Introduction to statistical testing

Day 5

Sophie Robert-Hayek

University of Lorraine

Py4SHS 2023

Introduction

Outline

Bibliography

- Haslwanter, T. (2016). *An Introduction to Statistics with Python: With Applications in the Life Sciences*. Germany: Springer International Publishing.
- Lehmann, E., Romano, J.P.(2022). *Testing Statistical Hypotheses*. Switzerland: Springer International Publishing.

Reminder on previous session

In previous sessions, we studied **Machine Learning** based approaches ...

Reminder on previous session

In previous sessions, we studied **Machine Learning** based approaches ...with the goal of building models able to **predict** given unseen data.

But sometimes we simply want to analyze a dataset with the goal of **understanding a phenomenon**.

Introduction

Given a dataset, we want to be able to be able to **answer** specific questions.

Introduction

Given a dataset, we want to be able to be able to **answer** specific questions.

For example:

- Are two texts written by the same author ?

Introduction

Given a dataset, we want to be able to be able to **answer** specific questions.

For example:

- Are two texts written by the same author ?
- Is there a gender based wage gap ?

Introduction

Given a dataset, we want to be able to be able to **answer** specific questions.

For example:

- Are two texts written by the same author ?
- Is there a gender based wage gap ?
- Is a medical treatment efficient for treating high cholesterol ?

Reminder on previous session

Question

In your opinion, what tools do we need to answer these questions ?

Reminder on previous session

Question

In your opinion, what tools do we need to answer these questions ?

We need:

- A dataset (called sample);
- A **metric** to measure the phenomenon;
- A way to compare **how these metrics differ**.

Introduction

For example, when analyzing the efficiency of a medication, we need:

- Two samples: one taking the medication, one not taking any (or a placebo);
- A metric we want to compare: average fasting insulin;
- At the simplest level, compare the two means, and if different, then the medication work !

...but could we systematize this approach for more objective ?
How can we tell if the different in mean is significant and not due to random flucutation in the data ?

Introduction

Statistical **tests** enable us to decide whether a dataset **sufficiently support a particular hypothesis**.

Introduction

Statistical **tests** enable us to decide whether a dataset **sufficiently support a particular hypothesis**.

Possible hypothesis:

- The mean cholesterol differs between placebo and non-placebo patients;

Introduction

Statistical **tests** enable us to decide whether a dataset **sufficiently support a particular hypothesis**.

Possible hypothesis:

- The mean cholesterol differs between placebo and non-placebo patients;
- The median salary differs between men and women;

Introduction

Statistical **tests** enable us to decide whether a dataset **sufficiently support a particular hypothesis**.

Possible hypothesis:

- The mean cholesterol differs between placebo and non-placebo patients;
- The median salary differs between men and women;
- The mean use of the word *and* differs between two texts;

Introduction

Statistical tests consist in:

- **testing**;
- an **hypothesis**;
- for a **selected population sample**.

Quantitative testing framework and vocabulary

Outline

General framework

Statistical **testing** gives us a framework to:

- Design an **hypothesis** that needs to be tested;

General framework

Statistical **testing** gives us a framework to:

- Design an **hypothesis** that needs to be tested;
- Compute a **test statistic** able to discriminate for or against this hypothesis;

General framework

Statistical **testing** gives us a framework to:

- Design an **hypothesis** that needs to be tested;
- Compute a **test statistic** able to discriminate for or against this hypothesis;
- Compare this test statistic to its expected value if the hypothesis is null;

General framework

Statistical **testing** gives us a framework to:

- Design an **hypothesis** that needs to be tested;
- Compute a **test statistic** able to discriminate for or against this hypothesis;
- Compare this test statistic to its expected value if the hypothesis is null;
- Conclude regarding the likelihood of the hypothesis being true: we **reject or not** the hypothesis.

General framework

We thus need to define:

- An **hypothesis**;
- A **test statistic**;
- **Significance levels**;

Hypothesis

Hypothesis

An hypothesis is a **statement about the parameters describing a population**, that we are not sure if it is true or not.

The goal of statistical testing is to:

- **Reject** the hypothesis is there is enough evidence against it;
- **Don't reject** the hypothesis **if there is not enough evidence against it**.

