the ThreatLab
Integrating Research Into Practice

Data Analysis Workbook for the Insider Threat Community

A Collaborative Workbook

November 2023

AUTHORS
Michael Hunter, Dean Cisler,
Whitney Fauth, Kirk Kennedy

DPAC-2023-295 | PERSEREC-PA-23-04

OPA | PERSEREC
OFFICE OF PEOPLE ANALYTICS | DEFENSE PERSONNEL AND SECURITY RESEARCH CENTER

Authors

DEFENSE PERSONNEL AND SECURITY RESEARCH CENTER
Andrée Rose

PERATON
Dean Cisler
Whitney Fauth
Michael Hunter
Kirk Kennedy

Sponsors

PERSEREC is a Department of Defense entity dedicated to improving the effectiveness, efficiency, and fairness of DoD personnel suitability, security, and reliability systems. PERSEREC is part of the Office of People Analytics (OPA), which is a component of the Defense Human Resources Activity (DHRA) under the Office of the Under Secretary of Defense (Personnel and Readiness).

Within the National Counterintelligence and Security Center (NCSC), the primary mission of the National Insider Threat Task Force (NITTF) is to develop a Government-wide insider threat program for deterring, detecting, and mitigating insider threats, including the safeguarding of classified information from exploitation.

The Office of the Secretary of Defense (OSD) is responsible for policy development, planning, resource management and program evaluation. OSD includes the offices of top civilian defense decision -makers with regard to personnel, weapons acquisition, research, intelligence, and fiscal policy, as well as offices the Secretary establishes to assist in carrying out assigned responsibilities.

Point of Contact

dodhra.threatlab@mail.mil

Suggested Citation

Hunter, M., Cisler, D., Fauth, W., Kennedy, K., Rose A. (2023). *Data Analysis Workbook for the Insider Threat Community: A Collaborative Workbook.* Monterey, CA: Defense Personnel and Security Research Center/Defense Personnel Analytics Center.

Contents

Learning Objectives and Assumptions

Learning Objectives

This analysis Workbook is comprised of two main sections and is designed to:

- provide a basic understanding of research design and statistical principles,
- demonstrate how to develop research questions based on data,
- offer a framework for thinking about and generating important research questions,
- demonstrate how to conduct basic statistical analyses and data visualization, and
- increase the Insider Threat (InT) hubs' workforce knowledge on the minimum data essential for meaningful reporting and analysis.

This Workbook should be used in conjunction with the Workbook's Excel worksheet, which includes simulated database variables along with additional tabs for creating tables, figures, and statistical calculations. The simulated data set along with the functions and graphics for this Workbook are available upon request (please contact: DODHRA.THREATLAB@MAIL.MIL). While the reader is expected to have some experience using Excel, there is supplemental information on the functions used in this worksheet to help guide the reader (See Appendix).

Assumptions

This Workbook requires the reader to have some (even minimal) knowledge in research design and statistics. If certain topics within this Workbook require more attention or prerequisite knowledge, we encourage the reader to review the additional resources and references provided, or research certain topics outside of this Workbook to establish a basic understanding of statistical methodology. The reader is encouraged throughout this Workbook to think critically about research questions and analysis to fully utilize the content and exercises. The reader should have a baseline familiarity with InT vocabulary and organizational requirements. The Workbook references key terms as footnotes to inform the reader.

PART I: Workbook Overview of Research Design and Statistical Principles

Introduction

To strengthen the InT community's ability to identify and improve their prevention and response efforts, Defense Personnel and Security Research Center's (PERSEREC) Threat Lab developed an analysis Workbook for use by the InT hubs. This Workbook can be used as a template for the InT hubs to develop their own analysis plan that is specific to agency needs and data holdings.

The Impact Potential of Quality Data Analysis

Data analysis is the practice of working with data to glean meaningful information, which can then be used to make informed decisions. Data analysis involves collecting, cleaning, analyzing, and interpreting

data. For the InT hubs, data analysis can be used to detect (through analysis) and prevent (through informed decision-making) the potential threats that DoD insiders may pose to the United States through espionage, terrorism, and unauthorized disclosure of national security information (DoD Office of the Inspector General, 2022). Data analysis can also identify gaps and deficiencies in DoD security programs, policies, and procedures.

Some of the data sources that can be used for data analysis within InT include personnel security, physical security, information security, law enforcement, counterintelligence, user activity monitoring, and information assurance. Data analysis also involves various techniques and tools, from basic frequency and categorical analyses to trend analysis, advanced data visualization, and machine learning (Zimmerman et al., 2018). Data analysis can help reveal patterns, trends, anomalies, and correlations that can indicate potential InTs or vulnerabilities (Government Accountability Office, 2015).

Data Analysis for Program Evaluation and Targeted Deterrence Campaigns

Data analysis is a necessary tool for program evaluation. First and foremost, data analysis for program evaluation can be employed to support the DoD Data Strategy vision of "becoming a data-centric organization that uses data at speed and scale for operational advantage and increased efficiency" (DoD, 2020). The objective is to inform decision-making across a broad spectrum of traditional and emerging functions (Anton et al., 2019). Data analysis can be performed to evaluate the capabilities, effectiveness, feasibility, and costs of proposed and alternative programs being considered for any variety of purposes. Data analysis in InT program evaluation can be used to develop targeted deterrence campaigns for the InT hubs by providing data-driven training and awareness resources on InT detection, mitigation, and reporting for personnel and stakeholders (Defense Counterintelligence and Security Agency/Center for Development of Security Excellence, n.d.).

Initial Steps to Complete Before Analyzing Data

Before conducting any analysis, consider these critical initial steps. These include developing your research question(s), identifying variables of interest, and deciding on the most appropriate statistical analysis to use on your data.

Develop Research Questions/Hypotheses

A hypothesis is a tentative statement that predicts the relationship between two or more variables (e.g., time and the number of InT reports). To develop a research hypothesis, follow these steps:

1. Ask a question that you want to answer within the scope of your data. The question should be focused, specific, and researchable within the constraints of your project and data holdings. For example: Is there a rise in specific types of InT reports over the course of the last 4 years?
2. Conduct necessary preliminary research to find out what is already known about the topic. Look for previous studies that can inform assumptions or theories about what your research will find. For example: Previous research has shown a decrease in the number of InT reports within specific DoD components, but not others in the last 4 years (note: this is a hypothetical example and not based on real data or previous research).

3. Formulate one or more hypotheses based on your expected answer to the question and the existing knowledge on the topic.
4. Refine your hypothesis to make it testable and specific. You might have to define and operationalize[1] the variables you measure and make some assumptions about how you will measure them. For example, a useful hypothesis might be: "Army will show a decrease in InT reports over the last 4 years, while Air Force and Navy will see an increase" (note: this is a hypothetical example and not based on real data or previous research).

Overview of Quantitative Analysis

Quantitative analyses collect and evaluate numeric, measurable data, such as the number of InT reports in a given year, rank, age, or gender to understand the behavior and performance of an individual or organization (i.e., variables that can be collected). Quantitative data and analysis can be used to find patterns and averages, make predictions, test relationships, and generalize results to wider populations.

One important concept within quantitative analysis is defining the population of interest and determining whether it is necessary to extract a sample from that population; that is, defining the population of interest (including time as feature of the population) into account and developing an appropriate sampling strategy are critical parts of effective quantitative analysis. In quantitative analysis, a population is the entire group that you want to draw conclusions from. For example, you might be interested in InT incidents among U.S. Navy military personnel; in this case, the population of interest would be all active duty and Reserve component personnel in the U.S. Navy. The population can be defined in terms of geographical location, age, or many other characteristics. A population's total size, when known, can be denoted as "*N*".

In our example, you would analyze a subset of the total Navy military population instead of the entire Navy (as you may not have access to the entire Navy population). As a subset, the size of your sample (or sample size denoted as "*n*") is always less than the total size of the population.

> A sample is a subset of the population from which data will be collected in an attempt to accurately represent the population. Sampling is used when it is not feasible to collect data on an entire population, which is often the case in research.

Population values are called parameters and sample values are called statistics. Parameters are precise but usually unknown values (because researchers often do not have access to the entire population of interest), while statistics are estimates about a population that have an associated margin of error[2]. To draw valid conclusions about a population from a sample, you need to use

[1] Operationalization is an important term within research design. It is the process of defining how a concept will be measured, observed, or manipulated within a particular study or data analysis. In other words, it is used to translate a theoretical, conceptual variable of interest into a set of specific operations or procedures that define the variable's meaning in a specific study or data analysis. Operationalization is particularly important for quantitative analyses, where variables need to be clearly defined and measured in order to test hypotheses and draw conclusions.

