5-探索性数据分析

探索性数据分析（exploratory data analysis，EDA）是由John Tukey提出并推广一个统计学概念，它是指通过分析数据集来概要的了解数据的主要特点，通常使用描述性统计比如趋中性、变异型等，以及一些可视化的方式。可以使用一些统计模型，但是在探索性数据分析中，我们主要考察的是在形式统计模型及假设验证之外，数据能够告诉我们一些新信息。

探索性数据分析和数据初步分析（initial data analysis，IDA）不同。数据初步分析，关注比较狭窄，主要是考察数据是否符合模型和假设验证的条件，以及处理缺失数据并进行一些转换。探索性数据分析包含了初步数据分析。

探索性数据分析的目的

揭示观察到的现象的起因提出假设

综合评价推断性统计模型的假设是否成立

验证统计工具和方法对选择是否正确

为进一步数据收集提供一定的基础

本章通过实验性数据来展示不同变量的探索性数据分析过程。由于数据本身分为分类型数据和连续性数据因此描述这些数据的变量也可以分为分类型变量和连续型变量。由于这两类变量所使用的描述性统计指标有所差异以及相应的可视化方法有所不同因此我们首先讨论翻译分类变量和单一连续进变量。之后我们再讨论两个分类变量以皮一个分类变量和一个连续变量以及两个连续性变量。

## Categorical variable

分类型变量

The best way to characterizing categorical variables is via the frequency. The \*frequency\* is the number of times a particular value for a variable (data item) has been observed to occur.

描述分类型变量最好的方法是通过频率频率是指某一个数值或变量观测值出现的次数

绝对频率或者相对频率平路的表示方法

平日的表示方法频率可以表示成不同的形式绝对频率是指某一个观测值出现的次数或者说技术

### Absolute v.s relative frequency?

The frequency of a value can be expressed in different ways:

\* The \*absolute frequency\* describes the number of times a particular value for a variable occurs. Or simply put, counts.

\* The \*relative frequency\* describes the number of times a particular value for a variable occurs in relation to the total number of values for that variable.It is calculated by dividing the absolute frequency by the total number of values for the variable. Ratios, rates, proportions and percentages are different ways of expressing relative frequencies.

相对频率是指对于某一个变量来说某一个观测值出现的次数和所有类型的观测值总数的比例关系。相对频率可以用绝对频率除以变量总观测值数。比如一个单词在一个语料库中出现了200次这就是一个绝对评述。如果这个语料库有1万个词那么这个单词的相对频率可以用200除以1万。因此相对频率可以表述成为比例百分比等。

- A \*ratio\* compares the frequency of one value for a variable with another value for the variable (1st value : 2nd value).200:10000

- A \*rate\* is a measurement of one value for a variable in relation to another measured quantity. 200/10000

- A \*proportion\* describes the share of one value for a variable in relation to a whole.It is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of values in the population.

0.02

- A \*percentage\* expresses a value for a variable in relation to a whole population as a fraction of one hundred.

2%

### Frequency distributions

\*Frequency distributions\* are visual displays that organise and present frequency counts so that the information can be interpreted more easily.

评述的分布频数的分布是一种可视化的方法可以用来展示评述使得理解更加容易最好

A frequency distribution of data can be shown in a table or graph. Some common methods of showing frequency distributions include \*frequency tables\*, \*bar charts\* or \*histograms\*.

频数分布的展示方法可以有频数表格、条状图和直方图三种。

频数表格是最简单的展示某一个只出现次数的方法。

通常分类型数据或者变量在R当中会存储，factor或者字符。我们可以使用count（）或者table（）来生成评书表格。

条状图当中每一列代表了一个分类型的变量条状图可以用来展示某一个列别的频数或者频率。解读条状图时调到高低长短反映了该类别这是评述。

## Continous variable

A continuous variable can take any of an infinite set of ordered values.

