

Whirlwind Review of Statistics

Ram R Miller MD, MSc

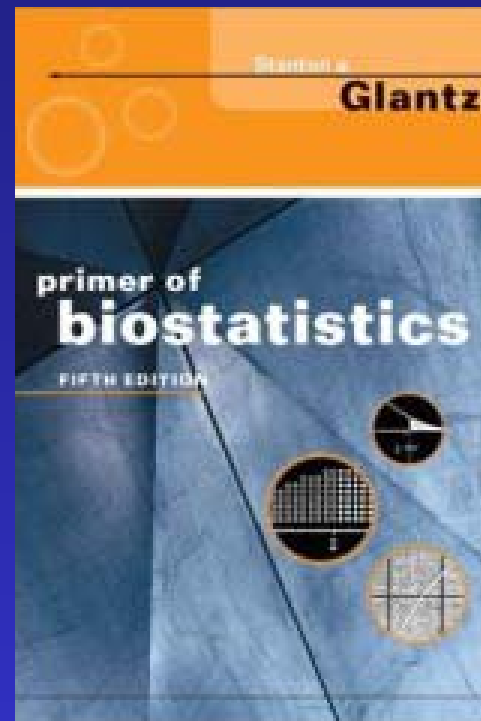
Division of Gerontology

Department of Epidemiology and
Preventive Medicine

rrmiller@epi.umaryland.edu

Recommended Reading

- Primer of Biostatistics
– Stanton A. Glantz



Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population				
Compare two independent groups				
Compare two dependent groups				
Association between variables (control for confounding)				
Compare more than two groups				

Types of Clinical Questions

- Describe a population:
 - What is the mean / average weight of people in the USA?
 - What is the outcome?
 - What values can it take?
 - This type of outcome is referred to as a *continuous* or *measurement* outcome

Types of Clinical Questions

- Describe a population:
 - What percentage of the US population is “overweight”
 - What is the outcome?
 - What values can it take?

Types of Clinical Questions

- This type of outcome is referred to as a *categorical* or *dichotomous* outcome
- Dichotomous means that it can be expressed as either **Yes** or **No**
 - Overweight
 - Hypertensive
 - Diabetic
 - Dead

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups				
Compare two dependent groups				
Association between variables (control for confounding)				
Compare more than two groups				

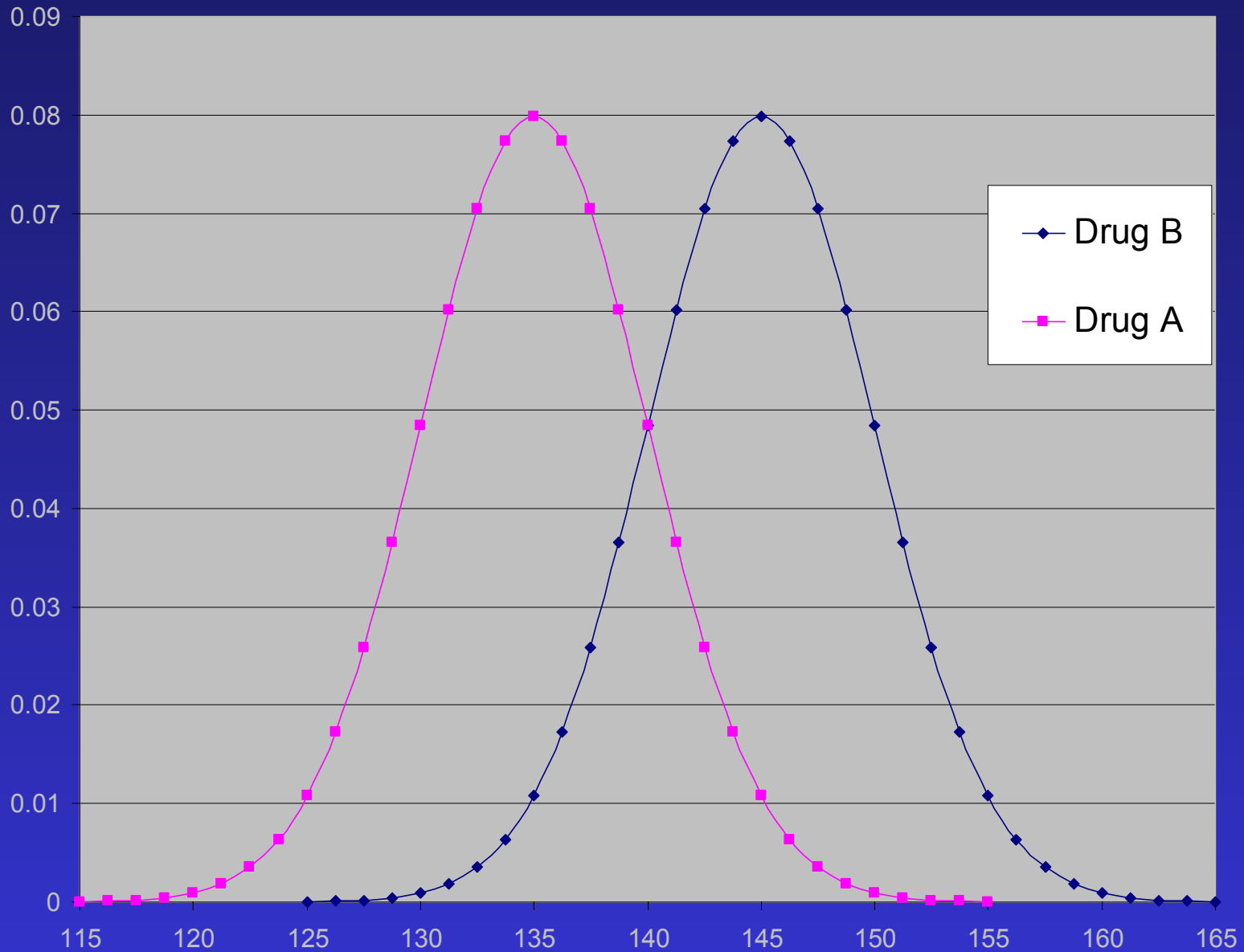
Types of Clinical Questions

- We might also want to compare two groups
 - Do people lose more weight on a low carb diet versus a low cal diet?
 - What is the outcome?
 - What values can it take?
 - The difference in weight lost is a continuous outcome

Types of Clinical Questions

- Compare two groups
 - Is there a greater percentage of overweight people in Maryland compared to Virginia ?
 - What is the outcome?
 - What values can it take?
 - This is another example of categorical / dichotomous outcome data

Effect of two drugs on BP



How do we compare two groups?

- Categorically:
 - Do more people taking drug A have high blood pressure compared to those taking drug B?
 - Relative risk
 - Odds ratio

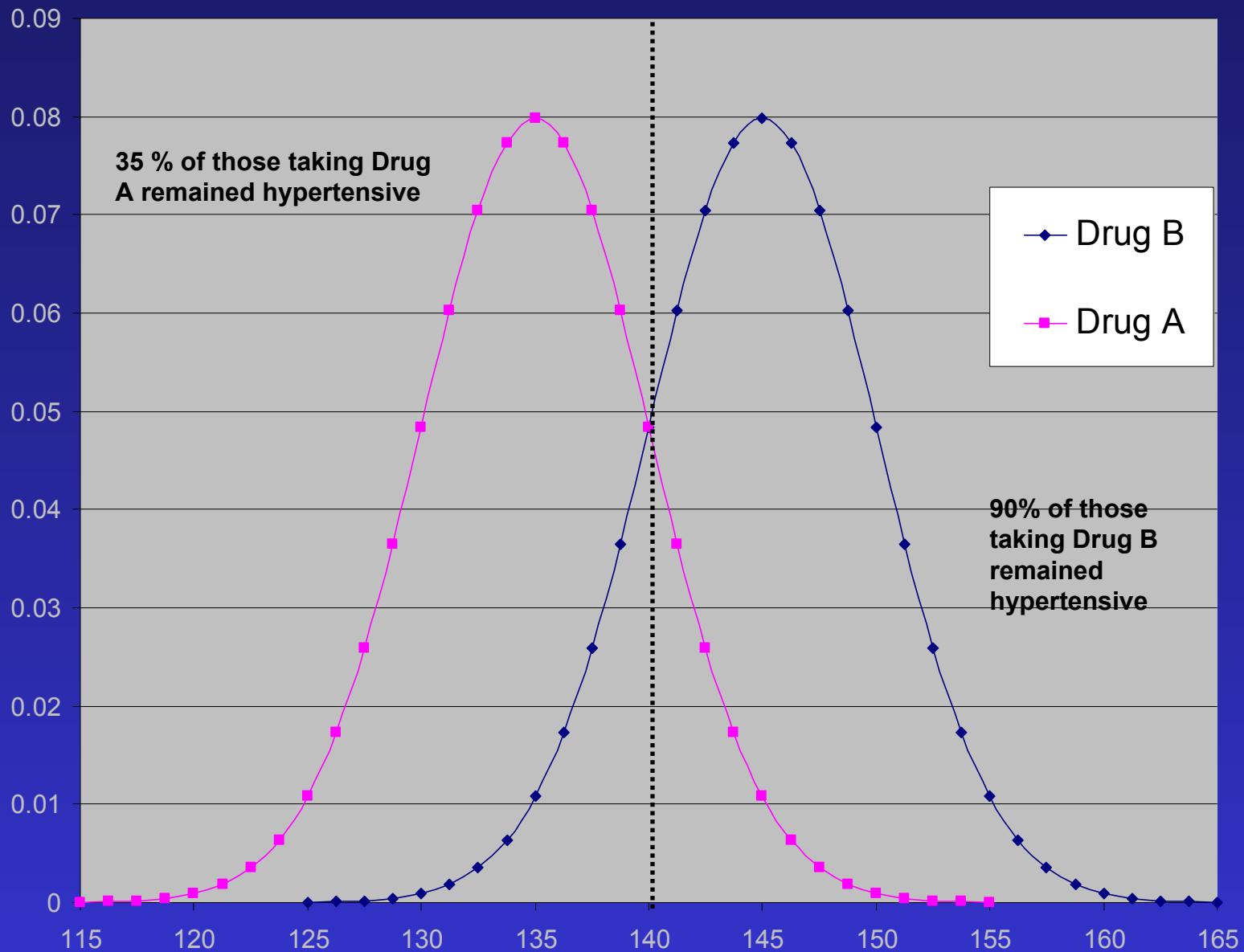
How do we compare two groups?

