DATA ANALYTICS AND MACHINE LEARNING WITH R

CONFIRMATORY DATA ANALYSIS

LUIS GUSTAVO NARDIN

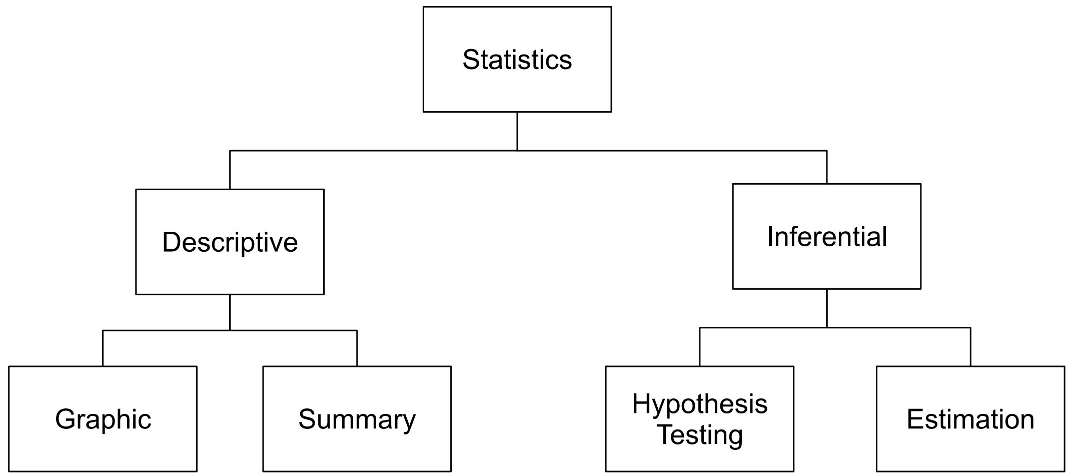
INTERNET TECHNOLOGY

BRANDENBURG UNIVERSITY OF TECHNOLOGY

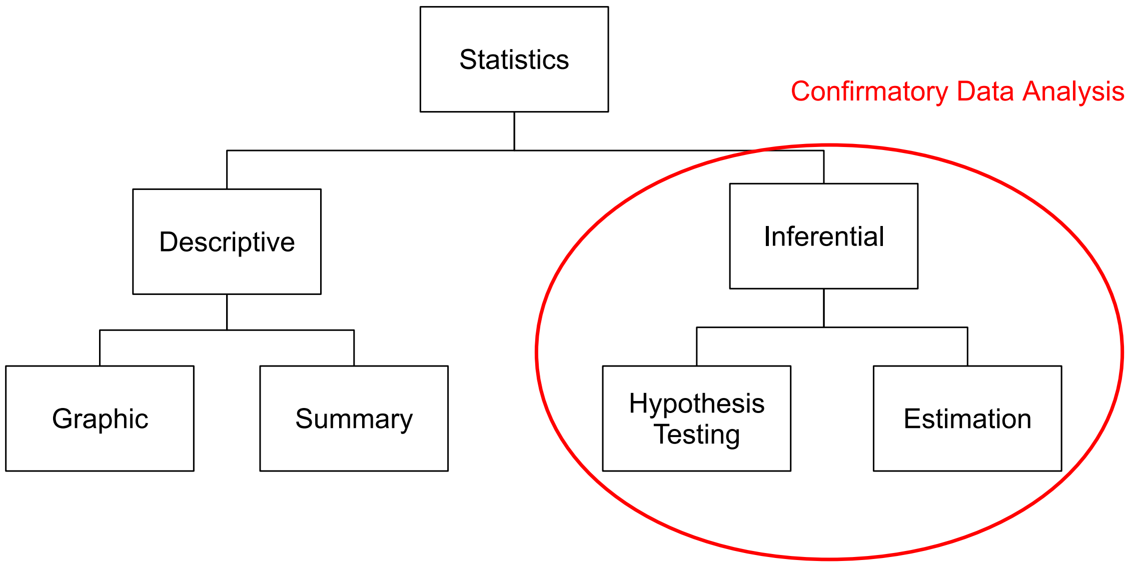
CONFIRMATORY DATA ANALYSIS

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

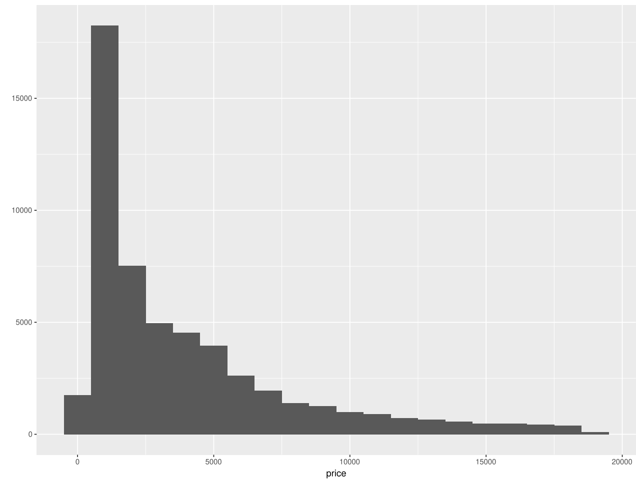


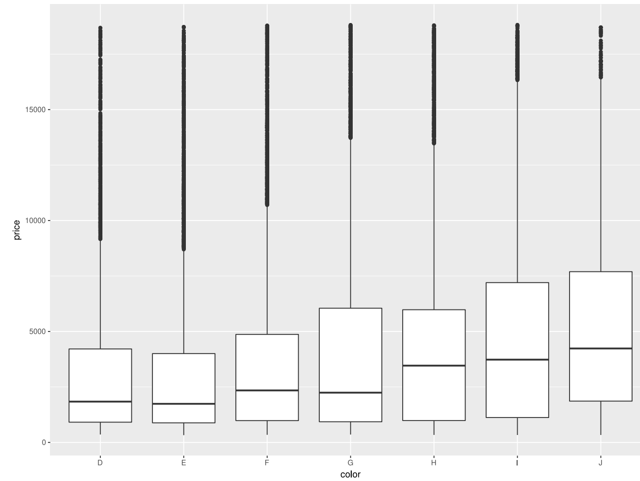
STATISTICAL INFERENCE

- 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=diamonds,
             geom="histogram", binwidth=1000)
```

```
> qplot(color, price, data=diamonds,
             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

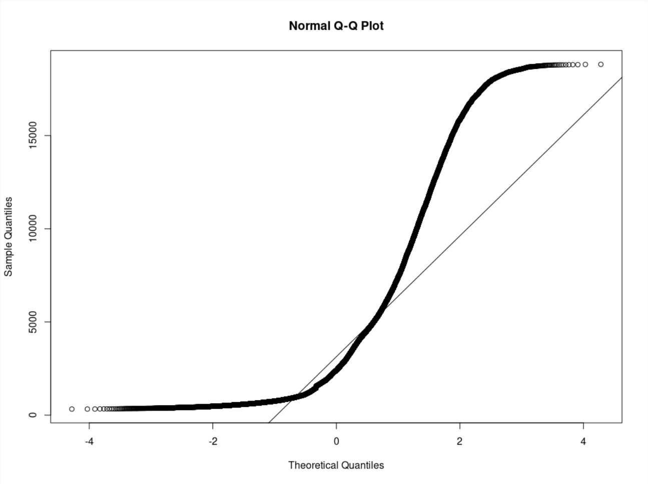
HYPOTHESIS TESTING

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[sample(5000, price)])

      Shapiro-Wilk normality test

data:  price[sample(5000, price)]
W = 0.65758, p-value < 2.2e-16

> shapiro.test(dE$price[sample(5000, dE$price)])

Shapiro-Wilk normality test

data: dE$price[sample(5000, dE$price)]
W = 0.88792, p-value = 9.927e-15

> shapiro.test(dJ$price[sample(5000, dJ$price)])

Shapiro-Wilk normality test

data: dJ$price[sample(5000, dJ$price)]
W = 0.88092, p-value = 2.278e-11
```

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
F = 0.56776, num df = 9796, denom df = 2807, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5347761 0.6021755
sample estimates:
ratio of variances
          0.567757
```

CORRELATION TEST

The correlation test (`cor.test`) determine 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$price, diamonds$carat)

