CPH Study Session - Biostatistics

Sumihiro (Sumi) Suzuki, PhD
Associate Professor and Chair
Department of Biostatistics and Epidemiology

University of North Texas Health Science Center
School of Public Health

January 19, 2017

Table of contents

1. Preface and Overview of the Content
2. Introduction and Summarizing Data
3. Inferential Statistics
4. Miscellaneous Topics

Preface and Overview of the Content

Today's presentation ...

- Overview of material based on the CPH Content Outline.
- Intended to be a refresher, not a comprehensive lecture.
- Assumes you have taken and passed an introductory biostatistics course.
- Focus on the main ideas and methods without much detailed justification or elaboration.

Items on the CPH Exam ...

- Of the 200 items, 30 are biostatistics items.
- Each item is mapped to one area of the CPH Biostatistics Content Outline. Study the Content Outline and not just material from your intro biostat course. Different schools cover different material.
- All items are multiple choice with one correct answer and three distractors. If there are two viable choices, choose the best one.
- No calculations required, all answers written out in equation form. E.g., if we donate a total of \$100 to 4 charities equally, how much does each charity get? Answer choice will be written as \$100/4 and not \$25.

CPH Content Outline ...

- Content Outline → https://www.nbphe.org/documents/CPH_Content_Outline_April_2014.pdf
- Note - there may be things on the exam you may have not covered before, but you will not fail the exam for getting any one item incorrect.
- As long as you can show a general competence in biostatistics, you will do fine for the biostat items.
- Outline is very broad. Let us first discuss what type of things you may want to know for each item on the outline.

Bios 1. Visualizing Data ...

- A. Data presentation - graphical representation of data.
  - Bar plots - for categorical data.
  - Histograms - for continuous and ordinal data.
  - Box (and whisker) plots - for continuous data possibly with outliers or skewed data.
- B. Kaplan Meier (curves) - survival data (time to event).
- C. Simple regression lines - linear relationship between independent (X) and dependent (Y) variable.

Bios 2. Descriptive Statistics ...

- A. Central tendencies - for continuous data: mean, median, mode.
- A. Variability - for continuous data: standard deviation, variance, range, interquartile range.
- B. Frequency - for categorical and ordinal data: counts, proportions or percentages (relative frequency).
- C. Percentiles - definition and concept; 1st, 2nd (median), 3rd quartiles.
- C. Standardized scores - Z-score, i.e.,

$$Z = \frac{X - \mu}{\sigma} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}.$$

Bios 3. Statistical Probability Distribution ...

- A. Normal - symmetric, bell-shaped, mean=median=mode, 68-95-99% rule, two parameters mean and variance, standard normal with mean 0 and variance 1.

- B. *T* distribution - symmetric like normal, use for *t*-tests for means when variance unknown.

- C. Binomial - two possible outcome (yes/no, success/failure, event/no event), counts number of events for a fixed number of trials. Two parameters, number of trials *n* and event probability *p*.

Bios 3. Statistical Probability Distribution ...

- D. Chi-square - skewed, used for test of independence/ homogeneity between two categorical variables.

- E. Poisson - counts number of events for infinite number of trials, used for rare events. One parameter, mean λ and mean=variance.

- F. F - skewed, used for ANOVA F-test in ANVOA and in linear regression.

Bios 4. Variables and Measurement Scales ...

- A. Qualitative vs. quantitative variables
  - Qualitative - categorical, dichotomous, ordinal variables.
  - Quantitative - continuous variables.
- B. Confounding - masks the true relationship, control with stratification or multiple regression.
- C. Effect modifiers - effect differs by levels of another variable, model with interaction in regression.
- D. Independent vs. dependent variables - outcome of interest is dependent, exposure or factor of interest is independent variable.
- E. Measurement scales - nominal, ordinal, interval, ratio.

Bios 5. Measurement ...

- Reliability - consistency of a measure, are similar results produced under similar conditions, Cronbach's alpha is one indicator of internal consistency.

- Validity - accuracy of a measure, does the result actually reflect the true measure, often difficult to know if a measure is valid.

