INSTITUTE
OF LAW,
POLITICS AND
DEVELOPMENT

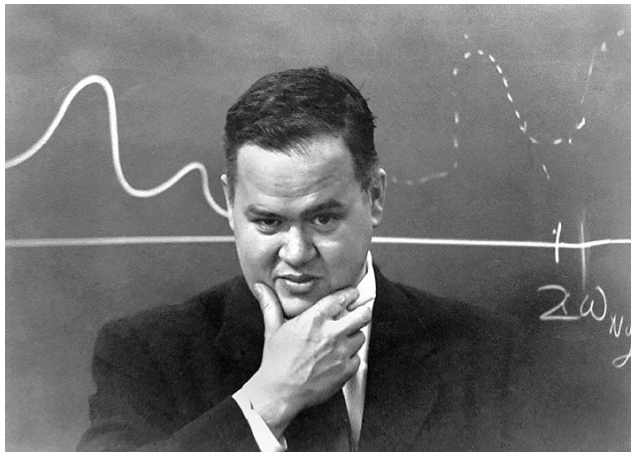
Sant'Anna
School of Advanced Studies – Pisa

Data Analysis for the Social Sciences with R

Exploratory Data Analysis (EDA)

Prof. Kevin Koehler
kevin.koehler@santannapisa.it

What is Exploratory Data Analysis (EDA)?

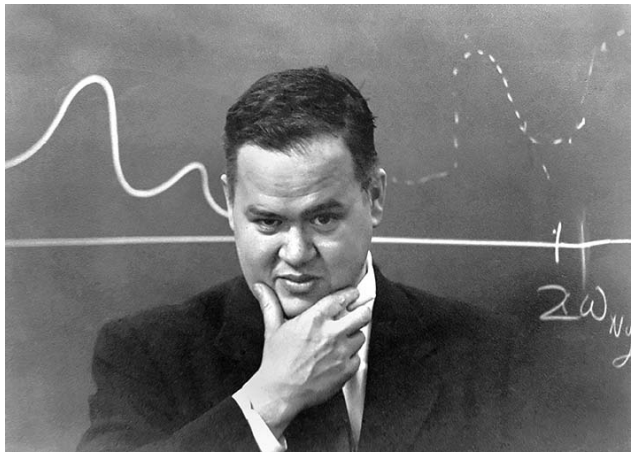


John Tukey (1915-2000)
Mathematical Statistician

EDA is about discovering *patterns of variation and covariation* in the data.

1. EDA is a- or pre-theoretical

What is Exploratory Data Analysis (EDA)?

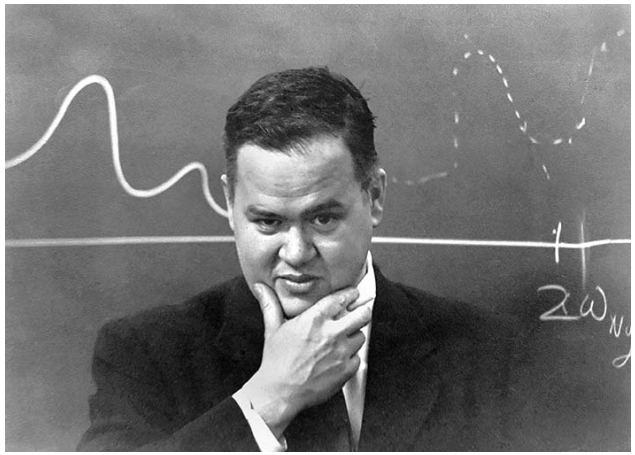


John Tukey (1915-2000)
Mathematical Statistician

EDA is about discovering *patterns of variation and covariation* in the data.

1. EDA is a- or pre-theoretical
2. You can (and should) use whatever tool comes to mind

What is Exploratory Data Analysis (EDA)?



John Tukey (1915-2000)
Mathematical Statistician

EDA is about discovering *patterns of variation and covariation* in the data.

1. EDA is a- or pre-theoretical
2. You can (and should) use whatever tool comes to mind
3. Limitation: Danger of **p-hacking**

The EDA toolbox

EDA is useful if you approach a new dataset for the first time. It allows you to form expectations about relationships which can be formalized as hypotheses and then tested.

