Analyzing Data- Where to Begin

Statistical Test Basics

Orthopaedic Residency Research Program 02/11/2022

Susan Thapa, MPH, PhD
Department of Orthopaedic Surgery

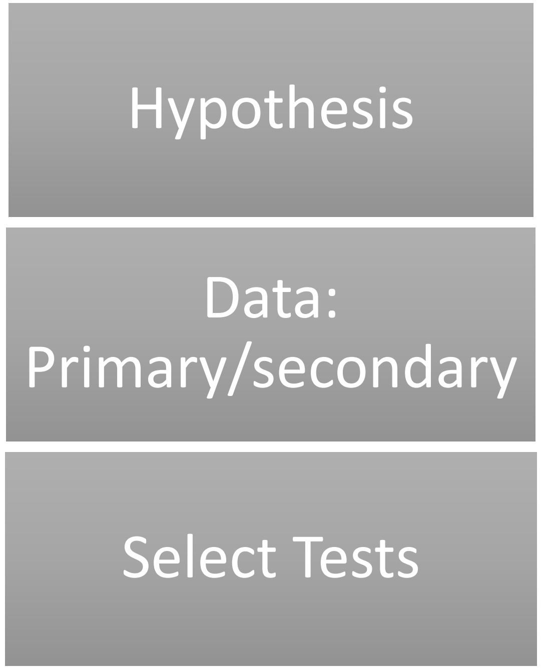
STATISTICAL TESTS

Uses:

- Compare groups
- Test hypothesis

Depends on:

- Data type- predictors, outcomes
- Data distribution

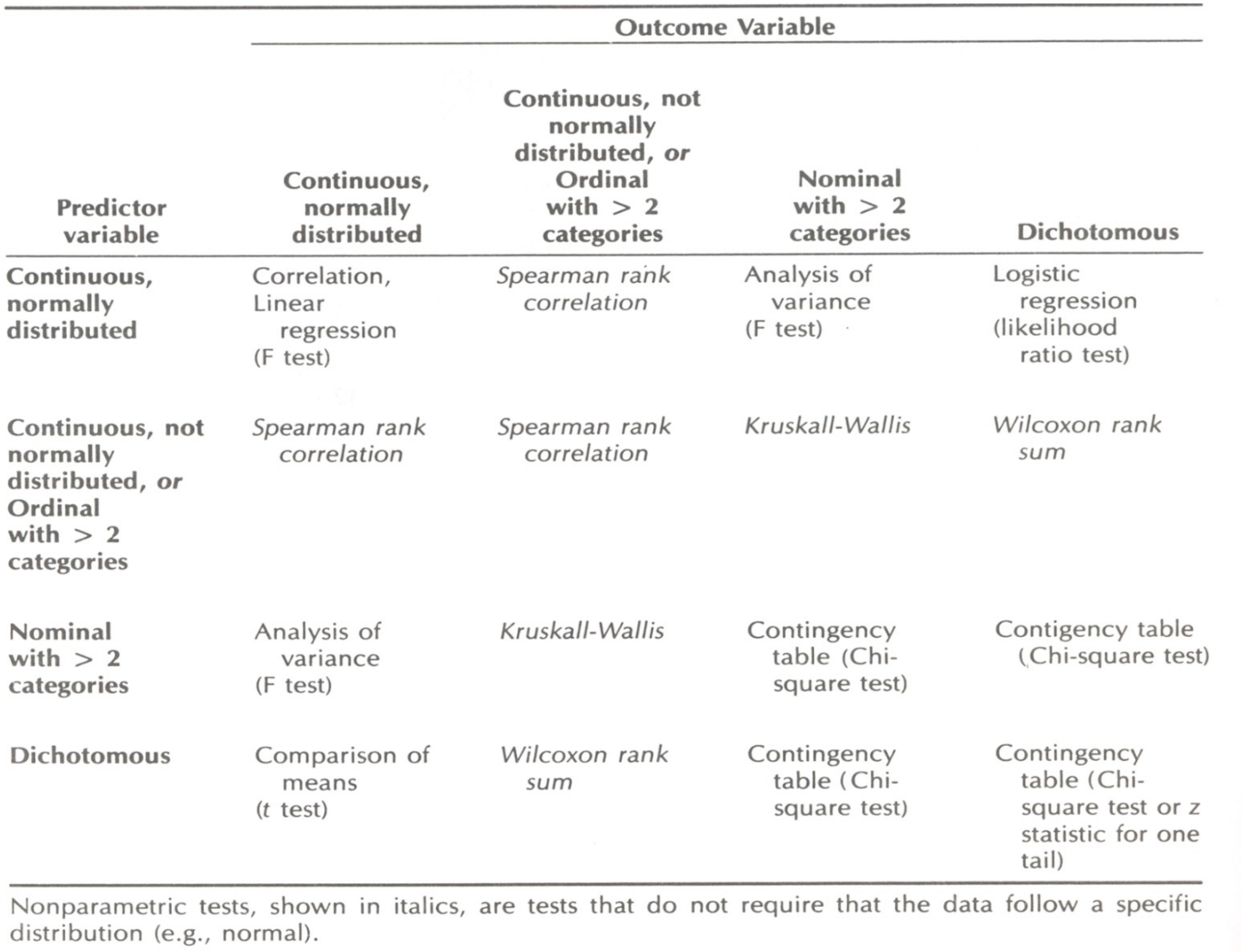


Hypothesis

Data: Primary/secondary

Select Tests

CHOOSING THE RIGHT TEST



	Outcome Variable			
Predictor variable	**Continuous, normally distributed**	**Continuous, not normally distributed, *or* Ordinal with > 2 categories**	**Nominal with > 2 categories**	**Dichotomous**
Continuous, normally distributed	Correlation, Linear regression (F test)	*Spearman rank correlation*	Analysis of variance (F test)	Logistic regression (likelihood ratio test)
Continuous, not normally distributed, *or* Ordinal with > 2 categories	*Spearman rank correlation*	*Spearman rank correlation*	*Kruskall-Wallis*	*Wilcoxon rank sum*
Nominal with > 2 categories	Analysis of variance (F test)	*Kruskall-Wallis*	Contingency table (Chi-square test)	Contigency table (Chi-square test)
Dichotomous	Comparison of means (*t* test)	*Wilcoxon rank sum*	Contingency table (Chi-square test)	Contingency table (Chi-square test or *z* statistic for one tail)

Nonparametric tests, shown in italics, are tests that do not require that the data follow a specific distribution (e.g., normal).