- Measurement:
 - What is the difference in the mean BP between people taking drug A and people taking drug B

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI)	RR, OR (CI)		
Compare two dependent groups				
Association between variables (control for confounding)				
Compare more than two groups				

Effect of two drugs on BP



RR of remaining hypertensive Drug B vs A

	Hypertensive	Not Hypertensive	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

$$\text{RR} = \frac{90 / 100}{35 / 100}$$
$$= 2.6$$

Chi-Square Test

- Used to test significance of Categorical Data
 - 2 X 2 data
 - 2 X 3 data
 - Drug A vs B for Hypertension, borderline hypertension and normotensive
 - 3 X 3 data
 - Drug A vs Drug B vs Drug C
 - Etc...

The Chi-Square Test

- Null Hypothesis: No difference in outcome across the groups
 - RR = 1
 - OR = 1

The Chi-Square Test

- How much do the data that we observe differ from what would be *expected* under the *Null Hypothesis*?

$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}}$$

Where the sum is taken over all the cells in the table

The Chi-Square Test

- Reject the null hypothesis if the X^2 value that you calculate is greater than a critical value
- Critical value is determined by the size of the table (2x2, 2x3, 3x3 etc...)

BP example

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A			100
Drug B			100
	125	75	200

Expected

BP example

Expected =

(Row total x Column total)

grand total

	HTN	No HTN	
Drug A			100
Drug B			100
	125	75	200

Expected

BP example

Expected =

(100 x 125)

200

	HTN	No HTN	
Drug A	62.5		100
Drug B			100
	125	75	200

Expected

BP example

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

$$\chi^2 = \frac{(35-62.5)^2}{62.5}$$

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

$$\chi^2 = \frac{(35-62.5)^2}{62.5} + \frac{(65-37.5)^2}{37.5}$$

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

$$\chi^2 = \frac{(35-62.5)^2}{62.5} + \frac{(65-37.5)^2}{37.5} + \frac{(90-62.5)^2}{62.5}$$

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

$$\chi^2 = \frac{(35-62.5)^2}{62.5} + \frac{(65-37.5)^2}{37.5} + \frac{(90-62.5)^2}{62.5} + \frac{(10-37.5)^2}{37.5}$$

	HTN	No HTN	
Drug A	35	65	100
Drug B	90	10	100
	125	75	200

Observed

	HTN	No HTN	
Drug A	62.5	37.5	100
Drug B	62.5	37.5	100
	125	75	200

Expected

$$\chi^2 = \frac{(35-62.5)^2}{62.5} + \frac{(65-37.5)^2}{37.5} + \frac{(90-62.5)^2}{62.5} + \frac{(10-37.5)^2}{37.5}$$

$$= 64.5$$

BP example

- For our blood pressure trial:
 - Chi – Square statistic = 64.5
- Compare this statistic to the critical value
 - For $P < 0.05$
 - Chi-Square > 3.84

BP example

- Here the Chi-Square statistic
 - $64.5 > 3.84$
- Therefore reject the null hypothesis:
 - The Relative Risk = 1 for high blood pressure, comparing those on Drug A to those on Drug B

Other Tests for Categorical Data

- Fisher's exact test:
 - Chi-square not appropriate for small samples, but Fisher's exact test is
 - Computationally more difficult but since most analysis is done on computers this is the preferred test for categorical data

Categorical Data on a Paired Sample

- Comparing one group before and after on a categorical outcome
 - Hypertensive (yes/no) before / after treatment
- Comparing matched samples on a categorical outcome
 - Case-control studies
 - Matched cohort studies

Categorical Data on a Paired Sample

- McNemar's test is used

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI)	RR, OR (CI) Chi-Square Test Fisher's Exact test		
Compare two dependent groups		McNemar's paired-sample chi-square test		
Association between variables (control for confounding)				
Compare more than two groups		Chi-Square Test		

Comparing two means

- The T-test
- Null hypothesis:
 - There is no difference in the means between the two groups

T-test

- Independent groups
 - Two separate groups of subjects
- Dependent groups
 - Same people compared under two separate conditions
 - Before / after
 - Paired sample T-test

Comparing Two Means: The T-test

- The t-statistic is a ratio between the observed effect and the standard error of the effect
- Like the Chi-square compare the observed t-statistic to a critical value to determine statistical significance
- For $P < 0.05$
- Critical value for $t = 2$

Comparing more than two groups

- Suppose we wanted to study the mean blood pressure across more than two groups
- Let say we wanted to study the effect of three drugs on mean blood pressure
- Would use a technique called **Analysis of Variance (ANOVA)**

ANOVA

- Null Hypothesis: all means are equal
- ANOVA compares the variance within each group to the variance between the groups
- If the null is true (all groups have the same mean), then the averages of each individual group are estimates of the same underlying mean and the between group variation can be predicted from the within group variation
- But if the between group variation is larger than expected then it is less likely that the null is true

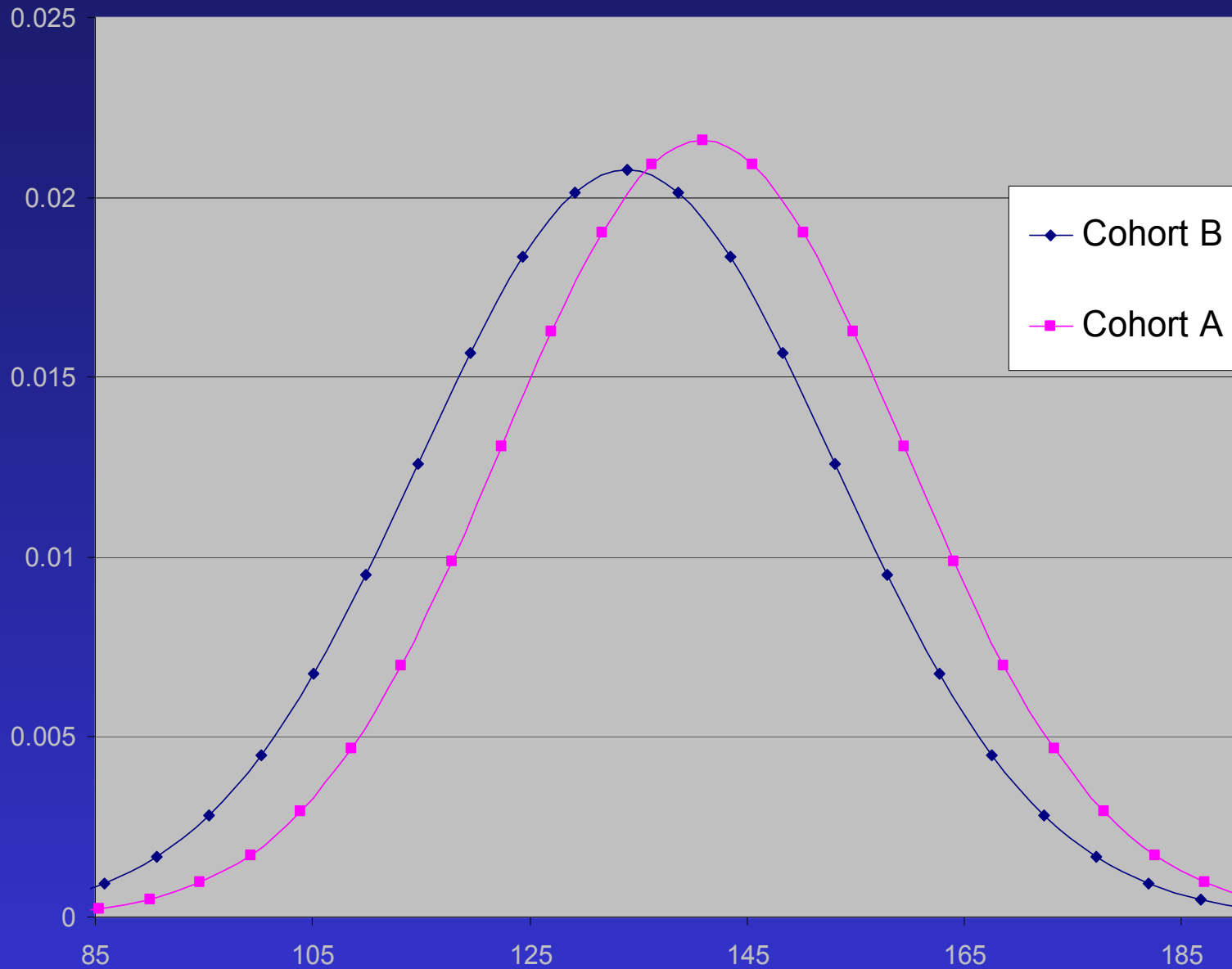
Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test		
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)				
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Correlation and Regression

