Research question

4 Types of research questions:

descriptive

Describe the teachers' perceptions of the newly implemented reading assessment program.

comparative

Are there differences in students' grades by gender (male vs. female)?

relationship

Is there a relationship between age and fitness level?

predictive

Do age, gender, and education predict income?

Informal data analysis

look at data in:

- QQ-plot (check for normality)
- histograms of dependent variables (check for normality)
- boxplots (to check for outliers)
- two way box-plot (if binary response variable)
- scatterplots (per predictive variable)

formal analysis:

definitions:

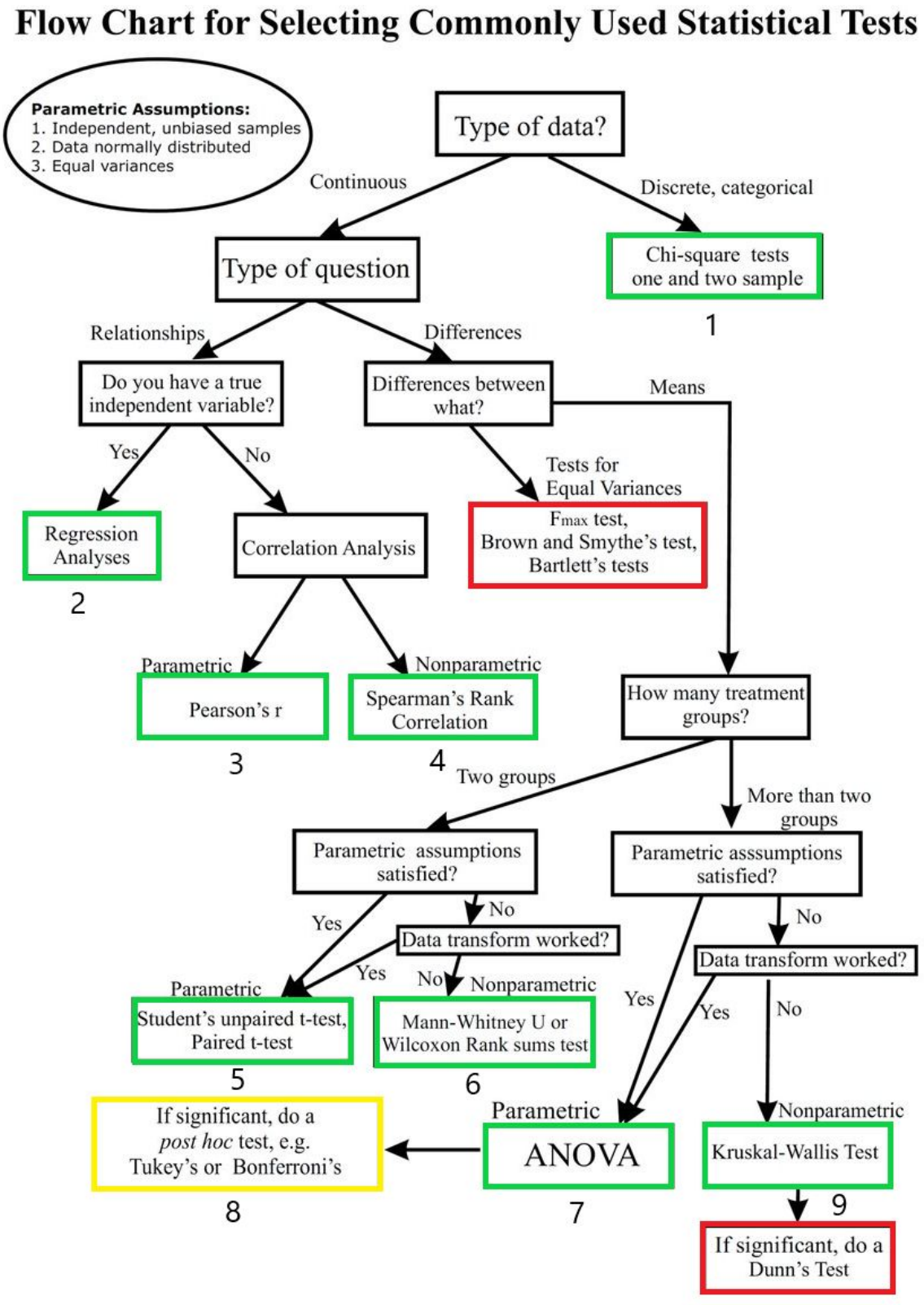
A parametric statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn.

