SNOW LEOPARD NETWORK

– Module 12 –

Fundamental Statistical Tools IV

24.06.2021

Anne Heloise Theo & Guillaume Demare

T-TEST AND ANOVA (RECAP)

(Unpaired) Two Sample T-test

- Are **two sample means** significantly different?

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

```
# two sample t-test (unpaired)
# note: group1 and group2 must be numeric vectors
t.test(x = group1, y = group2, var.equal = TRUE)
```

(One-Way) ANalysis Of VAriance (ANOVA)

- Are **at least two sample means** significantly different?

$$F = \frac{MSB}{MSW} = \frac{SSB/(c-1)}{SSW/(n-c)}$$

```
# anova
# note: my_data (data frame) here has two columns:
# var (numeric) and group (categorical)
fit <- aov(var ~ group, data = my_data)
summary(fit)
```

ASSUMPTIONS

- Data is continuous (i.e. interval or ratio)
- Data points (and categories) are independent
- Data is (approximately) normally distributed
- Variance between groups is homogenous

CORRELATION

- **Pearson correlation coefficient**

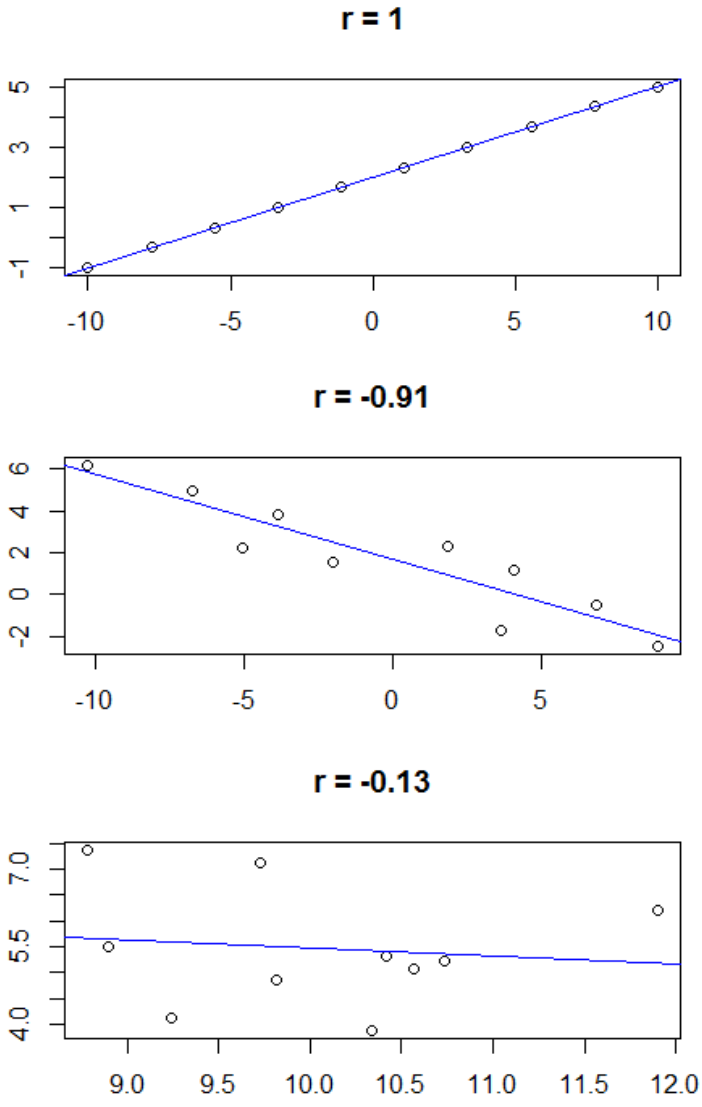
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

a: $(x_i - \bar{x})$ in the numerator; a: $(x_i - \bar{x})^2$ in the denominator

b: $(y_i - \bar{y})$ in the numerator; b: $(y_i - \bar{y})^2$ in the denominator

- Measures how strongly two sets of continuous data are correlated
- *r* ranges from -1 to +1

```
# calculate correlation coefficient for length vs weight
a = uncia$Length.cm - mean(uncia$Length.cm)
b = uncia$Weight.kg - mean(uncia$Weight.kg)
r = sum(a * b) / sqrt(sum(a^2) * sum(b^2))

# using the cor() function with method = "pearson"
cor(uncia$Length.cm, uncia$Weight.kg, method = "pearson")
```



IS IT SIGNIFICANT?

