



Using SPSS in Healthcare Settings

Jarlath Quinn – Analytics Consultant

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY



Just waiting for all attendees to join...

Using SPSS in Healthcare Settings

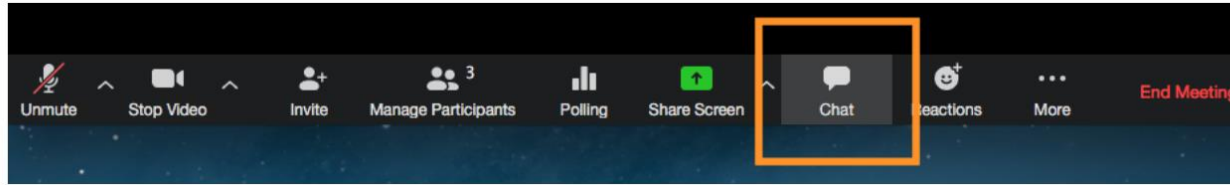
Jarlath Quinn – Analytics Consultant

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat panel – if we run out of time we will follow up with you.





- Gold accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open-source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry
- Deep experience of applied advanced analytics applications across sectors
 - Retail
 - Healthcare/Pharma
 - Finance/Insurance
 - Media/Telecoms
 - Utilities
 - FMCG
 - Charity/Housing/Government



Agenda

- Working with patient satisfaction data
- Interpreting Correlations
- Working with Decision Trees
- Estimating risk with Odds Ratios and Relative Risk Scores
- Exploring Survival Analysis



Let's take a look....



Correlations

		Current Salary
Educational Level (years)	Pearson Correlation	.661**
	Sig. (2-tailed)	<.001
	N	474

