ID740

Statistical Methods For Research I

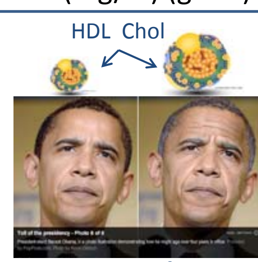
Exploratory Data Analysis (EDA)

Michael Griswold, PhD
Center of Biostatistics

THE UNIVERSITY OF MISSISSIPPI MEDICAL CENTER

A thinking question & some data

- Q?: How does HDL change as people age?
- Data: Jackson Heart Study (JHS)
  - 500 randomly sampled African American participants from Jackson, MS
  - Measures:
    - HDL: Fasting High Density Lipoprotein Chol (mg/dl) (good)
    - Age: in years (bad ☺)
  - What does the data look like?



Projection of aging after 4 years in office

Stata Command

```
list id hdl age in 1/5
```

Stata Output

```
     | id    hdl       age |
  1. |  1     55     54.99 |
  2. |  2     51     63.04 |
  3. |  3     62     59.23 |
  4. |  4     36     45.76 |
  5. |  5     38     40.95 |
```

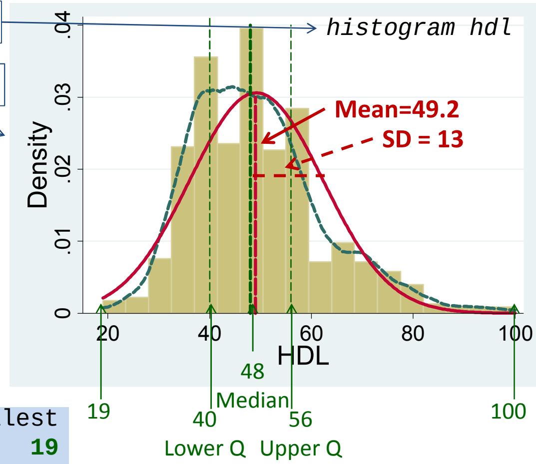
HDL

Stata Commands

Stata Output

summarize hdl, detail

```
HDL
Obs                         500
Mean                     49.176
Std. Dev.              13.04013

Variance               170.0451
Skewness               .8078762
Kurtosis               3.834666

        Percentiles        Smallest
  1%          26.5               19
  5%            32               21
 10%            35               21
 25%            40               23
 50%            48          Largest
 75%            56               91
 90%            68               94
 95%            74               96
 99%          88.5              100
```



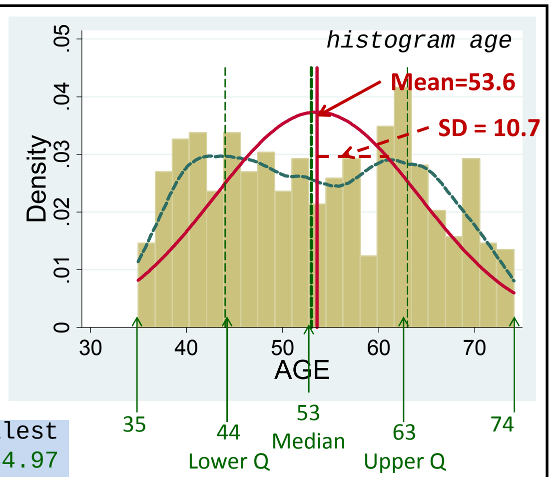
EDA Interpretations:
- Typical value approx 48 or 49
- slightly skewed, but fairly normal
- 50% of HDLs between 40 & 56
- 95% approx between 23 & 75

Q? How sure are we of these #s?

AGE

summarize age, detail

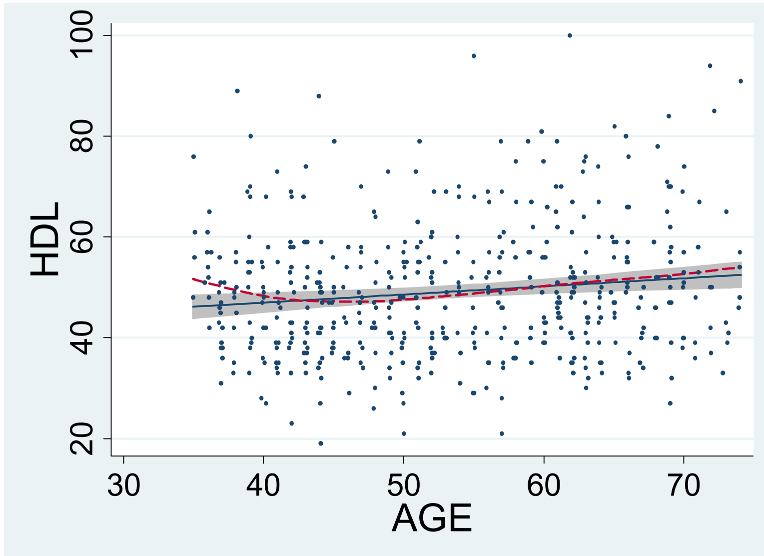
```
AGE
Obs                         500
Mean                    53.5834
Std. Dev.              10.68279

Variance                114.122
Skewness               .0745007
Kurtosis               1.829431

        Percentiles        Smallest
  1%         35.88            34.97
  5%        37.065            34.99
 10%        39.125            35.06
 25%         44.03             35.1
 50%        52.965          Largest
 75%        62.835            73.99
 90%         68.82            74.01
 95%         70.41            74.02
 99%        73.955            74.06
```



