Biostatistics Lunch & Learn Series

Statistical analysis:

What statistical methods are appropriate for my study design and data collected?

Southern California Clinical and Translational Science Institute:

Research Development and Team Science

Biostatistics, Epidemiology and Research Design (BERD)

February 15, 2018

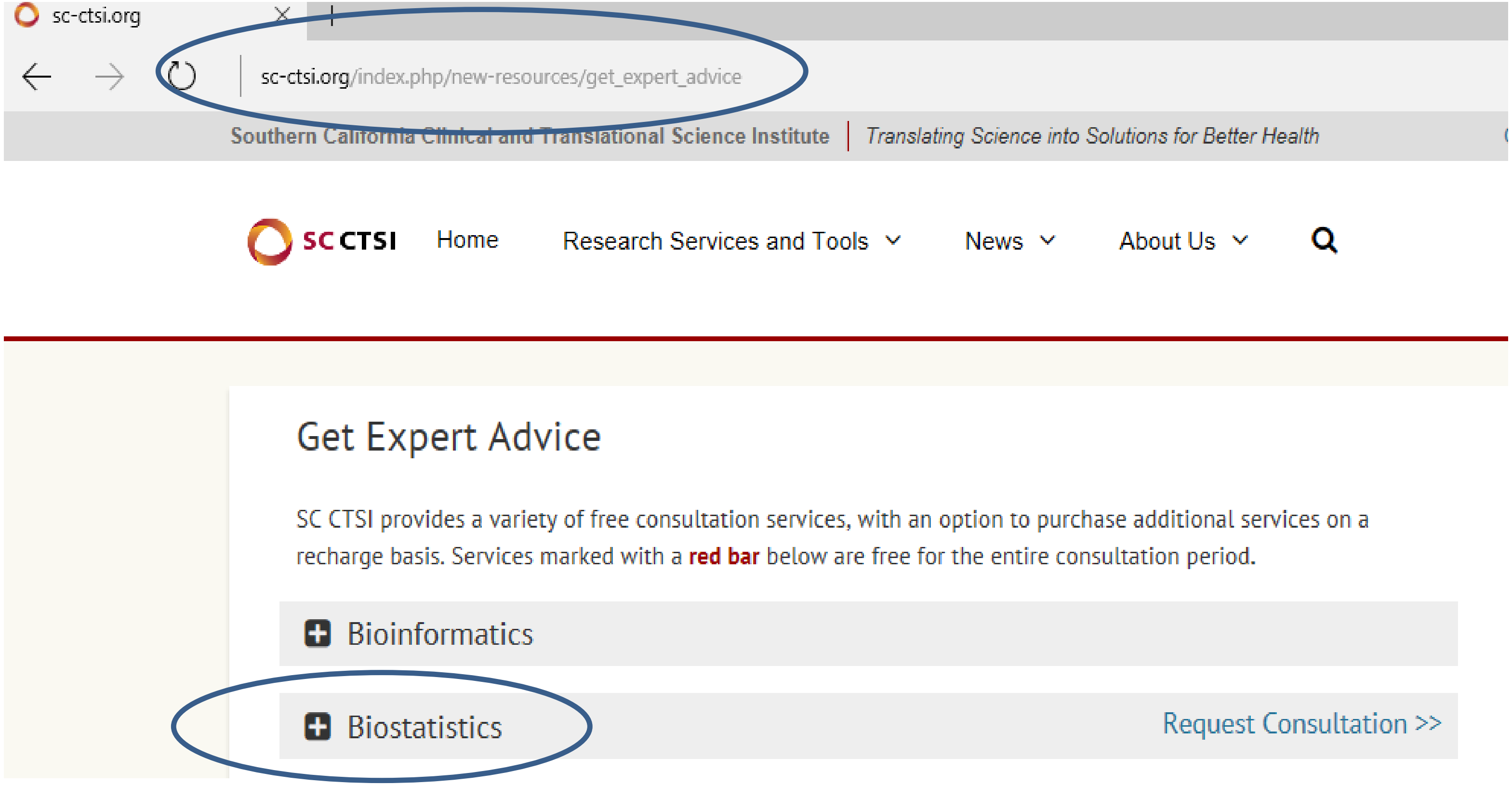
SC CTSI

Biostatistics, Epidemiology and Research Design (BERD)

Faculty: Wendy Mack, BERD Director
Christianne Lane, USC
Melissa Wilson, USC
Carolyn Wong, CHLA

Staff: Coleen Azen and Choo Phei Wei, CHLA
Caron Park and Melissa Koc, USC

SC CTSI


sc-ctsi.org
sc-ctsi.org/index.php/new-resources/get_expert_advice
Southern California Clinical and Translational Science Institute | Translating Science into Solutions for Better Health
SC CTSI
Home
Research Services and Tools
News
About Us

Get Expert Advice

SC CTSI provides a variety of free consultation services, with an option to purchase additional services on a recharge basis. Services marked with a **red bar** below are free for the entire consultation period.

⊞ Bioinformatics

⊞ Biostatistics — Request Consultation >>

Objectives

- Aug 24: Formulating a sound research question and study hypotheses: hypothesis testing

- Oct 19: Study designs and data collection strategies: scientific and logistical considerations in selecting the design to address your research question

- Dec 7: Sample size and study power: Why do I need so many subjects? What will my biostatistician need to know and how can I get that information?

- Today: Statistical analysis: What statistical methods are appropriate for my study design and data collected?

Reminder: Defining the Research Question and Hypothesis Testing

- What are the components of a good research question?

- How do I translate my research question to a statistical question (and hypothesis) that I can test?

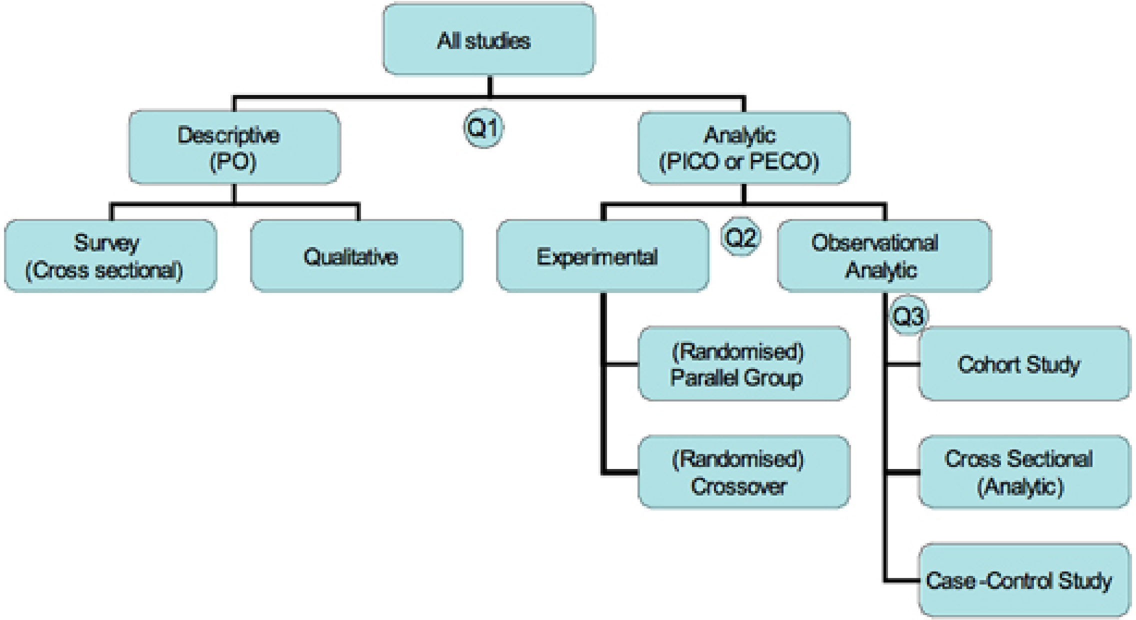
- What is statistical hypothesis testing? What does a p-value mean?

- How does the research question relate to study design? What alternative designs might be used to address my research question? (Today)

PICOT Criteria to Develop the Research Question

- **P Population**
  What specific population will you test the intervention in?
- **I Intervention (or Exposure)**
  What is the intervention/exposure to be investigated?
  Intervention (clinical trial); Exposure (observational study)
- **C Comparison Group**
  What is the main comparator to judge the effect of the intervention?
- **O Outcome**
  What will you measure, improve, affect?
- **T Time**
  Over what time period will outcome be assessed?

Spectrum of Study Designs



From Center for Evidence-Based Medicine (CEBM), University of Oxford
http://www.cebm.net/study-designs/

Estimating sample size for your study

- What data do you need to estimate sample size?
- How do you get the data you need?
- Implications for trial feasibility
- Resources for sample size estimation

Today’s Objectives:

- Understand the different types of data and how we might descriptively summarize such data
- Identify appropriate statistical methods to compare groups
- Identify appropriate statistical methods to evaluate associations among variables
- Understand survival time data and analysis methods including Kaplan-Meier lifetables and Cox regression
- Understand prediction models and associated concepts including ROC curves
- Understand screening concepts: sensitivity, specificity, etc.

Caveats

- This lecture will help you communicate with biostatisticians and others, as well as help you better interpret and critique research articles

- Know your limits and when to consult a biostatistician or other person with domain expertise.

- It is best to do so at the planning stage of your research!
  Is my research question appropriately specified? What is an appropriate and feasible study design to address the research question? Can I collect the appropriate data to test the research question? How should the data be analyzed and interpreted?

Statistical Analysis Plan

- Ties directly back to your research question, aims, hypotheses

- What are my dependent (outcome) variables? How are they measured? What type of variable are they? Am I measuring them just once (cross-sectional) or multiple times (longitudinal, repeated measures)?

- What are my independent variables (experimental interventions, control variables)? How are they measured? What types of variables are they?

- Given the above, what are appropriate methods of analysis?

Types of Research Data

- **Categorical**: falls into mutually exclusive categories
  - Nominal categorical: no natural order
    e.g., ethnicity, eye color, blood type
  - Ordinal categorical: categories have a natural order,
    e.g., socio-economic status, Likert scale data,
    educational level (elementary, high school, college)
  - Dichotomous, binary: only two categories
    e.g., dead/alive, hospitalized/released from ER, lung cancer/healthy

Types of Research Data

- **Continuous**: ordered numerical data that can theoretically take on any value

  - E*.g.*, height, weight, age, cholesterol level

  - Interval data: The interval between units have equivalent meaning across the scale (i.e., difference of SBP 130 vs 120 is the same difference as SBP 160 vs 150.