** . Correlation is significant at the 0.01 level (2-tailed).

Interpreting Correlations

Linear Correlation Scale

+1

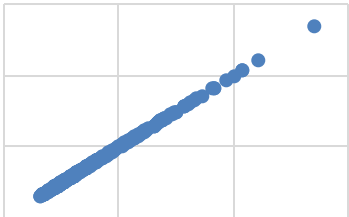
+0.5

0

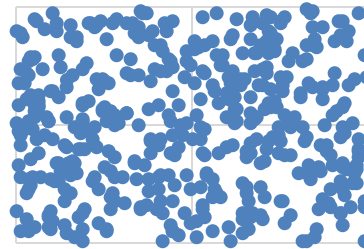
-0.5

-1

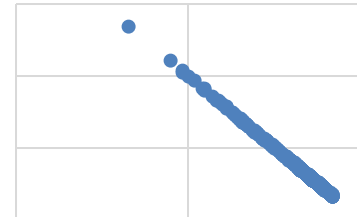
Perfect Positive Linear
Relationship



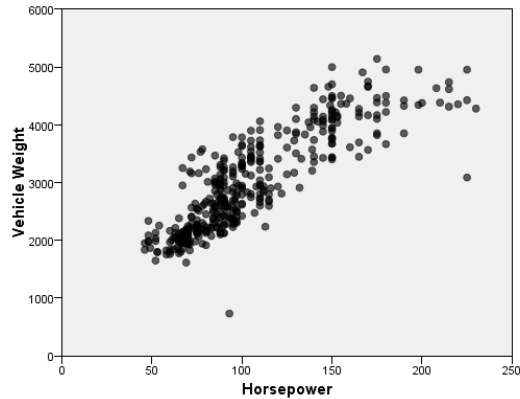
No Linear Relationship



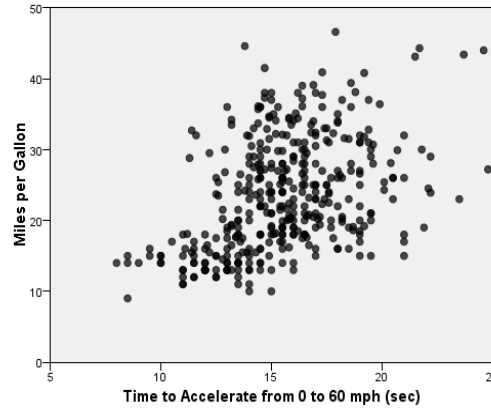
Perfect Negative
Linear Relationship



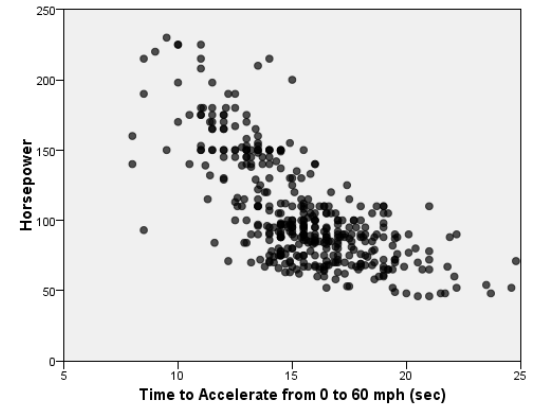
Pearson's r correlations



0.859

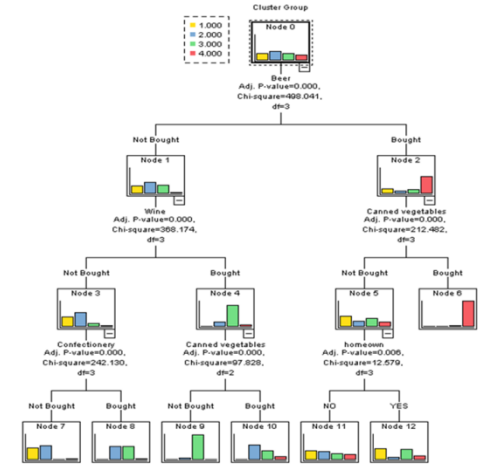


0.434



-.701

Pearson's r correlation coefficients



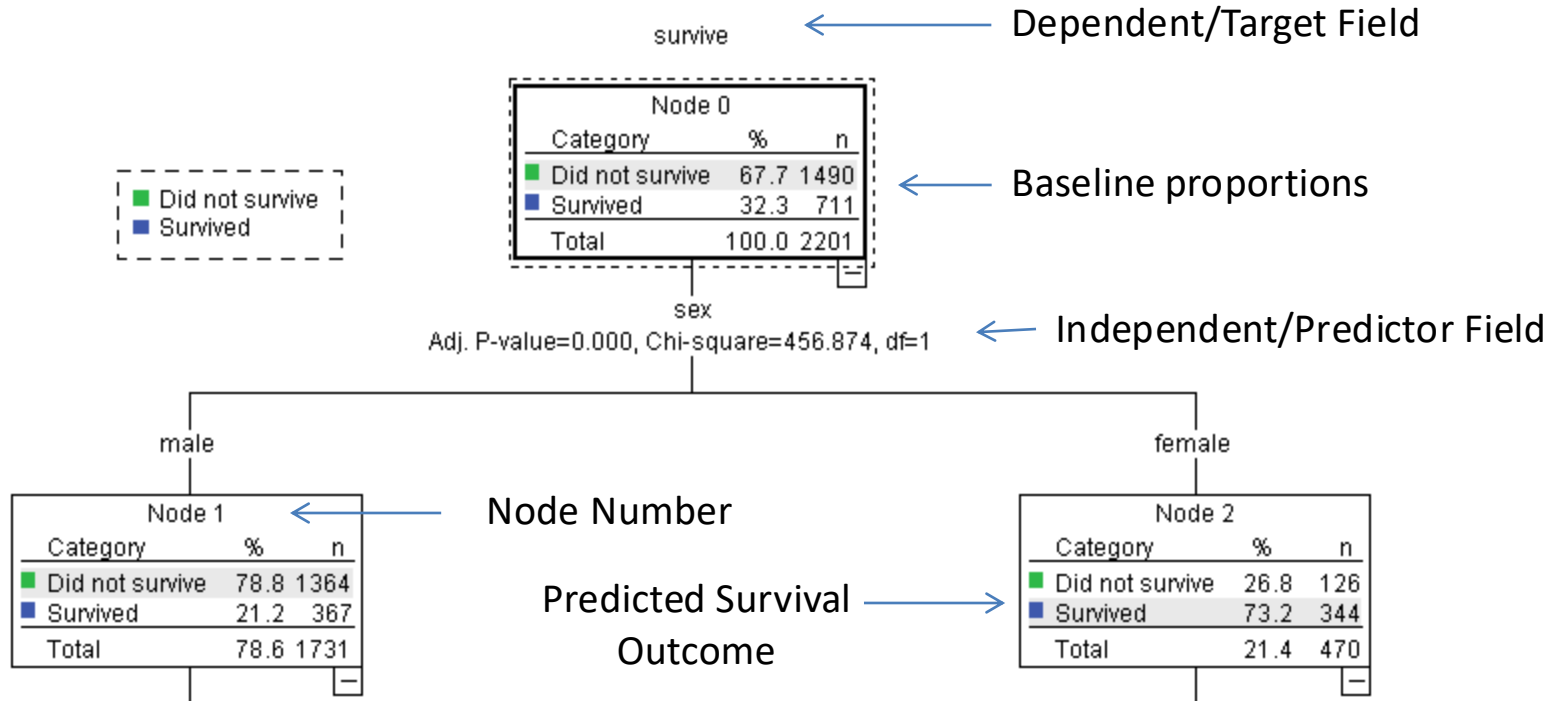
Working with Decision Trees

Why use Decision Trees?

- Decision trees can be used to
 - Build profiles of patients/staff/research subjects
 - Find key behavioural segments
 - Generate predictive models
- Decision Trees are especially popular because
 - they are fairly visual representations of models
 - relatively easy to understand

Decision Trees split targeted outcomes by key factors

...in this case the key factor is the sex of passenger



A Decision Tree based on the CHAID algorithm

C.H.A.I.D
Chi-Square
Automatic
Interaction
Detector

■ Did not survive
■ Survived

survive

Node 0		
Category	%	n
■ Did not survive	67.7	1490
■ Survived	32.3	711
Total	100.0	2201

sex

Adj. P-value=0.000, Chi-square=466.874, df=1

male

Node 1		
Category	%	n
■ Did not survive	78.8	1364
■ Survived	21.2	367
Total	78.6	1731

female

Node 2		
Category	%	n
■ Did not survive	26.8	126
■ Survived	73.2	344
Total	21.4	470

age

Adj. P-value=0.000, Chi-square=23.125, df=1

adult

Node 3		
Category	%	n
■ Did not survive	79.7	1329
■ Survived	20.3	338
Total	75.7	1667

child

Node 4		
Category	%	n
■ Did not survive	54.7	35
■ Survived	45.3	29
Total	2.9	64

1st

Node 5		
Category	%	n
■ Did not survive	2.8	4
■ Survived	97.2	141
Total	6.6	145

2nd; crew

Node 6		
Category	%	n
■ Did not survive	12.4	16
■ Survived	87.6	113
Total	5.9	129

3rd

Node 7		
Category	%	n
■ Did not survive	54.1	106
■ Survived	45.9	90
Total	8.9	196

class

Adj. P-value=0.000, Chi-square=37.988, df=3

1st

Node 8		
Category	%	n
■ Did not survive	67.4	118
■ Survived	32.6	57
Total	8.0	175

2nd

Node 9		
Category	%	n
■ Did not survive	91.7	154
■ Survived	8.3	14
Total	7.6	168

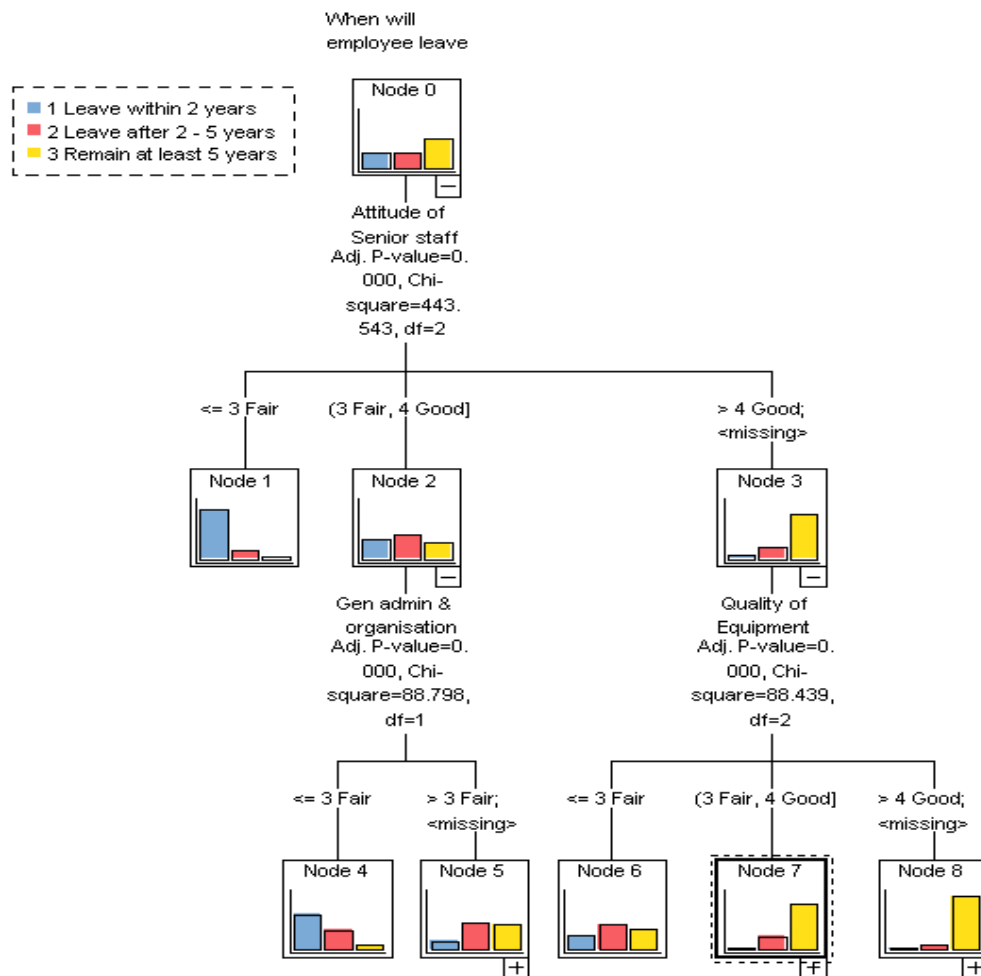
3rd

Node 10		
Category	%	n
■ Did not survive	83.8	387
■ Survived	16.2	75
Total	21.0	462

crew

Node 11		
Category	%	n
■ Did not survive	77.7	670
■ Survived	22.3	192
Total	39.2	862

