Experimentos e Análise de Dados

Prof. Moacir A. Ponti
www.icmc.usp.br/~moacir

Instituto de Ciências Matemáticas e de Computação – USP

Sumário

Amostragem e Variáveis

Censo vs Amostragem

- É muito raro ser necessário realizar **censo**
- **Amostragem** sempre implica em aceitar um erro, mas pode ser representativa

Variáveis

- Numéricas: discretas/contínuas
- Categóricas: ordinais/não-ordinais

Amostragem e Variáveis

Exemplo: verificar sal na panela

- Análise exploratória: **amostragem** (porque não censo?)
- Concluir se mais sal é necessário: **inferência**
- Amostra precisa ser **representativa**: aleatoriedade.

Viés de amostragem

Conveniência

Amostra facilmente acessível pelo pesquisador

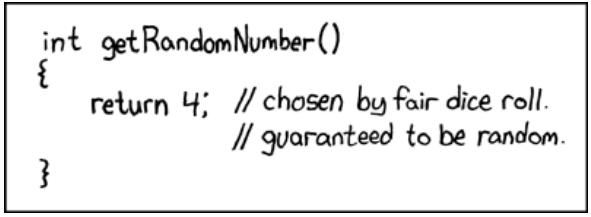
Abstenção

Apenas uma fração (não aleatória) da população responde

Resposta voluntária

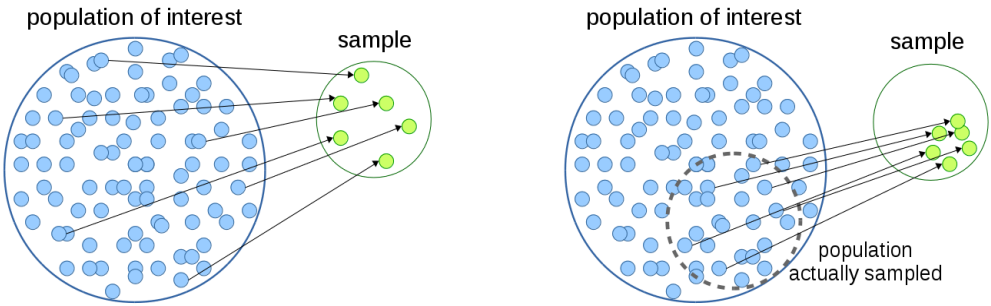
Participantes com opinião forte tem mais chance de responder (dentre os aleatoriamente selecionados)

Estratégias de amostragem



Agradedimentos à http://xkcd.com

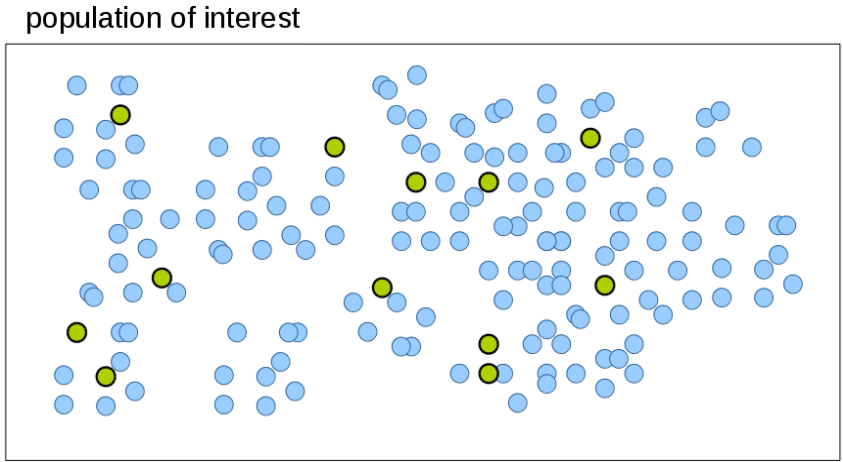
Viés de amostragem e i.i.d.



Nota: No caso à direita, não podemos considerar que a amostra é independente e identicamente distribuída (i.i.d.), comumente assumido por muitos métodos.

