



Rensselaer

why not change the world?®

Data Analysis: Concepts, Statistics, Visualizations, Regression

Ahmed Eleish

**Data Science – ITWS/CSCI/ERTH 4350/6350 Module 5,
September 30th, 2026**

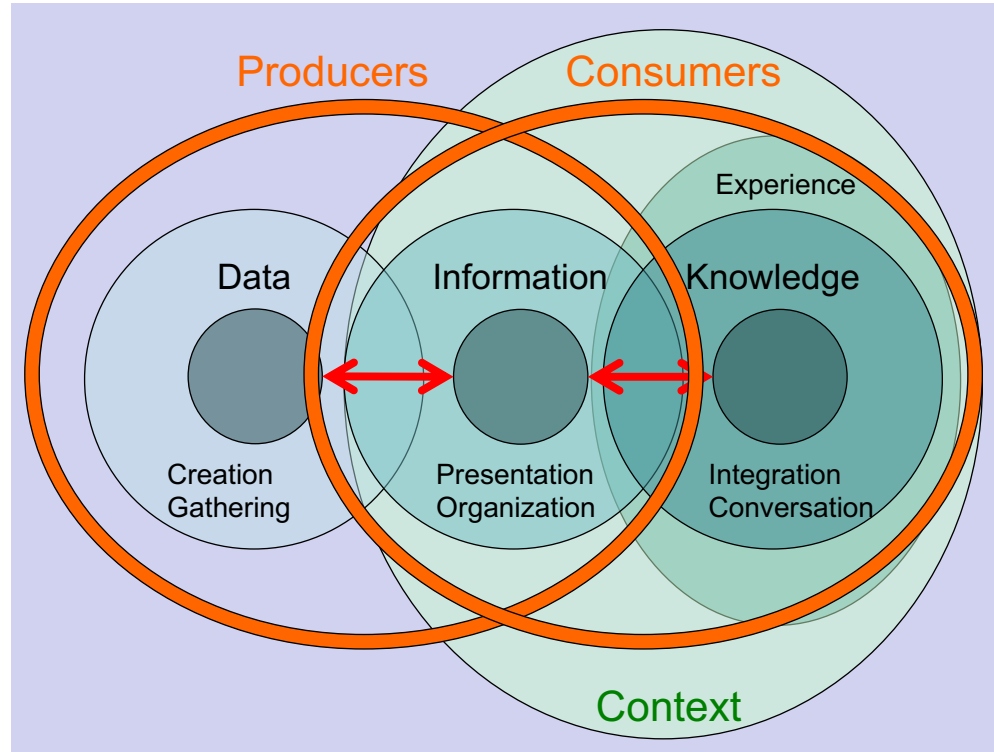


Contents

- Data analysis concepts
- Exploratory analysis
- Visualizations
- Statistics concepts and methods
- Regression analysis



Data-Information-Knowledge Ecosystem



Definitions

- **Data** - are encodings that represent the qualitative or quantitative attributes of a variable or set of variables.
- **Variable**: An attribute of a population or sample that can be measured.
- **Observation**: a measurement or set of measurements made under specific conditions. Each measurement is associated with a different variable. **An observation = a data point.**
- **Dataset** – a set of data points (observations) collected for a particular purpose.



Data

- Types of Data:
 - Qualitative (or Categorical)
 - Quantitative (numeric values)



Data

- Qualitative:

Qualitative (or Categorical) data are descriptive, but not numeric.

Example: your eye-color, your gender, color of a vehicle, your birthplace



Data

- Quantitative:

Quantitative data take on numeric values

- Discrete data take whole numbers (0,1,2,3..), representing countable things.

e.g. The number of times an employee is late to work, number of cars in the parking garage.

- Continuous data can take infinite possibilities within a range of numeric values (fractions, decimals are included..)

e.g. height of a person, weight of an apple, amount of time spent studying.



Levels of Measurement

Qualitative (or Categorical) data can be measured at the:

- Nominal Level: Values are just names, without any order (example: eye color)
- Ordinal Level: Values have some natural order, example: academic year (freshman, sophomore, ..), test grade, military rank

Quantitative (data like numeric values) can be measured at the:

- Interval Level: Numeric data with no natural zero point; intervals (differences) are meaningful but ratios are not; can be negative, e.g. Temperature in Fahrenheit (or Celsius) degrees 80F is 20F hotter than 60F, but it is not 150% as hot.
- Ratio Level: Numeric data for which there is true zero; both intervals and ratios are meaningful; cannot be negative, e.g. weight, length, duration..



Types of Data

Type of Data	Level of Measurement	Examples
Categorical	Nominal (categories with no inherent order)	Eye color, ethnicity, academic major
	Ordinal (categories with inherent order)	Grade, academic year
	Binary (2 categories; special case of above)	Electric charge
Quantitative	Discrete (whole numbers)	Size of household (ratio)
	Continuous (can in theory take on any value in a range, although necessarily recorded to a predetermined degree of prevision.	Temperature C°/F° (no absolute zero - interval)
		Age, height (ratio)



Special values in data

- Fill value
 - Default
 - Missing value
 - Error value
 - Not-a-number (NAN)
 - Infinity
 - Null (NA)
- These vary by language/platform and usually require special attention (removal, debugging, etc.)



Outlier

- An extreme, or atypical, data value in a sample.
- They should be considered carefully, before exclusion from analysis.
- For example, data values may be recorded erroneously, and hence they may need correction.
- However, in other cases they may just be surprisingly different, but not necessarily 'wrong'.