- We want to compare the blood pressure from two cohorts

SBP for Two Cohorts



T-test

- We perform a t-test on the means and get the following result:
 - Difference in the means = 7.0 mmHg
 - $P < 0.0001$

Could age be a confounder?



Regression Modeling

- The main use for regression modeling is to examine an effect while adjusting for confounding by other factors
- Doing this is never perfect
- It is always better to design a study properly to prevent confounding in the first place

Could Age be a confounder?

- In order to explore whether age might be a confounder in the relationship between the mean SBP in the two cohorts
- We first need to explore how age and SBP are related

Could Age be a confounder?

- Mean age in Cohort A = 71.6 years
 - Cohort A had higher BP
- Mean age in Cohort B = 65.7 years
 - Cohort B had lower BP

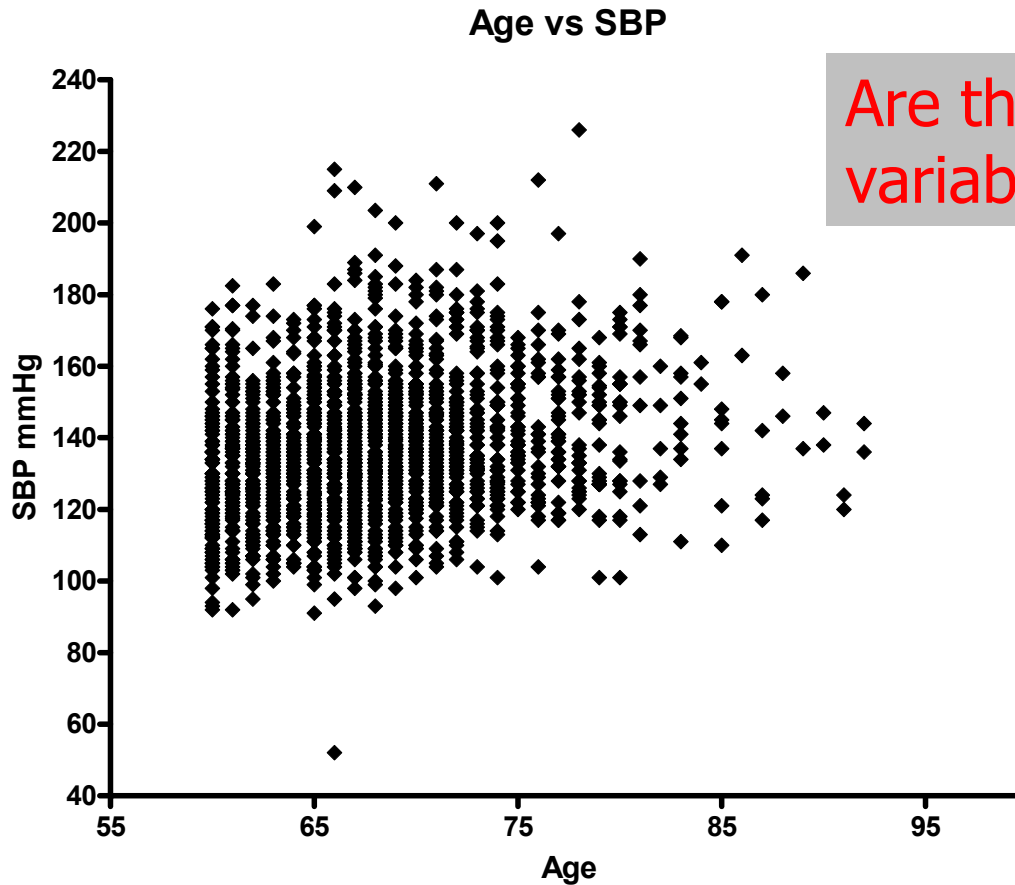
Could Age be a confounder?

- Age appears to be associated with the *exposure*
 - Cohort A is older than Cohort B

Could Age be a confounder?

- Is age associated with the *outcome*
 - Systolic BP in our example
- Does SBP increase with age?

Age vs. SBP



Are these two variables related?

Correlation

- Correlation modeling is a statistical technique employed to answer the question:
 - Is there a *LINEAR* relationship between two variables
 - Between age and systolic blood pressure in our example

Correlation

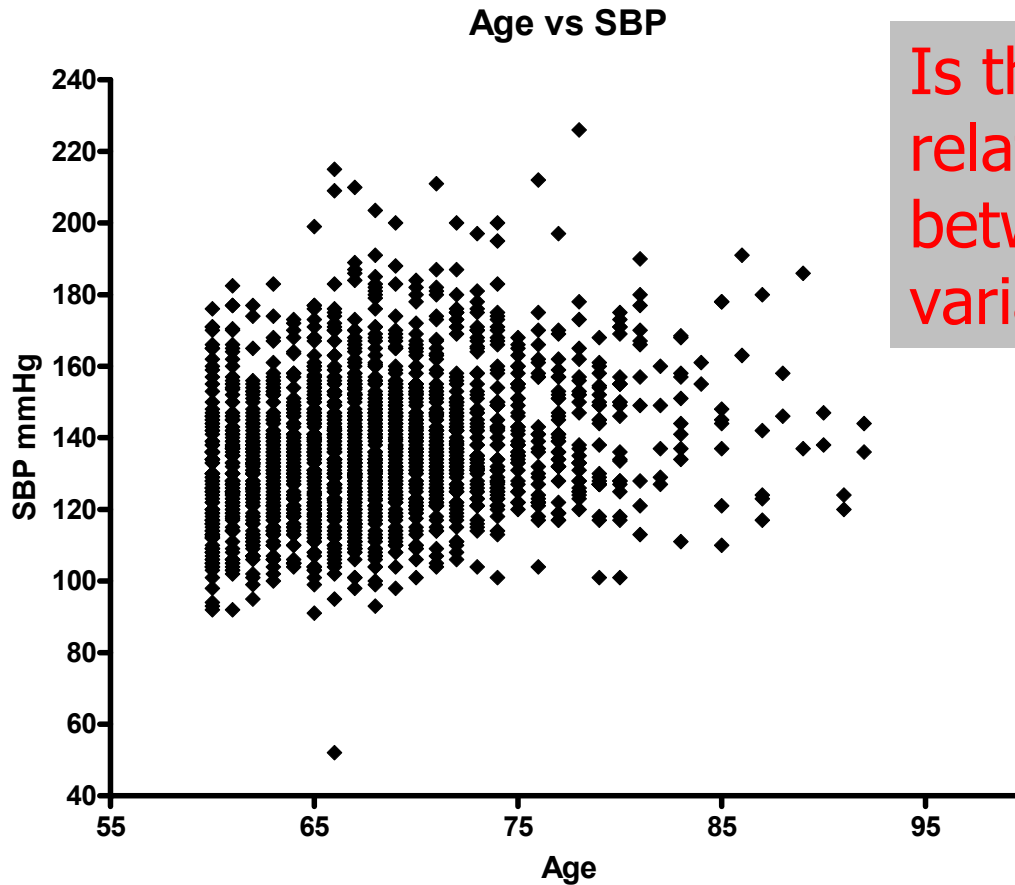
- The strength of the correlation is represented by the correlation coefficient r
- It depends on how much scatter there is
 - $|r| < 0.2$ Very weak correlation
 - $|r|$ between 0.2 and 0.5 weak correlation
 - $|r|$ between 0.5 and 0.8 moderate correlation
 - $|r| > 0.8$ strong