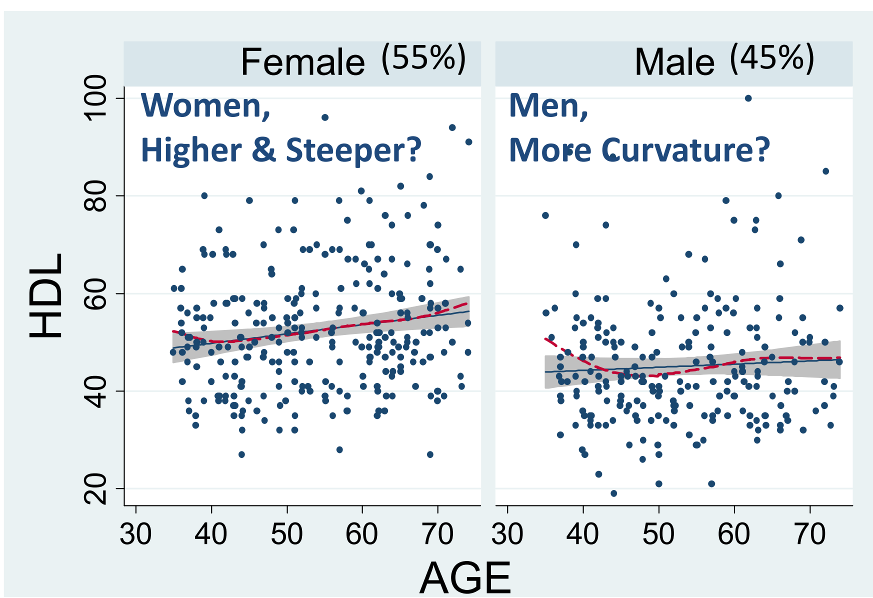
EDA Interpretations:
- Typical value approx 53
- More bimodal than normal
- 50% of Ages between 44 & 63
- 95% approx between 32 & 75??

Q? How sure are we of these #s?

Question: HDL vs AGE?



Question: HDL vs AGE – Gender Effects?

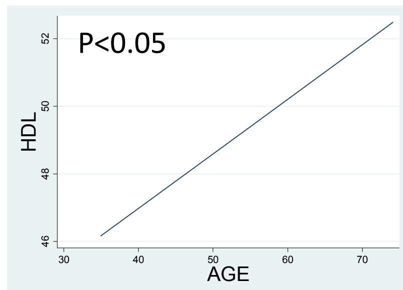


What do we usually get to see in lit?

- Unfortunately too common:

  "Age was significantly related to HDL, p<0.05"
- If you're lucky:

  "Age was significantly related to HDL; slope=0.16, p<0.05"
- If you're really lucky:



EDA lecture Overview

- Part 0: A thinking question – HDL vs Age
- Part 1: Intro and Basic EDA
  - Ideas & Basics
  - Some tools:
    - Histograms, Boxplots, Quantile-Quantile (Q-Q) plots
    - Scatterplots & Scatterplot matrices
    - Descriptive Statistics
- Part 2: Gaussian (Normal) distribution
  - Mean, variance, standard deviation.
  - Standardization / Z values.
- Part 3: Why can't we all just be normal?
  - Skewness, means vs medians, etc.

EDA Introduction: Doing research

- Research is a **process**. It is **iterative**. It is messy and full of **uncertainty**.
- Scientific Method: (inductive – from empirical data to theory):
  1. Define the question
  2. Gather information and resources (observe)
  3. Form hypothesis
  4. Perform experiment and collect data
  5. **Analyze this!** – *(this class)*
  6. Interpret data and draw conclusions that serve as a starting point for new hypothesis
  7. Publish results
  8. Retest (frequently done by other scientists)

Crawford S, Stucki L (1990), "Peer review and the changing research record", "J Am Soc Info Science", vol. 41, pp 223-228

