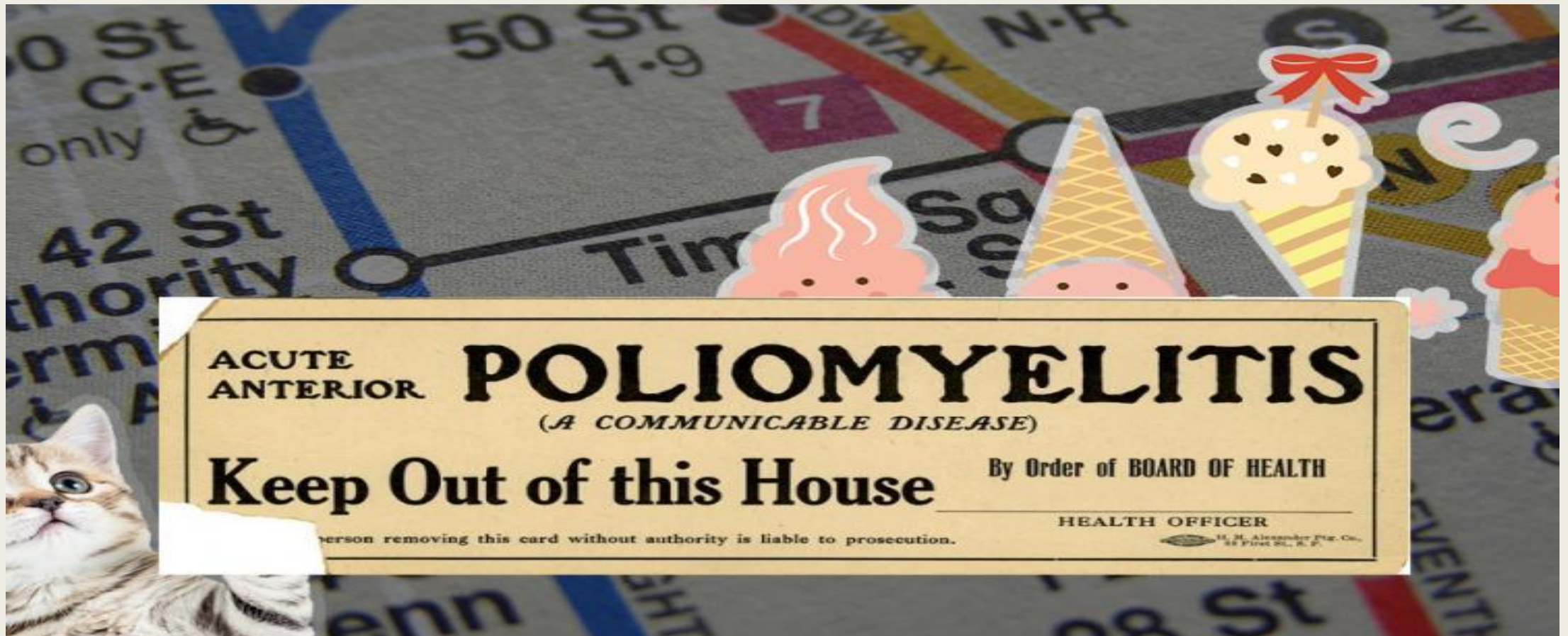


Michael A DeDonno Ph.D.

mddedonno@health.fau.edu

www.michaeldedonno.com

Why study statistics?



Why study statistics?

- 1977 - Study published in *Science* found an association between saccharin and cancer in rats, spurring the FDA to ban saccharin.
- Results in humans showed no clear evidence of an association.
 - *Experts soon found that the mechanism that led to cancers in rats was irrelevant in humans. FDA changed their position on saccharin.*



Why study statistics?

JAMA Neurology

Association of Proton Pump Inhibitors With Risk of Dementia A Pharmacoepidemiological Claims Data Analysis

Willy Gomm, PhD; Klaus von Holt, MD, PhD; Friederike Thomé, MSc; Karl Broich, MD; Wolfgang Maier, MD;
Anne Fink, MSc; Gabriele Doblhammer, PhD; Britta Haenisch, PhD

Why study statistics?

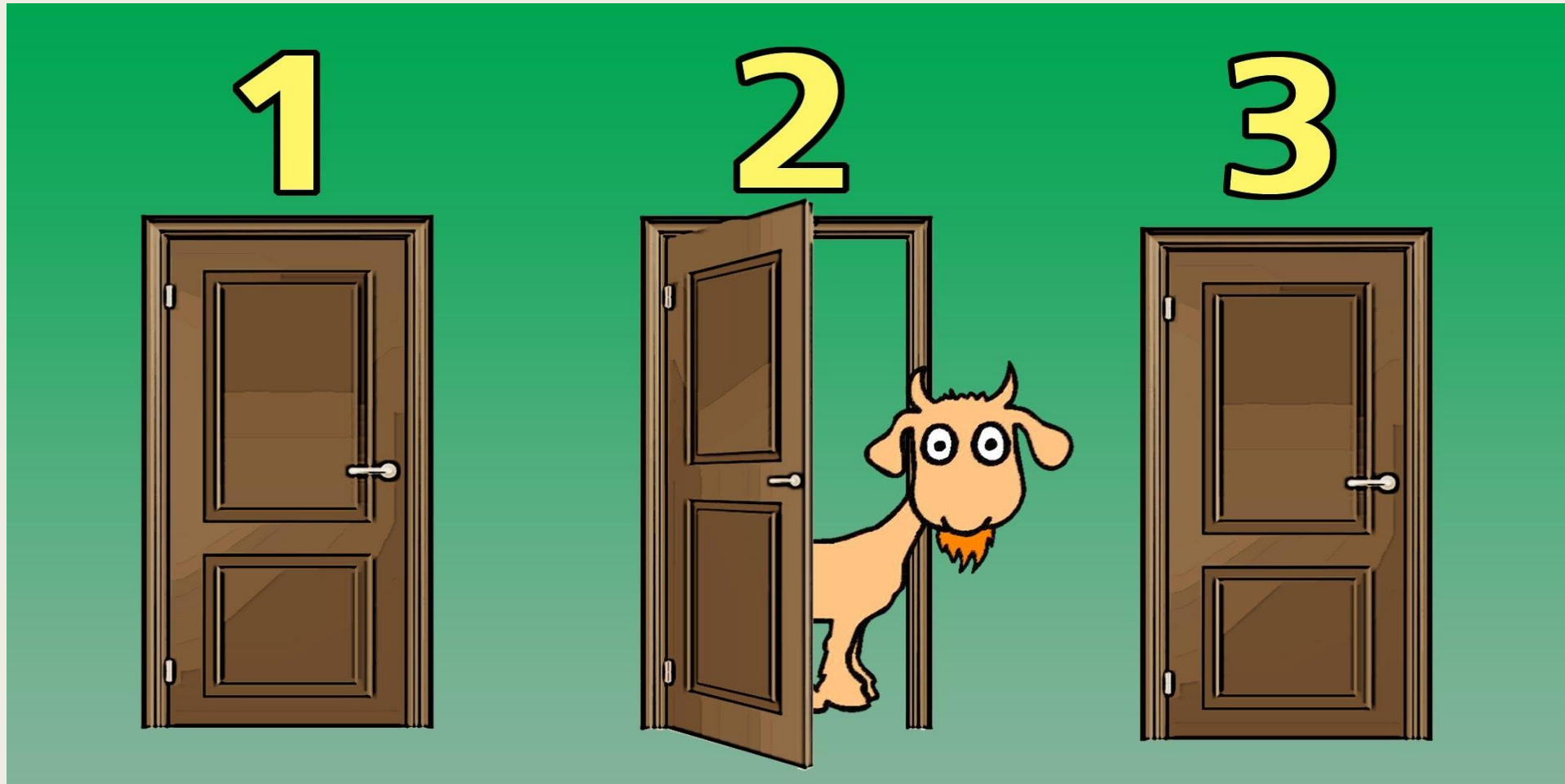


"Excellent health statistics - smokers are less likely to die of age related illness"

Why study statistics?



Why study statistics?





A white rectangular title card with four silver corner fasteners is mounted on a textured, light-brown fabric background. The card features the title 'The Art of Learning' in a black serif font, a thin horizontal line, and the subtitle 'via Multimedia' in a smaller black serif font.

The Art of Learning

via Multimedia

Why study statistics?



Why study statistics?

ON TEENAGERS, ADULT:

Statistics show that
teen pregnancy
drops off significantly
after age 25.

*Mary Anne Tebedo, Republican state senator from Colorado Springs
(contributed by Harry F. Ponce)*

MONDAY DECEMBER 1999

Why study statistics?

The Winchester Star

BRIEFS

Study Shows Frequent Sex Enhances Pregnancy Chances

By The Associated Press

BOSTON — A study that researchers say gives the best estimate ever of nature's window of female fertility comes to a not-so-startling conclusion: The best way to make babies is to have lots of sex.

Every other day is better than once a week. Daily is best.

The discovery, however, is not

ples don't want to use other forms of birth control.

Researchers say there are six days in every menstrual month when a woman can get pregnant.

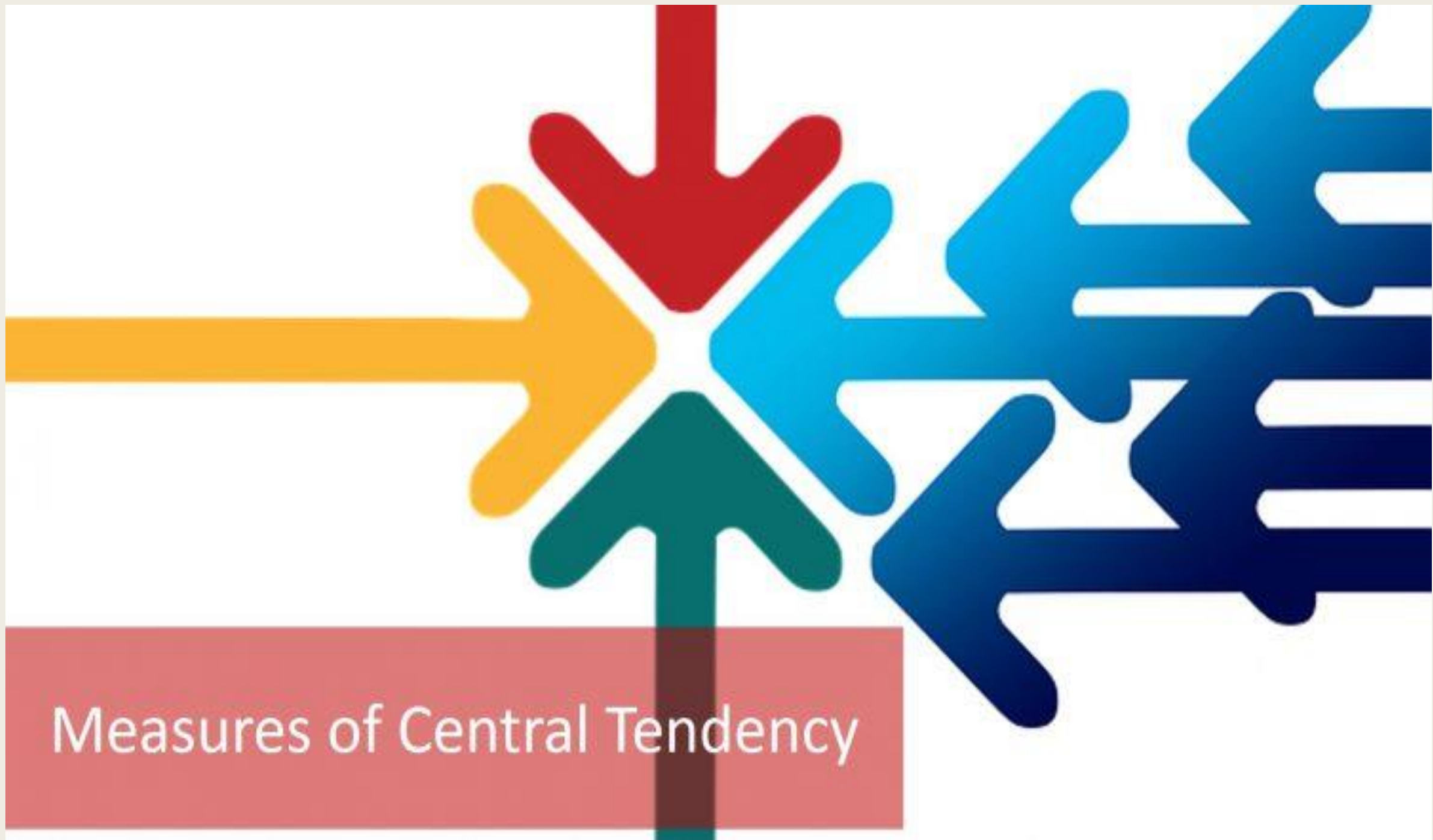
Until now, estimates of women's fertility ranged from two days in a menstrual cycle to 10 or more. The new study found that conception is possible if a woman has intercourse on the five days before ovulation as well as on the day

Agenda

- Central Tendency
- Variability
- Probability
- Percentile
- Correlation
- Regression

FAU
CHARLES E. SCHMIDT
COLLEGE OF MEDICINE
Florida Atlantic University





Measures of Central Tendency

Central Tendency

- Men
 - *Average height = 69.2 inches*
 - *Average weight = 195.7 lbs.*
 - *Waist circumference = 40 inches*
- Women
 - *Average height = 63.7 inches*
 - *Average weight = 168.5 lbs.*
 - *Waist circumference = 38 inches*
- *One number used to describe the entire sample or population - an average.*



“What fits your busy schedule better, exercising one hour a day or being dead 24 hours a day?”

