

MỤC LỤC

CHƯƠNG 1: THỐNG KÊ MÔ TẢ.....	3
1.1 Tiến hành nhập dữ liệu:	5
1.2 Tiến hành thống kê:	5
1.2.1. Bảng phân phối tần số, tần suất:	5
1.2.2. Đa giác tần số:	7
1.2.3. Các dạng biểu đồ:.....	9
1.2.4. Các giá trị đặc trưng của biến:	18
CHƯƠNG 2: LÝ THUYẾT ƯỚC LƯỢNG.....	20
3.1 Ước lượng khoảng cho giá trị trung bình	20
3.1.1 Trường hợp σ được biết	20
3.1.2. Trường hợp σ chưa biết.....	22
3.2. Ước lượng khoảng cho giá trị tỷ lệ	23
CHƯƠNG 3: KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ....	25
3.1. Kiểm định về giá trị trung bình:.....	25
3.2. Kiểm định về tỉ lệ:	27
3.3. Kiểm định về 2 giá trị trung bình	28
3.4. Kiểm định về 2 tỉ lệ:	29
Chương 4: HỒI QUY VÀ TƯƠNG QUAN	31

CHƯƠNG 1: THỐNG KÊ MÔ TẢ

Sử dụng số liệu về tỉ lệ dân số Việt Nam phân chia theo tỉnh/thành phố phân chia trong 2 năm 2021 và 2022.

Tỉnh/thành phố	Năm 2021	Năm 2022
An Giang	1.94	1.92
Bà Rịa - Vũng Tàu	1.19	1.18
Bạc Liêu	0.93	0.93
Bắc Giang	1.9	1.9
Bắc Kạn	0.33	0.33
Bắc Ninh	1.49	1.5
Bến Tre	1.32	1.3
Bình Dương	2.64	2.78
Bình Định	1.53	1.51
Bình Phước	1.04	1.04
Bình Thuận	1.27	1.26
Cà Mau	1.23	1.21
Cao Bằng	0.55	0.55
Cần Thơ	1.27	1.26
Đà Nẵng	1.21	1.23
Đắk Lắk	1.94	1.93
Đắk Nông	0.67	0.67
Điện Biên	0.63	0.64
Đồng Nai	3.22	3.27
Đồng Tháp	1.63	1.61
Gia Lai	1.59	1.6
Hà Giang	0.9	0.9
Hà Nam	0.89	0.88
Hà Nội	8.46	8.48
Hà Tĩnh	1.33	1.33
Hải Dương	1.97	1.96
Hải Phòng	2.1	2.1
Hậu Giang	0.74	0.73
Hoà Bình	0.88	0.88
Hưng Yên	1.3	1.3
Kiên Giang	1.78	1.76
Kon Tum	0.58	0.58

Khánh Hoà	1.27	1.26
Lai Châu	0.49	0.48
Lạng Sơn	0.81	0.81
Lào Cai	0.77	0.77
Lâm Đồng	1.34	1.34
Long An	1.75	1.74
Nam Định	1.86	1.89
Ninh Bình	1.02	1.02
Ninh Thuận	0.61	0.6
Nghệ An	3.46	3.44
Phú Thọ	1.53	1.52
Phú Yên	0.89	0.88
Quảng Bình	0.92	0.92
Quảng Nam	1.54	1.53
Quảng Ninh	1.37	1.37
Quảng Ngãi	1.26	1.25
Quảng Trị	0.66	0.65
Sóc Trăng	1.23	1.2
Sơn La	1.31	1.31
Tây Ninh	1.2	1.2
Tiền Giang	1.81	1.79
TP.Hồ Chí Minh	9.31	9.44
Tuyên Quang	0.81	0.81
Thái Bình	1.9	1.89
Thái Nguyên	1.34	1.34
Thanh Hoá	3.77	3.74
Thừa Thiên Huế	1.17	1.17
Trà Vinh	1.03	1.02
Vĩnh Long	1.04	1.03
Vĩnh Phúc	1.21	1.2
Yên Bái	0.86	0.85

(Nguồn: internet)

***Phân tích dữ liệu:**

Dữ liệu trên cho thấy rằng tỷ lệ dân số năm 2021 và năm 2022 có một số thay đổi. Tỷ lệ dân số năm 2021 dao động từ 0.33 đến 9.31, trong khi tỷ lệ dân số năm 2022 dao động từ 0.33 đến 9.44.

Tỷ lệ dân số tăng trưởng trung bình là khoảng 1.02% từ năm 2021 đến năm 2022. Có 27 tỉnh có tỷ lệ dân số tăng trưởng dương, trong đó có các tỉnh như An Giang, Đồng Nai, Bà Rịa - Vũng Tàu, Long An và Tiền Giang có tỷ lệ dân số tăng trưởng cao nhất.

Tương ứng với tỷ lệ dân số tăng trưởng thấp hơn là các tỉnh như Ninh Thuận, Khánh Hòa, Đắk Lắk và Gia Lai.

Mô hình phân tán của tỷ lệ dân số năm 2022 so với năm 2021 cho thấy sự phân bố không đều về tỷ lệ dân số tăng trưởng giữa các tỉnh.

Tóm lại, tỷ lệ dân số năm 2022 so với năm 2021 có một số thay đổi. Tỷ lệ dân số tăng trưởng trung bình là khoảng 1.02%, và có một số tỉnh có tỷ lệ dân số tăng trưởng cao hơn hoặc thấp hơn so với trung bình.

1.1 Tiến hành nhập dữ liệu:

-Sử dụng câu lệnh:

```
> nam2021 <- c(1.94, 1.19, 0.93, 1.90, 0.33, 1.49, 1.32, 2.64, 1.53, 1.04, 1.27, 1.23, 0.55, 1.27, 1.21, 1.94, 0.67, 0.63, 3.22, 1.63, 1.59, 0.90, 0.89, 8.46, 1.33, 1.97, 2.10, 0.74, 0.88, 1.30, 1.78, 0.58, 1.27, 0.49, 0.81, 0.77, 1.34, 1.75, 1.86, 1.02, 0.61, 3.46, 1.53, 0.89, 0.92, 1.54, 1.37, 1.26, 0.66, 1.23, 1.31, 1.20, 1.81, 9.31, 0.81, 1.90, 1.34, 3.77, 1.17, 1.03, 1.04, 1.21, 0.86)
```

```
> nam2022 <- c(1.92, 1.18, 0.93, 1.90, 0.33, 1.50, 1.30, 2.78, 1.51, 1.04, 1.26, 1.21, 0.55, 1.26, 1.23, 1.93, 0.67, 0.64, 3.27, 1.61, 1.60, 0.90, 0.88, 8.48, 1.33, 1.96, 2.10, 0.73, 0.88, 1.30, 1.76, 0.58, 1.26, 0.48, 0.81, 0.77, 1.34, 1.74, 1.89, 1.02, 0.60, 3.44, 1.52, 0.88, 0.92, 1.53, 1.37, 1.25, 0.65, 1.20, 1.31, 1.20, 1.79, 9.44, 0.81, 1.89, 1.34, 3.74, 1.17, 1.02, 1.03, 1.20, 0.85)
```

1.2 Tiến hành thống kê:

1.2.1. Bảng phân phối tần số, tần suất:

Bảng tần số:

- Năm 2021:

```
> bts2021<-table(nam2021)
> print(bts2021)
```

+Kết quả:

```
nam2021
0.33 0.49 0.55 0.58 0.61 0.63 0.66 0.67 0.74 0.77 0.81 0.86 0.88 0.89 0.9
 1    1    1    1    1    1    1    1    1    1    2    1    1    2    1
```

```
0.92 0.93 1.02 1.03 1.04 1.17 1.19 1.2 1.21 1.23 1.26 1.27 1.3 1.31 1.32 1.33 1.34 1.37 1.49 1.53 1.54 1.59
1 1 1 1 2 1 1 1 2 2 1 3 1 1 1 1 2 1 1 2 1 1

1.63 1.75 1.78 1.81 1.86 1.9 1.94 1.97 2.1 2.64 3.22 3.46 3.77 8.46 9.31
1 1 1 1 1 2 2 1 1 1 1 1 1 1 1
```

- Năm 2022:

```
> bts2022<-table(nam2022)
> print(bts2022)
```

+ Kết quả:

```
nam2022
0.33 0.48 0.55 0.58 0.6 0.64 0.65 0.67 0.73 0.77 0.81 0.85 0.88 0.9 0.92
1 1 1 1 1 1 1 1 1 1 2 1 3 1 1

0.93 1.02 1.03 1.04 1.17 1.18 1.2 1.21 1.23 1.25 1.26 1.3 1.31 1.33 1.34 1.37 1.5 1.51 1.52 1.53 1.6 1.61
1 2 1 1 1 1 3 1 1 1 3 2 1 1 2 1 1 1 1 1 1 1

1.74 1.76 1.79 1.89 1.9 1.92 1.93 1.96 2.1 2.78 3.27 3.44 3.74 8.48 9.44
1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
```

Bảng tần suất:

- Năm 2021:

```
> prop.table(table(nam2021))
```

+ Kết quả:

```
nam2021
0.33 0.49 0.55 0.58 0.61 0.63 0.66 0.67
0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302
0.74 0.77 0.81 0.86 0.88 0.89 0.9 0.92
0.01587302 0.01587302 0.03174603 0.01587302 0.01587302 0.03174603 0.01587302 0.01587302
0.93 1.02 1.03 1.04 1.17 1.19 1.2 1.21
0.01587302 0.01587302 0.01587302 0.03174603 0.01587302 0.01587302 0.01587302 0.03174603
1.23 1.26 1.27 1.3 1.31 1.32 1.33 1.34
0.03174603 0.01587302 0.04761905 0.01587302 0.01587302 0.01587302 0.01587302 0.03174603
1.37 1.49 1.53 1.54 1.59 1.63 1.75 1.78
0.01587302 0.01587302 0.03174603 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302
1.81 1.86 1.9 1.94 1.97 2.1 2.64 3.22
0.01587302 0.01587302 0.03174603 0.03174603 0.01587302 0.01587302 0.01587302 0.01587302
3.46 3.77 8.46 9.31
0.01587302 0.01587302 0.01587302 0.01587302
```

- Năm 2022:

```
> prop.table(table(nam2022))
```

+ Kết quả:

```
nam2022
0.33      0.48      0.55      0.58      0.6      0.64      0.65      0.67
0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302
0.73      0.77      0.81      0.85      0.88      0.9      0.92      0.93
0.01587302 0.01587302 0.03174603 0.01587302 0.04761905 0.01587302 0.01587302 0.01587302
1.02      1.03      1.04      1.17      1.18      1.2      1.21      1.23
0.03174603 0.01587302 0.01587302 0.01587302 0.01587302 0.04761905 0.01587302 0.01587302
1.25      1.26      1.3      1.31      1.33      1.34      1.37      1.5
0.01587302 0.04761905 0.03174603 0.01587302 0.01587302 0.03174603 0.01587302 0.01587302
1.51      1.52      1.53      1.6      1.61      1.74      1.76      1.79
0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302
1.89      1.9      1.92      1.93      1.96      2.1      2.78      3.27
0.03174603 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302 0.01587302
3.44      3.74      8.48      9.44
0.01587302 0.01587302 0.01587302 0.01587302
```

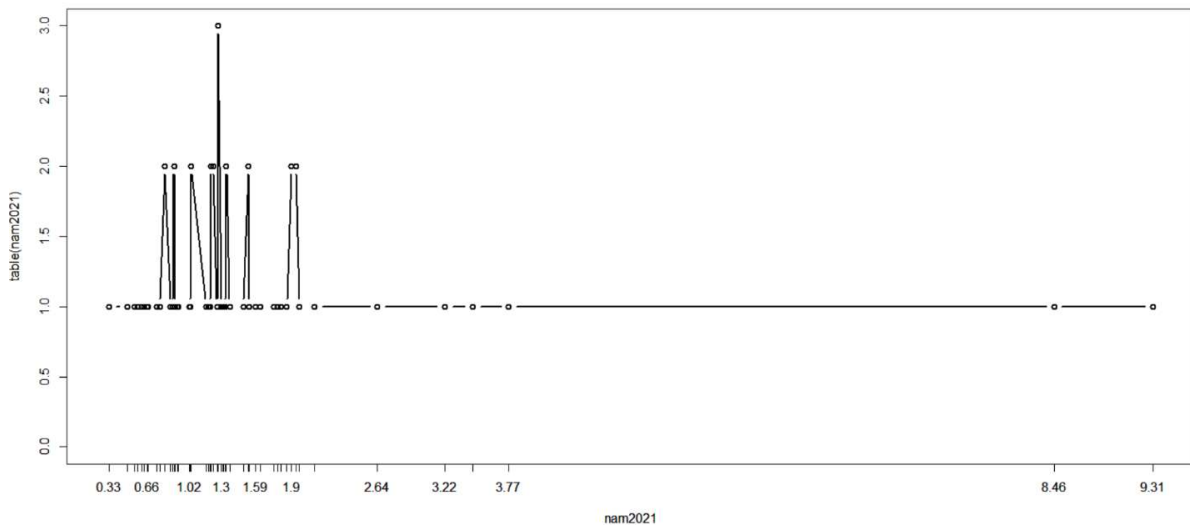
1.2.2. Đa giác tần số:

* Vẽ đa giác tần số:

- Năm 2021:

```
> plot(table(nam2021),type="b")
```

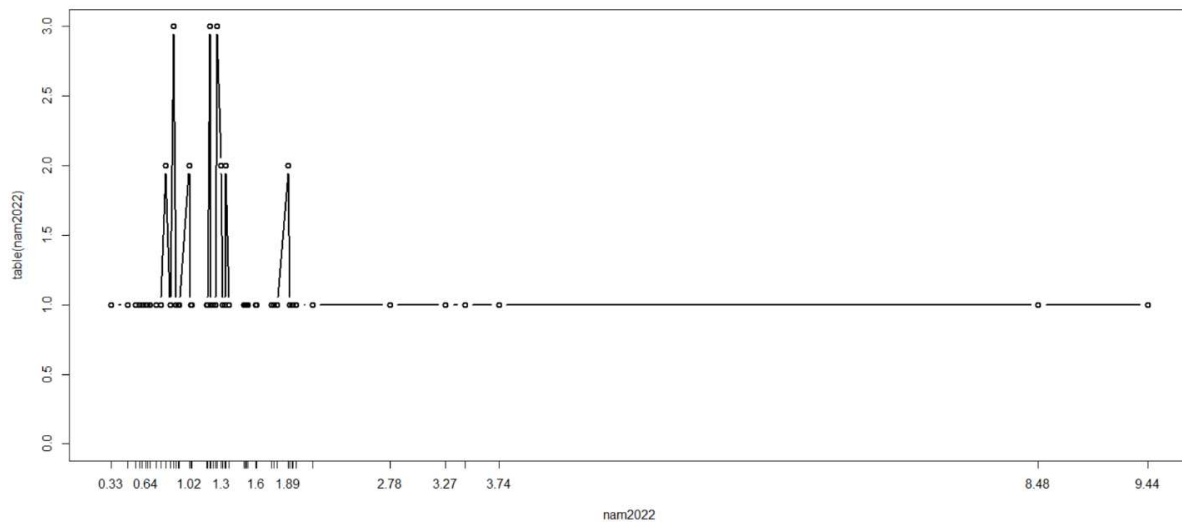
+ Kết quả:



- Năm 2022:

```
> plot(table(nam2022),type="b")
```

+Kết quả:

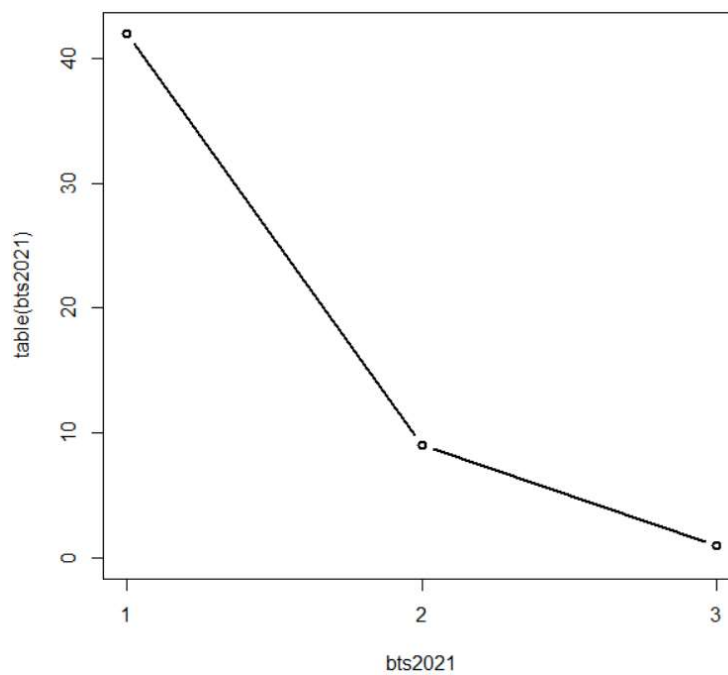


* Vẽ đa giác tần số theo khoảng 0,5:

- Năm 2021:

```
> plot(table(bts2021),type = "b")
```

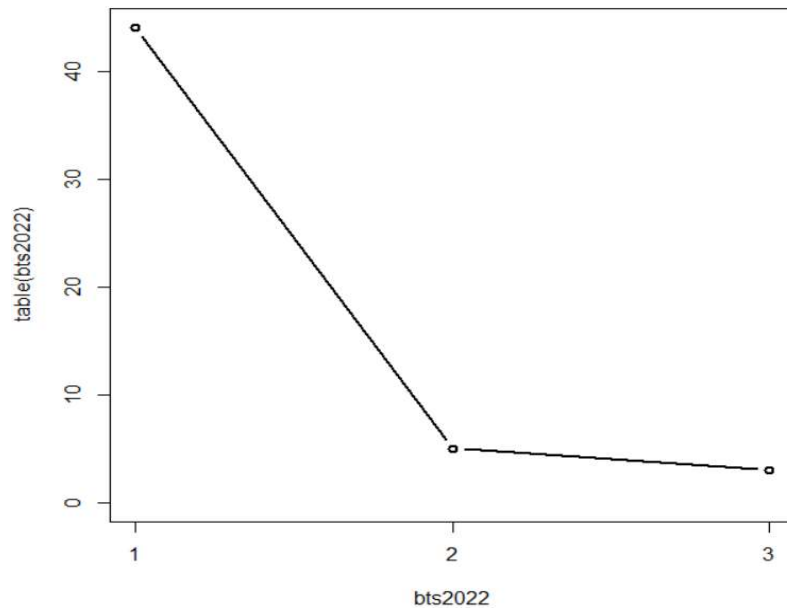
+ Kết quả:



- Năm 2022:

```
> plot(table(bts2022),type = "b")
```

+ Kết quả:



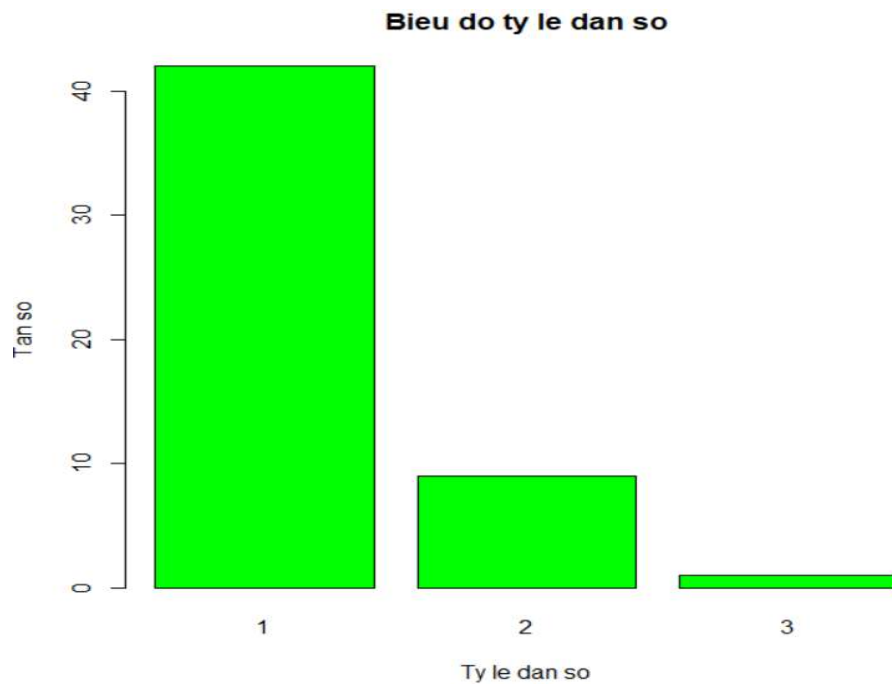
1.2.3. Các dạng biểu đồ:

* Biểu đồ cột của tần số:

- Năm 2021:

```
> barplot(table(bts2021),col="green",border = "black",main = "Bieu do ty le dan so",xlab = "ty le dan so",ylab = "Tan so")
```

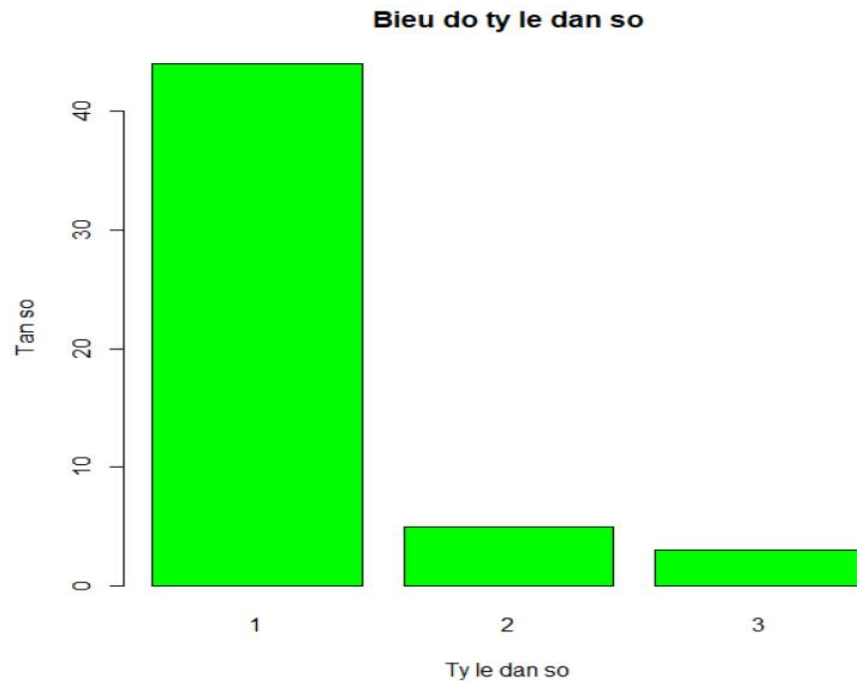
- Kết quả:



- Năm 2022:

```
> barplot(table(bts2022),col="green",border = "black",main = "Bieu do ty le dan so",xlab = "ty le dan so",ylab = "Tan so")
```

- Kết quả:



*Biểu đồ Histogram:

- Tính số khoảng chia K theo công thức Sturge:

+Năm 2021:

```
> so_gia_tri1 <- length(nam2021)
> k1 <- 1+log(so_gia_tri1, base=2)
> print(k1)
[1] 6.97728
```

+Năm 2022

```
> so_gia_tri2 <- length(nam2022)
> k2 <- 1+log(so_gia_tri2, base=2)
> print(k2)
[1] 6.97728
```

*Phân tích câu lệnh :

- `so_gia_tri1 <- length(nam2021)`: Đếm số lượng giá trị trong vector `nam2021` và lưu vào biến `so_gia_tri1`.

- `k1 <- 1 + log(so_gia_tri, base=2)`: Tính số lượng khoảng (bins) dựa trên số lượng giá trị của `nam2021`. Công thức này sử dụng công thức Sturges để ước lượng số lượng bins 15 trong histogram. Cụ thể, số lượng bins được tính bằng $1 + \log_2(n)$, trong đó n là số lượng giá trị. Kết quả được lưu vào biến `k1`.

- `print(k1)`: In ra số lượng bins đã tính được

+Năm 2021:

```
> do_rong_2021 <- (max(nam2021)- min(nam2021))/k1
```

```
> print(do_rong_2021)
```

```
[1] 1.287035
```

+Năm 2022:

```
> do_rong_2022 <- (max(nam2022)- min(nam2022))/k2
```

```
> print(do_rong_2022)
```

```
[1] 1.305666
```

*Phân tích câu lệnh :

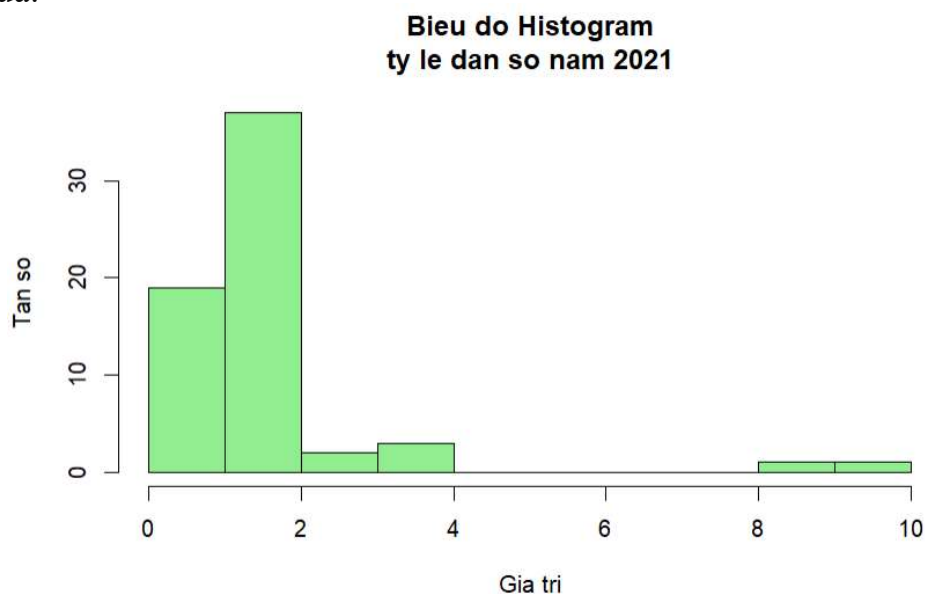
- `do_rong_2021 <- (max(nam2021) - min(nam2021))/k1`: Độ rộng của mỗi khoảng (bins) được tính bằng cách lấy hiệu của giá trị lớn nhất và giá trị nhỏ nhất trong dữ liệu `nam2021`, sau đó chia cho số lượng khoảng `k1`. Kết quả được lưu vào biến `do_rong_2021`.

- `print(do_rong_2021)`: In ra giá trị độ rộng của mỗi khoảng.

