Analysis of Data with Missing Values

EGI SAFITRI, S.MAT., M.SI

Outline

1. Background
2. Extent of missingness
3. Pattern of missingness
4. Strategies in dealing with missing data
5. Examples
6. Prevention of missing data
7. Summary

Background

- Missing data very common in research studies
- Best solution? Avoid them!!
- Not taught in many statistical courses
- Handling missing data
- Reporting of missing data

Background Cont.

- Preventing missing data
- Study designs: (1) longitudinal vs. cross-sectional, (2) randomized vs. observational studies

Coding Missing Data

- Often coded as values that are not possible, e.g. 999, -999
- If coded that way, make sure to specify them as missing in data analysis
- Sometimes such coding scheme developed to list different reasons of missing data
- If they are not important, safer to leave blank or enter “.”

Extent of Missing Data

- <1%, <5%, 10% or higher?
- By item/variable or by subject?
- Most values missing in one or a few variables?
- Missing values in one or a few primary variables?
- Missing values in one or a few secondary variables?
- Few values missing in several variables?

Pattern of Missing Data

- Item-level missingness
- Subject-level missingness
- Missing in outcome or predictor variable?
- Missing in continuous or categorical variable?
- Designed trials or designed surveys?
- Unstructured surveys?

Type of Missing Data

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

Missing Completely at Random (MCAR)

- Reasons: Lab error, road accident, bad weather, residential move, family emergency, inadvertently skipping questions
- Example: income and age, prob of missing data on income does not depend on income and age, i.e. participants of all ages likely to report income
- Little's MCAR test
- Non-significant test => MCAR

Missing at Random (MAR)

- Also called *ignorable* missingness
- Probability of missingness on Y does not depend on Y itself after controlling for other variables
- Example: prob of missingness on income depends on age (older more likely to report than younger), but participants within each group equally likely to report income, i.e. prob of missingness on income unrelated within an age group

Missing Not at Random (MNAR)

- Also called *nonignorable* missingness
- Missingness is not MCAR or MAR
- Probability of missingness on Y depends on values of Y itself, e.g. people with higher income do not report income (even after controlling for other factors)
- No statistical tests for MAR and MNAR, but can run some sensitivity analyses

Missing Data Patterns: Example (Polit, 2010)

	# Follow-Up	Missing data reason	Cotinine M, ng/mL (All Subjects)	Cotinine M (90 Subjects)
No missing	50 men 50 women	-	185.0	-
MCAR	45 men 45 women	Lab error, road accident, bad weather, residential move, family emergency	185.0	185.5
MAR	40 men 50 women	Male dropouts lost interest, men smoked >women, dropout unrelated to cotinine level	185.0	175.0
MNAR	40 men 50 women	Male dropouts resumed heavy smoking, embarrassed to continue, dropout related to cotinine level	185.0	165.0

Handling Missing Data

- Ignoring missing data:
  1. Pairwise deletion, e.g. bivariate correlation, also called *available-case analysis*
  2. Listwise or casewise deletion, e.g. multiple regression, also called *complete-case analysis*
- Ignoring missing data, but using all available data: GEE, mixed models, survival analysis
- Imputation: (1) single-imputation, (2) multiple imputation
- Sensitivity analysis, e.g. worst outcome

Single Imputation

- Imputation using a central tendency measure: Continuous-> Mean, Ordinal-> median, Nominal-> mode
- Subgroup imputation
- LOCF
- Regression
- Maximum likelihood
- Expectation maximization (EM) algorithm

Multiple Imputation, Newgard and Haukoos (2007)