Types of Research Data

- **Discrete:** countable, ordered numerical data that are whole numbers.

  - *E.g.*, # of students, # of strokes, # of hospital days, # of correct turns in a maze

  - **Sometimes**, discrete data can be analyzed as continuous. It is not always appropriate to analyze discrete data as continuous.

Types of Research Data

- **Survival time data**: Contains two components

  - If the subject/animal had the event (e.g., did they die?)

  - The last time the subject was observed

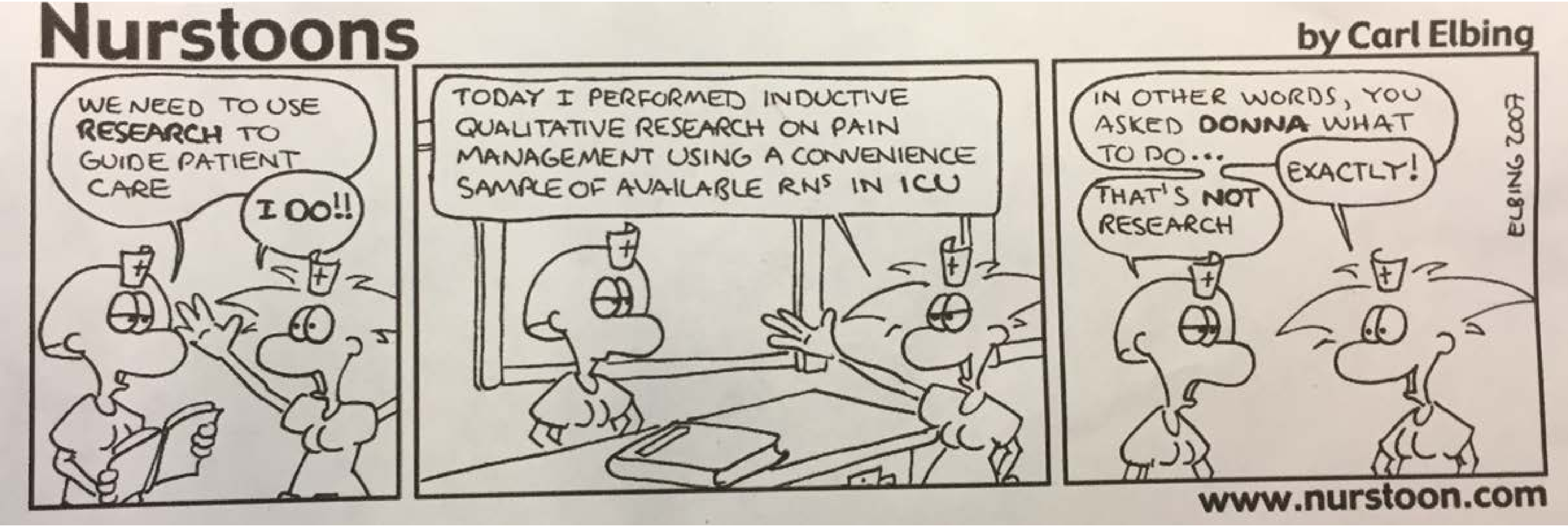
  E.g., Subject died at age 82
  Subject was alive at age 53 (last age observed on-study)
  Subject died 2.5 years after lung cancer diagnosis

Summarizing Data: Describing the Population

- Remember when we ask a research question and conduct a study to address that research question, our objective is to make an inference about a **population**, based on information contained in a **sample**

- The way we sample from the population influences:
  1) The precision of our estimates (variability)
  2) Our estimates themselves (may be subject to bias or error)

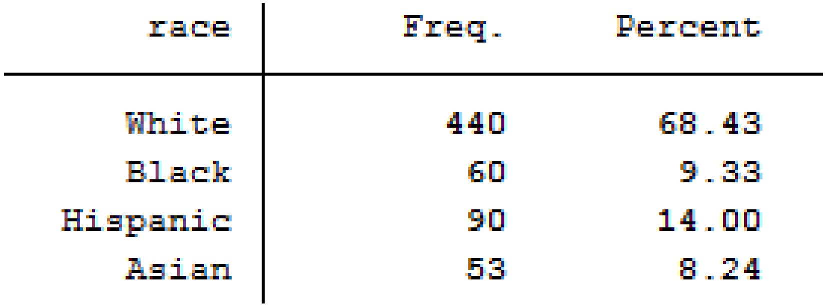
Sampling from where???


Nurstoons
by Carl Elbing
WE NEED TO USE RESEARCH TO GUIDE PATIENT CARE
I DO!!
TODAY I PERFORMED INDUCTIVE QUALITATIVE RESEARCH ON PAIN MANAGEMENT USING A CONVENIENCE SAMPLE OF AVAILABLE RNs IN ICU
IN OTHER WORDS, YOU ASKED DONNA WHAT TO DO...
EXACTLY!
THAT'S NOT RESEARCH
ELBING 2007
www.nurstoon.com

Summarizing Data: Describing the Population

- Methods for data summarizations depend on the type of data

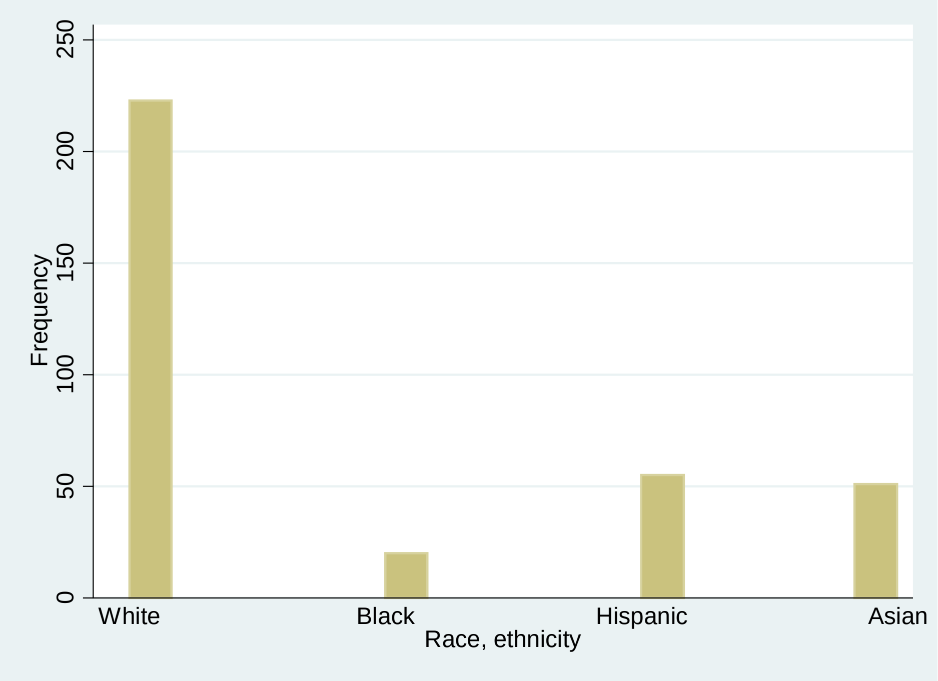
- Categorical data: usually summarized by frequency, percents



```
        race |      Freq.     Percent
-------------+-------------------------
       White |        440       68.43
       Black |         60        9.33
    Hispanic |         90       14.00
       Asian |         53        8.24
```

Summarizing Data: Describing the Population

- Categorical data: bar charts for counts or percents



Summarizing Data: Describing the Population

- Continuous data: describe by measures of **central tendency and spread**

- **Central tendency**: mean, median, mode

- **Spread**: variance, standard deviation, range, interquartile range

- **Percentiles** of the distribution:
  25th, 50th, 75th percentiles

Summarizing Continuous Data: Central Tendency

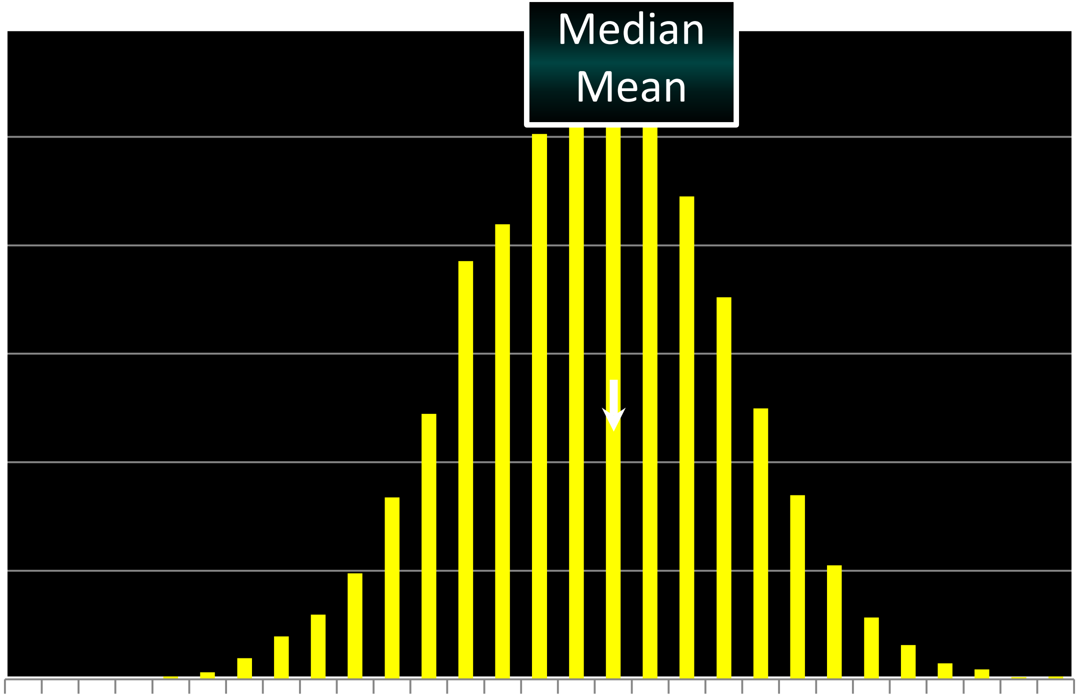
- The “middle” of the data
- Median: 50th percentile
- Mode: the most common value