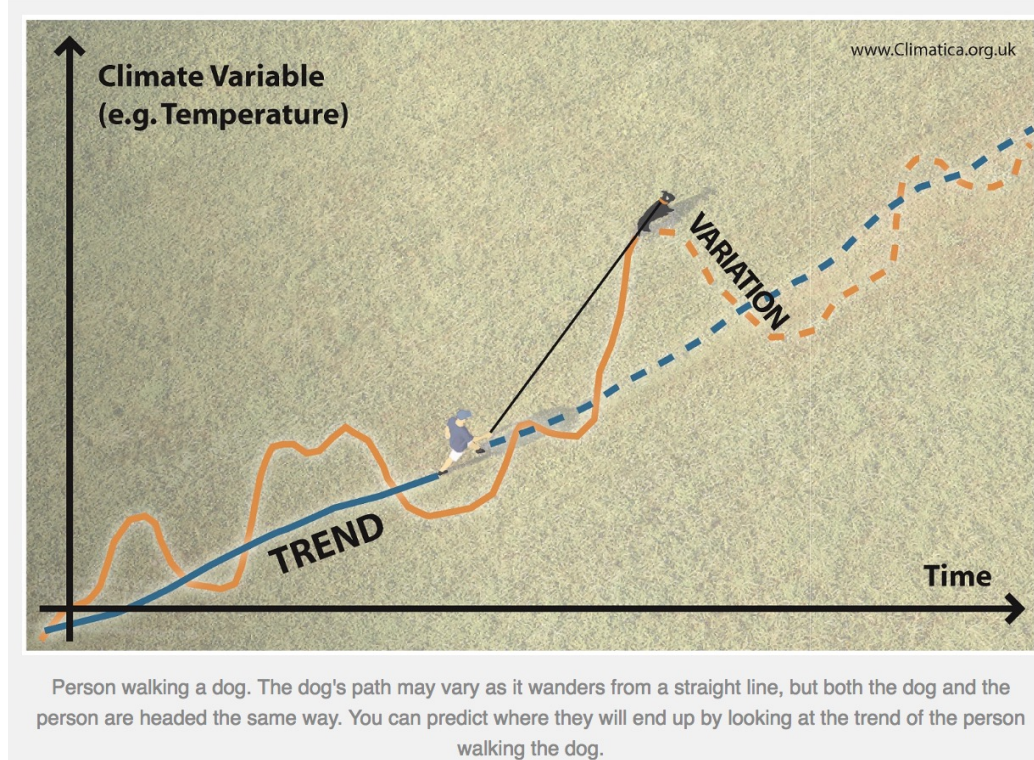


Variation

- **Variation** is the tendency of the values of a variable to change from measurement to measurement. You can see variation easily in real life; if you measure any continuous variable twice, you will get two different results. This is true even if you measure quantities that are constant, like the speed of light. Each of your measurements will include a small amount of error that varies from measurement to measurement.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham*, Chapter 5

Data Analysis: Chasing Trends



<http://climatica.org.uk/climate-science-information/uncertainty>

‘Signal to noise’

- Affects choices of analysis
- Affects interpretations
- Can be an indicator of dataset quality
- Signal and noise are context dependent
- Understanding accuracy and precision
 - Accuracy
 - Precision

Accurate vs. Precise



**High Accuracy
High Precision**



**Low Accuracy
High Precision**



**High Accuracy
Low Precision**



**Low Accuracy
Low Precision**

<http://climatica.org.uk/climate-science-information/uncertainty>

Exploratory Data Analytics (EDA)

- To use visualization and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle where you:
- Generate **questions** about your **data**.
- **Search** for **answers** by **visualizing**, **transforming**, and **modeling** your data.
- Use what you learn to refine your questions and/or generate **new questions**.

Exploratory Data Analytics (EDA)

- **EDA is not a formal process with a strict set of rules.** More than anything, **EDA is a state of mind.** During the initial phases of EDA you should feel free to investigate every idea that occurs to you.
- Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will hone in on a few particularly productive areas that you'll eventually write up and communicate to others.
- **EDA is an important part of any data analysis**, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data.
- **Data cleaning is just one application of EDA**

Resource: R for Data Science, *Garrett Golemund, Hadley Wickham*, Chapter 5

- *There are no routine statistical questions, only questionable statistical routines.*

~ Sir David Cox

- *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*

~John Tukey

Resource: R for Data Science, Garrett Grolemund, Hadley Wickham, Chapter 5

Exploratory Data Analytics (EDA)

- **Your goal during EDA is to develop an understanding of your data.**
- The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, **the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham*, Chapter 5

Exploratory Data Analytics (EDA)

creative process..

- **EDA is fundamentally a creative process.** And like most creative processes, the key to asking *quality* questions is to generate a large *quantity* of questions.
- At the beginning, It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham*, Chapter 5

Exploratory Data Analytics (EDA)

No rule on which Question..

- **There is no rule about which questions you should ask to guide your research.**
- However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:
 - 1) What type of variation occurs within my variables?**
 - 2) What type of covariation occurs between my variables?**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham*, Chapter 5

Visualizations

Getting started with analysis

- Examining the data visually
 - Which visualization techniques are suitable for one's data?
- Box plots
- Frequency plots
- Histograms
- Scatter plot diagrams
- ...



Why visualization?

- Reducing amount of data
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products

