ACH3553 - Estatística I

Curso de Gestão de Políticas Públicas

Aula 3

Alexandre Ribeiro Leichsenring
alexandre.leichsenring@usp.br

Índice

1. Organização dos dados
   - Distribuição de dados escalares não agrupados
2. Representação gráfica de dados

Índice

1 Organização dos dados
- Distribuição de dados escalares não agrupados

2 Representação gráfica de dados

Distribuição de dados escalares não agrupados

- Considere dados de notas de estatística em um semestre anterior.
- Quando os dados estão agrupados em classes: Ok!
- Para resumir as informações da base de dados não agrupados, são necessários outros procedimentos
- Exemplos:
  - Percentis
  - Medidas resumo (próxima aula)
  - Representações gráficas para dados desagrupados

Percentis

- Vamos supor que sua nota em estatística foi 7,7.
- Qual foi sua posição relativa no conjunto da classe?
- Para comparar seu desempenho com o desempenho de todos os demais, é preciso uma medida que relacione a sua nota com as notas de todos os demais
- O **percentil** pode ser entendido como uma classificação porcentual

Percentis

Exemplo

Vamos abrir a base **Notas Estatistica.RData** e examinar os dados das notas de estatística.

1. Vamos criar um *ranking* de notas
2. Vamos calcular "manualmente" o *ranking* porcentual
3. Vamos visualizar de maneira mais abrangente como seria o *ranking* porcentual (percentil) para cada nota tirada.

- Acessar os dados em: https://rstudio.cloud/project/448060.

Percentis

Alguns percentis famosos:

Decis

- Percentis de ordem 10
- Pontos (observações) que dividem a distribuição em 10 porções de tamanhos iguais
  - Se uma nota (observação) está classificada no primeiro decil (classificação percentil = 10), 10% dos casos se localizam nesse nível ou abaixo dele;
  - Se uma observação está no segundo decil, 20% das observações estão nesse nível ou abaixo dele;
  - . . .

Percentis

Quartis

- Percentis de ordem 25
- Pontos (observações) que dividem a distribuição em 4 partes
  - 1° quartil: 25% dos casos estão neste nível ou abaixo;
  - 2° quartil: 50% dos casos estão neste nível ou abaixo;
  - 3° quartil: 75% dos casos estão neste nível ou abaixo.

Percentis

Mediana

- Percentil de ordem 50
- Pontos (observações) que dividem a distribuição em 2 partes
  - Metade das observações é menor ou (igual) à mediana;
  - Metade das observações é maior do que a mediana;

Índice

1 Organização dos dados
- Distribuição de dados escalares não agrupados

2 Representação gráfica de dados

Representação gráfica de dados

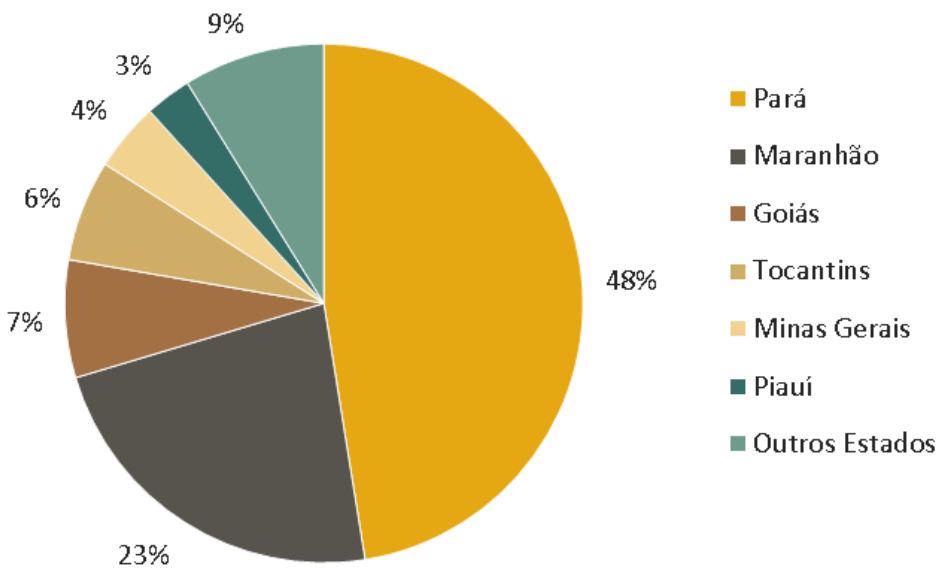
- Nem sempre a informação em forma de tabelas é facilmente apreendida
- Quando há uma quantidade grande de variáveis e categorias, a informação fica um pouco *escondida*
- Métodos de representação gráfica da informação são muito úteis para evidenciar a informação de uma tabela ou base de dados
- Os gráficos podem revelar padrões pouco evidentes

Gráficos para variáveis qualitativas

Gráfico de pizza

- Para dados nominais
- Distribuição de percentuais com soma 100
- Evidenciam diferenças nas porcentagens de categorias

Gráfico 1 - População residente por local de nascimento: Canaã dos Carajás



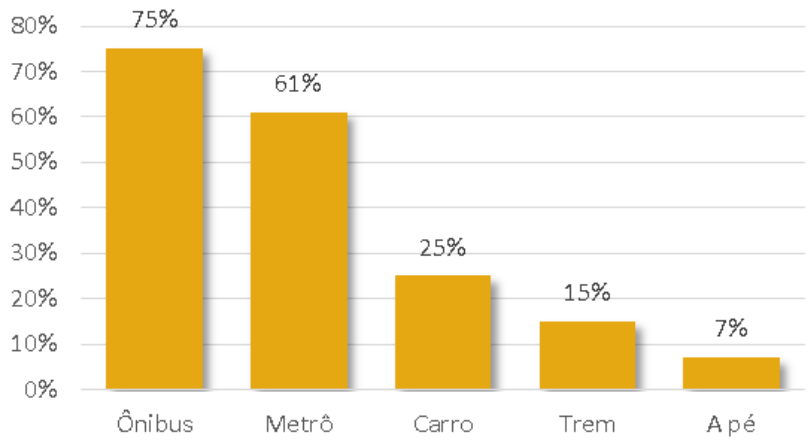
Fonte: IBGE / Censo Demográfico 2010.

