Data Analysis and Presentation

Lauren DiBiase, MS, CIC
Associate Director
Public Health Epidemiologist
Hospital Epidemiology
UNC Hospitals

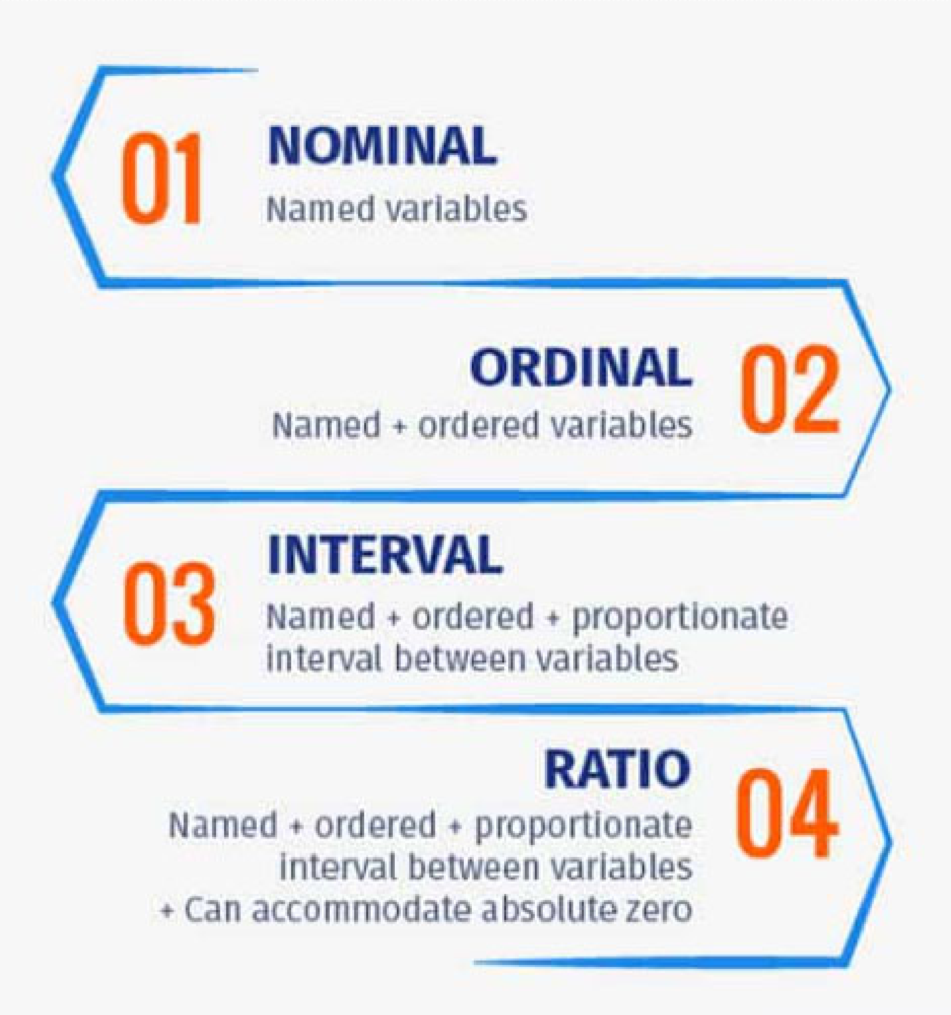
Types of Data

- **Discrete data** – counted in whole units(e.g., ventilator days)

- **Continuous data** – measurement of things with an infinite number of possible values between the minimum and maximum (e.g., temperature)

Counted data vs. measured data

Scales of Measurement



01 **NOMINAL**

Named variables

02 **ORDINAL**

Named + ordered variables

03 **INTERVAL**

Named + ordered + proportionate interval between variables

04 **RATIO**

Named + ordered + proportionate interval between variables + Can accommodate absolute zero

Nominal Scale

- Simplest level of measure
- Use of categories – mutually exclusive groups
- No order among classifications

  Example: Handwashing observations-compliant or non-compliant

Ordinal Scale

- Each category is distinct
- Each category has a relationship to each other

  Example: Cancer staging: 1, 2, 3

Equal Interval Scale

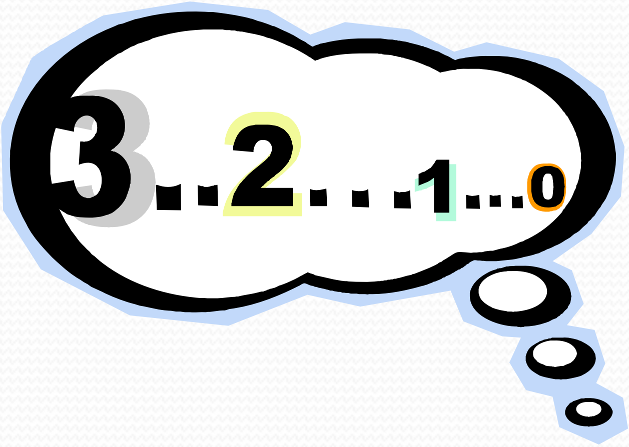
- Ordinal data
- Exact distance between any 2 points on the scale is known

  Example: Blood pressure

Ratio Scale

- Equal interval measurements that have a true zero point

  Example: Distance



Question

- The ICP fills out a survey after an educational program. After having learned about the product XYZ, how likely are you to consider implementing it in your hospital ?

Extremely unlikely 1 – 2 – 3 – 4 – 5 Extremely likely

What type of scale is this?

A. Nominal
B. Equal Interval
C. Continuous

Measures

- Absolute
  - Simplest type of measurement
  - Also known as counts or frequencies
  - e.g. there were 160 cases of *C. difficile* last year
- Relative
  - Includes a denominator
  - Useful for comparisons
  - e.g. there were 160 cases of *C. difficile* out of 120,000 patient days last year

What Makes a Rate?

THERE IS A FINE LINE BETWEEN NUMERATOR AND DENOMINATOR

1. Numerator (top number)
   - *e.g., number of infections*
2. Denominator (bottom number)
   - Represent the population at risk of becoming part of the numerator
   - Ideally, should incorporate time and can account for risk factors such as device use (e.g., device-days), length of stay (e.g., resident-days)
     - *e.g., number of residents [proportion]*
     - *e.g., number of resident-days, number of device-days [incidence density/rate]*
3. Time Frame
   - *e.g., day, week, month*

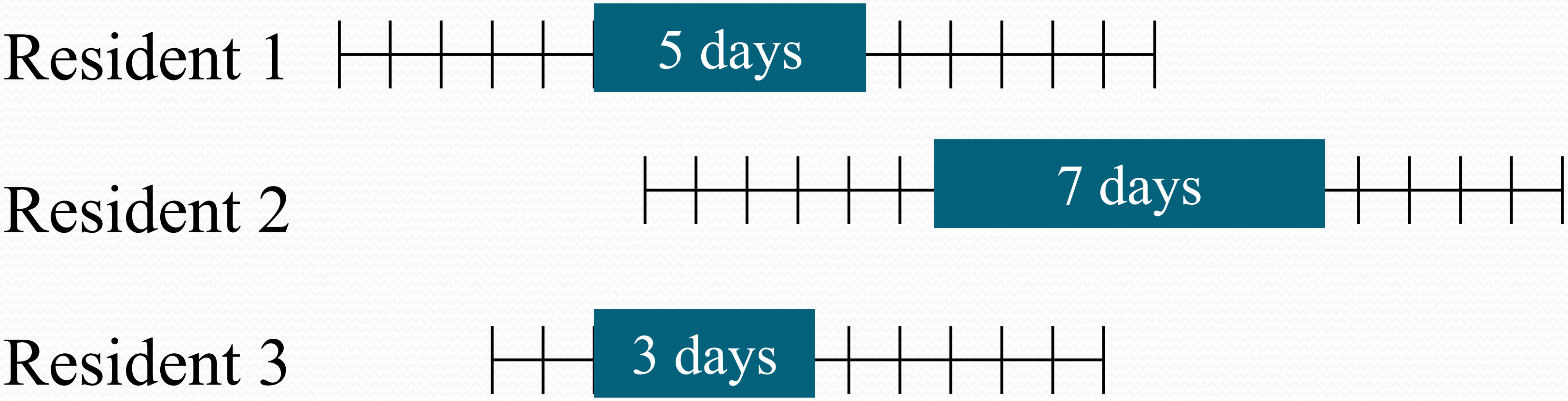
Denominators

- Represent the population *at risk* of becoming part of the numerator

- Often, the most difficult data to obtain, but essential for comparisons

- Ideally, should incorporate time and can account for risk factors such as device use (e.g., device-days), length of stay (e.g., resident-days)

What is a Resident/Device-Day?



Resident 1 — 5 days

Resident 2 — 7 days

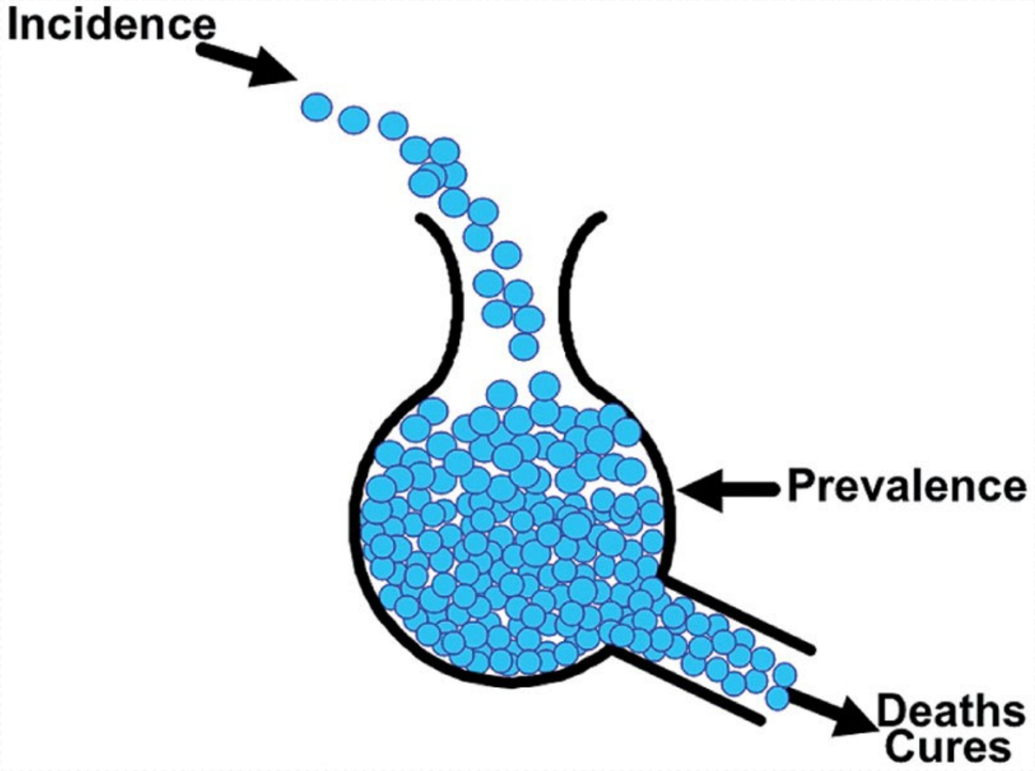
Resident 3 — 3 days

=15 resident-days, device-days, etc.

- More informative than simply saying “3 residents”

Rate Measures

- Prevalence
- Incidence
- Attack Rate



Prevalence

- Prevalence: the total number of cases of disease existing in a population at a point in time.
  - *e.g., # of MRSA cases per population* ***on*** *March 8*

$$\frac{\text{Count of } \textbf{existing} \text{ cases}}{\text{Number of people at risk}} \times \text{constant (e.g., 100 or 1000)} =$$

Incidence

- Incidence: the number of new cases of disease in a population over a period of time.
  - *e.g., # of **new** MRSA cases per population **during** March*

Count of **new** cases x constant (e.g., 100 or 1000) =
Number of people at risk

Question

- On June 1st, there were 25 surgical patients in the hospital. Two of these were post-op SSIs identified in May. During the month 5 additional SSIs were admitted. A total of 60 surgeries were performed in June. What is the numerator for a June incidence rate?

