Using SAS/STAT to evaluate the impact of business interventions: A gentle introduction to some frequently used tools

Melodie Rush
Senior Systems Engineer

CUSTOMER LOYALTY TEAM ▪ Support You Can Count On

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 SAS Base and SAS/Stat

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	0000000140	76		3. Gold	16000	4	0	0	0	0
2	0000000620	49		3. Gold	6000	5	0	0	0	0
3	0000000868	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	0000002313	68	Female	1. Tin	0.01	8	0	0	0	0
6	0000002771	72		4. Platinum	20759.81	3	0	0	0	0
7	0000003131	74	Female	1. Tin	0.01	8	0	0	0	0
8	0000003328	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	0000005886	43	Female	3. Gold	6000	1	1	1	0	1

Observations 22,223
Variables 14

Data

Variable	Description
ID	Unique Customer ID
DEMAFFL	Affluence grade on a scale from 1 to 30
DEMAGE	Age, in years
DEMCLUSTER	Type of Residential Neighborhood – 55 levels
DEMCLSUTERGROUP	Neighborhood group - 7 levels
GENDER	M=Male, F=Female
DEMGREGION	Demographic Region
LOYALTYSTATUS	Loyalty status: tin, silver, gold, or platinum
PROMSPEND	Total amount spent
PROMTIME	Time as loyalty card member
TARGETBUY	Organics purchased? 1=Yes, 0=No
TARGETAMT	Number of organic products purchased during promotion
PREPROMAMT	Number of organic products purchased before promotion
DIFF_AMT	TARGETAMT - PREPROMAMT

First Things First…Categorical Variables

The FREQ Procedure

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

Gender		
DemGender	Frequency	Percent
	2512	.
F	12149	61.64
M	5815	29.50
U	1747	8.86

First Things First…Categorical Variables

The FREQ Procedure

Organics Purchase Indicator		
TargetBuy	Frequency	Percent
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		
Loyalty Status	Frequency	Percent
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

