Accelerating Impact: Immersive Summer Bootcamp in Implementation Science and Biostatistics

Georgian Implementation Science Fogarty Training (GIFT) Program

Ilia State University & Yale University

Yale Center for Analytical Sciences

Descriptive Statistics

https://www.youtube.com/watch?v=PbODigCZqL8


EXTRAPOLATED
DIDN'T RANDOMIZE
R.I.P
ACCEPTED THE ALTERNATIVE
Deadly Sins

Table 1. Baseline Characteristics of the Patients with Asthma.*

Characteristic	First-Line Controller Therapy Trial		Add-on Therapy Trial	
	LTRA (N=148)	Inhaled Glucocorticoid (N=158)	LTRA (N=170)	LABA (N=182)
Mean age — yr	47.6±16.5	44.1±16.4	51.0±16.0	49.7±16.1
Age group — no. of patients (%)				
<16 yr	1 (1)	4 (3)	0	0
16–25 yr	17 (12)	17 (11)	12 (7)	17 (9)
26–35 yr	22 (15)	30 (19)	24 (14)	31 (17)
36–45 yr	27 (18)	34 (22)	30 (18)	25 (14)
46–55 yr	30 (20)	33 (21)	32 (19)	39 (21)
56–65 yr	28 (19)	22 (14)	35 (21)	38 (21)
≥66 yr	23 (16)	18 (11)	37 (22)	32 (18)
Female sex — no. of patients (%)	73 (49)	83 (53)	109 (64)	111 (61)
Race — no. of patients (%)†				
White	144 (97)	153 (97)	168 (99)	178 (98)
Other or not known	4 (3)	5 (3)	2 (1)	4 (2)
Smoking status — no. of patients/total no. (%)				
Current smoker	37/147 (25)	30/155 (19)	29/168 (17)	31/180 (17)
Former smoker	54/147 (37)	54/155 (35)	63/168 (38)	75/180 (42)
Never smoked	56/147 (38)	71/155 (46)	76/168 (45)	74/180 (41)
Current smoker >45 yr of age	15/147 (10)	11/155 (7)	16/168 (10)	16/180 (9)
Peak-expiratory-flow reversibility — % (no. of patients tested)	9.2±10.7 (128)	8.7±9.2 (142)	9.0±10.1 (163)	8.3±9.6 (170)
Time since asthma diagnosis — yr				
Median	8.5	10	11	11
Interquartile range	3–19	4–16.5	5–22.5	6–21
Assigned therapy — no. of patients/total no. (%)				
Montelukast	127/143 (89)		158/166 (95)	
Zafirlukast	16/143 (11)		8/166 (5)	
Beclomethasone		146/157 (93)		
Budesonide		8/157 (5)		
Fluticasone		3/157 (2)		
Salmeterol				167/181 (92)‡
Formoterol				14/181 (8)†
Dose of glucocorticoid — μg/day (no. of patients)§			425±351 (170)	451±390 (182)

- **Results:** The mean (± standard deviation [SD]) for total resources utilized at 90 days for in-network plus out-of-network services was less for the participants in the SC + CCTA group (**$10,134**; SD ± $14,239) versus the SC-only group **($16,579**; SD ± $19,148; p = **0.144**), as was the median for the SC + CCTA ($4,288) versus SC only ($12,148; p = 0.652; median difference = –$1,291; 95% confidence interval [CI] = –$12,219 to $1,100; p = 0.652)...

  Conclusions: Adding CCTA to the current ED risk stratification of ACPSs resulted in no difference in the quantity of resources utilized, but an increased diagnosis of CAD, and significantly less recidivism and rehospitalization over a 90-day follow-up period.

Common Biostatistical Problems and the Best Practices that Prevent Them

1. p-values for establishing negative results.
2. Misleading and vague phrasing.
3. Speculation about low power.
4. Exclusive reliance on intent-to-treat analysis.
5. Reliance on omnibus tests.
6. Overuse of multiple comparisons adjustments.
7. Entangled outcomes and predictors.

