1. Data Management and Visualization

Week 1: Selecting a research question

---

1. Choose a sample
2. Exploratory data analysis
3. Inferential data analysis

Week 2: Writing your first program - SAS or Python

What is a variable distribution?

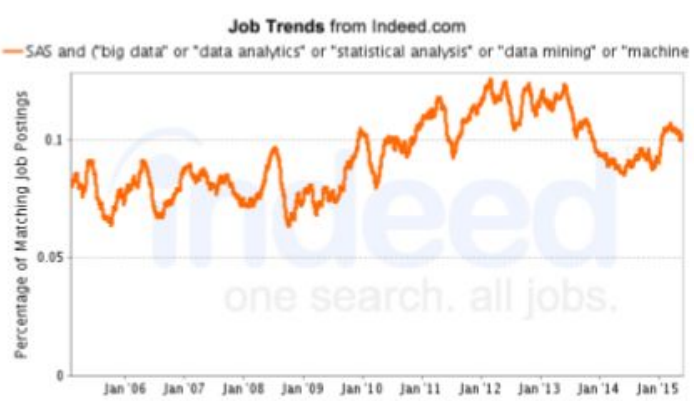
- What values the variable takes
- How often the variable takes those values

Python or SAS?

	SAS	Python
Description:	SAS has traditionally been the market leader in commercial data analysis. The software offers a huge array of statistical and analytic functions.	Python originated as an open source scripting language and though not initially used to conduct data analysis. Pandas and other specialized libraries are beginning to change that.
Cost:	Free for educators and students.	Free and open source
System specifications:	SAS Studio will be used for the majority of the specialization, hosted in the cloud, and requiring no downloads onto your machine. Machine learning algorithms presented in course 4 may require downloads of a free version of SAS Enterprise Miner.	Spyder, an open source cross-platform integrated development environment (IDE) for programming in the Python language, will be used.
Availability	Available worldwide excluding Myanmar, Cuba, Iran, North Korea, Sudan, Syria, and China	Available worldwide.

<table>
<tr><th></th><th>SAS</th><th>Python</th></tr>
<tr><td>Advancements:</td><td>SAS releases well tested updates in a controlled environment.</td><td>Python has open contributions and there are chances of errors in latest developments.</td></tr>
<tr><td>Ease of learning:</td><td>SAS is easy to learn. In addition to resources made available through this specialization, there are supporting websites of various universities and SAS has comprehensive documentation.</td><td>Python is known for its simplicity in the programming world. This remains true for data analysis as well. Documentation is improving.</td></tr>
<tr><td>Example Code:</td><td>proc freq; tables TAB12MDX;</td><td>c1 = data["TAB12MDX"].value_counts
(sort=False)
print (c1)

p1 = data["TAB12MDX"].value_counts
(sort=False, normalize=True)
print (p1)</td></tr>
<tr><td>Example Output:</td><td><table><tr><th>TAB12MDX</th><th>Frequency</th><th>Percent</th><th>Cumulative Frequency</th><th>Cumulative Percent</th></tr><tr><td>0</td><td>38131</td><td>88.49</td><td>38131</td><td>88.49</td></tr><tr><td>1</td><td>4962</td><td>11.51</td><td>43093</td><td>100.00</td></tr></table></td><td><pre>counts for TAB12MDX
0   38131
1   4962
dtype: int64
percentages for TAB12MDX
0   0.884854
1   0.115146
dtype: float64</pre></td></tr>
<tr><td>Employment Scenario</td><td>Globally, SAS is still the market leader in available corporate jobs.</td><td>Python is often considered the better option for start-ups and companies looking for cost efficiency.</td></tr>
</table>

	SAS	Python
Job Trends from Indeed.com		
Strengths	SAS has a strong brand and is a first class statistical modelling program.	The fact that Python is a general purpose programming language means that knowledge of Python can be useful for all types of programming work
Some limitations	Full license can be expensive outside of the educational environment.	Since the core language is small and excludes many standard scientific operations, such duties fall on third party libraries such as Pandas More work is still needed to make Python a first class statistical modeling environment, but it is on its way.

Symbol	Mnemonic Equivalent	Definition	Example
=	EQ	equal to	A=3
^=	NE	not equal to	a ne 3
¬=	NE	not equal to	
~=	NE	not equal to	
>	GT	greater than	Num>5
<	LT	less than	Num<8
>=	GE	greater than or equal to	Sales>=300
<=	LE	less than or equal to	Sales<=100

SAS Program for video examples (Module 2).sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
       CHECK321="Smoked Cigarettes in Past 12 Months"
       S3AQ3B1="Usual Smoking Frequency"
       S3AQ3C1="Usual Smoking Quantity";
/*subsetting the data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by IDNUM;
PROC FREQ; TABLES TAB12MDX CHECK321 S3AQ3B1 S3AQ3C1 AGE;
RUN;
```

---
week 2.sas

```
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data new; set mydata.gapminder;
/* I choose to put the full description here since not everyone is using the same data set */
label polityscore="2009 Democracy score (Polity)
Overall polity score from the Polity IV dataset, calculated by subtracting an autocracy score from a democracy
score.
The summary measure of a country's democratic and free nature.
-10 is the lowest value, 10 the highest.";
proc sort; by country; /* country is the ID in my data set */
proc freq; tables polityscore; /* show the frequency distribution for this variable */
run;
/*---------------------------------------------------------------------------------------------*/
/*
* Note: the following code is used to view descriptive statistics on some variables.
* This is instead of the method described in the course videos since my variables are continuous
* rather than descriptive, so a "proc freq" will not say much about my data
*/
/*---------------------------------------------------------------------------------------------*/
ods select BasicMeasures Quantiles;
proc univariate data=mydata.gapminder;
  var incomeperperson armedforcesrate;
run;
```

Week 3: Managing Data

Note: data management syntax goes after the DATA statement and before the PROC statements.

SAS programs week 3.sas

```
/*Program for NESARC data set*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new1; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
        CHECK321="Smoked Cigarettes in Past 12 Months"
        S3AQ3B1="Usual Smoking Frequency"
        S3AQ3C1="Usual Smoking Quantity";
IF AGE LE 20 THEN AGEGROUP=1; /*18, 19, 20 year olds*/
ELSE IF AGE LE 22 THEN AGEGROUP=2; /*21, 22 year olds*/
ELSE AGEGROUP=3; /*23, 24, 25 year olds*/
/*set missing data*/
IF S3AQ3B1=9 THEN S3AQ3B1=.;
/*Reverse code values*/
IF S3AQ3B1=1 THEN USFREQ=6;
ELSE IF S3AQ3B1=2 THEN USFREQ=5;
ELSE IF S3AQ3B1=3 THEN USFREQ=4;
ELSE IF S3AQ3B1=4 THEN USFREQ=3;
ELSE IF S3AQ3B1=5 THEN USFREQ=2;
ELSE IF S3AQ3B1=6 THEN USFREQ=1;
/*recode with more meaningful quantitative values*/
IF S3AQ3B1=1 THEN USFREQMO=30;
ELSE IF S3AQ3B1=2 THEN USFREQMO=22;
ELSE IF S3AQ3B1=3 THEN USFREQMO=14;
ELSE IF S3AQ3B1=4 THEN USFREQMO=5;
ELSE IF S3AQ3B1=5 THEN USFREQMO=2.5;
ELSE IF S3AQ3B1=6 THEN USFREQMO=1;
IF S3AQ3C1=99 THEN S3AQ3C1=.;
/*coding in valid data*/
```

```
IF S2AQ3 NE 9 AND S2AQ8A=. THEN S2AQ8A=11;
/*secondary variables estimating the number of cigaretted smoked in the past 30 days*/
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
/*subsetting the data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by IDNUM;
PROC PRINT; VAR USFREQMO S3AQ3C1 NUMCIGMO_EST;
PROC FREQ; TABLES TAB12MDX CHECK321 S3AQ3B1 S3AQ3C1 AGE USFREQMO NUMCIGMO_EST;
RUN;
/*Program for AddHealth data set*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new2; set mydata.addhealth_pds;
IF H1GI4 GE 6 then H1GI4=.;
IF H1GI6A GE 6 then H1GI6A=.;
IF H1GI6B GE 6 then H1GI6B=.;
IF H1GI6C GE 6 then H1GI6C=.;
IF H1GI6D GE 6 then H1GI6D=.;
NUMETHNIC=SUM(of H1GI4 H1GI6A H1GI6B H1GI6C H1GI6D);
IF NUMETHNIC GE 2 THEN ETHNICITY=1; /*multiple race/ethnic endorsed*/
ELSE IF H1GI4=1 THEN ETHNICITY=2; /*Hispanic or Latino*/
ELSE IF H1GI6A=1 THEN ETHNICITY=3; /*Black or African American*/
ELSE IF H1GI6B=1 THEN ETHNICITY=4; /*American Indian or Native American*/
ELSE IF H1GI6C=1 THEN ETHNICITY=5; /*Asian or Pacific Islander*/
ELSE IF H1GI6D=1 THEN ETHNICITY=6; /*White*/
PROC SORT; by AID;
PROC FREQ; TABLES H1GI4 H1GI6A H1GI6B H1GI6C H1GI6D NUMETHNIC ETHNICITY;
RUN;
```

---
week 3.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp
      from       work.gapminder as gapminder
                       left join work.popden as popden on gapminder.country = popden.country
                       left join work.pop as pop on gapminder.country = pop.country
                       left join work.surarea as surarea on gapminder.country = surarea.country
                       left join work.cpi as cpi on gapminder.country = cpi.country
                       left join work.hdi as hdi on gapminder.country = hdi.country
                       left join work.murder as murder on gapminder.country = murder.country
                       left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
```

```
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
      LIFEEXPECTANCY_RANK
```

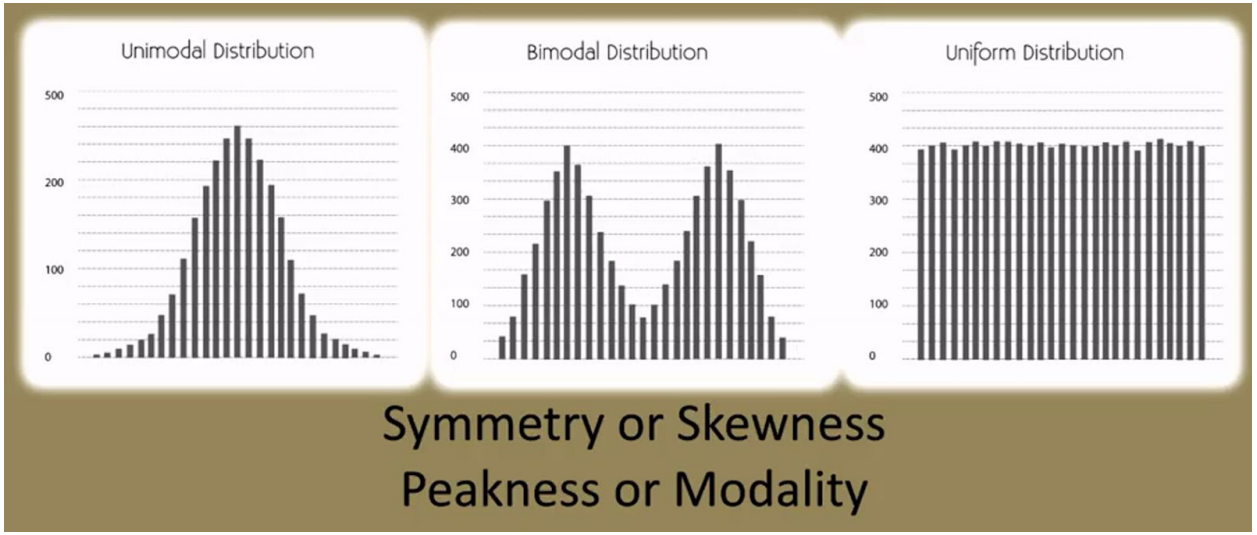
```
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = -1;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
```

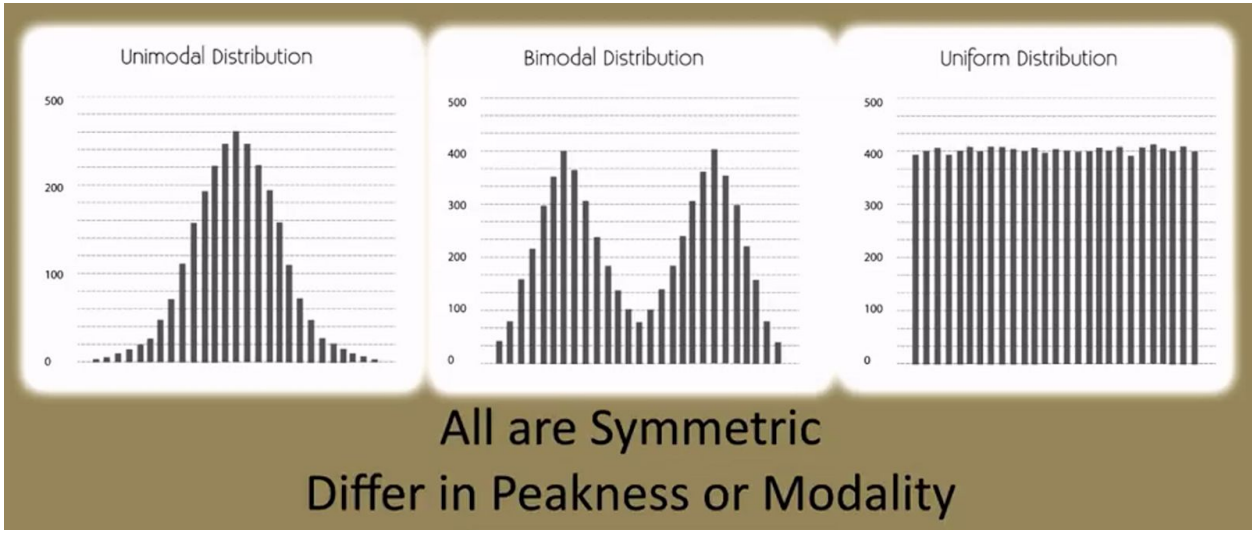
```
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = -1;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = -1;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = -1;
IF SUICIDEPER100TH <  4.983422  THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH >= 4.983422  AND  SUICIDEPER100TH <    8.262893   THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 8.262893  AND  SUICIDEPER100TH <    12.367980  THEN SUICIDEPER100TH_RANK = 3;
IF SUICIDEPER100TH >= 12.367980 THEN SUICIDEPER100TH_RANK = 4;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = -1;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = -1;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = -1;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = -1;
```

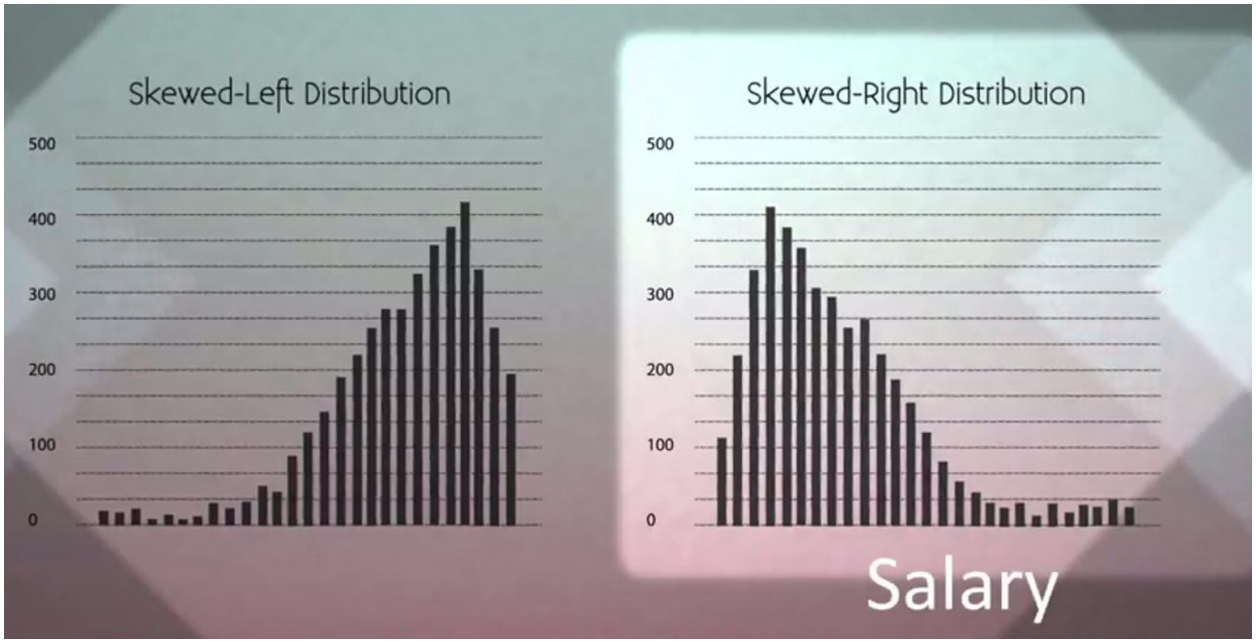
```
IF popden <  31.774  THEN popden_RANK = 1;
IF popden >= 31.774  AND  popden < 77.990  THEN popden_RANK = 2;
IF popden >= 77.990  AND  popden < 196.229 THEN popden_RANK = 3;
IF popden >= 196.229 THEN popden_RANK = 4;
IF popden = . THEN popden_RANK = -1;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = -1;
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = -1;

IF homicide <  1.715654  THEN homicide_RANK = 1;
IF homicide >= 1.715654  AND homicide < 6.111090  THEN homicide_RANK = 2;
IF homicide >= 6.111090  AND homicide < 19.004826 THEN homicide_RANK = 3;
IF homicide >= 19.004826 THEN homicide_RANK = 4;
IF homicide = . THEN homicide_RANK = -1;
IF milexpprcntgdp <  1.0752166 THEN milexpprcntgdp_RANK = 1;
IF milexpprcntgdp >= 1.0752166 AND  milexpprcntgdp < 1.4946096 THEN milexpprcntgdp_RANK = 2;
IF milexpprcntgdp >= 1.4946096 AND  milexpprcntgdp < 2.4681756 THEN milexpprcntgdp_RANK = 3;
IF milexpprcntgdp >= 2.4681756 THEN milexpprcntgdp_RANK = 4;
IF milexpprcntgdp = . THEN milexpprcntgdp_RANK = -1;

/* FREQUENCY PER VARIABLE */
PROC FREQ;
TABLES
      EMPLOYRATE_RANK
```

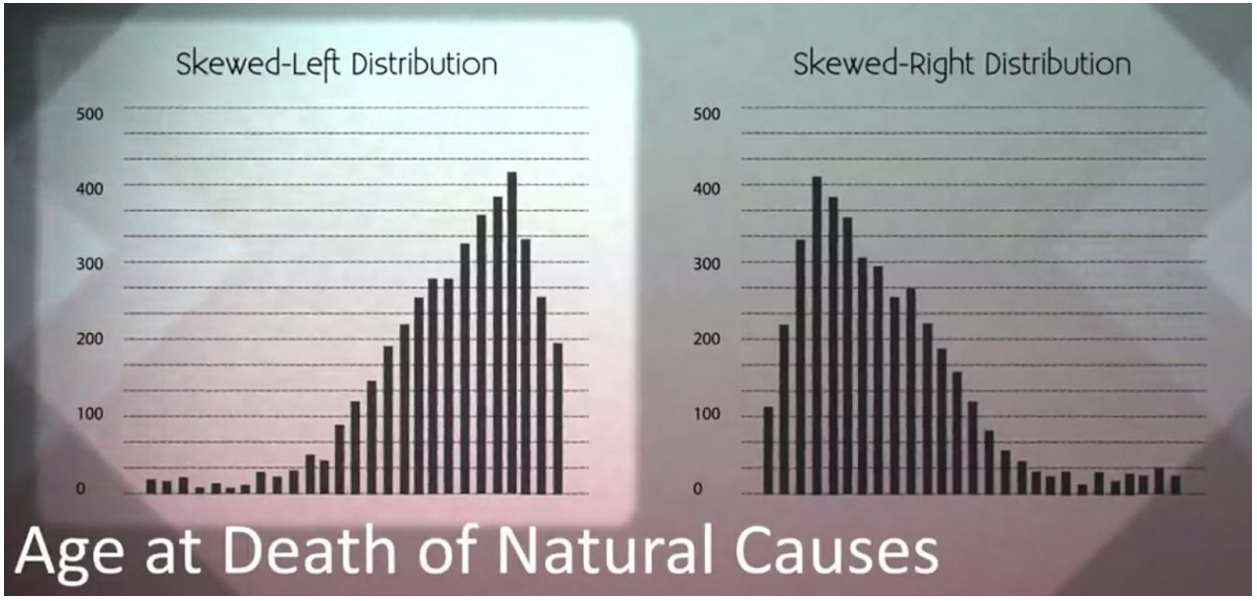
```
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH_RANK
      URBANRATE_RANK
      surarea_RANK
      population_RANK
      popden_RANK
      corruptionindex_RANK
      hdi_RANK
      homicide_RANK
      milexpprcntgdp_RANK;
RUN;
```

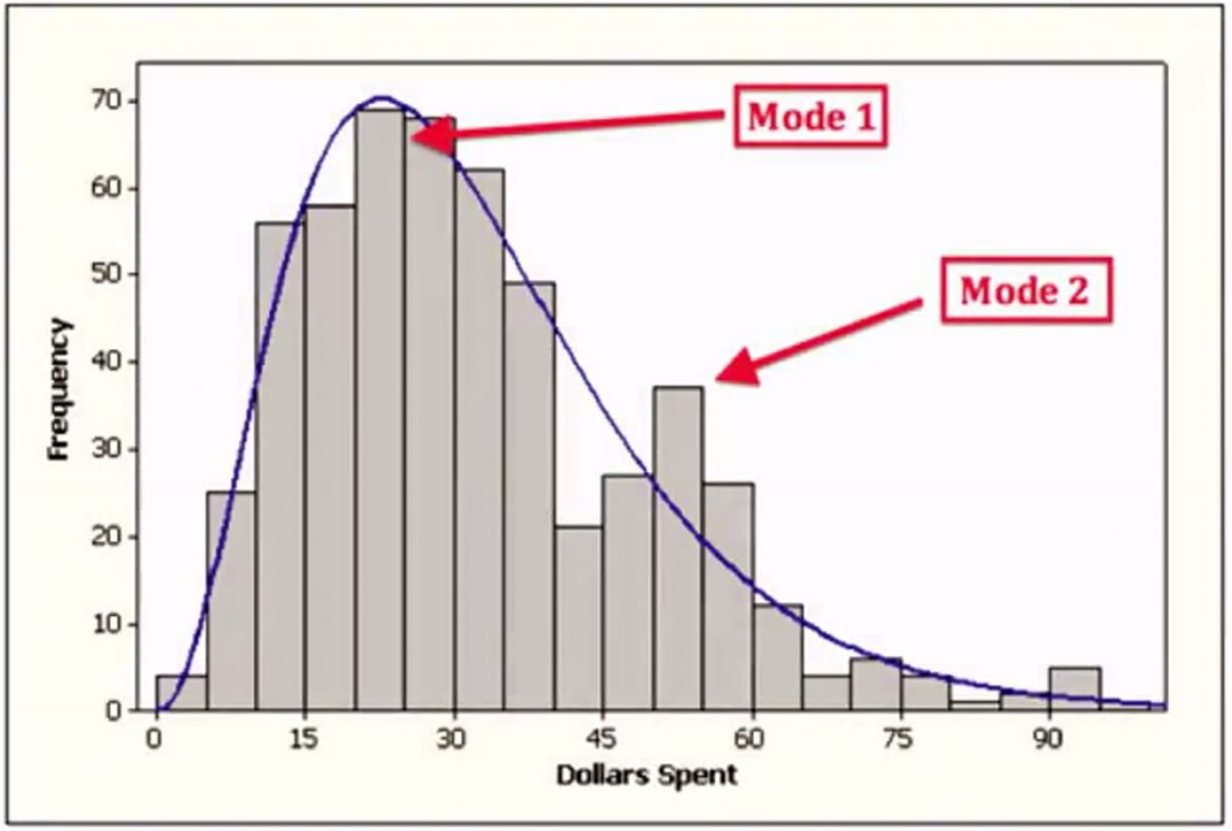
Week 4: Visualizing Data

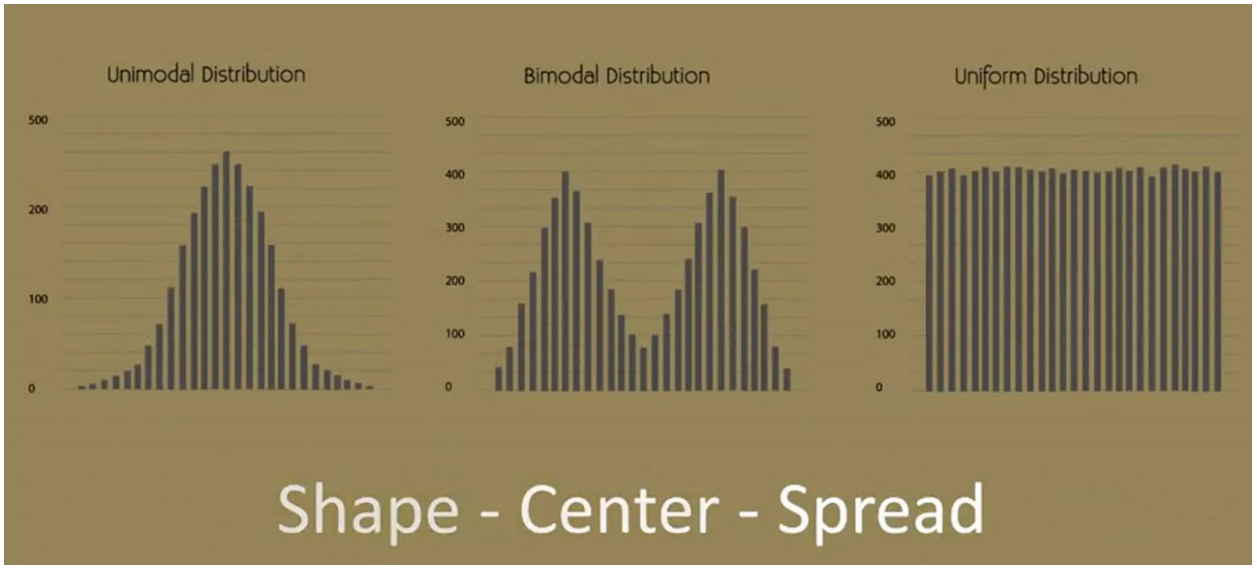




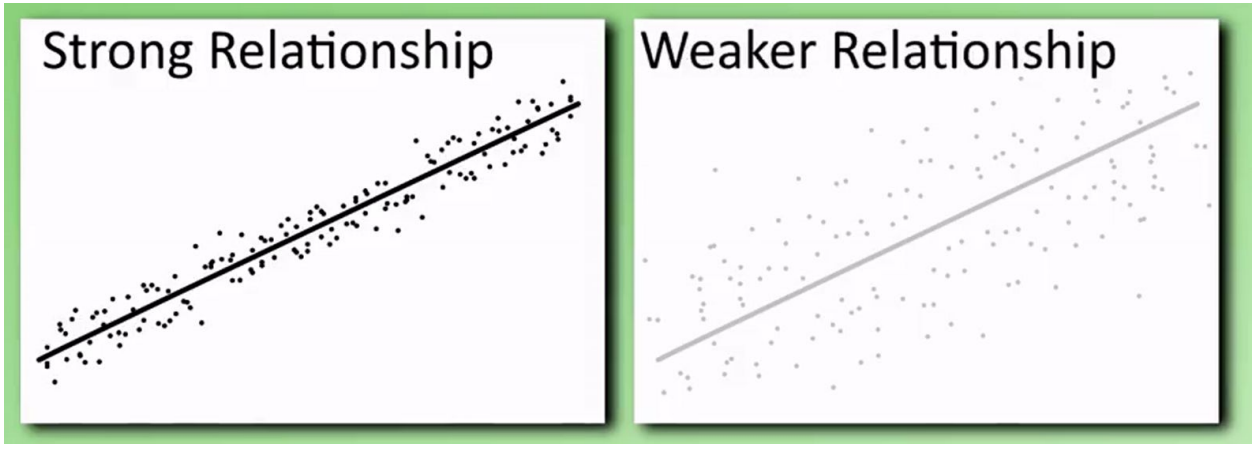

Skewed-Left Distribution
500
400
300
200
100
0
Skewed-Right Distribution
500
400
300
200
100
0
Salary

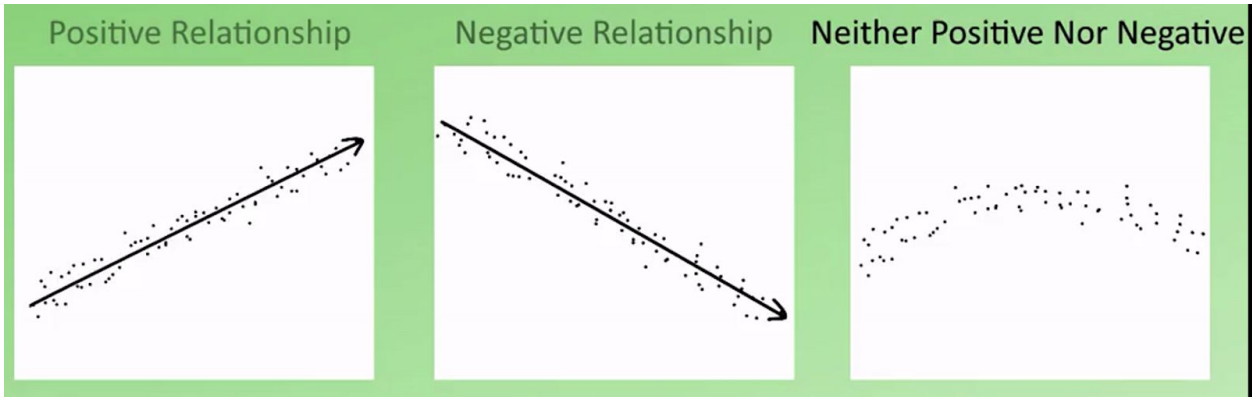

Skewed-Left Distribution
500
400
300
200
100
0
Skewed-Right Distribution
500
400
300
200
100
0
Age at Death of Natural Causes

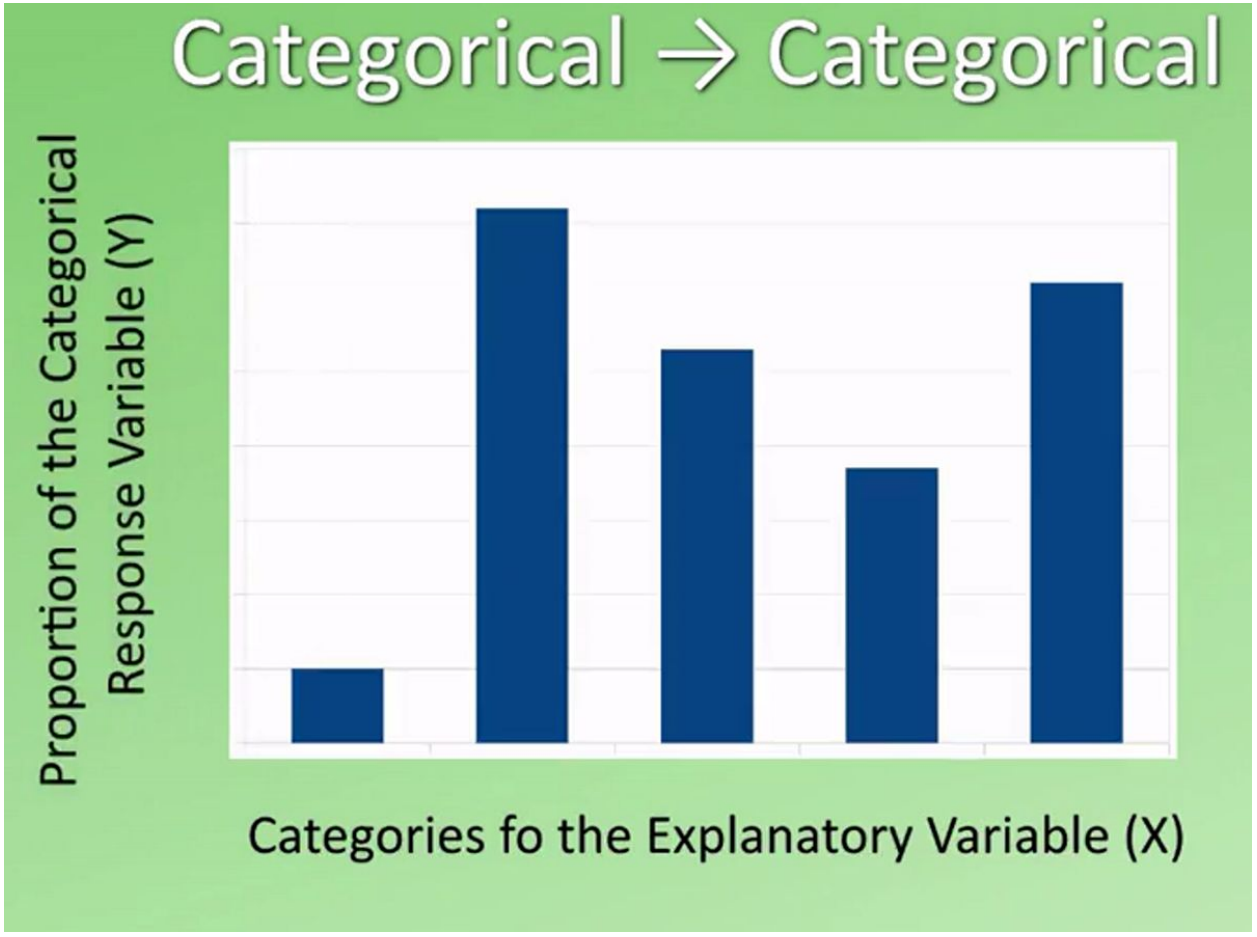

Mode 1
Mode 2
Frequency
Dollars Spent
0
10
20
30
40
50
60
70
0
15
30
45
60
75
90

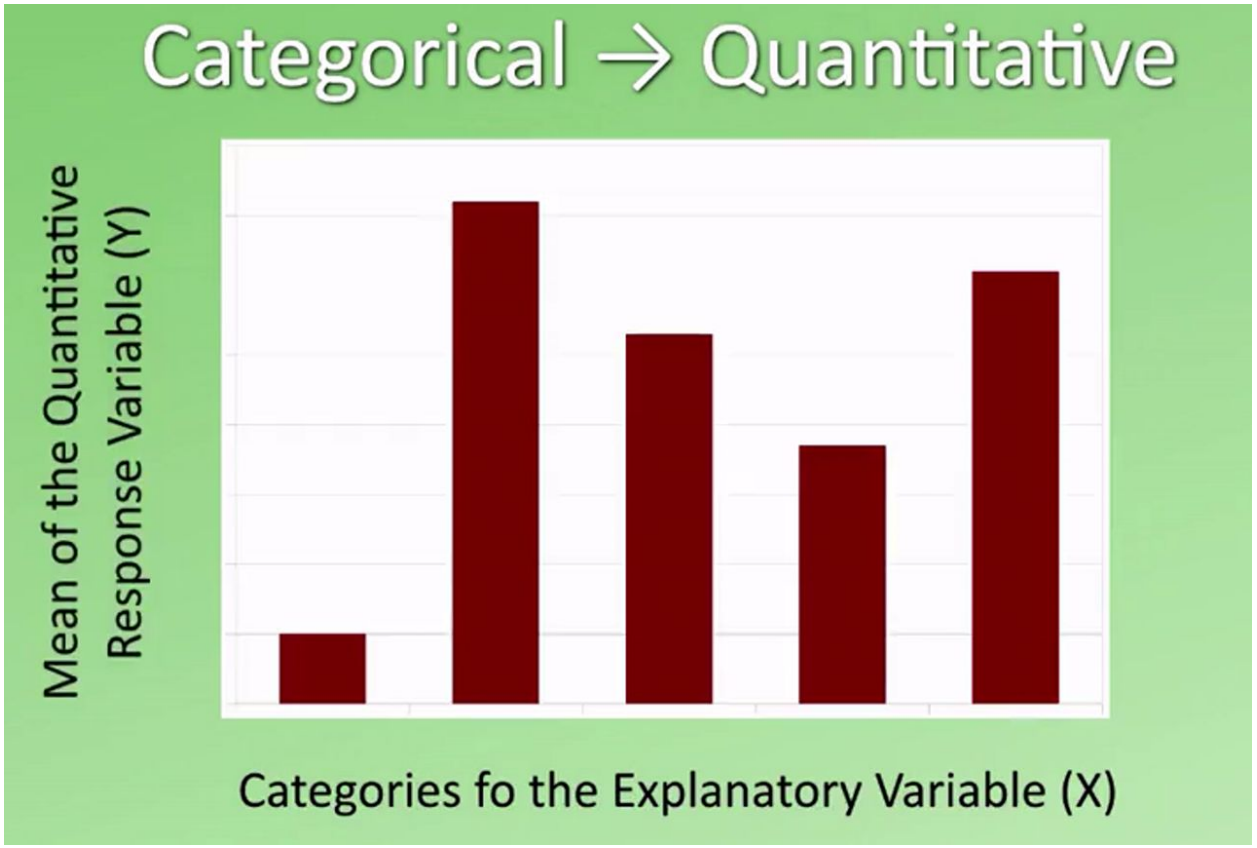

Unimodal Distribution
Bimodal Distribution
Uniform Distribution
500
400
300
200
100
0
Shape - Center - Spread

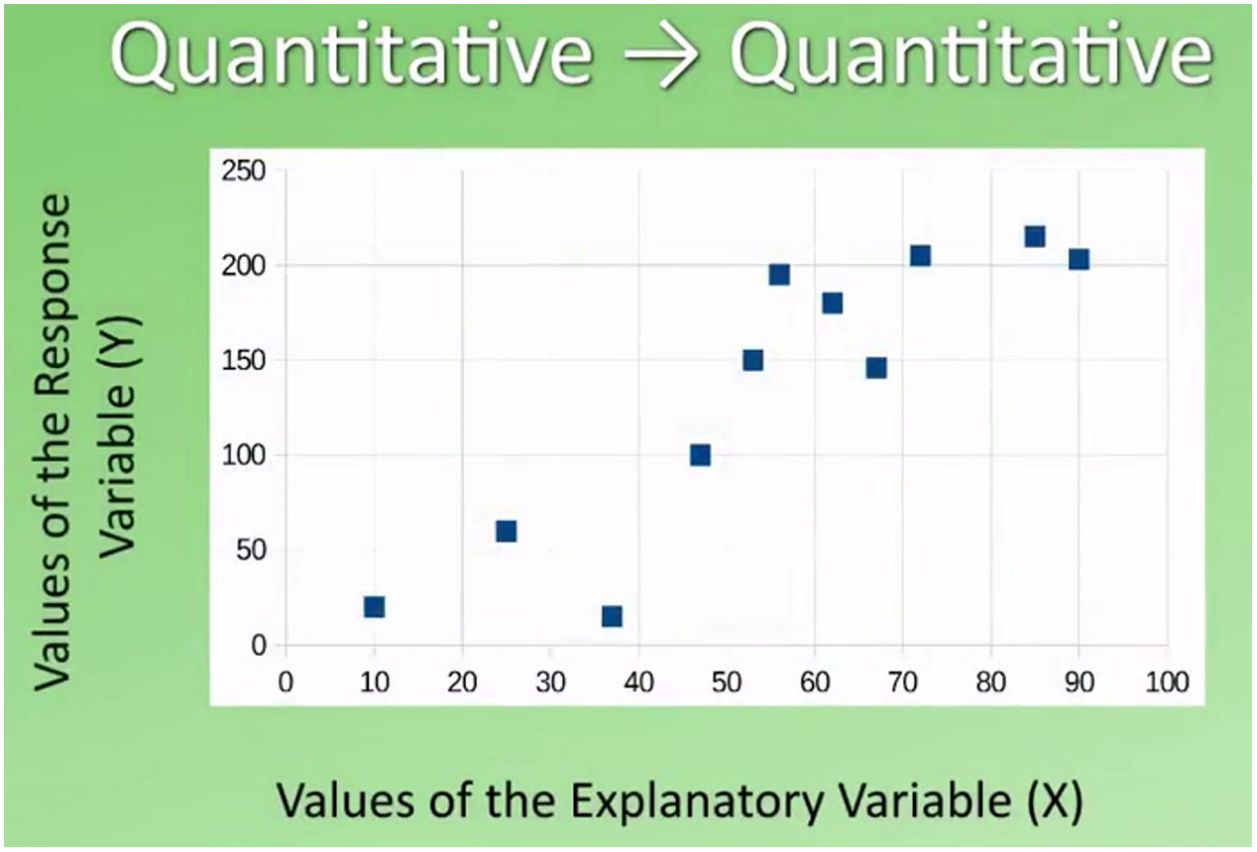
A categorical response variable should not have more than two levels coded as 0 and 1.




