Introduction to Statistical Analysis

Changyu Shen
Richard A. and Susan F. Smith Center for Outcomes Research in Cardiology
Beth Israel Deaconess Medical Center
Harvard Medical School
This pie chart shows how much pie I ate while making this chart.
Objectives

• Descriptive versus inferential procedures
• Parametric versus non-parametric inferential procedures
• Central limit theorem (CLT) and pivot quantity
• Inferential procedures
  – Categories
  – Univariate analysis
  – Bivariate analysis
  – Multivariate analysis
Descriptive versus Inferential Procedures

• A descriptive statistic usually is the “sample version” of the corresponding population parameter
  – Mean age in a sample is the sample version of the mean age in the population
  – The purpose is to get an approximate sense of the population
• An inference procedure formally addresses the uncertainty when using statistics to infer parameters
  – Interval estimation
  – Hypothesis testing
Descriptive Statistics

- Continuous variables
  - Measures of the center of the distribution
    - Mean
    - Median
  - Measures of the dispersion of the distribution
    - Standard deviation
    - Interquartile range
    - Range
  - Measures of symmetry
    - Skewness
  - Measures of “fatness” of the tail
    - Kurtosis
Descriptive Statistics

• Binary variables
  – Proportion (proportion can also be viewed as mean if event is coded as 1 and non-event is coded as 0)
  \[ SD = \sqrt{\text{proportion} \times (1 - \text{proportion})} \]

• Categorical variables
  – Proportions
Skewness
The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.

Negatively skewed distribution
or Skewed to the left
Skewness < 0

Normal distribution
Symmetrical
Skewness = 0

Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis
The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.

Platykurtic distribution
Low degree of peakedness
Kurtosis < 0

Normal distribution
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Kurtosis > 0
Childhood Obesity, Other Cardiovascular Risk Factors, and Premature Death

Paul W. Franks, Ph.D., Robert L. Hanson, M.D., M.P.H., William C. Knowler, M.D., Dr.P.H., Maurice L. Sievers, M.D., Peter H. Bennett, M.B., F.R.C.P., and Helen C. Looker, M.B., B.S.
Table 1. Baseline Characteristics of the Participants and Prevalence of Death before 55 Years of Age.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex — no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2397 (49.4)</td>
</tr>
<tr>
<td>Female</td>
<td>2460 (50.6)</td>
</tr>
<tr>
<td><strong>Age — yr</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>11.3±3.7</td>
</tr>
<tr>
<td>Range</td>
<td>5–19</td>
</tr>
<tr>
<td><strong>Age group — no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>5–9</td>
<td>2075 (42.7)</td>
</tr>
<tr>
<td>10–14</td>
<td>1913 (39.4)</td>
</tr>
<tr>
<td>15–19</td>
<td>869 (17.9)</td>
</tr>
<tr>
<td><strong>Body-mass index †</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>21.9±6.1</td>
</tr>
<tr>
<td>Range</td>
<td>12.4–55.3</td>
</tr>
<tr>
<td><strong>Obesity — no. (%)‡</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1394 (28.7)</td>
</tr>
<tr>
<td><strong>2-hr glucose — mmol/liter</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>5.5±1.2</td>
</tr>
<tr>
<td>Range</td>
<td>1.3–11.0</td>
</tr>
<tr>
<td><strong>Impaired glucose tolerance — no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>198 (4.1)</td>
</tr>
<tr>
<td><strong>Blood pressure — mm Hg</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Systolic</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>106±16</td>
</tr>
<tr>
<td>Range</td>
<td>58–196</td>
</tr>
<tr>
<td><strong>Diastolic</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>59±11</td>
</tr>
<tr>
<td>Range</td>
<td>6–110</td>
</tr>
<tr>
<td><strong>Hypertension — no. (%)§</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>607 (12.5)</td>
</tr>
<tr>
<td><strong>Total cholesterol — mmol/liter</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3.8±0.7</td>
</tr>
<tr>
<td>Range</td>
<td>1.6–11.2</td>
</tr>
<tr>
<td><strong>Hypercholesterolemia — no. (%)¶</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>182 (3.7)</td>
</tr>
<tr>
<td><strong>Follow-up period — yr</strong></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>23.9</td>
</tr>
<tr>
<td>Range</td>
<td>0.04–37.9</td>
</tr>
<tr>
<td><strong>Death — no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>From all causes</td>
<td>559 (11.5)</td>
</tr>
<tr>
<td>From endogenous causes</td>
<td>166 (3.4)</td>
</tr>
<tr>
<td>From external causes</td>
<td>393 (8.1)</td>
</tr>
<tr>
<td>Age</td>
<td>Person-Years of Follow-up</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>5–14 yr</td>
<td>20,066</td>
</tr>
<tr>
<td>15–24 yr</td>
<td>43,081</td>
</tr>
<tr>
<td>25–34 yr</td>
<td>31,163</td>
</tr>
<tr>
<td>35–44 yr</td>
<td>17,154</td>
</tr>
<tr>
<td>45–54 yr</td>
<td>4,646</td>
</tr>
</tbody>
</table>
Before data collection