Hypothesis

Question

In your opinion, what is one of the biggest methodological limit of this affirmation: ***Don't reject if there is not enough evidence against it*** ?

Not rejecting the hypothesis does not mean that the hypothesis is true ... !

all hypothesis are innocent until proven guilty :-)

Formulating Hypotheses

For statistical testing, we need to design **two mutually exclusive hypotheses**:

- **Null Hypothesis** (H_0): A statement of no effect or no difference.
- **Alternative Hypothesis** (H_1): A statement indicating the presence of an effect or difference, **often corresponding to the research question**.

Formulating Hypotheses

Returning to our medical example of cholesterol drugs:

- H_0 (null hypothesis): the patients with and without the medication have the same average cholesterol level;
- H_1 (alternative hypothesis): the patients with and without the medication have a different cholesterol level.

Formulating hypotheses

With mathematical notation, if:

- $\mathcal{S}_0$ the population with the placebo;
- $\mathcal{S}_1$ the population with the medication;

Formulating hypotheses

With mathematical notation, if:

- $\mathcal{S}_0$ the population with the placebo;
- $\mathcal{S}_1$ the population with the medication;
- μ_0 is the average cholesterol in $\mathcal{S}_0$;
- μ_1 is the average cholesterol in $\mathcal{S}_1$

- H_0: μ_0 and μ_1 do not differ;
- H_1: μ_0 and μ_1 differ.

Formulating hypothesis

Question

Given the following research question, *Does the use of the word* $\epsilon\nu$ *vary between text 1 and text 2* ?
Can you:

- Define the population $\mathcal{S}_1$ and $\mathcal{S}_0$?
- Define the statistics μ_1 and μ_0 to compute ?
- Design H_0 and H_1 ?

Not a single option ! Could be:

- **Population**: the sentences, $\mathcal{S}_0$ sentences from text 1 and $\mathcal{S}_1$ sentences from text 2;

Not a single option ! Could be:

- **Population**: the sentences, $\mathcal{S}_0$ sentences from text 1 and $\mathcal{S}_1$ sentences from text 2;
- **Statistics**: frequency of $\epsilon\nu$ in the sentences, μ_0 is the average frequency in the sentences of text 1 and μ_1 if the average frequency in text 2;
- **Hypothesis**:
  - H_0: the average use of $\epsilon\nu$ per 1000 words is the same in text 1 and text 2;
  - H_1: the average use of $\epsilon\nu$ per 1000 words is different in text 1 and text 2.

Formulating hypothesis

How can we measure if the difference between μ_0 and μ_1 is significant?

Test statistics

Test statistics

A **test statistic** is a quantity derived from the sample for statistical hypothesis testing.

This test statistic is:

- A numerical summary of a dataset that reduces the data to one value;
- That quantifies behaviors that would distinguish the null from the alternative hypothesis.

Test statistics

In practice, statistical testing consist in:

- Computing the test statistic;
- Compute the expected value of this test statistic if H_0 were true;
- Compare the actual measured statistic to the expected statistic.

Example test statistic

Among the most famous tests, two samples *t-test* (Student's test).

Student's test/Welsh's test

A two-sample Student's test tests **if the means of two populations are equal**.

Example test statistic

Among the most famous tests, two samples *t-test* (Student's test).

Student's test/Welsh's test

A two-sample Student's test tests **if the means of two populations are equal**.

The test statistic (Welsh) is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means
s_1 and s_2 are the sample standard deviations
n_1 and n_2 are the sample sizes.

Example test statistic

Under certain assumptions regarding data distribution (**and most of the time in large sample size**!), it is possible to infer the statistical law **followed by the statistic**.

Example test statistic

Under certain assumptions regarding data distribution (**and most of the time in large sample size**!), it is possible to infer the statistical law **followed by the statistic**.

