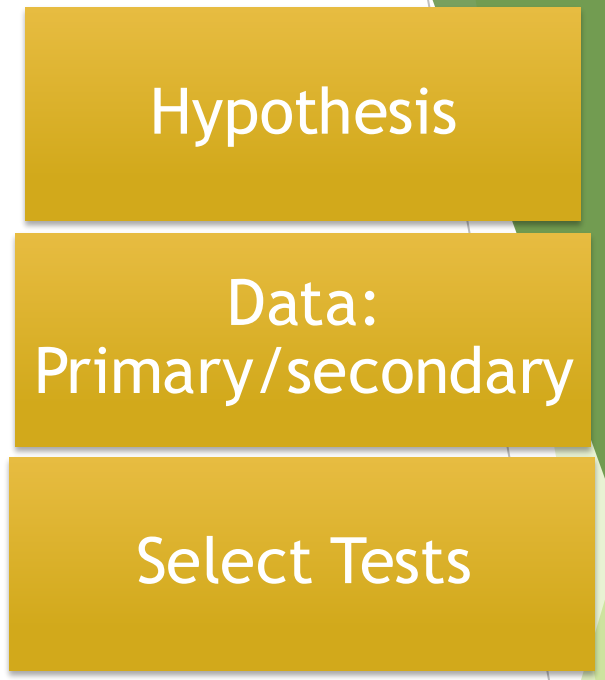
Analyzing Data Basics- Statistical tests

Ling Chen, PhD, MSPH
Institute for Informatics, Data Science and Biostatistics

Clinical Research



Uses:

- Make inference:

Compare groups, effects of treatment

- Prediction

Statistical modeling to predict clinical outcome for a future patient

Depends on:

- Data type of outcomes
- Data distribution of outcomes

TYPES OF DATA

Continuous:

- Blood pressure, age, BMI, Proximal Junctional Angle, PROMIS outcomes

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, employment status

  Categories are named but without specific orders

Be wary of non-independence among samples

- *Absurd Example*: You have two groups of patients and want to "prove" that cholesterol is higher in one group. You do the following:

  1. Select one subject from each group
  2. Take one blood sample from each subject
  3. Break each sample into 100 subsamples
  4. Yell "eureka" when a t-test reveals a significant difference between the two "groups", each of which has N = 100.

- You would never do the above because the lack of independence of the subsamples is obvious.

Be wary of non-independence among samples

- The **lack of independence will often be a problem** in experimental studies when there are **several samples selected from the same individual.** In such settings, the data analysis should always address this issue.

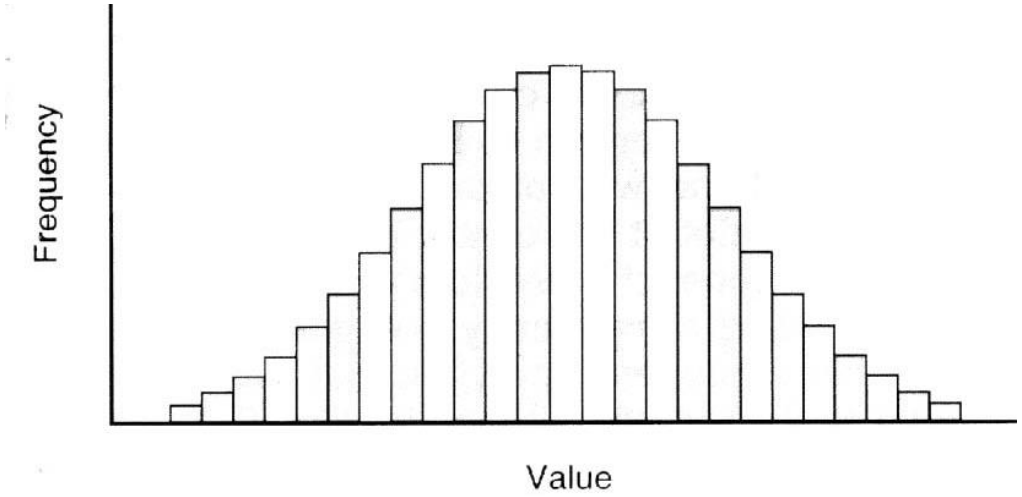
Choosing the right test

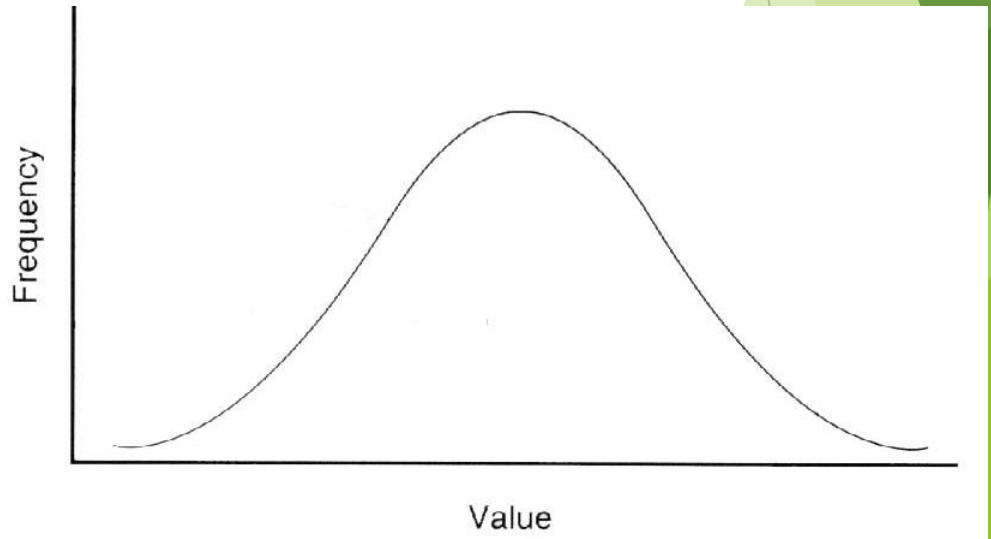
Outcome variable/ Dependent variables	Correlation between observations		Assumptions violated
	Independent /No correlation	Correlated	
Continuous	<u>Normal distribution</u> Independent T-Test ANOVA Linear correlation- Pearson's Linear regression	<u>Normal distribution</u> Paired T-test Repeated Measures ANOVA Linear mixed model	<u>Non-parametric/Non-normal</u> Mann Whitney test Kruskal- Wallis test Spearman correlation
Categorical	Chi-square test Logistic regression (OR) Poisson regression (RR, PR) Negative binomial regression (RR,PR)	McNemar test Conditional logistic regression GEE Modeling	Exact tests for small samples
Time-to-event	Hazard Ratio- Cox regression Kaplan Meier statistics/Survival	Cox regression for clustered data	Time dependent Cox regression for variables changing over time

