# Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Research process
- Basic experimental design ideas
- Analysis of variance
- ANOVA modeling with R

# Research process

- Scientific *question* of interest
- Decide *what data* to collect (and how)
- Collection and *analysis* of data
- Conclusions, generalizations : *inference* on the population
- *Communication* and dissemination of results

# Generic question :
## Does a 'treatment' have an 'effect' ?

Examples :

- Does smoking cause cancer, heart disease, *etc* ?
- Does eating oat bran lower cholesterol ?
- Does echinacea prevent illness ?
- Does exercise slow the aging process ?

Approach the question :

- One simple method for resolving this type of question is to *compare two groups* of study subjects :
    - *Control group :* gives a base level for comparison
    - *Treatment group :* group receiving the 'treatment'

# Types of studies

- A basic means to address this type of question involves *comparing two groups* of study subjects :
    - *Control group* : provides a baseline for comparison
    - *Treatment group* : group receiving the 'treatment'
- *Experimental study* : subjects assigned to groups by the investigator
    - *randomization* : protects against bias in assignment to groups
    - *'blind', 'double-blind'* : protects against bias in outcome assessment/measurement
    - *placebo* : artificial/fake treatment
- *Observational study* : subjects 'assign' themselves to groups
    - *confounder* : associated with both group membership/risk factor *and* with the outcome of interest

# A few comments

- With a well-planned and well executed controlled experiment, it is possible to infer *causality*
- This is *not possible* with observational studies due to the presence of confounders
- With confounding, it is not possible to tell whether the observed difference between groups is due to the *treatment* or to the *confounding factor*
- Not always possible to carry out an experiment, for *pratical* and *ethical* reasons

# Example : Hibernation

- General question : *How do changes in an animal's environment induce hibernation ?*
- What changes should be studied *? ?*
    - temperature
    - photoperiod (daylight duration)
- What measures to take ?
    - nerve enzymatic activity (Na+K+ATP-ase)
- What animal to study ?
    - golden hamster, 2 organs

# Specific question

- General question : *How do changes in an animal's environment induce hibernation ?*
- *Specific question :* What is the effect of changing daylight duration on the enzyme concentration of the sodium pump in two golden hamster organs ?

# Sources of variability

- Variability due to the conditions of interest (wanted)
    - Duration (long or short)
    - Organ (heart or brain)

- Variability of the response (NOT wanted) : measurement error

    - Preparation of the enzyme suspension
    - Instrument calibration/standardization

- Variability in experimental units (NOT wanted)
    - biological differences between hamsters
    - environmental differences

# Types of variability

- Systematic, expected (wanted)
- Random variation (can manage this)
- Systematic, unexpected (NOT wanted)
    - biased results
    - *e.g.*, what time the measurements are made

# Questions for the hibernation study

- *Long or short :* Is there an effect of daylight duration on enzyme concentration ?
- *Heart vs. Brain :* Are the concentrations different in the 2 organs ?
- *Interaction :* Is the *difference* in enzyme concentration (long/short) *different* for heart and brain ?
- *Hamsters :* Variability between hamsters ?
- *Measurement error :* What is the error due to the measurement process for enzyme concentration ?

# Experimental design – why do we care?

- Poor design *costs* :
    - time, money, ethical considerations
- To ensure *relevant data* are collected, and can be analyzed to test the scientific hypothesis/ question of interest
    - Decide *in advance* how data will be analyzed
    - 'Designing the experiment' $=$ 'Planning the analysis'
- **The design is about the biology**

# Common experimental designs

- *Completely randomized design* (CRD)
    - compares 2 (or more) *levels* of a single *factor*
    - analysis : 1-way anova (below)
- *Randomized Block Design* (RBD)
    - compares 2 (or more) *levels* of a single *factor*
    - observations in *blocks*
    - analysis : similar to unreplicated 2-way anova
- *(Full) Factorial design*
    - levels from multiple factors varied and studied *simultaneously*
    - can detect *interaction* between factors
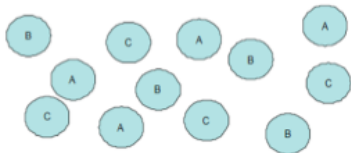    - analysis : 2-way (or multi-way) anova

# Completely randomized experiment

- Study subjects (experimental units) *homogeneous*
- *Randomized* to treatments (factor levels)
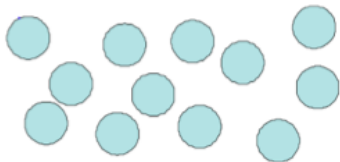


experimental units      randomized to treatments

# Data example : Blood coagulation time

- 24 animals
- Randomly assigned
- to 4 different diets
- Measured blood coagulation times from samples taken in a random order
- *As always*, the first step of analysis is **EXPLORATORY**