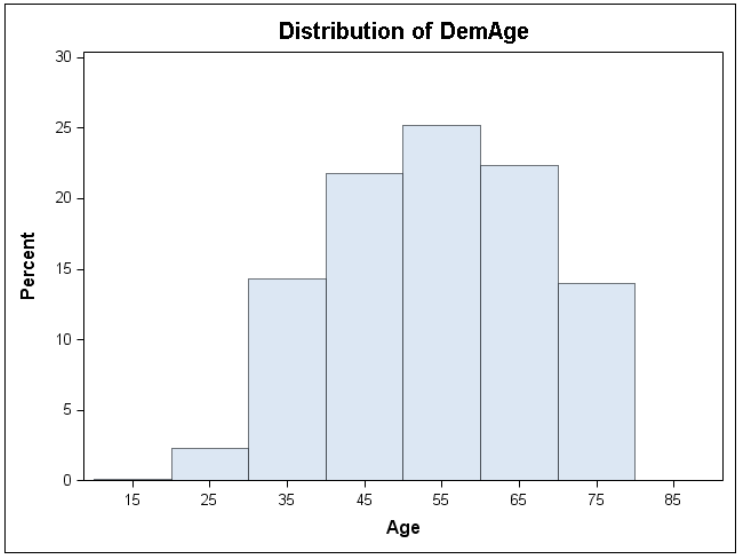
Frequency Missing = 4259

First Things First…Continuous Variables

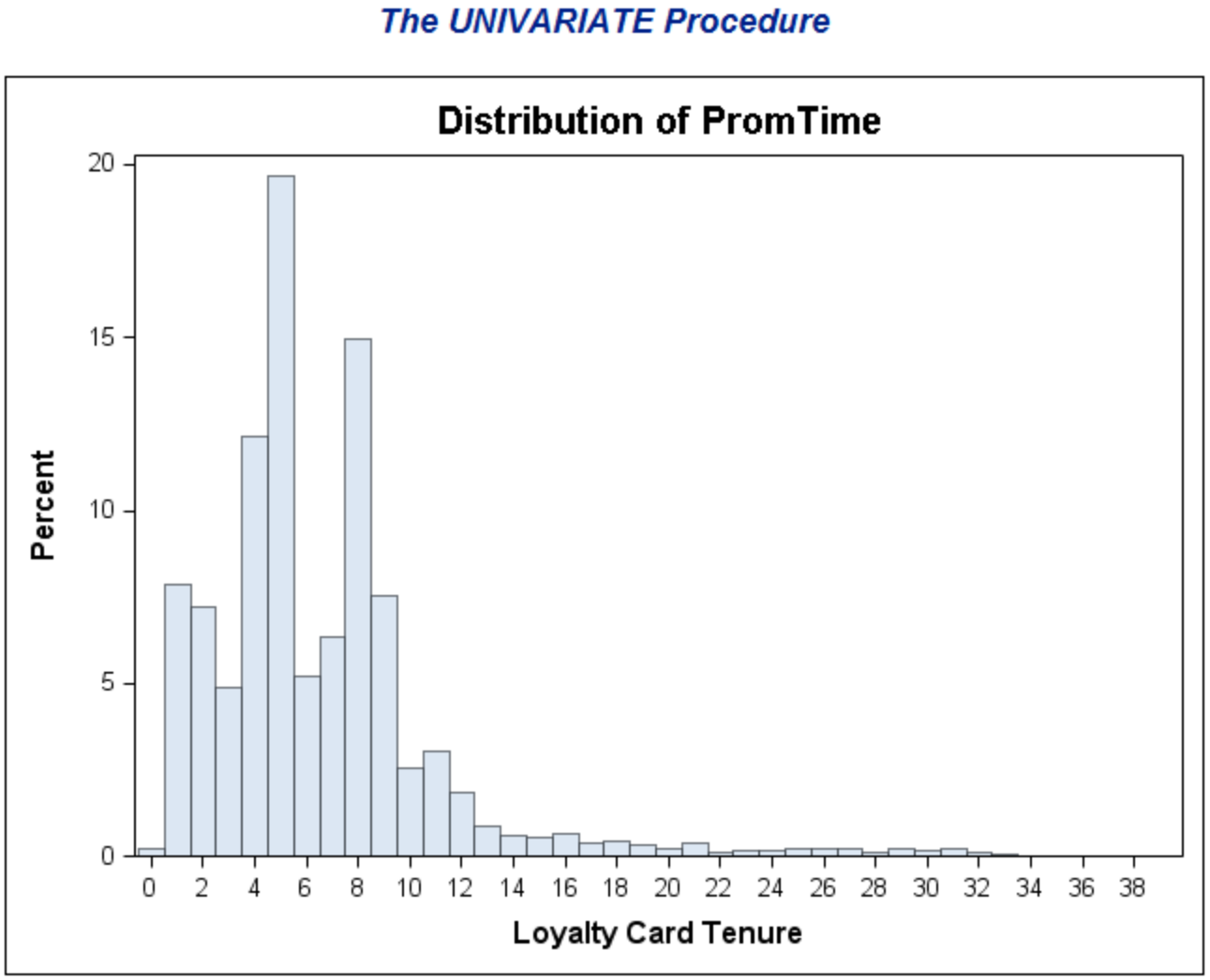
The MEANS Procedure

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	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