Positive Relationship
Negative Relationship
Neither Positive Nor Negative

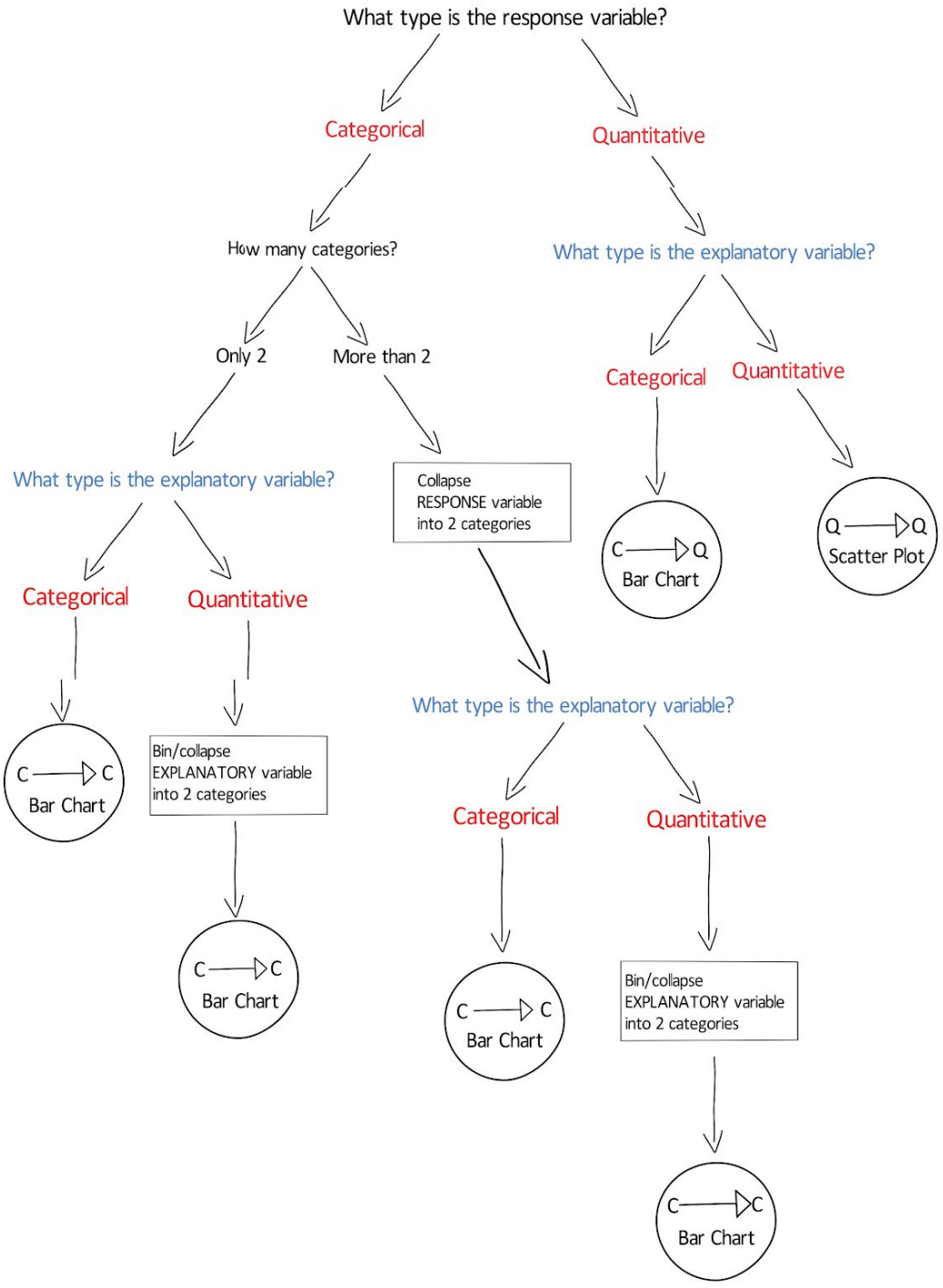

Categorical → Categorical
Proportion of the Categorical Response Variable (Y)
Categories fo the Explanatory Variable (X)


Categorical → Quantitative
Mean of the Quantitative Response Variable (Y)
Categories fo the Explanatory Variable (X)



Very Important!

⇒ Translation Syntax SPSS STATA SAS R 082115.pdf


What type is the response variable?
Categorical
Quantitative
How many categories?
What type is the explanatory variable?
Only 2
More than 2
Categorical
Quantitative
What type is the explanatory variable?
Collapse RESPONSE variable into 2 categories
C → Q Bar Chart
Q → Q Scatter Plot
Categorical
Quantitative
C → C Bar Chart
Bin/collapse EXPLANATORY variable into 2 categories
C → C Bar Chart
What type is the explanatory variable?
Categorical
Quantitative
C → C Bar Chart
Bin/collapse EXPLANATORY variable into 2 categories
C → C Bar Chart

SAS Program for video examples (Module 4).sas

```
/*program for NESARC data set*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
        CHECK321="Smoked Cigarettes in Past 12 Months"
        S3AQ3B1="Usual Smoking Frequency"
        S3AQ3C1="Usual Smoking Quantity";
IF AGE LE 20 THEN AGEGROUP=1; /*18, 19, 20 year olds*/
ELSE IF AGE LE 22 THEN AGEGROUP=2; /*21, 22 year olds*/
ELSE AGEGROUP=3; /*23, 24, 25 year olds*/
IF S3AQ3B1=9 THEN S3AQ3B1=.;
IF S3AQ3C1=99 THEN S3AQ3C1=.;
IF S3AQ3B1=1 THEN USFREQ=6;
ELSE IF S3AQ3B1=2 THEN USFREQ=5;
ELSE IF S3AQ3B1=3 THEN USFREQ=4;
ELSE IF S3AQ3B1=4 THEN USFREQ=3;
ELSE IF S3AQ3B1=5 THEN USFREQ=2;
ELSE IF S3AQ3B1=6 THEN USFREQ=1;
/*values for new variable USFREQ
1=once a month or less
2=2-3 days a month
3=1-2 days a week
4=3-4 days a week
5=5-6 days a week
6=every day*/
IF S3AQ3B1=1 THEN USFREQMO=30;
ELSE IF S3AQ3B1=2 THEN USFREQMO=22;
ELSE IF S3AQ3B1=3 THEN USFREQMO=14;
```

```
ELSE IF S3AQ3B1=4 THEN USFREQMO=5;
ELSE IF S3AQ3B1=5 THEN USFREQMO=2.5;
ELSE IF S3AQ3B1=6 THEN USFREQMO=1;
/*USFREQMO usual smoking days per month
1=once a month or less
2.5=2-3 days per month
5=1-2 days per week
14=3-4 days per week
22=5-6 days per week
30=everyday*/
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
PACKSPERMONTH=NUMCIGMO_EST/20;
IF PACKSPERMONTH LE 5 THEN PACKCATEGORY=3;
ELSE IF PACKSPERMONTH LE 10 THEN PACKCATEGORY=7;
ELSE IF PACKSPERMONTH LE 20 THEN PACKCATEGORY=15;
ELSE IF PACKSPERMONTH LE 30 THEN PACKCATEGORY=25;
ELSE IF PACKSPERMONTH GT 30 THEN PACKCATEGORY=58;
IF TAB12MDX=1 THEN SMOKEGRP=1; /*nicotine dependent*/
ELSE IF S3AQ3B1=1 THEN SMOKEGRP=2; /*daily smoker*/
ELSE SMOKEGRP=3; /*non daily smoker*/
IF S3AQ3B1=1 THEN DAILY=1;
ELSE IF S3AQ3B1 NE . THEN DAILY=0;
/*subsetting the data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by IDNUM;
PROC PRINT; VAR USFREQMO S3AQ3C1 NUMCIGMO_EST;
PROC FREQ; TABLES TAB12MDX CHECK321 S3AQ3B1 S3AQ3C1 AGE USFREQMO NUMCIGMO_EST;
PROC GCHART; VBAR TAB12MDX/Discrete type=PCT width=30; /*Categorical variable example*/
PROC GCHART; VBAR NUMCIGMO_EST/ type=PCT; /*Quantitatitve variable example*/
```

```
PROC UNIVARIATE; VAR NUMCIGMO_EST; /*Univariate PROC only appropriate to use for quantitative variables*/
PROC FREQ; TABLES PACKSPERMONTH;
PROC GCHART; VBAR TAB12MDX/Discrete type=PCT width=30; /*Categorical variable example*/
PROC GCHART; VBAR NUMCIGMO_EST/ type=PCT; /*Quantitatitve variable example*/
PROC UNIVARIATE; VAR NUMCIGMO_EST; /*Only appropriate to use for quantitative variables*/
PROC GCHART; VBAR PACKCATEGORY/discrete TYPE=mean SUMVAR=TAB12MDX;
PROC GCHART; VBAR ETHRACE2A/discrete type=mean SUMVAR=DAILY;
RUN;
/*Program for Gapminder data set*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new2; set mydata.gapminder;
IF incomeperperson eq . THEN incomegroup=.;
ELSE IF incomeperperson LE 744.239 THEN incomegroup=1;
ELSE IF incomeperperson LE 2553.496 THEN incomegroup=2;
ELSE IF incomeperperson LE 9425.236 THEN incomegroup=3;
ELSE IF incomeperperson GT 9425.236 THEN incomegroup=3;
PROC SORT; by COUNTRY;
PROC FREQ; TABLES incomegroup;
PROC UNIVARIATE; VAR urbanrate internetuserate;
PROC GPLOT; PLOT internetuserate*urbanrate;
PROC UNIVARIATE; VAR incomeperperson HIVrate;
PROC GPLOT; PLOT HIVrate*incomeperperson;
PROC GCHART; VBAR incomegroup/discrete type=mean SUMVAR=HIVrate;
RUN;
```

---
week 4.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
```

```
          OUT=murder REPLACE;

PROC IMPORT
          DATAFILE='/home/qallaf890/military expenditure.csv'
          OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
     create table mygapminder AS
     select          gapminder.*
                     ,surarea.surarea
                     ,pop.population
                     ,popden.popden
                     ,cpi.corruptionindex
                     ,hdi.hdi
                     ,murder.homicide
                     ,military.milexpprcntgdp
     from       work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
     set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
     COUNTRY='COUNTRY'
     INCOMEPERPERSON='GDP PER CAPITA'
```

```
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
```

```
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = -1;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
```

```
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = -1;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = -1;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = -1;
IF SUICIDEPER100TH <  4.983422  THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH >= 4.983422  AND  SUICIDEPER100TH <    8.262893   THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 8.262893  AND  SUICIDEPER100TH <    12.367980  THEN SUICIDEPER100TH_RANK = 3;
IF SUICIDEPER100TH >= 12.367980 THEN SUICIDEPER100TH_RANK = 4;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = -1;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = -1;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = -1;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
```

```
IF population = . THEN population_RANK = -1;
IF popden <  31.774  THEN popden_RANK = 1;
IF popden >= 31.774  AND  popden < 77.990  THEN popden_RANK = 2;
IF popden >= 77.990  AND  popden < 196.229 THEN popden_RANK = 3;
IF popden >= 196.229 THEN popden_RANK = 4;
IF popden = . THEN popden_RANK = -1;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = -1;
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = -1;

IF homicide <  1.715654  THEN homicide_RANK = 1;
IF homicide >= 1.715654  AND homicide < 6.111090  THEN homicide_RANK = 2;
IF homicide >= 6.111090  AND homicide < 19.004826 THEN homicide_RANK = 3;
IF homicide >= 19.004826 THEN homicide_RANK = 4;
IF homicide = . THEN homicide_RANK = -1;
IF milexpprcntgdp <  1.0752166 THEN milexpprcntgdp_RANK = 1;
IF milexpprcntgdp >= 1.0752166 AND  milexpprcntgdp < 1.4946096 THEN milexpprcntgdp_RANK = 2;
IF milexpprcntgdp >= 1.4946096 AND  milexpprcntgdp < 2.4681756 THEN milexpprcntgdp_RANK = 3;
IF milexpprcntgdp >= 2.4681756 THEN milexpprcntgdp_RANK = 4;
IF milexpprcntgdp = . THEN milexpprcntgdp_RANK = -1;

/* FREQUENCY PER VARIABLE */
/*
PROC FREQ;
```

```
TABLES
      EMPLOYRATE_RANK
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH_RANK
      URBANRATE_RANK
      surarea_RANK
      population_RANK
      popden_RANK
      corruptionindex_RANK
      hdi_RANK
      homicide_RANK
      milexpprcntgdp_RANK;
RUN;
*/
/* UNIVARIATE GRAPH */
PROC GCHART; VBAR homicide_RANK /DISCRETE TYPE=PERCENT WIDTH=30;
RUN;
PROC GCHART; VBAR corruptionindex_RANK /DISCRETE TYPE=PERCENT WIDTH=30;
RUN;
PROC GCHART; VBAR hdi_RANK /DISCRETE TYPE=PERCENT WIDTH=30;
RUN;
/* BIVARIATE GRAPH */
PROC GPLOT; PLOT homicide*POPULATION;
RUN;
PROC GPLOT; PLOT corruptionindex*POPULATION;
RUN;
PROC GPLOT; PLOT hdi*POPULATION;
RUN;
PROC GPLOT; PLOT INCOMEPERPERSON*POPULATION;
```

```
RUN;
```

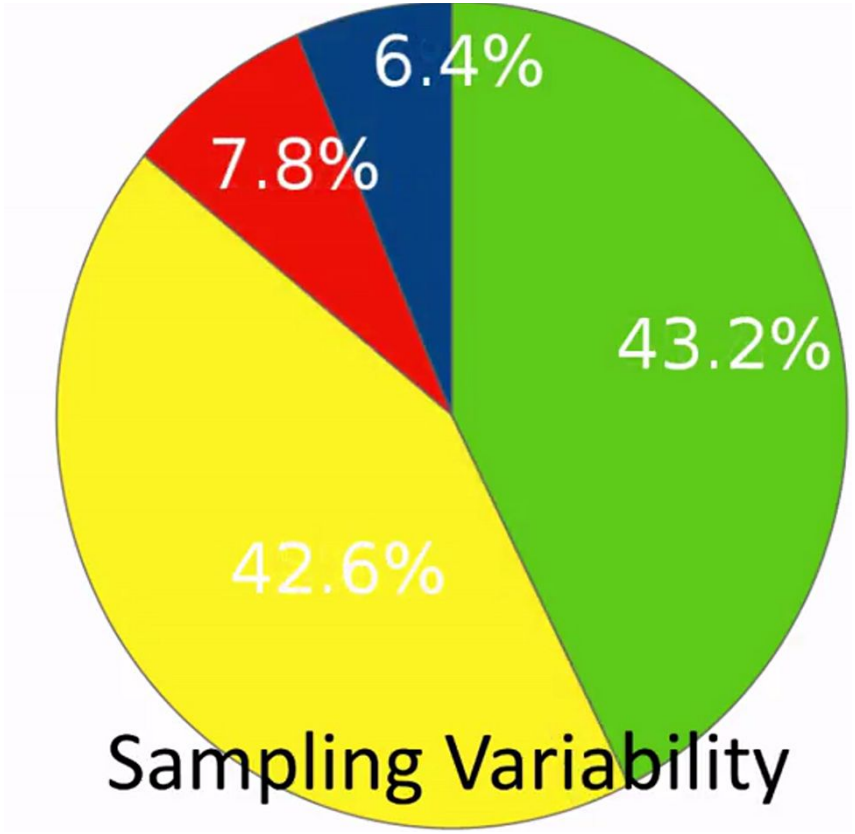
2. Data Analysis Tools

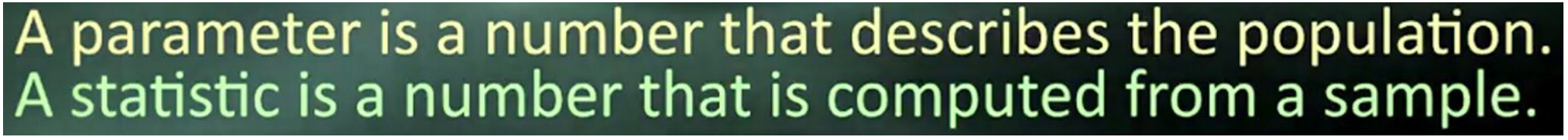
Week 1: Hypothesis Testing and ANOVA

Exploratory data analysis:

1. Examination of frequency distribution
2. Graphical representations
3. Calculations of center and spread

Statistical hypothesis testing: assessing the evidence provided by the data in favor or against each hypothesis about the population.





Parameters are typically unknown.

SAMPLING VARIABILITY:

The statistics of different samples of a population will differ somewhat.

Central Limit Theorem:

As long as adequately large samples AND an adequately large number of samples are used from a population, the distribution of the statistics of the samples will be normally distributed.

Hypothesis Testing

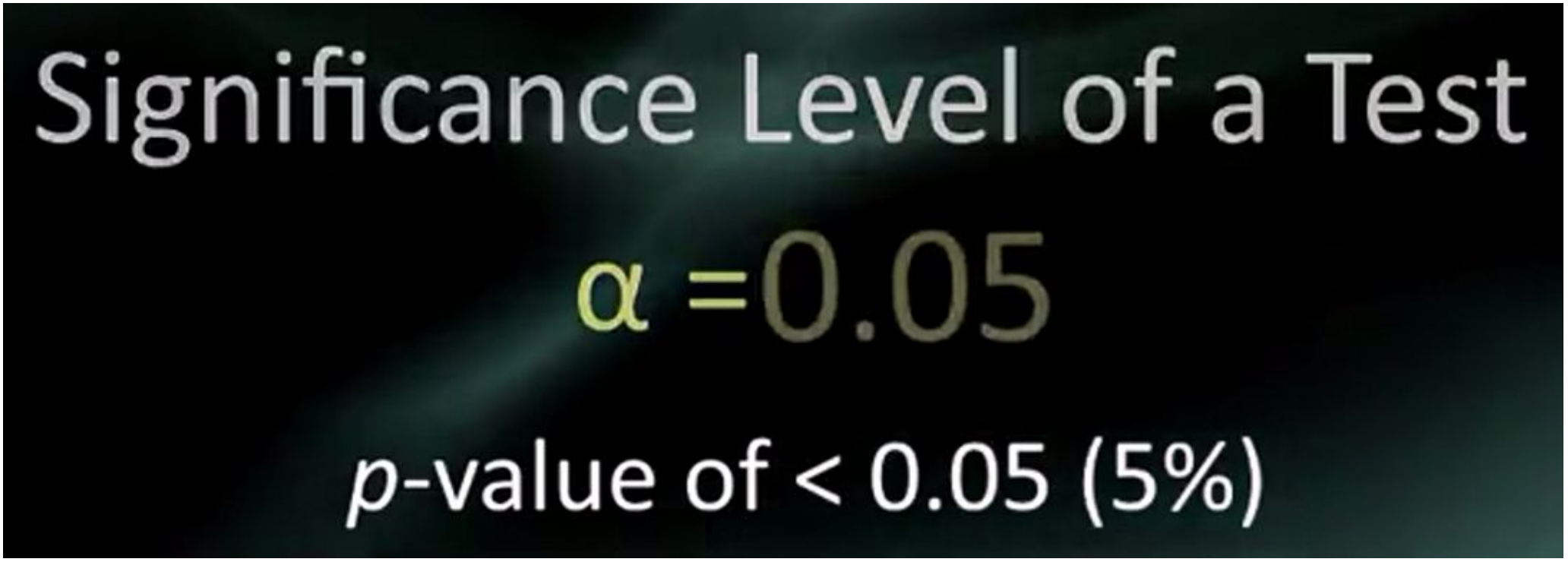
1. Specify the null (H_o)
   and the alternate (H_a) hypothesis
2. Choose a sample
3. Assess the evidence
4. Draw conclusions

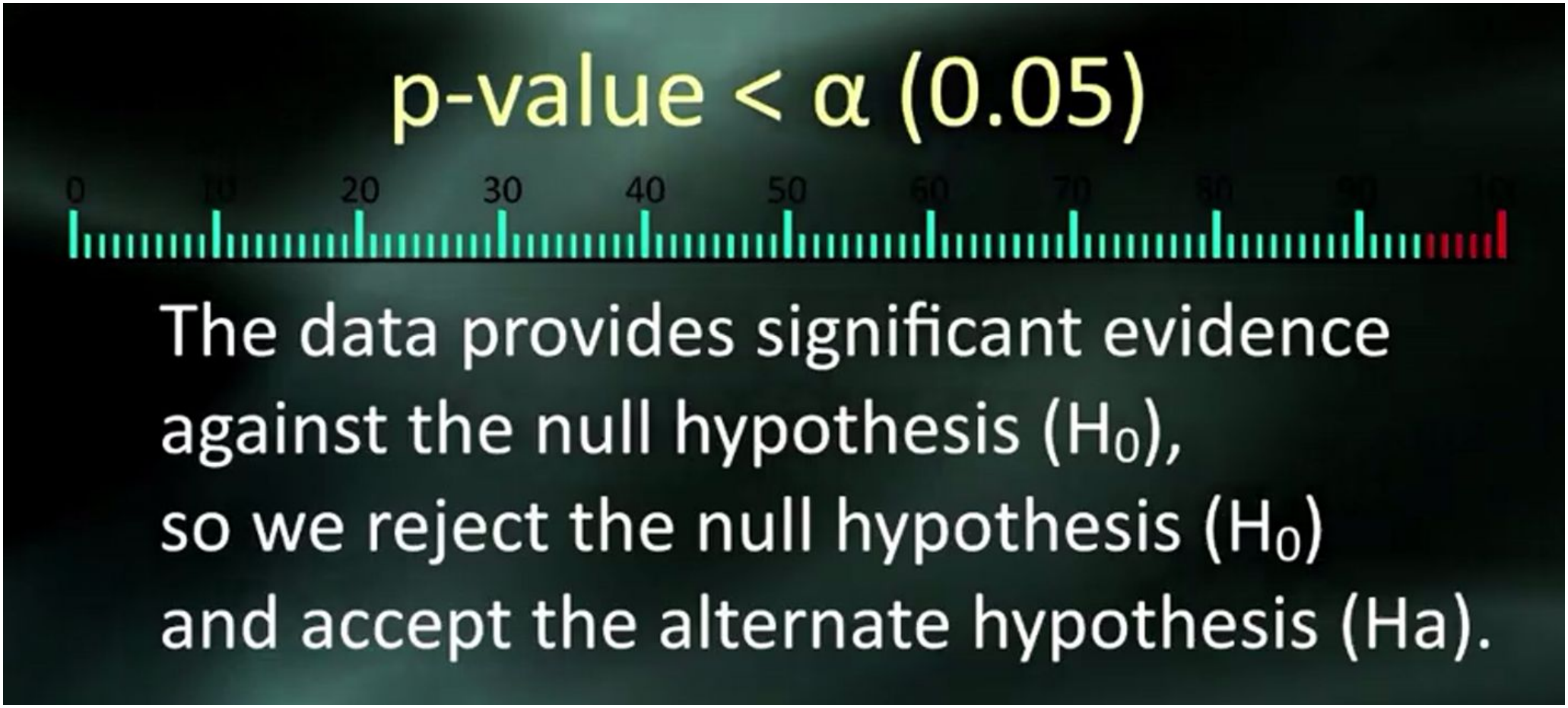
Definition: Assessing evidence provided by the data,
in favor of or against each hypothesis about the population.

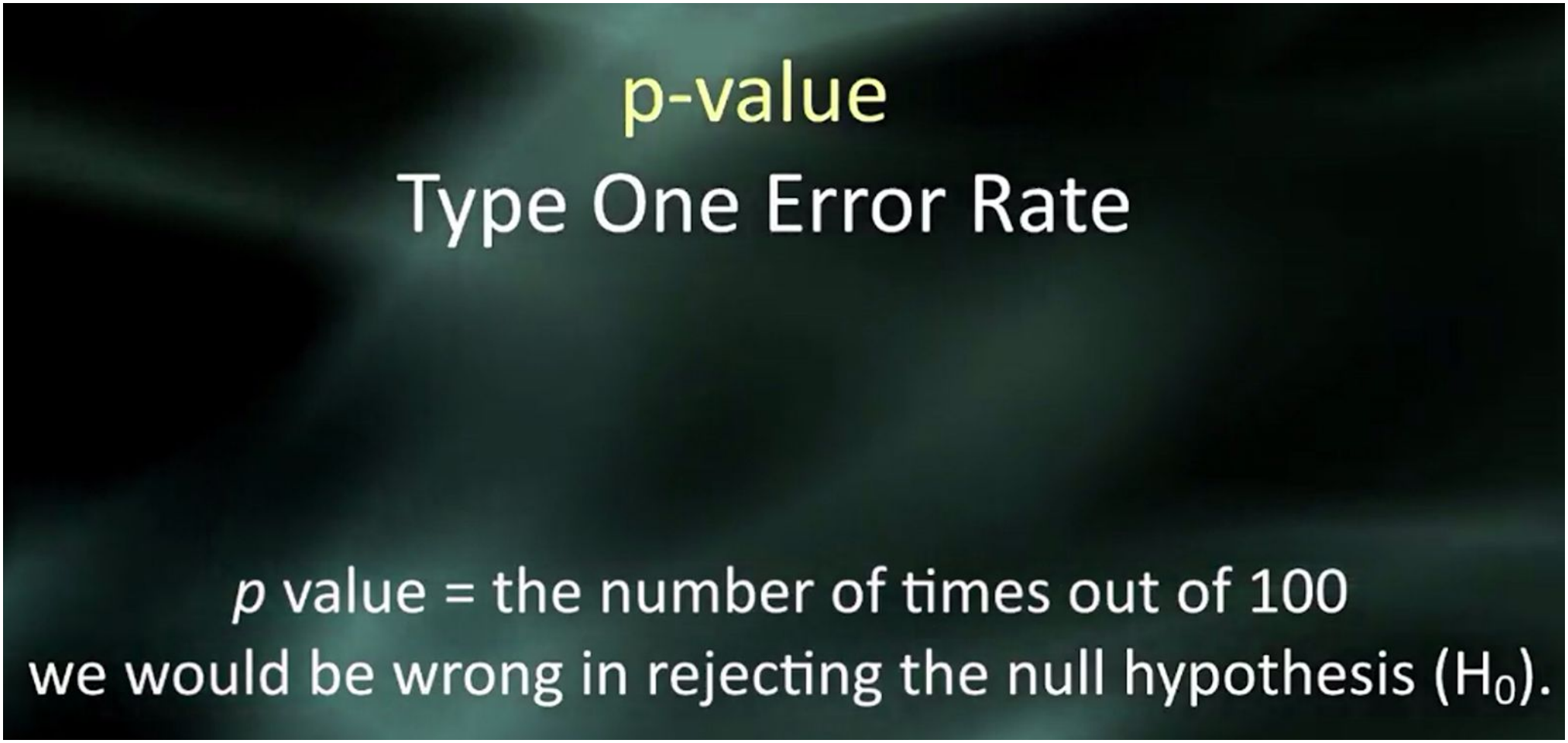
Two Opposing Hypotheses:

Null Hypothesis (H_o)—There is no difference.
Alternate Hypothesis(H_a)—There is a difference.

A result is statistically significant if it is unlikely to have occurred by chance.


Significance Level of a Test
α =0.05
p-value of < 0.05 (5%)

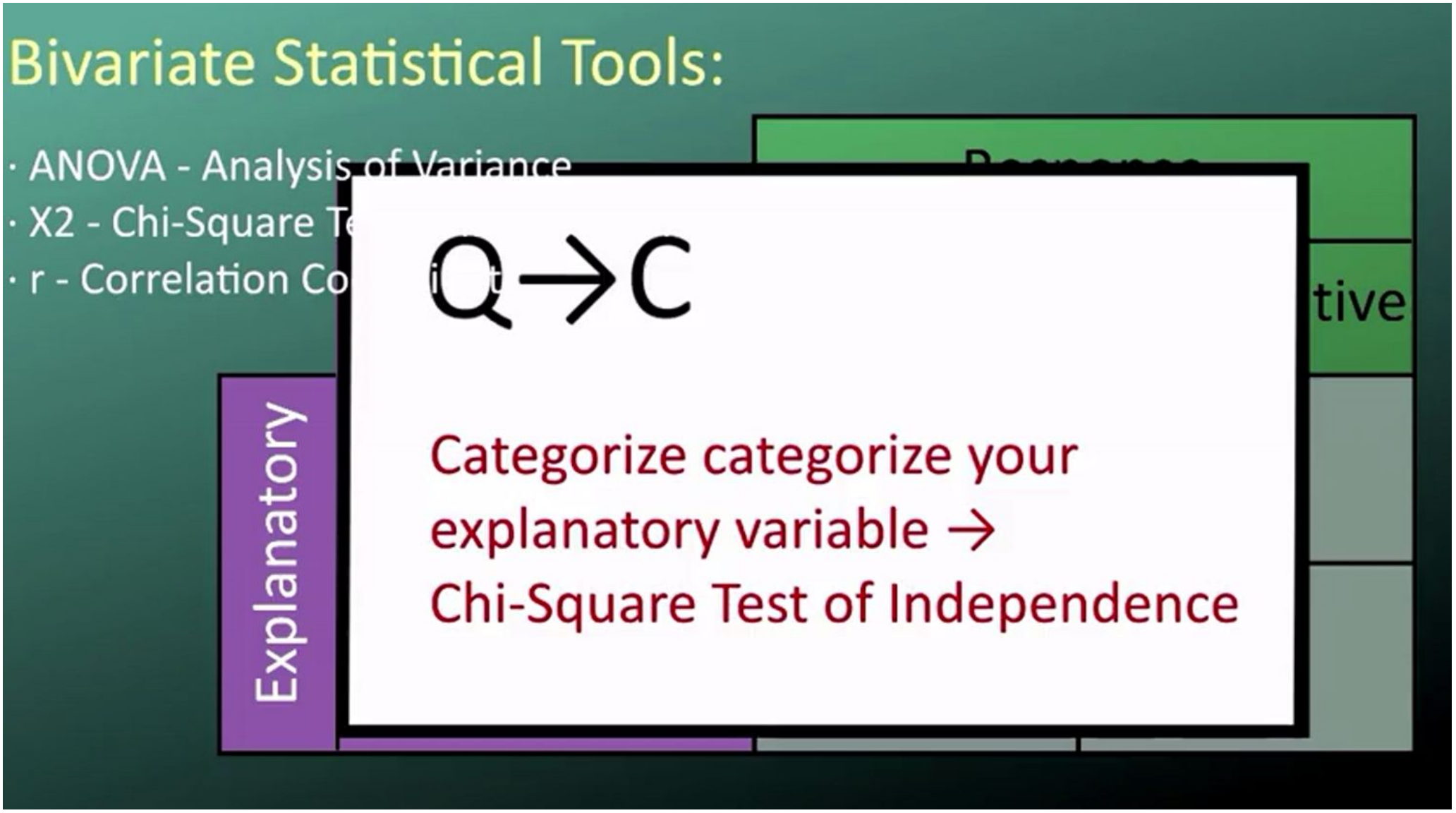

p-value < α (0.05)
0
10
20
30
40
50
60
70
80
90
100
The data provides significant evidence
against the null hypothesis (H_0),
so we reject the null hypothesis (H_0)
and accept the alternate hypothesis (Ha).


p-value
Type One Error Rate
p value = the number of times out of 100
we would be wrong in rejecting the null hypothesis (H_0).

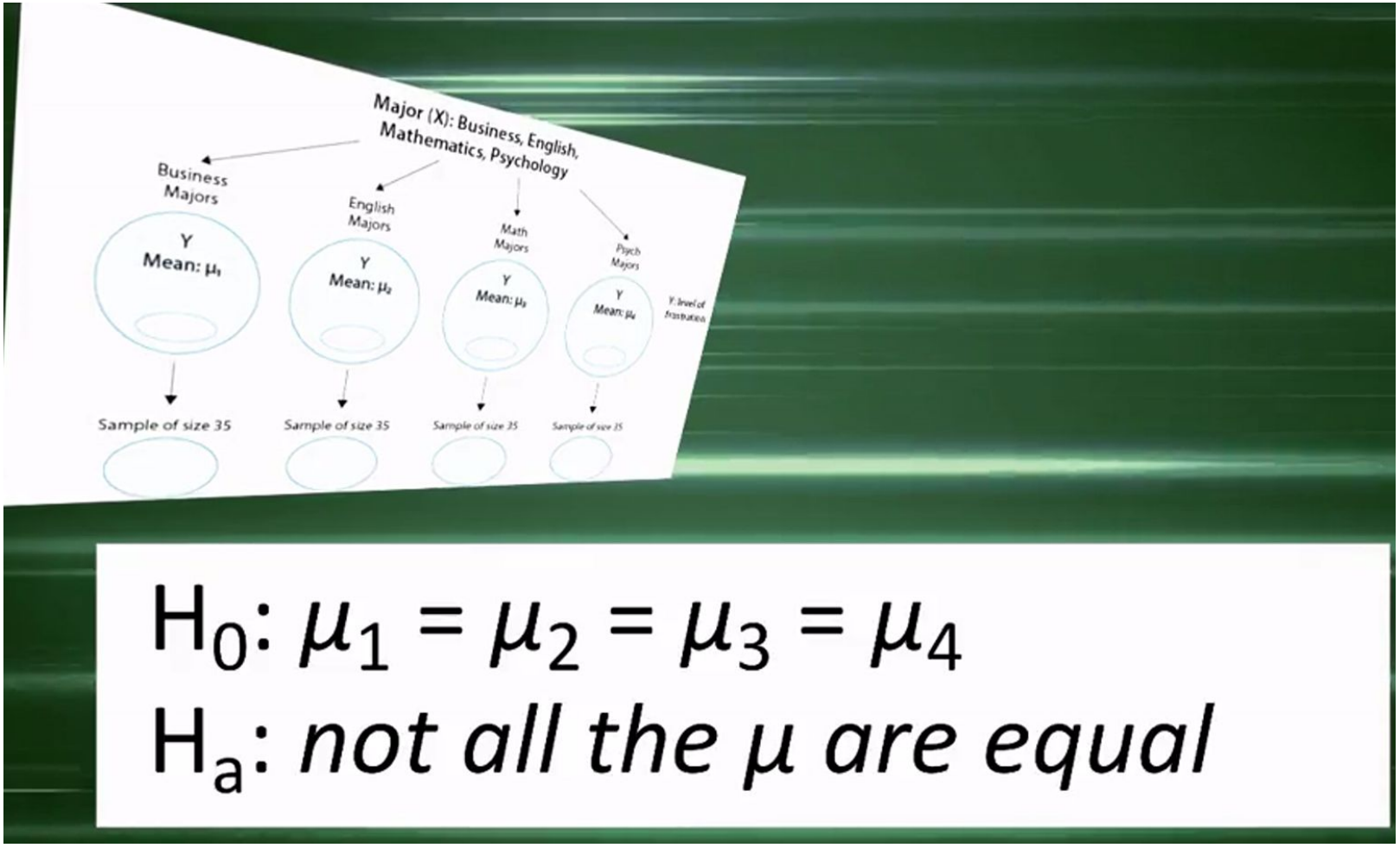
Bivariate Statistical Tools:

- ANOVA - Analysis of Variance
- X2 - Chi-Square Test of Independence
- r - Correlation Coefficient

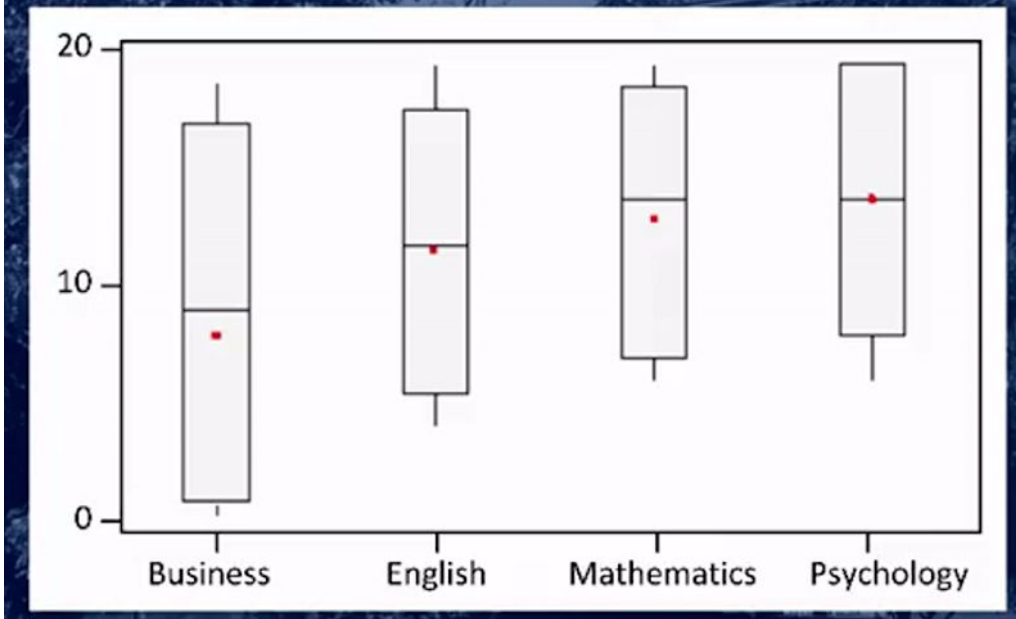
		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C Chi Square Test of Independence	C→Q Analysis of Variance (ANOVA)
	Quantitative	Q→C	Q→Q Pearson Correlation

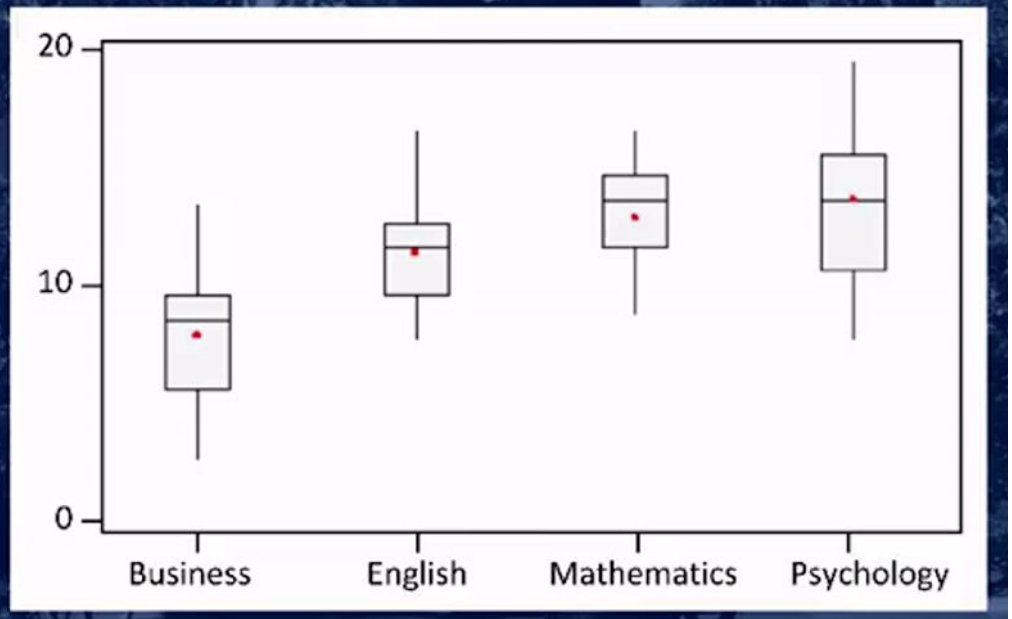

Bivariate Statistical Tools:
· ANOVA - Analysis of Variance
· X2 - Chi-Square Te
· r - Correlation Co
tive
Explanatory
Q→C
Categorize categorize your explanatory variable →
Chi-Square Test of Independence

Ideas behind anova



Are the differences among the sample means due to true differences among the population means or merely due to sampling variability?

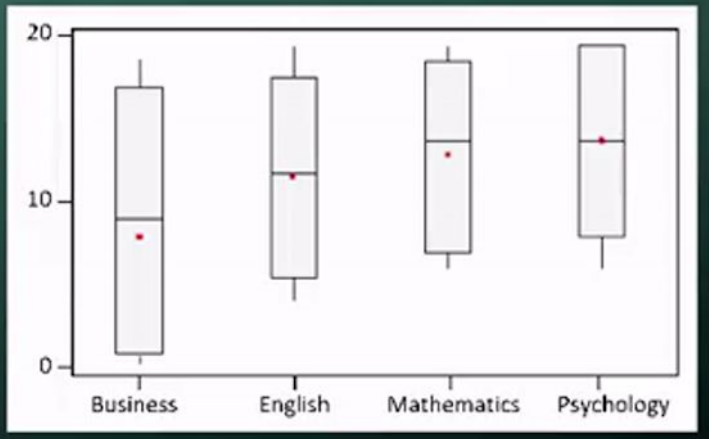


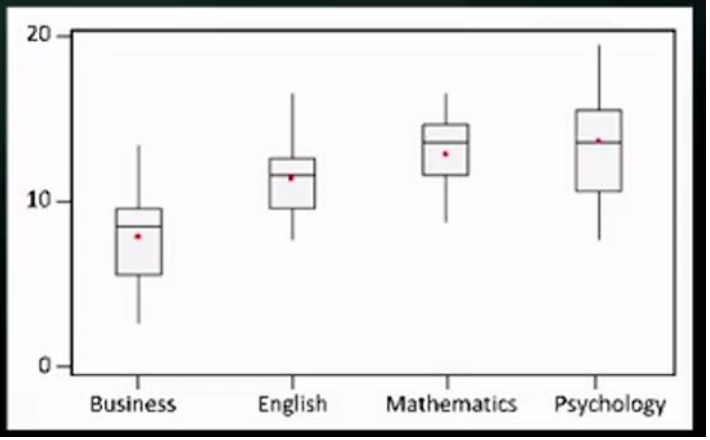


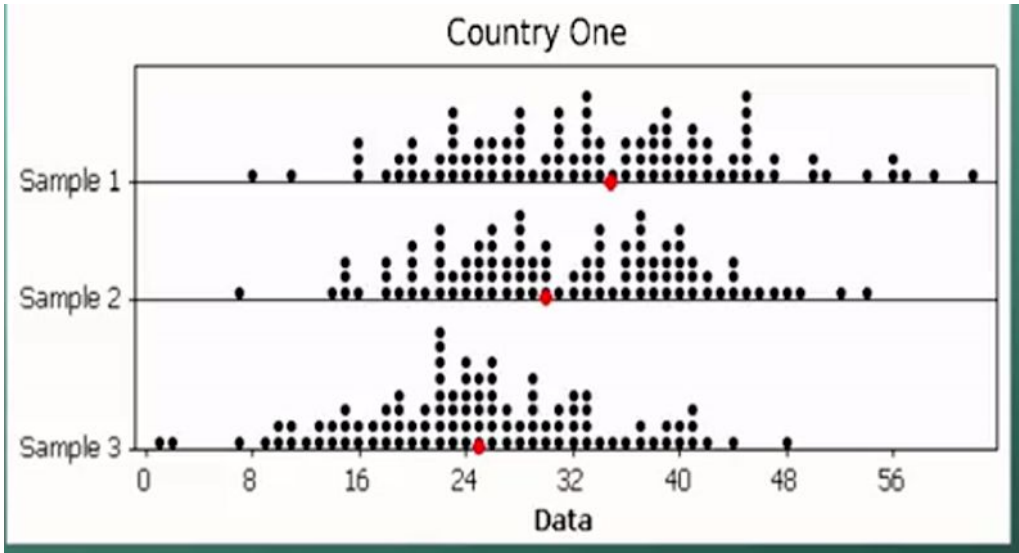
ANOVA F Test

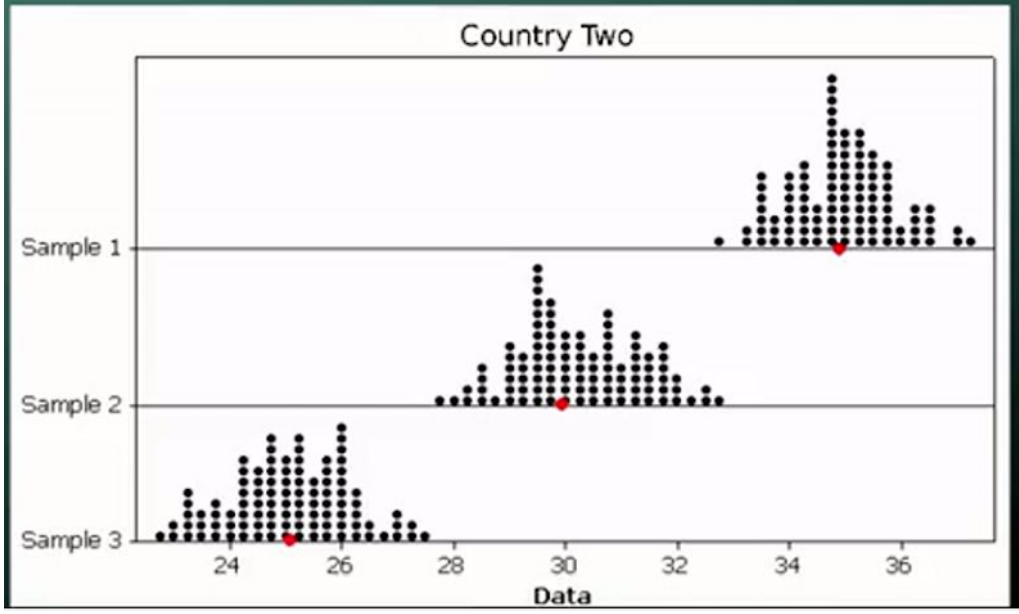
So F is the variation among sample means divided by the variation within groups.

$$F = \frac{\text{Variation Among Sample Means}}{\text{Variation Within Groups}}$$





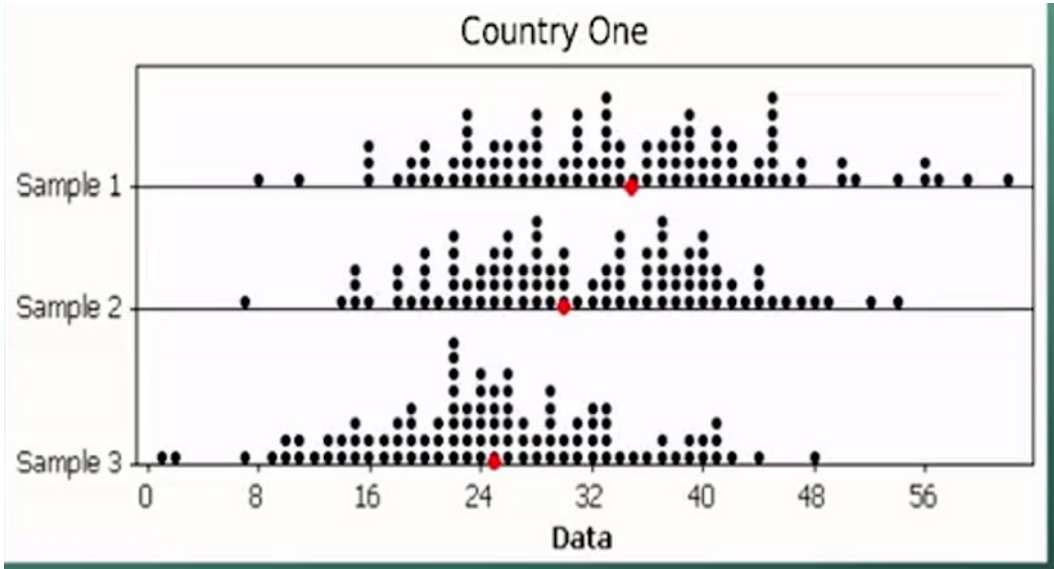


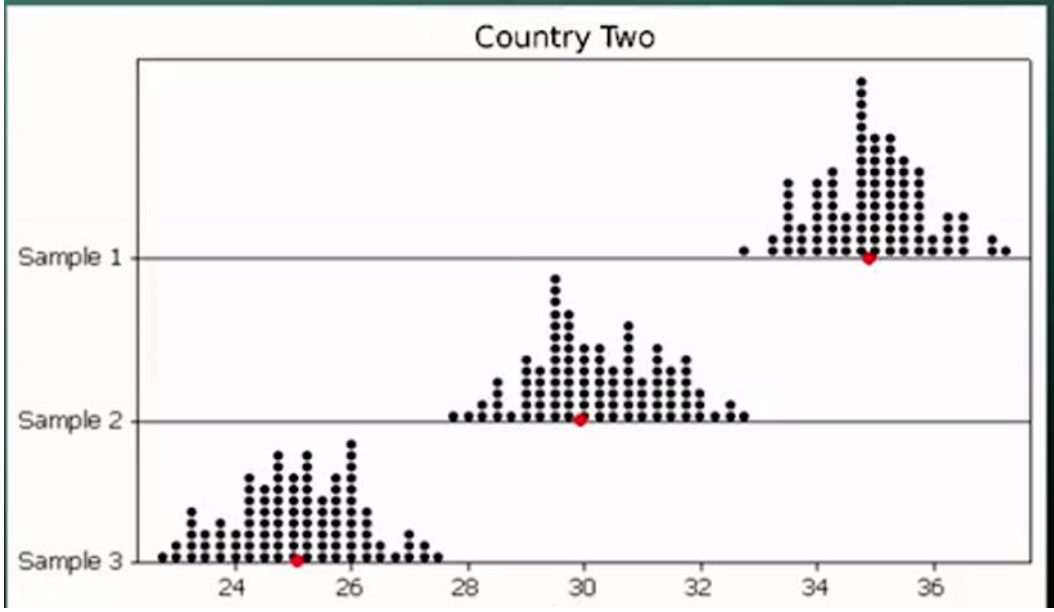


Variation within groups is large.

Differences or variation among the sample means could become negligible.

Data provides very little evidence against the null hypothesis.

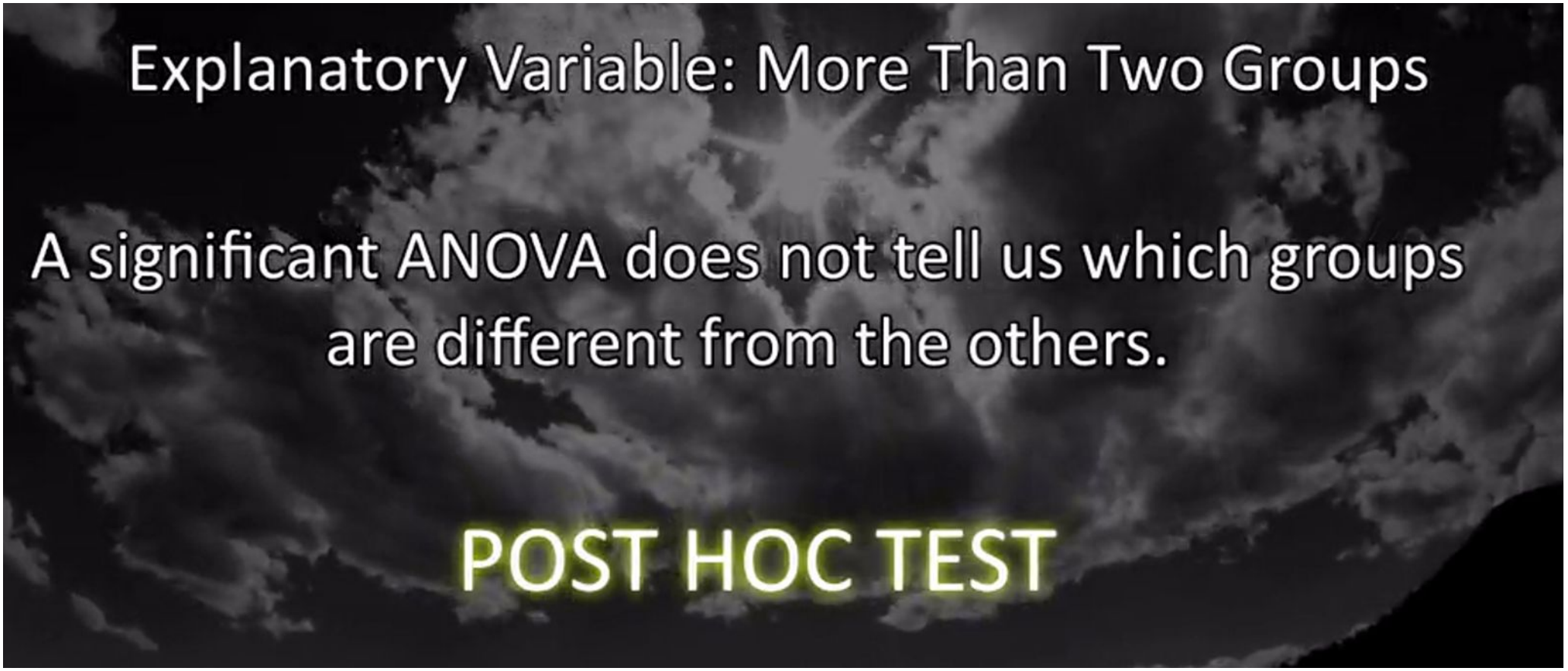


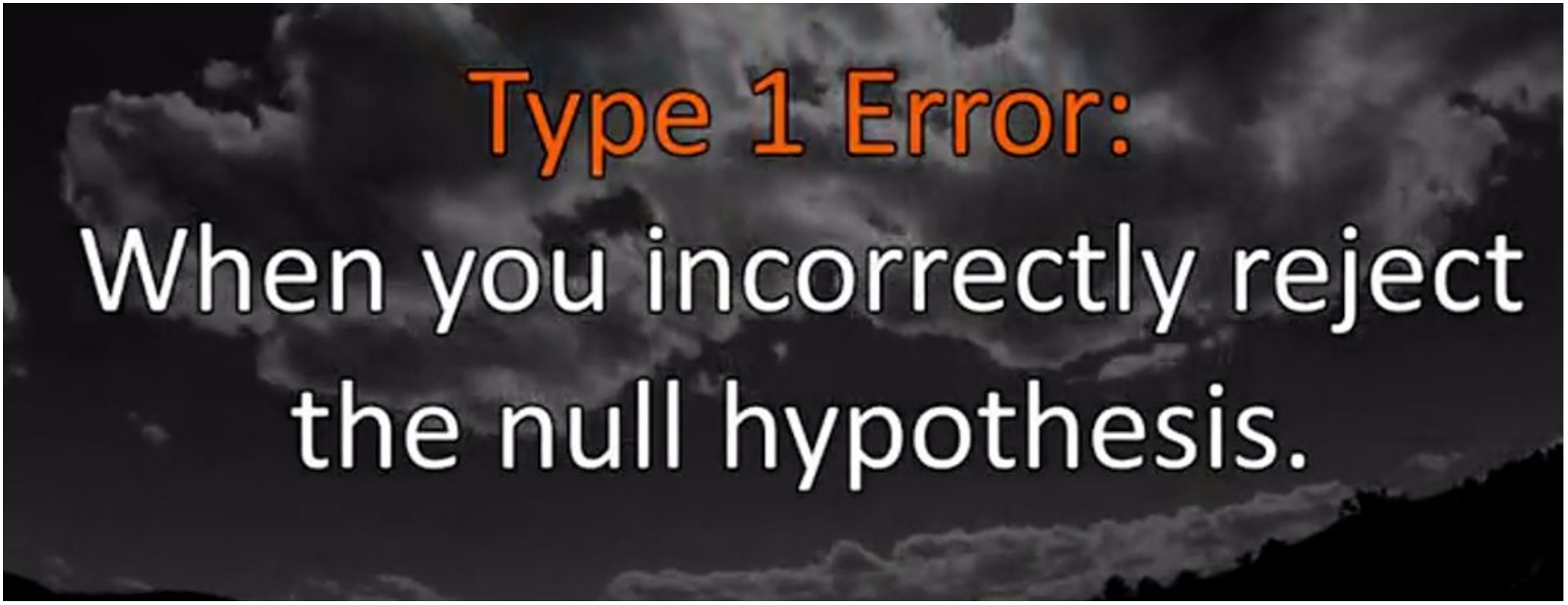


Variation within groups is small.

Variation among the sample means dominates.

Data provides very stronger evidence against the null hypothesis.

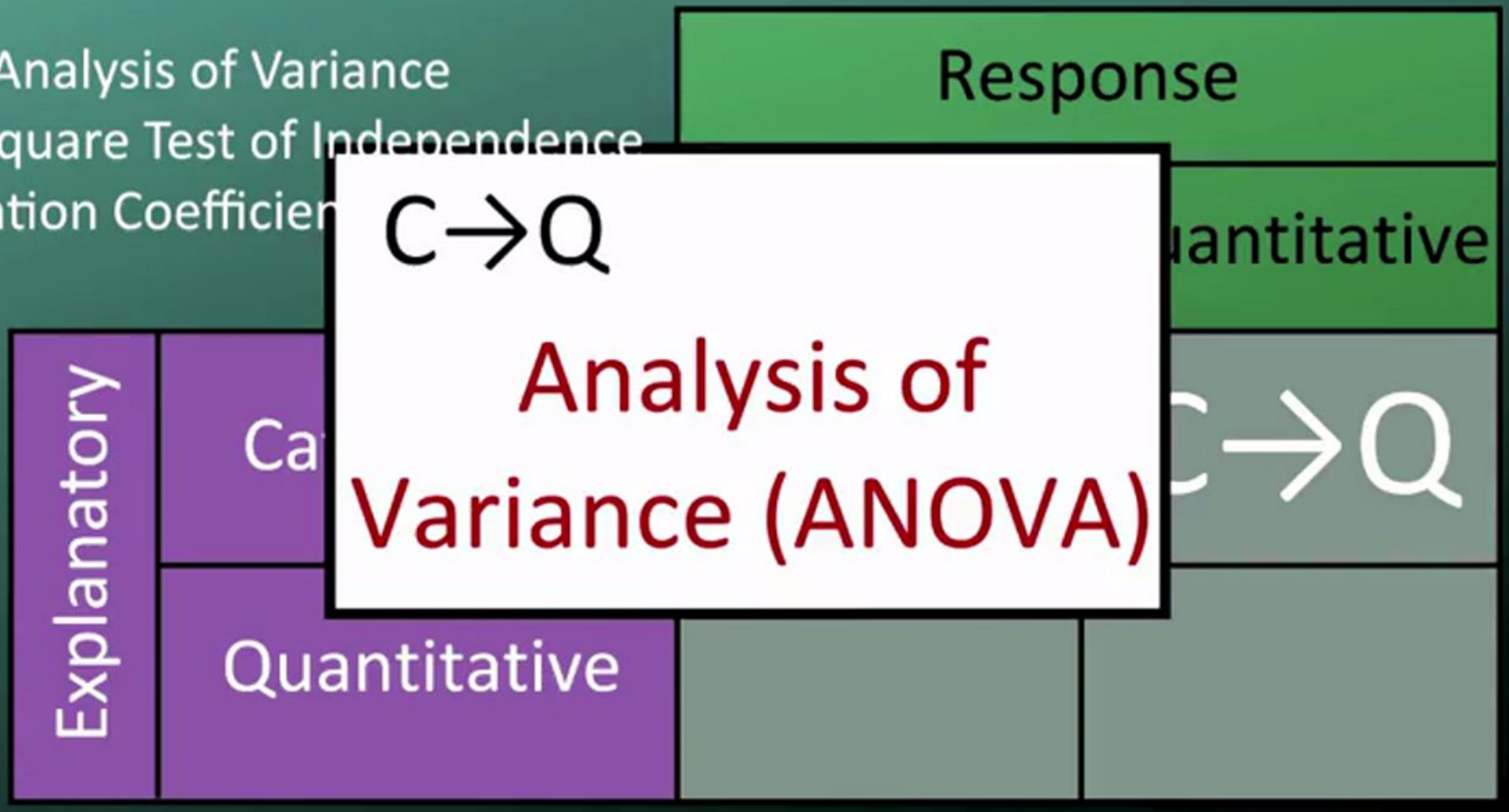

Explanatory Variable: More Than Two Groups
A significant ANOVA does not tell us which groups are different from the others.
POST HOC TEST


Type 1 Error:
When you incorrectly reject
the null hypothesis.

Means with the same letter are not significantly different.				
Duncan Grouping		Mean	N	ETHRACE2A
	A	368.96	1054	1
	A			
B	A	311.44	42	3
B				
B	C	259.59	210	2
B	C			
B	C	244.29	58	4
	C			
	C	220.07	333	5

Bivariate Statistical Tools:

- ANOVA - Analysis of Variance
- X2 - Chi-Square Test of Independence
- r - Correlation Coefficien



Examine differences in the mean response variable for each category of our explanatory variable.

Module1Program-ANOVA.sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
       CHECK321="Smoked Cigarettes in Past 12 Months"
       S3AQ3B1="Usual Smoking Frequency"
       S3AQ3C1="Usual Smoking Quantity";
/*Set appropriate missing data as needed*/
IF S3AQ3B1=9 THEN S3AQ3B1=.;
IF S3AQ3C1=99 THEN S3AQ3C1=.;
IF S3AQ3B1=1 THEN USFREQMO=30;
ELSE IF S3AQ3B1=2 THEN USFREQMO=22;
ELSE IF S3AQ3B1=3 THEN USFREQMO=14;
ELSE IF S3AQ3B1=4 THEN USFREQMO=5;
ELSE IF S3AQ3B1=5 THEN USFREQMO=2.5;
ELSE IF S3AQ3B1=6 THEN USFREQMO=1;
/*USFREQMO usual smoking days per month
1=once a month or less
2.5=2-3 days per month
5=1-2 days per week
14=3-4 days per week
22=5-6 days per week
30=everyday*/
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
PACKSPERMONTH=NUMCIGMO_EST/20;
IF PACKSPERMONTH LE 5 THEN PACKCATEGORY=3;
ELSE IF PACKSPERMONTH LE 10 THEN PACKCATEGORY=7;
ELSE IF PACKSPERMONTH LE 20 THEN PACKCATEGORY=15;
ELSE IF PACKSPERMONTH LE 30 THEN PACKCATEGORY=25;
```

```
ELSE IF PACKSPERMONTH GT 30 THEN PACKCATEGORY=58;
/*subsetting data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by IDNUM;
PROC ANOVA; CLASS MAJORDEPLIFE;
MODEL NUMCIGMO_EST=MAJORDEPLIFE;
MEANS MAJORDEPLIFE;
PROC ANOVA; CLASS ETHRACE2A;
MODEL NUMCIGMO_EST=ETHRACE2A;
MEANS ETHRACE2A/DUNCAN;
RUN;
```