[2] The margin of error in statistics is a measure of the uncertainty in the results (based on the data). It shows how much the results might differ from the true value of the population parameter that you are interested in estimating.

appropriate sampling methods that ensure your sample represents the population. There are different types of sampling methods, such as random sampling, stratified sampling, and cluster sampling. It is important to select a statistically reliable sample to reduce the margin of error in statistical estimates and ensure results best reflect the population of interest.

Types of Quantitative Variables

A variable is an attribute of data that can have different values. Variables can be classified based on the type of data they contain and the role they play in the research. To construct any quantitative analysis a researcher must be able to define the outcome and independent variables, which is briefly described in the next section.

Outcome and Independent Variables

An outcome variable represents the effect of change in pre-defined independent variable(s). It is the thing that you want to track, or improve, and it's always measured. The outcome variable is also known as a *dependent* variable because it *depends* on the value of independent variable(s). The independent variable, where there is often more than one, represents the cause of the change in the outcome variable, which is also called a predictor variable. For example, if you want to study how InT reports changed over time, the number (or frequency) of InT reports would be your outcome variable and time (in years) would be the independent variable. Another example would be if you want to study how certain potential risk indicators (PRIs)[3] affect the number of InT reports, the independent variable would be the specific PRIs and the outcome variable would be the number of actual InT reports associated with each PRI.

The type of variable determines how you can analyze and present your data using statistical methods and tools. There are four main types of quantitative variables that can be used in research and statistics: discrete, nominal, ordinal and continuous (Centers for Disease Control and Prevention, n.d.), and string variables[4].

- **discrete variables:** represent counts of individual items or values. Discrete variables can only have integer values that cannot be divided into smaller units. For example, the number of InT incident reports is a discrete variable (e.g., because there should not be 1.6 incident reports).
- **nominal variables:** represent groups with no rank or order between them, are also defined as categorical variables. While nominal variables can have numeric or nonnumeric values, the numbers do not represent any quantity or measurement. For example, organizational names, reporting thresholds, or gender are nominal variables.
- **ordinal variables:** represent groups that are ranked in a specific order. Ordinal variables can have numeric or nonnumeric values, but the numbers only indicate the position or rank of the

[3] PRIs are observable and reportable behaviors that may be exhibited by those at risk of becoming an InT. PRIs characteristics overlap with the adjudicative guidelines, which can also be used to determine the type and level of InT risk (see this resource for review: https://www.cdse.edu/Portals/124/Documents/student-guides/INT210-guide.pdf).

[4] While string variables are defined here, a string variable is actually a type of qualitative variable, which is described in the next section.

groups, not the magnitude of difference between them. For example, finishing place in a race (i.e., 1st, 2nd, or 3rd), education level, or DITMAC triage levels (i.e., blue, grey, purple, and brown) are ordinal variables.

- **continuous variables:** represent measurements of continuous or nonfinite values. Continuous variables can have any value in a range and can be divided into smaller units. For example, years, distance, age, or heart rate are continuous variables.

Types of Statistical Analyses

Statistical analysis is the process of (a) collecting and evaluating data to (b) answer research questions or test hypotheses by (c) using statistical methods and other data analysis techniques. Even if you are not currently conducting statistical analyses, you should collect and organize data as if you will eventually conduct statistical analyses. There are different types of statistical analyses depending on the purpose, data type(s), and research design of the study. Some of the common types of statistical analysis include Descriptive Statistics, Categorical Analyses, Measures of Association Analyses, and Trend Analyses.

Descriptive Statistics

Descriptive statistics involve summarizing and organizing the data using measures of central tendency (mean and/or median), variability (minimum, maximum, ranges, and standard deviations), and frequency to describe the main features and patterns of the data. Descriptive statistics do not make inferences or draw conclusions about the population, but only present the information within the sample. An example of descriptive statistics in the context of InT research is reporting the frequency (count) and annual cost (in US dollars) of insider attacks in different organizations, industries, and regions.

Categorical Analyses

Categorical analysis involves analyzing data that are divided into groups or categories, such as nominal or ordinal variables. Categorical variables are variables that can be divided into groups, such as gender, installation/location, PRIs, or reporting thresholds. Categorical statistics can include techniques from more basic analysis such as frequency tables, pie charts, bar charts, and cross-tabulation, to more advanced techniques such as odds ratios, relative risk, and multinomial regression. Categorical statistics help compare and contrast different groups or categories based on their characteristics or outcomes. For example, a categorical analysis could involve the classification of current types of insiders, levels of access, types of motivations behind an insider attack, and types of methods they used.

Frequency Analyses

To use frequencies in a categorical analysis, you need to count how many times each category or group occurs in your data. Frequencies are also called counts and are a very concise way of summarizing and organizing data. They are typically displayed as frequency tables, which shows the number of observations in each category or group. A frequency table usually has at least two columns: one for the values or categories and one for the frequencies. It may also show the proportion of the total population for each group. Proportions are especially valuable because you do not necessarily need to know how many people are in the total group to interpret or compare results with another group or population.

For example, Table 1 shows the frequency of (simulated data) InT reports in a 5-year period categorized by gender. Table 1 simply lists the categories (gender) and their corresponding frequencies (of reports) and proportions. In this example, there were 40 InT reports for females and 60 for males. For numerical variables, the frequency table may group the values into intervals that show the frequencies for each interval.

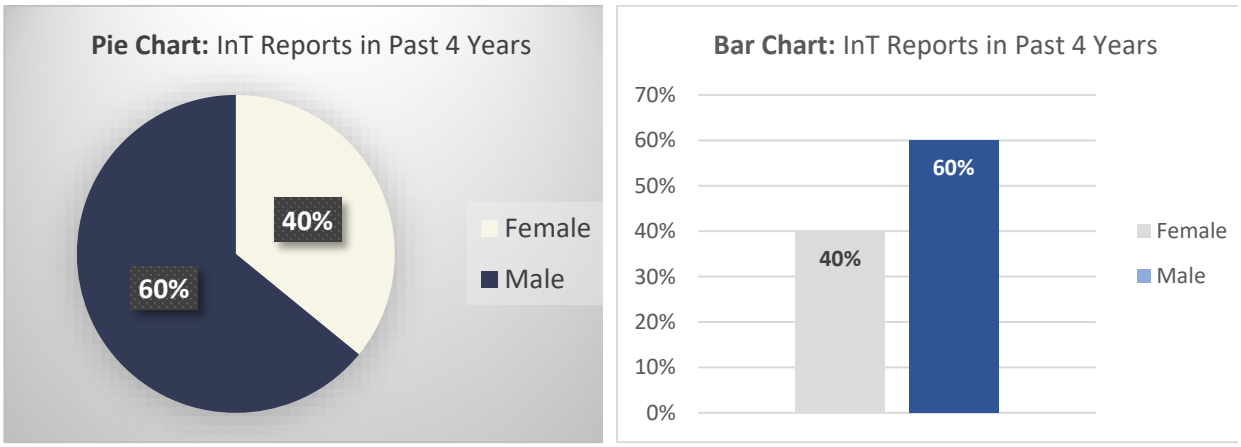
Table 1: Frequency Table of Simulated Data on Reporting Thresholds by Gender[1]

Gender	Number of InT Reports in Past 5 Years (*n*)	Proportion of InT Reports in Past 5 Years (%)
Female	40	40.0
Male	60	60.0
Total	100	100.0

[1] *This table was created in the Workbook's Excel worksheet.*

Frequencies can also be displayed using bar charts or pie charts as well. Pie charts are used to compare categorical data, such as gender, types of reporting thresholds, and job categories. The size of each slice is proportional to the relative size of each category out of the whole. Pie charts can be represented by labels indicating their relative (percentage) or absolute size (count or summary statistic). A bar graph is a chart that represents categorical data with rectangular bars. The height or length of each bar is proportional to the frequency count of the category it represents. Figure 1 displays a pie chart and a bar chart of the same frequency data presented in Table 1 where the pie chart (on the left) and bar chart (on the right) show the percent of security alerts for males and females. Pie charts are useful for showing the relationship of a variable across different parts to the whole, but pie charts are less effective for comparing exact numbers across different groups or populations.

Figure 1: Types of Graphs Showing the Percentage of InT Reports by Gender[1]



[1] This figure was created in the Workbook's Excel worksheet.