Reproducible Research Paradigm

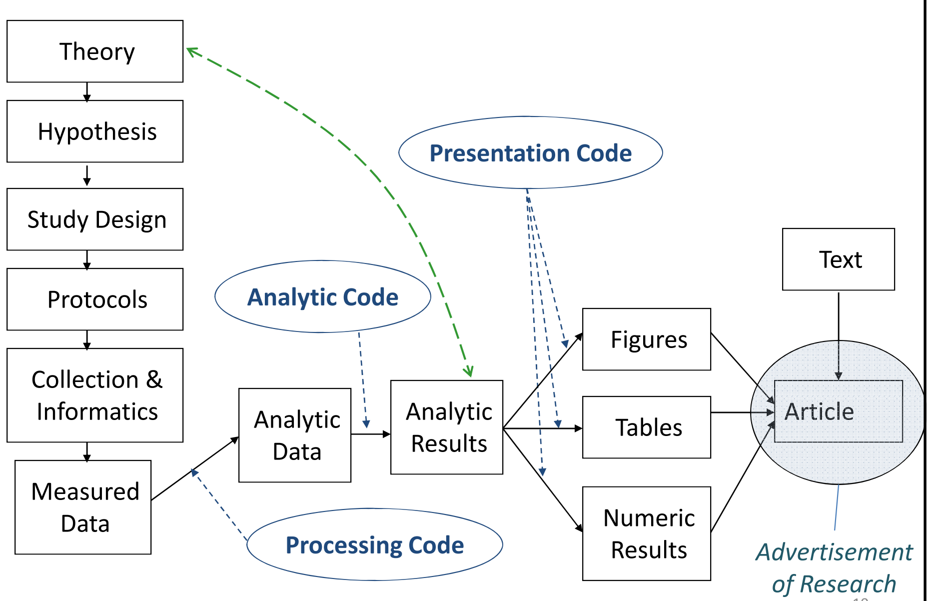


Figure by Peng & Griswold, Reproducible Research, JSM 2006

10

EDA Introduction: Data Analysis

- Data Analysis is also a **process**. It is **iterative**. It examines both patterns AND **uncertainty**.

General Data Analysis Steps:

1. Get the raw data in and clean it
2. **Exploratory Data Analysis (EDA)**
3. Initial estimation / inference
4. Determine primary models/output
5. Diagnostics

(lather, rinse, repeat)

Exploratory data analysis

John Wilder Tukey
1915-2000 (aged 85)

- "An approach to analysing data for the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses"
- Tukey held that too much emphasis in statistics was placed on hypothesis testing (confirmatory data analysis).
- Some objectives of EDA are to:
  - Suggest hypotheses about the causes of observed phenomena
  - Assess assumptions on which statistical inference will be based
  - Support the selection of appropriate statistical tools and techniques
  - Provide a basis for further data collection
- Many EDA techniques have been adopted into data mining

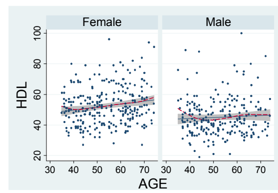
• *Conversation with John W. Tukey and Elizabeth Tukey*, Luisa T. Fernholz and Stephan Morgenthaler, *Statistical Science*, Volume 15, Number 1 (2000), 79-94.
• *Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught*, John W. Tukey *The American Statistician*, *34(1)*, (Feb., 1980), pp. 23-25.

Exploratory data analysis

Tukey argued data analysis involves 2 phases

1. **Exploratory** data analysis
   – Used to understand the data
     • to see patterns in the data
     • to find violations of statistical assumptions
   – Mostly graphical
   – Helps develop hypotheses
   – Data driven
2. 'Confirmatory' data analysis
   – Inferential Statistics
     • Estimates, confidence intervals, hypotheses tests, etc.
   – EDA and theory driven

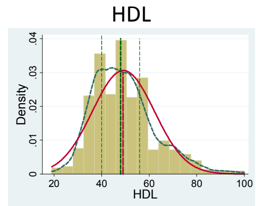
Recall:



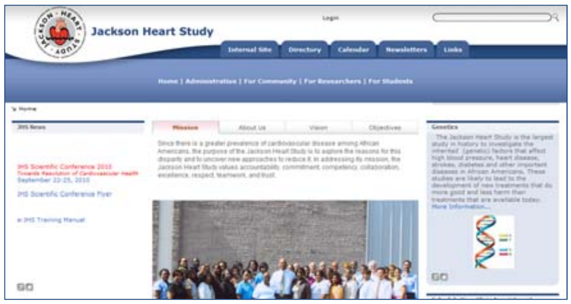
EDA helps Answer Questions

- EDA questions
  – What is a typical value?
  – What is the uncertainty / how spread out are the data?
  – What is a good distributional fit for the data?
  – What values may be inappropriate?
  – What are the relationships between two attributes?
  – Etc.
- You must build an understanding from EDA to effectively use any confirmatory statistical tools
  – Never run a regression without plotting the data first
  – Etc.
- The specific form of EDA depends on the data and questions at hand.

Some example data: JHS



- Jackson Heart Study:
- http://jhs.jsums.edu/jhsinfo/
- Purpose: explore the reasons for CVD disparity and uncover new approaches to reduce it.