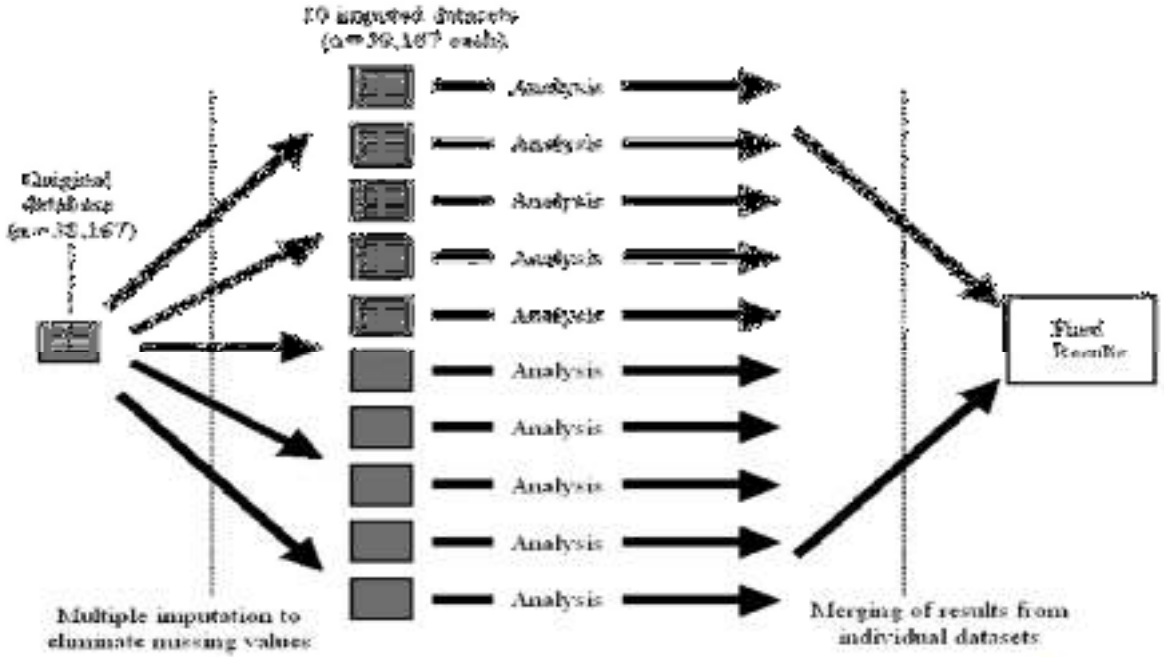


Figure 1. Overview of multiple imputation process with sample database (n = 38,167).

An Example of Sensitivity Analysis, Polit (2010)

	Listwise Deletion	Mean Imputation	Sub-group Mean	Regression Imputation	EM
N	1753	1904	1904	1904	1904
Mean, Imp (N=151)	-	45.50	45.52 (m=46.36, f=44.7)	46.43	46.38
SD, Imp	-	0.00	0.84	5.95	5.93
Mean, Full	45.50	45.50	45.50	45.57	45.57
SD, Full	13.83	13.27	13.28	13.37	13.49
R^2	0.185	0.170	0.171	0.198	0.197

Example: Mean Imputation

The Relationship Between Sleep and Physical Function in Community-Dwelling Adults

A Pilot Study

Rebecca Ann Lorenz, PhD; Chakra B. Budhathoki, PhD; Gurpreet K. Kalra, MS; Kathy C. Richards, PhD

More than 50% of community-dwelling adults have sleep complaints. Because aging is associated with decline in physical function, coexistent sleep difficulties may exacerbate functional decline. This pilot study explored the relationships between sleep, age, chronic disease burden, and physical function among 50 community-dwelling older adults. Findings revealed significant relationships between total sleep time and preclinical disability ($r = -0.33$, $P \leq .05$) and mobility difficulty ($r = -0.36$, $P \leq .05$). A regression analysis showed that total sleep time was significantly associated with mobility difficulty and preclinical disability, even after controlling for chronic disease burden. These findings suggest that total sleep time may be a catalyst for functional decline. **Key words:** *adults, physical function, sleep*

Example: Mode Imputation

Effect of Shared Governance on Nurse-Sensitive Indicator and Satisfaction Outcomes: An International Comparison

Karen Gabel Speroni, PhD, MHSA, BSN, RN
Kirsten Wisner, PhD, RNC-OB, CNS, C-EFM
Amy Stafford, DNP, RN, CIC, CMSRN
Fiona Haines, MCur RN, RM, Adv Mid Adv NeonSc, CPHQ
Majeda A. AL-Ruzzieh, PhD, RN
Cynthia Walters, DNP, RN, NE-BC
Chakra Budhathoki, PhD

Objective: Researchers examined associations between Index for Professional Nursing Governance (IPNG) scores and outcomes, by US and international hospitals.
Background: Nursing governance and effects on nurse-related outcomes are not well studied.
Methods: Associations were evaluated using average IPNG scores from 2170 RNs and nurse-sensitive indi-

Internationally, self-governance significantly outperformed traditional governance and shared governance for 5 of 12 (41.7%) outcomes (NSI = 2, patient satisfaction = 3).
Conclusions: Shared governance is a strategy that can be considered by nurse leaders for improving select outcomes.