**Parameter**
- Determine the parameter of interest

**Experiment**
- Determine the study design

**Procedure**
- Determine statistical method(s)

After data collection

**Statistic(s)**
- Compute statistic(s)

**Interpretation**
- Draw conclusion(s)
Formulating a Research Question

- **Parameter ($\theta$)**
  - Translate the research question into estimation/testing of population parameter(s)
    - e.g. For the evaluation of a medical intervention, how to measure efficacy?
    - If there are multiple choices, which to pick?
      - Interpretability
      - Variability
      - Practicality

- **Sometimes we needs to transfer $\theta$ to another quantity $\tau$**
  - It is easier to draw statistical inference of $\tau$
  - Transform back
Parametric Inferential Methods

• Assumptions about the underlying distributions of the variables (collectively called a model)
  – Shape of distributions, e.g. normal distribution, Poisson distribution
  – Relationship between the mean of a variable and the values of other variables, e.g. linear, quadratic
  – Correlation structure among variables

• Advantages
  – Convenience in defining the parameters of interest; e.g. slope in a linear regression
  – More efficient (better precision)
Parametric Inferential Methods

- Disadvantages
  - Model misspecification renders parameter not meaningful
  - Bias
Nonparametric Inferential Methods

• Fewer assumptions
• Many of them are based on ranks
  – Signed rank test (NP version of paired T)
  – Wilcoxon rank-sum/Mann-Whitney U test (NP version of two-sample T)
  – Kruskal-Wallis test (NP version of ANOVA)
  – Friedman test (NP version of repeated ANOVA)
• Advantage
  – Robust
• Disadvantage
  – Less efficient
Central Limit Theorem (CLT)

• CLT states that the sampling distribution of the sample mean approaches to a normal distribution as the sample size gets large.
• It is the foundational theorem of most of the widely adopted statistical inference procedures that rely on large sample sizes.
Pivot Quantity

• $\theta$: parameter of interest
• $\hat{\theta}$: a statistic that serves as a point estimator of $\theta$ (usually the sample version of $\theta$)
• $\widehat{SE}$: the estimated standard error of $\hat{\theta}$
• $Pivot = (\hat{\theta} - \theta)/\widehat{SE}$
• Unique feature: the sampling distribution of the pivot does not depend on $\theta$. 
Scenario 1

Population (mean SBP=120mmHg)

Sample size=21

Sample 1 \( \bar{T}_1=1.32 \)

Sample 2 \( \bar{T}_2=0.27 \)

Sample 3 \( \bar{T}_3=-0.89 \)

Sample \( m \) \( \bar{T}_m=-0.11 \)

\[ T = \frac{(\text{sample mean of SBP-120})}{\text{estimated standard error}} \]
Scenario 2

Population (mean SBP=115mmHg)

Sample 1
\( T_1 = 0.87 \)

Sample 2
\( T_2 = 1.84 \)

