1 Subject

Parametric hypothesis testing for comparison of two or more populations. Independent and dependent samples.

The tests for comparison of population try to determine if K (K \ge 2) samples come from the same underlying population according to a variable of interest (X). We talk parametric test when we assume that the data come from a type of probability distribution. Thus, the inference relies on the parameters of the distribution. For instance, if we assume that the distribution of the data is Gaussian, the hypothesis testing relies on mean or on variance.

We handle univariate test in this tutorial i.e. we have only one variable of interest. When we want to analyze simultaneously several variables, we talk about multivariate test.

Samples are independent if the response of the nth person in the second sample is not a function of the response of the nth person in the first sample. Independent samples are also called uncorrelated or unrelated samples. Dependent samples, also called related samples or correlated samples, occur when the response of the nth person in the second sample is partly a function of the response of the nth person in the first scheme; matched-pairs studies; etc.)¹.

The test that we describe in this tutorial can be used to compare process (e.g. is it that two machines produce the same diameter bolts), but it enables also to test the connection that can exist between a categorical variable and a quantitative variable (e.g. is that women drive slower on average than men on the roads).

2 Dataset

The CREDIT_APPROVAL.XLS² data file describes 50 households consisting of married couples, both actives, which asked for a credit to a bank. The available variables are:

Variable	Description
Sal.Homme	Logarithm of the male salary
Sal.Femme	Logarithm of the female salary
Rev.Tete	Logarithm of the income per head in the household
Age	Logarithm of the male age
Acceptation	Acceptation of the credit
Garantie.Supp	Additional guarantee is required
Emploi	Type of job held by the reference person of the household

All the continuous variables are potentially a variable of interest; the categorical variables are used to define the sub-populations.

¹See <u>http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm</u>

² <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_approval.xls</u>

3 Descriptive statistics and normality test

3.1 Importing the data file into Tanagra

The simplest way to launch Tanagra is to open the data file into Excel. We select the data range; then we click on the Tanagra menu installed with the TANAGRA.XLA add-in³. After we checked the coordinates of the selected cells, we click on OK button.



Tanagra is automatically launched. A new diagram is created. We verify that the data set has **50 observations** and **7 variables**.

³ See <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html</u>

🌋 TANAGRA 1.4.26 - [Da	itaset (tan7F.txt)]						
🟆 File Diagram Component	t Window Help						- 8 ×
D 🚅 🔲 🙀							
Ar	nalusis						
Dataset (tap7E tyt)		Dataset (descrip	otion			
		7 attribute(s) 50 example(s)					
		Attribute	Category	Informations			
		Sal.Homme	Continue	-			
		Sal.Femme	Continue	-			
		Rev.Tete	Continue	-			
		Age	Continue	-	>		=
		Acceptation	Discrete	2 values			
		Garantie.Supp	Discrete	3 values			
		Emploi	Discrete	2 values	J		
μ							
Data visualization	Statistics	Nonnarametric s	onents tatistics	Instance	selection	Feature construction	
Feature selection	Regression	Factorial ana	lusis	D	15	Clustering	
Sov learning	Meta-spy learning	Sov learning acce	ssment	500	vrina	Association	
	t ~ -	pho (earning asso	ssment	500	n ing	HSSOCIACION	
Correlation scatterplot	Scatterplot with label						
Scatterplot	View pultiple soatter	alot					
	ore on matciple scatter						
·							

3.2 Descriptive statistics

In a first step, we want to describe our dataset by computing some descriptive statistics indicators on quantitative variables. We add the DEFINE STATUS component into the diagram. We set Sal.Homme, Sal.Femme, Rev.Tete and Age as INPUT.



Then we insert the MORE UNIVARIATE CONT STAT component (STATISTICS tab). We click on the VIEW menu. We obtain the following results.

