ACADGILD

Mastering Data Science

Session 15 - Introduction to Statistics

Agenda

1. Statistics
2. Introduction to Basic Terms
3. Variables
4. Mean, Median and Mode
5. Measure of Dispersion
6. Range
7. Sample Variance
8. Standard Deviation
9. Population Vs. Sample
10. Chebysheff's Theorem
11. Law of Expected Values and Variance
12. Probability Density Function

Statistics

- Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to help in making more effective decisions.
- Statistical Analysis is implemented to manipulate, summarize and investigate data, so that useful decision-making information results are obtained.

Copyrights© 2017, AcadGild. All rights reserved

Types of Statistics

- Descriptive Statistics is a method of organizing, summarizing, and presenting data in an informative way.
- Inferential Statistics is a method which is used in determining something about a population on the basis of a sample.
  - Population - The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
  - Sample – A portion, or part, of the population of interest.

Copyrights© 2017, AcadGild. All rights reserved

Introduction to Basic Terms

- Population - A collection/set of individuals/objects/events whose properties are to be analyzed. There are two kinds:
  - Finite
  - Infinite
- Sample - A population subset.

Copyrights© 2017, AcadGild. All rights reserved

Introduction to Basic Terms

- Variable - A characteristic about each individual element of a population/sample.
- Data (singular) - A value of the associated variable with one element of a population/sample. This value may be a number, a word, or a symbol.
- Data (plural) - A set of values collected for the variable from each of the elements belonging to the sample.
- Experiment - A planned activity whose results yield a set of data.
- Parameter - A numerical value which summarizes the entire population data.
- Statistics - A numerical value which summarizes the sample data.

Copyrights© 2017, AcadGild. All rights reserved

Two Kinds of Variables

Qualitative, or Attribute, or Categorical, Variable

- A variable that categorizes or describes a population element.

Note: Arithmetic operations such as addition and averaging, are not meaningful for data resulting from a qualitative variable.