Paired means that both samples consist of the same test subjects. A paired t-test is equivalent to a one-sample t-test. Unpaired means that both samples consist of distinct test subjects. An unpaired t-test is equivalent to a two-sample t-test.

Useful packages

```
>  install.packages("UsingR")
>  install.packages("MASS")
>  install.packages("ggpubr")
>  install.packages("car")
```

What test to use



1) Chi-square test

one sample

research question

Is there any significant difference between observed proportions and the expected proportions?

hypothesis

accept Null hypothesis: There is no significant difference between the observed and expected values. (p>=0.05)

accept Alternative hypothesis: There is a significant difference between the observed and expected values. (p<0.05)

Test

```
> x <- observed values
> p <- expected values
> chisq.test(x, p)
```

two sample

Research question

Are the row variables and the column variables of a contingency table significantly associated?

Hypothesis

accept Null hypothesis: The row and column variables of the contingency table are independent. (p>=0.05)

accept Alternative hypothesis: The row and column variables of the contingency table are dependent. (p<0.05)

Test

```
> x <- contingency table
> chisq.test(x) # x is frequency table, not relative frequency
```

2) Regression analysis

linear regression

Research question

What is the relationship between the predictor values and the response variable?

Procedure

Check if all variables are close to normality:

```
> plot(density(x)) #for all parameters
```

Make a scatter plot per predictor value to check if there's linear relation:

```
> x <- all predictor variable values
> y <- all response variable values
> scatter.smooth(x, y, main="", xlab="", ylab="")
```

Make a boxplot of all variables to check for outliers:

```
> boxplot(x, names="variable", show.names=T)
```

Look if there is correlation (value closer to 0 mean they're not correlated):

```
> cor(x,y)
```

Make a linear model:

```
> model <- lm(y~x1+...+xn, data = data)
```

Find the best fit:

```
#package "MASS" for stepAIC
new.model <- stepAIC(model, direction="both")
```

Check if the linear model is significant:

```
> summary(new.model)
```

Check the p value for the F-statistic:

- If p<0.05, there is at least one predictor value that is significant

Check the p values of each predictor variable:

- If p<0.05: changes in the predictor's value are related to changes in the response variable.
- If p>= 0.05: changes in the predictor are not associated with changes in the response.

Optional :check the t-values of each predictor variable:

- A larger *t-value* indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better.

Check the adjusted R-squared value:

- The R-squared value gives us the proportion of variation of the response variable that can be explained by the model. (closer to 1 is more explained by model)

Check the F-value:

- A higher F-value means a better fit.

check if the model is a good fit with model diagnostics (see extra):

```
> layout(matrix(c(1,2,3,4),2,2))
> plot(new.model)
```

visually represent the fitness:

```
> plot(y  ,new.model$fitted, xlab="", ylab="", main="Fitted vs
real")
```

Conclusion:

Explain the relationship between the predictor variables and the response variable.

logistic regression (binary response variable)

Research question

What is the relationship between the predictor values and the response variable?

Procedure 1: test and train data

Make histograms to check for normality of parameters:

```
> plot(density(x))  #for all parameters
```

Make a boxplot of all variables to check for outliers:

```
> boxplot(x, names="variable", show.names=T)
```

Make boxplot pairs to see if there is relationship between response and predictor variable:

```
# do this for all predictor variables
> predvar0 <- data$varname[data$resvar==0]
> predvar1 <- data$varname[data$resvar==1]
> boxplot(predvar0, predvar1, names=c("0", "1"))
```

Split data into training and testing

```
> train <- data[1:800,]
> test <- data[801:889,]
```

Make a logistic model:

```
> model <- glm(y~x1+...+xn, data = train,
family=binomial(link='logit'))
```

Find the best fit:

```
#package “MASS” for stepAIC
> new.model <- stepAIC(model, direction="both")
> summary(new.model)
```

test the trained model on test data:

```
fitted.results <-
predict(new.model,newdata=subset(test,select=c(var1, …,
varn)),type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0) #round off results
> misClasificError <- mean(fitted.results != test$resvar)
> print(paste('Accuracy',1-misClasificError))
```

Interpret the accuracy:

everything above 0.8 is a good fit.

