Agenda for the session

Figure 1: Agenda for the session

Research questions and corresponding statistical tests

The most popular research questions include:

1. whether two variables (n = 2) are correlated (i.e., associated)

Ans: Correlation test between two variables

2. whether multiple variables (n > 2) are correlated

Ans: Correlation matrix between multiple variables

3. whether two groups (n = 2) of samples differ from each other

Ans: Comparing the means of two groups:

- `Student's t-test (parametric)`
- `Wilcoxon rank test (non-parametric)`

4. whether multiple groups (n >= 2) of samples differ from each other

Ans: Comparing the means of more than two groups

```
- ANOVA test (analysis of variance, parametric): extension of t-test to compare more than two groups.
- Kruskal-Wallis rank sum test (non-parametric): extension of Wilcoxon rank test to compare more than t
```

Tesing of variablility

5. whether the variability of two samples differ

Ans: Comparing the variances:

```
 - Comparing the variances of two groups: F-test (parametric)
 - Comparison of the variances of more than two groups: Bartlett's test (parametric), Levene's test (pa
 -and Fligner-Killeen test (non-parametric)
```

Statistical test requirements

Many of the statistical procedures including correlation, regression, t-test, and analysis of variance assume some certain characteristic about the data. Generally they assume that:

- **the data are normally distributed,**
- **and the variances of the groups to be compared are homogeneous (equal)**

These assumptions should be taken seriously to draw reliable interpretation and conclusions of the research. These tests - correlation, t-test and ANOVA - are called parametric tests, because their validity depends on the distribution of the data.

Pretest for statistical testing

- For large samples, parametric test can be used
- Use **Shapiro-Wilk's** significance test comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality

A simple trick to attain normality!!

A statistically acceptable trick is to form a large ($n > 100$) sample for analysis. The rest is with your statistical analysis tool.

Normality Test in R

Normality and the other assumptions made by parametric tests should be taken seriously to draw reliable interpretation and conclusions of the research.

Packages required

- *dplyr* for data manipulation
- *ggpubr* for an easy ggplot2-based data visualization

Example

Choose the ToothGrowth data and test for normality:

"'{r} # Store the data in the variable my_data my_data <- ToothGrowth set.seed(1234) dplyr::sample_n(my_data, 10)

```
 For large sample these pretests are laxuaries.However, to be consistent, normality can be checked by v
```

```
## Normality test

*Approach*:  Use a significance test comparing the sample distribution to a normal one in order to asce

There are several methods for normality test such as Kolmogorov-Smirnov (K-S) normality test and Shapir
```

```{r}
shapiro.test(my_data$len)
```

Info: From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

##Visual methods Density plot and Q-Q plot can be used to check normality visually.

Density Plot

{r} library("ggpubr") ggdensity(my_data$len, main = "Density plot of tooth length", xlab = "Tooth length")

Q-Q Plot

{r} ggqqplot(my_data$len)

qqPlot()

It's also possible to use the function qqPlot() [in car package]:

{r} library("car") qqPlot(my_data$len)

Takeaway from the session

-Always form good research questions and hypothesis for analysis.

- In case of parametric tests, always make sure that the group of data to be analyzed follows normal distribution.
- If normality can not be maintained, then try non- parametric tests.
- Always use large samples for statistical analysis.