The main types of tools are:

1. **Plotting** (of single variables or of variables against each other)

The EDA toolbox

EDA is useful if you approach a new dataset for the first time. It allows you to form expectations about relationships which can be formalized as hypotheses and then tested.

The main types of tools are:

1. **Plotting** (of single variables or of variables against each other)
2. **Numerical summaries** (descriptive statistics)

The EDA toolbox

EDA is useful if you approach a new dataset for the first time. It allows you to form expectations about relationships which can be formalized as hypotheses and then tested.

The main types of tools are:

1. **Plotting** (of single variables or of variables against each other)
2. **Numerical summaries** (descriptive statistics)
3. **Tests of association and difference** (with a central role for the notion of confidence intervals)

Today's class

Aim: conduct and systematically report on an EDA of the Tunisia survey

1. What are **confidence intervals**?
2. The EDA process
3. Reporting an EDA with R markdown

Who supported the Tunisian *autogolpe* of 25th July 2021?

Support for the Tunisian *autogolpe*

On **25 July 2021**, Tunisian president Kais Saied invoked Article 80 of the 2014 constitution, suspended parliament, and dismissed the government.

Support for the Tunisian *autogolpe*

On **25 July 2021**, Tunisian president Kais Saied invoked Article 80 of the 2014 constitution, suspended parliament, and dismissed the government.

We asked Tunisians whether they saw these events as a **necessary correction to Tunisia's democratic transition** or whether they, by contrast, **undermined the democratic transition**.

Support for the Tunisian *autogolpe*

On **25 July 2021**, Tunisian president Kais Saied invoked Article 80 of the 2014 constitution, suspended parliament, and dismissed the government.

We asked Tunisians whether they saw these events as a **necessary correction to Tunisia's democratic transition** or whether they, by contrast, **undermined the democratic transition**.

This gives us **two groups**:

1. Those who **support** July 25th (by answering that it was a necessary correction)
2. Those who **oppose** July 25th (by saying it undermines the transition)

Support for the Tunisian *autogolpe*

On **25 July 2021**, Tunisian president Kais Saied invoked Article 80 of the 2014 constitution, suspended parliament, and dismissed the government.

We asked Tunisians whether they saw these events as a **necessary correction to Tunisia's democratic transition** or whether they, by contrast, **undermined the democratic transition**.

This gives us **two groups**:

1. Those who **support** July 25th (by answering that it was a necessary correction)
2. Those who **oppose** July 25th (by saying it undermines the transition)

We want to find out **in what respects the two groups differ**.

How can we find out?

This is a typical use case for EDA

We will:

1. Plot the outcome variable

This is a typical use case for EDA

We will:

1. Plot the outcome variable
2. Plot various third variables across outcome categories

This is a typical use case for EDA

We will:

1. Plot the outcome variable
2. Plot various third variables across outcome categories
3. Test whether there are significant differences across outcome categories

Load the data (either from your hard drive, or from GitHub)

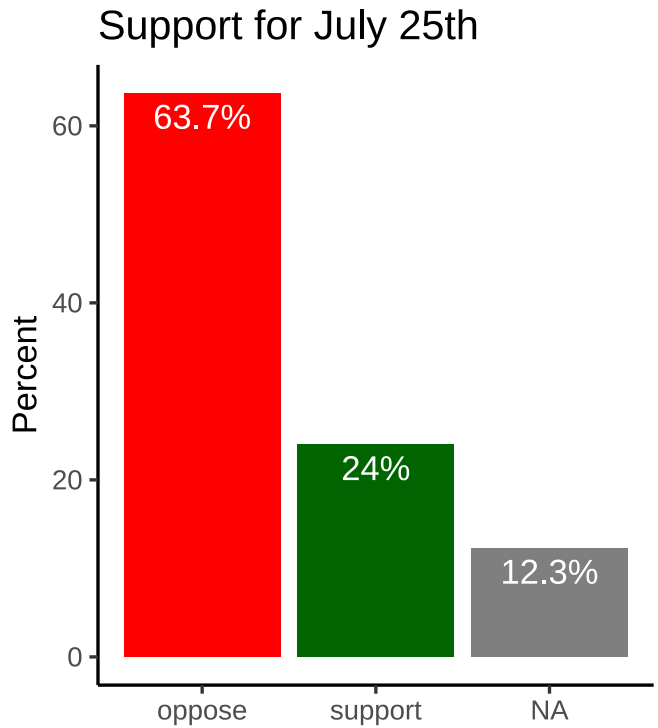
1. Plot the outcome variable

`july25`	What statement best characterizes the July 25 events in Tunisia? 1 = They represent a necessary correction to Tunisia's democratic transition 2 = They undermine the democratic transition 98 = Don't know. 99 = Declined to answer