We can also use it to model satisfaction





Smoker * History of Angina Crosstabulation

% within Smoker

		History of Angina		Total
		Yes	No	
Smoker	Yes	68.2%	31.8%	100.0%
	No	44.4%	55.6%	100.0%
Total		49.8%	50.2%	100.0%

Estimating risk with Odds Ratios and Relative Risk Scores

Measures of Risk

- **Measures of risk** play a crucial role in healthcare for several reasons:
 - Assessing treatment effects – measuring impact of treatment
 - Interpreting clinical studies – comparing treatment effects
 - Public health decision-making – e.g. evaluating vaccination programs
 - Risk communication – often a simpler way to communicate complexity
 - Identifying high-risk populations
- There are two key measures of risk in SPSS:
 - **Odds Ratio**
 - **Relative Risk Ratio**

Odds Ratios and Relative Risk Estimates

- These methods are both **Measures of Association**
- Rather than just looking at whether a relationship is 'statistically significant' we wish to measure the how strongly they are related
- In healthcare applications, these values can indicate if exposure to a factor can increase risk of an outcome or conversely, have a protective effect
- We can also use, confidence intervals to show if the effect is statistically significant
- One type of method is based on calculating *odds* and the other is based on *probabilities*

Odds vs Probability

- The **Probability** of getting a 6 is $1/6$
- So on 16.7% of occasions you will roll a 6 ($P = 0.167$)
- But the **Odds** of getting a 6 are $1/5$
- On average, for every 6 rolls, 5 of them will be a number *other than* 6
- So Odds are calculated differently from Probability



Odds Ratios vs Relative Risk Estimates

- **Odds Ratio**

- Ratio of the *odds* – i.e. the ratio of the odds of A in the presence of B and the odds of A in the absence of B
- Regarded as more difficult to interpret than relative risk

- **Relative Risk Ratio**

- Ratio of the *probabilities* – i.e. the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group

Relative Risk and Odds Ratio in SPSS

- In this example, we will look at the associated risk between smoking and angina
- Here the **risk factor** is smoking status and angina the **outcome**
- We can use the SPSS Crosstabs procedure to compute both the Relative Risk Factor and the Odds Ratio

What's the association between smoking on angina?

Smoker * History of Angina Crosstabulation

			History of Angina		Total
			Yes	No	
Smoker	Yes	Count	1552	723	2275
		% within Smoker	68.2%	31.8%	100.0%
	No	Count	3427	4298	7725
		% within Smoker	44.4%	55.6%	100.0%
Total		Count	4979	5021	10000
		% within Smoker	49.8%	50.2%	100.0%

Chi Squared - $P < .001$

What's the association between smoking on angina?

Outcome

Smoker * History of Angina Crosstabulation

			History of Angina		Total	
			Yes	No		
Risk Factor	Smoker	Yes	Count	1552	723	2275
			% within Smoker	68.2%	31.8%	100.0%
		No	Count	3427	4298	7725
			% within Smoker	44.4%	55.6%	100.0%
	Total		Count	4979	5021	10000
			% within Smoker	49.8%	50.2%	100.0%

Note the row and column positions of the two variables.
This usually makes it easier to interpret the results.

What's the association between smoking on angina?

Smoker * History of Angina Crosstabulation

			History of Angina		Total
			Yes	No	
Smoker	Yes	Count	1552	723	2275
		% within Smoker	68.2%	31.8%	100.0%
	No	Count	3427	4298	7725
		% within Smoker	44.4%	55.6%	100.0%
Total		Count	4979	5021	10000
		% within Smoker	49.8%	50.2%	100.0%

Note the row and column positions of the risk and outcome **categories**. This affects the calculation. *

Calculating the Odds Ratio

Smoker * History of Angina Crosstabulation

			History of Angina		Total
			Yes	No	
Smoker	Yes	Count	1552	723	2275
		% within Smoker	68.2%	31.8%	100.0%
	No	Count	3427	4298	7725
		% within Smoker	44.4%	55.6%	100.0%
Total		Count	4979	5021	10000
		% within Smoker	49.8%	50.2%	100.0%

$$1552 / 723 = 2.147$$

$$2.147 / 0.797 = \mathbf{2.69}$$

$$3427 / 4298 = 0.797$$

- The **odds** of having angina are 2.69 times greater for smokers than for non-smokers
- This is a measure of association, it is *not* a causal statement

Calculating the Relative Risk Ratio

Smoker * History of Angina Crosstabulation

			History of Angina		Total
			Yes	No	
Smoker	Yes	Count	1552	723	2275
		% within Smoker	68.2%	31.8%	100.0%
	No	Count	3427	4298	7725
		% within Smoker	44.4%	55.6%	100.0%
Total		Count	4979	5021	10000
		% within Smoker	49.8%	50.2%	100.0%

$$1552 / 2275 = 0.6822$$

$$0.6822 / 0.4436 = \mathbf{1.538}$$

$$3427 / 7725 = 0.4436$$

- The **probability** of having angina is 1.54 times greater for smokers than for non-smokers
- Whatever that **baseline** probability is.....



The Results in SPSS

Smoker * History of Angina Crosstabulation

			History of Angina		Total
			Yes	No	
Smoker	Yes	Count	1552	723	2275
		% within Smoker	68.2%	31.8%	100.0%
	No	Count	3427	4298	7725
		% within Smoker	44.4%	55.6%	100.0%
Total		Count	4979	5021	10000
		% within Smoker	49.8%	50.2%	100.0%

The Results in SPSS

- Although the Odds Ratio and Risk Ratio values are quite different here, generally speaking these two numbers start to become much more similar to each other when the proportion of people with the condition occurs relatively rarely (e.g. below 10%)

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Smoker (Yes / No)	2.692	2.438	2.972
For cohort History of Angina = Yes	1.538	1.481	1.597
For cohort History of Angina = No	.571	.536	.609
N of Valid Cases	10000		

The Results in SPSS

What about these values?