Bios 6 & 12. Estimation and Confidence Intervals

- A. Sampling theory and central limit theorem (CLT) - sample mean ($\bar{X}$) follows a normal distribution as long as the sample comes from a normal population or the sample size is large enough (CLT).
- B. Estimation of population parameters - Most conclusions we make is about an unknown feature of the population. To make conclusions we need to first estimate them, e.g., sample mean for population mean.
- 12. Confidence intervals - A sample statistic like the sample mean only gives one value for the estimate. To have some idea about the precision of this estimate, we use confidence intervals which give plausible values for the parameter with some level of confidence, e.g., 95% confidence interval.

Bios 7, 8, & 12. Testing, Probability, Interpretation

- A. Statistical test assumptions - normally distributed data, sample size large enough for central limit theorem.
- B. Level of significance - α level (usually 0.05), *p*-value.
- C. Decision errors and statistical power - type I error, type II error, power=one minus type II error rate.
- D. Tests for group means - Z-test, one-sample *t*-test, two-sample *t*-test, paired *t*-test, ANOVA *F*-test.
- E. Tests of proportions - chi-square tests, test of independence, equality of binomial proportions (Z-test).
- F. Goodness of fit test - chi-square test to determine whether data come from a hypothesized distribution, e.g., normal distribution.

Bios 9. Risks and Rates ...

- A. Odds ratio (OR) - measure of association between exposure and outcome, used more in retrospective studies, e.g., case-control.
- A. Relative risk (RR) - also measure of association, used more in prospective studies, e.g., cohort study.
- You are likely to see it in a 2×2 contingency table.
- For both ...
  - value of 1 means no association.
  - value greater than 1 means positive association.
  - value less than 1 means negative association.

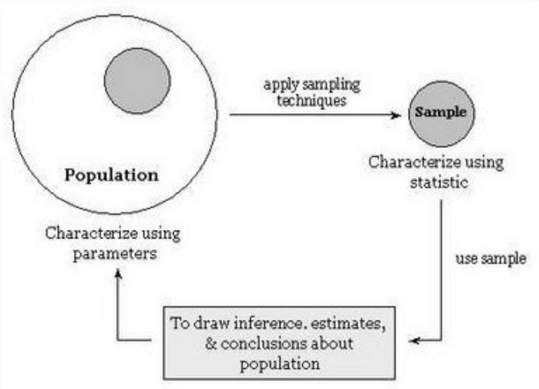
Bios 10. Correlation and Prediction Methods ...

- A. Correlation - association of two continuous variables.
- B. Simple linear regression - model linear relationship between two continuous variables.
- C. Multiple regression - model with one dependent variable and more than one independent variable. Usually one independent variable is the exposure of factor of interest, others are controlling variables, e.g., confounders.
- D. Logistic regression - dependent variable is dichotomous, produces odds ratios to show effect.
- E. Survival analysis - dependent variable is time to event, i.e., whether an event occurred (yes/no) and the time it took for it to occur. Model survival with Kaplan Meier curves and compare groups with log rank test.

Introduction and Summarizing Data

Goal of statistics ...

- To make statements about the population based on the sample (data).
- General process looks as follows.


Population
Characterize using parameters
apply sampling techniques
Sample
Characterize using statistic
use sample
To draw inference. estimates, & conclusions about population

Goal of statistics ...

- Within the process above, we usually do two things.
- Summarize the data numerically and graphically using descriptive statistics and graphs, usually known as **descriptive statistics**.
- Make statements about some feature about the population (parameter) after analyzing the data, usually known as **inferential statistics**.
- Various methods and techniques exist for both, but choosing the appropriate methods depends on the type of variable analyzed and what type of information do we get from the variables, i.e., what type of data do we have?

Measurement scale (levels of measurement) ...

- Classification that describes the nature of information within the values assigned to variables.
- Four levels with increasing levels of information: (lowest) **nominal, ordinal, interval, ratio** (highest).

Nominal: Values have no inherent order, values only used to distinguish categories, e.g., sex, race/ethnicity.

Ordinal: Values are ranked to give an order, but level of difference between ranks is not constant. E.g., a survey question where answers are 'bad', 'fair', 'good'. Here 'good' is better (higher ranked) than 'fair' but does not indicate how much better.

Interval: Distances between values are equally spaced to indicate the level of difference in ranks. E.g., temperature, a one degree increase has the same meaning for any given temperature. For an interval scale, the choice of 0 is arbitrary, e.g., 0 degrees Celsius and Fahrenheit are not the same.

