ANEXOS

Introducción a la Ciencia de Datos

Chi-square independence test

- For 2-way tables you can use `chisq.test()` to test independence of the row and column variable.
- By default, the **p**-value is calculated from the asymptotic chi-squared distribution of the test statistic.
- Optionally, the **p**-value can be derived via Monte Carlo simulation.

```
> (HairEye <- margin.table(HairEyeColor, c(1, 2)))
Eye
Hair Brown Hazel Green Blue
Black 68 15 5 20
Brown 119 54 29 84
Red 26 14 14 17
Blond 7 10 16 94

> chisq.test(HairEye)
Pearson's Chi-squared test
data: HairEye
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Fisher Exact Test of independence

- X must be a two-way contingency table in table form.
- Another form, fisher.test(X, Y) takes two categorical vectors of the same length. For tables larger than 22 the method can be computationally intensive (or can fail) if the frequencies are not small.

```
> fisher.test(GSStab)
Fisher's Exact Test for Count Data
data: GSStab
p-value = 0.03115
alternative hypothesis: two.sided
```

Testing for Normality: Shapiro-Wilks test

- To determine whether your data sample is normally distributed use the shapiro.test() function:

```
> shapiro.test(x)
Shapiro-Wilk normality test
data: x
W = 0.9651, p-value = 0.4151
9.13
```

- The large **p** -value suggests the underlying population could be normally distributed.
- a small **p** -value suggsest that it is unlikely that this sample came from a normal population:

Testing for Normality: Shapiro-Wilks test

- When you choose a test, you may be more interested in the normality in each sample. You can test both samples in one line using the tapply() function, like this:

```
> with(beaver, tapply(temp, activ, shapiro.test)
```

- The large **p** -value suggests the underlying population could be normally distributed.
- a small **p** -value suggsest that it is unlikely that this sample came from a normal population:

Performing Robust ANOVA (Kruskal–Wallis Test)

- Your data is divided into groups.
- Are there significant differences between these groups?

```
my_data <- PlantGrowth
```

- Is any significant difference between the average weights of plants in the 3 experimental conditions?

```
kruskal.test(weight ~ group, data = my_data)
Kruskal-Wallis rank sum test
data:  weight by group
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842
```

P-value is less than the significance level 0.05, we can conclude that there are significant differences between the treatment groups.

Multiple pairwise-comparison between groups

- There is a significant difference between groups, but which pairs of groups are different?.
- The function `pairwise.wilcox.test()` calculate pairwise comparisons between group levels with corrections for multiple testing.

```
pairwise.wilcox.test(PlantGrowth$weight, PlantGrowth$group,
p.adjust.method = "BH")
```

```
Pairwise comparisons using Wilcoxon rank sum test data: PlantGrowth
$weight and PlantGrowth$group
ctrl trt1
trt1 0.199 -
trt2 0.095 0.027
P value adjustment method: BH
```

only trt1 and trt2 are significantly different ($p < 0.05$)..

Dealing with non normality

Data transformation

- if you are looking at relationships between variables (e.g., regression) it is alright just to transform the problematic variable
- if you are looking at differences within variables (e.g., change in a variable over time) then you need to transform all levels of those variables.

Data transformation	Can correct for
log transformation	Positive skew, unequal variances
square root transformation	Positive skew, unequal variances
reciprocal transformation 1/(variable+1)	Positive skew, unequal variances
reverse score transformation	negative skew

Comparison of Statistical Analysis Tools for Normally and Non-Normally Distributed Data

Tools for Normally Distributed Data	R functions	Equivalent Tools for Non-Normally Distributed Data	R functions
T-test	`t.test()`	Mann-Whitney test; Mood's median test; Kruskal-Wallis test	`wilcox.test(); mod.medtest(); kruskal.test()`
ANOVA	`aov()`	Mood's median test; Kruskal-Wallis test	`mod.medtest(); kruskal.test()`
Paired t-test	`t.test(x, y, paired = TRUE)`	One-sample sign test	`SIGN.test()`
F-test; Bartlett's test	`var.test();bartlett.test()`	Levene's test	`Levene.test()`

Gracias…