💯 TANAGRA 1.4.26 - [More Univariate cont stat 1]								
Tile Diagram Component Window Help								- 8 ×
D 📽 🖬 😘								
Analysis								~
🖃 🥅 Dataset (tan7F.txt)				Results				
😑 🚰 Define status 1	Attribute	St	ats		Hi	istogram		
More Univariate cont stat 1		Stati	stics	Values	Count	Percent	Histogram	
Parameters	_	Average	7.4640	x < 6.2960	2	4.00%		
Execute		Median	7.4650	(20(0		2.00%	-	
VIEW		Std dev. [Coef of variation]	0.5619 [0.0753]	6.2960_=<_X_<_6.5720	1	2.00%	<u>.</u>	
1		MAD [MAD/STDDEV]	0.4024 [0.7162]	6.5720_=<_x_<_6.8480	3	6.00%		
1		Min*Max (Full range)	6.02*8.78 [2.76]	6.8480_=<_x_<_7.1240	3	6.00%		
x	Sal.Homme	1st*3rd quartile [Range]	7.26*7.83 [0.57]	7.1240_=<_x_<_7.4000	11	22.00%		
		Skewness (std-dev)	-0.2334 (0.3366)	7 4000 - 4 x 4 7 6760	15	30.00%		
		Kurtosis (std-dev)	0.6809 (0.6619)	//4000*_x_*_//0/00		30.00%		
				7.6760_=<_x_<_7.9520	7	14.00%		
							-	<u>U</u>
Data visualization Statistics	Nonnarametri	c statistics Instance	Components selection Feature (construction Feature	e selection	n	Regression	
Factorial analysis	Cluster	ing Sov le	arning Meta-si	ny learning Sny learning	mzasze se	ent	Scoring	
Association		ing option		er to an	.5			
🖽 ANOVA Randomized Blocks 🛛 🖾 Fisher's test 🔪 💊	ĢLe	evene's test	🕂 Normality Test	‡‡ Paired T-Test		🖄 Semi-j	partial Correlation	述 Univariate
Bartlett's test 👖 Group characterizat	ion 🖉 Li	near correlation	📶 One-way ANOVA	11 Paired V-Test		T-Tes	t	🛄 Univariate
🚓 Brown - Forsythe's test 🛛 🚺 Group exploration	<u>∧</u> ₩	ore Univariate cont stat	🗠 One-way MANOVA	🖉 Partial Correlation		T-Tes	t Unequal Variance	👥 Univariate
<								>

With such tools, we try primarily to detect anomalies (e.g. skewed or bimodal distribution, the existence of outliers, etc.). It seems that there is no need to worry too much here.

3.3 Normality test

The tests we present in this tutorial assume Gaussian distribution of variables of interest. Certainly they are more or less robust to this assumption. However, to ensure the viability of the results (and show you how to do), we decided to conduct this checking^{4,5}.

A first empirical evaluation was possible with the previous component. We know that the ratio between the MAD (mean absolute deviation) and the SD (standard deviation) is approximately 0.8 if the sample comes a Gaussian population⁶. Thus if the ratio differs from this value, we suspect that the underlying distribution is not Gaussian (the inverse is not true, if the ratio is 0.8, it does not mean that the underlying distribution is necessarily Gaussian).

We mainly use this empirical rule for detecting the strong deviations. On our dataset, it seems that none of the variables diverge strongly to the Gaussian distribution according this rule: we 0.7162 for « Sal.Homme »; 0.7846 for « Sal.Femme », 0.7986 for « Rev.Tete » and 0.7895 pour « Age ». Apart from "Sal.Homme, we are even surprisingly close to the reference value.

⁴ See <u>http://data-mining-tutorials.blogspot.com/2008/11/normality-test.html</u> for more details about the normality test.

⁵ In the context of comparison of population, performing a normality test of the whole sample before the comparison is not always relevant. Indeed, if the alternate hypothesis is true, the conditional distribution (in the sub-population) is actually Gaussian, but the distribution of the whole sample is not Gaussian (e.g. a bimodal distribution can be a characteristic a significant difference between the means of the sub populations). It is more suitable to perform the normality test within each subgroup.

⁶ More precisely $\sqrt{\frac{2}{\pi}} \approx 0.798$. See <u>http://en.wikipedia.org/wiki/Mean_absolute_deviation</u>

Now we proceed to a more formal evaluation using normality tests. We add the NORMALITY TEST (STATISTICS tab) into the diagram. We click on the VIEW menu.



At the 5% significance level, all the variables seem compatible to the Gaussian distribution.

4 Tests for independent samples

4.1 Comparison of two means with equal variances

We want to compare the female salary according to the credit acceptation. In other words, we want to know if the credit acceptation relies on the salary of the female⁷.

We are in a comparison of means scheme. We use the t-test of difference in means. We assume that the variances are the same in the two populations (acceptation of credit is yes or no)^{8,9}.

