



ILHAM-EC

Computational Data Analysis and Statistics in Agriculture (Part I)

Prof. Konstadinos MATTAS

Dr. Anastasios MICHAILIDIS

Training in Italy

Sassari, 22-24 January 2018

Co-funded by the
Erasmus+ Programme
of the European Union





Introduction to agricultural statistics using SPSS


- Why is statistics important in agriculture?
- S.P.S.S.
 - Data collection
 - Data entrance
 - Data description
 - Summary statistics
 - Reliability analysis
 - Validity test
 - Qualitative research
 - Categorical data analysis

Why is statistics important in agriculture?

- ❑ **Data and numerical information have played a very important role in the growth and development of agriculture, especially in the developed countries.**
- ❑ **In agrarian countries, like Greece and Egypt, the utility of agricultural statistics is even more important, though it has not been utilized adequately so far.**
- ❑ **In academic institutions ... research activities and findings (papers) have to be justified.**

Why is statistics important in agriculture?

- **The agriculture of a place is the result of many forces (physical, social, cultural, economic, institutional, technological, political and psychological) interacting upon each other and, therefore, the growth, development and problems of agriculture cannot be solved by fractional and isolated approaches.**
- In overcoming these problems, a multidisciplinary approach is required and a large body of data is to be incorporated in any project of research.

- 
-
- **In order to understand the nature of agricultural statistics more fully, they may be classified into the following major categories:**



Categories

- Land utilization and irrigation
- Forestry.
- Agricultural production
- Agricultural prices
- Socioeconomics, demographics and structural issues of rural life
- Marketing and agricultural economics
- Health issues
- Weather and climate
- Forecasts



The key areas utilizing agriculture statistics are

- ❑ health policy
- ❑ food security
- ❑ food safety
- ❑ natural resource use
- ❑ renewable energy production
- ❑ environmental economics and climate change
- ❑ crisis management during diseases and natural disasters
- ❑ long-term viability and competitiveness of agri-business and the agri-value chain
- ❑ rural development
- ❑ International competitiveness in trade



Why S.P.S.S.?

- SPSS is used extensively in business, government and academia.
- It is a statistical analysis package that allows the in depth analysis of large amounts of data.
- It is very useful for discovering correlations between different variables.
- It can be used to make statistically valid forecasts for future events or results.