TYPES OF DATA

Continuous:

- Blood pressure, age, BMI

Discrete: data split into different categories

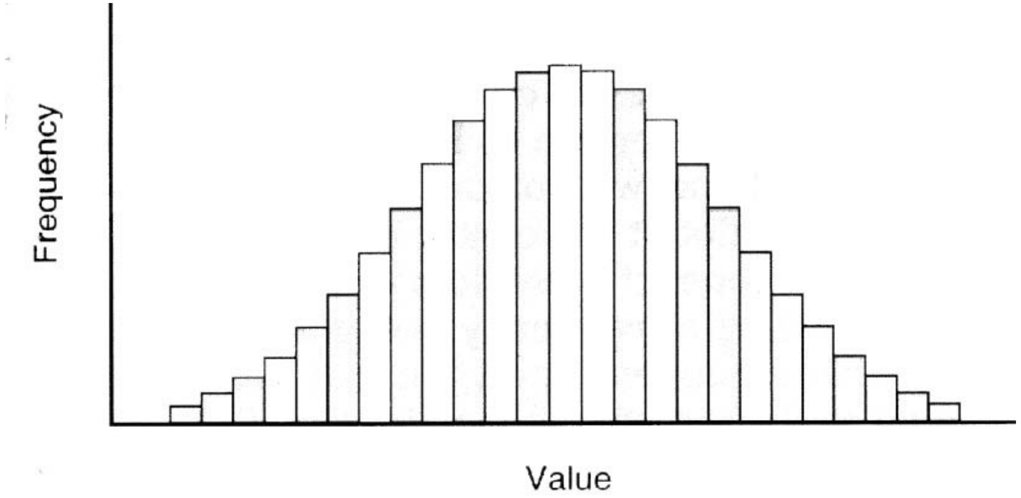
- Dichotomous: binary-yes/no; treatment/control, surgery failure vs success
- Ordinal: Age groups, Pain scale, Performance scale
- Nominal: Race, Gender, marital status

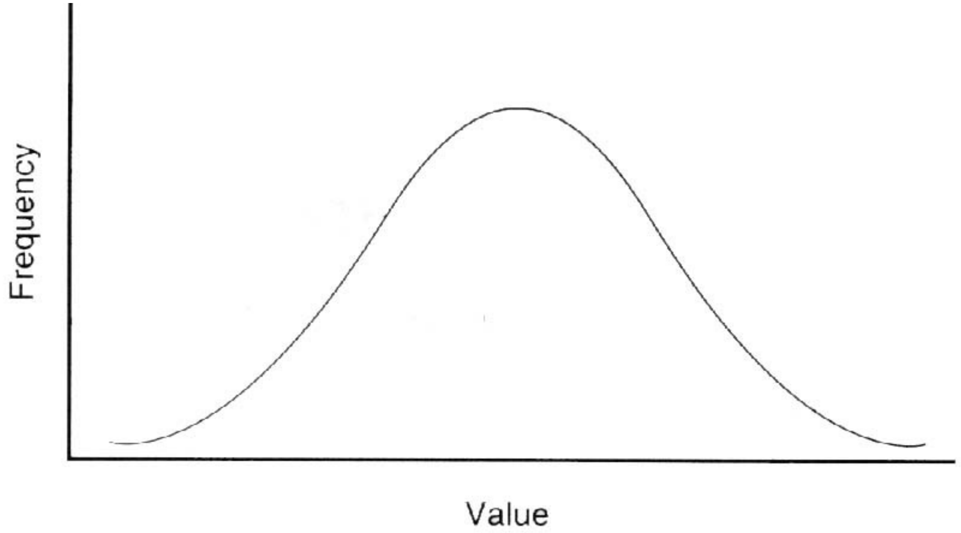
  Categories are named but without specific orders

Continuous Data

DISTRIBUTION OF CONTINUOUS DATA

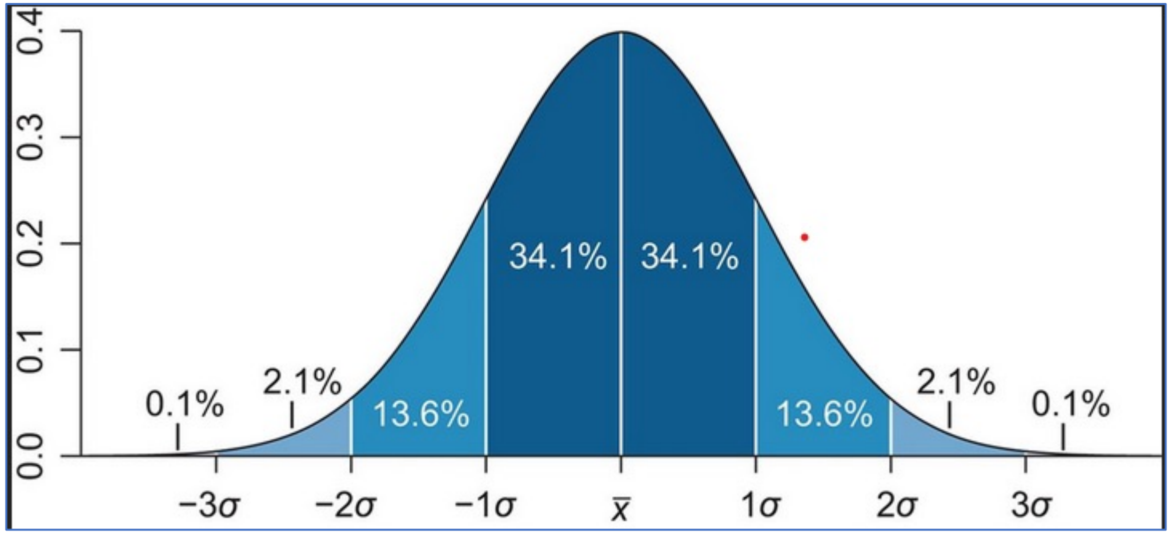
- Continuous data often represented as histograms
- Data ordered into bins often of equal size
- Shows the relative frequency of the data within each bin
- Allows selection of statistical tests
- Testing of assumptions of statistical tests





NORMAL FREQUENCY DISTRIBUTION

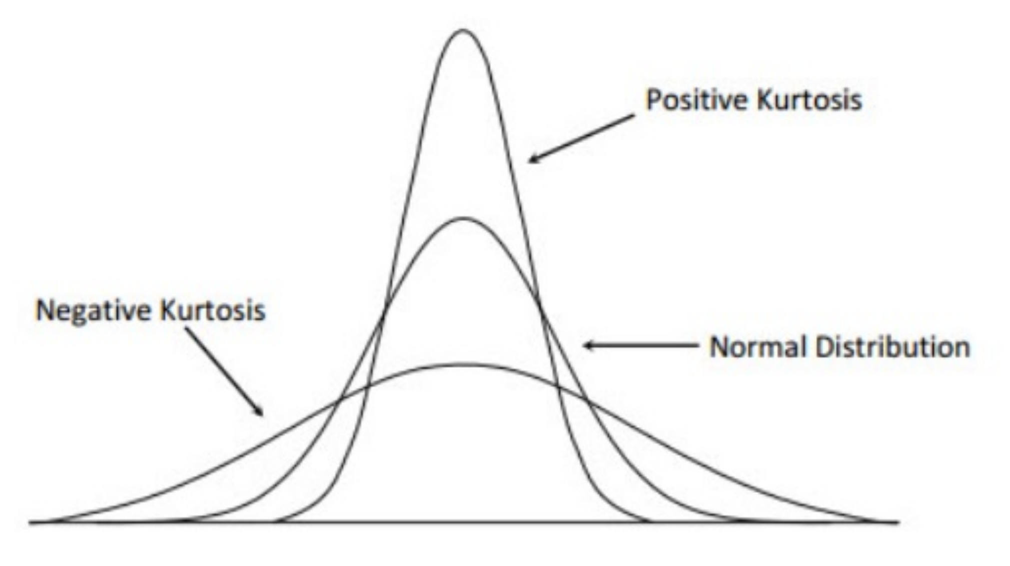
- Mean = Median = Mode
- Symmetrical: Skew = 0
- Kurtosis (vertical stretch)= 3