We add the DEFINE STATUS component into the diagram using the shortcut into the toolbar. We set SAL.FEMME as TARGET, ACCEPTATION as INPUT.

⁷ The opposite interpretation is not relevant. The credit acceptation cannot influence the salary of the female.

⁸ http://www.itl.nist.gov/div898/handbook/prc/section3/prc31.htm

⁹ <u>http://en.wikipedia.org/wiki/Student's_t-test</u>



Then, we insert the T-TEST component (STATISTICS tab). We click on the VIEW menu in order to obtain the results.

1.4.26 - [T-Test 1]								
Tile Diagram Component Window Help								_ 8 ×
Analysis				T-Te	st 1			
🖃 🏢 Dataset (tan7F.txt)				Param	eters			
🖻 🎇 Define status 1	Parameters							
More Univariate cont stat 1	Sort results r	10						
	-							
Tar Tart 1				Resi	utts			
Parameters	Attribute_Y	Attribute_X		Description	1			Statistical test
Execute			Value	Examples	Average	Std-dev	-	0.4063 / 0.1448 =
View	Sal.Femme A		oui	34	7,4394	0.5483	·	2.806296
/ T		Acceptation	non	16	7.0331	0.2615	d.f.	48.00
1			AU	50	7.3094	0.5099	p-value	0.007219
	Computation t Created at 14	ime : 0 ms. /07/2008 18:16:	13					
			Components					
Data visualization Statistics	-Nonp:	arametric statis	tics Instan	ce selection	Feature co	onstruction	Fea	ture selection
Regression Factorial analysis		PL5	Cli	ustering	Spv le	arning	Me	ta-spv learning
Spv learning assessment Scoring		Association						
🖾 Fisher's test 🛛 🛱 Levene's test		\land Normalit	y Test	17 Paired T-T	est	🕍 Semi	-partial Co	orrelation 🗷 🗷 Univariate
Group characterization 👘 🕍 Linear correls	tion	M One-way	ANOVA	tr ¹ Paired V-T	est	— 🛄 Т-Те	st	📶 Univariate
🔚 Group exploration 🛛 🖾 More Univaria	te cont stat	🗠 One-way	MANOVA	🖉 Partial Cor	relation	🛄 T-Te	st Unequa	I Variance 😐 Univariate
<								>

The t-statistic is t = 0.4063 / 0.1448 = 2.8063. The difference is significant at the 5% significance level since the p-value is 0.007219. The credit acceptation is related to the female salary in the household.

It must however take with caution this result. The conditional variances are quite different; they range from simple to double. And the numbers of instances are unbalanced. The condition of applicability of the test-statistic for equal variances does not seem satisfied. We must refine the analysis by introducing the t-test for unequal variances.

4.2 Comparison of two means – Unequal variances

This version does not use the pooled variance for the estimation of the variance of the deviation. The degree of freedom is also corrected. We add the T-TEST UNEQUAL VARIANCE component (STATISTICS tab) into the diagram. We obtain the following results.



In comparison to the previous statistic, the numerator is the same, but the denominator is different. We thus obtain t = 0.4063 / 0.1145 = 3.5476. The degree of freedom is d.f. = 47.96, so if we round it to the nearest integer, d.f. \approx 48. The p-value of the test is lower p-value = 0.00088. The salary gap according to the acceptance of credit is confirmed.

4.3 Comparison of means for K populations – Equal variances (ANOVA)

We can see the one-way analysis of variance (ANOVA) as a generalization of the comparison of means for more than two populations. We want to compare women's wages in 3 groups defined by the type of additional warranty used by borrowers (GARANTIE.SUPP). Initially, we consider that the variances are identical in groups.

We add the DEFINE STATUS component into the diagram. We set SAL.FEMME as TARGET, GARANTIE.SUP as INPUT.



We insert the ONE-WAY ANOVA component (STATISTICS tab). We obtain the following results.