Create an ROCR curve:

```
# package ROCR
> p <- predict(new.model,
newdata=subset(test,select=c(2,3,4,5,6,7,8)), type="response")
> pr <- prediction(p, test$resvar)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
```

Calculate area under curve for goodness of fit:

```
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
```

residual deviance test

```
> dev <- residual deviance of model
> dof <- degrees of freedom of model
> 1-pchisq(dev,dof)
```

accept null hypothesis (p>=0.05): data fits the proportions well

accept alternative hypothesis(p<0.05): reject null hypothesis, There is variation in the data that cannot be explained by neither the explanatory variables nor the inherent randomness in the data.

3) Pearson correlation test

Research question

Is there a linear correlation between variable x1 and x2?

Procedure

check if covariation is linear with a graph:

```
#use package “ggpubr”
> ggscatter(data, x="xvar", y="yvar", add="reg.line", conf.int=T,
cor.coef=T, cor.method="pearson", xlab=" ", ylab=" ")
```

check if data is normally distributed:

```
> shapiro.test(x) #normally distributed if p>=0.05
> shapiro.test(y) #normally distributed if p>=0.05
```

Test

```
> cor.test(x1,x2, method="pearson", [use = “complete.obs”])
```

if p<0.05: data is significantly correlated with a correlation coefficient of “cor”

if p>=0.05: data is not significantly correlated

4) spearman's rank correlation test

Research question

Is there a linear correlation between variable x1 and x2?

*x1 and x2 do not have to be normally distributed

Procedure

Check if covariation is linear with a graph:

```
#use package "ggpubr"
> ggscatter(data, x="xvar", y="yvar", add="reg.line", conf.int=T,
cor.coef=T, cor.method="pearson", xlab=" ", ylab=" ")
```

Test

```
> cor.test(x1,x2, method="spearman", [use = "complete.obs"])
```

If p<0.05: data is significantly correlated with a correlation coefficient of "rho"

If p>=0.05: data is not significantly correlated

rho value indication:

-1 Indicates strong negative correlation

0 Indicates no association between variables

1 Indicates a strong positive correlation

5) T-test (paired/unpaired)

Research question

Is there a significant difference in means between variable x1 variable x2?

Test

```
> t.test(x, y, paired = T/F, alternative = "two.sided")
```

If p<0.05: reject null hypothesis and accept the alternative hypothesis that the means of both samples are significantly different .

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in means exists.

6) Wilcoxon rank sum test (paired/unpaired)

Research question

Is there a significant difference in medians between variable x1 variable x2?

Test

```
> wilcox.test(x, y, paired = T/F, alternative = "two.sided")
```

If p<0.05: reject null hypothesis and accept the alternative hypothesis that the medians of both samples are significantly different .

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in medians exists.

7) ANOVA

Research question

Is there a significant difference in means of the response variable in different groups/conditions?

Test

```
> res <- aov(y~x, data=data)
> summary(res)
```

If p<0.05: reject null hypothesis and accept the alternative hypothesis that there is a significant difference in means between some groups.

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in means exists.

8) Tukey test

Research question

Which of the means of the different groups are significantly different?

Test

```
> res <- aov(y~x, data=data)
> TukeyHSD(res)
```

For every pair of groups/conditions:

If p<0.05: reject null hypothesis and accept the alternative hypothesis that the means of both samples are significantly different .

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in means of both samples exists.

9) Kruskal-wallis test

Research question

Is there a significant difference in means of the response variable in different groups/conditions?

Test

```
> res <- kruskal.test(y~x, data=data)
> summary(res)
```

If p<0.05: reject null hypothesis and accept the alternative hypothesis that there is a significant difference in means between some groups.

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in means exists.

if p is significant we can perform a multiple pairwise-comparison:

Research question

Which of the means of the different groups are significantly different?

Test

```
> pairwise.wilcox.test(y, x, p.adjust.method = “BH”)
```

For every pair of groups/conditions:

If p<0.05: reject null hypothesis and accept the alternative hypothesis that the means of both samples are significantly different .

If p>=0.05: can not reject the null hypothesis and thus can not conclude that a significant difference in means of both samples exists.

