# Basic Statistics for Surveillance

Lisa Stancill, MPH
Business Intelligence Analyst
Infection Prevention
UNC Hospitals

1

## What are Statistics?

The margin of error...

17 in every 100 people...

Men are at 3 times higher risk...

Numbers that describe the health of the population

1 in 9 children...

39% OF THE POPULATION...

The science used to interpret these numbers.

There is a statistically significant difference...

Risk of dying is 8 times higher among...

2


3

## Learning Objectives

| | |
|---|---|
| Describe Surveillance Data | Define these terms: rates, prevalence, incidence, mean, median, mode, standard deviation |
| Display and Interpret Surveillance Data | Compare bar graphs, line graphs, pie charts and tables |
| Determine the Significance of Changes to Surveillance Data | Describe benchmarks (internal vs. external), create control charts, define p-values and 95% CI |

4

## Descriptive Statistics

- Measures of Rates and Ratios
  - *Rate: How fast disease occurs in a population.*
  - *Ratio: How much disease compared to standard.*
- Measures of Central Tendency
  - *Central Tendency: How well the data clusters around an average value.*
- Measures of Dispersion
  - *Dispersion: How widely your data is spread from the average.*

5

## Absolute Measures

- Simplest type of measurement
- Also known as counts or frequencies
- Example:
  - LTC A: 25 residents with novel coronavirus
  - LTC B: 10 residents with novel coronavirus
- *Is COVID19 worse at LTC A?*

6

## Relative Measures

- Includes a denominator
- Useful for comparisons
- Examples:
  - 16 cases of *C. difficile* out of 1000 residents
  - 1 positive *C. difficile* test out of 7 samples tested

7

## Absolute versus Relative

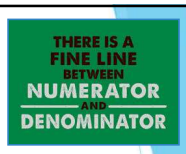
Example:

Novel coronavirus among LTC facility residents

- Absolute measures
  - LTC A: 25 residents ill
  - LTC B: 10 residents ill
- Relative measures
  - LTC A: 25 ill per 100 residents = 0.25 or 25%
  - LTC B: 10 ill per 25 residents = 0.40 or 40%

8

## What Makes a Rate?




1. Numerator (top number)
   - *e.g., number of infections*
2. Denominator (bottom number)
   - *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*

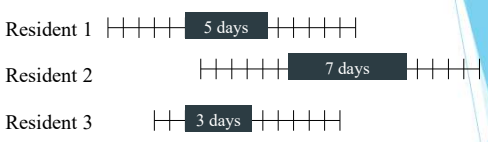
9

## 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)

10

## What is a Resident/Device-Day?



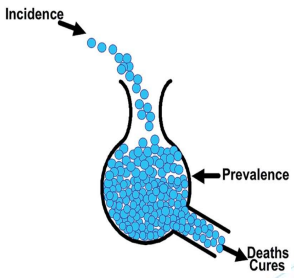

=15 resident-days, device-days, etc.

• More informative than simply saying "3 residents" since accounts for each resident's time of risk

11

## Rate Measures

- Prevalence
- Incidence
- Attack Rate




12

## 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*

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

13

## 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

14

## 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 expressed as a percent.
  - 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*

Count of new cases x **100** =
Number of people at risk

15

## Example 1:

- You perform surveillance for urinary tract infections (UTIs) in your 200 resident facility.
- During the 1st quarter of the year, you identify 3 new UTIs.
- During the 1st quarter, there were 180 residents in the facility with 12,000 resident-days.

16

## Example 1:

- In the first quarter, what was the UTI rate?
  - Incidence or prevalence?
  - Numerator?
  - Denominator?
  - Units?

17

## Example 1: Answers

- In the first quarter, what was the UTI rate?
  - Incidence or prevalence?
    - Incidence
  - Numerator?
    - 3
  - Denominator?
    - 180 residents or 12,000 resident days
  - Units?
    - "infections per 100 residents or infections per 1000 resident days"
  - ANSWER: 1.7 infections per 100 residents or 0.25 infections per 1000 resident days

18

## Example 1:

- You are concerned about the UTI rate so on April 7, you conduct a "spot check" on all of the residents of one area of the facility for a UTI.
- At that time with a census of 25, you review 20 charts and find 1 healthcare associated UTI.

