Cross Tab and Chi-Square Tests

Lecture

Cross Tab or Contingency Table (Pivot Table in Excel) (Business Questions)

- Used to answer business questions such as:
  - Consider two events: a name belongs to either list A or list B; a person either responds or does not respond to a direct mail offer. Are these two events **independent**?
  - Is there **any association** (relationship) between whether a person carries an unpaid balance (or not) on his/her credit card and the marital status (single, married, divorced/separated) of that person?
- Why do we need a test? Why can't we just look at the numbers and answer the business questions?

Cross-Tab Mechanics

- It produces a summary table that classifies each observation in a data set with respect to the two categorical variables.
- The entries in the summary table are typically *observed (O) counts* in different cell combinations.
    - Sometimes the counts are converted (most programs do it automatically) to percentages or probabilities.
- *Expected (E) counts* in in different cell combinations are calculated assuming the variables are independent
- The statistic used is *Chi-square statistic*

$$\chi^2 = \Sigma \frac{(E-O)^2}{E}.$$

- The df for the chi-square statistics is (r-1)(c-1)
- The p-value for the Chi-Square test is used to make your decision.

Procedure for Cross Tab and Chi-Square Test

- Select two categorical variables that you want to analyze
- Write hypotheses about association (or, independence) between the categorical variables
- Choose level of significance
- Use *Chi-square test* to test if the variables are independent in the population
- Make decisions about independence of variables based on the p-value from the Chi-Square test
- Interpret cross-tab using probabilities

More on Cross-tabs and Chi-Square

- Chi-square and its p-value is highly sensitive to sample size.
  - In most direct marketing applications with very large sample size, even a very small difference may become statistically significant!
- Statistical significance does not imply managerial significance(!)
  - That is, managers may choose not to take any action even if there is statistical significance in cross-tab.
  - Need to evaluate what’s the difference in probabilities and is that meaningful to warrant managerial action

Strength of Association

- One problem with cross-tab and chi-square test is that, while we can say there is a relationship between the variables, but we **can not say how strong or weak** that relationship is!!!
- Many measures of association exists for nominal variables such contingency coefficient, Phi coefficient, Cramer’s V, Lambda symmetric, uncertainty coefficient etc.
  - They generally have values between 0 to 1. So, it’s easy to make a judgement of whether the association is weak or strong
  - Interpretation : 0-0.3 is low, 0.3-0.7 is moderate and 0.7+ is high.
- Other measures of association include Kendall’s Tau, Gamma, etc.
  - These are more applicable for ordinal association

Cross-tab for a Continuous Variable?

- What if we want to relate a continuous variable (such as Age measured in years) and a categorical variable (such as whether a person responds to a direct mail offer or not) via cross-tab?
  - This may create problems because Age may have too many distinct values!
    - That will result in too many cells in the cross-tab and low (less than 5 which violates test assumptions) sample size per cell.
  - The solution is to first transform the continuous variable *AGE* into a discrete variable, such as *CAT_AGE* (less than 21, 21-30, 31-40, 41-50, 51-60, 61-70 and 71+) and then do cross-tab of *CAT_AGE* with Response to an offer.

Cross Tab and Chi-Square Tests

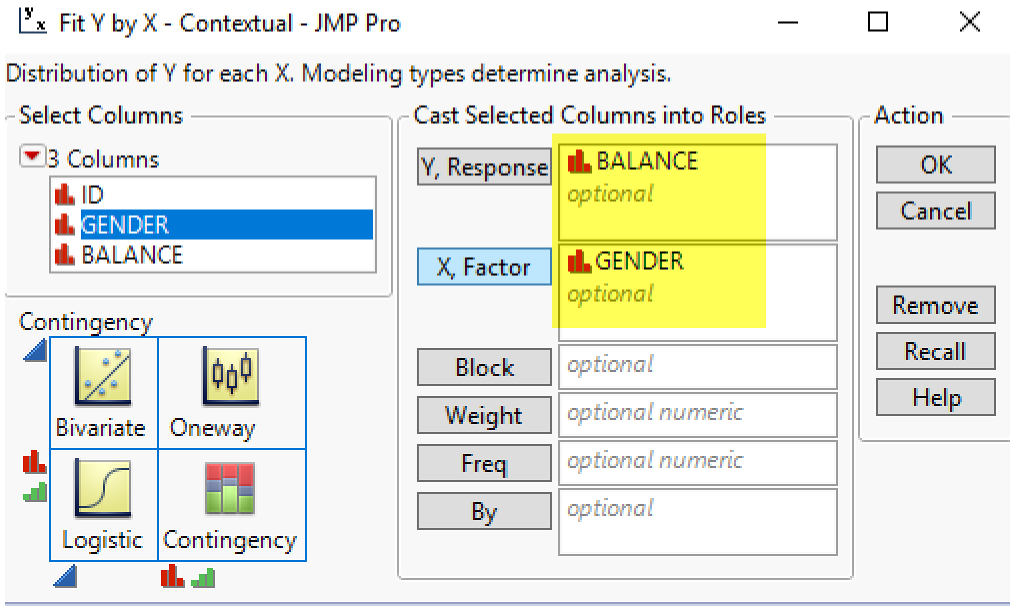
Demonstrations using JMP

Data Set: Credit_Balance_Data

- Variables in data set are:
  - ID : customer ID
  - Balance: whether a customer has carried an unpaid balance at least once in last 1 year on this credit card (Yes/No)
  - Gender: Male or Female
- In this example, we have two variables: Gender (**M/F**) and Balance (**Y/N**). So, each observation will be classified as belonging to one of the four combinations – **MY**, **MN**, **FY** and **FN**
- A cross-tab will count the number of observations in each of the four combinations and produce a summary table with those counts.

