Statistics

Mihaela Gerova

StudCee

October 24, 2016

Table of Contents

Disclaimer

What I DON'T know

- Some stuff about R
- Stuff that depends on the teacher
- Some complicated stuff

Disclaimer

What I DON'T know

- Some stuff about R
- Stuff that depends on the teacher
- Some complicated stuff

What I DO know

- A lot of stuff about statistics

Tips

- With "?" you can get information about functions and datasets

Tips

- With "?" you can get information about functions and datasets
- Make flowcharts of methods for data analysis to see how they relate to each other

Tips

- With "?" you can get information about functions and datasets
- Make flowcharts of methods for data analysis to see how they relate to each other
- Or check out some useful links (studcee.svcover.nl!)

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **One** variable. Data numerical or categorical?

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **One** variable. Data numerical or categorical?
  - numerical data. Use: *histogram*, *boxplot*, *stem-and-leaf plot* (not usually).

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **One** variable. Data numerical or categorical?
  - numerical data. Use: *histogram*, *boxplot*, *stem-and-leaf plot* (not usually).
  - categorical data. Use: *barplot*, *pie chart* (not preferred).

Exploratory Data Analysis: Which method should I use?

From data set 'survey'

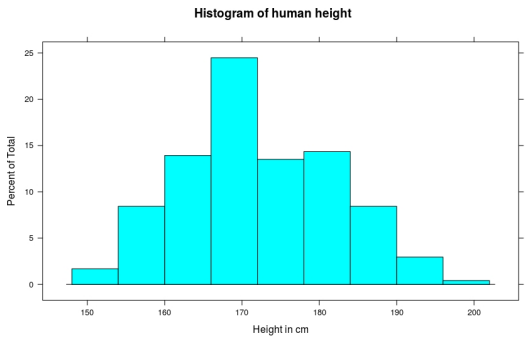


Figure: Histogram of human height

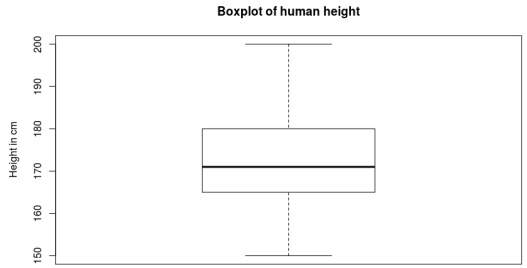


Figure: Boxplot of human height

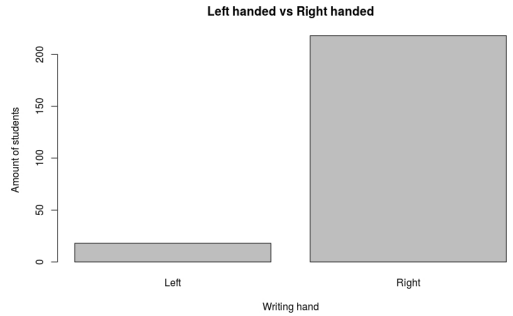


Figure: Barplot of writing hand

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **Two** variables. Data numerical or categorical?

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **Two** variables. Data numerical or categorical?
  - both categorical. Use: *multiple barplots*.

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **Two** variables. Data numerical or categorical?
  - both categorical. Use: *multiple barplots*.
  - one numerical, one categorical. Use: *multiple boxplots*.

Exploratory Data Analysis: Which method should I use?

We want to look at the distribution(s). Are we looking at one or two variables at a time?

- **Two** variables. Data numerical or categorical?
  - both categorical. Use: *multiple barplots*.
  - one numerical, one categorical. Use: *multiple boxplots*.
  - both numerical. Use: *scatterplot*. In this case, we're looking at a relationship, not at distributions.

Exploratory Data Analysis: Which method should I use?

