ANALYTICS FOR BUSINESS INTELLIGENCE

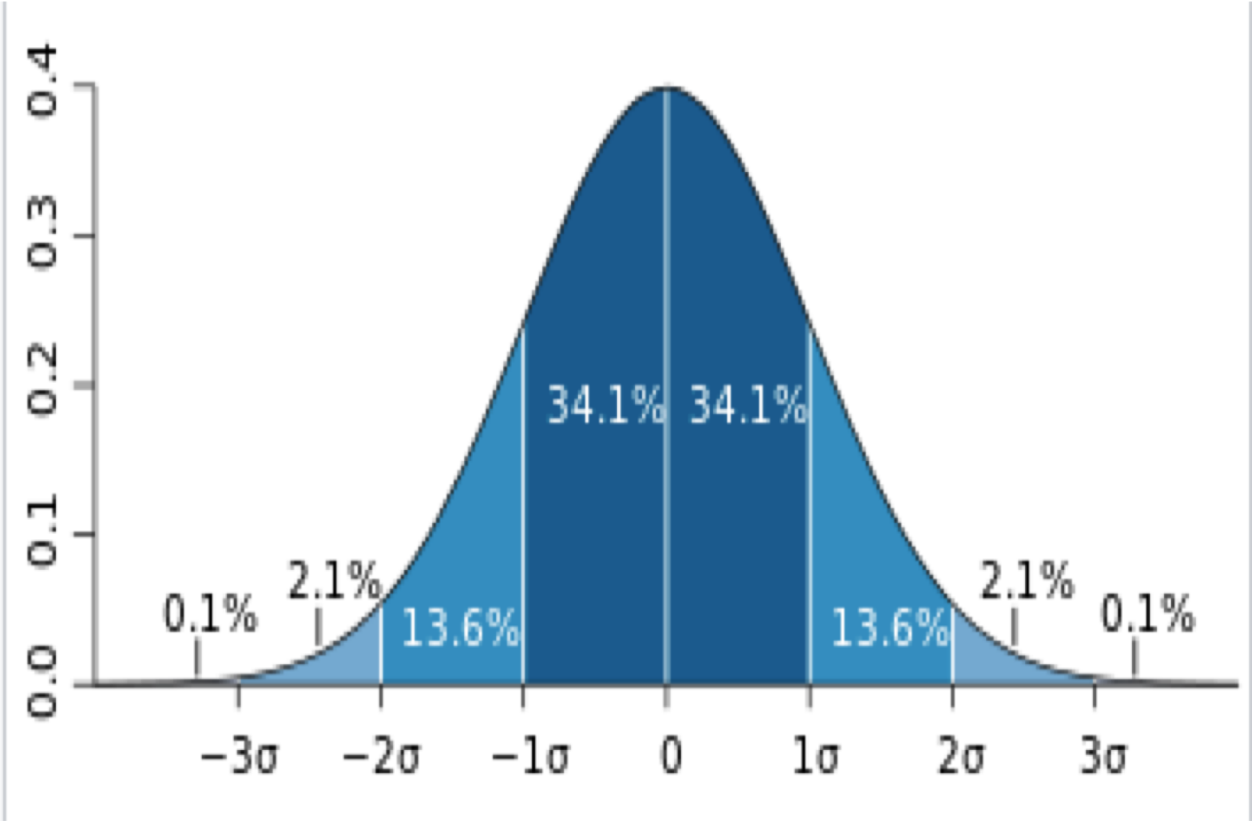
Fall 2021

Nuria Diaz-Tena

To Review:

1. **Normal Distributions**
2. **T Distributions**
3. **Types of comparisons**
   - **One sample**
   - **Two samples**
     - **Equal number of observations and standard error**
     - **Unequal number of observations**
     - **Unequal variances**
4. **Statistical Experiment Review**
5. **Regression Analysis**