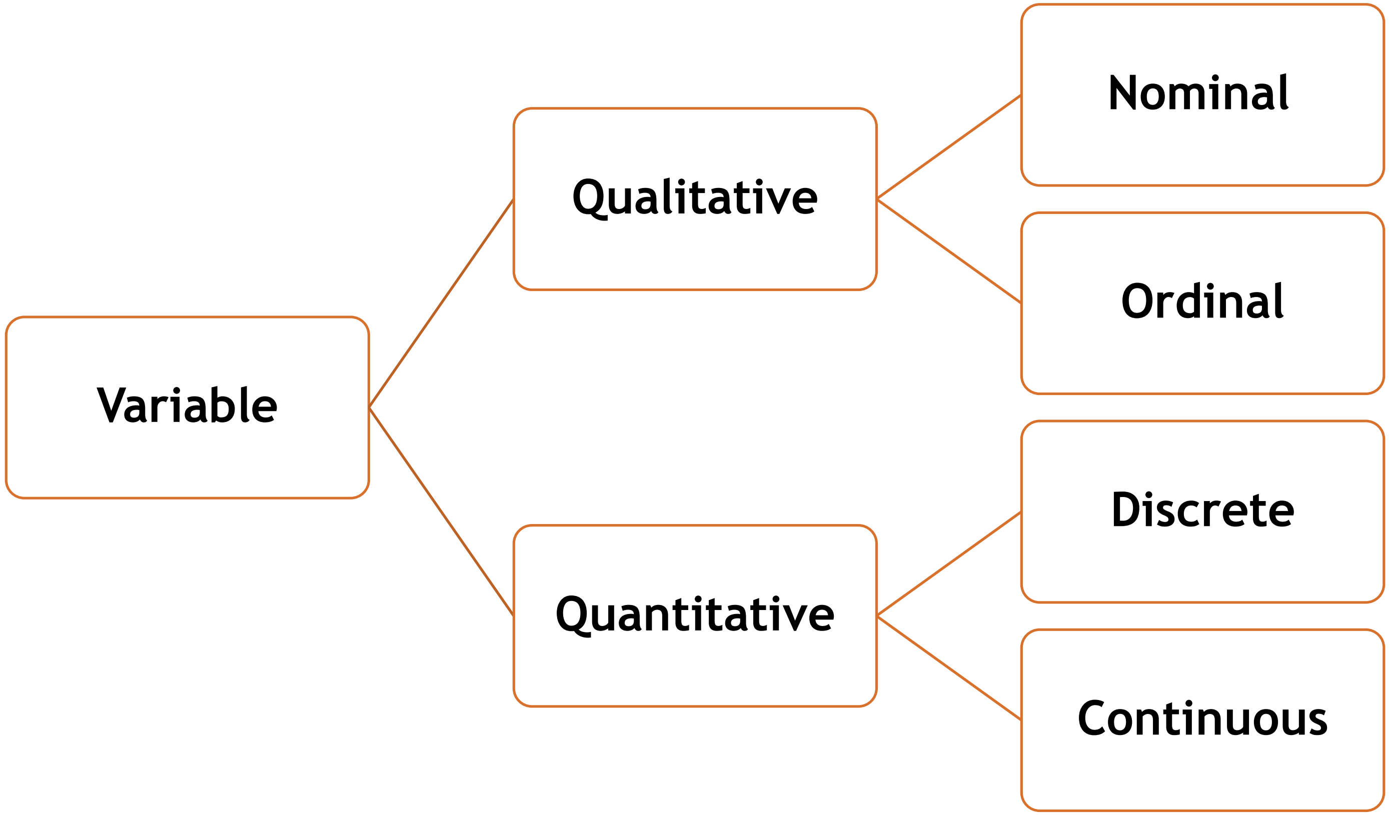
Quantitative, or Numerical, Variable

- A variable that quantifies a population element.

Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.

Copyrights© 2017, AcadGild. All rights reserved

Two Kinds of Variables



Copyrights© 2017, AcadGild. All rights reserved

Two Kinds of Variables

- Nominal Variable - A qualitative variable that categorizes (or describes, or names) a population element.
- Ordinal Variable - A qualitative variable that incorporates an ordered position or ranking.
- Discrete Variable - A quantitative variable that can assume a countable number of values.
  - This can assume values corresponding to the isolated points along a line interval.
  - There is a gap between any two values
- Continuous Variable - A quantitative variable that can assume an uncountable number of values.
  - This can assume any value along a line interval
  - Including every possible value between any two values

Copyrights© 2017, AcadGild. All rights reserved

The Mean

- Let **x1, x2, x3,…, xn** be the realized values of a random variable ‘X’, from a sample of size ‘n’.

The sample arithmetic mean is defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Copyrights© 2017, AcadGild. All rights reserved

Example

- The systolic blood pressure of seven middle aged men were as follows:

151, 124, 132, 170, 146, 124 and 113.

The Mean is $\bar{x} = \dfrac{(151+124+132+170+146+124+113)}{7}$

$$= 137.14$$

Copyrights© 2017, AcadGild. All rights reserved

Median and Mode

- The median for the sample data arranged in an increasing order is defined as:

  i. If “n” is an odd number - Middle value

  ii. If “n” is an even number - Midway between the two middle values

- The mode is the most commonly occurring value.

Copyrights© 2017, AcadGild. All rights reserved

Median and Mode

Example - n is odd

The re-ordered systolic blood pressure data seen earlier are:

113, 124, 124, 132, 146, 151, and 170.

- The Median is the middle value of the ordered data, i.e. 132.
- Two individuals have systolic blood pressure = 124 mm Hg, so the Mode is 124.

Copyrights© 2017, AcadGild. All rights reserved

Median and Mode

Example - n is even

Six men with high cholesterol participated in a study to investigate the effects of diet on cholesterol level. At the beginning of the study, their cholesterol levels (mg/dL) were as follows:

366, 327, 274, 292, 274 and 230

Rearrange the data in numerical order as follows:

230, 274, 274, 292, 327 and 366.

- The Median is half way between the middle two readings, i.e. (274+292) / 2 = 283.
- The mode between the two men having the same cholesterol level = 274.

Copyrights© 2017, AcadGild. All rights reserved

Mean Vs. Median

- If the histogram of the data is right-skewed then large sample values tend to inflate the mean.
- If the distribution is skewed then the median is not influenced by large sample values and is a better measure of centrality.