  A. 25

  B. 5

  C. 7

  D. 8.3

Attack Rate

- Attack Rate: the number of new cases of disease out of the population at risk.
  - Related to incidence but always uses 100 as the constant, so it is always expressed as a percentage.
  - Often used for outbreaks or clusters that occur over a short period of time
  - *e.g., % of residents with MRSA during outbreak in LTC A in March*

$$\frac{\text{Count of new cases}}{\text{Number of people at risk}} \quad \text{x} \quad \mathbf{100} \quad =$$

Question

- 15 persons were infected with Salmonella at a picnic where 75 ate potato salad. What was the attack rate of salmonella among those who ate potato salad?

A. 15%
B. 0.20
C. 18%
D. 20%

Mortality Rates

- Crude Mortality Rate:

$$\frac{\text{\# persons dying}}{\text{Population at risk}} \text{ x } k$$

- Cause-Specific Mortality Rate

$$\frac{\text{\# persons dying from a specific cause}}{\text{Population at risk}} \text{ x } k$$

- Case Fatality Rate

$$\frac{\text{\# persons dying from a specific disease}}{\text{\# of persons with the disease}} \text{ x } k$$

Constant “K” is usually 1000 or 100,000

Question

- During the winter of 2017, 645 persons died from influenza related illness in Columbus. The population of Columbus was 1.2 million. What was the crude mortality rate?

A. 54 per 100,000
B. 5.3 %
C. 54%
D. 0.005%
E. Unknown

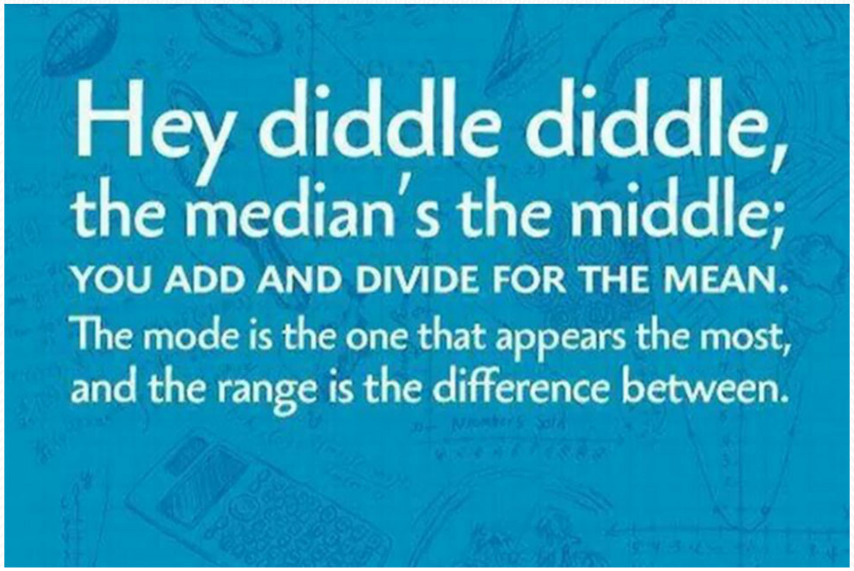
Question

- During the winter of 2017, 645 persons died from influenza related illness in Columbus. The population of Columbus was 1.2 million. What was the cause-specific mortality rate?

  A. 54 per 100,000
  B. 5.3 %
  C. 54%
  D. 0.005%
  E. Unknown

Measures of Central Tendency

- Mean: average of a group of numbers
- Median: middle number in an ordered group of numbers; also defined as the 50th percentile
- Mode: most common value in a group of numbers



Hey diddle diddle,
the median's the middle;
YOU ADD AND DIVIDE FOR THE MEAN.
The mode is the one that appears the most,
and the range is the difference between.

Measures of Dispersion

- Range: the largest value minus the smallest value
- Standard deviation: describes the variability or dispersion in the data set

Question

- What is the range for the following numbers?

2,3,4,5,8, 9, 10, 12, 14

Range = 14 – 2 = 12

- What is the mean?

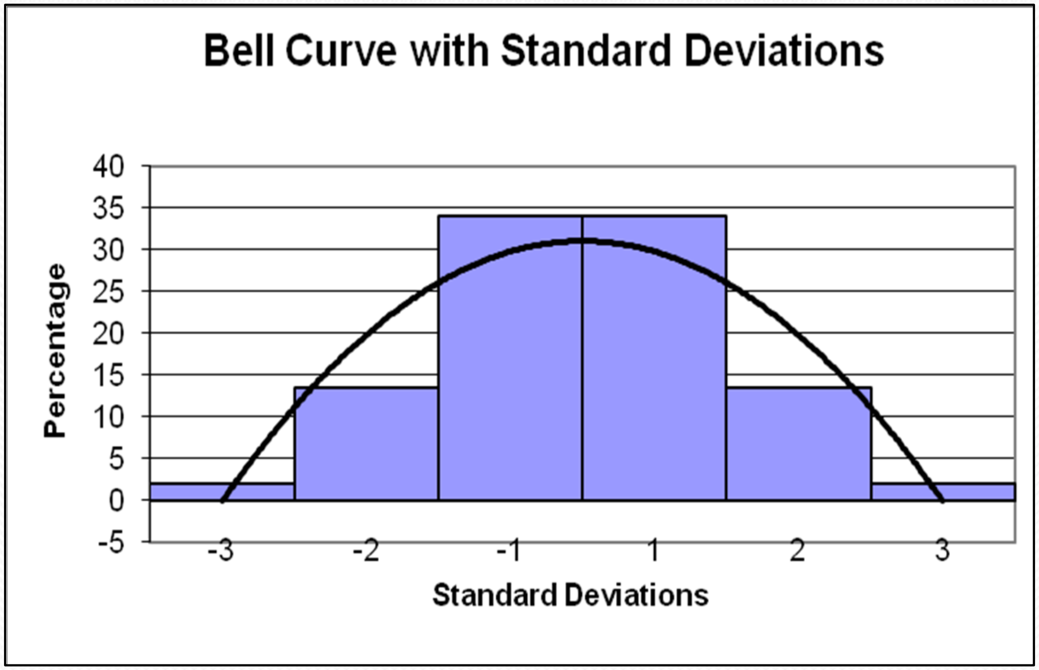
Mean = 67/9 = 7.44

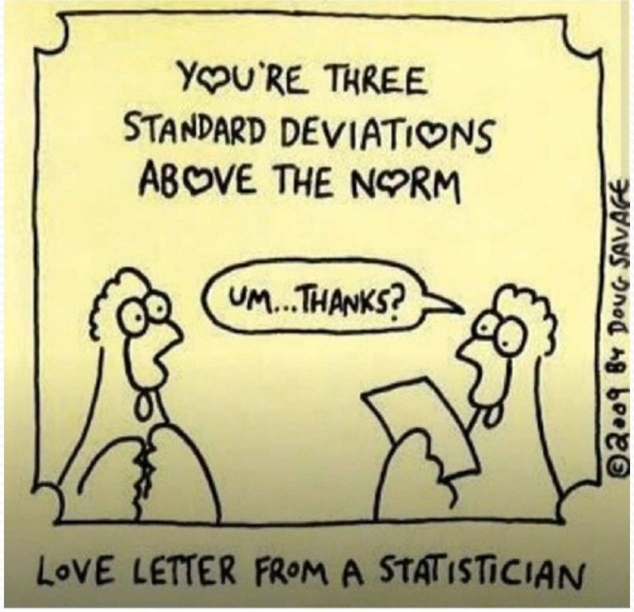
- What is the median?

Median = 8

Standard Deviation

- In a normally distributed data set, the spread of values is even on both sides of the mean




YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
©2009 BY DOUG SAVAGE
LOVE LETTER FROM A STATISTICIAN

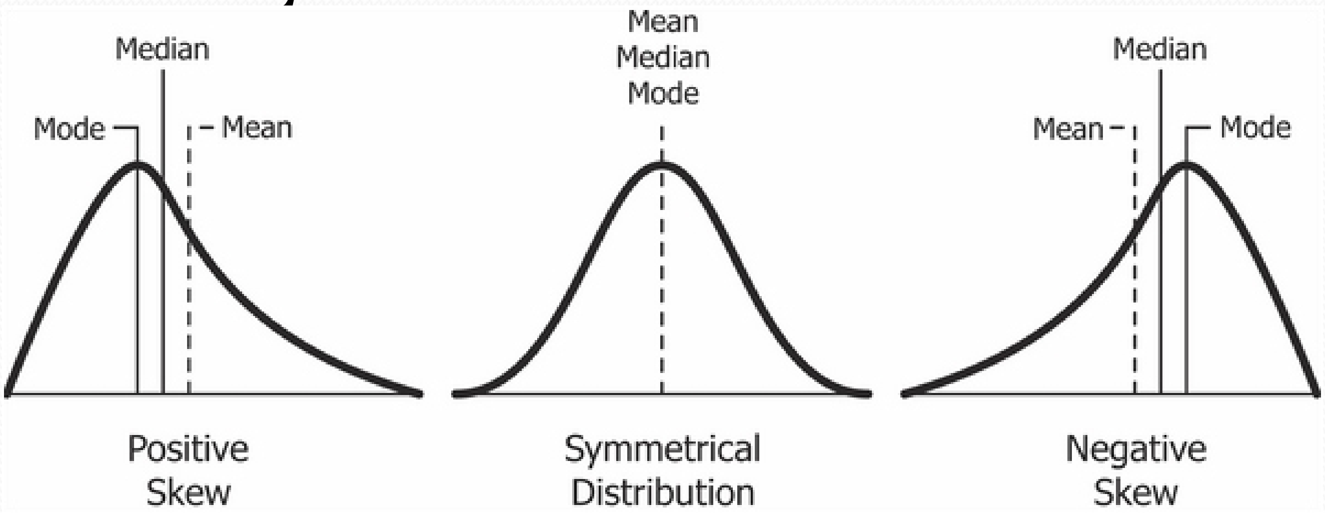
68% of values + 1 SD

95% of values + 2 SD

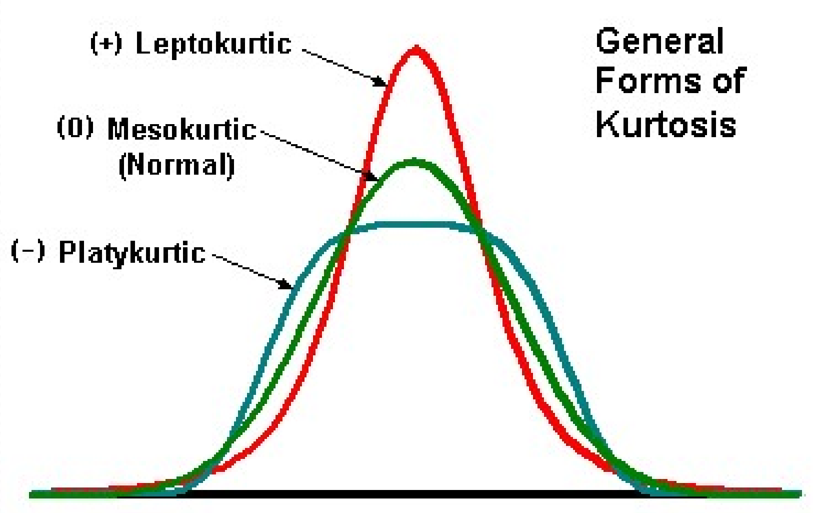
99% of values + 3 SD

Measures Frequency Distribution

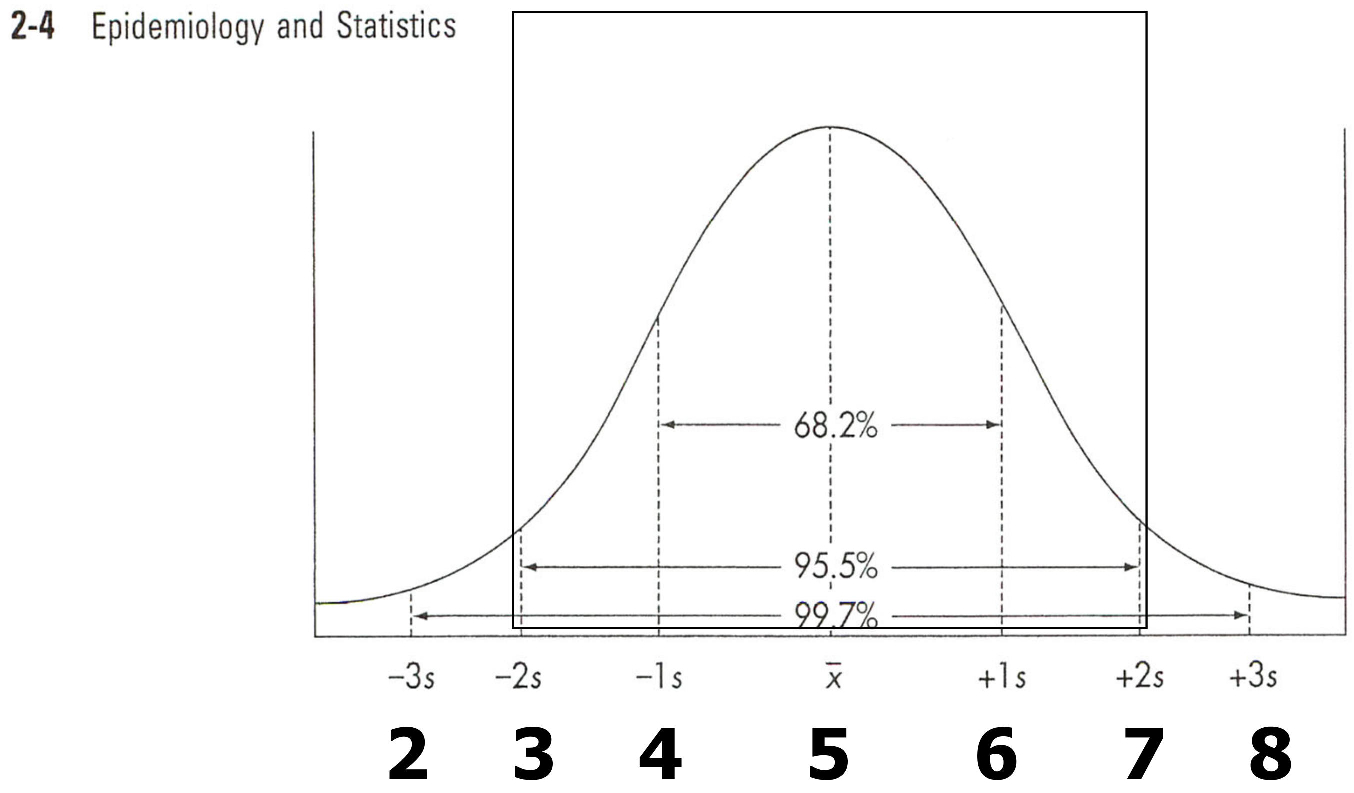
- Skewness – asymmetrical distribution



- Kurtosis – how flat or peaked a curve is



A study of the length of stay of patients with HAI showed an average excess stay of 5 days, with a standard deviation 1, what percentage of the patients had LOS between 3 and 7 days?



Question

- What percentage of patients had LOS between days 3 and 7 days?

A. 68.2%
B. 95.5%
C. 98.7%
D. 67.5%

Formulas

For Rates:

- # VAE/Vent Days X 1000
- # CLABSI/CL Days X 1000
- # CAUTI/Foley Catheter Days X 1000

For device utlization:

- # Device days/# Patient days

Question

- Using a device associated infection formula, calculate the rate for 1000 vent days:

4 cases of VAE

800 ventilator days

Question

- Calculate the device utilization rate for a facility which has had 800 vent days and 4000 patient days.

What Makes a Standardized Infection Ratio (SIR)?

1. Numerator (top number)

   =*number of observed infections*

2. Denominator (bottom number)

   =*number of expected or predicted infections*

- Number of predicted infections =

  calculated based on your hospital's number of procedures, device days, risk factors, nursing units

  compared to a standard infection rate (e.g., historical data, state data, national data)

Standardized Infection Ratio

- SIR = $\dfrac{\text{\# observed infections}}{\text{\# predicted infections}}$

- SIR >1.0 → *more infections than predicted*
- SIR <1.0 → *fewer infections than predicted*

- *~LOWER SIRs are BETTER~*

SIR Interpretations

- **SIR=1**
  - The number of infections is the same as the number of expected infections
  - No progress has been made in reducing infections since the baseline period or compared to another standard population (e.g., all NC, all US).

SIR Interpretations

- If the **SIR is less than 1**
  - Fewer infections than predicted based on standard or baseline data
  - Infection reduction/prevention compared to standard or baseline data
  - 1 minus the SIR = percent reduction:
    For example, a SIR of 0.80 means that there was a 20 percent reduction from the standard population or baseline time period

SIR Interpretations

- If the **SIR is greater than 1**
  - More infections than predicted based on standard or baseline data
  - Infections are increased compared to standard or baseline data
  - SIR minus 1 = percent increase:
    For example, a SIR of 1.25 means that there was a 25 percent increase from the standard population or baseline time period

Question

- CLABSI rate = 4 CLABSI/284 line days
- Predicted Infections = 0.50

- What is the SIR?
- How would you explain the SIR to your administrator?

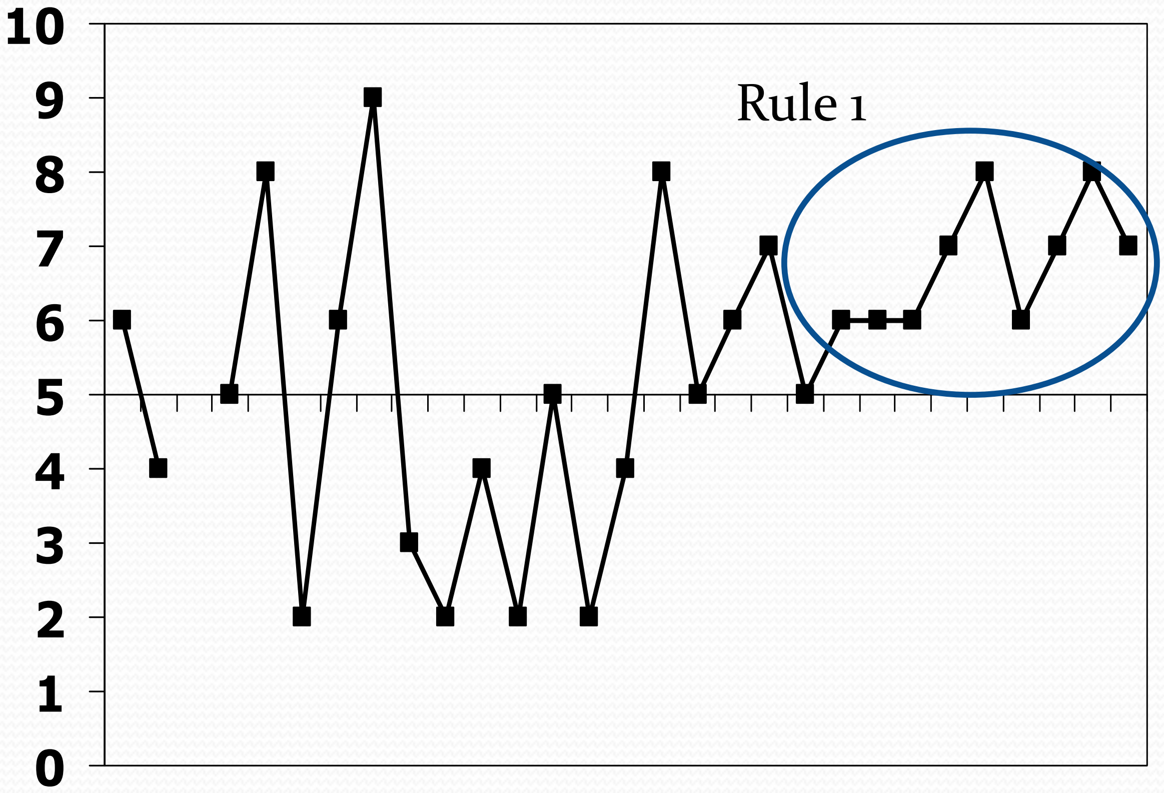
Determine the Significance-How?

- Practical Significance vs. Statistical Significance

- Make comparisons
  - *For example: over time, to other areas of facility, to other facilities (NHSN data)*
  - Remember to choose appropriate data for comparison (*i.e., same denominator units)*

- Apply a type of statistical test
  - *e.g., control charts (for time trends)*
  - P-values
  - 95% confidence intervals

Run charts



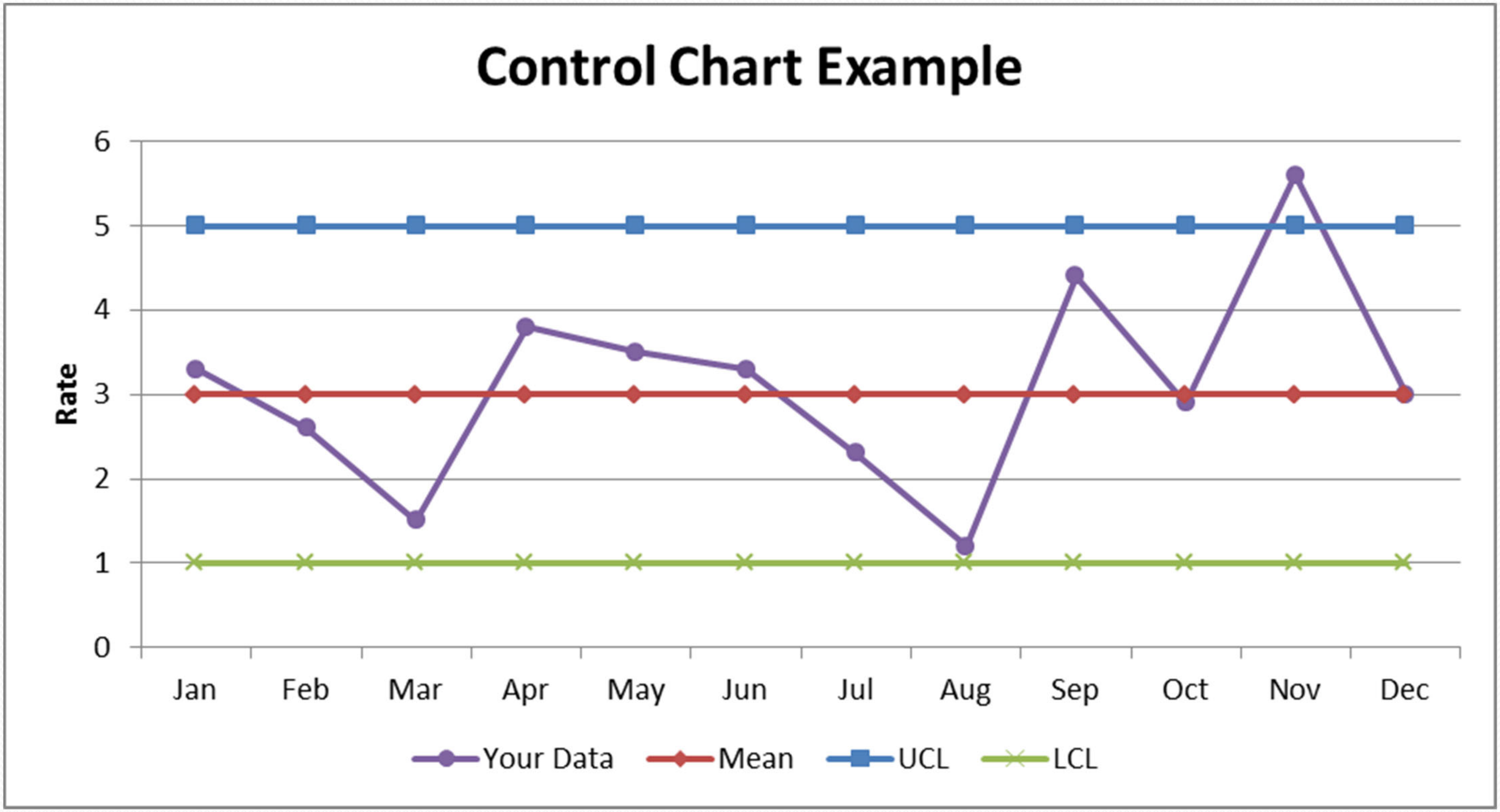
Rules used to detect variation

1. 7 or more consecutive points on either side of the median
2. 5 or more points either decreasing or increasing
3. 14 or more data points in a row going up or down