Bachetti, http://www.ctspedia.org/do/view/CTSpedia

…but first

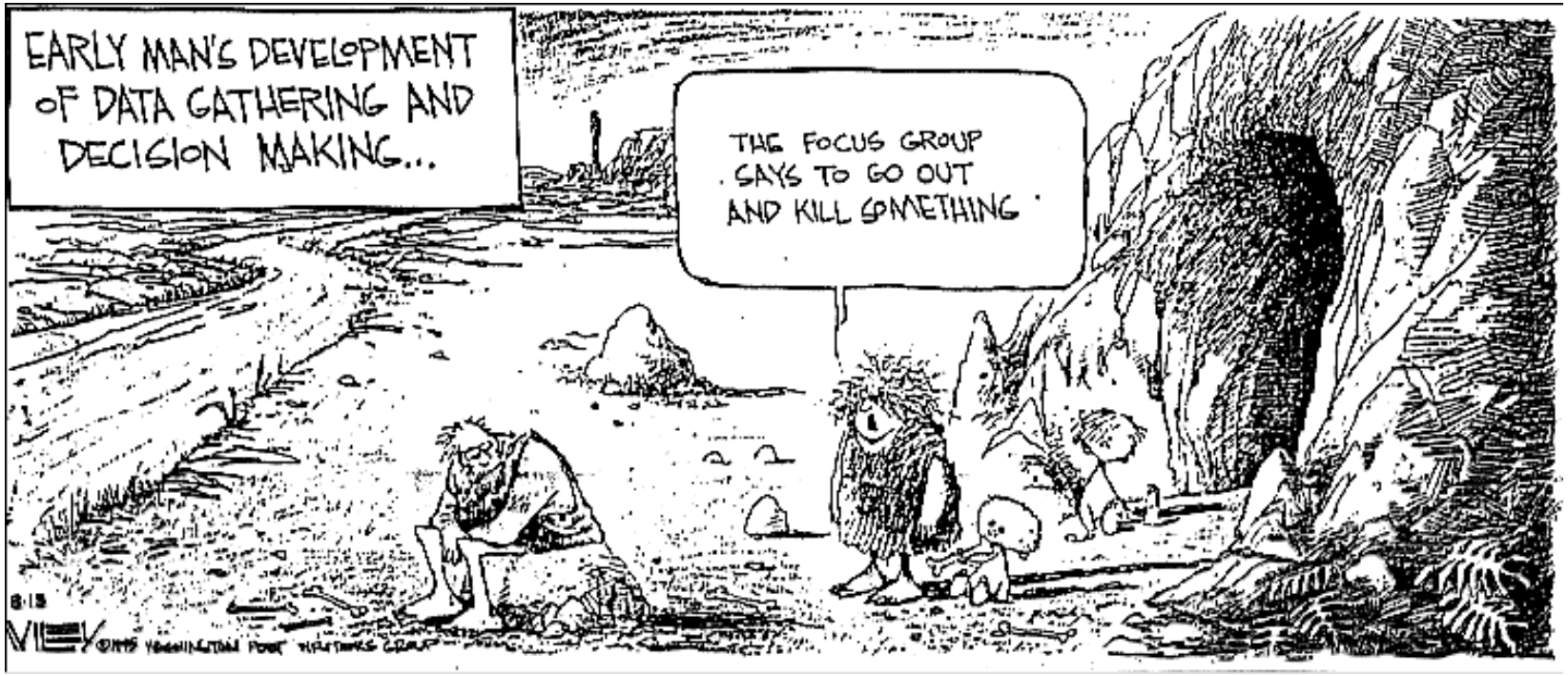
“ἀρχὴ παιδεύσεως ἡ τῶν ὀνομάτων ἐπίσκεψις”

The beginning of a right education is the examination of words

Antisthenis 5th – 4th century BC (in Epictitos)

Statistics Definition

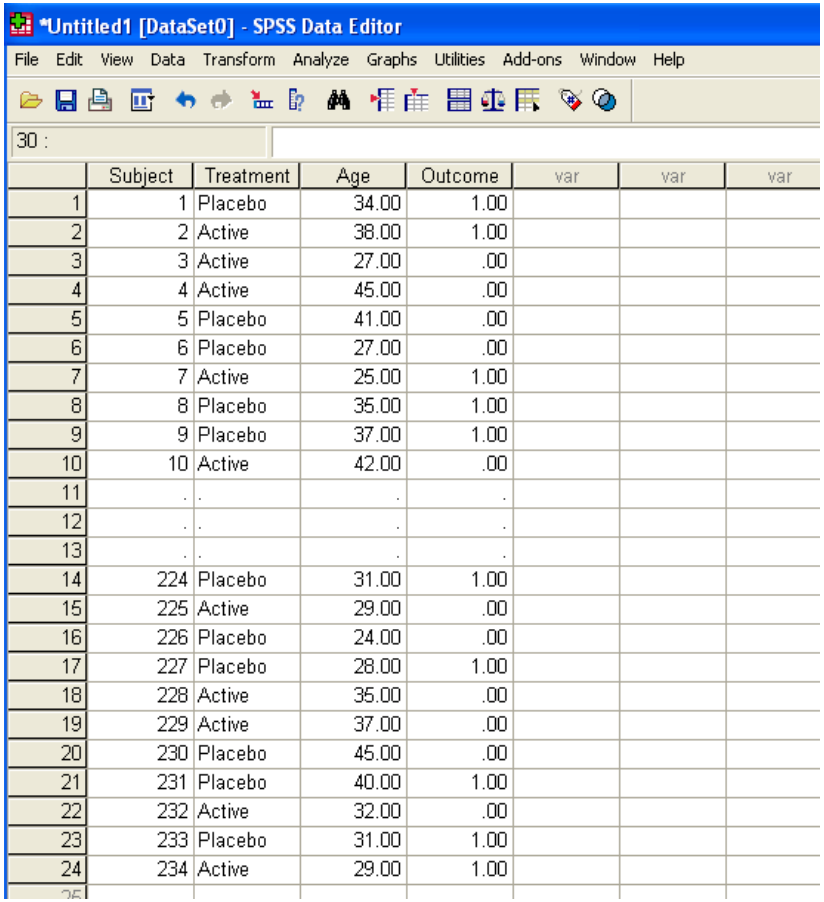
- The science and art of ***collecting***, ***summarizing*** and ***analyzing*** data with random variation