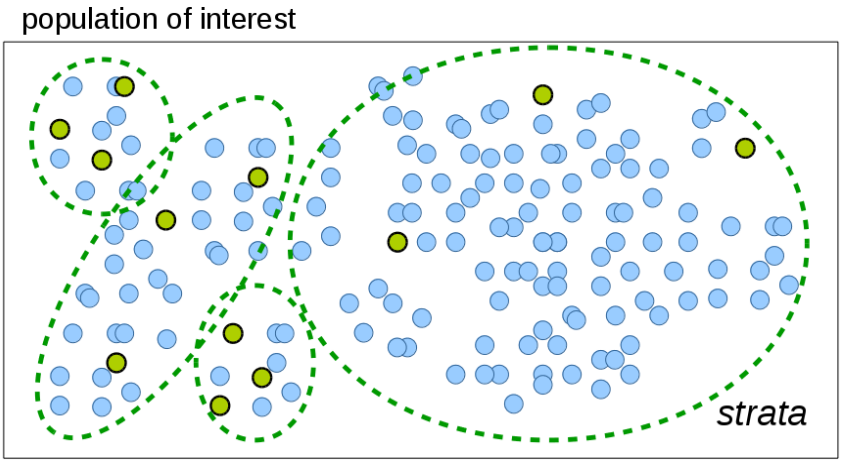
Estratégias de amostragem

Amostragem aleatória simples (Simple random sampling)



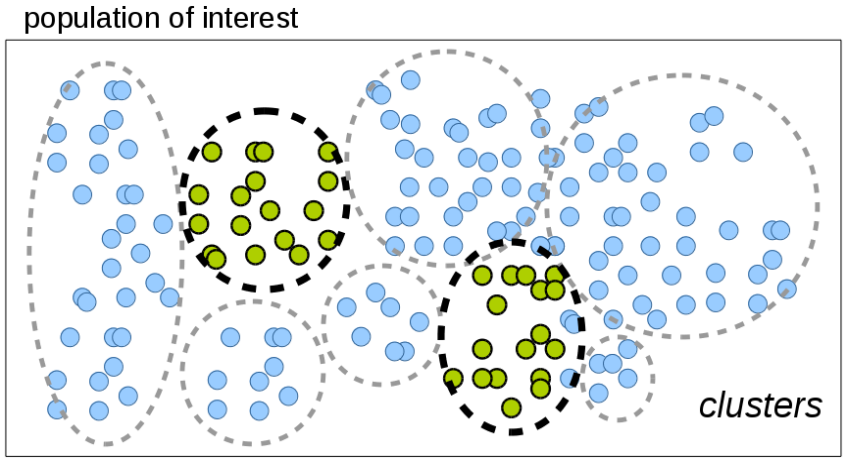
Estratégias de amostragem

Amostragem estratificada



Estratégias de amostragem

Amostragem por agrupamento



OBS: também pode ser feita amostragem aleatória simples dentro de cada cluster.

Sumário

Experimentos

Visam estabelecer relações causais, correlações ou comparações.

1. **Controle**: comparar intervenção com um grupo controle;
2. **Aleatorização**: distribuir sujeitos/exemplos de forma aleatória;
3. **Replicação**: coletar amostra suficiente, ou replicar estudo;
4. **Bloqueio**: bloquear por variáveis que possam afetar resultado.

Terminologia (pouco comum em computação): placebo, efeito placebo, estudo cego e duplo-cego.

Experimentos: amostragem e atribuição

Exemplos:

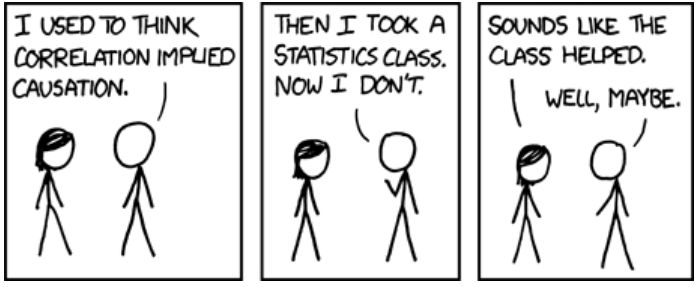
1. Método de segmentação com aplicação médica;
2. Projeto de uma nova tela de tinta eletrônica para facilitar a leitura;
3. Método para melhoria da segmentação de sentenças em fala.

Experimentos: amostragem e atribuição

(ideal)	Atribuição aleatória	Sem atribuição aleatória	(observacional)
Amostragem aleatória	Causal e generalizável	Não causal, generalizável	Generalização
Amostragem não aleatória	Causal, não generalizável	Não causal, não generalizável	Não generalização
(mais comum)	Causalidade	Associação	(indadequado)

Agradedimentos à Mine Çetinkaya-Rundel

Causalidade vs Correlação



Agradedimentos à `http://xkcd.com`

Sumário

Medidas e transformações

Medidas de centro e dispersão

- **Comum:** média e desvio padrão
- **Robustas:** mediana e IQR

Transformação

- Logaritmo, Raiz Quadrada.
- Normalização.