---
week 1.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
      set mydata.gapminder;

/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
            DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
            OUT=popden REPLACE;

PROC IMPORT
            DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
            OUT=pop REPLACE;

PROC IMPORT
            DATAFILE='/home/qallaf890/surface land.csv'
            OUT=surarea REPLACE;

PROC IMPORT
            DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
            OUT=cpi REPLACE;

PROC IMPORT
            DATAFILE='/home/qallaf890/Indicator_HDI.csv'
            OUT=hdi REPLACE;

PROC IMPORT
            DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
```

```
          OUT=murder REPLACE;

PROC IMPORT
          DATAFILE='/home/qallaf890/military expenditure.csv'
          OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
     create table mygapminder AS
     select          gapminder.*
                     ,surarea.surarea
                     ,pop.population
                     ,popden.popden
                     ,cpi.corruptionindex
                     ,hdi.hdi
                     ,murder.homicide
                     ,military.milexpprcntgdp
     from       work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
     set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
     COUNTRY='COUNTRY'
     INCOMEPERPERSON='GDP PER CAPITA'
```

```
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
```

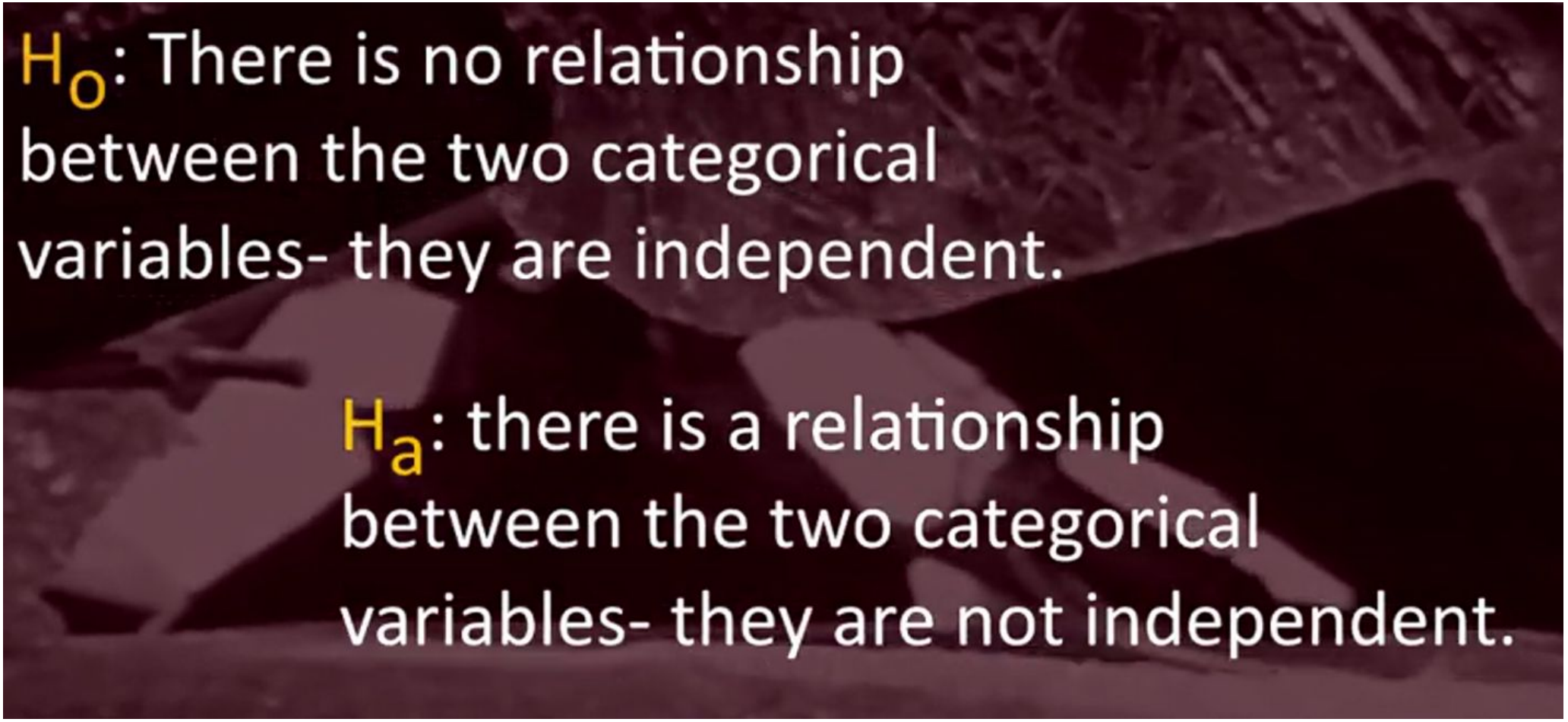
```
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
```

```
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF SUICIDEPER100TH <  4.983422  THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH >= 4.983422  AND  SUICIDEPER100TH <    8.262893   THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 8.262893  AND  SUICIDEPER100TH <    12.367980  THEN SUICIDEPER100TH_RANK = 3;
IF SUICIDEPER100TH >= 12.367980 THEN SUICIDEPER100TH_RANK = 4;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
```

```
IF population = . THEN population_RANK = .;
IF popden <  31.774  THEN popden_RANK = 1;
IF popden >= 31.774  AND  popden < 77.990  THEN popden_RANK = 2;
IF popden >= 77.990  AND  popden < 196.229 THEN popden_RANK = 3;
IF popden >= 196.229 THEN popden_RANK = 4;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;

IF homicide <  1.715654  THEN homicide_RANK = 1;
IF homicide >= 1.715654  AND homicide < 6.111090  THEN homicide_RANK = 2;
IF homicide >= 6.111090  AND homicide < 19.004826 THEN homicide_RANK = 3;
IF homicide >= 19.004826 THEN homicide_RANK = 4;
IF homicide = . THEN homicide_RANK = .;
IF milexpprcntgdp <  1.0752166 THEN milexpprcntgdp_RANK = 1;
IF milexpprcntgdp >= 1.0752166 AND  milexpprcntgdp < 1.4946096 THEN milexpprcntgdp_RANK = 2;
IF milexpprcntgdp >= 1.4946096 AND  milexpprcntgdp < 2.4681756 THEN milexpprcntgdp_RANK = 3;
IF milexpprcntgdp >= 2.4681756 THEN milexpprcntgdp_RANK = 4;
IF milexpprcntgdp = . THEN milexpprcntgdp_RANK = .;

/* FREQUENCY PER VARIABLE */
/*
PROC FREQ;
```

```
TABLES
      EMPLOYRATE_RANK
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH_RANK
      URBANRATE_RANK
      surarea_RANK
      population_RANK
      popden_RANK
      corruptionindex_RANK
      hdi_RANK
      homicide_RANK
      milexpprcntgdp_RANK;
RUN;
*/
/*
PROC ANOVA; CLASS homicide_RANK;
MODEL POPULATION = homicide_RANK;
MEANS homicide_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS corruptionindex_RANK;
MODEL POPULATION = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPULATION = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS homicide_RANK;
MODEL POPDEN = homicide_RANK;
```

```
MEANS homicide_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS corruptionindex_RANK;
MODEL POPDEN = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
*/
ods graphics off;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPDEN = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
ods graphics ON;
/*
PROC GPLOT; PLOT popden*hdi;
RUN;
*/
```


H_o: There is no relationship between the two categorical variables- they are independent.
H_a: there is a relationship between the two categorical variables- they are not independent.

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}}$$

Observed Count

Expected Count

Drank Alcohol in the Last 2 Hours			
Gender (x)	Yes	No	Total
Male	77 / 72.3	404 / 408.7	481
Female	16 / 20.7	122 / 117.3	138
Total	93	526	619

For a 2 by 2 case, $\chi^2 > 3.84$ is considered large.

The p value for the chi squared test of independence is the probability of getting counts like those observed, assuming that the two variables are not related.

Explanatory Variable > Two Levels:
χ^2 and p value do not provide insight into why null hypothesis can be rejected.

BonferroniAdjustment

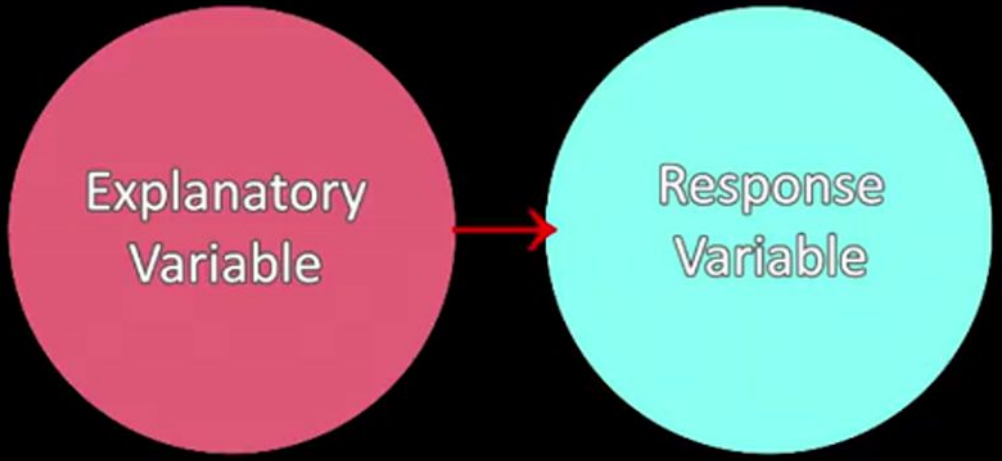
$$\frac{p}{c}$$

c = number of comparisons

Bonferroni Adjustment

χ^2 Test of Independence

$$\frac{.05}{\#\text{ comparisons}}$$



Module2Program-chi2.sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
       CHECK321="Smoked Cigarettes in Past 12 Months"
       S3AQ3B1="Usual Smoking Frequency"
       S3AQ3C1="Usual Smoking Quantity";
/*Set appropriate missing data as needed*/
IF S3AQ3B1=9 THEN S3AQ3B1=.;
IF S3AQ3C1=99 THEN S3AQ3C1=.;
IF S3AQ3B1=1 THEN USFREQMO=30;
ELSE IF S3AQ3B1=2 THEN USFREQMO=22;
ELSE IF S3AQ3B1=3 THEN USFREQMO=14;
ELSE IF S3AQ3B1=4 THEN USFREQMO=5;
ELSE IF S3AQ3B1=5 THEN USFREQMO=2.5;
ELSE IF S3AQ3B1=6 THEN USFREQMO=1;
/*USFREQMO usual smoking days per month
1=once a month or less
2.5=2-3 days per month
5=1-2 days per week
14=3-4 days per week
22=5-6 days per week
30=everyday*/
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
PACKSPERMONTH=NUMCIGMO_EST/20;
IF PACKSPERMONTH LE 5 THEN PACKCATEGORY=3;
ELSE IF PACKSPERMONTH LE 10 THEN PACKCATEGORY=7;
ELSE IF PACKSPERMONTH LE 20 THEN PACKCATEGORY=15;
ELSE IF PACKSPERMONTH LE 30 THEN PACKCATEGORY=25;
```

```
ELSE IF PACKSPERMONTH GT 30 THEN PACKCATEGORY=58;
/*subsetting data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by IDNUM;
/*PROC ANOVA; CLASS ETHRACE2A;
MODEL NUMCIGMO_EST=ETHRACE2A;
MEANS ETHRACE2A/DUNCAN;*/
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON1; SET NEW;
IF USFREQMO=1 OR USFREQMO=2.5;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON2; SET NEW;
IF USFREQMO=1 OR USFREQMO=6;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON3; SET NEW;
IF USFREQMO=1 OR USFREQMO=14;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON4; SET NEW;
IF USFREQMO=1 OR USFREQMO=22;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON5; SET NEW;
```

```
IF USFREQMO=1 OR USFREQMO=30;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON6; SET NEW;
IF USFREQMO=2.5 OR USFREQMO=6;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON7; SET NEW;
IF USFREQMO=2.5 OR USFREQMO=14;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON8; SET NEW;
IF USFREQMO=2.5 OR USFREQMO=22;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON9; SET NEW;
IF USFREQMO=2.5 OR USFREQMO=30;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON10; SET NEW;
IF USFREQMO=6 OR USFREQMO=14;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON11; SET NEW;
IF USFREQMO=6 OR USFREQMO=22;
```

```
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON12; SET NEW;
IF USFREQMO=6 OR USFREQMO=30;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON13; SET NEW;
IF USFREQMO=14 OR USFREQMO=22;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON1; SET NEW;
IF USFREQMO=14 OR USFREQMO=30;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
DATA COMPARISON15; SET NEW;
IF USFREQMO=22 OR USFREQMO=30;
PROC SORT; BY IDNUM;
PROC FREQ; TABLES TAB12MDX*USFREQMO/CHISQ;
RUN;
```

---
week 2.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
```

```
          OUT=murder REPLACE;

PROC IMPORT
          DATAFILE='/home/qallaf890/military expenditure.csv'
          OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
     create table mygapminder AS
     select          gapminder.*
                     ,surarea.surarea
                     ,pop.population
                     ,popden.popden
                     ,cpi.corruptionindex
                     ,hdi.hdi
                     ,murder.homicide
                     ,military.milexpprcntgdp
     from       work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
     set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
     COUNTRY='COUNTRY'
     INCOMEPERPERSON='GDP PER CAPITA'
```

```
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
```

```
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
```

```
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF SUICIDEPER100TH <  4.983422  THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH >= 4.983422  AND  SUICIDEPER100TH <    8.262893   THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 8.262893  AND  SUICIDEPER100TH <    12.367980  THEN SUICIDEPER100TH_RANK = 3;
IF SUICIDEPER100TH >= 12.367980 THEN SUICIDEPER100TH_RANK = 4;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
```

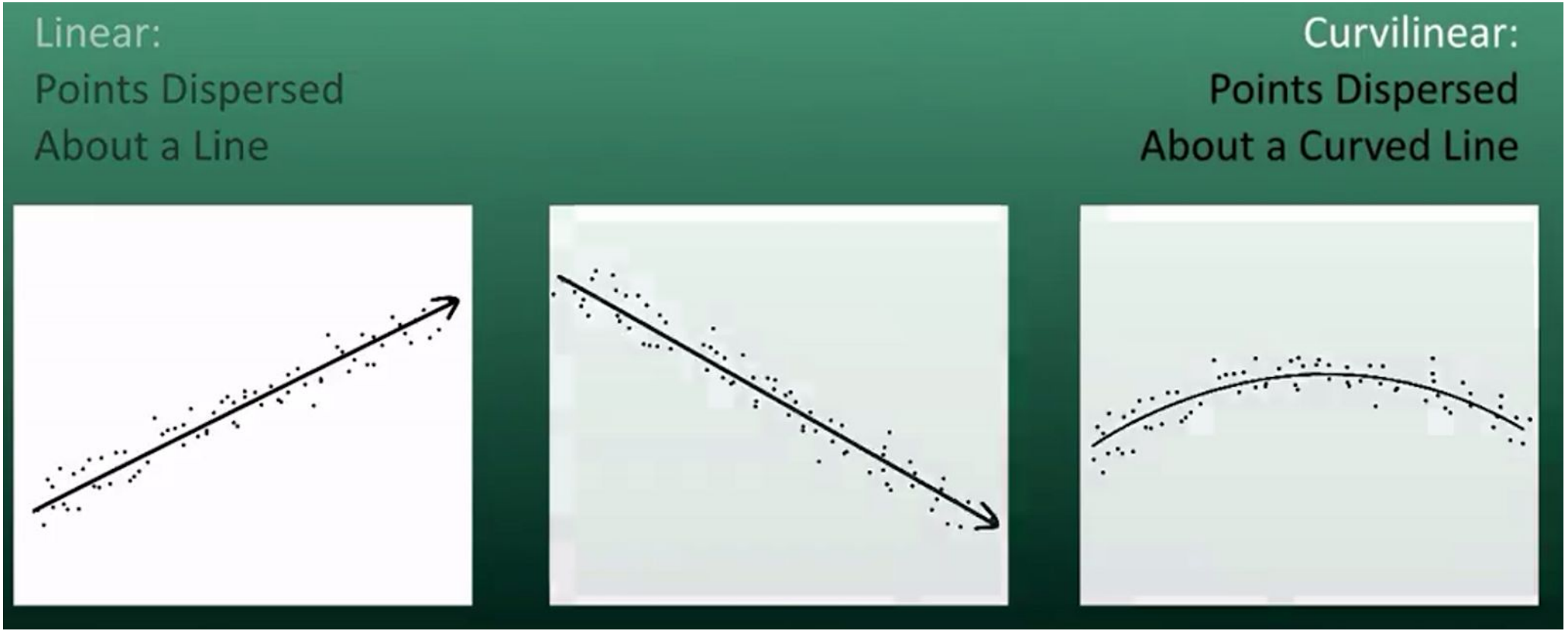
```
IF population = . THEN population_RANK = .;
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;

IF homicide <  1.715654  THEN homicide_RANK = 1;
IF homicide >= 1.715654  AND homicide < 6.111090  THEN homicide_RANK = 2;
IF homicide >= 6.111090  AND homicide < 19.004826 THEN homicide_RANK = 3;
IF homicide >= 19.004826 THEN homicide_RANK = 4;
IF homicide = . THEN homicide_RANK = .;
IF milexpprcntgdp <  1.0752166 THEN milexpprcntgdp_RANK = 1;
IF milexpprcntgdp >= 1.0752166 AND  milexpprcntgdp < 1.4946096 THEN milexpprcntgdp_RANK = 2;
IF milexpprcntgdp >= 1.4946096 AND  milexpprcntgdp < 2.4681756 THEN milexpprcntgdp_RANK = 3;
IF milexpprcntgdp >= 2.4681756 THEN milexpprcntgdp_RANK = 4;
IF milexpprcntgdp = . THEN milexpprcntgdp_RANK = .;

/* FREQUENCY PER VARIABLE */
/*
PROC FREQ;
TABLES
     EMPLOYRATE_RANK
```

```
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH_RANK
      URBANRATE_RANK
      surarea_RANK
      population_RANK
      popden_RANK
      corruptionindex_RANK
      hdi_RANK
      homicide_RANK
      milexpprcntgdp_RANK;
RUN;
*/
/*
PROC ANOVA; CLASS homicide_RANK;
MODEL POPULATION = homicide_RANK;
MEANS homicide_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS corruptionindex_RANK;
MODEL POPULATION = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPULATION = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS homicide_RANK;
MODEL POPDEN = homicide_RANK;
MEANS homicide_RANK/DUNCAN;
RUN;
```

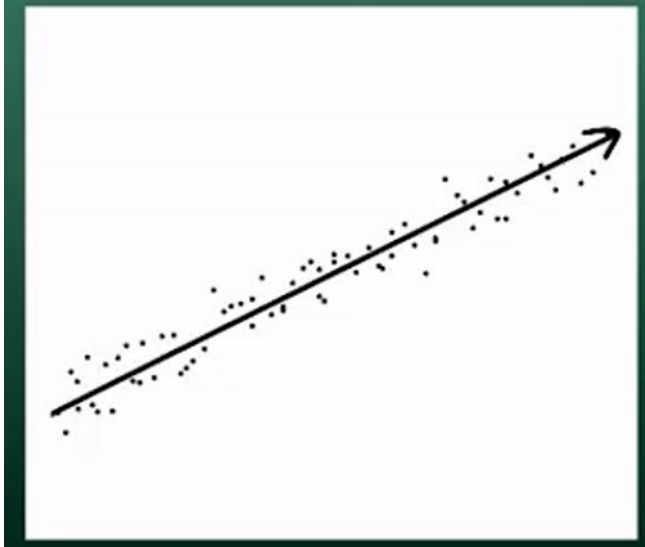
```
PROC ANOVA; CLASS corruptionindex_RANK;
MODEL POPDEN = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
*/
/*
ods graphics off;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPDEN = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
ods graphics ON;
*/
/*
PROC GPLOT; PLOT popden*hdi;
RUN;
*/
/*
proc univariate data = mygapminder plot;
run;
*/
DATA COMPARISON1; SET MYGAPMINDER;
IF HDI_RANK=1 OR HDI_RANK=2;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
DATA COMPARISON2; SET MYGAPMINDER;
IF HDI_RANK=1 OR HDI_RANK=3;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
```