Ratio: Values are equally spaced with an absolute zero point. E,g., height, an inch is always an inch, and 0 inches is the same as 0 centimeters.

- Important note - a variable with ratio scale is NOT the same as a ratio (a/b). The latter is simply the relationship between two numbers, e.g., odds ratio is the ratio of two odds. The former only indicates the level of information in a variable.

Types of variables ...

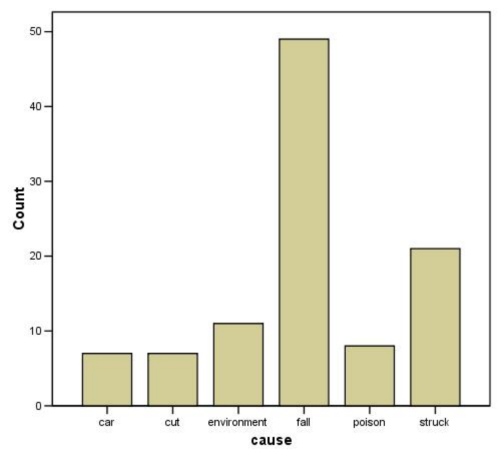
- Variables may be either **qualitative** or **quantitative**. However, this distinction is not very useful. Instead we use the following.
- **Categorical variable** - fixed number of outcomes (nominal), e.g., gender, race. Categorical variable with two possible outcomes is called a **dichotomous variable**.
- **Ordinal variable** - fixed number of outcomes (ordinal), e.g., socioeconomic status.
- **Continuous variable** - outcome (interval or ratio) may be any numerical value between a defined minimum and maximum, e.g., GPA is any number between 0.0 and 4.0.

Summarizing categorical/ordinal variables ...

- Use frequencies (counts of categories) or relative frequencies (percentages of categories).
- Present these in table format.
- Or graph them on a bar chart.
- Example. Six leading causes for a child (age 1-4) to visit an emergency room are a fall, being struck, environmental, poisoning, cuts, and car accidents. Number of cases for every 1000 children is as below.

Cause	fall	struck	envr.	poison	cut	car
Frequency	49	21	11	8	7	7
Rel. freq. (%)	47.5	20.4	10.7	7.8	6.8	6.8

- By graphing the frequencies (counts) in a bar chart ...



- We could have plotted relative frequencies in which case the vertical axis would have been proportions or percentages of the counts.

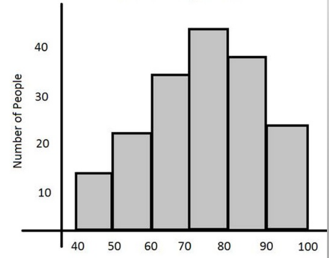
Summarizing continuous variables ...

- Central tendency - what happens in the “center” of the population or what is a typical value from the population? Estimate the population central tendency with sample statistics, i.e.,
    - mean ($\bar{X}$) = average of the sample.
    - median (2nd quartile) = middle value of the sample.
    - mode = most frequent value.

- Variability - how spread out are values in the population? Sample statistics for variability ...
    - std dev (s) = spread from mean in original units.
    - variance (s^2) = spread from mean in squared units.
    - range = maximum - minimum.
    - interquartile range (IQR) = 3rd - 1st quartile.

Graphical summary of continuous variables ...

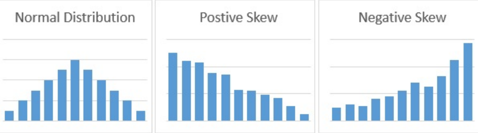
- Histogram - graphical representation of the distribution of (continuous or ordinal) data.



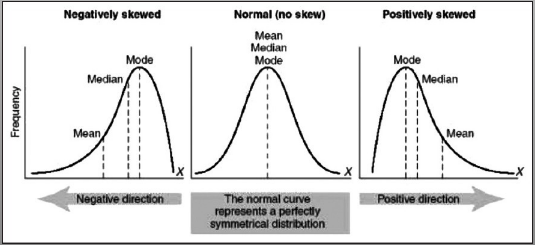
- Helpful in determining the shape of the data.
- Shape determines which numerical summary to use.

