

Handout Disclaimer

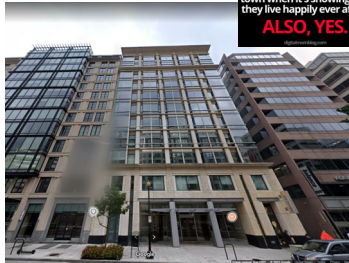
Disclaimer: the following document was distributed as a handout for the Analyzing Data for Beginner Series and is designed to enhance the sessions you have attended. NSH makes no representations to the factual correctness of any information contained herein. All of the content comprising this handout is the exclusive property of the presenter and the national society for histotechnology. It may not be copied, reproduced, distributed, displayed or transmitted without the consent of the presenter or the National Society for Histotechnology.



Analyzing Data for Beginners: Day 1

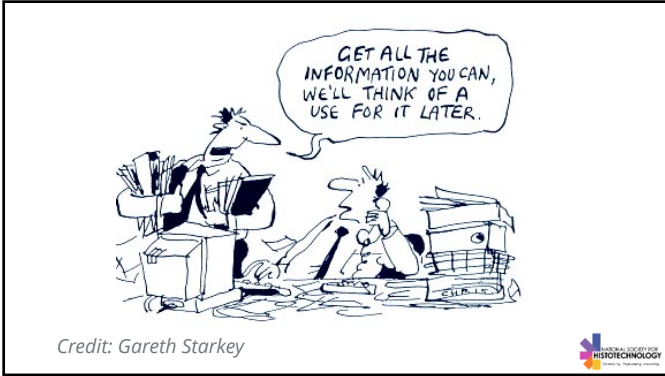
Connie Wildeman, MPA





Does every Hallmark Christmas movie have the same plot?
YES.
Am I still going to watch two people fall in love in a small town when it's snowing and they live happily ever after?
ALSO, YES.



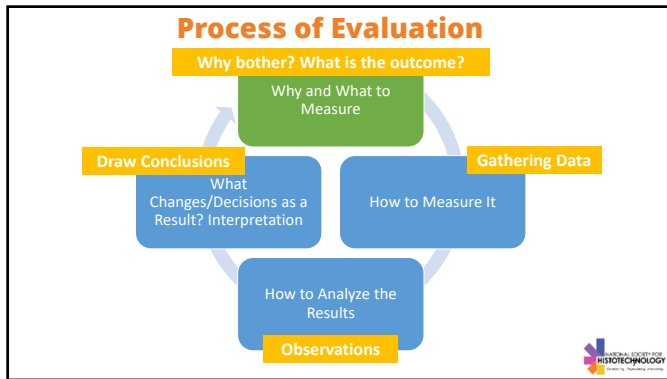


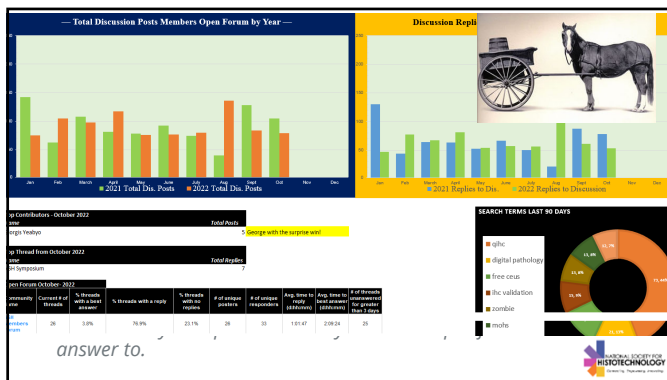
What we will cover today:

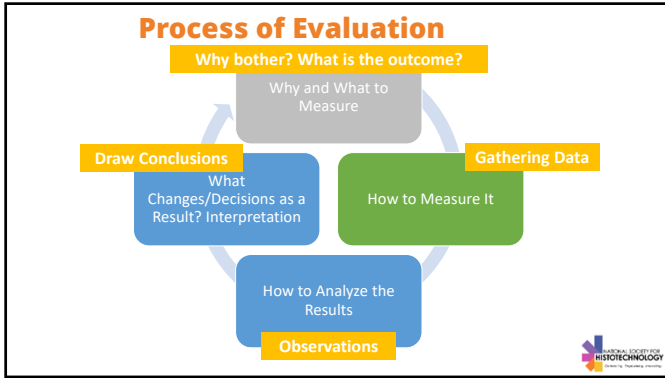
- Part 1: What is Data?
- Part 2: Preparing Data for Analysis
- Part 3: Analyzing Data

Part I: What is Data?

