DATA ANALYTICS AND MACHINE LEARNING WITH R

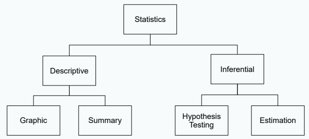
CONFIRMATORY DATA ANALYSIS

LUIS GUSTAVO NARDIN
INTERNET TECHNOLOGY
BRANDENBURG UNIVERSITY OF TECHNOLOGY

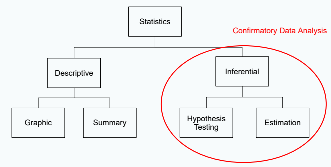
CONFIRMATORY DATA ANALYSIS

Confirmatory Data Analysis refers to an approach which, **subsequent to data acquisition**, proceeds with the **imposition of a prior model** and analysis, estimation, and testing model parameters using traditional statistical tools such as **significance, inference, and confidence**.

CONFIRMATORY DATA ANALYSIS



CONFIRMATORY DATA ANALYSIS

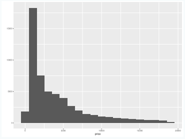


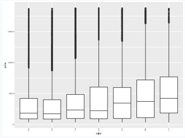
STATISTICAL INFERENCE

- Branch of statistics that allows to arrive at conclusions about a population through a sample of the population
- Measure the **effect** that some **input parameters** of the process generating the population have on features, or **output metrics**, of the process

EXPLORATORY
DATA ANALYSIS

GRAPHIC REPRESENTATION





```
> qplot(price, data=diamon
               geom="histogra
```

```
> qplot(color, price, data
               geom="boxplot"
```

SUMMARY STATISTICS

	carat	depth	table	price
Mean	0.79	61.75	57.46	3,933.00
Median	0.70	61.80	57.00	2,401.00
Standard Deviation	0.47	1.43	2.23	3,989.44
Minimum	0.2	43	43	326.00
Maximum	5.01	79	95	18,823.00

CONFIRMATORY DATA ANALYSIS

- Hypothesis Testing
- Regression
- Analysis of Variance

HYPOTHESIS TESTING

HYPOTHESIS TESTING

- Hypothesis testing is intended to confirm or validate some conjectures about the dataset under analysis
- These conjectures, or hypotheses, are related to the parameters of the probability distribution of the data

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

where H_0 is the *null hypothesis* and H_1 is the *Alternative hypothesis*

HYPOTHESIS TESTING

- Hypothesis testing tries to find evidence about the refutability of the null hypothesis using probability theory
- The null hypothesis is rejected if the data do not support it with "enough evidence," which is expressed in terms of significance level α
- 5% significance level ($\alpha = 0.05$) is a widely accepted value in most cases

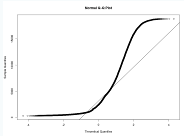
HYPOTHESIS TESTING

Test	Description
`shapiro.test`	Normality test
`var.test`	Compare two variances
`cor.test`	Correlation between two samples
`t.test`	Compare the means with normal errors
`wilcox.test`	Compare the means with non-normal errors
`prop.test`	Compare two proportions
`chisq.test`	Goodness-of-fit tests
`poisson.test`	Poisson distribution test
`binom.test`	Binomial distribution test

NORMALITY TEST

The **Shapiro-Wilk test** (`shapiro.test`) checks if a random sample comes from a normal distribution.

p-value lower than a threshold (e.g., 0.05) rejects the null hypothesis indicating that the values come from a normal distribution.




```
> qqnorm(diamonds$price)
> qqline(diamonds$price)
```

```
> attach(diamonds)
> shapiro.test(price[

      Shapiro-Wilk no

data:  price[sample(5
W = 0.65758, p-value

> shapiro.test(dE$pri
```

VARIANCE TEST

The **Fisher's F test** (`var.test`) compares the variances of two samples and checks whether they are significantly different.

```
> var(dE$price)
[1] 11183397
> var(dJ$price)
[1] 19697506
> var.test(dE$price, dJ$price)

 F test to compare two variances

data:  dE$price and dJ$price
```

CORRELATION TEST

The correlation test (`cor.test`) determines the significance of the correlation between the samples of two variables.

