# Using SAS/STAT<sup>®</sup>:

A Gentle Introduction to Some Frequently Used Tools

Melodie Rush Customer Loyalty Principal Data Scientist

LinkedIn: https://www.linkedin.com/in/melodierush\_ Twitter: @Melodie\_Rush



## Scenario

- You work for a supermarket and the supermarket is offering a new line of organic products. Management wants to determine which customers are likely to purchase these products.
- So you decided to send coupons to customers that are in your loyalty program so you can see which ones buy items from the new organic line.
- You have collected the data and now you need to determine information about your customers that have bought the organic line items.

## • You have Base SAS and SAS/STAT (*SAS/Graph is useful too!*)



## Data

Variable	Description
ID	Unique Customer ID
DEMAFFL	Affluence grade on a scale from 1 to
DEMAGE	Age, in years
DEMCLUSTER	Type of Residential Neighborhood -
DEMCLSUTERGROUP	Neighborhood group - 7 levels
GENDER	M=Male, F=Female
DEMGREGION	Demographic Region
LOYALTYSTATUS	Loyalty status: tin, silver, gold, or pl
PROMSPEND	Total amount spent
PROMTIME	Time as loyalty card member
TARGETBUY	Organics purchased? 1=Yes, 0=No
TARGETAMT	Number of organic products purcha
PREPROMAMT	Number of organic products purcha
DIFF_AMT	TARGETAMT - PREPROMAMT

Copyright © SAS Institute Inc. All rights reserved.





## Data SAS Code

**PROC PRINT** DATA=mydata.organics (OBS=**10**) OBS="Row number" LABEL; VAR ID DemAffl DemAge DemCluster DemClusterGroup DemGender DemReg PromClass PromSpend PromTime TargetBuy TargetAmt; RUN;



## Data SAS Enterprise Guide

Tas	ks Program To	ols			Ĩ	🗎 List Data Fin	st 10	rows for C:\Pu	ublic\Organics\orga
	Data	•				Data Options		Data	
	Describe	۲		List Data		Properties		Data source Task filter:	: C:\Public\Organic
	Graph	•	Σ	Summary Statistics Wizard	•			Variables to a	assign:
	ANOVA	•	Σ	Summary Statistics				Name Name	
	Regression	•		Summary Tables Wizard				1 Dem Affl 1 Dem Age	
	Multivariate			Summary Tables		1	Li Da	st Data First 1 ata	0 rows for C:\Publi
	Survival Analysis	•		List Report Wizard			Op Tit Pn	otions les operties	Rows to list
	Capability			Characterize Data					First n rows
	Control Charts		մհ	Distribution Analysis					Amount (n):
lín	Pareto Chart			One-Way Frequencies		[2000]			Print the row n
	Time Series	•		Table Analysis	ŀ	Preview			Column headir
	Data Mining	•							Print number of
	OLAP	►							Round values
	Task Templates	Þ							output)

C/Public/Organics/organics asa7bdd None gr: Taik roles:	ic\Organics\organics.sas7bdat	23	
None  gr: Task roles:  DemAft DemAft DemAge DemAge DemAge DemCluster  ows for C1/Public/Organics/organics.sas7bdat  Cptions  First n rows Heading direction Defaut Heading direction Defaut Heading direction Defaut Heading direction Defaut Defaut Defaut Defaut First n rows Row number Column width Defaut First Nimum Unform Unform Unform Unform Divide page into sections (no effect on HTML Spit labels at:  Divide page into sections (no effect on HTML Spit labels at:  Run Save Cancel Help SSCSS	C:\Public\Organics\organics.sas7bdat		
gr: Task roles: DemAil DemAge DemAge DemAge DemAge DemAge DemAge DemAge DemAge Demage Demage Demage Demage Demage Demage Demage Demage Default Hostcortal Minimum Outionn headings: Row number Column width Default Outical Print the row number Column headings: Row number Column headings: Row number Column headings Print number of rows Print number of rows Divide page into sections (no effect on HTML Spit labels at: Rum    Save Cancel Help	None		
ows for C\Public\Organics\organics.sas7bdat     Options     Pirst n rows     Amount (n):     10     Print the row number   Column width   Image: Default     Column width   Image: Default   Vertical     Vertical <th>ign: Task roles:</th> <th></th> <th></th>	ign: Task roles:		
Rows to lid       Heading direction         Pirst n rows <ul> <li>Horizontal</li> <li>Vertical</li> <li>Vertical</li> <li>Column heading:</li> <li>Row number</li> <li>Column width</li> <li>Default</li> <li>Full</li> <li>Minimum</li> <li>Uniform</li> <li>Uniform by</li> <li>Split labels at:</li> <li>Split labels</li></ul>	ows for C:\Public\Organics\organics.sas7bdat Options	Line Line & Sector Street	
Print the row number   Column width   © Default   © Column heading:   Row number   Print   Print number of rows   Olivide page into sections (no effect on HTML   Split labels at:     Run   Save   Cancel     Help	Rows to list First n rows Amount (n): 10	Heading direction Default Horizontal Vertical	
Divide page into sections (no effect on HTML Split labels at:	<ul> <li>Print the row number</li> <li>Column heading: Row number</li> <li>Use variable labels as column headings</li> <li>Print number of rows</li> <li>Round values before summing</li> </ul>	Column width <ul> <li>Default</li> <li>Full</li> <li>Minimum</li> <li>Uniform</li> <li>Uniform by</li> </ul>	
Run V Save Cancel Help	☑ Divide page into sections (no effect on HTML output)	Split labels at:	
222		Run 🔻 Save Cancel	Help Sas

Preview code

## Data

### Organics Data Table (first 10 rows)

Row number	Customer Loyalty ID	Age	Gender	Loyalty Status	Total Spend	Loyalty Card Tenure	Organics Purchase Indicator	Organics Purchase Count	PrePromAmt	Diff_Amt
1	000000140	76		3. Gold	16000	4	0	0	0	0
2	000000620	49		3. Gold	6000	5	0	0	0	0
3	000000868	70	Female	2. Silver	0.02	8	1	1	0	1
4	0000001120	65	Male	1. Tin	0.01	7	1	1	0	1
5	000002313	68	Female	1. Tin	0.01	8	0	0	0	0
6	0000002771	72		4. Platinum	20759.81	3	0	0	0	0
7	000003131	74	Female	1. Tin	0.01	8	0	0	0	0
8	000003328	62	Male	1. Tin	0.01	5	0	0	0	0
9	0000004529	62	Male	2. Silver	2038.76	3	0	0	0	0
10	000005886	43	Female	3. Gold	6000	1	1	1	0	1

Observations 22,223 Variables 14



## **First Things First...** Categorical Variables - SAS Code

**PROC FREQ** DATA=mydata.organics; TABLES TargetBuy TargetAmt PromClass DemGender; RUN;



# First Things First...

Categorical Variables - SAS Enterprise Guide

Tas	ks Program To	ools				
	Data	F			One-Way F	requencies Initial Values f
	Describe	×		List Data	Data Statistics	Data
	Graph	Þ	Σ	Summary Statistics Wizard	_ Plots Results Titles	Data source: C:\ Task filter: Non
	ANOVA	×	Σ	Summary Statistics	Properties	
	Regression	×		Summary Tables Wizard	-	Variables to assign: Name
	Multivariate	Þ		Summary Tables		(▲ ID () DemAffl
	Survival Analysis	Þ		List Report Wizard		100 Dem Age A Dem Cluster A Dem Cluster Group
	Capability	×		Characterize Data		DemGender
	Control Charts	×.	Jh	Distribution Analysis		A Dem TVReg
lîn	Pareto Chart		▦	One-Way Frequencies		100 Prom Spend 100 Prom Time
	Time Series	×		Table Analysis		TargetBuy
	Data Mining	•				The selection pane
	OLAP	Þ			Preview c	ode
	Task Templates	Þ				

or C:\Public\Organics\organics.sas7bdat
Public \Organics \organics.sas7bdat
Jp E Control C
enables you to choose different sets of options for the task.
Run 🔻 Save Cancel Help



## **First Things First...** Categorical Variables

### The FREQ Procedure

Organics Purchase Indicator					
TargetBuy Frequency Perce					
0	16718	75.23			
1	5505	24.77			