EARLY MAN'S DEVELOPMENT OF DATA GATHERING AND DECISION MAKING...
THE FOCUS GROUP SAYS TO GO OUT AND KILL SOMETHING
8-15

Study Results - 1

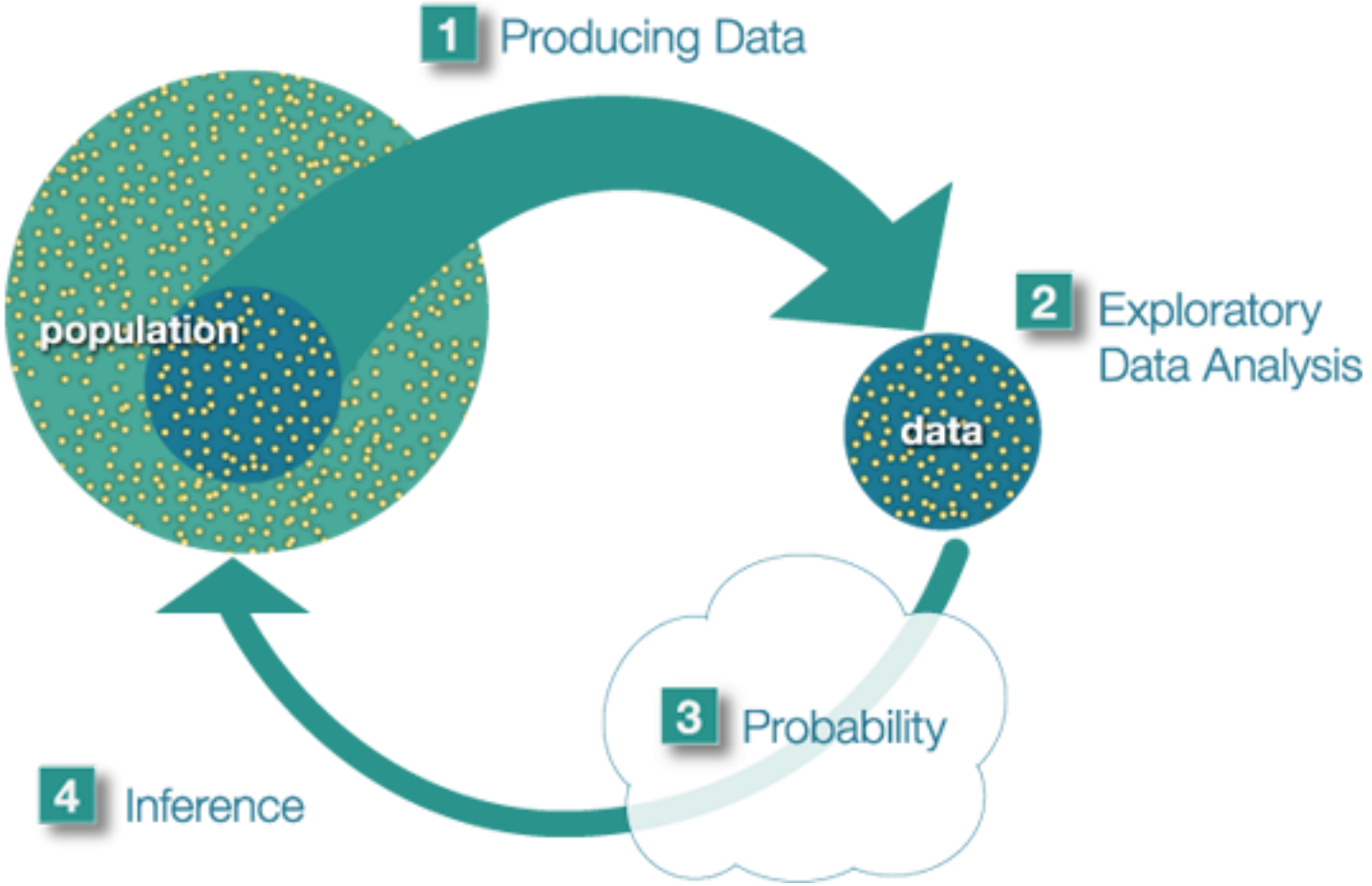
- A patient with a UTI is given a drink of cranberry juice and the infection goes away.
- Do you conclude that cranberry juice is the cure for UTI?