Some example data: JHS

Classroom dataset

- 500 deidentified randomly sampled participants
- Random noise added to variables
- Useful for getting to know JHS data & illustrating methods
- Can't be used for publishing but can request actual JHS research datasets

Variables

Name	Label
id	Subject ID
age	Age(yrs) at Baseline Clinic Visit
bmi	Body Mass Index Visit 1
bmigp	Body Mass Index Grouping Visit 1
glucose	Fasting Glucose
diabetes	Diabetes Status (Type I or II)
bpmed	Antihypertensive Medication
hdl	Fasting High Density Lipoprotein Choleste...
ldl	Fasting Low Density Lipoprotein Cholester...
tg	Fasting Triglyceride Level (mg/dl)
educgp	Education Level Group 2
chd	CHD
alcohol	Alcohol drinking in the past 12 month (Y/N)
egfr	eGFR based on the MDRD Formula
leptin	Leptin: The concentration of leptin in ng/...
systolic	19: Systolic (Computed Net Average)
diastolic	20: Diastolic(Computed Net Average)
currsmoke	3: Do you now smoke cigarettes
waist	3a: Waist to nearest cm
male	Gender==Male

JHS example data: First 10 obs

id	age	bmi	bmigp	glucose	diabetes	bpmed	hdl	ldl	tg
1	55.13	25.97	1	84	0	1	55	139	142
2	63.03	62.61	2	116	0	1	52	149	69
3	59.04	26.61	1	94	0	0	62	138	72
4	45.99	32.97	2	92	0	0	36	124	79
5	40.95	29.48	1	80	0	0	38	118	71
6	56.96	23.76	0	89	0	0	42	91	85
7	36.05	24.01	0	80	0	0	51	186	140
8	46.12	31.92	2	106	0	0	37	135	78
9	63.03	28.85	1	119	1	1	39	145	172
10	50	35	2	88	0	0	48	84	49

educgp	chd	alcohol	egfr	leptin	systolic	diasto~c	currsm~e
4	0	1	72.8	4.5	113	79	
2	0	0	87.09	52.1	137	75	
2	0	1	78.24	29.3	111	54	
2	0	1	110.92	22.9	107	79	N
3	0	1	84.26	23.2	121	81	
4	0	0	95.26	2.9	126	87	
3	0	1	94.68	5.3	119	90	
4	0	1	92.84	49.4	110	76	
1	0	0	77.23	17.2	144	78	N
2	0	0	91.26	41.3	118	79	

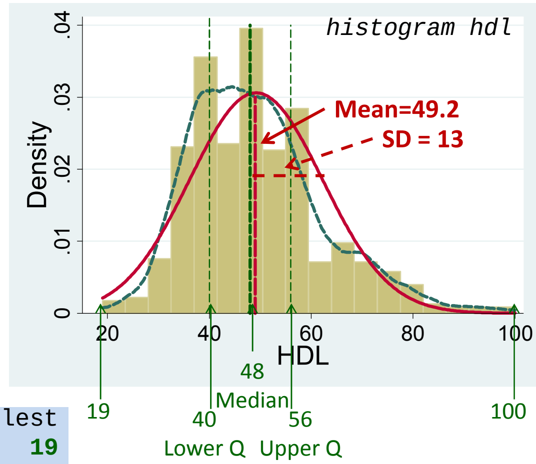
Some JHS Questions

- Some EDA questions
  - What is a typical value for HDL?
  - How spread out are the HDL?
  - What is a good distributional fit for HDL?
  - What values of HDL may be inappropriate?
  - Is there a relationship between HDL & Age?
  - Do these things vary across Gender?

HDL

summarize hdl, detail

```
HDL
Obs                      500
Mean                  49.176
Std. Dev.           13.04013

Variance            170.0451
Skewness            .8078762
Kurtosis            3.834666

      Percentiles        Smallest
 1%            26.5            19
 5%              32            21
10%              35            21
25%              40            23
50%              48       Largest
75%              56            91
90%              68            94
95%              74            96
99%            88.5           100
```



EDA Interpretations:
- Typical value approx 48 or 49
- slightly skewed, but fairly normal
- 50% of HDLs between 40 & 56
- 95% approx between 23 & 75

"Short" Digression on:

– Measures of Central Tendency –

(Means & Medians)

– Sample Versus Population –

– Distributional Shapes –

Sample Mean $\overline{X}$

The Average or Arithmetic Mean

Shorthand notation

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Add up data, then divide by sample size (*n*)
  - The sample size *n* is the number of observations (pieces of data)
- Example: *n* = 5 Systolic blood pressures (mmHg)

  $X_1 = 120$
  $X_2 = 80$
  $X_3 = 90$
  $X_4 = 110$
  $X_5 = 95$

$$\boxed{\overline{X} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99\,\text{mmHg}}$$

- Sensitive to extreme values
  - One data point could make a big change in sample mean
- Why is it called the *sample* mean?
  - To distinguish it from population mean

Population versus *Sample*

- *Population*—The entire group you want information about
  - For example: The blood pressure of all 18-year-old female college students in the United States
- *Sample*—A part of the population from which we actually collect information and draw conclusions about the whole population
  - For example: Sample of N=5 blood pressures 18-year-old female college students in the United States
- NOTE: The sample mean $\overline{X}$ is not the population mean μ
  - More on this in upcoming sampling distribution and statistical inference lectures

