Intro to Biostatistics: Fundamentals of Measurement

Branko Miladinovic, PhD
Luis Maldonado, MD

August 18, 2010

USF HEALTH
CENTER FOR EVIDENCE BASED MEDICINE

UNIVERSITY OF SOUTH FLORIDA

Life of a (medical) researcher

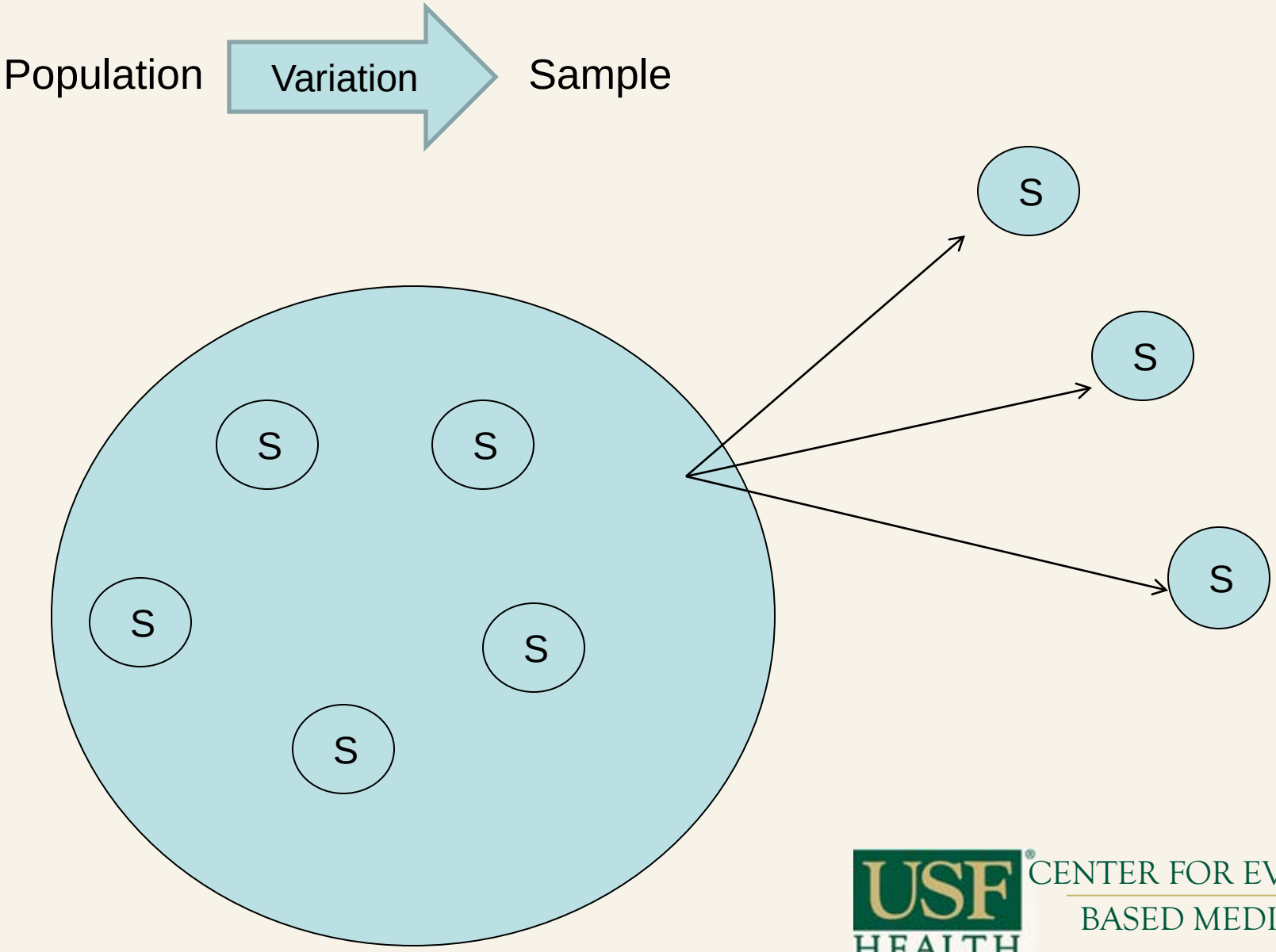
Greenhalgh, T. BMJ 1997;315:364-366

Objectives

- Describe basic biostatistics & their applications under Patient Intervention Comparison Outcome decision making paradigm.

- Interpret basic statistical results and their significance (estimation vs prediction)

Three Fundamental Concepts of Statistics:



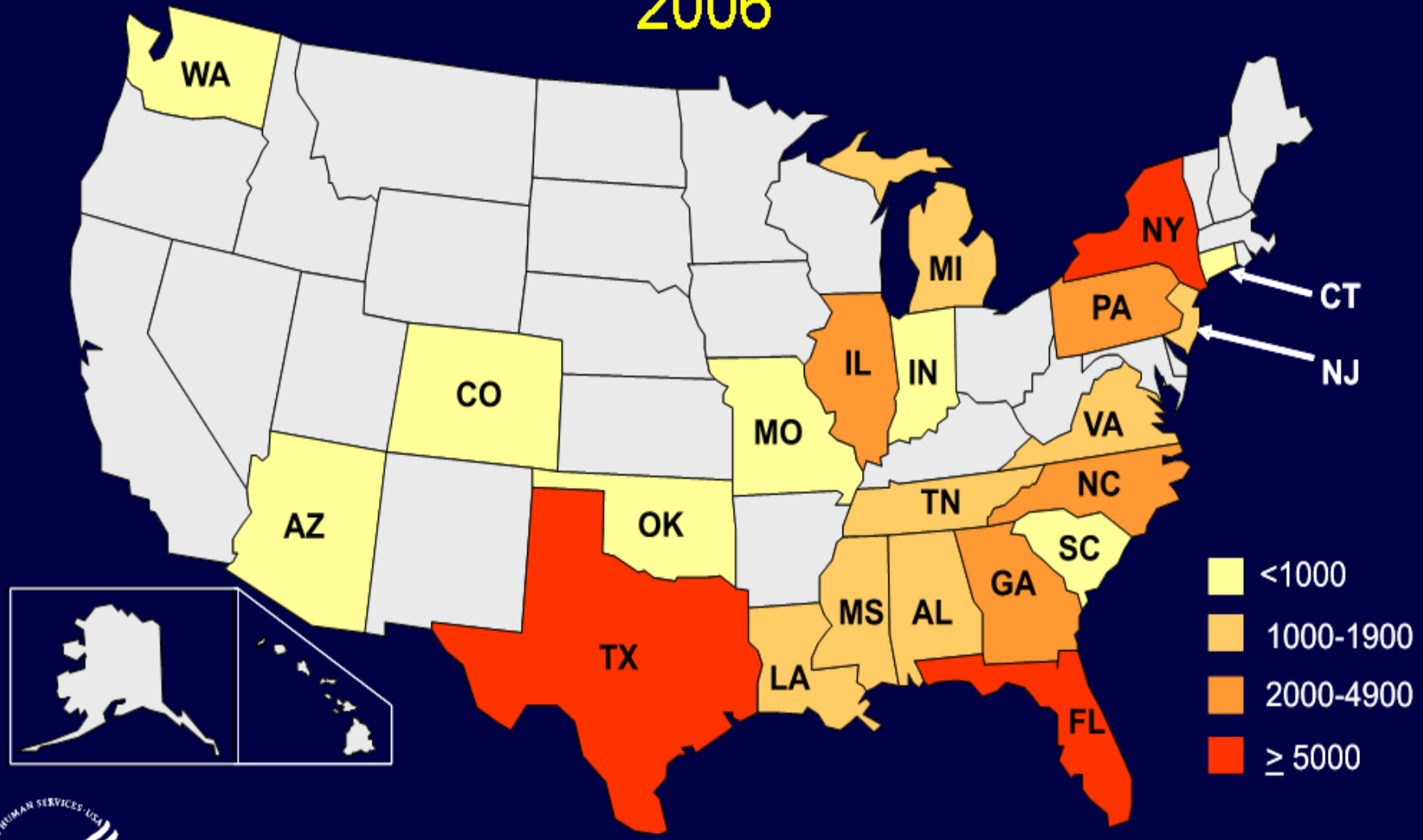
Consider the following statement by the CDC:

