Hypothesis testing

Descriptive analysis and basic statistics in biomedical studies using R and Markdown

Juan R Gonzalez
juanr.gonzalez@isglobal.org

BRGE - Bioinformatics Research Group in Epidemiology
ISGlobal - Barcelona Institute for Global Health
http://brge.isglobal.org

Hypothesis testing

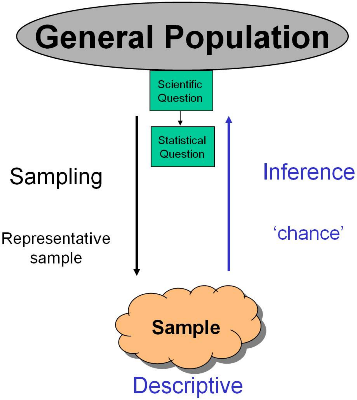


Figure 1: Scheme scientific research

Tests (continuous variables)

```r
library(Hmisc)
df <- spss.get("data/partoFin.sav", allow="_",
               datevars=c("dia_nac", "dia_entr", "ulti_lac"))
```

- One sample test

```r
t.test(df$peso, mu=4)
```

```

    One Sample t-test

data:  df$peso
t = -8.4635, df = 27, p-value = 4.471e-09
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.016260 3.400169
sample estimates:
mean of x
 3.208214
```

- Two independent sample test

```r
t.test(peso ~ sexo, data=df)
```

```

    Welch Two Sample t-test

data:  peso by sexo
t = 0.39385, df = 25.82, p-value = 0.6969
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3024945  0.4458278
sample estimates:
mean in group niño mean in group niña
          3.249167           3.177500
```

- Paired t-test

```r
t.test(df$horas_an, df$horas_de, paired = TRUE)
```

```

    Paired t-test

data:  df$horas_an and df$horas_de
t = 0.88662, df = 27, p-value = 0.3831
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.173414  2.959128
sample estimates:
mean of the differences
              0.8928571
```

- ANOVA (more than 2 groups)

```r
mod <- aov(peso ~ naci_ca, data=df)
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
naci_ca      2  0.818  0.4092   1.765  0.192
Residuals   25  5.798  0.2319
```

```r
mod <- aov(peso ~ naci_ca + sexo, data=df)
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
naci_ca      2  0.818  0.4092   1.704  0.203
sexo         1  0.035  0.0352   0.147  0.705
Residuals   24  5.763  0.2401
```

Post-hoc

- None

```
with(df, pairwise.t.test(peso, naci_ca , p.adjust="none"))
```

```

    Pairwise comparisons using t tests with pooled SD 

data:  peso and naci_ca 

           Española Otras
Otras      0.339    -    
Sudamérica 0.078    0.461

P value adjustment method: none 
```

- Bonferroni

```
with(df, pairwise.t.test(peso, naci_ca, p.adjust="bonf"))
```

```

    Pairwise comparisons using t tests with pooled SD 

data:  peso and naci_ca 

           Española Otras
Otras      1.00     -    
Sudamérica 0.23     1.00 

P value adjustment method: bonferroni 
```

- Tukey

```r
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = peso ~ naci_ca + sexo, data = df)

$naci_ca
                         diff        lwr       upr     p adj
Otras-Española      0.2171429 -0.3493323 0.7836181 0.6102379
Sudamérica-Española 0.4100000 -0.1564752 0.9764752 0.1885514
Sudamérica-Otras    0.1928571 -0.4612521 0.8469664 0.7446189

$sexo
                 diff        lwr       upr     p adj
niña-niño -0.07166667 -0.4578847 0.3145514 0.7051077
```

Outliers

Analyzing data with outliers may influence the value of a (non-robust) statistic. We can test the null-hypothesis that a variable does not contain **an** outlier. Under the assumption that the data are realizations of one and the same distribution, such a hypothesis can be tested by the Grubbs (1950) test. This test is based on the statistic $g = |suspectvalue - \bar{x}|/s$, where the suspect value is included for the computation of the mean $\bar{x}$ and the standard deviation s.

```r
library(outliers)
grubbs.test(df$peso)
```

```

	Grubbs test for one outlier

data:  df$peso
G = 2.52870, U = 0.75441, p-value = 0.1026
alternative hypothesis: highest value 4.46 is an outlier
```

Since the p-value is not lower than 0.05, the conclusion is that there are no evidences to reject the null- hypothesis of no outliers.

