BASIS OF STATISTICS AND HYPOTHESIS TESTING

ALEN DELIC, MS

ADAM DE HAVENON, MD, MS

PLAN FOR THE YEAR

- We will give 5-6 lectures covering some topics that will be basic and others more advanced
    - Basics of Statistics
    - Regression, Survival Analysis, and Model Fitting
    - Clinical Trial Design and Sample Size
    - Causal Inference #1
    - Causal Inference #2
    - Epidemiologic Approaches and Risk Scores

WHAT IS STATISTICS AND WHAT ARE ITS USES

- We can split statistics into two branches:
  - Descriptive statistics
    - Used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data
  - Inferential statistics
    - The use of samples of data to make inferences on the population
    - Involves the use of hypothesis testing via a large range of statistical methods, i.e regressions, survival analysis.

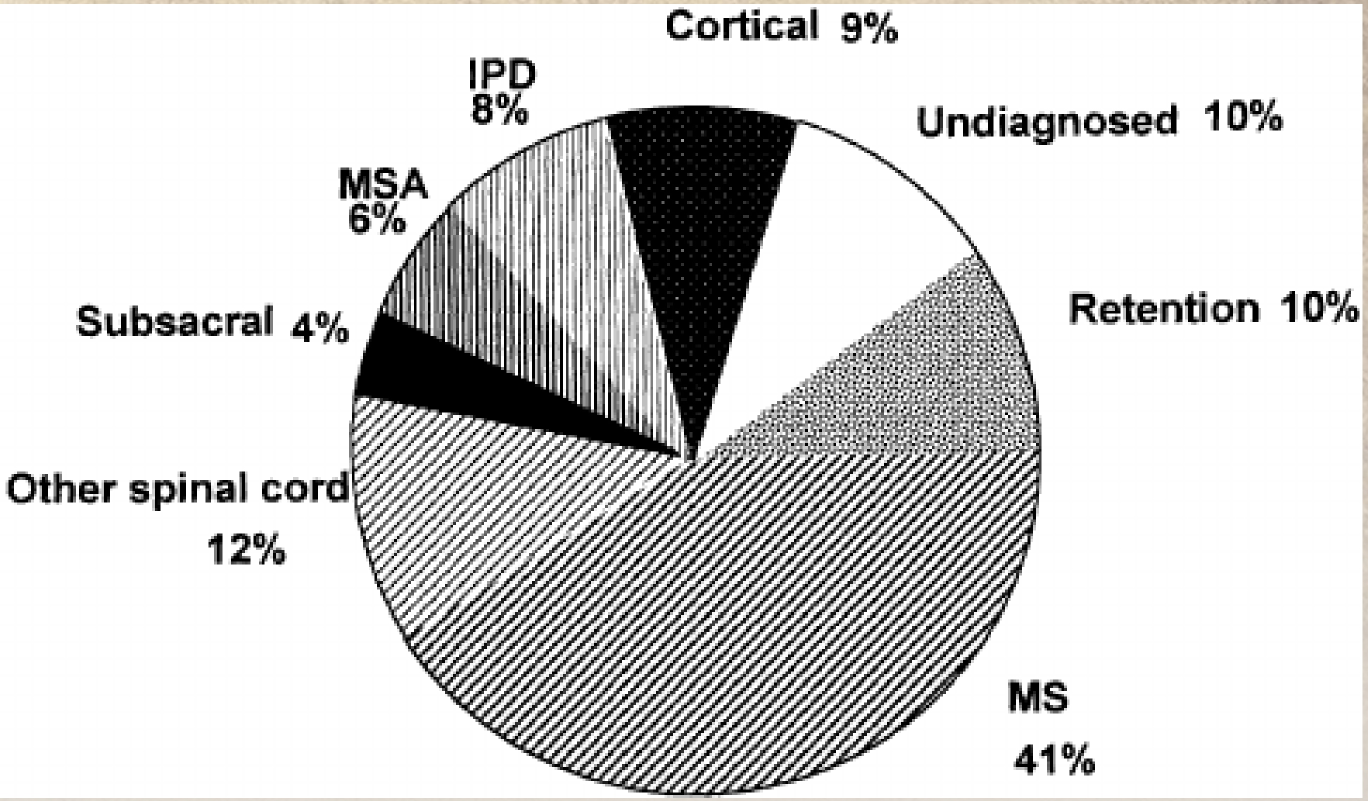
BASIS OF STATISTICS

- To begin diving into statistics, we first have to dive into the order of measurement scale.
- The choice of your test statistic or descriptive statistic will be based on the measurement scale of your data.
- When you determine which order your data is in, this will dwindle down the possible list of statistics that can be applied to it.

Measurement Scale	Combination of Elements	Description	Additional Notes
Nominal	Name	Unordered categories Cancer therapies: chemo, radiation, surgery	Can't do arithmetic
Dichotomous (binary)	Special case of nominal scale	Sex: Male or Female	Binary scale can be considered an interval scale with one category to represent 0
Ordinal	Name + order	Ordered categories Quality of Life: Bad, OK, Great	Can't do arithmetic
Interval (continuous)	Name + order + equal intervals + arbitrary zero point	Continuous measurement with arbitrary zero Body temperature: 0°F does not imply absence of temperature. Also no sense in saying that 75°F is 1.5x as hot as 50°F	Can do arithmetic. Test statistics equal for both interval and ratio.
Ratio	Name + order + equal intervals + absolute zero point	Continuous measurement with absolute zero Hematocrit: 0% is absence of hematocrit. 50% is 2x more than 25%	Can do arithmetic. Test statistics equal for both interval and ratio.

