acima.
- **Experimentation in software engineering - an introduction. Springer, 2000.** (WOHLIN *et. al.* 2012). Nesse livro será possível ler um capítulo sobre experimentos controlados voltados para software.
- **Reporting Experiments in Software Engineering** (JEDLITSCHKA *et. al.* 2008). Este artigo apresenta os itens que devem ser considerados no planejamento e na escrita de artigos sobre experimentos controlados.
- **Statistical Methods** (FREUND *et. al.* 2010). Esse livro explica os métodos de análise estatística que você precisará para analisar os dados do seu experimento.
- **Experimento em sistemas colaborativos** (WAINER, 2012). Este capítulo de livro apresenta conceitos sobre como realizar um experimento para avaliar o uso de um sistema colaborativo em relação a outro existente, ou ao não uso dele.

12 Artigos exemplos

- **In-game assessments increase novice programmers' engagement and level completion speed.** (LEE, KO e KWAN, 2013). Artigo relata dois experimentos controlados utilizando o delineamento completamente aleatorizado para avaliar engajamento e tempo para estudantes completarem tarefas em um jogo com e sem avaliação explícita de aprendizado.
- **An experimental evaluation of scaffolded educational games design for programming** (JANTAN e ALJUNID, 2012). Artigo sobre experimento controlado utilizando o delineamento experimental Completamente Aleatorizado.
- **Avaliação Empírica da Utilização de um Jogo para Auxiliar a Aprendizagem de Programação** (JESUS; RAABE, 2010). Artigo relata um quase-experimento (não houve aleatorização na alocação dos tratamentos aos participantes) utilizando três turmas de alunos para avaliar os benefícios de um jogo educacional, sendo uma delas usada como grupo de controle.
- **Investigating Video Classes Formats for Teaching Digital Game Programming in High School** (SILVA; SANTANA; ARANHA, 2019). Artigo apresenta aplicação do delineamento experimental Quadrado Latino com estudantes do ensino técnico para avaliar formatos de videoaulas.

13 Checklist

Resumidamente, para delinear um experimento controlado você terá que realizar as seguintes atividades:

- Identificar o objetivo do experimento e os tratamentos a serem investigados. Caso seja um tratamento fatorial, identificar as combinações a serem investigadas;
- Identificar procedimentos e necessidades do experimento em termos de materiais experimentais e participantes;
- Identificar variáveis dependentes ou de resposta, ou seja, aquelas cujos valores serão coletados ao longo do experimento e usados como evidência acerca de possíveis efeitos causados pelo fator tratamento;
- Identificar variáveis independentes ou explanatórias, ou seja, aquelas que podem influenciar nos valores das variáveis de resposta do experimento. Identificar quais delas são fatores de ruído ou de confundimento que devem ser controlados para não comprometer os resultados do experimento;
- Delinear o experimento, escolhendo um plano experimental existente ou aplicando conceitos como os de controle local, replicação e aleatorização para definir a configuração do experimento;
- Analisar as ameaças à validade do experimento, identificando a adequação do delineamento realizado ou necessidade de ajuste voltando-se ao passo anterior.

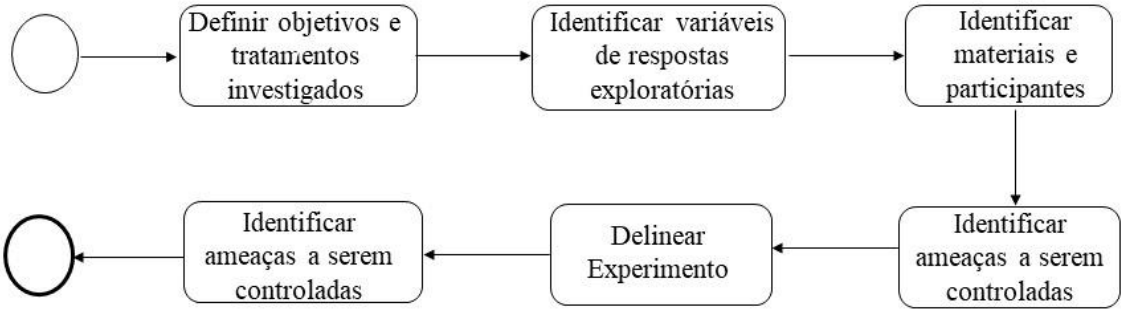


Figura 8. Atividades a serem realizadas para o delineamento de um experimento.

14 Referências

BOX, G.; HUNTER, J.; HUNTER, W. **Statistics for Experimenters: Design, Innovation, and Discovery**, 2nd Edition.Wiley-Interscience, 2005.

FREUND, R. J.; MOHR, D.; WILSON, W. J. **Statistical Methods**, 3 edição, Elsevier, 2010.

JANTAN, S. R.; Aljunid, S. A. **An experimental evaluation of scaffolded educational games design for programming**. In: IEEE Conference on Open Systems, Kuala Lumpur, 2012, pp. 1-6.