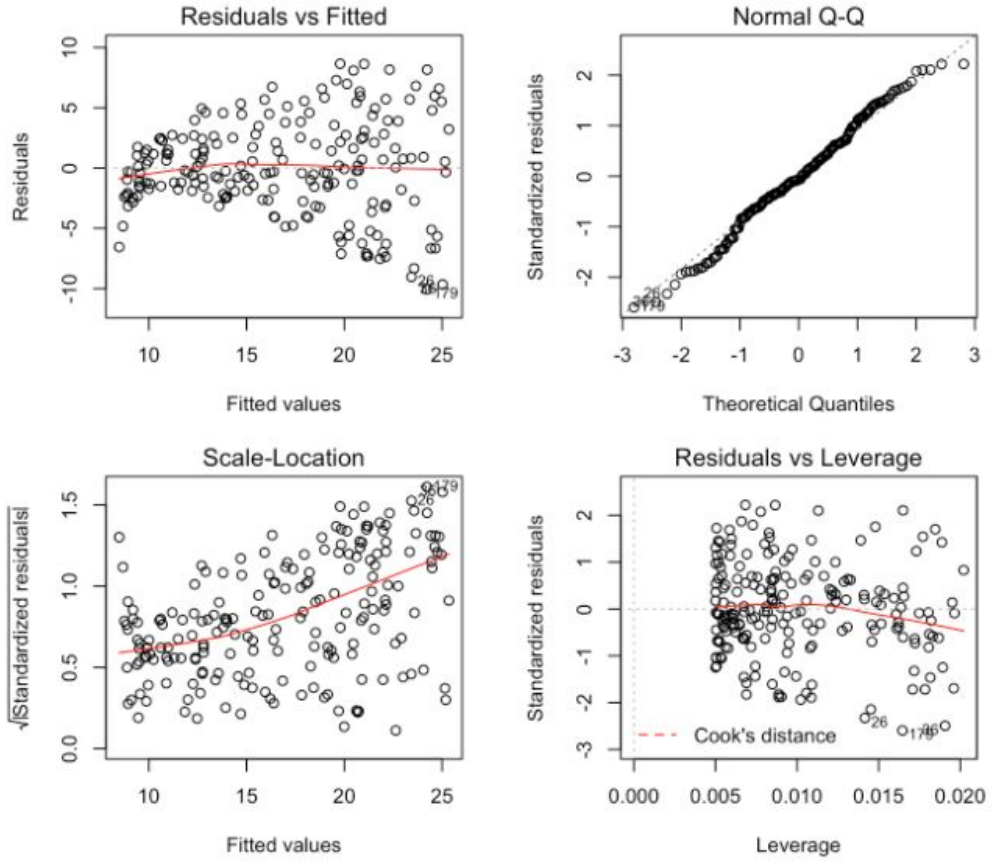
EXTRA

Predicting response variable value with given predictor variable values

```
> model <- lm(y~x1+...+xn, data=data)
> values=data.frame(x1=value, …, xn=value)
> predict(model, values, interval="predict")
```

Describing diagnostics of estimated linear model

```
> model <- lm(y~x1+...+xn, data=data)
> layout(matrix(c(1,2,3,4),2,2))
> plot(fit)
```



Residuals vs Fitted

Used to check linearity: if the red line is approximately horizontal we can assume a linear relation between predictors and and outcome variables.

Normal Q-Q

Used to check if the residuals are normally distributed. should follow a straight diagonal line.

Scale-location plot

Used to check if residuals are spread equally along ranges of the predictors. should be a horizontal red line with equally spread points. if the spread increases (heteroscedasticity) then you could use the log of the outcome:

```
> model <- lm(log(y)~x, data=data)
```

Residuals vs leverage

Used to check for influential data points on our model. If no data points are beyond the red-dotted line, there are no influential cases.