Data Set : Credit_Balance_Data

- Business Question: Is there any association between Balance and Gender ?
- Hypotheses for testing association:
  - Null: Balance and Gender, are independent (i.e., no association) in the population
  - Alternative: H_o is not true
- JMP > Analyze > Fit Y by X > Balance as Y, response > Gender as X, Factor > OK
- Click Red Triangle next to Contingency Analysis > Measures of Association



Fit Y by X - Contextual - JMP Pro

Distribution of Y for each X. Modeling types determine analysis.

Select Columns: 3 Columns — ID, GENDER, BALANCE

Cast Selected Columns into Roles

Y, Response: BALANCE (optional)

X, Factor: GENDER (optional)

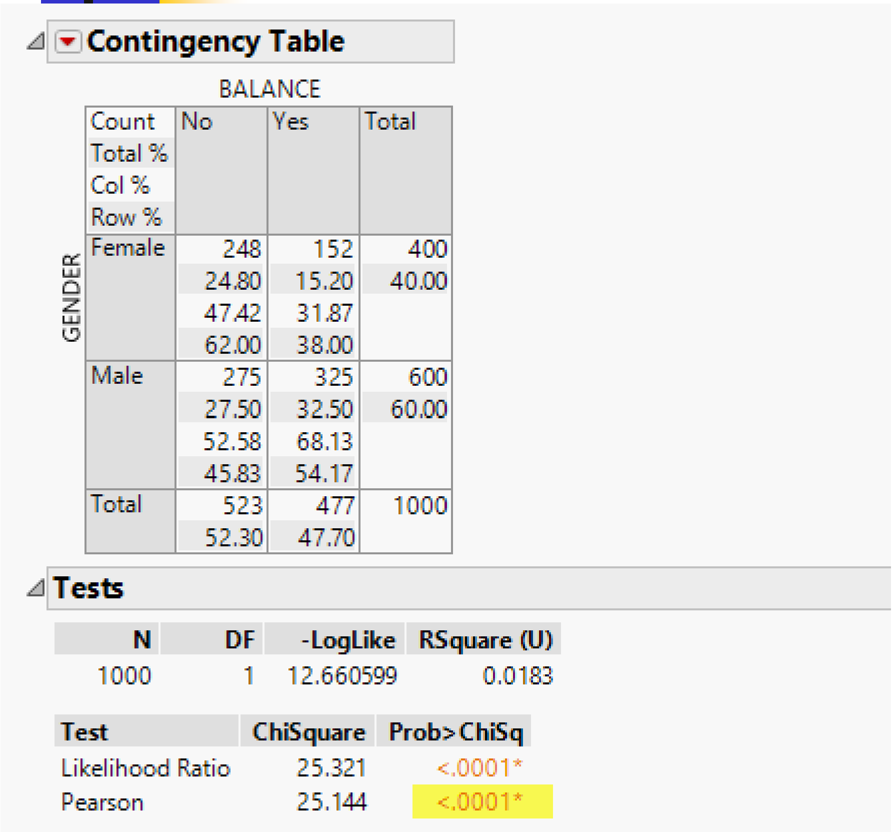
Block: optional

Weight: optional numeric

Freq: optional numeric

By: optional

Results of Chi-Square Test



Contingency Table

BALANCE

GENDER	Count Total % Col % Row %	No	Yes	Total
	Female	248 24.80 47.42 62.00	152 15.20 31.87 38.00	400 40.00
	Male	275 27.50 52.58 45.83	325 32.50 68.13 54.17	600 60.00
	Total	523 52.30	477 47.70	1000

Tests

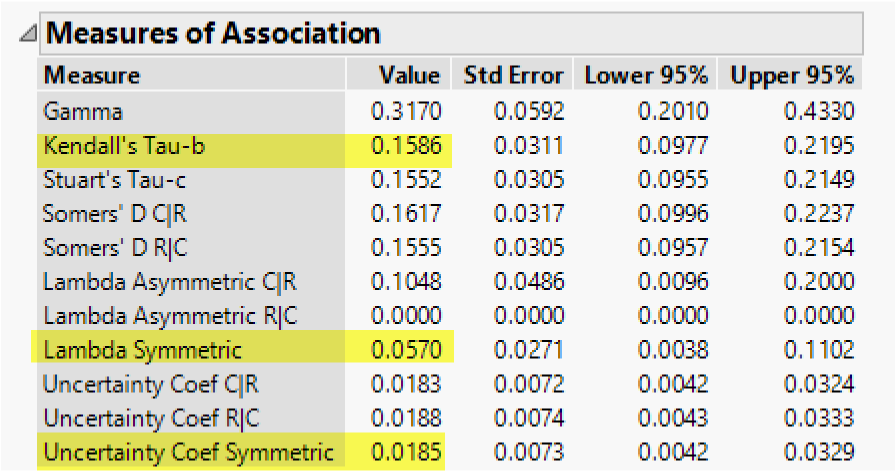
N	DF	-LogLike	RSquare (U)
1000	1	12.660599	0.0183

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	25.321	<.0001*
Pearson	25.144	<.0001*

If we pick a person at random what's the chance he/she will carry a credit card balance?