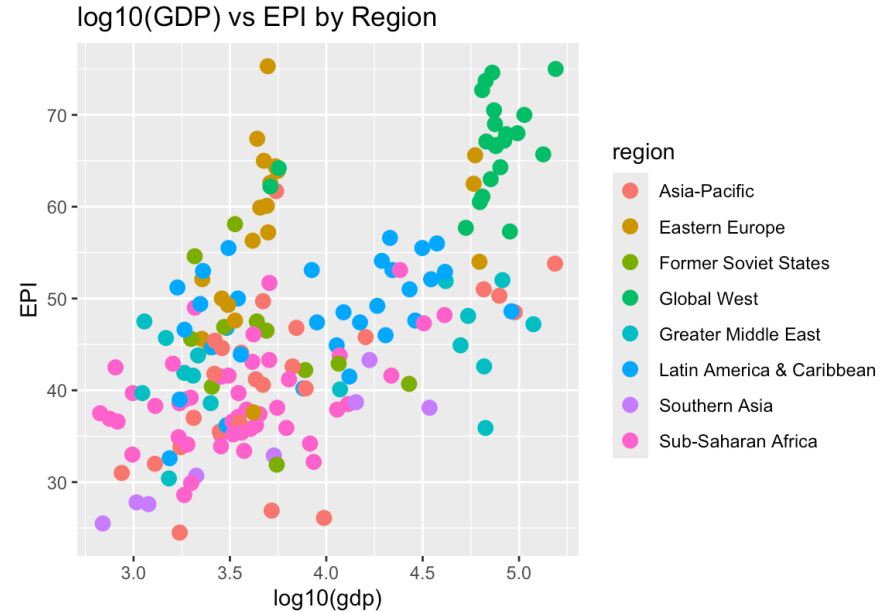
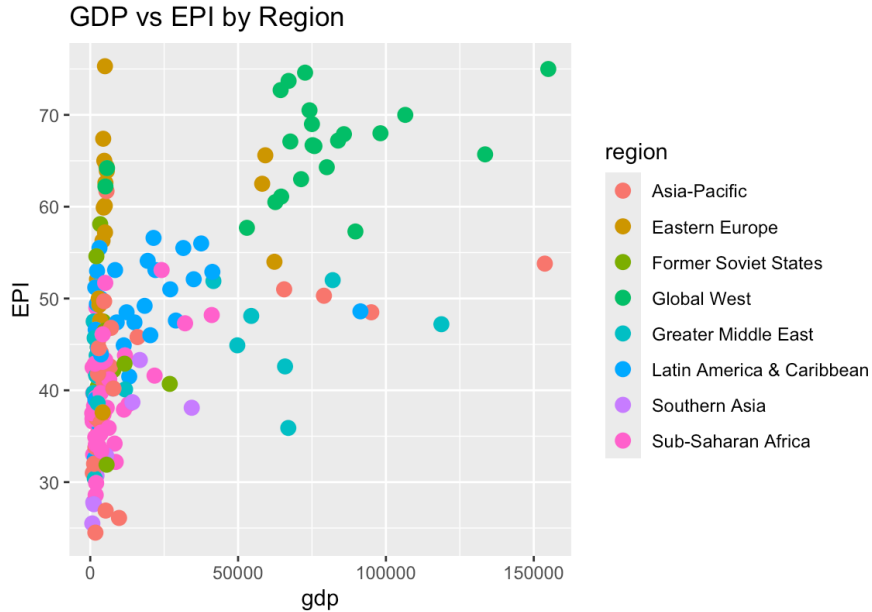


Visualizations

- Scatter Plot – Paired data (x,y)
- Describe the relationship between numerical variables.
- Make a note on the direction of the data points
 - Positive direction
 - Negative Direction
- Check for unusual observations
- See the relationship - Linear or Non-linear

Visualizations

- Scatter Plot – Paired data (x,y)

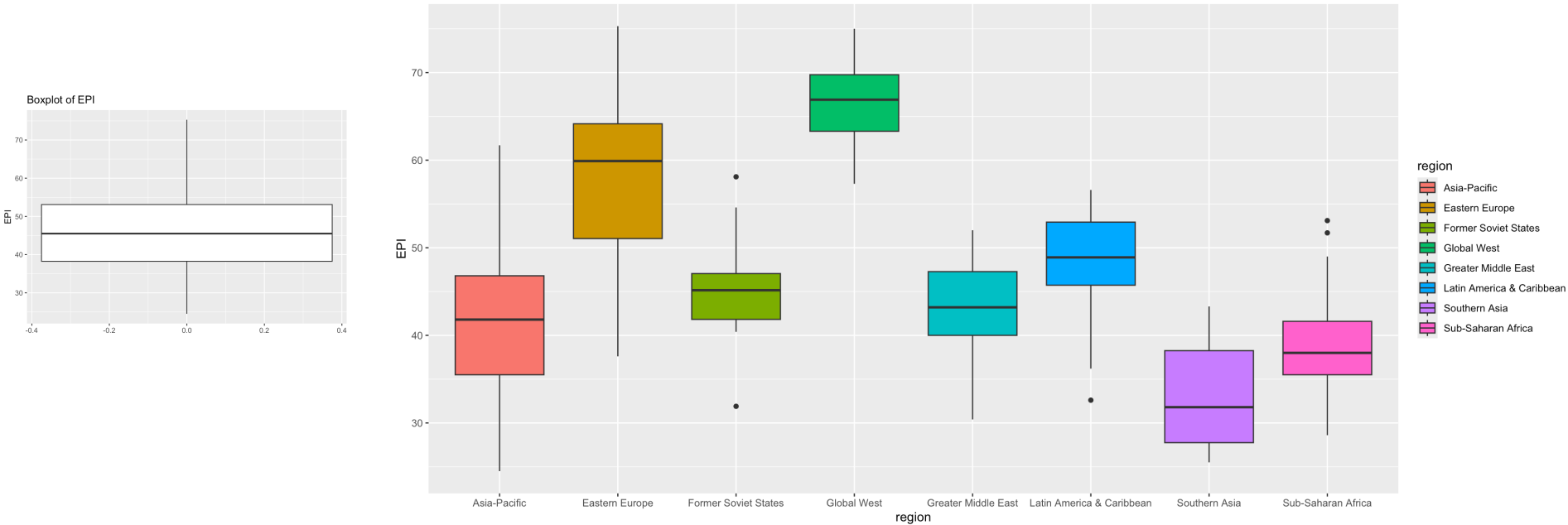


Visualizations

- Boxplot - box-and-whisker of a variable x
- Gives an overview of the range of the variable and where most of the observations are exist along that range
- Make a note of the minimum, maximum, quartiles
- Check for unusual observations (outliers)