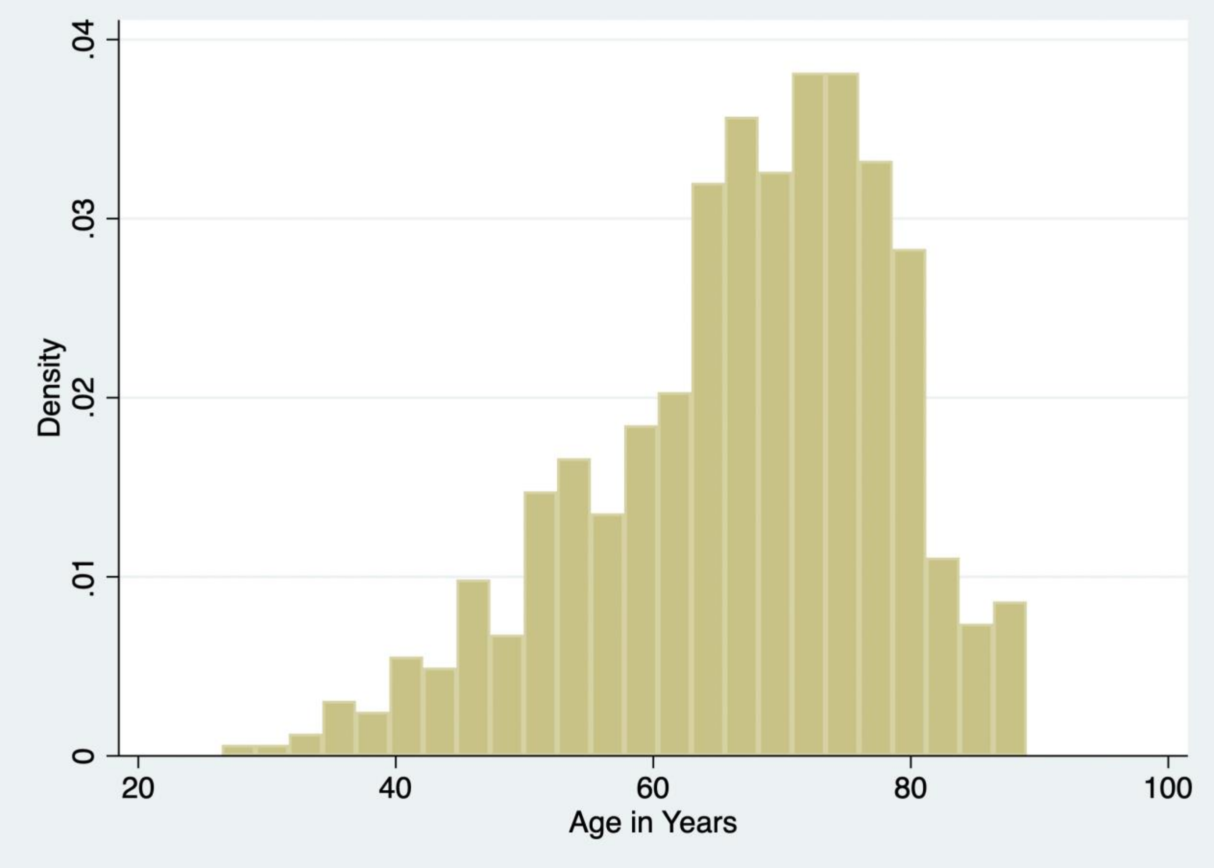
BASIS OF DESCRIPTIVE STATISTICS

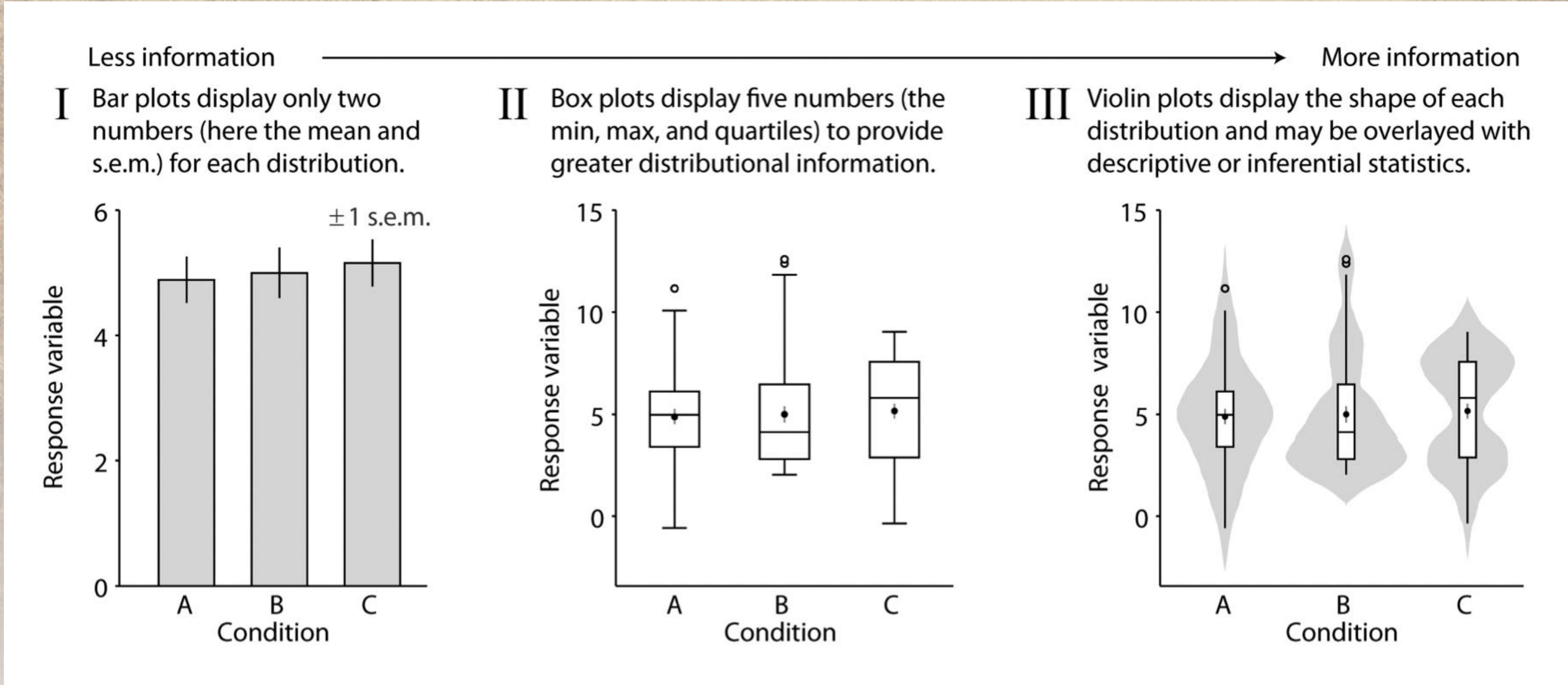
- For nominal and binary data, we are limited with the amount of descriptive stats we can report
  - Limited to counts and percentages. For graphical output, histograms and pie charts are mostly used.



