KDnuggets

Descriptive Statistics Key Terms, Explained

Previous post
Next post

Tags: Explained, Key Terms, Statistics

This is a collection of 15 basic descriptive statistics key terms, explained in easy to understand language, along with an example and some Python code for computing simple descriptive statistics.

By Matthew Mayo, KDnuggets.

Statistics, though a central set of tools for data science, are often overlooked in favor of more solidly technical skills like programming. Even machine learning learning algorithms, with their reliance on mathematical concepts such as algebra and calculus -- not to mention statistics! -- are often treated at a

higher level than is required to appreciate the underlying math, leading, perhaps, to "data scientists" who lack a fundamental understanding of one of the key aspects of their profession.

This post won't resolve the discrepancy between knowing and not knowing the absolute basics of statistics. However, if you are unable to fully understand the basic descriptive statistics terminology included herein, you are definitely lacking foundational knowledge that is needed to build a whole series of much more robust and useful professional concepts on top of.

So here is a collection of 15 basic descriptive statistics key terms, explained in easy to understand language. A comprehensive example follows, which includes a bit of Python code. Note that, as a basic introduction, mathematical representations and descriptions of the terms herein have been intentionally omitted.

1. Descriptive Statistics

Descriptive statistics are a collection of statistical tools which are used to quantitatively describe or summarize a collection of data. Descriptive statistics aim to summarize, and as such can be distinguished from inferential statistics, which are more predictive in nature.

2. Population

A population is a selected individual or group representing the full set of members of a certain group of interest.

3. Sample

A sample is a subset drawn from a larger population. If this drawing is accomplished in such a manner that each member of the population has an equitable chance of selection, the result is referred to as a **random sample**.

4. Parameter

A parameter is a value which is generated from a population. If I had all the data of all humans on Earth and generated the mean age, this value would be a parameter.

5. Statistic

A statistic is a value which is generated from a sample. If I calculated the mean age of a subset of humans on planet Earth (much more feasible), this value would be a statistic. Hence, the discipline of statistics.

6. Generalizability

Generalizability refers tot he ability to draw conclusions about the characteristics of the population as a whole based on the results of data collected from a sample. This is ability is not a given, and depends heavily on the nature of sample collection, sample size, and various other factors.

7. Distribution

A distribution is the arrangement of data by the values of one variable in order, from low to high. This arrangement, and its characteristics such as shape and spread, provide information about the underlying sample.

8. Mean

Mean, along with median and mode, is one of the 3 major measures of central tendency, which collectively evaluate an important and basic aspect of a distribution. The simple arithmetic average of a distribution of variable values (or scores), the mean provides a single, concise numerical summary of a distribution. The mean is also likely the most common statistics encountered in general research. Population mean is denoted μ, while sample mean is denoted $\bar{x}$.

9. Median

The median is the score of a distribution residing at the 50th percentile, separating the top and bottom 50 percent of scores. The median is useful for both splitting a set of distribution scores in half and helping to identify the skew of a distribution.

10. Mode

refers to a distribution with 2 modes.

11. Skew

When there are more scores toward one end of the distribution than the other, this results in skew. When the scores of a distribution are more clustered at the high end, the relatively fewer number of scores on the low end result in a tail, with the scenario being referred to as negative skew. Positive skew is when a distribution shows a tail at its high end.

In general, in a negatively skewed distribution, we would expect the mean to be less than the median, while in a positively skewed distribution, we would expect the mean to be greater than the median.

12. Range

One of the most important measures of dispersion, the range is the difference between the maximum and minimum values of a distribution.

13. Variance

Variance is the statistical average of the dispersion of scores in a distribution. Variance is not often used on its own, but can be a useful calculation on the way to a more descriptive statistical measurement, such as standard deviation.

14. Standard Deviation

The standard deviation of a distribution is the average deviation between individual distribution scores and the distribution's mean. Individually, the standard deviation provides a good measure of how spread out a disquisitions scores are. When considered alongside the mean, these 2 measures provides a good overview of the distribution of scores.