1. Create a variable coded "oppose" for opposition and "support" for support
2. Create a bar plot showing the percentage of respondents supporting and opposing

1. Plot the outcome variable

```r
tun22 <- tun22 %>%
  mutate(sup_july25=
           case_when(
             july25==1~"oppose",
             july25==2~"support",
             TRUE~NA)
         )
tun22 %>%
  count(sup_july25) %>%
  mutate(percent=n/sum(n)*100,
         label=paste0(round(percent,2),"%")) %>%
ggplot(aes(x = sup_july25,
           y = percent,
           fill = sup_july25)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = label),
            vjust = 1.5, color="white") +
  scale_fill_manual(values = c("oppose" = "red",
                               "support" = "darkgreen",
                               "NA" = "gray")) +
  ylab("Percent") +
  xlab("") +
  ggtitle("Support for July 25th") +
  theme_classic() +
  theme(legend.position = "none")
```



2.1 Support for July 25th vs. age

How can we graphically summarize the relationship between age and attitudes toward July 25th?

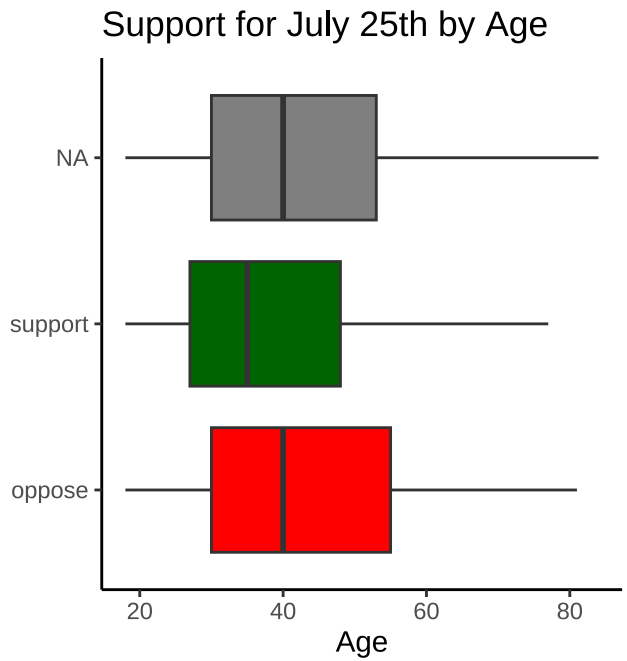
2.1 Support for July 25th vs. age

How can we graphically summarize the relationship between age and attitudes toward July 25th?

Since `age` is interval scaled, we create **boxplots** summarizing age across the differenc categories of the `sup_july25` variable.

2.1 Support for July 25th vs. age

```
ggplot(tun22, aes(x = age,
                  y = sup_july25,
                  fill = sup_july25)) +
  geom_boxplot() +
  scale_fill_manual(
    values = c(
      "oppose" = "red",
      "support" = "darkgreen",
      "NA" = "gray"
    )
  ) +
  ylab("") +
  xlab("Age") +
  ggtitle("Support for July 25th by Age") +
  theme_classic() +
  theme(legend.position = "none")
```



Is there an age difference?

How can we know?

Is there an age difference?

How can we know?

In a **frequentist** universe, one plausible answer is the following:

Is there an age difference?

How can we know?

In a **frequentist** universe, one plausible answer is the following:

We know that there is a real difference in age between supporters and opponents of July 25th if we take a large number of independent samples and find an age difference in at least 90% (or 95%, or 99%) of cases.

Is there an age difference?

How can we know?

In a **frequentist** universe, one plausible answer is the following:

We know that there is a real difference in age between supporters and opponents of July 25th if we take a large number of independent samples and find an age difference in at least 90% (or 95%, or 99%) of cases.

