Statistical Learning Overview.

W. Wang[1]

[1]Department of Mathematics
University of Houston

MATH 4323

Statistics

- Statistics is a branch of applied mathematics. It is the science concerned with studying and developing methods for collecting, describing, analyzing, interpreting, and presenting the data.
- The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory.
- Statistics is a highly interdisciplinary field. Statistical techniques are used in a wide range of types of scientific and social research:
  - Biostatistics
  - Econometrics
  - Engineering Statistics
  - Epidemiology
  - Political science
  - Many more..

Types of data

- **Population Data** is everything or everyone we want information about. It is a set of data that consists of all possible values pertaining to a certain set of observations or an investigation.
- **Sample Data** is a subset of the population that we have information from. It is just a small section of the population taken for the purpose of investigation.
- University of Houston is interested in how many students buy used books as opposed to new ones. They randomly choose 100 students at the student center to interview. Population? Sample?

Variables

- A **variable** is any characteristics, number, or quantity that can be measured or counted.
- Age, gender, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.
- It is called a **variable** because the value may vary between data units in a population, and may change in value over time.

Types of Variables

- **Categorical variables** place a case into one of several groups or categories. For example, clothing size, sex, eye color, religion, and favorite ice cream flavor.

- **Quantitative Variables** take numerical values for which arithmetic operations such as adding and averaging make sense. Quantitative variables can be further classified as either **discrete** or **continuous**.

Types of Quantitative Variables

- Discrete quantitative variables: Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of of which measured as whole units (i.e. 1, 2, 3 cars).
- Continuous quantitative variables: Observations can take any value between a certain set of real numbers. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.

Types of Relationships

- A **deterministic**(functional) relationship involves an exact relationship between two variables. For example, the relationship between degrees Fahrenheit and degrees Celsius:

$$Fahr = \frac{9}{5} Cels + 32$$

  if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit exactly.
- **Statistical** relationship: the relationship between the variables is not perfect. In general, the observations for a statistical realtion do not fall directly on the curve of relationship.

Examples of Statistical relationships

- Height and weight — as height increases, you'd expect the weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
- Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Types of Statistics

A statistic is a value that has been produced from a data collection, such as a summary measure, an estimate or projection. Statistical information is data that has been organised to serve a useful purpose.

- **Descriptive** statistics summarize data from a sample using indexes such as the mean or standard deviation, or using graphical descriptions such as bar chart, histogram, scatter-plot.

  Descriptive (or summary) statistics summarise the raw data and allow data users to interpret a dataset more easily.
  Descriptive statistics can describe the shape, center and spread of a dataset.

Types of Statistics

- **Inferential** statistics draw conclusions from data that are subject to random variation (e.g., observational errors, sampling variation).

  Inferential statistics are used to infer conclusions about a population from a sample of that population. Inferential statistics are the result of techniques that use the data collected from a sample to make generalisations about the whole population from which the sample was taken.
  - Estimation: Point estimate and confidence interval;
  - Hypothesis testing: recall those steps for hypothesis testing

Regression and Causality

- The existence of a statistical relation between the response variable Y and the predictor variable X does not imply in any way that Y depends causally on X.
- No matter how strong is the statistical relation between X and Y, **no cause-and-effect** pattern is necessarily implied by the regression model.
- For example, data on size of vocabulary (X) and writing speed (Y) for a sample of young children aged 5-10 will show a positive regression relation. Does this relation imply that an increase in vocabulary causes a faster writing speed?
- No, this relation does not imply that an increase in vocabulary causes a faster writing speed. There are other explanatory variables, such as age of the child and amount of education, affect both vocabulary and the writing speed. Obviously, older children have a large vocabulary and a faster writing speed.

Regression and Causality

- Even when a strong statistical relationship reflects causal conditions, the causal conditions may act in the **opposite** direction, from Y to X. For example, readings of the thermometer are collected at different known temperatures and regression analysis is performed to assess the prediction accuracy of the thermometer. For this purpose, what would be our response variable Y? and what is the predictor variable X?

  reading is the predictor variable; actual temp is the one to be predicted.
- The causal pattern does not go from X to Y, but in the oppsite direction: the actual temperature (Y) affects the thermometer reading (X).
- **Regression analysis** by itself provides no information about casual patterns and must be supplemented by additional analyses to obtain insights about causal relations.