In the case of Welsh's test:

- if $\mu_0 = \mu_1$;
- if the **samples are assumed to be normally distributed**;
- if the **samples are independant**

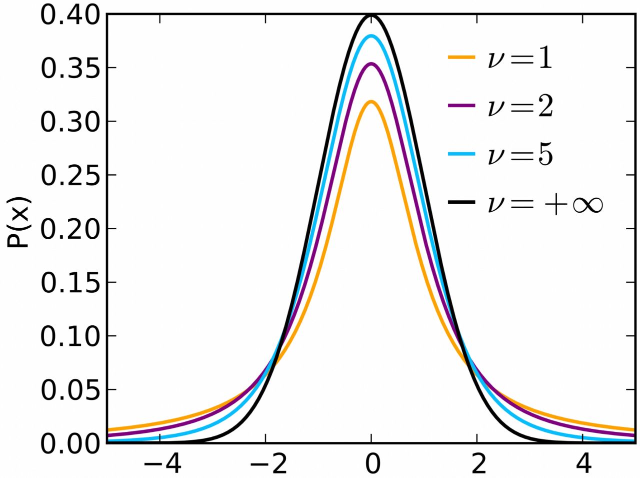
the t statistics follows a Student's distribution with

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

.

Example test statistic

We then compare how likely it is to **observe this value under this hypothesis**.



Significance level

Significance level

The **significance level** is the **probability** of rejecting the null hypothesis when it is true.

Significance level

Significance level

The **significance level** is the **probability** of rejecting the null hypothesis when it is true.

Common levels include $\alpha = 0.1$, $\alpha = 0.05$.

Critical value

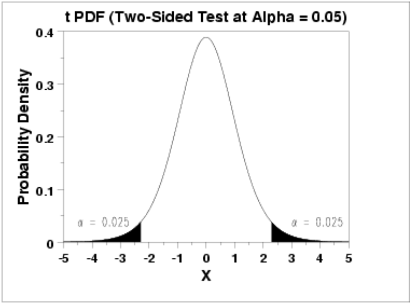
Given the significance level and the law followed by the statistics, one can compute **critical values**.

Critical values

Critical values are the boundaries of:

- the acceptance region of the test: where the null hypothesis is not rejected;
- the rejection region of the test: where the null hypothesis is rejected.

Critical value



P-values

Another possible approach is the use of **p-values**, that provide a **less clear-cut approach**.

p-value

The **p-value** in statistical testing is the probability of **obtaining test results at least as extreme as the observed results**, under the assumption that the null hypothesis is true.

P-values are then compared to the **significance level**. The p-value is compared to the **significance level** and if lower, H_0 is rejected.

Steps in Hypothesis Testing

1. State the null and alternative hypotheses;
2. Choose a significance level (α);
3. Select the appropriate test statistic;
4. Compute the acceptance and rejection regions;
5. Compute the test statistic and p-value;
6. Make a decision: reject or fail to reject H_0.

Example: use of $\epsilon\nu$

Let's practice this on our text dataset !

Can you remind me of the hypothesis we are trying to test ?

Example: use of $\epsilon\nu$

Let's practice this on our text dataset !

Can you remind me of the hypothesis we are trying to test ?

- H_0: the average use of $\epsilon\nu$ per 1000 words is the same in text 1 and text 2;
- H_1: the average use of $\epsilon\nu$ per 1000 words is different in text 1 and text 2.

Example: use of $\epsilon\nu$

Step 0: Pre-processing gives us the frequency vectors for each text.
Text 1: $8.1, 7.8, 8.5, 9.0, 8.2$ (frequencies per 1000 words)
Text 2: $7.5, 7.3, 7.8, 7.6, 7.4$ (frequencies per 1000 words)

Example use of $\epsilon\nu$

Step 1:

Example use of $\epsilon\nu$

Step 1: Compute the statistics.

Example use of $\epsilon\nu$

Step 1: Compute the statistics.

$$\bar{x}_1 = \frac{8.1 + 7.8 + 8.5 + 9.0 + 8.2}{5} = 8.32$$

$$\bar{x}_2 = \frac{7.5 + 7.3 + 7.8 + 7.6 + 7.4}{5} = 7.52$$

$$s_1 = \sqrt{\frac{(8.1 - 8.32)^2 + (7.8 - 8.32)^2 + (8.5 - 8.32)^2 + (9.0 - 8.32)^2 + (8.2 - 8.32)^2}{5 - 1}} =$$

$$s_2 = \sqrt{\frac{(7.5 - 7.52)^2 + (7.3 - 7.52)^2 + (7.8 - 7.52)^2 + (7.6 - 7.52)^2 + (7.4 - 7.52)^2}{5 - 1}} =$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{8.32 - 7.52}{\sqrt{\frac{0.47^2}{5} + \frac{0.19^2}{5}}} = \frac{0.8}{\sqrt{0.044 + 0.0072}} = \frac{0.8}{\sqrt{0.0512}} = \frac{0.8}{0.226} \approx 3.54$$

Example use of $\epsilon\nu$

Step 2:

Example use of $\epsilon\nu$

Step 2: Compute the distribution followed by the test statistic.

