Introduction to Statistics in R

Workshop 1 – June 17, 2025

CCM

Introduction to Statistics in R: Part 1

A two-part course on using R for statistical analysis for SickKids researchers and trainees. Part 2 will follow on Tuesday June 24.

Topics for Part 1 include:

- Data exploration and cleaning
- Data visualization
- Correlations, t-tests, and ANOVAs
- Linear regressions

Prerequisites: familiarity with RStudio

To register please RSVP at:
https://ccm20250617.eventbrite.com

Hands-on Bioinformatics Tutorials for Biologists

Tuesday June 17, 2025
9 AM – 12 PM

In person in Multimedia Room, PGCRL 3rd floor, or on Zoom

Organized by the Centre for Computational Medicine (ccm.sickkids.ca) at the SickKids Research Institute with support from other groups:

SickKids | Research Training Centre — www.sickkids.ca/research

Digital Research Alliance of Canada — alliancecan.ca

Compute Ontario — computeontario.ca

CCM Overview

Note that prior knowledge of R is recommended to get the most out of this workshop series.

Workshop 1	Data exploration and introductory statistics
Workshop 2	Generalized and hierarchical regressions

Check out other workshop series hosted by CCM at https://ccm.sickkids.ca/bioinformatics-training/

Table of contents

01

Data exploration and cleaning

Framingham Heart Study

One of the longest prospective epidemiological studies of cardiovascular disease and its risk factors.

Began in 1948 in Framingham, MA.

- Initially enrolled 5209 men and women aged 29-62 years old.
- Followed them over time, with assessments every 2 years.
- Data collection for the original cohort ended in 2014.
- Over time, the researchers incorporated offspring and their spouses into the study.
- Today, they have data from three generations of participants as well as more ethnically diverse cohorts.

Framingham Heart Study

Examinations included:

- Detailed medical history
- Physical exams
- Lab tests
- Lifestyle and habits
- Noninvasive imaging

Our dataset: subset of the FHS from Kaggle

- 4000+ records and 16 variables.

(https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data)

Variable	Description	Class/type
male	Sex (male or female)	Binary
age	Age	Continuous
education	Education (0-11 years, HS or GED, some uni, uni grad+)	Ordinal/categorical
currentSmoker	Whether the patient is currently a smoker	Binary
cigsPerDay	Number of cigarettes smoked per day, on average	Continuous
BPMeds	Whether the patient is on blood pressure medication	Binary
prevalentStroke	Whether the patient previously had a stroke	Binary
prevalentHyp	Whether the patient is hypertensive	Binary
diabetes	Whether the patient has diabetes	Binary
totChol	Total cholesterol level	Continuous
sysBP	Systolic blood pressure	Continuous
diaBP	Diastolic blood pressure	Continuous
BMI	Body mass index	Continuous
heartRate	Heart rate	Continuous
glucose	Glucose level	Continuous
TenYearCHD	10-year risk of coronary heart disease	Binary

Data exploration

Some common and/or important data cleaning steps include making sure that:

1. All variables are in the right format and/or correctly classified.
2. Duplicate or missing data are dealt with.
3. Variables of interest have been transformed appropriately.
4. Outliers are handled.
5. Data are validated.

Data exploration

1. Look at the data and its structure – are all the variables in the right format?

```
str()
```

2. Tabulate data

`table(df$column)` or `table(df$column1, df$column2, ...)`

3. Modify variables

```
mutate(column = function(column))
mutate(df, across(columns, function))
```

4. Transform variables – center, scale, log-transform

```
log(x, base = exp(1))
scale(x, center = TRUE, scale = TRUE)
```

Data exploration

5. Subset dataframes

`filter()` rows according to some condition

`select()` columns to keep

```
subset(df, subset = row_vals == “abc”, cols = X)
```

6. Combine vectors and dataframes

`cbind()` to join by columns

`rbind()` to join by rows

`merge()` to join dataframes

7. Reshape dataframes

`pivot_wider()` each row is a participant/patient

`pivot_longer()` each row is an observation, with multiple rows per participant

Data exploration

8. Handy functions, pipes, and operators

`ifelse()` for conditional element selection
`case_when()` as a vectorized version of multiple `ifelse()` statements, uses formula notation (more on that later)

`%>%` passes the result of the function on the left as input to the next function
`%in%` to compare vectors

Create a summary table

Use `summarize()` to create a dataframe with summaries of desired variables. For example, what if you're only interested in the educational attainment and smoking habits of people who go on to develop heart disease?

