Statistics

Descriptive statistics:

Statistics involving describing data. Involves summarizing and organizing data so they can be easily understood.

Inference statistics:

Complex set of procedures to draw conclusions over large populations with sample data.

Data

Numeric: wind speed, time duration, discrete etc.

Categorical: Car types, Binary, ordinal (ordered).

*------------------- 5/7/2024 Up to Page 5*

Exploratory Data Analysis

Data Structure:

* Rectangular (rows are records & columns are variable or features) and
* Non-Rectangular (spatial or graph)

In statistics we use mostly rectangular data:

**Estimates of Location**

An estimate of where most of the data is located (i.e., its central tendency)

|  |  |  |  |
| --- | --- | --- | --- |
| Key Term | Definition | Formula | Usage |
| Mean | Sum of all values/ number of values |  | average |
| Weighted Mean | Sum of all values times a weight / sum of weights |  | Some variables are intrinsically more variable than other and high variable observations are given lower weight. Ex: when taking average from multiple sensors giving lower weight for sensors that giving less accurate readings. |
| Median | The value such that one-half of data lies above and below |  | While calculating average household income in a city where bill gate lives the mean gives diff value where median gives right value no matter who is rich or not. |
| Percentile | The value such that P percentage of data lies below |  |  |
| Weighted Median | The value such that one half of the weighted sum lies above and below the sorted data. |  |  |
| Trimmed Mean | The average of all values after removing fixed number of extreme values | P smallest and largest values omitted | A trimmed means eliminate the influence of extreme values, EX: International diving the top score & bottom score from five judges are dropped and the final score is the average of scores from 3 remaining judges. This makes it difficult for a single judge to manipulate the scores. |

Note: Trimmed mean, Median and weighted median are robust to outliers.

Outlier: is a any value which is very distant from other values in data set and cause skewness.

*------------------- 5/8/2024 Up to Page 11*

[Refer Estimates of Location in Python Notebooks]

*------------------- 5/9/2024 Up to Page 12*

**Estimates of Variability**

Measures whether the data values are tightly clustered or spread out.

At the heart of the statistics lies variability:

* Measuring it
* Reducing it
* Distinguishing random from real variability
* Identifying the various sources of real variability
* Make decision out of it in presence.

|  |  |  |  |
| --- | --- | --- | --- |
| Key Terms | Why | Formula | Formula Explanation |
| Deviation | To calculate how individual item deviate from mean, useful to calculate Standard deviation | Mean – Individual Value | Simple distance calculation formula |
| Variance | How each data is varied from the mean |  | Its not express in square root as it already quantifies the spread of data in squared units. |
| Standard Deviation  (Population) means whole dataset | Square root of variance |  | Just calculating the mean of deviations (distance form mean), squaring to avoid negatives to cancel with positive. Square root to get back squares to their original unit. |
| Standard Deviation (sample) | Same as population but n-1  Dividing by 𝑛−1 instead of n results in a slightly larger value for the standard deviation, which better reflects the variability in the population from which the sample was drawn |  | This adjustment is known as Bessel's correction. The rationale behind this correction is to provide an unbiased estimate of the population standard deviation. |
| Mean Absolute Deviation (Manhattan Norm, l1 – norm) | Mean of absolute values of the deviation from the mean |  | SD emphasize the large deviation, but MAD won’t because its taking absolute values |
| Median Absolute Deviation from Median | Robustness to outliers |  | Median of absolute values of the deviation from the median |
| Percentile | A percentile tells you where a certain value falls within a dataset when arranged in ascending order. |  | Simple percentage calculation  For instance, if your score is at the 70th percentile, it means you've scored as well as or better than 70% of the other people in the dataset. |
| Interquartile range | The difference between 75th percentile and 25th percentile |  | This range represents where the bulk of the data lies, excluding the extremes. |

*------------------- 5/10/2024 Up to Page 14 (Additional notes pending)*

*------------------- 5/11/2024 Up to Page 14 (Additional notes updated with why and formula explanation)*

**Degrees of freedom and n or n – 1?**

whether you divide by n or n – 1. It is based on the premise that you want to make estimates about a population, based on a sample. If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate.

To fully explain why using n leads to a biased estimate involves the notion of degrees of freedom, which considers the number of constraints in computing an estimate. In this case, there are n – 1 degrees of freedom since there is one constraint: the standard deviation depends on calculating the sample mean. For most problems, data scientists do not need to worry about degrees of freedom.

Notes:

* Even for normal distribution the calculation of SD, MAD, Median AD are different.
* The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile).

*------------------- 5/12/2024 Up to Page 17*

[Refer Estimates of Variability in Python Notebooks]

*------------------- 5/13/2024 Up to Page 18*

**Explore Data Distribution**

|  |  |  |
| --- | --- | --- |
| Key Term | Usage | visual |
| Box Plot | * Helps to identify the quartiles, Interquartile, median, find outliers | Box Plot - Simply explained - DATAtab |
| Frequency Table | * Able to find how frequent or the count of category or intervals in a dataset. * Helps to summarize data. * Further helps to plot Bar and pie charts. | Frequency Table: How to Make & Examples - Statistics By Jim |
| Histogram | Visual representation of data distribution, the bar height says how frequence of the data occur.  Uses bins to represent frequency of observations. | Histograms Unveiled: Analyzing Numeric Distributions |
| Density Plot | A smoothed version of the histogram, often based on a kernel density estimate.  estimate the probability density function (PDF) of the underlying distribution, providing a smoothed representation of the data distribution.  When dataset is large and continuous data. | Density – from Data to Viz |

*------------------- 5/14/2024 Up to Page 19*

[Refer Estimates of Distribution in Python Notebooks]

*------------------- 5/15/2024 Up to Page 26*

**Exploring Binary and Categorical data**

|  |  |  |
| --- | --- | --- |
| Key Term | Definition | Example |
| Mode | The most commonly occurring category or value in a data set. | In most part of US, the mode of religious preference would be Christian |
| Expected value | When categories can be associated with numerical values, this gives an average value based on category’s probability of occurrence. |  |
| Bar chart | The proportion of each category plotted as bars |  |
| Pie chart | The proportion of each category plotted as wedges in a pie. |  |

*------------------- 5/16/2024 Up to Page 27*

[Refer Exploration of Binary and Categorical Data in Python Notebooks]

*------------------- 5/17/2024 Up to Page 28*

Expected Value

A marketer for a new cloud technology, for example, offers two levels of service, one priced at $300/month and another at $50/month. The marketer offers free webinars to generate leads, and the firm figures that 5% of the attendees will sign up for the $300 service, 15% will sign up for the $50 service, and 80% will not sign up for anything. This data can be summed up, for financial purposes, in a single “expected value,”

The expected value is calculated as follows:

1. Multiply each outcome by its probability of occurrence.

2. Sum these values

In the cloud service example, the expected value of a webinar attendee is thus $22.50 per month, calculated as follows:

EV = (0 .05) (300) + (0 .15) (50) + (0 .80) (0) = 22.5

*------------------- 5/18/2024 Up to Page 29*

**Probability**

For our purposes the probability of an event will happen is the proportion of the event will occur if the situation could be repeated over and over again, infinitely.

**Correlation**

In Exploratory Data Analysis the correlation will be performed among features or between feature and target variables.

If the highest value of X goes with highest value of Y then correlation of X and Y is positively correlated and also for vice versa.

If the highest value of X goes with lowest value of X then correlation of X and Y is negatively correlated and also for vice versa

*------------------- 5/19/2024 Up to Page 30*