Sample 3
\( T_3 = -1.14 \)

Sample \( m \)
\( T_m = -0.33 \)

Sample size = 21

\( T = \frac{\text{sample mean of SBP}-115}{\text{estimated standard error}} \)
Scenarios 1 and 2 have the same sampling distribution for T
Pivot Quantity

• Often the common distribution of \((\hat{\theta} - \theta) / \bar{SE}\) can be approximated by standard normal distribution based on CLT, then
  - We can make statement like
    \[
    \Pr \left[ l \leq \frac{\hat{\theta} - \theta}{\bar{SE}} \leq u \right] = 0.95 \rightarrow \Pr \left[ \hat{\theta} - u \times \bar{SE} \leq \theta \leq \hat{\theta} - l \times \bar{SE} \right] = 0.95 \text{ (confidence interval)}
    \]
  - We can compute the null distribution for the test statistic
    \[
    S = \frac{\hat{\theta} - \theta_0}{\bar{SE}} \text{ (hypothesis testing)}
    \]
Three Foundational Elements of Statistical Inference

Sampling distribution
(Define how uncertainty is measured)

CLT
Computation of the sampling distribution

Pivot
Bridge statistics and parameters
Procedures: Category “Normal” ("N")

- Sampling distribution of the pivot is standard normal or approximately standard normal
  - Estimation: \( \hat{\theta} \) and \( \hat{\theta} \pm c \times \hat{SE} \) (\( c = 1.96 \) for 95% CI)
  - Hypothesis testing:
    \[
    S = \frac{\hat{\theta} - \theta_0}{\hat{SE}}, \text{ two-sided: } |S| > c; \text{ one-sided: } S > c \text{ or } S < -c.
    \]
    \( c = 1.96 \) for \( \alpha = 0.05 \)
Standard Normal Distribution

“Bell Curve”

<table>
<thead>
<tr>
<th>Standard Deviation (σ)</th>
<th>Z-Score</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>-3</td>
<td>0.1%</td>
</tr>
<tr>
<td>-2.5</td>
<td>-2.5</td>
<td>0.5%</td>
</tr>
<tr>
<td>-2</td>
<td>-2</td>
<td>1.7%</td>
</tr>
<tr>
<td>-1.5</td>
<td>-1.5</td>
<td>4.4%</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>9.2%</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.5</td>
<td>15.0%</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>50%</td>
</tr>
<tr>
<td>+0.5</td>
<td>+0.5</td>
<td>84.1%</td>
</tr>
<tr>
<td>+1</td>
<td>+1</td>
<td>97.7%</td>
</tr>
<tr>
<td>+1.5</td>
<td>+1.5</td>
<td>99.9%</td>
</tr>
<tr>
<td>+2</td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>+2.5</td>
<td>+2.5</td>
<td></td>
</tr>
<tr>
<td>+3</td>
<td>+3</td>
<td></td>
</tr>
</tbody>
</table>

0.1% 0.5% 1.7% 4.4% 9.2% 15.0% 19.1% 19.1% 15.0% 9.2% 4.4% 1.7% 0.5% 0.1%
Procedures: Category “T”

• Sampling distribution of the pivot is T distribution
  – Estimation: \( \hat{\theta} \) and \( \hat{\theta} \pm c \times \widehat{SE} \)
  – Hypothesis testing:
    \[ S = \frac{\hat{\theta} - \theta_0}{\widehat{SE}}, \text{ two-sided: } |S| > c; \text{ one-sided: } S > c \text{ (or } S < -c). \]
    
  – \( c \) depends on confidence level/type I error and sample size
  – When sample size is large, T distribution is the same as standard normal distribution
Procedure: Category “Other” ("O")

• Alternative pivots and sampling distributions of the pivots
  – Chi-square distribution for the pivot used to infer population variance
  – F distribution for the pivot used to infer ratio of the variances of two populations
  – ...