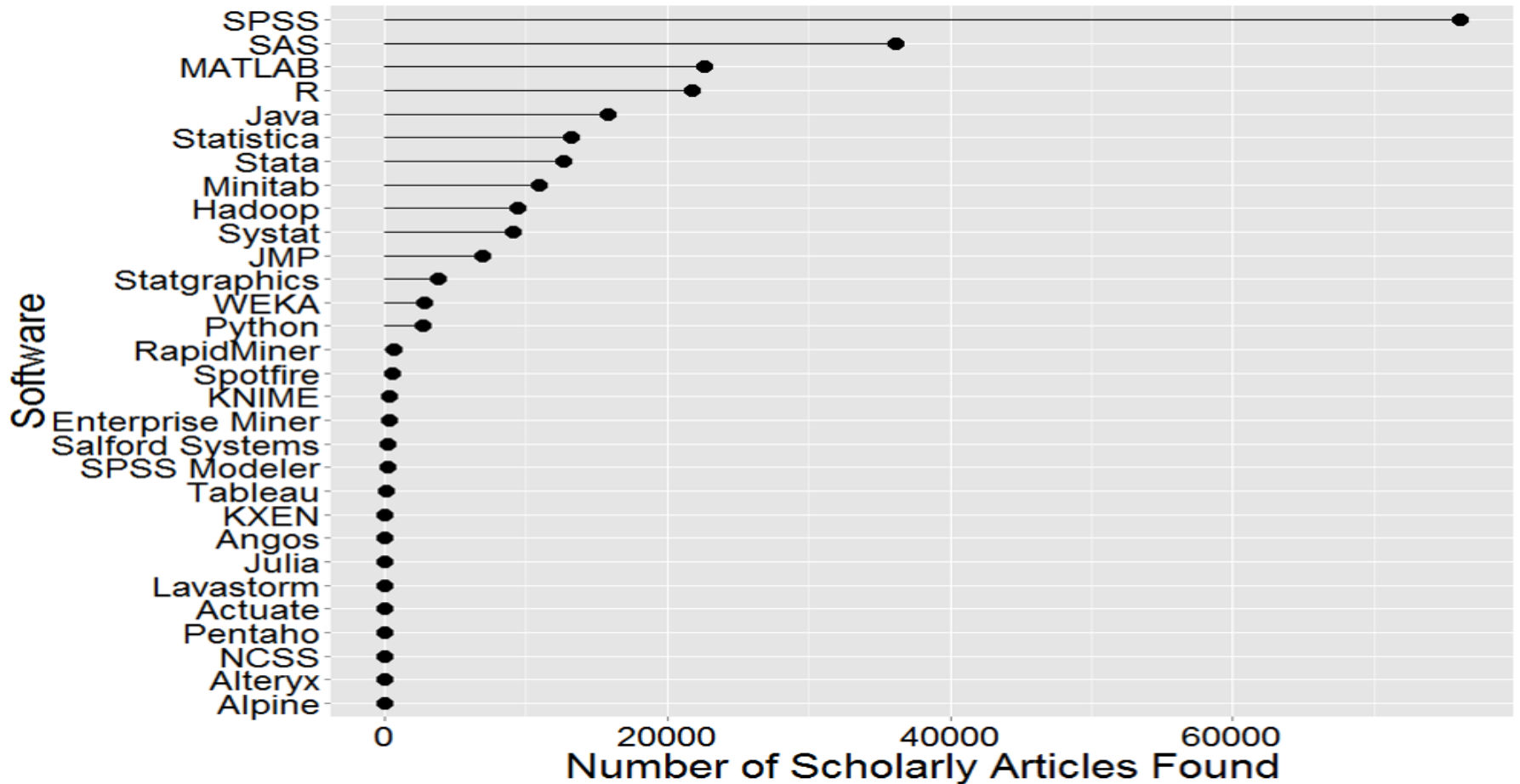
Benefits of SPSS

- Statistical analysis can be conducted using two main methods.
- One is simply by using a generalized spreadsheet or data management program such as MS Excel or through using a specialized statistical package such as SPSS.




Here are key reasons why SPSS is the best option to use

- Effective data management
- Wide range of options
- Better output organization
- Tables and figures
- Categorical data / Survey analysis
- No need for deep knowledge of statistics
- Publish a paper



SPSS led the pack with over 75,000 citations in scientific papers, which were culled through a search on Google Scholar.



S.P.S.S. is more suitable for
in depth data analysis.

Questionnaire

Sample questionnaire

Name: _____ Today's date: _____

Address: _____

City, state, zip: _____

Telephone: home (____) _____ - _____ Date of birth: _____

work (____) _____ - _____ Sex (*circle*): . Female Male

Background

1. Ethnic origin (*check only one*):

- White not Hispanic
- Black not Hispanic
- Hispanic

- Asian or Pacific Islander
- Filipino
- American Indian/Alaskan Native
- Other: _____



Designing questionnaires

- Start the questionnaire with a statement about the rationale behind the survey;
- knowing the context of the survey makes it easier for the respondents to provide meaningful answers



Statement of your rationale

For example:

- This is a survey of energy consumption within the domestic setting. It will gather data on the types of fuel-source used, the number of devices owned and the number of occasions each device is used within a specified timeframe. This data will be used to extrapolate peak-time energy demand.
- This questionnaire will gather data about public transport in Newark and surrounding villages. The data will be used to design new routes and services that will better meet the needs of the local community.
- This survey is looking at communication device interfaces and their suitability for use in the new communication technologies currently being developed. Your experience, either good or bad, of existing technology interfaces will help us to create better products in the future.

Demographic data

- It is quite common to begin with, questions about **demographic data**, statistics concerning human populations or a segments of the human population, broken down by age, gender, religion, ethnicity, marital status, income, post code, etc.
- Collect this data **if it is required**, but don't collect it for the sake of it. If you don't need it, don't ask for it.