Cross Tabulation

A cross tabulation (or crosstabs) is a matrixed summary of different combinations of independent and dependent variables. It is used to display the appropriate categories of interest in the data by two or more categorical variables. It shows the frequency of different combinations of categories. A Chi-square test (see section below) uses the cross tabulated data (a contingency table) to test whether there is a statistically significant[5] relationship between the variables in the table. Table 2 is a contingency table showing the frequency of males and females across five types of reporting thresholds, totaling 8 cells[6]. You can calculate row totals, column totals, and grand total as shown in Table 2. For example, out of the 100 people in this sample, 40 females had 22 personal conduct issues, while 60 males had 29 personal conduct issues.

Table 2: Crosstabs on Simulated Data on Gender by Reporting Thresholds[1]

Gender	Personal Conduct	Criminal Conduct	Unauthorized Disclosure	Serious Threat	Total (*n*)
Female	22	4	4	10	**40**
Male	29	14	9	8	**60**
Total (n)	**51**	**18**	**13**	**18**	**100**

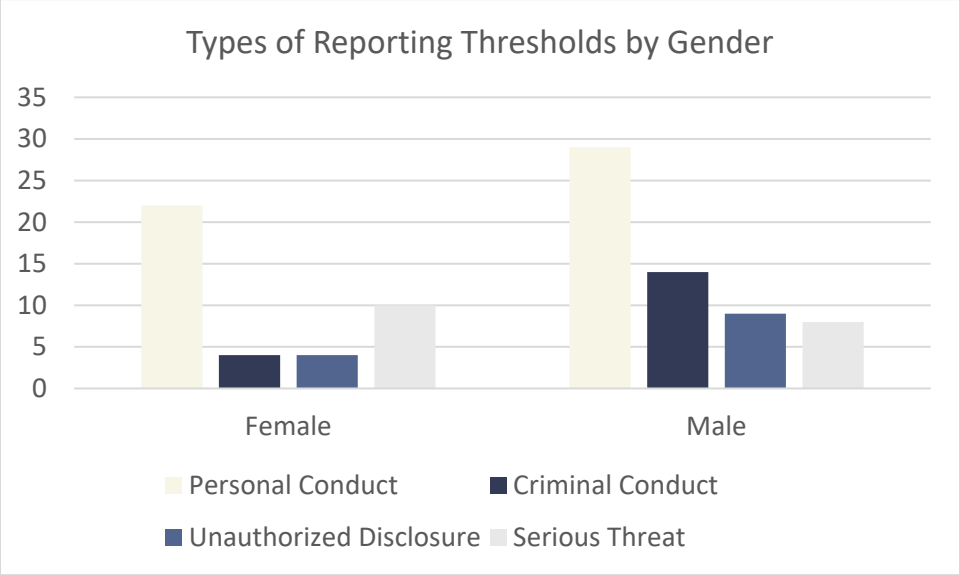
[1.] *This table was created in the Workbook's Excel worksheet.*

Data from contingency tables could also be represented as a bar graph. Bar graphs are also used to compare different categories and show the relationship between them. For example, Figure 2 below represented the data from Table 2. That is, it shows the frequency for each type of reporting threshold for males and females.

[5] Statistical significance is a determination about the null hypothesis, which suggests the results are due to chance alone. In other words, a null hypothesis is a statement that there is no effect or relationship between two variables. A p-value, which is the probability of obtaining a test statistic at least as extreme as the one that was actually observed (assuming that the null hypothesis is true), of less than 0.05 is often used to indicate statistical significance.

[6] A data cell is the intersection of a row and column; it shows the number of observations that belong to both categories.

Figure 2: Bar Chart of Frequency of the Types of Reporting Thresholds Across Gender[1]



[1.] *This figure was created in the Workbook's Excel worksheet.*

Chi-Square Significance Testing

A chi-square test (denoted χ^2) is a statistical test that assesses whether differences between frequency distributions[7] across categorical variables are statistically significant. Chi-square uses a statistical test called the goodness of fit[8] between expected and observed results[9]. The chi-square goodness of fit test is used to test whether the frequency distribution of a single categorical variable is different from expectations. For example, you can use this test to see if the number (or proportion) of employees with a reportable incident across different installations is equal or not. The chi-square test is also used to test whether two categorical variables are related to each other. You can use this test to see if there is a relationship between gender and types of reporting thresholds. The simulated data set provided in the

[7] A statistical distribution is a way of describing how a set of data or a random variable is spread out. It shows the possible values that the data or the variable can take and how often they occur. For example, if you roll a fair six-sided die many times, you can expect to see each number from 1 to 6 about one-sixth of the time.

[8] Goodness of fit is a statistical measure that determines how well the observed data fits with the model's expected values. It is used to evaluate the accuracy of a regression model or any other statistical model that predicts the dependent variable based on one or more independent (or predictor) variables. The goodness of fit is usually expressed as a percentage (e.g., R^2) or a probability value.

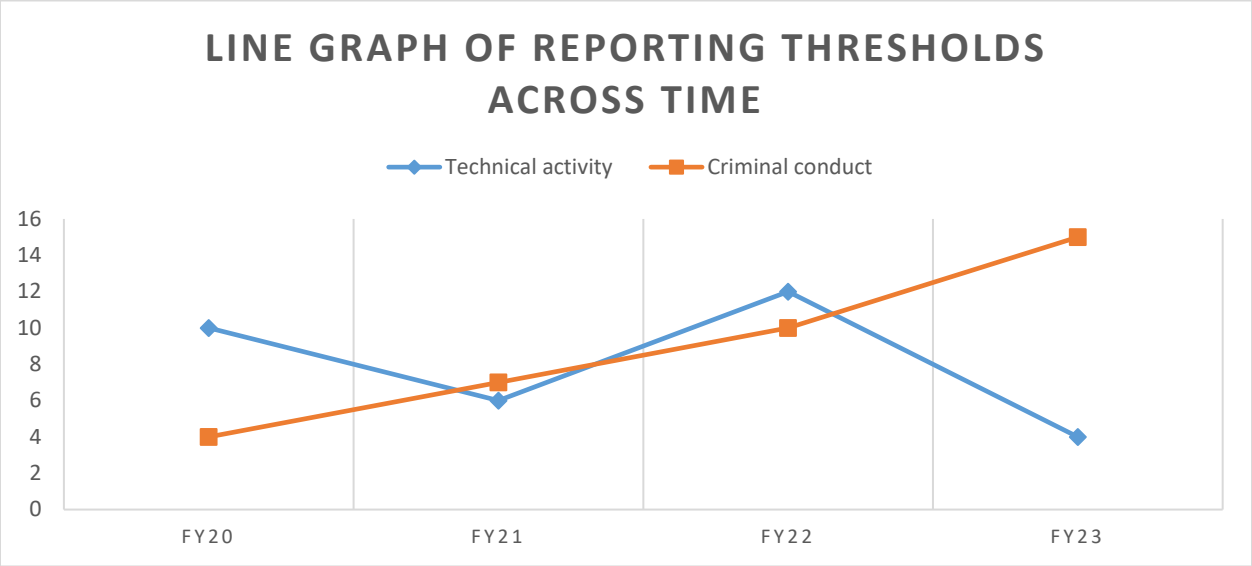
[9] Observed frequencies are the actual of each category in the data. They are obtained from the sample of interest and reflect the reality of the situation. For example, if we want to test whether a coin is fair, we can toss it 100 times and record how many times it lands on heads or tails. The observed frequencies are the number of heads and tails we get from the experiment. Expected frequencies, on the other hand, are the theoretical of each category that we would *expect* to see. They are calculated based on some assumptions or known information and reflect the ideal situation. For example, if we assume that the coin is fair, we expect to see 50 heads and 50 tails out of 100 tosses.

Workbook Appendix can calculate the chi-square statistic and there is a detailed description of this calculation.

Trend Analysis

Trend analysis involves analyzing changes or patterns in the data over time. Trend analysis helps us to identify trends, cycles, seasonality, or outliers in the data. Trend analysis can be presented graphically, or with statistical tests. A line graph or scatterplot can be used to show the trend over time. For example, in Figure 3, the x-axis represents time (or fiscal year), while the y-axis represents the frequency for each of the two selected reporting thresholds[10]. The line connects the data points, which represent the values of the variable at different points in time. Upon initial visual inspection of the two plotted lines, there is a clear linear trend for an increase in criminal activity across time, while there is no clear pattern or trend for technical activity across time.