# Compare distributions with boxplots

- We hope we *don't* see :
    - *outliers* – points outside the whiskers
    - *skewness* – asymmetrical boxes
    - *unequal variance* – clearly unequal box sizes
- *BUT :* don't over-interpret boxplots based on small *n*
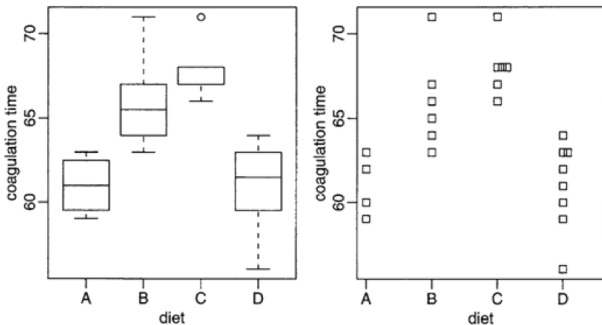
# Trees

- A study is conducted to investigate the growth of a certain type of tree at an elevation of 675 meters
- The variable of interest is the core measurement (in cm) for a 10 year period
- The theory is that the mean should be at least 1.75
- In a random sample of 10 measurements, the mean was 2 with an SD of 0.5

# Hypothesis test for this setup

- Identify the population parameter being tested
  *Here, the parameter being tested is the population mean core measurement $\mu$*

- Formulate the NULL and ALT hypotheses
  $H : \mu = 1.75$ (or $\mu \le 1.75$)
  $A : \mu > 1.75$

- Compute the Test Statistic (TS)
  $t = (2?1.75)/(.5/\sqrt{10}) = 1.58$

- Compute the *p*-value
  *Here,* $p = P(t_9 > 1.58) = 0.07$

- *Decision Rule :* REJECT *H* if the *p*-value $\le \alpha$
  *If we use* $\alpha = 0.05$, *the decision here will be DO NOT REJECT H (but just barely !)*

# More trees

- Now say we are interested in whether the mean core measurement is the same in trees at 675 meters and trees at 825 meters
- Assume that we have a random sample also of size 10 of trees at 825 meters, with a mean core measurement of 2.65 cm and SD 1.15 cm
- How might we test the null that the means are the same, against the alternative that they are different?

# Test for comparing two (independent) means : equal variances

- We want to compare the means of two sets of measures :
    - Group 1 (p. ex. 'control') : $x_1, \ldots, x_n$
    - Group 2 (p. ex. 'treatment') : $y_1, \ldots, y_m$
- We can *model* these data as :
    $$x_i = \mu + \epsilon_i; i = 1, \ldots, n;$$
    $$y_j = \mu + \Delta + \tau_i; j = 1, \ldots, m,$$

    where $\Delta$ signifies the effect of the treatment (compared to the 'control' group)
- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ or $A : \Delta > 0$ or $A : \Delta < 0$

# Equal variances, cont.

- $T = $ obs. diff. $/$ ES(obs. diff.) $= \dfrac{\Delta}{\sqrt{\widehat{Var(\hat\Delta)}}}$ ;

  $\hat\Delta = \bar{y} - \bar{x}$ ; $Var(\hat\Delta) = \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m} = \dfrac{n+m}{nm}\sigma^2$

- We assume that :
    - the variances of the 2 samples are *equal* :
      $Var(\epsilon) = Var(\tau)$
    - the observations are *independent*
    - *the 2 samples are independent*

- We can estimate the variances *separately* :
  $s_x^2 = ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)/(n-1)$
  $s_y^2 = ((y_1 - \bar{y})^2 + \cdots + (y_m - \bar{y})^2)/(m-1)$

- When the variances are *equal*, we can combine the two
  estimators : $s_p^2 = ((n-1)s_x^2 + (m-1)s_y^2)/(n+m-2)$

  $\Rightarrow t_{obs} = \dfrac{\bar{y} - \bar{x}}{\sqrt{s_p^2(n+m)/(nm)}} \sim t_{n+m-2}$ under $H$

# Trees one more time !

- You guessed it ! Now say we are also interested in trees at 975 meters as well
- Want to make a *three-way* comparison
- Have a random sample (size 10 again) and find the mean is 2.5 and the SD is 1
- How might we test the null that all three means are the same, against the alternative that at least one is different ?