Organics Purchase Count						
TargetAmt Frequency Percent						
0	16718	75.23				
1	4625	20.81				
2	715	3.22				
3	165	0.74				

Loyalty Status					
LoyaltyStatus Frequency Perce					
1. Tin	6487	29.19			
2. Silver	8572	38.57			
3. Gold	6324	28.46			
4. Platinum	840	3.78			

Gender						
DemGender Frequency Percen						
	2512	-				
F	12149	61.64				
М	5815	29.50				
U	1747	8.86				



## **First Things First...** Categorical Variables

### The FREQ Procedure

Organics Purchase Indicator				
TargetBuy Frequency Perce				
0	16718	75.23		
1	5505	24.77		

Organics Purchase Count						
TargetAmt Frequency Percent						
0	16718	75.23				
1	4625	20.81				
2	715	3.22				
3	165	0.74				

Loyalty Status						
LoyaltyStatus Frequency Perc						
1. Tin	6487	29.19				
2. Silver	8572	38.57				
3. Gold	6324	28.46				
4. Platinum	840	3.78				

Gender							
Gender	Frequency	Percent					
	4259						
Female	12149	67.63					
Male	5815	32.37					

### Frequency Missing = 4259



## **First Things First...** Continuous Variables – SAS Code

- **PROC MEANS** DATA=mydata.organics VARDEF=DF MEAN STD MIN MAX N NMISS; VAR PromSpend PromTime DemAge; RUN;
  - \* Use PROC UNIVARIATE to generate the histograms;
- TITLE1 "Summary Statistics";
- TITLE2 "Histograms";
- **PROC UNIVARIATE** DATA=mydata.organics NOPRINT;
  - VAR PromSpend PromTime DemAge;
    - HISTOGRAM ;
- RUN;



## **First Things First...** Continuous Variables - SAS Enterprise Guide

Tasks	Program T	ools				Σ Summary S	Statistics Initial Values for	C:\Public\	Organics\orga
	Data	Þ				Data Statistics	Data		
	Describe	Þ		List Data		Basic Percenti	les Data source: C:\	Public\Org	janics\organics
	Graph	•	Σ	Summary Stat	istics Wizard	Addition: Plots Results	al Task filter: Nor	Σ Su	mmary Statist
	ANOVA	•	Σ	Summary Stat	istics	Titles Properties	Variables to assign:	Da	ta atistics
	Regression Multivariate	•		Summary Tab	les Wizard		ID DemAffl		Basic Percentiles Additional
	Survival Analysis			Summary Tab	les		(12) DemAge (12) DemCluster (12) DemClusterGro	Pio Re u Titl	ts sults les
	Capability Control Charts	) }		Characterize D Distribution A	zard )ata nalysis		DemGender     A DemReg     DemTVReg     A PromClass     B PromSpend     D PromTime		penes
lîn -	Pareto Chart			One-Way Freq	juencies		100 TargetBuy 100 Target∆mt		
	Time Series Data Mining	+		Table Analysis			The selection pane	er	
	OLAP	•		Data Statistics	Plots	review	code		
	Task Templates	×		Percentiles Additional	Generated plots				
				Plots Results Titles Properties	<b>□ ■ ■ ■</b> ■ Box and white	sker			Preview code

Statistics > Basic     Basic statistics   Veriance   Variance   Mode   Degrees of freedom   Sum   Sum of weights   Number of missing values   The selection pane enables you to choose different sets of options for the task.	Initial Values for C/Public/Organics/	organics cas	Zbdat	t		23
Basic statistics          Mean       Maximum decimal places:         Standard deviation       1         Standard error       1         Variance       Iminimum         Minimum       Divisor for standard deviation and variance:         Mode       Degrees of freedom         Range       Sum of weights         Sum of weights       Number of missing values	Statistics > Basic	organics.sas	,/Dual			<u></u>
Mean Maximum decimal places: Standard deviation   Standard error   Variance   Minimum   Maximum   Sum   Sum of weights   Number of missing values      The selection pane enables you to choose different sets of options for the task.	Basic statistics					
Standard deviation Standard error Variance Minimum Maximum Mode Range Sum Sum of weights Number of observations Number of missing values The selection pane enables you to choose different sets of options for the task.	V Mean			Maximum decim	al places:	
Standard error         Variance         Minimum         Maximum         Mode         Degrees of freedom         Range         Sum         Sum of weights         Number of observations         Number of missing values	Standard deviation			1	*	
Vanance         Minimum       Divisor for standard deviation and variance:         Mode       Degrees of freedom         Range       Sum         Sum       Sum of weights         Number of observations       Number of missing values	Standard error					
Minimum       Divisor for standard deviation and variance:         Mode       Degrees of freedom         Range       Sum         Sum of weights       Number of observations         Number of missing values       Number of missing values	] Vanance					
Maximum         Mode         Degrees of freedom         Range         Sum         Sum of weights         Number of observations         Number of missing values	] Minimum			Divisor for stand variance:	lard deviation and	
Mode       Degrees of freedom         Range         Sum         Sum of weights         Number of observations         Number of missing values	] Maximum I Mada					E
Sum of weights Number of observations Number of missing values The selection pane enables you to choose different sets of options for the task.	Papaa			Degrees of fre	edom 👻	
Sum of weights Number of observations Number of missing values The selection pane enables you to choose different sets of options for the task.						
Number of observations Number of missing values The selection pane enables you to choose different sets of options for the task.	Sum of weights					
Number of missing values	Number of observations					
The selection pane enables you to choose different sets of options for the task.	Number of missing values					
The selection pane enables you to choose different sets of options for the task.						
The selection pane enables you to choose different sets of options for the task.						
The selection pane enables you to choose different sets of options for the task.						
The selection pane enables you to choose different sets of options for the task.						-
	The selection pane enables you to choos	e different set	s of options	for the task.		
	,					
						+
	J	<u> </u>				10lp



# First Things First...

## Continuous Variables

## The MEANS Procedure

Variable	Label	Mean	Std Dev	Minimum	Maximum	Ν	N Miss
PromSpend	Total Spend	4420.6	7559.0	0.0	296313.9	22223	0
PromTime	Loyalty Card Tenure	6.6	4.7	0.0	39.0	21942	281
DemAge	Age	53.8	13.2	18.0	79.0	20715	1508

The UNIVARIATE Procedure





## **First Things First...** Continuous Variables - SAS Code

**PROC UNIVARIATE** DATA = mydata.organics CIBASIC (TYPE=TWOSIDED ALPHA=0.05) MU0=0;

VAR PromSpend DemAge PromTime;

HISTOGRAM PromSpend DemAge PromTime / NORMAL W=1 L=1 COLOR=YELLOW MU=EST SIGMA=EST) CFRAME=GRAY CAXES=BLACK WAXIS=1 CBARLINE=BLACK CFILL=BLUE PFILL=SOLID ;

RUN;



## **First Things First...** Continuous Variables - SAS Enterprise Guide

Tas	ks Program Too	s		Data Distributions	Data
	Data Describe		List Data	Summary Normal Lognormal Exponential	Data source: C:\Public\Organics\c Task filter: None
	Graph	Σ	Summary Statistics Wizard	Weibull Beta	Variables to assign:
	ANOVA	Σ	Summary Statistics	Gamma Kemel	Name
	Regression Multivariate		Summary Tables Wizard Summary Tables	Plots Appearance Inset Tables Titles Properties	<ul> <li>ID</li> <li>ID</li> <li>DemAffl</li> <li>DemAge</li> <li>DemCluster</li> <li>DemClusterGroup</li> <li>DemGender</li> </ul>
	Capability	- =	List Report Wizard Characterize Data		DemReg     DemTVReg     PromClass     PromSpend
	Control Charts	ենե	Distribution Analysis		Prom Time
lm	Pareto Chart		One-Way Frequencies	Data	Distributions >
	Time Series		Table Analysis	Summary	
	Data Mining	•		Normal	Available distribution
	OLAP	•		Lognormal Exponentia	al 🔽 Normal
	Task Templates	•		Weibull Beta	🔳 Lognomal
				Gamma Kernel	🔲 Exponentia





# A picture is worth...

### The UNIVARIATE Procedure





# A picture is worth...

### The UNIVARIATE Procedure





## Is it Normal?

### The UNIVARIATE Procedure



Goodness-of-Fit Tests for Normal Distribution								
		Statistic	p Value					
orov-Smirnov	D	0.0493570	Pr > D	<0.010				
von Mises	W-Sq	10.8931792	Pr > W-Sq	<0.005				
n-Darling	A-Sq	80.1149533	Pr > A-Sq	<0.005				
					1			



## Just to clear things up...

# Base

- FREQ
- MEANS
- UNIVARIATE
- CORR

# STAT • TTEST

- NPAR1WAY
- ANOVA
- REG
- LOGISTIC



# Associations

Copyright © SAS Institute Inc. All rights reserved.