Note - If mean = median = mode then the data are said to be symmetrical.

For example,

- In the CK measurement study, the sample mean = 98.28.
- The median = 94.5, i.e. mean is larger than median indicating that mean is inflated by two large data values 201 and 203.

Copyrights© 2017, AcadGild. All rights reserved

Using Python to Calculate Measures of Central Tendency

```
import numpy as np
from random import randint

x = [randint(1, 10) for p in range(0, 10)]
print(x)
mean = np.mean(x)
print("The mean is {:.2f}".format(mean))
median = np.median(x)
print("The median is {:.2f}".format(median))
mode = max(set(x), key=x.count)
print("The mode is {:.2f}".format(mode))
```

```
[3, 9, 5, 10, 10, 1, 7, 8, 5, 5]
The mean is 6.30
The median is 6.00
The mode is 5.00
```

[Practice code]

Measures of Dispersion

➢ The concept Measures of Dispersion characterize how to spread out the distribution, i.e., how variable the data are.

➢ The commonly used dispersion measures include:

- Range
- Variance and Standard Deviation

Copyrights© 2017, AcadGild. All rights reserved

Range

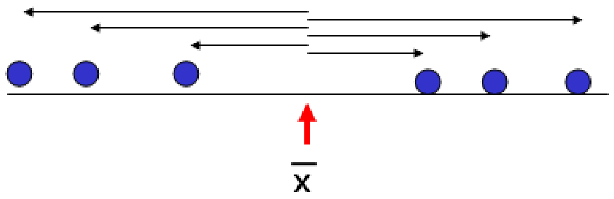
- ➢ The Range is the difference between the largest and the smallest observations in the sample.
- ➢ For example, the minimum and maximum blood pressure is 113 and 170 respectively. Hence the range is 57 mmHg
  - Easy to calculate;
  - Implemented for both “best” or “worst” case scenarios
  - Too sensitive for extreme values

Copyrights© 2017, AcadGild. All rights reserved

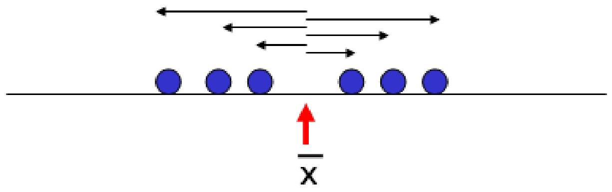
Sample Variance

- The sample variance, s^2, is the arithmetic mean of the squared deviations from the sample mean:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$



$>$



Copyrights© 2017, AcadGild. All rights reserved

Standard Deviation

- The sample standard deviation (s) is the square-root of the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

- The sample standard deviation has an advantage of being in the same units as the original variable (x).

Copyrights© 2017, AcadGild. All rights reserved

Using Python to Calculate Measures of Dispersion

```python
import numpy as np
from random import randint

x = [randint(1, 10) for p in range(0, 10)]
print(x)
variance = np.var(x)
print("The variance is {:.2f}".format(variance))
std = np.std(x)
print("The standard deviation is {:.2f}".format(std))
rng = max(x) - min(x)
print("The range is {:.2f}".format(rng))
```

```
[1, 10, 2, 9, 1, 4, 4, 2, 4, 6]
The variance is 9.01
The standard deviation is 3.00
The range is 9.00
```

Practice Code

Population Vs. Sample

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Vs.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Population Mean

Sample Mean

Copyrights© 2017, AcadGild. All rights reserved

Population Vs. Sample

	Population	Sample
Size	N	n
Mean		

Copyrights© 2017, AcadGild. All rights reserved

Population Vs. Sample

	Population	Sample
Size	N	n
Mean		
Variance		

Copyrights© 2017, AcadGild. All rights reserved

Population Vs. Sample

- The variance of a population is:

Population Mean

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Population Size

- The variance of a sample is:

Sample Mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Note! the denominator is sample size (n) minus one !

