Session 1

Introduction to Biostatistical analysis

4th March 2019

Chan Yiong Huak
Head, Biostatistics Unit
Yong Loo Lin School of Medicine
National University of Singapore

Data Types

1. Quantitative

    - Discrete : eg number of children
    - Continuous : eg age, SBP

2. Qualitative / Categorical

    - Nominal : eg race
    - Ordinal : eg pain severity

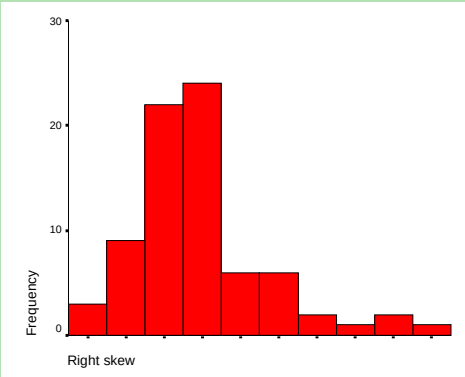
QUANTITATIVE DATA ANALYSIS

Parametric tests

How to check for Normality?

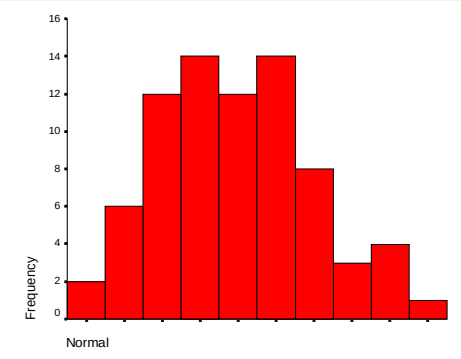
Normality Checking

Skewness



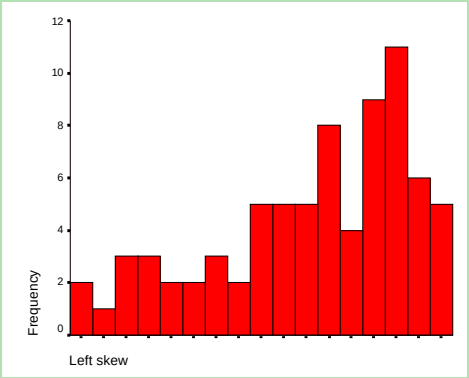
Right skew

Skew > 0



Normal