Use of Computers

- Data analysis often entails lengthy and tedious calculations, computers are usually utilized to perform the necessary calculations.
- Almost every statistics package for computers contains a data analysis component. While packages differ in many details, their basic analysis output tends to be quite similar.

Review:

Some basic results in probability ad statistics

- $\mu =$ population mean; $\bar{x} =$ sample mean
- $E(x) =$mean=average$= \mu$
- $E(x-\mu)^2 =$ variance $= \sigma^2 = E(x^2) - [E(x)]^2 = E(x^2) - \mu^2$
- $cov(x,y) = E((x-\mu_x)(y-\mu_y)) = E(xy) - E(x)E(y)$
- suppose $X, Y, Z, ..., T$ are random variables,
  $E(X+Y+Z+\cdots+T) = E(X)+E(Y)+E(Z)+\cdots+E(T)$
- $E(a+bX) = a+bE(X)$
- $var(a+bX) = b^2 var(X)$

Matrix Algebra

- A **matrix** is a rectangular collection of numbers. Generally, matrices are denoted as bold capital letters.
- The **dimension** of a matrix is expressed as number of rows $\times$ number of columns.
- A **vector** is a matrix with only one row (called a row vector) or only one column (called a column vector).
- An "ordinary" number can be thought of as a 1×1 matrix, also known as a **scalar**.

Supervised/Unsupervise Learning.

Statistical learning is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning.

Statistical learning refers to a vast set of tools for modeling and understanding complex datasets. These tools can be classified into

- **Supervised** Learning - building a statistical model to predict/estimate an output based on one or more inputs (output is supervised)
- **Unsupervised** Learning - building a statistical model to capture/infer a relationship between inputs, without being given any output (output is not supervised).

Supervised learning

- **Supervised** Learning algorithm is based on the observations of a series of examples in which each data input has been previously labeled. Through this learning process, it generates a function that links the input values to the desired output. It is used to build predictive models.
- In **Supervised** Learning, our goal is to build a model, with the labeled training data and use the model to make predictions about data that is not available or which is in the future.
- **Supervision** means that in our samples (the dataset), the desired output are already known as previously labeled.

Wage data.

Example. We examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. In particular, we wish to understand the effect that

- employee's age,
- employee's education,
- calendar year

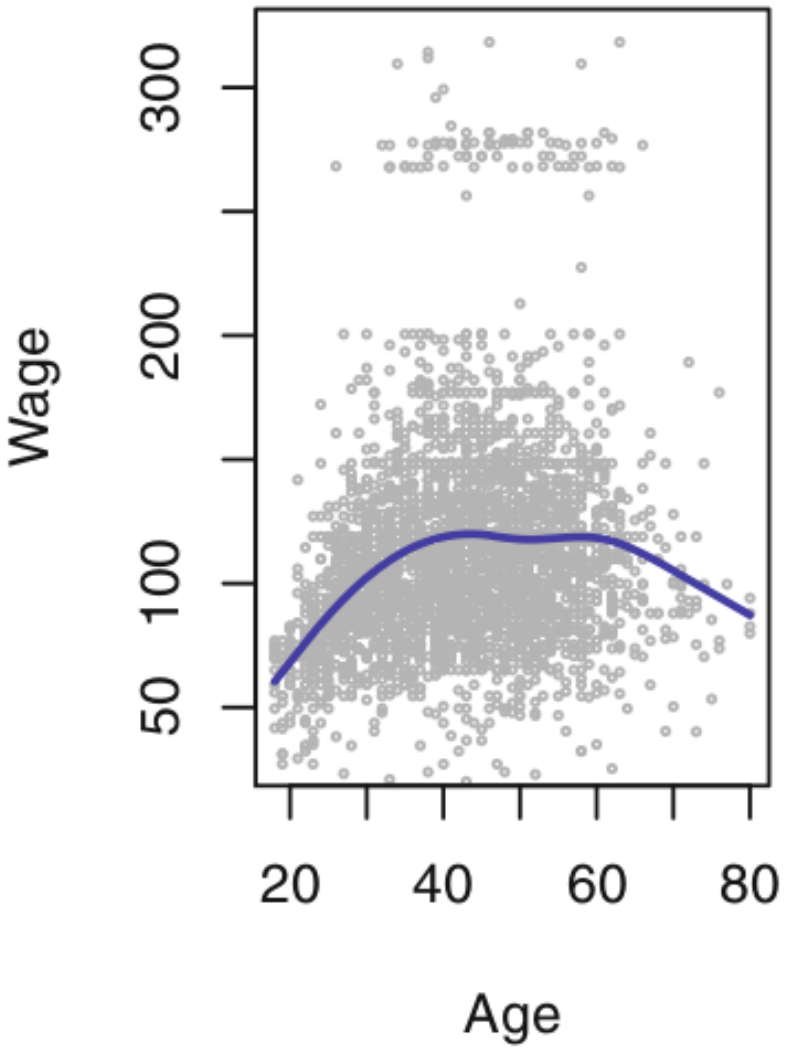
have on employee's wage.