Example: LOCF

Uridine supplementation in the treatment of HIV lipoatrophy: Results of ACTG 5229

Grace A McComsey[1], Ulrich A Walker[2], Chakra B Budhathoki[3], Zhaohui Su[3], Judith S Currier[4], Lisa Kosmiski[5], Linda G Naini[6], Stéphannie Charles[7], Kathy Medvik[1], and Judith A Aberg[8] on behalf of the AIDS Clinical Trials Group A5229 Team

[1] Case Western Reserve University and University Hospitals Case Medical Center, Cleveland, Ohio [2] Basel University Department of Rheumatology, Basel, Switzerland [3] Statistical and Data Analysis Center, Harvard School of Public Health, Boston, MA [4] University of California, Los Angeles, CA [5] University of Colorado, Denver, CO [6] ACTG Operations Center, Silver Spring, Maryland [7] Frontier Science Technology and Research Foundation, Amherst, NY, New York [8] University School of Medicine, New York, NY

Abstract

BACKGROUND—Lipoatrophy is prevalent on thymidine NRTIs (tNRTI). A pilot trial showed that uridine (NucleomaxX®) increased limb fat.

Example: MI

Reducing Preschool Behavior Problems in an Urban Mental Health Clinic: A Pragmatic, Non-Inferiority Trial

Deborah Gross, DNSc, RN, Harolyn M.E. Belcher, MD, MHS, Chakra Budhathoki, PhD, Mirian E. Ofonedu, PhD, LCSW-C, Daryl Dutrow, MBA, MSW, Melissa Kurtz Uveges, PhD, MAR, RN, Eric Slade, PhD

Objective: This pragmatic, randomized, non-inferiority trial compared the effectiveness and cost of group-based parent management training with mastery-based individual coaching parent management training in a low-income, predominantly African American sample.

Method: Parents seeking treatment for their 2- to 5-year-old children's behavior problems in an urban fee-for-service child mental health clinic were randomized to the Chicago Parent Program (CPP; n = 81) or Parent–Child Interaction Therapy (PCIT; n = 80). Consent followed clinic intake and diagnostic assessment and parent management training was delivered by clinicians employed at the clinic. Primary outcome measures were externalizing child behavior problems, assessed at baseline and postintervention follow-up, using the Child Behavior Checklist (CBCL) and average per-participant treatment cost.

Considerations to Decrease Missing Data

- Some attrition unavoidable
- Take attrition or drop-out into account in estimating sample size
- Analyze all randomized subjects in RCTs
- Try to increase response rate in surveys
- Ask questions that decrease refusal rate, e.g. exact income vs. income categories
- Logistical support for clinic visits if ethical

Considerations to Decrease Missing Data Cont.

- Study design
- Reduce drop outs
- One may discontinue assigned treatment, but try to keep them in the study follow-ups
- They can switch treatment, or completely discontinue
- Communication
- Training

Summary

- Missing data are common
- Data analysis plan should specify how missing data would be handled
- Better study designs
- Account for expected attrition in sample size estimation
- Better data analyses: imputation is common
- GEE, mixed models, survival analyses do not need imputation

Checking Your Data With Outlier Analyses

Agenda

What is an outlier?

Why are outlier analyses important for data validity and reliability?

What steps should states take after an outlier analysis?

How can states conduct and display an outlier analysis?

Discussion

What Is an Outlier?

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." (Hawkins, 1980)

Outlier analyses include investigating whether the data are valid or invalid.

States define what value or combination of values are outside the expected norm.

What Is an Outlier? (cont.)

Invalid outliers are the target of outlier analysis, as they represent errors in the data.

Valid outliers may appear to be outside the norm, but investigation demonstrates that the data are not in error.

Valid outliers may occur due to random variation, which occurs due to chance and is inherent in a system.

Why Are Outlier Analyses Important for Data Validity and Reliability?

Help identify errors in the data

Reveal systematic errors in data collection, coding, or entry

Identify LEAs/LLAs that are performing better or worse than the norm

Provide opportunities for understanding the factors behind high performance or providing targeted technical assistance where it is needed

What Steps Should States Take After an Outlier Analysis?

Investigate any identified outliers to understand why the data are so different from the norm.