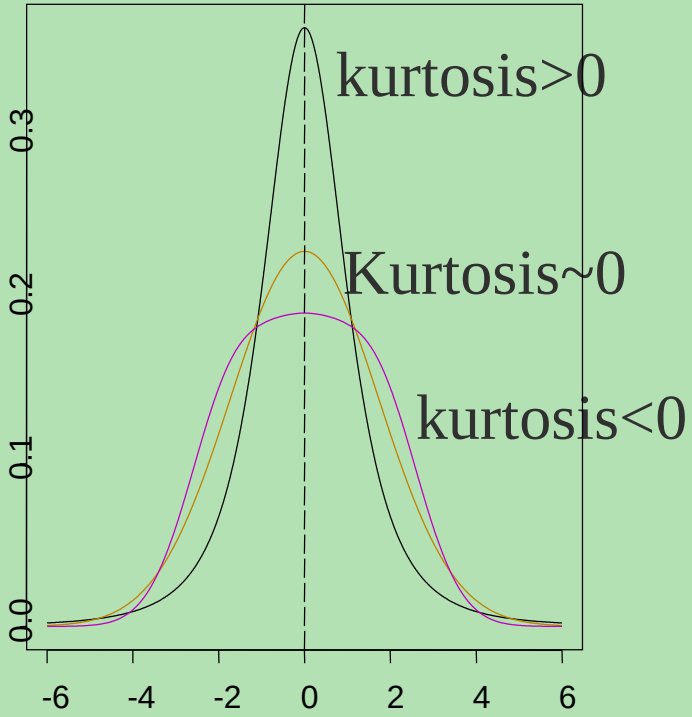
Skew ~ 0



Left skew

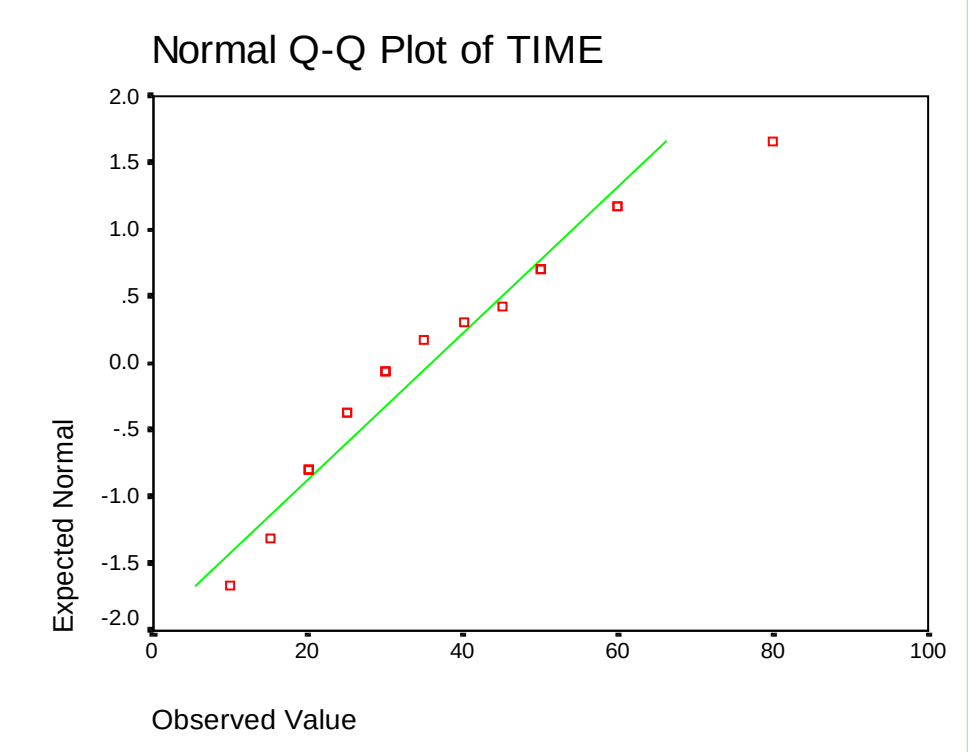
Skew < 0

Kurtosis



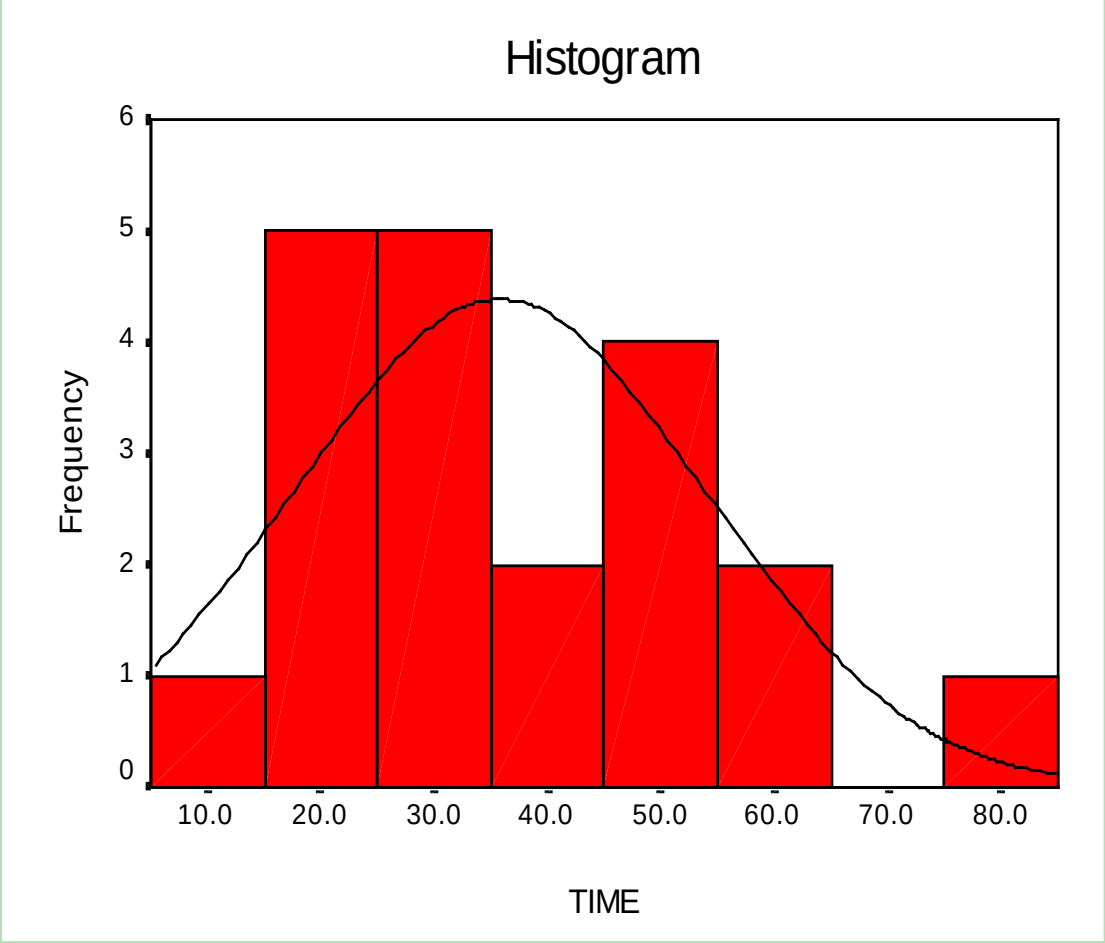
Normality Checking

Q-Q Plot



Normality Checking

Histogram



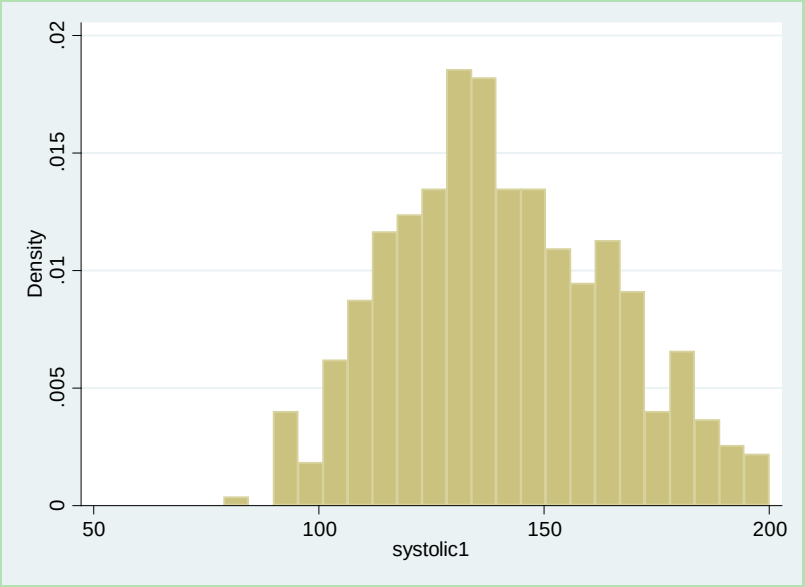
Normality Checking

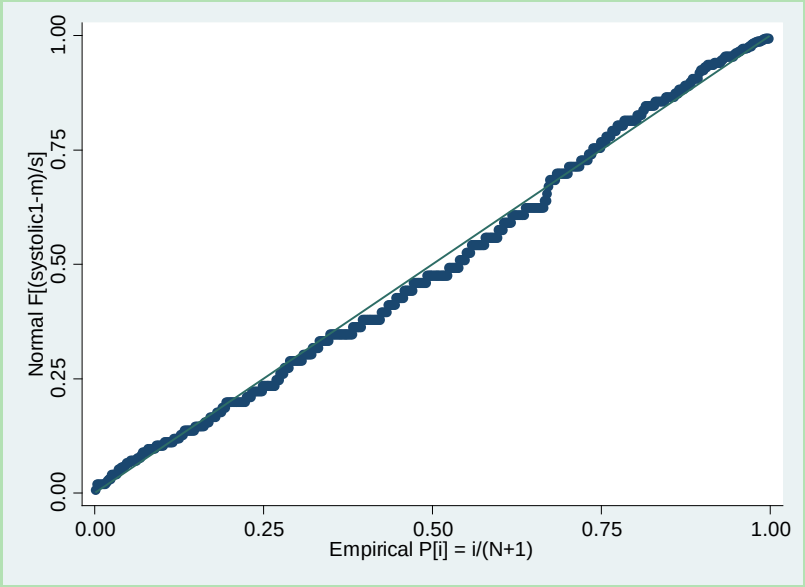
Formal tests

T

	K a			S		
	S	d	S	S	d	S
T	.	2	.	.	2	.

a L





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SBP	.049	500	.006	.990	500	.002

a. Lilliefors Significance Correction

Is the distribution normal?

Quantitative data

Normality & Homogeniety Assumptions Satisfied?

YES	**NO**
Parametric Tests	**Non-parametric Tests**
1 Sample T test Paired T test	Wilcoxon Signed-Rank test
2 Samples T test	Wilcoxon Rank Sum test / Mann-Whitney U test
One-way ANOVA Post-Hoc tests	Kruskal-Wallis test Bonferroni correction

Parametric Tests – analysis of the means

Non-Parametric – analysis of the median

1 Sample T / Wilcoxon Signed rank

Is the mean birth weight of our Singapore babies any different from the US babies.

Only have a 3.5kg from a US report.

Can have a sample of local babies

Paired T / Wilcoxon Signed Rank

Postulate that subjects will have a reduction in systolic BP after intervention.