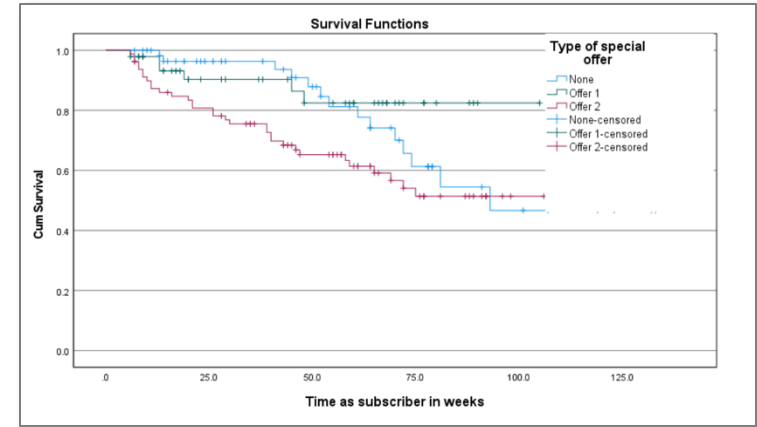
Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Smoker (Yes / No)	2.692	2.438	2.972
For cohort History of Angina = Yes	1.538	1.481	1.597
For cohort History of Angina = No	.571	.536	.609
N of Valid Cases	10000		

The Results in SPSS

- The 95% confidence intervals provide an indication as to how much these values are likely to vary from one sample to the next
- For both the Odds Ratio and the Relative Risk Estimate both of the intervals are positive so it's likely that this *increased risk effect* exists in the population

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Smoker (Yes / No)	2.692	2.438	2.972
For cohort History of Angina = Yes	1.538	1.481	1.597
For cohort History of Angina = No	.571	.536	.609
N of Valid Cases	10000		



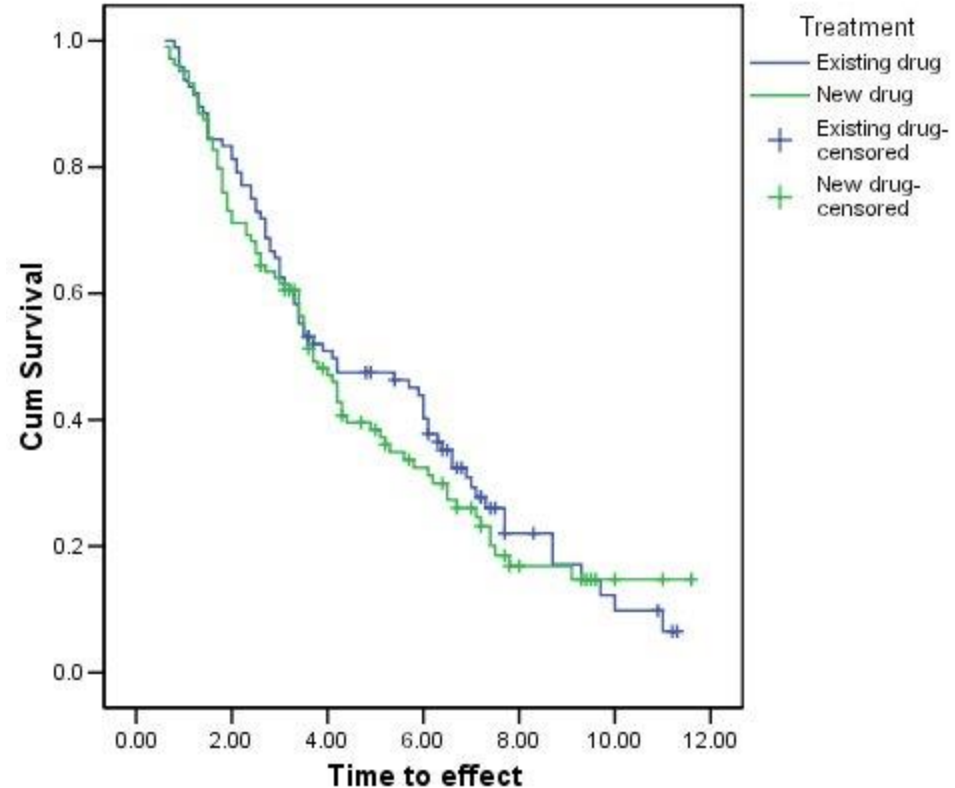
Performing Survival Analysis

Introducing Survival Analysis

- Survival analysis refers to a family of statistical procedures where the outcome variable of interest is *time until an event occurs*.
- It is commonly employed where researchers are interested in the modelling the effects of different treatments or conditions upon patient survival time.
- For this reason, survival analysis is a key technique used to analyse the efficacy of pharmaceuticals in studies of disease prognosis.

Introducing Survival Analysis

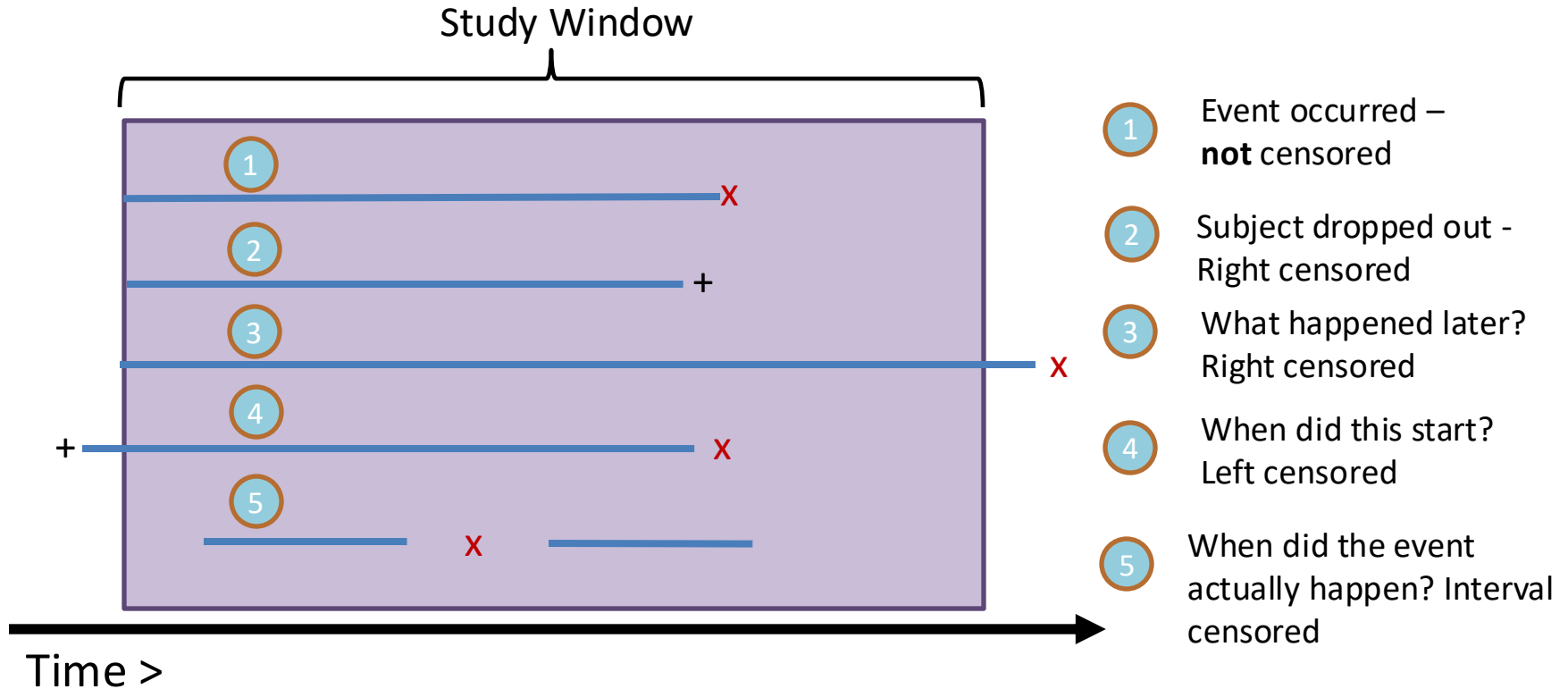
- Survival Analysis uses **Time-To-Event (TTE)** data
- These data are comprised of two key variables:
 1. An outcome/status variable indicating if the event has occurred yet
 2. A variable showing how much elapsed time has occurred before the event occurred or did not occur