BASIS OF DESCRIPTIVE STATISTICS

- For interval and ratio (both continuous), there are many more options to describe this type of data
  - Mean and Standard Deviation: Used when there is a relatively normal, bell-shaped distribution with minimal outliers that would skew data
  - Median and IQR: Used when there is a skewed distribution or high amount of outliers. Can be combined with box plot to show distribution of data.

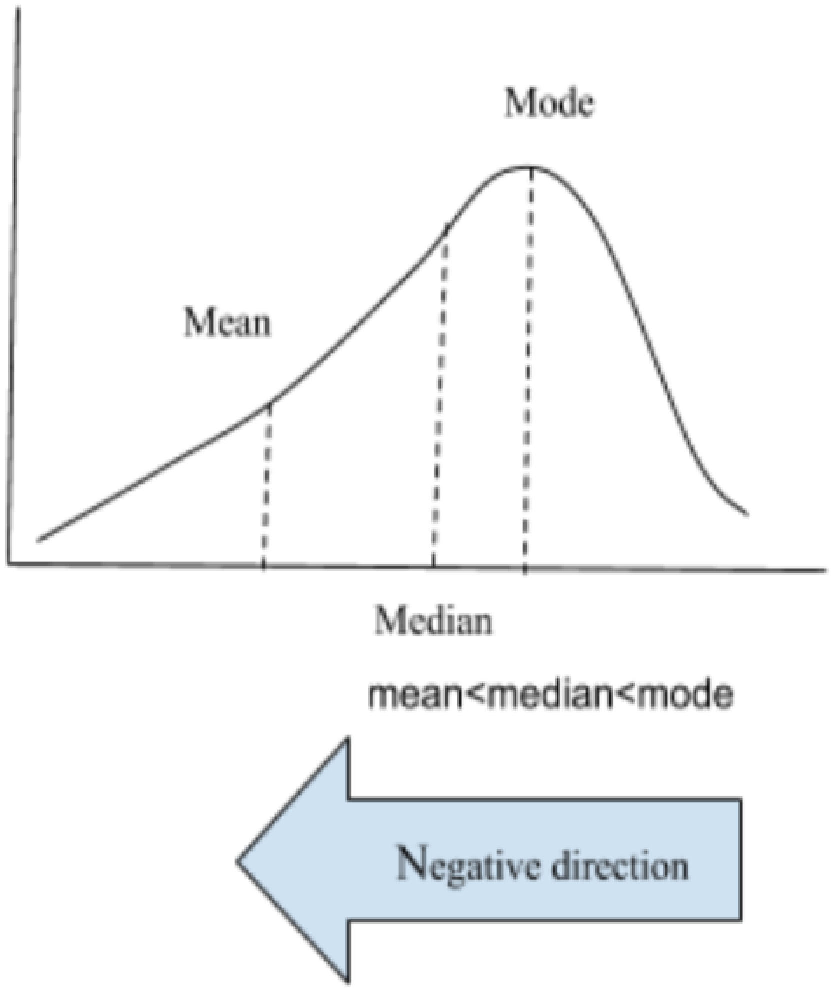




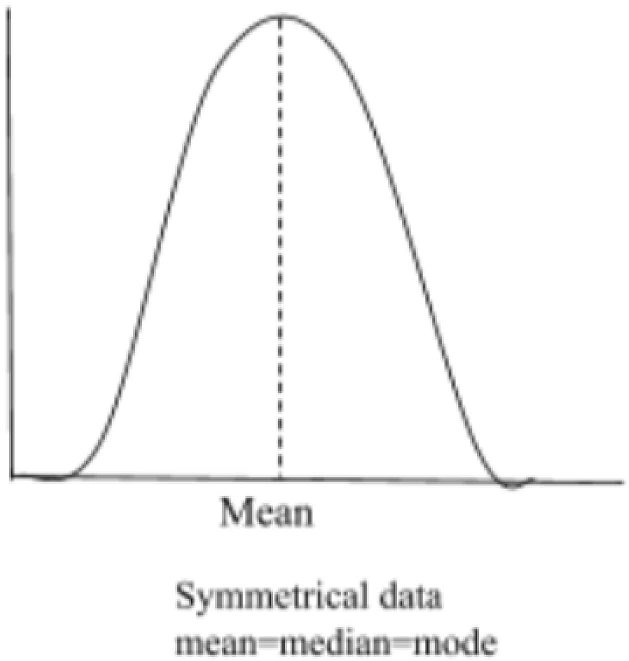
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4427844/pdf/nihms-687401.pdf