• Non-pivot based approaches
  – When the sampling distribution of a statistic only depends on one parameter (which is of interest), exact method is possible
  – Bootstrap
Univariate Analysis for Continuous Variables

• Population mean

\[ \hat{\theta} : \text{sample mean}; \ \hat{SE} : \text{estimated SEM} \]

a) If the variable is normally distributed: T test and T interval (category T)

b) If the sample size is large: Z test and Z interval (category N)

c) Other (category O)
Univariate Analysis for Continuous Variables

• Population median (for skewed distribution)
  – Signed test

  \[ H_0: \text{median}=10 \text{ versus } H_A: \text{median}>10 \]

  \[ \iff \]

  \[ H_0^*: \text{proportion of units with value } >10=0.5 \]

  \[ H_A^*: \text{proportion of units with value } >10>0.5 \]

  Essentially a test of proportion
Univariate Analysis for Binary Variables

- **Proportion**: a proportion can be viewed as a mean by coding “event” as 1 and “non-event” as 0
  - a) If the sample size is large: Z test and Z interval (category N)
  - b) Exact method based on binomial distribution for small sample sizes (category O)
Univariate Analysis for Time to Event Variables

• Time to event variables
  – Time to death
  – Time to MI

• Feature: potentially right censoring
e.g. Follow-up is not available after time t so we know subject is event free at time t but do not know exactly when event occurs after t

• Need special technique for analysis
Univariate Analysis for Time to Event Variables

• Survival probability beyond time $t$
  – $\hat{\theta}$: Kaplan-Meier estimator (category N)
  
  Key assumption: those who are censored at a given time point and those who are not censored have the same survival probability
Univariate Analysis for Time to Event Variables

<table>
<thead>
<tr>
<th>Patient</th>
<th>$t_i$ (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>2</td>
<td>5.5*</td>
</tr>
<tr>
<td>3</td>
<td>6.7</td>
</tr>
<tr>
<td>4</td>
<td>6.7</td>
</tr>
<tr>
<td>5</td>
<td>7.9*</td>
</tr>
<tr>
<td>6</td>
<td>8.4</td>
</tr>
<tr>
<td>7</td>
<td>8.4</td>
</tr>
<tr>
<td>8</td>
<td>8.4</td>
</tr>
<tr>
<td>9</td>
<td>10.3</td>
</tr>
<tr>
<td>10</td>
<td>alive</td>
</tr>
</tbody>
</table>

* censored

<table>
<thead>
<tr>
<th>Interval $[t_i, t_{i+1}]$</th>
<th>$n_i =$ # at risk at time $t_i$</th>
<th>$d_i =$ # deaths</th>
<th>$c_i =$ # censored</th>
<th>$1 - \frac{d_i}{n_i}$</th>
<th>$\hat{S}(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 3.2)</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>1.000</td>
</tr>
<tr>
<td>[3.2, 6.7)</td>
<td>$10 - 0 - 0 = 10$</td>
<td>1</td>
<td>1</td>
<td>0.90</td>
<td>0.900</td>
</tr>
<tr>
<td>[6.7, 8.4)</td>
<td>$10 - 1 - 1 = 8$</td>
<td>2</td>
<td>1</td>
<td>0.75</td>
<td>0.675</td>
</tr>
<tr>
<td>[8.4, 10.3)</td>
<td>$8 - 2 - 1 = 5$</td>
<td>3</td>
<td>0</td>
<td>0.40</td>
<td>0.270</td>
</tr>
<tr>
<td>[10.3, 12)</td>
<td>$5 - 3 - 0 = 2$</td>
<td>1</td>
<td>0</td>
<td>0.50</td>
<td>0.135</td>
</tr>
<tr>
<td>Study Ends</td>
<td>$2 - 1 - 0 = 1$</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>0.135</td>
</tr>
</tbody>
</table>
Prevalence of Diabetes among Men and Women in China