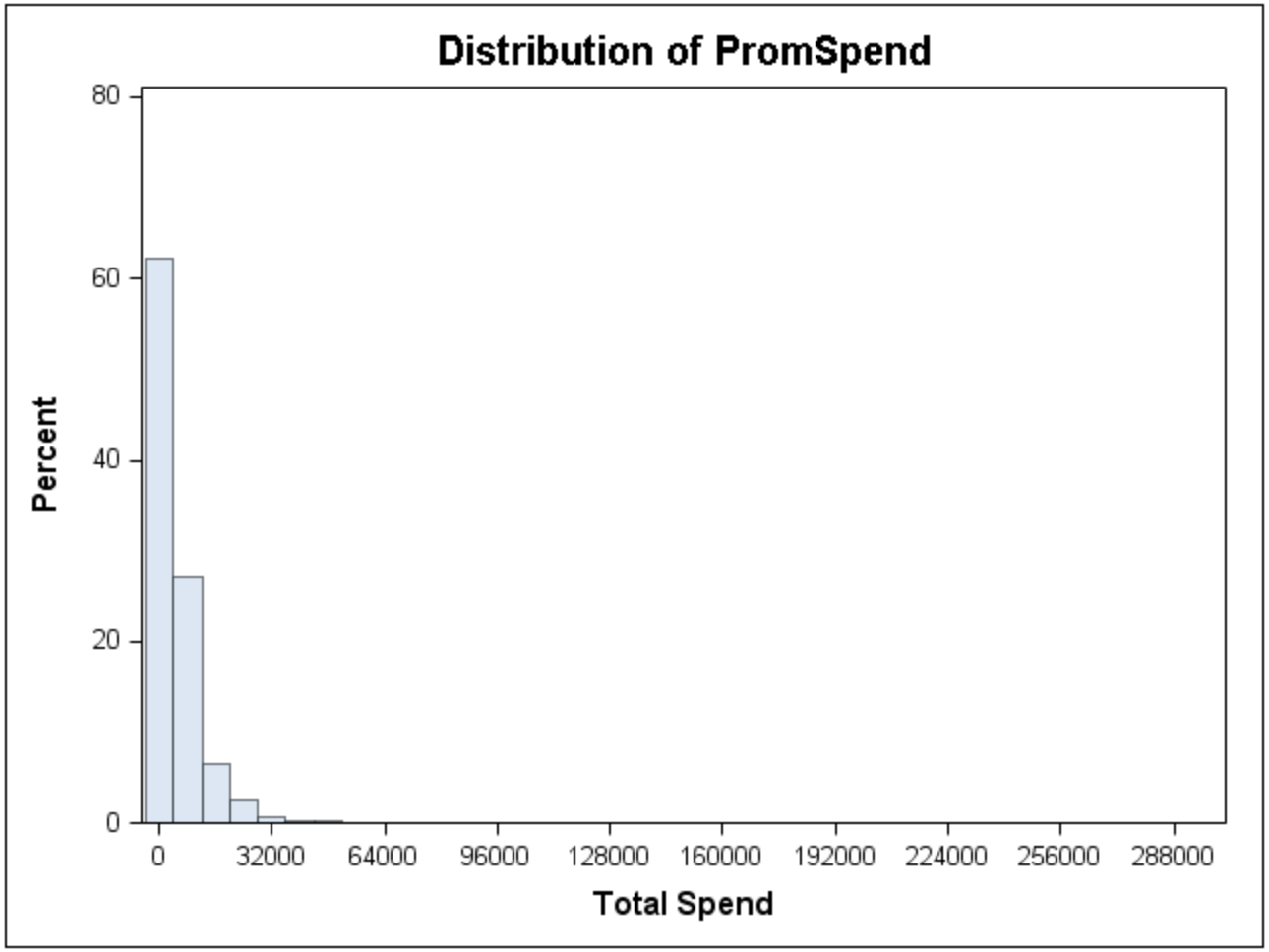
The UNIVARIATE Procedure



A picture is worth…

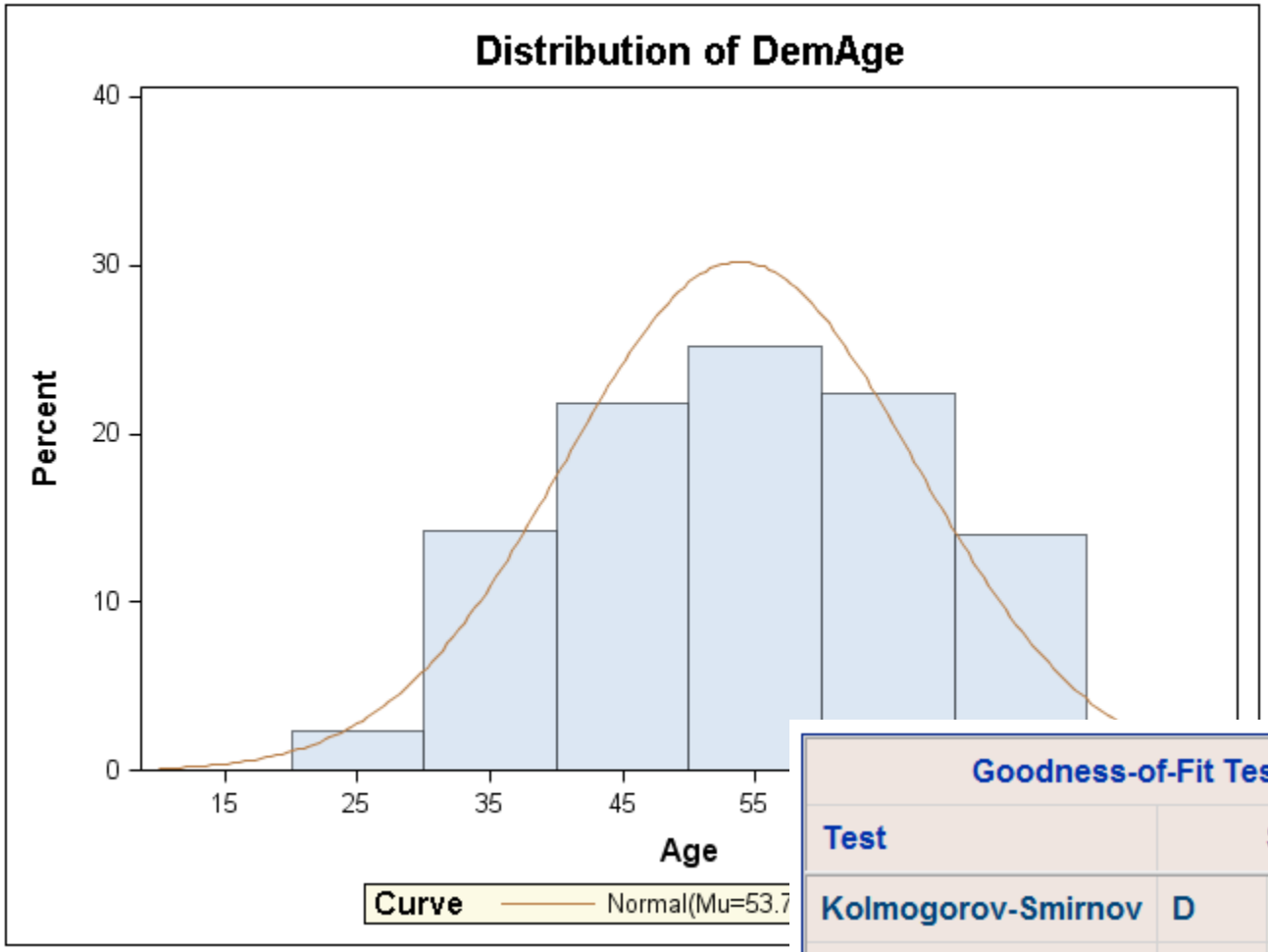


A picture is worth…



Is it Normal?

The UNIVARIATE Procedure



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0493570	Pr > D	<0.010
Cramer-von Mises	W-Sq	10.8931792	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	80.1149533	Pr > A-Sq	<0.005

Just to clear things up…

SAS Base

- FREQ
- MEANS
- UNIVARIATE

SAS/STAT

- T-Test
- NPAR1WAY
- ANOVA
- REG
- LOGISTIC

Base STAT

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

Alternative Hypothesis

- There is an association between GENDER and TARGETBUY
- The probability of purchasing organic items is different between males and females

Chi-Square Test

Base

No Association

- Observed frequencies = expected frequencies

Association

- Observed frequencies ≠ expected frequencies

The FREQ Procedure

Frequency Expected	Table of Gender by TargetBuy		
	TargetBuy(Organics Purchase Indicator)		
Gender(Gender)	0	1	Total
Female	7,944 8,652	4,205 3,497	12,149
Male	4,849 4,141	966 1,674	5,815
Total	12,793	5,171	17,964

Frequency Missing = 4259

SAS Code

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

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

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

Adding Chi-Square to our FREQ Output

Base

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Row Pct

Table of Gender by TargetBuy

Gender(Gender)	TargetBuy(Organics Purchase Indicator) 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

Frequency Missing = 4259

SAS Code

Base

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

Adding Chi-Square to our FREQ Output (continued)

Base

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
- 0 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 (continued)

Base

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	

When not 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	Purchased
1	A	yes
2	A	yes
3	A	yes
4	A	no
5	B	yes
6	B	no
7	B	no
8	B	no
9	B	no

Fisher's Exact Test

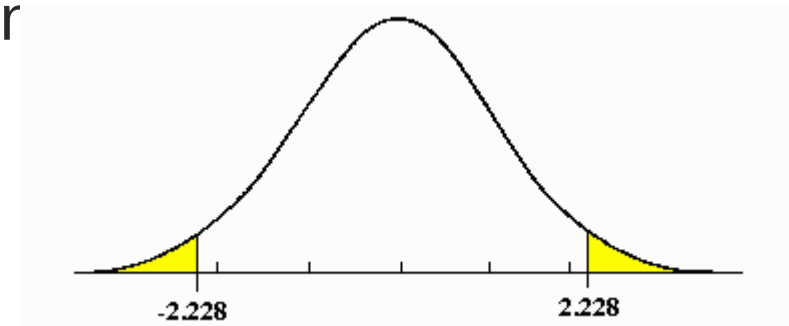
- Useful when sample sizes are small (less than 20-25 total)
- 2x2 tables
- Calculates probabilities by considering every possible table where the marginal (row and column totals remain fixed)
- Large datasets may require a prohibitive amount of time and memory for computing exact p-value.