"There were 56,300 new HIV infections in the US in 2006"

Q1: Define population, sources of variation, sample.

Q2: How informative is the statement?

Estimated Number* of New HIV Cases—22 States
2006



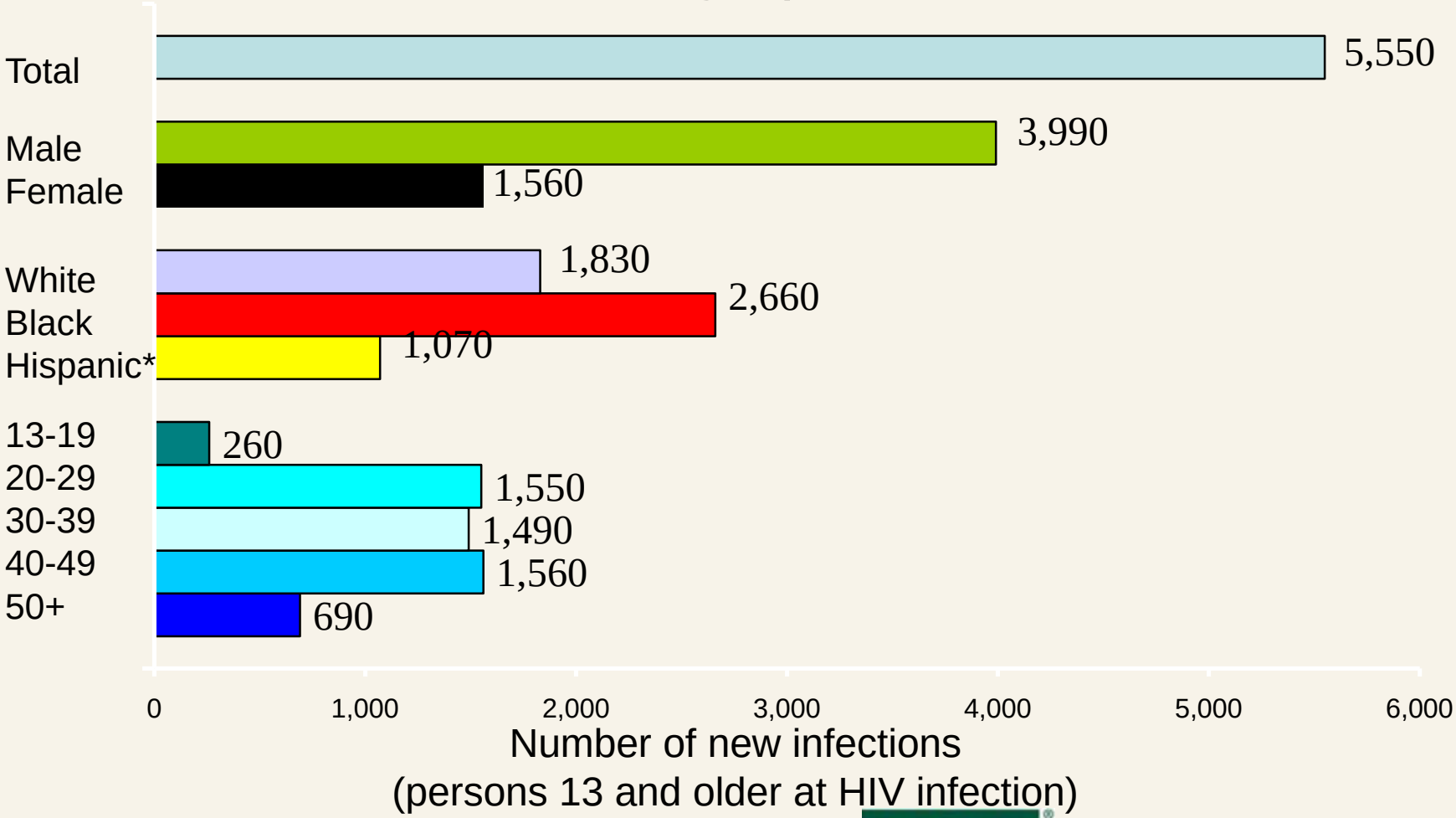
*Rounded to the nearest 100. Data have been adjusted for reporting delay.

The CDC reported:

- "Based on the stratified extrapolation approach the incidence of HIV in the US for 2006 was 56,300 new infections (with a 95% confidence interval of 48,200 to 64,500) "

- "The estimates were stratified for estimation by sex, race/ethnicity, age at infection and transmission category. There were 67 strata in total. The incidence estimates were adjusted for risk redistribution, and rounded to the nearest 100."

