



# Statistical Analysis on Depression

Katranitsiotis Panagiotis

# Motivation

- ▷ This subject is chosen because depression is one of the most common mental disorder.
- ▷ Even young people suffer from depression.
- ▷ Great impact on human's life.
- ▷ The purpose of this paper is to investigate the variables that affect depression.

# Research Questions

- ▷ Relation between “Depression” and “Employability” ( $X^2$ )
- ▷ If the “Number of depressed people” differs between males and females (t-test)
- ▷ If “Depression” is the commonest mental disorder comparing to other mental illnesses, such as Schizophrenia, Bipolar and Anxiety disorder (anova)
- ▷ Find the relations among the life conditions of people in non-urban areas and depression (linear and logistic regression)

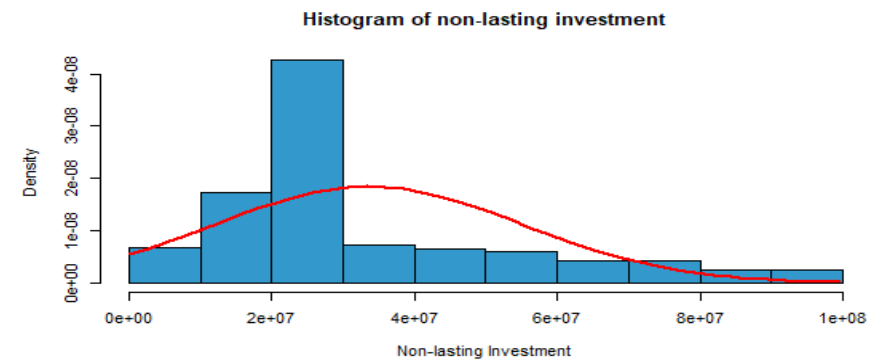
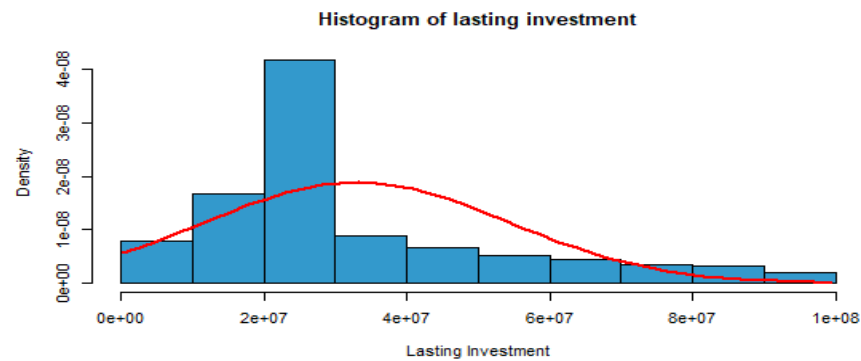
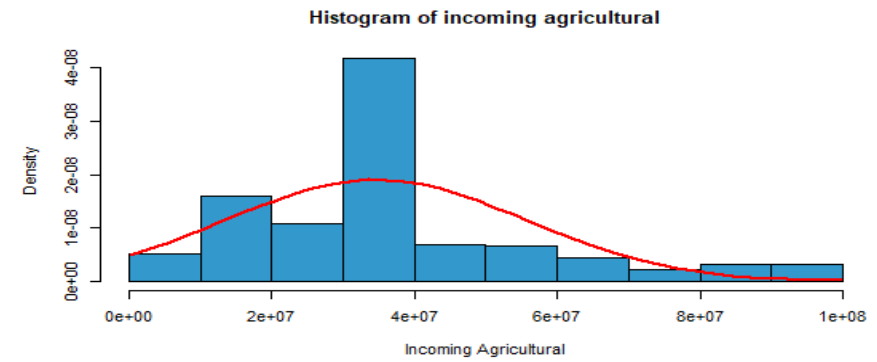
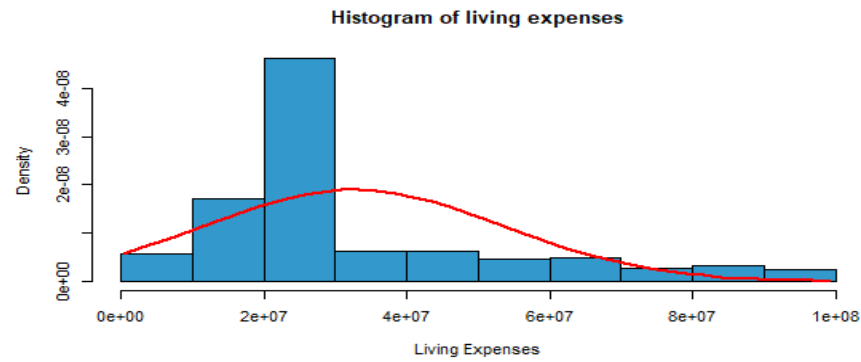
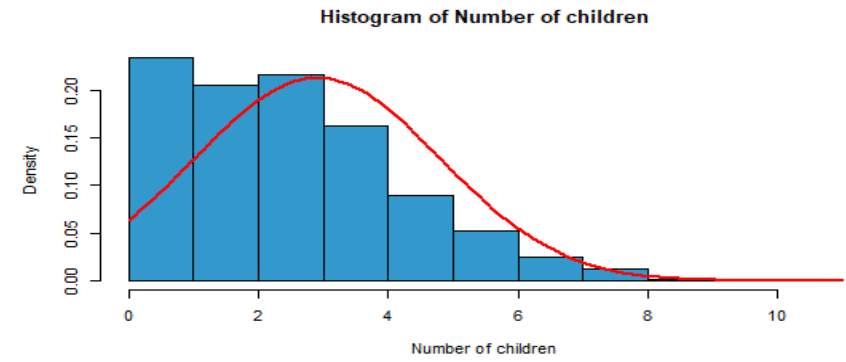
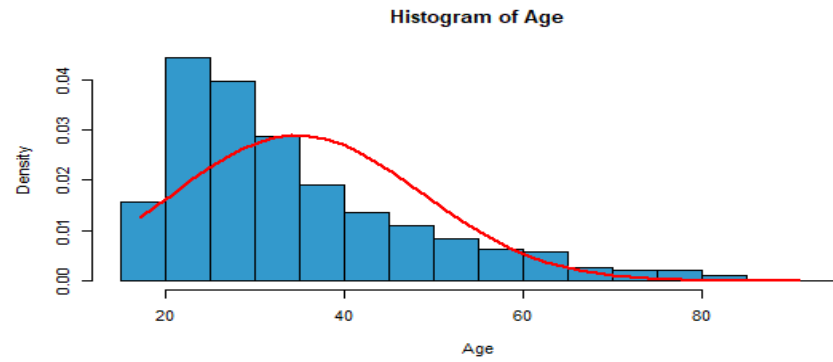
# Datasets