Figure 3: Preselected Reporting Thresholds across Fiscal Years FY20–FY23[1]



[1.] *This figure was created in the Workbook's Excel worksheet.*

Statistics can also be used to describe trends. For example, if the trend is assumed to be linear, regression analysis may be used to determine whether there is a statistically significant trend in the data. The most basic and common trend analysis uses linear regression, which is a statistical technique that can be used to model the relationship between two or more variables. Regression analysis is a very useful and popular tool in statistics and requires more in-depth knowledge about the actual

[10] The x and y axis are two important lines that make a graph. A graph consists of a horizontal axis (or a "x-axis") and a vertical axis (or a "y-axis") where data can be represented. There are different types of graphs that can be created using the x- and y-axis, such as line graphs, bar graphs, and scatter plots. Each type of graph has its own purpose and advantages for displaying data. For example, a line graph can show how the numerical value of a variable (y-axis) changes over time (x-axis).

computations and assumptions that go into any regression model[11], which is not covered in this Workbook.

Overview of Qualitative Data Analyses

Qualitative data analysis is the process of analyzing and interpreting qualitative data, which are nonnumeric, conceptual, and often based on user feedback in the form of written responses. Qualitative analysis aims to answer the 'why,' 'what,' and 'how' questions, and to identify patterns, themes, and explanations. It typically involves collecting data from interviews, open-ended responses in surveys, narratives/documents, audio, and video. Researchers typically use this method to discover why a group of people have a particular opinion or experience life in a certain way, by studying their motivations, emotions and behaviors.

Qualitative analysis may be supported by content coding of narrative or other qualitative inputs. Coding refers to the process by which labels or tags are assigned to segments of your data to represent meaningful concepts or categories. Determining the type of coding methods is essential in any qualitative analysis. Qualitative analysis can be divided into different categories, such as text analysis and content analysis, which are described in more detail below.

Text Analysis

Text analysis is the process of parsing texts in order to extract machine-readable facts from them. This type of analysis is intended to utilize data out of free text content, such as words, sentences, paragraphs, and documents; it can help identify main topics, themes, concepts, opinions, sentiments, emotions, and intentions of texts. Text analysis can also help discover hidden patterns, relationships, associations, and trends in large collections of texts. Text analysis can help classify text into polarity (positive or negative) states or topics (such as data breach types or organizational responses) that could provide useful information for insider threat detection and mitigation. For example, text analysis can help determine the sentiment of insiders towards their organization or coworkers using text data, or the topic of data breaches such as theft, sabotage, fraud, or espionage. Text analysis can be done manually or with software tools that use natural language processing (NLP) techniques[12]. However, because NLP is an automated process, it is important to note that NLP is not able to parse out context and meaning due to the complexity and nuance of human language[13].

Content Analysis

Content analysis is a qualitative analysis method that involves categorizing and classifying speech, written text, interviews, images, or other forms of communication. Content analysis is used to

---

[11] For more information on this topic please engage with additional material outside of this Workbook, such as https://dss.princeton.edu/online_help/analysis/regression_intro.htm.

[12] NLP is a branch of artificial intelligence that deals with understanding and generating natural language. NLP tools can perform various tasks related to text analysis.

[13] For more detailed information on text mining, please review this peer-reviewed article, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2217579/

determine the meaning, purpose, or effect of any type of communication by studying and evaluating the details, innuendoes, implications, themes, patterns, and other aspects of the content. This method can be both quantitative (focused on counting and measuring) and qualitative (focused on interpreting and understanding) and can be quite labor intensive as it involves manual coding of all text. For example, content analysis can be used to find relationships and patterns in how concepts are communicated in social media posts to help detect the presence and influence of ideological or political messages, or for mapping case reports into various reporting threshold categories.

Qualitative Coding

Developing a method for coding qualitative data (for both text and content analysis) is critically important. An important framework that is often used in qualitative research is the grounded theory approach. This approach involves identifying a research question and then collecting data from various sources, such as interviews, observations, documents, or artifacts. This approach involves comparing data within and across categories of information sources to identify patterns, themes, concepts, and relationships.

Qualitative coders are required to review the data to construct meaningful categories and iterate through the coded data to ensure the categories are mutually exclusive, or at least to the extent possible. This thematic approach ensures that the qualitative data falls within a single category. However, it is possible for a single data source to merge into multiple themes, especially when that data source is unstructured and complex (e.g., social media posts; see Rose & Hesse, 2015).

Once the data is coded and categorized, the data can be grouped into broader themes if needed. The best qualitative coding approach requires familiarization with the data to capture the essence of the content for generating themes. Many qualitative analyses are of an exploratory nature, in which case the coding approach emphasizes the identification of emerging themes and how they may relate to the existing scientific literature on the research topic. A coding method that is often used in grounded theory research is the three-stage coding process of open coding, axial coding, and selective coding (see Chun et al., 2019 and Qureshi & Ünlü, 2020 for more details on this approach).

PART II: Analysis of Real-Case Scenarios

This section of the Workbook demonstrates how to construct your research questions and offers “hands on” experience conducting selected statistical analyses and visualization techniques. While this section is intended to generally guide you on how to construct your own research questions (based on your current data holdings), we include and address three example research questions for demonstration purposes.

We recommended readers first review Part I of this Workbook to gain a basic understanding and vocabulary to engage with the material in Part II of the Workbook. The data used for this demonstration is fictitious data intended for educational purposes only and does not represent real circumstances. The simulated data set along with the functions and graphics for this section is available upon request (please contact: DODHRA.THREATLAB@MAIL.MIL).

Determine your Research Questions

As described in Part I of this Workbook, the first step to conducting an analysis is to define your research questions. You might ask whether there are research questions or topics that your InT hub is interested in exploring. For example, when a senior leader requests information or reports, the first step to take is to determine what question is being asked, what information is required to explore the topic of interest, and how information should be conveyed. For example, executive leaders might be interested in whether or not there are gender differences in the number of InT reports and the types of reporting thresholds. Based on this information, you can determine the research question(s) of the intended analysis. If you are reporting the number of specific InT reports by gender and reporting thresholds, consider the best method to organize and analyze your data to effectively deliver such information, which we provide in this section.

Datasets, Variables, and Inclusion/Exclusion Criteria

Once you identify your objectives, determine if the dataset[14] you are using contains the information you need to address the objective. For example, you may be using a dataset developed from cases retrieved from a InT case management system. Inspect the dataset to learn if it contains all the variables of interest required to address the objective, whether there is missing data, and then determine what information is available in order to address the objective. For our example, you will need information on gender as well as the InT reports and the types of reporting thresholds. Make sure you accurately define the variables you intend to analyze. *What is meant by reporting thresholds? Is the interest in identifying all possible reporting thresholds or just a subset?* Clearly define what reporting thresholds across gender represents in the actual data as this will impact your results.

Inclusion/Exclusion Criteria

Another important step to consider when preparing to conduct any analysis is who to include and exclude in your data analysis. Inclusion and exclusion criteria are characteristics that define the eligible and ineligible population for a research study. Inclusion criteria specify the necessary attributes of the prospective subjects, while exclusion criteria disqualify those who do not meet the requirements. Inclusion and exclusion criteria can be based on many different factors, such as age, gender, diagnosis, intervention, and outcome. Inclusion and exclusion criteria are important because they help researchers study the needs of a relatively homogeneous group[15] with precision and to ensure the internal validity[16]

---

[14] A dataset in statistics is a collection of data that is usually organized in a table, where each column represents a variable, and each row represents a record or an observation. A dataset can contain numerical, categorical, or mixed types of data, depending on the nature and purpose of the analysis.

[15] A homogeneous group within a study population is a group of individuals who share one or more characteristics that are relevant for the research question. For example, a homogeneous group could be composed of people of the same age, gender, ethnicity, diagnosis, or exposure to an intervention. Homogeneous groups are useful for research studies because they allow researchers to examine the specific effects of a variable or an intervention on a subpopulation that is more similar and less variable than the overall population.

[16] Internal validity refers to how well a study can establish a causal relationship between the variables of interest. It means that the results of the study are not influenced by other factors or explanations that could affect the outcome. For example, if researchers want to test whether a new drug can cure a disease, they need to make sure

and generalizability[17] of the study. The inclusion and exclusion criteria are often defined by the researcher, but may also be defined by stakeholders. They are determined after the research question is developed but before conducting the study or analysis. For the purposes of our practice analysis, we will only focus on the population of those with an InT report between FY20–FY23 and exclude all other InT reports outside of this time frame. Soe additional examples of inclusions/exclusion criteria could be based on gender, military component, types of reporting thresholds, etc.