Copyrights© 2017, AcadGild. All rights reserved

Population Vs. Sample

- The square root of the variance is termed as the Standard Deviation, thus:

  - The population Standard Deviation = $\sigma = \sqrt{\sigma^2}$

  - The Sample Standard deviation = $s = \sqrt{s^2}$

Copyrights© 2017, AcadGild. All rights reserved

Chebysheff's Theorem

- A more general interpretation of the standard deviation is derived from Chebysheff's Theorem, which applies to all shapes of histograms (except bell shaped).
- The proportion of observations in any sample that lie within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2} \quad for \quad k > 1$$

For k=2 (say), the theorem states that at least 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

Copyrights© 2017, AcadGild. All rights reserved

Two Types of Random Variables

Discrete Random Variable

- Takes on a countable number of values
- For example, values on the roll of dice: 2, 3, 4, …, 12

Continuous Random Variable

- Values are not discrete, not countable
- For example, time (30.1 minutes? 30.10000001 minutes?)

Analogy

- Integers are discrete, while Real Numbers are Continuous

Copyrights© 2017, AcadGild. All rights reserved

Laws of Expected Value

- $E(C) = C$
  - The expected Value of a Constant is just the value of the constant.

- $E(X + C) = E(X) + C$

- $E(CX) = cE(X)$
  - We can “pull” a constant out of the expected value expression (either as part of a sum with a random variable X or as a coefficient of random variable X).

Copyrights© 2017, AcadGild. All rights reserved

Laws of Variance

- $V(c) = 0$
  - The Variance of constant (c) is zero.

- $V(X + c) = V(X)$
  - The Variance of random variable and a constant is just the variance of the random variable (per 1 above).

- $V(cX) = c^2 V(X)$
  - The Variance of a random variable and a constant co-efficient is the co-efficient squared times in the variance of the random variable.

Copyrights© 2017, AcadGild. All rights reserved

Probability Density Functions

Unlike a discrete random variable, a continuous random variable is one that can assume an uncountable number of values.

- We cannot list the possible values because there is an infinite number of them.
- The probability of each individual value is virtually 0 as there is an infinite number of values

Copyrights© 2017, AcadGild. All rights reserved

Point Probabilities are Zero

If the probability of each individual value is virtually 0 then there is an infinite number of values.

Thus, we can determine the probability of a **range of values** only.

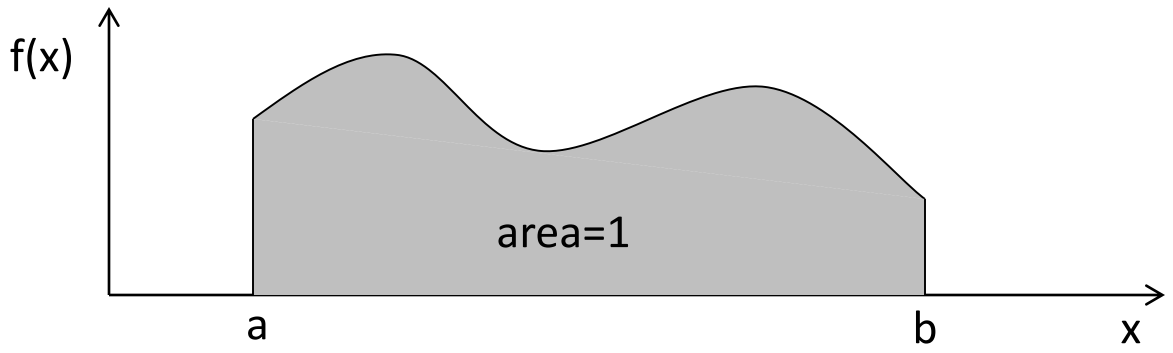
- For example, with a discrete random variable like tossing a die, it is meaningful to talk about P(X=5)
- In a **continuous** setting (e.g. with time as a random variable), the probability the random variable of interest say task length, takes exactly 5 minutes is infinitely small, hence P(X=5) = 0.

Copyrights© 2017, AcadGild. All rights reserved

Probability Density Function

A function f(x) is called a Probability Density Function over the range $a \le x \le b$ if it meets the following requirements:

1. $f(x) \ge 0$ for all x between a and b, and



2. The total area under the curve between a and b is 1.0

Copyrights© 2017, AcadGild. All rights reserved

Email us - support@acadgild.com