## Association

- An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable



# Tests of Association

## NULL HYPOTHESIS

 There is no association between GENDER and TARGETBUY

## **ALTERNATIVE HYPOTHESIS** There is an association

- TARGETBUY
- The probability of purchasing organic items is the same whether you are male or

female

between GENDER and

The probability of purchasing organic items is different between males and females



## Initial Analysis SAS Code

PROC FREQ DATA = mydata.organics2 ORDER=INTERNAL; TABLES Gender \* TargetBuy / FORMAT=COMMA8. NOROW NOCOL NOPERCENT EXPECTED NOCUM ALPHA=0.05;

RUN;



## Initial Analysis SAS Enterprise Guide

Tas	ks	Program	Tools		
	Da	ata	•		
	De	escribe	•		List Data
	Gr	raph	•	Σ	Summary Statistics W
	A	NOVA	F	Σ	Summary Statistics
	Re	egression	•		Summary Tables Wiz
	М	ultivariate	•		Summary Tables
	Su	irvival Analys	is 🕨		List Report Wizard
	Ca	apability	•		Characterize Data
	Co	ontrol Charts	•		Distribution Analysis.
lín	Pa	areto Chart			One-Way Frequencie
	Ti	me Series	•		Table Analysis
	Da	ata Mining	•		
	0	LAP	F		
	Ta	ask Template	s 🕨		





## Initial Analysis SAS Enterprise Guide

Table Analysis Initial C	crosstab for SASApp:MYDATA.ORGANICS2			Table Ana	lysis Initial Cro	rosstab for SASApp:MYDATA.ORGANICS2
Data Tables Cell Statistics Table Statistics Association Agreement Ordered Differences Trend Test	Data Data source: SASApp:MYDATA.ORGAN Task filter: None Variables to assign: Ta	IICS2	Edit	Data Tables Cell Statistic Table Statis Associa Agreem	cs stics stion nent d Differences	Tables       Preview:         Variables permitted in table:       Preview:         TargetBuy       TargetBuy
Results Cell Stat Results Table Stat Results Titles Properties	Name   ID   ID   DemAffl   DemCluster   DemClusterGroup   DemReg   DemTVReg   PromSpend			Trend To Computa Results Cell Stat Table Si Titles Properties		Tables to be generated:
Preview code	1     TargetBuy       1     TargetAmt       1     PrePromAmt       1     Diff Δmt	Data Tables Cell Statistics Table Statistics Association Agreement	Cell Statistics Available statistics Cumulative column percentages	entages		Gender by TargetBuy <select a="" begin="" defining="" new="" table="" to="">  Delete</select>
		Trend Test Computation Options Results Cell Stat Results Table Stat Results	<ul> <li>Column percentages</li> <li>Cell frequencies</li> <li>Cell percentages</li> <li>Missing value frequenci</li> <li>Cell contribution to Pear</li> </ul>	es son chi-square	code	Run 🔻 Save Cancel Help
		litles Properties	<ul> <li>Cell frequency deviation</li> <li>Expected cell frequency</li> <li>Percentage of total freq</li> <li>Include percentages in frequency</li> </ul>	from expected / uency the data set		S



# Categorical Variable Test

Copyright © SAS Institute Inc. All rights reserved



## Chi-Square Test

### The FREQ Procedure





Tal	Table of Gender by TargetBuy						
	TargetBuy(Organics Purchase Indicator)						
er)	0	1	Total				
ale	7,944 8,652	4,205 3,497	12,149				
ale	4,849 4,141	966 1,674	5,815				
	12,793	5,171	17,964				
F	Frequency Miss	ing = 4259					



## Chi-Square Test

Chi-square tests and the corresponding p-values

- Determine whether an association exists
- Do not measure the strength of an association
- Depend on and reflect the sample size

## o-values ts ociation



# p-value for Chi-Square Test

- Probability of observing a chi-square statistic at least as large as the one actually observed, given that there is not association between the variables
- Probability of the association you observe in the data occurring by chance



## Chi-Square Test SAS Code

PROC FREQ DATA = mydata.organics2 ORDER=INTERNAL; TABLES Gender \* TargetBuy / FORMAT=COMMA8. NOCOL NOPERCENT CELLCHI2 EXPECTED NOCUM CHISQ ALPHA=0.05;







## Chi-Square Test SAS Enterprise Guide

Task	cs Program To	ools	
	Data	F	
	Describe	×	🔟 List Data
	Graph	Þ	Σ Summary Statistics Wizard
	ANOVA	×	Σ Summary Statistics
	Regression	F	Summary Tables Wizard
	Multivariate	F	Summary Tables
	Survival Analysis	F	List Report Wizard
	Capability	F	Characterize Data
	Control Charts	Þ	Distribution Analysis
Шī	Pareto Chart		One-Way Frequencies
	Time Series	F	Table Analysis
	Data Mining	Þ	
	OLAP	Þ	Data
	Task Templates	Þ	Tables Cell Statistics

Data Tables Cell Statistics Table Statistics Association Agreement Trend Test Results Cell Stat Results Table Stat Results Titles Properties









# Adding Chi-Square to our FREQ Output

### The FREQ Procedure

Frequency	Table of Gender by TargetBuy							
Expected Cell Chi-Square		TargetBuy(O	rganics Purcha	ase Indicator)				
Row Pct	Gender(Gender)	0	1	Total				
	Female	7,944 8,652 57.915 65.39	4,205 3,497 143.28 34.61	12,149				
	Male	4,849 4,141 121 83.39	966 1,674 299.35 16.61	5,815				
	Total	12,793	5,171	17,964				
	F	requency Miss	sing = 4259					





# Adding Chi-Square to our FREQ Output (Continued)

## Statistics for Table of Gender by TargetBuy

Statistic	DF	Value	Prob
Chi-Square	1	621.5507	<.0001
Likelihood Ratio Chi-Square	1	662.3524	<.0001
Continuity Adj. Chi-Square	1	620.6729	<.0001
Mantel-Haenszel Chi-Square	1	621.5161	<.0001
Phi Coefficient		-0.1860	
Contingency Coefficient		0.1829	
Cramer's V		-0.1860	





# Strength of Association

Cramer's V Statistic

- -1 to 1 for 2 by 2 tables
- O to 1 for larger tables
- Values further away from 0 indicate the presence of a relatively strong association



# Adding Chi-Square to our FREQ Output Strength of Association – Cramer's V

## Statistics for Table of Gender by TargetBuy

Statistic	DF	Value	Prob
Chi-Square	1	621.5507	<.0001
Likelihood Ratio Chi-Square	1	662.3524	<.0001
Continuity Adj. Chi-Square	1	620.6729	<.0001
Mantel-Haenszel Chi-Square	1	621.5161	<.0001
Phi Coefficient		-0.1860	
Contingency Coefficient		0.1829	
Cramer's V		-0.1860	





# Fisher's Exact Test

Copyright © SAS Institute Inc. All rights reserved.


## When to use Chi-Square

- When more than 20% of cells have expected counts less than five
- In this case use Fisher's Exact Test



## **Example for Fisher's Exact Test**

Row number	Product	Pu
1	A	yes
2	A	yes
3	A	yes
4	A	no
5	в	yes
6	в	no
7	в	no
8	в	no
9	в	no

chased
;
;
;
;



## Fisher's Exact Test

- Useful when sample sizes are small (less than 20-25 total)
- 2x2 tables
- Calculates probabilities by considering every possible table where the marginal (row and column) totals remain fixed)
- Large datasets may require a prohibitive amount of time and memory for computing exact p-value.