Continuous Data

DISTRIBUTION OF CONTINUOUS DATA

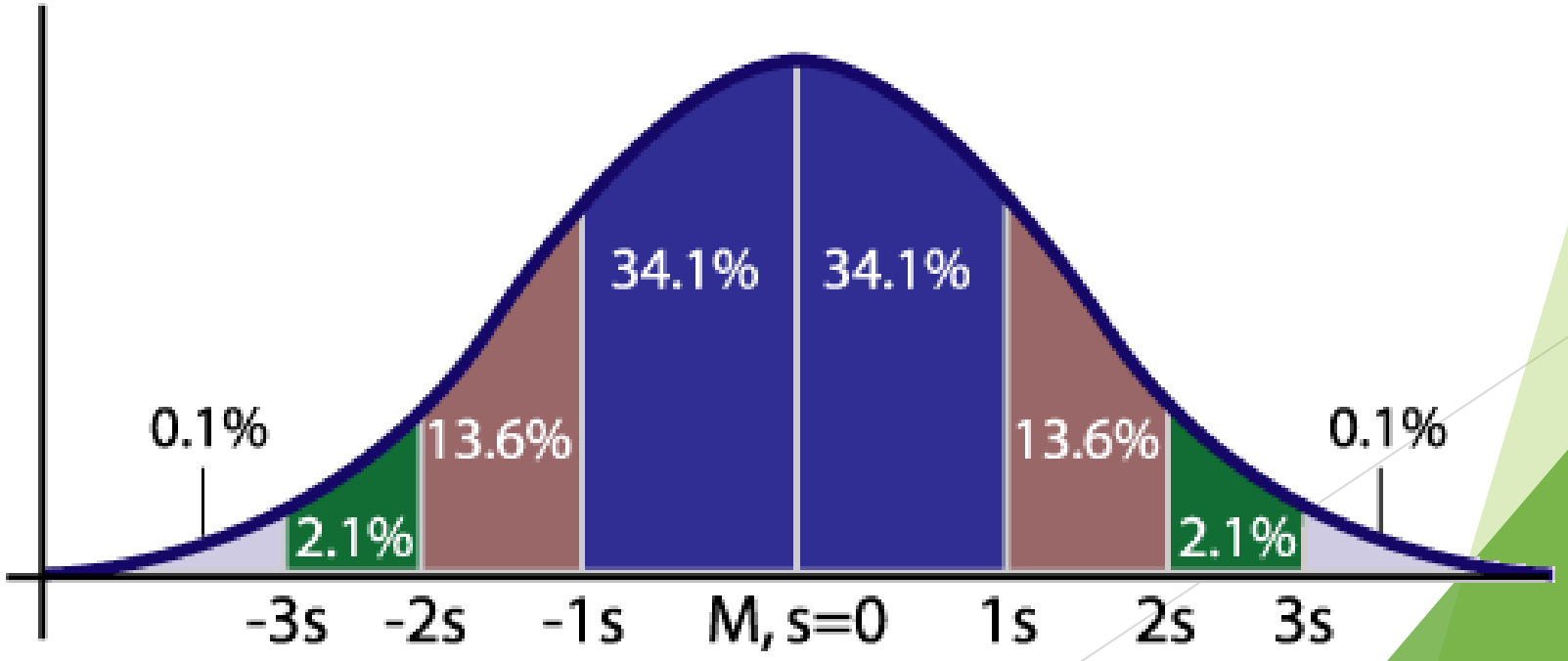
- Continuous data often represented as histograms
- Data ordered into bins often of equal size
- Shows the relative frequency of the data within each bin
- Allows for testing of assumptions and select statistical tests





NORMAL FREQUENCY DISTRIBUTION

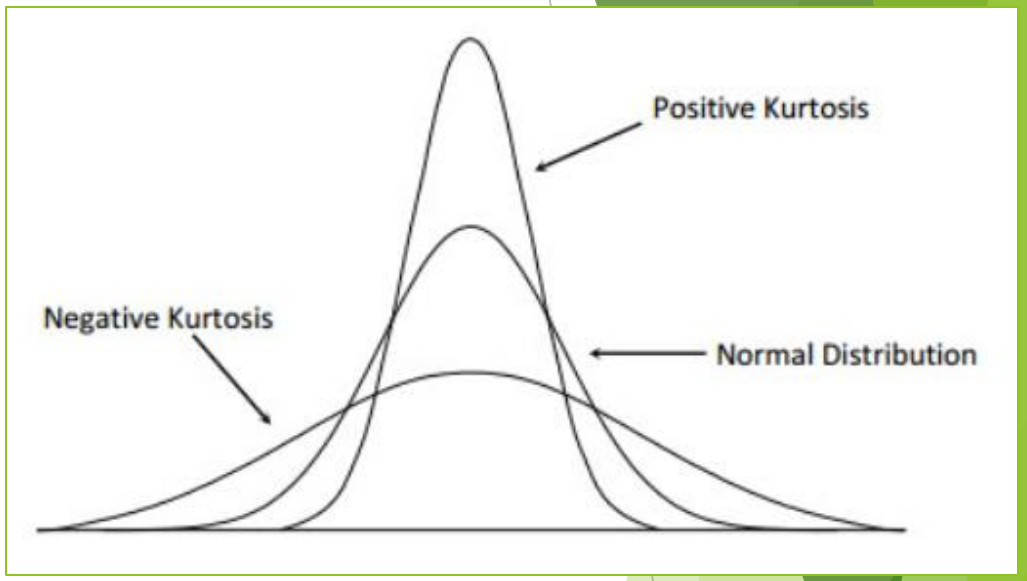
- Mean = Median = Mode
- Symmetrical: Skew = 0
- 95% of data are within 2*SD

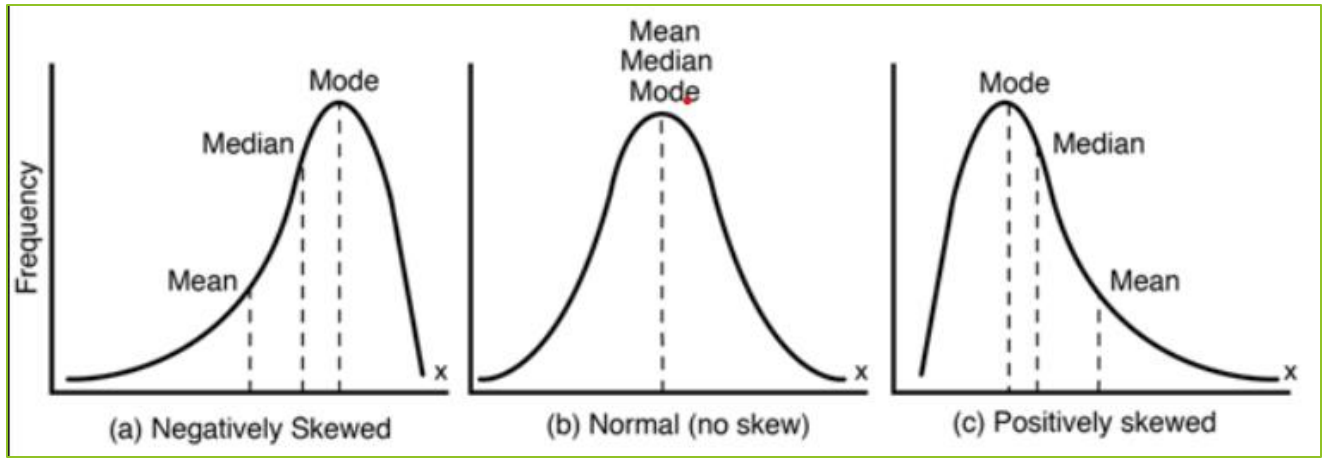


SHAPE OF FREQUENCY DISTRIBUTION

Skewness: (unbalanced) horizontal stretching

Kurtosis: vertical stretching





Tests for Continuous Outcomes

Normally Distributed/Parametric

- T-test
- ANOVA
- Correlation- Pearson's
- Linear regression

Non-Parametric

- Wilcoxon tests
- Kruskal Wallis test
- Correlation- Spearman’s
- Non-parametric regression