From data set 'survey'

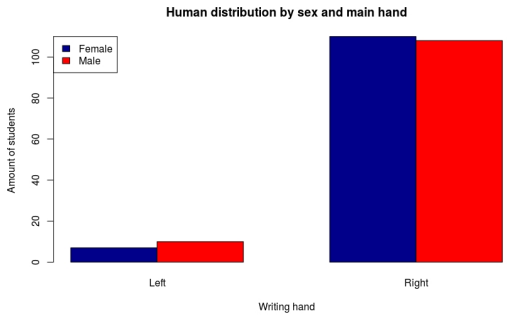


Figure: Barplot of main writing hand vs sex

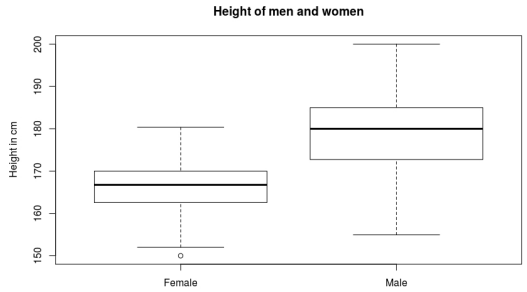


Figure: Boxplot of height of men and women

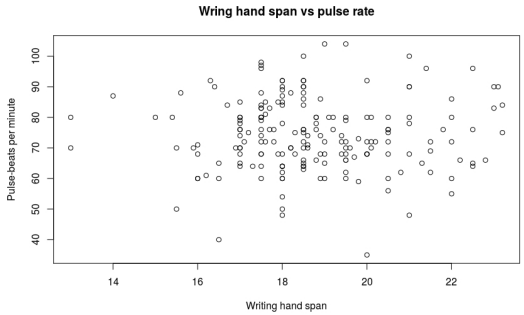


Figure: Scatterplot of writing hand span vs pulse rate

Formal Data Analysis: Which method should I use?

This depends on both the kind of data you're working with and the kind of research question you pose.
Kinds of data: numerical, binary, (non-binary) categorical.
Kinds of questions:

1. Is population 1's parameter equal to X?

Formal Data Analysis: Which method should I use?

This depends on both the kind of data you're working with and the kind of research question you pose.
Kinds of data: numerical, binary, (non-binary) categorical.
Kinds of questions:

1. Is population 1's parameter equal to X?
2. Is population 1 in some respect equal to population 2?

Formal Data Analysis: Which method should I use?

This depends on both the kind of data you're working with and the kind of research question you pose.
Kinds of data: numerical, binary, (non-binary) categorical.
Kinds of questions:

1. Is population 1's parameter equal to X?
2. Is population 1 in some respect equal to population 2?
3. Is variable A spread according to given probabilities?

Formal Data Analysis: Which method should I use?

This depends on both the kind of data you're working with and the kind of research question you pose.
Kinds of data: numerical, binary, (non-binary) categorical.
Kinds of questions:

1. Is population 1's parameter equal to X?
2. Is population 1 in some respect equal to population 2?
3. Is variable A spread according to given probabilities?
4. Are variables A and B independent?

Formal Data Analysis: Which method should I use?

This depends on both the kind of data you're working with and the kind of research question you pose.
Kinds of data: numerical, binary, (non-binary) categorical.
Kinds of questions:

1. Is population 1's parameter equal to X?
2. Is population 1 in some respect equal to population 2?
3. Is variable A spread according to given probabilities?
4. Are variables A and B independent?
5. Questions about relationships / predictions

FDA for binary data

- (Q1) Is the mean equal to X? Use a *one-sample test of proportion*

FDA for binary data

- (Q1) Is the mean equal to X? Use a *one-sample test of proportion*
  - **Ex:** Known poverty rate in 2000 is 11.3%. Sample of 50,000 in 2001; 5,850 (11.7%) indicate poverty.

FDA for binary data

- (Q1) Is the mean equal to X? Use a *one-sample test of proportion*
  - **Ex:** Known poverty rate in 2000 is 11.3%. Sample of 50,000 in 2001; 5,850 (11.7%) indicate poverty.
  - **Question:** Did rate of poverty increase?