## Fisher's Exact Test

- Null Hypothesis: No Association Alternative Hypothesis:
  - Two-Tailed
  - Left-tailed
  - Right-tailed





### **Fisher's Exact Test** SAS Code

PROC FREQ DATA = mydata.products ORDER=INTERNAL; TABLES Product \* Purchased / NOROW NOPERCENT CELLCHI2 EXPECTED NOCUM FISHER SCORES=TABLE ALPHA=0.05;

### RUN;





## **Fisher's Exact Test** SAS Enterprise Guide

Tasks		Program	Tools	
	Da	ata	•	
	De	escribe	×.	T
	Gr	aph	×	Σ
	A	AVOVA	×.	Σ
	Re	gression	×	
	М	ultivariate	•	
	Su	irvival Analys	is	
	Ca	apability	•	
	Co	ontrol Charts	×	
lín.	Pa	reto Chart		
	Ti	me Series	۱.	
	Da	ata Mining	•	
	OI	AP	►	
	Та	sk Template	s 🕨	

	List Data
	Summary Statistics Wizard
	Summary Statistics
	Summary Tables Wizard
	Summary Tables
	List Report Wizard
I	Characterize Data
I	Distribution Analysis
	One-Way Frequencies
	Table Analysis

Data Tables Cell Statistics Table Statistics Association Agreement Ordered Differences Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties

Table Statistics > Association
Tests of association
Chi-square tests (Including Pearson, likelihood ratio and Mantel-Haenszel chi-square tests and Fisher's exact test for 2x2 tables)
Exact p-values
Fisher's exact test for r x c tables



## **Fisher's Exact Test**

Frequency	Table o	f Product	by Purcha	ased	Statistic		Value	Prob
Expected Cell Chi-Square		P	Purchased Chi-Square		1	2.7225	0.0989	
Col Pct	Product	no	yes	Total	Likelihood Ratio Chi-Square	1	2.8626	0.0907
	Α	1	3	4	Continuity Adj. Chi-Square	1	0.9506	0.3296
	2.2222       1.7778         0.6722       0.8403         20.00       75.00         B       4       1		Mantel-Haenszel Chi-Square	1	2.4200	0.1198		
		20.00	75.00		Phi Coefficient		-0.5500	
		4	1	5	Contingency Coefficient		0.4819	
		2.7778 0.5378	2.2222 0.6722		Cramer's V		-0.5500	
		80.00	25.00		WARNING: 100% of the cells h	expected co	counts less	
	Total	5	4	9	than 5. Chi-Square may	not	be a valid te	ST.





## Fisher's Exact Test

**Fisher's Exact Test** 

Cell (1,1) Frequency (F)

Left-sided Pr <= F

Right-sided Pr >= F

**Table Probability (P)** 

Two-sided Pr <= P







## Ordinal Variables Test

Copyright © SAS Institute Inc. All rights reserved.



## Mantel Haenszel Chi-Square Test

- Good with Ordinal association
  - Means as one variable increases the other variable tends to increase or decrease
- Need to have one variable with more than 2 levels
- Does not measure strength of association

### Null Hypothesis:

There is no ordinal association between the row and column variables Alternative Hypothesis:

There is an ordinal association between the row and column variables.



## Mantel Haenszel Chi-Square Test SAS Code

PROC FREQ DATA = mydata.organics2 ORDER=INTERNAL; TABLES LoyaltyStatus \* TargetBuy / NOCOL NOPERCENT CELLCHI2 EXPECTED NOCUM CHISQ SCORES=TABLE ALPHA=0.05;

### RUN;

Base

Loyalty Status							
Loyalty Status	Frequency	Percent					
1. Tin	6487	29.19					
2. Silver	8572	38.57					
3. Gold	6324	28.46					
4. Platinum	840	3.78					



## Mantel Haenszel Chi-Square Test SAS Enterprise Guide

Tas	ks	Program	Tools		
	Data 🕨				
	De	escribe	Þ		List Data
	Gr	raph	•	Σ	Summary Statist
	A	AVOVA	Þ	Σ	Summary Statist
	Re	egression	•		Summary Table
	М	ultivariate	•		Summary Table
	Su	Survival Analysi	is 🕨		List Report Wiza
	Ca	apability	•	8000 8000	Characterize Dat
	Control C		۰.		Distribution Ana
h	Pa	areto Chart			One-Way Frequ
	Ti	me Series	Þ		Table Analysis
	Da	ata Mining	Þ		
	0	LAP	۲		
	Ta	ask Template	5 🕨		

Copyright © SAS Institute Inc. All rights reserved.





## Mantel Haenszel Chi-Square Test SAS Enterprise Guide

Data Tables	Data		Data Tables	Tables
Cell Statistics Table Statistics Association Agreement Ordered Differences Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties	Data source: SASApp:MYDATA.ORGANICS2 Task filter: None	Edit	Cell Statistics Table Statistics Association Agreement	Variables permitted in table: Preview:
	Variables to assign: Name M		Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties	Tables to be generated: LoyaltyStatus by TargetBuy <select a="" begin="" defining="" new="" table="" to=""></select>
	The selection pane enables you to choose different sets of options for the task.	Data Tables	Cell Statistics	The selection pane enables you to choose different sets of options for the task.
Preview code	Run 💌 Save Cancel	Cell Statistics Table Statistics Association Agreement	Available statistics	Run 🔻 Save Cancel Help
		Ordered Differences Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties	<ul> <li>Column percentages</li> <li>Cell frequencies</li> <li>Cell percentages</li> <li>Missing value frequencies</li> <li>Cell contribution to Pearson chi-square</li> <li>Cell frequency deviation from expected</li> <li>Expected cell frequency</li> </ul>	
			<ul> <li>Percentage of total frequency</li> <li>Include percentages in the data set</li> </ul>	<b>S</b> S2



## Mantel Haenszel Chi-Square Test

### The FREQ Procedure

Frequency	Table of Loyalty Status by TargetBuy					
Expected Cell Chi-Square	LovaltyStatus/Lovalty	TargetBuy(Organics Purchase Indicator)				
Row Pct	Status)	0	1	Total		
	1. Tin	4458 4880.1 36.503 68.72	2029 1606.9 110.86 31.28	6487		
	2. Silver	6460 6448.6 0.0202 75.36	2112 2123.4 0.0615 24.64	8572		
	3. Gold	5088 4757.4 22.968 80.46	1236 1566.6 69.751 19.54	6324		
	4. Platinum	712 631.92 10.149 84.76	128 208.08 30.82 15.24	840		
	Total	16718	5505	22223		





## Mantel Haenszel Chi-Square Test

### Statistics for Table of LoyaltyStatus by TargetBuy

Statistic	DF	Value	Prob
Chi-Square	3	281.1281	<.0001
Likelihood Ratio Chi-Square	3	283 1617	<.0001
Mantel-Haenszel Chi-Square	1	278.2499	<.0001
Phi Coefficient		0.1125	
Contingency Coefficient		0.1118	
Cramer's V		0.1125	





## Strength of Association

Spearman Correlation Statistic

- Range -1 to 1
- Values close to 1, relatively high degree of positive correlation
- Values close to -1, relatively high degree of negative correlation
- Only valid if both variables are ordinally scaled and in logical order