## ▷ Linear and Logistic Regression

- The dataset shown below is referred to depression in non-urban areas.
- Contains 1,429 observations of 23 variables.
- The ville\_id is not unique so this data cannot be used for t-test, anova and  $X^2$ .

```
[1] "Survey_id"      "ville_id"      "sex"           "Age"           "Married"       "Number_children"
[7] "education_level" "total_members" "gained_asset"  "durable_asset" "save_asset"    "living_expenses"
[13] "other_expenses" "incoming_salary" "incoming_own_farm" "incoming_business" "incoming_no_business" "incoming_agricultural"
[19] "farm_expenses"  "labor_primary"  "lasting_investment" "no_lasting_investmen" "depressed"
```

```
> head(df)
  survey_id ville_id sex Age Married Number_children education_level total_members gained_asset durable_asset save_asset living_expenses
1      926      91  1  28      1           4           10             5      28912201      22861940      23399979      26692283
2      747      57  1  23      1           3           8             5      28912201      22861940      23399979      26692283
3     1190     115  1  22      1           3           9             5      28912201      22861940      23399979      26692283
4     1065     97  1  27      1           2          10             4      52667108      19698904      49647648       397715
5      806      42  0  59      0           4          10             6      82606287      17352654      23399979      80877619
6      483      25  1  35      1           6          10             8      35937466       736707      23399979      30696127
 other_expenses incoming_salary incoming_own_farm incoming_business incoming_no_business incoming_agricultural farm_expenses labor_primary
1    28203066             0             0             0             0             30028818      31363432             0
2    28203066             0             0             0             0             30028818      31363432             0
3    28203066             0             0             0             0             30028818      31363432             0
4    44042267             0             1             0             1             22288055      18751329             0
5    74503502             1             0             0             0             53384566      20731006             1
6    11531066             0             1             0             1             22688441      18907036             0
 lasting_investment no_lasting_investmen depressed
1      28411718      28292707             0
2      28411718      28292707             1
3      28411718      28292707             0
4      7781123      69219765             0
5     20100562      43419447             0
6      4442561      76629095             0
```



# Datasets

- X-squared
- This dataset is a questionnaire about unemployment and mental illnesses.
- Contains 334 observations of 31 variables.
- For the  $X^2$  we will choose only the variables Employability and Depression.