Population versus *Sample*

Population
Population mean: μ
Population variability: σ^2

Sample
Sample mean: $\overline{X}$
Sample var: s^2

- $\overline{X}$ & s^2 are just estimates of μ & σ^2
- Q: How good are these estimates?
- A: It depends…
  (upcoming Stat Inf lecture)

Sample Median

- The median is the middle number
  - (50% lie above & 50% lie below: *ie 50th percentile*)
  - Example

$X_1 = 120$
$X_2 = 80$
$X_3 = 90$
$X_4 = 110$
$X_5 = 95$

Sorted Values

80 90 95 110 120

- The sample median is not sensitive to extreme values
  - Example: If 120 became 200, the median would remain the same, but the mean would increase from 99 to 115.

80 90 95 110 200

Sample Median

- If the sample size is an even number, median is an average (recall 50% above & below)

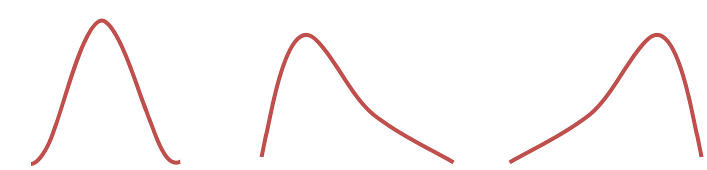
80 90 95 | 110 120 125

Median

$$\frac{95+110}{2} = 102.5\ mmHg$$

Shapes of the Distribution

- Three common shapes of data distributions:



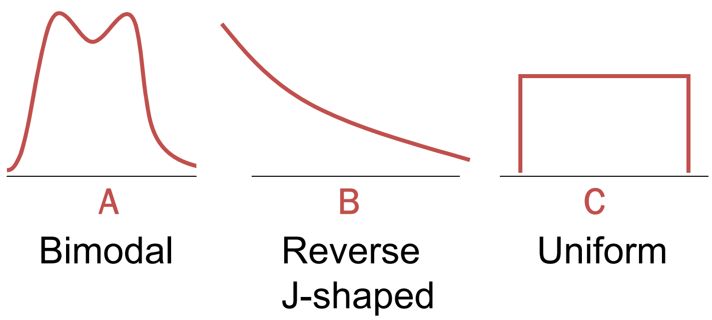
A
Symmetrical and bell shaped

B
Positively skewed or skewed to the right

C
Negatively skewed or skewed to the left

Shapes of the Distribution

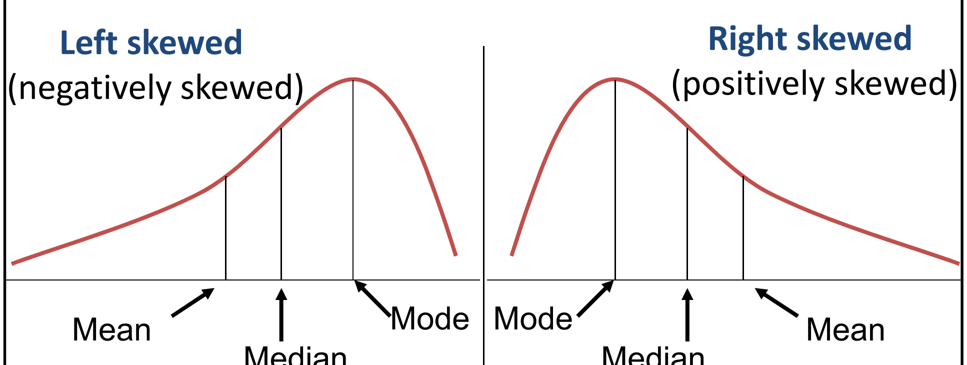
- Three less common shapes of frequency distributions:



Distribution Characteristics

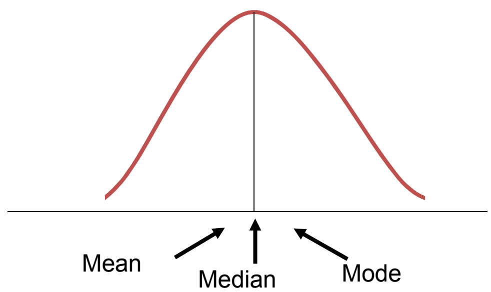
"Typical" values

- Mode: Peak(s)
- Median: Equal areas point
- Mean: Balancing point



Shapes of Distributions

- **Symmetric** (Right and left sides are mirror images)
  - Left tail looks like right tail
  - Mean = Median = Mode
- Normal distribution is a Symmetric Distribution



The Normal Distribution

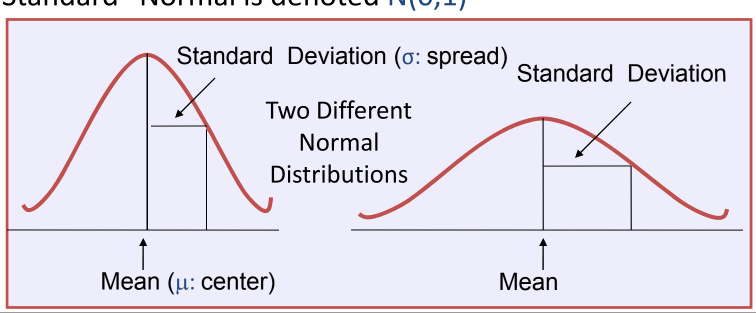
Q : Is every variable normally distributed?