19

## Example 1:

- On April 7th, what was the UTI infection rate at the time of your spot check?
  - Incidence or prevalence?
  - Numerator?
  - Denominator?
  - Units?

20

## Example 1: Answers

- In April, what was the UTI infection rate at the time of your spot check?
  - Incidence or prevalence?
    - Prevalence
  - Numerator?
    - 1
  - Denominator?
    - 20
  - Units?
    - "prevalent infections per 100 residents on April 7th"
  - ANSWER: 5 prevalent infections per 100 residents on April 7th.

21

## Example 1:

- You also routinely track counts of influenza-like illness in your 200 resident facility.
- During March, there is a cluster of influenza-like illness. In a short time period, 25 residents become ill and meet your case definition.
- During March, there were 180 residents in the facility with 5,000 resident-days.

22

## Example 1:

- What is the attack rate of influenza-like illness at your facility during March?
  - Numerator?
  - Denominator?
  - Units?

23

## Example 1: Answers

- What is the attack rate of influenza-like illness at your facility during March?
  - Numerator?
    - 25
  - Denominator?
    - 180
  - Units?
    - "percentage of residents who had influenza-like illness"
  - ANSWER: ***14%*** *of residents with influenza-like illness during outbreak in March*

24

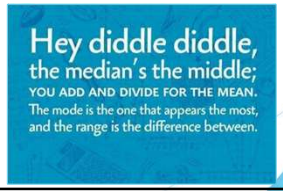
## Descriptive Statistics

- Measures of Rates and Ratios
  - *Rate: How fast disease occurs in a population*
  - *Ratio: How much disease compared to standard*
- Measures of Central Tendency
  - *Central Tendency: How well the data clusters around an average value*
- Measures of Dispersion
  - *Dispersion: How widely your data is spread from the average*

25

## Measures of Central Tendency

- Mean: average of a group of numbers
- Median: middle number in an ordered group of numbers
- Mode: most common value in a group of numbers




26

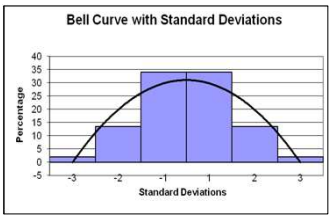
## Measures of Dispersion

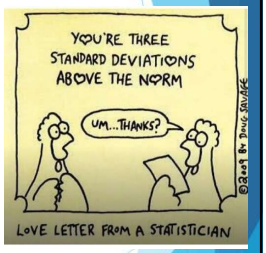
- Range: the largest value minus the smallest value
- Standard deviation: describes the variability or dispersion in the data set – tells you how spread out your data is

27

## Standard Deviation

- In a normally distributed data set,




68% of values $\pm$ 1 SD

95% of values $\pm$ 2 SD

99% of values $\pm$ 3 SD

28

## Example 2:

- Your administrator is becoming concerned that compliance with hand hygiene is not as high as it needs to be
- She has asked you to provide her with some data to confirm or disprove her suspicions

29

## Example 2:

- For the past year, once a month, you have been conducting hand hygiene audits in your facility – these are your monthly compliance results:
- 55%, 92%, 86%, 94%, 91%, 89%, 79%, 93%, 92%, 88%, 87%, 90%
- You decide as a first step to calculate the mean, median, mode and range of the monthly data to help describe hand hygiene compliance at your facility

30

## Example 2:

- What is the:
  - Mean?
  - Median?
  - Mode?
  - Range?

*HINT: 55%, 79%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 92%, 93%, 94%*

31

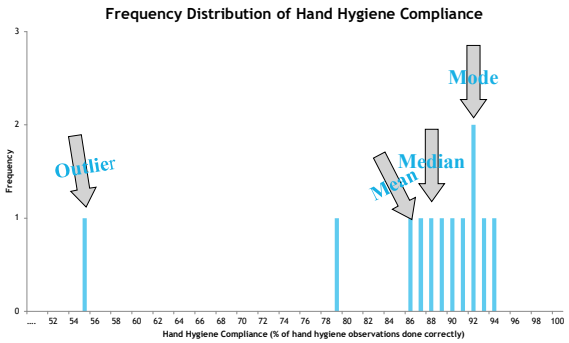
## Example 2: Answers

- What is the:
  - Mean?
    - 86.3%
  - Median?
    - 89.5%
  - Mode?
    - 92%
  - Range?
    - 39% *(94%[max]-55%[min])*
  - Standard Deviation?