Gráficos para variáveis qualitativas

Gráfico de barra

- Comparação de valores
- Distribuição de percentuais, mas cujas categorias tenham uma ordenação natural (por exemplo, escolaridade)
- Quando tenho pergunta com respostas múltiplas (possibilidade de mais de uma reposta): nesse caso, a soma pode extrapolar 100%, e fica artificial colocar numa pizza (que deve somar 100%)

Gráfico 3 *– Meio de locomoção da população*



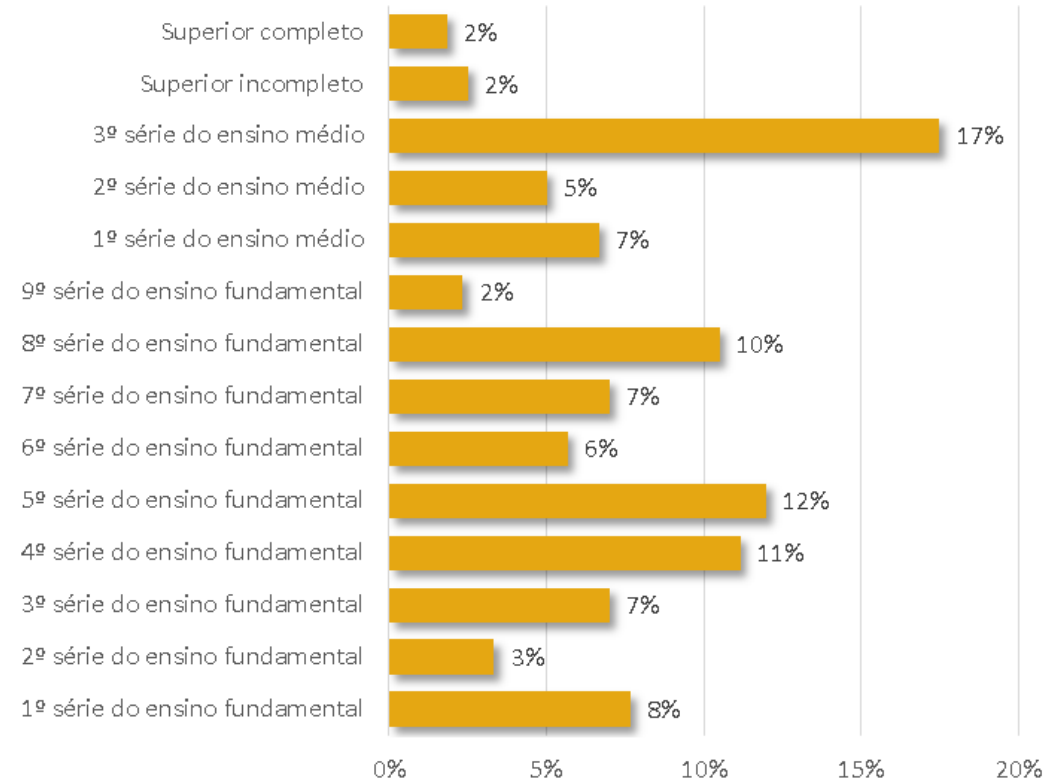
Fonte: Pesquisa origem e destino, 2014

Gráficos para variáveis qualitativas

Gráfico de barra

- Quando há muitas categorias, é adequado usar barras horizontais, o que permite melhor visualização dos nomes das categorias

Gráfico 4 - Escolaridade dos membros da família

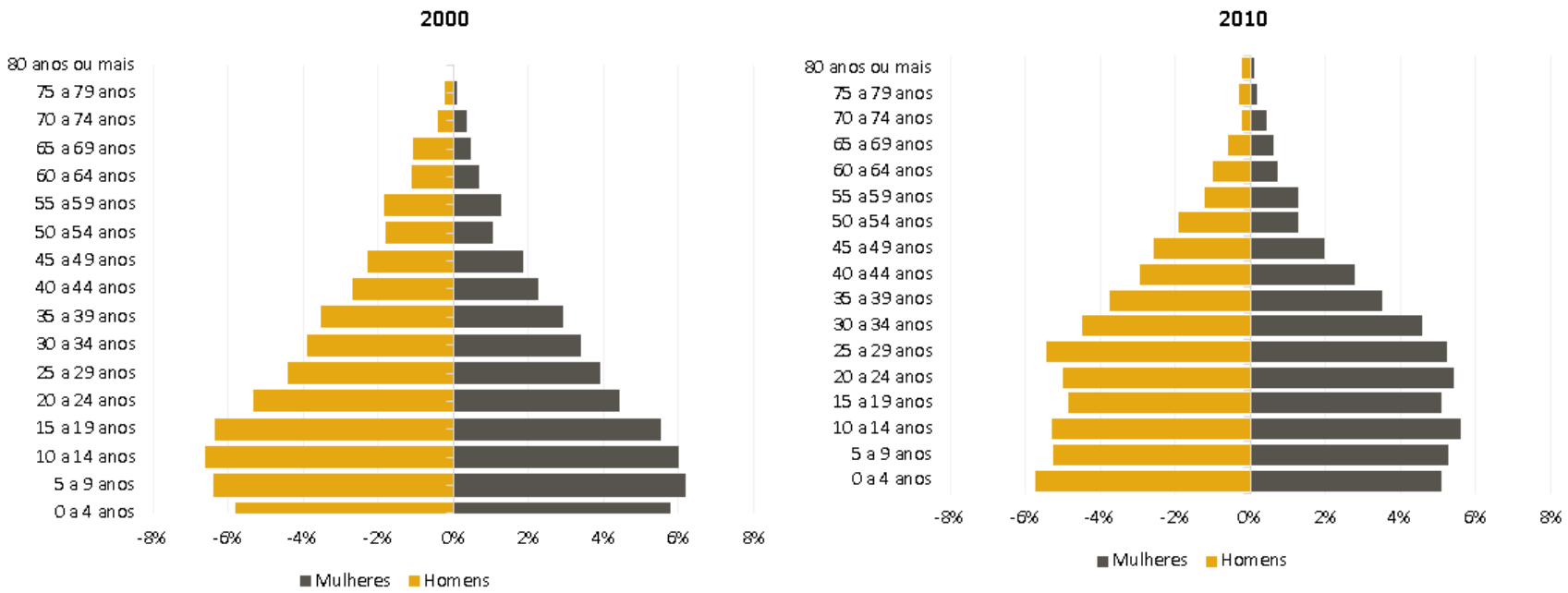


Fonte: IBGE/Pnad, 2011

Gráficos para variáveis qualitativas

Pirâmide etária

Gráfico 10 - População residente por sexo e idade – Canaã dos Carajás 2000 x 2010