If we pick a person at random what's the chance he is a male AND carries a credit card balance?

If we pick a **male** at random what's the chance he carries credit card balance?



Measures of Association

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.3170	0.0592	0.2010	0.4330
Kendall's Tau-b	0.1586	0.0311	0.0977	0.2195
Stuart's Tau-c	0.1552	0.0305	0.0955	0.2149
Somers' D C|R	0.1617	0.0317	0.0996	0.2237
Somers' D R|C	0.1555	0.0305	0.0957	0.2154
Lambda Asymmetric C|R	0.1048	0.0486	0.0096	0.2000
Lambda Asymmetric R|C	0.0000	0.0000	0.0000	0.0000
Lambda Symmetric	0.0570	0.0271	0.0038	0.1102
Uncertainty Coef C|R	0.0183	0.0072	0.0042	0.0324
Uncertainty Coef R|C	0.0188	0.0074	0.0043	0.0333
Uncertainty Coef Symmetric	0.0185	0.0073	0.0042	0.0329

Correlations Between Variables

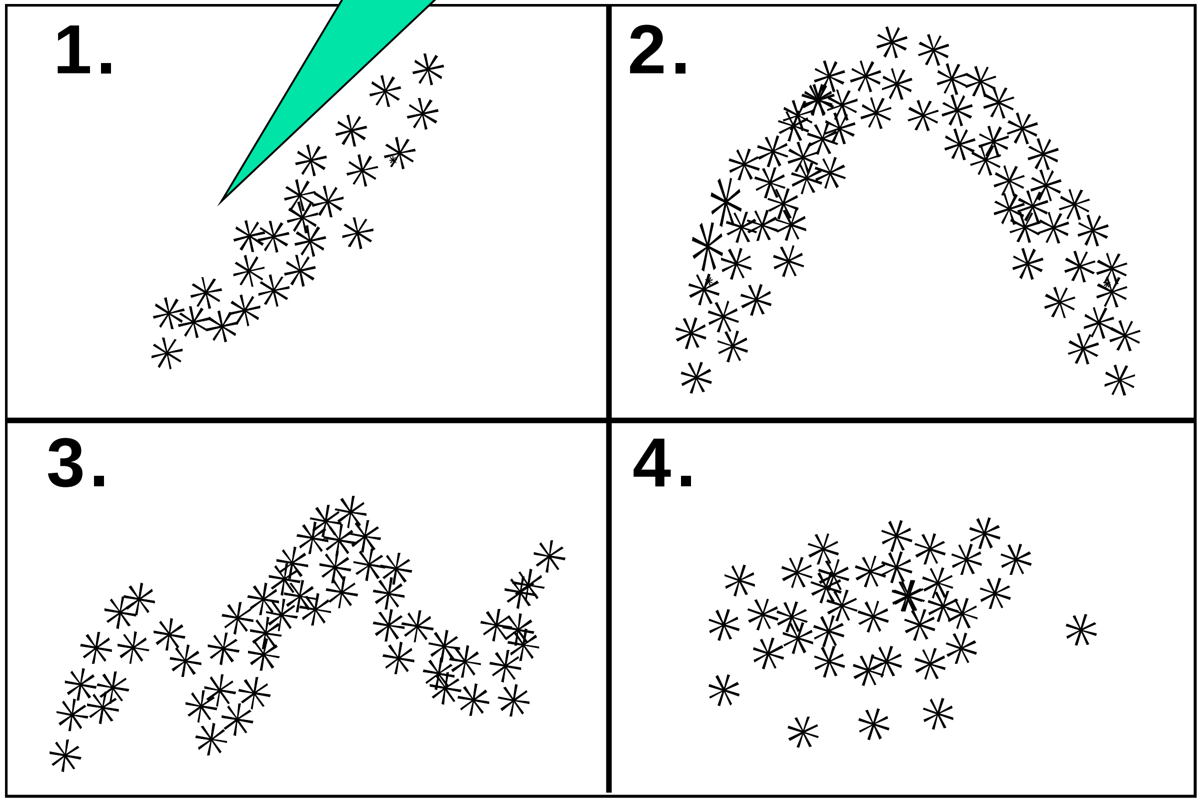
Lecture

Correlation (Business Questions)

- Is there a **relation** between age of a person and the person's income?
- Is there a **relation** between amount ($) we spend on advertising in a year and our sales revenue ($) for that year?
- Is there a **relation** between income of a person and the $ amount of the items the person ordered from a catalogue in last year?
- Why do we need a test? Why can't we just calculate correlation coefficient and answer the business questions?