Constructing a Statistical Process Chart

- Collect the data
- Calculate mean & SD
- Set up chart- draw horizontal line at:
  - **Mean**
  - **UCL – 2 or 3 SD above mean**
  - **LCL – 2 or 3 SD below mean**
- Enter data points
- Interpret data as “in control” or “out of control”

Statistical Process Control Charts



Question

- A Statistical Process Control Chart:

A. Analyzes the data for deviations from the pooled mean of the samples
B. Should be used only to display the data
C. Should be used only when a Pareto Chart is inconclusive
D. Should be used when data is discrete

Statistical Inference

- Does NOT prove association
- Statistically significant – highly unlikely that results occurred by chance
- Not statistically significant – results could easily be attributed to chance alone

Hypothesis Testing

- Null hypothesis: values are equal
- Alternative hypothesis: values differ

- These statements are mutually exclusive
  - They cover all possible outcomes
  - In the end, only one can be selected

p=value: The probability that the observed difference (or a more extreme one) was caused by random chance if the null hypothesis was true.

Hypothesis Testing: Types of Errors

- α - Type I Error – Probability of rejecting a true null hypothesis (no difference)

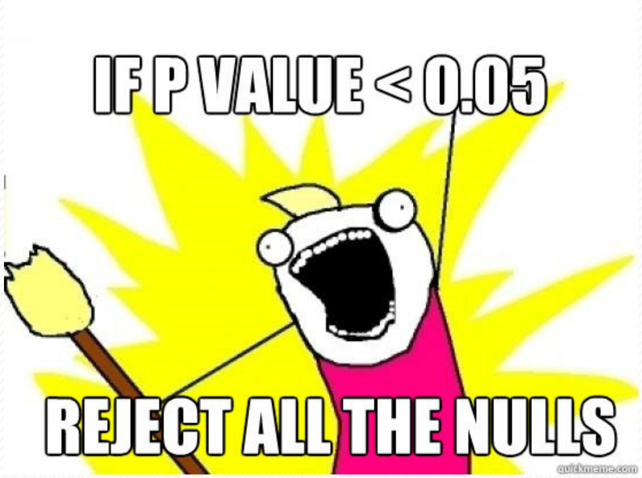
- β - Type II Error – Probability of not rejecting a false null hypothesis

P Value

- Probability that the difference does not reflect a true difference and is only due to chance.

- e.g., $P=0.05$ means that 95 out of 100 times your estimate was truly significant (another way to think about it - there is a 1 in 20 chance of committing a Type 1 (alpha) error

- Generally a level of $P<0.05$ is considered "statistically significant"



Power

- The ability of a test to detect a specified difference
- The ability to reject the null hypothesis when it is false
- Influenced by sample size

Question

- The probability of not rejecting a false null hypothesis is considered a(n):

  A. Type I error
  B. Type II error
  C. Alternative hypothesis
  D. Alpha error

Question

What is the probability of committing a Type I error if the P-value is 0.10?

A. 1 in 10

B. 1 in 100

C. 1 in 5

D. 1 in 20

Question

A pilot research study was conducted to compare the association between a new type of dressing and a unit's CLABSI rates. During the six month period prior to the intervention of the new dressing the unit's CLABSI rate was 2.06 per 1000 central line days. During the 6 months the dressing was trialed, the unit's CLABSI rate was 1.76 per 1000 central line days. The p-value was 0.03. What conclusion can be reached?

A. The new dressing may be associated with statistically significant lower CLABSI rates

B. The new dressing caused the decreased CLABSI rates

C. The new dressing should not be used

D. No significant statistical conclusions can be drawn from this pilot study

95% Confidence Intervals

- Means that you are 95% confident that the *true* average value lies within this interval.
- If spans the null value (1 for ratios), then not statistically significant

- Confidence interval size:
  - Wide: less confident with that estimate
  - Narrow: more confident with that estimate

- For comparisons:
  - Overlapping intervals suggest no significant difference
  - Non-overlapping intervals suggest significant differences

" I got the instructions from my Statistics Professor. He was 80% confident that the true location of the restaurant was in this neighborhood."

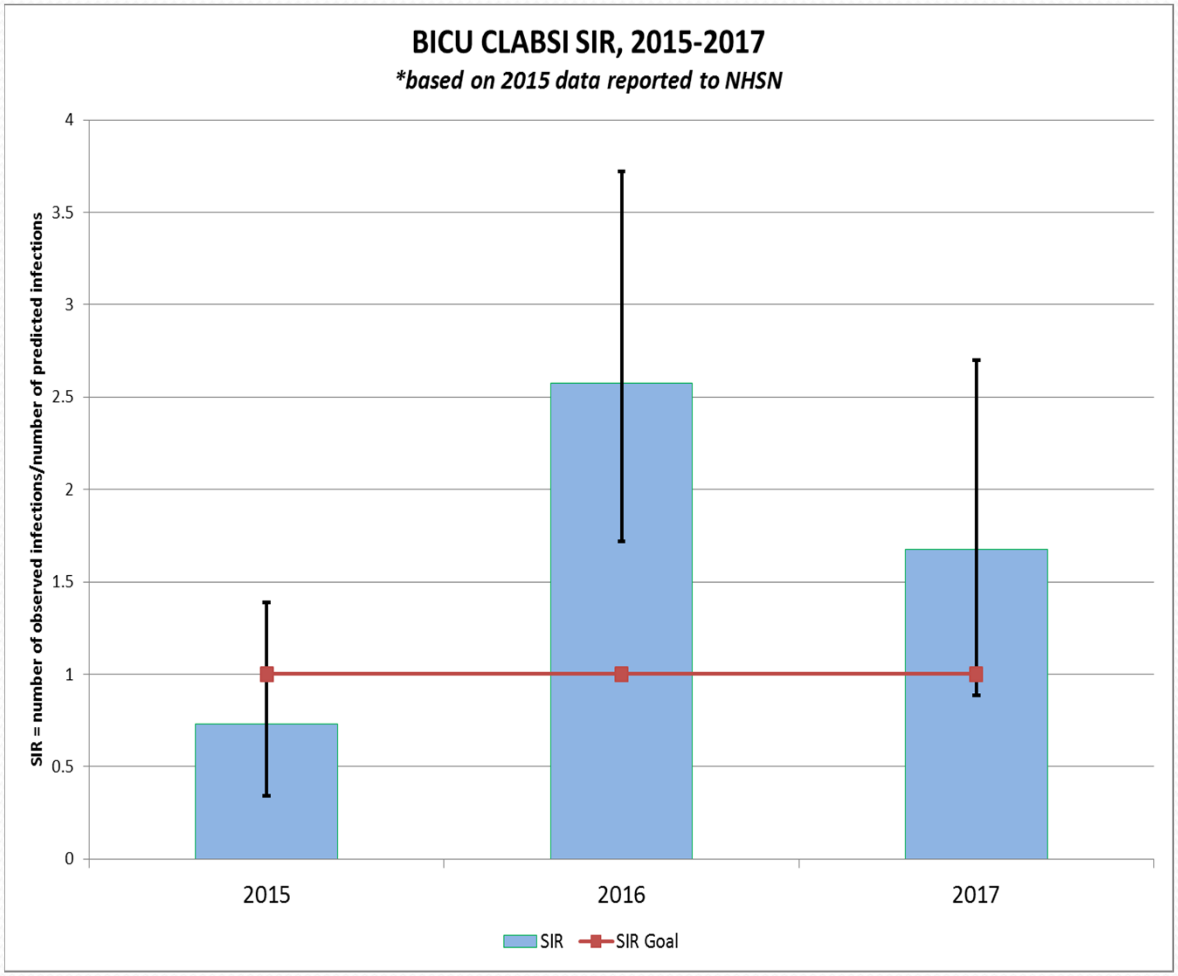
Question

- What year was the CLABSI SIR statistically significantly different from 1?

  A. 2015

  B. 2016

  C. 2017

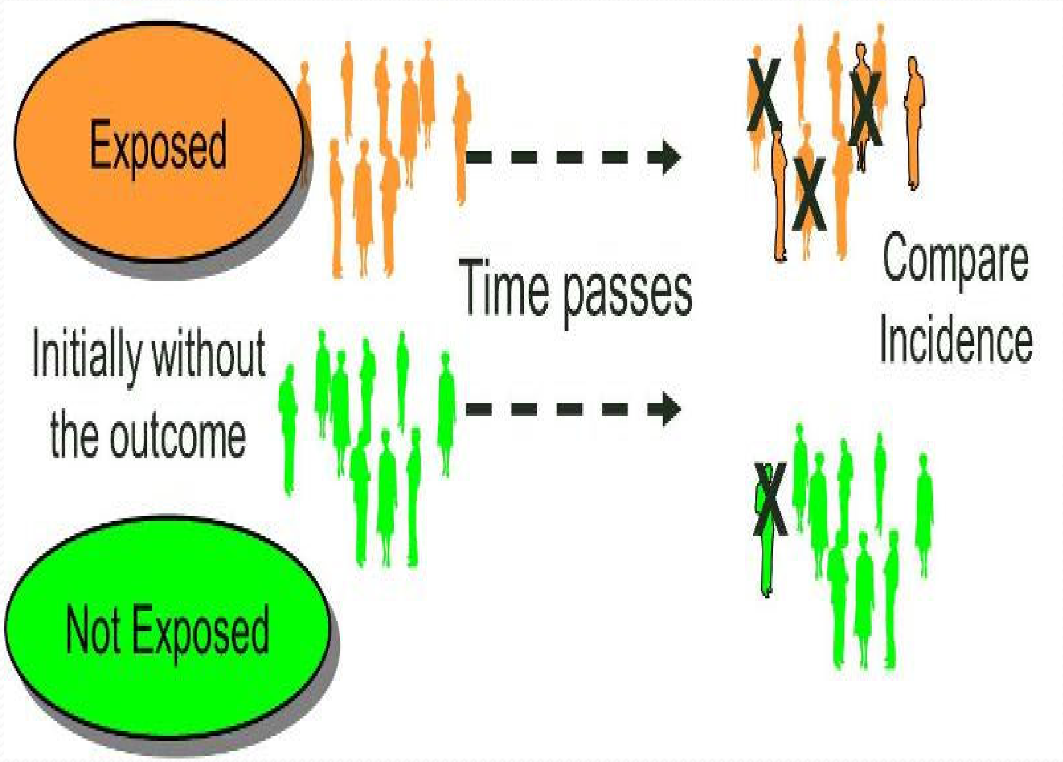


Common Study Designs

- Observational Studies
  - Descriptive –time, person, place
  - Analytic
    - **Cohort**
    - **Case control**
    - **Cross sectional – Prevalence**
- Experimental Studies
  - Natural
  - Planned -Clinical trials