# Mantel Haenszel Chi-Square Test with Spearman Correlation - SAS Enterprise Guide

Table Analysis Ordinal	Chi-Square for SASApp:MYDATA.ORGAN	IICS2	22	Table Analysis Ordinal	al Chi-Square for SASApp:MYDATA.ORGANICS2
Data Tables Cell Statistics Table Statistics Association	Data Data source: SASApp:MYDATA.ORGAI Task filter: None	VICS2	Edit	Data Tables Cell Statistics Table Statistics Association Agreement	Variables permitted in table:     Preview:       TargetBuy
Table Statistics Association Agreement Ordered Differences Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties	Variables to assign:	ask roles: Frequency count (Limit: 1) Group analysis by Table variables LoyaltyStatus TargetBuy different sets of options for the task.		Ordered Differences Trend Test Computation Options Results Cell Stat Results Table Stat Results Titles Properties	a       Image: Section pane enables you to choose different sets of options for the task.         Run       Save       Cancel       Help
			ei nep		
Data Tables Cell Statistic	Cell Statistics		Data Tables	Table Statistics > Association	1
Table Statis Associa	tics Available statistics	entages	Table Statistics	Tests of association	Measures of association
Agreem Ordered Trend T Comput Results Cell Stat	ent       Row percentages         Differences       Column percentages         est       Cell frequencies         ation Options       Cell percentages         t Results       Missing value frequencies	ow percentages olumn percentages ell frequencies ell percentages issing value frequencies		Chi-square tests Chi-square tests (Including Pearson, likelihood chi-square tests and Fisher's e	I ratio and Mantel-Haenszel exact test for 2x2 tables)
Table S Titles Properties	able Stat Results       Cell contribution to Pearson chi-square         Interview       Cell frequency deviation from expected         Interview       Expected cell frequency		Results Cell Stat Results	Exact p-values	Test that the measure equals zero Risk differences for 2 x 2 tables
	Percentage of total freq	uency the data set	Titles Properties	Hisner's exact test for r'x c table	Relative risk for 2 x 2 tables



# Mantel Haenszel Chi-Square Test with Spearman Correlation

Statistic	Value	ASE	95% Confidence Limits		
Gamma	-0.2072	0.0121	-0.2309	-0.1835	
Kendall's Tau-b	-0.1047	0.0062	-0.1168	-0.0926	
Stuart's Tau-c	-0.1057	0.0063	-0.1180	-0.0934	
Somers' D C R	-0.0773	0.0046	-0.0863	-0.0683	
Somers' D R C	-0.1418	0.0083	-0.1581	-0.1254	
Pearson Correlation	-0.1119	0.0065	-0.1247	-0.0991	
Spearman Correlation	-0.1121	0.0066	-0.1250	-0.0991	
Lambda Asymmetric C R	0.0000	0.0000	0.0000	0.0000	
Lambda Asymmetric R C	0.0000	0.0000	0.0000	0.0000	
Lambda Symmetric	0.0000	0.0000	0.0000	0.0000	
Uncertainty Coefficient C R	0.0114	0.0013	0.0088	0.0140	
Uncertainty Coefficient R C	0.0053	0.0006	0.0041	0.0065	
Uncertainty Coefficient Symmetric	0.0072	0.0008	0.0055	0.0089	





## Continuous Variable Test Correlations

Copyright © SAS Institute Inc. All rights reserved.



## Correlations

- Describes the relationship between 2 continuous variables
- Important to view data in scatter plot before you start analysis



## entinuous variables efore you start analysis



## Correlations Scatter Plot

- 2 dimensional graphs produced by plotting one variable against another
- Useful to
  - Explore the relationship between 2 variables
  - Locate outlying or unusual values
  - Identify possible trends
  - Identify a basic relationship of Y and X values
  - Communicate data analysis results















Copyright © SAS Institute Inc. All rights reserved.





## Correlations Scatter Plot Code

### **PROC GPLOT** DATA=mydata.organics2; PLOT PromSpend \* DemAge; RUN;

\*Proc GPLOT is in SAS/Graph, you can substitute PROC PLOT

Copyright © SAS Institute Inc. All rights reserved.





Tas	ks	Program	Tools		
	Da	ata	۱.		
	De	escribe	×.		
	Gr	aph	Þ		Bar Chart Wizard
	A	AVOVA	•		Bar Chart
	Re	gression	۱.	0	Pie Chart Wizard
	М	ultivariate	×	0	Pie Chart
	Su	ırvival Analysi	is 🕨	~	Line Plot Wizard
	Ca	apability	•	~	Line Plot
	Co	ontrol Charts	•		Scatter Plot
m	Pareto Chart			Scatter Plot Matrix	
	Ti	me Series	×		Area Plot
	Da	ata Mining	×	<b>M</b>	Bar-Line Chart
	OI	AP	•	80	Bubble Plot
	Та	sk Templates	•	•	Donut Chart
				1	Contour Plot
				₽₽₽	Box Plot
				*	Radar Chart
				5	Surface Plot
				H	Tile Chart
				-	Map Chart
				1	Open ODS Graphics Designer
				<u>a</u>	Show ODS Statistical Graph

\*Proc GPLOT is in SAS/Graph

### Graph



Scatter Plot Data	Data					
ppearance Plots Interpolations Axes General	Data source: C:\Public\DataMining\Organics\organics.sas7bdat Task filter: TargetBuy = 1					
Horizontal Axis	Columns to assign:		Task roles:			
Axis Major Ticks Minor Ticks Reference Lines Vertical Axis Axis Major Ticks Reference Lines Vertical Right Axis Axis Major Ticks Minor Ticks Reference Lines Legend Chart Area Titles Properties	Name ID DemAffl DemAge DemCluster DemClusterGroup DemGender DemReg DemTVReg PromClass PromSpend PromSpend DemTime DemTime TargetBuy TargetAmt		Horizontal (Limit: 1) Vertical (Limit: 1) PromSpend Vertical (Right) (Limit: 1) Group charts by	Summarize for each distinct horizontal value		
Preview code			Run 💌 Save	e Cancel Help		

### \*Proc GPLOT is in SAS/Graph

### Graph



## Correlations Scatter Plot



Image Source: eMathZone.com





## Correlations Pearson Correlation Coefficient

- Between -1 and 1
- Closer to either extreme, high degree of linear association between the two variables
- Close to O, no linear association
- Greater than O, positive linear association
- Less than O, negative linear association





## Correlations Hypothesis Test

- The parameter representing correlation is p
- p is estimated by the sample statistic r
- Null Hypothesis:
  - There is no association between the 2 variables,  $\rho = 0$
- Alternative Hypothesis:
  - There is an association between the 2 variables,  $\rho \neq 0$
  - Rejecting  $H_0$  indicates only great confidence that p is not exactly O
  - A p-value does not measure the magnitude of the association
  - Sample size affects the p-value





## Correlations SAS Code

### PROC CORR DATA=mydata.organics PEARSON; VAR DemAge PromSpend; RUN;

Copyright © SAS Institute Inc. All rights reserved.



S.Sas

Tas	ks P	rogram	Tools		
	Data		•		
	Desci	ribe	•		
	Grap	h	•		
	ANO	VA	Þ		
	Regre	ession	•		
	Multi	variate	×	Z	Correlations
	Survi	val Analys	is 🕨	1	Canonical Correlation
	Capa	bility	Þ	X	Principal Components.
	Cont	rol Charts	•	<u>L</u>	Factor Analysis
lín.	Paret	o Chart		35	Cluster Analysis
	Time	Time Series		×	Discriminant Analysis
	Data	Mining	•		
	OLAF	)	۲		
	Task	Template	s 🕨		







Data Options Results Output Data Titles Properties	Data							
	Data source: C:\Public\DataMining\Organics\organics.sas7bdat Task filter: None Edit							
	Variables to assign:	Task ro	les:					
	Name         ID         ID         DemAffl         DemAge         DemCluster         DemClusterGroup         DemGender         DemReg         DemTVReg         PromClass         PromSpend         Image: TargetBuy         TargetAmt	An O O O O O O O O O O O O O	alysis variables DemAge PromSpend melate with oup analysis by equency count (Limit: 1) rtial variables lative weight (Limit: 1)					
	The selection pane enables you to choo	ose different sets of option	is for the task.					





s	Options	
Data Correlation types Pearson Hoeffding Kendall Spearman		Pearson correlation options Cronbach's coefficient alpha Covariances Sums of squares and cross products Corrected sums of squares and cross products Suppress Pearson correlations from results
	<ul> <li>Fisher options</li> <li>Specify the value of alpha:</li> <li>0.05</li> <li>Specify the value rho0 in the null hypothesis:</li> <li>0</li> </ul>	Type of confidence limits: Two sided Use bias adjustment for constructing confidence levels
	Divisor for variance: Degrees of freedom	Omit rows with missing values for variables being correlated
view code		Run 🔻 Save Cancel Help





