SNOW LEOPARD NETWORK

– Module 12 –

Fundamental Statistical Tools III

17.06.2021

Anne Heloise Theo & Guillaume Demare

ASSUMPTION CHECK

- Most statistical tests (especially parametric tests) make a number of **assumptions**

- EXAMPLE: **Student's t-test assumptions**
  - Data values are continuous
  - Data is (approximately) normally distributed
  - Data points are independent
  - Samples have similar amount of variance

    i.e. homogeneity of variance

```
	Shapiro-Wilk normality test

data:  bci.wakhan
W = 0.98726, p-value = 0.7695

	Shapiro-Wilk normality test

data:  bci.altai
W = 0.98112, p-value = 0.718

	Bartlett test of homogeneity of variances

data:  BCI_2 by Location
Bartlett's K-squared = 2.3336, df = 1, p-value = 0.1266
```

```
## METHOD 1: graphical inspection

# histograms to check normality
hist(bci.wakhan)
hist(bci.altai)

# boxplot to check homogeneity of variance
boxplot(BCI_2 ~ Location, data = uncia)
```

```
## METHOD 2: statistical tests

# Shapiro-Wilk test to check normality
shapiro.test(bci.wakhan)
shapiro.test(bci.altai)

# Bartlett's test to check homogeneity of variance
bartlett.test(BCI_2 ~ Location, data = uncia)
```

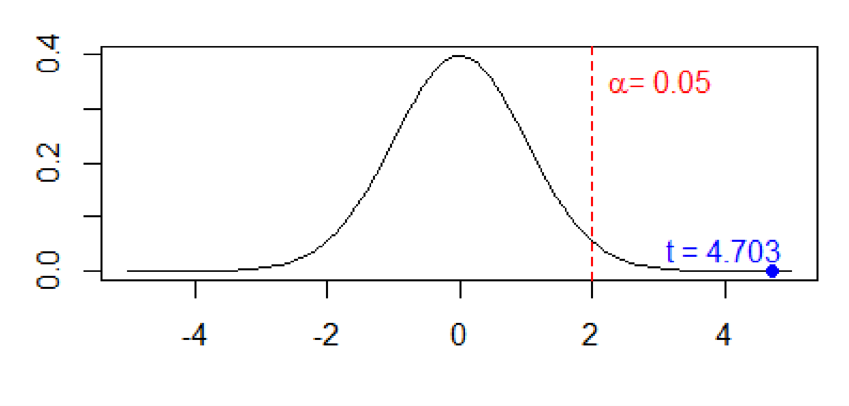
P-VALUE

- **Two sample t-test** to compare two sample **means**

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

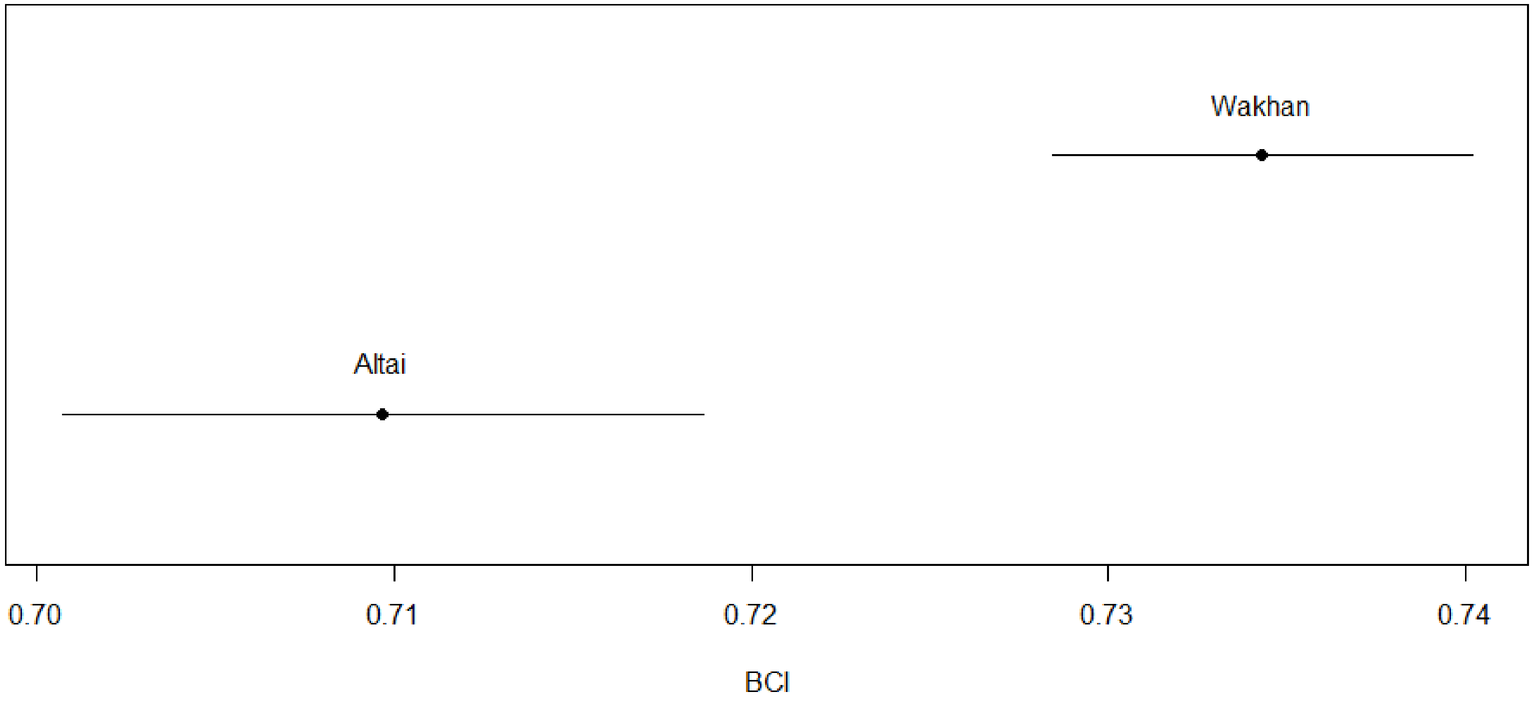
- Result: there is a **significant difference** between the two group means at the significance level α = 0.05

```
# t-test to compare mean BCI_2 between the two locations
bci.wakhan <- subset(uncia$BCI_2, uncia$Location == "Wakhan")
bci.altai <- subset(uncia$BCI_2, uncia$Location == "Altai")
t.test(x = bci.wakhan, y = bci.altai, alternative = "two.sided", var.equal = TRUE)

        Two Sample t-test

data:  bci.wakhan and bci.altai
t = 4.7027, df = 101, p-value = 8.155e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01422979 0.03499355
sample estimates:
mean of x mean of y
0.7342995 0.7096879
```