FDA for binary data

- (Q1) Is the mean equal to X? Use a *one-sample test of proportion*
  - **Ex:** Known poverty rate in 2000 is 11.3%. Sample of 50,000 in 2001; 5,850 (11.7%) indicate poverty.
  - **Question:** Did rate of poverty increase?
  - p-value = 0.004831

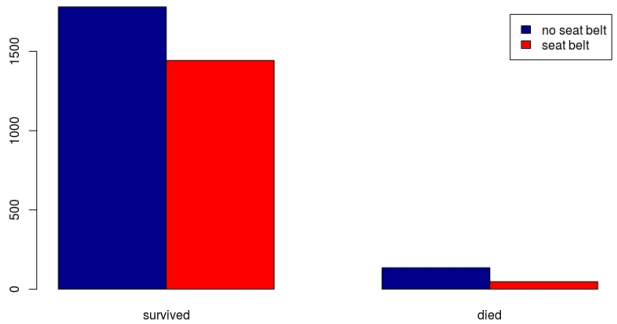
FDA for binary data

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *two-sample test of proportion*

FDA for binary data

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *two-sample test of proportion*
  - **Ex:** Given the following data, decide if the seat belt makes a difference in the chance of dying after a car accident.

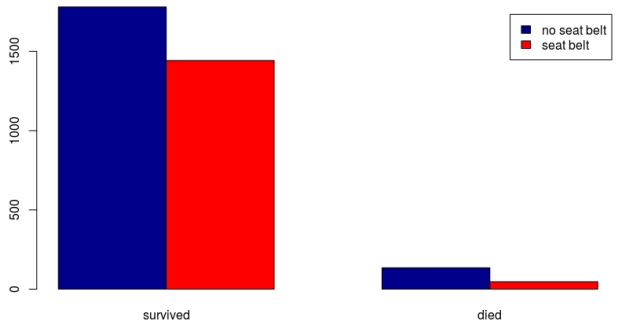
	Survived	Dead
No seat belt	1781	135
Seat belt	1443	47



FDA for binary data

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *two-sample test of proportion*
  - **Ex:** Given the following data, decide if the seat belt makes a difference in the chance of dying after a car accident.

	Survived	Dead
No seat belt	1781	135
Seat belt	1443	47

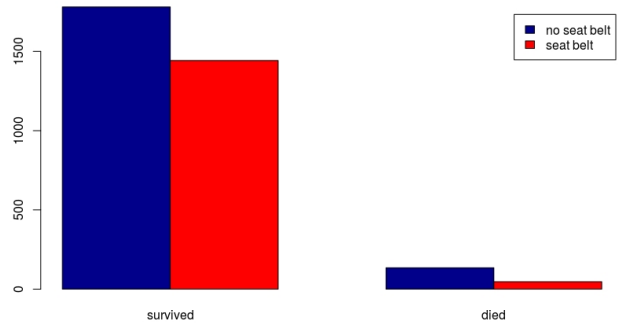


  - p-value = 8.105e-07

FDA for binary data

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *two-sample test of proportion*
  - **Ex:** Given the following data, decide if the seat belt makes a difference in the chance of dying after a car accident.

	Survived	Dead
No seat belt	1781	135
Seat belt	1443	47



  - p-value = 8.105e-07
- Example taken from:
  `http://www.dummies.com/programming/r/how-to-test-data-proportions-with-r/`

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q1) Is the mean equal to Number? Use a *one-sample t-test*

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q1) Is the mean equal to Number? Use a *one-sample t-test*
  - **Ex:** An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were: 0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q1) Is the mean equal to Number? Use a *one-sample t-test*
  - **Ex:** An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were: 0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418
  - **Question:** Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q1) Is the mean equal to Number? Use a *one-sample t-test*
  - **Ex:** An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were: 0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418
  - **Question:** Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?
  - p-value = 0.02927

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q1) Is the mean equal to Number? Use a *one-sample t-test*
  - **Ex:** An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were: 0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418
  - **Question:** Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?
  - p-value = 0.02927
- Example taken from
  `http://www.stat.columbia.edu/~martin/W2024/R2.pdf`

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *paired t-test* if the data are paired (two points per individual); else, use a *two-sample t-test*

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *paired t-test* if the data are paired (two points per individual); else, use a *two-sample t-test*
  - **Ex:** A researcher wants to investigate differences between the devices of two different companies. She subjects a number of devices from each company to various tests. The results of the tests are combined into a general performance score.

