DATA HANDLING AND ASSUMPTIONS

Making the Most of Your Data

Erik Kusch

erik.kusch@au.dk

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

25/03/2020

Why Care?

Biostatisticians often use 70% of their time to handle data and just 30% to actually analyse it.

Why care?

- Proper data collection and data handling ensure accurate results
- Proper data collection cuts down on data handling time
- Proper data handling will make reproducing an analysis much easier

What to consider?

- Which data format to use
- What kind of data to record
- How data values are recorded/stored
- What kind of data values are feasible

Why Care?

Biostatisticians often use 70% of their time to handle data and just 30% to actually analyse it.

Why care?

- Proper data collection and data handling ensure accurate results
- Proper data collection cuts down on data handling time
- Proper data handling will make reproducing an analysis much easier

What to consider?

- Which data format to use
- What kind of data to record
- How data values are recorded/stored
- What kind of data values are feasible

Why Care?

Biostatisticians often use 70% of their time to handle data and just 30% to actually analyse it.

Why care?

- Proper data collection and data handling ensure accurate results
- Proper data collection cuts down on data handling time
- Proper data handling will make reproducing an analysis much easier

What to consider?

- Which data format to use
- What kind of data to record
- How data values are recorded/stored
- What kind of data values are feasible

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Data Recording

Guidelines for data recording:

- When collecting categorical data, know what values the variables are allowed to take
- When collecting continuous data, know which range the variable values can fall into
- Make sure everyone involved in data collection is on the same page
- Make regular back-ups of your data set

I recommend:

- Preparing content-aware excel files for data entry
  - Only allow pre-defined values to be entered
  - Need some excel macro writing
- Using a cloud-service featuring version control for data storage

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.
→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.
→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!
→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.
→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

$\rightarrow$ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

$\rightarrow$ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

$\rightarrow$ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

$\rightarrow$ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

→ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

→ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

→ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

→ Reduces storage space and chances for errors.

Common Issues

The Decimals

Always use a *dot* to indicate decimals.

$\rightarrow$ It is the standard in science.

To `NA` Or Not To `NA`?

Never enter `NA` *values manually* into your data.

$\rightarrow$ They cause problems in `R`.

Entering `0`?

If a `0` value *has meaning* in your set-up, \textit{enter] it!

$\rightarrow$ Empty cells are interpreted as `NA` by `R`.

Redundancy Or Sparsity?

Don't clutter data with unnecessary data records.

$\rightarrow$ Reduces storage space and chances for errors.

Data Storing

R works very well with:

- excel files (*.xls, .xlsx, .csv*)

→ Easiest to handle outside of R, most storage-heavy

→ Make sure to provide co-workers with a master file before data collection to avoid cell formatting issues on different computers

- text files (*.txt*)

→ Difficult to handle outside of R, easy on storage

→ I advise against using these, formatting issues are far too common

- RDS files (*.rds*)

→ Impossible to handle outside of R, easy on storage

→ I **highly** recommend using these for every step of your work past initial data recording

Data Storing

R works very well with:

- excel files (*.xls, .xlsx, .csv*)

→ Easiest to handle outside of R, most storage-heavy

→ Make sure to provide co-workers with a master file before data collection to avoid cell formatting issues on different computers

- text files (*.txt*)

→ Difficult to handle outside of R, easy on storage

→ I advise against using these, formatting issues are far too common

- RDS files (*.rds*)

→ Impossible to handle outside of R, easy on storage

→ I **highly** recommend using these for every step of your work past initial data recording

Data Storing

R works very well with:

- excel files (*.xls, .xlsx, .csv*)

→ Easiest to handle outside of R, most storage-heavy

→ Make sure to provide co-workers with a master file before data collection to avoid cell formatting issues on different computers

- text files (*.txt*)

→ Difficult to handle outside of R, easy on storage

→ I advise against using these, formatting issues are far too common

- RDS files (*.rds*)

→ Impossible to handle outside of R, easy on storage

→ I **highly** recommend using these for every step of your work past initial data recording

Data Storing

R works very well with:

- excel files (*.xls, .xlsx, .csv*)

→ Easiest to handle outside of R, most storage-heavy

→ Make sure to provide co-workers with a master file before data collection to avoid cell formatting issues on different computers

- text files (*.txt*)

→ Difficult to handle outside of R, easy on storage

→ I advise against using these, formatting issues are far too common

- RDS files (*.rds*)

→ Impossible to handle outside of R, easy on storage

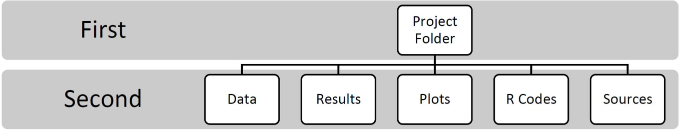
→ I **highly** recommend using these for every step of your work past initial data recording

Data Structure

I recommend a structure like the one below with at least two hierarchy levels.