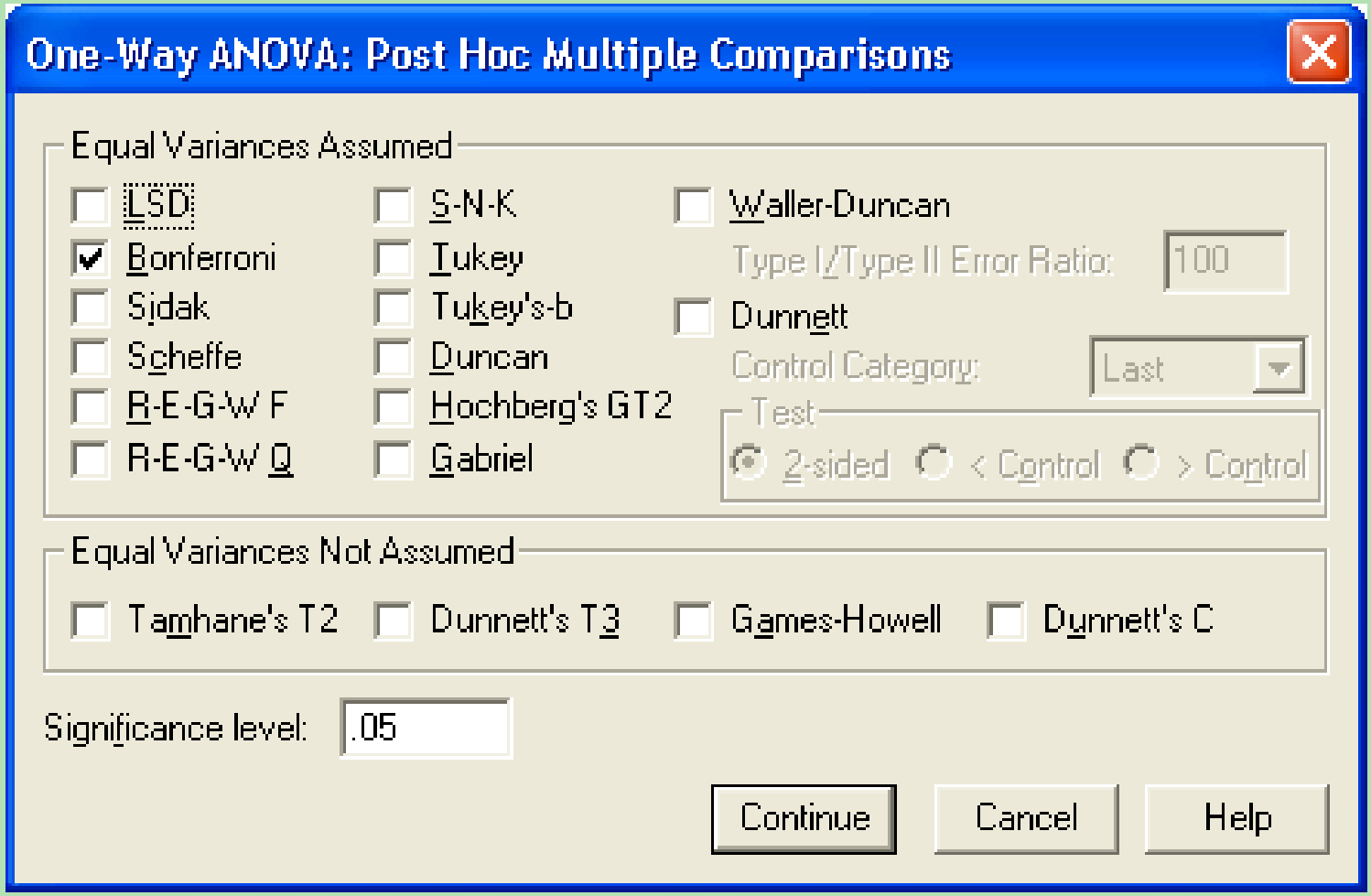
Two Samples T – Test / Mann Whitney U

Is the systolic BP reduction of the Active group significantly higher than the Control?

One-Way ANOVA / Kruskal Wallis

Are there any differences in the Systolic BP across races?

Post – Hoc options for multiple comparisons



Bonferroni
n(n-1)/2

n=3 : X 3
n=4 : X 6
n=5 : X 10
n=6 : X 15
etc

Sidak
Scheffe
Tukey

<table>
<tr><th colspan="2">Quantitative
Normal?</th><th>Qualitative</th></tr>
<tr><td><u>Yes</u>
Mean (sd)
Error bar</td><td><u>No</u>
Median
Box plot</td><td>n (%)
Pie chart
Bar chart</td></tr>
<tr><td colspan="2">Histogram</td><td rowspan="4">??</td></tr>
<tr><td>1 Sample T
Paired T</td><td>Wilcoxon Signed Rank</td></tr>
<tr><td>2 Sample T</td><td>Mann Whitney U</td></tr>
<tr><td>1 way ANOVA
Post-Hoc (Tukey, Sidak, Scheffe, Bonferroni)</td><td>Kruskal Wallis
Bonferroni ([n(n-1)/2])</td></tr>
</table>

QUALITATIVE DATA ANALYSIS

- QUALITATIVE (CATEGORICAL) DATA
  - Nominal
  - Ordinal

QUALITATIVE DATA

CONTINGENCY TABLE

	Variable 2		
Variable1	Factor 1	Factor 2	Factor 3
Factor A	n_{11}	n_{12}	n_{13}
Factor B	n_{21}	n_{22}	n_{23}

CONTINGENCY TABLE

Question

Is there evidence in the data for association between the categorical variables?

TEST OF INDEPENDENCE
CHI-SQUARE TEST

CONTINGENCY TABLE

Chi-Square Test

Is there a difference between
Groups A & B
in the success outcome rate?

CONTINGENCY TABLE (2 X 2)

treatment groups * success outcome Crosstabulation

			success outcome		
			no	yes	Total
treatment groups	A	Count	81	169	250
		% within treatment groups	32.4%	67.6%	100.0%
	B	Count	109	141	250
		% within treatment groups	43.6%	56.4%	100.0%
Total		Count	190	310	500
		% within treatment groups	38.0%	62.0%	100.0%

CONTINGENCY TABLE

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.655[b]	1	.010		
Continuity Correction [a]	6.188	1	.013		
Likelihood Ratio	6.674	1	.010		
Fisher's Exact Test				.013	.006
Linear-by-Linear Association	6.642	1	.010		
N of Valid Cases	500				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 95.00.

CONTINGENCY TABLE

Validity of Chi-Square Test

- No cell should have an expected count of less than 1
- No more than 20% of the cells should have an expected count less than 5

FISHER EXACT PROBABILITY TEST

STRENGTH OF ASSOCIATION

Measuring the Strength of Association

(only for 2 X 2 tables)

STRENGTH OF ASSOCIATION

The magnitude of the p-value does not indicate the strength of association between 2 categorical variables

STRENGTH OF ASSOCIATION

Odds Ratio (OR) : Prevalent Study design

Cross-sectional & Case-Control

Relative Risk (RR) : Incident Study design

Cohort Studies & Randomised Control Study

McNEMAR'S TEST
(Matched Case-Control Study)
1 to 1 matching

Not diabetic * diabetic Crosstabulation

Count

		diabetic		
		No AMI	AMI	Total
Not diabetic	No AMI	82	37	119
	AMI	16	9	25
Total		98	46	144

Matched Case-Control Study

1 to 1 matching (today) is a weak design

Group to group matching is the better design

<table>
<tr><td colspan="2">Quantitative
Normal?</td><td>Qualitative</td></tr>
<tr><td><u>Yes</u>
Mean (sd)
Error bar</td><td><u>No</u>
Median
Box plot</td><td>n (%)
Pie chart
Bar chart</td></tr>
<tr><td colspan="2">Histogram</td><td rowspan="4">Chi square / Fisher’s Exact
OR/ RR

