ECOM20001 Econometrics 1

Tutorial 2 Semester 1, 2022

Chin Quek

Department of Economics

Part 1: Visualising and Describing Data in R

- tute2_crime.csv

```
## Set the working directory for the tutorial file
setwd("your working directory")
```

Dataset

Dataset tute2_crime.csv has the following 5 variables:

- stateid: identifier for a US state
- vio: violent crime rate: incidents per 100,000 people
- rob: robbery rate: incidents per 100,000 people
- density: population per square mile of land
- avginc: real per capita personal income in the state

Reading in data

```
## Load the dataset from a comma separate value
data=read.csv("tute2_crime.csv")

## List the variables in the dataset named data
names(data)
```

```
## [1] "stateid" "vio"     "rob"     "dens"    "avginc"
```

```
## Dimension of the dataset: 45 observations (states), 5 variables
dim(data)
```

```
## [1] 45  5
```

Describing data

1. Discuss the sample means, standard deviations, minimums and maximums for each of the four main variables in the dataset: `vio`, `rob`, `density`, `avginc`.

- What does a "typical" state look like in the dataset? Focus on sample means in describing a typical state.

  Be sure to state the units of a variable to accurately describe what a typical state looks like.

- Discuss the minimum and maximum of each variable, highlighting the range of values that each variable takes on.
- How varied is the degree of violent crimes and robbery rates, population densities, and per capita incomes in the sample? How violent and robbery-filled is the worst state compared to the best state?

```r
## Using the summary() function
summary(data)      # Mean, Min, Max, Median, 25th/75th percentile
```

```
##     stateid         vio             rob              dens            avginc
##  Min.   : 1   Min.   : 66.9   Min.   :  8.8   Min.   :  1.086   Min.   :12.37
##  1st Qu.:12   1st Qu.:275.5   1st Qu.: 75.3   1st Qu.: 34.542   1st Qu.:13.92
##  Median :23   Median :382.8   Median :100.9   Median : 76.529   Median :15.80
##  Mean   :23   Mean   :431.5   Mean   :106.7   Mean   :105.656   Mean   :15.82
##  3rd Qu.:34   3rd Qu.:570.0   3rd Qu.:152.5   3rd Qu.:157.042   3rd Qu.:17.11
##  Max.   :45   Max.   :854.0   Max.   :240.8   Max.   :385.441   Max.   :20.27
```

Alternative R commands for descriptive statistics

```
sapply(data, mean)     # Means
```

```
##   stateid       vio       rob      dens    avginc
##  23.00000 431.48444 106.65556 105.65617  15.81649
```

```
sapply(data, median)  # Median
```

```
##   stateid       vio       rob      dens    avginc
##  23.00000 382.80000 100.90000  76.52950  15.79737
```

```
sapply(data, sd)       # Standard Deviation
```

```
##   stateid       vio       rob      dens    avginc
##  13.13393 209.54125  64.19275  97.66395   1.93695
```

Alternative R commands for descriptive statistics

```
sapply(data, min)      # Min
```

```
##   stateid      vio      rob     dens   avginc
##   1.00000 66.90000  8.80000  1.08610 12.37023
```

```
sapply(data, max)      # Max
```

```
##   stateid      vio      rob     dens   avginc
##   45.0000 854.0000 240.8000 385.4414  20.2728
```

```
sapply(data, quantile)
```

```
##       stateid   vio   rob     dens   avginc
## 0%          1  66.9   8.8   1.0861 12.37023
## 25%        12 275.5  75.3  34.5422 13.91905
## 50%        23 382.8 100.9  76.5295 15.79737
## 75%        34 570.0 152.5 157.0423 17.11416
## 100%       45 854.0 240.8 385.4414 20.27280
```

```r
## Descriptive Statistics: stargazer()
stargazer(data,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          type = "text",
          title = "Descriptive Statistics")
```

```
##
## Descriptive Statistics
## ========================================================
## Statistic N   Mean    St. Dev. Median   Min     Max
## --------------------------------------------------------
## stateid   45 23.000   13.134    23       1      45
## vio       45 431.484 209.541  382.800 66.900 854.000
## rob       45 106.656  64.193  100.900 8.800  240.800
## dens      45 105.656  97.664  76.530  1.086  385.441
## avginc    45 15.816   1.937   15.797  12.370 20.273
## --------------------------------------------------------
```