CONFIDENCE INTERVALS

- T-test: the two **sample means** (BCI) are significantly different at $\alpha = 0.05$
- There is no overlap between the 95% confidence intervals for the two estimates of **population means**



TYPE I & TYPE II ERRORS

- **Type I error** = rejection of a true null hypothesis
- **Type II error** = non-rejection of a false null hypothesis

- **How to avoid …**
  - increase sample size
  - think carefully about study design
  - set a different α
    e.g. 0.01 or 0.1
  - make sure that no key assumptions have been violated
  - choose a better suited analytical framework

		H_0 What the null hypothesis really is	
		TRUE	FALSE
H_0 What we think the null hypothesis is	TRUE	there really is **no** effect	**TYPE II ERROR** there actually is an effect
	FALSE	**TYPE I ERROR** there actually is **no** effect	there really is an effect

ONE-WAY ANOVA

- **Analysis of variance**
- To compare more than two sample **<u>means</u>**

Sum of squares (SST)

- **Variance:** $\frac{\sum(x-\bar{x})^2}{n-1}$
- SSW: $\sum(x-\bar{x})^2$
- SSB: $n\sum(\bar{x}-\bar{\bar{x}})^2$
- SST = SSW + SSB

F statistic

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

Degrees of freedom

- MSB: mean square between groups
- MSW: mean square within groups
- c: number of categories
- n: number of samples

ONE-WAY ANOVA

H_0: $\mu_1 = \mu_2 = \mu_3$: all means are equal

H_1 (alternative hypothesis): at least one population mean is different from the rest

Assumptions:

- Population is normally distributed
- Measurement is interval or ratio
- Categories are independent
- Samples are independent
- Homogeneity of variance

PAIRWISE T-TESTS

- Used to figure out **pair-wise differences** between groups
  i.e. post-hoc analysis

- The problem with multiple comparisons: false positives!

- **Bonferroni** correction: α/m

- **Holm** method/**Holm-Bonferroni** method: arrange significant results in ascending order of p-value and check significance against modified p-values:

$$P_k < \frac{\alpha}{m+1-k}$$

TWO-WAY ANOVA

- Used to test effect of **two grouping variables** simultaneously, as well as their interaction
- **Post-hoc analysis: Tukey's** honestly significant difference test

	Altai	Wakhan
Females	88	98
	90	99
	91	101
	92	103
Males	110	112
	111	116
	112	118
	113	119

```
# calculating anova
mod1 <- aov(Weight.kg ~ Location*Sex, data = uncia)
summary(mod1)

# Post-Hoc analysis
TukeyHSD(mod1, conf.level=.95)

# Plot Tukey
par(mar=c(4.1, 13, 4.1, 2.1))
plot(TukeyHSD(mod1, conf.level=.95), las = 2)
```