1. Summarize where data fits into the evaluation cycle.
2. Identify what data is present and available to you.
3. Define data
4. Classify types of data








Part I: What is Data?

1. Summarize where data fits into the evaluation cycle.
2. **Identify what data is present and available to you.**
3. Define data
4. Classify types of data

What Data Do You Already Have Access To?

How to Measure/Count

- LIS canned reports
- Emails
- Error logs
- Budgets (current and past)
- Page views
- Search terms
- Invoices
- Sign out logs
- Staff meeting agendas



**Okay, so data exists all over the place.
But how do I know what I am looking
at?**



Part I: What is Data?

1. Summarize where data fits into the evaluation cycle.
2. Identify what data is present and available to you.
3. **Define data**
4. Classify types of data



What is Data?

- **Factual information** (such as measurements or statistics) used as a basis for **reasoning**, **discussion**, or **calculation** (taken from Merriam-Webster)



Types of Data

Quantitative: its numerical in nature and can include things like, test scores, temperatures, click rates.

Qualitative: its descriptive in nature, like color, types of college degrees, and frequency can be calculated. Also known as categorical.



Identify if the following are Qualitative or Quantitative

1. The baby weighs 20 pounds
2. The workshop attendees rated the event highly effective
3. The sky is blue
4. There were 200 IHC requests this month
5. The lab is cold today
6. Joe is 6 foot 2



Data is...

- Factual information made up of qualitative and/or quantitative variables that is an important piece in the evaluation and reasoning process.



Data is not a decision



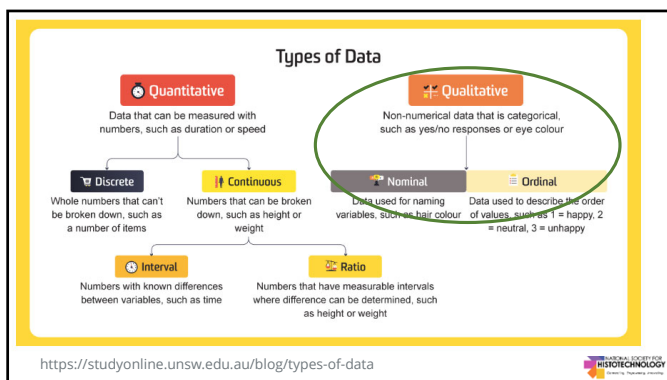
Part I: What is Data?

1. Summarize where data fits into the evaluation cycle.
2. Identify what data is present and available to you.
3. Define data
4. **Classify types of data**

Why does type of data matter?
It determines the kinds of statistical tests and measures that can be used.

How to Measure It



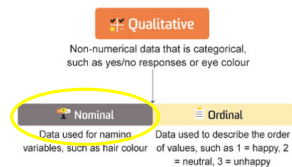


Nominal Data- Categorical

Nominal data are categorized according to labels which are purely descriptive—they don't provide any quantitative or numeric value. **Nominal data cannot be placed into any kind of meaningful order or hierarchy.**

Examples:

- Eye color
- Type of stain
- Gender
- Car color

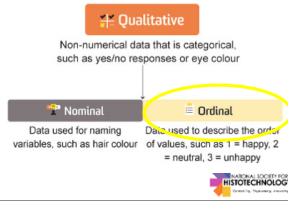


Ordinal Data- Categorical

Ordinal data are categorical (non-numeric) but may use numbers as labels - BUT CAN BE RANKED NATURALLY.