Cohort Studies

1. Population free of disease
2. Follow for exposure to risk factors
3. Measure risk factor exposures over time
4. Look for correlations between
   a. presence and absence of <u>disease</u>
   b. presence and absence of <u>exposure</u>

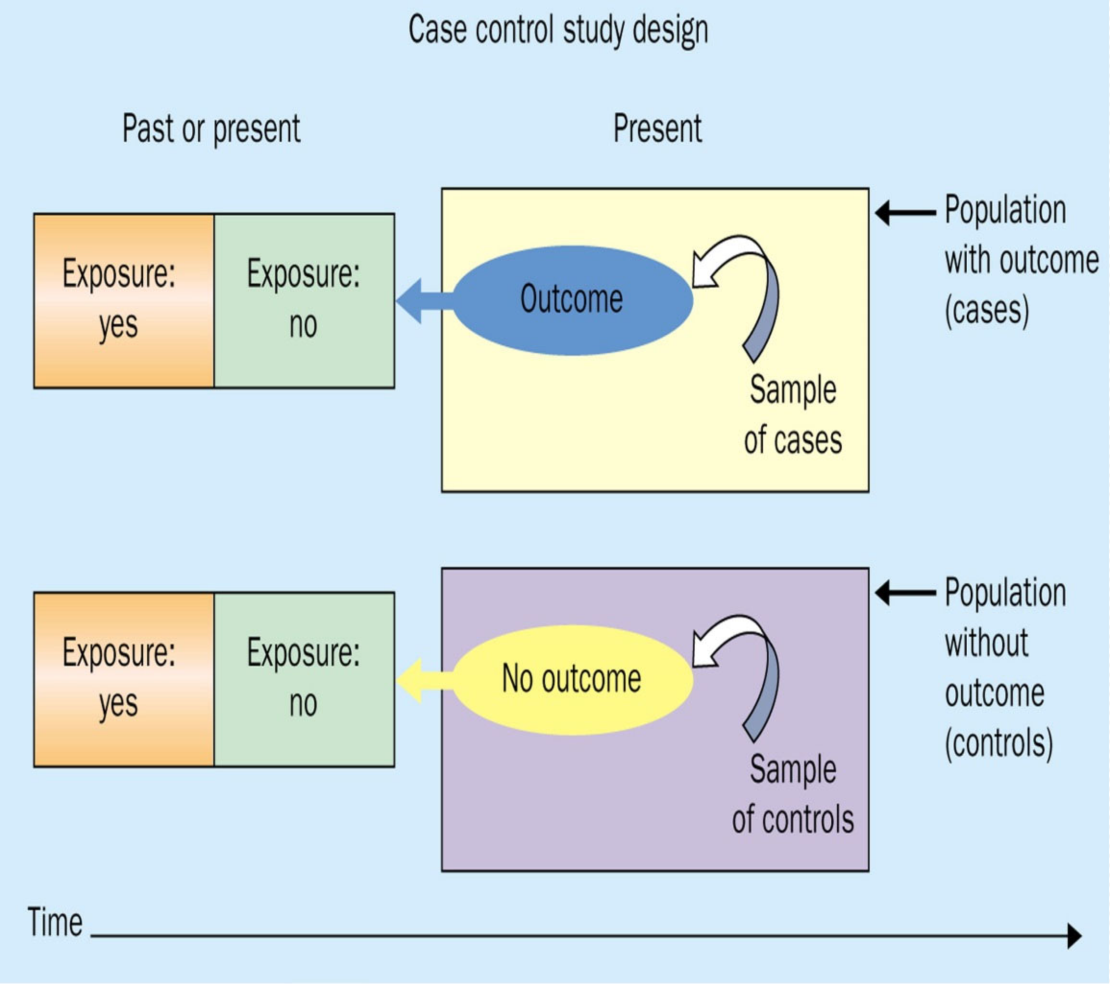


Cohort Studies

- Advantages
  - Clarify to temporal sequence
  - Facilitates study of rare exposures
  - Allow examination of multiple effects of single exposure
- Disadvantages
  - Large number of subjects
  - Time (think Framingham)
  - Expensive
  - Loss to follow-up

Case-Control Studies

- Retrospective
- Start with case of disease
- Match non-disease controls
- Look for differences in exposure levels

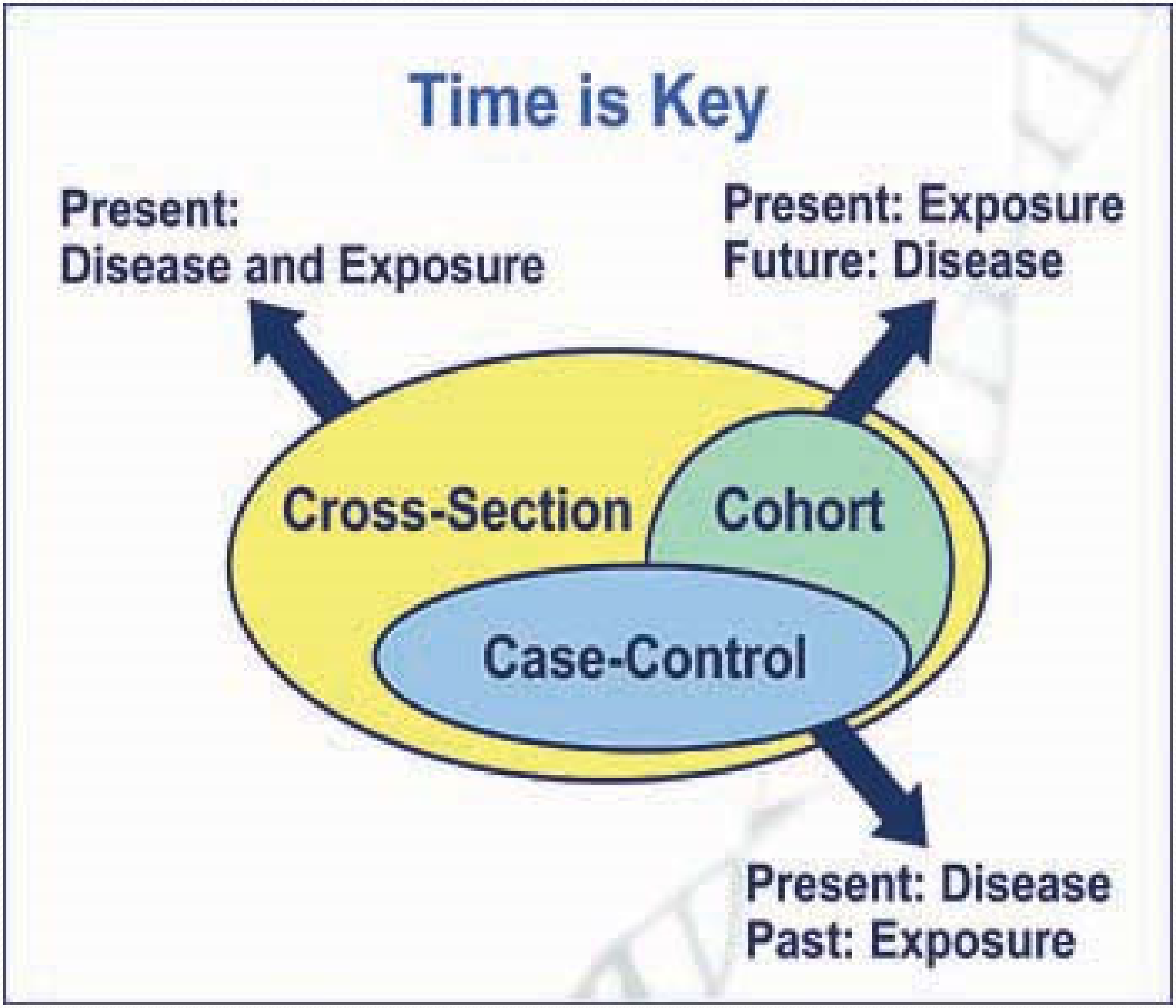


Case-Control Studies

- Advantages:
  - Less expensive
  - Quicker
  - Good for studying rare outcomes
- Disadvantages:
  - Limited **power**
  - Matches may be hard to find
  - Limited data available, especially as relates to exposure levels (recall bias)

Cross sectional – Prevalence

- Point Prevalence
- Period Prevalence


Time is Key
Present:
Disease and Exposure
Present: Exposure
Future: Disease
Cross-Section
Cohort
Case-Control
Present: Disease
Past: Exposure

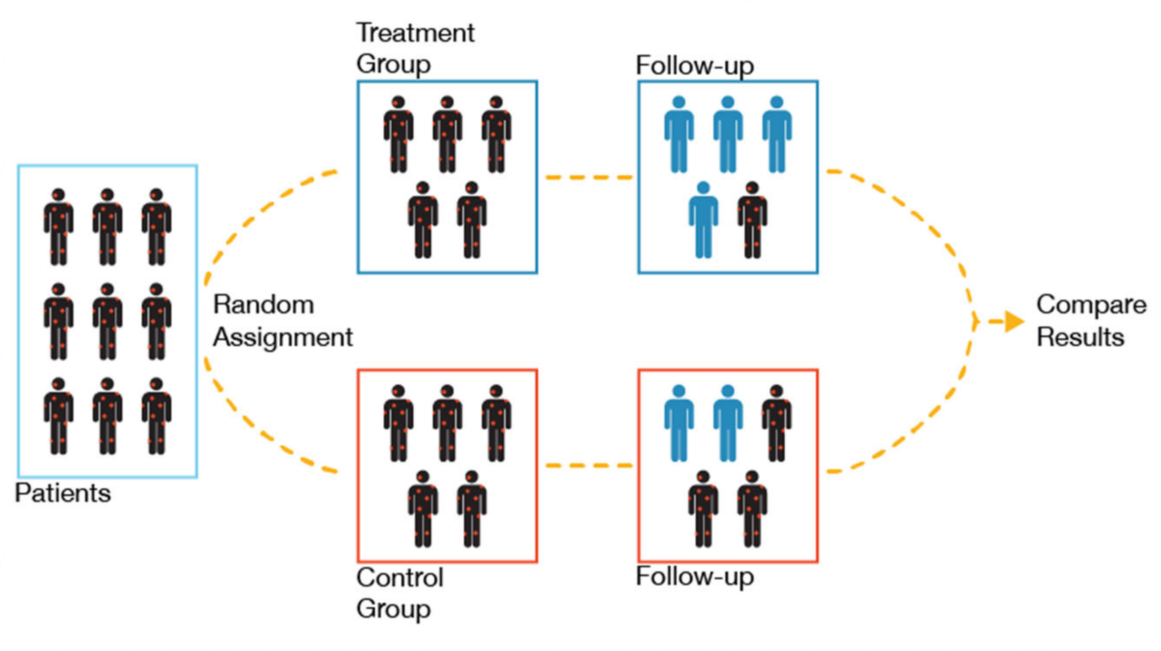
Question

- A hundred college freshmen were monitored for colds during the winter. 55 are smokers. 75% of the smokers had 2 or more colds. 20% of the non-smokers had 2 or more colds. What type of study was this?

   A. Case-control
   B. Cohort
   C. Cross-sectional
   D. Period prevalence

Experimental Studies

- Manipulate one or more factors
- Monitor outcomes of manipulated and non-manipulated
- True experiments – random
- Double blind – neither researcher or subject know which treatment group the subject is assigned



Types of Statistics

- **Descriptive**

  Techniques used to numerically describe the characteristics of a population or sample

- **Inferential**

  Techniques used to draw conclusions about a population based on a sample taken from the population

Two-by-Two Table

	Exposed	Not Exposed	Total
Disease	A	B	A + B
No Disease	C	D	C + D
Total	A + C	B + D	N

Measures of Association

- **Relative risk-** measures the strength of the association (Artificial, Indirect, or Causal)

Incidence rate of disease in exposed divided (÷)

by incidence of disease in unexposed

Measures of Association- Relative Risk

	Exposed	Not Exp.	Total
ill	4	1	5
well	10	10	20
	14	11	25

Incidence Rates:

4/14	1/11
0.29	0.09

RR = 0.29 / 0.09 = 3.2

Measures of Association

- **Odds Ratio-** probability of having a particular risk factor if a condition or disease is present, divided by the probability of having the risk factor if the disease or condition is not present.