Non-parametric tests

- Wilcoxon test (U Mann Withney)

```r
wilcox.test(peso ~ sexo, data=df)
```

```
    Wilcoxon rank sum test with continuity correction

data:  peso by sexo
W = 108.5, p-value = 0.5771
alternative hypothesis: true location shift is not equal to 0
```

- Krusdall-Wallis (More than two groups)

```r
kruskal.test(peso ~ naci_ca, data=df)
```

```
    Kruskal-Wallis rank sum test

data:  peso by naci_ca
Kruskal-Wallis chi-squared = 4.6217, df = 2, p-value = 0.09917
```

Two proporions

- Chi-square test

```
freq <- with(df, table(sexo, tip_par))
chisq.test(freq)
```

```

    Pearson's Chi-squared test with Yates' continuity correction

data:  freq
X-squared = 1.6106e-31, df = 1, p-value = 1
```

- Fisher test

```r
freq <- with(df, table(sexo, tip_par))
fisher.test(freq)
```

```
    Fisher's Exact Test for Count Data

data:  freq
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.06160374 9.21621060
sample estimates:
odds ratio
 0.8710761
```

Correlation

- Pearson correlation test

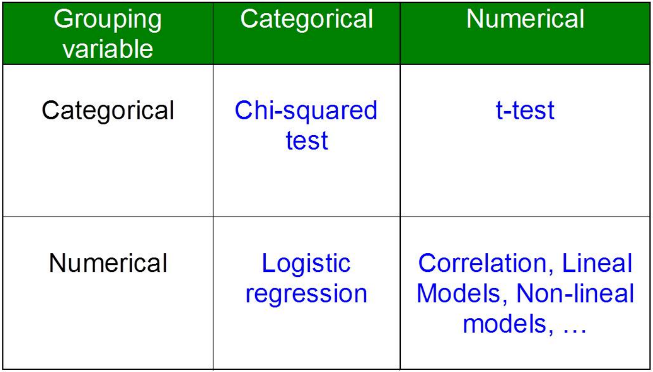
```r
cor.test(df$peso, df$edad)
```

```

    Pearson's product-moment correlation

data:  df$peso and df$edad
t = -2.7502, df = 26, p-value = 0.01069
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7202342 -0.1235120
sample estimates:
       cor
-0.4747143
```

- Spearman correlation test

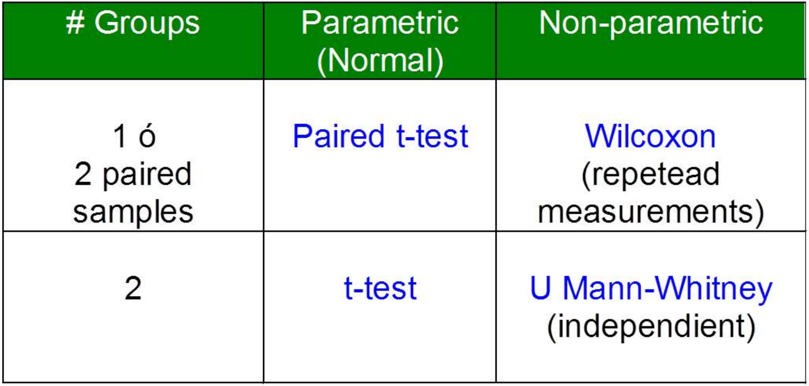
```r
cor.test(df$peso, df$edad, method="spearman")
```

```

    Spearman's rank correlation rho

data:  df$peso and df$edad
S = 5678.9, p-value = 0.002215
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.5541522
```



Grouping variable	Categorical	Numerical
Categorical	Chi-squared test	t-test
Numerical	Logistic regression	Correlation, Lineal Models, Non-lineal models, …