Study Results - 2



	Subject	Treatment	Age	Outcome	var	var	var
1	1	Placebo	34.00	1.00			
2	2	Active	38.00	1.00			
3	3	Active	27.00	.00			
4	4	Active	45.00	.00			
5	5	Placebo	41.00	.00			
6	6	Placebo	27.00	.00			
7	7	Active	25.00	1.00			
8	8	Placebo	35.00	1.00			
9	9	Placebo	37.00	1.00			
10	10	Active	42.00	.00			
11	.	.	.	.			
12	.	.	.	.			
13	.	.	.	.			
14	224	Placebo	31.00	1.00			
15	225	Active	29.00	.00			
16	226	Placebo	24.00	.00			
17	227	Placebo	28.00	1.00			
18	228	Active	35.00	.00			
19	229	Active	37.00	.00			
20	230	Placebo	45.00	.00			
21	231	Placebo	40.00	1.00			
22	232	Active	32.00	.00			
23	233	Placebo	31.00	1.00			
24	234	Active	29.00	1.00			

The Big Picture



http://bolt.mph.ufl.edu/6050-6052/preliminaries/the-big-picture/

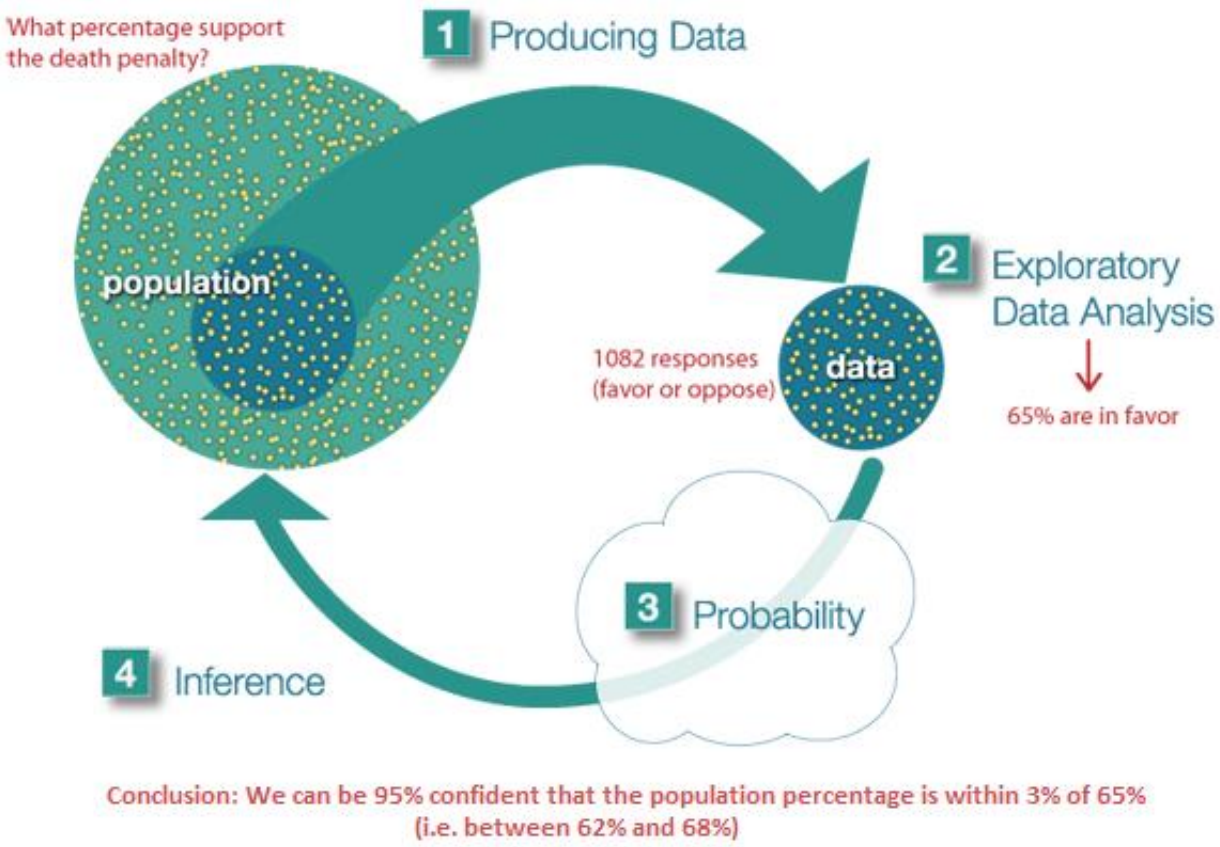
The Big Picture: Example

At the end of April 2005, a poll was conducted (by ABC News and the Washington Post), for the purpose of learning the opinions of U.S. adults about the death penalty.

1. Producing Data: A (representative) sample of 1,082 U.S. adults was chosen, and each adult was asked whether he or she favored or opposed the death penalty.

2. Exploratory Data Analysis (EDA): The collected data were summarized, and it was found that 65% of the sampled adults favor the death penalty for persons convicted of murder.

3 and 4. Probability and Inference: Based on the sample result (of 65% favoring the death penalty) and our knowledge of probability, it was concluded (with 95% confidence) that the percentage of those who favor the death penalty in the population is within 3% of what was obtained in the sample (i.e., between 62% and 68%). The following figure summarizes the example:



http://bolt.mph.ufl.edu/6050-6052/preliminaries/the-big-picture/

Intro to R

Types of Data

- Categorical (qualitative)
- Numerical (quantitative)

Categorical Data

- Values are categories
  - Dichotomous – two categories
    - Male/Female
    - Smoker/Non Smoker
  - Polytomous - >2 categories
    - Blood group
    - Non smoker/ex-smoker/light smoker/heavy smoker

Types of Categorical Data

- Nominal – no inherent order to the values
  - Blood type
  - Marital status

- Ordinal – values have an underlying order but numbers used are arbitrary
  - Smoking categories
  - Severity stages of cancer – I, II, III, IV

Numerical Data

- Data in which differences in values have meaning

Types of Numerical Data

