

ME 110 - Noções de Estatística

Nancy Lopes Garcia, Sala 209 - IMECC

nancy@ime.unicamp.br, www.ime.unicamp.br/~nancy

1 Processo de Aprendizagem

Todo processo de aprendizagem consiste em formular uma hipótese (H_1), deduzir as conseqüências desta hipótese a partir de *dedução*, comparar estas conseqüências com dados obtidos (experimentação ou pesquisa), se houver discrepâncias, utilizar os dados para formular novas hipóteses num processo de *indução* e assim por diante.

Exemplo 1.1 *Uma professora de escola primária pergunta às crianças o que acontecerá se ela misturar tinta guache das cores amarela e azul. Uma das crianças disse que H_1 : a mistura ficaria manchada. Quando a professora mistura as duas cores, obtém-se verde (dados), com isso as crianças aprendem.*

Exemplo 1.2 *As crianças querem obter a cor laranja mas somente possuem as cores azul, amarelo e vermelho.*

- H_1 : misturando-se azul e amarelo obtém-se laranja.
- Experimento 1: misturam-se as tintas de cores azul e amarelo.
- Dados: Obtém-se a cor verde.
- H_2 : misturando-se azul e vermelho obtém-se laranja.
- Experimento 2: misturam-se as tintas de cores azul e vermelho.
- Dados: Obtém-se a cor roxa.
- H_3 : misturando-se vermelho e amarelo obtém-se laranja.
- Experimento 3: misturam-se as tintas de cores vermelho e amarelo.
- Dados: Obtém-se a cor laranja.

Exemplo 1.3 *Um biólogo decide verificar as conseqüências de choques térmicos em um dos vetores de doenças de Chagas conhecido por *Panstrongylus megistus*.*

- H_1 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 35⁰ Celsius.
- Experimento 1: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 35⁰ Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.

- Dados: No final de 30 dias, morreram 15 insetos do grupo tratamento e 13 insetos do grupo controle.
- H_2 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 40^0 Celsius.
- Experimento 2: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 40^0 Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram 22 insetos do grupo tratamento e 15 insetos do grupo controle.
- H_3 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 45^0 Celsius.
- Experimento 3: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 45^0 Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram todos insetos do grupo tratamento e 11 insetos do grupo controle.

Exemplo 1.4 *Um biólogo decide verificar as conseqüências de choques térmicos em um dos vetores de doenças de Chagas conhecido por *Panstrongylus megistus*.*

- H_1 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 35^0 , 40^0 e 45^0 Celsius.
- Experimento 1: O biólogo toma 200 espécimes adultos e submete 50 deles a um choque de 35^0 Celsius (grupo tratamento 1), 50 deles a um choque de 40^0 Celsius (grupo tratamento 2), 50 deles a um choque de 45^0 Celsius (grupo tratamento 3) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram 14 insetos do grupo tratamento 1, 17 insetos no grupo tratamento 2, 50 insetos no grupo tratamento 3 e 13 insetos do grupo controle.

Exemplo 1.5 *A Data Folha quer determinar qual a popularidade do presidente Lula. Neste caso são entrevistadas 2000 pessoas em diversas cidades do Brasil.*

2 Dificuldades atenuadas através da utilização de métodos estatísticos

Há diversas fontes de onfundimento quando trabalhamos com experimentação e coleta de dados.

Eventos aleatórios Existem fenômenos nos quais há uma aleatoriedade intrínseca. Antes de realizar o experimento não é possível dizer qual será o resultado obtido. O mais simples destes fenômenos é o lançamento de uma moeda.

Antes de realizar o experimento é impossível (a menos que esta tenha duas caras ou duas coroas) prever qual será o resultado obtido.

Quando queremos determinar se uma moeda é ou não honesta podemos lançá-la 400 vezes. Será que mesmo sendo a moeda não viciada podemos obter 180 caras e 220 coroas?

Erro experimental O erro experimental é a variação produzida por fontes conhecidas ou desconhecidas. Efeitos importantes podem ser mascarados pelo erro experimental. Por outro lado, o erro experimental pode levar a indução de efeitos inexistentes.

Os efeitos do erro experimental podem ser grandemente minimizados através da utilização de amostragem adequada e técnicas de planejamento de experimentos.

Suponha que deseja-se determinar o conteúdo de um pacote de café coletado em um supermercado. Mesmo que pesemos o mesmo pacote 200 vezes na mesma balança, não obteremos 200 resultados idênticos. Dependendo da precisão da balança é provável obtermos 200 resultados distintos.