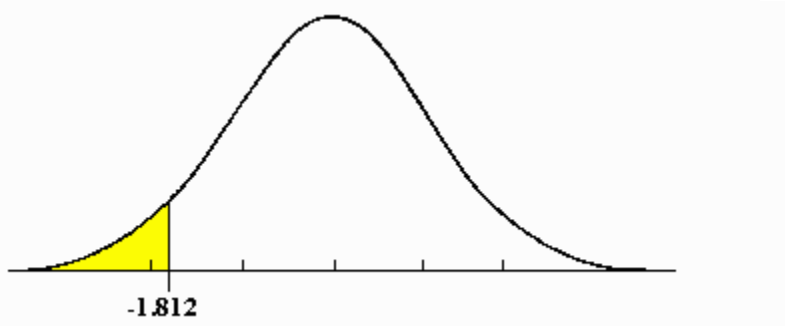
Fisher's Exact Test Hypothesis

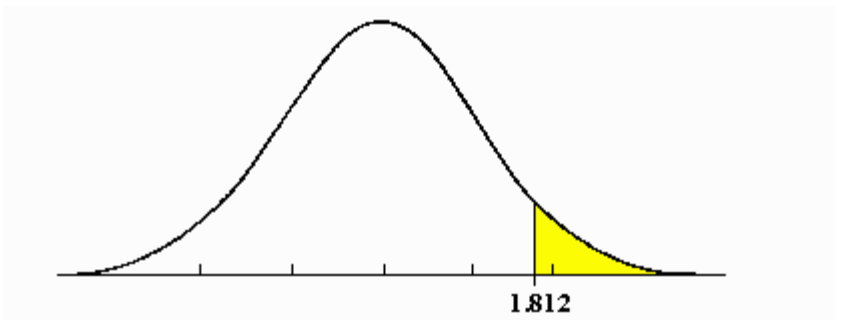
Null Hypothesis: No Association

Alternative Hypothesis:

- Two-Tailed
- Left-tailed
- Right-tailed







FREQ output

Base

Frequency Expected Cell Chi-Square Col Pct

Table of Product by Purchased

Product	Purchased		
	no	yes	Total
A	1 2.2222 0.6722 20.00	3 1.7778 0.8403 75.00	4
B	4 2.7778 0.5378 80.00	1 2.2222 0.6722 25.00	5
Total	5	4	9

Statistic	DF	Value	Prob
Chi-Square	1	2.7225	0.0989
Likelihood Ratio Chi-Square	1	2.8626	0.0907
Continuity Adj. Chi-Square	1	0.9506	0.3296
Mantel-Haenszel Chi-Square	1	2.4200	0.1198
Phi Coefficient		-0.5500	
Contingency Coefficient		0.4819	
Cramer's V		-0.5500	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

FREQ – Fisher's Exact Test

Base

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.1667
Right-sided Pr >= F	0.9921
Table Probability (P)	0.1587
Two-sided Pr <= P	0.2063

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

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

Base

The FREQ Procedure

Frequency Expected Cell Chi-Square Row Pct	Table of LoyaltyStatus by TargetBuy		
LoyaltyStatus(Loyalty Status)	TargetBuy(Organics Purchase Indicator)		
	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

Base

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

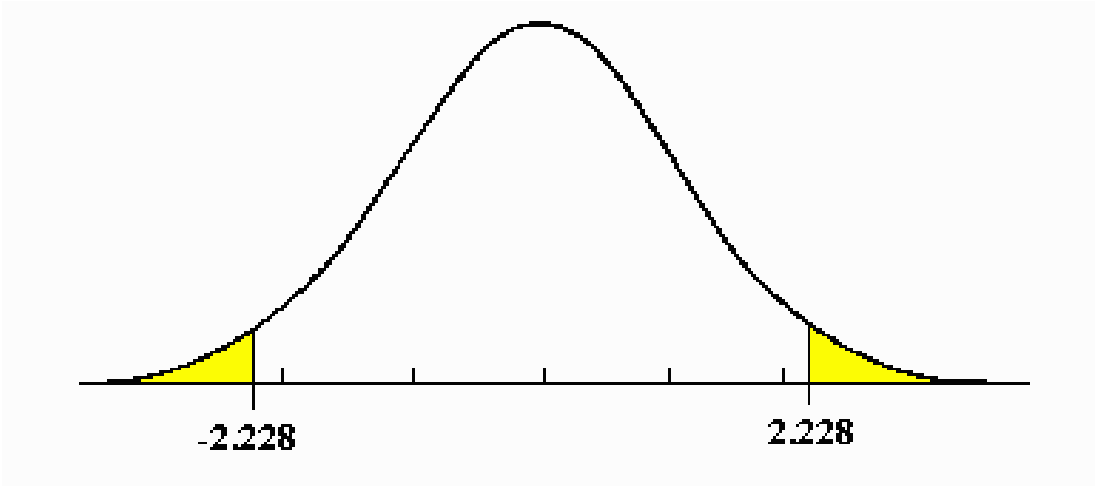
Mantel Haenszel Chi-Square test with Spearman Correlation

Base

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

t-test

1. One Sample
2. Two Sample
3. Paired



t-test -One Sample

Parametric test to compare sample mean with known value

- Null Hypothesis: H_0: μ = hypothesized value
- Alternative Hypothesis: H_a: μ ≠ hypothesized value

For our Example μ = 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=0.05
        H0 =47
        CI = EQUAL;