- Vẽ biểu đồ Histogram cho năm 2021:

```
> hist(nam2021,breaks = 10,col = "light green",main="Bieu do Histogram  
ty le dan so nam 2021",xlab = "Gia tri",ylab = "Tan so")
```

+Kết quả:



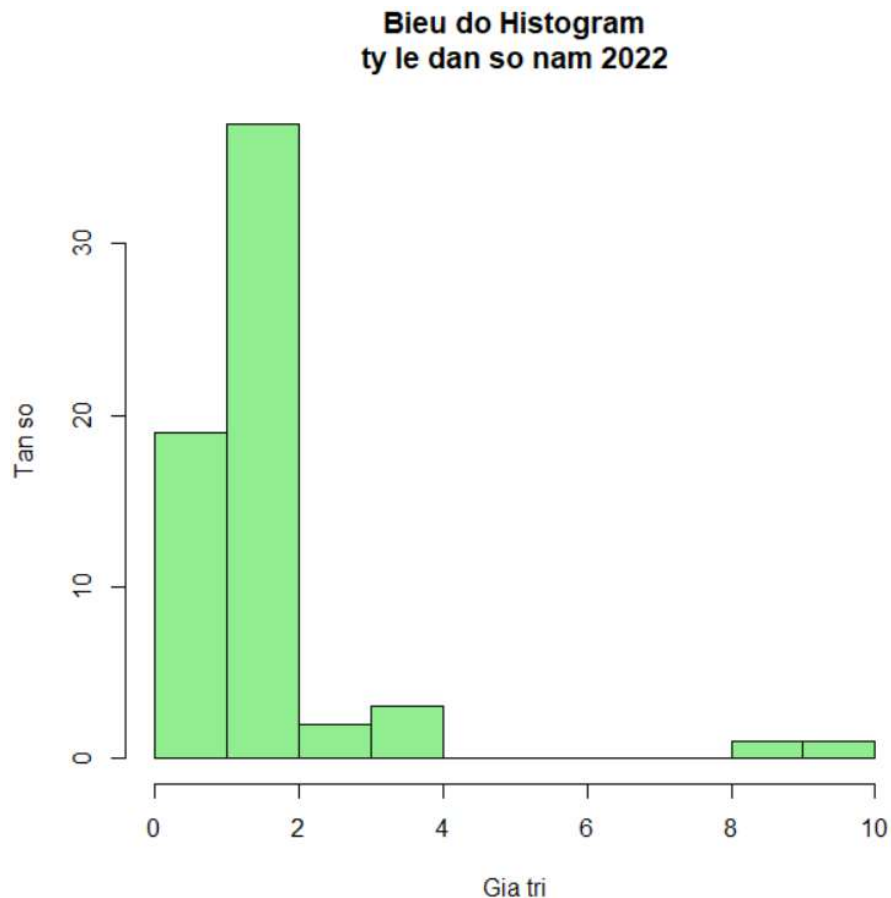
*Phân tích câu lệnh :

- nam2021: Dữ liệu bạn muốn vẽ histogram, trong trường hợp này là tỷ lệ dân số của năm 2021.
- col = "light green": Màu sắc của các cột trong histogram, ở đây là màu xanh lá nhạt
- main = "Bieu do Histogram ty le dan so nam2021": Tiêu đề của biểu đồ, giải thích nội dung của biểu đồ.
- xlab = "Gia tri": Nhãn trục hoành, giải thích ý nghĩa của trục hoành, trong trường hợp này là giá trị của dữ liệu.
- ylab = "Tan so": Nhãn trục tung, giải thích ý nghĩa của trục tung, trong trường hợp này là tần số (số lần xuất hiện của mỗi giá trị).
- breaks = 10 : Số lượng bins được chọn là 10

- Vẽ biểu đồ Histogram cho năm 2022:

```
> hist(nam2022,breaks = 10,col = "light green",main="Bieu do Histogram  
ty le dan so nam 2022",xlab = "Gia tri",ylab = "Tan so")
```

- Kết quả:



*Phân tích câu lệnh :

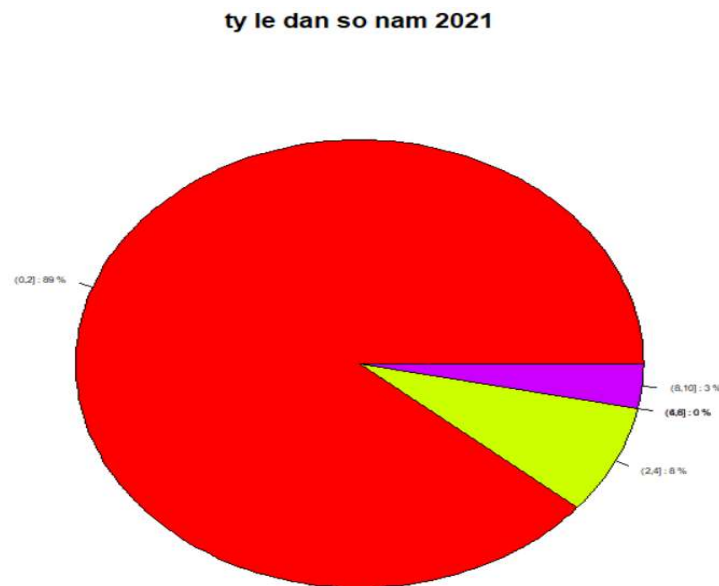
- nam2022: Dữ liệu bạn muốn vẽ histogram, trong trường hợp này là tỷ lệ dân số của năm 2022.
- col = "light green": Màu sắc của các cột trong histogram, ở đây là màu xanh
- main = "Biểu đồ Histogram tỷ lệ dân số của nam2022": Tiêu đề của biểu đồ, giải thích nội dung của biểu đồ.
- xlab = "Giá trị": Nhãn trục hoành, giải thích ý nghĩa của trục hoành, trong trường hợp này là giá trị của dữ liệu.
- ylab = "Tần số": Nhãn trục tung, giải thích ý nghĩa của trục tung, trong trường hợp này là tần số (số lần xuất hiện của mỗi giá trị).
- breaks = 10 : Số lượng bins được chọn là 10

* **Biểu đồ tròn:**

- Năm 2021:

```
>chiakhoang_2<-c("(0,2]", "(2,4]", "(4,6]", "(6,8]", "(8,10]")
>bangtanso_khoang_2_2021<-c(56,5,0,0,2)
>tile<-round(prop.table(bangtanso_khoang_2_2021),2)*100
>pie(bangtanso_khoang_2_2021,labels = paste(chiakhoang_2,":",tile,"%"),col =
rainbow(5),lty=1,main = "tỷ lệ dân số nam 2021",sub="Số liệu nam 2021",cex=0.5)
```

+Kết quả:



Số liệu nam 2021

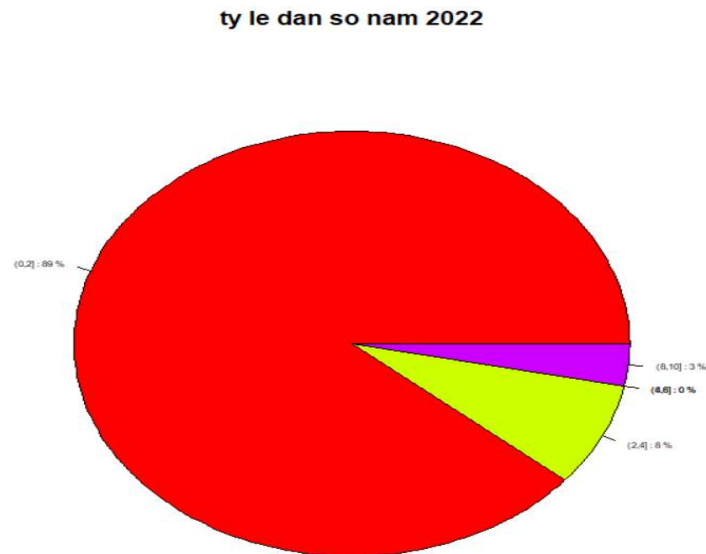
*Phân tích câu lệnh :

- `chiakhoang_2` là một vector chứa các khoảng giá trị.
- `bangtanso_khoang_2_2021` cho biết số phần tử nam2021 trong từng khoảng của `chiakhoang_2`
- Tile tính toán tỷ lệ phần trăm của mỗi khoảng giá trị.
- `pie(bangtanso_khoang_2_2021, ...)` tạo biểu đồ tròn với các thông số sau:
 - `labels = paste(chiakhoang_2, ":", Tile, "%")` thiết lập nhãn cho các phần, nó sẽ hiển thị tên của khoảng giá trị cùng với tỷ lệ phần trăm.
 - `col = rainbow(4)` sử dụng bảng màu cầu vồng với 4 màu để màu sắc của các phần.
 - `lty = 1` thiết lập kiểu đường là liền.
 - `main = "Ty le dan so nam 2021"` là tiêu đề của biểu đồ.
 - `sub = "So lieu nam 2021"` là phụ chú của biểu đồ.
 - `cex = 0.5` là kích thước của các nhãn, trong đây là 0.5 lần kích thước mặc định.