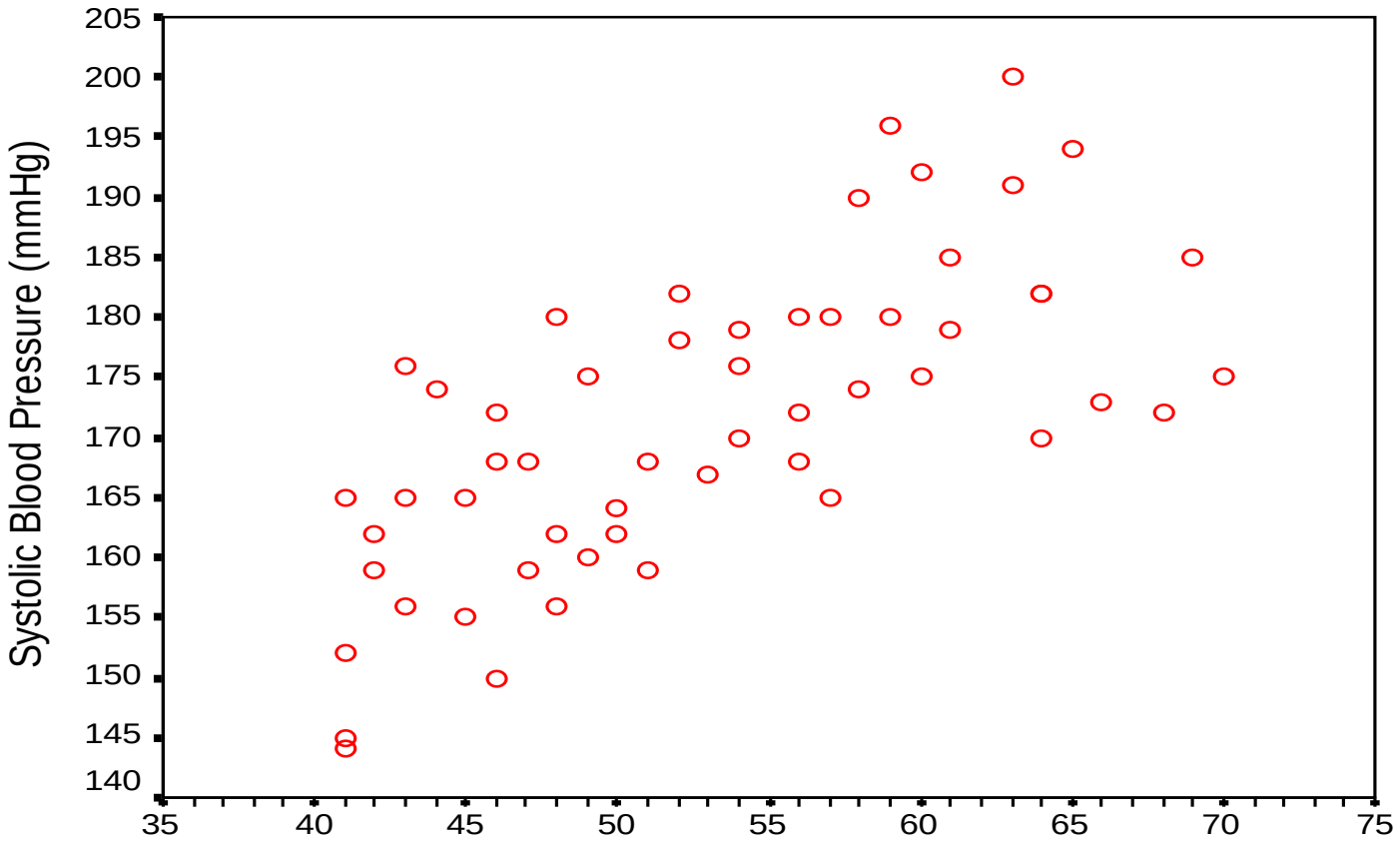
McNemar</td></tr>
<tr><td>1 Sample T
Paired T</td><td>Wilcoxon Signed Rank</td></tr>
<tr><td>2 Sample T</td><td>Mann Whitney U</td></tr>
<tr><td>1 way ANOVA
Post-Hoc (Tukey, Sidak, Scheffe, Bonferroni)</td><td>Kruskal Wallis
Bonferroni ([n(n-1)/2])</td></tr>
</table>

Correlational Analysis

Correlation

Describing linear association between 2 quantitative variables

(Systolic blood pressure and age of 55 hypertension patients)