3. Are you currently (*check only one*):

Married

Single

Separated

Divorced

Widowed



Sample size calculator

- <http://www.raosoft.com/samplesize.html>

Raosoft®

Sample size calculator

<p>What margin of error can you accept?</p> <p>5% is a common choice</p>	<input type="text" value="5"/> %	<p>The margin of error is the amount of error that you can tolerate. If 90% of respondents answer yes, while 10% answer no, you may be able to tolerate a larger amount of error than if the respondents are split 50-50 or 45-55.</p> <p>Lower margin of error requires a larger sample size.</p>
<p>What confidence level do you need?</p> <p>Typical choices are 90%, 95%, or 99%</p>	<input type="text" value="95"/> %	<p>The confidence level is the amount of uncertainty you can tolerate. Suppose that you have 20 yes-no questions in your survey. With a confidence level of 95%, you would expect that for one of the questions (1 in 20), the percentage of people who answer yes would be more than the margin of error away from the true answer. The true answer is the percentage you would get if you exhaustively interviewed everyone.</p> <p>Higher confidence level requires a larger sample size.</p>
<p>What is the population size?</p> <p>If you don't know, use 20000</p>	<input type="text" value="20000"/>	<p>How many people are there to choose your random sample from? The sample size doesn't change much for populations larger than 20,000.</p>
<p>What is the response distribution?</p> <p>Leave this as 50%</p>	<input type="text" value="50"/> %	<p>For each question, what do you expect the results will be? If the sample is skewed highly one way or the other, the population probably is, too. If you don't know, use 50%, which gives the largest sample size. See below under More information if this is confusing.</p>
<p>Your recommended sample size is</p>	<p>377</p>	<p>This is the minimum recommended size of your survey. If you create a sample of this many people and get responses from everyone, you're more likely to get a correct answer than you would from a large sample where only a small percentage of the sample responds to your survey.</p>

Online surveys with Vovici have completion rates of 66%!

Alternate scenarios

<p>With a sample size of</p>	<input type="text" value="100"/>	<input type="text" value="200"/>	<input type="text" value="300"/>	<p>With a confidence level of</p>	<input type="text" value="90"/>	<input type="text" value="95"/>	<input type="text" value="99"/>
<p>Your margin of error would be</p>	9.78%	6.89%	5.62%	<p>Your sample size would need to be</p>	267	377	643

Save effort, save time. Conduct your survey online with Vovici.

More information



Conduct the survey

- Once you have decided on your sample group you then need to decide on your data collection model.
- The main models are:
 - Personal Interviews
 - Telephone Surveys
 - Mail Surveys
 - Email surveys
 - Internet/web delivered surveys
- Each of these methods has advantages and disadvantages

Likert Scale

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
Scale Week is a worthwhile feature on The Research Bunker Blog.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
I would like to read more posts about survey rating scales.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Vance Marriner is, without a doubt, the most insightful contributor to The Research Bunker Blog.	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Strongly Disagree			Undecided			Strongly Agree
(1)	(2)	(3)	(4)	(5)	(6)	(7)

Strongly Disagree	Disagree	Moderately Disagree	Mildly Disagree	Undecided	Mildly Agree	Moderately Agree	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