连续型变量可以维1系列无限的纸

### Central tendency

趋中性

一个变量的趋中性描述的是用一个数值，整个数据集分布的中间。

A measure of central tendency attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

对于变量区中心的描述我们通常有三种指标众数中位数和平均数

众数是指在变量分布当中最常出现的指

正数相对于中位数和平均数的优点时数值型和分类型数据都可以用种树来描述。

种树的缺点是在一些类型的分布当中正数可能并不反映分布的中间指

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

\*Advantage:

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

\*Limitations:

1. In some distributions, the mode may not reflect the centre of the distribution very well.

>it is easy to see that the centre of the distribution is 57 years, but the mode is lower, at 54 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

2. more than one mode for the same distribution of data, bi-modal, or multi-modal. The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

在某些多峰值的分布当中可能存在不止一个种树因此使用中数来描述趋中性受到了限制

3. In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

同时对于一些连续变量来说有时并不存在中暑比如每一个数值都不相同的时候

#### Median

中位数是指将所有值按顺序或者逆序排序其中中间的纸。

- The median is the middle value in distribution when the values are arranged in ascending or descending order.

中位数将数据分布一分为二也就是分为两个50%如果数据维奇树那么中间的时就是中位数；如果数据有偶数各那么中位数就是中间两个值的平均数。

和平均数相比中位数的优点是我太容易受到奇异值或分布畸形的数据影响当分布并不对称时中位数比较能够准确的反映趋中性。中位数的缺点是不能够表示定名的数据因为那些数据没有办法排序

平均数是将所有的观测值求和之后在处于观测值的个数也就是算术平均数。平均数的优点是可以使用连续性的数据和离散型的数据。缺点是平均数不能够使用在分类型数据上因为那样的数据是没有办法被求和的另外平均数受到其意旨和数据分布影响比较大代表样本平均数代表总体平均数。

#### Mean

- The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

\* Advantage

The mean can be used for both continuous and discrete numeric data.

\* Limitations

1. The mean cannot be calculated for categorical data, as the values cannot be summed.

2. As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

The population mean is indicated by the Greek symbol µ (pronounced ‘mu’). When the mean is calculated on a distribution from a sample it is indicated by the symbol x̅ (pronounced X-bar).

### Measures of Spread

差异梁树

Dataset A: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Dataset B: 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The measures of central tendency for both datasets above are the same.However, if we look at the spread of the values, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data

如果我们观察上面两个数据集我们可以发现他们的趋中性指标是相同的但是我们仔细观察可以发现毕数据集的离散程度比a数据集更强因此我们在描述数据时既要考虑趋中性也要考虑差异性

描述差异性可以帮助我们理解数值和平均值之间的差异程度。

Summarising the dataset can help us understand the data, especially when the dataset is large. Measures of spread summarise the data in a way that shows how scattered the values are and how much they differ from the mean value.

The spread of the values can be measured for quantitative data, as the variables are numeric and can be arranged into a logical order with a low end value and a high end value.(ABS)

唱一行的描述差异性的描述是基于量化数据的。可以通过全剧4分卫方差和标准差来描述

Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.(ABS)

```{r}

dataset1 = c(4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8)

dataset2 = c(1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11)

dataset = data.frame(dataset1, dataset2)

dataset = gather(dataset, type, data)

ggplot(dataset, aes(data, fill = type)) + geom\_bar(position=position\_dodge())

```

#### Range

The range is the difference between the smallest value and the largest value in a dataset.

全剧是指数据集中最大值和最小值的差异。

```{r}

range(dataset1)

range(dataset2)

```

#### Quartiles

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

我们可以将一个排序的数据集平均分成4个部分，其中第一个4分卫就是从零到20%第2个4分卫是数据的中间也就是25%到50%或者是中位数

The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.

The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.

The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.

#### The interquartile range

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest. The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

```{r}

quantile(dataset1)

quantile(dataset2)

```

#### Variance vs. standard deviation

The variance and the standard deviation are measures of how close each observed data value is to the mean value.

In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation. The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.

The standard deviation is the square root of the variance.

The standard deviation of a normal distribution enables us to calculate confidence intervals. In a normal distribution, about 68% of the values are within one standard deviation either side of the mean and about 95% of the scores are within two standard deviations of the mean.