With an estimated population size of about 8,600,000 adult Tunisians, there are $(8,600,000)^{1,000} \approx 10^{6,934.5}$ different samples of $n = 1,000$

Is there an age difference?

How can we know?

In a **frequentist** universe, one plausible answer is the following:

We know that there is a real difference in age between supporters and opponents of July 25th if we take a large number of independent samples and find an age difference in at least 90% (or 95%, or 99%) of cases.

With an estimated population size of about 8,600,000 adult Tunisians, there are $(8,600,000)^{1,000} \approx 10^{6,934.5}$ different samples of $n = 1,000$

Assume we take 1,000 independent samples of 1,000 Tunisians each, calculate the average age of supporters and opponents, and report the percentage of samples for which we found a difference. **This is our level of confidence that there is a real difference**

The logic of t-tests

But we only have one sample.

The logic of t-tests

But we only have one sample.

So instead of drawing thousands of samples, we ask:

How likely is it to see an age difference this big between groups, purely by chance, if there were no real difference in the population?

The logic of t-tests

But we only have one sample.

So instead of drawing thousands of samples, we ask:

How likely is it to see an age difference this big between groups, purely by chance, if there were no real difference in the population?

This is where t-tests come in. A t-test asks:

- What is the observed difference in average age?
- How much variability is there within each group?
- Is the observed difference large enough, relative to this variability, to reject the idea that it is due to chance?

The logic of t-tests

But we only have one sample.

So instead of drawing thousands of samples, we ask:

How likely is it to see an age difference this big between groups, purely by chance, if there were no real difference in the population?

This is where t-tests come in. A t-test asks:

- What is the observed difference in average age?
- How much variability is there within each group?
- Is the observed difference large enough, relative to this variability, to reject the idea that it is due to chance?

If the probability of seeing a difference this large by chance is less than a defined threshold (called α), we conclude: It is unlikely this happened by chance. There likely is a real difference.

The logic of t-tests

We compare:

- **Observed difference** between sample means
- **Expected difference** under the null hypothesis ($H_0 : \mu_1 = \mu_2$)

By computing a **t-statistic**:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{x}_i$: sample mean
- s_i^2: sample variance
- n_i: sample size

The logic of t-tests

Assumptions:

1. **Independence**: Observations must be independent.
2. **Normality**: Data in each group should be approximately normally distributed (esp. for small samples).
3. **Equal variance** (for Student's t-test): Variances in both groups should be similar.

The logic of t-tests

Assumptions:

1. **Independence**: Observations must be independent.
2. **Normality**: Data in each group should be approximately normally distributed (esp. for small samples).
3. **Equal variance** (for Student's t-test): Variances in both groups should be similar.

p-value: Probability of observing data as extreme as ours, if H_0 were true.

1. If $p < \alpha$ (e.g., 0.05), we reject H_0 $\rightarrow$evidence of a significant difference.
2. If $p \geq \alpha$, we fail to reject H_0 $\rightarrow$no statistically significant difference.

The logic of t-tests

William Sealy Gosset, aka "Student"
1876-1937
Head-Brewer of Guinness
Developed small-sample methods for hypothesis testing

Student's t-test:

```
ttest <- t.test(age~sup_july25,
                data=tun22,
                var.equal=T)
```

The average age of supporters is 37.74, while that of opponents is 42.71. The difference is significant with $p = 1.1 \times 10^{-5}$.

The logic of t-tests

William Sealy Gosset, aka "Student"
1876-1937
Head-Brewer of Guinness
Developed small-sample methods for hypothesis testing

Student's t-test:

```
ttest <- t.test(age~sup_july25,
                data=tun22,
                var.equal=T)
```