15. Interquartile Range (IQR)

The IQR is the difference between the score delineating the 75th percentile and the 25th percentile, the third and first quartiles, respectively.

Below is a simple Python module for computing much of the descriptive statistics discussed above, followed by an example.

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats

def descriptive_stats(distribution):

        '''
        Compute and present simple descriptive stats for a distribution

        Parameters
        ----------
        distribution: list
                Distribution as a Python list
        '''

        # Convert distribution as numpy ndarray
        dist = np.array(distribution)

        print 'Descriptive statistics for distribution:\n', dist
        print 'Number of scores:', len(dist)
        print 'Number of unique scores:', len(np.unique(dist))
        print 'Sum:', sum(dist)
        print 'Min:', min(dist)
        print 'Max:', max(dist)
        print 'Range:', max(dist)-min(dist)
        print 'Mean:', np.mean(dist, axis=0)
        print 'Median:', np.median(dist, axis=0)
        print 'Mode:', scipy.stats.mode(dist)[0][0]
        print 'Variance:', np.var(dist, axis=0)
        print 'Standard deviation:', np.std(dist, axis=0)
        print '1st quartile:', np.percentile(dist, 25)
        print '3rd quartile:', np.percentile(dist, 75)
        print 'Distribution skew:', scipy.stats.skew(dist)

        plt.hist(dist, bins=len(dist))
        plt.yticks(np.arange(0, 6, 1.0))
        plt.title('Histogram of distribution scores')
```

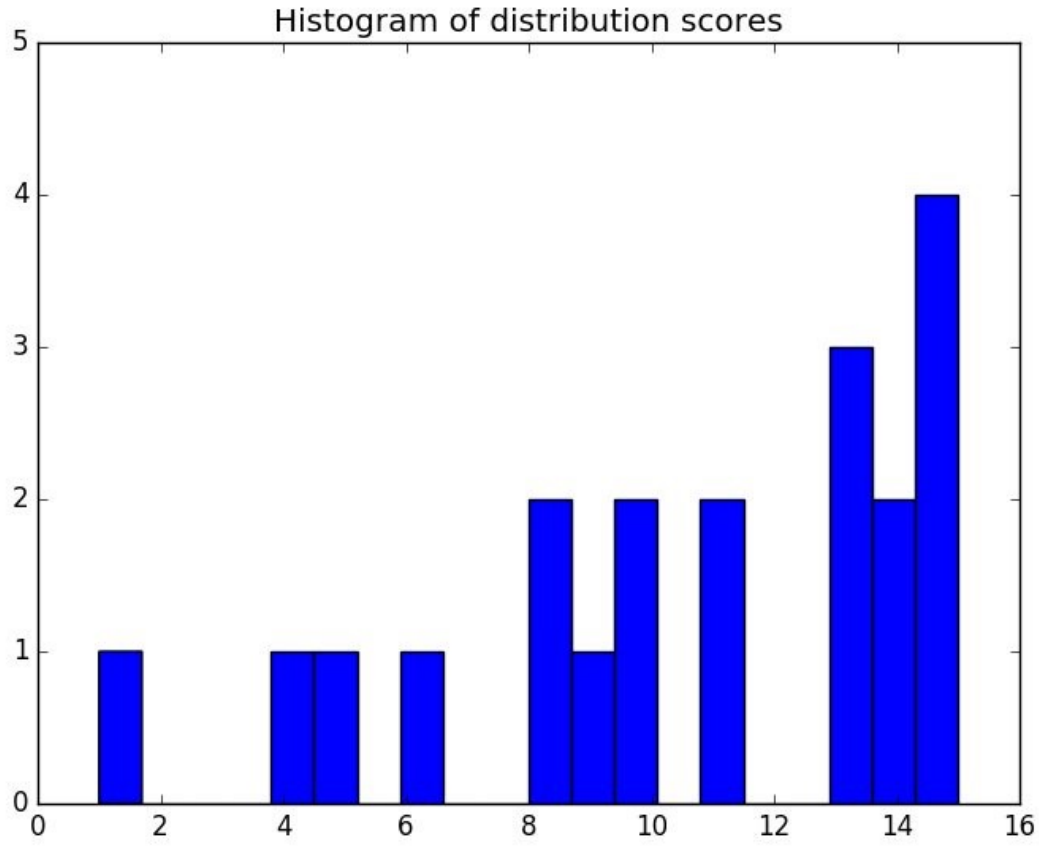
```
40  descriptive_stats([ 1, 4, 5, 6, 8, 8, 9, 10, 10, 11, 11, 13, 13, 13, 14, 14, 15, 15, 15, 15 ])
```

descriptive-stats.py hosted with ♥ by GitHub view raw