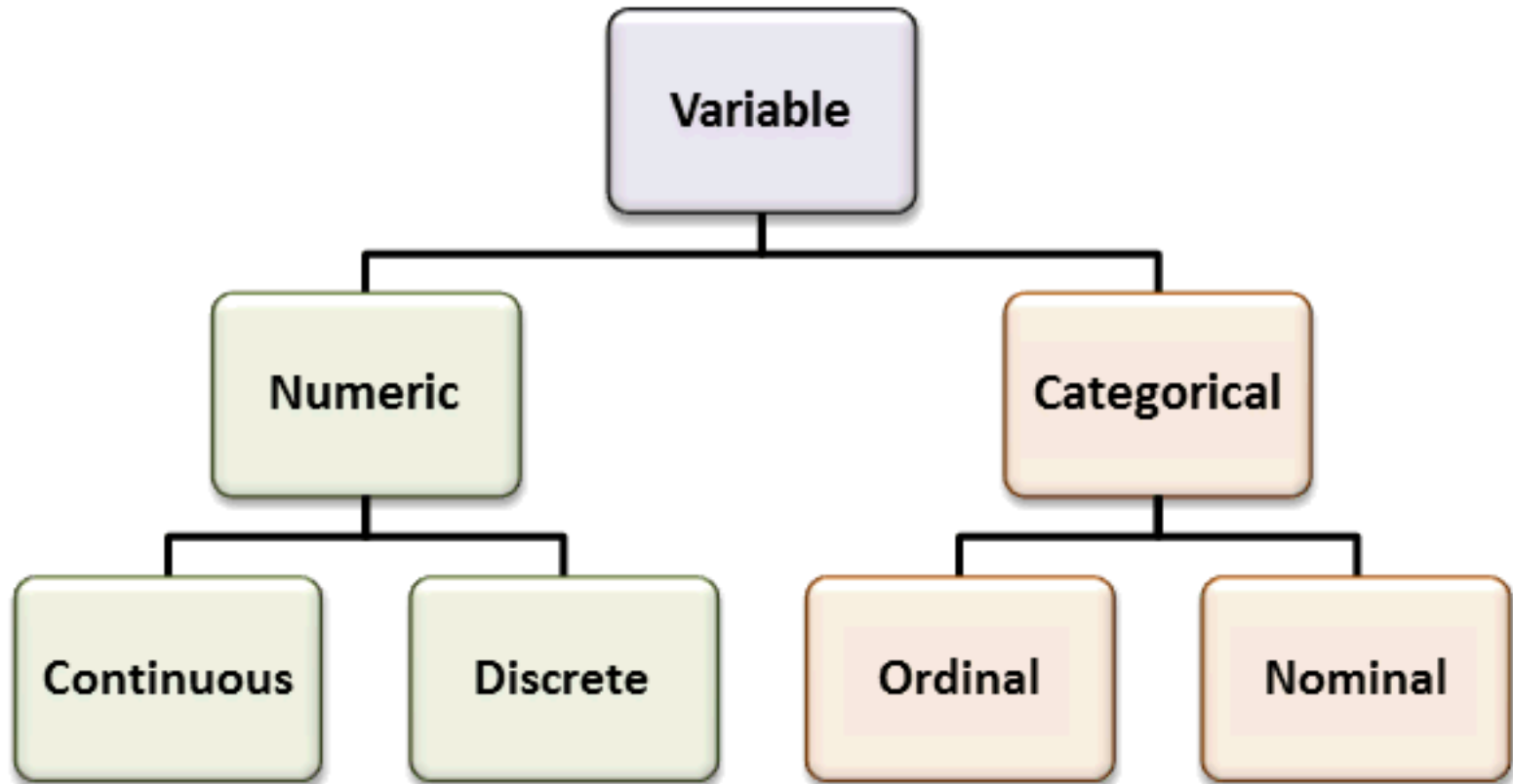
Likert type scale



	product A	product B
very satisfied	30	25
quite satisfied	30	20
neutral	20	40
quite dissatisfied	10	10
very dissatisfied	10	5
mean	3.6	3.5

A screenshot of a survey interface. On the left, a cartoon woman with blonde hair is gesturing. The question text reads: "How often during the day do you eat chocolate?". To the right of the question is a circular icon with a red arrow pointing to the right. Below the question are four orange buttons labeled "Hardly ever", "Sometimes", "Often", and "A lot". Below these buttons is a horizontal scale with four positions labeled 1, 2, 3, and 4. A red rectangular box highlights the positions 1 and 2.

Types of variables





Reliability and item analysis

- **Reliability refers to the extent to which a scale produces consistent results, if the measurements are repeated a number of times.**
- The analysis on reliability is called reliability analysis.
- Reliability analysis is determined by obtaining the proportion of systematic variation in a scale, which can be done by determining the association between the scores obtained from different administrations of the scale.
- Thus, if the association in reliability analysis is high, the scale yields consistent results and is therefore reliable.

Cronbach's Alpha (α) using SPSS Statistics

- Cronbach's alpha is the most common measure of internal consistency ("reliability").
- It is most commonly used when you have multiple Likert questions in a survey/questionnaire that form a scale and you wish to determine if the scale is reliable.



Example

- A researcher has devised a nine-question questionnaire to measure how safe people feel at work at an industrial complex.
- Each question was a 5-point Likert item from "strongly disagree" to "strongly agree".
- In order to understand whether the questions in this questionnaire all reliably measure the same latent variable (feeling of safety) (so a Likert scale could be constructed), a Cronbach's alpha was run on a sample size of 15 workers.