Source: Anthropometric Reference Data for Children and Adults: United States, 2011-2014, tables 4, 6, 10, 12, 19, 20

Central Tendency

Mode

- The value that occurs most often in a dataset

Median

- The middle value of a dataset

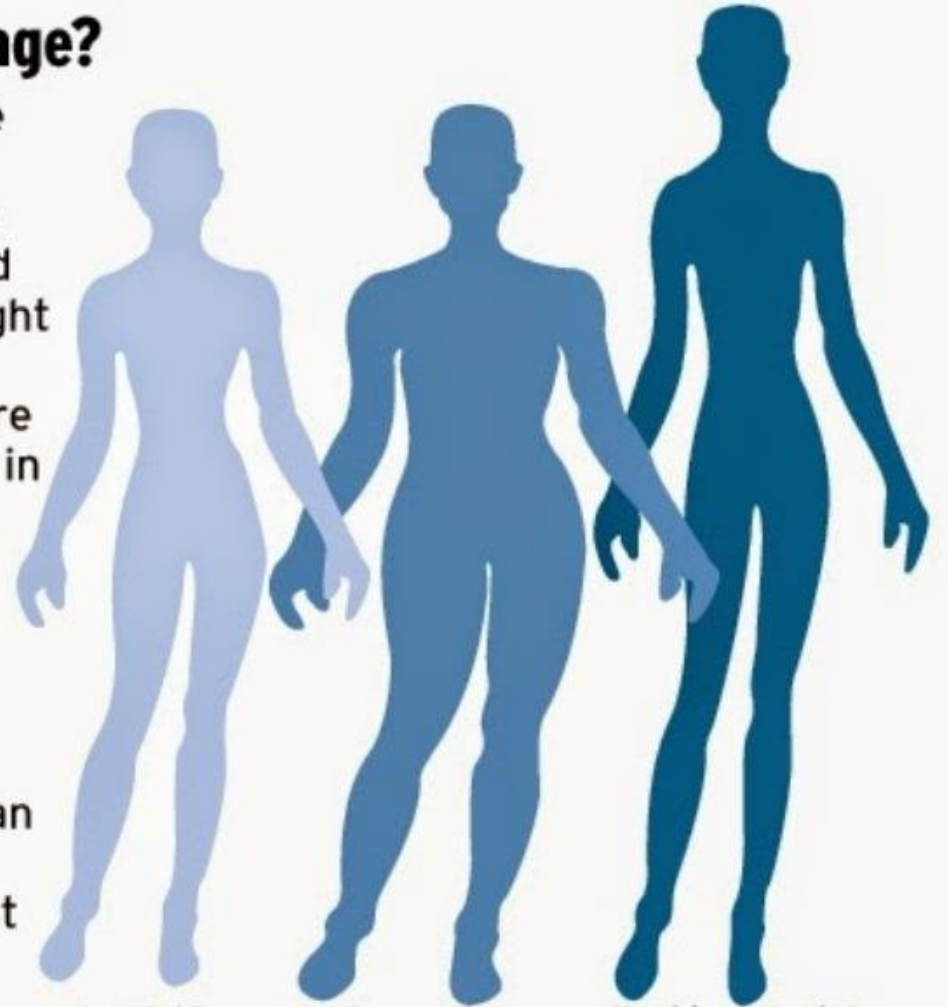
2, 3, 3, 3, 3, 4, 5, 5, 6, 6, 7, 7, 8, 9, 9, 10

Central Tendency

- Mean – The average of the numbers.
- Add all the numbers, divide by how many numbers in the dataset.
- $\text{Sum} \div \text{Count} = \text{Mean}$
- Sensitive to outliers

What's average?

The average size of American women over age 20 has increased very little in height in the past 50 years but by more than 30 percent in weight. Today's average fashion model is about the same weight as the average American woman was in 1962, but at least 8 inches taller.

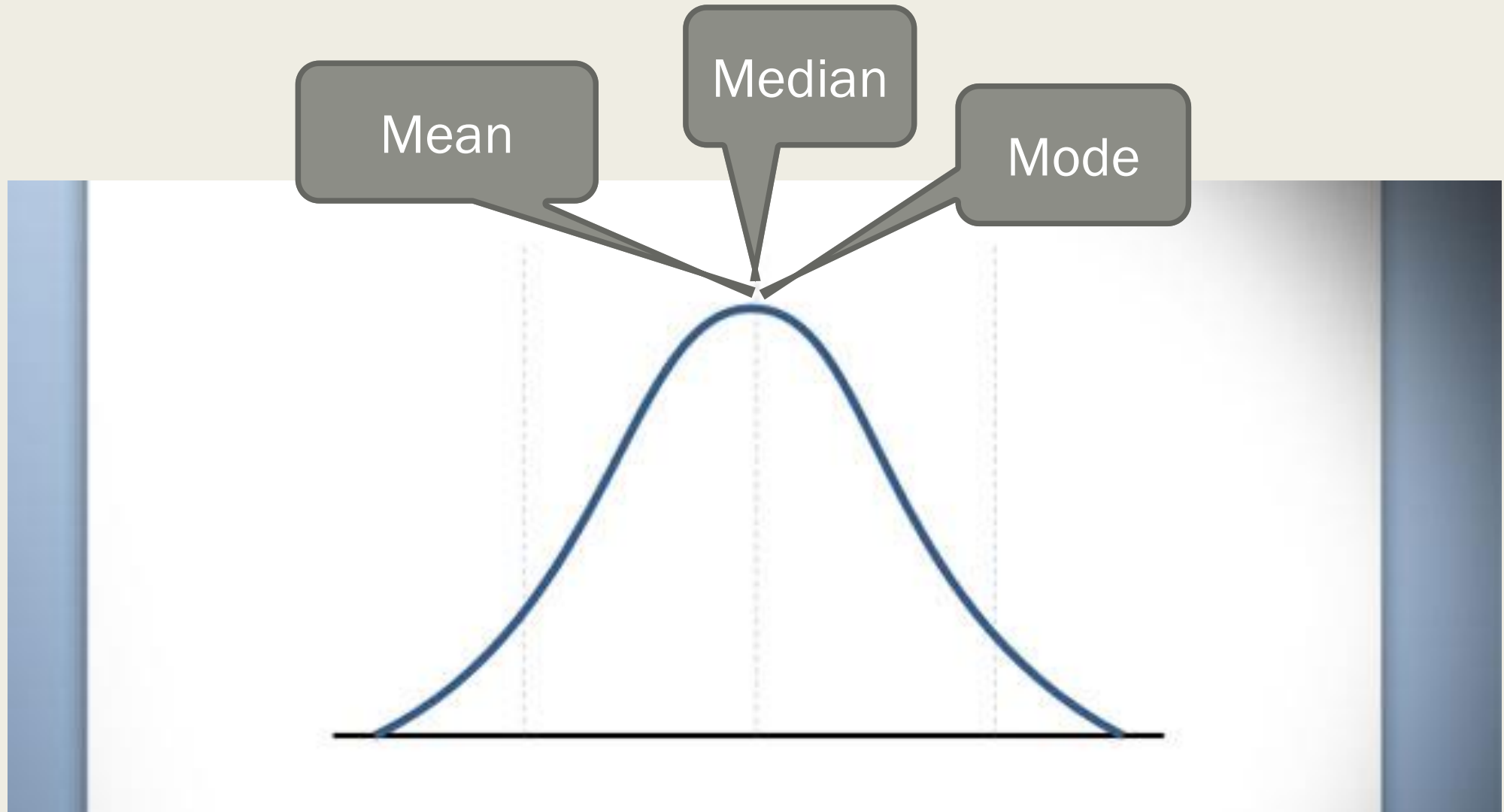


	In 1962	Current	Fashion model
Height	5'3"	5'3.8"	5'11"-6'0"
Waist	24-25	37	22-23
Dress size	8	14	0-1
Weight	120-125 lbs.	164.7 lbs.	90-120 lbs.

Sources:
WebMD,
Centers for
Disease Control and Prevention

Cindy O'Dell and Molly Zisk / The Register

Normal “Gaussian” distribution

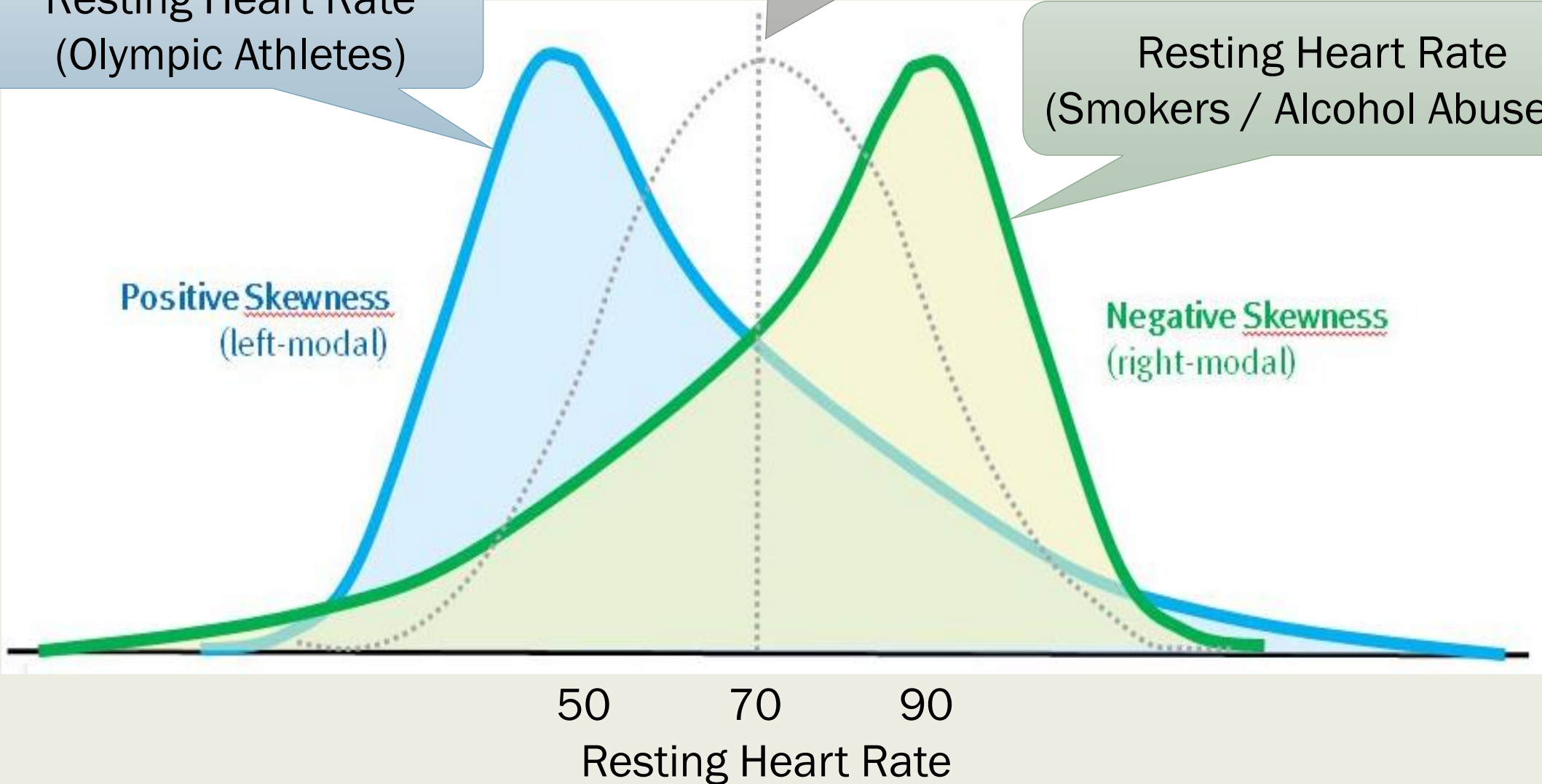


Skew of Data

Resting Heart Rate
(Olympic Athletes)

Average Resting Heart Rate
(Men 26-35)

Resting Heart Rate
(Smokers / Alcohol Abusers)