```
DATA COMPARISON3; SET MYGAPMINDER;
IF HDI_RANK=1 OR HDI_RANK=4;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
DATA COMPARISON4; SET MYGAPMINDER;
IF HDI_RANK=2 OR HDI_RANK=3;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
DATA COMPARISON5; SET MYGAPMINDER;
IF HDI_RANK=2 OR HDI_RANK=4;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
DATA COMPARISON6; SET MYGAPMINDER;
IF HDI_RANK=3 OR HDI_RANK=4;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES POPDEN_RANK*HDI_RANK/CHISQ;
RUN;
```

Week 3: Pearson Correlation



Pearson Correlation Coefficient



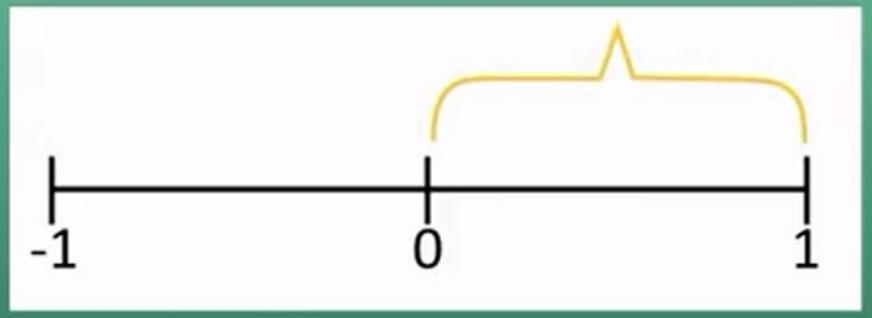
r

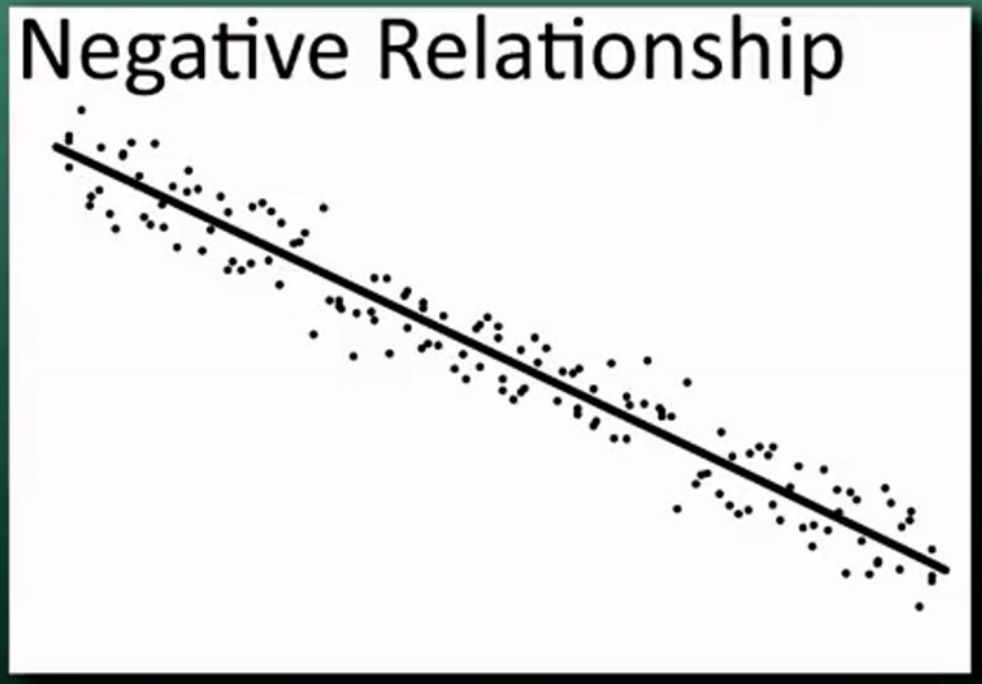
measures a linear relationship
between two quantitative variables.

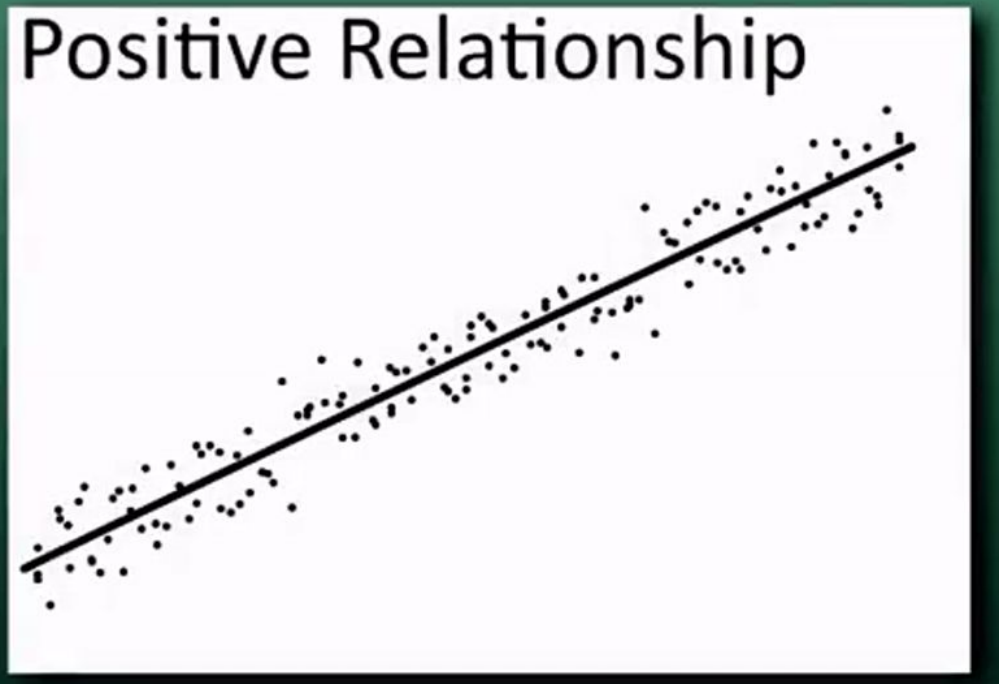
Numerical measure of a linear relationship between two quantitative variables:

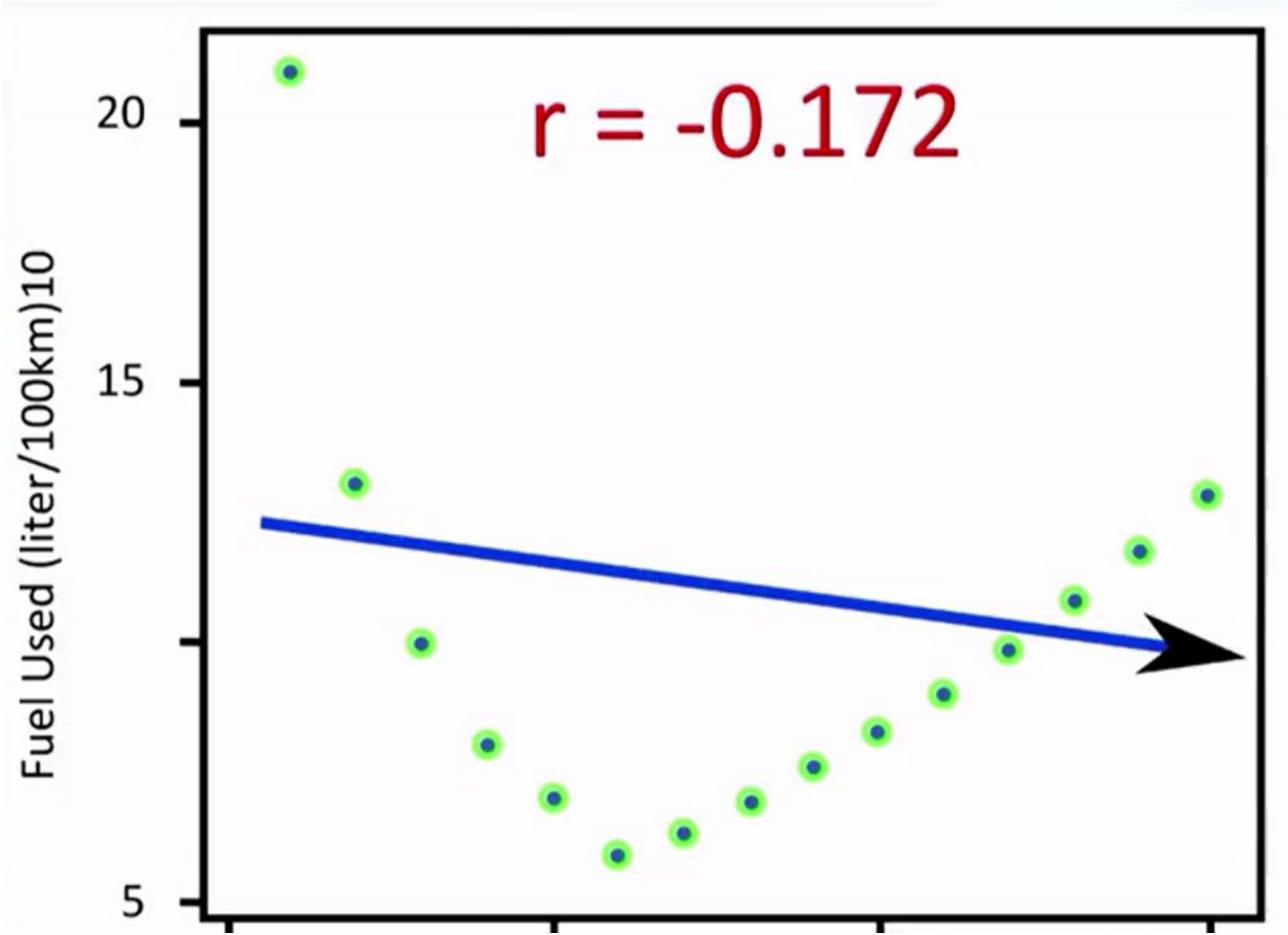
r

Pearson Correlation Coefficient


-1
0
1


Negative Relationship

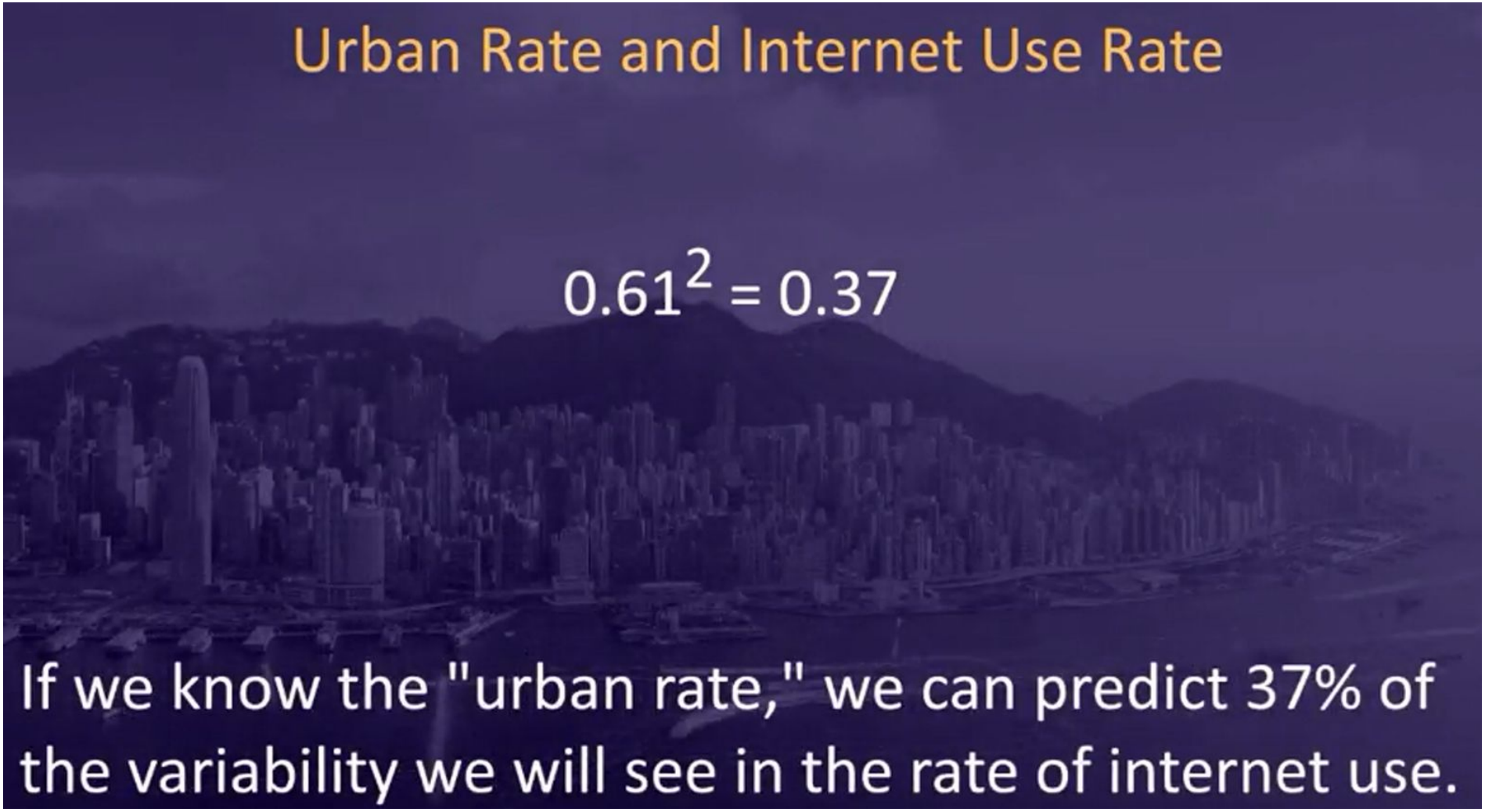

Positive Relationship


r = -0.172
20
15
5
Fuel Used (liter/100km)10

Interpret the scatter plot and the correlation!

An associated p-value is also calculated for the correlation coefficient. The p-value is significant when it is ≤ 0.05.

r^2 = the fraction of the variability of one variable that can be predicted by the other.



Module3Program-correlation.sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.gapminder;
IF incomeperperson eq . THEN incomegroup=.;
ELSE IF incomeperperson LE 744.239 THEN incomegroup=1;
ELSE IF incomeperperson LE 2553.496 THEN incomegroup=2;
ELSE IF incomeperperson LE 9425.236 THEN incomegroup=3;
ELSE IF incomeperperson GT 9425.236 THEN incomegroup=3;
PROC SORT; by COUNTRY;
PROC CORR; VAR urbanrate incomeperperson internetuserate;
RUN;
```

---
week 3.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp
      from       work.gapminder as gapminder
                       left join work.popden as popden on gapminder.country = popden.country
                       left join work.pop as pop on gapminder.country = pop.country
                       left join work.surarea as surarea on gapminder.country = surarea.country
                       left join work.cpi as cpi on gapminder.country = cpi.country
                       left join work.hdi as hdi on gapminder.country = hdi.country
                       left join work.murder as murder on gapminder.country = murder.country
                       left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
```

```
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
      LIFEEXPECTANCY_RANK
```

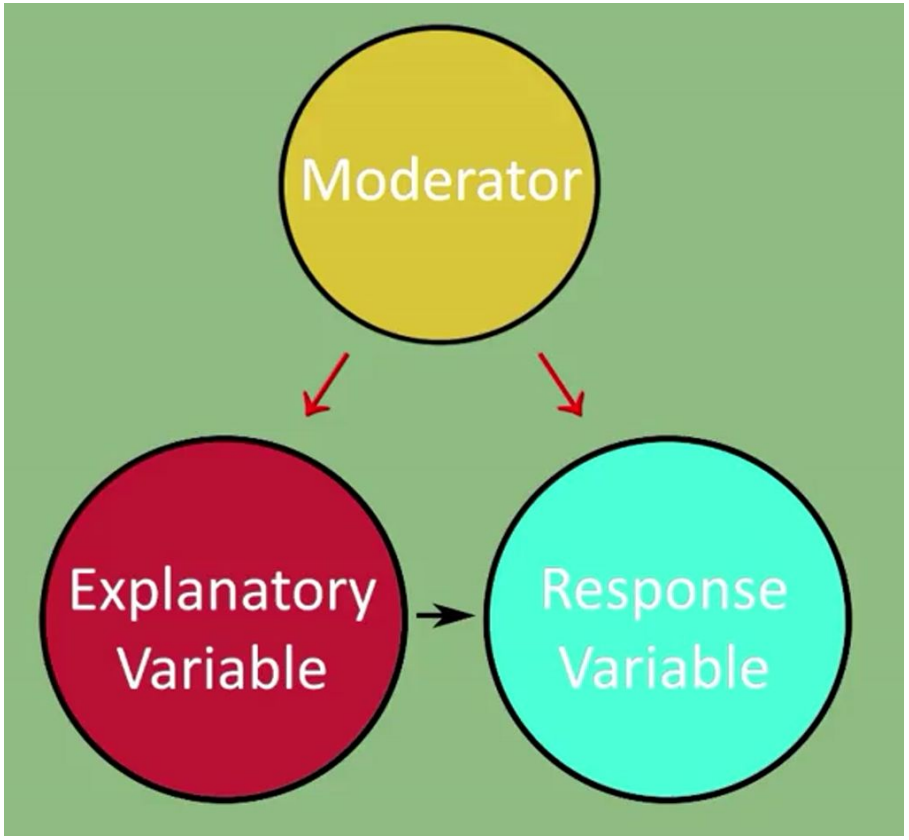
```
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
```

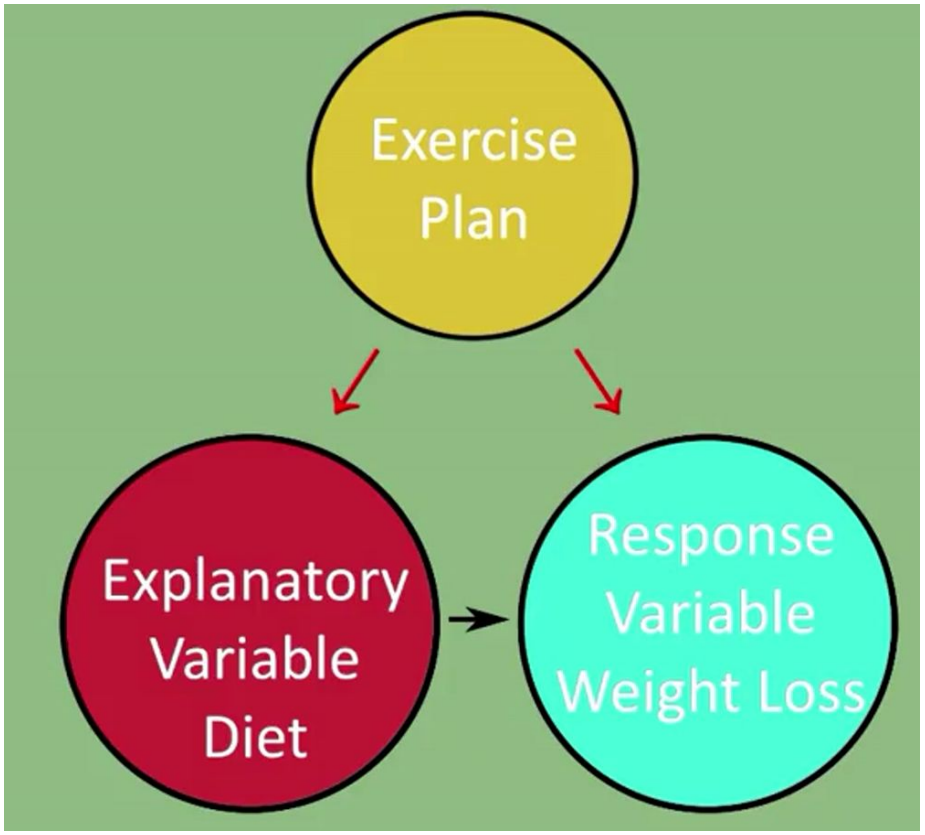
```
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF SUICIDEPER100TH <  4.983422  THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH >= 4.983422  AND  SUICIDEPER100TH <    8.262893   THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 8.262893  AND  SUICIDEPER100TH <    12.367980  THEN SUICIDEPER100TH_RANK = 3;
IF SUICIDEPER100TH >= 12.367980 THEN SUICIDEPER100TH_RANK = 4;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = .;
```

```
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;

IF homicide <  1.715654  THEN homicide_RANK = 1;
IF homicide >= 1.715654  AND homicide < 6.111090  THEN homicide_RANK = 2;
IF homicide >= 6.111090  AND homicide < 19.004826 THEN homicide_RANK = 3;
IF homicide >= 19.004826 THEN homicide_RANK = 4;
IF homicide = . THEN homicide_RANK = .;
IF milexpprcntgdp <  1.0752166 THEN milexpprcntgdp_RANK = 1;
IF milexpprcntgdp >= 1.0752166 AND  milexpprcntgdp < 1.4946096 THEN milexpprcntgdp_RANK = 2;
IF milexpprcntgdp >= 1.4946096 AND  milexpprcntgdp < 2.4681756 THEN milexpprcntgdp_RANK = 3;
IF milexpprcntgdp >= 2.4681756 THEN milexpprcntgdp_RANK = 4;
IF milexpprcntgdp = . THEN milexpprcntgdp_RANK = .;

/* FREQUENCY PER VARIABLE */
/*
PROC FREQ;
TABLES
      EMPLOYRATE_RANK
      INCOMEPERPERSON_RANK
```

```
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH_RANK
      URBANRATE_RANK
      surarea_RANK
      population_RANK
      popden_RANK
      corruptionindex_RANK
      hdi_RANK
      homicide_RANK
      milexpprcntgdp_RANK;
RUN;
*/
/*
PROC ANOVA; CLASS homicide_RANK;
MODEL POPULATION = homicide_RANK;
MEANS homicide_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS corruptionindex_RANK;
MODEL POPULATION = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPULATION = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS homicide_RANK;
MODEL POPDEN = homicide_RANK;
MEANS homicide_RANK/DUNCAN;
RUN;
PROC ANOVA; CLASS corruptionindex_RANK;
```

```
MODEL POPDEN = corruptionindex_RANK;
MEANS corruptionindex_RANK/DUNCAN;
RUN;
*/
/*
ods graphics off;
PROC ANOVA; CLASS hdi_RANK;
MODEL POPDEN = hdi_RANK;
MEANS hdi_RANK/DUNCAN;
RUN;
ods graphics ON;
*/
PROC CORR; DATA MYGAPMINDER;
RUN;
PROC GPLOT; PLOT corruptionindex*INCOMEPERPERSON;
RUN;
PROC GPLOT; PLOT HDI*LIFEEXPECTANCY;
RUN;
```


Moderator
Explanatory Variable
Response Variable


Exercise Plan
Explanatory Variable Diet
Response Variable Weight Loss

A third variable that effects the direction and or strength of the relation between your explanatory and response variable.

Is our explanatory variable associated with our response variable, for each population sub-group or each level of our third variable?

Module4Program-Moderation.sas

```
/*Example Using ANOVA*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.diet_exercise;
PROC SORT; BY EXERCISE;
PROC ANOVA; CLASS DIET;
MODEL WEIGHTLOSS=DIET;
MEANS DIET; BY EXERCISE
RUN;
/*Example Using Chi-Square*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.nesarc_pds;
LABEL TAB12MDX="Tobacco Dependence Past 12 Months"
       CHECK321="Smoked Cigarettes in Past 12 Months"
       S3AQ3B1="Usual Smoking Frequency"
       S3AQ3C1="Usual Smoking Quantity";
/*Set appropriate missing data as needed*/
IF S3AQ3B1=9 THEN S3AQ3B1=.;
IF S3AQ3C1=99 THEN S3AQ3C1=.;
IF S3AQ3B1=1 THEN USFREQMO=30;
ELSE IF S3AQ3B1=2 THEN USFREQMO=22;
ELSE IF S3AQ3B1=3 THEN USFREQMO=14;
ELSE IF S3AQ3B1=4 THEN USFREQMO=5;
ELSE IF S3AQ3B1=5 THEN USFREQMO=2.5;
ELSE IF S3AQ3B1=6 THEN USFREQMO=1;
/*USFREQMO usual smoking days per month
1=once a month or less
2.5=2-3 days per month
5=1-2 days per week
```

```
14=3-4 days per week
22=5-6 days per week
30=everyday*/
NUMCIGMO_EST=USFREQMO*S3AQ3C1;
PACKSPERMONTH=NUMCIGMO_EST/20;
IF PACKSPERMONTH LE 5 THEN PACKCATEGORY=3;
ELSE IF PACKSPERMONTH LE 10 THEN PACKCATEGORY=7;
ELSE IF PACKSPERMONTH LE 20 THEN PACKCATEGORY=15;
ELSE IF PACKSPERMONTH LE 30 THEN PACKCATEGORY=25;
ELSE IF PACKSPERMONTH GT 30 THEN PACKCATEGORY=58;
/*USQUAN: 0=nondaily smoking; 3=1-5 cigs/day; 8=6-10 cigs/day;
13=11-15 cigs/day; 18=16-20 cigs/day; 37=21-37 cigs/day*/
IF S3AQ3C1 NE 1 THEN USQUAN=0;
ELSE IF S3AQ3C1 GE 1 AND S3AQ3C1 LE 5 THEN USQUAN=3;
ELSE IF S3AQ3C1 GE 6 AND S3AQ3C1 LE 10 THEN USQUAN=8;
ELSE IF S3AQ3C1 GE 11 AND S3AQ3C1 LE 15 THEN USQUAN=13;
ELSE IF S3AQ3C1 GE 16 AND S3AQ3C1 LE 20 THEN USQUAN=18;
ELSE IF S3AQ3C1 GE 20 THEN USQUAN=37;
/*subsetting data to include only past 12 month smokers, age 18-25*/
IF CHECK321=1;
IF AGE LE 25;
PROC SORT; by MAJORDEPLIFE;
PROC FREQ; TABLES TAB12MDX*USQUAN/CHISQ;
BY MAJORDEPLIFE;
/*PACKCATEGORY PACKSPERMONTH TAB12MDX CHECK321 S3AQ31 S3AQ3C1;*/
PROC GCHART; VBAR USQUAN/discrete type=mean SUMVAR=TAB12MDX;
RUN;
/*Example Using Pearson Correlaton*/
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.gapminder;
IF incomeperperson EQ . THEN incomegrp=.;
```

```
ELSE IF incomeperperson LE 744.239 THEN incomegrp=1;
ELSE IF incomeperperson LE 9425.236 THEN incomegrp=2;
ELSE IF incomeperperson GE 9425.236 THEN incomegrp=3;
IF incomegrp NE .;
PROC SORT; by COUNTRY;
PROC SORT; by incomegrp;
PROC CORR; VAR urbanrate internetuserate; BY incomegrp;
RUN;
```

---
week 4.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
      set mydata.gapminder;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
ELSE IF INCOMEPERPERSON <=  744.239  THEN INCOMEPERPERSON_RANK = 1;
ELSE IF INCOMEPERPERSON <=  9425.326  THEN INCOMEPERPERSON_RANK = 2;
ELSE IF INCOMEPERPERSON  >  9425.326  THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON_RANK NE .;
PROC SORT; BY COUNTRY;
PROC SORT; BY INCOMEPERPERSON_RANK;
PROC CORR; VAR URBANRATE INTERNETUSERATE; BY INCOMEPERPERSON_RANK;
RUN;
PROC GPLOT;
      PLOT INTERNETUSERATE*URBANRATE;
      BY INCOMEPERPERSON_RANK;
RUN;
```

3. Regression Modeling in Practice

True Experiment

- Only one variable is manipulated
- Control group
- Random assignment


X causes Y
X is associated with Y

EXPERIMENTAL DATA

True Experiment

- Only one variable is manipulated
- Control group
- Random assignment

Quasi Experiment

treatment and control groups are pre-selected

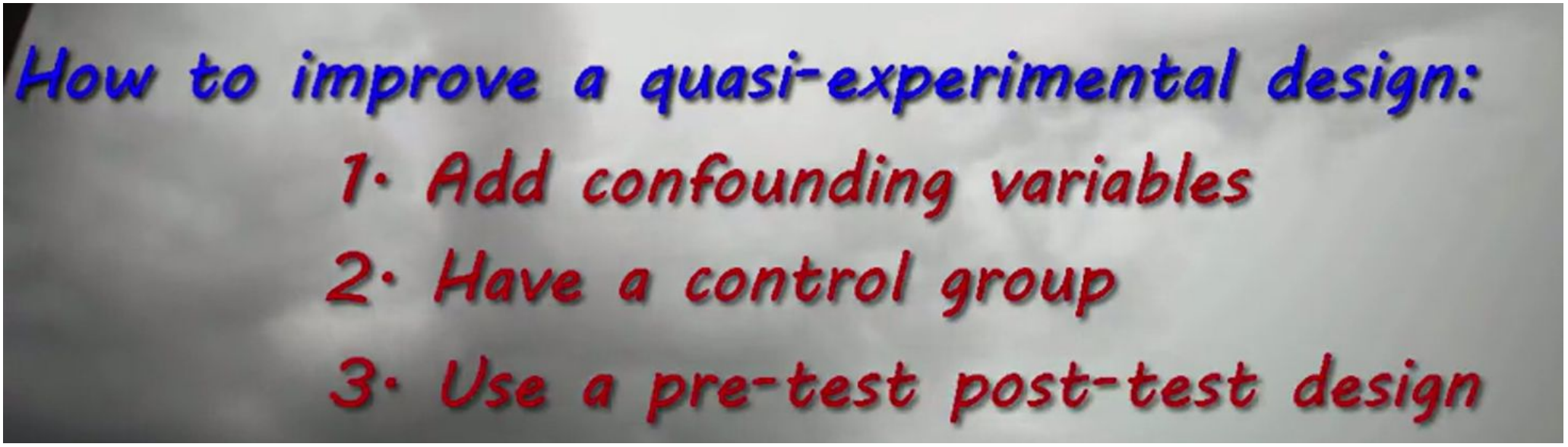
EXPERIMENTAL DATA

True Experiment

- Only one variable is manipulated
- Control group
- Random assignment

Quasi Experiment

- Only one variable is manipulated
- Control group
- ***No*** random assignment

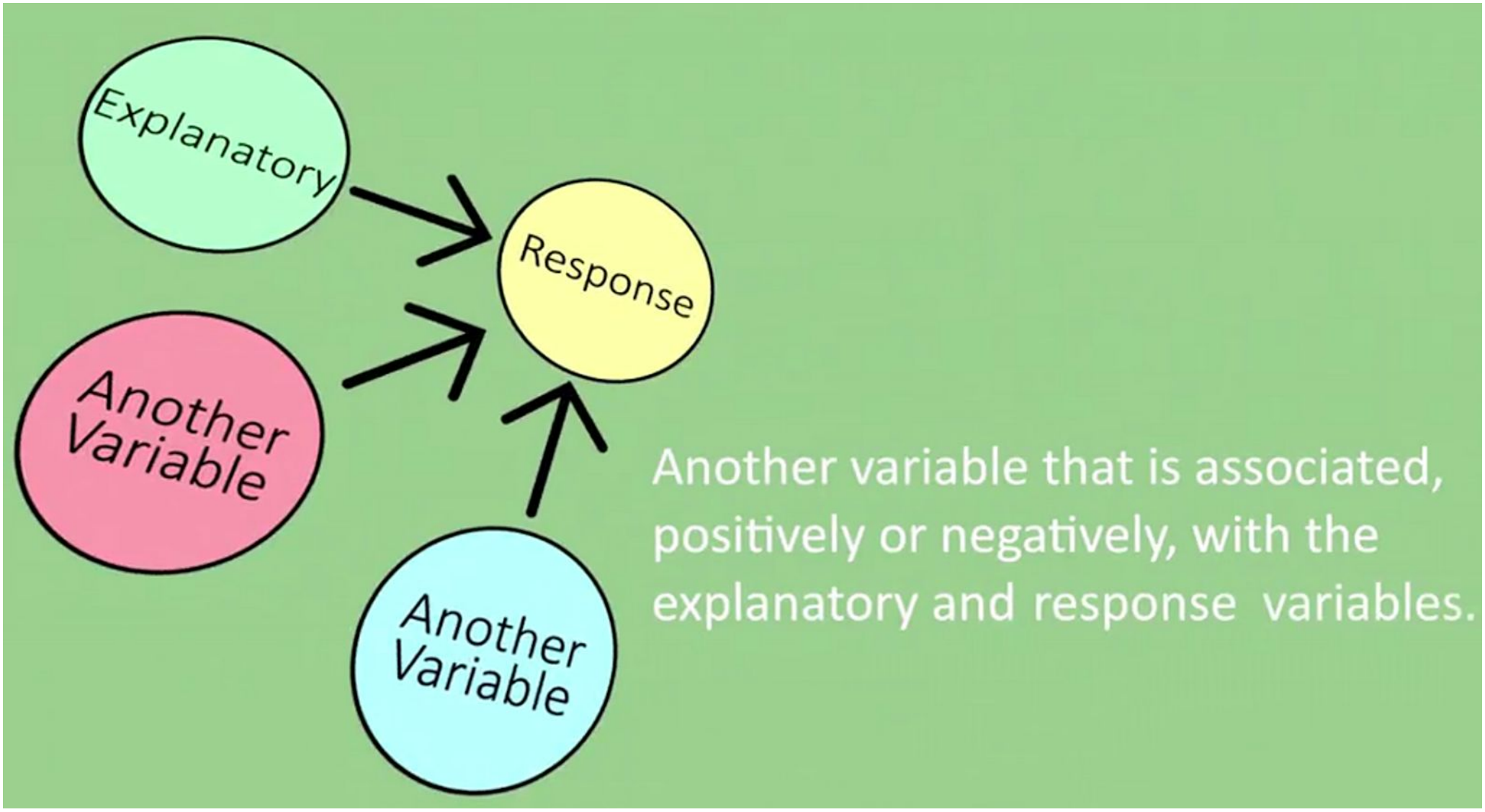

How to improve a quasi-experimental design:
1. Add confounding variables
2. Have a control group
3. Use a pre-test post-test design

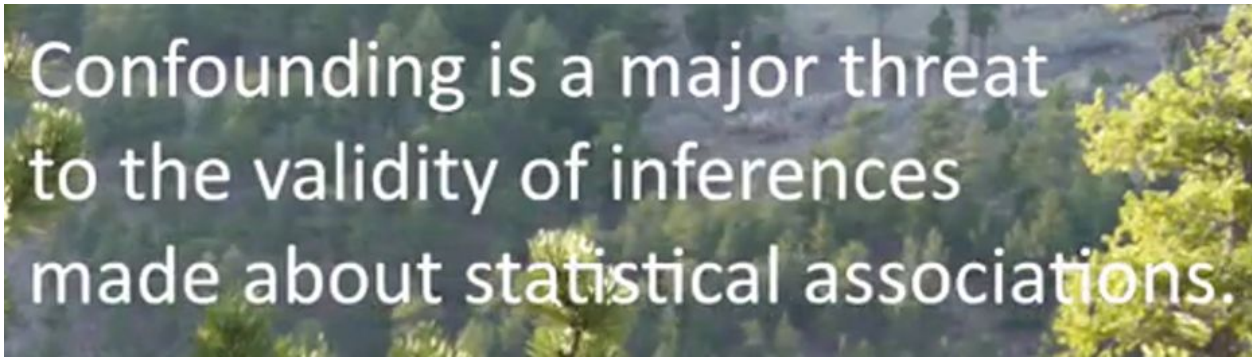
CONFOUNDING VARIABLE?

SERIOUSNESS OF THE FIRE

NUMBER OF FIREFIGHTERS
EXPLANATORY VARIABLE

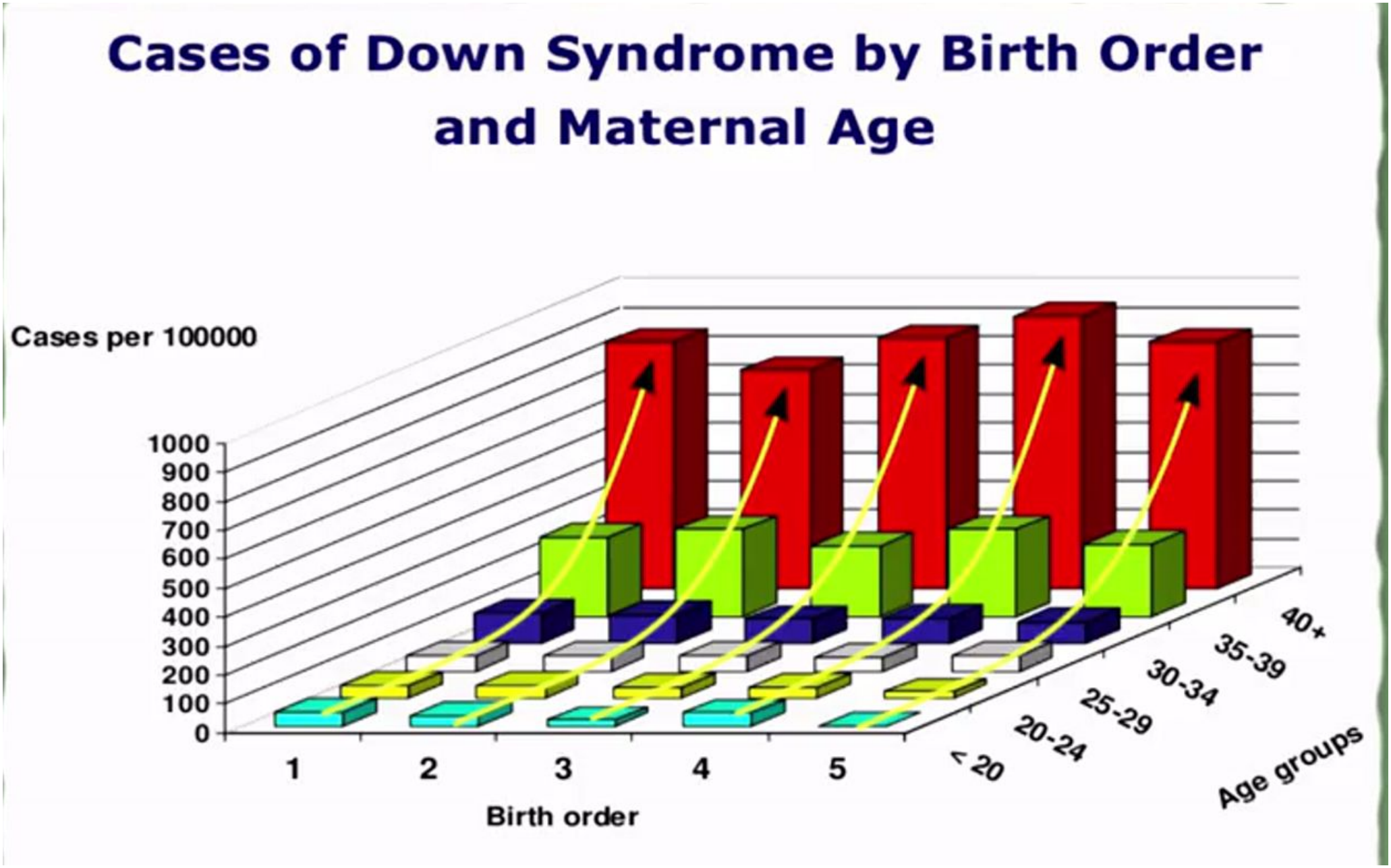
DAMAGE CAUSE BY FIRE
RESPONSE VARIABLE


Explanatory
Response
Another Variable
Another Variable
Another variable that is associated, positively or negatively, with the explanatory and response variables.

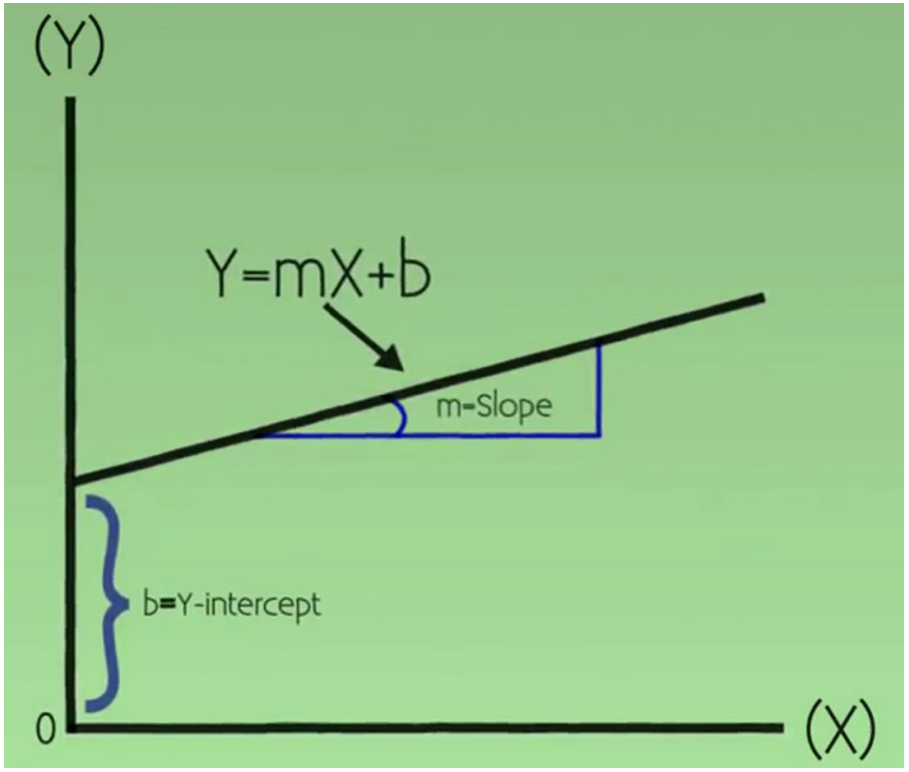

Confounding is a major threat
to the validity of inferences
made about statistical associations.


We want to demonstrate that
our association of interest is
significant even after controlling
for potential confounders.

Cases of Down Syndrome by Birth Order and Maternal Age



⇒ How-to-Write-About-Your-Data.pdf




```
PROC GLM; model
quantitativeresponse=explanatory;
```

Linear Regression Assumptions

Normality

Linearity

Homoscedasticity

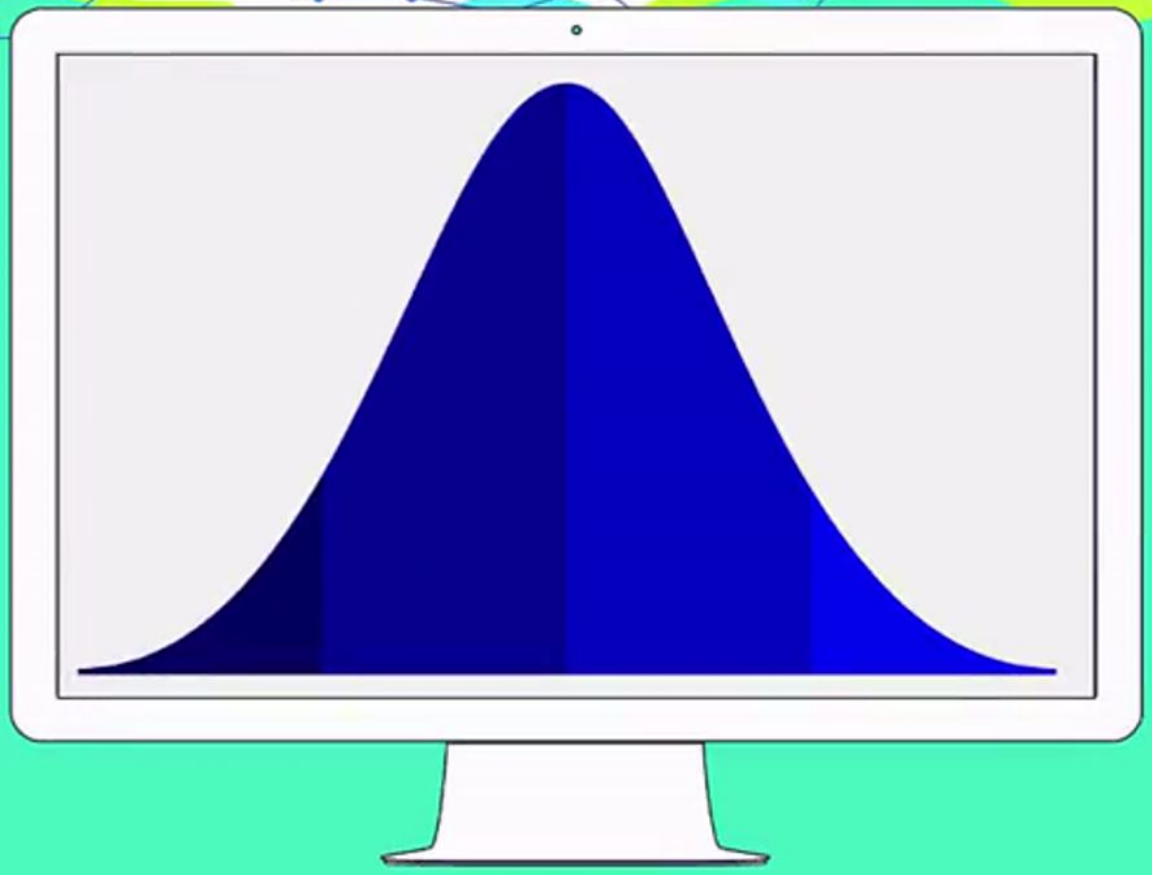
Independence

Multicollinearity

Outliers

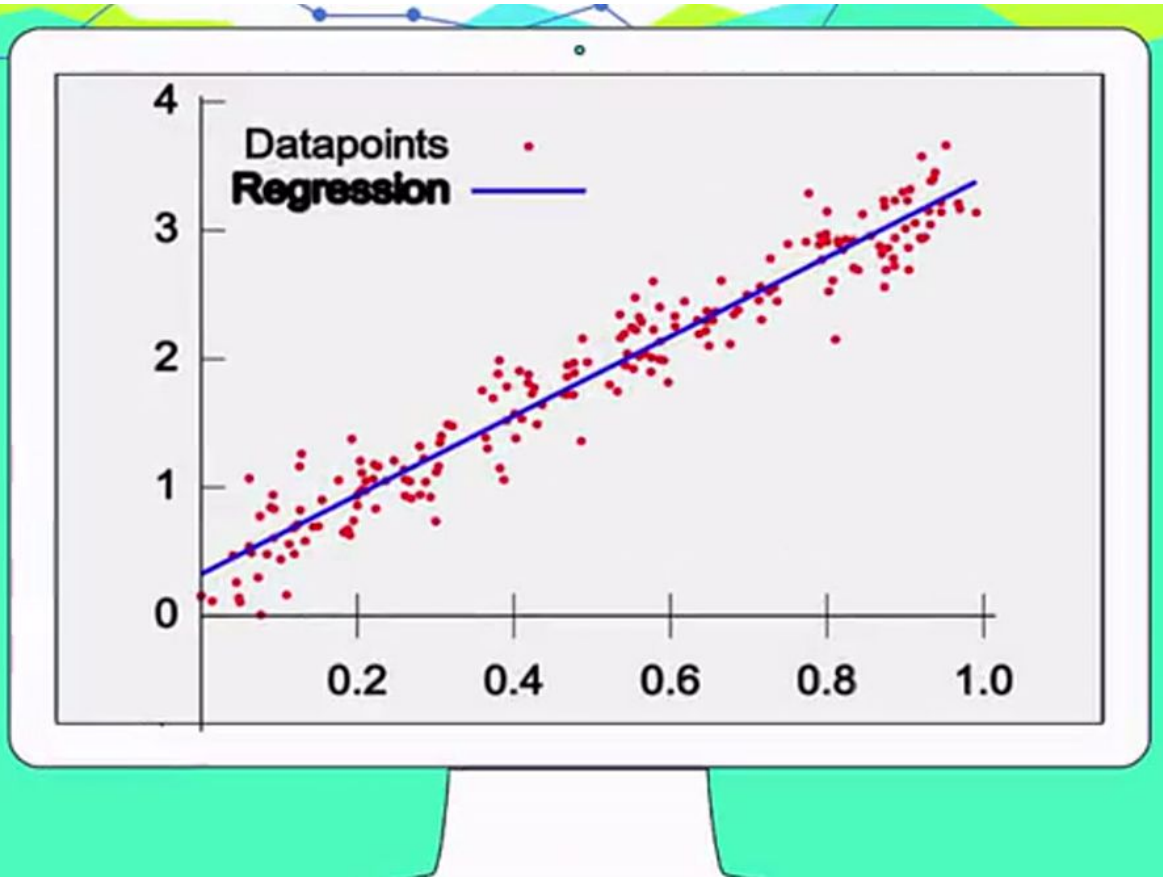
Normality

Residuals are normally distributed.