Relationships Between Continuous Variables

Pearson Correlation

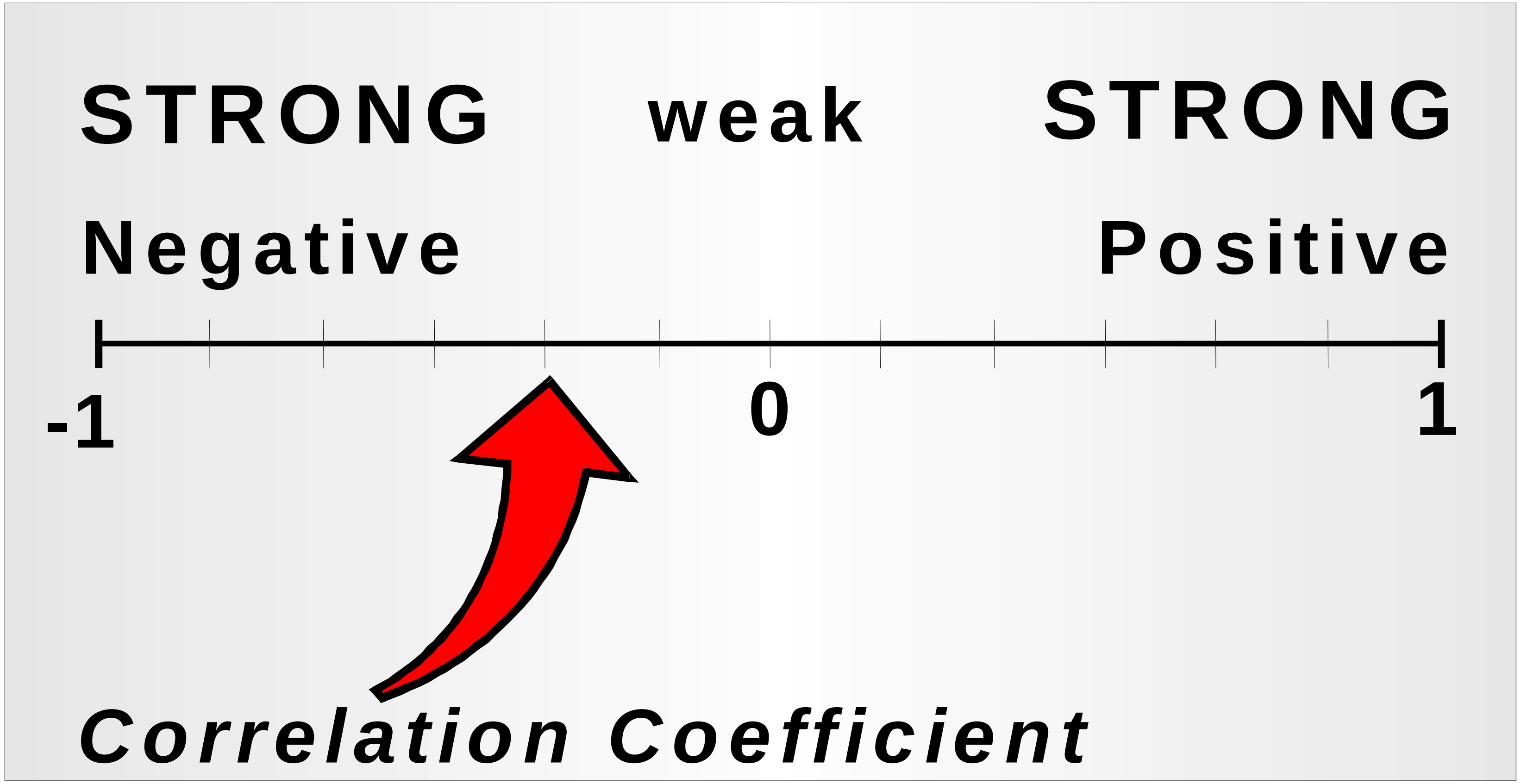


Pearson Correlation Coefficient Basics and Mechanics

- Measures **linear** (straight line-based) association (relationship) between two continuous variables
- Provides a summary statistic that shows both the *strength* as well as the *direction* of relationship between the variables
- Formula is:

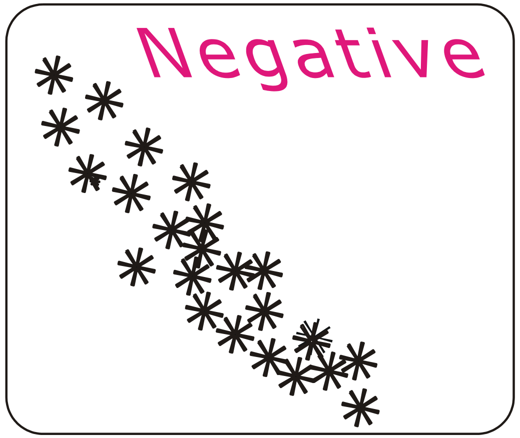
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

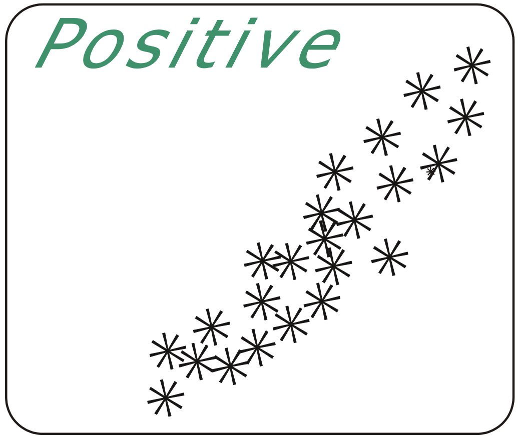
Pearson Correlation and Strength of Relationship between Variables

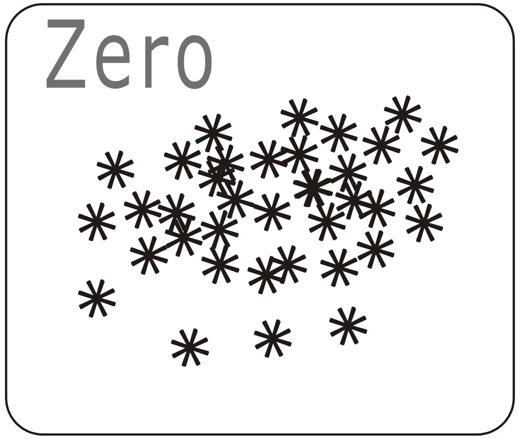


Rule of thumb: -0.3 to +0.3 low strength, -0.3 to -0.7 or +0.3 to +0.7 medium strength, -0.7 to -1 or +0.7 to +1 high strength of association