Examples:

- Grades/Marks (A,B,C,O,ME,etc)
- Education level
- Income level (low, middle, upper middle class, etc.)
- Satisfaction levels

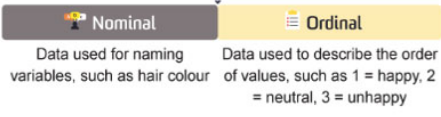


Priority	Block	Container	Task	Lab Responsible Pathologist
Embedded Priority	PS20	A1	PS20 A1-1	H&E
Embedded Priority	PS20	A1	PS20 A1-2	H&E
Embedded Priority	PS20	A1	PS20 A1-3	H&E
Embedded Priority	PS20	A1	PS20 A1-4	ERG
Embedded Priority	PS20	A1	PS20 A1-7	H&E
Embedded Priority	PS20	A1	PS20 A1-1	H&E FS
Embedded Priority	PS20	A1	PS20 A1-4	H&E
Embedded Priority	PS20	A1	PS20 A1-9	H&E
Embedded Priority	SS20	A1	SS20 A1-1	H&E
Embedded Priority	SS20	A1	SS20 A1-13	H&E
Embedded Priority	SS20	A1	SS20 A1-9	H&E
Embedded Priority	SS20	A1	SS20 A1-1	ERG
Embedded Priority	SS20	A2	SS20 A2-1	ERG
Embedded Priority	SS20	A3	SS20 A3-1	H&E
Embedded Priority	SS20	A4	SS20 A4-1	H&E
Embedded Priority	SS20	A5	SS20 A5-1	H&E
Embedded Priority	SS20	A6	SS20 A6-1	ERG
Embedded Routine Surgical	SS20	A1	SS20 A1-2	H&E FS
Embedded Routine Surgical	SS20	A10	SS20 A10-1	HER2 IHC
Embedded Routine Surgical	SS20	A11	SS20 A11-1	H&E
Embedded Routine Surgical	SS20	A12	SS20 A12-1	H&E
Embedded Routine Surgical	SS20	A2	SS20 A2-1	H&E
Embedded Routine Surgical	SS20	A3	SS20 A3-1	HER2 IHC
Embedded Routine Surgical	SS20	A4	SS20 A4-1	H&E
Embedded Routine Surgical	SS20	A5	SS20 A5-1	H&E



Qualitative

Non-numerical data that is categorical, such as yes/no responses or eye colour



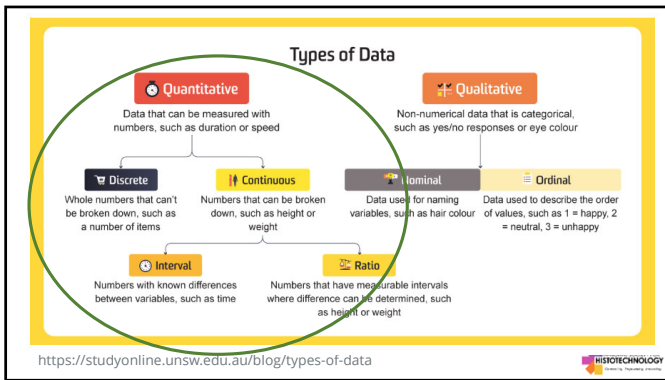
Is the Question Nominal or Ordinal?



1. Are you left handed or right handed?
2. How satisfied are you with your pizza delivery service? 1 – not satisfied, 2 – satisfied, 3 – very satisfied
3. What kind of house do you live in?
4. What kind of pet do you have?
5. Are you willing to work extra shifts? I am willing to work any extra shifts, I am willing to work extra shifts if given notice, I am willing to work extra shifts occasionally, I am not willing to work extra shifts.

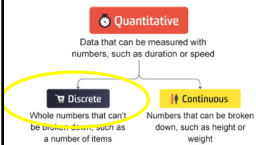
Remember, nominal has no hierarchy...





Discrete Data

Discrete data is numeric. They do not have to be whole numbers. **You count discrete data.**



Examples:

- Number of employees
- Number of IHC tests run
- Favorite ice cream flavor among a group (counted)
- Number of responses to a survey