Wenyiing Yang, M.D., Juming Lu, M.D., Jianping Weng, M.D., Weiping Jia, M.D., Linong Ji, M.D., Jianzhong Xiao, M.D., Ph.D., Zhongyan Shan, M.D., Jie Liu, M.D., Haoming Tian, M.D., Qiuhe Ji, M.D., Dalong Zhu, M.D., Jiapu Ge, M.D., Lixiang Lin, M.D., Li Chen, M.D., Xiaohui Guo, M.D., Zhigang Zhao, M.D., Qiang Li, M.D., Zhiguang Zhou, M.D., Guangliang Shan, M.D., Ph.D., and Jiang He, M.D., Ph.D., for the China National Diabetes and Metabolic Disorders Study Group*
Figure 1. Age-Specific Prevalences of Diabetes and Prediabetes among Chinese Adults 20 Years of Age or Older.

The prevalences of total diabetes (Panel A) and prediabetes (Panel C) among men and women are shown, according to age. The crude and age-standardized prevalences of total diabetes and prediabetes among men and women are shown in Panels B and D, respectively. Total diabetes includes both previously diagnosed diabetes and previously undiagnosed diabetes. Prediabetes was defined as impaired fasting glucose or impaired glucose tolerance. Error bars indicate 95% confidence intervals.
Bivariate Analysis: Two Continuous Variables

- Pearson correlation coefficient (PCC)

\[
r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}
\]

Where, \( \bar{X} = \text{mean of } X \) variable
\( \bar{Y} = \text{mean of } Y \) variable
Bivariate Analysis: Two Continuous Variables

• Pearson correlation coefficient (PCC)
  – Takes value in [-1,1]
  – Measures linear relationships
  – The two variables do NOT necessarily need to be normal
  – PCC=0 does not necessarily mean the two variables are independent
  – \( \hat{\theta} \): sample correlation coefficient
  – Inference is category N
Bivariate Analysis: Two Continuous Variables

• Simple linear regression
  – Model:
    \[ Y = \beta_0 + \beta_1 X + \varepsilon \]
  – \( \varepsilon \) is the error that is independence of \( X \)
  – Mean of \( \varepsilon \) is 0
  – Sometimes \( \varepsilon \) is assumed to be normal
  – Examine how changes of \( X \) affect the mean of \( Y \)
  – **Slope** (\( \beta_1 \)): the amount of change in the mean value of \( Y \) for every one unit increase in \( X \)
  – Point estimate: least square method
  – Inference is category N or T
\[ \mu_Y = E(Y) = \beta_0 + \beta_1 x \]

\[ Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i \]
Bivariate Analysis: Two Binary Variables

- **Odds ratio:** \(rac{\text{Odd}[Y=1|X=1]}{\text{Odd}[Y=1|X=0]}\)
  - Example: \(rac{\text{Odd}[\text{Death}|\text{Female}]}{\text{Odd}[\text{Death}|\text{Male}]}\)
  - Always greater than 0
  - Measure of the strength of association when event rate is relatively low (e.g. death rate is low)
  - \(\hat{\theta}\): sample odds ratio
  - Inference is category N or category O (exact method for small sample size)
Bivariate Analysis: Two Binary Variables

- Relative risk: \( \frac{\Pr[Y=1|X=1]}{\Pr[Y=1|X=0]} \)
  - Example: \( \frac{\Pr[\text{disease-free}|\text{vaccine}]}{\Pr[\text{disease-free}|\text{no vaccine}]} \)
  - Always greater than 0
  - Measure of the strength of association when event rate is relatively high (e.g. disease free rate is high)
  - \( \hat{\theta} \): sample relative risk
  - Inference is category N
Bivariate Analysis: Two Binary Variables

- **Risk difference**: $\Pr[Y = 1|X = 1] - \Pr[Y = 1|X = 0]$
  - Example: $\Pr[\text{Death}|\text{Female}] - \Pr[\text{Death}|\text{Male}]$
  - Takes value in [-1,1]
  - $\hat{\theta}$: sample risk difference
  - Inference is category N or O (equivalent)
Bivariate Analysis: One Continuous and One Binary Variables

