Master Project - Data Visualization

Chenghui Zheng

April 9, 2021

Overview

Overview:

- Introduction
- How to make a good graph?
  - Literature Review
  - Fundamentals of Data Visualization (Wilke 2019)
- Visual Inference Protocols
  - Literature Review
  - Line-up Q-Q plot test
- New Formalism of Visual Inference Examples

Introduction

Introduction

- Graph design for data analysis and presentation is largely unscientific. (Cleveland & McGill, 1984)
- Much of the early experimentation regarding the accuracy of graphical forms was based in psychophysics research on the perception of size and shape. (Vanderplas, Cook & Hofmann, 2020)
- Deficient data visuals can reduce the quality and impede the progress of scientific research. (Mason, 2019)

How to make a good graph?

How to make a good graph?

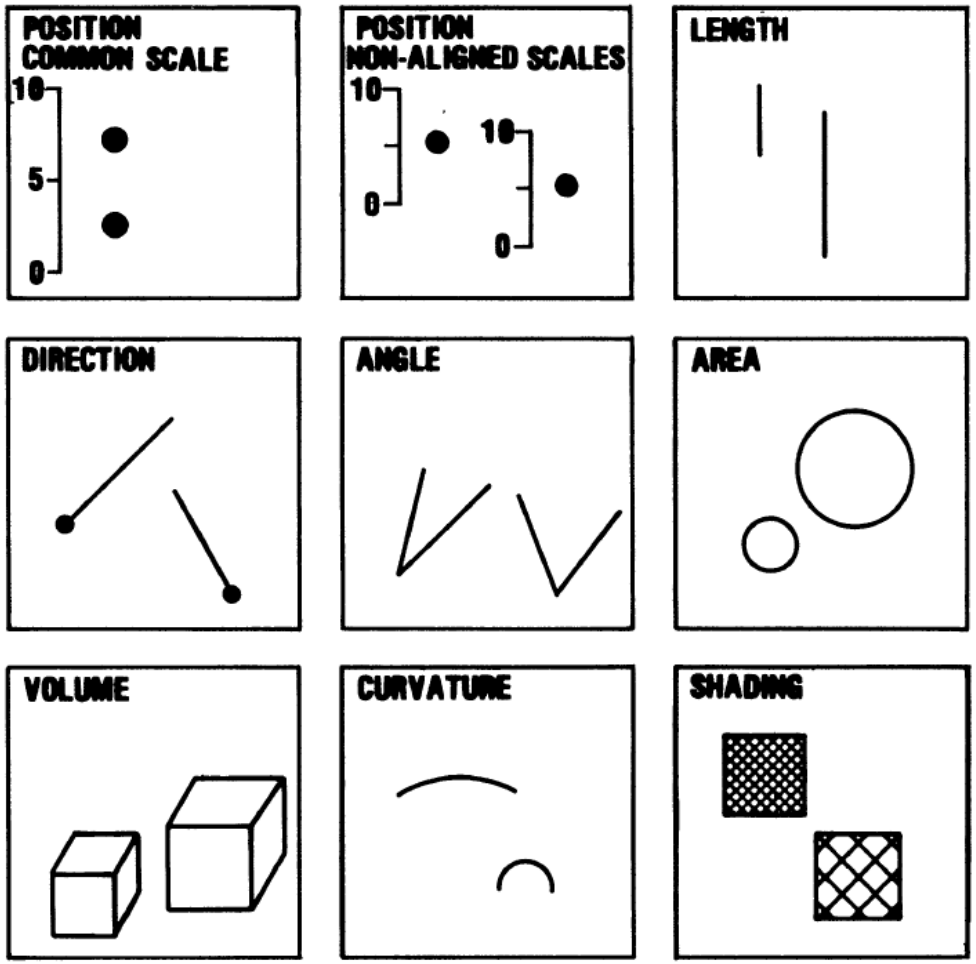
Literature Review

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods [Cleveland and McGill, 1984]

Graph Construction:

Elementary perceptual tasks: quantitative information extracted from a graph.

The list of 10 elementary perceptual tasks



The order of 10 elementary perceptual tasks

From most to least accurate:

1. Position along a common scale;
2. Positions along nonaligned scales;
3. Lengths, direction, angle;
4. Area;
5. Volume, curvature;
6. Shading, color saturation.