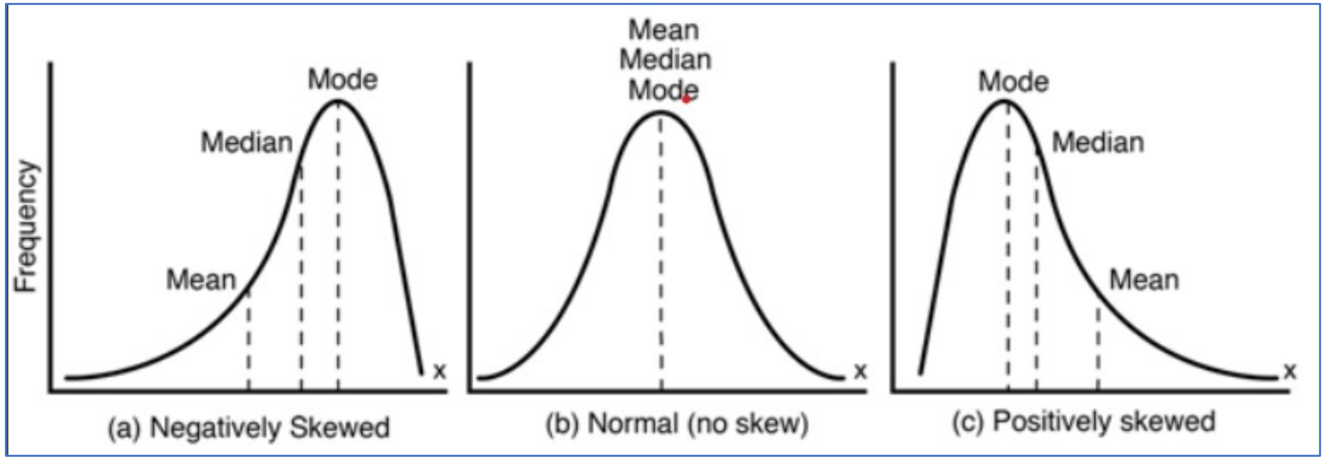


SHAPE OF FREQUENCY DISTRIBUTION

Skewness: (unbalanced) horizontal stretching

Kurtosis: vertical stretching





(a) Negatively Skewed (b) Normal (no skew) (c) Positively skewed

T-TEST and ANOVA

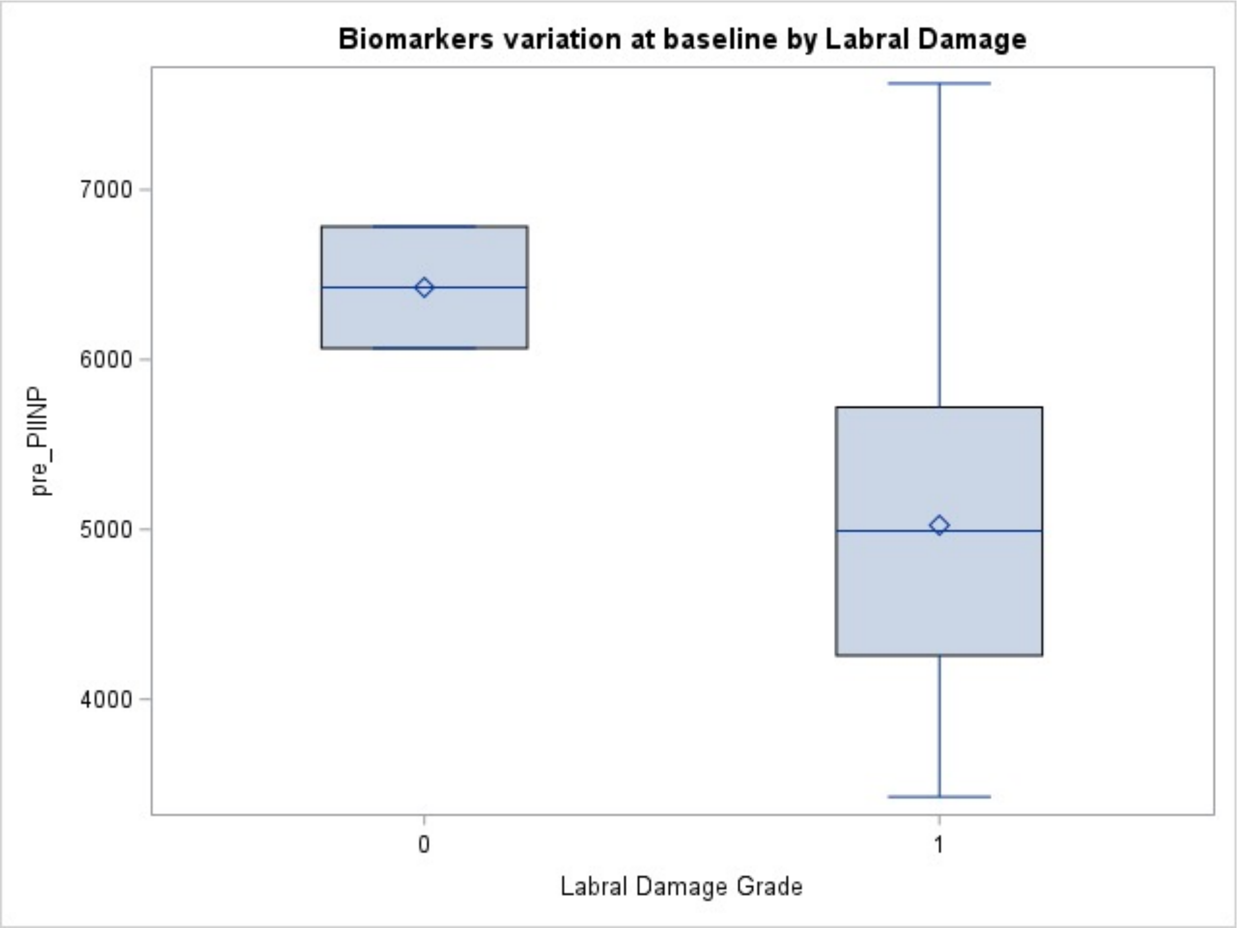
When to use:

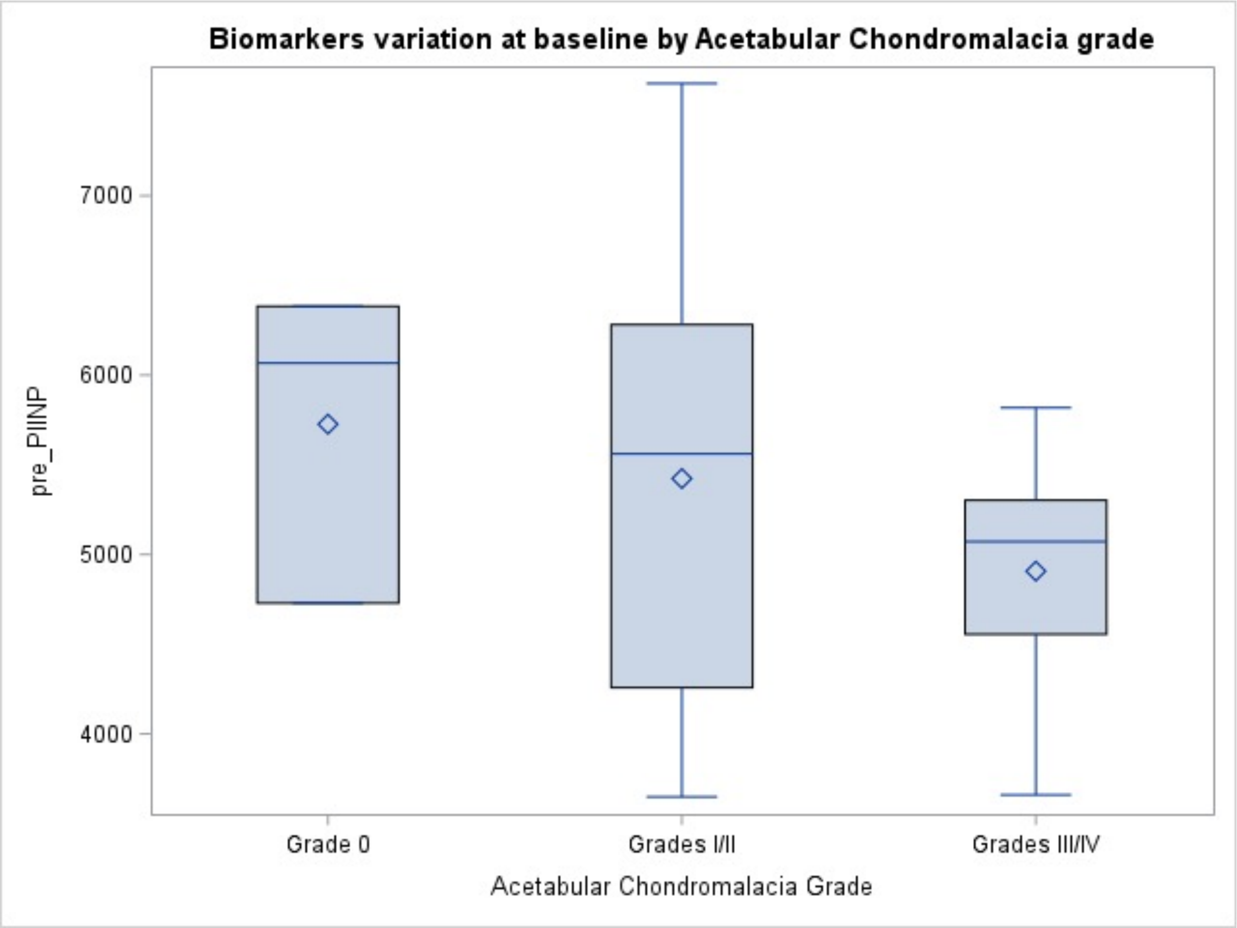
- Outcome-continuous; Predictor: Categorical
- Normal distribution
- Number of categories
  - T-test- predictors have 2 categories
  - ANOVA- predictors have >2 categories

Non-Parametric Alternatives

- T-test- e.g.: Mann-Whitney test
- ANOVA- e.g.: Kruskal-Wallis test

T-TEST and ANOVA





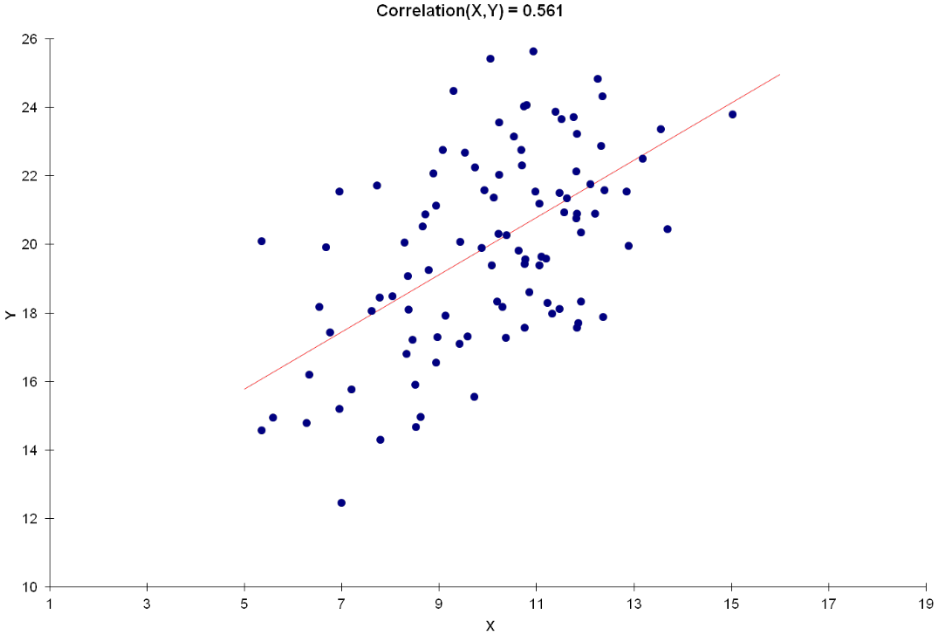
CORRELATION

Measures the strength of the linear relation between two continuous variables

Measures the tightness of a cluster about the fitted line

Correlation Coefficient

- Values range from -1 to +1
- Positive relation: positive coefficient
- Inverse/negative relation: negative coefficient
- 0: no correlation



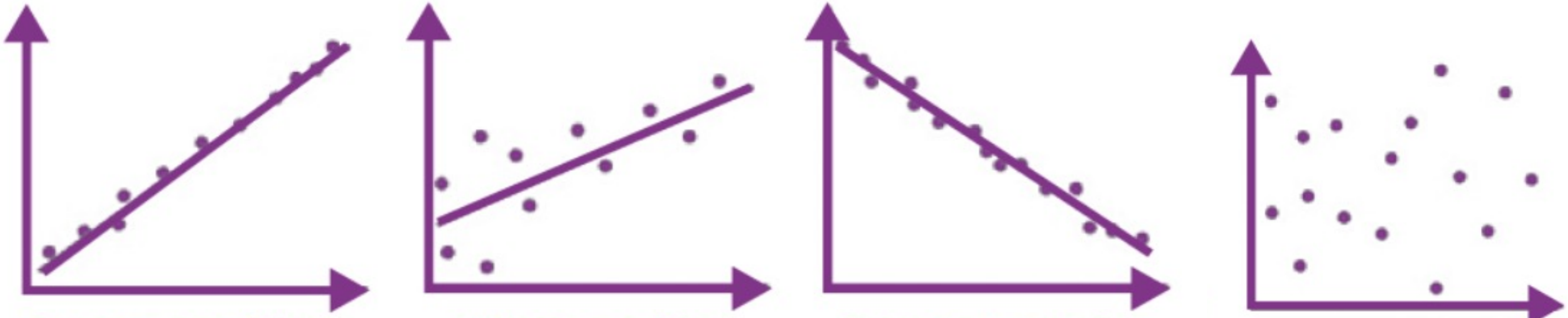
Methods:

- Pearson’s- Parametric
- Spearman’s- Non-parametric

Limitation

- Do not handle nonlinear relationships accurately
- Non-linear relationship may be characterized as null relationship