Correlação

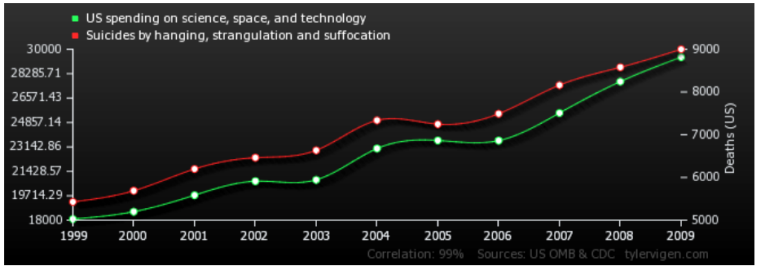
Medida estatística que indica a direção e a força da relação linear entre duas variáveis
Se correlação é $\neq 0$, então:

1. variável A causa B,
2. variável B causa A,
3. uma variável C causa A e B,
4. A causa C que por sua vez causa B, ou
5. não há relação entre A e B.

US spending on science, space, and technology

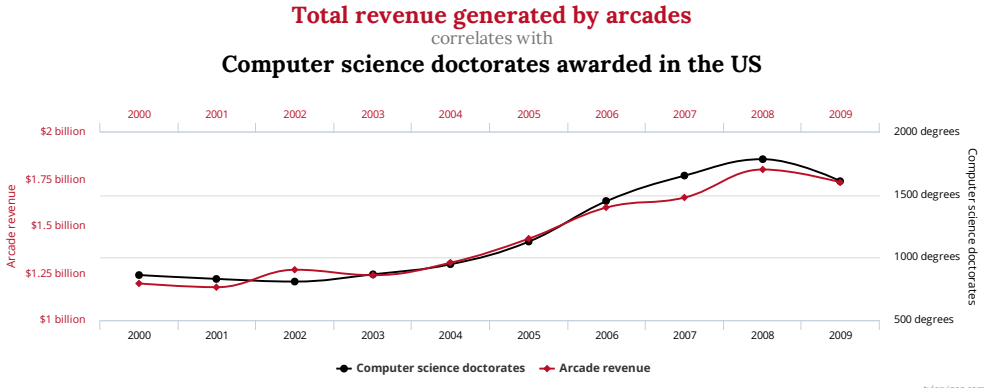
correlates with

Suicides by hanging, strangulation and suffocation



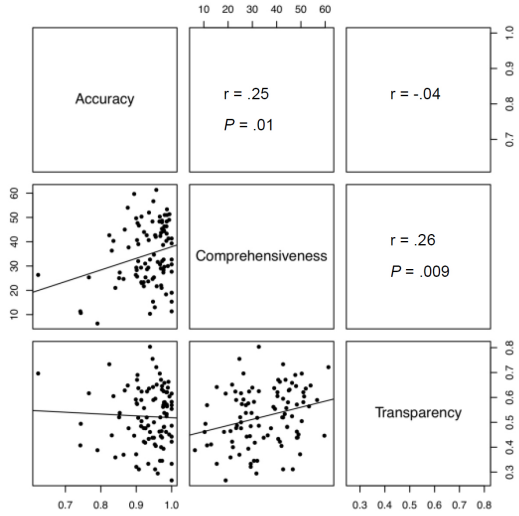
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000
Correlation: 0.992082											

Agradedimentos à http://tylervigen.com/



Agradedimentos à http://tylervigen.com/

Correlação vs Regressão Linear



McNally et al (2012) http://www.i-jmr.org/2012/1/e1/

Correlação vs Regressão Linear

- r (coeficiente de correlação) indica a direção e a força da relação linear entre duas variáveis
- r^2 indica a proporção da variação de uma variável explicada pela outra em um modelo de regressão linear simples

OBS: para $r = 0.25$, a correlação quadrada é $R^2 = 0.06$

Sumário

Teste de hipótese

1. Especifica **hipótese nula** e **hipótese alternativa**
2. Assume que a hipótese nula é **verdadeira** e calcula a **estatística de teste**
3. Calcula o **p-valor**: se a hipótese nula é verdadeira, qual a probabilidade de observarmos tão extremos quanto aquele que dispomos?
   - se o nível for inferior a um limiar α que define a probabilidade de cometer erro tipo I, rejeitar a hipótese nula;
   - do contrário, não rejeitar hipótese nula.

Teste de hipótese

Testes comumente utilizados:

- Teste t-Student (ou Teste t): comum para dados com distribuição Normal,
- Wilcoxon: não paramétrico, compara rankings entre dois conjuntos de dados,
- ANOVA: analisa multiplos conjuntos pela estatística F.
- Kruskal-Wallis: não -paramétrico

Comparação de múltiplos resultados (ex. datasets) e várias intervenções (ex. métodos)

- Friedman ranking

Teste de hipótese

Statisticians issue warning over misuse of P values

"Misuse of the P value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced..."

`http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503`