- Samples with **normal error** should use the Pearson's product moment correlation (`method="p"`)
- Samples with **non-normal error** should use the Pearson's product moment correlation (`method="k"` or `method="s"`)

Normal Error

```
> cor.test(diamonds$p

  Pearson's product-m

data:  diamonds$price
t = 551.41, df = 5393
alternative hypothesi

95 percent confidence
```

Non-normal Error

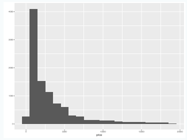
```
> cor.test(diamonds$p
  method="k")

  Kendall's rank corr

data:  diamonds$price
z = 288.02, p-value <
alternative hypothesi
```

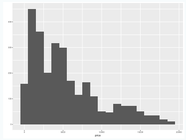
MEANS TEST

For example, are the mean price of the diamonds of color E and J different?

```
> dE <- diamonds[which(dia
> mean(dE$price)
[1] 3076.752
> qplot(price, data=dE, ge
    binwidth=1000)
```

```
> dJ <- diamonds[which(dia
> mean(dJ$price)
[1] 5323.818
> qplot(price, data=dJ, ge
    binwidth=1000)
```





MEANS TESTS

Usually it is necessary to perform some initial checking to identify whether the data complies with the assumptions of the statistical analysis to be performed.

For example, non-normality, outliers and serial correlation may invalidate inferences made by standard parametric tests.

MEANS TEST

Normal Error

`t.test`

```
> t.test(dE$price, dJ

  Welch Two Sample t-

data:  dE$price and d
t = -24.881, df = 376
alternative hypothesi

95 percent confidence
```

Non-normal Error

`wilcox.test`

```
> wilcox.test(dE$price, dJ

  Wilcoxon rank sum test w
    continuity correction

data:  dE$price and dJ$pri
W = 9232700, p-value < 2.2
alternative hypothesis: tr
                        no
```

MEANS TEST

A means hypothesis test can be used to verify if the mean price of diamonds of color E is greater than or less than the mean price of diamonds of color J.

```
> wilcox.test(dE$pric
  alternative = "grea

  Wilcoxon rank sum t
    continuity correc

data:  dE$price and d
W = 9232700, p-value
alternative hypothesi
```

```
> wilcox.test(dE$pric
  alternative = "less

  Wilcoxon rank sum t
    continuity correc

data:  dE$price and d
W = 9232700, p-value
alternative hypothesi
```

$H_0 : \mu \leq \mu_0$
$H_1 : \mu > \mu_0$

$H_0 : \mu \geq \mu_0$
$H_1 : \mu < \mu_0$

REGRESSION

REGRESSION

- Regression analysis is a set of statistical processes for estimating the relationship among two kinds of variables:
  - **Dependent** variables (or **responses**)
  - **Independent** variables (or **predictors**)

REGRESSION

The steps to perform a regression analysis

1. Define the relationship of interest
2. Collect data containing values for the dependent and independent variables
3. Build a regression model
4. Evaluate the regression model
5. Use the model to predict

1. INTEREST

“Interested in accurately predicting the price of diamonds based on one or more of their characteristics”

2. COLLECT DATA

- `diamonds` data set provides ~ 54000 diamonds entries from **http://www.diamondse.info/**
- Structure of the data frame
  - `help(diamonds)`
  - `str(diamonds)`
- 10 variables: `price`, `carat`, `cut`, `color`, `clarity`, `x`, `y`, `z`, `depth`, and `table`

3. BUILD MODEL

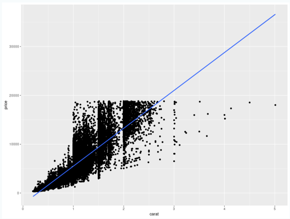
VARIABLES CORRELATION

Before building a model it is interesting to evaluate the correlation that exists between the dependent variable (e.g., price) with the individual independent variables (e.g., carat, depth, table, x, y, z) using the `cor` function.

Independent Variable	Correlation
`carat`	**0.92**
`depth`	-0.01
`table`	0.12
`x`	**0.88**
`y`	**0.86**
`z`	**0.86**

3. BUILD MODEL

GRAPHICAL ANALYSIS



3. BUILD A MODEL

LINEAR REGRESSION

- Linear regression is used to predict the value of a dependent variable Y based on the input independent variables X
- The generalized form of a mathematical equation representing a linear regression is

  $Y = B_1 + B_2 X + \varepsilon$

  where, B_1 is the intercept, B_2 is the slope, and ε is the error term (i.e., the part of Y the regression model is unable to explain) assumed to follow a normal distribution with a mean of zero and a standard deviation of σ .

3. BUILD MODEL

BUILD LINEAR MODEL

- The function used for building linear models is **`lm`** function
- The `lm` function takes in two main arguments
  1. Formula
  2. Data
- The data is typically a `data.frame` and the formula is commonly written out directly as the example below

```
lm( Y ~ X1, data=dataset )
lm( Y ~ X1 + X2, data=dataset )
lm( Y ~ X1 + X2 * X3, data=dataset )
```

where, + relate the main factors and * the interactions.