        VAR DemAge;
        BY TargetBuy;
RUN;
```

Can also calculated in PROC UNIVARIATE

t-test -One Sample

STAT

Organics Purchase Indicator=1

t Test

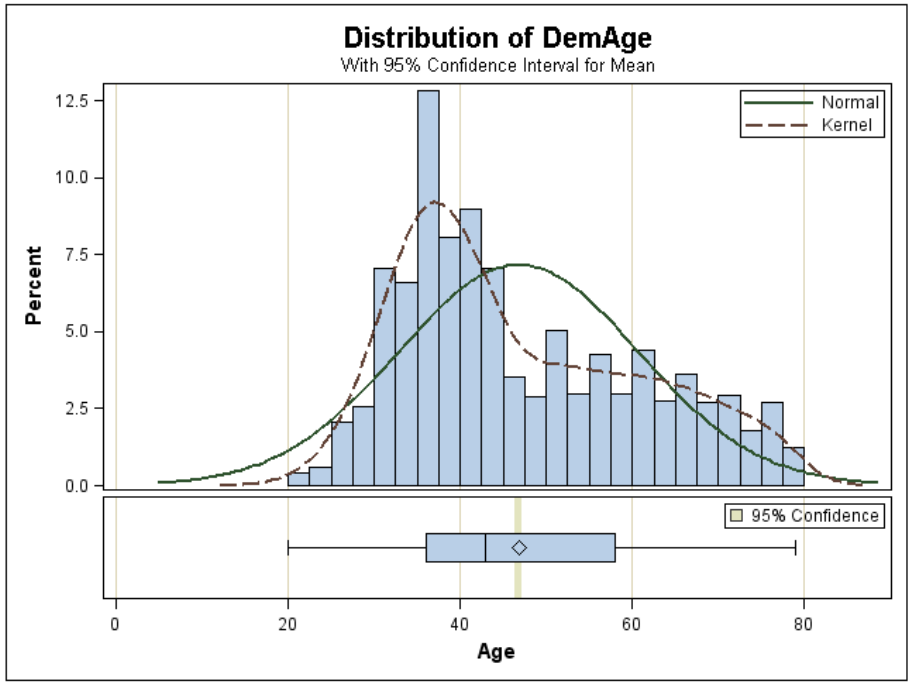
The TTEST Procedure

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

STAT

Organics Purchase Indicator=0

t Test

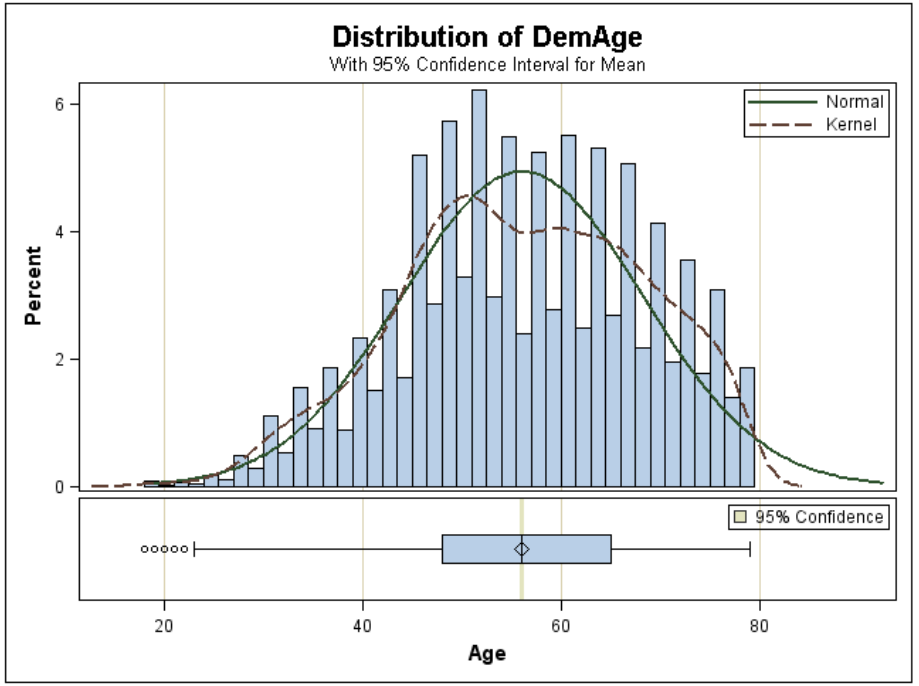
The TTEST Procedure

Variable: DemAge (Age)

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

Mean	95% CL Mean		Std Dev	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
μ_1 is amount Females spent during the promotion
μ_2 is amount Males spent during the promotion

Assumptions

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

t-test -One Sample

```
PROC TTEST
        DATA = mydata.organics
        PLOTS(ONLY)=SUMMARY
        ALPHA=0.05
        H0 =0
        CI = EQUAL;

