Extração de Dados

Revisão Sistemática e Meta-Análise

Marcelo M. Weber & Nicholas A. C. Marino

github.com/nacmarino/maR

Recapitulando

- **Revisão Sistemática**: "é uma síntese da pesquisa disponível em um tópico precisamente definido, usando métodos explícitos para identificar, selecionar, avaliar criticamente, e analisar os resultados relevantes". (*Koricheva et al, 2013*)
- **Meta-Análise**: "é a análise estatística de uma ampla coleção de resultados de estudos com o propósito de integrar a evidência disponível". (*Glass, 1976*)
- Uma meta-análise é um componente opcional da revisão sistemática.

Recapitulando

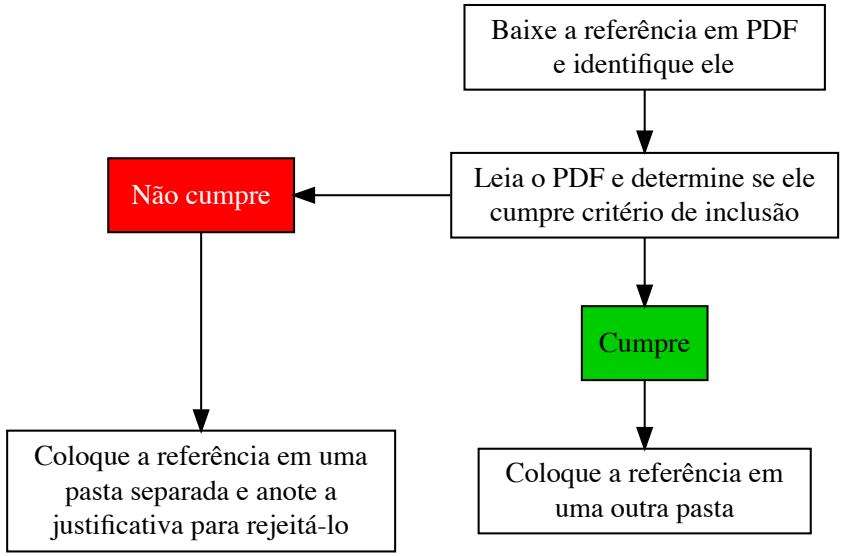
- Para os trabalhos que forem vistos:
  - Documente todos os passos e decisões;
  - Faça uma lista de todos os trabalhos vistos, com o status e informações relevantes de cada um deles, que os levaram a ser aceitos ou rejeitados.
- Cada trabalho visto deve receber um número de identificação.
- Cada linha recebe as informações de uma única observação.
- Em cada coluna, apenas um tipo de dado.

Recapitulando

id_estudo	autor	ano	revista	entra	observacao
1	Fulano et al	2013	Vovo Mafalda	sim	cumpre requisitos
2	Beltrano	2014	Tio Patinhas	sim	cumpre requisitos
3	Primano	2016	Turma da Monica	sim	informacoes no SM
4	Hermano et al	2010	Mickey	sim	multiplos niveis do tratamento
5	Ciclano & Juvano	2010	Galvalandia	nao	experimento nao replicado