Weblink

Correlation

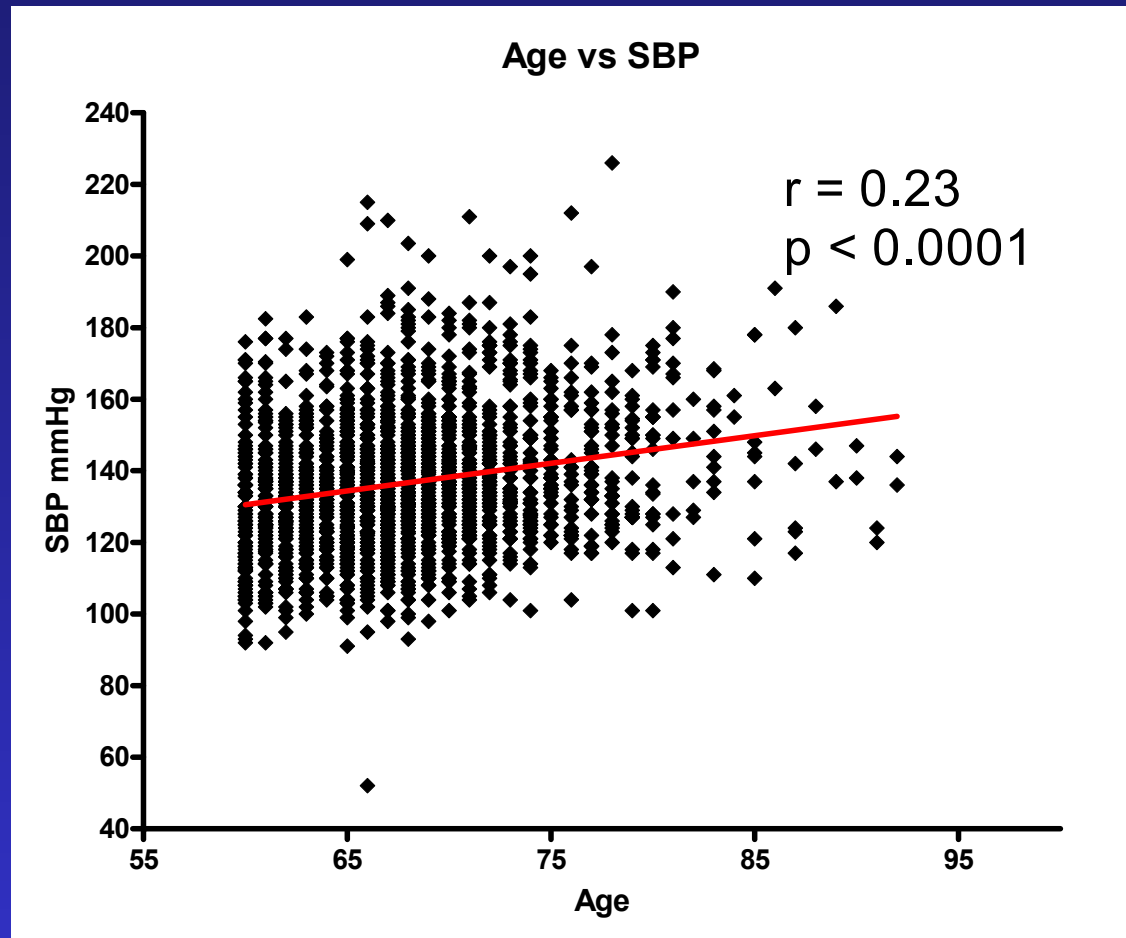
- The sign on the correlation coefficient is either positive or negative depending on whether:
 - The correlation is **positive**
 - For *increases* in one variable the other *increases*
 - The correlation is **negative**
 - For *increases* in one variable the other *decreases*

Correlation of age and SBP



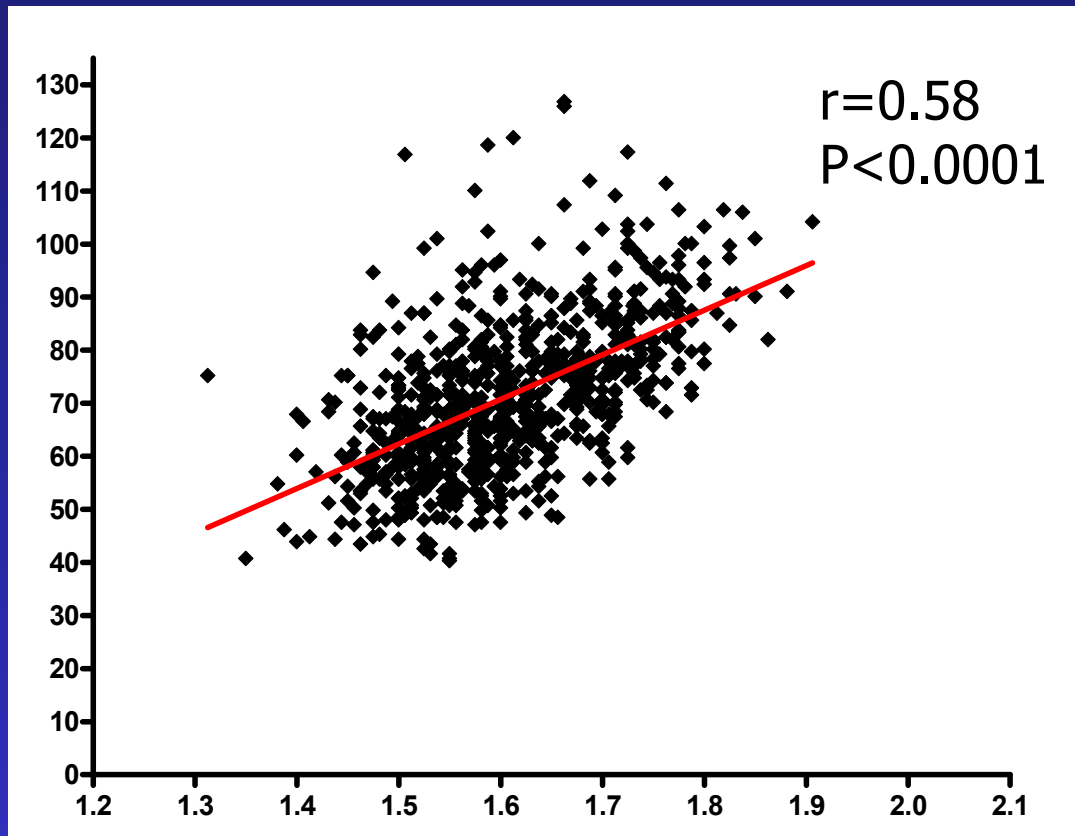
Is there a Linear relationship between these two variables?