Confundimento entre relação e causa Os dados representados na Figura 1 consistem da população da cidade de Oldenburg ao final de cada ano do período 1930–1936 versus o número de cegonhas observadas naquele ano. Embora, ninguém diria com base nesse gráfico que as cegonhas são a causa do aumento da população, este tipo de erro é feito com muita frequência em outros contextos. Considere outros exemplos, crianças com pés maiores soletram melhor, bairros com maiores taxas de divórcio tem taxas de mortalidade menores, países que acrescentam fluor na água potável têm maiores taxas de câncer. Embora, existam estudos que estabelecem todas estas relações, não podemos afirmar que exista causa e efeito. Nestes casos, provavelmente existe um terceiro fator não considerado na análise. Crianças com pés maiores tem maior habilidade em soletrar porque são mais velhas. Idade também é um fator a ser considerado no exemplo seguinte, como casais que são mais velhos são menos prováveis de se divorciarem e mais prováveis de morrerem do que aqueles em bairros onde o perfil demográfico é mais jovem. Nações que adicionam fluor na água são, em geral, mais ricas e mais conscienciosas em termos de saúde, e assim uma grande porcentagem de sua população vive o suficiente para adquirir câncer, a qual é uma doença predominantemente de idosos.

Complexidade dos efeitos estudados Suponha que uma nutricionista deseja estimar o efeito de 5 dietas para perder peso. Existe uma dieta padrão (controle) e as outras 4 dietas consistem em combinar 2 níveis de proteína e 2 níveis de carboidratos. Além disso, ela deseja estudar o efeito da dieta em homens e mulheres. Note que nesse caso, existem os efeitos de diversos fatores, proteínas, carboidratos, sexo e dieta/controle. Além disso, existe a interação entre os fatores. Por exemplo, as dietas fazem efeitos diferentes para homens e mulheres. Este tipo de problema pode ser “facilmente” contornado utilizando-se um bom planejamento de experimentos.

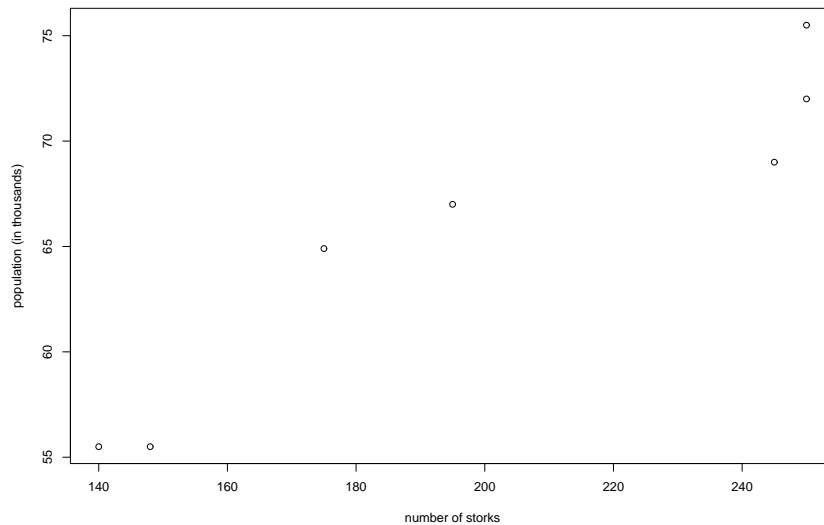
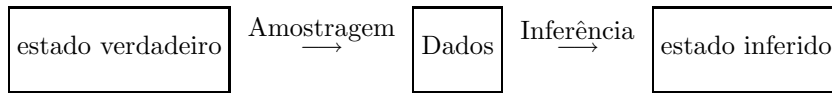


Figura 1: Gráfico da população de Oldenburg no final de cada ano no período de 1930–1936 versus o número de cegonhas avistadas durante o ano

3 Objetivo da Inferência



Exemplo 3.1 *Pesquisas de opinião*

A fim de descobrir quanto é impopular a CPMF inicialmente concebida a fim de arrecadar recursos para a saúde, a DataFolha entrevistou 2000 pessoas e descobriu que 1740 destas pessoas são contrárias à volta do imposto. Isto nos diz que, nesta amostra, 87% das pessoas são contrárias ao imposto. O que isto diz a respeito da população do Brasil? Será que a proporção de pessoas contrárias à CPMF é próxima de 87%? Se sim, qual a margem de erro desta estimativa? Será que há evidência significativa que o valor verdadeiro é maior que 85%?