Recapitulando

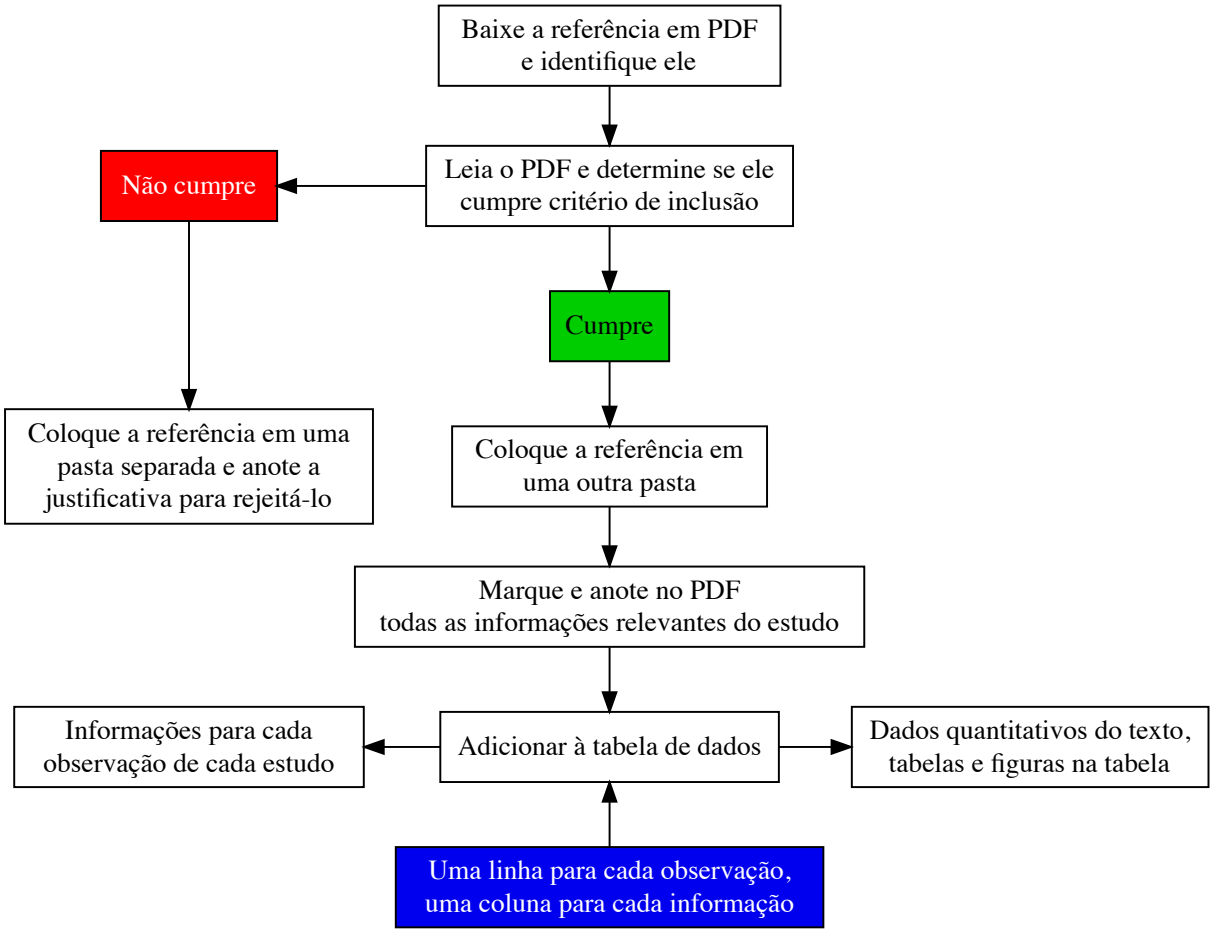
- Esperamos que todos estejam nesta fase.
- Com a lista de trabalhos que vão cumprem os critérios de inclusão, o passo seguinte é a extração dos dados.



Extração de Dados

- É uma das partes mais importantes de uma revisão sistemática - se não a parte mais importante;
- O tempo gasto aqui é o tempo que você nao vai gastar no futuro;
- Gaste tempo **planejando**:
  - O(s) critério(s) de inclusão para a extração de dados.
  - As informações que precisam ser extraídas de cada estudo.
  - O formato que cada variável extraída vai assumir na base de dados.
- Se estiver na dúvida, revisite o **PICO**.
- **Documente todas as decisões e escolhas**.

Fluxograma de Trabalho



Que informações extrair?

- O tipo de informação a ser extraída depende da natureza da sua pergunta:
  - **Informações básicas sobre o estudo**: localidade, coordenadas, clima,…
  - **Outras informações sobre o estudo**: tamanho da área amostrada, tipo de ecossistema, forma de amostragem, espécies envolvidas,…
  - **Informações sobre a manipulação de interesse**: desenho aditivo ou substitutivo, níveis da manipulação, espécies adicionadas,…
  - **Dados quantitativos**: médias, coeficientes de correlação, slopes, erros, tamanho amostral
  - **Outras informações relevantes** (informações sobre as espécies, informações sobre background do solo,…)

Mas o que são informações relevantes?