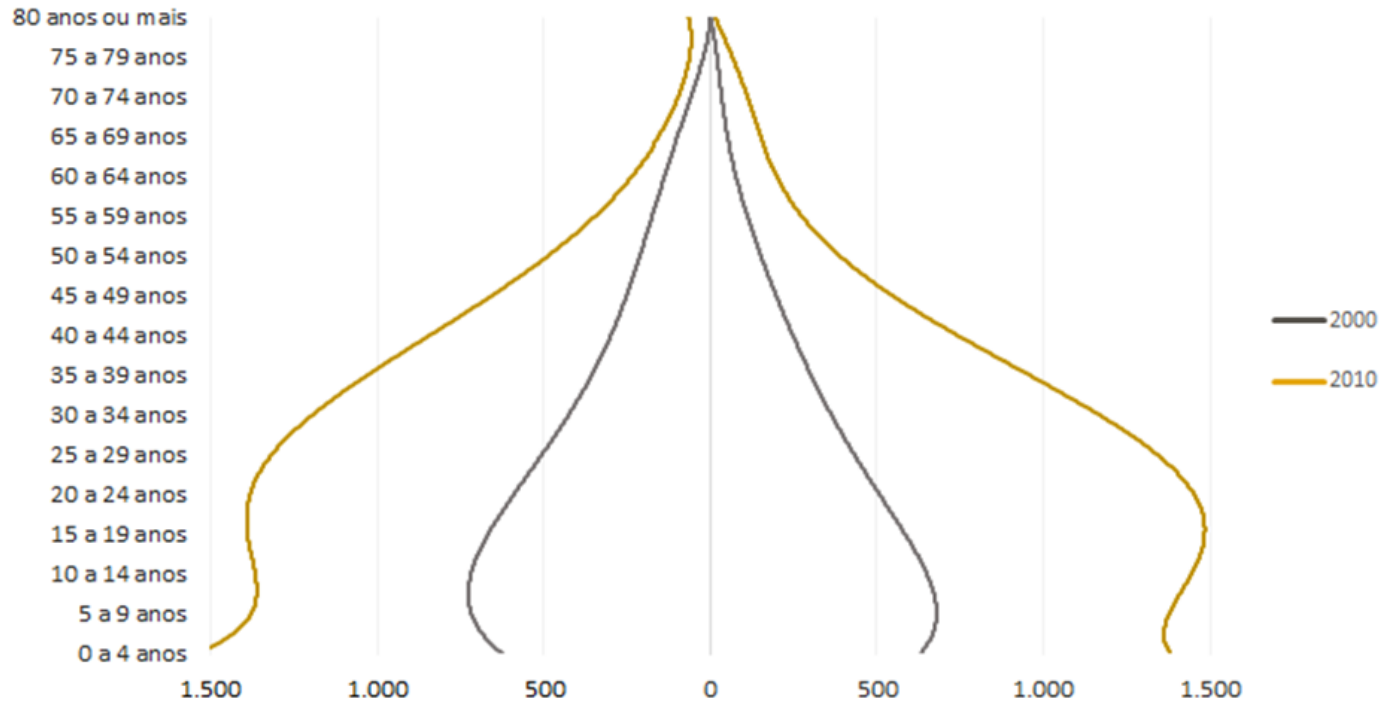


Fonte: IBGE/Censo Demográfico 2000 e 2010.

Gráficos para variáveis qualitativas

Pirâmide etária

Gráfico 11 - População residente por sexo e idade – Canaã dos Carajás 2000 x 2010



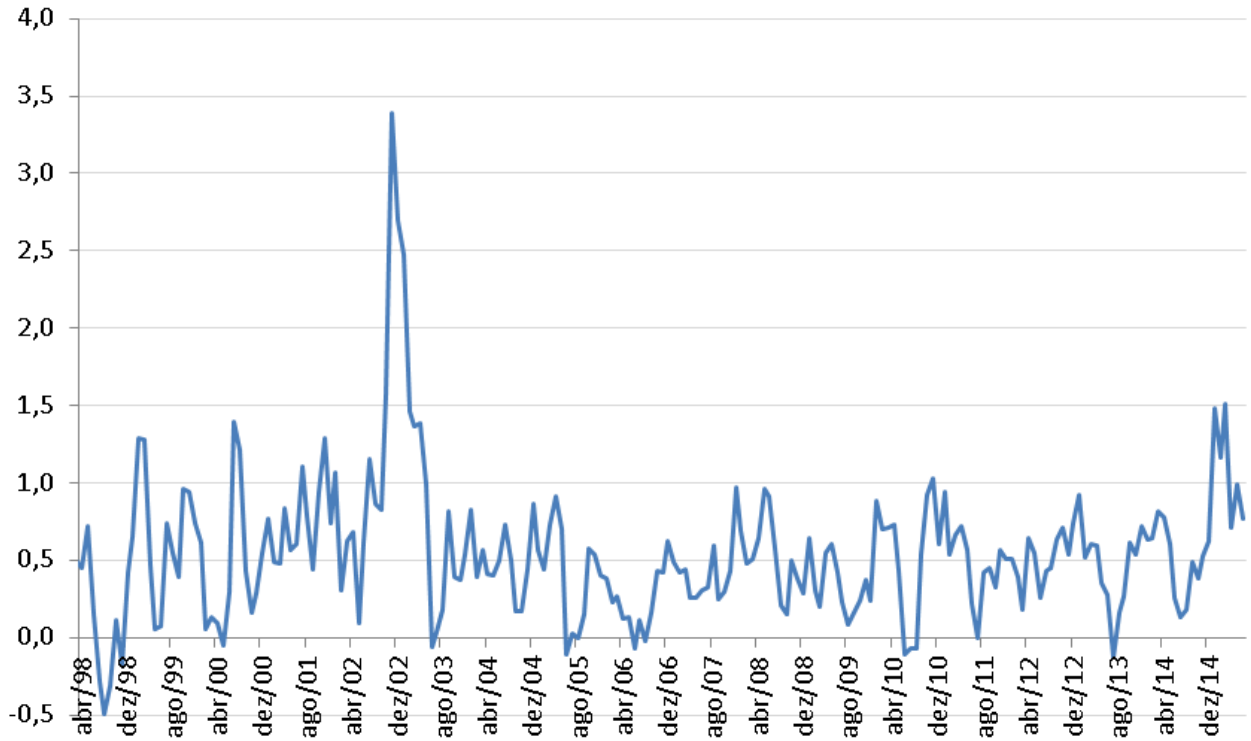
Fonte: IBGE/Censo Demográfico 2000 e 2010.

