



ĐẠI HỌC FPT

BÁO CÁO

**Nghiên Cứu Tỷ Lệ Gia Tăng Dân Số Và
Chênh Lệch Thu Nhập Bình Quân Đầu Người
Qua Các Năm
- 2011, 2015, 2020 -**

**Computer Project
MAS291- SE1609**

I. Giới thiệu

- **Giảng Viên:** Nguyễn Việt Anh

- **Lớp:** SE1609

- **Thành viên:**

STT	Họ và Tên	Mã sinh viên	Nội dung thực hiện
1	Nguyễn Văn Sỹ	HE150242	Tính toán kiểm tra giả định.
2	Trương Sỹ Đạt	HE153095	Tính toán hồi quy tuyến tính.
3	Đỗ Xuân Hiếu	HE153537	Tính toán khoảng tin cậy.
4	Ngô Doãn Luân	HE151070	Xử lý phân tích data, và thực hiện vẽ biểu đồ

- **Nguồn lấy dữ liệu:**

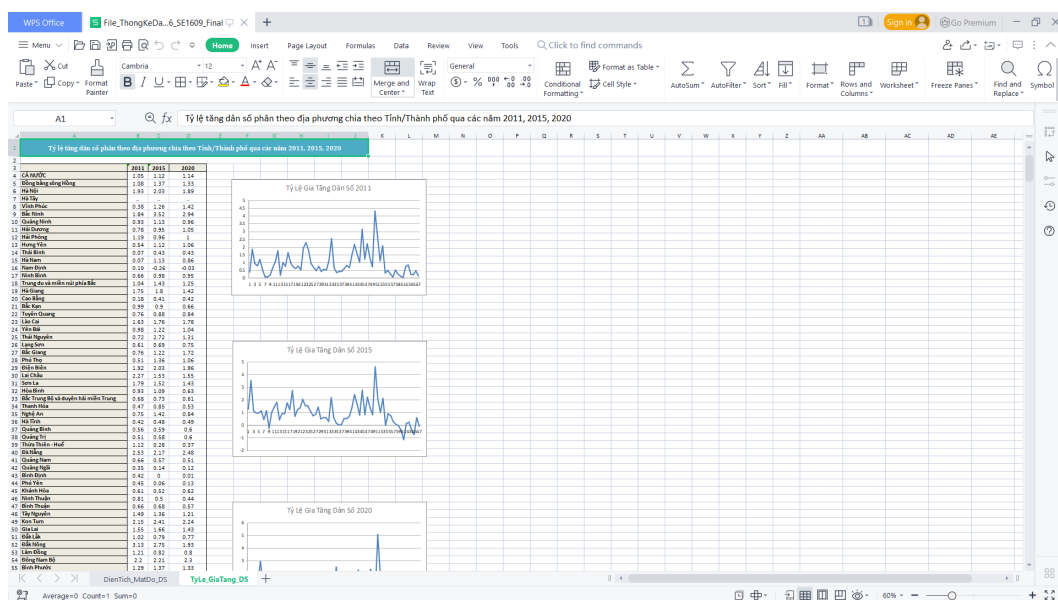
- Diện tích, dân số và mật độ dân số:

<https://www.gso.gov.vn/px-web-2/?pxid=V0201&theme=D%C3%A2n%20s%E1%BB%91%20v%C3%A0%20lao%20%C4%91%E1%BB%99ng>

- Thu nhập bình quân đầu người:

<https://www.gso.gov.vn/px-web-2/?pxid=V1129&theme=Y%20t%E1%BA%BF%2C%20v%C4%83n%20h%C3%B3a%20v%C3%A0%20%C4%91%E1%BB%9D%20s%E1%BB%91ng>

- Sau đó chuyển dữ liệu về Excel và bắt đầu xử lý, cũng như vẽ các biểu đồ liên quan :



Trích nguồn: Cổng thông tin điện tử Tổng cục Thống kê www.gso.gov.vn

Tổng cục Thống kê (tiếng Anh: General Statistics Office, viết tắt là GSO) là cơ quan trực thuộc Bộ Kế hoạch và Đầu tư, thực hiện chức năng tham mưu, giúp Bộ trưởng Bộ Kế hoạch và Đầu tư quản lý nhà nước về thống kê, điều phối hoạt động thống kê, tổ chức các hoạt động thống kê và cung cấp thông tin thống kê kinh tế - xã hội cho các cơ quan, tổ chức và cá nhân theo quy định của pháp luật.

