Statistics in R

Session-4

Contents

- Covariance, correlation
- Sample, Population
- Review
- quiz

Covariance

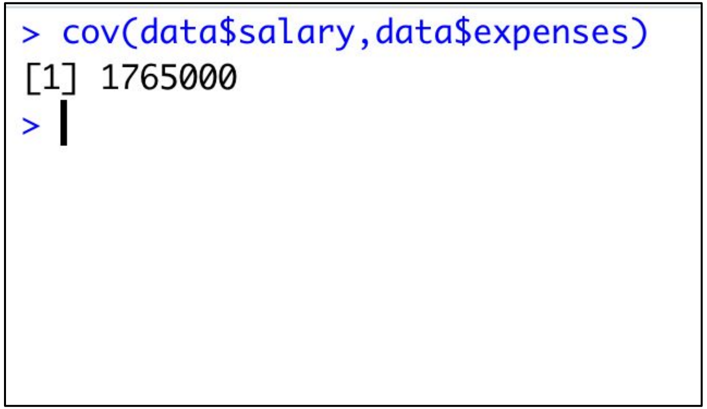
Covariance

- We know the variance (how data varies around its mean)
- Covariance tells us how much two variable varies together

name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Covariance

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$



```
> cov(data$salary,data$expenses)
[1] 1765000
> 
```

name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Covariance

name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

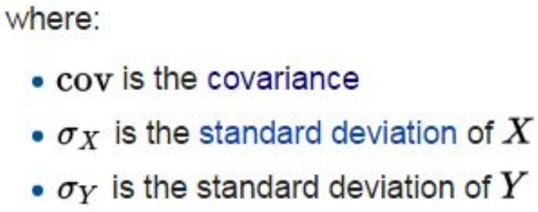
Covariance 1765000

name	age	salary	expenses
Mac	21	1500	60
Ravi	25	1800	80
David	27	1700	60
Moorey	43	3300	120
Nolan	33	2400	90

Covariance 17650

Correlation

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$



where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

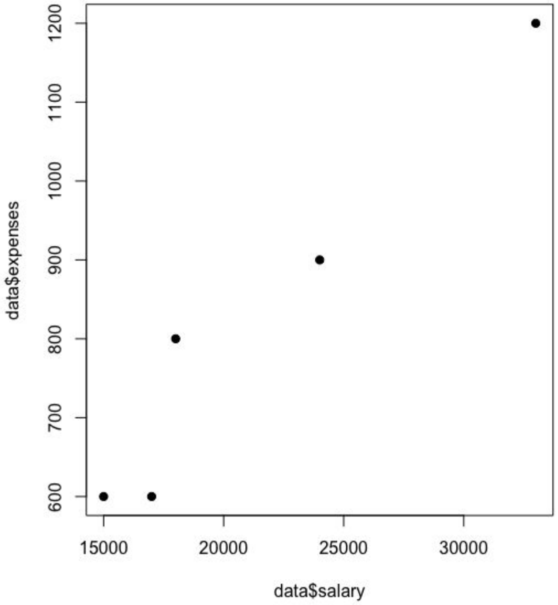
name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

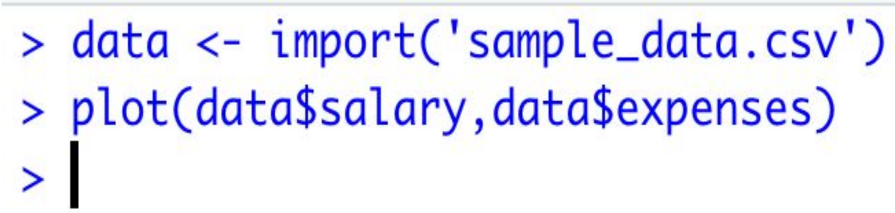
Correlation .97

name	age	salary	expenses
Mac	21	1500	60
Ravi	25	1800	80
David	27	1700	60
Moorey	43	3300	120
Nolan	33	2400	90

Correlation .97

Correlation (Some examples)