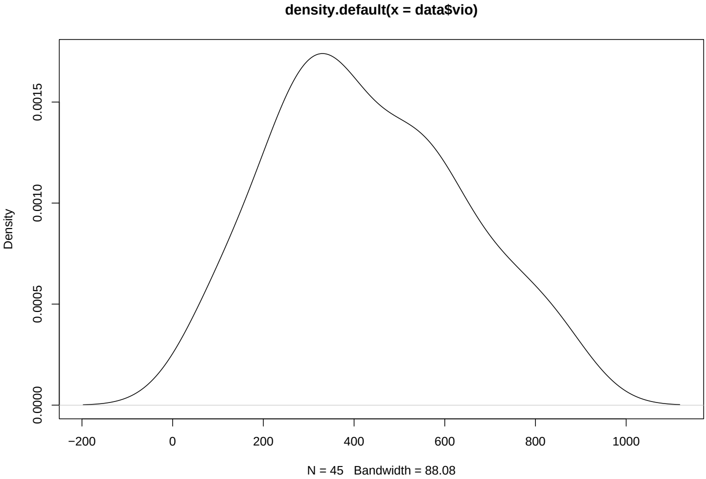
Probability densities

2. How do the respective probability densities of `vio`, `rob`, `density`, `avginc` look?

- Focus on their means, and skewness

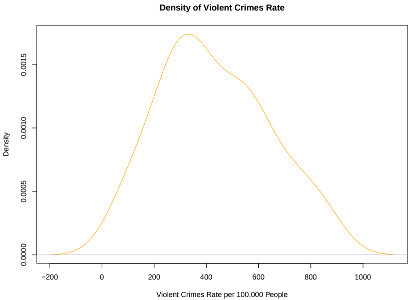
A simple plot of probability density

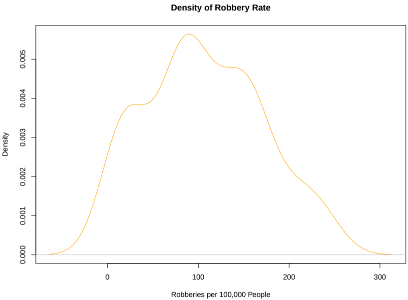
```
plot(density(data$vio))
```

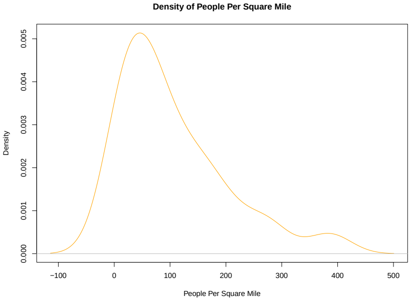


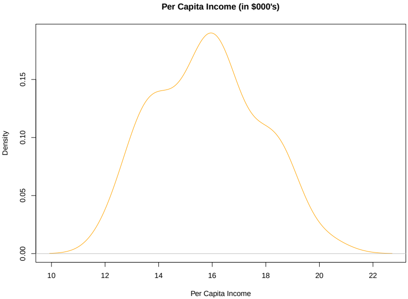
Fancy probability densities

```r
## Create probability densities for all relevant variables
plot(density(data$vio),
     main="Density of Violent Crimes Rate",
     xlab="Violent Crimes Rate per 100,000 People",
     ylab="Density", col="orange")

plot(density(data$rob),
     main="Density of Robbery Rate",
     xlab="Robberies per 100,000 People",
     ylab="Density", col="orange")

plot(density(data$dens),
     main="Density of People Per Square Mile",
     xlab="People Per Square Mile",
     ylab="Density", col="orange")

plot(density(data$avginc),
     main="Per Capita Income (in $000's)",
     xlab="Per Capita Income",
     ylab="Density", col="orange")
```


Density of Violent Crimes Rate
Density
0.0000
0.0005
0.0010
0.0015
−200
0
200
400
600
800
1000
Violent Crimes Rate per 100,000 People


Density of Robbery Rate
Density
0.000
0.001
0.002
0.003
0.004
0.005
0
100
200
300
Robberies per 100,000 People


Density of People Per Square Mile
Density
0.000
0.001
0.002
0.003
0.004
0.005
−100
0
100
200
300
400
500
People Per Square Mile


Per Capita Income (in $000's)
Density
0.00
0.05
0.10
0.15
10
12
14
16
18
20
22
Per Capita Income

Describing relationship between two variables

3. Comment on the 3 scatter plots below

Visually, does a relationship appear exist in each graph? If so, offer an **economic explanation** for why the relationship might exist.

There may be multiple explanations, so you may offer various explanations if you wish. But just one explanation is fine.

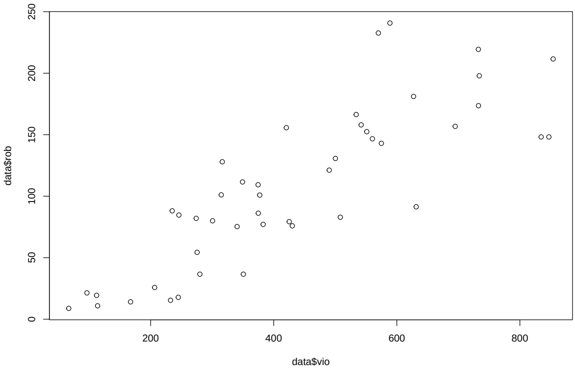
- Robbery vs Violence
- Robbery vs Per Capita Income
- Robbery vs People per Square Mile

Economic explanations

Economic explanations focus on the costs and benefits of a particular behaviour for explaining empirical patterns.