- Lý do lựa chọn đề tài:

Sau khi cả nhóm thảo luận cũng như tìm hiểu nhiều nguồn data khác trên mạng, nhóm em đã thống nhất lựa chọn đề tài nghiên cứu về ***Dân số*** và ***Thu nhập bình quân đầu người(GDP)*** qua các năm 2011, 2015, 2020.

Bắt nguồn từ thực tế diễn biến phức tạp do đại dịch Covid-19 gây ra “*Tổng cục Thống kê cho biết, trong quý III/2021 và 9 tháng năm 2021, thị trường lao động Việt Nam chịu ảnh hưởng nghiêm trọng do diễn biến phức tạp của dịch COVID-19 lần thứ 4 và thời gian giãn cách xã hội kéo dài.*” hầu hết các ngành kinh tế đều ghi nhận mức sụt giảm thu nhập bình quân đầu người so với quý trước ở ngưỡng đáng báo động (thu nhập bình quân đạt 6,2 triệu đồng/tháng giảm 13,5% so với quý II/2021). Từ đó nhóm em muốn tìm hiểu và đưa ra được sự thay đổi thu nhập bình quân đầu người từ các năm trước khi có dịch Covid-19 bùng phát, cùng với đó là tỷ lệ gia tăng mật độ dân số qua các năm khác nhau.

II. Thực hiện dự án (File Excel tính toán các bài đính kèm)

A. Tính khoảng tin cậy:

- Câu hỏi : Từ số liệu như trên, khoảng trung bình nào số dân trung bình của các tỉnh với độ tin cậy là 95% trong 3 năm 2011, 2021?
 - Trong file Excel, chúng em đã sử dụng các hàm =AVERAGE() và =STDEV() để tính trung bình của mẫu và độ lệch chuẩn của mẫu có được.
- Vì lý do Excel được tối ưu hóa cho việc tính toán thống kê, tránh được sự sai số và nhầm lẫn dữ liệu khi chúng ta tính toán bằng tay nên nhóm em xin phép được tận dụng luôn hàm có sẵn trong Excel để tiện cho việc tính toán.
- Kích thước mẫu (n): **63** (các tỉnh thành toàn quốc Việt Nam)
- Giá trị trung bình của mẫu (\bar{X}): sử dụng hàm “=AVERAGE(“vùng dữ liệu”)” để tính giá trị trung bình của dân số các năm:

	2011	2015	2020
Giá trị trung bình của mẫu	3985,65	4213	4728,3

- Độ lệch chuẩn của mẫu (s)(standard deviation):

độ lệch cũng được tính trên Excel bằng hàm “=STDEV(“vùng dữ liệu”)”

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

	2011	2015	2020
Độ lệch chuẩn của mẫu (S)	3925,361874	4240,590077	4987,011996

- Cách tính khoảng tin cậy:

Khác với giá trị trung bình hay độ lệch chuẩn, khoảng tin cậy không thể tính

trực tiếp bằng 1 hàm trên Excel mà được tính bằng các bước sau:

- Bước 1: Tính giá trị trung bình, độ lệch chuẩn (đã có được từ phần c))
- Bước 2: Tính giá trị t (alpha / 2) (1 giá trị được lấy từ bảng) t bằng hàm “=tinv(“độ tin cậy / 2”, “mức độ tự do”)”. Trong đó độ tin cậy được sử dụng trong project là 95% từ đó ta có alpha = 5%, một giá trị phổ biến trong bộ môn MAS291. Bên cạnh đó thì “mức độ tự do” được tính dễ dàng bằng cách lấy tổng số các cá thể trong mẫu được thu thập trừ đi 1.
- Áp dụng công thức khoảng tin cậy:

$$E = \left(t_{\frac{\alpha}{2}, n-1} \right) \times \frac{S}{\sqrt{n}}$$

- Ta dùng hàm excel sau đó tính được $E = 2,297142129x$

	2011	2015	2020
Giá trị trung bình	3985,65	4213	4728,3
E	1135,601853	1227,280644	1443,30463
Khoảng tin cậy	(2850,048147; 5121,251853)	(2985,719356; 5440,280644)	(3284,99537; 6171,60463)