# ANOVA

- Abbreviation for *AN*alysis *O*f *VA*riance (analyse de variance)
- But it's a test for a difference in *means*
- The idea :
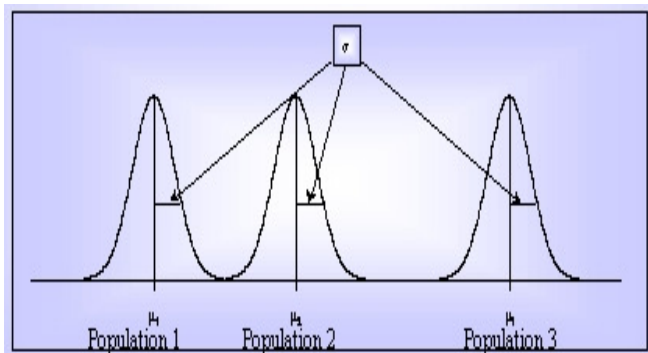
# Principle

- The variation (total sum of squared deviations) consists of *2 components*
    - *individual* fluctuations : variability *intra-group* (error)
    - *between group* fluctuations : variability *inter-group* (treatment)
- Variability inter-group > Variability intra-group
  ⇒ (at least) 2 means are (significantly) different
- General principle :
    - *Decompose* the total sum of squared deviations into its 2 (orthogonal) parts
    - *Test* if the MS*inter* (MSB) is (significantly) bigger than the MS*intra* (MSW, or MSE)

# Hypothesis tests

- Notation :
    - $k$ groups
    - $n_i$ individuals in group $i$
    - observations $x_{ij}$ (observation $j$ from group $i$)
- $H : \mu_1 = \mu_2 = \cdots = \mu_k$
  $A : \exists \, \mu_i \neq \mu_j$ (at least 1 mean is different from the others)
- ANOVA is a rather *robust* test (resultats not too influenced by small deviations from the assumptions

# Pairs of tests : why not ?

Why not start off by carrying out tests ($z$ or $t$) for each pair of samples ?

- For $m$ comparisons (independent), the probability of rejecting at least one $H$ can be expressed as $\alpha_m = 1 - (1 - \alpha)^m$ ; now for $\alpha = 0.05$ :
- 3 tests $\implies$ Type I error $= 0.14$
- 5 tests $\implies$ Type I error $= 0.23$
- 10 tests $\implies$ Type I error $= 0.4$
- 21 tests $\implies$ Type I error $= 0.66$

$\implies$ Type I error no longer controlled at level $\alpha = 0.05$
(anti-conservative/liberal test)

# The models

- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- Under $H$, the model is :

$$x_{ij} = \mu + \epsilon_{ij}$$

- Under $A$, the model is :

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

  where $\alpha_i$ ia *the effect of modality/level i* of facteur $A$ on the variable $X$

- For each model, we can derive an estimator for the residual variance

# Sum of squares

- Goal : test difference between means of two (or more) groups

    - Between SS measures the difference
- The difference must be measured relative to the variance within the groups
    - Within SS
- $F$-test : considers the ratio of $B/W$
- The larger $F$ is, the more significant the difference

# The ANOVA procedure

- Subdivide observed total sum of squares into several components
- Pic appropriate significance point for a chosen Type I error from an $F$ table
- Compare the observed components to test the NULL hypothesis

# Parameter estimation

- Under $H : x_{ij} = \mu + \epsilon_{ij}$ :

$$\hat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}, \quad n = \sum_{i=1}^{k} n_i$$

- Under $A : x_{ij} = \mu + \alpha_i + \epsilon_{ij}$ :

$$\hat{\mu} + \hat{\alpha}_i = \overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad n = \sum_{i=1}^{k} n_i,$$

which gives us $\hat{\alpha}_i = \overline{x}_i - \overline{x}$

$\hat{\epsilon}_{ij} = x_{ij} - \hat{x}_{ij} = x_{ij} - \hat{\mu} - \hat{\alpha}_i = x_{ij} - \overline{x} - (\overline{x}_i - \overline{x}) = x_{ij} - \overline{x}_i$

# Decomposition of the total variation

- The model under $A$ : $x_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- with estimators : $x_{ij} = \overline{x} + (\overline{x}_i - \overline{x}) + (x_{ij} - \overline{x}_i)$
- $\implies (x_{ij} - \overline{x}) = (\overline{x}_i - \overline{x}) + (x_{ij} - \overline{x}_i)$
- with sum of squares :
- $(x_{ij} - \overline{x})^2 = (\overline{x}_i - \overline{x})^2 + (x_{ij} - \overline{x}_i)^2 + 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i)$
- and sums for individuals $(j)$ :

$\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 =$

$n_i(\overline{x}_i - \overline{x})^2 + + \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 + 2(\overline{x}_i - \overline{x})\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)$