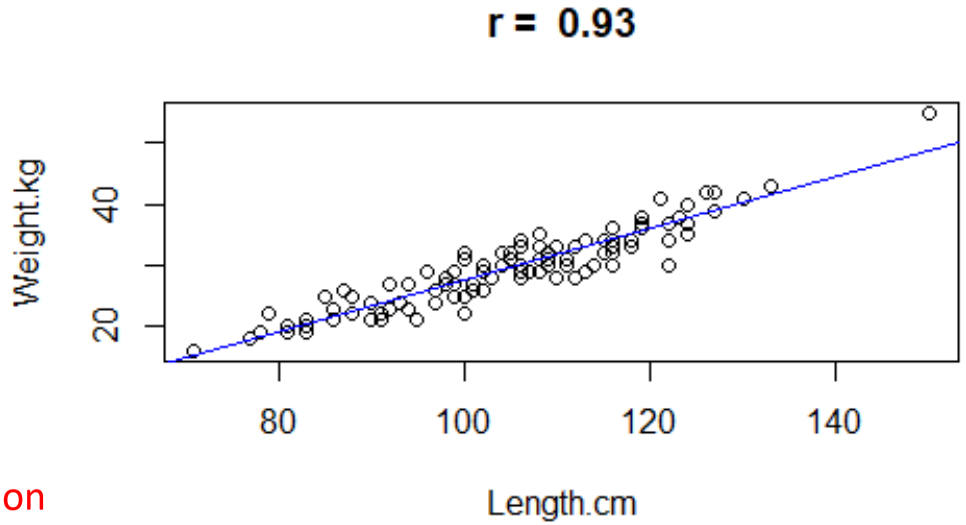
- Correlation coefficient describes the **sample**, so can we make inference about the **population**?
  --> is the correlation coefficient **significantly different from zero**?

- H_0: there is no relationship ($\rho = 0$)
- H_1: there is a significant relationship ($\rho \neq 0$)

- **t-test** for a correlation coefficient:

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

coefficient of determination




```
# calculate t test statistic
t = r * sqrt(nrow(uncia)- 2) / sqrt(1 - r^2)
# correlation test
cor.test(uncia$Weight.kg, uncia$Length.cm)
t = 25.545, df = 101, p-value < 2.2e-16
```

```
# calculate coefficient of determination
r^2
[1] 0.8659629
```

SIMPLE LINEAR REGRESSION

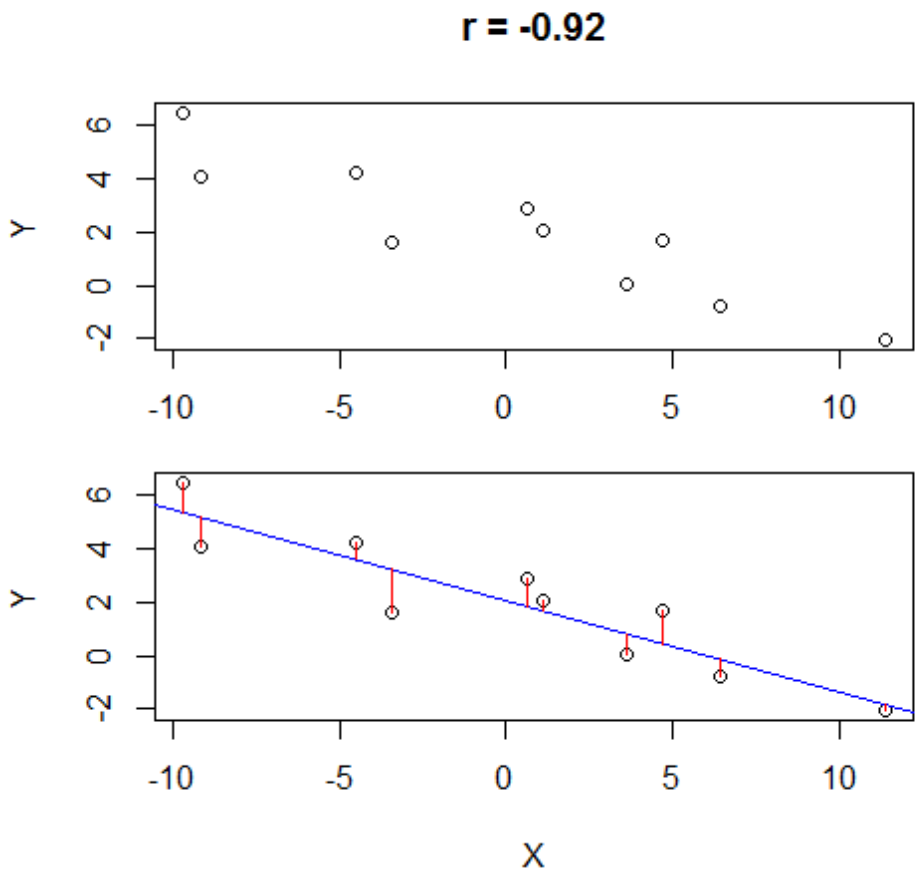
- **Simple linear regression** quantifies the relationship between two continuous variables: predictor and response