	TenYearCHD	education_class	n_participants	age	currentSmoker	prop_smokers	cigsPerDay
1	0	< High School	1397	51	645	0.46	8.39
2	0	College Graduate	403	47	204	0.51	9.09
3	0	High School Graduate	1106	47	598	0.54	9.67
4	0	Some College	599	48	275	0.46	7.70
5	1	< High School	323	56	160	0.50	9.93
6	1	College Graduate	70	53	36	0.51	11.97
7	1	High School Graduate	147	52	81	0.55	11.14
8	1	Some College	88	53	46	0.52	10.74

02

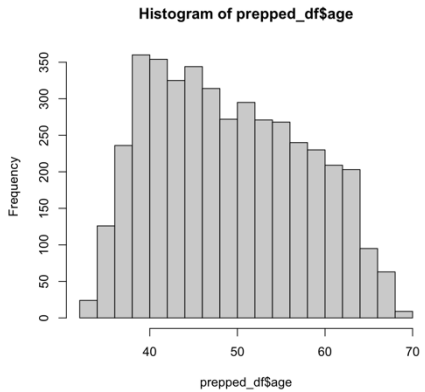
Data visualization

Base R vs ggplot2

Base R

Useful in a pinch but doesn't generate the prettiest figures.

`plot()` and `hist()` are useful for taking a quick look at your data.



ggplot2

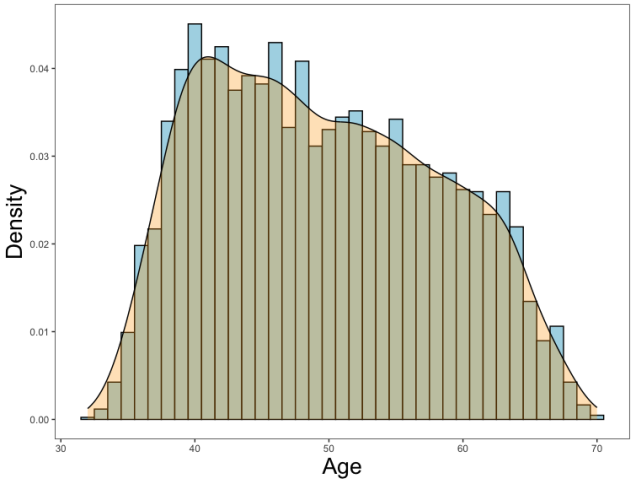
ggplot2 is a package based on the grammar of graphics.

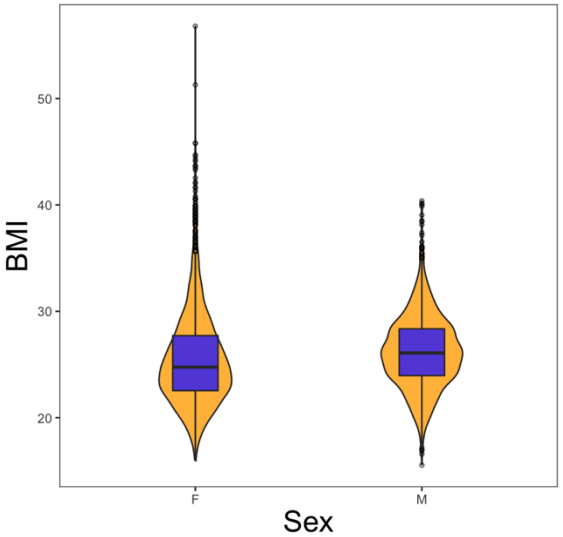
You can create your figure with some basic components:

1. **Data**
2. **Mapping**
3. **Layers**
4. Scales
5. Facets
6. Coordinates
7. Theme

Distributions

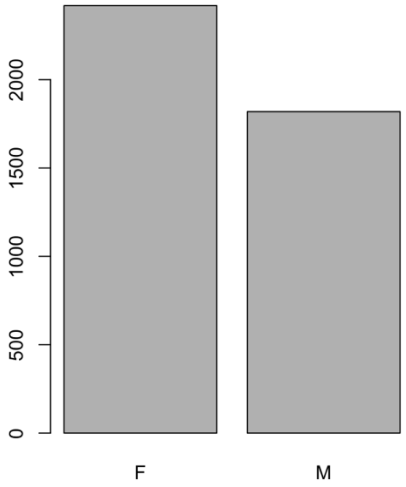
Look at the distribution of data using histograms, density plots, boxplots, and violin plots (among others).