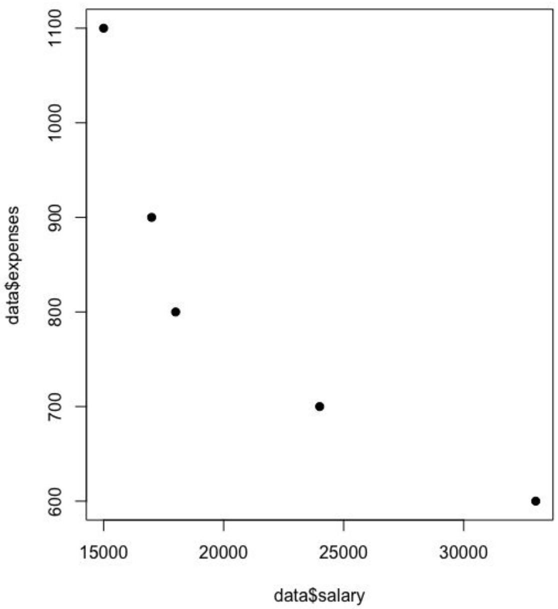




```
> data <- import('sample_data.csv')
> plot(data$salary,data$expenses)
>
```

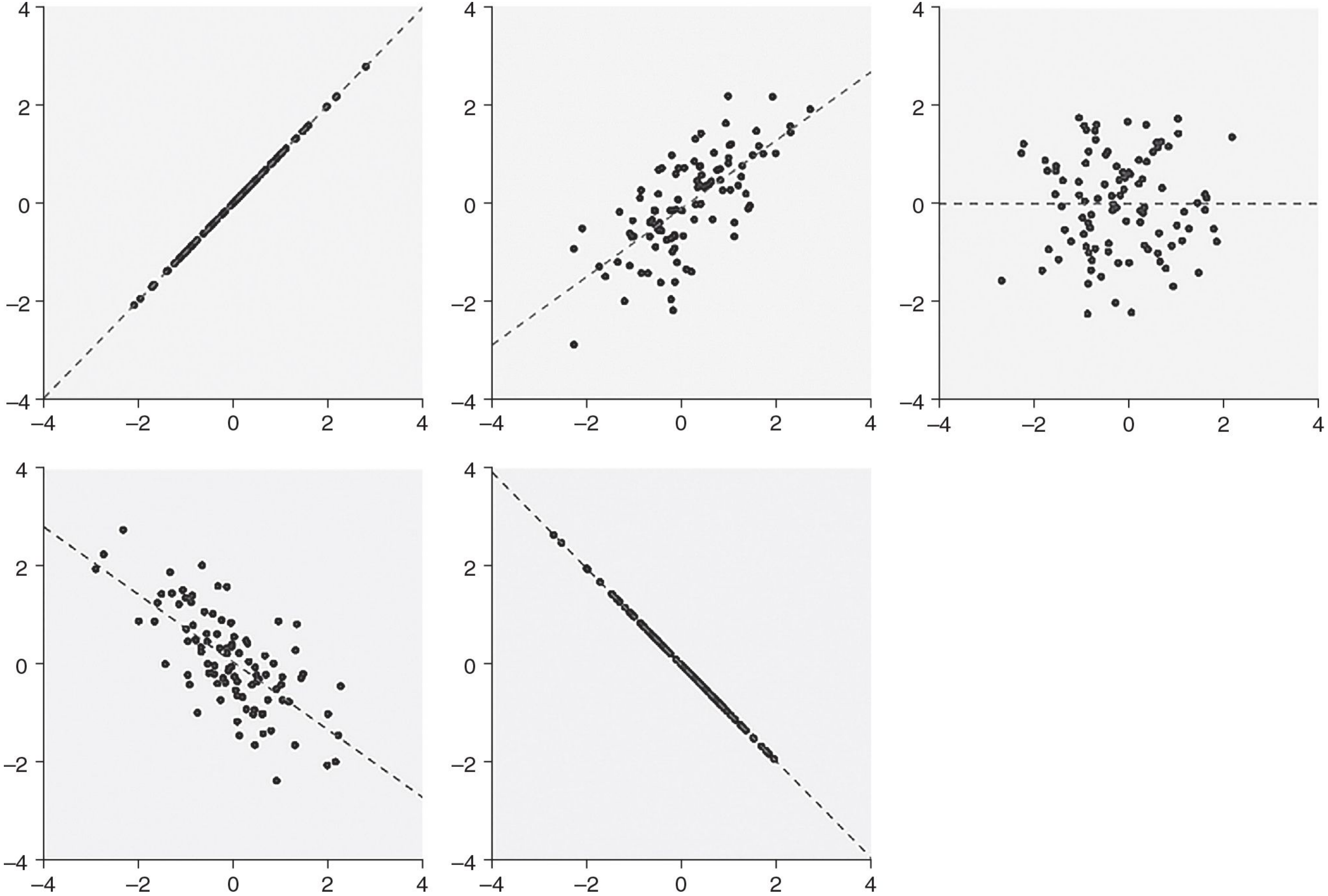
Correlation .97

Correlation (Some examples)



name	age	salary	expense s
Mac	21	15000	1100
Ravi	25	18000	800
David	27	17000	900
Moorey	43	33000	600
Nolan	33	24000	700

Correlation -.87



http://methods.sagepub.com/Reference//the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i15659.xml

Population sample

Population

Set of all objects, events, people.. under study