💯 TANAGRA 1.4.26 - [One-way ANOVA 1]											
💯 File Diagram Component Window Help								_ 8 ×			
Analysis			One-way	ANOVA 1				^			
🖃 🏢 Dataset (tan7F.txt)	UIIC-WAY ARUVA 1										
🚍 🚰 Define status 1	Danamotors		i ui ui								
- 🕂 More Univariate cont stat 1	Fai ameters										
- 🕂 Normality Test 1	SOLC LEZALCE HO										
🖨 🎦 Define status 2											
T-Test 1			Kes	uus			a				
T-Test Unequal Variance 1	ATTIDUTE_Y ATTIDUTE_X		Description	1			statistical test				
Define status 3		∀alue	Examples	Average	Std-dev	Va	riance decompositio	n			
		hypotheque	29	7.1762	0.4615	Source	Sum of square	d.f.			
	>	caution	5	7.3700	0.4684	BSS	1.3248	2			
View	Sal Femme Garantie Sunn	non	16	7,5319	0.5524	WSS	11.4175	47			
	Sam Simile Garancistapp	AU	50	7.3094	0.5099	TSS	12.7423	49			
N							Significance level				
N						Statistics	Value	Proba			
						Fisher's F	2.726664	0.075794 🧹			
		0									
Data visualization	Nonparametric st	atistics Instance	e selection	Feature	construction	Fe	ature selection				
Regression Factorial ana	lucis PIS	Clus	tering	Sov	learning	AA	ata-sny learning				
Coulearning accercment Cooring	Accocition						sea she coarring	1			
Sporte annung assessment Scoring	HSSOCIACION				* 30						
A Hisher's test	rene's test 🔨 🖄 Normality Test 📰 Paired T-Test 🖉 Semi-partial Correlation 🖄 Univariate										
Group exploration 🖉 Linear co	ur correlation Mill Une-way ANUVA Mill Paired V-Test Mill Integration Mill University and Mill University of the Antiper Mill Correlation Mill Tract Manager Mill University of the Antiper Mill Original An										
an oroup exploration Dic/More Uni	Nore Univariate contistat induced way (MANUVA 😰 Partial Correlation induced Variance 😶 Univariate i										
<								>			

We obtain the conditional means and the conditional standard deviation. The F-statistic is F = 2.7267. The deviation between means seems not significant at the 5% level since the p-value of the test is p-value = 0.07579.

Even if the conditional standard deviation seems similar, the size of the groups is highly unequal in our dataset. We know that the test is very sensitive to the Behrens-Fisher problem. It is more careful to implement the test which does not rely on the pooled variance. We use the Welch's version.

4.4 Comparison of means for K populations – Unequal variances

This version is adapted to the case where the variances are unequal within subgroups. We insert the WELCH ANOVA component (STATISTICS tab) into the diagram.



The F-statistic is 2.3446. The first degree of freedom is the number of groups minus 1. The second degree of freedom is more complicated to obtain. We round it to the nearest integer (d.f. $2 \approx 11$). The p-value of the test is 0.141856. The deviation between the means is not significant at the 5% level.

4.5 Comparison of 2 variances – F-Test

The tests for differences in variances are often used prior to the tests for differences in means. They would thus determine the most appropriate procedure. But their utilization is not limited to this task. The comparison of the variance can be the purpose of the analysis. For instance, we want to know if the disposal of the students in a classroom influences the disparity of their scores. In this context, the comparison of variances is at least as interesting as the comparison of means.

In our dataset, we want to compare the variances of the female wages according to the acceptation of the credit. We insert the FISHER'S TEST component (STATISTICS tab) into the diagram¹⁰, under the DEFINE STATUS 2 component. Thus, the target attribute is SAL.FEMME, and the input one is ACCEPTATION. We click on the VIEW menu.

¹⁰ http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm

TANAGRA 1.4.26 - [Fisher's test 1]										
File Diagram Component Window Help								-	5	
D 🗳 🔲 👪										
Analysis			F	isher's test 1						
🖃 🏢 Dataset (tan7F.txt)		Darameters								
😑 👪 Define status 1	Paramet	ers								
👫 More Univariate cont stat 1	Sort result	ts po								
Normality Test 1	Jore resu	6 110								
Define status 2				Results						
T-Test I Inequal Variance 1	Attribute	Y Attribute X		Description			Statistic	caltest		
Fisher's test 1			Value	Examples	Average	Std-dev	Te	st		
🔪 😑 🚰 Define s Parameters			oui	34	7,4394	0.5483	Fisher	4.3962		
Cne Execute	Sal.Femme	Acceptation	non	16	7.0331	0.2615	df	33/15		
Weld View		>	AU	50	7,3094	0.5099	p-value	0.0037		
N	_									
	Computati	on time : 0 ms.								
	Created at	: 14/07/2008 18:3	5:28						~	
N		(Components							
Data visualization Statistic	os	Nonparametric :	tatistics Instan	ce selection	Feature	e construction	n			
Feature selection 🔨 Regression	on	n Factorial analysis PLS Clustering								
Spv learning 💦 🔨 Meta-spv lea	arning	Spv learning ass	essment 9	Scoring	As	sociation				
🖽 ANOVA Randomized Blocks 🔪 🚠 Fisher's	test		ene's test	🕂 Normali	ty Test	‡l‡ P	aired T-T	est		
🕁 Bartlett's test 🛛 🚺 Group o	haracterizati	on 🕍 Line	ar correlation	📶 One-wa	y ANOVA	t <mark>i</mark> t [‡] P	aired V-T	est		
🚰 Brown - Forsythe's test 🛛 🚺 Group e	xploration	<u>i</u> Mor	e Univariate cont stat	🔄 One-wa	y MANOVA	🛃 P	artial Co	rrelation	1	
<									>	
			/						_	