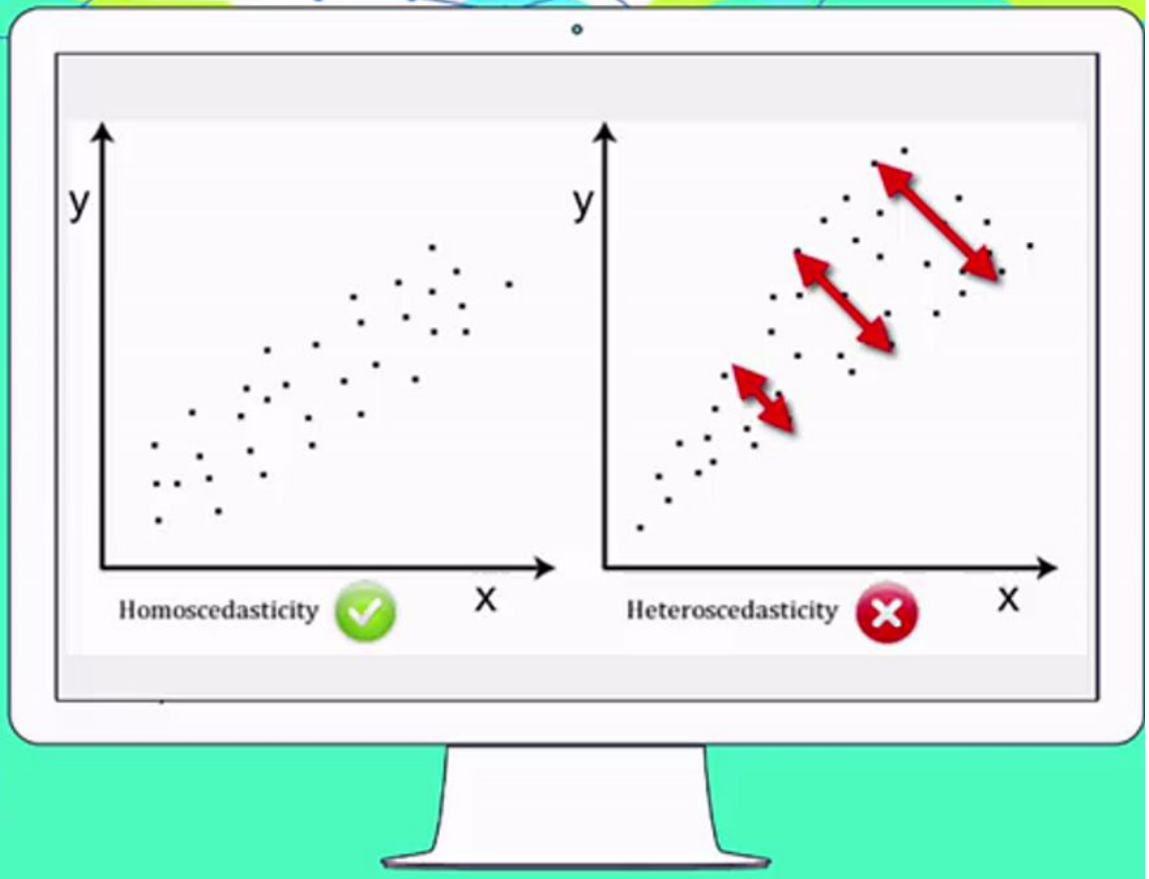
Linearty

Associations between explanatory variables and response variable are linear.



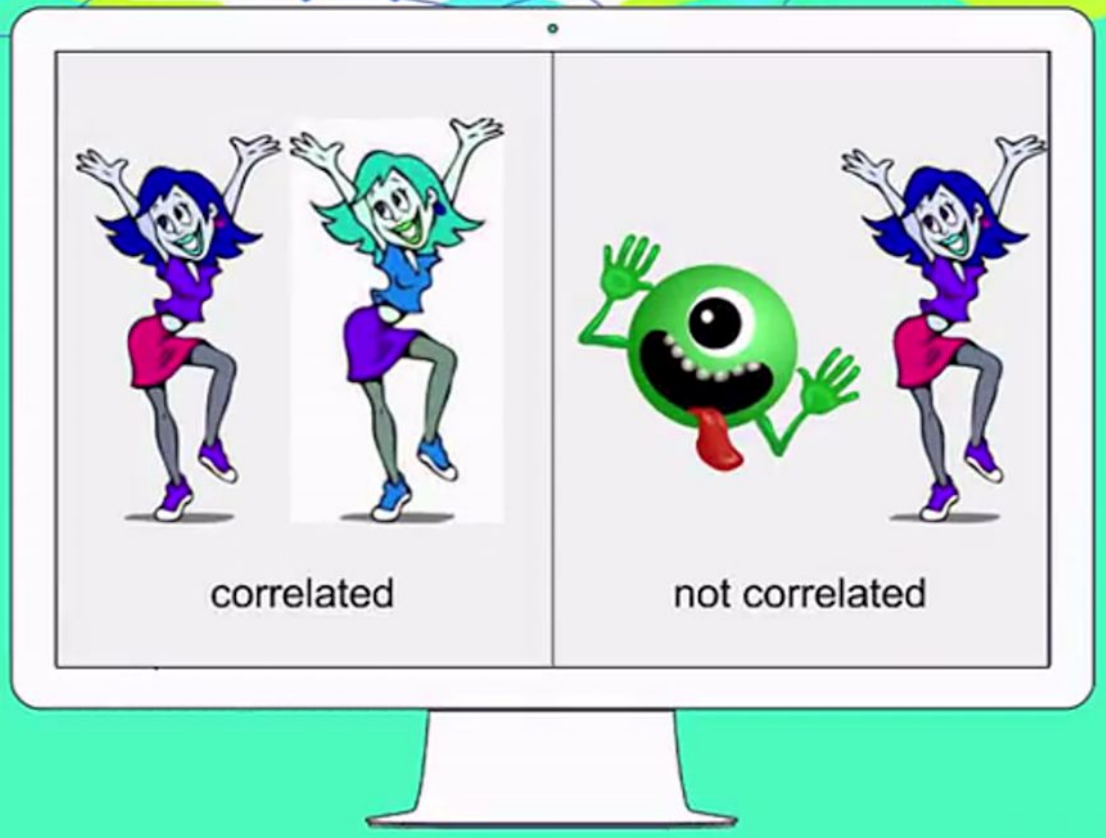
Homoscedasticity

The variability in the response variable is the same at all levels of the explanatory variable



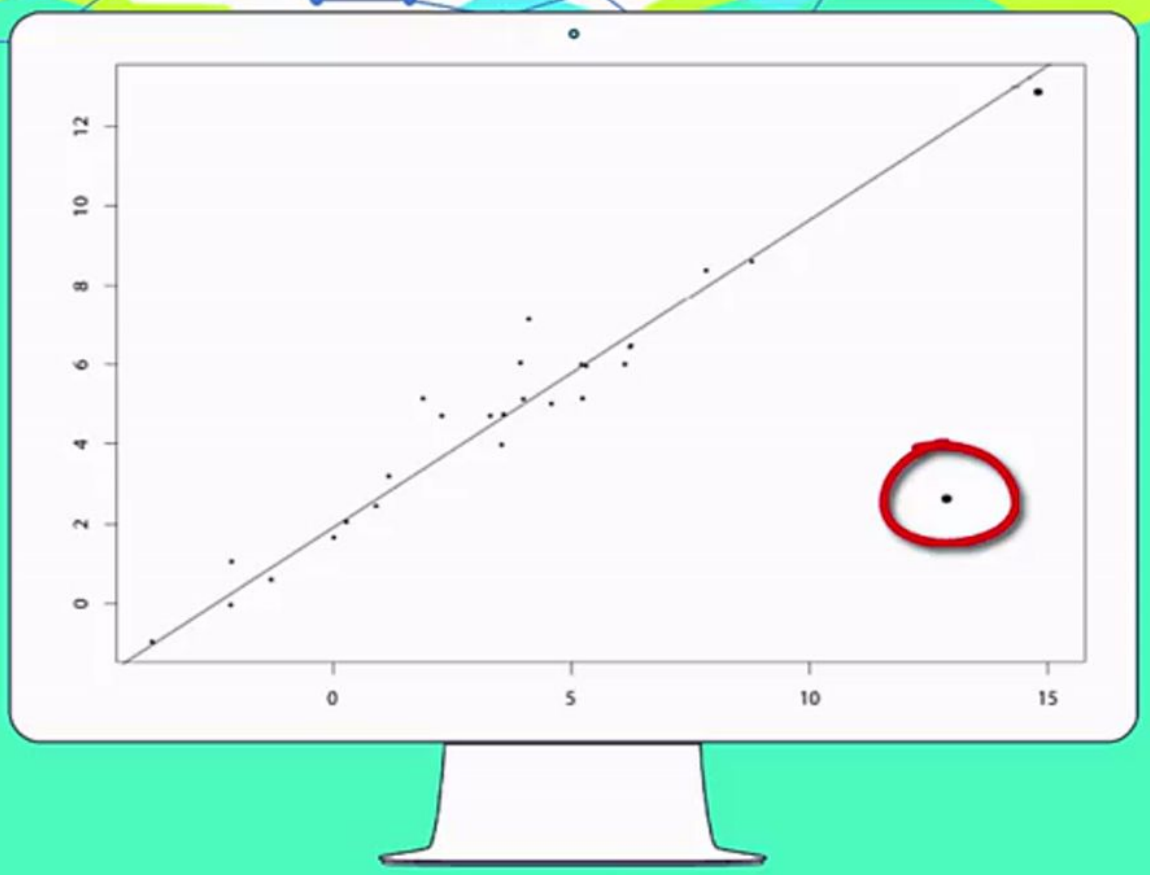
Independence

Observations are not correlated with each other.



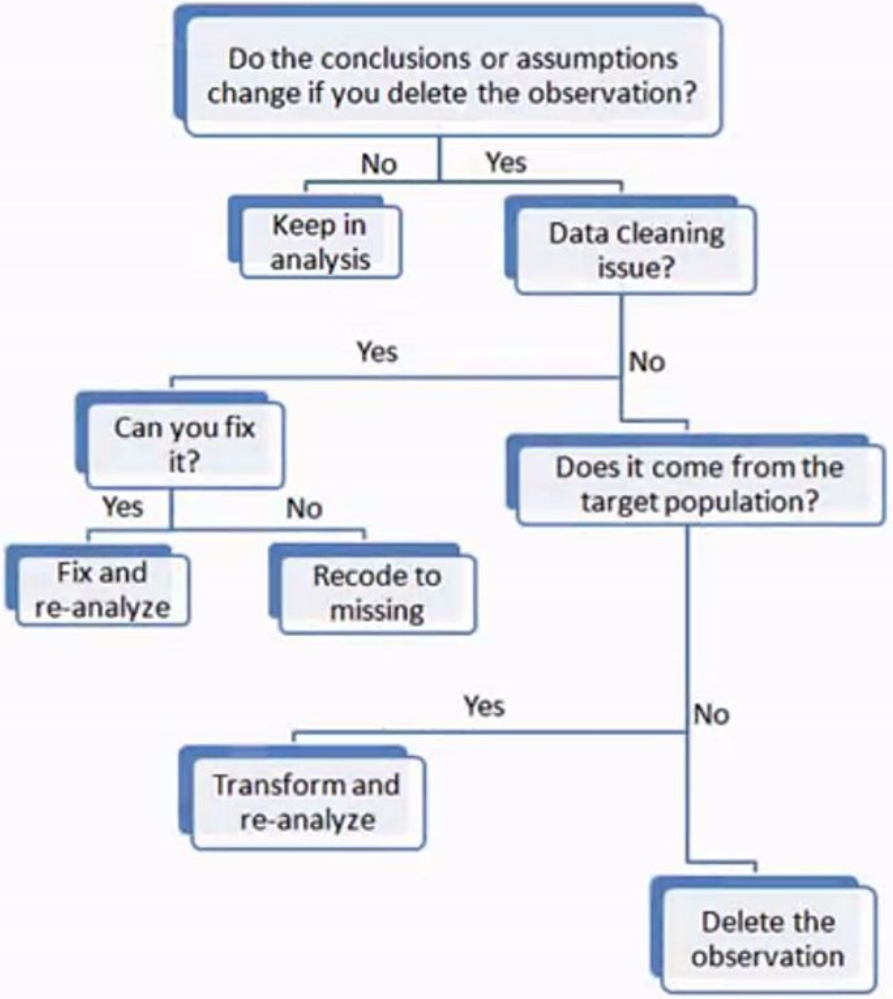
Outliers

One or more observations has extreme values on variable(s) relative to other observations.



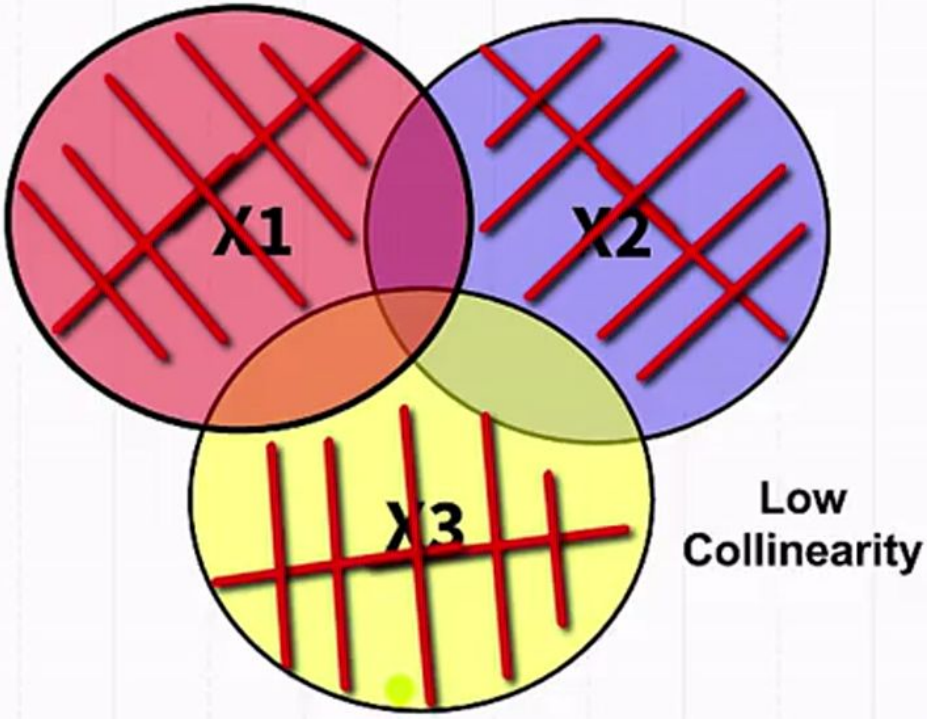
Q: WHAT DO YOU DO WITH OUTLIERS?

A: TRY USING THIS DECISION FLOWCHART

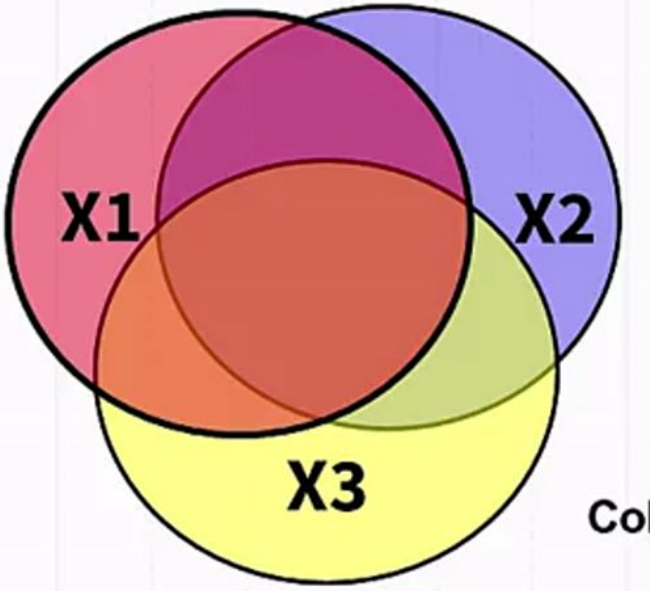


Multicollinearity

Explanatory variables are highly correlated with each other



Low Collinearity



High Collinearity

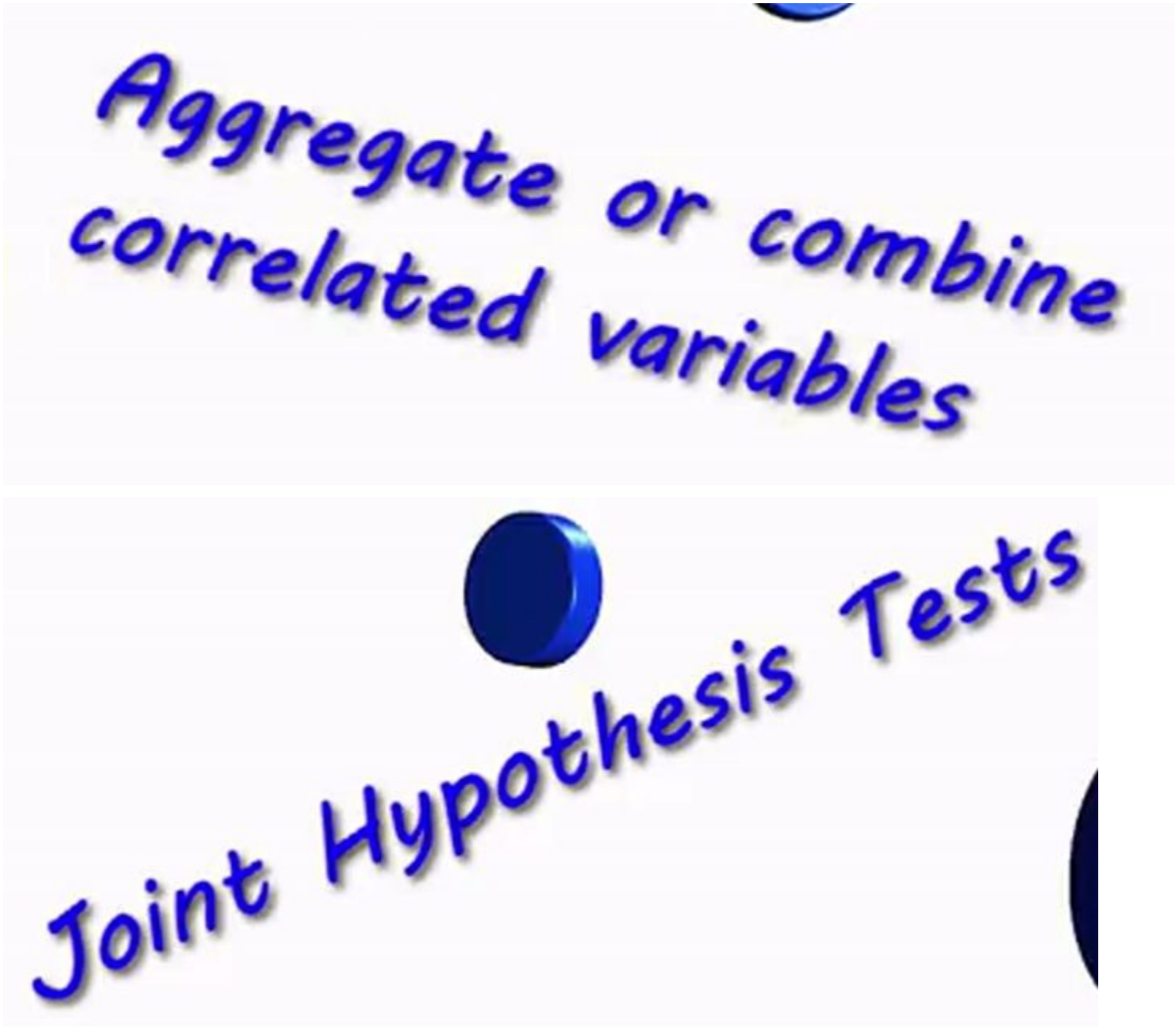
Multicollinearity

Signs:

1) highly associated explanatory variable not significant

2) negative regression coefficient that should be positive

3) taking out an explanatory variable drastically changes results


Aggregate or combine correlated variables
Joint Hypothesis Tests

Centering = subtracting the mean of a variable from the value of the variable

Mean = 0

Do not center the response variable

SAS-code-for-video-examples-_Gapminder_.sas

```
libname mydata "/courses/d1406ae5ba27fe300/c_3054" access=readonly;
data new; set mydata.gapminder;
run;
*************************************************************************************
BASIC LINEAR REGRESSION
*************************************************************************************;
* scatterplot with linear regression line;
proc sgplot;
 reg x=urbanrate y=internetuserate / lineattrs=(color=blue thickness=2);
 title "Scatterplot for the Association Between Urban Rate and Internet Use Rate";
 yaxis label="Female Employment Rate";
 xaxis label="Urbanization Rate";
run;
title;
* basic linear regression;
PROC glm;
model internetuserate=urbanrate/solution;
run;
```

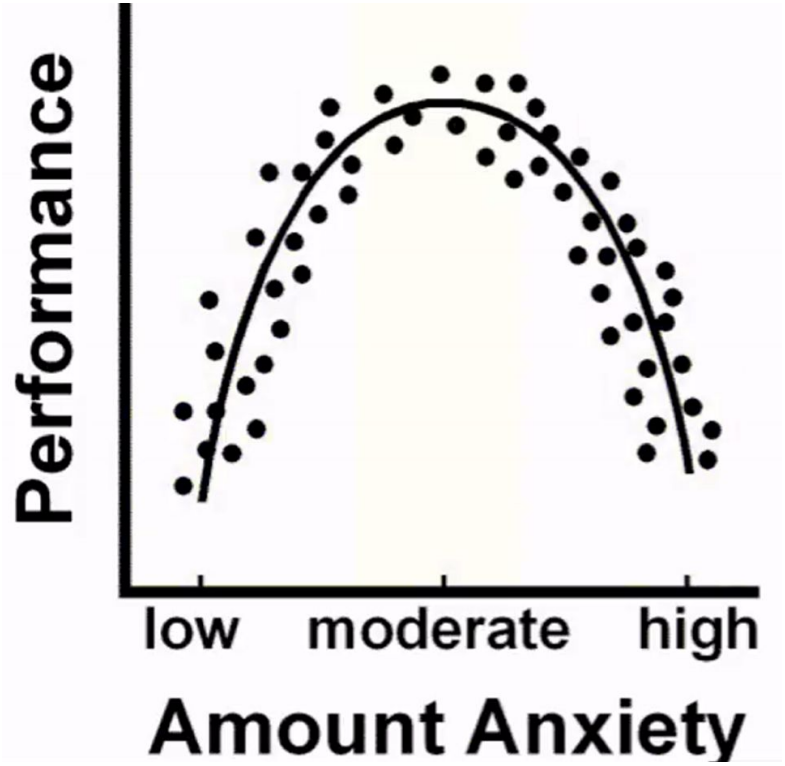
---
week 2.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
```

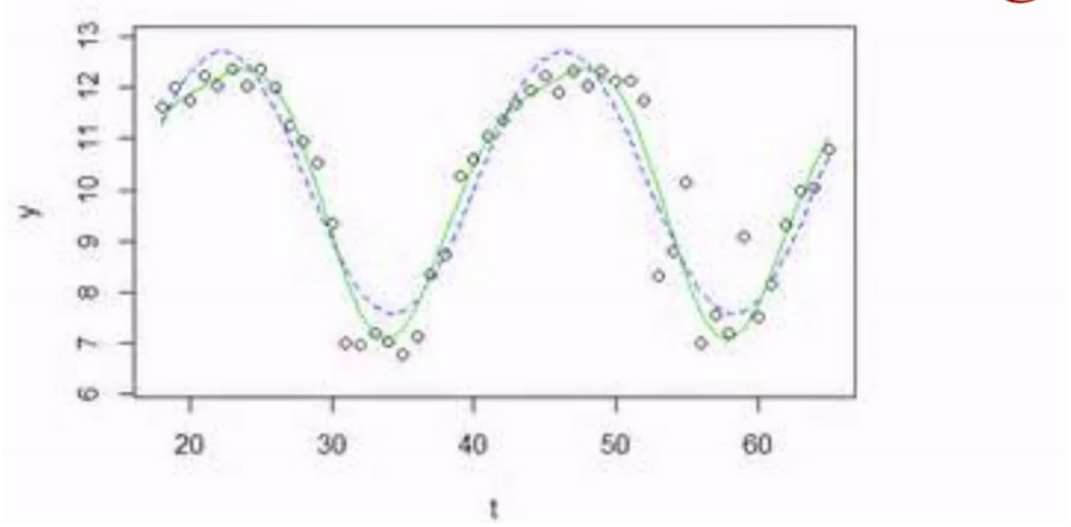
```
          OUT=murder REPLACE;

PROC IMPORT
          DATAFILE='/home/qallaf890/military expenditure.csv'
          OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
     create table mygapminder AS
     select          gapminder.*
                     ,surarea.surarea
                     ,pop.population
                     ,popden.popden
                     ,cpi.corruptionindex
                     ,hdi.hdi
                     ,murder.homicide
                     ,military.milexpprcntgdp
     from       work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
     set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
     COUNTRY='COUNTRY'
     INCOMEPERPERSON='GDP PER CAPITA'
```

```
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
/* TO FIND THE MEAN OF THE VARIABLES */
/*
proc univariate data=MYGAPMINDER;
run;
*/
PROC SQL;
CREATE TABLE WORK.MYGAPMINDER_2 AS
SELECT          *
                ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
```

```
                ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                ,(HIVRATE - 1.93544218) AS HIVRATE_2
                ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                ,(URBANRATE - 56.7693596) AS URBANRATE_2
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
FROM       WORK.MYGAPMINDER
;
QUIT;
/* TO CHECK THE CENTERING */
/*
proc univariate data=WORK.MYGAPMINDER;
run;
*/
PROC SORT; BY COUNTRY;
PROC GLM; MODEL CORRUPTIONINDEX = INCOMEPERPERSON_2;
```


Performance
low
moderate
high
Amount Anxiety

Overfitting



Bias-Variance Tradeoff

Specification: The process of developing a regression model

SAS-code-for-video-examples-_Gapminder_.sas

```
libname mydata "/courses/d1406ae5ba27fe300/c_3054" access=readonly;
data new; set mydata.gapminder;
run;
*************************************************************************************
POLYNOMIAL REGRESSION
*************************************************************************************;
* scatterplot with linear regression line femaleemployrate response variable;
proc sgplot;
 reg x=urbanrate y=femaleemployrate / lineattrs=(color=blue thickness=2) clm;
 yaxis label="Female Employment Rate";
 xaxis label="Urbanization Rate";
run;
* scatterplot with linear and quadratic regression line;
proc sgplot;
 reg x=urbanrate y=femaleemployrate / lineattrs=(color=blue thickness=2) degree=1 clm;
 reg x=urbanrate y=femaleemployrate / lineattrs=(color=green thickness=2) degree=2 clm;
 yaxis label="Female Employment Rate";
 xaxis label="Urbanization Rate";
run;
* centering quantitative explanatory variables;
data new2; set new;
if urbanrate ne . and femaleemployrate ne . and internetuserate ne .;
urbanrate_c=urbanrate-56.8410778;
internetuserate_c=internetuserate-34.2204688;
run;
proc means; var urbanrate internetuserate;
run;
* check coding;
```

```
proc means; var urbanrate_c internetuserate_c;
run;
* linear regression model;
PROC glm;
model femaleemployrate=urbanrate_c/solution clparm;
run;
* polynomial regression model;
PROC glm;
model femaleemployrate=urbanrate_c urbanrate_c*urbanrate_c/solution clparm;
run;
*********************************************************************************
EVALUATING MODEL FIT
*********************************************************************************;
* multiple regression adding internet use rate;
PROC glm;
model femaleemployrate=urbanrate_c urbanrate_c*urbanrate_c internetuserate_c/solution clparm;
run;
* request regression diagnostic plots;
PROC glm PLOTS(unpack)=all;
model femaleemployrate=urbanrate_c urbanrate_c*urbanrate_c
internetuserate_c/solution clparm;
output residual=res student=stdres out=results;
run;
* plot of standardized residuals for each observation;
proc gplot;
label stdres="Standardized Residual" country="Country";
plot stdres*country/vref=0;
run;
* using proc reg to get a partial regression plot;
* calculate quadratic terms;
data partial;
```

```
set new2;
urbanrate2=urbanrate_c*urbanrate_c;
run;
*partial regression plot;
PROC reg plots=partial;
model femaleemployrate=urbanrate urbanrate2
internetuserate/partial;
run;
```

---
week 3 - v2.sas

```
*************************************************************************************************
FETCHING DATA
*************************************************************************************************;
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp
      from       work.gapminder as gapminder
                       left join work.popden as popden on gapminder.country = popden.country
                       left join work.pop as pop on gapminder.country = pop.country
                       left join work.surarea as surarea on gapminder.country = surarea.country
                       left join work.cpi as cpi on gapminder.country = cpi.country
                       left join work.hdi as hdi on gapminder.country = hdi.country
                       left join work.murder as murder on gapminder.country = murder.country
                       left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
```

```
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
PROC SQL;
CREATE TABLE WORK.MYGAPMINDER_2 AS
SELECT          *
                ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
```

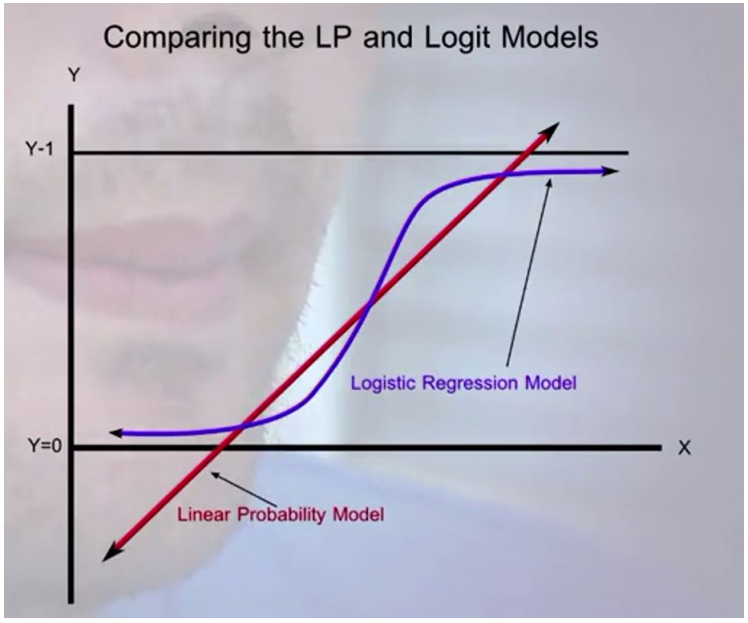
```
                ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                ,(HIVRATE - 1.93544218) AS HIVRATE_2
                ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                ,(URBANRATE - 56.7693596) AS URBANRATE_2
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
FROM       WORK.MYGAPMINDER
;
QUIT;
/*
PROC GPLOT;
PLOT  suicideper100th *    country    ;
PLOT  suicideper100th *    incomeperperson  ;
PLOT  suicideper100th *    alcconsumption   ;
PLOT  suicideper100th *    armedforcesrate  ;
PLOT  suicideper100th *    breastcancerper100th  ;
PLOT  suicideper100th *    co2emissions     ;
PLOT  suicideper100th *    femaleemployrate ;
PLOT  suicideper100th *    hivrate    ;
PLOT  suicideper100th *    internetuserate  ;
```

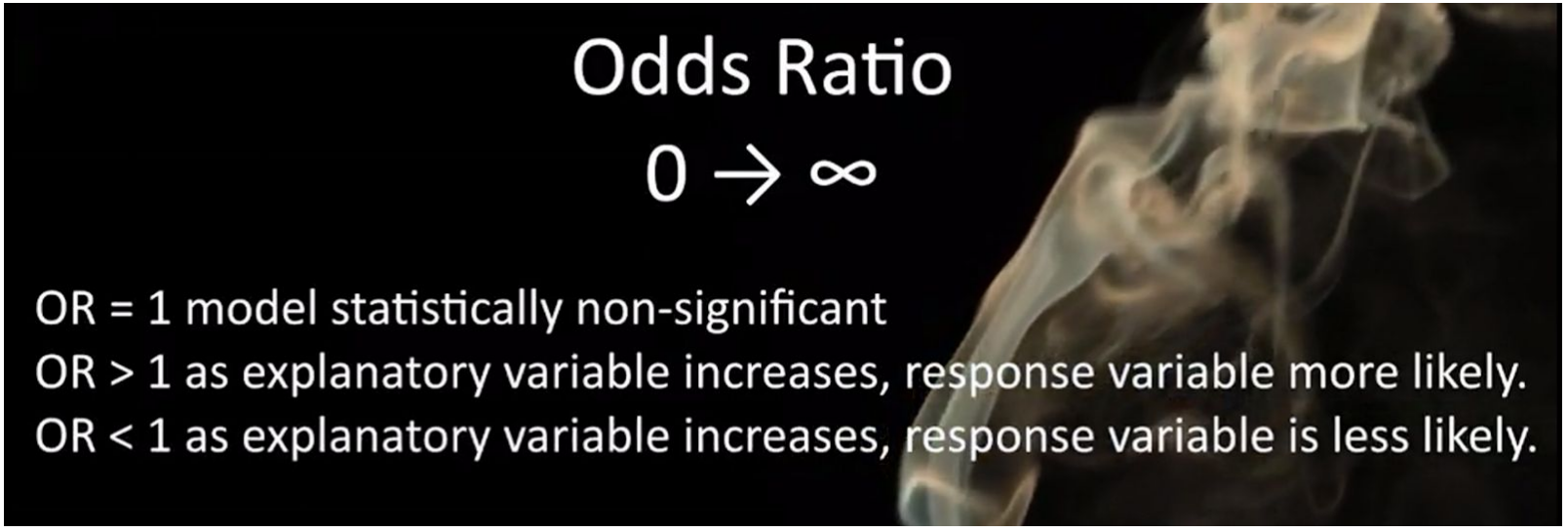
```
PLOT  suicideper100th *    lifeexpectancy   ;
PLOT  suicideper100th *    oilperperson     ;
PLOT  suicideper100th *    polityscore      ;
PLOT  suicideper100th *    relectricperperson    ;
PLOT  suicideper100th *    employrate ;
PLOT  suicideper100th *    urbanrate  ;
PLOT  suicideper100th *    surarea    ;
PLOT  suicideper100th *    population ;
PLOT  suicideper100th *    popden     ;
PLOT  suicideper100th *    corruptionindex  ;
PLOT  suicideper100th *    hdi   ;
PLOT  suicideper100th *    homicide   ;
PLOT  suicideper100th *    milexpprcntgdp   ;
RUN;
*/
* check coding;
/*
proc means; data WORK.MYGAPMINDER_2;
run;
*/
/*
data gapminder;
     set WORK.MYGAPMINDER_2;
run;
*/
*************************************************************************************
POLYNOMIAL REGRESSION
*************************************************************************************;
* scatterplot with linear and quadratic regression line;
proc sgplot;
 reg x=ALCCONSUMPTION y=SUICIDEPER100TH / lineattrs=(color=blue thickness=2) degree=1 clm;
```

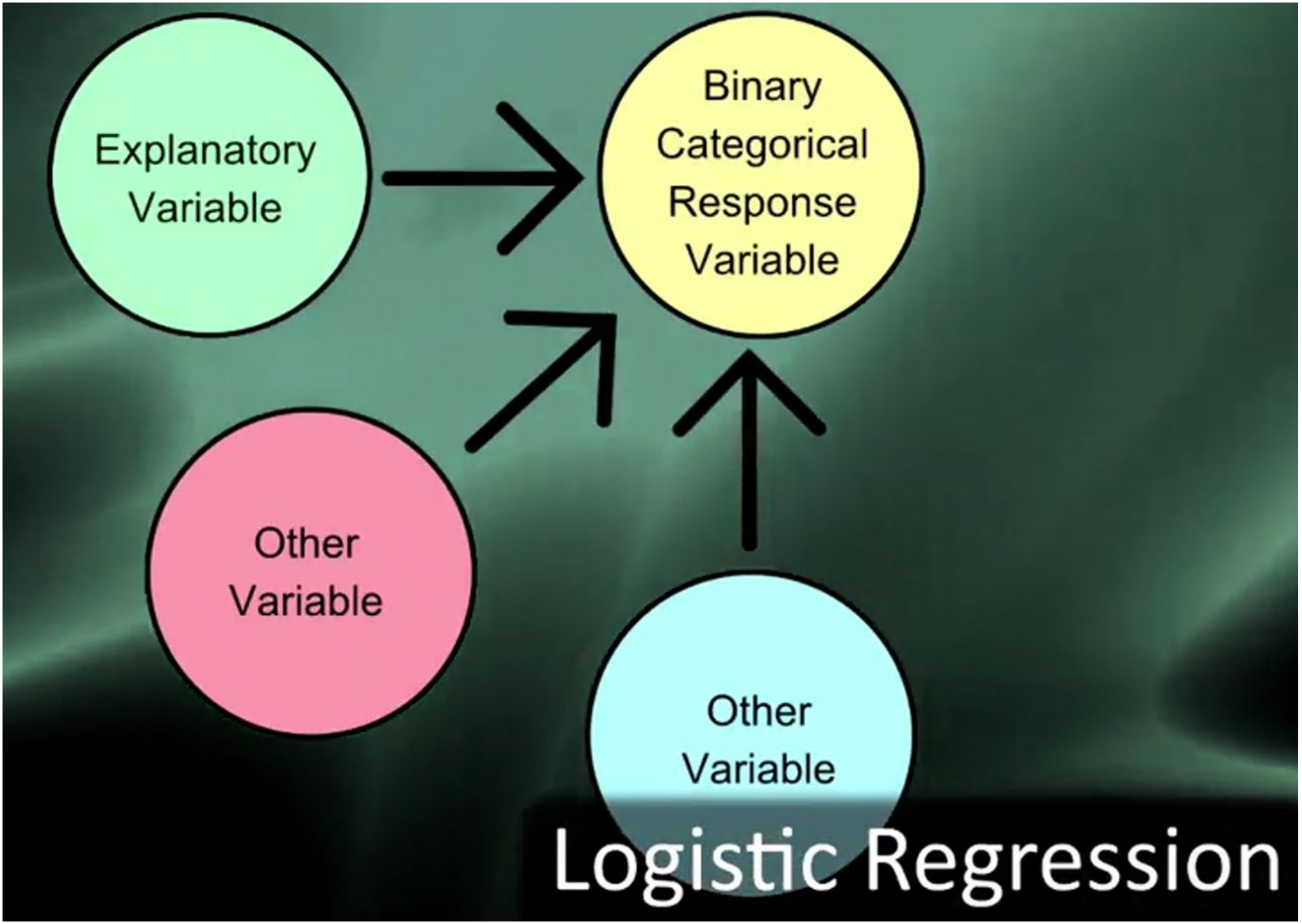
```
 reg x=ALCCONSUMPTION y=SUICIDEPER100TH / lineattrs=(color=green thickness=2) degree=2 clm;
 xaxis label='LITRES OF ALCOHOL CONSUMPTION';
 yaxis label='SUCIDE PER 100,000';
run;
* linear regression model;
PROC glm;
model SUICIDEPER100TH = ALCCONSUMPTION_2/solution clparm;
run;
* polynomial regression model;
PROC glm;
model SUICIDEPER100TH = ALCCONSUMPTION_2 ALCCONSUMPTION_2*ALCCONSUMPTION_2/solution clparm;
run;
*********************************************************************************
EVALUATING MODEL FIT
*********************************************************************************;
* multiple regression adding more variables;
PROC GLM;
MODEL SUICIDEPER100TH =
           ALCCONSUMPTION_2
           ALCCONSUMPTION_2*ALCCONSUMPTION_2
           URBANRATE_2
           HOMICIDE_2
           / SOLUTION CLPARM;
RUN;
* request regression diagnostic plots;
PROC glm PLOTS(unpack)=all;
MODEL SUICIDEPER100TH =
           ALCCONSUMPTION_2
           ALCCONSUMPTION_2*ALCCONSUMPTION_2
           URBANRATE_2
           HOMICIDE_2
```

```
          / SOLUTION CLPARM;
output residual=res student=stdres out=results;
run;
* plot of standardized residuals for each observation;
proc gplot;
label stdres="Standardized Residual" country="Country";
plot stdres*country/vref=0;
run;
* using proc reg to get a partial regression plot;
* calculate quadratic terms;
data partial;
set WORK.MYGAPMINDER_2;
ALCCONSUMPTION_sq = ALCCONSUMPTION_2*ALCCONSUMPTION_2
;
run;
*partial regression plot;
PROC reg plots=partial;
MODEL SUICIDEPER100TH =
          ALCCONSUMPTION_2
          ALCCONSUMPTION_sq
          URBANRATE_2
          HOMICIDE_2
/partial;
run;
```

Week 4: Logistic Regression






Explanatory Variable
Binary Categorical Response Variable
Other Variable
Other Variable
Logistic Regression

SAS-code-for-video-examples-_NESARC_.sas