Follow up with the LEAs/LLAs to determine the root cause of the outlying data.

For more information on examining root cause, states can review *Equity, Inclusion, and Opportunity: How to Address Success Gaps, White Paper.*

Investigative Questions

1. Are the outliers found in just one LEA or LLA?
2. Are the same outliers identified in more than one dataset?
3. Are multiple outliers commonly identified in the same LEAs/LLAs?
4. Are the LEAs/LLAs with outliers using non-standard data collection definitions?

Investigative Questions (cont'd)

5. Are the LEAs/LLAs with outliers using non-standard methods for aggregating the data?
6. Are the LEAs/LLAs with outliers using non-standard methods to collect the data?
7. Did the small *n* size affect the analysis?

How Can States Conduct and Display an Outlier Analysis?

There are several possible approaches to conduct outlier analyses.

The IDC Outlier Analyses: Step-by-Step Guide includes six different tutorials covering different methods states can use to identify and visualize outliers.

IDC also created IDEA Data Quality: Outlier Analysis Tool , an Excel-based tool states can use to identify outliers using the interquartile range approach described in the step-by-step tutorials.

Example Outlier Analysis

Steps to conduct analysis in the IDC Excel Tool

- Paste Here Tab
  - Report the measures in the columns.
  - Report the LEA/LLAs in the rows.
- Outlier Analysis
  - Results will be displayed on this tab.

Discussion Questions

Are the data of high quality?

Did you expect outliers where there aren't any?

Might there be outliers that are not identified?

Discussion Questions (cont'd)

What is occurring in these programs?

- Are the LEAs/LLAs with outliers using non-standard data collection definitions?
- Are the LEAs/LLAs with outliers using non-standard methods for aggregating the data?
- Are the LEAs/LLAs with outliers using non-standard methods to collect the data?
- Did the small *n* size affect the analysis?

Who Should Review the Data With the Data Manager?

Part C Coordinator

Program Director

Service Providers

How Can You Show Outlier Analyses?

Heat Maps in Excel

Alternative Quick Heat Maps

Dot Plots in Excel

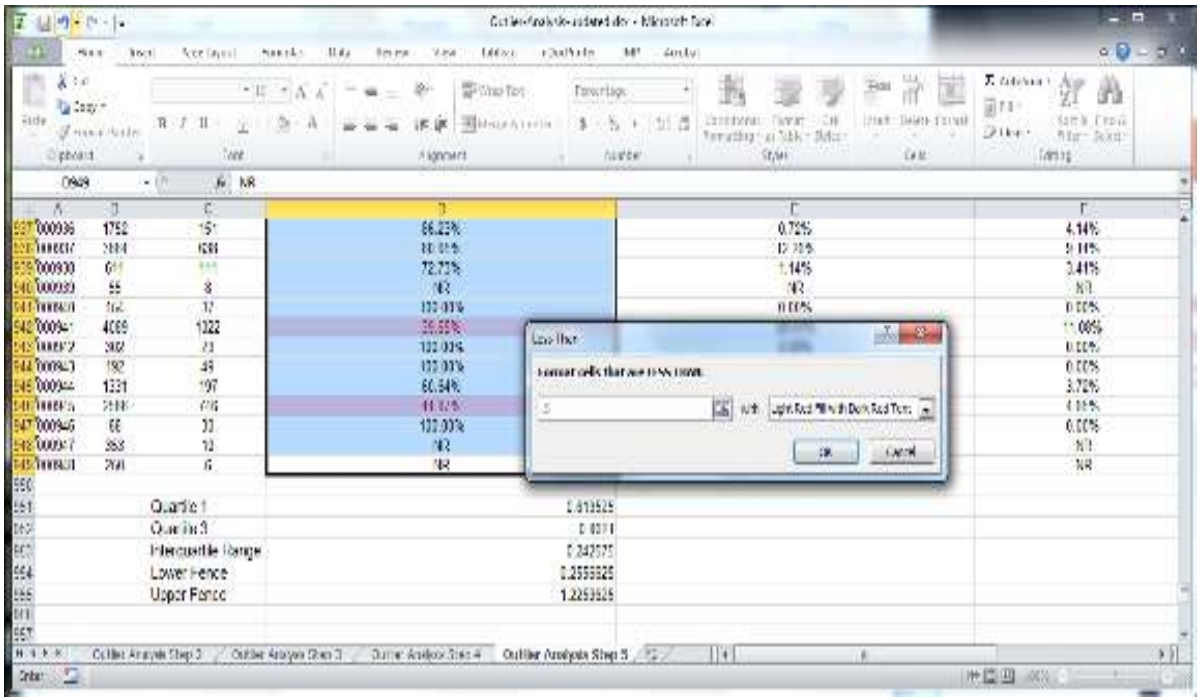
Dot Plots in Tableau

Heat Maps in Excel

Use conditional formatting in conjunction with sorting

Can be particularly important when you have multiple columns of data or lots of rows