A BRIEF TANGENT ON SKEW AND OUTLIERS

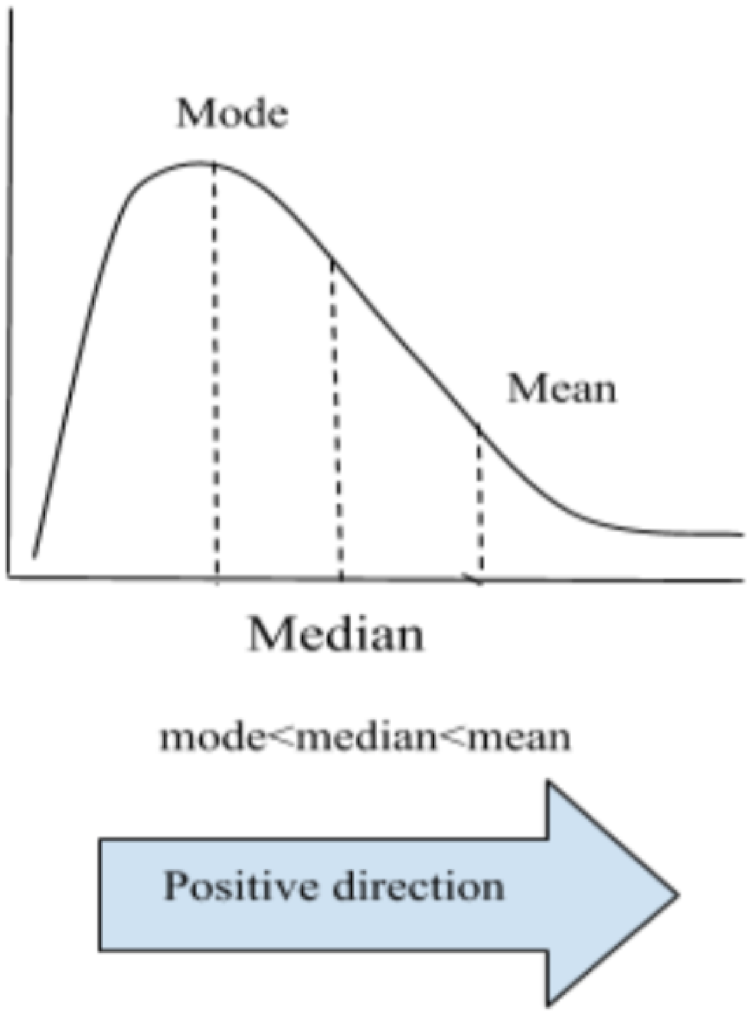
Left Skew

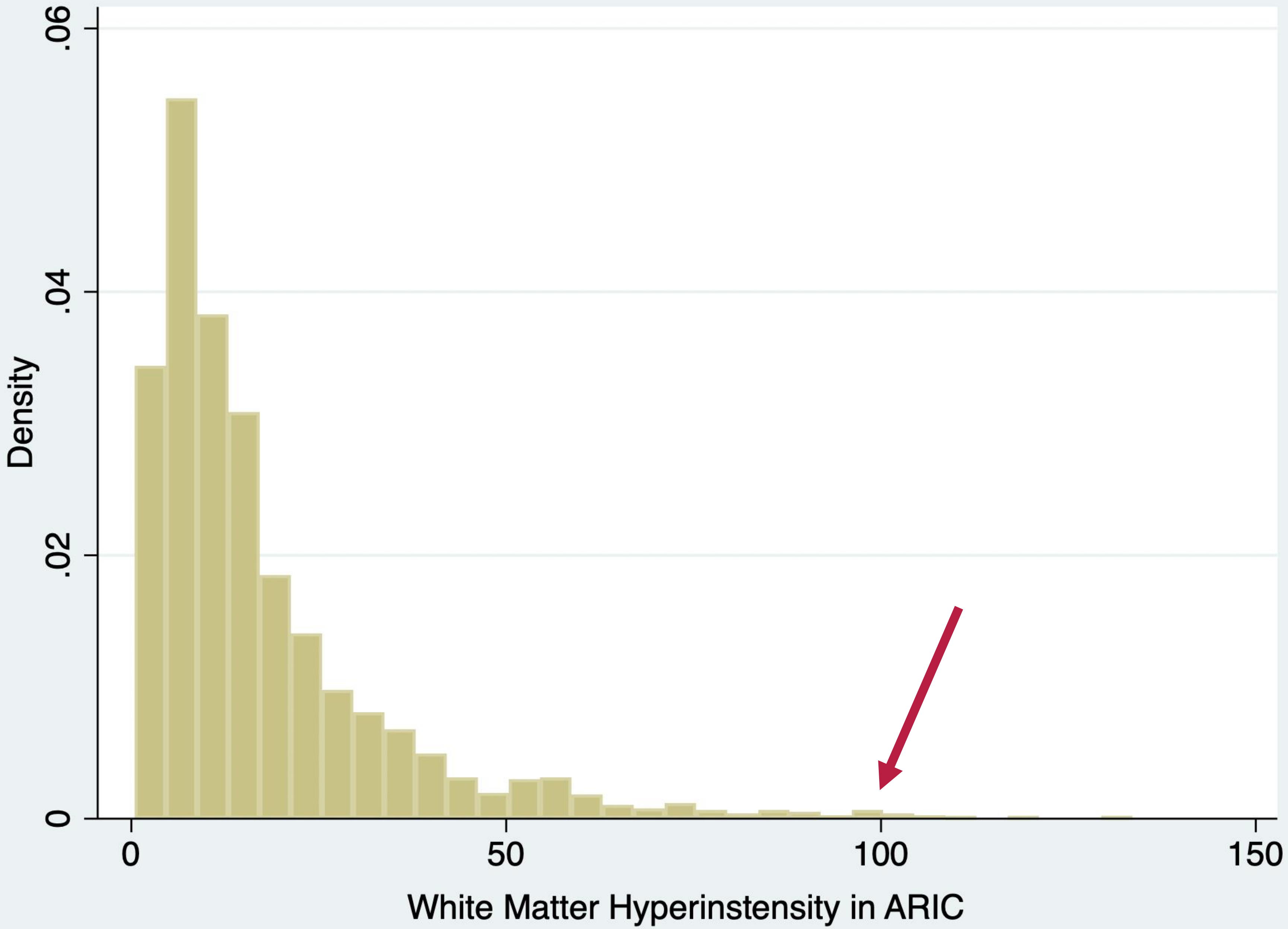


Symmetric Skew (No Skew)