  Mode can be used to describe both categorical and continuous data

- Mean: the “average”

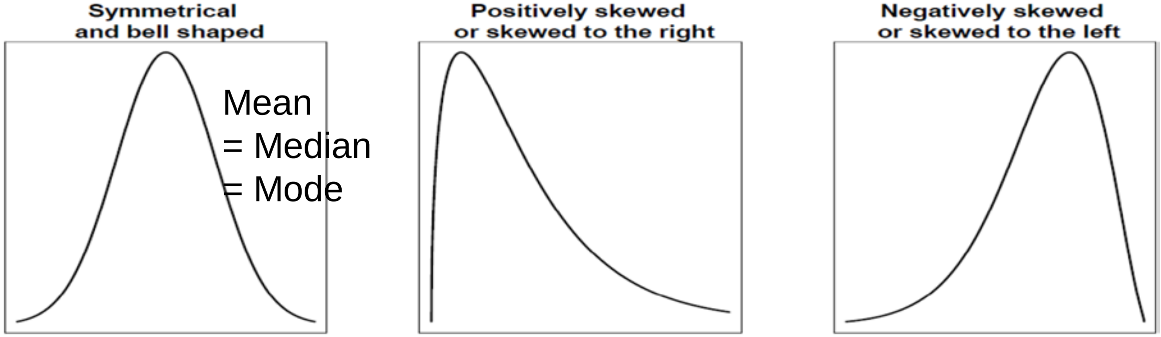
$$\overline{x} = \frac{\sum_i x_i}{n}$$

Summarizing Continuous Data: Central Tendency



25th percentile ------------------------ 75th percentile

Summarizing Continuous Data: Central Tendency



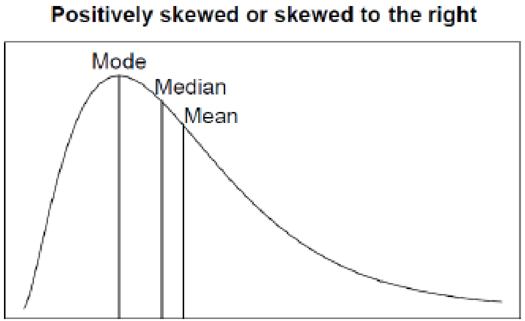
Mean
= Median
= Mode

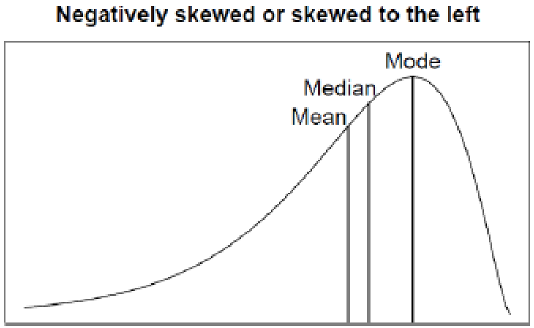
Positive skew/skewed to the right

- Longer tail in high values
- Mean > median > mode

Negative skew/skewed to the left

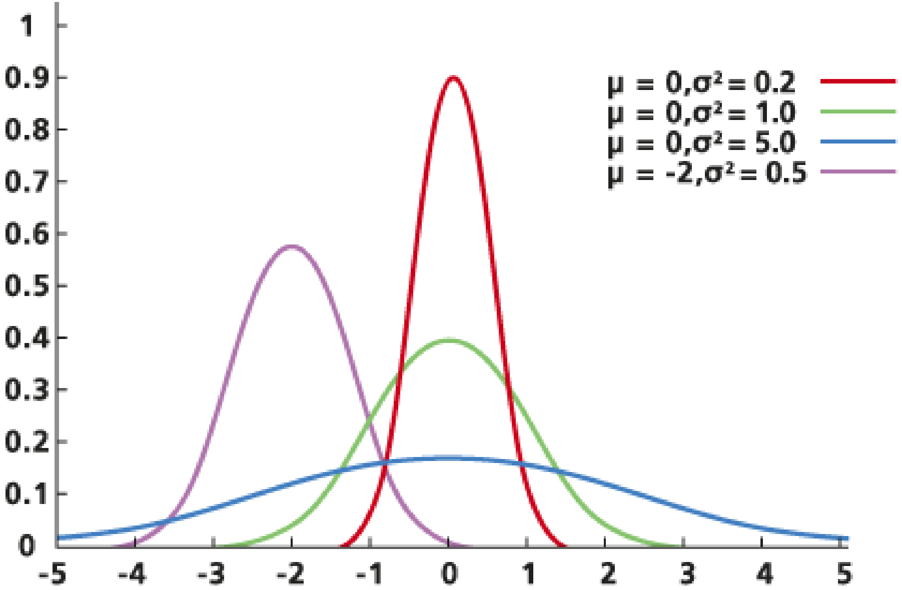
- Longer tail in low values
- Mode > Median > Mean





Summarizing Continuous Data: Spread

- Summarize continuous distributions by two characteristics: central tendency AND spread



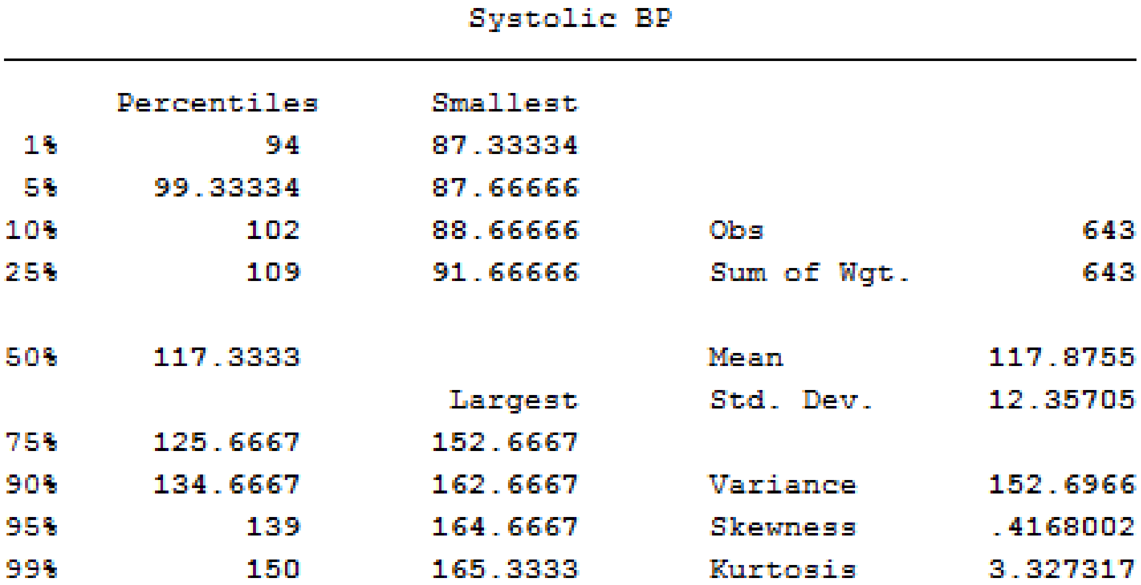
Summarizing Continuous Data: Spread

- Measures of spread
- Range: the difference between the largest and smallest value in the data
- Interquartile range: The difference between the 75th and 25th percentile values
- Variance: the average squared deviation from the mean

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n - 1}$$

- Standard Deviation: the square root of the variance

Summarizing Continuous Data: Spread



```
                         Systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           94       87.33334
 5%     99.33334       87.66666
10%          102       88.66666       Obs                 643
25%          109       91.66666       Sum of Wgt.         643

50%     117.3333                      Mean           117.8755
                        Largest       Std. Dev.      12.35705
75%     125.6667       152.6667
90%     134.6667       162.6667       Variance       152.6966
95%          139       164.6667       Skewness       .4168002
99%          150       165.3333       Kurtosis       3.327317
```

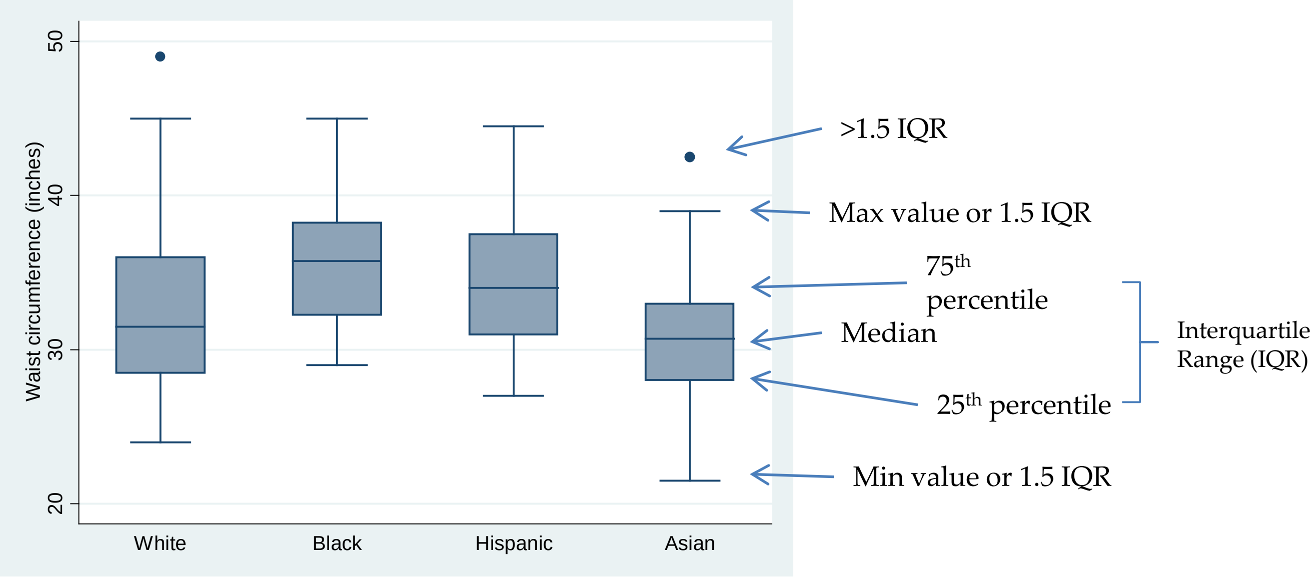
Summarizing Continuous Data: Stem and leaf

```
21* | 5          21.5 inches
22* | 5
23* | 5
24* | 005
25* | 00005555
26* | 00000005555
27* | 00000000000000000055558
28* | 0000000000000000000000055555555
29* | 000000000000000000055555555
30* | 00000000000000000000555555557
31* | 0000000000000000000055555555
32* | 0000000000000000000000055555555
33* | 00000000000000055555555
34* | 0000000000000005555555558
35* | 00000000000005558
36* | 00000000000000555555555
37* | 00000000000005555555
38* | 000000005555
39* | 000000035555
40* | 0000008
41* | 00055
42* | 005
43* | 00
44* | 005
45* | 00
46* |
47* |
48* |
49* | 0
```

Valid Value?