- Discrete – observations can only take certain numerical (integer) values
  - Number of children
  - Number of visits
  - Number of heartbeats/min

- Continuous (interval)- observations are on a continuum – not restricted to integers – in theory could be recorded to infinite decimal places
  - Age
  - BP
  - Weight

Censored Data

- The only information about some of the values is that they are above or below a certain point
  - Eg. data reported as below detection limit
  - Eg. Time to event – mixture of continuous and categorical
    - Follow-up time
    - Indicator of event status
- Requires specific analytic methods

A physician groups patients with systolic blood pressure (SBP) as

1= no hypertension if SBP<120;

2= pre-hypertension if 120≤SBP<140;

3=hypertension if SBP≥140.

Coding SBP into these 3 values is an example of which type of data?

A. Discrete Numeric
B. Continuous
C. Nominal
D. Ordinal

Type of insurance is an example of which type of data?

A. Discrete Numeric
B. Ordinal
C. Nominal
D. Continuous

Describing Data - Numerical

- Measures of Central Tendency – where is the middle of the distribution
  - Mean (arithmetic)
  - Median
  - Mode
  - Geometric Mean

Arithmetic Mean

$$\overline{X} = \frac{\Sigma x}{n}$$

- Example
  - X: 100 125 104 116 135 120 115 118 124 113
    - $\Sigma x = 1170$ n=10

$$\frac{\Sigma x}{n} = \frac{1170}{10} = 117$$

- Disadvantage – sensitive to extreme values

Arithmetic Mean

$$\overline{X} = \frac{\Sigma x}{n}$$

- Example
  - X: 100 125 104 116 135 120 115 118 124 ~~113~~ 345
    - Σx = ~~1170~~ 1402 n=10

$$\frac{\Sigma x}{n} = \frac{1402}{10} = 140.2$$

- Disadvantage – sensitive to extreme values

Median

- The middle observation
- To calculate
  - Arrange observations from lowest to highest
  - For odd # of observations the median is the middle #
  - For even # of observations the median is the average of 2 center values
- Example
  - X: 100 104 113 115 116 118 120 124 125 135

5 6

$$M = \frac{116+118}{2} = 117$$

Median

- The middle observation
- To calculate
  - Arrange observations from lowest to highest
  - For odd # of observations the median is the middle #
  - For even # of observations the median is the average of 2 center values
- Example
  - X: 100 104 ~~113~~ 115 116 118 120 124 125 135 345

$$M = \frac{118+120}{2} = 119$$

Mode

- Most frequent #
- Not often used – can have multiple modes
- Example
  - X: 100 104 113 115 116 118 120 124 125 135