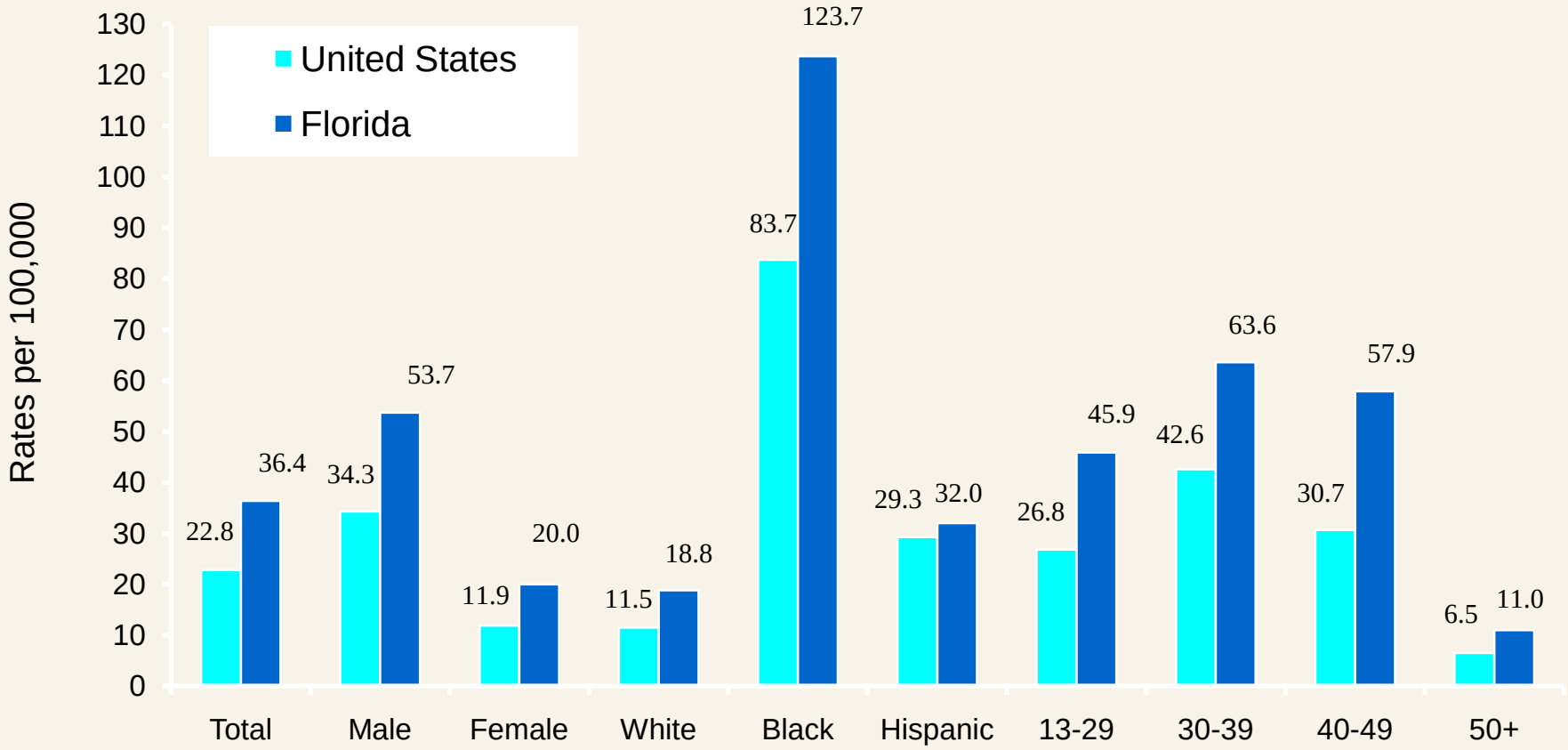
2006 Florida HIV Incidence Estimates Demographics



USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

HIV Incidence Estimates United States and Florida: 2006

(Rates per 100,000 in persons aged 13+ years)



In Florida, the Hispanic category includes Asian/Pacific Islanders, American Indian/Alaskan Natives, and mixed races. These races/ethnicities represent approximately 6% of the diagnosed cases in this subcategory.

Data Types

- Categorical: the data have “categories” instead of numeric values. (ex: male/female, disease/no disease, red/orange/yellow)

- Dichotomous: Categorical variable with only two possible categories.

- Continuous: this means the variable can take on a range of possible values. (weight, bp, height, etc)

Data Levels

- **Nominal:** Meaning- “name” There isn’t an order or value to the category. Examples: Purple vs. blue scrubs, TGH vs. SJH

- **Ordinal:** “order” Here, you can put some order to the categories, but you can’t give numeric value in between. Example: (bad, good, very good). This is commonly seen with Likert scales.

- **Ratio:** This is the highest level of data. You can categorize it, put it in order, give a value in between categories and there’s an absolute zero. Examples: Weight and Height.

Categorical and Continuous Data

- **Categorical data**: yes/no, male/female, disease/no disease

- **Continuous data**: weight, height, scores, blood values, etc.

Basic Descriptive Stats for Categorical Data

- Remember- you can't take an average of yes/no. (well, some people have *tried* to put that in papers…)

- So, how do we describe categorical data?
  - N, or n
  - Frequencies
  - Percentages
  - Prevalence/Incidence