THE PROBLEM WITH THE WORLD IS THAT THE COLLECTIONS OF STUFF IN IT ARE SO LARGE, IT'S HARD TO GET THE INFORMATION WE WANT:
VOTING POPULATIONS: WHAT PERCENTAGE FAVORS EACH CANDIDATE?
MANUFACTURED GOODS: WHAT PROPORTION WILL BE DEFECTIVE?
PICKLES: WHAT'S THEIR AVERAGE LENGTH?
THE PICKLE-JAR MAKERS NEED TO KNOW!

Gonick, L., Smith, W., & Smith, W. (1993). *The cartoon guide to statistics* (pp. 141-142). New York: HarperPerennial.

Population

Finding average test score of students in classroom A in School A

Finding average test score of students in school A

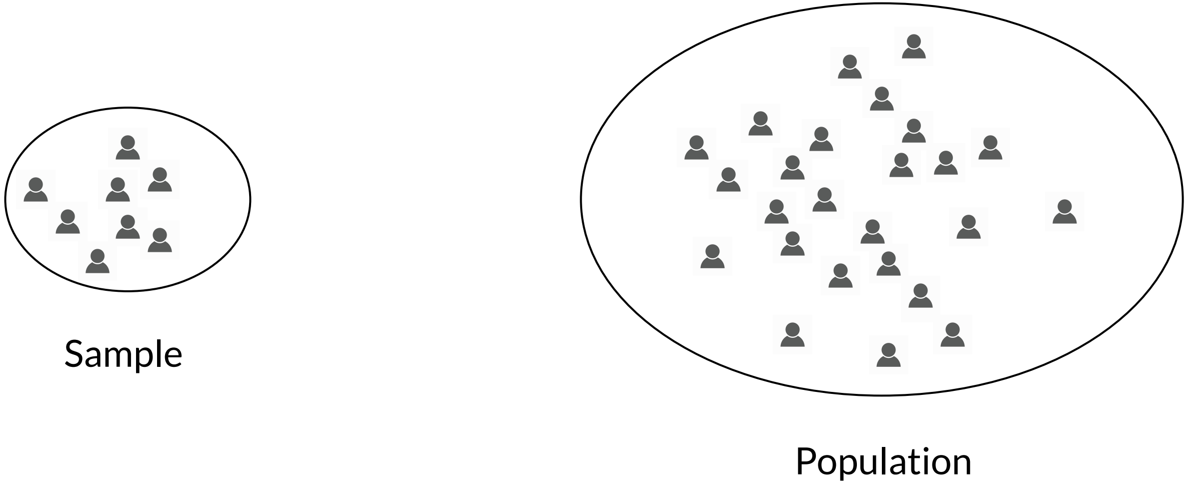
Finding average test score of students in country A