- Na sua cabeça…*"tudo pode ser potencialmente importante, e tudo influencia tudo"*.
  - Mas por quê *x*, *y* ou *z* podem ser importantes?
  - Como você espera que *a* influencie *b*?
- Foque na sua pergunta - a partir dela, você vai ter noção do que é importante extrair.
- **Você está testando uma hipótese**…o que a literatura diz sobre ela?
- Que outros corpos de teoria podem indicar quais informações são importantes?
- **Estar familiarizado com a área que você está revisando é fundamental**.

Quais informações extrair?

Variáveis correlacionadas em um mesmo estudo

- Diferentes medidas de diversidade, densidade ou biomassa;
- Diferentes formas de medir um processo (e.g., emissão de um gás);
- Diferentes formas de inferir o comportamento (e.g., visitação de plantas);
- Diferentes forma de medida um organismo (e.g., plantas vs animais);
- ...

Quais informações extrair?

Dados apresentados para durações diferentes

- **Medidas finais**: capturam todo o histórico do experimento/observação, mas também pode sofrer influência de outros fatores que não o desejado;
- **Medidas iniciais**: capturam a resposta inicial do experimento/observação, mas podem estar sujeitos à influência da estocasticidade e não refletir a tendência à longo prazo;
- **Integrar todas as medidas**: estimativa mais robusta, mas muito mais trabalhosa.

Quais informações extrair?

Múltiplas observações a partir de um estudo

- Diferentes níveis de um mesmo tratamento;
  - Gradiente de riqueza, de área, de intensidade, de concentração, de 'idade' das unidades experimentais.
- Diferentes unidades de observação-alvo em um mesmo estudo;
  - Espécies, indivíduos, ambientes,..., populações diferentes avaliadas no mesmo estudo;
  - Observações dependentes por virem do mesmo estudo;
  - Observações independentes por serem 'experimentos' diferentes.

Quais informações extrair?

Estudos multifatoriais

- Por exemplo, você quer saber qual o efeito da adição de nutrientes em uma variáveis resposta *x*, e um estudo manipula a concentração de nutrientes (baixa vs alta) e o disturbio (baixo vs alto) de forma fatorial.
- Uma opção é usar os níveis do outro fator como uma 'realidade' pararela: para cada um dos níveis do distúrbio, você vai ter um medida da adição de nutrientes;
- Outra opção é selecionar um dos níveis do segundo fator e trabalhar apenas com ela, para simplificar as coisas e reduzir ruído.
- Se esta for a sua pergunta, você também pode usar uma medida de effect size bifatorial (ou multifatorial, mas aqui complica a interpretação).

Quais informações extrair?

Outros casos

Como você encara dados do mesmo experimento/localidade apresentados em múltiplos estudos?

- *Salami Science*: mesmo experimento apresentado como uma série de artigos (normalmente) de menor impacto;
- *Pão Francês*: pequenos experimentos repetidos inúmeras vezes, podendo ser mais ou menos similares entre si;
- *De volta para o futuro*: resultados do mesmo trabalho descrito anteriormente, mas agora com *x* anos/meses/semanas/dias a mais de coleta.

Como registrar cada informação?

- **Regra de ouro**: uma observação por linha, um tipo de dado por coluna.
- Você não precisa registrar todas as informações em uma única tabela - eu, particularmente, sugiro usar uma estrutura de base de dados.
- Informações da inclusão do estudo:

Como registrar cada informação?

id_estudo	autor	ano	revista	entra	observacao
1	Fulano et al	2013	Vovo Mafalda	sim	cumpre requisitos
2	Beltrano	2014	Tio Patinhas	sim	cumpre requisitos
3	Primano	2016	Turma da Monica	sim	informacoes no SM
4	Hermano et al	2010	Mickey	sim	multiplos niveis do tratamento
5	Ciclano & Juvano	2010	Galvalandia	nao	experimento nao replicado

id_estudo	pais	especie	manipulacao	concentracao_n	concentracao_p
1	Brasil	araucaria angustifolia	np	50	50
2	Patopolis	theobroma cacao	n	50	0
3	Sao Paulo	handroanthus albus	n	50	0
4	Disneylandia	cecropia hololeuca	np	25	100

Dados quantitativos

- É a parte principal para quem vai fazer uma meta-análise.
- É a etapa da extração de dados que consumirá mais tempo de todo o processo.
- Importante registrar de onde veio cada dado extraído para a meta-análise.
- Mais importante ainda é determinar a qualidade do que você está extraindo: *garbage in, garbage out*.