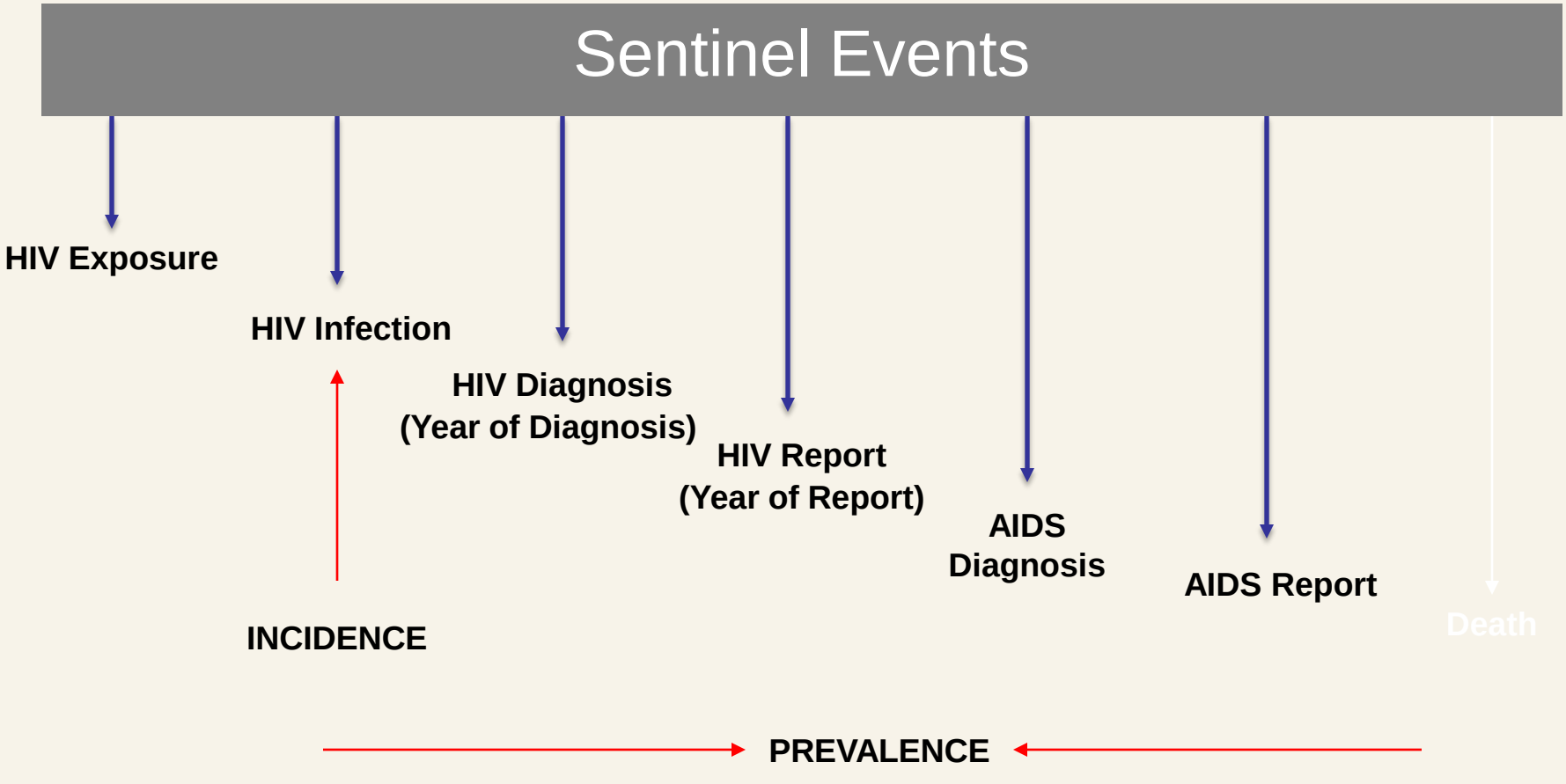
- Prevalence is the number of people living with HIV infection at the end of a given time period.

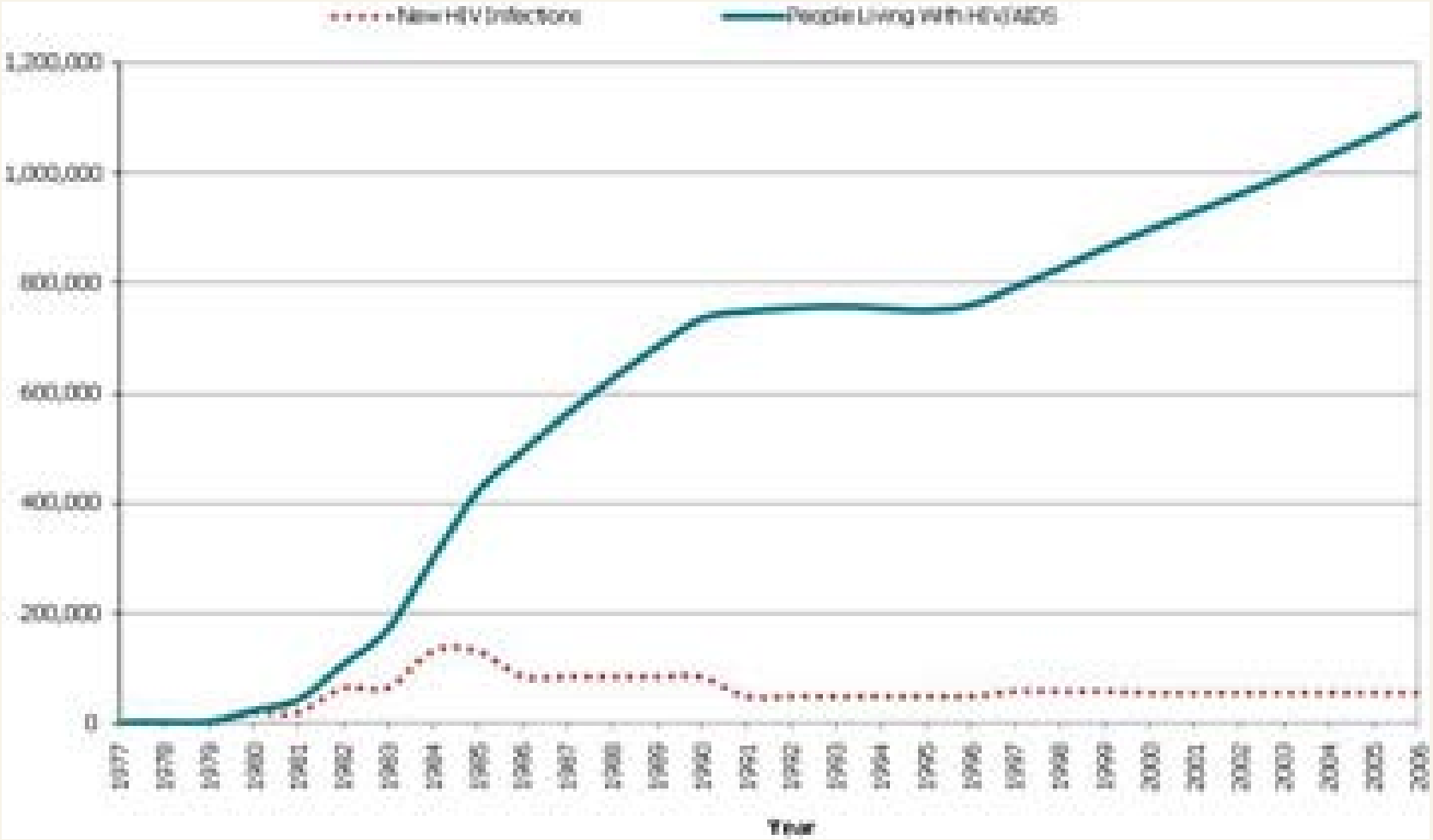
- Incidence is the number of new infections that occur during a given time period.

- Fatality Rate = Number of those who died within a given group.

Rates are often defined as person-year: If 100 people are followed for 6 years, we accumulate 600 person-years. If we observe 60 events, then the incidence rate is 60/600 = 0.1 events per person year.

