0: Introduction to Statistical Learning

```
$ echo "Data Science Institute"
```

Intro to course support team

- Technical Facilitator: Holly
- Learning Support Staff: Kasra
- Learning Support Staff: Amanda
- Learning Support Staff: Vishnou

Welcome!

- So far, we've focused primarily on coding - now we explore the relationship between coding and statistics as this will allow us to answer questions such as "should we spend more of the advertising budget on TV or the internet"?
- This learning module will include definitions, mathematical concepts and approaches that may be new for most participants
- The learning curve will feel steep - this is expected - don't be hard on yourself if it takes time to sink in
- This module experienced the most changes since last cohort -- e.g. focus on more hands-on notebooks over mathematical theory

Rules of Engagement

- Mute yourself unless you want to ask questions!
- Questions are encouraged - ask as we go - this is your time to understand these concepts. However, please save *advanced* questions to office hours or work periods because sometimes they are excessively time consuming and may confuse most of your fellow participants who are beginners.
- If you have questions during live learning session, please try to ask them in the chats first where two of our Learning Supports will monitor the chats and answer questions there without interrupting the course. We will also pause and take time answering questions during the live course.

Let's navigate the ASC repo

https://github.com/UofT-DSI/applying_statistical_concepts

What is Statistical Learning?

Types of Statistical Learning

Applying Statistical Learning

Statistics and data science

What is Statistical Learning?

There are a variety of definitions, but broadly, we consider data science to be an interdisciplinary approach to generating insight from data.

Statistics is focused on collecting and analyzing data. The discipline has developed robust methods for using samples of data to make broader claims.

Statistics and data science

What is Statistical Learning?

Data science relies on statistics for robust ways of dealing with data. And it relies on software engineering for robust ways of approaching coding.

In data science, we write code to implement statistical methods. We need to know aspects of both the implementation (in code), and the underlying statistical methods, that we need in order to develop insights from our data.

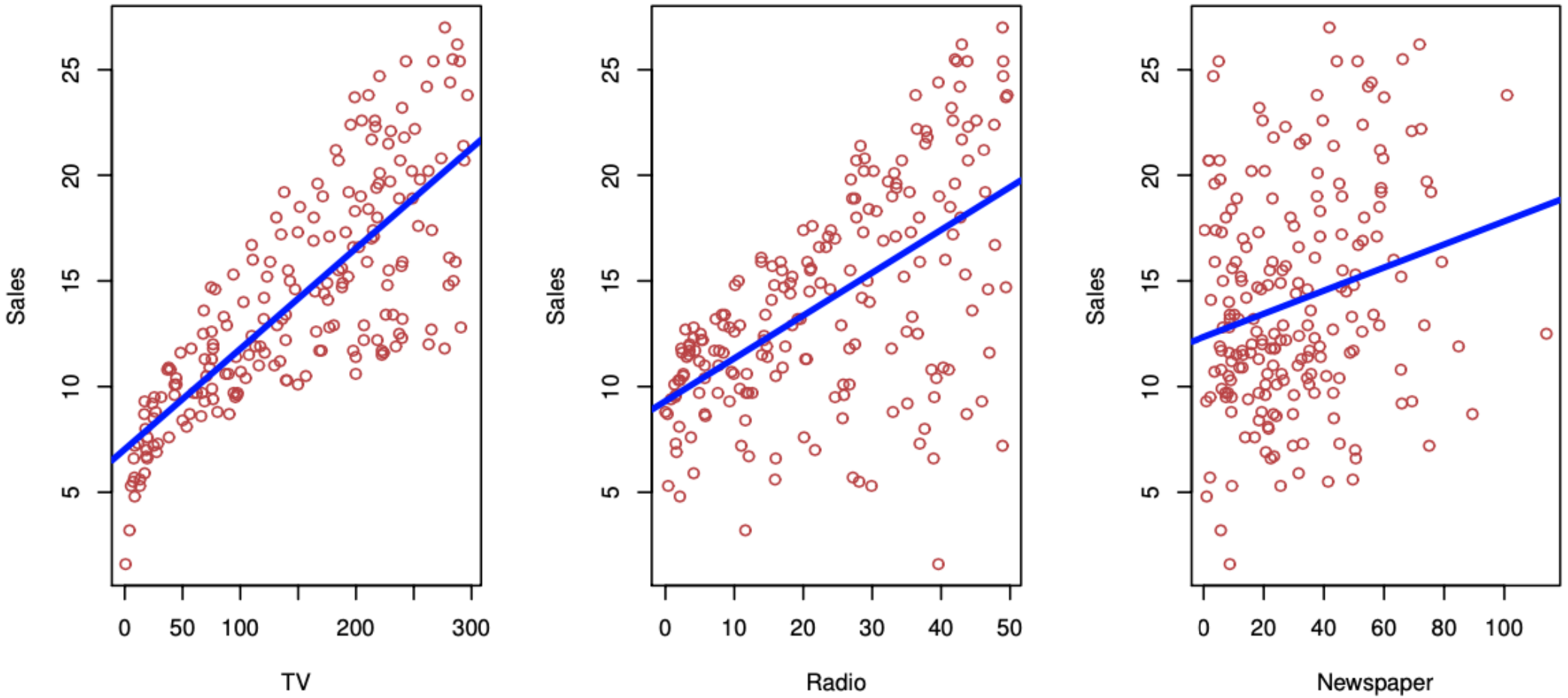
What is Statistical Learning?