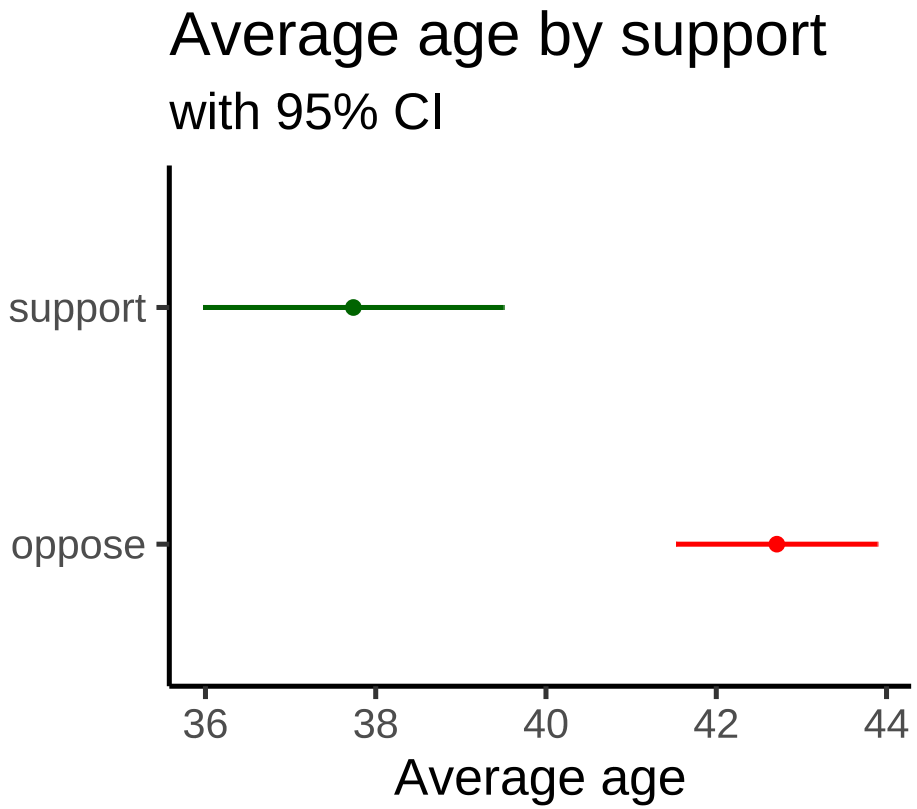
The average age of supporters is 37.74, while that of opponents is 42.71. The difference is significant with p = 1.1×10^{-5}.

Given unequal variance, we should use a Welch t-test (the results remain):

```
ttest <- t.test(age~sup_july25,
                data=tun22)
```

The logic of t-tests

William Sealy Gosset, aka "Student"
1876-1937
Head-Brewer of Guinness
Developed small-sample methods for hypothesis testing



2.2 Support for July 25th vs. populism

There are three items in the survey which are taken from standard batteries measuring populism (`mps`, `people`, and `officials`)

2.2 Support for July 25th vs. populism

There are three items in the survey which are taken from standard batteries measuring populism (`mps`, `people`, and `officials`)

First, note that the populism variables are coded inversely (i.e., smaller values mean more agreement).

This is counterintuitive and should be reversed:

2.2 Support for July 25th vs. populism

There are three items in the survey which are taken from standard batteries measuring populism (`mps`, `people`, and `officials`)

First, note that the populism variables are coded inversely (i.e., smaller values mean more agreement).

This is counterintuitive and should be reversed:

```
reverse_code <- function(x, min = 1, max = 5, na.threshold=90){
  x[x > na.threshold] <- NA
  if(min(x, na.rm = TRUE) < min | max(x, na.rm = TRUE) > max){
    warning("Warning: input is outside the range of the scale.")
  }
  return((max + min) - x)
}
tun22$mps <- reverse_code(tun22$mps)
tun22$people <- reverse_code(tun22$people)
tun22$officials <- reverse_code(tun22$officials)
```

2.2 Support for July 25th vs. MPs lose touch

`mps`	To what extent do you agree with the following statement: "Members of Parliament very quickly lose touch with ordinary people after they assume office." 1 = Agree Strongly 2 = Agree Somewhat 3 = Neither agree nor disagree 4 = Disagree Somewhat 5 = Disagree Strongly 98 = Don't Know 99 = Declined to answer

Considering how the `mps` variables is coded, create an appropriate plot which shows the distribution of agreement and disagreement across the categories of the `sup_july25` variable.