Creating and Merging a Dataset

In some cases, the information needed to answer a research question may be contained in more than one dataset. In this instance, you can create a new dataset by merging different datasets together. To do this correctly, first you need to make sure that the different datasets represent the same sample or population of interest. Next, make sure each individual has an assigned unique identifier (such as a participant code) to make sure the rows from the different datasets align. The unique identifier should be the same for the same person in each of the different datasets. Within the DoD, the DoD ID number is a common unique identifier; we strongly recommend avoiding the use of social security numbers or only using them if there is no other option pursuant to DODI 1000.30 (see section below for more about handling this type of data). Once all cases in your dataset have been assigned unique identifiers, you can use statistical software such as R or database software, or Microsoft Access, to merge the datasets.

Handling Personal Identification Information (PII)

If any of the datasets contain Personally Identifiable Information (PII), make sure to understand the policies and regulations at your agency on how to handle this type of data, or just exclude it from your analyzable dataset altogether. PII can be used to identify an individual and includes information such as first and last name, DoD ID number, social security number (SSN), date of birth, and home address. To reiterate, data use agreements and internal operating policy define whether PII should be removed from a dataset and a unique subject code used instead. Analysts should retain awareness of organizational, DoD, and Federal rules governing the use of unique identifiers (review, DoD, 2012, for additional guidance). This will allow you to keep track of a subject's data without using personal or private information and therefore protect and preserve any or all privacy laws. For the purposes of this Workbook, all PII was removed in our simulated dataset to further exemplify the importance of safeguarding PII in your data.

---

that the improvement in the patients' health is due to the drug and not something else, such as their diet, lifestyle, or placebo effect. Internal validity is important for scientific research because it allows us to draw accurate and reliable conclusions from the data.

[17] Generalizability in research design refers to the extent to which the findings of a study can be applied to other populations or settings. The generalizability of a study depends on the study design, data collection, and statistical analysis. Generalizability is important because it allows researchers to produce knowledge that can be useful and applicable for a wide range of contexts and people. However, generalizability often comes at a trade-off with internal validity. Therefore, researchers need to balance between achieving high generalizability and high internal validity in their studies.

Practice Analysis Using Simulated Data

As described earlier, for purposes of demonstration, we generated simulated data using variables from the DITMAC case management system, which is contained in the Workbook's Excel worksheet. While the data itself is fictitious, we are still able to talk about real-world scenarios where the research questions could be applied in a variety of situations. The simulated data set is structured so that each variable is represented as a unique column, with each row containing a data point. The simulated dataset covers information on the report number, submission type, fiscal year, date the report was sent, the reporting component, triage level, duty status, clearance eligibility, access level, gender, the location where the incident took place, the type of reporting thresholds, and PRIs. Within the Excel file (available upon request; contact DODHRA.THREATLAB@MAIL.MIL) containing the simulated data set, there are two additional tabs with functions and calculations for creating frequency tables, running crosstabs analyses and trends analyses, which are described in detail within the Appendix of this Workbook.

Data Dictionary

The purpose of developing a data dictionary for a research study is to provide a clear and consistent description of the data elements, such as variables, attributes, values, and units of measure that are used in the study. A data dictionary can help users understand and interpret the data, avoid data inconsistencies and errors, and ensure data quality and validity. A data dictionary can be created in various formats, such as a spreadsheet, a document, or a database. It can be updated and maintained throughout the research lifecycle to reflect any changes or additions to the data. A data dictionary can also be shared with other researchers or users to facilitate data discovery, access, reuse, and citation. Table 3 presents an example data dictionary based on the simulated data set with information on each variable including variable name, a description, and variable type.

Table 3: Data Dictionary for Simulated Dataset[1]

Variable Name	Variables defined in dataset	Variable Description	Variable Type
Case Number	CaseNum	A series of characters that represent a specific case number assigned to an individual. It can range from 1–16 characters long and includes letters, numbers, and special characters.	String
Submission Type	SubType	Indicates if the report is a case or a request for information	Categorical
Fiscal Year	FY	Represents the fiscal year for the security alert. Created based off the Date variable. For example, FY20 represents a report that falls within the date range Oct. 1, 2020 – Sept. 30, 2021.	Categorical

Variable Name	Variables defined in dataset	Variable Description	Variable Type
Date	Date	The date the report was submitted to DITMAC DSoS data.	Date (mm-dd-YYYY)
Reporting Component	RepComp	The name of the component submitting the case or request for information.	Categorical
Triage Level	TL	Indicator of the severity of the case.	Categorical
Duty Status (CIV/CTR/MIL)	DutyStat	Indicator if the person being reported on is a civilian (CIV), contractor (CTR), or military personnel (MIL).	Categorical
Clearance Eligibility	ClearElig	The type of security clearance held by the individual in the report.	Categorical
Access Level	AccLvl	The type of access held by the individual in the report.	Categorical
Gender	Gender	Identifies which gender is associated with a security alert. There are two variables to select: "Male" or "Female."	Categorical
Incident Location	Incident Loc	The physical location the incident happened.	Categorical
Reporting Thresholds	Report Thresh	The type of reporting threshold met for the security alert.	Categorical
PRI	PRI	The type of PRI assigned to the individual in the report.	Categorical

[1.] *This table was created based on the Workbook's Excel worksheet simulated data.*

Research Question 1a: How Many InT Reports Between FY20–FY23 Belong to Males Compared to Females?

Let's say you are interested in looking at whether there is a difference in the frequency of InT reports between men and women (who meet reporting thresholds). To begin answering this question, you may first conduct a frequency analysis to discern the proportion of males and females in your population of InT reports. This will tell you the proportion of Int reports received by gender. You can then graphically depict the ratio of males and females with InT reports using a pie chart and calculate a chi-square test to determine whether there is a statistically significant difference. For the purposes of this demonstration, we assume the ratio of males to females is expected to be 50/50 (but this could be different depending on the known information about the population).

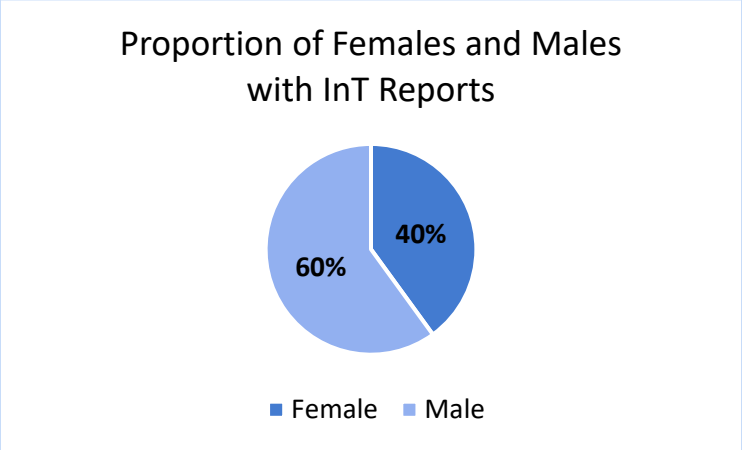
Research Question 1a: Methods

We calculated the frequencies of females and males who met any reporting threshold (i.e., any InT report). We constructed a pie chart using the calculated percentages. Next, we conducted chi-square goodness of fit analysis to test if the number of InT reports was of equal proportion between females and males.

Research Question 1a: Results

Frequency analysis of our sample showed that out of 100 InT reports for FY20–FY23, 40% are attributed to females and 60% are attributed to males. Figure 4 shows the proportion of males and females with a InT report.

Figure 4: Percentage of Security Alerts for Females and Males in FY20–FY23 (*n* = 100)[1]



[1] *This figure was created in the Workbook's Excel worksheet.*

A chi-square goodness of fit test indicated a statistically significant difference in the proportion of InT reports for males compared to females ($\chi^2 = 4$, $df = 1$, $p < 0.05$; reference worksheet for calculations).

Research Question 1a: Interpretation

Given our statistically significant chi-square results (on simulated data), we conclude there is a difference in the proportion of males and females with InT reports. Results show that, between FY20–FY23, a disproportionate number of InT reports involved males (60%) compared to females (40%).

Research Question 1b: Are there specific reporting thresholds that are more frequent in males compared to females?

From Research Question 1a, you were able to report on the percentage of InT reports based on gender and determine whether the percentage of males and females was significantly disproportionate. Research Question 1b requires further detail by incorporating data on the types of reporting thresholds by gender in our simulated dataset. To accomplish this, start by showing the frequency of males and females across the different reporting thresholds. This can be achieved by creating a crosstab, or contingency table that reports on the counts of males and females for each reporting threshold. Then

you can graphically depict the frequencies for males and females across each reporting threshold using a bar chart. The current research question requires us to determine whether or not differences exist between males and females for the different types of reporting thresholds. We will also calculate a chi-square statistic comparing the observed frequencies in each cell with the expected frequencies under the null hypothesis of no relationship between the variables. Below is an example write-up of the methods and results to address this research question.