CORRELATION



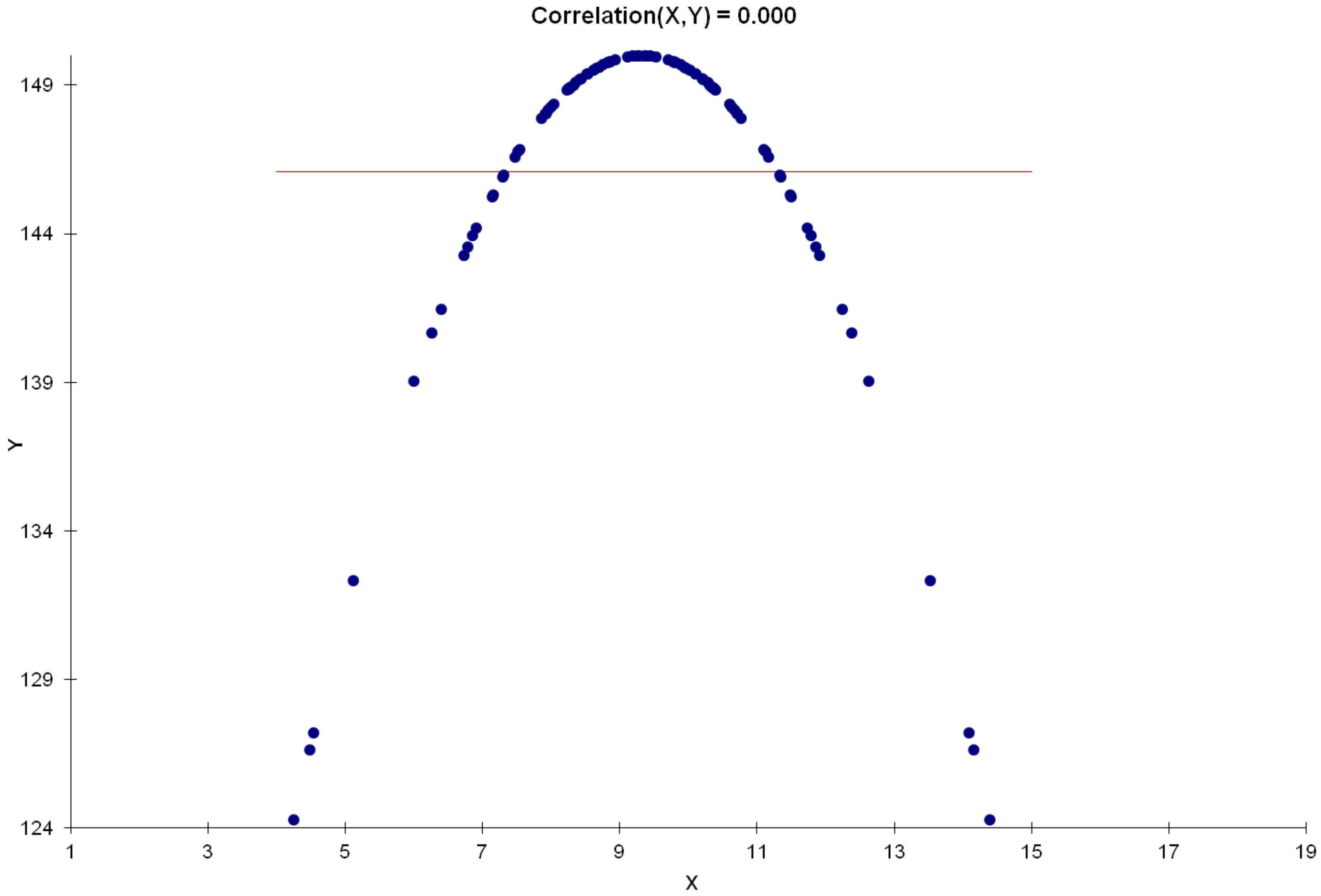
Strong positive correlation

Weak positive correlation

Strong negative correlation

No correlation

Non-Linear


Correlation(X,Y) = 0.000
149
144
139
134
129
124
Y
1
3
5
7
9
11
13
15
17
19
X

LINEAR REGRESSION

Variables

- Continuous outcome
- One predictor (simple linear regression)
- Two or more predictors/independent variables (Multivariate)

Uses

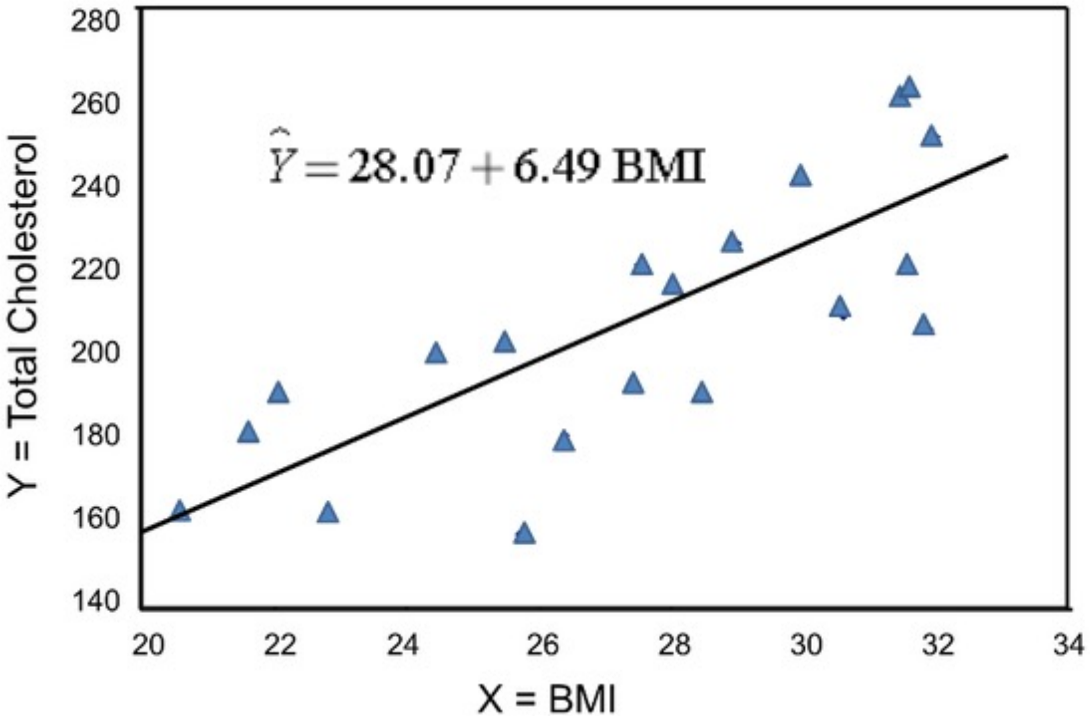
- Prediction
- Hypothesis testing
- Modeling Causal Relation

LINEAR REGRESSION

- **Example**: Assessing the association between BMI and total cholesterol
- **Regression equation:**

  Y = β0 + β1X1 + β2X2 + β3X3 + e

- **β coefficient:** directly used to estimate effect size
- **R2**- Variance explained by the model/independent variables