```
Descriptive statistics for distribution:

[ 1  4  5  6  8  8  9 10 10 11 11 13 13 13 14 14 15 15 15 15]

Number of scores: 20
Number of unique scores: 11
Sum: 210
Min: 1
Max: 15
Range: 14
Mean: 10.5
Median: 11.0
Mode: 15
Variance: 16.15
Standard deviation: 4.01870625948
1st quartile: 8.0
3rd quartile: 14.0
Distribution skew: -0.714152479663
```



Related:

- Machine Learning Key Terms, Explained
- Natural Language Processing Key Terms, Explained
- Deep Learning Key Terms, Explained

Previous post
Next post

Top Stories Past 30 Days

Most Popular	Most Shared

SHARES

1. **The 10 Statistical Techniques Data Scientists Need to Master**
2. **Top 10 Machine Learning Algorithms for Beginners**
3. **Deep Learning Specialization by Andrew Ng - 21 Lessons Learned**
4. **6 Books Every Data Scientist Should Keep Nearby**
5. **Top 10 Videos on Deep Learning in Python**
6. **Did Spark Really Kill Hadoop?**
7. **The 10 Algorithms Machine Learning Engineers Need to Know**

1. **Data Science, Machine Learning: Main Developments in 2017 and Key Trends in 2018**
2. **Deep Learning Specialization by Andrew Ng – 21 Lessons Learned**
3. **Using Deep Learning to Solve Real World Problems**
4. **Top Data Science and Machine Learning Methods Used in 2017**
5. **Big Data: Main Developments in 2017 and Key Trends in 2018**
6. **Why You Should Forget for-loop for Data Science Code and Embrace Vectorization**
7. **Understanding Deep Convolutional Neural Networks with a practical use-case in Tensorflow and Keras**

Latest News

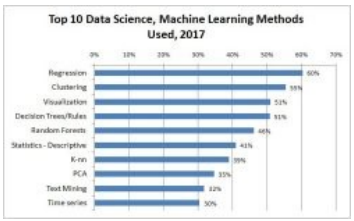
- Is Religion The Next Frontier For AI?
- An introduction to Monte Carlo Tree Search
- Computer Vision by Andrew Ng - 11 Lessons Learned
- DeepSchool.io: Deep Learning Learning
- Lehigh University: Data Architect
- Top Stories of 2017: 10 Free Must-Read Books for Machin...

Top Stories
Last Week

Most Popular

1. NEW **Top Data Science and Machine Learning Methods Used in 2017**

SHARES



2. Top 10 Machine Learning Algorithms for Beginners
3. The 10 Statistical Techniques Data Scientists Need to Master
4. NEW Transitioning to Data Science: How to become a data scientist, and how to create a data science team
5. NEW Today I Built a Neural Network During My Lunch Break with Keras
6. NEW Data Science, Machine Learning: Main Developments in 2017 and Key Trends in 2018
7. NEW Another Day in the Life of a Data Scientist

Most Shared

1. Data Science, Machine Learning: Main Developments in 2017 and Key Trends in 2018

2. Top Data Science and Machine Learning Methods Used in 2017
3. Another Day in the Life of a Data Scientist
4. The 10 Deep Learning Methods AI Practitioners Need to Apply
5. Machine Learning & Artificial Intelligence: Main Developments in 2017 and Key Trends in 2018
6. Transitioning to Data Science: How to become a data scientist, and how to create a data science team
7. How to Generate FiveThirtyEight Graphs in Python