Visualizations

- Boxplot - box-and-whisker of a variable x



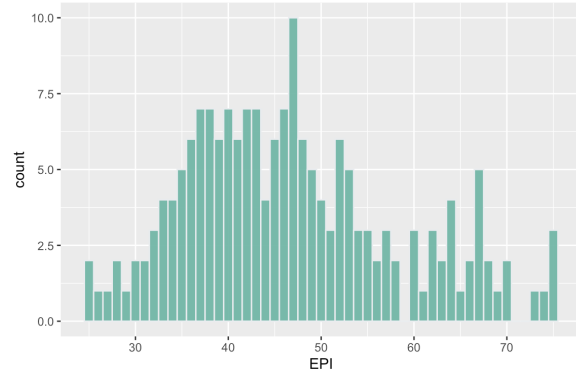
Visualizations

- Histogram – Distribution of variable X
 - Describe frequency of occurrence of values or ranges of values (bins) of x
 - Tune bin size parameter
 - Observe overall shape
 - Check for mixed distributions

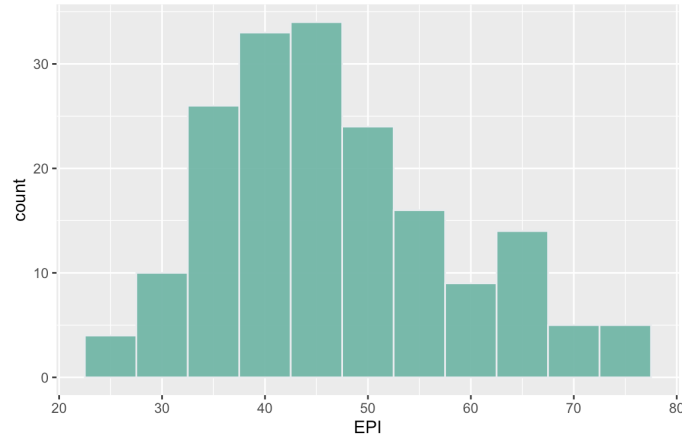
Visualizations

- Histogram – Distribution of variable X

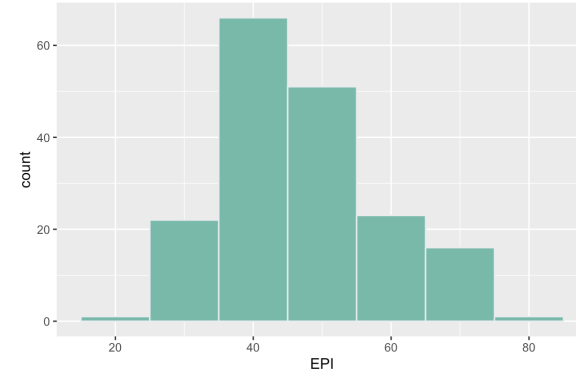
Histogram of EPI - Bin size = 1



Histogram of EPI - Bin size = 5



Histogram of EPI - Bin size = 10



Statistical Concepts

Populations and samples

- **Population** : The complete set of actual or potential elements about which inferences are made
- **Sample** : A subset of the population selected using some sampling methods.

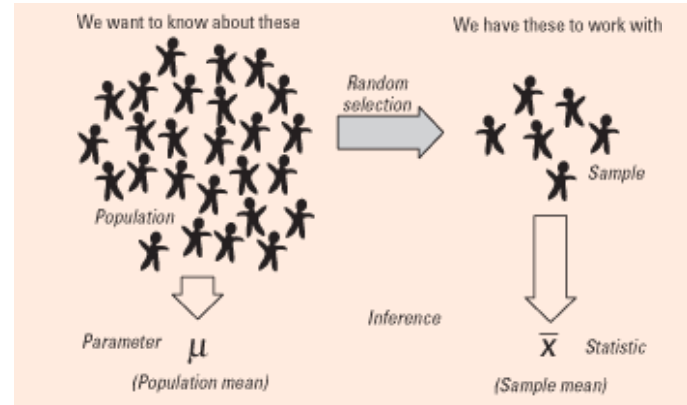


Image Credit/Reference: Quick Study Statistics

Sampling Types

- **Random Sampling**

- Sampling in which the data are collected using chance methods or random numbers.

- **Stratified Sampling**

- Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.

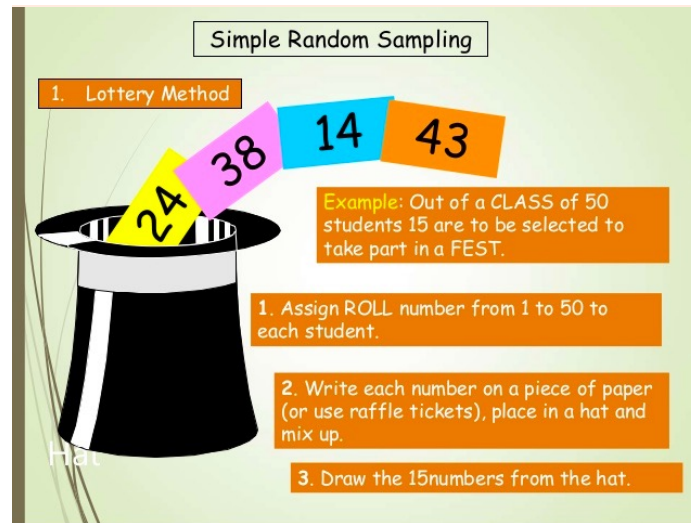
- **Cluster Sampling**

- Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are included in the sample.