-Năm 2022:

```
> bangtanso_khoang_2_2022<-c(56,5,0,0,2)
> tile<-round(prop.table(bangtanso_khoang_2_2022),2)*100
> pie(bangtanso_khoang_2_2021,labels = paste(chiakhoang_2,":",tile,"%"),col
= rainbow(5),lty=1,main="ty le dan so nam 2022",sub="so lieu nam 2022",cex=0.5)
```

+Kết quả:



So lieu nam 2022

*Phân tích câu lệnh :

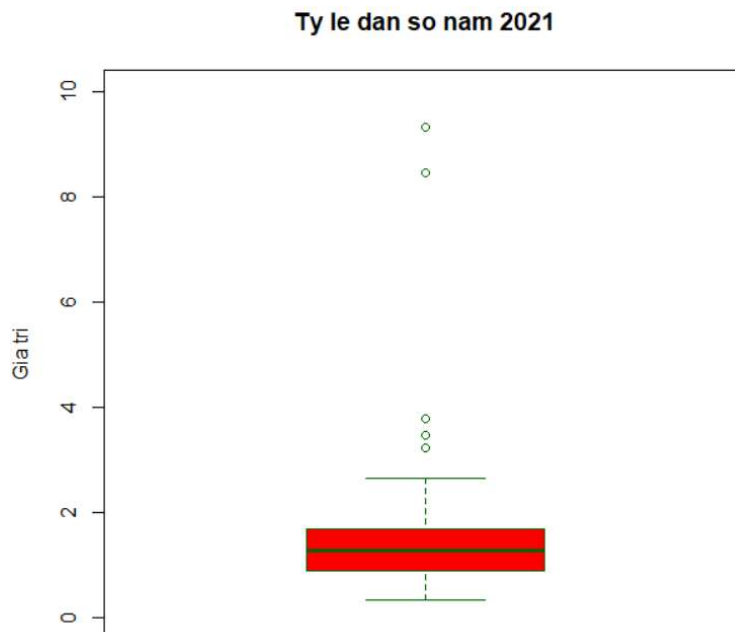
- bangtanso_khoang_2_2022 cho biết số phần tử nam2022 trong từng khoảng của chiakhoang_2
- Tile tính toán tỷ lệ phần trăm của mỗi khoảng giá trị.
- pie(bangtanso_khoang_2_2022, ...) tạo biểu đồ tròn với các thông số sau:
- labels = paste(chiakhoang_2, ":", Tile, "%") thiết lập nhãn cho các phần, nó sẽ hiển thị tên của khoảng giá trị cùng với tỷ lệ phần trăm.
- col = rainbow(5) sử dụng bảng màu cầu vồng với 5 màu để màu sắc của các phần.
- lty = 1 thiết lập kiểu đường là liền.
- main = "Ty le dan so nam 2022" là tiêu đề của biểu đồ.
- sub = "So lieu nam 2022" là phụ chú của biểu đồ.
- cex = 0.5 là kích thước của các nhãn, trong đây là 0.5 lần kích thước mặc định.

* **Biểu đồ hộp và râu:**

- Năm 2021:

```
> boxplot(nam2021,main="Ty le dan so nam 2021",border = "darkgreen",col="red",  
horiz=F,xlab="",ylab="Gia tri",ylim=c(0,10))
```

+Kết quả:



*Phân tích câu lệnh :

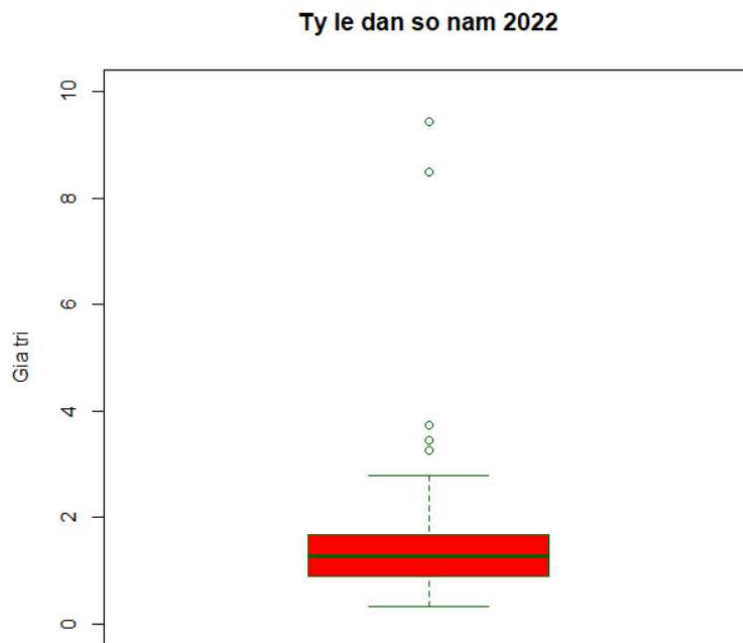
- nam2021: Dữ liệu cần được biểu diễn.

- `main = "Ty le dan so nam 2021"`: Tiêu đề của biểu đồ.
- `border = "darkgreen"`: Màu của đường viền của hộp.
- `col = "red"`: Màu của hộp.
- `horiz = FALSE`: Xác định xem biểu đồ có nằm ngang (TRUE) hay dọc (FALSE). Trong trường hợp này, biểu đồ được vẽ dọc.
- `xlab = ""`: Nhãn trục x.
- `ylab = "Gia tri"`: Nhãn trục y.
- `ylim = c(0, 10)`: Phạm vi của trục y từ 0 đến 10.

-Năm 2022:

```
> boxplot(nam2022,main="Ty le dan so nam 2021",border = "darkgreen",col="red",horiz=F,xlab="",ylab="Gia tri",ylim=c(0,10))
```

+Kết quả:



*Phân tích câu lệnh :

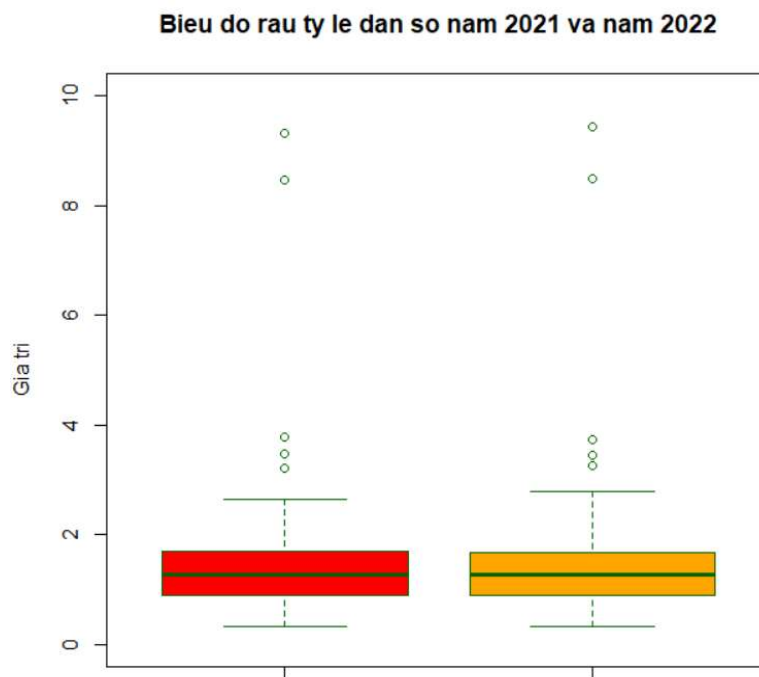
- `nam2022`: Dữ liệu cần được biểu diễn.
- `main = "Ty le dan so nam 2022"`: Tiêu đề của biểu đồ.
- `border = "darkgreen"`: Màu của đường viền của hộp.
- `col = "red"`: Màu của hộp.
- `horiz = FALSE`: Xác định xem biểu đồ có nằm ngang (TRUE) hay dọc (FALSE). Trong trường hợp này, biểu đồ được vẽ dọc.

- xlab = "": Nhãn trục x.
- ylab = "Gia tri": Nhãn trục y.
- ylim = c(0, 10): Phạm vi của trục y từ 0 đến 10.

- Biểu đồ so sánh phân phối tỷ lệ dân số giữa năm 2021 và 2022:

```
> boxplot(nam2021,nam2022,main="Bieu do rau ty le dan so nam 2021 va nam 2022",border="darkgreen",col = c("red","orange"),horiz=F,xlab="",ylab="Gia tri",ylim=c(0,10))
```

+Kết quả:



*Phân tích câu lệnh :

- nam2021, nam2022: Dữ liệu cần được biểu diễn.
- main = "Bieu do rau ty le dan so nam 2021 va nam 2022": Tiêu đề của biểu đồ.
- border = "darkgreen": Màu của đường viền của hộp.
- col = c("red", "orange"): Màu của hộp cho từng năm. Trong trường hợp này, đỏ cho năm 2021 và vàng cho năm 2022.
- horiz = FALSE: Xác định xem biểu đồ có nằm ngang (TRUE) hay dọc (FALSE). Trong trường hợp này, biểu đồ được vẽ dọc.
- xlab = "": Nhãn trục x.
- ylab = "Gia tri": Nhãn trục y.
- ylim = c(0, 10): Phạm vi của trục y từ 0 đến 10.