Research Question 1b: Method

The number of males and females that had an InT report across each of the reporting thresholds in the simulated data included: Personal Conduct, Criminal Conduct, Unauthorized Disclosure, and Serious Threat. The frequencies were calculated and a bar chart was created to display the differences between gender the number of males and females for each reporting threshold. Next, the totals for each of the rows and columns were calculated. Finally, the chi-square statistic was calculated along with the corresponding p-value to determine if there was a relationship among gender and the types of reporting thresholds.

Research Question 1b: Results

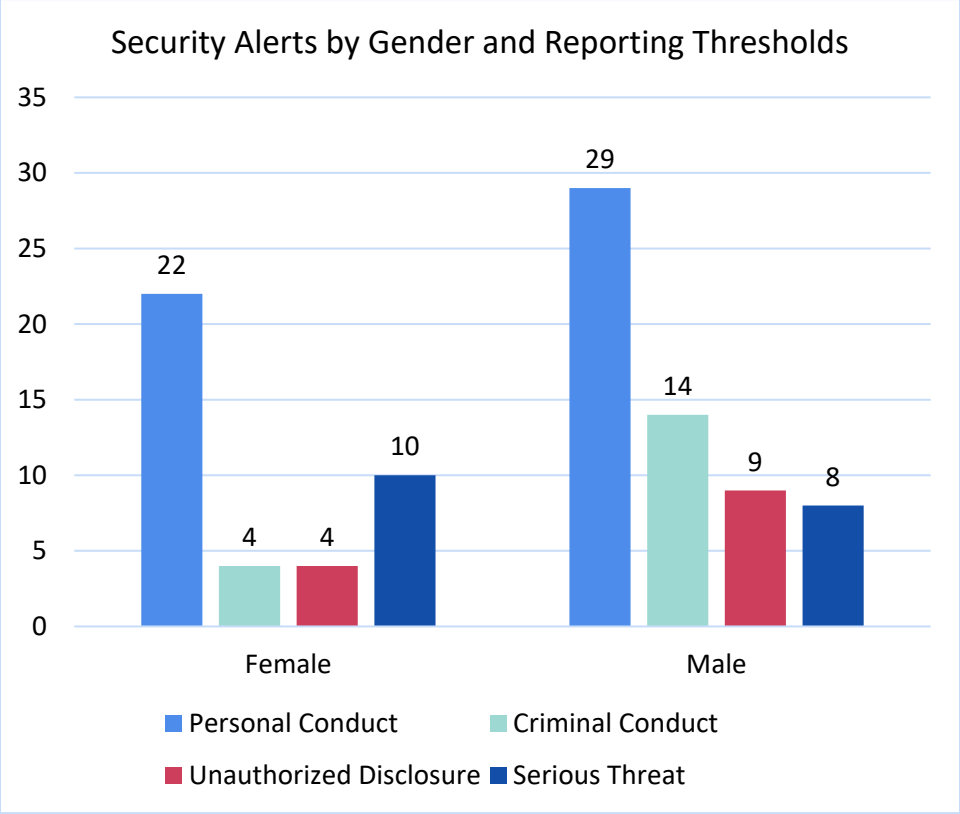
Table 4 displays the number of males and females across each of the different types of reporting thresholds. Figure 5 displays the bar graph for the frequency of males and females across each reporting threshold. In this simulation, the chi-square test showed that there was no statistical relationship between gender and the different types of reporting thresholds ($\chi^2 = 4.86$, $df = 3$, $p = 0.18$).

Table 4: Frequency of Females and Males for each Reporting Threshold

Gender[1]	Personal Conduct	Criminal Conduct	Unauthorized Disclosure	Serious Threat	Total
Female	22	4	4	10	**40**
Male	29	14	9	8	**60**
Total	**51**	**18**	**13**	**18**	**100**

[1] This table was created in the Workbook's Excel worksheet.

Figure 5: Frequency of Reporting Thresholds for Female's and Male's for FY20 – FY23 (*n* = 100)[1]



[1] *This figure was created in the Workbook's Excel worksheet.*

Research Question 1b: Interpretation

Given that we failed to reject the null hypothesis (based on the chi-square results), our results show that there is no statistically significant difference between males and females across each of the reporting thresholds.

Research Question 2: Did Psychological Condition and/or Abusive Conduct Change on an Annual Basis, Specifically Between FY20–FY23?

This research question is focused on looking at how the number of specific reporting thresholds, namely for psychological condition and abusive conduct, changed over a 4-year time period. The best way to assess change over time, is by running a trend analysis to visualize the changes in reporting thresholds of psychological condition and abusive conduct for our predefined time period and determine whether there is a linear trend by fitting a trend line and calculating a R^2 value for each of the reporting thresholds.

Research Question 2: Method

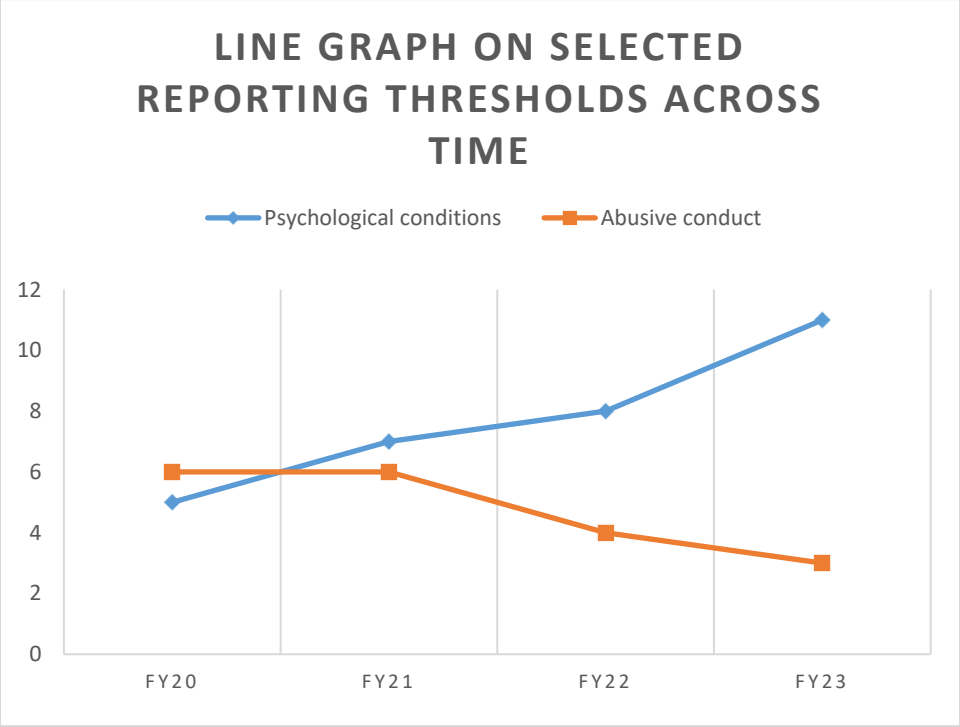
In order to quantify the annual occurrence of Psychological Condition and Abusive Conduct, we constructed a frequency table to display the total number of incidents with a reported PRI of abusive

conduct or psychological condition for each fiscal year. Next, we plotted a line graph of the calculated frequencies for FY20–FY23. We then fitted the linear trend lines for each of the reporting thresholds onto the graph with their corresponding calculated R^2 values to show the general trend and the goodness of fit of each of the trend lines to the actual data.

Research Question 2: Results

Figure 6 displays the data for psychological and abusive conduct reporting thresholds from FY20 through FY23. The number of incidents with a reporting threshold of abusive conduct appeared to decrease over time. However, the number of incidents with a reporting threshold of psychological conditions appeared to increase.

Figure 6: Preselected Reporting Thresholds across Fiscal Years FY20–FY23[1]



1. *This figure was created in the Workbook's Excel worksheet.*

Research Question 2: Interpretation

Within the specified time frame of FY20–FY23, there appears to be a meaningful linear trend, but different for each of the reporting thresholds. For the psychological condition reporting threshold, there is a strong positive linear relationship across time, such that the number of InT reports under the psychological condition threshold increased. In contrast, for the abusive conduct reporting threshold, there appears to be a negative linear relationship, such that the number of InT reports under the abusive conduct threshold decreased over time.