- Waist circumference (inches)
- Symmetry, skewness
- Outliers (valid values?)
- Look at central tendency and spread

Summarizing Continuous Data: Boxplots



>1.5 IQR

Max value or 1.5 IQR

75th percentile

Median

25th percentile

Interquartile Range (IQR)

Min value or 1.5 IQR

Providing an Estimate in the Presence of Spread (Confidence Intervals)

- Sample means, proportions, etc. are **estimates** of the population parameter from which we have sampled
- Repeated samples from the same population will give different estimates of the population mean, proportion, etc.
- Confidence intervals provide an estimate of likely values of the **true value of the population parameter**, given your sample
- 95% confidence interval: 95% of confidence intervals from repeated samples from the population will contain the true value of the population parameter
- Note the corollary: 5% of repeated samples will NOT include the true value of the population parameter

Providing an Estimate in the Presence of Spread (Confidence Intervals)

- Point estimate: The sample estimate (sample mean, etc.)
- In general, 95% CI = point estimate ± 1.96 SE(estimate)

- 95% confidence interval on a sample mean:

$$\bar{x} \pm 1.96(SEM)$$

where SEM (standard error of the mean) = SD/$\sqrt{n}$

Larger samples (larger n) will have narrower confidence intervals (more precise estimate of population parameter).

Providing an Estimate in the Presence of Spread (Confidence Intervals)

- Example: In a sample of postmenopausal women, mean SBP=120, SD=10
- If n=1000, 95% CI = 120±1.96(10/31.6) = 120±0.62
  = (119.4, 120.62)
- Contrast this to a sample of n=20
  95% CI = 120±1.96(10/4.5) = 120±4.4
  = (115.6, 124.4)

Providing an Estimate in the Presence of Spread (Confidence Intervals)

- 95% confidence interval on a sample proportion (p):

$$\bar{p} \pm 1.96(SE(p))$$

where SE(p) (standard error of the proportion) =

$$\sqrt{\frac{p(1-p)}{n}}$$

Again, larger samples (larger n) will have narrower confidence intervals (more precise estimate of population parameter).

Sample Distributions: Parametric vs Non-Parametric

- Not all distributions are normal, *i.e.,* bell–shaped

- Statistics fall into two broad categories
  - Parametric: assumes the data follow an underlying distribution
  - Non–parametric: also known as distribution–free statistics, do not assume an underlying distribution

- If your test statistic assumes a normal distribution, you cannot use it to analyze non–normally distributed data
- Fortunately, many parametric statistics are “robust” to deviation from the specified distribution

Testing Differences Among Groups

- To do our hypothesis testing, we need to decide on the appropriate statistical method (the test statistic to be used)
- The statistical method to be used depends on answers to the following:

  1) What **type of data** are you comparing between groups (continuous, categorical)?

  2) If the outcome is a continuous variable, what is its distribution (**normal, not normal**)?

  3) Are the data comparing **independent** groups (e.g., measures of cognition in persons with SBP<130 vs. persons with SBP>130) or are the data **paired/matched** in some way (e.g., measures of cognition in hypertensive persons, before and after a specific BP medication).

Testing Differences Among Groups

- Group comparisons by data type

- For **categorical** data, groups are compared with **chi-square tests** (testing if the proportions of subjects in categories differs between groups)

- For **continuous** data, groups are compared with parametric or non-parametric tests (depending on normality of data)
  - **Parametric** (normal outcome data): t-tests (2 groups), analysis of variance (>2 groups)
  - **Non-parametric** (non-normal): Wilcoxon rank sum

Testing Differences Among Groups

- Group comparisons for **matched/repeated** measures

- For **categorical** data, groups are compared with **chi-square tests** that incorporate the matching (McNemar’s test for proportions)

- For **continuous** data, groups are compared with parametric or non-parametric tests, incorporating the matched data
  - **Parametric** (normal outcome data): paired t-tests (2 groups), repeated measures analysis of variance (>2 groups)
  - **Non-parametric** (non-normal): signed rank test

Two independent group comparison: continuous data

- Normal (or fairly normal) outcome data: use independent sample (Student’s) t-test

$$\frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

- H0: mean group 1 = mean group 2

- H1: mean group 1 ≠ mean group 2

Two independent group comparison: continuous data

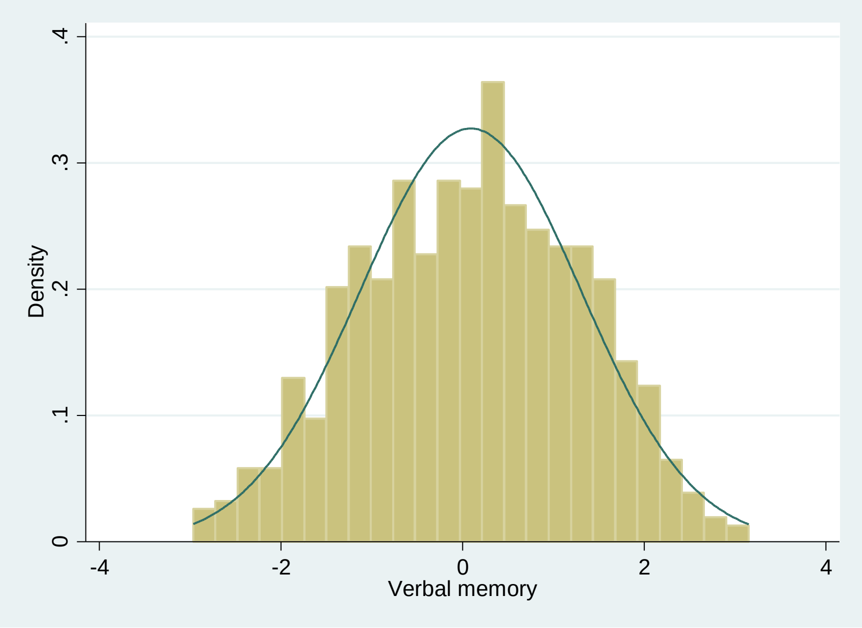
- Not normal outcome: Wilcoxon rank sum

- H0: median group 1 = median group 2
- H1: median group 1 ≠ median group 2

- Non-parametric tests are based on rankings of the data, rather than the values of the data. Ranks are invariant to skewness and other non-normalities of the data

- Rank data overall (irrespective of groups), then compare ranks between groups

Two independent group comparison: continuous data



Normal or not normal?

Two dependent group comparison: continuous data

- Normal (or fairly normal) outcome data: use paired t-test

$$\frac{\bar{d}}{SE(\ \bar{d})}$$

where d are the differences in the outcome value within pairs or within-subject (pre/post values)

- Paired designs remove between-subject variability. When possible, it is a far more powerful design, as within-subject (or within-pairs) is the only source of variability.

- H0: mean difference = 0; H1: mean difference ≠ 0

Group comparisons: categorical data

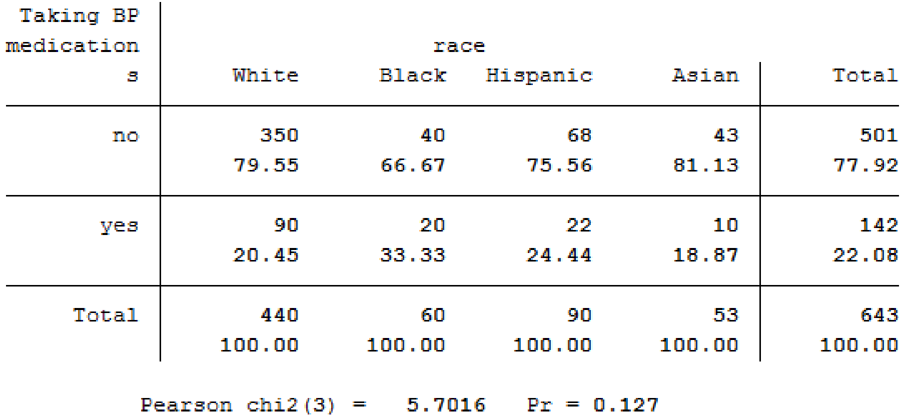
- Example: Use of BP medications by race

- H0: The proportions of postmenopausal women using BP medications does not differ by race (we can also say “BP medications and race are not associated”)

- H1: The proportions of postmenopausal women using BP medications does differ by race

Group comparisons: categorical data

- Table: 20.4% of white, 33.3% of black, 24.4% of Hispanic, 18.9% of Asian women taking BP medications
- Chi-square = 5.70, p-value=0.127
- P>0.05, so do not reject H0. Conclude that use of BP meds does not differ by race in postmenopausal women



```
 Taking BP |
medication |                     race
         s |     White      Black   Hispanic      Asian |     Total
-----------+--------------------------------------------+----------
        no |       350         40         68         43 |       501
           |     79.55      66.67      75.56      81.13 |     77.92
-----------+--------------------------------------------+----------
       yes |        90         20         22         10 |       142
           |     20.45      33.33      24.44      18.87 |     22.08
-----------+--------------------------------------------+----------
     Total |       440         60         90         53 |       643
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =   5.7016   Pr = 0.127
```

Survival Time Data

- Survival time data: Contains two components

  1) If the subject had the event (did the subject die?)