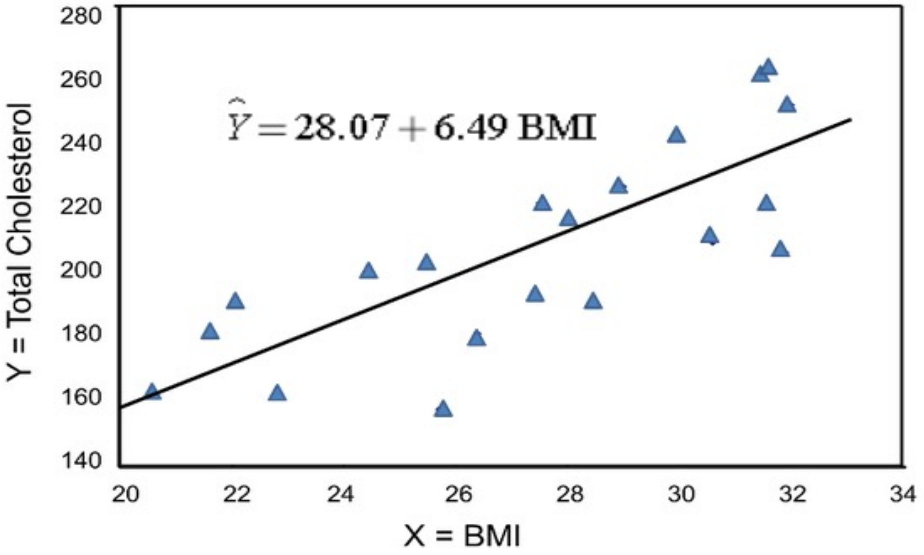
LINEAR REGRESSION ASSUMPTIONS

Independence of samples

Linear relation between dependent and predictors

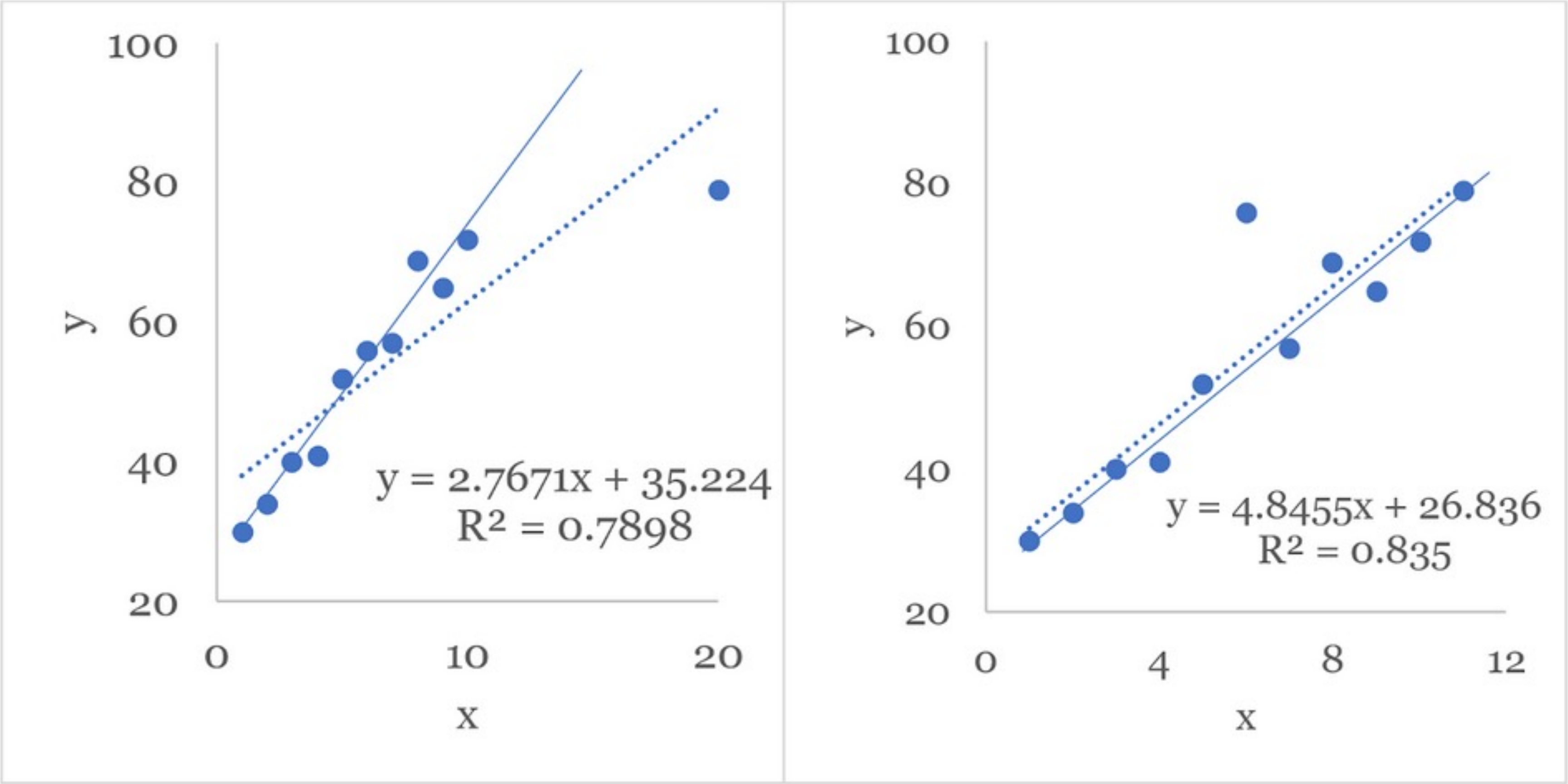
Normality of residuals

Homoscedasticity: stable amount of variance throughout range of values



OUTLIERS

- Outliers: very far above or below mean (extreme in the x or y axis)



Categorical/Discrete data

ODDS RATIOS, RISK RATIOS, HAZARDS RATIOS

Odds Ratios (OR)

- Case control designs
- Cohort/Follow-up studies when outcomes are rare (<10% prevalence)
- Method- Logistic Regression
- OR=1 no association

Risk Ratios (RR)

- Cohort/Follow-up studies when outcomes are common (>10% prevalence)
- Retrospective or Prospective cohorts
- Method- Poisson/Negative binomial Regressions
- RR=1 no association

Hazard Ratios (HR)

- Time-to-event data (*denominator is total follow-up time not total patients*)
- Cohort/Follow-up studies- retrospective or prospective
- Method- Cox Regression
- HR=1 no association

CONTINGENCY TABLES

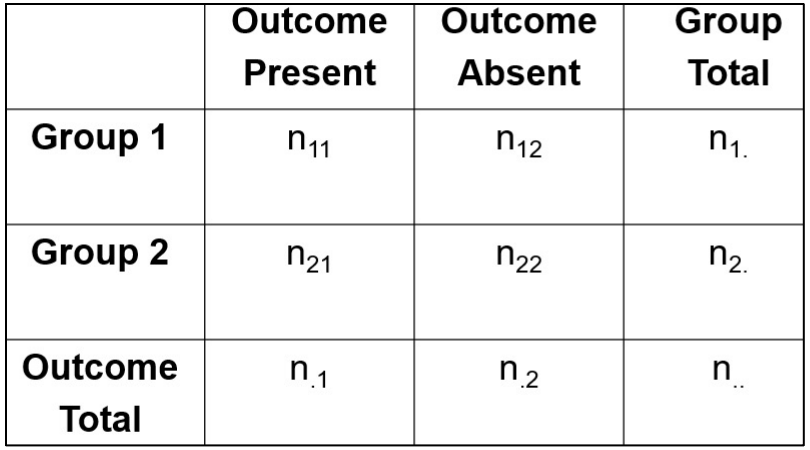
Uses:

- Unadjusted OR/RR

$$OR = \frac{odds_1}{odds_2} = \frac{n_{11} / n_{12}}{n_{21} / n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