Example (Wilke 2019): position along a common scale vs lengths

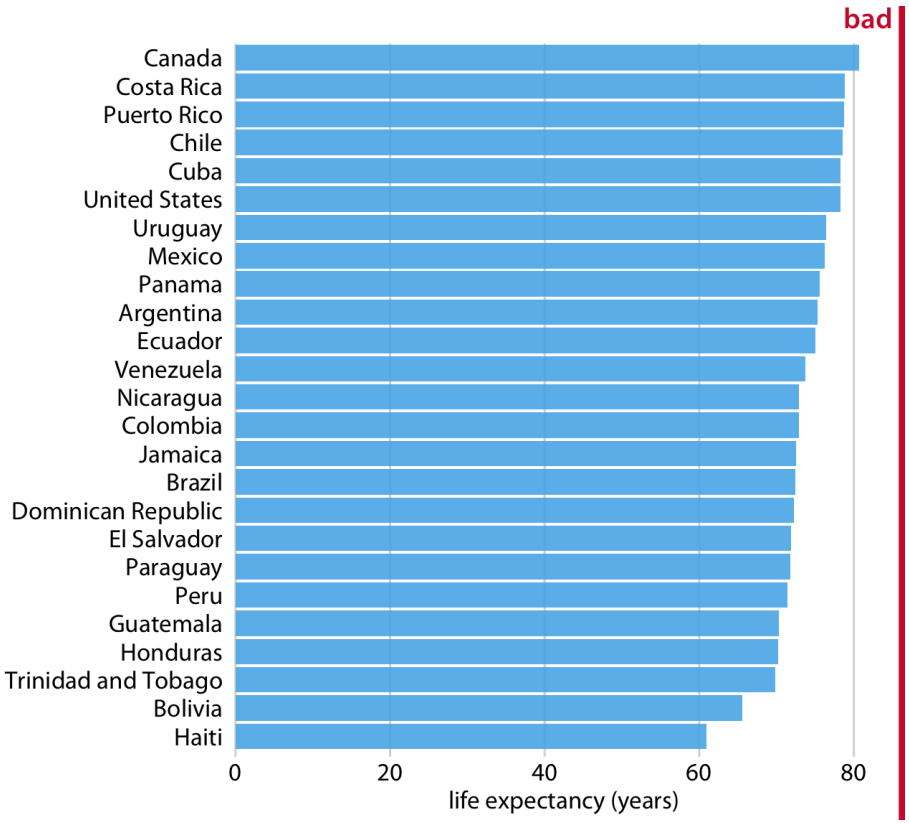


Figure: The bar plot of life expectancies of countries in 2007

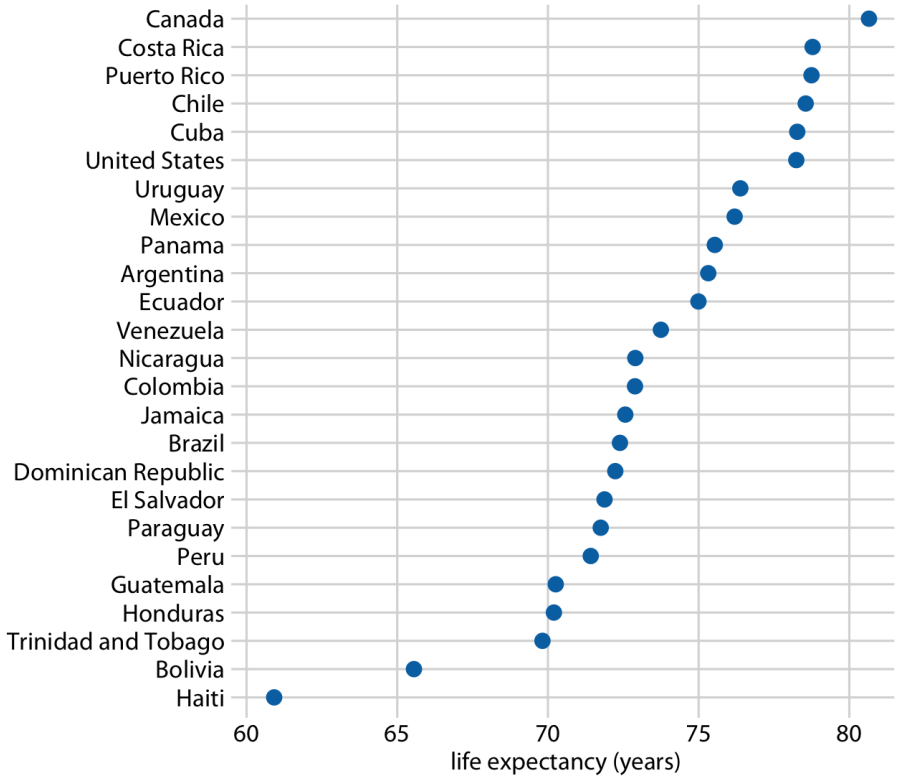


Figure: The dot plot of life expectancies of countries in 2007

Examples (Cleveland & McGill): angle vs positions

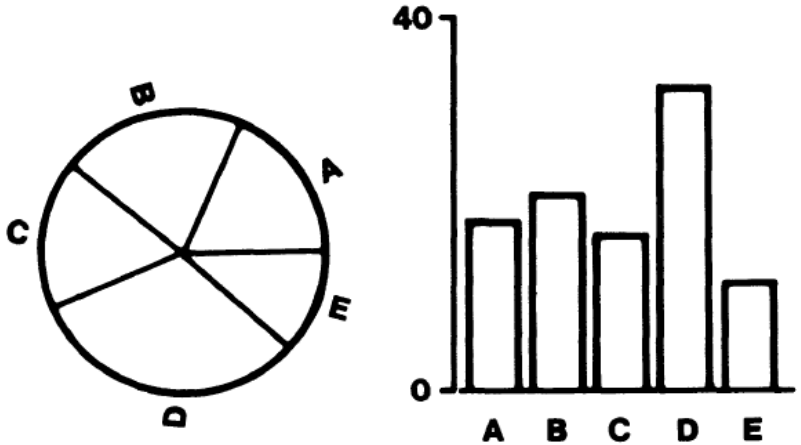


Figure: Comparison between pie chart and bar chart

How to make a good graph?

Literature Review

What makes a good graph? [Vanderplas, Cook & Hofmann, 2020]

Graphical testing methods $\begin{cases} \text{Explicitly structured graphical test} \\ \text{Implicitly structured graphical test} \end{cases}$

Explicit graphical tests require participants to answer specific questions about graphical objects. Implicit graphical tests require participants to identify both the purpose and function of the plot and use that information to evaluate the plot as shown. Line-up protocol is the implicit graphical test.

How to make a good graph?

Literature Review

What makes a good graph? [Vanderplas, Cook & Hofmann, 2020]

$$\text{Graphical testing methods} \begin{cases} \text{Explicitly structured graphical test} \\ \text{Implicitly structured graphical test} \end{cases}$$

Explicit graphical tests require participants to answer specific questions about graphical objects. Implicit graphical tests require participants to identify both the purpose and function of the plot and use that information to evaluate the plot as shown. Line-up protocol is the implicit graphical test.