In such setup, we consider

- age, education and calendar year to be our **predictor** (also called explanatory, independent) variables,
- wage - our **response** (also called dependent) variable.

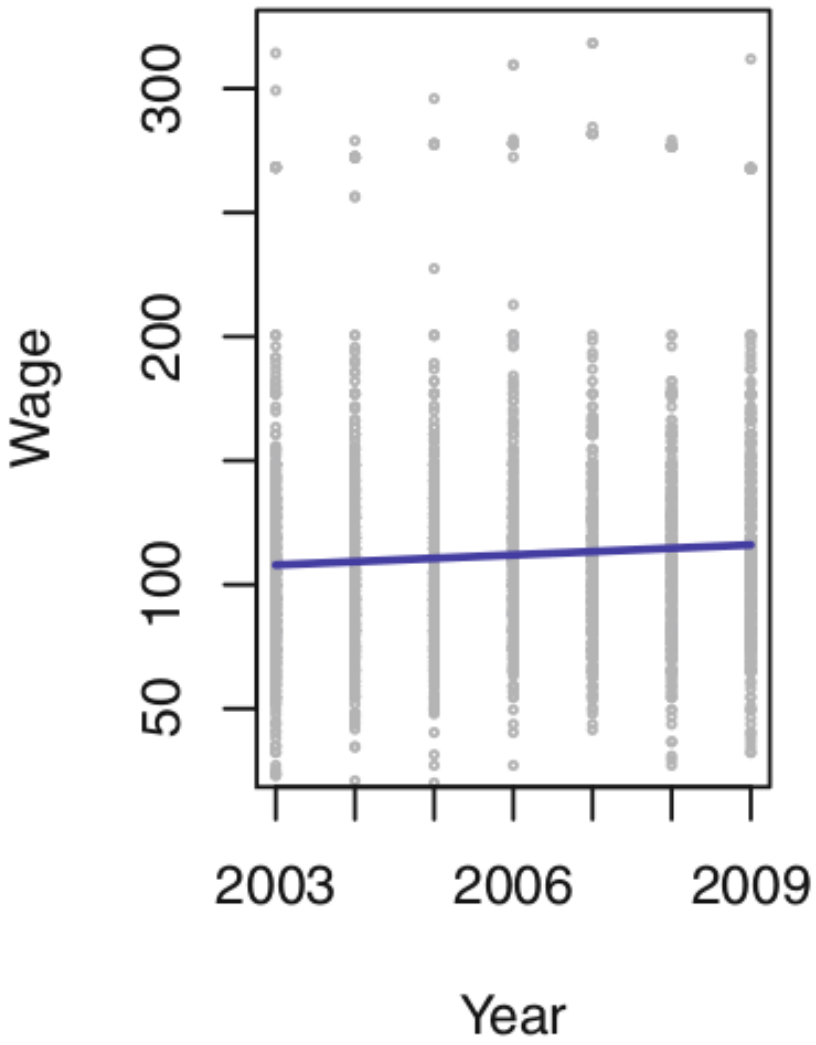
Prior to jumping into any statistical learning, first step of a self-respecting data scientist is to **visualize the data**.