Correlations2 for	C:\Public\DataMining\Organics\organics.sas7bdat	t	×
Options Results Output Data Titles Properties	Plots	<ul> <li>Results to display</li> <li>Show statistics for each variable</li> <li>Show significance probabilities associated with correlations</li> <li>Show correlations in decreasing order of magnitude</li> <li>Show n correlations per row variable:</li> </ul>	2
	Summary of correlations to calculate Number of variables to correlate: Total correlations to be calculated:	2 1	
Preview code		Run 💌 Save Cancel	- Help





## Correlations Hypothesis Test

### **Correlation Analysis**

### The CORR Procedure

2 Variables:

PromSpend DemAge

Simple Statistics								
Variable N Mean Std Dev Sum Minimum Maximum Label							Label	
PromSpend	22223	4421	7559	98238772	0.01000	296314	Total Spend	
DemAge	20715	53.79715	13.20605	1114408	18.00000	79.00000	Age	







## Correlations SAS code with Scatter Plot






lata Iptions	Data					
Results Output Data Titles Properties	Data source: C:\Public\DataMining\Organics\organics.sas7bdat Task filter: None					
	Variables to assign:	Task rol	es:			
	Name         ID         DemAffl         DemAge         DemCluster         DemClusterGroup         DemGender         DemReg         DemTVReg         PromClass         PromSpend         PromTime         TargetBuy         TargetAmt	Ana Ana O Cor O Fre O Rel Cor Cor Cor Cor Cor Cor Cor Cor	alysis variables DemAge PromSpend relate with oup analysis by quency count (Limit: 1) tial variables lative weight (Limit: 1)			
	The selection pane enables you to choose	different sets of option	s for the task.			





Code Prev	view for Task
Insert	Code
22	001T:
23	TITLE:
24	TITLE1 "Correlation Analysis";
25	FOOTNOTE;
26	FOOTNOTE1 "Generated by the SAS System (& SASSE
27	PROC CORR DATA=WORK.SORTTempTableSorted
28	PLOTS=(SCATTER MATRIX)
29	PEARSON
30	VARDEF=DF
31	;
32	VAR PromSpend DemAge;
33	RUN;
34	
35	/*
36	End of task code.
37	
38	RUN; QUIT;
39	<pre>% eg_conditional_dropds (WORK.SORTTempTableSorted)</pre>
40	TITLE; FOOTNOTE;
41	ODS GRAPHICS OFF;
42	
-	III







add user code or change existing user (	code.
adarah an stada an	A
COUDIE-CIICK TO INSERT CODE>	Table Coded
PROCIORR DATA=WORK.SURT	emplableSorted
PLOTS=(SCATTER MATRIX)	
PEARSON	
VARDEF=DF	
Adouble-click to insert code>	

Enter User Code	
PLOTS(MAXPOINTS=30000)	ОК
	Cancel

Insert	Code
22	QUIT;
23	TITLE;
24	TITLE1 "Cor
25	FOOTNOTE;
26	FOOTNOTE1 "
27	PROC CORR D
28	PLOTS= (
29	PEARSON
30	VARDEF=
31	
32	/* Start of
38	PLOTS (MAXPO
34	/* End of c
35	;
36	VAR Dem
37	RUN;
38	
39	/*
40	End of t
41	
42	RUN; QUIT;
43	% eq condit





# Correlations SAS code with Scatter Plot

## The CORR Procedure







# Correlations Scatter Plot Matrix

Scatter Plot Matrix



## Graph



## Correlations Scatter Plot Matrix Code

PROC SGSCATTER DATA=mydata.ORGANICS; TITLE "Scatter Plot Matrix"; MATRIX DemAffl DemAge PromSpend PromTime TargetBuy TargetAmt/ START=TOPLEFT ELLIPSE=(ALPHA=0.05 TYPE=PREDICTED) NOLEGEND;

RUN;

\*Proc SGSCATTER is in SAS/Graph.

Copyright © SAS Institute Inc. All rights reserved





Tas	ks	Program	Tools		
	Da	ata	•		
	De	escribe	•		
	Gr	aph	×		Bar Chart Wizard
	A	AVON	•		Bar Chart
	Re	Regression		0	Pie Chart Wizard
6	М	ultivariate	۲	0	Pie Chart
	Su	ırvival Analysis	•	~	Line Plot Wizard
	Capat Contr Pareto	apability	۲	~	Line Plot
		ontrol Charts	×	1	Scatter Plot
		reto Chart		12	Scatter Plot Matrix
	Ti	me Series	•		Area Plot
	Da	ata Mining	•	1	Bar-Line Chart
	OI	LAP	•	80	Bubble Plot
	Task Templates	sk Templates		•	Donut Chart
			•	3	Contour Plot
				₽₽₽	Box Plot
				$(\bigstar)$	Radar Chart
				₩	Surface Plot
					Tile Chart
				1	Map Chart
				<b>a</b>	Open ODS Graphics Designer
				<u></u>	Show ODS Statistical Graph

\* Proc SGSCATTER is in SAS/Graph

## Graph



Label: Scatter Plot Matrix			
Data source: C:\Public\D;	ataMining\Organics\organ		
Task filter: None 🚽			
Matrix variables: DemAffl,	DemAge, PromSpend, PromTi	me, TargetBuy	
Group variable: No group		-	
Chart title: "Scatter Plot M	atrix" 🖛		
Diagonal: Show variable n	ames, Down to the right 🕶		
Ellipse: Predicted, 95% co	nfidence 💌		
Legend: "Legend", Bottom	, No border 🕶		
Chart area: Default size, N	lo data tips 🔫		
Chart footnote: "Generate	d by the SAS System" 🝷		
Preview code	Bun _ ]	Save	0

## \*Proc SGSCATTER is in SAS/Graph

## Graph

L	x	Ŋ
Edit		
Help		

Sas

## Correlations SAS code with Scatter Plot Matrix



/\* Start of custom user code. \*/ PLOTS(MAXPOINTS=100000) End of custom user code. \*/

VAR PromTime DemAge TargetAmt PromSpend DemAffl;

RUN;





# Correlations SAS code with Scatter Plot Matrix







# Correlations VS Causation







# Correlation VS Causation







# Categorical and Continuous Variables

Copyright © SAS Institute Inc. All rights reserved



# t-test

# <u>Three types</u>1. One Sample2. Two Sample3. Paired





## t-test One Sample

Parametric test to compare sample mean with known value

- Null Hypothesis:  $H_0: \mu$  = hypothesized value
- Alternative Hypothesis:  $H_a: \mu \neq$  hypothesized value • For our Example  $\mu = 47$

Assumptions

- The data consist of independently chosen random samples
- The sample size is large



## t-test One Sample

PROC TTEST DATA = mydata.organics PLOTS(ONLY)=SUMMARY ALPHA=0.05HO = 47CI = EQUAL;VAR DemAge; BY TargetBuy; RUN;

## Can also calculated in PROC UNIVARIATE







## t-test One Sample - SAS Enterprise Guide

Tas	ks Program To	ools		
	Data	۲		
	Describe	•		
	Graph	۲		
	ANOVA	×	FI	t Test
	Regression	•	ň	One-Way ANOVA
	Multivariate	×	△	Nonparametric One-W
	Survival Analysis	×	<b>*</b>	Linear Models
	Capability	Þ	Z	Mixed Models
	Control Charts	•		
lín	Pareto Chart			
	Time Series	۲		
	Data Mining	۲		
	OLAP	•		
	Task Templates	Þ		







## t-test One sample - SAS Enterprise Guide

pple T-test Age = 47 for SASApp:MYDATA.ORGANICS2			t Test type Data	Data	
t Test type     Choose t Test type:     Two Sample     Image: Test type:     Image: Test type: Test type:     Image: Test type: T	App:MYDATA.ORGANICS2		t Test type Data Analysis Plots Titles Properties	Data         Data source:       SASApp:MYDATA.ORGANICS2         Task filter:       None         Variables to assign:       Task roles:         Name       Analysis variables         ID       DemAge         Image: DemAge       Image: DemAge         Image: DemCluster       Image: DemReg         Image: DemTVReg       Image: DemTVReg         Image: PromSpend       Image: DemAnt         Image: DemAnt       Image: DemAnt         Image: DemAn	Edit
Analysis Plots Titles Properties Ho = Standard deviation © Equal tailed © UMPU (Unifo Confidence leve Specify a new null P	t value for the numbypothesis: 47 a confidence untervals mly most powerful unbiased test) 1: 95% • nypothesis value. The default is 0.	ve Cancel Help	Preview code	t Test type       Data         Analysis       Plots         Titles       Types         Properties       Image: Summary plot         Image: Histogram       Image: Sumplet         Image: Confidence interval plot       Image: Summary plot         Image: Normal quantile-quantile (Q-Q) plot       Image: Summary plot         Image: Normal quantile-quantile (Q-Q) plot       Image: Summary plot	
Preview code	Run 💌 Sa	re Cancel Help			