Types of plots for visualization

Fundamentals of Data Visualization (Wilke 2019)

Visualizing amounts $\begin{cases} \text{grouped/stacked bar plot} \\ \text{dot plot/ heat map} \end{cases}$

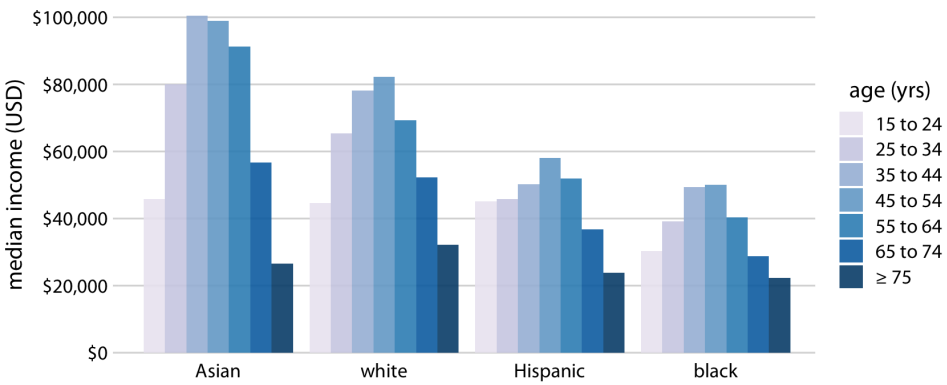


Figure: Median household annual income in 2016

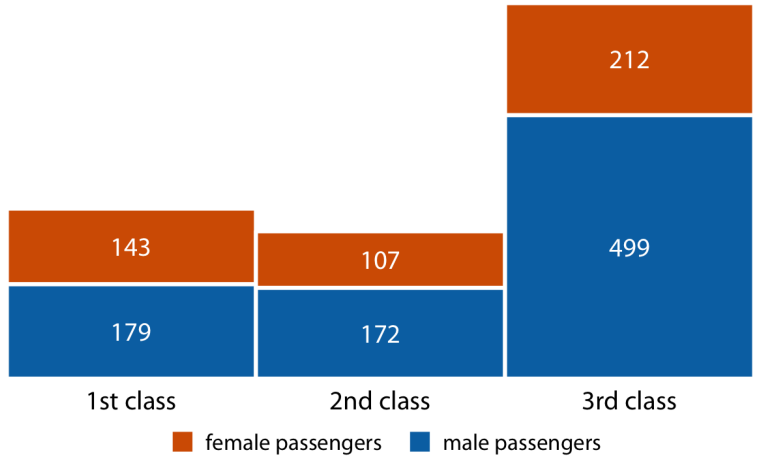


Figure: Number of passengers in Titanic

Types of plots for visualization

Fundamentals of Data Visualization (Wilke 2019)

Visualizing distribution { histogram, density plot

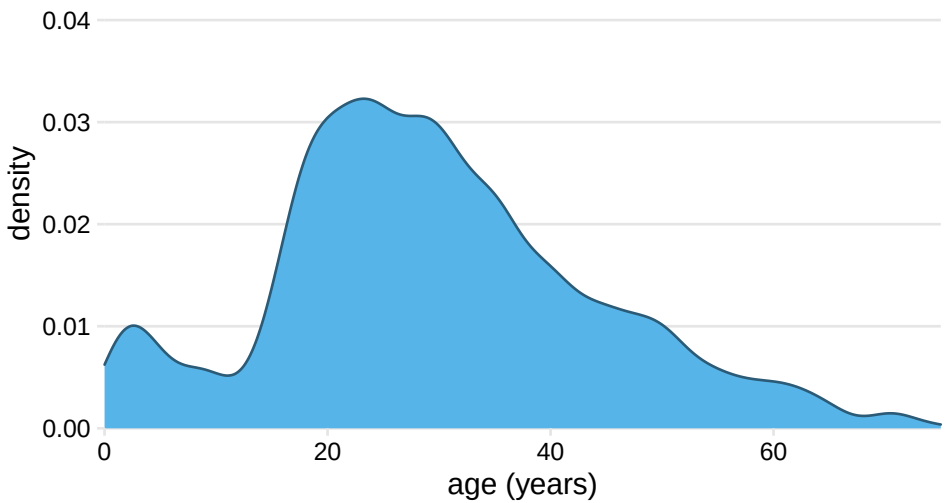


Figure: age distribution of Titanic passengers

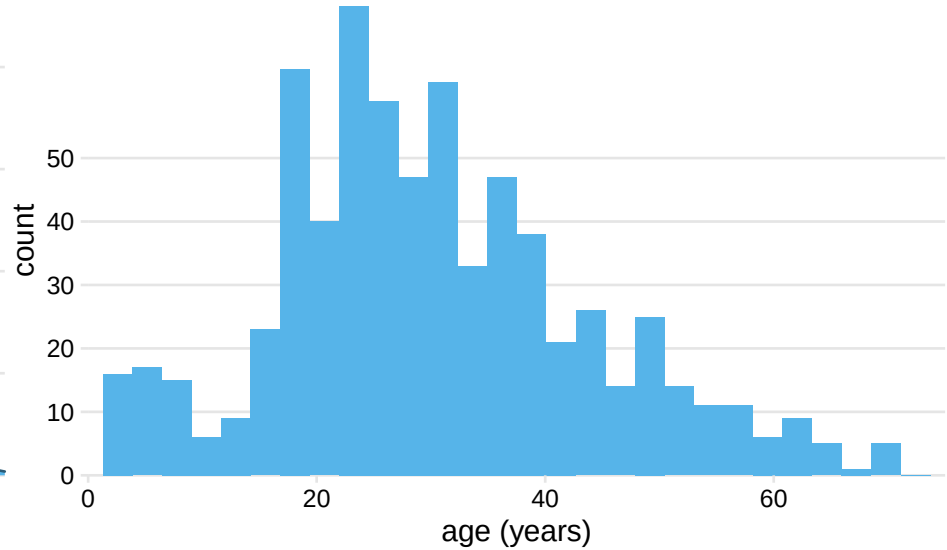


Figure: Age distribution of Titanic passengers

Types of plots for visualization

Fundamentals of Data Visualization (Wilke 2019)