Research Question 3: In the Last 12 Months, What are the Most Common Examples of PRIs in Cases Where the Potential for Workplace Violence is a Concern?

Depending on the type of data available, qualitative and quantitative methods can be used to address this research question. For the purposes of this Workbook, we will address this research question using qualitative and quantitative methods. Below are three simulated narratives describing a set of case reports in the past 12 months, which is all of the information available to us as the researcher. Our goal is to use the methods described in Part I to address this research question. Our focus will be to identify examples of PRIs associated with the potential for "workplace violence."[18]

The Set of Simulated Case Study Narratives for Research Question 3

Case 1: Vanessa Johnson

Vanessa Johnson is a civilian employee of an Army administrative organization housed on an Army base. She mostly keeps to herself in completing her work and has not formed social or personal connections with her coworkers or superiors. She often has conflicts with her colleagues, which has led to multiple heated confrontations in work meetings. Ms. Johnson has written-up after a heated conversation that ended with her saying that Mr. Smith would "get what's coming to him."

Ms. Johnson has also complained about lack of recognition for her performance and for being written up. She has received formal, written warnings after multiple instances of cutting corners and disregarding policies in the interest of meeting deadlines.

Ms. Johnson recently launched a fictional, serialized drama podcast that tells the story of an underappreciated worker who loses her job and then engages in a workplace attack. As the episodes progress, the characters become increasingly transparent stand-ins for Ms. Johnson's real-life coworkers.

[18] Workplace violence can include all violent behavior and threats of violence, as well as any conduct that can result in injury, damage property, induce a sense of fear, and otherwise impede the normal course of work.

Case 2: Joe Wilson

Joe Wilson has worked as a cleared IT Specialist for the Navy for 6 years and was recently passed over for a promotion. Mr. Wilson is a decorated former Navy SEAL who served 10 years before being honorably and medically discharged due to a back injury sustained during combat deployment.

In the past few months Mr. Wilson's behavior has grown more erratic and confrontational, and he has been involved in multiple verbal altercations with colleagues. A heated argument occurred after one of Mr. Wilson's colleagues reported witnessing Mr. Wilson taking his prescribed pain medication in excess.

Another of Mr. Wilson's colleagues reported increasingly threatening, paranoid-sounding, and anti-government social media posts on Mr. Wilson's personal pages. These posts center on Mr. Wilson's ex-wife and his manager. His posts include statements that his manager is monitoring him at home and has also written posts claiming his divorce was caused by the people he works with turning his ex-wife against him. Mr. Wilson posted his ex-wife's personal information online and encouraged people to harass her.

Case 3: Arnold Nolan

Arnold Nolan is a government security professional, working on an Air Force installation. Several months ago, Mr. Nolan lost a child custody dispute. Shortly after, he went to his ex-wife's home where they got into a heated argument outside the house that resulted in his being arrested for verbal assault. Mr. Nolan received a DUI around the same time the assault charges were brought against him.

In the time since losing custody of his children, Mr. Nolan has begun to express displeasure with his work situation. Coworkers mentioned that he had started complaining about the stress that the job put on him and on multiple occasions, has expressed complaints about "the people in charge around here."

Mr. Nolan was reported to Security by one coworker who said that in the last month, Mr. Nolan had repeatedly alluded to a recent mass shooting at a shopping mall. In doing so, Mr. Nolan showed a clear understanding of the details of the event and a level of interest that his coworker found concerning.

Research Question 3: Method

We conducted a content analysis using thematic coding methods, looking for themes of possible risk indicators. Each case was coded, and the relevant themes highlighted and counted by a rater. The number of cases including each identified theme were recorded.

Research Question 3: Results

Table 5 lists the narrative text that was coded for each of the PRIs identified in each case report, along with the frequency of cases they appeared in. Personal conduct was identified in all three cases (100%), while substance misuse and alcohol consumption were found in two (66.6%) of the cases. Psychological conditions and criminal conduct were identified for one case (33.3%).

Table 5: Identified PRIs of Cases Involved in Workplace Violence (*n* = 3)

PRIs	Relevant quotes from narratives	Frequency *n* (%)
Personal conduct	**"**She often has conflicts with her colleagues, which has led to multiple heated confrontations in work meetings."	3 (100)
	"[Mr. Wilson's] been involved in multiple verbal altercations with colleagues."	
	"[Mr. Nolan] has expressed complaints about 'the people in charge around here."	
Substance misuse and Alcohol consumption	**"**Mr. Nolan received a DUI around the same time as the assault charges were brought against him."	2 (66.6)
	"Mr. Wilson's … colleagues reported witnessing Mr. Wilson taking his prescribed pain medication in excess.**"**	
Criminal conduct	"[Mr. Nolan] went to his ex-wife's home where they got into a heated argument outside the house that resulted in his being arrested for verbal assault."	1 (33.3)
Psychological conditions	**"**One of Mr. Wilson's colleagues reported increasingly threatening, paranoid-sounding, and anti-government social media posts on Mr. Wilson's personal pages. These posts center on Mr. Wilson's ex-wife and manager. "	1 (33.3)

Research Question 3: Interpretation

Using thematic coding approach (described in Part I of Workbook), the most common PRIs were identified among three cases involving workplace violence (over the last 12 months). Two of the most frequent PRIs identified in cases involving workplace violence were personal conduct and substance misuse/alcohol consumption. All three cases (100%) had identified incidents involving personal conduct issues such as receiving written reprimands for ignoring policies. Also, two out of the three cases (66%)

identified incidents of substance/alcohol abuse issues. We concluded that the possible risk indicators of personal conduct and substance misuse/alcohol consumption were the most common examples that involve potential cases leading up to workplace violence in the last 12 months.

Closing Remarks

We hope you enjoyed this Workbook and that you learned the fundamental steps to formulating research questions that can be addressed with the data you have available to you and feel confident to conduct the analyses presented in Part II of the Workbook. We also encourage you to reference this Workbook and its supplemental materials in your future efforts to meet the specific needs and requirements of your agency and stakeholders. As mentioned throughout, please reach out to the PERSEREC Project Director, Andrée Rose (DODHRA.THREATLAB@MAIL.MIL), with questions or to request the supplemental material for this Workbook.

Appendix A

Understanding the Built-in Functions in the Simulated Dataset Results

In the simulated dataset, there are three separate tabs (located at the bottom of the worksheet). The three tabs are “Simulated Dataset”, “Chi-Square Tests”, and “Trend Analysis.” The “Simulated Dataset” tab contains the simulated data used throughout this Workbook. The “Chi-Square Tests” and “Trend Analysis” tabs contain the associated tables, figures, and statistical tests conducted in this Workbook, and described in more detail the subsequent sections. It is important to note that the tables and graphs will automatically update if there are any updates or changes to the simulated dataset.

The “Chi-Square” Tab

The “Chi-Square” tab contains the examples on how to construct frequency and contingency tables using Excel commands and formulas in Part I of the Workbook. For example, if calculating the number of InT reports for each gender, the frequency table uses Excel’s “COUNTIF” function. The “COUNTIF” function works by searching for any instance of a desired value within a specified range in the dataset. When counting the number of females with a security alert, the range including the gender column was defined within the “COUNTIF” function along with the value name (in this case “female”) to search for.

Constructing the contingency tables for the chi-square uses a command similar to the frequency tables described above. Instead of using the “COUNTIF” function, we used the “COUNTIFS” function to construct counts of each reporting threshold for males and females. Excel’s “COUNTIFS” function allows for the inclusion of two or more variables of interest. For example, when looking at the number of females with a reporting threshold for Personal Conduct, within the “COUNTIFS” function, the range for the first variable is defined along with the value to search for, then the range for the second variable is defined along with the value to search for in that category. Within our data set, the “COUNTIFS” function is searching for any mention of “female” within the “Gender” variable/column and any mention of “Personal Conduct” within our "Reporting thresholds” column. Then it produces a final count of all instances when both conditions are met. Finally, for the contingency tables, Excel’s “SUM” function was used to calculate the total for each column and each row.

To calculate the summary figures required for the chi-square test, we created two additional tables to perform the calculations required for the chi-square statistic. The table “Crosstabs: Expected Values” computes the expected values using Excel’s built in computation capabilities by multiplying the row total by the column total then divide by the grand total, which was described in more detail in Part I. The table labeled “Crosstabs: Chi-square Computation” will apply the formulas necessary to compute the chi-square statistic. Each of these frequency and crosstabs tables, and their formulas, can be altered to answer different questions, as long as the appropriate adjustments are made to ensure the appropriate variables are being captured within the data set.