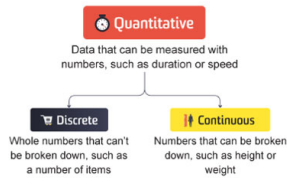
Discrete Data Example

Priority	Block	Container	Task	Lab Responsible Pathologist	Case Flags
Embedded Priority	PS20	A1	PS20 A1-1	H&E	RENAL
Embedded Priority	PS20	A1	PS20 A1-4	H&E	RENAL
Embedded Priority	PS20	A1	PS20 A1-7	H&E	RENAL
Embedded Priority	PS20	A1	PS20 A1-1	H&E	
Embedded Priority	PS20	A1	PS20 A1-4	ERG	
Embedded Priority	PS20	A1	PS20 A1-7	H&E	
Embedded Priority	PS20	A1	PS20 A1-1	H&E FS	
Embedded Priority	PS20	A1	PS20 A1-4	H&E	
Embedded Priority	PS20	A1	PS20 A1-7	H&E	
Embedded Priority	SS20	A1	SS20 A1-11	H&E	
Embedded Priority	SS20	A1	SS20 A1-13	H&E	
Embedded Priority	SS20	A1	SS20 A1-9	H&E	
Embedded Priority	SS20	A1	SS20 A1-1	ERG	
Embedded Priority	SS20	A2	SS20 A2-1	ERG	
Embedded Priority	SS20	A3	SS20 A3-1	H&E	Derm Service
Embedded Priority	SS20	A4	SS20 A4-1	H&E	Derm Service
Embedded Priority	SS20	A5	SS20 A5-1	H&E	Derm Service
Embedded Priority	SS20	A6	SS20 A6-1	ERG	Derm Service
Embedded Routine Surgical	SS20	A1	SS20 A1-2	H&E FS Permanent	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A10	SS20 A10-1	RES2 IHC	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A11	SS20 A11-1	H&E	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A12	SS20 A12-1	H&E	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A2	SS20 A2-1	H&E	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A3	SS20 A3-1	HER2 IHC	GYN Surg Service FS/TP Slides
Embedded Routine Surgical	SS20	A4	SS20 A4-1	H&E	GYN Surg Service FS/TP Slides

Goes from ordinal data to discrete!

Continuous Data

Continuous data is numeric and you measure it (height, weight, temperature). It can be measured over time and be put in the context of time.



Identify if the following are Discrete or Continuous

1. The baby weighs 7 lbs 6 ounces pounds
2. There were 200 IHC requests this month
3. The lab was 65 degree Fahrenheit
4. The lab has 6 employees
5. It took the staff 2 days and 3 hours to complete the safety training.



Remember, counting vs measuring...



Two Types of Continuous Data

Discrete

Numbers that can't be broken down, such as number of items

Continuous

Numbers that can be broken down, such as height or weight

Nominal

Data used for naming variables, like hair color

Interval

Numbers with known differences between variables, such as time

Interval:


- Year
- Credit score
- SAT test
- Temperature (C or F)

Ratio


Numbers that have measurable intervals where difference can be determined, such as height or weight

Ratio:


- Weight
- Error rates
- Crime rate
- Length of time




Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓



Part 1 Recap!



1. The baby weighs 7 lbs 6 ounces pounds
2. There were 200 IHC requests this month
3. The lab was 65 degree Fahrenheit
4. The lab has 6 employees
5. It took the staff 2 days and 3 hours to complete the safety training.



Part I: What is Data?

- ✓ Summarize where data fits into the evaluation cycle.
- ✓ Define data.
- ✓ Classify types of data.
- ✓ Identify what data is present and available to you.

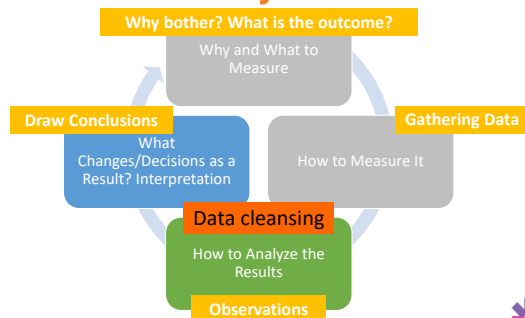


Part II: Preparing Data for Analysis

1. Explain the impact of dirty data
2. Perform a data cleanse



Evaluation Life Cycle



Preparing the Data: Cleansing

60-80% of a data scientists time is spent cleaning data.