A : Absolutely not

Q : Then why do we spend so much time on it?

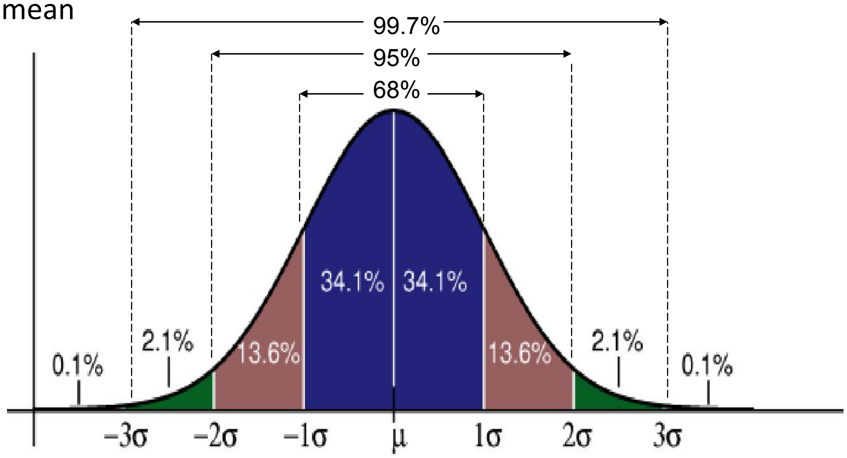
A : Some variables are normally distributed; a bigger reason is the "Central Limit Theorem" (more later)

- We can describe any Normal distribution with just 2 #'s: N(μ, σ^2)
- A "Standard" Normal is denoted N(0,1)



Normal Distribution: the 68-95-99 Rule

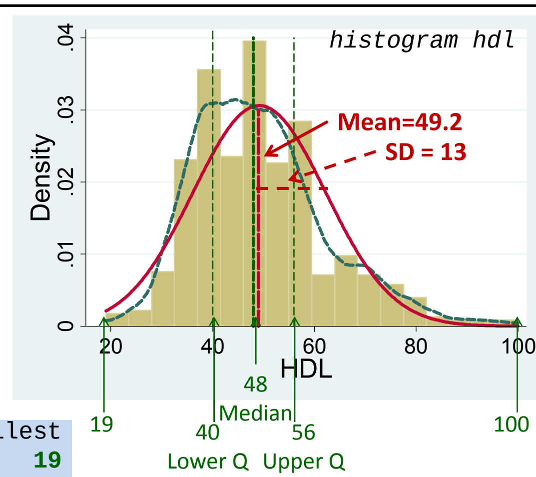
- **68%** of the data fall within **one** standard deviation of the mean
- **95%** of the data fall within **two** standard deviations of the mean
- **99.7%** of the data fall within **three** standard deviations of the mean



HDL

summarize hdl, detail

```
HDL
Obs                        500
Mean                    49.176
Std. Dev.             13.04013

Variance              170.0451
Skewness              .8078762
Kurtosis              3.834666

        Percentiles        Smallest
 1%          26.5              19
 5%            32              21
10%            35              21
25%            40              23
50%            48         Largest
75%            56              91
90%            68              94
95%            74              96
99%          88.5             100
```



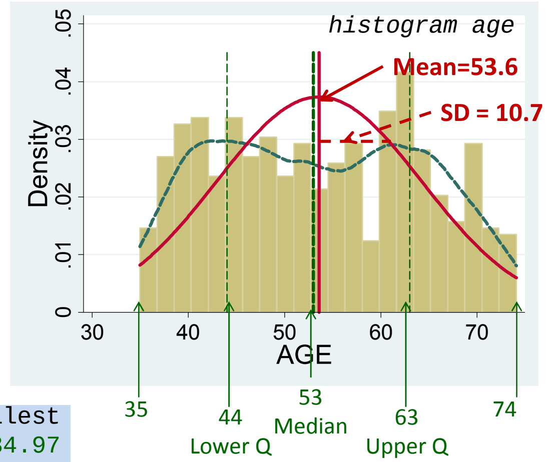
EDA Interpretations:
- Typical value approx 48 or 49
- slightly skewed, but fairly normal
- 50% of HDLs between 40 & 56
- 95% approx between 23 & 75

AGE

summarize age, detail

AGE	
Obs	500
Mean	**53.5834**
Std. Dev.	**10.68279**
Variance	114.122
Skewness	.0745007
Kurtosis	1.829431

	Percentiles	Smallest
1%	35.88	34.97
5%	37.065	34.99
10%	39.125	35.06
25%	44.03	35.1
50%	**52.965**	Largest
75%	62.835	73.99
90%	68.82	74.01
95%	70.41	74.02
99%	73.955	74.06



EDA Interpretations:
- Typical value approx 53
- More bimodal than normal
- 50% of Ages between 44 & 63
- 95% approx between 32 & 75**??**

Q? How sure are we of these #s?