$$y = \alpha + \beta\, x$$

- Method: ordinary least square (OLS)

$$\beta = \frac{n \sum(xy) - \sum x \sum y}{n \sum(x^2) - (\sum x)^2} \quad \text{(slope)}$$

$$\alpha = \frac{\sum y - \beta \sum x}{n} \quad \text{(intercept)}$$



SIMPLE LINEAR REGRESSION

- **Simple linear regression** quantifies the relationship between two continuous variables: predictor and response

$$y = \alpha + \beta\, x$$

- Method: ordinary least square (OLS)

$$\beta = \frac{n \sum(xy) - \sum x \sum y}{n \sum(x^2) - (\sum x)^2} \quad \text{(slope)}$$

$$\alpha = \frac{\sum y - \beta \sum x}{n} \quad \text{(intercept)}$$

```
# simple linear regression
fit1 <- lm(Weight.kg ~ Length.cm, data = uncia)
summary(fit1)

Call:
lm(formula = Weight.kg ~ Length.cm, data = uncia)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8779 -1.6441  0.0846  1.4293  6.2722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.75378    1.75831  -8.391 3.06e-13 ***
Length.cm     0.42321    0.01657  25.545  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.389 on 101 degrees of freedom
Multiple R-squared:  0.866, Adjusted R-squared:  0.8646
F-statistic: 652.5 on 1 and 101 DF,  p-value: < 2.2e-16
```

MULTIPLE LINEAR REGRESSION

- **Simple linear regression** quantifies the relationship between two continuous variables: predictor and response

$$y = \alpha + \beta\, x$$

- **Multiple linear regression** quantifies the relationship between one continuous dependent variable (aka the response) and two or more independent variables (aka the predictors)

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

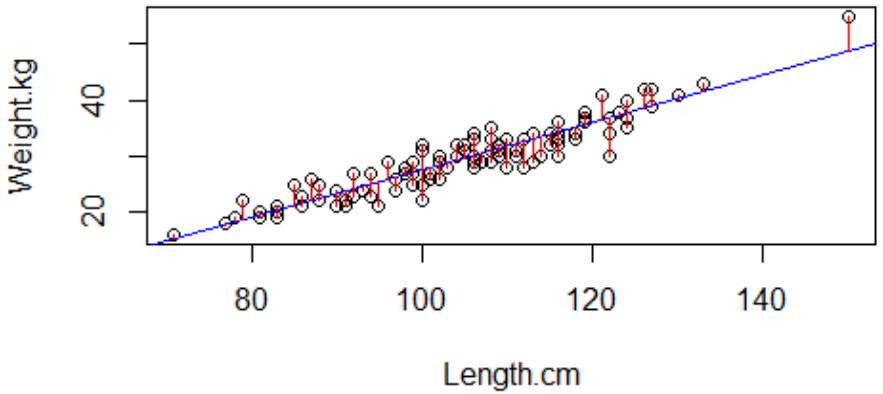
```
# multiple linear regression
fit2 <- lm(Weight.kg ~ Length.cm + Sex, data = uncia)
summary(fit2)

Call:
lm(formula = Weight.kg ~ Length.cm + Sex, data = uncia)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7106 -1.8340  0.0712  1.5567  5.8673

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.47423    2.06006  -7.997 2.31e-12 ***
Length.cm     0.44365    0.02096  21.165  < 2e-16 ***
SexM         -0.94024    0.59783  -1.573    0.119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.371 on 100 degrees of freedom
Multiple R-squared:  0.8692, Adjusted R-squared:  0.8666
F-statistic: 332.3 on 2 and 100 DF,  p-value: < 2.2e-16
```