    *can use programs like Excel to calculate*
    - 10.2%

32

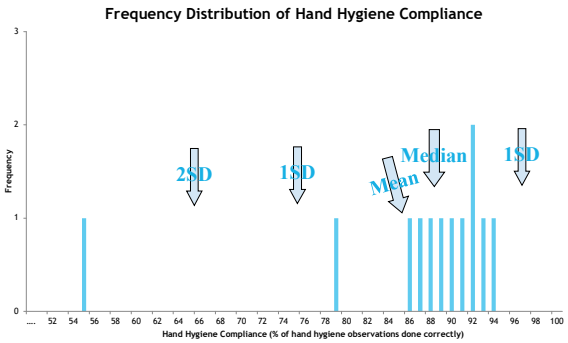
## Example 2: Central Tendency




*outlier: a value that falls outside the overall pattern.

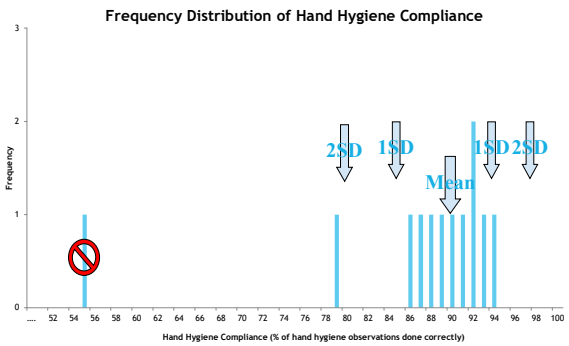
33

## Example 2: Dispersion



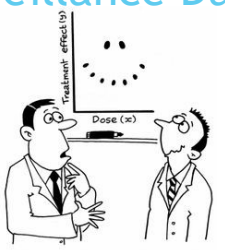

34

## Example 2: Dispersion




35

## Displaying Surveillance Data




"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

36

## Data Types

- Quantitative variables: numerical values
  - *(e.g., number of infections, number of residents)*
- Categorical variables: descriptive groups or categories
  - *(e.g., areas of the facility, gender, occupational groups)*
- *Data visualization is typically a graphical representation of these two types of data that allows you to see and understand trends, outliers and patterns in data*

37

## Displaying and Interpreting Surveillance Data

- Line lists
- Graphs: a visual representation of data on a coordinate system (e.g., two axes)
- Tables: a set of data arranged in rows and columns

38

## Line Lists

- Allow for record-level review of data
- Helpful way to standardize the data you want to routinely collect
- Helpful in pinpointing issues in data quality
- Can help inform rates or other summarized measures
- Can help identify trends

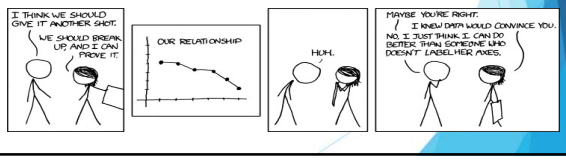
| Pt # | Name | Room # | Source | Organism | Culture date | Antibiotic | Date |
|---|---|---|---|---|---|---|---|
| 3695832 | | EW | Ucc U- Mod | Prot mir | 3-14 | | |
| | | EW 322 | Ucc U+M+ | Prot mir | | | |
| 0532210 | | EW 316 | cellulitis | | | cephalexin | 3-9 |
| | | EW 356 | Ucc – outside doc | | | cephalexin | 3-2 |
| | | EW 324 | ucc | | | cephalexin | 3-30 |
| | | EW 346 | sputum | | | amox | 3-10 |
| | | EW 308 | ucc | ecoli | | | |
| 7802490 | | JW 234 | Ucc U- Mod | Kleb pn, enta stau | 3-6 | | |
| | | JW 202 | wound | | | | |
| | | PW | eyes | | | tobra | 3-2 |
| 3887077 | | PW | Ucc U- M+ | ecoli | 3-2 | | |
| | | PW 122 | Cellulitis foot | | | clinda | 3-12 |
| 2475260 | | PW | Ucc U+Mod | Ecoli, ent | 3-12 | | |
| 4417105 | | PW | Ucc U-Mod | steno | 3-22 | | |
| 2259700 | | PW | wound | Prot mir | 3-5 | Sxi reported to FX | |
| 7899247 | | PW | Ucc U+M+ | ecoli | 3-30 | | |