2.2 Support for July 25th vs. MPs lose touch

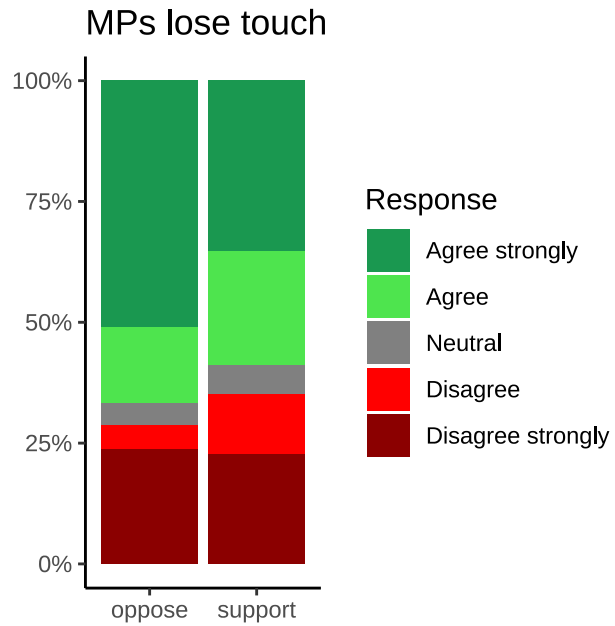
`mps`	To what extent do you agree with the following statement: "Members of Parliament very quickly lose touch with ordinary people after they assume office." 1 = Agree Strongly 2 = Agree Somewhat 3 = Neither agree nor disagree 4 = Disagree Somewhat 5 = Disagree Strongly 98 = Don't Know 99 = Declined to answer

Considering how the `mps` variables is coded, create an appropriate plot which shows the distribution of agreement and disagreement across the categories of the `sup_july25` variable.

There are different solutions, I went with a stacked bar graph (i.e., one bar oppose and one for support, each showing the distribution of responses.)

2.2 Support for July 25th vs. MPs lose touch

```
tun22 %>%
  filter(!is.na(sup_july25) & !is.na(mps)) %>%
  group_by(sup_july25, mps) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(sup_july25) %>%
  mutate(
    prop = n / sum(n),
    people = factor(mps,
                    levels = 1:5,
                    labels = c("Disagree strongly",
                               "Disagree",
                               "Neutral",
                               "Agree",
                               "Agree strongly"),
                    ordered = TRUE),
    people = forcats::fct_rev(mps)) %>%
  ggplot(aes(x = sup_july25, y = prop, fill = mps)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_fill_manual(values = c("#1a9850","#4ee44e",
                               "#808080","#ff0000",
                               "#8b0000")) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "MPs lose touch",
       x = " ", y = " ", fill = "Response") +
  theme_classic()
```



Respondents who support July 25th seem marginally *less* critical of MPs.

Is there a real difference in the assessment of MPs?

How can we know?

Is there a real difference in the assessment of MPs?

How can we know?

We cannot simply use a t-test, since our variable is not interval-scaled (but ordinal). The alternative is the **Wilcoxon rank-sum test** (or **Mann-Whitney U Test**).

Is there a real difference in the assessment of MPs?

How can we know?

We cannot simply use a t-test, since our variable is not interval-scaled (but ordinal). The alternative is the **Wilcoxon rank-sum test** (or **Mann-Whitney U Test**).

The core intuition is to use the ranks of the values, not the values themselves. This deals with the problem of ordinal scales, and is robust to outliers and non-normal distributions. Once we use rank-sums, the same frequentist considerations apply as for the t-test above.

The Wilcoxon rank-sum test

First, we rank all observations. Then we calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

Where

- R_1, R_2 are the sum of ranks for each group

The Wilcoxon rank-sum test

First, we rank all observations. Then we calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

Where

- R_1, R_2 are the sum of ranks for each group

Then, we take

$$U = min(U_1, U_2)$$

The Wilcoxon rank-sum test

Under the null hypothesis of no group differences, the expected value of U is:

$$E(U) = \frac{n_1 n_2}{2}$$

The Wilcoxon rank-sum test

Under the null hypothesis of no group differences, the expected value of U is:

$$E(U) = \frac{n_1 n_2}{2}$$

while the standard deviation (for the simplest case without ties) is:

$$SD(U) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The Wilcoxon rank-sum test

