💚

Học Excel: Phân tích dữ liệu

1. **Các khái niệm cơ bản về phân tích dữ liệu**

▼ Tính giá trị trung bình và giá trị trung vị

- Giá trị trung bình

```
= AVERAGE(x:y)
```

- Giá trị trung vị
  - Các số đã được sắp xếp theo thứ tự, trung vị là số nằm ở giữa một nhóm các số, có nghĩa là, phân nửa các số có giá trị lớn hơn số trung vị, còn phân nửa các số có giá trị bé hơn số trung vị

```
= MEDIAN(x:y)
```

- Mode
  - Là số xuất hiện nhiều nhất trong một nhóm các số.

```
= MODE(x:y)
```

▼ Đo lường max, min và các đặc điểm dữ liệu khác

- Min

```
= MIN(x:y)
```

- Max

```
= MAX(x:y)
```

- Các gốc phần tư của dữ liệu
  - Có hai cách tính theo hai phương pháp khác nhau, chọn cách nào cho phù hợp tùy theo nhu cầu, cụ thể:
  - Gốc phần tư thứ nhất

```
= QUARTILE(x:y,1)
= QUARTILE.INC(x:y,1)
= QUARTILE.EXC(x:y,1)
```

  - Gốc phần tư thứ hai (giá trị trung vị - median)

```
= QUARTILE(x:y,2)
= QUARTILE.INC(x:y,2)
= QUARTILE.EXC(x:y,2)
```

  - Gốc phần tư thứ ba

```
= QUARTILE(x:y,3)
= QUARTILE.INC(x:y,3)
= QUARTILE.EXC(x:y,3)
```

▼ Phân tích dữ liệu bằng cách sử dụng phương sai

- Bước 1: Tính sai số (giá trị - giá trị trung bình (mean))
- Bước 2: Tính bình phương sai số (giá trị sai số bình phương): Nhằm đảm bảo một giá trị dương cho phương sai
- Phương sai
  - Phương sai là một trong những thước đo của sai số
  - Phương sai của một tập hợp là giá trị trung bình của tất cả bình phương sai số

```
= AVERAGE(X:Y)
= VAR.P(X:Y)
```

Tính Var cho mẫu

```
= SUM(X:Y)/(COUNT(X:Y)-1)
= VAR.S(x:y)
```

- Độ lệch chuẩn
  - Là căn bậc hai của phương sai

```
= SQRT(AVERAGE(X:Y))
= STDEV.P(x:y)
```

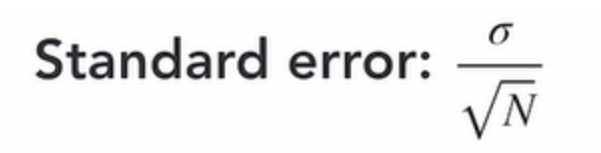
Độ lệch chuẩn cho mấu

```
= STDEV.S(x:y)
```

▼ Giới thiệu định lý giới hạn trọng tâm

▼ Phân tích dân số bằng cách sử dụng các mẫu dữ liệu

- Kĩ thuật nên làm
  - Thu thập mẫu càng lớp càng tốt
  - Cần ước lượng độ lệch chuẩn của dân số (dựa trên khảo sát, kiến thức)
  - Xác định độ tinh cậy mong muốn
  - Cần tính toán biên độ sai số, cách tính
    - Tính lỗi tiêu chuẩn



Trong đó:

$$\sigma : Độ lệch chuẩn$$

N: Là số mẫu

- Biên độ sai số là lỗi tiêu chuẩn nhân với điểm Z (Z là số độ lệch chuẩn so với giá trị trung bình - Có bảng tính Z)

• Standard deviation of 0.1 ounces

• 95% certainty (z-score of 1.96)