# Decomposition, cont.

- Thus, $2(\overline{x}_i \overline{x}) \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i) = 0$, since $\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i) = 0$ $(E[\epsilon_{ij}] = 0)$
- Therefore,

$$\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = n_i(\overline{x}_i - \overline{x})^2 + \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$$

- with the sums for the factor levels :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 + \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$$

$\implies SS_{total} = SS_{groups} + SS_{error}$

# Test principle

- 1-factor analysis of variance tests the effect of one *factor A* having $k$ modalities on the means of a *quantitative variable X*
- The tested hypotheses are :
  $H : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$ vs. $A : \exists\, \mu_i \neq \mu_j$
- Test if the ratio of 2 variance estimators is close to 1
- The variance estimators associated are :
  - *Total variance* : $SS_{total}/(n-1)$
  - *Variance due to factor A* ($MS_{trts}$) : $SS_{trts}/(k-1)$
    $\implies$ estimator of $\sigma^2$ if $H$ is true
  - *Residual variance* ($MS_{error}$) : $SS_{error}/(n-k)$
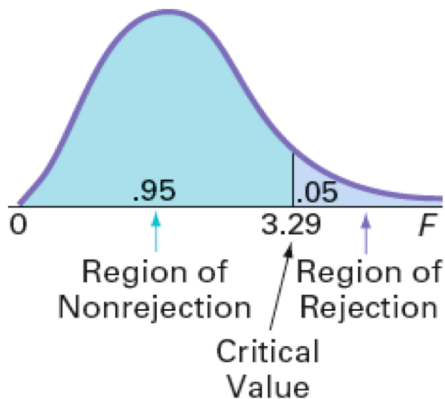    $\implies$ estimator of $\sigma^2$ whichever model

# Test statistic

- Under $H$, $SS_{trts}/(k-1)$ and $SS_{error}/(n-k)$
  $\Rightarrow$ *estimators of the same parameter $\sigma^2$*

- Thus (under $H$), the ratio $\frac{SS_{trts}/(k-1)}{SS_{error}/(n-k)} \approx 1$

- Under $A$, at least 1 $\alpha_i \neq 0$ and $SS_{error}/(n-k)$ is a unique estimator of $\sigma^2$ ; $SS_{trts}/(k-1) >> SC_{error}/(n-k)$

- Thus (under $A$), the ratio $\frac{SS_{trts}/(k-1)}{SS_{error}/(n-k)}$ *much larger than 1*

- $\Rightarrow$ $F$-Test *unilateral* in every case

- $F_{obs} = \frac{SS_{trts}/(k-1)}{SS_{error}/(n-k)} = MS_{trts}/MS_{erreur}$

- Test statistic is distributed according to a Fisher $F$ distribution, with $k-1$ (num) and $n-k$ (denom) degrees of liberty (df)

# ANOVA table

ANOVA table

| source | df | $SS$ | $MS$ ($=SS/$df) | $F$ | $p$-value |
|--------|-----|------|------------------|-----|-----------|
| treatments | $k-1$ | $SS_{trts}$ | $SS_{trts}/(k-1)$ | $MS_{trts}/MS_{error}$ | $P(F_{obs} >$ |
| error | $n-k$ | $SS_{error}$ | $SS_{error}/(n-k)(=\hat{\sigma}^2)$ | | $F_{k-1,n-k})$ |
| total (corr.) | $n-1$ | $SS_{total}$ | | | |



.95   .05

0        3.29        F

Region of Nonrejection / Region of Rejection

Critical Value

34 / 72

# What does it mean when we reject $H$ ?

- The null hypothesis $H$ is a joint (global) one : that *all* the population means are equal
- When we reject the null hypothesis, that does not mean that all the means are different ! !
- It means that *at least one* is different
- To know which is different, we can carry out '*post hoc*'/*a posteriori* tests (pairs of tests, for example – below)

# BREAK

# Model formulas in R

- A simple *model formula* in R looks something like :
  `yvar ~ xvar1 + xvar2 + xvar3`
- We could write this model (algebraically) as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- By default, an intercept is included in the model – you *don ?t* have to include a term in the model formula
- If you want to leave the intercept out :
  `yvar ~ -1 + xvar1 + xvar2 + xvar3`

# More on model formulas

- We can also include interaction terms in a model formula :
  `yvar ~ xvar1 + xvar2 + xvar3`
- Examples :
    - `yvar ~ xvar1 + xvar2 + xvar3 + xvar1 :xvar2`
    - `yvar ~ (xvar1 + xvar2 + xvar3)`$^2$
    - `yvar ~ (xvar1 * xvar2 * xvar3)`