Geometric Mean

$$GM_{\bar{x}} = \sqrt[n]{\prod x}$$

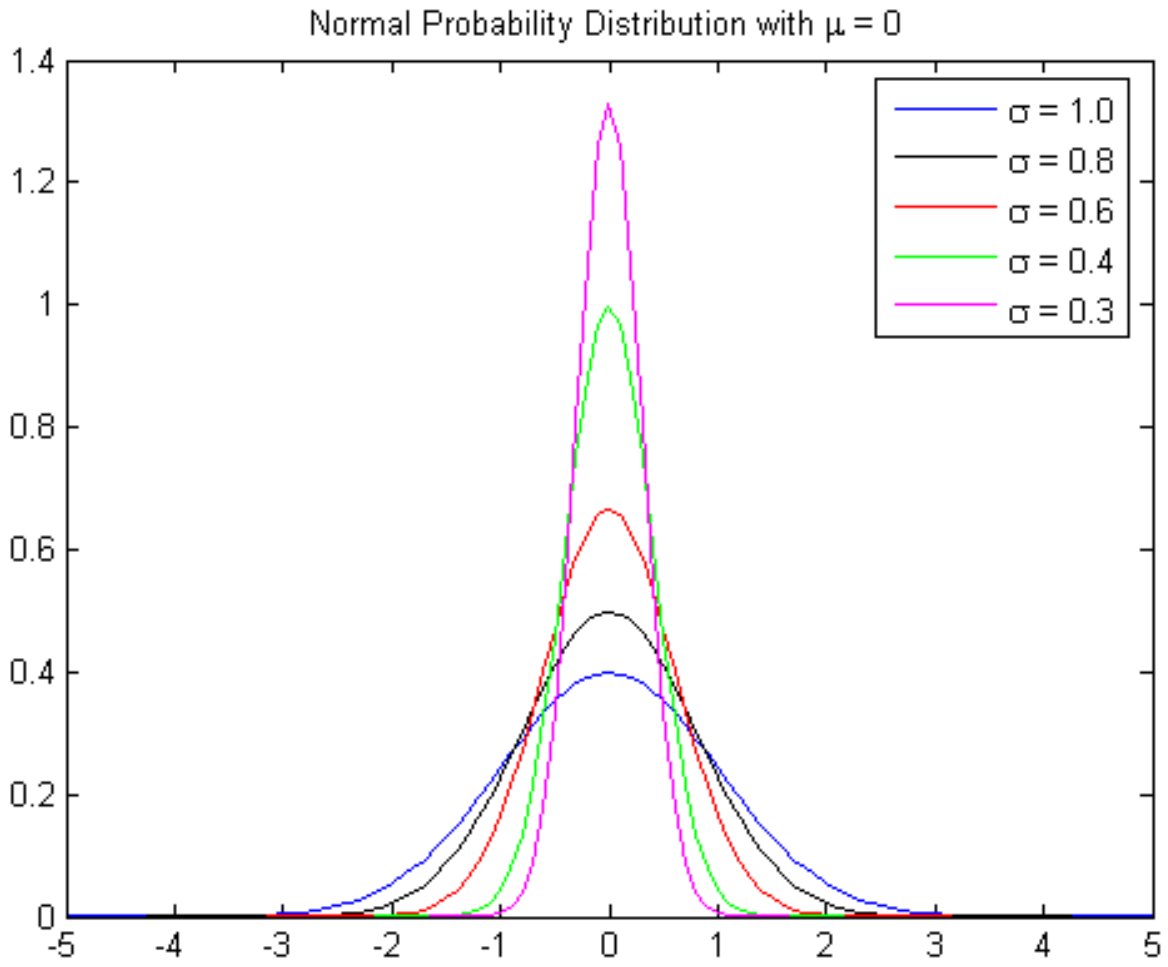
- Frequently used for data measured on a logarithmic scale
- Example
  - 100 104 ~~113~~ 115 116 118 120 124 125 135 345

$$GM_{\bar{x}} = \sqrt[10]{100 \times 104 \times 115 \times \ldots 345} = 130.37$$

Which measure of Central Tendency do you use?