Censored Data

- An often unavoidable issue with time-to-event data is that data points are often '**censored**'
- This refers to unknown circumstances where we, for example, don't know how long the patient survived or how much time elapsed before the event occurred
- It's an added complication of the analysis but one that is incorporated into the various calculations that estimate survival time

Censored Data in Research Studies



Key Output Types in Survival Analysis

1. Cumulative Survival Tables
2. Group Comparison Tables/Tests
3. Survival Plots

1

Survival Table						
	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	6.000	Dropped	.	.	1	188
2	6.000	Dropped	.989	.007	2	187
3	6.000	Current	.	.	3	188
4	7.000	Dropped	.	.	4	187
5	7.000	Dropped	.979	.	5	186
6	7.000	Current	.	.	6	185
7	7.000	Current	.	.	7	184
8	7.000	Current	.	.	8	183
9	8.000	Dropped	.	.	9	182
10	8.000	Dropped	.968	.	10	181
11	8.000	Current	.	.	11	180
12	8.000	Current	.	.	12	179
13	9.000	Dropped	.	.	13	178
14	9.000	Dropped	.957	.	14	177
15	9.000	Current	.	.	15	176
16	9.000	Current	.	.	16	175
17	9.000	Current	.	.	17	174

2

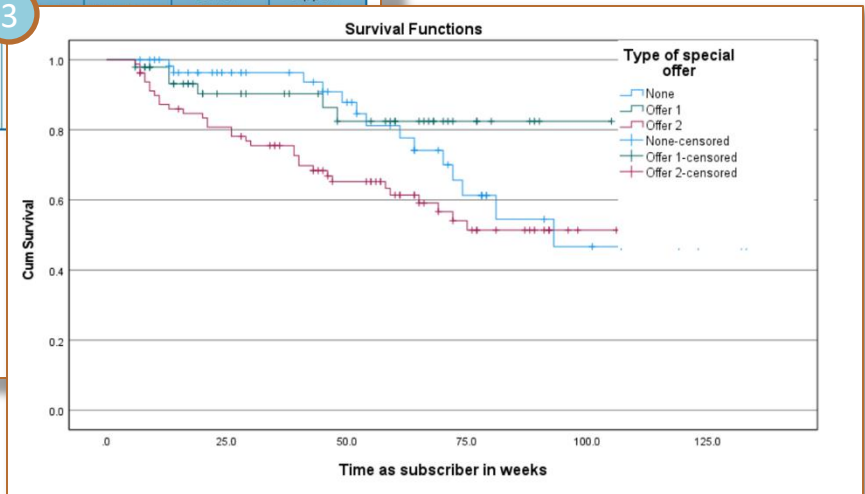
Means and Medians for Survival Time						
Type of special offer	Mean ^a				Median	
	Estimate	Std. Error	95% Confidence Interval			95% Confidence Interval
			Lower Bound	Upper Bound		
None	96.389	7.315	82.052	110.727		
Offer 1	91.521	5.068	81.588	101.455		
Offer 2	80.583	5.678	69.455	91.711		
Overall	94.302	4.339	85.797	102.806		

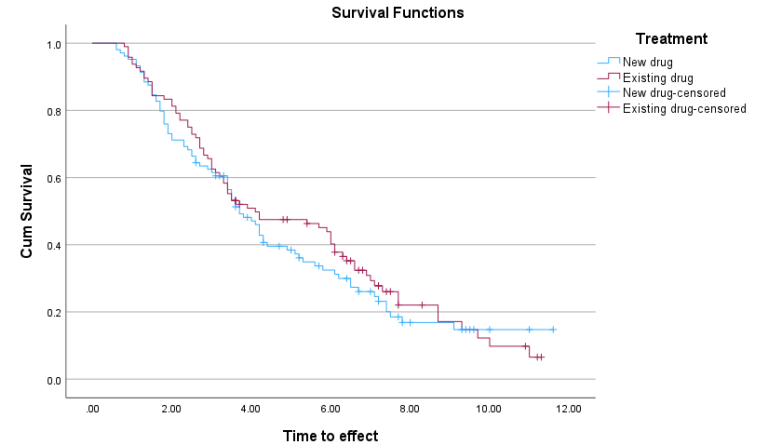
a. Estimation is limited to the largest survival time if it is censored.

Overall Comparisons			
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6.563	2	.038
Breslow (Generalized Wilcoxon)	10.151	2	.006
Tarone-Ware	8.730	2	.013

Test of equality of survival distributions for the different levels of Type of special offer.

3



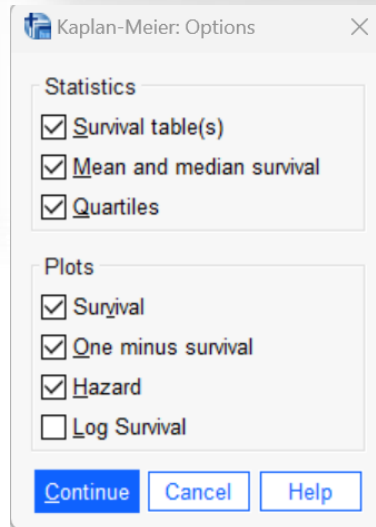
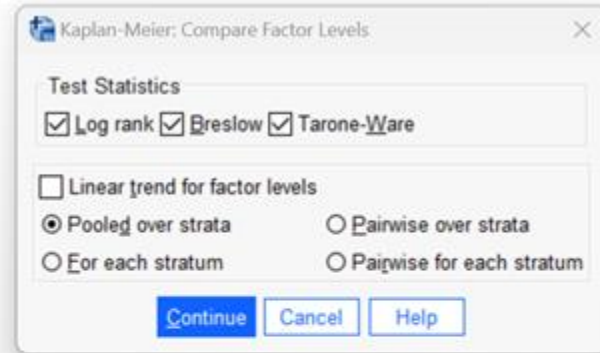


Kaplan Meier Survival Analysis

Kaplan-Meier Example

- Sample dataset containing 200 rows of data
- Each row represents a patient receiving anti-inflammatory medication for treating chronic arthritic pain
- The time-to-event variable is the time the medication takes to have an effect (if at all)
- The event variable denotes whether the effect occurred or the data were censored
- A key additional variable measures whether the treatment was administered with a new drug or a standard medication
- The sample file **pain_medication.sav** is included in the samples folder where SPSS Statistics is installed

Kaplan Meier - Example



Survival Table

Case Processing Summary

Treatment	Total N	N of Events	Censored	
			N	Percent
New drug	104	79	25	24.0%
Existing drug	96	74	22	22.9%
Overall	200	153	47	23.5%