Gráficos para variáveis quantitativas

Gráfico de série

- Para retratar série histórica

Gráfico 11 - Evolução da taxa básica mensal de juros no Brasil



Fonte: Banco Central do Brasil

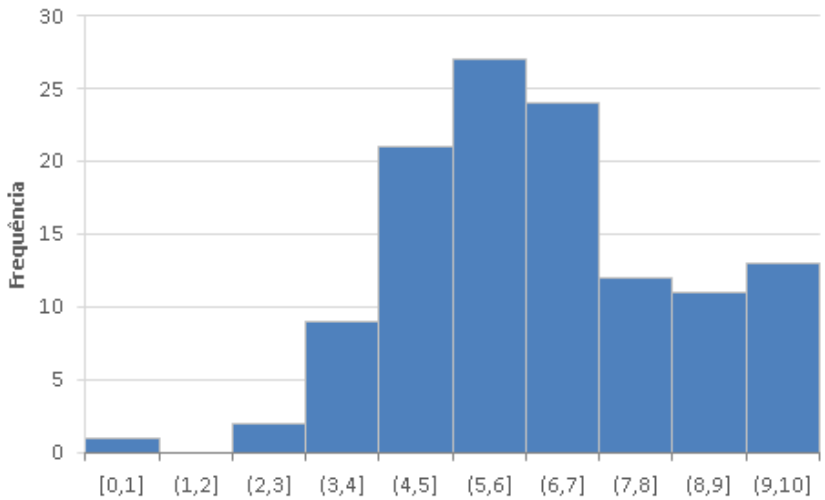
Gráficos para variáveis quantitativas

Histograma

O *Histograma* é um gráfico de barras para dados escalares

1. Os dados são convenientemente divididos em classes
2. Faz-se a contagem das frequências ou das porcentagens de cada classe no conjunto de dados
3. Constrói-se um gráfico de barras

   ▶ A altura da barra correspondente a cada classe é proporcional ao número de observações dentro da classe



Gráficos para variáveis quantitativas

Histograma

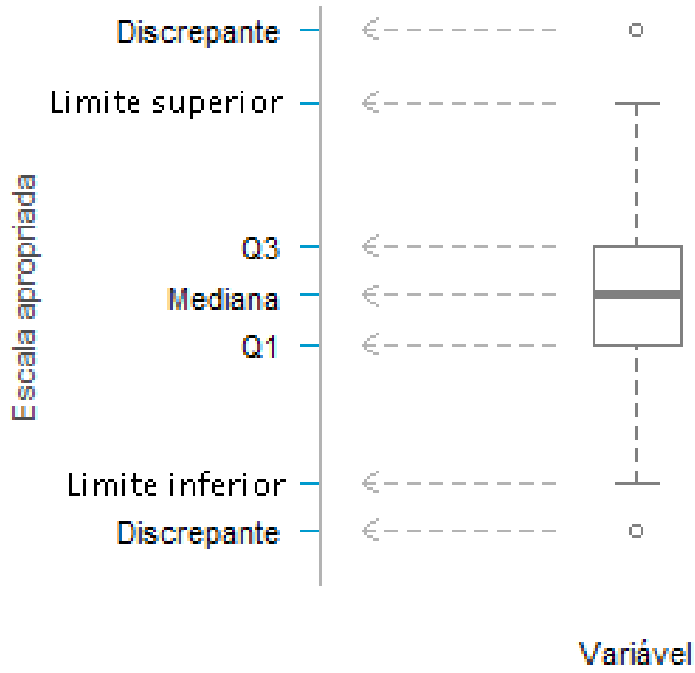
Exemplo

Vamos fazer um *Histograma* para os dados das notas de estatística da base Notas Estatistica.RData.

Boxplot

Aspectos da distribuição de um conjunto de dados através de 5 estatísticas resumo:

- Limite inferior (na ausência de valores discrepantes, coincide com o menor valor)
- Primeiro quartil $Q1$ (25° percentil)
- Mediana (50° percentil)
- Terceiro quartil $Q3$ (75° percentil)
- Limite superior (na ausência de valores discrepantes, coincide com o maior valor)



- Obs. Entre $Q1$ e $Q3$ (interior da caixa) estão 50% das observações

Boxplot

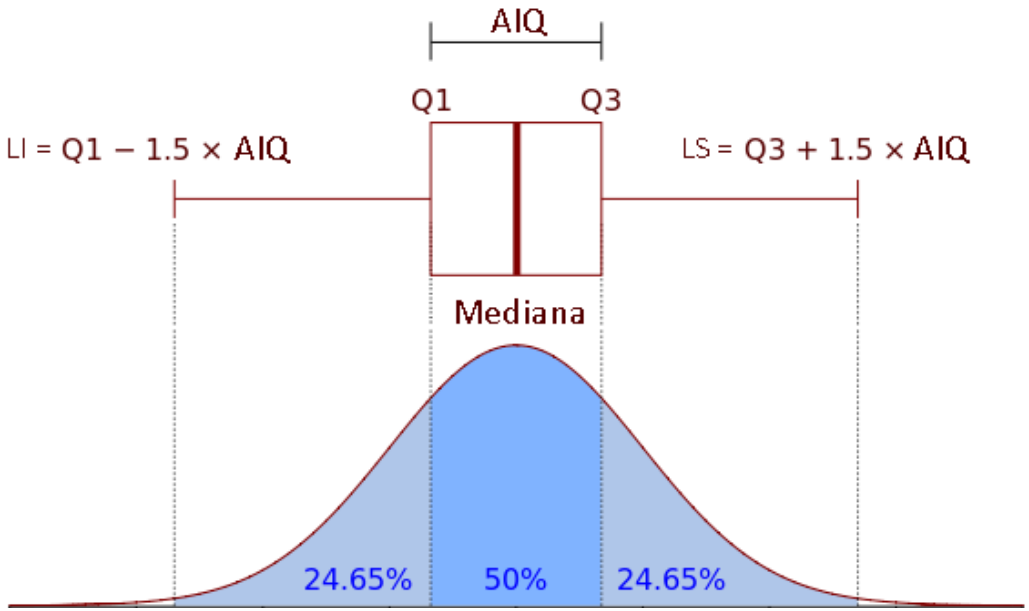
Valores discrepantes. O boxplot indica quais observações podem ser consideradas "discrepantes". As linhas verticais que partem para cima (até o limite superior, *LS*) e para baixo (até o limite inferior, *LI*) são chamadas *whiskers* em inglês, ou *fio do bigode*. Os valores para além dos fios do bigode são considerados valores discrepantes.

Para calcular o comprimento dos fios do bigode

1. Considere a Amplitude do Intervalo Interquartil, isto é, a distância entre o 1° e o 3° quartis:
$$AIQ = Q_3 - Q_1$$
2. O comprimento do fio do bigode é igual a $1,5 \times$ AIQ, ou seja, $1,5 \times (Q_3 - Q_1)$.
3. Dessa maneira, *LS* e *LI* são dados por:
$$\begin{aligned} LS &= Q_3 + 1,5 \times (Q_3 - Q_1) \\ LI &= Q_1 - 1,5 \times (Q_3 - Q_1) \end{aligned}$$
4. Se o valor máximo (mínimo) das observações for menor (maior) do que *LS* (*LI*), então o fio do bigode termina no valor máximo (mínimo)
5. Os pontos acima de *LS* e abaixo de *LI* devem ser marcados como valores discrepantes (conforme figura no *slide* anterior)

Boxplot

Os valores compreendidos entre *LI* e *LS* correspondem a aproximadamente 99,3 % dos dados da distribuição normal.