Visualizing nested proportions { Mosaic, Treemap

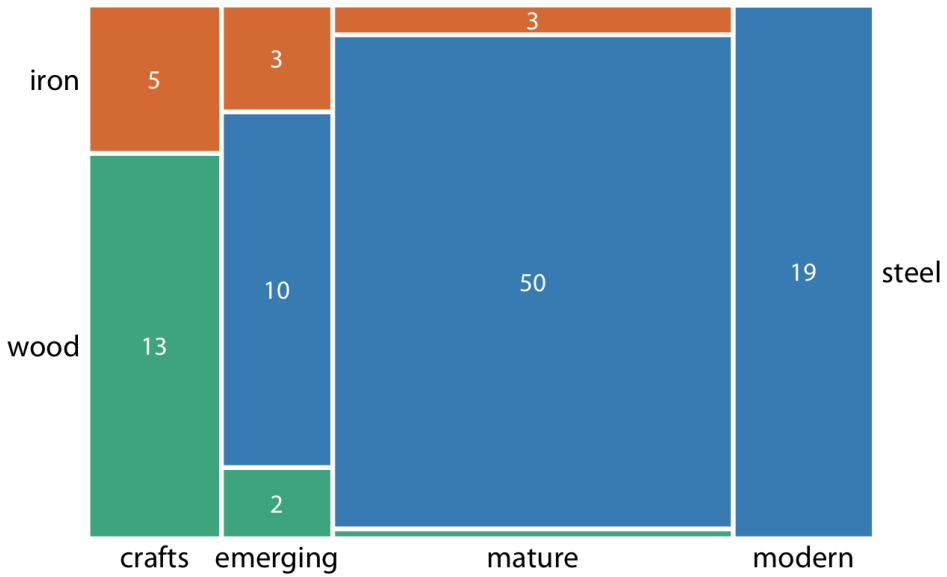


Figure: Mosaic for bridge construction materials & era

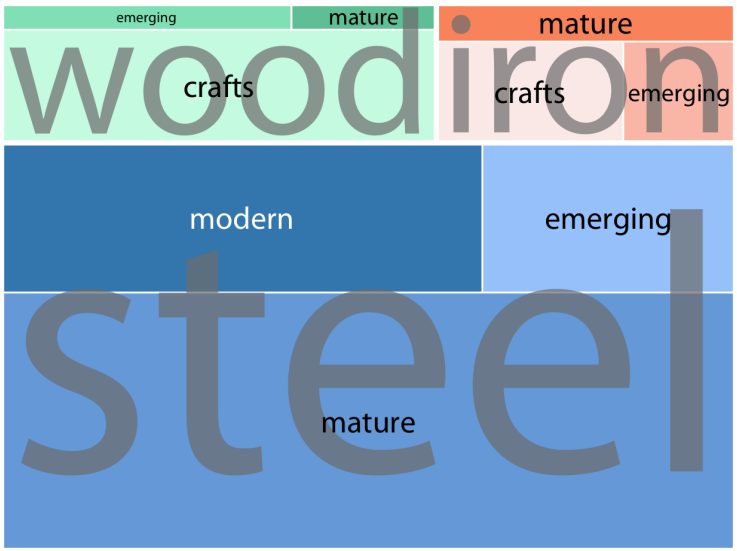


Figure: Treemap for ridge construction materials & era

Types of plots for visualization

Fundamentals of Data Visualization (Wilke 2019)

Visualizing associations { Scatterplot
slopegraph (paired data)