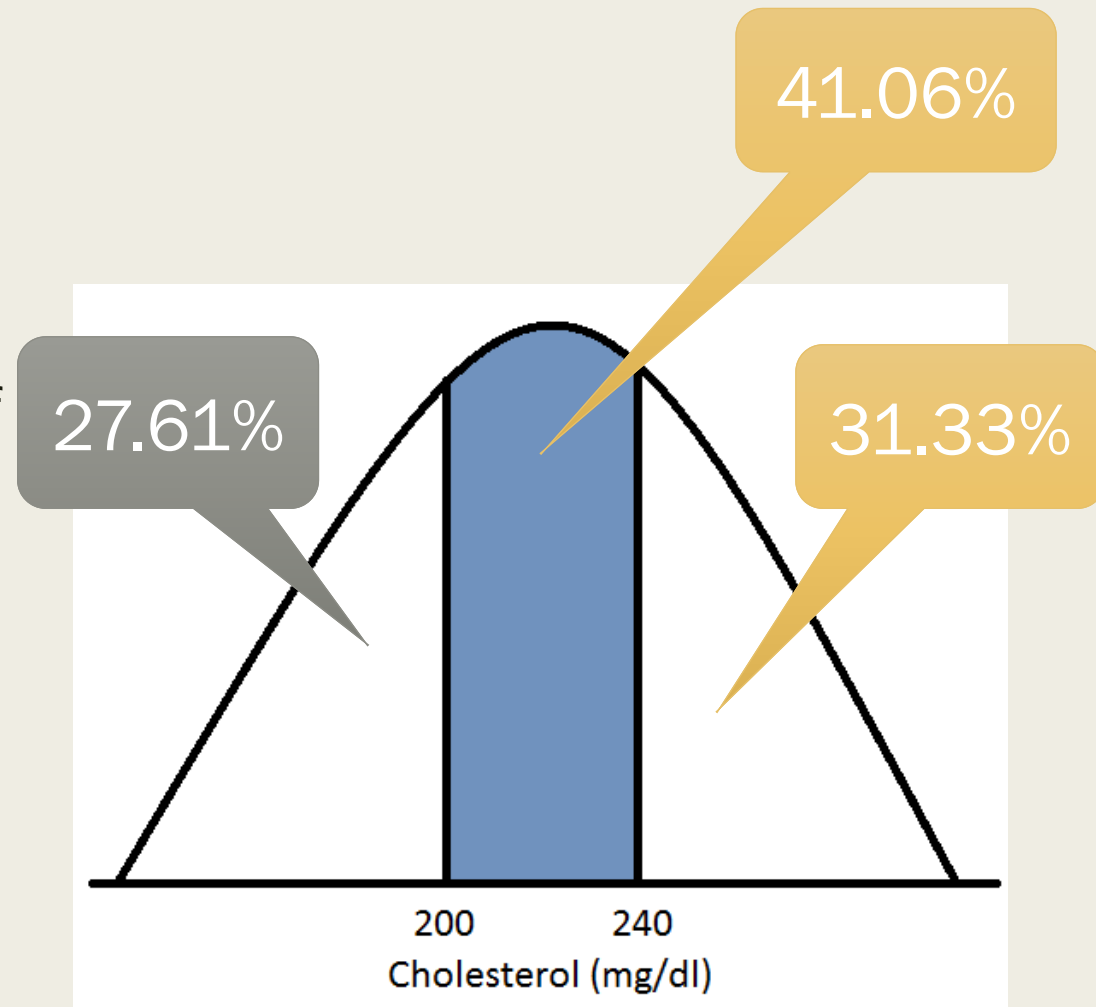


Probability & Continuous Variables

- Natural variability exists in almost all health related measures
 - *Plays a role in the definition of disease and pathology*
- How we define a particular condition? Depends on the variable of measure
 - *Discrete variable – Fixed number of values – easier to define*
 - Presence or Absence of a medical condition: e.g., bone fracture – yes/no
 - *Continuous variable – Infinite number of possible values – more difficult to define & have changed over time*
 - Blood pressure: How high is too high?
 - Testosterone: How low is too low?

Implications of a chosen threshold

- Cholesterol definitions
 - < 200 = *Acceptable*
 - $200-240$ = *Elevated or Borderline high*
 - > 240 = *High*
- Modeling total cholesterol in a population of middle-aged men using a normal distribution with a mean of 222mg ($\sigma = 37$).
- 72% of this population has Elevated or High cholesterol
- Some have advocated giving statins to virtually everybody over the age of 50.
- This of course redefines age as a pathology in-and-of-itself



Variability



Dispersion

- While indicators of central tendency provide a central value, does little to show the dispersion of the sample.
 - *The AvgHR of an Olympic athlete is 50. Do all Olympic athletes have a resting HR of 50?*
 - *There is of course a dispersion around the AvgHR.*



Dispersion Indicators popular in medicine

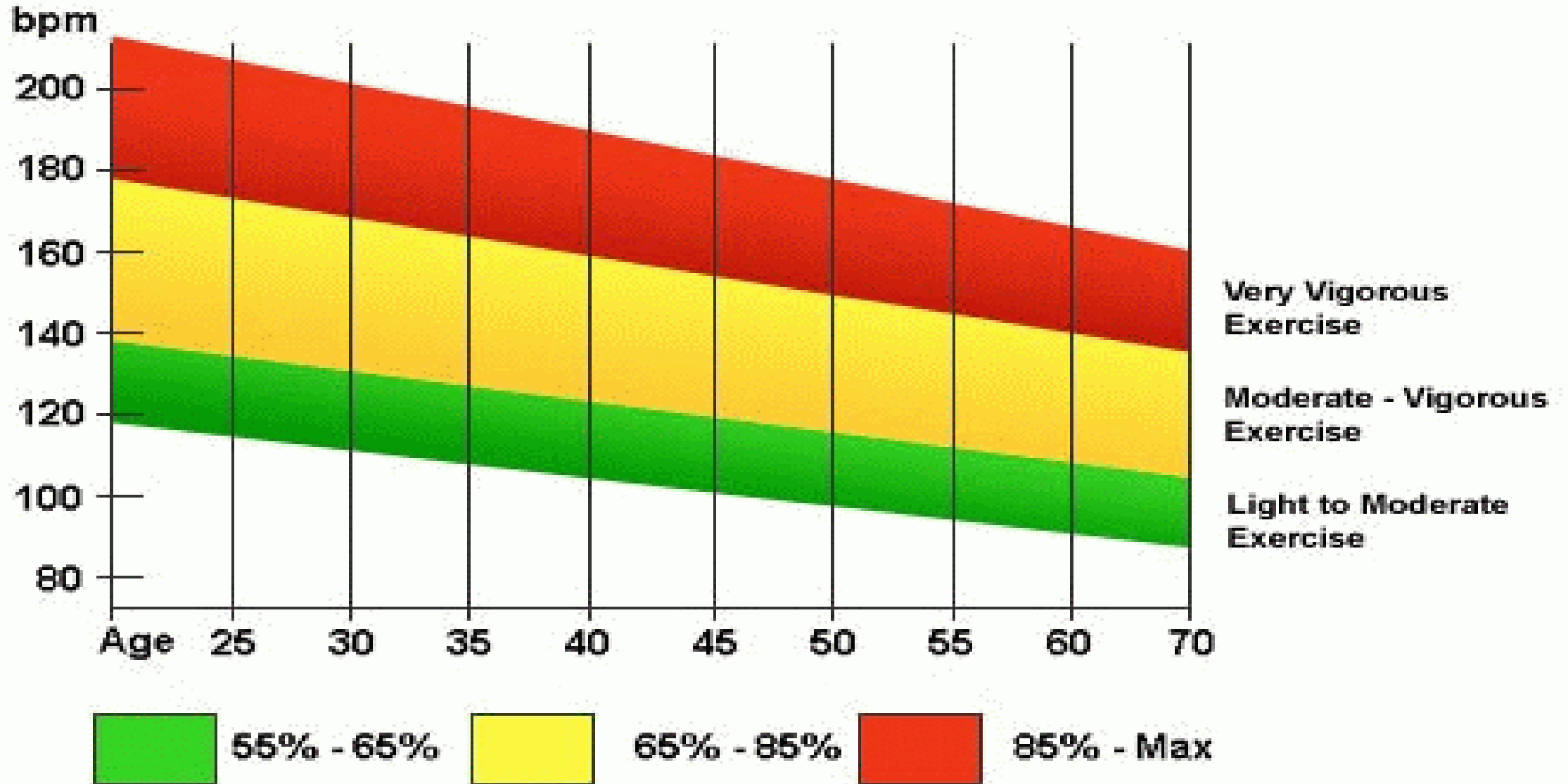
- Range
- Deviation from the mean
- Variance
- Standard Deviation (SD)
- Standard Error (SE)



Dispersion Indicator - Range

- Simple, direct way to indicate the spread of a collection of values
 - *The range in BMI (kg/m²) of the sample was 17.00 – 35.00*
 - *The range in avg HR of a sample of Olympic athletes was 41 – 72 bpm*
- Does not tell how the values are distributed
 - *E.g., Range of weight in US adults ~ 85 – 600lbs*
 - *Provides little value of average*

Dispersion Indicator - Range



Dispersion Indicator – Deviation from the mean

- Subtracting the mean from each value to obtain a deviation from the mean

Problematic. Some deviations will be positive, some negative, and the sum will **always** be zero

Heart Rate	Average HR	Deviation
74	70	4
84	70	14
58	70	-12
70	70	0
64	70	-6
$350 \div 5 = 70$		

Dispersion Indicator – Variance

- An indicator that corrects the problems we have with a deviation from the mean

Heart Rate	Average HR	Deviation	Sqr Deviation	Sqrd Dev
74	70	4	4 ²	16
84	70	14	14 ²	196
58	70	-12	12 ²	144
70	70	0	0 ²	0
64	70	-6	6 ²	36
				392

However, the variance value of 78.4 does little to aid in the understanding of dispersion in a real life

- Variance = Avg. of Sqrd. Deviations = $392/5 = 78.4$

Dispersion Indicator – Standard Deviation

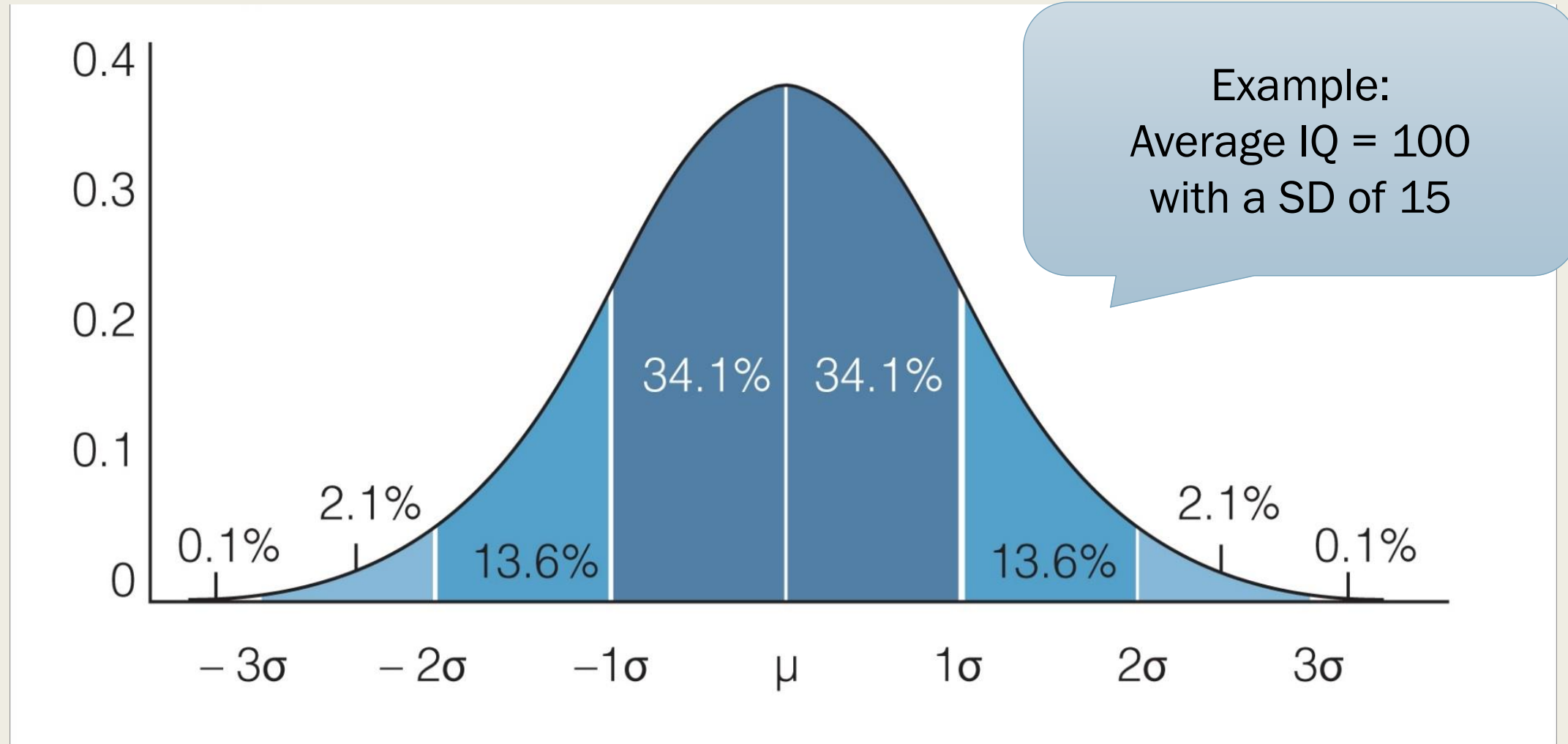
- An indicator (in original units) of the spread within a set of values

Heart Rate	Average HR	Deviation	Sqr Deviation	Sqrd Dev
74	70	4	4^2	16
84	70	14	14^2	196
58	70	-12	-12^2	144
70	70	0	0^2	0
64	70	-6	-6^2	36
				392

- Remember Variance = 78.4 (Sqrd. Dev.)
- We need to return the variance back into original units
- **Stand Deviation = Square root of 78.4 = 8.85**

*Something of
“real” value*

Dispersion Indicator – Standard Deviation



■ σ = Standard Deviation (SD)

Dispersion Indicator – Standard Deviation

Heart Rate	Average HR
74	70
84	70
58	70
70	70
64	70

- **Stand Deviation (SD) = 8.85**
- **1 SD below = 61.15 (70 - 8.85)**
- **1 SD above = 78.85 (70 + 8.85)**
- **3 of the 5 (60%) values are within 1 SD of the Mean**

Dispersion Indicator – Standard Error

- An Estimate at a specified confidence level (probability) the interval within which the population mean will fall.
- Example:
 - *In a study, six of the ten (60%) patients with severe mental illness were drug abusers*
 - *Physicians who encounter individuals with severe mental illness may see drug abuse in 60% of them.*
 - *Is 60% the real proportion of ALL patients with severe mental illness who abuse drugs?*