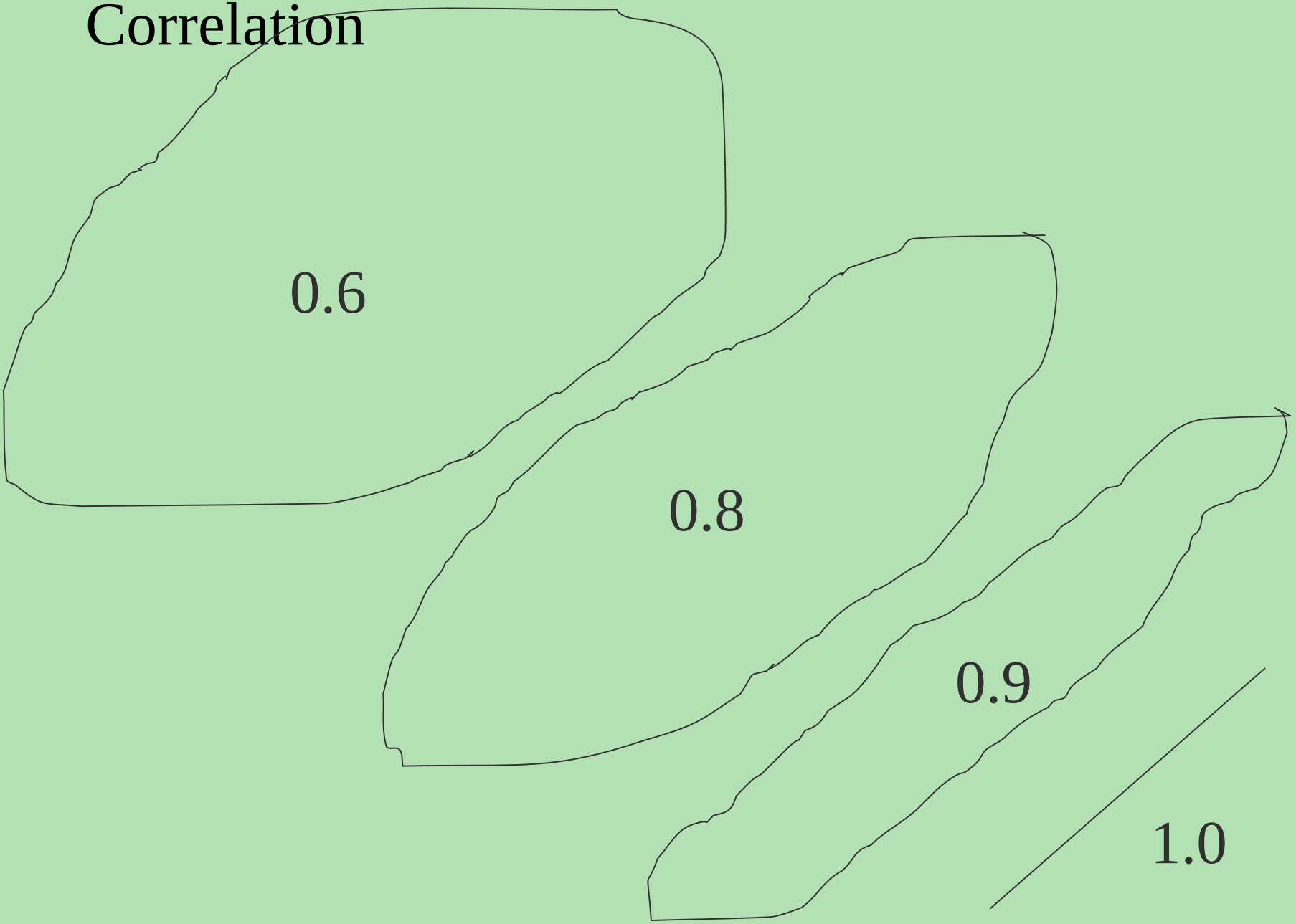
Correlation

Pearson's correlation coefficient (r)

- Measures linear relationship
- Range from -1 to +1
- The sign (+/-) indicates whether association is positive or negative
- Values close to 0 imply no linear association

Interpretation :
Greater than 0.9 – very strong
0.8 – 0.9 : strong
0.7 – 0.8 : moderate
0.6 – 0.7 : mild
0.5 – 0.6 : mild to poor
Less than 0.5 - poor

Correlation



Correlation

Pearson correlation assumptions

- both variables have to be Normal

Spearman’s rank correlation

Non-parametric method

Same interpretation as Pearson’s

Correlation

Correlations

		Systolic Blood Pressure (mmHg)	Age (years)
Systolic Blood Pressure (mmHg)	Pearson Correlation	1.000	.696**
	Sig. (2-tailed)	.	.000
	N	55	55
Age (years)	Pearson Correlation	.696**	1.000
	Sig. (2-tailed)	.000	.
	N	55	55

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

Correlations

			Systolic Blood Pressure (mmHg)	Age (years)
Spearman's rho	Systolic Blood Pressure (mmHg)	Correlation Coefficient	1.000	.717**
		Sig. (2-tailed)	.	.000
		N	55	55
	Age (years)	Correlation Coefficient	.717**	1.000
		Sig. (2-tailed)	.000	.
		N	55	55