Right Skew

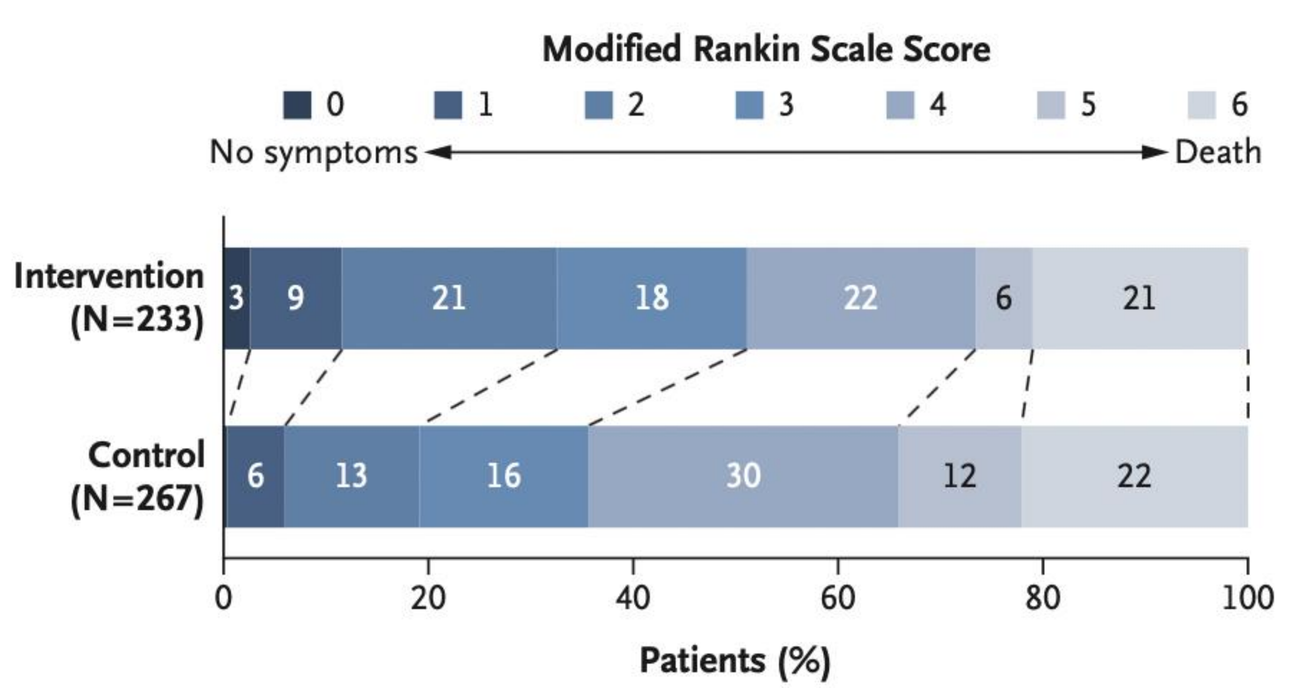




Density
0
.02
.04
.06
0
50
100
150
White Matter Hyperinstensity in ARIC

BASIS OF DESCRIPTIVE STATISTICS

- For ordinal data, there is a combination of metrics from nominal and continuous to describe this data
  - Median and IQR, perhaps range. Can also use relative counts and frequencies



DESCRIPTIVE STATS IN TABLE 1

Table 1. Demographic and Clinical Characteristics of the Patients at Baseline (Intention-to-Treat Population).*

Characteristic	Trial 1		Trial 2	
	Tirbanibulin (N=175)	Vehicle (N=176)	Tirbanibulin (N=178)	Vehicle (N=173)
Age — yr	69.5±8.6	70.2±9.4	69.1±8.7	70.2±8.9
Male sex — no. (%)	147 (84)	154 (88)	158 (89)	150 (87)
White race — no. (%)†	175 (100)	175 (99)	177 (99)	173 (100)
Fitzpatrick skin type I or II — no. (%)‡	123 (70)	142 (81)	126 (71)	120 (69)
Median count of actinic keratosis lesions (IQR)	6 (5–7)	6 (5–7)	6 (5–7)	6 (5–7)
Face:scalp ratio of patients with the specified application location — no.§	119:56	121:55	119:59	118:55
History of treatment for actinic keratosis on face or scalp — no. (%)	145 (83)	153 (87)	132 (74)	125 (72)
History of skin cancer — no. (%)	77 (44)	89 (51)	75 (42)	72 (42)

* Plus–minus values are means ±SD. The intention-to-treat population comprised all the patients who had undergone randomization. IQR denotes interquartile range.

† Race was determined by the investigator.

‡ Fitzpatrick skin types range from I to VI: type I indicates always burns, never tans; type II, usually burns, tans minimally; type III, sometimes mildly burns, tans uniformly; type IV, burns minimally, always tans well; type V, very rarely burns, tans very easily; and type VI, never burns.

§ Enrollment across patients was controlled to achieve a 2:1 ratio of facial:scalp treatment areas (i.e., to enroll twice as many patients with facial lesions as those with scalp lesions).

INTRODUCTION TO INFERENTIAL STATISTICS

- When we try to find associations between two variables, we go into the realm of hypothesis testing

- We will start in discussing what is a hypothesis test and how to interpret the results of one

INTRODUCTION TO INFERENTIAL STATISTICS

- Hypothesis testing
  - A researcher asks a question and we make a hypothesis about the outcome (Ex: Is Drug A an effective treatment for Disease B)
    - Null hypothesis: there is no effect of Drug A on Disease B
    - Alternative hypothesis: there is an effect of Drug A on Disease B