Segundo esse critério, se a variável em estudo tem distribuição próxima da distribuição Normal, valores discrepantes ocorrem com probabilidade 0,7%.

Referência: `https://pt.wikipedia.org/wiki/Diagrama_de_caixa`

Leitura

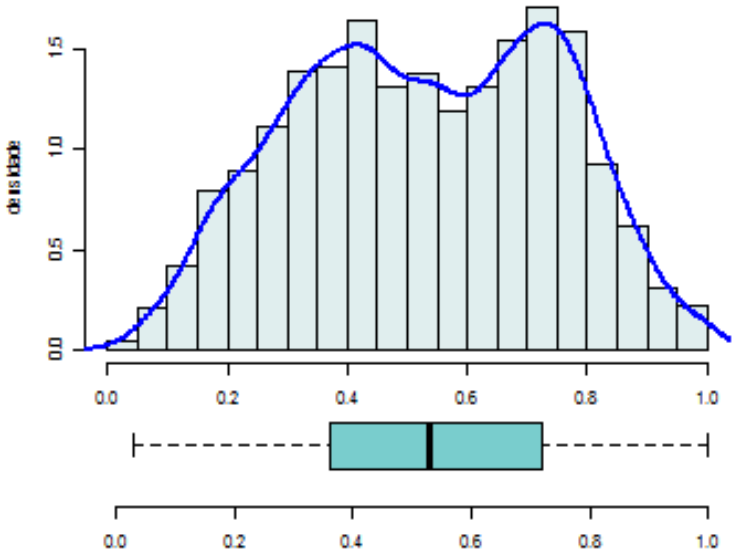
Na wikipedia tem uma boa referência sobre boxplot. Recomendo a leitura.

https://pt.wikipedia.org/wiki/Diagrama_de_caixa

Gráficos para variáveis quantitativas

Boxplot

No exemplo abaixo, temos um *boxplot* da proporção da população de cada município do Brasil que estava efetivamente inscrita no Cadastro Único em 2011.



Gráficos para variáveis quantitativas

Boxplot

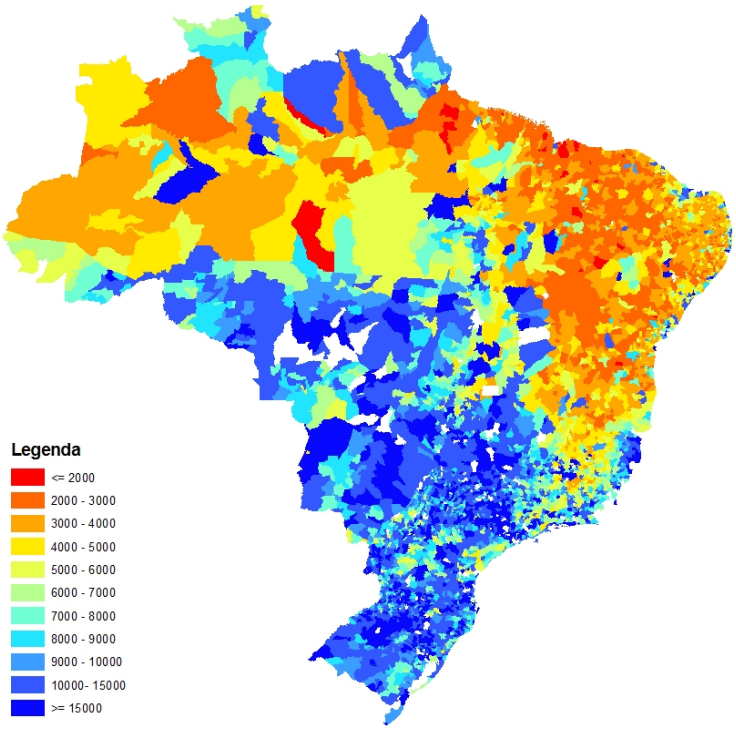
Exemplo

Fazer um *Boxplot* para os dados das notas de estatística.

Mapas

- Oferece um método sem paralelo de exploração de padrões geográficos.
- Recente desenvolvimento de ferramentas tecnológicos acessíveis
- Abundância de dados georreferenciados

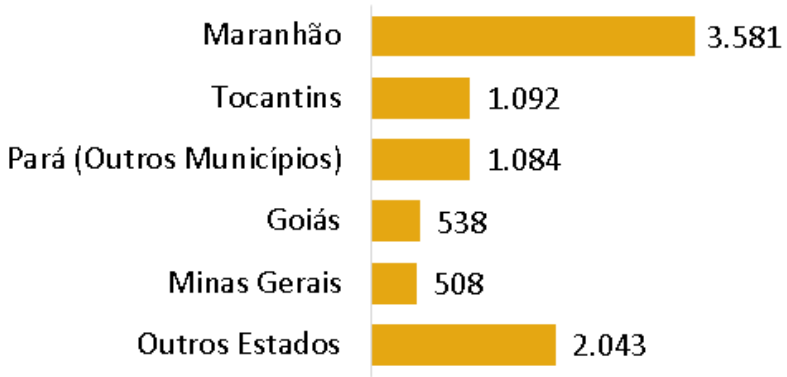
Mapa 1 - Distribuição do PIB *per capita* pelos municípios