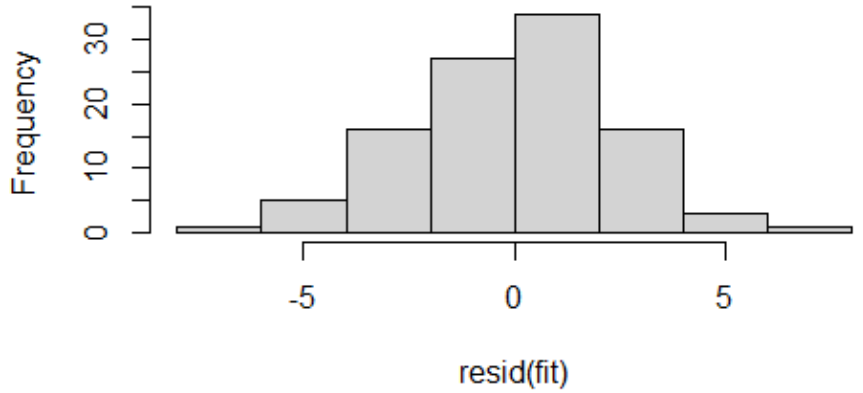
LINEAR MODEL ASSUMPTIONS

- **Linearity**
  Correct functional form as a **linear** equation

$$Weight = \alpha + \beta_1 Length$$

- **Constant residual variance** (aka homoscedasticity)
- **Independent error terms** (i.e. no autocorrelation!)
- Residuals are **normally distributed**






```
# plot residuals vs fitted values
plot(fit1, which = 1)
# check normality of errors
plot(fit1, which = 2)
hist(resid(fit))
```

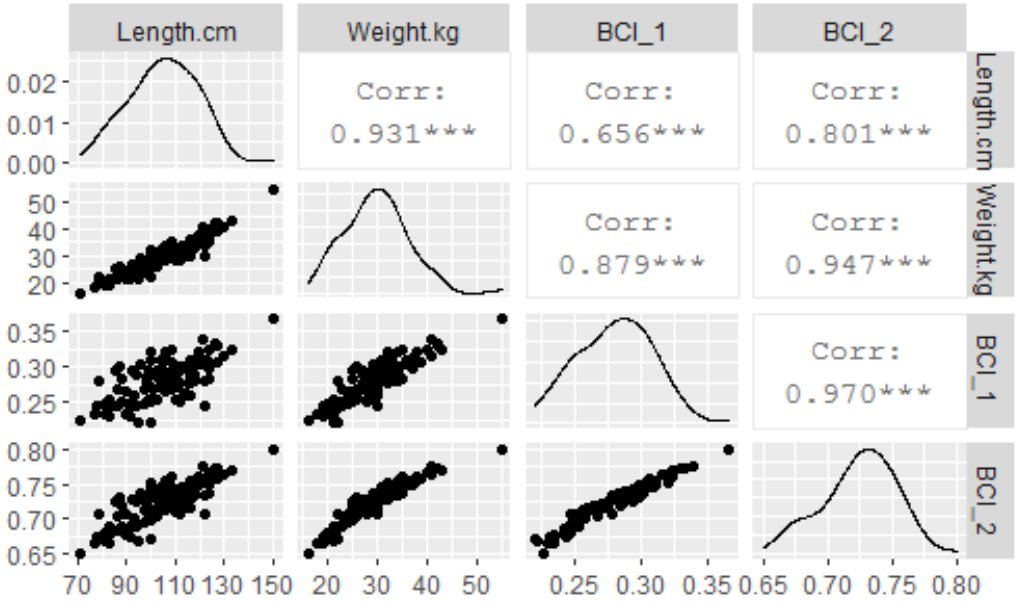
LINEAR MODEL ASSUMPTIONS

- **No multi-collinearity** between predictors in the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- **No omitted variables**