Dados quantitativos

1. Medida do Efeito:
   - Valores de 'média' para cada observação/tratamento;
   - Coeficientes de Correlação ou Slopes de Regressão;
   - Valores de resultados positivos e negativos;
   - Outras métricas.
2. Uma estimativa de erro (é fácil converter entre elas):
   - Variância;
   - Desvio Padrão;
   - Erro Padrão;
   - Intervalo de Confiança.
3. Tamanho Amostral.

Exemplo de uma tabela de dados quantitativos

id_estudo	fonte	media_controle	erro_controle	n_controle	tipo_erro_controle
1	Tabela 1	10	3.2	12	se
2	Figure 2a	6	2.1	10	sd
3	Texto	8	1.9	14	ci
4	Mat Sup Fig 1	20	0.4	20	se

media_tratamento	erro_tratamento	n_tratamento	tipo_erro_tratamento	boxplot
18	2.5	12	se	nao
12	0.9	10	sd	sim
10	1.5	14	ci	nao
21	0.6	20	se	nao

E se faltar algum dado quantitativo?

Pode ocorrer por diversos motivos, dentre eles:

- Dados foram apresentados muito mal (bad reporting);
- Dados não foram apresentados seguindo o desenho experimental;
- Dados não foram apresentados.

O que fazer:

- Entrar em contato com o(s) autor(es) do trabalho: nem sempre é o desejável, tampouco é garantia de conseguir os dados.
- Tentar algum tipo de imputação dos dados: você usa relações existentes na base de dados para 'predizer' qual é o valor que foi perdido.
- Excluir observação da base de dados: não é o desejável, mas é o que precisa ser feito às vezes;
- Usar uma medida de tamanho de efeito alternativa.

Como tirar os dados a partir de figuras?

- Tradicionalmente, isto era feito com um paquímetro.
- Existem softwares grátis que te permitem determinar as coordenadas de cada 'ponto' em uma figura (exemplo, mas existem muito mais):
  - ImageJ
  - DataThief (vou mostrar esse daqui a pouco)
  - WebPlotDigitizer
  - GraphClick
- O pacote *metagear* no R também tem uma ferramenta que serve para determinar os pontos em uma figura digitalizada.

E se houver mais de uma observação para um mesmo estudo?

- Se, por algum motivo, você vai usar múltiplas observações a partir do mesmo estudo, a forma de entrada de dados é a mesma que a descrita anteriormente (a regra de ouro vale sempre).
- A observação deve receber o mesmo número de identificação para a identidade do estudo.

id_estudo	autor	ano	revista	observacao	pais	especie	manipulacao	concentracao_n
1	Fulano et al	2013	Vovo Mafalda	cumpre requisitos	Brasil	araucaria angustifolia	np	50
2	Beltrano	2014	Tio Patinhas	cumpre requisitos	Patopolis	theobroma cacao	n	50
3	Primano	2016	Turma da Monica	informacoes no SM	Sao Paulo	handroanthus albus	n	50
4	Hermano et al	2010	Mickey	multiplos niveis do tratamento	Disneylandia	cecropia hololeuca	np	25
4	Hermano et al	2010	Mickey	multiplos niveis do tratamento	Disneylandia	cecropia hololeuca	n	25

Devemos dividir esforços?

- Se você é desconfiado, cricri, ou gosta de carregar o mundo nas costas, não.
- Se você acredita nos outros, sabe o valor de trabalhar em equipe, ou quer agilizar o processo, sim.
- No fim das contas, a escolha depende do tamanho da meta-análise e das pessoas disponíveis para ajudar.
- É importante registrar quem extraiu os dados de que trabalho.
- Existe um método para determinar o grau de concordância entre revisores.