T-TEST and ANOVA

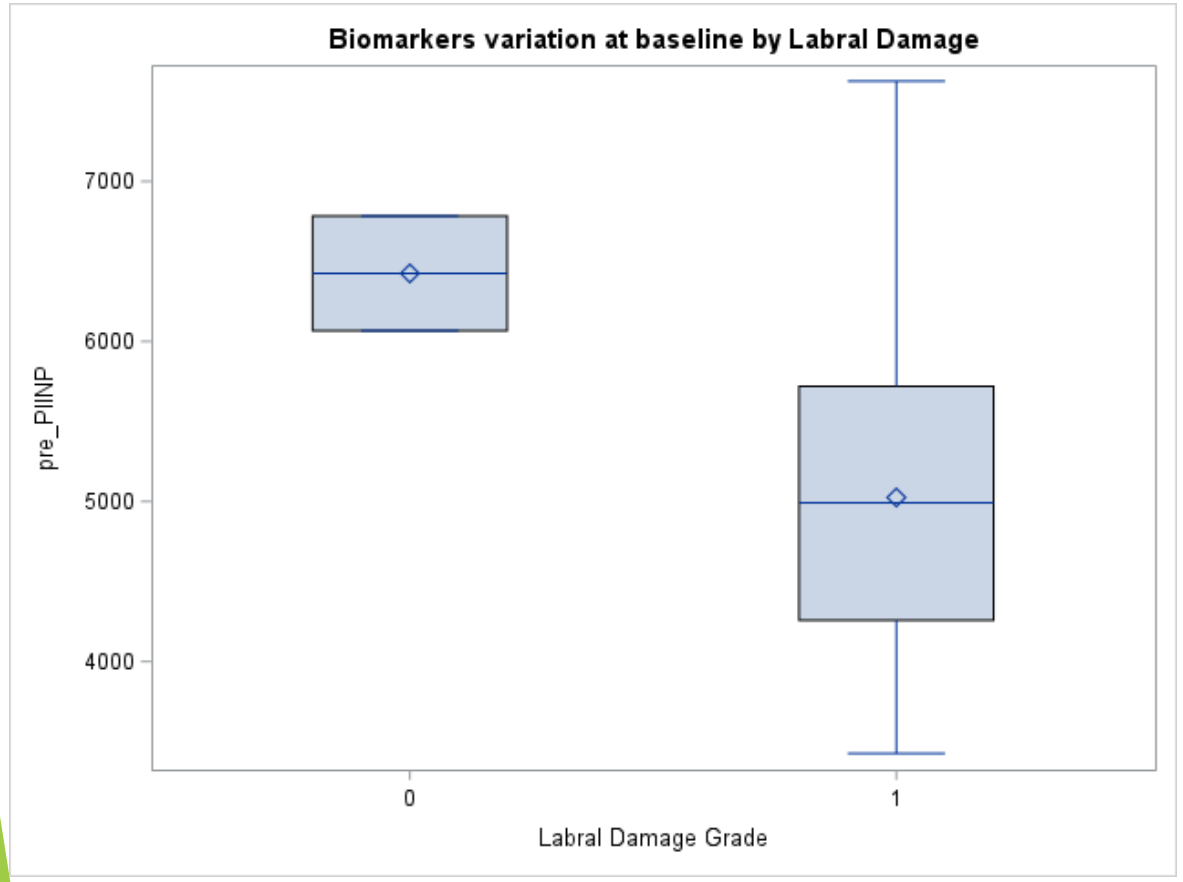
When to use:

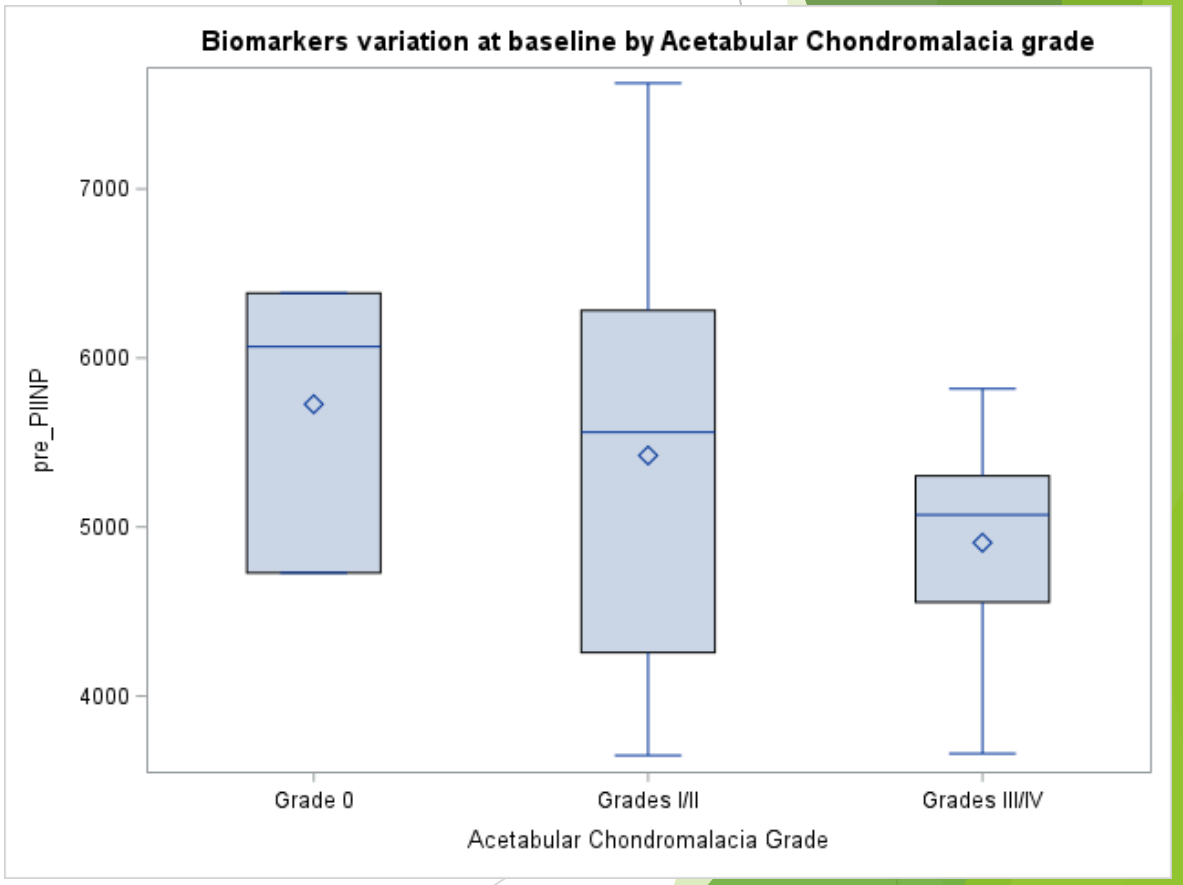
- Outcome-continuous
- Normal distribution of outcome
- Predictor/Independent variable: Categorical
  - Number of categories
    - T-test- predictors have 2 categories
    - ANOVA- predictors have >2 categories

