Statistics 2

I- Test of Hypothesis

→ One-Sample

- σ Known: Z-test: $Z = \frac{\bar{X} - H}{\sigma/\sqrt{n}}$
- σ unKnown: t-test: $t = \frac{\bar{X} - H}{S/\sqrt{n}}$ $df = n-1$

→ Two Samples:

1→ Independent

- σ Known: Z-test $z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- σ unKnown

  - σ Equal
    - pooled standard deviation
    $sp^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
    - $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{sp^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
    - $df = n_1+n_2-2$
  - σ unEqual
    - Adjust the degrees of freedom
    $df = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$
    - $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

2→ Dependent Samples → t-test: ① Compute mean and standard deviation for the sample differences

N.B!!
in this case we work on $Md : M_1 - M_2$

② $t = \frac{\bar{d}}{sd\sqrt{n}}$

$df = n - 1$

→ More than 2 Samples

Analysis of Variance (ANOVA)

One-way (one factor)

F-test

$SST\,df = K - 1$ (K: nb of treatment)

$SSE\,df = n - K$ (n: nb of observations)

$MST = SST / K-1$

$MSE = SSE / n-K$

$F = MST / MSE$

Two-way (two factors)

One-way + add blocking variable

. $SST\,df = K - 1$

. $SSB\,df = b - 1$ (b: nb of blocks)

. $SSE\,df = (K-1)(b-1)$

F (for treatments) = MST/MSE

F (for blocks) = MSB/MSE

Interaction

$SST\,df = K - 1$

$SSB\,df = b - 1$

$SSI\,df = (K-1)(b-1)$

$SSE\,df = n - Kb$

F (for interaction) = MSI/MSE

Pairs of Means

. Used when H_0 is rejected to find the pair of mean that differ

. $CI = (\bar{X}_1 - \bar{X}_2) \pm t\sqrt{MSE\left(\frac{1}{n1} + \frac{1}{n2}\right)}$ $df = n - K$

. if $0 \in CI$, no difference

. if $0 \notin CI$, difference between treatment means

→ Correlation Analysis

. Group of techniques to measure the relationship between two variables

. In addition to graphic techniques, we'll develop numerical measures to describe the relation.

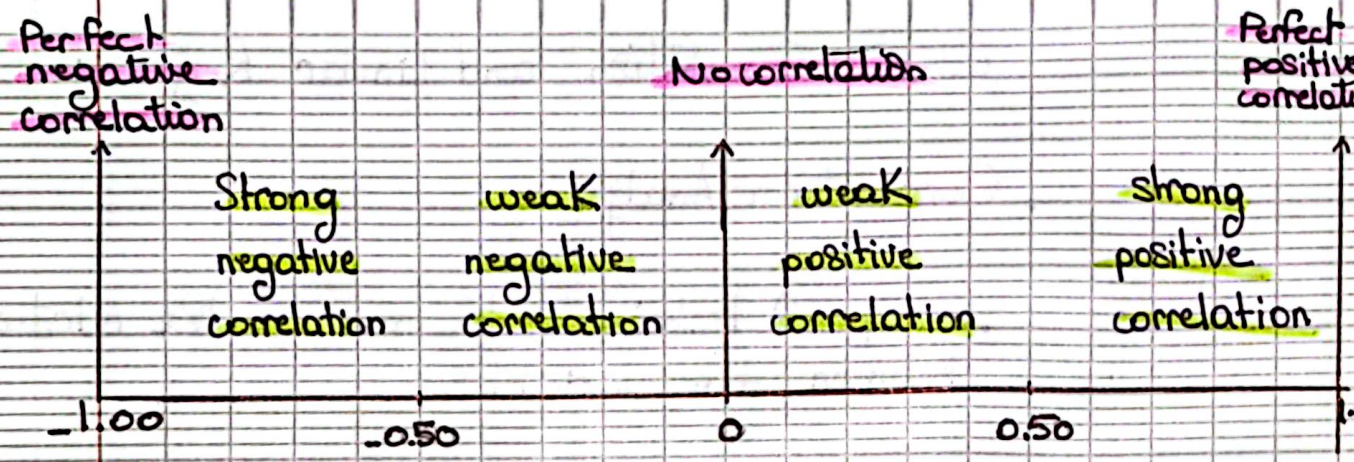
↳ Scatter Diagram

· Graphic tool used to portray the relationship between two variables
. Independent variable (predictor): X_axis
. Dependent variable (estimated) : Y_axis