Simple Random Sampling

A sample is selected so that each possible sample of the same size has an equal probability of being selected; used for most elementary inference

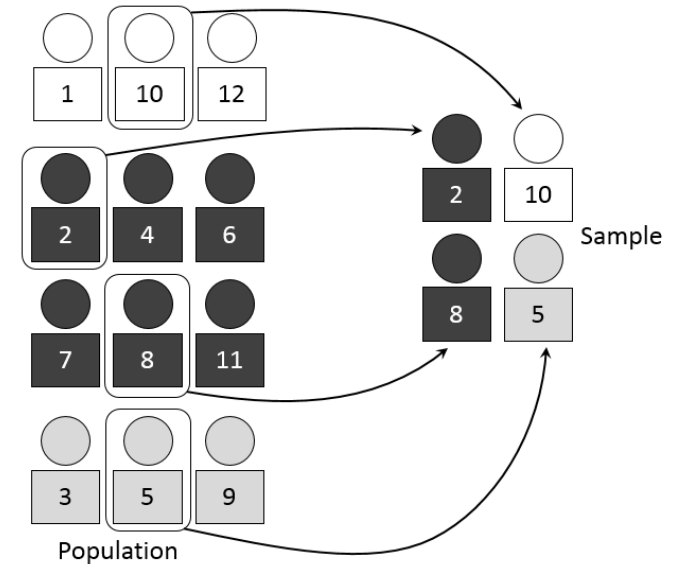


Courtesy: Quick Study Academic – Statistics www.quickstudy.com
Reference: Quick Study Statistics
Image Courtesy:
<https://www.slideshare.net/mohammedzuhairy1/sampling-techniques-64917617>

Stratified Sampling

The population is divided into strata based on some criterion, and a fixed number of elements from each stratum are selected for the sample (e.g. at random).

- Strata are internally homogenous.



https://en.wikipedia.org/wiki/Stratified_sampling

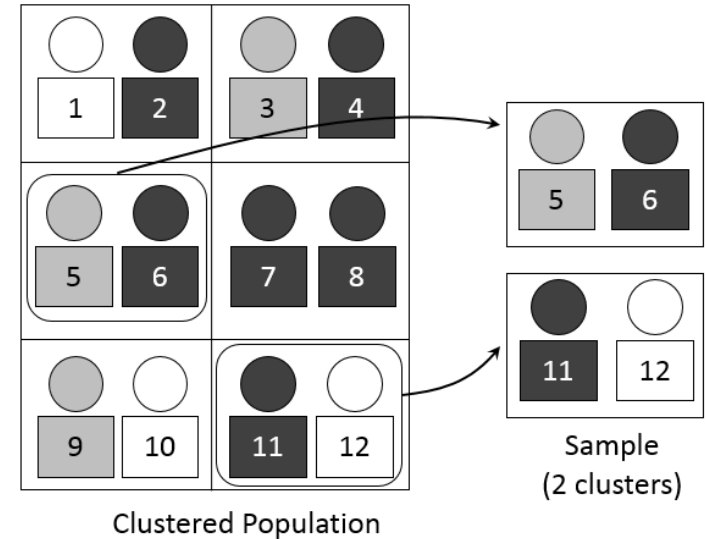
https://en.wikipedia.org/wiki/Stratified_sampling#/media/File:Stratified_sampling.PNG

Image Credit: Dan Kernler – license: [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/) - no changes

Cluster Sampling

The population is divided into groups called clusters; some clusters are randomly selected, and every member in them observed.

- Clusters are internally heterogenous.

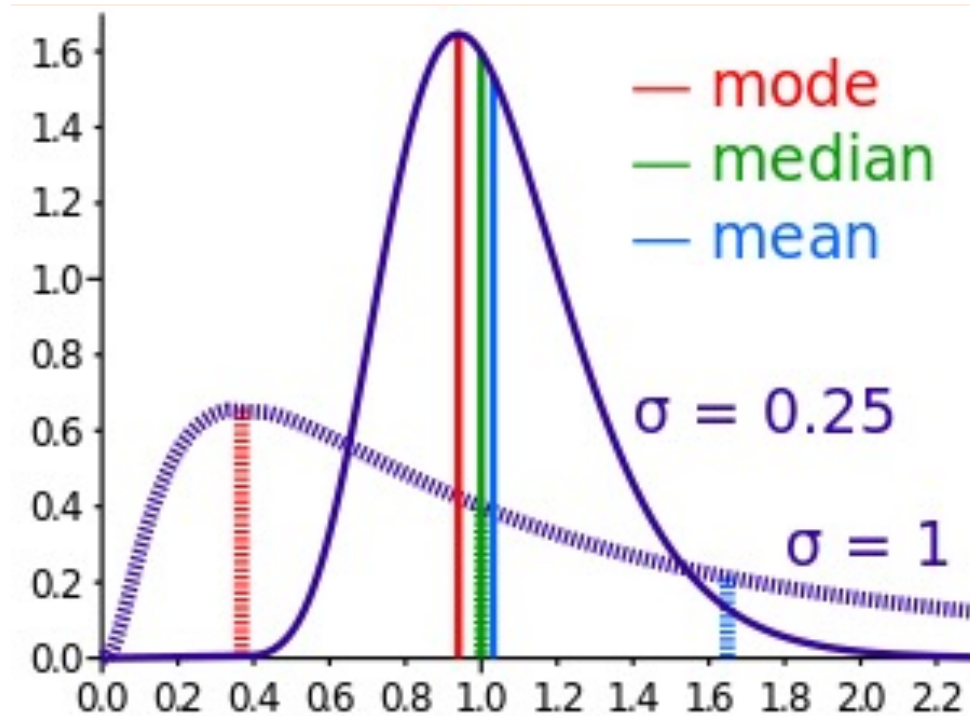


https://en.wikipedia.org/wiki/Cluster_sampling

https://en.wikipedia.org/wiki/Cluster_sampling#/media/File:Cluster_sampling.PNG

Image Credit : Dan Kernler – license: [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/) - no changes

Statistics - central tendency – mean, median, mode



Measures of Central Tendency

- Mean: The most commonly used measure of central tendency, commonly referred to as “Average”, sensitive to extreme values (sensitive to outliers)

- Population Mean

- Sample Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$
N = number of items in the population	n = number of items in the sample

Image Resource: <https://www.onlinemathlearning.com/population-mean.html> Courtesy: Quick Study Academic – Statistics www.quickstudy.com



Measures of Central Tendency

- **Median:** Value that divides the set in 2 so the same number of observations lie on each side of it.