Correlation, Scatter Plots and Direction of Relationship between Variable





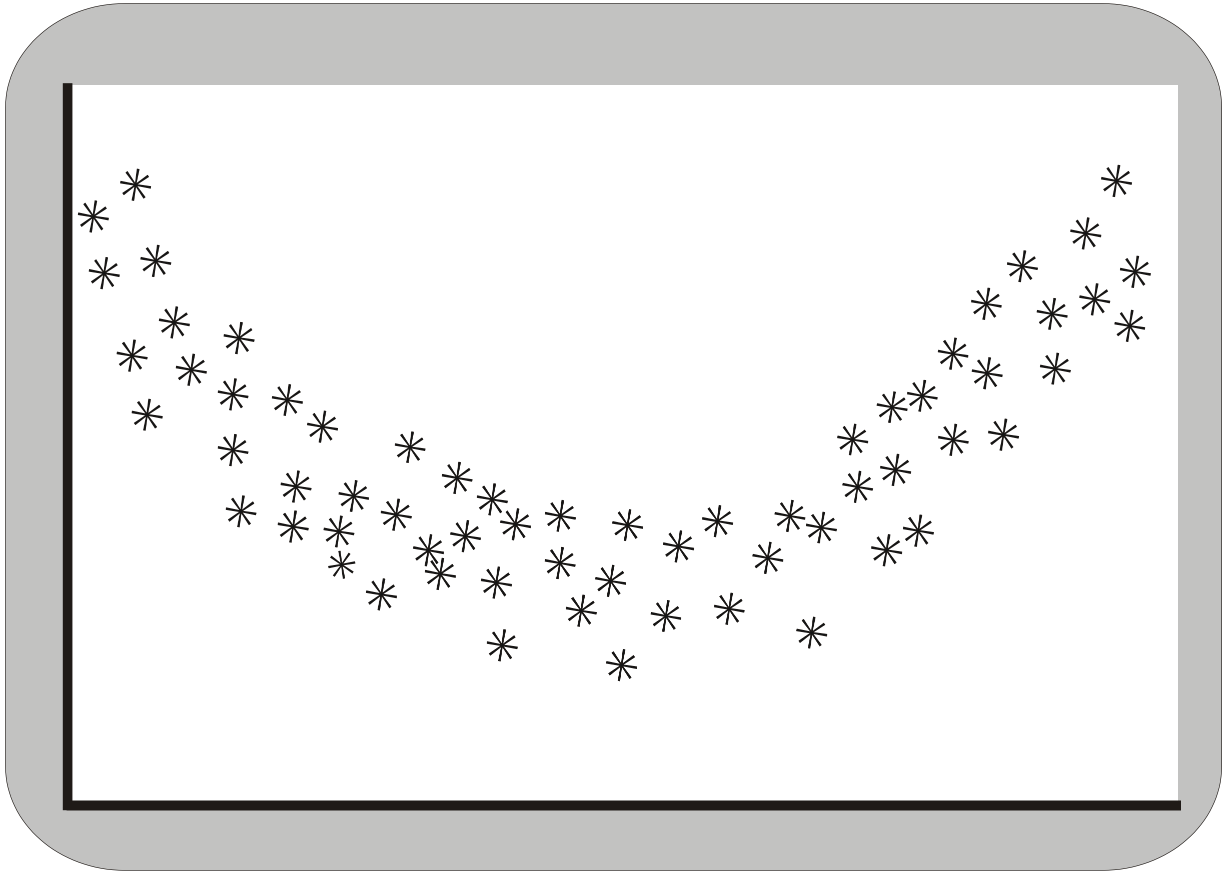


Potential Abuses of Correlation

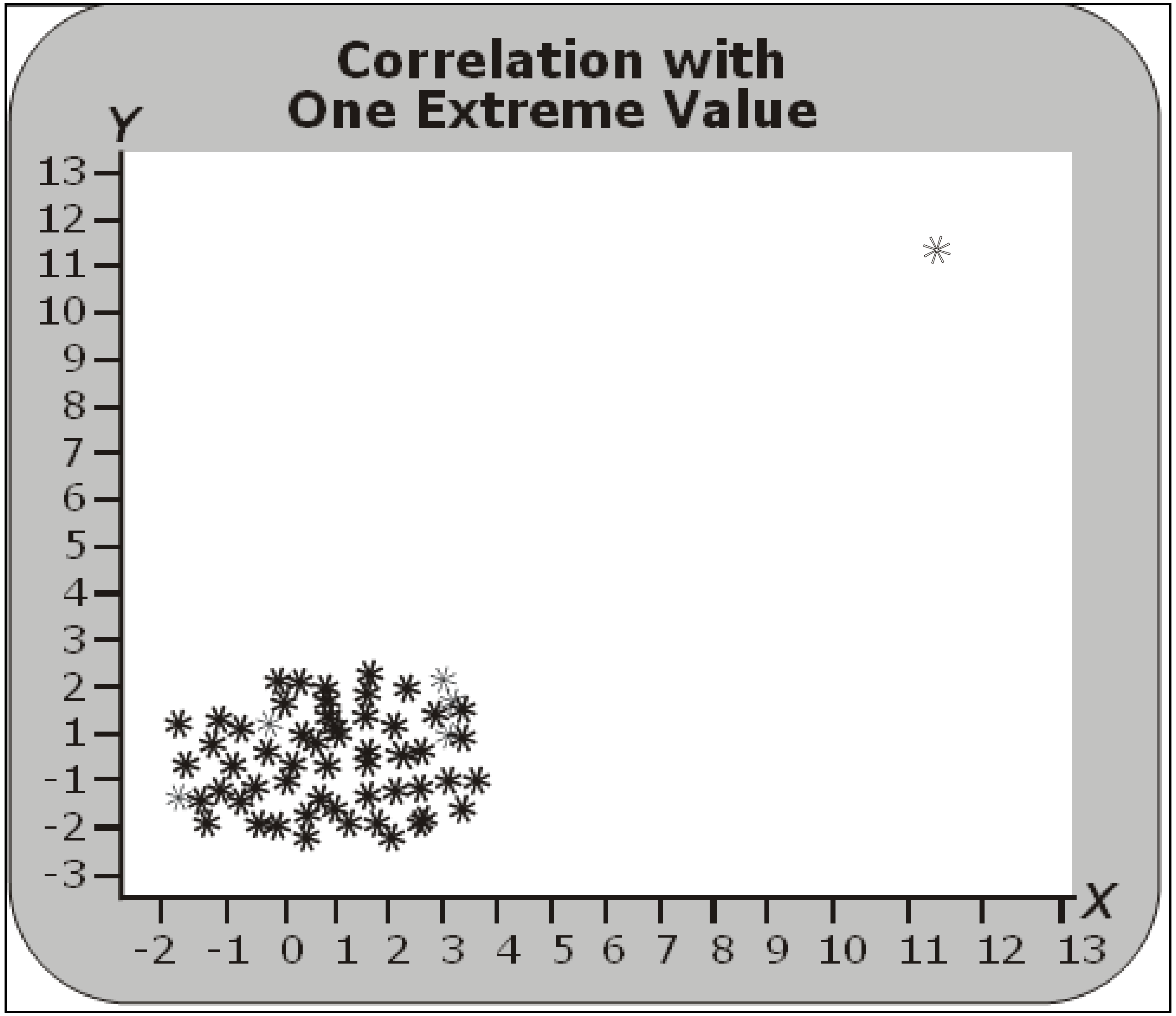
- Pearson Correlation measures *linear relation (association)* between two variables. Many people abuse correlation by doing one of the following:
- Conclude a **cause-and-effec**t between the two variables if they are correlated
- Conclude there is **no relationship** between two variables if the correlation coefficient is close to 0
- Fail to look at data and explore the **impact of extreme** values on correlation coefficient