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *paired t-test* if the data are paired (two points per individual); else, use a *two-sample t-test*
  - **Ex:** A researcher wants to investigate differences between the devices of two different companies. She subjects a number of devices from each company to various tests. The results of the tests are combined into a general performance score.

```
data1$Apples
 1] 78.45 80.32 82.75 78.04 79.86 80.23 81.23 79.58 83.44 79.76 80.72 81.70 79.32 78.20 83.09 76.00 81.52 80.06 81.75 80.75
data1$Tamtung
 1] 83.09 79.80 82.59 82.95 81.00 78.55 81.48 80.40 81.79 81.29 81.74 81.32 82.08 80.72 80.22 80.40 79.27 80.10 80.44 80.75
```

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *paired t-test* if the data are paired (two points per individual); else, use a *two-sample t-test*
  - **Ex:** A researcher wants to investigate differences between the devices of two different companies. She subjects a number of devices from each company to various tests. The results of the tests are combined into a general performance score.

```
data1$Apples
 1] 78.45 80.32 82.75 78.04 79.86 80.23 81.23 79.58 83.44 79.76 80.72 81.70 79.32 78.20 83.09 76.00 81.52 80.06 81.75 80.75
data1$Tamtung
 1] 83.09 79.80 82.59 82.95 81.00 78.55 81.48 80.40 81.79 81.29 81.74 81.32 82.08 80.72 80.22 80.40 79.27 80.10 80.44 80.75
```

  - **Question:** Is there a difference between the results of the two companies?

FDA for numerical data

Assumption t-test: Normal Distribution

- (Q2) Is the mean in population 1 equal to the mean in population 2? Use a *paired t-test* if the data are paired (two points per individual); else, use a *two-sample t-test*
  - **Ex:** A researcher wants to investigate differences between the devices of two different companies. She subjects a number of devices from each company to various tests. The results of the tests are combined into a general performance score.

```
data1$Apples
 1] 78.45 80.32 82.75 78.04 79.86 80.23 81.23 79.58 83.44 79.76 80.72 81.70 79.32 78.20 83.09 76.00 81.52 80.06 81.75 80.75
data1$Tamtung
 1] 83.09 79.80 82.59 82.95 81.00 78.55 81.48 80.40 81.79 81.29 81.74 81.32 82.08 80.72 80.22 80.40 79.27 80.10 80.44 80.75
```

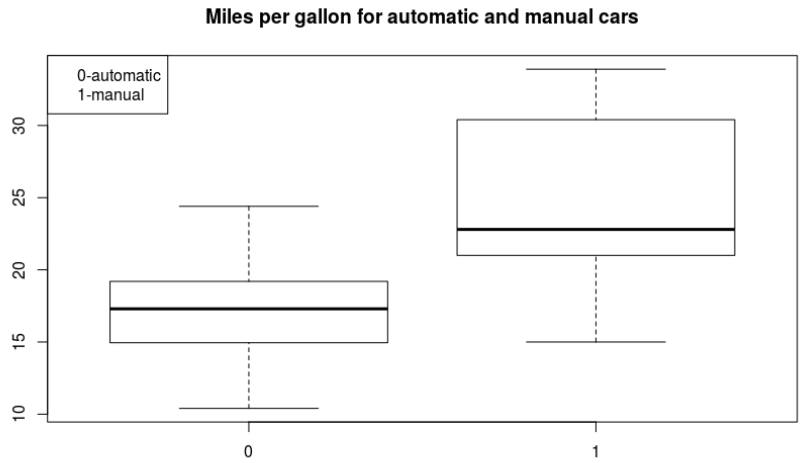
  - **Question:** Is there a difference between the results of the two companies?
  - p-value = 0.1836

FDA for numerical data

- (Q2, non-parametric) Do the data in polulation 1 and those in population 2 come from the same distribution? Use a *Wilcoxon Signed-Rank test* if the data are paired; else, use a *Wilcoxon Rank-Sum test*