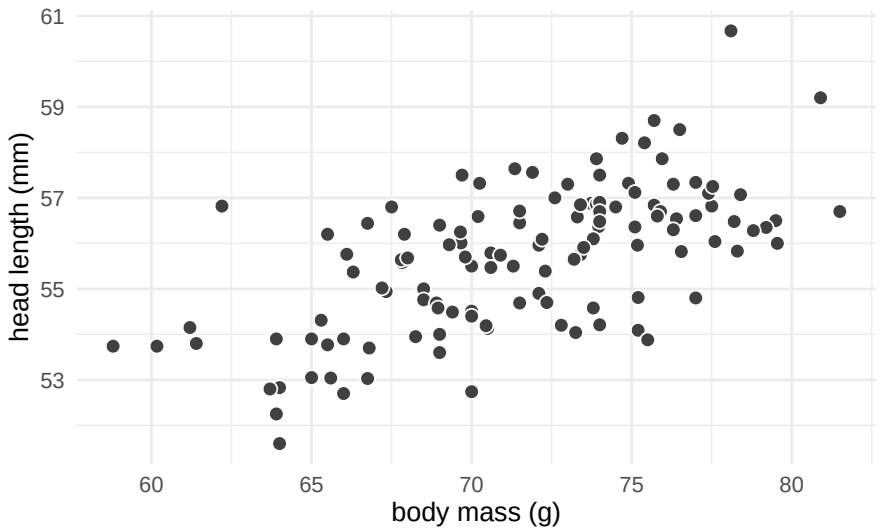


Figure: Head length versus body mass for blue jays

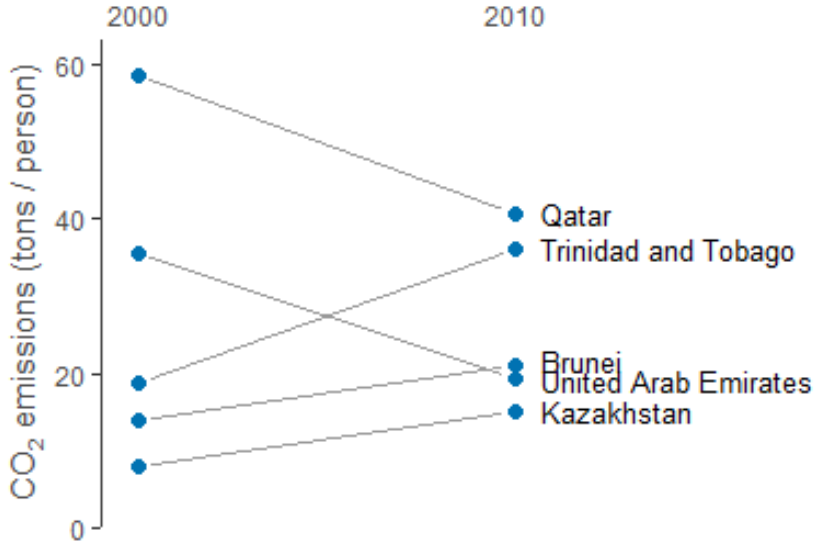


Figure: CO2 emission per capita in 2000 & 2010

Visual Inference Test Protocols

Visual inference test protocols

Literature review

Statistical inference for exploratory data analysis and model diagnostics [Buja.etc, 2009]

Rorschach protocol:
To measure a data analyst's tendency to over-interpret plots in which there is only spurious structure.

Line-up protocol:
To assess the significance of visual discoveries. The interest is the probability of singling out the actual data plot.

Visual inference test protocols

Literature review

Statistical inference for exploratory data analysis and model diagnostics [Buja.etc, 2009]

Rorschach protocol:
To measure a data analyst's tendency to over-interpret plots in which there is only spurious structure.

Line-up protocol:
To assess the significance of visual discoveries. The interest is the probability of singling out the actual data plot.