- Mean – symmetric distribution
- Median – to describe skewed distribution
- Mode – for a multimodal distribution
- Geometric Mean – logarithmic distribution

Measures of Spread



Measures of Spread

- Range
- Standard Deviation
- Coefficient of Variation
- Percentile
- Interquartile Range

Standard Deviation

- How observations cluster around the mean
- Mean Deviation

$$\frac{\sum(x-\bar{X})}{n}$$

  - Always = 0
- Population Variance

$$\sigma^2 = \frac{\sum(x-\bar{X})^2}{n}$$

Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n}}$$

Sample Standard Deviation

- Variance

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

- Standard Deviation

$$s = \sqrt{\frac{\Sigma(x - \bar{X})^2}{n-1}}$$

Standard Deviation Example

x	$(x-\bar{X})$	$(x-\bar{X})^2$
100	-17	289
125	8	64
104	-13	169
116	-1	1
135	18	324
120	3	9
115	-2	4
118	1	1
124	7	49
113	4	16
117		926

$$s^2 = \frac{926}{9} = 102.9$$

$$s = \sqrt{102.9} = 10.14$$

Properties of SD

- At least 75% of values always lie b/w mean±2SD
- If the population distribution is bell-shaped
  - 67% of obs lie b/w ±1SD
  - 95% of obs lie b/w ±2SD
  - 99.7% of obs lie b/w ±3SD

Coefficient of Variation

- Relative measure of standard deviation
- Can compare spread b/w variables not measured on the same scale

$$CV = \frac{s}{\bar{X}} \times 100\%$$

$$CV = \frac{10.14}{117} = 8.7\%$$

Percentiles

- A value below which a certain percentage of the values occur
  - Example: Median – 50^{th} percentile

Interquartile Range

- Interquartile Range – difference between the 25^{th} and 75^{th} percentile
  - Calculate the % of (sample size + 1)

$$25\% \times 11 = 2.75$$

$$75\% \times 11 = 8.25$$

  - X: 100 104 113 115 116 118 120 124 125 135

$$25th = 104 + 0.75 \times (113 - 104)$$

$$75th = 124 + 0.25 \times (125 - 124)$$

Which to use

- SD when reporting mean
- Percentile or IQR when reporting median
- Range to show existence of extreme values
- CV when comparing spread of 2 numerical variables measured on different scale

Introduce the study and dataset

MAYBE AN R EXAMPLE HERE TO GENERATE ALL THE DISTRIBUTION PARAMETERS

Describing Data - Categorical

- Proportion – number of obs with a given characteristic divided by the total number of observations
  - Example : 300 obese adolescents have an OGTT, 75 have impaired glucose tolerance

$$\frac{a}{a+b} = \frac{75}{300} = 0.25$$

Describing Data – Categorical

- Ratio – number with divided by the number without

$$\frac{a}{b} = \frac{75}{225} = 0.33$$

Describing Data – Categorical

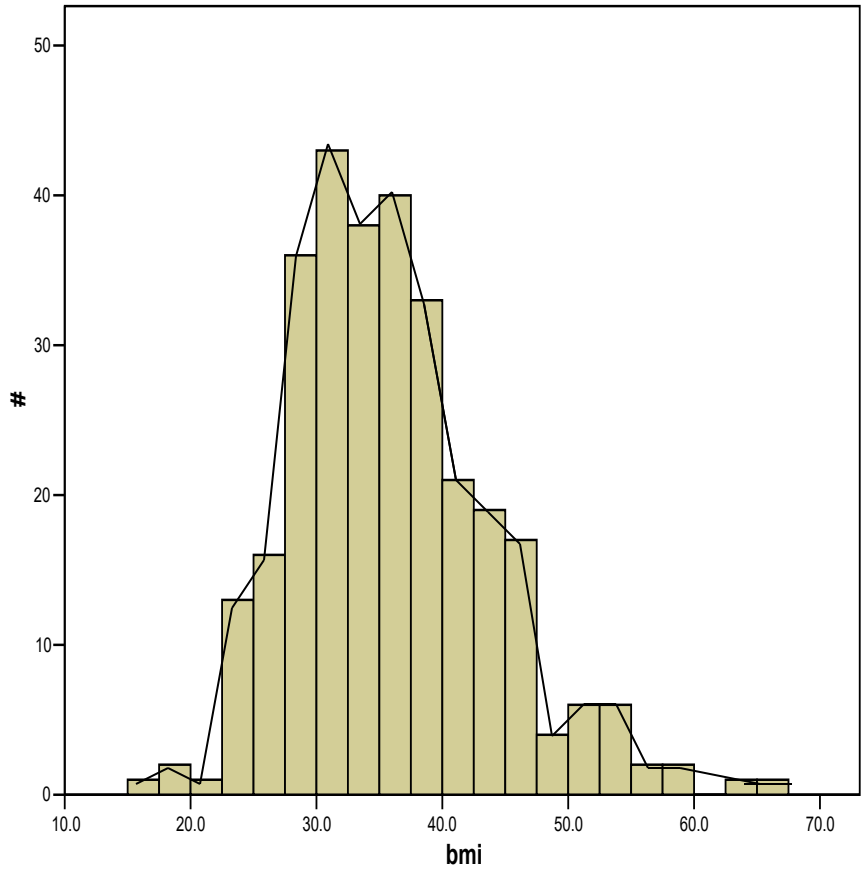
- Rates – proportion x base for a given period of time
  - Example: Follow 300 NGT children for 1 y and 10 develop IGT