check out **zstatistics.com** for more details on linear model assumptions and how to remedy!




```
# pairs() or plot()
pairs(uncia[c("Length.cm", "Weight.kg", "BCI_1", "BCI_2")])
plot(uncia[c("Length.cm", "Weight.kg", "BCI_1", "BCI_2")])
# correlogram
install.packages("GGally")
library(GGally)
ggpairs(uncia[c("Length.cm", "Weight.kg", "BCI_1", "BCI_2")])
```

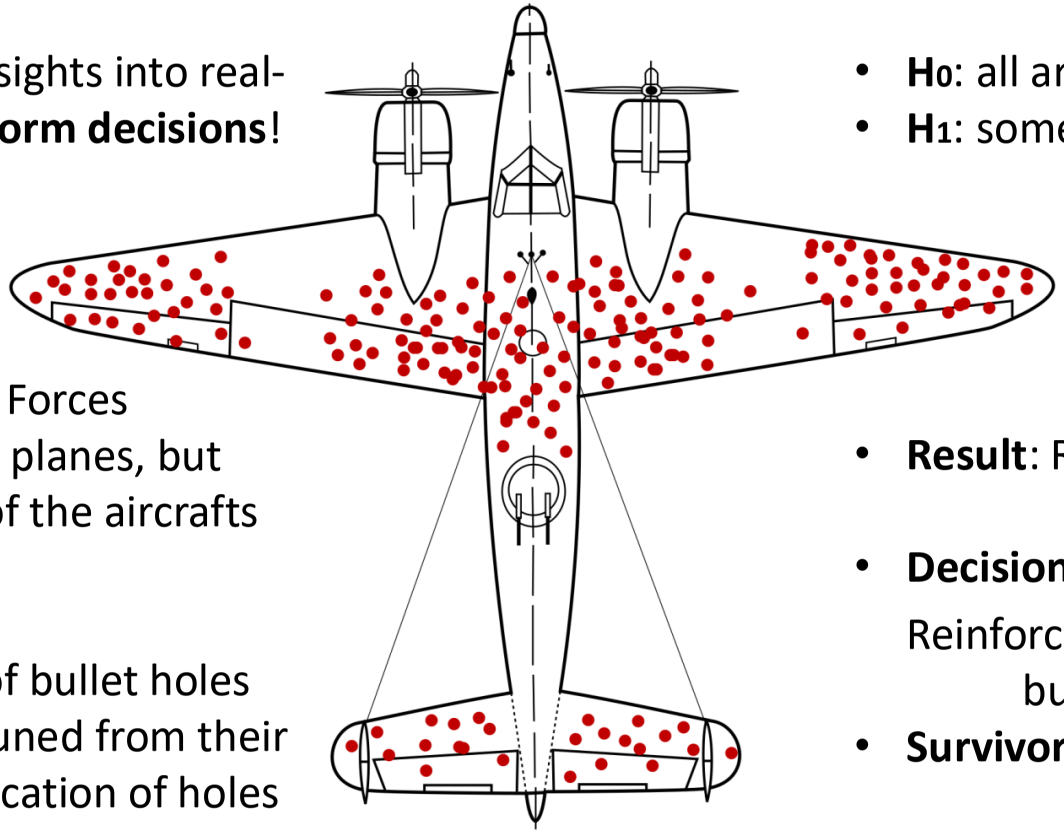
GOING FURTHER

- Statistics can provide insights into real-world processes and **inform decisions**!

Example:

- During WWII, the Allied Forces aimed to reinforce their planes, but only focusing on areas of the aircrafts that needed it the most

- **Data collection**: count of bullet holes for every plane that retuned from their mission, noting down location of holes



- **H_0**: all areas are hit the same
- **H_1**: some areas are hit more often

- **Result**: Reject null hypothesis

- **Decision**?

  Reinforce all areas WITHOUT bullet holes (Abraham Wald)

- **Survivor bias!**