Wage data: Visualization (Scatterplot).



- Wage increases until the age of ≈ 45, then tapers off and decreases.
- Blue line $\equiv$ average wage for a given age, makes this trend clearer. We could use it to predict employee's wage from their age.
- There is a significant amount of variability around average wage value $\Longrightarrow$ age alone is unlikely to provide an accurate prediction of employee's wage.

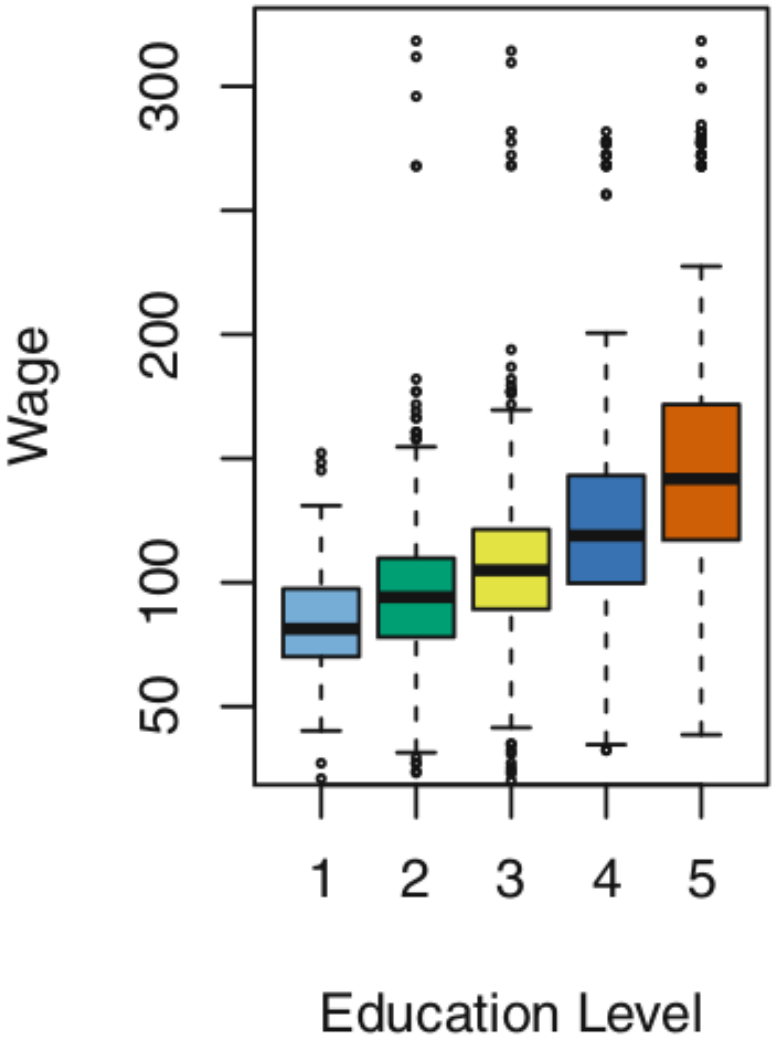
Wage data: Visualization (Scatterplot)



Wages increase from year 2003 to year 2009 by approximately $\$10,000$, in a roughly linear (or straight-line) fashion.

Nonetheless, this rise is very slight relative to the variability in the data.

Wage data: Visualization (Boxplot).



Wages are also typically greater for individuals with higher education levels:

employees with the lowest education level (*education* $=$ 1) tend to have substantially lower wages than those with the highest education level (*education* $=$ 5).

Some useful models (not in this course though...)

The most accurate prediction of a given employees' wage might be obtained by combining their age, education, and calendar year as predictors. For this, one could use such models as:

- Multiple Linear Regression,
- Trees/Random Forests,
- Neural Networks.