Next, we created the graphs for the frequency tables using Excel’s built in chart feature and are located in the Insert tab. To create a graph, highlight the data to graph and select the chart type. For the frequency tables, pie charts are a good option for readers to quickly assess the differences in

proportions between groups. Histograms[19] can be used to show differences in categories between groups, such as the differences in reporting thresholds by gender.

Table 3 shows the data dictionary for this simulated data set, and it serves as documentation on all the variables. It should include information for each variable such as: variable name, variable, variable description, and variable type.

The "Trends Analysis" Tab

This tab contains the same functions as described above to construct frequency table based on the variables in the simulated dataset, "Fiscal Year" and "Reporting Thresholds." This frequency table was then used to construct the line graphs and associated linear fit. Line graphs were used to show change over time, such as changes in frequency of InT reports involving psychological conditions for FY20–FY23. The R^2 value was also calculated and included in the charts, which is also known as the coefficient of determination, and is a measure of how well a linear regression model fits the data. It indicates the proportion of the variance in the dependent variable that is explained by the independent variable. In Excel, you can calculate the R^2 value for a linear fit by using the "RSQ" function in Excel. Alternatively, you can use the Add Trendline option in a scatter plot to display the R R^2 value on the chart, which was used in the worksheet. To do this, follow these steps:

1. Select the scatter plot that contains your data points.
2. Right-click on any data point and choose Add Trendline from the menu.
3. In the Format Trendline pane, select Linear as the trendline type.
4. Check the box that says Display R-squared value on chart.
5. The R^2 value will appear on the upper right corner of the chart.

The R^2 value for a linear fit can range from 0 to 1. A value of 0 means that the model does not explain any of the variation in the dependent variable. A value of 1 means that the model explains all of the variation in the dependent variable. Generally, a higher R^2value indicates a better fit, but it does not imply causation and it may be affected by outliers or other factors.

Appendix B

Legal and Procedural Requirements for Data Collection and Analysis Associated with Human Subjects

When collecting and analyzing data that pertains to human subjects, Federal law and DoD policies and procedures must be understood and correctly applied. Specific concerns include the following:

- Compliance with human subjects research principles
- Adherence to the Privacy Act of 1974 and eGovernment Act of 2002

[19] A histogram is a type of graph that shows the distribution of numerical data. It is made up of bars that represent the frequency or number of observations within different ranges of values. A histogram can help to visualize the shape, spread, and center of the data.

Analysts may be subject to civil and criminal penalties for willfully violating these laws and rules.

Human Subjects Research

Activities related to research involving human subjects, as defined in Part 219 of Title 32, Code of Federal Regulations (CFR), *Protection of Human Subjects*, are guided by the ethical principles set forth in the report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, *Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, (the "Belmont Report") and in compliance with all applicable Federal statutes and regulations (DHRA SOPs, 2019).

Collection and analysis of data about people may be considered human subjects research and may be subject to structured rules about how and for how long data can be collected, analyzed, used, and stored. Before embarking on any data collection effort for research purposes, analysts should work with their component's Human Research Protection Program (HRPP) to determine whether their data collection effort requires approval from an Institutional Review Board (IRB). DoD Instruction 3216.02, *Protection of Human Subjects and Adherence to Ethical Standards in DoD-Conducted and -Supported Research*, "establishes policy, assigns responsibilities, and provides procedures for the protection of human subjects and adherence to ethical standards in DoD-conducted and supported research."

We recommend analysts and scientist complete HRPP and IRB training available through the Collaborative Institutional Training Initiative (CITI).

The Privacy Act of 1974 (Public Law 93-579)

The Privacy Act of 1974 (codified in 5 U.S.C. 5 § 552a) was enacted to protect "the privacy of an individual [who] is directly affected by the collection, maintenance, use, and dissemination of personal information by Federal agencies." As such, the Privacy Act restricts the disclosure of personally identifiable information (PII) and enacts "fair information practices" that govern the collection, maintenance, use, and dissemination of PII.

A system of records (SOR) is a group of records, whatever the storage media (paper, electronic, etc.), under the control of a Federal agency or organization from which personal information about an individual is retrieved by the name of the individual, or by some other identifying number, symbol, or other identifying particular assigned, that is unique to the individual.

The Privacy Act requires DoD publish, in the Federal Register, a system of records notice (SORN) before a new SOR containing PII for individuals (who are citizens of the United States or aliens lawfully admitted for permanent residence) is established. The SORN establishes a 30-day comment period during which affected individuals may voice comments or concerns about the SOR and how data will be collected, stored, and used. Any employee who willfully maintains a SOR without filing a SORN may be subject to criminal penalties and fines.

The information collected for and maintained in your agency's insider threat database must be listed in the appropriate agency SORN. Furthermore, the data collected for your analysis should be listed under "Categories of Records in the System" found within this SORN.

DoD's *Introduction to the Privacy Act*[20], published by the Defense Privacy and Civil Liberties Office (n.d.) details Privacy Act rules and restrictions.

[20] https://dpcld.defense.gov/Portals/49/Documents/Privacy/2011 DPCLO_Intro_Privacy_Act.pdf

References

Anton, P. S., McKernan, M. Munson, K., Kallimani, J. G., Levedahl, A., Blickstein, I., Drezner, J. A., & Newberry, S. J. (2019). *Assessing the use of data analytics in Department of Defense acquisition*. RAND Corporation. https://www.rand.org/pubs/research_briefs/RB10085.html

Bell, M. L., Fiero, M., Horton, N. J., et al. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol* 14, 118 (2014). https://doi.org/10.1186/1471-2288-14-118

Centers for Disease Control and Prevention. (n.d.). *Lesson 4: Displaying public health data*. https://www.cdc.gov/csels/dsepd/ss1978/lesson4/section1.html

Chun Tie, Y., Birks, M., & Francis K. (2019, January 2). *Grounded theory research: A design framework for novice researchers*. SAGE Open Medicine. https://doi.org/10.1177/2050312118822927

Defense Counterintelligence and Security Agency/Center for Development of Security Excellence. (n.d.). *Insider Threat Toolkit*. https://www.cdse.edu/Training/Toolkits/Insider-Threat-Toolkit/

Department of Defense, Defense Privacy and Civil Liberties Division. (2019, March 22). *Department of Defense (DoD) Insider Threat Management and Analysis Center (DITMAC) and DoD component Insider Threat Records System, DUSDI 01 DoD* (84 FR 10803). https://dpcld.defense.gov/Portals/49/Documents/Privacy/SORNs/OSDJS/DUSDI-01-DoD.pdf

Department of Defense, Office of the Inspector General. (2022, September 28). *Audit of the DoD component Insider Threat Reporting to the DoD Insider Threat Management and Analysis Center* (DODIG-2022-141). https://www.dodig.mil/reports.html/Article/3175529/audit-of-the-dod-component-insider-threat-reporting-to-the-dod-insider-threat-m/

Department of Defense. (2012, August 1). *Reduction of social security number (SSN) use within DoD* (DoD Instruction 1000.30; Incorporating Change 2, November 30, 2022). https://dpcld.defense.gov/Portals/49/Documents/Privacy/Memorandum/DODI%201000.30.pdf#:~:text=All%20DoD%20personnel%20shall%20reduce%20or%20eliminate%20the,truncated%2C%20masked%2C%20partially%20masked%2C%20encrypted%2C%20or%20disguised%20SSNs

Department of Defense. (2020). *DoD data strategy*. https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF

Qureshi, H. A., & Ünlü, Z. (2020). *Beyond the paradigm conflicts: A four-step coding instrument for grounded theory*. International Journal of Qualitative Methods, *19*. https://doi.org/10.1177/1609406920928188

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation*, Vol. 7, No 4.

Rose, A. E. & Hesse, C. M. (2015). *Indicators of Suicide Found on Social Networks: Phase 2.* Monterey, CA: Defense Personnel and Security Research Center/Defense Manpower Data Center.

United States Government Accountability Office. (2015, June 2). *Insider threats: DOD should strengthen management and guidance to protect classified information and systems* (GAO-15-544). https://www.gao.gov/products/gao-15-544

Zimmerman, R., Friedman, G., Munshi, D., Richmond, D., & Jaros, S. (2018). *Modeling insider threat from the inside and outside: Individual and environmental factors examined using event history analysis*. Defense Personnel and Security Research Center/Office of People Analytics. https://www.dhra.mil/Portals/52/Documents/perserec/reports/TR-18-14_Modeling_Insider_Threat_From_the_Inside_and_Outside.pdf