**Table 2.** Unemployment and Mental Illness

	Variable Name	Type	Description
1.	I am currently employed at least part-time	Categorical	If the participant is employed (1) or not (0)
2.	Depression	Categorical	Whether the participant suffers from depression (1) or not (0)

# Datasets

- T-test

- This dataset is about depression for each gender.
- Contains 47,858 observations of 6 variables (table 3).
- For the performance of t-test the dataset need to be restructured (table 4).
- A new binary variable was created that refers to gender.
- We randomly chose the year 2017.

**Table 4.** Depression among Gender (restructured)

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
3.	Year	Integer	Year of each registration
4.	Percentage of population	Numeric	The number of the total population of each Entity
5.	Gender	Categorical	The gender of each participant: - Male - Female

**Table 3.** Depression among Gender

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
2.	Code	Character	The three-letter code of each entity
3.	Year	Integer	Year of each registration
4.	Prevalence - Depressive disorders - Sex: Male - Age: Standardized (Percent)	Numeric	The male percent of population that deal with depression
5.	Prevalence - Depressive disorders - Sex: Female - Age: Standardized (Percent)	Numeric	The female percent of population that deal with depression
6.	Total population (Gap-minder, HYDE & UN)	Numeric	The number of the total population of each Entity

T-test will be performed in “Percentage of population” ~ “Gender”

# Datasets

- ANOVA

- This dataset is about depression in each entity population among each illness.
- Contains 6,468 observations of 10 variables (table 5).
- We will only need mental illnesses:
  - Schizophrenia
  - Bipolar disorder
  - Anxiety Disorder
  - Depression

**Table 5.** Percentage of population for each Illness

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
2.	Code	Character	The three-letter code of each entity
3.	Year	Integer	Year of each registration
4.	Schizophrenia (%)	Numeric	The percentage of population of each entity suffering from schizophrenia
5.	Bipolar disorder (%)	Numeric	The percentage of population of each entity suffering from bipolar disorder
6.	Eating disorders (%)	Numeric	The percentage of population of each entity suffering eating disorders
7.	Anxiety disorders (%)	Numeric	The percentage of population of each entity suffering from anxiety disorder
8.	Drug use disorders (%)	Numeric	The percentage of population of each entity suffering from drug use disorder
9.	Depression (%)	Numeric	The percentage of population of each entity suffering from depression
10.	Alcohol use disorders (%)	Numeric	The percentage of population of each entity suffering from alcohol use disorder



