Visualizing numerical data

to identify the explanatory variable, we identify which of the two is suspected in affecting the other.

observational data → correlation (not causation)

Evaluating the relationship:

- direction - shape - strength - outliers

skewness, modality

histogram: provides a view of the data density

left skewed distribution: longer tail is on the left.

right skewed distribution: longer tail is on the right.

Modality:

unimodal: just one prominent peak.

bimodal: just two.

uniform: no prominent peaks.

multimodal: more than two.

Note: the ideal bin width depends on the data you're working with.

Visualization methods: histogram, dot plot, box plot, intensity map.

Measures of center

Mean: arithmetic average

Median: midpoint of the distribution

Mode: most frequent observation

Measures of spread

Variability in the data.

Variance: the average squared deviation from the mean
(s^2, σ^2)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ → deviation from the $\bar{x}$ for each observation.

Standard deviation: the average deviation around the mean.
(s, σ)

$$s = \sqrt{s^2}$$

Note: set with more data away from the center, is more variable.

Interquartile range: the range of the middle 50% of the data.

$$Q3 - Q1 = 75\% - 25\%$$

Robust Statistics: measures on which extreme observations have little effect.

	robust	non-robust
skewed	median	mean
	iqr	sd, range

symmetric

Transforming data

transformation: rescaling of the data using a function

- to see the data structure differently
- to reduce skew

Exploring Categorical variables

frequency table

bar plot

just categorical variable.

categories can be listed in any order.

contingency table

segmented bar plot

conditional frequency distributions

mosaic plot

shows the marginal distri

categorical vs numerical: side-by-side box plots.

Introduction to inference

Null hypothesis: "there is nothing going on"

- independent variables
- the difference was just by chance

Alternative hypothesis.

Conduct a hypothesis test under the assumption that the null hypothesis is true.

Note: if there is not enough evidence, we fail to reject the null hypothesis.

Making a decision: results from the simulations look

like the data, by chance (independent)

the probability of observing data under the assumption that the null hypothesis is true, is called the p-value.

Last modified: Jun 16, 2020