HIV/AIDS Surveillance





USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

Basic Stats: Descriptive Stats for continuous data

- **N, or n:** We need to know how many people were in the sample. Results drawn from a sample with n=5 aren't very likely to be reliable. However, a sample of n=100 will make you feel a little more comfortable.

- **Central tendency**: Mean, median, mode

- **Variation:** Standard deviation, variance, standard error

- Exercise 1: HIV-AIDS patients are at an increased risk of a variety of diseases. Assume that below are the numbers for TB patients screened for HIV-AIDS*.

Age (Years)	Number Screened	Number Positive	Prevalence (%)
1-20	50	17	34.00
21-40	161	73	45.30
41-60	38	14	36.83
>60	8	2	25.00
Total	257	106	41.24

*G. Pennap, S. Makpa & S. Ogbu : The Prevalence of HIV/AIDS Among Tuberculosis Patients In a Tuberculosis/Leprosy Referral Center in Alushi, Nasarawa State, Nigeria.. *The Journal of Epidemiology.* 2010 Volume 8 Number 1.

1. Assume the table reports the number of new cases at the end of the year (incidence). Calculate incidence rate for each group and interpret.

2. Calculate and interpret the absolute risk, relative risk and odds ratio between age groups 1-20 and 21-40.

- Ratio <1: Exposure is Protective
- Ratio =1: No Difference
- Ratio >1: Exposure is Risk Factor

2x2-table.xls

- Why Ratios? Because they are more sensitive to the baseline!

7% difference in survival between 97% and 90% is the same as a 7% survival difference between 14% and 7%.

But (odds) ratios are 97%/3% = 32.3, 90%/10% = 9 definitely not equal.

- Note: For rare events RR $\approx$ OR.

- Example 2: Calculate the effectiveness of the detection of HIV associated neurocognitive disorder (HAND) in 30 antiretroviral drug exposed persons.

True Disease\Rating	Normal (1)	Questionable (2)	Abnormal (3)	Total
Normal	8	4	4	16
Abnormal	3	3	8	14
Total	11	7	12	30

- Methodology: Graph (1 – specificity) on x-axis and sensitivity of y-axis and find the area under the ROC curve (Note there is no establish cut-off point so all three will need to be graphed).

- Area = Probability that the test is effective.

- Sensitivity is the probability that the symptom is present given that the person has a disease.

- Specificity is the probability that the symptom is not present given that the person does not have a disease.

- Receiver operating characteristic (ROC) curve is a plot of (1 – specificity) vs sensitivity. The area under the curve gives the probability that that a test accurately distinguishes between normal and abnormal.

Survival Analysis: time-to-event

Time to event : Is the time from entry into a study until a subject has particular outcome of interest (event):

- Time to death
- Time to relapse of a disease
- Time to recovery from illness
- Length of stay in a hospital

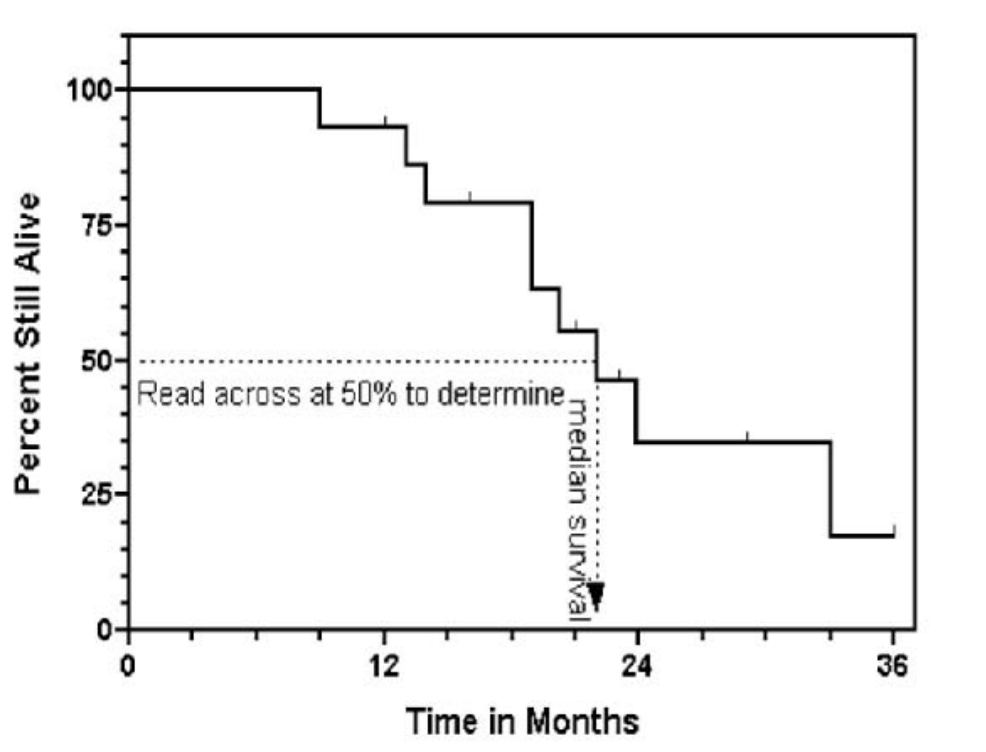
- **Kind of survival studies**
- Clinical trials
- Prospective cohort studies
- Retrospective cohort studies

- Survival function: S(t), the probabilitythat a person survives longer than some specified time ***t.***

- The 5 year survival probability is 30% means “70% of patients will die within the first 5 years”