Figure 2: Statistical Tests



# Groups	Parametric (Normal)	Non-parametric
1 ó 2 paired samples	Paired t-test	Wilcoxon (repetead measurements)
2	t-test	U Mann-Whitney (independient)

Figure 3: Testing means

Permutation tests (**Advanced**)

Some times parametric tests cannot be applied, since there is not known distribution. Therefore, Monte Carlo-based methods can be used instead:

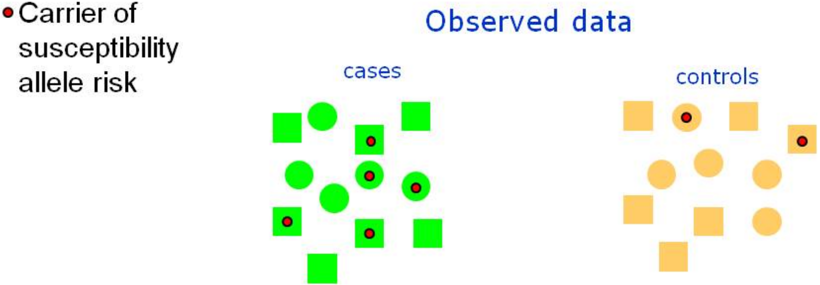


Figure 4: Permutation testing

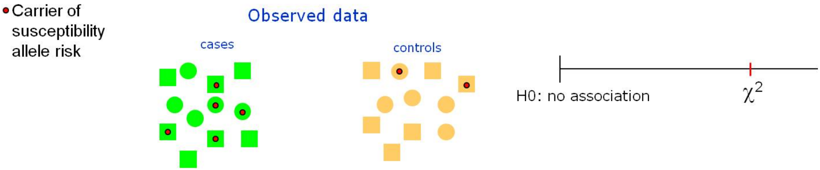


Figure 5: Permutation testing

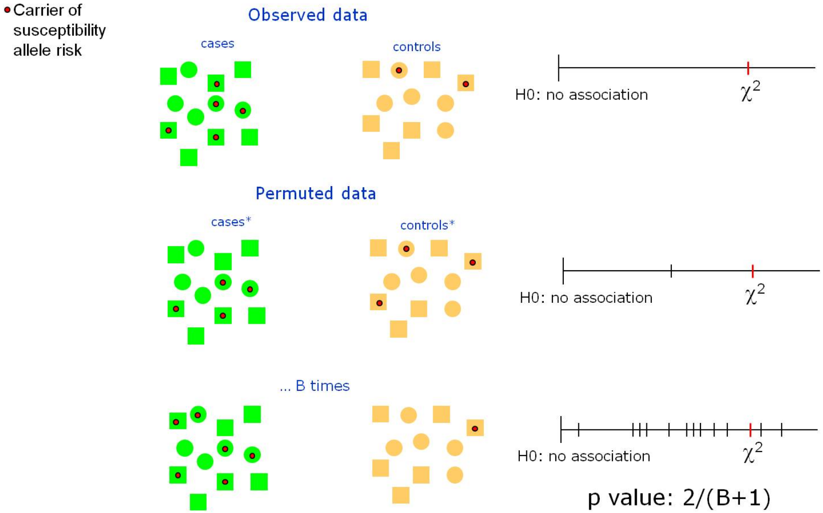


Figure 6: Permutation testing

Let's assume that you are interested in knowing whether there are differences in the *median absolute deviation* ($|x - \text{median}(x)|$) of breastfeeding weeks depending on the type of visits (intensive vs standard). How will you get a p-value?

```r
B <- 10000
stat.ref <- mad(df$sem_lac[df$tx=="Intensivo"]) -
            mad(df$sem_lac[df$tx=="Est?ndar"])
stat <- rep(NA, B)
for (i in 1:B) {
 tx.r <- sample(df$tx, nrow(df), replace=FALSE)
 stat[i] <- mad(df$sem_lac[tx.r=="Intensivo"]) -
            mad(df$sem_lac[tx.r=="Est?ndar"])
}
```

```r
df$tx
```

```
Regimen visitas asignado
 [1] Intensivo Intensivo Estándar  Intensivo Estándar  Estándar  Intensivo
 [8] Intensivo Estándar  Intensivo Estándar  Intensivo Estándar  Intensivo
[15] Estándar  Estándar  Intensivo Intensivo Estándar  Intensivo Estándar
[22] Intensivo Estándar  Intensivo Estándar  Intensivo Intensivo Estándar
Levels: Estándar Intensivo
```