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

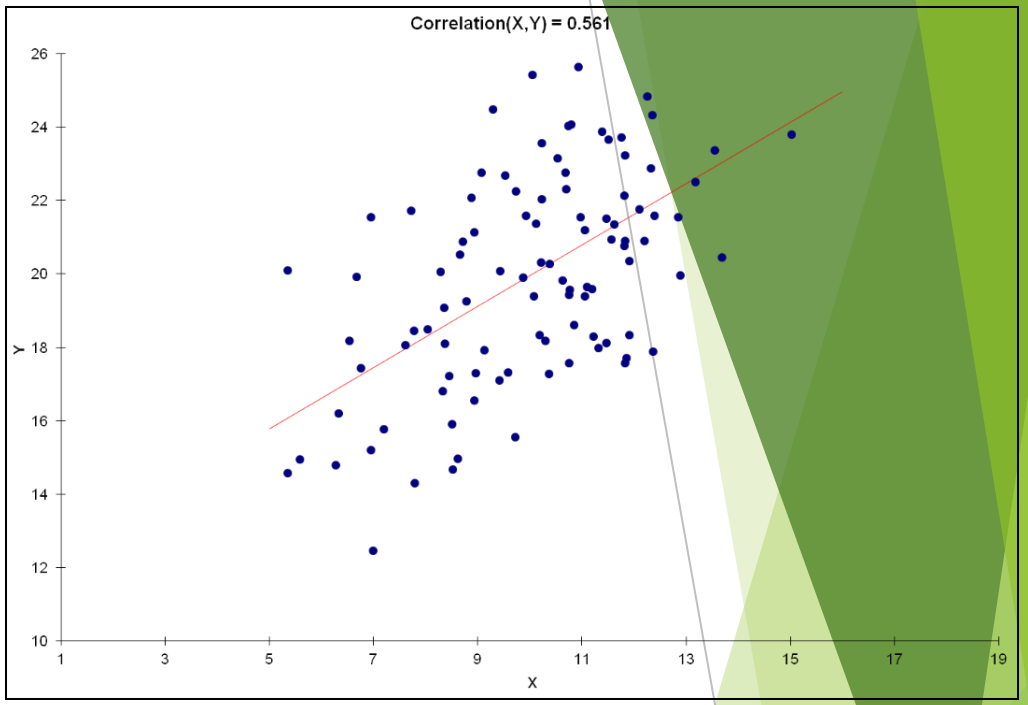
Method- Pearson's Correlation

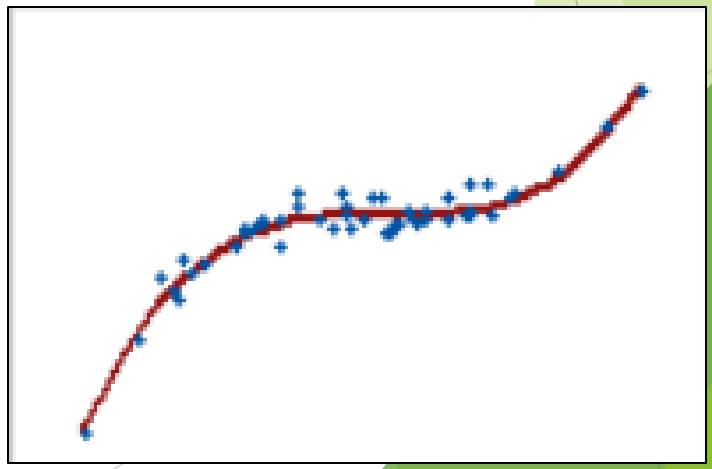
Correlation Coefficient

- Values range from -1 to +1
- Positive coefficient : positive relation
- Negative coefficient: inverse relation
- 0: no correlation

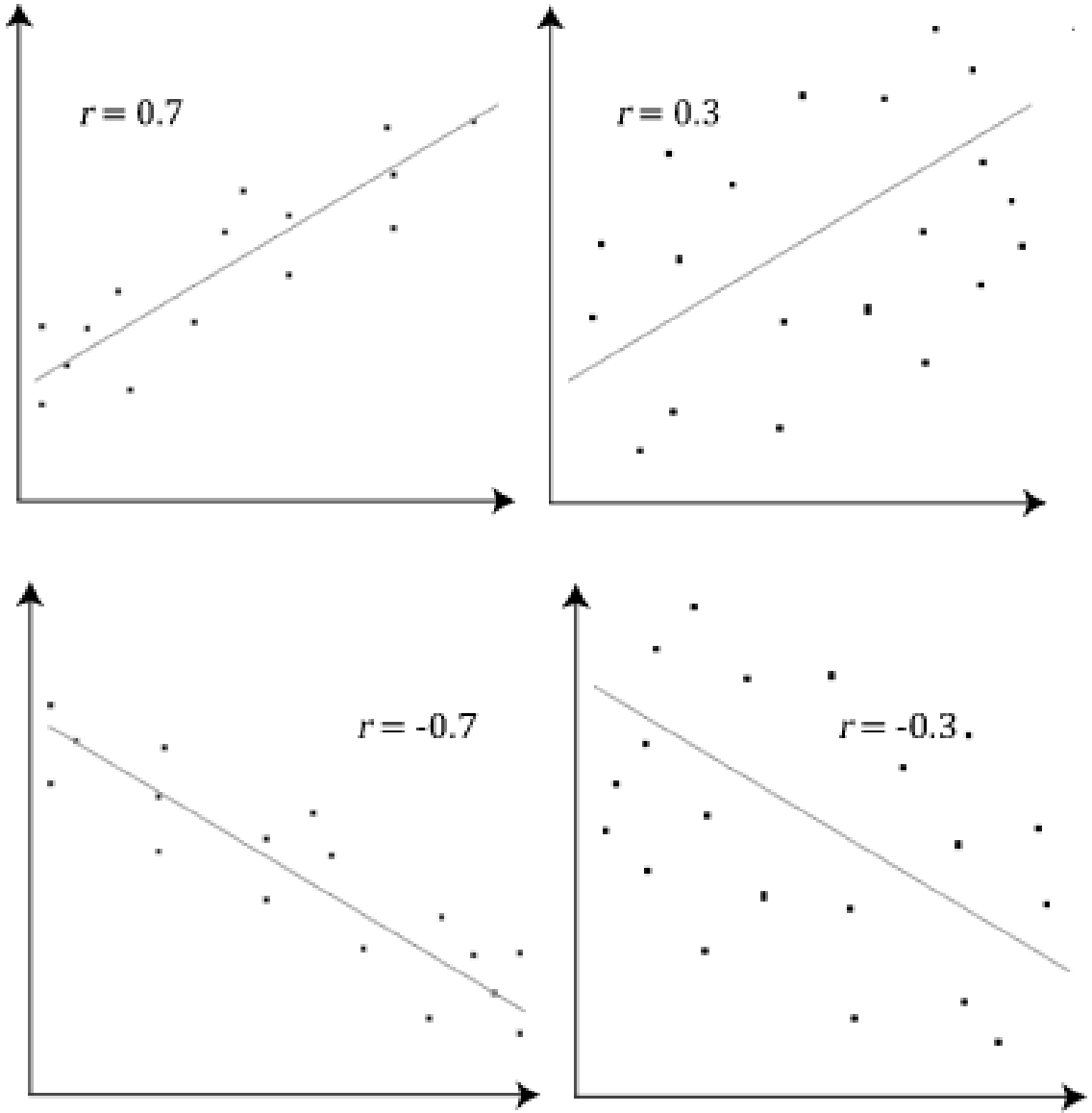
Non-Parametric:

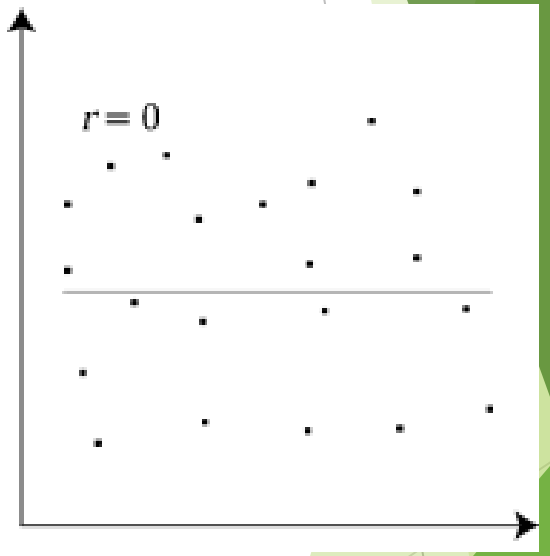
- Spearman's- Non-parametric (rank based)





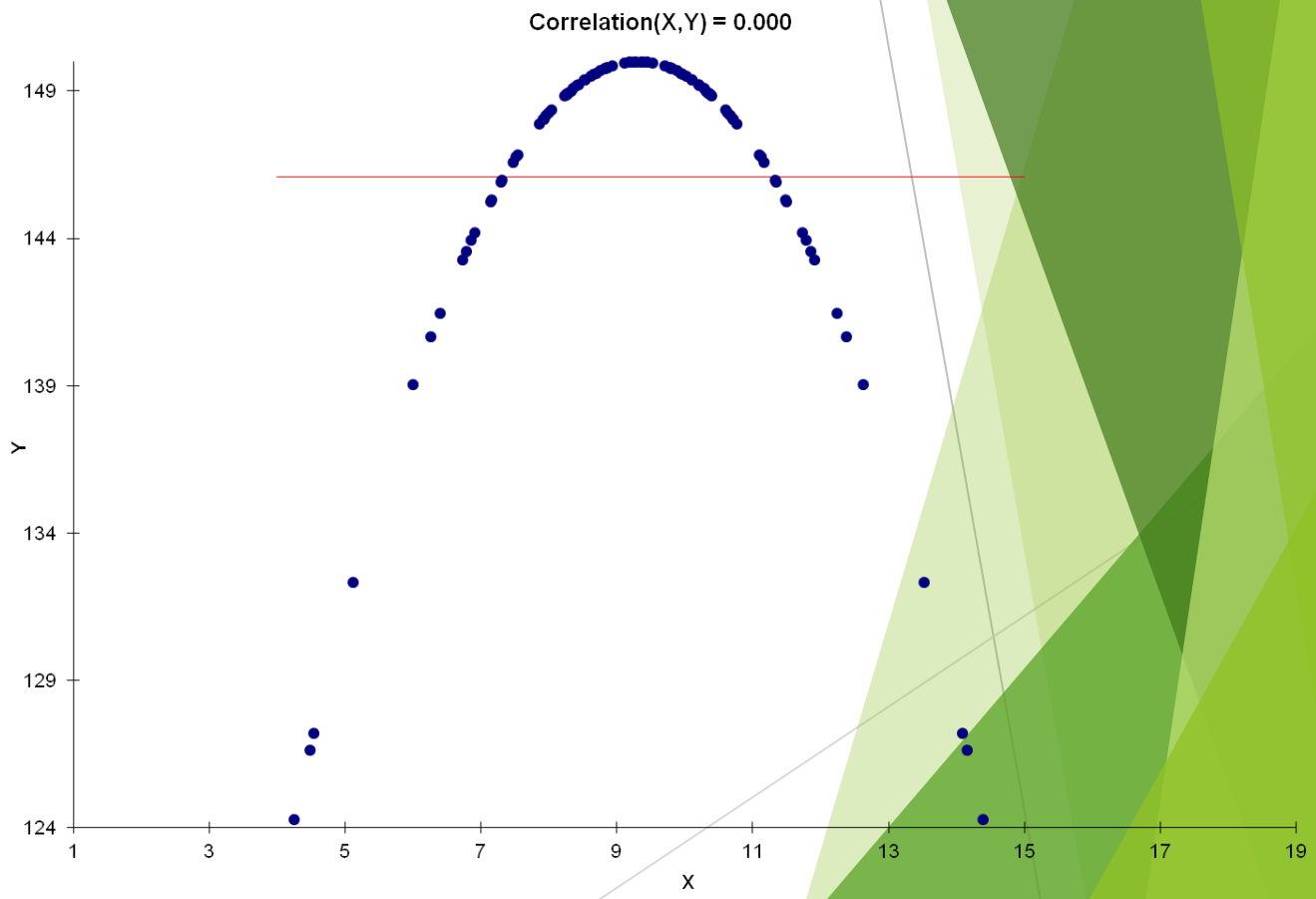
CORRELATION





Linear correlation- Limitation

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



LINEAR REGRESSION

Variables

- Continuous outcome, normally distributed
- One predictor (simple linear regression)
- Two or more predictors/independent variables (Multivariate)
- Independent variables can be categorical/continuous