# More on model formulas

- The generic form is `response ~ predictors`
- The predictors can be `numeric` or `factor`
- Other symbols to create formulas with *combinations of variables* (*e.g. interactions*)
    - `+` to *add* more variables
    - `−` to *leave out* variables
    - `:` to introduce *interactions* between two (or more) terms
    - `*` to include *both the interactions and all lower order terms* (`a*b` is the same as `a+b+a:b`)
    - `∧ n` *adds all terms* including interactions up to order *n*
    - `I( )` treats what's inside ( ) as a *mathematical expression*

# Tables of group means for chicks data

|  |  | Groundnut | Soybean | Mean |
|---|---|---|---|---|
| Level of | 0 | 6876 | 7452 | 7064 |
| protein | 1 | 6893 | 6961 | 7927 |
|  | 2 | 6719 | 6624 | 6671 |
| Mean |  | 6763 | 7012 | 6887 |

|  |  | G-nut | Soy | Level of protein | | | Mean |
|---|---|---|---|---|---|---|---|
|  |  |  |  | 0 | 1 | 2 |  |
| Level of | 0 | 6537 | 6752 | 6750 | 6595 | 6588 | 6644 |
| fish | 1 | 6989 | 7273 | 7379 | 7259 | 6755 | 7131 |
| Mean |  | 6763 | 7012 | 7064 | 6927 | 6671 | 6887 |

# Interpreting R output

```
> chicks.aov <- aov(Weight ~ House + Protein*LP*LS)
> summary(chicks.aov)
               Df   Sum Sq  Mean Sq F value     Pr(>F)
House           1   708297   708297 15.8153 0.0021705 **
Protein         1   373751   373751  8.3454 0.0147366 *
LP              2   636283   318141  7.1037 0.0104535 *
LS              1  1421553  1421553 31.7414 0.0001524 ***
Protein:LP      2   858158   429079  9.5808 0.0038964 **
Protein:LS      1     7176     7176  0.1602 0.6966078
LP:LS           2   308888   154444  3.4485 0.0687641 .
Protein:LP:LS   2    50128    25064  0.5596 0.5868633
Residuals      11   492640    44785
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

# R output for the coagulation example

```
> g <- lm(coag ~ diet, data=coagulation)

> summary(g)


Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
   Min     1Q Median     3Q    Max
 -5.00  -1.25   0.00   1.25   5.00

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.6706,        Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

# ASSUMPTIONS

- *Independence* : The $k$ groups (samples) are independent, as well as the individuals within groups ; the ensemble of the $n$ individuals are placed *at random* (*randomization*) between the $k$ modalities for the controled factor $A$, with $n_i$ individuals receiving treatment $i$.
- *Homoscedasticity* : The $k$ populations have the same variance ; the factor $A$ acts only on the *mean* of the variable $X$ and does not change its variance
- *Normality* : The variable studied follows a Normal distribution in the $k$ populations compared (or the CLT applied to the means if the $n_i$ are 'sufficiently large')

# Model assessment : Normality

- Boxplots of observations (or residuals) should be rather symmetric
- A graph of the sample mean vs. variaces should not display any pattern
- QQ-plot (normal) plot of the observations (or residuals) should form a straight line
- Check whether there are any unusual or influential values

# Model evaluation : Homogeneity of variance

- Boxplots of the observations should show similar variability
- Variability of the residuals should be similar in teh graph of residuals versus fitted values
- It is also possible to carry out formal hypothesis tests (e.g. Bartlett, Levene), but these are not useful for diagnosing problems

# Some diagnostic plots
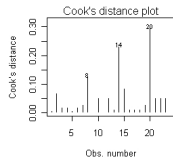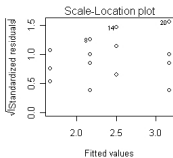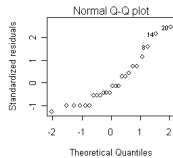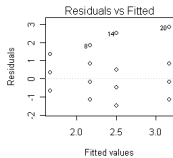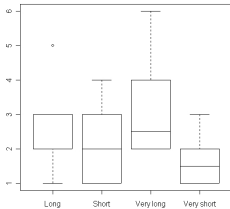
# Evaluation of the model : Independence

- Graphics : residuals vs. group mean, might indicated autocorrelation for example
- Normally, treat the question of independence during the conception of the expermient, for example using randomization or perhaps other methods

# ANOVA : after the test

- Once all the conditions for an ANOVA have been verified and the analysis carried out, two conclusions are possible :
    - we reject $H$
    - we do not have enough evidence to reject $H$
- If $H$ is not rejected, we conclude that there are not significant differences between group means
- If we DO reject $H$, typically we are interested in *identifying the modalities/factor levels* that are responsible for the significant result