  Probability of risk factor if disease present divided (÷) by probability of risk factor if disease not present

Measures of Association- Odds Ratio

	Smoke	No Smoke	Total
COPD	14 a	3 b	17
No COPD	12 c	18 d	30
	26	21	47

Odds Ratio:

14/3 **12/18**

4.66 **0.67**

OR = 4.66 / .67 = 7

OR= ad/bc

Causal Association

- **Strength**-disease rates higher with factor
- **Consistency**-reproducibility
- **Specificity**-association specific to one factor & one disease
- **Time Relationship**-exposure precedes onset of disease
- **Biological Gradient**-dose response: increased factor, increased disease

Causal Association

- **Plausibility**-should be biologically plausible
- **Coherence**-should be in accordance with other factors of disease, natural history
- **Experiment**-associations derived from experiments carry more weight
- **Analogy**-if similar association shown to be causal, assoc. more likely

Statistics suggest that an association exists

Types of Statistical Tests

- Parametric Tests
  - Population fits standard “bell” curve
  - Usually continuous, interval data
- Non-parametric Tests
  - Can be Nominal or Ordinal data
  - Population not required to fit “bell” curve

Parametric Tests

- Z-test
  - Used to test difference between the means
  - Sample size greater than 30
  - Population parameters known (S.D.)
- T-test
  - Used to test difference between means
  - Sample size is less than 30
  - Population parameters unknown

One Tailed vs. Two Tailed

- **One Tailed test** – concern is with difference in one direction from the mean (e.g., Do people with foleys have greater number of UTI's)?
- **Two Tailed test** –concern is with difference in any direction (e.g., cancer drug therapy)

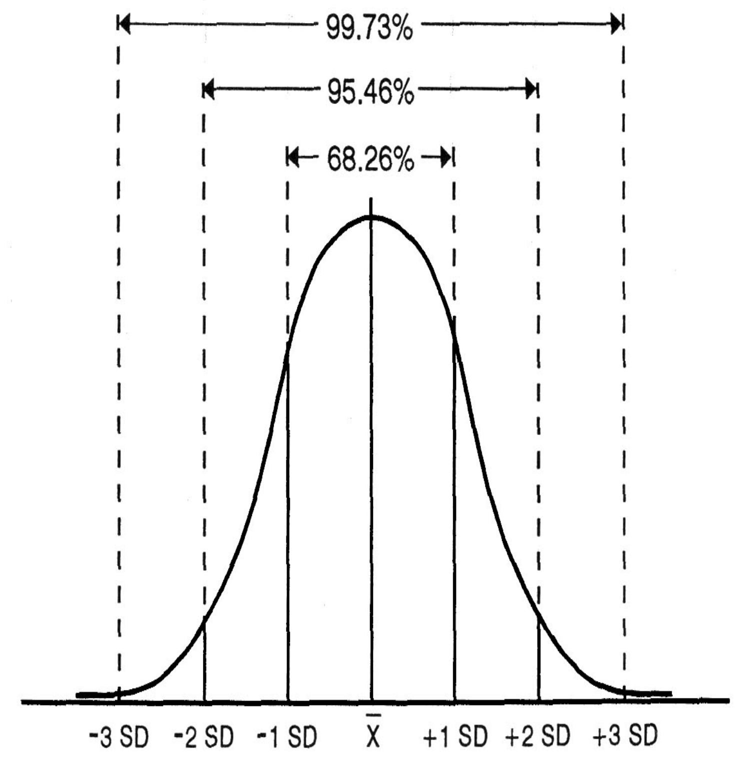


FIGURE 5.5. The normal distribution.

Non-Parametric Tests

- Used to determine if there are non-random associations between two categorical variables
- 2 X 2 contingency table
- Used to determine the P-value
- Does not require normal distribution

Chi-square Test

Fisher’s Exact Test

Chi-square Test

Start with 2 X 2 table with cells a, b, c, d

Chi-square= $$\frac{\mathbf{N[\,|ad\text{-}bc|\,\text{-}N/2]^2}}{\mathbf{(a+b)(c+d)(a+c)(b+d)}}$$

Alternatively,

$$\chi^2 = \Sigma\, (\mathrm{Oi} - \mathrm{Ei})^2 / \mathrm{Ei}$$

Take result to chi-square table to look up the P value: **If the resultant P-value is less than 0.05, then there is a statistically significant difference between the two classifications**

Fisher’s Exact Test

- Use to evaluate 2 X 2 table – variant of the chi-square
- Use if any value is below 30
- Fisher’s exact can be used when numbers in cells are imbalanced (i.e., 5 in one cell and 100 in another), can even have 0 in one cell
- Calculates the P-value directly

Question

- You have decided to compare your CLABSI rate to the published NHSN rate. What test will you use to compare?

A. 2 X 2 table

B. Chi-square

C. Fisher's exact

D. You need more information

Testing for Reliability

- Any test will give you one of 4 options as a result:
  1. True positive (those who test positive and DO have disease)
  2. True negative (those who test negative and do NOT have disease
  3. False positive (those who test positive and do NOT have disease)
  4. False negative (those who test negative and who DO have disease)

- Sensitivity and specificity are common statistical measures used to describe the properties of diagnostic tests

Sensitivity

If a person has a disease, how often will the test be positive (true positive rate)? (**accuracy of a positive result)**

Sensitivity Rate

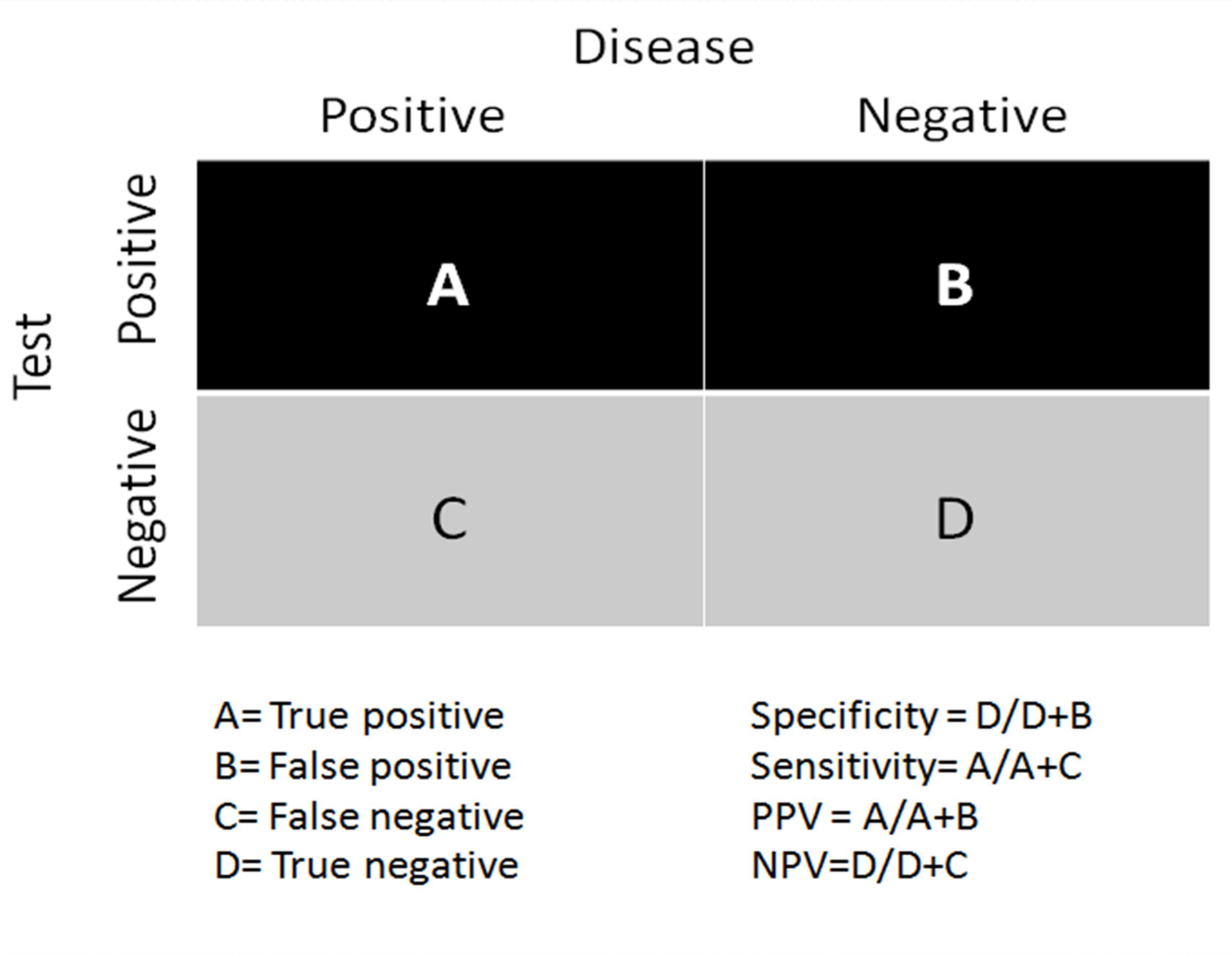
$$\frac{\text{\# of true positives}}{(\text{\# of true positives} + \text{\# false negatives})} \times 100$$

Specificity

If a person does not have the disease how often will the test be negative (true negative rate)? (**accuracy of negative result**)

Specificity Rate:

$$\frac{\text{\# of true negatives}}{(\text{\# true negatives + \# false positives})} \quad X \quad 100$$



		Disease	
		Positive	Negative
Test	Positive	A	B
	Negative	C	D

A= True positive
B= False positive
C= False negative
D= True negative

Specificity = D/D+B
Sensitivity= A/A+C
PPV = A/A+B
NPV=D/D+C

Note that when you are assessing predictive value, this is across the table (↔), sensitivity and specificity are assessed up and down the table (↕)

Question

Calculate the Sensitivity and Specificity for these data:

	Has Condition	
	YES	NO
Positive Test	40	30
Negative Test	10	70
Total	50	100

Sens = 40/50 = 80% **PPV = 40/70 =57%**

Spec = 70/100 = 70% **NPV = 70/80 =88%**

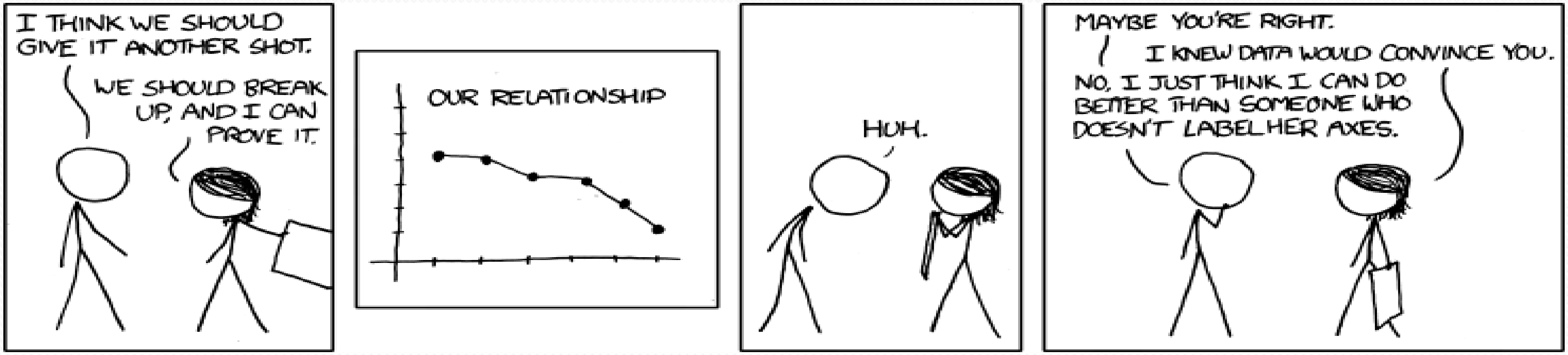
Graph Types

- Bar Charts – often used to display discrete data
  - *Comparison between categories*
- Pie Charts
  - *To show a percentage of a whole*
- Line Graphs – often used to display continuous data
  - *To show trends over time*
- Histogram
  - *Used to show a measurement of same variable over time – most often used in outbreak situations*