  2) The last time the subject was observed

  E.g., Subject died at age 82

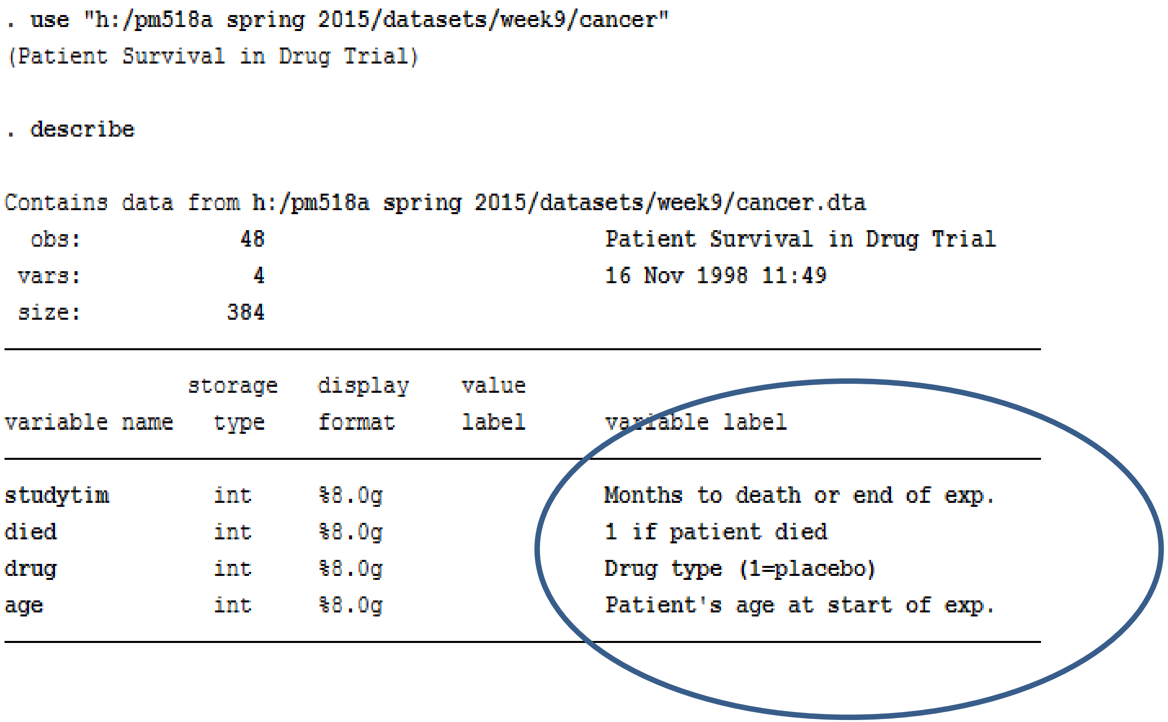
  Subject was alive at age 53 (last age observed on-study)

  Subject died 2.5 years after lung cancer diagnosis

Lifetables

- One method to analyze and graphically present survival data
- Can be used for a single sample, or group comparisons
- Compute and graph the probability of surviving to particular times over the study follow-up
- Example: Patient survival on a cancer clinical trial (n=48 patients)

Lifetables

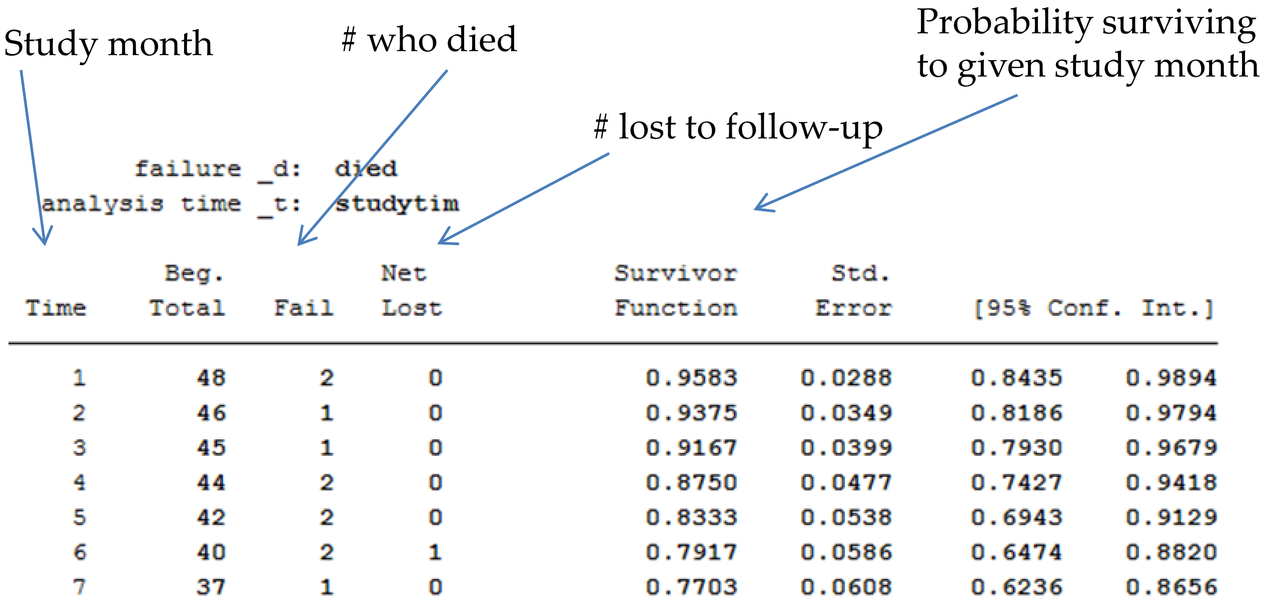


```
. use "h:/pm518a spring 2015/datasets/week9/cancer"
(Patient Survival in Drug Trial)

. describe

Contains data from h:/pm518a spring 2015/datasets/week9/cancer.dta
  obs:            48                          Patient Survival in Drug Trial
 vars:             4                          16 Nov 1998 11:49
 size:           384
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------------
studytim        int    %8.0g                 Months to death or end of exp.
died            int    %8.0g                 1 if patient died
drug            int    %8.0g                 Drug type (1=placebo)
age             int    %8.0g                 Patient's age at start of exp.
-------------------------------------------------------------------------------
```

Lifetables: Compute Survival Over Follow-up



Study month

who died

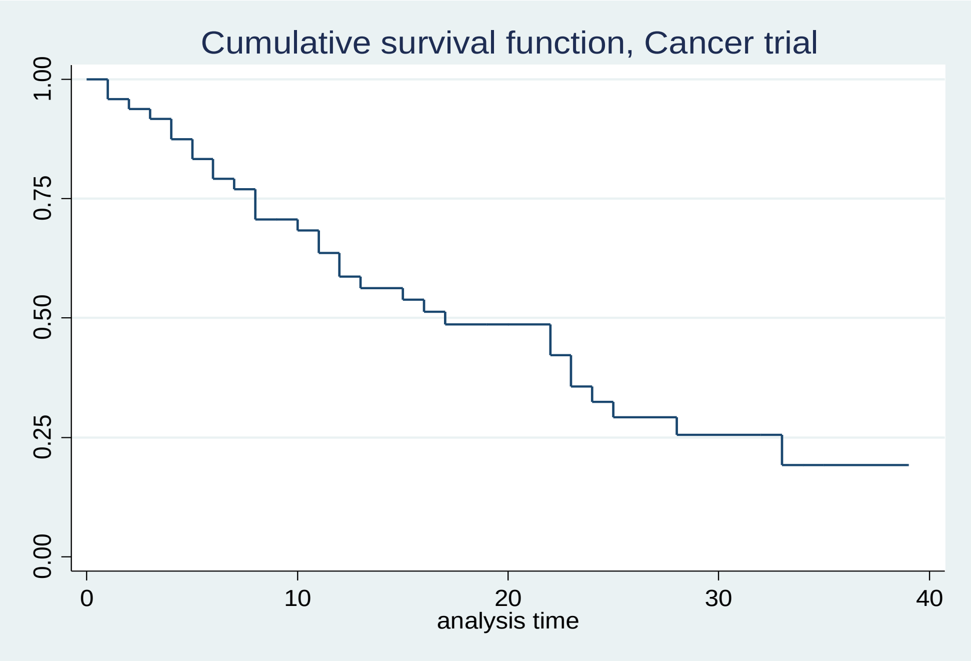
# lost to follow-up

Probability surviving
to given study month

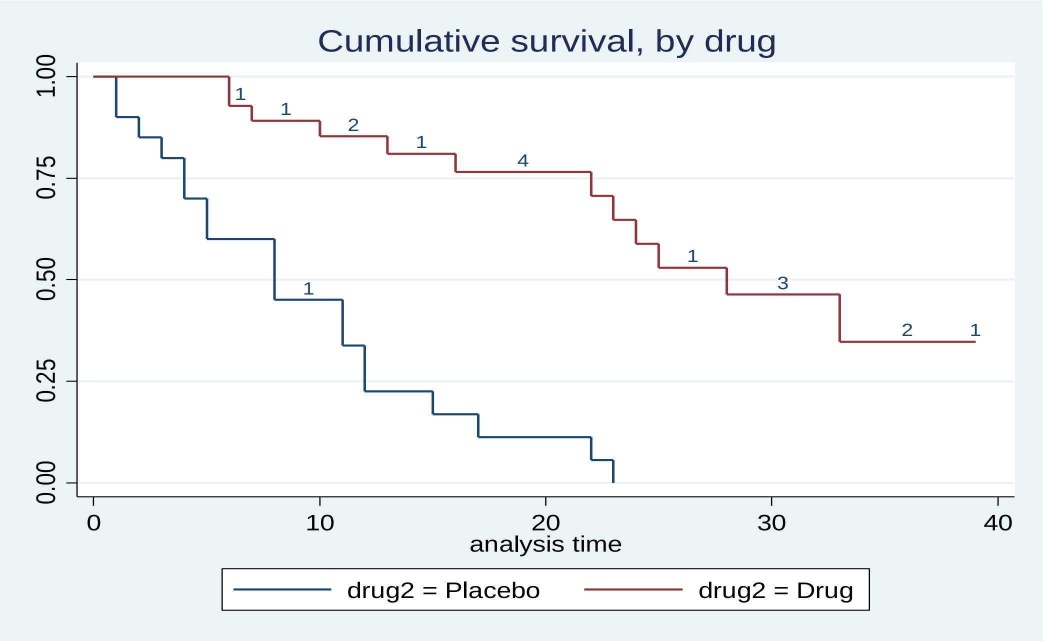
```
      failure _d:  died
analysis time _t:  studytim

             Beg.          Net              Survivor      Std.
   Time     Total   Fail   Lost             Function     Error     [95% Conf. Int.]
-----------------------------------------------------------------------------------
      1        48      2      0               0.9583    0.0288     0.8435    0.9894
      2        46      1      0               0.9375    0.0349     0.8186    0.9794
      3        45      1      0               0.9167    0.0399     0.7930    0.9679
      4        44      2      0               0.8750    0.0477     0.7427    0.9418
      5        42      2      0               0.8333    0.0538     0.6943    0.9129
      6        40      2      1               0.7917    0.0586     0.6474    0.8820
      7        37      1      0               0.7703    0.0608     0.6236    0.8656
```