# Multiple comparisons

- Comparing means of pairs of treatments
- Carried out *after* a significant ANOVA
- Types of comparisons
    - planned (*a priori*) : indpendent of the ANOVA results ; the theory predicts which treatments should be different
    - unplanned (*a posteriori*) : the comparisons are decided *based on the ANOVA results*
- $H : \mu_i = \mu_j$ vs. $A : \mu_i \neq \mu_j$
- Test statistic

$$t = \frac{\overline{y}_i - \overline{y}_j}{\sqrt{\hat{\sigma}^2 \left(1/n_i + 1/n_j\right)}}$$

- $(\hat{\sigma}^2 = MS_{error})$ ; df $= df_{error}$

# Bonferroni method – global control

- To maintain the global level $\alpha_e$ at level $\alpha$, we must *adjust $\alpha$ for each comparison by the total number of comparisons*
- In this way, $\alpha_e$ is *independent* of the number of comparisons
- Simplest method : method of Bonferroni

$$\alpha' = \alpha/k,$$

where $k$ = number of comparisons (tests)

- $p_{adj} = \min(kp, 1)$
- Bonferroni's method assures that the global level is *no larger than the desired level*
- (That property makes this method *conservative*, and thus less powerful than other methods, but it is applicable for any situation)

# Multiple comparisons : Tukey Honest Significant Differences

- Interested in simultaneous confidence intervals or tests for *differences in the mean outcome X* for *pairs of levels* of a factor
- To test all pairwise comparisons among means using the *Tukey HSD*, calculate HSD for each pair of means :

$$q_s = \frac{M_i - M_j}{\sqrt{MSW/n_{group}}},$$

  where $M_k$ is the mean of group $k$, $M_i > M_j$
- For hypothesis testing, compare $q_s$ to a $q$ value from the *studentized range distribution* (difference between largest and smallest sample means divided by pooled sample SD *sqrt2/n*)
- Reject the null at level $\alpha$ if $q_s > q_\alpha$
- CI : $(\overline{y}_i - \overline{y}_j) \pm \frac{q_{\alpha;k;N-k}}{\sqrt{2}} \hat{\sigma}_e \sqrt{2/n}$; $i, j = 1, \ldots, k$, $i \neq j$
- $k$ = number of populations ; $N$ = total sample size

# (Complete) Randomized block design

- Assume that the hamsters have come from 4 *different litters*, 2 hamsters per litter
- We expect that hamsters born in the same litter are *more similar* to each other than hamsters from a different litter
- For each pair of hamsters *randomly assign* short or long to one member of each pair
- Example (toss a fair coin, for example) :
  S, L // L, S // S, L // S, L
- Analysis : 2-way analysis of variance

# Replication, Randomization, Blocking

- These are the 'big three' of experimental design
- **Replication** – to reduce *random variation* of the test statistic; increases generalizability
- **Randomisation** – to reduce/remove *bias*
- **Blocking** – to reduce *unwanted variation*
- Idea here is that units within a block are similar to each other, but different *between* blocks
- 'Block what you can, randomize what you cannot'

# Factorial crossing

- Compare 2 (or more) sets of conditions in the *same experiment*
- Designs with factorial treatment structure allow you to measure *interaction* between two (or more) sets of conditions that influence the response
- Factorial designs may be either *observational* or *experimental*

# Interaction

- Interaction is very common (and very important) in science
- Interaction is a *difference of differences*
- Interaction is *present* if the effect of one factor *is different* for different levels of the other factor
- *Main effects can be difficult to interpret in the presence of interaction*, because the effect of one factor *depends* on the level of the other factor
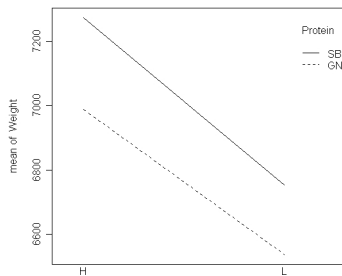
## Factorial experimental design and interaction

- Example : hibernation study
  - General question : *How do changes in an animal's environment induce hibernation ?*
  - *Specific question* : What is the effect of changing daylight duration on the enzyme concentration of the sodium pump in two golden hamster organs ?
- Compare two (or more) sets of conditions in the *same experiment* : long/ short <u>AND</u> heart/brain
- In this example, there are 4 combinations of conditions : – Long/Heart, Long/Brain, Short/Heart, Short/Brain
- Interaction = 'difference of differences'
- There is an *interaction* when the effect of the association of combined treatments is not the sum of treatment effects
- In the case of interaction, the effect of a treatment *varies according to whether it is associated with the other treatment*
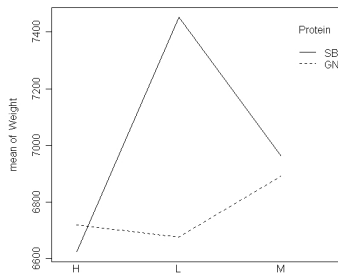- The interpretation of individual effects is more difficult when interation is present