Procedure via SPSS

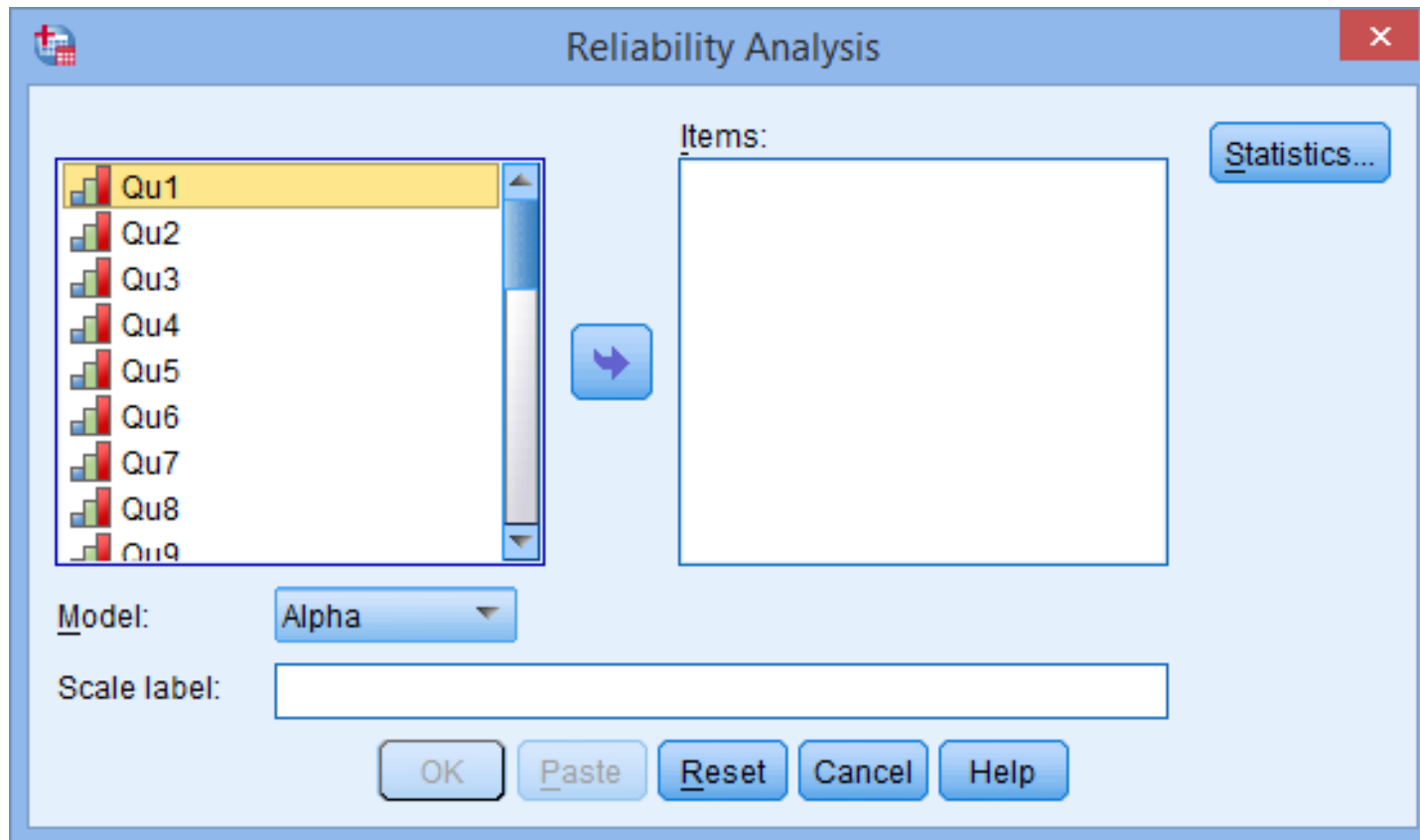
- In SPSS Statistics, the nine questions have been labelled Qu1 through to Qu9.
- The eight steps below show you how to check for internal consistency using Cronbach's alpha in SPSS Statistics.
- At the end of these eight steps, we show you how to interpret the results from your Cronbach's alpha.

Click Analyze > Scale > Reliability Analysis...
on the top menu, as shown below:

The screenshot shows the IBM SPSS Statistics Data Editor interface. The title bar reads "Cronbach's alpha.sav [DataSet5] - IBM SPSS Statistics Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. The Analyze menu is open, showing a list of options: Reports, Descriptive Statistics, Compare Means, General Linear Model, Generalized Linear Models, Mixed Models, Correlate, Regression, Loglinear, Classify, Dimension Reduction, Scale, Nonparametric Tests, Forecasting, Survival, Multiple Response, Simulation..., Quality Control, and ROC Curve... The Scale option is highlighted in yellow, and its sub-menu is open, showing Reliability Analysis... and Multidimensional Scaling (ALSCAL)... The Reliability Analysis... option is also highlighted in yellow, with a mouse cursor pointing to it. The data grid shows variables Qu1 through Qu8 and rows 1 through 18. The data values are as follows:

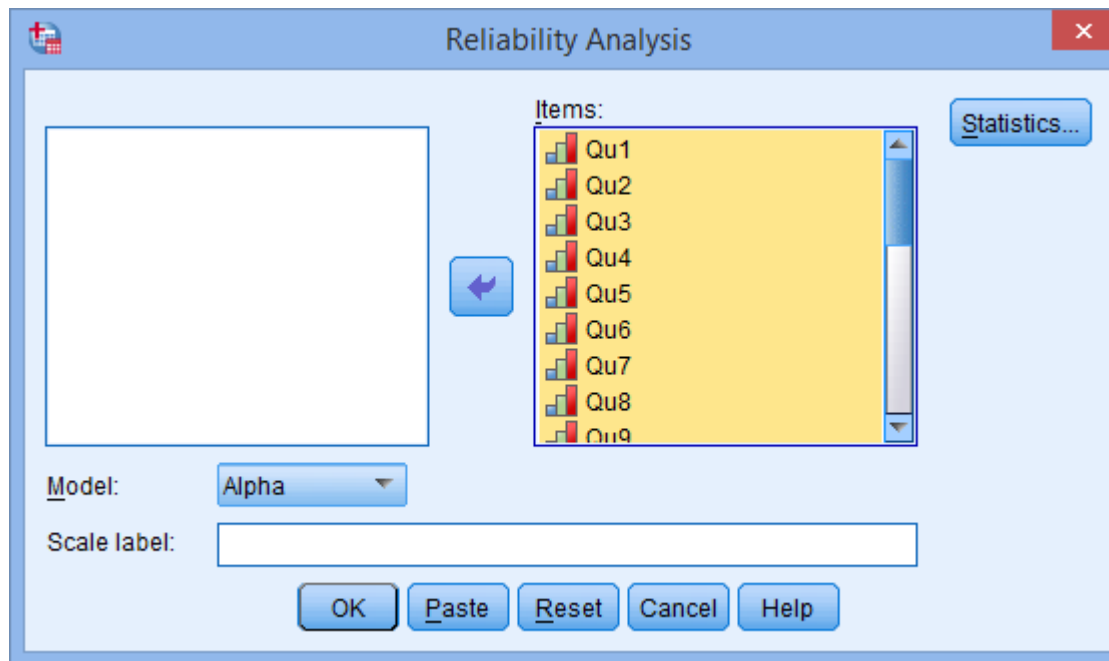
	Qu1	Qu2	Qu3	Qu4	Qu5	Qu6	Qu7	Qu8
1	3				4	4	4	4
2	6				2	3	3	3
3	5				3	3	2	3
4	4				5	6	5	6
5	4				5	6	5	6
6	6				3	3	2	3
7	6				2	3	4	4
8	5				3	4	4	5
9	6				3	4	4	3
10	6				3	2	3	2
11	5				3	3	2	3
12	6				2	3	4	4
13	6				3	4	4	5
14	6				4	5	4	4
15	5	6	3	3	3	3	2	3
16	4	6	3	3	2	3	4	4
17	6	6	3	3	3	3	2	3
18	6	3	2	3	2	3	4	4

You will be presented with the **Reliability Analysis** dialogue box, as shown below:

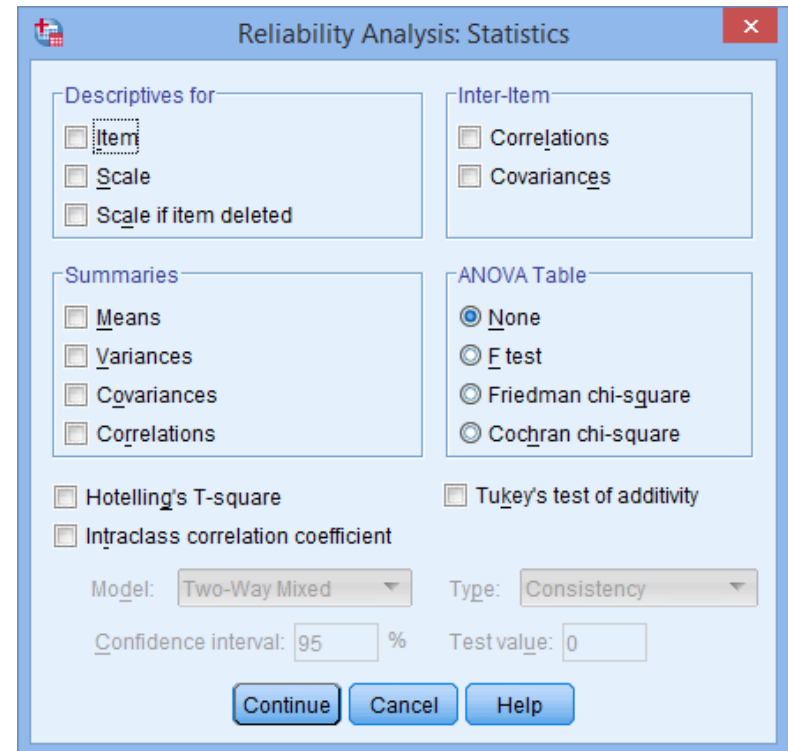


Transfer the variables Qu1 to Qu9 into the Items: box. You can do this by drag-and-dropping the variables into their respective boxes or by using the button.