Survival Table

Treatment		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
New drug	1	.600	Taken effect	.	.	1	103
	2	.600	Taken effect	.981	.013	2	102
	3	.700	Taken effect	.971	.016	3	101
	4	.800	Taken effect	.962	.019	4	100
	5	.900	Taken effect	.952	.021	5	99

Survival Table

Survival Table								
1		2	3	4 Cumulative Proportion Survival at the Time		5	6	7
Treatment		Time	Status	Estimate	Std. Error	N of Cumulative Events	N of Remaining Cases	
New drug	1	.600	Taken effect	.	.	1	103	
	2	.600	Taken effect	.981	.013	2	102	
	3	.700	Taken effect	.971	.016	3	101	
	4	.800	Taken effect	.962	.019	4	100	
	5	.900	Taken effect	.952	.021	5	99	

1. Treatment – New Drug or Existing Drug
2. Time to effect
3. Status (note: censored cases are not included in calculations)
4. Cumulative survival i.e. proportion of non-censored cases where effect has not yet occurred
5. Standard Error (an estimate of variance) of cumulative survival
6. Cumulative count of events thus far
7. Number of non-censored cases remaining where effect has not yet occurred

Other Output

Means and Medians for Survival Time

Treatment	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
New drug	4.867	.360	4.162	5.572	3.700	.292	3.128	4.272
Existing drug	5.185	.350	4.499	5.871	4.100	1.131	1.884	6.316
Overall	5.014	.252	4.520	5.507	3.900	.272	3.367	4.433

a. Estimation is limited to the largest survival time if it is censored.

Percentiles

Treatment	25.0%		50.0%		75.0%	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
New drug	7.100	.509	3.700	.292	1.900	.226
Existing drug	7.700	.648	4.100	1.131	2.400	.247
Overall	7.300	.371	3.900	.272	2.100	.196

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.379	1	.538
Breslow (Generalized Wilcoxon)	.748	1	.387
Tarone-Ware	.705	1	.401

Test of equality of survival distributions for the different levels of Treatment.

Three tests for comparing factor levels

All three of these tests are designed to compare the equality of survival distributions for different groups (levels).

- **Log rank** - All time points are weighted equally in this test.
- **Breslow** - Time points are weighted by the number of cases at risk at each time point. Less power when percentage of censored cases is large. Early events weighted more heavily than later events.
- **Tarone-Ware** - Time points are weighted by the square root of the number of cases at risk at each time point. It is more sensitive when the survival functions do not differ by a constant factor. Regarded as a compromise between the previous two tests.

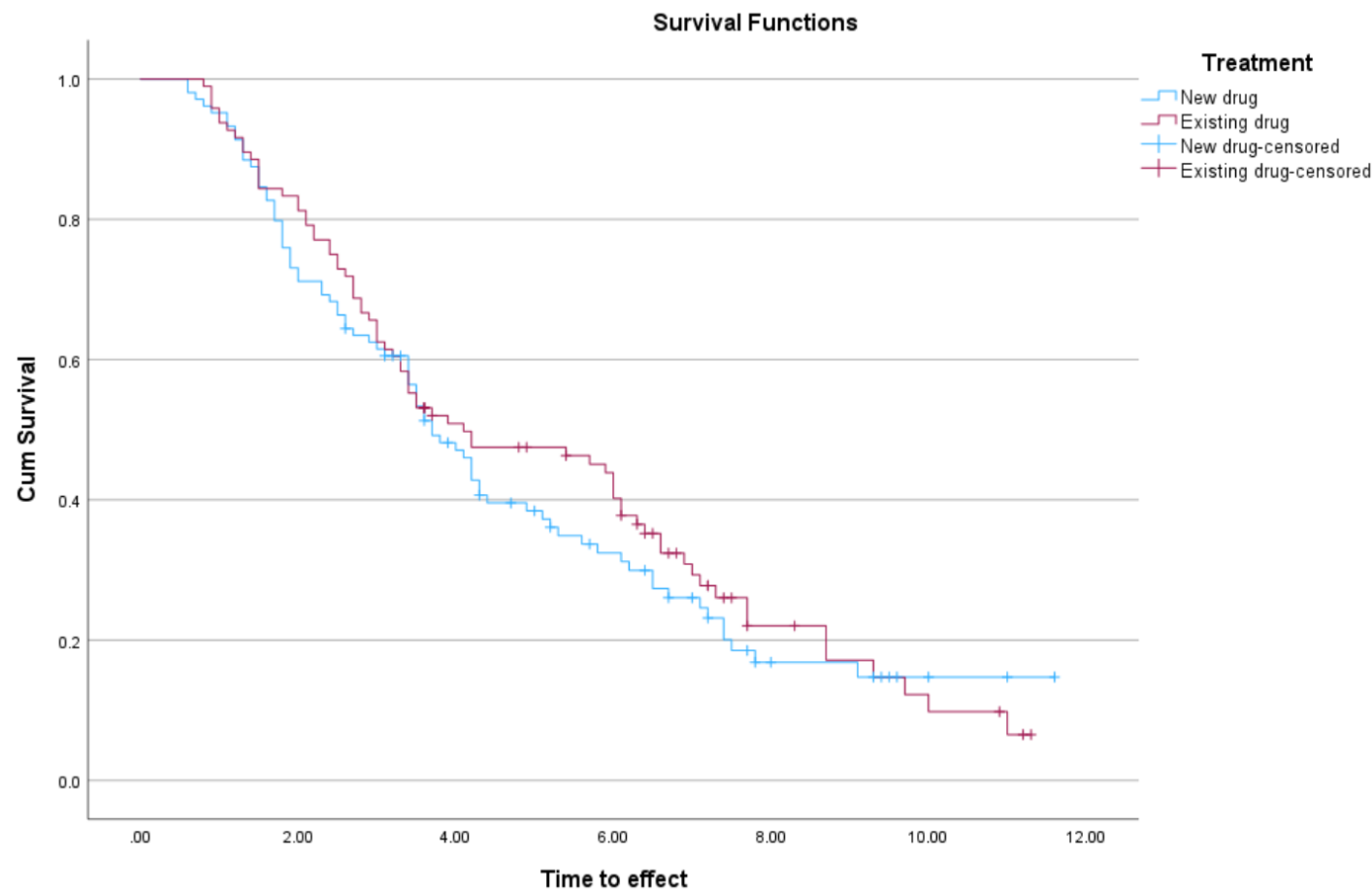
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.379	1	.538
Breslow (Generalized Wilcoxon)	.748	1	.387
Tarone-Ware	.705	1	.401