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

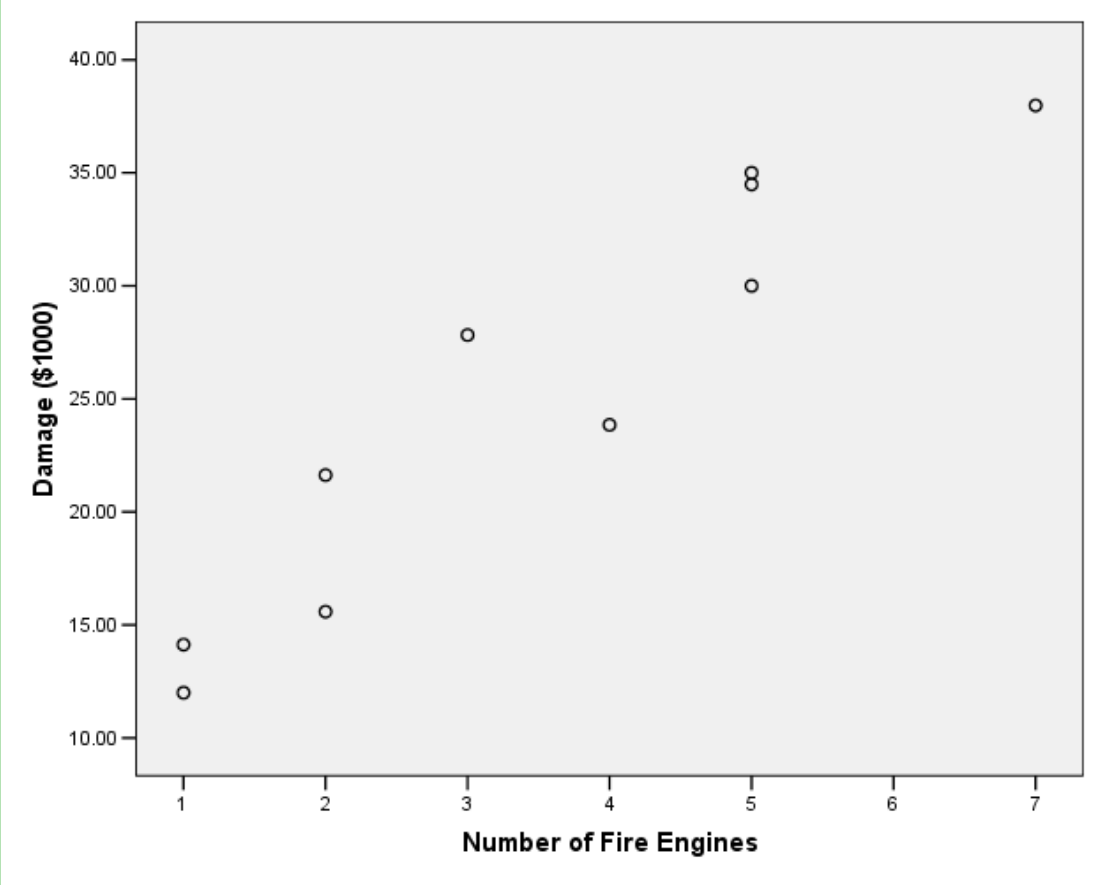
Correlation

Correlation is not Causation !!!

If two variables are highly correlated, it does not mean one causes the other.

Correlation

Interpret this scatter plot



Definitions

	Statistical	Medical
Univariate	1 y 1 or many x	1y 1x
Multivariate	Many y 1 or many x	1y Many x

Regression Modeling

Purpose of Regression Modeling

1. Descriptive
   - form and strength of the association between outcome and factors of interest

2. Adjustment
   - for covariates and confounders

3. Prediction
   - the future outcome

Linear Regression Analysis

1. Linear Regression

Dependent variable

Quantitative

Independent variables

Quantitative / Qualitative

Simple Linear Regression

Example

Given the systolic blood pressure (mmHg) and age (in years) of 55 hypertension patients, we are interested to determine whether there is a linear relationship between the 2 variables.

Mathematically, we can write their relationship as follows :

SBP (mmHg) = a + b*Age (years) + error term

Assumptions for error term :
mean = 0, constant variance and normally distributed.

Simple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	115.706	7.999		14.465	.000	99.662	131.7
	Age (yrs)	1.051	.149	.696	7.060	.000	.752	1.350

a. Dependent Variable: Systolic Blood Pressure (mmHg)

There is a 1.05 (95% CI 0.752, 1.35), $p < 0.001$, mmHg increase in SBP for each 1-year increase in age of the patient.

Multiple Regression

An extension of the simple regression model by including more than one predictor variable

SBP (mmHg) = a + b * Age (years)
+ c * Smoking status
+ error term

Coding for smoking status : 1 for Yes; 0 for No

Multiple Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	110.667	7.311		15.136	.000	95.996	125.3
	Age (yrs)	1.055	.134	.699	7.893	.000	.787	1.324
	Smoker (Yes)	8.274	2.234	.328	3.703	.001	3.791	12.758

a. Dependent Variable: Systolic Blood Pressure (mmHg)

Both age and smoking status are the significant risk factors.

1. There is a 1.06 mmHg (95% CI 0.79 to 1.32) increase in SBP for each 1-year increase in the patient's age ($p < 0.001$).

2. There is a 8.27 mmHg (95% CI 3.79 to 12.76) increase in SBP for patient who is a smoker compared to a non-smoker ($p = 0.001$).

Multiple Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	120.851	7.200		16.784	.000	106.4	135.3
	Age (years)	.829	.135	.549	6.147	.000	.558	1.101
	Smoker (Yes)	6.440	2.075	.255	3.104	.003	2.273	10.608
	Under-weight	-1.481	2.782	-.046	-.532	.597	-7.069	4.108
	Over-weight	8.232	2.450	.321	3.360	.001	3.311	13.152

a. Dependent Variable: Systolic Blood Pressure (mmHg)

Reference group: smoking status = no, weight group = normal-weight

Selection Methods

Methods for Selecting Variables

- Enter
- Forward
- Stepwise
- Backward
- Remove

Quan Normal?		Qual	
Yes T-tests	No Non-para	Chi-sq Fisher's exact OR|RR McNemar	
Agreement: Bland Altman/Kappa			
Correlation: Pearson/Spearman			
Linear Reg			

2. Logistic Regression

Dependent variable

Qualitative

Independent variables

Quantitative / Qualitative

Logistic Regression

This method is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables.

Logistic Regression

The coefficients in the logistic regression can be used to estimate odds ratios for each of the independent variables in the model.

Logistic Regression

Example, what lifestyle characteristics are risk factors for coronary heart disease(CHD)?

Given a sample of patients measured on smoking status, diet, exercise, alcohol use and CHD status, a model could be developed
using the four lifestyle variables to predict the presence or absence of CHD
in the sample of patients.

Logistic Regression

The model can then be used to derive estimates of the odds ratios for each factor, for example, how much likely smokers are to develop CHD than nonsmokers.

The model can also be used to predict the probability of a person to have CHD or not given the characteristic of his life-style.

Logistic regression only provide Odds ratios. To get multivariate Relative Risks, Poisson regression or Modified Cox is used.