39

## 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



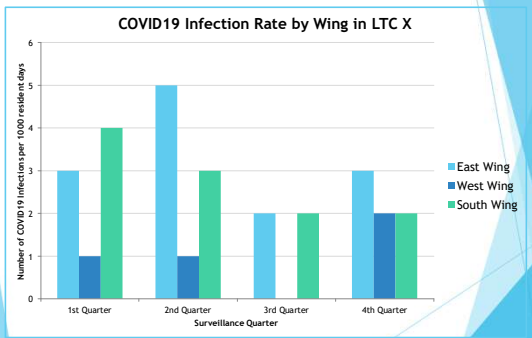

40

## Graph Types

- Bar Graphs
  - *E.g., Histograms (shown in previous example)*
  - *E.g., Comparison between categories*
  - *E.g., Epidemic Curves*
- Line Graphs
  - *E.g., To show trends over time*
- Pie Charts
  - *E.g., As a percentage of a whole*

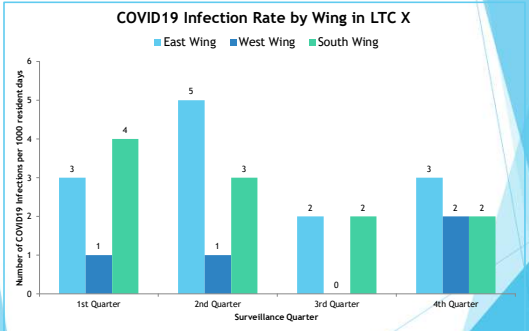
41

## Bar Graph



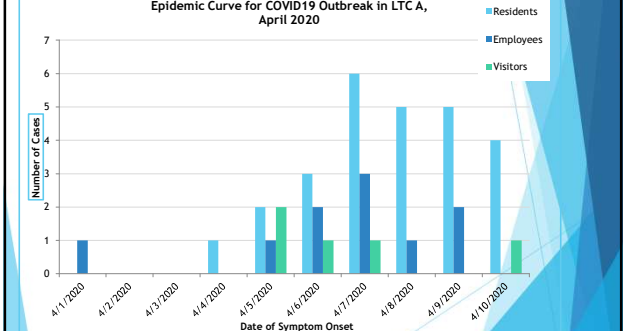

42

## Bar Graph



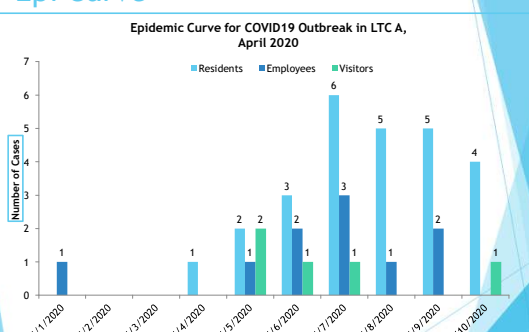

43

## Epi Curve



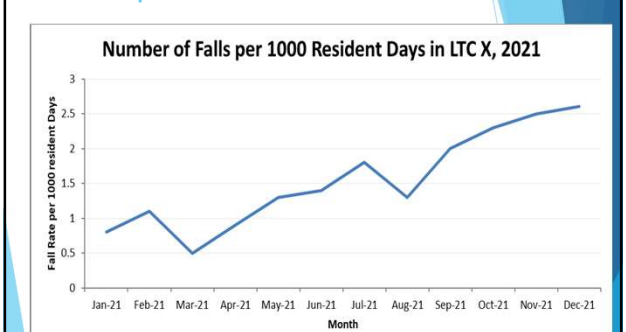

44

## Epi Curve



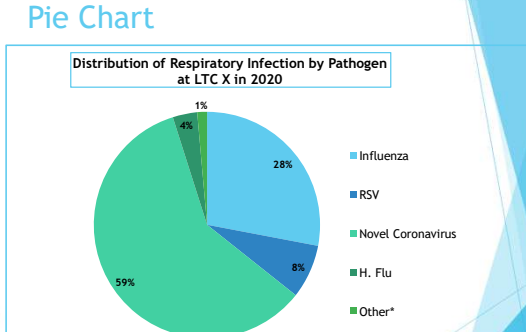

45

## Line Graph




46

## Pie Chart




47

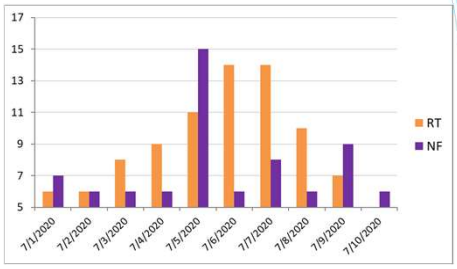
## Tables

**Number of UTIs by Age Group, LTC X, 2021**

| Age Group (Years) | Number of Cases |
|---|---|
| <50 | 0 |
| 51-60 | 2 |
| 61-70 | 7 |
| 71-80 | 6 |
| 81-90 | 3 |
| >90 | 1 |
| **Total** | **19** |

48

## *What's wrong with this graph?*




- *Missing a chart title*
- *Missing axis labels on the x and y axis*
- *The y axis starts at 5 (should start at 0)*
- *No explanation of abbreviations used in the legend*

49

## Interpreting Surveillance Data


50

## Why Analyze your Data?

- Provide feedback to internal stakeholders
- Analyzing data can help you identify areas that need improvement
- Reports can help inform prioritization and success of prevention activities
- Ultimately, these are YOUR data – you should know your data better than anyone else

51

## Checklist


- Before you begin analyzing your data, ask yourself these questions:
  - What data are you analyzing?
  - What is the time period of interest?
  - Why are you analyzing these data?
  - Who is the audience/stakeholders (and what do **they** want to see)?
    - Other IPs
    - Managers
    - Physicians