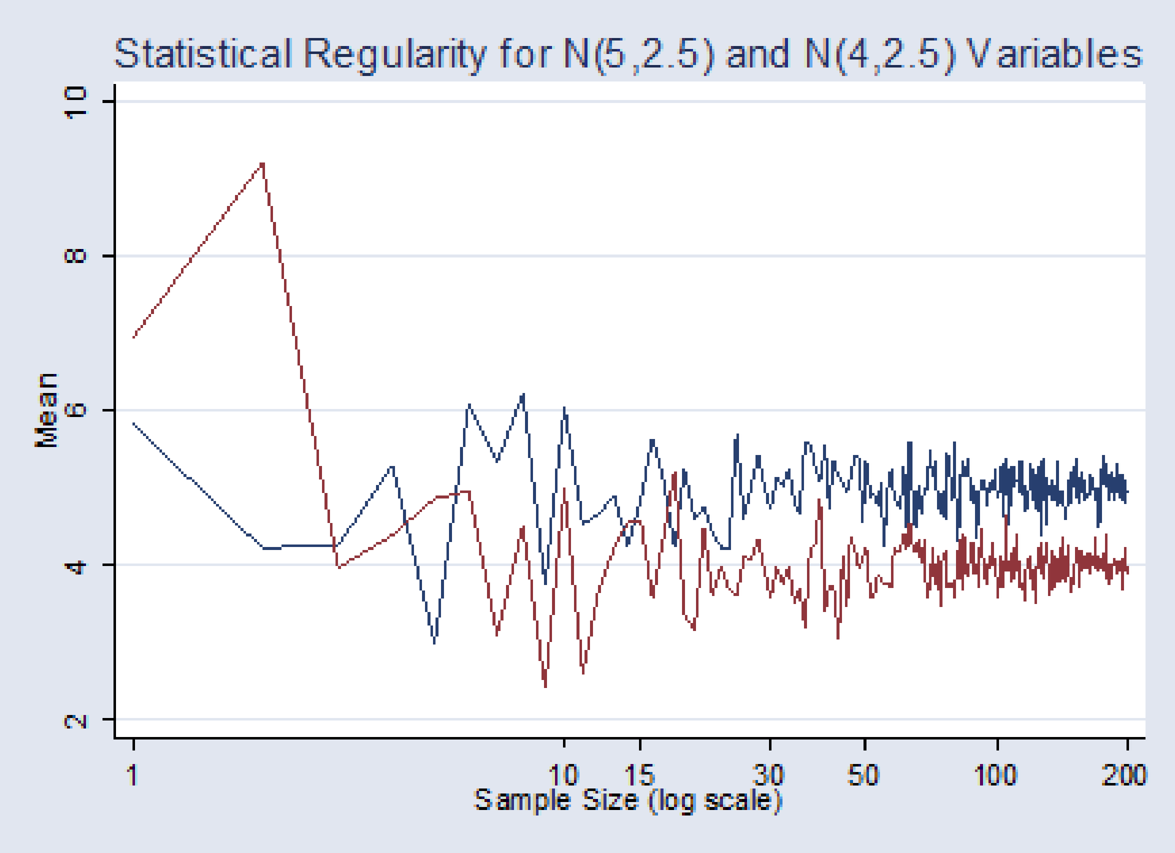
INTRODUCTION TO INFERENTIAL STATISTICS

- We look at the specific level of measurements of the variables and find the correct test to test the association of the relationship

  - Obtain data to enter into the test and see if we reach "statistical significance" by the interpretation of p-values and confidence intervals

INTRODUCTION TO INFERENTIAL STATISTICS

- As we obtain data, the concept of 'Statistical Regularity' and a relation to the Law of Large Numbers occurs = the signal-to-noise ratio becomes distinct as we gather more and more data



STATISTICAL SIGNIFICANCE

- P-values – used to determine if a sample estimate is significantly different from a hypothesized value.
  - The p-value is the probability that the observed effect within the study would have occurred by chance if, in reality, there was no true effect.
  - Standard conventions are set at $p = .01, .05$ and $.10$ for statistical significance

STATISTICAL SIGNIFICANCE

- Statistical significance is the likelihood that the results we obtained are due to chance.

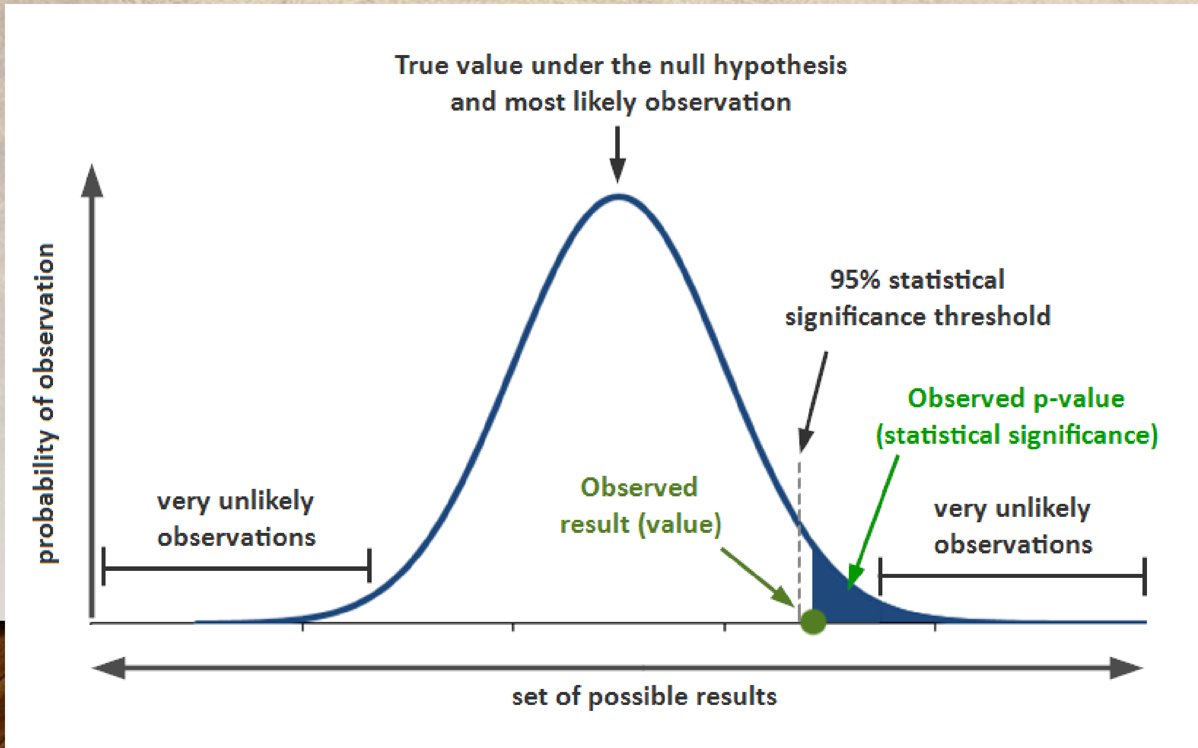
  - This is different from "clinical" significance as just because a result is statistically significant, does not imply it is clinically significant
    - Example: A statistically significant effect size ($p<.05$) of an experimental drug showing a 0.01% decrease in chance of disease may not be clinically significant

STATISTICAL SIGNIFICANCE

- Confidence Intervals – function of the sample mean estimate and sample mean standard deviation.
  - Different interpretation of significant for various types of statistics. i.e crossing 1.00 for ratio statistics and 0.00 for non-ratio statistics
  - Can be used to gauge the estimated effect size from the sample data
    - Example: In an adjusted regression model, Drug A was associated with a 12 point decrease in blood pressure [95% CI (8.2, 15.8)]

CALCULATING THE P-VALUE

- P – value = the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct
  - A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.