Uses

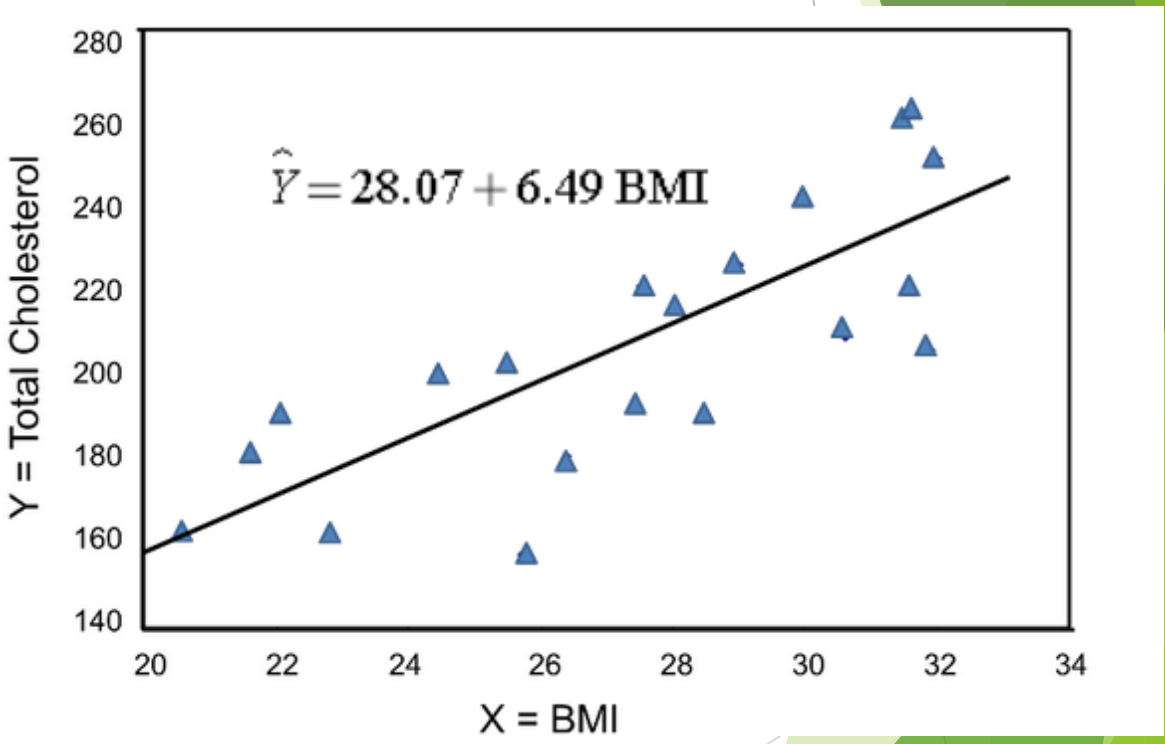
- Prediction
- Hypothesis testing

LINEAR REGRESSION

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

  $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + e$

- **β coefficient:** directly used as an estimate effect size
- **R^2**- Variance explained by the model/independent variables