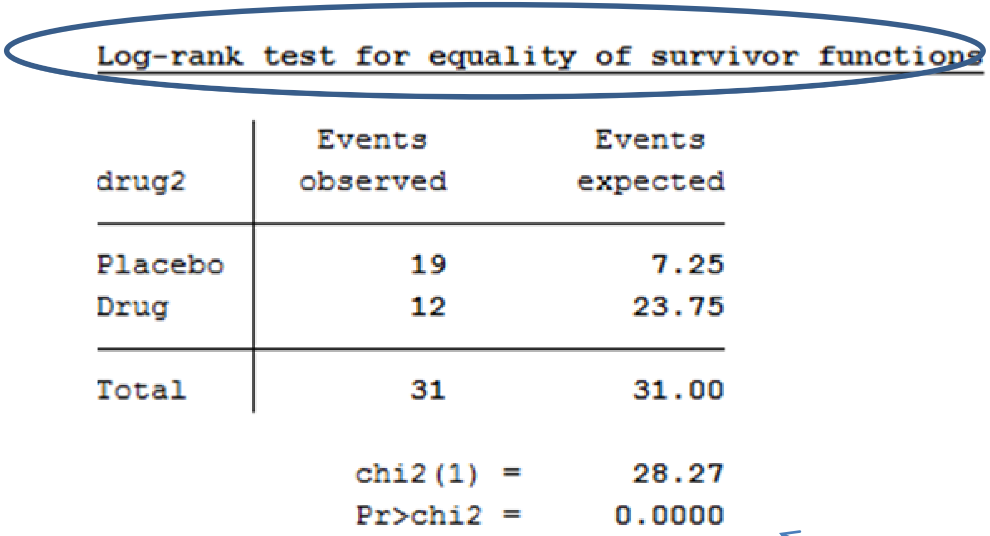
Lifetables Single Group: Graph Survival Over Follow-up



Lifetable Group Comparisons: Graph Survival Over Follow-up



Lifetables: Test for Group Differences in Survival Curves



```
Log-rank test for equality of survivor functions

             |   Events         Events
drug2        |  observed       expected
-------------+-------------------------
Placebo      |        19           7.25
Drug         |        12          23.75
-------------+-------------------------
Total        |        31          31.00

                   chi2(1) =      28.27
                   Pr>chi2 =     0.0000
```

H0: Placebo survival curve = Drug survival curve

HA: Placebo survival curve ≠ Drug survival curve

P<<0.05
Reject H0

Measures of Association

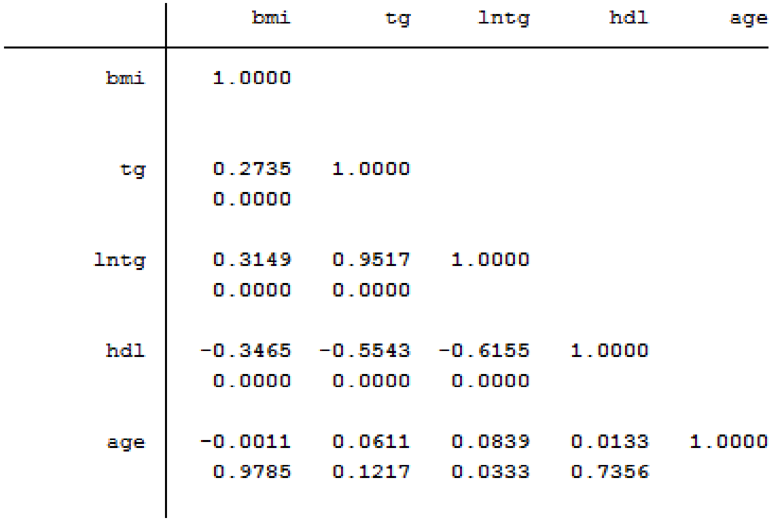
- Rather than evaluate group differences, we may want to just state how two or more variables are associated or correlated

- For continuous variables, the Pearson's correlation (r) is a simple measure of **linear** correlation

- Pearson's r assumes normal distribution of variables.
  Non-parametric (for non-normal data) version is Spearman's correlation

- For both Pearson and Spearman correlations, r ranges from -1 to 1, with 0 representing uncorrelated variables

Measures of Association

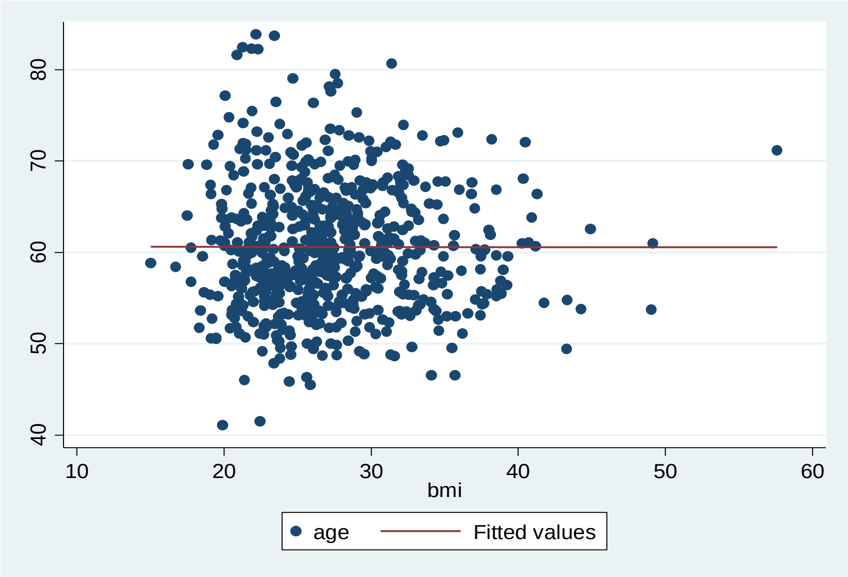


```
             |      bmi       tg     lntg      hdl      age
-------------+---------------------------------------------
         bmi |   1.0000
             |
             |
          tg |   0.2735   1.0000
             |   0.0000
             |
        lntg |   0.3149   0.9517   1.0000
             |   0.0000   0.0000
             |
         hdl |  -0.3465  -0.5543  -0.6155   1.0000
             |   0.0000   0.0000   0.0000
             |
         age |  -0.0011   0.0611   0.0839   0.0133   1.0000
             |   0.9785   0.1217   0.0333   0.7356
             |
```

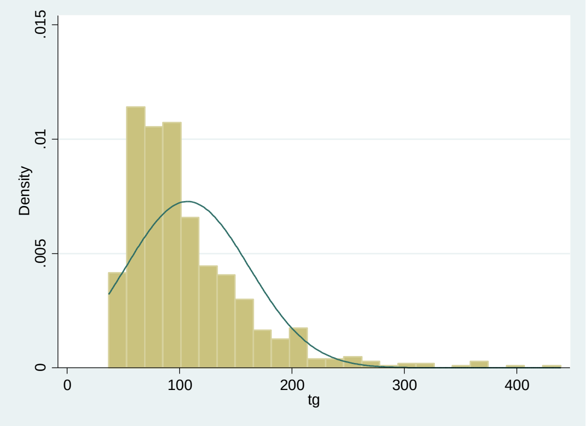
Top number = correlation, bottom number = p-value
H0: r = 0 (no correlation)

Measures of Association

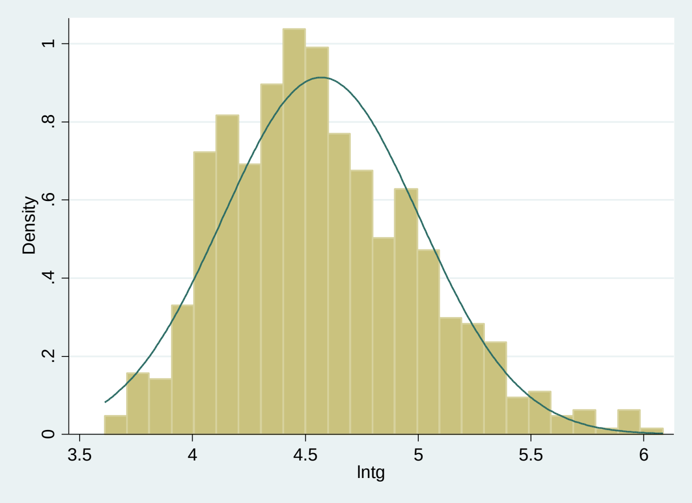
- Uncorrelated: age and BMI



Measures of Association



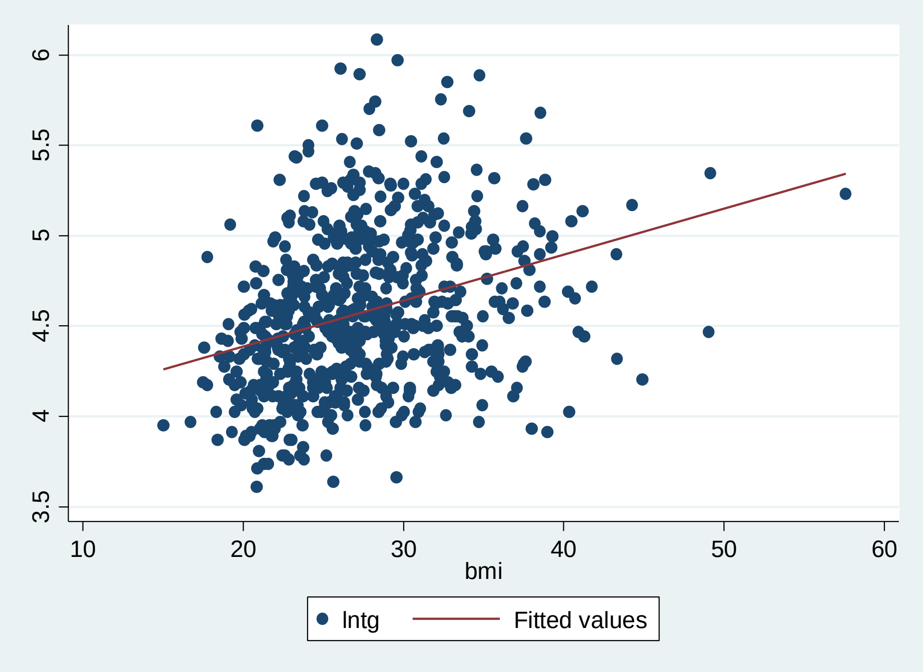
Triglycerides



Log(triglycerides)