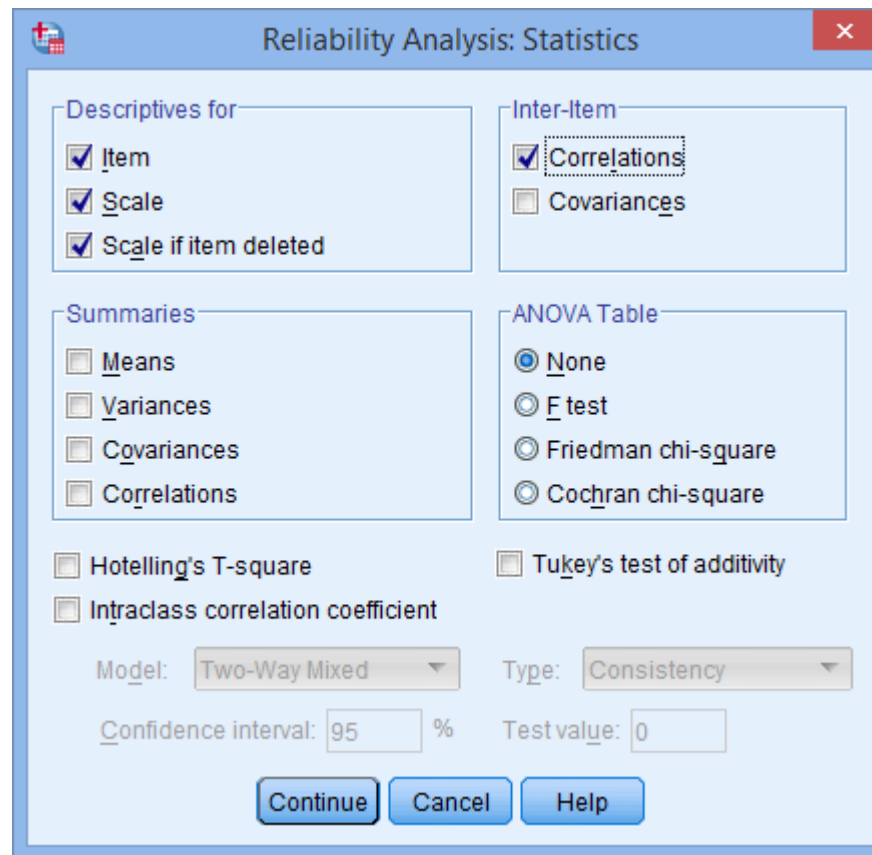
You will be presented with the following screen:



- Leave the **Model**: set as "Alpha", which represents Cronbach's alpha in SPSS Statistics.
- Click on the **Statistics...** , which will open the **Reliability Analysis: Statistics** dialog box, as shown below:



Select the Item, Scale and Scale if item deleted options in the –Descriptives for– area, and the Correlations option in the – Inter-Item– area, as shown below:





- Click the  button. This will return you to the **Reliability Analysis** dialogue box.

- Click the  button to generate the output.

SPSS Statistics Output for Cronbach's Alpha

- SPSS Statistics produces many different tables. The first important table is the **Reliability Statistics** table that provides the actual value for **Cronbach's alpha**, as shown below:

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.805	.796	9

From our example, we can see that **Cronbach's alpha is 0.805**, which indicates a **high level of internal consistency for our scale with this specific sample**

Item-Total Statistics

- The **Item-Total Statistics** table presents the "**Cronbach's Alpha if Item Deleted**" in the final column, as shown below:

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Qu1	24.20	45.029	.633	.588	.767
Qu2	23.93	47.352	.520	.651	.783
Qu3	24.07	46.638	.654	.899	.767
Qu4	23.40	47.114	.551	.823	.779
Qu5	23.60	51.257	.389	.573	.799
Qu6	24.47	50.695	.372	.693	.802
Qu7	24.07	45.210	.615	.777	.770
Qu8	24.20	56.457	.128	.791	.823
Qu9	24.07	45.210	.589	.610	.774



Validity test

- **Validity** refers to how well a test measures what it is purported to measure.
- While reliability is necessary, it alone is not sufficient. For a test to be reliable, it also needs to be valid. For example, if your scale is off by 1 kgr, it reads your weight every day with an excess of 1 kgr. The scale is reliable because it consistently reports the same weight every day, but it is not valid because it adds 1 kgr to your true weight. It is not a valid measure of your weight.



