Introduction to Data Science Essential Concepts

IDS Unit 1: Essential Concepts

Lesson 1: Data Trails

Data are a collection of recorded observations. Data are gathered by people and by sensors. Patterns in data can reveal previously unknown patterns in our world. Data play a large, and sometimes invisible, role in our lives.

Lesson 2: Stick Figures

Data consist of records of particular characteristics of people or objects. Data can be organized in many different ways, and some ways make it easier than others for achieving particular purposes.

Lesson 3: Data Structures

Variables record values that vary. By organizing data into rectangular format, we can easily see the characteristics of observations by reading across a row, or we can see the variability in a variable by reading down the column. Computers can easily process data when it is in rectangular format.

Lesson 4: The Data Cycle

A statistical investigation consists of cycling through the four stages of the Data Cycle; statistical questions are questions that address variability and are productive in that they motivate data collection, analysis, and interpretation. The Data Collection phase might consist of collecting data through Participatory Sensing or some other means, or it might consist of examining previously collected data to determine the quality of the data for answering the statistical questions. Data Analysis is almost always done on the computer and consists of creating relevant graphics and numerical summaries of the data. Data Interpretation is involved with using the analysis to answer the statistical questions.

Lesson 5: So Many Questions

Statistical questions address variability.

Lesson 6: What Do I Eat? [The Data Cycle: Consider Data]

After raising statistical questions, we examine and record data to see if the questions are appropriate.

Lesson 7: Setting the Stage [The Data Cycle: Collect Data]

In Participatory Sensing, we humans behave as if we are robot sensors, collecting data whenever a "trigger" event occurs. Our ability to learn about the patterns in our life through these data depends on our being reliable data collectors.

Lesson 8: Tangible Plots [The Data Cycle: Analyze Data]

Distributions organize data for us by telling us (a) which values of a variable were observed, and (b) how many times the values were observed (their frequency).

Lesson 9: What Is Typical?

The “center” of a distribution is a deliberately vague term, but it is one way to answer the subjective question "what is a typical value?" The center could be the perceived balancing point or the value that approximately cuts the area of the distribution in half.

Lesson 10: Making Histograms

Histograms can be created through the use of an algorithm. The distributions displayed in a histogram can be classified using the technical terms for the shapes of distributions. Learning to describe routine tasks through an algorithm is an important component of computational thinking.

Lesson 11: What Shape Are You In?

Identifying the shape of a histogram is part of the **interpret** step of the Data Cycle.

Lesson 12: Exploring Food Habits

Once Participatory Sensing data has been collected, the Dashboard and PlotApp perform the analysis step of the Data Cycle, though humans need to tell the computer which plots to examine.

Lesson 13: RStudio Basics

The computer has a syntax, and it can only understand if you speak its language.

Lesson 14: Variables, Variables, Variables

To examine whether two (or more) variables are related, we can plot their distributions on the same graph.

Lesson 15: Americans’ Time on Task

Learning to examine other analyses is an important part of statistical thinking.

Lesson 16: Categorical Associations

A two-way table is a summary of the association/relationship between two categorical variables. Joint relative frequencies answer questions of the form "what proportion of the people/objects had *this* value on the first variable and *this* value on the second?"

Lesson 17: Interpreting Two-Way Tables

Marginal (relative) frequencies tell us about the distribution of a single variable. Conditional relative frequencies tell us about the distribution of one variable when "subsetting" the other.

IDS Unit 2: Essential Concepts

Lesson 1: What Is Your True Color?

Students will understand that the 'typical' value is a value that can represent the entire group, even though we know that not all members of the group share the same value.

Lesson 2: What Does Mean Mean?

The center of a distribution is the 'typical' value. One way of measuring the center is with the mean, which finds the balancing point of the distribution. The mean gives us the typical value, but does not tell the whole story. We need a way to measure the variability to understand how observations might differ from the typical value.

Lesson 3: Median In the Middle

Another measure of center is the median, which can also be used to represent the typical value of a distribution. The median is preferred for skewed distributions or when there are outliers, because it better matches what we think of as 'typical.'

Lesson 4: How Far Is It from Typical?

MAD measures the variability in a sample of data - the larger the value, the greater the variability. More precisely, the MAD is the typical distance of observations from the mean. There are other measures of spread as well, notably the standard deviation and the interquartile range (IQR).

Lesson 5: Human Boxplots

A common statistical question is “How does this group compare to that group?” This is a hard question to answer when the groups have lots of variability. One approach is to compare the centers, spreads, and shapes of the distributions. Boxplots are a useful way of comparing distributions from different groups when all of the distributions are unimodal (one hump).

Lesson 6: Face Off

Writing (and saying) precise comparisons between groups in which variability is present based on the (a) center, (b) spread, (c) shape, and (d) unusual outcomes help to make statements in context of the data. Actual comparison statements should use terms such as "less than," "about the same as," etc.

Lesson 7: Plot Match

Boxplots are an alternative visualization of histograms or dot plots. They capture most, but not all, of the features we can see in a dot plot or histogram.