Fonte: IBGE

Uso combinado de gráficos

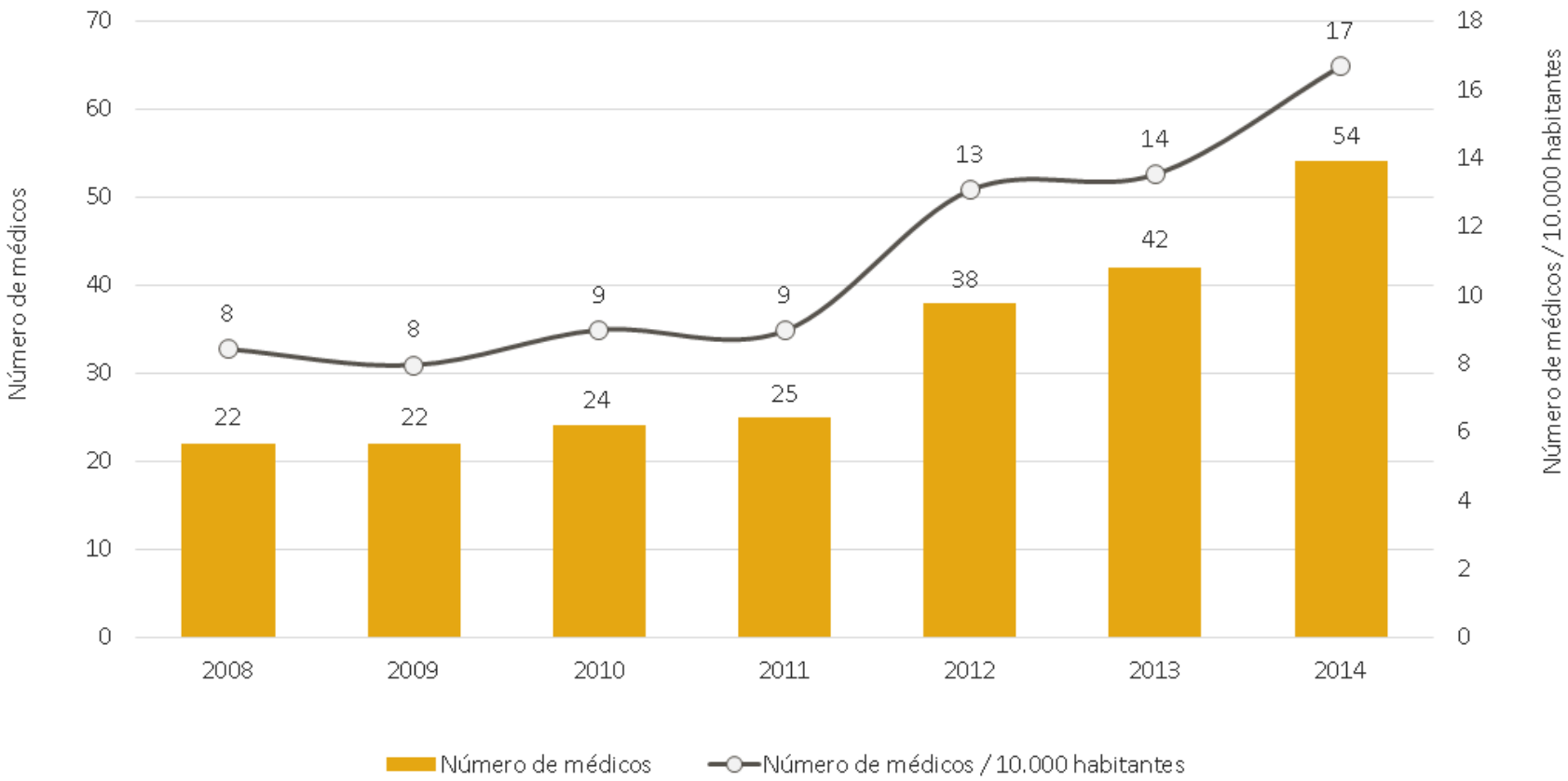
Gráfico 5 - População residente há menos de 10 anos ininterruptos por lugar de residência anterior: Canaã dos Carajás - 2010



Fonte: IBGE/ Censo Demográfico, 2000 e 2010.

Uso combinado de gráficos

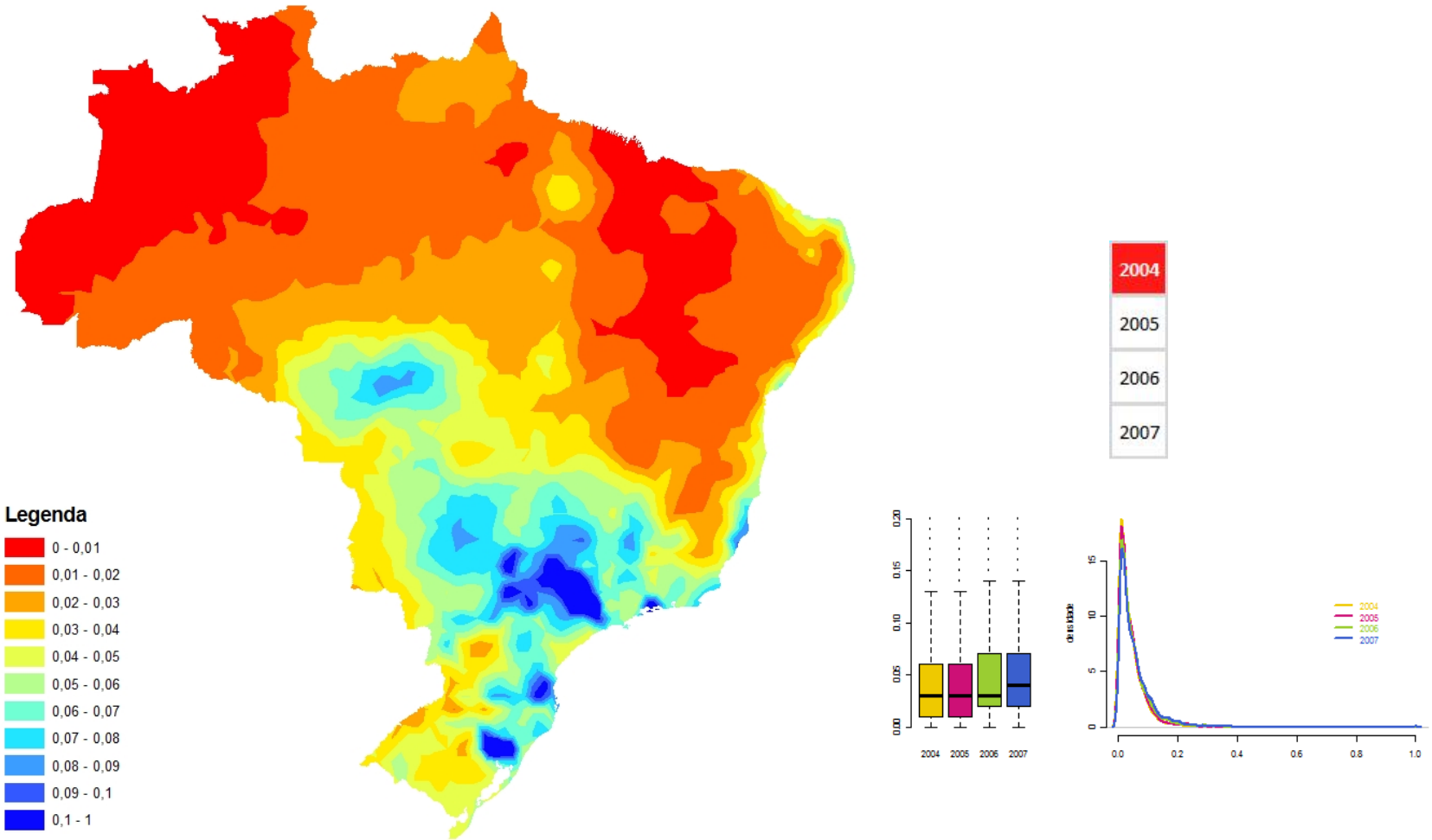
Gráfico 6 - Número de médicos por 10.000 habitantes: Canaã dos Carajás - 2008/2014



Fonte: MS / Cadastro Nacional dos Estabelecimentos de Saúde (CNES); IBGE/Contagem da população, 2007; IBGE/Censo Demográfico, 2010; IBGE/Estimativas da população, 2006, 2008-2009 e 2011-2014.