```
libname mydata "/courses/d1406ae5ba27fe300/c_3054" access=readonly;
*************************************************************************************************
DATA MANAGEMENT
*************************************************************************************************;
data new; set mydata.NESARC_PDS;
if s3aq3c1=99 then s3aq3c1=.;
if tab12mdx=1 then nicotinedep=1; else nicotinedep=0;
rename s3aq3c1=numbercigsmoked;
/*********************************************/
/**********MORE COMPLICATED AGGREGATION********/
/*********************************************/
array one(*)
s3aq3b1 ever_daily s3aq51 s3aq8b12
S3AQ8B11 S3AQ8B7A S3AQ8B7B S3AQ8B7C S3AQ8B7D S3AQ8B7E S3AQ8B7F S3AQ8B7G S3AQ8B7H
S3AQ8B7J S3AQ8B13 S3AQ8B6 S3AQ8B1 S3AQ8B5 S3AQ8B2 S3AQ8B3 S3AQ8B4 S3AQ8B14;
do I=1 to dim(one);
if one(I) eq 9 then one(I)=.;
end;
/*array two(*)
s6q1 s6q2 s6q3 s6q61 s6q62 s6q63 s6q64 s6q65 s6q66 s6q67 s6q68 s6q69 s6q610 s6q611
s6q612 s6q613 s6q7;
do II=1 to dim (two);
if two(II) eq 9 then two(II)=.;
end;*/
/*qualifying panic aggregate variable*/
if s6q1=1 and s6q2=1 or s6q2=1 and s6q3=1 or s6q3=1 and s6q61=1 or s6q61=1 and s6q62=1 or s6q62=1
and s6q63=1 or s6q63=1 and s6q64=1 or s6q64=1 and s6q65=1 or s6q65=1 and s6q66=1 or s6q66=1
and s6q67=1 or s6q67=1 and s6q68=1 or s6q68=1 and s6q69=1 or s6q69=1 and s6q610=1 or s6q610=1
```

```
and s6q611=1 or s6q611=1 and s6q612=1 or s6q612=1 and s6q613=1 or s6q613=1 and s6q7=1
or s6q7=1 then panic=1; else panic=0;
if check321=1; /*SMOKED IN THE PAST YEAR*/
if s3aq3b1=1; /*USUALLY SMOKED DAILY IN THE PAST YEAR*/
if age le 25; /*AGE 18 TO 25*/
if s6q1=9 then s6q1=.;
if s6q2=9 then s6q2=.;
if s6q3=9 then s6q3=.;
if s6q61=9 then s6q61=.;
if s6q62=9 then s6q62=.;
if s6q63=9 then s6q63=.;
if s6q64=9 then s6q64=.;
if s6q65=9 then s6q65=.;
if s6q66=9 then s6q66=.;
if s6q67=9 then s6q67=.;
if s6q68=9 then s6q68=.;
if s6q69=9 then s6q69=.;
if s6q610=9 then s6q610=.;
if s6q611=9 then s6q611=.;
if s6q612=9 then s6q612=.;
if s6q613=9 then s6q613=.;
if s6q7=9 then s6q7=.;
/*Current Tolerance criteria #1 DSM-IV*/
if S3AQ8B11 eq 1 or S3AQ8B12 eq 1 then ctobcrit1=1;
else if S3AQ8B11 eq 2 and S3AQ8B12 eq 2 then ctobcrit1=0;
/*Current 8 WITHDRAWAL SUB-SYMPTOMS IN DSM-IV*/
CWITHDR_count=0;
     if S3AQ8B7A eq 1 then CWITHDR_count=CWITHDR_count+1; /*depressed mood*/
     if S3AQ8B7B eq 1 then CWITHDR_count=CWITHDR_count+1; /*insomnia*/
     if S3AQ8B7C eq 1 then CWITHDR_count=CWITHDR_count+1; /*difficulty concentrating*/
     if S3AQ8B7D eq 1 then CWITHDR_count=CWITHDR_count+1; /*increased appetite or weight gain*/
```

```
   if S3AQ8B7E eq 1 then CWITHDR_count=CWITHDR_count+1; /*irritability, anger and frustration*/
   if S3AQ8B7F eq 1 then CWITHDR_count=CWITHDR_count+1; /*anxiety*/
     if S3AQ8B7G eq 1 then CWITHDR_count=CWITHDR_count+1; /*anxiety*/
   if S3AQ8B7H eq 1 then CWITHDR_count=CWITHDR_count+1; /*restlessness*/

/*Current Withdrawal criteria #2 DSM-IV*/
if CWITHDR_count>=4 or S3AQ8B7J=1 then Ctobcrit2=1;
else if CWITHDR_count lt 4 and S3AQ8B7J=2 then Ctobcrit2=0;
/*Current Larger amount or longer period criteria #3 DSM-IV*/
if S3AQ8B13 eq 1 then Ctobcrit3=1;
else if S3AQ8B13 eq 2 then Ctobcrit3=0;
/*Current Cut down criteria #4 DSM-IV*/
if S3AQ8B6 eq 1 or S3AQ8B1 eq 1 then Ctobcrit4=1;
else if S3AQ8B6 eq 2 and S3AQ8B1 eq 2 then Ctobcrit4=0;
/*Current Substance activities criteria #5 DSM-IV*/
if S3AQ8B5 eq 1 then Ctobcrit5=1;
else if S3AQ8B5 eq 2 then Ctobcrit5=0;
/*Current Reduce activities criteria #6 DSM-IV*/
if S3AQ8B2 eq 1 or S3AQ8B3 eq 1 then Ctobcrit6=1;
else if S3AQ8B2 eq 2 and S3AQ8B3 eq 2 then Ctobcrit6=0;
/*Current use continued despite knowledge of physical or psychological problem criteria #7 DSM-IV*/
if S3AQ8B4=1 or S3AQ8B14=1 then Ctobcrit7=1;
else if S3AQ8B4=2 and S3AQ8B14=2 then Ctobcrit7=0;
/*CURRENT DSM-IV NICOTINE DEPENDENCE SYMPTOM COUNT*/
NDSYMPTOMS=sum (of Ctobcrit1 Ctobcrit2 Ctobcrit3 Ctobcrit4
       Ctobcrit5 Ctobcrit6 Ctobcrit7);
* ETHNICITY/RACE VARIABLE;
if ethrace2a=5 then ethrace=0; * hispanic;
if ethrace2a=1 then ethrace=1; * white;
if ethrace2a=2 then ethrace=2; * black;
if (ethrace2a=3 or ethrace2a=4) then ethrace=3; * other;
```

```
run;
*******************************************************************************************
END DATA MANAGEMENT
*******************************************************************************************;
*******************************************************************************************
CATEGORICAL EXPLANATORY VARIABLES RE-VISITED (3+ CATEGORIES)
*******************************************************************************************;
* centering quantitative number of cigarettes smoked (also age for later regression);
* print mean;
PROC MEANS;
var numbercigsmoked age;
run;
* centering (subtract mean);
data new2;
set new;
numbercigsmoked_c = numbercigsmoked - 13.3642586;
age_c = age - 21.6053030;
run;
* check coding;
PROC MEANS;
var numbercigsmoked_c age_c;
run;
*adding 4 category race-ethicity variable;
PROC GLM;
class ethrace (ref="0");
model NDSymptoms=dyslife majordeplife numbercigsmoked_c age_c SEX
ethrace/solution;
run;
* change the reference group to non-Hispanic White;
PROC GLM;
class ethrace (ref="1");
```

```
model NDSymptoms=dyslife majordeplife numbercigsmoked_c age_c SEX
ethrace/solution;
run;
*******************************************************************************************
LOGISTIC REGRESSION
*******************************************************************************************;
Proc logistic descending; model nicotinedep=SOCPDLIFE;
run;
* adding depression;
Proc logistic descending; model nicotinedep=SOCPDLIFE MAJORDEPLIFE;
run;
```

---
week 4.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp

                            ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                  ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                  ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                  ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                  ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                  ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                  ,(HIVRATE - 1.93544218) AS HIVRATE_2
                  ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                  ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                  ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                  ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                  ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                  ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                  ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                  ,(URBANRATE - 56.7693596) AS URBANRATE_2
```

```
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
      from      work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
```

```
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
keep
      COUNTRY
      EMPLOYRATE
      EMPLOYRATE_RANK
      INCOMEPERPERSON
      INCOMEPERPERSON_RANK
      ARMEDFORCESRATE
      ARMEDFORCESRATE_RANK
      LIFEEXPECTANCY
      LIFEEXPECTANCY_RANK
      SUICIDEPER100TH
      SUICIDEPER100TH_RANK
      URBANRATE
      URBANRATE_RANK
      surarea
      surarea_RANK
      population
      population_RANK
      popden
```

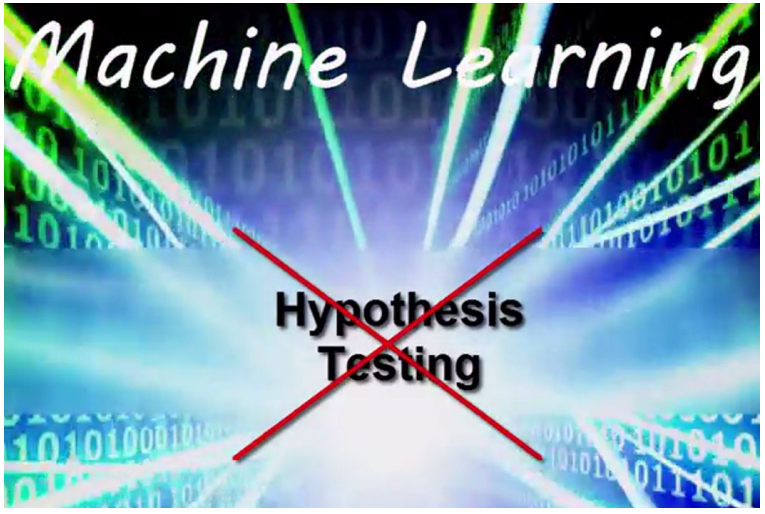
```
      popden_RANK
      corruptionindex
      corruptionindex_RANK
      hdi
      hdi_RANK
      homicide
      homicide_RANK
      milexpprcntgdp
      milexpprcntgdp_RANK
      ALCCONSUMPTION
      ;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF SUICIDEPER100TH <  16  THEN SUICIDEPER100TH_RANK = 0;
IF SUICIDEPER100TH >= 16 THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
/* UNIVARIATE GRAPH */
PROC GCHART; VBAR SUICIDEPER100TH;
RUN;
PROC FREQ;
TABLES
      SUICIDEPER100TH;
RUN;
*******************************************************************************************
LOGISTIC REGRESSION
*******************************************************************************************;
PROC LOGISTIC DESCENDING; MODEL SUICIDEPER100TH_RANK=ALCCONSUMPTION;
RUN;
```

4. Machine Learning for Data Analysis


Machine Learning
Describe Associations
Search For Patterns
Make Predictions

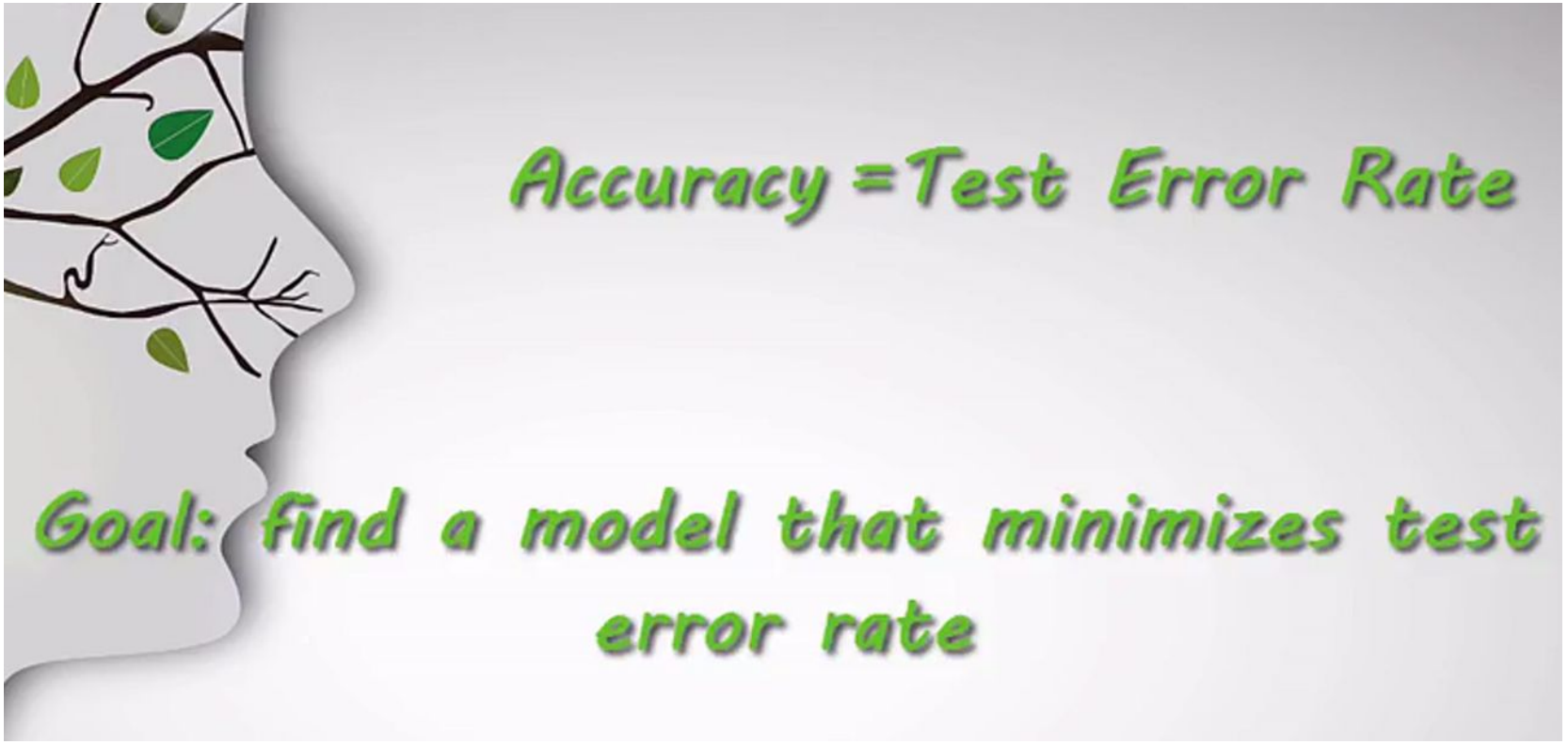

Machine Learning
Unsupervised Learning


Machine Learning
Supervised Learning


Machine Learning
Hypothesis
Testing


Machine Learning
Training Set


Machine Learning
Test Set


Accuracy = Test Error Rate
Goal: find a model that minimizes test error rate

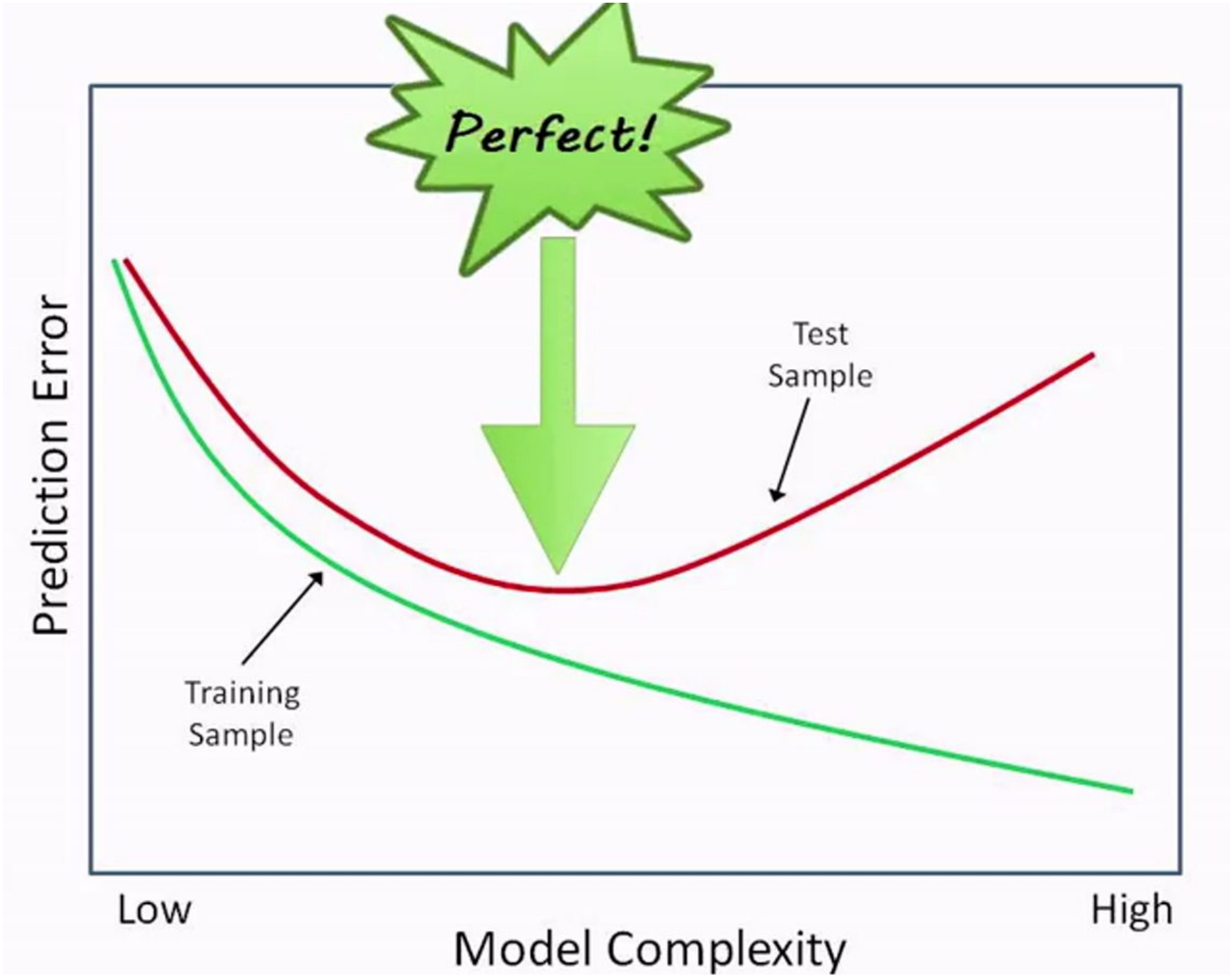
Linear Regression

Accuracy=mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

Variance = change in parameter estimates across different data sets

Bias = how far off model estimated values are from true values


Perfect!
Test
Sample
Training
Sample
Prediction Error
Low
High
Model Complexity

Confusion Matrix

Test Sample Nicotine Dependence Classification

Predicted \ Actual	Yes	No	Total
Yes	184	91	275
No	32	685	717
Total	216	776	992

Test error rate = % misclassified = 12%

Growing the Tree

Binary splits maximize correct classification

All cut-points are tested

Subgroups showing similar outcomes are generated

Validating the Tree

Cross-validation guards againts overfit

A random subset is tested and only "branches" that improve the classification are retained

Selected sub-tree is the lowest probability of misclassification

Grow Statement

Start with a large tree based on the randomly selected training sample

Specify criterion using the GROW statement to minimize the node's error

Entropy is the most common choice when growing a classification tree

The growth process continues until the tree reaches a maximum depth of 10

Prune Statement

The PRUNE statement specifies the method for pruning a large tree into a smaller subtree

The most common method is pruning through Cost-complexity

Cost-complexity requested for a target variable by specifing "Prune Cost-complexity"

This algorithm makes a trade-off between complexity and error rate

Strengths of Decision Trees

Selects from a large number of variables and interactions that are most important

Easy to interpret and visualize

Can handle large data sets and predict both binary categorical target variables and quantitative target variables

Limitations Decision Trees

Small changes in the data can lead to different splits that undermines the interpretability of the model

Decision trees are not very reproducible on future data

Decision-Tree-Program.sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.treeaddhealth;
PROC SORT; BY AID;
ods graphics on;
proc hpsplit seed=15531;
class TREG1 BIO_SEX HISPANIC WHITE BLACK NAMERICAN ASIAN
  alcevr1 marever1 cocever1 inhever1 Cigavail EXPEL1 ;
model TREG1 =AGE BIO_SEX HISPANIC WHITE BLACK NAMERICAN ASIAN alcevr1 ALCPROBS1
 marever1 cocever1 inhever1 DEVIANT1 VIOL1 DEP1 ESTEEM1 PARPRES PARACTV
 FAMCONCT schconn1 Cigavail PASSIST EXPEL1 GPA1;
grow entropy;
prune costcomplexity;

RUN;
```

---
week 1.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp

                            ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                  ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                  ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                  ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                  ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                  ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                  ,(HIVRATE - 1.93544218) AS HIVRATE_2
                  ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                  ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                  ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                  ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                  ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                  ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                  ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                  ,(URBANRATE - 56.7693596) AS URBANRATE_2
```

```
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
      from      work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
```

```
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
/* DATA MANAGEMENT STEP */
IF SUICIDEPER100TH <  16  THEN SUICIDEPER100TH_RANK = 2;
IF SUICIDEPER100TH >= 16 THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
```

```
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = .;
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
```

```
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;
*******************************************************************************************
DECISION TREE
*******************************************************************************************;
PROC SORT; BY COUNTRY;
ods graphics on;
proc hpsplit seed = 12345;
class SUICIDEPER100TH_RANK
EMPLOYRATE_RANK
INCOMEPERPERSON_RANK
ARMEDFORCESRATE_RANK
LIFEEXPECTANCY_RANK
URBANRATE_RANK
surarea_RANK
population_RANK
popden_RANK
corruptionindex_RANK
hdi_RANK
;
model SUICIDEPER100TH_RANK =
ALCCONSUMPTION
EMPLOYRATE_RANK
INCOMEPERPERSON_RANK
ARMEDFORCESRATE_RANK
LIFEEXPECTANCY_RANK
URBANRATE_RANK
surarea_RANK
```

```
population_RANK
popden_RANK
corruptionindex_RANK
hdi_RANK
;
grow entropy;
prune costcomplexity;

RUN;
```

Random Forests

Splits on only ONE variable in a node

Variable with largest association with Target among candidate explanatory variables

ONLY among those variables that were randomly selected

Random Forests

First a subset of explanatory variables is selected at random

Next the node is split with the BEST variable of the subset

After this node is split, a new list of eligible variables is selected at random

Random Forests

This continues until the tree is fully grown

Ideally, there will be only one observation in each termnial node

Random Forests

Eligible variable set will be different
from node to node

Important variables eventually make it to the tree

Their relative success in predicting the
target variable will get them more "votes"

Each Tree is Grown by:

A subset of the explanatory variables at each node

AND

A random subset of the sample for each tree in the forest

Bagged and Unbagged Data

Bagged

Out of Bag

Important!

Trees generated are not themselves interpreted

They are used collectively to rank the importance of variables in predicting the target of interest

Like Decision Trees

Random forests are a data mining algorithm that can select important variables

The target variable can be categorical or quantitative

The explanatory variables can be a combination of categorical and quantitative

Unlike Decision Trees

The results of random forests generalize well to new data since the strongest signals emerge by growing many trees

Small changes in data do NOT impact te results of random forests

Weakness

The results are less satisfying because trees are not interpreted

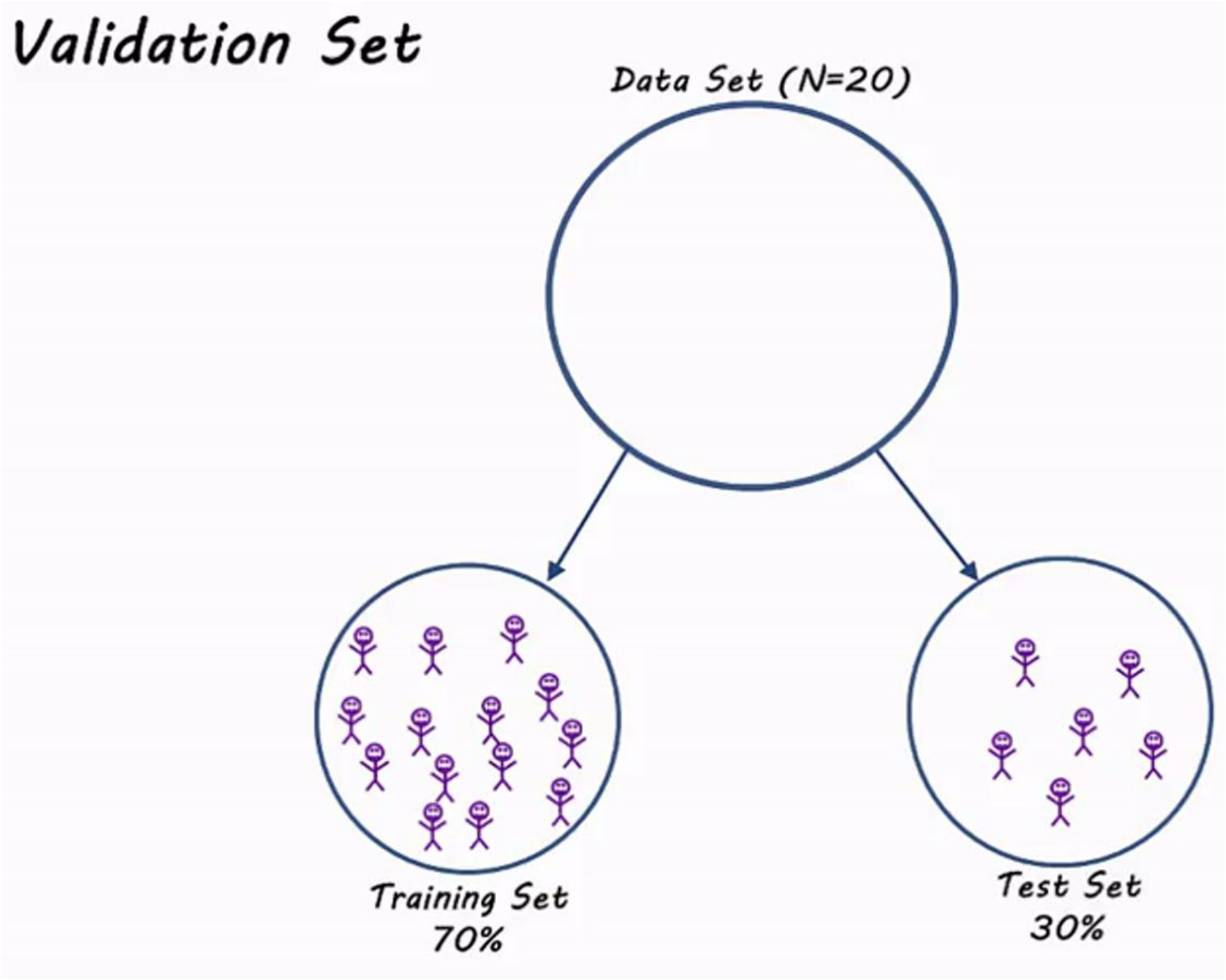
The forest is used to rank the importance of variables in predicting a target


Validation and Cross
Training set model estima


Validation and Cross Validation
How do we know which model is
the best model??


Validation and Cross Validation
Need to estimate test error

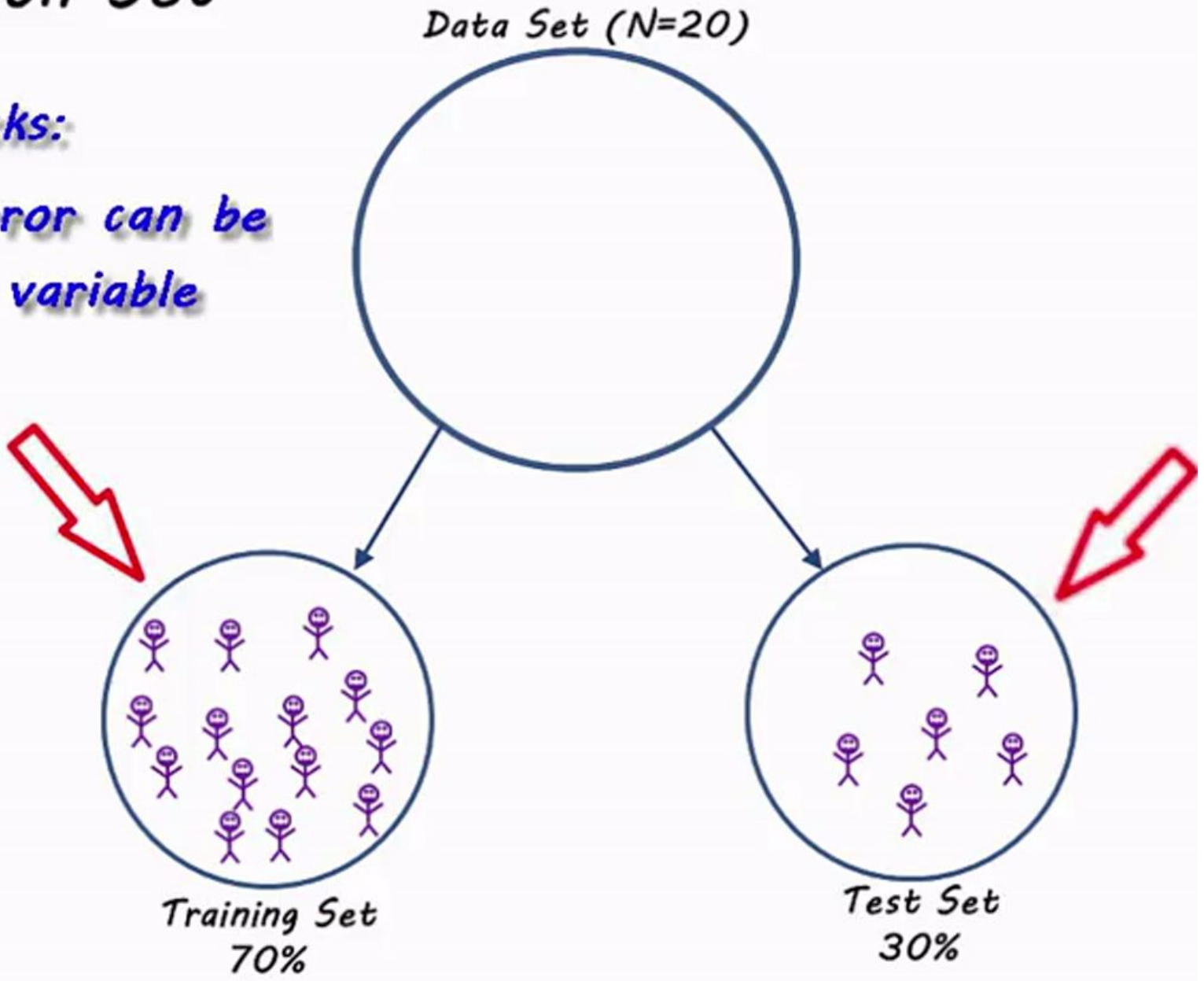
Validation Set



Validation Set

Drawbacks:

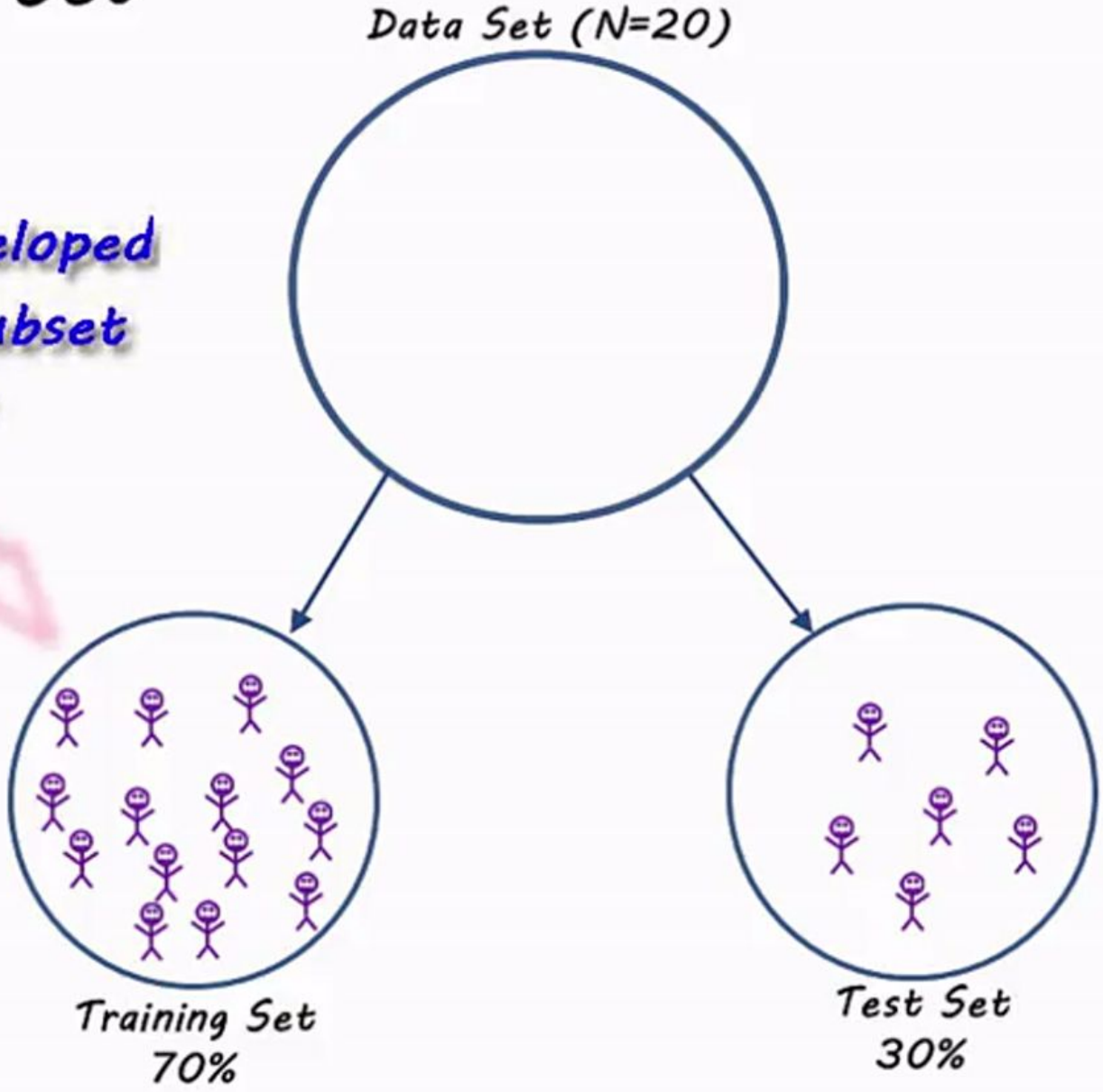
- Test error can be highly variable



Validation Set

Drawbacks:

- Model developed on only a subset of the data



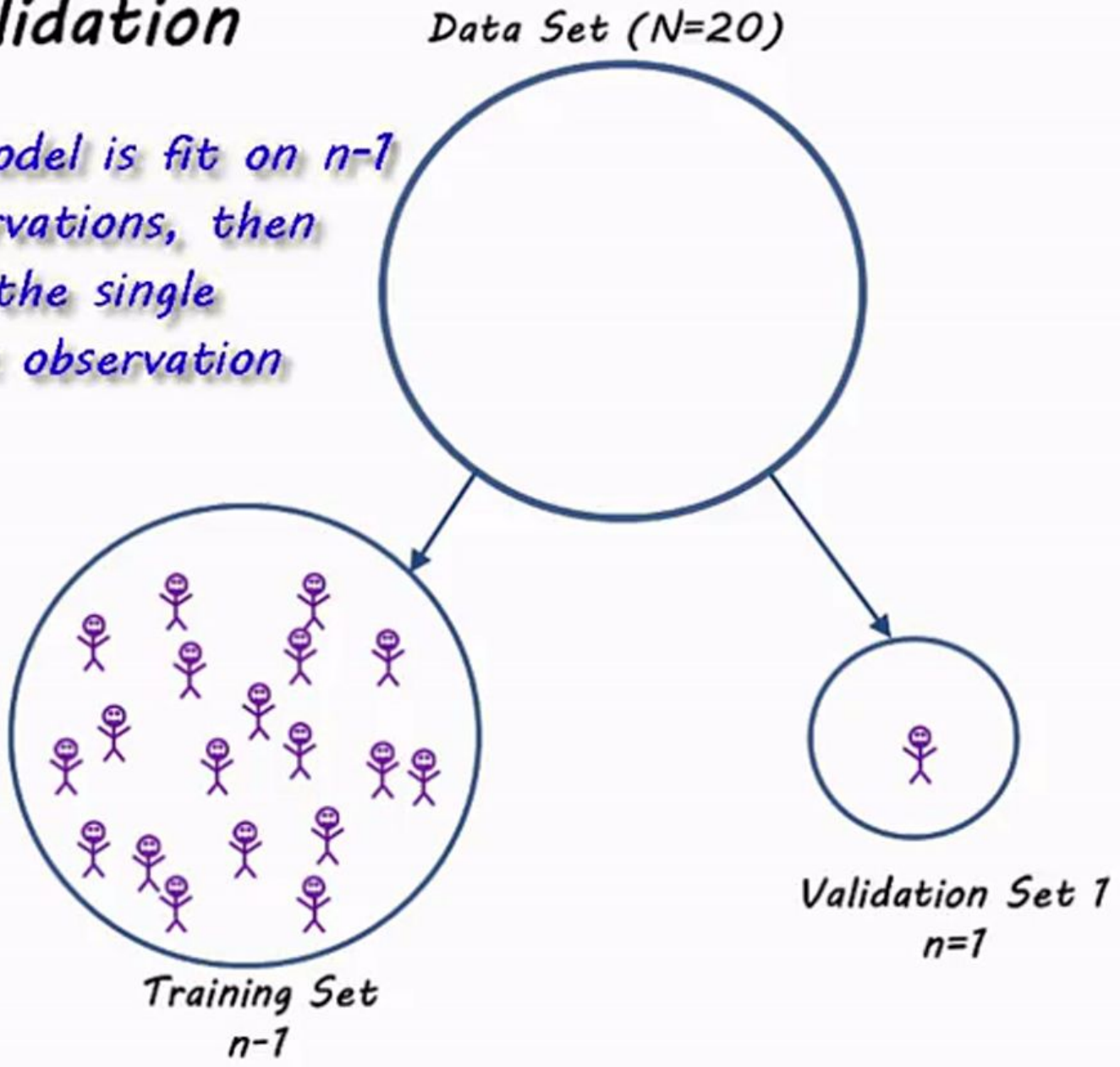
Cross Validation

- Goal is to define a data set to "test" the model during the training phase

- Validation Set

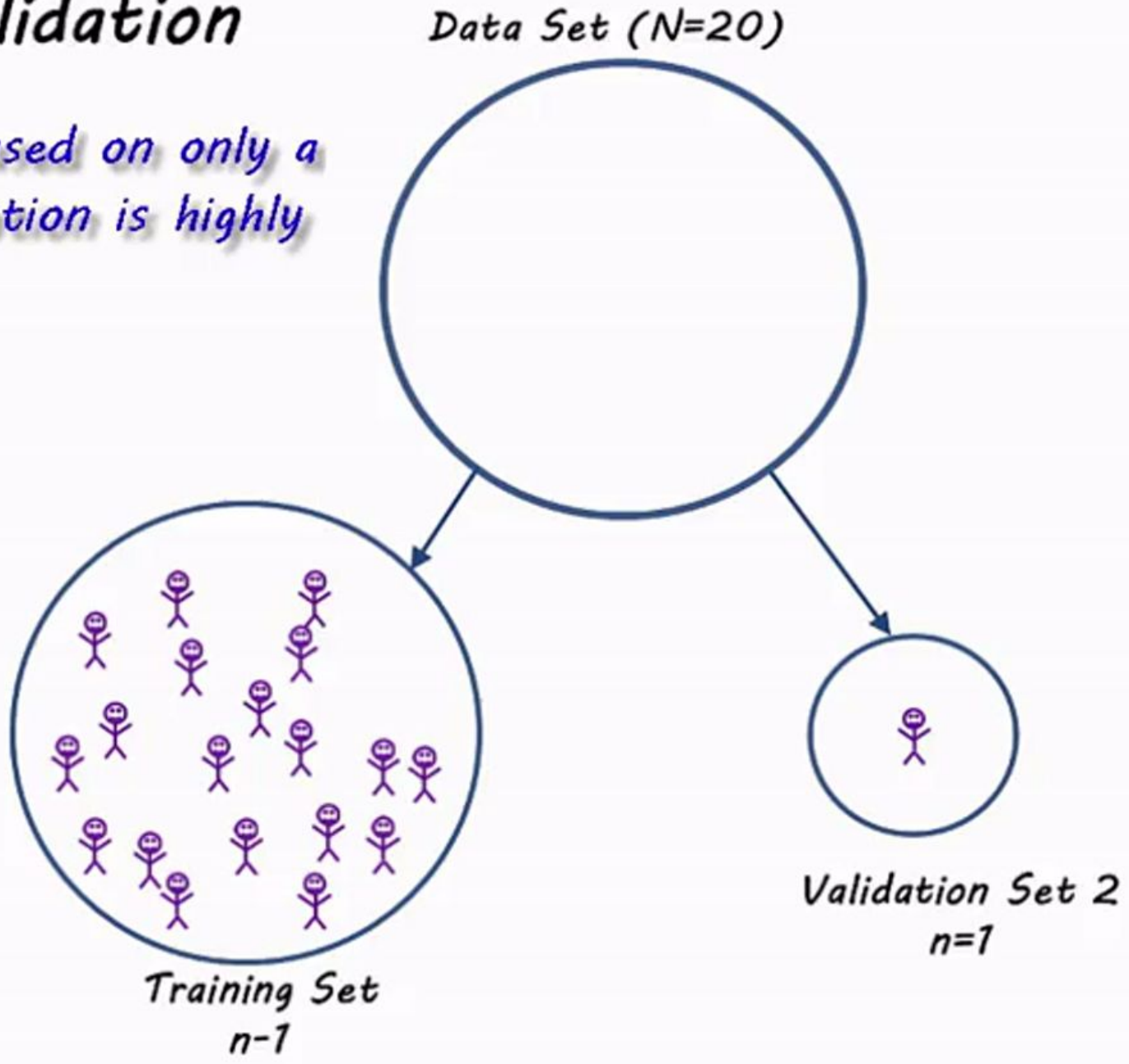
Leave One Out Cross Validation

- Statistical model is fit on n-1 training observations, then validated on the single validation set observation