The F-statistic is F = 4.3962. The degrees of freedom are 33 and 15. The p-value of the test is 0.0037. The deviation between the conditional variances is significant. About the comparison of means above, the test for unequal variance seems the most suitable.

4.6 Comparison of K variances – Bartlett's test

The generalization for the comparison of K (K \ge 2) variances is the Bartlett's test. We want to compare the variances of the female salary according to the guarantee. We insert the BARTLETT'S TEST (STATISTICS tab) into the diagram.

TANAGRA 1.4.26 - [Bartlett's test 1]												
Tile Diagram Component Window Help									-	Ξ×		
Analysis	Bartlett's test 1											
🖃 🏢 Dataset (tan7F.txt)					Parameters							
🖃 🎎 Define status 1	Baran	eters								'		
- 🕂 More Univariate cont stat 1	raran	icicits										
🖳 🕂 Normality Test 1	Sort res	ults no										
🖃 🏧 Define status 2												
🛄 T-Test 1					Results							
	Attribu	te_Y Attribute_X			Description	l i		Statistic	altest			
Fisher's test 1			۱	/alue	Examples	Average	Std-dev	Tes	t			
🖨 🏰 Define status 3		emme Garantie.Supp	hyp	otheque	29	7.1762	0.4615	Pooled var.	0.2429			
One-way ANOVA 1	Sal.Fem		C	aution	5	7.3700	0.4684	Bartlett's T	0.6390			
Butlett's tost 1		>		>		non	16	7.5319	0.5524	df	2	
bartiett's test 1				AU	50	7.3094	0.5099	p-value	0.7265			
										~		
			Compor	ients								
Data visualization Statist	cs	Nonparametric st	atistics	Instance	e selection	Feature	construction					
Feature selection Regress	ion	Factorial anal	ysis	F	PLS	Clu	stering					
Spv learning 🔰 Meta-spv le	arning	rning Spy learning assessment Scoring Association										
🔢 ANOVA Randomized Blocks 👘 🏹 Fisher's	sher's test 🖄 Normality Test 🖽 Paired T-Test											
🛱 Bartlett's test 🖌 👖 Group	characterization 🛛 🕍 Linear correlation 🙀 One-way ANOVA 🔢 🖬 Paired V-Test											
🚰 Brown - Forsythe's test 🛛 🔠 Group	xploration 🗹 More Univariate cont stat 🕍 One-way MANOVA 🛛 🖄 Partial Correlation											
										>		
										.::		

The Bartlett's statistic is T = 0.6390, the p-value is 0.7265. The differences in variances are not significant at the 5% level.

4.7 Tests for difference in variances – More robust tests

The F-Test and the Bartlett's test are very sensitive to the departures to the normality assumption¹¹. It is more suitable to use a more robust procedure such as the Levene test for equality of ($K \ge 2$) variances¹². It is reliable even if we deal with a skewed distribution. It is as powerful as the Bartlett's test when we deal with a Gaussian distribution.