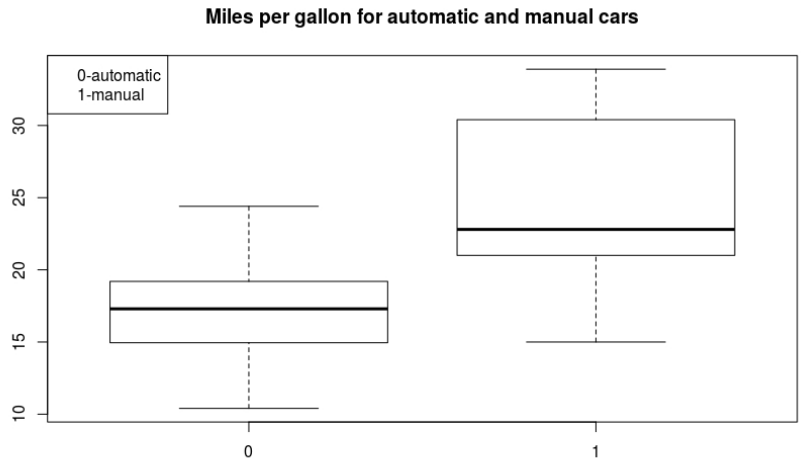
FDA for numerical data

- (Q2, non-parametric) Do the data in polulation 1 and those in population 2 come from the same distribution? Use a *Wilcoxon Signed-Rank test* if the data are paired; else, use a *Wilcoxon Rank-Sum test*
  - **Ex:** From data set 'mtcars'. We want to know if the gas mileage data of manual and automatic transmissions in 'mtcars' have identical data distribution.



FDA for numerical data

- (Q2, non-parametric) Do the data in polulation 1 and those in population 2 come from the same distribution? Use a *Wilcoxon Signed-Rank test* if the data are paired; else, use a *Wilcoxon Rank-Sum test*
  - **Ex:** From data set 'mtcars'. We want to know if the gas mileage data of manual and automatic transmissions in 'mtcars' have identical data distribution.



- p-value = 0.001871

FDA for categorical data

- (Q3) Is the variable spread according to given probabilities? Use a *chi-squared test for given probabilities*

FDA for categorical data

- (Q3) Is the variable spread according to given probabilities? Use a *chi-squared test for given probabilities*
  - **Ex:** Given the amount of murders per day in the week, decide if there is equal probability of murder happening in town.

Mon	Tue	Wed	Thur	Fri	Sat	Sun
53	42	51	45	36	37	45

FDA for categorical data

- (Q3) Is the variable spread according to given probabilities? Use a *chi-squared test for given probabilities*
  - **Ex:** Given the amount of murders per day in the week, decide if there is equal probability of murder happening in town.

Mon	Tue	Wed	Thur	Fri	Sat	Sun
53	42	51	45	36	37	45

  - X-squared $= 13.319$, df $= 6$, p-value $= 0.03824$

FDA for categorical data

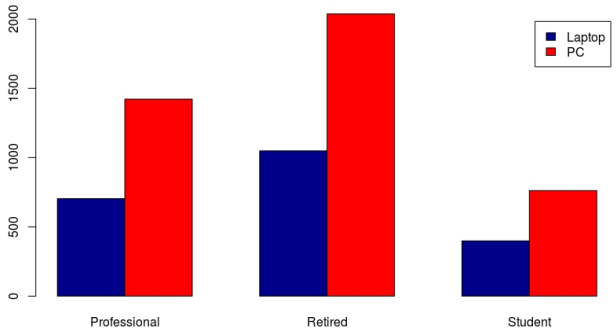
- (Q4) Are the variables independent? Use a *chi-square test for independence*

FDA for categorical data

- (Q4) Are the variables independent? Use a *chi-square test for independence*
  - **Ex:** A researcher of a computer magazine collected data from customers on type of device they are using (Laptop or PC) and their working status (Student, Professional, Retired).

Chi-square test for independence

Assumption: Counts in the table are more than 5.
Suppose you have analytically explored the data and got the following results:



	Professional	Retired	Student
Laptop	704	1049	399
PC	1422	2037	762

Chi-square test for independence

Assumption: Counts in the table are more than 5.