Line-up test

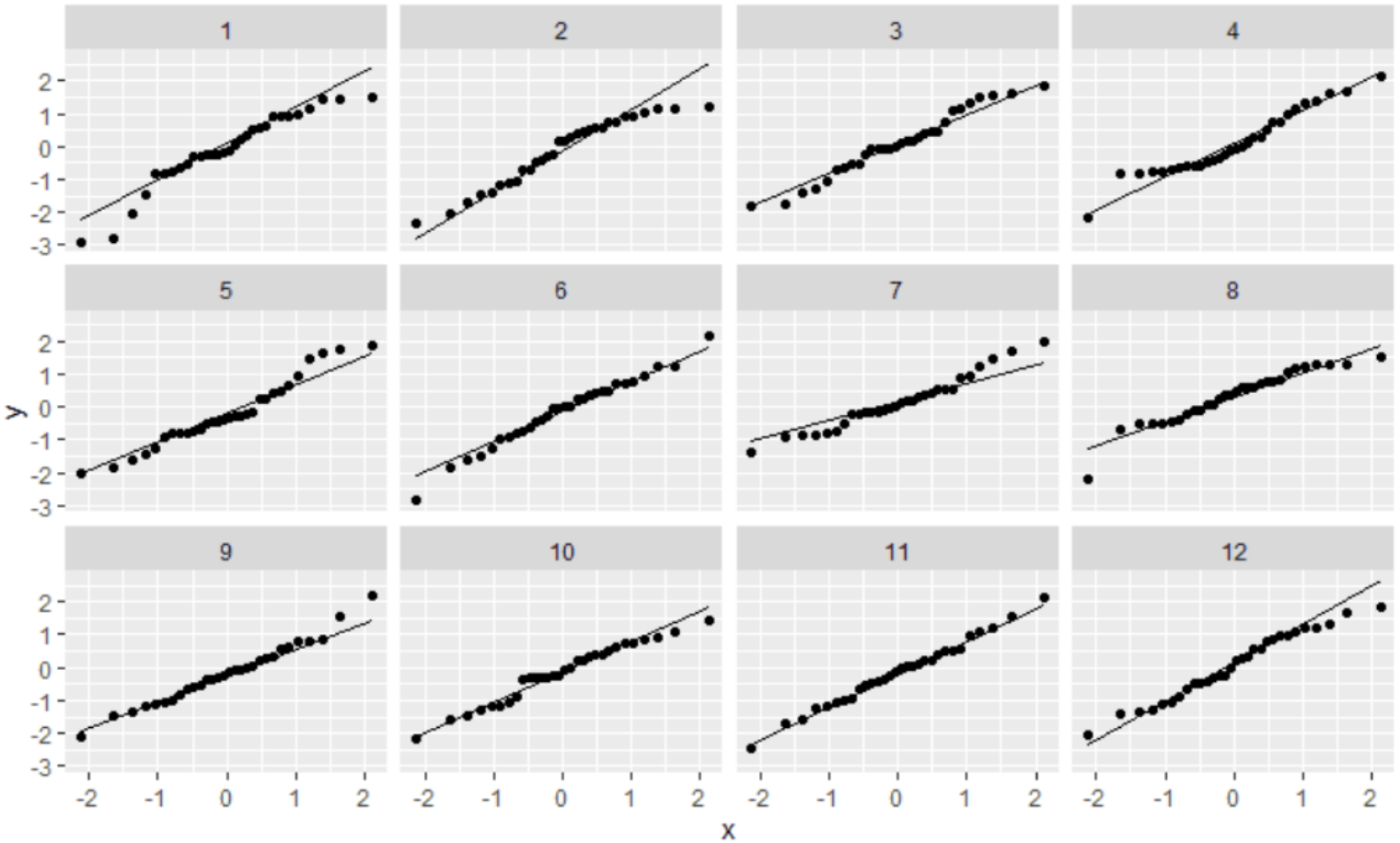


Figure: Line-up normal q-q test for contaminated normal

Line-up test

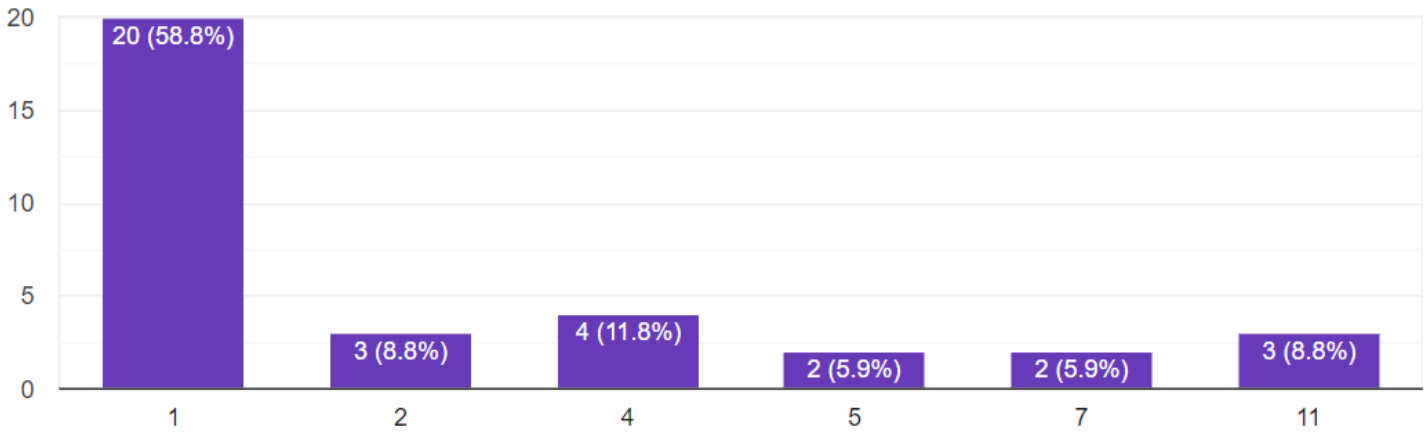


Figure: Result of Line-up normal q-q test for contaminated normal

Not all data points will be perfectly aligned with the normal q-q line, even if they are generated from N(0,1).

Line-up test

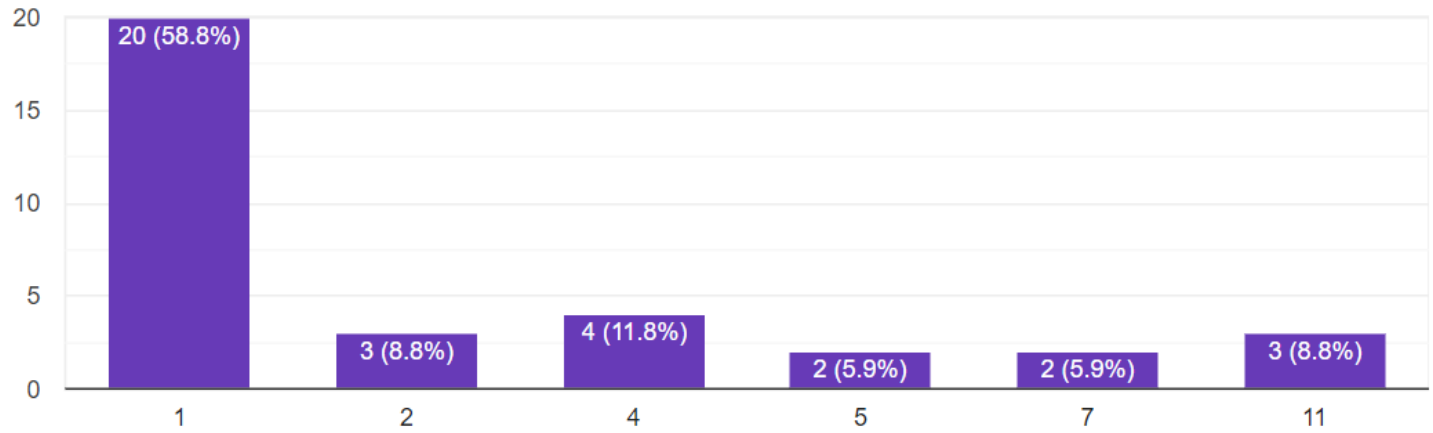


Figure: Result of Line-up normal q-q test for contaminated normal

Not all data points will be perfectly aligned with the normal q-q line, even if they are generated from N(0,1).