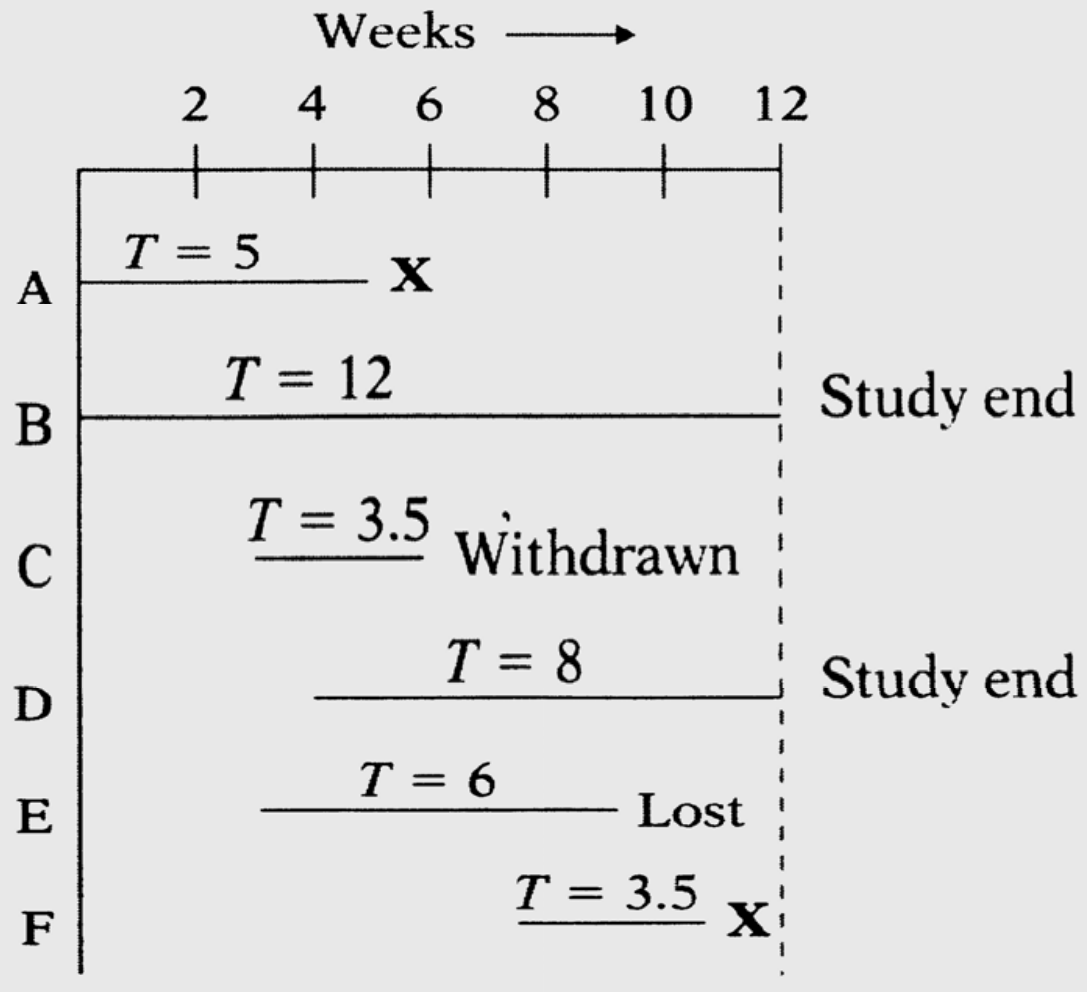
Survival curve: median survival time

- Median survival time: the time at which half the subjects have died and half are still alive. It isn't estimable if the survival curve doesn't reach to 50% or below.

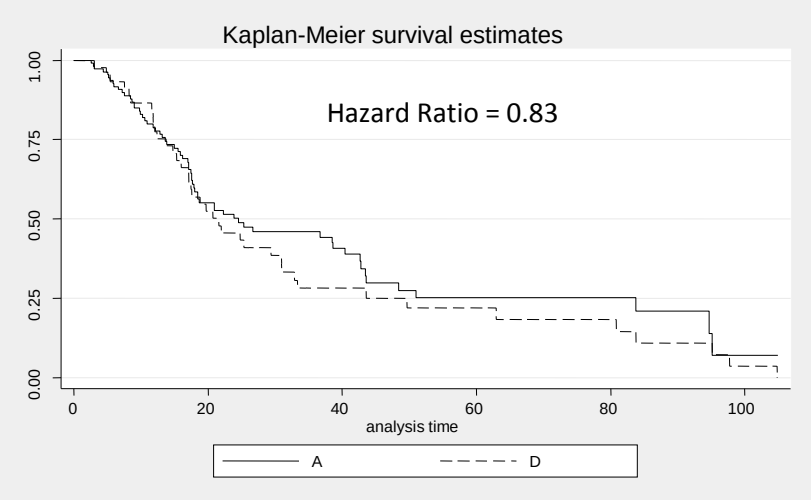


USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

- Censored: subject does not experience event of interest (lost to follow-up, withdrawal).



Example: RCT comparing the effectiveness of two drugs A (experimental) and D (control) (P-value = 0.3).



Interpretation:

- √The hazard of the event for those in the treatment group is only about 83% of the hazard for those in the placebo group;
- √ The hazard at any time for patients on the treatment group is 83% that of patients in the placebo group;
- X the survival time in the treatment group is 83% longer than that in the placebo group. HR can't be directly translated into information about the duration of time until event.
- X the patients in the control group died 17% faster than that in the treatment group.

Example 2: Research Question:

- USF Medical School wants to implement a new curriculum for first year medical students that addresses issues specific to HIV-AIDS patients.

- Of particular interest are students' knowledge, attitude and experience with HIV-AIDS patients.

- Q: How do we test whether the curriculum is successful?

- Answer: Design a pre-post curriculum survey that will test the changes in student knowledge, attitude and experience.

P =

I =

C =

O =

Sampling Methods

- Simple Random
- Systematic (every kth person)
- Convenience
- Consecutive
- Stratified
- Combinations

- What sort of questions should we ask?

- Knowledge questions should be asked as T/F.

Why?

- Attitude/Experience should be on a Likert 1-5 scale.

Why?

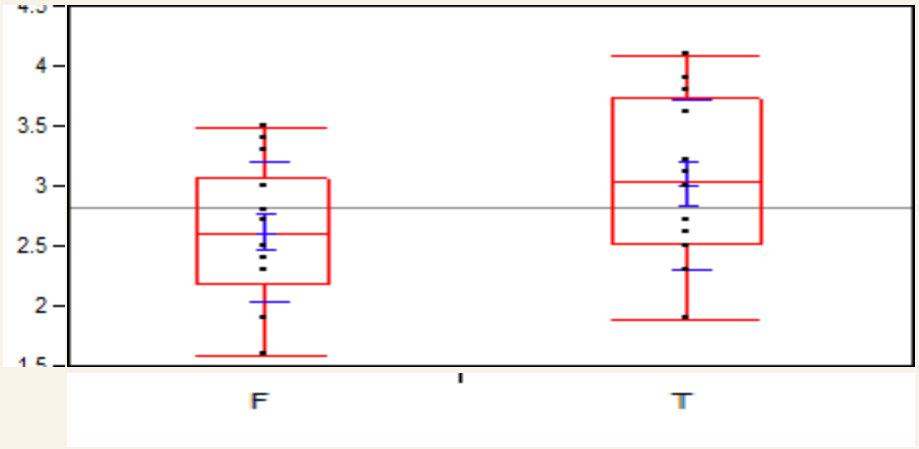
A few examples:

- Q: You can get HIV from shaking hands or drinking from the same glass of an infected person?