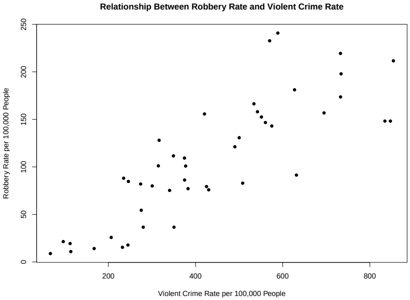
A simple scatter plot

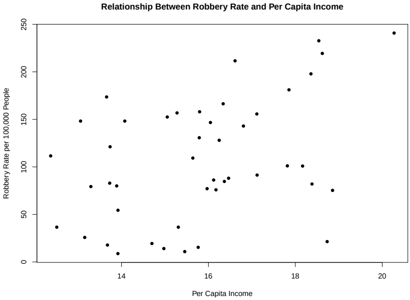
```r
## Create scatter plot for vio (x-variable) vs rob (y-variable)
plot(data$vio,data$rob)
```

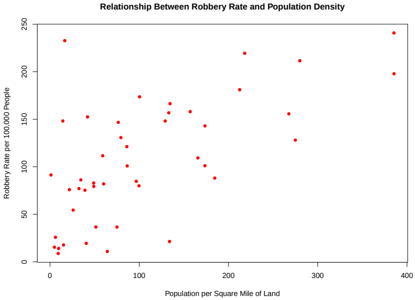


Fancy scatter plots

```
plot(data$vio,data$rob,
     main="Relationship Between Robbery Rate and Violent Crime Rate",
     xlab="Violent Crime Rate per 100,000 People",
     ylab="Robbery Rate per 100,000 People",
     pch=16)

plot(data$avginc,data$rob,
     main="Relationship Between Robbery Rate and Per Capita Income",
     xlab="Per Capita Income",
     ylab="Robbery Rate per 100,000 People",
     pch=16)

plot(data$dens,data$rob,
     main="Relationship Between Robbery Rate and Population Density",
     xlab="Population per Square Mile of Land",
     ylab="Robbery Rate per 100,000 People",
     col="red",
     pch=16)
```


Relationship Between Robbery Rate and Violent Crime Rate
Robbery Rate per 100,000 People
0
50
100
150
200
250
200
400
600
800
Violent Crime Rate per 100,000 People


Relationship Between Robbery Rate and Per Capita Income
Robbery Rate per 100,000 People
0
50
100
150
200
250
14
16
18
20
Per Capita Income


Relationship Between Robbery Rate and Population Density
Robbery Rate per 100,000 People
0
50
100
150
200
250
0
100
200
300
400
Population per Square Mile of Land

To be clear: All "explanations" are just hypotheses and none of them are proven from a simple scatter plot.

- There are potentially many other hypotheses.
- Later in ECOM20001, and throughout ECOM30002: Econometrics 2, we will develop empirical approaches to unpack these various explanations for correlations found in scatter plots.

Part 2: Summation Practice Problems

1. Show the following equality is true

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

2. Show the following equality is true:

$$n\overline{x} = \sum_{i=1}^{n} x_i$$

3. Show the following equality is true

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

4. Show the following equality is true

$$\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Workings to Part 2

1.

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_1 - \sum \overline{x} \\
&= \sum_{i=1}^{n} x_1 - n\overline{x} \\
&= \sum_{i=1}^{n} x_1 - n\frac{\sum_{i=1}^{n} x_i}{n} \\
&= \sum_{i=1}^{n} x_1 - \sum_{i=1}^{n} x_1 = 0
\end{aligned}
$$

2

$$
\begin{aligned}
n\bar{x} &= \sum_{i=1}^{n} x_i \\
&= n\frac{\sum_{i=1}^{n} x_i}{n} \\
&= \sum_{i=1}^{n} x_i
\end{aligned}
$$

Notice how you can manipulate summations $\sum_{i=1}^{n} x_i$ and multiply them by $\frac{n}{n}$ to get means and sample sizes e.g. :

$$\sum_{i=1}^{n} x_i = \frac{n}{n}\sum_{i=1}^{n} x_i = n\overline{x}$$

3.

$$\begin{aligned}
\sum (x_i - \overline{x})^2 &= \sum \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right) \\
&= \sum x_i^2 - \sum (2\overline{x}x_i) + \sum \left(\overline{x}^2\right) \\
&= \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 \\
&= \sum x_i^2 - 2\overline{x}n\overline{x} + n\overline{x}^2 \\
&= \sum x_i^2 - n\overline{x}^2
\end{aligned}$$

In line 3, you could also multiply the term $2\overline{x}\sum x_i$ by $\frac{n}{n}$ which would give the result is the same as the above.

4.

$$
\begin{aligned}
\sum (x_i - \overline{x})(y_i - \overline{y}) &= \sum (x_i y_i - \overline{x} y_i - \overline{y} x_i + \bar{x}\bar{y}) \\
&= \sum (x_i y_i) - \sum (\overline{x} y_i) \\
&\quad - \sum (\overline{y} x_i) + \sum (\bar{x}\bar{y}) \\
&= \sum x_i y_i - \overline{x} \sum y_i - \overline{y} \sum x_i + n\bar{x}\bar{y} \\
&= \sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum x_i y_i - n\bar{x}\bar{y}
\end{aligned}
$$

the terms $\bar{x} \sum y_i$ and $\bar{y} \sum x_i$ in the 3rd line could also be divided by $\frac{n}{n}$ yielding the same result