Also, to reflect the non-linear wage vs age relationship, one could use:

- Polynomial Regression,
- Splines.

Unfortunately, this course will focus on a different set of tasks and models.

Stock market data: Classification Task.

The *Wage* data involves predicting a continuous (else known as quantitative), output value $\Longrightarrow$ **regression** task.

In certain cases we wish to predict a categorical (else known as qualitative) output value $\Longrightarrow$ **classification** task.

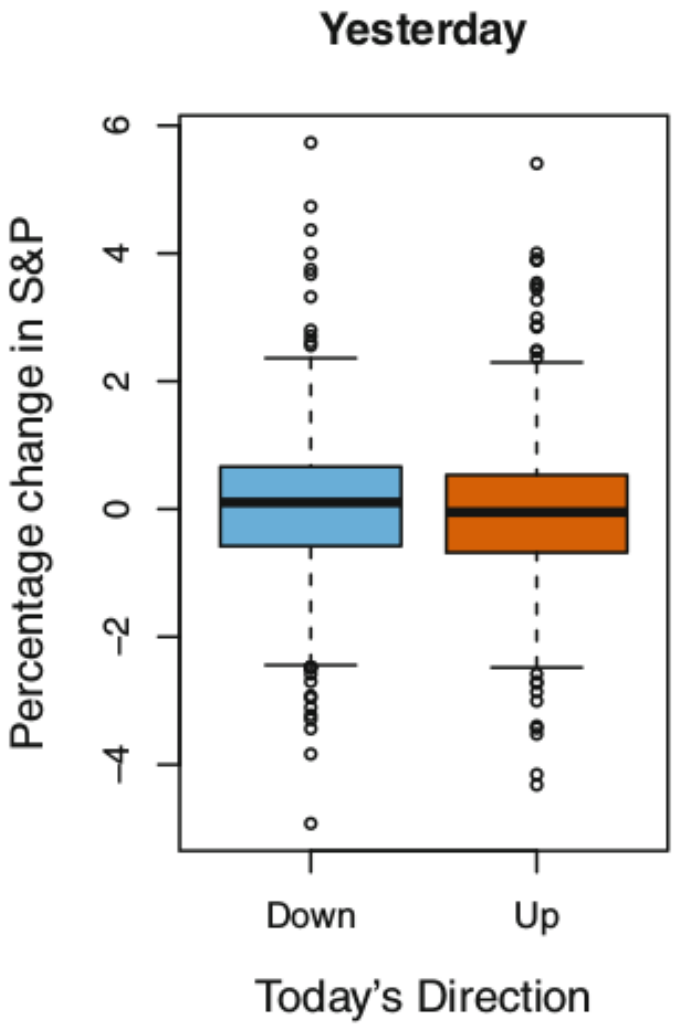
Example. *Smarket* data set contains the daily movements in the Standard & Poor's 500 (S&P) stock index over a 5-year period between 2001 and 2005.

Goal:

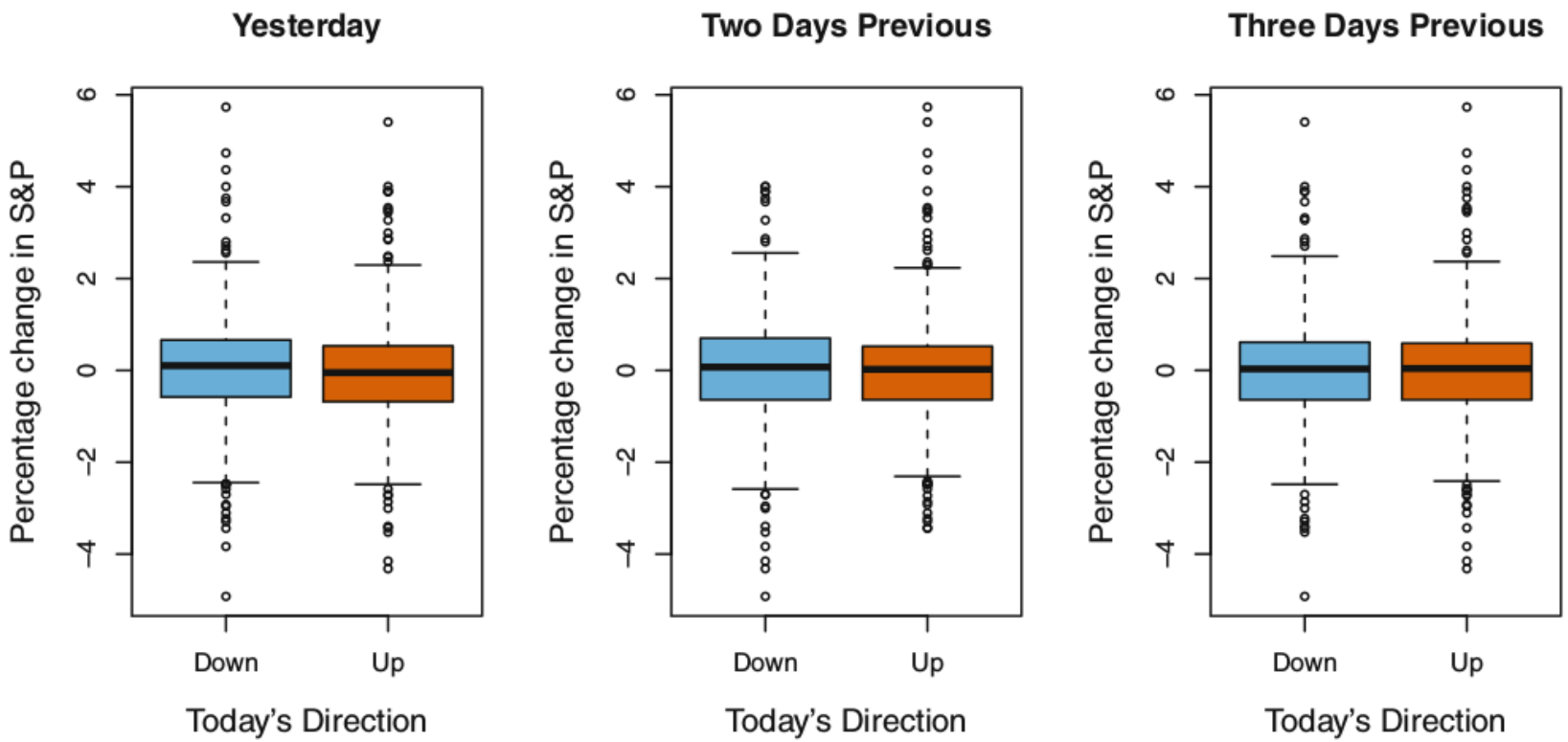
- given the past 5 days' %-changes in the index (predictors);
- predict if index will increase or decrease today (outcome).

Hence, we don't predict a numerical value, but only whether stock market will go up or down $\Longrightarrow$ categorical response, with two categories $\Longrightarrow$ classification task.

Stock market data: Visualization (Boxplot).



Stock market data: Visualization (Boxplot).



No clear pattern detected.

Nevertheless, later we explore a method to produce a better than random guess (50%) performance of market movement prediction.

Supervised Learning: Regression and Classification.

Both previous examples - *Wage* and *Smarket* data - dealt with **supervised** learning task (Why?). In particular,

- *Wage* corresponded to **regression** task (Why? Provide other potential data examples for regression tasks.)
- *Smarket* corresponded to **classification** task (Why? Provide other potential data examples for classification tasks.)

Supervised learning implies data sets with **both** input and output variables.

However, there's plenty of important problems where we **only** observe input variables, with no corresponding output $\Longrightarrow$ **unsupervised** learning.

Unsupervised Learning.

Examples of unsupervised learning:

- Marketing: given demographic & spending information (predictors) on customers, proceed to group them by similarity.
- Gene Expression data: given 6830 gene expression measurements (predictors) for 64 cell lines, proceed to group the cell lines by similarity of their gene expressions.

In both of those examples, no observable response variable is given. Those tasks are known as **clustering**

Gene Expression example.