Line-up protocol

Validation of Visual Statistical Inference, Applied to Linear Models
[Majumder, Hofmann & Cook, 2013]

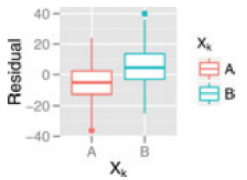
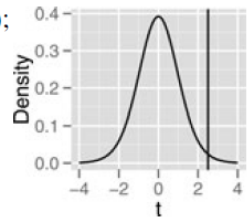
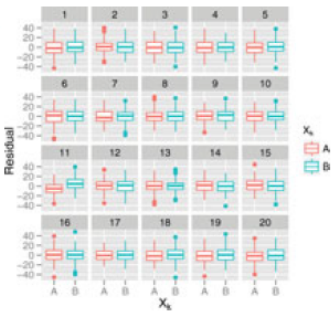
	Conventional inference	Lineup protocol
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$
	↓	↓
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	Actual T is extreme	Actual plot is identifiable

Figure: Comparison between conventional and visual inference

New Formalism of Visual Inference

Visual Inference

In the `ggplot` implementation of the grammar of graphics, each layer of the visual statistic is composed of *statistical transformation* and *geometric elements*.

Layers of visual statistic

The statistical transformation, `stat`, or $S(\boldsymbol{X})$, maps the columns of the data table $\boldsymbol{X}$ to a lower-dimensional summary.
The geometric elements, `geom` G, create graphical objects for the plots, such as polygons, area, lines, points, etc.

Thus, a visual statistic V can be formally expressed as:

$$V = G \circ S(\boldsymbol{X})$$

Visual Inference

In the `ggplot` implementation of the grammar of graphics, each layer of the visual statistic is composed of *statistical transformation* and *geometric elements*.

Layers of visual statistic

The statistical transformation, `stat`, or $S(\boldsymbol{X})$, maps the columns of the data table $\boldsymbol{X}$ to a lower-dimensional summary.
The geometric elements, `geom` G, create graphical objects for the plots, such as polygons, area, lines, points, etc.

Thus, a visual statistic V can be formally expressed as:

$$V = G \circ S(\boldsymbol{X})$$

Visual Inference

In the `ggplot` implementation of the grammar of graphics, each layer of the visual statistic is composed of *statistical transformation* and *geometric elements*.

Layers of visual statistic

The statistical transformation, `stat`, or $S(\boldsymbol{X})$, maps the columns of the data table $\boldsymbol{X}$ to a lower-dimensional summary.
The geometric elements, `geom` G, create graphical objects for the plots, such as polygons, area, lines, points, etc.

Thus, a visual statistic V can be formally expressed as:

$$V = G \circ S(\boldsymbol{X})$$

Visual Inference

A visual statistic V will also be augmented with other elements such as annotations, context, rulers or other navigation devices. We denote these by $A(\boldsymbol{W})$ where $\boldsymbol{W}$ is auxiliary data sometimes but not necessarily the original data $\boldsymbol{X}$. The additional augmented visual statistic $\tilde{V}$ is now expressed as

$$\tilde{V} = V \circ (A|V).$$

A further augmentation, $\mathcal{A}$, that does not depend on $\boldsymbol{X}$ or $\boldsymbol{W}$ is used to create an attractive plot. The visual statistic in its complete form is

$$V_{complete} = \tilde{V} \circ \mathcal{A} = V \circ (A|V) \circ \mathcal{A}.$$

Visual Inference

A visual statistic V will also be augmented with other elements such as annotations, context, rulers or other navigation devices. We denote these by $A(\boldsymbol{W})$ where $\boldsymbol{W}$ is auxiliary data sometimes but not necessarily the original data $\boldsymbol{X}$. The additional augmented visual statistic $\tilde{V}$ is now expressed as

$$\tilde{V} = V \circ (A|V).$$

A further augmentation, $\mathcal{A}$, that does not depend on $\boldsymbol{X}$ or $\boldsymbol{W}$ is used to create an attractive plot. The visual statistic in its complete form is

$$V_{complete} = \tilde{V} \circ \mathcal{A} = V \circ (A|V) \circ \mathcal{A}.$$

Examples:

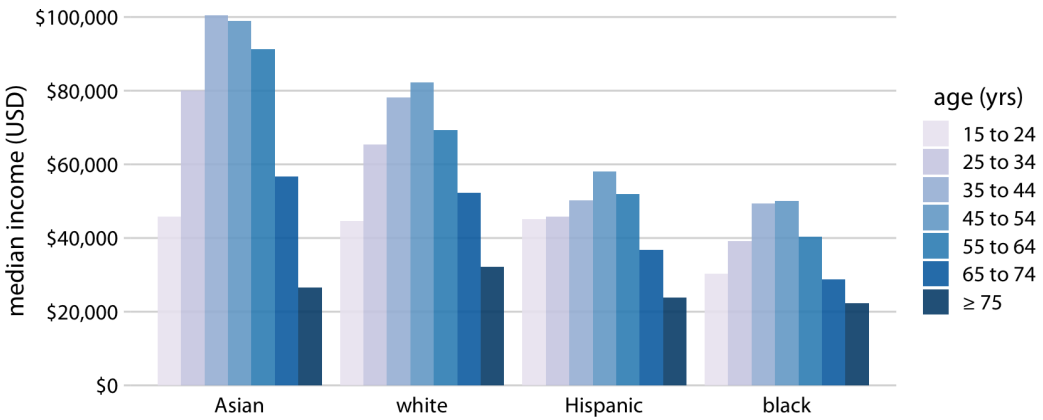


Figure: Grouped bar plot for median household annual income in 2016

age	race	median_income
15 to 24	black	30267
25 to 34	black	39176
35 to 44	black	49336
45 to 54	black	50103
55 to 64	black	40363
65 to 74	black	28697
>74	black	22302
15 to 24	asian	45809
25 to 34	asian	80098
35 to 44	asian	100443