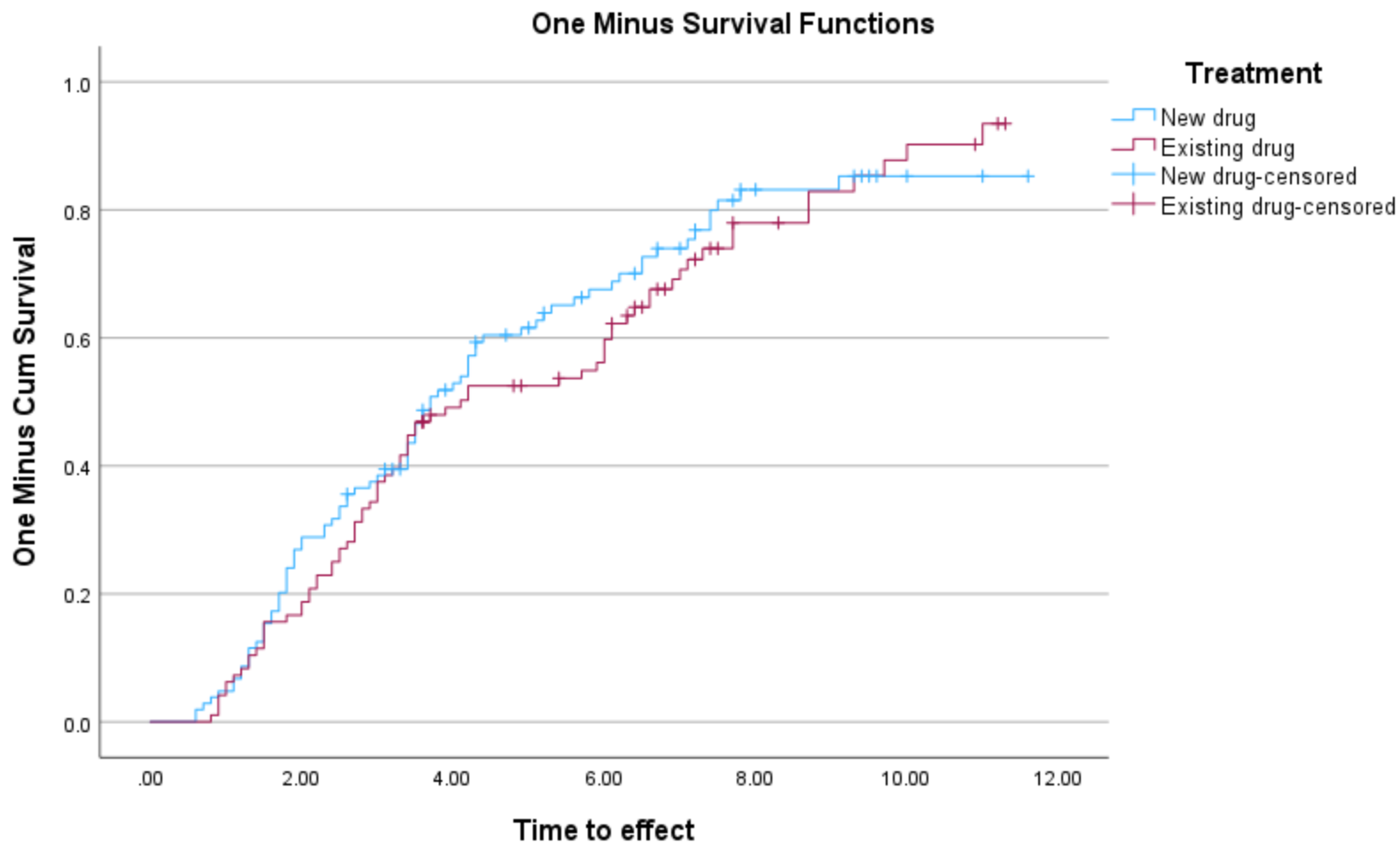
Test of equality of survival distributions for the different levels of Treatment.



Survival Function Chart



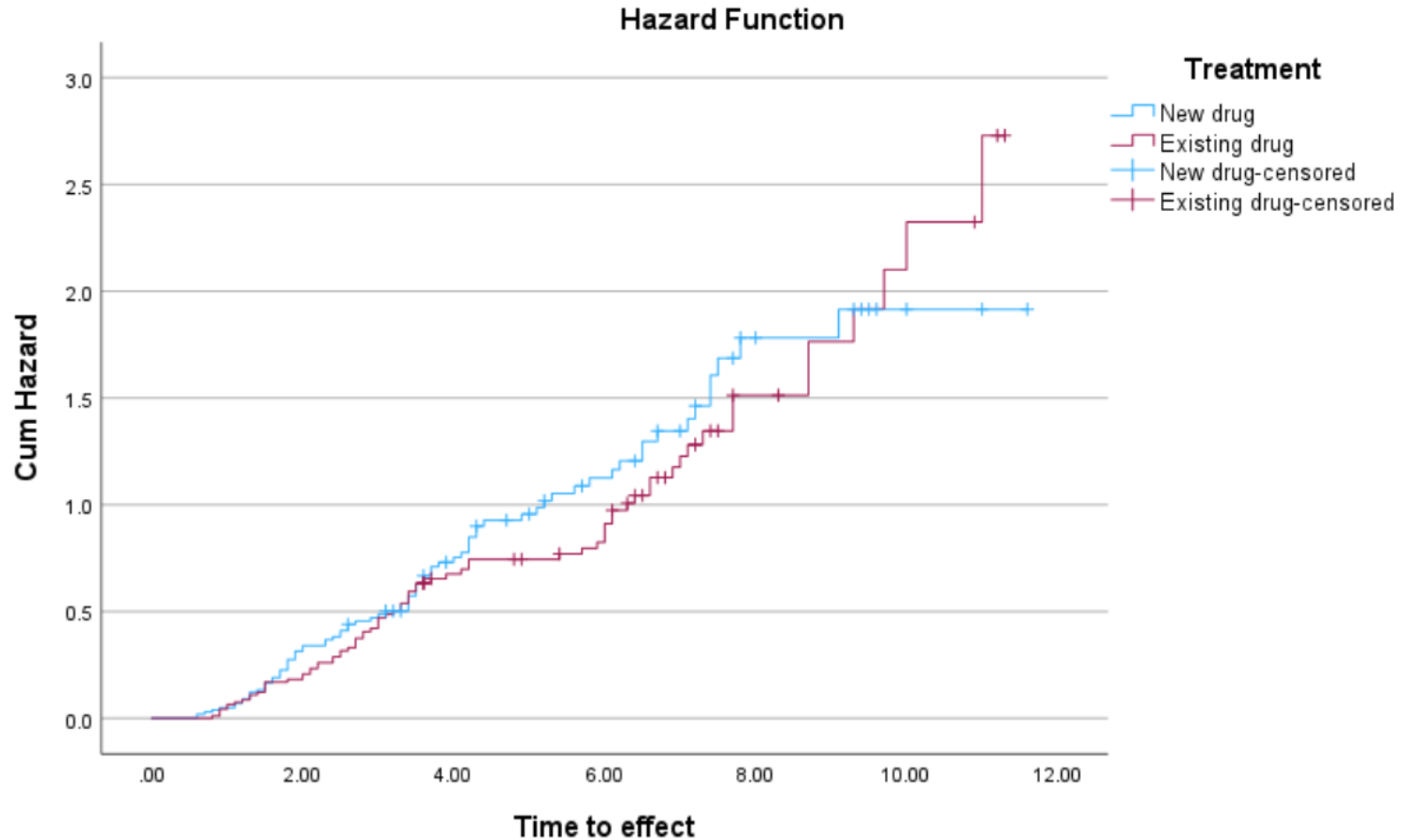
One-Minus Survival Function



Survival Function vs Hazard Function

- The **Survival** function denotes the probability that a person or subject will 'survive' beyond a given timepoint based on the rate of occurrence in a sample of data
- But the **Hazard** function, denotes the *rate* of failure or non-survival occurring at a given timepoint based on the fact that the subjects have made it this far without encountering the event of interest
- The **Cumulative Hazard** function however measures the expected *rate* of occurrence in a given timepoint so the values can exceed 1.0. If the periodicity was measured in years for example, one could reach a timepoint, where a given event might be expected to have occurred 1.5 times on average by that stage.