# Interaction plot

### pas d'interaction



### interaction

# Advantages of factorial experiments

- *More efficient* (powerful) than a series of experiments studying one factor at a time
- Permits estimation of *interaction* between sets of conditions that may affect the response
- All data are used for effect estimation

# 2-way ANOVA

- Simulataneous study of a factor $A$ with $I$ levels and a factor $B$ with $J$ levels
- For each pair of levels $(A, B)$ :
    - we have a sample
    - all samples are of the *same size* $n$ (balanced design)
- Suppositions :
    - the populations studies are Normally distributed
    - the population variances are all equal (homoscedasticity)
    - the samples are taken randomly and independently in the populations

# Complete model

- The *complete model* : with interaction
- $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0$, $Var(\epsilon_{ijk}) = \sigma^2$, $Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ si $(ijk) \neq (i'j'k')$

ANOVA table

| source | df | SS | MS | F |
|--------|-----|-----|-----|-----|
| A | $I-1$ | $nJ\sum_{i=1}^{I}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$ | $SS_A/df_A$ | $MS_A/MS_{err}$ |
| B | $J-1$ | $nI\sum_{j=1}^{J}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$ | $SS_B/df_B$ | $MS_B/MS_{err}$ |
| AB | $(I-1)(J-1)$ | $n\sum_{j=1}^{J}\sum_{i=1}^{I}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$ | $SS_{AB}/df_{AB}$ | $MS_{AB}/MS_{err}$ |
| error | $IJ(n-1)$ | $\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{ij\cdot})^2$ | $SS_{err}/df_{err}$ | |
| total (corr.) | $nIJ-1$ | $\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{\cdots})^2$ | | |

# Hypothesis tests

- Test for interaction
  $H : \gamma_{ij} = 0, \ i = 1, \ldots, I, \ j = 1, \ldots, J$
- Test statistic :
  $F_{AB} = MS_{AB}/MS_{error} \sim F_{(I-1)(J-1), IJ(n-1)}$ under $H$
- Test for effect of factor $A$
  $H : \alpha_i = 0, \ i = 1, \ldots, I$
- Test statistic :
  $F_A = MS_A/MS_{error} \sim F_{I-1, IJ(n-1)}$ sous $H$
- Test for effect of factor $B$
  $H : \beta_j = 0, \ j = 1, \ldots, J$
- Test statistic :
  $F_B = MS_B/MS_{error} \sim F_{J-1, IJ(n-1)}$ sous $H$

# Additive model

- The *additive model* : without interactions
- $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
- $E[\epsilon_{ijk}] = 0$, $Var(\epsilon_{ijk}) = \sigma^2$, $Cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$ id $(ijk) \neq (i'j'k')$

ANOVA Table

| source | df | SS | MS | F |
|--------|-----|-----|-----|-----|
| A | $I-1$ | $nJ\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2$ | $SS_A/df_A$ | $MS_A/MS_{err}$ |
| B | $J-1$ | $nI\sum_{j=1}^{J}(\bar{y}_{.j.} - \bar{y}_{...})^2$ | $SS_B/df_B$ | $MS_B/MS_{err}$ |
| error | $nIJ-I-J+1$ | $\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$ | $SS_{err}/df_{err}$ | |
| total (corr.) | $nIJ-1$ | $\sum_{k=1}^{n}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{...})^2$ | | |

# Indicator variables for the model

- The matrix form for the linear model :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- According to the form of the matrix $\boldsymbol{X}$, we are in the case of :

    - *linear regression* ($\boldsymbol{X}$ is then comprised of the constant 1 and $p$ explanatory variables), or
    - *factorial model* ($\boldsymbol{X}$ is comprised of **indicator variables associated with the levels** of the factor(s))
    - *ancova* ($\boldsymbol{X}$ is comprised of both qualitative and quantitative variables)