What are some ways to improve validity?

- ❑ Make sure your goals and objectives are clearly defined and operationalized.
- ❑ Match your assessment measure to your goals and objectives. Additionally, have the test reviewed by a group of experts to obtain feedback from an outside party who is less invested in the instrument.
- ❑ Get experts involved; have the experts look over the assessment for troublesome wording, or other difficulties.
- ❑ If possible, compare your measure with other measures, or data that may be available.



Statistical hypotheses

- A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.
- A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis.
- The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of "no effect" or "no difference".
- The alternative hypothesis is the statement you want to be able to conclude is true.



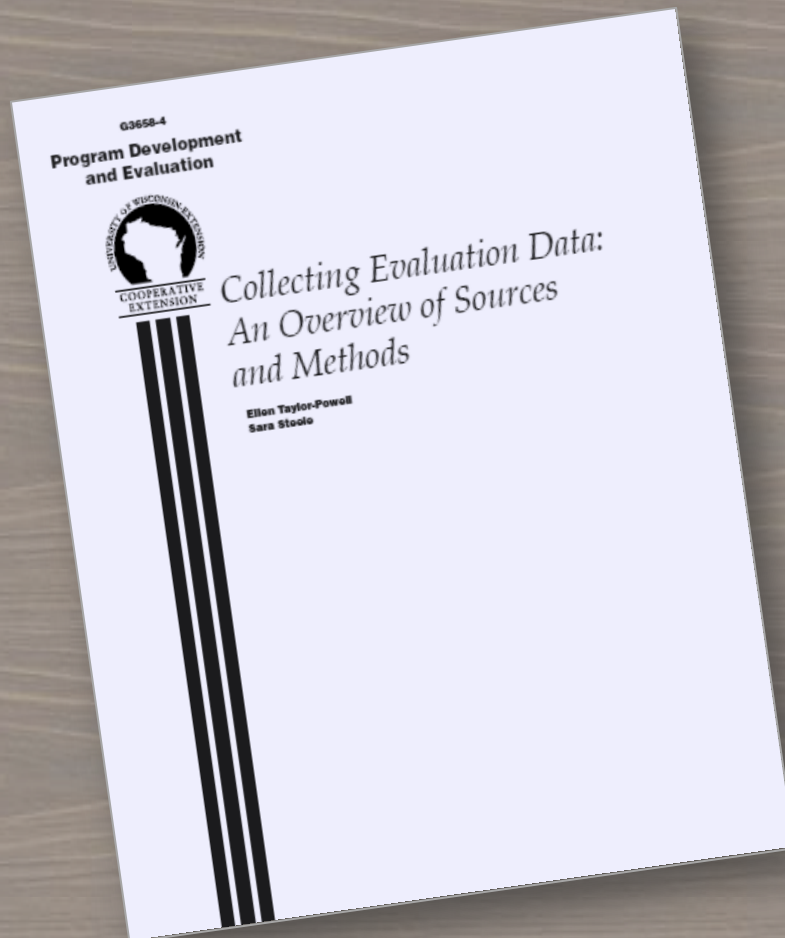
p-value

- Based on the sample data, the test determines whether to reject the null hypothesis. You use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point that you define, then you can reject the null hypothesis.
- A common misconception is that statistical hypothesis tests are designed to select the more likely of two hypotheses. Instead, a test will remain with the null hypothesis until there is enough evidence (data) to support the alternative hypothesis.



Examples

- Examples of questions you can answer with a hypothesis test include:
 - Do male and female differ in their level of agreement?
 - Do educational level differ in terms of income
 - Is there any statistical relation between age and innovative behavior?



Collecting data involves:

- 1) **Sources** - where you will get the information; and
- 2) **Methods** - how you will collect /gather the information

Use this booklet for help

<http://learningstore.uwex.edu/pdf/G3658-04.pdf>

Thank you !!!

tassosm@auth.gr



<http://rural-lab.agro.auth.gr/ilham-ec1.pdf>