    - Administrative


52

## Data Analysis: Interpreting the Results

**Examine trends over time**

- Assess patterns to determine temporality
- Identify acute or unusual events which require immediate follow-up

**Assess which risk groups are being most affected – allows you to target your prevention efforts**

53

## Determine the Significance of your data - How?

**Practical/Clinical 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) - **is there special cause variation?***

**Other statistical tests and measures**

P-values

95% confidence intervals

54

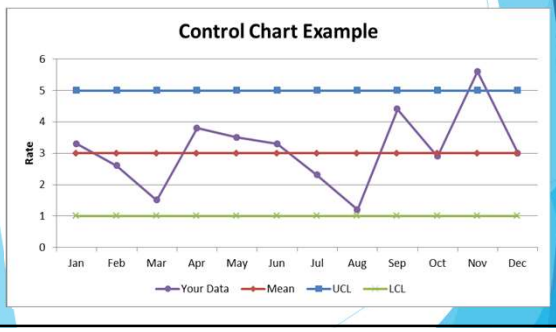
## Internal Benchmarks

- Compare current results to your own prior results
- Best way to chart your own progress over time
  - Select feasible and stretch goals
- Note when interventions took place
- Use when there is no external benchmark

55

## Control Charts

- Tool to help determine when infection rates are out of range -user sets control limits. *How high is TOO high?*




56

## Control Chart Example 3:

| MONTH | 2020 UTI Rate | Moving Range |
|---|---|---|
| JAN | 4.5 | --- |
| FEB | 3.2 | 1.3 |
| MAR | 3.6 | 0.4 |
| APR | 3.5 | |
| MAY | 3.0 | |
| JUNE | 4.0 | |
| JULY | 4.1 | |
| AUG | 4.6 | |
| SEPT | 4.8 | |
| OCT | 5.2 | |
| NOV | 5.7 | |
| DEC | 6.5 | |

1. Find the mean of the UTI rates for the last year
2. Calculate the moving ranges (subtract month 1 from 2, month 2 from 3…) and take absolute values (no negative values)
3. Calculate the mean of the moving ranges

57

## Control Chart Example 3 Answers:

| MONTH | 2020 UTI Rate | Moving Range |
|---|---|---|
| JAN | 4.5 | --- |
| FEB | 3.2 | 1.3 |
| MAR | 3.6 | 0.4 |
| APR | 3.5 | 0.1 |
| MAY | 3.0 | 0.5 |
| JUNE | 4.0 | 1.0 |
| JULY | 4.1 | 0.1 |
| AUG | 4.6 | 0.5 |
| SEPT | 4.8 | 0.2 |
| OCT | 5.2 | 0.4 |
| NOV | 5.7 | 0.5 |
| DEC | 6.5 | 0.8 |