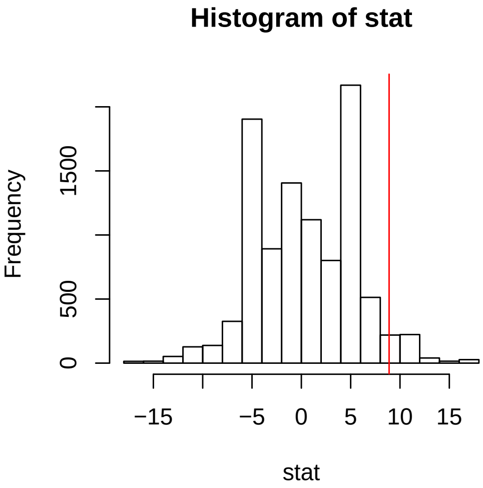
```r
tx.r
```

```
Regimen visitas asignado
 [1] Estándar  Intensivo Estándar  Intensivo Intensivo Estándar  Intensivo
 [8] Intensivo Estándar  Intensivo Intensivo Intensivo Estándar  Intensivo
[15] Intensivo Estándar  Estándar  Intensivo Estándar  Estándar  Estándar
[22] Estándar  Intensivo Estándar  Intensivo Intensivo Intensivo Estándar
Levels: Estándar Intensivo
```

```r
pvalue <- sum(stat>stat.ref)/(B+1)
pvalue
```

```
[1] 0.03049695
```

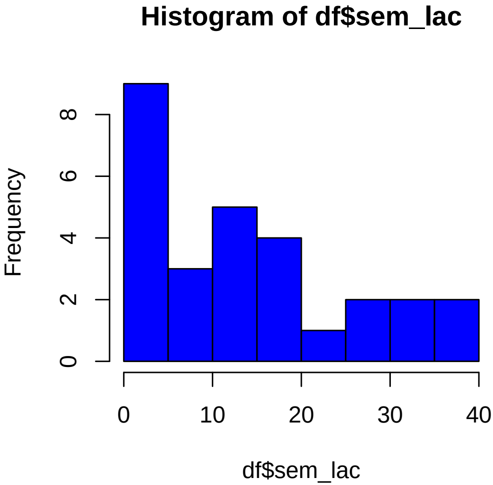
```
hist(stat)
abline(v=stat.ref, col="red")
```



```r
t.test(sem_lac ~ tx, data=df)
```

```

	Welch Two Sample t-test

data:  sem_lac by tx
t = -3.7597, df = 25.083, p-value = 0.0009121
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.009082  -6.139636
sample estimates:
 mean in group Estándar mean in group Intensivo
               6.692308               20.266667
```

```r
hist(df$sem_lac, col="blue")
```



Session info

```
sessionInfo()
```

```
R version 3.4.1 (2017-06-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 16299)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] outliers_0.14   Hmisc_4.0-3     ggplot2_2.2.1   Formula_1.2-2
[5] survival_2.41-3 lattice_0.20-35

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.12        compiler_3.4.1      pillar_1.1.0
 [4] RColorBrewer_1.1-2  plyr_1.8.4          base64enc_0.1-3
 [7] tools_3.4.1         rpart_4.1-11        digest_0.6.12
[10] evaluate_0.10.1     tibble_1.4.2        gtable_0.2.0
[13] htmlTable_1.9       checkmate_1.8.3     rlang_0.1.6
[16] Matrix_1.2-10       yaml_2.1.16         gridExtra_2.3
[19] stringr_1.3.0       knitr_1.20          cluster_2.0.6
[22] htmlwidgets_0.9     rprojroot_1.3-2     grid_3.4.1
[25] nnet_7.3-12         data.table_1.10.4   foreign_0.8-69
[28] rmarkdown_1.8       latticeExtra_0.6-28 magrittr_1.5
```