$$\frac{10}{300} \times 1000 = 33.3 \, per 1000 \, py$$

Describing Data - Graphs

- Histograms
- Stem and Leaf Graph
- Box-plot

Histogram



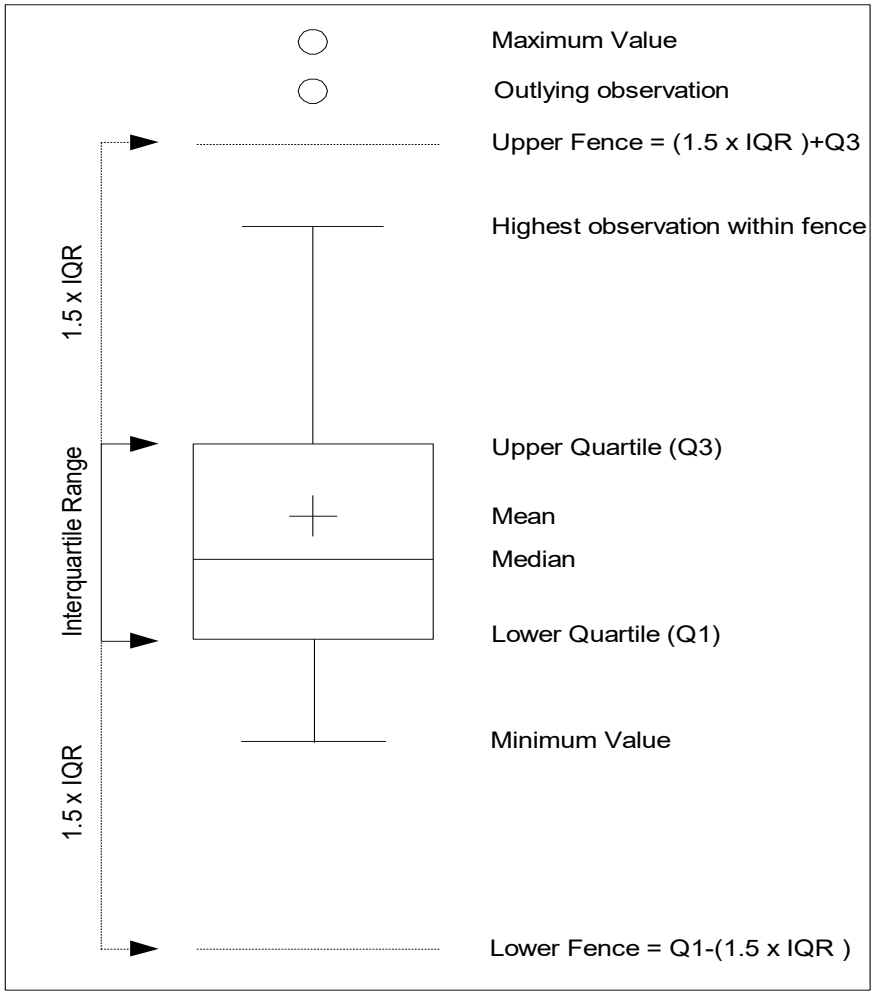
- Area of each rectangle represents frequency of occurrence
  - Should therefore have equal intervals
- If join midpoints of the tops of rectangles – frequency polygon

Stem and Leaf Plot

```
bmi Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00        1 .  &
    14.00        2 .  233344
    52.00        2 .  55666777777888899999999
    81.00        3 .  000000011111122222222223333344444444
    73.00        3 .  55555555666666667777777888899999999
    40.00        4 .  0000011112333334444
    21.00        4 .  5555666789
    12.00        5 .  0034&
     6.00 Extremes    (>=56)

 Stem width:   10
 Each leaf:       2 case(s)


 & denotes fractional leaves.
```

- Modification of histogram allows all actual values to be shown

Boxplot



- Box represents 25th and 75th percentiles
- Center line is Median
- '+' is the mean
- O is outlying observation (more than 1.5 IQR from 25 or 75th percentiles)

MAYBE AN R EXAMPLE HERE TO GENERATE EXAMPLES USING CATEGORICAL VARIABLES