Potential Dangers with Assumptions & no EDA

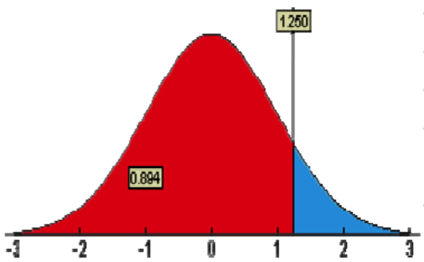
- The rule says that ***if a population is normally distributed***, then approximately 95% of the population will be within 2 SD of μ
- It doesn't guarantee that exactly 95% of your **sample** of data will fall within 2 SD of $\bar{x}$

34

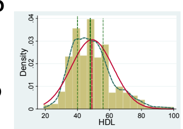
Last Digression: Standard Normal Scores

- Idea: How many standard deviations away from the mean are you?
- Standard Score (Z) = $\dfrac{\text{Value} - \text{mean}}{\text{Standard deviation}}$

Pr(Z<1.25) = 89.4% (area under the curve)

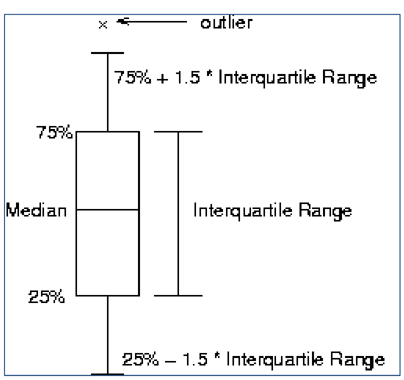


- Why is this really useful?
- Consider HDL ~ N(49.2, 13)
- Q: Pr(HDL > 65.5 mg/dl)?
- Standardize HDL value
  Z = (65.5-49.2)/13 ≈ 1.25
- Pr(HDL > 65.5)
  = Pr(Z > 1.25) = 89.4%

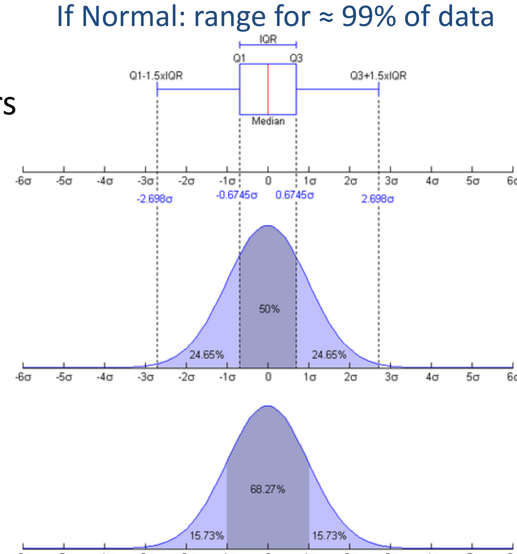


Back to EDA: Boxplots (HDL example)

- 5 # distb summary:
  - Q1, Q2 (med), Q3
  - Lower, upper whiskers
- Visualize Distb.



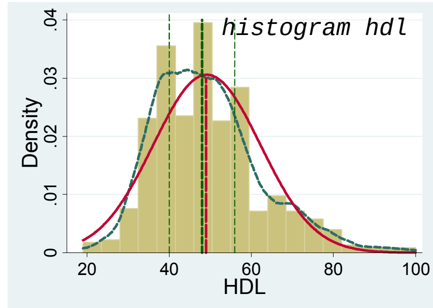
Why 1.5*IQR?
If Normal: range for ≈ 99% of data



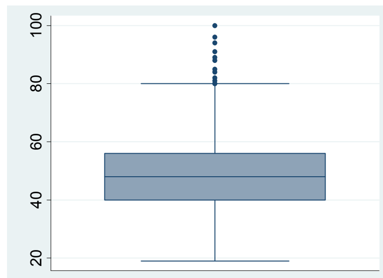
HDL

summarize hdl, detail

```
HDL
Obs                     500
Mean                 49.176
Std. Dev.          13.04013

Variance           170.0451
Skewness           .8078762
Kurtosis           3.834666

        Percentiles      Smallest
 1%          26.5              19
 5%            32              21
10%            35              21
25%            40              23
50%            48         Largest
75%            56              91
90%            68              94
95%            74              96
99%          88.5             100
```