		Null hypothesis is	
		TRUE	FALSE
The null hypothesis was	rejected ($P < \alpha$)	**Type I error**, false positive probability = α	true positive probability = $1 - \beta$ (**power of the test**)
	not rejected ($P \geq \alpha$)	true negative probability = $1 - \alpha$	**Type II error**, false negative probability = β

CHOICE OF SIGNIFICANCE TESTS (NO CONFOUNDING)

Level of Measurement of Dependent Variable	Two Independent Groups	Three or more Independent Groups	Two Correlated* Samples	Three or more Correlated* Samples
Dichotomous	chi-square or Fisher's exact test	chi-square or Fisher-Freeman-Halton test	McNemar test	Cochran Q test
Unordered Categorical	chi-square or Fisher-Freeman-Halton test	chi-square or Fisher-Freeman-Halton test	Stuart-Maxwell test	Multiplicity adjusted Stuart-Maxwell tests#
Ordered categorical	Wilcoxon-Mann-Whitney (WMW) test	Old School***: Kruskal-Wallis analysis of variance (ANOVA) New School***: multiplicity adjusted WMW tests	Wilcoxon sign rank test	Old School# Friedman two-way ANOVA by ranks New School# Mulitiplicity adjusted Wilcoxon sign rank tests
Continuous	independent groups t-test	Old school***: oneway ANOVA New school***: multiplicity adjusted independent groups t tests	paired t-test	mixed effects linear regression
Censored: time to event	log-rank test	Multiplicity adjusted log-rank test	Shared-frailty Cox regression	Shared-frailty Cox regression

CHOICE OF SIGNIFICANCE TESTS (WITH CONFOUNDING)

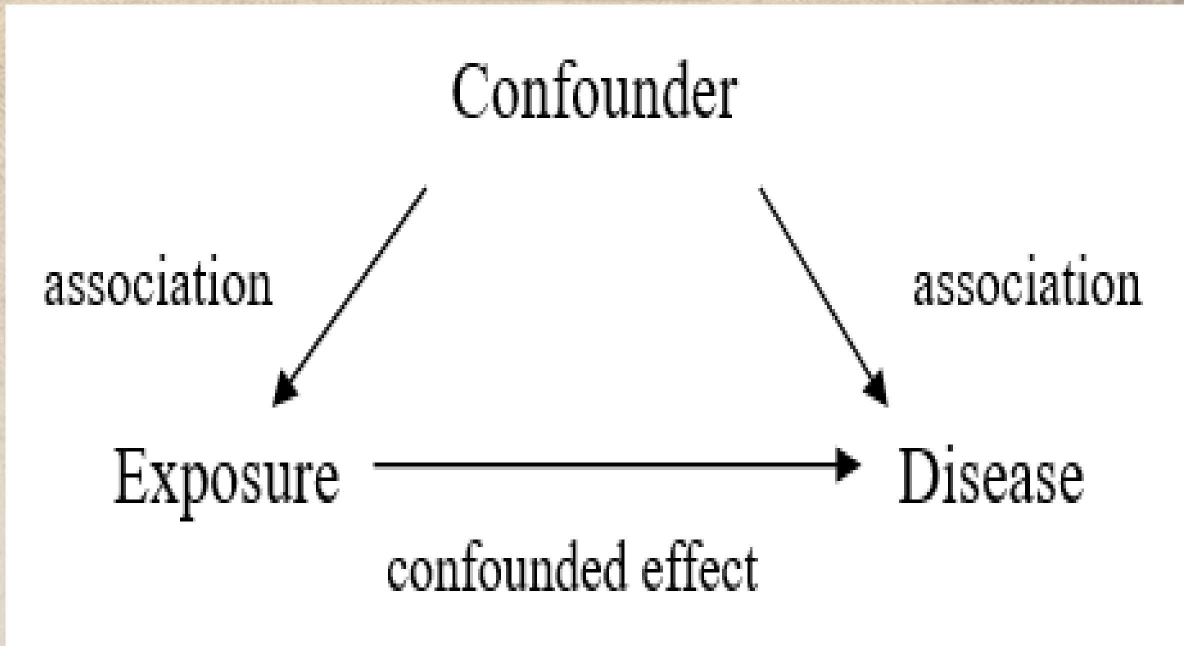
Level of Measurement of Dependent Variable	Two Independent Groups	Three or more Independent Groups	Two Correlated* Samples	Three or more Correlated* Samples
Dichotomous	logistic regression	logistic regression & consider need for multiplicity adjustment	conditional logistic regression, or mixed effects logistic regression	mixed effects logistic regression
Unordered categorical	multinomial logistic regression	multinomial logistic regression & consider need for multiplicity adjustment	General linear mixed model#	General linear mixed model#
Ordered categorical	ordinal logistic regression	ordinal logistic regression & consider need for multiplicity adjustment	mixed effects ordinal logistic regression	mixed effects ordinal logistic regression
Continuous	linear regression	linear regression & consider need for multiplicity adjustment	mixed effects linear regression	mixed effects linear regression
Censored: time to event	Cox regression	Cox regression & consider need for multiplicity adjustment	shared-frailty Cox regression	shared-frailty Cox regression

INTRODUCTION TO REGRESSION

- Regression aims to estimate the effect size of the main association while maintaining the levels of potential confounders constant.
- Common terms encountered in regression:
  - Dependent variable – Outcome variable – Response Variable (example: stroke)
  - Independent variable – Predictor variable – Covariate (example: blood pressure)
  - Coefficient – Parameter Estimate – Effect Size
  - P-Value and 95% Confidence Interval

INTRODUCTION TO REGRESSION

- Comparing groups by simple tests such as t-tests and chi-squared tests can be inefficient because there is likely confounding occurring by one or more variables
  - Confounding is the "mixing or confusion of effects". This induces bias into the main association.



A TALE OF TWO MODELS

```
Logistic regression                                     Number of obs =     76,940
                                                        LR chi2(2)    =   12168.45
                                                        Prob > chi2   =     0.0000
Log likelihood = -44058.651                             Pseudo R2     =     0.1213
```

Good_outcome	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
NIHSS_baseline	.8450461	.0016836	-84.50	0.000	.8417526	.8483524
Endovascular_thrombectomy	1.906714	.0775697	15.86	0.000	1.760583	2.064973
_cons	1.351414	.015456	26.33	0.000	1.321457	1.382049

LINEAR REGRESSION EXAMPLE

- The dataset we'll use explores the relationship between several variables and pulmonary function measured with forced expiratory volume (FEV) in a sample of 654 children, aged 3 – 19
  - Outcome (Independent) Variable – FEV
  - Predictor (Dependent) Variables - height, age, smoking status
- We'll begin with a simple Student T-test (not accounting for confounders) and then building a regression model that accounts for confounding