Exemplo 3.2 *Pesquisas de opinião.*

A fim de verificar a popularidade do governo Lula após um ano de mandato, a Datafolha realizou uma pesquisa entre os dias 8 a 12 e no dia 15 de dezembro em 396 cidades brasileiras. Foram entrevistados 12.180 brasileiros, e a margem de erro máxima para o levantamento é de dois pontos percentuais, para mais ou para menos. O desempenho do presidente Luiz Inácio Lula da Silva é aprovado por 42%, considerado regular por 41% e ruim ou péssimo por 15%.

Exemplo 3.3 *Confiabilidade*

Uma indústria produtora de fornos de microondas gostaria de saber quanto tempo deve-se dar de garantia a seus produtos de modo que somente 1% de seus produtos devam ser reparados no prazo de garantia. Através de testes acelerados, pode-se obter os tempos de vida de 50 fornos de microondas e com base nestes dados estimar o prazo de garantia.

Exemplo 3.4 *Controle de qualidade*

Uma indústria é fornecedora de rolamentos para indústria automobilística, os limites de especificação para rolamentos são 15mm e 18mm, peças fora da especificação são refugadas. Devemos então com base em uma amostra descobrir um intervalo de valores plausíveis para os rolamentos de modo que possamos saber se “grande” parte dos rolamentos produzidos estejam dentro dos limites de especificação. Se descobrimos que este intervalo é de (15.8; 17.2) podemos ficar tranquilos, mas se este intervalo for (12.0; 20.0) devemos nos preocupar em diminuir a variabilidade do processo.

Exemplo 3.5 *Melhoramento industrial*

Uma indústria de sucos concentrados está desenvolvendo um novo procedimento para retirar a água do suco de laranja de modo que o suco reconstituído seja mais agradável ao paladar. O sabor do suco varia em uma escala de 1 à 10. Pelo procedimento atual este índice é uma variável aleatória normalmente distribuída com média 7 e desvio padrão 1. Como descobrir se o novo método é realmente melhor?

Exemplo 3.6 *Calibração*

A resistência à tração é uma característica internacionalmente utilizada para classificar a qualidade de ferros fundidos cinzentos, entretanto além do seu alto custo, pode ser impossível de ser determinada em alguns casos. Outro método de se determinar a resistência de ferros fundidos é a resistência à compressão entre cunhas. Como relacionar as duas medidas?

4 Passos de uma boa análise estatística

- Descrição do problema
- Coleta de dados – Amostragem e Planejamento de Experimentos
- Inferência
- Conclusão

4.1 Descrição do problema

Todo problema estatístico se inicia com uma boa descrição do problema. Suponha, por exemplo, que um pesquisador decida que seria interessante estudar a “Saúde dos alunos da Unicamp”. Pois bem, este é um tema muito vago o qual geraria um questionário vago ou muito complicado.

Por que devemos escolher bem a questão de interesse?

1. Com uma questão específica em mente pode-se elaborar um questionário mais preciso e focado nas questões realmente importantes. Além disso, imagine que o pesquisador depois de coletar os dados descubra que ele realmente está interessado no efeito do esporte sobre a pressão arterial dos estudantes e ele esqueceu de medir a pressão, ou de anotar qual o tipo de atividade física de cada aluno entrevistado.
2. Nunca se deve formular as questões de interesse após a leitura dos dados obtidos. Isto invalida as conclusões. Um pesquisador não ético poderia escolher responder somente as perguntas que lhe interessam e as quais ele teria respostas que confirmassem suas conjecturas.

4.2 Coleta de dados

A coleta de dados é uma das fases mais importantes de um experimento. Somente se os dados forem coletados de forma correta, as conclusões são válidas. Também a forma como os dados devem ser analisados dependem de como foram coletados. Existem basicamente duas maneiras de se coletar dados que podem ser analisados estatisticamente: amostragem e experimentos planejados.

População e amostra: O objetivo de uma investigação científica é descobrir (entender, estudar) alguma característica de certa população. Como, em geral, é impossível ou impraticável examinar toda a população, examinamos parte e com base nestes dados fazer inferências a respeito de toda população. Temos que distinguir entre população alvo e população amostrada.