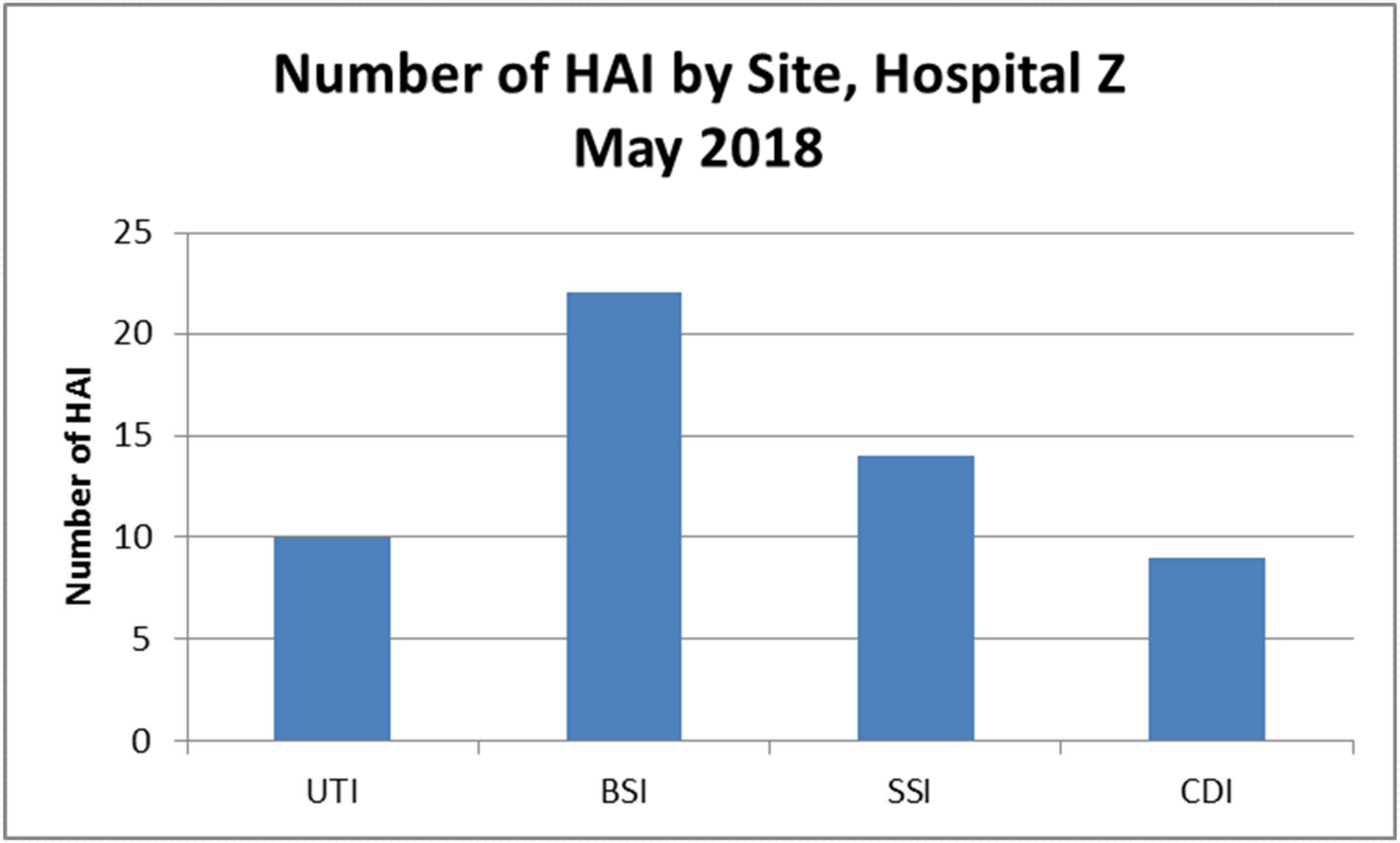
Features of Graphs and Tables

Graphs and tables should be self-explanatory!

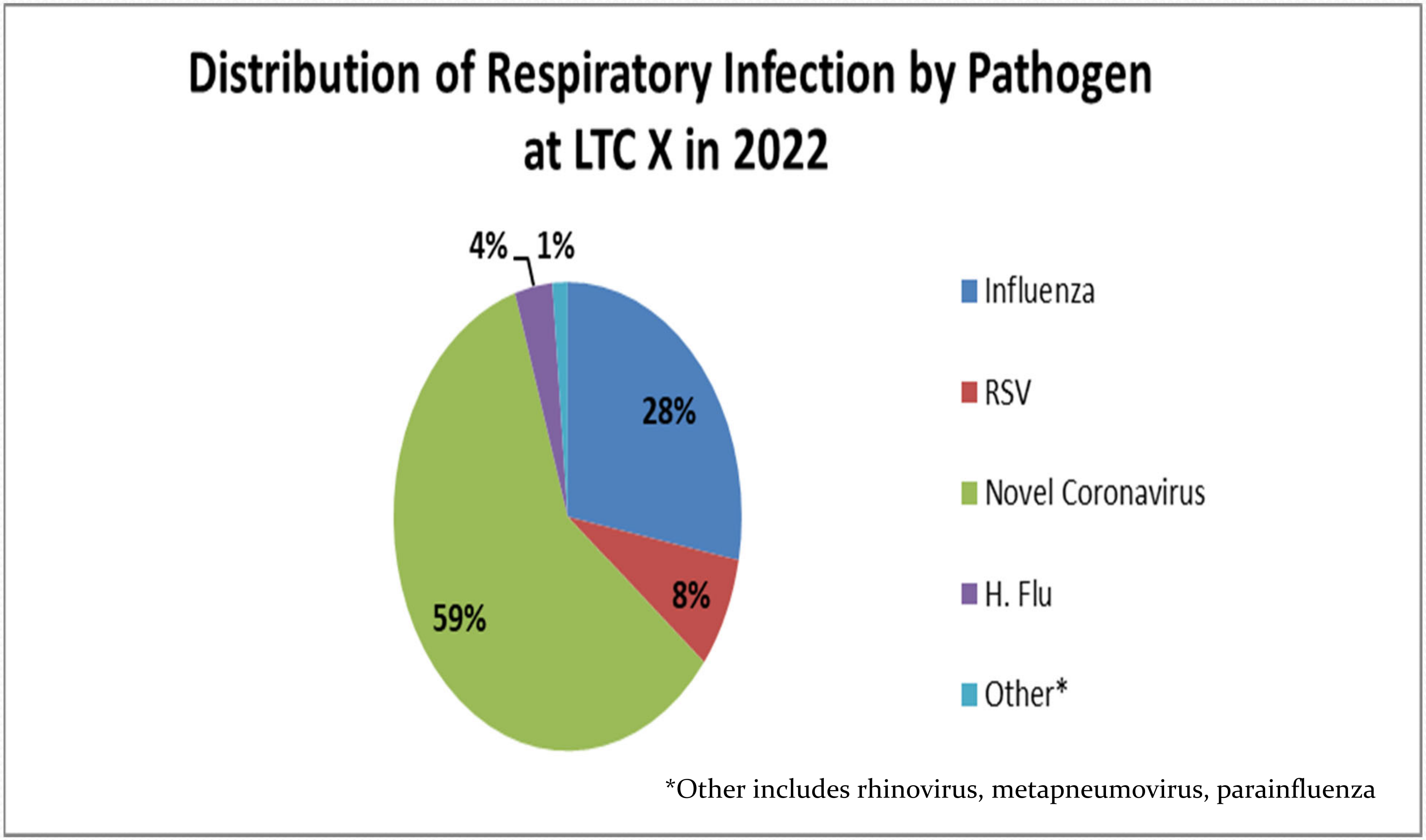
- Clear, concise title: describes person, place, time
- Informative labels: axes, rows, columns
- Appropriate intervals for axes
- Coded and labeled legends or keys
- Use footnotes to:
  - Explain codes, abbreviations, and symbols
  - Note exclusions
  - Note data source


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

Bar Chart


Number of HAI by Site, Hospital Z
May 2018
Number of HAI
25
20
15
10
5
0
UTI
BSI
SSI
CDI