Weak Correlation

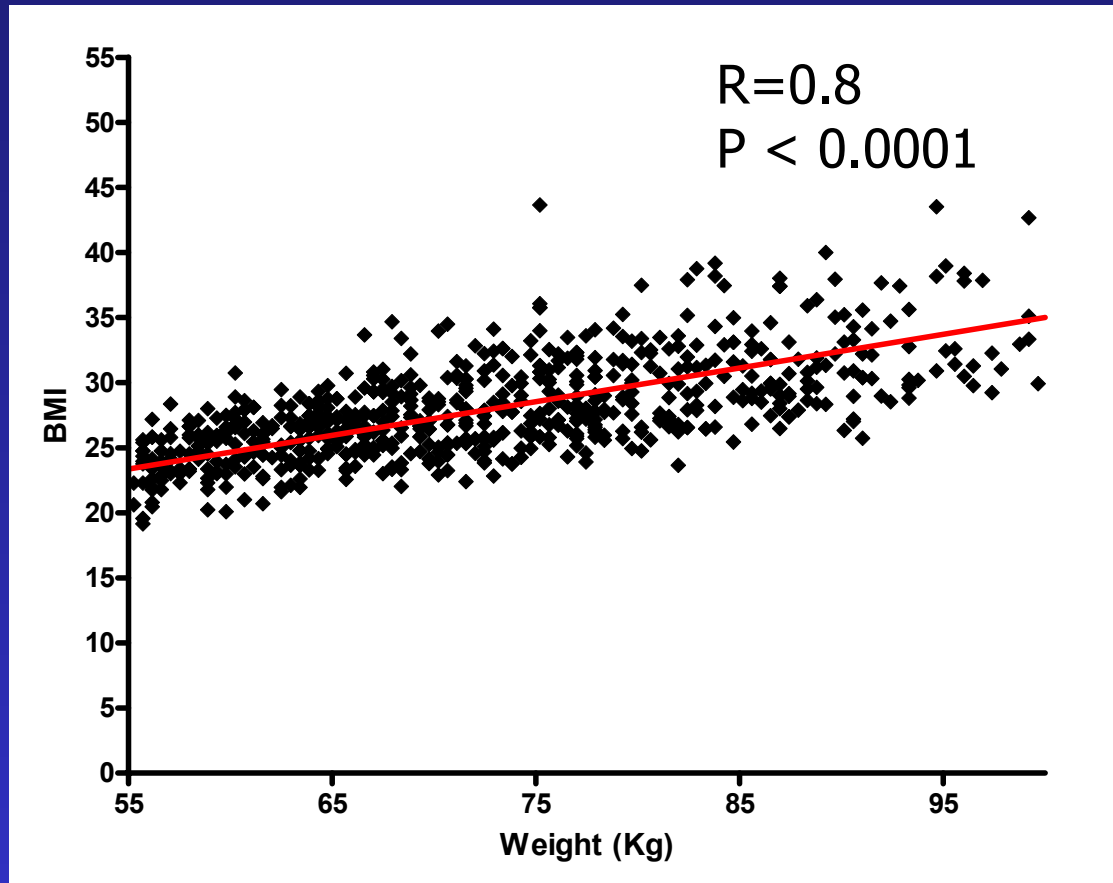


Height vs. Weight

Moderate Correlation



BMI and Weight Strong Correlation

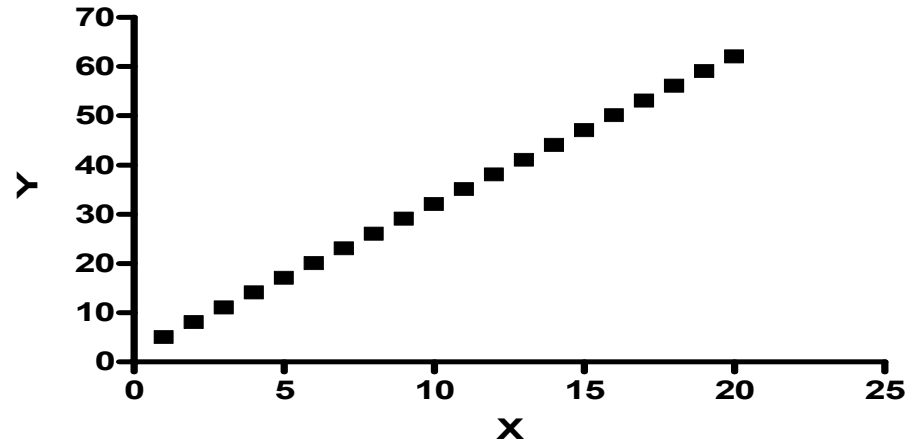


Correlation

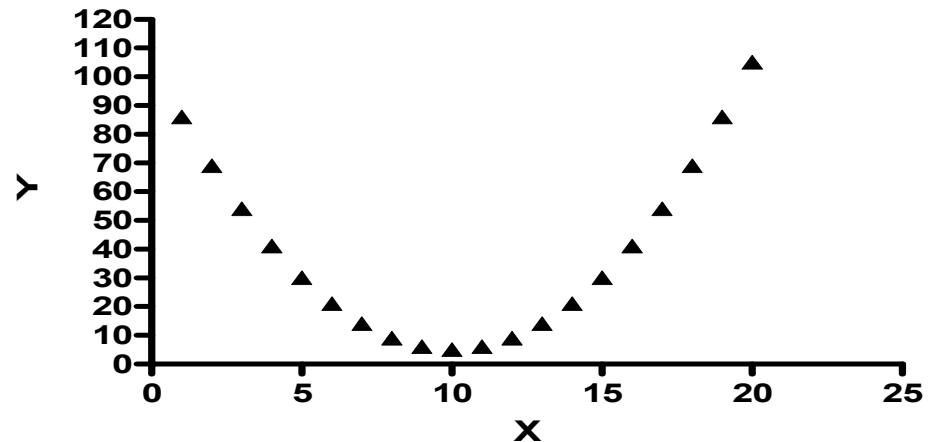
- Very important to remember that correlation analyses look for **LINEAR** relationships between the variables
- May fail to detect an important association if the relationship is not linear

Correlation

- Linear:
 $y = mx + b$



- Non linear:
 $y = mx^2 + b$



Weblink

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test		
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation			
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Correlation

- Correlation modeling identifies whether an association exists between two variables and how strong the association is
- But how much does one variable change when the other is changed
- What is the slope of the line

Regression

- Regression modeling can quantify this association
- There are two main applications of regression modeling:
 - To examine an effect while adjusting for confounding by another factor
 - What is the difference in blood pressure between the two cohorts, adjusting for age?
 - For prediction
 - If we know someone's age, can we predict what their blood pressure would be?

Examples

- If we run a linear regression to examine the difference in the systolic blood pressure between the two cohorts:

$$SBP = m_{\text{cohort}} \times \text{cohort} + b$$

Results of the linear regression

- The computer gives us the following result:

$$\text{SBP} = - 7 \times \text{cohort} + 140.9$$

- Here the variable "cohort" take on a 0 for cohort A and a 1 for cohort B
- The coefficient for "cohort" would give us the difference in the mean blood pressure between the two cohorts
- In this example cohort B (when the variable "cohort" = 1) has a SBP that is 7mmHg lower than cohort A (when cohort = 0)

Results of the linear regression

$$\text{SBP} = - 7 \times \text{cohort} + 140.9$$

- Notice that the intercept is the mean blood pressure in cohort A that we saw earlier (when the cohort variable = 0)

How do we adjust for confounding?

- If we have more than one variable in the regression model, the model will give you the effect of changes in each variable, while it holds the others constant
- $Y = m_1X_1 + m_2X_2 + \dots + m_zX_z + b$

Blood Pressure Example

Adjusting for Confounding

- Now if we added a term in our regression for cohort A or cohort B we would get this:

$$SBP = m_{age} \times age + m_{cohort} \times cohort + b$$

SBP adjusted for age

- The results of this regression are as follows:

$$\text{SBP} = 0.62 \times \text{age} - 3.4 \times \text{cohort} + 96.6$$

- So where before we saw a mean difference of 7 mmHg between the two groups, adjusting for age the the mean difference in only 3.4 mmHg

Multiple Regression

- The previous regression is an example of multiple regression
- The results yield the effect of each predictor variable adjusting for all the other predictor variables

How good is the model?

- If we know the values of the predictor (independent) variables
 - Age, BMI, Cohort
- How good are we at predicting the outcome (dependent) variable?
 - Systolic blood pressure

How good is the model?

- The *Coefficient of Determination* - R^2
- Is a measure of how much of the variability in the dependent variable is predicted by the model
- In other words how good are we at predicting Y (SBP in our example) if we know X (Age in our example)

How good is the model?