Students in classroom A in school A

Students in all classrooms in school A

Students in all all schools in country A

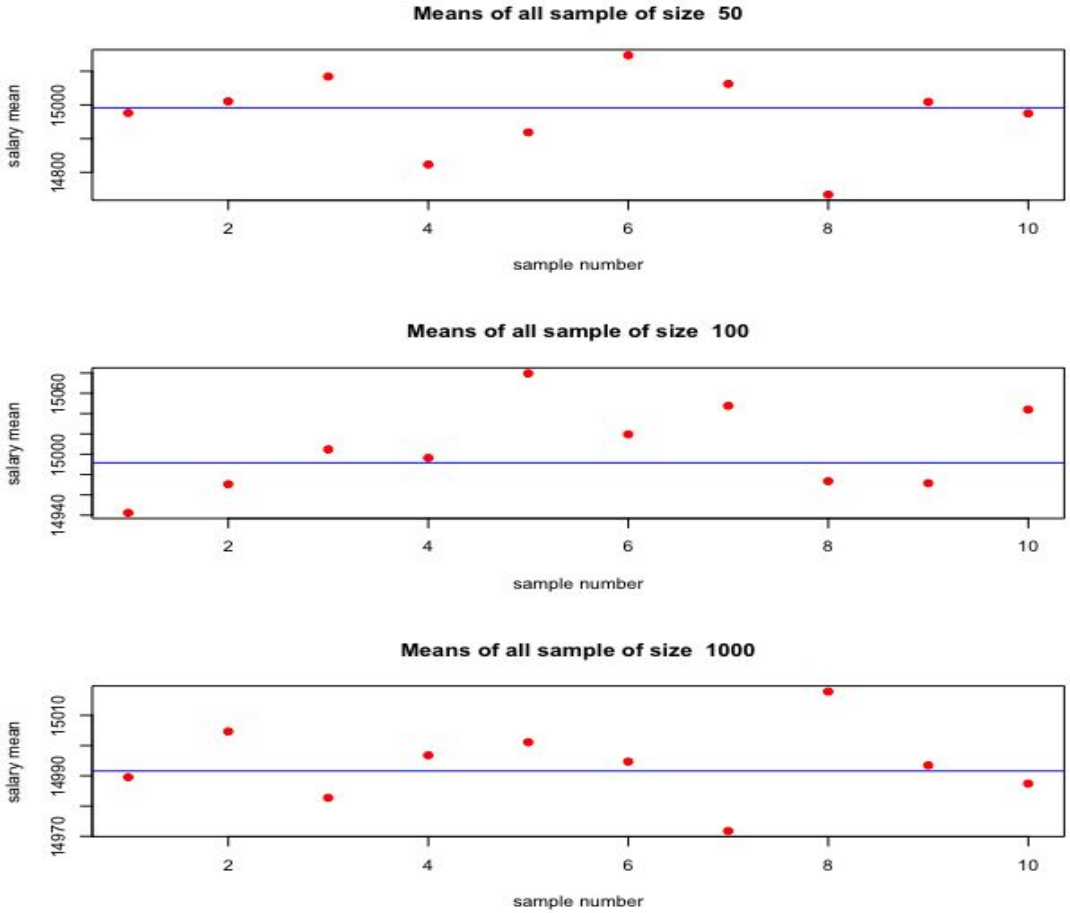
Sample



Sample

Population

Samples from assignment



Review

Session-1

- Compiler, Interpreter
- Data types in R
- Arithmetic operation in R
- Variable in R
- R script (taking input from user and performing computation)

Session-2

- If else
- Vector
  - Create, access
- Descriptive statistics
  - Distribution

Session-3

- Central tendency
- Standard deviation
- DataFrame
- Rio package (installing & loading)
- Basic ops
  - str()
  - dim()
  - colnames()
  - cor(), cov(), sd()