• Treat the continuous variable as outcome (the binary variable defines the group)
  – Paired data
    • Mean of the difference: paired T test/T CI (category T) or Z test/Z CI (category N)
    • Median of the difference: Wilcoxon signed-rank test (category N)
  – Non-paired data
    • Mean of the difference: T test/T CI (category T) or Z test/Z CI (category N)
    • Median of the difference: Wilcoxon rank-sum test (category N)
Bivariate Analysis: One Continuous and One Binary Variables

- Treat the binary variable as outcome, the analysis becomes logistic regression (discussed later)
Bivariate Analysis: One of the Two is Time to Event Variable

• Usually the time to event variable is the outcome
• When the other variable is binary
  – Log rank test to test if the two survival curves are the same at every time point (category N and O).
• Cox proportional hazard regression model (discussed later)
Amiodarone or an Implantable Cardioverter–Defibrillator for Congestive Heart Failure

Figure 1. Kaplan–Meier Estimates of Death from Any Cause.
CI denotes confidence interval.
Multivariate Analysis: Regression Model

• A statistical model is a set of assumptions about the underlying stochastic process that generated the data

• Regression model describes the relationship between dependent random variables and independent variables (observed or unobserved)

• Regression model is a general framework that covers many statistical methods (e.g. T test, ANOVA)
Objectives of Regression Models

• Understand association between a set of independent variables and the dependent variables
• Causal inference
• Predictions
Continuous Outcome: Linear Regression

- Model:
  \[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \]
  - \( \varepsilon \) is the error that is independence of all X
  - Mean of \( \varepsilon \) is 0
  - Sometimes \( \varepsilon \) is assumed to be normal

- \( \beta_i \) measures the amount of change in the mean value of Y for every one unit increase in \( X_i \) when other X’s are fixed

- Point estimate: least square method

- Inference is category N or T

- Once \( \beta \)’s are estimated, the model can be used for prediction
Binary Outcome: Logistic Regression

• Model:
  \[ \ln \text{Odd}(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \]

• \( \beta_i \) measures the amount of change in the odd of \( Y=1 \) (at the logarithm scale) for every one unit increase in \( X_i \) when other \( X \)'s are fixed; when \( X_i \) is binary, \( \exp(\beta_i) \) is the odds ratio

• Point estimate: maximum likelihood estimation

• Inference is category N or O

• Once \( \beta \)'s are estimated, the model can be used for prediction
Time to Event Outcome: Cox Model

• Hazard

\[ h(t) = \frac{\# \text{ of death at time } t}{\# \text{ alive right before time } t} \]

• Model

\[ h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \]

• Interpretation
  – \( h_0(t) \) is the “baseline hazard” corresponding to the group of subjects with all \( X \) equal to 0
  – The essence of the model is that the \( X \)’s affect the hazard (at the logarithm scale) in a linear manner
Time to Event Outcome: Cox Model

- Cox model is also called the proportional hazard model.

Consider the model with one binary covariate:

\[ h(t) = h_0(t) \exp(\beta_1 X_1) \]

The hazard ratio of the group with \( X_1 = 1 \) to the group with \( X_1 = 0 \) is \( \exp(\beta_1) \), which does not depend on \( t \).

- \( \beta_i \) measures the amount of change in the hazard (at the logarithm scale) for every one unit increase in \( X_i \) when other \( X \)'s are fixed; when \( X_i \) is binary, \( \exp(\beta_i) \) is the hazard ratio.

- Point estimate: maximum partial likelihood estimation
- Inference is category N
- Once \( \beta \)'s are estimated, the model can be used for prediction.
Summary

• Descriptive versus inferential procedures
  – Descriptive: “rough” idea of the population
  – Inferential: formally address uncertainty

• Parametric versus non-parametric inference procedures
  – Parametric: more efficient; less robust
  – Nonparametric: more robust; less efficient
Summary

• Three elements of statistical inference
  – Sampling distribution
  – Central limit theorem
  – Pivot

• Analysis
  – Parameter(s)
  – Assumptions
  – Procedures
  – Interpretations