Example use of $\epsilon\nu$

Step 2: Compute the distribution followed by the test statistic.

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{0.47^2}{5} + \frac{0.19^2}{5}\right)^2}{\frac{\left(\frac{0.47^2}{5}\right)^2}{4} + \frac{\left(\frac{0.19^2}{5}\right)^2}{4}}$$

$$\text{d.f.} = \frac{(0.044 + 0.0072)^2}{\frac{0.001936}{4} + \frac{0.00001444}{4}} = \frac{0.0512^2}{0.000484 + 0.00000361} = \frac{0.00262144}{0.00048761} \approx 5.38$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{8.32 - 7.52}{\sqrt{\frac{0.47^2}{5} + \frac{0.19^2}{5}}} = \frac{0.8}{\sqrt{0.044 + 0.0072}} = \frac{0.8}{\sqrt{0.0512}} = \frac{0.8}{0.226} \approx 3.54$$

Example use of $\epsilon\nu$

Step 3:

Example use of $\epsilon\nu$

Step 3: Determine the critical value for significance level 0.05.

Example use of $\epsilon\nu$

Step 3: Determine the critical value for significance level 0.05.

For a two-tailed test with a significance level of $\alpha = 0.05$ and degrees of freedom $\approx$ 5.38, the critical value $t_{\alpha/2,d.f.} \approx 2.571$.

Example use of $\epsilon\nu$

Step 4:

Example use of $\epsilon\nu$

Step 4: Make a decision regarding hypothesis.

Example use of $\epsilon\nu$

Step 4: Make a decision regarding hypothesis.
If the critical value is 2.571 and the computed statistic is 3.54, what do you do with H_0?

The statistic is above the critical value, and H_0 is rejected and we keep the alternative hypothesis *the average use of $\epsilon\nu$ per 1000 words is different in text 1 and text 2.*

Example use of $\epsilon\nu$

Step 5:

Example use of $\epsilon\nu$

Step 5: Optionally, compute the p-value.

P-values are computed using the cumulative distributive function.

Example use of $\epsilon\nu$

Step 5: Optionally, compute the p-value.

P-values are computed using the cumulative distributive function.

For $t = 3.54$ and d.f. ≈ 5.38, the two-tailed p-value is approximately $p \approx 0.014$.

Example use of $\epsilon\nu$

Step 5: Optionally, compute the p-value.

P-values are computed using the cumulative distributive function.

For $t = 3.54$ and d.f. ≈ 5.38, the two-tailed p-value is approximately $p \approx 0.014$.

How do you interprete this result ?

Typology of statistical tests

Outline

Parametric and non-parametric statistical tests

- **Parametric tests**: make assumptions regarding the distribution of the sample data;

Parametric and non-parametric statistical tests

- **Parametric tests**: make assumptions regarding the distribution of the sample data;
- **Non parametric tests**: do not make any assumptions on the dataset.

Parametric and non-parametric statistical tests

- **Parametric tests**: make assumptions regarding the distribution of the sample data;
- **Non parametric tests**: do not make any assumptions on the dataset.

Question

In the case of Welsh's test, is it a parametric or non-parametric test ?

Parametric test hypothesis validation

Parametric tests make *a priori* assumptions regarding the behavior of the sample (often that samples:

1. **Normality:** Data should follow a normal distribution.
2. **Homogeneity of variances:** Variances across groups should be equal.
3. **Independence:** Observations should be independent of each other.

In the lab, checking these assumptions will be written as bonus for simplicity's sake.

Assumptions of Parametric Tests

Violating these assumptions can **invalidate the conclusions** of the tests altogether.