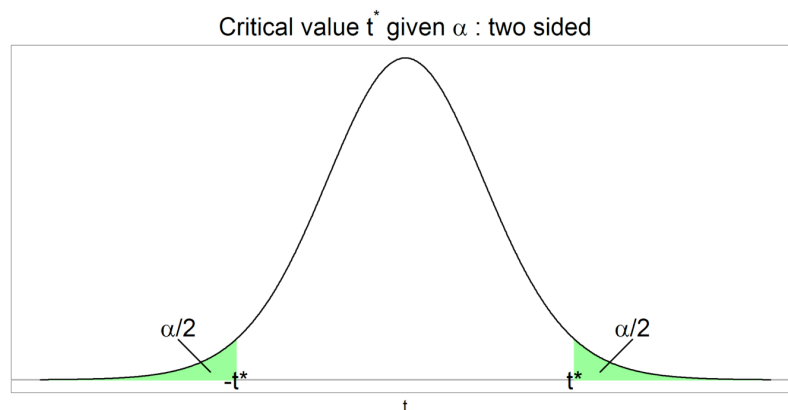
- Chúng em đã có được khoảng tin cậy trên bảng Excel là:

B. Kiểm tra giả định:

- Kiểm tra giả định mật độ dân cư tập trung ở các thành phố lớn có phản ánh được mức sống trung bình của thành phố đó hay không dựa vào GDP thu nhập bình quân đầu người mật độ dân số thay đổi như thế nào 2011 - 2020.
- Xây dựng các giả định H_0 , H_1 . Như giả định ở trên, giả định nhóm em hướng tới dân số của từng địa phương ảnh hưởng bởi thành phố đó có mức sống cao nên dân cư tập chung đông đúc. H_0 mật độ dân số phản ánh thu nhập bình quân đầu người, H_1 mức thu nhập bình quân không thực sự phản ánh được mật độ dân số có tăng của địa phương đó không mà phụ thuộc do diện tích địa phương đó.
- Tỷ lệ H_0 H_1 sẽ là 50% \Rightarrow ta có p: **0.05** ($H_0 = 0,05$ | $H_1 = 0,05$)
- Ta tính được $p\text{ hat} = x/n = \mathbf{0,476}$
 - ☐ x : Lấy ra ngẫu nhiên 30 tỉnh
 - ☐ n : Kích thước mẫu

$$\text{Test statistic } Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Ta tính được $z_0 = 15,521$
- Giá trị trái $-1,96$, giá trị phải $1,96$



Hypothesis testing	
alpha	5%
n	63
x	30
p hat	0,476
z test	15,521
z value (right)	1.96
z value (left)	-1.96

- Sau đó ta áp dụng công thức sau để tìm ra khoảng tin cậy:

C.I for p:

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Constructing confidence interval

C.I value (right)	0.60
C.I value (left)	0,35

- Tỷ lệ thu được khả năng giả thiết có thể xảy ra sẽ từ **35% - 60%**

C. Hồi quy tuyến tính

- Scatter plot (biểu đồ phân tán): Ta chọn 2 quần thể là số dân trung bình và thu nhập bình quân đầu người/tháng.
- Ta add dữ liệu vào excel tính toán và vẽ biểu đồ.
- Kết quả thu được như hình bên dưới với phương trình đường hồi quy tuyến tính có giá trị là năm 2011:
- Kết quả thu được như hình bên dưới với phương trình đường hồi quy tuyến tính có giá trị là năm 2020:
- Ta sẽ sử dụng các công thức sau để tính ra S_{xx} , S_{xy} , SS_T . Nhưng trước tiên ta cần tính được các hệ Số X_i , X_{i2} , Y_i , Y_{i2} bằng các hàm có sẵn trong Excel **=SUMSQ()** và **=SUM()²**. Sau khi tính ra S_{xx} , S_{xy} , SS_T thì ta có thể tính được (β_0) và (β_1) . Ta có được kết quả như sau:

$$\text{Correlation coefficient } R = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SS_T}}$$

- Ta sẽ dùng công thức sau để tính ra được hệ số tương quan:

Correlation coefficient	
R	0,7216112439

1. Sự liên quan giữa mật độ dân số và thu nhập trung bình có hay không

Năm 2011

Năm 2020

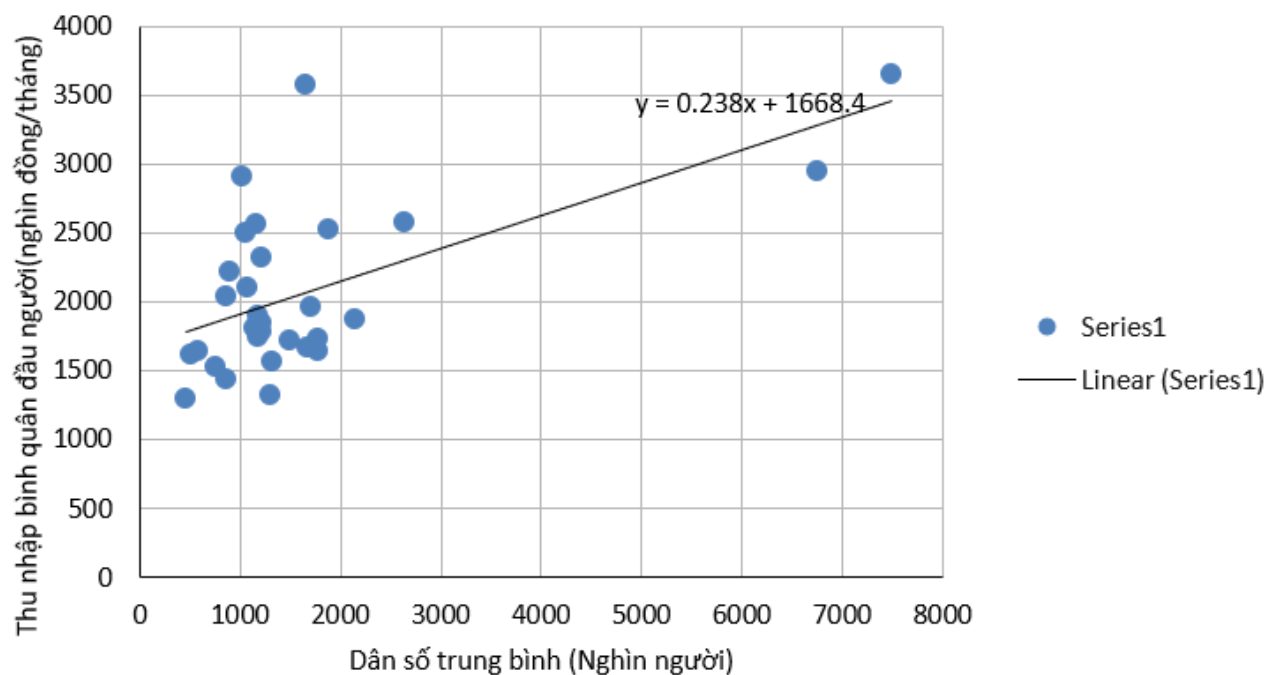
Estimated regression line		Estimated regression line	
n	30	n	30
sum (Xi^2)	154424347,6	sum (Xi^2)	216525127,8
(sum Xi)^2	1874015442	(sum Xi)^2	3182494265
sum (Yi^2)	138666174	sum (Yi^2)	546609407
(sum Yi)^2	3839785156	(sum Yi)^2	15148447626
sum (Xi*Yi)	120262193,7	sum (Xi*Yi)	275082158,3
sumXi * sumYi	3101472659	sumXi * sumYi	6943331167
Sxx	91957166,18	Sxx	110441985,6
Sxy	16879771,73	Sxy	43637786,11
SSt	10673335,47	SSt	41661152,76
beta1 hat	0,18356124	beta1 hat	0,39511954
beta0 hat	1.800,65507	beta0 hat	3.468,24358

Giả sử mô hình tuyến tính là hợp lý. Ta có phương trình tuyến tính đơn giản:

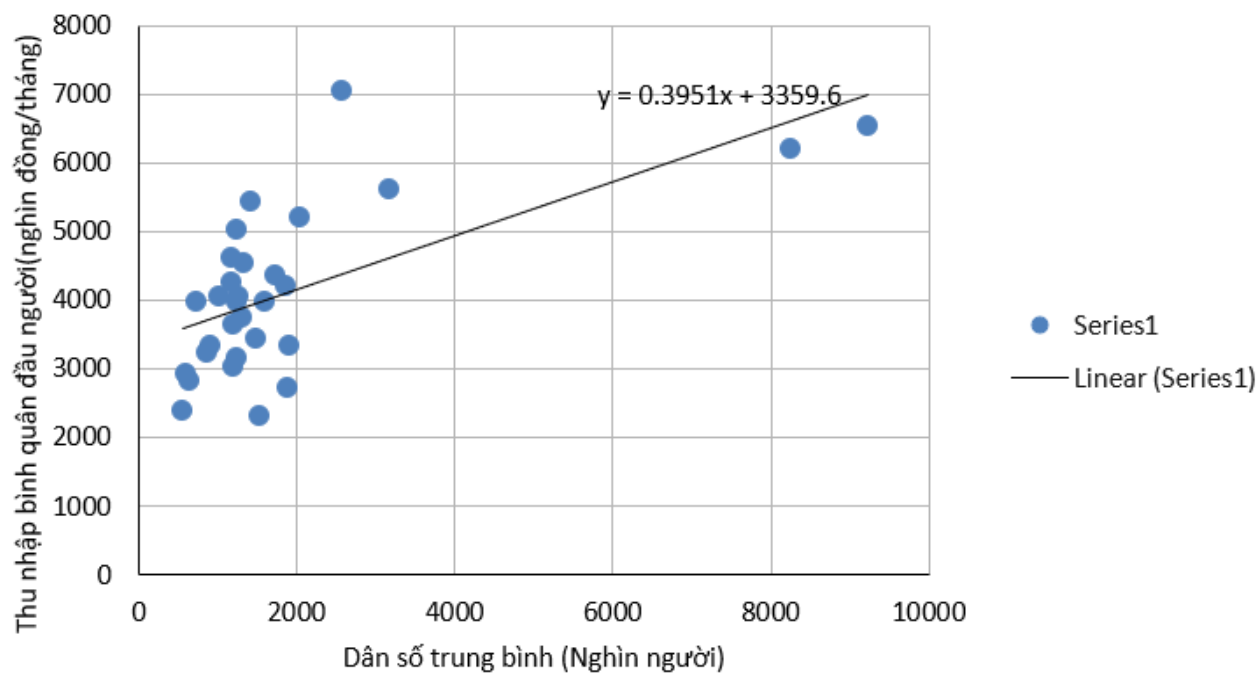
$$y = \beta_0 + \beta_1 x$$

Từ số liệu thu thập được với Y đại diện cho số thu nhập bình quân đầu người (nghìn vnd) và X đại diện cho mật độ dân số của các tỉnh, ta có ước tính bình phương nhỏ nhất của các hệ số trong mô hình hồi quy tuyến tính đơn giản:

Thu nhập bình quân đầu người(nghìn đồng/tháng) so với Dân số trung bình (Nghìn người) - 2011



Thu nhập bình quân đầu người(nghìn đồng/tháng) so với Dân số trung bình (Nghìn người) - 2020



Áp dụng kiểm kiểm định giả thiết để kiểm tra mức độ hợp lí của mô hình tuyến tính đơn giản:

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_1 \neq 0$$

Sử dụng biến t-Test với mức độ tin cậy $\alpha = 0.05$


$$t_{\frac{\alpha}{2}, n-2} = 2.048$$

Từ số liệu ta có thể ước tính mức độ tương quan giữa dân số trung bình và thu nhập bình quân đầu người của 30 tỉnh thành năm 2011 và 2020 theo phương trình tuyến tính: 2011 có $y = 0.238x + 1668.4$ và 2020 có $y = 0.3951x + 3359.6$

IV. Đánh giá kết quả nghiên cứu

- Các thành viên hoàn thành nhiệm vụ 100% và nộp bài đúng hạn
- Qua kết quả cho thấy thì mật độ dân cư tập trung đông ở các thành phố do tập trung nhiều hoạt động kinh tế nên thu nhập bình quân đầu người cũng nhỉnh hơn so với các địa phương khác.

Phụ Lục

-  File_ThongKeDanSo_Gr6_SE1609.xlsx (Hoặc file Excel đính kèm)
- Nguồn data chính: www.gso.gov.vn

- End -