# Datasets

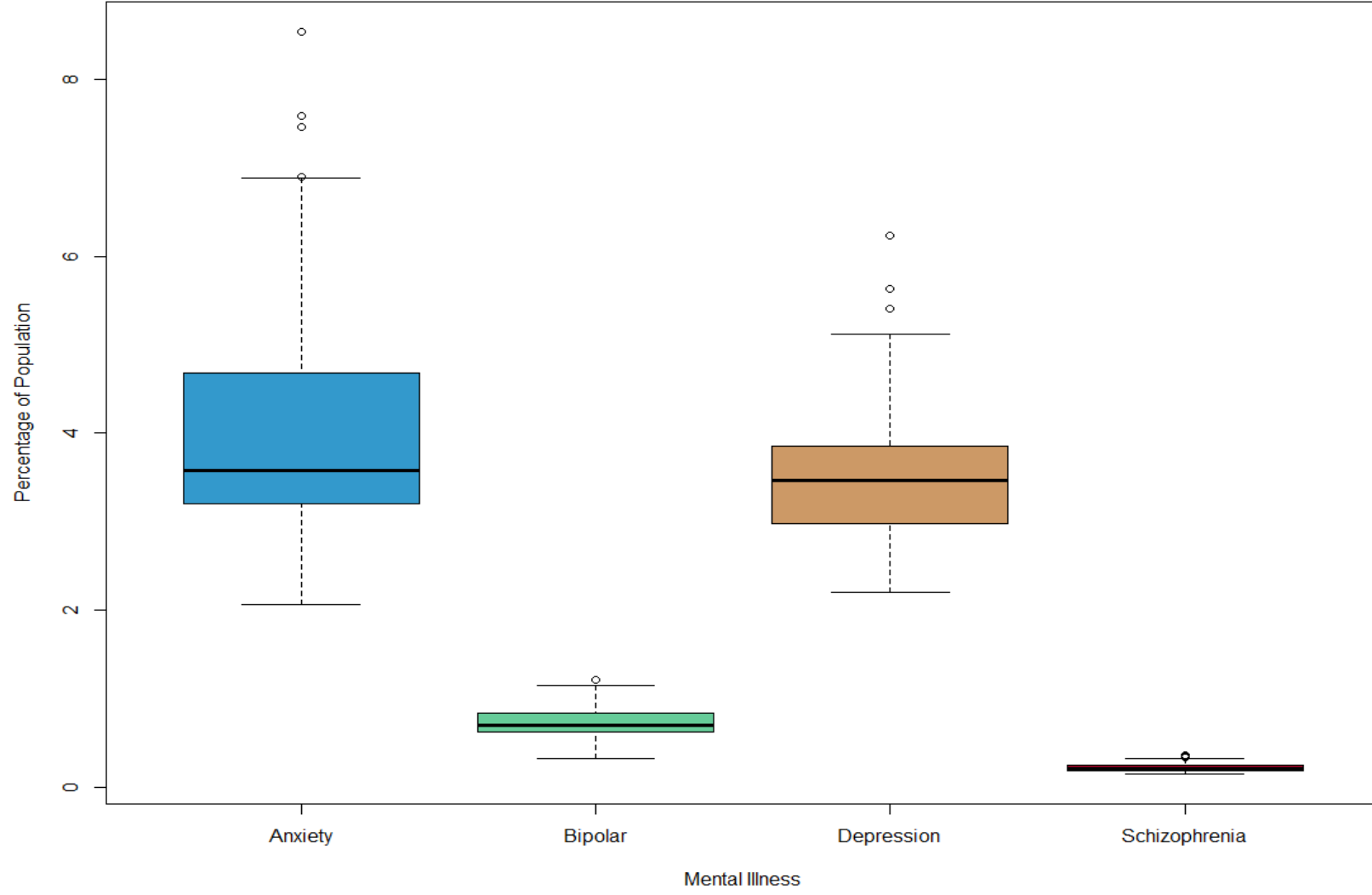
- ANOVA

- The data needed to be transformed in order to create the categorical variable Mental Illness.
- We randomly chose the year 2017.
- Anova will be used for the variables “Percentage of population” ~ “Mental Illness”

```
> head(anova.df)
# A tibble: 6 x 4
  Entity      Year Percentage_of_population Mental_illness
  <chr>    <dbl>          <dbl>    <fct>
1 Afghanistan  2017         0.166 schizophrenia
2 Albania      2017         0.201 schizophrenia
3 Algeria      2017         0.198 schizophrenia
4 American Samoa 2017         0.249 schizophrenia
5 Andean Latin America 2017         0.201 schizophrenia
6 Andorra      2017         0.264 schizophrenia
```

**Table 6.** Percentage of population exclusively among each Mental Illnesses (restructured)

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
3.	Year	Integer	Year of each registration
4.	Percentage of population	Numeric	The number of the total population of each Entity
5.	Mental Illness	Categorical	The mental illness that each participant suffers: - Schizophrenia - Bipolar disorder - Anxiety disorders - Depression



# First Insights: X-squared

- On the Crosstabulation Matrix we observe that there is no significant difference among the employed and unemployed depressed people. However, most of the non-depressed people are employed.

	Non-Depressed	Depressed
Unemployed	73	34
Employed	175	52

**Fig. 5** Employability and Depression crosstabulation matrix

- Performing  $X^2$ :  $p\text{-value} > 0.05$  so we accept the null hypothesis.