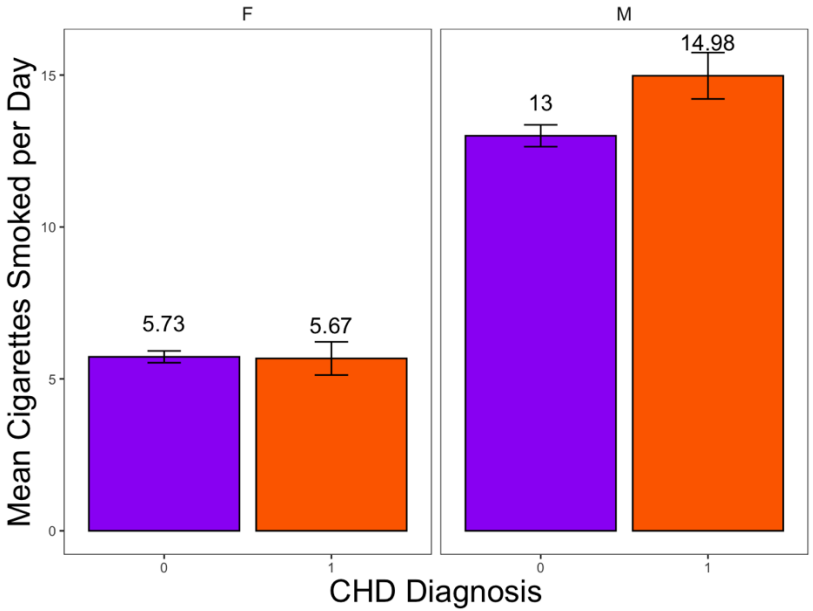




Bar Charts

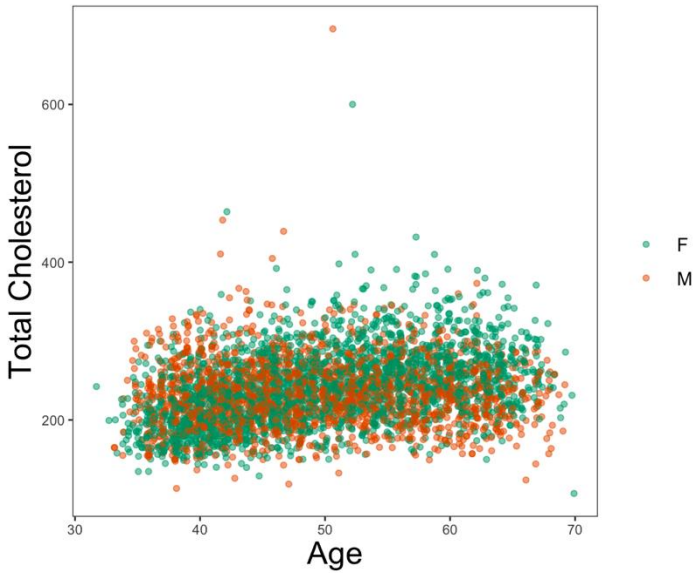
Compare groups using bar charts.