1.2.4. Các giá trị đặc trưng của biến:

- Năm 2021:

+Trung bình mẫu:

```
> mean(nam2021)
[1] 1.587143
```

+Trung vị của các giá trị:

```
> median(nam2021)
[1] 1.27
```

+Các giá trị mode và vị trí của các giá trị mode này trong table(x):

```
> which(table(nam2021)==max(table(nam2021)))
1.27
27
```

+Giá trị nhỏ nhất, lớn nhất của các giá trị:

```
> range(nam2021)
[1] 0.33 9.31
```

+Phương sai của các giá trị:

```
> var(nam2021)
[1] 2.217805
```

+Độ lệch chuẩn:

```
> sd(nam2021)
[1] 1.48923
```

+Giá trị nhỏ nhất, giá trị lớn nhất, giá trị trung bình, các tứ phân vị của các giá trị:

```
> summary(nam2021)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.330   0.895   1.270   1.587   1.690   9.310
```

Trung vị = 1,270 ; Mode = 1,587; Min = 0,330; Max = 9,310

-Năm 2022:

+Trung bình mẫu:

```
> mean(nam2022)
[1] 1.586984
```

+Trung vị của các giá trị:

```
> median(nam2022)
[1] 1.26
```

+Các giá trị mode và vị trí của các giá trị mode này trong table(x):

```
> which(table(nam2022)==max(table(nam2021)))
```

```
0.88  1.2  1.26
 13   22   26
```

+Giá trị nhỏ nhất, lớn nhất của các giá trị:

```
> range(nam2022)
```

```
[1] 0.33 9.44
```

+Phương sai của các giá trị:

```
> var(nam2022)
```

```
[1] 2.261999
```

+Độ lệch chuẩn:

```
> sd(nam2022)
```

```
[1] 1.503994
```

+Giá trị nhỏ nhất, giá trị lớn nhất, giá trị trung bình, các tứ phân vị của các giá trị:

```
> summary(nam2022)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.330	0.890	1.260	1.587	1.675	9.440

Trung vị =1,260; Mode = 1,587; Min = 0,330; Max = 9,440

CHƯƠNG 2: LÝ THUYẾT ƯỚC LƯỢNG

Mục tiêu:

Sinh viên thực hành được trên phần mềm R các nội dung sau

- Tìm khoảng ước lượng cho giá trị trung bình của tập chính
- Tìm khoảng ước lượng cho giá trị tỉ lệ của tập chính

Để thực hành phần ước lượng, chúng ta cần cài đặt gói lệnh “BSDA” với cấu trúc

```
> install.packages("BSDA")  
> library(BSDA)
```

Chú ý rằng, gói lệnh này chỉ thực hiện ở lần đầu tiên, trong các lần tiếp theo ta chỉ cần gọi thư viện “BSDA”.

3.1 Ước lượng khoảng cho giá trị trung bình

3.1.1 Trường hợp σ được biết

Tỉ lệ dân số phân theo tỉnh/ thành phố năm 2021(nam2021) được giả sử là tuân theo phân phối chuẩn với độ lệch tiêu chuẩn là 0.8 . Người ta thống kê tỷ lệ dân số của năm 2021 và thu được dữ liệu:

1.94, 1.19, 0.93, 1.90, 0.33, 1.49, 1.32, 2.64, 1.53, 1.04, 1.27, 1.23, 0.55, 1.27, 1.21, 1.94, 0.67, 0.63, 3.22, 1.63, 1.59, 0.90, 0.89, 8.46, 1.33, 1.97, 2.10, 0.74, 0.88, 1.30, 1.78, 0.58, 1.27, 0.49, 0.81, 0.77, 1.34, 1.75, 1.86, 1.02, 0.61, 3.46, 1.53, 0.89, 0.92, 1.54, 1.37, 1.26, 0.66, 1.23, 1.31, 1.20, 1.81, 9.31, 0.81, 1.90, 1.34, 3.77, 1.17, 1.03, 1.04, 1.21, 0.86

Tìm khoảng tin cậy 90% cho tỉ lệ dân số trung bình theo tỉnh/thành phố năm 2021.

BÀI LÀM

Tóm tắt bài toán

$$\gamma = 90\%$$

$$\sigma = 0.8$$

$$\text{Thống kê ước lượng : } T = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$$

Thực hiện kiểm định trên R:

```
> z.test(nam2021, sigma.x = 0.8, conf.level = 0.9)
```

- x: véc tơ dữ liệu mẫu
- sigma.x: lệch tiêu chuẩn mẫu
- conf.level: số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.9

Phân tích và kết quả:

One-sample z-Test

```
data:  nam2021
z = 15.747, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 1.421357 1.752929
sample estimates:
mean of x
 1.587143
```

- Kết quả trên cho ta tỉ lệ dân số trung bình của 63 tỉnh/thành $\bar{X} = 1.587143$
- Kết quả về khoảng tin cậy

```
90 percent confidence interval:
 1.421357 1.752929
```

Vậy ta có khoảng tin cậy 90% cho tỉ lệ dân số trung bình của các tỉnh thành là $[1.421357; 1.752929]$

3.1.2. Trường hợp σ chưa biết

Tỉ lệ dân số phân theo tỉnh/ thành phố năm 2021(nam2021) được giả sử là tuân theo phân phối chuẩn với độ lệch tiêu chuẩn chưa biết. Người ta thống kê tỷ lệ dân số của năm 2021 và thu được dữ liệu:

1.94, 1.19, 0.93, 1.90, 0.33, 1.49, 1.32, 2.64, 1.53, 1.04, 1.27, 1.23, 0.55, 1.27, 1.21, 1.94, 0.67, 0.63, 3.22, 1.63, 1.59, 0.90, 0.89, 8.46, 1.33, 1.97, 2.10, 0.74, 0.88, 1.30, 1.78, 0.58, 1.27, 0.49, 0.81, 0.77, 1.34, 1.75, 1.86, 1.02, 0.61, 3.46, 1.53, 0.89, 0.92, 1.54, 1.37, 1.26, 0.66, 1.23, 1.31, 1.20, 1.81, 9.31, 0.81, 1.90, 1.34, 3.77, 1.17, 1.03, 1.04, 1.21, 0.86

Tìm khoảng tin cậy 90% cho tỉ lệ dân số trung bình theo tỉnh/thành phố năm 2021.

BÀI LÀM

Tóm tắt bài toán

$\gamma = 90\%$

σ chưa biết

Thống kê ước lượng : $T = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$

Thực hiện kiểm định

> `t.test(nam2021, conf.level = 0.9)`

`x`

véc tơ dữ liệu mẫu

`conf.level`

số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng,
mặc định là 0.95

Phân tích và kết quả

```
> t.test(nam2021, conf.level = 0.9)

One Sample t-test

data:  nam2021
t = 8.4591, df = 62, p-value =
6.343e-12
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 1.273845 1.900440
sample estimates:
mean of x
 1.587143
```

- Kết quả trên cho ta tỉ lệ dân số trung bình của 63 tỉnh/thành $\bar{x} = 1.587143$

- Kết quả về khoảng tin cậy

```
90 percent confidence interval:
 1.273845 1.900440
```

Vậy ta có khoảng tin cậy 90% cho tỉ lệ dân số trung bình của các tỉnh thành là [1.273845; 1.900440]

3.2. Ước lượng khoảng cho giá trị tỷ lệ

Khảo sát tỉ lệ dân số gồm 63 tỉnh thành, thấy có 24 tỉnh thành có tỉ lệ thấp hơn so với tỉ lệ dân số trung bình. Hãy ước chừng khoảng tin cậy 90% cho tỷ lệ của tỉ lệ dân số thấp hơn so với tỉ lệ dân số trung bình.

BÀI LÀM

Tóm tắt bài toán

$\gamma = 90\%$

$n = 63$

$x = 24$

$\text{conf. level} = 0,9$

$\text{correct} = \text{FALSE}$ (vì $n \cdot f = 24 > 10$; $n(1 - f) = 39 > 10$ thỏa mãn).

Thống kê ước lượng:
$$T = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}$$

Thực hiện kiểm định

```
> prop.test(x=24, n=63, conf.level = 0.9, correct = F)
```

x	số lần cá thể có dấu hiệu T trong mẫu quan sát
n	kích thước mẫu quan sát
conf.level	số thuộc [0,1] chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95
correct	tham số dạng logic chỉ xem có hay không sự điều chỉnh liên tục Yates. Với $f = \frac{x}{n}$, nếu điều kiện $nf \geq 10$; $n(1-f) \geq 10$ thỏa mãn thì là <code>correct = FALSE</code> , trái lại <code>correct = TRUE</code> (mặc định <code>correct = TRUE</code>)

Phân tích và kết quả

```
> prop.test(x=24, n=63, conf.level = 0.9, correct = F)
```

1-sample proportions test without continuity correction

```
data: 24 out of 63, null probability 0.5
X-squared = 3.5714, df = 1, p-value = 0.05878
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.2871900 0.4845187
sample estimates:
      p
0.3809524
```

Vậy khoảng tin cậy 90% cho tỉ lệ tổng thể những ổ trục có bề mặt thô hơn thông số kỹ thuật là [0.2871900, 0.4845187].

CHƯƠNG 3: KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

Mục tiêu:

Sinh viên thực hành được trên phần mềm R các nội dung sau

- Thực hiện về kiểm định giá trị trung bình
- Thực hiện về kiểm định giá trị tỉ lệ

Để thực hành phần kiểm định, chúng ta cần cài đặt gói lệnh “BSDA” và “readxl”