Under the null hypothesis of no group differences, the expected value of U is:

$$E(U) = \frac{n_1 n_2}{2}$$

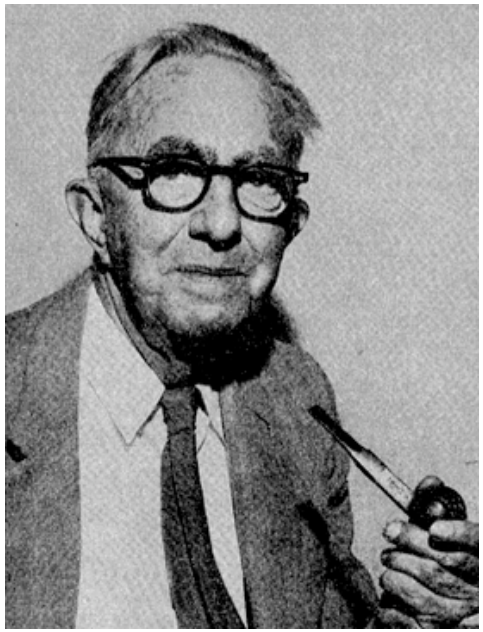
while the standard deviation (for the simplest case without ties) is:

$$SD(U) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

This allows us to calculate a z-score (for large samples):

$$z = \frac{U - E(U)}{SD(U)}$$

The Wilcoxon rank-sum test



Frank Wilcoxon, 1892-1965
Chemist and statistician

The Wilcoxon rank-sum test:

```
w_test <- wilcox.test(mps~sup_july25,
                      data=tun22,
                      exact=FALSE)
```

We can conclude that the difference we observed is significant with p = 0.002803.

In other words, opponents of July 25th are more critical of MPs than supporters.

Reporting EDA with R markdown

Sant'Anna
School of Advanced Studies – Pisa

Reporting EDA with R markdown

1. Type `install.packages("rmarkdown")`
2. Open an R `markdown` file
3. Name it and choose an output format (use PDF for today)

The empty file contains some useful basic information on Markdown. For more go to https://rmarkdown.rstudio.com/.

Reporting EDA with R markdown

```yaml
---
title: "Exploratory Data Analysis"
author: "Kevin Koehler"
date: "2025-05-03"
format:
  pdf:
    toc: TRUE
    number-sections: true
    link-citations: true
engine: knitr
header-inlcudes:
  - \usepackage{minipage}
  - \usepackage{array}
  - \usepackage{float}
---
```

The part between the three horizontal dashes is called the **YAML header** (YAML stands for **Y**et **A**nother **M**arkdown **L**anguage).

Reporting EDA with R markdown

```yaml
---
title: "Exploratory Data Analysis"
author: "Kevin Koehler"
date: "2025-05-03"
format:
  pdf:
    toc: TRUE
    number-sections: true
    link-citations: true
engine: knitr
header-inlcudes:
  - \usepackage{minipage}
  - \usepackage{array}
  - \usepackage{float}
---
```

The part between the three horizontal dashes is called the **YAML header** (YAML stands for **Y**et **A**nother **M**arkdown **L**anguage).

You can include LaTeX packages in the header (for later use in the document itself)

Reporting EDA with R markdown

You can include text and R code in the same document.

Reporting EDA with R markdown

You can include text and R code in the same document.

Text can be written in specific markdown code, or in LaTeX. Here are some basic commands:

1. Headings are created with `# Heading 1`, `## Heading 2`, `### Heading 3`, etc.
2. Text can be **bolded** using either `**bold text**` (markdown) or `\textbf{bold text}` (LaTeX)
3. The equivalent for *italics* is either `*italics*` (markdown) or `\textit{italics}` (LaTeX)
4. Include image files with `![](your_image_file){options}`

Reporting EDA with R markdown

Code can be included either **inline** or as **code chunks**

Reporting EDA with R markdown

Code can be included either **inline** or as **code chunks**

Inline:

```
The average value of age is `r round(mean(tun22$age, na.rm=T),2)`
```

Reporting EDA with R markdown

Code can be included either **inline** or as **code chunks**

Inline:

```
The average value of age is `r round(mean(tun22$age, na.rm=T),2)`
```

Code chunk

````
```{r summary_stats, echo=F, results='asis'}
stargazer(as.data.frame(tun22),
          type = "latex",
          header = F,
          title = "Variables and summary statistics",
          label = "sum_stats")
```
````