- ***The median is less sensitive to extreme values***

- For an even number, it is the average of middle two values.

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$
= **4.5**

Courtesy: Quick Study Academic – Statistics www.quickstudy.com Image Resource:
<https://en.wikipedia.org/wiki/Median>



Measures of Central Tendency

- **Mode:** Observation that occurs with the greatest frequency.



Courtesy: Quick Study Academic – Statistics
www.quickstudy.com Image Resource:
<https://www.mathsisfun.com/definitions/mode.html>

Variance/Covariance

- Variance and Covariance are measures of the “spread” of a set of points around their center of mass (mean).
 - Variance is measure of the deviation from the mean for points in one dimension.
 - Covariance is a measure of how much multiple dimensions vary from their means with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained.
- The covariance between one dimension and itself is the variance

Variance/Covariance

Example:

$$\begin{aligned} - \text{VAR}(X) &= \frac{(12-36)^2 + (20-36)^2 + \dots + (22-36)^2}{6} \\ &= 862.333 \\ - \text{COV}(X,Y) &= \frac{(12-36)*(1.7-4) + \dots + (35-36)*(2.5-4)}{6} \\ &= 77.283 \end{aligned}$$

X	Y
12	1.7
20	3.4
54	6.8
14	0.9
94	8.5
35	4.2
20	2.5

$$\text{VAR}(X) = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

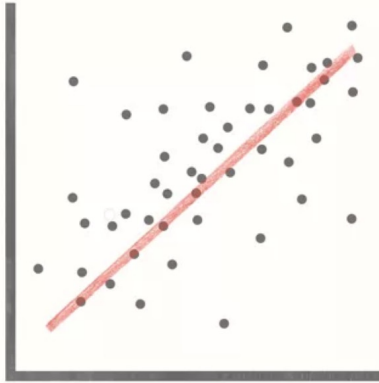
$$\text{COV}(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Correlation

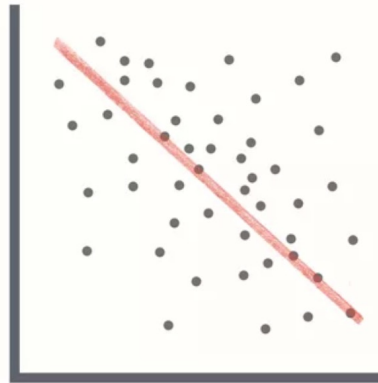
- One measure of the strength of the association between two numerical variables is correlation.
- Correlation describes the strength of the linear association between two variables.
- Correlation coefficient is between -1 and +1
-1 indicates a perfect negative linear association and +1 indicates a perfect positive linear association. A correlation coefficient of 0, indicates that there is no linear relationship in the two variables. -0.1 and +0.1, indicates no linear relationship *or a very weak* linear relationship
- Correlation coefficient is sensitive to outliers.
- Correlation coefficient is unitless.

Reference(s): <https://www.investopedia.com/terms/c/correlationcoefficient.asp>
<https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp>

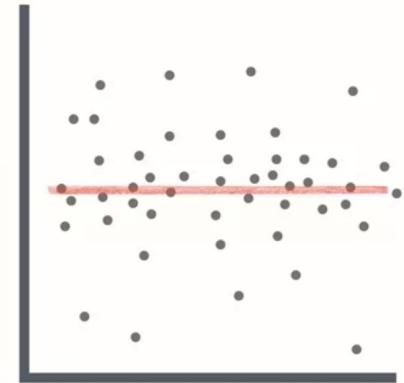
Correlation...



Positive Correlation



Negative Correlation



No Correlation

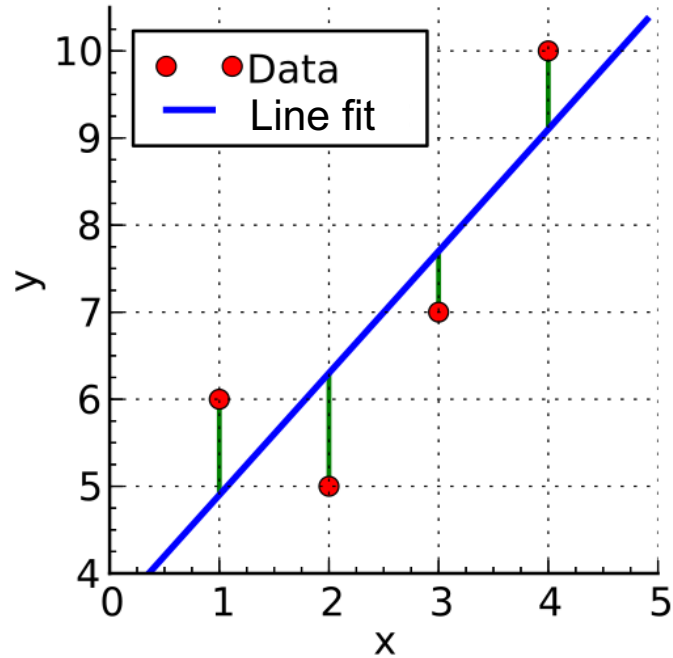
Image/Photo Credit: <https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp>