- Some graphical examples on next few slides will demonstrate the last two abuses

Missing Another Type of Relationship

Curvilinear Relationship



Effect of One Extreme Data Value

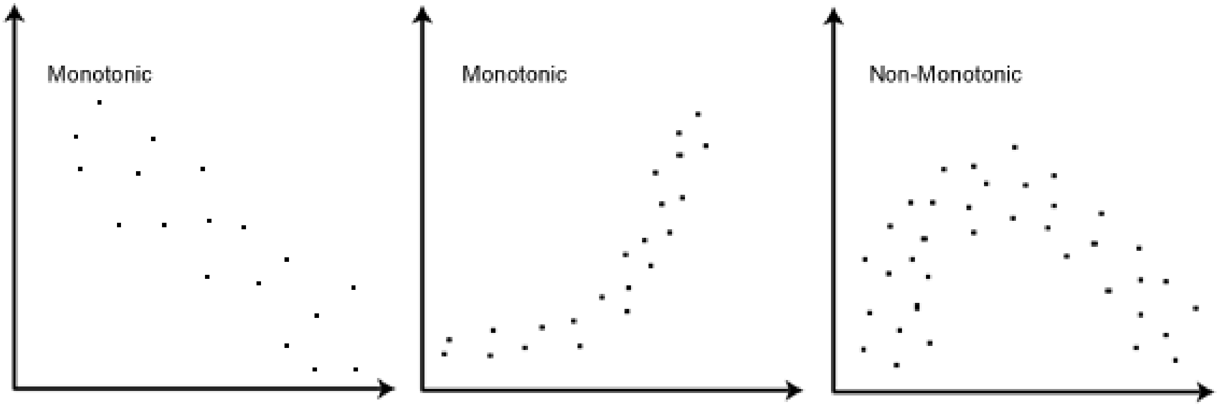


Procedure for Pearson Correlation Tests

- Select two continuous variables that you want to analyze
- Write hypotheses about linear relationship between the two variables
- Choose level of significance
- Test if the correlation coefficient is equal to 0 in the population
- Make decisions about linear relationship between variables based on the p-value of the correlation test.
- Interpret strength and direction of relationship using magnitude and sign of correlation coefficient.