Normal Distribution

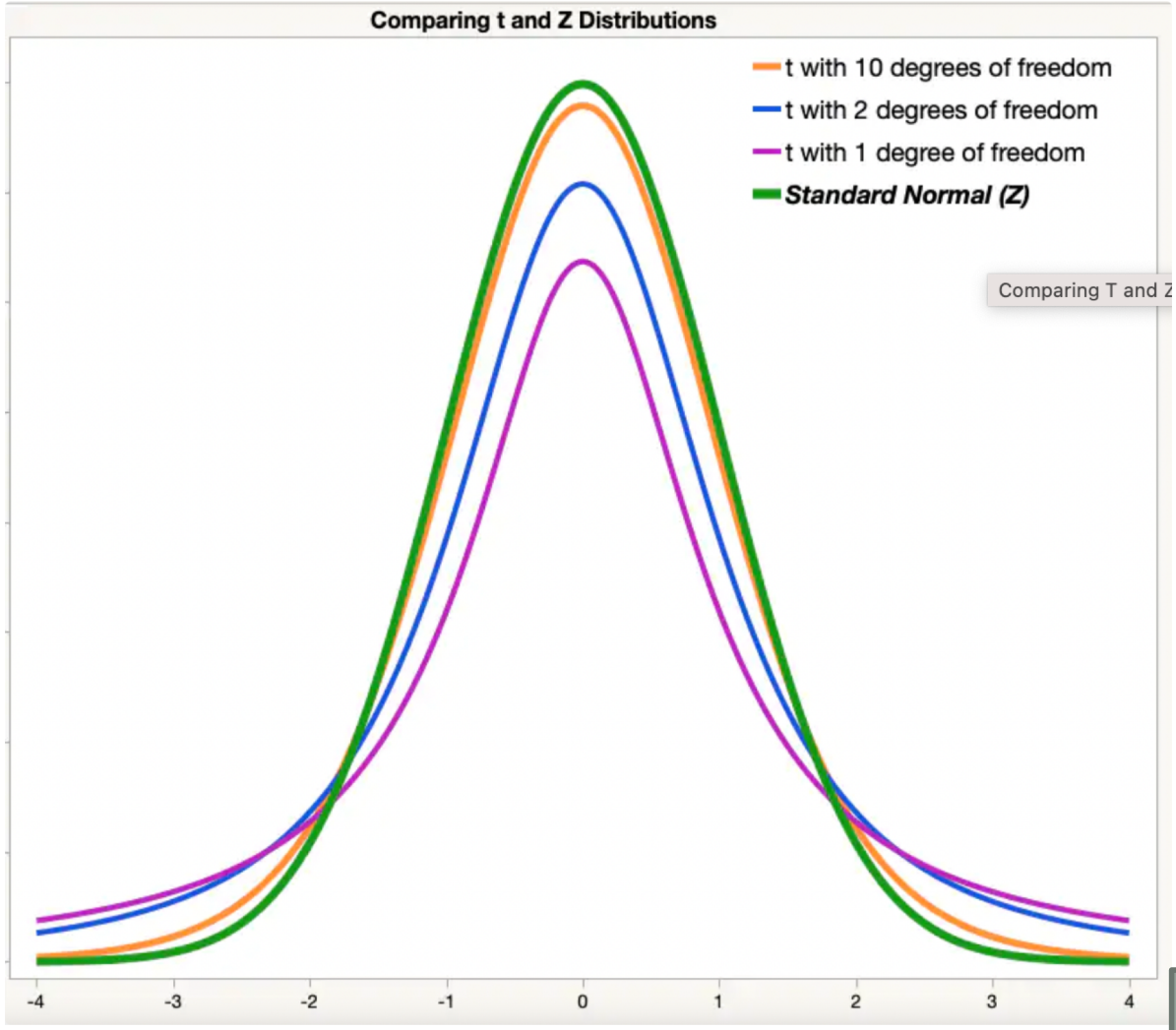


$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter mu is the mean, and the parameter sigma is its standard deviation.