3. BUILD MODEL

```
> lm1 <- lm( price ~ carat, data=diamonds )
> print(lm1)

Call:
lm(formula = price ~ carat, data = diamonds)

Coefficients:
(Intercept)        carat
      -2256         7756
```

price = -2256 + 7756 * carat

4. EVALUATE MODEL

```
> summary(lm1)

Call:
lm(formula = price ~ carat, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max 
-18585.3   -804.8    -18.9    537.4  12731.7 
```

4. EVALUATE MODEL

CHECKING FOR STATISTICAL SIGNIFICANCE

p-value

- Check the p-value for the model (bottom right) and for the individual independent variables (right column under Coefficients
- * indicates the statistical significance level
- The model and independent variables are statistically significant only when they are less than the statistical significance level (e.g., $\alpha = 0.05$)

4. EVALUATE MODEL

CHECKING FOR STATISTICAL SIGNIFICANCE

t-value

- A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance
- Pr(>|t|) or p-value is the probability that t-value as high or higher than the observed value when the Null Hypothesis (the B coefficient is equal to zero or that there is no relationship) is true
  - If the Pr(> |t|) is low, the coefficients are significant (significantly different from zero).
  - If the Pr(>|t|) is high, the coefficients are not significant

4. EVALUATE MODEL

CHECKING FOR STATISTICAL SIGNIFICANCE

t-Statistic = β − coefficient Std.Error

```
> modelSummary <- summary(lm1)
> modelCoeffs <- modelSummary$coefficients
> beta.estimate <- modelCoeffs["carat", "Estima
> std.error <- modelCoeffs["carat", "Std. Error
> t_value <- beta.estimate/std.error
> p_value <- 2*pt(-abs(t_value), df=nrow(diamon

> t_value
[1] 551.4081
```

4. EVALUATE MODEL

CHECKING FOR STATISTICAL SIGNIFICANCE

It is absolutely important for the model to be statistically significant before proceed and use it to predict (or estimate) the dependent variable, otherwise, the confidence in predicted values from that model reduces and may be construed as an event of chance.

4. EVALUATE MODEL

R-SQUARED

R-Squared explains the proportion of variation in the dependent (response) variable that has been explained by this model.

$$R^2 = 1 - \frac{SSE}{SST}$$

where, SSE is the sum of squared errors given by $SSE = \sum_i^n (y_i - \hat{y_i})^2$ and $SST = \sum_i^n (y_i - \bar{y_i})^2$ is the sum of squared total. Here, $\hat{y_i}$ is the fitted value for observation *i* and $\bar{y}$ is the mean of Y.

4. EVALUATE MODEL

ADJUST R-SQUARED

- As you add more X variables to your model, the R-Squared value of the new bigger model will always be greater than that of the smaller subset.
- Whatever new variable we add can only add to the variation that was already explained.
- Adj R-Squared penalizes total value for the number of terms (i.e., predictors) in your model.

$$R_{adj}^2 = 1 - \frac{MSE}{MST}$$

where, MSE is the mean squared error given by $MSE = \frac{SSE}{(n-q)}$ and $MST = \frac{SST}{(n-1)}$ is the mean squared total, where *n* is the number of observations and *q* is the number of coefficients in the model.

4. EVALUATE MODEL

STANDARD ERROR AND F-STATISTIC

Both standard errors and F-statistic are measures of goodness of fit.

$$\text{Std.Error} = \text{MSE} = \frac{\text{SSE}}{n - q}$$

$$\text{F-statistic} = \frac{\text{MSR}}{\text{MSE}}$$

where, *n* is the number of observations, *q* is the number of coefficients and MSR is the mean square regression, calculated as,

$$\text{MSR} = \frac{\sum_{i}^{n} (y_i - \bar{\hat{y}})}{q - 1} = \frac{\text{SST} - \text{SSE}}{q - 1}$$

4. EVALUATE MODEL

AIC AND BIC

The Akaike's information criterion - AIC (Akaike, 1974) and the Bayesian information criterion - BIC (Schwarz, 1978) are measures of the goodness of fit of an estimated statistical model and can also be used for model selection.

```
> AIC(lm1)
[1] 945466.5
> BIC(lm1)
[1] 945493.2
```

When comparing multiple models, the model with the **lowest AIC and BIC score is preferred**.