  Pearson's product-moment correlation

data:  diamonds$price and diamonds$carat
t = 551.41, df = 53938, p-value < 2.2e-16
alternative hypothesis: true correlation is
                        not equal to 0
95 percent confidence interval:
 0.9203098 0.9228530
sample estimates:
      cor
0.9215913
```

Non-normal Error

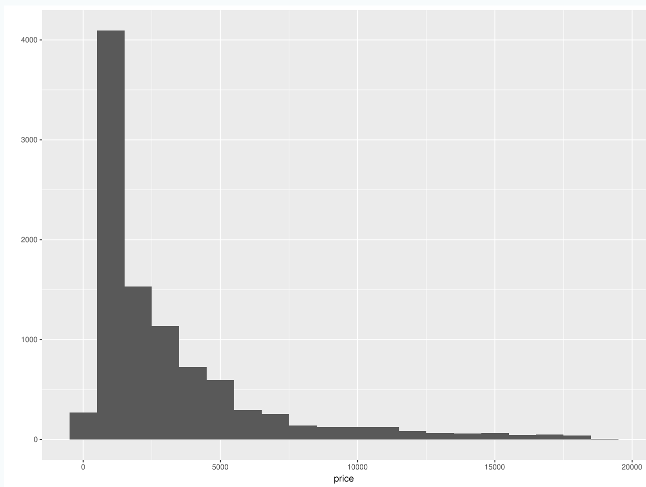
```
> cor.test(diamonds$price, diamonds$carat,
  method="k")

  Kendall's rank correlation tau

data:  diamonds$price and diamonds$carat
z = 288.02, p-value < 2.2e-16
alternative hypothesis: true tau is
                        not equal to 0
sample estimates:
      tau
0.8341049
```

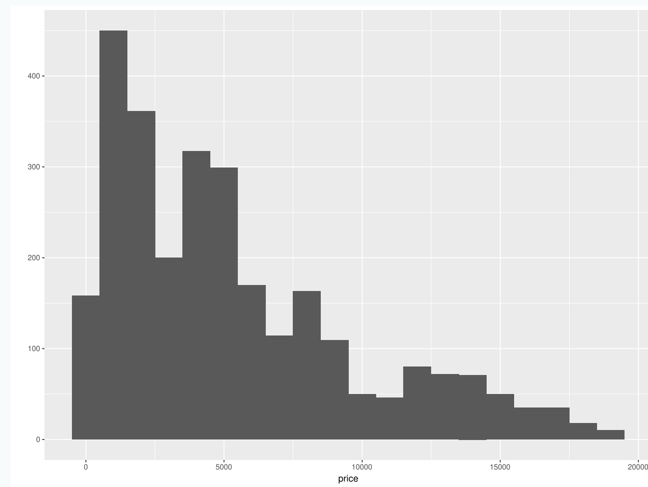
MEANS TEST

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

```
> dE <- diamonds[which(diamonds$color==E),]
> mean(dE$price)
[1] 3076.752
> qplot(price, data=dE, geom="histogram",
    binwidth=1000)
```

```
> dJ <- diamonds[which(diamonds$color==E),]
> mean(dJ$price)
[1] 5323.818
> qplot(price, data=dJ, geom="histogram",
    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$price, dJ$price)

  Welch Two Sample t-test

data:  dE$price and dJ$price
t = -24.881, df = 3766.3, p-value < 2.2e-16
alternative hypothesis: true difference in
                        means is not equal to 0
95 percent confidence interval:
 -2424.131 -2070.000
sample estimates:
mean of x mean of y
 3076.752  5323.818
```

Non-normal Error

`wilcox.test`

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

  Wilcoxon rank sum test with
    continuity correction

data:  dE$price and dJ$price
W = 9232700, p-value < 2.2e-16
alternative hypothesis: true location shift is
                        not equal to 0
```

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$price, dJ$price,
  alternative = "greater")

  Wilcoxon rank sum test with
    continuity correction

data:  dE$price and dJ$price
W = 9232700, p-value = 1
alternative hypothesis: true location shift
                        is greater than 0
```

```
> wilcox.test(dE$price, dJ$price,
  alternative = "less")

  Wilcoxon rank sum test with
    continuity correction

data:  dE$price and dJ$price
W = 9232700, p-value < 2.2e-16
alternative hypothesis: true location shift
                        is less than 0
```

$H0 : \mu \leq \mu 0$

$H1 : \mu > \mu 0$

$H0 : \mu \geq \mu 0$

$H1 : \mu < \mu 0$

REGRESSION

ANALYSIS OF VARIANCE