Lesson 8: How Likely Is It?

Probability is an area about which we humans have poor intuition. Probability measures a long-run proportion: 50% chance means the event happens 50% of the time if you repeated it forever. When we don't repeat forever, we see variability.

Lesson 9: Bias Detective

In the short-term, actual outcomes of chance experiments vary from what is 'ideal.' An ideal die has equally likely outcomes. But that does not mean we will see exactly the same number of one dots, two dots, etc.

Lesson 10: Marbles, Marbles…

There are two ways of sampling data that model real-life sampling situations: with and without replacement. Larger samples tend to be closer to the "true" probability.

Lesson 11: This AND/OR That

What does "A or B" mean versus "A and B" mean? These are compound events and two-way tables can be used to calculate probabilities for them.

Lesson 12: Don't Take My Stress Away!

Generating statistical questions is the first step in a Participatory Sensing campaign. Research and observations help create applicable campaign questions.

Lesson 13: The Horror Movie Shuffle

We can "shuffle" data based on categorical variables. The statistic we use is the difference in proportions. The distribution we form by shuffling represents what happens if chance were the only factor at play. If the actual observed difference in proportions is near the center of this shuffling distribution, then we would conclude that chance is a good explanation for the difference. But if it is extreme (in the tails or off the charts), then we should conclude that chance is NOT to blame. Sometimes, the apparent difference between groups is caused by chance.

Lesson 14: The Titanic Shuffle

We can also "shuffle" data based on numerical variables. The statistic we use is the difference in means. The distribution we form by this form of shuffling still represents what happens if chance were the only factor at play. When differences are small, we suspect that they might be due to chance. When differences are big, we suspect they might be 'real.'

Lesson 15: Tangible Data Merging

We can enhance the context of a statistical problem by merging related data sets together. To merge data, each data set must have a "unique identifier" that tells us how to match up the lines of the data.

Lesson 16: What Is Normal?

The Normal curve, also called the Gaussian distribution and the "bell curve," is a model that describes many real-life distributions and is usually called the Normal Model.

Lesson 17: A Normal Measure of Spread

The standard deviation is another measure of spread. This is commonly used by statisticians

because of its role in common models and distributions, such as the Normal Model.

Lesson 18: Shuffling with Normal

Z-scores allow us a way to measure how extreme a value is, regardless of the units of measurement. Usually, z-scores will range between -3 and +3, and so values that are at or more extreme than -3 or +3 standard deviations are considered large.

IDS Unit 3: Essential Concepts

Lesson 1: Anecdotes vs. Data

Data beat anecdotes. In science, we need to closely examine the quality of evidence in order to make sound conclusions. Anecdotes can contain personal bias, might be carefully selected to represent a particular point of view, and, in general, may be completely different from the general trend.

Lesson 2: What is an Experiment?

Science is often concerned with the question "What causes things to happen?" To answer this, controlled experiments are required. Controlled experiments have several key features: (1) there is a treatment variable and a response variable, and we wish to see if the treatment causes a change that we can measure with the response variable; (2) There is a comparison/control group; (3) Subjects are assigned randomly to treatment or control (randomized assignment); (4) Subjects are not aware of which group they are in (a 'blind'). This may require the use of a placebo for those in the control group; and (5) those who measure the response variable do not know which group the subjects were in (if both 4 and 5 are satisfied, this is a 'double blind' experiment).

Lesson 3: Let's Try an Experiment!

Randomized assignment is required to determine cause-and-effect.

Lesson 4: Predictions, Predictions

Designing an experiment requires making many decisions, including what to measure and how to measure it.

Lesson 5: Time Perception Experiment

Designing and carrying out an experiment helps us answer specific statistical questions of interest.

Lesson 6: Observational Studies

Observational studies are those for which there is no intervention applied by researchers.

Lesson 7: Observational Studies vs. Experiments

Experiments are not always possible because of various factors such as ethics, cost limitations, and feasibility.

Lesson 8: Monsters that Hide in Observational Studies

Confounding factors/variables make it difficult to determine a cause-and-effect relation between two variables.

Lesson 9: Survey Says…

Surveys ask simple, straightforward questions in order to collect data that can be used to answer statistical questions. Writing such questions can be hard (but fun)!

Lesson 10: We're So Random

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson 11: The Gettysburg Address

Statistics vary from sample to sample. If the typical value across many samples is equal to the population parameter, the statistic is 'unbiased.' Bias means that we tend to "miss the mark." If we don't do random sampling, we can get biased estimates.

Lesson 12: Bias in Survey Sampling

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson 13: The Confidence Game

We can estimate population parameters. This means that we can give an estimate "plus or minus" some amount that we are confident contains the true value (the population parameter).

Lesson 14: How Confident Are You?

We can estimate population parameters. This means that we can give an estimate "plus or minus" some amount that we are confident contains the true value (the population parameter).

Lesson 15 Ready, Sense, Go!