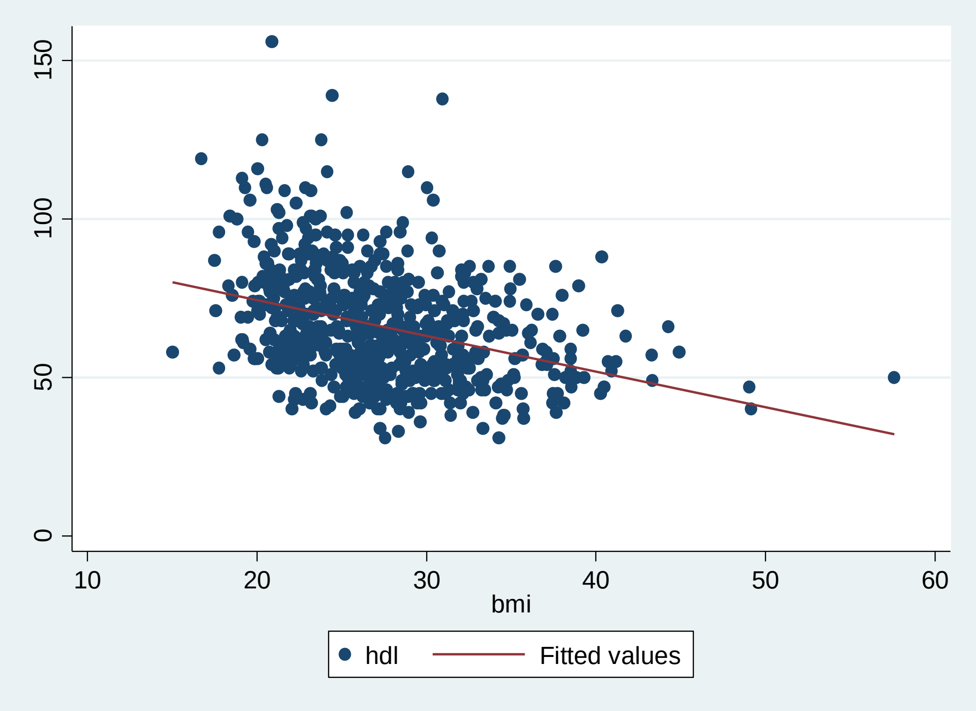
Measures of Association

- Positively correlated: Triglycerides (log transform) and BMI



Measures of Association

- Negatively correlated: HDL and BMI



SC CTSI

Coefficient of Determination (R^2)

- Square of correlation coefficient
- Proportion of variability in Y (e.g., HDL) that can be explained by its linear correlation with X (e.g., BMI)
- $r = -0.3465$
- $R^2 = 0.12$ (12% of variation in HDL can be explained by its linear correlation with BMI)

Linear Regression

- Describes the LINEAR relationship between two variables.
- With Y a continuous variable:

  $Y = a + bX$

- Y = “dependent variable”
- X = “independent variable”
- Estimate a = intercept (predicted value of Y when X = 0)

Linear Regression

- Y = a + bX

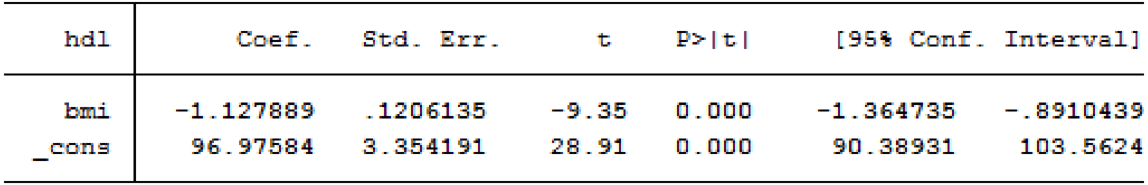
- Estimate b = slope (linear association between X and Y; predicted change in Y per unit change in X)

  H0: b (slope) = 0 (no linear association between X and Y)

- Direction of b reflects the correlation (r)
  b < 0 indicates a negative association
  b > 0 indicates a positive association

Linear Regression

- In our BMI, HDL example



```
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |  -1.127889   .1206135    -9.35   0.000    -1.364735   -.8910439
       _cons |   96.97584   3.354191    28.91   0.000     90.38931    103.5624
------------------------------------------------------------------------------
```

HDL = 96.98 – 1.13 BMI

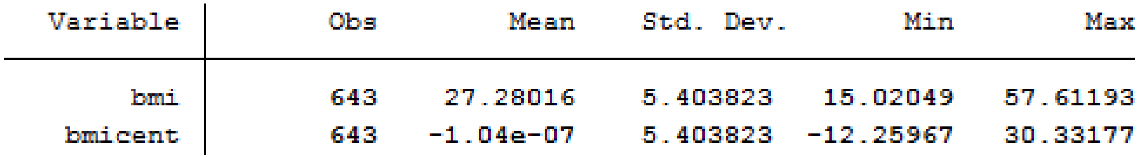
Slope: HDL decreases by 1.13 (mg/dL) per unit (kg/m^2) of BMI

P-value for H0: slope = 0 is <0.001

Intercept? HDL = 96.98 for persons with BMI=0 !!!!

Linear Regression

- To make some better sense of the intercept



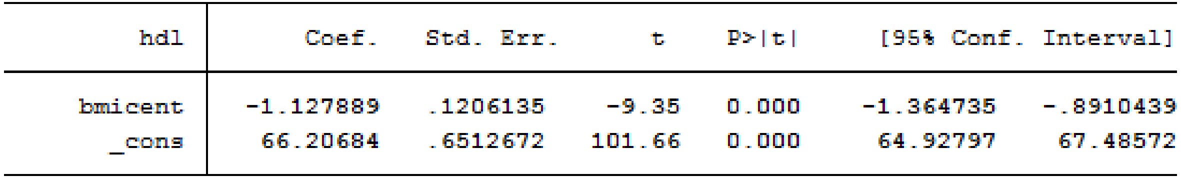
```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         bmi |       643    27.28016    5.403823   15.02049   57.61193
     bmicent |       643   -1.04e-07    5.403823  -12.25967   30.33177
```

BMICENT = BMI – mean(BMI)

BMICENT = 0 when person is at the mean level of BMI (when BMI = 27.28)

Linear Regression

- To make some better sense of the intercept



```
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmicent |  -1.127889    .1206135    -9.35   0.000    -1.364735   -.8910439
       _cons |   66.20684    .6512672   101.66   0.000     64.92797    67.48572
------------------------------------------------------------------------------
```

BMICENT = BMI – mean(BMI)

HDL = 66.21 – 1.13 BMI

HDL decreases by 1.13 (mg/dL) per unit (kg/m^2) of BMI

P-value for H0: slope = 0 is <0.001

Intercept: HDL = 66.21 for persons with BMICENT = 0
(i.e., when BMI = 27.28)

Multiple Linear Regression

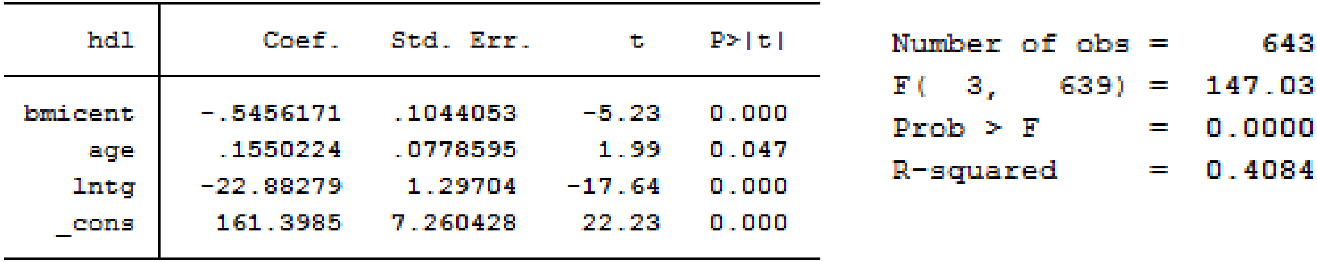
- Linear association model with a continuous outcome (dependent) variable, multiple independent variables

- $Y = a + b_1X_1 + b_2X_2 + \ldots$

- Coefficient of determination (R^2) is the proportion of variation in Y that can be explained by all of the X independent variables

Multiple Linear Regression

- HDL example



```
         hdl |      Coef.   Std. Err.       t    P>|t|        Number of obs =      643
-------------+----------------------------------------        F(  3,    639) =   147.03
     bmicent |  -.5456171   .1044053    -5.23   0.000         Prob > F      =   0.0000
         age |   .1550224   .0778595     1.99   0.047         R-squared     =   0.4084
        lntg |  -22.88279    1.29704   -17.64   0.000
       _cons |   161.3985   7.260428    22.23   0.000
```

HDL = 161.40 - 0.546(bmicent) + 0.155(age) = 22.88(lntg)

H0: bmicent slope = 0, p <0.001
H0: age slope = 0, p = 0.047
H0: lntg slope=0, p <0.001
R^2 = 0.4084 (40.84% of variance in HDL is explained by its linear relationships with BMI, age and triglycerides)

Other Regression Models

- There are many types of such regression models. The type of regression model used depends on what type of data the **outcome (dependent)** variable is. You must select the correct regression approach to match your dependent variable!