Spearman Rank Order Correlation

- The **Spearman correlation** between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

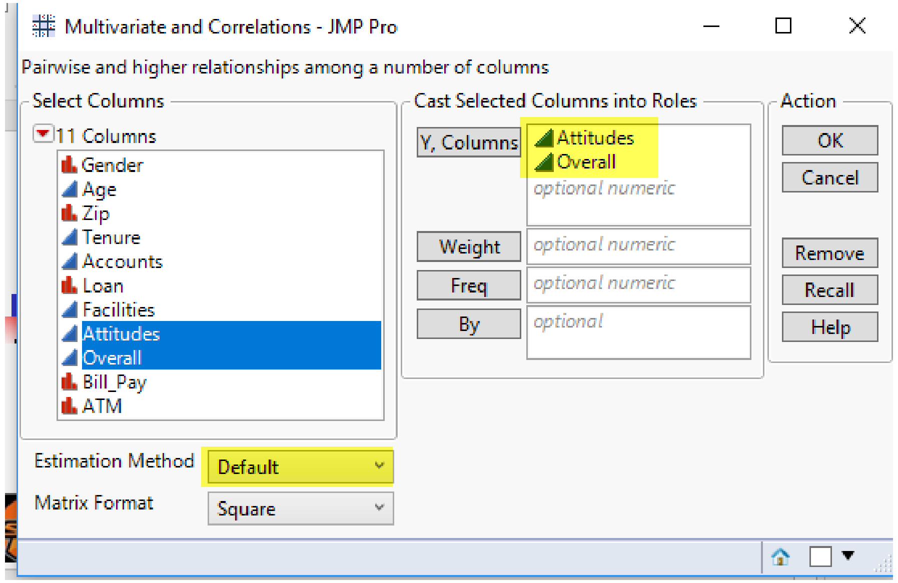


Correlations

Demonstrations using JMP

Data Set : Customer Survey

- Business Question: Is there any association (relation) between how customers feel about “employees’ attitude” and “overall satisfaction” ?
- Hypotheses for testing association:
  - Null: Attitudes and Overall are not related in the population
  - Alternative: H_o is not true
- Do it the quick and dirty approach (wrong!)
- JMP > Analyze > Multivariate Methods > Multivariate > Select Attitudes and Overall as Y, Columns > Click OK
- Click red triangle next to Multivariate > Select Correlation Probability



Multivariate and Correlations - JMP Pro

Pairwise and higher relationships among a number of columns

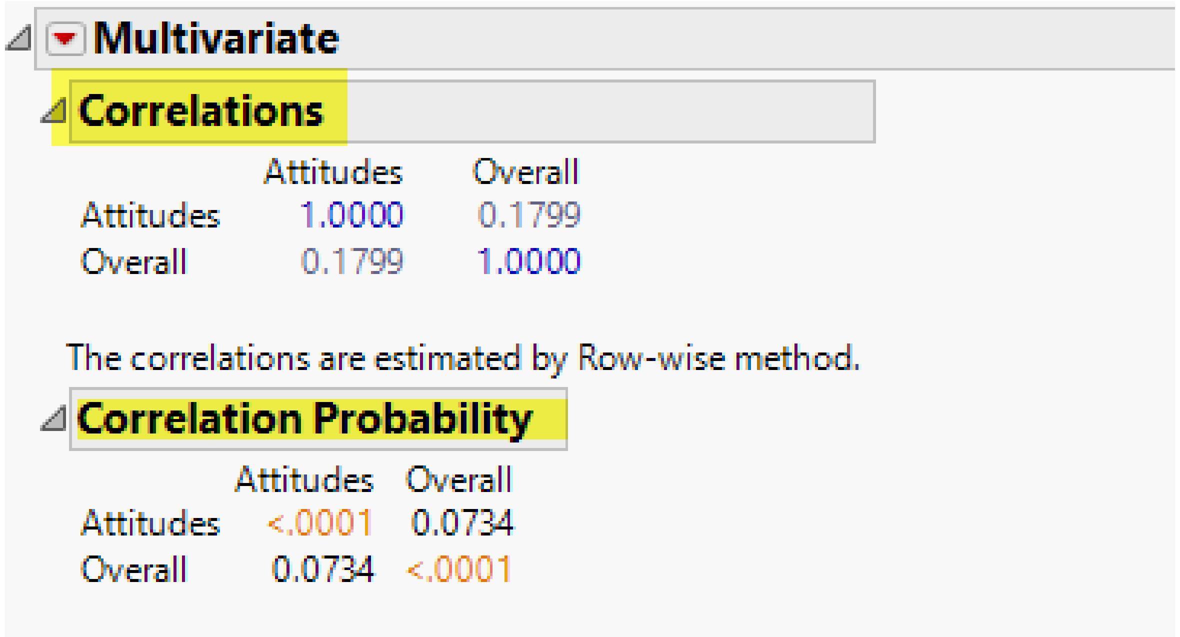
Select Columns: 11 Columns: Gender, Age, Zip, Tenure, Accounts, Loan, Facilities, Attitudes, Overall, Bill_Pay, ATM

Cast Selected Columns into Roles:
Y, Columns: Attitudes, Overall
Weight: optional numeric
Freq: optional numeric
By: optional