Heat Maps in Excel (cont'd)



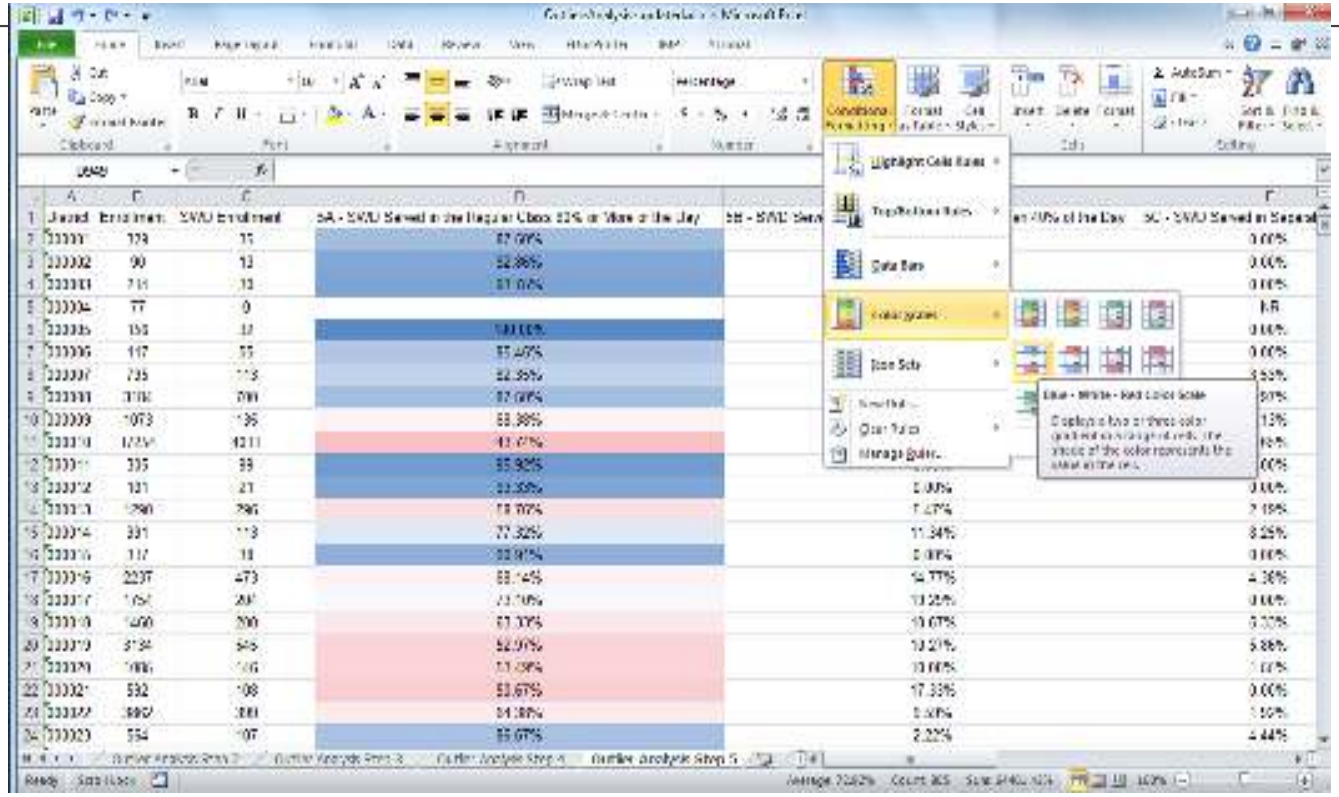
Alternative Quick Heat Maps

A quicker approach to heat maps in Excel is using the Color Scales feature found under the Conditional Formatting drop-down menu.

This will automatically color a set of selected cells based on the range of values.

Data bars and icon sets can also be used to quickly identify possible outliers.