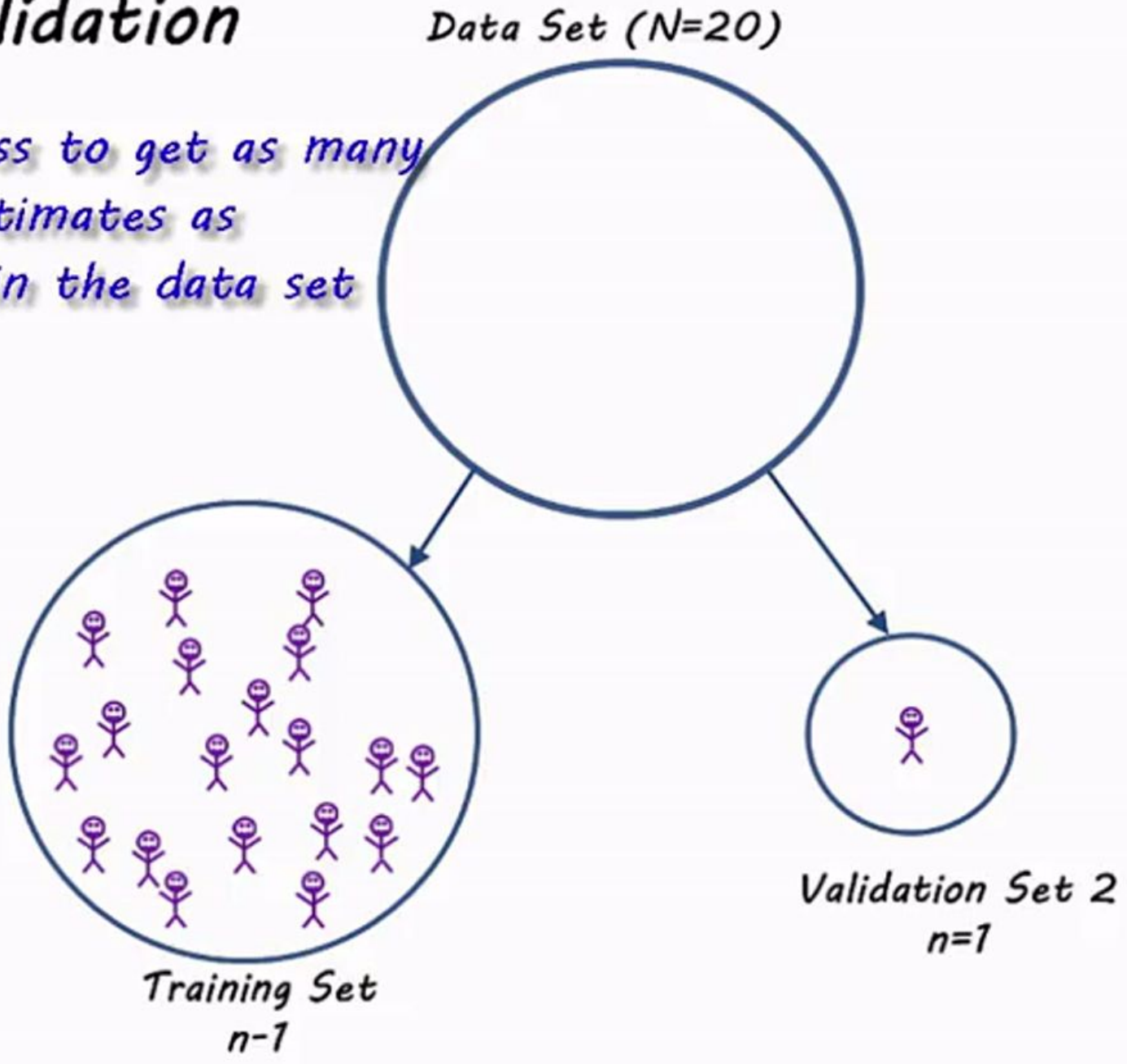
Leave One Out Cross Validation

- Test error based on only a single observation is highly variable



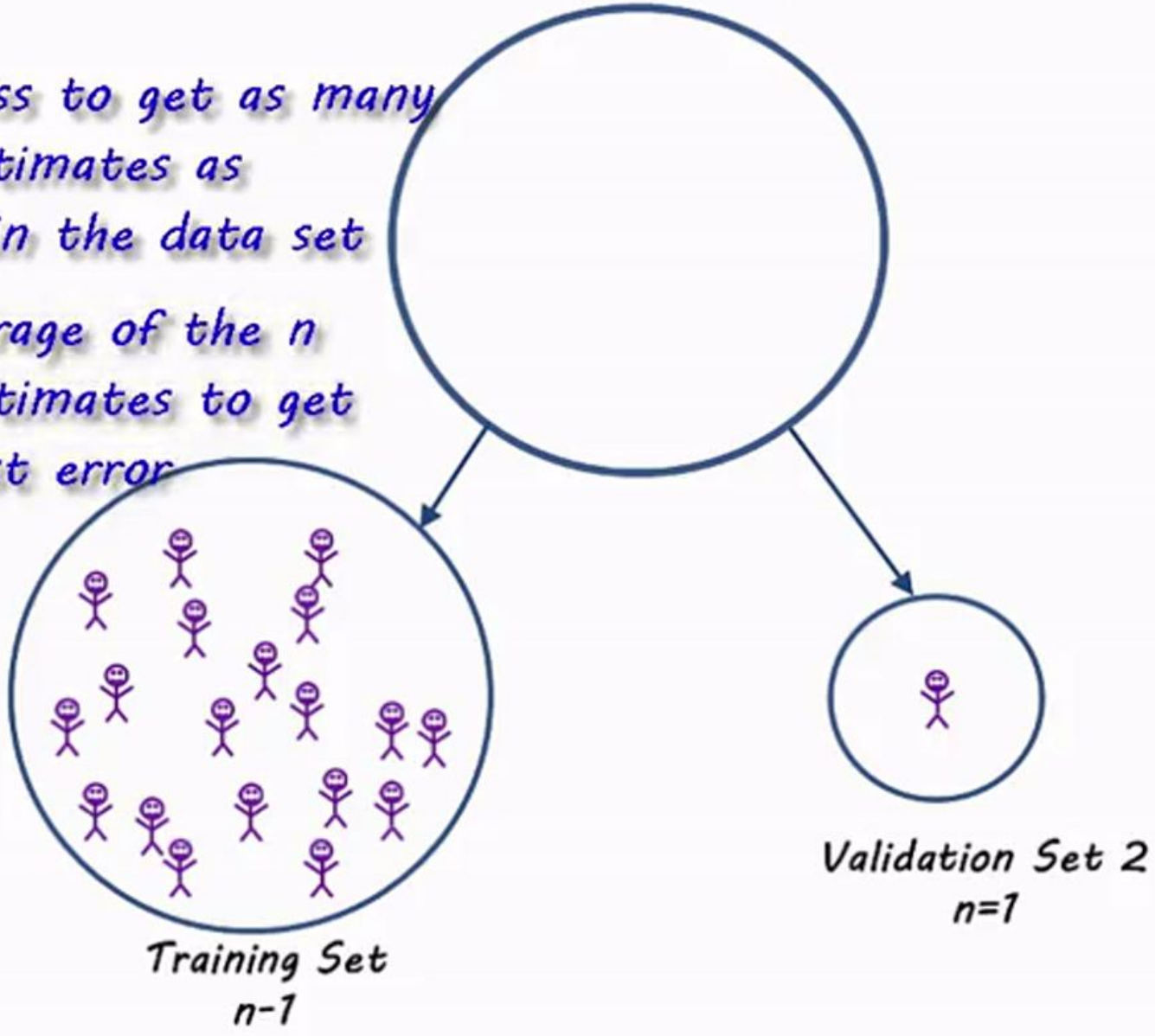
Leave One Out Cross Validation

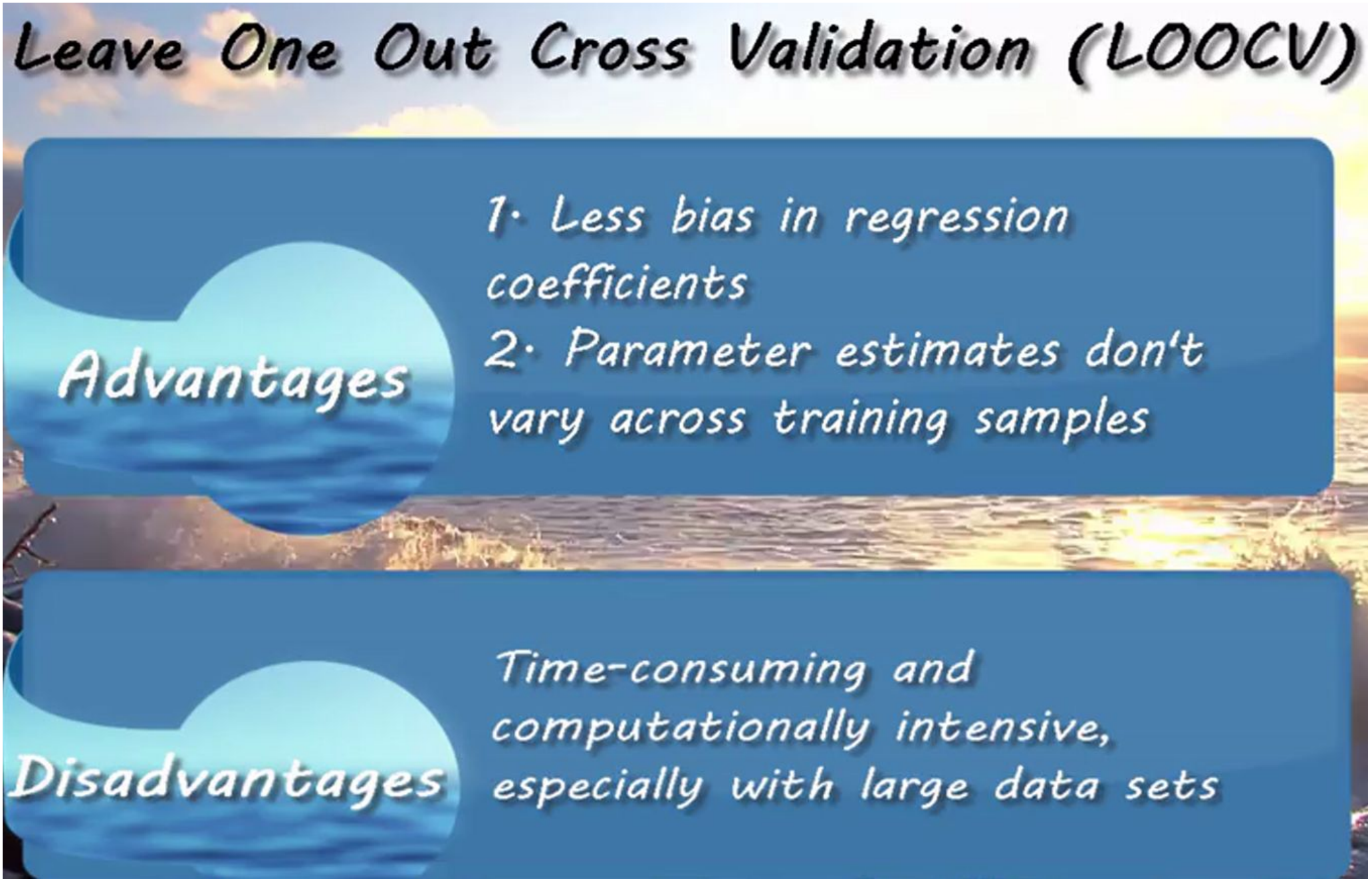
- Repeat process to get as many test error estimates as observations in the data set



Leave One Out Cross Validation

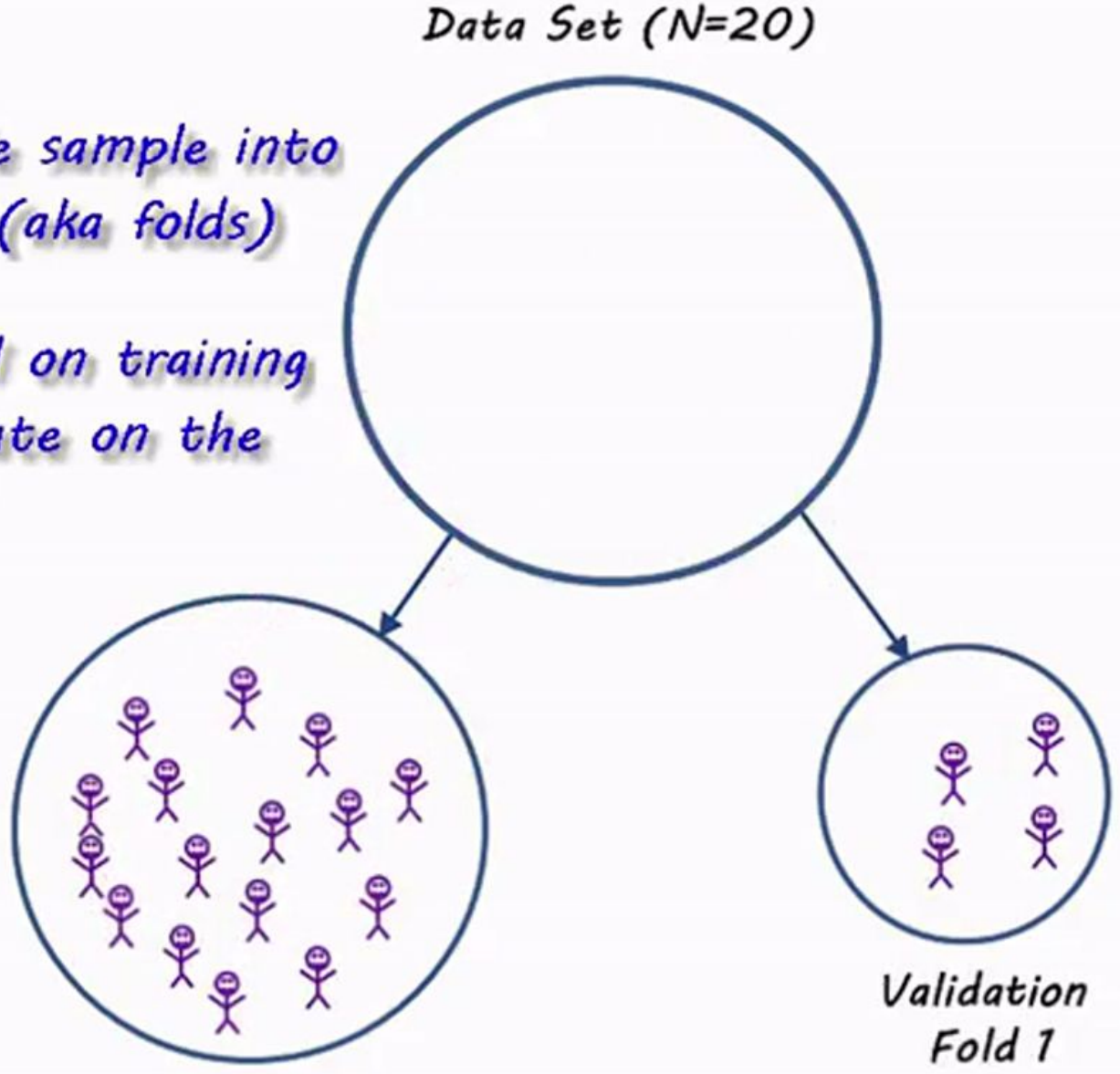
- Repeat process to get as many test error estimates as observations in the data set
- Compute average of the n test error estimates to get an overall test error estimate




Leave One Out Cross Validation (LOOCV)
Advantages
1. Less bias in regression coefficients
2. Parameter estimates don't vary across training samples
Disadvantages
Time-consuming and computationally intensive, especially with large data sets

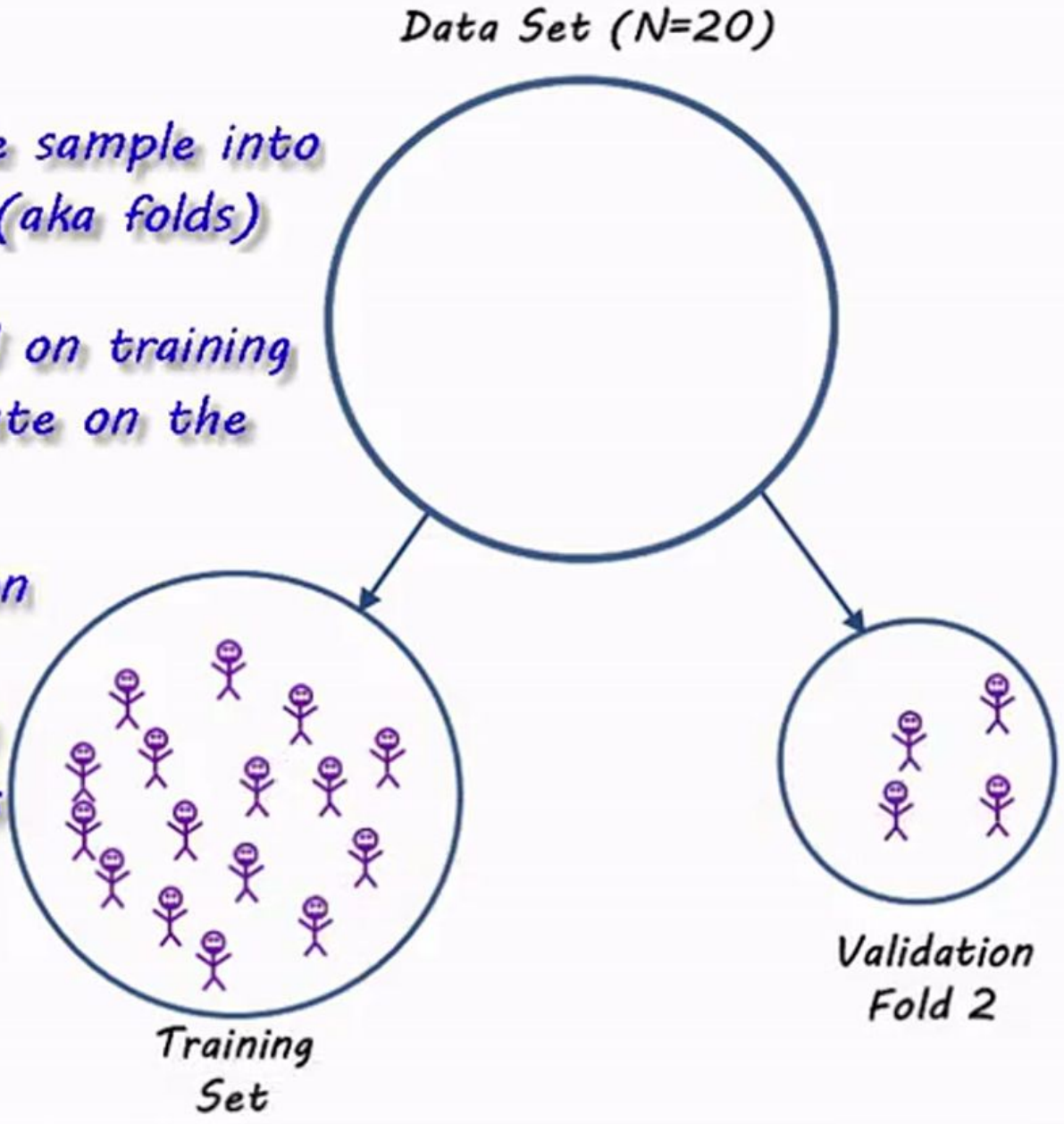
K-fold Cross Validation

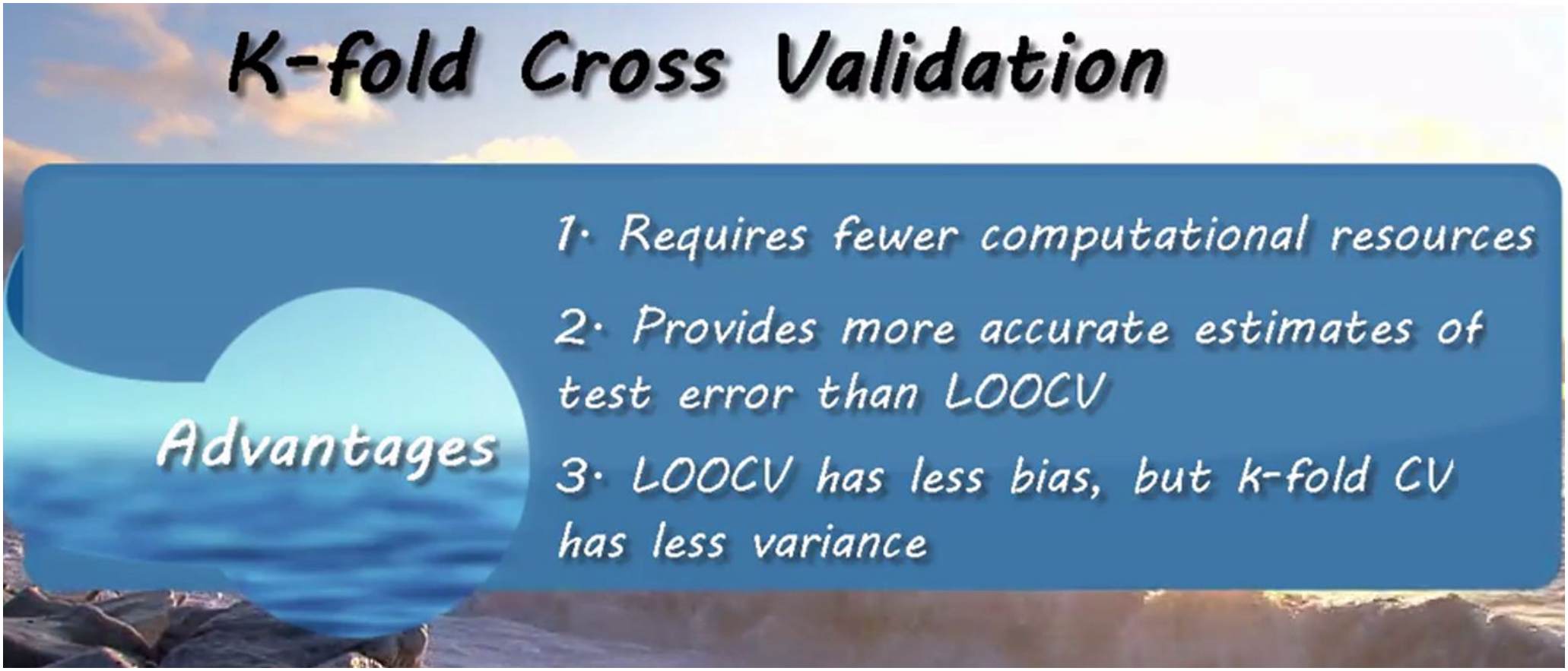
- Randomly divide sample into k equal groups (aka folds)
- Estimate model on training set, then validate on the validation fold



K-fold Cross Validation

- Randomly divide sample into k equal groups (aka folds)
- Estimate model on training set, then validate on the validation fold
- Repeat process on remaining folds, and average test error across folds




K-fold Cross Validation
Advantages
1. Requires fewer computational resources
2. Provides more accurate estimates of test error than LOOCV
3. LOOCV has less bias, but k-fold CV has less variance


Optimal: k=5 or k=10

Random-Forest-Program.sas

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
DATA new; set mydata.treeaddhealth;
PROC SORT; BY AID;
PROC HPFOREST;
target TREG1/level=nominal;
input BIO_SEX HISPANIC WHITE BLACK NAMERICAN ASIAN alcevr1 MARever1 cocever1 inhever1
      Cigavail PASSIST EXPEL1 /level=nominal;
input age DEVIANT1 VIOL1 DEP1 ESTEEM1 PARPRES PARACTV
      FAMCONCT schconn1 GPA1 /level=interval;

RUN;
```

Week 2.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp

                            ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                  ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                  ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                  ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                  ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                  ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                  ,(HIVRATE - 1.93544218) AS HIVRATE_2
                  ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                  ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                  ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                  ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                  ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                  ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                  ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                  ,(URBANRATE - 56.7693596) AS URBANRATE_2
```

```
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
      from      work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
```

```
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
/* DATA MANAGEMENT STEP */
IF SUICIDEPER100TH <  16  THEN SUICIDEPER100TH_RANK = 0;
IF SUICIDEPER100TH >= 16 THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
```

```
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = .;
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
```

```
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;
*******************************************************************************************
RANDOM FOREST
*******************************************************************************************;
PROC HPFOREST;
target SUICIDEPER100TH_RANK/level=nominal;
input EMPLOYRATE_RANK
INCOMEPERPERSON_RANK
ARMEDFORCESRATE_RANK
LIFEEXPECTANCY_RANK
URBANRATE_RANK
surarea_RANK
population_RANK
popden_RANK
corruptionindex_RANK
hdi_RANK /level=nominal;
input ALCCONSUMPTION /level=ORDINAL;

RUN;
```

Week 3: Lasso Regression

Shrinkage & Selection

Shrinkage — Constraint on parameters that shrinks coefficients toward zero

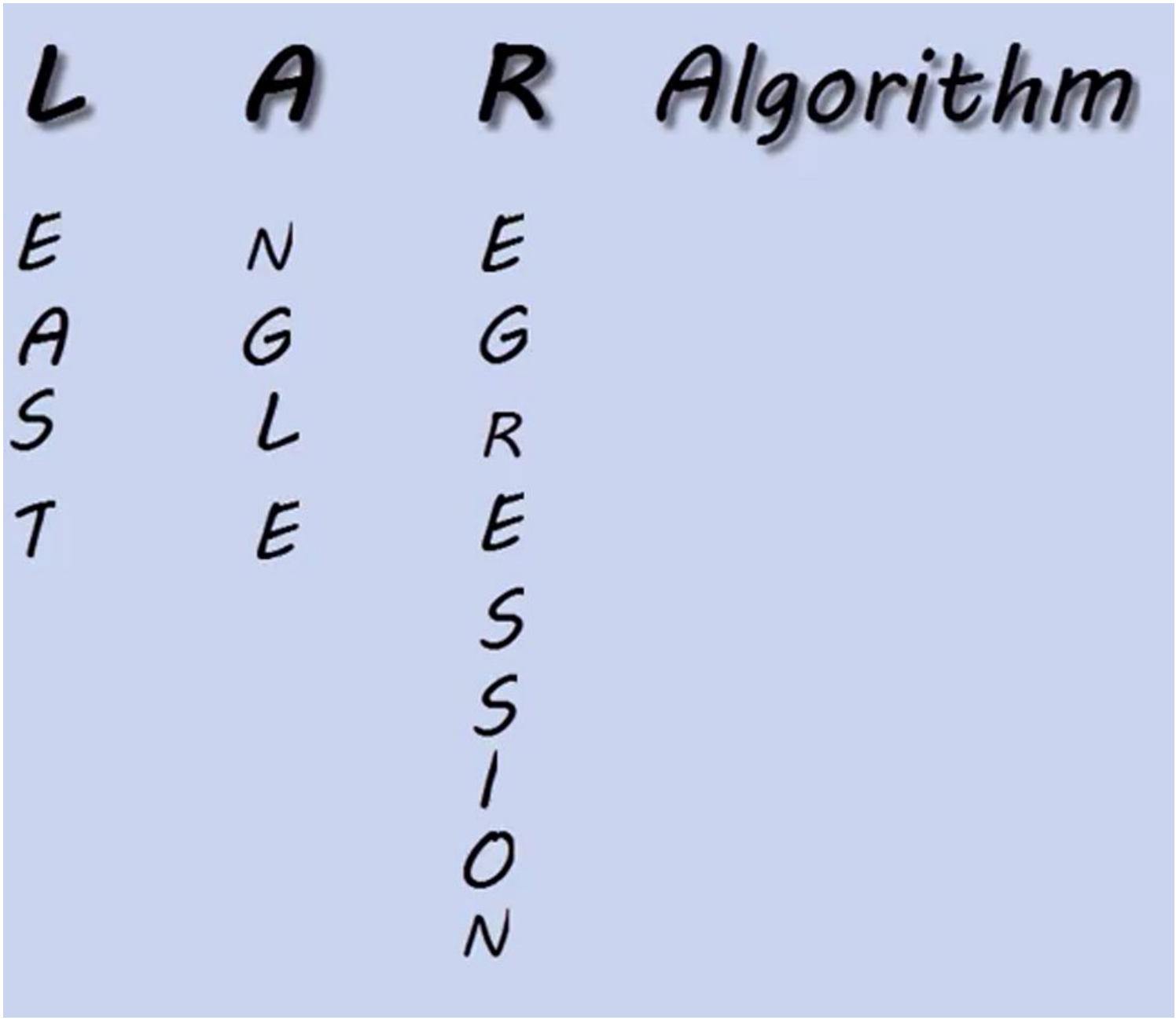
Selection — Identifies the most important variables associated with the response variable


Why Use Lasso Regression?
Greater prediction accuracy
Increase model interpretability

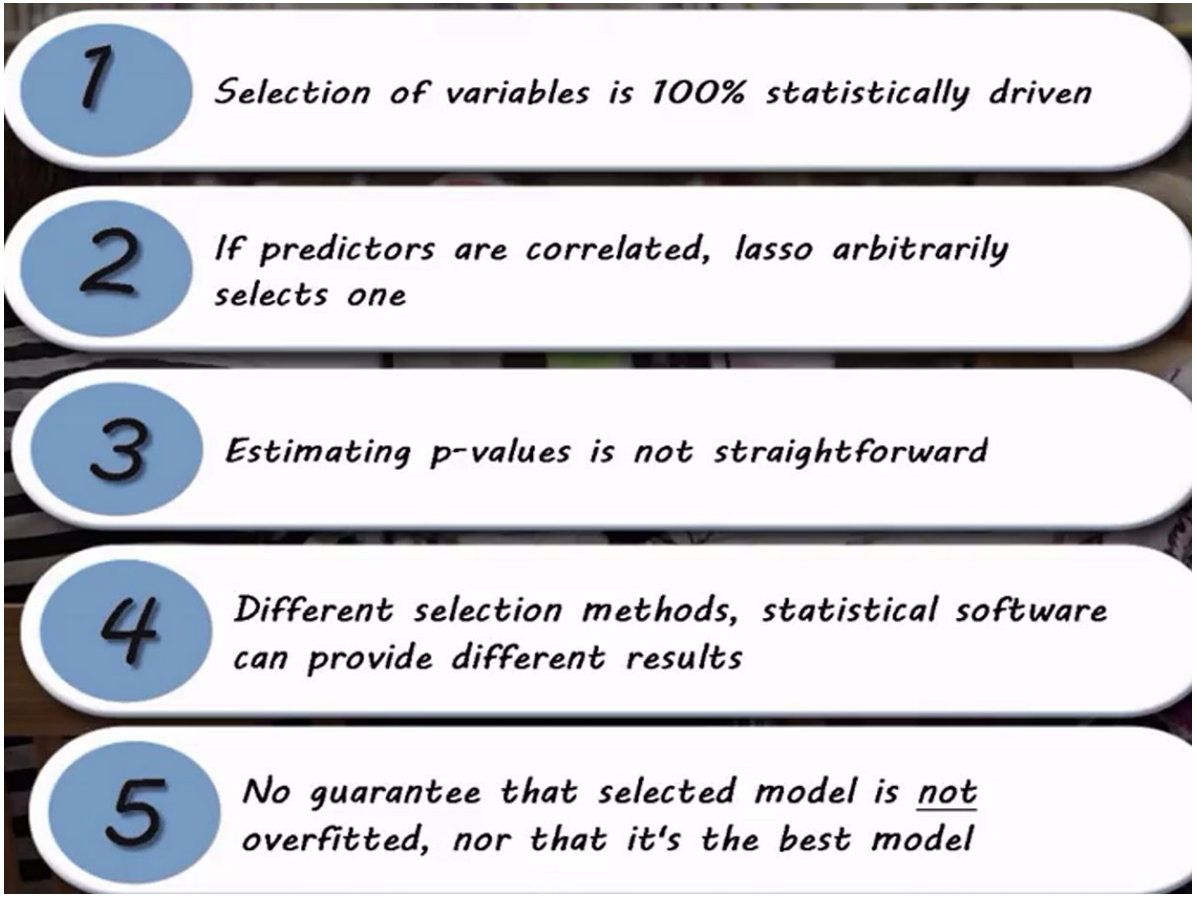
Why Use Lasso Regression?

Tuning parameter: lambda (λ)

- When λ=0 then it's OLS regression
- Bias increases and variance decreases as λ increases


L A R Algorithm
EAST
NGLE
EGRESSION

Lasso Regression Limitations




Machine Learning
+
Human
Intervention
+
Independent
Application

SAS-code-for-video-examples.sas

```
libname mydata "/courses/d1406ae5ba27fe300" access=readonly;
*************************************************************************************************
DATA MANAGEMENT
*************************************************************************************************;
data new;
set mydata.tree_addhealth;
if bio_sex=1 then male=1;
if bio_sex=2 then male=0;
* delete observations with missing data;
if cmiss(of _all_) then delete;
run;
ods graphics on;
* Split data randomly into test and training data;
proc surveyselect data=new out=traintest seed = 123
samprate=0.7 method=srs outall;
run;
* lasso multiple regression with lars algorithm k=10 fold validation;
proc glmselect data=traintest plots=all seed=123;
    partition ROLE=selected(train='1' test='0');
    model schconn1 = male hispanic white black namerican asian alcevr1 marever1 cocever1
    inhever1 cigavail passist expel1 age alcprobs1 deviant1 viol1 dep1 esteem1 parpres paractv
    famconct gpa1/selection=lar(choose=cv stop=none) cvmethod=random(10);
run;
```

---
week 3.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp

                            ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                  ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                  ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                  ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                  ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                  ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                  ,(HIVRATE - 1.93544218) AS HIVRATE_2
                  ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                  ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                  ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                  ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                  ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                  ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                  ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                  ,(URBANRATE - 56.7693596) AS URBANRATE_2
```

```
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
      from      work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
```

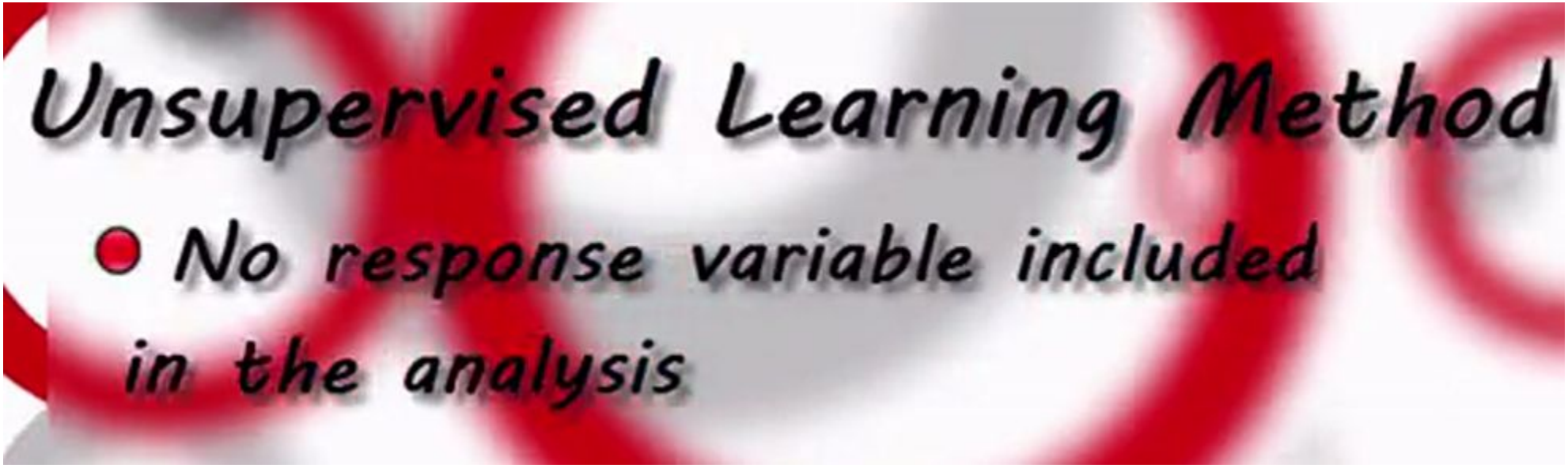
```
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
/* DATA MANAGEMENT STEP */
IF SUICIDEPER100TH <  16  THEN SUICIDEPER100TH_RANK = 0;
IF SUICIDEPER100TH >= 16 THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
```

```
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = .;
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
```

```
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;

*************************************************************************************************
LASSO REGRESSION
*************************************************************************************************;

* delete observations with missing data;
if cmiss(of _all_) then delete;
run;
ods graphics on;
* Split data randomly into test and training data;
proc surveyselect data=MYGAPMINDER out=traintest seed = 123
samprate=0.7 method=srs outall;
run;
* lasso multiple regression with lars algorithm k=10 fold validation;
proc glmselect data=traintest plots=all seed=123;
    partition ROLE=selected(train='1' test='0');
    model SUICIDEPER100TH_RANK =
ALCCONSUMPTION
EMPLOYRATE_RANK
INCOMEPERPERSON_RANK
ARMEDFORCESRATE_RANK
LIFEEXPECTANCY_RANK
URBANRATE_RANK
surarea_RANK
population_RANK
popden_RANK
```