Quan Normal?		Qual	
Yes T-tests	No Non-para	Chi-sq Fisher's exact OR|RR McNemar	
Agreement: Bland Altman/Kappa			
Correlation: Pearson/Spearman			
Linear Reg		Logistic Reg (OR) Poisson / Modified Cox (RR)	

3. Survival Analysis

Dependent variable

Quantitative – time to event

Independent variables

Quantitative / Qualitative

SURVIVAL ANALYSIS

Survival analysis describes the analysis of data that correspond to the time from a well-defined time origin until the occurrence of some particular event of interest or end-point

SURVIVAL ANALYSIS

Survival Time

The time of a major outcome variable from randomisation to a specified critical event

Outcomes

- Duration - time from randomisation to relapse
- Pressure sore - time to development
- Survival - time from randomisation until death in cancer patients

SURVIVAL ANALYSIS

Although survival time is a continuous variable, one cannot use the standard tests for analysis.

There are 2 reasons for this :

- The distribution of survival times is unlikely to be Normal and it may not be possible to find a transformation

- The presence of ***censored*** observations

SURVIVAL ANALYSIS

Censored observations arise in cases for which

- the critical event has not yet occurred
- lost to follow-up
- other interventions offered
- event occurred but unrelated cause

The time from randomisation to the last date the case was examined is known as the ***censored survival time.***

SURVIVAL ANALYSIS

Comparison of 2 Survival Curves

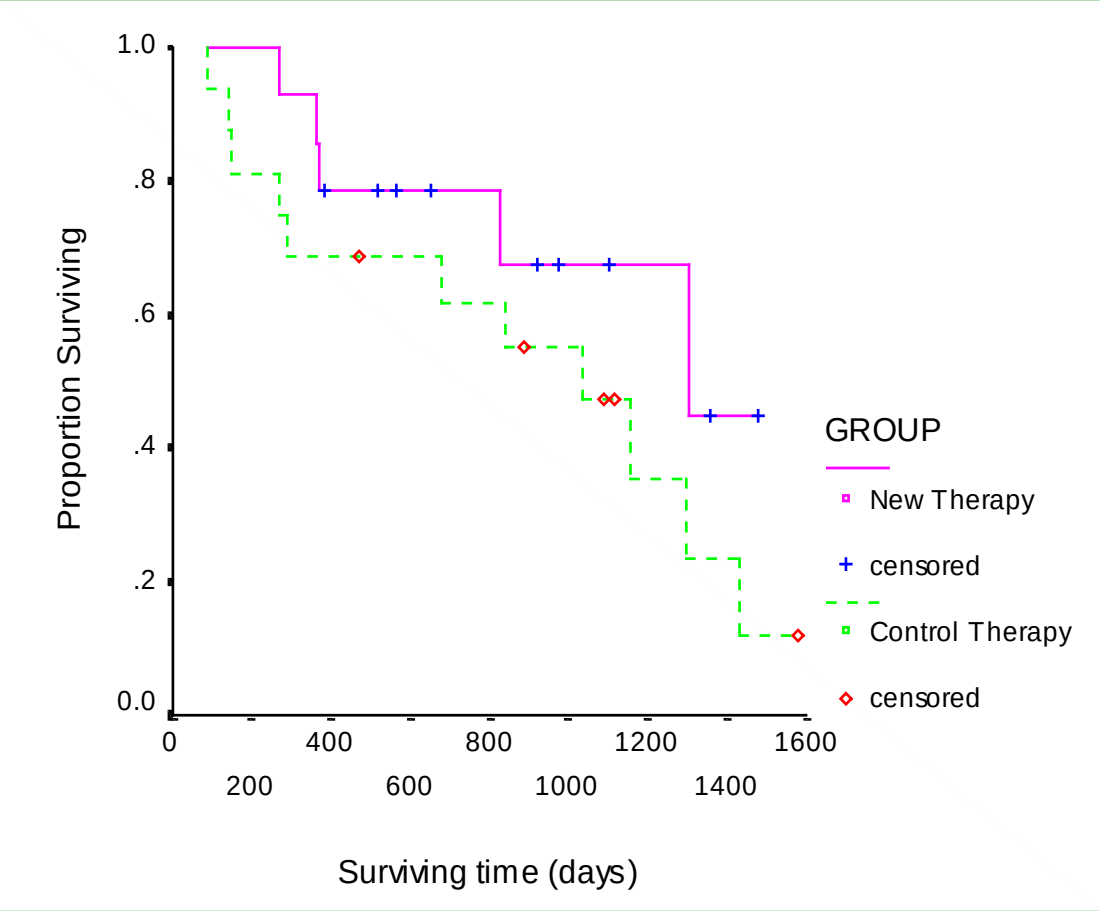
Example : Survival by treatment group

(Control vs New Therapy)

of 30 patients recruited by a cervical cancer trial.

SURVIVAL ANALYSIS

Comparison of two survival curves



Control Therapy

Mean = 887
Median = 1037

New Therapy

Mean = 1120
Median = 1307

SURVIVAL ANALYSIS

Log Rank Test

	Total	Number Events	Number Censored	Percent Censored
control	16	11	5	31.25
active	14	5	9	64.29
Overall	30	16	14	46.67

SURVIVAL ANALYSIS

Log Rank Test

Test Statistics for Equality of Survival Distributions for GROUP

	Statistic	df	Significance
Log Rank	1.68	1	.1947

SURVIVAL ANALYSIS

Cox regression (Proportional Hazards)

Uses the **Hazard Function** to estimate the relative risk of ‘failure’. This function is a rate and is an estimate of the potential for ‘death’ per unit time at a particular instant, given that the case has ‘survived’ until that instant.

Quan Normal?		Qual	Time
Yes T-tests	No Non-para	Chi-sq Fisher's exact OR|RR McNemar	Kaplan Meier (log rank) Conditional logistic
Agreement: Bland Altman/Kappa			
Correlation: Pearson/Spearman			
Linear Reg		Logistic Reg (OR) Poisson / Modified Cox (RR)	Cox Reg (HR)

The End

Thank You