4.7.1 The Levene test

TANAGRA 1.4.26 - [Levene's test 1] 📅 File Diagram Component Window Help - 3) D 🛩 🖬 🎎 Analysis Levene's test 1 🖃 🏢 Dataset (tan7F.txt) 🚊 🙀 Define status 1 Parameters 🕂 More Univariate cont stat 1 Sort results no 🕂 Normality Test 1 🚊 🎦 Define status 2 🗽 T-Test 1 🗽 T-Test Unequal Variance 1 Attribute_Y Attribute_X Description Statistical test 🚠 Fisher's test 1 Examples Average Value Std-dev Test 🔁 Define status 3 hypotheque 29 7.1762 0.4615 Levene's 0.321175 📶 One-way ANOVA 1 5 7.3700 0.4684 caution Sal.Femme Garantie.Supp 🚻 Welch ANOVA 1 2/47 16 7.5319 0.5524 non -> 🕁 Bartlett's test 1 p-value 0.726876 All 50 7.3094 0.5099 🚠 Levene's test 1 -Computation time : 0 ms < 1 > Created at 14/07/2008 18:41:15 Components Data visualization Nonparametric statistics Instance selection Feature construction Statistics Regression Feature selection Factorial analysis PLS Clustering Meta-spv learning Spv learning assessment Association Spv learning Scoring ANOVA Randomized Blocks 🚠 Fisher's test 🕂 Normality Test ±≓ Paired T-Test 🖾 Levene's test 🕁 Bartlett's test Group characterization 🖄 Linear correlation 📶 One-way ANOVA 1 Paired V-Test 🔆 Brown - Forsythe's test 🔚 Group exploration 🔣 More Univariate cont stat 🛛 🛃 One-way MANOVA 🕍 Partial Correlation >

We insert the LEVENE'S TEST component (STATISTICS tab) into the diagram.

The Levene's statistic is W = 0.321175. Under the null hypothesis (the variances are equals), it follows a Fisher distribution with (2; 47) degrees of freedom. The p-value of the test is p-value = 0.726876. The differences in variances are not significant at the 5% level.

4.7.2 The Brown-Forsythe's test

This version seems the most robust among the tests for equality of variances¹³. We insert the BROWN-FORSYTHE'S TEST component (STATISTICS tab) into the diagram.

¹¹ http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm

¹² http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

¹³ <u>http://en.wikipedia.org/wiki/Brown–Forsythe_test</u>

👷 TANAGRA 1.4.26 - [Brown - Forsythe's test 1]											
💇 File Diagram Component Window Help 📃 🖉											
🗅 🛩 🖬 🎎											
Analysis Brown - Forsythe's test 1											
🖃 🏢 Dataset (tan7F.txt)		Parameters									
🖨 🚰 Define status 1											
- 🖟 More Univariate cont stat 1	raram	eters									
🖳 🕂 Normality Test 1	Sort res	ults no									
🖃 🙀 Define status 2											
🗽 T-Test 1					Results						
🗽 T-Test Unequal Variance 1	Attribut	te_Y Attribute_X			Description	I		Statisti	caltest	=	
			١	/alue	Examples	Average	Std-dev	Τe	est		
🖻 – 🚰 Define status 3			hyp	otheque	29	7.1762	0.4615	Brown &			
One-way ANOVA 1	Sal Femr	Sal.Femme Garantie.Supp	Ci	aution	5	7.3700	0.4684	Forsythe's	0.326772		
Welch ANOVA 1	Jan on		samenine darance.sapp		non	16	7.5319	0.5524	W		
Bartlett's test 1				AU	50	7.3094	0.5099	df	2/47		
Levene's test 1								p-value	0.722874		
Brown - Forsythe's test 1-											
	Computa	tion time : 0 ms.								~	
			Compon	ients							
Data visualization Statistic	s	Nonparametric st	atistics	Instance	e selection	Feature	construction				
Feature selection Regression	n	Factorial analy	vsis	F	PLS	Clu	stering				
Spv learning Meta-spv lea	rning	Spv learning asses	sment	Sc	oring	Ass	ociation				
📰 ANOVA Randomized Blocks 🛛 🚠 Fisher's t	間ANOVA Randomized Blocks 조·Fisher's test 조·Levene's test 《 Normality Test 印 Paired T-Test										
🚰 Bartlett's tast 🚺 Group characterization 🖉 linear correlation 🔐 One-way ANOVA 🔢 Paired V-Test											
🚰 Brown - Forsythe's test 🛛 🔢 Group e:	exploration 🖾 More Univariate cont stat 🛃 One-way MANOVA 📝 Partial Correlation										
5 - 16										_	

5 Tests for dependent samples

The aim of the utilization of the dependent samples is to reduce (to control) the uncertainty due to within group variation. It enables to increase the sensitivity of the test¹⁴.

We say also matched-pairs samples or paired samples.