```
corruptionindex_RANK
hdi_RANK /selection=lar(choose=cv stop=none) cvmethod=random(10);
run;
```

Week 4: K-Means Cluster Analysis

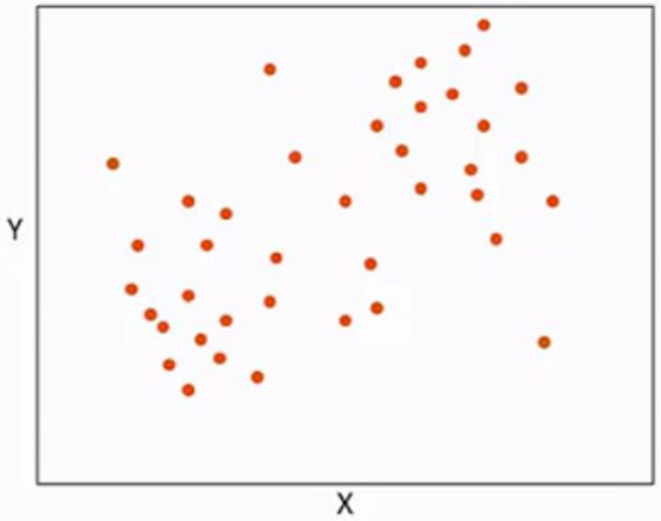



Less variance
within clusters
More variance
between clusters

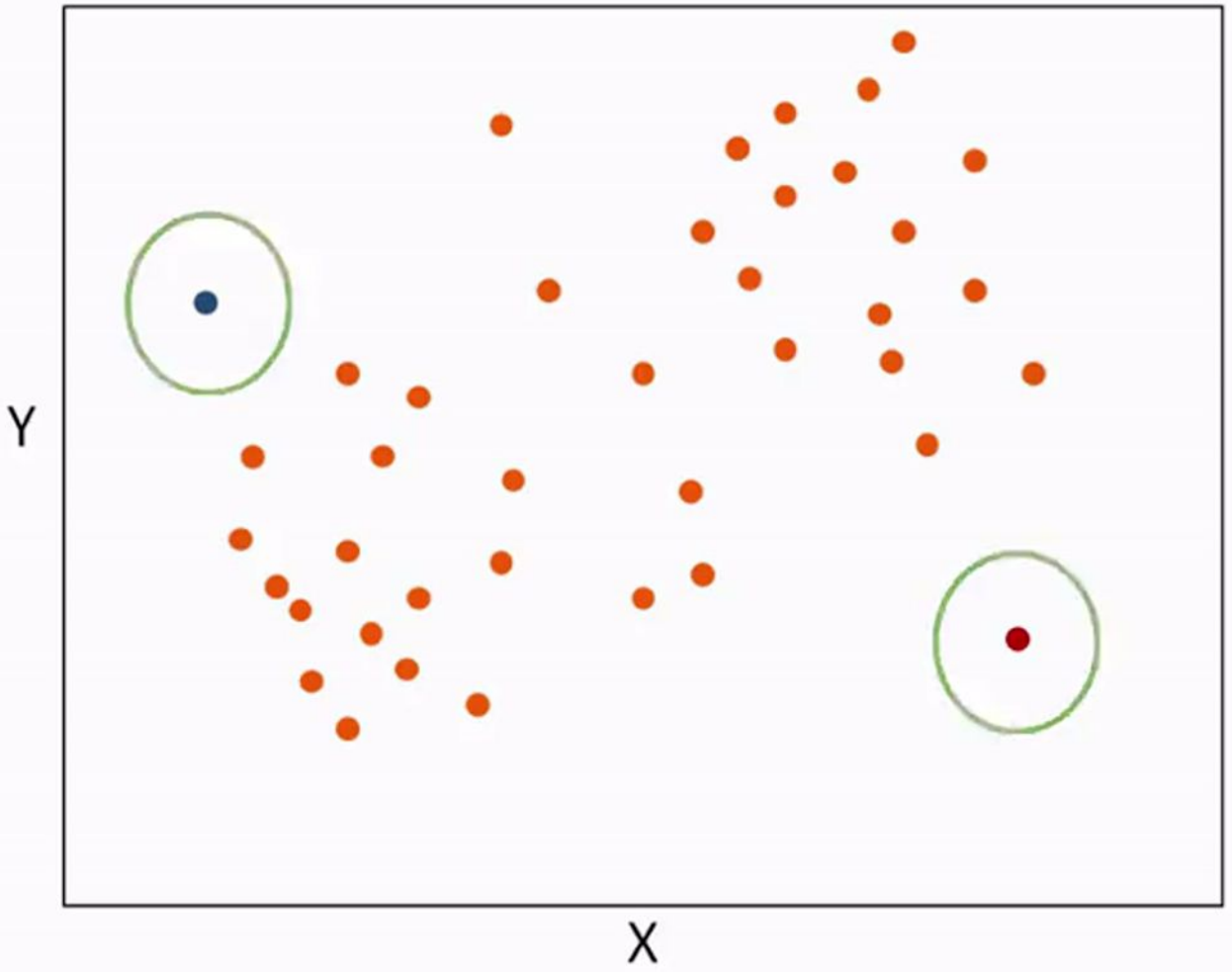

Data Reduction
Reduce # of variables

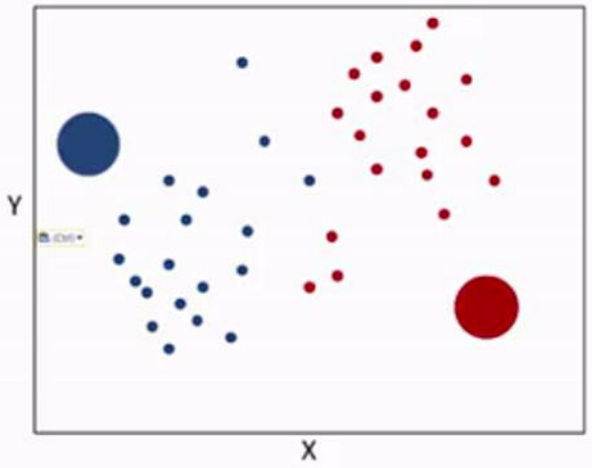
K-Means Cluster Analysis

- Creates a p-dimensional space where p is the number of input variables
- Euclidean distance

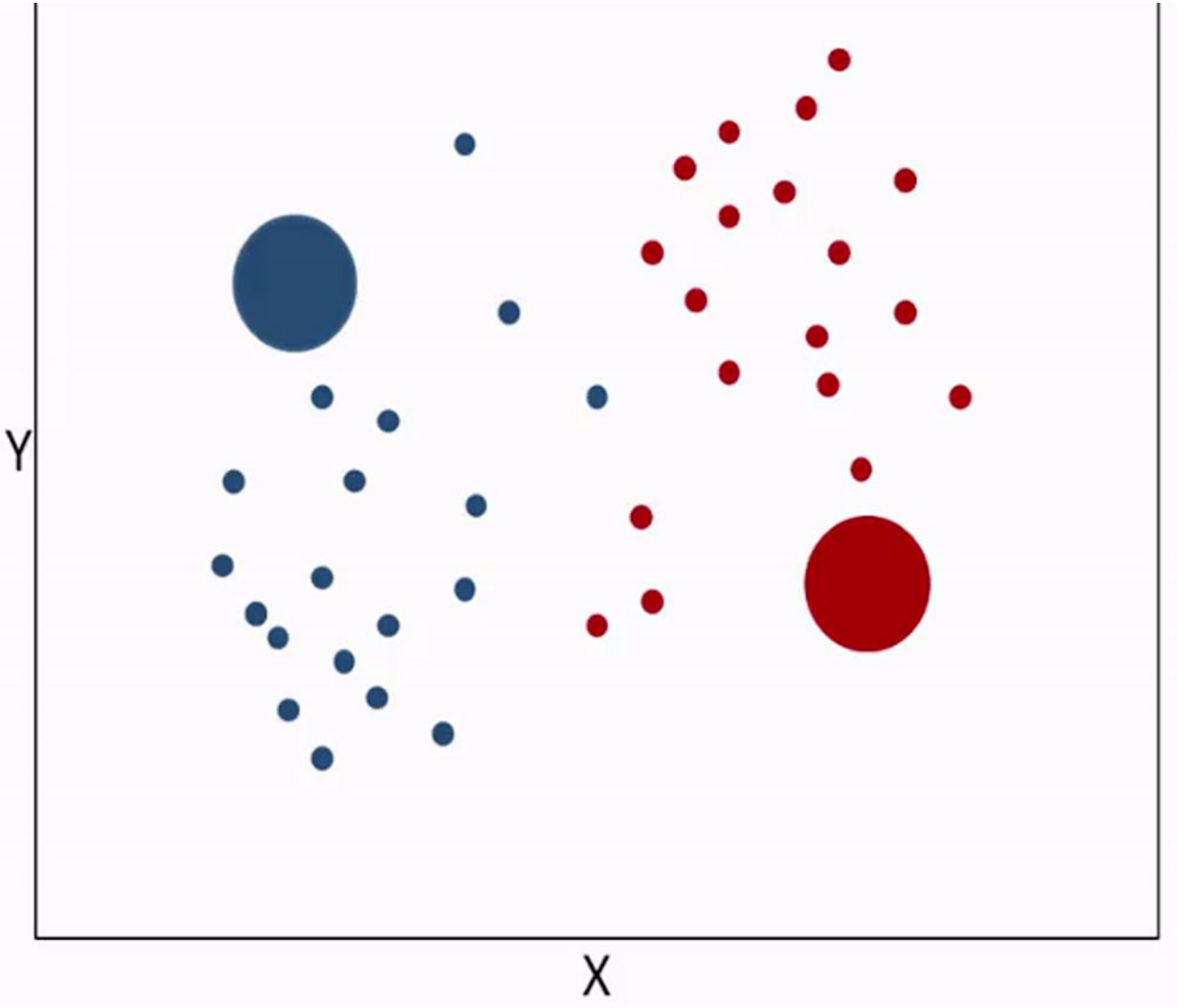


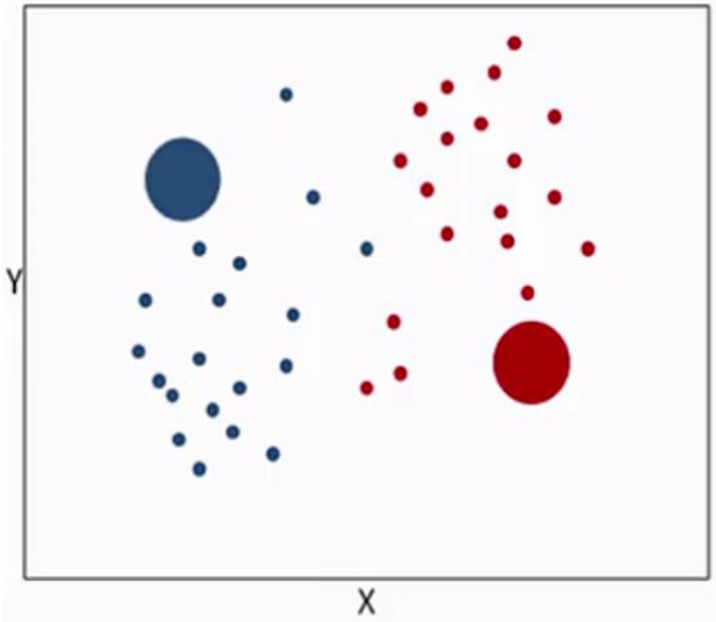
Step 1: randomly choose 2 points as the initial centroids (aka seeds) and calculate distance between points and the centroids



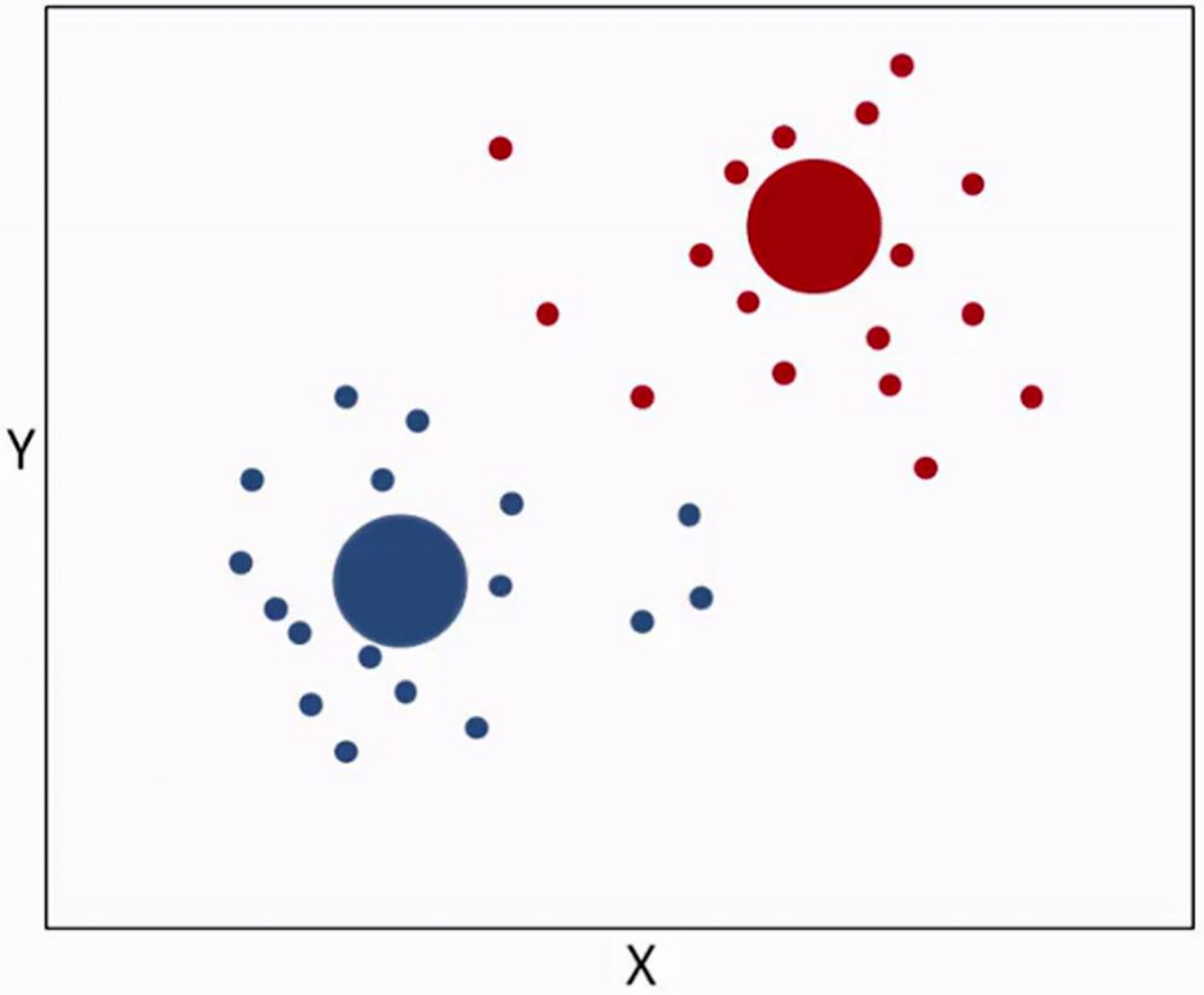


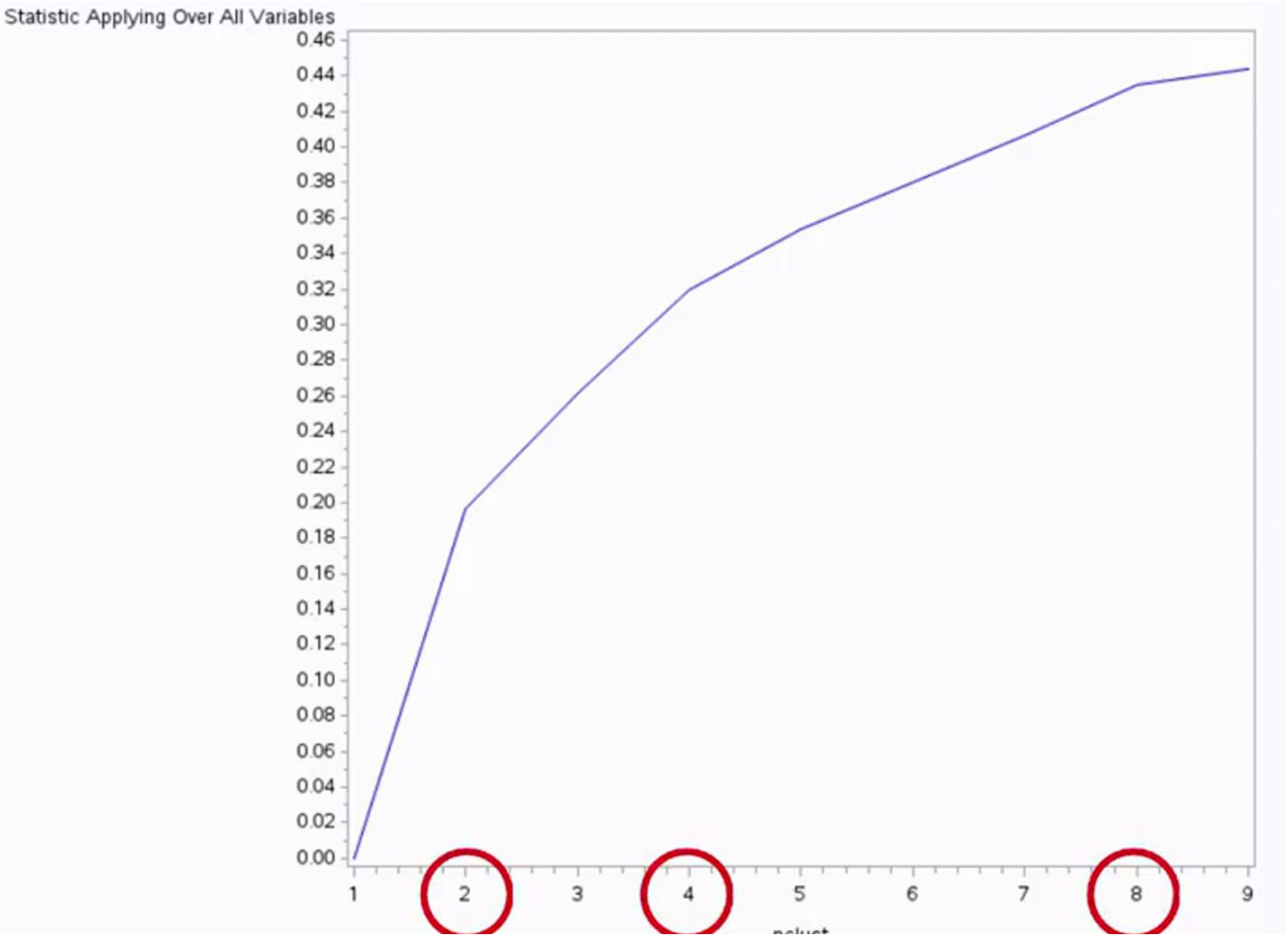
Step 2: each centroid is recalculated based on the location of the points that were assigned to it





Step 3: process is repeated until the location of the centroids doesn't change very much

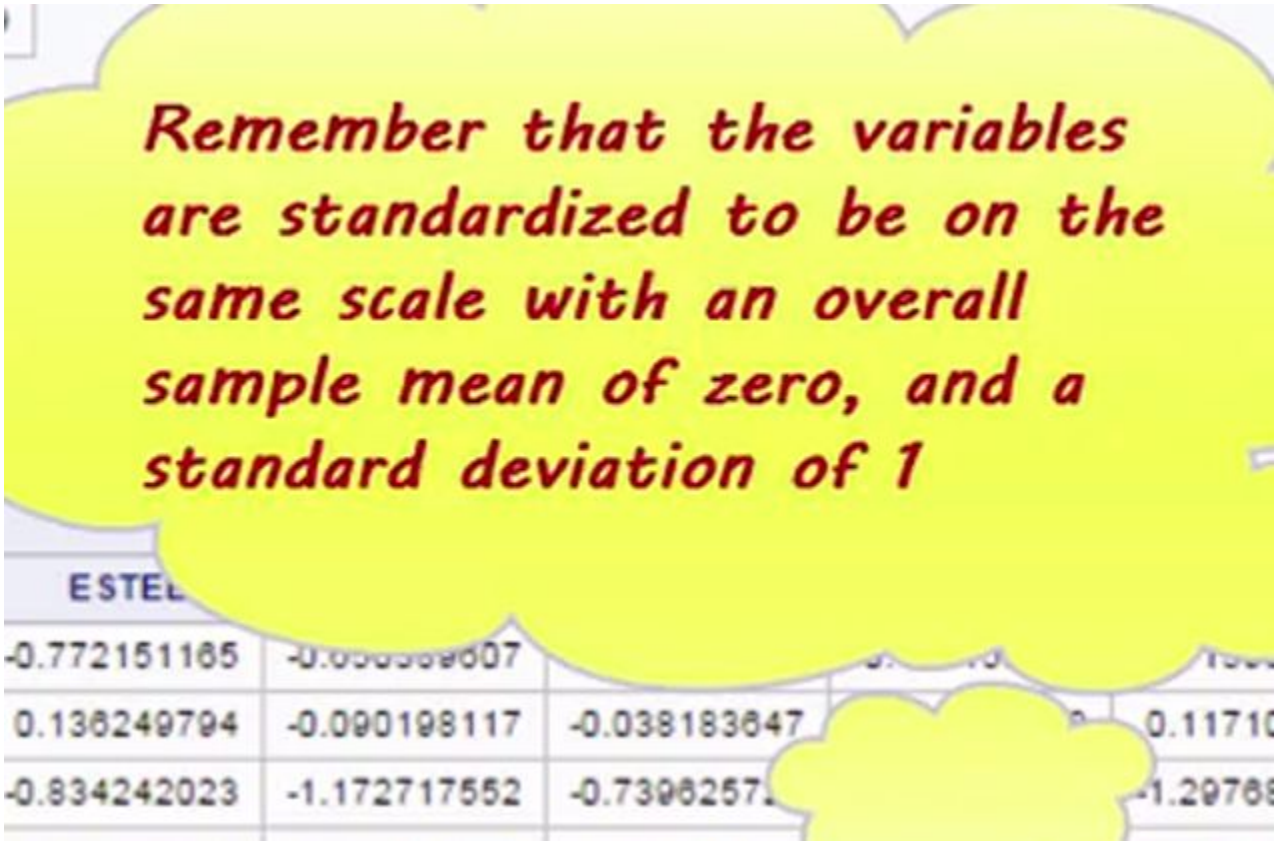


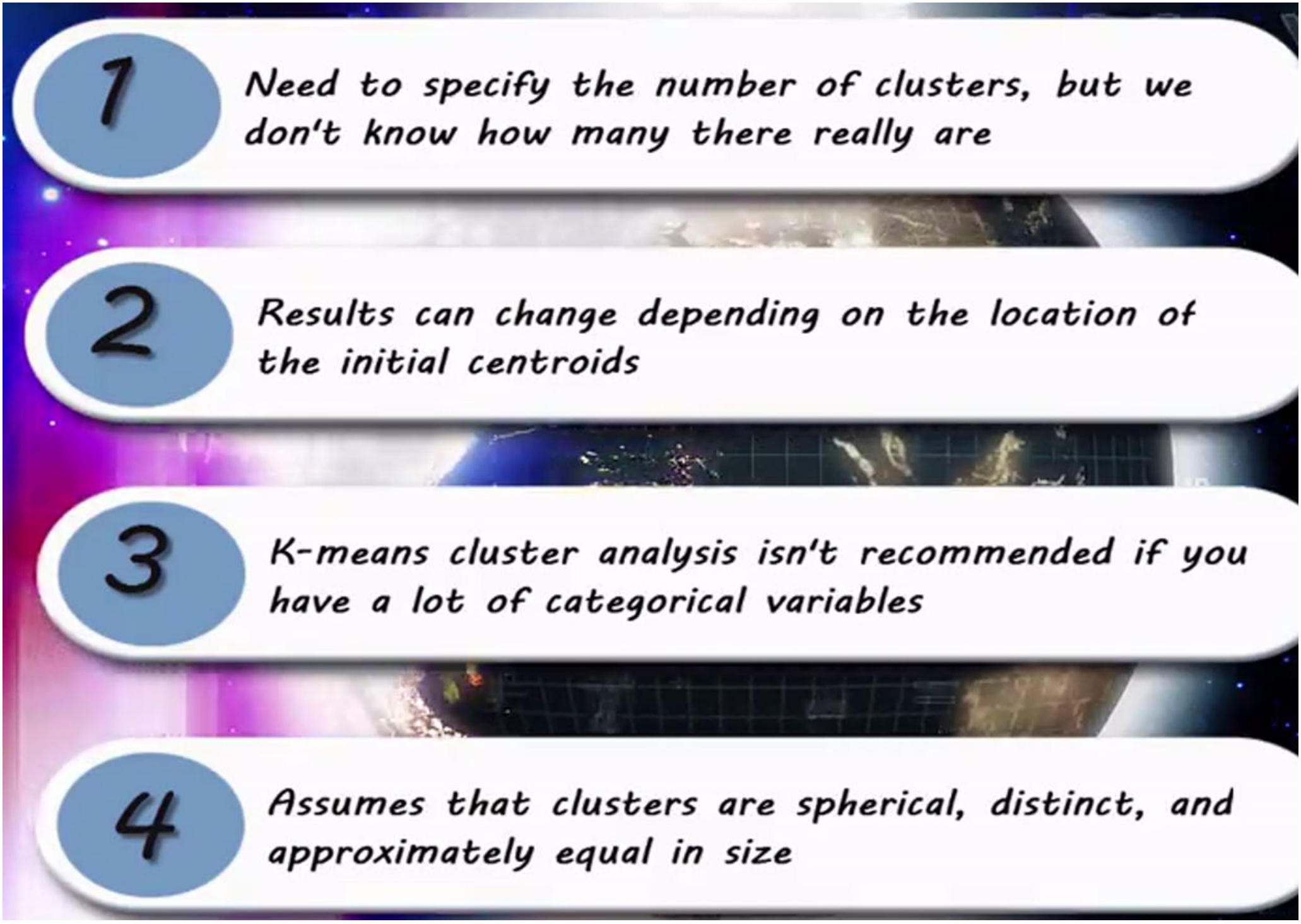

Statistic Applying Over All Variables
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9

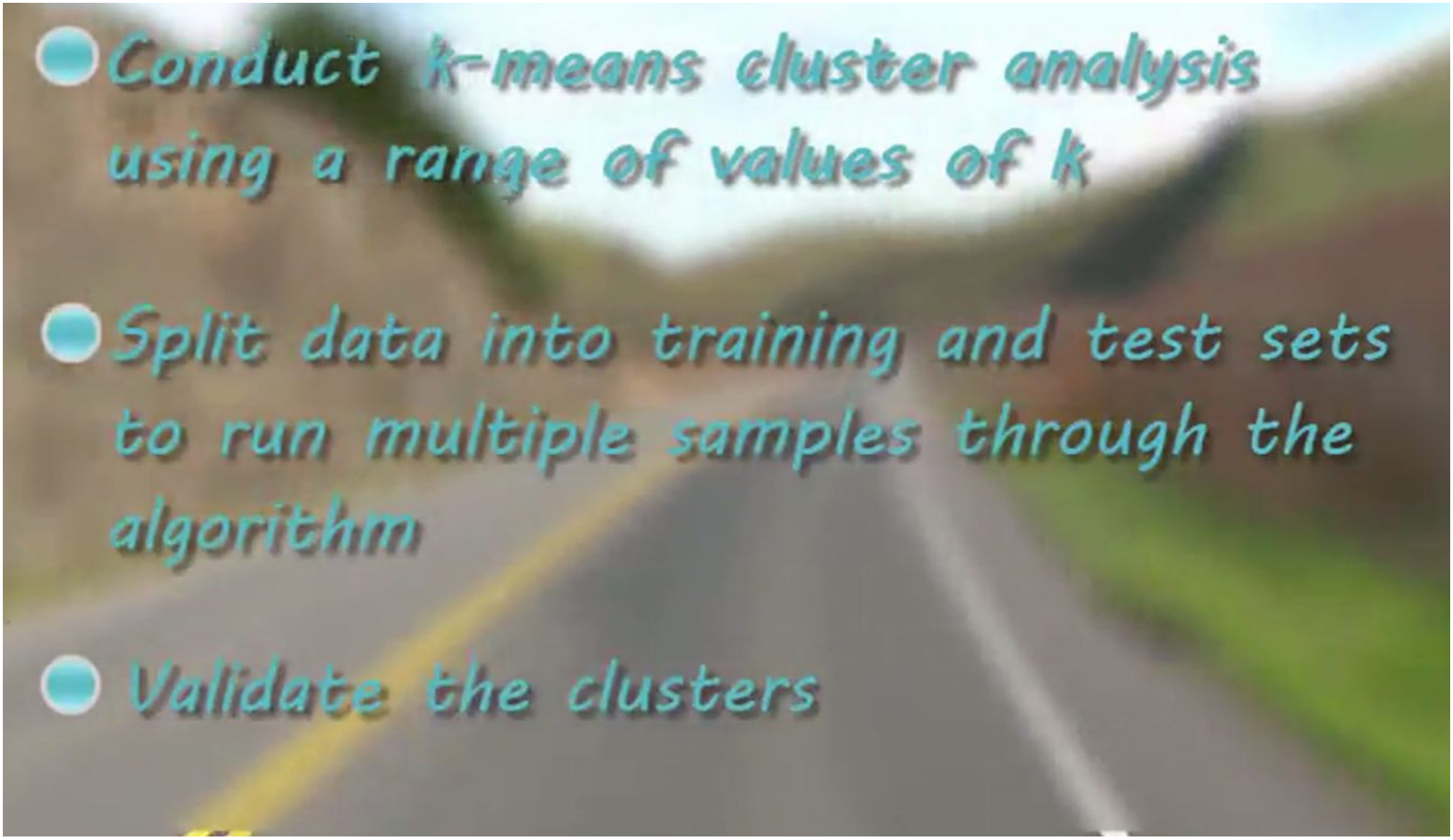
Canonical Discriminant Analysis

```
dep1 esteem1 schconn1
```

- Creates a smaller number of variables
- Linear combinations of clustering variables
- Canonical variables are ordered by proportion of variance accounted for
- Majority of variance is accounted for by first few canonical variables


Remember that the variables are standardized to be on the same scale with an overall sample mean of zero, and a standard deviation of 1
-0.772151165
0.136249794
-0.090198117
-0.038183647
-0.834242023
-1.172717552


1
Need to specify the number of clusters, but we don't know how many there really are
2
Results can change depending on the location of the initial centroids
3
K-means cluster analysis isn't recommended if you have a lot of categorical variables
4
Assumes that clusters are spherical, distinct, and approximately equal in size


Conduct k-means cluster analysis using a range of values of k
Split data into training and test sets to run multiple samples through the algorithm
Validate the clusters

SAS-code-for-video-examples.sas

```
libname mydata "/courses/d1406ae5ba27fe300" access=readonly;

*************************************************************************************************
DATA MANAGEMENT
*************************************************************************************************;
data clust;
set mydata.tree_addhealth;
* create a unique identifier to merge cluster assignment variable with
the main data set;
idnum=_n_;
keep idnum alcevr1 marever1 alcprobs1 deviant1 viol1 dep1 esteem1 schconn1
   parpres paractv famconct gpa1;

* delete observations with missing data;
if cmiss(of _all_) then delete;
run;
ods graphics on;
* Split data randomly into test and training data;
proc surveyselect data=clust out=traintest seed = 123
samprate=0.7 method=srs outall;
run;
data clus_train;
set traintest;
if selected=1;
run;
data clus_test;
set traintest;
if selected=0;
```

```
run;
* standardize the clustering variables to have a mean of 0 and standard deviation of 1;
proc standard data=clus_train out=clustvar mean=0 std=1;
var alcevr1 marever1 alcprobs1 deviant1 viol1 dep1 esteem1 schconn1
   parpres paractv famconct;
run;
%macro kmean(K);
proc fastclus data=clustvar out=outdata&K. outstat=cluststat&K. maxclusters= &K. maxiter=300;
var alcevr1 marever1 alcprobs1 deviant1 viol1 dep1 esteem1 schconn1
   parpres paractv famconct;
run;
%mend;
%kmean(1);
%kmean(2);
%kmean(3);
%kmean(4);
%kmean(5);
%kmean(6);
%kmean(7);
%kmean(8);
%kmean(9);
* extract r-square values from each cluster solution and then merge them to plot elbow curve;
data clus1;
set cluststat1;
nclust=1;
if _type_='RSQ';
keep nclust over_all;
run;
data clus2;
set cluststat2;
nclust=2;
```

```
if _type_='RSQ';
keep nclust over_all;
run;
data clus3;
set cluststat3;
nclust=3;
if _type_='RSQ';
keep nclust over_all;
run;
data clus4;
set cluststat4;
nclust=4;
if _type_='RSQ';
keep nclust over_all;
run;
data clus5;
set cluststat5;
nclust=5;
if _type_='RSQ';
keep nclust over_all;
run;
data clus6;
set cluststat6;
nclust=6;
if _type_='RSQ';
keep nclust over_all;
run;
data clus7;
set cluststat7;
nclust=7;
if _type_='RSQ';
```

```
keep nclust over_all;
run;
data clus8;
set cluststat8;
nclust=8;
if _type_='RSQ';
keep nclust over_all;
run;
data clus9;
set cluststat9;
nclust=9;
if _type_='RSQ';
keep nclust over_all;
run;
data clusrsquare;
set clus1 clus2 clus3 clus4 clus5 clus6 clus7 clus8 clus9;
run;
* plot elbow curve using r-square values;
symbol1 color=blue interpol=join;
proc gplot data=clusrsquare;
plot over_all*nclust;
Run;

*******************************************************************************
further examine cluster solution for the number of clusters suggested by the elbow curve
*******************************************************************************
* plot clusters for 4 cluster solution;
proc candisc data=outdata4 out=clustcan;
class cluster;
var alcevr1 marever1 alcprobs1 deviant1 viol1 dep1 esteem1 schconn1
   parpres paractv famconct;
```

```
run;
proc sgplot data=clustcan;
scatter y=can2 x=can1 / group=cluster;
run;
* validate clusters on GPA;
* first merge clustering variable and assignment data with GPA data;
data gpa_data;
set clus_train;
keep idnum gpa1;
run;
proc sort data=outdata4;
by idnum;
run;
proc sort data=gpa_data;
by idnum;
run;
data merged;
merge outdata4 gpa_data;
by idnum;
run;
proc sort data=merged;
by cluster;
run;
proc means data=merged;
var gpa1;
by cluster;
run;
proc anova data=merged;
class cluster;
model gpa1 = cluster;
means cluster/tukey;
```

```
run;
```

---
week 4.sas

```
/* COURSERA GAPMINDER DATA */
libname mydata "/courses/d1406ae5ba27fe300 " access=readonly;
data gapminder;
     set mydata.gapminder;
/* IMPORTING ADDITIONAL DATA (source: https://www.gapminder.org/) */
PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_population density (per square km).csv'
           OUT=popden REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator_total population with projections.csv'
           OUT=pop REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/surface land.csv'
           OUT=surarea REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/indicator ti cpi 2009.csv'
           OUT=cpi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Indicator_HDI.csv'
           OUT=hdi REPLACE;

PROC IMPORT
           DATAFILE='/home/qallaf890/Homicide age adjusted indicator LIVE -05 20100919.csv'
           OUT=murder REPLACE;
```

```
PROC IMPORT
           DATAFILE='/home/qallaf890/military expenditure.csv'
           OUT=military REPLACE;
/* JOINING THE DATA USING SQL */
proc sql;
      create table mygapminder AS
      select           gapminder.*
                       ,surarea.surarea
                       ,pop.population
                       ,popden.popden
                       ,cpi.corruptionindex
                       ,hdi.hdi
                       ,murder.homicide
                       ,military.milexpprcntgdp

                            ,(INCOMEPERPERSON - 8740.96608) AS INCOMEPERPERSON_2
                  ,(ALCCONSUMPTION - 6.68941176) AS ALCCONSUMPTION_2
                  ,(ARMEDFORCESRATE - 1.44401628) AS ARMEDFORCESRATE_2
                  ,(BREASTCANCERPER100TH - 37.4028902) AS BREASTCANCERPER100TH_2
                  ,(CO2EMISSIONS - 5033261622) AS CO2EMISSIONS_2
                  ,(FEMALEEMPLOYRATE - 47.5494381) AS FEMALEEMPLOYRATE_2
                  ,(HIVRATE - 1.93544218) AS HIVRATE_2
                  ,(INTERNETUSERATE - 35.6327158) AS INTERNETUSERATE_2
                  ,(LIFEEXPECTANCY - 69.7535236) AS LIFEEXPECTANCY_2
                  ,(OILPERPERSON - 1.48408516) AS OILPERPERSON_2
                  ,(POLITYSCORE - 3.68944099) AS POLITYSCORE_2
                  ,(RELECTRICPERPERSON - 1173.17899) AS RELECTRICPERPERSON_2
                  ,(SUICIDEPER100TH - 9.64083901) AS SUICIDEPER100TH_2
                  ,(EMPLOYRATE - 58.6359551) AS EMPLOYRATE_2
                  ,(URBANRATE - 56.7693596) AS URBANRATE_2
```

```
                ,(SURAREA - 677459.604) AS SURAREA_2
                ,(POPULATION - 33730861.5) AS POPULATION_2
                ,(POPDEN - 468.994722) AS POPDEN_2
                ,(CORRUPTIONINDEX - 4.02349398 ) AS CORRUPTIONINDEX_2
                ,(HDI - 0.66335593) AS HDI_2
                ,(HOMICIDE - 11.5500871) AS HOMICIDE_2
      from      work.gapminder as gapminder
                     left join work.popden as popden on gapminder.country = popden.country
                     left join work.pop as pop on gapminder.country = pop.country
                     left join work.surarea as surarea on gapminder.country = surarea.country
                     left join work.cpi as cpi on gapminder.country = cpi.country
                     left join work.hdi as hdi on gapminder.country = hdi.country
                     left join work.murder as murder on gapminder.country = murder.country
                     left join work.military as military on gapminder.country = military.country;
quit;
DATA mygapminder;
      set work.mygapminder;
/* GIVING DESCRIPTIONS TO VARIABLES */
LABEL
      COUNTRY='COUNTRY'
      INCOMEPERPERSON='GDP PER CAPITA'
      ALCCONSUMPTION='LITRES OF ALCOHOL CONSUMPTION'
      ARMEDFORCESRATE='ARMED FORCES % OF TOTAL LABOR'
      BREASTCANCERPER100TH='FEMALE BREAST CANCER PER 100,000'
      CO2EMISSIONS='TOTAL AMOUNT OF CO2 EMISSIONS (IN METRIC TONS)'
      FEMALEEMPLOYRATE='% OF FEMALE POPULATION EMPLOYED'
      EMPLOYRATE='% OF POPULATION EMPLOYED'
      HIVRATE='% ESTIMATED HIV PREVALENCE'
      INTERNETUSERATE='INTERNET USERS (PER 100)'
      LIFEEXPECTANCY='LIFE EXPECTANCY AT BIRTH'
      OILPERPERSON='OIL CONSUMPTION PER CAPITA (TONNES PER YEAR AND PERSON)'
```

```
      POLITYSCORE='DEMOCRACY SCORE MINUS AUTOCRACY SCORE'
      RELECTRICPERPERSON='RESEDENTIAL ELECTRICITY CONSUMPTION PER PERSON (KWH)'
      SUICIDEPER100TH='SUCIDE PER 100,000'
      URBANRATE='URBAN POPULATION (% OF TOTAL)'
      surarea='SURFACE AREA (IN KM^2)'
      population='TOTAL POPULATION'
      popden='POPULATION DENSITY (PER SQAURE KM)'
      corruptionindex='CORRUPTION PERCEPTION INDEX'
      hdi='HUMAN DEVELOPMENT INDEX'
      homicide='MURDER, AGE ADJUSTED, PER 100,000'
      milexpprcntgdp='MILITARY EXPENDITURE (% OF GDP)'
      ;
/* DATA MANAGEMENT STEP */
IF SUICIDEPER100TH <  16  THEN SUICIDEPER100TH_RANK = 0;
IF SUICIDEPER100TH >= 16 THEN SUICIDEPER100TH_RANK = 1;
IF SUICIDEPER100TH = . THEN SUICIDEPER100TH_RANK = .;
/* DATA MANAGEMENT STEP
NOTE: these are based on the following quantiles (<%25, <%50, <%75, >=%75)
*/
IF EMPLOYRATE <   51.2 THEN EMPLOYRATE_RANK = 1;
IF EMPLOYRATE >=  51.2 AND  EMPLOYRATE <  58.7 THEN EMPLOYRATE_RANK = 2;
IF EMPLOYRATE >=  58.7 AND  EMPLOYRATE <  65.0 THEN EMPLOYRATE_RANK = 3;
IF EMPLOYRATE >=  65.0 THEN EMPLOYRATE_RANK = 4;
IF EMPLOYRATE = . THEN EMPLOYRATE_RANK = .;
IF INCOMEPERPERSON <  744.239  THEN INCOMEPERPERSON_RANK = 1;
IF INCOMEPERPERSON >= 744.239  AND  INCOMEPERPERSON <  2553.496 THEN INCOMEPERPERSON_RANK = 2;
IF INCOMEPERPERSON >= 2553.496 AND  INCOMEPERPERSON <  9425.326 THEN INCOMEPERPERSON_RANK = 3;
IF INCOMEPERPERSON >= 9425.326 THEN INCOMEPERPERSON_RANK = 4;
IF INCOMEPERPERSON = . THEN INCOMEPERPERSON_RANK = .;
IF ARMEDFORCESRATE <   0.478489 THEN ARMEDFORCESRATE_RANK = 1;
IF ARMEDFORCESRATE >=  0.478489 AND  ARMEDFORCESRATE < 0.930638  THEN ARMEDFORCESRATE_RANK = 2;
```

```
IF ARMEDFORCESRATE >=  0.930638 AND  ARMEDFORCESRATE < 1.613217  THEN ARMEDFORCESRATE_RANK = 3;
IF ARMEDFORCESRATE >=  1.613217 THEN ARMEDFORCESRATE_RANK = 4;
IF ARMEDFORCESRATE = . THEN ARMEDFORCESRATE_RANK = .;
IF LIFEEXPECTANCY <  64.228 THEN LIFEEXPECTANCY_RANK = 1;
IF LIFEEXPECTANCY >= 64.228 AND  LIFEEXPECTANCY < 73.131  THEN LIFEEXPECTANCY_RANK = 2;
IF LIFEEXPECTANCY >= 73.131 AND  LIFEEXPECTANCY < 76.640  THEN LIFEEXPECTANCY_RANK = 3;
IF LIFEEXPECTANCY >= 76.640 THEN LIFEEXPECTANCY_RANK = 4;
IF LIFEEXPECTANCY = . THEN LIFEEXPECTANCY_RANK = .;
IF URBANRATE <  36.82 THEN URBANRATE_RANK = 1;
IF URBANRATE >= 36.82 AND URBANRATE < 57.94   THEN URBANRATE_RANK = 2;
IF URBANRATE >= 57.94 AND URBANRATE < 74.50   THEN URBANRATE_RANK = 3;
IF URBANRATE >= 74.50 THEN URBANRATE_RANK = 4;
IF URBANRATE = . THEN URBANRATE_RANK = .;
IF surarea <  18580  THEN surarea_RANK = 1;
IF surarea >= 18580  AND  surarea < 112620   THEN surarea_RANK = 2;
IF surarea >= 112620 AND  surarea < 488100   THEN surarea_RANK = 3;
IF surarea >= 488100 THEN surarea_RANK = 4;
IF surarea = . THEN surarea_RANK = .;
IF population <  882863   THEN population_RANK = 1;
IF population >= 882863   AND  population < 6412560  THEN population_RANK = 2;
IF population >= 6412560  AND  population < 22555046 THEN population_RANK = 3;
IF population >= 22555046 THEN population_RANK = 4;
IF population = . THEN population_RANK = .;
IF popden <  1032  THEN popden_RANK = 1;
IF popden >= 1032  THEN popden_RANK = 2;
IF popden = . THEN popden_RANK = .;
IF corruptionindex <  2.4 THEN corruptionindex_RANK = 1;
IF corruptionindex >= 2.4 AND corruptionindex < 3.3  THEN corruptionindex_RANK = 2;
IF corruptionindex >= 3.3 AND corruptionindex < 5.2  THEN corruptionindex_RANK = 3;
IF corruptionindex >= 5.2 THEN corruptionindex_RANK = 4;
IF corruptionindex = . THEN corruptionindex_RANK = .;
```

```
IF hdi <  0.522 THEN hdi_RANK = 1;
IF hdi >= 0.522 AND hdi < 0.698  THEN hdi_RANK = 2;
IF hdi >= 0.698 AND hdi < 0.793  THEN hdi_RANK = 3;
IF hdi >= 0.793 THEN hdi_RANK = 4;
IF hdi = . THEN hdi_RANK = .;
*******************************************************************************************
K-MEANS CLUSTER ANALYSIS
*******************************************************************************************;
data clust;
set MYGAPMINDER;
* create a unique identifier to merge cluster assignment variable with
the main data set;
idnum=_n_;
keep
idnum
incomeperperson
alcconsumption
armedforcesrate
breastcancerper100th
co2emissions
femaleemployrate
hivrate
internetuserate
lifeexpectancy
oilperperson
polityscore
relectricperperson
suicideper100th
employrate
urbanrate
surarea
```

```
population
popden
corruptionindex
hdi
homicide
milexpprcntgdp;

* delete observations with missing data;
if cmiss(of _all_) then delete;
run;
ods graphics on;
* Split data randomly into test and training data;
proc surveyselect data=clust out=traintest seed = 123
samprate=0.7 method=srs outall;
run;
data clus_train;
set traintest;
if selected=1;
run;
data clus_test;
set traintest;
if selected=0;
run;
* standardize the clustering variables to have a mean of 0 and standard deviation of 1;
proc standard data=clus_train out=clustvar mean=0 std=1;
var incomeperperson
alcconsumption
armedforcesrate
breastcancerper100th
co2emissions
femaleemployrate
```

```
hivrate
internetuserate
lifeexpectancy
oilperperson
polityscore
relectricperperson
suicideper100th
employrate
urbanrate
surarea
population
popden
corruptionindex
hdi
homicide
milexpprcntgdp;
run;
%macro kmean(K);
proc fastclus data=clustvar out=outdata&K. outstat=cluststat&K. maxclusters= &K. maxiter=300;
var incomeperperson
alcconsumption
armedforcesrate
breastcancerper100th
co2emissions
femaleemployrate
hivrate
internetuserate
lifeexpectancy
oilperperson
polityscore
relectricperperson
```

```
suicideper100th
employrate
urbanrate
surarea
population
popden
corruptionindex
hdi
homicide
milexpprcntgdp;
run;
%mend;
%kmean(1);
%kmean(2);
%kmean(3);
%kmean(4);
%kmean(5);
%kmean(6);
%kmean(7);
%kmean(8);
%kmean(9);
* extract r-square values from each cluster solution and then merge them to plot elbow curve;
data clus1;
set cluststat1;
nclust=1;
if _type_='RSQ';
keep nclust over_all;
run;
data clus2;
set cluststat2;
nclust=2;
```

```
if _type_='RSQ';
keep nclust over_all;
run;
data clus3;
set cluststat3;
nclust=3;
if _type_='RSQ';
keep nclust over_all;
run;
data clus4;
set cluststat4;
nclust=4;
if _type_='RSQ';
keep nclust over_all;
run;
data clus5;
set cluststat5;
nclust=5;
if _type_='RSQ';
keep nclust over_all;
run;
data clus6;
set cluststat6;
nclust=6;
if _type_='RSQ';
keep nclust over_all;
run;
data clus7;
set cluststat7;
nclust=7;
if _type_='RSQ';
```

```
keep nclust over_all;
run;
data clus8;
set cluststat8;
nclust=8;
if _type_='RSQ';
keep nclust over_all;
run;
data clus9;
set cluststat9;
nclust=9;
if _type_='RSQ';
keep nclust over_all;
run;
data clusrsquare;
set clus1 clus2 clus3 clus4 clus5 clus6 clus7 clus8 clus9;
run;
* plot elbow curve using r-square values;
symbol1 color=blue interpol=join;
proc gplot data=clusrsquare;
plot over_all*nclust;
Run;

*******************************************************************************
further examine cluster solution for the number of clusters suggested by the elbow curve
*******************************************************************************
* plot clusters for 3 cluster solution;
proc candisc data=outdata3 out=clustcan;
class cluster;
var incomeperperson
alcconsumption
```

```
armedforcesrate
breastcancerper100th
co2emissions
femaleemployrate
hivrate
internetuserate
lifeexpectancy
oilperperson
polityscore
relectricperperson
suicideper100th
employrate
urbanrate
surarea
population
popden
corruptionindex
hdi
homicide
milexpprcntgdp;
run;
proc sgplot data=clustcan;
scatter y=can2 x=can1 / group=cluster;
run;
* validate clusters on alcconsumption;
* first merge clustering variable and assignment data with alcconsumption data;
data alcconsumption_data;
set clus_train;
keep idnum alcconsumption;
run;
proc sort data=outdata3;
```

```
by idnum;
run;
proc sort data=alcconsumption_data;
by idnum;
run;
data merged;
merge outdata3 alcconsumption_data;
by idnum;
run;
proc sort data=merged;
by cluster;
run;
proc means data=merged;
var alcconsumption;
by cluster;
run;
proc anova data=merged;
class cluster;
model alcconsumption = cluster;
means cluster/tukey;
run;
```