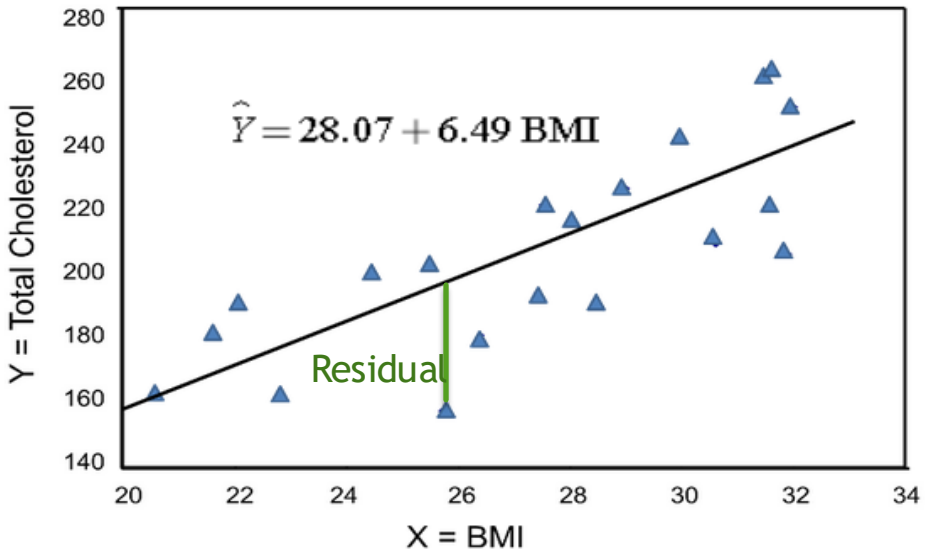
LINEAR REGRESSION ASSUMPTIONS

Independence of samples

Linear relation between dependent and predictors

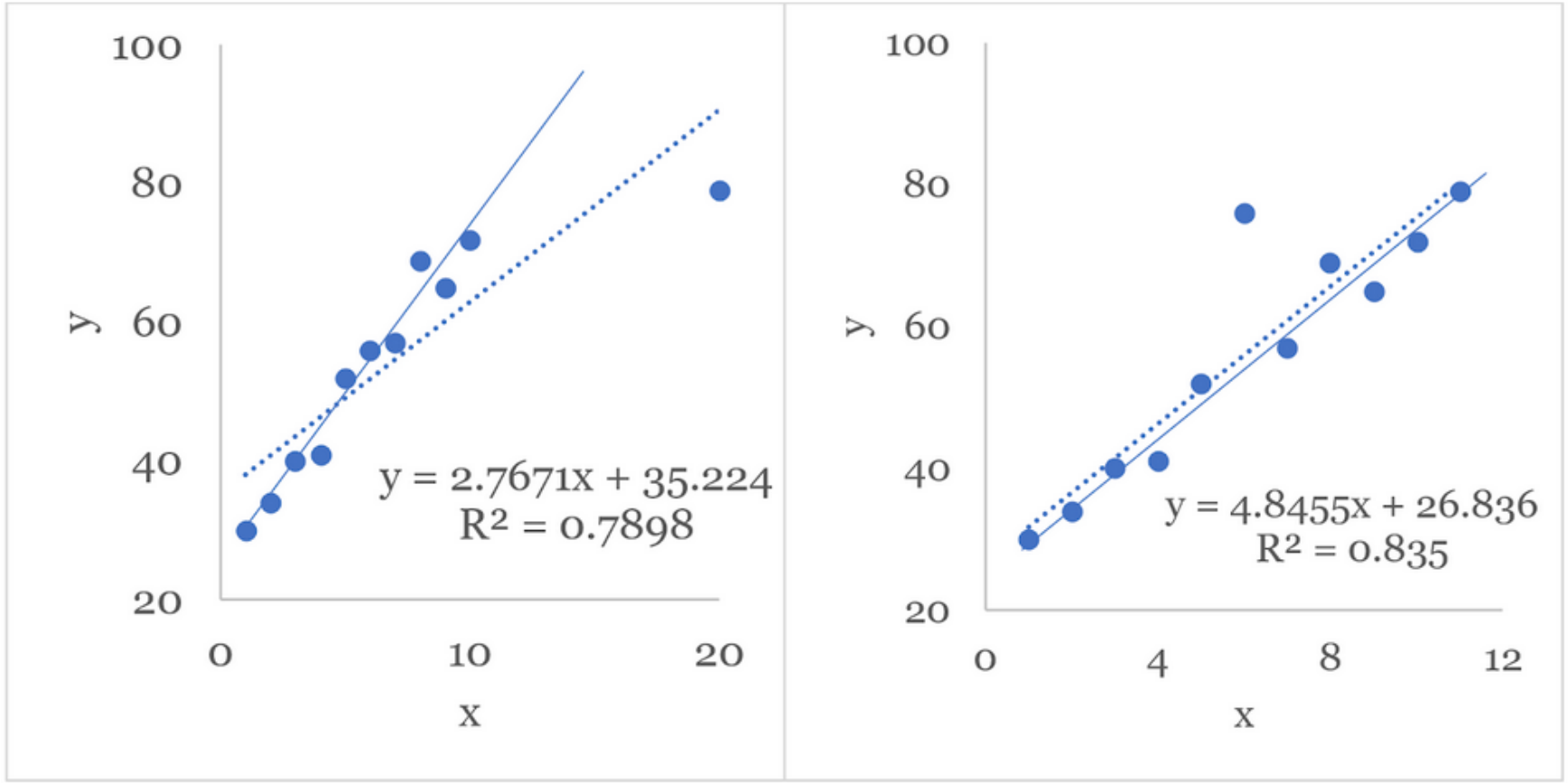
Normality of residuals

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

Tests for categorical outcomes

Comparing proportions

- Chi-sq tests- significance tests
- Multiple categories of outcomes/predictors
- Exact tests for smaller sample sizes

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

ODDS RATIOS, HAZARDS RATIOS

Odds Ratios (OR)

- Case control designs
- Best used in a cross-sectional study
- Method- Logistic Regression

Hazard Ratios (HR)

- Time-to-event data
- Cohort/Follow-up studies- retrospective or prospective
- Method- Cox Regression

LOGISTIC REGRESSION

Variables

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

Uses

- Prediction
- Hypothesis testing
- Modeling Causal Relation

LOGISTIC REGRESSION

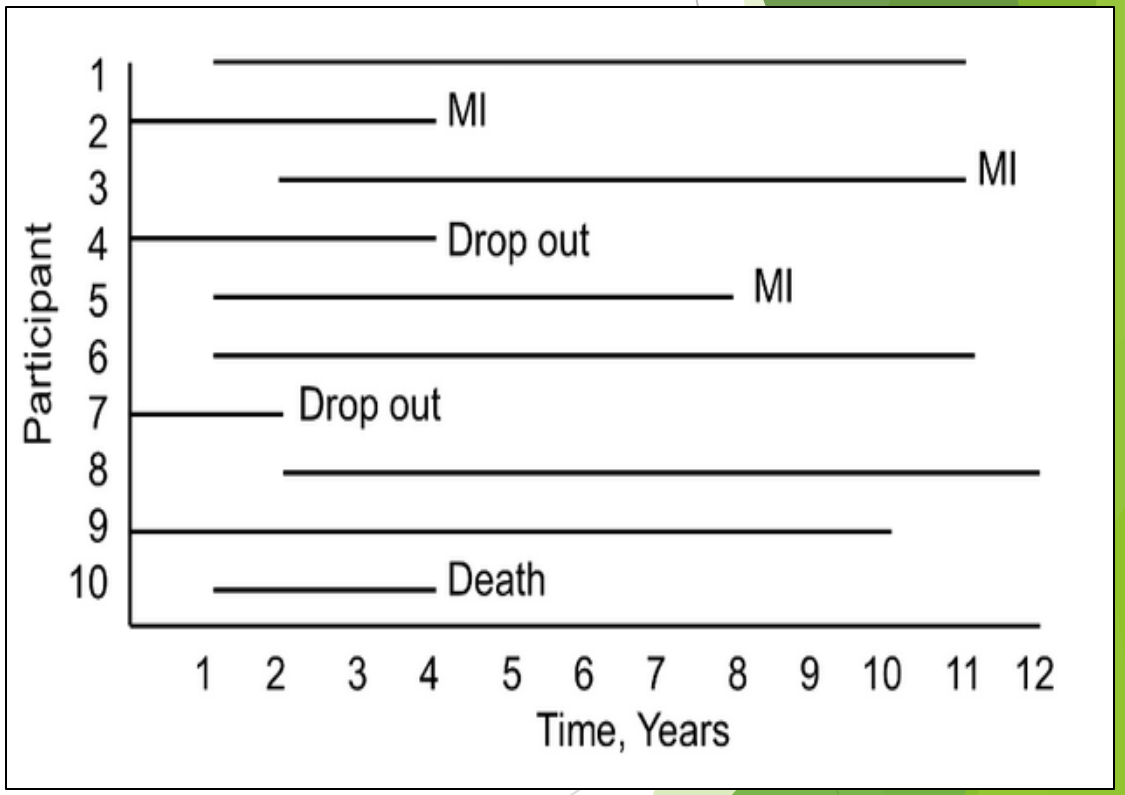
- **Regression equation:**

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

- **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

- Time-to-event 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 experienced the event of interest during the study period
- The follow up time for study participants

Survival Analysis Terms:

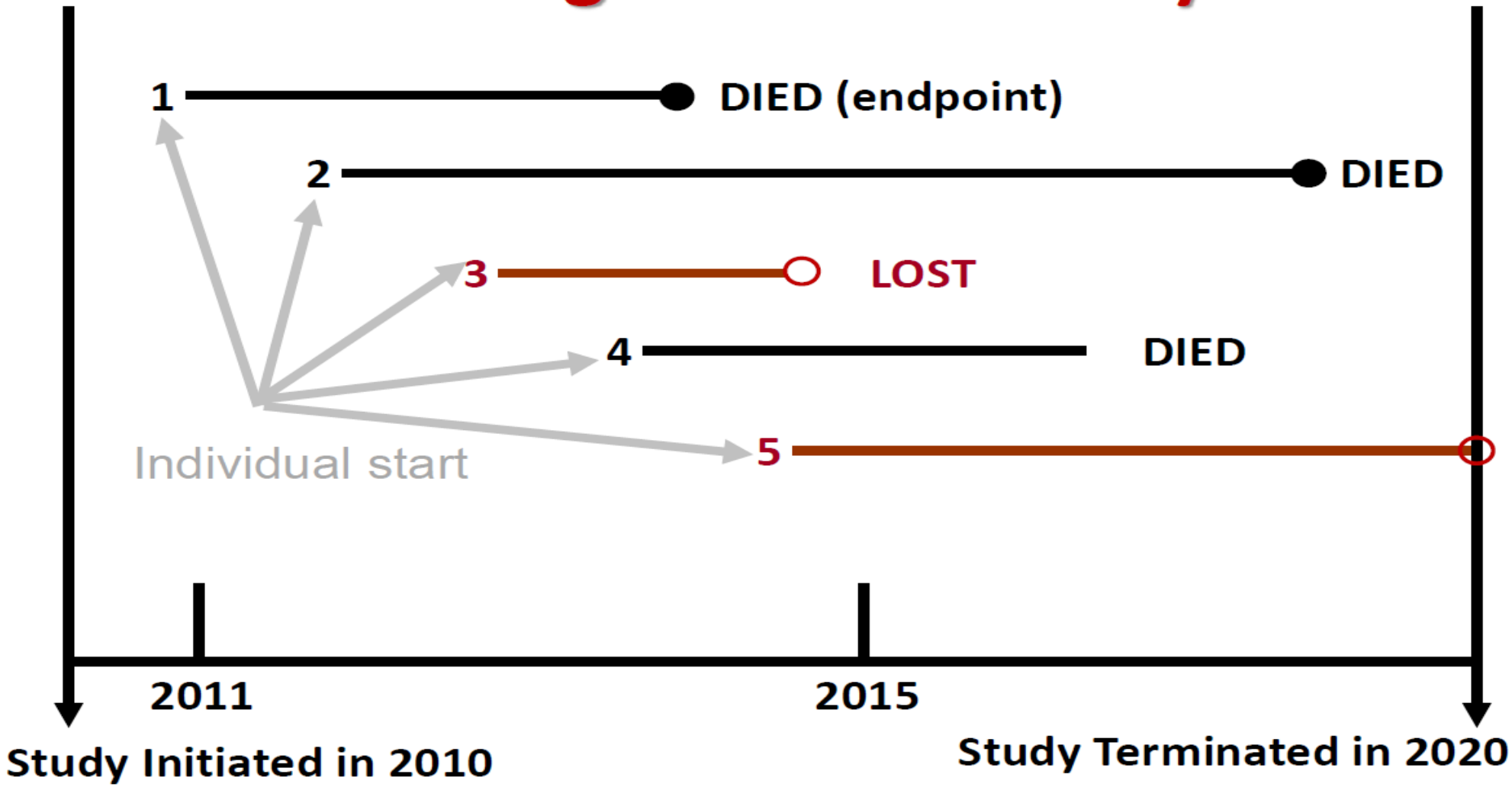
Time-to-event:

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

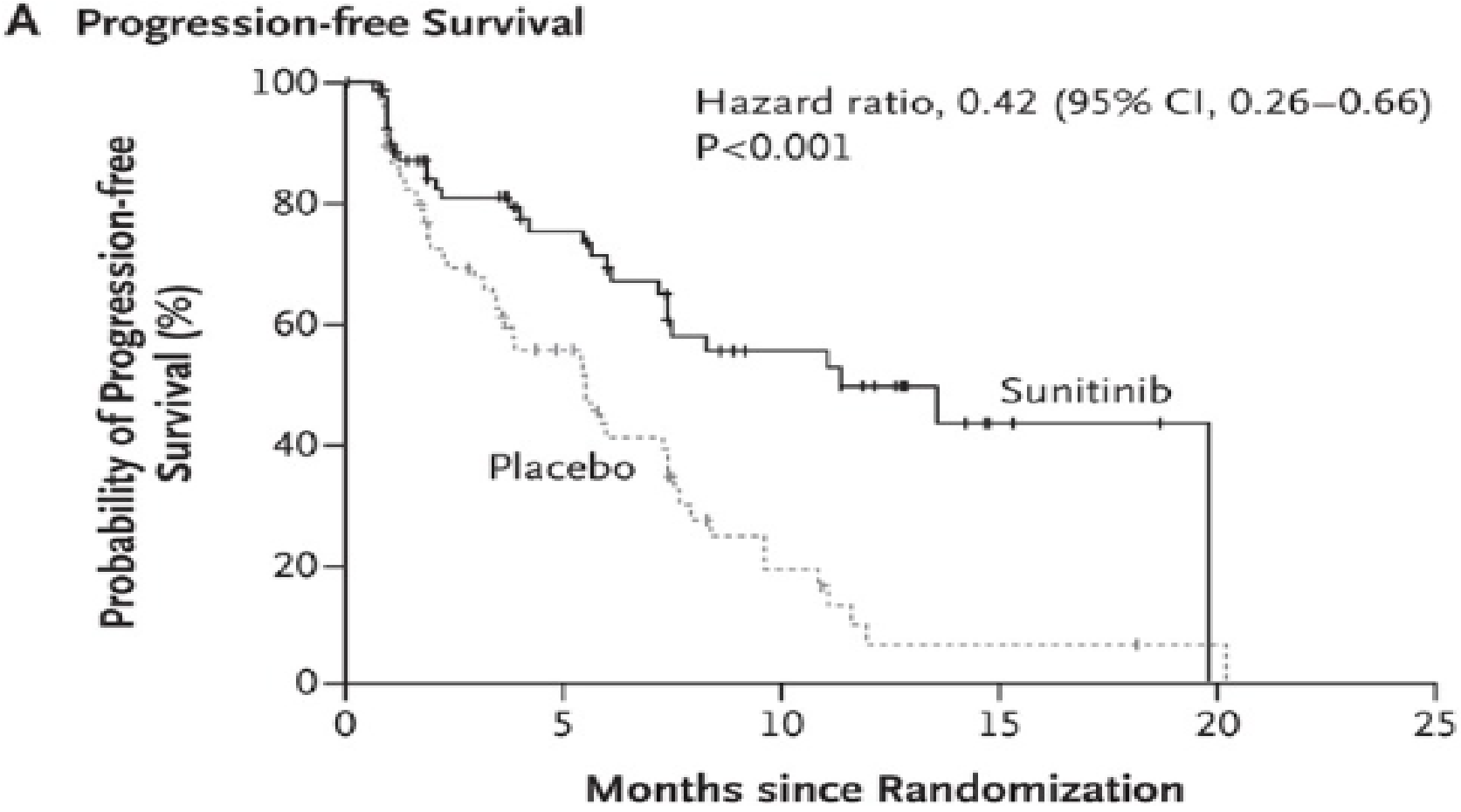
Censoring:

- Lost to follow up
- Drop out of the study
- Relocate
- End of study

Censoring in survival analysis



SURVIVAL ANALYSIS- Comparing 2 groups



Log-rank test:
Test for differencein survival between treated and control

BUILDING REGRESSION MODELS

Stepwise selection

independent variables entered or removed according to some criterion

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

Backward selection

All independent variables entered into the model, then remove least significant predictor, one at a time.

Forward selection

Start with the major risk factor, then add other independent variable and confounders

CORRELATED OUTCOMES, REPEATED MEASURES, MATCHING

Additional adjustments to models- e.g.:

- Paired t test/Repeated Measures ANOVA for continuous variables
- Conditional Logistic Regression- matching
- Generalized estimating equations (GEE)- correlated categorical outcomes
  - Logistic
  - Poisson
  - Neg binomial
- Cox Regression for Clustered data

Take aways

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

Questions?

Email: lingchen@wustl.edu