População alvo: É aquela em que estamos interessados em estudar. No Exemplo 3.2, a população alvo são todos os eleitores do Brasil.

População amostrada: É aquela da qual retiramos a nossa amostra. No Exemplo 3.2, a população amostrada são adultos que moram nas 396 cidades selecionadas.

Se a população amostrada não é a mesma que a população alvo, as conclusões (inferências) obtidas através da amostra só são válidas para a população amostrada.

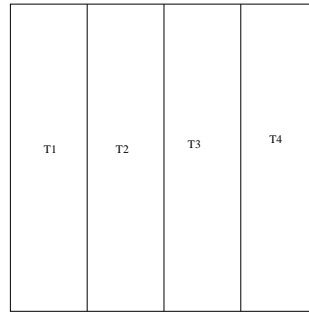


Figura 2: Alocação de 4 tratamentos a 4 UE

Pergunta: Como selecionar uma amostra?

- O tamanho da amostra é fixado através do nível de precisão desejado.
- Os requisitos de uma amostra é que seja finita, representativa da população, aleatória e não sistemática.

A amostragem mais representativa de uma população é a **Amostragem Aleatória Simples (AAS)**. Neste procedimento de amostragem todos os elementos da população têm a mesma chance de serem escolhidos para fazerem parte da amostra.

Como realizar uma AAS? **Método do chapéu:** “Escreva todos os nomes dos indivíduos da população em pedaços idênticos de papéis. Coloque os nomes dentro do chapéu, embaralhe bem todos os nomes e retire aqueles nomes que farão parte da amostra.”

Claro que existem procedimentos computacionais que simulam este método, mas o importante é manter em mente qual a filosofia que está por trás de um bom procedimento de amostragem.

Experimentos planejados Suponha que você gostaria de estudar qual o efeito de 4 tipos de adubação (Adubo orgânico (T1), adubo químico (T2), 10-10-10 (T3) e nenhum adubo (T4)) na produção de tomates. Antes de sair plantando dezenas de hectares com os adubos o melhor é estudar o efeito em pequena escala. Tome, por exemplo, um lote de 10m X 10m. Pense em dividir este lote em sub-lotes e plantar tomates em cada sub-lote sujeito a cada um dos adubos, no final, a produção de cada sub-lote será medida e comparada. Cada sub-lote neste caso, se chama de *Unidade Experimental* (UE).

Entretanto, esta não parece ser uma idéia tão boa, pois para cada tratamento teremos somente uma resposta e não há como avaliar o erro experimental, isto é, o efeito provocado pelos possíveis fatores que estão agindo no processo, mas que não foram incluídos no estudo. Sendo assim, gostaríamos de ter mais de uma replicação do experimento: mais de um ensaio em cada condição experimental (tratamento).

Para isto poderíamos dividir o lote em 16 partes de aplicar o tratamento T_i a n_i UE's, onde $i = 1, 2, 3, 4$, de modo que $n_1 + n_2 + n_3 + n_4 = 16$. Se $n_1 = n_2 = n_3 = n_4 = 4$ dizemos que o experimento é *balanceado*.

T1	T3	T1	T2
T4	T2	T4	T3
T3	T4	T1	T2
T4	T3	T2	T1

Figura 3: Alocação de 4 tratamentos a 16 UE

Rio

T1	T2	T4	T3
T2	T3	T1	T4
T3	T4	T2	T1
T4	T1	T3	T2

Figura 4: Alocação de 4 tratamentos a 16 UE em blocos

Pergunta importante: Como distribuir os tratamentos nas diversas UE's?

Resposta: Através de aleatorização.

Experimentos completamente aleatorizados Quando utilizamos AAS para distribuir os tratamentos à UE's. Por exemplo, escrever números de 1 a 16 e colocar em um “chapéu”, retirar os números sequencialmente e aplicar T1 aos n_1 primeiros lotes retirados, etc. Suponha que queremos um experimento balanceado e os números sorteados foram: 16, 3, 1, 11, 4, 6, 12, 15, 4, 6, 9, 2, 14, 8, 10, 5, 7, 13. Assim, nosso desenho experimental seria:

Imagine agora que temos a situação onde um rio passa por um dos lados do campo. Nesse caso, talvez a distribuição apresentada na Figura 3 não seja adequada e deveríamos utilizar uma distribuição das UE's em subgrupos mais homogêneos (*blocos*).