Linear Regression

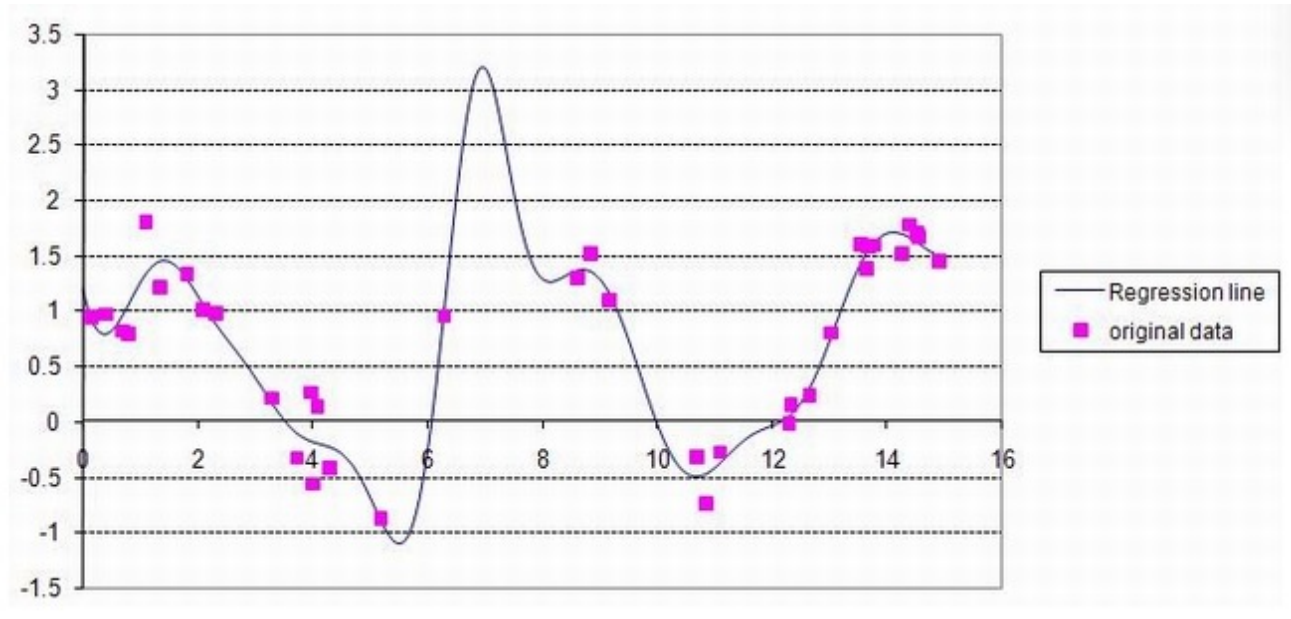
Regression in Statistics

- Regression is a statistical process for *estimating* the relationships among continuous numerical variables.
- In regression analysis the focus is on the relationship between a dependent variable and one or more independent variables.
- Independent variables are also called *predictors*, *covariates*.. these are **inputs**.
- Dependent variables are also called *target* or *response* variables, these are **outputs**.
- Estimation is often done by constraining an objective function.
- Must be tested for significance, confidence.

Regression



Regression - when it gets complex...



Real-World Uses

- Sales forecasting
- Price prediction
- Demand estimation
- Performance modeling



Linear Regression

- Fitting covariate and response data to a line is referred to as linear regression.

Definitions

Coefficient: numerical value that quantifies the relationship between predictor and response

Intercept: The expected value of the response variable when the value of the predictor variable is 0.

Slope: the average increase in Y associated with a one-unit increase in X.

Equation of line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

β_0 : intercept (Y when X = 0) • β_1 : slope (change in Y per unit change in X) • ϵ : error term

- Multiple methods for finding the best fit Line.

Some assumptions for linear regression

- **Linearity:** Relationship between predictor and response is linear
- **Independence:** Observations are independent (independently measured/sampled)
- **No multicollinearity:** Predictors not highly correlated
- **Homoscedasticity:** Constant variance of errors, i.e. errors/uncertainties are evenly spread out over range of inputs
- **Normality:** Errors are normally distributed

Least Squares Method

- Minimize sum of squared residuals
 - **Minimize $\Sigma(y_i - \hat{y}_i)^2$**
- Also called Ordinary Least Squares (OLS)
- Penalizes large errors more heavily
- Produces unique solution: line that minimizes total squared vertical distances from points

Least Squares Method

Equation of line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Let n be a positive integer. For a given data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$,

- we obtain the intercept β_0 and slope β_1 using the least squares method.
- Residual Sum of Squares (RSS), the i th residual $e_i = y_i - \hat{y}_i$

$$\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2$$

Or

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Least Squares Method

RSS can be re-written as:

$$L = \sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

$\hat{y}_i$
↓

Sum of squared distances between (x_i, y_i) and $(x_i, \widehat{\beta}_0 + \widehat{\beta}_1 x_i)$ over $i = 1, \dots, n$

Least Squares Method

$$L = \sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

- We partially differentiate L by β_0 and β_1 and let them be equal to zero, we obtain the following equations:

$$\frac{\partial L}{\partial \widehat{\beta}_0} = -2 \left(\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \right) = 0 \quad \text{Eq(1)}$$

$$\frac{\partial L}{\partial \widehat{\beta}_1} = -2 \left(\sum_{i=1}^n x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \right) = 0 \quad \text{Eq(2)}$$

Where the partial derivative is calculated by differentiating each variable and regarding the other variables as constants. In this case, β_1 and β_0 are regarded as constants when differentiating L by β_0 and β_1 respectively.

Least Squares Method

- By solving Eq (1) and Eq (2) when:

$$\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0 \quad \text{Eq(3)}$$

i.e., $x_1 = x_2 = \dots = x_N$ is not true.

where: $\bar{x}$ is the mean of X and $\bar{y}$ is the mean of Y.