Sensors are another data collection method. Unlike what we have seen so far, sensors do not involve humans (much). They collect data according to an algorithm.

Lesson 16: Does it have a Trigger?

A key feature that distinguishes the way sensors collect data from more traditional approaches is that sensors collect data when a 'trigger' event occurs. In Participatory Sensing, this event is something we humans agree upon beforehand. Every time that trigger happens, we collect data.

Lesson 17: Creating Our Own Participatory Sensing Campaign

Creating a Participatory Sensing Campaign requires that survey questions must be completed whenever they are "triggered". Research questions provide an overall direction in Participatory Sensing Campaign.

Lesson 18: Evaluating Our Own Participatory Sensing Campaign

Statistical questions guide a Participatory Sensing Campaign so that we can learn about a community or ourselves. These Campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson 19: Implementing Our Own Participatory Sensing Campaign

Practicing data collection prior to implementation allows optimization of a Participatory Sensing Campaign.

Lesson 20: Online Data-ing

We stretch students' conception of data, to help them see that many web pages present information that can be turned into data.

Lesson 21: Learning to Love XML

XML is a programming language that we use with our campaigns. We create basic XML "tags" in the code, which help us store data in a format we understand.

Lesson 22: Changing Orientation

Converting XML to spreadsheet format helps us better understand and view our data.

IDS Unit 4: Essential Concepts

Lesson 1: Water Usage

Data can be used to make predictions. Official data sets rely on censuses or random samples and can be used to make generalizations. On the other hand, data from Participatory Sensing campaigns are not random and rely on the sensors, in our case, humans, to be gathered and limits the ability to generalize.

Lesson 2: Exploring Water Usage

Exploring different data sets can give us insight about the same processes. Information from an official data set compared with a Participatory Sensing data set can yield more information than one data set alone. Research questions provide an overall direction to make comparisons between data sets.

Lesson 3: Evaluating and Implementing a Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be evaluated before implementing to make

sure they are reasonable and ethically sound.

Lesson 4: Learning About Our Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be tried before implementing to make sure they are collecting the data they are meant to collect and refined accordingly.

Lesson 5: Statistical Predictions using One Variable

Anyone can make a prediction. But statisticians measure the success of their predictions. This lesson encourages the classroom to consider different measures of success.

Lesson 6: Statistical Predictions by Applying the Rule

If we use the squared residuals rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute error rule, then the median of the current data is the best prediction of future values.

Lesson 7: Statistical Predictions Using Two Variables

When predicting values of a variable y, and if y is associated with x, then we can get improved predictions by using our knowledge about x. Basically, we "subset" the data for a given value of x, and use the mean y for those subset values. If the resulting means follow a trend, we can model this trend to generalize to as-yet unseen values of x.

Lesson 8: What's the Trend?

Associations are important because they help us make better predictions; the stronger the trend, the better the prediction we can make. "Better" in this case means that our mean squared residuals can be made smaller.

Lesson 9: Spaghetti Line

We can often use a straight line to summarize a trend. "Eye balling" a straight line to a scatterplot is one way to do this.

Lesson 10: Predicting Values

The regression line can be used to make good predictions about values of y for any given value of x. This works for exactly the same reason the mean works well for one variable: the predictions will make your score on the mean squared residuals as small as possible.

Lesson 11: How Strong Is It?

A high absolute value for correlation means a strong linear trend. A value close to 0 means a weak linear trend.

Lesson 12: More Variables to Make Better Predictions

We can use scatterplots to assess which variables might lead to strong predictive models.

Sometimes using several predictors in one model can produce stronger models.

Lesson 13: Combination of Variables

If multiple predictors are associated with the response variable, a better predictive model will be produced, as measured by the mean absolute error.

Lesson 14: Improving your Model

If a linear model is fit to a non-linear trend, it will not do a good job of predicting. For this reason, we need to identify non-linear trends by looking at a scatterplot or the model needs to match the trend.

Lesson 15: The Growth of Landfills

Modeling does not always have to produce an equation. Instead, we can create models to answer real-world problems related to our community.

Lesson 16: Exploring Trash via the Dashboard

Exploring the IDS Dashboard provides a visual approach to data analysis.

Lesson 17: Exploring Trash via RStudio

RStudio can be used to verify initial results/findings from data analysis done via the IDS Dashboard.

Lesson 18: Grow Your Own Classification Tree

Many data sets have multiple predictors and are very non-linear. We can still use this data, but need to model it differently, such as in a decision tree. Decision trees are a useful tool for classifying observations into groups.

Lesson 19: Data Scientists or Doctors?

We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

Lesson 20: Where Do I Belong?

We can identify groups, or "clusters," in data based on a few characteristics. For example, it is easy to classify a classroom into males and females, but what if you only knew each student's arm span? How well could you classify their genders now?

Lesson 21: Our Class Network

Networks are made when observations are interconnected. In a social setting, we can examine how different people are connected by finding relationships between other people in a network.