- Q: Physicians in private practice do not have a responsibility to treat HIV-AIDS patients.

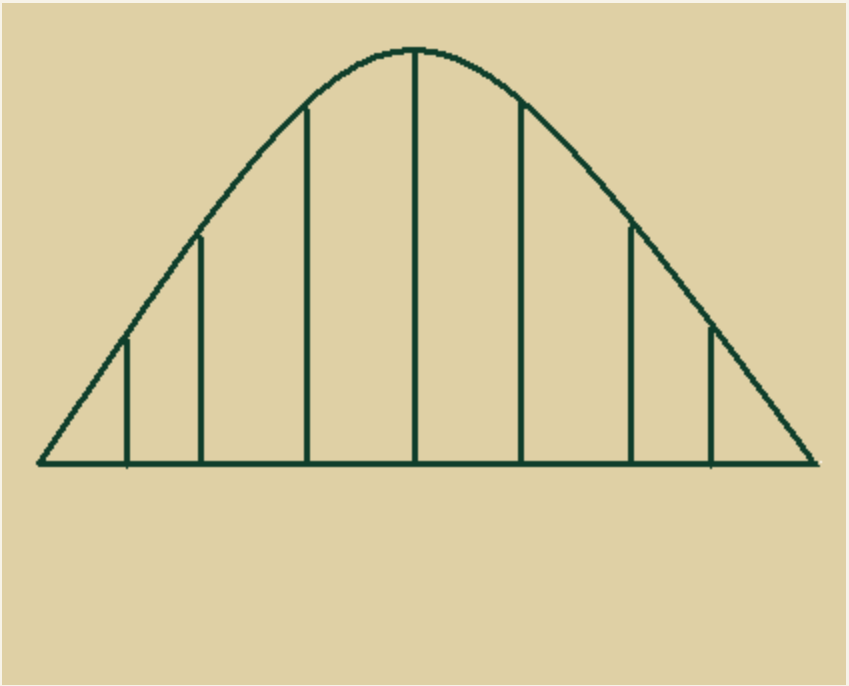
- Q: I feel comfortable around HIV-AIDS infected people.

- Compare experience and knowledge outcomes.



Level	Number	Mean	Std Dev
F	14	2.61	0.58
T	16	3.01	0.69

The Bell Curve!
(The Normal Probability Distribution)



<u>Important Characteristics</u>:

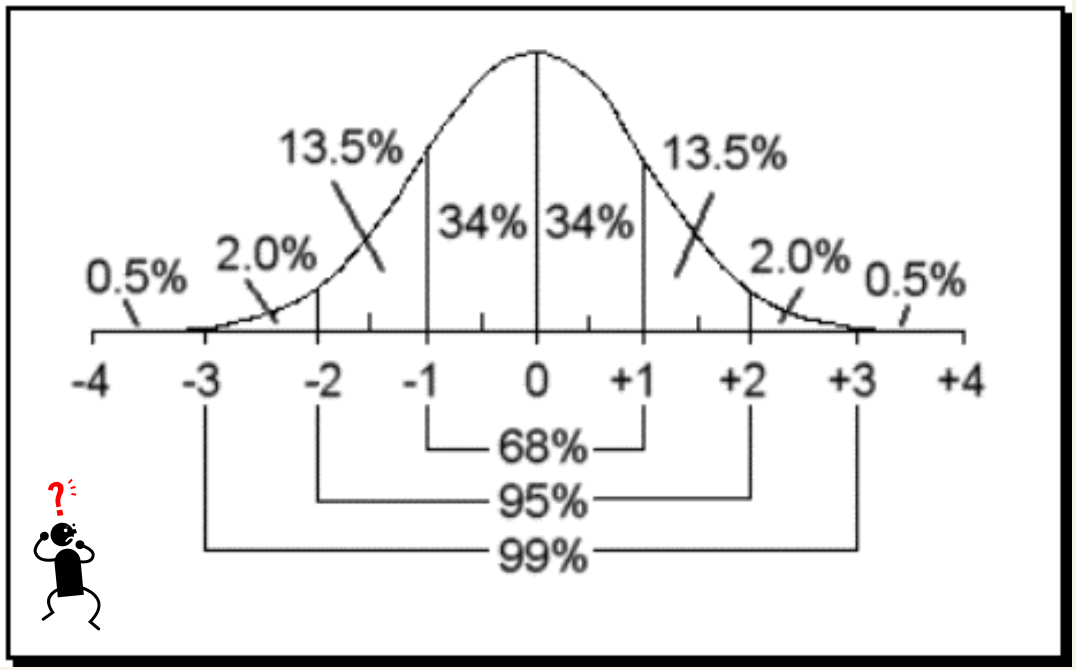
-Mean, median and mode have the same value

-Bell-shaped & symmetric around mean

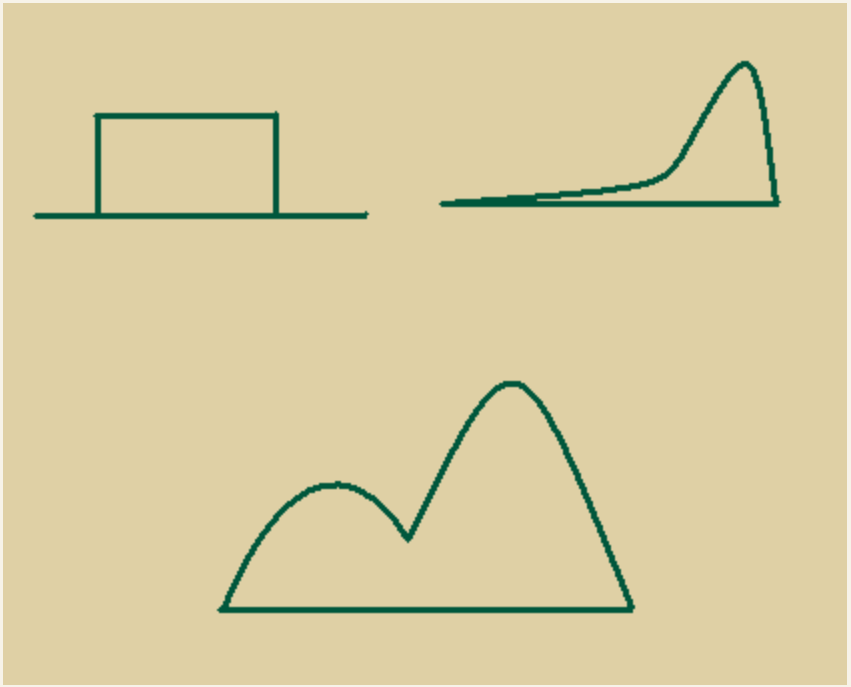
-Total area under the curve =1

AKA: Parametric distribution (*parameterized* by mean and standard deviation), Gaussian Distribution