Distribution of data ...

- Common shapes of histograms are normal (symmetric), positive skew, and negative skew.



- Central tendencies vary depending on shape.

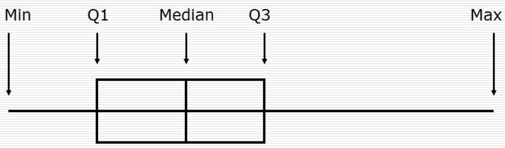


Properties of central tendency ...

- Symmetric - mean and median are the same.
- Positive (right) skew - mean $>$ median.
- Negative (left) skew - mean $<$ median.
- Notice - the median seems to capture the “middle” the best in all three cases. This is because the median is more robust (not as affected) by the distribution of the data and/or any extreme observations in the data (outliers).
- So, if the underlying distribution is not symmetric or if there are outliers, use the median instead of the mean. In other cases use the mean.

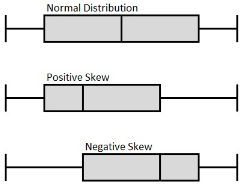
If skewed or if there are outliers ...

- Consider using box (and whisker) plot.



- Plot shows, the sample minimum, maximum, 1st, 2nd, and 3rd quartiles.

- Skewness affects the box plot ...



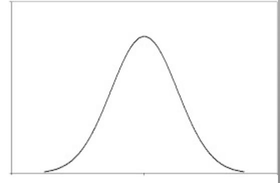
A few more descriptive statistics ...

- Percentile - the kth percentile is a value where $k\%$ of all other values fall below.
- Example. If you score in the 90th percentile on a test, that means you did better than 90% of the people who took the exam.
- The 1st, 2nd, and 3rd quartiles are the 25th, 50th, and 75th percentiles, respectively.

Inferential Statistics

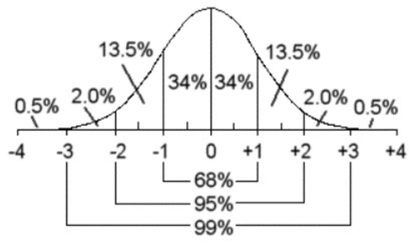
Normal Distribution ...

- Descriptive analysis are great for summarizing and presenting data, but the true strength of statistics is in inferring conclusions from the data.
- To do so, we must assume some underlying distribution for the population.
- Most common distribution for most analysis involving continuous data is the **normal** distribution.



Properties of Normal Distribution ...

- Mean $=$ median $=$ mode
- Symmetric about the mean, i.e., area to the left of the mean is 0.5 and area to the left is 0.5.
- About 68% of the values within (mean $\pm$ 1 std dev)
- About 95% of the values within (mean $\pm$ 2 std dev)
- About 99% of the values within (mean $\pm$ 3 std dev)



Z-Scores ...

- Normal distribution is so commonly used because it is easy to apply.
- For example, if we have an observation X from a normal distribution with mean 100 and variance 25, then

$$Z = \frac{X - (\text{mean})}{\text{sd}} = \frac{X - 100}{5}$$

  has normal distribution with mean 0 and variance 1, i.e., standard normal distribution.

- The transformation produces a quantity called a standardized score or Z-score.
- So, we can transform any normal variable into a standard normal variable.

Central Limit Theorem ...

- Even if distribution is not normal, a large sample size guarantees that the sample mean ($\bar{X}$) is normal.
- Then by standardizing the sample mean, i.e.,

  $$Z = \frac{\bar{X} - (\text{mean})}{\text{std error}}$$

  the distribution becomes approximately normal, again for a large sample size ($n \geq 30$).
- This type of standardization is how test statistics are computed when doing hypothesis tests.

Break Time ...

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

15-minute break

Types of Statistical Inference ...

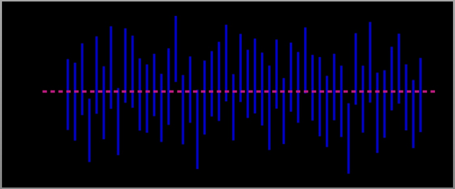
- Estimation - purpose is to estimate a feature of the population (parameter). Descriptive statistic gives point estimate. We can use a confidence interval to give an indication of precision.