5.1 Paired sample t-test¹⁵

We want to know if in a household, the man tends to have a different salary than his wife.

It is important not to implement a test for independent samples i.e. compare the mean of wages for men with the mean for women. Indeed, **the comparison must be done within each household**. We are in a test scheme for paired samples.

We insert again the DEFINE STATUS component into the diagram. We set SAL.HOMME as target, SAL.FEMME as INPUT.

¹⁴ <u>http://www.jerrydallal.com/LHSP/paired.htm</u>

¹⁵ <u>http://www.stat.psu.edu/online/development/stat500/lesson10/lesson10_04.html</u>

http://www.statsdirect.com/help/parametric_methods/ptt.htm

http://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples

TANAGRA 1.4.26 - IBrown - Forsythe's test 1]	Dofine attribute statures	
Tile Diagram Component Window Help	Parameters	_ = = ×
Analysis Analysis Dataset (tan7F.txt) C. Define status 1 Normality Test 1 Define status 2 L. T. T. T. t. 1	Attributes : C Sal Femme C Sal Femme C Age C Age D Acceptation D Garantie Supp E Emploi	
Attribute_Y T-Test Unequal Variance 1 Fisher's test 1 Fisher's test 1 Fisher's variance 1 Fisher's test 1 Fisher's variance 1 Fi	Image: Clear selection OK Clear selection Define attribute statuses	Statistical test dev Test 0.4615 Brown 8. 0.4684 Forsythe's 0.326772 W 0.5524 df 2/47 0.5099 p-value 0,722874
Another Statistics No Data visualization Statistics No Feature selection Regression Spy learning Spy AnovA Randomized Blocks Fisher's test Computation test Spy	Parameters Attributes : C Sal Homme C Bel Ferme R Rev Tete C Age D Acceptation G Garantie Supp Emploi	uction § n t⊟ Paired T-Test tH Daired V Test
Brown - Forsythe's test	E E Clear selection OK Cancel Help	Partial Correlation

We add the PAIRED T-TEST component (STATISTICS tab) into the diagram.



The mean of the difference between the male and female salaries within households is D.AVG = 0.1546. The test statistic is t = 3.86972, with a p-value = 0.000322. The difference is significant.

5.2 Comparison of variances for 2 paired samples

We can apply the same approach for the comparison of variances. We insert the PAIRED V-TEST component (STATISTICS tab) into the diagram.



The standard deviation of the male salary is 0.561863; for the female, it is 0.509948. The ratio between the variances is $(0.561863)^2 / (0.509948)^2 = 1.213974$. But the male and female salaries are highly correlated with r = 0.865438. The test statistic is then t = 1.342750. The p-value of the test is 0.185669. The difference between the variances is not significant at the 5% level.

5.3 Comparison of means for K paired samples

We can extent the comparison of means for two dependent samples to the comparison for K > 2 paired samples. In the design of experiments, we say "randomized complete block design"^{16,17,18}.

We want to implement the ANOVA RANDOMIZED BLOCKS (STATISTICS tab) on the comparison of wages (section 5.1). We can thus compare the results of the two approaches.

The definition of the status of the variable is a little bit different for this component. We add the DEFINE STATUS component. We set both SAL.HOMME and SAL.FEMME as input.

Note: We can set as many variables as we want in INPUT. Simply they must be continuous variables.

¹⁶ <u>http://en.wikipedia.org/wiki/Randomized_block_design</u>

¹⁷ http://itl.nist.gov/div898/handbook/pri/section3/pri332.htm

¹⁸ <u>http://www.jerrydallal.com/LHSP/ranblock.htm</u>



We insert the ANOVA RANDOMIZED BLOCKS component into the diagram.



What interests us is the primary source of variability explained by the "treatment" i.e. the difference between SAL.HOMME and SAL.FEMME. The test statistic F = 14.9747 follows a Fisher distribution at (1, 49) degrees of freedom under the null hypothesis. Again, the difference is significant with a p-value of 0.00032.

Equivalence with the paired t-test for K = 2 populations. We know that there is a relation between the Student (T) and the Fisher (F) distributions. Indeed, $[T(m)]^2 \equiv F(1,m)$. About our tests, we observe that $\sqrt{F} = \sqrt{17.9747} = 3.86972 = t$. It is the value of t-statistic for the paired t-test (section 5.1). The approaches are totally equivalent for K = 2.