graph box hdl

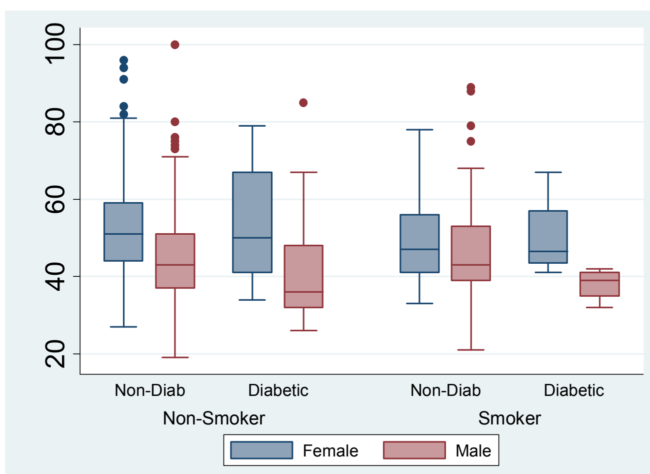


Q: Why is this useful?

HDL across subgroups

- Q: How do the values of HDL compare across different subgroups?

Boxplots can be useful EDA tools especially for comparisons across groups

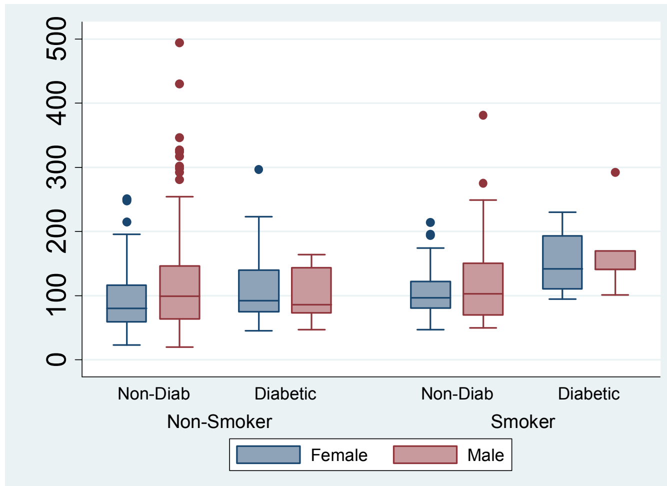


Triglycerides across subgroups

- Q: How do the values of TG compare across different subgroups?

Can you visualize the distributions?

Do you think TG is approx Normal?

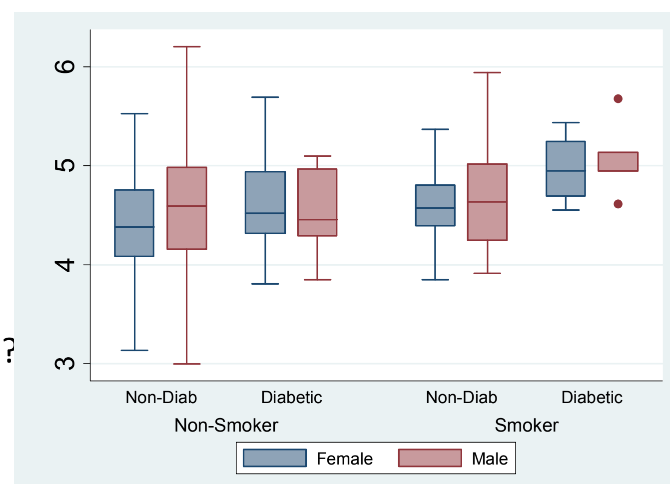


Transformations

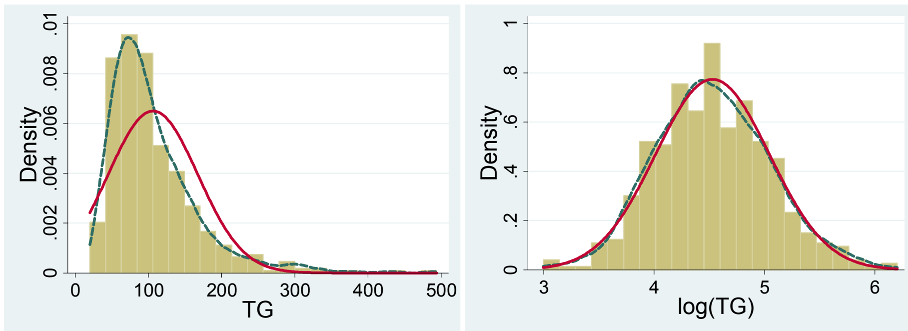
- Sometimes it can be useful to use a transformation to better fit distb assumptions

log(TG)

Do you think these look more Normal?



TG vs log(TG)



- Sometimes transformations are helpful and sometimes they are not.
- It all depends on the questions of interest
- “Congress is not interested in how many log(dollars) are spent on health care”

Summary

- Examine all your variables thoroughly and carefully before you begin analysis
- Use visual displays whenever possible
- Use EDA to show the story that addresses good questions

Recommended Reading

- Anything by Tukey, especially Exploratory Data Analysis (Tukey, 1997)
- Anything by Cleveland, especially Visualizing Data (Cleveland, 1993)
- Visual Display of Quantitative Information (Tufte, 1983)