5. PREDICT

Use the `predict` function to predict new values using the model build.

```
> newData <- data.frame( carat=0.91 )
> predict( lm1, newData)
       1
4801.987
```

EXERCISE

- Build different models to predict the diamonds price and determine the best model among those.

ANALYSIS OF VARIANCE

- Analysis of Variance (ANOVA) test differences between two or more group means
- ANOVA test is centered around the different sources of variation (variation between and within group) in a typical variable
- A primarily ANOVA test provides evidence of the existence of the mean equality between the group
- This statistical method is an extension of the *t-test*
- It is used in a situation where the factor variable has more than one group.

ANALYSIS OF VARIANCE

`Factor` refers to a **categorical quantity** under examination in an experiment as a possible cause of variation in the **response variable**.

`Levels` refer to the **categories, measurements, or strata of a factor** of interest in the experiment.

ASSUMPTIONS

- The observations are obtained **independently and randomly** from the population defined by the factor levels
- The data of each factor level are **normally distributed**
- These normal populations have a **common variance** (*Levene's* test can be used to check this.)

DIAMONDS

```
> diamonds
# A tibble: 53,940 x 10
   carat cut       color clarity depth table pr
 1 0.23  Ideal     E     SI2      61.5    55
 2 0.21  Premium   E     SI1      59.8    61
 3 0.23  Good      E     VS1      56.9    65
 4 0.290 Premium   I     VS2      62.4    58
 5 0.31  Good      J     SI2      63.3    58
 6 0.24  Very Good J     VVS2     62.8    57
```

ANALYSIS OF VARIANCE

A analysis of variance is a technique that **partitions the total sum of squares of deviations** of the observations about their mean into portions associated with independent variables in the experiment and a **portion associated with error**

ANOVA

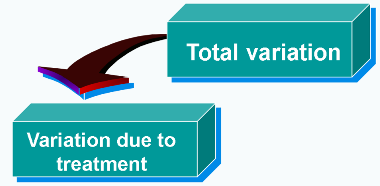
- One-Way ANOVA helps us understand the relationship between one continuous dependent variable and one categorical independent variable
- When we have two independent categorical variable we need to use Two-Way ANOVA.
- When we have more than two categorical independent variables we need to use N way ANOVA.

ONE-WAY ANOVA

PARTITION TOTAL VARIATION

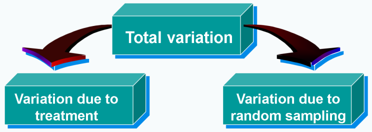
Total variation

ONE-WAY ANOVA
PARTITION TOTAL VARIATION

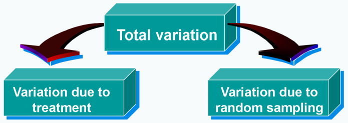


ONE-WAY ANOVA

PARTITION TOTAL VARIATION



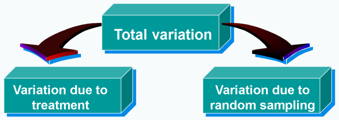
ONE-WAY ANOVA
PARTITION TOTAL VARIATION



Sum of Squares Among
Sum of Squares Between
Sum of Squares Treatment
Among Groups Variation

ONE-WAY ANOVA

PARTITION TOTAL VARIATION

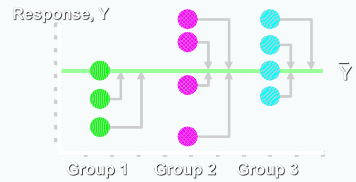


Sum of Squares Among
Sum of Squares Between
Sum of Squares Treatment (SST)
Among Groups Variation

Sum of Squares Within
Sum of Squares Error (SSE)
Within Groups Variation

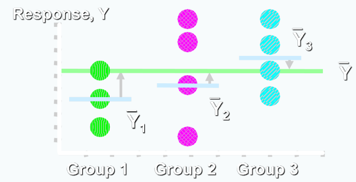
ONE-WAY ANOVA
TOTAL VARIATION

SS Total = Y 11 - Y - 2 + Y 21 - Y - 2 + ... + Y ij - Y - 2



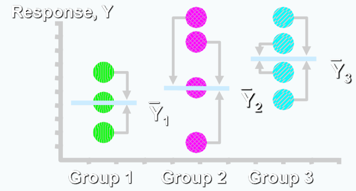
ONE-WAY ANOVA
TREATMENT VARIATION

$$SST = n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2 + \ldots + n_p(\bar{Y}_p - \bar{Y})^2$$