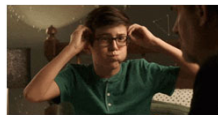


- Data Scrubbing: "The procedure of modifying or removing incomplete, incorrect, inaccurately formatted, or repeated data in a database." (Technopedia).
- AKA: Cleaning, scrubbing, dirty data, unclean data
- Extractions of data or databases



Connie's Customized Checklist

1. Remove duplicates (if you are not counting frequency)



Duplicate Values

- When you have one qualified response entered more than one time.
- CRMs, CMXs, etc typically have a duplicate value finder or preventer – but they are far from perfect.

Tip: Have protocols/SOPs in place to remove duplicates. Staff turnover and people unfamiliar with your database may not understand what to do when they encounter a duplicate and that can cause dirty data!



Formatting Issues

- Inconsistencies with data fields.
- Open ended fields will always need attention.

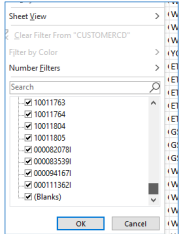
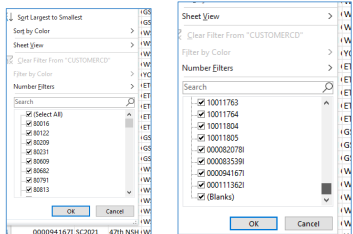
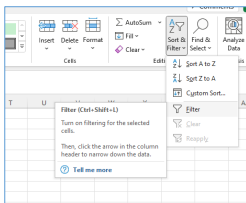
Tip: Consider drop downs, and prefilled items when you can. Test with people to see if there are issues that are preventable (i.e. are dropdowns missing fields, etc.)



0000835391	SC2021	47th NSH + WS10	Histogon\NSH	09/14/21	1 A	Allison Eck; Brad Flowers
0000835391	SC2021	47th NSH + WS11	Adhesive NSH	09/14/21	1 A	David Prine
0000835391	SC2021	47th NSH + WS21	Diagnosti NSH	15-Sep	1 A	Richard Ormscher
0000835391	SC2021	47th NSH + WS32	Discovery NSH	09/14/21	1 A	Kim Pickard; Michele Levitt
0000835391	SC2021	47th NSH + WS40	Using Th\NSH	09/16/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH + WS11	Adhesive NSH	09/14/21	1 A	David Prine
0000835391	SC2021	YOGADAY Morning \NSH		09/15/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH + YOGADAY Morning \NSH		09/15/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH + YOGADAY Morning \NSH		09/16/21	1 A	Gerelyn Henry
0000820781	SC2021	47th NSH + ET05	Stains Be\NSH	09/15/21	0.5 A	Jean Mitchell; Surena Becraft
SC2021	47th NSH + ET01	Modificat\NSH		09/21/21	0.5 A	Elizabeth Chilipala
SC2021	47th NSH + ET01	Modificat\NSH		09/25/21	0.5 A	Elizabeth Chilipala
SC2021	47th NSH + ET01	Modificat\NSH		09/27/21	0.5 A	Elizabeth Chilipala
SC2021	47th NSH + ET01	Modificat\NSH		09/30/21	0.5 A	Elizabeth Chilipala
SC2021	47th NSH + ET02	Chemical NSH		09/15/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/15/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/20/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/24/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/25/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/27/21	0.5 A	Steven Goodman
SC2021	47th NSH + ET02	Chemical NSH		09/27/21	0.5 A	Steven Goodman



Excel Trick: Filter or Sort



Connie's Customized Checklist

1. Remove duplicates (if you are not counting frequency)
2. Identify formatting issues:
 - Formatting for quantitative entries like dates, decimals
 - Formatting for categorical information like states, names, etc.
3. Find missing data and fill in or remove.



Missing Data

Why bother? What is the outcome?

Why and What to Measure

- Data fields that are not complete.
- "Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions."

Tip: Collection desires vs realities (think about the amount of time it would take a person to enter fully the information, or accurately)