- **Question:** Is the type of device dependents upon working status?

Chi-square test for independence

Assumption: Counts in the table are more than 5.

- **Question:** Is the type of device dependents upon working status?
- X-squared = 0.66238, df = 2, p-value = 0.7181

Questions about relationships

Is there a linear relationship between A and B? Can we predict A from B? What is the best model to predict the response variable?

ANOVA: What is it?

- ANOVA is a generalization of the t-test

ANOVA: What is it?

- ANOVA is a generalization of the t-test
- t-test: difference between two groups? $\rightarrow$ ANOVA: difference between X groups?

ANOVA: What is it?

- ANOVA is a generalization of the t-test
- t-test: difference between two groups? $\rightarrow$ ANOVA: difference between X groups?
- H_0: $\mu_1 = ... = \mu_X$. H_a: not H_0

ANOVA: What is it?

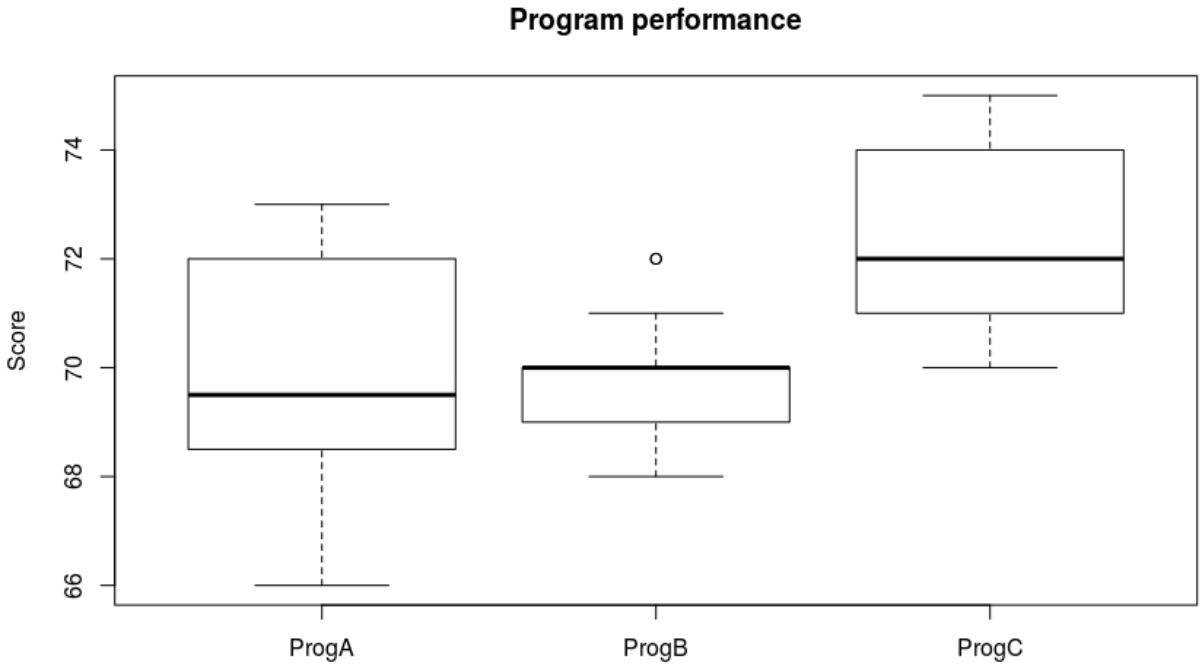
- ANOVA is a generalization of the t-test
- t-test: difference between two groups? $\rightarrow$ ANOVA: difference between X groups?
- H_0: $\mu_1 = ... = \mu_X$. H_a: not H_0
- Find out where the difference lies with Tukey's Honestly Significant Difference test

Example One-Way Anova

- **Ex:** Recently three artificial intelligence research groups developed their own computer program to simulate intelligent behaviour. A computer science professor assigns each of the three programs (A, B and C) to 20 of her students each to score.