Alternative Quick Heat Maps



Dot Plots in Excel

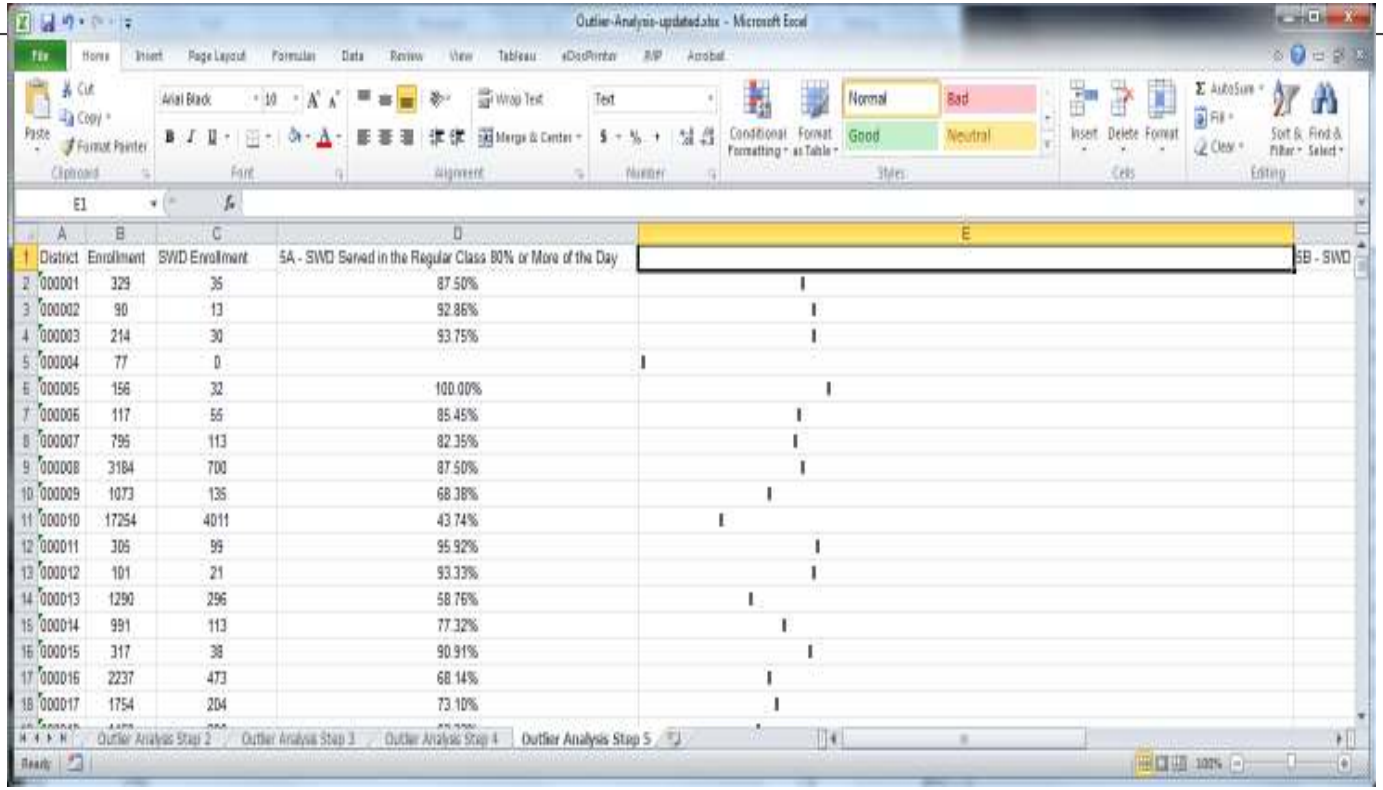
Use an in-cell formula that will create a simple dot plot next to your data

Look for data points that don't seem to fit with the others

Use an Excel formula to create an in-cell chart; dot plots will always remain in line with the data

Excel's standard chart functions can also assist in identifying outliers but will ultimately be disconnected from the original data.