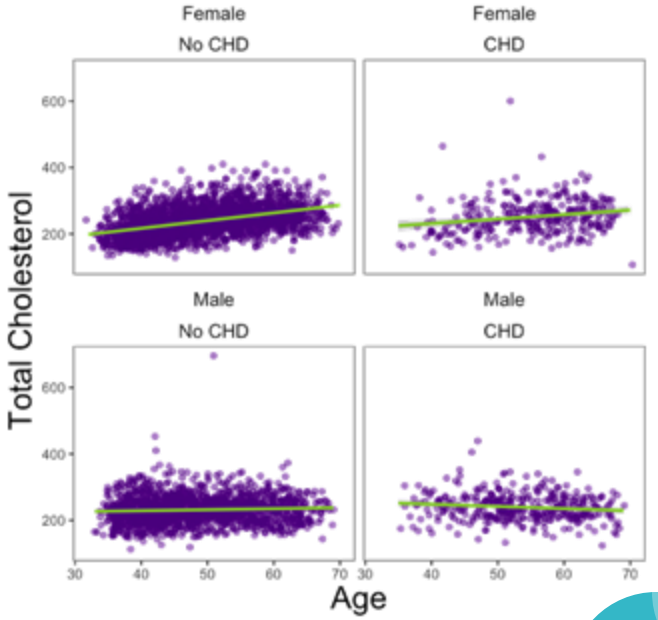




Scatterplots

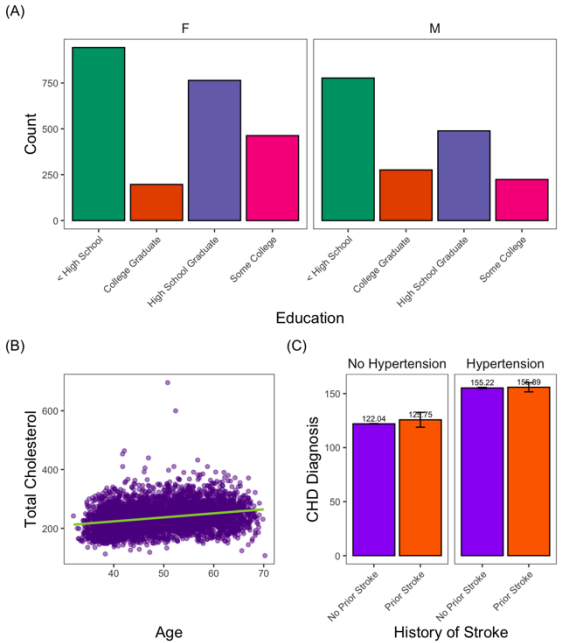
Plot the relationship between two continuous variables using scatterplots.





Panel figures

You might have multiple plots that you want to put into a single figure with panels. You can do this using the `patchwork()` package.