LINEAR REGRESSION EXAMPLE

- To quickly show how linear regression is related to a t-test and how the parameter estimates of a linear regression can be interpreted, we start with looking at the relationship between FEV and sex in both a t-test and regression
  - The difference of the t-test is equal to the parameter estimate of the linear regression

```
. ttest fev, by(male)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     318     2.45117    .0362111     .645736    2.379925    2.522414
       1 |     336    2.812446    .0547507    1.003598    2.704748    2.920145
---------+--------------------------------------------------------------------
combined |     654     2.63678    .0339047    .8670591    2.570204    2.703355
---------+--------------------------------------------------------------------
    diff |           -.3612766    .0663963               -.491653   -.2309002
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.4412
Ho: diff = 0                                     degrees of freedom =      652

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

```
. regress fev male

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =     29.61
       Model |  21.3239848         1  21.3239848   Prob > F        =    0.0000
    Residual |  469.595849       652  .720239032   R-squared       =    0.0434
-------------+----------------------------------   Adj R-squared   =    0.0420
       Total |  490.919833       653  .751791475   Root MSE        =    .84867

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .3612766   .0663963     5.44   0.000     .2309002     .491653
       _cons |    2.45117    .047591    51.50   0.000      2.35772     2.54462
------------------------------------------------------------------------------
```

MULTIVARIABLE LINEAR REGRESSION EXAMPLE

```
. regress fev smoker

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =     41.79
       Model |   29.569683         1   29.569683   Prob > F        =    0.0000
    Residual |   461.35015       652  .707592255   R-squared       =    0.0602
-------------+----------------------------------   Adj R-squared   =    0.0588
       Total |  490.919833       653  .751791475   Root MSE        =    .84119

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .7107189   .1099426     6.46   0.000     .4948346    .9266033
       _cons |   2.566143   .0346604    74.04   0.000     2.498083    2.634202
------------------------------------------------------------------------------
```

- Still using the FEV dataset, we now show a full model with multiple predictor variables in the equation
  - Predictors variables available in dataset: sex, smoking status, age and height
  - For our mock paper, we want to hypothesize that smoking is associated with lower FEV. We want to build a model that shows that. We start with a simple regression of FEV and smoking status.
- In this model, being a smoker is associated with higher FEV. This is opposite of what we want to see and what intuition tells us. We need to explore if there is confounding present.

MULTIVARIABLE LINEAR REGRESSION EXAMPLE

- Adding age and height to the model, we still get an estimate that is worth exploring further
  - The estimates for the height and age predictor variables are significant. Also have the proper direction as what intuition would say
  - The estimate for smoking status has turned into the right direction but we're still not seeing significance so let's explore further

```
. regress fev smoker height age

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(3, 650)       =    715.69
       Model |  376.837002         3  125.612334   Prob > F        =    0.0000
    Residual |  114.082831       650  .175512048   R-squared       =    0.7676
-------------+----------------------------------   Adj R-squared   =    0.7665
       Total |  490.919833       653  .751791475   Root MSE        =    .41894

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |  -.1102319   .0600175    -1.84   0.067    -.2280834    .0076196
      height |   .1090947   .0047196    23.12   0.000     .0998272    .1183622
         age |    .059741   .0095634     6.25   0.000     .0409621      .07852
       _cons |  -4.616007   .2238833   -20.62   0.000    -5.055629   -4.176385
------------------------------------------------------------------------------
```

MULTIVARIABLE LINEAR REGRESSION EXAMPLE

- Height is directly related to lung size and since smoking occurs more in later teenage years, smoking status acts as a surrogate for lung size
  - Restricting to older children creates a more homogenous lung size in the sample

```
tab age smoker
```

Age (years)	Smoking Status never smo	ever smok	Total
3	2	0	2
4	9	0	9
5	28	0	28
6	37	0	37
7	54	0	54
8	85	0	85
9	93	1	94
10	76	5	81
11	81	9	90
12	50	7	57
13	30	13	43
14	18	7	25
15	9	10	19
16	6	7	13
17	6	2	8
18	4	2	6
19	1	2	3
Total	589	65	654

MULTIVARIABLE LINEAR REGRESSION EXAMPLE

- No smokers under age of 9 so we should restrict the analysis to only those older than 9 years
  - We now see that we have achieved significance in all 3 predictor variables with proper effect directions

```
. regress fev smoker height age if age>9

      Source |       SS           df       MS      Number of obs   =       345
-------------+----------------------------------   F(3, 341)       =    178.10
       Model |  120.042531         3  40.014177   Prob > F        =    0.0000
    Residual |  76.6153415       341 .224678421   R-squared       =    0.6104
-------------+----------------------------------   Adj R-squared   =    0.6070
       Total |  196.657872       344 .571679862   Root MSE        =      .474

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |  -.1810955   .0696684    -2.60   0.010    -.3181293   -.0440616
      height |   .1399885   .0074785    18.72   0.000     .1252786    .1546983
         age |   .0726024   .0141029     5.15   0.000     .0448627    .1003421
       _cons |  -6.758471   .4477109   -15.10   0.000    -7.639094   -5.877848
------------------------------------------------------------------------------
```

MULTIVARIABLE LINEAR REGRESSION EXAMPLE

- To use your regression output and find a specified value of the outcome variable for a specific individual, we use the variable coefficients
  - Example: What is the FEV for a 52" tall 15 year old who smokes?
  - FEV = Constant + [smoking status]*(-0.181) + [height]*(0.140) + [age]*(0.073)
  - FEV = -6.76 + [1]*(-0.181) + [52]*(0.140) + [15]*(0.073) = 1.43

```
. regress fev smoker height age if age>9

      Source |       SS           df       MS      Number of obs   =       345
-------------+----------------------------------   F(3, 341)       =    178.10
       Model |  120.042531         3  40.014177   Prob > F        =    0.0000
    Residual |  76.6153415       341  .224678421   R-squared       =    0.6104
-------------+----------------------------------   Adj R-squared   =    0.6070
       Total |  196.657872       344  .571679862   Root MSE        =      .474

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |  -.1810955   .0696684    -2.60   0.010    -.3181293   -.0440616
      height |   .1399885   .0074785    18.72   0.000     .1252786    .1546983
         age |   .0726024   .0141029     5.15   0.000     .0448627    .1003421
       _cons |  -6.758471   .4477109   -15.10   0.000    -7.639094   -5.877848
------------------------------------------------------------------------------
```