↳ Correlation Coefficient (r)

. A measure of the stregth of the linear relationship between two variables

. it ranges from -1.00 to 1.00
. IF it's 0, there's no association
. A value near 1.00 => positive correlation
. A value near -1.00 => negative correlation



$$r = \frac{\sum (X - \bar{X})(y - \bar{y})}{(n-1)\, s_x\, s_y}$$

$\bar{X}$: mean of independent var
$\bar{y}$: mean of dependent var
n: nb of observation
s_x / s_y: standard deviation

Testing the Significance of r

- We test the significance of r to determine whether the observed correlation between two variables in a sample is statistically significant or simply due to chance.

- So we will le ρ represent the correlation in the population and conduct a hypothesis test to find out.

Step 1 : State Ho and H1

Ho : $\rho \leq 0$ / $\rho \geq 0$ / $\rho = 0$

H1 : $\rho > 0$ / $\rho < 0$ / $\rho \neq 0$

Step 2: Select the level of significance

$\alpha = \ldots$ (given)

Step 3: Select the test statistic

we use t test

Step 4: Formulate the decision rule

. Find the critical value from t table

(α, df = n - 2)

reject Ho if t ...

Step 5: Make the decision

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

reject or not reject Ho

Step 6: Interpret

There is / there isn't correlation with respect to the nb of sales calls made and the number of copiers sold in the population salespeople

→ Regression Analysis

. In regression analysis we estimate one variable based on another variable

. Dependent variable: being estimated
. Independent variable: used to estimate the dependent variable

. Regression equation :
Equation that expresses the linear relationship between two variable

$$\hat{y} = a + bx$$

$\hat{y}$: estimated value of y for a selected x
a : constant or intercept
b : slope of the fitted line
x : value of the independent variable

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$a = \bar{y} - b\bar{x}$$

Regression Equation Slop Test

. For a regression eq, the slope is tested for significance
. We test the hypothesis that the slope of the line in the population is 0
. reject Ho => there is no relationship between the two variable
. df in this test → n-2

Step1: State H1 and Ho
Ho : $\beta \leq 0$ / $\beta \geq 0$ / $\beta = 0$
H1 : $\beta > 0$ / $\beta < 0$ / $\beta \neq 0$

Step 2: Select the level of significance
α = ... (given)
Step 3: Select the test statistic
we use t-test

Step 4 : Formulate the decision rule
. Find the critical value from t-table
(α = ... / df = n-2 / one-tail or 2 tail)

reject Ho if t...

Step 5: make the decision
$t = \frac{b - 0}{Sb}$ reject or accept Ho

The Standard Error of Estimate

A measure of the dispersion, or scatter of the observed values around the line of regression for a given value of x

$$S_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

- if $S_{y,x}$ is small, this indicates that the data are relatively close to the regression line. and the regression eq can be used
- if it is large, the data are widely scattered around the regression line and the regression eq will not provide a precise estimate of y.

Coefficient of Determination

- The square of the correlation coefficient (r)
- ranges from 0 to 1.0
- the proportion of the variation in the dependent variable y that is explained by the variation in the independen variable x

ex: 74.8% of the variation in the nb of copiers sold is explained by the variation in sales calls

Relationship among r, r^2, and $s_{y,x}$

- The Standard Error of Estimation measure how close the actual value are to the regression line.
  - → when it is small, the two variables are closly related

- The correlation coefficient measures the strength of the linear association between two variables.
  - → when points on the scatter diagram are close to the line, & r tends to be large.

- Therefore, the correlation coefficient and the standard error of estimation are inversely related.

Confidence and Prediction Intervals

confidene interval: predict the mean value of y for a given x

$$\hat{y} \pm t\, s_{y,x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2}}$$

prediction interval: predict an individual y for a given value of x

$$\hat{y} \pm t\, s_{y,x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$