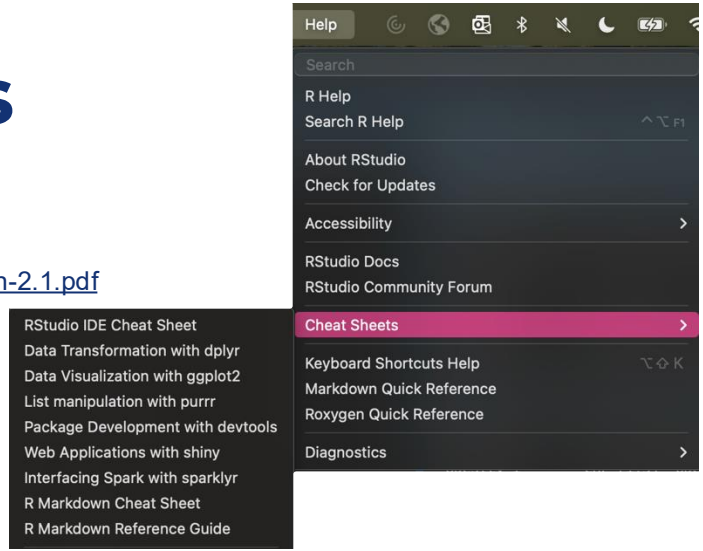


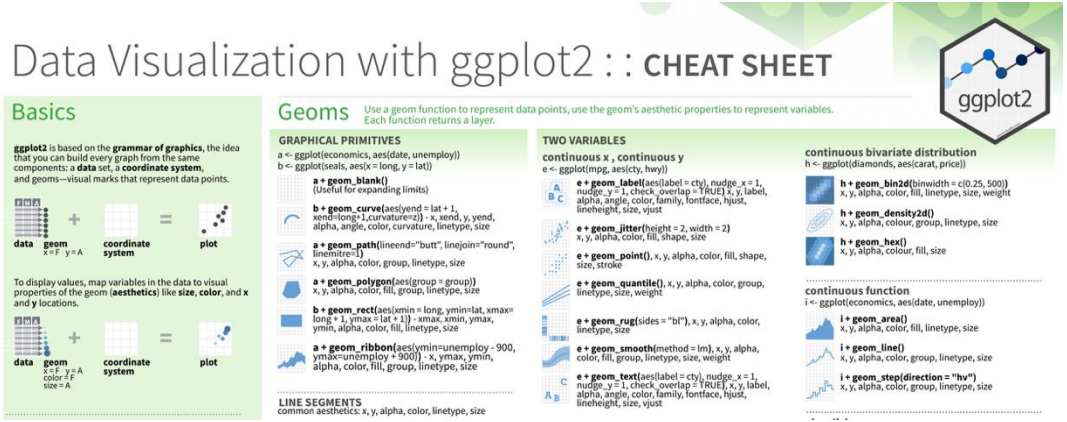
Resources

ggplot2 cheat sheet:
https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf

color palette reference:
https://sape.inf.usi.ch/quick-reference/ggplot2/colour

R Studio cheat sheets: Help > Cheat Sheets > …





03

Correlations and t-tests

Pearson Correlations

A Pearson correlation measures the strength of the linear relationship between two continuous variables.

```
cor(x, y, use = “complete.obs”, method = “pearson”)
```

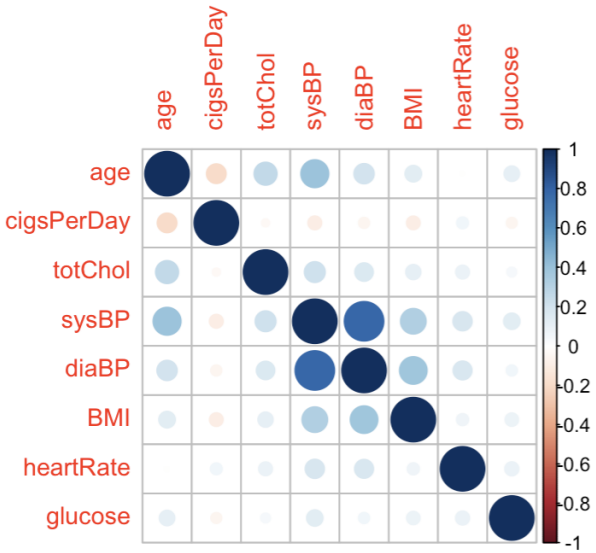
- `use = “complete.obs”` removes rows where at least one of the values is NA/missing and this is not specified, could also specify `“na.or.complete”` (output is NA if there are no complete cases)

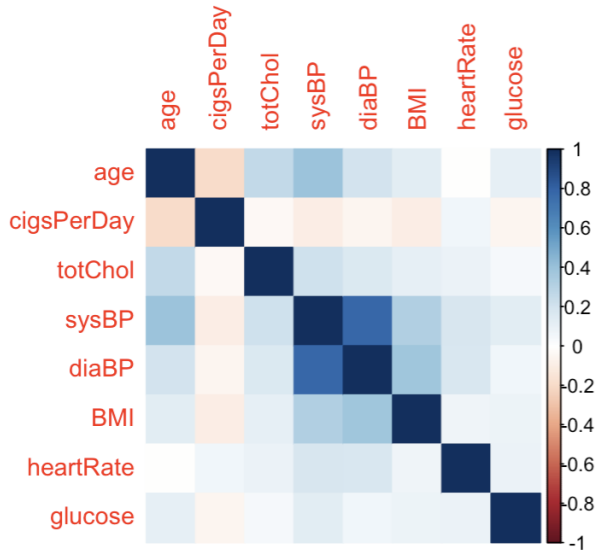
```
cor.test(x, y, use = “complete.obs”, method = “pearson”)
```

- provides r, p-value, and confidence intervals.

Correlation Plots

Correlation plots allow you to visualize correlations between all the numeric variables in your dataframe.