# Example : ToothGrowth

- "The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."
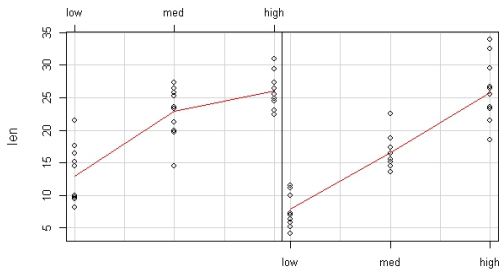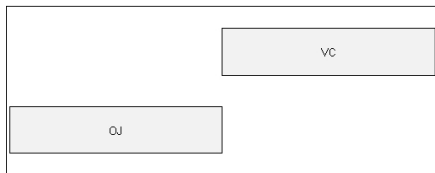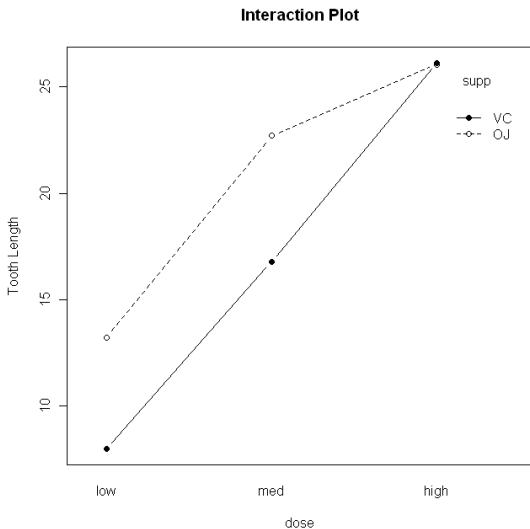


**Boxplots of Tooth Growth Data**
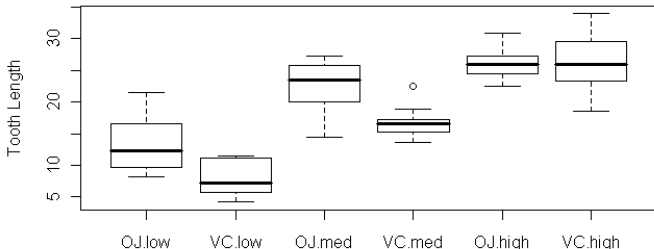
# Example, cont : Graphics



ToothGrowth data: length vs dose, given type of supplement

# Example, cont. : Interaction plot



**Interaction Plot**

# Example, cont : ANOVA table output

```
> aov.out = aov(len ~ supp * dose, data=ToothGrowth)
> summary(aov.out)

            Df Sum Sq Mean Sq F value   Pr(>F)
supp         1  205.3   205.3  15.572 0.000231 ***
dose         2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose    2  108.3    54.2   4.107 0.021860 *
Residuals   54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

# Unbalanced designs

- When all sample sizes are equal, the main effects and interactions can be estimated independently *independently*
- That ;s because of the orthogonality of the sub-spaces that correspond to the different model effects
- This is no longer the case when the sample sizes are different (unbalanced case) :
  *SSModel ≠ SSA + SSB + SSAB*
- For an unbalanced design, effect estimation must be *adjusted* (for the other effects in the model) : the estimated values depend on the other terms in the model and their order of entry
- We can no longer carry out tests $F = \frac{MS_x}{MSerror}$
- We must carry out sub-model tests

|     | L    | M    | H    |
|-----|------|------|------|
| VC  | 4.2  | 16.5 |      |
|     | 11.5 | 16.5 | 23.6 |
|     | 7.3  | 15.2 | 18.5 |
|     |      | 17.3 |      |
| OJ  | 15.2 |      | 25.5 |
|     | 21.5 | 19.7 | 26.4 |
|     | 17.6 | 23.3 | 22.4 |
|     | 9.7  |      | 24.5 |

# Example, cont. : supp 1st

```
> # full interaction model with
> # supp entering first
>
> fit1 <-
    lm(len ~ supp + doselev + supp:doselev,
      data=toothun)
> anova(fit1)
Analysis of Variance Table

Response: len
              Df Sum Sq Mean Sq F value    Pr(>F)
supp           1 174.46  174.46 17.3664 0.0011049
doselev        2 375.75  187.87 18.7012 0.0001495
supp:doselev   2  17.70    8.85  0.8808 0.4377931
Residuals     13 130.60   10.05
```

# Example, cont : doselev 1st

```
> # full interaction model with doselev
> # entering first
>
> fit2 <-
    lm(len ~ doselev + supp + supp:doselev,
      data=toothun)
> anova(fit2)
Analysis of Variance Table

Response: len
              Df Sum Sq Mean Sq F value    Pr(>F)
doselev        2 396.08  198.04 19.7131 0.0001158
supp           1 154.13  154.13 15.3428 0.0017685
doselev:supp   2  17.70    8.85  0.8808 0.4377931
Residuals     13 130.60   10.05
```

# Summary : numerical and graphical analysis

- Design plot
- Boxplots of outcome for each factor level
- Interaction plots
- Write out model, assumptions, de
  ne all parameters
- ANOVA table
- Plots for assumption checking/model assessment
- Example of full analysis at :
  https://www.guru99.com/r-anova-tutorial.html