• 40 measurements

$$• 1.96 \times \frac{0.1}{\sqrt{40}} = 0.03$$

▼ Xác định và giảm thiểu các nguồn lỗi

2. **Trực quan hóa dữ liệu**

▼ Nhóm dữ liệu bằng cách sử dụng biểu đồ

- Các bước:
  - Có thể chọn phạm vi dữ liệu
  - Chọn tab Insert → Biểu đồ mong muốn
  - Thay đổi các trục nhấn chuột phải ở bất kì nơi nào trên trục muốn thay đổi →

▼ Xác định các mối quan hệ bằng cách sử dụng biểu đồ phân tán XY

- Cách thực hiện như cách tạo một biểu đồ ở trên, sau đó chọn biểu đồ Scatter

▼ Trực quan hóa dữ liệu bằng thang đo Logarit

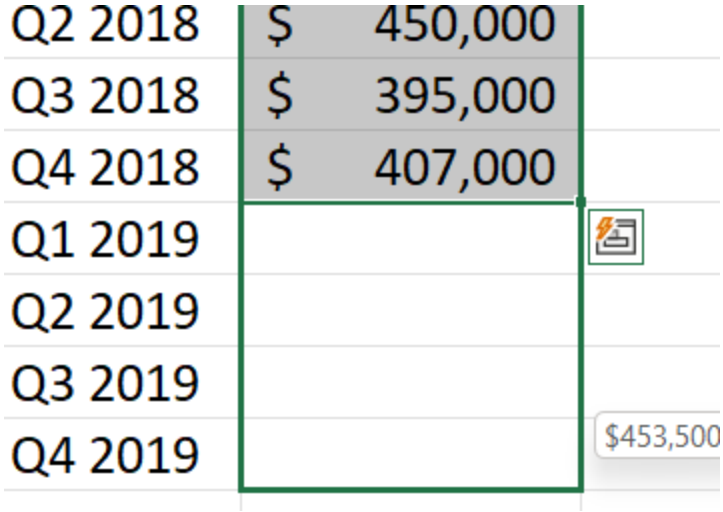
- Được dùng cho các bộ dữ liệu có chữ số lớn

▼ Thêm đường xu hướng vào biểu đồ

- Muốn kiểm tra các xu hướng
- Tạo biểu đồ → nhấp chuột vào ô góc trên bên phải của biểu đồ → chọn Trendline

▼ Dự báo kết quả trong tương lai

- Dự đoán kết quả dựa trên các kết quả đã biết trước đó
- Có thể dự đoán bằng trend



- Dự đoán bằng hàm
  - FORECAST.LINEAR: Dự báo tuyến tính (gồm ba tham số: đối tượng dự đoán, dữ liệu phụ thuộc, dữ liệu độc lập)

▼ Tính toán trung bình

- Sử dụng hàm Average

3. **Kiểm tra một giả thuyết**

▼ Hình thành giả thuyết

- Khi hình thành một giả thuyết là đang đưa ra một phỏng đoán có học thức về một mối quan hệ giữa hai bộ dữ liệu

▼ Giải thích kết quả phân tích

▼ Xem xét các giới hạn của việc kiểm tra giả thuyết

4. **Sử dụng phân phối dữ liệu**

▼ Sử dụng phân phối chuẩn

- Phân phối chuẩn được xác định bằng hai giá trị
  - Giá trị trung bình
  - Độ lệch chuẩn : Là mức chênh lệch trong dữ liệu
- Tính xác suất
  - Sử dụng hàm NORM.DIST (cần cung cấp 4 đối số: giá trị cần tính, giá trị trung bình, độ lệch chuẩn, true: nếu muốn là hàm phân phối tích lũy còn flase là ngược lại
- Tìm điểm dừng (trong đó 71% giá trị sẽ xảy ra dưới một điểm cụ thể)
  - Sử dùng hàm NORM.INV (cần 3 đối số: xác suất, giá trị trung bình, độ lệch chuẩn)

▼ Sử dụng phân phối hàm mũ

- Tính xác suất trong phân phối hàm mũ
  - Sử dụng hàm EXPON.DIST(giá trị cần tính, giá trị lambda (= 1/mean), chọn sử dùng phân phối tích lũy hay là không(true, flase))
- Nếu muốn tìm xác suất giữa hai thời điểm
  - Ví dụ: Xác suất của 10s - 15s
    - Bằng xác suất 15s - xác suất tích lũy trong 10s