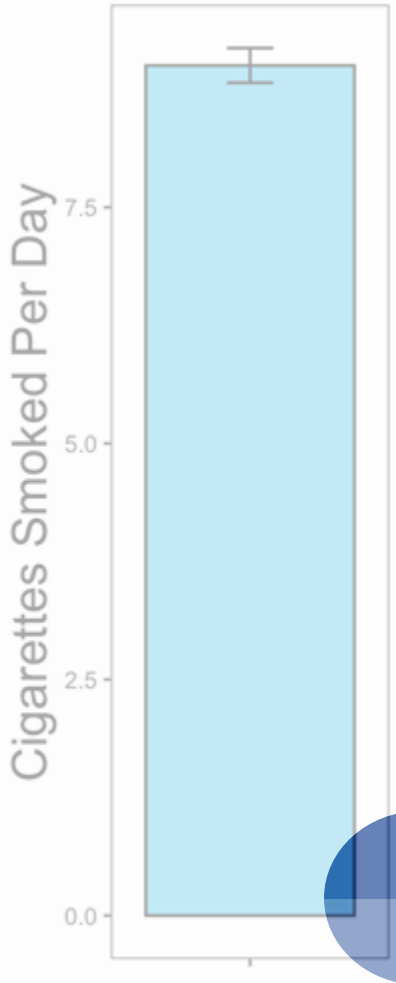


One-way T-tests

Compare your sample mean to the population mean, 0, or another value.

```
t.test(df$outcome, mu = 0)
```

Are the number of cigarettes smoked by our sample significantly different from 0? Different from an average of 10 cigarettes smoked per day?