Example One-Way Anova

- **Ex:** Recently three artificial intelligence research groups developed their own computer program to simulate intelligent behaviour. A computer science professor assigns each of the three programs (A, B and C) to 20 of her students each to score.



Example One-Way Anova

- **Question:** Is there difference between the performance of the programs?

Example One-Way Anova

- **Question:** Is there difference between the performance of the programs?
- Anova results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Program	2	88.73	41.87	16.81	1.83e-06
Residuals	57	142.00	2.49		

Example One-Way Anova

- **Question:** Is there difference between the performance of the programs?
- Anova results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Program	2	88.73	41.87	16.81	1.83e-06
Residuals	57	142.00	2.49		

- Tukey's Honestly Significant Difference test results:

	diff	lwr	upr	p adj
ProgB-ProgA	-0.2	-1.401096	1.001096	0.9154357
ProgC-ProgA	2.4	1.198904	3.601096	0.0000339
ProgC-ProgB	2.6	1.398904	3.801096	0.0000080

ANOVA: In other words...

- ...We can pose the research question as such: "Does Y significantly differ by X?", where Y is numerical and X is categorical
- So: looks a lot like regression

Regression

- "Can we predict Y from X (and Z)?", where Y is numerical and:

Regression

- "Can we predict Y from X (and Z)?", where Y is numerical and:
- X is numerical: Linear regression

Regression

- "Can we predict Y from X (and Z)?", where Y is numerical and:
- X is numerical: Linear regression
- X is categorical: +/- ANOVA

Regression

- "Can we predict Y from X (and Z)?", where Y is numerical and:
- X is numerical: Linear regression
- X is categorical: +/- ANOVA
- X is numerical, Z is categorical: ANCOVA

Regression

- "Can we predict Y from X (and Z)?", where Y is numerical and:
- X is numerical: Linear regression
- X is categorical: +/- ANOVA
- X is numerical, Z is categorical: ANCOVA
- X is transformed to X^2: Quadratic regression

- Special case if Y is binary: Logistic regression

Writing a report

Introduction

Let's introduce our dataset.

- What is our population?

Introduction

Let's introduce our dataset.

- What is our population?
- What is our sample (size)?

Introduction

Let's introduce our dataset.

- What is our population?
- What is our sample (size)?
- What are our variables? What type are they?

Introduction

Let's introduce our dataset.

- What is our population?
- What is our sample (size)?
- What are our variables? What type are they?
- How were the data collected?

Introduction

Let's introduce our dataset.

- What is our population?
- What is our sample (size)?
- What are our variables? What type are they?
- How were the data collected?
- Research question: e.g. "Can we predict [y] from [x]?"

Exploratory analysis

Let's take a look at our data.

- What do our distributions look like?

Exploratory analysis

Let's take a look at our data.

- What do our distributions look like?
- Do there seem to be relevant relationships between variables?

Formal analysis

Let's do relevant research.

1. Choose relevant test based on (I) type of data, (II) research question and (III) whether the assumptions are justified.

Formal analysis

Let's do relevant research.

1. Choose relevant test based on (I) type of data, (II) research question and (III) whether the assumptions are justified.
2. Report relevant results. These include p-values, test statistics (R^2, t, ...) and possibly estimates (e.g. of β_0).

Conclusion

Let's draw a conclusion.

- Just draw a conclusion based on the hypotheses and the p-values. That's it. Let's ignore any problems.

Discussion

Let's discuss our research.

- Those problems from the other slide...

Discussion

Let's discuss our research.

- Those problems from the other slide...
- Are there problems with the initial research? Not enough data? Data not representative? Third variable problems? Etc. etc.

Discussion

Let's discuss our research.

- Those problems from the other slide...
- Are there problems with the initial research? Not enough data? Data not representative? Third variable problems? Etc. etc.
- Are there problems with your analysis? Any chance the assumptions don't hold? What are the possible impacts on the results? Etc. etc.

SPAM

We're still looking for members! Questions and applications: Send a mail to studcee@svcover.nl

$\sim$ *Tokei wa tanoshidesu!!!* $\sim$