Kappa assessment

- Observado: grau de concordância entre dois revisores.

	Aceito	Rejeitado	Total
Aceito	35	20	55
Rejeitado	5	9	14
Total	40	29	69

- Esperado ao acaso: (ΣLinha * ΣColuna)/ΣTotal

	Aceito	Rejeitado	Total
Aceito	31.88	23.11	55
Rejeitado	8.11	5.88	14
Total	40.00	29.00	69

- Número de vezes em que ambos concordaram:
  - Observado: 35 + 9 = 44
  - Ao acaso: 31.88 + 5.88 = 37.76

Kappa assessment

- K = (concordância observada - concordância esperada)/(numero total de observacoes - concordânca esperada)

```
(44 - 37.76)/(69 - 37.76)
```

```
## [1] 0.1997439
```

- Baixa concordância entre revisores merece atenção.
- **Documente todas as decisões e escolhas, e relate:**
  - **se extração de dados foi feita por uma única pessoa ou uma equipe;**
  - **se feito por uma equipe, como você lidou com um possível viés individual.**

Como tirar dados de boxplot?

- Hozo et al, 2005, BMC Medical Research Technology, Estimating the mean and variance from the median, range, and the size of a sample

```
# a: mínimo; m: mediana; b: máximo; n: tamanho da amostra
box_size <- function(a,m,b,n) {
  mn_small <- (a+2*m+b)/4
  mn_with_n <- (a+2*m+b)/4+(a-2*m+b)/(4*n)
  s <- sqrt((a*a+m*m+b*b+(n-3)*((a+m)^2+(m+b)^2)/8-n*mn_small*mn_small)/(n-1))
  s_form <- ((((a-(2*m)+b)^2)/4)+((b-a)^2))/12
  sd_form <- sqrt(s_form)
  s_range_4 <- (b-a)/4
  s_range_6 <- (b-a)/6
  sample_size <- n
  median_data <- m
  calculated <- c(mn_small, mn_with_n, median_data, s, s_form, sd_form, s_range_4, s_range_6, sample_size)
  names(calculated) <- c("Mean", "Mean with n", "Median", "SD with n",
                         "Variance", "SD", "Range 4", "Range 6", "Sample Size")
  return(calculated)
}
```

Para a média a partir do boxplot

- Se o n < 25:

```
mn_small <- (a+2*m+b)/4
```

- Se o n > 25:

```
  mn_with_n <- (a+2*m+b)/4+(a-2*m+b)/(4*n)
```

Para a variância a partir do boxplot

- Se o n < 15

```
s_form <- ((((a-(2*m)+b)^2)/4)+((b-a)^2))/12
sd_form <- sqrt(s_form)
```

- Se o 15 < n < 70

```
s_range_4 <- (b-a)/4
```

- Se o n > 70

```
s_range_6 <- (b-a)/6
```

Resumindo

- O planejamento da extração de dados é fundamental para o sucesso da sua revisão sistemática ou meta-análise: o tempo gasto aqui é tempo bem gasto;
- A sua pergunta vai guiar grande parte da escolha das informações a serem extraídas;
- Ao criar sua planilha da revisão sistemática ou meta-análise tente aproveitar ao máximo da estrutura de uma base de dados;
- Não se esqueça da regra de ouro: uma linha, uma observação; uma coluna, um tipo informação.
- **O mais importante: documente todas as decisões e escolhas que você fizer aqui**.

Literatura Recomendada

1. Hozo et al, 2005, BMC Medical Research Technology, Estimating the mean and variance from the median, range, and the size of a sample

2. Borer et al, 2009, Bull Ecol Soc Am, Some simple guidelines for effective data management

3. Zimmerman, 2008, Sci Tech Human Val, New knowledge from old data - the role of standards in the sharing and reuse of ecological data

4. Whitlock, 2010, Trends Ecol Evol, Data archiving in ecology and evolution - best practices

5. Curtis et al, 2013, Extraction and critical appraisal of data, In: Handbook of meta-analysis in ecology and evolution (Capítulo 5)