Assumptions of Parametric Tests

Violating these assumptions can **invalidate the conclusions** of the tests altogether.

How can we check that the requirements are met by the data so that **our conclusions are valid** ?

Assumptions of Parametric Tests

Validation methods:

- **Normality:**
  - Visual inspection (Q-Q plots, histograms)
  - Statistical tests (Shapiro-Wilk, Kolmogorov-Smirnov)
- **Homogeneity of Variances:**
  - Levene's test
  - Bartlett's test
  - Fligner-Killeen test
- **Independence:**
  - Study design considerations
  - Durbin-Watson test for autocorrelation

Assumptions of Parametric Tests

Validation methods:

- **Normality:**
  - Visual inspection (Q-Q plots, histograms)
  - Statistical tests (Shapiro-Wilk, Kolmogorov-Smirnov)
- **Homogeneity of Variances:**
  - Levene's test
  - Bartlett's test
  - Fligner-Killeen test
- **Independence:**
  - Study design considerations
  - Durbin-Watson test for autocorrelation

For simplicity's sake, verification of hypothesis have been added to the bonus part of the lab (**but they are not bonus in real life!**).

Univariate testing

- Examines a single variable.
- Focuses on understanding the characteristics and patterns within that variable.
- **Common tests**: t-test, welsh test, chi-square test for goodness of fit, statistical distribution test (Shapiro-Wilk ...).

Multivariate Testing Overview

- Examines multiple variables simultaneously.
- Focuses on the relationships and interactions between variables.
- **Common tests**: MANOVA, multiple regression, factor analysis.

Example

Question

Can you give from your own work experience examples that could be analyzed ?

Practical application: the authorship of Colossian

Outline

Python and statistical tests

While Python is not as developed as R for statistical testing, there are several very interesting libraries that **provide the most popular statistical tests**:

- The Scipy library
- The statsmodel library

Today, we will focus on using scipy !

The question of the authorship of Colossian

13 Pauline epistles are considered to be canonical in the New Testament (Christian Scripture), but modern scholarship considers that:

The question of the authorship of Colossian

13 Pauline epistles are considered to be canonical in the New Testament (Christian Scripture), but modern scholarship considers that:

- 7 letters are considered to be **authentic** letters;
- 3 letters are almost unanimously considered to be written by different authors in Paul's name;
- 3 letters are strongly debated.

The question of the authorship of Colossian

13 Pauline epistles are considered to be canonical in the New Testament (Christian Scripture), but modern scholarship considers that:

- 7 letters are considered to be **authentic** letters;
- 3 letters are almost unanimously considered to be written by different authors in Paul's name;
- 3 letters are strongly debated.

The epistle to the Colossians is one of the very debated epistle, and today **we will see together if statistical testing can give us answers**.

A big thanks to Jermo for his idea and methodology :-)

The question of the authorship of Colossian

Stylometry consists in the study of the style of an author, that could **characterize their way of expressing themselves**.

It is say to be **computational** when it is done automatically through computer tools.

The question of the authorship of Colossian

An often used approach to stylometry is the study of **particles and conjunctions** (also used **functional words**).

Our goal today is to assess if the Colossian epistle has **a different frequency of functional words** than the authentic letters.

and you will use all of the skills we learned this week to do it almost alone :-)

Bonus: Linear regression analysis and Colossians

In his study, Jermo used **linear regression models** to infer the relationship between Colossians and the other authentic letters.

Linear Regression models the relationship between two variables as a **linear equation** of observed data X to explain response y.

The model assumes that: $Y = \beta_0 + \beta_1 X + \epsilon$
where β_0 is the **intercept**, β_1 is the **slope**, ϵ **normally distributed error terms**.

Bonus: Linear regression analysis and Colossians

In a subset of the study:

X = total number of tokens. y = particle use.

Epistle	**Number of Tokens**	**Number of particles**
Romans	850	50
1 Corinthian	670	20
2 Corinthian	450	10
Colossian	310	20

We model the particle use as:

$$\text{particle\_use} = \beta_0 + \beta_1 \times \text{token\_number} + \epsilon$$

Fitting the Model

Fitting the model

Fitting a linear regression model consists in finding β_0 and β_1 that **minimize the sum of the squared differences between observed and predicted values**.

