**Getting To Know Your Data**

1. Tính giá trị trung bình: AVERAGE()
2. Tính giá trị ở giữa tập hợp: MEDIAN()
3. Giá trị xuất hiện nhiều nhất: MODE()
4. Hàm tính phương sai của toàn bộ tập hợp: VARP()
5. Hàm tính độ lệch chuẩn của toàn bộ tập hợp: STDEVP()
6. Tính tứ phân vị của một tập hợp: QUARTILE(Data, thứ tự phân vị %1,2,3,4% )
7. Tính tương đương chuẩn hóa của một biến ngẫu nhiên %Tính giá trị Z\_core =(giá trị - trung bình)/độ lệch chuẩn% : STANDARDIZE(Giá trị, TB, độ lệch chuẩn)

**Statistical Data Visualization**

1. **Biểu đồ tần số.**

Vào Insert -> Chart -> Histogram

Is the data "normally" distributed?

Now you want to use statistics to verify if the distribution is approximately symmetrical or "normally distributed". To do so, you will calculate the skew and kurtosis.

Kurtosis: Calculated using KURT() %Hệ số nhọn của tập dữ liệu%. Measures how the tails behave. It identifies how values are concentrated around the mean and how they trail away from it in the tails.

Skew: Calculated using SKEW() % Tính độ lệch của tập dữ liệu%. Measures how symmetrical the distribution is. 0 means the distribution is exactly symmetrical. Values above or below 0 indicate that there are more values above or below the mean.

-> Trị tuyệt đối 2 hệ số KURT() và SKEW() <2 thì dữ liệu có phân phối chuẩn

1. **Biểu đồ phân tán**

Vào Insert -> Chart -> Scatter

Tìm hệ số tương quan giữa X và Y %tích Pearson%: CORREL() -> Giá trị này trong khoảng [-1;1]

Spreadsheets have 3 formulas for calculating the y-intercept and slope given two variables.

SLOPE() %Độ dốc của đường hồi quy tuyến tính của dữ liệu%- will return the slope of a trend line or linear regression representing the linear change in one unit to another.

INTERCEPT() % giá trị y cũa đường hồi quy tuyến tính khi x =0) - returns the value where the trendline will intersect the y-axis.

LINEST() %trả về hàm xu hướng tuyến tính phù hợp nhất thông qua pp bình phương nhỏ nhất% - calculates both the slope & the intercept of two variables using the least-squares method.

1. **Biều đồ cột**

Vào Insert -> Chart -> Bar

The difference is subtle but helps your audience in different ways.

Stacked: helps you focus on the proportion of the total that is a particular class. You will visually understand the amount of the total that represents competitive online auctions.

Side-by-side: lets the audience understand the values in relation to each other. You will visually interpret how similar each class values are.

**Statistical Hypothesis Testing**

1. **Sampling in Spreadsheets**

Population: An entire collection of observations or events

Sample: A subset from a population meant to represent the population

There are many ways to sample data. Sampling choices affect the statistics and as a result, what you learn about the data population. A popular method for sampling is to do it randomly.

This data has an extra column of randomly generated numbers that was created with the RAND() function, which can be empty and returns numbers between 0 and 1. Once the second column has been created, sort the data array by the randomly created column. Then when constructing visuals or stats based on the sample, you can control the cell range. In this exercise you will randomly sample 20 observations of monthly US airline passenger miles, and then create a histogram of this sample to understand the distribution.

Many analytics professionals believe a sample size of 30 or more is large enough to get trustworthy descriptive stats.

Central Limit Theorem: If a sample size from an independent, random variable is large enough, then the sampling distribution will be normal or nearly normal.

1. **Comparing samples with a t-test\_ Kiểm định sự khác biệt 2 tổng thể (mẫu nhỏ n<30)**

As you just learned sample independence can be tested with a T.TEST()!

T.TEST(range1, range2, tails, type) = p-value

Range 1: Mẫu dữ liệu hoặc nhóm ô đầu tiên để xem xét cho hàm t-test.

Range 2: Mẫu dữ liệu hoặc nhóm ô thứ hai để xem xét cho hàm t-test.

Tails: Quy định số bên của phân phốI ( 1 phía hay 2 phía)

Type: quy định kiểu t-test ( có cùng phương sai hay khác phương sai; 1 là khác, 2 là cùng)

Trả về xác suất được liên kết với hàm t-test là p\_value . Xác định xem hai mẫu có khả năng xuất phát từ hai tập hợp gốc có cùng giá trị trung bình hay không.

Nếu p-value < alpha thì bác bỏ H0

1. **Hypothesis Testing with the Z-test\_ Kiểm định 1 tổng thể dùng z-test (n>=30)**

Z.TEST(dữ\_liệu, giá\_trị, [độ\_lệch\_chuẩn]) -> Cho ra p\_value

dữ\_liệu :Mảng hoặc dải ô có chứa tập dữ liệu cần xem xét.

giá\_trị: Thống kê kiểm định cần sử dụng trong phép thử Z.

độ\_lệch\_chuẩn - [tùy chọn]: Độ lệch chuẩn cần giả định cho phép thử Z. Nếu độ lệch chuẩn không được cung cấp, độ lệch chuẩn của dữ liệu sẽ được sử dụng.

Nếu Kiểm định 1 phía: p-value < alpha %P\_value cutoff% thì bác bỏ H0

Kiểm định 2 phía (H0=): p-value <alpha/2 %P\_value cutoff% thì bác bỏ H0

1. **Hypothesis Testing with the Chi-squared test\_ Kiểm định giả thuyết với kiểm định Chi bình phương**

Khi bạn có các quan sát trước và một mẫu mới, bạn chỉ có hai lựa chọn ... sự khác biệt trong nhóm mới bắt nguồn từ các biến thể lấy mẫu ngẫu nhiên %KHÔNG CÓ SỰ KHÁC BIỆT% HOẶC rằng trên thực tế có một sự khác biệt có ý nghĩa

Hai lựa chọn này được sử dụng cho giả thuyết H0 và giả thuyết thay thế H1. Hơn nữa, như trước khi kiểm tra chi bình phương sẽ cung cấp xác suất rằng hai nhóm là độc lập hoặc thực sự khác nhau.

Số mẫu n>5

CHITEST(dải\_ô\_đã\_quan\_sát, dải\_ô\_dự\_kiến) -> Cho ra p\_value

Nếu p-value < alpha %P\_value cutoff% thì bác bỏ H0