Dispersion Indicator – Standard Error

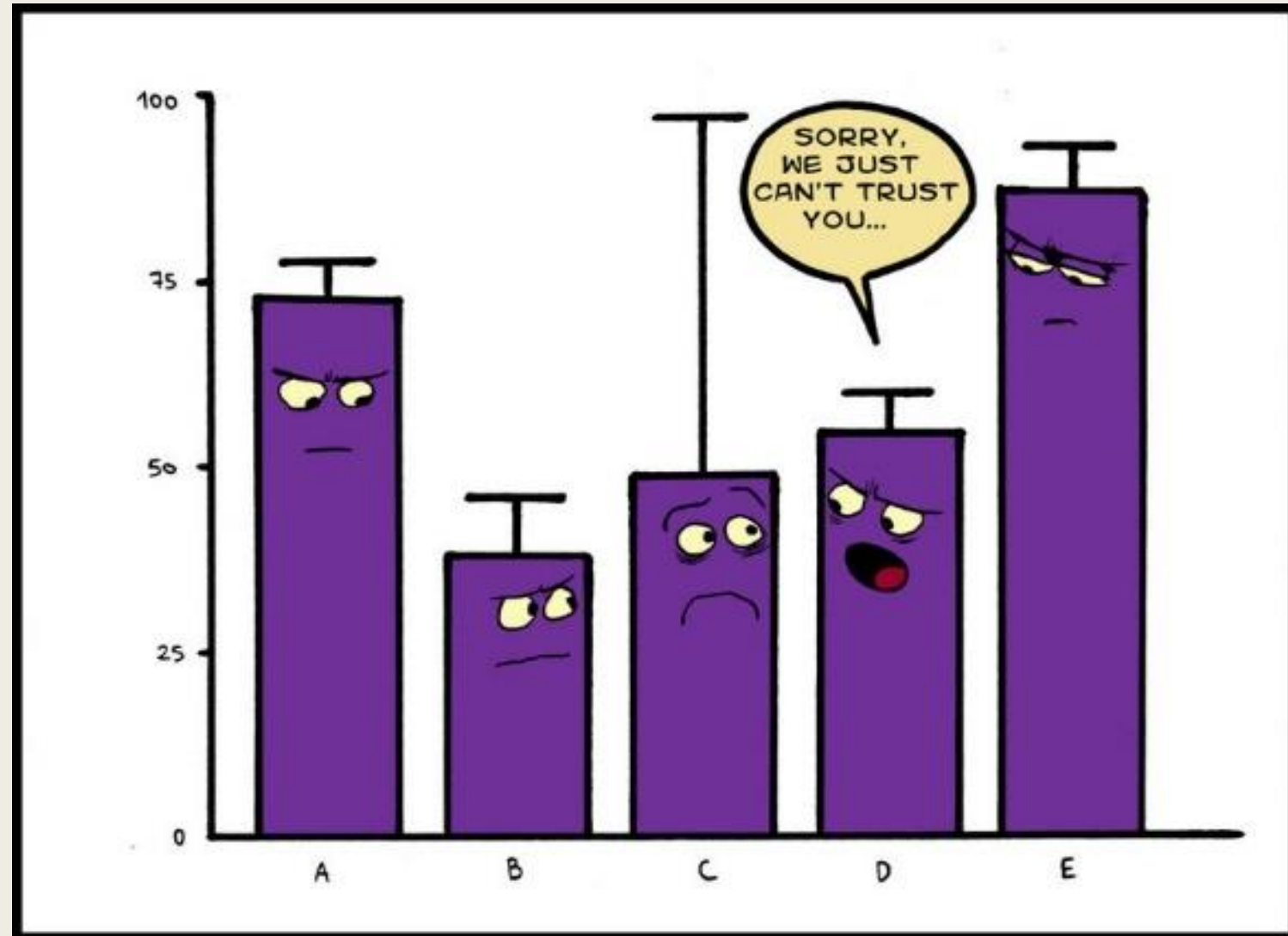
- Example (cont.):

- *Six of the ten (60%) patients with severe mental illness were drug abusers*
- *We can calculate an interval (SE_p: Standard Error of the proportion) at a specified confidence level (e.g., 95%) that would include the real proportion of individuals with a severe mental illness who are drug abusers*
- *Based on the example above, SE_p = 0.3 to 0.9*
 - So the real proportion at a 95% confidence is somewhere between 30% and 90%

The expanse of the interval may explain the absence of SE_p in some medical reports.

Dispersion Indicator – Standard Error

- What impacts Standard Error calculations?
 - *Desired degree of confidence (68%, 95%, 99%)*
 - **Sample size** – *As sample size increases, SE interval tightens (more precise)*

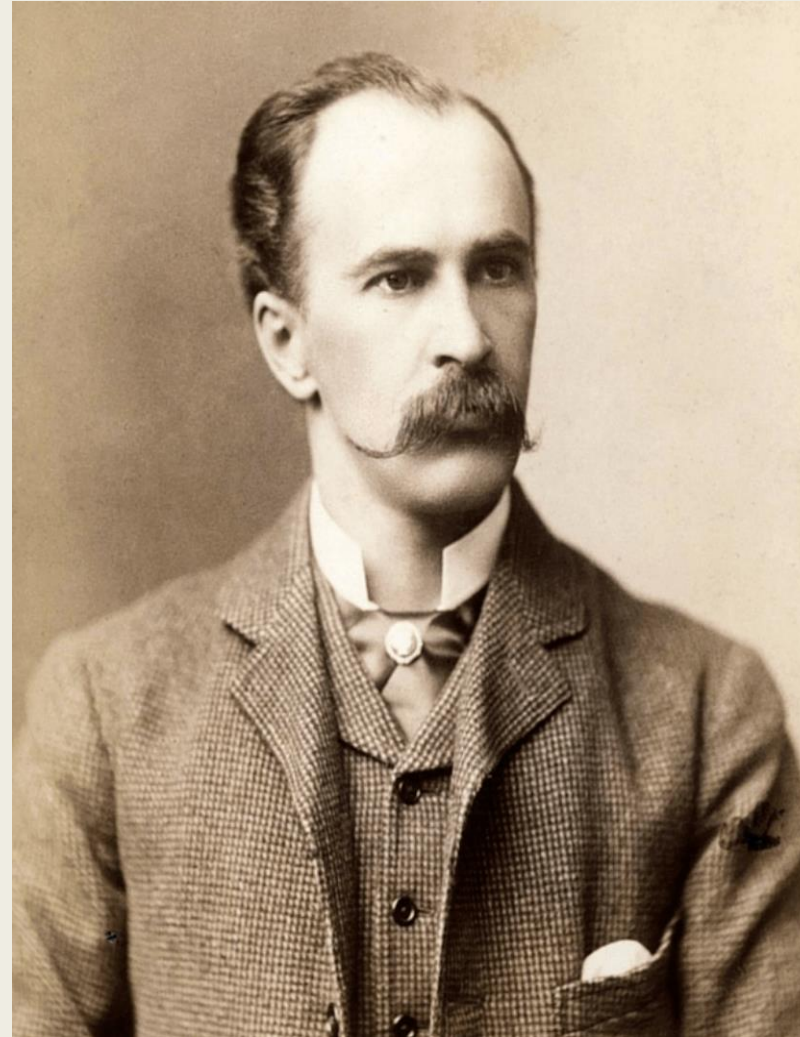


Probability



Probability

- General notation for probability
 - P denotes probability
 - A , B , and C denote specific events
 - $P(A)$ denotes the “probability of event A occurring”



Medicine is a
science of
uncertainty and
an art of
probability.

William Osler

Probability

- Two general approaches to finding probabilities
 - *Relative Frequency Approximation of Probability*
 - *Classical Approach to Probability*
- Results in values between 0 and 1
 - $(0 \leq P(A) \leq 1)$
- Probabilities as Percentages
 - *Mathematically, probability of 0.25 is equivalent to 25%*
 - *Medical journals almost universally express probabilities as decimals*

Relative Frequency Approximation of Probability

- Find the probability of dying when making a skydiving jump
- In 2015, there were about 3,000,000 skydiving jumps, 21 deaths
- $P(\text{Skydiving death}) = \frac{\text{Number of skydiving deaths}}{\text{Total number of jumps}} = \frac{21}{3,000,000}$
 $= 0.000007$

Classical Approach to Probability

- **Note:** *Requires equally likely outcomes*
- Find the probability of getting three children all of the same gender when three children are born
- The sample space {bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg}
- Assume all events are equally likely
- Eight equally likely outcomes, 2 outcomes with children of same gender
- $P(\text{three children of the same gender}) = 2/8 = 1/4$ or 0.25

Common mistake # 1 in finding a probability value

- Mistaken belief that if something happens more frequently than normal, then it will happen less frequently in the future. *Gamblers fallacy*
 - *Example: Theory of Progressive betting – If playing blackjack (21), and you lose a hand, you should double your bet to recoup your lost money due to the probability of winning increases, after a loss.*



Common mistake #2 in finding a probability value

- Find the probability that a high school driver, texted while driving during the previous 30 days
- In a study, it was found that 3,785 texted while driving in the previous 30 days, while 4,720 did not.
- Solution: $3785 / 4720 = 0.802$
- Correct: sum the total number of drivers in the sample: $3785 + 4720 = 8505$
- $P(\text{texting while driving}) = \frac{\text{Number of drivers who texted while driving}}{\text{Total number of drivers in the sample}}$
 $= 3785 / 8505 = 0.445$

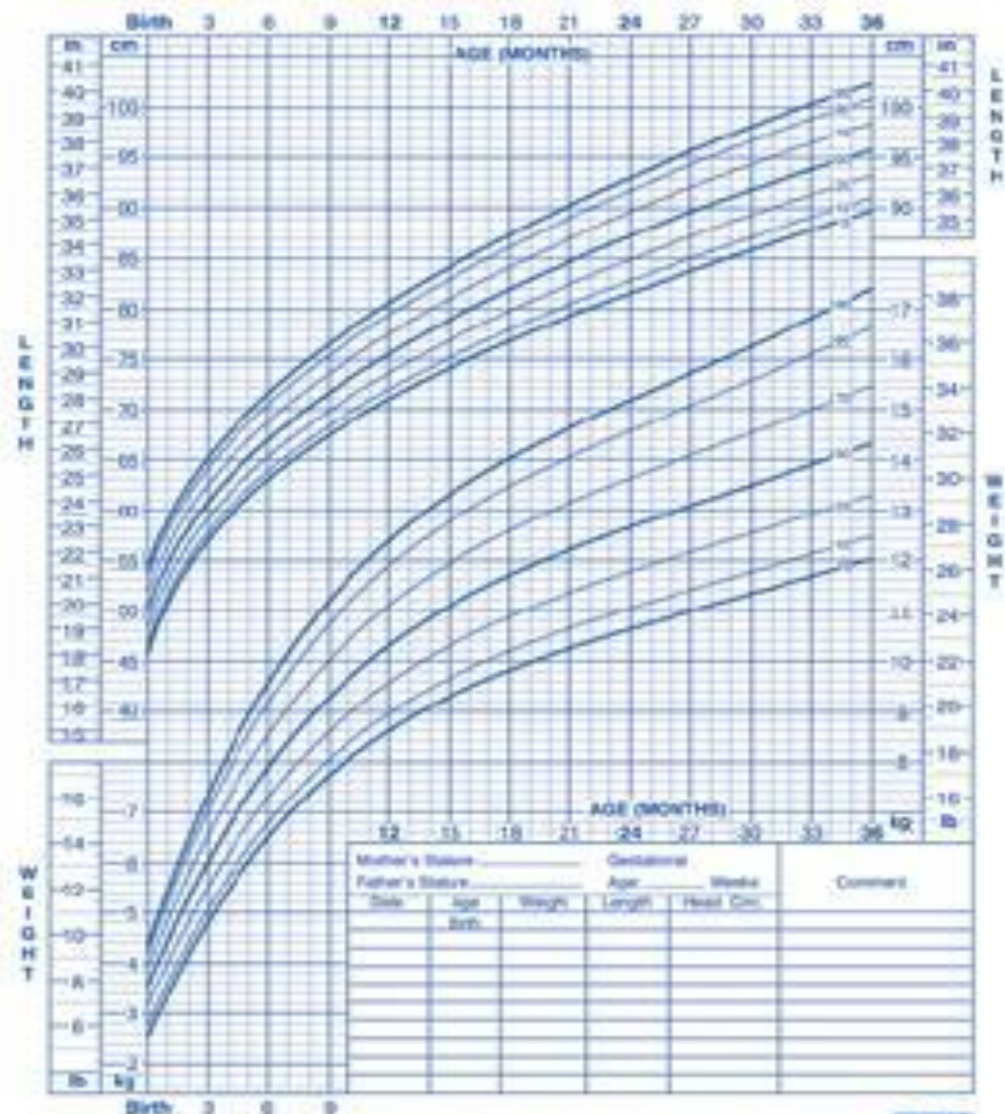
Percentile



Birth to 36 months: Boys
Length-for-age and Weight-for-age percentiles

NAME _____

RECORD # _____



Published May 20, 2002 (revised 4/2014)

REVISION: Developed by the National Center for Health Statistics in collaboration with
the National Center for Chronic Disease Prevention and Health Promotion (2002)

Web: www.cdc.gov/nchs/nhanes



www.healthypeople.gov

Percentile

Males, Birth – Weight (in kilograms) Age (in months)