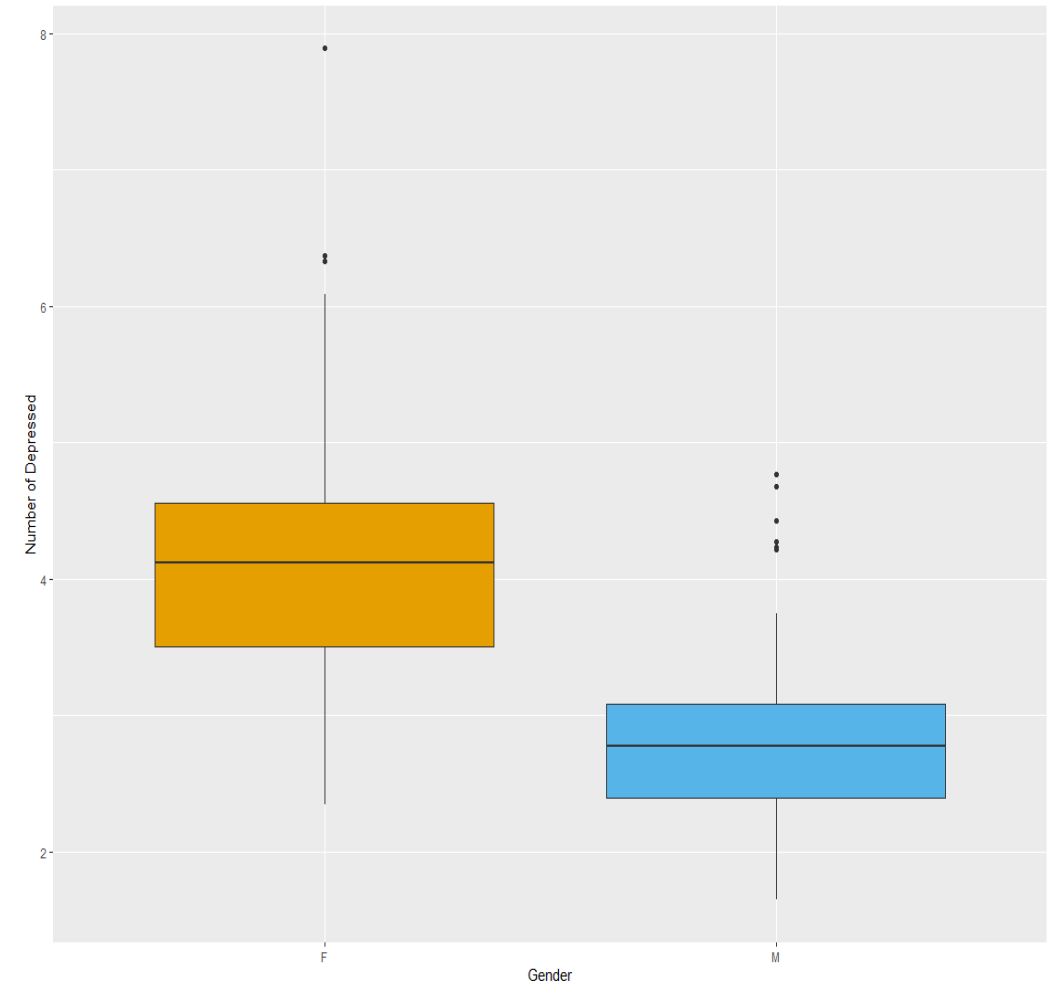
```
Pearson's Chi-squared test with Yates' continuity correction  
  
data:  t  
x-squared = 2.5456, df = 1, p-value = 0.1106
```

**Fig. 13**  $X^2$  for testing the independency of Depression and Employability

- Consequently, unemployability and depression are independent variables.

# First Insights: T-test

- ▷ Question: If the number of depressed people differs between males and females.
- ▷ The data are independent and normally distributed (considering the central limit theorem).
- ▷ In the dataset, it seems that the number of depressed Women is larger than Men



# First Insights: T-test

- ▶ For the population, t-test is performed (with 95% confidence):

```
welch Two Sample t-test
data: Number_of_Depressed by Gender
t = 20.62, df = 393.71, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.176849 1.424917
sample estimates:
mean in group F mean in group M
 4.092908      2.792025
```

**Fig. 16** T-test – Number of Depressed among Males and Females

- ▶  $p\text{-value} < 0.05$  which means that the null hypothesis is rejected.
- ▶ Consequently, the number of depressed differs between gender.
- ▶ Considering the F statistics or the confidence interval, it is concluded that the mean value of Females is larger than Males.

Thank You!