- **Continuous** outcome: **linear** regression – do independent (X) variables relate to **the levels** of Y? (e.g., levels of HDL cholesterol)

- **Dichotomous** outcome: **logistic** regression – do independent (X) variables relate to **the probability** that Y=1 (vs Y=0)? (e.g., that a mouse survived versus died within 30 days after experimental exposure)

Other Regression Models

- **Ordinal** outcome: **ordinal** logistic regression – do independent (X) variables relate to **the probability** that Y = higher compared to lower level? (e.g., animal behavior is frozen, moving but unorganized, moving and organized)

- **Nominal** outcome: **multinomial** logistic regression – do independent (X) variables relate to the probability that Y = category 1 (vs category 2, 3, etc.)? (e.g., subject healthy, MI, stroke)

Other Regression Models

- **Count** outcome: **Poisson or negative binomial** regression – do independent (X) variables relate to the count Y (e.g., # of hospital days, # of incorrect turns in a maze)

- **Survival** outcome: Cox (proportional hazards) or other “survival” regression – do independent (X) variables relate to the event rate? (e.g., rate of incident dementia among an initially cognitively healthy population)

Uses of Regression Models:
Association vs. Prediction

- The two primary uses of regression models are in association and prediction

- **Association**: Research question and hypotheses relate to the association between the dependent (outcome) variable and **specific** independent variable(s)

- Objective: Do a good job at estimating the magnitude of the association (e.g., the slope) and making inferences about that association

- Does BMI relate to the levels of HDL cholesterol? What is the direction and magnitude of the association?

Uses of Regression Models: Association

- Use multivariable regression models to adjust for other independent variables that might:
  - **Explain** the association
    When I adjust for age, is the association of BMI with HDL still statistically significant?
  - **Confound** (bias) the association of interest
    When I adjust for age, does the slope estimate for BMI with HDL change?
  - **Modify** the association of interest
    Is the association (slope) estimate for BMI with HDL different in persons <60 vs 60 and older?

Uses of Regression Models: Prediction

- In contrast to association models**, prediction models** are not concerned with estimating specific associations

- Objective of prediction models: find the set of independent variables (X) that do the best job of predicting the dependent (outcome) variable

- Uses: Clinical decisions, who will benefit from a treatment, identifying high risk patients, diagnostics

Uses of Regression Models: Prediction

- Prediction models heavily rely on multivariable (many) independent variables, as a single independent variable is usually not a good predictor of an outcome variable

- Along with the prediction model, one must assess the adequacy of prediction: compare the “predicted” outcome (from your model) to the actual value of the outcome for each subject. How well does the model do?

Uses of Regression Models: Prediction

- Prediction models are usually **over-optimistic**. The prediction model was specifically developed to match the observed outcome variable as closely as possible **IN YOUR SAMPLE**.

- However, when applied to an **independent** dataset, the prediction model does not do as well. It is essential that predictive models be evaluated and validated in independent samples. **(internal validity)**

  How well does your model do in predicting outcomes in a new sample of subjects from the same population?

Uses of Regression Models: Prediction

- Also, be careful about applying a prediction models to populations that were not part of the model development sample. **(external validity)**

  If you developed a great predictive model for fracture risk in postmenopausal women, do not expect it to be applicable to premenopausal women.

  We apply the Framingham 10-year coronary risk model to everybody!

Classification

- Classification of patients (generally into two categories) based on:

  1) **Predictive model** (e.g., multivariable predictive model for probability of mortality (mortality risk) in burn patients)

  2) **Value of a laboratory variable/biomarker** (e.g., for diagnosis, to identify persons at high risk for diabetes, burn patients at high risk for mortality)

Classification

- For a continuous variable Y (e.g., predicted probability of death, HbA1c), we can calculate patient classification characteristics at different cutpoints (c)

- Sensitivity = true positive rate = P(Y>c | D)
  Proportion of diseased (or whatever the outcome to be predicted) that have a value of Y greater than the cutpoint

- Specificity = true negative rate = P(Y<c | no D)
  Proportion of non-diseased (persons without the outcome to be predicted) that have a value of Y less than the cutpoint

SC CTSI

Classification

- False positive rate = 1 – specificity = P(Y>c | no D)
  Proportion of non-diseased that have a value of Y greater than the cutpoint

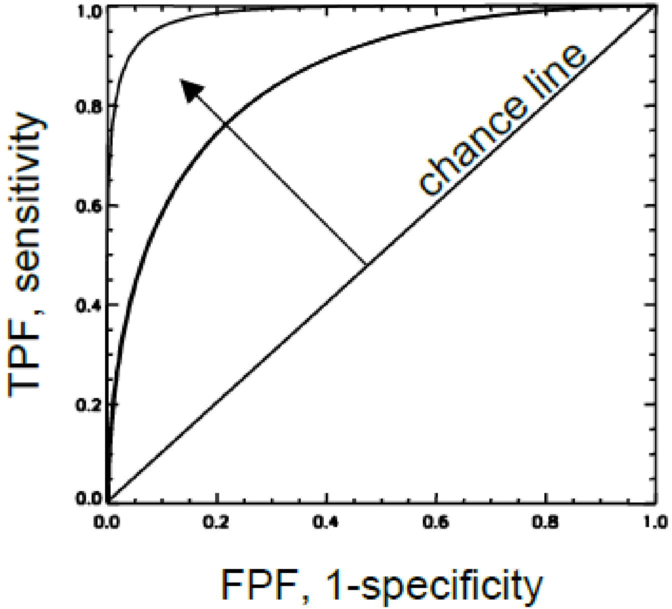
Classification

- For a continuous laboratory variable (or a model-predicted probability of disease), we can compute the sensitivity and specificity for many levels of cutpoints over the range of the variable

- A receiver operating characteristic (ROC) curve, is a graphical representation of this, plotting sensitivity (true positive rate) versus 1 – specificity (false positive rate) over values of c (possible cutpoints)

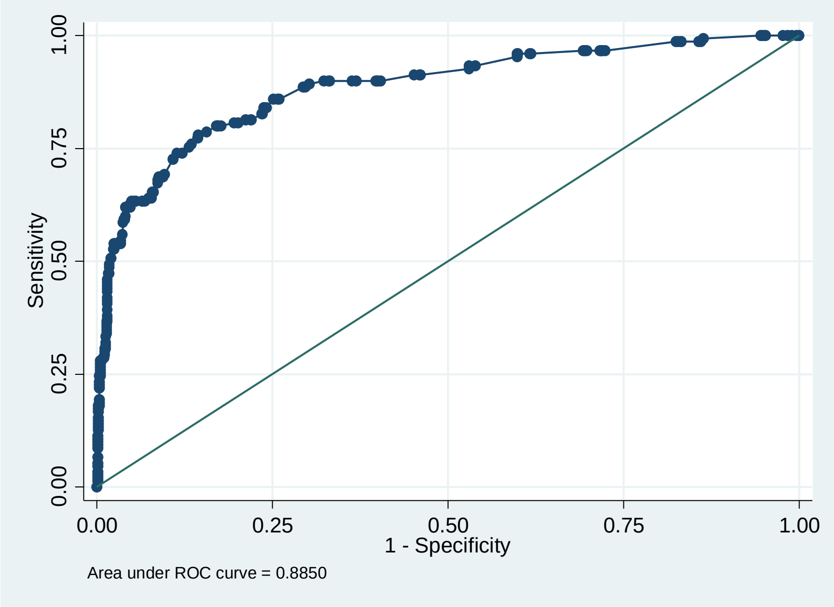
- The area under the ROC curve (AUC) is a measure of how well the variable is classifying (dead vs alive, diabetes vs not, etc.)
  AUC = 1 Perfect classification
  AUC = 0.5 No better than chance

Classification



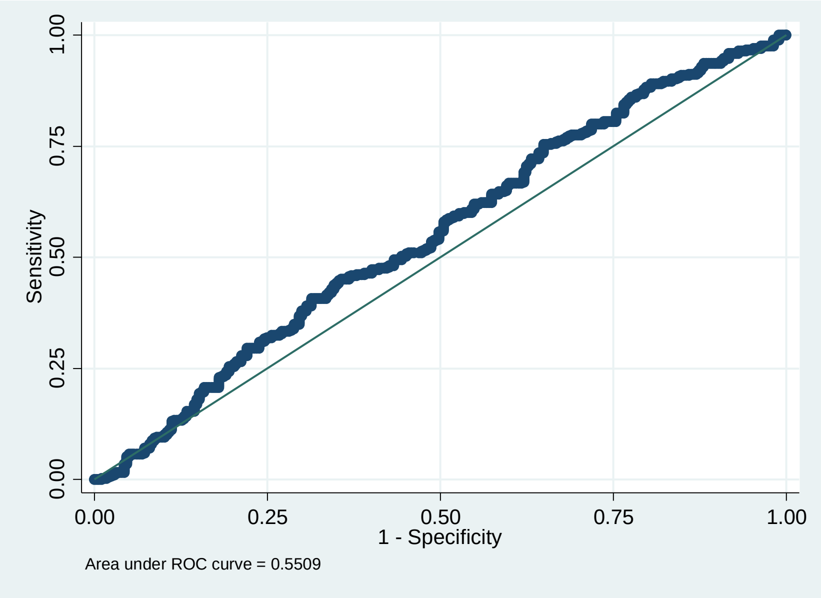
Note a tradeoff between sensitivity and specificity. As one increases, the other will decrease

Classification



Example above: Using percent burn area to classify die/survive

Classification



Example above: Using BMI to classify high LDL (>130, <=130)

Summary and caveats

- This is obviously a very broad overview of an array of analytic methods that may or may not be appropriate to your data.

- Think about your data:
  - What type of data do I have (continuous, ordinal, dichotomous, normal, non-normal)?
  - What are my hypotheses (group comparisons? Correlations? Associations? Predictions?)
  - What possible analytic approaches might be appropriate to my data, to test my hypotheses?

Summary and caveats

- We have NOT covered analytic methods for correlated outcome data, for example arising from:
  - Longitudinal data: repeatedly measured in the same subject/animal over time
  - Correlated units (families, classrooms, etc.)

  There are regression techniques for such correlated data, similar to those that we have summarized above, with regression techniques specific to the type of dependent variable (continuous, dichotomous, etc.).

  Be aware of the possible correlations in your outcome data in developoing your analytic plan and in talking with your statistician.

CTSI Biostatistics (BERD): a resource for you at USC

- Biostatisticians to help you with study design, sample size estimation, data management plan, statistical analyses, and summarizations of your methods and results

- Recharge center

- To request a consult:

  http://sc-ctsi.org/index.php/new-resources/get_expert_advice