Age	3rd Percentile	5th Percentile	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile	95th Percentile	97th Percentile
0	2.35	2.52	2.77	3.15	3.53	3.87	4.17	4.34	4.44
0.5	2.79	2.96	3.20	3.59	4.00	4.38	4.71	4.91	5.03
1.5	3.61	3.77	4.02	4.42	4.87	5.32	5.72	5.96	6.12
2.5	4.34	4.50	4.75	5.18	5.67	6.17	6.63	6.92	7.10
3.5	4.99	5.15	5.41	5.86	6.39	6.94	7.46	7.78	7.99
4.5	5.57	5.74	6.013	6.48	7.04	7.63	8.20	8.55	8.79
5.5	6.09	6.27	6.55	7.04	7.63	8.26	8.87	9.25	9.51
6.5	6.56	6.74	7.03	7.54	8.16	8.82	9.47	9.88	10.16

regression
dependent

independent
coefficient

linear

association
estimate
Moment

analysis

correlation

predictors
positive
factors
explanatory

negative
response

collaboration

variables

variable

variable

sample

Product

rate

Pearson

Correlation

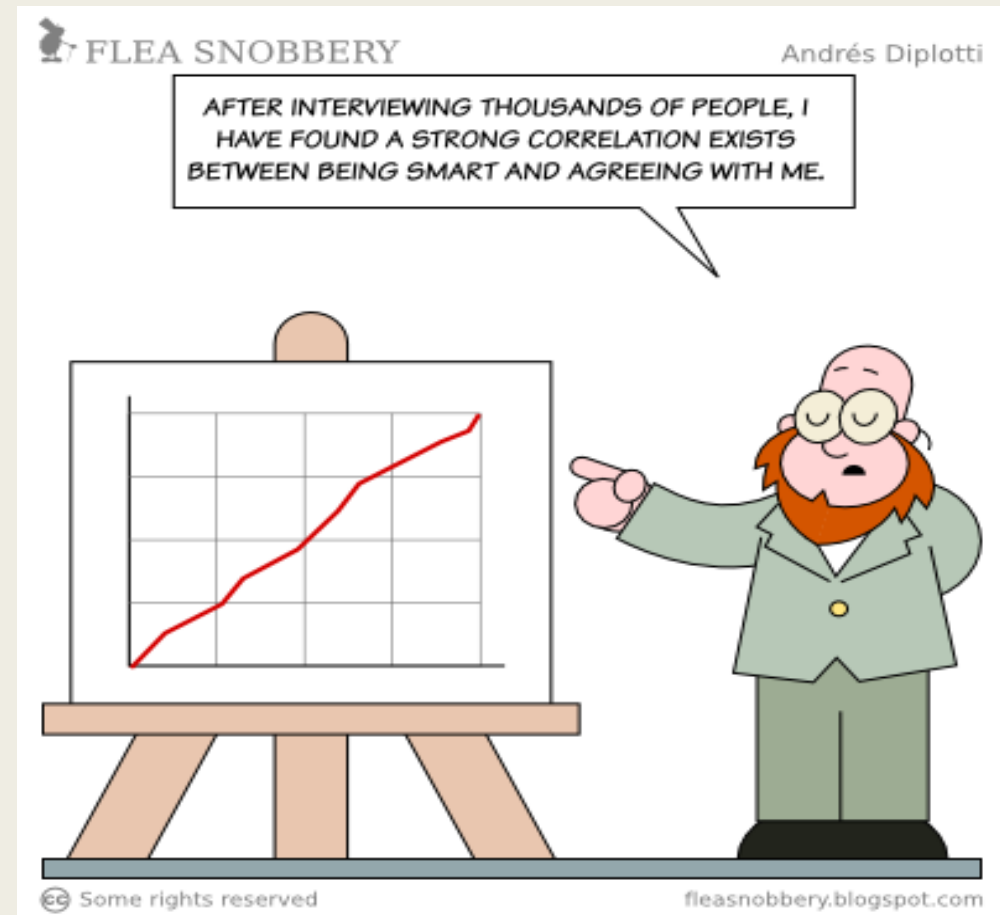
- Correlation - Measure two continuous variables and want to quantify how consistently they vary together
- The stronger the correlation, the more likely to accurately estimate the value of one variable from the other



“Doctor, there is a correlation between violence and the low-fat diet. Right now, I’d *kill* for a doughnut!”

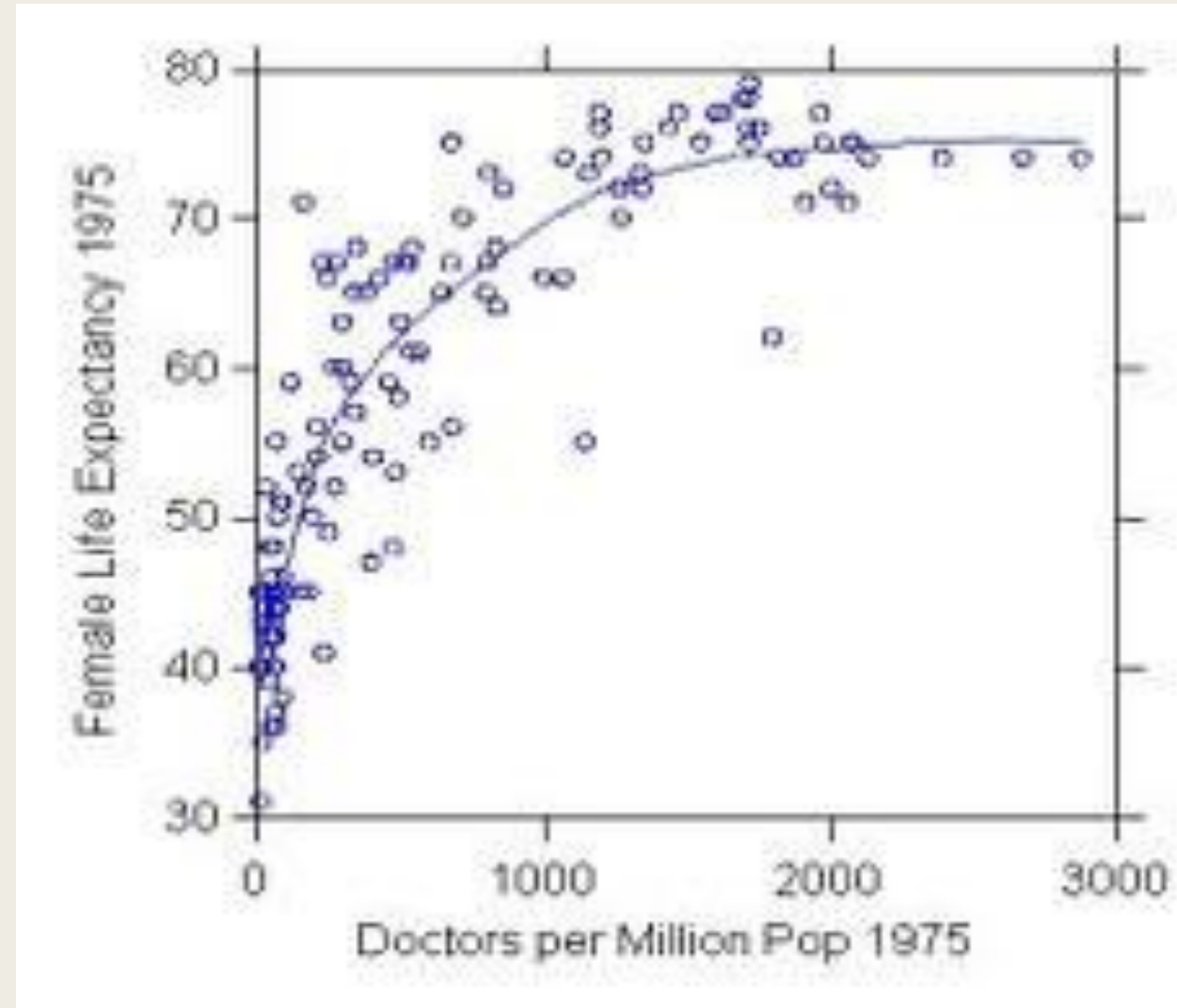
Patterns of Correlation

- Linear correlation
- Curvilinear correlation
- No correlation
- Positive correlation
- Negative correlation



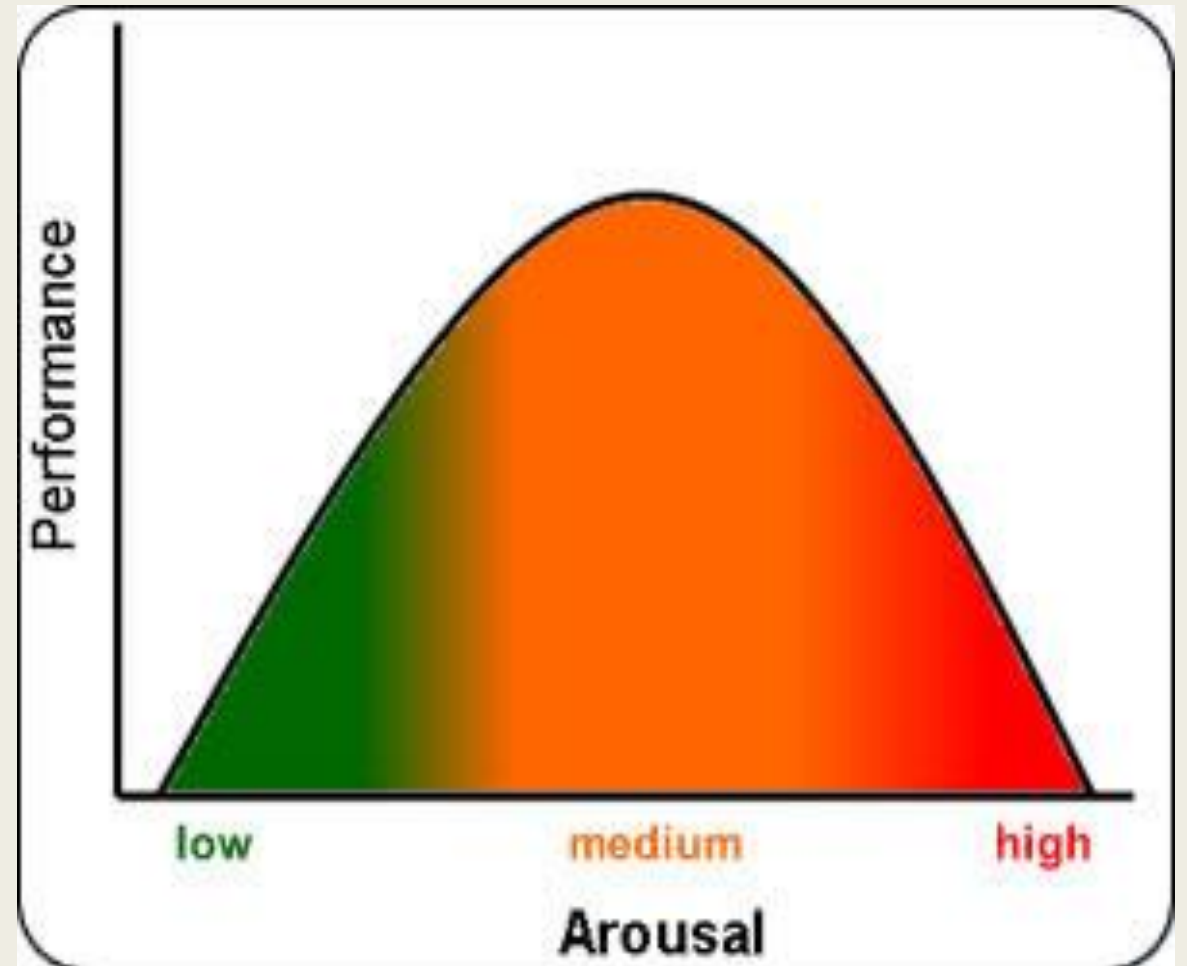
Curvilinear Correlation

- Female life expectancy and Doctors per million of population
- Test performance and average hours of sleep per night