Z and T Distribution



R using ttest
Excel

Sampling Learning Objectives

- Understand differences between **populations** and **samples**
- Understand how **samples** can go wrong
- Understand how **bias** is defined
- Identify potential **sources of bias** in a sample
- Understand how to best obtain **random samples**
- Understand there is **variability in samples** and how it's affected by a sample's attributes
- Estimate **margins of error and confidence intervals**
- Identify **sampling frames** and how they might differ from populations
- Identify **potential sources of bias** that can occur even in well-designed studies and how they can be managed

Errors in Sampling

Sampling errors are errors caused by the act of taking a sample. They cause sample results to be different from the results of a census.

Nonsampling errors are errors not related to the act of selecting a sample from the population. They can be present even in a census.

1. Sampling Errors

A. Random Sampling Error

- Deviation between the statistic and parameter
- Caused by chance in selecting a random sample
- ONLY error accounted for in the margin of error in a confidence statement

B. Bad Sampling Methods

- Convenience and voluntary response samples
- An incomplete sampling frame can cause **undercoverage,** where certain groups of the population are left out.

2. Nonsampling Errors

A. Processing Errors

- Mistakes in mechanical tasks such as arithmetic or data entry

B. Poorly Worded Questions

- Question is slanted to favor one response over the other

C. Response Error

- Response from an individual in the survey that is inaccurate from lying, bad memory, etc.

D. Nonresponse Error

- Failure to obtain data from an individual selected for a sample

Stratified Random Sampling

Step 1: Divide the sampling frame into

distinct groups of individuals, called ***strata***.

- Choose strata because you have an interest in the groups or because the individuals within each group are similar.

Step 2: Take a separate SRS in each stratum and combine these to make up the complete sample.

The Challenge of Internet Surveys

Using the Internet for “Web surveys” is becoming increasingly popular.

Advantages:

- Easy to collect large amounts of data
- Costs less money
- Allows for delivery of multimedia content

Disadvantages:

- Not easy to do well
- Voluntary response, undercoverage, nonresponse

Collecting Data from Open Source conversations

There is already lots of opinions about lots of topics on twitter, blogs, etc.

Can we use this data to gather information?

Course Planning

Lecture	Date	Topic	Assignments	Wg
1	9/01/21	Introduction, Visualizations & Exploratory Data Analysis - Chapter 1		
2	9/15/21	**Int**roduction to R		
3	9/22/21	**Statistical Experiments**	**Homework #1**	5%
4	**9/29/21**	**Regression – Chapter 2 and 3**	**Homework #2**	**5%**
5	10/06/21	Bayesian Models		
6	10/13/21	Models Review	**Project #1**	15%
	10/20/21	Midterm Exam	Midterm	20%
7	10/27/21	Classification Models – Chapter 4		
8	11/03/21	Cross Validation and Advance Classification – Chapter 5 to 9	Homework #3	5%
9	11/10/21	Advance Classification	Homework #4	5%
10	11/17/21	Support Vector Machines - Chapter 9		
		Thanksgiving Recess		
	11/29/21	Review of Machine Learning Models		
11	12/01/21	Clustering - Chapter 10		
12	12/06/21	Project Discussion		
13	12/08/21	Final Exam Preparation	Project #2	15%
	12/22/21	Final Exam	Final	30%

Project:

Project # 1

Life expectancy

1. What is the best model to predict life expectancy with the data provided in the following two links?

- https://www.worldometers.info/demographics/life-expectancy/#countries-ranked-by-life-expectancy
- https://www.worldometers.info/world-population/population-by-country/

2. Which other variables could you include in the model to improve predictions of life expectancy?

Median Value of Own houses in Boston area

- What I the best model to predict the median value of the houses in the Boston area?

What is a Statistical Model?

$Y = f(X) + \varepsilon$ actual

$\hat{Y} = f(X)$ predicted

$E(\hat{Y} - Y)^2 = [f(X) - \hat{f}(X)]^2 + Var(\varepsilon)$

input variables

independent variables

feature variables

Predictors

Drivers

output variable

response

dependent variable

Error Term

Which predictors are associated with the response?