```
>install.packages("BSDA")  
>install.packages("readxl")
```

3.1. Kiểm định về giá trị trung bình:

Bài toán 1: Đã biết σ

Tỷ lệ dân số trong phân theo tỉnh/ thành phố năm 2021(nam2021) được giả sử là tuân theo phân phối chuẩn với độ lệch tiêu chuẩn là 0.8 . Người ta thống kê tỷ lệ dân số của năm 2021 và thu được dữ liệu:

1.94, 1.19, 0.93, 1.90, 0.33, 1.49, 1.32, 2.64, 1.53, 1.04, 1.27, 1.23, 0.55, 1.27, 1.21, 1.94, 0.67, 0.63, 3.22, 1.63, 1.59, 0.90, 0.89, 8.46, 1.33, 1.97, 2.10, 0.74, 0.88, 1.30, 1.78, 0.58, 1.27, 0.49, 0.81, 0.77, 1.34, 1.75, 1.86, 1.02, 0.61, 3.46, 1.53, 0.89, 0.92, 1.54, 1.37, 1.26, 0.66, 1.23, 1.31, 1.20, 1.81, 9.31, 0.81, 1.90, 1.34, 3.77, 1.17, 1.03, 1.04, 1.21, 0.86

Có thể kết luận rằng dân số trung bình năm 2021 lớn hơn 1.75% với mức ý nghĩa 10% hay không?

*** Bài toán kiểm định:**

- Giả thuyết H_0 : $\mu = 1.75\%$
- Đối thuyết H_1 : $\mu > 1.75\%$

*** Thực hiện kiểm định trên R:**

```
z.test(nam2021, alternative ="greater",mu = 1.75,sigma.x=0.8,conf.level=0.9)
```

- alternative: chuỗi kí tự chỉ đối thuyết,ở đây ta dùng lớn hơn, “greater”.
- mu: giá trị trung bình
- sigma.x: độ lệch tiêu chuẩn
- conf.level: độ tin cậy cho khoảng tin cậy đc trả về

* Phân tích kết quả và kết luận:

```
> z.test(nam2021, alternative = "greater", mu = 1.75, sigma.x=0.8, conf.level=0.9)
```

One-sample z-Test

```
data: nam2021
z = -1.6158, p-value = 0.9469
alternative hypothesis: true mean is greater than 1.75
90 percent confidence interval:
 1.457975      NA
sample estimates:
mean of x
 1.587143
```

- Kết quả trên cho ta một số thông tin sau:

+ Giá trị thống kê: $z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} = -1.6158$

+ Trị số p- của bài toán là p-value = 0.9469

+ Tỷ lệ dân số trung bình năm 2021 trong mẫu = 1.587143

→ **Kết luận:** Vì p-value > α nên ta không bác bỏ H₀ thay bằng H₁. Do đó, không thể cho rằng tỷ lệ dân số năm 2021 lớn hơn 1.75% với mức ý nghĩa 10%.

Bài toán 2: Chưa biết σ

Tỷ lệ dân số trong phân theo tỉnh/ thành phố năm 2022(nam2022) được giả sử là tuân theo phân phối chuẩn. Người ta thống kê tỷ lệ dân số của năm 2022 và thu được dữ liệu:

1.92, 1.18, 0.93, 1.90, 0.33, 1.50, 1.30, 2.78, 1.51, 1.04, 1.26, 1.21, 0.55, 1.26, 1.23, 1.93, 0.67, 0.64, 3.27, 1.61, 1.60, 0.90, 0.88, 8.48, 1.33, 1.96, 2.10, 0.73, 0.88, 1.30, 1.76, 0.58, 1.26, 0.48, 0.81, 0.77, 1.34, 1.74, 1.89, 1.02, 0.60, 3.44, 1.52, 0.88, 0.92, 1.53, 1.37, 1.25, 0.65, 1.20, 1.31, 1.20, 1.79, 9.44, 0.81, 1.89, 1.34, 3.74, 1.17, 1.02, 1.03, 1.20, 0.85

Có thể kết luận rằng tỷ lệ dân số trung bình năm 2022 bằng 1.75% với mức ý nghĩa 10% hay không?

* Bài toán kiểm định:

- Giả thuyết H₀ : μ = 1.75%

- Đối thuyết H₁: μ ≠ 1.75%

* Thực hiện kiểm định trên

```
> t.test(nam2022, alternative = "two.sided", mu = 1.75, , conf.level=0.9)
```

* Phân tích kết quả và kết luận:

```
> t.test(nam2022, alternative = "two.sided", mu = 1.75, conf.level=0.9)
```

One Sample t-test

```
data: nam2022
t = -0.86031, df = 62, p-value = 0.3929
alternative hypothesis: true mean is not equal to 1.75
90 percent confidence interval:
 1.270581 1.903388
sample estimates:
mean of x
 1.586984
```

- Kết quả trên cho ta một số thông tin sau:

+ Giá trị thống kê: $z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} = -0.86031$

+ Trị số p- của bài toán là p - value = 0.3929

+ Dân số trung bình năm 2022 trong mẫu = 1.586984

→ **Kết luận:** Vì p-value > α nên ta không bác bỏ H₀ thay bằng H₁. Do đó, không thể cho rằng tỷ lệ dân số năm 2021 không bằng 1.75% với mức ý nghĩa 10%.

3.2. Kiểm định về tỉ lệ:

Bài toán: Có ý kiến cho rằng tỷ lệ dân số trên 2% năm 2022 là 20%. Theo khảo sát của tổng cục thống kê thu được dữ liệu:

1.92, 1.18, 0.93, 1.90, 0.33, 1.50, 1.30, 2.78, 1.51, 1.04, 1.26, 1.21, 0.55, 1.26, 1.23, 1.93, 0.67, 0.64, 3.27, 1.61, 1.60, 0.90, 0.88, 8.48, 1.33, 1.96, 2.10, 0.73, 0.88, 1.30, 1.76, 0.58, 1.26, 0.48, 0.81, 0.77, 1.34, 1.74, 1.89, 1.02, 0.60, 3.44, 1.52, 0.88, 0.92, 1.53, 1.37, 1.25, 0.65, 1.20, 1.31, 1.20, 1.79, 9.44, 0.81, 1.89, 1.34, 3.74, 1.17, 1.02, 1.03, 1.20, 0.85

Nhận thấy có 7 tỉnh/ thành phố và khu vực có dân số trên 2. Có thể kết luận dân số trên 2 năm 2022 nhỏ hơn 20% hay không với mức ý nghĩa 10% ?

* Bài toán kiểm định:

- Giả thuyết H₀ : p=0.2

- Đối thuyết H₁: p < 0.2

* Thực hiện kiểm định trên R:

```
> prop.test(x=7,n=63,p=0.2,alternative="less",conf.level=0.9,correct= FALSE)
```

- x: vector chỉ số lần thành công trong mỗi mẫu
- n: vector chỉ số lần thử nghiệm trong mỗi mẫu
- p: vector chỉ xác suất thành công
- alternative: chuỗi kí tự chỉ đối thuyết, ở đây ta dùng “less” tức nhỏ hơn
- conf.level: độ tin cậy cho khoảng tin cậy được trả về
- correct=FALSE nếu $np_0 \geq 5$, $n(1 - p_0) \geq 5$

* Phân tích kết quả và kết luận:

```
> prop.test(x=7,n=63,p=0.2,alternative = "less",conf.level = 0.9,correct = FALSE)
```

1-sample proportions test without continuity correction

```
data: 7 out of 63, null probability 0.2
X-squared = 3.1111, df = 1, p-value = 0.03888
alternative hypothesis: true p is less than 0.2
90 percent confidence interval:
 0.0000000 0.1720501
sample estimates:
               p
0.1111111
```

Kết quả cho thấy giá trị p-value < 0.2 nên ta bác bỏ H_0 thay bằng H_1

→ **Kết luận:** Tỷ lệ dân số trên 2% năm 2022 nhỏ hơn 20%

3.3. Kiểm định về 2 giá trị trung bình

Bài toán: Theo dữ liệu của tổng cục thống kê về tỷ lệ dân số năm 2021(nam2021) và năm 2022(nam2022) được giả sử là tuân theo phân phối chuẩn với mức ý nghĩa 5% có thể nói rằng tỷ lệ dân số của năm 2021(nam2021) nhỏ hơn năm 2022(nam2022) hay không?

* Bài toán kiểm định:

- Giả thuyết H_0 : $\mu_1 = \mu_2$
- Đối thuyết H_1 : $\mu_1 < \mu_2$

* Thực hiện kiểm định trên R

```
> t.test(nam2021,nam2022,alternative = "less",conf.level = 0.95)
```

*** Phân tích kết quả và kết luận:**

```
> t.test(nam2021,nam2022,alternative = "less",conf.level = 0.95)
```

Welch Two Sample t-test