- Hypothesis testing - start with some statement about a parameter. Use the data to determine whether the statement can be rejected.

- In general, a confidence interval and hypothesis test have equivalent information, i.e., if the null value (value under H_0) is in the confidence interval, then the test will not reject (not significant).

Confidence Intervals (CI) ...

- Think of it as plausible estimates for the parameter.
- So, practically, if a 95% confidence interval for the mean is $CI = [10, 15]$, any value between 10 to 15 are 'good' estimates for the population mean.
- However, technically, it means that if we were to repeatedly construct similar CI's using different samples from the same population, about 95% of those intervals will cover the true parameter.



Hypothesis testing ...

- Most common class of statistical inference.
- Start with two contradicting statements and try to find evidence against one in favor of the other.
- Statements are called the **null hypothesis** (denoted by H_0) and **alternative hypothesis** (denoted by H_1 or H_A).
- Decide whether there is enough evidence (data) to reject the H_0 in favor of H_A.

Decision making ...

- Base on data, either decide to reject or not reject H_0.
- Decision leads to one of four scenarios.

	Do not reject H_0	Reject H_0
H_0 is true	Correct	Type I error
H_A is true	Type II error	Correct
	not enough evidence	enough evidence

- Type I error rate is often denoted as α.
- Type II error rate is often denoted as β.
- Rate at which a test correctly rejects is known as the **power** of the test, denoted as $1 - \beta$.

Usual testing approach ...

- Construct H_0 and H_A.
- Assume an acceptable rate at which type I error can occur. This is called the **significance level** of the test and the standard value is $\alpha = 0.05$.
- Choose appropriate test and construct a test statistic.
- Compute a *p*-value.
- Compare *p*-value to α and make a decision.

p-value ...

- Formally, *p*-value is the probability to observe a value of the test statistic at least as ‘extreme’ as what was actually observed. Here, extreme is often used to represent evidence against H_0.
- The *p*-value is used to measure the significance of the test, i.e., is there enough evidence against H_0 to reject it.
- If so, the test is said to be significant, and if not, the test is said to be not significant.
- A *p*-value less than α indicates that there is enough evidence to reject the null hypothesis.

p-value ... misconceptions

- A *p*-value merely indicates the chances of the result you saw (test statistic) whenever H_0 is true.
- Low p-value means either that H_0 is true and a highly improbable event has occurred or that the H_0 is false. Nothing more, nothing less.
- As such, *p*-value ...
  - is NOT the probability of making a type I error.
  - does NOT indicate the size or importance of the observed effect.

Choosing a test (statistical method) ...

- When conducting a test, the most fundamental question is, "which one should we used?"
- Answer depends on the type of dependent and independent variable.
- And/or answer depends on the parameter of interest.
- For example, in testing whether two group means are different, we can think about this in one of two ways.
  - (1) Dependent = continuous, and independent = dichotomous (two groups).
  - (2) Two parameters, mean of 1st group and mean of 2nd group, i.e., test H_0: mean1 = mean2.

  Either way, the correct test is a two-sample test of means.

Tests for group means ...

- Scenario - want to know is there a difference in population means between several groups.
- If only two groups and population variance known use a two-sample Z-test.
- If only two groups and population variance unknown use a two-sample (unpaired) t-test.
- If more than two groups use ANOVA F-test.
- All tests H_0: group means same vs. H_A: different.

Tests for proportions ...

- Scenario - want to know if the frequency of categories of one variable depend on the categories of another.

- Or want to know if the distribution of a categorical dependent variable different based on levels of a categorical independent variable.

- Often times data organized in a contingency table, e.g.,

	hemoglobin (g/100 ml)			
ethnicity	≥ 10.0	9.0-9.9	< 9.0	row total
White	80	100	20	200
Black	99	190	96	385
other	70	30	10	110
column total	249	320	126	695

Given a $r \times c$ contingency table ...

- Test H_0: variables are independent (no association) vs. H_A: variables are dependent (associated).
- Use a chi-square (χ^2) test; statistic computed from observed and expected counts.
- Degrees of freedom of the test is $(r-1)(c-1)$.
- Same test may also be called test for homogeneity.

Goodness of fit test ...

- Another application of chi-square tests.

- Scenario - want to know does the sample come from a hypothesized distribution.