▼ Sử dụng phân phối đồng đều

- Cách tạo các giá trị ngẫu nhiên
  - Sử dụng hàm RANDBETWEEN
- Tính xác suất
  - 1/ tổng các giá trị
- Muốn tạo các giá trị số thực ngẫu nhiên từ 0 đến 1

- Sử dụng hàm rand, cụ thể nhập: = rand()

▼ Sử dụng phân phối nhị thức

- Tính xác suất của một giá trị nào đó
  - Sử dụng hàm BINOM.DIST(cần 4 đối số: giá trị xét, số mẫu, xác suất thành công, chọn xác suất tích lũy hay không(true, false))

▼ Sử dụng phân phối Poisson

- Phân phối Poisson dùng để ước tính một hoạt động trong một thời gian nhất định. Giá trị cần thiết trong phân phối Poisson là mean
  - Sử dụng hàm POISSON.DIST

5. **Đo lường hiệp phương sai và tương quan**

▼ Hình dung hiệp phương sai có ý nghĩa gì

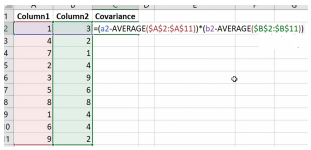
- Công thức tính hiệp phương sai

$$\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

  - Đối với mỗi điểm dữ liệu trong hai tập dữ liệu sẽ thấy được độ lệch của nó so với giá trị trung bình
- Giải thích các giá trị hiệp phương sai
  - Kết quả bằng 0: Các tập dữ liệu không liên quan với nhau
  - Kết quả dương: Có mối quan hệ với nhau, giá trị càng lớn càng có mối quan hệ mạnh mẽ hơn
  - Kết quả âm: Có quan hệ nghịch đảo nhau

▼ Tính hiệp phương sai giữa hai cột dự liệu

- Cách 1:
  - Bước 1: Tính từng phép nhân ở tử



- Bước 2: Tính tổng và hoàn thiện phép tính ở tử số
- Bước 3: Chia tử số cho số mẫu

- Cách 2: Sử dụng hàm COVARIANCE.P hoặc COVERIANCE.S

▼ Tính hiệp phương sai giữa nhiều cặp cột

▼ Hình dung mối tương quan có ý nghĩa gì

- Công thức tính độ tương quan

$$\frac{\Sigma(x-\bar{x})(y-\bar{y}}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}$$

- Ý nghĩa của kết quả mang lại:
  - Kết quả bằng 0 cho biết dữ liệu không liên quan về nhau
  - 0 ≤ x < 1: Tương quan thuận, dữ liệu có xu hướng di chuyển theo cùng một chiều
  - -1 ≤ x < 0: Tương quan nghịch, dữ liệu có xu hướng đi ngược chiều nhau

▼ Tính toán mối tương quan giữa hai cột dữ liệu

- Sử dụng hàm CORREL

▼ Tính tương quan giữa nhiều cặp cột

6. **Thực hiện phân tích Bayes**

▼ Giới thiệu phân tích bayes

- Thống kê mô tả cung cấp thông tin thực tế về dữ liệu của bạn
- Quy tắc Bayes

- Có thể kết hợp độ chính xác, độ xác thực sai và tỉ lệ cơ bản để tìm kết quả

▼ Tạo ma trận phân loại

▼ Tính xác suất Bayes trong Excel

Given Circumstances				
Base Rate (Green)	85%		Reported Green	Reported Blue
Accuracy	80%	Actually Green	68.00%	17.00%
		Actually Blue	3.00%	12.00%
Probability of Green	85%			
Probability of Blue	15%		Prob. Cab is Blue When Reported Blue	Prob. Cab is Green When Reported Green
			41%	96%
Probability correct	80%			
Probability incorrect	20%			

▼ Cập nhật phân tích Bayes của bạn