- The *Coefficient of Determination* - R^2
- Takes on the values between 0 and 1
- A value of 1 implies that if we know all the predictor variables we can perfectly predict all the values of the outcome variable
- Ideally want to see $R^2 > 0.5$

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test		
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation Linear Regression			
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Logistic regression

- Can be used similarly to linear regression
 - Examine an association
 - Control for confounding
- But for dichotomous outcomes
 - High blood pressure (Yes / no)
 - Death (yes / no)
 - Cured (yes / no)

Logistic regression

- Yields the odds ratio for each predictor
- What is the OR for hypertension (SBP > 140) for **each** increased year of age
- What is the OR for hypertension for **each** increased point of BMI
- What is the OR for hypertension for those in cohort B versus cohort A

Blood pressure example

- OR for Age 1.05
 - The increased odds of hypertension for each year of age
- OR for BMI 1.07
 - The increased odds of hypertension for each increased point of BMI
- OR for cohort B 0.77
 - The decreased odds of hypertension for those in cohort B compared to cohort A

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test		
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation Linear Regression	Logistic regression		
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Survival Analyses

- Analyses that include the time to an event occurrence are broadly referred to as survival analyses
- Death is not always the outcome of interest
- More correctly called “Time to Event Analysis”

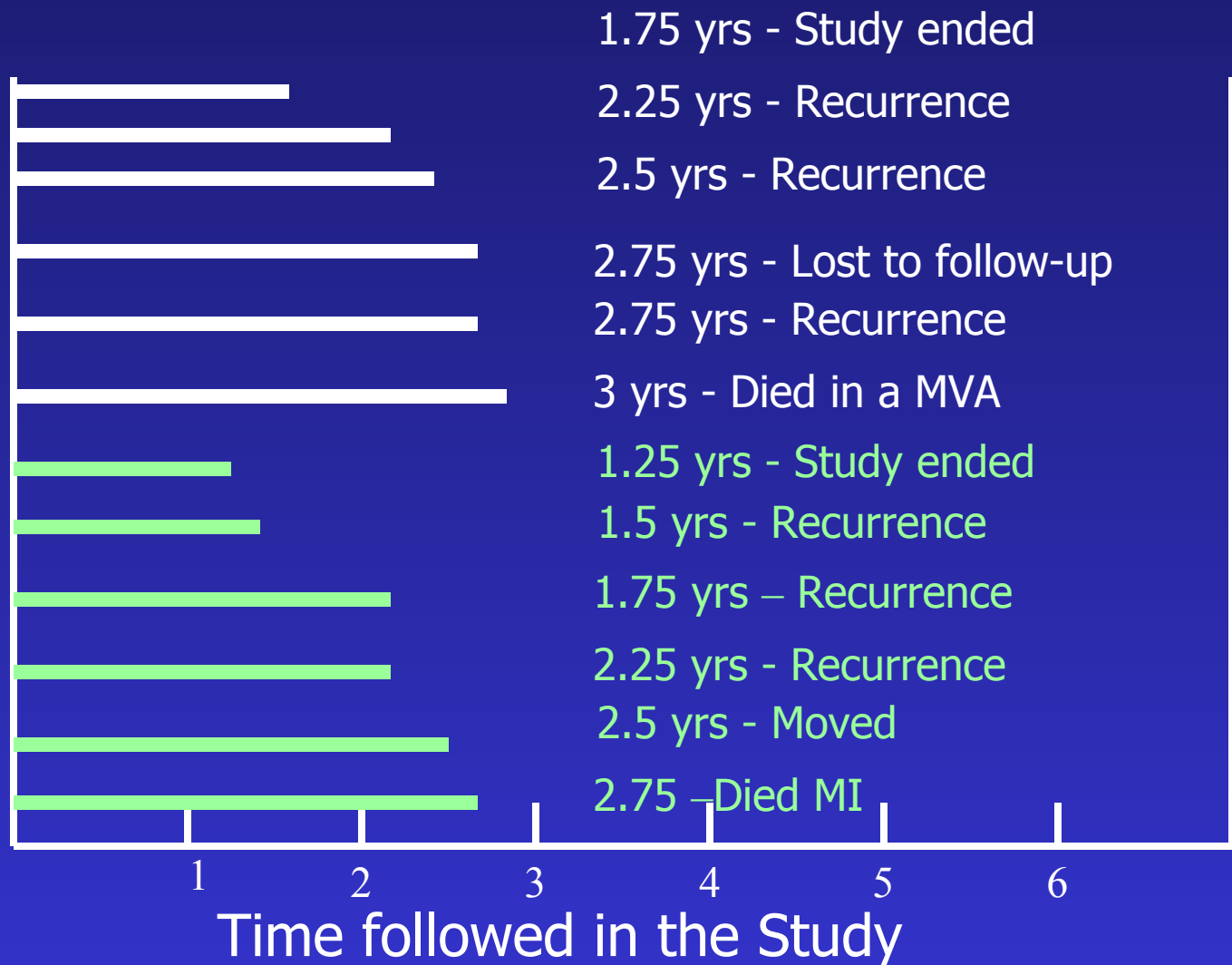
Survival Analysis

- Medical studies often involve this type of data:
 - Cancer recurrence
 - Discharge from the hospital
 - Wound healing
 - Death

Cancer Therapy Example



Cancer Therapy Example



Cancer Therapy Example

- You can clearly see that:
 - We have different follow-up for subjects in this study
 - Because of loss to follow-up or the end of the study not all events (Cancer recurrence) are observed
- This is known as *Censored Data*

Assumptions for survival analyses

- In order to analyze this type of data we need to make the following assumptions:
 - 1) Subjects that enter the study at different times have the same survival distribution
 - 2) The reason subjects are censored is unrelated to the outcome

How do we analyze this type of data?

- Person time analysis
 - 3 recurrences occurred in the new treatment group in 15 person – years of follow up
 - = 0.2 recurrences per person-year
 - = 20 recurrences per 100 person-years
 - 3 recurrences occurred in the old treatment group in 12.75 person – years of follow up
 - = 0.24 recurrences per person-year
 - = 24 recurrences per 100 person-years

Person time analysis

- These rates are called *Incidence Density Rates*
- If we divide one rate by the other we obtain the relative risk
- $RR = 0.2 / 0.24$
 $= 0.83$
- Those getting the new treatment regimen have a 0.83 times risk of recurrence
- We can calculate a P value and a confidence interval on the relative risk

Additional Assumption for Person-Time Analyses

- Constant Rate Assumption:
 - The risk of the event occurring remains constant over the study period
 - The risk for seeing the outcome is the same for one person followed for two years as for two people followed for one year

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)		
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test	Person-time analyses	
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation Linear Regression	Logistic regression		
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

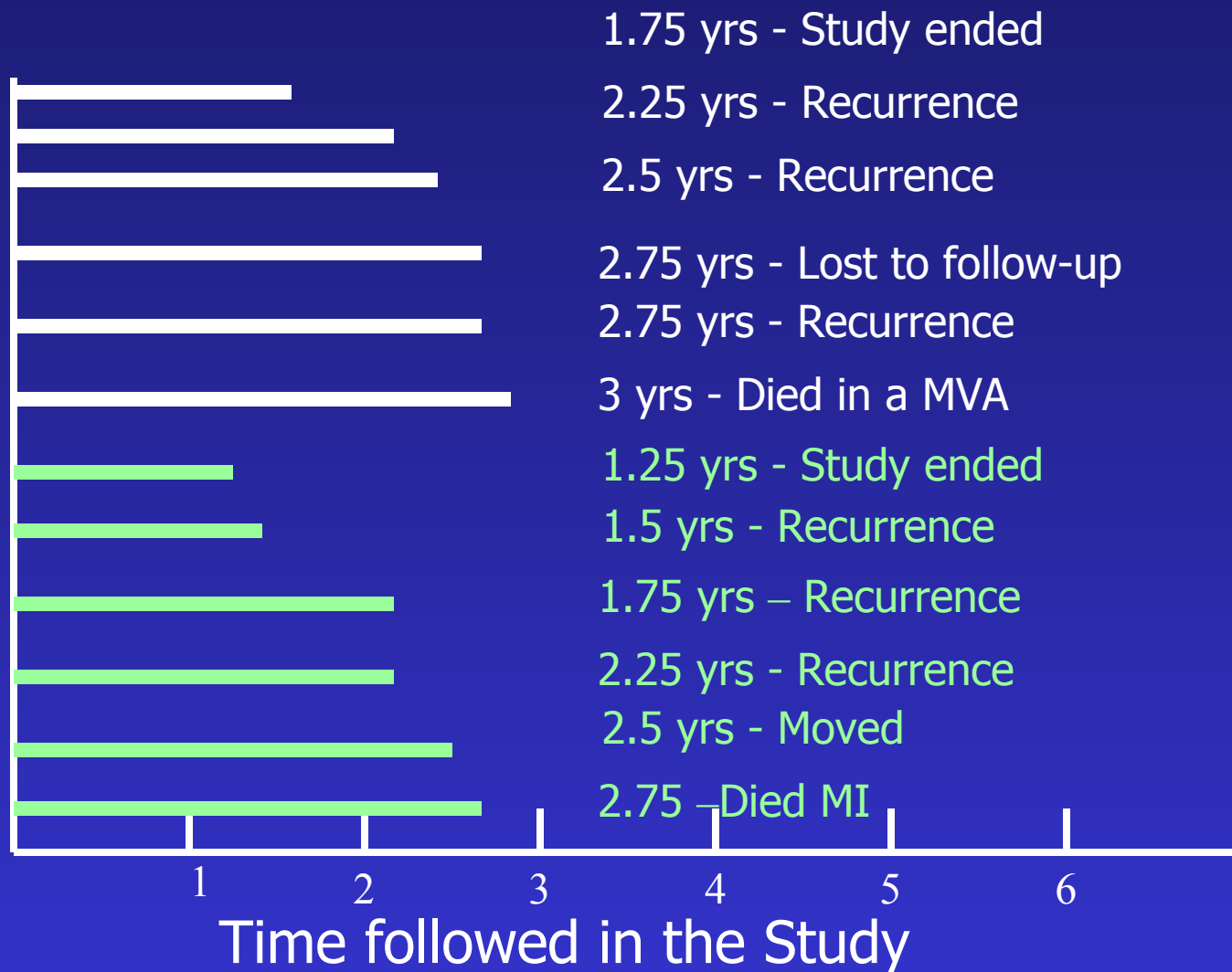
Survival Curve

- Plots the percent of individuals *at risk* for the endpoint at any given time point
 - At risk means those who are free from the endpoint at that time
- Can also plot the percent who have had the endpoint
- Does not rely on the constant rate assumption