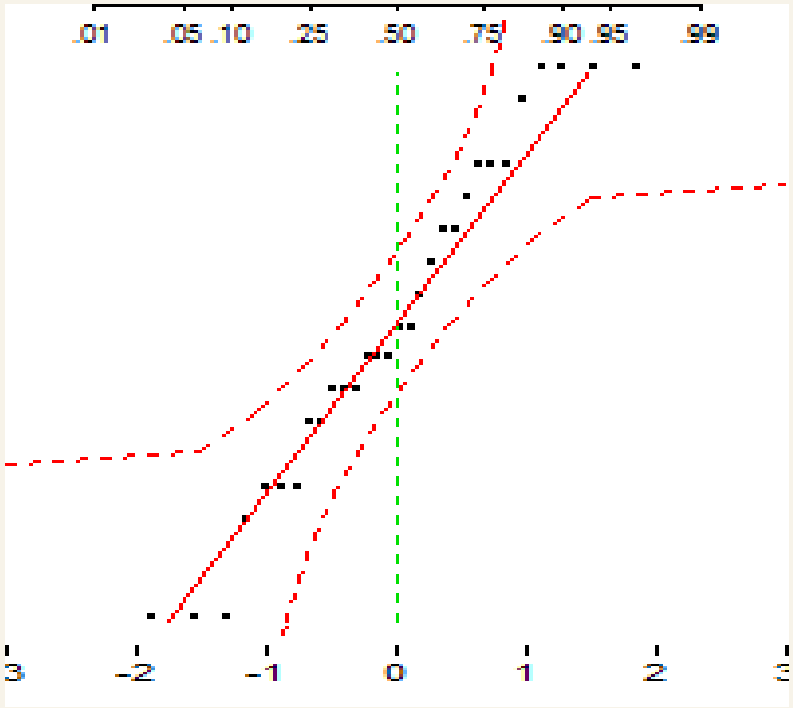
The Empirical Rule



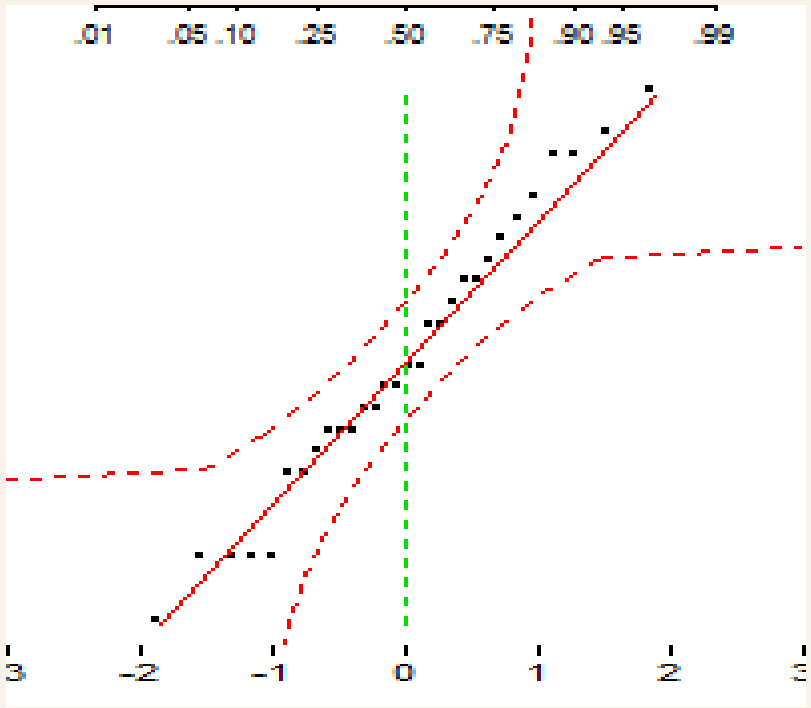
Non-Normal (non-parametric) Distributions



- Normality can be ascertained using quantile-quantile (QQ) plots. Deviations from the straight line suggest data not normal.



QQ plot for Attitude



QQ plot for Experience.

Correlation vs Regression

Correlation is a measure of a linear or non-linear relationship between two continuous variables.

Correlation Coeff. Value	Direction and Strength of Correlation
-1	Perfectly Negative
-0.5	Moderately Negative
-0.2	Weakly Negative
0	No Association
+ 0.2	Weakly Positive
+ 0.5	Moderately Positive
+ 1	Perfectly Positive

- Regression analysis evaluates the relative impact of a predictor variable X on a particular outcome Y.

$$Y = a + bX + e$$

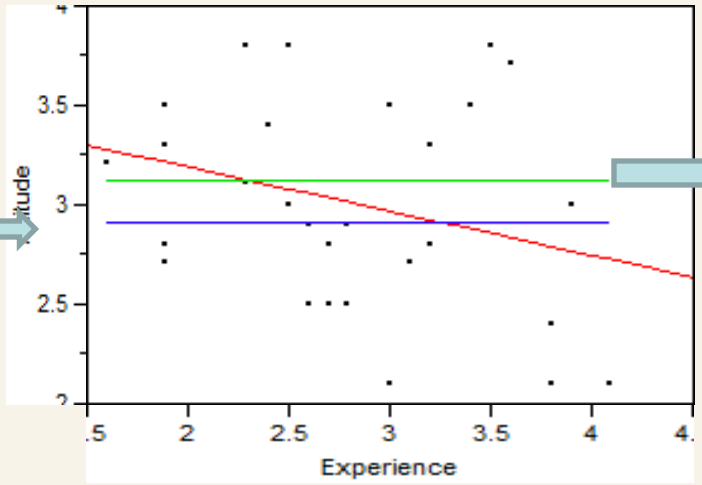
- Can be linear or non-linear (multiple linear, logistic…etc)

Simple linear regression if Y continuous

Simple logistic regression if Y is dichotomous.

Baseline response to 30 question in each survey domain:

mean correct knowledge



mean incorrect knowledge

Attitude = 3.6419388 - 0.222384 Experience

R-squared = 0.079 (Expressed as percent signifies percent variability in Y explained by the variability in X)

Linear Corr Coefficient = - 0.28

- Let

  Outcome Y = correct/incorrect on knowledge (T/F).
  Predictor X = experience score

- Using logistic regression we get odds ratio OR = 2.73,

which means that for each increase of 1 on the experience score, the odds of getting a correct knowledge answer increase by 2.73.

Questions ?