Hazard Function Chart





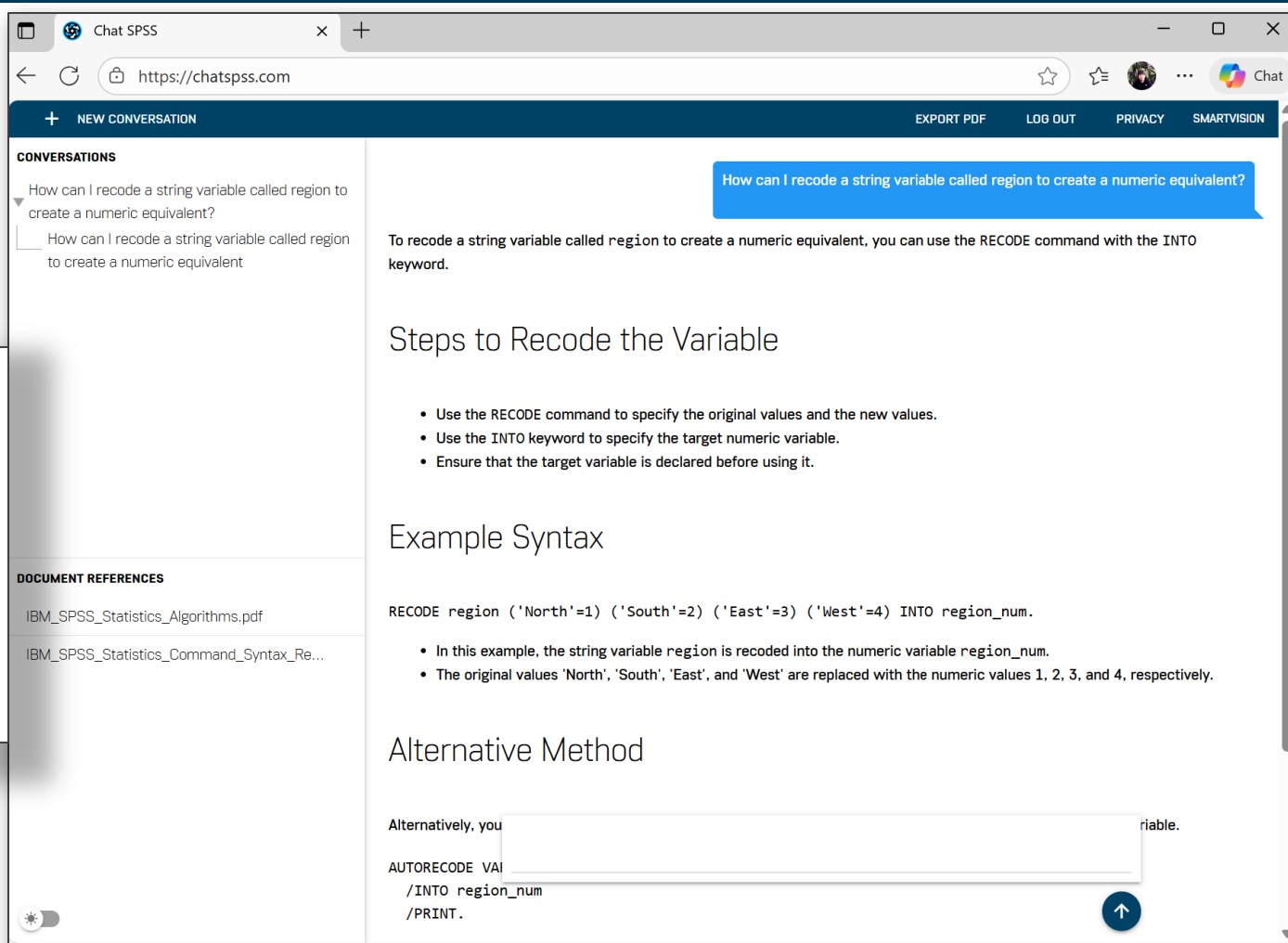
Before we go

Useful Video Resources

- [Getting started with SPSS](#)
- [Understanding Significance Tests](#)
- [Understanding Correlation Values](#)
- [Getting started with Decision Trees](#)
- [Working with Odds Ratios and Relative Risk Estimates](#)
- [Understanding Survival Analysis](#)
- [How to perform meta analysis in SPSS Statistics](#)



ChatSPSS is a free generative AI chatbot developed by Smart Vision Europe Ltd based on official IBM documentation of SPSS procedures



The screenshot shows the ChatSPSS web interface. The browser address bar displays <https://chatspss.com/>. The interface includes a navigation bar with links for "NEW CONVERSATION", "EXPORT PDF", "LOG OUT", "PRIVACY", and "SMARTVISION".

CONVERSATIONS

How can I recode a string variable called region to create a numeric equivalent?

How can I recode a string variable called region to create a numeric equivalent?

DOCUMENT REFERENCES

- IBM_SPSS_Statistics_Algorithms.pdf
- IBM_SPSS_Statistics_Command_Syntax_Re...

How can I recode a string variable called region to create a numeric equivalent?

To recode a string variable called **region** to create a numeric equivalent, you can use the **RECODE** command with the **INTO** keyword.

Steps to Recode the Variable

- Use the **RECODE** command to specify the original values and the new values.
- Use the **INTO** keyword to specify the target numeric variable.
- Ensure that the target variable is declared before using it.

Example Syntax

```
RECODE region ('North'=1) ('South'=2) ('East'=3) ('West'=4) INTO region_num.
```


- In this example, the string variable **region** is recoded into the numeric variable **region_num**.
- The original values 'North', 'South', 'East', and 'West' are replaced with the numeric values 1, 2, 3, and 4, respectively.

Alternative Method

Alternatively, you can use the **AUTORECODE** command.

```
AUTORECODE VAI  
/INTO region_num  
/PRINT.
```

Smart Vision Europe: Services and Expertise



We have decades of experience providing guidance, training and consultancy in the delivery of effective data science initiatives.

Working with Smart Vision Europe

- **Sourcing Software**
 - You can buy your analytical software from us often with discounts
 - Assist with selection, pilot, implementation & support of analytical tools
 - <http://www.sv-europe.com/buy-spss-online/>
- **Training**
 - Formal classroom/virtual training
 - Online self-paced training resources
- **Advice and Support**
 - ‘No strings attached’ technical and business advice relating to analytics
 - Tracked technical support services around the IBM SPSS product line



Contact us:

+44 (0)207 786 3568

info@sv-europe.com

Twitter: @sveurope

[Follow us on Linked In](#)

[Sign up for our Newsletter](#)



Thank you