## t-test One Sample

## **Organics Purchase Indicator=1**

## t Test

The TTEST Procedure

Variable: DemAge (Age)

N	Mean	Std Dev	Std Err	Minimum	Maximum
5100	46.8063	13.9384	0.1952	20.0000	79.0000

Mean	95% CL Mean		Std Dev	95% CL	Std Dev
46.8063	46.4236	47.1889	13.9384	13.6731	14.2143

DF	t Value Pr >  t	
5099	-0.99	0.3210







## t-test One Sample

## Organics Purchase Indicator=0

## t Test

## The TTEST Procedure

Variable: DemAge (Age)

N	Mean	Std Dev	Std Err	Minimum	Maximum
15615	56.0804	12.1137	0.0969	18.0000	79.0000

Mean	95% C	95% CL Mean		95% CL	Std Dev
56.0804	55.8904	56.2705	12.1137	11.9808	12.2496

DF	t Value	Pr >  t
15614	93.67	<.0001







## t-test Two Sample

Parametric test to compare two independent samples

- Null Hypothesis:  $H_0: \mu_1 = \mu_2$
- Alternative Hypothesis:  $H_a: \mu_1 \neq \mu_2$
- For our Example
- $\mu_1$  is amount Females spent during the promotion
  - $\mu_2$  is amount Males spent during the promotion

## Assumptions

- Independent Observations
- Normally distributed responses for each group
- Equal variances for each group

Copyright © SAS Institute Inc. All rights reserved.



ng the promotion g the promotion

## t-test Two Sample

PROC TTEST DATA = mydata.organics PLOTS(ONLY)=SUMMARY ALPHA=0.05HO = OCI = EQUAL;CLASS Gender; VAR PromSpend; RUN;







## t-test Two Sample - SAS Enterprise Guide

Tas	sks	Program	Tools		
	Da	ata	•		
	De	escribe	•		
	G	raph	•		
	A	NOVA	×.	Ы	t Test
	Re	egression	•	<i>й</i> с	One-Way ANOVA
	Μ	ultivariate	•	杰	Nonparametric One-W
	Su	urvival Analys	is 🕨	₩.	Linear Models
	Ca	apability	×	Z	Mixed Models
	C	ontrol Charts	•		
1	Pa	areto Chart			
	Ti	me Series	Þ		
	Da	ata Mining	•		
	0	LAP	Þ		
	Ta	ask Template	s 🕨		



## Way ANOVA...



## t-test Two sample - SAS Enterprise Guide

D Sample T-test Gender for SASApp:MYDATA.ORGANICS2	t Test type Data	Data	
st type a lysis s s	Analysis Plots Titles Properties	Data source: SASApp:MYDATA.ORGANICS2 Task filter: None	Edit
s perties Two Sample One Sample One Sample One Sample One Sample Trest Gender for SASApp:MYDATA.ORGANICS2		Variables to assign:       Task roles:         Name       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)         Image: Classification variable (Limit: 1)       Image: Classification variable (Limit: 1)	
Preview code     Preview code     Preview code     Properties     Null hypothesis   Decity the test value for the sull hypothesis:   Properties     Ho =     Sta fard deviation confidence intervals   Equal tailed   UMPU (Uniformity most powerful unbiased test)   Confidence level:      95%, •	Preview code	Image: Second state of the second s	
Preview code Run Save Cancel Help			





## t-test Two Sample

## The TTEST Procedure

### Variable: PromSpend (Total Spend)



Copyright © SAS Institute Inc. All rights reserved.



Dev	Std Err	Minimum	Maximum
28.8	64.6762	0.0100	239542
89.6	103.5	0.0100	296314
3.6	117.7		

5% CL Mean		Std Dev	95% CL	Std Dev
3.7	4387.3	7128.8	7040.3	7219.6
9.7	4795.4	7889.6	7748.8	8035.7
2.8	-101.3	7383.6	7308.1	7460.8
1.2	-92.8713			

es	DF	t Value	Pr >  t
	17962	-2.82	0.0048
1	10480	-2.72	0.0065

/ of Variances				
Den DF	F Value	Pr > F		
12148	1.22	<.0001		



# t-test Two Sample







## t-test Paired

Parametric test to compare repeat measures on the same subject

- Null Hypothesis:  $H_0: \mu_{post} = \mu_{pre}$ ٠
- Alternative Hypothesis:  $H_a$ :  $\mu_{post} \neq \mu_{pre}$ •
- For our Example
- $\mu_{post}$  is amount of organic items bought during promotion
- $\mu_{\text{pre}}$  is amount of organic items bought before promotion

## Assumptions

- The subjects are selected randomly
- The distribution of the sample mean differences is normal



## t-test Paired

PROC TTEST DATA = mydata.organics PLOTS(ONLY)=SUMMARY ALPHA=0.05HO = OCI = EQUAL;PAIRED TargetAmt \* PrePromAmt; RUN;





Tas	ks	Program	Tools		
	Da	ata	•		
	De	escribe	•		
	Gr	raph	•		
	A	AVOVA	•	<u>F-I</u>	t Test
	Re	egression	•	蘆	One-Way ANOVA
	Μ	ultivariate	•	盃	Nonparametric One-
	Su	urvival Analys	is 🕨	*	Linear Models
	Ca	apability	•	×	Mixed Models
	Co	ontrol Charts	•		
1	Pa	areto Chart			
	Ti	me Series	Þ		
	Da	ata Mining	×		
	O	LAP	•		
	Ta	ask Templates	; •		



-Way ANOVA...



Paired T-test Amount Bought for SASAp	pp:MYDATA.ORGANICS2	Paired 1-test 4	Amount Bought for SASApp:MYDATA.ORGANICS2
Paired T-test Amount Bought for SASAp         Test type         Data         Analysis         Plots         Titles         Properties         Image: Choose t Test type:         Image: Transmission of the test type:         Image: Choose t Test type:         Image: Transmission of test type:         Image: Test type:	wo Sample	t Test type Data Analysis Plots Titles Properties	Data         Data source:       SASApp:MYDATA.ORGANICS2         Task filter:       None         Variables to assign:       Task roles:         Name       Paired variables (Limit: 2)         ID       Paired variables (Limit: 2)         ID       Target Amt         ID       Februariables         ID       Februariables <t< th=""></t<>
Preview code	Ine Sample     T-test Gender for SASApp:MYDATA.ORGANICS2     Analysis     Null protinesis   Specify the test value for the null roothesis:   Ho = 0     Standard to wintee confidence intervals   If Equal tailed   UMPU (Unformly most powerful unbiased test)   Confidence level:   95%	Preview code	DemAge     DemCusterGroup     DemCusterGroup     DemCusterGroup     DemReg     DemR
[ <sup>2220</sup> ] Days Same	Ten Run Ten Save Cancel Help		

Copyright © SAS Institute Inc. All rights reserved.





## The TTEST Procedure

## Difference: TargetAmt - PrePromAmt

N	Mean	Std Dev	Std Err	Minimum	Maximum
22223	0.1798	0.6690	0.00449	-2.0000	3.0000

Mean	95% C	L Mean	Std Dev	95% CL	Std Dev
0.1798	0.1710	0.1886	0.6690	0.6628	0.6752

DF	t Value	Pr >  t	
22222	40.06	<.0001	









ePromAmt				
——— No ——— Ke	rmal rnel			
	ence l			
o				
	1			



# Nonparametric test

Copyright © SAS Institute Inc. All rights reserved.