Uso combinado de gráficos

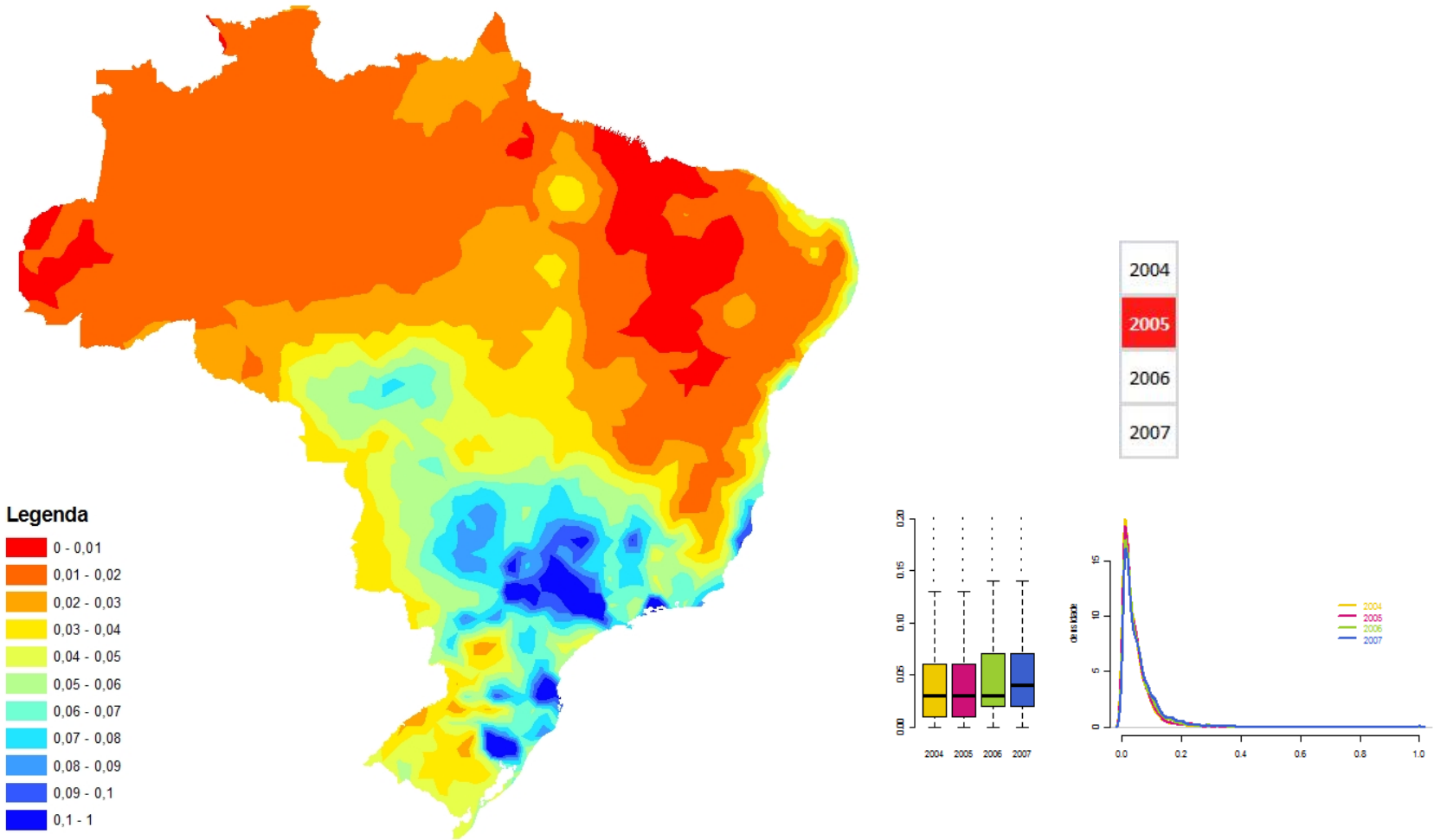
Mapa 1 - Proporção da população CadÚnico com emprego formal



Fonte: MTE/Rais, MDS/CadÚnico

Uso combinado de gráficos

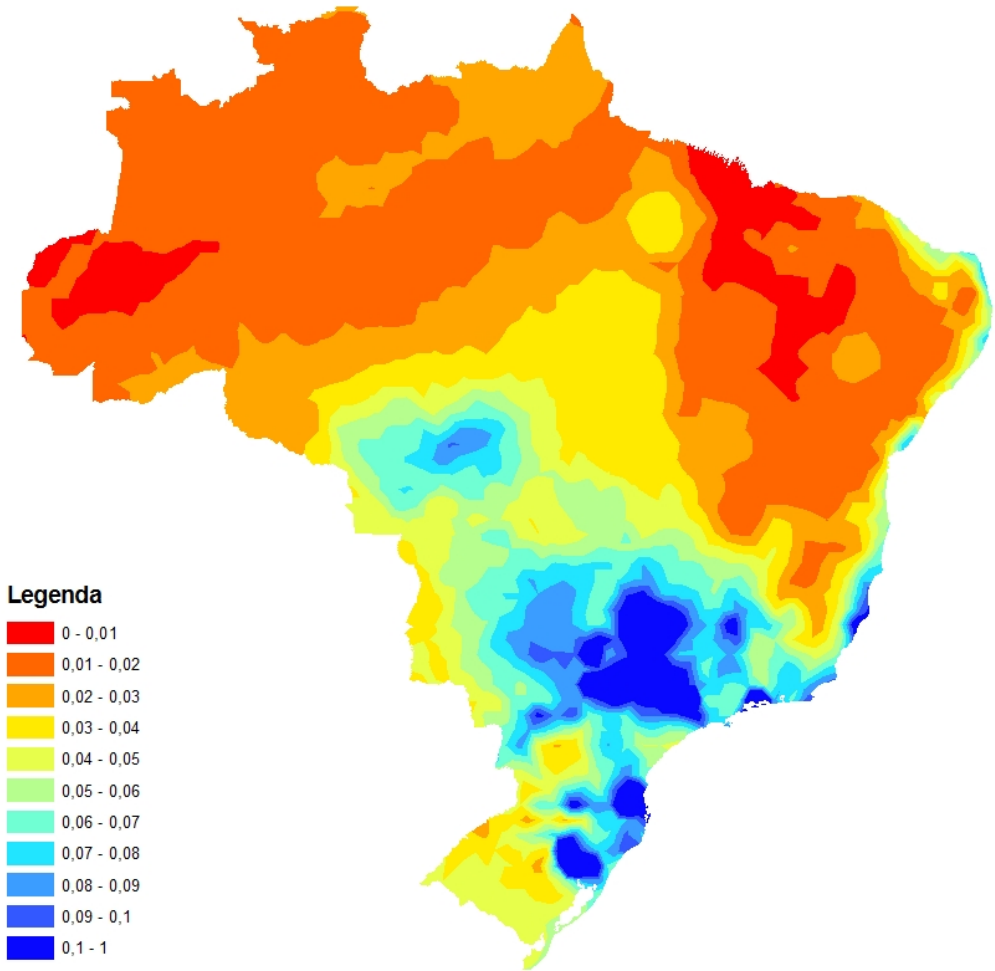
Mapa 1 - Proporção da população CadÚnico com emprego formal