Quiz

Dplyr

Dataframe

Select employees who are under 21
Compute the savings of each employee

name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Dplyr

Select

Filter

Summarize

```
data_object  %>%
    function()
```

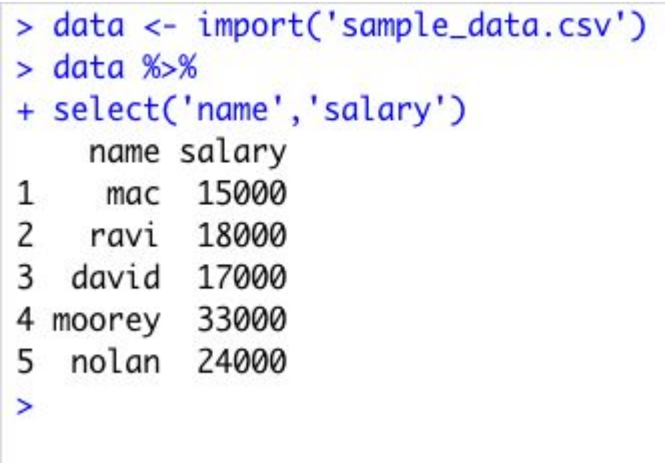
name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Dplyr

Select

data_object %>%
function()

Select all columns or first two columns or **(salary, expenses)** column



```
> data <- import('sample_data.csv')
> data %>%
+ select('name','salary')
    name salary
1    mac  15000
2   ravi  18000
3  david  17000
4 moorey  33000
5  nolan  24000
>
```

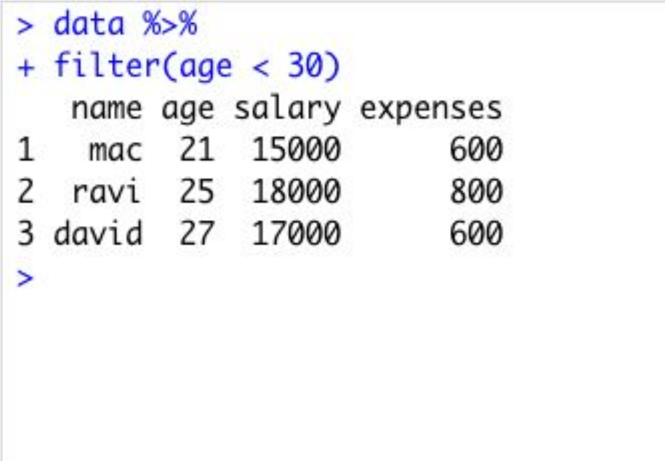
name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Dplyr

Filter

```
data_object  %>%
    function()
```

Select all employes who are under 30



```
> data %>%
+ filter(age < 30)
   name age salary expenses
1   mac  21  15000      600
2  ravi  25  18000      800
3 david  27  17000      600
>
```

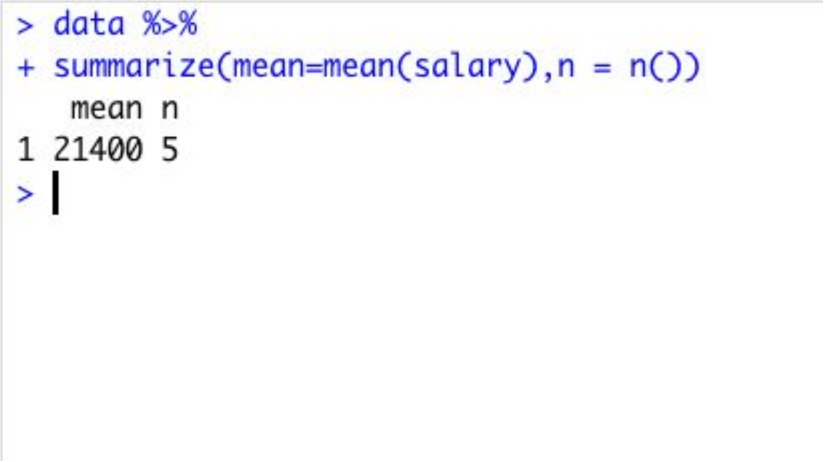
name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Dplyr

Summarize

```
data_object  %>%
     function()
```

Compute mean for employee salary



```
> data %>%
+ summarize(mean=mean(salary),n = n())
   mean n
1 21400 5
> 
```

name	age	salary	expenses
Mac	21	15000	600
Ravi	25	18000	800
David	27	17000	600
Moorey	43	33000	1200
Nolan	33	24000	900

Thank you

http://bit.ly/content-r

Practice

@datacamp

http://bit.ly/practice-r