* [Overview](#gjdgxs)
* [Statistic](#30j0zll)
  + [Continuous vs Categorical Features](#1fob9te)
* [Statistics Review](#3znysh7)

# CSE 450 - Machine Learning & Data Mining

## Preparation Reading 02: Descriptive Statistics

## Overview

In this reading we'll review some basic statistics you'll need for machine learning.

Estimated Reading Time

Plan on around 60 to 90 minutes for this preparation reading, which contains a mix of online and textbook reading.

## Statistic

Think of a "statistic" as a piece of information that tells you something about a set of data. Broadly speaking, there are two types of statistics:

Descriptive Statistic

A calculation on a set of data that allows us to describe the data.

Inferential Statistic

A calculation on a set of data that allows us to make predictions (or inferences) about the data.

We'll be working mainly with descriptive statistics in this course.

Calculating and visualizing descriptive statistics about a data set we're interested in will allow us to decide which machine learning algorithms are most appropriate to use, and how to set those algorithms up for the best results.

### Continuous vs Categorical Features

When analyzing a feature of our data, one of the first things we need to know is whether or not we are looking at a continuous or categorical feature.

Continuous Feature

A property of the data ($x$) that can take on a range of quantitative values between two numeric limits: $ x\_{min} \leq x \leq x\_{max} $.

If we were to measure the height of everyone on campus, we might find that their heights span a range between 100 cm and 200 cm. We would say that height is a continuous feature of our data, with a minimum value of 100 cm and a maximum value of 200 cm.

Other examples of continuous features for this dataset might include the student's weight, age, GPA, number of credits completed, hours of Netflix watched each week, and bank account balance.

Warning

A mathematician might take issue with calling the measurement in the above example "continuous", if we're only allowing whole centimeters (101 cm, 104 cm, 198 cm, etc...).

To a mathematician, this would be a *discrete* measurement, since we're jumping from whole number to whole number. A *continuous* measurement would require that we also allow an infinite number of fractional values (101.2343 cm, 104.5 cm, etc...).

In machine learning, we will usually call discrete numeric values like this "continuous features" even though they aren't necessarily mathematically continuous.

If instead of measuring their heights, we asked students which academic class they were in (Freshman, Sophomore, Junior, or Senior), we would be measuring a *categorical feature*.

Categorical Feature

A property of the data that can have one or more qualitative values.

Other examples of categorical measurements for this dataset might include the student's favorite flavor of ice cream, their eye color, their relationship status, and whether or not they've completed an internship.

Features in Disguise

Beware of categorical features that *look* like continuous features. For example, if students in our survey live in three different apartment complexes, we might ask "Which apartment complex do you live in, 1, 2, or 3?".

Even though this is technically a numeric value between two limits, it is really a categorical feature because there's no quantitative difference between the apartment complexes. Later on, we'll see how to make sure our machine learning algorithms know that a feature really is a categorical feature, even when it uses numeric values.

Even more tricky is the case of features that could go either way. In our first categorical feature example, we could have asked "Which academic year are you currently in, 1, 2, 3, or 4?".

There might actually be an important relationship in our data that relies on the fact that senior students have been at school 4 times longer than freshman students. In cases like these, we sometimes have to try treating a feature both ways to see which gives us better results.

## Statistics Review

In *Appendix A* of *Machine Learning and Predictive Data Analytics*, read the following sections:

* A.1 - Descriptive Statistics for Continuous Features
* A.2 - Descriptive Statistics for Categorical Features
* A.3 - Populations and Samples