The only files allowed in your first hierarchy level are:

- R master file
- Manuscript master file



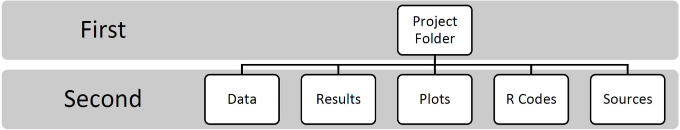
Additionally, make sure to **back-up your project folder frequently** and use **version control** on it.

Data Structure

I recommend a structure like the one below with at least two hierarchy levels.

The only files allowed in your first hierarchy level are:

- R master file
- Manuscript master file



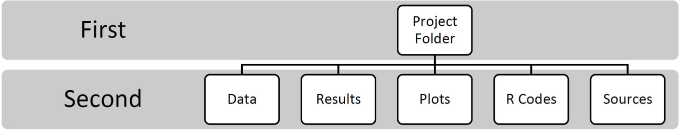
Additionally, make sure to **back-up your project folder frequently** and use **version control** on it.

Data Structure

I recommend a structure like the one below with at least two hierarchy levels.

The only files allowed in your first hierarchy level are:

- R master file
- Manuscript master file



Additionally, make sure to **back-up your project folder frequently** and use **version control** on it.

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

The README File

Using the **README file**, one can identify what information is contained within the data set and thus decide:

- What type/class a data record should be of
- Which variables may be redundant
- Which data records exceed their variable-specific feasible thresholds
- Where to get comparative data sets from

Data Mining should then focus on:

- *Identifying problems* within the data records
- *Explorative* data analyses

Mining in R - Numbers or Visualisations?

For data mining, one may wish to enlist the use of Descriptive Statistics & Data Visualization:

Descriptive Statistics:

- `summary()`
- `str()`
- iterative sub-setting and inspection

Data Visualizations:

- Histograms (`hist()`)
- Scatter plots (`ggplot2 Package`)

The `R` package `skimr` offers the function `skim()` to do all of this in one line of code.

Holistic data mining is best **achieved using a combination** of data visualizations tools and parameters of descriptive statistics!

Mining in `R` - Numbers or Visualisations?

For data mining, one may wish to enlist the use of Descriptive Statistics & Data Visualization:

Descriptive Statistics:

- `summary()`
- `str()`
- iterative sub-setting and inspection

Data Visualizations:

- Histograms (`hist()`)
- Scatter plots (`ggplot2 Package`)

The `R` package `skimr` offers the function `skim()` to do all of this in one line of code.

Holistic data mining is best **achieved using a combination** of data visualizations tools and parameters of descriptive statistics!

Mining in `R` - Numbers or Visualisations?

For data mining, one may wish to enlist the use of Descriptive Statistics & Data Visualization:

Descriptive Statistics:

- `summary()`
- `str()`
- iterative sub-setting and inspection

Data Visualizations:

- Histograms (`hist()`)
- Scatter plots (`ggplot2 Package`)

The `R` package `skimr` offers the function `skim()` to do all of this in one line of code.

Holistic data mining is best **achieved using a combination** of data visualizations tools and parameters of descriptive statistics!

Mining in `R` - Numbers or Visualisations?

For data mining, one may wish to enlist the use of Descriptive Statistics & Data Visualization:

Descriptive Statistics:

- `summary()`
- `str()`
- iterative sub-setting and inspection

Data Visualizations:

- Histograms (`hist()`)
- Scatter plots (`ggplot2 Package`)

The `R` package `skimr` offers the function `skim()` to do all of this in one line of code.

Holistic data mining is best **achieved using a combination** of data visualizations tools and parameters of descriptive statistics!

Mining in `R` - Numbers or Visualisations?

For data mining, one may wish to enlist the use of Descriptive Statistics & Data Visualization:

Descriptive Statistics:

- `summary()`
- `str()`
- iterative sub-setting and inspection

Data Visualizations:

- Histograms (`hist()`)
- Scatter plots (`ggplot2 Package`)

The `R` package `skimr` offers the function `skim()` to do all of this in one line of code.

Holistic data mining is best **achieved using a combination** of data visualizations tools and parameters of descriptive statistics!

Recording Data Collection - The README File

Documenting data recording is just as important as proper data collection!

To do so, one usually uses a **README** file containing the following:

- Project Name and Summary
- Primary contact information
- Your name and title (if you aren't the primary contact)
- Other people working on the project
- Location of data and supporting info
- Organization and naming conventions used for the data
- Any previous work on the project and where its located
- Funding information

This file is always **saved in conjunction with the actual data set**!

Recording Data Collection - The README File

Documenting data recording is just as important as proper data collection!

To do so, one usually uses a **README** file containing the following:

- Project Name and Summary
- Primary contact information
- Your name and title (if you aren't the primary contact)
- Other people working on the project
- Location of data and supporting info
- Organization and naming conventions used for the data
- Any previous work on the project and where its located
- Funding information

This file is always **saved in conjunction with the actual data set**!