- Example, is a 6-sided die fair? Count the number of 1's, 2's, etc. observed. Compare observed and expected (assuming a fair die) using a chi-square test.

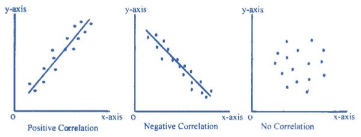
- For continuous data, divide data into intervals, then compare observed and expected using a chi-square test.

Measures of association ...

- Two common questions when doing analysis are “is there an effect?” and “if so, how much”.
- For continuous independent and dependent variables use **correlation**.
- For dichotomous independent and dependent variables use either **relative risk** or **odds ratio**.

Correlation ...

- Strength of linear relationship between two continuous variables is represented by a parameter called the correlation coefficient.
- Correlation coefficient ranges between -1 and 1.
- If it is 0, then variables are uncorrelated (no association).
- If it is positive, then variables are positively correlated.
- If it is negative, then variables are negatively correlated.
- Equivalent to simple linear regression.



Relative Risk ...

- To analyze a prospective study, we summarize the data into a 2×2 contingency table.

	outcome/disease		
risk factor	yes	no	total
yes (with risk)	a	b	$a+b$
no (without risk)	c	d	$c+d$
total	$a+c$	$b+d$	n

- The **relative risk** (RR) is

$$\begin{aligned} RR &= \frac{\text{risk of getting the disease with the risk factor}}{\text{risk of getting the disease without the risk factor}} \\ &= \frac{a/(a+b)}{c/(c+d)}. \end{aligned}$$

Odds Ratio ...

- For a retrospective study, it is often more meaningful to analyze anther quantity called the **odds ratio**.

	outcome/disease		
risk factor	case	control	total
yes (with risk)	a	b	$a+b$
no (without risk)	c	d	$c+d$
total	$a+c$	$b+d$	n

- The odds ratio is

$$
\begin{aligned}
OR &= \frac{\text{odds of having the disease with the risk}}{\text{odds of having the disease without the risk}} \\
&= \frac{a/b}{c/d} = \frac{ad}{bc}.
\end{aligned}
$$

Interpretation of RR and OR ...

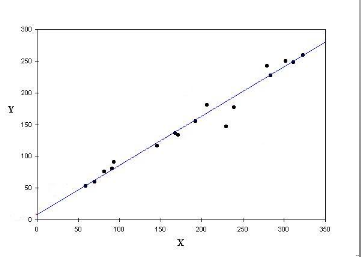
- Both RR and OR have similar interpretation.
- If 1 then there is no association.
- If greater than 1 then there is positive association.
- If less than 1 then there is negative association.
- If 1 is included in a confidence interval, then the OR or RR is not significant. Otherwise it is significant.
- For example, an $OR = 1.5$ with 95% $CI = [1.2, 2.1]$ is significant at $\alpha = 0.05$.
- But, $RR = 1.2$ with $CI = [0.7, 1.3]$ is not because it contains 1.

Linear regression ...

- Model the relationship between independent (X) and dependent (Y) variable.
- The model is a line with intercept β_0 and slope β_1,

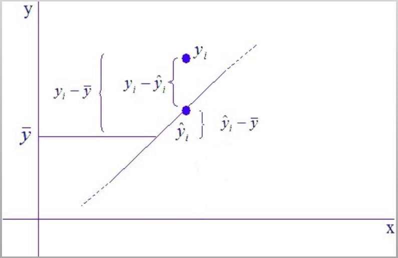
$$Y = \beta_0 + \beta_1 X.$$

- Data are points on scatter plot. Use the methods of least squares to find a line that fits well.



Fit of the line ...

- We can measure how well the line does in fitting the data.
- Graphically we have that ...



- Coefficient of determination, R^2, quantifies how much variability is explained by the line.
- Sample correlation coefficient squared is R^2.

Interpretation ... slope, β_1

- When X increases by 1 unit, Y changes by β_1.
- If $\beta_1 > 0$ then X and Y are directly proportional, and variables have positive association.
- If $\beta_1 < 0$ then X and Y are inversely proportional, and variables have negative association.
- If $\beta_1 = 0$, then Y does not depend on X at all, meaning variables not related.
- Usually β_0 is not of interest.