- Find the mean of the UTI rates.
  =4.4
- Calculate the moving ranges
  See table
- Calculate the mean of the moving ranges.
  =0.5

58

## Control Chart Example 3:

- Calculate Upper Control limit= Mean + (**2.66** x Mean of Moving Range)
- Calculate Lower Control limit= Mean - (**2.66** x Mean of Moving Range)

In this example:

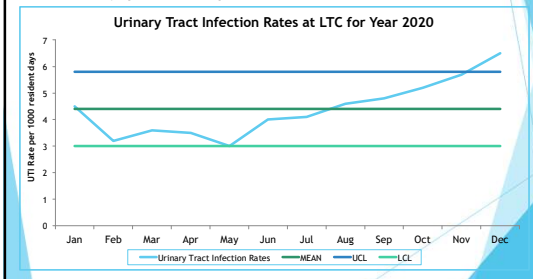
UCL = 4.4 + (2.66 x 0.5) = 5.8

LCL = 4.4 - (2.66 x 0.5) = 3.0

59

## Control Chart Example 3:

- Draw horizontal lines at the mean, UCL and LCL based on your historical data
- Then graph your current data and use the limits to identify potential problems.




60

## Interpretation of Other Statistical Tests *(more advanced topic)*

- Consider your calculated infection rate to be an estimation of the true rate.

  Why an estimation?
  - You may only do surveillance on a sample of residents in your facility.
  - If surveillance activities were repeated by other ICPs, your numerators may vary slightly based on interpretation of case definitions, available clinical information in the chart, etc.

61

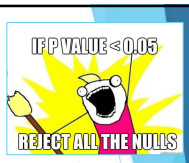
## Hypotheses

- 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.

62

## Other Statistical Tests: 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 is truly significant (and not due to chance)
- Generally a level of P<0.05 is considered "statistically significant"

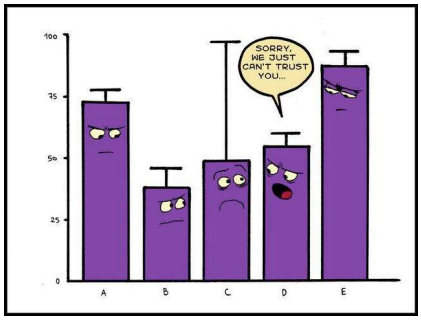
63


64

## Other Statistical Tests: 95% Confidence Interval

- Means that you are 95% confident that the *true* average value lies within this interval
- 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

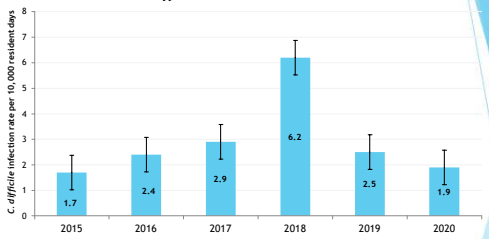
65




66

## 95% Confidence Interval Example:

**C. difficile Rates in LTC X 2015-2020**




*Is the C. difficile infection rate at this LTC in 2018 statistically significantly different than the C. difficile infection rate in other years?*

*YES - the 95% CI do not overlap*

*Is the C. difficile infection rate at this LTC in 2020 statistically significantly different than the rate in 2019?*

*NO- the 95% CI overlap*

67

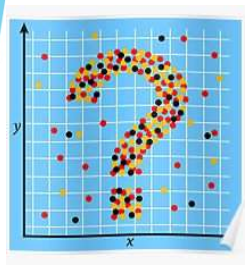
## Learning Objectives

| | |
|---|---|
| Describe Surveillance Data | Define these terms: rates, prevalence, incidence, mean, median, mode, standard deviation |
| Display and Interpret Surveillance Data | Compare bar graphs, line graphs, pie charts and tables |
| Determine the Significance of Changes to Surveillance Data | Describe benchmarks (internal vs. external), create control charts, define p-values and 95% CI |

68

*"The world cannot be understood without numbers. But the world cannot be understood with numbers alone."*

*-Hans Rosling*




*Thank you!*

69

## *Online Excel Resources*

www.excel-easy.com

https://excelexposure.com/

https://www.thoughtco.com/excel-formulas-step-by-step-tutorial-3123636

https://www.gcflearnfree.org/excel2016/sorting-data/1/

70