        CLASS Gender;
        VAR PromSpend;
RUN;
```

t-test –Two Sample

t Test

The TTEST Procedure

Variable: PromSpend (Total Spend)

Descriptive Statistics

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	12149	4260.5	7128.8	64.6762	0.0100	239542
Male	5815	4592.6	7889.6	103.5	0.0100	296314
Diff (1-2)		-332.0	7383.6	117.7		

Go with Unequal Variance Test

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		4260.5	4133.7	4387.3	7128.8	7040.3	7219.6
Male		4592.6	4389.7	4795.4	7889.6	7748.8	8035.7
Diff (1-2)	Pooled	-332.0	-562.8	-101.3	7383.6	7308.1	7460.8
Diff (1-2)	Satterthwaite	-332.0	-571.2	-92.8713			

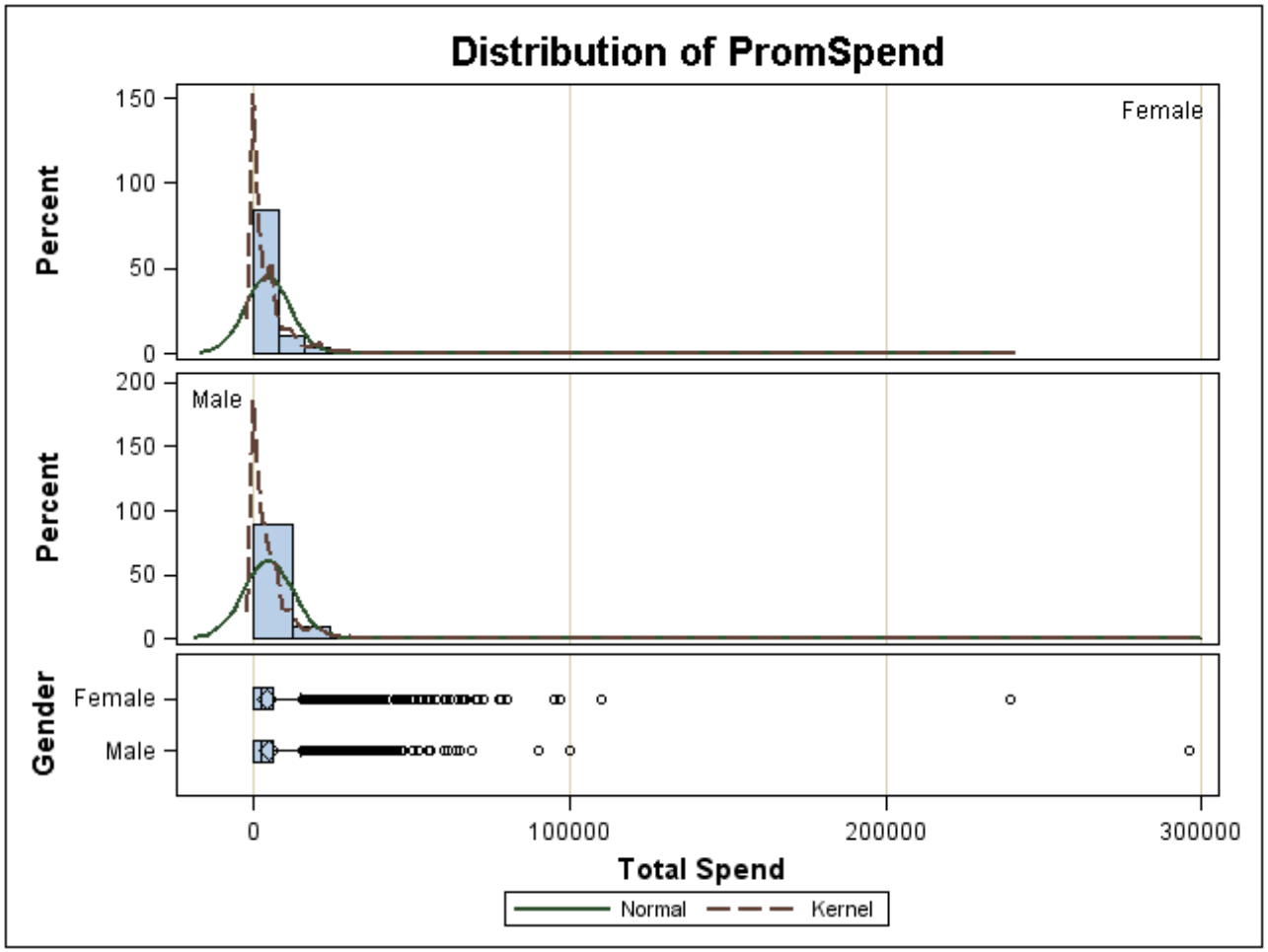
Go with Unequal Variance Test

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	17962	-2.82	0.0048
Satterthwaite	Unequal	10480	-2.72	0.0065

Equality of Variance

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	5814	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: $\mu_{post} = \mu_{pre}$
- Alternative Hypothesis: H_a: $\mu_{post} \neq \mu_{pre}$

For our Example