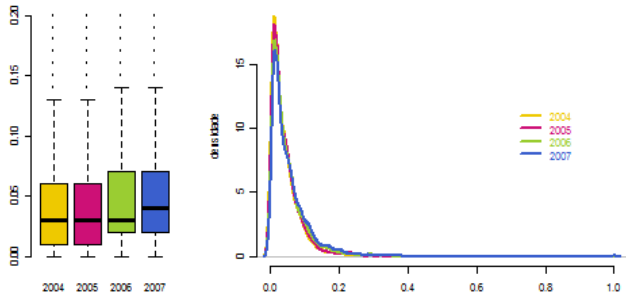


Fonte: MTE/Rais, MDS/CadÚnico

Uso combinado de gráficos

Mapa 1 - Proporção da população CadÚnico com emprego formal

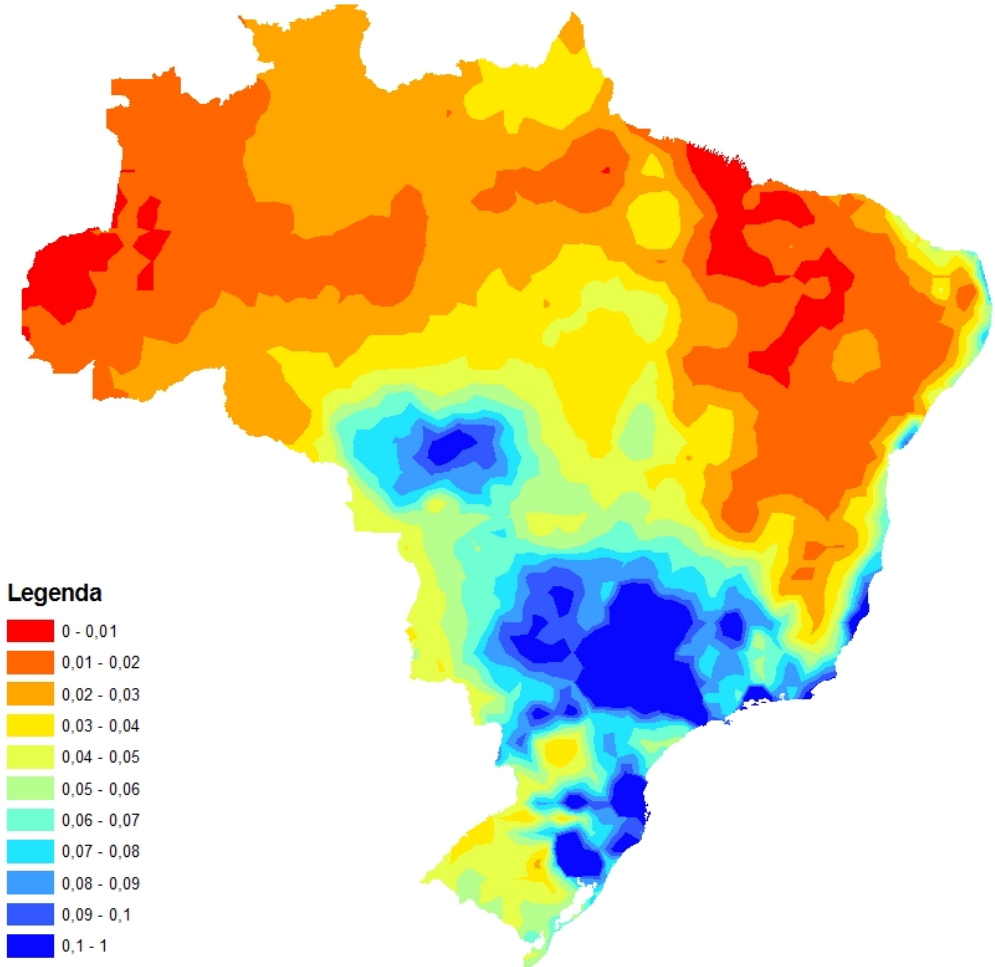


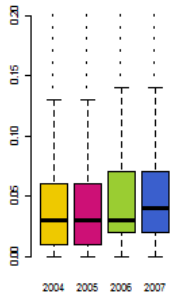


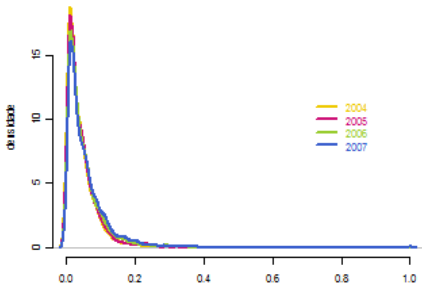
Fonte: MTE/Rais, MDS/CadÚnico

Uso combinado de gráficos

Mapa 1 - Proporção da população CadÚnico com emprego formal







Fonte: MTE/Rais, MDS/CadÚnico

Representação gráfica de dados

Recomendações gerais

- O gráfico deve ter um título explicativo (tanto quanto possível)
- Interessante usar rótulos de dados para destacar uma ou outra informação (se há muitas categorias, o uso de rótulos pode "poluir" o gráfico)
- Sempre mencionar a fonte dos dados. Exemplo:
  - **Fonte**: IBGE/Censo Agrícola (2006)
- Ordenação das respostas: sempre ordenar as respostas
  - Em caso de ordenação natural das respostas (série, porte, etc), usar a ordenação natural
  - Caso contrário, ordenar por frequência ou %
- Quando há muitas categorias, verificar conveniência de agregar um número delas em uma categorias "Outros" (digamos aquelas categorias que juntas somam 5% ou 10%)