0000835391	SC2021	47th NSH WS10	Histogon\NSH	09/14/21	1 A	Allison Eck; Brad Flowers
0000835391	SC2021	47th NSH WS11	Adhesive NSH	09/14/21	1 A	David Prine
0000835391	SC2021	47th NSH WS21	Diagnost\NSH	15-Sep	1 A	Richard Ormesher
0000835391	SC2021	47th NSH WS32	Discovery NSH	09/16/21	1 A	Kim Pickard; Michele Levitt
0000835391	SC2021	47th NSH WS40	Using Tit\NSH	09/16/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH WS11	Adhesive NSH	09/14/21	1 A	David Prine
0000835391	SC2021	47th NSH YOGADAY Morning	\NSH	09/15/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH YOGADAY Morning	\NSH	09/15/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH YOGADAY Morning	\NSH	09/16/21	1 A	Gerelyn Henry
0000835391	SC2021	47th NSH ET05	Stains Be\NSH	09/15/21	0.5 A	Jean Mitchell; Surena Becraft
0000835391	SC2021	47th NSH ET01	Stains Be\NSH	09/21/21	0.5 A	Elizabeth Chlipala
0000835391	SC2021	47th NSH ET01	Stains Be\NSH	09/25/21	0.5 A	Elizabeth Chlipala
0000835391	SC2021	47th NSH ET01	Modificat\NSH	09/27/21	0.5 A	Elizabeth Chlipala
0000835391	SC2021	47th NSH ET01	Modificat\NSH	09/30/21	0.5 A	Elizabeth Chlipala
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/15/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/15/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/20/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/24/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/25/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/27/21	0.5 A	Steven Goodman
0000835391	SC2021	47th NSH ET02	Chemical NSH	09/27/21	0.5 A	Steven Goodman

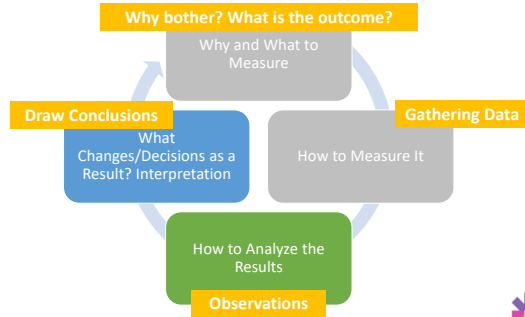


Part III: Analyzing Data

1. Interpret measures of central tendency.
2. Analyze data



Evaluation Life Cycle



Two Types of Quantitative Analysis

- Descriptive: focus on describing a sample of a population.
- Inferential: makes predictions based on the sample about the ENTIRE population.



"In other words, we use one group of statistical methods – descriptive statistics – to investigate the slice of cake, and another group of methods – inferential statistics – to draw conclusions about the entire cake."



-Karyn Warren, PhD



Frequency

The rate at which something occurs or is repeated (over time or in a sample).



Central Tendency

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- A single measure may not give you the full picture because extreme values can impact each measurement.



Mean

- Average
- Total all numbers in the data set and divide by the number of responses.



Median

- The value that appears in the center of the data set.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list



Mode

- The number that appears the most in the data set.



Range

- Range: describes the difference between the smallest and largest value.
- To calculate this, you first need to use numeric codes to represent each grade, i.e. A = 1, A- = 2, B = 3, etc. The range would be $5 - 1 = 4$. So in this simple example, the range is 4. This is an easy calculation to carry out. The range is useful because it offers a basic understanding of how spread out the values in a dataset are.



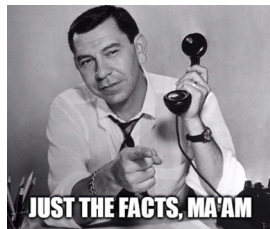
Why Central Tendency Matters

13: Other (please describe): Which type of POL/Specialty Lab?	14: Non-Registered HT/HTL: How many full time personnel work in your lab with each of the following job functions? Leave blank if no staff perform a particular function.	14: Registered HT/HTL: How many full time personnel work in your lab with each of the following job functions? Leave blank if no staff perform a particular function.	14: Laboratory Assistant: How many full time personnel work in your lab with each of the following job functions? Leave blank if no staff perform a particular function.
MEDIAN	2	4	2
MODE	1	1	1
MEAN	4.158054711	6.45	4.506527415



OBSERVATION - Analysis

- Observation = FACTS
- Analysis is just what the numbers or information say.