We use optimization method to find:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Evaluating the Model

Like any Machine Learning model, the quality of the model needs to be **evaluated** ...

Evaluating the Model

Like any Machine Learning model, the quality of the model needs to be **evaluated** ...

Can anyone remind me of **the approach we used for classification** ?

Evaluating the Model

Like any Machine Learning model, the quality of the model needs to be **evaluated** ...

Can anyone remind me of **the approach we used for classification** ?

- Accuracy
- Precision
- Recall

Bonus: Linear regression analysis and Colossians

Here we are doing regression, and standard approach include:

Bonus: Linear regression analysis and Colossians

Here we are doing regression, and standard approach include:

- **R-squared**: Proportion of variance explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Bonus: Linear regression analysis and Colossians

Here we are doing regression, and standard approach include:

- **R-squared**: Proportion of variance explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- **RMSE (Root Mean Squared Error)**: Measures the average magnitude of the error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

Bonus: Linear regression analysis and Colossians

Here we are doing regression, and standard approach include:

- **R-squared**: Proportion of variance explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- **RMSE (Root Mean Squared Error)**: Measures the average magnitude of the error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

What is in your opinion the optimal value of R-squared and RMSE ?

Bonus: Linear regression analysis and Colossians

Linear regression analysis is a **parametric** models, which means that there are assumptions that need to be validated in order to be able to trust the model.

Bonus: Linear regression analysis and Colossians

Linear regression analysis is a **parametric** models, which means that there are assumptions that need to be validated in order to be able to trust the model.

- **Linearity**: The relationship between X and Y is linear.
- **Independence**: Observations are independent.
- **Homoscedasticity**: Constant variance of the errors/the residuals.
- **Normality**: The errors/the residuals are normally distributed.

Usually, the checking of assumption can be performed visually.

Bonus: Linear regression analysis and Colossians

After **model quality evaluation** and **assumption checking**, we can use the model for **outlier detection**.

Bonus: Linear regression analysis and Colossians

After **model quality evaluation** and **assumption checking**, we can use the model for **outlier detection**.

Prediction interval

A prediction interval provides a **range within which we expect a future observation to fall**, given a specific value of the independent variable(s).

The prediction interval for a new observation Y_{new} given X_{new} is:

$$\hat{Y}_{new} \pm t_{\alpha/2,n-2} \cdot \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Bonus: Linear regression analysis and Colossians

The methodology is the following:

- Fit a regression model on the authentic epistles using various X dataset;
- Evaluate quality of said model;
- Check if Colossian fits within the prediction interval: if it fits, **no deviation**, otherwise **deviation**.

In the example lab, we will use *functional words* count.

Bonus: Linear regression analysis and Colossians

In Python, we rely on `statsmodels`:

- `statsmodels` is a Python module that provides classes and functions for the estimation of many different statistical models;
- It includes tools for performing linear regression, among other statistical analyses;
- It allows for detailed output and in-depth statistical testing.

Bonus: Linear regression analysis and Colossians

- Make the right imports:

```
import statsmodels.formula.api as smf
```

- Define the model:

```
model = smf.OLS('Y ~ X1 + 1', data=data)
```

- Fit the model:

```
results = model.fit()
```

Bonus: Linear regression analysis and Colossians

- Obtain a summary of the model:

```
print(results.summary())
```

- This provides detailed statistics, including coefficients, R-squared, p-values, and more.

Bonus: Linear regression analysis and Colossians

- Define new data for prediction (**here, Colossians**):

```
new_data = pd.DataFrame({'X1': [value1]})
```

- Make predictions:

```
predictions = results.get_prediction(new_data)
```

Bonus: Linear regression analysis and Colossians

- Get summary frame including prediction intervals:

```
prediction_summary = predictions.
summary_frame(alpha=0.05)
print(prediction_summary)
```

- The summary frame includes:
  - `mean`: Predicted mean value
  - `mean_se`: Standard error of the mean prediction
  - `mean_ci_lower`, `mean_ci_upper`: Confidence interval for the mean prediction
  - `obs_ci_lower`, `obs_ci_upper`: Prediction interval for the new observation