Logistic regression ...

- Linear regression can only be used when dependent variable (Y) is continuous.
- When dependent variable is dichotomous (1 or 0), an analogous method is **logistic regression**.
- We model the probability (p) of getting a 1, i.e.,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X.$$

- So popular in public health because e^{β_1} is the odds ratio when X increases by 1 unit.

Multiple regression ...

- Purpose of (linear or logistic) regression is to model the relationship between dependent and independent variables.
- However, there may be other variables that affect this relationship, e.g., confounders.
- Including these in the model will improve the model.
- Any regression model with more than one independent variable is known as a **multiple regression** model, e.g.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

Survival analysis ... terms and idea

- **Survival analysis** is a collection of statistical procedures used for outcome that is time until an event.
- **Time** means years, months, weeks, days, etc. from the beginning of follow-up until the event for an individual.
- Alternatively, time may refer to the age of an individual when the event occurs.
- **Event** means death, disease incidence, relapse from remission, recovery, or any other occurrence of interest.

Some applications for survival analysis ...

- Study that follows leukemia patients in remission over several weeks to see how long they stay in remission.
- Study that follows a disease-free cohort of individuals over several years to see who develops heart disease.
- A parolee's time until rearrest.
- Heart transplant patient's time until death.

Censored data ...

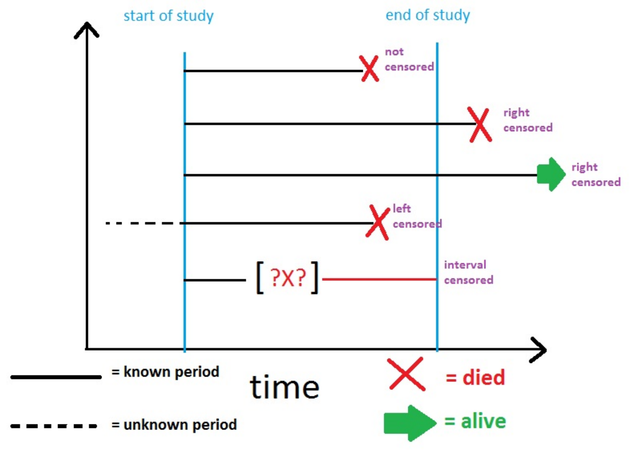
- In survival analysis we must consider a key analytical problem called **censoring**.
- Censoring occurs when exact survival time is unknown.
- In general, there are three reasons for censoring.
    - The study ends before an individual experiences the event, e.g., a leukemia patient may stay in remission even after the end of the study.
    - An individual is **lost to follow-up** during the study, e.g., a parolee may flee and will not be able to be located.
    - An individual is **withdrawn from the study**, a person in the disease-free cohort may die of a car accident before developing heart disease.

Three types of censoring ...

- Most survival data are **right-censored**, i.e., we know when the survival time starts, but do not know when or if the event occurs, usually due to one of the three reasons mentioned above.
- **Left-censored** data occur when the start of the survival period is unknown, e.g., the survival time of an HIV patient may start at infection, but the person cannot enter the study until he/she first tests positive.
- **Interval censored** data occur when the exact time of the event is unknown within the interval. This occurs in studies, where subjects are not monitored continuously.

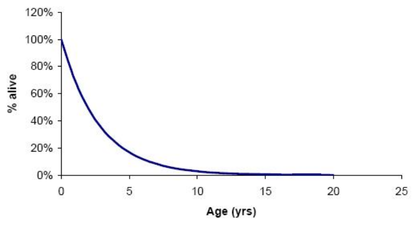
Illustration ... censoring

- Graphically ...

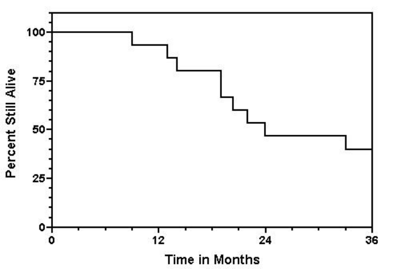


Why use survival analysis ...

- Goal of survival analysis is to analyze the **survival experience** of the population of interest.
- Survival experience is captured by a **survival function** or equivalently a **survival curve**.
- In theory, survival curves are continuous and smooth.