Imagine we work in a marketing agency. Our client wants to know whether the money spent on advertising is leading to sales, and which advertising channels results in the most sales. They want to increase their sales and need to determine how to spend the money the right way to drive that increase. Suppose we want to figure out the **relationship between how the advertising budget is spent and sales** in order to increase sales for a client.

- There are three types of advertising: TV, radio, and newspaper which can be labelled $X_1, X_2, and\ X_3$ respectively.
- The advertising budgets (in thousands of dollars) are independent, or ***predictor variables***, which we label X (the horizontal axis).
- The number of sales (in thousands of units) is the dependent, or ***response variable***, which we label Y (the vertical axis).

What is Statistical Learning?

The sales in relation to each of the advertising budgets are shown along with a simple fitted line for the relationships.



What is Statistical Learning?

We want to find the relationship between the predictor variables (budget) and the response variable (sales). This relationship can be described as a function f. In reality the relationship is complicated, and cannot be perfectly described. We are using this function to model the relationship. The difference between the actual value and the estimation of that value can be described as a random error term ϵ.

This relationship between X and Y can be written as:

$$Y = f(X_1, X_2, X_3) + \epsilon$$

Statistical learning is summarized by the set of approaches which are used to estimate f.

Types of Statistical Learning

Prediction vs Inference

There are two main reasons why we want to model to estimate f:

1. If we want to know what sales can be expected for a given advertising budget? ***What response is expected given a set of predictors.*** This is **prediction**.
2. If we want to know to what extent sales volume is related to the advertising budget? ***How the response variable is affected by changes in the predictors.*** This is **inference**.

Types of Statistical Learning

Prediction

Prediction problems focus on the response Y. They can arise when the ***predictor variables X are known but the response Y is not easily obtained***. We use "^" to denote estimates. That is, $\hat{Y}$ is an estimate for Y and $\hat{f}$ is an estimate for f.

The accuracy of our prediction, $\hat{Y}$, depends on two types of errors: 1) those that we can potentially control, influence or **reduce** and 2) those that we cannot control or **reduce**. The **reducible error** is the error that we need to focus on as an analyst. But there is alway some **irreducible error**: the random error associated with the true response $Y = f(x) + \epsilon$. (Even if $\hat{f} = f, \hat{Y}$ will still have error associated with its prediction since ϵ is not a function of X.)

Our focus is on making predictions for Y using $\hat{f}$.

Types of Statistical Learning

Inference

Inference problems focus on predictors X. They can arise when both the ***predictor variables*** X ***and the response*** Y ***are known*** and we want to know how they are related.

The accuracy of our inference depends on how exactly we can estimate $\hat{f}$. It may depend on: 1) understanding which predictors are more important than others, 2) how does the response change (positively or negatively) given changes in the predictors, and 3) does the response change linearly or non-linearly given changes in the predictors.

Our focus is on finding the true form of f.

Applying Statistical Learning

How do we estimate f?

Assume that we have n observations in our data set. The standard approach is to split the data set into training data and testing data.

- **training data** is used to train or teach the model we are using to estimate f.
- **testing data** is used to test the accuracy of the resulting estimate for f on new data.

Applying Statistical Learning

Supervised vs Unsupervised Learning

- **Supervised learning** involves models for predicting a response based on predictor variables.
    - Examples of supervised learning models are linear regression and classification. These models are the primary focus of this learning module.
- **Unsupervised learning** refers to models used to investigate features associated with observations
    - There is no response variable to predict, instead the goal is to understand the relationship between variables or observations.
    - An example of this is clustering.

Applying Statistical Learning

Regression vs Classification Problems

- Variables can be either qualitative or quantitative.
  - **Quantitative** variables have numerical values (ex: age, monetary value, etc.)
  - **Qualitative** variables are categorical values (ex: {small, medium, large} or {yes, no})
- ***Problems that involve quantitative response variables are ♦ regression ♦ problems.***
- ***Problems that involve qualitative response variables are ♦ classification ♦ problems.***
- This is a bit of a generalization (logistic regression is a classification method but its output is numerical so can be thought of as a regression method as well)