What is the relationship between the response and each predictor?

Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Simple Model – Linear Regression

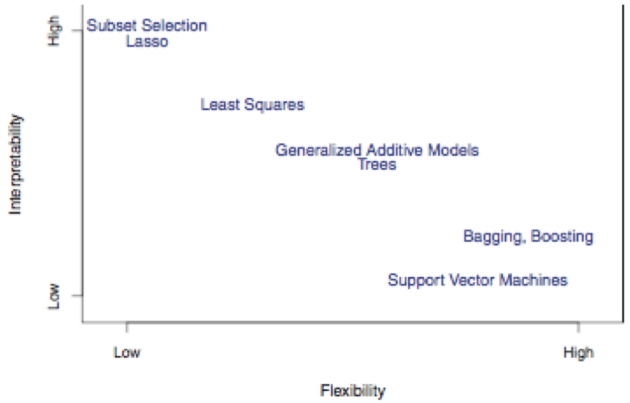
$Y = a + b*(X) + \varepsilon$ actual

Types of Models

Model Form

Parametric Model

$f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$ linear model



Non-Parametric Model

f(X) = any model form of x flexible model

Response Type

Regression Model

Numerical Response

- Linear Regression
- K-Nearest Neighbor
- Boosting

Classification Model

Categorical Response

- Logistic Regression
- K-Nearest Neighbor
- Boosting

Linear Regression

$Y = f(X) + \varepsilon$

actual $Y \approx \beta_0 + \beta_1 X. + \varepsilon$

$\hat{Y} = \hat{f}(X)$

predicted $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$

Estimate Intercept and slope by minimizing the

Residual Sum of Squares (RSS)

All formulas in Chapter 3.

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

H_0 : There is no relationship between X and Y $\quad H_0 : \beta_1 = 0$

H_a : There is some relationship between X and Y $\quad H_a : \beta_1 \neq 0,$

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)},$$

P-value – the probability of observing any number equal to |t| or larger in absolute value, assuming $\beta_1 = 0$.

Error Discussion

Differences and relationships between:

1. RSS –Residual Sum of Squares
2. MSE - Mean Square Error
3. RSE - Residual Standards Error = sqrt(RSS/(n − 2))
4. R-Square = 1- TSS/RSS
5. Adjusted R-Square
6. AIC
7. MAPE – “Mean Absolute Percent Error”
8. Correlation

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

AIC = n log(Var) +2K
K- number of predictors

* TSS : Total Sum Squares (Sum of square difference between y and mean(y))

Model Diagnostics, Significance and Fit

- **Check Residual plot:**
- Are the residuals around 0 with no pattern?
  - If not, there may be a better form instead of linear
- Are the residuals of equal value in all independent values?
  - If not, the residuals have heteroscedasticity instead of homoscedasticity and it is suggested to use logs.
- Outliers.
  - Check for outliers on the residual plot. Do this outliers affect the predictions?
- High-leverage points. Very influential points
  - Check the leverage versus residual plot
- Collinearity
  - Check correlations to know the best predictors for he model
  - Are the p-values less than 5%?

Multiple Linear Regression

$Y = f(X) + \varepsilon$ actual

$Y = B_0 + B_1 * X_1 + B_2 * X_2 + \ldots + B_n * X_n + \text{Error}$

More than One Driver

input variables

independent variables

feature variables

Predictors

output variable

response

dependent variable

Error Term

Reducing error term

Reducible error

Irreducible error

Variance

Which predictors are associated with the response?

What is the relationship between the response and each predictor?

Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

R using for linear regression

Model Accuracy on Testing Set or Hold-out period

- The Model is estimated on the Training Set (80%)
- The Model is evaluated on the Testing Set (20%)or Hold-Out Period (Time Series data – most recent Period)

Regression - Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Classifier – Accuracy

To Do:

Project # 1

Read Chapter 3