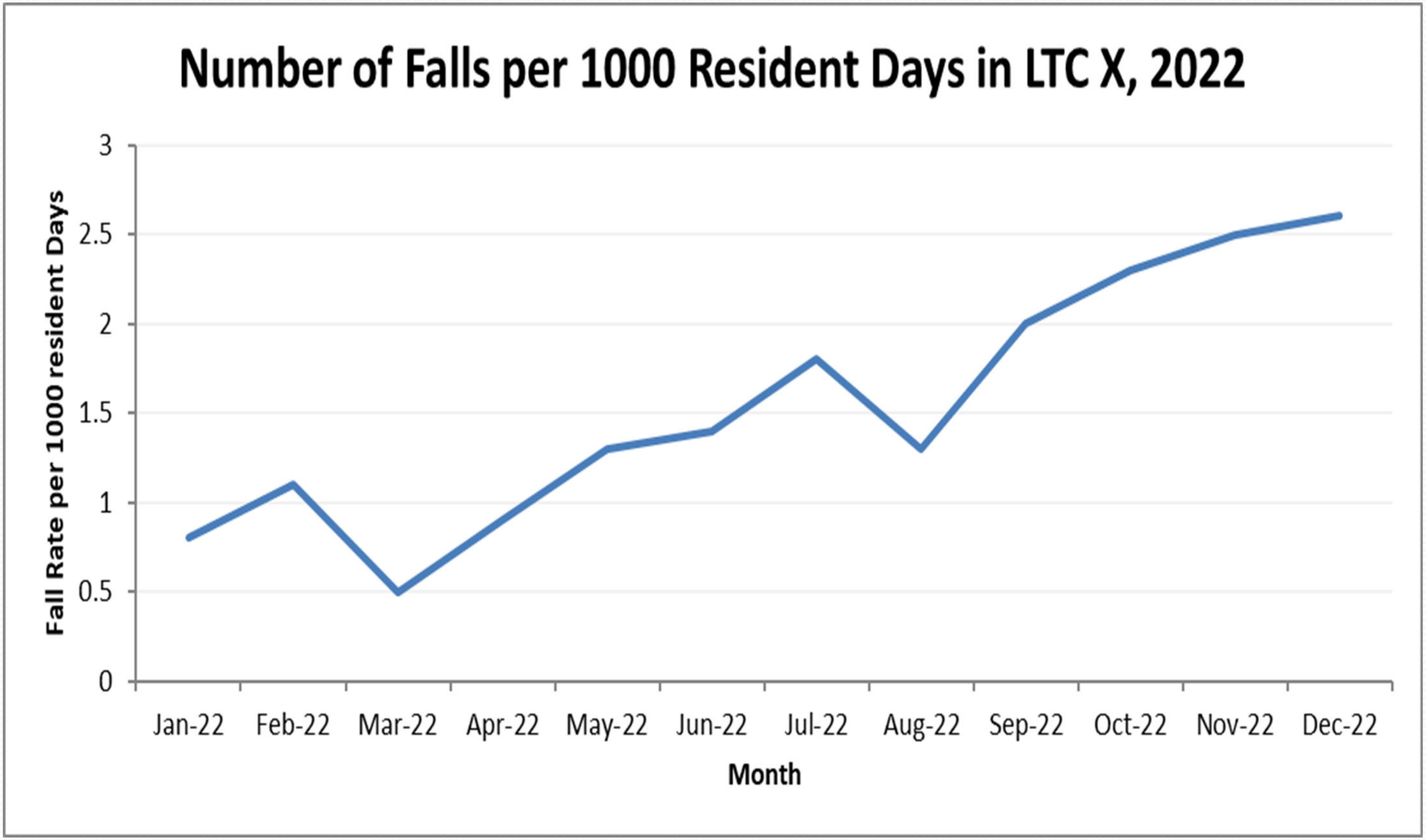
Pie Chart



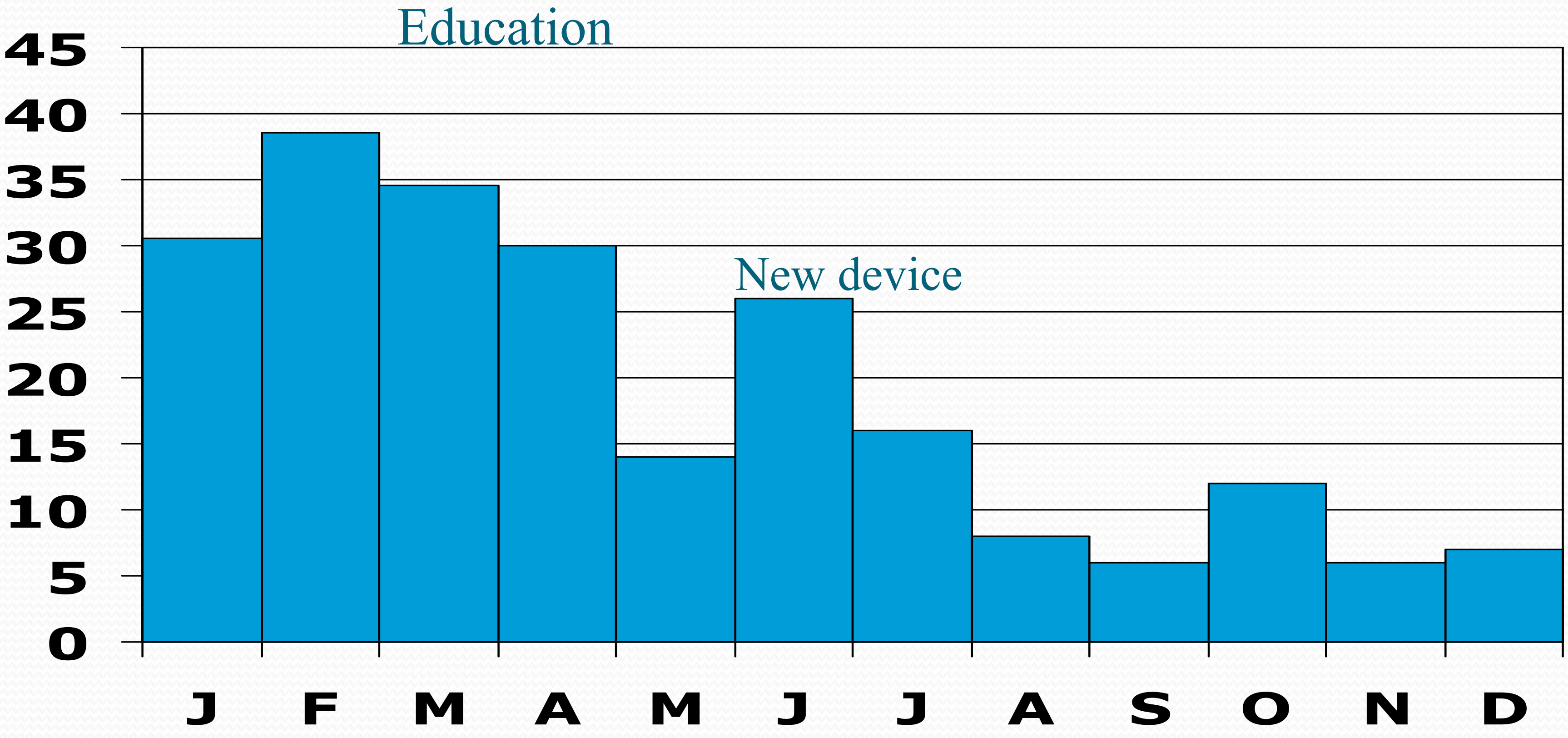
Distribution of Respiratory Infection by Pathogen at LTC X in 2022

*Other includes rhinovirus, metapneumovirus, parainfluenza

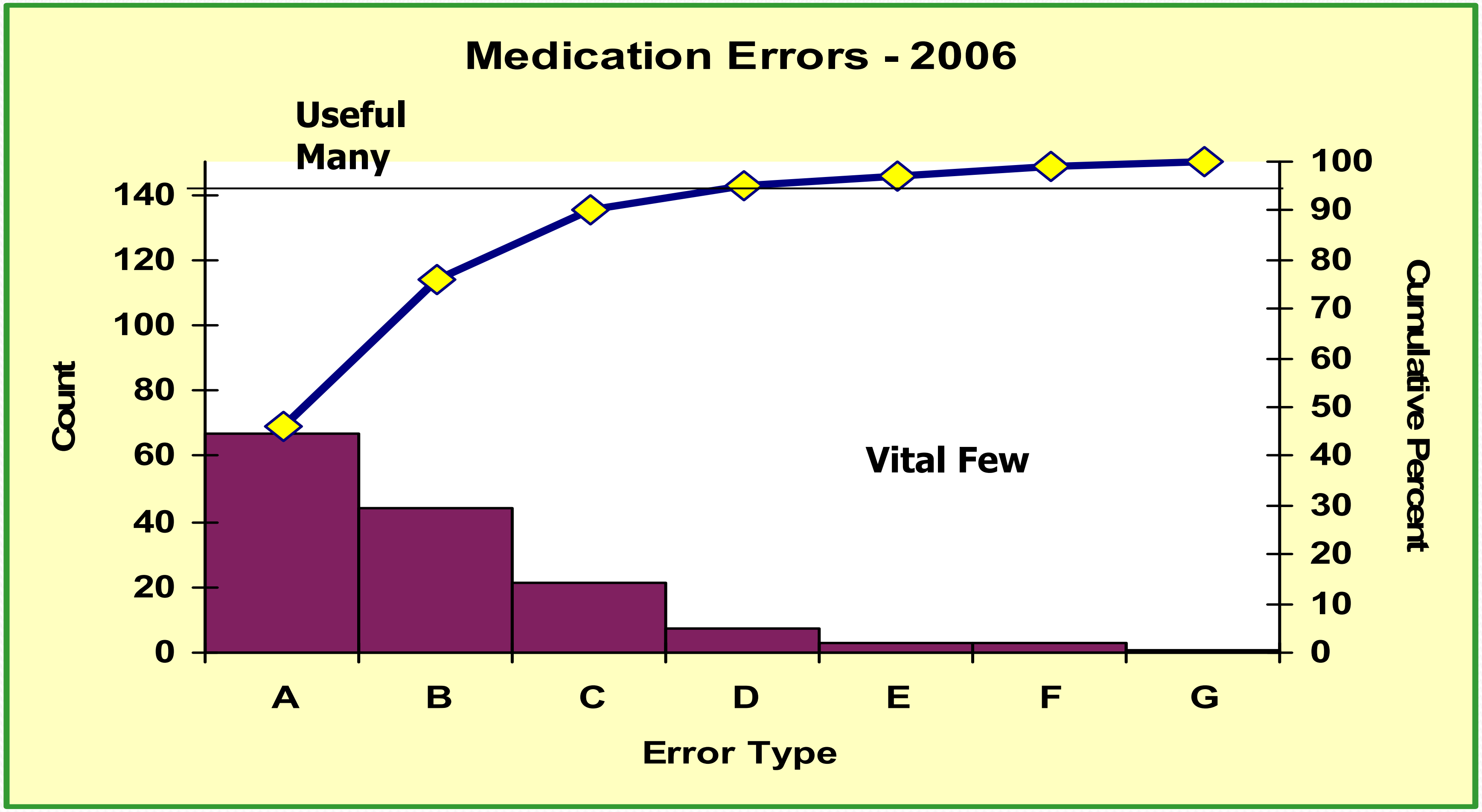
Line Graph


Number of Falls per 1000 Resident Days in LTC X, 2022
Fall Rate per 1000 resident Days
3
2.5
2
1.5
1
0.5
0
Jan-22
Feb-22
Mar-22
Apr-22
May-22
Jun-22
Jul-22
Aug-22
Sep-22
Oct-22
Nov-22
Dec-22
Month

Histogram



Pareto Chart



What 20% of the errors are causing 80% of the problems (80/20 rule)?

Question

- What type of chart/graph could you use to BEST display discrete causes of medication errors and the cumulative percentage of all errors?

  A. Bar chart
  B. Line graph
  C. Pareto chart
  D. Pie chart

Epidemic Curve

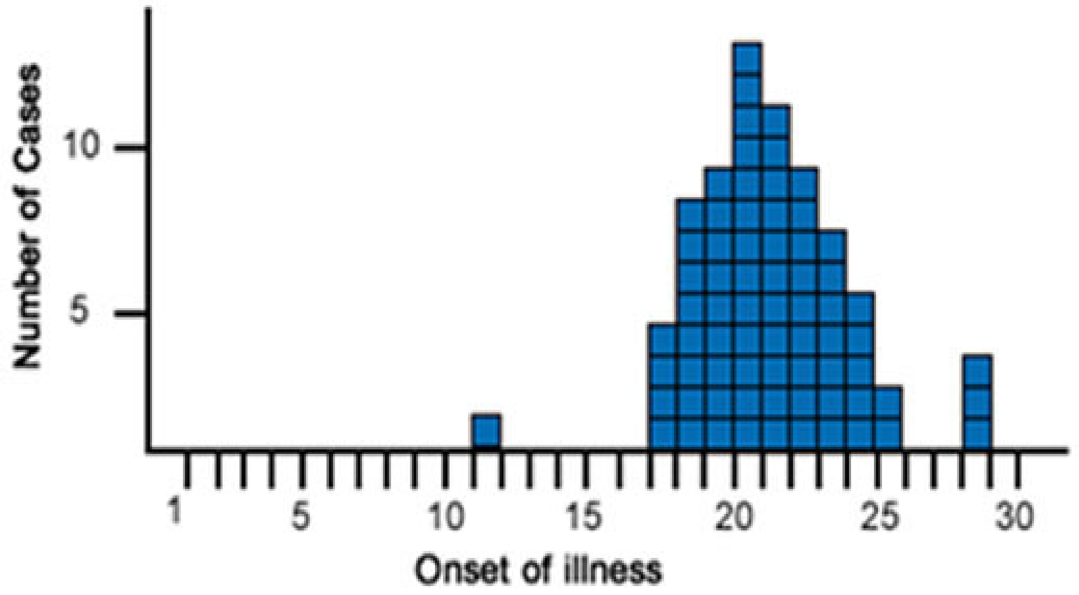
- Useful visualization of onset of illness among cases associated with an outbreak
  - Distribution of cases over time
  - Magnitude
  - Pattern of spread
  - Likely time of exposure
  - Outliers

Epidemic Curve

- Point Source Outbreak – persons exposed over brief time to same source (e.g. single meal or event) – number of cases rise rapidly and fall gradually
- Continuous Common Source – persons exposed to same source but exposure is prolonged over period of days, weeks or longer – curve rises gradually and may plateau
- Propagated Outbreak – no common source, spread person-to-person – curve has progressively taller peaks

Question

Based on the epidemic curve, what is the most *likely* source of this outbreak?



A. Widespread contamination of a food product
B. An item served during catered lunch
C. An ill healthcare worker with norovirus

Thank you!

Any Questions?