Figure: Data table for median household annual income in 2016

Example:

```
ggplot(income_df, aes(x = race, y = median_income, fill = age)) +
  geom_col(position = "dodge", alpha = 0.9) +
  scale_y_continuous(
    expand = c(0, 0),
    name = "median income (USD)",
    breaks = c(0, 20000, 40000, 60000, 80000, 100000),
    labels = c("$0", "$20,000", "$40,000", "$60,000", "$80,000", "$100,000")
  ) +
  scale_fill_manual(values = colors_seven, name = "age (yrs)") +
  coord_cartesian(clip = "off") +
  xlab(label = NULL) +
  theme_minimal() +
  theme(
    axis.line.x = element_blank(),
    axis.ticks.x = element_blank(),
    legend.title.align = 0.5
  ) -> p_income_age_dodged
p_income_age_dodged
```

Figure: Code for grouped bar plot

Example:

The bar plot is generated by mapping race, $X_{.2}$, to x axis, median income, $X_{.3}$, to y axis and age group, $X_{.1}$, to *fill*. The `geom` is `col(position = "dodge")` since we want the heights of the bars to represent the median income of each age group in each race.

$$V_C = G_C \circ S_C, G_C = \text{col}(\text{position} = \text{“dodge”}),$$

$$S_C(X) = (X_{.2}, X_{.3}, X_{.1}) \rightarrow SG_C = (x, y, \mathit{fill}).$$

The augmentation, $\mathcal{A}$, includes the `theme`, `labels`, `breaks`,etc. The final plot is expressed as

$$V_{complete} = V_C \circ \mathcal{A}.$$

Example:

The bar plot is generated by mapping race, $X_{.2}$, to x axis, median income, $X_{.3}$, to y axis and age group, $X_{.1}$, to *fill*. The `geom` is `col(position = "dodge")` since we want the heights of the bars to represent the median income of each age group in each race.

$$V_C = G_C \circ S_C, G_C = \text{col}(\text{position} = \text{“dodge”}),$$

$$S_C(X) = (X_{.2}, X_{.3}, X_{.1}) \rightarrow SG_C = (x, y, \mathit{fill}).$$

The augmentation, $\mathcal{A}$, includes the `theme, labels, breaks,`etc. The final plot is expressed as

$$V_{complete} = V_C \circ \mathcal{A}.$$

References

- A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, andH. Wickham. Statistical inference for exploratory data analysis and modeldiagnostics. *Philosophical Transactions of the Royal Society A: Mathemati-cal, Physical and Engineering Sciences*, 367(1906):4361–4383, Nov. 2009.
- W. S. Cleveland and R. McGill. Graphical perception: theory, experimen-tation, and application to the development of graphical methods.*Journal of the American Statistical Association*, 79(387), 1984.
- M. Majumder, H. Hofmann, and D. Cook. Validation of visual statisti-cal inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, Sept. 2013.

References

- B. Mason. Why scientists need to be better at data visualization.*Knowable Magazine*, Nov. 2019.
- S. Vanderplas, D. Cook, and H. Hofmann. Testing statistical charts: what makes a good graph? *Annual Review of Statistics and Its Application*,7(1):61–88, Mar. 2020.
- S. Vanderplas, C. R¨ottger, D. Cook, and H. Hofmann. Statistical significance calculations for scenarios in visual inference. Nov. 2020.
- H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inferencefor infovis. *IEEE Transactions on Visualization and Computer Graphics*,16(6):973–979, Nov. 2010.
- Wilke, C. O. Fundamentals of data visualization. 2019

Graphical testing methods

Explicitly structured graphical test:

- Preattentive perception;
  -searching tasks : size, shape, angle and color
- Attention mediated;
  - Direct observation;
  - Psychophysics methodology;
    - Constant stimuli;
      - repeatedly presented with charts and a particular question
    - Method of adjustment
      - adjust the stimuli interactively
  - Think aloud
    - Concurrent think-aloud(CTA);
    - Retrospective think-aloud(RTA)

Implicitly structured graphical test: Line-up protocol

Graphical testing methods

Explicitly structured graphical test:

- Preattentive perception;
  -searching tasks : size, shape, angle and color
- Attention mediated;
  - Direct observation;
  - Psychophysics methodology;
    - Constant stimuli;
      - repeatedly presented with charts and a particular question
    - Method of adjustment
      - adjust the stimuli interactively
  - Think aloud
    - Concurrent think-aloud(CTA);
    - Retrospective think-aloud(RTA)

Implicitly structured graphical test: Line-up protocol

Graphical testing methods

Explicitly structured graphical test:

- Preattentive perception;
  -searching tasks : size, shape, angle and color
- Attention mediated;
  - Direct observation;
  - Psychophysics methodology;
    - Constant stimuli;
      - repeatedly presented with charts and a particular question
    - Method of adjustment
      - adjust the stimuli interactively
  - Think aloud
    - Concurrent think-aloud(CTA);
    - Retrospective think-aloud(RTA)

Implicitly structured graphical test: Line-up protocol

Graphical testing methods

Explicitly structured graphical test:

- Preattentive perception;
  -searching tasks : size, shape, angle and color
- Attention mediated;
  - Direct observation;
  - Psychophysics methodology;
    - Constant stimuli;
      - repeatedly presented with charts and a particular question
    - Method of adjustment
      - adjust the stimuli interactively
  - Think aloud
    - Concurrent think-aloud(CTA);
    - Retrospective think-aloud(RTA)

Implicitly structured graphical test: Line-up protocol

Graphical testing methods

Explicitly structured graphical test:

- Preattentive perception;
  -searching tasks : size, shape, angle and color
- Attention mediated;
  - Direct observation;
  - Psychophysics methodology;
    - Constant stimuli;
      - repeatedly presented with charts and a particular question
    - Method of adjustment
      - adjust the stimuli interactively
  - Think aloud
    - Concurrent think-aloud(CTA);
    - Retrospective think-aloud(RTA)

Implicitly structured graphical test: Line-up protocol