$$RR = \frac{Risk1}{Risk2} = \frac{n_{11}/n_1}{n_{21}/n_2}$$

- Chi-sq tests- significance tests
- One predictor at a time; no adjustment

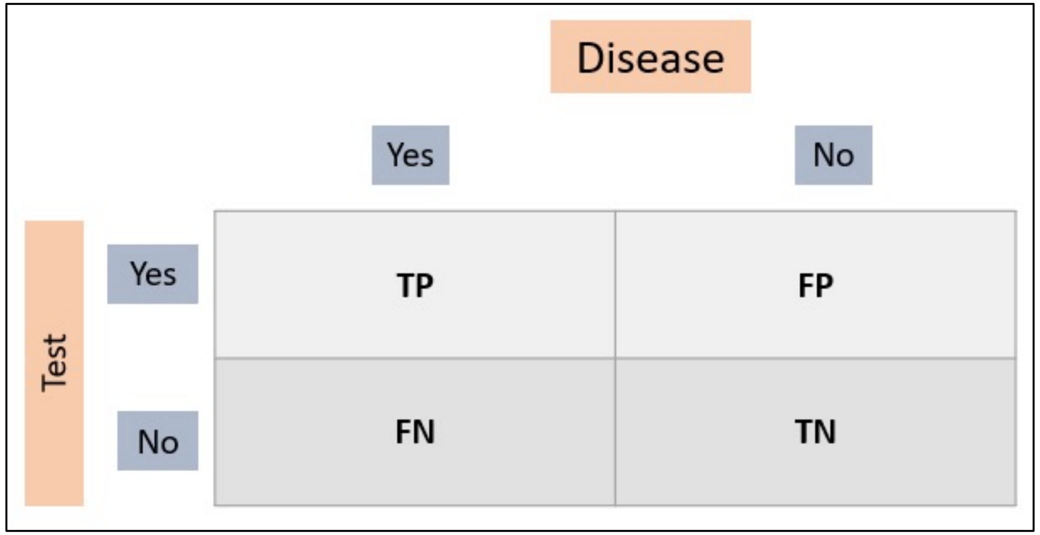


	Outcome Present	Outcome Absent	Group Total
Group 1	n_{11}	n_{12}	$n_{1.}$
Group 2	n_{21}	n_{22}	$n_{2.}$
Outcome Total	$n_{.1}$	$n_{.2}$	$n_{..}$

CONTINGENCY TABLES

Uses:

- Sensitivity/Specificity, etc. in diagnostic tests



		Disease	
		Yes	No
Test	Yes	TP	FP
	No	FN	TN

TP: True Positive, FP: False Positive
TN: True Negative, FN: False Negative

- Sensitivity (SN)
  - % with disease who test positive
  - = TP/(TP+FN)

- Specificity (SP)
  - % without disease who test negative
  - = TN/(FP+TN)

- Positive predictive value (PPV)
  - % positive test results that are true positives
  - = TP/(TP+FP)

- Negative predictive value (NPV)
  - % negative test results that are true negatives
  - = TN/(FN+TN)

LOGISTIC REGRESSION

Variables

- Categorical outcome
  - e.g.: Surgery failure/success
  - Ordinal/nominal- ordinal logistic regression

Uses

- Prediction
- Hypothesis testing
- Modeling Causal Relation

.

LOGISTIC REGRESSION

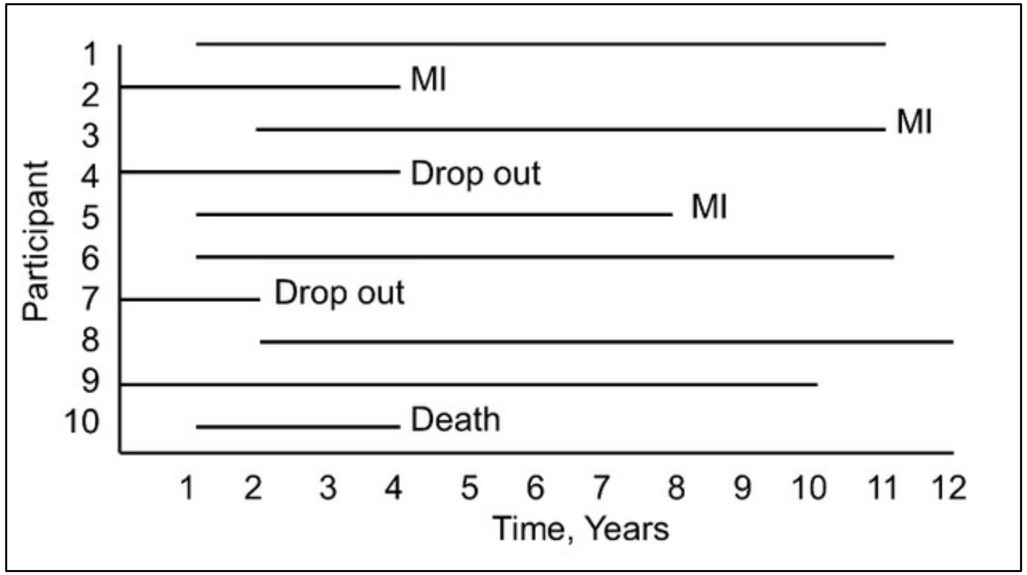
- **Regression equation:**

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

- **Effect Size: e^{β}** (Odds Ratio) used to estimate effect sizes
- **C- statistics**- Discriminatory power of the model
- Often continuous predictors are dichotomized at “cut-points” chosen to maximize discriminatory power

COX REGRESSION/SURVIVAL ANALYSIS

- Categorical outcome
- Follow-up time available and varies between observations/study participants
- Effect Sizes: e^{β} (Hazard Ratio) used to estimate effect sizes



COX REGRESSION/SURVIVAL ANALYSIS

- Whether or not a participant suffers the event of interest during the study period
- The follow up time for study participants

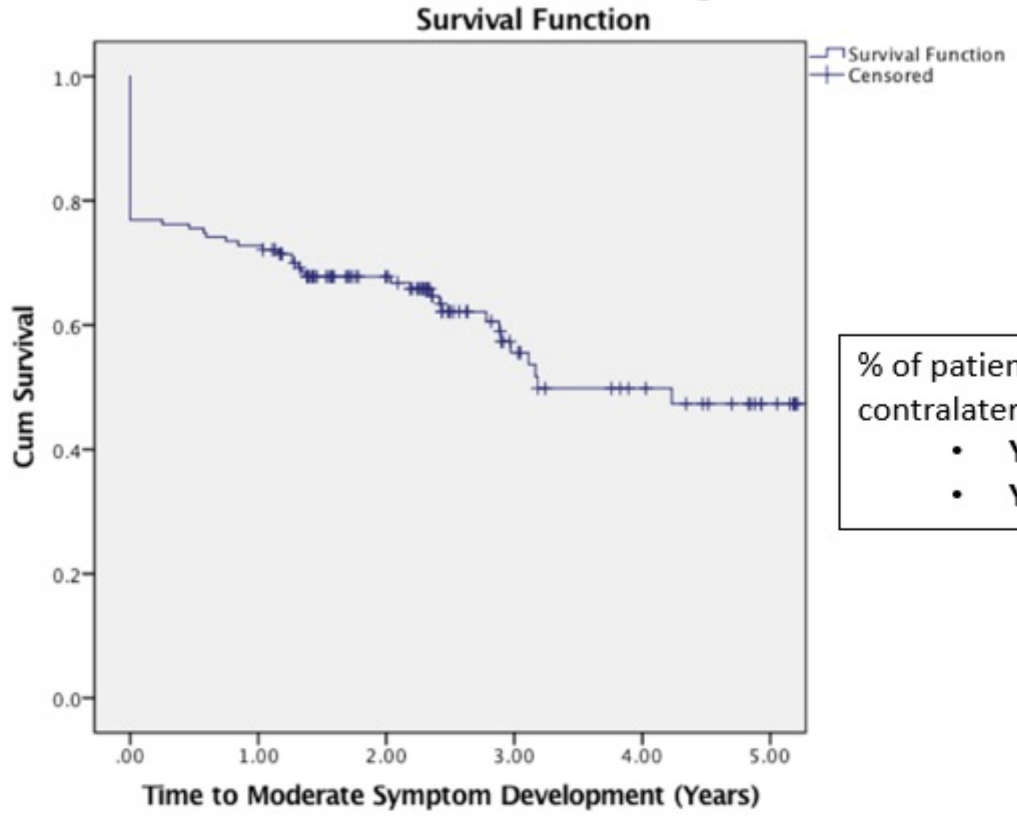
Survival Analysis Terms:

Time-to-event:

- The time from entry into a study to development of outcome

Censoring:

- Lost to follow up
- Drop out of the study
- Study ends before death/outcome

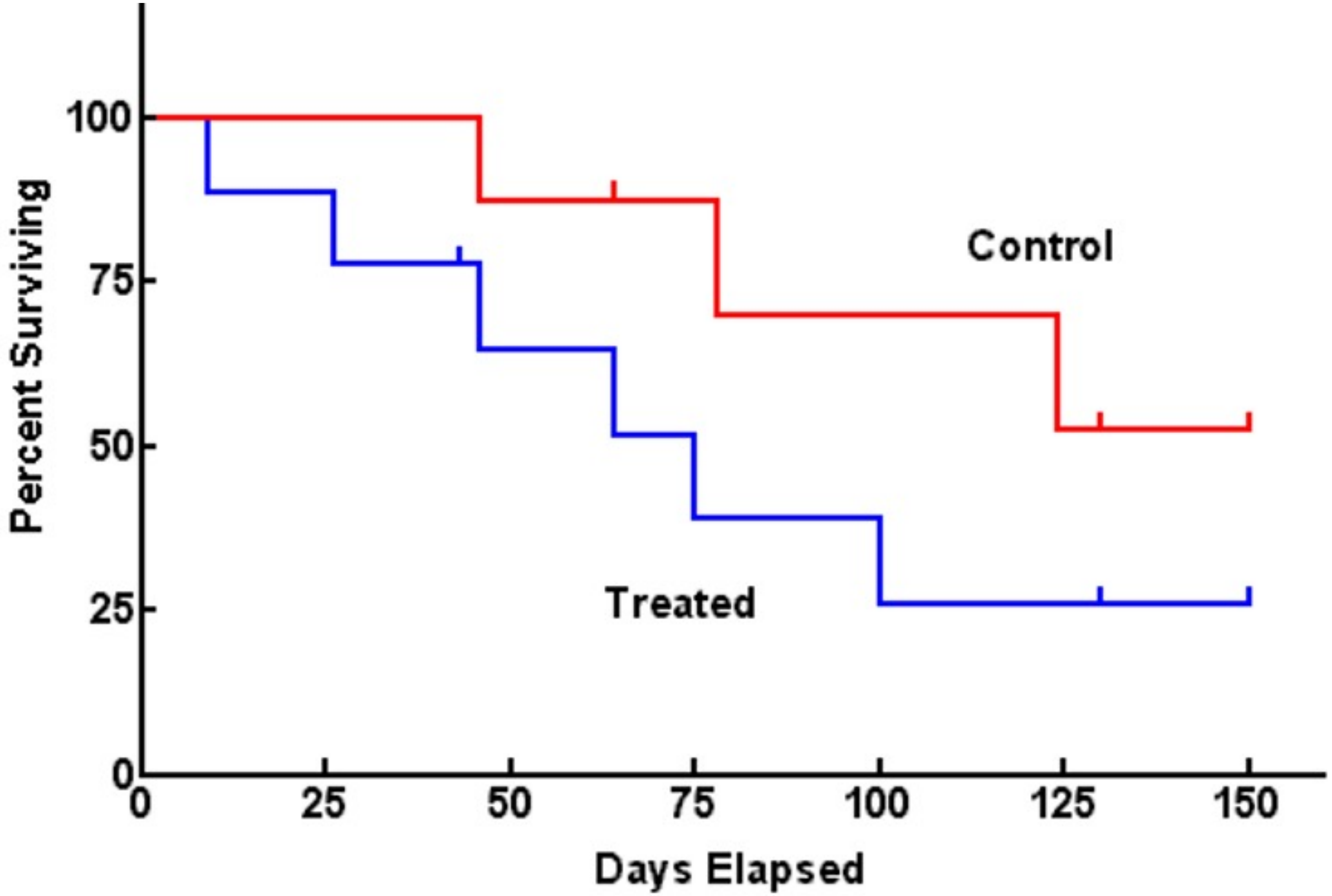


% of patients asymptomatic on contralateral hip:

- **Year 3**: 56%
- **Year 4**: 48%

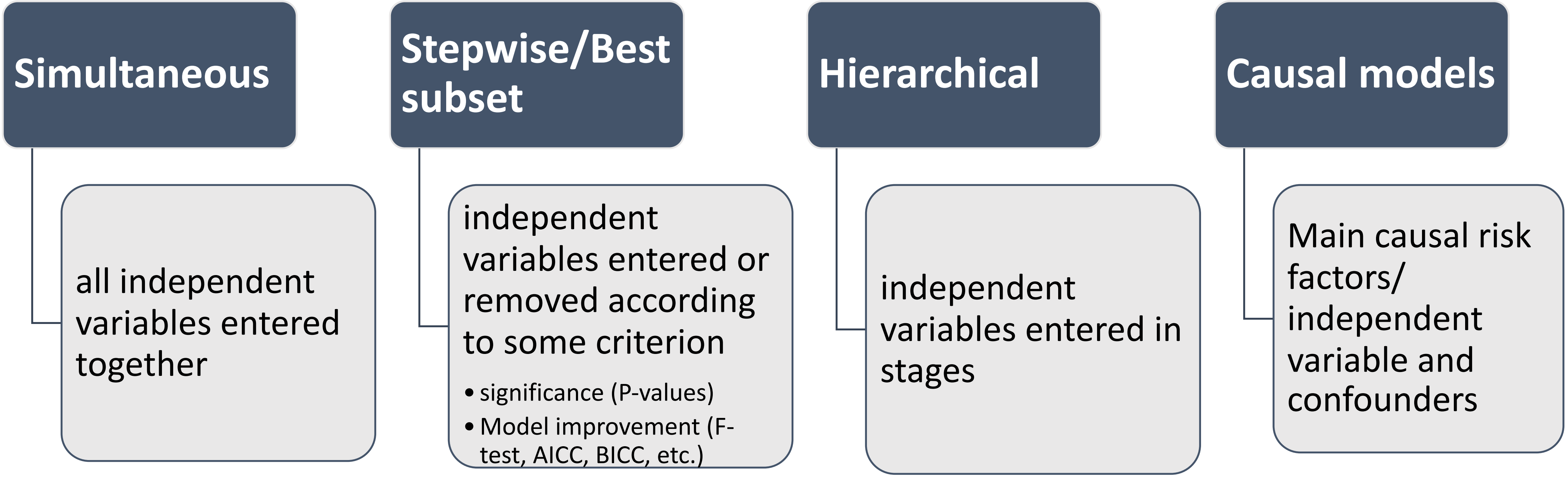
Fig: Kaplan-Meier analysis of symptom-free survival of the contralateral hip (*Louer et al, 2020*)

SURVIVAL ANALYSIS- Comparing 2 groups



Log-rank test:
Test of significance for difference between treated and control

BUILDING REGRESSION MODELS



Simultaneous

all independent variables entered together

Stepwise/Best subset

independent variables entered or removed according to some criterion

- significance (P-values)
- Model improvement (F-test, AICC, BICC, etc.)

Hierarchical

independent variables entered in stages

Causal models

Main causal risk factors/ independent variable and confounders

CORRELATED OUTCOMES, REPEATED MEASURES, MATCHING

Additional adjustments to regression models- e.g.:

- Repeated Measures ANOVA
- Conditional Logistic Regression- matching
- Generalized estimating equations (GEE)- correlated outcomes
    - Logistic
    - Poisson
    - Neg binomial
- Cox Regression for Clustered data

Summary

- Data type
- Data distribution
- Study Design
- Additional data characters:
    - Correlated
    - Matched
    - Repeated

Questions?