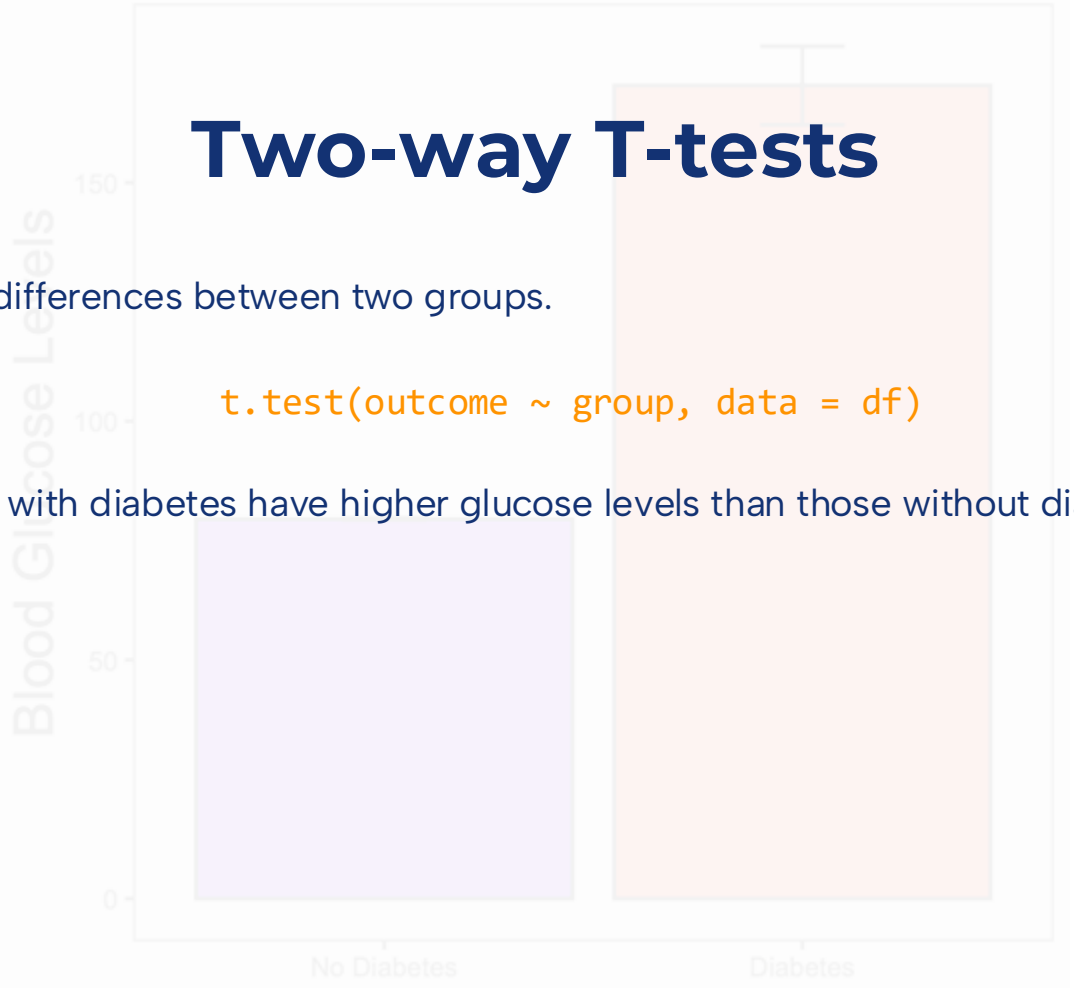


Two-way T-tests

Compare differences between two groups.

```
t.test(outcome ~ group, data = df)
```

Do people with diabetes have higher glucose levels than those without diabetes?



04

Linear regressions

Simple Linear Regressions

Formula syntax for conducting linear regressions in R.

Syntax	Description
`outcome ~ predictor1 + predictor2 + …`	Main effects of predictor 1, predictor 2, etc.
`outcome ~ predictor1 + predictor2 + predictor1:predictor2`	Main effects of predictor 1 and predictor 2, and the interaction between predictor 1 and 2
`outcome ~ predictor1 * predictor2`	Main effects of predictor 1 and predictor 2, and the interaction between predictor 1 and predictor 2
`outcome ~ predictor1 + predictor1:predictor2`	Main effect of predictor 1 only, and the interaction between predictor 1 and predictor 2
`outcome ~ predictor1:predictor2`	The interaction between predictor 1 and predictor 2 only

Main effect models

Examine relationships between predictors and outcomes of all classes.

Are age and blood pressure related to blood glucose levels?

- Both predictors are continuous
- Outcome variable is also continuous

Are smoking status and use of BP medications related to cholesterol levels?

- Both predictors are categorical/binary
- Outcome variable is continuous

Syntax	Description
`outcome ~ predictor1 + predictor2 + …`	Main effects of predictor 1, predictor 2, etc.

Interaction effects

How is the relationship between a predictor and an outcome *moderated* by another variable?

How are sex and cholesterol levels related to blood pressure? Does the relationship between cholesterol levels and blood pressure vary between males and females?

- The predictors are categorical/binary and continuous
- Outcome variable is continuous

Syntax	Description
`outcome ~ predictor1 + predictor2 + predictor1:predictor2`	Main effects of predictor 1 and predictor 2, and the interaction between predictor 1 and 2
`outcome ~ predictor1 * predictor2`	Main effects of predictor 1 and predictor 2, and the interaction between predictor 1 and predictor 2
`outcome ~ predictor1 + predictor1:predictor2`	Main effect of predictor 1 only, and the interaction between predictor 1 and predictor 2
`outcome ~ predictor1:predictor2`	The interaction between predictor 1 and predictor 2 only

Resources

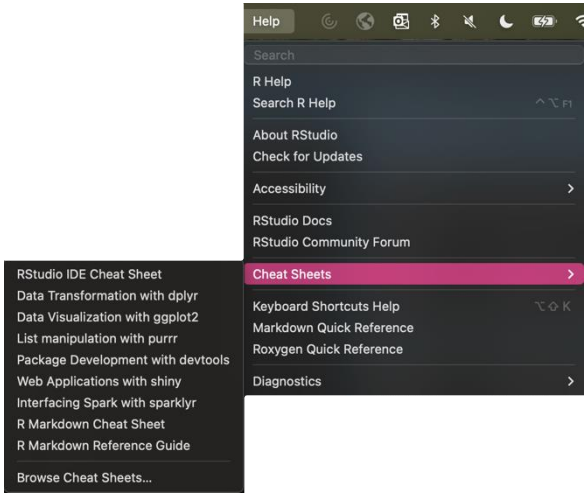
Plotting main effects:

https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_marginal_effects.html

Plotting interactions:

https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_interactions.html

More cheat sheets on R Studio:



Thanks!

Do you have any questions?

shireen.parimoo@sickkids.ca

CREDITS: This presentation template was created by **Slidesgo,** and includes icons by **Flaticon**, and infographics & images by **Freepik**