```
data: nam2021 and nam2022
t = 0.00059525, df = 123.99, p-value =
0.5002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 0.4420787
sample estimates:
mean of x mean of y
 1.587143  1.586984
```

- Kết quả cho thấy giá trị $p\text{-value} > 0.05$ nên ta chưa bác bỏ H_0

→ **Kết luận:** Chưa có cơ sở nói rằng tỷ lệ dân số năm 2021(nam2021) nhỏ hơn năm 2022(nam2022)

3.4. Kiểm định về 2 tỉ lệ:

Bài toán: Theo dữ liệu của tổng cục thống kê dân số năm 2021(nam2021) và năm 2022(nam2022) ta có bảng sau:

Năm	Số tỉnh/thành phố và khu vực	Số tỉnh/thành phố và khu vực có tỷ lệ dân số trên 1.93%
2021	63	10
2022	63	8

Với mức ý nghĩa $\alpha = 5\%$ có thể nói rằng tỉ lệ số tỉnh/ thành phố và khu vực có tỷ lệ dân số trên 1.93 của năm 2021 lớn hơn của năm 2022 hay không?

*** Bài toán kiểm định:**

- Giả thuyết $H_0 : p_1 = p_2$

- Đối thuyết $H_1 : p_1 > p_2$

*** Thực hiện kiểm định trên R:**

```
prop.test(c(10,8),c(63,63),alternative = "greater",conf.level = 0.95)
x=c(10,8),n=c(63,63)
```

*** Phân tích kết quả và kết luận:**

```
> prop.test(c(10,8),c(63,63),alternative = "greater",conf.level = 0.95)
```

2-sample test for equality of proportions with continuity correction

```
data: c(10, 8) out of c(63, 63)
X-squared = 0.064815, df = 1, p-value = 0.3995
alternative hypothesis: greater
95 percent confidence interval:
 -0.0865747  1.0000000
sample estimates:
 prop 1    prop 2 
0.1587302 0.1269841
```

- Kết quả cho thấy giá trị $p\text{-value} > 0.05$ nên ta chưa bác bỏ H_0

→ **Kết luận:** Vậy *chưa có cơ sở kết luận* tỉ lệ số tỉnh/ thành phố và khu vực có tỷ lệ dân số trên 1.93% của năm 2021 lớn hơn

Chương 4: HỒI QUY VÀ TƯƠNG QUAN

Bài toán: Sử dụng dữ liệu về tỉ lệ dân số biến động ở mỗi tỉnh/thành Việt Nam trong năm 2021 và năm 2022.

```
> nam2021<-c(1.94, 1.19, 0.93, 1.90, 0.33, 1.49, 1.32, 2.64, 1.53, 1.04, 1.27, 1.23, 0.55, 1.27, 1.21, 1.94, 0.67, 0.63, 3.22, 1.63, 1.59, 0.90, 0.89, 8.46, 1.33, 1.97, 2.10, 0.74, 0.88, 1.30, 1.78, 0.58, 1.27, 0.49, 0.81, 0.77, 1.34, 1.75, 1.86, 1.02, 0.61, 3.46, 1.53, 0.89, 0.92, 1.54, 1.37, 1.26, 0.66, 1.23, 1.31, 1.20, 1.81, 9.31, 0.81, 1.90, 1.34, 3.77, 1.17, 1.03, 1.04, 1.21, 0.86)
```

```
> nam2022<-c(1.92, 1.18, 0.93, 1.90, 0.33, 1.50, 1.30, 2.78, 1.51, 1.04, 1.26, 1.21, 0.55, 1.26, 1.23, 1.93, 0.67, 0.64, 3.27, 1.61, 1.60, 0.90, 0.88, 8.48, 1.33, 1.96, 2.10, 0.73, 0.88, 1.30, 1.76, 0.58, 1.26, 0.48, 0.81, 0.77, 1.34, 1.74, 1.89, 1.02, 0.60, 3.44, 1.52, 0.88, 0.92, 1.53, 1.37, 1.25, 0.65, 1.20, 1.31, 1.20, 1.79, 9.44, 0.81, 1.89, 1.34, 3.74, 1.17, 1.02, 1.03, 1.20, 0.85)
```

- Tìm hệ số tương quan:

```
> cor(nam2021, nam2022)
[1] 0.9998755
```

- Tìm hàm quy hồi tuyến tính của năm 2022 theo năm 2021

```
> reg <- lm(nam2022~nam2021)
> reg
Call:
lm(formula = nam2022 ~ nam2021)

Coefficients:
(Intercept)      nam2021 
   -0.01569        1.00979
```

→ $y = -0.01569 + 1.00979x$ → Hàm hồi quy tuyến tính

*Phân tích câu lệnh :

- `lm`: Đây là hàm trong R để thực hiện mô hình hồi quy tuyến tính
- `nam2022~nam2021`: diễn đạt mối quan hệ giữa biến phụ thuộc và biến độc lập. Trong trường hợp này, `nam2021` được sử dụng làm biến độc lập để dự đoán `nam2022`. Dấu `~` chỉ ra mối quan hệ giữa các biến
- `reg <-`: Đây là phần gán kết quả của hàm `lm` vào một đối tượng mới được đặt tên là `reg`. Bạn có thể đặt tên cho đối tượng kết quả bất kỳ.
- Kết quả tính toán của `lm` cho thấy $B_0 = -0.01569$ và $B_1 = 1.00979$. Nói cách khác, với hai thông số này, chúng ta có thể ước tính tỷ lệ dân số thành phố năm 2022 cho bất cứ tỉnh/ địa phương nào trong khoảng tỉnh/ địa phương của mẫu bằng phương trình tuyến tính :

* Tìm khoảng tin cậy:

- Tìm khoảng tin cậy 70% cho nam2022 với nam2021= 15%

```
> predict(reg, newdata = data.frame(nam2021 = 15), interval = "prediction", level = 0.7)
      fit      lwr      upr
1 15.13113 15.093 15.16927
```

+ Giá trị dự đoán cho biến phụ thuộc nam2022 là 15.13113

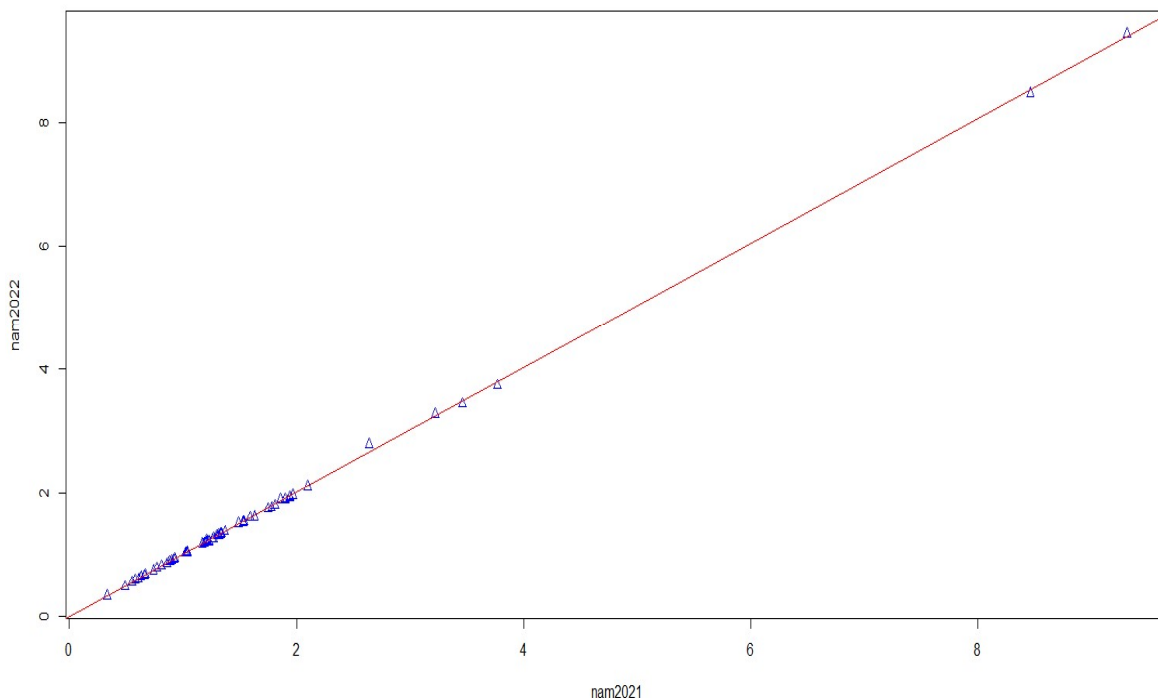
+ Giới hạn dưới của khoảng dự đoán với mức tin cậy là 70% là 15.093

+ Giới hạn trên của khoảng dự đoán với mức tin cậy là 70% là 15.16927

- Vẽ biểu đồ phân tán đường hồi quy:

```
> plot(nam2022 ~ nam2021, pch=2, col='blue')
```

```
> abline(reg,col='red',lwd=1)
```



Phân tích câu lệnh trên

```
> plot(nam2022 ~ nam2021, pch=2, col='blue')
```

```
> abline(reg,col='red',lwd=1)
```

- pch=2: Đặc điểm của điểm dữ liệu trên biểu đồ. Trong trường hợp này, pch=2 tạo ra các điểm hình tam giác.

- col='blue': Màu sắc của điểm dữ liệu trên biểu đồ. Trong trường hợp này, chúng ta sử dụng màu xanh.

- abline(reg, ...): Thêm đường hồi quy vào biểu đồ.

- reg: Đây là đối tượng mô hình hồi quy tuyến tính đã được ước lượng.
- col='red': Màu của đường hồi quy. Trong trường hợp này, chúng ta sử dụng màu đỏ.
- lwd=9: Độ dày của đường hồi quy. Trong trường hợp này, chúng ta sử dụng độ dày là 9

---Hết---