# Nonparametric Analysis

Nonparametric analysis are those that rely only on the assumption that the observations are independent

A nonparametric test is appropriate when

- The data contains valid outliers
- The data is skewed
- The response variable is ordinal and not contiguous



## Nonparametric Analysis PROC NPAR1WAY

- The rank of each data point is used instead of the raw data
  - Rank from smallest to largest
  - In the event of a tie the ranks are averaged
- For 2 level variables Wilcoxon rank-sum test is used
- For more than 2 levels Kruskal-Wallis test is used



Nonparametric Analysis PROC NPAR1WAY

Null Hypothesis:  $H_0$ : all populations are identical with respect to scale, shape, and location

Alternative Hypothesis:  $H_a$ : all populations are not identical with respect to scale, shape, and location

- Only assumption is that you have independent observations
- Used with ordinal, interval and ratio measurement variables




## PROC NPAR1WAY 2 levels

PROC NPAR1WAY DATA=organics2 WILCOXON MEDIAN; VAR Diff\_Amt; CLASS Gender; RUN;





# PROC NPAR1WAY2 levels - SAS Enterprise Guide

Tas	ks Program	Tools		
	Data			
	Describe	۱.		
	Graph			
	ANOVA	×	<u>F-I</u>	t Test
	Regression	•	蘆	One-Way ANOVA
	Multivariate	•	△	Nonparametric One-W
	Survival Analys	is 🕨	*	Linear Models
	Capability	►	*	Mixed Models
	Control Charts	۱.		
lín.	Pareto Chart			
	Time Series	۰.		
	Data Mining	۲		
	OLAP	Þ		
	Task Templates	5 F		



Way ANOVA...



## PROC NPAR1WAY 2 level - SAS Enterprise Guide

Data Analysis Exact p-values Results Titles Properties	Data Data source: SASApp:MYDATA.ORGANICS2 Task filter: None	Edit	
	Variables to assign: Name ID DemAffl DemAge DemCluster DemCluster DemCluster DemCluster DemReg DemReg DemTVReg PromSpend PromTime PrepromAmt Task roles: Task roles: Dependent variables Dependent variab	Nonparametri	tric One-Way ANOVA GENDER for SASApp:MYDATA.ORGANICS2
Preview code	Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of the selection pane enables you to choose different sets of options for the task.         Image: Contract of task of ta	Data Analysis Exact p-values Results Titles Properties	Analysis s Test scores Wilcoxon Calculate empirical distribution function statistics (EDF)
			<ul> <li>Include missing values as a class level</li> <li>Savage</li> <li>Van der Waerden</li> <li>Ansari-Bradley</li> <li>Klotz</li> <li>Mood</li> <li>Siegel-Tukey</li> <li>Raw data</li> <li>NOTE: The test scores must be checked in order to enable the exact p-values on the "Exact p-values" page and to enable the statistics on the "Results" page.</li> </ul>
		Preview code	ade Run V Save Cancel Help





## PROC NPAR1WAY 2 levels

### The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Diff_Amt Classified by Variable Gender					
Gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Female	12149	114836825	109128393	271388.619	9452.36847
Male	5815	46524805.5	52233237.5	271388.619	8000.82640
Average scores were used for ties.					



Wilcoxon Two-Sample Test				
Statistic	46524805.5000			
Normal Approximation				
Z	-21.0342			
One-Sided Pr < Z	<.0001			
Two-Sided Pr >  Z	<.0001			
t Approximation				
One-Sided Pr < Z	<.0001			
Two-Sided Pr >  Z	<.0001			
Z includes a continuity correction of 0.5.				



## PROC NPAR1WAY 2 levels







## PROC NPAR1WAY > 2 levels

proc sort data=mydata.organics2 out=organics2; by loyaltyStatus;

PROC NPAR1WAY DATA=organics2 WILCOXON MEDIAN; VAR Diff Amt; CLASS LoyaltyStatus;

RUN;

Copyright © SAS Institute Inc. All rights reserved.





# PROC NPAR1WAY > 2 levels - SAS Enterprise Guide

Tas	ks Prog	ram	Tools		
	Data		۱.		
	Describe		×.		
	Graph		•		
	ANOVA		×	H	t Test
	Regressio	on	×	<i>й</i> ч	One-Way ANOVA
	Multivari	iate	×	△	Nonparametric One
5	Survival	Analysi	s ⊧	*	Linear Models
	Capabilit	y	•	Ż	Mixed Models
	Control (	Charts	•		
	Pareto C	hart			
	Time Ser	ies	×		
	Data Mir	ning	•		
	OLAP		F		
	Task Ten	nplates	F		



### . e-Way ANOVA...



# PROC NPAR1WAY > 2 level - SAS Enterprise Guide

🖄 Nonparametric (	One-Way ANOVA PROMCLASS for SASApp:MYDATA.ORGANICS2	
Data Analysis	Data	
Exact p-values Results Titles Properties	Data source: SASApp:MYDATA.ORGANICS2 Task filter: None	Edit
Preview code	Variables to assign:       Task roles:         Name       ID         ID       Dependent variables         ID       Diff_Amt         ID       DemAffl         ID       Optimum         ID       DemAge         ID       DemAge         ID       DemAge         ID       DemCluster         ID       DemClusterGroup         ID       Image Counce         Image Counce       Image Counce         Image Counce <td>Data Analysis       Analysis         Exact p-values Results       Test scores         Vilcoxon       Vilcoxon         Properties       Wedan         Include missing values as a class level         Savage         Van der Waerden         Ansan-Bradley         Klotz         Mood         Siegel-Tukey         Raw data         NOTE: The test scores must be checked in order to enable the exact p-values on the "Exact p-values" page and to enable the statistics on the "Results" page.</td>	Data Analysis       Analysis         Exact p-values Results       Test scores         Vilcoxon       Vilcoxon         Properties       Wedan         Include missing values as a class level         Savage         Van der Waerden         Ansan-Bradley         Klotz         Mood         Siegel-Tukey         Raw data         NOTE: The test scores must be checked in order to enable the exact p-values on the "Exact p-values" page and to enable the statistics on the "Results" page.
		Preview code    Run  Save  Cancel  Help





.....

## PROC NPAR1WAY > 2 Levels

### The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Diff_Amt Classified by Variable LoyaltyStatus					
Loyalty Status	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1. Tin	6487	76643119.0	72083544.0	352231.776	11814.8788
2. Silver	8572	95164849.5	95252064.0	377122.743	11101.8257
3. Gold	6324	66574842.5	70272288.0	349574.903	10527.3312
4. Platinum	840	8559165.0	9334080.0	147752.001	10189.4821
Average scores were used for ties.					

Kruskal-Wallis Test			
Chi-Square 225.1912			
DF	3		
Pr > Chi-Square	<.0001		





## PROC NPAR1WAY > 2 Levels







## Resources

### Where to learn more

Copyright © SAS Institute Inc. All rights reserved.



## Resources Courses and Tutorials

### Public SAS Courses

- Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression FREE
- SAS Enterprise Guide: ANOVA, Regression, and Logistic Regression •

### Online Tutorials

- **Basic Statistics Tutorials** 
  - t-tests
  - Tests of Association
    - Pearson Chi-Square
    - Mantel-Haenszel Chi-Square
  - Nonparametric Analysis



## Resources **Documentation and Videos**

- <u>SAS/STAT Support Learn, Documentation, Support</u>
- What's New Documentation
  - <u>http://support.sas.com/documentation/whatsnew/</u>
- STAT, IML, OR, ETS Papers
- Statistical Procedures SAS Community
- Frequently Asked-for Statistics
- Videos
  - Youtube.com
    - <u>http://www.youtube.com/playlist?list=PL0B05D53A5E101AA6</u>
  - Video portal to the STAT and OR focus area.
    - http://support.sas.com/rnd/app/video/index.html



# Online. Everyday.

## SAS Online Community

https://communities.sas.com/

Copyright © SAS Institute Inc. All rights reserved.

"I always learn something new when I post in this forum. Just what I needed ... "





## Thank you for your time and attention! Using SAS/STAT®: A Gentle Introduction to Some Frequently Used Tools



Connect with me: LinkedIn: <u>https://www.linkedin.com/in/melodierush</u> Twitter: @Melodie\_Rush

Copyright © SAS Institute Inc. All rights reserved.