μ_{post} is amount of organic items bought during promotion

μ_{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 –Two Sample

t Test

The TTEST Procedure

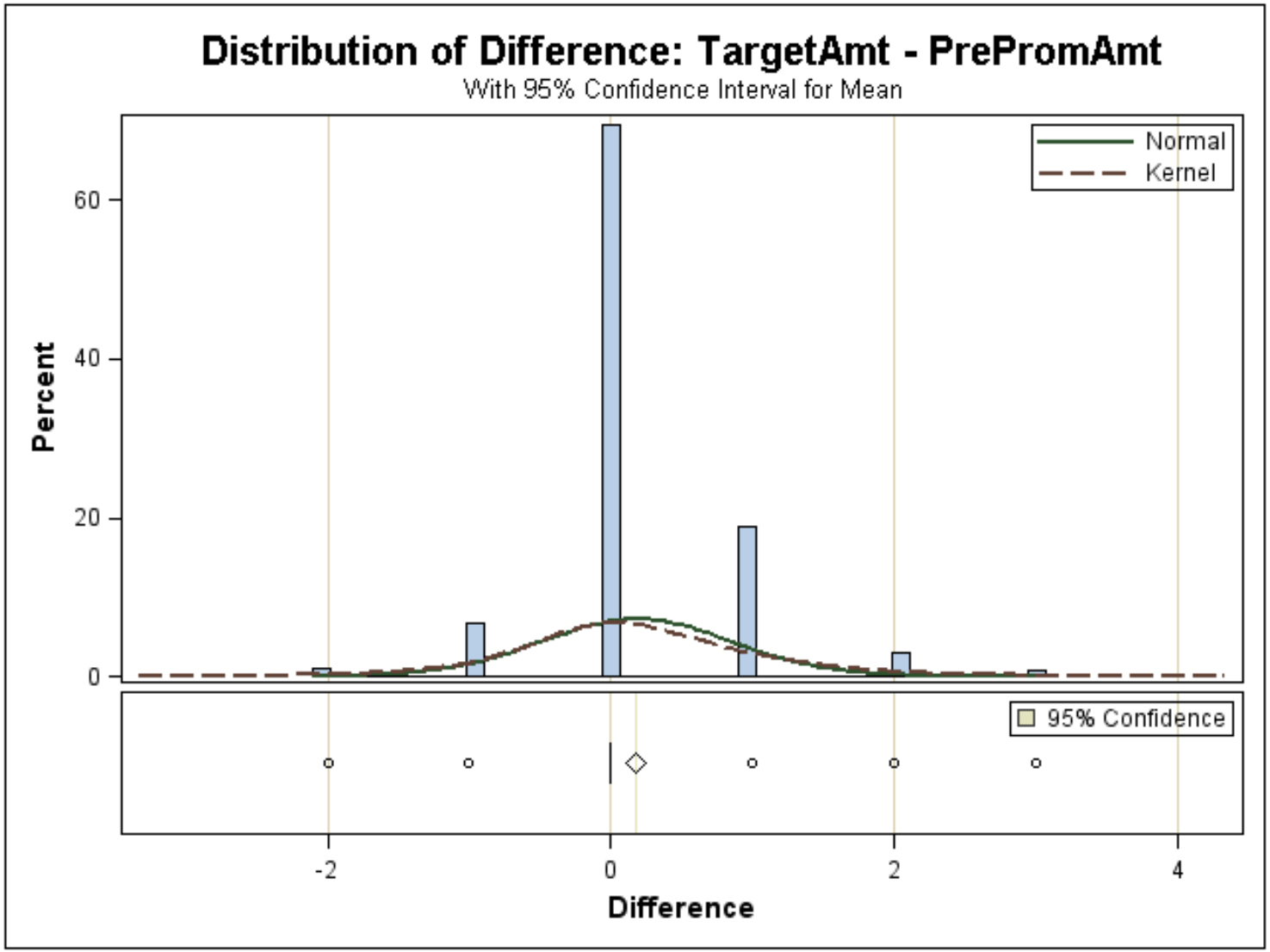
Difference: TargetAmt - PrePromAmt

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

Mean	95% CL 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

t-test –Two Sample



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

- 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

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 NPAR1WAY – 2 levels

STAT

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Diff_Amt Classified by Variable Gender

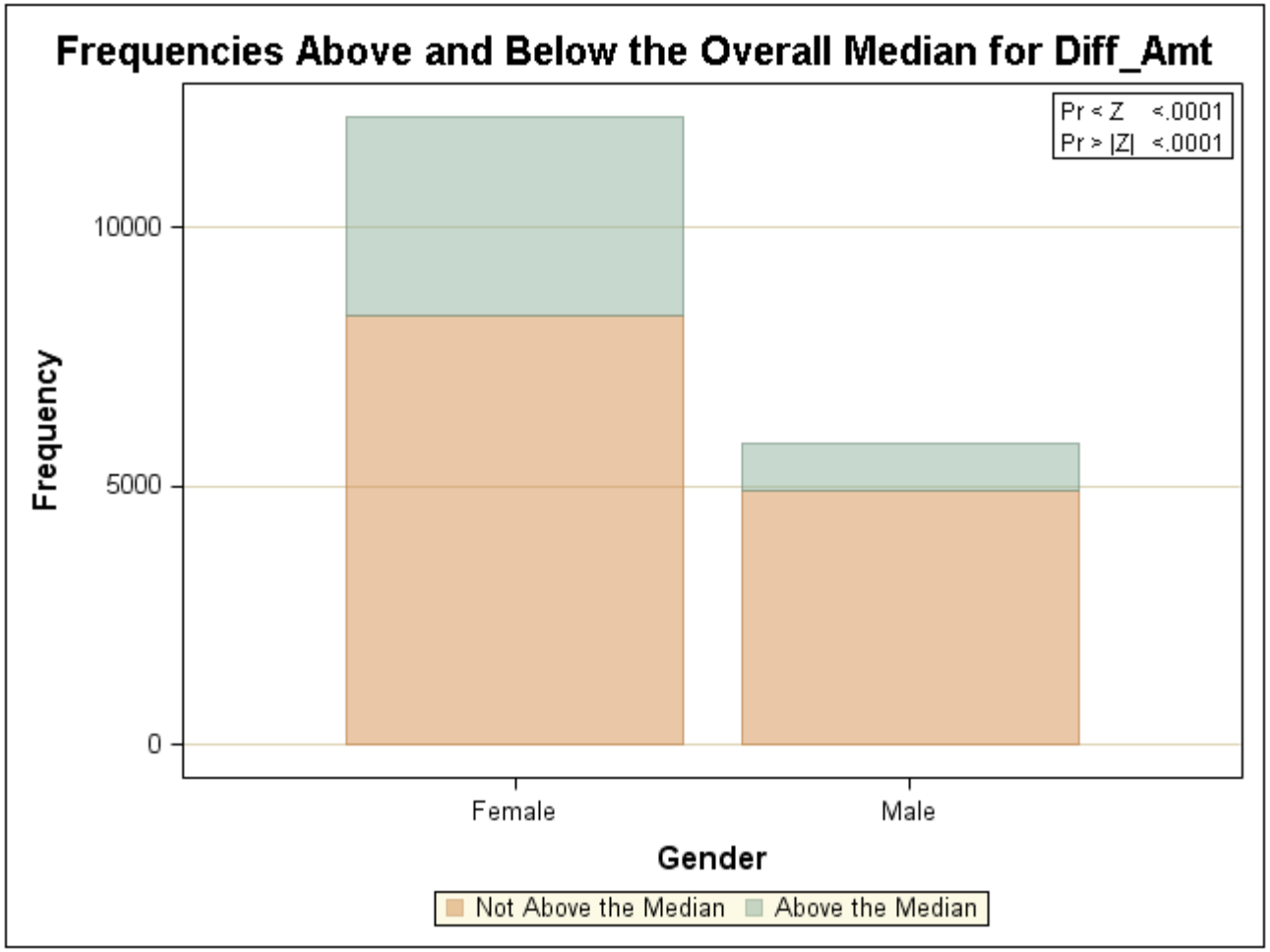
Gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Female	12149	114836825	109128393	271388.619	9452.36847
Male	5815	46524805.5	52233237.5	271388.619	8000.82640

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

STAT

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

PROC NPAR1WAY DATA=organics2 WILCOXON MEDIAN;
        VAR Diff_Amt;
        CLASS LoyaltyStatus;

RUN;
```