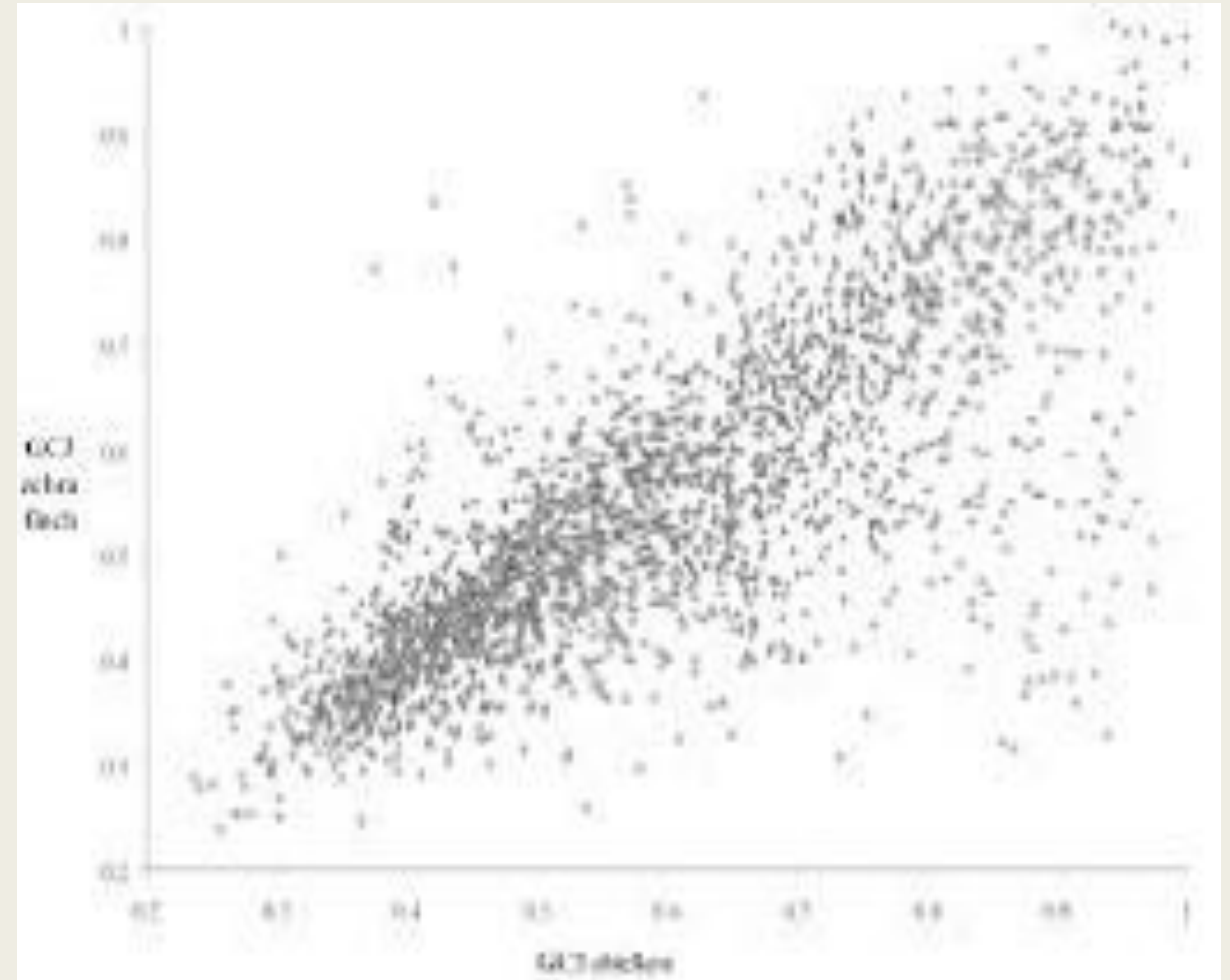
Yerkes-Dodson Law

- Psychologists Robert M. Yerkes and John Dillingham Dodson
- Performance increases with physiological and mental arousal, but only to a point, over aroused results in decreased performance



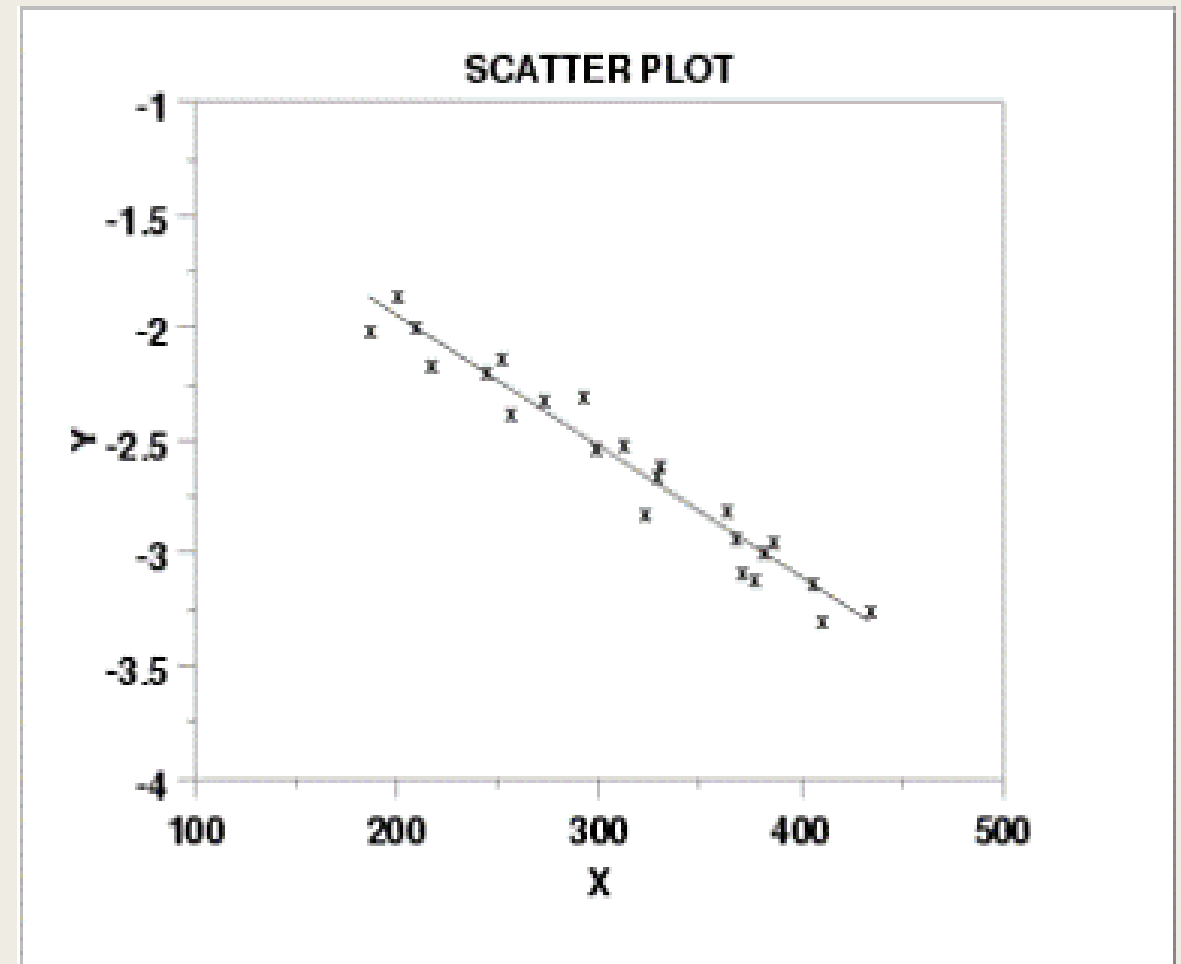
Positive Correlation (+1)

- Depression and Anxiety
- Weight and Height
- Optimism and Healthy Heart



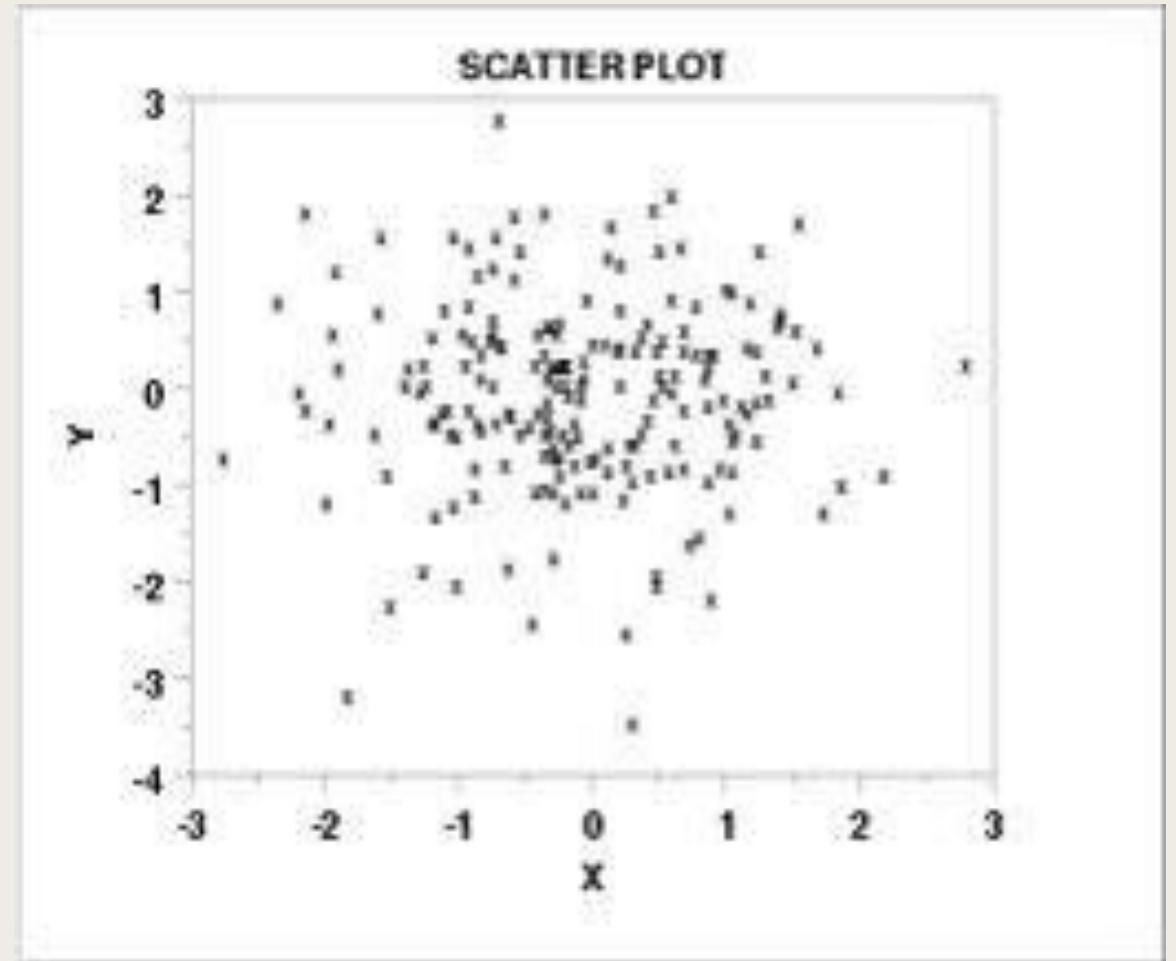
Negative Correlation (-1)

- Self-esteem and Depression
- Medical errors in a hospital and quality of the quality assurance program in place
- Hours working and cognitive ability



No Correlation

- Favorite color and IQ
- Hair color and personality
- Height and education curriculum



Pearson product-moment correlation coefficient

- PPMCC, or PCC, or Pearson's r
- Measure of linear dependence between two variables
- Developed by Karl Pearson based on related idea introduced by Francis Galton
- Value between -1 and $+1$



Pearson's r

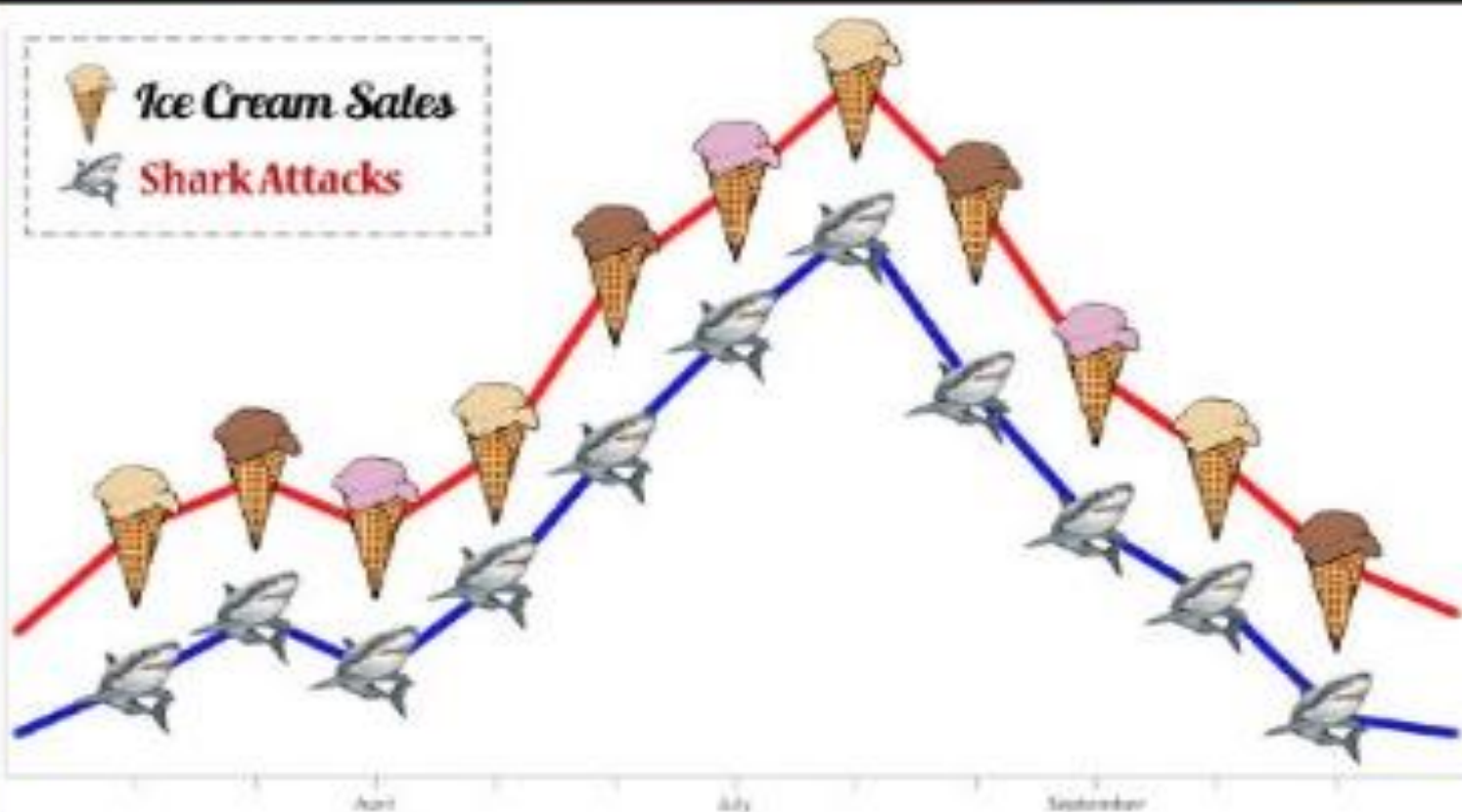
- Shows the strength of the linear relationship between two variables, symbolized by r
 - *Perfect Positive correlation: $r = +1$*
 - *No correlation: $r = 0$*
 - *Perfect Negative correlation: $r = -1$*

- What's a stronger correlation? - 0.87 or + 0.68

BACK TO THE FUTURE

- You're on a pediatrics rotation and seeing a mother and her child. The mother states that she recently read on the Internet that there's a correlation between eating ice cream and swimming pool accidents. She's worried because her child swims at the local pool on a regular basis, and that he eats ice cream at the pool concession stand once in a while.
- She asks you if she should keep her child from eating ice cream at the pool to avoid accidents.
- You respond by saying...