Dot Plots in Excel



District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD
000001	329	35	87.50%	
000002	90	13	92.86%	
000003	214	30	93.75%	
000004	77	0		
000005	156	32	100.00%	
000006	117	55	85.45%	
000007	795	113	82.35%	
000008	3184	700	87.50%	
000009	1073	135	68.38%	
000010	17254	4011	43.74%	
000011	305	99	95.92%	
000012	101	21	93.33%	
000013	1290	296	58.76%	
000014	991	113	77.32%	
000015	317	38	90.91%	
000016	2237	473	68.14%	
000017	1754	204	73.10%	

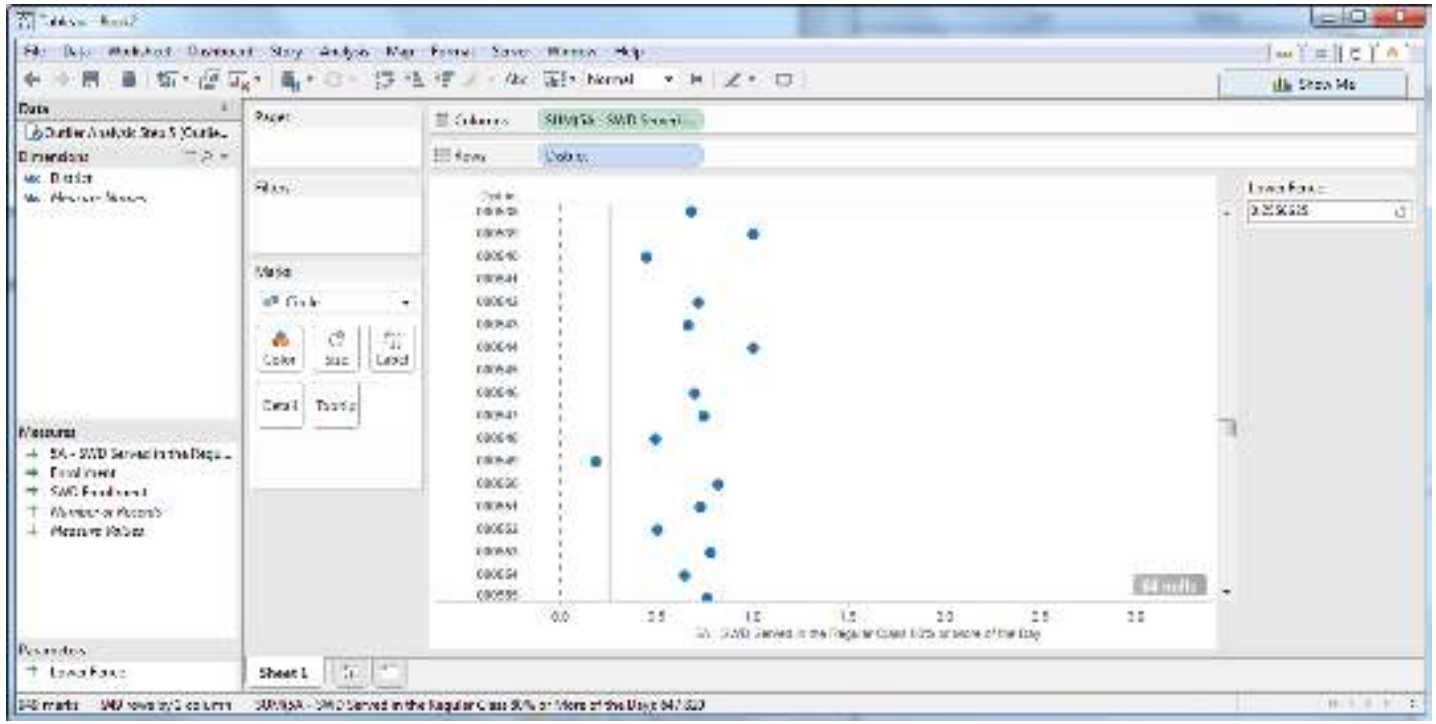
Dot Plots in Tableau

If you have a lot of data, using an interactive visualization program like Tableau can be useful.

You can quickly and easily visualize hundreds of rows and multiple measures.

If you are visualizing Public Data, you can use Tableau Public for free. For private data, you would need at least a personal version of the Tableau desktop license.

Dot Plots in Tableau



Discussion—How to Use Outlier Data for Program Improvement

Why might there be so many more referrals in one LLA? (e.g., co-location of programs)?

Are there fewer programs in some LLAs to which to refer children who exit Part C?

Is there a need for better information about programs to which to refer children who exit?

Is there lack of clarity about the referral process?

Who Should Be Involved?

Part C Coordinator

Program Director

Service Providers

Other