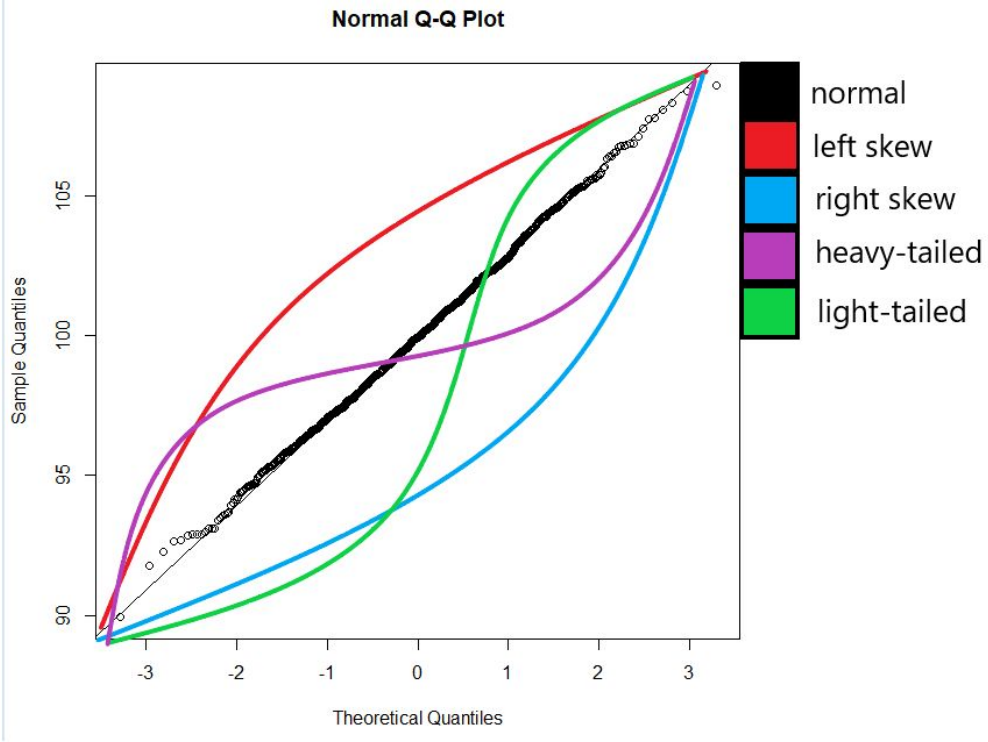
QQ-plot

Check if both samples from same distribution:

```
> qqplot(x,y)
```

Check for normality with a 1 way QQ-plot:

```
> qqnorm(x)
> qqline(x)
```



importing data

```
> data1 <- read.delim(file.choose(), sep=" ")
> data1 <- read.table(file.choose())
```

changing values of a column

```
> data$column <- as.character(data$column)
> data$column[data$column=="word1"] <- "word2"
> data$column <- as.factor(data$column)
```

make a frequency table

frequency:

```
> data.freq <- table(data$column1, data$column2)
```

relative frequency

```
> data.relfreq <- round(prop.table(data.freq, 1), 4)
or
> data.relfreq <- round(prop.table(data.freq, 2), 4)
```

plotting multiple histograms

```
layout(matrix(c(1,2,3,4),2,2))
for(col in 2:ncol(data)) {
x <- names(data)[col]
y <- paste("Histogram of", x, sep=" ")
hist(data[,col], main=y, xlab=x)
}
#layout(1) when done
```

plotting multiple boxplots

```
layout(matrix(c(1,2,3,4),2,2))
for(col in 2:ncol(data)) {
x <- names(data)[col]
y <- paste("Boxplot of", x, sep=" ")
boxplot(data[,col], main=y, ylab=x)
}
#layout(1) when done
```

removing NA's:

```
> data <- na.omit(data)
```

Points of discussion

- Is there enough data?
- is the experiment too specific?
- are the groups different from each other?
- were the measurements precise?
- is the sample randomly chosen?

links

1) Chi-square test
http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r
2) Regression analysis (1=linear, 2=logistic)
http://r-statistics.co/Linear-Regression.html
https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/
3) Pearson correlation test + 4) spearman's rank correlation test
http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
5) T-test (paired/unpaired)
http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r
6) Wilcoxon rank sum test (paired/unpaired)
http://www.sthda.com/english/wiki/wiki.php?title=unpaired-two-samples-wilcoxon-test-in-r
7) ANOVA + 8) Tukey test
http://www.sthda.com/english/wiki/one-way-anova-test-in-r
9) Kruskal-wallis test
http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

diagnostics of linear model
http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/