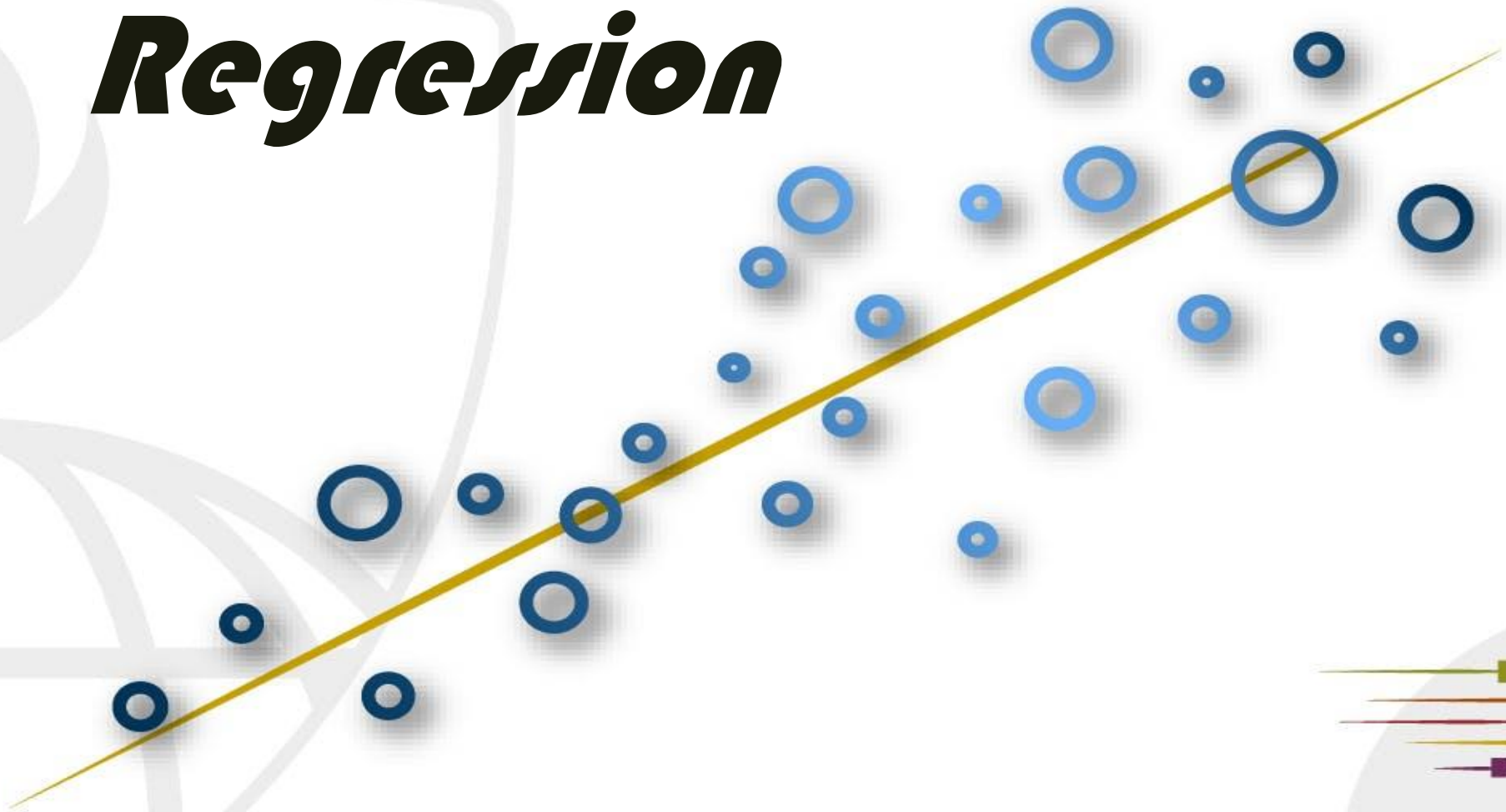
Correlation is not Causation



WARNING!

Restricting the number of ice creams sold reduces the likelihood of shark attack.

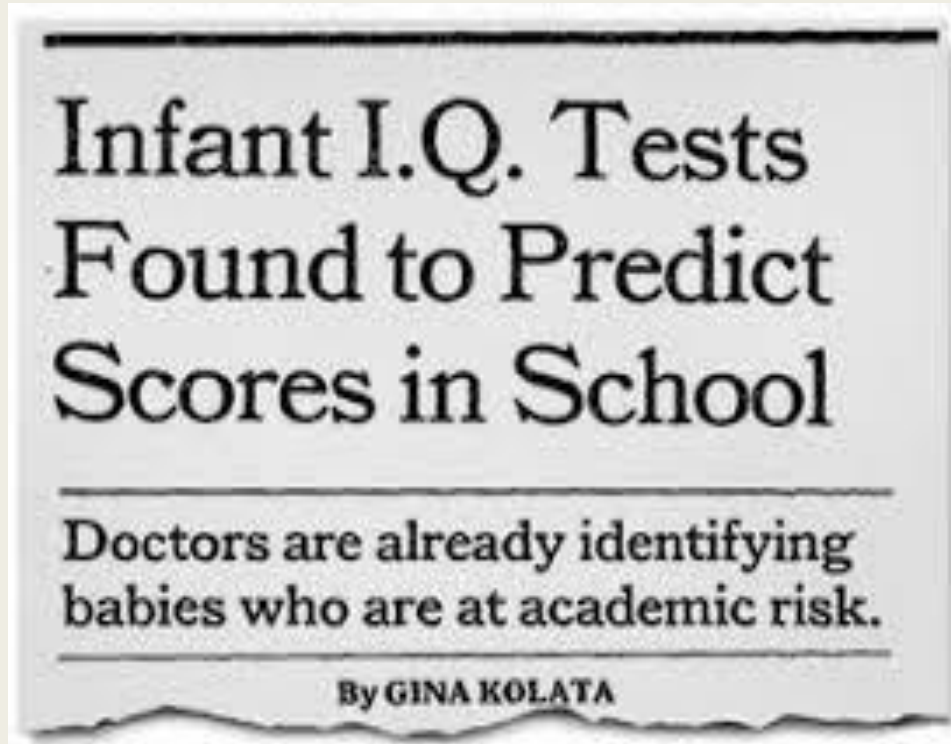
Regression



Regression

- Used to measure the relationship between two variables
 - *Prediction and a cause and effect relationship*
 - *Does one variable change in a consistent manner with another variable?*
 - *x = independent variable (cause)*
 - *y = dependent variable (effect)*

Regression ~ Prediction



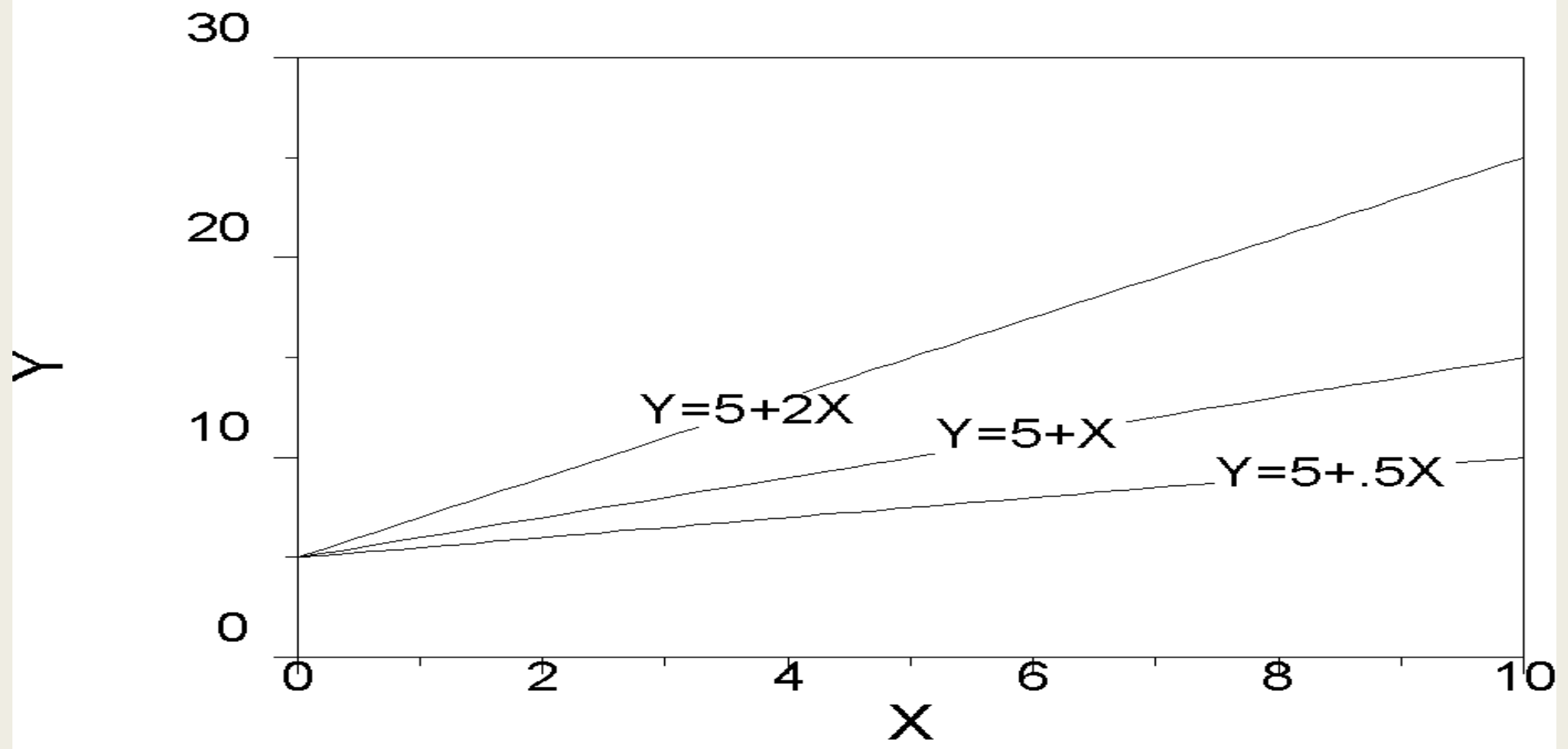
Regression

- In simple regression
 - *We predict scores on one variable from the scores on a second variable (bivariate prediction).*
 - *The variable we are predicting is called the criterion variable (DV) and is referred to as Y.*
 - *The variable we are basing our predictions on is called the predictor variable (IV) and is referred to as X.*