Kaplan-Meier Plots

- A specific type of survival curve
- Calculates the proportion survival at each event occurrence (Cancer recurrence in our example)

Cancer Therapy Example



Cancer Therapy Example

Time (Years)	Outcome	# who had the event during period	# at risk at the start of the period	Proportion who did not have event during this period	Probability of survival until this period
1.75	Ended				
2.25	Recurrence				
2.5	Recurrence				
2.75	Lost to F/U				
2.75	Recurrence				
3	Died MVA				

Cancer Therapy Example

Time (Years)	Outcome	# who had the event during period	# at risk at the start of the period	Proportion who did not have event during this period	Probability of survival until this period
1.75	Ended	0	6	1.0	1.0
2.25	Recurrence				
2.5	Recurrence				
2.75	Lost to F/U				
2.75	Recurrence				
3	Died MVA				

Cancer Therapy Example

Time (Years)	Outcome	# who had the event during period	# at risk at the start of the period	Proportion who did not have event during this period	Probability of survival until this period
1.75	Ended	0	6	1.0	1.0
2.25	Recurrence	1	5	4/5	$1.0 * 4/5 = 4/5$
2.5	Recurrence				
2.75	Lost to F/U				
2.75	Recurrence				
3	Died MVA				

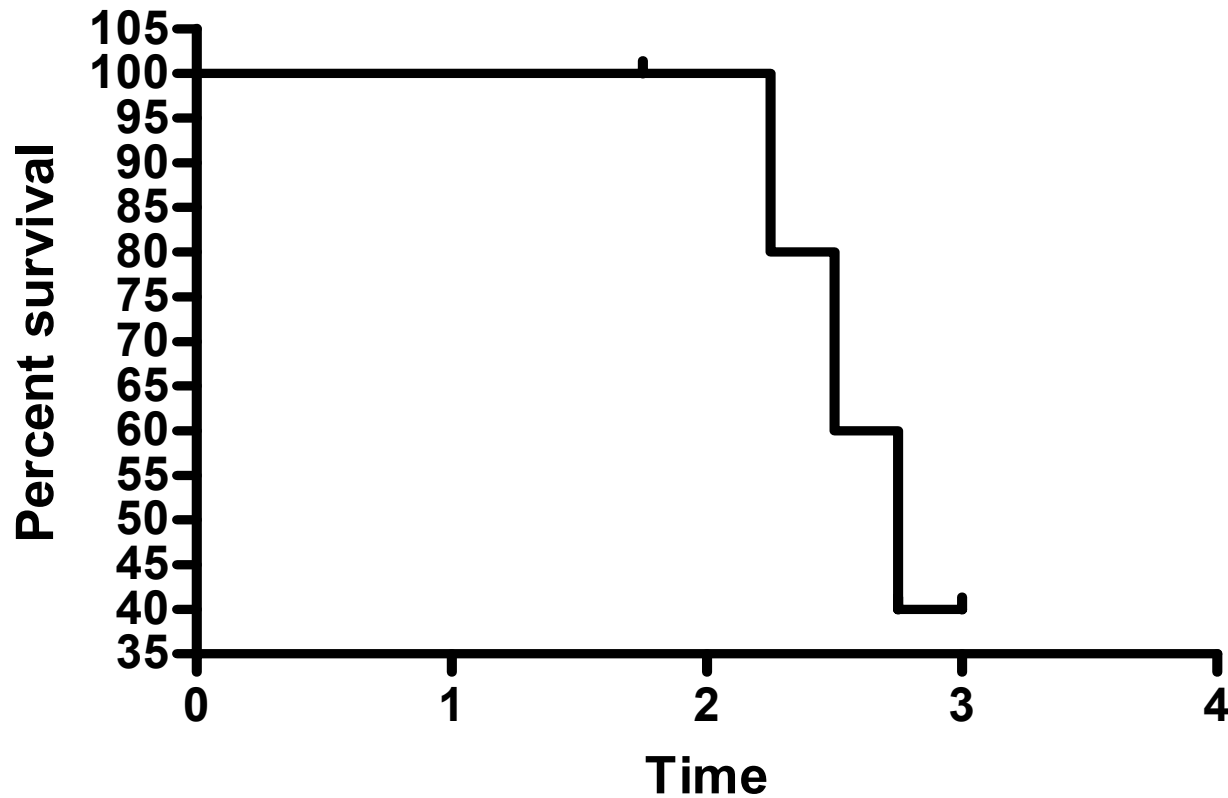
Cancer Therapy Example

Time (Years)	Outcome	# who had the event during period	# at risk at the start of the period	Proportion who did not have event during this period	Probability of survival until this period
1.75	Ended	0	6	1.0	1.0
2.25	Recurrence	1	5	4/5	$1.0 * 4/5 = 4/5$
2.5	Recurrence	1	4	3/4	$1.0 * 4/5 * 3/4 = 3/5$
2.75	Lost to F/U				
2.75	Recurrence				
3	Died MVA				

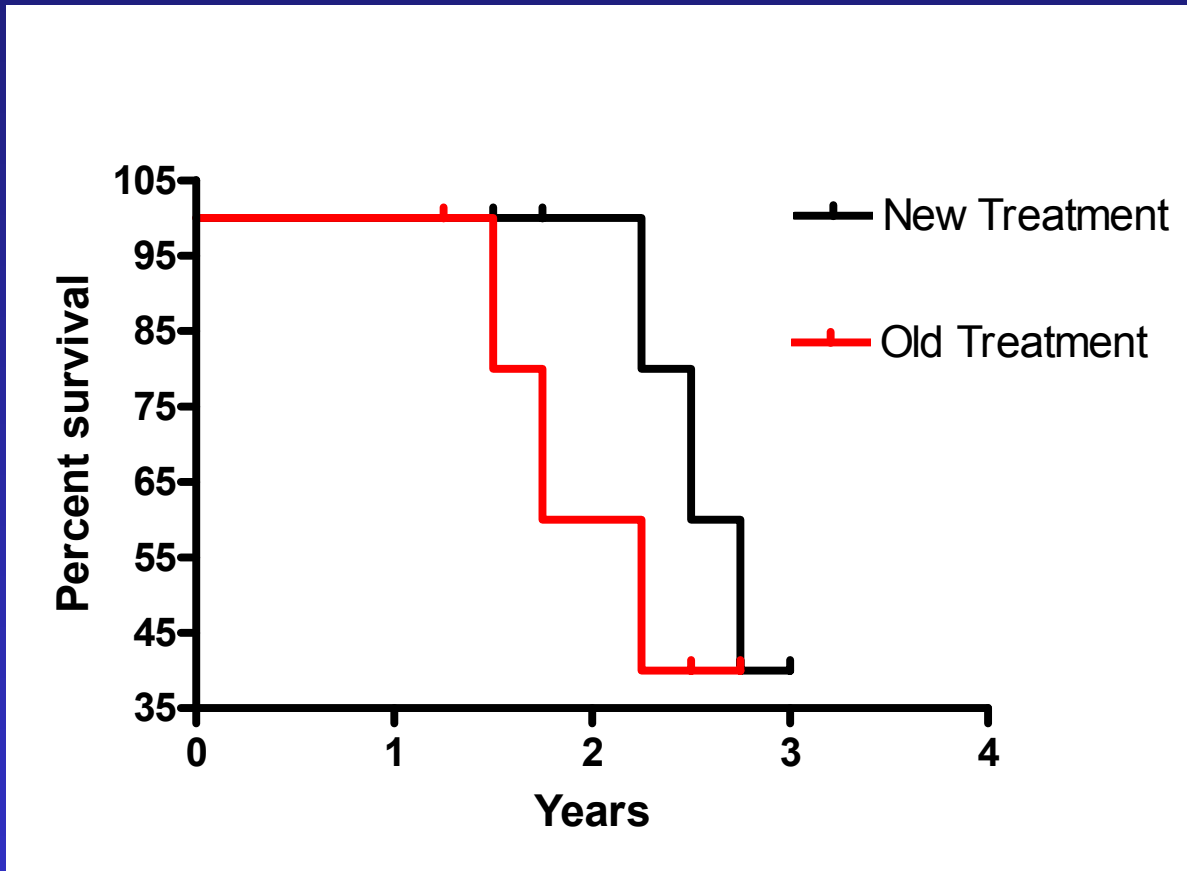
Cancer Therapy Example

Time (Years)	Outcome	# who had the event during period	# at risk at the start of the period	Proportion who did not have event during this period	Probability of survival until this period
1.75	Ended	0	6	1.0	1.0
2.25	Recurrence	1	5	$4/5$	$1.0 * 4/5 = 4/5$
2.5	Recurrence	1	4	$3/4$	$1.0 * 4/5 * 3/4 = 3/5$
2.75	Lost to F/U				
2.75	Recurrence	1	3	$2/3$	$3/5 * 2/3 = 2/5$
3	Died MVA	0	1	$1/1$	$2/5$