Recording Data Collection - The README File

Documenting data recording is just as important as proper data collection!

To do so, one usually uses a **README** file containing the following:

- Project Name and Summary
- Primary contact information
- Your name and title (if you aren't the primary contact)
- Other people working on the project
- Location of data and supporting info
- Organization and naming conventions used for the data
- Any previous work on the project and where its located
- Funding information

This file is always **saved in conjunction with the actual data set**!

Data Sharing

Open science conduct is essential and you should (read *have to* as a student/employee of Aarhus University) share your data & coding to ensure **reproducibility** of your work:

Peer-to-Peer:
- Raw data
- Code
- You may just as well point your peers to public repositories

Public:
- Raw data
- Code
- Html visualizations (`shiny`, `mapview`)
- Websites

Aarhus Guideline: Store data on the Ecoinf/Biochange data server. NOT on the computational server. Read more here.

Data Sharing

Open science conduct is essential and you should (read *have to* as a student/employee of Aarhus University) share your data & coding to ensure **reproducibility** of your work:

Peer-to-Peer:

- Raw data
- Code
- You may just as well point your peers to public repositories

Public:

- Raw data
- Code
- Html visualizations (`shiny`, `mapview`)
- Websites

Aarhus Guideline: Store data on the Ecoinf/Biochange data server. NOT on the computational server. Read more here.

Data Sharing

Open science conduct is essential and you should (read *have to* as a student/employee of Aarhus University) share your data & coding to ensure **reproducibility** of your work:

Peer-to-Peer:
- Raw data
- Code
- You may just as well point your peers to public repositories

Public:
- Raw data
- Code
- Html visualizations (`shiny`, `mapview`)
- Websites

Aarhus Guideline: Store data on the Ecoinf/Biochange data server. NOT on the computational server. Read more here.

Data Sharing

Open science conduct is essential and you should (read *have to* as a student/employee of Aarhus University) share your data & coding to ensure **reproducibility** of your work:

Peer-to-Peer:
- Raw data
- Code
- You may just as well point your peers to public repositories

Public:
- Raw data
- Code
- Html visualizations (`shiny`, `mapview`)
- Websites

Aarhus Guideline: Store data on the Ecoinf/Biochange data server. NOT on the computational server. Read more here.

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Assumptions

Statistical tests rely on individual *statistical assumptions*. Most prominent:

- **Normality**: Data follow a normal distribution
- **Randomness**: Data are truly random
- **Independence**: Data are independent
- **Homogeneity of variances**: Data from separate groups have same variance
- **Linearity**: Data have linear relationship

Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

The QQ Plot In Theory

Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

- Base assumption: The data is normally distributed
- If p-value $<$ chosen significance level, the data is **not** normally distributed
- Very sensitive to sample size

The QQ Plot In Theory

Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

- Base assumption: The data is normally distributed
- If p-value $<$ chosen significance level, the data is **not** normally distributed
- Very sensitive to sample size

The QQ Plot In Theory

- Method for comparing two probability distributions by plotting their quantiles against each other
- If the two distributions being compared are similar, the plot will show the line $y = x$.
- Compare the data distribution to the normal distribution

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

$\rightarrow$ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

→ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

$\rightarrow$ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

$\rightarrow$ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

→ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

→ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

→ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

$\rightarrow$ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

$\rightarrow$ Fixing this *after data collection* is **almost impossible**!

Independence

Theory:

- Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable).
- A dependence is a connection between/within the data.
- The assumption of independence relies on the absence of any connection in your data that haven't been accounted for in your approach (accounting for it is difficult).

Independent data:

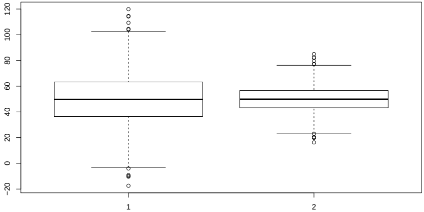
- *Between Groups*
  Groups of data records should be pulled from different individuals.
- *Within Groups*
  Data values within the same group are not to influence one another.
- *Within Individuals*
  Data values recorded for one individual should not influence each other. This is often an issue with repeated measurement approaches.

→ Fixing this *after data collection* is **almost impossible**!

Homogeneity of Variances

Particularly important for t-Tests and ANOVAs

- **Assumption**: Data from separate groups have same variance
- **Test**: `leveneTest()` in the `car` package.

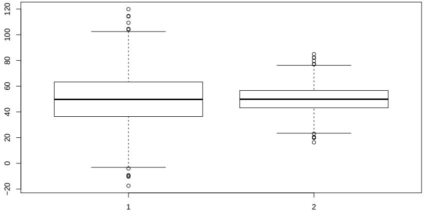



```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1     337 <2e-16 ***
##       1998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Homogeneity of Variances

Particularly important for t-Tests and ANOVAs

- **Assumption**: Data from separate groups have same variance
- **Test**: `leveneTest()` in the `car` package.




```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1     337 <2e-16 ***
##       1998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```