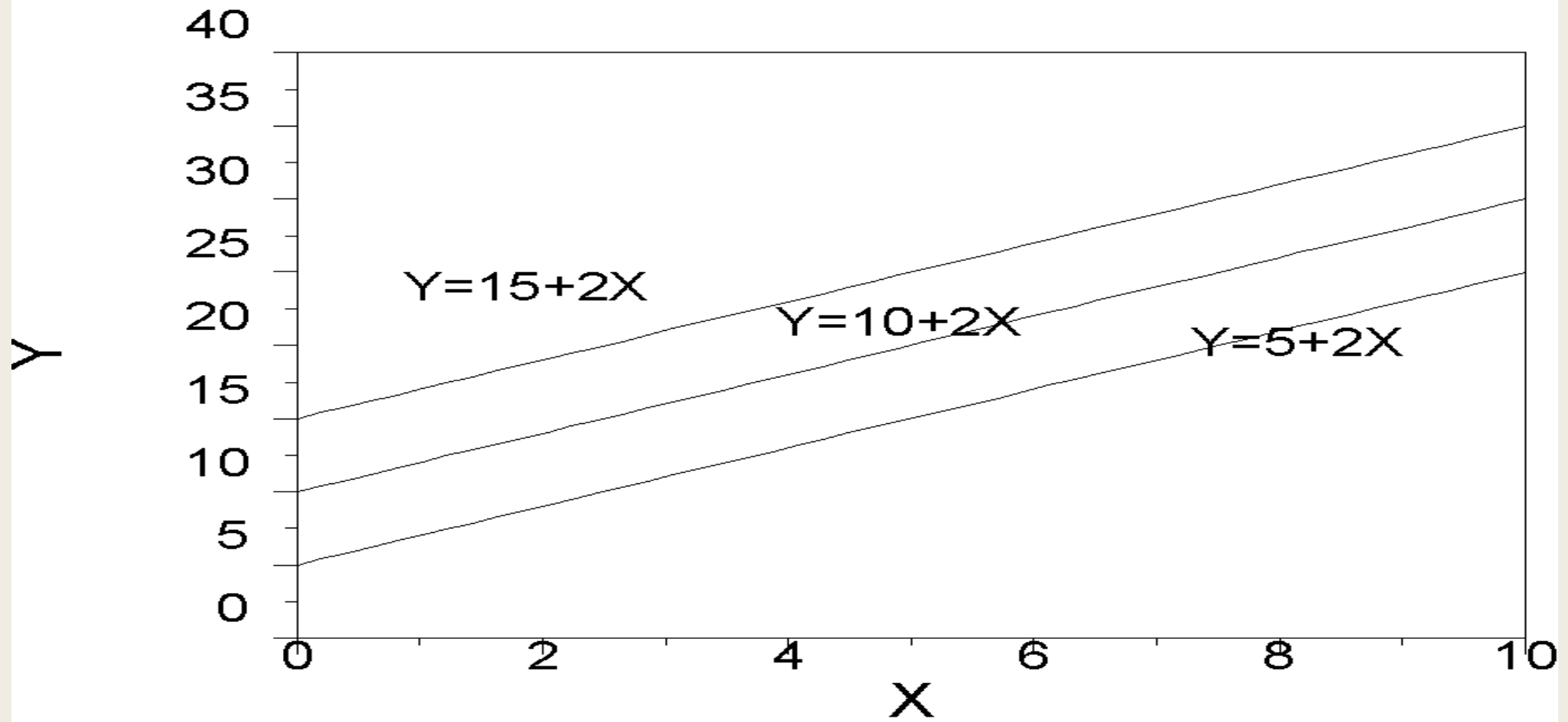
Regression

- Regression equation: $Y = a + (b)(X)$ where
 - $a = y$ intercept (point where $x = 0$ and the line passes through the y -axis)
 - $b =$ Regression coefficient, slope of the line
- The slope indicates the nature of the correlation
 - Positive = y increases as x increases
 - Negative = y decreases as x increases
 - 0 = no correlation
 - Same as Pearson's correlation
 - No relationship between the variables

Changing the Slope



Changing the Y Intercept



Regression


- *When there is only one predictor variable, the prediction method is called simple regression.*
 - *Does exercise (predictor) lead to lower resting heart rate (criterion)?*
 - *Does smoking (predictor) lead to decreased lung capacity (criterion)?*

Regression

- *When there are multiple predictor variables, the prediction method is called multiple regression.*
- How does smoking, drinking alcohol, over eating, inactivity, poor diet (predictors) impact mental health (criterion)?
- How does alcohol consumption and Thorazine (Chlorpromazine) impact lung capacity?

“Practice isn’t the thing you do once you’re good. It’s the thing you do that makes you good.” - Malcolm Gladwell

A group of investigators, describe a linear association between calcium content of the aortic valve cusps as measured *in vivo* and the diameter of the aortic opening. They report a correlation coefficient of -0.45 and a p value of 0.001. which of the following is the best interpretation of the results reported by the investigators?

- A. Alpha-error level is set too
- B. Sample size is too low for drawing definite conclusion
- C. Calcium deposition causes narrowing of the aortic valve opening
-  D. As calcium content of the cusps increases the aortic valve diameter decreases
- E. As the aortic valve diameter decreases the calcium content of the cusps decreases

Choice C is inappropriate as it suggests causation

An ICU patient has an intraarterial cannula placed after cardiac surgery to monitor systolic blood pressure (SBP). Twenty four SBP values are recorded over a period of six hours, with a maximum value of 141 mmHg, and a minimum value of 96 mmHg. If the next SBP recording is 220mmHg, which of the following is most likely to remain unchanged?

A. Mean

 Mode

C. Range


D. Variance

E. Standard deviation

A patient with severe heart failure is placed in the ICU and undergoes invasive hemodynamic monitoring. Over the next hour, the recorded values of his pulmonary artery wedge pressure are 26 mmHg, 20mmHg, 20mmHg, 27mmHg, 14mmHg, and 27mmHg. Which of the following is the median of the recorded values?

A. 20

B. 22

 23

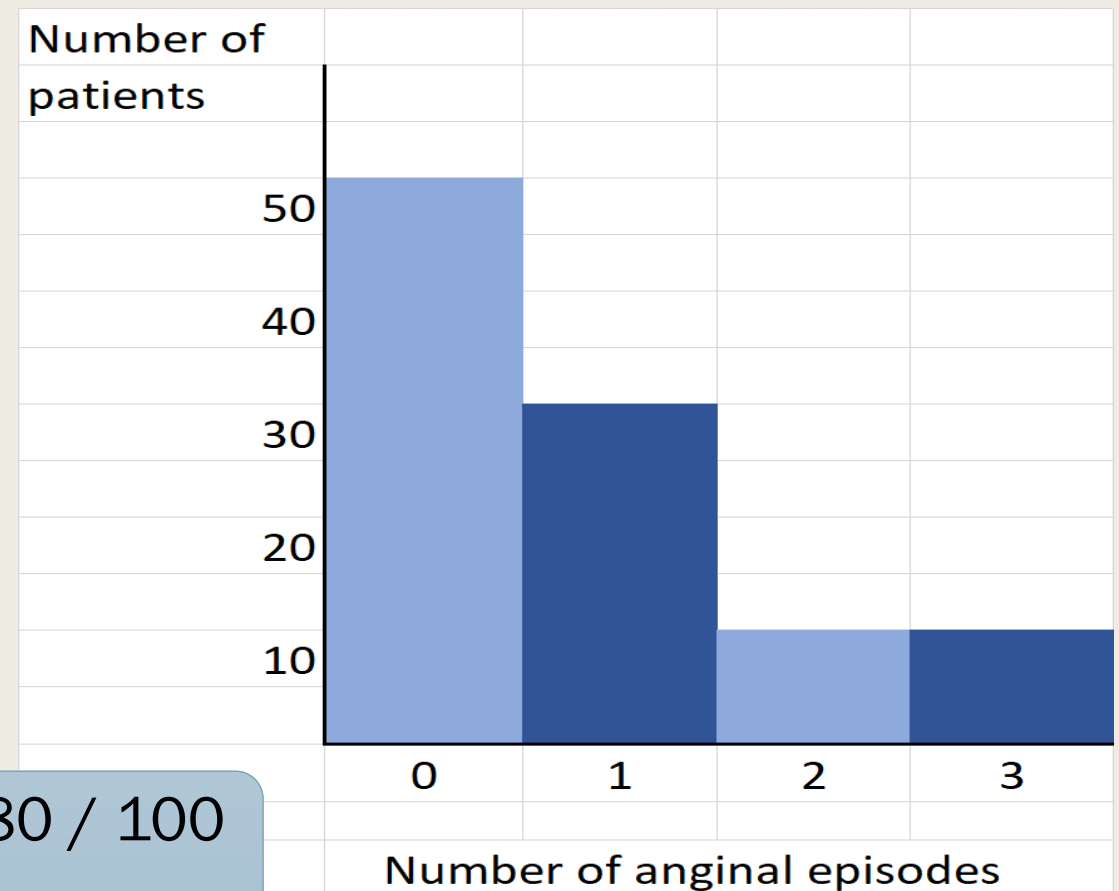
D. 24

E. 26

In an experimental study, patients suffering from stable angina are treated with a new beta-blocker. The number of angina episodes experienced by the patients on the 30th day of treatment is shown in the table below

■ What is the average number of angina episodes experienced by the patients?

- 😊 A. Between 0 and 1
- B. 1
- C. Between 1 and 2
- D. 2
- E. Between 2 and 3



$$(50 * 0) + (30 * 1) + (10 * 2) + (10 * 3) = 80 / 100 = .80$$

A fictitious disease, Head2Toes Fever, is slowly taking over the entire FAU medical student community (3,000 students). In 2015, the prevalence of this disease within this community was 10%. In 2016, 270 new cases of Head2Toes Fever were reported. What is incidence of this gripping disease in 2016?

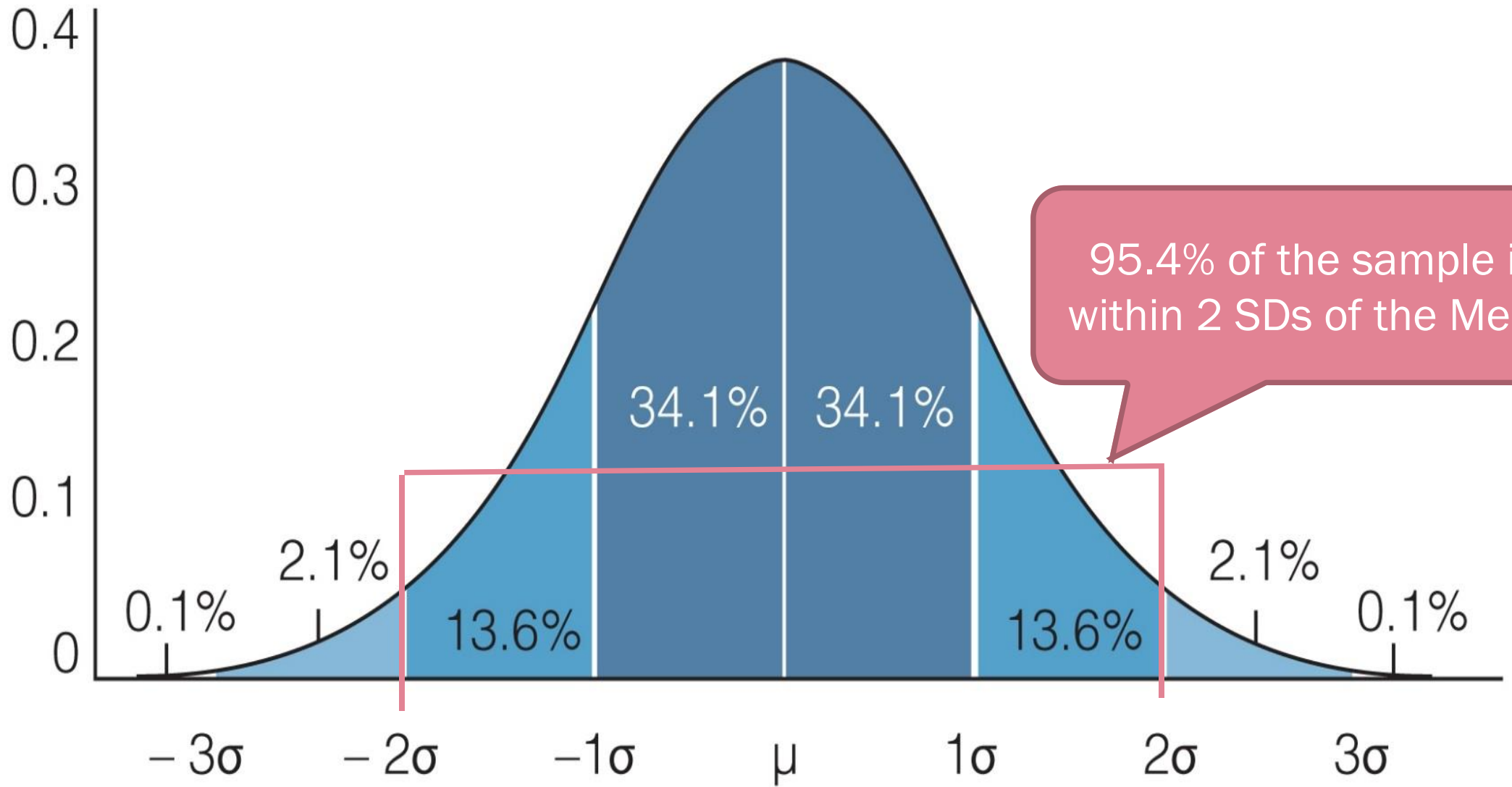
- A. 5%
- B. 9%
- C. 10%
- D. 15%
- E. 19%

(# of new cases) / (# at risk). For 2016, there are 270 new cases and 2700 people at risk; therefore, the incidence is 10%.

Do not forget to limit the incidence value to only those at risk. The 300 infected from 2015 (10% of 3000) are no longer at risk for the disease—they already have it!

A research study of 100 patients shows that their calcium levels range from 8.8 - 15.1 milligrams/deciliter (mg/dL), with a mean of 12.1 mg/dL. The calcium levels fall in a normal distribution, with a standard deviation of 1.0 mg/dL. Based on this study, we know that the percentage of calcium values below 10.1 is approximately

- A. 1%
- B. 2%
- C. 5%
- D. 8%
- E. 16%



95.4% of the sample is within 2 SDs of the Mean

10.1 11.1 12.1 13.1 14.1 mg/dL

A medical student seeks to examine the effect of sleep on passing an exam. She gives a survey to her classmates right before the exam to find out whether or not they slept the previous night, and then obtains data on the exam scores. This project is an example of what kind of study design?

- A. Quasi-experimental
- B. Longitudinal
- C. Historical Cohort
- D. Prospective Cohort
- E. Randomized Controlled

Two girls were born to the same mother, on the same day, at the same time, in the same month and year and yet they're not twins. How can this be?

- They're in a set of triplets

What is the car's parking spot number?



A close-up photograph of Michael Phelps swimming in a pool. He is wearing a black swim cap with an American flag design and blue-tinted goggles. His mouth is open in a shout or exertion, and water is splashing around his head. The background is a clear blue pool.

GOALS SHOULD NEVER BE EASY.

MICHAEL PHELPS
QUOTE-O-MATIC.COM