ONE-WAY ANOVA ERROR VARIATION

SSE = Y 11 - Y - 1 2 + Y 21 - Y - 2 2 + ... + Y pj - Y - p 2



ONE-WAY ANOVA TEST STATISTIC

- Test Statistic
  - F = MST MSE = SST / p - 1 SSE / n - p
  - MST is Mean Square for Treatment
  - MSE is Mean Square for Error
- Degree of Freedom
  - $v_1 = p - 1$
  - $v_2 = n - p$
  - p = Number of Groups or levels
  - n = Total sample size

ONE-WAY ANOVA SUMMARY TABLE

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square (Variance)	F
Treatment	p - 1	SST	MST = SST/(p-1)	MST/MSE
Error	n - p	SSE	MSE=SSE/(n-p)	
Total	n - 1	SS(Total)=SST+SSE		

ONE-WAY ANOVA IN R

- Our objective is to test the following assumption:
  - H0: There is no difference in price average of diamonds cut between group
  - H5: The price average is different for at least one diamonds cut group

ONE-WAY ANOVA IN R

- Test for Normality `shapiro.test`

```
> shapiro.test(diamonds[diamonds$cut == "Ideal",]$price[1:50

  Shapiro-Wilk normality test

data:  diamonds[diamonds$cut == "Ideal", ]$price[1:5000]
W = 0.91671, p-value < 2.2e-16
```

- Test for common Variance `levene.test` (package `lawstat`)

```
> levene.test(diamonds$price, diamonds$cut)

  modified robust Brown-Forsythe Levene-type test based
  on the absolute deviations from the median

data:  diamonds$price
Test Statistic = 123.6, p-value < 2.2e-16
```

ONE-WAY ANOVA IN R

- Levels of factor

```
> levels(diamonds$cut)
[1] "Fair"      "Good"      "Very Good" "Premium"   "Ideal"
```

- Run the `aov`

```
> fit <- aov(price ~ cut, data=diamonds)
```

- Analysis of the results `summary.aov`

```
> summary.aov(fit)
               Df    Sum Sq   Mean Sq F value Pr(>F)
cut             4 1.104e+10 2.760e+09   175.7 <2e-16 ***
Residuals   53935 8.474e+11 1.571e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value is lower than the usual threshold of 0.05. You are confident to say there is a statistical difference between the groups, indicated by the "***".

SUMMARY TABLE

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square (Variance)	F
Treatment	5 - 1 = 4	1.104e+10	2.760e+09	175.7
Error	53940 - 5 = 53935	8.474e+11	1.571e+07	
Total	53940 - 1 = 53939	8.5844e+11		

PAIRWISE COMPARISON

- The one-way ANOVA does not inform which group has a different mean, fir such perform the *Tukey HSD* test
- The Tukey HSD ("honestly significant difference" or "honest significant difference") test is a statistical tool used to determine if the **relationship between two sets of data is statistically significant**
- The Tukey HSD test is invoked when you need to determine if the **interaction among three or more variables** is mutually statistically significant, which unfortunately is not simply a sum or product of the individual levels of significance.

PAIRWISE COMPARISON IN R

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price ~ cut, data = diamonds

$cut
                      diff        lwr
Good-Fair       -429.89331 -740.44880  -119
```

TWO-WAY ANOVA

- A two-way ANOVA test adds another group variable to the formula. It is identical to the one-way ANOVA test, but the formula changes to
  $y = x1 + x2$
- Hypothesis
  - H0: The means are equal for both variables (i.e., factor variable)
  - H3: The means are different for both variables

TWO-WAY ANOVA IN R

```
> fit <- aov(price ~ cut + color, data=diamonds)
> summary.aov(fit)
               Df    Sum Sq   Mean Sq F value Pr(>F)
cut             4 1.104e+10 2.760e+09   181.1 <2e-16
color           6 2.551e+10 4.251e+09   278.9 <2e-16
Residuals   53929 8.219e+11 1.524e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

You can conclude that both cut and color are statistically different from 0. You can reject the NULL hypothesis and confirm that changing the cut or the color impact the price.

EXERCISE
SPRUCE MOTH TRAP

- Data
  - Response: number of spruce moths found in trap after 48 hours
  - Factor 1: Location of trap in tree (top branches, middle branches, lower branches, ground)
  - Factor 2: Type of lure in trap (scent, sugar, chemical)

EXERCISE

SPRUCE MOTH TRAP

- Activities
  - Load the data
  - Create summary statistics for location
  - Create summary statistics for type of lure
  - Create boxplots for each category
  - Check for normality
  - Check for equality of variance
  - Perform ANOVA
  - Evaluate the contribution of each variable