- We can obtain:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{Eq(4)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{Eq(5)}$$

→ Equation of line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Full derivation: <https://www.youtube.com/watch?v=ewnc1cXJmGA>

Linear Models are Estimates

True relationship between X and Y:

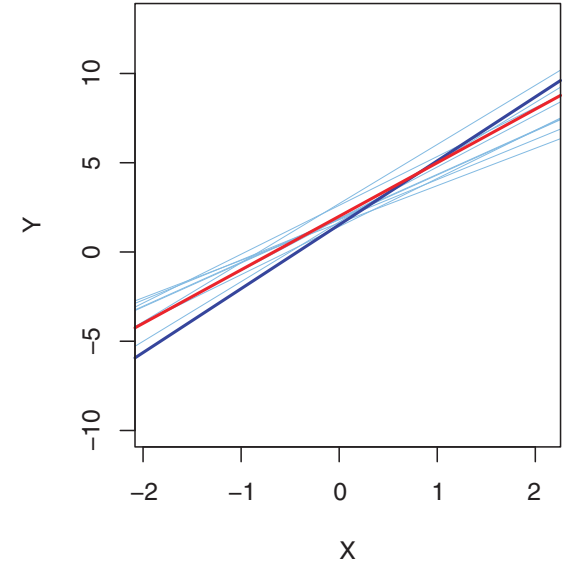
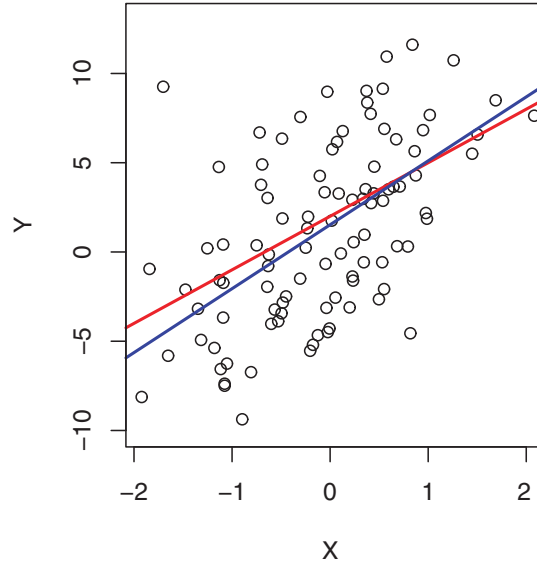
$$Y = 2 + 3X + \epsilon$$

- Where ϵ is a mean-zero random error

Dark Red: true relationship

Dark Blue: least squares regression line

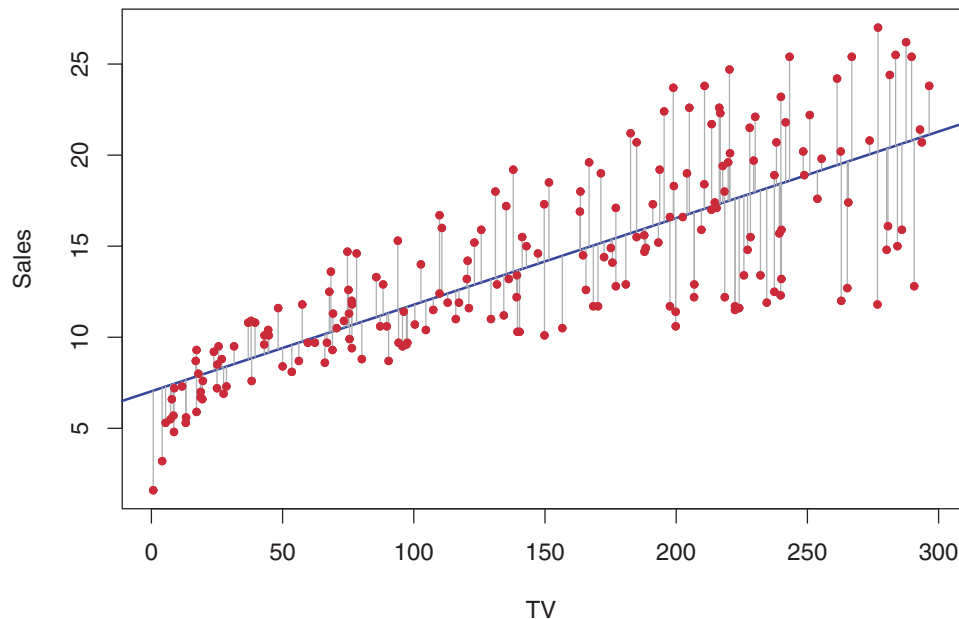
Light Blue: least squares regression lines for multiple new random datasets generated using the same model



*Remember, we are estimating models based on samples to learn about populations...

Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of \$



Evaluating Linear Models

1. Assessing the coefficient estimates

1.1 Values of coefficients >> their Std. errors

$$SE(\hat{\beta}_1) = \frac{RSE}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

1.2. High t-statistic

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

1.3. Low p-value

Hypothesis (more TV ads → more sales)

X ~~H0~~: There is no relationship between X and Y

✓ ~~Ha~~: There is some relationship between X and Y

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$RSE = \sqrt{\frac{1}{n-2} \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Evaluating Linear Models

2. Assessing the model's fit

Residual Standard Error

- Mean sales $\approx 14,000$ units
- $RSE = 3.26 = 3,260$ units

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

R^2

- Measures the how much of the variability in Y can be explained using X (as a predictor)
- has a value between 0,1

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad TSS = \sum_{i=1}^n (y_i - \bar{y}_i)^2$$

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Readings

Select 1 from:

Data on the Web and Linked Data

- Facilitating Discovery of Public Datasets [\[56\]](#)
- Tom Heath Linked Data Tutorial (2009) [\[57\]](#)
- Relational Databases on the Semantic Web, Tim Berners-Lee [\[58\]](#)
- A Survey of Current Approaches for Mapping of Relational Databases to RDF [\[59\]](#)
- On directly mapping relational databases to RDF and OWL [\[60\]](#)

And select 1 from:

Data Mining

- Brief Introduction to Data Mining [\[30\]](#)
- Longer Introduction to Data Mining and slide sets (read the [introduction](#)) [\[31\]](#)

2 total :-)

Thanks!

Work on your data collection!!