Estimation Method: Default

Matrix Format: Square

Results of Correlation



Multivariate

Correlations

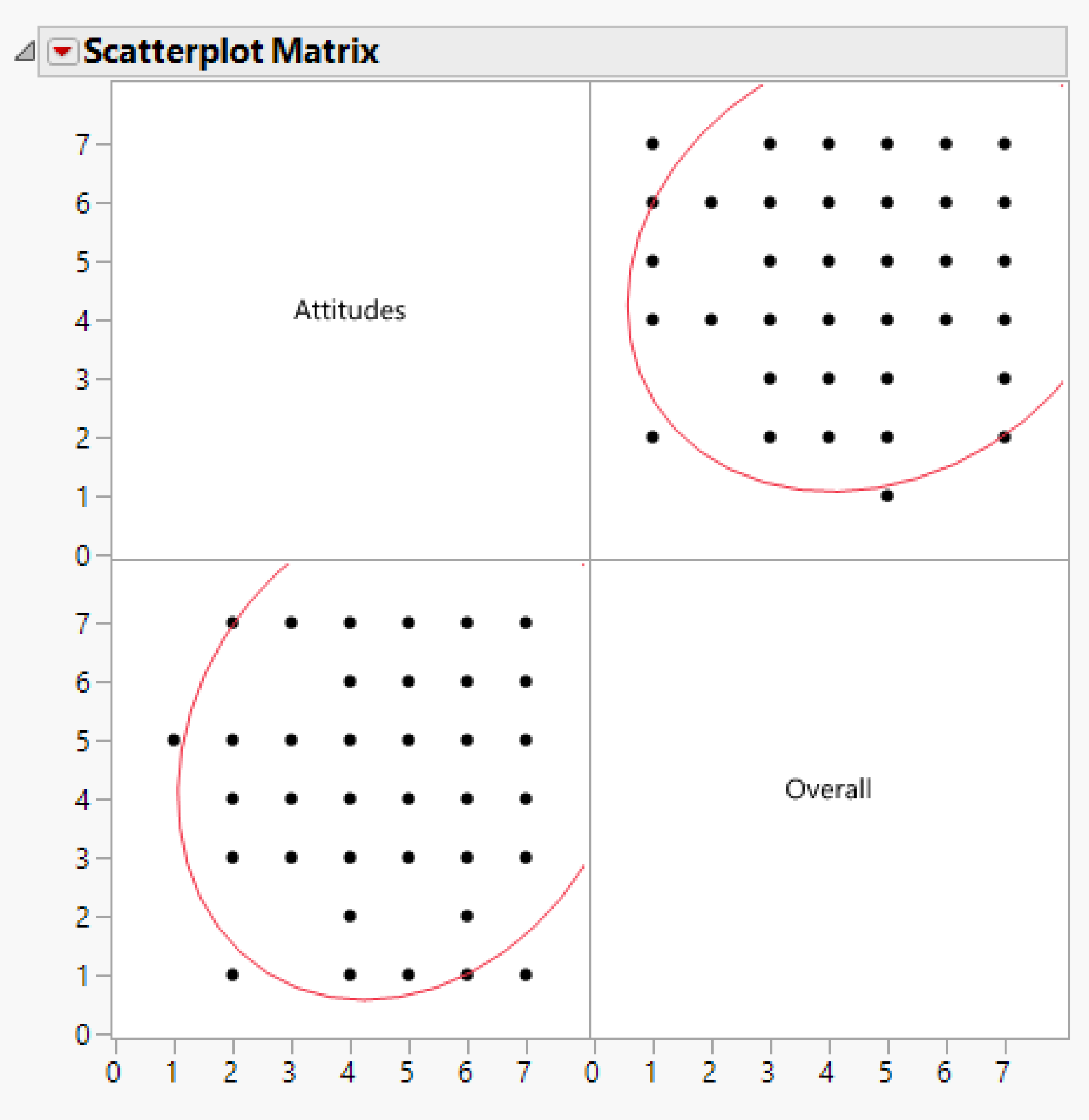
	Attitudes	Overall
Attitudes	1.0000	0.1799
Overall	0.1799	1.0000

The correlations are estimated by Row-wise method.

Correlation Probability

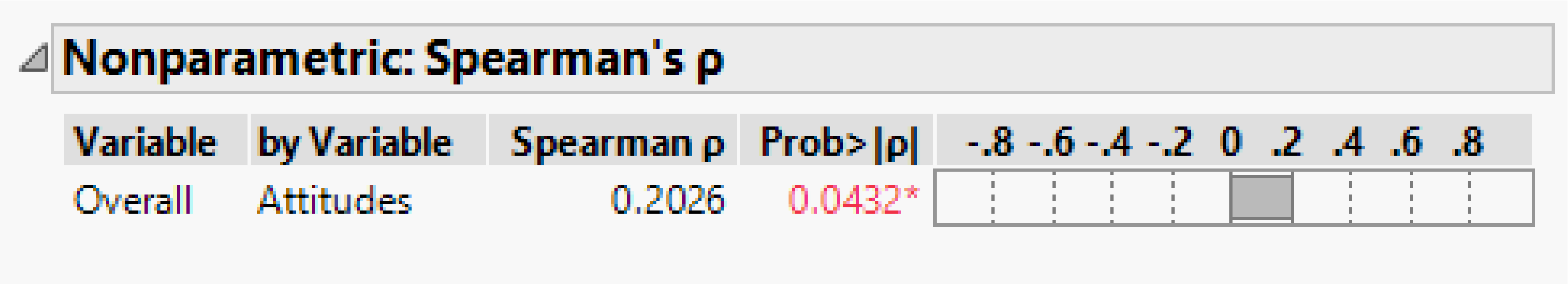
	Attitudes	Overall
Attitudes	<.0001	0.0734
Overall	0.0734	<.0001

Look at Plots



Do it Right!

- Red triangle next to Multivariate > Non parametric correlations > Spearman’s



Nonparametric: Spearman's ρ

Variable	by Variable	Spearman ρ	Prob>|ρ|
Overall	Attitudes	0.2026	0.0432*