Kaplan-Meier Plot for New Treatment Group



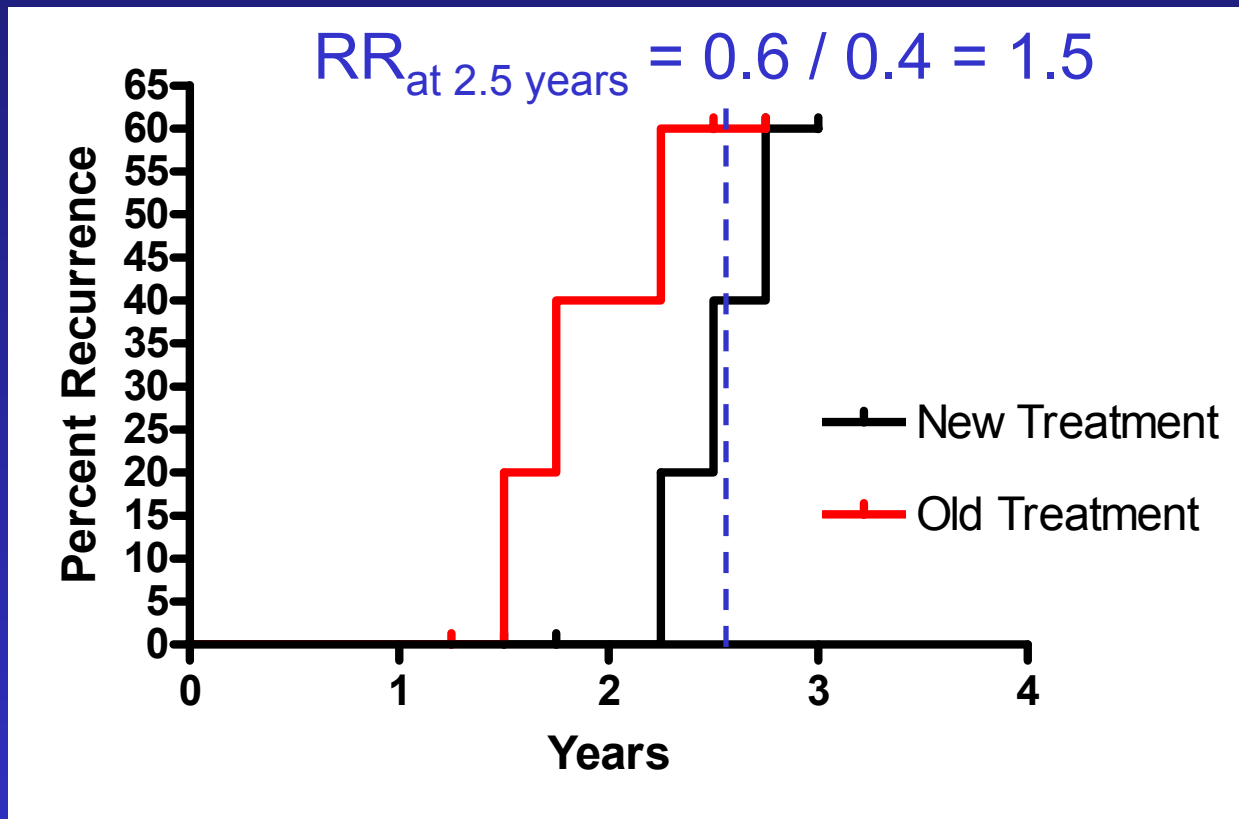
Kaplan-Meier Plot for Both Treatment Groups



Kaplan-Meier Plot

- Kaplan Meier curves display the cumulative survival
- If we instead plotted (1-survival) we would get a cumulative incidence plot
- The relative risk at any time point is the ratio of the cumulative incidence of the two curves

Kaplan-Meier Plot for Incidence



Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)	Kaplan – Meier Survival Curve	
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher's Exact test	Person-time analyses	
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation Linear Regression	Logistic regression		
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Proportional Hazards Regression

- Also known as Cox Proportional Hazards Regression
- A form of regression analysis in which the hazard rate is the dependent variable
- Does not rely on the constant rate assumption

Proportional Hazards Regression

- Begins with the assumption that for the outcome of interest, there is a "baseline" survival curve
- This can be thought of as the survival curve of a hypothetical "completely average" individual
- This is the only part of the model that depends on time

Proportional Hazards Regression

- The proportional hazard model assumes that the effect of each predictor variable is to multiply the “baseline” hazard rate by a constant
 - In other words the predictors either increase the survival (bend the “average” survival curve up) or decrease the survival (bend the “average” survival curve down)
- This constant is a relative risk
- Knowing the relative risks we could construct a “customized” survival curve for any combination of predictor values

Proportional Hazards Regression

- Like linear and logistic regression proportional hazards regression enables us to:
 - Examine an association between a predictor variable and an outcome
 - Control for confounding

An example of Proportional Hazards Regression

- What is the effect of the following on risk of death?
 - Age
 - Female sex
 - Coronary artery disease
 - Congestive heart failure
 - Diabetes
 - Stroke

Effect on risk of death from The Proportional Hazards Regression

- Age RR 0.99 (0.98, 1.0)
- Female sex RR 0.74 (0.63, 0.86)
- CAD RR 1.3 (1.1, 1.5)
- CHF RR 2.2 (1.8, 2.7)
- Diabetes RR 1.2 (1.0, 1.5)
- Stroke RR 2.0 (1.6, 2.4)

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)	Kaplan – Meier Survival Curve	
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher’s Exact test	Person-time analyses	
Compare two dependent groups	Difference in means (CI) Paired sample T-test			
Association between variables (control for confounding)	Correlation Linear Regression	Logistic regression	Cox Proportional Hazards regression	
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		

Non-Parametric Data

- All of the previous techniques assume that the data are approximated by a normal distribution (parametric)
- If this is not true then need to use other techniques
 - Ordinal data (Pain on a scale of 1 to 10)
 - Counts

Non-Parametric Techniques

- Typically use the **rank** of the observations rather than the observations themselves
- Retains the relative size of the observations but does not make any assumptions about the sample distributions

Non-Parametric Techniques

- Procedure in general:
 - Rank all observations according to magnitude: 1-smallest etc...
 - Compute the sum of the ranks (T) in the smaller sample
 - Compare this sum T to a critical value of T to test the null hypothesis

Non-Parametric Techniques

- The previous is an example of
 - The Wilcoxon Rank Sum test
 - Mann-Whitney U Test

Non-Parametric Techniques

Parametric Procedure	Nonparametric Procedure
Independent sample t-test	Wilcoxon Rank Sum test Mann-Whitney U test
Paired sample t-test	Wilcoxon signed rank test
Correlation Coefficient	Spearman rank correlation coefficient

Summary Table

Goal	Measurement Data	Categorical Data	Survival Data	Non-Normal Data
Describe the Population	Mean (CI)	Proportion (CI)	Kaplan – Meier Survival Curve	Median
Compare two independent groups	Difference in means (CI) Two-sample T-test	RR, OR (CI) Chi-Square Test Fisher’s Exact test	Person-time analyses	Wilcoxon Rank Sum test Mann-Whitney U test
Compare two dependent groups	Difference in means (CI) Paired sample T-test			Wilcoxon signed rank test
Association between variables (control for confounding)	Correlation Linear Regression	Logistic regression	Cox Proportional Hazards regression	Spearman rank correlation coefficient
Compare more than two groups	Analysis of variance (ANOVA)	Chi-Square Test		Kruskal-Wallis Test