JURISTO, N.; MORENO, A. **Basics of Software Engineering Experimentation**. Springer Publishing Company, Incorporated, 2010.

JEDLITSCHKA, A.; CIOLKOWSKI, M.; PFAHL, D. Reporting Experiments in Software Engineering. In: SHULL, F.; SINGER, J.; SJØBERG, D. I. K. (eds) **Guide to Advanced Empirical Software Engineering**. Springer, London, 2008.

JESUS, E. A.; RAABE, A. L. A. Avaliação Empírica da Utilização de um Jogo para Auxiliar a Aprendizagem de Programação. In: XXI Simpósio Brasileiro de Informática na Educação, João Pessoa - PB, 2010.

KUEHL, R. Design of Experiments: Statistical Principles of Research Design and Analysis. 2nd Edition, Duxbury Press, 1999.

LEE, M. J.; KO, A. J.; KWAN, I. **In-game assessments increase novice programmers' engagement and level completion speed**. In: Proceedings of the ninth annual international ACM conference on International computing education research (ICER '13), 2013.

SELTMAN, H. J. **Experimental Design and Analysis**. 2018. Online at: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

SILVA, T. R.; SANTANA, A. O.; ARANHA, E. H. S. **Investigating Video Classes Formats for Teaching Digital Game Programming in High School**. In: Simpósio Brasileiro de Informática na Educação - SBIE, Brasília - DF, 2019.

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. **Experimental and quasi-experimental designs for generalized causal inference**. 2a edição, Houghton Mifflin, Boston, 2002.

WAINER, J. **Experimento em sistemas colaborativos.** In: PIMENTEL, M.; FUKS, H. (Org.). Sistemas Colaborativos. Rio de Janeiro, Brazil: Elsevier, 2012. p. 405–432.

WOHLIN, C.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; RUNESON, P.; WESSLÉN, A. **Experimentation in software engineering - an introduction**. Springer, 2000.

15 Exercícios

1. Considere que você foi alocado para delinear um experimento para verificar o efeito da complexidade do tema da aula no aprendizado dos alunos de um determinado curso. Considere ainda que esses temas são classificados como de baixa, média e alta complexidade, e que serão aplicados pré e pós testes para se avaliar o aprendizado antes e depois da aplicação do tratamento. Identifique:

a) Possíveis fatores de ruído ou confundimento.
b) Possíveis delineamentos experimentais, controlando nenhum, um ou até dois fatores de ruídos identificados.
c) A classificação: experimento ou quasi-experimento.

2. Como você ajustaria os possíveis delineamentos do experimento da questão

anterior para investigar não só o efeito da complexidade do tema, mas também a metodologia de aula do professor, de forma conjunta?

3. Considere que você tem acesso às turmas do ensino fundamental de 10 escolas, sendo que metade delas aplica uma determinada metodologia para ensinar pensamento computacional, e a outra metade faz uso de outra metodologia. Considerando que você quer comparar essas metodologias de acordo com o aprendizado proporcionado, medido por pré e pós-testes reconhecidos pela comunidade acadêmica, identifique:

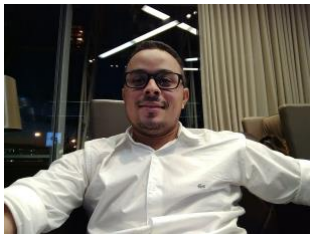
a) Possíveis fatores de ruído ou confundimento.
b) Possíveis delineamentos experimentais, controlando nenhum, um ou até dois fatores de ruídos identificados.
c) A classificação: experimento ou quasi-experimento.

Sobre os autores

Eduardo Henrique da Silva Aranha

http://lattes.cnpq.br/9520477461031645

Doutor (2009) em Ciência da Computação pelo Centro de Informática da Universidade Federal de Pernambuco (CIn/UFPE), possui graduação (1999) e mestrado (2002) também em Ciência da Computação pela Universidade Federal de Pernambuco. Tem experiência como professor no ensino superior desde 2002, bem como experiência na gerência e desenvolvimento de software desde 1997. Atualmente é professor do Departamento de Informática e Matemática Aplicada da UFRN e coordena o Laboratório de Pesquisa em Games e Educação do Instituto Metrópole Digital da UFRN.



Thiago Reis da Silva

http://lattes.cnpq.br/9776112478293682

Doutor (2017) em Ciência da Computação pela Universidade Federal do Rio Grande do Norte (UFRN), Mestre (2012) em Ciência da Computação pela Universidade do Estado do Rio Grande do Norte (UERN) e Graduado (2010) em Sistemas de Informação pela Universidade Federal do Piauí (UFPI). É integrante do Laboratório de Pesquisa em Games e Educação do Instituto Metrópole Digital da UFRN, onde pesquisa sobre Jogos Digitais, Ensino de Programação e Engenharia de Software. Atualmente é Professor do Instituto Federal de Educação, Ciência e Tecnologia do Maranhão - IFMA.