Example. Focusing on the gene expression data, where we have

- $n = 64$ cell lines (observations),
- $p = 6830$ gene expression measurements (variables)

When we just had two variables, e.g. *wage* and *age*, it was easy to plot the observations.

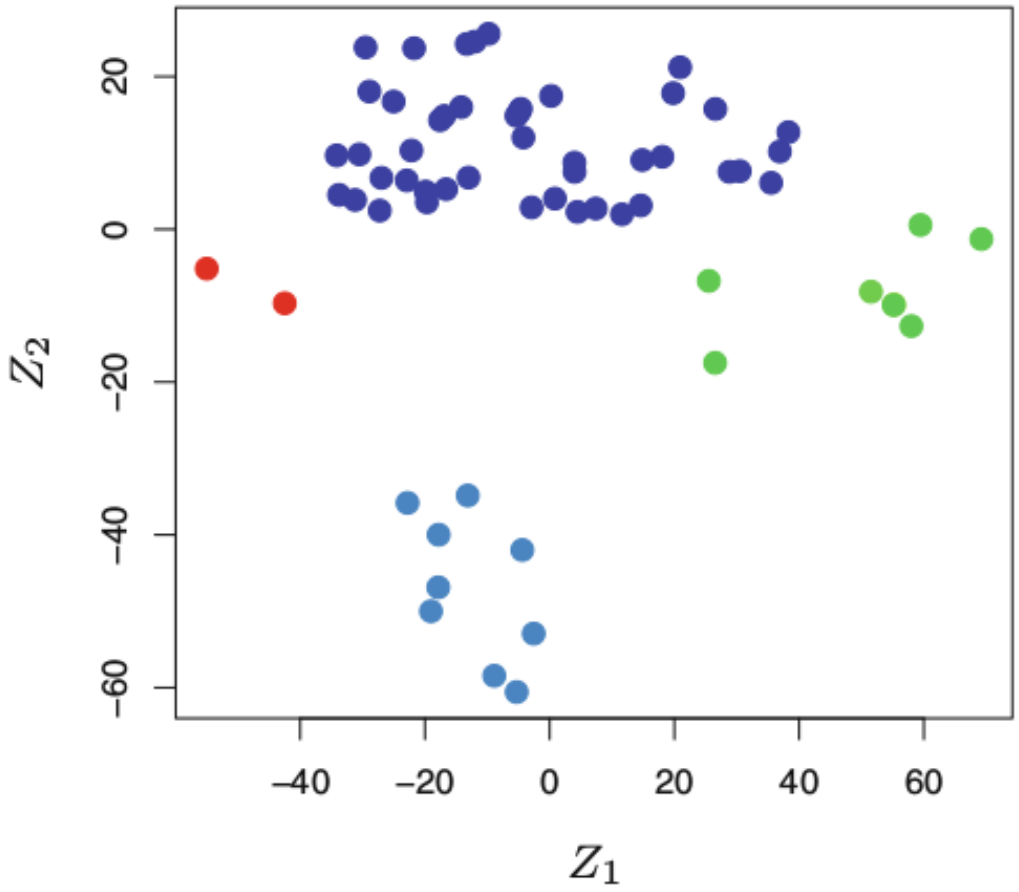
With $p = 6830$, it is virtually impossible to visualize the observations.

This issue can be addressed via **principal component analysis (PCA)**, which

- shrinks the total number p of variables down to a set of **few** principal components (PCs),
- with those PCs retaining as much information from all p variables as possible.

Gene Expression example: PCA.

Example (cont'd). Using just the first two PCs, Z_1 and Z_2 (which amounts to a considerable loss of information, but still):



Deciding on the number of clusters is often a difficult problem. But here we suggest $\approx$ 4 groups, each marked with separate color.

Now we could examine each cluster for types of cancer and their relationship to gene expression levels.

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the *regression problem*, Y is quantitative (e.g price, blood pressure).
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- *There is much overlap* — both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on *large scale* applications and *prediction accuracy*.
  - Statistical learning emphasizes *models* and their interpretability, and *precision* and *uncertainty*.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".
- Machine learning has the upper hand in *Marketing!*