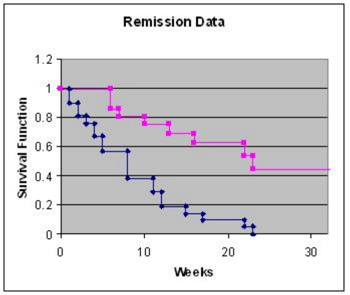
- In practice survival curves are estimated as a step function using a method known as **Kaplan Meier** estimator.



- Anytime there is a step down, it means at least one event occurred at that time.
- The estimated curve usually does not decrease all the way down to zero, because the data are censored before every subject experiences the event.

Comparing survival curves ...

- A common application is to compare survival experiences of two groups.
- E.g., time in remission for leukemia patients, one receiving a new treatment other receiving standard treatment.



- Visually, the curves above look to be different, but ...
- We would like to know are the two survival experiences significantly different?
- Test the hypothesis

$$\begin{gathered} H_0\text{: survival curves are the same} \\ \text{vs.} \\ H_A\text{: survival curves are different.} \end{gathered}$$

- Use a **log rank test**. If test rejects, the curves are significantly different.
- The method works for more than 2 groups as well.

Miscellaneous Topics

Reliability of a measurement ...

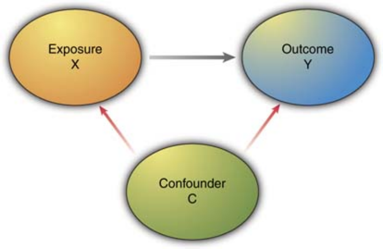
- Overall consistency of a measure.
- Measure has high reliability if similar results are produced under similar conditions.
- A common value to quantify reliability is **Cronbach's alpha**, varying from 0 to 1.
- Higher values of Cronbach's alpha indicates higher internal consistency.
- High reliability does not necessary mean the measure is accurate, i.e., not necessarily **valid**.

Validity of a measurement ...

- Assessment of the degree to which a measure represents it is supposed to measure.
- Measure could be reliable but not valid.
- For example, suppose a person weighing 200 lbs get on a scale 20 times.
  - Each time the scale reads 250 lbs.
  - This scale is highly reliable as it gives the same measure under the same conditions.
  - It is not valid because the true weight is 200 lbs.
- In general, an unreliable measure cannot be valid.

Confounding ...

- **Confounding variable** is a extraneous variable that distorts the true effect of the independent variable (exposure) on the dependent variable (outcome).



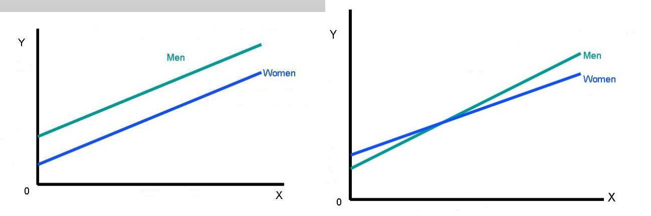
- For example, calorie intake may be positively associated to BMI, but it may be confounded by the amount of physical activity.

Ways of controlling confounding ...

- **Stratification** - Conduct separate analysis for each level of a confounding variable, e.g., one analysis for only males and another for females.
  - Need large enough sample size for each strata to have enough subjects.
  - Need to categorize continuous confounder.
  - Difficult to control when there are multiple confounders.
- **Regression** - Include the confounding variable(s) as additional independent variable in regression, e.g., $(\text{dep. var.}) = \beta_0 + \beta_1(\text{main ind. var.}) + \beta_2(\text{confounder})$.
  - Can control for more than one confounder.
  - Confounder can be continuous or categorical.

Effect modifiers ...

- **Effect modification** occurs when the effect of an independent variable on the dependent variable differs depending on the level of a third variable. This variable is called an **effect modifier**.



- First graph shows same effect for men and women, so no effect modification.
- Second shows different effects, sex is an effect modifier.
- Use interaction in regression to model.

Counting Distribution ...

- Binomial distribution models number of events out of *n* observations.
- Poisson distribution models number of events out of infinite (in theory) observations.
- In practice number of observations will not be infinite, so when to use Poisson?
- Use Poisson when the event is rare or when modeling number of events over space or time.

Thank You ...

Any Questions?