PROC NPAR1WAY

STAT

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Diff_Amt Classified by Variable Loyalty Status

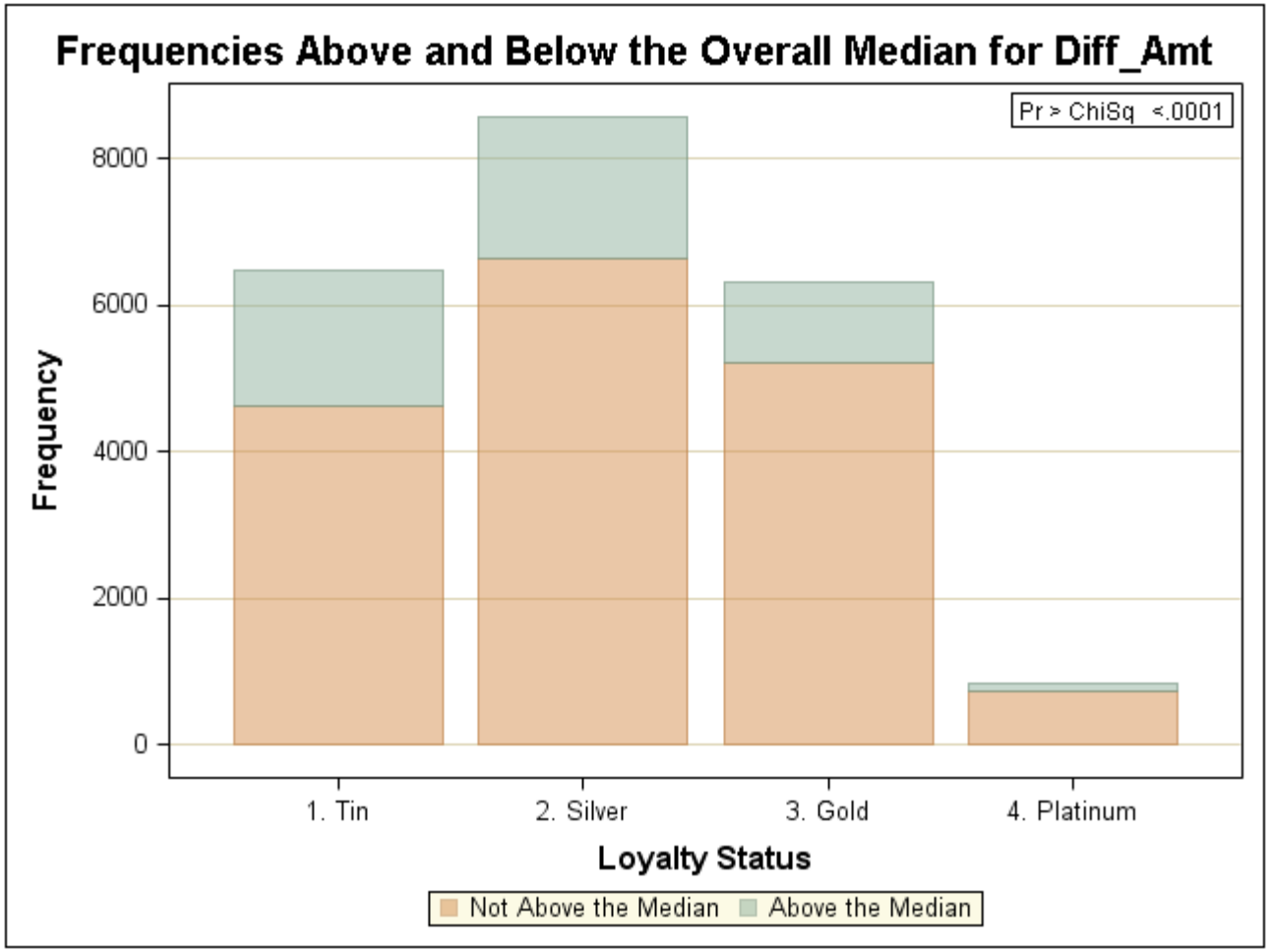
Loyalty Status	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1. Tin	6487	76643119.0	72083544.0	352231.776	11814.8788
2. Silver	8572	95164849.5	95252064.0	377122.743	11101.8257
3. Gold	6324	66574842.5	70272288.0	349574.903	10527.3312
4. Platinum	840	8559165.0	9334080.0	147752.001	10189.4821

Average scores were used for ties.

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

PROC NPAR1WAY

STAT



Bonus Section from Customer Loyalty

Technical Team made up of 16 members

- 250 years at SAS
- 342 years of SAS experience

Bonus Section from Customer Loyalty

1. Support.sas.com
2. Don’t carry excess baggage in programs
   - Use Keep and Drop options on DATA and SET statements
   - Take advantage of SAS indexes when possible
   - Use First. and Last. In data steps
3. Creating dummy variables in PROC SQL
4. Use multiple @’s in front of macro variables
5. SAS Enterprise Guide

Resources

Public SAS Courses

- Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Online Tutorials

- SAS Online Resources for Statistics Education
  - t-tests
  - Tests of Association
    - » Pearson Chi-Square
    - » Mantel-Haenszel Chi-Square
  - Nonparametric Analysis

It's where you'll find the latest information on:

- New training courses and services
- Special offers and discounts
- The latest course schedules
- New training locations
- Events and conferences
- SAS certification news
- And, much more.

Everything you need – in one place.
Visit and bookmark it today.

CUSTOMER LOYALTY TEAM • Support You Can Count On

sas THE POWER TO KNOW.

Online Resources

- Exact Methods in the NPAR1WAY Procedure
- Support.sas.com summary and SAS/Stat Documentation
- SAS/Stat 9.2 User's Guide – The NPAR1WAY Procedure
- SAS/Stat 9.3 User's Guide – The NPAR1WAY Procedure
- An Overview of Non-parametric Tests in SAS: When, Why and How

General SAS Resources

- What's New in SAS 9.3 Book
  - http://support.sas.com/documentation/cdl/en/whatsnew/64209/HTML/default/viewer.htm#bookInfo.htm
- SAS/Stat Newsletter
  - http://support.sas.com/community/newsletters/index.html
- Stat, IML, OR, ETS Papers
- Discussion Forum
- Videos
  - Youtube.com

    http://www.youtube.com/playlist?list=PL0B05D53A5E101AA6
  - Video portal to the STAT and OR focus area.

    http://support.sas.com/rnd/app/video/index.html

Thank you for using SAS!

www.sas.com