ASCP BOC QUALIFICATION STATISTICS: 2006 - 2021							
QUALIFICATION TYPE	YEAR	MEAN	RANGE OF SCORES	TOTAL # TAKING QUALIFICATION	TOTAL PASS	TOTAL FAIL	TOTAL QUALIFIED
IMMUNOHISTOCHEMISTRY (QIHC)	2006	455	240 - 640	60	42	18	30%
	2007	455	250 - 650	52	42	10	19%
	2008	458	260 - 620	70	58	12	17%
	2009	471	350 - 650	76	62	14	18%
	2010	499	270 - 680	88	80	8	9%
	2011	463	210 - 680	84	69	15	18%
	2012	475	220 - 680	104	84	20	19%
	2013	433	220 - 620	123	74	49	40%
	2014	430	100 - 680	133	72	61	46%
	2015	455	217 - 906	113	57	56	50%
	2016	459	250 - 814	124	66	58	47%
	2017	399	212 - 759	131	46	85	35%
	2018	410	189 - 759	135	57	78	58%
	2019	428	236 - 759	135	67	68	50%
	2020	418	212 - 759	131	64	67	51%
	2021	408	217 - 749	141	50	91	65%

- More test takers passed the QIHC by percent in 2010 than any other year measured. (continuous data example)
- The total number of test takers in 2021 was 141. (discrete data example)

Remember, counting vs measuring...



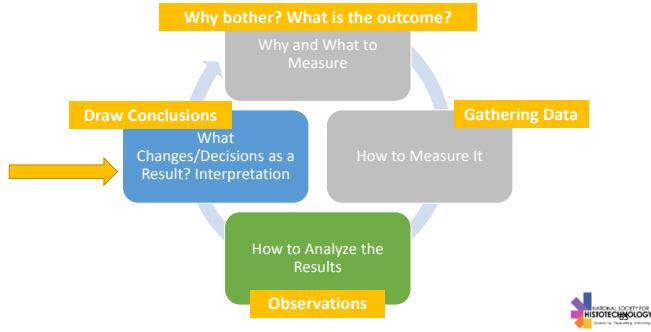
Why can't I make conclusions like...

The QIHC exam is too hard.

- Because the data doesn't show us that. Maybe the test takers didn't have enough study tools, maybe the technology is outdated for taking the exam.
- More importantly, we aren't at that step yet!



Evaluation Life Cycle



Instructions


- Looking at the data set, type in an observation about the data.



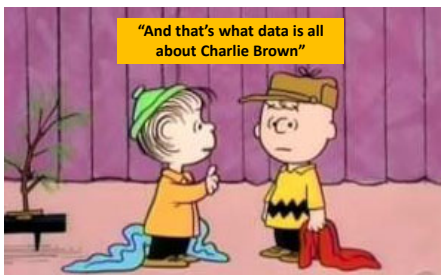

	Total	Hospital	Private Independent
Sample Size	1,088	809	279
Grossing Questions			
Who performs grossing in your lab?			
Non-Registered HT/HTL	9%	7%	15%
Registered HT/HTL	31%	26%	44%
Laboratory Assistant	3%	2%	5%
Laboratory Technician	8%	7%	11%
Grossing Assistant	22%	20%	27%
Supervisor	9%	9%	11%
Manager	3%	3%	3%
Pathologist	31%	35%	21%
Pathologist's Assistant	53%	62%	25%
Resident	12%	16%	2%
Other (please specify)	2%	2%	2%
What are the average number of cassettes grossed per hour?			
Mean	46.9	46.1	48.7
Median	30.0	30.0	35.0
What are the average number of containers grossed per hour?			
Mean	27.6	23.3	36.1

Diversity of Thought Makes Data Better!

- We all look at things differently, even numbers
- Using many sets of eyes and brains can analyze data better



"And that's what data is all about Charlie Brown"

For Your Reading Pleasure

- <https://uxdesign.cc/data-for-design-f33fd7419cc8>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7886543/>
- <https://towardsdatascience.com/9-common-mistakes-beginner-data-scientists-make-91255ddd1311>



References

- Evaluation, A Systematic Approach: Seventh Edition, Rossi, Lipsey, and Freeman
- DataFlog, 10 Really Cool Data Cartoons You Have to See! <https://dataflog.com/read/10-really-cool-data-cartoons-you-have-to-see/>
- Merriam-Webster Dictionary, 2022
- [ASCP QIHC Pass Rates – published Jan 18, 2022](#)
- 2022 NSH Workload Study
- Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi: 10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.